IMPERIAL

Department of Mathematics
Imperial College
180 Queen's Gate, London SW7 2AZ

Rank-based hypothesis testing in the presence of missing data

Yijin Zeng

February 19, 2026

A thesis submitted to Imperial College London
for the degree of Doctor of Philosophy (PhD)

Statement of originality

I certify that the research presented in this thesis is my own PhD research carried out under the supervision of Dr Dean Bodenham and Prof. Niall Adams. The results presented here have not been previously submitted to Imperial College London or any other academic institution for a degree or diploma, or any other certification.

Signed: Yijin Zeng Date: 26 September 2025

Copyright declaration

The copyright of this thesis rests with the author. Unless otherwise indicated, its contents are licensed under a Creative Commons Attribution-Non Commercial 4.0 International Licence (CC BY-NC).

Under this licence, you may copy and redistribute the material in any medium or format. You may also create and distribute modified versions of the work. This is on the condition that: you credit the author and do not use it, or any derivative works, for a commercial purpose.

When reusing or sharing this work, ensure you make the licence terms clear to others by naming the licence and linking to the licence text. Where a work has been adapted, you should indicate that the work has been changed and describe those changes.

Please seek permission from the copyright holder for uses of this work that are not included in this licence or permitted under UK Copyright Law.

Abstract

Nearly all hypothesis testing methods are designed solely for data that are completely observed. Despite the prevalence of missing data problems, no standard approach exists to address this issue. This thesis proposes several rank-based hypothesis testing methods for distinct, univariate data in the presence of missing data, without making any assumptions of the missing data.

The importance and difficulty of addressing the missing data problem properly motivate the research in this thesis. Standard hypothesis testing methods cannot be implemented directly in the presence of missing data. Common missing data methods, such as case deletion and imputation methods, risk increasing the Type I error beyond a specific significance level. Meanwhile, the vast literature on this topic addresses this issue only under restrictive assumptions by constraining the missing data to specific missingness mechanisms.

The key contributions of this thesis are the development of several rank-based hypothesis testing methods, which, unlike existing methods, can accommodate the missing values in a manner without making any assumptions of the missing data. We prove that our proposed methods control the Type I error regardless of the values of the missing data. Simulation results demonstrate that these methods have good statistical power, typically when less than 15–20% of the data are missing, while all the other methods fail to control the Type I error.

Acknowledgements

First and foremost, I would like to sincerely thank my supervisors, Dean Bodenham and Niall Adams, for their supervision over the past four years. Thank you for the countless meetings, talks, and emails. Your guidance went beyond my PhD, and your patience has made this work possible.

I am very grateful to my parents and my family for their love and encouragement. Your faith in me kept me motivated throughout my studies. Your understanding and help made it possible for me to focus on this work. I enjoyed our phone time every weekend, and I apologize for all the missed calls while I was busy.

I am also very fortunate to have spent the last four years with my partner, Yidi. You understand me and kept me entertained during my PhD. Our times walking together in parks, and sometimes busy streets in London, are precious memories, and I cannot imagine a better time than the one I have with you. I was stressed sometimes during my PhD, but thinking of you makes me happy and relaxed.

I have been very fortunate to share an office with a group of incredible colleagues and friends, including Victor, Alex, Michael, Maja, Jixin, and many others. Thank you for all the lunches in the SCR, breaks on the Queen's Lawn, and countless discussions about all kinds of questions and topics. I have always enjoyed our time together and learned a great deal from you. You were the motivation for me to come into the office, and the reason I enjoyed my PhD.

Finally, I am grateful to Imperial College London and the StatML CDT for funding my studies over the last four years.

Contents

1. Introduction **10**

2. Background **14**
2.1. Hypothesis testing . . . 14
2.1.1. Two-sample hypothesis testing . . . 15
2.1.2. Independence testing . . . 16
2.1.3. Type I error rate and statistical power . . . 16
2.1.4. Significance level and p-value . . . 18
2.2. Rank-based hypothesis testing methods . . . 20
2.2.1. Wilcoxon-Mann-Whitney test . . . 20
2.2.2. Ansari-Bradley test . . . 21
2.2.3. Spearman's footrule independence test . . . 22
2.2.4. Other rank-based test statistics and testing methods . . . 23
2.3. Missing data mechanisms and missing data methods . . . 25
2.3.1. Missingness mechanisms . . . 25
2.3.2. Common missing data methods . . . 26
2.3.3. Relevant work . . . 28

3. Location test with arbitrarily missing data **29**
3.1. Bounds of the Wilcoxon-Mann-Whitney test statistic . . . 29
3.2. Wilcoxon-Mann-Whitney test with missing data . . . 31
3.2.1. Rejection conditions to decide the significance . . . 31
3.2.2. p-value of the proposed method . . . 33
3.3. Maximum allowable proportion of missing data . . . 37
3.4. Power analysis when data are missing completely at random . . . 39
3.5. Simulations for the location test . . . 43
3.5.1. As the proportion of missing data varies . . . 43
3.5.2. As the sample size varies . . . 46
3.5.3. As the mean value varies . . . 48
3.6. Application to clinical trial data . . . 50
3.7. Conclusion . . . 52

4. Scale test with arbitrarily missing data **53**
4.1. Bounds of the Ansari-Bradley test statistic . . . 54
4.1.1. Intuition for the bounds and main results . . . 54
4.1.2. Definitions and notation . . . 54
4.1.3. Main results and sketch of the proof . . . 55

4.2. Ansari-Bradley test with missing data . . . 57
4.3. Simulations for the scale test . . . 58
4.3.1. As the proportion of missing data varies . . . 58
4.3.2. As the sample size varies . . . 60
4.3.3. As the scale value varies . . . 62
4.4. Extension to location-scale test with missing data . . . 64
4.4.1. Evaluation of the location-scale test without missing data . . . 67
4.5. Simulations for the location-scale test . . . 69
4.5.1. As the proportion of missing data varies . . . 69
4.5.2. As the sample size varies . . . 71
4.5.3. As the location and scale values vary . . . 72
4.6. Application to hepatitis C virus data . . . 74
4.7. Conclusion . . . 76

5. Independence test with arbitrarily missing data 77
5.1. Bounds of Spearman's footrule . . . 78
5.1.1. Intuition for the bounds and main results . . . 78
5.1.2. Definitions and notation . . . 79
5.1.3. Main results for the lower bounds and sketch of the proof . . . 80
5.1.4. Main results for the upper bounds and sketch of the proof . . . 81
5.2. Algorithms for computing the bounds . . . 83
5.2.1. Algorithms for the tight lower bound . . . 83
5.2.2. Algorithms for the tight upper bound . . . 84
5.3. Bounds of Kendall's τ . . . 84
5.4. Simulations for the bounds . . . 85
5.4.1. The bounds when data are missing completely at random . . . 85
5.4.2. The bounds when data are missing not at random . . . 87
5.5. Spearman's footrule independence test with missing data . . . 88
5.6. Simulations for the independence test . . . 90
5.6.1. As the proportion of missing data varies . . . 91
5.6.2. As the sample size varies . . . 93
5.6.3. As the correlation coefficient varies . . . 93
5.7. Application to precipitation data . . . 94
5.8. Conclusion . . . 98

6. Extensions of the location test with missing data 100
6.1. Location test in the presence of missing data with tied observations . . . 101
6.1.1. Bounds of the Wilcoxon-Mann-Whitney statistic with ties . . . 101
6.1.2. Bounds of the p-value with ties . . . 103
6.1.3. Simulations for the proposed location test with ties and closed support . . . 105
6.2. A variant of the location test method with missing data . . . 109
6.2.1. Bounding the test statistic as a random variable . . . 109
6.2.2. p-value of the proposed variant method . . . 112

6.2.3. Simulations for the variant location method 113
6.3. Conclusion 117

7. Conclusion **118**

A. Proofs of results in Chapter 3 **121**
A.1. Bounds of sum of ranks 121
A.2. Maximum allowable proportion of missing data 125
A.3. Power for data missing completely at random 127
A.4. Asymptotic power for data missing completely at random 129

B. Proofs of results in Chapter 4 **136**
B.1. Lower bound with single value missing 136
B.2. Lower bound with multiple missing values in X 147
B.3. Lower bound with multiple missing values in Y 167
B.4. Lower and upper bounds under general missingness case 176

C. Proofs of results in Chapter 5 **184**
C.1. Different missingness cases 184
C.2. Some important lemmas 186
C.3. Tight lower bounds of Spearman's footrule 189
C.3.1. Lower bound under missing case I 189
C.3.2. Lower bound under missing case II 210
C.3.3. Lower bound under missing case III 257
C.3.4. Lower bound under general missingness case 258
C.4. Tight upper bound of Spearman's footrule 264
C.4.1. Upper bound under missing case I 264
C.4.2. Upper bound under missing case II 279
C.4.3. Upper bound under missing case III 280
C.4.4. Upper bound under general missingness case 290
C.5. Efficient algorithms for calculating tight upper bound 295
C.5.1. Algorithm for tight upper bound under missing case I 295
C.5.2. Algorithm for tight upper bound under missing case II 303
C.5.3. Algorithm for tight upper bound under missing case III 311
C.5.4. Algorithm for tight upper bound under general missingness case 321

D. Proofs of results in Chapter 6 **338**
D.1. Bounds of the Wilcoxon-Mann-Whitney statistic with ties 338
D.2. Issue of controlling p-value with ties 343
D.3. Bounds of variance 346
D.4. Bounds of the p-value 349
D.5. Bounds of the Wilcoxon-Mann-Whitney statistic as a random variable 352
D.6. p-value of the variant method 355

E. Additional simulation results **358**
E.1. Additional simulation results for Chapter 3 . . . 358
E.1.1. Additional results for data missing completely at random . . . 358
E.1.2. Additional results for data missing not at random . . . 360
E.1.3. Gamma-distributed data . . . 361
E.1.4. A different missingness mechanism . . . 363
E.2. Additional simulation results for Chapter 4 . . . 364
E.2.1. Scale testing for data missing completely at random . . . 364
E.2.2. Scale testing for data missing not at random . . . 366
E.2.3. Scale testing for Gamma-distributed data . . . 367
E.2.4. Scale testing for a different missingness mechanism . . . 369
E.2.5. Location-scale testing for data missing completely at random . . . 370
E.2.6. Location-scale testing for data missing not at random . . . 372
E.2.7. Location-scale testing for Gamma-distributed data . . . 374
E.2.8. Location-scale testing for a different missingness mechanism . . . 375
E.3. Additional simulation results for Chapter 5 . . . 376
E.3.1. Bounds with missing data . . . 376
E.3.2. As the proportion of missing data varies . . . 381
E.3.3. As the sample size varies . . . 392
E.3.4. A different missingness mechanism . . . 397
E.3.5. Gamma-distributed data . . . 400

Bibliography **402**

Glossary

$D(X, Y)$ Spearman's footrule distance between two samples X and Y. 22

$L(X, Y)$ Lepage test statistic between two samples X and Y. 23

$R(x, Z)$ Rank of a univariate value $x \in \mathbb{R}$ in sample Z. 20

$R_{X|Z}$ Sum of ranks of values in sample X considered in sample Z. 20

$T(X, Y)$ Ansari-Bradley test statistic between two samples X and Y. 22

$W(X, Y)$ Wilcoxon-Mann-Whitney test statistic between two samples X and Y. 21

$\Phi(x)$ Cumulative distribution function of a standard normal distribution. 31

$\mathcal{R}_\alpha$ Rejection region with a significance level α. 18

$\mathrm{I}(A)$ Indicator function such that $\mathrm{I}(A) = 1$ if A is true; otherwise $\mathrm{I}(A) = 0$. 20

$\rho_{\mathsf{Scale}}(X, Y)$ Spearman's ρ correlation between two samples X and Y. 23

$\tau_{\mathsf{Scale}}(X, Y)$ Kendall's τ correlation between two samples X and Y. 24

$p(t(x, y))$ p-value of a test statistic $t(x, y)$. 19

i.i.d Independent and identically distributed. 15

MAR Missing at random. 26

MCAR Missing completely at random. 25

MNAR Missing not at random. 26

1. Introduction

The problem of univariate, nonparametric hypothesis testing with missing values has not received enough attention given its importance. In general, standard hypothesis testing methods do not take missing data into account, and the testing results cannot be determined directly when a subset of the data is missing.

In certain cases, it is possible to make assumptions for the missing data, and the testing result might then be taken as valid. For example, in clinical trials, it frequently occurs that participants drop out before the trial concludes, resulting in missing data for those participants. Several analyses of clinical trial data (Feldman et al., 1993; Margulies et al., 2016) have dealt with missing values by assigning them the worst-possible ranks, or by first assuming they have the worst possible values and then converting these values to ranks, as Lachin (1999) sought to formalise.

However, assumptions for missing data are usually empirical and untestable using the observed data alone. Moreover, it is not always possible to make plausible assumptions; the missing data assumptions are often made primarily to facilitate reaching testing conclusions, without any accompanying analysis (Mathur, 2023), or based mainly on their simplicity to use (Staudt et al., 2022).

When the missing data mechanisms are unclear or too complicated to be analysed, common practices are to ignore all missing values or impute them with plausible values, following which the data are treated as if complete and standard hypothesis testing methods can be applied. However, with only a few exceptions, such as when the data are missing randomly, i.e. each data point have the same probability to be missing, these practices are often invalid as they risk increasing the Type I error rate beyond a specified significance level. To mitigate the problems of these methods, some analyses, such as that in Bakris et al. (2015), consider multiple missing data methods, including ignoring missing data and imputing missing data using different values, and conclude the testing result with more confidence since all these methods agree with each other.

One common misconception when dealing with missing data is that a small proportion of missing data may be safely ignored, with some suggesting such a threshold to be 5% (Heymans and Twisk, 2022; Schafer, 1999) and others indicating 10% (Bennett, 2001). While such rules of thumb might be reasonable in certain applied settings, such as survey research when the data are known to be missing completely at random, we show that for hypothesis testing problems, even with 1% of the data missing, the results can be invalid if the missing data are not taken into account properly when the data are missing not at random, as illustrated by Figure 3.5 in this thesis.

There is no standard approach to hypothesis testing in the presence of missing data. This thesis is therefore dedicated to the development of hypothesis testing methods that can accommodate arbitrary missing data. We only assume the data to be univariate and distinct, and make no

assumptions about the missing data. The key contribution of our methods is that they always control the Type I error regardless of the values of the missing data, while retaining reasonable statistical power when less than 15–20% of the data are missing.

The core idea of the proposed methods is to consider all possible imputations of the missing data, and reject the null hypothesis only when all possible imputed test statistics lead to significant results. For continuous-valued data, considering all possible imputations is an intractable problem. The approaches taken in this thesis are to consider all possible imputed *rank statistics*, rather than the values of imputations directly. By considering ranks, our problem could then be solved by performing a combinatorial search to evaluate the significance level of each possible rank statistic. However, directly evaluating all possible rank statistics is computationally intractable even for moderate sample sizes.

The methods proposed in this thesis *avoid* the combinatorial search by constructing mathematically tight bounds of the test statistics in the presence of missing data. The bounds are tight, meaning that they cannot be improved based on the observed data alone. Based on the bounds of the test statistics, the bounds of the p-value of the test statistics are derived. The proposed methods then reject the null hypothesis when all possible p-values are below a significance level. In other words, when all possible test statistics are significant. We prove that the proposed methods control the Type I error regardless of the values of the missing data. To the best of our knowledge, no other methods in the literature address the same problem directly as we do in this thesis.

The cost of always controlling the Type I error is that our methods become conservative when a large proportion of data are missing. In fact, we theoretically show that one of our methods will never reject the null hypothesis when more than 30% of the data are missing from both samples, regardless of the sample sizes or alternatives.

While this result may seem undesirable, we argue that when no assumptions about the missing data can be made, our methods are optimal in the sense that the bounds are tight; thus, they cannot be improved given the observed data alone. When assumptions about the missing data can be made, it is possible to make the bounds tighter and achieve more powerful testing methods. Section 6.1 considers the additional assumption that the missing data are restricted to a bounded space, and shows that when this assumption is included, our methods have greater test power.

Furthermore, when the proportion of missing data is moderate, e.g., 10%, our methods retain good statistical power. This is empirically shown by various numerical simulation results in this thesis. We observe that our methods often have power close to 1 given moderate sample sizes and differences between the two samples. On the other hand, all other common missing data methods fail to control the Type I error even when only small portion of the data is missing. We show that the Type I error of these methods can asymptotically converge to 1 as sample sizes increase, even when only 1% of the data are missing.

Finally, our methods can also be interpreted as a sensitivity analysis (Cook and Zea, 2020b; Thabane et al., 2013) that considers all possible assumptions on the missing data. As mentioned earlier, our methods take all possible imputations into account by providing bounds of the p-value. Hence, if a significant result is obtained by our methods, then this result is consistent among all possible missing data assumptions, since every possible imputation leads to a significant result. If, however, our methods fail to reject the null hypothesis, there are two possible outcomes: either

all possible missing data assumptions lead to insignificant results when the lower bound of the p-value is greater than α, or a subset of assumptions leads to significant results when the lower bound of the p-value is less than or equal to α, but the upper bound is larger than α. Hence, our methods provide a theoretically rigorous answer to whether different approaches to handling missing data lead to different conclusions.

Description of the thesis

Chapter 2 provides background required for this thesis, including hypothesis testing and several rank-based testing methods. In addition, the missing data problems are discussed following the framework of Rubin (1976), which divides the missingness mechanisms into three cases, namely missing completely at random, missing at random, and missing not at random. Common missing data methods, such as case deletion and mean imputation are also discussed. Relevant work for hypothesis testing in the presence of missing data is reviewed at the end of the chapter.

Chapter 3 proposes a novel location testing method in the presence of missing data based on the Wilcoxon-Mann-Whitney test. Tight lower and upper bounds of the Wilcoxon-Mann-Whitney test statistic in the presence of missing data are derived. These bounds are then used to construct the rejection region, and the bounds of the p-value for the proposed method. This chapter analyses the proposed method by deriving the maximum proportion of missing data for which the proposed method can still yield significant results, and studying the statistical power of the proposed method in the special case when the data are missing completely at random. Numerical simulations are performed for evaluating the Type I error and statistical power of the proposed method. Finally, this chapter concludes with a real-world application of the proposed method to clinical trial data, where a subset of the data is missing.

Chapter 4 extends the framework in Chapter 3 to the Ansari-Bradley test for the scale testing problem. Specifically, the tight lower and upper bounds of the Ansari-Bradley test statistic, and corresponding p-values are derived. The proposed method then rejects the null hypothesis only if all possible p-values are smaller than a significance level α. The proposed scale testing method in this chapter is then combined with the proposed location testing method from Chapter 3 using the Holm-Bonferroni method (Holm, 1979) for location-scale testing. Numerical simulations are performed in this chapter, and show similar results to those obtained in Chapter 3. The proposed location-scale testing method is illustrated using a hepatitis C virus dataset.

Chapter 5 considers the independence testing in the presence of missing data based on Spearman's footrule statistic for measuring statistical association. Similar to Chapter 3 and 4, this chapter first derives the tight lower and upper bounds of Spearman's footrule, and then applies these bounds to obtain the bounds of the p-value. Efficient algorithms for obtaining these bounds are provided. Based on the tight bounds of Spearman's footrule, we construct non-trivial bounds for Kendall's τ, and later evaluate these bounds using numerical simulations. Similar numerical experiments to those considered in Chapter 3 and Chapter 4 are conducted in this chapter for evaluating the Type I error and statistical power of the proposed independence testing methods. The proposed method in this chapter is illustrated on a real-world precipitation dataset.

Chapter 6 extends the proposed location testing method from Chapter 3 in two different directions. The first extension considers the proposed location testing method in the case where

data can be tied and are defined on a potentially closed support. The second extension considers a different approach for always controlling the Type I error in the presence of missing data. The simulation results demonstrate that the two extensions considered in this chapter improve the statistical power of the proposed location testing method from Chapter 3.

Chapter 7 concludes this thesis with a summary of contributions, and discusses potential future work.

Appendices provide technical proofs required for this thesis, and include additional simulation results for the proposed methods.

Contributions of the thesis

The work in this thesis has led to several publications and manuscripts that are currently under review.

- Chapter 3 describes the work in Zeng et al. (2024a).
- Chapter 4 presents the work in Zeng et al. (2025c).
- Chapter 5 covers the work in Zeng et al. (2025b).
- Additionally, Zeng et al. (2025a) propose a two-sample testing method for multivariate data to detect any distributional shift based on the Maximum Mean Discrepancy test (Gretton et al., 2012) with the Laplacian kernel, although this work falls outside the scope of this thesis.

Software packages implementing the proposed methods in this thesis are publicly available in R.

- *wmwm* (Zeng et al., 2024b) provides the proposed location test in Chapter 3 and handles ties as discussed in Section 6.1.
- *abwm* (Zeng, 2025a) implements the proposed scale test in Chapter 4.
- *bosfr* (Zeng, 2025b) implements the proposed independence test in Chapter 5.

2. Background

This chapter reviews standard definitions and methodologies used throughout this thesis. Section 2.1 discusses two-sample hypothesis testing, independence testing, and the concepts of Type I error rate and statistical power. Based on these concepts, Section 2.2 further introduces several rank-based hypothesis testing methods, in particular the Wilcoxon-Mann-Whitney test, the Ansari-Bradley test, and Spearman's footrule independence test. As we will discuss in Chapter 3, 4, and 5, these three testing methods are extended to deal with arbitrarily missing data.

Missing data problems are discussed in Section 2.3, following the framework of Rubin (1976), which divides the missingness mechanisms into three cases, namely missing completely at random (MCAR), missing at random (MAR) and missing not at random (MNAR). Section 2.3 also reviews several common missing data methods such as case deletion and mean imputation, and discusses potential issues in applying these methods for hypothesis testing when the data can be missing, in particular when the data are MNAR.

2.1. Hypothesis testing

Hypothesis testing is a fundamental statistical approach used to determine whether the data provide sufficient evidence to reject a given hypothesis. One of the earliest instances of hypothesis testing dates back to 1710, when John Arbuthnot analysed birth records in London (Hacking, 2016). The modern framework of hypothesis testing is primarily attributed to the foundational work of Fisher (1970), and Neyman and Pearson (1933).

The problem of testing a hypothesis arises naturally in many different contexts. For example, in clinical trials, researchers are often interested in determining whether a new drug or a new treatment method improves the health of the treated patients (Bakris et al., 2015; Schwartz et al., 2001). A common objective in machine learning is to assess whether a new model outperforms an existing one (Chen et al., 2022b). Hypothesis testing is widely used in various distinct areas such as marketing (Siroker and Koomen, 2015), psychology (Strack et al., 1988), and finance (Campbell et al., 1998).

In general, hypothesis testing starts by forming two competing hypotheses: the so-called *null hypothesis*, and the *alternative hypothesis*, conventionally denoted by H_0 and H_1, respectively. The null hypothesis is often the statement of "no effect" or "no difference", while the alternative hypothesis is the statement one tries to support using the data. If the testing method decides the data provide sufficient evidence against the null hypothesis H_0, then we have evidence to reject the null hypothesis. If not, we conclude that we cannot reject the null hypothesis, or we fail to reject the null hypothesis.

Almost all hypothesis testing methods assume the data are complete and cannot take missing data into account. This thesis focuses on developing novel two-sample hypothesis testing and

independence hypothesis testing methods in the presence of missing data for *univariate*, *distinct* data in which ties are absent. An extension of one of our methods that can deal with ties is proposed in Section 6.1. However, for all other parts of this thesis, we shall assume the data are distinct.

2.1.1. Two-sample hypothesis testing

The two-sample hypothesis testing problem considers two univariate, distinct samples $X = \{x_1, \ldots, x_n\}$, and $Y = \{y_1, \ldots, y_m\}$ of independent and identically distributed (i.i.d) observations from unknown cumulative distribution functions F and G, respectively.

A common two-sample hypothesis testing problem is to test whether $F = G$, or more formally

$$H_0 : F = G, \text{ versus } H_1 : F \neq G.$$

In certain cases, we assume that the cumulative distribution functions F and G can differ only in location, and the goal is to test whether a difference in location exists. That is, one assumes $G = F(x + \Delta),\ \forall x \in \mathbb{R}$, and tests the hypothesis

$$H_0 : \Delta = 0, \text{ versus } H_1 : \Delta \neq 0. \tag{2.1}$$

This is normally referred to as a two-sample *location* test.

Similarly, one could assume F and G differ only in their scale: $G = F(\Gamma x),\ \forall x \in \mathbb{R}$, and the goal is to test that

$$H_0 : \Gamma = 1, \text{ versus } H_1 : \Gamma \neq 1. \tag{2.2}$$

This is known as a two-sample *scale* test.

A natural extension of the above location test and scale test is to allow F and G to differ both in their location and scale, i.e., $G = F(\Gamma x + \Delta),\ \forall x \in \mathbb{R}$, and test

$$H_0 : \Delta = 0 \text{ and } \Gamma = 1, \text{ versus } H_1 : \Delta \neq 0, \text{ or } \Gamma \neq 1. \tag{2.3}$$

This test is known as a *location-scale* test.

While all the testing problems considered above share the same null hypothesis $H_0 : F = G$, they are distinguished by their alternative hypotheses. In other words, each problem specifies a distinct direction of departure from the null hypothesis. In this thesis, we will focus on the rank-based methods for these problems, since they are non-parametric, and robust to outliers (Lehmann and D'Abrera, 1975). Furthermore, as we will demonstrate in this thesis, rank-based test statistics allow for the derivation of exact bounds in the presence of missing data.

For each testing problem, different rank-based methods would be considered. Informally, if several methods perform similarly under the null hypothesis, the preferred method is the one that most effectively detects the specific departure when it exists. In Section 2.1.3, we formalize these concepts, defining the performance of a test under the null hypothesis as its Type I error rate, and its ability to detect a false null hypothesis as its statistical power.

It is also worth mentioning that the tests defined in (2.1), (2.2), and (2.3) are known as two-

sided tests, where one tests whether the parameters of interest, Δ or Γ, in our notation, are different from the specified values in the null hypothesis, without regard to the direction of the difference. Hence, the parameters under the alternative hypothesis can be either larger or smaller than the specified values.

In contrast, one-sided tests are used to test whether the parameters are larger than the specified values in the null hypothesis, or whether the parameters are smaller than the specified values in the null hypothesis. For instance, the one-sided location test for deciding whether $y_1, \ldots, y_m$ are stochastically larger than $x_1, \ldots, x_n$ can be formulated as:

$$H_0 : \Delta = 0, \text{ versus } H_1 : \Delta > 0.$$

Note that compared to the two-sided location test in (2.1), the alternative hypothesis for the one-sided test does not include cases where $\Delta < 0$.

In this thesis, we focus on two-sided testing problems. However, the results and methodologies developed in this thesis can be adapted in a straightforward manner to the corresponding one-sided scenarios.

2.1.2. Independence testing

The independence testing problem differs from the two-sample testing problem: instead of comparing two samples of independent random observations, the independence testing problem considers a single sample of paired observations

$$(x_1, y_1), \ldots, (x_n, y_n) \overset{\text{i.i.d.}}{\sim} F_{X,Y},$$

where $F_{X,Y}$ is an unknown two-dimensional cumulative distribution function.

Let F and G denote the marginal cumulative distribution functions of $X = \{x_1, \ldots, x_n\}$ and $Y = \{y_1, \ldots, y_n\}$, respectively. The objective of independence testing is to determine whether X and Y are statistically independent. That is, to test the hypothesis

$$H_0 : F_{X,Y} = F \cdot G, \text{ versus } H_1 : F_{X,Y} \neq F \cdot G. \tag{2.4}$$

2.1.3. Type I error rate and statistical power

Prior to discussing specific methods for performing hypothesis testing, we introduce the definitions of Type I error rate and statistical power. These are two fundamental concepts in hypothesis testing and are used throughout this thesis. Thus, we shall define them carefully below.

To start, we clarify the mechanism for a test method to reject the null hypothesis by introducing the concept of rejection region. A hypothesis testing method typically construct a test statistic $t(X, Y) \in \mathbb{R}$. The decision to reject the null hypothesis is governed by a rejection region $\mathcal{R} \subset \mathbb{R}$ such that the null hypothesis is rejected if and only if the the test statistic satisfies $t(X, Y) \in \mathcal{R}$, i.e. falls into the rejection region.

Using the rejection region, we now define the notions of Type I and Type II errors as follows:

Definition 2.1.1 (Type I error)**.** A Type I error occurs when the null hypothesis H_0 is rejected, i.e. $t(X, Y) \in \mathcal{R}$, while H_0 is true.

Definition 2.1.2 (Type II error)**.** A Type II error occurs when the null hypothesis H_0 is not rejected, i.e. $t(X, Y) \notin \mathcal{R}$, even though H_0 is false.

According to the definitions, Type I and Type II errors are defined in the context of a single hypothesis test. In other words, they are used to describe the outcome of a single hypothesis test. On the contrary, the Type I error rate and the statistical power of a test quantify the *probabilities* of this test making a Type I error and a Type II error. To describe these two concepts, it is more convenient to consider $X = \{X_1, \ldots, X_n\}$ and $Y = \{Y_1, \ldots, Y_m\}$ as sets of random variables, rather than observations in the following of this section. First, we need to define the power function as following

Definition 2.1.3 (Power function)**.** Let $\mathcal{D}$ denote the true underlying data generation distributions. For a two sample testing problem, let $\mathcal{D} = (F, G)$ denote the pair of true underlying distributions of $X = \{X_1, \ldots, X_n\}$ and $Y = \{Y_1, \ldots, Y_m\}$. For an independence test, let $\mathcal{D} = F_{X,Y}$ denote the true joint distribution of $(X_1, Y_1), \ldots, (X_n, Y_n)$. Then, consider a hypothesis test defined by a test statistic $t(X, Y)$ and a rejection region $\mathcal{R} \subset \mathbb{R}$. The power function is defined as the probability that the test statistic falls into the rejection region under the distribution $\mathcal{D}$:

$$\beta(\mathcal{D}) = \Pr\left(t(X, Y) \in \mathcal{R} \mid \mathcal{D}\right).$$

Additionally, for a hypothesis testing problem, we separate all possible data generating distributions into null space Θ_0 and alternative space Θ_1 according to its corresponding null and alternative hypothesis. Specifically, we define

Definition 2.1.4 (Null and alternative spaces)**.** Let Ω be the class of all admissible data generating distributions. For a two sample testing problem, $\Omega = \{(F, G) | F, G \in \mathcal{F}\}$, where $\mathcal{F}$ is the class of all continuous cumulative distribution functions. For an independence testing problem, $\Omega = \{F_{X,Y} | F_{X,Y} \in \mathcal{G}\}$, where $\mathcal{G}$ is the class of all bivariate cumulative distribution functions with continuous marginals. The null space Θ_0 is the subset of distributions for which the null hypothesis holds:

$$\Theta_0 = \{\mathcal{D} \in \Omega \mid H_0 \text{ is true under } \mathcal{D}\}.$$

Similarly, the alternative space Θ_1 is defined as:

$$\Theta_1 = \{\mathcal{D} \in \Omega \mid H_1 \text{ is true under } \mathcal{D}\}.$$

Applying the definition of the power function, and the null and alternative spaces, we are now ready to define the Type I error rate and statistical power of a test.

Definition 2.1.5 (Type I error rate)**.** The Type I error rate of a test is the maximum probability of rejecting the null hypothesis when it is true. Formally, it is the supremum of the power function over the null space:

$$\sup_{\mathcal{D} \in \Theta_0} \beta(\mathcal{D})$$

Definition 2.1.6 (Statistical power)**.** The statistical power of a test is the probability of correctly rejecting the null hypothesis when the alternative hypothesis holds. It is given by the value of the power function

$$\beta(\mathcal{D}), \text{ where } \mathcal{D} \in \Theta_1.$$

According to the definitions, the Type I error rate is the maximum probability of rejecting the null hypothesis whenever it holds. Hence, the probability of the testing method rejects the null hypothesis will be controlled by the Type I error rate as long as H_0 holds, regardless of the specific underlying distributions $\mathcal{D}$. In contrast, statistical power depends on the specific data underlying distributions. This means that under the alternative hypothesis H_1, the probability of rejection varies as a function of the specific distributions (F and G for the two-sample problem, or the joint distribution $F_{X,Y}$ for independence testing). For instance, as shown in Table 3.4, the power of our proposed location test increases with the value of the location shift parameter Δ, defined in Equation (2.1).

The Type I error rate and the statistical power quantify the probabilities that a test commits a Type I error or a Type II error, respectively. By convention, the Type I error rate is often simply referred to as the Type I error (without "rate"), and the statistical power is referred to as the power (without "statistical"). To avoid confusion between the Type I error defined in Definition 2.1.1 and the Type I error rate defined in Definition 2.1.5, we adopt the following convention:

- We use the phrase *"making a Type I error"* exclusively to refer to the event of rejecting a true null hypothesis, as defined in Definition 2.1.1.
- References to *"Type I error"*, without the verb "making" will refer to the probability rate defined in Definition 2.1.5.

2.1.4. Significance level and p-value

Last section defines the Type I error and power of a testing method. Ideally, a testing method will reject the null hypothesis H_0 only when H_0 is indeed false and H_1 is correct, and fail to reject the null hypothesis when H_0 is correct. This would then lead to a perfect test method with Type I error equals to 0 and power 1. In practice, however, such a test rarely exists because the distributions of the random samples under the null and alternative hypotheses often overlap, making perfect testing impossible.

Consequently, the objective of a hypothesis testing is to design procedures that maintain a low Type I error while achieving high statistical power. Most hypothesis testing methods accomplish this by maximizing the power while controlling the Type I error under a pre-specified small value α, known as the *significance level*. Informally, the significance level determines the risk of making the Type I error we are willing to take while rejecting the null hypothesis. A larger significance level can often achieve larger power, but also result in higher Type I error. On the other hand, a small significance level means it is less likely to make a Type I error, but could sacrifice the power of the test.

For a given pre-specified significance level α, a rejection region $\mathcal{R}_\alpha$ is chosen such that it controls

the Type I error. This is typically expressed as:

$$\Pr\left(t(X,Y) \in \mathcal{R}_\alpha \mid H_0\right) \leq \alpha.$$

Since the null hypothesis H_0 is composite, the "conditional on H_0" notation implies that the inequality hold for *any* distribution $\mathcal{D} \in \Theta_0$. The rejection region $\mathcal{R}_\alpha$ in this case is determined by the significance level α, and the distribution of the test statistic $t(X,Y)$ under the null hypothesis. Theories for determining $\mathcal{R}_\alpha$ are beyond the scope of this thesis, but they are discussed in Lehmann and Romano (1986, Section 3).

As discussed earlier, the test then rejects the null hypothesis if the test statistic falls within the rejection region $\mathcal{R}_\alpha$; otherwise, the null hypothesis is not rejected. When the null hypothesis is rejected, we also say the test result is *significant* at the significance level α. Consequently, if the null hypothesis is not rejected, we say the test result is not significant at the significance level α.

In this thesis, we will only deal with the rejection region $\mathcal{R}_\alpha$ of the following form

$$\mathcal{R}_\alpha = (-\infty, L_\alpha) \cup (R_\alpha, \infty)$$

where L_α and R_α are such that

$$\Pr\left(t(X,Y) < L_\alpha | H_0\right) \leq \alpha/2, \text{ and } \Pr\left(t(X,Y) > R_\alpha | H_0\right) \leq \alpha/2. \tag{2.5}$$

Hence, the null hypothesis is rejected when the test statistic is smaller than L_α or larger than R_α.

The approach described above, which rejects the null hypothesis when the test statistic falls into the rejection region $\mathcal{R}_\alpha$, yields in only binary outcomes: either the null hypothesis is rejected, or it is not. In other words, the testing result is classified as either significant or not significant at the significance level α.

An equivalent, but more informative testing approach is to first compute the so-called p-value

$$p(t(x,y)) = 2\min\left\{\Pr\left(t(X,Y) \leq t(x,y)|H_0\right), \Pr\left(t(X,Y) \geq t(x,y)|H_0\right)\right\}, \tag{2.6}$$

and then reject the null hypothesis only when the p-value is less than or equal to the significance level α.

By definition, the p-value $p(t(x,y))$ represents the smallest significance level at which the test result would be declared significant. Compared with the rejection region approach, the p-value offers several potential advantages, as discussed by Lehmann and Romano (1986). First, reporting the p-value allows others to apply their own significance thresholds and thereby reach conclusions tailored to their chosen α levels. Second, unlike the binary rejection rule based on the rejection region $\mathcal{R}_\alpha$, the p-value quantifies the strength of evidence against the null hypothesis. In general, a smaller p-value represents stronger evidence against the null hypothesis. Thus, in practice, it is often preferred to report the p-value rather than simply reporting whether the testing result is significant at the significance level α.

We end this section by noting that the rejection region and the p-value, defined in Equation (2.5) and (2.6), respectively, depend on the distribution of the test statistic $t(X,Y)$ under the

null hypothesis. Hence, this thesis often reviews, and studies the distribution of a test statistic under the corresponding null hypothesis. Although it is often possible to obtain test results by using numerical simulation-based approaches, such as permutation tests (Lehmann and Romano, 1986, Chapter 5), these methods are not a primary focus of this thesis.

2.2. Rank-based hypothesis testing methods

This section reviews several rank-based hypothesis testing methods, in particular the Wilcoxon-Mann-Whitney test, the Ansari-Bradley test, and Spearman's footrule independence test. As mentioned previously, the three methods are fundamental components for the proposed methods in this thesis. Other relevant rank-based hypothesis testing methods, such as the Lepage location-scale test, Spearman's rank and Kendall's τ independence test are also discussed.

The concept of rank is central in this thesis. Hence we define it carefully below:

Definition 2.2.1. Suppose Z is a finite set, or vector of n distinct real-valued observations. Then for any $x \in Z$, the rank of x in Z is defined as

$$R(x, Z) = \sum_{z \in Z} \mathrm{I}\left(z \leq x\right),$$

where $\mathrm{I}(A)$ is the indicator function such that $\mathrm{I}\left(A\right) = 1$ if A is true; otherwise $\mathrm{I}\left(A\right) = 0$.

In other words, the rank of x in the set or vector Z with distinct values is the number of elements in Z that are less than or equal to x. For example, consider a set $Z = \{2.5, 3.1, 4.6\}$, then the rank of 3.1 in Z is 2 because 2.5 and 3.1 are both less than or equal to 3.1, while only 4.6 is strictly larger than 3.1.

Note that Definition 2.2.1 requires that Z consists of distinct values. The definition of rank is slightly different when the values in Z can be tied as shown in Definition 6.1.1, and will be considered only in Section 6.1.

2.2.1. Wilcoxon-Mann-Whitney test

Let $X = \{x_1, \cdots, x_n\}$, and $Y = \{y_1, \cdots, y_m\}$ be two samples of distinct real-valued observations. One of the most frequently used methods for two-sample location testing problem, as described in (2.1), is the Wilcoxon-Mann-Whitney test (Mann and Whitney, 1947), also known as the two-sample Wilcoxon test (Wilcoxon, 1992), and the Wilcoxon rank-sum test (Lehmann and D'Abrera, 1975).

The Wilcoxon-Mann-Whitney test starts by computing the Wilcoxon-Mann-Whitney test statistic. To define it formally, we first define the sum of ranks of the values in X as follows.

Definition 2.2.2. Suppose X and Z are two sets of univariate values, with $X \subset Z$. Then $R_{X|Z}$ denotes the sum of the ranks of the values in X, where the ranking is considered in Z, i.e.,

$$R_{X|Z} = \sum_{x \in X} R(x, Z).$$

Given the Definition 2.2.2, the Wilcoxon-Mann-Whitney test statistic is defined as

Definition 2.2.3. For samples X and Y, the Wilcoxon-Mann-Whitney test statistic $W(X, Y)$ is defined as

$$W(X, Y) = R_{X|X\cup Y} - n(n+1)/2.$$

To apply the Wilcoxon-Mann-Whitney test statistic for hypothesis testing, the distribution of the test statistic $W(X, Y)$ under the null hypothesis is required, as discussed in Section 2.1.4.

Under the null hypothesis that X and Y are independent random samples from identical continuous distribution, the exact distribution of $W(X, Y)$ can be computed recursively as demonstrated in Mann and Whitney (1947). For large sample sizes, calculating the exact distribution could become computationally impractical and the normal approximation to the exact distribution is often used.

Mann and Whitney (1947) proved that the distribution of $W(X, Y)$ approximately follows a normal distribution with mean $nm/2$ and variance $nm(n+m+1)/12$ under the null hypothesis. The approximation becomes more accurate when the sample sizes n, m get larger. One rule of thumb is to use the normal approximation when both n and m are larger than 50 (R Core Team, 2024).

2.2.2. Ansari-Bradley test

As before, denote $X = \{x_1, \ldots, x_n\}$ and $Y = \{y_1, \ldots, y_m\}$ as two samples of distinct real-valued observations of sizes n and m, respectively. One common method for testing scale difference between X and Y as in (2.2) is the Ansari-Bradley test, proposed by Ansari and Bradley (1960).

Similar to the Wilcoxon-Mann-Whitney test, the Ansari-Bradley test is also a rank-based testing method. To define it properly, let us denote $Z = X \cup Y$, $N = n + m$, and

$$\tilde{N} = \lfloor (N+1)/2 \rfloor,$$

where $\lfloor (N+1)/2 \rfloor$ is the largest integer less than or equal to $(N+1)/2$. The Ansari-Bradley test statistic is then defined as follows by Ansari and Bradley (1960):

$$\sum_{i=1}^{n} \left\{ \mathrm{I}\left(R(x_i, Z) \leq \tilde{N}\right) R(x_i, Z) + \mathrm{I}\left(R(x_i, Z) > \tilde{N}\right) \left(N + 1 - R(x_i, Z)\right) \right\}.$$

An equivalent, but perhaps more intuitive definition of the Ansari-Bradley test statistic is used by Lepage (1971):

$$\frac{n(N+1)}{2} - \sum_{i=1}^{n} \left| R(x_i, Z) - \frac{1}{2}(N+1) \right|.$$

This definition shows that the Ansari-Bradley test statistic measures the sum of absolute deviation of $R(x_i, Z)$ from $(N+1)/2$, over all $i = 1, \ldots, n$.

In this thesis, we take the following definition of Ansari-Bradley test statistic for notation convenience:

Definition 2.2.4. Suppose $X = \{x_1, \ldots, x_n\}$ and $Y = \{y_1, \ldots, y_m\}$ are two sets of univariate,

distinct values. Then the Ansari-Bradley test statistic $T(X, Y)$ is defined as

$$T(X,Y) = \sum_{i=1}^{n} \left| R(x_i, X \cup Y) - \frac{1}{2}(n+m+1) \right|.$$

The distribution of $T(X, Y)$ is studied by Ansari and Bradley (1960). They provide recursive equations for computing the exact distribution of $T(X, Y)$. Moreover, they show that, under the null hypothesis, $T(X, Y)$ converges in distribution to a normal distribution with mean

$$\mu = \begin{cases} nN/4 & \text{when } N \text{ is even,} \\ n(N^2-1)/(4N) & \text{when } N \text{ is odd,} \end{cases} \tag{2.7}$$

and variance

$$\sigma^2 = \begin{cases} nm(N^2-4)/\{48(N-1)\} & \text{when } N \text{ is even,} \\ nm(N+1)(N^2+3)/(48N^2) & \text{when } N \text{ is odd.} \end{cases} \tag{2.8}$$

Fahoome and Sawilowsky (2000) suggest that when both n, m are larger than 15, the above normal approximation can be used for testing with the significance level $\alpha = 0.05$; while for a $\alpha = 0.01$, both n, m should be larger than 29. The R Core Team (2024) applies the normal approximation when both $n, m \geq 50$.

2.2.3. Spearman's footrule independence test

As discussed in Section 2.1.2, the independence test considers one paired sample, rather than two samples of observations. Consequently, it is more convenient to consider X and Y as vectors, rather than sets. Spearman's footrule is a rank-based statistical association measure, and can be used for independence testing.

In order to define Spearman's footrule, we first introduce a few useful notation. For any n-dimensional vector X, let $X(i)$ denote the component of X at index $i \in [n]$. Thus, an n-dimensional vector X can be written as $X = (X(1), \ldots, X(n))$. Define $\mathcal{V}^n$ to be the set of n-dimensional vectors with distinct real values, i.e. if $X \in \mathcal{V}^n$ then $X \in \mathbb{R}$ and all components of X are distinct.

Using above notation, Spearman's footrule is defined as

Definition 2.2.5. Suppose $X, Y \in \mathcal{V}^n$. Spearman's footrule between X and Y is defined as

$$D(X,Y) = \sum_{i=1}^{n} |R(X(i), X) - R(Y(i), Y)|.$$

Spearman's footrule $D(X, Y)$ was proposed by Spearman (1906) as an easy but precise measure of association between two vectors X and Y. Spearman (1906) describes the method as analogous to using one's foot to measure distance, highlighting its intuitive and straightforward nature, thus the name "footrule."

To apply Spearman's footrule $D(X, Y)$ for independence testing, the distribution of $D(X, Y)$ under the null hypothesis is required. The exact distribution of Spearman's footrule can be

computed following the algorithm proposed by Salama and Quade (2002). The approximation of the exact distribution is studied by Diaconis and Graham (1977). They show that under the null hypothesis, the distribution of $D(X,Y)$ converges in distribution to a normal distribution with mean and variance equal to $(n+1)(n-1)/3$ and $(n+1)(2n^2+7)/45$, respectively (Ury and Kleinecke, 1979).

Salama and Quade (1990) study the accuracy of the normal approximation. Their results show that when the sample size of X and Y is at least 40, the normal approximation performs reasonably well: the difference in skewness between the exact and normal approximation distribution of Spearman's footrule is -0.075 and the difference in kurtosis is -0.074.

2.2.4. Other rank-based test statistics and testing methods

In this section, we further review several rank-based testing methods that are relevant to the methods we discussed in Section 2.2.1, 2.2.2, and 2.2.3.

Lepage test

Lepage test (Lepage, 1971) is a rank-based two-sample hypothesis testing method for the location-scale testing problem, as discussed in (2.3).

The Lepage test first computes the Lepage test statistic, which is the sum of squared standardised Wilcoxon-Mann-Whitney test statistic defined in Definition 2.2.3, and squared standardised Ansari-Bradley test statistic defined in Definition 2.2.4. To be more precise, the Lepage test statistic is defined as

$$L(X,Y) = \left(\frac{W(X,Y)-\mu_w}{\sigma_w}\right)^2 + \left(\frac{T(X,Y)-\mu_t}{\sigma_t}\right)^2,$$

where μ_w and σ_w^2 are the mean and variance of the Wilcoxon-Mann-Whitney test statistic $W(X,Y)$ under the null; μ_t and σ_t^2 are the mean and variance of the Ansari-Bradley test statistic $T(X,Y)$ under the null.

Spearman's rank correlation

Spearman's rank correlation (Spearman, 1906) is a rank-based measure of association between two paired vectors X and Y. The definition of Spearman's rank correlation is based on the squared distance between ranks. To define it formally, let us first denote the squared distance between ranks as

$$\rho(X,Y) = \sum_{i=1}^{n} |R(X(i),X) - R(Y(i),Y)|^2. \tag{2.9}$$

Then scaling $\rho(X,Y)$ between $[-1,1]$ leads to the standard definition of Spearman's rank correlation

$$\rho_{\text{Scale}}(X,Y) = 1 - \rho(X,Y)/\{n(n^2-1)\}.$$

Note that the $\rho(X, Y)$ defined above is closely related to the Spearman's footrule $D(X, Y)$ defined in Definition 2.2.5:

$$D(X, Y) = \sum_{i=1}^{n} |R(X(i), X) - R(Y(i), Y)|.$$

While Spearman's footrule is based on the absolute differences between ranks, Spearman's rank correlation relies on squared differences. These two approaches have been compared by Spearman (1906), Berry and Mielke (1997), and Chen et al. (2023).

Kendall's τ correlation coefficient

Consider again paired values $(X(1), Y(1)), \ldots, (X(n), Y(n))$. Another commonly used measure of association between X and Y is known as Kendall's τ correlation coefficient (Kendall, 1948).

Kendall's τ correlation coefficient is also based on ranks. However, unlike Spearman's footrule and Spearman's rank correlation, which are defined based on the sum of rank differences between corresponding elements of X and Y, Kendall's τ correlation coefficient is based on the concept of the so-called discordance.

More formally, any two pairs $(X(i), Y(i))$ and $(X(j), Y(j))$, where $i < j$, are said to be concordant if either one of the following two conditions (i), (ii) is true:

$$(i) : X(i) < X(j), \text{ and } Y(i) < Y(j),$$
$$(ii) : X(i) > X(j), \text{ and } Y(i) > Y(j).$$

Otherwise the two pairs are said to be discordant. The Kendall's τ correlation coefficient is then computed by first counting the total number of discordant pairs:

$$\tau(X, Y) = \sum_{i>j} \{\mathrm{I}(X(i) > X(j))\,\mathrm{I}(Y(i) < Y(j)) + \mathrm{I}(X(i) < X(j))\,\mathrm{I}(Y(i) > Y(j))\}, \tag{2.10}$$

and then scaling $\tau(X, Y)$ between $[-1, 1]$:

$$\tau_{\text{Scale}}(X, Y) = 1 - 4\tau(X, Y)/\{n(n-1)\}.$$

Kendall's τ is closely related to Spearman's footrule. Diaconis and Graham (1977, Theorem 2) proved that they differ by at most a constant factor, as shown by the following result.

Theorem 2.2.6 (Diaconis and Graham (1977))**.** Suppose $X, Y \in \mathcal{V}^n$. Then, the values of Spearman's footrule and Kendall's τ between X and Y are such that

$$\tau(X, Y) \leq D(X, Y) \leq 2\tau(X, Y).$$

Other rank-based testing methods

Besides the methods enumerated above, there are other rank-based testing methods for location, scale, or independence testing problems, although they less relevant to the proposed methods in the thesis. Notable examples include the Van der Waerden (der Waerden, 1952) and median

tests (Brown and Mood, 1951) for location; the Mood test (Mood, 1954) for scale; the Cucconi test (Cucconi, 1968) for location-scale differences; and Hoeffding's D (Hoeffding, 1994) for independence.

In the thesis, we mainly focus on the Wilcoxon-Mann-Whitney, Ansari-Bradley, and Spearman's footrule statistics and extend these methods to the cases where the data might be missing. While it is possible to extend other rank-based methods for incorporating missing data following the framework in the thesis, we choose these three test statistics since they have good statistical power (Genest et al., 2010; Lehmann and D'Abrera, 1975; Luigi and Nikitin, 1999), and we can prove their exact mathematical bounds in the presence of missing data as we will show in the following chapters. While it is possible to derive the bounds for other rank-based test statistics such as Kendall's τ, these derivations are non-trivial and can be considered as future work.

2.3. Missing data mechanisms and missing data methods

In this section, we discuss the missing data problem following the framework of Rubin (1976), where the missingness mechanisms are divided into three cases according to whether the probability of missing depends on observed or missing data. Common missing data methods, such as case deletion and mean imputation, are also discussed in this section.

2.3.1. Missingness mechanisms

A framework about missingness mechanisms is founded by Rubin (1976), where the missingness mechanisms are divided into three cases, namely missing completely at random (MCAR), missing at random (MAR), and missing not at random (MNAR).

Following Rubin (1976), consider a vector of univariate real-valued observations $z = (z_1, \cdots, z_N)$. Let $\iota = (\iota_1, \cdots, \iota_N)$ be an indicator function of z such that ι_i takes value 1 if z_i is missing, and 0 if z_i is observed. The core idea of Rubin (1976) is to admit ι as a probabilistic phenomenon (Schafer and Graham, 2002). In other words, admitting ι is a realised value of a random vector I, although we may not be able to explicitly specify the distribution of I.

The definitions of MCAR, MAR, and MNAR are based on whether the distribution of missingness mechanisms I depends on the observed data or the missing data.

Missing completely at random

Suppose z is realised values of a random vector Z. Let $f_\theta(\iota|z)$ denote the probability of ι given $Z = z$, where θ denotes any unknown parameters of the distribution. Then, the missingness mechanism is missing completely at random (MCAR) if

$$f_\theta(\iota|z) = f_\theta(\iota|\widetilde{z}), \forall \iota, z, \widetilde{z}.$$

In such cases, the missingness mechanism I is independent of the values of missing and observed data. The implication of MCAR is that whether an observation is missing is completely random, and does not depend on the values of the data in z.

Statistical inferences, such as hypothesis testing, can often be justified when the data are

MCAR if we first ignore all the missing values. However, this assumption is often considered to be too strict, and is rarely satisfied in practice (Schafer and Graham, 2002).

Missing at random

We now consider the case when data are missing at random. Denote z' as a sub-vector of z including all observed data in z, i.e. including all z_i such that $\iota_i = 0$. Then, the missingness mechanism is missing at random (MAR) if

$$f_\theta(\iota|z) = f_\theta(\iota|\widetilde{z}), \forall \iota, z, \widetilde{z} \text{ such that } z' = \widetilde{z}'.$$

In other words, the missingness mechanism I is independent of the values of missing data, but can depend on the observed values z'.

By definition, MAR is a generalization of MCAR; that is, while all MCAR data satisfy the MAR condition, the converse is not necessarily true. While MAR mechanisms are frequently modeled in multivariate data, such as in Cheung (2005) and Lee (1997), constructing a scenario where data are MAR but not MCAR is neither trivial nor standard in the univariate setting as considered in this thesis. To the best of our knowledge, no existing literature addresses MAR specifically for univariate data. Consequently, we restrict our scope in this thesis to missing data that are either MCAR or missing not at random (MNAR), defined as below.

Missing not at random

If the missingness mechanism is neither MCAR nor MAR, the missingness mechanism is defined as MNAR. This mechanism suggests that the probability of missing depends on the missing values. In this case, valid statistical inference cannot be made based on the observed data only, and the missingness mechanism has to be specified by the researcher (Dong and Peng, 2013).

In this thesis, we perform numerical simulations and show that when the data are MNAR, common missing data methods such as case deletion, mean imputation and hot deck imputation are invalid for hypothesis testing, as they can lead to the Type I error exceeding the given significance level.

When dealing with missing data, another challenge is to identify which missingness mechanisms the data hold. For multivariate data, it is possible to test whether the data are MCAR (Berrett, 2024; Bordino and Berrett, 2025; Little, 1988) However, it is impossible to distinguish between MAR and MNAR using the observed data alone (White et al., 2011). This problem becomes even more challenging when the data are univariate. To the best of our knowledge, there is no testing method to distinguish between MCAR, MAR and MNAR when the data are univariate.

2.3.2. Common missing data methods

We now review several different common missing data methods. As mentioned previously, these methods usually assume the data are missing according to specific missingness mechanisms, such as missing completely at random (MCAR), or missing not at random (MNAR).

Table 2.1.: Illustrative example of case deletion, mean imputation, and hot deck imputation methods. The complete data for samples X and Y are shown first. In the observed data, a subset of X and Y is unobserved (denoted by *). Case deletion ignores the missing data. Mean imputation imputes the missing data as the mean of the observed values, while hot deck imputation imputes the missing data by uniformly sampled values from all observed values.

Sample X							
Complete	5.2	3.1	4.6	2.1	2.4	4.3	1.8
Observed	*	3.1	4.6	*	2.4	4.3	*
Case deletion		3.1	4.6		2.4	4.3	
Mean Imp.	**3.6**	3.1	4.6	**3.6**	2.4	4.3	**3.6**
Hot deck Imp.	**4.3**	3.1	4.6	**3.1**	2.4	4.3	**4.3**

Sample Y							
Complete	2.3	3.7	1.1	5.6	7.3	2.0	9.6
Observed	2.3	3.7	*	5.6	*	2.0	9.6
Case deletion	2.3	3.7		5.6		4.3	9.6
Mean Imp.	2.3	3.7	**5.1**	5.6	**5.1**	4.3	9.6
Hot deck Imp.	2.4	3.7	**9.6**	5.6	**2.4**	4.3	9.6

Imp. stands for imputation.

Case deletion

Case deletion is perhaps the most frequently used approach for dealing with missing data. This method ignores all values that are not observed and uses only observed data for testing. As we mentioned in Section 2.3.1, this approach can be justified when data are MCAR. However, when the data are MAR or MNAR, simply ignoring the missing data can lead to invalid testing results as it risks increasing the Type I error above a pre-specified significance level.

Single imputation

Single imputation methods refer to a type of missing data methods where each missing observation is imputed by a single value (Schafer and Graham, 2002). When all missing values are imputed using the mean or median of the observed data, the methods are known to be mean imputation and median imputation, respectively. Another commonly used single imputation method is hot deck imputation, which imputes each missing value with a uniformly sampled observation from all observed values. Table 2.1 provides a simple example for illustrating the case deletion, and the mean, hot deck imputation methods.

Multiple imputation and expectation-maximization algorithm

When the distributions of the data and the missingness mechanism are known, it is possible to apply likelihood-based approaches such as multiple imputation (Rubin, 1996) or expectation-maximization algorithm (Dempster et al., 1977). These methods are generally preferred to the case deletion and single imputation methods, since they often provide more valid results (Dong and Peng, 2013). However, this thesis does not pursue likelihood-based approaches. For overviews of these methods, we refer to the work by Baraldi and Enders (2010); Little and Rubin (2019); Rubin (2003); Rubin and Schenker (1986); Schafer (1999); Schafer and Graham (2002).

Sensitivity analysis

When a set of assumptions about missing data are believed to be appropriate, a sensitivity analysis (Cook and Zea, 2020a; Goldberg et al., 2021; Little and Yau, 1996; Thabane et al., 2013) is often performed to evaluate whether a consensus about the testing results can be established.

If the testing results are consistent across multiple assumptions, the findings are considered robust and can be interpreted with greater confidence. Conversely, if the results differ under alternative assumptions, the conclusion might be considered sensitive to the assumptions and should be interpreted with more caution. For example, in the analysis of clinical trial data with missing observations, Bakris et al. (2015) applied several missing data methods, including mean imputation and hot deck imputation, and rejected the null hypothesis with more confidence as all methods produced consistent, significant results.

2.3.3. Relevant work

Various authors have suggested modifications and refinements to rank-based tests to accommodate missing values, usually when the missingness mechanisms are known. Cheung (2005) and Lee (1997) adapt the Wilcoxon-Mann-Whitney test when the data are assumed MAR, and covariates exist to estimate the probability of each value to be missing. On the other hand, Lachin (1999) considers the Wilcoxon-Mann-Whitney test under a special MNAR case, where the missing data are assumed to have the largest values. In a sequential testing context, calculating the minimum and maximum possible Wilcoxon-Mann-Whitney statistics has been considered by Alling (1963) and Halperin and Ware (1974), under the assumption that the unobserved values are larger than the observed values. A related line of research based on the Wilcoxon-Mann-Whitney test concerns time-censored data (Gehan, 1965; Matsouaka and Betensky, 2015), although this is different to the case of missing values.

Common rank-based measures of statistical association used for independence testing include Spearman's rank correlation and Kendall's τ (Lehmann and D'Abrera, 1975). Although less widely used, Spearman's footrule has been shown to exhibit higher statistical power than Spearman's rank correlation and Kendall's τ under certain alternatives (Chen et al., 2023; Genest et al., 2010; Luigi and Nikitin, 1999).

In the presence of missing data, Alvo and Cabilio (1995) define Spearman's rank correlation and Kendall's τ as the conditional expectations given the observed data, and propose new testing methods under the assumption that incomplete rank vectors are uniformly distributed over all compatible permutations. Extending this approach, Alvo and Charbonneau (1997) adapt Spearman's footrule for trend testing under the same assumption.

Several studies have studied the bounds of Spearman's footrule under different missing data scenarios. However, to the best of our knowledge, none addresses the problem under general missingness patterns as we do. One frequently studied special case is the "top-k lists" problem (Dwork et al., 2001; Fagin et al., 2003; Lin, 2010; Quade and Salama, 2006; Salama and Quade, 2004; Sen et al., 2003), where all missing values are assumed to exceed any observed values, so only the smallest ranks are observed. Papaioannou and Loukas (1984) derive a lower bound of Spearman's footrule when the data are distinct and each pair is either fully observed or missing, and later extend to the cases where ties exist (Loukas and Papaioannou, 1991; Speevak, 2017).

3. Location test with arbitrarily missing data

This chapter presents research work by Zeng et al. (2024a). In this work, we consider the *location* two-sample testing problem described in Section 2.1.1, and propose a new rank-based approach for the testing problem in the presence of missing data without making assumptions about the missingness mechanism. Our approach is a theoretical extension of the Wilcoxon-Mann-Whitney test reviewed in Section 2.2.1 with controlled Type I error by providing exact bounds for the test statistic after accounting for all possible missing values.

We first derive the exact bounds of the Wilcoxon-Mann-Whitney test statistic in Section 3.1, assuming only the data are distinct. Based on the bounds, Section 3.2 proposes a testing approach with controlled Type I error, regardless of the values of missing data. The exact upper bound on the maximum proportion of data that can be missing while allowing the proposed method to yield a significant result is derived in Section 3.3. The power analysis of the proposed method is included in Section 3.4, under the assumption that the data are missing completely at random (MCAR).

Simulation results in Section 3.5 demonstrate that our method has good power, typically for cases of 10–20% missing data, while common missing data approaches such as case deletion, mean imputation, and median imputation fail to control the Type I error. We also illustrate our method on complex clinical trial data in which patients' withdrawal from the trial leads to missing values.

Two extensions of the work in this section are further discussed in Chapter 6, where we consider (1) cases where the data may contain ties and are defined on a potentially closed support, and (2) a different approach for always controlling the Type I error without making missing data assumptions.

Note that the first three sections in this chapter, Section 3.1, 3.2, and 3.3, consider only the observations of samples. Therefore, we use the capital letters, such as X and Y to represent the set of observations. However, in Section 3.4 for power analysis of the proposed method, we work with random variables. Consequently, capital letters in that section denote either individual random variables or sets of random variables.

3.1. Bounds of the Wilcoxon-Mann-Whitney test statistic

This section provides exact bounds of the Wilcoxon-Mann-Whitney test statistic in the presence of missing data when the data are assumed distinct. The bounds are exact, meaning that they cannot be further improved given the data alone.

To start, we prove the bounds of the sum of ranks of X in $X \cup Y$ with missing data.

Proposition 3.1.1. Suppose that $X = \{x_1, \cdots, x_n\}$ and $Y = \{y_1, \cdots, y_m\}$ are samples of distinct, real-valued observations. Suppose that $X' \subset X$ is a subset of n' values in X, and suppose that $Y' \subset Y$ is a subset of m' values in Y. Defining $Z = X \cup Y$ and $Z' = X' \cup Y'$ and supposing only Z' is known, the minimum and maximum possible rank sum of X, over all possible values for the $d = n - n' + m - m'$ observations in $Z \setminus Z'$, are

$$\min_{Z \setminus Z'} R_{X|Z} = R_{X'|Z'} + (n - n')(n + n' + 1)/2, \tag{3.1}$$

$$\max_{Z \setminus Z'} R_{X|Z} = R_{X'|Z'} + \{n(n + 2m + 1) - n'(n' + 2m' + 1)\}/2, \tag{3.2}$$

where $R_{X|Z}$ denotes the sum of ranks of X in Z as defined in Definition 2.2.2.

Proof. The proof is provided in Appendix A.1. □

Remark 3.1.2. As demonstrated in the proof of Proposition 3.1.1, the rank sum $R_{X|Z}$ is *minimized* when all missing values in X take the smallest possible ranks, while all missing values in Y take the largest possible ranks. Conversely, the rank sum is *maximized* when this assignment is reversed: all missing values in X take the largest possible ranks, while all missing values in Y take the smallest possible ranks.

After bounding the sum of ranks of X in Z, the Wilcoxon-Mann-Whitney test statistic can be bounded according to its definition provided in Definition 2.2.3.

Theorem 3.1.3. Suppose that $X = \{x_1, \cdots, x_n\}$ and $Y = \{y_1, \cdots, y_m\}$ are samples of distinct, real-valued observations. Suppose that $X' \subset X$ is a subset of n' values in X, and suppose that $Y' \subset Y$ is a subset of m' values in Y. Defining $Z = X \cup Y$ and $Z' = X' \cup Y'$ and supposing only Z' is known, the minimum and maximum possible Wilcoxon-Mann-Whitney statistic $W(X, Y)$, over all possible values for the $d = n - n' + m - m'$ observations in $Z \setminus Z'$, are

$$\min_{Z \setminus Z'} W(X, Y) = W(X', Y'),$$
$$\max_{Z \setminus Z'} W(X, Y) = W(X', Y') + nm - n'm'.$$

Proof. According to the definition of the Wilcoxon-Mann-Whitney test statistic in Definition 2.2.3, we have $W(X, Y) = R_{X|X \cup Y} - n(n + 1)/2$. Hence,

$$\begin{aligned}
\min_{Z \setminus Z'} W(X, Y) &= \min_{Z \setminus Z'} R_{X|Z} - n(n + 1)/2 \\
&= {}^{\text{Equation (3.1)}}\ R_{X'|Z'} + (n - n')(n + n' + 1)/2 - n(n + 1)/2 \\
&= R_{X'|Z'} - n'(n' + 1)/2 \\
&= W(X', Y').
\end{aligned}$$

Similarly,

$$\begin{aligned}\max_{Z\setminus Z'} W(X,Y) &= \max_{Z\setminus Z'} R_{X|Z} - n(n+1)/2 \\ &= {}^{\text{Equation (3.2)}}\ R_{X'|Z'} + \{n(n+2m+1) - n'(n'+2m'+1)\} - n(n+1)/2 \\ &= R_{X'|Z'} - n'(n'+1)/2 + nm - n'm' \\ &= W(X',Y') + nm - n'm'.\end{aligned}$$

This completes the proof. □

3.2. Wilcoxon-Mann-Whitney test with missing data

This section develops a two-sample testing method in the presence of missing data using the bounds of the Wilcoxon-Mann-Whitney test statistic we derived in Section 3.1. The key idea of our method is to reject the null hypothesis only when all possible test statistics are significant. An equivalent approach, which rejects the null hypothesis by considering the bounds of the p-values of the Wilcoxon-Mann-Whitney test, is also proposed.

3.2.1. Rejection conditions to decide the significance

This subsection considers rejection conditions that allow us to reject the null hypothesis when all possible test statistics are significant. To start, recall that under the null hypothesis H_0, the Wilcoxon-Mann-Whitney test statistic $W(X,Y)$ asymptotically follows a normal distribution with mean $\mu = nm/2$ and variance $\sigma^2 = nm(n+m+1)/12$, as discussed in Section 2.2.1. When the sample sizes n, m are large enough for normal approximation, e.g. both $n, m \geq 50$, the Wilcoxon-Mann-Whitney test rejects the null hypothesis H_0 if

$$(W(X,Y) - \mu)/\sigma \geq \Phi^{-1}(1-\alpha/2) \tag{3.3}$$

$$\text{or } (W(X,Y) - \mu)/\sigma \leq \Phi^{-1}(\alpha/2) \tag{3.4}$$

for any given significance level $0 < \alpha < 1$, where $\Phi(x)$ denotes the cumulative distribution function of the standard normal distribution. Denote the rejection region $\mathcal{R}_\alpha$ with respect to α as

$$\mathcal{R}_\alpha = (-\infty, \Phi^{-1}(\alpha/2)) \cup (\Phi^{-1}(1-\alpha/2), \infty).$$

Then it is equivalent to reject the null hypothesis if the Wilcoxon-Mann-Whitney test statistic $W(X,Y)$ falls into the rejection region, i.e.

$$(W(X,Y) - \mu)/\sigma \in \mathcal{R}_\alpha.$$

Suppose only a subset of data $X' \subset X, Y' \subset Y$ are observed, where either $X' \neq X$ or $Y' \neq Y$. Then $W(X,Y)$ can not be computed directly, and one cannot assess the Inequalities (3.3) or (3.4). However, following Theorem 3.1.3, we can still compute the minimum and maximum

possible Wilcoxon-Mann-Whitney test statistic in the presence of missing data

$$W_{\min}(X,Y) = \min_{Z \setminus Z'} W(X,Y) = W(X',Y'),$$
$$\text{and } W_{\max}(X,Y) = \max_{Z \setminus Z'} W(X,Y) = W(X',Y') + nm - n'm',$$

where $W(X',Y')$ is the Wilcoxon-Mann-Whitney test statistic computed using only observed data X' and Y'.

After computing $W_{\min}(X,Y)$ and $W_{\max}(X,Y)$, we propose to reject the null hypothesis when all possible Wilcoxon-Mann-Whitney test statistics fall into the rejection region $\mathcal{R}_\alpha$. This is done by rejecting the null hypothesis when either one of the following two conditions is true.

Condition 3.2.1.

$$\frac{W_{\min}(X,Y) - \mu}{\sigma} \geq \Phi^{-1}\left(1 - \frac{\alpha}{2}\right),$$

where $\mu = nm/2$ and $\sigma^2 = nm(n+m+1)/12$.

Condition 3.2.2.

$$\frac{W_{\max}(X,Y) - \mu}{\sigma} \leq \Phi^{-1}\left(\frac{\alpha}{2}\right),$$

where $\mu = nm/2$ and $\sigma^2 = nm(n+m+1)/12$.

To verify that all possible test statistics are significant when Condition 3.2.1 or Condition 3.2.2 is true, let W denote any possible Wilcoxon-Mann-Whitney test statistic. Since $W_{\max}(X,Y) \geq W \geq W_{\min}(X,Y)$, and $\Phi^{-1}(x)$ is a monotonic increasing function,

$$\text{Condition 3.2.1} \implies \frac{W - \mu}{\sigma} \geq \Phi^{-1}\left(1 - \frac{\alpha}{2}\right),$$
$$\text{Condition 3.2.2} \implies \frac{W - \mu}{\sigma} \leq \Phi^{-1}\left(\frac{\alpha}{2}\right).$$

Thus, if either Condition 3.2.1 or Condition 3.2.2 is true, any possible Wilcoxon-Mann-Whitney test statistic W must fall into the rejection region $\mathcal{R}_\alpha$.

In particular, the Wilcoxon-Mann-Whitney test statistic $W(X,Y)$, which can be computed if the missing data were fully observed, must fall into the rejection region, i.e.,

$$\text{Condition 3.2.1 or Condition 3.2.2 are true} \implies \frac{W(X,Y) - \mu}{\sigma} \in \mathcal{R}_\alpha.$$

An important consequence is that if the null hypothesis is rejected when either Conditions 3.2.1 or Condition 3.2.2 holds, the Type I error is no larger than α regardless the missing data, since we have

$$\Pr\left(\text{Condition 3.2.1 or Condition 3.2.2 are true}|H_0\right) \leq \Pr\left(\frac{W(X,Y) - \mu}{\sigma} \in \mathcal{R}_\alpha | H_0\right) = \alpha.$$

This result is central to our method. Hence we summarise it more formally as follows:

Theorem 3.2.3. The proposed method, which rejects the null hypothesis whenever Condition 3.2.1 or Condition 3.2.2 holds, has Type I error no larger than the Wilcoxon-Mann-Whitney test when all data are observed.

As discussed in Section 2.2.1, in the absence of missing data, the distribution of the Wilcoxon-Mann-Whitney test statistic can be derived by only requiring the assumption that the data are independent and identically distributed according to a common continuous distribution. Therefore, under this fairly mild assumption, in the absence of missing data, the Type I error of the Wilcoxon-Mann-Whitney test will be controlled.

Furthermore, in the case of discrete data where ties may be present, the exact distribution of the Wilcoxon-Mann-Whitney test statistic can still be computed. When the two samples are sufficiently large (say, $n, m \geq 50$), a normal distribution approximates the true distribution very well. This case with ties will be considered in Section 6.1.1.

3.2.2. p-value of the proposed method

As discussed above, our method rejects the null hypothesis if Condition 3.2.1 or Condition 3.2.2 holds, and otherwise does not reject it. In this section, we present an equivalent but more informative approach: we derive bounds for the p-value and reject the null hypothesis whenever the maximum possible p-value is smaller than the significance level.

To start, the p-value of the Wilcoxon-Mann-Whitney test is given by:

$$p(W(X,Y)) = 2\min\left\{\Phi\left(\frac{W(X,Y)-\mu}{\sigma}\right), 1-\Phi\left(\frac{W(X,Y)-\mu}{\sigma}\right)\right\}. \tag{3.5}$$

In the presence of missing data, while $p(W(X,Y))$ can not be computed directly, one can compute the corresponding p-values of the minimum and maximum Wilcoxon-Mann-Whitney test statistic

$$p(W_{\min}(X,Y)) = 2\min\left\{\Phi\left(\frac{W_{\min}(X,Y)-\mu}{\sigma}\right), 1-\Phi\left(\frac{W_{\min}(X,Y)-\mu}{\sigma}\right)\right\},$$
$$\text{and } p(W_{\max}(X,Y)) = 2\min\left\{\Phi\left(\frac{W_{\max}(X,Y)-\mu}{\sigma}\right), 1-\Phi\left(\frac{W_{\max}(X,Y)-\mu}{\sigma}\right)\right\}.$$

One caveat here is to consider $p(W_{\min}(X,Y))$ and $p(W_{\max}(X,Y))$ as the minimum and the maximum possible p-value of the Wilcoxon-Mann-Whitney test directly. Instead, we have the following result regarding the bounds of the p-values of the Wilcoxon-Mann-Whitney test in the presence of missing data.

Proposition 3.2.4. Suppose that $X = \{x_1, \cdots, x_n\}$ and $Y = \{y_1, \cdots, y_m\}$ are samples of distinct, real-valued observations. Suppose that $X' \subset X$ is a subset of n' values in X, and suppose that $Y' \subset Y$ is a subset of m' values in Y. Defining $Z = X \cup Y$ and $Z' = X' \cup Y'$ and supposing only Z' is known, then the p-value of the Wilcoxon-Mann-Whitney test $p(W(X,Y))$

is bounded such that

$$p(W(X,Y)) \geq \min\{p(W_{\min}(X,Y)), p(W_{\max}(X,Y))\},$$
$$p(W(X,Y)) \leq \begin{cases} 1, & \text{if } (W_{\min}(X,Y)-\mu)(W_{\max}(X,Y)-\mu) \leq 0, \\ \max\{p(W_{\min}(X,Y)), p(W_{\max}(X,Y))\}, & \text{otherwise.} \end{cases}$$

Proof. We first prove that $p(W(X,Y)) \geq \min\{p(W_{\min}(X,Y)), p(W_{\max}(X,Y))\}$. Note that if $W(X,Y) < \mu$, we have

$$\Phi\left(\frac{W(X,Y)-\mu}{\sigma}\right) < \Phi(0) = \frac{1}{2}.$$

Hence,

$$\begin{aligned} p(W(X,Y)) &= 2\min\left\{\Phi\left(\frac{W(X,Y)-\mu}{\sigma}\right), 1-\Phi\left(\frac{W(X,Y)-\mu}{\sigma}\right)\right\} \\ &= 2\Phi\left(\frac{W(X,Y)-\mu}{\sigma}\right). \end{aligned}$$

Since $W_{\min}(X,Y) \leq W(X,Y)$, and $\Phi(x)$ is a monotonic increasing function, we have

$$p(W(X,Y)) \geq 2\Phi\left(\frac{W_{\min}(X,Y)-\mu}{\sigma}\right) \geq p(W_{\min}(X,Y)).$$

Following the same approach, we can show that if $W(X,Y) \geq \mu$, we have

$$p(W(X,Y)) \geq 2\left(1-\Phi\left(\frac{W_{\max}(X,Y)-\mu}{\sigma}\right)\right) \geq p(W_{\max}(X,Y)).$$

Hence, we have proven that

$$p(W(X,Y)) \geq \min\{p(W_{\min}(X,Y)), p(W_{\max}(X,Y))\}.$$

Next, we prove that

$$p(W(X,Y)) \leq \begin{cases} 1, & \text{if } (W_{\min}(X,Y)-\mu)(W_{\max}(X,Y)-\mu) \leq 0, \\ \min\{p(W_{\min}(X,Y)), p(W_{\max}(X,Y))\}, & \text{otherwise.} \end{cases}$$

By definition, we have $p(W(X,Y)) \leq 1$. Hence the result must be true if $(W_{\min}(X,Y) - \mu)(W_{\max}(X,Y)-\mu) \leq 0$.

If, however, $(W_{\min}(X,Y)-\mu)(W_{\max}(X,Y)-\mu) > 0$, then we must have either

$$W_{\min}(X,Y), W_{\max}(X,Y) > \mu,$$
$$\text{or } W_{\min}(X,Y), W_{\max}(X,Y) < \mu.$$

Suppose $W_{\min}(X,Y), W_{\max}(X,Y) > \mu$. Since $W(X,Y) > W_{\min}(X,Y)$, we have $W(X,Y) > \mu$.

Subsequently,

$$\begin{aligned} p(W(X,Y)) &= 2\left(1-\Phi\left(\frac{W(X,Y)-\mu}{\sigma}\right)\right) \\ &\leq 2\left(1-\Phi\left(\frac{W_{\min}(X,Y)-\mu}{\sigma}\right)\right) \\ &= p(W_{\min}(X,Y)) \\ &\leq \max\{p(W_{\min}(X,Y)), p(W_{\max}(X,Y))\}. \end{aligned}$$

Similarly, when $W_{\min}(X,Y), W_{\max}(X,Y) < \mu$, we can use the same approach and prove that

$$p(W(X,Y)) \leq p(W_{\max}(X,Y)) \leq \max\{p(W_{\min}(X,Y)), p(W_{\max}(X,Y))\}.$$

This completes the proof. □

Following Proposition 3.2.4, we denote

$$p_{\min}(X,Y) = \min\{p(W_{\min}(X,Y)), p(W_{\max}(X,Y))\}$$

and

$$p_{\max}(X,Y) = \begin{cases} 1, & \text{if } (W_{\min}(X,Y)-\mu)(W_{\max}(X,Y)-\mu) \leq 0, \\ \max\{p(W_{\min}(X,Y)), p(W_{\max}(X,Y))\}, & \text{otherwise,} \end{cases}$$

as the minimum and maximum possible p-values of the Wilcoxon-Mann-Whitney test statistic in the presence of missing data.

Proposition 3.2.4 provides all possible p-values of the Wilcoxon-Mann-Whitney test when the data are not entirely observed. An equivalent approach of rejecting the null hypothesis when Condition 3.2.1 or Condition 3.2.2 holds is to reject the null hypothesis when the maximum possible value of p-value is smaller than the significance level α, as demonstrated by the following result:

Proposition 3.2.5. Suppose that $X = \{x_1, \cdots, x_n\}$ and $Y = \{y_1, \cdots, y_m\}$ are partially observed samples of distinct, real-valued observations. Then for any given significance level $\alpha < 1$,

$$\text{Condition 3.2.1 or 3.2.2 hold} \iff p_{\max}(X,Y) \leq \alpha,$$

where $p_{\max}(X,Y)$ is defined in Proposition 3.2.4. Hence, the p-value of the proposed location testing method is the maximum possible p-value $p_{\max}(X,Y)$ over all possible values of missing data.

Proof. We first prove that " $\Longrightarrow$ " holds. Suppose Condition 3.2.1 or 3.2.2 holds. Then, if Condition 3.2.1 is true,

$$\begin{aligned} & W_{\min}(X,Y), W_{\max}(X,Y) > \mu \\ \Longrightarrow\ & p_{\max}(X,Y) = \max\{p(W_{\min}(X,Y)), p(W_{\max}(X,Y))\}. \end{aligned}$$

According to Condition 3.2.1, we also have

$$\Phi\left(\frac{W_{\min}(X,Y)-\mu}{\sigma}\right) \geq \Phi\left(\Phi^{-1}\left(1-\frac{\alpha}{2}\right)\right) = 1-\frac{\alpha}{2},$$
$$\Phi\left(\frac{W_{\max}(X,Y)-\mu}{\sigma}\right) \geq \Phi\left(\Phi^{-1}\left(1-\frac{\alpha}{2}\right)\right) = 1-\frac{\alpha}{2}.$$

Hence,

$$\begin{aligned} p(W_{\min}(X,Y)) &= 2\min\left\{\Phi\left(\frac{W_{\min}(X,Y)-\mu}{\sigma}\right), 1-\Phi\left(\frac{W_{\min}(X,Y)-\mu}{\sigma}\right)\right\} \\ &\leq \alpha, \end{aligned}$$

and

$$\begin{aligned} p(W_{\max}(X,Y)) &= 2\min\left\{\Phi\left(\frac{W_{\max}(X,Y)-\mu}{\sigma}\right), 1-\Phi\left(\frac{W_{\max}(X,Y)-\mu}{\sigma}\right)\right\} \\ &\leq \alpha. \end{aligned}$$

Therefore, when Condition 3.2.1 is true, we have

$$p_{\max}(X,Y) = \max\{p(W_{\min}(X,Y)), p(W_{\max}(X,Y))\} \leq \alpha.$$

The result when Condition 3.2.2 is true can be proved using the same approach. Thus, it is omitted here. Hence, we have proved that " $\Longrightarrow$ " is true.

We now prove that " $\Longleftarrow$ " holds. If $p_{\max}(X,Y) \leq \alpha$, then we have

$$(W_{\min}(X,Y)-\mu)(W_{\max}(X,Y)-\mu) > 0,$$

and

$$\max\{p(W_{\min}(X,Y)), p(W_{\max}(X,Y))\} \leq \alpha.$$

Note that

$$(W_{\min}(X,Y)-\mu)(W_{\max}(X,Y)-\mu) > 0$$

implies either

$$W_{\min}(X,Y), W_{\max}(X,Y) > \mu, \text{ or } W_{\min}(X,Y), W_{\max}(X,Y) < \mu.$$

Suppose $W_{\min}(X,Y), W_{\max}(X,Y) < \mu$. We then have

$$p(W_{\min}(X,Y)) = 2\Phi\left(\frac{W_{\min}(X,Y)-\mu}{\sigma}\right),$$

and

$$p(W_{\max}(X,Y)) = 2\Phi\left(\frac{W_{\max}(X,Y)-\mu}{\sigma}\right).$$

Since

$$\max\{p(W_{\min}(X,Y)), p(W_{\max}(X,Y))\} \leq \alpha,$$

we have

$$\Phi\left(\frac{W_{\min}(X,Y)-\mu}{\sigma}\right), \Phi\left(\frac{W_{\max}(X,Y)-\mu}{\sigma}\right) \leq \frac{\alpha}{2}.$$

Hence, we have shown Condition 3.2.2

$$\frac{W_{\min}(X,Y)-\mu}{\sigma} \leq \Phi^{-1}\left(\frac{\alpha}{2}\right) \text{ and } \frac{W_{\max}(X,Y)-\mu}{\sigma} \leq \Phi^{-1}\left(\frac{\alpha}{2}\right)$$

is true.

Following the same approach, when $W_{\min}(X,Y), W_{\max}(X,Y) < \mu$, we can show Condition 3.2.1 is true. Hence, we have proven " $\Longleftarrow$ ". This completes our proof. □

3.3. Maximum allowable proportion of missing data

The proposed method is based on the bounds of the Wilcoxon-Mann-Whitney test statistics in the presence of missing data, which are determined by both the values of observed data, and the proportion of missing data, as demonstrated by Theorem 3.1.3.

Informally, the bounds lead to significant testing results when the proportion of missing data are appropriate and the observed values provide enough evidence. This section considers the maximum proportion of missing data such that the proposed method can still yield significant results. The probability of rejecting the null hypothesis of the proposed method under different alternative hypothesis is studied in next Section 3.4, assuming the data are missing completely at random (MCAR).

The main result of this section is stated as

Proposition 3.3.1. Suppose that $X = \{x_1, , \cdots, x_n\}$ and $Y = \{y_1, \cdots, y_m\}$ are samples of distinct real-valued observations and $X' \subset X$ and $Y' \subset Y$ are subsets of distinct values with size n' and m', respectively, which are observed. Suppose that n, m are sufficiently large enough for applying normal approximation in the Wilcoxon-Mann-Whitney test. Then for any significance threshold $\alpha \in (0,1)$, if

$$n'm'/(nm) < 1/2 + \sqrt{(1+n+m)/(12nm)}\Phi^{-1}(1-\alpha/2), \tag{3.6}$$

where $\Phi(x)$ denotes the cumulative distribution function of the standard normal distribution, the proposed method will not reject the null hypothesis, regardless of the observed values in X' and Y'.

Proof. The proof is included in Appendix A.2. □

Consider fixed sample sizes n, m and a significance level α. The Proposition 3.3.1 asserts that the proposed method will not reject the null hypothesis for any observed values, when the number of observed values n', m' are small enough, such that $n'm'/(nm)$ is smaller than the right hand side of Inequality (3.6).

A result obtained directly from Proposition 3.3.1 is

Corollary 3.3.2. If X and Y both have at least 30% of data missing, the proposed method will not reject the null hypothesis.

Proof. Let $n'/n = 0.7$ and $m'/m = 0.7$. Then,

$$n'm'/nm = 0.49 < 1/2 < 1/2 + \sqrt{(1+n+m)/(12nm)}\Phi^{-1}(1-\alpha/2).$$

According to Proposition 3.3.1, the proposed method will not reject the null hypothesis, regardless of the values of missing data.

□

Proposition 3.3.1 demonstrates that the proposed method will not reject the null hypothesis when there is a large proportion of missing data, e.g. when both samples have at least 30% data missing, as shown in Corollary 3.3.2. On the other hand, we show that if the proportion of missing data is not subject to the Inequality (3.6), the proposed method gets chances to reject the null hypothesis H_0.

Proposition 3.3.3. Suppose that $X = \{x_1, , \cdots, x_n\}$ and $Y = \{y_1, \cdots, y_m\}$ are samples of distinct real-valued observations and $X' \subset X$ and $Y' \subset Y$ are subsets of distinct values with size n' and m', respectively, which are observed. Suppose that n, m are sufficiently large enough for applying normal approximation in the Wilcoxon-Mann-Whitney test. Then for any significance threshold $\alpha \in (0, 1)$, if

$$n'm'/(nm) \geq 1/2 + \Phi^{-1}(1-\alpha/2)\{(n+m+1)/(12nm)\}^{1/2}, \tag{3.7}$$

then sets X' and Y' exist for which the proposed method will yield a significant p-value.

Proof. Given $\alpha \in (0, 1)$, suppose we have n, m with $n' \leq n$ and $m' \leq m$ satisfying Inequality (3.7). A well-known result regarding the Wilcoxon-Mann-Whitney is that when all values in X' are smaller than all values in Y', and all values are distinct,

$$0 = W(X', Y').$$

The goal is to prove that if Inequality (3.7) holds then sets X' and Y' exist for which the proposed method will yield a significant p-value. Let us therefore choose sets X' and Y' where all values in X' are smaller than all values in Y', and all values are distinct, and so $W(X', Y') = 0$.

We proceed by rearranging Inequality (3.7), starting by multiplying both sides by nm:

$$\begin{aligned}
&\frac{n'm'}{nm} \geq \frac{1}{2} + \Phi^{-1}\left(1-\frac{\alpha}{2}\right)\left\{\frac{n+m+1}{12nm}\right\}^{\frac{1}{2}} \\
\iff & n'm' \geq \frac{nm}{2} + \Phi^{-1}\left(1-\frac{\alpha}{2}\right)\left\{\frac{n+m+1}{12nm}\right\}^{\frac{1}{2}} \\
\iff & n'm' - \frac{nm}{2} \geq \Phi^{-1}\left(1-\frac{\alpha}{2}\right)\left\{\frac{n+m+1}{12nm}\right\}^{\frac{1}{2}} \\
\iff & \left(n'm' - \frac{nm}{2}\right) / \left\{\frac{n+m+1}{12nm}\right\}^{\frac{1}{2}} \geq \Phi^{-1}\left(1-\frac{\alpha}{2}\right) \\
\iff & \left(\frac{nm}{2} - n'm'\right) / \left\{\frac{n+m+1}{12nm}\right\}^{\frac{1}{2}} \leq -\Phi^{-1}\left(1-\frac{\alpha}{2}\right). \qquad (3.8)
\end{aligned}$$

Recall that the statistic $W(X,Y)$ follows a normal distribution with mean $\mu = nm/2$ and variance $\sigma^2 = nm(n+m+1)/12$. If $(W_{\max}(X,Y)-\mu)/\sigma < \Phi^{-1}(\alpha/2)$, then Condition 3.2.1 is true, and the proposed method will declare a significant result. Noting $\Phi^{-1}(\alpha/2) = -\Phi^{-1}(1-\alpha/2)$, if

$$\frac{(W_{\max}(X,Y)-\mu)}{\sigma} < -\Phi^{-1}\left(1-\frac{\alpha}{2}\right),$$

then the proposed method will yield a significant result.

We recall from Theorem 3.1.3 that the $W_{\max}(X,Y)$ is given by

$$W_{\max}(X,Y) = W(X',Y') + nm - n'm',$$

but since we earlier assumed $W(X',Y') = 0$, then $W_{\max}(X,Y) = nm - n'm'$. Then

$$\begin{aligned}
\frac{(W_{\max}(X,Y)-\mu)}{\sigma} &= \frac{nm - n'm' - \frac{nm}{2}}{\left\{\frac{nm(n+m+1)}{12}\right\}^{\frac{1}{2}}} \\
&= \frac{nm/2 - n'm'}{\left\{\frac{nm(n+m+1)}{12}\right\}^{\frac{1}{2}}} \\
\implies \frac{(W_{\max}(X,Y)-\mu)}{\sigma} &\leq -\Phi^{-1}\left(1-\frac{\alpha}{2}\right), \qquad (3.9)
\end{aligned}$$

where the last line follows from Inequality (3.8). Note Inequality (3.9) implies that the proposed method will yield a significant p-value, which completes the proof. □

Proposition 3.3.3 demonstrates that the bounds of maximum allowable proportion of missing data given by Proposition 3.3.1 is tight, since if Inequality (3.6) does not hold, there always exists observed values X' and Y' that can lead the proposed method to yield significant results.

3.4. Power analysis when data are missing completely at random

This section provides theoretical results estimating the power of the proposed method when data are assumed to be missing completely at random (MCAR). Consequently, we change our notation slightly and consider sets of random variables rather than observations.

Our proposed method tests against location shifts as shown in Equation (2.1) in Section 2.1.1.

We assume the distributions for the two samples are such that $G = F(x + \Delta)$. In the following Proposition 3.4.1, we show that the power of our proposed method depends on the sample size n and m, the significance level α, the distribution F and the location shift parameter Δ, where F and Δ determine the power through the parameters p_1, p_2, and p_3. Table 3.4 compares the theoretical power derived in Proposition 3.4.1 with empirical results from numerical simulations. These estimates assume a standard normal distribution for F and a significance level of $\alpha = 0.05$, across varying values of the shift parameter Δ.

Proposition 3.4.1. Let $X = \{X_1, \cdots, X_n\}$ and $Y = \{Y_1, \cdots, Y_m\}$ be sets of random variables identically and independently distributed according to continuous distributions F and G, respectively. Let $X' \subset X$ and $Y' \subset Y$ denote the subsets of distinct random variables with sizes n' and m', respectively, which can be observed. Assume that the observations of the random variables in the complements $X \setminus X'$ and $Y \setminus Y'$ are missing completely at random (MCAR), and $0 < \Pr(X_1 < Y_1) < 1$. Suppose that n, m are sufficiently large enough for applying normal approximation in the Wilcoxon-Mann-Whitney test. Then, for any given significance level $\alpha \in (0, 1)$, the proposed method yields a significant p-value with probability approximately equal to

$$\Phi((L - \mu')/\sigma') + 1 - \Phi((R - \mu')/\sigma'),$$

where Φ is the cumulative distribution function of the standard normal distribution, and L, R, μ', σ' are

$$L = \sigma\Phi^{-1}(\alpha/2) + \mu - nm + n'm', \quad R = \sigma\Phi^{-1}(1 - \alpha/2) + \mu,$$
$$\mu' = m'n'p_1, \ (\sigma')^2 = m'n'p_1(1 - p_1) + m'n'(n' - 1)(p_2 - p_1^2) + n'm'(m' - 1)(p_3 - p_1^2),$$

and $\mu = nm/2$, $\sigma^2 = nm(n + m + 1)/12$, and

$$p_1 = \Pr(X_1 < Y_1), \ p_2 = \Pr(X_1 < Y_1 \text{ and } X_1 < Y_2), \ p_3 = \Pr(X_1 < Y_1 \text{ and } X_2 < Y_1).$$

Proof. The proof is provided in the Appendix A.3, based on results in Lehmann and D'Abrera (1975, Section 2.3). □

Remark 3.4.2. The condition $0 < \Pr(X_1 < Y_1) < 1$ required in the above proposition is necessary for the rescaled Wilcoxon-Mann-Whitney statistic to tend to the standard normal distribution asymptotically (Lehmann and D'Abrera, 1975, Appendix, Example 20). If this condition does not hold, i.e. $\Pr(X_1 < Y_1) \in \{0, 1\}$, then one of the distributions F or G lies entirely to the left of the other distribution (Lehmann and D'Abrera, 1975, Section 2.3).

Example 3.4.3. The result of Proposition 3.4.1 may appear to be complex at first glance. We therefore provide an example to demonstrate how Proposition 3.4.1 can be applied for estimating the power of the proposed method when data are missing completely at random (MCAR).

Suppose $X_1, \ldots, X_n \overset{\text{i.i.d.}}{\sim} F$, $Y_1, \ldots, Y_m \overset{\text{i.i.d.}}{\sim} G$, where $F = \mathrm{N}(0, 1)$, and $G = \mathrm{N}(1, 1)$. Suppose also the sample sizes $n = 100, m = 100$, the data are MCAR, and the number of observed data in each sample is $n' = 90, m' = 90$. For estimating the power, we consider the significance level $\alpha = 0.05$.

To compute the power of the proposed method, we first compute the probabilities p_1, p_2, p_3 defined in Proposition 3.4.1. When the data are assumed normally distributed, the probabilities can be obtained following the Equations (2.26), (2.27), (2.28) in Lehmann and D'Abrera (1975):

$$\begin{aligned} p_1 &= \Phi\left(1/\sqrt{2}\right) \approx 0.76, \\ p_2 = p_3 &= \Pr\left(Z < 1/\sqrt{2} \text{ and } Z' < 1/\sqrt{2}\right) \approx 0.63, \end{aligned}$$

where Z, Z' are joint standard normal distributions with a correlation coefficient of $1/2$.

Next, for given sample sizes $n, m = 100$, observed number of data $n', m' = 90$, and the significance level $\alpha = 0.05$, we compute other statistics required by Proposition 3.4.1 for deciding the power of the proposed method:

$$\begin{aligned} \sigma &= \sqrt{nm(n+m+1)/12} \approx 409.26, \\ \mu &= nm/2 = 5000, \\ \mu' &= n_1 m_1 p_1 \approx 6158.0, \\ \sigma' &= m'n'p_1(1-p_1) + m'n'(n'-1)(p_2 - p_1^2) + n'm'(m'-1)(p_3 - p_1^2) \approx 286.0358, \\ L &= \sigma\Phi^{-1}(0.025) + \mu - nm + n'm' \approx 2297.8, \\ R &= \sigma\Phi^{-1}(0.975) + \mu \approx 5802.1. \end{aligned}$$

Finally, according to Proposition 3.4.1, the power of the proposed method equals to

$$\begin{aligned} \Phi\left(\frac{L-\mu'}{\sigma'}\right) + 1 - \Phi\left(\frac{R-\mu'}{\sigma'}\right) &\approx \Phi(-13.49542) + 1 - \Phi(-1.244161) \\ &\approx 0.893. \end{aligned}$$

In other words, given the alternative hypothesis, the sample sizes, the number of observed data, and the significance level considered in this example, the probability of the proposed method yielding a significant p-value approximately equals 0.893.

Table 3.4 shows the estimated power obtained from Monte Carlo simulations compared to the theoretically-estimated power from Proposition 3.4.1, when data are missing completely at random (MCAR), as the proportion of data missing, s, varies. The table shows that the theoretical results closely match those obtained in simulation, at least for this case where the data are sampled from normal distributions.

This section ends with the following result which shows that, in the case where data are MCAR, the probability of the proposed method yielding a significant p-value tends to either 0 or 1, depending on the proportion of missing data, as the total sample sizes increase. The assumptions are the same as for Proposition 3.4.1.

Proposition 3.4.4. Let $X = \{X_1, \cdots, X_n\}$ and $Y = \{Y_1, \cdots, Y_m\}$ be sets of random variables identically and independently distributed according to continuous distributions F and G, respectively. Let $X' \subset X$ and $Y' \subset Y$ denote the subsets of distinct random variables with sizes n' and m', respectively, which can be observed. Assuming that the random variables in the complements $X \setminus X'$ and $Y \setminus Y'$ are missing completely at random, and $0 < \Pr(X_1 < Y_1) < 1$, and additionally

Table 3.1.: Estimated probabilities of rejections of the proposed method given by 1000 Monte Carlo replications and theoretical approximation given by Proposition 3.4.1 (in parentheses); data are missing completely at random (MCAR)

		Data sparsity					
		$s = 0$	$s = 5\%$	$s = 10\%$	$s = 15\%$	$s = 20\%$	$s = 30\%$
n, m	Δ						
20	0	0.06 (0.05)	0.01 (0.01)	0.00 (0.00)	0.00 (0.00)	0.00 (0.00)	0.00 (0.00)
	0.5	0.30 (0.31)	0.10 (0.10)	0.01 (0.01)	0.00 (0.00)	0.00 (0.00)	0.00 (0.00)
	1	0.85 (0.85)	0.54 (0.53)	0.17 (0.16)	0.00 (0.01)	0.00 (0.00)	0.00 (0.00)
	2	1.00 (1.00)	1.00 (1.00)	0.95 (0.96)	0.34 (0.32)	0.00 (0.00)	0.00 (0.00)
50	0	0.06 (0.05)	0.00 (0.00)	0.00 (0.00)	0.00 (0.00)	0.00 (0.00)	0.00 (0.00)
	0.5	0.68 (0.67)	0.33 (0.23)	0.02 (0.02)	0.00 (0.00)	0.00 (0.00)	0.00 (0.00)
	1	1.00 (1.00)	0.96 (0.95)	0.54 (0.52)	0.08 (0.04)	0.00 (0.00)	0.00 (0.00)
	2	1.00 (1.00)	1.00 (1.00)	1.00 (1.00)	1.00 (0.99)	0.08 (0.10)	0.00 (0.00)
100	0	0.06 (0.05)	0.00 (0.00)	0.00 (0.00)	0.00 (0.00)	0.00 (0.00)	0.00 (0.00)
	0.5	0.93 (0.93)	0.46 (0.45)	0.03 (0.03)	0.00 (0.00)	0.00 (0.00)	0.00 (0.00)
	1	1.00 (1.00)	1.00 (1.00)	0.90 (0.89)	0.12 (0.12)	0.00 (0.00)	0.00 (0.00)
	2	1.00 (1.00)	1.00 (1.00)	1.00 (1.00)	1.00 (1.00)	0.78 (0.76)	0.00 (0.00)
200	0	0.05 (0.05)	0.00 (0.00)	0.00 (0.00)	0.00 (0.00)	0.00 (0.00)	0.00 (0.00)
	0.5	1.00 (1.00)	0.77 (0.78)	0.04 (0.05)	0.00 (0.00)	0.00 (0.00)	0.00 (0.00)
	1	1.00 (1.00)	1.00 (0.95)	1.00 (1.00)	0.36 (0.35)	0.00 (0.00)	0.00 (0.00)
	2	1.00 (1.00)	1.00 (1.00)	1.00 (1.00)	1.00 (1.00)	1.00 (1.00)	0.00 (0.00)
300	0	0.05 (0.05)	0.00 (0.00)	0.00 (0.00)	0.00 (0.00)	0.00 (0.00)	0.00 (0.00)
	0.5	1.00 (1.00)	0.92 (0.92)	0.06 (0.06)	0.00 (0.00)	0.00 (0.00)	0.00 (0.00)
	1	1.00 (1.00)	1.00 (1.00)	1.00 (1.00)	0.59 (0.58)	0.00 (0.00)	0.00 (0.00)
	2	1.00 (1.00)	1.00 (1.00)	1.00 (1.00)	1.00 (1.00)	1.00 (1.00)	0.00 (0.00)
500	0	0.05 (0.05)	0.00 (0.00)	0.00 (0.00)	0.00 (0.00)	0.00 (0.00)	0.00 (0.00)
	0.5	1.00 (1.00)	0.99 (0.99)	0.09 (0.10)	0.00 (0.00)	0.00 (0.00)	0.00 (0.00)
	1	1.00 (1.00)	1.00 (1.00)	1.00 (1.00)	0.89 (0.88)	0.00 (0.00)	0.00 (0.00)
	2	1.00 (1.00)	1.00 (1.00)	1.00 (1.00)	1.00 (1.00)	1.00 (1.00)	0.00 (0.00)
1000	0	0.05 (0.05)	0.00 (0.00)	0.00 (0.00)	0.00 (0.00)	0.00 (0.00)	0.00 (0.00)
	0.5	1.00 (1.00)	1.00 (1.00)	0.20 (0.21)	0.00 (0.00)	0.00 (0.00)	0.00 (0.00)
	1	1.00 (1.00)	1.00 (1.00)	1.00 (1.00)	1.00 (1.00)	0.00 (0.00)	0.00 (0.00)
	2	1.00 (1.00)	1.00 (1.00)	1.00 (1.00)	1.00 (1.00)	1.00 (1.00)	0.00 (0.00)
10000	0	0.06 (0.05)	0.00 (0.00)	0.00 (0.00)	0.00 (0.00)	0.00 (0.00)	0.00 (0.00)
	0.5	1.00 (1.00)	1.00 (1.00)	1.00 (1.00)	0.00 (0.00)	0.00 (0.00)	0.00 (0.00)
	1	1.00 (1.00)	1.00 (1.00)	1.00 (1.00)	1.00 (1.00)	0.00 (0.00)	0.00 (0.00)
	2	1.00 (1.00)	1.00 (1.00)	1.00 (1.00)	1.00 (1.00)	1.00 (1.00)	0.00 (0.00)

s data sparsity for both X and Y; n, m sample sizes for X and Y respectively; Δ value of mean shift such that all samples in X generated from $\mathrm{N}(0, 1)$ and Y from $\mathrm{N}(\Delta, 1)$.

assuming $m' \leq n'$ and

$$
\begin{aligned}
&m'/n' \to \lambda' \text{ when } n, m \to \infty, \text{ where } 0 < \lambda' \leq 1, \\
&n'/n \to \lambda_x, \text{ when } n, m \to \infty, \text{ where } 0 < \lambda_x \leq 1, \\
&m'/m \to \lambda_y, \text{ when } n, m \to \infty, \text{ where } 0 < \lambda_y \leq 1 \\
&\lambda_x \lambda_y (p_1 - 1) + 1/2 \neq 0 \text{ and } \lambda_x \lambda_y p_1 - 1/2 \neq 0, \text{ where } p_1 = \Pr(X_1 < Y_1).
\end{aligned}
$$

As $n, m \to \infty$, with probability

$$
p = \begin{cases} 0, & \text{if } \lambda_x \lambda_y (p_1 - 1) + 1/2 > 0 \text{ and } \lambda_x \lambda_y p_1 - 1/2 < 0, \\ 1, & \text{otherwise,} \end{cases}
$$

the proposed method will yield a significant p-value.

Proof. The proof is provided in Appendix A.4, based on similar results in Lehmann and D'Abrera (1975, Section 3, Chapter 2). □

3.5. Simulations for the location test

This section performs numerical simulations for investigating the Type I error and statistical power of the proposed method when data are missing. The performance of our method is also compared with the Wilcoxon-Mann-Whitney test when the missing data have been imputed using either mean imputation or hot deck imputation, or when the missing data are ignored, which we call case deletion, or when the missing values are known, i.e. the data is complete.

3.5.1. As the proportion of missing data varies

We first consider experiments comparing the performance of the proposed method and other methods with varying proportion of missing data s, while other factors that can potentially affect the experiments results, including sample sizes, missingness mechanisms, and the distributions of the data, are fixed for each experiment.

Missing completely at random (MCAR)

The first experiment considers the case where data are missing completely at random (MCAR). Observations in X are sampled independently from a $\mathrm{N}(0, 1)$ distribution, while observations in Y are sampled independently from a $\mathrm{N}(0, 1)$ distribution to evaluate the Type I error, and from a $\mathrm{N}(1, 1)$ distribution to evaluate the statistical power. A proportion $s \in [0, 0.3]$ of the observations $X = \{x_1, \ldots, x_n\}$ are selected completely at random to be marked as missing. The same proportion s of $Y = \{y_1, \ldots, y_m\}$ are selected completely at random to be marked as missing. Sample sizes $n = m = 100$ are considered.

Figure 3.1 shows that the Type I error is not controlled by either hot deck imputation or mean imputation, although it is controlled for the proposed method and for the case deletion case when the missing data are ignored. On the other hand, all methods have good power, except

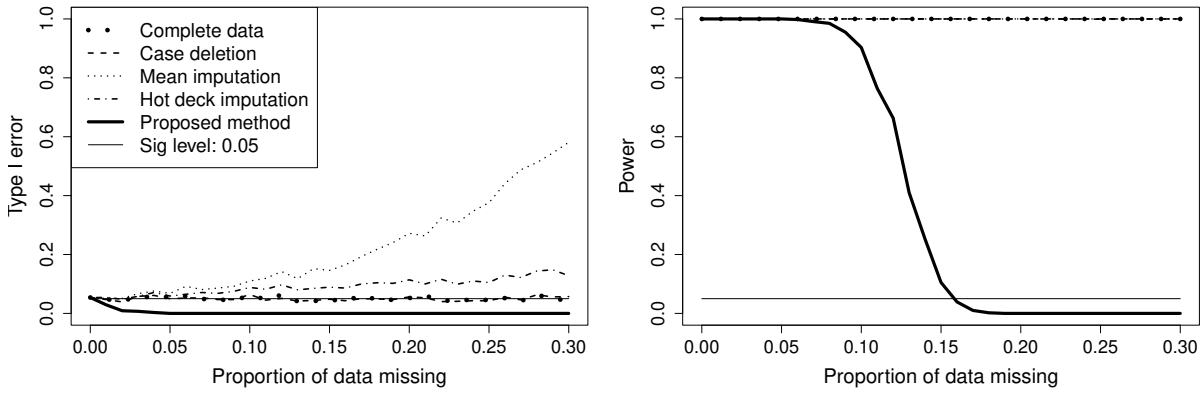


Figure 3.1.: The Type I error and statistical power of the proposed method and the standard Wilcoxon-Mann-Whitney test after the missing data is either known or has been imputed or ignored as the proportion of missing data increases. The data is missing completely at random (MCAR). (Left) Type I error: $\mathrm{N}(0,1)$ vs $\mathrm{N}(0,1)$; (Right) Power: $\mathrm{N}(0,1)$ vs $\mathrm{N}(1,1)$. For both figures, a significance threshold of $\alpha = 0.05$ has been used and the total sample sizes are $n = 100$, $m = 100$, and 1000 trials were used.

that the power for the proposed method decreases significantly when more than 10% of the data is missing.

For this experiment, $n = m = 100$, but Figure E.1 in Appendix E.1.1 considers an experiment with different sample sizes $n, m = 500$, which shows the increasing power of the proposed method. Furthermore, if there is a larger difference between the two samples for the power experiment, for example if the second sample follows a $\mathrm{N}(2,1)$ distribution rather than a $\mathrm{N}(1,1)$ distribution, then the proposed method can still have good power for over 20% of the data is missing; see Figure E.2 in Appendix E.1.1.

Missing not at random (MNAR)

The second experiment is the same as the first, except in this experiment the data are missing not at random (MNAR). The missingness mechanism is as follows: if s is the proportion of observations to be missing, then for any observation $x_i \in \{x_1, \ldots, x_n\}$, the probability of x_i being missing is

$$\Pr\left(x_i \text{ is missing}\right) = \begin{cases} \min(1, sn/\sum_{j=1}^{n} \mathrm{I}\left(x_j < 0\right)), & \text{if } x_i < 0, \\ \max(0, sn/\sum_{j=1}^{n} \mathrm{I}\left(x_j < 0\right) - 1), & \text{otherwise,} \end{cases} \tag{3.10}$$

where $\mathrm{I}\left(A\right)$ is the indicator function for the event A. In other words, when there are more negative values in X than the number of missing values in X, i.e. $sn/\sum_{j=1}^{n} \mathrm{I}\left(x_j < 0\right) < 1$, only values smaller than 0 in X can be missing. Otherwise all values smaller than 0 in X will be missing, and values larger than 0 will be randomly missing.

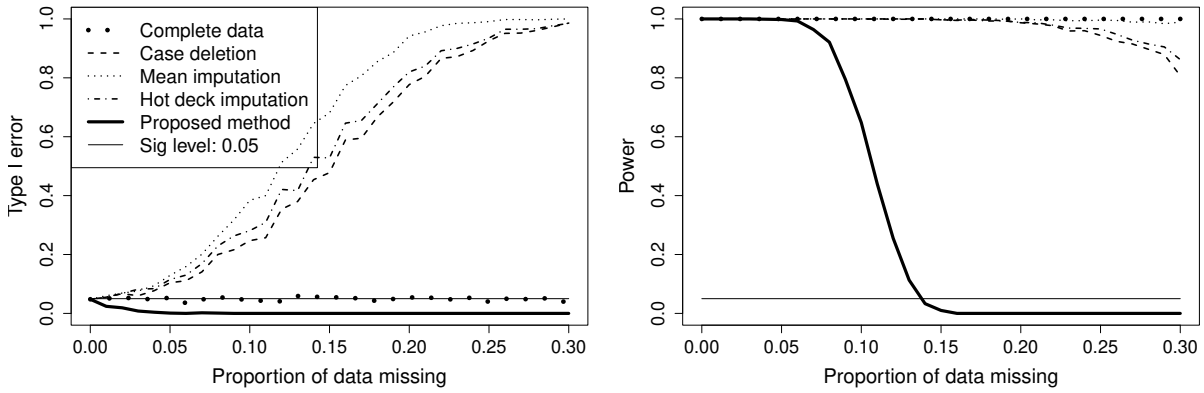


Figure 3.2.: The Type I error and statistical power of the proposed method and the standard Wilcoxon-Mann-Whitney test after the missing data is either known or has been imputed or ignored as the proportion of missing data increases. The data is missing not at random (MNAR), according to the mechanism in Equation (3.10), and (3.11). (Left) Type I error: $\mathrm{N}(0,1)$ vs $\mathrm{N}(0,1)$; (Right) Power: $\mathrm{N}(0,1)$ vs $\mathrm{N}(1,1)$. For both figures, a significance threshold of $\alpha = 0.05$ has been used and the total sample sizes are $n = 100$, $m = 100$, and 1000 trials were used.

For $y_i \in \{y_1, \ldots, y_m\}$, a similar missingness mechanism

$$\Pr\left(y_i \text{ is missing}\right) = \begin{cases} \min(1, sm/\sum_{j=1}^{m} \mathrm{I}\left(y_j > 0\right)), & \text{if } y_i > 0, \\ \max(0, sm/\sum_{j=1}^{m} \mathrm{I}\left(y_j > 0\right) - 1), & \text{otherwise,} \end{cases} \tag{3.11}$$

is used. When there are more positive values in Y than the number of missing values in Y, i.e. $sm/\sum_{j=1}^{m} \mathrm{I}\left(y_j > 0\right) < 1$, only values larger than 0 in Y can be missing. Otherwise all values larger than 0 in Y will be missing, and values smaller than 0 will be randomly missing.

Figure 3.2 shows that when the data from both samples are MNAR and follow the above missingness mechanism, then the Type I error is not controlled for either of the imputation methods or when the missing data are ignored. This relatively simple example illustrates the potential peril of not taking missing data into account. On the other hand, the proposed method controls the Type I error rate for this case. The statistical power of these methods appears similar to that for the missing completely at random (MCAR) case in Figure 3.1; all methods have good power, although the proposed method's power decreases as the proportion of missing data increases beyond 10% of the total.

While no formal proof is provided for the power of the proposed method under MNAR, the intuition for the power of the proposed method is that when the proportion of missing data is moderate (e.g., 5-10%), the bias introduced by the missingness mechanism is relatively small compared to the magnitude of the location shift Δ. Consequently, the observed data still retain sufficient information to detect the location difference. However, as the proportion of missing data increases, the observed data become both increasingly sparse and more severely biased. Hence, the missing data obscures the location shift, causing the power of the test to approach zero, as illustrated in Figure 3.2.

Additional experiments in Appendix E.1.2 shows the power of the proposed method increases

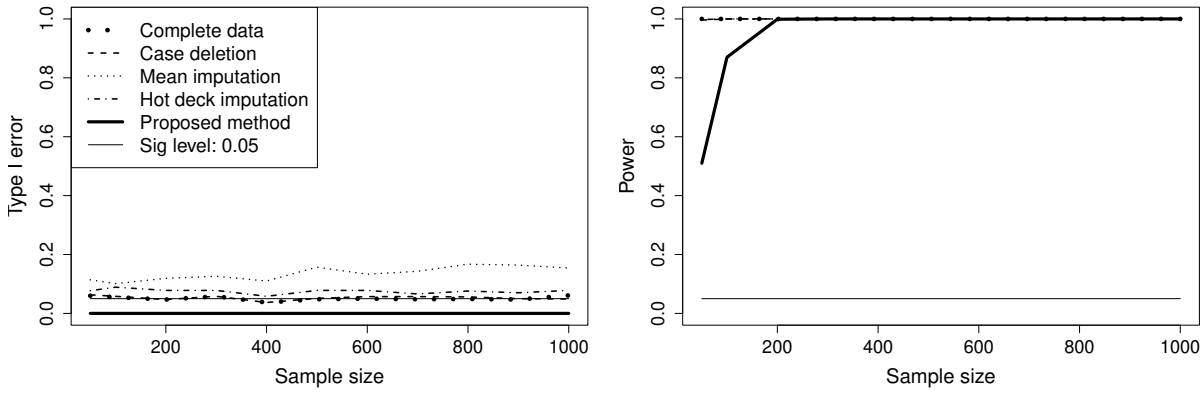


Figure 3.3.: The Type I error and statistical power of the proposed method and the standard Wilcoxon-Mann-Whitney test after the missing data is either known or has been imputed or ignored as the sample sizes increase. The data is missing completely at random (MCAR). (Left) Type I error: $\mathrm{N}(0,1)$ vs $\mathrm{N}(0,1)$; (Right) Power: $\mathrm{N}(0,1)$ vs $\mathrm{N}(1,1)$. For both figures, a significance threshold of $\alpha = 0.05$ has been used and the proportion of missing data for both samples is $s = 0.1$, and 1000 trials were used.

for larger sample sizes, and/or when the data in the second sample are observations from a $\mathrm{N}(2,1)$ distribution.

3.5.2. As the sample size varies

We again consider the data to be either MCAR or MNAR following the mechanism in Equation (3.10), and (3.11), but now as the sample sizes n, m increase, with fixed proportion of missing data $s = 0.1$.

Missing completely at random (MCAR)

The first experiment considers data to be MCAR. As before, $X = \{x_1, \ldots, x_n\}$ are sampled independently from a $\mathrm{N}(0,1)$ distribution, and observations in $Y = \{y_1, \ldots, y_m\}$ are sampled independently from a $\mathrm{N}(0,1)$ distribution to evaluate the Type I error, and from a $\mathrm{N}(1,1)$ distribution to evaluate the statistical power. The sample sizes $n \in \{50, 100, 200, \ldots, 1000\}$ are considered, with $m = n$. Then $s = 0.1$ of the observations $X = \{x_1, \ldots, x_n\}$ are selected completely at random to be marked as missing, and the same proportion s of $Y = \{y_1, \ldots, y_m\}$ are selected completely at random to be marked as missing.

Figure 3.3 shows that mean imputation and hot deck imputation fail to control the Type I error for sample sizes considered in the experiment, while case deletion and the proposed method control the Type I error. All methods have good power. The power of the proposed method increases with sample sizes. When the sample size is 200, the power of the proposed method is close to 1.

Missing not at random (MNAR)

We now consider two experiments similar to those in Figure 3.3, but with data that are MNAR, following the mechanisms specified in Equations (3.10) and (3.11).

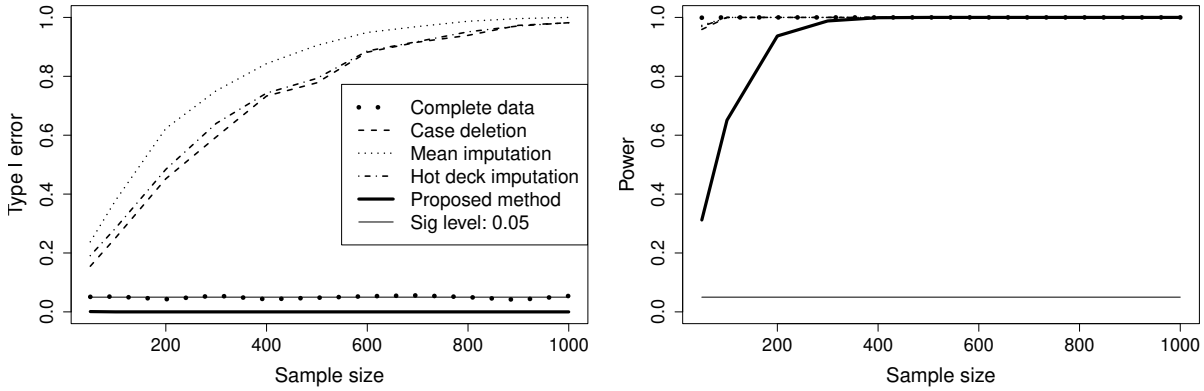


Figure 3.4.: The Type I error and statistical power of the proposed method and the standard Wilcoxon-Mann-Whitney test after the missing data is either known or has been imputed or ignored as the sample sizes increase. The data is missing not at random (MNAR), according to the mechanism in Equation (3.10), and (3.11). (Left) Type I error: $\mathrm{N}(0,1)$ vs $\mathrm{N}(0,1)$; (Right) Power: $\mathrm{N}(0,1)$ vs $\mathrm{N}(1,1)$. For both figures, a significance threshold of $\alpha = 0.05$ has been used and the proportion of missing data for both samples is $s = 0.1$, and 1000 trials were used.

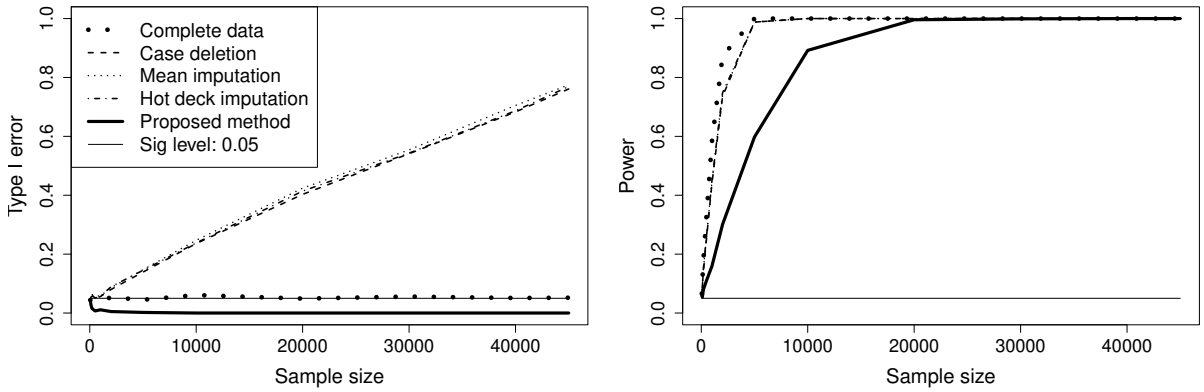


Figure 3.5.: The Type I error and statistical power of the proposed method and the standard Wilcoxon-Mann-Whitney test after the missing data is either known or has been imputed or ignored as the sample sizes increase. The data is missing not at random (MNAR), according to the mechanism in Equation (3.10), and (3.11). (Left) Type I error: $\mathrm{N}(0,1)$ vs $\mathrm{N}(0,1)$; (Right) Power: $\mathrm{N}(0,1)$ vs $\mathrm{N}(0.1,1)$. For both figures, a significance threshold of $\alpha = 0.05$ has been used and the proportion of missing data for both samples is $s = 0.01$, and 1000 trials were used.

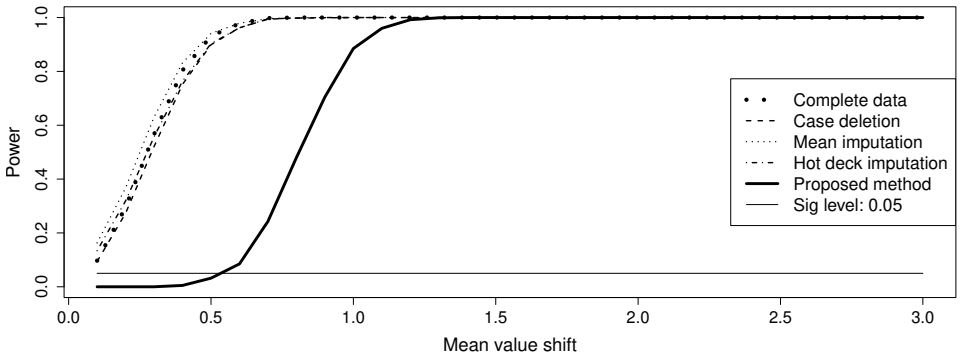


Figure 3.6.: Statistical power of the proposed method and the standard Wilcoxon-Mann-Whitney test after the missing data is either known or has been imputed or ignored as the sample sizes increase. The data is missing completely at random (MCAR). Data in samples X and Y are generated from $\mathrm{N}(0, 1)$, and $\mathrm{N}(\Delta, 1)$, respectively, with Δ denoting the mean value shift. A significance threshold of $\alpha = 0.05$ has been used, the total sample sizes are $n = m = 100$, the proportion of missing data for both samples is $s = 0.1$, and 1000 trials were used.

Figure 3.4 further highlights the potential peril of not taking the missing data into account: the Type I error of case deletion, mean imputation and hot deck imputation methods appear to be asymptotically converge to 1 as the sample size increases with fixed proportion of missing data. In contrast, the proposed method controls the Type I error across all sample sizes considered. The power of all methods is similar to the results obtained in Figure 3.3 when the data are MCAR.

Figure 3.5 provides a similar experiment, but with only 1% of the data missing, and with sample Y generated from $\mathrm{N}(0.1, 1)$ to evaluate power. The results show that, even with only 1% missing data, case deletion, and imputation-based methods cannot control the Type I error. By contrast, the proposed method always controls the Type I error, and can detect small differences between X and Y.

3.5.3. As the mean value varies

The following of this section considers two experiments for evaluating the power of the proposed method with varying mean values of the distribution generating the sample Y. The sample sizes, and the proportion of missing data for both X and Y are fixed. Other methods considered in this section before are also included.

Missing completely at random (MCAR)

The first experiment again assumes the data are MCAR. Observations in the sample X are sampled independently from a $\mathrm{N}(0, 1)$ distribution, while observations in sample Y are sampled independently from a $\mathrm{N}(\Delta, 1)$, where $\Delta \in \{0.1, 0.2, \ldots, 3.0\}$. Sample sizes $n = m = 100$ are used.

Subsequently, $s = 0.1$ of the observations $X = \{x_1, \ldots, x_n\}$ are selected completely at random to be marked as missing, and the same proportion s of $Y = \{y_1, \ldots, y_m\}$ are selected completely at random to be marked as missing.

Figure 3.6 shows that under the settings we considered in this experiment, the proposed method

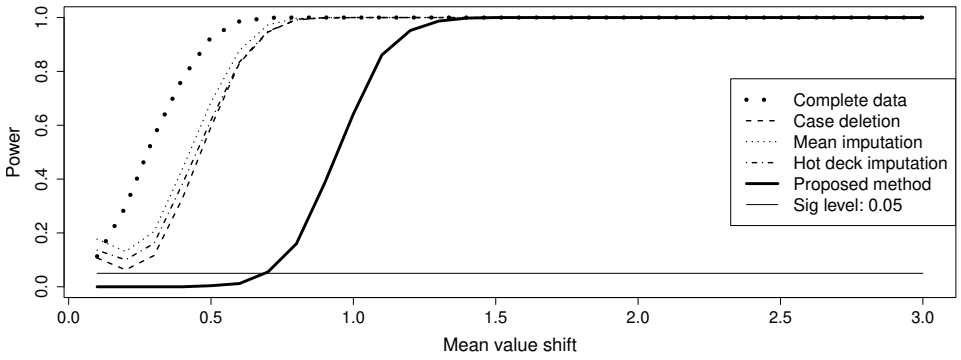


Figure 3.7.: Statistical power of the proposed method and the standard Wilcoxon-Mann-Whitney test after the missing data is either known or has been imputed or ignored as the sample sizes increase. The data is missing not at random, according to the mechanism in Equation (3.10), and (3.11). Data in samples X and Y are generated from $\mathrm{N}(0, 1)$, and $\mathrm{N}(\Delta, 1)$, respectively, with Δ denoting the mean value shift. A significance threshold of $\alpha = 0.05$ has been used, the total sample sizes are $n = m = 100$, the proportion of missing data for both samples is $s = 0.1$, and 1000 trials were used.

starts to have power when the mean value shift Δ is greater than 0.5. When the shift $\Delta = 1$, the power of the proposed method is close to 1. The power of all other methods increases when $\Delta > 0$, and are close to 1 and when $\Delta = 0.5$.

Missing not at random (MNAR)

This experiment considers a similar case to Figure 3.6, but the data are MNAR according to the missingness mechanism specified by Equation (3.10), and (3.11).

Figure 3.7 shows similar results to Figure 3.6, although all methods appear to require a larger mean value shift Δ to achieve power equals to 1. The power of the proposed method is approximately 0.6 when the mean value shift $\Delta = 1$, and appears to converges to 1 when $\Delta = 1.3$.

Overall, this section evaluates the Type I error and statistical power of the proposed method for location testing in the presence of missing data, and compares it with the Wilcoxon-Mann-Whitney test when the missing data have been imputed using mean or hot deck imputation, or when the data are ignored, or when the data is known.

The simulation results show that the proposed method controls the Type I error, while imputation methods and case deletion have larger Type I error than the significance level α, especially when the data are missing not at random (MNAR). For fixed proportion of missing data, the Type I error of methods using imputations and case deletion appears to asymptotically converge to 1 as the sample size increases, at least under the experiment settings considered for Figure 3.4.

The proposed method appears to have good statistical power, provided moderate total sample sizes n, m for X, Y, mean value differences of the distributions generating samples X and Y, and the proportion of missing data. For example, when $n = m = 100$, samples X and Y are generated from $\mathrm{N}(0, 1)$, and $\mathrm{N}(1, 1)$, respectively, and 10% data in both X and Y are missing, the proposed method has power close to 90% and 60% when the data are MCAR, and MNAR, respectively, as illustrated by Figure 3.1 and Figure 3.2.

Table 3.2.: Treatment groups in clinical trial NCT01874431. The number of patients in each treatment group, by dosage level, that start and complete the trial, with the percentage missing shown.

Finerenone dosage	Started	Completed	% missing
0 mg (Placebo)	94	90	4.3
1.25 mg	96	90	6.3
2.5 mg	92	87	5.4
5 mg	100	89	11.0
7.5 mg	97	89	8.2
10 mg	98	89	9.2
15 mg	125	114	8.8
20 mg	119	112	5.8

Greater power of the proposed method can be achieved, when the sample sizes n, m increases, and/or when larger mean value differences of distributions generating X and Y are provided, as demonstrated by experiments in Section 3.5.2 and Section 3.5.3.

While experiments considered in this section are restricted to normal distributed data, the proposed method can be applied to data generated from any continuous distribution, since we only require data to be distinct.

Additional experiments which consider non-normal distributed data are included in Appendix E.1.3, where the data in both X and Y are generated from the gamma distribution, rather than the normal distribution, although the results are similar. Appendix E.1.4 considers a MNAR mechanism that is different than the mechanism specified by Equation (3.10) and (3.11), and also obtains similar results.

3.6. Application to clinical trial data

In this section we demonstrate the proposed method to analyse clinical trial data that contains missing values.

Patients with type 2 diabetes may suffer from kidney disease, or *diabetic nephropathy*. Albumin is a protein in the blood that a healthy kidney will not allow to pass into the urine (NIDDK, 2016). However, a damaged kidney may allow albumin to pass into the urine, and a high level of albumin in the urine may indicate kidney disease. Persistent *albuminuria* is defined as a urinary albumin-creatinine ratio (UACR) that is greater than or equal to 30 mg/g (Bakris et al., 2015).

The clinical trial NCT01874431 (Bayer, 2022) assessed the efficacy of different doses of the drug finerenone (BAY 94-8862) for reducing the urinary albumin-creatinine ratios of patients with type 2 diabetes and diabetic nephropathy. This dataset is available from Bayer via the Vivli platform (Vivli, 2023). There were 1501 patients enrolled in the trial, of whom 821 were randomly allocated to eight different treatment groups while the remainder were excluded. One treatment group received a placebo, while the other seven groups each received finerenone at a different dosage level. The number of patients in each treatment group is shown in Table 3.2.

The patients' urinary albumin-creatinine ratios were measured at the start of the trial (baseline value) as well as on a further five visits, the fifth and final visit being on day 90 of the trial. The

Table 3.3.: Results from statistical analysis, with p-values

	1.25 mg	2.5 mg	5 mg	7.5 mg	10 mg	15 mg	20 mg
Ignoring missing	0.196	0.208	0.264	0.002	3.9×10^{-4}	7.4×10^{-6}	6.5×10^{-8}
Proposed method	1.000	1.000	1.000	0.112	0.072	0.011	1.4×10^{-4}
Proposed, Holm-corrected	1.000	1.000	1.000	0.450	0.358	0.064	9.4×10^{-4}

difference in patients' urinary albumin-creatinine ratios at the start and end of the trial was compared across the different treatment groups and the placebo by Bakris et al. (2015) using an ANCOVA model and an F-test. They found a significant improvement for the treatment groups with finerenone dosage levels of 7.5 mg, 10 mg, 15 mg and 20 mg, which in this case means that the obtained p-values are 0.004 or less. However, not all patients who start the trial completed the trial; there is a drop-out rate of between 4% and 11% for each treatment group; see Table 3.2 for details. In other words, there are *missing data* in the final urinary albumin-creatinine ratio measurements. In the original analysis, these missing data were imputed by using the last recorded value from an earlier visit as the imputed value. A sensitivity analysis was then performed by repeating the primary analysis using different imputation methods; see Bakris et al. (2015) for details.

We now analyse these data, avoiding any imputation, using our proposed method. Let X be the set of placebo group measurements which consists of $n = 94$ patients. For $i = 1, \ldots, 94$, let $x_i = (c_i - s_i)/s_i$, where s_i is patient i's baseline (starting) urinary albumin-creatinine ratio measurement and c_i is patient i's day 90 (completion) corresponding measurement. The x_i values measure the proportional increase or decrease in urinary albumin-creatinine ratio. Four of the patients in the placebo group are missing c_i values, and so four x_i values are missing values. Next, let Y_1 be the similarly transformed values for the first treatment group with finerenone dosage 1.25 mg, and let $Y_2, \ldots, Y_7$ be the data for the remaining treatment groups. Again, several values in $Y_1, \ldots, Y_7$ will be missing data. We now use the proposed method to compare the X values with the Y_j values for $j = 1, \ldots, 7$.

If each element in X is an observation of a random variable U and each element in Y_j is an observation of a random variable V, then the null hypothesis is that $U = V$. Since the alternative hypothesis is that the treatment lowers the patients' urinary albumin-creatinine ratios, the alternative hypothesis is $U > V$, and so the p-values are one-sided. The obtained p-values are shown in the second row of Table 3.3, with the p-values for the 15 mg and 20 mg treatment groups computed as 0.011 and 1.4×10^{-4}, respectively.

To correct for multiple hypothesis testing, we use the Holm method (Holm, 1979) to compute the adjusted p-values; this results in the p-values 0.064 and 9.4×10^{-4} for the 15 mg and 20 mg treatment groups, respectively. The Holm method is preferred since it is both more powerful than the Bonferroni method and does not impose any independence assumptions on the p-values, unlike other approaches.

Table 3.3 shows all the computed p-values for our proposed method and, for comparison, the p-values computed when ignoring missing data This table shows two interesting features. First, our proposed method yields a p-value for the 20 mg treatment group that is significant at the 0.001 significance level, after correction for multiple hypothesis testing. However, we do not obtain

significant (corrected) p-values for the other treatment groups. The second interesting feature is that the Wilcoxon-Mann-Whitney test on the data, ignoring missing values, gives significant p-values for the 7.5 mg, 10 mg, 15 mg and 20 mg treatment groups, as in the analysis in Bakris et al. (2015).

To summarise these results, our proposed method obtains a significant result only for the 20 mg treatment group and notably this result is robust to any form of missingness. However, our purpose in providing the analysis of this dataset is to illustrate the impact of missing data in a complex study, rather than to provide clinical recommendations.

3.7. Conclusion

This chapter presents the work of Zeng et al. (2024a). We begin by deriving tight bounds of the Wilcoxon-Mann-Whitney test statistics in the presence of missing data, assuming only the data are distinct. Based on the bounds, we propose to reject the null hypothesis only when all possible test statistics are significant. We further derive the bounds of the p-value of the Wilcoxon-Mann-Whitney test in the presence of missing data. We show that, when the maximum possible p-value is smaller than the significance level α, all test statistics are significant, and the proposed method rejects the null hypothesis. Importantly, we prove that our proposed method controls the Type I error, regardless of the values of missing data.

We further analyse the proposed method by deriving the exact upper bound on the maximum proportion of data that can be missing while allowing the proposed method to yield a significant outcome. We show that, when both samples have at least 30% of their data missing, the proposed method does not yield significant result, regardless of the observed values. When the data are assumed to be MCAR, the approximate and asymptotic power of our method are also derived.

The simulation results demonstrate that our method has good statistical power when less than 10–20% of the data missing. When the data are missing not at random (MNAR), common missing data methods such as case deletion, mean imputation and hot deck imputation fail to control the Type I error, while our approach always controls the Type I error. Finally, we illustrate our method on a clinical trial dataset in which patients' withdrawal leads to missing values.

The central idea of this thesis is to first bound the test statistic and then reject the null only when all possible statistics are significant. The next chapter extends this framework to two-sample scale testing with missing data, beginning by deriving tight bounds for the Ansari–Bradley statistic.

Acknowledgments

The analysis in Section 3.6 is based on research using data from data contributor Bayer that has been made available through Vivli, Inc (Vivli, 2023). Vivli has not contributed to or approved, and is not in any way responsible for, the contents of this thesis. The Vivli platform can be accessed at https://vivli.org/.

4. Scale test with arbitrarily missing data

This chapter describes the work in Zeng et al. (2025c). We consider the two-sample *scale* testing problem in the presence of missing data, without making any assumptions about the missingness mechanisms. The key idea of this chapter is similar to that in Chapter 3: we consider the exact bounds of the Ansari-Bradley test statistic in the presence of missing data, and propose to reject the null hypothesis only when all possible test statistics are significant.

In Chapter 3, we propose a valid location two-sample testing method in the presence of missing data. However, in practice, a change in location is often accompanied by a change in variability (Marozzi, 2013; Neuhäuser, 2001; Zhang et al., 2009). Moreover, detecting heteroscedasticity between two samples is often important, as revealing differences in variability may carry practical significance. For instance, in clinical trials, the presence of heteroscedasticity can indicate a treatment effect (Murakami and Neuhäuser, 2025). Zhu et al. (2019) report that the intra-individual variability of total cholesterol could be a risk factor of cardiovascular mortality, irrespective of mean total cholesterol level. This chapter is therefore motivated to develop valid scale and location-scale two sample testing methods in the presence of missing data, without relying on missingness assumptions.

The two-sample scale testing problem, as described previously in Section 2.1.1, is closely related to the two-sample variance testing used to determine whether the variances of two populations are equal or significantly different. As mentioned in Section 2.2.2, one common scale testing method is the Ansari-Bradley test. Similar to the Wilcoxon-Mann-Whitney test as considered in Chapter 3 for the location test, the Ansari-Bradley test is also based on rank, and is a non-parametric test such that the distribution of the test statistic under the null hypothesis does not depend on the distribution of the data.

Section 4.1 considers exact bounds of the Ansari-Bradley test statistic in the presence of missing data. The proofs for these bounds appear to be more complicated than the proofs for the bounds of the Wilcoxon-Mann-Whitney test statistic. Hence Section 4.1 only sketches the proofs, and presents the main results for the bounds, while all the technical details can be found in Appendix B. Section 4.2 then considers the conditions for rejecting the null hypothesis using the bounds derived, and discusses the bounds of the p-values. Section 4.3 performs numerical simulations for evaluating the Type I error, and power of the proposed method, and compares them with common missing data methods as considered earlier in Section 3.5. We obtained similar results to those in Section 3.5: the proposed methods have good power until more than 10-20% data are missing, while other methods fail to control the Type I error.

In Section 4.4, we introduce a new location-scale testing method for missing data by combining the location test from Chapter 3 with the scale test developed in this chapter, using the Holm-Bonferroni method. The method is further evaluated through numerical simulations in Section 4.5 and illustrated in Section 4.6 using a hepatitis C virus dataset with missing values.

4.1. Bounds of the Ansari-Bradley test statistic

This section considers the tight bounds of the Ansari-Bradley test statistic in the presence of missing data. We provide intuition for the main results and the sketch for proofs, while all technical details are provided in Appendix B.

4.1.1. Intuition for the bounds and main results

Before formally state the main results for the bounds, we provide an intuitive overview for the main result as shown in Theorem 4.1.3. First, let us recall that the Ansari-Bradley statistic is defined as follows in Definition 2.2.4:

$$T(X,Y) = \sum_{i=1}^{n} \left| R(x_i, X \cup Y) - \frac{1}{2}(N+1) \right|,$$

where $N = n + m$ is the total sample size. Suppose both X and Y contain missing data. Then, the minimum possible Ansari-Bradley test statistic, i.e. lower bound, will be taken when the ranks of the missing data in X take values close to "center" $(N+1)/2$, while the ranks of the missing data in Y are as far as possible from this center.

To be more specific, we prove that the minimum possible Ansari-Bradley test statistic between X and Y will be taken when the ranks of all the missing values in X take the consecutive integers symmetrically around $(N+1)/2$, while the ranks of all missing values in Y take either the minimum, or the maximum integers in all possible ranks $\{1, \ldots, N\}$. The maximum possible Ansari-Bradley test statistic can then be derived from the lower bound directly, based on the fact that the sum $T(X,Y) + T(Y,X)$ is a constant for a fixed sample size n, m.

While the intuition is straightforward, the formal proof involves several technical complications and requires us to handle them carefully:

- We can not assign ranks to $(N+1)/2$ when N is even. Therefore, the missing values of X must be distributed symmetrically around the center using the closest available integers. Definition 4.1.1 handles these parity cases.
- For the missing values of Y, we must decide how many go to the "left" extreme, i.e. smallest ranks, versus the "right" extreme, i.e. largest ranks. Definition 4.1.2 introduces a function $f(k)$ to evaluate the cost of placing k values at the maximum rank, allowing us to minimize over all valid splits.

The main result Theorem 4.1.3, formalizes the intuition and incorporates the technical complications discussed above. The theorem treats the four cases in Definition 4.1.1 separately and explicitly evaluates $f(k)$ for all valid values of k.

4.1.2. Definitions and notation

As we discussed above, to formally state the main results for the bounds, we first make the following two definitions. Let the sample sizes of X and Y be n and m, respectively. Let n' and m' denote the number of observed values in X and Y. The total sample size of X and Y is $N = n + m$, and the number of missing values in X is $n - n'$. Our first definition classifies

all possible configurations of the total sample size and the number of missing values into four distinct cases, based on the parity of N and $n - n'$.

Definition 4.1.1. Let $N \geq n \geq n'$ be positive integers. We define the following four cases $\mathcal{C}(1), \mathcal{C}(2), \mathcal{C}(3)$ and $\mathcal{C}(4)$ according to the parity of N and $n - n'$:

$$\begin{aligned}
\mathcal{C}(1): &\ N \text{ is odd but } n - n' \text{ is even,} \\
\mathcal{C}(2): &\ N \text{ is odd and } n - n' \text{ is odd,} \\
\mathcal{C}(3): &\ N \text{ is even but } n - n' \text{ is odd,} \\
\mathcal{C}(4): &\ N \text{ is even and } n - n' \text{ is even.}
\end{aligned}$$

Let Y' denote the set of all observed values in Y. Our second definition introduces a function that is important for formulating the final results for the bounds.

Definition 4.1.2. Suppose $X = \{x_1, \ldots, x_n\}$ and $Y' = \{y_1, \ldots, y_{m'}\}$ are samples of distinct real values, and $m \in \mathbb{N}$ is a positive integer such that $m \geq m'$. Denote $Z' = X \cup Y'$, $N' = (n+m'+1)/2$, and $M' = (m - m')/2$. For any given $k \in \{0, \ldots, m - m'\}$, denote $a = \min\{0, k - M'\}$, $b = \max\{0, k - M'\}$. Further, for any $i \in \{1, \ldots, n\}$, let $A_i = \mathrm{I}\left(R(x_i, Z') < N' + a\right), B_i = \mathrm{I}\left(N' + a \leq R(x_i, Z') \leq N' + b\right)$, and $C_i = \mathrm{I}\left(R(x_i, Z') > N' + b\right)$. Then, for given X, Y', m, $f_{X,Y',m}(k)$ is defined as a function of $k \in \{0, \ldots, m - m'\}$ such that

$$f_{X,Y',m}(k) = (k - M') \sum_{i=1}^{n} (A_i - C_i) + \mathrm{sgn}(k - M') \sum_{i=1}^{n} B_i \left(2N' + k - M' - 2R(x_i, Z')\right),$$

where $\mathrm{sgn}(x) = 1$ if $x \geq 0$, otherwise $\mathrm{sgn}(x) = -1$.

4.1.3. Main results and sketch of the proof

Using Definition 4.1.1 and 4.1.2, we state our main results for the bounds of the Ansari-Bradley test statistic in the presence of missing data as follows.

Theorem 4.1.3. Suppose $X = \{x_1, \ldots, x_n\}$ and $Y = \{y_1, \ldots, y_m\}$ are samples of distinct real values. Let $X' \subset X$ and $Y' \subset Y$ be observed subsets with sizes $|X'| = n'$, and $|Y'| = m'$, respectively. Define $a = \max\{(m - n' - 2m' + 1)/2, 0\}$, $b = \min\{(m + n' + 1)/2, m - m'\}$, $c = \max\{(m - n' - 2m')/2, 0\}$ and $d = \min\{(m + n')/2, m - m'\}$. Then, the minimum possible Ansari-Bradley test statistic over all possible missing values is:

$$\min_{\substack{X \setminus X' \in \mathbb{R}^{n-n'}, \\ Y \setminus Y' \in \mathbb{R}^{m-m'}}} T(X, Y) = T(X', Y') + \begin{cases}
\min\limits_{k \in \{a, \ldots, b\}} f_{X',Y',m}(k) + (n^2 - n'^2)/4 & \text{if } \mathcal{C}(1), \\
\min\limits_{k \in \{c, \ldots, d\}} f_{X',Y',m}(k) + (n^2 - n'^2 - 1)/4 & \text{if } \mathcal{C}(2), \\
\min\limits_{k \in \{a, \ldots, b\}} f_{X',Y',m}(k) + (n^2 - n'^2 + 1)/4 & \text{if } \mathcal{C}(3), \\
\min\limits_{k \in \{c, \ldots, d\}} f_{X',Y',m}(k) + (n^2 - n'^2)/4 & \text{if } \mathcal{C}(4),
\end{cases}$$

and the maximum possible Ansari-Bradley test statistic is:

$$\max_{\substack{X\setminus X'\in\mathbb{R}^{n-n'},\\ Y\setminus Y'\in\mathbb{R}^{m-m'}}} T(X,Y) = -\min_{\substack{X\setminus X'\in\mathbb{R}^{n-n'},\\ Y\setminus Y'\in\mathbb{R}^{m-m'}}} T(Y,X) + \begin{cases} N^2/4 & N \text{ is even,} \\ (N^2-1)/4 & N \text{ is odd.} \end{cases} \tag{4.1}$$

Sketch of proof. We only provide a sketch of the proof here, while all the technical details can be found in Appendix B. This theorem proves the tight upper and lower bounds of the Ansari-Bradley test statistic in the presence of missing data. First, note that the sum of the Ansari-Bradley test statistics in both directions $T(X,Y) + T(Y,X)$ is a constant that depends only on the sample sizes n and m. Therefore, it is sufficient to work on the minimum possible Ansari-Bradley test statistic, and the upper bound can be found according to Equation (4.1).

To derive the lower bound, we start by considering the case where only values in X can be missing, while all values in Y are observed. We first demonstrate that, for the case when one single value in X is missing, the Ansari-Bradley test statistic takes the minimum value when the rank of the missing value takes $(1+N)/2$, when N is odd, or $N/2$, when N is even. We then extend this result to the case where multiple values in X are missing: we prove that the Ansari-Bradley test statistic takes the minimum possible value when the values of the missing data take the consecutive ranks centred around $(1+N)/2$.

To further generalise our results to the case when values in Y can be missing, we again start by considering the cases where only values in Y can be missing, while all values in X are observed. When one single value in Y is missing, we prove that the Ansari-Bradley test statistic takes the minimum value when the rank of the missing value in Y takes either the smallest or the largest possible rank, i.e. 1 or N. This result is subsequently generalised to the case where multiple values in Y can be missing. The Ansari-Bradley test statistic is shown to take the minimum value when all missing values in Y take the ranks at both extremes of $\{1,\dots,N\}$. In other words, the ranks of the missing values of Y must be

$$\{1,\dots,m'-k\} \cup \{N-k+1,\dots,N\},$$

where $k \in \{0,\dots,m-m'\}$ is unknown.

For deciding the optimal k to minimise the Ansari-Bradley test statistic, we propose to evaluate all $k \in \{0,\dots,m-m'\}$. For a given k, the value of the Ansari-Bradley statistic is shown to be equal to

$$T(X,Y_k) = T(X,Y') + f_{X,Y',m}(k) + \text{constant}, \tag{4.2}$$

where Y' denotes the observed values in Y, the constant term depends only on the sample sizes n and n', and the function $f_{X,Y',m}(k)$ is defined in Definition 4.1.2. Then, computing $f_{X,Y',m}(k)$ for all $k \in \{0,\dots,m'\}$ allows us to identify the minimum possible Ansari-Bradley test statistic.

Finally, for the general case where values in both X and Y may be missing, we show that the minimum possible value of the Ansari-Bradley statistic is achieved when the missing values in X take consecutive ranks centred around $(1+N)/2$, and the missing values in Y take both extremes of $\{1,\dots,N\}$. This completes the sketch of the proof.

□

Remark 4.1.4. The computational complexity of applying Theorem 4.1.3 for the lower and upper bounds of the Ansari-Bradley test statistic is analysed as follows. To compute the lower bound of the Ansari-Bradley test statistic, $T(X', Y')$ can be computed in linear time once the combined set of observed values $X' \cup Y'$ has been ranked, which requires $\mathcal{O}((n'+m')\log(n'+m'))$ time. Given the ranked values, the computation of $f_{X',Y',m}(k)$ for a fixed k requires a cost of time $O(n')$. Since $b-a, d-c \leq m-m'$, evaluating the minimum of $f_{X',Y',m}(k)$ over the ranges $k \in \{a, \ldots, b\}$ or $k \in \{c, \ldots, d\}$ yields a worst-case cost of $\mathcal{O}\{(m-m')n'\}$. Therefore, Theorem 4.1.3 enables the computation of the lower bound of $T(X, Y)$ with a total computational complexity

$$\mathcal{O}\{(n'+m')\log(n'+m') + (m-m')n'\} = \mathcal{O}(N^2).$$

The upper bound of $T(X, Y)$ requires the same computational complexity.

4.2. Ansari-Bradley test with missing data

This section develops a two-sample scale testing method in the presence of missing data by applying the bounds of the Ansari-Bradley test statistic derived in Theorem 4.1.3. The core idea parallels that of the location test with missing data in Section 3.2: the proposed scale test rejects the null hypothesis only when all possible Ansari-Bradley statistics are significant.

As discussed in Section 3.2, this approach relies on rejection conditions to determine whether all test statistics are significant. Equivalently, one can use the exact bounds of the p-values of the test statistic to decide the significance of the result. This section presents the results directly, since the technical details of the proofs are the same as those in Section 3.2.

To start, let μ and σ denote the mean and variance of the Ansari-Bradley test statistic under the null hypothesis, as defined in Equations (2.7) and (2.8), respectively. Let $T_{\min}(X, Y)$ and $T_{\max}(X, Y)$ denote the minimum and maximum possible Ansari-Bradley statistics, which can be computed using Theorem 4.1.3. The null hypothesis is then rejected if one of the following conditions holds:

Condition 4.2.1.

$$\frac{T_{\min}(X,Y) - \mu}{\sigma} \geq \Phi^{-1}\left(1 - \frac{\alpha}{2}\right),$$

where μ and σ are defined in Equation (2.7) and (2.8), respectively.

Condition 4.2.2.

$$\frac{T_{\max}(X,Y) - \mu}{\sigma} \leq \Phi^{-1}\left(\frac{\alpha}{2}\right),$$

where μ and σ are defined in Equation (2.7) and (2.8), respectively.

The proposed method, which rejects the null hypothesis when either Condition 4.2.1 or Condition 4.2.2 holds, always control the Type I error regardless of the values of missing data. More formally, we state the result as follows.

Theorem 4.2.3. The proposed method, which rejects the null hypothesis whenever Condition 4.2.1 or Condition 4.2.2 holds, has Type I error no larger than the Ansari-Bradley test when all data are observed.

The bounds of the p-value of the Ansari-Bradley test in the presence of missing data are similar to the bounds of the p-value of the Wilcoxon-Mann-Whitney test in Proposition 3.2.4.

Proposition 4.2.4. Suppose that $X = \{x_1, \cdots, x_n\}$ and $Y = \{y_1, \cdots, y_m\}$ are samples of distinct, real-valued observations. Suppose that $X' \subset X$ is a subset of n' values in X, and $Y' \subset Y$ is a subset of m' values in Y. Define $Z = X \cup Y$, $Z' = X' \cup Y'$ and suppose only Z' is known. Then the p-value of the Ansari-Bradley test $p(T(X,Y))$ is bounded such that

$$p(T(X,Y)) \geq p_{\min}(X,Y) = \min\{p(T_{\min}(X,Y)), p(T_{\max}(X,Y))\},$$
$$p(T(X,Y)) \leq p_{\max}(X,Y) = \begin{cases} 1, & \text{if } (T_{\min}(X,Y)-\mu)(T_{\max}(X,Y)-\mu) \leq 0, \\ \max\{p(T_{\min}(X,Y)), p(T_{\max}(X,Y))\}, & \text{otherwise,} \end{cases}$$

where p-value of $T(X,Y)$ is defined as

$$p(T(X,Y)) = 2\min\left\{\Phi\left(\frac{T(X,Y)-\mu}{\sigma}\right), 1-\Phi\left(\frac{T(X,Y)-\mu}{\sigma}\right)\right\}.$$

Finally, we can show that rejecting the null hypothesis using the rejection conditions Condition 4.2.1 and 4.2.2 is equivalent to rejecting using the maximum possible p-values.

Proposition 4.2.5. Suppose that $X = \{x_1, \cdots, x_n\}$ and $Y = \{y_1, \cdots, y_m\}$ are partially observed samples of distinct, real-valued observations. Then for any given significance level $\alpha < 1$,

$$\text{Condition 4.2.1 or 4.2.2 hold} \iff p_{\max}(X,Y) \leq \alpha,$$

where $p_{\max}(X,Y)$ is defined in Proposition 4.2.4. Hence, the p-value of the proposed scale testing method is the maximum possible p-values $p_{\max}(X,Y)$.

4.3. Simulations for the scale test

This section presents numerical simulations to evaluate the Type I error and statistical power of the proposed scale testing method in the presence of missing data. We compare its performance with that of the Ansari-Bradley test applied after mean imputation, hot deck imputation, or case deletion.

4.3.1. As the proportion of missing data varies

We first consider experiments comparing the performance of the proposed method and other methods with varying proportion of missing data s, while sample sizes, missingness mechanisms, and the distributions of the data, are fixed for each experiment.

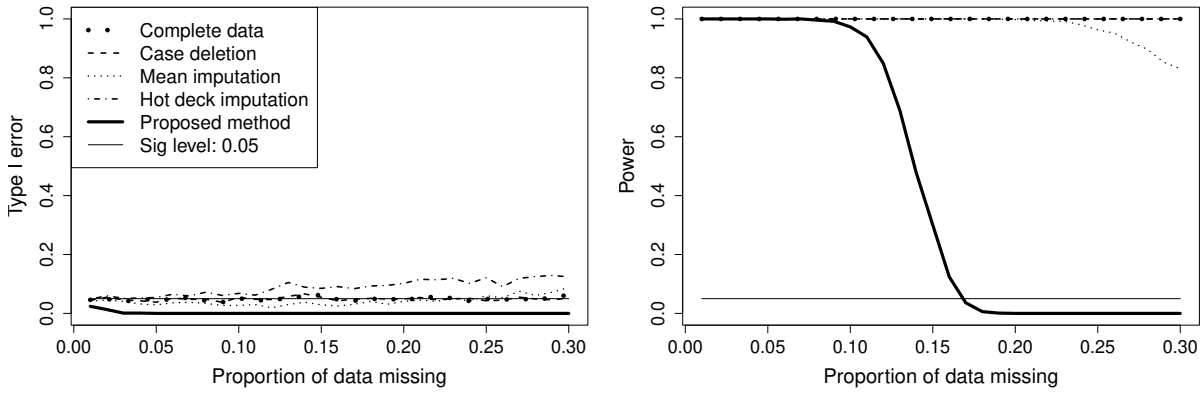


Figure 4.1.: The Type I error and statistical power of the proposed method and the standard Ansari-Bradley test after the missing data is either known or has been imputed or ignored as the proportion of missing data increases. The data is missing completely at random (MCAR). (Left) Type I error: $\mathrm{N}(0,1)$ vs $\mathrm{N}(0,1)$; (Right) Power: $\mathrm{N}(0,1)$ vs $\mathrm{N}(0,\sigma^2)$, with the scale parameter $\sigma = 3$. For both figures, a significance threshold of $\alpha = 0.05$ has been used and the total sample sizes are $n = 100$, $m = 100$, and 1000 trials were used.

Missing completely at random (MCAR)

The first experiment considers the case where data are missing completely at random (MCAR). The data are generated following the same approach used for generating Figure 3.1, except that the observations in Y are sampled independently from $\mathrm{N}(0,\sigma^2)$ with $\sigma = 3$ for evaluating the statistical power.

Figure 4.1 shows that when data are MCAR, the Type I error is controlled by the proposed method, the mean imputation and the case deletion method when the missing data are ignored. However, hot deck imputation method fails to control the Type I error. On the other hand, all methods have good statistical power, except that the power for the proposed method decreases significantly when more than 10% of the data is missing.

Missing not at random (MNAR)

The second experiment is similar to the first, except that here the data are missing not at random (MNAR). The missingness mechanism is defined as follows. Let s denote the proportion of observations to be missing. For any $x_i \in \{x_1, \ldots, x_n\}$, the probability of x_i being missing is

$$\Pr\left(x_i \text{ is missing}\right) = \begin{cases} \min(1, sn/\sum_{j=1}^{n} \mathrm{I}\left(|x_j| < 1\right)), & \text{if } |x_i| < 1, \\ \max(0, sn/\sum_{j=1}^{n} \mathrm{I}\left(|x_j| < 1\right) - 1), & \text{otherwise,} \end{cases} \tag{4.3}$$

where $\mathrm{I}\left(A\right)$ denotes the indicator function of event A. In other words, if the number of values in X with absolute value smaller than one exceeds the number of missing values (i.e., $sn/\sum_{j=1}^{n} \mathrm{I}\left(|x_j| < 1\right) < 1$), then only those values with $|x_i| < 1$ may be missing. Otherwise, all such values will be missing, and additional missingness will occur randomly among the values with $|x_i| \geq 1$.

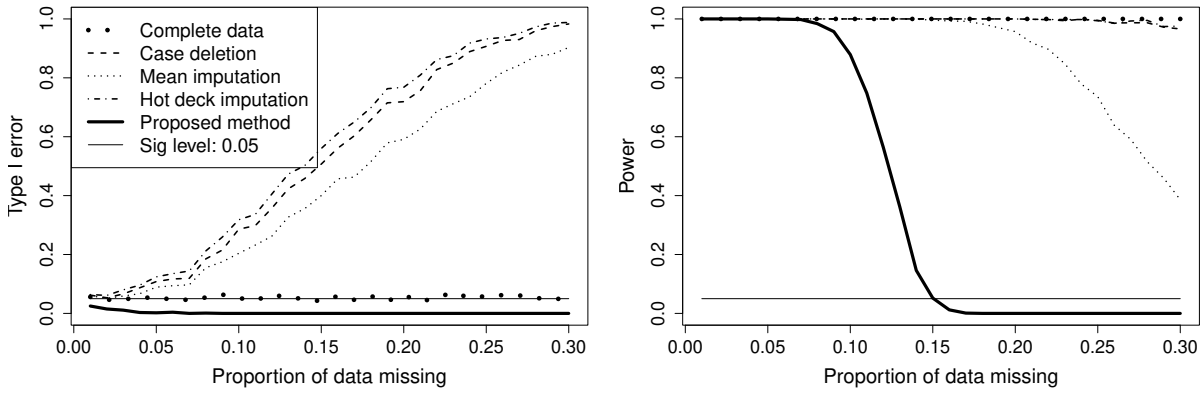


Figure 4.2.: The Type I error and statistical power of the proposed method and the standard Ansari-Bradley test after the missing data is either known or has been imputed or ignored as the proportion of missing data increases. The data is missing not at random (MNAR), according to the mechanism in Equation (4.3), and (4.4). (Left) Type I error: $\mathrm{N}(0,1)$ vs $\mathrm{N}(0,1)$; (Right) Power: $\mathrm{N}(0,1)$ vs $\mathrm{N}(0,\sigma)$, with scale parameter $\sigma = 3$. For both figures, a significance threshold of $\alpha = 0.05$ has been used and the total sample sizes are $n = 100$, $m = 100$, and 1000 trials were used.

For $y_i \in \{y_1, \ldots, y_m\}$, a symmetric missingness mechanism is used:

$$\Pr\left(y_i \text{ is missing}\right) = \begin{cases} \min(1, sm/\sum_{j=1}^{m} \mathrm{I}\left(|y_j| > 1\right)), & \text{if } |y_i| > 1, \\ \max(0, sm/\sum_{j=1}^{m} \mathrm{I}\left(|y_j| > 1\right) - 1), & \text{otherwise,} \end{cases} \tag{4.4}$$

so that missingness depends on whether values in Y exceed one in absolute value. To be specific, if the number of values with $|y_i| > 1$ exceeds the number of missing values (i.e., $sm/\sum_{j=1}^{m} \mathrm{I}\left(|y_j| > 1\right) < 1$), then only those values with $|y_i| > 1$ may be missing. Otherwise, all such values will be missing, and additional missingness will occur randomly among the values with $|y_i| \leq 1$.

These mechanisms, given in Equations (4.3) and (4.4), are chosen to induce different missingness patterns for X and Y.

Figure 4.2 shows that when both samples are MNAR and follow the missingness mechanisms described above, the Type I error is not controlled by either the imputation methods or the case deletion method. This simple example highlights the risks of not taking missing data into account properly. In contrast, the proposed method successfully controls the Type I error rate in this setting. The statistical power of all methods is comparable to that in the missing completely at random (MCAR) case shown in Figure 4.1: they all achieve good power, although the power of the proposed method declines once the proportion of missing data exceeds 10%.

4.3.2. As the sample size varies

This section again considers data that are either MCAR or MNAR, following the mechanisms in Equations (4.3) and (4.4), but now examines the case where the sample sizes n and m increase while the proportion of missing data is fixed at $s = 0.1$.

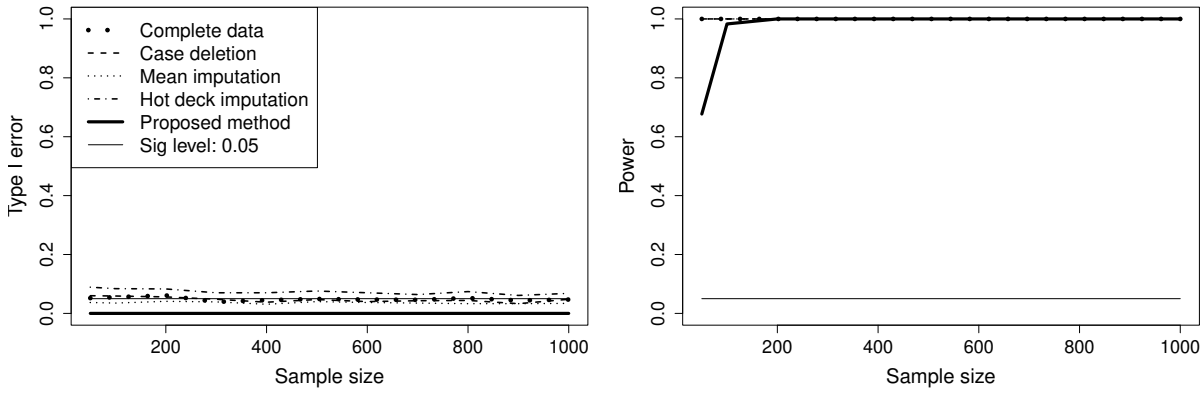


Figure 4.3.: The Type I error and statistical power of the proposed method and the standard Ansari-Bradley test after the missing data is either known or has been imputed or ignored as the sample sizes increase. The data is missing completely at random (MCAR). (Left) Type I error: $\mathrm{N}(0,1)$ vs $\mathrm{N}(0,1)$; (Right) Power: $\mathrm{N}(0,1)$ vs $\mathrm{N}(0,\sigma^2)$, with the scale parameter $\sigma = 3$. For both figures, a significance threshold of $\alpha = 0.05$ has been used and the proportion of missing data for both samples is $s = 0.1$, and 1000 trials were used.

Missing completely at random (MCAR)

The first experiment considers the data to be MCAR. As before, $X = \{x_1, \ldots, x_n\}$ are sampled independently from a $\mathrm{N}(0,1)$ distribution, and observations in $Y = \{y_1, \ldots, y_m\}$ are sampled independently from a $\mathrm{N}(0,1)$ distribution to evaluate the Type I error, and from a $\mathrm{N}(0,\sigma^2)$ with $\sigma = 3$ to evaluate the statistical power. The sample sizes $n \in \{50, 100, 200, \ldots, 1000\}$ are considered, with $m = n$. Then $s = 0.1$ of the observations $X = \{x_1, \ldots, x_n\}$ are selected completely at random to be marked as missing, and the same proportion s of $Y = \{y_1, \ldots, y_m\}$ are selected completely at random to be marked as missing.

Figure 4.3 shows that mean imputation and hot deck imputation fail to control the Type I error for the sample sizes considered, whereas case deletion and the proposed method control the Type I error. All methods achieve good power, with the power of the proposed method increasing as the sample size grows. At a sample size of 200, its power is nearly 1.

Missing not at random (MNAR)

This experiment considers a similar case to Figure 3.3, but the data in this experiment are MNAR, following the mechanism in given by Equation (4.3), and Equation (4.4).

Figure 4.4 further highlights the potential peril of not taking the missing data into account: the Type I error of case deletion, mean imputation, and hot deck imputation methods appears to converge asymptotically to 1 as the sample size increases under a fixed proportion of missing data. In contrast, the proposed method controls the Type I error across all sample sizes considered. The power of all methods in this experiment is similar to that observed in Figure 4.3 for the MCAR case.

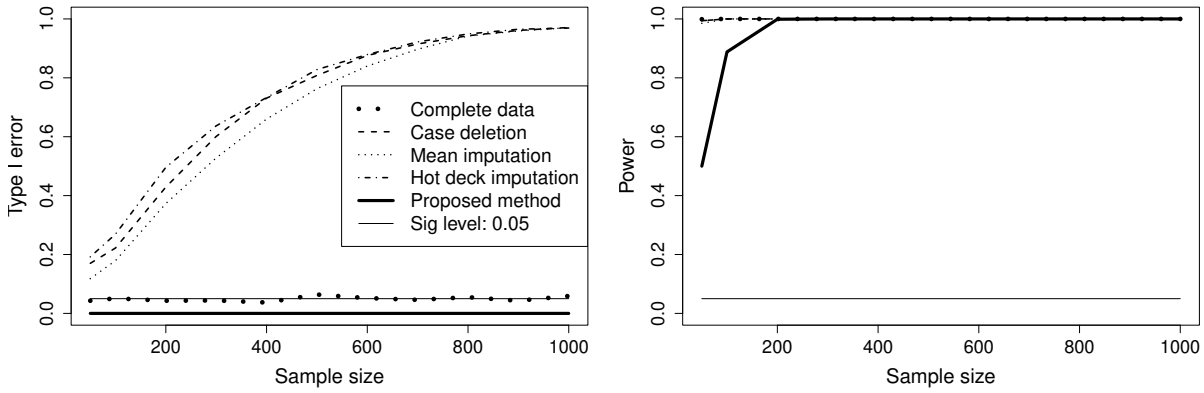


Figure 4.4.: The Type I error and statistical power of the proposed method and the standard Ansari-Bradley test after the missing data is either known or has been imputed or ignored as the sample sizes increase. The data is missing not at random (MNAR), according to the mechanism in Equation (4.3), and (4.4) (Left) Type I error: $\mathrm{N}(0,1)$ vs $\mathrm{N}(0,1)$; (Right) Power: $\mathrm{N}(0,1)$ vs $\mathrm{N}(0,\sigma^2)$ with the scale parameter $\sigma = 3$. For both figures, a significance threshold of $\alpha = 0.05$ has been used and the proportion of missing data for both samples is $s = 0.1$, and 1000 trials were used.

4.3.3. As the scale value varies

Finally, this section considers two experiments to evaluate the power of the proposed method under varying scale parameters σ of the distribution generating sample Y. The sample sizes and the proportion of missing data in both X and Y are kept fixed.

Missing completely at random (MCAR)

The first experiment assumes the data are MCAR. Observations in sample X are drawn independently from a $\mathrm{N}(0,1)$ distribution, while those in sample Y are drawn independently from a $\mathrm{N}(0,\sigma^2)$ with $\sigma \in 1.0, 1.1, \ldots, 4.0$. The sample sizes are fixed at $n = m = 100$, with a missing data proportion of $s = 0.1$.

Figure 4.5 shows that the proposed method begins to gain power once the scale parameter σ exceeds 1.7. At $\sigma = 3$, its power is nearly 1. The power of the other methods also increases with σ and approaches 1 when $\sigma \approx 1.7$.

Missing not at random (MNAR)

We next consider an experiment similar to that in Figure 4.5, but with the data being MNAR according to the mechanisms in Equations (4.3) and (4.4).

The results, shown in Figure 4.6, are similar to those in Figure 4.5, although in this case all methods require a larger scale parameter σ to achieve power close to 1. The proposed method reaches power near 1 at $\sigma = 3.5$, while the other methods achieve similar power at $\sigma = 2$.

Overall, this section evaluates the Type I error and statistical power of the proposed scale testing method in the presence of missing data. This method is compared with the Ansari-Bradley test when the missing data have been imputed using mean or hot deck imputation, or when the data are ignored, or when the data are known, i.e. without missing data.

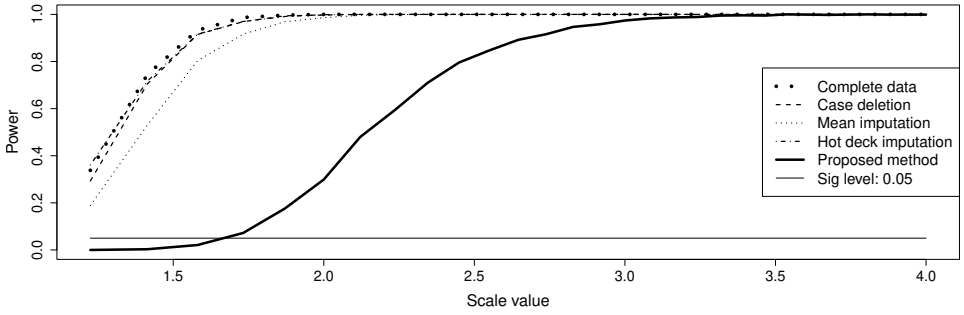


Figure 4.5.: Statistical power of the proposed method and the standard Ansari-Bradley test after the missing data is either known or has been imputed or ignored as the sample sizes increase. The data is missing completely at random (MCAR). Data in samples X and Y are generated from $\mathrm{N}(0, 1)$, and $\mathrm{N}(0, \sigma^2)$, respectively, with σ denoting the scale value. A significance threshold of $\alpha = 0.05$ has been used, the total sample sizes are $n = m = 100$, the proportion of missing data for both samples is $s = 0.1$, and 1000 trials were used.

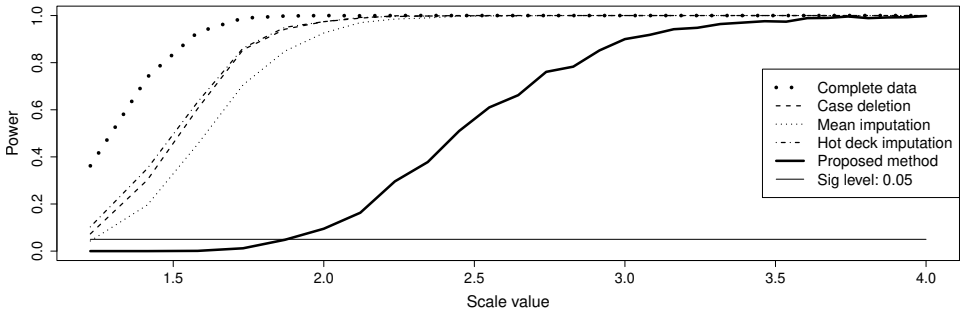


Figure 4.6.: Statistical power of the proposed method and the standard Ansari-Bradley test after the missing data is either known or has been imputed or ignored as the sample sizes increase. The data is missing not at random (MNAR), according to the mechanism in equation (4.3), and (4.4). Data in samples X and Y are generated from $\mathrm{N}(0, 1)$, and $\mathrm{N}(0, \sigma^2)$, respectively, with σ denoting the scale value. A significance threshold of $\alpha = 0.05$ has been used, the total sample sizes are $n = m = 100$, the proportion of missing data for both samples is $s = 0.1$, and 1000 trials were used.

The simulation results in this section are similar to that obtained in Section 3.5: the proposed method controls the Type I error, while imputation methods and case deletion fail to do so, especially when the data are missing not at random (MNAR). The problem of the imputation and case deletion methods appear to be more severe as the sample size increases. For example, Figure 4.4 shows that the Type I error of these methods asymptotically converge to 1 with increasing sample size.

The power of the proposed method drops when more than 10% of data are missing, as shown in Figure 4.1 and 4.2. This is because our method rejects the null hypothesis only when all possible test statistics are significant across all possible values of imputations. When more data are missing, there are more possible different combinations of imputations. Hence the upper and lower bounds of the Ansari-Bradley test statistics increases, and drops, respectively. As a result, the corresponding p-value of the proposed method increases, reducing its power.

The power of the proposed method also depends on sample size and the scale parameter σ. Greater power is achieved with larger sample sizes or larger scale parameters σ of the distribution generating Y, as demonstrated in Sections 4.3.2 and 4.3.3.

Additional experiments are provided in the appendices. Appendix E.2.3 considers non-normal data, where both X and Y are Gamma-distributed. Appendix E.2.4 examines an alternative MNAR mechanism, different from those specified in Equations (4.3) and (4.4).

4.4. Extension to location-scale test with missing data

In this section, we develop a new location-scale testing method that accounts for all possible missing values by combining the scale testing method proposed in this chapter with the location testing method from Chapter 3, using the Holm-Bonferroni correction.

We begin by reviewing the Holm-Bonferroni method, a widely used procedure for controlling the Type I error in multiple hypothesis testing. Specifically, suppose we are testing k hypotheses with p-values $p_1, \ldots, p_k$. For a chosen significance level $\alpha \in (0, 1)$, the Holm-Bonferroni procedure proceeds as follows:

- Sort the p-values in ascending order:

$$p_{(1)} \leq \cdots \leq p_{(k)}.$$

  Let $H_{(1)}, \ldots, H_{(k)}$ be the corresponding null hypotheses.

- For each $i \in \{1, \ldots, k\}$, compare $p_{(i)}$ with the threshold:

$$p_{(i)} \leq \frac{\alpha}{k - i + 1}.$$

- Find the smallest index t such that

$$p_{(t)} > \frac{\alpha}{k - t + 1}.$$

- Reject all null hypotheses $H_{(1)}, \ldots, H_{(t-1)}$ and do not reject $H_{(t)}, \ldots, H_{(k)}$.

Then, the probability of rejecting at least one true null hypothesis among $H_{(1)}, \ldots, H_{(k)}$, i.e., making at least one Type I error, is at most the significance level α.

Applying the Holm-Bonferroni method, a location-scale testing procedure based on the Wilcoxon-Mann-Whitney test and the Ansari-Bradley test can be constructed as follows:

- Compute the p-values of the Wilcoxon-Mann-Whitney test and the Ansari-Bradley test.
- Sort the p-values in ascending order:

$$p_{(1)} \leq p_{(2)}.$$

- Compare $p_{(1)}$ with the thresholds $\alpha/2$. If the inequality $p_{(1)} \leq \alpha/2$ holds, we reject the null hypothesis, otherwise we do not reject the null hypothesis.

As the goal for the location-scale testing problem is to determine if there is location or scale difference, we reject the null hypothesis so long as the null hypothesis for one of the location test or the scale test is rejected. Thus, it suffices to check whether $p_{(1)} \leq \alpha/2$ using the Holm-Bonferroni method: if $p_{(1)} \leq \alpha/2$, then at least one hypothesis is rejected, regardless of the value of $p_{(2)}$; conversely, if $p_{(1)} > \alpha/2$, then both hypotheses are not rejected.

Further, since the null hypothesis is rejected when $p_{(1)} \leq \alpha/2$, we can define the p-value of this location-scale testing method as $p_{\text{Holm}} = \min\{1, 2p_{(1)}\}$, where $p_{(1)}$ is the minimum of the p-values for the Wilcoxon-Mann-Whitney test and the Ansari-Bradley test. More formally, we define

Definition 4.4.1. Let X and Y be samples of distinct real values. Denote the p-values of the Wilcoxon-Mann-Whitney test and the Ansari-Bradley test on X and Y as p_{WMW} and p_{ABT}, respectively. The combined p-value for location-scale testing using the Holm-Bonferroni method is defined as

$$p_{\text{Holm}} = \min\{1, 2p_{(1)}\}, \text{ where } p_{(1)} = \min\{p_{\text{WMW}}, p_{\text{ABT}}\}.$$

Following Definition 4.4.1, the null hypothesis of the location-scale test based on the Holm-Bonferroni method is rejected whenever the combined p-value p_{Holm} is smaller than the significance level α.

To construct a valid location-scale testing method in the presence of missing data, we propose bounding the combined p-value p_{Holm} using the bounds of the Ansari-Bradley p-value established in Proposition 3.2.4, and the bounds of the Wilcoxon-Mann-Whitney p-value from Chapter 3. Specifically, we obtain the following result:

Proposition 4.4.2. Let X and Y be samples of distinct real values, with only subsets of their observations available. Denote $p_{\text{ABT,min}}$ and $p_{\text{ABT,max}}$ as the minimum and maximum possible p-values of the Ansari-Bradley test, and $p_{\text{WMW,min}}$ and $p_{\text{WMW,max}}$ as the minimum and maximum possible p-values of the Wilcoxon-Mann-Whitney test. Then the p-value of the combined location-scale test, constructed by the Holm-Bonferroni procedure, is bounded as follows:

$$\min\{1, 2\min\{p_{\text{ABT,min}}, p_{\text{WMW,min}}\}\} \leq p_{\text{Holm}} \leq \min\{1, 2\min\{p_{\text{ABT,max}}, p_{\text{WMW,max}}\}\}.$$

Proof. To start, recall that the p-value of the combined location-scale test base on the Wilcoxon-Mann-Whitney test and the Ansari-Bradley test is defined in Definition 4.4.1 as

$$p_{\text{Holm}} = \min\{1, 2p_{(1)}\}, \text{ where } p_{(1)} = \min\{p_{\text{WMW}}, p_{\text{ABT}}\}.$$

Since

$$p_{\text{WMW,min}} \leq p_{\text{WMW}}, \text{ and } p_{\text{ABT,min}} \leq p_{\text{ABT}},$$

we have

$$\begin{aligned} & 2\min\{p_{\text{ABT,min}}, p_{\text{WMW,min}}\} \leq 2\min\{p_{\text{WMW}}, p_{\text{ABT}}\} = 2p_{(1)}, \\ \implies & \min\{1, 2\min\{p_{\text{ABT,min}}, p_{\text{WMW,min}}\}\} \leq p_{\text{Holm}}. \end{aligned}$$

Similarly, since

$$p_{\text{WMW}} \leq p_{\text{WMW,max}}, \text{ and } p_{\text{ABT}} \leq p_{\text{ABT,max}},$$

we have

$$\begin{aligned} & 2p_{(1)} = 2\min\{p_{\text{WMW}}, p_{\text{ABT}}\} \leq 2\min\{p_{\text{ABT,max}}, p_{\text{WMW,max}}\} \\ \implies & p_{\text{Holm}} \leq \min\{1, 2\min\{p_{\text{ABT,max}}, p_{\text{WMW,max}}\}\}. \end{aligned}$$

This completes the proof. □

Following Proposition 4.4.2, the proposed location-scale testing method based on the Holm-Bonferroni method rejects the null hypothesis when

$$\min\{1, 2\min\{p_{\text{ABT,max}}, p_{\text{WMW,max}}\}\} \leq \alpha.$$

Then, since the p-value p_{Holm} controls the Type I error, the proposed method also controls the type I error, without assuming missingness mechanisms.

Furthermore, since our location-scale testing method is based on the Holm-Bonferroni method, we can distinguish whether the significant result is driven by a location difference, a scale difference, or both. The specific rejection criteria are defined as follows:

- if $p_{\text{WMW,max}} \leq \alpha/2$, but $p_{\text{ABT,max}} > \alpha$, then only the location test is significant.
- if $p_{\text{ABT,max}} \leq \alpha/2$, but $p_{\text{WMW,max}} > \alpha$, then only the scale test is significant.
- if $p_{\text{WMW,max}} \leq \alpha/2$, $p_{\text{ABT,max}} \leq \alpha$, or $p_{\text{ABT,max}} \leq \alpha/2$, $p_{\text{WMW,max}} \leq \alpha$ then both the location and scale test are significant.
- if $\min\{p_{\text{WMW,max}}, p_{\text{ABT,max}}\} > \alpha/2$, then neither test is significant.

In addition to the Holm-Bonferroni method, one might consider alternative approaches for combining different test results, such as Stouffer's (Stouffer et al., 1949) or Fisher's tests (Fisher, 1970). However, these methods assume that the p-value of the tests are independent. Hence they

are not appropriate for our purpose of combing the Wilcoxon-Mann-Whitney and the Ansari-Bradley test results.

4.4.1. Evaluation of the location-scale test without missing data

Section 4.4 proposed a location-scale testing with missing data method by combining the Wilcoxon-Mann-Whitney test and the Ansari-Bradley test using the Holm-Bonferroni method. However, as we mentioned previously in Section 2.2.4, a more common method for the location-scale testing problem is the Lepage test. The Lepage test statistic is defined as the sum of the squared Wilcoxon-Mann-Whitney test statistic and the Ansari-Bradley test statistic after scaling:

$$L(X,Y) = \left(\frac{W(X,Y)-\mu_w}{\sigma_w}\right)^2 + \left(\frac{T(X,Y)-\mu_t}{\sigma_t}\right)^2,$$

where $W(X,Y)$ denotes the Wilcoxon-Mann-Whitney test statistic, $T(X,Y)$ denotes the Ansari-Bradley test statistic, and μ_w, μ_t and σ_w^2, σ_t^2 are the mean and variance of the Wilcoxon-Mann-Whitney test statistic and the Ansari-Bradley test statistic under the null hypothesis.

An alternative approach for constructing the location-scale testing with missing data method is to adapt the Lepage test following the same framework considered in Sections 3.2 and 4.2, based on the tight bounds for the test statistics in the presence of missing data. However, obtaining the bounds for the Lepage test statistic may be technically challenging. Moreover, it is not immediately clear that the Lepage test constructed such would have greater power than the proposed method based on the Holm-Bonferroni method.

We cannot directly compare the two approaches directly in the presence of missing data, as the bounds of the Lepage test statistic in the presence of missing data are not available. Consequently, we evaluate the two approaches empirically *without* the missing data.

We consider experiments where the sample X is drawn from a distribution $F(x)$, while Y is generated by first sampling $Z \sim F(x)$ and then applying the transformation $Y = Z/a + b$. We evaluate the proposed method against the Lepage test across various choices of a, b, and distribution $F(x)$.

Table 4.1 shows that both the Lepage test and the proposed method control the Type I error, while have comparable power under various alternative hypotheses. When the data are normally distributed, the Lepage test seems to be more powerful when $a \neq 1$ and $b \neq 0$. However, when $a = 1$, $b \neq 0$, or $a \neq 1$, $b = 0$, the proposed method appears to be slightly more powerful.

Apart from the standard normal distribution, we also consider the skewed normal distribution (Azzalini, 1985) to incorporate the case when the skewness of the distribution is non-zero. The skewed normal distribution is a family of distributions that extends the normal distribution by incorporating an additional shape parameter λ to regulate the skewness. Specifically, the density function of a random variable Z following skewed normal distribution is defined as

$$\phi(z;\lambda) = 2\phi(z)\Phi(\lambda z),\ z \in \mathbb{R} \tag{4.5}$$

where ϕ and Φ are the standard normal density and distribution, respectively.

When the data follow a skewed normal distribution, the proposed method appears to have similar or better performance than the Lepage test when $a \neq 1$ and $b \neq 0$. Otherwise the

Table 4.1.: Estimated Type I error and power of the Lepage test (Lep) and the proposed method (PM) based on Holm-Bonferroni method without missing data. The sample X is drawn from distribution $F(x)$, while Y is generated by first sampling $Z \sim F(x)$ and then applying the transformation $Z/a + b$. The sample sizes for X and Y are n and m, respectively. When $a = 1$ and $b = 0$, the table reports the estimated Type I error; otherwise, it reports the estimated power. A significance level of $\alpha = 0.05$ is used. Each entry corresponds to the proportion of null hypothesis rejections over 1000 Monte Carlo replications.

$F(x)$	Parameters		$n = m = 50$		$n = m = 75$		$n = m = 100$		$n = m = 125$	
	a	b	Lep	PM	Lep	PM	Lep	PM	Lep	PM
Normal	1	0	0.039	0.040	0.049	0.050	0.055	0.053	0.039	0.035
	1	0.25	0.167	0.168	0.228	0.241	0.311	0.306	0.390	0.398
	1	0.5	0.556	0.568	0.753	0.755	0.873	0.879	0.931	0.941
	1	0.75	0.906	0.916	0.979	0.980	0.999	0.999	1.000	1.000
	1.25	0	0.173	0.169	0.218	0.221	0.326	0.331	0.389	0.403
	1.5	0	0.461	0.466	0.671	0.676	0.821	0.811	0.903	0.901
	1.75	0	0.750	0.771	0.917	0.920	0.972	0.979	0.992	0.995
	1.25	0.25	0.334	0.293	0.494	0.428	0.626	0.554	0.730	0.669
	1.5	0.5	0.887	0.835	0.984	0.961	0.996	0.991	1.000	0.999
	1.75	0.75	1.000	0.999	1.000	1.000	1.000	1.000	1.000	1.000
Skewed	1	0	0.041	0.037	0.045	0.046	0.052	0.046	0.052	0.048
	1	0.25	0.492	0.466	0.663	0.651	0.798	0.777	0.901	0.892
	1	0.5	0.974	0.972	0.999	0.997	1.000	1.000	1.000	1.000
	1	0.75	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
	1.25	0	0.227	0.216	0.301	0.270	0.414	0.380	0.488	0.440
	1.5	0	0.584	0.526	0.793	0.740	0.917	0.877	0.955	0.923
	1.75	0	0.867	0.822	0.979	0.947	0.997	0.991	1.000	0.998
	1.25	0.25	0.853	0.876	0.965	0.971	0.998	0.998	0.998	0.998
	1.5	0.5	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
	1.75	0.75	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
Student's t	1	0	0.042	0.041	0.052	0.052	0.053	0.041	0.053	0.053
	1	0.25	0.132	0.138	0.173	0.178	0.218	0.212	0.252	0.260
	1	0.5	0.385	0.396	0.572	0.591	0.710	0.727	0.797	0.823
	1	0.75	0.736	0.742	0.902	0.913	0.972	0.974	0.991	0.991
	1.25	0	0.143	0.144	0.198	0.192	0.241	0.241	0.315	0.299
	1.5	0	0.356	0.366	0.521	0.510	0.620	0.627	0.745	0.763
	1.75	0	0.631	0.635	0.806	0.817	0.925	0.928	0.963	0.971
	1.25	0.25	0.237	0.214	0.361	0.327	0.471	0.408	0.539	0.465
	1.5	0.5	0.762	0.686	0.927	0.882	0.978	0.947	0.993	0.986
	1.75	0.75	0.983	0.971	0.998	0.996	1.000	1.000	1.000	1.000

Lep, Lepage test. PM, Proposed method. Normal, standard normal distribution; Skewed Normal, skewed normal distribution with shape parameter $\lambda = 4$, as illustrated in Equation (4.5); Student's t, Student's t distribution with degrees of freedom $\nu = 3$.

proposed method seems to have slightly less power.

We also consider the cases when the data following student's t-distributions to incorporate the case when the distributions have heavy tails, The conclusion for this case is similar to that when the data are normally distributed.

The experimental results in Table 4.1 show that, when the data are fully observed, the Lepage test and the proposed method perform similarly across a range of alternatives. These findings might suggest that, in the presence of missing data, the proposed method could achieve power comparable to that of an approach based on bounds of the Lepage test statistic, although such an implication cannot be guaranteed.

In next section, we further evaluate the Type I error and statistical power of the proposed location-scale testing method in the presence of missing data.

4.5. Simulations for the location-scale test

This section presents numerical simulations to evaluate the Type I error and statistical power of the proposed location-scale testing method in the presence of missing data. We also compare its performance with that of the combined Ansari-Bradley and Wilcoxon-Mann-Whitney tests using the Holm-Bonferroni method, after dealing with the missing data by mean imputation, hot deck imputation, or case deletion.

4.5.1. As the proportion of missing data varies

First, we evaluate the performance of the proposed method and other methods as the proportion of missing data s increases, while keeping the sample sizes, missingness mechanisms, and data distributions fixed.

Missing completely at random (MCAR)

The first experiment considers the case where data are missing completely at random (MCAR). The data are generated using the same procedure as in Figure 3.1, except that the observations in Y are drawn independently from $\mathrm{N}(\mu, \sigma^2)$ with $\mu = 1$ and $\sigma = 3$ to evaluate statistical power.

Figure 4.7 shows that Type I error is not controlled when using hot deck or mean imputation. In contrast, both the proposed method and the case deletion method, which ignores missing data, successfully control the Type I error. However, the power of the proposed location-scale testing method declines noticeably once the proportion of missing data exceeds 10%.

Missing not at random (MNAR)

The second experiment is similar to the first experiment, except that now the data are missing not at random (MNAR) according to the missingness mechanisms specified in Equation (4.3) and Equation (4.4).

Figure 4.8 shows that only the proposed method controls the Type I error when the data are MNAR following the above missingness mechanism, while imputation methods and case deletion method fail to control the Type I error. The power for all the methods are similar to that in Figure 4.7.

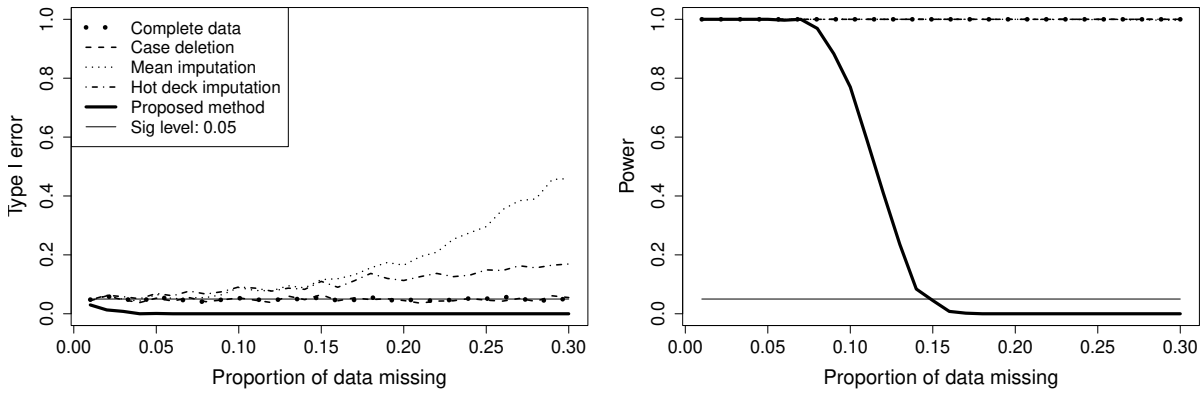


Figure 4.7.: The Type I error and statistical power of the proposed location-scale testing method, and the combination of the Ansari-Bradley test and the Wilcoxon-Mann-Whitney test using the Holm-Bonferroni method after the missing data is either known or has been imputed or ignored as the proportion of missing data increases. The data are missing completely at random (MCAR). (Left) Type I error: $\mathrm{N}(0,1)$ vs $\mathrm{N}(0,1)$; (Right) Power: $\mathrm{N}(0,1)$ vs $\mathrm{N}(\mu,\sigma^2)$, with the location parameter $\mu = 1$, and the scale parameter $\sigma = 3$. For both figures, a significance threshold of $\alpha = 0.05$ has been used and the total sample sizes are $n = 100$, $m = 100$, and 1000 trials were used.

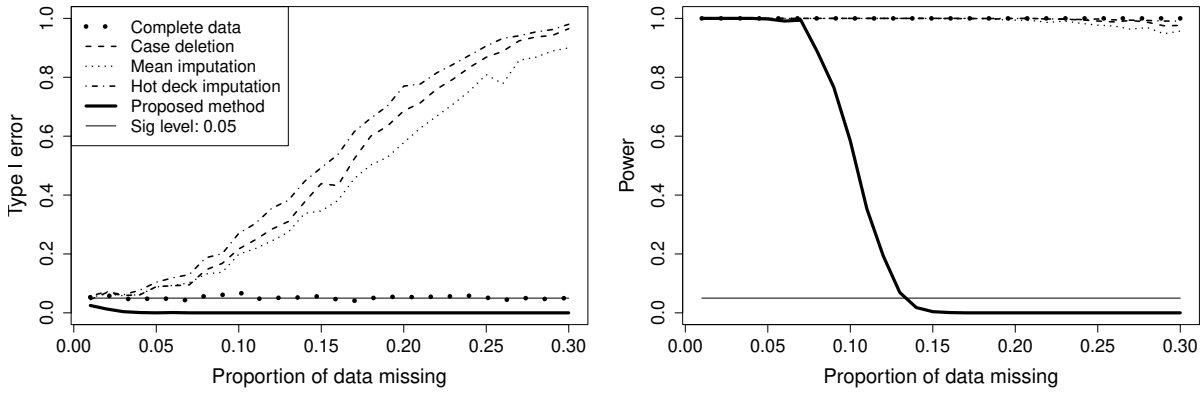


Figure 4.8.: The Type I error and statistical power of the proposed location-scale testing method and the combination of the Ansari-Bradley test and the Wilcoxon-Mann-Whitney test using the Holm-Bonferroni method after the missing data is either known or has been imputed or ignored as the proportion of missing data increases. The data are missing not at random (MNAR), according to the mechanism in Equation (4.3), and (4.4). (Left) Type I error: $\mathrm{N}(0,1)$ vs $\mathrm{N}(0,1)$; (Right) Power: $\mathrm{N}(0,1)$ vs $\mathrm{N}(\mu,\sigma)$, with the location parameter $\mu = 1$ and the scale parameter $\sigma = 3$. For both figures, a significance threshold of $\alpha = 0.05$ has been used and the total sample sizes are $n = 100$, $m = 100$, and 1000 trials were used.

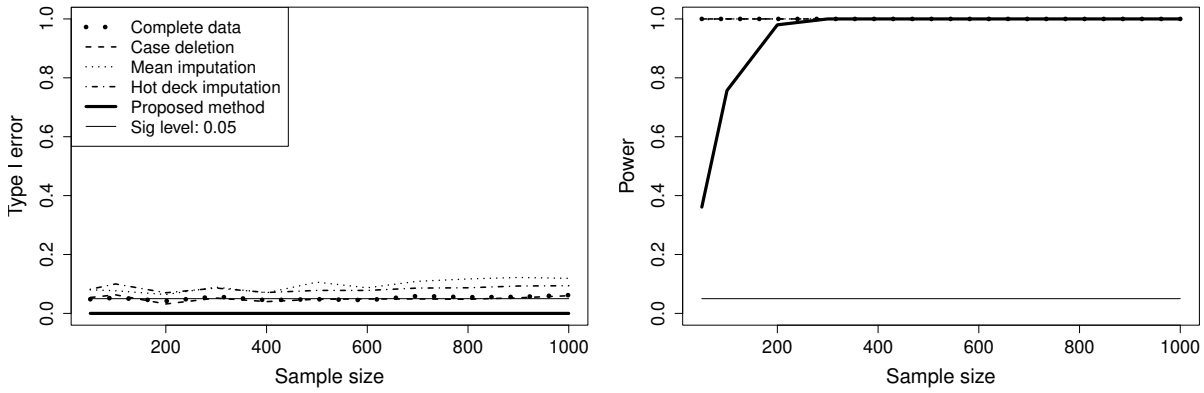


Figure 4.9.: The Type I error and statistical power of the proposed location-scale testing method, and the combination of the Ansari-Bradley test and the Wilcoxon-Mann-Whitney test using the Holm-Bonferroni method after the missing data is either known or has been imputed or ignored as the sample size increases. The data is missing completely at random (MCAR). (Left) Type I error: $\mathrm{N}(0,1)$ vs $\mathrm{N}(0,1)$; (Right) Power: $\mathrm{N}(0,1)$ vs $\mathrm{N}(\mu,\sigma^2)$, with the location parameter $\mu = 1$ and the scale parameter $\sigma = 3$. For both figures, a significance threshold of $\alpha = 0.05$ has been used and the proportion of missing data for both samples is $s = 0.1$, and 1000 trials were used.

4.5.2. As the sample size varies

We now evaluate the performance of the proposed method when the sample sizes n, m increase, with fixed proportion of missing data $s = 0.1$.

Missing completely at random (MCAR)

We again consider the case where the data are MCAR. The sample $X = x_1, \ldots, x_n$ is drawn independently from $\mathrm{N}(0,1)$, while $Y = y_1, \ldots, y_m$ is sampled independently from $\mathrm{N}(0,1)$ for evaluating Type I error, and from $\mathrm{N}(\mu,\sigma)$ with $\mu = 1$ and $\sigma = 3$ for evaluating power. A proportion $s = 10\%$ of observations in both X and Y are then randomly marked as missing.

Figure 4.9 shows that mean imputation and hot-deck imputation fail to control the Type I error for the sample sizes considered, whereas the proposed method and the case deletion method control the Type I error. The power of the proposed method increases with sample size, reaching 1 when the sample size is around 400.

Missing not at random (MNAR)

We now consider an experiment similar to that in Figure 4.9, but with data that are MNAR, following the mechanisms specified in Equations (4.3) and (4.4).

Figure 4.10 shows that the Type I error of case deletion, mean imputation, and hot-deck imputation methods asymptotically converges to 1 as the sample size increases with 10% missing data. In contrast, the proposed method maintains control of the Type I error. The power of all methods is similar to the results obtained in Figure 4.9.

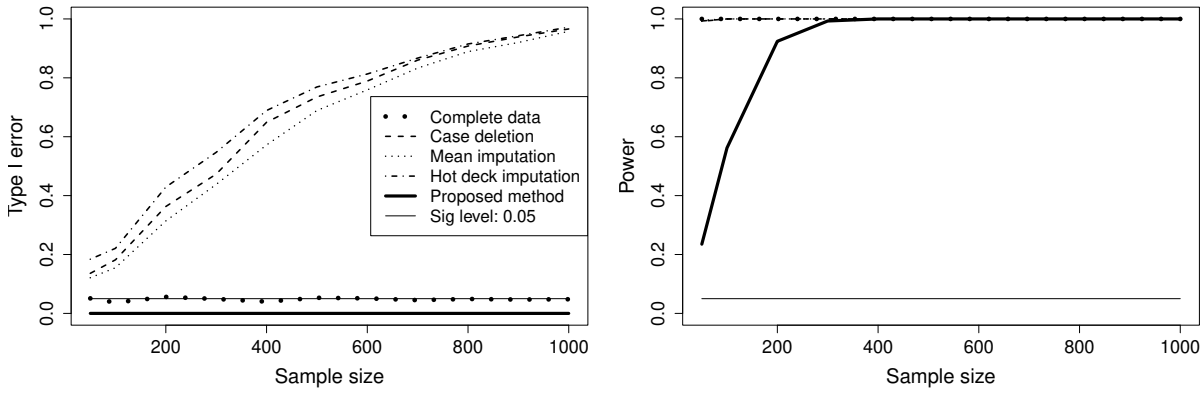


Figure 4.10.: The Type I error and statistical power of the proposed location-scale testing method, and the combination of the Ansari-Bradley test and the Wilcoxon-Mann-Whitney test using the Holm-Bonferroni method after the missing data is either known or has been imputed or ignored as the sample size increases. The data is missing not at random (MNAR), according to the mechanism in Equation (4.3), and (4.4). (Left) Type I error: $\mathrm{N}(0,1)$ vs $\mathrm{N}(0,1)$; (Right) Power: $\mathrm{N}(0,1)$ vs $\mathrm{N}(\mu,\sigma^2)$ with the location parameter $\mu = 1$ and scale parameter $\sigma = 3$. For both figures, a significance threshold of $\alpha = 0.05$ has been used and the proportion of missing data for both samples is $s = 0.1$, and 1000 trials were used.

4.5.3. As the location and scale values vary

This section considers two experiments for evaluating the power of the proposed method with varying mean and scale values of the distribution generating the sample Y. The sample sizes are fixed at $n = m = 100$, with a missing data proportion of $s = 0.1$ for both X and Y.

Missing completely at random (MCAR)

The first experiment again assumes that the data are MCAR. Observations in sample X are drawn independently from a $\mathrm{N}(0,1)$ distribution, while those in sample Y are drawn independently from a $\mathrm{N}(\gamma/2,(1+\gamma)^2)$ distribution, with $\gamma \in \{0.1, 0.2, \ldots, 6.0\}$.

Subsequently, a proportion $s = 0.1$ of the observations in $X = \{x_1, \ldots, x_n\}$ are selected completely at random and marked as missing, and the same proportion s of observations in $Y = \{y_1, \ldots, y_m\}$ are similarly marked as missing.

Figure 4.11 shows that under the settings we considered in this experiment, the proposed method begins to gain power when $\gamma = 0.8$, and when $\gamma = 3$, the power of the proposed method is approaches 1. The power of all other methods increases as soon as $\gamma > 0$, and are close to 1 and when $\gamma = 1$.

Missing not at random (MNAR)

This experiment considers a similar case to Figure 4.5, but the data are MNAR according to the missingness mechanism specified by Equation (4.3), and (4.4).

Figure 4.12 shows results similar to those in Figure 4.11, although all methods require a larger γ to achieve power equal to 1. The power of the proposed method reaches approximately 1 when

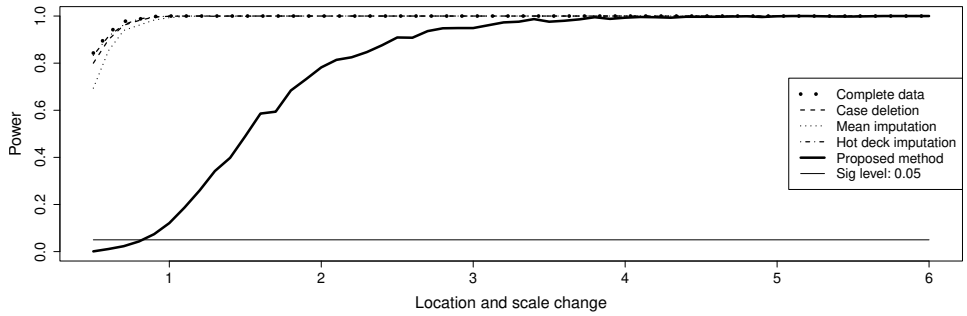


Figure 4.11.: The Type I error and statistical power of the proposed location-scale testing method and the combination of the Ansari-Bradley test and the Wilcoxon-Mann-Whitney test using the Holm-Bonferroni method after the missing data is either known or has been imputed or ignored, as the location and scale values increase. The data is missing completely at random (MCAR). Data in samples X and Y are generated from $\mathrm{N}(0,1)$, and $\mathrm{N}(\gamma/2,(1+\gamma)^2)$, respectively, with γ denoting the shift parameter controlling the shift in both location and scale. A significance threshold of $\alpha = 0.05$ has been used, the total sample sizes are $n = m = 100$, the proportion of missing data for both samples is $s = 0.1$, and 1000 trials were used.

$\gamma = 3.5$.

This section performs numerical simulations to evaluate the Type I error and statistical power of the proposed location-scale method in the presence of missing data. The proposed method is compared with the combination of the Wilcoxon-Mann-Whitney test and the Ansari-Bradley test using the Holm-Bonferroni procedure, after missing data are dealt with by mean imputation, hot-deck imputation, or case deletion.

The simulation results show that the proposed method controls the Type I error, while imputation methods and case deletion yield Type I errors above the significance level α, particularly when the data are missing not at random (MNAR). For a fixed proportion of missing data, the Type I error of imputation and case deletion methods appears to asymptotically converge to 1 as the sample size increases, at least under the experimental settings considered in Figure 4.10.

The proposed method also demonstrates good statistical power, provided moderate total sample sizes n, m for X and Y, along with sufficient mean and scale differences in the distributions generating X and Y, and a reasonable proportion of missing data. For example, when $n = m = 100$, X is sampled from $\mathrm{N}(0,1)$ and Y from $\mathrm{N}(\mu,\sigma^2)$ with $\mu = 1$ and $\sigma = 3$, and 10% of data in both X and Y are missing, the proposed method achieves power close to 80% under MCAR and 50% under MNAR, as shown in Figures 4.7 and 4.8.

The power of the proposed method increases further with larger sample sizes n, m, or larger mean differences in the distributions generating X and Y, as demonstrated in Section 4.5.2 and Section 4.3.3.

Although the experiments in this section are restricted to normally distributed data, the proposed method applies to data generated from any continuous distribution, as it only requires distinct observations. Additional experiments with non-normal data are included in Appendix E.1.3, where both X and Y are sampled from gamma distributions instead of normal distributions. Appendix E.1.4 examines a different MNAR mechanism from that specified in Equations (4.3) and

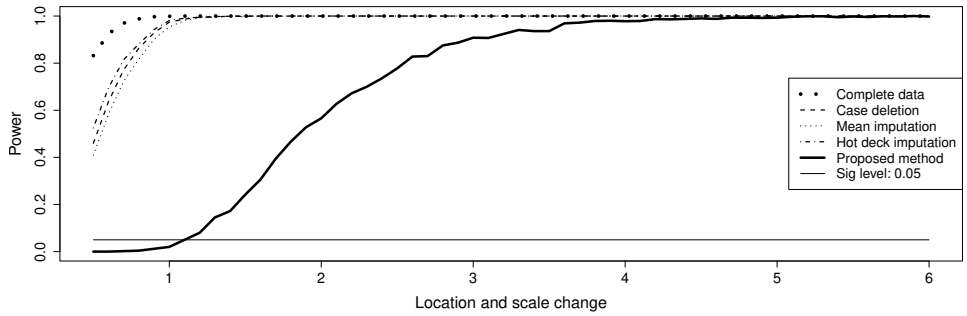


Figure 4.12.: The Type I error and statistical power of the proposed location-scale method and the combination of the Ansari-Bradley test and the Wilcoxon-Mann-Whitney test using the Holm-Bonferroni method after the missing data is either known or has been imputed or ignored, as the location and scale values increase. The data is missing not at random (MNAR), according to the mechanism in Equation (4.3), and (4.4). Data in samples X and Y are generated from $\mathrm{N}(0, 1)$, and $\mathrm{N}(\gamma/2, (1+\gamma)^2)$, respectively, with γ denoting the shift parameter controlling the shift in both location and scale. A significance threshold of $\alpha = 0.05$ has been used, the total sample sizes are $n = m = 100$, the proportion of missing data for both samples is $s = 0.1$, and 1000 trials were used.

(4.4), and produces similar outcomes.

4.6. Application to hepatitis C virus data

To further illustrate our proposed location-scale testing method, we apply it to a hepatitis C virus (HCV) dataset that contains missing values. HCV infection is one of the most important causes of chronic liver disease worldwide (Yağanoğlu, 2022). The infection is known to alter lipid metabolism (Gyamfi et al., 2019), and patients often exhibit reduced serum cholesterol (CHOL) levels compared to healthy controls, with levels declining further as liver injury progresses (Liou et al., 2022).

To investigate this effect, we analyse the publicly available HCV dataset from the UCI Machine Learning Repository (Lichtinghagen et al., 2020), which contains data from 615 individuals with 14 features. These features include demographic variables such as age and gender, as well as a panel of laboratory measurements for blood donors (without HCV infection) and HCV patients at different disease stages (hepatitis, fibrosis, and cirrhosis), including CHOL, and other liver function biomarkers.

While most studies, such as Corey et al. (2009) and Sheridan et al. (2022), emphasise the mean differences in CHOL levels across liver disease stages, our primary focus in this section is to evaluate whether the levels of CHOL differ significantly in either location or scale among HCV patients at three distinct disease stages: hepatitis, fibrosis, and cirrhosis. A difference in location means that one group tends to have overall higher or lower CHOL levels than another group. A difference in scale means that CHOL levels are more spread out in one group than another.

Briefly, hepatitis represents the early stage of liver injury characterised by inflammation, fibrosis reflects the progressive stage marked by accumulation of scar tissue, and cirrhosis denotes the

late stage with advanced, irreversible scarring and significant loss of liver function (Khullar and Firpi, 2015).

The data for the three groups are shown in Table 4.2. Note that the original data set retains only two decimal place, resulting in several tied observations. For the purpose of this example, we do not adjust for ties; instead, we add small random perturbations at the third decimal place to break them.

Table 4.2.: Cholesterol (CHOL) levels of patients at different stages of HCV infection: hepatitis, fibrosis and cirrhosis. Hepatitis represents the early stage of liver injury characterised by inflammation, fibrosis reflects the progressive stage marked by accumulation of scar tissue, and cirrhosis denotes the late stage with advanced, irreversible scarring and significant loss of liver function (Khullar and Firpi, 2015). Missing values are indicated by an asterisk *.

Hepatitis	6.008	3.903	3.206	4.280	3.201	3.606	6.110	6.281	5.490	5.372	4.140	5.734	4.894
(early)	9.679	4.075	4.659	7.046	5.238	3.094	6.154	5.728	5.308	4.674	4.624		
Fibrosis	5.421	4.209	4.018	4.220	3.503	5.202	4.705	4.009	5.859	4.645	4.453	4.523	4.709
(progress)	4.818	4.316	6.195	4.898	3.107	4.208	5.006	*					
Cirrhosis	3.205	4.829	5.346	3.932	3.719	*	4.107	5.501	3.601	4.085	3.545	2.400	1.432
(late)	3.922	3.687	3.615	3.516	4.885	*	3.296	5.170	3.820	3.783	3.987	4.518	6.308
	3.026	3.633	4.206	5.305									

Notice that one value is missing in the fibrosis group and two values are missing in the cirrhosis group. Since these groups contain 21 and 30 observations, respectively, the corresponding proportions of missing data for the two groups are approximately 5% and 7%. The reason that the values are missing is not explained by its source (Lichtinghagen et al., 2020).

Table 4.3.: p-values of the proposed location-scale testing method, and the combined Ansari-Bradley test and Wilcoxon-Mann-Whitney test using the Holm-Bonferroni method after mean imputation, hot deck imputation, or ignoring missing data.

Methods	Hep. vs Fib.	Hep. vs Cir.	Fib. vs Cir.
Case Deletion	0.0991	0.0070	0.0277
Mean Imputation	0.0832	0.0060	0.0161
Hot Deck Imputation	0.0803	0.0045	0.0135
Prop. (min)	0.0573	0.0026	0.0063
Prop. (max)	0.2203	0.0404	0.2389

Hep.: Hepatitis. Fib.: Fibrosis. Cir.: Cirrhosis. Prop.: proposed method. Proposed method provides both the minimum and maximum possible p-values according to Proposition 4.4.2.

To examine whether the CHOL levels differs either in location or scale across the three groups of HCV patients, we apply the combined Ansari-Bradley test and Wilcoxon-Mann-Whitney test with the Holm-Bonferroni method following mean imputation, hot deck imputation, or case deletion for missing data. The p-values of these methods are computed according to Definition 4.4.1. The proposed method provides both the minimum and maximum possible p-values in the presence of missing data, according to Proposition 4.4.2. The results are summarised in Table 4.3.

Using the proposed method, we find that CHOL levels in the hepatitis and fibrosis groups are not significantly different at the $\alpha = 0.05$ significance level, regardless of the missing value

present in the fibrosis group. In contrast, CHOL levels for the hepatitis and cirrhosis groups are significantly different in either location or scale, and this conclusion is unaffected by the missing data. Finally, for the fibrosis versus cirrhosis comparison our method does not yield significant result, since the testing result depends on the values of missing data. By contrast, all other missing data methods yield significant results for this comparison.

In summary, our method identifies a significant difference only between the hepatitis and cirrhosis groups, corresponding to a comparison of patients with early-stage versus late-stage liver disease. Notably, this conclusion is robust to any possible values of the missing data. We emphasise that the purpose of this section is to illustrate the impact of missing data and to demonstrate the utility of the proposed method, rather than to provide clinical recommendations.

4.7. Conclusion

This chapter introduces novel methods for scale testing and location-scale testing in the presence of missing data. We first derive the exact bounds of the Ansari-Bradley test statistic under missing data, assuming only that the data are distinct. Based on these bounds, we propose to reject the null hypothesis only when all possible test statistics are significant. The proposed scale testing method is then combined with the location testing method from Chapter 3 using the Holm-Bonferroni procedure to address the location-scale testing problem with missing data.

Numerical simulations in this chapter yield results similar to those in Section 3.5. When the data are missing not at random (MNAR), common approaches such as case deletion, mean imputation, and hot-deck imputation fail to control the Type I error, whereas our method successfully controls the Type I error. Furthermore, the proposed methods exhibit good statistical power when approximately 10% of the data are missing. Finally, for illustrating the practical utility of our methods, we apply the proposed location-scale testing method to a hepatitis C virus dataset with missing values.

5. Independence test with arbitrarily missing data

This chapter describes the work in (Zeng et al., 2025b). We consider the *independence* testing problem in the presence of missing data when the data are assumed to be distinct. As discussed in Section 2.1.2, this problem considers a single sample of paired observations

$$(x_1, y_1), \ldots, (x_n, y_n) \overset{\text{i.i.d.}}{\sim} F_{X,Y},$$

and the goal is to decide whether the distributions generating $X = \{x_1, \ldots, x_n\}$ and $Y = \{y_1, \ldots, y_n\}$ are statistically independent.

Common rank-based statistical association measures applied for independence testing include Spearman's rank correlation and Kendall's τ, as discussed in Section 2.2.4. A closely related, but less widely known measure is Spearman's footrule, which has seen a resurgence of interest since Diaconis and Graham (1977) established its asymptotic normality under the null hypothesis (see Section 2.2.3). Spearman's footrule is a natural choice for testing independence (Chen et al., 2023; Genest et al., 2010; Luigi and Nikitin, 1999), and has been shown to have greater statistical power than Spearman's rank correlation and Kendall's τ under certain alternatives (Chen et al., 2023; Genest et al., 2010; Luigi and Nikitin, 1999).

This chapter focuses on Spearman's footrule for independence testing in the presence of missing data, without making any assumptions of the missingness mechanisms or missingness patterns. The key idea for the independence testing method in this chapter is the same as those in Chapter 3 and 4: we first derive the tight bounds for the test statistic in the presence of missing data across all possible imputations, and then propose to reject the null hypothesis only when all possible imputations produce significant testing results.

Section 5.1 provides main results for the tight lower and upper bounds of Spearman's footrule with missing data. We only sketch the proofs in Section 5.1, while all the technical details can be found in Chapter C in the Appendix. For better illustrating the main results in Section 5.1, two simple examples are provided.

Section 5.2 discusses algorithms for computing the tight lower and upper bounds. The procedure for the lower bound is provided in Algorithm 3 in Section C.3.2 of the Appendix, while the procedure for the upper bound is given in Algorithm 8 in Section C.5.4 of the Appendix. The computational complexity of the algorithms for the tight lower and upper bounds is $\mathcal{O}(n^2)$, and $\mathcal{O}(n^3)$, respectively, as discussed in Remark C.3.14 and Remark C.5.21.

In Section 5.3, we derive non-trivial lower and upper bounds of Kendall's τ using our tight bounds of Spearman's footrule. These bounds are further evaluated in Section 5.4 using numerical simulations when the data are missing completely at random (MCAR) and missing not at random (MNAR).

A novel independence testing method for data with missing values is proposed in Section 5.5, based on our tight bounds for Spearman's footrule. The key idea is the same as that introduced in Section 3.2: we reject the null hypothesis only when all possible test statistics are significant. To this end, we provide rejection conditions to determine whether all test statistics are significant and also derive bounds for the p-value.

The performance of our proposed independence testing method is evaluated in Section 5.6 using numerical simulations. The results show that our method has good statistical power, when less than 10–15% of the data are missing. Importantly, our method controls the Type I error regardless of the values of missing data, while all other methods fail to control the Type I error when the data are MNAR. In the end of this chapter, we illustrate our method on real-world data by comparing the dependence of yearly precipitation levels in the United Kingdom, Germany and the Unites States.

Although this chapter focuses on Spearman's footrule, the proposed framework could essentially be applied to Kendall's τ and Spearman's rank correlation. However, calculating the exact bounds for these statistics under missing data is non-trivial. While deriving their bounds under specific missingness cases are feasible, such as when only sample X contains missing data. However, it is challenging to derive their bounds for general missingness patterns. To our knowledge, such general bounds have not been established in the existing literature. We choose Spearman's footrule because it is a powerful test statistic for testing independence, and we can derive precise bounds in the presence of missing data, as we will demonstrate in this chapter.

5.1. Bounds of Spearman's footrule

This section presents the tight lower and upper bounds of Spearman's footrule under the assumption that the data are distinct. We only sketch the proofs in this section, while all the technical details can be found in Chapter C of the Appendix.

5.1.1. Intuition for the bounds and main results

As before, we provide intuition for the bounds and the main results presented in this chapter before diving into formal mathematical results in Theorem 5.1.1 and 5.1.5. We start by recalling the definition of Spearman's footrule in Definition 2.2.1:

$$D\left(X,Y\right)=\sum_{i=1}^{n}|R(X(i),X)-R(Y(i),Y)|.$$

The definition implies that the statistic is minimized when the paired ranks are equal and maximized when the paired ranks occupy opposite extremes of all possible ranks. However, proving the bounds is non-trivial, as we cannot optimize each pair in isolation. The ranks of the data depends on each other, so assigning a rank to one missing value could shift the ranks of other data points in the sample. Hence, the precise conditions for the minimum and maximum bounds are not immediately obvious.

We employ different strategies for finding the lower and upper bounds. For proving the result Theorem 5.1.1 for the lower bound, we sequentially construct imputations of the partially observed

vectors, each with a non-increasing value of Spearman's footrule. This is explained in more detail in the sketch of proof for the theorem. Intuitively, to minimize the Spearman's footrule, we force alignment of all pairs with missing data. We show that each pair with missing data should have equal ranks as shown in condition (i) and (ii) in Theorem 5.1.1. Furthermore, for the pairs where both values are missing, we require the values to be larger than all other components in the same sample, as shown in condition (3) in Theorem 5.1.1.

For deriving the result Theorem 5.1.5 for the upper bound, we start by considering the set of all possible values of Spearman's footrule for imputations of X and Y, and then the size of this set is gradually reduced using several constraints. Definition 5.1.2 and 5.1.3 consider the pairs with only one value is missing, while the other one is observed. Definition 5.1.2 requires all the missing values to be either smaller or larger than all observed values. Definition 5.1.3 is based on Definition 5.1.2, and further requires the missing values to be ordered according to the value of the paired values. Definition 5.1.4 considers the pairs where both X and Y are missing, and requires the data in X to be smaller or larger than all observed data, and the sum of ranks of each pair to be $n+1$. Based on these constraints, Theorem 5.1.5 shows that the maximum Spearman footrule is taken when

- all missing values in X with paired data observed: (1) take values either smaller or larger than the observed X values, and (2) arranged in reverse order relative to their paired values;
- similarly, all missing values in Y with paired data observed: (1) take values either smaller or larger than the observed Y values, and (2) arranged in reverse order relative to their paired values;
- for pairs where both X and Y are missing: (1) the values assigned to X are smaller or larger than all other entries in X, including both observed values in X and the missing values in X with paired data observed, and (2) the sum of the ranks for each pair equals $n+1$.

5.1.2. Definitions and notation

For describing our results, we first introduce some necessary definitions and notation. While Chapter 3 and Chapter 4 deal with sets of data in two samples X and Y, this chapter focuses on paired *vectors* (X, Y).

For any n-dimensional vector X, let $X(i)$ denote the component of X at index $i \in [n]$. Thus, an n-dimensional vector X can be written as $X = (X(1), \ldots, X(n))$. For any $n \in \mathbb{N}$, let $[n]$ denote the set $\{1, \ldots, n\}$.

This chapter only considers vectors with distinct real values, i.e. no ties. Define $\mathcal{V}^n$ to be the set of n-dimensional vectors with distinct real values; that is, if $X \in \mathcal{V}^n$ then $X \in \mathbb{R}$ and all components of X are distinct.

If $X \in \mathcal{V}^n$ and $U \subset [n]$ is a set of indices, we say that $X_* \in \mathcal{V}^n$ is an *imputation* of X if $X_*(i) = X(i)$ for all $i \in [n] \setminus U$, and the *imputed values* are the components $X_*(u)$ for all $u \in U$.

For any subset $S \subset [n]$ of indices, we use the notation $(X(l))_{l \in S}$ to denote a vector including the components of X corresponding to the indices in set S, with the order of components in $(X(l))_{l \in S}$ following the order of components in X. For example, if $n = 5$ and $S = \{1, 3, 4\}$, we then have $(X(l))_{l \in S} = (X(1), X(3), X(4))$.

Let $X, Y \in \mathcal{V}^n$, which may be only partially observed. We use the notation U to denote the set of all indices where $X(i)$ is missing but $Y(i)$ is observed. Furthermore, we use the notation V to denote the set of all indices where $X(i)$ is observed but $Y(i)$ is missing. Additionally, we use the notation W to denote the set of all indices where both $X(i)$ and $Y(i)$ are missing.

Finally, using the definition of rank in Definition 2.2.1, the Spearman's footrule is defined formally as:

$$D(X, Y) = \sum_{i=1}^{n} |R(X(i), X) - R(Y(i), Y)|.$$

5.1.3. Main results for the lower bounds and sketch of the proof

Using the definitions and notation introduced in Section 5.1.2, we now present the main results for the tight lower bounds of Spearman's footrule in the presence of missing data, without making any assumptions about the missingness mechanisms or patterns.

Theorem 5.1.1. Suppose $X, Y \in \mathcal{V}^n$ and U, V, $W \subset [n]$ are pairwise disjoint subsets. Suppose $X_*, Y_* \in \mathcal{V}^n$ are imputations of X, Y for indices $U \cup W$, and $V \cup W$, respectively. If the following three conditions all hold

$$\begin{aligned}
&\text{(i) } R(X_*(i), X) = R(Y_*(i), Y), \text{ for all } i \in U \cup V \cup W, \\
&\text{(ii) } \min_{j \in W} X_*(j) > \max_{j \in [n] \setminus W} X_*(j), \\
&\text{(iii) } \min_{j \in W} Y_*(j) > \max_{j \in [n] \setminus W} Y_*(j),
\end{aligned}$$

then

$$D(X_*, Y_*) \leq D(X, Y),$$

and for any other imputations $X', Y' \in \mathcal{V}^n$ of X, Y for indices $U \cup W$, $V \cup W$, respectively,

$$D(X_*, Y_*) \leq D\left(X', Y'\right).$$

Sketch of proof. We provide only sketch of proof here, while all the technical details can be found in Section C.3 in the Appendix.

We start our proof by noting that the general missingness case, i.e. all possible missingness configurations, can be shown as a combination of three different cases, namely missing case I, II, and III.

The three missingness cases are defined as follows. The missingness pattern is missing case I if all missing values are either all in X or all in Y. Without loss of generality, we assume that only X contains missing values when we refer to missing case I in the following. The missingness pattern is defined as missing case II if for any pair $(X(i), Y(i))$, $i \in [n]$, at most one value is missing. In other words, we cannot have both $X(i)$ and $Y(i)$ missing for any particular i. Finally, the missingness pattern is missing case III if for any pair $(X(i), Y(i))$, $i \in [n]$, the two values are either both observed, or both missing.

To prove Theorem 5.1.1, which establishes the lower bound under the general missingness case, we construct imputation vectors that achieve the lower bounds in cases I, II, and III separately, and then combine these results.

For case I, the strategy is to sequentially construct imputations $X_1, X_2, X_3 \in \mathcal{V}^n$ of X for the set of indices U, where values in X are unobserved. We show that each imputation leads to a non-increasing Spearman's footrule:

$$D(X,Y) \geq D(X_1,Y) \geq D(X_2,Y) \geq D(X_3,Y),$$

and X_3 satisfies the desired conditions:

$$X_3(i) = X(i) \text{ for } i \in [n] \setminus U, \text{ and } R(X_3(i), X_*) = R(Y(i), Y) \text{ for } i \in U.$$

Hence, the lower bound of Spearman's footrule $D(X,Y)$ equals $D(X_3,Y)$, which can be computed using Algorithm 2. The strategy for case II follows similarly.

For case III, we show that the lower bound provided in Theorem 1 of Papaioannou and Loukas (1984) is tight, and we further provide specific conditions under which this bound is achieved.

By combining the results for cases I, II, and III, we establish Theorem 5.1.1. Thus, the problem of obtaining the tight lower bound of Spearman's footrule reduces to finding imputations that satisfy conditions (i), (ii), and (iii). We prove that such imputations always exist, and provide an algorithm for computing the tight lower bounds in Algorithm 3. This completes the sketch of proof.

□

To illustrate Theorem 5.1.1, Table 5.1 presents a simple example where both X and Y contain missing values, and the imputations X_*, Y_* satisfy the three conditions specified in Theorem 5.1.1. Consequently, these imputations yield the minimum possible value of Spearman's footrule across all admissible imputations.

5.1.4. Main results for the upper bounds and sketch of the proof

We now turn to the tight upper bound of Spearman's footrule under the assumption that the data are distinct. To present our results, we first introduce three definitions.

The first definition specifies a set of vectors $Z \in \mathcal{V}^n$ such that, for a subset $U \subset T \subset [n]$, all components of Z indexed by U are either smaller than, or larger than, all components of Z indexed by $T \setminus U$.

Definition 5.1.2. Let $U \subset T \subset [n]$ be a subset of indices. Define

$$\mathcal{Z}(U,T,n) = \{Z \in \mathcal{V}^n | \text{ for any } i \in U,\ Z(i) > \max_{j \in T \setminus U} Z(j) \text{ or } Z(i) < \min_{j \in T \setminus U} Z(j)\}.$$

The second definition introduces a subset of $\mathcal{Z}(U,T,n)$ by further ordering the components of Z in U according to the values of the corresponding values in Y.

Definition 5.1.3. Suppose $Y \in \mathcal{V}^n$ and let $U \subset T \subset [n]$. Define

$$\mathcal{X}(Y;U,T,n) = \{Z \in \mathcal{Z}(U,T,n) | \text{ for any } i,j \in U,\ Z(i) > Z(j) \text{ if } Y(i) < Y(j)\}.$$

Table 5.1.: (Left) Example ranked data under general missingness case, where we do not make any assumptions of the missingness patterns. (Right) Example ranked vectors of imputations X_*, Y_*. The imputed components are underlined, and the imputed values of a, b, d, e, h satisfying the three conditions in Theorem 5.1.1. This imputation therefore gives a lower bound of Spearman's footrule.

	a	b	c	d	e	f	g	h
$R(X, X)$	5	1	4	*	*	3	2	*
$R(Y, Y)$	*	*	1	2	3	4	5	*

	a	b	c	d	e	f	g	h
$R(X_*, X_*)$	7	1	6	$\underline{3}$	$\underline{4}$	5	2	$\underline{8}$
$R(Y_*, Y_*)$	$\underline{7}$	$\underline{1}$	2	3	4	5	6	$\underline{8}$

Finally, by applying Definition 5.1.3, the third definition introduces a set of paired vectors $(Z_1, Z_2) \in (\mathcal{Z}(W, [n], n), \mathcal{V}^n)$, such that for every component in W, the sum of the paired ranks is equal to $n + 1$.

Definition 5.1.4. Let $W \subset [n]$ be a subset of indices. Define $\mathcal{Y}(W, n)$ as

$$\mathcal{Y}(W, n) = \{(Z_1, Z_2) \in (\mathcal{Z}(W, [n], n), \mathcal{V}^n) | R(Z_1(i), Z_1) + R(Z_2(i), Z_2) = n + 1,\ i \in W\}.$$

Using Definition 5.1.2, 5.1.3 and 5.1.4, we now present the main results for the tight upper bounds of Spearman's footrule.

Theorem 5.1.5. Suppose $X, Y \in \mathcal{V}^n$, and let U, V, $W \subset [n]$ be pairwise disjoint subsets. Then, there exist $(X_*, Y_*) \in (\mathcal{X}(Y; U, [n] \setminus W, n), \mathcal{X}(X; V, [n] \setminus W, n)) \cap \mathcal{Y}(W, n)$ of imputations X and Y for indices $U \cup W$ and $V \cup W$, respectively, such that $D(X, Y) \leq D(X_*, Y_*)$. Furthermore, consider any imputation $X', Y' \in \mathcal{V}^n$ of X, Y for indices $U \cup W$ and $V \cup W$, respectively. Then we have $D(X', Y') \leq D(X_*, Y_*)$.

Sketch of proof. We provide only sketch of proof here, while all the technical details can be found in Section C.4 in the Appendix.

Similar to the proof of Theorem 5.1.1 for the tight lower bound, the proof of Theorem 5.1.5 for the tight upper bound under the general missingness pattern is also obtained by first proving the results under missing case I, II and III, separately. These results are then combined together for proving Theorem 5.1.5.

While the lower bound is obtained by sequentially constructing imputations of the partially observed vectors with non-increasing Spearman's footrule, the upper bound is derived by first considering the set of all possible values of Spearman's footrule for imputations of X and Y, and then narrowing this set using several constraints.

Specifically, under missing case I, where only X contains missing values, we show that it is sufficient to evaluate imputations within $\mathcal{X}(Y; U, T, n)$, as defined in Definition 5.1.3, in order to obtain the tight upper bound. In this case, only $m_1 + 1$ imputations need to be considered, where m_1 is the number of missing values in X.

A similar conclusion holds under missing case II. For this case, we prove that it is sufficient to evaluate $(m_1 + 1)(m_2 + 1)$ imputations, where m_1 and m_2 are the numbers of missing values in X and Y, respectively.

For missing case III, we show that it is sufficient to evaluate all paired imputations in $\mathcal{Y}(W, n)$, as defined in Definition 5.1.4. This reduces the computation to only $m_3 + 1$ paired imputations, where m_3 is the number of components in which both X and Y are unobserved.

The above results for missing case I, II and III are then combined together for showing that it is sufficient to evaluate all pairs of imputations in $(\mathcal{X}(Y;U,[n]\setminus W,n),\mathcal{X}(X;V,[n]\setminus W,n))\cap\mathcal{Y}(W,n)$ for obtaining the tight upper bounds for Spearman's footrule, as formally stated in Theorem 5.1.5. This result implies that only $(m_1+1)(m_2+1)(m_3+1)$ pairs of imputations need to be evaluated, where m_1 is the number of indices with missing values only in X, m_2 is the number of indices with missing values only in Y, and m_3 is the number of indices with missing values in both X and Y. This completes the sketch of proof.

□

Remark 5.1.6. Theorem 5.1.5 requires us to evaluate $(m_1+1)(m_2+1)(m_3+1)$ different pairs of imputations. Evaluating all these imputations directly would cost computational time $\mathcal{O}((m_1+1)(m_2+1)(m_3+1)n\log n)$, since computing Spearman's footrule once requires computational time $\mathcal{O}(n\log n)$. A more efficient algorithm for the tight upper bound with computational complexity $\mathcal{O}(n\log n+\min\{m_1,m_2\}^2+m_1m_2m_3)$ is provided in Algorithm 8.

If m_1, m_2 and m_3 are all proportional to n, then this means that finding the tight upper bound is $\mathcal{O}(n^3)$. While this may be computationally expensive for large n, without any of the work above, considering all possible imputations would take

$$\binom{n}{m_1+m_3}\binom{n}{m_2+m_3}(m_1+m_3)!(m_2+m_3)!$$

steps, which is computationally infeasible even for moderate n. For example, when $n=100$ and $m_1=m_2=m_3=5$, the number of possible imputations is approximately 3.95×10^{39}.

To further illustrate the results of Theorem 5.1.5, we include a simple example in Table 5.2. While one could naively consider all $(8\times7\times6)^2=112,896$ possible imputations to find the imputed pair of vectors that maximises Spearman's footrule, this theorem shows we only need to consider 18 imputations to do so.

5.2. Algorithms for computing the bounds

Theorem 5.1.1 provides conditions of the imputations to obtain the minimum possible Spearman's footrule. On the other hand, Theorem 5.1.5 demonstrates that only a small subset of imputations needs to be evaluated to achieve the maximum. Based on these results, this section discusses algorithms for computing the tight lower and upper bounds of Spearman's footrule in the presence of missing data.

5.2.1. Algorithms for the tight lower bound

Theorem 5.1.1 shows that in order to obtain the minimum possible Spearman's footrule, it is sufficient to construct imputations that satisfy conditions (i), (ii) and (iii) as defined in the theorem.

The existence of such imputations follows directly from Proposition C.3.13, which proves the existence of such imputations under missing case II, i.e. when $\Psi=\emptyset$. Furthermore, Proposition C.3.31 shows that the minimum possible Spearman's footrule under general missingness case is equal to the minimum possible Spearman's footrule after ignoring all pairs where values in both X and Y are unobserved. Hence, we only need to consider the lower bound when the data

Table 5.2.: (Left) Example ranked data under general missingness case. (Right) Imputed rank vectors (X_*, Y_*) that maximise Spearman's footrule. The imputed components are underlined. There are 112,896 possible imputations according to Remark 5.1.6, but by using Theorem 5.1.5 we only need to consider 18 of these imputations to find the maximum.

	a	b	c	d	e	f	g	h
$R(X, X)$	5	1	4	*	*	3	2	*
$R(Y, Y)$	*	*	1	2	3	4	5	*

	a	b	c	d	e	f	g	h
$R(X_*, X_*)$	6	2	5	$\underline{8}$	$\underline{7}$	4	3	$\underline{1}$
$R(Y_*, Y_*)$	$\underline{1}$	$\underline{7}$	2	3	4	5	6	$\underline{8}$

are missing under missing case II, where values in both X and Y can be missing, but for any pair $(X(i), Y(i))$, $i \in [n]$, at most one value is missing.

The proof of Proposition C.3.13 constructs such imputations. This is provided in Algorithm 3. The computational complexity of the algorithm is $\mathcal{O}(n \log n + n(m_1 + m_2))$, as discussed in more detail in Remark C.3.14.

5.2.2. Algorithms for the tight upper bound

Theorem 5.1.5 shows that to obtain the tight upper bound of Spearman's footrule, it is sufficient to evaluate all possible pairs of imputations in the set $(\mathcal{X}(Y; U, [n] \setminus W, n), \mathcal{X}(X; V, [n] \setminus W, n)) \cap \mathcal{Y}(W, n)$. This result allows us to evaluate $(m_1+1)(m_2+1)(m_3+1)$ different pairs of imputations, where m_1 is the number of components where only values in X are unobserved, m_2 is the number of components where only values in Y are unobserved, and m_3 is the number of components where values in both X and Y are unobserved. This is discussed in more details in Remark C.4.20.

A naive approach would compute Spearman's footrule directly for each of the $(m_1 + 1)(m_2 + 1)(m_3+1)$ imputations, requiring computational time $\mathcal{O}(m_1 m_2 m_3 n \log n)$, since evaluating Spearman's footrule once costs $\mathcal{O}(n \log n)$. However, it is not necessary to re-rank all components and recompute Spearman's footrule from scratch each time the ranks of the missing values in X and Y are adjusted. Exploiting this, Algorithm 8 achieves a reduced complexity of $\mathcal{O}(n \log n + \min\{m_1, m_2\}^2 + m_1 m_2 m_3)$.

5.3. Bounds of Kendall's τ

We have established tight bounds for Spearman's footrule in the presence of missing data under general missingness cases. A natural question is whether similar results can be obtained for other rank-based association measures, such as Spearman's rank correlation ρ and Kendall's τ.

It is possible to derive tight bounds of Kendall's τ under special missingness patterns, such as missing case III, as demonstrated by Papaioannou and Loukas (1984). However, extending these results to obtain efficient methods for computing tight bounds under the general missingness case is considerably more difficult. For Spearman's rank correlation ρ, no tight bounds appear to exist in the literature, either under specific or general missingness patterns.

While the tight bounds of Kendall's τ could be difficult to obtain under the general missingness cases, we now demonstrate that the bounds of Spearman's footrule can be applied for constructing non-trivial bounds of Kendall's τ.

To start, let $X, Y \in \mathcal{V}^n$, and define Kendall's τ coefficient as in Equation (2.10). Diaconis and Graham (1977, Theorem 2) showed that Kendall's τ and Spearman's footrule D differ by at most a constant factor.

Theorem 5.3.1 (Diaconis and Graham (1977))**.** Suppose $X, Y \in \mathcal{V}^n$. Then, the values of Spearman's footrule and Kendall's τ between X and Y are such that

$$\tau(X, Y) \leq D(X, Y) \leq 2\tau(X, Y).$$

Suppose $X, Y \in \mathcal{V}^n$ are partially observed. Let $D_{\min}(X, Y)$ and $D_{\max}(X, Y)$ denote the exact lower and upper bounds of Spearman's footrule between X and Y, as given by Theorems 5.1.1 and 5.1.5, respectively. Since $D_{\min}(X, Y) \leq D(X, Y) \leq D_{\max}(X, Y)$, Theorem 5.3.1 implies that

$$D_{\min}(X, Y)/2 \leq \tau(X, Y) \leq D_{\max}(X, Y). \tag{5.1}$$

In the next section, we empirically investigate these bounds on Kendall's τ through numerical simulations.

5.4. Simulations for the bounds

This section provides numerical simulations that evaluate the performance of our tight bounds of Spearman's footrule, and the bounds of Kendall's τ coefficient provided in Inequality (5.1). For comparison, we also investigate the behaviour of the methods when (i) the complete data are used and (ii) the missing values are ignored. When the missing values are ignored, the rank correlation coefficients are computed using only the pairs where both components from X and Y are observed. This section considers cases when the coefficient equals to 0, while additional simulation results for coefficients of -0.5 and 0.5, instead of 0, are provided in Section E.3.1 of the Appendix.

To make the results directly comparable, we first scale all the rank correlation coefficients between $[-1, 1]$. For scaling Spearman's footrule, we define

$$D_{\text{Scale}}(X, Y) = 1 - 3D(X, Y)/(n^2 - 1). \tag{5.2}$$

When n is odd, $D_{\text{Scale}}(X, Y) \in [-0.5, 1]$, but when n is even, $D_{\text{Scale}}(X, Y)$ is in the range $[-0.5\{1 + 3/(n^2 - 1)\}, 1]$ (Kendall, 1948). For scaling Spearman's ρ and Kendall's τ, we define

$$\begin{aligned} \rho_{\text{Scale}}(X, Y) &= 1 - \rho(X, Y)/\{n(n^2 - 1)\}, \\ \tau_{\text{Scale}}(X, Y) &= 1 - 4\tau(X, Y)/\{n(n - 1)\}, \end{aligned} \tag{5.3}$$

where $\tau(X, Y)$ and $\rho(X, Y)$ are given in Equation (2.9) and (2.10), respectively. Both $\rho_{\text{Scale}}(X, Y)$ and $\tau_{\text{Scale}}(X, Y)$ take values in $[-1, 1]$ (Kendall, 1948).

5.4.1. The bounds when data are missing completely at random

In the first simulation, the data are assumed to be missing completely at random (MCAR). First, X, Y are generated such that each pair $(X(i), Y(i))$ is an independent random sample from a

two-dimensional standard normal distribution with covariance coefficient 0; in other words. X and Y are independent. The sample size for X and Y is $n = 100$. Then, for each given proportion $s \in \{0, 0.05, \ldots, 0.30\}$ a subset of indices T in $\{1, 2, \ldots, n\}$ of size $\lfloor s \cdot n \rfloor$ is selected to be the set of components that are either partially observed or missing. For each $i \in T$, there is probability $1/3$ such that $X(i)$ is missing but $Y(i)$ is observed, probability $1/3$ such that $Y(i)$ is missing but $X(i)$ is observed, and probability $1/3$ such that both $X(i)$ and $Y(i)$ are missing. In any of these three cases, where at least one of $X(i)$ and $Y(i)$ is missing, we say that $(X(i), Y(i))$ is an incomplete pair.

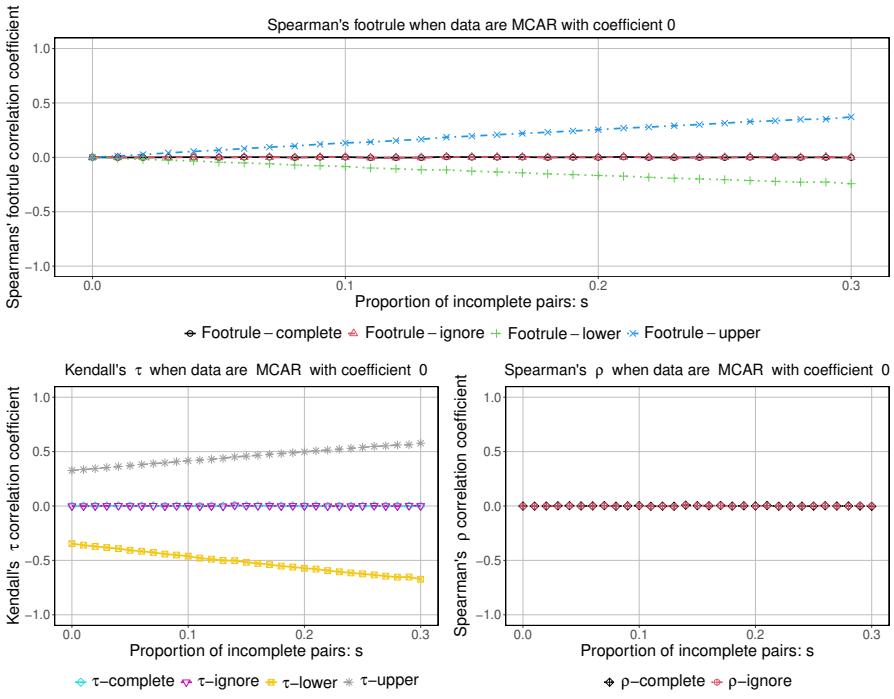


Figure 5.1.: Rank correlation coefficients when data are missing completely at random (MCAR). The subset of indices T containing all components with at least one missing value is selected randomly. For each $i \in T$, there is probability $1/3$ such that $X(i)$ is missing but $Y(i)$ is observed, probability $1/3$ such that $Y(i)$ is missing but $X(i)$ is observed, and probability $1/3$ such that both $X(i)$ and $Y(i)$ are missing. X, Y are generated such that $(X, Y) \overset{\text{i.i.d.}}{\sim} \mathrm{N}(0, I_2)$. The methods are described in Table 5.3. The sample size for X and Y is $n = 100$. Results represent the average of 1000 Monte Carlo simulations.

Subsequently, the bounds of Spearman's footrule for partially observed X and Y will be computed using Theorems 5.1.1 and 5.1.5, while the bounds for Kendall's τ for partially observed data will be computed using Inequality (5.1). We do not have bounds for Spearman's ρ for partially observed data, but we compute values for all three rank correlation coefficients in the cases where the complete data is used or missing values are ignored. Figure 5.1 presents the results, which are an average over 1000 Monte Carlo simulations.

Figure 5.1 shows that when the (independent) data are MCAR, ignoring the missing values will result in the methods producing unbiased estimates of the rank correlation coefficients, i.e. Footrule-ignore, τ-ignore and ρ-ignore are all approximately zero. Similarly, when the complete data are used, the three rank correlation coefficients will be unbiased, i.e. Footrule-complete,

Table 5.3.: Description of the rank correlation coefficients or their bounds, often in the presence of missing data, shown in Figure 5.1–5.2.

Coefficient/bound	Description
Footrule-upper	Upper bound for Spearman's footrule D, when data is partially observed.
Footrule-lower	Lower bound for Spearman's footrule D, when data is partially observed.
Footrule-ignore	Spearman's footrule D, when any missing data is ignored.
Footrule-complete	Spearman's footrule D, when data is fully observed.
τ-upper	Upper bound for Kendall's τ coefficient, when data is partially observed.
τ-lower	Lower bound for Kendall's τ coefficient, when data is partially observed.
τ-ignore	Kendall's τ coefficient, when any missing data is ignored.
τ-complete	Kendall's τ coefficient, when the data is fully observed.
ρ-ignore	Spearman's rank correlation ρ, when any missing data is ignored.
ρ-complete	Spearman's rank correlation ρ, when the data is fully observed.

τ-complete and ρ-complete are all approximately zero.

The range of bounds of Spearman's footrule, labelled Footrule-lower and Footrule-upper, increases gradually with the increasing proportion of incomplete pairs s. When $s = 0.3$, the lower and upper bounds of Spearman's footrule are around -0.2 and 0.4, respectively.

For Kendall's τ coefficient, when the proportion of incomplete pairs is 0, the lower and upper bounds of Kendall's τ, labelled by τ-lower and τ-upper, range from approximately -0.3 to 0.3. The bounds become wider as the proportion of incomplete pairs increases. When $s = 0.3$, the lower and upper bounds of Kendall's τ are approximately -0.6 and 0.5, respectively. It appears that the bounds of Spearman's footrule and Kendall's τ coefficient increase or decrease linearly, depending on the proportion s of missing components.

5.4.2. The bounds when data are missing not at random

The second simulation is designed in the same way as the first, except now the data are missing not at random (MNAR), rather than being MCAR. There are many ways to defined a MNAR mechanism; we propose one approach here, and consider a second approach in Section E.3.4 in the Appendix. Here, the index set $T \subset \{1, \ldots, n\}$ of the missing components is chosen now depends on the values of X and Y. Let $q = \sum_{i=1}^{n} \mathrm{I}\left(X(i)Y(i) > 0\right)$ be the number of pairs of components in X and Y such that their product $X(i)Y(i) > 0$. Each index $i \in [n]$ is selected to be in the set T of indices of missing components according to the following probability conditional on $C = |T|$, the size of the set T:

$$p\left(i \in T \,|\, C = \lfloor s \cdot n \rfloor\right) = \begin{cases} \min\left\{1, sn/q\right\}, & \text{if} X(i)Y(i) > 0, \\ \max\{0, (sn-q)/(n-q)\}, & \text{otherwise,} \end{cases} \tag{5.4}$$

for any given s. If $sn < q$, then $\min\left\{1, sn/q\right\} < 1$ and $\max\{0, (sn-q)/(n-q)\} = 0$, and so some components with $X(i)Y(i) > 0$ will be fully observed, and no components with $X(i)Y(i) \leq 0$ will be partially observed. However, if $sn \geq q$, then all components with $X(i)Y(i) > 0$ will be partially observed and, moreover, if $sn > q$ then some components with $X(i)Y(i) \leq 0$ will be partially observed. The probability is conditional in order to ensure we have subsets of indices of the desired cardinality $C = \lfloor s \cdot n \rfloor$.

Figure 5.2 shows that when the data are MNAR with the missingness mechanism specified in (5.4), then Footrule-ignore, τ-ignore and ρ-ignore produce biased estimates of their rank corre-

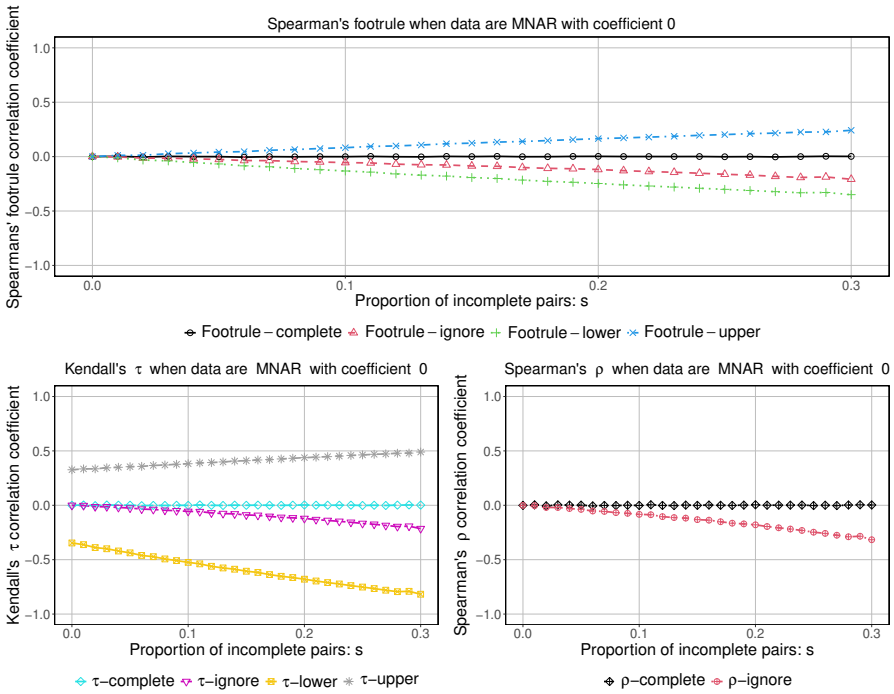


Figure 5.2.: Rank correlation coefficients when data are missing completely at random (MNAR). The subset of indices T containing all components with at least one missing value is selected according to Equation (5.4). For each $i \in T$, there is probability $1/3$ such that $X(i)$ is missing but $Y(i)$ is observed, probability $1/3$ such that $Y(i)$ is missing but $X(i)$ is observed, and probability $1/3$ such that both $X(i)$ and $Y(i)$ are missing. X, Y are generated such that $(X, Y) \overset{\text{i.i.d.}}{\sim} \mathrm{N}(0, I_2)$. The methods are described in Table 5.3. The sample size for X and Y is $n = 100$. Results represent the average of 1000 Monte Carlo simulations.

lation coefficients. Furthermore, as the proportion of incomplete pairs, denoted by s, increases, the bias of Footrule-ignore, τ-ignore and ρ-ignore also increases. On the other hand, Footrule-lower, Footrule-upper, τ-lower and τ-upper have similar performance compared to when data are MCAR in Figure 5.1.

5.5. Spearman's footrule independence test with missing data

We now apply the bounds of Spearman's footrule to independence testing with missing data, following a similar approach to that in Section 3.2 and Section 4.2. The null hypothesis is rejected only if all possible imputations yield significant test statistics.

As discussed in Section 3.2, this approach relies on rejection conditions to determine whether all test statistics are significant. Equivalently, the exact bounds of the p-values can be used to assess the significance of the result. In this section, we present the results directly, as the technical details of the proofs are similar to those in Section 3.2.

We begin by reviewing the distribution of Spearman's footrule under the null hypothesis. A well-known result (Diaconis and Graham, 1977, Theorem 1) shows that, under the null, the Spearman's footrule is approximately normally distributed. The exact mean and variance of this distribution are given by $\mu = (n+1)(n-1)/3$ and $\sigma^2 = (n+1)(2n^2+7)/45$, respectively (Ury and

Kleinecke, 1979). The accuracy of this approximation was examined by Salama and Quade (1990), who showed that for sample sizes $n \geq 40$, the normal approximation is reasonably accurate: the differences in skewness and kurtosis between the exact and approximate distributions are -0.075 and -0.074, respectively.

Next, we state the conditions for determining whether all possible values of Spearman's footrule are significant. Let $D_{\min}(X, Y)$ and $D_{\max}(X, Y)$ denote the minimum and maximum possible values of Spearman's footrule, which can be computed from Theorems 5.1.1 and 5.1.5, respectively. For a given significance level α, the null hypothesis is rejected if at least one of the following conditions is satisfied.

Condition 5.5.1.

$$\frac{D_{\min}(X, Y) - \mu}{\sigma} \geq \Phi^{-1}\left(1 - \frac{\alpha}{2}\right),$$

where $\mu = (n+1)(n-1)/3$ and $\sigma^2 = (n+1)(2n^2+7)/45$.

Condition 5.5.2.

$$\frac{D_{\max}(X, Y) - \mu}{\sigma} \leq \Phi^{-1}\left(\frac{\alpha}{2}\right),$$

where $\mu = (n+1)(n-1)/3$ and $\sigma^2 = (n+1)(2n^2+7)/45$.

More formally, we have the following result.

Theorem 5.5.3. The proposed method, which rejects the null hypothesis whenever Condition 5.5.1 or Condition 5.5.2 holds, has Type I error no larger than the Spearman's footrule independence test when all data are observed.

We now present the bounds of the p-values for Spearman's footrule in the presence of missing data. To begin, we define the p-value of Spearman's footrule. Suppose the sample size n is sufficiently large for the normal approximation to apply (e.g., $n \geq 40$). Let F_n denote the cumulative distribution function of a normal distribution with mean $(n+1)(n-1)/3$ and variance $(n+1)(2n^2+7)/45$. The p-value of Spearman's footrule is then given by

$$p(D(X, Y)) = 2 \min F_n(D(X, Y)), , 1 - F_n(D(X, Y)). \tag{5.5}$$

suppose $X, Y \in \mathcal{V}^n$ are only partially observed, and let $D_{\min}(X, Y)$ and $D_{\max}(X, Y)$ denote the minimum and maximum possible values of Spearman's footrule between X and Y, respectively. The bounds of the p-value are then determined by the following result.

Proposition 5.5.4. Suppose $X, Y \in \mathcal{V}^n$ are partially observed. Assume n is sufficiently large for normal approximation to be used. Let $D_{\min}(X, Y)$ and $D_{\max}(X, Y)$ be the minimum and maximum possible values of Spearman's footrule between X and Y, respectively. Denote $p_1 = p(D_{\min}(X, Y))$ and $p_2 = p(D_{\max}(X, Y))$, where $p(\cdot)$ is defined in (5.5). Furthermore, define $p_{\min}(X, Y) = \min\{p_1, p_2\}$ and define

$$p_{\max}(X, Y) = \begin{cases} \max\{p_1, p_2\}, & \text{if } (D_{\min} - \mu)(D_{\max} - \mu) \geq 0, \\ 1, & \text{otherwise,} \end{cases}$$

Table 5.4.: Description of the independence testing methods, often in the presence of missing data, shown in Figure 5.3–5.6.

Testing method	Description
Proposed	Using p-values computed from Footrule-upper and Footrule-lower.
Footrule-ignore	Using p-value of Spearman's footrule, when missing data is ignored.
Footrule-complete	Using p-value of Spearman's footrule, when the data is fully observed.
Footrule-mean	Using p-value of Spearman's footrule, with MI[1] for missing values.
Footrule-median	Using p-value of Spearman's footrule, with MEI[2] for missing values.
Footrule-hot deck	Using p-value of Spearman's footrule, using HDI[3] for missing values.
τ-ignore	Using p-value of Kendall's τ, when missing data is ignored.
τ-complete	Using p-value of Kendall's τ, when the data is fully observed.
τ-mean	Using p-value of Kendall's τ, with MI[1] for missing values.
τ-median	Using p-value of Kendall's τ, using MEI[2] for missing values.
τ-hot deck	Using p-value of Kendall's τ, using HDI[3] for missing values.
ρ-ignore	Using p-value of Spearman's ρ, when missing data is ignored.
ρ-complete	Using p-value of Spearman's ρ, when the data is fully observed.
ρ-mean	Using p-value of Spearman's ρ, with MI[1] for missing values.
ρ-median	Using p-value of Spearman's ρ, with MEI[2] for missing values.
ρ-hot deck	Using p-value of Spearman's ρ, with HDI[3] for missing values.
Alvo and Cabilio's ρ	Using p-value of estimate of Spearman's rank correlation ρ, from Alvo and Cabilio (1995).
Alvo and Cabilio's τ	Using p-value of estimate of Kendall's τ coefficient, from Alvo and Cabilio (1995).

[1] MI: Mean Imputation. [2] MEI: Median Imputation. [3] HDI: Hot Deck Imputation.

where $\mu = (n+1)(n-1)/3$. Then, the p-value of $D(X,Y)$ is bounded such that $p(D(X,Y)) \in [p_{\min}(X,Y), p_{\max}(X,Y)]$.

Finally, we show that rejecting the null hypothesis based on Conditions 5.5.1 and 5.5.2 is equivalent to rejecting it based on the maximum possible p-value.

Proposition 5.5.5. Suppose $X, Y \in \mathcal{V}^n$ are partially observed. Then for any given significance level $\alpha < 1$,

$$\text{Condition 5.5.1 or 5.5.2 hold} \iff p_{\max}(X,Y) \leq \alpha,$$

where $p_{\max}(X,Y)$ is defined in Proposition 5.5.4. Hence, the p-value of the proposed independence testing method is the maximum possible p-values $p_{\max}(X,Y)$.

An analytic power calculation is not provided in this chapter due to the complexity of the exact bounds for Spearman's footrule. Although these bounds can be computed using Algorithms 3 and 8, they do not seem to possess simple closed forms, making formal derivation difficult.

In next section, we perform numerical simulations to investigate the power of the proposed method under different sample sizes, proportion of missing data, and the degrees of departure of the null hypothesis. The simulations also highlight the potential peril of not taking missing data into account when the data are missing not at random (MNAR).

5.6. Simulations for the independence test

This section performs simulations to investigate the Type I error and the statistical power of the proposed method for independence testing when data are partially observed.

For comparison with the proposed method, we also consider the Spearman's footrule, Kendall's τ and Spearman's ρ, in the case where the missing data are ignored. For this case, the correlation coefficients are computed only using pairs where both components from X and Y are observed. We also consider these three methods when imputation methods are used; we consider mean imputation, median imputation and hot deck imputation (Schafer and Graham, 2002).

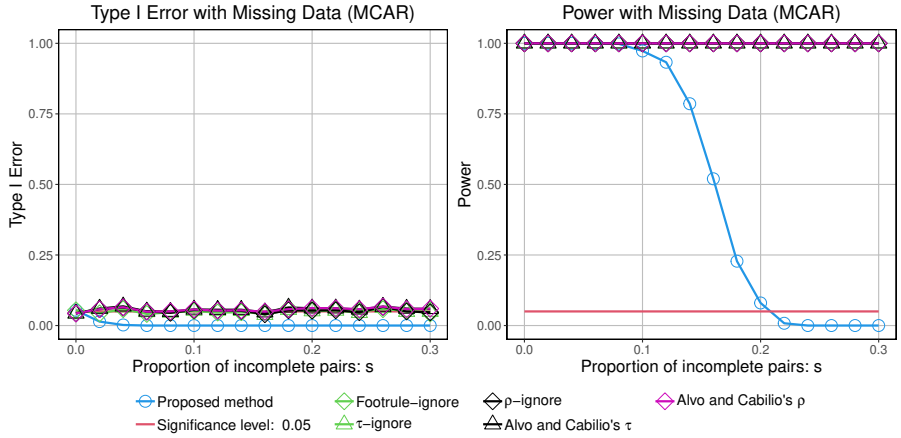


Figure 5.3.: Statistical Type I error and power of the proposed method, as the proportion of missing data s increases, where $s \in \{0.00, 0.02, \ldots, 0.30\}$. Other methods that ignore missing data, and Alvo and Cabilio's ρ and Alvo and Cabilio's τ methods are also considered. These methods are described in Table 5.4. The data is missing completely at random (MCAR). The subset of indices T containing all components with at least one missing value is selected randomly. For each $i \in T$, there is probability $1/3$ such that $X(i)$ is missing but $Y(i)$ is observed, probability $1/3$ such that $Y(i)$ is missing but $X(i)$ is observed, and probability $1/3$ such that both $X(i)$ and $Y(i)$ are missing. (Left) Type I error: $(X, Y) \overset{\text{i.i.d.}}{\sim} \mathrm{N}(0, I_2)$; (Right) Power: $(X, Y) \overset{\text{i.i.d.}}{\sim} \mathrm{N}(0, \Sigma)$, where $\Sigma = \begin{pmatrix} 1 & \gamma \\ \gamma & 1 \end{pmatrix}$, with covariance coefficient $\gamma = 0.5$. For both figures, a significance level $\alpha = 0.05$ is used and the sample size $n = 200$. The results in the figures are average of 1000 trials.

Two independence testing methods that take missing data into account are also considered. These methods (Alvo and Cabilio, 1995) are defined as the conditional expectations of Spearman's ρ and Kendall's τ, respectively, given the observed ranks, under the assumption that the ranks of the missing values are uniformly distributed. In the following, we will refer to the two methods as Alvo and Cabilio's ρ and Alvo and Cabilio's τ methods, respectively.

5.6.1. As the proportion of missing data varies

In the first experiment, we explore how the methods perform as the proportion s of missing data increases. More precisely, we consider s to be the proportion of incomplete pairs, where we recall that $(X(i), Y(i))$ is an incomplete pair if at least one of $X(i)$ or $Y(i)$ is missing.

We start by considering the case when the data are missing completely at random (MCAR), as the proportion of incomplete pairs increases. For evaluating the Type I error of all methods, the data are generated following the same distribution and missingness mechanism for Figure 5.1, as described in Section 5.4.1, with sample size $n = 500$. The significance level $\alpha = 0.05$ is used for all test methods. For evaluating the power, the data are generated in the same way, except that

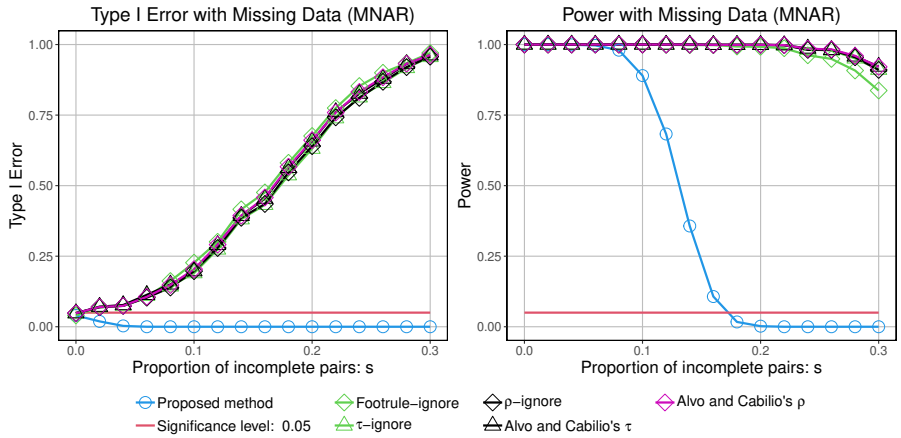


Figure 5.4.: Statistical Type I error and power of the proposed method as the proportion of missing data s increases, where $s \in \{0.00, 0.02, \ldots, 0.30\}$. Other methods that ignore missing data, and Alvo and Cabilio's ρ and Alvo and Cabilio's τ methods are also considered. These methods are described in Table 5.4. The data is missing completely at random (MNAR). The subset of indices T containing all components with at least one missing value is selected according to Equation (5.4). For each $i \in T$, there is probability 1/3 such that $X(i)$ is missing but $Y(i)$ is observed, probability 1/3 such that $Y(i)$ is missing but $X(i)$ is observed, and probability 1/3 such that both $X(i)$ and $Y(i)$ are missing. (Left) Type I error: $(X, Y) \overset{\text{i.i.d.}}{\sim} \mathrm{N}(0, I_2)$; (Right) Power: $(X, Y) \overset{\text{i.i.d.}}{\sim} \mathrm{N}(0, \Sigma)$, where $\Sigma = \begin{pmatrix} 1 & \gamma \\ \gamma & 1 \end{pmatrix}$, with covariance coefficient $\gamma = 0.5$. For both figures, a significance level $\alpha = 0.05$ is used and the sample size $n = 200$. The results in the figures are average of 1000 trials.

the covariance matrix is now $\Sigma = \begin{pmatrix} 1 & \gamma \\ \gamma & 1 \end{pmatrix}$, with covariance coefficient $\gamma = 0.5$. The appendices contains results for the same experiment, but with different values of α, different values of the covariance coefficient γ, and different sample sizes n.

Figure 5.3 shows that all methods appear to control the Type I error when the data are MCAR. All methods appear to have good statistical power, but the power of the proposed method starts to drop when the proportion of incomplete pairs is larger than 10%, and when 22% of pairs are incomplete, the power is 0.

We next consider the case when the data are missing not at random (MNAR). The parameters of this experiment are the same as for the MCAR case, but now the missingness mechanism is as described in Section 5.4.2.

Figure 5.4 shows that when data are MNAR, only the proposed method controls the Type I error, while the other methods fail to control the Type I error when at least 4% of the data is incomplete. When the proportion of incomplete pairs is 10%, the methods that ignore the missing data and Alvo and Cabilio's ρ and τ methods have Type I error close to 25%. Meanwhile, all methods have good power with small proportion of incomplete pairs. The power of the proposed method starts to drop when the proportion s at least 8%, and drops to 0 when the proportion s is larger than 18%.

We also investigate how imputation methods perform in the experiment for Figure 5.4, when data are MNAR. We consider using mean imputation, median imputation and hot deck imputation Schafer and Graham (2002) for Spearman's footrule, Kendall's τ and Spearman's ρ. The results are shown in Figure 5.5, which demonstrates that these imputation methods also fail to

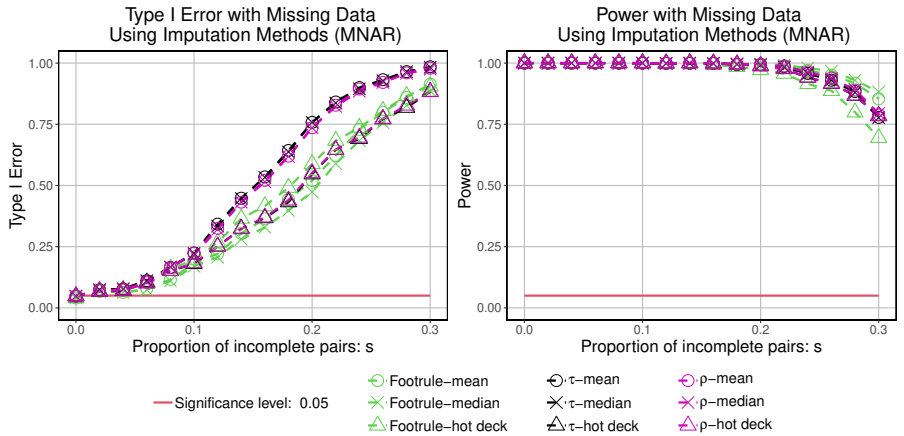


Figure 5.5.: Statistical Type I error and power of imputation methods as the proportion of missing data s increases, where $s \in \{0.00, 0.02, \ldots, 0.30\}$. The imputation methods impute missing data using either mean, median or randomly selected (hot deck) values of observed data. These methods are described in Table 5.4. The data is missing completely at random (MNAR). The subset of indices T containing all components with at least one missing value is selected according to Equation (5.4). For each $i \in T$, there is probability $1/3$ such that $X(i)$ is missing but $Y(i)$ is observed, probability $1/3$ such that $Y(i)$ is missing but $X(i)$ is observed, and probability $1/3$ such that both $X(i)$ and $Y(i)$ are missing. (Left) Type I error: $(X, Y) \overset{\text{i.i.d.}}{\sim} \mathrm{N}(0, I_2)$; (Right) Power: $(X, Y) \overset{\text{i.i.d.}}{\sim} \mathrm{N}(0, \Sigma)$, where $\Sigma = \begin{pmatrix} 1 & \gamma \\ \gamma & 1 \end{pmatrix}$, with covariance coefficient $\gamma = 0.5$. For both figures, a significance level $\alpha = 0.05$ is used and the sample size $n = 200$. The results in the figures are average of 1000 trials.

control the Type I error rate when at least 4% of the data is incomplete.

5.6.2. As the sample size varies

We again consider the above case when the data are missing not at random, but now as the sample size n increases, with fixed proportion of missing pairs $s = 0.1$, and fixed correlation coefficient $\gamma = 0.5$ for the alternative hypothesis. Figure 5.6 shows that the proposed method still controls the Type I error as the sample size increases. Furthermore, while the proposed method is not very powerful for smaller sample sizes, as the sample size n increases, its power goes to 1. On the other hand, the other approaches have good power, but fail to control the Type I error, and their Type I error increases as n increases.

5.6.3. As the correlation coefficient varies

We now consider the case as the correlation coefficient γ increases, with fixed proportion of missing pairs $s = 0.1$, sample size $n = 200$ and significance level $\alpha = 0.05$. The results are shown in Figure 5.7 and 5.8 for MCAR and MNAR, respectively.

When the data are MCAR, Figure 5.7 shows that the power of the proposed method increases with the correlation coefficient, and equal to 1 when the coefficient is around 0.5. All other methods have power close to 1 when the coefficient is around 0.3.

The results shown in Figure 5.8 when data are MNAR are similar to that of Figure 5.7. However, note that when the correlation equals to 0, the Type I error of the proposed method is

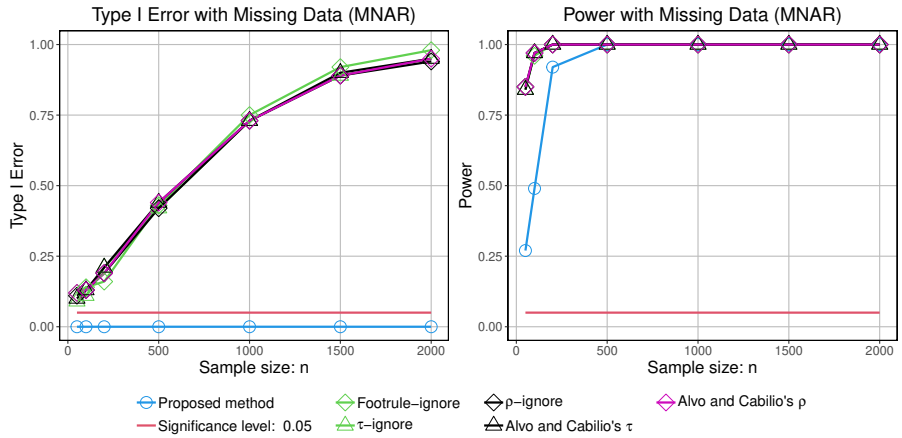


Figure 5.6.: Statistical Type I error and power of the proposed method as the sample size n increases, where $n \in \{40, 70, 100, 200, 300, 500\}$. Other methods that ignore missing data, and Alvo and Cabilio's ρ and Alvo and Cabilio's τ methods are also considered. These methods are described in Table 5.4. The data is missing completely at random (MNAR). The subset of indices T containing all components with at least one missing value is selected according to Equation (5.4). For each $i \in T$, there is probability $1/3$ such that $X(i)$ is missing but $Y(i)$ is observed, probability $1/3$ such that $Y(i)$ is missing but $X(i)$ is observed, and probability $1/3$ such that both $X(i)$ and $Y(i)$ are missing. (Left) Type I error: $(X, Y) \overset{\text{i.i.d.}}{\sim} \mathrm{N}(0, I_2)$; (Right) Power: $(X, Y) \overset{\text{i.i.d.}}{\sim} \mathrm{N}(0, \Sigma)$, where $\Sigma = \begin{pmatrix} 1 & \gamma \\ \gamma & 1 \end{pmatrix}$, with covariance coefficient $\gamma = 0.5$. For both figures, a significance level $\alpha = 0.05$ is used and the proportion of missing pairs $s = 0.1$. The results in the figures are average of 1000 trials.

controlled, while all other methods fail to do so.

This section evaluates the Type I error and statistical power of the proposed independence testing method in the presence of missing data. Our proposed method is based on the tight bounds of Spearman's footrule, and take into account all possible imputations. The method is also compared with Spearman's footrule, Kendall's τ, and Spearman's ρ, when data are imputed or ignored. Two independence testing methods (Alvo and Cabilio, 1995) that take missing data into account are also considered.

The results in this section show that our proposed method has good statistical power, typically when the proportion of pairs containing missing data is below 15%, while all the other methods fail to control the Type I error.

More simulations results with different coefficients, sample sizes, proportion of missing data, and significance levels are provided in Section E.3 in the Appendix. In particular, Section E.3.4 considers another missing not at random (MNAR) missingness mechanism that is different to the one specified in Equation (5.4). Experiments when the data are Gamma-distributed, rather than Normal-distributed are included in Section E.3.5. The results for these simulations are similar to the results shown in this section.

5.7. Application to precipitation data

To further demonstrate the bounds of Spearman's footrule and the proposed independence testing method, we consider the yearly precipitation data in the United Kingdom and the United States

Table 5.5.: Precipitation per year in the United States, the United Kingdom, and Germany from 1901 to 2024 in mm. The 15 largest values from 1901 to 1960 for all three countries are marked with a special character #.

Year	US	UK	Germany	Year	US	UK	Germany
1901	697.06	1027.95	664.64	1963	672.30	1055.03	623.64
1902	754.66	956.38	653.80	1964	745.47	993.67	597.15
1903	722.52	#1404.49	703.70	1965	751.32	1198.87	869.67
1904	684.67	1029.85	611.83	1966	690.38	1255.44	878.63
1905	#786.07	1012.00	709.44	1967	744.53	1271.55	752.27
1906	#791.10	1153.20	698.59	1968	740.80	1138.39	757.61
1907	750.46	1144.31	631.46	1969	744.68	1073.23	650.36
1908	735.85	1066.47	614.17	1970	730.64	1186.76	796.17
1909	758.11	1138.46	718.46	1971	751.72	999.48	551.69
1910	637.14	1188.59	#775.08	1972	771.79	1028.40	612.24
1911	722.56	1085.72	548.90	1973	836.40	1011.52	637.48
1912	#764.20	#1261.04	#771.28	1974	755.73	1222.67	780.57
1913	731.36	1057.05	665.10	1975	805.18	985.27	611.20
1914	714.45	1180.02	736.35	1976	668.61	1050.19	536.24
1915	#792.61	1148.94	720.32	1977	750.74	1162.17	743.71
1916	727.02	#1280.92	#783.90	1978	746.53	1125.13	703.89
1917	643.91	1119.24	656.78	1979	815.13	1238.50	739.47
1918	725.75	1216.47	688.96	1980	720.66	1240.28	760.58
1919	#771.54	1125.07	694.11	1981	756.37	1240.57	875.56
1920	760.60	1209.69	662.09	1982	833.13	1284.87	630.49
1921	712.57	971.51	534.05	1983	846.85	1144.41	704.61
1922	740.98	1108.57	#801.00	1984	767.93	1180.51	726.48
1923	#772.19	#1311.30	719.50	1985	753.82	1194.23	669.03
1924	681.35	#1258.66	692.48	1986	782.84	1259.90	749.24
1925	678.07	1142.69	713.51	1987	737.04	1143.88	802.17
1926	#762.60	1172.29	#773.13	1988	677.00	1251.56	780.52
1927	#770.19	#1295.67	#793.67	1989	745.73	1118.84	620.32
1928	748.13	#1326.36	677.35	1990	807.93	1319.55	718.96
1929	758.69	1138.89	587.21	1991	801.98	1092.06	595.51
1930	664.41	#1260.31	#793.52	1992	787.23	1294.52	714.82
1931	699.44	1183.50	#789.56	1993	807.85	1237.25	780.32
1932	758.54	1173.61	647.93	1994	771.66	1311.97	777.91
1933	687.96	919.98	595.52	1995	787.07	1153.78	757.45
1934	662.58	1141.88	610.31	1996	812.05	1025.44	601.25
1935	738.38	#1241.53	728.08	1997	781.31	1126.54	647.92
1936	697.39	1131.09	731.36	1998	823.67	1370.87	799.33
1937	750.72	1104.08	714.15	1999	729.70	1324.77	744.33
1938	742.77	#1264.57	662.99	2000	720.22	1399.04	733.33
1939	687.55	1148.73	#830.41	2001	735.28	1105.69	789.60
1940	755.95	1144.86	754.25	2002	729.92	1402.60	889.55
1941	#789.74	1022.15	#756.31	2003	762.21	961.72	555.63
1942	#771.15	1127.81	621.46	2004	806.34	1218.10	720.01
1943	686.88	1139.79	574.43	2005	759.90	1137.37	702.60
1944	758.45	1172.23	#767.20	2006	747.04	1186.31	681.06
1945	#811.68	1099.04	723.55	2007	722.11	1207.77	874.14
1946	760.41	1214.18	700.97	2008	764.87	1328.89	704.13
1947	740.81	1063.01	613.33	2009	780.43	1309.18	730.71
1948	#761.18	#1279.64	674.82	2010	766.11	1068.64	774.65
1949	752.07	1129.20	613.19	2011	769.46	1194.00	668.45
1950	740.70	#1283.03	#786.34	2012	726.76	1376.21	702.39
1951	#764.73	#1278.19	701.44	2013	802.29	1160.69	721.10
1952	683.24	1080.42	735.74	2014	780.06	1399.18	699.96
1953	714.42	1004.90	566.39	2015	860.09	1317.82	649.47
1954	667.70	#1380.14	#789.89	2016	790.36	1196.33	688.48
1955	703.43	967.69	699.89	2017	818.70	1208.01	808.22
1956	654.85	1094.22	738.34	2018	860.00	1144.82	552.78
1957	#806.31	1163.27	706.28	2019	872.32	1296.07	661.65
1958	747.72	1211.87	#799.91	2020	780.19	1359.22	648.50
1959	#763.34	1060.89	521.46	2021	791.96	1171.36	753.38
1960	733.13	#1285.78	#798.16	2022	748.90	1176.40	595.02
1961	779.90	1160.06	765.67	2023	777.06	1449.59	849.78
1962	716.56	1085.69	648.92	2024	817.97	1306.22	841.89

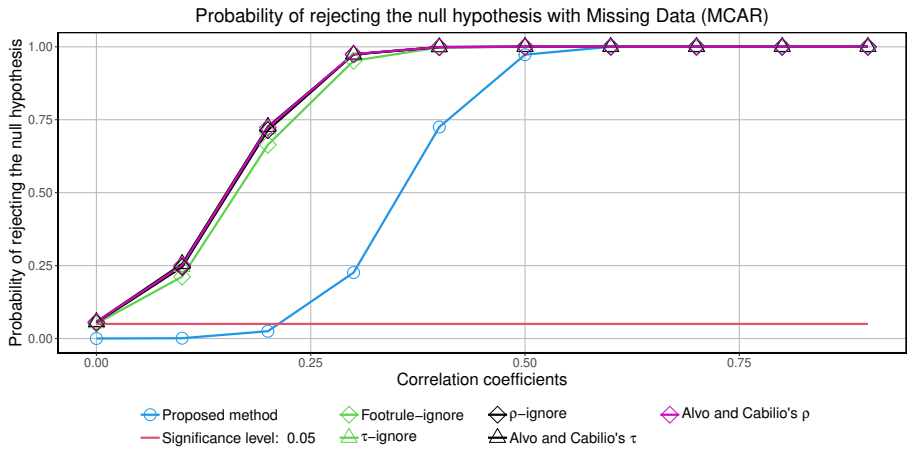


Figure 5.7.: Statistical Type I error and power of the proposed method, methods that ignore missing data or use complete data, and Alvo and Cabilio's ρ and Alvo and Cabilio's τ methods, as the correlation coefficient γ increases, where $\gamma \in \{0, 0.1, \ldots, 0.9\}$. These methods are described in Table 5.4. The data is missing completely at random (MCAR). The subset of indices T containing all components with at least one missing value is selected randomly. For each $i \in T$, there is probability 1/3 such that $X(i)$ is missing but $Y(i)$ is observed, probability 1/3 such that $Y(i)$ is missing but $X(i)$ is observed, and probability 1/3 such that both $X(i)$ and $Y(i)$ are missing. $(X, Y) \overset{\text{i.i.d.}}{\sim} \mathrm{N}(0, \Sigma)$, where $\Sigma = \begin{pmatrix} 1 & \gamma \\ \gamma & 1 \end{pmatrix}$, with γ denoting covariance coefficient. For both figures, a significance level $\alpha = 0.05$ is used and the proportion of missing pairs $s = 0.1$. The results in the figures are average of 1000 trials.

and Germany from 1901 to 2024. The data are sourced from Harris et al. (2020), providing monthly precipitation measurements on high-resolution (0.5×0.5 degree) grids across the globe.

These values are summed across the months of the year, and then averaged across each country, to give the annual precipitation values in Table 5.5. More specifically, each measurement gives the precipitation for year $i \in \mathcal{I}$, month $j \in \mathcal{J}$ and grid index $t \in \mathcal{T}$. For each year i and grid index t, we sum across $j \in \mathcal{J}$ to get the annual precipitation for that grid point. We then average these grid point measurements across the subset of $\mathcal{T}$ for a particular country, to get that country's average precipitation for year i.

Note that the original dataset retains only one decimal place, resulting in several tied observations. In this example, we do not account for ties. We add small random noise at the second decimal place to break all ties.

Based on the complete dataset, precipitation in the United Kingdom and Germany exhibits a significant dependency with significance level $\alpha = 0.05$. The estimated Spearman's ρ, Kendall's τ, and the scaled Spearman's footrule are 0.5037, 0.3472, and 0.3171, respectively, with corresponding p-values of 3.94×10^{-9}, 1.07×10^{-8}, and 2.36×10^{-8}. On the other hand, precipitation in the United States does not yield significantly associated with that in the United Kingdom according to the test statistics: the estimated Spearman's ρ, Kendall's τ, and the scaled Spearman's footrule are 0.1527, 0.1044, and 0.1044, respectively, with corresponding p-values of 9.04×10^{-2}, 8.56×10^{-2}, and 6.59×10^{-2}, respectively. Similarly, the precipitation in the United States is not significantly associated with that in Germany according to the testing results: the estimated Spearman's ρ, Kendall's τ, and the scaled Spearman's footrule are -0.0298, -0.0181, and -0.0166,

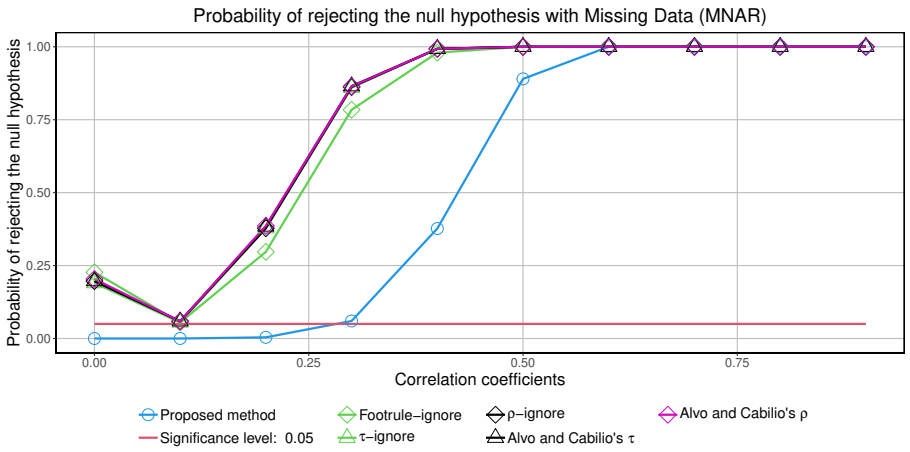


Figure 5.8.: Statistical Type I error and power of the proposed method, methods that ignore missing data or use complete data, and Alvo and Cabilio's ρ and Alvo and Cabilio's τ methods, as the correlation coefficient γ increases, where $\gamma \in \{0, 0.1, \ldots, 0.9\}$. These methods are described in Table 5.4. The data is missing not at random (MNAR). The subset of indices T containing all components with at least one missing value is selected according to Equation (5.4). For each $i \in T$, there is probability 1/3 such that $X(i)$ is missing but $Y(i)$ is observed, probability 1/3 such that $Y(i)$ is missing but $X(i)$ is observed, and probability 1/3 such that both $X(i)$ and $Y(i)$ are missing. $(X, Y) \overset{\text{i.i.d.}}{\sim} \mathrm{N}(0, \Sigma)$, where $\Sigma = \begin{pmatrix} 1 & \gamma \\ \gamma & 1 \end{pmatrix}$, with γ denoting covariance coefficient. For both figures, a significance level $\alpha = 0.05$ is used and the proportion of missing pairs $s = 0.1$. The results in the figures are average of 1000 trials.

respectively, with corresponding p-values of 7.42×10^{-1}, 7.66×10^{-1}, and 7.71×10^{-1}, respectively. All the results are summarised in Table 5.6.

Suppose now that the data are missing informatively. In other words, the data are missing not at random (MNAR). We assume that the largest 15 values between 1901 to 1950 are missing for all three countries.

For analysing the associations of the precipitation data between all three countries with missing data, we apply all the test methods listed in Table 5.4. The estimated coefficients and testing results for each method are summarised in Table 5.6.

As previously discussed, when the data are complete, the association of the precipitations between the United Kingdom and Germany is statistically significant with significance level $\alpha = 0.05$ according to Footrule-complete, τ-complete, and ρ-complete. When the data are missing, all methods reject the null hypothesis. The proposed method reports tight bounds of the Spearman's footrule and p-values with missing data. The minimum and maximum possible Spearman's footrule are 0.1188 and 0.4876, respectively, and the corresponding p-values range from 9.01×10^{-18} to 3.6×10^{-2}. Since the maximum possible p-value is smaller than the given significance level, all possible imputations lead to significant results. Hence, the proposed method rejects the null hypothesis.

The association of the precipitation data between the United Kingdom and the Unites States are not statistically significant with significance level $\alpha = 0.05$ according to the complete data. However, when the data are missing as described, all methods, except the proposed method, rejects the null hypothesis. The proposed method reports the Spearman's footrule to lie within

Table 5.6.: Estimated coefficients and testing results for analysing the statistical association for the precipitation data in the United Kingdom (U.K.), the United States (U.S.) and Germany.

Method	U.K. and Germany		U.K. and U.S.		U.S. and Germany	
	est. coef.	p-value	est. coef.	p-value	est. coef.	p-value
ρ-complete	0.5037	3.94×10^{-9}	0.1527	9.04×10^{-2}	-0.0298	7.42×10^{-1}
τ-complete	0.3472	1.07×10^{-8}	0.1044	8.56×10^{-2}	-0.0181	7.66×10^{-1}
F-complete	0.3171	2.36×10^{-8}	0.1044	6.59×10^{-2}	-0.0166	7.71×10^{-1}
Prop. (min)	0.1188	9.01×10^{-18}	-0.004	9.09×10^{-12}	-0.1160	2.15×10^{-6}
Prop. (max)	0.4876	3.6×10^{-2}	0.3873	1.00×10^{0}	0.2690	1.00×10^{0}
ρ-ignore	0.5372	2.75×10^{-9}	0.3392	3.33×10^{-4}	0.1029	2.87×10^{-1}
τ-ignore	0.3772	6.13×10^{-9}	0.2355	2.84×10^{-4}	0.0720	2.67×10^{-1}
F-ignore	0.3535	5.32×10^{-9}	0.2535	2.84×10^{-5}	0.0798	1.87×10^{-1}
ρ-mean	0.5311	2.21×10^{-10}	0.2720	2.24×10^{-3}	0.0807	3.73×10^{-1}
τ-mean	0.3636	3.25×10^{-9}	0.1936	1.63×10^{-3}	0.0566	3.57×10^{-1}
F-mean	0.3495	7.57×10^{-10}	0.2047	3.12×10^{-4}	0.0849	1.35×10^{-1}
ρ-median	0.5437	6.78×10^{-11}	0.2918	1.01×10^{-3}	0.0676	4.56×10^{-1}
τ-median	0.3773	8.75×10^{-10}	0.2088	6.93×10^{-4}	0.0453	4.62×10^{-1}
F-median	0.3750	4.01×10^{-11}	0.2273	6.25×10^{-5}	0.0761	1.80×10^{-1}
ρ-hot deck	0.4969	4.34×10^{-9}	0.2251	1.20×10^{-2}	0.0208	8.18×10^{-1}
τ-hot deck	0.3417	1.93×10^{-8}	0.1548	1.09×10^{-2}	0.0145	8.12×10^{-1}
F-hot deck	0.3153	2.82×10^{-8}	0.1629	4.11×10^{-3}	0.0201	7.23×10^{-1}
A & C's ρ	0.5569	6.56×10^{-10}	0.2871	1.45×10^{-3}	0.0712	4.30×10^{-1}
A & C's τ	0.3966	3.51×10^{-10}	0.2057	1.14×10^{-3}	0.0487	4.41×10^{-1}

Prop.: proposed method. Proposed method provides both the minimum and maximum possible Spearman's footrule and p-value. F-ignore: Footrule-ignore. F-mean: Footrule-mean. F-median: Footrule-median. F-hot deck: Footrule-hot deck. All methods are described in Table 5.4.

the interval [-0.004, 0.3873], with the corresponding p-value ranging from 9.09×10^{-12} to 1. Since the maximum possible p-value is larger than the significant level α, the proposed method does not reject the null hypothesis. The wide range of the p-values in this example indicates high uncertainty of the testing results from missing data, and emphasises the importance of accounting for missingness.

Finally, the association of the precipitation data between the United States and the Germany are not statistically significant with significance level $\alpha = 0.05$ according to the complete data. When the data are missing as described, all methods do not reject the null hypothesis. The proposed method reports the range of the Spearman's footrule to lie within [-0.1160, 0.2690], and the p-value in $[2.15 \times 10^{-6}, 1]$.

5.8. Conclusion

In this chapter, we derive the tight lower and upper bounds of Spearman's footrule with missing data, when the data are assumed to be distinct. Algorithms with computational complexities $\mathcal{O}(n^2)$ and $\mathcal{O}(n^3)$ are provided for computing the tight lower and upper bounds, respectively.

Using our bounds, we derive non-trivial lower and upper bounds of Kendall's τ. These bounds are all evaluated using numerical simulations. The bounds of Spearman's footrule also allows us to construct a novel two-sample independence testing method for data with missing values, following the similar approach to our proposed location, and scale testing with missing data methods.

Simulation results show that our proposed independence testing method has good statistical power when the proportion of missing data is smaller than 15%. Importantly, our method controls the Type I error, while all other methods by ignoring or imputing the missing data fail to do so, especially when the data are missing not at random (MNAR). The results highlight the peril

of not taking missing data into account correctly, and demonstrate that our proposed method is particularly useful when the data are missing.

6. Extensions of the location test with missing data

In Chapter 3, 4 and 5, we have considered the tight lower and upper bounds of the Wilcoxon-Mann-Whitney test statistic, the Ansari-Bradley test statistic, and Spearman's footrule in the presence of missing data for the location testing, scale testing, and independence testing, respectively. Our methods make no assumptions about the missingness mechanisms and missingness patterns, and only require the data to be distinct, i.e. no ties are allowed.

However, for discrete-valued data, it is possible that two or more observations can be equal. In such cases, one can introduce small random perturbations to each tied observation to break the ties while preserving the relative ordering of the data. This approach could lead to consistent testing results when the proportion of tied observations is small (Amerise and Tarsitano, 2015). However, when the proportion is large, this approach may depend on the perturbations introduced, and could lead to distinct testing results.

A more common method for dealing with ties and allowing deterministic testing results is known as the mid-rank method (Lehmann and D'Abrera, 1975). This method assigns the same rank to the tied observations, while keeping the sum of all ranks unchanged regardless of the number of ties. We describe this method in more details in Section 6.1.1.

Section 6.1 describes the work in Zeng et al. (2024a, Section 5), which considers the Wilcoxon-Mann-Whitney test in the presence of missing data for the cases where data can be tied in a potentially closed support and the mid-rank method is applied for dealing with ties. The tight lower and upper bounds of the Wilcoxon-Mann-Whitney test statistics under such cases are derived in Section 6.1.1. The bounds of the p-values are provided in Section 6.1.2. For evaluating this method when the data can be tied, and are defined in a closed support, we perform numerical simulations in Section 6.1.3.

Up to this point, all the proposed methods are based on the bounds of the test statistics of realised, real-valued observations. In other words, we determine the tight lower and upper bounds for the realised value of the test statistic. However, it is also possible to take a different perspective and establish bounds for the test statistic *prior* to the data are realised. To be more specific, we treat the test statistic in the presence of missing data as a random variable. Our goal is then to construct other random variables that, by design, are guaranteed to be less than or equal to, or greater than or equal to, the test statistic for every possible realisation of the data.

Once such bounding random variables are constructed, we study their distributions under the null hypothesis, and propose a novel method for rejecting the null hypothesis using the observed data alone. Importantly, we prove that this approach controls the Type I error regardless of the missingness mechanisms.

This approach is explored for the Wilcoxon-Mann-Whitney test in Section 6.2. We first con-

struct the bounding random variables for the Wilcoxon-Mann-Whitney test statistic in Section 6.2.1, and then study the distributions of the bounds under the null hypothesis. Using these results, we derive the p-values of the proposed method in 6.2.2. This new approached is evaluated in Section 6.2.3 and is shown to have better power than the location testing method proposed in Chapter 3.

It is possible to obtain similar results for the Ansari-Bradley test statistic and Spearman's footrule. However, obtaining these results might be technical challenging, and could be considered as future work.

6.1. Location test in the presence of missing data with tied observations

In this section, the location testing with missing data method proposed in Chapter 3 is extended to the cases where the data can be tied and are defined in a potentially closed support. We start by bounding the Wilcoxon-Mann-Whitney test statistic under such cases.

6.1.1. Bounds of the Wilcoxon-Mann-Whitney statistic with ties

As we mentioned previously, for discrete-valued data, it is possible that two or more observations can be equal. Hence, in this section, we shall consider X, Y and $Z = X \cup Y$ to be *multisets*, rather than sets.

We follow the practice of mid-rank method for dealing with ties: tied values would all be assigned the same rank, e.g. if $z_1, z_2 \in Z$ are such that $z_1 = z_2$, then $\text{rank}_Z(z_1) = \text{rank}_Z(z_2)$; if there were exactly $k-1$ smaller values than z_1 and z_2, and these two values were distinct, then z_1 and z_2 would occupy ranks k and $k+1$, but if $z_1 = z_2$, then these two values are each assigned the arithmetic mean rank $\{k + (k+1)\}/2 = k + 1/2$. The same argument holds for cases when more than two values are equal. To be more specific, we have the following definition.

Definition 6.1.1. Suppose Z is a multiset consisting of real-valued, potentially tied observations. Then for any $x \in Z$, the rank of x in Z is defined as

$$R(x, Z) = \sum_{z \in Z} \mathrm{I}\left(z < x\right) + \left(\sum_{z \in Z} \mathrm{I}\left(z = x\right) + 1 \right) /2$$

According to the definition, when there is an odd number of equal values then their midranks will be an integer, and for an even number of equal values then their midranks will be a half-integer. This is consistent with the definition of rank in Definition 2.2.1 when all the elements are distinct. Suppose $Z = \{z_1, \ldots, z_N\}$ consists of N observations. A nice property of using mid-rank method is that the identity

$$\sum_{i=1}^{N} R(z_i, Z) = N(N+1)/2$$

always holds, regardless of the number of ties (Lehmann and D'Abrera, 1975).

When considering ties, we also consider the case when the data are defined in a support Ω to limit the space from which the data are observed.

Remark 6.1.2. As an example of why it may be necessary to consider the support Ω, suppose X is known to consist of 5 values in the set $\Omega = \{1, 2, 3, 4, 5\}$, and $X' \subset X$ is observed as $\{2, 3, 5\}$. The two missing values in $X \setminus X'$ could both be 1, to potentially minimise any sum of ranks of X within some larger multiset. If, however, the support is not taken into account and the two missing values in $X \setminus X'$ take values smaller than 1, the sum of ranks of X could be smaller than the minimum possible sum of ranks of X when the support is taken into account. This could result the lose of power of the proposed method, as we will show in Section 6.1.3.

However, the support Ω need not be discrete or bounded, for example, in the space $\Omega = [0, \infty)$ there could be multiple missing values which could potentially have had the value 0. We could also consider $\Omega = \mathbb{R}$, in which case the missing values could take any real value.

Following the above discussions, one can extend Theorem 3.1.3 to the case where ties are allowed, and the data are defined in a potentially closed support Ω.

Proposition 6.1.3. Suppose that $X = \{x_1, \ldots, x_n\}$ and $Y = \{y_1, \ldots, y_m\}$ are samples of observations, which need not necessarily be distinct, from a space $\Omega \subset \mathbb{R}$. Suppose that $X' \subset X$ and $Y' \subset Y$ are sub-multisets with sizes n' and m', respectively, which are observed. Denote $Z' = X' \cup Y'$. Then the Wilcoxon-Mann-Whitney statistic $W(X, Y)$ is bounded as follows:

$$\min_{Z \setminus Z'} W(X, Y) = W(X', Y') + \left(\left|\partial Y'_a\right| (n - n') + \left|\partial X'_b\right| (m - m')\right)/2,$$
$$\max_{Z \setminus Z'} W(X, Y) = W(X', Y') + (mn - m'n') - \left(\left|\partial X'_a\right| (m - m') + \left|\partial Y'_b\right| (n - n')\right)/2.$$

where

$$\left|\partial X'_a\right| = \begin{cases} \sum\limits_{x \in X'} \mathrm{I}(x = \min \Omega) & \text{if } \min \Omega \text{ exists,} \\ 0 & \text{otherwise,} \end{cases}$$
$$\left|\partial X'_b\right| = \begin{cases} \sum\limits_{x \in X'} \mathrm{I}(x = \max \Omega) & \text{if } \max \Omega \text{ exists,} \\ 0 & \text{otherwise,} \end{cases}$$
$$\left|\partial Y'_a\right| = \begin{cases} \sum\limits_{y \in Y'} \mathrm{I}(y = \min \Omega) & \text{if } \min \Omega \text{ exists,} \\ 0 & \text{otherwise,} \end{cases}$$
$$\left|\partial Y'_b\right| = \begin{cases} \sum\limits_{y \in Y'} \mathrm{I}(y = \max \Omega) & \text{if } \max \Omega \text{ exists,} \\ 0 & \text{otherwise.} \end{cases}$$

Proof. The proof is provided in Section D.1 in the Appendix. □

Remark 6.1.4. The result above covers cases where the space Ω is discrete and bounded, e.g. $\{0, 1, 2\}$, or continuous and bounded, e.g. $[0, 1]$, or bounded only on one side, e.g. $\mathbb{Z}^+ = \{1, 2, \ldots\}$, or $[0, \infty)$, or not bounded, e.g. $\mathbb{R}$, in which case the result reduces to Theorem 3.1.3.

Remark 6.1.5. The presence of ties and potentially closed support affect our strategy of proof for the bounds compared to that shown in Theorem 3.1.3. As discussed in Remark 3.1.2, the

bounds of the Wilcoxon-Mann-Whitney test statistic without ties are achieved when all missing values are strictly smaller or larger than all observed data. However, when ties are permitted and the support is closed, missing values cannot theoretically exceed the support boundaries. Consequently, they may be forced to take values equal to the observed extremes rather than strictly surpassing them.

Compared to the bounds shown in Theorem 3.1.3 when the data are assumed to be distinct, the lower and upper bounds in Proposition 6.1.3 include the following additional terms

$$\left(\left|\partial Y_a'\right|(n-n') + \left|\partial X_b'\right|(m-m')\right)/2$$

and

$$\left(\left|\partial X_a'\right|(m-m') + \left|\partial Y_b'\right|(n-n')\right)/2,$$

respectively. Note that these terms

- can be non-zero only when the support has finite lower or upper boundaries. Consequently, if the support is unbounded, the bounds of the test statistic remain identical to that when ties are not allowed.
- cannot be negative. Hence, they can only tighten the bounds, i.e., shrink the interval of possible test statistics.
- increases with the number of observed values lying on the boundaries. Hence, the bounds explicitly depends on the number of observed values lie on the boundaries of the support.

Intuitively, these are because when the observed values are lie on the support boundaries, the missing data are constraint from taking values strictly less or larger than them. This restriction limits the possible ranks the missing values can take, thereby constraining the range of the test statistic.

6.1.2. Bounds of the p-value with ties

Proposition 6.1.3 provides the bounds of the statistic $W(X,Y)$ when the data can be tied and are defined in a potentially closed support. For proposing a testing method based on the bounds, we still need to construct the bounds for the p-value of $W(X,Y)$.

One might consider constructing the bounds of the p-values following the similar approach in Proposition 3.2.4 when the data are not tied. However, this approach might not always lead to the bounds of the p-value, since now the variance can take on different values depending on the number of ties, as explained below.

When there are ties, the Wilcoxon-Mann-Whitney test statistic for multisets $X = \{X_1, \cdots, X_n\}$ and $Y = \{Y_1, \cdots, Y_m\}$ still asymptotically follows a normal distribution with mean $\mu = nm/2$, but the variance σ^2 takes the ties into account; if $X \cup Y$ contains e distinct values, with multiplicities $d_1, \ldots, d_e$, where d_i is the ith smallest distinct value (so $d_1 + \cdots + d_e = n + m$), then Lehmann and D'Abrera (1975) gives the variance as

$$\sigma^2(X,Y) = nm(n+m+1)/12 - nm\{12(n+m)(n+m-1)\}^{-1}\sum_{i=1}^{e}(d_i^3 - d_i). \tag{6.1}$$

Comparing this with the variance $nm(n+m+1)/12$ when there are no ties, the second term on the right hand side in the above equation essentially corrects for ties and reduces the variance.

Since now the variance can take on different values depending on the number of ties, as shown in Equation (6.1), the bounds of the p-value constructed according to Proposition 3.2.4 when data are not tied might not always control the p-value, as illustrated by the following result.

Proposition 6.1.6. Suppose that $X = \{x_1, \ldots, x_n\}$ and $Y = \{y_1, \ldots, y_m\}$ are samples of real-valued observations, which need not necessarily be distinct. Suppose that $X' \subset X$ and $Y' \subset Y$ are sub-multisets with sizes n' and m', respectively, which are known. Define $Z = X \cup Y$ and $Z' = X' \cup Y'$ and suppose that $Z \setminus Z'$ is unknown. Suppose X_1, Y_1 and X_2, Y_2 are such that $X' \subset X_i$ and $Y' \subset Y_i$ for $i = 1, 2$ and

$$W(X_1, Y_1) = \min_{Z \setminus Z'} W(X, Y), \qquad W(X_2, Y_2) = \max_{Z \setminus Z'} W(X, Y).$$

Then, among candidates for X, Y, it is possible X_3, Y_3 exist such that $X' \subset X_3$ and $Y' \subset Y_3$ and

$$W(X_1, Y_1) \leq W(X_3, Y_3) \leq W(X_2, Y_2),$$

but moreover, after defining $\mu = nm/2$, the standardised statistics are such that

$$(W(X_3, Y_3) - \mu)/\sigma(X_3, Y_3) \notin [(W(X_1, Y_1) - \mu)/\sigma(X_1, Y_1), (W(X_2, Y_2) - \mu)/\sigma(X_2, Y_2)],$$

and consequently $p_3 \notin [p_{\min}(X, Y), p_{\max}(X, Y)]$, where $p_{\min}(X, Y)$, and $p_{\max}(X, Y)$ are defined according to Proposition 3.2.4 and p_3 is the p-values for $W(X_3, Y_3)$.

Proof. The proof is provided in Section D.2 in the Appendix. □

Remark 6.1.7. Proposition 6.1.6 shows that when the data can be tied, the bounds of p-values derived in Proposition 3.2.4 are no longer valid. This is because the variance of the test statistic is no loner a constant of the sample sizes, but a value that depends on the number of ties. Hence, while $W(X_3, Y_3)$ can be bounded by $[W(X_1, Y_1), W(X_2, Y_2)]$, we can still have

$$(W(X_3, Y_3) - \mu)/\sigma(X_3, Y_3) \notin [(W(X_1, Y_1) - \mu)/\sigma(X_1, Y_1), (W(X_2, Y_2) - \mu)/\sigma(X_2, Y_2)],$$

depends on the variances $\sigma(X_1, Y_1)$, $\sigma(X_2, Y_2)$ and $\sigma(X_3, Y_3)$.

In order to deal with the issue raised by Proposition 6.1.6, we start by computing the minimum and maximum possible values for the variance $\sigma^2(X, Y)$ in the presence of ties.

Proposition 6.1.8. Suppose that $X = \{x_1, \ldots, x_n\}$ and $Y = \{y_1, \ldots, y_m\}$ are samples of observations, which need not necessarily to be distinct, from a space $\Omega \subset \mathbb{R}$. Suppose that $X' \subset X$ and $Y' \subset Y$ are sub-multisets with sizes n' and m', respectively, which are known, and that $X' \cup Y'$ contains e' distinct values with multiplicities $d'_1, \ldots, d'_{e'}$ and $d'_1 \leq \cdots \leq d'_{e'}$. Define

$$\sigma^2_{\max}(X, Y) = nm(n+m+1)/12 - nm\{12(n+m)(n+m-1)\}^{-1} \sum_{i=1}^{e'} \{(d'_i)^3 - d'_i\},$$
$$\sigma^2_{\min}(X, Y) = \sigma^2_{\max}(X, Y) - nm\{12(n+m)(n+m-1)\}^{-1}\{d^3_{\max} - d_{\max} - (d'_{e'})^3 + d'_{e'}\},$$

where $d_{\max} = d'_{e'} + n + m - n' - m'$, then $\sigma^2(X,Y)$, the variance of Wilcoxon-Mann-Whitney statistic $W(X,Y)$ as defined in Equation 6.1, is bounded by

$$\sigma^2_{\min}(X,Y) \leq \sigma^2(X,Y) \leq \sigma^2_{\max}(X,Y).$$

Proof. The proof is provided in Section D.3 in the Appendix. □

Extending results from the previous section, we can apply the proposed method to cases where there may be tied observations and the sample space may have closed support, by using the following result.

Corollary 6.1.9. Suppose that $X = \{x_1, \ldots, x_n\}$ and $Y = \{y_1, \ldots, y_m\}$ are samples of observations, which need not necessarily be distinct, from a space $\Omega \subset \mathbb{R}$. Suppose that $X' \subset X$ and $Y' \subset Y$ are sub-multisets with sizes n' and m', respectively, which are known, and compute $|\partial X'_a|$, $|\partial X'_b|$, $|\partial Y'_a|$ and $|\partial Y'_b|$ from Proposition 6.1.3 and $\sigma_{\min}(X,Y)$, $\sigma^2_{\max}(X,Y)$ from Proposition 6.1.8, and define $\mu = nm/2$,

$$\begin{aligned}
&W_{\min}(X,Y) = W(X',Y') + \left|\partial Y'_a\right|(n-n')/2 + \left|\partial X'_b\right|(m-m')/2,\\
&W_{\max}(X,Y) = W(X',Y') + (mn - m'n') - \left|\partial X'_a\right|(m-m') + \left|\partial Y'_b\right|(n-n'),\\
&p_1 = 2\Phi(-\left|W_{\min}(X,Y) - \mu\right|/\sigma_{\min}(X,Y)),\ p_2 = 2\Phi(-\left|W_{\max}(X,Y) - \mu\right|/\sigma_{\max}(X,Y)),\\
&p_3 = 2\Phi(-\left|W_{\max}(X,Y) - \mu\right|/\sigma_{\min}(X,Y)),\ p_4 = 2\Phi(-\left|W_{\min}(X,Y) - \mu\right|/\sigma_{\max}(X,Y)).
\end{aligned}$$

Then, the the p-value of $W(X,Y)$ is bounded as follows

$$\begin{cases}
p_1 \leq p \leq p_2, & \text{if } W_{\min}(X,Y) - \mu < 0 \text{ and } W_{\max}(X,Y) - \mu < 0,\\
\min\{p_1, p_3\} \leq p \leq 1, & \text{if } W_{\min}(X,Y) - \mu < 0 \text{ and } W_{\max}(X,Y) - \mu \geq 0,\\
p_3 \leq p \leq p_4, & \text{if } W_{\min}(X,Y) - \mu \geq 0 \text{ and } W_{\max}(X,Y) - \mu \geq 0.
\end{cases} \tag{6.2}$$

Proof. The proof is provided in Section D.4 in the Appendix. □

According to Corollary 6.1.9, given a significance threshold $\alpha \in (0,1)$ and defining $p = \max\{p_2, p_4\}$, if $p < \alpha$ and the two terms $(W_{\min}(X,Y) - \mu)$ and $(W_{\min}(X,Y) - \mu)$ have the same sign, then the data will yield a significant result, regardless of the missing values in $(X \cup Y) \setminus (X' \cup Y')$.

6.1.3. Simulations for the proposed location test with ties and closed support

We ran numerical simulations to demonstrate the advantage of accounting for ties and closed support when using the proposed method instead of using the original version from Chapter 3.

The first experiment is similar to the one shown in Figure 3.2, but uses Poisson distributions, rather than normal distributions. The data in first sample X are independent observations of a Pois(1) random variable with data missing completely at random (MCAR). The data in the second sample Y are independent observations of a Pois(Δ) random variable with data missing not at random (MNAR); only values greater than 0 can be missing, according to the mechanism in Equation (3.11). For the Type I error plot $\Delta = 1$ and for the plot of statistical power, $\Delta = 3$.

The results in Figure 6.1 show that the modified method has an improvement in statistical power over the original proposed method which doesn't account for ties or closed support.

Figure 6.2 is similar to Figure 6.1, but shows decreased power when the second sample consists of observations sampled from a Pois(2) distribution. In Figure 6.3, on the other hand, the second sample of observations are sampled from a Pois(4) distribution. In this case, the proposed method, which takes ties and closed support into account has good power when up to 20% of the data is missing.

In Figure 6.4, the data for both the Type I error and power plots is the same as for Figure 6.3, but we show the proposed method taking ties into account, but *not* taking the closed support into account. In this case, the performance is almost exactly the same as for the proposed method, which does not take ties or closed support into account. This shows the importance of taking the closed support into account, as illustrated in Figure 6.1, 6.2 and 6.3, since Poisson-distributed observations have a lower bound of 0.

In Figure 6.5, the data is the same as for Figure 6.1, but now with the sample sizes increased to $n = m = 1000$. We see an increase in power in the proposed method(s), as well as an increase in the Type I error in the methods which use imputation methods or case deletion.

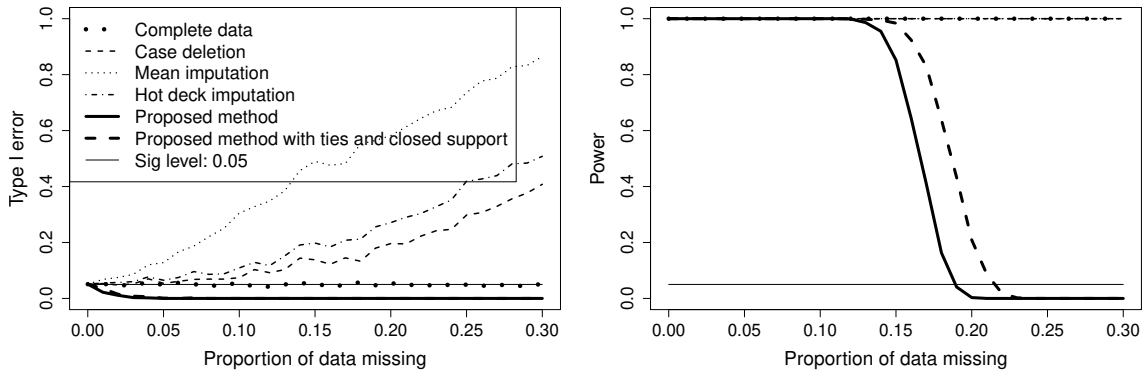


Figure 6.1.: The Type I error and statistical power of the proposed method and the standard Wilcoxon-Mann-Whitney test after the missing data is either known or has been imputed or ignored as the proportion of missing data increases. The data for the first sample X is missing completely at random (MCAR) but the data for the second sample Y is missing not at random (MNAR), according to the mechanism in Equation (3.11). (Left) Type I error: Pois(1) vs Pois(1); (Right) Power: Pois(1) vs Pois(3). For both figures, a significance threshold of $\alpha = 0.05$ has been used and the total sample sizes are $n = 100$, $m = 100$, and 1000 trials were used.

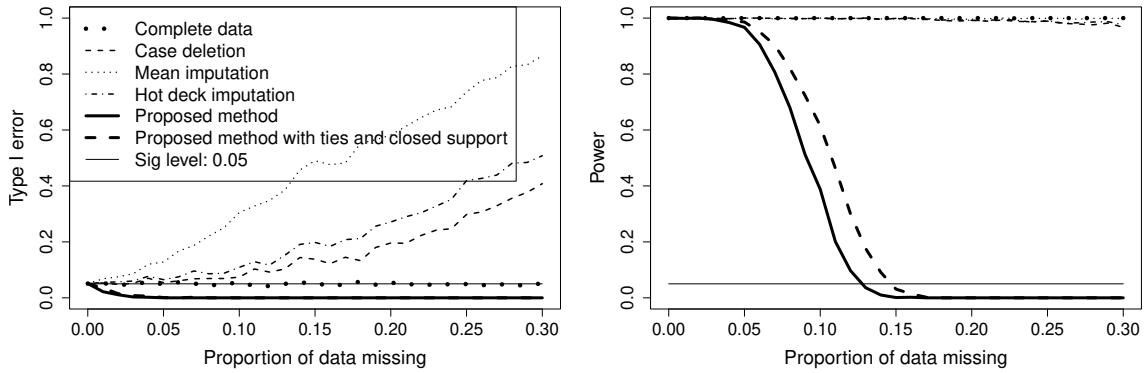


Figure 6.2.: The Type I error and statistical power of the proposed method and the standard Wilcoxon-Mann-Whitney test after the missing data is either known or has been imputed or ignored as the proportion of missing data increases. The data for the first sample X is missing completely at random (MCAR) but the data for the second sample Y is missing not at random (MNAR), according to the mechanism in Equation (3.11). (Left) Type I error : Pois(1) vs Pois(1); (Right) Power: Pois(1) vs Pois(2). For both figures, a significance threshold of $\alpha = 0.05$ has been used and the total sample sizes are $n = 100$, $m = 100$, and 1000 trials were used.

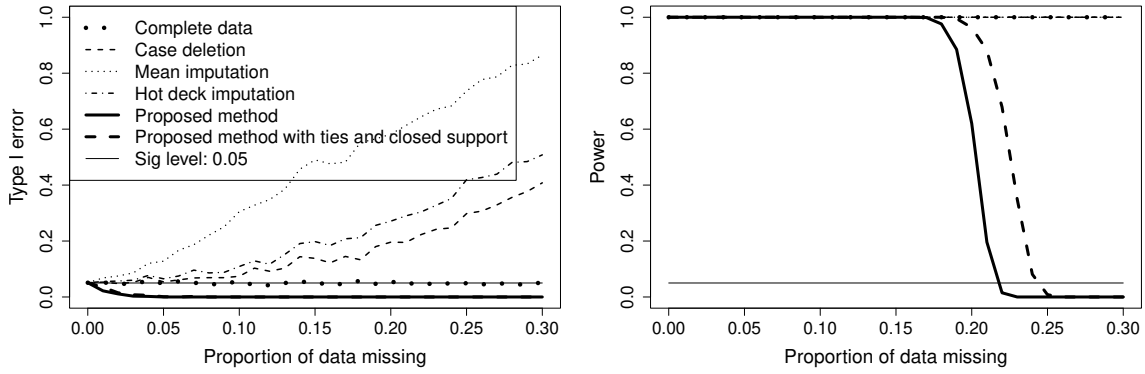


Figure 6.3.: The Type I error and statistical power of the proposed method and the standard Wilcoxon-Mann-Whitney test after the missing data is either known or has been imputed or ignored as the proportion of missing data increases. The data for the first sample X is missing completely at random (MCAR) but the data for the second sample Y is missing not at random (MNAR), according to the mechanism in Equation (3.11). (Left) Type I error : Pois(1) vs Pois(1); (Right) Power: Pois(1) vs Pois(4). For both figures, a significance threshold of $\alpha = 0.05$ has been used and the total sample sizes are $n = 100$, $m = 100$, and 5000 trials were used.

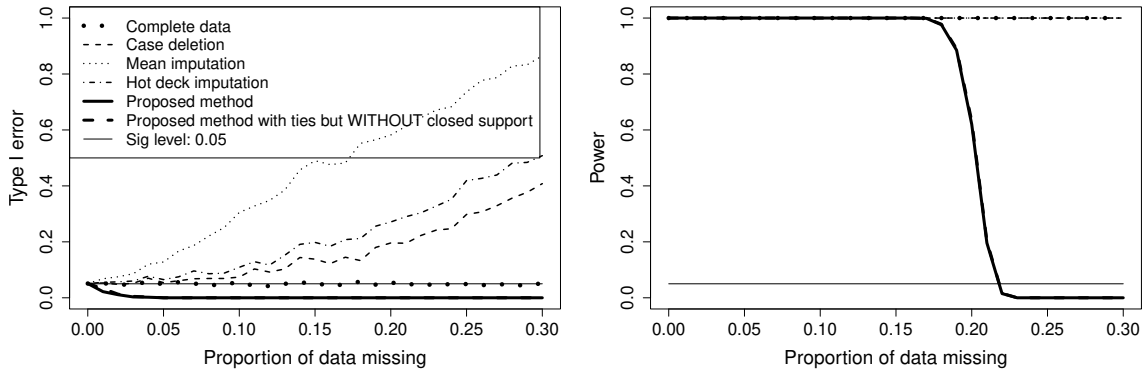


Figure 6.4.: The Type I error and statistical power of the proposed method and the standard Wilcoxon-Mann-Whitney test after the missing data is either known or has been imputed or ignored as the proportion of missing data increases. The data for the first sample X is missing completely at random (MCAR) but the data for the second sample Y is missing not at random (MNAR), according to the mechanism in Equation (3.11). We also plot the modified method, but **without** taking closed support into account. (Left) Type I error : Pois(1) vs Pois(1); (Right) Power: Pois(1) vs Pois(4). For both figures, a significance threshold of $\alpha = 0.05$ has been used and the total sample sizes are $n = 100$, $m = 100$, and 1000 trials were used.

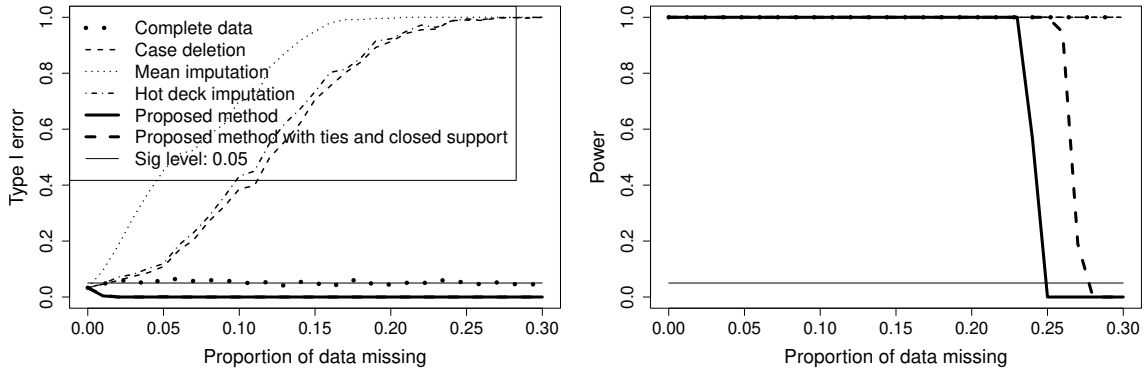


Figure 6.5.: The Type I error and statistical power of the proposed method and the standard Wilcoxon-Mann-Whitney test after the missing data is either known or has been imputed or ignored as the proportion of missing data increases. The data for the first sample X is missing completely at random (MCAR) but the data for the second sample Y is missing not at random (MNAR), with only observations greater than 0 are possibly missing. (Left) Type I error : Pois(1) vs Pois(1); (Right) Power: Pois(1) vs Pois(4). For both figures, a significance threshold of $\alpha = 0.05$ has been used and the total sample sizes are $n = 1000$, $m = 1000$, and 1000 trials were used.

6.2. A variant of the location test method with missing data

As discussed earlier, all the proposed methods considered so far are based on the bounds of the test statistics of realised, real-valued observations. In this section, we take a different perspective and construct bounds for the Wilcoxon-Mann-Whitney test statistic *prior* to the data are realised. In other words, the test statistic in the presence of missing data is treated as a random variable.

Our goal in this section is to construct other random variables that, by design, are guaranteed to be less than or equal to, or greater than or equal to, the Wilcoxon-Mann-Whitney test statistic with missing data for every possible realisation of the data. Once such bounding random variables are constructed, we study their distributions under the null hypothesis, and propose a novel method for rejecting the null hypothesis using the observed data alone. Importantly, we prove that this approach controls the Type I error regardless of the missingness mechanisms.

For our purposes in this section, X and Y will be used to denote sets of *random variables*, rather than observations.

We start this section by providing bounds of the test statistic in the presence of missing data, when the statistic is taken as a random variable.

6.2.1. Bounding the test statistic as a random variable

Suppose $X = \{X_1, \ldots, X_n\}$ and $Y = \{Y_1, \ldots, Y_m\}$ are independent and identically distributed random variables following continuous cumulative distribution functions F, and G, respectively. The Wilcoxon-Mann-Whitney test statistic is defined as

$$W(X, Y) = \sum_{i=1}^{n} R(X, X \cup Y) - n(n+1)/2.$$

Once variables in X and Y are all observed, one can obtain the testing result by first computing the p-value of the test statistic as the probability of more "extremes" events to happen under the null hypothesis, i.e.,

$$p(W(X,Y)) = 2\min\left\{\Phi\left(\frac{W(X,Y)-\mu}{\sigma}\right), 1-\Phi\left(\frac{W(X,Y)-\mu}{\sigma}\right)\right\},$$

where $\mu = nm/2$ and $\sigma^2 = nm(m+m+1)/12$. Then, the null hypothesis is rejected if the p-value is less than or equal to a pre-specified significance level α. Notably, this procedure relies on the asymptotic result that, under the null hypothesis, $W(X, Y)$ is approximately normally distributed with mean μ and variance σ^2, as discussed in Section 2.2.1. Without this result, this procedure is not guaranteed to control the Type I error below α.

Suppose only n' number of random variables in X, and m' number of random variables in Y will be observed, and denote them as X', Y', respectively. One could still compute the Wilcoxon-Mann-Whitney test statistic using only the observed variables $W(X', Y')$. However, the distribution of $W(X', Y')$ under the null hypothesis now depends on the missingness mechanism. In other words, how the observed n' and m' random variables are selected. See Section 2.3.1 for more discussions about the missingness mechanisms.

For example, suppose the n', m' random variables in X' and Y' are selected randomly, i.e.

the data are missing completely at random (MCAR), then one can show that $W(X', Y')$ still follows a normal distribution, with mean $n'm'/2$ and variance $n'm'(n' + m' + 1)/12$ under the null hypothesis. However, if larger values in X are more likely to be observed than smaller values in X, then this is missing not at random (MNAR), and the distribution of $W(X', Y')$ could *not* follow the same normal distribution, even when the null hypothesis is true.

When the distribution of $W(X', Y')$ under the null hypothesis is unknown, it is challenging to directly constructing a valid p-value. In Chapter 3, we propose one method for obtaining valid p-values based on the bounds of Wilcoxon-Mann-Whitney test statistic after X and Y are realised.

In this section, we propose to construct a valid p-value by first bounding the Wilcoxon-Mann-Whitney test statistic $W(X', Y')$ as a random variable, as shown by the following result.

Theorem 6.2.1. Suppose $X = \{X_1, \ldots, X_n\}$ and $Y = \{Y_1, \ldots, Y_m\}$ are continuous random variables. Suppose $X' = \{X'_1, \ldots, X'_{n'}\}$ is a non-empty subset of X and $Y' = \{Y'_1, \ldots, Y'_{m'}\}$ is a non-empty subset of Y. Let us order $X_1, \ldots, X_n$ from small to large as $X_{(1)}, \ldots, X_{(n)}$ and denote

$$X_* = X \setminus \{X_{(n'+1)}, \ldots, X_{(n)}\},\ X^* = X \setminus \{X_{(1)}, \ldots, X_{(n-n')}\}.$$

Similarly, let us order $Y_1, \ldots, Y_m$ from small to large as $Y_{(1)}, \ldots, Y_{(m)}$ and denote

$$Y_* = Y \setminus \{Y_{(m'+1)}, \ldots, Y_{(m)}\},\ Y^* = Y \setminus \{Y_{(1)}, \ldots, Y_{(m-m')}\}.$$

Then, the Wilcoxon-Mann-Whitney test statistic $W(X', Y')$ is bounded as follows

$$W(X_*, Y^*) \leq W(X', Y') \leq W(X^*, Y_*).$$

Proof. The proof is provided in Section D.5 in the Appendix. □

While the above theorem considers X and Y as random variables, the result holds when they are observed samples. We take them explicitly as random variables in the theorem to make the notation in this section consistent as we will need to investigate the distributions of the bounds $W(X_*, Y^*)$ and $W(X^*, Y_*)$ in Proposition 6.2.2.

Note the above theorem does not impose any constraints on the subsets X' and Y', which can be considered as variables that will be observed. Hence, the result holds regardless of the missingness mechanism. After proving this result, we proceed to study the distribution of the test statistics $W(X_*, Y^*)$ and $W(X^*, Y_*)$ under the null hypothesis.

Deriving these distributions in closed form appears to be technically challenging. However, it is possible to approximate their distributions through numerical simulations, based on the following result.

Proposition 6.2.2. Suppose $(P_1, \ldots, P_n, Q_1, \ldots, Q_m)$ is a uniform random permutation of $\{1, \ldots, N\}$ with $N = n + m$. Denote $P = \{P_1, \ldots, P_n\}$, $Q = \{Q_1, \ldots, Q_m\}$. Order $P_1, \ldots, P_n$ from small to large as $P_{(1)}, \ldots, P_{(n)}$, and $Q_1, \ldots, Q_m$ from small to large as $Q_{(1)}, \ldots, Q_{(m)}$. Denote

$$P_* = P \setminus \{P_{(n'+1)}, \ldots, P_{(n)}\},\ P^* = P \setminus \{P_{(1)}, \ldots, P_{(n-n')}\},$$
$$\text{and } Q_* = Q \setminus \{Q_{(m'+1)}, \ldots, Q_{(m)}\},\ Q^* = Q \setminus \{Q_{(1)}, \ldots, Q_{(m-m')}\}.$$

Then, suppose $X = \{X_1, \ldots, X_n\}$ and $Y = \{Y_1, \ldots, Y_m\}$ are independent continuous random variables following the same distribution. We have

$$W(X_*, Y^*) \stackrel{d}{=} W(P_*, Q^*), \text{ and } W(X^*, Y_*) \stackrel{d}{=} W(P^*, Q_*),$$

where $W(X_*, Y^*)$ and $W(X^*, Y_*)$ are defined in Theorem 6.2.1.

Proof. To start, denote $Z = X \cup Y$. Since $X = \{X_1, \ldots, X_n\}$ and $Y = \{Y_1, \ldots, Y_m\}$ are independent continuous random variables, the probability of ties is zero. Further, as X and Y follow the same distribution, the ranks of random variables

$$(R(X_1, Z), \ldots, R(X_n, Z), R(Y_1, Z), \ldots, R(Y_m, Z))$$

is uniformly distributed on all permutations of $\{1, \ldots, N\}$ with $N = n + m$. In other words, we have

$$(R(X_1, Z), \ldots, R(X_n, Z), R(Y_1, Z), \ldots, R(Y_m, Z)) \stackrel{d}{=} (P_1, \ldots, P_n, Q_1, \ldots, Q_m).$$

After ordering, we further have,

$$(R(X_{(1)}, Z), \ldots, R(X_{(n)}, Z), R(Y_{(1)}, Z), \ldots, R(Y_{(m)}, Z)) \stackrel{d}{=} (P_{(1)}, \ldots, P_{(n)}, Q_{(1)}, \ldots, Q_{(m)}).$$

Note that

$$\begin{aligned} W(X_*, Y^*) &= \sum_{i=1}^{n'} \sum_{j=m-m'+1}^{m} \mathrm{I}\left(Y_{(j)} \leq X_{(i)}\right) - n'(n'+1) \\ &= \sum_{i=1}^{n'} \sum_{j=m-m'+1}^{m} \mathrm{I}\left(R(Y_{(j)}, Z) \leq R(X_{(i)}, Z)\right) - n'(n'+1), \end{aligned}$$

is a Borel-measurable function of $R(X_{(1)}, Z), \ldots, R(X_{(n')}, Z), R(Y_{(m-m'+1)}, Z), \ldots, R(Y_{(m)}, Z)$. Similarly, we have

$$W(P_*, Q^*) = \sum_{i=1}^{n'} \sum_{j=m-m'+1}^{m} \mathrm{I}\left(Q_{(j)} \leq P_{(i)}\right) - n(n+1)$$

is the same function of $P_{(1)}, \ldots, P_{(n')}, Q_{(m-m'+1)}, \ldots, Q_{(m)}$. Hence, we have

$$W(X_*, Y^*) \stackrel{d}{=} W(P_*, Q^*).$$

Similarly, we can show that $W(X^*, Y_*) \stackrel{d}{=} W(P^*, Q_*)$, which completes our proof. □

Following Proposition 6.2.2, we can obtain the approximate distributions of $W(X_*, Y^*)$ and $W(X^*, Y_*)$ under the null hypothesis using numerical simulations. Algorithm 1 summarise the steps for generating independent random samples following the same distribution as $W(X_*, Y^*)$ under the null hypothesis.

Following Algorithm 1, the approximate distribution of $W(X_*, Y^*)$ under the null hypothesis

Algorithm 1 Generate independent random samples following the same distribution as $W(X_*, Y^*)$ under the null hypothesis

Input: Integers n, m, n', m' where $n \geq n'$, $m \geq m'$. The number of samples $B \in \mathbb{N}$.

Output: B independent samples $\{W_i(P_*, Q^*)\}_{i=1}^B$, each following the same distribution as $W(X_*, Y^*)$ under the null.

1: Let $N = n + m$.

2: **for** $k \in \{1, \ldots, B\}$ **do**

3: Generate a random permutation of $\{1, \ldots, N\}$, take the first n as $P = (P_1, \ldots, P_n)$, and the last m as $Q = (Q_1, \ldots, Q_m)$, where $n + m = N$.

4: Order $P_1, \ldots, P_n$ from small to large as $P_{(1)}, \ldots, P_{(n)}$, and $Q_1, \ldots, Q_m$ from small to large as $Q_{(1)}, \ldots, Q_{(m)}$.

5: Let $P_* = P \setminus \{P_{(n'+1)}, \ldots, P_{(n)}\}$ include the smallest $n - n'$ values in P.

6: Let $Q^* = Q \setminus \{Q_{(1)}, \ldots, Q_{(m-m')}\}$ include the largest $m - m'$ values in Q.

7: Compute the Wilcoxon-Mann-Whitney statistic between P_* and Q^* as

$$W_k(P_*, Q^*) = \sum_{i=1}^{n'} \sum_{j=m-m'+1}^{m} \mathrm{I}\left(Q_{(j)} \leq P_{(i)}\right) - n(n+1).$$

8: **end for**

9: **return** The B samples $\{W_i(P_*, Q^*)\}_{i=1}^B$.

is

$$\hat{F}(t) = \sum_{i=1}^{B} \mathrm{I}\left(W_i(P_*, Q^*) \leq t\right)/B.$$

The approximate distribution of $W(X^*, Y_*)$ can be obtained following the similar approach in the algorithm by replacing P_* and Q^* as P^* and Q_*, respectively.

6.2.2. p-value of the proposed variant method

Theorem 6.2.1 shows that the Wilcoxon-Mann-Whitney test statistic after ignoring missing data is bounded such that $W(X_*, Y^*) \leq W(X', Y') \leq W(X^*, Y_*)$. In Proposition 6.2.2, we further investigate the distributions of $W(X_*, Y^*)$ and $W(X^*, Y_*)$ under the null hypothesis and propose a numerical simulation approach for approximating the distributions in Algorithm 1.

Based on these results, we now propose a novel Wilcoxon-Mann-Whitney test method in the presence of missing data with controlled Type I error regardless of the missingness mechanism. The key idea of our proposed method is to consider the probability of events that are more "extreme" than observing $W(X', Y')$ under the null hypothesis, regardless of the missingness mechanisms. Our main result is summarised as the following

Proposition 6.2.3. Suppose $X = \{X_1, \ldots, X_n\}$ and $Y = \{Y_1, \ldots, Y_m\}$ are independent continuous random variables following the same distribution, $X' = \{X'_1, \ldots, X'_{n'}\}$ is a non-empty subset of X and $Y' = \{Y'_1, \ldots, Y'_{m'}\}$ is a non-empty subset of Y. Let $\{W_i(P_*, Q^*)\}_{i=1}^B$ and

$\{W_i(P^*, Q_*)\}_{i=1}^B$ be independent random variables generated according to Algorithm 1. Define

$$p_1 = \frac{1 + \sum_{i=1}^B \mathrm{I}\left(W_i(P_*, Q^*) \leq W(X', Y')\right)}{B+1},$$
$$\text{and } p_2 = \frac{1 + \sum_{i=1}^B \mathrm{I}\left(W_i(P^*, Q_*) \geq W(X', Y')\right)}{B+1}.$$

Then, we have $\Pr(p_1 \leq \alpha/2) \leq \alpha/2$ and $\Pr(p_2 \leq \alpha/2) \leq \alpha/2$ for a given significance level $\alpha \in (0, 1)$. Therefore, defining the p-value of the test statistic $W(X', Y')$ as

$$p = \min\{2\min\{p_1, p_2\}, 1\}, \tag{6.3}$$

we further have $\Pr(p \leq \alpha) \leq \alpha$.

Proof. The proof is in Section D.6 of the Appendix. □

Remark 6.2.4. According to Proposition 6.2.3, the probability that the p-value in Equation (6.3) is smaller or equal to a significance level $\alpha \in (0, 1)$ is bounded by α. Therefore, rejecting the null hypothesis whenever the p-value is less than or equal to α ensures that the proposed method controls the Type I error, without requiring any assumptions about the missingness mechanism.

6.2.3. Simulations for the variant location method

We run numerical simulations for evaluating the Type I error and statistical power of the variant method proposed in this section, and compare it with the original proposed location testing method from Chapter 3.

The first experiment is similar to the one considered in Figure 3.2 when the proportion of missing data is increasing. The results in Figure 6.6 show that the variant method, like the original proposed location testing method, successfully controls the Type I error. However, the variant method demonstrates higher statistical power than the original method. The experiment in Figure 6.7 follows a similar setup to Figure 3.2, but with the mean of the distribution of Y increased to 2. Under this setting, the variant method outperforms the original method by larger gap: the variant method maintains strong power even when roughly 25% of the data are missing, whereas the original method fails to produce significant results.

Figure 6.8 considers a similar experiment to that considered in 3.4 when the sample size is increasing, but the proportion of missing data are fixed at 10%. It shows that the variant method of this section is more powerful than the original proposed location testing method from Chapter 3 across all sample sizes considered in this experiment. Both methods control the Type I error, while all other methods fail to do so.

Figure 6.9 considers a similar experiment as the one for Figure 3.7, when the mean value of the distribution of sample Y increases with 10% of data missing. The results show that the variant method exhibits higher statistical power than the original method across different mean values of sample Y. Figure 6.10 considers a similar experiment, but with 20% of data missing, rather than 10%, it shows that the advantage of variant method over the original method increases with larger proportion of missing data. Notably, other missing data methods exhibit significant drops

in power when the mean shift is small. The power of these methods starts recovering when the mean value of the distribution of Y is larger than 0.5.

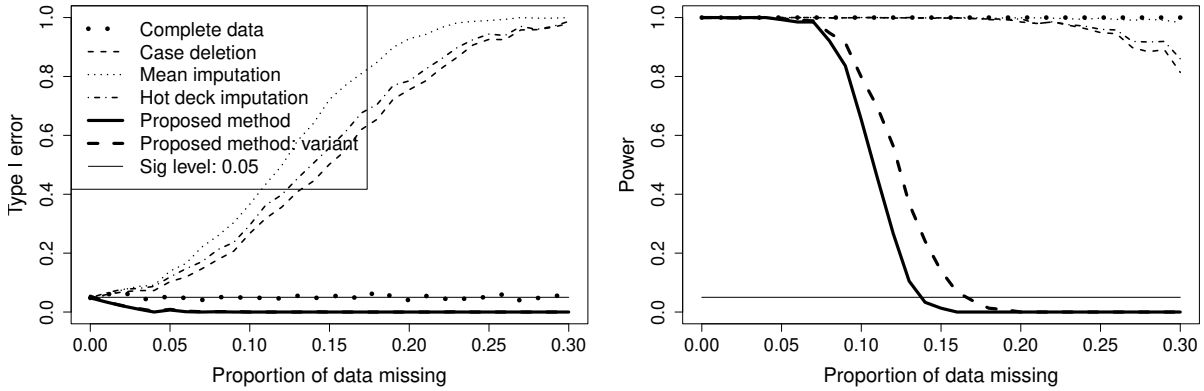


Figure 6.6.: The Type I error and statistical power of the variant of the proposed location testing method, the original proposed location testing method, and the standard Wilcoxon-Mann-Whitney test after the missing data is either known or has been imputed or ignored as the proportion of missing data increases. The data is missing not at random (MNAR), according to the mechanism in Equation (3.10), and (3.11). (Left) Type I error: $\mathrm{N}(0,1)$ vs $\mathrm{N}(0,1)$; (Right) Power: $\mathrm{N}(0,1)$ vs $\mathrm{N}(1,1)$. For both figures, a significance threshold of $\alpha = 0.05$ has been used and the total sample sizes are $n = 100$, $m = 100$, and 1000 trials were used.

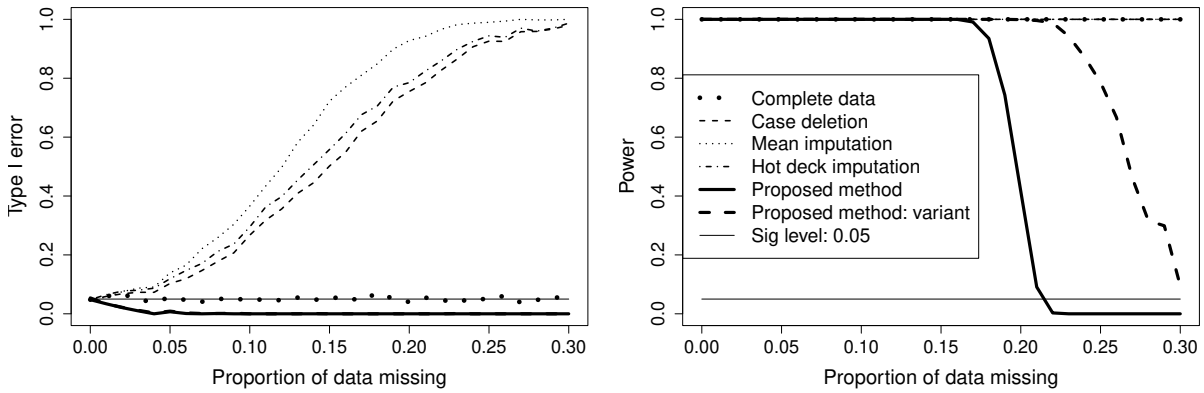


Figure 6.7.: The Type I error and statistical power of the variant of the proposed location testing method, the original proposed location testing method, and the standard Wilcoxon-Mann-Whitney test after the missing data is either known or has been imputed or ignored as the proportion of missing data increases. The data is missing not at random (MNAR), according to the mechanism in Equation (3.10), and (3.11). (Left) Type I error: $\mathrm{N}(0,1)$ vs $\mathrm{N}(0,1)$; (Right) Power: $\mathrm{N}(0,1)$ vs $\mathrm{N}(2,1)$. For both figures, a significance threshold of $\alpha = 0.05$ has been used and the total sample sizes are $n = 100$, $m = 100$, and 1000 trials were used.

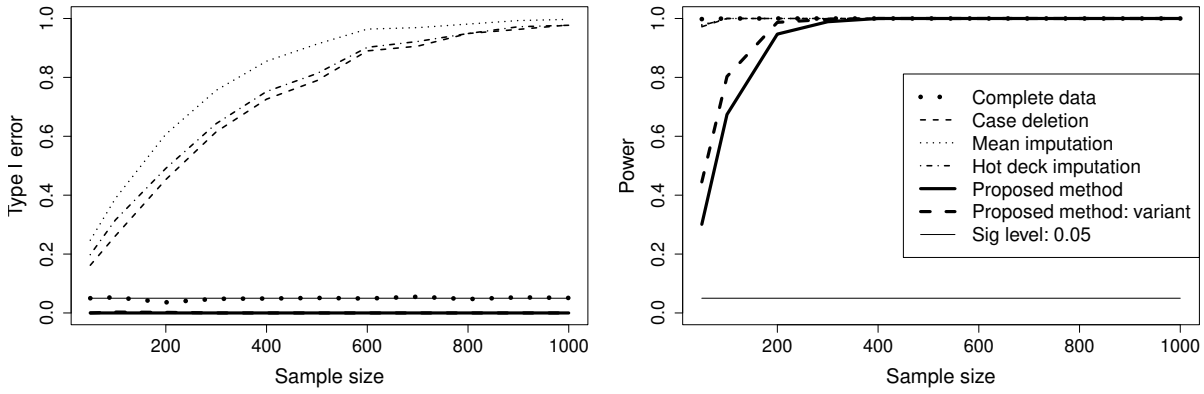


Figure 6.8.: The Type I error and statistical power of the variant of the proposed location testing method, the original proposed location testing method, and the standard Wilcoxon-Mann-Whitney test after the missing data is either known or has been imputed or ignored as the sample sizes increase. The data is missing not at random (MNAR), according to the mechanism in Equation (3.10), and (3.11). (Left) Type I error: $\mathrm{N}(0,1)$ vs $\mathrm{N}(0,1)$; (Right) Power: $\mathrm{N}(0,1)$ vs $\mathrm{N}(1,1)$. For both figures, a significance threshold of $\alpha = 0.05$ has been used and the proportion of missing data for both samples is $s = 0.1$, and 1000 trials were used.

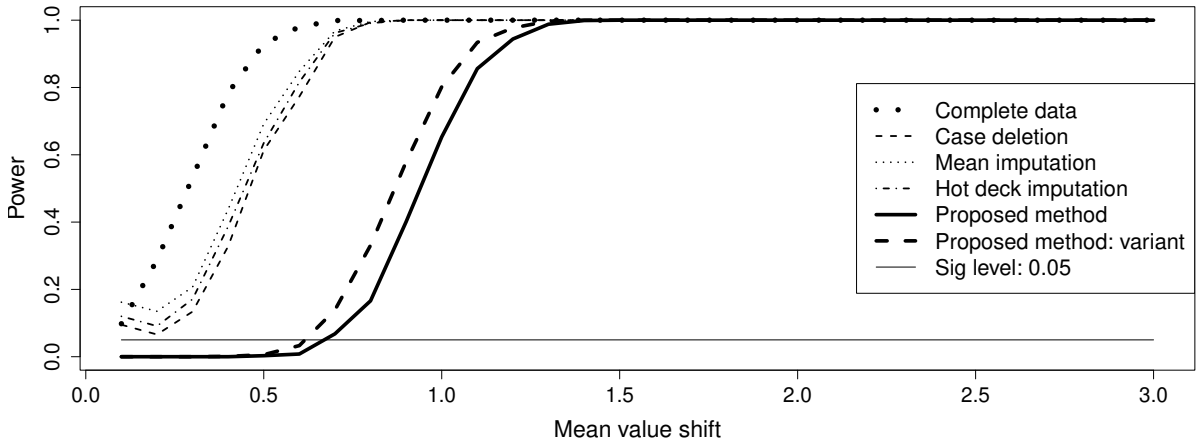


Figure 6.9.: Statistical power of the variant of the proposed location testing method, the original proposed location testing method, and the standard Wilcoxon-Mann-Whitney test after the missing data is either known or has been imputed or ignored as the mean value increases. The data is missing not at random, according to the mechanism in Equation (3.10), and (3.11). Data in samples X and Y are generated from $\mathrm{N}(0, 1)$, and $\mathrm{N}(\Delta, 1)$, respectively, with Δ denoting the mean value shift. A significance threshold of $\alpha = 0.05$ has been used, the total sample sizes are $n = m = 100$, the proportion of missing data for both samples is $s = 0.1$, and 1000 trials were used.

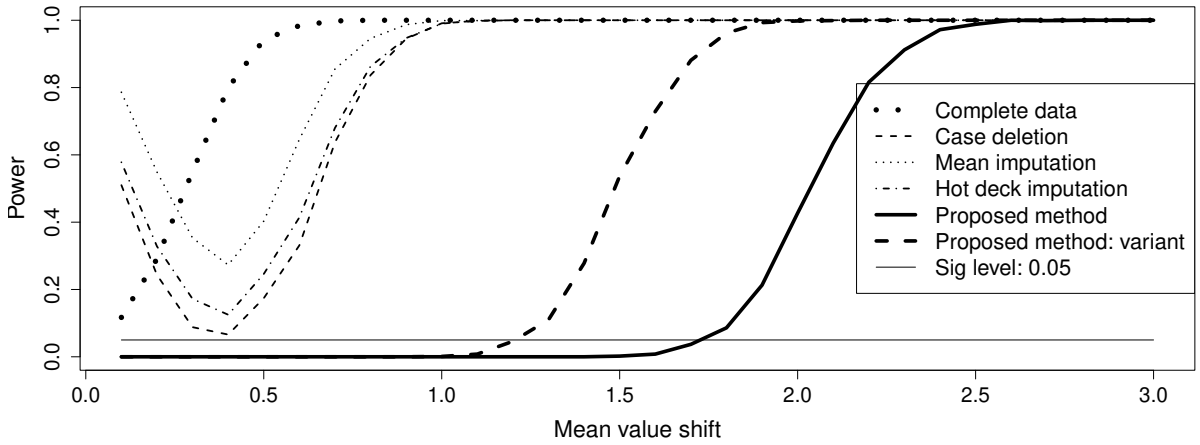


Figure 6.10.: Statistical power of the variant of the proposed location testing method, the original proposed location testing method, and the standard Wilcoxon-Mann-Whitney test after the missing data is either known or has been imputed or ignored as the mean value increases. The data is missing not at random (MNAR), according to the mechanism in Equation (3.10), and (3.11). Data in samples X and Y are generated from $\mathrm{N}(0, 1)$, and $\mathrm{N}(\Delta, 1)$, respectively, with Δ denoting the mean value shift. A significance threshold of $\alpha = 0.05$ has been used, the total sample sizes are $n = m = 100$, the proportion of missing data for both samples is $s = 0.2$, and 1000 trials were used.

6.3. Conclusion

This section considers two different extensions of the proposed location testing method from Chapter 3 for obtaining higher statistical power. Importantly, both extensions still make no assumptions of the missingness mechanisms, and are shown to be able to control the Type I error regardless of values of missing data.

The first extension considers the proposed location testing method when the data can be tied and are defined in potentially closed support. The bounds of the Wilcoxon-Mann-Whitney test statistic and the corresponding bounds of the p-value are derived. We perform numerical simulations and demonstrate that, by taking into account the ties and closed support, the power of the proposed method can be improved.

The second extension considers the Wilcoxon-Mann-Whitney test statistic with missing data as a random variable, rather than a realised observation as considered for the original proposed method. We bound the Wilcoxon-Mann-Whitney test statistic using other random variables with fixed distributions under the null, and prove that these distributions can be approximated using numerical simulations. Based on these results, we propose a new testing method based on the Wilcoxon-Mann-Whitney test, and prove that this method also controls the Type I error without the need to assume the missingness mechanisms. Simulation results demonstrate that this variant method has higher statistical power than the original method, especially when the proportion of missing data are large. See Figure 6.10 as an example.

7. Conclusion

Addressing the missing data problem in hypothesis testing is important yet challenging. When the missing data are not handled correctly, the testing results could have a Type I error rate higher than a pre-specified significance level. This problem does not vanish with larger sample sizes; on the contrary, the Type I error could asymptotically converges to 1 as sample sizes increase, even when only 1% of the data are missing. See Figure 3.5 in this thesis for an example. The missingness mechanism behind the missing data is often unknown, making the proper treatment of missing data difficult.

This thesis proposes several rank-based hypothesis testing methods in the presence of missing data, without making any assumptions about the missingness mechanisms. Specifically, we adapt the Wilcoxon-Mann-Whitney test, the Ansari-Bradley test, and Spearman's footrule to handle missing data for the location, scale, and independence tests, respectively. An important property of these methods is that the Type I error is always controlled, regardless of the values of missing data.

To summarise the contributions of this thesis:

- We derive the tight lower and upper bounds of the Wilcoxon-Mann-Whitney test statistic in the presence of missing data. Based on the bounds, the corresponding bounds of the p-value of the test results are constructed. We prove that, by rejecting the null hypothesis when all possible p-values are significant, the proposed method controls the Type I error regardless of the values of the missing data. The statistical power of our method is studied. We show that when more than 30% of the data are missing in both samples X and Y, the method does not produce any significant result, regardless of the observed values. The approximate and asymptotic statistical power of the proposed method are also investigated, under the assumption that the data are missing completely at random (MCAR). We perform various numerical experiments for evaluating the Type I error and statistical power of the proposed method, and compare it with other common missing data methods. Finally, our method is illustrated using clinical trial data that contains missing values.

- We extend the framework for the Wilcoxon-Mann-Whitney test to the Ansari-Bradley test for the scale testing problem. Tight lower and upper bounds of the Ansari-Bradley test statistic are derived, and then applied to obtain the corresponding bounds of the p-value. By rejecting the null hypothesis when the maximum possible p-value is smaller than the significance level α, we propose a new scale testing method in the presence of missing data. This method is then combined with the proposed location testing method using the Holm-Bonferroni method, resulting in a location-scale testing method that can handle missing data. The performance of the proposed scale and location-scale tests is evaluated through extensive numerical simulations. A hepatitis C virus dataset is considered for illustrating

the practical utility of the proposed location-scale testing method.

- We construct tight lower and upper bounds for Spearman's footrule in the presence of missing data, and propose corresponding algorithms for efficiently computing these bounds. Based on these results, we further derive non-trivial bounds of the Kendall's τ coefficient. The bounds of the p-value of Spearman's footrule are also constructed. By rejecting the null hypothesis whenever the maximum possible p-value is smaller than α, we introduce a novel independence testing method that accommodates missing data without making any missingness assumptions. The Type I error rate and statistical power of the proposed method are evaluated and compared with other common approaches for missing data through extensive numerical simulations. We illustrate our method on real-world data by comparing the dependence of yearly precipitation levels in the United Kingdom, Germany and the Unites States.

- We extend the proposed location testing method based on the Wilcoxon-Mann-Whitney test to cases where data can be tied and defined on a potentially closed support. The bounds of the Wilcoxon-Mann-Whitney test statistic, and its corresponding variance are derived, and then used to derive the bounds of the corresponding p-value. We perform various numerical simulations and demonstrate that by taking into account ties and closed support, the power of the proposed location testing method can be further increased.

- We consider a variant of the proposed location testing method by treating the Wilcoxon-Mann-Whitney test statistic in the presence of missing data as a random variable, rather than a realised value. We propose to bound the test statistic using other variables with fixed distributions under the null hypothesis. We prove that these distributions can be approximated using numerical simulations, and subsequently apply this result for constructing a novel location testing method with controlled Type I error. This variant of the proposed method is evaluated using numerical simulations, and is shown to be more powerful than the original proposed location testing method.

Future Work

While there could be multiple different directions for extending the work in this thesis, the following four topics represent promising areas to be explored.

Consideration of ties

The proposed methods in this thesis assume the data are distinct. For discrete-valued data, however, two or more observations could be equal, so the data can be tied. Chapter 6 extends the proposed location testing method to cases where data can be tied and defined on a potentially closed support, and demonstrates that by taking ties and the closed support into account, the power of the proposed method can be improved. Following the similar approach used in Chapter 6, it is possible to extend the proposed scale testing method and independence testing method to handle ties and closed support.

Bounding test statistics as random variables

The proposed methods in this thesis are based on the bounds of the corresponding test statistics when the variables are realised. In other words, the test statistics are considered as realised real values. In Chapter 6, we show that it is possible to take a different perspective and bound the test statistic in the presence of missing data as a random variable, prior to the data are realised. Chapter 6 explores this methodology for the Wilcoxon-Mann-Whitney test, and shows that this methodology improves the power of the original proposed location testing method quite significantly in certain cases, while also controlling the Type I error for any missingness mechanism. It is possible to extend this methodology to the Ansari-Bradley test and Spearman's footrule independence test. The difficulty in extending this methodology is constructing random variables for bounding the test statistic, and study the distributions of these bounding variables under the null hypothesis.

Extend the framework to other rank-based methods

This thesis focuses mainly on constructing new methods based on the Wilcoxon-Mann-Whitney test, the Ansari-Bradley test, and Spearman's footrule for the location, scale, and independence tests in the presence of missing data. However, the framework in this thesis could be extended to other rank-based hypothesis testing methods. Examples include the Lepage test for the location-scale testing and Spearman's ρ for the independence testing. The main challenge in such extensions lies in deriving tight bounds for the corresponding test statistics in the presence of missing data.

Multivariate hypothesis testing with missing data

While this thesis focuses on univariate data, the framework can be extended to multivariate hypothesis testing. Such an extension requires deriving bounds for the relevant test statistics in the presence of missing data. Multivariate rank-based statistics are natural candidates for this approach (Chen et al., 2022a; Zhou et al., 2024). However, rank-based statistics are not necessary: any statistic that admits valid bounds with missing data can be considered. For example, Zeng et al. (2025a) extend the framework to the Maximum Mean Discrepancy (MMD) test (Gretton et al., 2012) when the Laplacian kernel is applied.

A. Proofs of results in Chapter 3

This chapter provides proof for results used in Chapter 3.

A.1. Bounds of sum of ranks

Proposition A.1.1. Suppose that $X = \{x_1, \cdots, x_n\}$ and $Y = \{y_1, \cdots, y_m\}$ are samples of distinct, real-valued observations. Suppose that $X' \subset X$ is a subset of n' values in X, and suppose that $Y' \subset Y$ is a subset of m' values in Y. Defining $Z = X \cup Y$ and $Z' = X' \cup Y'$ and supposing only Z' is known, the minimum and maximum possible rank sum of X, over all possible values for the $d = n - n' + m - m'$ observations in $Z \setminus Z'$, are

$$\min_{Z \setminus Z'} R_{X|Z} = R_{X'|Z'} + (n - n')(n + n' + 1)/2 \tag{A.1}$$

$$\max_{Z \setminus Z'} R_{X|Z} = R_{X'|Z'} + \{n(n + 2m + 1) - n'(n' + 2m' + 1)\}/2, \tag{A.2}$$

where $R_{X|Z}$ denotes the sum of ranks of X in Z as defined in Definition 2.2.2.

Proof. To start, let us denote $X \setminus X' = X_*$, and $Y \setminus Y' = Y_*$. In other words, X_* is the set that includes all unobserved values in X, and similarly, Y_* is the set that includes all unobserved values in Y.

According to the definition of the sum of ranks in Definition 2.2.2,

$$\begin{aligned} R_{X|Z} &= \sum_{x \in X} R(x, Z) \\ &= \sum_{x \in X'} R(x, Z) + \sum_{x \in X_*} R(x, Z) \\ &= R_{X'|Z} + R_{X_*|Z}. \end{aligned}$$

Consider any $x \in X'$. Since $Z = X \cup Y = X \cup Y' \cup Y_*$,

$$R(x, Z) = R(x, X) + R(x, Y') + R(x, Y_*).$$

It follows that

$$R_{X'|Z} = \underbrace{\sum_{x \in X'} R(x, X)}_{=C_1} + \underbrace{\sum_{x \in X'} R(x, Y')}_{=C_2} + \underbrace{\sum_{x \in X'} R(x, Y_*)}_{=C_3},$$

and

$$R_{X_*|Z} = \underbrace{\sum_{x\in X_*} R(x,X)}_{=D_1} + \underbrace{\sum_{x\in X_*} R(x,Y')}_{=D_2} + \underbrace{\sum_{x\in X_*} R(x,Y_*)}_{=D_3}.$$

Hence,

$$R_{X|Z} = C_1 + C_2 + C_3 + D_1 + D_2 + D_3. \tag{A.3}$$

Notice that

$$\begin{aligned} C_1 + D_1 &= \sum_{x\in X'} R(x,X) + \sum_{x\in X_*} R(x,X) \\ &= \sum_{x\in X} R(x,X) \\ &= (1+n)n/2 = n^2/2 + n/2. \end{aligned}$$

Also, since $Z' = X' \cup Y'$,

$$\begin{aligned} R_{X'|Z'} &= \sum_{x\in X'} R(x,Z') \\ &= \sum_{x\in X'} R(x,X') + \sum_{x\in X'} R(x,Y') \\ &= (1+n')n'/2 + C_2. \end{aligned}$$

Thus,

$$C_2 = R_{X'|Z'} - (1+n')n'/2.$$

Since all values in Z' are observed, C_2 is a constant determined solely by observed values.

Put the above results of $C_1 + D_1$ and C_1 back into Equation (A.3), we obtain that

$$\begin{aligned} R_{X|Z} &= R_{X'|Z'} - (1+n')n'/2 + n^2/2 + n/2 + C_3 + D_2 + D_3 \\ &= R_{X'|Z'} + (n-n')(n+n'+1)/2 + C_3 + D_2 + D_3. \end{aligned} \tag{A.4}$$

The following of this proof derives the bounds of $R_{X|Z}$ by considering the bounds of the terms C_3, D_2 and D_3 separately.

First, recall that

$$C_3 = \sum_{x\in X'} R(x,Y_*) = \sum_{x\in X'} \sum_{y\in Y_*} \mathrm{I}\left(y \le x\right).$$

Notice that

$$C_3 \ge 0.$$

Further, when any $y \in Y^*$ is lager than any $x \in X'$, i.e.

$$\begin{aligned} &\min Y_* > \max X', \\ \Longrightarrow\ & C_3 = 0. \end{aligned}$$

Similarly, recall that

$$D_2 = \sum_{x \in X_*} R(x, Y') = \sum_{x \in X_*} \sum_{y \in Y'} \mathrm{I}\left(y \leq x\right).$$

We have

$$D_2 \geq 0,$$

and when any $y \in Y'$ is larger than any $x \in X_*$, i.e.

$$\begin{aligned} &\min Y' > \max X_*, \\ \Longrightarrow\ & D_2 = 0. \end{aligned}$$

Finally, we recall that

$$D_3 = \sum_{x \in X_*} R(x, Y_*) = \sum_{x \in X_*} \sum_{y \in Y_*} \mathrm{I}\left(y \leq x\right).$$

We have

$$D_3 \geq 0,$$

and when any $y \in Y_*$ is larger than any $x \in X_*$, i.e.

$$\begin{aligned} &\min Y_* > \max X_*, \\ \Longrightarrow\ & D_3 = 0. \end{aligned}$$

Hence, we have shown that

$$C_3, D_2, D_3 \geq 0,$$

where the "=" holds when the following three conditions all hold:

$$\begin{aligned} &\min Y_* > \max X', \\ &\min Y' > \max X_*, \\ &\min Y_* > \max X_*. \end{aligned}$$

Note that when

$$\min Y_* > \max Z' \text{ and } \max X_* < \min Z',$$

the above three conditions hold. In other words, if any unobserved values in $y \in Y_*$ is larger than all observed values in $z \in Z'$ while any unobserved values in $x \in X_*$ is smaller than all observed values in $z \in Z'$, we have

$$\begin{aligned} R_{X|Z} &= {}^{\text{Equation (A.4)}}\ R_{X'|Z'} + (n-n')(n+n'+1)/2 + C_3 + D_2 + D_3 \\ &= R_{X'|Z'} + (n-n')(n+n'+1)/2 + 0 + 0 + 0 \\ &= R_{X'|Z'} + (n-n')(n+n'+1)/2. \end{aligned}$$

Hence, we have

$$\min_{Z \setminus Z'} R_{X|Z} = R_{X'|Z'} + (n-n')(n+n'+1)/2,$$

which proves Equation (A.1).

To prove Equation (A.2), notice that we can apply the result of Equation (A.1) and show that

$$\min_{Z \setminus Z'} R_{Y|Z} = R_{Y'|Z'} + \underbrace{(m-m')(m+m'+1)/2}_{=E_1}.$$

Since

$$R_{Z|Z} = R_{X|Z} + R_{Y|Z} = \underbrace{(1+n+m)(n+m)/2}_{=E_2}$$

is a constant, when $R_{Y|Z}$ takes its minimum among all possible Z^*, $R_{X|Z}$ takes its maximum, i.e.,

$$\begin{aligned} \max_{Z \setminus Z'} R_{X|Z} &= R_{Z|Z} - \min_{Z \setminus Z'} R_{Y|Z} \\ &= E_2 - R_{Y'|Z'} - E_1. \end{aligned}$$

Notice that

$$\begin{aligned} R_{Y'|Z'} + R_{X'|Z'} &= \underbrace{(1+n'+m')(n'+m')/2}_{=E_3} \\ \Longrightarrow R_{Y'|Z'} &= E_3 - R_{X'|Z'}. \end{aligned}$$

Then, we further have

$$\max_{Z \setminus Z'} R_{X|Z} = E_2 - E_3 + R_{X'|Z'} - E_1.$$

We can verify that

$$
\begin{aligned}
&E_2 - E_3 - E_1 \\
&= (1+n+m)(n+m)/2 - (1+n'+m')(n'+m')/2 - (m-m')(m+m'+1)/2 \\
&= \{n+m+n^2+m^2+2nm - (1+n'+m')(n'+m') - m^2 - mm' - m + mm' + m'^2 + m'\}/2 \\
&= \{n+n^2+2nm - n' - m' - n'^2 - m'^2 - 2n'm' + m'^2 + m'\}/2 \\
&= \{n+n^2+2nm - (n' + n'^2 + 2n'm')\}/2 \\
&= \{n(n+2m+1) - n'(n'+2m'+1)\}/2.
\end{aligned} \tag{A.5}
$$

Hence, we have

$$\max_{Z \setminus Z'} R_{X|Z} = R_{X'|Z'} + \{n(n+2m+1) - n'(n'+2m'+1)\}/2.$$

This proves Equation (A.2), and completes our proof. □

A.2. Maximum allowable proportion of missing data

The following proposition, while not stated in the main paper, is needed to prove Proposition 3.3.1.

Proposition A.2.1. Suppose $X = \{x_1, x_2, \cdots, x_n\}$ and $Y = \{y_1, y_2, \cdots, y_m\}$ are samples of distinct real-valued observations and $X' \subset X$, $Y' \subset Y$ are subsets of distinct values with size n' and m' respectively, which are observed. Suppose that n, m are sufficiently large enough for applying normality approximation in Wilcoxon-Mann-Whitney test. Then, for any significance level $\alpha \in (0,1)$, Condition 3.2.1 or Condition 3.2.2 is true if and only if either of the following two inequalities are true

$$
\begin{aligned}
W(X', Y') &> \Phi^{-1}(1-\alpha/2)\{nm(n+m+1)/12\}^{1/2} + nm/2, \\
W(X', Y') &< \Phi^{-1}(\alpha/2)\{nm(n+m+1)/12\}^{1/2} - nm/2 + n'm',
\end{aligned}
$$

where $\Phi(x)$ is cumulative distribution function of the standard normal distribution.

Proof. Recall that Condition 3.2.1 is defined as

$$\frac{W_{\min}(X,Y) - \mu}{\sigma} \geq \Phi^{-1}\left(1 - \frac{\alpha}{2}\right) \text{ and } \frac{W_{\max}(X,Y) - \mu}{\sigma} \geq \Phi^{-1}\left(1 - \frac{\alpha}{2}\right).$$

Since $W_{\min}(X,Y) \leq W_{\max}(X,Y)$, and $\Phi^{-1}(x)$ is a monotonic increasing function, hence Condition 3.2.1 is equivalent to

$$
\begin{aligned}
&\frac{W_{\min}(X,Y) - \mu}{\sigma} \geq \Phi^{-1}\left(1 - \frac{\alpha}{2}\right) \\
\iff &W_{\min}(X,Y) \geq \sigma\Phi^{-1}\left(1 - \frac{\alpha}{2}\right) + \mu \\
\iff &W(X', Y') > \Phi^{-1}(1-\alpha/2)\{nm(n+m+1)/12\}^{1/2} + nm/2.
\end{aligned}
$$

Similarly, recall that Condition 3.2.2 is defined as

$$\frac{W_{\min}(X,Y)-\mu}{\sigma} \leq \Phi^{-1}\left(\frac{\alpha}{2}\right) \text{ and } \frac{W_{\max}(X,Y)-\mu}{\sigma} \leq \Phi^{-1}\left(\frac{\alpha}{2}\right).$$

Since $W_{\min}(X,Y) \leq W_{\max}(X,Y)$, and $\Phi^{-1}(x)$ is a monotonic increasing function, hence Condition 3.2.2 is equivalent to

$$\begin{aligned}
&\frac{W_{\max}(X,Y)-\mu}{\sigma} \leq \Phi^{-1}\left(\frac{\alpha}{2}\right) \\
\iff & W_{\max}(X,Y) \leq \sigma\Phi^{-1}\left(\frac{\alpha}{2}\right)+\mu \\
\iff & W(X',Y')+nm-n'm' < \Phi^{-1}(\alpha/2)\{nm(n+m+1)/12\}^{1/2}+nm/2 \\
\iff & W(X',Y') < \Phi^{-1}(\alpha/2)\{nm(n+m+1)/12\}^{1/2}-nm/2-n'm'.
\end{aligned}$$

This completes our proof. □

The following well-known result regarding the Wilcoxon-Mann-Whitney statistic will be useful to prove Proposition 3.3.1.

Lemma A.2.2. If X and Y are samples of n and m values, respectively, with all values distinct, then the Wilcoxon-Mann-Whitney test statistic $W(X,Y)$ is bounded by $0 \leq W(X,Y) \leq nm$.

Proof. By definition, $W(X,Y) = R_{X|X\cup Y} - n(n+1)/2$, and

$$R_{X|X\cup Y} = \sum_{x\in X} R(x, X\cup Y).$$

The rank sum $R_{X|X\cup Y}$ is minimised when the values in X occupy ranks $1,\ldots,n$ in $X\cup Y$, and so $R_{X|X\cup Y} = n(n+1)/2$, which implies $W(X,Y)=0$ is the minimum possible value for $W(X,Y)$. On the other hand, $R_{X|X\cup Y}$ is maximised when the values in X occupy ranks $n+1,\ldots,n+m$ in $X\cup Y$, which implies $R_{X|X\cup Y} = nm+n(n+1)/2$, which implies $W(X,Y)=nm$ is the maximum possible value for $W(X,Y)$. □

We are finally ready to provide Proposition 3.3.1 in the main paper which is restated here:

Proposition A.2.3. Suppose that $X=\{x_1,,\cdots,x_n\}$ and $Y=\{y_1,\cdots,y_m\}$ are samples of distinct real-valued observations and $X'\subset X$ and $Y'\subset Y$ are subsets of distinct values with size n' and m', respectively, which are observed. Suppose that n,m are sufficiently large enough for applying normal approximation in the Wilcoxon-Mann-Whitney test. Then for any significance threshold $\alpha\in(0,1)$, if

$$n'm'/(nm) < 1/2+\sqrt{(1+n+m)/(12nm)}\Phi^{-1}(1-\alpha/2), \tag{A.6}$$

where $\Phi(x)$ denotes the cumulative distribution function of the standard normal distribution, the proposed method will not reject the null hypothesis, regardless of the observed values in X' and Y'.

Proof. Suppose the conditions of the proposition are satisfied, and

$$n'm'/(nm) < 1/2+\Phi^{-1}(1-\alpha/2)\{(n+m+1)/(12nm)\}^{1/2}. \tag{A.7}$$

Then

$$
\begin{aligned}
& n'm'/(nm) < 1/2 + \Phi^{-1}(1-\alpha/2)\{(n+m+1)/(12nm)\}^{1/2} \\
\iff & n'm' < nm/2 + \Phi^{-1}(1-\alpha/2)\{nm(n+m+1)/12\}^{1/2} \\
\iff & 0 < -n'm' + nm/2 + \Phi^{-1}(1-\alpha/2)\{nm(n+m+1)/12\}^{1/2} \\
\iff & 0 > n'm' - nm/2 - \Phi^{-1}(1-\alpha/2)\{nm(n+m+1)/12\}^{1/2} \\
\iff & W(X',Y') \geq 0 > n'm' - nm/2 - \Phi^{-1}(1-\alpha/2)\{nm(n+m+1)/12\}^{1/2} \\
\iff & W(X',Y') > \Phi^{-1}(\alpha/2)\{nm(n+m+1)/12\}^{1/2} - nm/2 + n'm', \qquad \text{(A.8)}
\end{aligned}
$$

since $\Phi^{-1}(1-x) = -\Phi^{-1}(x)$, for $x \in (0,1)$, and using Lemma A.2.2 applied to X' and Y', i.e. $0 \leq W(X',Y') \leq n'm'$. Similarly,

$$
\begin{aligned}
& n'm'/(nm) < 1/2 + \Phi^{-1}(1-\alpha/2)\{(n+m+1)/(12nm)\}^{1/2} \\
\iff & n'm' < nm/2 + \Phi^{-1}(1-\alpha/2)\{nm(n+m+1)/12\}^{1/2} \\
\iff & W(X',Y') \leq n'm' < nm/2 + \Phi^{-1}(1-\alpha/2)\{nm(n+m+1)/12\}^{1/2} \\
\iff & W(X',Y') < \Phi^{-1}(1-\alpha/2)\{nm(n+m+1)/12\}^{1/2} + nm/2. \qquad \text{(A.9)}
\end{aligned}
$$

Given that if Inequality (A.7) is true it implies both Inequalities (A.8) and (A.9) are true, this means that neither of the inequalities in Proposition A.2.1 are true, which implies that neither Conditions 3.2.1 nor Condition 3.2.2 is true. Hence, the proposed will not reject the null hypothesis. □

A.3. Power for data missing completely at random

The following proposition provides results for the power of the proposed method, when the data are assumed to be missing completely at random.

Proposition A.3.1. Let $X = \{X_1, \cdots, X_n\}$ and $Y = \{Y_1, \cdots, Y_m\}$ be sets of random variables identically and independently distributed according to continuous distributions F and G, respectively. Let $X' \subset X$ and $Y' \subset Y$ denote the subsets of distinct random variables with sizes n' and m', respectively, which can be observed. Assume that the observations of the random variables in the complements $X \setminus X'$ and $Y \setminus Y'$ are missing completely at random (MCAR), and $0 < \Pr(X_1 < Y_1) < 1$. Suppose that n, m are sufficiently large enough for applying normal approximation in the Wilcoxon-Mann-Whitney test. Then, for any given significance level $\alpha \in (0,1)$, the proposed method yields a significant p-value with probability approximately equal to

$$\Phi((L-\mu')/\sigma') + 1 - \Phi((R-\mu')/\sigma'),$$

where Φ is the cumulative distribution function of the standard normal distribution, and L, R, μ', σ'

are

$$L = \sigma\Phi^{-1}(\alpha/2) + \mu - nm + n'm', \quad R = \sigma\Phi^{-1}(1-\alpha/2) + \mu,$$
$$\mu' = m'n'p_1,\ (\sigma')^2 = m'n'p_1(1-p_1) + m'n'(n'-1)(p_2 - p_1^2) + n'm'(m'-1)(p_3 - p_1^2),$$

and $\mu = nm/2$, $\sigma^2 = nm(n+m+1)/12$, and

$$p_1 = \Pr(X_1 < Y_1),\ p_2 = \Pr(X_1 < Y_1 \text{ and } X_1 < Y_2),\ p_3 = \Pr(X_1 < Y_1 \text{ and } X_2 < Y_1).$$

Proof. The proof is based on the results in Section 3, Chapter 2 in Lehmann and D'Abrera (1975).

Since the missingness mechanism is assumed missing completely at random for both X and Y, for all observed variables in X and Y, we must have X' and Y' are also identical and independently distributed according to F and G, respectively.

Then, from (2.14) in Lehmann and D'Abrera (1975), since F, G are both continuous cumulative distribution function and $0 \leq P(X_1 < Y_1) \leq 1$, $W(X', Y')$ approximately follows a normal distribution, i.e.

$$(W(X', Y') - \mu')/\sigma' \sim \mathrm{N}(0, 1) \tag{A.10}$$

approximately, where $\mu' = E(W(X', Y'))$ and $\sigma' = (\mathrm{var}(W(X', Y')))^{1/2}$, and from the Equations (2.17) and (2.21) in Lehmann and D'Abrera (1975) respectively, we have

$$\mu' = m'n'p_1,$$
$$\sigma'^2 = m'n'p_1(1-p_1) + m'n'(n'-1)(p_2 - p_1^2) + n'm'(m'-1)(p_3 - p_1^2),$$

where

$$p_1 = \Pr(X_1 < Y_1), p_2 = \Pr(X_1 < Y_1 \text{ and } X_1 < Y_2), p_3 = \Pr(X_1 < Y_1 \text{ and } X_2 < Y_1).$$

In Proposition A.2.1, we showed that the proposed method yields a significant p-value if and only if either of the following two inequalities are true

$$W(X', Y') < \Phi^{-1}(\alpha/2)\{nm(n+m+1)/12\}^{1/2} - nm/2 + n'm',$$
$$W(X', Y') > \Phi^{-1}(1-\alpha/2)\{nm(n+m+1)/12\}^{1/2} + nm/2,$$

Recall that $\sigma = \{nm(n+m+1)/12\}^{1/2}$ and $\mu = nm/2$. We can denote the right hand sides of the above two inequalities as $L = \sigma\Phi^{-1}(\alpha/2) + \mu - nm + n'm'$, and $R = \sigma\Phi^{-1}(1-\alpha/2) + \mu$, respectively. Subsequently,

$$\begin{aligned}
&\Pr(\text{Proposed method yields a significant } p\text{-value}) \\
&= \Pr\big(W(X', Y') < L \text{ or } W(X', Y') > R\big) \\
&= \Pr\big(W(X', Y') < L\big) + \Pr\big(W(X', Y') > R\big).
\end{aligned}$$

Notice that, according to the approximation result in (A.10),

$$\Pr\left(W(X',Y') < L\right) = \Pr\left((W(X',Y') - \mu')/\sigma' < (L-\mu')/\sigma'\right) \approx \Phi\left\{(L-\mu')/\sigma'\right\},$$
$$\Pr\left(W(X',Y') > R\right) = \Pr\left((W(X',Y') - \mu')/\sigma' > (R-\mu')/\sigma'\right) \approx 1 - \Phi(R-\mu')/\sigma'.$$

Hence,

$$\begin{aligned}&\Pr\left(\text{Proposed method yields a significant } p\text{-value}\right)\\ &\approx \Phi\left((L-\mu')/\sigma'\right) + 1 - \Phi\left((R-\mu')/\sigma'\right),\end{aligned}$$

which completes our proof. □

A.4. Asymptotic power for data missing completely at random

Proposition A.3.1 provides approximate power of the proposed method for finite sample sizes when the data are missing completely at random. However, the following proposition considers asymptotic power of the proposed method when the sample sizes are infinite.

Proposition A.4.1. Let $X = \{X_1, \cdots, X_n\}$ and $Y = \{Y_1, \cdots, Y_m\}$ be sets of random variables identically and independently distributed according to continuous distributions F and G, respectively. Let $X' \subset X$ and $Y' \subset Y$ denote the subsets of distinct random variables with sizes n' and m', respectively, which can be observed. Assuming that the random variables in the complements $X \setminus X'$ and $Y \setminus Y'$ are missing completely at random (MCAR), and $0 < \Pr(X_1 < Y_1) < 1$, and additionally assuming $m' \leq n'$ and

$$m'/n' \to \lambda' \text{ when } n, m \to \infty, \text{ where } 0 < \lambda' \leq 1, \tag{A.11}$$
$$n'/n \to \lambda_x, \text{ when } n, m \to \infty, \text{ where } 0 < \lambda_x \leq 1, \tag{A.12}$$
$$m'/m \to \lambda_y, \text{ when } n, m \to \infty, \text{ where } 0 < \lambda_y \leq 1 \tag{A.13}$$
$$\lambda_x\lambda_y(p_1 - 1) + 1/2 \neq 0 \text{ and } \lambda_x\lambda_y p_1 - 1/2 \neq 0, \text{ where } p_1 = \Pr(X_1 < Y_1). \tag{A.14}$$

As $n, m \to \infty$, with probability

$$p = \begin{cases} 0, & \text{if } \lambda_x\lambda_y(p_1-1) + 1/2 > 0 \text{ and } \lambda_x\lambda_y p_1 - 1/2 < 0, \\ 1, & \text{otherwise,} \end{cases} \tag{A.15}$$

the proposed method will yield a significant p-value.

Proof. For any given n, m, n', m', since the data are assumed to be missing completely at random (MCAR), the observed variables X' and Y' are identically and independently distributed according to F and G, respectively.

Since $0 < \lambda' \leq 1$ and $0 < \Pr(X_1 < Y_1) < 1$, the Corollary 5 in Lehmann and D'Abrera (1975) (see Appendix: U-Statistic Corollary 5 in Lehmann and D'Abrera (1975)) can be applied, and we obtain that

$$(W(X',Y') - \mu')/\sigma' \to^d \mathrm{N}(0,1), \text{ when } n, m \to \infty, \tag{A.16}$$

with

$$\mu' = m'n'p_1, \tag{A.17}$$

$$\sigma' = \{m'n'p_1(1-p_1) + m'n'(n'-1)(p_2-p_1^2) + n'm'(m'-1)(p_3-p_1^2)\}^{1/2}. \tag{A.18}$$

Notice that

$$m/n = \frac{m}{m'}\frac{m'}{n'}\frac{n'}{n}.$$

According to (A.11), (A.12) and (A.13), we have limitations

$$\lim_{n,m\to\infty} m/m' \to 1/\lambda_y, \text{where } 0 < \lambda_y \leq 1,$$
$$\lim_{n,m\to\infty} m'/n' \to \lambda', \text{where } 0 < \lambda' \leq 1,$$
$$\lim_{n,m\to\infty} n'/n \to \lambda_x, \text{where } 0 < \lambda_x \leq 1.$$

Hence, the limitation of m/n exists and

$$\lim_{n,m\to\infty} m/n = \left(\lim_{n,m\to\infty} m/m'\right)\left(\lim_{n,m\to\infty} m'/n'\right)\left(\lim_{n,m\to\infty} n'/n\right) = \lambda'\lambda_x/\lambda_y.$$

For notation convenience, let us denote

$$\lambda = \lambda'\lambda_x/\lambda_y.$$

Let us now consider the probability of the proposed method yielding a significant p-value. Proposition A.2.1 asserts that

$$\begin{aligned}&\Pr\left(\text{Proposed method yields a significant } p\text{-value}\right)\\ &= \Pr\left(W(X',Y') \leq \sigma\Phi^{-1}(\alpha/2) + \mu - nm + n'm'\right) + \Pr\left(W(X',Y') \geq \sigma\Phi^{-1}(1-\alpha/2) + \mu\right),\end{aligned}$$

where $\mu = nm/2$ and $\sigma^2 = nm(n+m+1)/12$. Further,

$$\begin{aligned}&\Pr\left(\text{Proposed method yields a significant } p\text{-value}\right)\\ &= \Pr\left(W(X',Y') - \mu + nm - n'm' \leq \sigma\Phi^{-1}(\alpha/2)\right) + \Pr\left(W(X',Y') - \mu \geq \sigma\Phi^{-1}(1-\alpha/2)\right)\\ &= \Pr\left(\frac{(W(X',Y') - \mu + nm - n'm')}{\sigma} \leq \Phi^{-1}(\alpha/2)\right) + \Pr\left(\frac{W(X',Y') - \mu}{\sigma} \geq \Phi^{-1}(1-\alpha/2)\right)\\ &= \Pr\left(\frac{\sigma'}{\sigma}\frac{W(X',Y') - \mu'}{\sigma'} + \frac{\mu' - \mu + nm - n'm'}{\sigma} \leq \Phi^{-1}(\alpha/2)\right)\\ &+ \Pr\left(\frac{\sigma'}{\sigma}\frac{W(X',Y') - \mu'}{\sigma'} + \frac{\mu' - \mu}{\sigma} \geq \Phi^{-1}(1-\alpha/2)\right).\end{aligned}$$

Denote

$$a_1(n,m,n'm') = (\mu' - \mu + nm - n'm')/\sigma, \text{ and } a_2(n,m,n',m') = (\mu' - \mu)/\sigma.$$

Then

$$\begin{aligned}
&\Pr\left(\text{Proposed method yields a significant } p\text{-value}\right) \\
&= \Pr\left((\sigma'/\sigma)(W(X',Y') - \mu')/\sigma' + a_1(n,m,n'm') \leq \Phi^{-1}(\alpha/2)\right) \\
&+ \Pr\left((\sigma'/\sigma)(W(X',Y') - \mu')/\sigma' + a_2(n,m,n'm') \geq \Phi^{-1}(1-\alpha/2)\right).
\end{aligned} \tag{A.19}$$

Our goal is the limitation of the Equation (A.19) when $n, m \to \infty$. We start by considering the limitation of σ'/σ. Recall that $\sigma^2 = nm(n+m+1)/12$, and σ' is defined in Equation (A.18). Then

$$\left(\frac{\sigma'}{\sigma}\right)^2 = 12\frac{m'n'p_1(1-p_1) + m'n'(n'-1)\left(p_2 - p_1^2\right) + n'm'(m'-1)\left(p_3 - p_1^2\right)}{nm(n+m+1)}.$$

Divide the numerator and denominator of the above equation both by n^2m,

$$\left(\frac{\sigma'}{\sigma}\right)^2 = 12\frac{\frac{m'}{m}\frac{n'}{n}\frac{1}{n}p_1(1-p_1) + \frac{m'}{m}\frac{n'}{n}(\frac{n'}{n} - \frac{1}{n})(p_2 - p_1^2)\frac{n'}{n}\frac{m'}{m}(\frac{m'}{n} - \frac{1}{n})(p_3 - p_1^2)}{1 + (m/n) + (1/n)}.$$

Replacing m'/n as $(m'/m)(m/n)$,

$$\left(\frac{\sigma'}{\sigma}\right)^2 = 12\frac{\frac{m'}{m}\frac{n'}{n}\frac{1}{n}p_1(1-p_1) + \frac{m'}{m}\frac{n'}{n}(\frac{n'}{n} - \frac{1}{n})(p_2 - p_1^2)\frac{n'}{n}\frac{m'}{m}(\frac{m'}{m}\frac{m}{n} - \frac{1}{n})(p_3 - p_1^2)}{1 + (m/n) + (1/n)}.$$

Following limitations (A.11), (A.12), and (A.13), and recall that we have shown that $\lim_{n,m\to\infty} m/n \to \lambda$,

$$\begin{aligned}
&\lim_{n,m\to\infty}\left(\frac{\sigma'}{\sigma}\right)^2 \to 12\frac{0 + \lambda_x^2\lambda_y(p_2 - p_1^2) + \lambda_x^2\lambda_y\lambda'(p_3 - p_1^2)}{1+\lambda} \\
\implies &\lim_{n,m\to\infty}\left(\frac{\sigma'}{\sigma}\right) \to 12\left(\frac{0 + \lambda_x^2\lambda_y(p_2 - p_1^2) + \lambda_x^2\lambda_y\lambda'(p_3 - p_1^2)}{1+\lambda}\right)^{1/2} := C.
\end{aligned}$$

According to asymptotic result (A.16), $(W(X',Y') - \mu')/\sigma'$ convergences in distribution to $\mathrm{N}(0,1)$. Since σ'/σ convergences to a constant C, then according to Slutsky's theorem,

$$\lim_{n,m\to\infty}(\sigma'/\sigma)(W(X',Y') - \mu')/\sigma' \to^d N(0,1) \tag{A.20}$$

Next, we consider other terms in Equation (A.19). Recall that μ', μ and σ are defined as

$$\mu' = m'n'p_1,\ \mu = nm/2,\ \sigma = \{nm(n+m+1)/12\}^{1/2}.$$

We can therefore rewrite $a_1(n,m,n',m')$ and $a_2(n,m,n',m')$ in Equation (A.19) as

$$\begin{aligned}
a_1(n,m,n',m') &= (\mu' - \mu + nm - n'm')/\sigma \\
&= (m'n'p_1 - nm/2 + nm - n'm')/\{nm(n+m+1)/12\}^{1/2} \\
&= \{(n'/n)(m'/m)(p_1 - 1) + 1/2\}/\{(n+m+1)/(12nm)\}^{1/2},
\end{aligned} \tag{A.21}$$

and

$$\begin{aligned} a_2(n,m,n',m') &= (\mu' - \mu)/\sigma \\ &= (m'n'p_1 - nm/2)/\{nm(n+m+1)/12\}^{1/2} \\ &= \{(m'/m)(n'/n)p_1 - 1/2\}/\{(n+m+1)/(12nm)\}^{1/2}, \end{aligned} \tag{A.22}$$

respectively. Then, according to limitations (A.12) and (A.13) such that $\lim_{n,m\to\infty} n'/n \to \lambda_x$ and $\lim_{n,m\to\infty} m'/m \to \lambda_y$, respectively, we note that for the numerator of Equation (A.21),

$$\lim_{n,m\to\infty}\{(n'/n)(m'/m)(p_1-1)+1/2\} \to \lambda_x\lambda_y(p_1-1)+1/2, \tag{A.23}$$

and for the numerator of Equation (A.22)

$$\lim_{n,m\to\infty}\{(m'/m)(n'/n)p_1 - 1/2\} \to \lambda_x\lambda_y p_1 - 1/2. \tag{A.24}$$

Note that the denominators of both Equation (A.21) and (A.22)

$$\lim_{n,m\to\infty}\{(n+m+1)/(12nm)\}^{1/2} \to 0^+. \tag{A.25}$$

Combining Equation (A.23) and (A.25),

$$\lim_{n,m\to\infty} a_1(n,m,n'm') \to +\infty \text{ if } \lambda_x\lambda_y(p_1-1)+1/2 > 0, \tag{A.26}$$

$$\lim_{n,m\to\infty} a_1(n,m,n'm') \to -\infty \text{ if } \lambda_x\lambda_y(p_1-1)+1/2 < 0. \tag{A.27}$$

Similarly, combining Equation (A.24) and (A.25),

$$\lim_{n,m\to\infty} a_2(n,m,n'm') \to +\infty \text{ if } \lambda_x\lambda_y p_1 - 1/2 > 0, \tag{A.28}$$

$$\lim_{n,m\to\infty} a_2(n,m,n'm') \to -\infty \text{ if } \lambda_x\lambda_y p_1 - 1/2 < 0. \tag{A.29}$$

Recall that we assume in Equation (A.14) that $\lambda_x\lambda_y(p_1-1)+1/2 \neq 0$ and $\lambda_x\lambda_y p_1 - 1/2 \neq 0$. Hence, we can categorise all possible interactions among $\lambda_x, \lambda_y, p_1$ into the following 4 cases:

$$\begin{aligned} \text{Case I} &: \lambda_x\lambda_y(p_1-1)+1/2 < 0 \text{ and } \lambda_x\lambda_y p_1 - 1/2 < 0, \\ \text{Case II} &: \lambda_x\lambda_y(p_1-1)+1/2 > 0 \text{ and } \lambda_x\lambda_y p_1 - 1/2 < 0, \\ \text{Case III} &: \lambda_x\lambda_y(p_1-1)+1/2 < 0 \text{ and } \lambda_x\lambda_y p_1 - 1/2 > 0, \\ \text{Case IV} &: \lambda_x\lambda_y(p_1-1)+1/2 > 0 \text{ and } \lambda_x\lambda_y p_1 - 1/2 > 0. \end{aligned}$$

The following proof obtains the limitations of Equation (A.19) under Case I, II, III and IV, separately.

Case I. Let us consider Case I first, i.e.

$$\lambda_x\lambda_y(p_1-1)+1/2 < 0 \text{ and } \lambda_x\lambda_y p_1 < 1/2.$$

Notice that this case is possible. For example, we can consider $\lambda_x = \lambda_y = 0.9$ and $p_1 = 0.1$, then $\lambda_x\lambda_y(p_1 - 1) = -0.9^3 = -0.729 < -0.5$, and $\lambda_x\lambda_y p_1 = 0.081 < 1/2$.

When Case I holds, according to Equation (A.27) and (A.29),

$$\lim_{n,m\to\infty} a_1(n, m, n'm') \to -\infty, \lim_{n,m\to\infty} a_2(n, m, n'm') \to -\infty.$$

Recall that in Equation (A.20), we have shown that

$$\lim_{n,m\to\infty} (W(X', Y') - \mu')/\sigma' \cdot (\sigma'/\sigma) \to^d C\mathrm{N}(0, 1),$$

with C being a constant. Then

$$\lim_{n,m\to\infty} \Pr\left((\sigma'/\sigma)(W(X', Y') - \mu')/\sigma' + a_1(n, m, n'm') \le \Phi^{-1}(\alpha/2)\right) \to 1$$
$$\lim_{n,m\to\infty} \Pr\left((\sigma'/\sigma)(W(X', Y') - \mu')/\sigma' + a_2(n, m, n'm') \ge \Phi^{-1}(1 - \alpha/2)\right) \to 0.$$

Put the results back into Equation (A.19), we obtain for Case I

$$\begin{aligned}
&\lim_{n,m\to\infty} \Pr\left(\text{Proposed method yields a significant } p\text{-value}\right) \\
&= \lim_{n,m\to\infty} \Pr\left((\sigma'/\sigma)(W(X', Y') - \mu')/\sigma' + a_1(n, m, n'm') \le \Phi^{-1}(\alpha/2)\right) \\
&+ \lim_{n,m\to\infty} \Pr\left((\sigma'/\sigma)(W(X', Y') - \mu')/\sigma' + a_2(n, m, n'm') \ge \Phi^{-1}(1 - \alpha/2)\right) = 1.
\end{aligned}$$

Case II. Let us now consider Case II, i.e.

$$\lambda_x\lambda_y(p_1 - 1) + 1/2 > 0 \text{ and } \lambda_x\lambda_y p_1 < 1/2.$$

Notice that this case is possible. For example, we can consider $\lambda_x = \lambda_y = 0.5$ and $p_1 = 0.5$, then $\lambda_x\lambda_y(p_1 - 1) = -0.5^3 = -0.125 > -0.5$, and $\lambda_x\lambda_y p_1 = 0.0125 < 1/2$.

When Case II holds, according to Equation (A.26) and (A.29)

$$\lim_{n,m\to\infty} a_1(n, m, n'm') \to +\infty, \lim_{n,m\to\infty} a_2(n, m, n'm') \to -\infty.$$

Recall that we have shown in Equation (A.20) that

$$\lim_{n,m\to\infty} (\sigma'/\sigma)(W(X', Y') - \mu')/\sigma' \to^d C\mathrm{N}(0, 1),$$

with C being a constant. Then

$$\lim_{n,m\to\infty} \Pr\left((\sigma'/\sigma)(W(X', Y') - \mu')/\sigma' + a_1(n, m, n'm') \le \Phi^{-1}(\alpha/2)\right) \to 0$$
$$\lim_{n,m\to\infty} \Pr\left((\sigma'/\sigma)(W(X', Y') - \mu')/\sigma' + a_2(n, m, n'm') \ge \Phi^{-1}(1 - \alpha/2)\right) \to 0,$$

Put the results back into Equation (A.19), we obtain for Case II

$$\begin{aligned}
&\lim_{n,m\to\infty} \Pr\left(\text{Proposed method yields a significant } p\text{-value}\right) \\
&= \lim_{n,m\to\infty} \Pr\left((\sigma'/\sigma)(W(X',Y')-\mu')/\sigma' + a_1(n,m,n'm') \le \Phi^{-1}(\alpha/2)\right) \\
&+ \lim_{n,m\to\infty} \Pr\left((\sigma'/\sigma)(W(X',Y')-\mu')/\sigma' + a_2(n,m,n'm') \ge \Phi^{-1}(1-\alpha/2)\right) = 0.
\end{aligned}$$

Case III. Let us now consider case III, i.e.

$$\lambda_x\lambda_y(p_1-1)+1/2<0 \text{ and } \lambda_x\lambda_y p_1 > 1/2.$$

Notice that

$$\begin{aligned}
\lambda_x\lambda_y(p_1-1)+1/2<0 &\iff \lambda_x\lambda_y(1-p_1)-1/2>0 \\
&\iff \lambda_x\lambda_y-\lambda_x\lambda_y p_1 > 1/2 \\
&\iff \lambda_x\lambda_y > \lambda_x\lambda_y p_1 + 1/2 \\
&\iff \lambda_x\lambda_y > 1.
\end{aligned}$$

Hence the Case III is impossible provided $0<\lambda_x,\lambda_y\le 1$.

Case IV. Let us consider Case IV, i.e.

$$\lambda_x\lambda_y(p_1-1)+1/2>0 \text{ and } \lambda_x\lambda_y p_1 > 1/2.$$

Notice that this case is possible. For example, we can consider $p_1=0.9$ and $\lambda_x=\lambda_y=0.9$. Then $\lambda_x\lambda_y(p_1-1)=-0.081>-1/2$, and $\lambda_x\lambda_y p_1 = 0.9^3 = 0.729 > 1/2$.

When Case IV holds, according to Equation (A.26) and (A.28)

$$\lim_{n,m\to\infty} a_1(n,m,n'm') \to +\infty, \lim_{n,m\to\infty} a_2(n,m,n'm') \to +\infty.$$

Recall that we have shown in Equation (A.20) that

$$\lim_{n,m\to\infty} (\sigma'/\sigma)(W(X',Y')-\mu')/\sigma' \to^d C\mathrm{N}(0,1),$$

with C being a constant. We then have

$$\begin{aligned}
&\lim_{n,m\to\infty} \Pr\left((\sigma'/\sigma)(W(X',Y')-\mu')/\sigma' + a_1(n,m,n'm') \le \Phi^{-1}(\alpha/2)\right) \to 0 \\
&\lim_{n,m\to\infty} \Pr\left((\sigma'/\sigma)(W(X',Y')-\mu')/\sigma' + a_2(n,m,n'm') \ge \Phi^{-1}(1-\alpha/2)\right) \to 1,
\end{aligned}$$

Put the results back into Equation (A.19), we obtain for Case IV

$$\begin{aligned}
&\lim_{n,m\to\infty} \Pr\left(\text{Proposed method yields a significant } p\text{-value}\right) \\
&= \lim_{n,m\to\infty} \Pr\left((\sigma'/\sigma)(W(X',Y')-\mu')/\sigma' + a_1(n,m,n'm') \le \Phi^{-1}(\alpha/2)\right) \\
&+ \lim_{n,m\to\infty} \Pr\left((\sigma'/\sigma)(W(X',Y')-\mu')/\sigma' + a_2(n,m,n'm') \ge \Phi^{-1}(1-\alpha/2)\right) = 1.
\end{aligned}$$

Overall, among all possible 4 cases, we have shown Case III is impossible, Case II leads to the rejection probability to 0 when $n, m \to \infty$, and Case I, Case IV both leads to the rejection probability to 1 when $n, m \to \infty$. Thus, if Case II holds, the rejection probability goes to 0 when $n, m \to \infty$, otherwise the probability goes to 1. Notice that the Case II can be written as

$$\lambda_x \lambda_y (1 - p_1) < 1/2, \lambda_x \lambda_y p_1 < 1/2.$$

Hence, we conclude our results. □

B. Proofs of results in Chapter 4

This chapter presents technical results required in Chapter 4. We prove the tight bounds of the Ansari-Bradley test statistic in the presence of missing data, when data are assumed to be distinct.

B.1. Lower bound with single value missing

We begin our proof by considering the case where only one single value in X is not observed. The main result for this section is Proposition B.1.6.

We first prove the following lemma, which considers the implications if a single value $x \in X$ is replaced with another value x^* where the rank of x^* is equal to the rank of x plus or minus one.

Lemma B.1.1. Suppose $X = \{x_1, \ldots, x_n\}$ and $Y = \{y_1, \ldots, y_m\}$ are samples of distinct real values. Let $x \in X$ be any element in X. Denote $X' = X \setminus \{x\}$. Consider any real value x^* that is distinct from all values in $X' \cup Y$, and denote $X_* = \{x^*\} \cup X'$. Then, if $R(x, X \cup Y) = R(x^*, X_* \cup Y) + 1$ we have $x > x^*$, and $1 \geq \sum_{z \in X'} \mathrm{I}(x > z > x^*) \geq 0$. On the other hand, if $R(x, X \cup Y) = R(x^*, X_* \cup Y) - 1$, we have $x^* > x$, and $1 \geq \sum_{z \in X'} \mathrm{I}(x^* > z > x) \geq 0$.

Proof. To start, let us assume $R(x, X \cup Y) = R(x^*, X_* \cup Y) + 1$. According to the definition of rank, $R(x, X \cup Y) = \sum_{z \in X \cup Y} \mathrm{I}(x \geq z) = 1 + \sum_{z \in X' \cup Y} \mathrm{I}(x \geq z)$. Since all values in $X \cup Y$ are distinct, we have $\mathrm{I}(x \geq z) = \mathrm{I}(x > z)$ for any $z \in X' \cup Y$. Hence, $R(x, X \cup Y) = 1 + \sum_{z \in X' \cup Y} \mathrm{I}(x > z)$. Similarly, we can show that $R(x^*, X_* \cup Y) = 1 + \sum_{z \in X' \cup Y} \mathrm{I}(x^* > z)$. Subsequently, since $R(x, X \cup Y) = R(x^*, X_* \cup Y) + 1$, we have

$$\begin{aligned} & \sum_{z \in X' \cup Y} \mathrm{I}(x > z) = \sum_{z \in X' \cup Y} \mathrm{I}(x^* > z) + 1 \\ \implies & x > x^*. \end{aligned}$$

Further, we have

$$\begin{aligned} & \sum_{z \in X'} \mathrm{I}(x > z) + \sum_{z \in Y} \mathrm{I}(x > z) = \sum_{z \in X'} \mathrm{I}(x^* > z) + \sum_{z \in Y} \mathrm{I}(x^* > z) + 1 \\ \implies & \sum_{z \in X'} \mathrm{I}(x > z) - \sum_{z \in X'} \mathrm{I}(x^* > z) = \sum_{z \in Y} \mathrm{I}(x^* > z) - \sum_{z \in Y} \mathrm{I}(x > z) + 1 \\ \implies & \sum_{z \in X'} (\mathrm{I}(x > z) - \mathrm{I}(x^* > z)) = \sum_{z \in Y} (\mathrm{I}(x^* > z) - \mathrm{I}(x > z)) + 1. \end{aligned}$$

Since $x > x^*$, we have

$$\sum_{z \in X'} (\mathrm{I}(x > z) - \mathrm{I}(x^* > z)) \geq 0, \text{ and } \sum_{z \in Y} (\mathrm{I}(x^* > z) - \mathrm{I}(x > z)) \leq 0.$$

Therefore, we have $1 \geq \sum_{z \in X'} (\mathrm{I}(x > z) - \mathrm{I}(x^* > z)) \geq 0$, which is equivalent to

$$1 \geq \sum_{z \in X'} \mathrm{I}(x > z > x^*) \geq 0.$$

Hence, we completes our proof when $R(x, X \cup Y) = R(x^*, X_* \cup Y) + 1$.

For the case when $R(x, X \cup Y) = R(x^*, X_* \cup Y) - 1$, note that we have $R(x^*, X_* \cup Y) = R(x, X \cup Y) + 1$, then using the result we have already proved above, we have $x^* > x$, and $1 \geq \sum_{z \in X'} \mathrm{I}(x^* > z > x) \geq 0$, which completes our proof. □

According to Lemma B.1.1, if $x \in X$ is replaced with another value x^* where the rank of x^* is equal to the rank of x plus one, i.e. $R(x, X \cup Y) = R(x^*, X_* \cup Y) - 1$, then we must have the value of x^* is strictly greater than x, and perhaps less immediately obvious, there exists at most one value $z \in X'$ that is greater than x but smaller than x^*, i.e. $1 \geq \sum_{z \in X'} \mathrm{I}(x^* > z > x) \geq 0$. However, when the rank of x^* is equal to the rank of x minus one, the opposite conclusion can be drawn from Lemma B.1.1.

Applying Lemma B.1.1, we proceed by analysing how the values of the Ansari-Bradley test statistic $T(X, Y)$ changes when $x \in X$ is replaced by x^*, still assuming that the rank of x^* is equal to the rank of x plus or minus one. Additionally, we assume that the ranks of x and x^* are either both greater than $(N+1)/2$, or both smaller than $(N+1)/2$.

Proposition B.1.2. Suppose $X = \{x_1, \ldots, x_n\}$ and $Y = \{y_1, \ldots, y_m\}$ are samples of distinct real values. Let $x \in X$ be any element in X. Denote $X' = X \setminus \{x\}$. Consider any real value x^* that is distinct from all values in $X' \cup Y$, and denote $X_* = \{x^*\} \cup X'$. Assume that $(R(x, X \cup Y) - (N+1)/2)(R(x^*, X_* \cup Y) - (N+1)/2) \geq 0$. Then, if case

$$\text{(I)}: \ |R(x, X \cup Y) - (N+1)/2| = |R(x^*, X_* \cup Y) - (N+1)/2| + 1$$

holds, we have $T(X, Y) \geq T(X_*, Y)$. However, if case

$$\text{(II)}: \ |R(x, X \cup Y) - (N+1)/2| = |R(x^*, X_* \cup Y) - (N+1)/2| - 1$$

holds, we have $T(X, Y) \leq T(X_*, Y)$.

Proof. We first prove the result for case (I). Then, we show that the result for case (II) follows directly from the result for case (I).

Without loss of generality, assume $x = x_1$. According to the definition of the Ansari-Bradley

test statistic, we have

$$T(X,Y) = \sum_{i=1}^{n} \left| R(x_i, X \cup Y) - \frac{1}{2}(N+1) \right|$$
$$= \left| R(x_1, X \cup Y) - \frac{1}{2}(N+1) \right| + \sum_{i=2}^{n} \left| R(x_i, X \cup Y) - \frac{1}{2}(N+1) \right|.$$

By the definition of rank, for any $i \in \{2, \ldots, n\}$, $R(x_i, X \cup Y) = R(x_i, X' \cup Y) + \mathrm{I}(x_i \geq x_1)$. Since $x_i \neq x_1$, we further have $R(x_i, X \cup Y) = R(x_i, X' \cup Y) + \mathrm{I}(x_i > x_1)$. Hence,

$$T(X,Y) = \left| R(x_1, X \cup Y) - \frac{1}{2}(N+1) \right| + \sum_{i=2}^{n} \left| R(x_i, X' \cup Y) + \mathrm{I}(x_i > x_1) - \frac{1}{2}(N+1) \right|$$
$$= \left| R(x_1, X \cup Y) - \frac{1}{2}(N+1) \right| + \sum_{i=2}^{n} \mathrm{I}(x_i > x_1) \left| R(x_i, X' \cup Y) + 1 - \frac{1}{2}(N+1) \right|$$
$$+ \sum_{i=2}^{n} \mathrm{I}(x_i < x_1) \left| R(x_i, X' \cup Y) - \frac{1}{2}(N+1) \right|.$$

Similarly, we can show that

$$T(X_*, Y) = \left| R(x^*, X_* \cup Y) - \frac{1}{2}(N+1) \right| + \sum_{i=2}^{n} \mathrm{I}(x_i > x^*) \left| R(x_i, X' \cup Y) + 1 - \frac{1}{2}(N+1) \right|$$
$$+ \sum_{i=2}^{n} \mathrm{I}(x_i < x^*) \left| R(x_i, X' \cup Y) - \frac{1}{2}(N+1) \right|.$$

For notation convenience, for any $i \in \{2, \ldots, n\}$, let us denote

$$\alpha_i = \left| R(x_i, X' \cup Y) + 1 - \frac{1}{2}(N+1) \right|, \ \beta_i = \left| R(x_i, X' \cup Y) - \frac{1}{2}(N+1) \right|.$$

Then, we have

$$\begin{aligned} & T(X_*, Y) - T(X,Y) \\ &= \left| R(x^*, X_* \cup Y) - \frac{1}{2}(N+1) \right| - \left| R(x_1, X \cup Y) - \frac{1}{2}(N+1) \right| \\ &+ \sum_{i=2}^{n} \left(\mathrm{I}(x_i > x^*) - \mathrm{I}(x_i > x_1) \right) \alpha_i + \sum_{i=2}^{n} (\mathrm{I}(x_i < x^*) - \mathrm{I}(x_i < x_1)) \beta_i. \end{aligned} \tag{B.1}$$

Notice that

$$|R(x_1, X \cup Y) - (N+1)/2| = |R(x^*, X_* \cup Y) - (N+1)/2| + 1.$$

Hence,

$$\begin{aligned} & T(X_*, Y) - T(X,Y) \\ &= -1 + \sum_{i=2}^{n} \left(\mathrm{I}(x_i > x^*) - \mathrm{I}(x_i > x_1) \right) \alpha_i + \sum_{i=2}^{n} (\mathrm{I}(x_i < x^*) - \mathrm{I}(x_i < x_1)) \beta_i. \end{aligned}$$

Since $(R(x_1, X \cup Y) - (N+1)/2)(R(x^*, X_* \cup Y) - (N+1)/2) \geq 0$, we have either the following two cases holds:

$$\text{(i)} : R(x_1, X \cup Y) > R(x^*, X_* \cup Y) \geq \frac{1}{2}(N+1), \text{ and } R(x_1, X \cup Y) = R(x^*, X_* \cup Y) + 1,$$
$$\text{(ii)} : \frac{1}{2}(N+1) \geq R(x^*, X_* \cup Y) > R(x_1, X \cup Y) \text{ and } R(x_1, X \cup Y) = R(x^*, X_* \cup Y) - 1.$$

Suppose case (i) holds, then according to Lemma B.1.1, we have $x_1 > x^*$. Hence, for any $i \in \{2, \ldots, n\}$,

$$\begin{aligned}
&\mathrm{I}(x_i > x^*) - \mathrm{I}(x_i > x_1) = \mathrm{I}(x_i < x^*) - \mathrm{I}(x_i < x_1) = 0, && \text{if } x_i > x_1, \\
&\mathrm{I}(x_i > x^*) - \mathrm{I}(x_i > x_1) = 1,\ \mathrm{I}(x_i < x^*) - \mathrm{I}(x_i < x_1) = -1, && \text{if } x_1 > x_i > x^*, \\
&\mathrm{I}(x_i > x^*) - \mathrm{I}(x_i > x_1) = \mathrm{I}(x_i < x^*) - \mathrm{I}(x_i < x_1) = 0, && \text{if } x_i < x^*.
\end{aligned}$$

Therefore, we have $T(X_*, Y) - T(X, Y) = -1 + \sum_{i=2}^{n} \mathrm{I}(x_1 > x_i > x^*)(\alpha_i - \beta_i)$. According to Lemma B.1.1, we have $1 \geq \sum_{i=2}^{n} \mathrm{I}(x_1 > x_i > x^*) \geq 0$. In other words, there is at most one x_i greater than x^* but smaller than x_1. By definition, we also have that $1 \geq \alpha_i - \beta_i \geq -1$ for any $i \in \{2, \ldots, n\}$. Hence, we have proved $T(X_*, Y) \leq T(X, Y)$ under case (i).

Similarly, suppose case (ii) holds, then according to Lemma B.1.1, we have $x_1 < x^*$. Hence, for any $i \in \{2, \ldots, n\}$, we have

$$\begin{aligned}
&\mathrm{I}(x_i > x^*) - \mathrm{I}(x_i > x_1) = \mathrm{I}(x_i < x^*) - \mathrm{I}(x_i < x_1) = 0, && \text{if } x_i > x^*, \\
&\mathrm{I}(x_i > x^*) - \mathrm{I}(x_i > x_1) = -1,\ \mathrm{I}(x_i < x^*) - \mathrm{I}(x_i < x_1) = 1, && \text{if } x^* > x_i > x_1, \\
&\mathrm{I}(x_i > x^*) - \mathrm{I}(x_i > x_1) = \mathrm{I}(x_i < x^*) - \mathrm{I}(x_i < x_1) = 0, && \text{if } x_i < x_1.
\end{aligned}$$

Therefore, we have $T(X_*, Y) - T(X, Y) = -1 + \sum_{i=2}^{n} \mathrm{I}(x^* > x_i > x_1)(\beta_i - \alpha_i)$. According to Lemma B.1.1, we have $1 \geq \sum_{i=2}^{n} \mathrm{I}(x^* > x_i > x_1) \geq 0$. In other words, there is at most one x_i greater than x_1 but smaller than x^*. By definition, we also have that $1 \geq \beta_i - \alpha_i \geq -1$ for any $i \in \{2, \ldots, n\}$. Hence, we have proved $T(X_*, Y) \leq T(X, Y)$ under case (ii), which completes our proof for the case (I) when

$$|R(x_1, X \cup Y) - (N+1)/2| = |R(x^*, X_* \cup Y) - (N+1)/2| + 1.$$

For the case (II) when

$$|R(x, X \cup Y) - (N+1)/2| = |R(x^*, X_* \cup Y) - (N+1)/2| - 1.$$

Notice that

$$|R(x^*, X_* \cup Y) - (N+1)/2| = |R(x, X \cup Y) - (N+1)/2| + 1.$$

Then using the result for the case (I), we have $T(X_*, Y) \geq T(X, Y)$, which completes our proof. □

Under the assumption that the ranks of x and x^* are either both greater than $(N+1)/2$, or both smaller than $(N+1)/2$, Proposition B.1.2 asserts that when the absolute distance between

the rank of the replaced value x^* and $(N+1)/2$ is one plus the absolute distance between the rank of original value x and $(N+1)/2$, i.e. $|R(x^*, X_* \cup Y) - (N+1)/2| = |R(x, X \cup Y) - (N+1)/2| + 1$, the Ansari-Bradley test statistic is not decreasing.

Proposition B.1.2 appears to suggest that larger absolute distance between the rank of the replaced value x^* and $(N+1)/2$ leads to larger value of the Ansari-Bradley test statistic. We shall prove this argument in Proposition B.1.4. However, before doing so, we introduce the following lemma, which shows that when considering the values of the Ansari-Bradley test statistic after replacing a set of values, only the ranks of original values and replaced values matter.

Lemma B.1.3. Suppose $X = \{x_1, \ldots, x_n\}$ and $Y = \{y_1, \ldots, y_m\}$ are samples of distinct real values. Let $\tilde{X} = \{\tilde{x}_1, \ldots, \tilde{x}_{n-n'}\}$ be a non-empty subset of X. Denote $X' = X \setminus \tilde{X}$. Consider a set $\hat{X} = \{x_1^*, \ldots, x_{n-n'}^*\}$ of real values, denote $X_* = \hat{X} \cup X'$, and assume that $(X_*, Y) \in \mathcal{D}(n, m)$. Then, if

$$\{R(\tilde{x}_1, X \cup Y), \ldots, R(\tilde{x}_{n-n'}, X \cup Y)\} = \left\{R(x_1^*, X_* \cup Y), \ldots, R(x_{n-n'}^*, X_* \cup Y)\right\},$$

we have $T(X, Y) = T(X_*, Y)$.

Proof. To start, let us denote $X' = \{x_1', \ldots, x_{n'}'\}$, $Z = X \cup Y$, and $Z^* = X_* \cup Y$. Then, according to the definition of the Ansari-Bradley test statistic,

$$\begin{aligned} T(X, Y) &= \sum_{i=1}^{n} \left| R(x_i, Z) - \frac{1}{2}(N+1) \right| \\ &= \sum_{i=1}^{n'} \left| R(x_i', Z) - \frac{1}{2}(N+1) \right| + \sum_{i=1}^{n-n'} \left| R(\tilde{x}_i, Z) - \frac{1}{2}(N+1) \right|. \end{aligned}$$

Similarly, we have

$$T(X_*, Y) = \sum_{i=1}^{n'} \left| R(x_i', Z^*) - \frac{1}{2}(N+1) \right| + \sum_{i=1}^{n-n'} \left| R(x_i^*, Z^*) - \frac{1}{2}(N+1) \right|.$$

Since $\{R(\tilde{x}_1, Z), \ldots, R(\tilde{x}_{n-n'}, Z)\} = \left\{R(x_1^*, Z^*), \ldots, R(x_{n-n'}^*, Z^*)\right\}$, we have

$$\sum_{i=1}^{n-n'} \left| R(\tilde{x}_i, Z) - \frac{1}{2}(N+1) \right| = \sum_{i=1}^{n-n'} \left| R(x_i^*, Z^*) - \frac{1}{2}(N+1) \right|.$$

Thus, in order to prove $T(X, Y) = T(X_*, Y)$, it is sufficient to show that

$$R(x_i', Z) = R(x_i', Z^*), \text{ for any } i \in \{1, \ldots, n'\}.$$

Denote $Z' = X' \cup Y$. Let $x \in X'$ be any value in X', and denote $R(x, Z') = k'$, then $R(x, Z)$ depends only on $\{R(\tilde{x}_1, Z), \ldots, R(\tilde{x}_{n-n'}, Z)\}$, such that

$$R(x, Z) = k \iff \sum_{i=1}^{n-n'} \mathrm{I}\left(R(\tilde{x}_i, Z) < k\right) = k - k'.$$

For the " $\Longrightarrow$ ", note that $R(x, Z') = k'$ implies k' elements in Z' smaller than x, and $R(x, Z)$ implies k elements in Z smaller than x. Then there are $k - k'$ elements in X' smaller than x, hence the result. For the " $\Longleftarrow$ ", let us assume $\tilde{x}_1 < \ldots < \tilde{x}_{n-n'}$, then

$$\begin{aligned}
&\sum_{i=1}^{n-n'} \mathrm{I}\left(R(\tilde{x}_i, Z) < k\right) = k - k' \\
\Longleftrightarrow\ &\{R(\tilde{x}_1, Z), \ldots, R(\tilde{x}_{k-k'}, Z)\} < k, \text{ and } \{R(\tilde{x}_{k-k'+1}, Z), \ldots, R(\tilde{x}_{n-n'}, Z)\} \geq k \\
\Longleftrightarrow\ &\{R(\tilde{x}_1, Z'), \ldots, R(\tilde{x}_{k-k'}, Z')\} < k', \text{ and } \{R(\tilde{x}_{k-k'+1}, Z'), \ldots, R(\tilde{x}_{n-n'}, Z')\} \geq k'.
\end{aligned}$$

Since $R(x, Z') = k'$, there are $k - k'$ elements in X' smaller than x. Hence, we have proven the " $\Longleftarrow$ ".

Similarly, we have

$$R(x, Z^*) = k \iff \sum_{i=1}^{n-n'} \mathrm{I}\left(R(x_i^*, Z^*) < k\right) = k - k'.$$

Since $\{R(\tilde{x}_1, Z), \ldots, R(\tilde{x}_{n-n'}, Z)\} = \left\{R(x_1^*, Z^*), \ldots, R(x_{n-n'}^*, Z^*)\right\}$, we have

$$R(x, Z) = R(x, Z^*),$$

which completes the proof. □

Lemma B.1.3 stresses that when considering the values of Ansari-Bradley test statistic, only the ranks of the values matter. With Lemma B.1.3, we now proceed to prove the following proposition which generalises Proposition B.1.2. We continue to assume that the ranks of x and x^* are either both greater than $(N+1)/2$, or both smaller than $(N+1)/2$. That is, $(R(x, X \cup Y) - (N+1)/2)(R(x^*, X_* \cup Y) - (N+1)/2) \geq 0$. However, unlike Proposition B.1.2, the following proposition does not require the rank of x^* to differ from that of x by exactly one.

Proposition B.1.4. Suppose $X = \{x_1, \ldots, x_n\}$ and $Y = \{y_1, \ldots, y_m\}$ are samples of distinct real values. Let $\tilde{X} \subset X$ be a non-empty subset of X. Denote $X' = X \setminus \tilde{X}$. Consider any real value x^* that is distinct to all values in $X' \cup Y$, and denote $X_* = \{x^*\} \cup X'$. Assume that $(R(x, X \cup Y) - (N+1)/2)(R(x^*, X_* \cup Y) - (N+1)/2) \geq 0$. Then, if case

$$\text{(I)}: \ |R(x, X \cup Y) - (N+1)/2| \geq |R(x^*, X_* \cup Y) - (N+1)/2|$$

holds, we have $T(X, Y) \geq T(X_*, Y)$. However, if case

$$\text{(II)}: |R(x, X \cup Y) - (N+1)/2| \leq |R(x^*, X_* \cup Y) - (N+1)/2|$$

holds, we have $T(X, Y) \leq T(X_*, Y)$.

Proof. We first prove the result for case (I). Then, we show that the result for case (II) follows directly from the result for case (I).

Since $(R(x, X \cup Y) - (N+1)/2)(R(x^*, X_* \cup Y) - (N+1)/2) \geq 0$, then we have one of the

following two cases holds:

$$\text{(i) } R(x, X \cup Y) \geq R(x^*, X_* \cup Y) \geq \frac{1}{2}(N+1),$$
$$\text{or (ii) } R(x, X \cup Y) \leq R(x^*, X_* \cup Y) \leq \frac{1}{2}(N+1).$$

Note that if $R(x, X \cup Y) = R(x^*, X_* \cup Y)$, then we have $T(X, Y) = T(X_*, Y)$ according to Lemma B.1.3. Hence, we only need to consider the cases when $R(x^*, X_* \cup Y) \neq R(x, X \cup Y)$.

Suppose case (i) holds. That is, assume that $R(x, X \cup Y) > R(x^*, X_* \cup Y) \geq \frac{1}{2}(N+1)$. Then, without loss of generality, assume that $R(x, X \cup Y) = R(x^*, X_* \cup Y) + a$, where $a \in \mathbb{N}^+$ is a positive integer. Let $x_1^*, \ldots, x_a^*$ be real numbers distinct to samples in $X' \cup Y$ such that

$$R(x_k^*, \{x_k^*\} \cup X' \cup Y) = R(x, X \cup Y) - k, \text{for any } k \in \{1, \ldots, a\}.$$

Denote $x = x_0^*$. Then, for any $k \in \{0, \ldots, a\}$, we have

$$R(x_k^*, \{x_k^*\} \cup X' \cup Y) \geq \frac{1}{2}(N+1),$$

and for any $k \in \{1, \ldots, a\}$,

$$R(x_k^*, \{x_k^*\} \cup X' \cup Y) = R(x_{k-1}^*, \{x_{k-1}^*\} \cup X' \cup Y) - 1.$$

Subsequently, by applying Proposition B.1.2 between each $\{x_k^*\} \cup X', Y$ and $\{x_{k-1}^*\} \cup X', Y$ for $k \in \{1, \ldots, a\}$, it follows that

$$T(\{x_0^*\} \cup X', Y) \geq \cdots \geq T(\{x_a^*\} \cup X', Y).$$

Since $x = x_0^*$, we have $T(\{x_0^*\} \cup X', Y) = T(X, Y)$. Further, since

$$R(x_a^*, \{x_a^*\} \cup X' \cup Y) = R(x, X \cup Y) - a = R(x^*, X_* \cup Y),$$

then according to Lemma B.1.3, we have $T(\{x_a^*\} \cup X', Y) = T(X_*, Y)$. Therefore,

$$T(X, Y) = T(\{x_0^*\} \cup X', Y) \geq \cdots \geq T(\{x_a^*\} \cup X', Y) = T(X_*, Y).$$

which proves our result when case (i) holds. Case (ii) can be proved similarly, thus it is omitted here.

Subsequently, for case (II), note that

$$|R(x, X \cup Y) - (N+1)/2| \leq |R(x^*, X_* \cup Y) - (N+1)/2|$$

is equivalent to

$$|R(x^*, X_* \cup Y) - (N+1)/2| \geq |R(x, X \cup Y) - (N+1)/2|.$$

Then, by applying the result for case (I), we have $T(X_*, Y) \geq T(X, Y)$, which completes our

proof. □

Proposition B.1.4 asserts that when the absolute distance between the rank of the replaced value x^* and $(N+1)/2$ is larger than the absolute distance between the rank of the original value x and $(N+1)/2$, i.e. $|R(x^*, X_* \cup Y) - (N+1)/2| \geq |R(x, X \cup Y) - (N+1)/2|$, the values of Ansari-Bradley test statistic with x^* is greater or equal to that with x. This result generalises Proposition B.1.2, by dropping the assumption that the rank of x^* differs from that of x by exactly one. However, Proposition B.1.4 still assumes that the ranks of x and x^* are either both greater than $(N+1)/2$, or both smaller than $(N+1)/2$.

In order to obtain more general results regarding the values of the Ansari-Bradley test statistic with one single value missing. We first prove the following lemma. Unlike Proposition B.1.2 and Proposition B.1.4, the following lemma does not assume the ranks of x and x^* are both larger or both smaller than $(N+1)/2$.

Lemma B.1.5. Suppose $X = \{x_1, \ldots, x_n\}$ and $Y = \{y_1, \ldots, y_m\}$ are samples of distinct real values. Denote $N = n + m$ and assume N is even. Suppose $x \in X$ is any value in X and denote $X' = X \setminus \{x\}$. Consider any real value x^* that is distinct to all values in $X' \cup Y$, and denote $X_* = \{x^*\} \cup X'$. Then, if

$$|R(x^*, X_* \cup Y) - (N+1)/2| = |R(x, X \cup Y) - (N+1)/2| = 1/2$$

we have $T(X, Y) = T(X_*, Y)$.

Proof. To start, if $R(x^*, X_* \cup Y) = R(x, X \cup Y)$. Then, we have our desired result $T(X, Y) = T(X_*, Y)$ according to Lemma B.1.3. However, if $R(x^*, X_* \cup Y) \neq R(x, X \cup Y)$, since

$$|R(x^*, X_* \cup Y) - (N+1)/2| = |R(x, X \cup Y) - (N+1)/2| = 1/2,$$

we have one of the following two cases is true

$$\begin{aligned} &\text{(i) } R(x^*, X_* \cup Y) = (N+1)/2 - 1/2,\ R(x, X \cup Y) = (N+1)/2 + 1/2, \\ \text{or } &\text{(ii) } R(x^*, X_* \cup Y) = (N+1)/2 + 1/2,\ R(x, X \cup Y) = (N+1)/2 - 1/2. \end{aligned}$$

In the following, we consider only case (i). Once case (i) has been established, observe that case (ii) is equivalent to

$$R(x, X \cup Y) = (N+1)/2 - 1/2,\ R(x^*, X_* \cup Y) = (N+1)/2 + 1/2.$$

Hence, we have $T(X_*, Y) = T(X, Y)$ from case (i).

Without loss of generality, let us assume that $x = x_1$. Then, according to the definition of the Ansari-Bradley test statistic,

$$\begin{aligned} T(X, Y) &= \sum_{i=1}^{n} \left| R(x_i, X \cup Y) - \frac{1}{2}(N+1) \right| \\ &= \left| R(x_1, X \cup Y) - \frac{1}{2}(N+1) \right| + \sum_{i=2}^{n} \left| R(x_i, X \cup Y) - \frac{1}{2}(N+1) \right|. \end{aligned}$$

By the definition of rank, we have $R(x_i, X \cup Y) = R(x_i, X' \cup Y) + \mathrm{I}(x_i \geq x_1)$ for any $i \in \{2, \ldots, n\}$. Since $x_i \neq x_1$, we further have $R(x_i, X \cup Y) = R(x_i, X' \cup Y) + \mathrm{I}(x_i > x_1)$. Hence,

$$\begin{aligned} T(X, Y) &= \left| R(x_1, X \cup Y) - \frac{1}{2}(N+1) \right| + \sum_{i=2}^{n} \left| R(x_i, X' \cup Y) + \mathrm{I}(x_i > x_1) - \frac{1}{2}(N+1) \right| \\ &= \left| R(x_1, X \cup Y) - \frac{1}{2}(N+1) \right| + \sum_{i=2}^{n} \mathrm{I}(x_i > x_1) \left| R(x_i, X' \cup Y) + 1 - \frac{1}{2}(N+1) \right| \\ &\quad + \sum_{i=2}^{n} \mathrm{I}(x_i < x_1) \left| R(x_i, X' \cup Y) - \frac{1}{2}(N+1) \right|. \end{aligned}$$

Similarly, we have

$$\begin{aligned} T(X_*, Y) &= \left| R(x^*, X_* \cup Y) - \frac{1}{2}(N+1) \right| + \sum_{i=2}^{n} \mathrm{I}(x_i > x^*) \left| R(x_i, X' \cup Y) + 1 - \frac{1}{2}(N+1) \right| \\ &\quad + \sum_{i=2}^{n} \mathrm{I}(x_i < x^*) \left| R(x_i, X' \cup Y) - \frac{1}{2}(N+1) \right|. \end{aligned}$$

For notation convenience, for any $i \in \{2, \ldots, n\}$, let us denote

$$\alpha_i = \left| R(x_i, X' \cup Y) + 1 - \frac{1}{2}(N+1) \right|, \text{ and } \beta_i = \left| R(x_i, X' \cup Y) - \frac{1}{2}(N+1) \right|. \tag{B.2}$$

Subsequently,

$$\begin{aligned} &T(X_*, Y) - T(X, Y) \\ &= \left| R(x^*, X_* \cup Y) - \frac{1}{2}(N+1) \right| - \left| R(x_1, X \cup Y) - \frac{1}{2}(N+1) \right| \\ &+ \sum_{i=2}^{n} \left(\mathrm{I}(x_i > x^*) - \mathrm{I}(x_i > x_1) \right) \alpha_i + \sum_{i=2}^{n} (\mathrm{I}(x_i < x^*) - \mathrm{I}(x_i < x_1)) \beta_i. \end{aligned} \tag{B.3}$$

Since case (i) is true, we have

$$R(x^*, X_* \cup Y) = \frac{1}{2}(N+1) - \frac{1}{2}, \; R(x_1, X \cup Y) = \frac{1}{2}(N+1) + \frac{1}{2}.$$

Thus,

$$\left| R(x^*, X_* \cup Y) - \frac{1}{2}(N+1) \right| - \left| R(x_1, X \cup Y) - \frac{1}{2}(N+1) \right| = 0.$$

Hence,

$$T(X_*, Y) - T(X, Y) = \sum_{i=2}^{n} \left(\mathrm{I}(x_i > x^*) - \mathrm{I}(x_i > x_1) \right) \alpha_i + \sum_{i=2}^{n} (\mathrm{I}(x_i < x^*) - \mathrm{I}(x_i < x_1)) \beta_i$$

following Equation (B.3). Since $R(x_1, X \cup Y) = R(x^*, X_* \cup Y) + 1$, we have $x_1 > x^*$ according

to Lemma B.1.1. Hence, for any $i \in \{2, \ldots, n\}$,

$$\begin{aligned}
&\mathrm{I}(x_i > x^*) - \mathrm{I}(x_i > x_1) = \mathrm{I}(x_i < x^*) - \mathrm{I}(x_i < x_1) = 0, && \text{if } x_i > x_1 \\
&\mathrm{I}(x_i > x^*) - \mathrm{I}(x_i > x_1) = 1,\ \mathrm{I}(x_i < x^*) - \mathrm{I}(x_i < x_1) = -1, && \text{if } x_1 > x_i > x^* \\
&\mathrm{I}(x_i > x^*) - \mathrm{I}(x_i > x_1) = \mathrm{I}(x_i < x^*) - \mathrm{I}(x_i < x_1) = 0, && \text{if } x_i < x^*.
\end{aligned}$$

This gives us

$$T(X_*, Y) - T(X, Y) = \sum_{i=2}^{n} \mathrm{I}(x_1 > x_i > x^*)(\alpha_i - \beta_i).$$

According to Lemma B.1.1, we further have

$$1 \geq \sum_{i=2}^{n} \mathrm{I}(x_1 > x_i > x^*) \geq 0.$$

If $\sum_{i=2}^{n} \mathrm{I}(x_1 > x_i > x^*) = 0$, then we have

$$T(X_*, Y) - T(X, Y) = 0,$$

which gives us the desired result.

However, if $\sum_{i=2}^{n} \mathrm{I}(x_1 > x_i > x^*) = 1$, then there exists one and only one value $x_j \in X'$ such that $x_1 > x_j > x^*$. Hence, we have

$$T(X_*, Y) - T(X, Y) = \alpha_j - \beta_j \tag{B.4}$$

Since $x_1 > x_j$, we have $R(x_1, X \cup Y) > R(x_j, X \cup Y)$. Recall that $R(x_1, X \cup Y) = (N+1)/2 + 1/2$. Hence,

$$\frac{1}{2}(N+1) + \frac{1}{2} > R(x_j, X \cup Y).$$

Additionally, since $x^* < x_j$, it follows that

$$\frac{1}{2}(N+1) - \frac{1}{2} = R(x^*, X_* \cup Y) \leq R(x_j, X \cup Y).$$

Thus, we have

$$R(x_j, X \cup Y) = \frac{1}{2}(N+1) - \frac{1}{2}.$$

Further, notice that

$$R(x_j, X \cup Y) = R(x_j, X' \cup Y) + \mathrm{I}(x_j > x_1).$$

Since $x_1 > x_j > x^*$, we have $\mathrm{I}(x_j > x_1) = 0$, which means

$$R(x_j, X' \cup Y) = R(x_j, X \cup Y) = \frac{1}{2}(N+1) - \frac{1}{2}.$$

Then, by the definition of α_j and β_j in Equation (B.2), we have $\alpha_j = \beta_j = 1/2$, which proves the result $T(X_*, Y) = T(X, Y)$ according to Equation (B.4). This completes our proof. □

We are now ready to prove the final result concerning the lower bounds of the Ansari-Bradley test statistic when one single value in X is missing. This result is proved by applying mainly Proposition B.1.4, and Lemma B.1.5.

Proposition B.1.6. Suppose $X = \{x_1, \ldots, x_n\}$ and $Y = \{y_1, \ldots, y_m\}$ are samples of distinct real values. Denote $N = n + m$. Suppose $x \in X$ is a value in X and denote $X' = X \setminus \{x\}$. Consider a real value x^* that is distinct to all values in $X' \cup Y$, and denote $X_* = \{x^*\} \cup X'$. Then, if

$$\left| R(x^*, X_* \cup Y) - \frac{1}{2}(N+1) \right| = \begin{cases} 0, & \text{when} \quad N \text{ is odd,} \\ 1/2, & \text{when} \quad N \text{ is even,} \end{cases}$$

we have $T(X, Y) \geq T(X_*, Y)$.

Proof. We first prove the case when N is odd. In other words, we show that when $R(x^*, X_* \cup Y) = (N+1)/2$, we have $T(X, Y) \geq T(X_*, Y)$.

Suppose $R(x, X \cup Y) = (N+1)/2$, then we have $R(x^*, X_* \cup Y) = R(x, X \cup Y)$. According to Lemma B.1.3, we have $T(X, Y) = T(X_*, Y)$, which proves our result. However, if $R(x, X \cup Y) > (N+1)/2$, or $R(x, X \cup Y) < (N+1)/2$, then we have $T(X, Y) \geq T(X_*, Y)$ according to Proposition B.1.4. Hence, we have shown that we always have $T(X, Y) \geq T(X_*, Y)$ for the case when N is odd.

We now consider the case when N is even. That is, if $|R(x^*, X_* \cup Y) - (N+1)/2| = 1/2$, we have $T(X, Y) \geq T(X_*, Y)$.

Let x_1^* and x_2^* be values that are distinct to data in $X' \cup Y$, and suppose that

$$R(x_1^*, \{x_1^*\} \cup X' \cup Y) = \frac{1}{2}(N+1) + \frac{1}{2}, \text{ and } R(x_2^*, \{x_2^*\} \cup X' \cup Y) = \frac{1}{2}(N+1) - \frac{1}{2}.$$

Then, according to Lemma B.1.5, we have

$$T(X_*, Y) = T(\{x_1^*\} \cup X', Y) = T(\{x_2^*\} \cup X', Y).$$

Suppose $R(x, X \cup Y) = (N+1)/2 - 1/2$, or $R(x, X \cup Y) = (N+1)/2 + 1/2$. By applying Lemma B.1.5 directly, we have $T(X, Y) = T(X_*, Y)$.

However, if $R(x, X \cup Y) > (N+1)/2 + 1/2$, then according to Proposition B.1.4,

$$T(X, Y) \geq T(\{x_1^*\} \cup X', Y) = T(X_*, Y).$$

Finally, if $R(x, X \cup Y) < (N+1)/2 - 1/2$, then by applying Proposition B.1.4 again,

$$T(X, Y) \geq T(\{x_2^*\} \cup X', Y) = T(X_*, Y).$$

This completes our proof. □

B.2. Lower bound with multiple missing values in X

In this section, we generalise the results in Proposition B.1.6 to the cases where multiple values in X can be missing. The main results for this section is Theorem B.2.12.

We start by proving the following result:

Lemma B.2.1. Suppose $X = \{x_1, \ldots, x_n\}$ and $Y = \{y_1, \ldots, y_m\}$ are samples of distinct real values. Let $\tilde{X} = \{\tilde{x}_1, \ldots, \tilde{x}_{n-n'}\}$ be a non-empty subset of X, and denote $X' = X \setminus \tilde{X}$. Suppose $\hat{X} = \{x_1^*, \ldots, x_{n-n'}^*\}$, and denote $X_* = \hat{X} \cup X'_*$. Assume that all values in $X_* \cup X'$ are distinct. Subsequently, for any $x \in X'$, if

$$\begin{aligned} R(x, X \cup Y) &> \max\{R(\tilde{x}_1, X \cup Y), \ldots, R(\tilde{x}_{n-n'}, X \cup Y)\}, \\ \text{and } R(x, X \cup Y) &> \max\{R(x_1^*, X_* \cup Y), \ldots, R(x_{n-n'}^*, X_* \cup Y)\}, \end{aligned}$$

we have $R(x, X_* \cup Y) = R(x, X \cup Y)$.

Proof. First, we show that $x > \max \tilde{X}$ and $x > \max \hat{X}$. The inequality $x > \max \tilde{X}$ comes directly from the assumption that

$$R(x, X \cup Y) > \max\{R(\tilde{x}_1, X \cup Y), \ldots, R(\tilde{x}_{n-n'}, X \cup Y)\}.$$

The inequality $x > \max \hat{X}$ is not immediately obvious, so we prove it by contradiction. Suppose instead that $\max \hat{X} > x$. Without loss of generality, assume $x_1^* = \max \hat{X} > x$. Subsequently, according to the definition of rank,

$$\begin{aligned} R(x_1^*, X_* \cup Y) &= \sum_{i=1}^{n-n'} \mathrm{I}\,(x_1^* \geq x_i^*) + \sum_{z \in X' \cup Y} \mathrm{I}\,(x_1^* \geq z) \\ &= n - n' + \sum_{z \in X' \cup Y} \mathrm{I}\,(x_1^* \geq z) \\ &\geq n - n' + \sum_{z \in X' \cup Y} \mathrm{I}\,(x \geq z) \\ &= \sum_{i=1}^{n-n'} \mathrm{I}\,(x \geq \tilde{x}_i) + \sum_{z \in X' \cup Y} \mathrm{I}\,(x \geq z) = R(x, X \cup Y), \end{aligned}$$

which contracts the assumption that

$$R(x, X \cup Y) > \max\{R(x_1^*, X_* \cup Y), \ldots, R(x_{n-n'}^*, X_* \cup Y)\}.$$

Therefore, it must hold that $x > \max X_*$.

Now we prove that $R(x, X_* \cup Y) = R(x, X \cup Y)$ using the above conclusions. Note that by the definition of rank,

$$\begin{aligned} R(x, X \cup Y) &= R(x, X' \cup Y) + \sum_{i=1}^{n-n'} \mathrm{I}\,(x \geq \tilde{x}_i) \\ \implies R(x, X \cup Y) &= R(x, X' \cup Y) + n - n', \end{aligned}$$

where the " $\implies$ " holds because $x > \max \tilde{X}$. Similarly,

$$\begin{aligned} R(x, X_* \cup Y) &= R(x, X' \cup Y) + \sum_{i=1}^{n-n'} \mathrm{I}\left(x \geq x_i^*\right), \\ \implies R(x, X_* \cup Y) &= R(x, X' \cup Y) + n - n', \end{aligned}$$

where the " $\implies$ " holds because $x > \max X_*$. Hence, we have shown our desired result $R(x, X_* \cup Y) = R(x, X \cup Y)$. □

The following lemma provides the same conclusion as Lemma B.2.1, but with different conditions.

Lemma B.2.2. Suppose $X = \{x_1, \ldots, x_n\}$ and $Y = \{y_1, \ldots, y_m\}$ are samples of distinct real values. Let $\tilde{X} = \{\tilde{x}_1, \ldots, \tilde{x}_{n-n'}\}$ be a non-empty subset of X, and denote $X' = X \setminus \tilde{X}$. Suppose $\hat{X} = \{x_1^*, \ldots, x_{n-n'}^*\}$, and denote $X_* = \hat{X} \cup X'_*$. Assume that all values in $X_* \cup X'$ are distinct. Subsequently, for any $x \in X'$, if

$$\begin{aligned} R(x, X \cup Y) &< \min\{R(\tilde{x}_1, X \cup Y), \ldots, R(\tilde{x}_{n-n'}, X \cup Y)\}, \\ \text{and } R(x, X \cup Y) &< \min\{R(x_1^*, X_* \cup Y), \ldots, R(x_{n-n'}^*, X_* \cup Y)\}, \end{aligned}$$

we have $R(x, X_* \cup Y) = R(x, X \cup Y)$.

Proof. The proof is similar to the proof for Lemma B.2.2. First, we show that $x < \min \tilde{X}$ and $x < \min \hat{X}$. The first inequality $x < \min \tilde{X}$ comes directly from the assumption that

$$R(x, X \cup Y) < \min\{R(\tilde{x}_1, X \cup Y), \ldots, R(\tilde{x}_{n-n'}, X \cup Y)\},$$

The inequality $x < \min \hat{X}$ is not immediately obvious, so we prove it by contradiction. Suppose instead that $x > \min \hat{X}$. Without loss of generality, assume $x_1^* = \min X_* < x$. Subsequently, according to the definition of rank,

$$\begin{aligned} R(x_1^*, X_* \cup Y) &= \sum_{i=1}^{n-n'} \mathrm{I}\left(x_1^* \geq x_i^*\right) + \sum_{z \in X' \cup Y} \mathrm{I}\left(x_1^* \geq z\right) \\ &= 1 + \sum_{z \in X' \cup Y} \mathrm{I}\left(x_1^* \geq z\right) \\ &\leq 1 + \sum_{z \in X' \cup Y} \mathrm{I}\left(x \geq z\right) \\ &< 0 + \sum_{z \in X' \cup Y} \mathrm{I}\left(x \geq z\right) \\ &= \sum_{i=1}^{n-n'} \mathrm{I}\left(x \geq \tilde{x}_i\right) + \sum_{z \in X' \cup Y} \mathrm{I}\left(x \geq z\right) = R(x, X \cup Y), \end{aligned}$$

which contracts the assumption that

$$R(x, X \cup Y) < \min\{R(x_1^*, X_* \cup Y), \ldots, R(x_{n-n'}^*, X_* \cup Y)\}.$$

Hence, we must have $x < \min X_*$.

Then, we prove $R(x, X_* \cup Y) = R(x, X \cup Y)$ using the above results. According to the definition of rank, we have

$$\begin{aligned} & R(x, X \cup Y) = R(x, X' \cup Y) + \sum_{i=1}^{n-n'} \mathrm{I}\left(x \geq \tilde{x}_i\right) \\ \Longrightarrow\ & R(x, X \cup Y) = R(x, X' \cup Y), \end{aligned}$$

where the " $\Longrightarrow$ " holds because $x < \min \tilde{X}$. Similarly, we have

$$\begin{aligned} & R(x, X_* \cup Y) = R(x, X' \cup Y) + \sum_{i=1}^{n-n'} \mathrm{I}\left(x \geq x_i^*\right) \\ \Longrightarrow\ & R(x, X_* \cup Y) = R(x, X' \cup Y), \end{aligned}$$

where the " $\Longrightarrow$ " holds because $x < \min \hat{X}$. Hence, we have shown our desired result $R(x, X_* \cup Y) = R(x, X \cup Y)$. □

The following lemma provides similar results for Lemma B.2.1 and B.2.2, but is concerning with the rank of x in observed values $R(x, X' \cup Y)$, rather than the rank of x in all values $R(x, X \cup Y)$.

Lemma B.2.3. Suppose $X = \{x_1, \ldots, x_n\}$ and $Y = \{y_1, \ldots, y_m\}$ are samples of distinct real values, X' is a non-empty subset of X, and $|X'| = n' < n$. Let $\hat{X} = \{x_1^*, \ldots, x_{n-n'}^*\}$ be a set of distinct real values, and denote $X_* = \hat{X} \cup X'$. Assume all values in $X_* \cup Y$ are distinct. Subsequently, for any $x \in X'$,

$$\begin{aligned} & R(x, X' \cup Y) < \min\{R(x_1^*, X_* \cup Y), \ldots, R(x_{n-n'}^*, X_* \cup Y)\} \\ \Longleftrightarrow\ & R(x, X_* \cup Y) = R(x, X' \cup Y). \end{aligned}$$

Proof. We first prove the " $\Longrightarrow$ ". Without loss of generality, let us assume that $x_1^* = \min X_*$. By definition, $R(x_1^*, X_* \cup Y) = \sum_{i=1}^{n-n'} \mathrm{I}\left(x_1^* \geq x_i^*\right) + \sum_{z \in X' \cup Y} \mathrm{I}\left(x_1^* \geq z\right)$. Since $x_1^* \notin X' \cup Y$, we further have

$$\begin{aligned} R(x_1^*, X_* \cup Y) &= \sum_{i=1}^{n-n'} \mathrm{I}\left(x_1^* \geq x_i^*\right) + \sum_{z \in X' \cup Y} \mathrm{I}\left(x_1^* > z\right) \\ &= 1 + \sum_{z \in X' \cup Y} \mathrm{I}\left(x_1^* > z\right). \end{aligned}$$

Hence, $\sum_{z \in X' \cup Y} \mathrm{I}\left(x_1^* > z\right) = R(x_1^*, X_* \cup Y) - 1$. For any $x \in X'$, since

$$R(x, X' \cup Y) < \min\{R(x_1^*, X_* \cup Y), \ldots, R(x_{n-n'}^*, X_* \cup Y)\},$$

we have

$$\sum_{z \in X' \cup Y} \mathrm{I}\left(x_1^* > z\right) > R(x, X' \cup Y) - 1 = \sum_{z \in X' \cup Y} \mathrm{I}\left(x \geq z\right) - 1 = \sum_{z \in X' \cup Y} \mathrm{I}\left(x > z\right).$$

Thus, $x_1^* = \min X_* > x$, where $x \in X'$ is any value in X'. Using this result, we have

$$R(x, X_* \cup Y) = \sum_{i=1}^{n-n'} \mathrm{I}\,(x \geq x_i^*) + \sum_{z \in X' \cup Y} \mathrm{I}\,(x \geq z) = \sum_{z \in X' \cup Y} \mathrm{I}\,(x \geq z) = R(x, X' \cup Y),$$

which proves the " $\Longrightarrow$ ".

We now prove the " $\Longleftarrow$ ". For any $x \in X'$, notice that

$$\begin{aligned} R(x, X_* \cup Y) &= \sum_{i=1}^{n-n'} \mathrm{I}\,(x \geq x_i^*) + \sum_{z \in X' \cup Y} \mathrm{I}\,(x \geq z) \\ &= \sum_{i=1}^{n-n'} \mathrm{I}\,(x \geq x_i^*) + R(x, X' \cup Y). \end{aligned}$$

Hence,

$$\begin{aligned} & R(x, X_* \cup Y) = R(x, X' \cup Y) \\ \Longrightarrow & \sum_{i=1}^{n-n'} \mathrm{I}\,(x \geq x_i^*) = 0 \\ \Longrightarrow & x < \min\{x_1^*, \ldots, x_{n-n'}^*\} \\ \Longrightarrow & R(x, X_* \cup Y) < \min\{R(x_1^*, X_* \cup Y), \ldots, R(x_{n-n'}^*, X_* \cup Y)\}. \end{aligned}$$

By the definition of rank, we have $R(x, X' \cup Y) \leq R(x, X_* \cup Y)$. Hence, we have

$$R(x, X' \cup Y) < \min\{R(x_1^*, X_* \cup Y), \ldots, R(x_{n-n'}^*, X_* \cup Y)\},$$

which proves the " $\Longleftarrow$ ". □

The following lemma considers the opposite condition as considered in Lemma B.2.3.

Lemma B.2.4. Suppose $X = \{x_1, \ldots, x_n\}$ and $Y = \{y_1, \ldots, y_m\}$ are samples of distinct real values and $X' \subset X$ is a subset of X, with sample size $|X'| = n' < n$. Let $\hat{X} = \{x_1^*, \ldots, x_{n-n'}^*\}$ be a set of real values, and denote $X_* = \hat{X} \cup X'$. Assume that all values in $X_* \cup Y$ are distinct. Then, for any $x \in X'$,

$$\begin{aligned} & R(x, X' \cup Y) \geq \min\{R(x_1^*, X_* \cup Y), \ldots, R(x_{n-n'}^*, X_* \cup Y)\} \\ \Longleftrightarrow\ & R(x, X_* \cup Y) > R(x, X' \cup Y). \end{aligned} \tag{B.5}$$

Additionally, suppose that the set $R(x_1^*, X_* \cup Y), \ldots, R(x_{n-n'}^*, X_* \cup Y)$ consists of consecutive integers. Then, the following equivalence holds:

$$\begin{aligned} & R(x, X' \cup Y) \geq \min\{R(x_1^*, X_* \cup Y), \ldots, R(x_{n-n'}^*, X_* \cup Y)\} \\ \Longleftrightarrow\ & R(x, X_* \cup Y) = R(x, X' \cup Y) + n - n'. \end{aligned} \tag{B.6}$$

Proof. The " $\Longleftarrow$ " of (B.5) follows directly from Lemma B.2.3. For the " $\Longrightarrow$ ", we have $R(x, X_* \cup Y) \neq R(x, X' \cup Y)$ according to Lemma B.2.3. By the definition of rank, we also have

$R(x, X_* \cup Y) \geq R(x, X' \cup Y)$. Hence, we have $R(x, X_* \cup Y) > R(x, X' \cup Y)$, which proves the " $\Longrightarrow$ ".

We now prove the " $\Longrightarrow$ " of (B.6). For any $x \in X'$ such that

$$R(x, X' \cup Y) \geq \min\{R(x_1^*, X_* \cup Y), \ldots, R(x_{n-n'}^*, X_* \cup Y)\},$$

we have $R(x, X' \cup Y) \geq R(x_1^*, X_* \cup Y)$. By the definition of rank, $R(x, X_* \cup Y) \geq R(x, X' \cup Y)$. Hence, $R(x, X_* \cup Y) \geq R(x_1^*, X_* \cup X' \cup Y)$. Further, since $\{R(x_1^*, X_* \cup Y), \ldots, R(x_{n-n'}^*, X_* \cup Y)\}$ consists of consecutive integers, and all values in $X_* \cup Y$ are distinct, it follows that

$$R(x, X_* \cup Y) > \max\{R(x_1^*, X_* \cup Y), \ldots, R(x_{n-n'}^*, X_* \cup Y)\}.$$

Hence, $x > \max\{x_1^*, \ldots, x_{n-n'}^*\}$. Then, we have

$$\begin{aligned} R(x, X_* \cup Y) &= \sum_{i=1}^{n-n'} \mathrm{I}\,(x \geq x_i^*) + \sum_{z \in X' \cup Y} \mathrm{I}\,(x \geq z) \\ &= n - n' + \sum_{z \in X' \cup Y} \mathrm{I}\,(x \geq z) \\ &= n - n' + R(x, X' \cup Y), \end{aligned}$$

which proves the " $\Longrightarrow$ " of (B.6).

We now prove the " $\Longleftarrow$ " of (B.6). Let $x \in X'$ be any value such that

$$R(x, X_* \cup Y) = R(x, X' \cup Y) + n - n'.$$

By the definition of rank,

$$\begin{aligned} R(x, X_* \cup Y) &= \sum_{i=1}^{n-n'} \mathrm{I}\,(x \geq x_i^*) + \sum_{z \in X' \cup Y} \mathrm{I}\,(x \geq z) \\ &= \sum_{i=1}^{n-n'} \mathrm{I}\,(x > x_i^*) + R(x, X' \cup Y). \end{aligned}$$

Hence,

$$\begin{aligned} & R(x, X_* \cup Y) = n - n' + R(x, X' \cup Y) \\ \Longrightarrow & \sum_{i=1}^{n-n'} \mathrm{I}\,(x > x_i^*) = n - n' \\ \Longrightarrow & x > \max\{x_1^*, \ldots, x_{n-n'}^*\}. \end{aligned}$$

Without loss of generality, let us assume that $x_1^* = \min X_*$. Subsequently, since

$$R(x_1^*, X_* \cup Y) = \sum_{i=1}^{n-n'} \mathrm{I}\,(x_1^* \geq x_i^*) + \sum_{z \in X' \cup Y} \mathrm{I}\,(x_1^* \geq z) = 1 + \sum_{z \in X' \cup Y} \mathrm{I}\,(x_1^* > z)\,,$$

and

$$R(x, X' \cup Y) = \sum_{z \in X' \cup Y} \mathrm{I}\,(x \geq z) = 1 + \sum_{z \in X' \cup Y} \mathrm{I}\,(x > z)\,,$$

we have

$$R(x, X' \cup Y) \geq R(x_1^*, X_* \cup X' \cup Y),$$

which proves the " $\Longleftarrow$ " of (B.6). □

The following lemma shows that if we replace values in X with ranks in the right hand side of Equation (B.7) with values taking ranks in the right hand side of Equation (B.8), the value of the Ansari-Bradley statistic is not changed. Note that we now assume that the total sample size $N = n + m$ is odd.

Lemma B.2.5. Suppose $X = \{x_1, \ldots, x_n\}$ and $Y = \{y_1, \ldots, y_m\}$ are samples of distinct real values. Let $N = n + m$ and assume N is odd. Suppose $X' \subset X$ is a non-empty subset of X such that $|X'| = n'$ is odd. Let $\tilde{X} = X \setminus X'$, and denote $\tilde{X} = \{\tilde{x}_1, \ldots, \tilde{x}_{n-n'}\}$. Consider a set $\hat{X} = \{x_1^*, \ldots, x_{n-n'}^*\}$ of real values, and denote $X_* = \hat{X} \cup X'$. Assume that all values in $X_* \cup Y$ are distinct. Subsequently, if

$$\begin{aligned} &\{R(\tilde{x}_1, X \cup Y), \ldots, R(\tilde{x}_{n-n'}, X \cup Y)\} \\ &= \left\{ -\frac{n-n'}{2} + 1 + \frac{1}{2}(N+1), \ldots, \frac{n-n'}{2} + \frac{1}{2}(N+1) \right\}, \end{aligned} \tag{B.7}$$

and

$$\begin{aligned} &\left\{R(x_1^*, X_* \cup Y), \ldots, R(x_{n-n'}^*, X_* \cup Y)\right\} \\ &= \left\{ -\frac{n-n'}{2} + \frac{1}{2}(N+1), \ldots, \frac{n-n'}{2} - 1 + \frac{1}{2}(N+1) \right\}, \end{aligned} \tag{B.8}$$

we have $T(X, Y) = T(X_*, Y)$.

Proof. To start, let us denote $N' = (N+1)/2 - (n-n')/2$. By the definition of the Ansari-Bradley test statistic,

$$\begin{aligned} T(X_*, Y) &= \sum_{x \in X_*} \left| R(x, X_* \cup Y) - \frac{1}{2}(N+1) \right| + \sum_{i \in X'} \left| R(x, X_* \cup Y) - \frac{1}{2}(N+1) \right|, \\ \text{and } T(X, Y) &= \sum_{x \in \tilde{X}} \left| R(x, X \cup Y) - \frac{1}{2}(N+1) \right| + \sum_{i \in X'} \left| R(x, X \cup Y) - \frac{1}{2}(N+1) \right|. \end{aligned}$$

Note that

$$\begin{aligned}\sum_{x\in X_*}\left|R(x,X_*\cup Y)-\frac{1}{2}(N+1)\right| &= \left|-\frac{n-n'}{2}+1\right|+\ldots+0+\ldots+\left|\frac{n-n'}{2}\right|,\\ \sum_{x\in \tilde{X}}\left|R(x,X\cup Y)-\frac{1}{2}(N+1)\right| &= \left|-\frac{n-n'}{2}\right|+\ldots+0+\ldots+\left|\frac{n-n'}{2}-1\right|\\ &= \left|-\frac{n-n'}{2}+1\right|+\ldots+0+\ldots+\left|\frac{n-n'}{2}\right|.\end{aligned}$$

Hence, $\sum_{x\in X_*}\left|R(x,X_*\cup Y)-\frac{1}{2}(N+1)\right|=\sum_{x\in \tilde{X}}\left|R(x,X\cup Y)-\frac{1}{2}(N+1)\right|$. Then, in order to prove our result, it is sufficient to show that

$$\sum_{x\in X'}\left|R(x,X_*\cup Y)-\frac{1}{2}(N+1)\right|=\sum_{x\in X'}\left|R(x,X\cup Y)-\frac{1}{2}(N+1)\right|.$$

Notice that

$$\begin{aligned}\sum_{x\in X'}\left|R(x,X_*\cup Y)-\frac{1}{2}(N+1)\right| &= \sum_{x\in X'}\mathrm{I}\left(R(x,X'\cup Y)<N'\right)\left|R(x,X_*\cup Y)-\frac{1}{2}(N+1)\right|\\ &\quad+\sum_{x\in X'}\mathrm{I}\left(R(x,X'\cup Y)\geq N'\right)\left|R(x,X_*\cup Y)-\frac{1}{2}(N+1)\right|.\end{aligned}$$

According to Lemma B.2.3, we have that

$$\begin{aligned}&\sum_{x\in X'}\mathrm{I}\left(R(x,X'\cup Y)<N'\right)\left|R(x,X_*\cup Y)-\frac{1}{2}(N+1)\right|\\ &=\sum_{x\in X'}\mathrm{I}\left(R(x,X'\cup Y)<N'\right)\left|R(x,X'\cup Y)-\frac{1}{2}(N+1)\right|.\end{aligned}$$

Meanwhile, since $\left\{R(x_1^*,X_*\cup Y),\ldots,R(x_{n-n'}^*,X_*\cup Y)\right\}=\{N',\ldots,N'+n-n'-1\}$ consists of consecutive integers, then according to Lemma B.2.4,

$$\begin{aligned}&\sum_{x\in X'}\mathrm{I}\left(R(x,X'\cup Y)\geq N'\right)\left|R(x,X_*\cup Y)-\frac{1}{2}(N+1)\right|\\ &=\sum_{x\in X'}\mathrm{I}\left(R(x,X'\cup Y)\geq N'\right)\left|R(x,X'\cup Y)+n-n'-\frac{1}{2}(N+1)\right|.\end{aligned}$$

Hence, we have

$$\begin{aligned}
&\sum_{x\in X'}\left|R(x,X_*\cup Y)-\frac{1}{2}(N+1)\right| \\
&=\sum_{x\in X'}\mathrm{I}\left(R(x,X'\cup Y)<N'\right)\left|R(x,X'\cup Y)-\frac{1}{2}(N+1)\right| \\
&+\sum_{x\in X'}\mathrm{I}\left(R(x,X'\cup Y)\geq N'\right)\left|R(x,X'\cup Y)+n-n'-\frac{1}{2}(N+1)\right| \\
&=\sum_{x\in X'}\mathrm{I}\left(R(x,X'\cup Y)<N'\right)\left(\frac{1}{2}(N+1)-R(x,X'\cup Y)\right) \\
&+\sum_{x\in X'}\mathrm{I}\left(R(x,X'\cup Y)\geq N'\right)\left(R(x,X'\cup Y)+n-n'-\frac{1}{2}(N+1)\right).
\end{aligned} \tag{B.9}$$

Similarly, by applying Lemma B.2.3 and Lemma B.2.4, we can show that

$$\begin{aligned}
&\sum_{x\in X'}\left|R(x,X\cup Y)-\frac{1}{2}(N+1)\right| \\
&=\sum_{x\in X'}\mathrm{I}\left(R(x,X'\cup Y)<N'+1\right)\left|R(x,X'\cup Y)-\frac{1}{2}(N+1)\right| \\
&+\sum_{x\in X'}\mathrm{I}\left(R(x,X'\cup Y)\geq N'+1\right)\left|R(x,X'\cup Y)+n-n'-\frac{1}{2}(N+1)\right| \\
&=\sum_{x\in X'}\mathrm{I}\left(R(x,X'\cup Y)<N'+1\right)\left(\frac{1}{2}(N+1)-R(x,X'\cup Y)\right) \\
&+\sum_{x\in X'}\mathrm{I}\left(R(x,X'\cup Y)\geq N'+1\right)\left(R(x,X'\cup Y)+n-n'-\frac{1}{2}(N+1)\right).
\end{aligned}$$

Notice that

$$\begin{aligned}
&\sum_{x\in X'}\mathrm{I}\left(R(x,X'\cup Y)<N'+1\right)\left(\frac{1}{2}(N+1)-R(x,X'\cup Y)\right) \\
&=\sum_{x\in X'}\mathrm{I}\left(R(x,X'\cup Y)<N'\right)\left(\frac{1}{2}(N+1)-R(x,X'\cup Y)\right) \\
&+\sum_{x\in X'}\mathrm{I}\left(N'\leq R(x,X'\cup Y)<N'+1\right)\left(\frac{1}{2}(N+1)-R(x,X'\cup Y)\right).
\end{aligned}$$

Meanwhile,

$$\begin{aligned}
&\sum_{x\in X'}\mathrm{I}\left(R(x,X'\cup Y)\geq N'+1\right)\left(R(x,X'\cup Y)+n-n'-\frac{1}{2}(N+1)\right) \\
&=\sum_{x\in X'}\mathrm{I}\left(R(x,X'\cup Y)\geq N'\right)\left(R(x,X'\cup Y)+n-n'-\frac{1}{2}(N+1)\right) \\
&-\sum_{x\in X'}\mathrm{I}\left(N'+1>R(x,X'\cup Y)\geq N'\right)\left(R(x,X'\cup Y)+n-n'-\frac{1}{2}(N+1)\right).
\end{aligned}$$

Hence, we can further decompose $\sum_{i\in X'}\left|R(x,X\cup Y)-\frac{1}{2}(N+1)\right|$ as follows

$$\begin{aligned}
&\sum_{i\in X'}\left|R(x,X\cup Y)-\frac{1}{2}(N+1)\right| \\
&=\sum_{i\in X'}\mathrm{I}\left(R(x,X'\cup Y)<N'\right)\left(\frac{1}{2}(N+1)-R(x,X'\cup Y)\right) \\
&+\sum_{i\in X'}\mathrm{I}\left(N'\leq R(x,X'\cup Y)<N'+1\right)\left(\frac{1}{2}(N+1)-R(x,X'\cup Y)\right) \\
&+\sum_{i\in X'}\mathrm{I}\left(R(x,X'\cup Y)\geq N'\right)\left(R(x,X'\cup Y)+n-n'-\frac{1}{2}(N+1)\right) \\
&-\sum_{i\in X'}\mathrm{I}\left(N'+1>R(x,X'\cup Y)\geq N'\right)\left(R(x,X'\cup Y)+n-n'-\frac{1}{2}(N+1)\right).
\end{aligned}$$

Recall that in Equation (B.9), we showed that

$$\begin{aligned}
&\sum_{x\in X'}\left|R(x,X_*\cup Y)-\frac{1}{2}(N+1)\right| \\
&=\sum_{x\in X'}\mathrm{I}\left(R(x,X'\cup Y)<N'\right)\left(\frac{1}{2}(N+1)-R(x,X'\cup Y)\right) \\
&+\sum_{x\in X'}\mathrm{I}\left(R(x,X'\cup Y)\geq N'\right)\left(R(x,X'\cup Y)+n-n'-\frac{1}{2}(N+1)\right).
\end{aligned}$$

Hence, we further have

$$\begin{aligned}
&\sum_{x\in X'}\left|R(x,X\cup Y)-\frac{1}{2}(N+1)\right| \\
&=\sum_{x\in X'}\left|R(x,X_*\cup Y)-\frac{1}{2}(N+1)\right| \\
&+\sum_{x\in X'}\mathrm{I}\left(N'\leq R(x,X'\cup Y)<N'+1\right)\left(\frac{1}{2}(N+1)-R(x,X'\cup Y)\right) \\
&-\sum_{x\in X'}\mathrm{I}\left(N'+1>R(x,X'\cup Y)\geq N'\right)\left(R(x,X'\cup Y)+n-n'-\frac{1}{2}(N+1)\right).
\end{aligned}$$

If $\sum_{i\in X'}\mathrm{I}\left(N'\leq R(x,X'\cup Y)<N'+1\right)=0$ then we have

$$\sum_{x\in X'}\left|R(x,X\cup Y)-\frac{1}{2}(N+1)\right|=\sum_{x\in X'}\left|R(x,X_*\cup Y)-\frac{1}{2}(N+1)\right|,$$

which proves our result. Otherwise, we have

$$\sum_{x\in X'}\mathrm{I}\left(N'\leq R(x,X'\cup Y)<N'+1\right)=1,$$

and $R(x, X' \cup Y) = N'$. Subsequently,

$$\begin{aligned}
&\sum_{x \in X'} \mathrm{I}\left(N' \leq R(x, X' \cup Y) < N' + 1\right) \left(\frac{1}{2}(N+1) - R(x, X' \cup Y)\right) \\
&- \sum_{x \in X'} \mathrm{I}\left(N' + 1 > R(x, X' \cup Y) \geq N'\right) \left(R(x, X' \cup Y) + n - n' - \frac{1}{2}(N+1)\right) \\
&= \left(\frac{1}{2}(N+1) - N'\right) - \left(N' + n - n' - \frac{1}{2}(N+1)\right) \\
&= \left(\frac{1}{2}(N+1) + \frac{n-n'}{2} - \frac{1}{2}(N+1)\right) - \left(-\frac{n-n'}{2} + \frac{1}{2}(N+1) + n - n' - \frac{1}{2}(N+1)\right) \\
&= \frac{n-n'}{2} - \frac{n-n'}{2} = 0.
\end{aligned}$$

Hence, we obtain our desired result

$$\sum_{x \in X'} \left| R(x, X \cup Y) - \frac{1}{2}(N+1) \right| = \sum_{x \in X'} \left| R(x, X_* \cup Y) - \frac{1}{2}(N+1) \right|.$$

□

The following proposition considers the ranks of imputation which would allow us to take the lower bound of the Ansari-Bradley statistic, when the total sample size $N = n + m$ is odd.

Proposition B.2.6. Suppose $X = \{x_1, \ldots, x_n\}$ and $Y = \{y_1, \ldots, y_m\}$ are samples of distinct real values. Denote $N = n+m$ and assume N is odd. Let $X' \subset X$ be a non-empty subset of X with size $|X'| = n'$. Consider a set $\hat{X} = \{x_1^*, \ldots, x_{n-n'}^*\}$ of real values, and denote $X_* = \hat{X} \cup X'$. Assume that all values in $X_* \cup Y$ are distinct, and denote $\mathcal{R} = \left\{R(x_1^*, X_* \cup Y), \ldots, R(x_{n-n'}^*, X_* \cup Y)\right\}$. Subsequently, if $n - n'$ is even, then we have $T(X_*, Y) \leq T(X, Y)$, provided that either

$$\begin{aligned}
\mathcal{R} &= \left\{-\frac{n-n'}{2} + 1 + \frac{1}{2}(N+1), \ldots, \frac{n-n'}{2} + \frac{1}{2}(N+1)\right\}, \\
\text{or } \mathcal{R} &= \left\{-\frac{n-n'}{2} + \frac{1}{2}(N+1), \ldots, \frac{n-n'}{2} - 1 + \frac{1}{2}(N+1)\right\}.
\end{aligned}$$

However, if $n - n'$ is odd then we have $T(X_*, Y) \leq T(X, Y)$, provided that

$$\mathcal{R} = \left\{-\frac{n-n'-1}{2} + \frac{1}{2}(N+1), \ldots, \frac{n-n'-1}{2} + \frac{1}{2}(N+1)\right\}.$$

Proof. We prove the proposition by mathematical induction on $n - n'$. Consider the base case when $n - n' = 1$. We are given that $R(x_1^*, X_* \cup Y) = (N+1)/2$, which immediately follows that $|R(x_1^*, X_* \cup Y) - (N+1)/2| = 0$. Hence, we have $T(X_*, Y) \leq T(X, Y)$ according to Proposition B.1.6, which proves the base case.

Let $k \in \{1, \ldots, n-1\}$, and suppose that the statement of Proposition B.2.6 holds for $n - n' = k$. We show that the proposition also holds for $n - n' = k + 1$. We prove this result for the following

two cases:

$$\text{case(I)} : k \text{ is odd},$$
$$\text{and case(II)} : k \text{ is even},$$

separately.

Suppose case (I) holds. We show that the proposition also holds for $n-n' = k+1$. Specifically, we show that $T(X_*, Y) \leq T(X, Y)$ provided either

$$\mathcal{R} = \left\{ -\frac{n-n'}{2} + 1 + \frac{1}{2}(N+1), \ldots, \frac{n-n'}{2} + \frac{1}{2}(N+1) \right\},$$
$$\text{or } \mathcal{R} = \left\{ -\frac{n-n'}{2} + \frac{1}{2}(N+1), \ldots, \frac{n-n'}{2} - 1 + \frac{1}{2}(N+1) \right\},$$

where $n-n' = k+1$. By Lemma B.2.5, these two cases yield the same value of the Ansari-Bradley test statistic $T(X_*, Y)$. Therefore, denote

$$\mathcal{R}_1 = \left\{ -\frac{n-n'}{2} + 1 + \frac{1}{2}(N+1), \ldots, \frac{n-n'}{2} + \frac{1}{2}(N+1) \right\},$$
$$\text{or } \mathcal{R}_2 = \left\{ -\frac{n-n'}{2} + \frac{1}{2}(N+1), \ldots, \frac{n-n'}{2} - 1 + \frac{1}{2}(N+1) \right\},$$

it is sufficient to show $T(X_*, Y) \leq T(X, Y)$ when $\mathcal{R} = \mathcal{R}_1$ or $\mathcal{R} = \mathcal{R}_2$.

Let us denote $\tilde{X} = X \setminus X'$, and $\tilde{X} = \{\tilde{x}_1, \ldots, \tilde{x}_{k+1}\}$. Consider a set $Z^* = \{z_1^*, \ldots, z_k^*\}$ of distinct real values, and denote $Z = Z^* \cup \{\tilde{x}_1\} \cup X'$. Assume that all values in $Z \cup Y$ are distinct, and

$$\{R(z_1^*, Z \cup Y), \ldots, R(z_k^*, Z \cup Y)\} = \left\{ -\frac{k-1}{2} + \frac{1}{2}(N+1), \ldots, \frac{k-1}{2} + \frac{1}{2}(N+1) \right\}.$$

By the induction hypothesis, i.e. the Proposition B.2.6 is correct when $n-n' = k$, it follows that

$$T(Z, Y) \leq T(X, Y). \tag{B.10}$$

Since $z_1^*, \ldots, z_k^*$ takes the ranks between $-(k-1)/2 + (N+1)/2$ to $(k-1)/2 + (N+1)/2$, the rank of $\tilde{x}_1$ in $Z \cup Y$ is either smaller than $-(k-1)/2 + (N+1)/2$ or greater than $(k-1)/2 + (N+1)/2$. In other words, we have either

$$R(\tilde{x}_1, Z \cup Y) < -\frac{k-1}{2} + \frac{1}{2}(N+1), \text{ or } R(\tilde{x}_1, Z \cup Y) > \frac{k-1}{2} + \frac{1}{2}(N+1).$$

Below we prove that we have our desired result $T(X_*, Y) \leq T(X, Y)$ for both cases.

Suppose $R(\tilde{x}_1, Z \cup Y) < -(k-1)/2 + (N+1)/2$. Then, consider a real value x^* that is distinct to values in $Z^* \cup X' \cup Y$. Denote $Z' = \{x^*\} \cup Z^* \cup X'$, and assume that

$$R(x^*, Z' \cup Y) = -\frac{k-1}{2} - 1 + \frac{1}{2}(N+1).$$

Notice that $R(x^*, Z' \cup Y) < (N+1)/2$ and $R(\tilde{x}_1, Z \cup Y) < (N+1)/2$, and

$$\left| R(x^*, Z' \cup Y) - \frac{1}{2}(N+1) \right| \leq \left| R(\tilde{x}_1, Z \cup Y) - \frac{1}{2}(N+1) \right|.$$

According to Proposition B.1.4, we have

$$T(Z', Y) \leq T(Z, Y) \leq^{\text{Inequality (B.10)}} T(X, Y).$$

To complete our prove for this case, it then suffices to show that $T(X_*, Y) = T(Z', Y)$.

Recall that for any $i \in \{1, \ldots, k\}$, we have

$$R(z_i^*, Z \cup Y) > -\frac{k-1}{2} - 1 + \frac{1}{2}(N+1).$$

Then, for any $i \in \{1, \ldots, k\}$,

$$\begin{aligned} R(z_i^*, Z \cup Y) &\geq -\frac{k-1}{2} + \frac{1}{2}(N+1) > R(\tilde{x}_1, Z \cup Y), \\ \text{and } R(z_i^*, Z \cup Y) &> R(x^*, Z' \cup Y). \end{aligned}$$

According to Lemma B.2.1, for any $i \in \{1, \ldots, k\}$,

$$\begin{aligned} & R(z_i^*, Z \cup Y) = R(z_i^*, Z' \cup Y) \\ \Longrightarrow & \{R(z_1^*, Z' \cup Y), \ldots, R(z_k^*, Z' \cup Y)\} = \left\{ -\frac{k-1}{2} + \frac{1}{2}(N+1), \ldots, \frac{k-1}{2} + \frac{1}{2}(N+1) \right\}. \end{aligned}$$

Hence, we have

$$\mathcal{R}_2 = R(x^*, Z' \cup Y) \cup \{R(z_1^*, Z' \cup Y), \ldots, R(z_k^*, Z' \cup Y)\}.$$

According to Lemma B.1.3, we have

$$T(X_*, Y) = T(Z', Y),$$

which completes our proof for this case when $R(\tilde{x}_1, Z \cup Y) < -(k-1)/2 + (N+1)/2$.

The case when $R(\tilde{x}_1, Z \cup Y) > (k-1)/2 + (N+1)/2$ can be proved similarly. Consider a real value x^* that is distinct to all values in $Z^* \cup X' \cup Y$. Denote $Z' = \{x^*\} \cup Z^* \cup X'$, and assume that

$$R(x^*, Z' \cup Y) = \frac{k-1}{2} + 1 + \frac{1}{2}(N+1).$$

Note that $R(x^*, Z' \cup Y) \geq (N+1)/2$ and $R(\tilde{x}_1, Z \cup Y) \geq (N+1)/2$, and

$$\left| R(x^*, Z' \cup Y) - \frac{1}{2}(N+1) \right| \leq \left| R(\tilde{x}_1, Z \cup Y) - \frac{1}{2}(N+1) \right|.$$

Then, according to Proposition B.1.4,

$$T(Z', Y) \leq T(Z, Y) \leq^{\text{Inequality (B.10)}} T(X, Y).$$

To complete our prove for this case, it then suffices to show that $T(X_*, Y) = T(Z', Y)$.

Recall that for any $i \in \{1, \ldots, k\}$,

$$R(z_i^*, Z \cup Y) < \frac{k-1}{2} + 1 + \frac{1}{2}(N+1).$$

Then for any $i \in \{1, \ldots, k\}$, we have

$$\begin{aligned} R(x^*, Z' \cup Y) &= \frac{k-1}{2} + 1 + \frac{1}{2}(N+1) > R(z_i^*, Z \cup Y), \\ \text{and } R(\tilde{x}_1, Z \cup Y) &> R(z_i^*, Z \cup Y). \end{aligned}$$

According to Lemma B.2.2, for any $i \in \{1, \ldots, k\}$,

$$\begin{aligned} & R(z_i^*, Z \cup Y) = R(z_i^*, Z' \cup Y) \\ \Longrightarrow & \{R(z_1^*, Z' \cup Y), \ldots, R(z_k^*, Z' \cup Y)\} = \left\{-\frac{k-1}{2} + \frac{1}{2}(N+1), \ldots, \frac{k-1}{2} + \frac{1}{2}(N+1)\right\}. \end{aligned}$$

Hence, we have

$$\mathcal{R}_1 = R(x^*, Z' \cup Y) \cup \{R(z_1^*, Z' \cup Y), \ldots, R(z_k^*, Z' \cup Y)\}.$$

According to Lemma B.1.3, we have

$$T(X_*, Y) = T(Z', Y),$$

which completes our proof for this case when $R(\tilde{x}_1, Z \cup Y) > (k-1)/2 + (N+1)/2$.

Suppose case (II) holds. In other words, k is even. We show that the proposition also holds for $n - n' = k + 1$. Specifically, we show that $T(X_*, Y) \leq T(X, Y)$ provided

$$\mathcal{R} = \left\{-\frac{n-n'-1}{2} + \frac{1}{2}(N+1), \ldots, \frac{n-n'-1}{2} + \frac{1}{2}(N+1)\right\},$$

where $n - n' = k + 1$. As before, let us denote $\tilde{X} = X \setminus X'$, and $\tilde{X} = \{\tilde{x}_1, \ldots, \tilde{x}_{k+1}\}$. Then, we show that $T(X_*, Y) \leq T(X, Y)$ for the following two cases separately:

$$\begin{aligned} \text{(i)} : R(\tilde{x}_1, \{\tilde{x}_1\} \cup X' \cup Y) &< -\frac{k}{2} + 1 + \frac{1}{2}(N+1), \\ \text{(ii)} : R(\tilde{x}_1, \{\tilde{x}_1\} \cup X' \cup Y) &\geq -\frac{k}{2} + 1 + \frac{1}{2}(N+1). \end{aligned}$$

Suppose case (i) holds, i.e., assume that $R(\tilde{x}_1, \{\tilde{x}_1\} \cup X' \cup Y) < -k/2 + 1 + (N+1)/2$. Consider a set $Z^* = \{z_1^*, \ldots, z_k^*\}$ of distinct real values, and denote $Z = Z^* \cup \{\tilde{x}_1\} \cup X'$. Assume that all

values in Z are distinct, and

$$\{R(z_1^*, Z \cup Y), \ldots, R(z_k^*, Z \cup Y)\} = \left\{-\frac{k}{2} + 1 + \frac{1}{2}(N+1), \ldots, \frac{k}{2} + \frac{1}{2}(N+1)\right\}.$$

By the induction hypothesis, i.e. the Proposition B.2.6 is correct when $n - n' = k$, we have

$$T(Z, Y) \leq T(X, Y). \tag{B.11}$$

Notice that

$$R(\tilde{x}_1, \{\tilde{x}_1\} \cup X' \cup Y) < -\frac{k}{2} + 1 + \frac{1}{2}(N+1) = \min\{R(z_1^*, Z \cup Y), \ldots, R(z_k^*, Z \cup Y)\}.$$

According to Lemma B.2.3, we have

$$R(\tilde{x}_1, Z \cup Y) = R(\tilde{x}_1, \{\tilde{x}_1\} \cup X' \cup Y) < -\frac{k}{2} + 1 + \frac{1}{2}(N+1).$$

Now consider a real value x^*. Denote $Z' = Z^* \cup \{x^*\} \cup X'$. Assume that values in Z' are distinct and

$$R(x^*, Z' \cup Y) = -k/2 + (N+1)/2.$$

Notice that $R(\tilde{x}_1, Z \cup Y) \leq (N+1)/2$ and $R(x^*, Z' \cup Y) \leq (N+1)/2$, and $R(\tilde{x}_1, Z \cup Y) \leq R(x^*, Z' \cup Y)$. Then according to Proposition B.1.4, we have

$$T(Z', Y) \leq T(Z, Y) \leq^{\text{Inequality (B.11)}} T(X, Y).$$

Then, for proving our result under this case, it suffices to prove $T(X_*, Y) = T(Z', Y)$.

Notice that for any $i \in \{1, \ldots, k\}$,

$$R(z_i^*, Z \cup Y) \geq -\frac{k}{2} + 1 + \frac{1}{2}(N+1).$$

Then, for any $i \in \{1, \ldots, k\}$, we have

$$\begin{aligned} R(x^*, Z' \cup Y) &= -\frac{k}{2} + \frac{1}{2}(N+1) < R(z_i^*, Z \cup Y) \\ R(\tilde{x}_1, Z \cup Y) &< -\frac{k}{2} + 1 + \frac{1}{2}(N+1) \leq R(z_i^*, Z \cup Y). \end{aligned}$$

According to Lemma B.2.1, for any $i \in \{1, \ldots, k\}$,

$$\begin{aligned} &R(z_i^*, Z' \cup Y) = R(z_i^*, Z \cup Y) \\ \Longrightarrow &\{R(z_1^*, Z' \cup Y), \ldots, R(z_k^*, Z' \cup Y)\} = \left\{-\frac{k}{2} + 1 + \frac{1}{2}(N+1), \ldots, \frac{k}{2} + \frac{1}{2}(N+1)\right\}. \end{aligned}$$

Hence, we have

$$\mathcal{R} = \{R(x^*, Z' \cup Y)\} \cup \{R(z_1^*, Z' \cup Y), \ldots, R(z_k^*, Z' \cup Y)\}.$$

According to Lemma B.1.3, we have

$$T(X_*, Y) = T(Z', Y),$$

which completes our proof for case (i).

Case (ii) can be proved similarly. Suppose case (ii) holds, i.e., assume that $R(\tilde{x}_1, \{\tilde{x}_1\} \cup X' \cup Y) \geq -k/2 + 1 + (N+1)/2$. Consider a set $Z^* = \{z_1^*, \ldots, z_k^*\}$ of distinct real values, and denote $Z = Z^* \cup \{\tilde{x}_1\} \cup X'$. Assume that all values in Z are distinct, and

$$\{R(z_1^*, Z \cup Y), \ldots, R(z_k^*, Z \cup Y)\} = \left\{-\frac{k}{2} + \frac{1}{2}(N+1), \ldots, \frac{k}{2} - 1 + \frac{1}{2}(N+1)\right\}.$$

By the induction hypothesis, i.e. the Proposition B.2.6 is correct when $n - n' = k$, we have

$$T(Z, Y) \leq T(X, Y). \tag{B.12}$$

Notice that

$$R(\tilde{x}_1, \{\tilde{x}_1\} \cup X' \cup Y) \geq -\frac{k}{2} + 1 + \frac{(N+1)}{2} \geq \min\{R(z_1^*, Z \cup Y), \ldots, R(z_k^*, Z \cup Y)\}.$$

Then according to Lemma B.2.4, we have

$$R(\tilde{x}_1, Z \cup Y) = R(\tilde{x}_1, \{\tilde{x}_1\} \cup X' \cup Y) + k \geq \frac{k}{2} + 1 + \frac{1}{2}(N+1).$$

Now consider a real value x^*. Denote $Z' = Z^* \cup \{x^*\} \cup X'$. Assume that values in Z' are distinct and

$$R(x^*, Z' \cup Y) = \frac{k}{2} + \frac{1}{2}(N+1).$$

Notice that $R(x^*, Z' \cup Y) \geq (N+1)/2$, and $R(\tilde{x}_1, Z \cup Y) \geq (N+1)/2$ and $R(x^*, Z' \cup Y) \leq R(\tilde{x}_1, Z \cup Y)$. Then, according to Proposition B.1.4, we have

$$T(Z', Y) \leq T(Z, Y) \leq^{\text{Inequality (B.12)}} T(X, Y).$$

Then, for proving our result under this case, it suffices to prove $T(X_*, Y) = T(Z', Y)$.

Notice that for any $i \in \{1, \ldots, k\}$,

$$R(z_i^*, Z \cup Y) \leq \frac{k}{2} - 1 + \frac{1}{2}(N+1).$$

Thus, for any $i \in \{1, \ldots, k\}$,

$$R(x^*, Z' \cup Y) = \frac{k}{2} + \frac{1}{2}(N+1) > R(z_i^*, Z \cup Y),$$
$$\text{and } R(\tilde{x}_1, Z \cup Y) \geq \frac{k}{2} + 1 + \frac{1}{2}(N+1) > R(z_i^*, Z \cup Y).$$

According to Lemma B.2.2, for any $i \in \{1, \ldots, k\}$,

$$\begin{aligned} & R(z_i^*, Z \cup Y) = R(z_i^*, Z' \cup Y) \\ \implies & \left\{R(z_1^*, Z' \cup Y), \ldots, R(z_k^*, Z' \cup Y)\right\} = \left\{-\frac{k}{2} + \frac{1}{2}(N+1), \ldots, \frac{k}{2} - 1 + \frac{1}{2}(N+1)\right\}. \end{aligned}$$

Hence, we have

$$\mathcal{R} = \{R(x^*, Z' \cup Y)\} \cup \{R(z_1^*, Z' \cup Y), \ldots, R(z_k^*, Z' \cup Y)\}.$$

According to Lemma B.1.3, we have

$$T(X_*, Y) = T(Z', Y),$$

which proves our result under case (i), and completes our proof. □

Recall that Lemma B.2.5 and Proposition B.2.6 both assume that the total sample size $N = n + m$ is odd. The following Lemma B.2.7 and Proposition B.2.8 provides similar results when the total sample size N is even.

The proofs for Lemma B.2.7 and Proposition B.2.8 are similar to that of Lemma B.2.5 and Proposition B.2.6, respectively, differing only in notation. Thus, the proofs are omitted here.

Lemma B.2.7. Suppose $X = \{x_1, \ldots, x_n\}$ and $Y = \{y_1, \ldots, y_m\}$ are samples of distinct real values. Let $N = n + m$ and assume N is even. Suppose $X' \subset X$ is a non-empty subset of X such that $|X| = n'$ is odd. Let $\tilde{X} = X \setminus X'$ and denote $\tilde{X} = \{\tilde{x}_1, \ldots, \tilde{x}_{n-n'}\}$. Consider a set $\tilde{X} = \{x_1^*, \ldots, x_{n-n'}^*\}$ of real values, and denote $X_* = \tilde{X} \cup X'$. Assume that values in $X_* \cup X'$ are distinct. Then, if

$$\begin{aligned} & \{R(\tilde{x}_1, X \cup Y), \ldots, R(\tilde{x}_{n-n'}, X \cup Y)\} \\ & = \left\{-\frac{n-n'}{2} + 1 + \frac{1}{2}(N+1), \ldots, \frac{n-n'}{2} + \frac{1}{2}(N+1)\right\}, \end{aligned}$$

and

$$\begin{aligned} & \left\{R(x_1^*, X_* \cup Y), \ldots, R(x_{n-n'}^*, X_* \cup Y)\right\} \\ & = \left\{-\frac{n-n'}{2} + \frac{1}{2}(N+1), \ldots, \frac{n-n'}{2} - 1 + \frac{1}{2}(N+1)\right\}, \end{aligned}$$

we have $T(X, Y) = T(X_*, Y)$.

Proof. The proof follows the same approach as that of Lemma B.2.5, differing only in notation. Thus, it is omitted here. □

Proposition B.2.8. Suppose $X = \{x_1, \ldots, x_n\}$ and $Y = \{y_1, \ldots, y_m\}$ are samples of distinct real values. Let $N = n+m$ and assume N is even. Let X' be a non-empty subset of X and denote $|X'| = n'$. Consider a set $\hat{X} = \{x_1^*, \ldots, x_{n-n'}^*\}$ of distinct real values, and denote $X_* = \hat{X} \cup X'$. Assume all values in $X_* \cup Y$ are distinct. Denote $\mathcal{R} = \left\{R(x_1^*, X_* \cup Y), \ldots, R(x_{n-n'}^*, X_* \cup Y)\right\}$.

Subsequently, if $n - n'$ is even, we have $T(X_*, Y) \leq T(X, Y)$, provided

$$\mathcal{R} = \left\{ -\frac{n - n' - 1}{2} + \frac{1}{2}(N+1), \ldots, \frac{n - n' - 1}{2} + \frac{1}{2}(N+1) \right\}.$$

Similarly, if $n - n'$ is odd, we have $T(X_*, Y) \leq T(X, Y)$, provided that either

$$\mathcal{R} = \left\{ -\frac{n - n'}{2} + \frac{1}{2}(N+1), \ldots, \frac{n - n'}{2} - 1 + \frac{1}{2}(N+1) \right\},$$
$$\text{or } \mathcal{R} = \left\{ -\frac{n - n'}{2} + 1 + \frac{1}{2}(N+1), \ldots, \frac{n - n'}{2} + \frac{1}{2}(N+1) \right\}.$$

Proof. The proof follows the same approach as that of Proposition B.2.6, differing only in notation. Thus, it is omitted here. □

Proposition B.2.6 and Proposition B.2.8 provides the ranks of the imputations for taking the minimum value of the Ansari-Bradley statistic. Below we consider the minimum value of the Ansari-Bradley after the missing data are imputed according to Proposition B.2.6 and Proposition B.2.8.

We first make the following definition, which classifies all possible configurations of the total sample sizes $N = n + m$ and the number of missing values based on the parity of N and $n - n'$.

Definition B.2.9. Let $N \geq n \geq n'$ be positive integers. We define the following four cases $\mathcal{C}(1), \mathcal{C}(2), \mathcal{C}(3)$ and $\mathcal{C}(4)$ according to the parity of N and $n - n'$:

$$\begin{aligned}
&\mathcal{C}(1): \ N \text{ is odd but } n - n' \text{ is even,} \\
&\mathcal{C}(2): \ N \text{ is odd and } n - n' \text{ is odd,} \\
&\mathcal{C}(3): \ N \text{ is even but } n - n' \text{ is odd,} \\
&\mathcal{C}(4): \ N \text{ is even and } n - n' \text{ is even.}
\end{aligned}$$

Then, we define the following four sets, which are useful for Proposition B.2.11.

Definition B.2.10. Let $N \geq n \geq n'$ be positive integers, we denote $\mathcal{S}(1), \mathcal{S}(2), \mathcal{S}(3), \mathcal{S}(4)$ as

$$\begin{aligned}
\mathcal{S}(1) &= \left\{ -\frac{n - n'}{2} + \frac{1}{2}(N+1), \ldots, \frac{n - n'}{2} - 1 + \frac{1}{2}(N+1) \right\}, \\
\mathcal{S}(2) &= \left\{ -\frac{n - n' - 1}{2} + \frac{1}{2}(N+1), \ldots, \frac{n - n' - 1}{2} + \frac{1}{2}(N+1) \right\}, \\
\mathcal{S}(3) &= \left\{ -\frac{n - n'}{2} + \frac{1}{2}(N+1), \ldots, \frac{n - n'}{2} - 1 + \frac{1}{2}(N+1) \right\}, \\
\mathcal{S}(4) &= \left\{ -\frac{n - n' - 1}{2} + \frac{1}{2}(N+1), \ldots, \frac{n - n' - 1}{2} + \frac{1}{2}(N+1) \right\},
\end{aligned}$$

respectively.

Using Definition B.2.9 and Definition B.2.10, the following proposition gives the explicit form of the Ansari-Bradley statistic after imputing the missing values in X according to Propositions B.2.6 and B.2.8, across all cases defined in Definition B.2.9.

Proposition B.2.11. Suppose $X = \{x_1, \ldots, x_n\}$ and $Y = \{y_1, \ldots, y_m\}$ are samples of distinct real values and $X' \subset X$ is a non-empty subset of X with sample size $|X'| = n'$. Denote $N = n+m$. Consider a set $\hat{X} = \{x_1^*, \ldots, x_{n-n'}^*\}$ of distinct real values, and denote $X_* = \hat{X} \cup X'$. Assume that all values in $X_* \cup Y$ are distinct and denote $\mathcal{R} = \{R(x_1^*, X_* \cup Y), \ldots, R(x_{n-n'}^*, X_* \cup Y)\}$. Subsequently, we have

$$T(X_*, Y) = \begin{cases} T(X', Y) + (n^2 - n'^2)/4 & \text{if } \mathcal{C}(1) \text{ holds, and } \mathcal{R} = \mathcal{S}(1), \\ T(X', Y) + (n^2 - n'^2 - 1)/4 & \text{if } \mathcal{C}(2) \text{ holds, and } \mathcal{R} = \mathcal{S}(2), \\ T(X', Y) + (n^2 - n'^2 + 1)/4 & \text{if } \mathcal{C}(3) \text{ holds, and } \mathcal{R} = \mathcal{S}(3), \\ T(X', Y) + (n^2 - n'^2)/4 & \text{if } \mathcal{C}(4) \text{ holds, and } \mathcal{R} = \mathcal{S}(4), \end{cases}$$

Proof. To start, let us denote $X' = \{x_1', \ldots, x_{n'}'\}$. Below, we only consider the first case when $\mathcal{C}(1)$ holds, and $\mathcal{R} = \mathcal{S}(1)$, because the other three cases can be proved following the same approach.

By the definition of the Ansari-Bradley test statistic, we have

$$T(X_*, Y) = \sum_{i=1}^{n-n'} \left| R(x_i^*, X_* \cup Y) - \frac{N+1}{2} \right| + \sum_{i=1}^{n'} \left| R(x_i', X_* \cup Y) - \frac{N+1}{2} \right|.$$

Since $\mathcal{R} = \mathcal{S}(1)$,

$$\sum_{i=1}^{n-n'} \left| R(x_i^*, X_* \cup Y) - \frac{N+1}{2} \right| = \left| -\frac{n-n'}{2} \right| + \ldots + \left| \frac{n-n'}{2} - 1 \right| = \frac{(n-n')^2}{4}.$$

Hence, in order to prove our result under this case, it is sufficient to show that

$$\begin{aligned} \sum_{i=1}^{n'} \left| R(x_i', X_* \cup Y) - \frac{N+1}{2} \right| &= T(X', Y) + \frac{n^2 - n'^2}{4} - \frac{(n-n')^2}{4} \\ &= T(X', Y) + \frac{(n-n')n'}{2}. \end{aligned}$$

In order to prove our result under this case, we first show the two equivalence (B.13) and (B.14) below.

Note that

$$\min \mathcal{S}(1) = -\frac{n-n'}{2} + \frac{1}{2}(N+1) = \frac{-n+n'+n+m+1}{2} = \frac{n'+m+1}{2}.$$

Then, according to Lemma B.2.3, for any $i \in \{1, \ldots, n'\}$, we have

$$R(x_i', X' \cup Y) < \frac{n'+m+1}{2} \iff R(x_i', X_* \cup Y) = R(x_i', X' \cup Y).$$

Further, for any $i \in \{1, \ldots, n'\}$, we have

$$R(x_i', X' \cup Y) < \frac{n'+m+1}{2} \iff R(x_i', X_* \cup Y) < \frac{N+1}{2}. \tag{B.13}$$

The " $\Longrightarrow$ " follows from the fact that $R(x_i', X_* \cup Y) = R(x_i', X' \cup Y)$ and $n' + m < N$. For the

" $\Longleftarrow$ ", since values of X_* take all the ranks in $\mathcal{S}(1)$,

$$\begin{aligned} & R(x_i', X_* \cup Y) < (N+1)/2 \\ \Longrightarrow\ & R(x_i', X_* \cup Y) < \min \mathcal{S}(1) = -\frac{n-n'}{2} + \frac{1}{2}(N+1) = \frac{n'+m+1}{2}. \end{aligned}$$

Notice that $\mathcal{R} = \mathcal{S}(1)$ consists of consecutive integers, and $\min \mathcal{S}(1) = (n'+m+1)/2$ as we have shown before. Then according to Lemma B.2.4, for any $i \in \{1, \ldots, n'\}$,

$$R(x_i', X' \cup Y) \geq \frac{n'+m+1}{2} \iff R(x_i', X_* \cup Y) = R(x_i', X' \cup Y) + n - n'.$$

Further, for any $i \in \{1, \ldots, n'\}$, we have

$$R(x_i', X' \cup Y) \geq \frac{n'+m+1}{2} \iff R(x_i', X_* \cup Y) \geq \frac{N+1}{2}. \tag{B.14}$$

The " $\Longrightarrow$ " follows from the fact that $R(x_i', X_* \cup Y) = R(x_i', X' \cup Y) + n - n'$ and $(n'+m+1)/2 + n - n' \geq (N+1)/2$. For the " $\Longleftarrow$ ", since values of X_* take all the ranks in $\mathcal{S}(1)$, we have

$$\begin{aligned} & R(x_i', X_* \cup Y) \geq \frac{N+1}{2} \\ \Longrightarrow\ & R(x_i', X_* \cup Y) > \max \mathcal{S}(1) = \frac{n-n'}{2} - 1 + \frac{1}{2}(N+1). \end{aligned}$$

Since $R(x_i', X_* \cup Y) = R(x_i', X' \cup Y) + \sum_{i=1}^{n-n'} \mathrm{I}\left(x_i' \geq x_i^*\right)$,

$$\begin{aligned} & R(x_i', X' \cup Y) = R(x_i', X_* \cup Y) - \sum_{i=1}^{n-n'} \mathrm{I}\left(x_i' \geq x_i^*\right) \\ \Longrightarrow\ & R(x_i', X' \cup Y) > \frac{n-n'}{2} - 1 + \frac{1}{2}(N+1) - (n-n') \\ \Longrightarrow\ & R(x_i', X' \cup Y) \geq \frac{n'+m+1}{2}. \end{aligned}$$

After proving (B.13) and (B.14), we first decompose $\sum_{i=1}^{n'} |R(x_i', X_* \cup Y) - (N+1)/2|$ as such

$$\begin{aligned} \sum_{i=1}^{n'} \left| R(x_i', X_* \cup Y) - \frac{N+1}{2} \right| = & \sum_{i=1}^{n'} \mathrm{I}\left(R(x_i', X_* \cup Y) < \frac{N+1}{2} \right) \left(\frac{N+1}{2} - R(x_i', X_* \cup Y) \right) \\ & + \sum_{i=1}^{n'} \mathrm{I}\left(R(x_i', X_* \cup Y) \geq \frac{N+1}{2} \right) \left(R(x_i', X_* \cup Y) - \frac{N+1}{2} \right). \end{aligned}$$

For notation convenience, we denote $N' = (n'+m+1)/2$ and

$$r_i = R(x_i', X' \cup Y), \text{ for any } i \in \{1, \ldots, n'\}.$$

Then, by applying (B.13) and (B.14), we further have

$$\sum_{i=1}^{n'} \left| R(x'_i, X_* \cup Y) - \frac{N+1}{2} \right| = \sum_{i=1}^{n'} \mathrm{I}\left(r_i < N'\right) \left(\frac{N+1}{2} - r_i \right) + \sum_{i=1}^{n'} \mathrm{I}\left(r_i \geq N'\right) \left(r_i + n - n' - \frac{N+1}{2} \right).$$

Notice that

$$\frac{N+1}{2} = \frac{n+m+1}{2} = \frac{n'+m+1}{2} + \frac{n-n'}{2} = N' + \frac{n-n'}{2},$$
$$\text{and } n - n' - \frac{N+1}{2} = \frac{2n-2n'-n-m-1}{2} = -\frac{n'+m+1}{2} + \frac{n-n'}{2} = -n' + \frac{n-n'}{2}.$$

Hence, we have

$$\begin{aligned}
&\sum_{i=1}^{n'} \left| R(x'_i, X_* \cup Y) - \frac{N+1}{2} \right| \\
&= \sum_{i=1}^{n'} \mathrm{I}\left(r_i < N'\right) \left(N' - r_i + \frac{n-n'}{2} \right) + \sum_{i=1}^{n'} \mathrm{I}\left(r_i \geq N'\right) \left(r_i - N' + \frac{n-n'}{2} \right) \\
&= \sum_{i=1}^{n'} \mathrm{I}\left(r_i < N'\right) \left(N' - r_i \right) + \sum_{i=1}^{n'} \mathrm{I}\left(r_i \geq N'\right) \left(r_i - N' \right) + \frac{(n-n')n'}{2} \\
&= \sum_{i=1}^{n'} \left| r_i - N' \right| + \frac{(n-n')n'}{2} = T(X', Y) + \frac{(n-n')n'}{2},
\end{aligned}$$

which completes the proof of our result under Case (1). □

Finally, we combine the results of Propositions B.2.6, B.2.8 and B.2.11, and concludes this section. The following theorem provides the lower bound of the Ansari-Bradley statistic in the presence of missing data when only values in X can be missing.

Theorem B.2.12. Suppose $X = \{x_1, \ldots, x_n\}$ and $Y = \{y_1, \ldots, y_m\}$ are samples of distinct real values, and $X' \subset X$ is a subset of X with sample size $|X| = n'$, which are observed. Then, the minimum possible Ansari-Bradley test statistic, across all possible values of missing data, is given as follows:

$$\min_{X \setminus X' \in \mathbb{R}^{n-n'}} T(X, Y) = T(X_*, Y) = \begin{cases} T(X', Y) + (n^2 - n'^2)/4 & \text{if } \mathcal{C}(1) \text{ or } \mathcal{C}(4) \text{ holds,} \\ T(X', Y) + (n^2 - n'^2 - 1)/4 & \text{if } \mathcal{C}(2) \text{ holds,} \\ T(X', Y) + (n^2 - n'^2 + 1)/4 & \text{if } \mathcal{C}(3) \text{ holds.} \end{cases}$$

Proof. Denote $N = n + m$. Then, the parity of N and $n - n'$ satisfy either $\mathcal{C}(1)$, $\mathcal{C}(2)$, $\mathcal{C}(3)$ and $\mathcal{C}(4)$. Let us assume that $\mathcal{C}(k)$ holds, where $k \in \{1, 2, 3, 4\}$. Let $\hat{X} = \{x_1^*, \ldots, x_{n-n'}^*\}$ be a set of distinct real values, and denote that $X_* = \hat{X} \cup X'$. Assume that all values in $X_* \cup Y$ are distinct and

$$\{R(x_1^*, X_* \cup Y), \ldots, R(x_{n-n'}^*, X_* \cup Y)\} = \mathcal{S}(k).$$

Then, according to Proposition B.2.6 and B.2.8, we have

$$T(X_*, Y) \leq T(X, Y),$$

for any possible values of $X \setminus X'$. Hence, it follows that

$$T(X_*, Y) \leq \min_{X \setminus X' \in \mathbb{R}^{n-n'}} T(X, Y).$$

Be definition, we have

$$\min_{X \setminus X' \in \mathbb{R}^{n-n'}} T(X, Y) \leq T(X_*, Y).$$

Hence,

$$\min_{X \setminus X' \in \mathbb{R}^{n-n'}} T(X, Y) = T(X_*, Y),$$

where

$$T(X_*, Y) = \begin{cases} T(X', Y) + (n^2 - n'^2)/4 & \text{if } \mathcal{C}(1) \text{ or } \mathcal{C}(4) \text{ holds,} \\ T(X', Y) + (n^2 - n'^2 - 1)/4 & \text{if } \mathcal{C}(2) \text{ holds,} \\ T(X', Y) + (n^2 - n'^2 + 1)/4 & \text{if } \mathcal{C}(3) \text{ holds.} \end{cases}$$

according to Proposition B.2.11. Hence, we conclude our result. □

B.3. Lower bound with multiple missing values in Y

The previous section considers the tight lower bound of the Ansari-Bradley test statistic when multiple values can be missing in X. This section considers similar problem, but when multiple values in Y can be missing, while all values in X are observed. The main results in this section is Theorem B.3.8.

We first show that the sum of the Ansari-Bradley test statistics in both directions $T(X, Y) + T(Y, X)$ is a constant depending only on the total sample sizes $N = n + m$.

Lemma B.3.1. Suppose $X = \{x_1, \ldots, x_n\}$ and $Y = \{y_1, \ldots, y_m\}$ are samples of distinct real values. Denote $N = n + m$. Then, the sum of the Ansari-Bradley test statistics $T(X, Y)$ and $T(Y, X)$ is a constant of sample size N. More specifically,

$$T(X, Y) + T(Y, X) = \sum_{i=1}^{N} \left| i - \frac{N+1}{2} \right| = \begin{cases} N^2/4 & N \text{ is even,} \\ (N^2 - 1)/4 & N \text{ is odd.} \end{cases}$$

Proof. By definition of the Ansari-Bradley test statistic, we have

$$T(X,Y) = \sum_{i=1}^{n} \left| R(x_i, X \cup Y) - \frac{N+1}{2} \right|,$$
$$\text{and } T(Y,X) = \sum_{i=1}^{m} \left| R(y_i, X \cup Y) - \frac{N+1}{2} \right|.$$

Since X and Y are samples of distinct values, each value in X and Y take a distinct rank in $\{1, \ldots, N\}$. Hence, we have our desired result

$$T(X,Y) + T(Y,X) = \sum_{i=1}^{N} \left| i - \frac{N+1}{2} \right|.$$

When N is odd,

$$\begin{aligned}
&\sum_{i=1}^{N} \left| i - \frac{1}{2}(N+1) \right| \\
&= \left(\frac{1}{2}(N+1) - 1 \right) + \ldots + \left(\frac{1}{2}(N+1) - \frac{1}{2}(N+1) \right) + \ldots + \left(N - \frac{1}{2}(N+1) \right) \\
&= \left(\frac{1}{2}(N-1) \right) + \ldots + 0 + \ldots + \left(\frac{1}{2}(N-1) \right) \\
&= \left(1 + \frac{1}{2}(N-1) \right) \frac{1}{2}(N-1) \\
&= \frac{N^2 - 1}{4}.
\end{aligned}$$

However, when N is even,

$$\begin{aligned}
&\sum_{i=1}^{N} \left| i - \frac{1}{2}(N+1) \right| \\
&= \left(\frac{1}{2}(N+1) - 1 \right) + \ldots + \left(\frac{1}{2}(N+1) - \frac{1}{2}N \right) + \frac{1}{2} + \ldots + \frac{1}{2}(N-1) \\
&= \left(\frac{1}{2} + \frac{1}{2}(N-1) \right) \frac{1}{2}N \\
&= \frac{N^2}{4}.
\end{aligned}$$

Hence, we complete our proof. □

This lemma shows that in order to minimise the Ansari-Bradley statistic $T(X,Y)$ when values in Y can be missing, we only need to maximise $T(Y,X)$ since $T(X,Y) + T(Y,X)$ is a constant.

We approach this problem by first considering the case when only one value in Y can be missing.

Lemma B.3.2. Suppose $X = \{x_1, \ldots, x_n\}$ and $Y = \{y_1, \ldots, y_m\}$ are samples of distinct real values. Denote $N = n + m$. Suppose $y \in Y$ is a value in Y, and denote $Y' = Y \setminus \{y\}$. Consider two real values y^*, y_*, and denote $Y_* = \{y^*\} \cup Y'$, and $Y_* = \{y_*\} \cup Y'$. Then, if

$$y_* < \min X \cup Y', \text{ and } y^* > \max X \cup Y',$$

we have $T(Y, X) \leq \max\{T(Y, X_*), T(Y, X_*)\}$.

Proof. This result follows from Proposition B.1.4. Suppose $y \geq (N+1)/2$. Then, according to Proposition B.1.4, we have $T(Y, X) \leq T(Y_*, X)$. However, if $y < (N+1)/2$, then according to Proposition B.1.4, we have $T(Y, X) \leq T(Y_*, X)$. Overall, we have

$$T(Y, X) \leq \max\{T(Y, X_*), T(Y, X_*)\}$$

as required, which completes our proof. □

Then, we generalise the result in Lemma B.3.2 into the cases where multiple values in Y can be missing. We first make the following definition:

Definition B.3.3. Suppose X is a sample of distinct real values. We define $\mathcal{X}(X, t)$ to be a collection of sets of t distinct values such that if $\hat{X} \in \mathcal{X}(X, t)$, then all values in $\hat{X}$ are either smaller than $\min X$ or larger than $\max X$. That is,

$$\mathcal{X}(X, t) = \{\{x_1^*, \ldots, x_t^*\} : \ x_i^* \notin [\min X, \max X], \text{ and } x_i^* \neq x_j^* \text{ for } i \neq j\}.$$

Applying Definition B.3.3, we generalise Lemma B.3.2 into the following result:

Lemma B.3.4. Suppose $X = \{x_1, \ldots, x_n\}$ and $Y = \{y_1, \ldots, y_m\}$ are samples of distinct real values. Let $Y' \subset Y$ be a non-empty subset of Y with sample size $|Y'| = m'$. Then there exist $\hat{Y} \in \mathcal{X}(X \cup Y', m - m')$ such that $T(\hat{Y} \cup Y', X) \geq T(Y, X)$.

Proof. We prove this result by mathematical induction on $m - m'$. The base case when $m - m' = 1$ is proved in Lemma B.3.2 since $\{y^*\}, \{y_*\} \in \mathcal{X}(X \cup Y', 1)$. For the induction step, let $k \in \{1, \ldots, m-1\}$, and assume the lemma is correct when $m - m' = k$. We show that it still holds when $m - m' = k + 1$.

Denote $Y \setminus Y' = \tilde{Y} = \{\tilde{y}_1, \ldots, \tilde{y}_{k+1}\}$. By assumption that Lemma B.3.4 is correct when $m - m' = k$, there exist $Y_* \in \mathcal{X}(X \cup \{\tilde{y}_1\} \cup Y', k)$ such that

$$T(Y_* \cup \{\tilde{y}_1\} \cup Y', X) \geq T(Y, X).$$

Further, consider two values y^*, y_* such that

$$y_* < \min X \cup Y_* \cup Y', \text{ and } y^* > \max X \cup Y_* \cup Y'.$$

Then according to Lemma B.3.2, we have

$$\begin{aligned} &\max\{T(\{y_*\} \cup Y_* \cup Y', X), T(\{y^*\} \cup Y_* \cup Y', X)\} \geq T(Y_* \cup \{\tilde{y}_1\} \cup Y', X) \\ \implies &\max\{T(\{y_*\} \cup Y_* \cup Y', X), T(\{y^*\} \cup Y_* \cup Y', X)\} \geq T(Y, X). \end{aligned}$$

Since $Y_* \in \mathcal{X}(X \cup \{\tilde{y}_1\} \cup Y', k)$, and we have that $y_* < \min Y_* \cup Y' \cup X$ and $y^* > \max Y_* \cup Y' \cup X$. Hence we have $\{y^*\} \cup Y_* \in \mathcal{X}(Y' \cup X, k+1)$, and $\{y_*\} \cup Y_* \in \mathcal{X}(Y' \cup X, k+1)$, which completes our proof. □

The following Definition B.3.5 further defines subsets of the set $\mathcal{X}(X, t)$ in Definition B.3.3 subjecting to a integer k.

Definition B.3.5. Suppose X is a sample of distinct real values, and $\hat{X} \in \mathcal{X}(X,t)$, where t is a positive integer. We define $\mathcal{Y}(X,t,k)$ to be a subset of $\mathcal{X}(X,t)$ such that if $\hat{X} \in \mathcal{Y}(X,t,k)$, we have $\hat{X} \in \mathcal{X}(X,t)$, and $\sum_{x\in\hat{X}} \mathrm{I}\left(x > \max X\right) = k$. In other words,

$$\mathcal{Y}(X,t,k) = \mathcal{X}(X,t) \cap \left\{ \{x_1^*,\ldots,x_t^*\} : \sum_{i=1}^{t} \mathrm{I}\left(x_i^* > \max X\right) = k \right\}.$$

The following Definition B.3.6 defines a function which is important for stating the results in Proposition B.3.7.

Definition B.3.6. Suppose $X = \{x_1,\ldots,x_n\}$ and $Y' = \{y_1,\ldots,y_{m'}\}$ are samples of distinct real values, and $m \in \mathbb{N}$ is a positive integer such that $m \geq m'$. Denote $Z' = X \cup Y'$, $N' = (n+m'+1)/2$, and $M' = (m-m')/2$. For any given $k \in \{0,\ldots,m-m'\}$, denote $a = \min\{0, k-M'\}$, $b = \max\{0,k-M'\}$. For any $i \in \{1,\ldots,n\}$, let us denote $A_i = \mathrm{I}\left(R(x_i,Z') < N'+a\right), B_i = \mathrm{I}\left(N'+a \leq R(x_i,Z') \leq N'+b\right)$, and $C_i = \mathrm{I}\left(R(x_i,Z') > N'+b\right)$. Then, we define

$$f_{X,Y',m}(k) = (k-M')\sum_{i=1}^{n}(A_i - C_i) + \mathrm{sgn}(k-M')\sum_{i=1}^{n} B_i\left(2N'+k-M'-2R(x_i,Z')\right),$$

where $\mathrm{sgn}(x) = 1$ if $x \geq 0$, otherwise $\mathrm{sgn}(x) = -1$.

The following proposition provides explicit form of the Ansari-Bradley statistic when the missing values in Y are imputed with $\hat{Y} \in \mathcal{Y}(X \cup Y', m-m', k)$.

Proposition B.3.7. Suppose $X = \{x_1,\ldots,x_n\}$ and $Y = \{y_1,\ldots,y_m\}$ are samples of distinct real values and $Y' \subset Y$ is a subset of Y with sample size $|Y'| = m'$. Then, if $\hat{Y} \in \mathcal{Y}(X \cup Y', m-m', k)$, we have $T(X,Y_*) = T(X,Y') + f_{X,Y',m}(k)$, where $Y_* = \hat{Y} \cup Y'$, and $f_{X,Y',m}(k)$ is defined in Definition B.3.6.

Proof. Since $\hat{Y} \in \mathcal{Y}(X \cup X', m-m', k)$, according to the definition of $\hat{Y}$ in Definition B.3.5, we have $y_i^* < \min X \cup Y'$ or $y_i^* > \max X \cup Y'$, for any $i \in \{1,\ldots,m-m'\}$. Notice that we also have $k = \sum_{i=1}^{m-m'} \mathrm{I}\left(y_i^* > \max X \cup Y'\right)$. In other words, the ranks of values of $\hat{Y}$ take k maximum, and $m-m'-k$ minimum distinct values in $\{1,\ldots,N\}$, i.e.,

$$\left\{R(y_1^*, X \cup Y_*),\ldots,R(y_{m-m'}^*, X \cup Y_*)\right\} = \left\{1,\ldots,m-m'-k\right\} \cup \left\{N-k+1,\ldots,N\right\}.$$

Let us divide $\hat{Y}$ into $\hat{Y}_1$ and $\hat{Y}_2$ such that values in $\hat{Y}_1$ take the minimum ranks, while values in $\hat{Y}_2$ take the maximum ranks. That is,

$$\hat{Y}_1 = \left\{ y \in \hat{Y} | 1 \leq R(y, X \cup Y_*) \leq m-m'-k \right\},$$
$$\text{and } \hat{Y}_2 = \left\{ y \in \hat{Y} | N-k+1 \leq R(y, X \cup Y_*) \leq N \right\}.$$

According to the definition of the Ansari-Bradley test statistic,

$$\begin{aligned} T(X, Y_*) &= \sum_{i=1}^{n} \left| R(x_i, X \cup Y_*) - \frac{1}{2}(N+1) \right| \\ &= \sum_{i=1}^{n} \left| R(x_i, X \cup Y') + \sum_{y \in \hat{Y}} \mathrm{I}\left(x_i > y\right) - \frac{1}{2}(N+1) \right| \\ &= \sum_{i=1}^{n} \left| R(x_i, X \cup Y') + \sum_{y \in \hat{Y}_1} \mathrm{I}\left(x_i > y\right) + \sum_{y \in \hat{Y}_2} \mathrm{I}\left(x_i > y\right) - \frac{1}{2}(N+1) \right|. \end{aligned}$$

By definitions of $\hat{Y}_1$ and $\hat{Y}_2$, we have

$$\sum_{y \in \hat{Y}_1} \mathrm{I}\left(x_i > y\right) = m - m' - k, \text{ and } \sum_{y \in \hat{Y}_2} \mathrm{I}\left(x_i > y\right) = 0.$$

Hence,

$$T(X, Y_*) = \sum_{i=1}^{n} \left| R(x_i, X \cup Y') + m - m' - k - \frac{1}{2}(N+1) \right|.$$

Notice that

$$\frac{1}{2}(N+1) = \frac{1}{2}(n + m - m' + m' + 1) = \frac{1}{2}(n + m' + 1) + \frac{1}{2}(m - m').$$

Hence, we further have

$$\begin{aligned} T(X, Y_*) &= \sum_{i=1}^{n} \left| R(x_i, X \cup Y') + m - m' - k - \frac{1}{2}(n + m' + 1) - \frac{1}{2}(m - m') \right| \\ &= \sum_{i=1}^{n} \left| R(x_i, X \cup Y') - \frac{1}{2}(n + m' + 1) + \frac{1}{2}(m - m') - k \right|. \end{aligned}$$

For notation ease, let us denote $N' = (n + m' + 1)/2$, $M' = (m - m')/2$ and

$$\chi_i = \left| R(x_i, X \cup Y') - N' + M' - k \right|, \text{ for any } i \in \{1, \ldots, n\}.$$

Denote $a = \min\{0, k - M'\}$, and $b = \max\{0, k - M'\}$. For any $i \in \{1, \ldots, n\}$, also denote that

$$\begin{aligned} A_i &= \mathrm{I}\left(R(x_i, X \cup Y') < N' + a\right), \\ B_i &= \mathrm{I}\left(N' + a \le R(x_i, X \cup Y') \le N' + b\right), \\ C_i &= \mathrm{I}\left(R(x_i, X \cup Y') > N' + b\right). \end{aligned}$$

Note that $A_i + B_i + C_i = 1$. Thus, we have

$$T(X, Y_*) = \sum_{i=1}^{n} A_i \chi_i + \sum_{i=1}^{n} B_i \chi_i + \sum_{i=1}^{n} C_i \chi_i. \tag{B.15}$$

Below we prove our results for the following two cases separately:

$$\text{Case (i)} : k - M' \geq 0, \quad \text{Case (ii)} : k - M' < 0.$$

Suppose case (i) holds. That is, assume that $k - M' \geq 0$. Then, we have $a = 0$ and $b = k - M'$. For the first term on the right hand side of the Equation (B.15), notice that

$$A_i = 1 \implies R(x_i, X \cup Y') < N' \implies R(x_i, X \cup Y') < N' + k - M'.$$

Hence, when $A_i = 1$, we have $\chi_i = (N' - M' + k - R(x_i, X \cup Y'))$. Therefore,

$$\begin{aligned}\sum_{i=1}^{n} A_i \chi_i &= \sum_{i=1}^{n} A_i \left(N' - M' + k - R(x_i, X \cup Y')\right) \\ &= \sum_{i=1}^{n} A_i \left(N' - R(x_i, X \cup Y')\right) + \sum_{i=1}^{n} A_i \left(k - M'\right).\end{aligned}$$

Similarly, for the second term on the right hand side of the Equation (B.15), when $B_i = 1$, we have

$$R(x_i, X \cup Y') \leq N' + k - M' \implies \chi_i = \left(N' - M' + k - R(x_i, X \cup Y')\right).$$

Hence,

$$\begin{aligned}\sum_{i=1}^{n} B_i \chi_i &= \sum_{i=1}^{n} B_i \left(N' - M' + k - R(x_i, X \cup Y')\right) \\ &= \sum_{i=1}^{n} B_i \left(R(x_i, X \cup Y') - N'\right) + \sum_{i=1}^{n} B_i \left(2N' - M' + k - 2R(x_i, X \cup Y')\right).\end{aligned}$$

Similarly, for the third term on the right hand side of the Equation (B.15), when $C_i = 1$, we have

$$R(x_i, X \cup Y') > N' + k - M' \implies \chi_i = \left(R(x_i, X \cup Y') - N' + M' - k\right).$$

Hence, we have

$$\begin{aligned}\sum_{i=1}^{n} C_i \chi_i &= \sum_{i=1}^{n} C_i \left(R(x_i, X \cup Y') - N' + M' - k\right) \\ &= \sum_{i=1}^{n} C_i \left(R(x_i, X \cup Y') - N'\right) + \sum_{i=1}^{n} C_i \left(M' - k\right).\end{aligned}$$

Combining the above results for the first, second and third terms on the right hand side of the

Equation (B.15) together,

$$\begin{aligned}
T(X, Y_*) = &\sum_{i=1}^{n} A_i \left(N' - R(x_i, X \cup Y')\right) + \sum_{i=1}^{n} A_i \left(k - M'\right) \\
&+ \sum_{i=1}^{n} B_i \left(R(x_i, X \cup Y') - N'\right) + \sum_{i=1}^{n} B_i \left(2N' - M' + k - 2R(x_i, X \cup Y')\right) \\
&+ \sum_{i=1}^{n} C_i \left(R(x_i, X \cup Y') - N'\right) + \sum_{i=1}^{n} C_i \left(M' - k\right).
\end{aligned}$$

According to the definition of the Ansari-Bradley test statistic,

$$\begin{aligned}
T(X, Y') &= \sum_{i=1}^{n} \left| R(x_i, X \cup Y') - \frac{1}{2}(n + m' + 1) \right| \\
&= \sum_{i=1}^{n} \mathrm{I}\left(R(x_i, X \cup Y') < N'\right) \left(N' - R(x_i, X \cup Y')\right) \\
&+ \sum_{i=1}^{n} \mathrm{I}\left(R(x_i, X \cup Y') \geq N'\right) \left(R(x_i, X \cup Y') - N'\right) \\
&= \sum_{i=1}^{n} A_i \left(N' - R(x_i, X \cup Y')\right) + \sum_{i=1}^{n} (B_i + C_i) \left(R(x_i, X \cup Y') - N'\right).
\end{aligned}$$

Therefore, we have

$$\begin{aligned}
T(X, Y_*) &= T(X, Y') + \sum_{i=1}^{n} A_i \left(k - M'\right) \\
&\quad + \sum_{i=1}^{n} B_i \left(2N' - 2M' + k - 2R(x_i, X \cup Y')\right) + \sum_{i=1}^{n} C_i \left(M' - k\right) \\
&= T(X, Y') + (k - M') \sum_{i=1}^{n} (A_i - C_i) + \sum_{i=1}^{n} B_i \left(2N' + k - 2M' - 2R(x_i, X \cup Y')\right) \\
&= T(X, Y') + f_{X, Y', m}(k),
\end{aligned}$$

where the last equation holds because $k - M' \geq 0$. This completes our proof for case (i). The result for case (ii) can be proved similarly.

Suppose case (ii) holds. That is, assume that $k - M' \leq 0$. Then we have $a = k - (m - m')/2$, and $b = 0$. Hence, for the first term on the right hand side of the Equation (B.15), we have

$$\begin{aligned}
\sum_{i=1}^{n} A_i \chi_i &= \sum_{i=1}^{n} A_i \left(N' - M' + k - R(x_i, X \cup Y')\right) \\
&= \sum_{i=1}^{n} A_i \left(N' - R(x_i, X \cup Y')\right) + \sum_{i=1}^{n} A_i \left(k - M'\right).
\end{aligned}$$

Similarly, for the second term on the right hand side of the Equation (B.15),

$$\begin{aligned}\sum_{i=1}^{n} B_i\chi_i &= \sum_{i=1}^{n} B_i \left(R(x_i, X\cup Y') - N' + M' - k\right)\\ &= \sum_{i=1}^{n} B_i \left(N' - R(x_i, X\cup Y')\right) + \sum_{i=1}^{n} B_i \left(2R(x_i, X\cup Y') - 2N' + M' - k\right),\end{aligned}$$

and for the second term on the right hand side of the Equation (B.15),

$$\begin{aligned}\sum_{i=1}^{n} C_i\chi_i &= \sum_{i=1}^{n} C_i \left(R(x_i, X\cup Y') - N' + M' - k\right)\\ &= \sum_{i=1}^{n} C_i \left(R(x_i, X\cup Y') - N'\right) + \sum_{i=1}^{n} C_i \left(M' - k\right).\end{aligned}$$

Hence, we have

$$\begin{aligned}T(X, Y_*\cup Y') = &\sum_{i=1}^{n} A_i \left(N' - R(x_i, X\cup Y')\right) + \sum_{i=1}^{n} A_i \left(k - M'\right)\\ &+ \sum_{i=1}^{n} B_i \left(N' - R(x_i, X\cup Y')\right) + \sum_{i=1}^{n} B_i \left(2R(x_i, X\cup Y') - 2N' + M' - k\right)\\ &+ \sum_{i=1}^{n} C_i \left(R(x_i, X\cup Y') - N'\right) + \sum_{i=1}^{n} C_i \left(M' - k\right).\end{aligned}$$

Notice that

$$\begin{aligned}T(X, Y') &= \sum_{i=1}^{n} \left|R(x_i, X\cup Y') - \frac{1}{2}(n + m' + 1)\right|\\ &= \sum_{i=1}^{n} \mathrm{I}\left(R(x_i, X\cup Y') \le N'\right)\left(N' - R(x_i, X\cup Y')\right)\\ &+ \sum_{i=1}^{n} \mathrm{I}\left(R(x_i, X\cup Y') > N'\right)\left(R(x_i, X\cup Y') - N'\right)\\ &= \sum_{i=1}^{n} (A_i + B_i)\left(N' - R(x_i, X\cup Y')\right) + \sum_{i=1}^{n} C_i \left(R(x_i, X\cup Y') - N'\right).\end{aligned}$$

Therefore, we have

$$\begin{aligned}T(X, Y_*) &= T(X, Y') + \sum_{i=1}^{n} A_i \left(k - M'\right)\\ &\quad + \sum_{i=1}^{n} B_i \left(2R(x_i, X\cup Y') - 2N' + M' - k\right) + \sum_{i=1}^{n} C_i \left(M' - k\right)\\ &= T(X, Y') + (A_i - C_i)\sum_{i=1}^{n} \left(k - M'\right) - \sum_{i=1}^{n} B_i \left(2N' + k - 2M' - 2R(x_i, X\cup Y')\right)\\ &= T(X, Y') + f_{X,Y',m}(k),\end{aligned}$$

where the last equation holds because $k - M' < 0$. This completes our proof for case (ii). $\square$

We complete our results for the lower bounds of the Ansari-Bradley statistic when multiple values in Y can be missing using the following result:

Theorem B.3.8. Suppose $X = \{x_1, \ldots, x_n\}$ and $Y = \{y_1, \ldots, y_m\}$ are samples of distinct real values and $Y' \subset Y$ is a subset of Y with sample size $|Y'| = m'$, which are observed. Then, the minimum possible Ansari-Bradley test statistic, across all unobserved values, is given as follows:

$$\min_{Y \setminus Y' \in \mathbb{R}^{m-m'}} T(X, Y) = T(X, Y') + \min_{k \in \{0, \ldots, m-m'\}} f_{X,Y',m}(k),$$

where $f_{X,Y',m}(k)$ is defined in Definition B.3.6.

Proof. To start, let us assume that when the unobserved values $Y \setminus Y' = \tilde{Y}$, the Ansari-Bradley test statistic takes the minimum value. In other words,

$$T(X, \tilde{Y} \cup Y') = \min_{Y \setminus Y' \in \mathbb{R}^{m-m'}} T(X, Y).$$

Now, consider $m - m' + 1$ different sets $\hat{Y}^{(1)}, \ldots, \hat{Y}^{(m-m'+1)} \in \mathcal{X}(X \cup Y', m - m')$, such that

$$k = \sum_{y \in \hat{Y}^{(k)}} \mathrm{I}\left(y > \max X \cup Y'\right), \text{ where } k \in \{1, \ldots, m - m' + 1\}.$$

By definition, we have

$$T(X, \tilde{Y} \cup Y') = \min_{Y \setminus Y' \in \mathbb{R}^{m-m'}} T(X, Y) \leq \min\{T(X, Y' \cup \hat{Y}^{(1)}), \ldots, T(X, Y' \cup \hat{Y}^{(m-m'+1)})\}.$$

Below we show that

$$T(X, \tilde{Y} \cup Y') \geq \min\{T(X, Y' \cup \hat{Y}^{(1)}), \ldots, T(X, Y' \cup \hat{Y}^{(m-m'+1)})\}.$$

Note that according to Lemma B.3.4, there exist a set $\hat{Y} \in \mathcal{X}(X \cup Y', m - m')$ such that $T(\hat{Y} \cup Y', X) \geq T(\tilde{Y} \cup Y', X)$. By Lemma B.3.1,

$$T(\hat{Y} \cup Y', X) + T(X, \hat{Y} \cup Y') = T(X, \tilde{Y} \cup Y') + T(\tilde{Y} \cup Y', X) = \sum_{i=1}^{N} \left| i - \frac{N+1}{2} \right|.$$

Then we also have

$$\begin{aligned} & T(X, \hat{Y} \cup Y') \leq T(X, \tilde{Y} \cup Y') \\ \implies & \min\{T(X, Y' \cup \hat{Y}^{(1)}), \ldots, T(X, Y' \cup \hat{Y}^{(m-m'+1)})\} \leq T(X, \tilde{Y} \cup Y'). \end{aligned}$$

Hence, we have our desired result

$$\begin{aligned} T(X, \tilde{Y} \cup Y') &= \min\{T(X, Y' \cup \hat{Y}^{(1)}), \ldots, T(X, Y' \cup \hat{Y}^{(m-m'+1)})\} \\ &= T(X, Y') + \min_{k \in \{0, \ldots, m-m'\}} f_{X,Y',m}(k), \end{aligned}$$

where the last equation follows from Proposition B.3.7. □

B.4. Lower and upper bounds under general missingness case

This section combines the results from Section B.2 and B.3 when only values in X can be missing, and only values in Y can be missing, respectively. We derive the tight lower and upper bounds of the Ansari-Bradley test statistic in the presence of missing data without assuming any missingness patterns, as shown in Theorem B.4.7 in the end of this section.

We start by proving the following lemma, which shows that if we replace a subset $\tilde{Y}$ of Y such that for any $y \in \tilde{Y}$, $R(y, X \cup Y) \leq (N+1)/2$ with $\hat{Y} = \mathcal{Y}(X \cup Y', m - m', 0)$, the values of the Ansari-Bradley statistic can only decrease or remain unchanged.

Lemma B.4.1. Suppose $X = \{x_1, \ldots, x_n\}$ and $Y = \{y_1, \ldots, y_m\}$ are samples of distinct real values. Denote $N = n + m$. Suppose $\tilde{Y} \subset Y$ is a non-empty subset of Y such that for any $y \in \tilde{Y}$, $R(y, X \cup Y) \leq (N+1)/2$. Denote $Y' = Y \setminus \tilde{Y}$, and $|Y'| = m'$. Then, if $\hat{Y} = \mathcal{Y}(X \cup Y', m - m', 0)$, we have $T(X, Y_*) \leq T(X, Y)$, where $Y_* = \hat{Y} \cup Y'$.

Proof. We prove the result by mathematical induction on $m - m'$. Let us denote $\hat{Y} = \{y_1^*, \ldots, y_{m-m'}^*\}$. For the base case when $m - m' = 1$, since $\hat{Y} = \mathcal{Y}(X \cup Y', 1, 0)$, we have $y_1^* < \max X \cup Y'$. Hence, $R(y_1^*, X \cup Y_*) = 1$. Then, according to Proposition B.1.4, we have $T(Y, X) \leq T(Y_*, X)$. By applying Lemma B.3.1, we have our desired result $T(X, Y_*) \leq T(X, Y)$.

Let $k \in \{1, \ldots, m-1\}$, and suppose that Lemma B.4.1 holds for $m - m' = k$. We show that the lemma also holds when $m - m' = k + 1$.

Denote $\tilde{Y} = \{\tilde{y}_1, \ldots, \tilde{y}_{k+1}\}$, and without loss of generality, let us assume $\tilde{y}_1 = \max \tilde{Y}$. Consider a set $\hat{Z} = \{z_1^*, \ldots, z_k^*\}$ and assume that $\hat{Z} \in \mathcal{Y}(X \cup Y' \cup \{\tilde{y}_1\}, k, 0)$. For notation convenience, denote $Z = \hat{Z} \cup Y' \cup \{\tilde{y}_1\}$. Then, by the assumption that Lemma B.4.1 is correct when $m - m' = k$, we have

$$T(X, Z) \leq T(X, Y). \tag{B.16}$$

Since $\tilde{y}_1 = \max \tilde{Y}$, we have

$$R(\tilde{y}_1, X \cup Y) > \max\{R(\tilde{y}_2, X \cup Y), \ldots, R(\tilde{y}_{k+1}, X \cup Y)\} \geq k,$$

where the last " $\geq$ " holds because $\{R(\tilde{y}_2, X \cup Y), \ldots, R(\tilde{y}_{k+1}, X \cup Y)\}$ are k distinct values in $\{1, \ldots, N\}$. Then, since $\hat{Z} \in \mathcal{Y}(X \cup Y' \cup \{\tilde{y}_1\}, k, 0)$, we have

$$\begin{aligned}
& z_i^* < \min X \cup Y' \cup \{\tilde{y}_1\}, \text{ for any } i \in \{1, \ldots, k\} \\
\implies & \{R(z_1^*, X \cup Z), \ldots, R(z_k^*, X \cup Z)\} = \{1, \ldots, k\} \\
\implies & R(\tilde{y}_1, X \cup Y) > \max\{R(z_1^*, X \cup Z), \ldots, R(z_k^*, X \cup Z)\}.
\end{aligned}$$

Then, according to Lemma B.2.2,

$$R(\tilde{y}_1, X \cup Z) = R(\tilde{y}_1, X \cup Y) < \frac{1}{2}(N+1).$$

Next, consider any real value $\{y^*\} \in \mathcal{Y}(Y' \cup \hat{Z}, 1, 0)$, and denote $Z' = Y' \cup \hat{Z} \cup \{y^*\}$. Then, by applying the lemma when $m - m' = 1$, we have

$$T(X, Z') \leq T(X, Z) \leq^{\text{Inequality (B.16)}} T(X, Y).$$

Now, in order to finish our proof, we only need to show that $\hat{Z} \cup \{y^*\} \in \mathcal{Y}(X \cup Y', k+1, 0)$.

Since $\hat{Z} \in \mathcal{Y}(X \cup Y' \cup \{\tilde{y}_1\}, k, 0)$, we have $z_i^* < \min X \cup Y'$, for any $i \in \{1, \ldots, k\}$. Combining this result with $\{y^*\} \in \mathcal{Y}(Y' \cup \hat{Z}, 1, 0)$, hence we have $\hat{Z} \cup \{y^*\} \in \mathcal{Y}(X \cup Y', k+1, 0)$, which completes our proof. □

The following lemma provides the same conclusion as Lemma B.4.1, but now we replace the subset $\tilde{Y}$ of Y such that for any $y \in \tilde{Y}$, $R(y, X \cup Y) \geq (N+1)/2$ with $\hat{Y} = \mathcal{Y}(X \cup Y', m - m', 0)$,

Lemma B.4.2. Suppose $X = \{x_1, \ldots, x_n\}$ and $Y = \{y_1, \ldots, y_m\}$ are samples of distinct real values. Denote $N = n + m$. Suppose $\tilde{Y} \subset Y$ is a non-empty subset of Y such that for any $y \in \tilde{Y}$, $R(y, X \cup Y) \geq (N+1)/2$. Denote $Y' = Y \setminus \tilde{Y}$, and $|Y'| = m'$. Then, if $\hat{Y} = \mathcal{Y}(X \cup Y', m - m', m - m')$, we have $T(X, Y_*) \leq T(X, Y)$, where $Y_* = \hat{Y} \cup Y'$.

Proof. This result can be prove following the same approach for proving Lemma B.4.1. Thus it is omitted here. □

The following definition is used in Proposition B.4.4.

Definition B.4.3. Suppose X and Y are samples of distinct values, t_1 and t_2 are positive integers, and $k \in \{1, 2, 3, 4\}$. Then we define

$$\mathcal{Z}(X, Y, t_1, t_2, k) = \left\{ (\hat{X}, \hat{Y}) : \mathcal{R} = \mathcal{S}(k), \hat{Y} \in \mathcal{X}(X_* \cup Y, t_2) \right\},$$

where $\mathcal{R} = \{R(x_1^*, X_* \cup Y_*), \ldots, R(x_{t_1}^*, X_* \cup Y_*)\}$, $\hat{X} = \{x_1^*, \ldots, x_{t_1}^*\}$, $X_* = \hat{X} \cup X$, and $Y_* = \hat{Y} \cup Y$.

Applying Definition B.4.3, we show that there exist imputations $\hat{X}, \hat{Y} \in \mathcal{Z}(X', Y', n - n', m - m', i)$ that minimises the Spearman's footrule, when N and $n - n'$ satisfy condition $\mathcal{C}(i)$ for $i = \{1, 2, 3, 4\}$.

Proposition B.4.4. Suppose $X = \{x_1, \ldots, x_n\}$ and $Y = \{y_1, \ldots, y_m\}$ are samples of distinct real values. Let $X' \subset X, Y' \subset Y$ be observed subsets of X, and Y with sample sizes $|X'| = n'$, and $|Y'| = m'$. Denote $N = n + m$, and suppose N and $n - n'$ satisfy condition $\mathcal{C}(i)$, where $i \in \{1, 2, 3, 4\}$. Then there exist $\hat{X}, \hat{Y} \in \mathcal{Z}(X', Y', n - n', m - m', i)$ such that $T(X_*, Y_*) \leq T(X, Y)$, where $X_* = \hat{X} \cup X'$, and $Y_* = \hat{Y} \cup Y'$.

Proof. To start, let $\hat{X} = \{x_1^*, \ldots, x_{n-n'}^*\}$ be a set of distinct real values. Denote $X_* = \hat{X} \cup X'$. Assume that values in $X_* \cup Y$ are distinct, and

$$\{R(x_1^*, X_* \cup Y), \ldots, R(x_{n-n'}^*, X_* \cup Y)\} = \mathcal{S}(i).$$

Then, according to Proposition B.2.6 and Proposition B.2.8, we have

$$T(X_*, Y) \leq T(X, Y), \tag{B.17}$$

for any $i \in \{1,2,3,4\}$.

Denote $\tilde{Y} = Y \setminus Y'$. Let $s_* = \min \mathcal{S}(i)$ and $s^* = \max \mathcal{S}(i)$. Since values in X_* take the ranks between s_* and s^*, for any $y \in \tilde{Y}$, we have $R(y, X_* \cup Y) \geq s^* + 1$ or $R(y, X_* \cup Y) \leq s_* - 1$. By definition of $\mathcal{S}(i)$, for any $i \in \{1,2,3,4\}$, we have

$$\begin{aligned} s_* &\leq -\frac{n-n'}{2} + 1 + \frac{1}{2}(N+1) \leq -\frac{1}{2} + 1 + \frac{1}{2}(N+1) = \frac{1}{2} + \frac{1}{2}(N+1), \\ \text{and } s^* &\geq \frac{n-n'-1}{2} + \frac{1}{2}(N+1) \geq \frac{1}{2}(N+1). \end{aligned}$$

Hence, for any $y \in \tilde{Y}$, we have either

$$R(y, X_* \cup Y) \geq s^* + 1 > \frac{1}{2}(N+1), \text{ or } R(y, X_* \cup Y) \leq s_* - 1 < \frac{1}{2}(N+1).$$

Thus, we can partition $\tilde{Y} = \tilde{Y}_1 \cup \tilde{Y}_2$, where

$$\begin{aligned} \tilde{Y}_1 &= \left\{ y | y \in \tilde{Y}, R(y, X_* \cup Y) < \frac{1}{2}(N+1) \right\}, \\ \tilde{Y}_2 &= \left\{ y | y \in \tilde{Y}, R(y, X_* \cup Y) > \frac{1}{2}(N+1) \right\}. \end{aligned}$$

Denote $|\tilde{Y}_1| = l_1$, and $|\tilde{Y}_2| = l_2$. Without loss of generality, denote $\tilde{Y}_1 = \{\tilde{y}_1, \dots \tilde{y}_{l_1}\}$, $\tilde{Y}_2 = \{\tilde{y}_{l_1+1}, \dots \tilde{y}_{l_1+l_2}\}$, and $\tilde{Y} = \{\tilde{y}_1, \dots \tilde{y}_{l_1}, \tilde{y}_{l_1+1}, \dots, \tilde{y}_{l_1+l_2}\}$.

Suppose $\hat{Y}_1 \in \mathcal{Y}(X_* \cup (Y \setminus \tilde{Y}_1), l_1, 0)$. Then, according to Lemma B.4.1,

$$T(X_*, (Y \setminus \tilde{Y}_1) \cup \hat{Y}_1) \leq T(X_*, Y) \leq^{\text{Inequality (B.17)}} T(X, Y). \tag{B.18}$$

Since for any $y \in \tilde{Y}_1$, $R(y, X_* \cup Y) \leq s_* - 1$, and for any $x \in \hat{X}$, $R(x, X_* \cup Y) \geq s_*$. Hence, for any $x \in \hat{X}$, $R(x, X_* \cup Y) > \max\{R(y, X_* \cup Y) : y \in \tilde{Y}_1\}$. Since $|\tilde{Y}_1| = l_1$, and values in $\tilde{Y}_1$ takes distinct values in $\{1, \dots, N\}$, we have

$$R(x, X_* \cup Y) > \max\{R(y, X_* \cup Y) : y \in \tilde{Y}_1\} \geq l_1.$$

Note that $\hat{Y}_1 \in \mathcal{Y}(X_* \cup (Y \setminus \tilde{Y}_1), l_1, 0)$. Hence, for any $y \in \tilde{Y}_1$, we have $y < \min X_* \cup (Y \setminus \tilde{Y}_1)$, which implies

$$\left\{ R(\tilde{y}_1, X_* \cup (Y \setminus \tilde{Y}_1) \cup \hat{Y}_1), \dots, R(\tilde{y}_{l_1}, X_* \cup (Y \setminus \tilde{Y}_1) \cup \hat{Y}_1) \right\} = \{1, \dots, l_1\}.$$

Hence, for any $x \in \hat{X}$, we also have

$$R(x, X_* \cup Y) > \left\{ R(\tilde{y}_1, X_* \cup (Y \setminus \tilde{Y}_1) \cup \hat{Y}_1), \dots, R(\tilde{y}_{l_1}, X_* \cup (Y \setminus \tilde{Y}_1) \cup \hat{Y}_1) \right\}.$$

Then, by applying Lemma B.2.1,

$$R(x, X_* \cup Y) = R(x, X_* \cup (Y \setminus \tilde{Y}_1) \cup \hat{Y}_1), \text{ for any } x \in \hat{X}.$$

Therefore

$$\begin{aligned}&\left\{R(x_1^*, X_* \cup (Y \setminus \tilde{Y}_1) \cup \hat{Y}_1), \ldots, R(x_{n-n'}^*, X_* \cup (Y \setminus \tilde{Y}_1) \cup \hat{Y}_1)\right\}\\&= \{R(x_1^*, X_* \cup Y), \ldots, R(x_{n-n'}^*, X_* \cup Y)\}\\&= \mathcal{S}(i).\end{aligned}$$

Similarly, by applying Lemma B.2.1, we have

$$R(\tilde{y}, X_* \cup (Y \setminus \tilde{Y}_1) \cup \hat{Y}_1) = R(\tilde{y}, X_* \cup Y) > \frac{1}{2}(N+1), \text{ for any } y \in \tilde{Y}_2.$$

Suppose $\hat{Y}_2 \in \mathcal{Y}(X_* \cup Y' \cup \hat{Y}_1, l_2, l_2)$. Then, according to Lemma B.4.2,

$$T(X_*, Y' \cup \hat{Y}_1 \cup \hat{Y}_2) \leq T(X_*, (Y \setminus \tilde{Y}_1) \cup \hat{Y}_1) \leq^{\text{Inequality (B.18)}} T(X, Y).$$

Subsequently, in order to complete our proof, we need to show that $\hat{Y}_1 \cup \hat{Y}_2 \in \mathcal{X}(X_* \cup Y', l_1 + l_2)$, and $\left\{R(x_1^*, X_* \cup Y' \cup \hat{Y}_1 \cup \hat{Y}_2), \ldots, R(x_{n-n'}^*, X_* \cup Y' \cup \hat{Y}_1 \cup \hat{Y}_2)\right\} = \mathcal{S}(i)$.

For the first part, notice that $\hat{Y}_1 \in \mathcal{Y}(X_* \cup (Y \setminus \tilde{Y}_1), l_1, 0)$. Hence for any $y \in \hat{Y}_1$, we have $y < \min X_* \cup Y'$. Note also $\hat{Y}_2 \in \mathcal{Y}(X_* \cup Y' \cup \hat{Y}_1, l_2, l_2)$. Hence for any $y \in \hat{Y}_2$, we have $y > \max X_* \cup Y'$. Therefore, we have $\hat{Y}_1 \cup \hat{Y}_2 \in \mathcal{X}(X_* \cup Y', l_1 + l_2)$.

For the second part, note that $R(\tilde{y}, X_* \cup (Y \setminus \tilde{Y}_1) \cup \hat{Y}_1) > (N+1)/2$ for any $y \in \tilde{Y}_2$, and $\left\{R(x_1^*, X_* \cup (Y \setminus \tilde{Y}_1) \cup \hat{Y}_1), \ldots, R(x_{n-n'}^*, X_* \cup (Y \setminus \tilde{Y}_1) \cup \hat{Y}_1)\right\} = \mathcal{S}(i)$ takes consecutive ranks between s_* and s^*, where $s_* \leq (N+1)/2 + 1/2$. Since ranks are distinct values in $\{1, \ldots, N\}$, for any $x \in \hat{X}$, we have

$$R(x, X_* \cup (Y \setminus \tilde{Y}_1) \cup \hat{Y}_1) < \min\{R(y, X_* \cup (Y \setminus \tilde{Y}_1) \cup \hat{Y}_1) : y \in \tilde{Y}_2\}.$$

Note also that $\hat{Y}_2 \in \mathcal{Y}(X_* \cup Y' \cup \hat{Y}_1, l_2, l_2)$. By definition of $\mathcal{Y}$, we have

$$\begin{aligned}&y > \max X_* \cup Y' \cup \hat{Y}_1, \text{for any } y \in \hat{Y}_2\\\implies &\min\{R(y, X_* \cup Y' \cup \hat{Y}_1 \cup \hat{Y}_2) : y \in \tilde{Y}_2\} > R(x, X_* \cup (Y \setminus \tilde{Y}_1) \cup \hat{Y}_1), \text{ for any } x \in \hat{X}.\end{aligned}$$

Then, by applying Lemma B.2.2, we obtain desired result

$$\left\{R(x_1^*, X_* \cup Y' \cup \hat{Y}_1 \cup \hat{Y}_2), \ldots, R(x_{n-n'}^*, X_* \cup Y' \cup \hat{Y}_1 \cup \hat{Y}_2)\right\} = \mathcal{S}(i).$$

This completes our proof □

Before we present the final results, we prove the following two lemmas. The first lemma considers imputations $\hat{X}, \hat{Y} \in \mathcal{Z}(X', Y', n-n', m-m', i)$, and provides the minimum and maximum possible number of imputed values in Y that are larger than all values in $\hat{X} \cup X' \cup Y'$. In other words, the lemma provides lower and upper bounds for $k = \sum_{i=1}^{m-m'} \mathrm{I}\left(y_i^* > \max \hat{X} \cup X' \cup Y'\right)$.

Lemma B.4.5. Suppose $X = \{x_1, \ldots, x_n\}$ and $Y = \{y_1, \ldots, y_m\}$ are samples of distinct real values. Let $X' \subset X$ and $Y' \subset Y$ be observed subsets with sizes $|X'| = n'$ and $|Y'| = m'$. Denote $N = n + m$, and suppose $N, n - n'$ satisfy condition $\mathcal{C}(i)$, where $i \in \{1, 2, 3, 4\}$. Suppose $\hat{X}, \hat{Y} \in \mathcal{Z}(X', Y', n-n', m-m', i)$. Denote $\hat{Y} = \{y_1^*, \ldots, y_{m-m'}^*\}$, and $k = \sum_{i=1}^{m-m'} \mathrm{I}\left(y_i^* > \max \hat{X} \cup X' \cup Y'\right)$,

then we have

$$\begin{cases} a \le k \le b & \text{if } i \in \{1,3\}, \\ c \le k \le d & \text{if } i \in \{2,4\}, \end{cases}$$

where a, b, c, d are constants of sample sizes defined as

$$\begin{aligned} a &= \max\left\{(m - n' - 2m' + 1)/2, 0\right\}, \\ b &= \min\left\{(m + n' + 1)/2, m - m'\right\}, \\ c &= \max\left\{(m - n' - 2m')/2, 0\right\}, \\ \text{and } d &= \min\left\{(m + n')/2, m - m'\right\}. \end{aligned}$$

Proof. To start, consider the cases when $i \in \{1,3\}$. That is, suppose N and $n - n'$ satisfy $\mathcal{C}(1)$ or $\mathcal{C}(3)$, and $\hat{X}, \hat{Y} \in \mathcal{Z}(X', Y', n - n', m - m', 1)$, or $\hat{X}, \hat{Y} \in \mathcal{Z}(X', Y', n - n', m - m', 3)$. Denote $X_* = \hat{X} \cup X'$, $Y_* = \hat{Y} \cup Y'$, and $\mathcal{R} = \{R(x_1^*, X_* \cup Y_*), \ldots, R(x_{n-n'}^*, X_* \cup Y_*)\}$. Then, by definition of $\mathcal{Z}$, we have

$$\mathcal{R} = \mathcal{S}(i) = \left\{-\frac{n - n'}{2} + \frac{1}{2}(N+1), \ldots, \frac{n - n'}{2} - 1 + \frac{1}{2}(N+1)\right\},$$

where $i = \{1,3\}$. Denote $s_* = -(n - n')/2 + (N+1)/2$, and $s^* = \{(n - n')/2 + (N+1)/2, \ldots, N\}$, as $\min \mathcal{R}$ and $\max \mathcal{R}$, respectively. Let

$$\hat{Y}_1 = \left\{y \in \hat{Y} : y > \max X_* \cup Y'\right\}.$$

Then $|\hat{Y}_1| = k$. By definition of $\hat{Y}_1$, for any $y \in \hat{Y}_1$,

$$s^* < R(y, X_* \cup Y_*) \le N.$$

Note that

$$|\{s_* + 1, \ldots N\}| = N - s^* = \frac{2N - n + n' + 2 - N - 1}{2} = \frac{m + n' + 1}{2},$$

and since each value in $\hat{Y}_1$ take different values in $\{s_*+1, \ldots, N\}$, we have $k = |\hat{Y}_1| \le (m+n'+1)/2$. By definition of $\tilde{Y}_1$, we also have $k \le m - m'$. Hence, $k \le b$.

Using similar approach, we can show that $a \le k$. Let

$$\hat{Y}_2 = \left\{y \in \hat{Y} : y < \min X_* \cup Y'\right\}.$$

Since $\hat{X}, \hat{Y} \in \mathcal{Z}(X', Y', n - n', m - m', 1)$, or $\hat{X}, \hat{Y} \in \mathcal{Z}(X', Y', n - n', m - m', 3)$, we have $\hat{Y} \in \mathcal{X}(X_* \cup Y', m - m')$. By definition, for any $y \in \hat{Y}$, we have $y < \min X_* \cup Y'$ or $y > \max X_* \cup Y'$. Hence, $|\hat{Y}_2| = m - m' - k$. Also, by definition of $\hat{Y}_2$, for any $y \in \hat{Y}_2$,

$$1 \le R(y, X_* \cup Y_*) < s_*.$$

Note that

$$|\{1,\ldots,s_*-1\}| = s_* = -\frac{n-n'}{2} + \frac{1}{2}(N+1).$$

and since each value in $\hat{Y}_2$ take different values in $\{1,\ldots,s_*-1\}$, we have

$$\begin{aligned} |\hat{Y}_2| \leq |\{1,\ldots,s_*-1\}| &= -\frac{n-n'}{2} + \frac{1}{2}(N+1) - 1 \\ \implies k \geq m - m' + \frac{n-n'}{2} - \frac{1}{2}(N+1) + 1 &= \frac{m-n'-2m'+1}{2}. \end{aligned}$$

By definition, we also have $k \geq 0$. Hence, $k \geq a$, which proves our cases when $i \in \{1,3\}$.

The cases when $i \in \{2,4\}$ can be proved similarly. Thus, it is omitted here. □

The following lemma provides explicit form for the Ansari-Bradley statistic after being imputed with $(\hat{X},\hat{Y}) \in \mathcal{Z}(X',Y',n-n',m-m',i)$.

Lemma B.4.6. Suppose $X = \{x_1,\ldots,x_n\}$ and $Y = \{y_1,\ldots,y_m\}$ are samples of distinct real values. Let $X' \subset X, Y' \subset Y$ be subsets of X, Y with sample sizes $|X'| = n'$, and $|Y'| = m'$. Denote $N = n+m$, and suppose N and $n-n'$ satisfy $\mathcal{C}(i)$, where $i \in \{1,2,3,4\}$. Suppose $(\hat{X},\hat{Y}) \in \mathcal{Z}(X',Y',n-n',m-m',i)$, and denote $\sum_{i=1}^{m-m'} \mathrm{I}\left(y_i^* > \max \hat{X} \cup X' \cup Y'\right) = k$. Subsequently, we have

$$T(X_*,Y_*) = \begin{cases} T(X',Y') + f_{X',Y',m}(k) + (n^2-n'^2)/4 & \text{if} \quad i \in \{1,4\}, \\ T(X',Y') + f_{X',Y',m}(k) + (n^2-n'^2-1)/4 & \text{if} \quad i = 2, \\ T(X',Y') + f_{X',Y',m}(k) + (n^2-n'^2+1)/4 & \text{if} \quad i = 3, \end{cases}$$

where $X_* = X' \cup \hat{X}$ and $Y_* = \hat{Y} \cup Y'$.

Proof. To start, by applying Proposition B.2.11, we have

$$T(X_*,Y_* \cup Y') = \begin{cases} T(X',Y_*) + (n^2-n'^2)/4, & \text{if} \quad i \in \{1,4\}, \\ T(X',Y_*) + (n^2-n'^2-1)/4 & \text{if} \quad i = 2, \\ T(X',Y_*) + (n^2-n'^2+1)/4 & \text{if} \quad i = 3. \end{cases}$$

Then, by applying Proposition B.3.7 we have

$$T(X',Y_*) = T(X',Y') + f_{X',Y',m}(k).$$

Combining the above to results, we have our desired results. □

Finally, we prove the main results for the tight lower and upper bounds of the Ansari-Bradley statistics in the presence of missing data.

Theorem B.4.7. Suppose $X = \{x_1,\ldots,x_n\}$ and $Y = \{y_1,\ldots,y_m\}$ are samples of distinct real values. Let $X' \subset X$ and $Y' \subset Y$ be observed subsets with sizes $|X'| = n'$ and $|Y'| = m'$. Define $a = \max\{(m-n'-2m'+1)/2, 0\}$, $b = \min\{(m+n'+1)/2, m-m'\}$, $c = \max\{(m-n'-2m')/2, 0\}$ and $d = \min\{(m+n')/2, m-m'\}$. Then, the minimum possible Ansari-Bradley test statistic

over all possible missing values is:

$$\min_{\substack{X\setminus X'\in\mathbb{R}^{n-n'},\\ Y\setminus Y'\in\mathbb{R}^{m-m'}}} T(X,Y) = T(X',Y') + \begin{cases} \min\limits_{k\in\{a,\ldots,b\}} f_{X',Y',m}(k) + (n^2-n'^2)/4 & \text{if } \mathcal{C}(1),\\ \min\limits_{k\in\{c,\ldots,d\}} f_{X',Y',m}(k) + (n^2-n'^2-1)/4 & \text{if } \mathcal{C}(2),\\ \min\limits_{k\in\{a,\ldots,b\}} f_{X',Y',m}(k) + (n^2-n'^2+1)/4 & \text{if } \mathcal{C}(3),\\ \min\limits_{k\in\{c,\ldots,d\}} f_{X',Y',m}(k) + (n^2-n'^2)/4 & \text{if } \mathcal{C}(4), \end{cases}$$

and the maximum possible Ansari-Bradley test statistic is:

$$\max_{\substack{X\setminus X'\in\mathbb{R}^{n-n'},\\ Y\setminus Y'\in\mathbb{R}^{m-m'}}} T(X,Y) = -\min_{\substack{X\setminus X'\in\mathbb{R}^{n-n'},\\ Y\setminus Y'\in\mathbb{R}^{m-m'}}} T(Y,X) + \begin{cases} N^2/4 & N \text{ is even},\\ (N^2-1)/4 & N \text{ is odd}. \end{cases}$$

Proof. For notation convenience, let us denote

$$T_{\min}(X,Y) = \min_{\substack{X\setminus X'\in\mathbb{R}^{n-n'},\\ Y\setminus Y'\in\mathbb{R}^{m-m'}}} T(X,Y), \text{ and } T_{\max}(X,Y) = \max_{\substack{X\setminus X'\in\mathbb{R}^{n-n'},\\ Y\setminus Y'\in\mathbb{R}^{m-m'}}} T(X,Y).$$

To start, denote $N = n+m$. Suppose $N, n-n'$ satisfy $\mathcal{C}(i)$, where $i \in \{1,2,3,4\}$. Let us assume that when $X \setminus X' = \tilde{X}$, and $Y \setminus Y' = \tilde{Y}$, the Ansari-Bradley test statistic takes the minimum values, i.e.,

$$T(X' \cup \tilde{X}, Y' \cup \tilde{Y}) = T_{\min}(X,Y).$$

Consider $(\hat{X},\hat{Y}) \in \mathcal{Z}(X',Y',n-n',m-m',i)$, and denote $\hat{X} = \{x_1^*,\ldots,x_{n-n'}^*\}$, $\hat{Y} = \{y_1^*,\ldots,y_{m-m'}^*\}$, and $k = \sum_{i=1}^{m-m'} \mathrm{I}\left(y_i^* > \max \hat{X} \cup X' \cup Y'\right)$. Then, by Lemma B.4.6,

$$T(X_*,Y_*) = \begin{cases} T(X',Y') + f_{X',Y',m}(k) + (n^2-n'^2)/4, & \text{if } \mathcal{C}(1), \text{ or } \mathcal{C}(4)\\ T(X',Y') + f_{X',Y',m}(k) + (n^2-n'^2-1)/4, & \text{if } \mathcal{C}(2)\\ T(X',Y') + f_{X',Y',m}(k) + (n^2-n'^2+1)/4, & \text{if } \mathcal{C}(3) \end{cases}$$

By Lemma B.4.5, when $N, n-n'$ satisfy $\mathcal{C}(1)$ or $\mathcal{C}(3)$, $k \in \{a,\ldots,b\}$; while $N, n-n'$ satisfy $\mathcal{C}(2)$ or $\mathcal{C}(4)$, $k \in \{c,\ldots,d\}$. Without loss of generality, let us assume that $\hat{Y}$ takes the k that minimises $f_{X',Y',m}(k)$. In order words,

$$T(X_*,Y_*) = \begin{cases} T(X',Y') + \min\limits_{k\in\{a,\ldots,b\}} f_{X',Y',m}(k) + (n^2-n'^2)/4, & \text{if } \mathcal{C}(1)\\ T(X',Y') + \min\limits_{k\in\{c,\ldots,d\}} f_{X',Y',m}(k) + (n^2-n'^2-1)/4, & \text{if } \mathcal{C}(2)\\ T(X',Y') + \min\limits_{k\in\{a,\ldots,b\}} f_{X',Y',m}(k) + (n^2-n'^2+1)/4, & \text{if } \mathcal{C}(3)\\ T(X',Y') + \min\limits_{k\in\{c,\ldots,d\}} f_{X',Y',m}(k) + (n^2-n'^2)/4, & \text{if } \mathcal{C}(4) \end{cases}$$

By definition, we have

$$T_{\min}(X,Y) \le T(X_*,Y_*)$$

However, according to Proposition B.4.4, there exist $(\hat{X}_1, \hat{Y}_1) \in \mathcal{Z}(X', Y', n - n', m - m', i)$, such that

$$T(X' \cup \hat{X}_1, Y' \cup \hat{Y}_1) \leq T(X' \cup \tilde{X}, Y' \cup \tilde{Y}).$$

Applying Lemma B.4.6, we have

$$T(X_*, Y_*) \leq T(X' \cup \hat{X}_1, Y' \cup \hat{Y}_1).$$

Combining the above results,

$$\begin{aligned} & T(X_*, Y_*) \leq T(X' \cup \hat{X}_1, Y' \cup \hat{Y}_1) \leq T(X' \cup \tilde{X}, Y' \cup \tilde{Y}) = T_{\min}(X, Y) \leq T(X_*, Y_*) \\ \Longrightarrow\ & T(X_*, Y_*) = T_{\min}(X, Y). \end{aligned}$$

Then, according to Lemma B.3.1, we have

$$T(X, Y) = -T(Y, X) + \begin{cases} N^2/4 & N \text{ is even}, \\ (N^2 - 1)/4 & N \text{ is odd}. \end{cases}$$

Hence,

$$T_{\max}(X, Y) = -T_{\min}(Y, X) + \begin{cases} N^2/4 & N \text{ is even}, \\ (N^2 - 1)/4 & N \text{ is odd}. \end{cases}$$

This completes our proof. □

C. Proofs of results in Chapter 5

This chapter provides proofs for Chapter 5 in the main text. Our objective is to establish the tight lower and upper bounds of Spearman's footrule in the presence of missing data, where the components of X and Y can be missing in any configuration.

C.1. Different missingness cases

If a vector contains missing values, we will refer to it as partially observed, while a vector with no missing values is fully observed. We start by considering three specific cases which, when considered together, cover any possible case of a pair of partially observed vectors.

- Missing case I. All missing values are either all in X or all in Y. In other words, either X or Y is fully observed, while the other vector is partially observed.
- Missing case II. For any pair $(X(i), Y(i))$, $i \in [n]$, at most one value is missing. Therefore, both X and Y may contain missing data, and if $X(i)$ is missing, the paired $Y(i)$ must be observed. On the other hand, if $Y(i)$ is missing, the paired $X(i)$ must be observed.
- Missing case III. For any pair $(X(i), Y(i))$, $i \in [n]$, the two values are either both observed, or both missing. Therefore, both X and Y may contain missing values, and if $X(i)$ is missing, the paired $Y(i)$ is also missing. Similarly, if $Y(i)$ is missing, the paired $X(i)$ is also missing.

Note that these three missing cases are not mutually exclusive. In particular, the missing case I can be considered as a special case of missing case II, given that the latter includes the case where only one vector of X and Y can have missing values.

The three cases above are special cases of missingness patterns for a pair of vectors X, Y. We consider the case for which any missingness pattern is possible for X, Y to be the *general missingness case*. Fortunately, the following result shows that a pair of vectors subject to a general missingness pattern can always be rearranged into a concatenation of subvectors subject to missing case II and missing case III.

Lemma C.1.1. Suppose $X, Y \in \mathcal{V}^n$ are only partially observed and subject to a general missingness pattern. Then there exist subvectors X_2, X_3, Y_2, Y_3 and a permutation σ such that the concatenated subvectors (X_2, X_3) and (Y_2, Y_3) are equal to the permutations of X and Y under σ, respectively, i.e.

$$(X_2, X_3) = (X(\sigma(1)), \ldots, X(\sigma(n))), \qquad (Y_2, Y_3) = (Y(\sigma(1)), \ldots, Y(\sigma(n))).$$

Furthermore, the pair X_2, Y_2 is subject to missing case II, while the pair X_3, Y_3 is subject to missing case III.

Table C.1.: (Left) Example data under general missing case, where we do not make any assumptions of the missingness patterns. (Right) Example data after permutation, the subvectors including the first four indices are missing according to missing case II, while the last two indices are missing according to missing case III.

	a	b	c	d	e	f	g	h
X	5.2	*	4.6	2.1	*	3.3	1.8	*
Y	*	*	1.2	*	3.6	2.4	*	*

	a	d	g	e	c	f	b	h
X	5.2	1.8	6	*	4.6	3.3	*	*
Y	*	*	*	3.6	1.2	2.4	*	*

Proof. Denote the number of pairs where both $(X(i), Y(i))$ are observed as n_1, and the corresponding indices as $(\sigma(1), \ldots, \sigma(n_1))$; the number of pairs where only $X(i)$ is observed but $Y(i)$ is missing as m_1, and the corresponding indices as $(\sigma(n_1+1), \ldots, \sigma(n_1+m_1))$; the number of pairs where only $Y(i)$ is observed but $X(i)$ is missing as m_2, and the corresponding indices as $(\sigma(n_1+m_1+1), \ldots, \sigma(n_1+m_1+m_2))$. Finally, denote the indices of the pairs such that both $(X(i), Y(i))$ are missing as $(\sigma(n_1+m_1+m_2+1), \ldots, \sigma(n_1+m_1+m_2+m))$.

Then, since each pair either: observes both $X(i)$ and $Y(i)$, observes one of $X(i)$ or $Y(i)$, does not observe $X(i)$ nor $Y(i)$, we have $n_1+m_1+m_2+m = n$ and $\sigma = (\sigma(1), \ldots, \sigma(n))$ is a permutation of indices $\{1, \ldots, n\}$.

Further, since the subvectors

$$(X(\sigma(1)), \ldots, X(\sigma(n-m))), (Y(\sigma(1)), \ldots, Y(\sigma(n-m)))$$

includes only pairs where at most one pairs is missing, the subvectors are missing according to missing case II.

Since the subvectors

$$(X(\sigma(n-m+1)), \ldots, X(\sigma(n))), (Y(\sigma(n-m+1)), \ldots, Y(\sigma(n)))$$

includes only pairs where both values are missing, the subvectors are missing according to missing case III.

Let us define $X_2 = (X(\sigma(1)), \ldots, X(\sigma(n-m)))$, $X_3 = (X(\sigma(n-m+1)), \ldots, X(\sigma(n)))$, $Y_2 = (Y(\sigma(1)), \ldots, Y(\sigma(n-m)))$, and $Y_3 = (Y(\sigma(n-m+1)), \ldots, Y(\sigma(n)))$. This then completes our proof. □

For illustrating the result, we provide one example in Table C.1, where the data under general missingness case are composed of missing case II and missing case III after permutation. As we mentioned before, missing case I is a special case for missing case II and will be useful for proving the bounds for missing case II.

Let $X, Y \in \mathcal{V}^n$, which may be only partially observed. We use the notation U to denote the set of all indices where $X(i)$ is missing but $Y(i)$ is observed. Furthermore, we use the notation V to denote the set of all indices where $X(i)$ is observed but $Y(i)$ is missing. Additionally, we use the notation W to denote the set of all indices where both $X(i)$ and $Y(i)$ are missing.

Finally, $R(X, X)$ is used to denote the vector of ranks of all components in X, i.e.

$$R(X, X) = (R(X(1), X), \ldots, R(X(n), X)).$$

For example, if $X = (7, 2, 5)$, then $R(X, X) = (3, 1, 2)$. Note that, in practice, we often first convert the vectors X and Y into the rank vectors $R(X, X)$ and $R(Y, Y)$, respectively, and work with these rank vectors directly.

C.2. Some important lemmas

Before proving our results for the bounds of Spearman's footrule, we first provide several lemmas that will be used repeatedly throughout the proofs.

Lemma C.2.1. Suppose $X \in \mathcal{V}^n$ and $O, O' \subset [n]$ are non-empty subsets of indices such that $O \subset O'$. Denote $|O' \setminus O| = k$. Then, for any $i \in O$, we have

$$R(X(i), (X(l))_{l\in O}) + k \geq R(X(i), (X(l))_{l\in O'}) \geq R(X(i), (X(l))_{l\in O}).$$

Proof. Let $i \in O$ be a fixed index. Then, according to the definition of rank, we have

$$\begin{aligned} R(X(i), (X(l))_{l\in O'}) &= \sum_{l\in O'} \mathrm{I}\left(X(l) \leq X(i)\right) \\ &= \sum_{l\in O} \mathrm{I}\left(X(l) \leq X(i)\right) + \sum_{l\in O'\setminus O} \mathrm{I}\left(X(l) \leq X(i)\right) \\ &= R(X(i), (X(l))_{l\in O}) + \sum_{l\in O'\setminus O} \mathrm{I}\left(X(l) \leq X(i)\right). \end{aligned}$$

Since $|O' \setminus O| = k$, we have $k \geq \sum_{l\in O'\setminus O} \mathrm{I}\left(X(l) \leq X(i)\right) \geq 0$. Thus, it follows that

$$R(X(i), (X(l))_{l\in O}) + k \geq R(X(i), (X(l))_{l\in O'}) \geq R(X(i), (X(l))_{l\in O}),$$

which completes our proof. □

Lemma C.2.2. Suppose $X \in \mathcal{V}^n$, and $U, O \subset [n]$ are non-empty disjoint subsets of indices such that $U \cup O = [n]$. Then, for any $i \in O$, if either of the following two conditions hold:

$$\begin{aligned} &\text{(i)} : X(j) > X(i), \text{ for any } j \in U, \\ \text{or } &\text{(ii)} : R(X(j), X) > R(X(i), (X(l))_{l\in O}), \text{ for any } j \in U, \end{aligned}$$

we have $R(X(i), X) = R(X(i), (X(l))_{l\in O})$.

Proof. First, we show that if the condition (i) is true, then we have $R(X(i), X) = R(X(i), (X(l))_{l\in O})$.

To start, let $i \in O$ be a fixed index such that the condition (i) is true, i.e., for any $j \in U$, $X(j) > X(i)$. According to the definition of rank, we have

$$R(X(i), X) = \sum_{l\in[n]} \mathrm{I}\left(X(l) \leq X(i)\right).$$

Since U and O are non-empty disjoint subsets of $[n]$ such that $U \cup O = [n]$, then we have

$$R(X(i), X) = \sum_{l\in O} \mathrm{I}\left(X(l) \leq X(i)\right) + \sum_{l\in U} \mathrm{I}\left(X(l) \leq X(i)\right).$$

Now, according to the condition (i), we have $X(j) > X(i)$, for any $j \in U$. Hence,

$$\sum_{l \in U} \mathrm{I}\left(X(l) \leq X(i)\right) = 0,$$

which follows that

$$R(X(i), X) = \sum_{l \in O} \mathrm{I}\left(X(l) \leq X(i)\right) = R(X(i), (X(l))_{l \in O}).$$

Thus, we complete our prove when the condition (i) is true.

Next, we show that if the condition (ii) is true, we have $R(X(i), X) = R(X(i), (X(l))_{l \in O})$. Let $i \in O$ be any fixed index such that the condition (ii) is true, i.e., for any $j \in U$, $R(X(j), X) > R(X(i), (X(l))_{l \in O})$. We prove the result for the condition (ii) by showing that the condition (ii) implies the condition (i), i.e.,

$$R(X(j), X) > R(X(i), (X(l))_{l \in O}), \text{ for } j \in U \implies X(j) > X(i), \text{ for } j \in U.$$

Without loss of generality, denote u as the index of the minimum value in $(X(l))_{l \in U}$, i.e.,

$$X(u) = \min(X(l))_{l \in U}.$$

Then, all components with indices in $U \setminus \{u\}$ are larger than $X(u)$. Hence, we have

$$\sum_{l \in U} \mathrm{I}\left(X(l) \leq X(u)\right) = \mathrm{I}\left(X(u) \leq X(u)\right) = 1.$$

Subsequently, since U and O are non-empty disjoint subsets of $[n]$ such that $U \cup O = [n]$, then according to the definition of rank, we have

$$\begin{aligned} R(X(u), X) &= \sum_{l \in O} \mathrm{I}\left(X(l) \leq X(u)\right) + \sum_{l \in U} \mathrm{I}\left(X(l) \leq X(u)\right) \\ &= \sum_{l \in O} \mathrm{I}\left(X(l) \leq X(u)\right) + \mathrm{I}\left(X(u) \leq X(u)\right) \\ &= \sum_{l \in O \cup \{u\}} \mathrm{I}\left(X(l) \leq X(u)\right) \\ &= R(X(u), (X(l))_{l \in O \cup \{u\}}). \end{aligned} \tag{C.1}$$

Next, according to Lemma C.2.1, we have

$$R(X(i), (X(l))_{l \in O}) + 1 \geq R(X(i), (X(l))_{l \in O \cup \{u\}}).$$

According to the condition (ii), we have $R(X(j), X) > R(X(i), (X(l))_{l \in O})$, for any $j \in U$. Notice that $u \in U$, then we have

$$\begin{aligned} & R(X(u), X) > R(X(i), (X(l))_{l \in O}) \\ \implies & R(X(u), X) \geq R(X(i), (X(l))_{l \in O}) + 1, \end{aligned}$$

where the " $\Longrightarrow$ " holds because $R(X(i),(X(l))_{l\in O})$ and $R(X(u),X)$ are both integers. Hence, we have

$$R(X(u),X) \geq R(X(i),(X(l))_{l\in O}) + 1 \geq R(X(i),(X(l))_{l\in O\cup\{u\}}).$$

According to (C.1), we further have

$$R(X(u),(X(l))_{l\in O\cup\{u\}}) = R(X(u),X) \geq R(X(i),(X(l))_{l\in O\cup\{u\}}).$$

Since $X \in \mathcal{V}^n$ is a vector of distinct real values, and $u \notin O$, then we have

$$\begin{aligned} & R(X(i),(X(l))_{l\in O\cup\{u\}}) \neq R(X(u),(X(l))_{l\in O\cup\{u\}}) \\ \Longrightarrow\ & R(X(u),(X(l))_{l\in O\cup\{u\}}) > R(X(i),(X(l))_{l\in O\cup\{u\}}) \\ \Longrightarrow\ & X(u) > X(i). \end{aligned}$$

Recall that $X(u) = \min(X(l))_{l\in U}$. Therefore, we have $X(j) > X(i)$ for any $j \in U$.

Then, by applying the result when the condition (i) holds, we obtain

$$R(X(i),X) = R(X(i),(X(l))_{l\in O}),$$

which completes our proof. □

Lemma C.2.3. Suppose $X_1, X_2 \in \mathcal{V}^n$, and U, O are non-empty disjoint subsets of $[n]$ such that $U \cup O = [n]$. Then, if X_2 is an imputation of X_1 for indices U and $R(X_1(i),X_1) = R(X_2(i),X_2)$, for any $i \in U$, we have $R(X_1,X_1) = R(X_2,X_2)$.

Proof. To start, let us denote

$$U = \{u_1,\cdots,u_m\}, \text{ and } O = \{o_1,\cdots,o_{n-m}\}.$$

For both $j = 1,2$, since $X_j \in \mathcal{V}^n$ is a vector of n distinct real numbers, the ranks of all components in X_j form a permutation of $\{1,\cdots,n\}$, i.e.,

$$\{R(X_j(1),X_j),\cdots,R(X_j(n),X_j)\} = \{1,\cdots,n\}, \text{ for both } j = \{1,2\}.$$

Let us denote

$$S_j = \{1,\cdots,n\} \setminus \{R(X_j(u_1),X_j),\cdots,R(X_j(u_m),X_j)\}, \text{ for both } j = \{1,2\}.$$

Then, since $U \cup O = [n] = \{1,2,\cdots,n\}$, we have

$$\{R(X_j(o_1),X_j),\cdots,R(X_j(o_{n-m}),X_j)\} = S_j, \text{ for both } j = 1,2.$$

Since $R(X_1(i),X_1) = R(X_2(i),X_2)$, for any $i \in U$, we have $S_2 = S_1$.

Notice that for any fixed $i \in \{1,\cdots,n-m\}$, and $j = 1,2$, the order of $R(X_j(o_i),X_j)$ in S_j equals to the order of $X_j(o_i)$ in $\{X_j(o_1),\ldots,X_j(o_{n-m})\}$. Since X_2 is an imputation of X_1 for

indices U, then according to the definition of imputations, for any $i \in \{1, 2, \cdots, n-m\}$, we have $X_1(o_i) = X_2(o_i)$. Hence, the order of $X_1(o_i)$ in $\{X_1(o_1), \ldots, X_1(o_{n-m})\}$ equals to the order of $X_2(o_i)$ in $\{X_2(o_1), \ldots, X_1(o_{n-m})\}$. Thus, the order of $R(X_1(o_i), X_1)$ in S_1 equals to the order of $R(X_2(o_i), X_2)$ in S_2. Furthermore, since $S_1 = S_2$, we have

$$R(X_1(o_i), X_1) = R(X_2(o_i), X_2),$$

where $i \in \{1, \cdots, n-m\}$. Since we also have $R(X_1(i), X_1) = R(X_2(i), X_2)$, for any $i \in U$, and $O \cup U = [n]$, then we have

$$R(X_1(i), X_1) = R(X_2(i), X_2), \text{ for any } i \in [n].$$

In order words, we have $R(X_1, X_1) = R(X_2, X_2)$. This completes our proof. □

C.3. Tight lower bounds of Spearman's footrule

In what follows, we first provide the tight lower bounds of Spearman's footrule. In Section C.3.1 and C.3.2, the problem for establishing exact lower bounds under missing case I and II are discussed, respectively. Algorithms 2 and 3 provide computationally efficient methods for computing these bounds. The lower bound under missing case III is discussed in Papaioannou and Loukas (1984), although it not shown there that the bound is tight. In Proposition C.3.31, we show this bound is tight. Finally, by synthesising the results developed for missing case I, II and III, we provide the lower bounds in the general missingness case.

C.3.1. Lower bound under missing case I

We begin by carefully studying missing case I, where the missing values are assumed to be either all in X or all in Y. Without loss of generality, we proceed by assuming only X contains missing values.

We first prove Proposition C.3.5, which investigates the conditions for taking the minimum Spearman's footrule when one component $X(u)$ in X is replaced by a new value $X_*(u)$. The following four lemmas will be useful for proving the proposition.

Lemma C.3.1. Suppose $X, X_* \in \mathcal{V}^n$ and let X_* be an imputation of X for an index $u \in [n]$. Then, if $R(X_*(u), X_*) = R(X(u), X) + 1$, we have $X_*(u) > X(u)$ and

$$\sum_{i \in [n] \setminus \{u\}} \mathrm{I}\left(X(i) > X(u)\right) \mathrm{I}\left(X(i) < X_*(u)\right) = 1.$$

However, if $R(X_*(u), X_*) = R(X(u), X) - 1$, we have $X_*(u) < X(u)$ and

$$\sum_{i \in [n] \setminus \{u\}} \mathrm{I}\left(X(i) > X_*(u)\right) \mathrm{I}\left(X(i) < X(u)\right) = 1.$$

Proof. To start, according to the definition of rank, we have

$$R(X(u), X) = \sum_{i\in[n]} \mathrm{I}\left(X(i) \le X(u)\right) = \mathrm{I}\left(X(u) \le X(u)\right) + \sum_{i\in[n]\setminus\{u\}} \mathrm{I}\left(X(i) \le X(u)\right).$$

Since $X \in \mathcal{V}^n$ is a vector of distinct real values, we have

$$R(X(u), X) = 1 + \sum_{i\in[n]\setminus\{u\}} \mathrm{I}\left(X(i) < X(u)\right). \tag{C.2}$$

Similarly, we can obtain

$$R(X_*(u), X_*) = 1 + \sum_{i\in[n]\setminus\{u\}} \mathrm{I}\left(X_*(i) < X_*(u)\right).$$

Since X_* is an imputation of X for the index u, then according to the definition of imputations, we have $X_*(i) = X(i)$ for any $i \in [n] \setminus \{u\}$. Hence, we further have

$$R(X_*(u), X_*) = 1 + \sum_{i\in[n]\setminus\{u\}} \mathrm{I}\left(X(i) < X_*(u)\right). \tag{C.3}$$

Subsequently, if $R(X_*(u), X_*) = R(X(u), X) + 1$, by combining (C.2) and (C.3), we have

$$\sum_{i\in[n]\setminus\{u\}} \mathrm{I}\left(X(i) < X_*(u)\right) = \sum_{i\in[n]\setminus\{u\}} \mathrm{I}\left(X(i) < X(u)\right) + 1.$$

Hence, it can be seen that $X_*(u) > X(u)$.

Now, since $X \in \mathcal{V}^n$ is a vector of distinct real values, we further have

$$\begin{aligned}
&\sum_{i\in[n]\setminus\{u\}} \mathrm{I}\left(X(i) < X_*(u)\right) - \sum_{i\in[n]\setminus\{u\}} \mathrm{I}\left(X(i) < X(u)\right) = 1 \\
\iff &\sum_{i\in[n]\setminus\{u\}} \mathrm{I}\left(X(i) < X(u)\right) + \sum_{i\in[n]\setminus\{u\}} \mathrm{I}\left(X(u) < X(i) < X_*(u)\right) \\
&- \sum_{i\in[n]\setminus\{u\}} \mathrm{I}\left(X(i) < X(u)\right) = 1 \\
\iff &\sum_{i\in[n]\setminus\{u\}} \mathrm{I}\left(X(u) < X(i) < X_*(u)\right) = 1 \\
\iff &\sum_{i\in[n]\setminus\{u\}} \mathrm{I}\left(X(i) > X(u)\right) \mathrm{I}\left(X(i) < X_*(u)\right) = 1,
\end{aligned}$$

which proves our results when $R(X_*(u), X_*) = R(X(u), X) + 1$.

The case when $R(X_*(u), X_*) = R(X(u), X) - 1$ can be proved similarly. Combining (C.2) and (C.3), we have

$$\sum_{i\in[n]\setminus\{u\}} \mathrm{I}\left(X(i) < X_*(u)\right) = \sum_{i\in[n]\setminus\{u\}} \mathrm{I}\left(X(i) < X(u)\right) - 1,$$

which follows $X_*(u) < X(u)$.

Since $X_* \in \mathcal{V}^n$ is a vector of distinct real values and $X_*(u) \neq X_*(i) = X(i)$ for any $i \in [n] \setminus \{u\}$,

we have

$$\begin{aligned}
&\sum_{i\in[n]\setminus\{u\}} \mathrm{I}\left(X(i) < X(u)\right) - \sum_{i\in[n]\setminus\{u\}} \mathrm{I}\left(X(i) < X_*(u)\right) = 1 \\
\iff &\sum_{i\in[n]\setminus\{u\}} \mathrm{I}\left(X(i) < X_*(u)\right) + \sum_{i\in[n]\setminus\{u\}} \mathrm{I}\left(X_*(u) < X(i) < X(u)\right) \\
&- \sum_{i\in[n]\setminus\{u\}} \mathrm{I}\left(X(i) < X_*(u)\right) = 1 \\
\iff &\sum_{i\in[n]\setminus\{u\}} \mathrm{I}\left(X_*(u) < X(i) < X(u)\right) = 1 \\
\iff &\sum_{i\in[n]\setminus\{u\}} \mathrm{I}\left(X(i) > X_*(u)\right) \mathrm{I}\left(X(i) < X(u)\right) = 1,
\end{aligned}$$

which proves our results when $R(X_*(u), X_*) = R(X(u), X) - 1$, and completes our proof. □

Lemma C.3.2. Suppose $X, Y \in \mathcal{V}^n$ and let $X_* \in \mathcal{V}^n$ be an imputation of X for an index $u \in [n]$. For any $i \in [n] \setminus \{u\}$, let us denote

$$\begin{aligned}
\xi_i &= \left|R(X(i), (X(l))_{l\in[n]\setminus\{u\}}) + 1 - R(Y(i), Y)\right|, \\
\text{and } \psi_i &= \left|R(X(i), (X(l))_{l\in[n]\setminus\{u\}}) - R(Y(i), Y)\right|.
\end{aligned}$$

Then, we have

$$\begin{aligned}
D\left(X_*, Y\right) - D\left(X, Y\right) = &\left|R(X_*(u), X) - R(Y(u), Y)\right| - \left|R(X(u), X) - R(Y(u), Y)\right| \\
&+ \sum_{i\in[n]\setminus\{u\}} \mathrm{I}\left(X(i) > X_*(u)\right) \mathrm{I}\left(X(i) < X(u)\right) \left(\xi_i - \psi_i\right) \\
&+ \sum_{i\in[n]\setminus\{u\}} \mathrm{I}\left(X(i) < X_*(u)\right) \mathrm{I}\left(X(i) > X(u)\right) \left(\psi_i - \xi_i\right).
\end{aligned}$$

Proof. For convenience, let us assume (after relabelling) $u = 1$, and denote

$$\alpha = \left|R(X(1), X) - R(Y(1), Y)\right|, \text{ and } \beta = \left|R(X_*(1), X) - R(Y(1), Y)\right|.$$

According to the definition of Spearman's footrule, we have

$$\begin{aligned}
D\left(X, Y\right) &= \sum_{i=1}^{n} \left|R(X(i), X) - R(Y(i), Y)\right| \\
&= \left|R(X(1), X) - R(Y(1), Y)\right| + \sum_{i=2}^{n} \left|R(X(i), X) - R(Y(i), Y)\right| \\
&= \alpha + \sum_{i=2}^{n} \left|R(X(i), X) - R(Y(i), Y)\right|.
\end{aligned}$$

According to the definition of rank, for any $i \in \{2, \ldots, n\}$, we have

$$\begin{aligned} R(X(i), X) &= \sum_{l=1}^{n} \mathrm{I}\left(X(1) \leq X(i)\right) \\ &= \sum_{l=2}^{n} \mathrm{I}\left(X(1) \leq X(i)\right) + \mathrm{I}\left(X(1) \leq X(i)\right) \\ &= R(X(i), (X(l))_{l \in [n] \setminus \{1\}}) + \mathrm{I}\left(X(1) \leq X(i)\right). \end{aligned}$$

Since $X \in \mathcal{V}^n$ is a vector of distinct values, $\mathrm{I}\left(X(1) \leq X(i)\right) = \mathrm{I}\left(X(1) < X(i)\right)$. Hence, for any $i \in \{2, \ldots, n\}$, we have

$$R(X(i), X) = R(X(i), (X(l))_{l \in [n] \setminus \{1\}}) + \mathrm{I}\left(X(1) < X(i)\right).$$

Therefore, it follows that

$$D\left(X, Y\right) = \alpha + \sum_{i=2}^{n} \left| R(X(i), (X(l))_{l \in [n] \setminus \{1\}}) + \mathrm{I}\left(X(1) < X(i)\right) - R(Y(i), Y) \right|.$$

Subsequently,

$$\begin{aligned} D\left(X, Y\right) &= \alpha + \sum_{i=2}^{n} \mathrm{I}\left(X(1) < X(i)\right) \left| R(X(i), (X(l))_{l \in [n] \setminus \{1\}}) + 1 - R(Y(i), Y) \right| \\ &\quad + \sum_{i=2}^{n} \mathrm{I}\left(X(1) > X(i)\right) \left| R(X(i), (X(l))_{l \in [n] \setminus \{1\}}) - R(Y(i), Y) \right| \\ &= \alpha + \sum_{i=2}^{n} \mathrm{I}\left(X(i) > X(1)\right) \xi_i + \sum_{i=2}^{n} \mathrm{I}\left(X(i) < X(1)\right) \psi_i. \end{aligned}$$

Similarly, we can show that

$$\begin{aligned} D\left(X_*, Y\right) &= \beta + \sum_{i=2}^{n} \mathrm{I}\left(X_*(1) < X_*(i)\right) \left| R(X_*(i), (X_*(l))_{l \in [n] \setminus \{1\}}) + 1 - R(Y(i), Y) \right| \\ &\quad + \sum_{i=2}^{n} \mathrm{I}\left(X_*(1) > X_*(i)\right) \left| R(X_*(i), (X_*(l))_{l \in [n] \setminus \{1\}}) - R(Y(i), Y) \right|. \end{aligned}$$

Since X_* is an imputation of X for the index 1, then according to the definition of imputations, we have $X_*(i) = X(i)$ for any $i \in [n] \setminus \{1\}$. Thus, we have

$$\begin{aligned} D\left(X_*, Y\right) &= \beta + \sum_{i=2}^{n} \mathrm{I}\left(X(i) > X_*(1)\right) \left| R(X(i), (X(l))_{l \in [n] \setminus \{1\}}) + 1 - R(Y(i), Y) \right| \\ &\quad + \sum_{i=2}^{n} \mathrm{I}\left(X(i) < X_*(1)\right) \left| R(X(i), (X(l))_{l \in [n] \setminus \{1\}}) - R(Y(i), Y) \right| \\ &= \beta + \sum_{i=2}^{n} \mathrm{I}\left(X(i) > X_*(1)\right) \xi_i + \sum_{i=2}^{n} \mathrm{I}\left(X(i) < X_*(1)\right) \psi_i. \end{aligned}$$

Now, we have

$$\begin{aligned} D\left(X_*, Y\right) - D\left(X, Y\right) = \beta - \alpha + \sum_{i=2}^{n} \left\{ \mathrm{I}\left(X(i) > X_*(1)\right) - \mathrm{I}\left(X(i) > X(1)\right) \right\} \xi_i \\ + \sum_{i=2}^{n} \left\{ \mathrm{I}\left(X(i) < X_*(1)\right) - \mathrm{I}\left(X(i) < X(1)\right) \right\} \psi_i. \end{aligned} \tag{C.4}$$

Next, since $X, X_* \in \mathcal{V}^n$ are both vectors of distinct real values, then for any $i \in \{2, \cdots, n\}$, we have $X(i) \neq X(1)$, and $X_*(i) \neq X_*(1)$. Further, since X_* is an imputation of X for the index 1, we have $X_*(i) = X(i)$ for any $i \in \{2, \ldots, n\}$. Hence, we also have $X(i) \neq X_*(1)$ for any $i \in \{2, \cdots, n\}$. Subsequently, for any $i \in \{2, \cdots, n\}$, we have

$$\begin{aligned} &\mathrm{I}\left(X(i) > X_*(1)\right) - \mathrm{I}\left(X(i) > X(1)\right) \\ &= \mathrm{I}\left(X(i) > X_*(1)\right) \left\{ \mathrm{I}\left(X(i) < X(1)\right) + \mathrm{I}\left(X(i) \geq X(1)\right) \right\} \\ &- \mathrm{I}\left(X(i) > X(1)\right) \left\{ \mathrm{I}\left(X(i) < X_*(1)\right) + \mathrm{I}\left(X(i) \geq X_*(1)\right) \right\} \\ &= \mathrm{I}\left(X(i) > X_*(1)\right) \mathrm{I}\left(X(i) < X(1)\right) + \mathrm{I}\left(X(i) > X_*(1)\right) \mathrm{I}\left(X(i) \geq X(1)\right) \\ &- \mathrm{I}\left(X(i) > X(1)\right) \mathrm{I}\left(X(i) < X_*(1)\right) - \mathrm{I}\left(X(i) > X(1)\right) \mathrm{I}\left(X(i) \geq X_*(1)\right) \end{aligned}$$

By rearranging the above terms, we further have

$$\begin{aligned} &\mathrm{I}\left(X(i) > X_*(1)\right) - \mathrm{I}\left(X(i) > X(1)\right) \\ &= \mathrm{I}\left(X(i) > X_*(1)\right) \mathrm{I}\left(X(i) < X(1)\right) - \mathrm{I}\left(X(i) > X(1)\right) \mathrm{I}\left(X(i) < X_*(1)\right) \\ &+ \mathrm{I}\left(X(i) > X_*(1)\right) \mathrm{I}\left(X(i) \geq X(1)\right) - \mathrm{I}\left(X(i) > X(1)\right) \mathrm{I}\left(X(i) \geq X_*(1)\right) \end{aligned}$$

Since for any $i \in \{2, \ldots, n\}$, we have $X(i) \neq X(1), X_*(1)$. Hence,

$$\begin{aligned} \mathrm{I}\left(X(i) > X_*(1)\right) &= \mathrm{I}\left(X(i) \geq X_*(1)\right), \\ \mathrm{I}\left(X(i) \geq X(1)\right) &= \mathrm{I}\left(X(i) > X(1)\right), \end{aligned}$$

which means

$$\mathrm{I}\left(X(i) > X_*(1)\right) \mathrm{I}\left(X(i) \geq X(1)\right) - \mathrm{I}\left(X(i) > X(1)\right) \mathrm{I}\left(X(i) \geq X_*(1)\right) = 0$$

Therefore, we have

$$\begin{aligned} &\mathrm{I}\left(X(i) > X_*(1)\right) - \mathrm{I}\left(X(i) > X(1)\right) \\ &= \mathrm{I}\left(X(i) > X_*(1)\right) \mathrm{I}\left(X(i) < X(1)\right) - \mathrm{I}\left(X(i) < X_*(1)\right) \mathrm{I}\left(X(i) > X(1)\right). \end{aligned} \tag{C.5}$$

Similarly, for any $i \in \{2, \cdots, n\}$, we have

$$\begin{aligned}
&\mathrm{I}\left(X(i) < X_*(1)\right) - \mathrm{I}\left(X(i) < X(1)\right) \\
&= \mathrm{I}\left(X(i) < X_*(1)\right)\left\{\mathrm{I}\left(X(i) > X(1)\right) + \mathrm{I}\left(X(i) \leq X(1)\right)\right\} \\
&- \mathrm{I}\left(X(i) < X(1)\right)\left\{\mathrm{I}\left(X(i) > X_*(1)\right) + \mathrm{I}\left(X(i) \leq X_*(1)\right)\right\} \\
&= \mathrm{I}\left(X(i) < X_*(1)\right)\mathrm{I}\left(X(i) > X(1)\right) + \mathrm{I}\left(X(i) < X_*(1)\right)\mathrm{I}\left(X(i) \leq X(1)\right) \\
&- \mathrm{I}\left(X(i) < X(1)\right)\mathrm{I}\left(X(i) > X_*(1)\right) - \mathrm{I}\left(X(i) < X(1)\right)\mathrm{I}\left(X(i) \leq X_*(1)\right).
\end{aligned}$$

By rearranging the above terms, we further have

$$\begin{aligned}
&\mathrm{I}\left(X(i) < X_*(1)\right) - \mathrm{I}\left(X(i) < X(1)\right) \\
&= \mathrm{I}\left(X(i) < X_*(1)\right)\mathrm{I}\left(X(i) > X(1)\right) - \mathrm{I}\left(X(i) < X(1)\right)\mathrm{I}\left(X(i) > X_*(1)\right) \\
&+ \mathrm{I}\left(X(i) < X_*(1)\right)\mathrm{I}\left(X(i) \leq X(1)\right) - \mathrm{I}\left(X(i) < X(1)\right)\mathrm{I}\left(X(i) \leq X_*(1)\right).
\end{aligned}$$

Since for any $i \in \{2, \ldots, n\}$, we have $X(i) \neq X(1), X_*(1)$. Hence,

$$\begin{aligned}
\mathrm{I}\left(X(i) < X_*(1)\right) &= \mathrm{I}\left(X(i) \leq X_*(1)\right), \\
\mathrm{I}\left(X(i) \leq X(1)\right) &= \mathrm{I}\left(X(i) < X(1)\right),
\end{aligned}$$

which means

$$\mathrm{I}\left(X(i) < X_*(1)\right)\mathrm{I}\left(X(i) \leq X(1)\right) - \mathrm{I}\left(X(i) < X(1)\right)\mathrm{I}\left(X(i) \leq X_*(1)\right) = 0.$$

Therefore, we have

$$\begin{aligned}
&\mathrm{I}\left(X(i) < X_*(1)\right) - \mathrm{I}\left(X(i) < X(1)\right) \\
&= \mathrm{I}\left(X(i) < X_*(1)\right)\mathrm{I}\left(X(i) > X(1)\right) - \mathrm{I}\left(X(i) < X(1)\right)\mathrm{I}\left(X(i) > X_*(1)\right).
\end{aligned} \tag{C.6}$$

Putting (C.5) and (C.6) back into (C.4), we obtain

$$\begin{aligned}
D\left(X_*, Y\right) - D\left(X, Y\right) = \beta - \alpha &+ \sum_{i=2}^{n} \mathrm{I}\left(X(i) > X_*(1)\right)\mathrm{I}\left(X(i) < X(1)\right)\xi_i \\
&- \sum_{i=2}^{n} \mathrm{I}\left(X(i) < X_*(1)\right)\mathrm{I}\left(X(i) > X(1)\right)\xi_i \\
&+ \sum_{i=2}^{n} \mathrm{I}\left(X(i) < X_*(1)\right)\mathrm{I}\left(X(i) > X(1)\right)\psi_i \\
&- \sum_{i=2}^{n} \mathrm{I}\left(X(i) > X_*(1)\right)\mathrm{I}\left(X(i) < X(1)\right)\psi_i.
\end{aligned}$$

That is,

$$D(X_*, Y) - D(X, Y) = \beta - \alpha + \sum_{i=2}^{n} \mathrm{I}(X(i) > X_*(1))\,\mathrm{I}(X(i) < X(1))\,(\xi_i - \psi_i) + \sum_{i=2}^{n} \mathrm{I}(X(i) < X_*(1))\,\mathrm{I}(X(i) > X(1))\,(\psi_i - \xi_i).$$

This completes our proof. □

Lemma C.3.3. Suppose $X, Y \in \mathcal{V}^n$, and let $X_* \in \mathcal{V}^n$ be an imputation of X for an index $u \in [n]$. Then, if one of the two following conditions holds

$$\begin{aligned} &\text{(i)} : R(X(u), X) > R(Y(u), Y) \text{ and } R(X_*(u), X_*) = R(X(u), X) - 1, \\ \text{or } &\text{(ii)} : R(X(u), X) < R(Y(u), Y) \text{ and } R(X_*(u), X_*) = R(X(u), X) + 1, \end{aligned}$$

we have $D(X_*, Y) \leq D(X, Y)$. However, if one of the two following conditions holds

$$\begin{aligned} &\text{(iii)} : R(X(u), X) \geq R(Y(u), Y) \text{ and } R(X_*(u), X_*) = R(X(u), X) + 1, \\ \text{or } &\text{(iv)} : R(X(u), X) \leq R(Y(u), Y) \text{ and } R(X_*(u), X_*) = R(X(u), X) - 1, \end{aligned}$$

we have $D(X_*, Y) \geq D(X, Y)$.

Proof. For convenience, let us assume (after relabelling) $u = 1$, and denote

$$\alpha = |R(X(1), X) - R(Y(1), Y)|, \text{ and } \beta = |R(X_*(1), X) - R(Y(1), Y)|.$$

For any $i \in \{2, \cdots, n\}$, let us denote

$$\begin{aligned} \xi_i &= \left|R(X(i), (X(l))_{l \in [n] \setminus \{1\}}) + 1 - R(Y(i), Y)\right|, \\ \text{and } \psi_i &= \left|R(X(i), (X(l))_{l \in [n] \setminus \{1\}}) - R(Y(i), Y)\right|. \end{aligned}$$

Then, for any $i \in \{2, \cdots, n\}$, if

$$R(X(i), (X(l))_{l \in [n] \setminus \{1\}}) - R(Y(i), Y) \geq 0,$$

we have

$$\begin{aligned} \psi_i &= R(X(i), (X(l))_{l \in [n] \setminus \{1\}}) - R(Y(i), Y), \\ \text{and } \xi_i &= R(X(i), (X(l))_{l \in [n] \setminus \{1\}}) + 1 - R(Y(i), Y). \end{aligned}$$

Hence we have $\xi_i = \psi_i + 1$. However, if

$$R(X(i), (X(l))_{l \in [n] \setminus \{1\}}) - R(Y(i), Y) < 0,$$

we have

$$\begin{aligned}\psi_i &= R(Y(i), Y) - R(X(i), (X(l))_{l\in[n]\setminus\{1\}}),\\ \text{and } \xi_i &= R(Y(i), Y) - R(X(i), (X(l))_{l\in[n]\setminus\{1\}}) - 1.\end{aligned}$$

Hence we have $\xi_i = \psi_i - 1$. Therefore, we have

$$\xi_i - \psi_i \in \{-1, 1\}, \text{ for any } i \in \{2, \cdots, n\}. \tag{C.7}$$

Next, according to Lemma C.3.2, we have

$$\begin{aligned}D(X_*, Y) - D(X, Y) = \beta - \alpha &+ \sum_{i=2}^{n} \mathrm{I}(X(i) > X_*(1))\, \mathrm{I}(X(i) < X(1))\, (\xi_i - \psi_i)\\ &+ \sum_{i=2}^{n} \mathrm{I}(X(i) < X_*(1))\, \mathrm{I}(X(i) > X(1))\, (\psi_i - \xi_i).\end{aligned}$$

For notation ease, let us denote

$$\begin{aligned}\Sigma_1 &= \sum_{i=2}^{n} \mathrm{I}(X(i) > X_*(1))\, \mathrm{I}(X(i) < X(1))\, (\xi_i - \psi_i),\\ \text{and } \Sigma_2 &= \sum_{i=2}^{n} \mathrm{I}(X(i) < X_*(1))\, \mathrm{I}(X(i) > X(1))\, (\psi_i - \xi_i).\end{aligned}$$

Then we have

$$D(X_*, Y) - D(X, Y) = \beta - \alpha + \Sigma_1 + \Sigma_2. \tag{C.8}$$

Suppose the condition

$$\text{(i)} : R(X(1), X) > R(Y(1), Y) \text{ and } R(X_*(1), X_*) = R(X(1), X) - 1$$

holds. Then we have

$$\begin{aligned}&R(X(1), X) - R(Y(1), Y) > 0\\ \Longrightarrow\ &\alpha = |R(X(1), X) - R(Y(1), Y)| = R(X(1), X) - R(Y(1), Y),\end{aligned}$$

and

$$\begin{aligned}&R(X_*(1), X_*) - R(Y(1), Y) = R(X(1), X) - 1 - R(Y(1), Y) \geq 0\\ \Longrightarrow\ &\beta = |R(X_*(1), X_*) - R(Y(1), Y)| = R(X(1), X) - 1 - R(Y(1), Y).\end{aligned}$$

Hence, we have

$$\beta = \alpha - 1.$$

Meanwhile, since $R(X_*(1), X_*) = R(X(1), X) - 1$ and X_* is an imputation of X for the index 1,

then according to Lemma C.3.1, we have

$$X_*(1) < X(1) \implies \mathrm{I}\left(X(i) < X_*(1)\right)\mathrm{I}\left(X(i) > X(1)\right) = 0,$$

for any $i \in \{2, \ldots, n\}$. Therefore,

$$\Sigma_2 = \sum_{i=2}^{n} \mathrm{I}\left(X(i) < X_*(1)\right)\mathrm{I}\left(X(i) > X(1)\right)\left(\psi_i - \xi_i\right) = 0.$$

According to Lemma C.3.1, we also have

$$\sum_{i=2}^{n} \mathrm{I}\left(X(i) > X_*(1)\right)\mathrm{I}\left(X(i) < X(1)\right) = 1.$$

According to (C.7), we have $\xi_i - \psi_i \in \{-1, 1\}$, which then follows that

$$\begin{aligned}\Sigma_1 &= \sum_{i=2}^{n} \mathrm{I}\left(X(i) > X_*(1)\right)\mathrm{I}\left(X(i) < X(1)\right)\left(\xi_i - \psi_i\right) \\ &\leq \sum_{i=2}^{n} \mathrm{I}\left(X(i) > X_*(1)\right)\mathrm{I}\left(X(i) < X(1)\right) = 1.\end{aligned}$$

Recall that $\beta = \alpha - 1$. Subsequently, following (C.8), we have

$$D\left(X_*, Y\right) - D\left(X, Y\right) = -1 + \Sigma_1 + \Sigma_2 \leq -1 + 1 + 0 = 0,$$

which proves our results when the condition (i) holds. The other three cases can be proved similarly.

Suppose the condition

$$\text{(ii)} : R(X(1), X) < R(Y(1), Y) \text{ and } R(X_*(1), X_*) = R(X(1), X) + 1$$

holds. Then we have

$$\begin{aligned}&R(X(1), X) - R(Y(1), Y) < 0 \\ \implies &\alpha = |R(X(1), X) - R(Y(1), Y)| = R(Y(1), Y) - R(X(1), X),\end{aligned}$$

and

$$\begin{aligned}&R(X_*(1), X_*) - R(Y(1), Y) = R(X(1), X) + 1 - R(Y(1), Y) \leq 0 \\ \implies &\beta = |R(X_*(1), X_*) - R(Y(1), Y)| = R(Y(1), Y) - R(X(1), X) - 1.\end{aligned}$$

Hence, we have

$$\beta = \alpha - 1.$$

Since $R(X_*(1), X_*) = R(X(1), X) + 1$, and X_* is an imputation of X for the index 1, then

according to Lemma C.3.1, we have

$$X_*(1) > X(1) \implies \mathrm{I}\left(X(i) > X_*(1)\right)\mathrm{I}\left(X(i) < X(1)\right) = 0,$$

for any $i \in \{2, \ldots, n\}$. Therefore,

$$\Sigma_1 = \sum_{i=2}^{n} \mathrm{I}\left(X(i) > X_*(1)\right)\mathrm{I}\left(X(i) < X(1)\right)\left(\xi_i - \psi_i\right) = 0.$$

According to Lemma C.3.1, we also have

$$\sum_{i=2}^{n} \mathrm{I}\left(X(i) < X_*(1)\right)\mathrm{I}\left(X(i) > X(1)\right) = 1.$$

According to (C.7), we have $\xi_i - \psi_i \in \{-1, 1\}$, which follows that

$$\begin{aligned}\Sigma_2 &= \sum_{i=2}^{n} \mathrm{I}\left(X(i) < X_*(1)\right)\mathrm{I}\left(X(i) > X(1)\right)\left(\psi_i - \xi_i\right) \\ &\leq \sum_{i=2}^{n} \mathrm{I}\left(X(i) < X_*(1)\right)\mathrm{I}\left(X(i) > X(1)\right) = 1.\end{aligned}$$

Recall that $\beta = \alpha - 1$. Subsequently, following (C.8), we have

$$D\left(X_*, Y\right) - D\left(X, Y\right) = -1 + \Sigma_1 + \Sigma_2 \leq -1 + 0 + 1 = 0,$$

which proves our results when the condition (ii) holds.

Suppose the condition

$$\text{(iii)} : R(X(1), X) \geq R(Y(1), Y) \text{ and } R(X_*(1), X_*) = R(X(1), X) + 1$$

holds. Then we have

$$\begin{aligned}&R(X(1), X) - R(Y(1), Y) \geq 0 \\ \implies &\alpha = |R(X(1), X) - R(Y(1), Y)| = R(X(1), Y) - R(Y(1), X),\end{aligned}$$

and

$$\begin{aligned}&R(X_*(1), X_*) - R(Y(1), Y) = R(X(1), X) + 1 - R(Y(1), Y) > 0 \\ \implies &\beta = |R(X_*(1), X_*) - R(Y(1), Y)| = R(X(1), Y) - R(Y(1), X) + 1.\end{aligned}$$

Hence, we have

$$\beta = \alpha + 1.$$

Since $R(X_*(1), X_*) = R(X(1), X) + 1$ and X_* is an imputation of X for the index 1, then

according to Lemma C.3.1, we have

$$X_*(1) > X(1) \implies \mathrm{I}\left(X(i) > X_*(1)\right)\mathrm{I}\left(X(i) < X(1)\right) = 0,$$

for any $i \in \{2, \ldots, n\}$/ Therefore, we have

$$\Sigma_1 = \sum_{i=2}^{n} \mathrm{I}\left(X(i) > X_*(1)\right)\mathrm{I}\left(X(i) < X(1)\right)\left(\xi_i - \psi_i\right) = 0.$$

According to Lemma C.3.1, we also have

$$\sum_{i=2}^{n} \mathrm{I}\left(X(i) < X_*(1)\right)\mathrm{I}\left(X(i) > X(1)\right) = 1.$$

According to (C.7), we have $\xi_i - \psi_i \in \{-1, 1\}$, which follows that

$$\begin{aligned}\Sigma_2 &= \sum_{i=2}^{n} \mathrm{I}\left(X(i) < X_*(1)\right)\mathrm{I}\left(X(i) > X(1)\right)\left(\psi_i - \xi_i\right) \\ &\geq \sum_{i=2}^{n} \mathrm{I}\left(X(i) < X_*(1)\right)\mathrm{I}\left(X(i) > X(1)\right)\cdot(-1) = -1.\end{aligned}$$

Recall that $\beta = \alpha + 1$. Following (C.8), we have

$$D\left(X_*, Y\right) - D\left(X, Y\right) = 1 + \Sigma_1 + \Sigma_2 \geq 1 + 0 - 1 = 0,$$

which proves our results when the condition (iii) holds.

Finally, suppose the condition

$$\text{(iv)} : R(X(1), X) \leq R(Y(1), Y) \text{ and } R(X_*(1), X_*) = R(X(1), X) - 1$$

holds. Then we have

$$\begin{aligned}&R(X(1), X) - R(Y(1), Y) \leq 0, \\ \implies &\alpha = |R(X(1), X) - R(Y(1), Y)| = R(Y(1), X) - R(X(1), Y),\end{aligned}$$

and

$$\begin{aligned}&R(X_*(1), X_*) - R(Y(1), Y) = R(X(1), X) - 1 - R(Y(1), Y) < 0, \\ \implies &\beta = |R(X_*(1), X_*) - R(Y(1), Y)| = R(Y(1), X) - R(X(1), Y) + 1.\end{aligned}$$

Hence, we have

$$\beta = \alpha + 1.$$

Since $R(X_*(1), X_*) = R(X(1), X) - 1$ and X_* is an imputation of X for index 1, then according

to Lemma C.3.1, we have

$$X_*(1) < X(1) \implies \mathrm{I}\left(X(i) < X_*(1)\right) \mathrm{I}\left(X(i) > X(1)\right) = 0,$$

for any $i \in \{2, \ldots, n\}$. Therefore,

$$\Sigma_2 = \sum_{i=2}^{n} \mathrm{I}\left(X(i) < X_*(1)\right) \mathrm{I}\left(X(i) > X(1)\right) \left(\psi_i - \xi_i\right) = 0.$$

According to Lemma C.3.1, we also have

$$\sum_{i=2}^{n} \mathrm{I}\left(X(i) > X_*(1)\right) \mathrm{I}\left(X(i) < X(1)\right) = 1.$$

According to (C.7), we have $\xi_i - \psi_i \in \{-1, 1\}$, which follows that

$$\begin{aligned}
\Sigma_1 &= \sum_{i=2}^{n} \mathrm{I}\left(X(i) > X_*(1)\right) \mathrm{I}\left(X(i) < X(1)\right) \left(\xi_i - \psi_i\right) \\
&\geq \sum_{i=2}^{n} \mathrm{I}\left(X(i) > X_*(1)\right) \mathrm{I}\left(X(i) < X(1)\right) \cdot (-1) = -1.
\end{aligned}$$

Recall that $\beta = \alpha + 1$. Following (C.8), we have

$$D\left(X_*, Y\right) - D\left(X, Y\right) = 1 + \Sigma_1 + \Sigma_2 \geq 1 - 1 + 0 = 0,$$

which proves our results when the condition (iv) holds. □

Lemma C.3.4. Suppose $X \in \mathcal{V}^n$ and $u \in [n]$ is an index. Let r be an integer between 1 to n. Then, there exists an imputation $X_* \in \mathcal{V}^n$ of X for the index u such that $R(X_*(u), X_*) = r$.

Proof. For notation ease, let us assume (after relabelling) that $u = 1$ and $X(2), \cdots, X(n)$ are ordered such that $X(2) < \cdots < X(n)$.

Suppose $r = 1$. Let $X_*(i) = X(i)$ for any $i \in \{2, \cdots, n\}$ and let $X_*(1)$ be a real number such that $X_*(1) < X(2)$. Then X_* is an imputation of X for the index u. Meanwhile, since $X_*(1) < X(2)$, we have

$$\begin{aligned}
&X_*(1) < X(2) < \cdots < X(n) \\
\implies &X_*(1) < X_*(2) < \cdots < X_*(n).
\end{aligned}$$

Thus, according to the definition of rank, we have

$$R(X_*(1), X_*) = \sum_{i=1}^{n} \mathrm{I}\left(X_*(i) \leq X_*(1)\right) = 1.$$

Therefore, $R(X_*(u), X_*) = 1 = r$, which proves our result when $r = 1$.

Suppose $1 < r < n$. Let $X_*(i) = X(i)$ for any $i \in \{2, \cdots, n\}$, and let $X_*(1)$ be a real number such that $X(r) < X_*(1) < X(r+1)$. Then X_* is an imputation of X for the index u. Meanwhile,

since $X(r) < X_*(1) < X(r+1)$, we have

$$\begin{aligned} &X(2) < \cdots < X(r) < X(1) < X(r+1) < \cdots < X(n) \\ \Longrightarrow\ &X_*(2) < \cdots < X_*(r) < X_*(1) < X_*(r+1) < \cdots < X_*(n). \end{aligned}$$

Thus, according to the definition of rank, we have

$$R(X_*(1), X_*) = \sum_{i=1}^{n} \mathrm{I}\left(X_*(i) \leq X_*(1)\right) = r.$$

Therefore, $R(X_*(u), X_*) = r$, which proves our result when $1 < r < n$.

Suppose $r = n$. Let $X_*(i) = X(i)$ for any $i \in \{2, \cdots, n\}$ and let $X_*(1)$ be a real number such that $X_*(1) > X(n)$. Then, X_* is an imputation of X for the index u. Meanwhile, since $X_*(1) > X(n)$, we have

$$\begin{aligned} &X_*(1) > X(n) > \cdots > X(2) \\ \Longrightarrow\ &X_*(1) > X_*(n) > \cdots > X_*(2). \end{aligned}$$

Thus, according to the definition of rank, we have

$$R(X_*(1), X_*) = \sum_{i=1}^{n} \mathrm{I}\left(X_*(i) \leq X_*(1)\right) = n.$$

Therefore, $R(X_*(u), X_*) = n$, which proves our result when $r = n$, and completes our proof.

□

We are now ready to prove Proposition C.3.5.

Proposition C.3.5. Suppose $X, Y \in \mathcal{V}^n$ and let $X_* \in \mathcal{V}^n$ be an imputation of X for an index $u \in [n]$. If we choose $X_*(u)$ such that $R(X_*(u), X_*) = R(Y(u), Y)$, then we have $D\left(X_*, Y\right) \leq D\left(X, Y\right)$. Moreover, for any other imputation $X' \in \mathcal{V}^n$ of X for the index u, we have $D\left(X_*, Y\right) \leq D\left(X', Y\right)$.

Proof. Notice that there are two statements in Proposition C.3.5:

(i) Suppose $X, Y \in \mathcal{V}^n$ and let $X_* \in \mathcal{V}^n$ be an imputation of X for an index $u \in [n]$. If $R(X_*(u), X_*) = R(Y(u), Y)$, then we have $D\left(X_*, Y\right) \leq D\left(X, Y\right)$.

(ii) Suppose $X, Y \in \mathcal{V}^n$ and let $X_* \in \mathcal{V}^n$ be an imputation of X for an index $u \in [n]$. If $R(X_*(u), X_*) = R(Y(u), Y)$, then for any other imputation $X' \in \mathcal{V}^n$ of X for index u, we have $D\left(X_*, Y\right) \leq D\left(X', Y\right)$.

Below, we first show that the statement (i) is true. Then we prove the statement (ii) is true using the statement (i).

First, we prove the statement (i) is true. Suppose $R(X(u), X) = R(Y(u), Y)$. Then, we have

$$R(X(u), X) = R(Y(u), Y) = R(X_*(u), X).$$

Since $X_* \in \mathcal{V}^n$ is an imputation of X for the index $u \in [n]$, then according to Lemma C.2.3, we have $R(X_*, X_*) = R(X, X)$, which implies that

$$D\left(X_*, Y\right) = D\left(X, Y\right).$$

Hence, we have proved the statement (i) when $R(X(u), X) = R(Y(u), Y)$.

Suppose $R(X(u), X) > R(Y(u), Y)$, denote $R(X(u), X) = R(Y(u), Y) + a$, where $a \in \mathbb{N}$ and $a > 0$. According to Lemma C.3.2, we can find vectors $X_1, X_2, \cdots, X_a \in \mathcal{V}^n$ such that each vector is an imputation of X for the index $u \in [n]$ and

$$R(X_k(u), X_k) = R(X(u), X) - k, \text{ for any } k \in \{1, \cdots, a\}.$$

Denote $X_0 = X$. Then, for any $k \in \{1, \cdots, a\}$, we have

$$R(X_{k-1}(u), X_{k-1}) = R(X(u), X) - (k-1) > R(X(u), X) - a = R(Y(u), Y),$$

and

$$R(X_k(u), X_k) = R(X_{k-1}(u), X_{k-1}) - 1.$$

Subsequently, by applying Lemma C.3.3 between each X_k, Y and X_{k-1}, Y, where $k \in \{1, \cdots, a\}$, we have

$$D\left(X, Y\right) = D\left(X_0, Y\right) \geq D\left(X_1, Y\right) \geq \cdots \geq D\left(X_a, Y\right).$$

Notice that

$$R(X_a(u), X_a) = R(X(u), X) - a = R(Y(u), Y) = R(X_*(u), X_*).$$

Since X_a is an imputation of X_* for the index u, then according to Lemma C.2.3, we have $R(X_*, X_*) = R(X_a, X_a)$, which gives

$$D\left(X_*, Y\right) = D\left(X_a, Y\right) \leq D\left(X, Y\right).$$

This proves the statement (i) when $R(X(u), X) > R(Y(u), Y)$.

The case when $R(X(u), X) < R(Y(u), Y)$ can be proved similarly. Hence, we have shown the statement (i) is true.

Next, we show that the statement (ii) is true using the statement (i). Let $X' \in \mathcal{V}^n$ be an imputation of X for the index u. According to Lemma C.3.2, we can find an imputation $X'_* \in \mathcal{V}^n$ of X' for the index $u \in [n]$, and $R(X'_*(u), X_*) = R(Y(u), Y)$. Then, according to the statement (i), we have

$$D\left(X'_*, Y\right) \leq D\left(X', Y\right).$$

Since $X_* \in \mathcal{V}^n$ is an imputation of X for the index $u \in [n]$, and $R(X_*(u), X_*) = R(Y(u), Y)$, then according to Lemma C.2.3, we have $R(X'_*, X'_*) = R(X_*, X_*)$, which implies $D\left(X'_*, Y\right) =$

$D(X_*, Y)$. Thus, we have

$$D(X_*, Y) = D\left(X'_*, Y\right) \leq D\left(X', Y\right).$$

This proves the statement (ii), and completes our proof. □

This proposition demonstrates that if one component of X is changed, Spearman's footrule will be minimised when the rank of that component equals to the rank of its paired component, i.e. $R(X_*(u), X_*) = R(Y(u), Y)$. For example, consider vectors $X = (1.1,\ *,\ 3.3)$, and $Y = (3.9,\ 2.8,\ 1.7)$, where $*$ denotes a missing data in X. Then, the proposition suggests that the minimum Spearman's footrule between X and Y is achieved when the missing entry in X is assigned rank 2, matching the rank of the corresponding value 2.8 in Y.

When considering Proposition C.3.5, a natural concern would be whether the required imputation X_* actually exists. To this end, we verify the existence of such imputations X_* of X, in the more general case where multiple components are imputed so that the ranks of all imputed components match the ranks of their paired components in Y.

Lemma C.3.6. Suppose $X, Y \in \mathcal{V}^n$ and $U \subset [n]$ is a subset of indices. Then there exists $X_* \in \mathcal{V}^n$ such that

$$X_*(i) = X(i) \text{ for } i \in [n] \setminus U, \text{ and } R(X_*(i), X_*) = R(Y(i), Y) \text{ for } i \in U. \tag{C.9}$$

Proof. We first consider the case when $U = [n]$. Let $X_* = Y$, then according to the definition of imputations, we have that $X_* \in \mathcal{V}^n$ is an imputation of X for $U = [n]$. Since $X_* = Y$, we have $R(X_*, X_*) = R(Y, Y)$. Hence, Lemma C.3.6 is true when $U = [n]$.

However, if $U = \emptyset$, we can let $X_* = X$. Then according to the definition of imputations, $X_* \in \mathcal{V}^n$ is an imputation of X for $U = \emptyset$. Hence, Lemma C.3.6 is true when $U = \emptyset$.

Now, we only need to prove Lemma C.3.6 when $U \neq [n]$ and $U \neq \emptyset$.

For any fixed $n \in \mathbb{N}$, let $P(k)$ be the statement of lemma C.3.6 when $|U| = k$. We prove $P(k)$ holds for any $k \in \{1, \ldots, n-1\}$ by induction on k.

Base Case: We show $P(1)$ is true. Suppose $|U| = 1$. Then $P(1)$ is true according to Lemma C.3.2.

Induction Step: We show the implication $P(1), P(k) \implies P(k+1)$ for any $k \in \{1, \ldots, n-2\}$.

Suppose $|U| = k+1$, and denote $m = k+1$. Without loss of generality, let us assume (after relabelling) $U = \{1, \cdots, m\}$ and $Y(1), \cdots, Y(m)$ such that $Y(1) < \cdots < Y(m)$.

Notice that $|U \setminus \{1\}| = |U| - 1 = m - 1 = k$. Then since $P(k)$ is true, there exists $X_1 \in \mathcal{V}^n$ such that

$$X_1(i) = X(i), \quad i \in [n] \setminus (U \setminus \{1\}), \tag{C.10}$$

$$R(X_1(i), X_1) = R(Y(i), Y), \quad i \in U \setminus \{1\}. \tag{C.11}$$

According to Lemma C.2.1, for any $i \in U \setminus \{1\}$, we have

$$R(X_1(i), (X_1(l))_{l \in [n] \setminus \{1\}}) \geq R(X_1(i), X_1) - 1 = R(Y(i), Y) - 1,$$

where the last equation holds according to (C.11). Then, since $Y(1) < \cdots < Y(m)$, we have $R(Y(1), Y) < \cdots < R(Y(m), Y)$. Hence, for any $i \in U \setminus \{1\}$, we have

$$\begin{aligned} & R(X_1(i), (X_1(l))_{l\in[n]\setminus\{1\}}) \geq R(Y(i), Y) - 1 > R(Y(1), Y) - 1 \\ \implies & R(X_1(i), (X_1(l))_{l\in[n]\setminus\{1\}}) \geq R(Y(1), Y). \end{aligned} \tag{C.12}$$

Since $P(1)$ is true, then there exists $X_2 \in \mathcal{V}^n$ such that

$$X_2(i) = X_1(i), \quad i \in [n] \setminus \{1\}, \tag{C.13}$$

$$R(X_2(1), X_2) = R(Y(1), Y). \tag{C.14}$$

According to Lemma C.2.1, for any $i \in U \setminus \{1\}$, we have

$$R(X_2(i), X_2) \geq R(X_2(i), (X_2(l))_{l\in[n]\setminus\{1\}}).$$

Notice that according to (C.13), for any $i \in U \setminus \{1\}$, we have

$$R(X_2(i), (X_2(l))_{l\in[n]\setminus\{1\}}) = R(X_1(i), (X_1(l))_{l\in[n]\setminus\{1\}}).$$

Hence, for any $i \in U \setminus \{1\}$, we have

$$\begin{aligned} R(X_2(i), X_2) \geq R(X_1(i), (X_1(l))_{l\in[n]\setminus\{1\}}) &\geq^{(C.12)} R(Y(1), Y) \\ &=^{(C.14)} R(X_2(1), X_2). \end{aligned}$$

Since $X_2 \in \mathcal{V}^n$ is a vector of distinct real values, for any $i \in U \setminus \{1\}$ we have $R(X_2(i), X_2) \neq R(X_2(1), X_2)$. Hence, for any $i \in U \setminus \{1\}$,

$$\begin{aligned} R(X_2(i), X_2) \geq R(X_2(1), X_2) &\implies R(X_2(i), X_2) > R(X_2(1), X_2) \\ &\implies X_2(i) > X_2(1). \end{aligned}$$

According to Lemma C.2.2, we then have

$$R(X_2(1), (X_2(i))_{i\in[n]\setminus(U\setminus\{1\})}) = R(X_2(1), X_2) =^{(C.14)} R(Y(1), Y). \tag{C.15}$$

Now, since $P(k)$ is true, there exists $X_3 \in \mathcal{V}^n$ such that

$$X_3(i) = X_2(i), \quad i \in [n] \setminus (U \setminus \{1\}), \tag{C.16}$$

$$R(X_3(i), X_3) = R(Y(i), Y), \quad i \in U \setminus \{1\}. \tag{C.17}$$

Since $Y(1) < \cdots < Y(m)$, we have $R(Y(1), Y) < \cdots < R(Y(m), Y)$. Then, for any $i \in U \setminus \{1\}$,

it follows that

$$\begin{aligned} R(X_3(i), X_3) &=^{(\text{C.17})} R(Y(i), Y) > R(Y(1), Y) \\ &=^{(\text{C.15})} R(X_2(1), (X_2(l))_{l\in[n]\setminus(U\setminus\{1\})}) \\ &=^{(\text{C.16})} R(X_3(1), (X_3(l))_{l\in[n]\setminus(U\setminus\{1\})}). \end{aligned}$$

Subsequently, according to Lemma C.2.2, we have

$$R(X_3(1), X_3) = R(X_3(1), (X_3(l))_{l\in[n]\setminus(U\setminus\{1\})}) = R(Y(1), Y).$$

Combining this result with (C.17), it follows that

$$R(X_3(i), X_3) = R(Y(i), Y), \text{ for any } i \in U.$$

Notice that for any $i \in [n] \setminus U$, we have

$$X_3(i) =^{(\text{C.16})} X_2(i) =^{(\text{C.13})} X_1(i) =^{(\text{C.10})} X(i).$$

Thus, we have found an imputation $X_3 \in \mathcal{V}^n$ of X for the indices U such that

$$X_3(i) = X(i), i \in [n] \setminus U, \text{ and } R(X_3(i), X_3) = R(Y(i), Y) \text{ for all } \ i \in U.$$

This proves $P(k+1)$ is true, and completes our proof. □

The proof for Lemma C.3.6 constructs X_* sequentially by considering the order of the values in the paired vector $(Y(u))_{u\in U}$. The first step is to identify u_1, the index in U with the minimum rank in Y, defined as $u_1 = \text{argmin}\{R(Y(u), Y)|u \in U\}$. Then the component $X(u_1)$ is imputed so that its rank satisfies

$$R(X(u_1), (X(i))_{i\in\{[n]\setminus U\}\cup\{u_1\}}) = R(Y(u_1), Y).$$

Following this imputation, the ranks of all observed values in X are updated. U is then updated to $U \setminus \{u_1\}$, and this process is repeated until U is empty. Algorithm 2 below formally describes the procedure of the proof of Lemma C.3.6. Note that in Algorithm 2 we work with the rank vectors $R(Y, Y)$ and $R(X_*, X_*)$, rather than the actual vectors Y and X_*, which suffices for our purpose of computing the value of Spearman's footrule $D(X_*, Y)$.

Remark C.3.7. The computational complexity of Algorithm 2 is analysed as follows. Ranking all observed components in X and Y in line 1 requires $\mathcal{O}(n \log n)$ steps. Inside the while loop, identifying u_1 in line 3 requires the computational complexity $\mathcal{O}(n)$, and updating ranks of observed components between line 5 and line 8 requires the computational complexity $\mathcal{O}(n)$. Denote $m_1 = |U|$. The while loop run m_1 times. Since each iteration is $\mathcal{O}(n)$, the computational complexity of loop is $\mathcal{O}(m_1 n)$. Overall, the computational complexity for Algorithm 2 is $\mathcal{O}(n \log n + m_1 n)$.

To further emphasise that, when computing Spearman's footrule, it is the ranks $R(X_*, X_*)$ of the imputed vector X_* that are important, rather than the imputed values $(X_*(u))_{u\in U}$ themselves,

Algorithm 2 An Algorithm for Computing the Exact Lower Bound of Spearman's Footrule Under Missing case I

Input: $X, Y \in \mathcal{V}^n$, where X may be partially observed and Y is fully observed.
Output: Minimum possible Spearman's footrule distance between X and Y.

Rank all observed components in X and Y.
while $U \neq \emptyset$ **do**
  Identify $u_1 = \operatorname{argmin}\{R(Y(i), Y) | i \in U\}$.
  Let $R(X(u_1), (X(i))_{i\in\{[n]\setminus U\}\cup\{u_1\}}) = R(Y(u_1), Y)$.
  for $i \in [n] \setminus U$ **do**
    Denote $r_i = R(X(i), (X(i))_{i\in[n]\setminus U})$
    Update $R(X(i), (X(j))_{j\in\{[n]\setminus U\}\cup\{u_1\}}) \leftarrow r_i + \mathrm{I}\left(r_i \geq R(Y(u_1), Y)\right)$.
  end for
  Update $U \leftarrow U \setminus \{u_1\}$.
end while
Output $D(X, Y)$.

we have the following lemma which will be useful in proving later results.

Lemma C.3.8. Suppose $X, Y \in \mathcal{V}^n$ and $U \subset [n]$ is a subset of indices. Let $X_1, X_2 \in \mathcal{V}^n$ both be imputations of X for indices U. Then, if for any $i \in U$,

$$R(X_1(i), X_1) = R(X_2(i), X_2) = R(Y(i), Y),$$

we have $D(X_1, Y) = D(X_2, Y)$.

Proof. Since $X_1, X_2 \in \mathcal{V}^n$ are both imputations of X for U, for any $i \in [n] \setminus U$, we have

$$X_1(i) = X_2(i) = X(i).$$

Therefore, according to the definition of imputations, X_2 is an imputation of X_1 for indices U. Since for any $i \in U$, we have

$$R(X_1(i), X_1) = R(X_2(i), X_2).$$

Then, according to Lemma C.2.3, we have

$$R(X_1, X_1) = R(X_2, X_2).$$

Hence,

$$D(X_1, Y) = D(X_2, Y).$$

This completes our proof. □

We are now ready to state the main result for missing case I, which is an extension of Proposition C.3.5.

Theorem C.3.9. Suppose $X, Y \in \mathcal{V}^n$, and $U \subset [n]$ is a subset of indices. Let $X_* \in \mathcal{V}^n$ be an imputation of X for indices U such that $R(X_*(i), X_*) = R(Y(i), Y)$ for any $i \in U$. Then, we

have $D(X_*, Y) \leq D(X, Y)$. Moreover, for any other imputation $X' \in \mathcal{V}^n$ of X for U, we have $D(X_*, Y) \leq D(X', Y)$.

Proof. Notice that there are two statements in Theorem C.3.9:

(i) Suppose $X, Y \in \mathcal{V}^n$ and let $X_* \in \mathcal{V}^n$ be an imputation of X for indices $U \subset [n]$. If $R(X_*(i), X_*) = R(Y(i), Y)$ for any $i \in U$, then we have $D(X_*, Y) \leq D(X, Y)$.

(ii) Suppose $X, Y \in \mathcal{V}^n$ and let $X_* \in \mathcal{V}^n$ be an imputation of X for indices $U \in [n]$. If $R(X_*(i), X_*) = R(Y(i), Y)$, for any $i \in U$, then for any other imputation $X' \in \mathcal{V}^n$ of X for indices U, we have $D(X_*, Y) \leq D(X', Y)$.

Below, we first show that the statement (i) is true. Then, we prove the statement (ii) is true using the statement (i).

First, we show that the statement (i) is true. Denote $m = |U|$. When $m = n$, we have $R(X_*, X_*) = R(Y, Y) \implies D(X_*, Y) = 0$. According to the definition of Spearman's footrule, we have $D(X, Y) \geq 0$. Hence, we have $D(X, Y) \geq D(X_*, Y) = 0$. Therefore, we have shown the statement (i) is true when $m = n$.

However, when $m = 0$, then according to the definition of imputations, we have $X = X_*$. Then, we have $D(X_*, Y) = D(X, Y)$, which proves the statement (i) when $m = 0$.

Now, in order to prove the statement (i), we only need to consider cases when $1 \leq m \leq n-1$. For any fixed $n \in \mathbb{N}$, let $P(k)$ be the statement of statement (i) when $|U| = k$. We prove $P(k)$ holds for any $k \in \{1, \ldots, n-1\}$ by induction on k.

Base Case: Notice that $P(1)$ is true according to Proposition C.3.5.

Induction Step: We show the implication $P(1), P(k) \implies P(k+1)$ for any $k \in \{1, \ldots, n-2\}$

For convenience, let us assume (after relabelling) $U = \{1, \cdots, m\}$, and $Y(1) < \cdots < Y(m)$. According to Lemma C.3.6, we can find a vector $X_1 \in \mathcal{V}^n$ such that

$$X_1(i) = X(i), \quad \text{for any } i \in [n] \setminus (U \setminus \{1\}), \tag{C.18}$$

$$R(X_1(i), X_1) = R(Y(i), Y), \quad \text{for any } i \in U \setminus \{1\}. \tag{C.19}$$

That is, X_1 is an imputation of X for indices $U \setminus \{1\}$ such that (C.19) holds. Notice that $|U \setminus \{1\}| = |U| - 1 = k + 1 - 1 = k$. Then since $P(k)$ is true, we have

$$D(X_1, Y) \leq D(X, Y).$$

According to Lemma C.3.6, we can find a vector $X_2 \in \mathcal{V}^n$ such that

$$X_2(i) = X_1(i), \quad \text{for any } i \in [n] \setminus \{1\}, \tag{C.20}$$

$$R(X_2(1), X_2) = R(Y(1), Y). \tag{C.21}$$

That is, X_2 is an imputation of X_1 for index 1 such that (C.21) holds. Subsequently, since $P(1)$ is true, we have

$$D(X_2, Y) \leq D(X_1, Y).$$

According to Lemma C.2.1, for any $i \in [n] \setminus \{1\}$, we have

$$R(X_2(i), X_2) \geq R(X_2(i), (X_2(l))_{l\in[n]\setminus\{1\}}) =^{(C.20)} R(X_1(i), (X_1(l))_{l\in[n]\setminus\{1\}}).$$

Subsequently, applying Lemma C.2.1 again, for any $i \in U \setminus \{1\}$, we have

$$R(X_1(i), (X_1(l))_{l\in[n]\setminus\{1\}}) \geq R(X_1(i), X_1) - 1 =^{(C.19)} R(Y(i), Y) - 1.$$

Since $Y(1) < \cdots < Y(m)$ and $U = \{1, \cdots, m\}$, then for any $i \in U \setminus \{1\}$, we have

$$\begin{aligned} & R(X_1(i), (X_1(l))_{l\in[n]\setminus\{1\}}) \geq R(Y(i), Y) - 1 > R(Y(1), Y) - 1 \\ \implies & R(X_1(i), (X_1(l))_{l\in[n]\setminus\{1\}}) \geq R(Y(1), Y). \end{aligned}$$

Hence, for any $i \in U \setminus \{1\}$, we have

$$R(X_2(i), X_2) \geq R(X_1(i), (X_1(l))_{l\in[n]\setminus\{1\}}) \geq R(Y(1), Y) =^{(C.21)} R(X_2(1), X_2).$$

Since $X_2 \in \mathcal{V}^n$ is a vector of distinct real values, for any $i \in U \setminus \{1\}$, we further have

$$\begin{aligned} R(X_2(i), X_2) \neq R(X_2(1), X_2) &\implies R(X_2(i), X_2) > R(X_2(1), X_2) \\ &\implies X_2(i) > X_2(1). \end{aligned}$$

Subsequently, according to Lemma C.2.2,

$$R(X_2(1), (X_2(i))_{i\in[n]\setminus(U\setminus\{1\})}) = R(X_2(1), X_2) =^{(C.21)} R(Y(1), Y). \tag{C.22}$$

According to Lemma C.3.6, we can find a vector $X_3 \in \mathcal{V}^n$ such that

$$X_3(i) = X_2(i), \quad \text{for any } i \in [n] \setminus (U \setminus \{1\}), \tag{C.23}$$

$$R(X_3(i), X_3) = R(Y(i), Y), \quad \text{for any } i \in U \setminus \{1\}. \tag{C.24}$$

That is, X_3 is an imputation of X_2 for indices $U \setminus \{1\}$ such that (C.24) holds. Notice that $|U \setminus \{1\}| = |U| - 1 = k + 1 - 1 = k$. Then since $P(k)$ is true, we have

$$D(X_3, Y) \leq D(X_2, Y).$$

Since $Y(1) < \cdots < Y(m)$, we have $R(Y(i), Y) > R(Y(1), Y)$ for any $i \in U \setminus \{1\}$. Hence, for any $i \in U \setminus \{1\}$, we have

$$\begin{aligned} R(X_3(i), X_3) =^{(C.24)} R(Y(i), Y) &> R(Y(1), Y) \\ &=^{(C.22)} R(X_2(1), (X_2(i))_{i\in[n]\setminus(U\setminus\{1\})}). \end{aligned}$$

Further, according to (C.23),

$$R(X_2(1), (X_2(l))_{l\in[n]\setminus(U\setminus\{1\})}) = R(X_3(1), (X_3(l))_{l\in[n]\setminus(U\setminus\{1\})}).$$

For any $i \in U \setminus \{1\}$, we then have

$$R(X_3(i), X_3) > R(X_2(1), (X_2(l))_{l\in[n]\setminus(U\setminus\{1\})}) = R(X_3(1), (X_3(l))_{l\in[n]\setminus(U\setminus\{1\})}).$$

By applying Lemma C.2.2, we have

$$\begin{aligned} R(X_3(1), X_3) &= R(X_3(1), (X_3(l))_{l\in[n]\setminus(U\setminus\{1\})}) \\ &= R(X_2(1), (X_2(l))_{l\in[n]\setminus(U\setminus\{1\})}) =^{(C.22)} R(Y(1), Y). \end{aligned}$$

Combining this result with (C.24), we have

$$R(X_3(i), X_3) = R(Y(i), Y), \text{ for any } i \in U.$$

Notice that for any $i \in [n] \setminus U$, we have

$$X_3(i) =^{(C.23)} X_2(i) =^{(C.20)} X_1(i) =^{(C.18)} X(i).$$

Since $X_* \in \mathcal{V}^n$ is an n-dimensional vector of distinct real values such that

$$\begin{aligned} X_*(i) &= X(i), \quad \text{for any } i \in [n] \setminus U, \\ R(X_*(i), X_*) &= R(Y(i), Y), \quad \text{for any } i \in U. \end{aligned}$$

Then according to Lemma C.2.3, we have

$$R(X_3, X_3) = R(X_*, X_*) \implies D(X_3, Y) = D(X_*, Y).$$

Notice that we have

$$D(X_*, Y) = D(X_3, Y) \leq D(X_2, Y) \leq D(X_1, Y) \leq D(X, Y).$$

Hence, we have shown $P(k+1)$ is true, this completes our proof for the statement (i).

Next, we show that the statement (ii) is true using the statement (i). According to Lemma C.3.6, we can find a vector $X'_* \in \mathcal{V}^n$ such that it is an imputation of X' for indices $U \subset [n]$, and $R(X'_*(i), X'_*) = R(Y(i), Y)$ for any $i \in U$. Then, according to the statement (i), we have

$$D(X'_*, Y) \leq D(X', Y).$$

Since $X_* \in \mathcal{V}^n$ is an imputation of X for the index $U \subset [n]$, and $R(X_*(i), X_*) = R(Y(i), Y)$ for any $i \in U$. Then, according to Lemma C.2.3, we have $R(X'_*, X'_*) = R(X_*, X_*)$, which means $D(X'_*, Y) = D(X_*, Y)$. Therefore, we have

$$D(X_*, Y) = D(X'_*, Y) \leq D(X', Y).$$

This proves the statement (ii) and completes our proof. □

Theorem C.3.9 provides the conditions under which the imputation X_* will minimise Spearman's footrule under missing case I. Moreover, Lemma C.3.6 shows that X_* exists and so this

lower bound is tight.

C.3.2. Lower bound under missing case II

This section focuses on the lower bound of Spearman's footrule under missing case II, where $X, Y \in \mathcal{V}^n$ may contain unobserved samples, but for each pair $(X(i), Y(i))$, at least one value is observed. We shall use $U \subset [n]$ to denote the set of indices where $X(i)$ is missing but $Y(i)$ is observed, and $V \subset [n]$ will denote the set of indices where $X(i)$ is observed but $Y(i)$ is missing. Given this notation, consider the following two conditions:

Condition C.3.10. $X_*, Y_* \in \mathcal{V}^n$ are imputations of X and Y for indices $U, V \subset [n]$ respectively, such that $U \cap V = \emptyset$ and

$$X_*(i) = X(i) \text{ for all } i \in [n] \setminus U, \text{ and } Y_*(j) = Y(j) \text{ for all } j \in [n] \setminus V.$$

Condition C.3.11. $X_*, Y_* \in \mathcal{V}^n$ are such, for indices $U, V \subset [n]$,

$$R(X_*(i), X_*) = R(Y_*(i), Y_*), i \in U \cup V.$$

Condition C.3.10 ensures that X_* and Y_* are imputations of X and Y satisfying missing case II. Condition C.3.11 ensures the ranks of the imputed components of X_* and Y_* are equal to the ranks of their paired components.

It is proved in Theorem C.4.13 that the minimum possible value of Spearman's footrule between X and Y equals $D(X_*, Y_*)$ when X_* and Y_* satisfy both Conditions C.3.10 and C.3.11. While finding X_* and Y_* that satisfy Condition C.3.10 is straightforward, it is less obvious that we can always find imputations that satisfy both Conditions C.3.10 and C.3.11. Hence, we first prove the existence of such imputations in Proposition C.3.13.

To start, we first provide the following lemma, which is obtained directly by applying Lemma C.3.6.

Lemma C.3.12. Suppose $X \in \mathcal{V}^n$ and $U \subset [n]$ is a subset of indices. Denote $m = |U|$, and let $(\sigma(l))_{l \in U}$ be a vector of any m distinct values in $[n]$. Then, there exist an imputation $X_* \in \mathcal{V}^n$ of X for U such that $R(X_*(i), X_*) = \sigma(i), \text{ for any } i \in U.$

Proof. Let $Y \in \mathcal{V}^n$ be a vector such that $R(Y(i), Y) = \sigma(i), \text{ for any } i \in U$. Subsequently, according to Lemma C.3.6, there exist an imputation $X_* \in \mathcal{V}^n$ of X for indices U such that

$$R(X_*(i), X_*) = R(Y(i), Y) = \sigma(i), \text{ for any } i \in U.$$

This completes our proof. □

We are now ready to prove Proposition C.3.13.

Proposition C.3.13. Suppose $X, Y \in \mathcal{V}^n$, and $U, V \subset [n]$ are disjoint subsets of $[n]$ such that $U \cap V = \emptyset$ and $U \cup V \neq \emptyset$. Then, there exist imputations $X_*, Y_* \in \mathcal{V}^n$ satisfying both Conditions C.3.10 and C.3.11. In other words, there exist imputations $X_*, Y_* \in \mathcal{V}^n$ of X and Y for U and V, respectively, such that $R(X_*(i), X_*) = R(Y_*(i), Y_*)$ for any $i \in U \cup V$.

Proof. For any given $n \in \mathbb{N}$, let $P(k)$ be the statement of Proposition C.3.13 when $|U|+|V| = k$. We prove $P(k)$ holds for any $k \in \{1, \ldots, n\}$ by induction on k.

Base Case: We show $P(1)$ holds. If $|U|+|V| = 1$, it is either $U = \emptyset, |V| = 1$ or $|U| = 1, V = \emptyset$. Suppose $U = \emptyset, |V| = 1$. Then, let $X_* = X$. According to the definition of imputations, $X_* \in \mathcal{V}^n$ is an imputation of X for $U = \emptyset$. Further, according to Lemma C.3.6, we can find an imputation $Y_* \in \mathcal{V}^n$ of Y for indices V such that $R(Y_*(i), Y_*) = R(X(i), X)$ for any $i \in V$. Notice that $X_* = X$. Hence, we have $R(Y_*(i), Y_*) = R(X_*(i), X_*)$ for any $i \in V$. Recall that $U = \emptyset$. Therefore, we have $R(Y_*(i), Y_*) = R(X_*(i), X_*)$ for any $i \in U \cup V$. Hence, we have shown $P(1)$ is true when $U = \emptyset, |V| = 1$. The case when $|U| = 1, V = \emptyset$ can be proved similarly.

Induction Step: We show the implication $P(k) \implies P(k+1)$ for any $k \in \{1, \ldots, n-1\}$.

Suppose $|U| + |V| = k + 1$. Then, we have either

$$\text{(i)} : |U| = 0, |V| = k+1 \text{ or } |U| = k+1, |V| = 0,$$

or

$$\text{(ii)} : |U| > 0, |V| > 0, |U| + |V| = k+1,$$

is true. Below, we consider the case (i) and the case (ii) separately.

First, suppose the case (i) : $|U| = 0, |V| = k+1$ or $|U| = k+1, |V| = 0$ is true. We consider $|U| = 0, |V| = k+1$. Since $|U| = 0$, we have $U = \emptyset$. Let $X_* = X$. According to the definition of imputations, $X_* \in \mathcal{V}^n$ is an imputation of X for $U = \emptyset$. Further, according to Lemma C.3.6, we can find an imputation $Y_* \in \mathcal{V}^n$ of Y for indices V such that $R(Y_*(i), Y_*) = R(X(i), X)$ for any $i \in V$. Notice that $X_* = X$. Hence, we have $R(Y_*(i), Y_*) = R(X_*(i), X_*)$ for any $i \in V$. Recall that $U = \emptyset$. Therefore, we have $R(Y_*(i), Y_*) = R(X_*(i), X_*)$ for any $i \in U \cup V$. Hence, we prove the result when $|U| = 0, |V| = k+1$. The case when $|U| = k+1, |V| = 0$ can be proved similarly. Hence, we have shown $P(k+1)$ is true under the case (i).

Now, we consider the case (ii) : $|U| > 0, |V| > 0, |U| + |V| = k+1$. Without loss of generality, let us assume (after relabelling) $V = \{1, \cdots, t\}$, $U = \{t+1, \cdots, t+m\}$, $X(1) < \cdots < X(t)$, and $Y(t+1) < \cdots < Y(t+m)$. Further, let us assume

$$R(X(1), (X(l))_{l \in [n] \setminus U}) \leq R(Y(t+1), (Y(l))_{l \in [n] \setminus V}).$$

However, if $R(X(1), (X(l))_{l \in [n] \setminus U}) > R(Y(t+1), (Y(l))_{l \in [n] \setminus V})$, we can then switch the label for X and Y and relabel the relevant components.

Denote

$$\alpha = R(X(1), (X(l))_{l \in [n] \setminus U}). \tag{C.25}$$

Since $X(1) < \cdots < X(t)$ and $Y(t+1) \cdots < Y(t+m)$, we have

$$R(X(1), (X(l))_{l \in [n] \setminus U}) < \cdots < R(X(t), (X(l))_{l \in [n] \setminus U}), \tag{C.26}$$

$$\text{and } R(Y(t+1), (Y(l))_{l \in [n] \setminus V}) < \cdots < R(Y(t+m), (Y(l))_{l \in [n] \setminus V}), \tag{C.27}$$

respectively. Further, since $R(X(1),(X(l))_{l\in[n]\setminus U}) \le R(Y(t+1),(Y(l))_{l\in[n]\setminus V})$, we have

$$\alpha = R(X(1),(X(l))_{l\in[n]\setminus U}) \le R(Y(i),(Y(l))_{l\in[n]\setminus V}), \text{ for any } i \in U. \tag{C.28}$$

Since $X \in \mathcal{V}^n$ is a vector of distinct real values, then $\{R(X(i),(X(l))_{l\in[n]\setminus U}) : i \in V\}$ is a set of t distinct values in $[n]$. Subsequently, according to Lemma C.3.12, there exist $Y_1 \in \mathcal{V}^n$ such that

$$Y_1(i) = Y(i), \quad i \in [n] \setminus V, \tag{C.29}$$
$$R(Y_1(i), Y_1) = R(X(i),(X(l))_{l\in[n]\setminus U}), \quad i \in V. \tag{C.30}$$

That is, Y_1 is an imputation of Y for indices $V \in [n]$ such that (C.30) holds.

Recall that $V = \{1,\cdots,t\}$. Let us assume $V \setminus \{1\} \neq \emptyset$, i.e., $t > 1$. Then, for any $i \in V \setminus \{1\}$, we have

$$\begin{aligned} &R(Y_1(i), Y_1) =^{(C.30)} R(X(i),(X(l))_{l\in[n]\setminus U}) \\ &>^{(C.26)} R(X(1),(X(l))_{l\in[n]\setminus U}) =^{(C.30)} R(Y_1(1), Y_1) \\ \Longrightarrow\ & Y_1(i) > Y_1(1), \text{ for any } i \in V \setminus \{1\}. \end{aligned}$$

Subsequently, by applying Lemma C.2.2, we have

$$R(Y_1(1),(Y_1(l))_{l\in[n]\setminus(V\setminus\{1\})}) = R(Y_1(1), Y_1).$$

Notice the above equation still holds when $V \setminus \{1\} = \emptyset$. Recall that $\alpha = R(X(1),(X(l))_{l\in[n]\setminus U})$. Hence, we further have

$$\begin{aligned} R(Y_1(1),(Y_1(l))_{l\in[n]\setminus(V\setminus\{1\})}) &= R(Y_1(1), Y_1) \\ &=^{(C.30)} R(X(1),(X(l))_{l\in[n]\setminus U}) = \alpha. \end{aligned} \tag{C.31}$$

Since U and V are non-empty disjoint subsets of $[n]$, we have $U \subset [n] \setminus V$. Hence, (C.29) implies $Y_1(i) = Y(i)$ for any $i \in U$. Thus, for any $i \in U$, we have

$$R(Y_1(i),(Y_1(l))_{l\in[n]\setminus V}) = R(Y(i),(Y(l))_{l\in[n]\setminus V}).$$

Since $[n] \setminus V \subset [n] \setminus (V \setminus \{1\})$, we can apply Lemma C.2.1 and get that, for any $i \in U$,

$$\begin{aligned} R(Y_1(i),(Y_1(l))_{l\in[n]\setminus(V\setminus\{1\})}) &\ge R(Y_1(i),(Y_1(l))_{l\in[n]\setminus V}) \\ &= R(Y(i),(Y(l))_{l\in[n]\setminus V}) \\ &\ge^{(C.28)} R(X_1(1),(X_1(l))_{l\in[n]\setminus U}) = \alpha. \end{aligned}$$

Recall that $Y_1 \in \mathcal{V}^n$ is a vector of distinct real numbers and $1 \notin U = \{t+1,\ldots,t+m\}$, where $1 \le t, m$. Hence, for any $i \in U$, we have

$$R(Y_1(i),(Y_1(l))_{l\in[n]\setminus(V\setminus\{1\})}) \neq R(Y_1(1),(Y_1(l))_{l\in[n]\setminus(V\setminus\{1\})}) =^{(C.31)} \alpha.$$

Therefore, for any $i \in U$, we have

$$R(Y_1(i), (Y_1(l))_{l\in[n]\setminus(V\setminus\{1\})}) > \alpha. \tag{C.32}$$

Since $|V \setminus \{1\}| = |V| - 1 = m - 1$ and $|U| = t$, we have $|V \setminus \{1\}| + |U| = m - 1 + t = k$. Since $P(k)$ is true, then there exist $X_2, Y_2 \in \mathcal{V}^n$ such that

$$X_2(i) = X(i), \text{ for any } i \in [n] \setminus U, \tag{C.33}$$

$$Y_2(i) = Y_1(i), \text{ for any } i \in [n] \setminus (V \setminus \{1\}), \tag{C.34}$$

$$\text{and } R(X_2(i), X_2) = R(Y_2(i), Y_2), \text{ for any } i \in U \cup (V \setminus \{1\}). \tag{C.35}$$

That is, X_2, Y_2 are imputations of X and Y_1 for U and $V \setminus \{1\}$, respectively, such that (C.35) holds. According to (C.34), we have

$$R(Y_2(1), (Y_2(l))_{l\in[n]\setminus(V\setminus\{1\})}) = R(Y_1(1), (Y_1(l))_{l\in[n]\setminus(V\setminus\{1\})}) =^{(C.31)} \alpha. \tag{C.36}$$

Since $V = \{1, \ldots, t\}$, we have $[n] \setminus V = \{t+1, \ldots, n\}$, and $[n] \setminus (V \setminus \{1\}) = \{1, t+1, \ldots, n\}$. Then, since $U = \{t+1, \ldots, t+m\}$, we have $U \subset [n] \setminus V$, and $U \subset [n] \setminus (V \setminus \{1\})$. Hence, according to (C.34), we have for any $i \in U$,

$$R(Y_2(i), (Y_2(l))_{l\in[n]\setminus(V\setminus\{1\})}) = R(Y_1(i), (Y_1(l))_{l\in[n]\setminus(V\setminus\{1\})}) >^{(C.32)} \alpha.$$

According to (C.33), we have $X_2(1) = X(1)$, and

$$R(X_2(1), (X_2(l))_{l\in[n]\setminus U}) = R(X(1), (X(l))_{l\in[n]\setminus U}) = \alpha.$$

Hence, for any $i \in U$

$$R(Y_2(i), (Y_2(l))_{l\in[n]\setminus(V\setminus\{1\})}) > \alpha = R(X_2(1), (X_2(l))_{l\in[n]\setminus U}).$$

According to Lemma C.2.1, for any $i \in U$, we have

$$R(Y_2(i), Y_2) \geq R(Y_2(i), (Y_2(l))_{l\in[n]\setminus(V\setminus\{1\})}),$$

which implies that

$$R(Y_2(i), Y_2) > \alpha = R(X_2(1), (X_2(l))_{l\in[n]\setminus U}).$$

Subsequently, for any $i \in U$, we have

$$R(X_2(i), X_2) =^{(C.35)} R(Y_2(i), Y_2) > \alpha = R(X_2(1), (X_2(l))_{l\in[n]\setminus U}).$$

According to Lemma C.2.2, we then have

$$R(X_2(1), X_2) = R(X_2(1), (X_2(l))_{l\in[n]\setminus U}) = \alpha. \tag{C.37}$$

Since $V \subset [n] \setminus U$, (C.33) gives $X_2(i) = X(i)$ for any $i \in V$, and

$$R(X_2(i), (X_2(l))_{l\in[n]\setminus U}) = R(X(i), (X(l))_{l\in[n]\setminus U}), \quad i \in V.$$

Recall that $V = \{1, \ldots, t\}$. Let us assume $V \setminus \{1\} \neq \emptyset$. Then, for any $i \in V \setminus \{1\}$, we have

$$R(X_2(i), (X_2(l))_{l\in[n]\setminus U}) = R(X(i), (X(l))_{l\in[n]\setminus U}).$$

Since $X(t) > \ldots > X(1)$, we have

$$\begin{aligned} R(X_2(i), (X_2(l))_{l\in[n]\setminus U}) &= R(X(i), (X(l))_{l\in[n]\setminus U}) \\ &> R(X(1), (X(l))_{l\in[n]\setminus U}) =^{(C.25)} \alpha. \end{aligned}$$

According to Lemma C.2.1, for any $i \in V \setminus \{1\}$, we have

$$R(X_2(i), X_2) \geq R(X_2(i), (X_2(l))_{l\in[n]\setminus U}) > \alpha.$$

Then, for any $i \in V \setminus \{1\}$, we have

$$R(Y_2(i), Y_2) =^{(C.35)} R(X_2(i), X_2) > \alpha =^{(C.36)} R(Y_2(1), (Y_2(l))_{l\in[n]\setminus(V\setminus\{1\})}).$$

According to Lemma C.2.2, we have

$$R(Y_2(1), Y_2) = R(Y_2(1), (Y_2(l))_{l\in[n]\setminus(V\setminus\{1\})}).$$

Notice that the above equation still holds when $V \setminus \{1\} = \emptyset$.

Subsequently, we have

$$R(Y_2(1), Y_2) = R(Y_2(1), (Y_2(l))_{l\in[n]\setminus(V\setminus\{1\})}) =^{(C.36)} \alpha =^{(C.37)} R(X_2(1), X_2).$$

Combining this result with (C.35), we obtain

$$R(X_2(i), X_2) = R(Y_2(i), Y_2), \text{ for any } i \in U \cup V. \tag{C.38}$$

Recall that we have

$$Y_2(i) =^{(C.34)} Y_1(i) =^{(C.30)} Y(i), \text{ for any } i \in V,$$

and

$$X_2(i) =^{(C.33)} X(i), \text{ for any } i \in [n] \setminus U.$$

Hence, we have found imputations $X_2, Y_2 \in \mathcal{V}^n$ of X and Y for U and V, respectively, such that (C.38) holds. Therefore, we have shown $P(k+1)$ is true. This proves our results when the case (ii) holds and completes our proof. □

The proof of Proposition C.3.13 follows a similar idea to the proof of Lemma C.3.6. The

construction of X_* and Y_* is summarised in Algorithm 3 below.

Algorithm 3 An algorithm for computing the exact lower bound of Spearman's footrule under missing case II and general missingness case

Input: $X, Y \in \mathcal{V}^n$, where both vectors may be partially observed under missing case II.
Output: Minimum possible Spearman's footrule distance between X and Y.

1: Ignore all components where values in both X and Y are missing.
2: Rank all observed components in X, Y.
3: **while** $U \neq \emptyset$ and $V \neq \emptyset$ **do**
4: Let $u_1 = \text{argmin}\left\{R(X(i), (X(j))_{j\in[n]\setminus U}) | i \in V\right\}$.
5: Let $u_2 = \text{argmin}\left\{R(Y(i), (Y(j))_{j\in[n]\setminus V}) | i \in U\right\}$.
6: **if** $R(X(u_1), (X(j))_{j\in[n]\setminus U}) \leq R(Y(u_2), (Y(j))_{j\in[n]\setminus V})$ **then**
7: Let $R(Y(u_1), (Y(j))_{j\in\{[n]\setminus V\}\cup\{u_1\}}) = R(X(u_1), (X(j))_{j\in[n]\setminus U})$
8: **for** $i \in [n] \setminus V$ **do**
9: Denote $R(Y(i), (Y(j))_{j\in[n]\setminus V}) = s_i$.
10: Update

$$R(Y(i), (Y(j))_{j\in\{[n]\setminus V\}\cup\{u_1\}}) \leftarrow s_i + \mathrm{I}\left(s_i \geq R(X(u_1), (X(j))_{j\in[n]\setminus U})\right).$$

11: **end for**
12: Update $V \leftarrow V \setminus \{u_1\}$.
13: **else**
14: Let $R(X(u_2), (X(j))_{j\in\{[n]\setminus U\}\cup\{u_2\}}) = R(Y(u_2), (Y(j))_{j\in[n]\setminus V})$
15: **for** $i \in [n] \setminus U$ **do**
16: Denote $R(X(i), (X(j))_{j\in[n]\setminus U}) = r_i$.
17: Update

$$R(X(i), (X(j))_{j\in\{[n]\setminus U\}\cup\{u_2\}}) \leftarrow r_i + \mathrm{I}\left(r_i \geq R(Y(u_2), (Y(j))_{j\in[n]\setminus V})\right).$$

18: **end for**
19: Update $U \leftarrow U \setminus \{u_2\}$.
20: **end if**
21: **end while**
22: **if** $U \neq \emptyset$ or $V \neq \emptyset$ **then**
23: Run Algorithm 2 when $U \neq \emptyset$. Reverse X and Y and run Algorithm 2 when $V \neq \emptyset$.
24: **end if**

Remark C.3.14. The computational complexity of Algorithm 3 is analysed as follows. Ranking all observed components in X and Y in line 4takes $\mathcal{O}(n \log n)$ steps. Inside the while loop, the computational complexity for finding the minimum ranks in line 4 and 5 is $\mathcal{O}(n)$, and for updating ranks of observed components from line 8 to line 11, and line 15 to line 18 is $\mathcal{O}(n)$. Let $|U| = m_1$, and $|V| = m_2$. The loop can run up to $m_1 + m_2 - 1$ times. Since each iteration is $\mathcal{O}(n)$, the computational complexity for the while loop is $\mathcal{O}(n(m_1+m_2))$. The computational complexity for Algorithm 2, called in line 23, after ranking all observed components is $\mathcal{O}(nm_1)$ when $U \neq \emptyset$ and $\mathcal{O}(nm_2)$ when $V \neq \emptyset$, according to Remark C.3.7. Therefore, the total computation complexity for Algorithm 3 is $\mathcal{O}(n \log n + n(m_1 + m_2))$.

We now prove Proposition C.3.23, which shows that all imputations X_* and Y_* satisfying both Conditions C.3.10 and C.3.11 yield the same value of Spearman's footrule. First, we prove several lemmas which will be useful for proving the proposition.

Table C.2.: (Left) Values of (X, Y) under missing case II, and an example of two imputations (X_1, Y_1) and (X_2, Y_2) such that the ranks of a, b, d, e are the same for both imputations, while preserving the relative ranks of the observed elements. (Right) Ranks of (X, Y) and imputations (X_1, Y_1) and (X_2, Y_2). The imputed components are underlined. Proposition C.3.13 guarantees the existence of such imputations, while Algorithm 3 provides a method for constructing such imputations. Proposition C.3.23 shows that all such imputations yield the same value of Spearman's footrule. Theorem C.3.29 further demonstrates that these imputations minimise Spearman's footrule.

Values	a	b	c	d	e	f	g	h
X	6.3	2.4	5.5	*	*	4.7	3.8	1.8
Y	*	*	1.1	2.4	3.5	4.2	5.8	6.7
X_1	6.3	2.4	5.5	2.9	3.1	4.7	3.8	1.8
Y_1	8.2	2.1	1.1	2.4	3.5	4.2	5.8	6.7
X_2	6.3	2.4	5.5	2.0	2.3	4.7	3.8	1.8
Y_2	8.2	4.1	1.1	2.4	3.5	4.2	5.8	6.7

Ranks	a	b	c	d	e	f	g	h
$R(X, X)$	6	2	5	*	*	4	3	1
$R(Y, Y)$	*	*	1	2	3	4	5	6
$R(X_1, X_1)$	8	2	7	$\underline{3}$	$\underline{4}$	6	5	1
$R(Y_1, Y_1)$	$\underline{8}$	$\underline{2}$	1	3	4	5	6	7
$R(X_2, X_2)$	8	4	7	$\underline{2}$	$\underline{3}$	6	5	1
$R(Y_2, Y_2)$	$\underline{8}$	$\underline{4}$	1	2	3	5	6	7

Lemma C.3.15. Suppose $X_1, X_2 \in \mathcal{V}^n$, and $O \subset [n]$ is a non-empty subset of indices such that $O \neq [n]$. Suppose there exist an index $k \in [n] \setminus O$ such that

$$R(X_1(k), (X_1(l))_{l \in \{k\} \cup O}) = R(X_2(k), (X_2(l))_{l \in \{k\} \cup O}),$$

Then, for any $k_1, k_2 \in O$, if

$$R(X_1(k_1), (X_1(l))_{l \in O}) = R(X_2(k_2), (X_2(l))_{l \in O}),$$

we have

$$R(X_1(k_1), (X_1(l))_{l \in \{k\} \cup O}) = R(X_2(k_2), (X_2(l))_{l \in \{k\} \cup O}).$$

Proof. To start, let us denote

$$\alpha = R(X_1(k), (X_1(l))_{l \in \{k\} \cup O}) = R(X_2(k), (X_2(l))_{l \in \{k\} \cup O}).$$

Suppose $k_1, k_2 \in O$ are such that

$$R(X_1(k_1), (X_1(l))_{l \in O}) = R(X_2(k_2), (X_2(l))_{l \in O}).$$

Then, we have either

$$\begin{aligned} \text{case (i)} : \alpha > R(X_1(k_1), (X_1(l))_{l \in O}) = R(X_2(k_2), (X_2(l))_{l \in O}), \\ \text{or case (ii)} : \alpha \leq R(X_1(k_1), (X_1(l))_{l \in O}) = R(X_2(k_2), (X_2(l))_{l \in O}), \end{aligned}$$

is true. In the following, we prove our result when the case (i) and the case (ii) are true separately.

Suppose the case

$$\text{(i)} : \alpha > R(X_1(k_1), (X_1(l))_{l \in O}) = R(X_2(k_2), (X_2(l))_{l \in O})$$

is true. Then it follows

$$\begin{aligned}\alpha = R(X_1(k), (X_1(l))_{l\in\{k\}\cup O}) > R(X_1(k_1), (X_1(l))_{l\in O}),\\ \text{and } \alpha = R(X_2(k), (X_2(l))_{l\in\{k\}\cup O}) > R(X_2(k_2), (X_2(l))_{l\in O}).\end{aligned}$$

Applying Lemma C.2.2 to $(X_1(l))_{l\in\{k\}\cup O}$ and $(X_2(l))_{l\in\{k\}\cup O}$ separately, we have

$$\begin{aligned}R(X_1(k_1), (X_1(l))_{l\in\{k\}\cup O}) = R(X_1(k_1), (X_1(l))_{l\in O}),\\ \text{and } R(X_2(k_2), (X_2(l))_{l\in\{k\}\cup O}) = R(X_2(k_2), (X_2(l))_{l\in O}),\end{aligned}$$

respectively. Since $R(X_1(k_1), (X_1(l))_{l\in O}) = R(X_2(k_2), (X_2(l))_{l\in O})$, then we have

$$R(X_1(k), (X_1(l))_{l\in\{k\}\cup O}) = R(X_2(k), (X_2(l))_{l\in\{k\}\cup O}),$$

which proves our result when the case (i) holds.

If, however, the case

$$\text{(ii)} : \alpha \leq R(X_1(k_1), (X_1(l))_{l\in O}) = R(X_2(k_2), (X_2(l))_{l\in O})$$

is true, we can apply Lemma C.2.1 and get

$$\begin{aligned}R(X_1(k_1), (X_1(l))_{l\in\{k\}\cup O}) &\geq R(X_1(k_1), (X_1(l))_{l\in O})\\ &\geq \alpha = R(X_1(k), (X_1(l))_{l\in\{k\}\cup O})\end{aligned}$$

and

$$\begin{aligned}R(X_2(k_2), (X_2(l))_{l\in\{k\}\cup O}) &\geq R(X_2(k_2), (X_2(l))_{l\in O})\\ &\geq \alpha = R(X_2(k), (X_2(l))_{l\in\{k\}\cup O}).\end{aligned}$$

Since X_1 and X_2 are both vectors of distinct real values and $k \neq k_1, k_2$, we have

$$\begin{aligned}&R(X_1(k_1), (X_1(l))_{l\in\{k\}\cup O}) \neq R(X_1(k), (X_1(l))_{l\in\{k\}\cup O})\\ \Longrightarrow\ &R(X_1(k_1), (X_1(l))_{l\in\{k\}\cup O}) > R(X_1(k), (X_1(l))_{l\in\{k\}\cup O})\\ \Longrightarrow\ &X_1(k_1) > X_1(k),\end{aligned}$$

and

$$\begin{aligned}&R(X_2(k_2), (X_2(l))_{l\in\{k\}\cup O}) \neq R(X_2(k), (X_2(l))_{l\in\{k\}\cup O})\\ \Longrightarrow\ &R(X_2(k_2), (X_2(l))_{l\in\{k\}\cup O}) > R(X_2(k), (X_1(l))_{l\in\{k\}\cup O})\\ \Longrightarrow\ &X_2(k_2) > X_2(k).\end{aligned}$$

Subsequently, according to the definition of rank and notice that $X_1, X_2 \in \mathcal{V}^n$ are vectors of

distinct real values, we can show that

$$\begin{aligned} R(X_1(k_1), (X_1(l))_{l\in\{k\}\cup O}) &= R(X_1(k_1), (X_1(l))_{l\in O}) + \mathrm{I}\left(X_1(k_1) > X_1(k)\right) \\ &= R(X_1(k_1), (X_1(l))_{l\in O}) + 1, \end{aligned}$$

and

$$\begin{aligned} R(X_2(k_2), (X_2(l))_{l\in\{k\}\cup O}) &= R(X_2(k_2), (X_2(l))_{l\in O}) + \mathrm{I}\left(X_2(k_2) > X_2(k)\right) \\ &= R(X_2(k_2), (X_2(l))_{l\in O}) + 1. \end{aligned}$$

Recall that $R(X_1(k_1), (X_1(l))_{l\in O}) = R(X_2(k_2), (X_2(l))_{l\in O})$. Hence, we have

$$R(X_1(k), (X_1(l))_{l\in\{k\}\cup O}) = R(X_2(k), (X_2(l))_{l\in\{k\}\cup O}).$$

This proves our results when the case (ii) is true, and completes our proof.

□

Lemma C.3.16. Suppose $X_1, X_2 \in \mathcal{V}^n$, and $O_1, O_2 \subset [n]$ are non-empty disjoint subset of indices such that $O_1, O_2 \neq \emptyset$ and $O_1 \cap O_2 = \emptyset$. Suppose

$$R(X_1(i), (X_1(l))_{l\in O_1\cup O_2}) = R(X_2(i), (X_2(l))_{l\in O_1\cup O_2}), \text{ for any } i \in O_2$$

Then, for any $k_1, k_2 \in O_1$, if

$$R(X_1(k_1), (X_1(l))_{l\in O_1}) = R(X_2(k_2), (X_2(l))_{l\in O_1}),$$

we have

$$R(X_1(k_1), (X_1(l))_{l\in O_1\cup O_2}) = R(X_2(k_2), (X_2(l))_{l\in O_1\cup O_2}).$$

Proof. For any given $n \in \mathbb{N}$, let $P(k)$ be the statement of Lemma C.3.16 when $|O_2| = k$. We prove $P(k)$ is true for any $k \in \{1, \ldots, n-1\}$ by induction on k.

Base Case: The case $P(1)$ is shown in Lemma C.3.15.

Induction Step: We show the implication $P(1), P(k) \implies P(k+1)$ for any $k \in \{1, \ldots, n-2\}$.

Let us now consider the case when $|O_2| = k+1$. Without loss of generality, let us assume (after relabelling) $O_2 = \{1, \ldots, k+1\}$. Then according to the definition of ranks and notice that $X_1, X_2 \in \mathcal{V}^n$ are vectors of distinct real values, for any $i \in O_2$, and $j = 1, 2$, we can show that

$$R(X_j(i), (X_j(l))_{l\in O_1\cup\left(O_2\setminus\{k\}\right)}) = R(X_j(i), (X_j(l))_{l\in O_1\cup O_2}) - \mathrm{I}\left(X_j(i) > X_j(k)\right).$$

Notice that for any $i \in O_2$, and $j = 1, 2$, we have

$$\mathrm{I}\left(X_j(i) > X_j(k)\right) = \mathrm{I}\left(R(X_j(i), (X_j(l))_{l\in O_1\cup O_2}) > R(X_j(k), (X_j(l))_{l\in O_1\cup O_2})\right).$$

Hence, for any $i \in O_2$, and $j = 1, 2$, we have

$$R(X_j(i), (X_j(l))_{l\in O_1\cup\left(O_2\setminus\{k\}\right)}) = R(X_j(i), (X_j(l))_{l\in O_1\cup O_2}) \\ - \mathrm{I}\left(R(X_j(i), (X_j(l))_{l\in O_1\cup O_2}) > R(X_j(k), (X_j(l))_{l\in O_1\cup O_2})\right).$$

Since

$$R(X_1(i), (X_1(l))_{l\in O_1\cup O_2}) = R(X_2(i), (X_2(l))_{l\in O_1\cup O_2}), \text{ for any } i \in O_2,$$

then for any $i \in O_2 \setminus \{k\}$, we have

$$R(X_1(i), (X_1(l))_{l\in O_1\cup\left(O_2\setminus\{k\}\right)}) = R(X_2(i), (X_2(l))_{l\in O_1\cup\left(O_2\setminus\{k\}\right)}).$$

Notice that $|\left(O_2 \setminus \{k\}\right)| = k$, and $R(X_1(k_1), (X_1(l))_{l\in O_1}) = R(X_2(k_2), (X_2(l))_{l\in O_1})$. Then, since $P(k)$ is true, we have

$$R(X_1(k_1), (X_1(l))_{l\in O_1\cup\left(O_2\setminus\{k\}\right)}) = R(X_2(k_2), (X_2(l))_{l\in O_1\cup\left(O_2\setminus\{k\}\right)}). \tag{C.39}$$

Next, since

$$R(X_1(i), (X_1(l))_{l\in O_1\cup O_2}) = R(X_2(i), (X_2(l))_{l\in O_1\cup O_2}), \text{ for any } i \in O_2,$$

we have

$$R(X_1(k), (X_1(l))_{l\in\left\{O_1\cup\left(O_2\setminus\{k\}\right)\right\}\cup\{k\}}) = R(X_2(k), (X_2(l))_{l\in\left\{O_1\cup\left(O_2\setminus\{k\}\right)\right\}\cup\{k\}}).$$

Notice that (C.39) is true, and $P(1)$ is true. Then, we have

$$R(X_1(k_1), (X_1(l))_{l\in O_1\cup O_2}) = R(X_2(k_2), (X_2(l))_{l\in O_1\cup O_2}).$$

Hence, we have shown $P(k+1)$ is true. Thus, we complete our proof.

□

Lemma C.3.17. Suppose $X_1, X_2 \in \mathcal{V}^n$ and $O \subset [n]$ is a non-empty subset of indices. Let $O_1, O_2 \subset [n]$ be non-empty subsets of indices such that $O_1 \cup O_2 = O$, $O_1 \cap O_2 \neq \emptyset$, and $O_1 \setminus O_2 \neq \emptyset$. Suppose

$$R(X_1(i), (X_1(l))_{l\in O_1}) = R(X_2(i), (X_2(l))_{l\in O_1}), \text{ for any } i \in O_1 \cap O_2, \tag{C.40}$$

and $\left(R(X_1(i), (X_1(l))_{l\in O})\right)_{i\in O_2}$ is a permutation of $\left(R(X_2(i), (X_2(l))_{l\in O})\right)_{i\in O_2}$. Then, for any $k \in O_1 \setminus O_2$, if

$$R(X_1(k), (X_1(l))_{l\in O_1}) = R(X_2(k), (X_2(l))_{l\in O_1}), \tag{C.41}$$

we have

$$R(X_1(k),(X_1(l))_{l\in O}) = R(X_2(k),(X_2(l))_{l\in O}).$$

Proof. To start, since $R(X_1(i),(X_1(l))_{l\in O_1}) = R(X_2(i),(X_2(l))_{l\in O_1})$ for any $i \in O_1 \cap O_2$, and $R(X_1(k),(X_1(l))_{l\in O_1}) = R(X_2(k),(X_2(l))_{l\in O_1})$, we have

$$\begin{aligned}&\sum_{i\in O_1\cap O_2} \mathrm{I}\left(R(X_1(i),(X_1(l))_{l\in O_1}) < R(X_1(k),(X_1(l))_{l\in O_1})\right)\\ &= \sum_{i\in O_1\cap O_2} \mathrm{I}\left(R(X_2(i),(X_2(l))_{l\in O_1}) < R(X_2(k),(X_2(l))_{l\in O_1})\right).\end{aligned}$$

Notice that for any $i \in O_1 \cap O_2$, we have

$$\begin{aligned}\mathrm{I}\left(R(X_1(i),(X_1(l))_{l\in O_1}) < R(X_1(k_1),(X_1(l))_{l\in O_1})\right) &= \mathrm{I}\left(X_1(i) < X_1(k_1)\right),\\ \text{and } \mathrm{I}\left(R(X_2(i),(X_2(l))_{l\in O_1}) < R(X_2(k_1),(X_2(l))_{l\in O_1})\right) &= \mathrm{I}\left(X_2(i) < X_2(k_1)\right).\end{aligned}$$

Thus, we have

$$\sum_{i\in O_1\cap O_2} \mathrm{I}\left(X_1(i) < X_1(k)\right) = \sum_{i\in O_1\cap O_2} \mathrm{I}\left(X_2(i) < X_2(k)\right).$$

Then, since $X_1, X_2 \in \mathcal{V}^n$ are both vectors of distinct real values and $k \notin O_1 \cap O_2$, we have

$$\sum_{i\in O_1\cap O_2} \mathrm{I}\left(X_1(i) \le X_1(k)\right) = \sum_{i\in O_1\cap O_2} \mathrm{I}\left(X_2(i) \le X_2(k)\right). \tag{C.42}$$

Subsequently, according to the definition of rank, we have

$$\begin{aligned}&R(X_1(k),(X_1(l))_{l\in O_1\setminus O_2})\\ &= \sum_{l\in O_1\setminus O_2} \mathrm{I}\left(X_1(l) \le X_1(k)\right)\\ &= \sum_{l\in O_1} \mathrm{I}\left(X_1(l) \le X_1(k)\right) - \sum_{l\in O_1\cap O_2} \mathrm{I}\left(X_1(l) \le X_1(k)\right)\\ &= R(X_1(k),(X_1(l))_{l\in O_1}) - \sum_{l\in O_1\cap O_2} \mathrm{I}\left(X_1(l) \le X_1(k)\right).\end{aligned}$$

According to (C.41) and (C.42), we further have

$$\begin{aligned}
&R(X_1(k),(X_1(l))_{l\in O_1\setminus O_2})\\
&=^{(\text{C.41})} R(X_2(k),(X_2(l))_{l\in O_1}) - \sum_{l\in O_1\cap O_2} \mathrm{I}\left(X_1(l)\le X_1(k)\right)\\
&=^{(\text{C.42})} R(X_2(k),(X_2(l))_{l\in O_1}) - \sum_{l\in O_1\cap O_2} \mathrm{I}\left(X_2(l)\le X_2(k)\right)\\
&= \sum_{l\in O_1} \mathrm{I}\left(X_2(l)\le X_2(k)\right) - \sum_{l\in O_1\cap O_2} \mathrm{I}\left(X_2(l)\le X_2(k)\right)\\
&= \sum_{l\in O_1\setminus O_2} \mathrm{I}\left(X_2(l)\le X_2(k)\right)\\
&= R(X_2(k),(X_2(l))_{l\in O_1\setminus O_2}).
\end{aligned} \tag{C.43}$$

As $\left(R(X_1(i),(X_1(l))_{l\in O})\right)_{i\in O_2}$ is a permutation of $\left(R(X_2(i),(X_2(l))_{l\in O})\right)_{i\in O_2}$, then there exist a permutation $(\sigma(i))_{i\in O_2}$ of O_2 such that

$$R(X_1(\sigma(i)),(X_1(l))_{l\in O}) = R(X_2(i),(X_2(l))_{l\in O}),\ \text{for any } i\in O_2. \tag{C.44}$$

Let $X_3\in\mathcal{V}^n$ be a permutation of X_1 such that

$$\begin{aligned}
X_3(i) &= X_1(i),\ \text{for any } i\in[n]\setminus O_2,\\
\text{and } X_3(i) &= X_1(\sigma(i)),\ \text{for any } i\in O_2.
\end{aligned}$$

Notice that $O_1\setminus O_2\subset[n]\setminus O_2$, hence we have $(X_1(l))_{l\in O_1\setminus O_2} = (X_3(l))_{l\in O_1\setminus O_2}$. Next, since $k\in O_1\setminus O_2$, we have $X_1(k)=X_3(k)$. Hence, we have

$$R(X_3(k),(X_3(l))_{l\in O_1\setminus O_2}) = R(X_1(k),(X_1(l))_{l\in O_1\setminus O_2}). \tag{C.45}$$

According to (C.43), we further have

$$R(X_3(k),(X_3(l))_{l\in O_1\setminus O_2}) = R(X_2(k),(X_2(l))_{l\in O_1\setminus O_2}). \tag{C.46}$$

Next, notice that $O = O_2\cup(O\setminus O_2)$, where $O\setminus O_2\subset[n]\setminus O_2$. Then, since $X_3(i) = X_1(\sigma(i))$, for any $i\in O_2$, and $X_3(i)=X_1(i)$, for any $i\in[n]\setminus O_2$, we have $(X_1(l))_{l\in O}$ is a permutation of $(X_3(l))_{l\in O}$. Hence, for any $i\in O_2$, we have

$$R(X_3(i),(X_3(l))_{l\in O}) = R(X_1(\sigma(i)),(X_1(l))_{l\in O}).$$

According to (C.41), we further have

$$R(X_3(i),(X_1(l))_{l\in O}) = R(X_2(i),(X_2(l))_{l\in O}),\ \text{for any } i\in O_2. \tag{C.47}$$

Notice that (C.46) and (C.47) is true. Then, according to Lemma C.3.16, we have

$$R(X_3(k),(X_3(l))_{l\in O}) = R(X_2(k),(X_2(l))_{l\in O}).$$

Since $(X_1(l))_{l\in O}$ is a permutation of $(X_3(l))_{l\in O}$, and $X_1(k) = X_3(k)$, we have

$$R(X_3(k), (X_3(l))_{l\in O}) = R(X_1(k), (X_1(l))_{l\in O}).$$

Thus,

$$R(X_1(k), (X_1(l))_{l\in O}) =^{(\text{C.45})} R(X_2(k), (X_2(l))_{l\in O}).$$

This completes our proof.

□

Lemma C.3.18. Suppose $X, Y \in \mathcal{V}^n$, and let $V = \{1, \cdots, t\}$, $U = \{t+1, \cdots, t+m\}$ be subsets of indices, where $t, m \geq 1$ and $t+m \leq n-1$. Suppose $X(t) > \ldots > X(1)$, $Y(t+m) > \ldots > Y(t+1)$ and $R(Y(t+1), (Y(l))_{l\in[n]\setminus V}) > R(X(1), (X(l))_{l\in[n]\setminus U})$ Then, if $R(X(i), X) = R(Y(i), Y)$ for any $i \in U \cup V$, we have $R(X(1), (X(l))_{l\in[n]\setminus U}) = R(X(1), X) = R(Y(1), Y)$.

Proof. To start, according to Lemma C.2.1, for any $i \in U \cup V$, we have

$$R(X(i), X) = R(Y(i), Y) \geq R(Y(i), (Y(l))_{l\in[n]\setminus V}).$$

Since it is assumed that $Y(t+m) > \ldots > Y(t+1)$, then for any $i \in U = \{t+1, \ldots, t+m\}$, we have

$$R(Y(t+m), (Y(l))_{l\in[n]\setminus V}) > \ldots > R(Y(t+1), (Y(l))_{l\in[n]\setminus V}).$$

Hence, for any $i \in U$, we have

$$R(X(i), X) = R(Y(i), Y) \geq R(Y(i), (Y(l))_{l\in[n]\setminus V}) \geq R(Y(t+1), (Y(l))_{l\in[n]\setminus V}).$$

Since

$$R(Y(t+1), (Y(l))_{l\in[n]\setminus V}) > R(X(1), (X(l))_{l\in[n]\setminus U}),$$

then for any $i \in U$, we have

$$R(X(i), X) > R(X(1), (X(l))_{l\in[n]\setminus U}).$$

According to Lemma C.2.2, we further have

$$R(X(1), (X(l))_{l\in[n]\setminus U}) = R(X(1), X).$$

Since for any $i \in U \cup V$, $R(X(i), X) = R(Y(i), Y)$, we have

$$R(X(1), (X(l))_{l\in[n]\setminus U}) = R(X(1), X) = R(Y(1), Y).$$

This completes our proof. □

Lemma C.3.19. Suppose $X, Y \in \mathcal{V}^n$, and let $V = \{1, \cdots, t\}$, $U = \{t+1, \cdots, t+m\}$ be subsets of indices, where $t, m \geq 1$ and $t+m \leq n-1$. Suppose $X(t) > \ldots > X(1)$, $Y(t+m) > \ldots > Y(t+1)$ and $R(Y(t+1), (Y(l))_{l\in[n]\setminus V}) = R(X(1), (X(l))_{l\in[n]\setminus U})$. Denote

$$\alpha = R(Y(t+1), (Y(l))_{l\in[n]\setminus V}) = R(X(1), (X(l))_{l\in[n]\setminus U}). \tag{C.48}$$

Define a and b as

$$\begin{aligned} a &= \max\{i \in \{0, \cdots, t-1\} \mid R(X(1+i), (X(l))_{l\in[n]\setminus U}) = \alpha + i\}, \\ \text{and } b &= \max\{i \in \{0, \cdots, m-1\} \mid R(Y(t+1+i), (Y(l))_{l\in[n]\setminus V}) = \alpha + i\}, \end{aligned}$$

respectively. Then, if $R(X(i), X) = R(Y(i), Y)$, for any $i \in U \cup V$, we have

$$\alpha + i \leq R(X(1+i), X) \leq \alpha + b + i + 1, \quad i \in \{0, \cdots, a\}, \tag{C.49}$$

$$\text{and } \alpha + i \leq R(Y(t+1+i), Y) \leq \alpha + a + i + 1, \quad i \in \{0, \cdots, b\}. \tag{C.50}$$

Proof. We only prove inequality (C.50) is true, since inequality (C.49) can be proved similarly, or be proved in the same way after switching the labels between X and Y, and relabelling the relevant components of the data.

According to the definitions of a and b, we have

$$R(X(1+i), (X(l))_{l\in[n]\setminus U}) = \alpha + i, \text{ for any } i \in \{0, \cdots, a\}, \tag{C.51}$$

$$\text{and } R(Y(t+1+i), (Y(l))_{l\in[n]\setminus V}) = \alpha + i, \text{ for any } i \in \{0, \cdots, b\}. \tag{C.52}$$

Subsequently, according to Lemma C.2.1, we have

$$R(Y(t+1+i), Y) \geq R(Y(t+1+i), (Y(l))_{l\in[n]\setminus V}) = \alpha + i, \text{ for any } i \in \{0, \cdots, b\},$$

which proves the left-hand side of (C.50).

In the following, we prove the right-hand side of (C.50) is true.

To start, let us consider the case when $a = t - 1$. According to Lemma C.2.1, for any $i \in \{0, \cdots, b\}$, we have

$$R(Y(t+1+i), Y) \leq t + R(Y(t+1+i), (Y(l))_{l\in[n]\setminus V}).$$

Then, we have

$$R(Y(t+1+i), Y) \leq^{(C.52)} t + \alpha + i = \alpha + i + a + 1.$$

Hence, we have shown that the right-hand side of (C.50) is true when $a = t - 1$.

Now we consider the cases when $a < t - 1$. For any fixed non-negative integer $b \geq 0$, let $P(k)$ be the statement that

$$P(k) : R(Y(t+1+i), Y) \leq \alpha + a + i + 1 \text{ for any } i \leq k.$$

We prove $P(k)$ holds for any $k \in \{0, \ldots, b\}$ by induction on k.

Base Case: We prove $P(0)$ holds. Since $a < t-1$, it follows $a+1 < t$ and $a+2 \leq t$. According to the definition of a, we have

$$R(X(a+1), (X(l))_{l\in[n]\setminus U}) = \alpha + a,$$

and

$$R(X(a+2), (X(l))_{l\in[n]\setminus U}) \neq \alpha + a + 1.$$

Since $a+2 \leq t$ and $X(t) > \ldots > X(1)$, we have

$$R(X(a+2), (X(l))_{l\in[n]\setminus U}) > \alpha + a.$$

Thus, we have

$$R(X(a+2), (X(l))_{l\in[n]\setminus U}) > \alpha + a + 1. \tag{C.53}$$

Since $a+2 \leq t$, we have $\{a+2, \cdots, t\} \neq \emptyset$, and $\{a+2, \cdots, t\} \subset V$. Then, since $R(X(i), X) = R(Y(i), Y)$, for any $i \in U \cup V$, we have

$$R(X(i), X) = R(Y(i), Y), \text{ for any } i \in \{a+2, \cdots, t\}.$$

According to Lemma C.2.1, for any $i \in \{a+2, \cdots, t\}$, we have

$$R(X(i), X) \geq R(X(i), (X(l))_{l\in[n]\setminus U}).$$

Hence, for any $i \in \{a+2, \cdots, t\}$, it follows

$$R(Y(i), Y) = R(X(i), X) \geq R(X(i), (X(l))_{l\in[n]\setminus U}).$$

Since $X(t) > \ldots > X(1)$, we further have for any $i \in \{a+2, \cdots, t\}$,

$$R(X(i), (X(l))_{l\in[n]\setminus U}) \geq R(X(a+2), (X(l))_{l\in[n]\setminus U}).$$

Therefore, for any $i \in \{a+2, \cdots, t\}$, we have

$$\begin{aligned} R(Y(i), Y) \geq R(X(i), (X(l))_{l\in[n]\setminus U}) &\geq R(X(a+2), (X(l))_{l\in[n]\setminus U}) \\ &>^{(C.53)} \alpha + a + 1. \end{aligned}$$

According to (C.49), we have

$$\alpha + a + 1 = R(Y(t+1), (Y(l))_{l\in[n]\setminus V}) + a + 1.$$

According to Lemma C.2.1, we have

$$R(Y(t+1),(Y(l))_{l\in[n]\setminus V})+a+1\geq R(Y(t+1),(Y(l))_{l\in\{1,\ldots,a+1\}\cup\{t+1,\ldots,n\}}).$$

Hence, for any $i\in\{a+2,\cdots,t\}$,

$$R(Y(i),Y)>\alpha+a+1\geq R(Y(t+1),(Y(l))_{l\in\{1,\ldots,a+1\}\cup\{t+1,\ldots,n\}}).$$

Subsequently, according to Lemma C.2.2, we have

$$R(Y(t+1),Y)=R(Y(t+1),(Y(l))_{l\in\{1,\ldots,a+1\}\cup\{t+1,\ldots,n\}})\leq\alpha+a+1,$$

which proves $P(0)$ holds.

Induction Step: We show the implication $P(k)\implies P(k+1)$ for any $k\in\{0,\ldots,b-1\}$.

Since $P(k)$ is true, we have

$$R(Y(t+1+i),Y)\leq\alpha+i+a+1,\quad i\in\{0,\cdots,k\}.$$

Notice that $0\leq k\leq b-1\leq m-2$. We have $\{t+1,\ldots,t+k+1\}\subset\{t+1,\ldots,t+m-1\}\subset U$. Since $R(X(i),X)=R(Y(i),Y)$ for any $i\in U\cup V$, we have for any $i\in\{0,\cdots,k\}$,

$$R(X(t+1+i),X)=R(Y(t+1+i),Y).$$

Hence, for any $i\in\{0,\cdots,k\}$, we have

$$\begin{aligned}
&\alpha+i+a+1\\
&\geq R(Y(t+1+i),Y)\\
&=R(X(t+1+i),X)\\
&=\sum_{l\in[n]\setminus U}\mathrm{I}\left(X(l)\leq X(t+1+i)\right)+\sum_{l\in U}\mathrm{I}\left(X(l)\leq X(t+1+i)\right)\\
&=R(X(t+1+i),(X(l))_{l\in[n]\setminus U})+\sum_{l\in\{t+1,\cdots,t+m\}}\mathrm{I}\left(X(l)\leq X(t+1+i)\right).
\end{aligned}\tag{C.54}$$

Since $Y(t+1)<\cdots<Y(t+m)$, we have

$$R(Y(t+1),Y)<\cdots<R(Y(t+m),Y).$$

Then, since $R(X(i),X)=R(Y(i),Y)$ for any $i\in U\cup V$, we have

$$\begin{aligned}
&R(X(t+1),X)<\cdots<R(X(t+m),X)\\
\implies&X(t+1)<\cdots<X(t+m)\\
\implies&\sum_{l\in\{t+1,\cdots,t+m\}}\mathrm{I}\left(X(l)\leq X(t+1+i)\right)=i+1,\quad i\in\{0,\cdots,m-1\}.
\end{aligned}$$

Since $k \leq b-1 \leq m-2$, for any $i \in \{0,1,\cdots,k\}$, we have

$$\sum_{l\in\{t+1,\cdots,t+m\}} \mathrm{I}\left(X(l) \leq X(t+1+i)\right) = i+1.$$

Thus, for any $i \in \{0,\cdots,k\}$, we have

$$\begin{aligned}
&\alpha+i+a+1 \geq^{(\mathrm{C}.54)} R(X(t+1+i),(X(l))_{l\in[n]\setminus U}) + i + 1 \\
\Longrightarrow\ & R(X(t+1+i),(X(l))_{l\in[n]\setminus U}) \leq \alpha + a, \\
\Longrightarrow\ & R(X(t+1+i),(X(l))_{l\in[n]\setminus U}) < \alpha+a+1 <^{(\mathrm{C}.53)} R(X(a+2),(X(l))_{l\in[n]\setminus U}).
\end{aligned}$$

Hence, we have $X(t+1+i) < X(a+2)$ for any $i \in \{0,\cdots,k\}$, which implies that

$$\sum_{i\in\{t+1,\cdots,t+1+k\}} \mathrm{I}\left(X(i) \leq X(a+2)\right) = k+1.$$

Since $U = \{t+1,\cdots,t+m\}$ and $k < m$, we have

$$\begin{aligned}
&R(X(a+2),(X(l))_{l\in\{t+1,\cdots,t+1+k\}\cup([n]\setminus U)}) \\
&= \sum_{i\in\{t+1,\cdots,t+1+k\}} \mathrm{I}\left(X(i) \leq X(a+2)\right) + \sum_{i\in[n]\setminus U} \mathrm{I}\left(X(i) \leq X(a+2)\right) \\
&= k+1+R(X(a+2),(X(l))_{l\in[n]\setminus U}) \\
&>^{(\mathrm{C}.53)} \alpha+a+2+k.
\end{aligned} \tag{C.55}$$

Since $X(1) < \cdots < X(t)$, we have

$$R(X(1),X) < \cdots < R(X(t),X).$$

Then, since $R(X(i),X) = R(Y(i),Y)$ for any $i \in U \cup V$, we have

$$R(Y(1),Y) < \cdots < R(Y(t),Y) \implies Y(1) < \cdots Y(t).$$

Then, for any $i \in \{a+2,\cdots,t\}$, we have

$$R(Y(i),Y) \geq R(Y(a+2),Y).$$

Since $R(X(i),X) = R(Y(i),Y)$ for any $i \in U \cup V$, we have

$$R(Y(i),Y) \geq R(Y(a+2),Y) = R(X(a+2),X).$$

According to Lemma C.2.1, we have

$$\begin{aligned}
R(X(a+2),X) &\geq R(X(a+2),(X(l))_{l\in\{t+1,\cdots,t+1+k\}\cup([n]\setminus U)}) \\
&>^{(\mathrm{C}.55)} \alpha+a+2+k \\
&=^{(\mathrm{C}.52)} R(Y(t+k+2),(Y(l))_{l\in[n]\setminus V}) + a + 1.
\end{aligned}$$

Applying Lemma C.2.1 again, we have

$$\begin{aligned}&R(Y(t+k+2),(Y(l))_{l\in[n]\setminus V})+a+1\\&\geq R(Y(t+k+2),(Y(l))_{l\in\{1,\cdots,a+1\}\cup\{t+1,\cdots,n\}}).\end{aligned}$$

Hence, for any $i\in\{a+2,\cdots,t\}$, it follows that

$$R(Y(i),Y)\geq R(Y(a+2),Y)>R(Y(t+k+2),(Y(l))_{l\in\{1,\cdots,a+1\}\cup\{t+1,\cdots,n\}}).$$

Thus, according to Lemma C.2.2, we have

$$\begin{aligned}R(Y(t+k+2),Y)&=R(Y(t+k+2),(Y(l))_{l\in\{1,\cdots,a+1\}\cup\{t+1,\cdots,n\}})\\&\leq\alpha+a+k+2.\end{aligned}$$

That is,

$$R(Y(t+1+k+1),Y)\leq\alpha+k+1+a+1,$$

which proves the right-hand side of (C.50) when $i=k+1$. Hence, we have shown $P(k+1)$ is true. Overall, we have shown that the right-hand side of (C.50) is true. This completes our proof.

□

Lemma C.3.20. Following Lemma C.3.19, let $\mathbb{S},\mathbb{A}_1,\mathbb{A}_2\subset[n]$ be subsets of indices such that

$$\begin{aligned}\mathbb{S}&=\{1,\cdots,a+1\}\cup\{t+1,\cdots,t+b+1\},\\\mathbb{A}_1&=V\setminus\{1,\cdots,a+1\}=\{1,\cdots,t\}\setminus\{1,\cdots,a+1\},\\\text{and }\mathbb{A}_2&=U\setminus\{t+1,\cdots,t+b+1\}=\{t+1,\cdots,t+m\}\setminus\{t+1,\cdots,t+b+1\},\end{aligned}$$

respectively. Then, $\big(R(X(i),(X(l))_{l\in[n]\setminus\mathbb{A}_2})\big)_{i\in\mathbb{S}}$ and $\big(R(Y(i),(Y(l))_{l\in[n]\setminus\mathbb{A}_1})\big)_{i\in\mathbb{S}}$ are both permutations of $\{\alpha,\cdots,\alpha+a+b+1\}$. Further, for any $i\in\mathbb{S}$, we have

$$R(X(i),(X(l))_{l\in[n]\setminus\mathbb{A}_2})=R(Y(i),(Y(l))_{l\in[n]\setminus\mathbb{A}_1}).$$

Proof. According to Lemma C.3.19, we have

$$\begin{aligned}&\alpha+i\leq R(X(1+i),X)\leq\alpha+b+i+1,\text{ for any }i\in\{0,\cdots,a\}\\\Longrightarrow\;&\alpha\leq R(X(1+i),X)\leq\alpha+b+a+1,\text{ for any }i\in\{0,\cdots,a\}\\\Longrightarrow\;&\alpha\leq R(X(i),X)\leq\alpha+b+a+1,\text{ for any }i\in\{1,\cdots,a+1\},\end{aligned}\tag{C.56}$$

and

$$\begin{aligned}&\alpha+i\leq R(Y(t+1+i),Y)\leq\alpha+a+i+1,\text{ for any }i\in\{0,1,2,\cdots,b\}\\\Longrightarrow\;&\alpha\leq R(Y(t+1+i),Y)\leq\alpha+a+b+1,\text{ for any }i\in\{0,1,2,\cdots,b\}\\\Longrightarrow\;&\alpha\leq R(Y(i),Y)\leq\alpha+a+b+1,\text{ for any }i\in\{t+1,\cdots,t+1+b\}.\end{aligned}\tag{C.57}$$

Then, according to the definition of b, we have $b \leq m-1$. Hence, we have $\{t+1, \cdots, t+1+b\} \subset U$. Since $R(X(i), X) = R(Y(i), Y)$ for any $i \in U \cup V$, then according to (C.57), we have

$$\alpha \leq R(X(i), X) \leq \alpha + a + b + 1 \text{ for any } i \in \{t+1, \cdots, t+1+b\}.$$

Combining this with (C.56), then for any $i \in \{1, \cdots, a+1\} \cup \{t+1, \cdots, t+1+b\}$, we have

$$\alpha \leq R(X(i), X) \leq \alpha + a + b + 1. \tag{C.58}$$

That is,

$$\alpha \leq R(X(i), X) \leq \alpha + a + b + 1, \text{ for any } i \in \mathbb{S}.$$

Notice that

$$\begin{aligned} |\mathbb{S}| &= |\{1, \cdots, a+1\} \cup \{t+1, \cdots, t+b+1\}| = a + b + 2, \\ \text{and } \{\alpha, \cdots, \alpha + a + b + 1\} &= a + b + 2. \end{aligned}$$

Since $X \in \mathcal{V}^n$ is a vector of distinct real values, $\big(R(X(i), X)\big)_{i \in \mathbb{S}}$ is a permutation of $a+b+2$ distinct integers. Hence, $\big(R(X(i), X)\big)_{i \in \mathbb{S}}$ is a permutation of $\{\alpha, \cdots, \alpha + a + b + 1\}$.

Since $Y(t+1) < \cdots < Y(t+m)$, we have

$$R(Y(t+1), Y) < \ldots < R(Y(t+m), Y).$$

Recall that $R(X(i), X) = R(Y(i), Y)$ for any $i \in U \cup V$. Thus,

$$\begin{aligned} & R(X(t+1), X) < \ldots < R(X(t+m), X) \\ \Longrightarrow\ & X(t+1) < \ldots < X(t+m). \end{aligned}$$

Since $\mathbb{A}_2 = U \setminus \{t+1, \cdots, t+b+1\} = \{t+1, \cdots, t+m\} \setminus \{t+1, \cdots, t+b+1\}$, then if $\mathbb{A}_2 \neq \emptyset$, we have for any $i \in \mathbb{A}_2$,

$$X(i) > X(t+b+1) \implies R(X(i), X) > R(X(t+b+1), X) \geq^{(C.58)} \alpha. \tag{C.59}$$

Since $\big(R(X(i), X)\big)_{i \in \mathbb{S}}$ is a permutation of $\{\alpha, \cdots, \alpha + a + b + 1\}$, we have

$$R(X(i), X) \notin \{\alpha, \alpha+1, \cdots, \alpha + a + b + 1\}, \text{ for any } i \in [n] \setminus \mathbb{S}.$$

Notice that $\mathbb{A}_2 = U \setminus \{t+1, \cdots, t+b+1\}$, and $[n] \setminus \mathbb{S} = [n] \setminus \big(\{1, \cdots, a+1\} \cup \{t+1, \cdots, t+b+1\}\big) = \big(V \setminus \{1, \cdots, a+1\}\big) \cup \big(U \setminus \{t+1, \cdots, t+b+1\}\big) \cup \big([n] \setminus (V \cup U)\big)$. Hence, $\mathbb{A}_2 \subset [n] \setminus \mathbb{S}$. Therefore, we have

$$R(X(i), X) \notin \{\alpha, \alpha+1, \cdots, \alpha + a + b + 1\}, \text{ for any } i \in \mathbb{A}_2. \tag{C.60}$$

According to (C.59), when $\mathbb{A}_2 \neq \emptyset$, we have $R(X(i), X) \geq \alpha$ for any $i \in \mathbb{A}_2$. Combining with

(C.60), we further have

$$R(X(i), X) > \alpha + a + b + 1, \text{ for any } i \in \mathbb{A}_2.$$

According to (C.58), we then have

$$\begin{aligned} &R(X(i), X) > \alpha + a + b + 1 \geq R(X(l), X), \text{ for any } i \in \mathbb{A}_2, l \in \mathbb{S} \\ \Longrightarrow\ &X(i) > X(l) \text{ for any } i \in \mathbb{A}_2, l \in \mathbb{S}. \end{aligned}$$

Subsequently, according to Lemma C.2.2, for any $i \in \mathbb{S}$,

$$R(X(i), (X(l))_{l \in [n] \setminus \mathbb{A}_2}) = R(X(i), X).$$

Notice that the above equation still holds when $\mathbb{A}_2 = \emptyset$. Then, since $\big(R(X(i), X)\big)_{i \in \mathbb{S}}$ is a permutation of $\{\alpha, \cdots, \alpha + a + b + 1\}$, $\big(R(X(i), (X(l))_{l \in [n] \setminus \mathbb{A}_2})\big)_{i \in \mathbb{S}}$ is also a permutation of $\{\alpha, \cdots, \alpha + a + b + 1\}$.

Similarly, we can show that $\big(R(Y(i), Y)\big)_{i \in \mathbb{S}}$ is a permutation of $\{\alpha, \cdots, \alpha + a + b + 1\}$ and

$$R(Y(i), (Y(l))_{l \in [n] \setminus \mathbb{A}_1}) = R(Y(i), Y) \text{ for any } i \in \mathbb{S}.$$

Since $\big(R(Y_i, Y)\big)_{i \in \mathbb{S}}$ is a permutation of $\{\alpha, \cdots, \alpha + a + b + 1\}$, $\big(R(Y_i, (Y(l))_{l \in [n] \setminus \mathbb{A}_1})\big)_{i \in \mathbb{S}}$ is also a permutation of $\{\alpha, \cdots, \alpha + a + b + 1\}$. Further, since for any $i \in \mathbb{S} \subset U \cup V$, we have $R(X_i, X) = R(Y_i, Y)$, it follows that for any $i \in \mathbb{S}$, we have

$$R(X_i, (X(l))_{l \in [n] \setminus \mathbb{A}_2}) = R(Y_i, (Y(l))_{l \in [n] \setminus \mathbb{A}_1}).$$

This completes our proof. □

Lemma C.3.21. Suppose $n \geq 2$ is a positive integer. Let $V = \{1, \ldots, t\}, U = \{t+1, \ldots, t+m\} \subset [n]$ be subsets of $[n]$ such that $t, m \geq 1$ and $t + m \leq n$. Consider two integers a and b such that $t - 1 \geq a \geq 0$, and $m - 1 \geq b \geq 0$, and define

$$\begin{aligned} &\mathbb{S} = \{1, \ldots, a + 1\} \cup \{t + 1, \ldots, t + b + 1\}, \\ &\mathbb{A}_1 = V \setminus \{1, \ldots, a + 1\} = \{1, \ldots, t\} \setminus \{1, \ldots, a + 1\}, \\ &\mathbb{A}_2 = U \setminus \{t + 1, \ldots, t + b + 1\} = \{t + 1, \ldots, t + m\} \setminus \{t + 1, \ldots, t + b + 1\}. \end{aligned}$$

Denote $O = [n] \setminus (U \cup V) = [n] \setminus \{1, \ldots, t + m\}$. Then, we have

$$\begin{aligned} &[n] \setminus \mathbb{A}_2 = \mathbb{S} \cup \mathbb{A}_1 \cup O,\ \mathbb{S}, [n] \setminus U \subset [n] \setminus \mathbb{A}_2,\ \mathbb{S} \cap ([n] \setminus U) \neq \emptyset,\ \mathbb{S} \cup ([n] \setminus U) = [n] \setminus \mathbb{A}_2, \\ \text{and } &[n] \setminus \mathbb{A}_1 = \mathbb{S} \cup \mathbb{A}_2 \cup O,\ \mathbb{S}, [n] \setminus V \subset [n] \setminus \mathbb{A}_1,\ \mathbb{S} \cap ([n] \setminus V) \neq \emptyset,\ \mathbb{S} \cup ([n] \setminus V) = [n] \setminus \mathbb{A}_1. \end{aligned}$$

Proof. We first prove that $[n] \setminus \mathbb{A}_2 = \mathbb{S} \cup \mathbb{A}_1 \cup O$, Suppose $\mathbb{A}_2 \neq \emptyset$. Then, we have $\mathbb{A}_2 =$

$\{t+1,\ldots,t+m\} \setminus \{t+1,\ldots,t+1+b\} = \{t+2+b,\ldots,t+m\}$. Hence,

$$\begin{aligned}[n] \setminus \mathbb{A}_2 &= \big(\{1,\ldots,t\} \cup \{t+1,\ldots,t+m\} \cup O\big) \setminus \{t+2+b,\ldots,t+m\} \\ &= \{1,\ldots,t\} \cup \big(\{t+1,\ldots,t+m\} \setminus \{t+b+2,\ldots,t+m\}\big) \cup O \\ &= \{1,\ldots,t\} \cup \{t+1,\ldots,t+b+1\} \cup O.\end{aligned}$$

Since $a \leq t-1$, we have $\{1,\ldots,a+1\} \subset \{1,\ldots,t\}$. Then,

$$\begin{aligned}[n] \setminus \mathbb{A}_2 &= \{1,\ldots,a+1\} \cup \big(\{1,\ldots,t\} \setminus \{1,\ldots,a+1\}\big) \cup \{t+1,\ldots,t+b+1\} \cup O \\ &= \{1,\ldots,a+1\} \cup \mathbb{A}_1 \cup \{t+1,\ldots,t+b+1\} \cup O \\ &= \{1,\ldots,a+1\} \cup \{t+1,\ldots,t+b+1\} \cup \mathbb{A}_1 \cup O \\ &= \mathbb{S} \cup \mathbb{A}_1 \cup O,\end{aligned}$$

which proves $[n] \setminus \mathbb{A}_2 = \mathbb{S} \cup \mathbb{A}_1 \cup O$ when $\mathbb{A}_2 \neq \emptyset$.

However, when $\mathbb{A}_2 = \emptyset$, we have

$$\{t+1,\ldots,t+b+1\} = U \implies t+b+1 = t+m \implies b = m-1.$$

Hence, $\mathbb{S} = \{1,\ldots,a+1\} \cup \{t+1,\ldots,t+m\}$. Then, we have

$$\begin{aligned}\mathbb{S} \cup \mathbb{A}_1 \cup O &= \{1,\ldots,a+1\} \cup \{t+1,\ldots,t+m\} \cup \mathbb{A}_1 \cup O \\ &= \{1,\ldots,a+1\} \cup \mathbb{A}_1 \cup \{t+1,\ldots,n\} \\ &= \{1,\ldots,a+1\} \cup \big(\{1,\ldots,t\} \setminus \{1,\ldots,a+1\}\big) \cup \{t+1,\ldots,n\} \\ &= [n] = [n] \setminus \mathbb{A}_2,\end{aligned}$$

which proves $[n] \setminus \mathbb{A}_2 = \mathbb{S} \cup \mathbb{A}_1 \cup O$ when $\mathbb{A}_2 = \emptyset$. Therefore, we have shown $[n] \setminus \mathbb{A}_2 = \mathbb{S} \cup \mathbb{A}_1 \cup \{t+m+1,\ldots,n\}$.

Next, we show $\mathbb{S}, [n] \setminus U \subset [n] \setminus \mathbb{A}_2$. Suppose $\mathbb{A}_2 \neq \emptyset$, we have $\mathbb{A}_2 = \{t+1,\ldots,t+m\} \setminus \{t+1,\ldots,t+1+b\} = \{t+2+b,\ldots,t+m\}$. Then

$$\begin{aligned}[n] \setminus \mathbb{A}_2 &= [n] \setminus \{t+2+b,\ldots,t+m\} \\ &= \{1,\ldots,t\} \cup \{t+1,\ldots,t+1+b\} \cup O.\end{aligned}$$

Since $\mathbb{S} = \{1,\ldots,a+1\} \cup \{t+1,\ldots,t+1+b\}$, we have $\mathbb{S} \subset [n] \setminus \mathbb{A}_2$. However, suppose $\mathbb{A}_2 = \emptyset$. Then we have $[n] \setminus \mathbb{A}_2 = [n]$ and $\mathbb{S} \subset [n] \setminus \mathbb{A}_2$. Therefore, we have shown $\mathbb{S} \subset [n] \setminus \mathbb{A}_2$. Meanwhile, according to the definition of $\mathbb{A}_2$, we have $\mathbb{A}_2 \subset U$, which gives $[n] \setminus U \subset [n] \setminus \mathbb{A}_2$. Hence, we have shown $[n] \setminus U \subset [n] \setminus \mathbb{A}_2$.

Next, we show $\mathbb{S} \cap ([n] \setminus U) \neq \emptyset$. Since $U = \{t+1,\ldots,t+m\}$, we have $[n] \setminus U = (U \cup V \cup O) \setminus U = V \cup O = \{1,\ldots,t\} \cup O$. Notice that $\mathbb{S} = \{1,\ldots,a+1\} \cup \{t+1,\ldots,t+b+1\}$, where $t-1 \geq a \geq 0$, and $m-1 \geq b \geq 0$. Since $\{1,\ldots,t\} \cap \{1,\ldots,a+1\} = \{1,\ldots,a+1\} \neq \emptyset$, we have $\mathbb{S} \cap ([n] \setminus U) \neq \emptyset$.

Next, we show $\mathbb{S} \cup ([n] \setminus U) = [n] \setminus \mathbb{A}_2$. Suppose $\mathbb{A}_2 \neq \emptyset$. Then, we have $\mathbb{A}_2 = \{t+1,\ldots,t+$

$m\} \setminus \{t+1, \ldots, t+1+b\} = \{t+2+b, \ldots, t+m\}$. Hence,

$$\begin{aligned}
[n] \setminus \mathbb{A}_2 &= \big(\{1, \ldots, t\} \cup \{t+1, \ldots, t+m\} \cup O\big) \setminus \{t+2+b, \ldots, t+m\} \\
&= \{1, \ldots, t\} \cup \big(\{t+1, \ldots, t+m\} \setminus \{t+b+2, \ldots, t+m\}\big) \cup O \\
&= \{1, \ldots, t\} \cup \{t+1, \ldots, t+b+1\} \cup O.
\end{aligned}$$

Since $t-1 \geq a \geq 1$, we have $\{1, \ldots, t\} = \{1, \ldots, t\} \cup \{1, \ldots, a+1\}$. Then, we have

$$\begin{aligned}
[n] \setminus \mathbb{A}_2 &= \{1, \ldots, t\} \cup \{1, \ldots, a+1\} \cup \{t+1, \ldots, t+b+1\} \cup O \\
&= \{1, \ldots, a+1\} \cup \{t+1, \ldots, t+b+1\} \cup \{1, \ldots, t\} \cup O \\
&= \mathbb{S} \cup (\{1, \ldots, t\} \cup O) = \mathbb{S} \cup ([n] \setminus U).
\end{aligned}$$

However, when $\mathbb{A}_2 = \emptyset$, we have

$$\{t+1, \ldots, t+b+1\} = U \implies t+b+1 = t+m \implies b = m-1.$$

Then, we have $\mathbb{S} = \{1, \ldots, a+1\} \cup \{t+1, \ldots, t+m\}$. Hence, we have

$$\begin{aligned}
\mathbb{S} \cup ([n] \setminus U) &= \{1, \ldots, a+1\} \cup \{t+1, \ldots, t+m\} \cup ([n] \setminus \{t+1, \ldots, t+m\}) \\
&= [n] = [n] \setminus \mathbb{A}_2.
\end{aligned}$$

Therefore, we have shown $\mathbb{S} \cup ([n] \setminus U) = [n] \setminus \mathbb{A}_2$.

Now, we show $[n] \setminus \mathbb{A}_1 = \mathbb{S} \cup \mathbb{A}_2 \cup O$. Suppose $\mathbb{A}_1 \neq \emptyset$. Then, we have $\mathbb{A}_1 = \{1, \ldots, t\} \setminus \{1, \ldots, a+1\} = \{a+2, \ldots, t\}$. Subsequently, it follows that

$$\begin{aligned}
[n] \setminus \mathbb{A}_1 &= \big(\{1, \ldots, t\} \cup \{t+1, \ldots, t+m\} \cup O\big) \setminus \{a+2, \ldots, t\} \\
&= \big(\{1, \ldots, t\} \setminus \{a+2, \ldots, t\}\big) \cup \{t+1, \ldots, t+m\} \cup O \\
&= \{1, \ldots, a+1\} \cup \{t+1, \ldots, t+m\} \cup O.
\end{aligned}$$

Since $b \leq m-1$, we have $\{t+1, \ldots, t+b+1\} \subset \{t+1, \ldots, t+m\}$. Then, we have

$$\begin{aligned}
&[n] \setminus \mathbb{A}_1 \\
&= \{1, \ldots, a+1\} \cup (\{t+1, \ldots, t+m\} \setminus \{t+1, \ldots, t+b+1\}) \cup \{t+1, \ldots, t+b+1\} \cup O \\
&= \{1, \ldots, a+1\} \cup \mathbb{A}_2 \cup \{t+1, \ldots, t+b+1\} \cup O \\
&= \{1, \ldots, a+1\} \cup \{t+1, \ldots, t+b+1\} \cup \mathbb{A}_2 \cup O \\
&= \mathbb{S} \cup \mathbb{A}_2 \cup O,
\end{aligned}$$

which proves $[n] \setminus \mathbb{A}_1 = \mathbb{S} \cup \mathbb{A}_2 \cup \{t+m+1, \ldots, n\}$ when $\mathbb{A}_1 \neq \emptyset$.

However, when $\mathbb{A}_1 = \emptyset$, we have

$$\{1, \ldots, a+1\} = V \implies a+1 = t \implies a = t-1.$$

Hence, $\mathbb{S} = \{1, \ldots, t\} \cup \{t+1, \ldots, t+b+1\}$. Then,

$$\begin{aligned}
&\mathbb{S} \cup \mathbb{A}_2 \cup O \\
&= \{1, \ldots, t\} \cup \{t+1, \ldots, t+b+1\} \cup \mathbb{A}_2 \cup O \\
&= \{1, \ldots, t\} \cup \{t+1, \ldots, t+b+1\} \cup \big(\{t+1, \ldots, t+m\} \setminus \{t+1, \ldots, t+b+1\}\big) \cup O \\
&= \{1, \ldots, t\} \cup \{t+1, \ldots, t+m\} \cup O \\
&= [n] = [n] \setminus \mathbb{A}_1,
\end{aligned}$$

which proves $[n] \setminus \mathbb{A}_1 = \mathbb{S} \cup \mathbb{A}_2 \cup \{t+m+1, \ldots, n\}$ when $\mathbb{A}_1 = \emptyset$. Therefore, we have shown $[n] \setminus \mathbb{A}_1 = \mathbb{S} \cup \mathbb{A}_2 \cup \{t+m+1, \ldots, n\}$.

Next, we show $\mathbb{S}, [n] \setminus V \subset [n] \setminus \mathbb{A}_1$. Suppose $\mathbb{A}_1 \neq \emptyset$, we have $\mathbb{A}_1 = \{1, \ldots, t\} \setminus \{1, \ldots, a+1\} = \{a+2, \ldots, t\}$. Then

$$\begin{aligned}
[n] \setminus \mathbb{A}_1 &= [n] \setminus \{a+2, \ldots, t\} \\
&= \{1, \ldots, a+1\} \cup \{t+1, \ldots, n\}.
\end{aligned}$$

Since $\mathbb{S} = \{1, \ldots, a+1\} \cup \{t+1, \ldots, t+b+1\}$, we have $\mathbb{S} \subset [n] \setminus \mathbb{A}_1$. However, suppose $\mathbb{A}_1 = \emptyset$. Then, we have $[n] \setminus \mathbb{A}_1 = [n]$ and $\mathbb{S} \subset [n] \setminus \mathbb{A}_1$. Therefore, we have shown $\mathbb{S} \subset [n] \setminus \mathbb{A}_1$.

Then, according to the definition of $\mathbb{A}_1$, we have $\mathbb{A}_1 \subset V$, which gives $[n] \setminus V \subset [n] \setminus \mathbb{A}_1$. Hence, we have shown $[n] \setminus V \subset [n] \setminus \mathbb{A}_1$.

Next, we show $\mathbb{S} \cap ([n] \setminus V) \neq \emptyset$. Since $V = \{1, \ldots, t\}$, we have $[n] \setminus V = \{t+1, \ldots, n\}$. Notice that $\mathbb{S} = \{1, \ldots, a+1\} \cup \{t+1, \ldots, t+b+1\}$, where $t - 1 \geq a \geq 0$, and $m - 1 \geq b \geq 0$. Then, since $\{t+1, \ldots, t+b+1\} \cap \{t+1, \ldots, n\} = \{t+1, \ldots, t+b+1\} \neq \emptyset$. We have $\mathbb{S} \cap ([n] \setminus V) \neq \emptyset$.

Finally, we show $\mathbb{S} \cup ([n] \setminus V) = [n] \setminus \mathbb{A}_1$. When $\mathbb{A}_1 \neq \emptyset$, we have $\mathbb{A}_1 = \{1, \ldots, t\} \setminus \{1, \ldots, a+1\} = \{a+2, \ldots, t\}$. Then

$$\begin{aligned}
[n] \setminus \mathbb{A}_1 &= \big(\{1, \ldots, t\} \cup \{t+1, \ldots, t+m\} \cup O\big) \setminus \{a+2, \ldots, t\} \\
&= \big(\{1, \ldots, t\} \setminus \{a+2, \ldots, t\}\big) \cup \{t+1, \ldots, t+m\} \cup O \\
&= \{1, \ldots, a+1\} \cup \{t+1, \ldots, t+m\} \cup O.
\end{aligned}$$

Since $b \leq m-1$, we have $\{t+1, \ldots, t+b+1\} \subset \{t+1, \ldots, t+m\}$, which gives $\{t+1, \ldots, t+m\} = \{t+1, \ldots, t+m\} \cup \{t+1, \ldots, t+b+1\}$. Then, we have

$$\begin{aligned}
[n] \setminus \mathbb{A}_1 &= \{1, \ldots, a+1\} \cup \{t+1, \ldots, t+m\} \cup \{t+1, \ldots, t+b+1\} \cup O \\
&= \{1, \ldots, a+1\} \cup \{t+1, \ldots, t+b+1\} \cup \{t+1, \ldots, t+m\} \cup O \\
&= \mathbb{S} \cup ([n] \setminus V).
\end{aligned}$$

Hence, we have shown $\mathbb{S} \cup ([n] \setminus V) = [n] \setminus \mathbb{A}_1$ when $\mathbb{A}_1 \neq \emptyset$. However, suppose $\mathbb{A}_1 = \emptyset$. We have

$$\{1, \ldots, a+1\} = V \implies a+1 = t \implies a = t-1.$$

Then, we have $\mathbb{S} = \{1, \ldots, t\} \cup \{t+1, \ldots, t+b+1\}$. Hence, we have

$$\begin{aligned}\mathbb{S} \cup ([n] \setminus V) &= \{1, \ldots, t\} \cup \{t+1, \ldots, t+b+1\} \cup ([n] \setminus \{1, \ldots, t\}) \\ &= [n] = [n] \setminus \mathbb{A}_1.\end{aligned}$$

Therefore, we have shown $\mathbb{S} \cup ([n] \setminus V) = [n] \setminus \mathbb{A}_1$. This completes our proof. □

The following result is important for proving Proposition C.3.23.

Proposition C.3.22. Suppose $X_1, Y_1, X_2, Y_2 \in \mathcal{V}^n$, and $U, V \subset [n]$ are disjoint subsets of indices such that $U \cap V = \emptyset$. Then, if

$$R(X_1(i), (X_1(l))_{l \in [n] \setminus U}) = R(X_2(i), (X_2(l))_{l \in [n] \setminus U}), \text{ for any } i \in [n] \setminus U, \tag{C.61}$$

$$R(Y_1(i), (Y_1(l))_{l \in [n] \setminus V}) = R(Y_2(i), (Y_2(l))_{l \in [n] \setminus V}), \text{ for any } i \in [n] \setminus V, \tag{C.62}$$

and $R(X_j(i), X_j) = R(Y_j(i), Y_j)$, for any $i \in U \cup V, j \in \{1, 2\}$, we have $D(X_1, Y_1) = D(X_2, Y_2)$.

Proof. For any given $n \in \mathbb{N}$, let $P(k)$ be the statement of Proposition C.3.22 when $|U| + |V| = k$. We prove $P(k)$ holds for any $k \in \{0, \ldots, n\}$ by induction on k.

Base Case: we prove $P(0)$ is true. Since $|U| = |V| = 0$, we have $U = V = \emptyset$. Then, according to (C.61) and (C.62), we have $R(X_1, X_1) = R(X_2, X_2)$ and $R(Y_1, Y_1) = R(Y_2, Y_2)$, respectively. Hence, we have $D(X_1, Y_1) = D(X_2, Y_2)$. This proves $P(0)$ is true.

Induction Step: We show the implication

$$P(i), \text{ for any } i \leq k \implies P(k+1)$$

for any $k \in \{0, \ldots, n-1\}$. Assume the induction hypothesis: when $|U| + |V| = i \leq k$, the cases $P(i)$ is true.

Suppose $|U| + |V| = k + 1$. Then we have either

$$\begin{aligned}&\text{case (I)} : |U| = k+1, |V| = 0 \text{ or } |U| = 0, |V| = k+1, \\ \text{or } &\text{case (II)} : |U| > 0, |V| > 0, |U| + |V| = k+1,\end{aligned}$$

is true.

Suppose the case (I): $|U| = k+1, |V| = 0$ or $|U| = 0, |V| = k+1$ holds. Without loss of generality, let us assume $|U| = k+1$ and $|V| = 0$. The other case then can be proved in the same way after switching the labels of X and Y.

Let us assume (after relabelling) $U = \{1, \cdots, m\}$, and $Y_1(1) < \ldots < Y_1(m)$. According to (C.62), the order of $Y_1(1), \ldots, Y_1(m)$ equals to the order of $Y_2(1), \ldots, Y_2(m)$. Hence, we have $Y_2(1) < \cdots < Y_2(m)$. Thus, for both $j = 1, 2$, we have $Y_j(1), \ldots, Y_j(m)$. Then, we have

$$R(Y_j(1), Y_j) < \cdots < R(Y_j(m), Y_j), \text{ for both } j = 1, 2.$$

Since $R(X_j(i), X_j) = R(Y_j(i), Y_j)$, for any $i \in U \cup V, j \in \{1, 2\}$, we further have

$$R(X_j(1), X_j) < \cdots < R(X_j(m), X_j) \implies X_j(1) < \cdots < X_j(m).$$

Then, according to Lemma C.2.2, we have

$$R(X_j(1),(X_j(l))_{l\in\{1\}\cup([n]\setminus U)}) = R(X_j(1),X_j), j\in\{1,2\}.$$

Since $R(X_j(i),X_j) = R(Y_j(i),Y_j),$ for any $i\in U\cup V, j\in\{1,2\}$, we have

$$R(X_j(1),(X_j(l))_{l\in\{1\}\cup([n]\setminus U)}) = R(X_j(1),X_j) = R(Y_j(1),Y_j), j\in\{1,2\}.$$

Next, since $V=\emptyset$, and $Y_1(i)=Y_2(i)$, for any $i\in[n]\setminus V$, then we have $Y_1=Y_2$. Hence, $R(Y_1(1),Y_1)=R(Y_2(1),Y_2)$. Then, it follows that

$$R(X_1(1),(X_1(l))_{l\in\{1\}\cup([n]\setminus U)}) = R(X_2(1),(X_2(l))_{l\in\{1\}\cup([n]\setminus U)}).$$

According to (C.61), we have

$$R(X_1(i),(X_1(l))_{l\in[n]\setminus U}) = R(X_2(i),(X_2(l))_{l\in[n]\setminus U}), \text{ for any } i\in[n]\setminus U.$$

Subsequently, according to Lemma C.3.15, we have

$$R(X_1(i),(X_1(l))_{l\in\{1\}\cup([n]\setminus U)}) = R(X_2(i),(X_2(l))_{l\in\{1\}\cup([n]\setminus U)}), \text{ for any } i\in[n]\setminus U.$$

Therefore, we have shown that for any $i\in\{1\}\cup([n]\setminus U)$, we have

$$R(X_1(i),(X_1(l))_{l\in\{1\}\cup([n]\setminus U)}) = R(X_2(i),(X_2(l))_{l\in\{1\}\cup([n]\setminus U)}).$$

Notice that $V=\emptyset$. Hence, we have $|V|+|[n]\setminus(\{1\}\cup([n]\setminus U))| = 0+|U\setminus\{1\}| = m-1=k$. Meanwhile, notice that according to (C.62), we have $R(Y_1,Y_1)=R(Y_2,Y_2)$. Also, we have $R(X_j(i),X_j)=R(Y_j(i),Y_j)$, for any $i\in(U\cup V)\setminus\{1\}, j\in\{1,2\}$. Then, since $P(k)$ is true, we have $D(X_1,Y_1)=D(X_2,Y_2)$. This proves $P(k+1)$ when case (I) is true.

Suppose the case (II): $|U|>0, |V|>0, |U|+|V|=k+1$ is true. Since $|U|>0, |V|>0$, we have $U\neq\emptyset$, and $V\neq\emptyset$. Without loss of generality, let us assume (after relabelling) $V=\{1,\cdots,t\}$, $U=\{t+1,\cdots,t+m\}$, $X_1(1)<\cdots<X_1(t)$, and $Y_1(t+1)<\cdots<Y_1(t+m)$. According to (C.61), and (C.62), the order of $X_1(1),\ldots,X_1(t)$ and the order of $Y_1(1),\ldots,Y_1(m)$ equals to the order of $X_2(1),\ldots,X_2(t)$ and the order of $Y_2(1),\ldots,Y_2(m)$, respectively. Hence, we have $X_j(1)<\cdots<X_j(t)$, and $Y_j(t+1)<\cdots<Y_j(t+m)$ for both $j=1,2$.

Since $U=\{t+1,\cdots,t+m\}$ and $V=\{1,\cdots,t\}$, where $t,m\geq 1$, we have $1\in[n]\setminus U$ and $t+1\in[n]\setminus V$. Then, according to (C.61) and (C.62), we have

$$\begin{aligned} &R(X_1(1),(X_1(l))_{l\in[n]\setminus U}) = R(X_2(1),(X_2(l))_{l\in[n]\setminus U}),\\ \text{and } &R(Y_1(t+1),(Y_1(l))_{l\in[n]\setminus V}) = R(Y_2(t+1),(Y_2(l))_{l\in[n]\setminus V}), \end{aligned}$$

respectively. Let us denote

$$\begin{aligned} \alpha &= R(X_1(1),(X_1(l))_{l\in[n]\setminus U}) = R(X_2(1),(X_2(l))_{l\in[n]\setminus U}),\\ \beta &= R(Y_1(t+1),(Y_1(l))_{l\in[n]\setminus V}) = R(Y_2(t+1),(Y_2(l))_{l\in[n]\setminus V}). \end{aligned}$$

Notice that it is either $\alpha \leq \beta$ or $\alpha \geq \beta$. Let us assume $\alpha \leq \beta$. However, if $\alpha \geq \beta$, we can switch the labels between X_j and Y_j for both $j \in \{1, 2\}$, and relabel the relevant components of the data.

In the following, we are going to consider the following two cases

$$\begin{aligned} &\text{case (i)} : \alpha < \beta, \\ \text{and } &\text{case (ii)} : \alpha = \beta, \end{aligned}$$

separately.

Suppose the case (i): $\alpha < \beta$ holds. In other words,

$$R(X_j(1), (X_j(l))_{l \in [n] \setminus U}) < R(Y_j(t+1), (Y_j(l))_{l \in [n] \setminus V}), \text{ for both } j = 1, 2.$$

According to Lemma C.3.18, we have

$$R(X_j(1), (X_j(l))_{l \in [n] \setminus U}) = R(X_j(1), X_j) = R(Y_j(1), Y_j), \text{ for both } j = 1, 2. \tag{C.63}$$

Since $X_j(1) < \cdots < X_j(t)$ for both $j = 1, 2$, we have

$$R(X_j(1), X_j) < \cdots < R(X_j(t), X_j), \text{ for both } j = 1, 2.$$

Then, since $R(X_j(i), X_j) = R(Y_j(i), Y_j), \text{ for any } i \in U \cup V, j \in \{1, 2\}$, we have

$$\begin{aligned} &R(Y_j(1), Y_j) < \cdots < R(Y_j(t), Y_j), \text{ for both } j = 1, 2 \\ \Longrightarrow\ &Y_j(1) < Y_j(2) < \cdots < Y_j(t) \text{ for both } j = 1, 2. \end{aligned}$$

Notice that $V = \{1, \ldots, t\}$. Next, according to Lemma C.2.2, for both $j = 1, 2$, we have

$$R(Y_j(1), (Y_j(l))_{l \in \{1\} \cup ([n] \setminus V)}) = R(Y_j(1), Y_j) =^{(\text{C.63})} R(X_j(1), (X_j(l))_{l \in [n] \setminus U}).$$

According to (C.61), we have

$$R(X_1(1), (X_1(l))_{l \in [n] \setminus U}) = R(X_2(1), (X_2(l))_{l \in [n] \setminus U}).$$

Then, it follows

$$R(Y_1(1), (Y_1(l))_{l \in \{1\} \cup ([n] \setminus V)}) = R(Y_2(1), (Y_2(l))_{l \in \{1\} \cup ([n] \setminus V)}).$$

According to (C.62), for any $i \in [n] \setminus V$,

$$R(Y_1(i), (Y_1(l))_{l \in [n] \setminus V}) = R(Y_2(i), (Y_2(l))_{l \in [n] \setminus V}).$$

Subsequently, according to Lemma C.3.15, for any $i \in [n] \setminus V$, we have

$$R(Y_1(i), (Y_1(l))_{l \in \{1\} \cup ([n] \setminus V)}) = R(Y_2(i), (Y_2(l))_{l \in \{1\} \cup ([n] \setminus V)}).$$

Therefore, we have shown that for any $i \in ([n] \setminus V) \cup \{1\}$,

$$R(Y_1(i), (Y_1(l))_{l\in\{1\}\cup([n]\setminus V)}) = R(Y_2(i), (Y_2(l))_{l\in\{1\}\cup([n]\setminus V)}).$$

Notice that $|U| + |V \setminus \{1\}| = |U| + |V| - 1 = k$. Meanwhile, notice that according to (C.61), we have $R(X_1(i), (X_1(l))_{l\in[n]\setminus U}) = R(X_2(i), (X_2(l))_{l\in[n]\setminus U})$, for any $i \in [n] \setminus U$. Also, we have $R(X_j(i), X_j) = R(Y_j(i), Y_j)$, for any $i \in (U \cup V) \setminus \{1\}, j \in \{1, 2\}$. Then, since $P(k)$ is true, we have $D(X_1, Y_1) = D(X_2, Y_2)$. This proves $P(k+1)$ when the case (i) holds.

Now, suppose the case (ii): $\alpha = \beta$ holds. In other words,

$$\alpha = \beta = R(X_j(1), (X_j(l))_{l\in[n]\setminus U}) = R(Y_j(t+1), (Y_j(l))_{l\in[n]\setminus V}),$$

for $j \in \{1, 2\}$. Define

$$\begin{aligned} a &= \max\{i \in \{0, \cdots, t-1\} \mid R(X_1(1+i), (X_1(l))_{l\in[n]\setminus U}) = \alpha + i\}, \\ \text{and } b &= \max\{i \in \{0, \cdots, m-1\} \mid R(Y_1(t+1+i), (Y_1(l))_{l\in[n]\setminus V}) = \alpha + i\}. \end{aligned}$$

Next, define

$$\begin{aligned} \mathbb{S} &= \{1, \cdots, a+1\} \cup \{t+1, \cdots, t+1+b\}, \\ \mathbb{A}_1 &= V \setminus \{1, \cdots, a+1\} = \{1, \cdots, t\} \setminus \{1, \cdots, a+1\}, \\ \text{and } \mathbb{A}_2 &= U \setminus \{t+1, \cdots, t+1+b\} = \{t+1, \cdots, t+m\} \setminus \{t+1, \cdots, t+1+b\}. \end{aligned}$$

Then, according to Lemma C.3.20, for $j = 1, 2$, $\big(R(X_j(i), (X_j(l))_{l\in[n]\setminus\mathbb{A}_2})\big)_{i\in\mathbb{S}}$, and $\big(R(Y_j(i), (Y_j(l))_{l\in[n]\setminus\mathbb{A}_1})\big)_{i\in\mathbb{S}}$ are permutations of $\{\alpha, \cdots, \alpha + a + b + 1\}$, and

$$R(X_j(i), (X_j(l))_{l\in[n]\setminus\mathbb{A}_2}) = R(Y_j(i), (Y_j(l))_{l\in[n]\setminus\mathbb{A}_1}), \text{ for any } i \in \mathbb{S}.$$

Thus, $\big(R(X_1(i), (X_1(l))_{l\in[n]\setminus\mathbb{A}_2})\big)_{i\in\mathbb{S}}$ is a permutation of $\big(R(X_2(i), (X_2(l))_{l\in[n]\setminus\mathbb{A}_2})\big)_{i\in\mathbb{S}}$. Then, there exist a permutation $(\sigma(l))_{l\in\mathbb{S}}$ of $\mathbb{S}$ such that

$$R(X_2(i), (X_2(l))_{l\in[n]\setminus\mathbb{A}_2}) = R(X_1(\sigma(i)), (X_1(l))_{l\in[n]\setminus\mathbb{A}_2}), \text{ for any } i \in \mathbb{S}. \tag{C.64}$$

Since $R(X_j(i), (X_j(l))_{l\in[n]\setminus\mathbb{A}_2}) = R(Y_j(i), (Y_j(l))_{l\in[n]\setminus\mathbb{A}_1})$, for any $i \in \mathbb{S}, j \in \{1, 2\}$, then we have

$$\begin{aligned} R(Y_2(i), (Y_2(l))_{l\in[n]\setminus\mathbb{A}_1}) &= R(X_2(i), (X_2(l))_{l\in[n]\setminus\mathbb{A}_2}) \\ &=^{(C.64)} R(X_1(\sigma(i)), (X_1(l))_{l\in[n]\setminus\mathbb{A}_2}) \\ &= R(Y_1(\sigma(i)), (Y_1(l))_{l\in[n]\setminus\mathbb{A}_1}). \end{aligned} \tag{C.65}$$

Next, define $O = [n] \setminus (U \cup V)$. Then, according to Lemma C.3.21, we have

$$[n] \setminus \mathbb{A}_2 = \mathbb{S} \cup \mathbb{A}_1 \cup O, \tag{C.66}$$

$$[n] \setminus \mathbb{A}_1 = \mathbb{S} \cup \mathbb{A}_2 \cup O, \tag{C.67}$$

$$\mathbb{S}, [n] \setminus U \subset [n] \setminus \mathbb{A}_2,\ \mathbb{S} \cap ([n] \setminus U) \neq \emptyset,\ \mathbb{S} \cup ([n] \setminus U) = [n] \setminus \mathbb{A}_2, \tag{C.68}$$

$$\text{and } \mathbb{S}, [n] \setminus V \subset [n] \setminus \mathbb{A}_1,\ \mathbb{S} \cap ([n] \setminus V) \neq \emptyset,\ \mathbb{S} \cup ([n] \setminus V) = [n] \setminus \mathbb{A}_1. \tag{C.69}$$

According to (C.61), we have

$$R(X_1(i), (X_1(l))_{l \in [n] \setminus U}) = R(X_2(i), (X_2(l))_{l \in [n] \setminus U}),\ \text{for any } i \in \mathbb{S} \cap ([n] \setminus U). \tag{C.70}$$

Note that $\big(R(X_1(i), (X_1(l))_{l \in [n] \setminus \mathbb{A}_2})\big)_{i \in \mathbb{S}}$ is a permutation of $\big(R(X_2(i), (X_2(l))_{l \in [n] \setminus \mathbb{A}_2})\big)_{i \in \mathbb{S}}$, and according to (C.61), we have

$$R(X_1(i), (X_1(l))_{l \in [n] \setminus U}) = R(X_2(i), (X_2(l))_{l \in [n] \setminus U}) \text{ for any } i \in ([n] \setminus U) \setminus \mathbb{S}.$$

According to (C.68), we have $\mathbb{S}, [n] \setminus U \subset [n] \setminus \mathbb{A}_2,\ \mathbb{S} \cap ([n] \setminus U) \neq \emptyset$, and $\mathbb{S} \cup ([n] \setminus U) = [n] \setminus \mathbb{A}_2$. Then, we can apply Lemma C.3.17 and get

$$R(X_1(i), (X_1(l))_{l \in [n] \setminus \mathbb{A}_2}) = R(X_2(i), (X_2(l))_{l \in [n] \setminus \mathbb{A}_2}),\ \text{for any } i \in ([n] \setminus U) \setminus \mathbb{S}. \tag{C.71}$$

According to (C.65), $R(Y_2(i), (Y_2(l))_{l \in [n] \setminus \mathbb{A}_1})$ is a permutation of $R(Y_1(i), (Y_1(l))_{l \in [n] \setminus \mathbb{A}_1})$. Meanwhile, according to (C.62), we have

$$R(Y_1(i), (Y_1(l))_{l \in [n] \setminus V}) = R(Y_2(i), (Y_2(l))_{l \in [n] \setminus V}),\ \text{for any } i \in ([n] \setminus V) \setminus \mathbb{S}.$$

According to (C.69), we have $\mathbb{S}, [n] \setminus V \subset [n] \setminus \mathbb{A}_1,\ \mathbb{S} \cap ([n] \setminus V) \neq \emptyset$, and $\mathbb{S} \cup ([n] \setminus V) = [n] \setminus \mathbb{A}_1$. Subsequently, according to Lemma C.3.17, we have

$$R(Y_1(i), (Y_1(l))_{l \in [n] \setminus \mathbb{A}_1}) = R(Y_2(i), (Y_2(l))_{l \in [n] \setminus \mathbb{A}_1}),\ \text{for any } i \in ([n] \setminus V) \setminus \mathbb{S}. \tag{C.72}$$

Next, consider vectors $X_3, Y_3 \in \mathcal{V}^n$ such that

$$X_3(i) = X_1(\sigma(i)),\ \text{for any } i \in \mathbb{S},\ \text{and } X_3(i) = X_1(i),\ \text{for any } i \in [n] \setminus \mathbb{S};$$
$$Y_3(i) = Y_1(\sigma(i)),\ \text{for any } i \in \mathbb{S},\ \text{and } Y_3(i) = Y_1(i),\ \text{for any } i \in [n] \setminus \mathbb{S},$$

where $(\sigma(l))_{l \in \mathbb{S}}$ such that (C.64) is a permutation of $\mathbb{S}$. Since X_3 and Y_3 are permutations of X_1 and Y_1 according to indices, we have

$$D(X_3, Y_3) = D(X_1, Y_1). \tag{C.73}$$

Meanwhile, since $X_3(i) = X_1(i),\ \text{for any } i \in [n] \setminus \mathbb{S}$, and X_3 is a permutation of X_1, we have

$$R(X_3(i), X_3) = R(X_1(i), X_1),\ \text{for any } i \in (U \cup V) \setminus \mathbb{S}.$$

Similarly, since $Y_3(i) = Y_1(i),\ \text{for any } i \in [n] \setminus \mathbb{S}$, and Y_3 is a permutation of Y_1, we have

$$R(Y_3(i), Y_3) = R(Y_1(i), Y_1),\ \text{for any } i \in (U \cup V) \setminus \mathbb{S}.$$

Then, since $R(X_1(i), X_1) = R(Y_1(i), Y_1)$ for any $i \in U \cup V$, we have

$$R(X_3(i), X_3) = R(Y_3(i), Y_3),\ \text{for any } i \in (U \cup V) \setminus \mathbb{S}.$$

Since we also have $R(X_2(i), X_2) = R(Y_2(i), Y_2)$ for any $i \in U \cup V$, then if follows that

$$R(X_j(i), X_j) = R(Y_j(i), Y_j),\ \text{for any } i \in (U \cup V) \setminus \mathbb{S}, j \in \{2, 3\}.$$

Notice that according to the definition of $\mathbb{S}, \mathbb{A}_1$ and $\mathbb{A}_2$, we have $(U \cup V) \setminus \mathbb{S} = \mathbb{A}_1 \cup \mathbb{A}_2$. Hence, we have

$$R(X_j(i), X_j) = R(Y_j(i), Y_j),\ \text{for any } i \in \mathbb{A}_1 \cup \mathbb{A}_2, j \in \{2, 3\}. \tag{C.74}$$

According to (C.66), we have $[n] \setminus \mathbb{A}_2 = \mathbb{S} \cup \mathbb{A}_1 \cup O$. According to the definition of $\mathbb{S}$, $\mathbb{A}_1$, O, we have $\mathbb{S} \cap \mathbb{A}_1 = \emptyset$, and $\mathbb{S} \cap O = \emptyset$, which gives $(\mathbb{A}_1 \cup O) \cap \mathbb{S} = \emptyset$. Then, we have $\mathbb{A}_1 \cup O \subset [n] \setminus \mathbb{S}$. Since $X_3(i) = X_1(\sigma(i)),\ \text{for any } i \in \mathbb{S},\ \text{and } X_3(i) = X_1(i),\ \text{for any } i \in [n] \setminus \mathbb{S}$, then for any $i \in [n] \setminus \mathbb{A}_2$, we have

$$\begin{aligned} X_3(i) &= X_1(\sigma(i)),\ \text{if } i \in \mathbb{S}, \\ X_3(i) &= X_1(i),\ \text{if } i \in [n] \setminus \mathbb{S}. \end{aligned}$$

Hence, $(X_3(l))_{l \in [n] \setminus \mathbb{A}_2}$ is a permutation of $(X_1(l))_{l \in [n] \setminus \mathbb{A}_2}$. Then, we have

$$R(X_3(i), (X_3(l))_{l \in [n] \setminus \mathbb{A}_2}) = R(X_1(\sigma(i)), (X_1(l))_{l \in [n] \setminus \mathbb{A}_2}),\ \text{for any } i \in \mathbb{S}.$$

According to (C.64), we further have for any $i \in \mathbb{S}$,

$$\begin{aligned} R(X_3(i), (X_3(l))_{l \in [n] \setminus \mathbb{A}_2}) &= R(X_1(\sigma(i)), (X_1(l))_{l \in [n] \setminus \mathbb{A}_2}) \\ &= R(X_2(i), (X_2(l))_{l \in [n] \setminus \mathbb{A}_2}). \end{aligned}$$

Next, since $(X_3(l))_{l \in [n] \setminus \mathbb{A}_2}$ is a permutation of $(X_1(l))_{l \in [n] \setminus \mathbb{A}_2}$, and for any $i \in ([n] \setminus U) \setminus \mathbb{S}$, we have $X_3(i) = X_1(i)$. Then, for any $i \in ([n] \setminus U) \setminus \mathbb{S}$, we have

$$\begin{aligned} R(X_3(i), (X_3(l))_{l \in [n] \setminus \mathbb{A}_2}) &= R(X_1(i), (X_1(l))_{l \in [n] \setminus \mathbb{A}_2}) \\ &=^{(C.71)} R(X_2(i), (X_2(l))_{l \in [n] \setminus \mathbb{A}_2}). \end{aligned}$$

Hence, we have for any $i \in ([n] \setminus U) \cup \mathbb{S}$, we have

$$R(X_3(i), (X_3(l))_{l \in [n] \setminus \mathbb{A}_2}) = R(X_2(i), (X_2(l))_{l \in [n] \setminus \mathbb{A}_2}).$$

According to (C.68), we have $\mathbb{S} \cup ([n] \setminus U) = [n] \setminus \mathbb{A}_2$. Hence, it follows that

$$R(X_3(i), (X_3(l))_{l\in[n]\setminus\mathbb{A}_2}) = R(X_2(i), (X_2(l))_{l\in[n]\setminus\mathbb{A}_2}), \text{ for any } i \in [n] \setminus \mathbb{A}_2. \tag{C.75}$$

Similarly, according to (C.67), we have $[n] \setminus \mathbb{A}_1 = \mathbb{S} \cup \mathbb{A}_2 \cup O$. According to the definition of $\mathbb{S}$, $\mathbb{A}_2$, O, we have $\mathbb{S} \cap \mathbb{A}_2 = \emptyset$, and $\mathbb{S} \cap O = \emptyset$, which gives $(\mathbb{A}_2 \cup O) \cap \mathbb{S} = \emptyset$. Then, we have $\mathbb{A}_2 \cup O \subset [n] \setminus \mathbb{S}$. Since $Y_3(i) = Y_1(\sigma(i)),$ for any $i \in \mathbb{S},$ and $Y_3(i) = Y_1(i),$ for any $i \in [n] \setminus \mathbb{S},$ then for any $i \in [n] \setminus \mathbb{A}_1$, we have

$$\begin{aligned} Y_3(i) &= Y_1(\sigma(i)), \text{ if } i \in \mathbb{S}, \\ Y_3(i) &= Y_1(i), \text{ if } i \in [n] \setminus \mathbb{S}. \end{aligned}$$

Hence, $(Y_3(l))_{l\in[n]\setminus\mathbb{A}_1}$ is a permutation of $(Y_1(l))_{l\in[n]\setminus\mathbb{A}_1}$. Then, we have

$$R(Y_3(i), (Y_3(l))_{l\in[n]\setminus\mathbb{A}_1}) = R(Y_1(\sigma(i)), (Y_1(l))_{l\in[n]\setminus\mathbb{A}_1}), \text{ for any } i \in \mathbb{S}.$$

According to (C.65), we further have for any $i \in \mathbb{S}$,

$$\begin{aligned} R(Y_3(i), (Y_3(l))_{l\in[n]\setminus\mathbb{A}_1}) &= R(Y_1(\sigma(i)), (Y_1(l))_{l\in[n]\setminus\mathbb{A}_1}) \\ &= R(Y_2(i), (Y_2(l))_{l\in[n]\setminus\mathbb{A}_1}). \end{aligned}$$

Since $(Y_3(l))_{l\in[n]\setminus\mathbb{A}_1}$ is a permutation of $(Y_1(l))_{l\in[n]\setminus\mathbb{A}_1}$, and for any $i \in ([n] \setminus V) \setminus \mathbb{S}$, we have $Y_3(i) = Y_1(i)$. Then, we have

$$\begin{aligned} R(Y_3(i), (Y_3(l))_{l\in[n]\setminus\mathbb{A}_1}) &= R(Y_1(i), (Y_1(l))_{l\in[n]\setminus\mathbb{A}_1}) \\ &=^{(C.72)} R(Y_2(i), (Y_2(l))_{l\in[n]\setminus\mathbb{A}_1}). \end{aligned}$$

Hence, for any $i \in ([n] \setminus V) \cup \mathbb{S}$, we have

$$R(Y_3(i), (Y_3(l))_{l\in[n]\setminus\mathbb{A}_1}) = R(Y_2(i), (Y_2(l))_{l\in[n]\setminus\mathbb{A}_1}).$$

According to (C.69), we have $\mathbb{S} \cup ([n] \setminus V) = [n] \setminus \mathbb{A}_1$. Hence, it follows that

$$R(Y_3(i), (Y_3(l))_{l\in[n]\setminus\mathbb{A}_1}) = R(Y_2(i), (Y_2(l))_{l\in[n]\setminus\mathbb{A}_1}), \text{ for any } i \in [n] \setminus \mathbb{A}_1. \tag{C.76}$$

Notice that (C.72), (C.75) and (C.76) are true, and $|\mathbb{A}_1| + |\mathbb{A}_2| = (|V| - 1 - a) + (|U| - 1 - b) = |V| + |U| - 2 - a - b = k - 1 - a - b \leq k - 1 \leq k$. Then, since $P(i)$ is true for any $i \leq k$, we have

$$D(X_3, Y_3) = D(X_2, Y_2) =^{(C.73)} D(X_1, Y_1).$$

Therefore, we have shown $P(k+1)$ is true. This completes our proof. □

We now present Proposition C.3.23, which is proved by applying Proposition C.3.22.

Proposition C.3.23. Suppose $X, Y \in \mathcal{V}^n$ and $U, V \subset [n]$ are disjoint subsets of indices such that $U \cap V = \emptyset$. Suppose X_1, Y_1 satisfy Conditions C.3.10 and C.3.11, and X_2, Y_2 also satisfy Conditions C.3.10 and C.3.11. In other words, $X_1, X_2 \in \mathcal{V}^n$ are imputations of X for indices

U, $Y_1, Y_2 \in \mathcal{V}^n$ are imputations of Y for indices V, and $R(X_j(i), X_j) = R(Y_j(i), Y_j)$, for any $i \in U \cup V, j \in \{1, 2\}$. Then, we have $D(X_1, Y_1) = D(X_2, Y_2)$.

Proof. Since $X_1, X_2 \in \mathcal{V}^n$ are imputations of X for indices U, then according to the definition of imputations, we have $X_1(i) = X_2(i)$ for any $i \in [n] \setminus U$. Hence, we have

$$R(X_1(i), (X_1(l))_{l\in[n]\setminus U}) = R(X_2(i), (X_2(l))_{l\in[n]\setminus U}), \text{ for any } i \in [n] \setminus U.$$

Similarly, since $Y_1, Y_2 \in \mathcal{V}^n$ are imputations of Y for indices V, then we have $Y_1(i) = Y_2(i)$ for any $i \in V$, which follows

$$R(Y_1(i), (Y_1(l))_{l\in[n]\setminus V}) = R(Y_2(i), (Y_2(l))_{l\in[n]\setminus V}), \text{ for any } i \in [n] \setminus V.$$

Since $R(X_j(i), X_j) = R(Y_j(i), Y_j)$, for any $i \in U \cup V, j \in \{1, 2\}$, then according to Proposition C.3.22, we have $D(X_1, Y_1) = D(X_2, Y_2)$. This completes our proof. □

Table C.2 provides an example of two pairs of imputations (X_1, Y_1) and (X_2, Y_2) for partially observed data (X, Y) that satisfy Conditions C.3.10 and C.3.11, and one can compute $D(X_1, Y_1) = D(X_2, Y_2) = 14$. While the equivalence of missing case I found in Lemma C.3.6 can be proved in a relatively straightforward approach, the proof of Proposition C.3.23 is much more challenging. For missing case I, the condition in (C.9) explicitly provides the ranks of all components in U for both X_1 and X_2. On the other hand, Condition C.3.11 for missing case II only requires the ranks of all components of X_j, Y_j for indices $U \cup V$ to be matched with their paired components, for $j \in \{1, 2\}$. However, the rank of component $X_1(i)$ in X_1 is not necessary equal to the rank of $X_2(i)$ in X_2, for $i \in U \cup V$. Similarly, it is not necessarily true for all $i \in U \cup V$ that $R(Y_1(i), Y_1) = R(Y_2(i), Y_2)$.

In Theorem C.3.29, we show that the minimum possible Spearman's footrule is achieved when imputations X_* and Y_* satisfy both Conditions C.3.10 and C.3.11.

Theorem C.3.29 provides the conditions under which the imputation X_*, Y_* will minimise Spearman's footrule under missing case II. Moreover, Proposition C.3.13 shows that X_*, Y_* exists and so this lower bound is tight.

To further illustrate the results presented in this section for missing case II, we include a simple example in Table C.2. This example demonstrates how Proposition C.3.13, Algorithm 3, Proposition C.3.23, and Theorem C.3.29 can be jointly applied to obtain the minimum Spearman's footrule under Missing Case II. One can demonstrate that $D(X_*, Y_*) = 14$ in this example is the minimum possible value for this statistic among the $\binom{8}{2} \cdot \binom{8}{2} = 784$ possible imputations via an exhaustive enumeration.

We now prove Theorem C.3.29 for the tight lower bounds of Spearman's footrule under missing case II. First, we prove some lemmas which will be useful for proving the theorem.

Lemma C.3.24. Suppose $X, Y \in \mathcal{V}^n$. Let $V = \{1, \ldots, t\}, U = \{t+1, \ldots, t+m\} \subset [n]$ be subsets of indices such that $t, m \geq 1$. Suppose $X(1) < \ldots < X(t)$ and $Y(t+1) < \ldots < Y(t+m)$. Assume $R(X(1), (X(l))_{l\in[n]\setminus U}) = R(Y(t+1), (Y(l))_{l\in[n]\setminus V})$, and $Y(1) > Y(t+1)$. Denote

$$\alpha = R(X(1), (X(l))_{l\in[n]\setminus U}) = R(Y(t+1), (Y(l))_{l\in[n]\setminus V}). \tag{C.77}$$

Let X_1, Y_1 be imputations of X, Y for indices U and $V \setminus \{1\}$, respectively. Then, if $R(X_1(i), X_1) = R(Y_1(i), Y_1)$ for any $i \in U \cup (V \setminus \{1\})$, we have

$$\begin{aligned} R(Y_1(t+1), (Y_1(l))_{l\in[n]\setminus V}) &= R(Y_1(t+1), Y_1) \\ &= R(Y_1(t+1), (Y_1(l))_{l\in[n]\setminus(V\setminus\{1\})}) = \alpha, \end{aligned} \tag{C.78}$$

and

$$\begin{aligned} R(X_1(1), (X_1(l))_{l\in[n]\setminus U}) &= R(X_1(t+1), X_1) \\ &= R(X_1(t+1), (X_1(l))_{l\in[n]\setminus(U\setminus\{t+1\})}) = \alpha. \end{aligned} \tag{C.79}$$

Additionally, we have $X_1(1) < \ldots < X_1(t)$, and $Y_1(t+1) < \ldots < Y_1(t+m)$.

Proof. Since X_1, Y_1 are imputations of X, and Y for indices U and $V \setminus \{1\}$, respectively, we have

$$X_1(i) = X(i), \text{ for any } i \in [n] \setminus U, \tag{C.80}$$

$$\text{and } Y_1(i) = Y(i), \text{ for any } i \in [n] \setminus (V \setminus \{1\}). \tag{C.81}$$

We first show that $X_1(1) < \ldots < X_1(t)$, and $Y_1(t+1) < \ldots < Y_1(t+m)$. Notice that $\{1, \ldots, t\} \subset [n] \setminus U$, and $\{t+1, t+m\} \subset [n] \setminus V \subset [n] \setminus (V \setminus \{1\})$. Then according to (C.80) and (C.81), we have $X_1(i) = X(i)$, for any $i \in \{1, \ldots, t\}$ and $Y_1(i) = Y(i)$, for any $i \in \{t+1, \ldots, t+m\}$, respectively. Then, since $X(1) < \ldots < X(t)$ and $Y(t+1) < \ldots < Y(t+m)$, we have $X_1(1) < \ldots < X_1(t)$, and $Y_1(t+1) < \ldots < Y_1(t+m)$.

Next, we show that (C.78) is true. To start, according to the definition of rank, we have

$$\begin{aligned} &R(Y(t+1), (Y(l))_{l\in\{1\}\cup([n]\setminus V)}) \\ &= \sum_{l\in\{1\}\cup([n]\setminus V)} \mathrm{I}\left(Y(l) \le Y(t+1)\right) \\ &= \mathrm{I}\left(Y(1) \le Y(t+1)\right) + \sum_{l\in[n]\setminus V} \mathrm{I}\left(Y(l) \le Y(t+1)\right). \end{aligned}$$

Since $Y(1) > Y(t+1)$, we have $\mathrm{I}\left(Y(1) \le Y(t+1)\right) = 0$. Hence, we have

$$\begin{aligned} R(Y(t+1), (Y(l))_{l\in\{1\}\cup([n]\setminus V)}) &= \sum_{l\in[n]\setminus V} \mathrm{I}\left(Y(l) \le Y(t+1)\right) \\ &= R(Y(t+1), (Y(l))_{l\in[n]\setminus V}) =^{(\mathrm{C}.77)} \alpha. \end{aligned}$$

Since $\{1\} \cup ([n] \setminus V) = \{1\} \cup \{t+1, \ldots, n\} = [n] \setminus (\{1, \ldots, t\} \setminus \{1\}) = [n] \setminus (V \setminus \{1\})$, we have $(Y(l))_{l\in\{1\}\cup([n]\setminus V)} = (Y(l))_{l\in[n]\setminus(V\setminus\{1\})}$. Then, it follows that

$$R(Y(t+1), (Y(l))_{l\in\{1\}\cup([n]\setminus V)}) = R(Y(t+1), (Y(l))_{l\in[n]\setminus(V\setminus\{1\})}).$$

Hence, we have

$$R(Y(t+1), (Y(l))_{l\in[n]\setminus(V\setminus\{1\})}) = \alpha. \tag{C.82}$$

Since $U = \{t+1, \ldots, t+m\}$, where $t, m \geq 1$, we have $1 \in [n] \setminus U$. Then, according to (C.80), we have $X_1(1) = X(1)$. Notice that (C.80) also means that $(X_1(l))_{l\in[n]\setminus U} = (X(l))_{l\in[n]\setminus U}$. Hence, we have

$$R(X_1(1), (X_1(l))_{l\in[n]\setminus U}) = R(X(1), (X(l))_{l\in[n]\setminus U}) =^{(C.77)} \alpha.$$

Similarly, since $V = \{1, 2, \ldots, t\}$, where $t \geq 1$, we have $t+1 \in [n] \setminus (V \setminus \{1\})$. Then, according to (C.81), we have $Y_1(t+1) = Y(t+1)$. Notice that (C.81) also means that $(Y_1(l))_{l\in[n]\setminus(V\setminus\{1\})} = (Y(l))_{l\in[n]\setminus(V\setminus\{1\})}$. Hence, we have

$$R(Y_1(t+1), (Y_1(l))_{l\in[n]\setminus(V\setminus\{1\})}) = R(Y(t+1), (Y(l))_{l\in[n]\setminus(V\setminus\{1\})}) =^{(C.82)} \alpha.$$

Notice that $[n]\setminus V \subset [n]\setminus(V\setminus\{1\})$. Then, according to (C.81), we have $Y_1(i) = Y(i),$ for any $i \in [n] \setminus V$. Hence, we have $(Y_1(l))_{l\in[n]\setminus V} = (Y(l))_{l\in[n]\setminus V}$. Then, it follows that

$$R(Y_1(t+1), (Y_1(l))_{l\in[n]\setminus V}) = R(Y(t+1), (Y(l))_{l\in[n]\setminus V}) =^{(C.77)} \alpha.$$

Combining the above results, we have

$$\begin{aligned} R(X_1(1), (X_1(l))_{l\in[n]\setminus U}) &= R(Y_1(t+1), (Y_1(l))_{l\in[n]\setminus V}) \\ &= R(Y_1(t+1), (Y_1(l))_{l\in[n]\setminus(V\setminus\{1\})}) = \alpha. \end{aligned} \tag{C.83}$$

Next, suppose $V \setminus \{1\} \neq \emptyset$. Then according to Lemma C.2.1, we have

$$R(X_1(i), X_1) \geq R(X_1(i), (X_1(l))_{l\in[n]\setminus U}), \text{ for any } i \in V \setminus \{1\}.$$

Since $R(X_1(i), X_1) = R(Y_1(i), Y_1)$ for any $i \in U \cup (V \setminus \{1\})$, then we have

$$R(Y_1(i), Y_1) = R(X_1(i), X_1) \geq R(X_1(i), (X_1(l))_{l\in[n]\setminus U}), \text{ for any } i \in V \setminus \{1\}.$$

Since $X_1(1) < \ldots < X_1(t)$, then for any $i \in V \setminus \{1\}$, we further have

$$\begin{aligned} R(Y_1(i), Y_1) &\geq R(X_1(i), (X_1(l))_{l\in[n]\setminus U}) \\ &> R(X_1(1), (X_1(l))_{l\in[n]\setminus U}) \\ &=^{(C.83)} R(Y_1(t+1), (Y_1(l))_{l\in[n]\setminus(V\setminus\{1\})}) =^{(C.83)} \alpha. \end{aligned}$$

Hence, for any $i \in V \setminus \{1\}$, we have

$$R(Y_1(i), Y_1) > R(Y_1(t+1), (Y_1(l))_{l\in[n]\setminus(V\setminus\{1\})}).$$

Subsequently, according to Lemma C.2.2, we have

$$R(Y_1(t+1), Y_1) = R(Y_1(t+1), (Y_1(l))_{l\in[n]\setminus(V\setminus\{1\})}) =^{(C.83)} \alpha. \tag{C.84}$$

In particular, the above equations still hold when $V \setminus \{1\} = \emptyset$. Combining (C.83) and (C.84), we

have

$$\begin{aligned} R(Y_1(t+1),(Y_1(l))_{l\in[n]\setminus V}) &= R(Y_1(t+1),(Y_1(l))_{l\in[n]\setminus(V\setminus\{1\})}) \\ &= R(Y_1(t+1),Y_1)=\alpha. \end{aligned}$$

This proves (C.78).

Next, we show that (C.79) is true. Suppose $U\setminus\{t+1\}\neq\emptyset$. Since $R(X_1(i),X_1)=R(Y_1(i),Y_1)$ for any $i\in U\cup(V\setminus\{1\})$, we have

$$R(X_1(i),X_1)=R(Y_1(i),Y_1),\ \text{for any } i\in U\setminus\{t+1\}.$$

According to Lemma C.2.1, we have

$$R(Y_1(i),Y_1)\geq R(Y_1(i),(Y_1(l))_{l\in[n]\setminus V}),\ \text{for any } i\in U\setminus\{t+1\}.$$

Further, since $Y_1(t+1)<\ldots<Y_1(t+m)$, we have

$$R(Y_1(i),(Y_1(l))_{l\in[n]\setminus V})>R(Y_1(t+1),(Y_1(l))_{l\in[n]\setminus V}),\ \text{for any } i\in U\setminus\{t+1\}.$$

Hence, we have

$$R(X_1(i),X_1)>R(Y_1(t+1),(Y_1(l))_{l\in[n]\setminus V})=^{(\text{C.77})}\alpha,\ \text{for any } i\in U\setminus\{t+1\}.$$

Next, notice that $R(X_1(i),X_1)=R(Y_1(i),Y_1)$ for any $i\in U\cup(V\setminus\{1\})$, and $t+1\in U$. Then we have

$$R(X_1(t+1),X_1)=R(Y_1(t+1),Y_1)=^{(\text{C.84})}\alpha.$$

Then, according to Lemma C.2.2, we have

$$R(X_1(t+1),(X_1(l))_{l\in[n]\setminus(U\setminus\{t+1\})})=R(X_1(t+1),X_1)=\alpha. \tag{C.85}$$

In particular, the above equations still hold when $U\setminus\{t+1\}=\emptyset$. Combining (C.83) and (C.85), we have

$$\begin{aligned} R(X_1(1),(X_1(l))_{l\in[n]\setminus U}) &= R(X_1(t+1),(X_1(l))_{l\in[n]\setminus(U\setminus\{t+1\})}) \\ &= R(X_1(t+1),X_1)=\alpha. \end{aligned}$$

This proves (C.79), and completes our proof.

□

Lemma C.3.25. Following Lemma C.3.24, let X_2 and Y_2 be imputations of X_1, Y_1 for indices $U\setminus\{t+1\}$ and V, respectively, such that $R(X_2(i),X_2)=R(Y_2(i),Y_2)$ for any $i\in(U\setminus\{t+1\})\cup V$. Then, we have $R(X_2(i),X_2)=R(Y_2(i),Y_2)$, for any $i\in U\cup V$.

Proof. According to the definition of imputations, since X_2 and Y_2 are imputations of X_1, Y_1

or indices $U \setminus \{t+1\}$ and V, respectively, we have

$$X_2(i) = X_1(i), \text{ for any } i \in [n] \setminus (U \setminus \{t+1\}), \tag{C.86}$$

$$\text{and } Y_2(i) = Y_1(i), \text{ for any } i \in [n] \setminus V. \tag{C.87}$$

According to the definition of Lemma C.3.24, we have $X_1(1) < \ldots < X_1(t)$ and $Y_1(t+1) < \ldots < Y_1(t+m)$. Notice that $\{1, \ldots, t\} \subset [n] \setminus U \subset [n] \setminus (U \setminus \{t+1\})$, and $\{t+1, t+m\} \subset [n] \setminus V$. Then according to (C.86) and (C.87), we have $X_2(i) = X_1(i)$, for any $i \in \{1, \ldots, t\}$ and $Y_2(i) = Y_1(i)$, for any $i \in \{t+1, \ldots, t+m\}$, respectively. Then, since $X_1(1) < \ldots < X_1(t)$ and $Y_1(t+1) < \ldots < Y_1(t+m)$, we have $X_2(1) < \ldots < X_2(t)$ and $Y_2(t+1) < \ldots < Y_2(t+m)$.

Since $t+1 \in [n] \setminus (U \setminus \{t+1\})$, then according to (C.86), we have $X_2(t+1) = X_1(t+1)$. Notice that (C.86) also gives $(X_2(l))_{l \in [n] \setminus (U \setminus \{t+1\})} = (X_1(l))_{l \in [n] \setminus (U \setminus \{t+1\})}$. Hence, we have

$$\begin{aligned} R(X_2(t+1), (X_2(l))_{l \in [n] \setminus (U \setminus \{t+1\})}) &= R(X_1(t+1), (X_1(l))_{l \in [n] \setminus (U \setminus \{t+1\})}) \\ &=^{(C.79)} \alpha. \end{aligned}$$

Next, since $V = \{1, \ldots, t\}$, we have $t+1 \in [n] \setminus V$. Then, according to (C.87), we have $Y_2(t+1) = Y_1(t+1)$. Notice that (C.87) also gives $(Y_2(l))_{l \in [n] \setminus V} = (Y_1(l))_{l \in [n] \setminus V}$. Hence, we have

$$R(Y_2(t+1), (Y_2(l))_{l \in [n] \setminus V}) = R(Y_1(t+1), (Y_1(l))_{l \in [n] \setminus V}) =^{(C.78)} \alpha.$$

Combining the above results, we have

$$R(X_2(t+1), (X_2(l))_{l \in [n] \setminus (U \setminus \{t+1\})}) = R(Y_2(t+1), (Y_2(l))_{l \in [n] \setminus V}) = \alpha. \tag{C.88}$$

Suppose $U \setminus \{t+1\} \neq \emptyset$. Then, according to Lemma C.2.1, we have

$$R(Y_2(i), Y_2) \geq R(Y_2(i), (Y_2(l))_{l \in [n] \setminus V}), \text{ for any } U \setminus \{t+1\}.$$

Since $R(X_2(i), X_2) = R(Y_2(i), Y_2)$ for any $i \in (U \setminus \{t+1\}) \cup V$, we have $R(X_2(i), X_2) = R(Y_2(i), Y_2)$ for any $i \in U \setminus \{t+1\}$. Then, it follows that

$$R(X_2(i), X_2) = R(Y_2(i), Y_2) \geq R(Y_2(i), (Y_2(l))_{l \in [n] \setminus V}), \text{ for any } U \setminus \{t+1\}.$$

Since $Y_2(t+1) < \ldots < Y_2(t+m)$, then for any $i \in U \setminus \{t+1\} = \{t+2, \ldots, t+m\}$, we have

$$\begin{aligned} R(X_2(i), X_2) &\geq R(Y_2(i), (Y_2(l))_{l \in [n] \setminus V}) \\ &> R(Y_2(t+1), (Y_2(l))_{l \in [n] \setminus V}) \\ &=^{(C.88)} R(X_2(t+1), (X_2(l))_{l \in [n] \setminus (U \setminus \{t+1\})}). \end{aligned}$$

Then, according to Lemma C.2.2, we have

$$R(X_2(t+1), X_2) = R(X_2(t+1), (X_2(l))_{l \in [n] \setminus (U \setminus \{t+1\})}) =^{(C.88)} \alpha. \tag{C.89}$$

In particular, the above equation still holds when $U \setminus \{t+1\} = \emptyset$.

Next, notice that $[n] \setminus U \subset [n] \setminus (U \setminus \{t+1\})$. Then, according to Lemma C.2.1, we have

$$R(X_2(1), (X_2(l))_{l\in[n]\setminus(U\setminus\{t+1\})}) \geq R(X_2(1), (X_2(l))_{l\in[n]\setminus U}).$$

Since $U = \{t+1, \ldots, t+m\}$, where $t, m \geq 1$, we have $1 \in [n] \setminus (U \setminus \{t+1\})$. Hence, according to (C.86), we have $X_2(1) = X_1(1)$. Meanwhile, notice that $[n] \setminus U \subset [n] \setminus (U \setminus \{t+1\})$. Then (C.86) also gives that $(X_2(l))_{l\in[n]\setminus U} = (X_1(l))_{l\in[n]\setminus U}$. Hence, we have

$$\begin{aligned} R(X_2(1), (X_2(l))_{l\in[n]\setminus U}) &= R(X_1(1), (X_1(l))_{l\in[n]\setminus U}) \\ &=^{(C.78)} \alpha =^{(C.89)} R(X_2(t+1), (X_2(l))_{l\in[n]\setminus(U\setminus\{t+1\})}). \end{aligned}$$

Therefore, we have

$$R(X_2(1), (X_2(l))_{l\in[n]\setminus(U\setminus\{t+1\})}) \geq R(X_2(t+1), (X_2(l))_{l\in[n]\setminus(U\setminus\{t+1\})}).$$

Since $X_2 \in \mathcal{V}^n$ is a vector of distinct real values,

$$\begin{aligned} &R(X_2(1), (X_2(l))_{l\in[n]\setminus(U\setminus\{t+1\})}) \neq R(X_2(t+1), (X_2(l))_{l\in[n]\setminus(U\setminus\{t+1\})}) \\ \Longrightarrow\ &R(X_2(1), (X_2(l))_{l\in[n]\setminus(U\setminus\{t+1\})}) > R(X_2(t+1), (X_2(l))_{l\in[n]\setminus(U\setminus\{t+1\})}) =^{(C.89)} \alpha. \end{aligned}$$

Since $X_2(1) < \ldots < X_2(t)$. For any $i \in V = \{1, \ldots, t\}$, we have

$$R(X_2(i), (X_2(l))_{l\in[n]\setminus(U\setminus\{t+1\})}) \geq R(X_2(1), (X_2(l))_{l\in[n]\setminus(U\setminus\{t+1\})}) > \alpha.$$

Then, according to Lemma C.2.1, for any $i \in V = \{1, \ldots, t\}$, we have

$$\begin{aligned} R(X_2(i), X_2) &\geq R(X_2(i), (X_2(l))_{l\in[n]\setminus(U\setminus\{t+1\})}) \\ &> \alpha =^{(C.88)} R(Y_2(t+1), (Y_2(l))_{l\in[n]\setminus V}). \end{aligned}$$

Since $R(X_2(i), X_2) = R(Y_2(i), Y_2)$ for any $i \in (U\setminus\{t+1\})\cup V$, we have $R(Y_2(i), Y_2) = R(X_2(i), X_2)$ for any $i \in V$. Then, it follows that

$$R(Y_2(i), Y_2) = R(X_2(i), X_2) > R(Y_2(t+1), (Y_2(l))_{l\in[n]\setminus V}), \text{ for any } i \in V.$$

Subsequently, according to Lemma C.2.2, we have

$$R(Y_2(t+1), Y_2) = R(Y_2(t+1), (Y_2(l))_{l\in[n]\setminus V}) =^{(C.87)} \alpha =^{(C.88)} R(X_2(t+1), X_2).$$

Hence, we obtain $R(Y_2(t+1), Y_2) = R(X_2(t+1), X_2)$. Since $R(X_2(i), X_2) = R(Y_2(i), Y_2)$ for any $i \in (U \setminus \{t+1\}) \cup V$, then we have

$$R(X_2(i), X_2) = R(Y_2(i), Y_2), \text{for any } i \in U \cup V.$$

This completes our proof. □

Lemma C.3.26. Suppose $X, Y \in \mathcal{V}^n$. Let $V = \{1, \ldots, t\}, U = \{t+1, \ldots, t+m\} \subset [n]$ be subsets

of indices such that $t, m \geq 1$. Suppose $X(1) < \ldots < X(t)$ and $Y(t+1) < \ldots < Y(t+m)$, and assume $R(Y(t+1), (Y(l))_{l\in\{1\}\cup([n]\setminus V)}) > R(X(1), (X(l))_{l\in[n]\setminus U})$. Let us denote

$$\alpha = R(X(1), (X(l))_{l\in[n]\setminus U}). \tag{C.90}$$

Let X_1 and Y_1 be imputations of X and Y for U and $V \setminus \{1\}$ such that $R(X_1(i), X_1) = R(Y_1(i), Y_1)$ for any $i \in (V \setminus \{1\}) \cup U$. Then, when $V \setminus \{1\} \neq \emptyset$, for any $V \setminus \{1\}$,

$$R(Y_1(i), (Y_1(l))_{l\in[n]\setminus\{1\}}) \geq R(X_1(1), (X_1(l))_{l\in[n]\setminus U}) = R(X_1(1), X_1) = \alpha.$$

Additionally, we have $X_1(1) < \ldots < X_1(t)$, and $Y_1(t+1) < \ldots < Y_1(t+m)$.

Proof. Since X_1, Y_1 are imputations of X, and Y for indices U and $V \setminus \{1\}$, respectively, we have

$$X_1(i) = X(i), \text{ for any } i \in [n] \setminus U, \tag{C.91}$$

$$\text{and } Y_1(i) = Y(i), \text{ for any } i \in [n] \setminus (V \setminus \{1\}). \tag{C.92}$$

To start, we first show that $X_1(1) < \ldots < X_1(t)$, and $Y_1(t+1) < \ldots < Y_1(t+m)$. Notice that $\{1, \ldots, t\} \subset [n] \setminus U$, and $\{t+1, t+m\} \subset [n] \setminus V \subset [n] \setminus (V \setminus \{1\})$. Then according to (C.91) and (C.92), we have $X_1(i) = X(i)$, for any $i \in \{1, \ldots, t\}$ and $Y_1(i) = Y(i)$, for any $i \in \{t+1, \ldots, t+m\}$, respectively. Then, since $X(1) < \ldots < X(t)$ and $Y(t+1) < \ldots < Y(t+m)$, we have $X_1(1) < \ldots < X_1(t)$, and $Y_1(t+1) < \ldots < Y_1(t+m)$.

Since $U = \{t+1, \ldots, t+m\}$, where $t \geq 1$, we have $1 \in [n] \setminus U$. Then according to (C.91), we have $X_1(1) = X(1)$. Notice that (C.91) also gives $(X_1(l))_{l\in[n]\setminus U} = (X(l))_{l\in[n]\setminus U}$. Hence, we have

$$R(X_1(1), (X_1(l))_{l\in[n]\setminus U}) = R(X(1), (X(l))_{l\in[n]\setminus U}) =^{(C.90)} \alpha. \tag{C.93}$$

Similarly, since $V = \{1, \ldots, t\}$, where $t \geq 1$, we have $t+1 \in [n] \setminus (V \setminus \{1\})$. Hence, according to (C.92), we have $Y_1(t+1) = Y(t+1)$. Since $[n] \setminus (V \setminus \{1\}) = \{1\} \cup ([n] \setminus V)$, then (C.92) also gives that $(Y_1(l))_{l\in\{1\}\cup([n]\setminus V)} = (Y(l))_{l\in\{1\}\cup([n]\setminus V)}$. Hence, we have

$$R(Y_1(t+1), (Y_1(l))_{l\in\{1\}\cup([n]\setminus V)}) = R(Y(t+1), (Y(l))_{l\in\{1\}\cup([n]\setminus V)}).$$

Since $R(Y(t+1), (Y(l))_{l\in\{1\}\cup([n]\setminus V)}) > R(X(1), (X(l))_{l\in[n]\setminus U})$, we further have

$$R(Y_1(t+1), (Y_1(l))_{l\in\{1\}\cup([n]\setminus V)}) > R(X(1), (X(l))_{l\in[n]\setminus U}) =^{(C.90)} \alpha. \tag{C.94}$$

Next, notice that according to (C.92), we have

$$R(X_1(i), X_1) = R(Y_1(i), Y_1), \text{ for any } i \in U.$$

According to Lemma C.2.1, for any $i \in U$, we have

$$R(Y_1(i), Y_1) \geq R(Y_1(i), (Y_1(l))_{l\in[n]\setminus(V\setminus\{1\})}).$$

Since $Y_1(t+m) > \cdots > Y_1(t+1)$, for any $i \in U$, we further have

$$R(Y_1(i), (Y_1(l))_{l\in[n]\setminus(V\setminus\{1\})}) \geq R(Y_1(t+1), (Y_1(l))_{l\in[n]\setminus(V\setminus\{1\})}) >^{(C.94)} \alpha.$$

Hence, we have

$$R(X_3(i), X_3) > \alpha, \text{ for any } i \in U.$$

Notice that according to (C.93), we have

$$R(X_1(1), (X_1(l))_{l\in[n]\setminus U}) = \alpha.$$

Subsequently, according to Lemma C.2.2, we have

$$R(X_1(1), X_1) = R(X_1(1), (X_1(l))_{l\in[n]\setminus U}) = \alpha. \tag{C.95}$$

Suppose $V \setminus \{1\} \neq \emptyset$. Then for any $i \in V \setminus \{1\}$, according to Lemma C.2.1, we have

$$R(Y_1(i), (Y_1(l))_{l\in[n]\setminus\{1\}}) \geq R(Y_1(i), Y_1) - 1.$$

Since $R(X_1(i), X_1) = R(Y_1(i), Y_1)$ for any $i \in (V \setminus \{1\}) \cup U$, we have $R(Y_1(i), Y_1) = R(X_1(i), X_1)$ for any $i \in V \setminus \{1\}$. Then, it follows that

$$R(Y_1(i), (Y_1(l))_{l\in[n]\setminus\{1\}}) \geq R(Y_1(i), Y_1) - 1 = R(X_1(i), X_1) - 1, \text{for } i \in V \setminus \{1\}.$$

Next, notice that $X_1(1) < \ldots < X_1(t)$. Hence, for any $i \in V \setminus \{1\}$, we further have

$$R(X_1(i), X_1) - 1 > R(X_1(1), X_1) - 1 =^{(C.95)} R(X_1(1), (X_1(l))_{l\in[n]\setminus U}) - 1.$$

Thus, we have

$$R(Y_1(i), (Y_1(l))_{l\in[n]\setminus\{1\}}) > R(X_1(1), (X_1(l))_{l\in[n]\setminus U}) - 1, \text{ for any } i \in V \setminus \{1\}.$$

Then, for any $i \in V \setminus \{1\}$, it follows that

$$R(Y_1(i), (Y_1(l))_{l\in[n]\setminus\{1\}}) \geq R(X_1(1), (X_1(l))_{l\in[n]\setminus U}) =^{(C.95)} R(X_1(1), X) = \alpha.$$

This completes our proof. □

Lemma C.3.27. Following Lemma C.3.26, let Y_2 be an imputation of Y_1 for the index 1 such that $R(Y_2(1), Y_2) = R(X_1(1), X_1)$. Then, we have

$$R(Y_2(i), (Y_2(l))_{l\in\{1\}\cup([n]\setminus V)}) > R(Y_2(1), (Y_2(l))_{l\in[n]\setminus(V\setminus\{1\})}) = \alpha, \text{ for any } i \in U.$$

Proof. To start, since $Y_2(1)$ is an imputation of Y_1 for the index 1, we have

$$Y_2(i) = Y_1(i), \text{ for any } i \in [n] \setminus \{1\}. \tag{C.96}$$

Suppose $V \setminus \{1\} \neq \emptyset$. Then for any $i \in V \setminus \{1\}$, according to Lemma C.2.1, we have

$$R(Y_2(i), Y_2) \geq R(Y_2(i), (Y_2(l))_{l\in[n]\setminus\{1\}}).$$

According to (C.96), we have $Y_2(i) = Y_1(i)$ for any $i \in V\setminus\{1\}$, and $(Y_1(l))_{l\in[n]\setminus\{1\}} = (Y_2(l))_{l\in[n]\setminus\{1\}}$. Hence, we have

$$R(Y_2(i), Y_2) \geq R(Y_2(i), (Y_2(l))_{l\in[n]\setminus\{1\}}) = R(Y_1(i), (Y_1(l))_{l\in[n]\setminus\{1\}}), \text{ for } i \in V \setminus \{1\}.$$

According to Lemma C.3.26, we have

$$R(Y_1(i), (Y_1(l))_{l\in[n]\setminus\{1\}}) \geq R(X_1(1), X_1), \text{ for any } i \in V \setminus \{1\}.$$

Since $R(X_1(1), X_1) = R(Y_2(1), Y_2)$, we then have

$$R(Y_2(i), Y_2) \geq R(Y_2(1), Y_2), \text{ for any } i \in V \setminus \{1\}.$$

Since $Y_2 \in \mathcal{V}^n$ is a vector of distinct real values, we have for any $i \in V \setminus \{1\}$,

$$R(Y_2(i), Y_2) \neq R(Y_2(1), Y_2).$$

Thus, it follows that

$$R(Y_2(i), Y_2) > R(Y_2(1), Y_2) \implies Y_2(i) > Y_2(1), \text{ for any } i \in V \setminus \{1\}.$$

Subsequently, according to Lemma C.2.2, we have

$$R(Y_2(1), Y_2) = R(Y_2(1), (Y_2(l))_{l\in[n]\setminus(V\setminus\{1\})}).$$

In particular, the above equation still holds when $V\setminus\{1\} = \emptyset$. Further, notice that $R(Y_2(1), Y_2) = R(X_1(1), X_1)$, and according to Lemma C.3.26, $R(X_1(1), X_1) = \alpha$. Hence, we have

$$R(Y_2(1), Y_2) = R(Y_2(1), (Y_2(l))_{l\in[n]\setminus(V\setminus\{1\})}) = \alpha. \tag{C.97}$$

Next, according to Lemma C.2.1, for any $i \in U$,

$$R(Y_2(i), Y_2) \geq R(Y_2(i), (Y_2(l))_{l\in[n]\setminus V}).$$

Since $V = \{1, \ldots, t\}, U = \{t+1, \ldots, t+m\}$, where $t, m \geq 1$, it follows $[n] \setminus V \subset [n] \setminus \{1\}$ and $U \subset [n] \setminus \{1\}$. Thus, according to (C.96), we have $Y_2(i) = Y_1(i)$ for any $i \in U$, and $(Y_2(l))_{l\in[n]\setminus V} = (Y_1(l))_{l\in[n]\setminus V}$. Then, we have

$$R(Y_2(i), (Y_2(l))_{l\in[n]\setminus V}) = R(Y_1(i), (Y_1(l))_{l\in[n]\setminus V}), \text{ for any } i \in U.$$

Further, since $[n] \setminus V \subset [n] \setminus (V \setminus \{1\})$ and $U \subset [n] \setminus (V \setminus \{1\})$, then according to (C.92), we have

$Y_1(i) = Y(i)$ for any $i \in U$, and $(Y_1(l))_{l\in[n]\setminus V} = (Y(l))_{l\in[n]\setminus V}$. Hence, we have

$$R(Y_1(i), (Y_1(l))_{l\in[n]\setminus V}) = R(Y(i), (Y(l))_{l\in[n]\setminus V}), \text{ for any } i \in U.$$

Hence, we have

$$R(Y_2(i), Y_2) \geq R(Y(i), (Y(l))_{l\in[n]\setminus V}), \text{ for any } i \in U.$$

Further, since $Y(t+1) < \ldots < Y(t+m)$, for any $i \in U$, we have

$$R(Y_2(i), Y_2) \geq R(Y(i), (Y(l))_{l\in[n]\setminus V}) \geq R(Y(t+1), (Y(l))_{l\in[n]\setminus V}).$$

Notice that we have

$$R(Y(t+1), (Y(l))_{l\in\{1\}\cup([n]\setminus V)}) > R(X(1), (X(l))_{l\in[n]\setminus U}) =^{(\text{C.90})} \alpha.$$

Then, it follows that

$$\begin{aligned} & R(Y(t+1), (Y(l))_{l\in\{1\}\cup([n]\setminus V)}) \geq \alpha + 1 \\ \Longrightarrow\ & R(Y(t+1), (Y(l))_{l\in\{1\}\cup([n]\setminus V)}) - 1 \geq \alpha. \end{aligned}$$

According to Lemma C.2.1, we have

$$R(Y(t+1), (Y(l))_{l\in[n]\setminus V}) \geq R(Y(t+1), (Y(l))_{l\in\{1\}\cup([n]\setminus V)}) - 1.$$

Hence, we have

$$R(Y(t+1), (Y(l))_{l\in[n]\setminus V}) \geq \alpha.$$

Then, we have

$$R(Y_2(i), Y_2) \geq \alpha, \text{ for any } i \in U$$

According to (C.97), we have $R(Y_2(1), Y_2) = \alpha$. Since $Y_2 \in \mathcal{V}^n$ is a vector of distinct real values, then for any $i \in U$, we have

$$\begin{aligned} & R(Y_2(i), Y_2) \neq R(Y_2(1), Y_2) = \alpha \\ \Longrightarrow\ & R(Y_2(i), Y_2) > R(Y_2(1), Y_2) \\ \Longrightarrow\ & Y_2(i) > Y_2(1) \\ \Longrightarrow\ & R(Y_2(i), (Y_2(l))_{l\in\{1\}\cup([n]\setminus V)}) > R(Y_2(1), (Y_2(l))_{l\in[n]\setminus(V\setminus\{1\})}) =^{(\text{C.97})} \alpha. \end{aligned}$$

This completes our proof. □

Lemma C.3.28. Following Lemma C.3.26 and Lemma C.3.27, let X_3 and Y_3 be imputations of X_1, Y_2 for U and $V \setminus \{1\}$, respectively, such that $R(X_3(i), X_3) = R(Y_3(i), Y_3)$ for any $i \in (V \setminus \{1\}) \cup U$. Then, we have $R(X_3(i), X_3) = R(Y_3(i), Y_3)$, for any $i \in U \cup V$.

Proof. To start, since X_3 and Y_3 are imputations of X_1, Y_2 for U and $V \setminus \{1\}$, respectively, we have

$$X_3(i) = X_1(i), \text{ for any } i \in [n] \setminus U, \tag{C.98}$$

$$Y_3(i) = Y_2(i), \text{ for any } i \in [n] \setminus (V \setminus \{1\}). \tag{C.99}$$

Recall that $U = \{t+1, \ldots, t+m\}$, where $t, m \geq 1$. Hence, $1 \in [n] \setminus U$. Then according to (C.98), we have $X_3(1) = X_1(1)$. Notice that (C.98) also gives that $(X_3(l))_{l \in [n] \setminus U} = (X_1(l))_{l \in [n] \setminus U}$. Hence, we have

$$R(X_3(1), (X_3(l))_{l \in [n] \setminus U}) = R(X_1(1), (X_1(l))_{l \in [n] \setminus U}) =^{(\text{C.90})} \alpha.$$

Similarly, since $1 \in [n] \setminus (V \setminus \{1\})$, according to (C.99), we have $Y_3(1) = Y_2(1)$. Notice that $[n] \setminus (V \setminus \{1\} = \{1\}) \cup ([n] \setminus V)$. Then (C.99) also gives that $(Y_3(l))_{l \in \{1\} \cup ([n] \setminus V)} = (Y_2(l))_{l \in \{1\} \cup ([n] \setminus V)}$. Hence, we have

$$R(Y_3(1), (Y_3(l))_{l \in \{1\} \cup ([n] \setminus V)}) = R(Y_2(1), (Y_2(l))_{l \in \{1\} \cup ([n] \setminus V)}). \tag{C.100}$$

According to Lemma C.3.27, we have

$$R(Y_2(1), (Y_2(l))_{l \in \{1\} \cup ([n] \setminus V)}) = \alpha.$$

Hence

$$R(Y_3(1), (Y_3(l))_{l \in \{1\} \cup ([n] \setminus V)}) = \alpha.$$

Combining the above results, we have

$$R(X_3(1), (X_3(l))_{l \in [n] \setminus U}) = R(Y_3(1), (Y_3(l))_{l \in \{1\} \cup ([n] \setminus V)}) = \alpha. \tag{C.101}$$

Suppose $V \setminus \{1\} \neq \emptyset$, then since $R(X_3(i), X_3) = R(Y_3(i), Y_3)$ for any $i \in (V \setminus \{1\}) \cup U$, we have

$$R(Y_3(i), Y_3) = R(X_3(i), X_3), \text{ for any } i \in V \setminus \{1\}.$$

According to Lemma C.3.26, we have $X_1(1) < \ldots < X_1(t)$. Notice that from (C.98), we have $X_1(i) = X_3(i)$ for any $i \in V = \{1, \ldots, t\}$. Hence, we further have $X_3(1) < \ldots < X_3(t)$. Subsequently,

$$R(X_3(i), X_3) > R(X_3(1), X_3), \text{ for any } i \in V \setminus \{1\}.$$

According to Lemma C.2.1, we further have

$$R(X_3(1), X_3) \geq R(X_3(1), (X_3(l))_{l \in [n] \setminus U}) =^{(\text{C.101})} R(Y_3(1), (Y_3(l))_{l \in \{1\} \cup ([n] \setminus V)}).$$

Hence, we have

$$R(Y_3(i), Y_3) > R(Y_3(1), (Y_3(l))_{l\in\{1\}\cup([n]\setminus V)}), \text{ for any } i \in V \setminus \{1\}.$$

Subsequently, according to Lemma C.2.2, we have

$$R(Y_3(1), Y_3) = R(Y_3(1), (Y_3(l))_{l\in[n]\setminus(V\setminus\{1\})}) = \alpha. \tag{C.102}$$

In particular, the above equations still hold when $V \setminus \{1\} = \emptyset$.

Since $R(X_3(i), X_3) = R(Y_3(i), Y_3)$ for any $i \in (V \setminus \{1\}) \cup U$, then for any $i \in U$, we have

$$R(X_3(i), X_3) = R(Y_3(i), Y_3).$$

Then, according to Lemma C.2.1, for any $i \in U$, we have

$$R(Y_3(i), Y_3) \geq R(Y_3(i), (Y_3(l))_{l\in\{1\}\cup([n]\setminus V)}).$$

Hence, for any $i \in U$, it follows that

$$R(X_3(i), X_3) \geq R(Y_3(i), (Y_3(l))_{l\in\{1\}\cup([n]\setminus V)}).$$

Since $U \subset [n] \setminus V \subset [n] \setminus (V \setminus \{1\})$, then according to (C.99), we have $Y_3(i) = Y_2(i)$ for any $i \in U$. Notice that $[n] \setminus (V \setminus \{1\}) = \{1\} \cup ([n] \setminus V)$. Then (C.99) also gives that $(Y_3(l))_{l\in\{1\}\cup([n]\setminus V)} = (Y_2(l))_{l\in\{1\}\cup([n]\setminus V)}$. Hence, we have

$$R(Y_3(i), (Y_3(l))_{l\in\{1\}\cup([n]\setminus V)}) = R(Y_2(i), (Y_2(l))_{l\in\{1\}\cup([n]\setminus V)}), \text{ for any } i \in U.$$

Combining the above results, for any $i \in U$, we have

$$R(X_3(i), X_3) \geq R(Y_2(i), (Y_2(l))_{l\in\{1\}\cup([n]\setminus V)}).$$

According to Lemma C.3.27, we further have

$$R(Y_2(i), (Y_2(l))_{l\in\{1\}\cup([n]\setminus V)}) > \alpha =^{(\text{C.101})} R(X_3(1), (X_3(l))_{l\in[n]\setminus U}).$$

Hence, we have

$$R(X_3(i), X_3) > R(X_3(1), (X_3(l))_{l\in[n]\setminus U}), \text{ for any } i \in U.$$

Subsequently, according to Lemma C.2.2, we have

$$R(X_3(1), X_3) = R(X_3(1), (X_3(l))_{l\in[n]\setminus U}) = \alpha = R(Y_3(1), Y_3).$$

Combining this result with (C.99), we have

$$R(X_3(i), X_3) = R(Y_3(i), Y_3), \text{ for any } i \in U \cup V.$$

This completes our proof. □

We are now ready to prove Theorem C.3.29.

Theorem C.3.29. Suppose $X, Y \in \mathcal{V}^n$, and $U, V \subset [n]$ are disjoint subsets of indices such that $U \cap V = \emptyset$. Suppose $X_*, Y_* \in \mathcal{V}^n$ satisfy both Conditions C.3.10 and C.3.11. In other words, $X_*, Y_* \in \mathcal{V}^n$ are imputations of X, Y for indices U and V, and $R(X_*(i), X_*) = R(Y_*(i), Y_*)$, for any $i \in U \cup V$. Then, we have $D(X_*, Y_*) \leq D(X, Y)$. Furthermore, for any other imputations $X'_*, Y'_* \in \mathcal{V}^n$ of X, Y for indices U, V, respectively, we have $D(X_*, Y_*) \leq D(X', Y')$.

Proof. Notice that there are two statements in Theorem C.3.29:

(1) : Suppose $X, Y \in \mathcal{V}^n$, and $U, V \subset [n]$ are disjoint subsets of indices such that $U \cap V = \emptyset$. Then if $X_*, Y_* \in \mathcal{V}^n$ are imputations of X, Y for indices U and V, respectively, and $R(X_*(i), X_*) = R(Y_*(i), Y_*)$, for any $i \in U \cup V$, we have $D(X_*, Y_*) \leq D(X, Y)$.

(2) : Suppose $X, Y \in \mathcal{V}^n$, and $U, V \subset [n]$ are disjoint subsets of indices such that $U \cap V = \emptyset$. Suppose $X_*, Y_* \in \mathcal{V}^n$ are imputations of X, Y for indices U and V, respectively, and $R(X_*(i), X_*) = R(Y_*(i), Y_*)$, for any $i \in U \cup V$. Then for any other imputations $X'_*, Y'_* \in \mathcal{V}^n$ of X, Y for indices U, V, respectively, we have $D(X_*, Y_*) \leq D(X', Y')$.

Below, we first show that the statement (1) is true. Then we prove the statement (2) using the statement (1).

First, we show the statement (1) is true. To begin with, let us consider the case when $|U| + |V| = n$. Since $U \cap V = \emptyset$, then $|U| + |V| = n$ means $U \cup V = [n]$. Since $R(X_*(i), X_*) = R(Y_*(i), Y_*)$, for any $i \in U \cup V$, we have $R(X_*, X_*) = R(Y_*, Y_*)$. Hence $D(X_*, Y_*) = 0$. According to the definition of Spearman's footrule, we have $D(X, Y) \geq 0 = D(X_*, Y_*)$, which proves the statement (1) when $|U| + |V| = n$.

When $|U| + |V| = 0$, then according to the definition of imputations, we have $X = X_*$ and $Y = Y_*$. Hence we have $D(X_*, Y_*) = D(X, Y)$. This proves statement (1) when $|U| + |V| = 0$.

Then, we only need to prove statement (1) when $n - 1 \geq |U| + |V| \geq 1$. For any given $n \in \mathbb{N}$, let $P(k)$ be the statement of the statement (1) when $|U| + |V| = k$. We prove $P(k)$ holds for any $k \in \{1, \ldots, n-1\}$ by induction on k.

Base Case: We show $P(1)$ holds. Suppose $|U| + |V| = 1$. Then we have either $|U| = 0, |V| = 1$ or $|U| = 1, |V| = 0$. For both cases, the statement $P(1)$ is true according to Theorem C.3.9.

Induction Step: We show the implication $P(k) \implies P(k+1)$ for any $k \in \{1, \ldots, n-2\}$.

Suppose $|U| + |V| = k + 1$. Then, we have either

$$\begin{aligned} &\text{case (I) : } |U| = k+1, |V| = 0 \text{ or } |U| = 0, |V| = k+1, \\ \text{or } &\text{case (II) : } |U| > 0, |V| > 0, |U| + |V| = k+1, \end{aligned}$$

is true.

Suppose the case (I) : $|U| = k+1, |V| = 0$ or $|U| = 0, |V| = k+1$ is true. Then the statement $P(k+1)$ is true according to Theorem C.3.9.

Suppose the case (II) : $|U| > 0, |V| > 0, |U| + |V| = k+1$ is true. Without loss of generality, let us assume (after relabelling) $V = \{1, \ldots, t\}$, $U = \{t+1, \ldots, t+m\}$, $X(1) < \ldots < X(t)$, and $Y(t+1) < Y(t+2) < \ldots < Y(t+m)$.

Notice that for $R(X(1),(X(l))_{l\in[n]\setminus U})$ and $R(Y(t+1),(Y(l))_{l\in[n]\setminus V})$, we have either

$$\begin{aligned} &R(X(1),(X(l))_{l\in[n]\setminus U}) \geq R(Y(t+1),(Y(l))_{l\in[n]\setminus V}),\\ \text{or } &R(X(1),(X(l))_{l\in[n]\setminus U}) \leq R(Y(t+1),(Y(l))_{l\in[n]\setminus V}), \end{aligned}$$

is true. Let us assume $R(X(1),(X(l))_{l\in[n]\setminus U}) \leq R(Y(t+1),(Y(l))_{l\in[n]\setminus V})$ is true. However, when $R(X(1),(X(l))_{l\in[n]\setminus U}) \geq R(Y(t+1),(Y(l))_{l\in[n]\setminus V})$ holds, we can switch the labels between X and Y, and relabel the relevant components of the data.

Let us denote

$$\alpha = R(X(1),(X(l))_{l\in[n]\setminus U}). \tag{C.103}$$

Then, since $\alpha = R(X(1),(X(l))_{l\in[n]\setminus U}) \leq R(Y(t+1),(Y(l))_{l\in[n]\setminus V})$, we have either

$$\begin{aligned} &\text{case (i)} : \alpha = R(Y(t+1),(Y(l))_{l\in[n]\setminus V}),\ \text{and } Y(1) > Y(t+1),\\ &\text{case (ii)} : \alpha = R(Y(t+1),(Y(l))_{l\in[n]\setminus V}),\ \text{and } Y(1) < Y(t+1),\\ \text{or } &\text{case (iii)} : \alpha < R(Y(t+1),(Y(l))_{l\in[n]\setminus V}), \end{aligned}$$

is true. In the following, we are going to consider the three cases separately.

Suppose the case (i) is true. In other words,

$$\alpha = R(Y(t+1),(Y(l))_{l\in[n]\setminus V}), \tag{C.104}$$

$$\text{and } Y(1) > Y(t+1). \tag{C.105}$$

Then, according to Proposition C.3.13, there exist $X_1, Y_1 \in \mathcal{V}^n$ such that

$$X_1(i) = X(i),\ \text{for any } i \in [n]\setminus U, \tag{C.106}$$

$$Y_1(i) = Y(i),\ \text{for any } i \in [n]\setminus (V\setminus\{1\}), \tag{C.107}$$

$$R(X_1(i),X_1) = R(Y_1(i),Y_1) \text{ for any } i \in U\cup(V\setminus\{1\}). \tag{C.108}$$

In other words, X_1, Y_1 are imputations of X, and Y for indices U and $V\setminus\{1\}$, respectively, such that the equation (C.108) holds.

Since U and V are disjoint such that $U\cap V = \emptyset$, then $U\cap(V\setminus\{1\}) = \emptyset$, i.e. U and $V\setminus\{1\}$ are also disjoint. Notice that $|U| + |V\setminus\{1\}| = m + t - 1 = k$. Then, since $P(k)$ is true, we have

$$D(X_1,Y_1) \leq D(X,Y). \tag{C.109}$$

Next, applying Proposition C.3.13 again, we can find $X_2, Y_2 \in \mathcal{V}^n$ such that

$$X_2(i) = X_1(i),\ \text{for any } i \in [n]\setminus (U\setminus\{t+1\}), \tag{C.110}$$

$$Y_2(i) = Y_1(i),\ \text{for any } i \in [n]\setminus V, \tag{C.111}$$

$$R(X_2(i),X_2) = R(Y_2(i),Y_2) \text{ for any } i \in (U\setminus\{t+1\})\cup V. \tag{C.112}$$

In other words, X_2 and Y_2 are imputations of X_1, Y_1 for indices $U\setminus\{t+1\}$ and V, respectively,

such that (C.112) is true.

Notice that U and V are disjoint, i.e. $U \cap V = \emptyset$. Hence, $(U \setminus \{t+1\}) \cap V = \emptyset$. Next, since $|U \setminus \{t+1\}| + |V| = m + t - 1 = k$, we can apply $P(k)$ and get

$$D\left(X_2, Y_2\right) \leq D\left(X_1, Y_1\right). \tag{C.113}$$

According to Lemma C.3.25, we have

$$R(X_2(i), X_2) = R(Y_2(i), Y_2), \text{for any } i \in U \cup V.$$

Notice that

$$\begin{aligned} X_2(i) &=^{(\text{C.110})} X_1(i) =^{(\text{C.106})} X(i) \quad \text{for any } i \in [n] \setminus U, \\ \text{and } Y_2(i) &=^{(\text{C.111})} Y_1(i) =^{(\text{C.107})} Y(i) \quad \text{for any } i \in [n] \setminus V. \end{aligned}$$

In other words, $X_2, Y_2 \in \mathcal{V}^n$ are imputations of X, Y for indices U and V, respectively, such that $R(X_2(i), X_2) = R(Y_2(i), Y_2)$, for any $i \in U \cup V$. Since $X_*, Y_* \in \mathcal{V}^n$ are imputations of X, Y for indices U and V, respectively, such that $R(X_*(i), X_*) = R(Y_*(i), Y_*)$, for any $i \in U \cup V$, then according to Proposition C.3.22, we have $D\left(X_2, Y_2\right) = D\left(X_*, Y_*\right)$. Thus,

$$D\left(X_*, Y_*\right) = D\left(X_2, Y_2\right) \leq^{(\text{C.113})} D\left(X_1, Y_1\right) \leq^{(\text{C.109})} D\left(X, Y\right),$$

which proves $P(k+1)$ when the case (i) holds.

Suppose the case (ii) or the case (iii) holds. In other words, we have either

$$\alpha = R(Y(t+1), (Y(l))_{l \in [n] \setminus V}), \text{ and } Y(1) < Y(t+1), \tag{C.114}$$

$$\text{or } \alpha < R(Y(t+1), (Y(l))_{l \in [n] \setminus V}). \tag{C.115}$$

According to the definition of rank, we have

$$\begin{aligned} &R(Y(t+1), (Y(l))_{l \in \{1\} \cup ([n] \setminus V)}) \\ &= \sum_{i \in \{1\} \cup ([n] \setminus V)} \mathrm{I}\left(Y(i) \leq Y(t+1)\right) \\ &= \mathrm{I}\left(Y(1) \leq Y(t+1)\right) + \sum_{i \in [n] \setminus V} \mathrm{I}\left(Y(i) \leq Y(t+1)\right) \\ &= \mathrm{I}\left(Y(1) \leq Y(t+1)\right) + R(Y(t+1), (Y(l))_{l \in [n] \setminus V}). \end{aligned}$$

Notice that $\mathrm{I}\left(Y(1) \leq Y(t+1)\right) \geq 0$. Hence, we have

$$R(Y(t+1), (Y(l))_{l \in \{1\} \cup ([n] \setminus V)}) \geq R(Y(t+1), (Y(l))_{l \in [n] \setminus V}).$$

Next, suppose (C.115) is true. We further have

$$\begin{aligned} R(Y(t+1), (Y(l))_{l \in \{1\} \cup ([n] \setminus V)}) &\geq R(Y(t+1), (Y(l))_{l \in [n] \setminus V}) \\ &> \alpha =^{(\text{C.103})} R(X(1), (X(l))_{l \in [n] \setminus U}). \end{aligned}$$

However, suppose (C.114) is true. Then, we have

$$\begin{aligned}
&R(Y(t+1),(Y(l))_{l\in\{1\}\cup([n]\setminus V)})\\
&=\mathrm{I}\left(Y(1)\leq Y(t+1)\right)+R(Y(t+1),(Y(l))_{l\in[n]\setminus V})\\
&=1+\alpha>\alpha=^{(\mathrm{C}.103)}R(X(1),(X(l))_{l\in[n]\setminus U}).
\end{aligned}$$

Hence, when the case (ii) or the case (iii) is true, we have

$$R(Y(t+1),(Y(l))_{l\in\{1\}\cup([n]\setminus V)})>\alpha=R(X(1),(X(l))_{l\in[n]\setminus U}). \tag{C.116}$$

According to Proposition C.3.13, there exist $X_3, Y_3 \in \mathcal{V}^n$ such that

$$X_3(i)=X(i),\ \text{for any } i\in[n]\setminus U, \tag{C.117}$$
$$Y_3(i)=Y(i),\ \text{for any } i\in[n]\setminus(V\setminus\{1\}), \tag{C.118}$$
$$R(X_3(i),X_3)=R(Y_3(i),Y_3)\ \text{for any } i\in(V\setminus\{1\})\cup U. \tag{C.119}$$

That is, X_3 and Y_3 are imputations of X and Y for U and $V\setminus\{1\}$ such that (C.119) is true.

Notice that U and V are disjoint, i.e. $U\cap V=\emptyset$. Hence, $(U\setminus\{t+1\})\cap V=\emptyset$. Next, since $|U\setminus\{t+1\}|+|V|=m+t-1=k$, we can apply $P(k)$ and get

$$D\left(X_3,Y_3\right)\leq D\left(X,Y\right). \tag{C.120}$$

Notice that (C.116) is true. Then, according to Lemma C.3.26, when $V\setminus\{1\}\neq\emptyset$, for any $i\in V\setminus\{1\}$, we have

$$R(Y_3(i),(Y_3(l))_{l\in[n]\setminus\{1\}})\geq R(X_3(1),(X_3(l))_{l\in[n]\setminus U})=R(X_3(1),X_3)=\alpha. \tag{C.121}$$

Next, according to Lemma C.3.6, there exists $Y_4\in\mathcal{V}^n$ such that

$$Y_4(i)=Y_3(i),\ \text{for any } i\in[n]\setminus\{1\}, \tag{C.122}$$
$$R(Y_4(1),Y_4)=R(X_3(1),X_3). \tag{C.123}$$

In other words, Y_4 is an imputation of Y_3 of the index 1 such that (C.123) is true. Then, according to Theorem C.3.9, we have

$$D\left(X_3,Y_4\right)\leq D\left(X_3,Y_3\right). \tag{C.124}$$

According to Lemma C.3.27, for any $i\in U$, we have

$$R(Y_4(i),(Y_4(l))_{l\in\{1\}\cup([n]\setminus V)})>R(Y_4(1),(Y_4(l))_{l\in[n]\setminus(V\setminus\{1\})})=\alpha. \tag{C.125}$$

Next, according to Proposition C.3.13, there exist $X_5, Y_5 \in \mathcal{V}^n$ such that

$$X_5(i) = X_3(i), \text{ for any } i \in [n] \setminus U, \tag{C.126}$$
$$Y_5(i) = Y_4(i), \text{ for any } i \in [n] \setminus (V \setminus \{1\}), \tag{C.127}$$
$$R(X_5(i), X_5) = R(Y_5(i), Y_5) \text{ for any } i \in (V \setminus \{1\}) \cup U. \tag{C.128}$$

In other words, X_5 and Y_5 are imputations of X_3, Y_4 for U and $V \setminus \{1\}$, respectively, such that (C.128) is true.

Since U and V are disjoint, U and $V \setminus \{1\}$ are also disjoint. Notice that $|U| + |V \setminus \{1\}| = m + t - 1 = k$. Then, according to $P(k)$, we have

$$D(X_5, Y_5) \leq D(X_3, Y_4). \tag{C.129}$$

According to Lemma C.3.28, we have

$$R(X_5(i), X_5) = R(Y_5(i), Y_5), \text{ for any } i \in U \cup V.$$

Combining (C.126) and (C.117), we have

$$X_5(i) =^{(\text{C.126})} X_3(i) =^{(\text{C.117})} X(i), \text{ for any } i \in [n] \setminus U.$$

Combining (C.127), (C.122), and (C.118), we have

$$Y_5(i) =^{(\text{C.127})} Y_4(i) =^{(\text{C.122})} Y_3(i) =^{(\text{C.118})} Y(i), \text{ for any } i \in [n] \setminus V.$$

In other words, $X_5, Y_5 \in \mathcal{V}^n$ are imputations of X, Y for indices U and V, respectively, such that $R(X_5(i), X_5) = R(Y_5(i), Y_5)$, for any $i \in U \cup V$. Since $X_*, Y_* \in \mathcal{V}^n$ are imputations of X, Y for indices U and V, respectively, such that $R(X_*(i), X_*) = R(Y_*(i), Y_*)$, for any $i \in U \cup V$, then according to Proposition C.3.22, we have $D(X_5, Y_5) = D(X_*, Y_*)$. Thus,

$$D(X_*, Y_*) = D(X_5, Y_5) \leq^{(\text{C.129})} D(X_3, Y_4) \leq^{(\text{C.124})} D(X_3, Y_3) \leq^{(\text{C.120})} D(X, Y).$$

which proves $P(k+1)$ when the case (ii) or the case (iii) holds. This completes our proof for the statement (1).

Next, we prove the statement (2) using the statement (1). According to Proposition C.3.13, we can find $X'_*, Y'_* \in \mathcal{V}^n$ such that X'_*, Y'_* are imputations of X', Y' for indices U, and V, respectively, and $R(X'_*(i), X'_*) = R(Y'_*(i), Y'_*)$ for any $i \in U \cup V$. Then, according to the statement (1), we have

$$D(X'_*, Y'_*) \leq D(X', Y').$$

Next, since $X_*, Y_* \in \mathcal{V}^n$ are imputation of X, Y for U and V, respectively, then we have

$$X_*(i) = X(i), \text{ for any } i \in [n] \setminus U,$$
$$Y_*(i) = Y(i), \text{ for any } i \in [n] \setminus V.$$

Further, since $X'_*, Y'_* \in \mathcal{V}^n$ such that X'_*, Y'_* are imputations of X', Y' for indices U, and V, respectively, we have

$$X'_*(i) = X_*(i) = X(i), \text{ for any } i \in [n] \setminus U,$$
$$Y'_*(i) = Y_*(i) = Y(i), \text{ for any } i \in [n] \setminus V.$$

In other words, $X'_*, Y'_* \in \mathcal{V}^n$ are also imputations of X, Y for indices U, and V, respectively. Notice that for X_* and Y_*, we have $R(X_*(i), X_*) = R(Y_*(i), Y_*)$, for any $i \in U \cup V$. Also, for X'_* and Y'_*, we have $R(X'_*(i), X'_*) = R(Y'_*(i), Y'_*)$ for any $i \in U \cup V$. Then, according to Proposition C.3.23, we have $D(X'_*, Y'_*) = D(X_*, Y_*)$. Hence, we have

$$D(X_*, Y_*) = D\left(X'_*, Y'_*\right) \leq D\left(X', Y'\right).$$

This proves the statement (2) and completes our proof. □

C.3.3. Lower bound under missing case III

Given two vectors $X, Y \in \mathcal{V}^n$, recall that in missing case III, for any pair of components $(X(i), Y(i))$ at index $i \in [n]$, the two values are either both observed or both missing. This case for Spearman's footrule has been previously considered in the literature, and the following result is proved in Theorem 1 in Papaioannou and Loukas (1984).

Theorem C.3.30 (Papaioannou and Loukas (1984))**.** Suppose $X, Y \in \mathcal{V}^n$ and $W \subset [n]$. For the subvectors, $X' = (X(i))_{i \in [n] \setminus W}$ and $Y' = (Y(i))_{i \in [n] \setminus W}$, it can be shown that $D(X', Y') \leq D(X, Y)$.

While this result provides a lower bound of Spearman's footrule, it does not directly prove that this bound is the minimum possible value of the Spearman's footrule. The following result fills this gap by identifying the conditions under which this lower bound is tight.

Proposition C.3.31. Suppose $X, Y \in \mathcal{V}^n$ and $W \subset [n]$. Define the subvectors $X' = (X(i))_{i \in [n] \setminus W}$ and $Y' = (Y(i))_{i \in [n] \setminus W}$. If the following three conditions all hold,

$$\text{(i) } R(X(i), X) = R(Y(i), Y), \text{ for all } i \in W,$$
$$\text{(ii) } \min_{j \in W} X(j) > \max_{j \in [n] \setminus W} X(j),$$
$$\text{(iii) } \min_{j \in W} Y(j) > \max_{j \in [n] \setminus W} Y(j),$$

then $D(X, Y) = D(X', Y')$.

Proof. To start, according to the definition of Spearman's footrule, we have

$$D(X, Y) = \sum_{i \in W} \left| R(X(i), X) - R(Y(i), Y) \right| + \sum_{i \in [n] \setminus W} \left| R(X(i), X) - R(Y(i), Y) \right|.$$

According to the condition (i), for any $i \in W$, we have $R(X(i), X) = R(Y(i), Y)$. Then,

$$\sum_{i \in W} \left| R(X(i), X) - R(Y(i), Y) \right| = 0,$$

which follows that

$$D\left(X, Y\right) = \sum_{i \in [n] \setminus W} \left| R(X(i), X) - R(Y(i), Y) \right|.$$

Next, according to the condition (ii), we have $\min(X(l))_{l \in W} > \max(X(l))_{l \in [n] \setminus W}$. Then, according to Lemma C.2.2, for any $i \in [n] \setminus W$, we have

$$R(X(i), X) = R(X(i), X').$$

Similarly, according to the condition (iii), we have $\min(Y(l))_{l \in W} > \max(Y(l))_{l \in [n] \setminus W}$. Then, according to Lemma C.2.2, for any $i \in [n] \setminus W$, we have

$$R(Y(i), Y) = R(Y(i), Y').$$

Subsequently, we have

$$\begin{aligned} D\left(X, Y\right) &= \sum_{i \in [n] \setminus W} \left| R(X(i), X) - R(Y(i), Y) \right| \\ &= \sum_{i \in [n] \setminus W} \left| R(X(i), X') - R(Y(i), Y') \right| = D\left(X', Y'\right). \end{aligned}$$

This completes our proof. □

C.3.4. Lower bound under general missingness case

We are now ready to provide the lower bound of Spearman's footrule in the general case, where the missingness pattern of data is a combination of missing case I, II, and III. The previous results can be combined to give the following result.

Theorem C.3.32. Suppose $X, Y \in \mathcal{V}^n$ and U, V, $W \subset [n]$ are pairwise disjoint subsets. Suppose $X_*, Y_* \in \mathcal{V}^n$ are imputations of X, Y for indices $U \cup W$, and $V \cup W$, respectively. If the following three conditions all hold

$$\begin{aligned} &\text{(i) } R(X_*(i), X) = R(Y_*(i), Y), \text{ for all } i \in U \cup V \cup W, \\ &\text{(ii) } \min_{j \in W} X_*(j) > \max_{j \in [n] \setminus W} X_*(j), \\ &\text{(iii) } \min_{j \in W} Y_*(j) > \max_{j \in [n] \setminus W} Y_*(j), \end{aligned}$$

then

$$D\left(X_*, Y_*\right) \leq D\left(X, Y\right),$$

and for any other imputations $X', Y' \in \mathcal{V}^n$ of X, Y for indices $U \cup W$, $V \cup W$, respectively,

$$D\left(X_*, Y_*\right) \leq D\left(X', Y'\right).$$

Proof. Notice that there are two statements in Theorem C.3.32:

(1) : Suppose $X, Y \in \mathcal{V}^n$, and $U, V, W \subset [n]$ are pairwise disjoint subsets of indices. If $X_*, Y_* \in \mathcal{V}^n$ are imputations of X, Y for indices U and V, respectively, and condition (i), (ii) and (iii) all hold, then we have $D\left(X_*, Y_*\right) \leq D\left(X, Y\right)$.

(2) : Suppose $X, Y \in \mathcal{V}^n$, and $U, V, W \subset [n]$ are pairwise disjoint subsets of indices. If $X_*, Y_* \in \mathcal{V}^n$ are imputations of X, Y for indices U and V, respectively, and condition (i), (ii) and (iii) all hold, then for any other imputations $X'_*, Y'_* \in \mathcal{V}^n$ of X, Y for indices U, V, respectively, we have $D\left(X_*, Y_*\right) \leq D\left(X', Y'\right)$.

Below, we first show the statement (1) is true. Then we prove the statement (2) using the statement (1).

First, we show the statement (1) is true. Suppose $W = \emptyset$ and $U \cup V = \emptyset$. Then according to the definition of imputations, we have $X_* = X$ and $Y_* = Y$. Hence, we have $D\left(X_*, Y_*\right) = D\left(X, Y\right)$, which proves the statement (1) when $W = \emptyset$ and $U \cup V = \emptyset$.

Suppose $W = \emptyset$ and $U \cup V \neq \emptyset$. Then $X_*, Y_* \in \mathcal{V}^n$ are imputations of X, Y for indices U, and V, respectively, such that $R(X_*(i), X) = R(Y_*(i), Y),$ for any $i \in U \cup V \cup W$. Then, according to Theorem C.3.29, we have $D\left(X_*, Y_*\right) \leq D\left(X, Y\right)$, which proves the statement (1) when $W = \emptyset$ and $U \cup V \neq \emptyset$.

Suppose $U \cup V = \emptyset$ and $W \neq \emptyset$. Then since (1), (2) and (3) all hold, we have $X_*, Y_* \in \mathcal{V}^n$ are imputations of X, Y for indices W such that

$$\begin{aligned}
&R(X_*(i), X) = R(Y_*(i), Y), \text{ for any } i \in W, \\
&\min(X_*(l))_{l \in W} > \max(X_*(l))_{l \in [n] \setminus W}, \\
&\min(Y_*(l))_{l \in W} > \max(Y_*(l))_{l \in [n] \setminus W}.
\end{aligned}$$

Denote $X' = (X(l))_{l \in [n] \setminus W}$, and $Y' = (Y(l))_{l \in [n] \setminus W}$. Then, according to Proposition C.3.31, we have

$$D\left(X_*, Y_*\right) = D\left(X', Y'\right).$$

Notice that according to Theorem C.3.31, we have

$$D\left(X', Y'\right) \leq D\left(X, Y\right).$$

Hence, we have $D\left(X_*, Y_*\right) \leq D\left(X, Y\right)$, which proves the statement (1) when $W \neq \emptyset$ and $U \cup V = \emptyset$.

Next, suppose $W \neq \emptyset$ and $U \cup V \neq \emptyset$. Denote $a = |W|$. Without loss of generality, let us assume (after relabelling) $W = \{n - a + 1, \ldots, n\}$. Denote $n' = n - a$. Then we have $[n] \setminus W = [n']$. Denote $X' = (X(l))_{l \in [n] \setminus W}, Y' = (Y(l))_{l \in [n] \setminus W}, X'_* = (X_*(l))_{l \in [n] \setminus W}$, and $Y'_* = (Y_*(l))_{l \in [n] \setminus W}$. Then, we have $X', Y', X'_*, Y'_* \in \mathcal{V}^{n'}$.

Since $X_*, Y_* \in \mathcal{V}^n$ are imputations of X, Y for indices $U \cup W$, and $V \cup W$, respectively, then we have

$$\begin{aligned} X_*(i) = X(i), \text{ for any } i \in [n] \setminus (U \cup W), \\ \text{and } Y_*(i) = Y(i), \text{ for any } i \in [n] \setminus (V \cup W). \end{aligned}$$

Notice that $[n] \setminus (U \cup W) = ([n] \setminus W) \setminus U$, and $[n] \setminus (V \cup W) = ([n] \setminus V) \setminus U$. Hence, we have

$$\begin{aligned} X_*(i) = X(i), \text{ for any } i \in ([n] \setminus W) \setminus U, \\ \text{and } Y_*(i) = Y(i), \text{ for any } i \in ([n] \setminus W) \setminus V. \end{aligned}$$

Therefore, X'_*, Y'_* are imputations of X', Y' for induces $U, V \subset [n] \setminus W$, respectively.

Since the condition (ii) is true, then according to Lemma C.2.2, for any $i \in U \cup V$, we have

$$R(X_*(i), X_*) = R(X_*(i), (X_*(l))_{l \in [n] \setminus W}) = R(X_*(i), X'_*).$$

Similarly, since the condition (iii) is true, then according to Lemma C.2.2, for any $i \in U \cup V$, we have

$$R(Y_*(i), Y_*) = R(Y_*(i), (Y_*(l))_{l \in [n] \setminus W}) = R(Y_*(i), Y'_*).$$

According to the condition (i), for any $i \in U \cup V$, we have

$$R(X_*(i), X_*) = R(Y_*(i), Y_*).$$

Hence, for any $i \in U \cup V$, we have

$$R(X_*(i), X'_*) = R(Y_*(i), Y'_*). \tag{C.130}$$

Since X'_*, Y'_* are imputations of X', Y' for indices $U, V \subset [n']$, respectively, then when $n' \geq 2$, according to Theorem C.3.29, we have

$$D\left(X'_*, Y'_*\right) \leq D\left(X', Y'\right).$$

However, when $n' = 1$, since $U \cup V \neq \emptyset$, and $U \cup V \subset [n] \setminus W = [n']$, we have $U \cup V = [n']$. Then according to condition (ii), we have $D\left(X'_*, Y'_*\right) = 0$. Hence, it also follows that

$$D\left(X'_*, Y'_*\right) \leq D\left(X', Y'\right).$$

Next, notice that according to Proposition C.3.31, we have

$$D\left(X_*, Y_*\right) = D\left(X'_*, Y'_*\right),$$

and according to Theorem C.3.31, we have

$$D\left(X', Y'\right) \leq D\left(X, Y\right).$$

Hence, we have

$$D\left(X_*, Y_*\right) = D\left(X'_*, Y'_*\right) \leq D\left(X', Y'\right) \leq D\left(X, Y\right).$$

This completes our proof for the statement (1).

Next, we prove the statement (2) is true using the statement (1). Suppose $W = \emptyset$ and $U \cup V = \emptyset$. Then according to the definition of imputations, we have $X_* = X_1 = X$ and $Y_* = Y_1 = Y$. Hence, we have $D\left(X_*, Y_*\right) = D\left(X_1, Y_1\right)$. This completes statement (2) when $W = \emptyset$ and $U \cup V = \emptyset$.

Suppose $W = \emptyset$ and $U \cup V \neq \emptyset$. Then $X_*, Y_* \in \mathcal{V}^n$ are imputations of X, Y for indices U, and V, respectively, such that $R(X_*(i), X_*) = R(Y_*(i), Y_*)$, for any $i \in U \cup V$. Meanwhile, $X_1, Y_1 \in \mathcal{V}^n$ are imputations of of X, Y for indices U, V. Then, according to Theorem C.3.29, we have $D\left(X_*, Y_*\right) \leq D\left(X_1, Y_1\right)$. This completes the statement (2) when $W = \emptyset$ and $U \cup V \neq \emptyset$.

Suppose $W \neq \emptyset$ and $U \cup V = \emptyset$. Let us denote $X' = (X(l))_{l \in [n] \setminus W}$, and $Y' = (Y(l))_{l \in [n] \setminus W}$. Then, according to Proposition C.3.31, we have

$$D\left(X_*, Y_*\right) = D\left(X', Y'\right).$$

Next, let us denote $X'_1 = (X'_1(l))_{l \in [n] \setminus W}$, and $Y'_1 = (Y'_1(l))_{l \in [n] \setminus W}$. Then, according to Theorem C.3.31, we have

$$D\left(X'_1, Y'_1\right) \leq D\left(X_1, Y_1\right).$$

Since $X_1, Y_1 \in \mathcal{V}^n$ are imputations of X, Y for indices W, we have $X'_1 = X'$ and $Y'_1 = Y'$. Thus,

$$D\left(X', Y'\right) = D\left(X'_1, Y'_1\right).$$

Combining the above results, we have

$$D\left(X_1, Y_1\right) \geq D\left(X'_1, Y'_1\right) = D\left(X', Y'\right) = D\left(X_*, Y_*\right).$$

This completes our proof when $W \neq \emptyset$ and $U \cup V = \emptyset$.

Next, suppose $W \neq \emptyset$ and $U \cup V \neq \emptyset$. Denote $a = |W|$. Without loss of generality, let us assume (after relabelling) $W = \{n - a + 1, \ldots, n\}$. Denote $n' = n - a$. Then we have $[n] \setminus W = [n']$. Denote $X' = (X(l))_{l \in [n] \setminus W}$, $Y' = (Y(l))_{l \in [n] \setminus W}$, $X'_1 = (X'_1(l))_{l \in [n] \setminus W}$, $Y'_1 = (Y'_1(l))_{l \in [n] \setminus W}$, $X'_* = (X_*(l))_{l \in [n] \setminus W}$, $Y'_* = (Y_*(l))_{l \in [n] \setminus W}$. Then, we have $X', Y', X'_1, Y'_1, X'_*, Y'_* \in \mathcal{V}^{n'}$.

Let us first show $D\left(X_1, Y_1\right) \geq D\left(X_*, Y_*\right)$ when $n' = 1$. Suppose $n' = 1$. Then $X'_* = X_*(1)$ and $Y'_* = Y_*(1)$. According to the definition of rank, we have $R(X'_*, X'_*) = R(Y'_*, Y'_*) = 1$. Hence, $D\left(X'_*, Y'_*\right) = 0$.

Next, since the conditions (i), (ii) and (iii) all hold. Then according to Proposition C.3.31, we have $D\left(X_*, Y_*\right) = D\left(X'_*, Y'_*\right) = 0$. Hence, we have $D\left(X_1, Y_1\right) \geq 0 = D\left(X_*, Y_*\right)$.

In the following, we show $D\left(X_1, Y_1\right) \geq D\left(X_*, Y_*\right)$ is true when $n' \geq 2$.

First, we show that X'_* and Y'_* are imputations of X'_1, Y'_1 for U and V, respectively, such that $R(X'_*(i), X'_*) = R(Y'_*(i), Y'_*)$, for any $i \in U \cup V$.

Notice that $X_*, Y_* \in \mathcal{V}^n$ are imputations of X, Y for indices $U \cup W$, and $V \cup W$, respectively. Then we have $X_*(i) = X(i)$ for any $i \in [n] \setminus (U \cup W)$, and $Y_*(i) = Y(i)$ for any $i \in [n] \setminus (V \cup W)$.

Notice that $[n] \setminus (U \cup W) = ([n] \setminus W) \setminus U = [n'] \setminus U$ and $[n] \setminus (V \cup W) = ([n] \setminus W) \setminus V = [n'] \setminus V$. Then, we have $X_*(i) = X(i)$ for any $i \in [n'] \setminus U$, and $Y_*(i) = Y(i)$ for any $i \in [n'] \setminus V$.

According to the definition of X'_*, Y'_*, X', Y', we have $X'_*(i) = X(i) = X'(i)$ for any $i \in [n'] \setminus U$ and $Y'_*(i) = Y(i) = Y'(i)$ for any $i \in [n'] \setminus V$. Hence, X'_* and Y'_* are imputations of X'_1, Y'_1 for U and V, respectively.

Since the condition (ii) is true, then according to Lemma C.2.2, for any $i \in U \cup V$, we have

$$R(X_*(i), X_*) = R(X_*(i), (X_*(l))_{l \in [n] \setminus W}) = R(X_*(i), X'_*).$$

Similarly, since the condition (iii) is true, then according to Lemma C.2.2, for any $i \in U \cup V$, we have

$$R(Y_*(i), Y_*) = R(Y_*(i), (Y_*(l))_{l \in [n] \setminus W}) = R(Y_*(i), Y'_*).$$

According to condition (i), for any $i \in U \cup V$, we have

$$R(X_*(i), X_*) = R(Y_*(i), Y_*).$$

Hence, for any $i \in U \cup V$, we have

$$R(X_*(i), X'_*) = R(Y_*(i), Y'_*).$$

According to the definition of X'_*, Y'_*, we have

$$R(X'_*(i), X'_*) = R(Y'_*(i), Y'_*), \text{ for any } i \in U \cup V.$$

Next, according to Proposition C.3.13, we can find imputations $X'_2, Y'_2 \in \mathcal{V}^{n'}$ of X'_1, Y'_1 for U and V, respectively, such that $R(X'_2(i), X'_1) = R(Y'_2(i), Y'_1)$ for any $i \in U \cup V$. Since X'_* and Y'_* are imputations of X'_1, Y'_1 for U and V, respectively, such that $R(X'_*(i), X'_*) = R(Y'_*(i), Y'_*)$, for any $i \in U \cup V$. Then, according to Proposition C.3.23, we have

$$D\left(X'_2, Y'_2\right) = D\left(X'_*, Y'_*\right).$$

Notice that the conditions (i), (ii), and (iii) all hold. Then according to Proposition C.3.31, we have

$$D\left(X'_*, Y'_*\right) = D\left(X_*, Y_*\right).$$

Hence, we have

$$D\left(X_*, Y_*\right) = D\left(X'_2, Y'_2\right). \tag{C.131}$$

Next, let $X_2, Y_2 \in \mathcal{V}^n$ be vectors such that the following five conditions all hold

$$X_2(i) = X_2'(i), \text{ for any } i \in [n] \setminus W, \tag{C.132}$$
$$Y_2(i) = Y_2'(i), \text{ for any } i \in [n] \setminus W, \tag{C.133}$$
$$R(X_2(i), X_2) = R(Y_2(i), Y_2), \text{ for any } i \in W, \tag{C.134}$$
$$\min(X_2(l))_{l \in W} > \max(X_2(l))_{l \in [n] \setminus W}, \tag{C.135}$$
$$\text{and } \min(Y_2(l))_{l \in W} > \max(Y_2(l))_{l \in [n] \setminus W}. \tag{C.136}$$

Combining (C.132) and (C.133), we have $X_2(i) = X_2'(i)$ and $Y_2(i) = Y_2'(i)$ for any $i \in [n'] \setminus W = [n']$. Further, since $X_2', Y_2' \in \mathcal{V}^{n'}$ are imputations of X_1', Y_1' for U and V, respectively, we have $X_2'(i) = X_1'(i)$ for any $i \in [n'] \setminus U$ and $Y_2'(i) = Y_1'(i)$ for any $i \in [n'] \setminus V$. Hence, we have

$$X_2(i) = X_2'(i) = X_1'(i) = X_1(i), \text{ for any } i \in [n'] \setminus U,$$
$$\text{and } Y_2(i) = Y_2'(i) = Y_1'(i) = Y_1(i), \text{ for any } i \in [n'] \setminus V.$$

Notice that $[n'] = [n] \setminus W$. Hence, we have $[n'] \setminus U = [n] \setminus (W \cup U)$, and $[n'] \setminus V = [n] \setminus (W \cup V)$. Therefore, it follows that

$$X_2(i) = X_2'(i) = X_1'(i) = X_1(i), \text{ for any } i \in [n] \setminus (W \cup U),$$
$$\text{and } Y_2(i) = Y_2'(i) = Y_1'(i) = Y_1(i), \text{ for any } i \in [n] \setminus (W \cup V).$$

In other words, X_2 and Y_2 are imputations of X_1, Y_1 for indices $W \cup U$ and $W \cup V$, respectively. Notice that (C.134), (C.135) and (C.136) are true. Then according to Theorem C.3.32, we have

$$D(X_1, Y_1) \geq D(X_2, Y_2).$$

Since (C.134), (C.135) and (C.136) are true, then according to Proposition C.3.31, we have

$$D(X_2, Y_2) = D\left(X_2', Y_2'\right).$$

Hence, we have

$$D(X_1, Y_1) \geq D\left(X_2', Y_2'\right) =^{(\text{C.131})} D(X_*, Y_*).$$

This completes proof for the statement (2), and finishes our proof. □

It is straightforward to show the existence of such X_*, Y_* in Theorem 5.1.1 by applying Proposition C.3.13. Hence the lower bounds for Spearman's footrule in the general missing case are tight.

For the imputations X_* and Y_* in Theorem 5.1.1, Proposition C.3.31 gives the equation $D(X_*, Y_*) = D\left((X_*(i))_{i \in [n] \setminus W}, (Y_*(i))_{i \in [n] \setminus W}\right)$, which suggests that, for indices in W, the components of both X and Y can be ignored for the purposes of calculating the lower bound. Hence, the minimum possible value of Spearman's footrule in the general missing case can be determined using Algorithm 3 after ignoring all components with indices in W. This algorithm has complexity $\mathcal{O}(n \log n + n(m_1 + m_2))$, as discussed in Remark C.3.14.

Table 5.1 provides a simple example where X and Y are missing under general case. The imputations X_*, Y_* satisfy the three conditions required in Theorem 5.1.1. Hence, the imputations X_*, Y_* results the minimum possible Spearman's footrule across all possible imputations.

C.4. Tight upper bound of Spearman's footrule

This section provides tight upper bounds of Spearman's footrule in the presence of missing data. We consider missing case I, II and III separately, and then combine our results to obtain the upper bounds in general missingness case.

C.4.1. Upper bound under missing case I

We start by considering missing case I for $X, Y \in \mathcal{V}^n$, where the missing components are assumed to be either all in X or all in Y. Without loss of generality, let us assume only X contains missing components while all values in Y are observed.

Suppose only a single component $X(u)$ is missing. The following Proposition C.4.2 is complementary to Proposition C.3.5, which considers the lower bound for missing case I. We start our proof for Proposition C.4.2 by providing the following lemma.

Lemma C.4.1. Suppose $X, Y \in \mathcal{V}^n$, and let X_1, X_2 be imputations of X for an index $u \in [n]$ such that $R(X_1(u), X_1) = 1$, and $R(X_2(u), X_2) = n$. Then, if

$$\text{(i)}: \ R(X(u), X) \geq R(Y(u), Y)$$

holds, we have $D(X, Y) \leq D(X_2, Y)$. However, if

$$\text{(ii)}: \ R(X(u), X) \leq R(Y(u), Y)$$

holds, we have $D(X, Y) \leq D(X_1, Y)$.

Proof. We prove the results for the case (i) and (ii) separately.

Suppose case (i) : $R(X(u), X) \geq R(Y(u), Y)$ is true. Then, we have either $R(X(u), X) = n$ or $R(X(u), X) < n$. Suppose $R(X(u), X) = n$. Then, we have

$$R(X(u), X) = n = R(X_2(u), X_2).$$

Since X_2 is an imputation of X for the index u, then according to Lemma C.2.3, we have $R(X_2, X_2) = R(X, X)$. Hence, we have $D(X, Y) = D(X_2, Y)$.

Suppose $R(X(u), X) < n$. Denote $R(X(u), X) = n - b$, where $b \in \mathbb{N}$ and $b > 0$. Then, according to Lemma C.3.2, there exist $V_1, \ldots, V_b \in \mathcal{V}^n$ such that for any $k \in \{1, \ldots, b\}$, V_k is an imputation of X for the index u and

$$R(V_k(u), V_k) = R(X(u), X) + k.$$

Define $V_0 = X$. Then, for any $k \in \{1, \ldots, b\}$, we have

$$R(V_{k-1}(u), V_{k-1}) = R(X(u), X) + (k-1) \geq R(X(u), X) \geq R(Y(u), Y),$$
$$\text{and } R(V_k(u_1), V_k) = R(V_{k-1}(u_1), V_{k-1}) + 1.$$

By applying Lemma C.3.3 for each pair V_k, Y and V_{k-1}, Y, where $k \in \{1, \ldots, b\}$, we have

$$D(V_0, Y) \leq D(V_1, Y) \leq \ldots \leq D(V_b, Y). \tag{C.137}$$

Notice that

$$R(V_b(u), V_b) = R(X(u), X) + b = n = R(X_2(u), X_2).$$

Since V_b and X_2 are both imputation of X for the index u, then according to Lemma C.2.3, we have

$$D(X_2, Y) = D(V_b, Y).$$

According to (C.137), we further have

$$D(X_2, Y) \geq D(V_0, Y).$$

Since $V_0 = X$, then we have $D(X, Y) \leq D(X_2, Y)$. This completes our proof when the case (i) is true.

When the case (ii) is true, the results can be proved similarly. Suppose the case (ii) is true, then we have either $R(X(u), X) = 1$, or $R(X(u), X) > 1$ is true. Suppose $R(X(u), X) = 1$. Then,

$$R(X(u), X) = 1 = R(X_1(u), X).$$

Since X_1 is an imputation of X for the index u, then according to Lemma C.2.3 we have $R(X_1, X_1) = R(X, X)$. Hence, we have $D(X, Y) = D(X_1, Y)$.

Suppose $R(X(u), X) > 1$. Denote $R(X(u_1), X) = 1 + a$, where $a \in \mathbb{N}$ and $a > 0$. Then, according to Lemma C.3.2, there exist $V_1, \ldots, V_a \in \mathcal{V}^n$ such that for any $k \in \{1, \ldots, a\}$, V_k is an imputation of X for the index u and

$$R(V_k(u), V_k) = R(X(u), X) - k.$$

Define $V_0 = X$. Then, for any $k \in \{1, \ldots, a\}$, we have

$$R(V_{k-1}(u), V_{k-1}) = R(X(u), X) - (k-1) \leq R(X(u), X) \leq R(Y(u), Y),$$
$$\text{and } R(V_k(u), V_k) = R(V_{k-1}(u), V_{k-1}) - 1.$$

By applying Lemma C.3.3 for each pair V_k, Y and V_{k-1}, Y, where $k \in \{1, \ldots, a\}$, we have

$$D\left(V_0, Y\right) \leq D\left(V_1, Y\right) \leq \cdots \leq D\left(V_a, Y\right). \tag{C.138}$$

Notice that

$$R(V_a(u), V_a) = R(X(u), X) - a = 1 = R(X_1(u), X_1).$$

Since V_a and X_1 are both imputation of X for the index u, then according to Lemma C.2.3, we have

$$D\left(X_1, Y\right) = D\left(V_a, Y\right).$$

According to (C.138), we further have

$$D\left(X_1, Y\right) \geq D\left(V_0, Y\right).$$

Since $V_0 = X$, then we have $D\left(X_1, Y\right) \geq D\left(X, Y\right)$. This completes our proof. □

By applying C.4.1, we can prove Proposition C.4.2.

Proposition C.4.2. Suppose $X, Y \in \mathcal{V}^n$, and let X_1, X_2 be imputations of X for an index $u \in [n]$ such that $R(X_1(u), X_1) = 1$, and $R(X_2(u), X_2) = n$. Then,

$$D\left(X, Y\right) \leq \max\{D\left(X_1, Y\right), D\left(X_2, Y\right)\},$$

and for any imputation $X' \in \mathcal{V}^n$ of X for index u,

$$D\left(X', Y\right) \leq \max\{D\left(X_1, Y\right), D\left(X_2, Y\right)\}.$$

Proof. According to Lemma C.4.1, if $R(X(u), X) \geq R(Y(u), Y)$, we have $D\left(X, Y\right) \leq D\left(X_1, Y\right) \leq \max\{D\left(X_1, Y\right), D\left(X_2, Y\right)\}$. However, if $R(X(u), X) \leq R(Y(u), Y)$, we have $D\left(X, Y\right) \leq D\left(X_2, Y\right) \leq \max\{D\left(X_1, Y\right), D\left(X_2, Y\right)\}$. Hence, we have

$$D\left(X, Y\right) \leq \max\{D\left(X_1, Y\right), D\left(X_2, Y\right)\}.$$

Next, we show that $D\left(X', Y\right) \leq \max\{D\left(X_1, Y\right), D\left(X_2, Y\right)\}$. According to Lemma C.3.2, there exist imputations $X_3, X_4 \in \mathcal{V}^n$ of X' for the index u such that $R(X_3(u), X_3) = 1$, and $R(X_4(u), X_4) = n$.

Then, according to Lemma C.4.1, if $R(X'(u), X) \geq R(Y(u), Y)$, we have $D\left(X', Y\right) \leq D\left(X_3, Y\right) \leq \max\{D\left(X_3, Y\right), D\left(X_4, Y\right)\}$. However, if $R(X'(u), X) \leq R(Y(u), Y)$, we have $D\left(X', Y\right) \leq D\left(X_4, Y\right) \leq \max\{D\left(X_3, Y\right), D\left(X_4, Y\right)\}$. Hence, we have $D\left(X', Y\right) \leq \max\{D\left(X_3, Y\right), D\left(X_4, Y\right)\}$.

Notice that X_3, X_4 are both imputations of X' for the index u, and X' is the imputation of X for index u. Hence, according to the definition of imputations, we can show that X_3, X_4 are both imputation of X for the index u. Notice that X_1, X_2 are imputation of X for the index u such that $R(X_1(u), X_1) = 1$, and $R(X_2(u), X_2) = n$. Then, according to Lemma C.2.3, we have $R(X_3, X_3) = R(X_1, X_1)$, and $R(X_4, X_4) = R(X_2, X_2)$. Hence, we have $D\left(X_3, Y\right) = D\left(X_1, Y\right)$

and $D(X_4, Y) = D(X_2, Y)$. Notice that $D(X', Y) \leq \max\{D(X_3, Y), D(X_4, Y)\}$. Therefore, we have

$$D\left(X', Y\right) \leq \max\{D(X_1, Y), D(X_2, Y)\}.$$

This completes our proof. □

When a single component of X is missing, Proposition C.4.2 shows that the maximum possible value of Spearman's footrule is achieved when the missing component is imputed to have rank either 1 or n.

To extend Proposition C.4.2 to cases when multiple components of X may be missing, we first make the following definition.

Definition C.4.3. Let $U \subset T \subset [n]$ be a subset of indices. Define

$$\mathcal{Z}(U, T, n) = \{Z \in \mathcal{V}^n | \text{ for any } i \in U,\ Z(i) > \max_{j \in T \setminus U} Z(j) \text{ or } Z(i) < \min_{j \in T \setminus U} Z(j)\}.$$

It will often happen that $T = [n]$ in $\mathcal{Z}(U, T, n)$, but this will not always be the case. This definition allows us to obtain the maximum possible Spearman's footrule by considering rank vectors of all possible imputations X_* of X in the set $\mathcal{Z}(U, [n], n)$, as shown in Proposition C.4.5.

For proving Proposition C.4.5, we first show the following result.

Lemma C.4.4. Let $U \subset [n]$ be a non-empty subset of indices, and suppose $X_1 \in \mathcal{Z}(U, [n], n)$. Let $V \subset [n]$ be a non-empty subset of indices such that $V \cap U = \emptyset$. Then, if $X_2 \in \mathcal{Z}(V, [n], n)$ is an imputation of X_1 for V, we have $X_2 \in \mathcal{Z}(U \cup V, [n], n)$.

Proof. Since $X_1 \in \mathcal{Z}(U, [n], n)$, we have

$$X_1(i) > \max_{j \in [n] \setminus U} X_1(j), \text{ or } X_1(i) < \min_{j \in [n] \setminus U} X_1(j), \text{ for any } i \in U.$$

Hence, we have

$$X_1(i) > \max_{j \in [n] \setminus (U \cup V)} X_1(j), \text{ or } X_1(i) < \min_{j \in [n] \setminus (U \cup V)} X_1(j), \text{ for any } i \in U. \tag{C.139}$$

Next, since X_2 is an imputation of X_1 for indices V, we have

$$X_2(i) = X_1(i), \text{ for any } i \in [n] \setminus V.$$

Notice that $[n] \setminus (U \cup V) \subset [n] \setminus V$. Hence, we have

$$\begin{aligned} & X_2(i) = X_1(i), \text{ for any } i \in [n] \setminus (V \cup U) \\ \implies & \max_{j \in [n] \setminus (U \cup V)} X_2(j) = \max_{j \in [n] \setminus (U \cup V)} X_1(j), \text{ and } \min_{j \in [n] \setminus (U \cup V)} X_2(j) = \min_{j \in [n] \setminus (U \cup V)} X_1(j). \end{aligned}$$

Meanwhile, since $U \cap V = \emptyset$, we have $U \subset [n] \setminus V$. Hence, we have

$$X_2(i) = X_1(i), \text{ for any } i \in U.$$

Then, according to (C.139), we have

$$X_2(i) > \max_{j\in[n]\setminus(U\cup V)} X_2(j), \text{ or } X_2(i) < \min_{j\in[n]\setminus(U\cup V)} X_2(j), \text{ for any } i \in U. \tag{C.140}$$

Since $X_2 \in \mathcal{Z}(V, [n], n)$, then we have

$$X_2(i) > \max_{j\in[n]\setminus V} X_2(j), \text{ or } X_2(i) < \min_{j\in[n]\setminus V} X_2(j), \text{ for any } i \in V.$$

Notice that $U \cap V = \emptyset$. Then, we have $U \subset [n] \setminus V$. Hence, we have

$$X_2(i) > \max_{j\in U} X_2(j), \text{ or } X_2(i) < \min_{j\in U} X_2(j), \text{ for any } i \in V.$$

According to (C.140), we have

$$\max_{j\in U} X_2(j) > \max_{j\in[n]\setminus(U\cup V)} \text{ or } \min_{j\in U} X_2(j) < \min_{j\in[n]\setminus(U\cup V)} X_2(j).$$

Hence, we further have

$$X_2(i) > \max_{j\in[n]\setminus(U\cup V)} X_2(j), \text{ or } X_2(i) < \min_{j\in[n]\setminus(U\cup V)} X_2(j), \text{ for any } i \in V.$$

Combining this result with (C.140), we have

$$X_2(i) > \max_{j\in[n]\setminus(U\cup V)} X_2(j), \text{ or } X_2(i) < \min_{j\in[n]\setminus(U\cup V)} X_2(j), \text{ for any } i \in V \cup U.$$

This completes our proof.

□

We are now ready to prove Proposition C.4.5.

Proposition C.4.5. Suppose $X, Y \in \mathcal{V}^n$, and let $U \subset [n]$. Then, there exists an imputation $X_* \in \mathcal{Z}(U, [n], n)$ of X for indices U such that $D(X, Y) \leq D(X_*, Y)$, and for any other imputation $X' \in \mathcal{V}^n$ of X for indices U, $D(X', Y) \leq D(X_*, Y)$.

Proof. In order to show Proposition C.4.5, we first show the following statement is true:

$\mathbb{S}$: Suppose $X, Y \in \mathcal{V}^n$, and let $U \subset [n]$. Then, there exists an imputation $X_* \in \mathcal{Z}(U, [n], n)$ of X for indices U such that $D(X, Y) \leq D(X_*, Y)$.

Then we prove Proposition C.4.5 using statement $\mathbb{S}$.

First, we show the statement $\mathbb{S}$ is true. Suppose $U = \emptyset$, then according to the definition of $\mathcal{Z}([n], [n], n)$ and imputations, we have $X_* = X$. Hence, we have $D(X_*, Y) = D(X, Y)$. This proves the statement $\mathbb{S}$ when $U = \emptyset$.

Suppose $U = [n]$. Then, according to the definition of $\mathcal{Z}([n], [n], n)$ and imputations, any vector $X_* \in \mathcal{V}^n$ is such that $X_* \in \mathcal{Z}([n], [n], n)$ and X_* is an imputation of X for indices $U = [n]$. Let $X_* = X$, then we have $D(X_*, Y) = D(X, Y)$. This proves the statement $\mathbb{S}$ when $U = [n]$.

In the following, we only prove the statement $\mathbb{S}$ when $U \neq [n]$ and $U \neq \emptyset$.

For any given $n \in \mathbb{N}$, let $P(k)$ be the statement $\mathbb{S}$ when $|U| = k$. We prove $P(k)$ is true for any $k \in \{1, \ldots, n-1\}$ by induction on k.

Base Case: We show $P(1)$ is true. Suppose $|U| = 1$. Let us denote $U = \{u\}$.

Then, let X_1, X_2 be imputations of X for the index $u \in [n]$ such that $R(X_1(u), X_1) = 1$, and $R(X_2(u), X_2) = n$. Since $R(X_1(u), X_1) = 1$, we have

$$X_1(u) < \min_{j \in T \setminus U} X_1(j).$$

Hence, $X_1 \in \mathcal{Z}(U, [n], n)$. Similarly, since $R(X_2(u), X_2) = n$, we have

$$X_2(u) > \max_{j \in T \setminus U} X_2(j).$$

Hence, $X_2 \in \mathcal{Z}(U, [n], n)$.

Next, according to Proposition C.4.2, we have

$$D\left(X', Y\right) \leq \max\{D\left(X_1, Y\right), D\left(X_2, Y\right)\}.$$

Hence, we have shown that $P(1)$ is true.

Induction Steps: We show the implication $P(1), P(k) \implies P(k+1)$ for any $k \in \{1, \ldots, n-2\}$.

Denote $m = k+1$. Without loss of generality, let us assume (after relabelling) $U = \{1, 2, \cdots, m\}$. Then, since $P(k)$ is true, there exists an imputation $X_1 \in \mathcal{Z}(U \setminus \{1\}, [n], n)$ of X for indices $U \setminus \{1\}$ such that $D\left(X, Y\right) \leq D\left(X_1, Y\right)$.

Since $P(1)$ is true, there exists an imputation $X_2 \in \mathcal{Z}(\{1\}, [n], n)$ of X_1 for the index $\{1\}$ such that $D\left(X_1, Y\right) \leq D\left(X_2, Y\right)$.

Then, we have $D\left(X, Y\right) \leq D\left(X_2, Y\right)$. Further, according to Lemma C.4.4, we have $X_2 \in \mathcal{Z}(U, [n], n)$.

Next, since X_1 is an imputation of X for indices $U \setminus \{1\}$, we have $X_1(i) = X(i)$ for any $i \in [n] \setminus U$. Further, since X_2 is an imputation of X_1 for the index $\{1\}$, we have $X_2(i) = X_1(i)$ for any $i \in [n] \setminus U$. Hence, we have $X_2(i) = X(i)$ for any $i \in [n] \setminus U$. In other words, X_2 is an imputation of X for indices U,

Thus, $X_2 \in \mathcal{Z}(U, [n], n)$ is an imputation of X for indices U such that $D\left(X, Y\right) \leq D\left(X_2, Y\right)$. Hence, we have shown $P(k+1)$ is true. This completes our proof for the statement $\mathbb{S}$.

Next, we prove Proposition C.4.5 using the Statement $\mathbb{S}$. Let us denote

$$S = \{D\left(V, Y\right) : V \in \mathcal{Z}(U, [n], n) \text{ is an imputation of } X \text{ for indices } U\}.$$

Then, the carnality of S is finite and there exist $X_* \in \mathcal{Z}(U, [n], n)$ of X for indices U such that $D\left(X_*, Y\right) = \max S$.

We now show that $D\left(X, Y\right) \leq D\left(X_*, Y\right)$. According to statement $\mathbb{S}$, there exist imputation $X_1 \in \mathcal{Z}(U, [n], n)$ of X for indices U such that $D\left(X, Y\right) \leq D\left(X_1, Y\right)$. Then, according to the definition of S and X_*, we have $D\left(X_1, Y\right) \leq D\left(X_*, Y\right)$. Hence, we have $D\left(X, Y\right) \leq D\left(X_*, Y\right)$.

Next, we show that $D\left(X', Y\right) \leq D\left(X_*, Y\right)$. According to statement $\mathbb{S}$, there exist imputation $X_2 \in \mathcal{Z}(U, [n], n)$ of X' for indices U such that $D\left(X', Y\right) \leq D\left(X_2, Y\right)$. Since X' is an imputation of X for indices U, then X_2 is also an imputation of X for indices U. Then, according to the

definition of X_*, we have $D(X_2, Y) \leq D(X_*, Y)$. Overall, we have $D(X', Y) \leq D(X_*, Y)$. This completes our proof.

□

We briefly consider the number of possible $X_* \in \mathcal{Z}(U, [n], n)$, up to equivalent rank vectors. An imputation $X_* \in \mathcal{Z}(U, [n], n)$ is such that all components in U are either larger or smaller than all components at indices $[n] \setminus U$. Let $m_1 = |U|$ and let r be the number of components smaller than all components in $[n] \setminus U$. Then, there are $m_1 - r$ components of X_* that are larger than its all components with indices in $[n] \setminus U$.

For the r smaller components with indices in U, each component can take a distinct rank between 1 and r. Hence, there are $r!$ possibilities for the ranks of these components. Similarly, for the $m_1 - r$ larger components, each component can take a distinct rank between $n-(m_1-r)+1$ and n. Hence, there are $(m_1 - r)!$ possible ranks for these larger components. Since $0 \leq r \leq m_1$, the number of possible rank vectors of imputations $X_* \in \mathcal{Z}(U, [n], n)$ is given by $\sum_{r=0}^{m_1} r!(m_1-r)!$.

Even for a moderately small m_1, such as $m_1 = 20$, the number of permutations can exceed 5.14×10^{18}. Therefore, further refinement of Proposition C.4.5 is necessary to determine the maximum possible Spearman's footrule efficiently. We therefore consider ordering the components of imputations X_* with indices in U according to the values of their paired components, as shown in Theorem C.4.10. We start by introducing the following lemma.

Lemma C.4.6. Suppose $X, Y \in \mathcal{V}^n$, and assume $Y(1) < \ldots < Y(n)$. Let X_1 be a permutation of X such that $X_1(1) > \ldots > X_1(n)$. Then, we have

$$\sum_{i=1}^{n} |X(i) - Y(i)| \leq \sum_{i=1}^{n} |X_1(i) - Y(i)|. \tag{C.141}$$

Proof. Let $P(k)$ be the statement of Lemma C.4.6 when $n = k$. We prove that $P(k)$ is true for any $k \in \mathbb{N}$ such that $k \geq 2$ by induction on k.

Base Case: We show $P(2)$ is true. Suppose $n = 2$. Then, we have $X = (X(1), X(2))$. If $X(1) > X(2)$, then $X_1 = (X(1), X(2))$. Hence, we have $X = X_1$ and then (C.141) holds.

Thus, it is sufficient to prove (C.141) when $X(1) < X(2)$. Suppose $X(1) < X(2)$. Then, we have $X_1 = (X(2), X(1))$. Therefore, the right-hand side of (C.141) is such that

$$|X_1(1) - Y(1)| + |X_1(2) - Y(2)| = |X(2) - Y(1)| + |X(1) - Y(2)|. \tag{C.142}$$

Since $Y(1) < Y(2)$, and $X(1) < X(2)$, then there are following 6 cases of the order between $Y(1), Y(2), X(1)$, and $X(2)$:

$$\begin{aligned}
&(1) : Y(1) < Y(2) \leq X(1) < X(2),\\
&(2) : Y(1) \leq X(1) \leq Y(2) \leq X(2),\\
&(3) : X(1) \leq Y(1) < Y(2) \leq X(2),\\
&(4) : Y(1) \leq X(1) < X(2) \leq Y(2),\\
&(5) : X(1) \leq Y(1) \leq X(2) \leq Y(2),\\
&(6) : X(1) < X(2) \leq Y(1) < Y(2).
\end{aligned}$$

In the following, we are going to consider the 6 cases separately.

Suppose the case $(1) : Y(1) < Y(2) \leq X(1) < X(2)$ is true. Then, the left-hand side of (C.141) is

$$|X(1) - Y(1)| + |X(2) - Y(2)| = X(1) - Y(1) + X(2) - Y(2).$$

The right-hand side of (C.141) is

$$\begin{aligned} |X_1(1) - Y(1)| + |X_1(2) - Y(2)| &=^{(\text{C.142})} |X(2) - Y(1)| + |X(1) - Y(2)| \\ &= X(2) - Y(1) + X(1) - Y(2) \\ &= X(1) - Y(1) + X(2) - Y(2). \end{aligned}$$

Then, the right-hand side of (C.141) minus the left-hand side of (C.141) equals to

$$(X(1) - Y(1) + X(2) - Y(2)) - (X(1) - Y(1) + X(2) - Y(2)) = 0.$$

Hence, we have shown (C.141) is true when the case (1) is true.

Next, suppose the case $(2) : Y(1) \leq X(1) \leq Y(2) \leq X(2)$ is true. Then, the left-hand side of (C.141) is

$$\begin{aligned} |X(1) - Y(1)| + |X(2) - Y(2)| &= X(1) - Y(1) + X(2) - Y(2) \\ &= (X(2) - Y(1)) + (X(1) - Y(2)). \end{aligned}$$

The right-hand side of (C.141) is

$$\begin{aligned} |X_1(1) - Y(1)| + |X_1(2) - Y(2)| &=^{(\text{C.142})} |X(2) - Y(1)| + |X(1) - Y(2)| \\ &= X(2) - Y(1) + Y(2) - X(1) \\ &= (X(2) - Y(1)) + (Y(2) - X(1)). \end{aligned}$$

Subsequently, the right-hand side of (C.141) minus the left-hand side of (C.141) equals to

$$\begin{aligned} &(X(2) - Y(1)) + (Y(2) - X(1)) - (X(2) - Y(1)) - (X(1) - Y(2)) \\ &= (Y(2) - X(1)) - (X(1) - Y(2)) = 2Y(2) - 2X(1) \geq 0. \end{aligned}$$

Hence, we have shown (C.141) is true when the case (2) is true.

Next, suppose the case $(3) : X(1) \leq Y(1) < Y(2) \leq X(2)$ is true. Then, the left-hand side of (C.141) is

$$\begin{aligned} |X(1) - Y(1)| + |X(2) - Y(2)| &= Y(1) - X(1) + X(2) - Y(2) \\ &= (X(2) - X(1)) + (Y(1) - Y(2)). \end{aligned}$$

The right-hand side of (C.141) is

$$\begin{aligned}|X_1(1) - Y(1)| + |X_1(2) - Y(2)| &=^{(\text{C.142})} |X(2) - Y(1)| + |X(1) - Y(2)| \\ &= X(2) - Y(1) + Y(2) - X(1) \\ &= (X(2) - X(1)) + (Y(2) - Y(1)).\end{aligned}$$

Subsequently, the right-hand side of (C.141) minus the left-hand side of (C.141) equals to

$$\begin{aligned}&(X(2) - X(1)) + (Y(2) - Y(1)) - (X(2) - X(1)) - (Y(1) - Y(2)) \\ &= (Y(2) - Y(1)) - (Y(1) - Y(2)) = 2Y(2) - 2Y(1) > 0.\end{aligned}$$

Hence, we have shown (C.141) is true when the case (3) is true.

Next, suppose the case (4) : $Y(1) \leq X(1) < X(2) \leq Y(2)$ is true. Then, the left-hand side of (C.141) is

$$\begin{aligned}|X(1) - Y(1)| + |X(2) - Y(2)| &= X(1) - Y(1) + Y(2) - X(2) \\ &= (Y(2) - Y(1)) + (X(1) - X(2)).\end{aligned}$$

The right hand side of (C.141) is,

$$\begin{aligned}|X_1(1) - Y(1)| + |X_1(2) - Y(2)| &=^{(\text{C.142})} |X(2) - Y(1)| + |X(1) - Y(2)| \\ &= X(2) - Y(1) + Y(2) - X(1) \\ &= (Y(2) - Y(1)) + (X(2) - X(1)).\end{aligned}$$

Subsequently, the right hand side of (C.141) minus the left hand side of (C.141) equals

$$\begin{aligned}&(Y(2) - Y(1)) + (X(2) - X(1)) - (Y(2) - Y(1)) - (X(1) - X(2)) \\ &= (X(2) - X(1)) - (X(1) - X(2)) = 2X(2) - 2X(1) > 0.\end{aligned}$$

Hence, we have shown (C.141) is true when the case (4) is true.

Next, suppose the case (5) : $X(1) \leq Y(1) \leq X(2) \leq Y(2)$ is true. Then, the left hand side of (C.141) is

$$\begin{aligned}|X(1) - Y(1)| + |X(2) - Y(2)| &= Y(1) - X(1) + Y(2) - X(2) \\ &= (Y(2) - X(1)) + (Y(1) - X(2)).\end{aligned}$$

The right hand side of (C.141) is,

$$\begin{aligned}|X_1(1) - Y(1)| + |X_1(2) - Y(2)| &=^{(\text{C.142})} |X(2) - Y(1)| + |X(1) - Y(2)| \\ &= X(2) - Y(1) + Y(2) - X(1) \\ &= (Y(2) - X(1)) + (X(2) - Y(1)).\end{aligned}$$

Subsequently, the right hand side of (C.141) minus the left hand side of (C.141) equals to

$$\begin{aligned}&(Y(2)-X(1))+(X(2)-Y(1))-(Y(2)-X(1))-(Y(1)-X(2))\\&=(X(2)-Y(1))-(Y(1)-X(2))=2X(2)-2Y(1)\geq 0.\end{aligned}$$

Hence, we have shown (C.141) is true when the case (5) is true.

Next, suppose the case (6) : $X(1) < X(2) \leq Y(1) < Y(2)$ is true. Then, the left hand side of (C.141) is

$$|X(1)-Y(1)|+|X(2)-Y(2)|=Y(1)-X(1)+Y(2)-X(2).$$

The right hand side of (C.141) is,

$$\begin{aligned}|X_1(1)-Y(1)|+|X_1(2)-Y(2)| &=^{(\text{C.142})} |X(2)-Y(1)|+|X(1)-Y(2)|\\&=Y(1)-X(2)+Y(2)-X(1)\\&=Y(1)-X(1)+Y(2)-X(2).\end{aligned}$$

Subsequently, the right hand side of (C.141) minus the left hand side of (C.141) equals to

$$(Y(1)-X(1)+Y(2)-X(2))-(Y(1)-X(1)+Y(2)-X(2))=0$$

Hence, we have shown (C.141) is true when case (6) is true. This completes our proof for $P(2)$.

Induction Step: We show the implication $P(2), P(k) \implies P(k+1)$ for any integer $k \geq 2$.

Let X_2 be a permutation of X such that $X_2(1) > \ldots > X_2(n-1)$ and $X_2(n) = X(n)$. Then, since $P(k)$ is true, we have

$$\sum_{i=1}^{n-1}|X(i)-Y(i)| \leq \sum_{i=1}^{n-1}|X_2(i)-Y(i)|.$$

Since $X_2(n) = X(n)$, we have

$$|X(n)-Y(n)| = |X_2(n)-Y(n)|.$$

Then, we have

$$\sum_{i=1}^{n}|X(i)-Y(i)| \leq \sum_{i=1}^{n}|X_2(i)-Y(i)|. \tag{C.143}$$

In the following, we are going to show $P(k+1)$ is true when

$$X(n) < \min\{X(1),\ldots,X(n-1)\}$$

and

$$X(n) > \min\{X(1),\ldots,X(n-1)\},$$

separately.

Suppose $X(n) < \min\{X(1), \ldots, X(n-1)\}$. Then, since $X_2(1), \ldots, X_2(n-1)$ is a permutation of $\{X(1), \ldots, X(n-1)\}$, we have

$$\min\{X(1), \cdots, X(n-1)\} = \min\{X_2(1), \cdots, X_2(n-1)\}.$$

Next, since $X_2(n) = X(n)$, then we further have

$$X_2(n) = X(n) < \min\{X(1), \cdots, X(n-1)\} = \min\{X_2(1), \cdots, X_2(n-1)\}.$$

Since X_2 is a permutation of X such that $X_2(1) > \ldots > X_2(n-1)$, and $X_2(n) < \min\{X_2(1), \cdots, X_2(n-1)\}$, we have X_2 is a permutation of X such that $X_2(1) > \ldots > X_2(n)$.

Since X_1 is a permutation of X such that $X_1(1) > \ldots > X_1(n)$, then we have $X_2 = X_1$. According to (C.143), we have

$$\sum_{i=1}^{n} |X(i) - Y(i)| \leq \sum_{i=1}^{n} |X_1(i) - Y(i)|.$$

Hence, we have shown $P(k+1)$ when $X(n) < \min\{X(1), \ldots, X(n-1)\}$.

However, suppose

$$X(n) > \min\{X(1), \ldots, X(n-1)\}.$$

Then, since $X_2(1), \cdots, X_2(n-1)$ is a permutation of $\{X(1), \cdots, X(n-1)\}$, we have

$$\min\{X(1), \cdots, X(n-1)\} = \min\{X_2(1), \cdots, X_2(n-1)\}.$$

Since $X_2(n) = X(n)$, then we further have

$$X_2(n) = X(n) > \min\{X(1), \ldots, X(n-1)\} = \min\{X_2(1), \ldots, X_2(n-1)\}.$$

Further, since $X_2(1) > \ldots X_2(n-1)$, we have

$$X_2(n) > X_2(n-1) = \min\{X_2(1), \ldots, X_2(n-1)\},$$

and

$$X_2(n-1) = \min\{X_2(1), \ldots, X_2(n)\}. \tag{C.144}$$

Next, let X_3 be a permutation of X_2 such that $X_3(i) = X_2(i)$ for any $i \in \{1, \ldots, n-2\}$, $X_3(n-1) = X_2(n)$, and $X_3(n) = X_2(n-1)$. Then, since $X_2(n) > X_2(n-1)$, we have $X_3(n-1) > X_3(n)$.

Since $P(2)$ is true, we have

$$\begin{aligned} &|X_2(n-1) - Y(n-1)| + |X_2(n) - Y(n)| \\ &\leq |X_3(n-1) - Y(n-1)| + |X_3(n) - Y(n)|. \end{aligned}$$

Since $X_3(i) = X_2(i)$ for any $i \in \{1, \ldots, n-2\}$, we have

$$\sum_{i=1}^{n-2} |X_2(i) - Y(i)| = \sum_{i=1}^{n-2} |X_3(i) - Y(i)|.$$

Then, we have

$$\sum_{i=1}^{n} |X_2(i) - Y(i)| \leq \sum_{i=1}^{n} |X_3(i) - Y(i)|. \tag{C.145}$$

Next, let X_4 be a permutation of X_3 such that $X_4(1) > \ldots > X_4(n-1)$ and $X_4(n) = X_3(n)$. Then, since $P(k)$ is true, we have

$$\sum_{i=1}^{n-1} |X_3(i) - Y(i)| \leq \sum_{i=1}^{n-1} |X_4(i) - Y(i)|.$$

Since $X_4(n) = X_3(n)$, we have

$$|X_3(n) - Y(n)| = |X_4(n) - Y(n)|.$$

Thus, we have

$$\sum_{i=1}^{n} |X_3(i) - Y(i)| \leq \sum_{i=1}^{n} |X_4(i) - Y(i)|. \tag{C.146}$$

According to (C.144), we have $X_2(n-1) = \min\{X_2(1), \ldots, X_2(n)\}$. Since X_3 is a permutation of X_2 and $X_3(n) = X_2(n-1)$, then we have $X_3(n) = \min\{X_3(1), \ldots, X_3(n)\}$.

Next, since X_4 is a permutation of X_3, and $X_4(n) = X_3(n)$, we have $X_4(n) = \min\{X_4(1), \ldots, X_4(n)\}$. Notice that $X_4(1) > \ldots > X_4(n-1)$. Hence, we have $X_4(1) > \ldots > X_4(n)$.

Since X_4 is a permutation of X_3, X_3 is a permutation of X_2, X_2 is a permutation of X, we have X_4 is a permutation of X. Then, since X_1 is a permutation of X, and $X_1(1) > \ldots > X_1(n)$, we have $X_4 = X_1$. Hence, we have

$$\begin{aligned}
\sum_{i=1}^{n} |X_1(i) - Y(i)| &= \sum_{i=1}^{n} |X_4(i) - Y(i)| \\
&\geq^{\text{(C.146)}} \sum_{i=1}^{n} |X_3(i) - Y(i)| \\
&\geq^{\text{(C.145)}} \sum_{i=1}^{n} |X_2(i) - Y(i)| \\
&\geq^{\text{(C.143)}} \sum_{i=1}^{n} |X(i) - Y(i)|.
\end{aligned}$$

Thus, we have shown $P(k+1)$ is true. This completes our proof.

□

By applying Lemma C.4.6, we can show the following result.

Lemma C.4.7. Suppose $X, Y \in \mathcal{V}^n$ and for $m \leq n$ assume $U = \{1, \ldots, m\}$ and $Y(1) < \cdots < Y(m)$. Let X' be an imputation of X for indices U. Suppose X_* is also an imputation of X for indices U, and $(X_*(i))_{i \in U}$ is a permutation of $(X'(i))_{i \in U}$ such that $X_*(1) > \cdots > X_*(m)$. Then, we have $D(X', Y) \leq D(X_*, Y)$.

Proof. Since X' is an imputation of X for indices U, we have $X'(i) = X(i)$ for any $i \in [n] \setminus U$. Similarly, since X'_* is an imputation of X for indices U, we have $X_*(i) = X(i)$ for any $i \in [n] \setminus U$. Hence, we have $X_*(i) = X'(i)$ for any $i \in [n] \setminus U$.

Further, since $(X_*(i))_{i \in U}$ is a permutation of $(X'(i))_{i \in U}$, then we have X_* is a permutation of X'. Hence, according to the definition of rank, for any $i \in [n] \setminus U$, we have

$$\begin{aligned} R(X_*(i), X_*) &= \sum_{j=1}^{n} \mathrm{I}\left(X_*(j) \leq X_*(i)\right) \\ &= \sum_{j=1}^{n} \mathrm{I}\left(X'(j) \leq X_*(i)\right). \end{aligned}$$

Since $X_*(i) = X'(i)$ for any $i \in [n] \setminus U$, we further have

$$R(X_*(i), X_*) = \sum_{j=1}^{n} \mathrm{I}\left(X'(j) \leq X'(i)\right) = R(X'(i), X'), \text{ for any } i \in [n] \setminus U.$$

Then, we have

$$\sum_{i=m+1}^{n} |R(X_*(i), X_*) - R(Y(i), Y)| = \sum_{i=m+1}^{n} \left|R(X'(i), X') - R(Y(i), Y)\right|. \tag{C.147}$$

Next, since X_* is a permutation of X', then according to the definition of rank, for any $i \in U$, we have

$$R(X_*(i), X_*) = \sum_{j=1}^{n} \mathrm{I}\left(X_*(j) \leq X_*(i)\right) = \sum_{j=1}^{n} \mathrm{I}\left(X'(j) \leq X_*(i)\right).$$

Since $(X_*(i))_{i \in U}$ is a permutation of $(X'(i))_{i \in U}$, then there exist a permutation $(\sigma(l))_{l \in U}$ of U such that

$$X_*(i) = X'(\sigma(i)), \text{ for any } i \in U.$$

Hence, we have

$$R(X_*(i), X_*) = \sum_{j=1}^{n} \mathrm{I}\left(X'(j) \leq X'(\sigma(i))\right) = R(X'(\sigma(i)), X'), \text{ for any } i \in U.$$

In other words, $(R(X_*(i), X_*))_{i \in U}$ is a permutation of $(R(X'(i), X'))_{i \in U}$. Since $X_*(1) > \ldots > X_*(m)$, we have $R(X_*(1), X_*) > \ldots > R(X_*(m), X_*)$.

Then, according to Lemma C.4.6, we have

$$\sum_{i=1}^{m} |R(X_*(i), X_*) - R(Y(i), Y)| \geq \sum_{i=1}^{m} \left|R(X'(i), X') - R(Y(i), Y)\right|.$$

Combining this result with (C.147), we have

$$\begin{aligned} &\sum_{i=1}^{n} |R(X_*(i), X_*) - R(Y(i), Y)| \geq \sum_{i=1}^{n} \left|R(X'(i), X') - R(Y(i), Y)\right| \\ \implies &D\left(X_*, Y\right) \geq D\left(X', Y\right). \end{aligned}$$

This completes our proof.

□

Consider any imputation X' of X for indices U, Lemma C.4.7 suggests the values of Spearman's footrule will always be larger or equal if the components of X' with indices in U are further arranged inversely according to their paired components. More formally, let us define

Definition C.4.8. Suppose $Y \in \mathcal{V}^n$ and let $U \subset T \subset [n]$. Define

$$\mathcal{X}(Y; U, T, n) = \{Z \in \mathcal{Z}(U, T, n) | \text{ for any } i, j \in U,\ Z(i) > Z(j) \text{ if } Y(i) < Y(j)\}.$$

Then, combining Proposition C.4.5 and Lemma C.4.7, Theorem C.4.10 will allow us to efficiently determine an imputation X_* that maximises Spearman's footrule. For proving Theorem C.4.10, we first need to show the following result.

Lemma C.4.9. Suppose $X, Y \in \mathcal{V}^n$, and for $m \leq n$ assume $U = \{1, \ldots, m\}$ and $Y(1) < \ldots < Y(m)$. Suppose $X_1 \in \mathcal{Z}(U, [n], n)$ is an imputation of X for U. If X_2 is an imputation of X for indices U such that $(X_2(i))_{i \in U}$ is a permutation of $(X_1(i))_{i \in U}$, and $X_2(1) > \ldots > X_2(m)$, then we have $X_2 \in \mathcal{X}(Y; U, [n], n)$.

Proof. Since $X_1 \in \mathcal{Z}(U, [n], n)$, then we have

$$X_1(i) < \min_{j \in [n] \setminus U} X_1(j), \text{ or } X_1(i) > \max_{j \in [n] \setminus U} X_1(j), \text{ for any } i \in U.$$

Since X_2 is an imputation of X for U, we have $X_2(i) = X(i)$ for any $i \in [n] \setminus U$. Then, since X_1 is an imputation of X for U, we have $X_1(i) = X(i)$ for any $i \in [n] \setminus U$. Hence, we have $X_1(i) = X_2(i)$ for any $i \in [n] \setminus U$. Thus, we have $\min_{j \in [n] \setminus U} X_1(j) = \min_{j \in [n] \setminus U} X_2(j)$ and $\max_{j \in [n] \setminus U} X_1(j) = \max_{j \in [n] \setminus U} X_2(j)$. Then, we have

$$X_1(i) < \min_{j \in [n] \setminus U} X_2(j), \text{ or } X_1(i) > \max_{j \in [n] \setminus U} X_2(j), \text{ for any } i \in U.$$

Since $(X_2(i))_{i \in U}$ is a permutation of $(X_1(i))_{i \in U}$, then we have

$$X_2(i) < \min_{j \in [n] \setminus U} X_2(j), \text{ or } X_2(i) > \max_{j \in [n] \setminus U} X_2(j), \text{ for any } i \in U.$$

Thus, we have $X_2 \in \mathcal{Z}(U,[n],n)$. Notice that $Y(1) < \ldots < Y(m)$ and $X_2(1) > \ldots > X_2(m)$. Hence, we have for any $i,j \in U,\ X_2(i) > X_2(j)$ if $Y(i) < Y(j)$. Therefore, we have shown $X_2 \in \mathcal{X}(Y;U,[n],n)$. This completes our proof. □

By applying Lemma C.4.7, Lemma C.4.9, and Proposition C.4.5, we can show the following result.

Theorem C.4.10. Suppose $X, Y \in \mathcal{V}^n$ and let $U \subset [n]$. Then, there exists an imputation $X_* \in \mathcal{X}(Y;U,[n],n)$ of X for indices U such that $D(X,Y) \leq D(X_*,Y)$, and for any other imputation $X' \in \mathcal{V}^n$ of X for indices U, $D(X',Y) \leq D(X_*,Y)$.

Proof. Without loss of generality, let us assume (after relabelling) $U = \{1,\ldots,m\}$ and $Y(1) < \ldots < Y(m)$. Let us denote

$$S = \{D(V,Y) : V \in \mathcal{X}(Y;U,[n],n) \text{ is an imputation of } X \text{ for indices } U\}.$$

The carnality of S is finite and there exist imputation $X_* \in \mathcal{X}(Y;U,[n],n)$ of X for indices U such that $D(X_*,Y) = \max S$.

We first show that $D(X,Y) \leq D(X_*,Y)$. According to Proposition C.4.5, there exist imputation $X_1 \in \mathcal{Z}(U,[n],n)$ of X for indices U such that $D(X,Y) \leq D(X_1,Y)$.

Next, let X_2 be an imputation of X for indices U such that $(X_2(i))_{i\in U}$ is a permutation of $(X_1(i))_{i\in U}$, and $X_2(1) > \ldots > X_2(m)$. Then, according to Lemma C.4.7, we have $D(X_1,Y) \leq D(X_2,Y)$.

According to Lemma C.4.9, we have $X_2 \in \mathcal{X}(Y;U,[n],n)$. Then, according to the definition of X_*, we have $D(X_2,Y) \leq D(X_*,Y)$. Hence, we have $D(X,Y) \leq D(X_*,Y)$.

Next, we show that $D(X',Y) \leq D(X_*,Y)$. According to Proposition C.4.5, there exist imputation $X_3 \in \mathcal{Z}(U,[n],n)$ of X' for indices U such that $D(X',Y) \leq D(X_3,Y)$.

Since X' is an imputation of X for indices U, then X_3 is also an imputation of X for indices U. Let X_4 be a imputation of X for indices U such that $(X_4(i))_{i\in U}$ is a permutation of $(X_3(i))_{i\in U}$, and $X_4(1) > \ldots > X_4(m)$. Then, according to Lemma C.4.7, we have $D(X_3,Y) \leq D(X_4,Y)$.

According to Lemma C.4.9, we have $X_4 \in \mathcal{X}(Y;U,[n],n)$. Then, according to the definition of X_*, we have $D(X_4,Y) \leq D(X_*,Y)$. Hence, we have $D(X',Y) \leq D(X_*,Y)$. This completes our proof.

□

Compared with Proposition C.4.5, Theorem C.4.10 further narrows down the scope of possible imputations X_* to the set $\mathcal{X}(Y;U,[n],n)$ for achieving the maximum possible Spearman's footrule.

Remark C.4.11. Suppose an imputation $X_* \in \mathcal{X}(Y;U,[n],n)$. Then $X_* \in \mathcal{Z}(U,[n],n)$, which means all components of X_* in U are either larger or smaller than all components in $[n] \setminus U$. Let $m_1 = |U|$ and let r be the number of components smaller than all components in $[n] \setminus U$. Then, there are $m_1 - r$ components of X_* that are larger than its all components with indices in $[n] \setminus U$. The ranks of the r smaller components in U take all values between 1 and r, and the ranks of the $m_1 - r$ larger components in U take all values between $n - (m_1 - r) + 1$ and n. Crucially, the rank of each component in U is decided by the rank of its paired component. Since $0 \leq r \leq m_1$, the number of possible rank vectors for imputations $X_* \in \mathcal{X}(Y;U,[n],n)$ is $m_1 + 1$.

Computing the value of Spearman's footrule between X_* and Y, for one possible X_*, requires computational complexity $\mathcal{O}(n \log n)$. Thus, computing Spearman's footrule for $X_* \in \mathcal{X}(Y; U, [n], n)$ with the distinct (m_1+1) rank vectors has computational complexity of $\mathcal{O}(nm_1 \log n)$. However, it is not necessary to re-rank all components and re-calculate the value of Spearman's footrule every time the ranks of components of X_* in U are altered. Section C.5.1 provides an efficient algorithm for computing the maximum Spearman's footrule under missing case I with computational cost $\mathcal{O}(m_1 + n \log n)$.

C.4.2. Upper bound under missing case II

This section considers the upper bound of Spearman's footrule under missing case II, where both X and Y may contain missing components, but each pair $(X(i), Y(i))$ has *at most one* value missing.

First, we show the following result.

Lemma C.4.12. Suppose $X, Y \in \mathcal{V}^n$ and $U, V \subset [n]$ are subsets of indices such that $U \cap V = \emptyset$. Suppose X_1 is an imputation of X for U. If $Y_1 \in \mathcal{X}(X_1; V, [n], n)$, we have $Y_1 \in \mathcal{X}(X; V, [n], n)$.

Proof. Since $Y_1 \in \mathcal{X}(X_1; V, [n], n)$, then we have $Y_1 \in \mathcal{Z}(V, [n], n)$, and for any $i, j \in V$, we have $Y_i > Y_j$ if $X_1(i) < X_1(j)$.

Next, since X_1 is an imputation of X for U, then we have $X_1(i) = X(i)$ for any $i \in [n] \setminus U$. Then, since $U \cap V = \emptyset$, we have $V \subset [n] \setminus U$. Hence, we have $X_1(i) = X(i)$ for any $i \in V$.

Thus, for any $i, j \in V$, we have $Y_i > Y_j$ if $X(i) < X(j)$. Therefore, we have shown $Y_1 \in \mathcal{X}(X; V, [n], n)$. This completes our proof. □

Combining Lemma C.4.12, and Theorem C.4.10, we can show the the main results for the upper bounds under missing case II.

Theorem C.4.13. Suppose $X, Y \in \mathcal{V}^n$ and $U, V \subset [n]$ are disjoint subsets of indices such that $U \cap V = \emptyset$. Then, there exist imputations $(X_*, Y_*) \in (\mathcal{X}(Y; U, [n], n), \mathcal{X}(X; V, [n], n))$ of X, Y for indices U and V, respectively, such that $D(X, Y) \leq D(X_*, Y_*)$. Furthermore, consider any imputation $X', Y' \in \mathcal{V}^n$ of X, Y for indices U and V, respectively, we have $D(X', Y') \leq D(X_*, Y_*)$.

Proof. To start, let us denote

$$S = \{D(V, W) : (V, W) \in (\mathcal{X}(Y; U, [n], n), \mathcal{X}(X; V, [n], n)) \\ \text{are imputations of } X, Y \text{ for indices } U \text{ and } V, \text{respectively}\}.$$

Then, the carnality of S is finite and there exist imputations

$$(X_*, Y_*) \in (\mathcal{X}(Y; U, [n], n), \mathcal{X}(X; V, [n], n))$$

of X, Y for indices U and V, respectively, such that $D(X_*, Y_*) = \max S$.

We first show that $D(X, Y) \leq D(X_*, Y_*)$. According to Theorem C.4.10, there exist an This completes our proof.imputation $X_1 \in \mathcal{X}(Y; U, [n], n)$ of X for indices U such that $D(X, Y) \leq$

$D(X_1, Y)$. Applying Theorem C.4.10 again, there exist an imputation $Y_1 \in \mathcal{X}(X_1; V, [n], n)$ of Y for indices V such that $D(X_1, Y) \leq D(X_1, Y_1)$. Hence, we have $D(X, Y) \leq D(X_1, Y_1)$.

According to Lemma C.4.12, we have $Y_1 \in \mathcal{X}(X; V, [n], n)$. Notice that $X_1 \in \mathcal{X}(Y; U, [n], n)$. Then, according to the definition of (X_*, Y_*), we have $D(X_1, Y_1) \leq D(X_*, Y_*)$. Hence, we have $D(X, Y) \leq D(X_*, Y_*)$.

Next, we show that $D(X', Y') \leq D(X_*, Y_*)$. According to Theorem C.4.10, there exist an imputation $X_2 \in \mathcal{X}(Y'; U, [n], n)$ of X' for indices U such that $D(X', Y') \leq D(X_2, Y')$.

Since X' is an imputation of X for U, we have X_2 is also an imputation of X for U. Meanwhile, notice that Y' is an imputation of Y for indices V, then according to Lemma C.4.12, we have $X_2 \in \mathcal{X}(Y; U, [n], n)$.

Applying Theorem C.4.10 again, there exist an imputation $Y_2 \in \mathcal{X}(X_2; V, [n], n)$ of Y' for indices V such that $D(X_2, Y') \leq D(X_2, Y_2)$. Hence, we have $D(X', Y') \leq D(X_1, Y_2)$.

Since Y' is an imputation of Y for V, then according to the definition of imputations, we can show that Y_2 is also an imputation of Y for V. Meanwhile, notice that X_2 is an imputation of X for indices U, then according to Lemma C.4.12, we have $Y_2 \in \mathcal{X}(X; V, [n], n)$.

Then, according to the definition of (X_*, Y_*), we have $D(X_2, Y_2) \leq D(X_*, Y_*)$. Hence, we have $D(X', Y') \leq D(X_*, Y_*)$. This completes our proof.

□

If we denote $m_1 = |U|$ and $m_2 = |V|$, then the number of possible rank vectors of imputations $X_* \in \mathcal{X}(Y; U, [n], n)$ is $m_1 + 1$ and the number of possible rank vectors of imputations $Y_* \in \mathcal{X}(X; V, [n], n)$ is $m_2 + 1$, following reasoning similar to that in Remark C.4.11 for missing case I. Hence, the number of possible combinations of rank vectors of imputations $(X_*, Y_*) \in (\mathcal{X}(Y; U, [n], n), \mathcal{X}(X; V, [n], n))$ is $(m_1 + 1)(m_2 + 1)$.

C.4.3. Upper bound under missing case III

This section provides the upper bound of Spearman's footrule under missing case III, where for each pair $(X(i), Y(i))$, the two values are either *both* observed or *both* missing.

To start, we consider the case where there is only one pair $(X(u), Y(u))$ where both values are missing. The following result is proved by applying Lemma C.4.1 and Proposition C.4.2.

Proposition C.4.14. Suppose $X, Y \in \mathcal{V}^n$, and let $u \in [n]$. Suppose $X_1, Y_1 \in \mathcal{V}^n$ and $X_2, Y_2 \in \mathcal{V}^n$ are imputations of X and Y for the index u such that

$$R(X_1(u_1), X_1) = 1, \text{ and } R(Y_1(u_1), Y_1) = n,$$
$$R(X_2(u_1), X_2) = n, \text{ and } R(Y_2(u_1), Y_2) = 1.$$

Then, $D(X, Y) \leq \max\{D(X_1, Y_1), D(X_2, Y_2)\}$.

Proof. To start, since $R(Y_2(u), Y_2) = 1$, then we have $R(X(u), X) \geq R(Y_2(u), Y_2)$. Hence, according to Lemma C.4.1, we have $D(X, Y_2) \leq D(X_2, Y_2)$.

Similarly, since $R(Y_1(u), Y_1) = n$, we have $R(X(u), X) \leq R(Y_1(u), Y_1)$. Hence, according to Lemma C.4.1, we have $D(X, Y_1) \leq D(X_1, Y_1)$.

Hence, we have $\max\{D(X,Y_1), D(X,Y_2)\} \leq \max\{D(X_1,Y_1), D(X_2,Y_2)\}$. According to Proposition C.4.2, we have $D(X,Y) \leq \max\{D(X,Y_1), D(X,Y_2)\}$. Thus, we further have $D(X,Y) \leq \max\{D(X_1,Y_1), D(X_2,Y_2)\}$. This completes our proof. □

When one pair $(X(u), Y(u))$ is missing, Proposition C.4.14 gives similar results as in Proposition C.4.2, offering two possibilities for determining the maximum possible value of Spearman's footrule.

Furthermore, to achieve the maximum possible Spearman's footrule, Proposition C.4.14 requires the paired ranks of the imputations for the index u to be at opposite ends of the sequence $1, \ldots, n$. For example, if the imputed rank of $X(u_1)$ is 1, then the imputed rank of its paired component must be n; conversely, if the imputed rank of $X(u_1)$ is n, then the rank of its paired component must be 1. Let X_* and Y_* be imputations of X and Y, respectively. This condition can be expressed as

$$R(X_*(u), X_*) + R(Y_*(u), Y_*) = n + 1.$$

It will be useful to define a set of indices for which this condition is satisfied.

Definition C.4.15. Let $W \subset [n]$ be a subset of indices. Define $\mathcal{Y}(W,n)$ as

$$\mathcal{Y}(W,n) = \{(Z_1,Z_2) \in (\mathcal{Z}(W,[n],n), \mathcal{V}^n) | R(Z_1(i),Z_1) + R(Z_2(i),Z_2) = n+1,\ i \in W\}.$$

Using Definition C.4.15, we can state Theorem C.4.18, which concludes the upper bounds of Spearman's footrule under missing case III. For proving the theorem, we will first need to prove the following two lemmas.

Lemma C.4.16. Suppose $X_1 \in \mathcal{V}^n$, and for $2 \leq m < n$, denote $W' = \{n-m+1, \ldots, n-1\}$ as a subset of indices. Suppose $X_1 \in \mathcal{Z}(W', [n], n)$, and define

$$W_1' = \left\{ i \in W' : X_1(i) > \max_{l \in [n] \setminus W'} X_1(l) \right\}, W_2' = \left\{ i \in W' : X_1(i) < \min_{l \in [n] \setminus W'} X_1(l) \right\}.$$

Suppose X_2 is an imputation of X_1 for the index n. Then if $R(X_2(n), X_2) = n$, we have $R(X_2(i), X_2) = R(X_1(i), X_1) - 1$, for any $i \in W_1'$, and $R(X_2(i), X_2) = R(X_1(i), X_1)$, for any $i \in W_2'$. However, if $R(X_2(n), X_2) = 1$, we have $R(X_2(i), X_2) = R(X_1(i), X_1)$, for any $i \in W_1'$, and $R(X_2(i), X_2) = R(X_1(i), X_1) + 1$, for any $i \in W_2'$.

Proof. First, we show that if $R(X_2(n), X_2) = n$, then we have $R(X_2(i), X_2) = R(X_1(i), X_1) - 1$, for any $i \in W_1'$, and $R(X_2(i), X_2) = R(X_1(i), X_1)$, for any $i \in W_2'$.

According to the definition of rank, for any $i \in W'$, we have

$$\begin{aligned} R(X_2(i), X_2) &= \sum_{j=1}^{n} \mathrm{I}(X_2(j) \leq X_2(i)) \\ &= \sum_{j=1}^{n-1} \mathrm{I}(X_2(j) \leq X_2(i)) + \mathrm{I}(X_2(n) \leq X_2(i)). \end{aligned}$$

Since $R(X_2(n), X_2) = n$, we have $\mathrm{I}\left(X_2(n) \leq X_2(i)\right) = 0$, for any $i \in W'$. Then, for any $i \in W'$, we have

$$R(X_2(i), X_2) = \sum_{j=1}^{n-1} \mathrm{I}\left(X_2(j) \leq X_2(i)\right) = R(X_2(i), (X_2(l))_{l\in[n-1]}). \tag{C.148}$$

Since X_2 is an imputation of X_1 for the index n, then we have $X_2(i) = X_1(i)$ for any $i \in [n] \setminus \{n\} = [n-1]$. Hence, $(X_2(l))_{l\in[n-1]}$ is a permutation of $(X_1(l))_{l\in[n-1]}$. Meanwhile, since $W' = \{n-m+1, \ldots, n-1\} \subset [n-1]$, we have $X_2(i) = X_1(i)$, for any $i \in W'$. Then, it follows that

$$R(X_2(i), (X_2(l))_{l\in[n-1]}) = R(X_1(i), (X_1(l))_{l\in[n-1]}), \text{ for any } i \in W'. \tag{C.149}$$

Using the definition of ranks and notice that $X_1 \in \mathcal{V}^n$ is a vector of distinct real values, we can show that

$$R(X_1(i), (X_1(l))_{l\in[n-1]}) = R(X_1(i), X_1) - \mathrm{I}\left(X_1(n) < X_1(i)\right), \text{ for any } i \in W'.$$

If $i \in W_1'$, then according to the definition of W_1', we have $X_1(i) > \max_{l\in[n]\setminus W'} X_1(l)$. Since $n \in [n] \setminus W'$, we have $X_1(i) > X_1(n)$. Hence, we have

$$R(X_1(u_1), (X_1(l))_{l\in[n-1]}) = R(X_1(u_1), X_1) - 1, \text{ for any } i \in W_1'. \tag{C.150}$$

However, if $i \in W_2'$, according to the definition of W_2', we have $X_1(i) < \min_{l\in[n]\setminus W'} X_1(l)$. Since $n \in [n] \setminus W'$, we have $X_1(i) < X_1(n)$. Hence, we have

$$R(X_1(u_1), (X_1(l))_{l\in[n-1]}) = R(X_1(u_1), X_1), \text{ for any } i \in W_2'. \tag{C.151}$$

Thus, for any $i \in W_1'$, we have

$$\begin{aligned} R(X_2(i), X_2) &=^{(\mathrm{C.148})} R(X_2(i), (X_2(l))_{l\in[n-1]}) \\ &=^{(\mathrm{C.149})} R(X_1(i), (X_1(l))_{l\in[n-1]}) \\ &=^{(\mathrm{C.150})} R(X_1(i), X_1) - 1. \end{aligned}$$

Similarly, for any $i \in W_1'$, we have

$$\begin{aligned} R(X_2(i), X_2) &=^{(\mathrm{C.148})} R(X_2(i), (X_2(l))_{l\in[n-1]}) \\ &=^{(\mathrm{C.149})} R(X_1(i), (X_1(l))_{l\in[n-1]}) \\ &=^{(\mathrm{C.151})} R(X_1(i), X_1). \end{aligned}$$

Hence, we have shown that when $R(X_2(n), X_2) = n$, we have $R(X_2(i), X_2) = R(X_1(i), X_1) - 1$, for any $i \in W_1'$, and $R(X_2(i), X_2) = R(X_1(i), X_1)$, for any $i \in W_2'$.

Similarly, we can show that if $R(X_2(n), X_2) = 1$, we have $R(X_2(i), X_2) = R(X_1(i), X_1)$, for any $i \in W_1'$, and $R(X_2(i), X_2) = R(X_1(i), X_1) + 1$, for any $i \in W_2'$.

According to the definition of rank, for any $i \in W$, we have

$$\begin{aligned} R(X_2(i), X_2) &= \sum_{j=1}^{n} \mathrm{I}\left(X_2(j) \leq X_2(i)\right) \\ &= \sum_{j=1}^{n-1} \mathrm{I}\left(X_2(j) \leq X_2(i)\right) + \mathrm{I}\left(X_2(n) \leq X_2(i)\right). \end{aligned}$$

Since $R(X_2(n), X_2) = 1$, we have $\mathrm{I}\left(X_2(n) \leq X_2(i)\right) = 1$, for any $i \in W'$. Then, we have

$$\begin{aligned} R(X_2(u_1), X_2) &= \sum_{i=1}^{n-1} \mathrm{I}\left(X_2(i) \leq X_2(u_1)\right) + 1 \\ &= R(X_2(u_1), (X_2(l))_{l\in[n-1]}) + 1, \text{ for any } i \in W'. \end{aligned} \tag{C.152}$$

Since X_2 is an imputation of X_1 for the index n, then we have $X_2(i) = X_1(i)$ for any $i \in [n] \setminus \{n\} = [n-1]$. Hence, $(X_2(l))_{l\in[n-1]}$ is a permutation of $(X_1(l))_{l\in[n-1]}$. Meanwhile, since $W' = \{n-m+1, \ldots, n-1\} \subset [n-1]$, we have $X_2(i) = X_1(i)$, for any $i \in W'$. Then, it follows that

$$R(X_2(i), (X_2(l))_{l\in[n-1]}) = R(X_1(i), (X_1(l))_{l\in[n-1]}), \text{ for any } i \in W'. \tag{C.153}$$

Using the definition of ranks and notice that $X_2 \in \mathcal{V}^n$ is a vector of distinct real values, we can show that

$$R(X_1(i), (X_1(l))_{l\in[n-1]}) = R(X_1(i), X_1) - \mathrm{I}\left(X_1(n) < X_1(i)\right), \text{ for any } i \in W'.$$

If $i \in W_1'$, then according to the definition of W_1', we have $X_1(i) > \max_{l\in[n]\setminus W'} X_1(l)$. Since $n \in [n] \setminus W'$, we have $X_1(i) > X_1(n)$. Hence, we have

$$R(X_1(u_1), (X_1(l))_{l\in[n-1]}) = R(X_1(u_1), X_1) - 1, \text{ for any } i \in W_1'. \tag{C.154}$$

However, if $i \in W_2'$, then according to the definition of W_2', we have $X_1(i) < \min_{l\in[n]\setminus W'} X_1(l)$. Since $n \in [n] \setminus W'$, we have $X_1(i) < X_1(n)$. Hence, we have

$$R(X_1(u_1), (X_1(l))_{l\in[n-1]}) = R(X_1(u_1), X_1), \text{ for any } i \in W_2'. \tag{C.155}$$

Thus, for any $i \in W_1'$, we have

$$\begin{aligned} R(X_2(i), X_2) &=^{(\text{C.152})} R(X_2(i), (X_2(l))_{l\in[n-1]}) + 1 \\ &=^{(\text{C.153})} R(X_1(i), (X_1(l))_{l\in[n-1]}) + 1 \\ &=^{(\text{C.154})} R(X_1(i), X_1). \end{aligned}$$

Similarly, for any $i \in W_2'$, we have

$$\begin{aligned} R(X_2(i), X_2) &=^{(\text{C.152})} R(X_2(i), (X_2(l))_{l\in[n-1]}) + 1 \\ &=^{(\text{C.153})} R(X_1(i), (X_1(l))_{l\in[n-1]}) + 1 \\ &=^{(\text{C.154})} R(X_1(i), X_1) + 1. \end{aligned}$$

Hence, we have shown that if $R(X_2(n), X_2) = 1$, we have $R(X_2(i), X_2) = R(X_1(i), X_1)$, for any $i \in W_1'$, and $R(X_2(i), X_2) = R(X_1(i), X_1) + 1$, for any $i \in W_2'$. This completes our proof. □

Lemma C.4.17. Suppose $X, Y \in \mathcal{V}^n$, and for $2 \leq m < n$ assume $W = \{n-m+1, \ldots, n\}$, and $W' = \{n-m+1, \ldots, n-1\}$. Suppose $(X_1, Y_1) \in \mathcal{Y}(W', n)$ are imputations of X and Y for indices W', and $(X_2, Y_2) \in \mathcal{Y}(\{n\}, n)$ are imputations of X_1 and Y_1 for the index n. Then X_2, Y_2 are imputations of X and Y for indices W such that $(X_2, Y_2) \in \mathcal{Y}(W, n)$.

Proof. First, we show that X_2, Y_2 are imputations of X and Y for indices W. Since X_1, Y_1 are imputations of X and Y for indices W', then according to the definitions of imputations, we have $X_1(i) = X(i)$ and $Y_1(i) = Y(i)$ for any $i \in [n] \setminus W'$.

Since $W' \subset W$, we have $[n] \setminus W \subset [n] \setminus W'$. Hence, we have

$$X_1(i) = X(i), \text{ and } Y_1(i) = Y(i), \text{ for any } i \in [n] \setminus W.$$

Since X_2, Y_2 are imputations of X_1 and Y_1 for the index n, then we have $X_2(i) = X_1(i)$ and $Y_2(i) = Y_1(i)$ for any $i \in [n] \setminus \{n\}$. Since $\{n\} \subset W$, then we have $[n] \setminus W \subset [n] \setminus \{n\}$. Hence, we have

$$\begin{aligned} &X_2(i) = X_1(i), \text{ and } Y_2(i) = Y_1(i), \text{ for any } i \in [n] \setminus W \\ \Longrightarrow\ &X_2(i) = X(i), \text{ and } Y_2(i) = Y(i), \text{ for any } i \in [n] \setminus W. \end{aligned}$$

This proves that X_2, Y_2 are imputations of X and Y for indices W.

Now, we show that $(X_2, Y_2) \in \mathcal{Y}(W, n)$. To start, since $(X_1, Y_1) \in \mathcal{Y}(W', n)$, then we have $X_1 \in \mathcal{Z}(W', [n], n)$. Similarly, since $(X_2, Y_2) \in \mathcal{Y}(\{n\}, n)$, we have $X_2 \in \mathcal{Z}(\{n\}, [n], n)$. Then, according to Lemma C.4.4, we have $X_2 \in \mathcal{Z}(W, [n], n)$. Hence, we have $(X_2, Y_2) \in (\mathcal{Z}(W, [n], n), \mathcal{V}^n)$. Then, in order to show that $(X_2, Y_2) \in \mathcal{Y}(W, n)$, it is sufficient to show that

$$R(X_2(i), X_2) + R(Y_2(i), Y_2) = n + 1, \text{ for any } i \in W.$$

Since $(X_2, Y_2) \in \mathcal{Y}(\{n\}, n)$, we have

$$R(X_2(n), X_2) + R(Y_2(n), Y_2) = n + 1. \tag{C.156}$$

Then, we only need to show that

$$R(X_2(i), X_2) + R(Y_2(i), Y_2) = n + 1, \text{ for any } i \in W'. \tag{C.157}$$

Below, we prove (C.157) is true.

Since $X_2 \in \mathcal{Z}(\{n\}, [n], n)$, then we have

$$X_2(n) > \max_{l \in [n-1]} X_2(l), \text{ or } X_2(n) < \min_{l \in [n-1]} X_2(l).$$

Hence, we have either

$$R(X_2(n), X_2) = n, \text{ or } R(X_2(n), X_2) = 1.$$

Combining with (C.156), we have either

$$\begin{aligned} \text{case (1)}: \ & R(X_2(n), X_2) = n, \text{ and } R(Y_2(n), Y_2) = 1, \\ \text{or case (2)}: \ & R(X_2(n), X_2) = 1, \text{ and } R(Y_2(n), Y_2) = n, \end{aligned}$$

is true.

Suppose case (1) is true. In other words, $R(X_2(n), X_2) = n$, and $R(Y_2(n), Y_2) = 1$.

Since $X_1 \in \mathcal{Z}(W', [n], n)$, then we have

$$X_1(i) > \max_{l \in [n] \setminus W'} X_1(l), \text{ or } X_1(i) < \min_{l \in [n] \setminus W'} X_1(l), \text{ for any } i \in W'.$$

Let us define

$$\begin{aligned} W_1' &= \left\{ i \in W' : X_1(i) > \max_{l \in [n] \setminus W'} X_1(l) \right\}, \\ \text{and } W_2' &= \left\{ i \in W' : X_1(i) < \min_{l \in [n] \setminus W'} X_1(l) \right\}. \end{aligned}$$

Then, we have $W_1' \cup W_2' = W'$. Denote $n_1' = |W_1'|$ and $n_2' = |W_2'|$.

Next, since X_2 is an imputation of $X_1 \in \mathcal{Z}(W', [n], n)$ such that $R(X_2(n), X_2) = n$. Then according to Lemma C.4.16, we have

$$R(X_2(i), X_2) = R(X_1(i), X_1) - 1, \text{ for any } i \in W_1', \tag{C.158}$$

$$\text{and } R(X_2(i), X_2) = R(X_1(i), X_1), \text{ for any } i \in W_2'. \tag{C.159}$$

For any $i \in W_1'$, since $X_1(i) > \max_{l \in [n] \setminus W'} X_1(l)$, then we have

$$R(X_1(i), X_1) \in \{n, \ldots, n - n_1' + 1\}, \text{ for any } i \in W_1'.$$

Further, since $(X_1, Y_1) \in \mathcal{Y}(W', n)$, then $R(X_1(i), X_1) + R(Y_1(i), Y_1) = n + 1$ for any $i \in W'$. Hence, we have

$$\begin{aligned} & R(Y_1(i), X_1) \in \{1, \ldots, n_1'\}, \text{ for any } i \in W_1' \\ \Longrightarrow\ & Y_1(i) < \min_{l \in [n] \setminus W'} Y_1(l), \text{ for any } i \in W_1'. \end{aligned} \tag{C.160}$$

Similarly, for any $i \in W_2'$, since $X_1(i) < \min_{l\in[n]\setminus W'} X_1(l)$, then we have

$$R(X_1(i), X_1) \in \{1, \ldots, n_2'\}, \text{ for any } i \in W_2'.$$

Further, since $(X_1, Y_1) \in \mathcal{Y}(W', n)$, then $R(X_1(i), X_1) + R(Y_1(i), Y_1) = n + 1$ for any $i \in W'$. Hence, we have

$$\begin{aligned} & R(Y_1(i), X_1) \in \{n, \ldots, n - n_2' + 1\}, \text{ for any } i \in W_2' \\ \Longrightarrow\ & Y_1(i) > \max_{l\in[n]\setminus W'} Y_1(l), \text{ for any } i \in W_2'. \end{aligned} \tag{C.161}$$

Combining (C.160) and (C.161), we have $Y_1 \in \mathcal{Z}(W', [n], n)$. Notice that Y_2 is an imputation of $Y_1 \in \mathcal{Z}(W', [n], n)$ for the index n such that $R(Y_2(n), X_2) = 1$. Then, according to Lemma C.4.16, we have

$$R(Y_2(i), Y_2) = R(Y_1(i), Y_1) + 1, \text{ for any } i \in W_1', \tag{C.162}$$

$$\text{and } R(Y_2(i), Y_2) = R(Y_1(i), Y_1), \text{ for any } i \in W_2'. \tag{C.163}$$

Combining (C.158), (C.159), (C.162) and (C.162), for any $i \in W'$, we have

$$R(X_2(i), X_2) + R(Y_2(i), Y_2) = R(X_1(i), X_1) + R(Y_1(i), Y_1).$$

Since $(X_1, Y_1) \in \mathcal{Y}(W', n)$, then for any $i \in W'$, we have

$$R(X_1(i), X_1) + R(Y_1(i), Y_1) = n + 1.$$

Hence, for any $i \in W'$, we have

$$R(X_2(i), X_2) + R(Y_1(i), Y_1) = n + 1.$$

This proves (C.157) when the case (1) is true.

When the case (2) is true, (C.157) can be proved similarly. Suppose the case (2) is true. In other words, suppose $R(X_2(n), X_2) = 1$, and $R(Y_2(n), Y_2) = n$.

Since $X_1 \in \mathcal{Z}(W', [n], n)$, then we have

$$X_1(i) > \max_{l\in[n]\setminus W'} X_1(l), \text{ or } X_1(i) < \min_{l\in[n]\setminus W'} X_1(l), \text{ for any } i \in W'.$$

Let us define

$$\begin{aligned} W_1' &= \left\{ i \in W' : X_1(i) > \max_{l\in[n]\setminus W'} X_1(l) \right\}, \\ \text{and } W_2' &= \left\{ i \in W' : X_1(i) < \min_{l\in[n]\setminus W'} X_1(l) \right\}. \end{aligned}$$

Then, we have $W_1' \cup W_2' = W'$. Denote $n_1' = |W_1'|$ and $n_2' = |W_2'|$.

Next, since X_2 is an imputation of $X_1 \in \mathcal{Z}(W', [n], n)$ such that $R(X_2(n), X_2) = 1$. Then

according to Lemma C.4.16, we have

$$R(X_2(i), X_2) = R(X_1(i), X_1), \text{ for any } i \in W_1', \tag{C.164}$$
$$\text{and } R(X_2(i), X_2) = R(X_1(i), X_1) + 1, \text{ for any } i \in W_2'. \tag{C.165}$$

For any $i \in W_1'$, since $X_1(i) > \max_{l \in [n] \setminus W'} X_1(l)$, then we have

$$R(X_1(i), X_1) \in \{n, \ldots, n - n_1' + 1\}, \text{ for any } i \in W_1'.$$

Further, since $(X_1, Y_1) \in \mathcal{Y}(W', n)$, then $R(X_1(i), X_1) + R(Y_1(i), Y_1) = n + 1$ for any $i \in W'$. Hence, we have

$$\begin{aligned} & R(Y_1(i), X_1) \in \{1, \ldots, n_1'\}, \text{ for any } i \in W_1' \\ \Longrightarrow\ & Y_1(i) < \min_{l \in [n] \setminus W'} Y_1(l), \text{ for any } i \in W_1'. \end{aligned} \tag{C.166}$$

Similarly, for any $i \in W_2'$, since $X_1(i) < \min_{l \in [n] \setminus W'} X_1(l)$, then we have

$$R(X_1(i), X_1) \in \{1, \ldots, n_2'\}, \text{ for any } i \in W_2'.$$

Further, since $(X_1, Y_1) \in \mathcal{Y}(W', n)$, then $R(X_1(i), X_1) + R(Y_1(i), Y_1) = n + 1$ for any $i \in W'$. Hence, we have

$$\begin{aligned} & R(Y_1(i), X_1) \in \{n, \ldots, n - n_2' + 1\}, \text{ for any } i \in W_2' \\ \Longrightarrow\ & Y_1(i) > \max_{l \in [n] \setminus W'} Y_1(l), \text{ for any } i \in W_2'. \end{aligned} \tag{C.167}$$

Combining (C.166) and (C.167), we have $Y_1 \in \mathcal{Z}(W', [n], n)$. Notice that Y_2 is an imputation of $Y_1 \in \mathcal{Z}(W', [n], n)$ such that $R(Y_2(n), X_2) = n$. Then, according to Lemma C.4.16, we have

$$R(Y_2(i), Y_2) = R(Y_1(i), Y_1), \text{ for any } i \in W_1', \tag{C.168}$$
$$\text{and } R(Y_2(i), Y_2) = R(Y_1(i), Y_1) - 1, \text{ for any } i \in W_2'. \tag{C.169}$$

Combining (C.164), (C.165), (C.168) and (C.169), for any $i \in W'$, we have

$$R(X_2(i), X_2) + R(Y_2(i), Y_2) = R(X_1(i), X_1) + R(Y_1(i), Y_1).$$

Since $(X_1, Y_1) \in \mathcal{Y}(W', n)$, then for any $i \in W'$, we have

$$R(X_1(i), X_1) + R(Y_1(i), Y_1) = n + 1.$$

Hence, for any $i \in W'$, we have

$$R(X_2(i), X_2) + R(Y_1(i), Y_1) = n + 1.$$

This proves (C.157) when the case (2) is true. Hence, we complete our proof.

□

We are now ready to prove Theorem C.4.18.

Theorem C.4.18. Suppose $X, Y \in \mathcal{V}^n$ and $W \subset [n]$. Then there exist imputations $(X_*, Y_*) \in \mathcal{Y}(W, n)$ of X and Y for indices W such that $D(X, Y) \leq D(X_*, Y_*)$. Furthermore, consider any other imputations $X', Y' \in \mathcal{V}^n$ of X, Y for indices W. Then $D(X', Y') \leq D(X_*, Y_*)$.

Proof. To prove Theorem C.4.18, we first prove the following statement

$\mathbb{S}$: Suppose $X, Y \in \mathcal{V}^n$ and $W \subset [n]$. Then there exist imputations $(X_*, Y_*) \in \mathcal{Y}(W, n)$ of X and Y for indices W such that $D(X, Y) \leq D(X_*, Y_*)$.

Then we will prove Theorem C.4.18 using statement $\mathbb{S}$.

First, we prove the statement $\mathbb{S}$ is true. Suppose $W = \emptyset$. Then according to the definition of imputations, we have $X_* = X$ and $Y_* = Y$. Hence, we have $D(X, Y) = D(X_*, Y_*)$. This proves the statement $\mathbb{S}$ when $W = \emptyset$.

Suppose $W = [n]$. Then, let $X_* \in \mathcal{V}^n$ be a vector such that $X_*(1) > \ldots > X_*(n)$. In other words, $X_* \in \mathcal{V}^n$ is a vector such that

$$R(X_*(i), X_*) = n + 1 - i, \text{ for any } i \in [n].$$

Let $Y_* \in \mathcal{V}^n$ be a vector such that $Y_*(1) < \ldots < Y_*(n)$. In other words, $Y_* \in \mathcal{V}^n$ is a vector such that

$$R(Y_*(i), Y_*) = i, \text{ for any } i \in [n].$$

Hence, we have

$$R(X_*(i), X_*) + R(Y_*(i), Y_*) = n + 1, \text{ for any } i \in [n].$$

Thus, X_*, Y_* are imputations of X, Y for $W = [n]$, and $(X_*, Y_*) \in \mathcal{Y}(W, n)$.

Next, let us assume (after relabelling) $Y_1 < \ldots < Y_n$. Then according to the definition of Spearman's footrule, we have

$$\begin{aligned} D(X, Y) &= \sum_{i=1}^{n} |R(X(i), X) - R(Y(i), Y)| \\ &= \sum_{i=1}^{n} |R(X(i), X) - i|. \end{aligned}$$

Since $Y_*(1) < \ldots < Y_*(n)$, we have

$$\begin{aligned} D(X_*, Y_*) &= \sum_{i=1}^{n} |R(X_*(i), X_*) - R(Y_*(i), Y_*)| \\ &= \sum_{i=1}^{n} |R(X_*(i), X_*) - i|. \end{aligned}$$

Notice that $(R(X(i), X))_{i \in [n]}$ and $(R(X_*(i), X_*))_{i \in [n]}$ are both permutations of $[n]$. Then, we have that $(R(X_*(i), X_*))_{i \in [n]}$ is a permutation of $(R(X(i), X))_{i \in [n]}$. Since $X_*(1) > \ldots > X_*(n)$,

then we have $R(X_*(1), X_*) > \ldots > R(X_*(n), X_*)$. Hence, according to Lemma C.4.6, we have

$$D(X_*, Y_*) \geq D(X, Y).$$

This proves the statement $\mathbb{S}$ is true when $W = [n]$.

In the following, we prove the statement $\mathbb{S}$ is true when $W \neq \emptyset$ and $W \neq [n]$.

For any given $n \in \mathbb{N}$, let $P(k)$ be the statement $\mathbb{S}$ when $|W| = k$. We prove $P(k)$ is true for any $k \in \{1, \ldots, n-1\}$ by induction on k.

Base Case: We prove $P(1)$ is true. When $|W| = 1$, let us denote $W = u$. Suppose $X_1, Y_1 \in \mathcal{V}^n$ and $X_2, Y_2 \in \mathcal{V}^n$ are imputations of X and Y for the index u such that

$$R(X_1(u), X_1) = 1, \text{ and } R(Y_1(u), Y_1) = n,$$
$$R(X_2(u), X_2) = n, \text{ and } R(Y_2(u), Y_2) = 1.$$

Then, according to Proposition C.4.14, we have $D(X, Y) \leq \max\{D(X_1, Y_1), D(X_2, Y_2)\}$.

Since $R(X_1(u), X_1) = 1$, we have $X_1(u) < X_1(i)$ for any $i \in [n] \setminus \{u\}$. Notice that $W = \{u\}$. Then, we have $X_1 \in \mathcal{Z}(W, [n], n)$. Since $R(X_1(u), X_1) + R(Y_1(u), Y_1) = n + 1$, we have $(X_1, Y_1) \in \mathcal{Y}(W, n)$.

Similarly, since $R(X_2(u), X_2) = n$, we have $X_2(u) > X_2(i)$ for any $i \in [n] \setminus \{u\}$. Notice that $W = \{u\}$. Then, we have $X_2 \in \mathcal{Z}(W, [n], n)$. Since $R(X_2(u), X_2) + R(Y_2(u), Y_2) = n + 1$, we have $(X_2, Y_2) \in \mathcal{Y}(W, n)$.

Notice that $D(X, Y) \leq \max\{D(X_1, Y_1), D(X_2, Y_2)\}$, $(X_1, Y_1) \in \mathcal{Y}(W, n)$ and $(X_2, Y_2) \in \mathcal{Y}(W, n)$. Hence, we have shown $P(1)$ is true.

Induction Step: We show the implication $P(1), P(k) \implies P(k+1)$ for any $k \in \{1, \ldots, n-2\}$.

Denote $m = k + 1$. Without loss of generality, let us assume (after relabelling) $W = \{n - m + 1, \cdots, n\}$. Define $W' = \{n - m + 1, \cdots, n - 1\}$.

Notice that $|W'| = |W| - 1 = k$. Then, since $P(k)$ is true, there exist imputations $(X_1, Y_1) \in \mathcal{Y}(W', n)$ of X and Y for indices W' such that $D(X, Y) \leq D(X_1, Y_1)$.

Next, since $P(1)$ is true, then there exist imputations $(X_2, Y_2) \in \mathcal{Y}(\{n\}, n)$ of X_1 and Y_1 for the index n such that $D(X_1, Y_1) \leq D(X_2, Y_2)$.

Hence, we have $D(X, Y) \leq D(X_2, Y_2)$. According to Lemma C.4.17, X_2, Y_2 are imputations of X and Y for indices W such that $(X_2, Y_2) \in \mathcal{Y}(W, n)$. Hence, we have shown $P(k+1)$ is true. This completes our proof for the statement $\mathbb{S}$.

Next, we prove Theorem C.4.18 is true using the statement $\mathbb{S}$. let us denote

$$S = \{D(V_1, V_2) : (V_1, V_2) \in \mathcal{Y}(W, n) \text{ are imputations of } X, Y \text{ for indices } W\}.$$

Then, the carnality of S is finite and there exist imputations $(X_*, Y_*) \in \mathcal{Y}(W, n)$ of X, Y for indices W such that $D(X_*, Y_*) = \max S$.

We now show that $D(X, Y) \leq D(X_*, Y_*)$. According to statement $\mathbb{S}$, there exist imputations $(X_1, Y_1) \in \mathcal{Y}(W, n)$ of X, Y for indices W such that $D(X, Y) \leq D(X_1, Y_1)$. According to the definition of S, we have $D(X_1, Y_1) \in S$. Hence, we have $D(X_1, Y_1) \leq \max S = D(X_*, Y_*)$. Thus, we have $D(X, Y) \leq D(X_*, Y_*)$.

Next, we show that $D(X', Y') \leq D(X_*, Y_*)$. According to the statement $\mathbb{S}$, there exist im-

putation $(X_2, Y_2) \in \mathcal{Y}(W, n)$ of X', Y' for indices W such that $D(X', Y') \leq D(X_2, Y_2)$. Since X', Y' are imputations of X and Y for indices W, then X_2, Y_2 are imputations of X, Y for indices W. Then, according to the definition of S, we have $D(X_2, Y_2) \in S$. Hence, we have $D(X_2, Y_2) \leq \max S = D(X_*, Y_*)$. Thus, we have $D(X', Y') \leq D(X_*, Y_*)$. This completes our proof. □

Theorem C.4.18 establishes the existence of imputations $(X_*, Y_*) \in \mathcal{Y}(W, n)$ of X and Y that will maximise Spearman's footrule $D(X, Y)$. In order to find this pair (X_*, Y_*), we could enumerate all possible imputations. If $X_*, Y_* \in \mathbb{R}$, there would be uncountable many imputations. But, as already mentioned, we only need to consider X_* and Y_* to be *rank* vectors, since the computation of Spearman's footrule only uses ranks. Therefore, the problem is reduced to only considering rank vector imputations, and these can be enumerated. However, different rank vector imputations may lead to the same value of Spearman's footrule statistic. Therefore, we will only count the number of possible *distinct values* of Spearman's footrule statistic using (ranked) imputations $(X_*, Y_*) \in \mathcal{Y}(W, n)$.

For any imputations $(X_*, Y_*) \in \mathcal{Y}(W, n)$ of X and Y for indices W, we have $X_* \in \mathcal{Z}(W, [n], n)$ using Definition 5.1.4. Therefore, according to Definition 5.1.2 of $\mathcal{Z}(W, [n], n)$, either $X_*(i) > \max_{j \in [n] \setminus W} X_*(j)$ or $X_*(i) < \min_{j \in [n] \setminus W} X_*(j)$, for any $i \in W$.

Denote $m_3 = |W|$ and $r = \sum_{i \in W} \mathrm{I}\left(X_*(i) < (X_*(l))_{l \in [n] \setminus W}\right)$, i.e. the number of components of X_* with indices in W smaller than all components of X_* with indices in $[n] \setminus W$. Then, there are $m_3 - r$ number of components of X_* with indices in W that are larger than all components of X_* with indices in $[n] \setminus W$.

In fact, any imputations $(X_*, Y_*) \in \mathcal{Y}(W, n)$ of X and Y for indices W with the same value of r give the same value of Spearman's footrule. For the moment, let $r \in \{1, 2, \ldots, m_3 - 1\}$. The r number of components of X_* with indices in W smaller than all components of X_* with indices in $[n] \setminus W$ take the ranks between between 1 and r, and the $m_3 - r$ components of X_* with indices in W larger than all components of X_* with indices in $[n] \setminus W$ take the ranks between $n - (m_3 - r) + 1$ and n. According to the definition of $\mathcal{Y}(W, n)$, the ranks of components of Y_* in W are decided by the equation $R(X_*(i), X_*) + R(Y_*(i), Y_*) = n + 1$. Hence, without changing the values of Spearman's footrule, we can reorder the pairs of components of X_* and Y_* in W, so that the components of X_* in W are arranged from small to large as $1, \ldots, r, n - (m_3 - r) + 1, \ldots, n$, and then the components of Y_* in W are arranged as $n, \ldots, n + 1 - r, m_3 - r, 1$.

In the preceding paragraph, we considered $r \in \{1, 2, \ldots, m_3 - 1\}$, but it is also possible that $r = 0$ or $r = m_3$, in which case either all components of X_* with indices in W are smaller or larger than all other components, which can be considered similarly.

Since r ranges between 0 to m_3, there are at most $m_3 + 1$ possible values of Spearman's footrule statistic, given by all possible imputations $(X_*, Y_*) \in \mathcal{Y}(W, n)$.

The above procedure shows how to construct imputations (X_*, Y_*) for each of these possible $m_3 + 1$ values of Spearman's footrule. This is detailed in Algorithm 7 in Section C.5.1.

C.4.4. Upper bound under general missingness case

This section considers the upper bound of Spearman's footrule under the general missingness case, where we do not make any assumptions about the missingness pattern. The previous results for

missing case I, II, and III can be combined to give the following result.

Theorem C.4.19. Suppose $X, Y \in \mathcal{V}^n$, and let U, V, $W \subset [n]$ be pairwise disjoint subsets. Then, there exist $(X_*, Y_*) \in (\mathcal{X}(Y; U, [n] \setminus W, n), \mathcal{X}(X; V, [n] \setminus W, n)) \cap \mathcal{Y}(W, n)$ of imputations X and Y for indices $U \cup W$ and $V \cup W$, respectively, such that $D(X, Y) \leq D(X_*, Y_*)$. Furthermore, consider any imputation $X', Y' \in \mathcal{V}^n$ of X, Y for indices $U \cup W$ and $V \cup W$, respectively, we have $D(X', Y') \leq D(X_*, Y_*)$.

Proof. To prove Theorem C.4.19, we first prove the following statement

$\mathbb{S}$: Suppose $X, Y \in \mathcal{V}^n$, and let U, V, $W \subset [n]$ be pairwise disjoint subsets. Then, there exist $(X_*, Y_*) \in (\mathcal{X}(Y; U, [n] \setminus W, n), \mathcal{X}(X; V, [n] \setminus W, n)) \cap \mathcal{Y}(W, n)$ of imputations X and Y for indices $U \cup W$ and $V \cup W$, respectively, such that $D(X, Y) \leq D(X_*, Y_*)$.

Then we will prove Theorem C.4.19 using the statement $\mathbb{S}$.

According to Theorem C.4.13, there exist imputations

$$(X_1, Y_1) \in (\mathcal{X}(Y; U, [n], n), \mathcal{X}(X; V, [n], n))$$

of X, Y for indices U and V, respectively, such that $D(X, Y) \leq D(X_1, Y_1)$.

Further, according to Theorem C.4.18, there exist imputations $(X_2, Y_2) \in \mathcal{Y}(W, n)$ of X_1 and Y_1 for indices W such that $D(X_1, Y_1) \leq D(X_2, Y_2)$. Hence, we have $D(X, Y) \leq D(X_2, Y_2)$.

Then, in order to prove our result, it is sufficient to show that X_2, Y_2 are imputations of X and Y for indices $U \cup W$ and $V \cup W$, respectively, and are such that $(X_2, Y_2) \in (\mathcal{X}(Y; U, [n] \setminus W, n), \mathcal{X}(X; V, [n] \setminus W, n)) \cap \mathcal{Y}(W, n)$.

First, we show that X_2, Y_2 are imputations of X and Y for indices $U \cup W$ and $V \cup W$, respectively. Since X_1, Y_1 are imputations of X, Y for indices U and V, respectively, we have

$$\begin{aligned} & X_1(i) = X(i), \text{ for any } [n] \setminus U, \text{ and } Y_1(i) = Y(i), \text{ for any } [n] \setminus V \\ \Longrightarrow\ & X_1(i) = X(i), \text{ for any } [n] \setminus (U \cup W), \text{ and } Y_1(i) = Y(i), \text{ for any } [n] \setminus (V \cup W). \end{aligned}$$

Further, since X_2, Y_2 are imputations of X_1 and Y_1 for indices W, we have

$$\begin{aligned} & X_2(i) = X_1(i), \text{ for any } [n] \setminus W, \text{ and } Y_2(i) = Y_1(i), \text{ for any } [n] \setminus W \\ \Longrightarrow\ & X_2(i) = X_1(i), \text{ for any } [n] \setminus (U \cup W), \text{ and } Y_2(i) = Y_1(i), \text{ for any } [n] \setminus (V \cup W). \end{aligned}$$

Hence, we have

$$X_2(i) = X(i), \text{ for any } [n] \setminus (U \cup W), \text{ and } Y_2(i) = Y(i), \text{ for any } [n] \setminus (V \cup W).$$

In other words, X_2, Y_2 are imputations of X and Y for indices $U \cup W$ and $V \cup W$, respectively.

Next, we show $(X_2, Y_2) \in (\mathcal{X}(Y; U, [n] \setminus W, n), \mathcal{X}(X; V, [n] \setminus W, n)) \cap \mathcal{Y}(W, n)$. Since $(X_2, Y_2) \in \mathcal{Y}(W, n)$, we only need to show that $(X_2, Y_2) \in (\mathcal{X}(Y; U, [n] \setminus W, n), \mathcal{X}(X; V, [n] \setminus W, n))$.

We now show $X_2 \in \mathcal{X}(Y; U, [n] \setminus W, n)$. Since $X_1 \in \mathcal{X}(Y; U, [n], n)$, then according to the

definition of $\mathcal{X}(Y;U,[n],n)$, we have

$$\begin{aligned} & X_1(i) > \max_{j\in[n]\setminus U} X_1(j), \text{ or } X_1(i) < \min_{j\in[n]\setminus U} X_1(j), \text{ for any } i \in U \\ \Longrightarrow & X_1(i) > \max_{j\in[n]\setminus(U\cup W)} X_1(j), \text{ or } X_1(i) < \min_{j\in[n]\setminus(U\cup W)} X_1(j), \text{ for any } i \in U. \end{aligned}$$

Since X_2 is an imputation of X_1 for W, we have $X_2(i) = X_1(i)$ for any $i \in [n]\setminus W$. Since $[n]\setminus(U\cup W) \subset [n]\setminus W$, we have $\max_{j\in[n]\setminus(U\cup W)} X_1(j) = \max_{j\in[n]\setminus(U\cup W)} X_2(j)$ and $\min_{j\in[n]\setminus(U\cup W)} X_1(j) = \min_{j\in[n]\setminus(U\cup W)} X_2(j)$. Meanwhile, since $U \in [n]\setminus W$, we have $X_2(i) = X_1(i)$ for any $i \in U$. Hence, we have

$$X_2(i) > \max_{j\in[n]\setminus(U\cup W)} X_2(j), \text{ or } X_2(i) < \min_{j\in[n]\setminus(U\cup W)} X_2(j), \text{ for any } i \in U.$$

Thus, we have shown $X_2 \in \mathcal{Z}(U,[n]\setminus W,n)$.

Next, since $X_1 \in \mathcal{X}(Y;U,[n],n)$, then according to the definition of $\mathcal{X}(Y;U,[n],n)$, we have

$$X_1(i) > X_1(j), \text{ if } Y(i) < Y(j), \text{ for any } i,j \in U.$$

Since $X_2(i) = X_1(i)$ for any $i \in U$, then we have

$$X_2(i) > X_2(j), \text{ if } Y(i) < Y(j), \text{ for any } i,j \in U.$$

Hence, we have $X_2 \in \mathcal{X}(Y;U,[n]\setminus W,n)$.

Similarly, we can show that $Y_2 \in \mathcal{X}(Y;V,[n]\setminus W,n)$. Since $Y_1 \in \mathcal{X}(X;V,[n],n)$, then according to the definition of $\mathcal{X}(X;V,[n],n)$, we have

$$\begin{aligned} & Y_1(i) > \max_{j\in[n]\setminus V} Y_1(j), \text{ or } Y_1(i) < \min_{j\in[n]\setminus V} Y_1(j), \text{ for any } i \in V \\ \Longrightarrow & Y_1(i) > \max_{j\in[n]\setminus(V\cup W)} X_1(j), \text{ or } Y_1(i) < \min_{j\in[n]\setminus(V\cup W)} Y_1(j), \text{ for any } i \in V. \end{aligned}$$

Since Y_2 is an imputation of Y_1 for W, we have $Y_2(i) = Y_1(i)$ for any $i \in [n]\setminus W$. Since $[n]\setminus(V\cup W) \subset [n]\setminus W$, we have $\max_{j\in[n]\setminus(V\cup W)} Y_1(j) = \max_{j\in[n]\setminus(V\cup W)} Y_2(j)$ and $\min_{j\in[n]\setminus(V\cup W)} Y_1(j) = \min_{j\in[n]\setminus(V\cup W)} Y_2(j)$. Meanwhile, since $V \in [n]\setminus W$, we have $Y_2(i) = Y_1(i)$ for any $i \in V$. Hence, we have

$$Y_2(i) > \max_{j\in[n]\setminus(V\cup W)} Y_2(j), \text{ or } Y_2(i) < \min_{j\in[n]\setminus(V\cup W)} Y_2(j), \text{ for any } i \in V.$$

Thus, we have shown $Y_2 \in \mathcal{Z}(V,[n]\setminus W,n)$.

Next, since $Y_1 \in \mathcal{X}(X;V,[n],n)$, then according to the definition of $\mathcal{X}(X;V,[n],n)$, we have

$$Y_1(i) > Y_1(j), \text{ if } X(i) < X(j), \text{ for any } i,j \in V.$$

Since $Y_2(i) = Y_1(i)$ for any $i \in V$, then we have

$$Y_2(i) > Y_2(j), \text{ if } X(i) < X(j), \text{ for any } i,j \in V.$$

Hence, we have $Y_2 \in \mathcal{X}(X; V, [n] \setminus W, n)$. This completes our proof for statement $\mathbb{S}$.

Next, we prove Theorem C.4.19 is true using statement $\mathbb{S}$. Let us denote

$$S = \{D(Z_1, Z_2) : (Z_1, Z_2) \in (\mathcal{X}(Y; U, [n] \setminus W, n), \mathcal{X}(X; V, [n] \setminus W, n)) \cap \mathcal{Y}(W, n) \\ \text{are imputations of } X, Y \text{ for indices } U \cup W \text{ and } V \cup W, \text{respectively}\}.$$

The carnality of S is finite and there exist imputations

$$(X_*, Y_*) \in (\mathcal{X}(Y; U, [n] \setminus W, n), \mathcal{X}(X; V, [n] \setminus W, n)) \cap \mathcal{Y}(W, n)$$

of X, Y for indices $U \cup W$ and $V \cup W$, respectively, such that $\max S = D(X_*, Y_*)$.

We now show that $D(X, Y) \leq D(X_*, Y_*)$. According to the statement $\mathbb{S}$, there exist imputations

$$(X_1, Y_1) \in (\mathcal{X}(Y; U, [n] \setminus W, n), \mathcal{X}(X; V, [n] \setminus W, n)) \cap \mathcal{Y}(W, n)$$

of X, Y for indices $U \cup W$ and $V \cup W$, respectively such that $D(X, Y) \leq D(X_1, Y_1)$. According to the definition of S, we have $D(X_1, Y_1) \in S$. Hence, we have $D(X_1, Y_1) \leq \max S = D(X_*, Y_*)$. Thus, we have $D(X, Y) \leq D(X_*, Y_*)$.

Next, we show that $D(X', Y') \leq D(X_*, Y_*)$. According to statement $\mathbb{S}$, there exist imputations

$$(X_2, Y_2) \in (\mathcal{X}(Y'; U, [n] \setminus W, n), \mathcal{X}(X'; V, [n] \setminus W, n)) \cap \mathcal{Y}(W, n)$$

of X', Y' for indices $U \cup W$ and $V \cup W$, respectively such that

$$D(X', Y') \leq D(X_2, Y_2). \tag{C.170}$$

Below, we show that $D(X_2, Y_2) \in S$. In other words, we show that X_2, Y_2 are imputations of X and Y for indices $U \cup W$, and $V \cup W$, respectively such that $(X_2, Y_2) \in (\mathcal{X}(Y; U, [n] \setminus W, n), \mathcal{X}(X; V, [n] \setminus W, n)) \cap \mathcal{Y}(W, n)$.

Since X', Y' are imputations of X and Y for indices $U \cup W$, and $V \cup W$, respectively, then X_2, Y_2 are also imputations of X and Y for indices $U \cup W$, and $V \cup W$, respectively.

Since $X_2 \in \mathcal{X}(Y'; U, [n] \setminus W, n)$, then according to the definition of $\mathcal{X}(Y'; U, [n] \setminus W, n)$, we have

$$X_2(i) > X_2(j), \text{ if } Y'(i) < Y'(j), \text{ for any } i, j \in U.$$

Since Y' is an imputation of Y for indices $V \cup W$, we have

$$Y'(i) = Y(i), \text{ for any } i \in [n] \setminus (V \cup W).$$

Then, since $U \subset [n] \setminus (V \cup W)$, we have

$$Y'(i) = Y(i), \text{ for any } i \in U.$$

Hence, we have

$$X_2(i) > X_2(j), \text{ if } Y(i) < Y(j), \text{ for any } i, j \in U.$$

Since $X_2 \in \mathcal{X}(Y'; U, [n] \setminus W, n)$, then according to the definition of $\mathcal{X}(Y'; U, [n] \setminus W, n)$, we have $X_2 \in \mathcal{Z}(U, [n] \setminus W, n)$. Hence, we have $X_2 \in \mathcal{X}(Y; U, [n] \setminus W, n)$.

Similarly, we can show $Y_2 \in \mathcal{X}(X; V, [n] \setminus W, n)$. Since $Y_2 \in \mathcal{X}(X'; V, [n] \setminus W, n)$, then according to the definition of $\mathcal{X}(X'; V, [n] \setminus W, n)$, we have

$$Y_2(i) > Y_2(j), \text{ if } X'(i) < X'(j), \text{ for any } i, j \in V.$$

Since X' is an imputation of X for indices $U \cup W$, we have

$$X'(i) = X(i), \text{ for any } i \in [n] \setminus (U \cup W).$$

Then, since $V \subset\subset [n] \setminus (U \cup W)$, we have

$$X'(i) = X(i), \text{ for any } i \in V.$$

Hence, we have

$$Y_2(i) > Y_2(j), \text{ if } X(i) < X(j), \text{ for any } i, j \in V.$$

Since $Y_2 \in \mathcal{X}(X'; V, [n] \setminus W, n)$, then according to the definition of $\mathcal{X}(X'; V, [n] \setminus W, n)$, we have $Y_2 \in \mathcal{Z}(V, [n] \setminus W, n)$. Hence, we have $Y_2 \in \mathcal{X}(X; V, [n] \setminus W, n)$.

Next, since $(X_2, Y_2) \in \mathcal{Y}(W, n)$, we have $(X_2, Y_2) \in (\mathcal{X}(Y; U, [n] \setminus W, n), \mathcal{X}(X; V, [n] \setminus W, n)) \cap \mathcal{Y}(W, n)$ Then, according to the definition of S, we have $D(X_2, Y_2) \in S$. Hence, we have $D(X_2, Y_2) \leq \max S = D(X_*, Y_*)$. Thus, we have

$$D\left(X', Y'\right) \leq^{(\text{C.170})} D(X_2, Y_2) \leq D(X_*, Y_*).$$

This completes our proof. □

Remark C.4.20. The number of possible rank vectors of imputations X_* and Y_* for achieving the maximum possible Spearman's footrule under the general missing case can be analysed by combining the analysis for missing cases I, II, and III.

Let us denote $|U| = m_1, |V| = m_2, |W| = m_3$. Consider any pair of imputations $(X_*, Y_*) \in \mathcal{Y}(W, n) \cap (\mathcal{X}(Y; U, [n] \setminus W, n), \mathcal{X}(X; V, [n] \setminus W, n))$ of X and Y for indices $U \cup W$ and $V \cup W$, respectively, where U and V are disjoint from W. In other words, consider a pair of imputations $(X_*, Y_*) \in \mathcal{Y}(W, n)$, which means the components in W satisfy the condition in Definition 5.1.4, and that X_* and Y_* satisfy the conditions of Definition 5.1.3 for the relevant components with indices in U and V, respectively.

Now suppose the components of X_* with indices in U and the components of Y_* with indices in V are fixed. Then, following the discussion of Theorem C.4.18, there are $m_3 + 1$ permutations of ranks of the imputations X_* and Y_* for indices in W that would result in different values of Spearman's footrule.

Next, suppose the components X_* and the components of Y_* with indices in W are fixed. Then there are m_1+1 and m_2+1 permutations of ranks of the imputations X_* and Y_* with indices in U and V, respectively, that result in different values of Spearman's footrule, following a similar argument to Remark C.4.11.

Hence, considering all imputations for indices $U \cup W$ and $V \cup W$, the total number of imputations to consider in order to find the maximum value of Spearman's footrule is $(m_1+1)(m_2+1)(m_3+1)$.

An algorithm for computing the tight upper bounds of Spearman's footrule with computational complexity $\mathcal{O}(n \log n + \min\{m_1, m_2\}^2 + m_1 m_2 m_3)$ is provided in Algorithm 8 in Section C.5.4.

C.5. Efficient algorithms for calculating tight upper bound

This section provides efficient algorithms for calculating exact upper bounds of Spearman's footrule under missing case I, II, III and general missingness case.

C.5.1. Algorithm for tight upper bound under missing case I

We provide an efficient algorithm for calculating exact upper bounds of Spearman's footrule under missing case I. To start, we prove the following lemma:

Lemma C.5.1. Suppose $X, Y \in \mathcal{V}^n$ and for $m_1 \in [n]$, let $U = \{u_1, \ldots, u_{m_1}\} \subset [n]$ be a subset of indices. Suppose $Y(u_1) < \ldots < Y(u_{m_1})$, and assume $X \in \mathcal{X}(Y; U, [n], n)$ is a vector such that $\sum_{i \in U} \mathrm{I}\left(X(i) < \min(X(l))_{l \in [n] \setminus U}\right) = r$, where $r \in \{0, \ldots, m_1\}$. Then, for any $i \in \{1, \ldots, m_1\}$, we have

$$R(X(u_i), X) = \begin{cases} n - i + 1, & \text{if } r = 0, \\ \mathrm{I}\left(u_i \le u_{m_1 - r}\right)(n - i + 1) + \mathrm{I}\left(u_i > u_{m_1 - r}\right)(m_1 - i + 1), & \text{if } m_1 > r > 0, \\ m_1 - i + 1, & \text{if } r = m_1. \end{cases}$$

Proof. To start, since $X \in \mathcal{X}(Y; U, [n], n)$, then according to the definition of $\mathcal{X}(Y; U, [n], n)$, we have

$$X(i) > \max_{j \in [n] \setminus U} X(j), \text{ or } X(i) < \min_{j \in [n] \setminus U} X(j), \text{ for any } i \in U. \tag{C.171}$$

Since $Y(u_1) < \ldots < Y(u_{m_1})$, then according to the definition of $\mathcal{X}(Y; U, [n], n)$, we also have

$$R(X(u_1), X) > \ldots > R(X(u_{m_1}), X). \tag{C.172}$$

Below, we prove our result when $r = 0$, $m_1 > r > 0$ and $r = m_1$, separately.

Suppose $r = 0$, then according to (C.171), we have

$$\begin{aligned} & X(i) > \max_{j \in [n] \setminus U} X(j), \text{ for any } i \in U \\ \implies & R(X(i), X) \in \{n, \ldots, n - m_1 + 1\}, \text{ for any } i \in U \\ \implies & {}^{(C.172)} R(X(u_i), X) = n - i + 1 \text{ for any } i \in \{1, \ldots, m_1\}. \end{aligned}$$

This proves our result when $r = 0$.

Similarly, suppose $m_1 > r > 0$, then according to (C.171), we have

$$\begin{aligned}
&\sum_{i\in U} \mathrm{I}\left(X(i) < \min_{j\in[n]\setminus U} X(j)\right) = r, \text{ and } \sum_{i\in U} \mathrm{I}\left(X(i) > \max_{j\in[n]\setminus U} X(j)\right) = m_1 - r \\
\Longrightarrow\ & R(X(i), X) \in \{1, \ldots, r\} \cup \{n, \ldots, n - m_1 + r + 1\}, \text{ for any } i \in U \\
\Longrightarrow\ & {}^{(C.172)} R(X(u_i), X) = n - i + 1 \text{ for any } \ i \in \{1, \ldots, m_1 - r\}, \\
& \text{and } R(X(u_i), X) = m_1 - i + 1 \text{ for any } \ i \in \{m_1 - r + 1, \ldots, m_1\} \\
\Longrightarrow\ & R(X(u_i), X) = \mathrm{I}\left(u_i \leq u_{m_1 - r}\right)(n - i + 1) + \mathrm{I}\left(u_i > u_{m_1 - r}\right)(m_1 - i + 1).
\end{aligned}$$

This proves our result when $m_1 > r > 0$.

Finally, suppose $r = m_1$, then according to (C.171), we have

$$\begin{aligned}
& X(i) < \min_{j\in[n]\setminus U} X(j), \text{ for any } i \in U \\
\Longrightarrow\ & R(X(i), X) \in \{1, \ldots, r\}, \text{ for any } i \in U \\
\Longrightarrow\ & {}^{(C.172)} R(X(u_i), X) = m_1 - i + 1 \text{ for any } i \in \{1, \ldots, m_1\}.
\end{aligned}$$

This proves our result when $r = m_1$ and completes our proof. □

Then, we prove the following lemma:

Lemma C.5.2. Suppose $X \in \mathcal{V}^n$, and $U \subset [n]$ is a subset of indices. Let $X_* \in \mathcal{Z}(U, [n], n)$ be an imputation of X for indices U such that

$$\sum_{i\in U} \mathrm{I}\left(X_*(i) < \min(X_*(l))_{l\in[n]\setminus U}\right) = r.$$

Then, we have

$$R(X_*(i), X_*) = R(X(i), (X(l))_{l\in[n]\setminus U}) + r, \text{ for any } i \in [n] \setminus U.$$

Proof. According to the definition of $\mathcal{Z}(U, [n], n)$, we have

$$X_*(i) > \max_{j\in[n]\setminus U} X_*(j), \text{ or } X_*(i) < \min_{j\in[n]\setminus U} X_*(j), \text{ for any } i \in U.$$

Then, since $\sum_{i\in U} \mathrm{I}\left(X_*(i) < \min(X_*(l))_{l\in[n]\setminus U}\right) = r$, we have r components of X_* with indices U smaller than any components of X_* with indices $[n] \setminus U$, and other components of X_* with indices U larger than any components of X_* with indices $[n] \setminus U$. Hence, we have

$$\sum_{j\in U} \mathrm{I}\left(X_*(j) \leq X_*(i)\right) = r, \text{ for any } i \in [n] \setminus U. \tag{C.173}$$

Next, according to the definition of rank, for any $i \in [n] \setminus U$, we have

$$\begin{aligned} R(X_*(i), X_*) &= \sum_{j=1}^{n} \mathrm{I}\left(X_*(j) \leq X_*(i)\right) \\ &= \sum_{j \in U} \mathrm{I}\left(X_*(j) \leq X_*(i)\right) + \sum_{j \in [n] \setminus U} \mathrm{I}\left(X_*(j) \leq X_*(i)\right) \\ &=^{(\mathrm{C}.173)} r + \sum_{j \in [n] \setminus U} \mathrm{I}\left(X_*(j) \leq X_*(i)\right). \end{aligned}$$

Since X_* is an imputation of X for U, we have $X_*(i) = X(i)$ for any $i \in [n] \setminus U$. Hence, we have

$$\sum_{j \in [n] \setminus U} \mathrm{I}\left(X(j) \leq X(i)\right) = \sum_{j \in [n] \setminus U} \mathrm{I}\left(X(j) \leq X(i)\right), \text{ for any } i \in [n] \setminus U.$$

Then, for any $i \in [n] \setminus U$, we have

$$\begin{aligned} R(X_*(i), X_*) &= r + \sum_{j \in [n] \setminus U} \mathrm{I}\left(X(j) \leq X(i)\right) \\ &= r + R(X(i), (X(l))_{l \in [n] \setminus U}). \end{aligned}$$

This completes our proof. □

We are now ready to prove the first main results for missing case I:

Proposition C.5.3. Suppose $X, Y \in \mathcal{V}^n$ and for $m_1 \in [n]$, let $U = \{1, \ldots, m_1\} \subset [n]$ be a subset of indices. Suppose $Y(1) < \ldots < Y(m_1)$, and let ${X_*}^{(r)} \in \mathcal{X}(Y; U, [n], n)$ be an imputation of X for indices in U such that

$$\sum_{i \in U} \mathrm{I}\left({X_*}^{(r)}(i) < \min({X_*}^{(r)}(l))_{l \in [n] \setminus U}\right) = r.$$

For any $i \in [n] \setminus U$, denote $d_i = R(Y(i), Y) - R(X(i), (X(j))_{j \in [n] \setminus U})$, and for any $r \in \{0, \ldots, m_1\}$, $i \in \{1, \ldots, m_1\}$, denote $q_i^{(r)} = \mathrm{I}\left(i \leq m_1 - r\right)(n - i + 1) + \mathrm{I}\left(i > m_1 - r\right)(m_1 - i + 1)$. Then, for any $r \in \{0, \ldots, m_1 - 1\}$, we have

$$D(X_*^{(r)}, Y) = \sum_{i \in U} \left| q_i^{(r)} - R(Y(i), Y) \right| + \sum_{i \in [n] \setminus U} |r - d_i|.$$

Proof. To start, according to the definition of Spearman's footrule, we have

$$\begin{aligned} D(X_*^{(r)}, Y) &= \sum_{i=1}^{n} \left| R({X_*}^{(r)}(i), X_*^{(r)}) - R(Y(i), Y) \right| \\ &= \sum_{i=1}^{m_1} \left| R({X_*}^{(r)}(i), X_*^{(r)}) - R(Y(i), Y) \right| + \sum_{i \in [n] \setminus U} \left| R({X_*}^{(r)}(i), X_*^{(r)}) - R(Y(i), Y) \right|. \end{aligned}$$

In order to prove our result, it is then sufficient to show the following two equations hold:

$$(1): \sum_{i=1}^{m_1} \left| R(X_*{}^{(r)}(i), X_*^{(r)}) - R(Y(i), Y) \right| = \sum_{i=1}^{m_1} \left| q_i^{(r)} - R(Y(i), Y) \right|,$$
$$\text{and } (2): \sum_{i \in [n] \setminus U} \left| R(X_*{}^{(r)}(i), X_*^{(r)}) - R(Y(i), Y) \right| = \sum_{i \in [n] \setminus U} |r - d_i|.$$

Below, we show equations (1) and (2) are true separately.

First, we show that equation (1) is true.

Since $X_*{}^{(r)} \in \mathcal{X}(Y; U, [n], n)$, and $Y(1) < \ldots < Y(m_1)$, then according to Lemma C.5.1, for any $i \in \{1, \ldots, m_1\}$, and $r \in \{0, \ldots, m_1\}$, we have

$$R(X_*{}^{(r)}(i), X_*^{(r)}) = \begin{cases} n - i + 1, & \text{if } r = 0, \\ q_i^{(r)}, & \text{if } m_1 > r > 0, \\ m_1 - i + 1, & \text{if } r = m_1, \end{cases}$$

where $q_i^{(r)} = \mathrm{I}(i \leq m_1 - r)(n - i + 1) + \mathrm{I}(i > m_1 - r)(m_1 - i + 1)$. Notice that when $r = 0$ and $r = m_1$, $R(X(i), X) = q_i^{(r)}$ is still true for any $i = 1, \ldots, m_1$. Hence, we have

$$R(X(i), X) = q_i^{(r)}, \text{ for any } i \in U, r \in \{0, \ldots, m_1\}.$$

Thus, for any $i \in U, r \in \{0, \ldots, m_1\}$, we have

$$\sum_{i=1}^{m_1} \left| R(X_*{}^{(r)}(i), X_*^{(r)}) - R(Y(i), Y) \right| = \left| q_i^{(r)} - R(Y(i), Y) \right|.$$

This proves equation (1).

Next, we show that equation (2) is true.

Since $X_*^{(r)} \in \mathcal{X}(Y; U, [n], n)$, then according to the definition of $\mathcal{X}(Y; U, [n], n)$, we also have $X_*^{(r)} \in \mathcal{Z}(U, [n], n)$. Notice that $X_*^{(r)}$ is an imputation of X for U, and

$$\sum_{i \in U} \mathrm{I}\left(X_*{}^{(r)}(i) < \min(X_*{}^{(r)}(l))_{l \in [n] \setminus U} \right) = r.$$

Then, according to Lemma C.5.2, for any $i \in [n] \setminus U$, we have

$$R(X_*{}^{(r)}(i), X_*^{(r)}) = r + R(X(i), (X(l))_{l \in [n] \setminus U}), \text{ for any } i \in [n] \setminus U. \tag{C.174}$$

Hence, we have

$$\begin{aligned} &\sum_{i \in [n] \setminus U} \left| R(X_*{}^{(r)}(i), X_*^{(r)}) - R(Y(i), Y) \right| \\ &= \sum_{i \in [n] \setminus U} \left| R(X(i), (X(l))_{l \in [n] \setminus U}) + r - R(Y(i), Y) \right| \\ &= \sum_{i \in [n] \setminus U} |r - d_i|. \end{aligned}$$

This proves equation (2), and completes our proof. □

Before showing the second main result for missing case I, we prove the following two lemmas.

Lemma C.5.4. Suppose $X = (X(u_1), \ldots, X(u_n))$ is a vector of integers, where $U = \{u_1, \ldots, u_n\}$ is a set of indices such that $u_1 < \ldots < u_n$. Then, we have

$$\sum_{i \in U} |X(i) + 1| - \sum_{i \in U} |X(i)| = 2 \sum_{i \in U} \mathrm{I}\left(X(i) \geq 0\right) - n.$$

Proof. To start, for any $i \in U$, if $X(i) \geq 0$, we have

$$|X(i) + 1| - |X(i)| = 1.$$

If, however $X(i) < 0$, then since X is a vector of integers, we have $X(i) + 1 \leq 0$. Hence, we have

$$|X(i) + 1| - |X(i)| = -1.$$

Therefore, we have

$$\begin{aligned}
\sum_{i \in U} |X(i) + 1| - \sum_{i \in U} |X(i)| &= \sum_{i \in U} \left(|X(i) + 1| - |X(i)|\right) \\
&= \sum_{i \in U} \mathrm{I}\left(X(i) \geq 0\right) - \sum_{i \in U} \mathrm{I}\left(X(i) < 0\right) \\
&= \sum_{i \in U} \mathrm{I}\left(X(i) \geq 0\right) - \left(n - \sum_{i \in U} \mathrm{I}\left(X(i) \geq 0\right)\right) \\
&= 2 \sum_{i \in U} \mathrm{I}\left(X(i) \geq 0\right) - n.
\end{aligned}$$

This completes our proof. □

Lemma C.5.5. Suppose $U = \{1, \ldots, m_1\}$, and $n \in \mathbb{N}$ is a number such that $n > m_1$. For any $i \in \{1, \ldots, m_1\}$, $r \in \{0, \ldots, m_1\}$, define

$$q_i^{(r)} = \mathrm{I}\left(i \leq m_1 - r\right)\left(n - i + 1\right) + \mathrm{I}\left(i > m_1 - r\right)\left(m_1 - i + 1\right).$$

Suppose $Y(1), \ldots, Y(m_1)$ are real numbers. Then, for any $r \in \{0, \ldots, m_1 - 1\}$, we have

$$\begin{aligned}
&\sum_{i=1}^{m_1} \left|q_i^{(r+1)} - Y(i)\right| \\
&= \sum_{i=1}^{m_1} \left|q_i^{(r)} - Y(i)\right| + \left|r + 1 - Y(m_1 - r)\right| - \left|n - m_1 + r + 1 - Y(m_1 - r)\right|.
\end{aligned}$$

Proof. We show our result for $r = 0$, $m_1 - 1 > r > 0$, and $r = m_1 - 1$ separately.

When $r = 0$, we have

$$\sum_{i\in U}\left|q_i^{(r)} - Y(i)\right| = \sum_{i=1}^{m_1}|n - i + 1 - Y(i)|,$$
$$\sum_{i\in U}\left|q_i^{(r+1)} - Y(i)\right| = \sum_{i=1}^{m_1-1}|n - i + 1 - Y(i)| + |1 - Y(i)|.$$

Hence, we have

$$\begin{aligned}&\sum_{i\in U}\left|q_i^{(r+1)} - Y(i)\right| - \sum_{i\in U}\left|q_i^{(r)} - Y(i)\right| \\ &= |1 - Y(m_1)| - |n - m_1 + 1 - Y(m_1)| \\ &= |r + 1 - Y(m_1 - r)| - |n - m_1 + r + 1 - Y(m_1 - r)|\,.\end{aligned}$$

This completes our proof when $r = 0$.

Similarly, when $m_1 - 1 > r > 0$, we have

$$\sum_{i\in U}\left|q_i^{(r)} - Y(i)\right| = \sum_{i=1}^{m_1-r}|n - i + 1 - Y(i)| + \sum_{i=m_1-r+1}^{m_1}|m_1 - i + 1 - Y(i)|,$$

and

$$\sum_{i\in U}\left|q_i^{(r+1)} - Y(i)\right| = \sum_{i=1}^{m_1-r-1}|n - i + 1 - Y(i)| + \sum_{i=m_1-r}^{m_1}|m_1 - i + 1 - Y(i)|.$$

Hence, we have

$$\begin{aligned}&\sum_{i\in U}\left|q_i^{(r+1)} - Y(i)\right| - \sum_{i\in U}\left|q_i^{(r)} - Y(i)\right| \\ &= |m_1 - (m_1 - r) + 1 - Y(m_1 - r)| - |n - (m_1 - r) + 1 - Y(m_1 - r)| \\ &= |r + 1 - Y(m_1 - r)| - |n - m_1 + r + 1 - Y(m_1 - r)|\,.\end{aligned}$$

This completes our proof when $m_1 - 1 > r > 0$.

Similarly, when $r = m_1 - 1$, we have

$$\sum_{i\in U}\left|q_i^{(r+1)} - Y(i)\right| = |n - 1 + 1 - Y(1)| + \sum_{i=2}^{m_1}|m_1 - i + 1 - Y(i)|,$$
$$\sum_{i\in U}\left|q_i^{(r)} - Y(i)\right| = \sum_{i=1}^{m_1}|m_1 - i + 1 - Y(i)|.$$

Hence, we have

$$\begin{aligned}&\sum_{i\in U}\left|q_i^{(r+1)} - Y(i)\right| + \sum_{i\in U}\left|q_i^{(r)} - Y(i)\right| \\ &= |m_1 - 1 + 1 - Y(1)| - |n - 1 + 1 - Y(1)| \\ &= |r + 1 - Y(m_1 - r)| - |n - m_1 + r + 1 - Y(m_1 - r)|\,.\end{aligned}$$

This completes our proof when $r = m_1 - 1$. Hence, we have shown our result. □

Now, we are ready to prove the second main result for missing case I.

Proposition C.5.6. Suppose $X, Y \in \mathcal{V}^n$ and for $m_1 \in [n]$, let $U = \{1, \ldots, m_1\} \subset [n]$ be a subset of indices. Suppose $Y(1) < \ldots < Y(m_1)$, and let ${X_*}^{(r)} \in \mathcal{X}(Y; U, [n], n)$ be an imputation of X for indices in U such that

$$\sum_{i \in U} \mathrm{I}\left({X_*}^{(r)}(i) < \min({X_*}^{(r)}(l))_{l \in [n] \setminus U}\right) = r.$$

For any $i \in [n] \setminus U$, denote $d_i = R(Y(i), Y) - R(X(i), (X(j))_{j \in [n] \setminus U})$, and for any $r \in \{0, \ldots, m_1 - 1\}$, denote $s_r = \sum_{i \in [n] \setminus U} \mathrm{I}(d_i \leq r)$. Then, for any $r \in \{0, \ldots, m_1 - 1\}$, we have

$$D(X_*^{(r+1)}, Y) = D(X_*^{(r)}, Y) + 2s_r - n + m_1 + C^{(r)},$$

where $C^{(r)} = |r + 1 - R(Y(m_1 - r), Y)| - |n - m_1 + r + 1 - R(Y(m_1 - r), Y)|$.

Proof. To start, according to Proposition C.5.3, we have

$$D(X_*^{(r)}, Y) = \sum_{i \in U} \left| q_i^{(r)} - R(Y(i), Y) \right| + \sum_{i \in [n] \setminus U} |r - d_i|,$$
$$\text{and } D(X_*^{(r+1)}, Y) = \sum_{i \in U} \left| q_i^{(r+1)} - R(Y(i), Y) \right| + \sum_{i \in [n] \setminus U} |r + 1 - d_i|,$$

where $q_i^{(r)} = \mathrm{I}(i \leq m_1 - r)(n - i + 1) + \mathrm{I}(i > m_1 - r)(m_1 - i + 1)$ for any $i \in \{1, \ldots, m_1\}$ and $r \in \{0, \ldots, m_1\}$. Then, in order to prove our result, it is sufficient to show that

$$\sum_{i \in [n] \setminus U} |r + 1 - d_i| - \sum_{i \in [n] \setminus U} |r - d_i| = 2s_r - n + m_1,$$

and

$$\sum_{i \in U} \left| q_i^{(r+1)} - R(Y(i), Y) \right| - \sum_{i \in U} \left| q_i^{(r)} - R(Y(i), Y) \right|$$
$$= |r + 1 - R(Y(m_1 - r), Y)| - |n - m_1 + r_1 + 1 - R(Y(m_1 - r), Y)|.$$

According to Lemma C.5.4, we have

$$\begin{aligned} \sum_{i \in [n] \setminus U} |r + 1 - d_i| - \sum_{i \in [n] \setminus U} |r - d_i| &= 2 \sum_{i \in [n] \setminus U} \mathrm{I}(r - d_i \geq 0) - (n - m_1) \\ &= 2s_r - n + m_1. \end{aligned}$$

Next, notice that $q_i^{(r)} = \mathrm{I}(i \leq m_1 - r)(n - i + 1) + \mathrm{I}(i > m_1 - r)(m_1 - i + 1)$, for any $i \in$

$\{1,\ldots,m_1\}$ and $r \in \{0,\ldots,m_1\}$. Then according to Lemma C.5.5, we have

$$\begin{aligned}
&\sum_{i\in U}\left|q_i^{(r+1)} - R(Y(i),Y)\right| - \sum_{i\in U}\left|q_i^{(r)} - R(Y(i),Y)\right| \\
&= |r+1-Y(m_1-r)| - |n-m_1+r+1-Y(m_1-r)| \\
&= C^{(r)}.
\end{aligned}$$

This completes our proof. □

Algorithm 4 An efficient algorithm for computing cumulative counts.

Input: A sequence of integers $Z = (Z(1),\ldots,Z(n))$, two integer number $m_1 \le m_2$.
Output: $s_r = \sum_{i\in[n]} \mathrm{I}(Z(i) \le r)$ for any $r \in \{m_1,\ldots,m_2\}$.

1: Initialise $s_{m_1} = 0$. If $m_2 > m_1$, initialise $t_i = 0$ for any $i \in \{m_1+1,\ldots,m_2\}$.
2: **for** $i = 1,\ldots,n$ **do**
3: **if** $Z(i) \le m_1$ **then**
4: $s_0 = s_0 + 1$.
5: **end if**
6: **if** $m_1 < Z(i) \le m_2$ **then**
7: $t_{Z(i)} = t_{Z(i)} + 1$.
8: **end if**
9: **end for**
10: If $m_2 > m_1$, compute $s_{r+1} = s_r + t_{r+1}$ for any $r \in \{m_1,\ldots,m_2-1\}$.
11: Return s_r for any $r \in \{m_1,\ldots,m_2\}$.

Remark C.5.7. The computational complexity of Algorithm 4 is analysed as follows. Each step between line 3 and line 8 takes constant time $\mathcal{O}(1)$, and since the loop runs n times, the computational complexity for the loop is $\mathcal{O}(n)$. The computation complexity for line 10 is $\mathcal{O}(m_2 - m_1)$. Therefore, the overall computational complexity for Algorithm 4 is $\mathcal{O}(n + m_2 - m_1)$.

Algorithm 5 An efficient algorithm for computing exact upper bounds of Spearman's footrule under missing case I.

Input: $X, Y \in \mathcal{V}^n$, where either X or Y is fully observed, while the other is partially observed.
Output: Maximum possible Spearman's footrule distance between X and Y.

1: If X is fully observed while Y is partially observed, then change the label of X and Y.
2: Denote m_1 as the number of missing components in X. If $m_1 = n$, return $\sum_{i=1}^n |i-(n-i+1)|$.
3: Rank all observed components in X and Y.
4: Relabel X and Y such that $(X(1),\ldots,X(m_1))$ are unobserved and $Y(1) < \ldots < Y(m_1)$.
5: Let $U = \{1,2,\ldots,m_1\}$. For $i \in [n] \setminus U$, let $d_i = R(Y(i),Y) - R(X(i),(X(l))_{l\in[n]\setminus U})$.
6: Run Algorithm 4 computing $s_r = \sum_{i\in[n]\setminus U} \mathrm{I}(d_i \le r)$ for any $r = \{0,\ldots,m_1-1\}$.
7: For any $r \in \{0,\ldots,m_1-1\}$, compute

$$C^{(r)} = |r+1-R(Y(m_1-r),Y)| - |n-m_1+r+1-R(Y(m_1-r),Y)|.$$

8: Initialise $D_0 = \sum_{i=1}^{m_1} \left|(n-i+1) - R(Y(i),(Y(l))_{l\in[n]})\right| + \sum_{i\in[n]\setminus U} |d_i|$.
9: For $r \in \{0,\ldots,m_1-1\}$, compute $D_{r+1} = D_r + 2s_r - n + m_1 + C^{(r)}$.
10: Return $\max\{D_0,\ldots,D_{m_1}\}$.

Remark C.5.8. A few comments of Algorithm 5 are made below:

In Algorithm 5, the initialisation D_0 computed in line 8 equals to $D(X_*^{(0)}, Y)$ defined in Proposition C.5.3. Spearman's footrule D_{r+1} and D_r equal to $D(X_*^{(r+1)}, Y)$ and $D(X_*^{(r)}, Y)$, respectively, and are updated according to Proposition C.5.6.

By computing $\{D_0, \ldots, D_{m_1}\}$, Algorithm 5 finds all possible Spearman's footrule values $D(X_*, Y)$, where $X_* \in \mathcal{X}(Y; U, [n], n)$ is imputation of X for indices U. See Remark 3 in the main paper for explanation. Hence, according to Theorem C.4.10, the algorithm guarantees to find the maximum possible Spearman's footrule between X and Y.

The computational complexity of Algorithm 5 is analysed as follows. Ranking and relabelling all observed components in X and Y in line 3 requires $\mathcal{O}(n \log n)$ steps. Using these rankings, in line 5 computing each d_i is $\mathcal{O}(1)$, and so overall line 5 is $\mathcal{O}(n - m_1)$. According to Remark C.5.7, the computational complexity of running Algorithm 4 in line 6 is $\mathcal{O}(n+m_1) = \mathcal{O}(n)$. In line 7, each iteration of the for loop is $\mathcal{O}(1)$, and since the loop runs m_1 times, the computational complexity for the loop is $\mathcal{O}(m_1)$. Line 8 and 9 takes $\mathcal{O}(1)$ and $\mathcal{O}(m_1)$ steps, respectively. Therefore, the overall computational complexity for Algorithm 5 is $\mathcal{O}(n \log n)$.

C.5.2. Algorithm for tight upper bound under missing case II

This subsection provides an efficient algorithm for computing exact upper bounds of Spearman's footrule under missing case II, where X and Y might be partially observed, but for any pair $(X(i), Y(i))$, where $i \in [n]$, at least one value is observed. First, we show the following result:

Proposition C.5.9. Suppose $X, Y \in \mathcal{V}^n$ and for $m_1, m_2 \in [n]$ such that $m_1 + m_2 < n$, let $V = \{1, \ldots, m_2\}$ and $U = \{m_2+1, \ldots, m_2+m_1\}$ be subsets of indices. Suppose $X(1) < \ldots < X(m_2)$, and $Y(m_2+1) < \ldots < Y(m_2+m_1)$. Let $({X_*}^{(r_1)}, Y_*^{(r_2)}) \in (\mathcal{X}(Y; U, [n], n), \mathcal{X}(X; V, [n], n))$ be imputations of X, Y for indices U and V, respectively, such that

$$\sum_{i \in U} \mathrm{I}\left({X_*}^{(r_1)}(i) < \min({X_*}^{(r_1)}(l))_{l \in [n] \setminus U}\right) = r_1,$$
$$\text{and } \sum_{i \in V} \mathrm{I}\left(Y_*^{(r_2)}(i) < \min(Y_*^{(r_2)}(l))_{l \in [n] \setminus V}\right) = r_2.$$

For any $i \in [n] \setminus (U \cup V)$, denote $d_i = R(Y(i), (Y(l))_{l \in [n] \setminus V}) - R(X(i), (X(l))_{l \in [n] \setminus U})$. Then, for any $r_1 \in \{0, \ldots, m_1\}$, and $r_2 \in \{0, \ldots, m_2\}$, we have

$$\begin{aligned} D(X_*^{(r_1)}, Y_*^{(r_2)}) = &\sum_{i \in V} \left| R(X(i), (X(l))_{l \in [n] \setminus U}) + r_1 - q_i^{(r_2)} \right| \\ &+ \sum_{i \in U} \left| p_{i-m_2}^{(r_1)} - R(Y(i), (Y(l))_{l \in [n] \setminus V}) - r_2 \right| + \sum_{i \in [n] \setminus (U \cup V)} |r_1 - r_2 - d_i| . \end{aligned}$$

where $q_i^{(r_2)} = \mathrm{I}(i \le m_2 - r_2)(n-i+1) + \mathrm{I}(i > m_2 - r_2)(m_2 - i + 1)$, and $p_i^{(r_1)} = \mathrm{I}(i \le m_1 - r_1)(n - i + 1) + \mathrm{I}(i > m_1 - r_1)(m_1 - i + 1)$.

Proof. To start, according to the definition of Spearman's footrule, we have

$$\begin{aligned} D(X_*^{(r_1)}, Y_*^{(r_2)}) = &\sum_{i\in V} \left| R(X_*{}^{(r_1)}(i), X_*^{(r_1)}) - R(Y_*^{(r_2)}(i), Y_*^{(r_2)}) \right| \\ &+ \sum_{i\in U} \left| R(X_*{}^{(r_1)}(i), X_*^{(r_1)}) - R(Y_*^{(r_2)}(i), Y_*^{(r_2)}) \right| \\ &+ \sum_{i\in[n]\setminus(U\cup V)} \left| R(X_*{}^{(r_1)}(i), X_*^{(r_1)}) - R(Y_*^{(r_2)}(i), Y_*^{(r_2)}) \right|. \end{aligned}$$

Hence, in order to prove our results, it is then sufficient to show the following three equations all hold:

$$\begin{aligned} (1): &\sum_{i\in V} \left| R(X_*{}^{(r_1)}(i), X_*^{(r_1)}) - R(Y_*^{(r_2)}(i), Y_*^{(r_2)}) \right| \\ &= \sum_{i\in V} \left| R(X(i), (X(l))_{l\in[n]\setminus U}) + r_1 - q_i^{(r_2)} \right|, \\ (2): &\sum_{i\in U} \left| R(X_*{}^{(r_1)}(i), X_*^{(r_1)}) - R(Y_*^{(r_2)}(i), Y_*^{(r_2)}) \right| \\ &= \sum_{i\in U} \left| p_i^{(r_1)} - R(Y(i), (Y(l))_{l\in[n]\setminus V}) - r_2 \right|, \\ \text{and } (3): &\sum_{i\in[n]\setminus(U\cup V)} \left| R(X_*{}^{(r_1)}(i), X_*^{(r_1)}) - R(Y_*^{(r_2)}(i), Y_*^{(r_2)}) \right| \\ &= \sum_{i\in[n]\setminus(U\cup V)} |r_1 - r_2 - d_i|. \end{aligned}$$

Below, we show equations (1), (2) and (3) are true separately.

We first show equation (1) is true. Since $X_*^{(r_1)} \in \mathcal{X}(Y; U, [n], n)$, then according to the definition of $\mathcal{X}(Y; U, [n], n)$, we have $X_*^{(r_1)} \in \mathcal{Z}(U, [n], n)$. Since

$$\sum_{i\in U} \mathrm{I}\left(X_*{}^{(r_1)}(i) < \min(X_*{}^{(r_1)}(l))_{l\in[n]\setminus U} \right) = r_1,$$

and $X_*^{(r_1)}$ is an imputation of X for U, then according to Lemma C.5.2, we have

$$R(X_*{}^{(r_1)}(i), X_*^{(r_1)}) = R(X(i), (X(l))_{l\in[n]\setminus U}) + r_1, \text{ for any } i \in [n] \setminus U. \tag{C.175}$$

Next, since $Y_*^{(r_2)} \in \mathcal{X}(X; V, [n], n)$, and $X(1) < \ldots < X(m_2)$, then according to Lemma C.5.1, for any $i \in \{1, \ldots, m_2\}$, we have

$$R(Y_*^{(r_2)}(i), Y_*^{(r_2)}) = \begin{cases} n - i + 1, & \text{if } r_2 = 0, \\ q_i^{(r_2)}, & \text{if } m_2 > r_2 > 0, \\ m_2 - i + 1, & \text{if } r = m_2, \end{cases}$$

where $q_i^{(r_2)} = \mathrm{I}(i \le m_2 - r_2)(n - i + 1) + \mathrm{I}(i > m_2 - r_2)(m_2 - i + 1)$. Notice that when $r_2 = 0$

and $r_2 = m_2$, $R(Y_*^{(r_2)}(i), Y_*^{(r_2)}) = q_i^{(r_2)}$ is still true for any $i \in \{1, \ldots, m_2\}$. Hence, we have

$$R(Y_*^{(r_2)}(i), Y_*^{(r_2)}) = q_i^{(r_2)}, \text{ for any } i \in V, r_2 \in \{0, \ldots, m_2\}. \tag{C.176}$$

Then, combining (C.175), and (C.176), we have

$$\begin{aligned}
&\sum_{i \in V} \left| R(X_*{}^{(r_1)}(i), X_*^{(r_1)}) - R(Y_*^{(r_2)}(i), Y_*^{(r_2)}) \right| \\
&= \sum_{i \in V} \left| R(X(i), (X(l))_{l \in [n] \setminus U}) + r_1 - q_i^{(r_2)} \right|,
\end{aligned}$$

which proves equation (1).

Similarly, we can show equation (2) is true. Since $Y_*^{(r_2)} \in \mathcal{X}(X; V, [n], n)$, then according to the definition of $\mathcal{X}(X; V, [n], n)$, we have $Y_*^{(r_2)} \in \mathcal{Z}(V, [n], n)$. Since

$$\sum_{i \in V} \mathrm{I}\left(Y_*^{(r_2)}(i) < \min(Y_*^{(r_2)}(l))_{l \in [n] \setminus V} \right) = r_2,$$

and $Y_*^{(r_2)}$ is an imputation of Y for V, then according to Lemma C.5.2, we have

$$R(Y_*^{(r_2)}(i), Y_*^{(r_2)}) = R(Y(i), (Y(l))_{l \in [n] \setminus V}) + r_2, \text{ for any } i \in [n] \setminus V. \tag{C.177}$$

Next, since $X_*^{(r_1)} \in \mathcal{X}(Y; U, [n], n)$, and $Y(m_2 + 1) < \ldots < Y(m_2 + m_1)$, then according to Lemma C.5.1, for any $i \in \{1, \ldots, m_2\}$, we have

$$R(X_*{}^{(r_1)}(m_2 + i), X_*^{(r_1)}) = \begin{cases} n - i + 1, & \text{if } r_1 = 0, \\ p_i^{(r_1)}, & \text{if } m_1 > r_1 > 0, \\ m_1 - i + 1, & \text{if } r = m_1, \end{cases}$$

where $p_i^{(r_1)} = \mathrm{I}\left(m_2 + i \leq m_2 + m_1 - r_1\right)(n - i + 1) + \mathrm{I}\left(m_2 + i > m_2 + m_1 - r_1\right)(m_1 - i + 1) = \mathrm{I}\left(i \leq m_1 - r_1\right)(n - i + 1) + \mathrm{I}\left(i > m_1 - r_1\right)(m_1 - i + 1)$. Notice that when $r_1 = 0$ and $r_1 = m_1$, $R(X_*{}^{(r_1)}(m_2 + i), X_*^{(r_1)}) = p_i^{(r_1)}$ is still true for any $i \in \{1, \ldots, m_2\}$. Hence, we have

$$R(X_*{}^{(r_1)}(m_2 + i), X_*^{(r_1)}) = p_i^{(r_1)}, \text{ for any } i \in \{1, \ldots, m_1\}, r_1 \in \{0, \ldots, m_1\}.$$

In other words, for any $i \in \{m_2 + 1, \ldots, m_2 + m_1\}, r_1 \in \{0, \ldots, m_1\}$, we have

$$R(X_*{}^{(r_1)}(i), X_*^{(r_1)}) = p_{i - m_2}^{(r_1)}. \tag{C.178}$$

Then, combining (C.177), and (C.178), we have

$$\begin{aligned}
&\sum_{i \in U} \left| R(X_*{}^{(r_1)}(i), X_*^{(r_1)}) - R(Y_*^{(r_2)}(i), Y_*^{(r_2)}) \right| \\
&= \sum_{i \in V} \left| p_{i - m_2}^{(r_1)} - R(Y(i), (Y(l))_{l \in [n] \setminus V}) - r_2 \right|,
\end{aligned}$$

which proves equation (2).

We now show equation (3) is true. Since $Y_*^{(r_2)} \in \mathcal{X}(X; V, [n], n)$, then according to the definition of $\mathcal{X}(X; V, [n], n)$, we have $Y_*^{(r_2)} \in \mathcal{Z}(V, [n], n)$. Since

$$\sum_{i \in V} \mathrm{I}\left(Y_*^{(r_2)}(i) < \min(Y_*^{(r_2)}(l))_{l \in [n] \setminus V}\right) = r_2,$$

and $Y_*^{(r_2)}$ is an imputation of Y for V, then according to Lemma C.5.2, we have

$$R(Y_*^{(r_2)}(i), Y_*^{(r_2)}) = R(Y(i), (Y(l))_{l \in [n] \setminus V}) + r_2, \text{ for any } i \in [n] \setminus V. \tag{C.179}$$

Then, combining (C.175) and (C.179), we have

$$\begin{aligned}
&\sum_{i \in [n] \setminus (U \cup V)} \left| R(X_*{}^{(r_1)}(i), X_*^{(r_1)}) - R(Y_*^{(r_2)}(i), Y_*^{(r_2)}) \right| \\
&= \sum_{i \in [n] \setminus (U \cup V)} \left| R(X(i), (X(l))_{l \in [n] \setminus U}) + r_1 - R(Y(i), (Y(l))_{l \in [n] \setminus V}) - r_2 \right| \\
&= \sum_{i \in [n] \setminus (U \cup V)} |r_1 - r_2 - d_i| .
\end{aligned}$$

This proves equation (3), and completes our proof. □

Next, we show the following result.

Proposition C.5.10. Suppose $X, Y \in \mathcal{V}^n$ and for $m_1, m_2 \in [n]$ such that $m_1 + m_2 < n$, let $V = \{1, \ldots, m_2\}$ and $U = \{m_2 + 1, \ldots, m_2 + m_1\}$ be subsets of indices. Suppose $X(1) < \ldots < X(m_2)$, and $Y(m_2 + 1) < \ldots < Y(m_2 + m_1)$. Let $(X_*{}^{(r_1)}, Y_*^{(r_2)}) \in (\mathcal{X}(Y; U, [n], n), \mathcal{X}(X; V, [n], n))$ be imputations of X, Y for indices U and V, respectively, such that

$$\begin{aligned}
\sum_{i \in U} \mathrm{I}\left(X_*{}^{(r_1)}(i) < \min(X_*{}^{(r_1)}(l))_{l \in [n] \setminus U}\right) &= r_1, \\
\text{and } \sum_{i \in V} \mathrm{I}\left(Y_*^{(r_2)}(i) < \min(Y_*^{(r_2)}(l))_{l \in [n] \setminus V}\right) &= r_2.
\end{aligned}$$

For any $r_1 \in \{0, \ldots, m_1\}$ and $r_2 \in \{0, \ldots, m_2\}$, define

$$\begin{aligned}
S^{(r_1, r_2)} &= \sum_{i \in V} \mathrm{I}\left(R(X(i), (X(l))_{l \in [n] \setminus U}) + r_1 - q_i^{(r_2)} \geq 0\right), \\
\text{and } R^{(r_1, r_2)} &= \sum_{i \in U} \mathrm{I}\left(p_{i - m_2}^{(r_1)} - R(Y(i), (Y(l))_{l \in [n] \setminus V}) - r_2 - 1 \geq 0\right),
\end{aligned}$$

where $q_i^{(r_2)} = \mathrm{I}(i \leq m_2 - r_2)(n - i + 1) + \mathrm{I}(i > m_2 - r_2)(m_2 - i + 1)$ and $p_i^{(r_1)} = \mathrm{I}(i \leq m_1 - r_1)(n - i + 1) + \mathrm{I}(i > m_1 - r_1)(m_1 - i + 1)$. For any $i \in [n] \setminus (U \cup V)$, denote $d_i = R(Y(i), (Y(l))_{l \in [n] \setminus V}) - R(X(i), (X(l))_{l \in [n] \setminus U})$, and for any $i \in \{-m_2, \ldots, m_1 - 1\}$, denote $s_i = \sum_{l \in [n] \setminus (U \cup V)} \mathrm{I}(d_l \leq i)$. Denote $n' = n - m_1 - m_2$. Then, for any $r_1 \in \{0, \ldots, m_1 - 1\}$, $r_2 \in \{0, \ldots, m_2\}$, we have

$$D(X_*^{(r_1 + 1)}, Y_*^{(r_2)}) = D(X_*^{(r_1)}, Y_*^{(r_2)}) + 2s_{r_1 - r_2} + n' + 2S^{(r_1, r_2)} + m_2 + C_1^{(r_1, r_2)}, \tag{C.180}$$

and for any $r_1 \in \{0, \ldots, m_1\}$, $r_2 \in \{0, \ldots, m_2 - 1\}$, we have

$$\begin{aligned} D(X_*^{(r_1)}, Y_*^{(r_2+1)}) &= D(X_*^{(r_1)}, Y_*^{(r_2)}) \\ &\quad - 2s_{r_1-r_2-1} + n' - 2R^{(r_1,r_2)} + m_1 + C_2^{(r_1,r_2)}, \end{aligned} \tag{C.181}$$

where

$$\begin{aligned} C_1^{(r_1,r_2)} &= |r_1 + 1 - R(Y(m_2 + m_1 - r_1), (Y(l))_{l\in[n]\setminus V}) - r_2| \\ &\quad - |n - m_1 + r_1 + 1 - R(Y(m_2 + m_1 - r_1), (Y(l))_{l\in[n]\setminus V}) - r_2|, \\ C_2^{(r_1,r_2)} &= |r_2 + 1 - R(X(m_2 - r_2), (X(l))_{l\in[n]\setminus U}) - r_1| \\ &\quad - \left|n - m_2 + r_2 + 1 - R(X(m_2 - r_2), (X(l))_{l\in[n]\setminus U}) - r_1\right|. \end{aligned}$$

Proof. We first show (C.180) is true. According to Proposition C.5.9, we have

$$\begin{aligned} D(X_*^{(r_1)}, Y_*^{(r_2)}) &= \sum_{i\in V} \left|R(X(i), (X(l))_{l\in[n]\setminus U}) + r_1 - q_i^{(r_2)}\right| \\ &\quad + \sum_{i\in U} \left|p_{i-m_2}^{(r_1)} - R(Y(i), (Y(l))_{l\in[n]\setminus V}) - r_2\right| \\ &\quad + \sum_{i\in[n]\setminus(U\cup V)} |r_1 - r_2 - d_i|, \end{aligned}$$

and

$$\begin{aligned} D(X_*^{(r_1+1)}, Y_*^{(r_2)}) &= \sum_{i\in V} \left|R(X(i), (X(l))_{l\in[n]\setminus U}) + r_1 + 1 - q_i^{(r_2)}\right| \\ &\quad + \sum_{i\in U} \left|p_{i-m_2}^{(r_1+1)} - R(Y(i), (Y(l))_{l\in[n]\setminus V}) - r_2\right| \\ &\quad + \sum_{i\in[n]\setminus(U\cup V)} |r_1 + 1 - r_2 - d_i|. \end{aligned}$$

In order to prove (C.180), it is then sufficient to show the following three equations all hold:

$$\begin{aligned} &\sum_{i\in V} \left|R(X(i), (X(l))_{l\in[n]\setminus U}) + r_1 + 1 - q_i^{(r_2)}\right| \\ &- \sum_{i\in V} \left|R(X(i), (X(l))_{l\in[n]\setminus U}) + r_1 - q_i^{(r_2)}\right| = 2S^{(r_1,r_2)} + m_2, \end{aligned} \tag{C.182}$$

$$\begin{aligned} &\sum_{i\in U} \left|p_{i-m_2}^{(r_1+1)} - R(Y(i), (Y(l))_{l\in[n]\setminus V}) - r_2\right| \\ &- \sum_{i\in U} \left|p_{i-m_2}^{(r_1)} - R(Y(i), (Y(l))_{l\in[n]\setminus V}) - r_2\right| = C_1^{(r_1,r_2)}, \end{aligned} \tag{C.183}$$

and

$$\sum_{i\in[n]\setminus(U\cup V)} |r_1 + 1 - r_2 - d_i| - \sum_{i\in[n]\setminus(U\cup V)} |r_1 - r_2 - d_i| = 2s_{r_1-r_2} + n'. \tag{C.184}$$

First, we show (C.182) is true. According to Lemma C.5.4, we have

$$\begin{aligned}
&\sum_{i\in V}\left|R(X(i),(X(l))_{l\in[n]\setminus U})+r_1+1-q_i^{(r_2)}\right| \\
&-\sum_{i\in V}\left|R(X(i),(X(l))_{l\in[n]\setminus U})+r_1-q_i^{(r_2)}\right| \\
&=2\sum_{i\in V}\mathrm{I}\left(R(X(i),(X(l))_{l\in[n]\setminus U})+r_1-q_i^{(r_2)}\geq 0\right)+m_2 \\
&=2S^{(r_1,r_2)}+m_2.
\end{aligned}$$

This proves (C.182).

Next, we show (C.183) is true. Notice that

$$\begin{aligned}
&\sum_{i\in U}\left|p_{i-m_2}^{(r_1+1)}-R(Y(i),(Y(l))_{l\in[n]\setminus V})-r_2\right| \\
&-\sum_{i\in U}\left|p_{i-m_2}^{(r_1)}-R(Y(i),(Y(l))_{l\in[n]\setminus V})-r_2\right| \\
&=\sum_{i=m_2+1}^{m_2+m_1}\left|p_{i-m_2}^{(r_1+1)}-R(Y(i),(Y(l))_{l\in[n]\setminus V})-r_2\right| \\
&-\sum_{i=m_2+1}^{m_2+m_1}\left|p_{i-m_2}^{(r_1)}-R(Y(i),(Y(l))_{l\in[n]\setminus V})-r_2\right| \\
&=\sum_{i=1}^{m_1}\left|p_{i}^{(r_1+1)}-R(Y(m_2+i),(Y(l))_{l\in[n]\setminus V})-r_2\right| \\
&-\sum_{i=1}^{m_1}\left|p_{i}^{(r_1)}-R(Y(m_2+i),(Y(l))_{l\in[n]\setminus V})-r_2\right|.
\end{aligned}$$

Then, according to Lemma C.5.5, we have

$$\begin{aligned}
&\sum_{i\in U}\left|p_{i}^{(r_1+1)}-R(Y(i),(Y(l))_{l\in[n]\setminus V})-r_2\right| \\
&-\sum_{i\in U}\left|p_{i}^{(r_1)}-R(Y(i),(Y(l))_{l\in[n]\setminus V})-r_2\right| \\
&=|r_1+1-R(Y(m_2+m_1-r_1),(Y(l))_{l\in[n]\setminus V})-r_2| \\
&-|n-m_1+r_1+1-R(Y(m_2+m_1-r_1),(Y(l))_{l\in[n]\setminus V})-r_2| \\
&=C_1^{(r_1,r_2)}.
\end{aligned}$$

This proves (C.183).

Next, we show (C.184) is true. According to Lemma C.5.4, we have

$$\begin{aligned}
&\sum_{i\in[n]\setminus(U\cup V)} |r_1+1-r_2-d_i| - \sum_{i\in[n]\setminus(U\cup V)} |r_1-r_2-d_i| \\
&= 2\sum_{i\in[n]\setminus(U\cup V)} \mathrm{I}\left(r_1-r_2-d_i\geq 0\right) + n - m_1 - m_2 \\
&= 2\sum_{i\in[n]\setminus(U\cup V)} \mathrm{I}\left(d_i\leq r_1-r_2\right) + n' \\
&= 2t_{r_1-r_2} + n'.
\end{aligned}$$

This proves (C.184) and completes our proof for (C.180).

Similarly, we can prove (C.181) is true. According to Proposition C.5.9, we have

$$\begin{aligned}
D(X_*^{(r_1)}, Y_*^{(r_2)}) = &\sum_{i\in V} \left| R(X(i), (X(l))_{l\in[n]\setminus U}) + r_1 - q_i^{(r_2)} \right| \\
&+ \sum_{i\in U} \left| p_i^{(r_1)} - R(Y(i), (Y(l))_{l\in[n]\setminus V}) - r_2 \right| \\
&+ \sum_{i\in[n]\setminus(U\cup V)} |r_1-r_2-d_i|,
\end{aligned}$$

and

$$\begin{aligned}
D(X_*^{(r_1)}, Y_*^{(r_2+1)}) = &\sum_{i\in V} \left| R(X(i), (X(l))_{l\in[n]\setminus U}) + r_1 - q_i^{(r_2+1)} \right| \\
&+ \sum_{i\in U} \left| p_i^{(r_1)} - R(Y(i), (Y(l))_{l\in[n]\setminus V}) - r_2 - 1 \right| \\
&+ \sum_{i\in[n]\setminus(U\cup V)} |r_1-r_2-1-d_i|.
\end{aligned}$$

Then, in order to prove (C.181), it is sufficient to show the following three equations all hold:

$$\begin{aligned}
&\sum_{i\in V} \left| R(X(i), (X(l))_{l\in[n]\setminus U}) + r_1 - q_i^{(r_2+1)} \right| \\
&- \sum_{i\in V} \left| R(X(i), (X(l))_{l\in[n]\setminus U}) + r_1 - q_i^{(r_2)} \right| = C_2^{(r_1,r_2)},
\end{aligned} \tag{C.185}$$

$$\begin{aligned}
&\sum_{i\in U} \left| p_i^{(r_1)} - R(Y(i), (Y(l))_{l\in[n]\setminus V}) - r_2 - 1 \right| \\
&- \sum_{i\in U} \left| p_i^{(r_1)} - R(Y(i), (Y(l))_{l\in[n]\setminus V}) - r_2 \right| = m_1 - 2R^{(r_1,r_2)},
\end{aligned} \tag{C.186}$$

and

$$\sum_{i\in[n]\setminus(U\cup V)} |r_1-r_2-d_i| - \sum_{i\in[n]\setminus(U\cup V)} |r_1-r_2-d_i-1| = n' - 2t_{r_1-r_2-1}. \tag{C.187}$$

First, we show (C.185) is true. Notice that

$$q_i^{(r_2)} = \mathrm{I}\left(i \leq m_2 - r_2\right)(n - i + 1) + \mathrm{I}\left(i > m_2 - r_2\right)(m_1 - i + 1),$$

for any $i \in \{1, \ldots, m_2\}$ and $r_2 \in \{0, \ldots, m_2\}$. Then, denote $u_i = i$ for any $i = 0, \ldots, m_2$. We have

$$q_{u_i}^{(r_1)} = \mathrm{I}\left(u_i \leq u_{m_2 - r_2}\right)(n - i + 1) + \mathrm{I}\left(u_i > u_{m_2 - r_2}\right)(m_2 - i + 1).$$

Then, according to Lemma C.5.5, we have

$$\begin{aligned}
&\sum_{i \in V} \left| R(X(i), (X(l))_{l \in [n] \setminus U}) + r_1 - q_i^{(r_2+1)} \right| \\
&- \sum_{i \in V} \left| R(X(i), (X(l))_{l \in [n] \setminus U}) + r_1 - q_i^{(r_2)} \right| \\
&= \sum_{i \in V} \left| q_i^{(r_2+1)} - R(X(i), (X(l))_{l \in [n] \setminus U}) - r_1 \right| \\
&- \sum_{i \in V} \left| q_i^{(r_2)} - R(X(i), (X(l))_{l \in [n] \setminus U}) - r_1 \right| \\
&= \left| r_2 + 1 - R(X(m_2 - r_2), (X(l))_{l \in [n] \setminus U}) - r_1 \right| \\
&- \left| n - m_2 + r_2 + 1 - R(X(m_2 - r_2), (X(l))_{l \in [n] \setminus U}) - r_1 \right| \\
&= C_2^{(r_1, r_2)}.
\end{aligned}$$

This proves (C.185).

Next, we show (C.186) is true. According to Lemma C.5.4, we have

$$\begin{aligned}
&\sum_{i \in U} \left| p_i^{(r_1)} - R(Y(i), (Y(l))_{l \in [n] \setminus V}) - r_2 - 1 \right| \\
&- \sum_{i \in U} \left| p_i^{(r_1)} - R(Y(i), (Y(l))_{l \in [n] \setminus V}) - r_2 \right| \\
&= m_1 - 2 \sum_{i \in U} \mathrm{I}\left(p_i^{(r_1)} - R(Y(i), (Y(l))_{l \in [n] \setminus V}) - r_2 - 1 \geq 0 \right) \\
&= m_1 - 2R^{(r_1, r_2)}.
\end{aligned}$$

This proves (C.186).

Next, we show (C.187) is true. According to Lemma C.5.4, we have

$$\begin{aligned}
&\sum_{i \in [n] \setminus (U \cup V)} |r_1 - r_2 - d_i| - \sum_{i \in [n] \setminus (U \cup V)} |r_1 - r_2 - d_i - 1| \\
&= n - m_1 - m_2 - 2 \sum_{i \in [n] \setminus (U \cup V)} \mathrm{I}\left(r_1 - r_2 - d_i - 1 \geq 0\right) \\
&= n' - 2 \sum_{i \in [n] \setminus (U \cup V)} \mathrm{I}\left(d_i \leq r_1 - r_2 - 1\right) \\
&= n' - 2t_{r_1 - r_2 - 1}.
\end{aligned}$$

This proves (C.187) and completes our proof for (C.181). Hence, we finish our proof.

□

Remark C.5.11. A few comments of Algorithm 6 are made below:

In Algorithm 6, the initialisation $D_{0,0}$ computed in line 10 equals to $D(X_*^{(0)}, Y_*^{(0)})$ defined in Proposition C.5.9. Spearman's footrule D_{r_1+1,r_2}, D_{r_1,r_2+1}, and D_{r_1,r_2} equal to $D(X_*^{(r_1+1)}, Y_*^{(r_2)})$, $D(X_*^{(r_1)}, Y_*^{(r_2+1)})$, and $D(X_*^{(r_1)}, Y_*^{(r_2)})$, respectively, and are updated according to Proposition C.5.10.

By computing $\{D_{0,0}, \ldots, D_{m_1,m_2}\}$, Algorithm 6 finds all possible Spearman's footrule values $D(X_*, Y_*)$, where $X_* \in \mathcal{X}(Y; U, [n], n)$, $Y_* \in \mathcal{X}(X; V, [n], n)$ are imputations of X, Y for indices U, and V, respectively. See discussions after Theorem 2.17 for explanations. Hence, according to Theorem C.4.13, the algorithm guarantees to find the maximum possible Spearman's footrule between X and Y.

The computational complexity of Algorithm 6 is analysed as follows. Ranking and relabelling all observed components in X and Y in line 2 requires $\mathcal{O}(n \log n)$ steps. Using these rankings, in line 4 computing each d_i is $\mathcal{O}(1)$, and so overall line 4 is $\mathcal{O}(n - m_1 - m_2)$. Line 5 and line 6 takes $\mathcal{O}(m_2^2)$ and $\mathcal{O}(m_1)$ steps, respectively.

According to Remark C.5.7, the computational complexity of running Algorithm 4 in line 7 and line 8 is $\mathcal{O}(m_1 + m_2)$, and $\mathcal{O}(m_2)$, respectively. In line 9, each iteration of running Algorithm 4 within the for loop is $\mathcal{O}(m_1)$. Since the loop runs m_2 times, the computational complexity for the loop is $\mathcal{O}(m_1 m_2)$.

Line 10 requires $\mathcal{O}(n)$ steps using the ranks of observed components. Line 13, line 14, line 17, and line 18 all take $\mathcal{O}(1)$ steps, and since the loop in 12 runs m_1 times, the computational complexity for the loop is $\mathcal{O}(m_1)$. Further, since the loop in 11 runs $m_2 + 1$ times, the computational complexity for the loop is $\mathcal{O}(m_1 m_2)$. Therefore, the overall computational complexity for Algorithm 6 is $\mathcal{O}(n \log n + m_1 m_2 + m_2^2)$. Note that, however, if $m_1 < m_2$, we can swap X and Y. Hence, the computational complexity for find the tight upper bounds of Spearman's footrule under missing case II can be reduced to $\mathcal{O}(n \log n + \min\{m_1, m_2\}^2 + m_1 m_2)$.

C.5.3. Algorithm for tight upper bound under missing case III

This section provides an efficient algorithm for computing exact upper bounds of Spearman's footrule under missing case III, where X and Y might be partially observed, and for any pair $(X(i), Y(i))$, where $i \in [n]$, the two value are either both observed, or both missing.

To start, we prove the following lemma:

Lemma C.5.12. Suppose $X, Y \in \mathcal{V}^n$ and for $0 < m_3 < n$, let $W = \{n - m_3 + 1, \ldots, n\} \subset [n]$ be a subset of indices. If $(X, Y) \in \mathcal{Y}(W, n)$ is such that

$$\sum_{i \in W} \mathrm{I}\left(X^{(r)}(i) < \min(X^{(r)}(l))_{l \in [n] \setminus W}\right) = r,$$

Algorithm 6 An efficient algorithm for computing exact upper bounds of Spearman's Footrule under missing case II

Input: $X, Y \in \mathcal{V}^n$, where X and Y might be partially observed, but for any pair $(X(i), Y(i))$, where $i \in [n]$, at least one value is observed.

Output: Maximum possible Spearman's footrule distance between X and Y.

1: Denote the number of unobserved components in X, Y as m_1 and m_2, respectively. If at least $m_1 = 0$ or $m_2 = 0$ is true, run Algorithm 5. If $m_1 + m_2 = n$, then return $\sum_{i=1}^{n} |i - (n-i+1)|$.

2: Rank all observed in X and Y. Relabel X and Y such that $(Y(1), \ldots, Y(m_2))$, and $(X(m_2+1), \ldots, X(m_2+m_1))$ are unobserved, and $X(1) < \ldots < X(m_2)$, $Y(m_1) < \ldots < Y(m_2+m_1)$.

3: Let $V = \{1, \ldots, m_2\}$, and $U = \{m_2+1, \ldots, m_2+m_1\}$.

4: For $i \in [n] \setminus (U \cup V)$, let $d_i = R(Y(i), (Y(l))_{l\in[n]\setminus V}) - R(X(i), (X(l))_{l\in[n]\setminus U})$.

5: For any $i \in \{1, \ldots, m_2\}$ and $r_2 \in \{0, \ldots, m_2\}$, let $q_i^{(r_2)} = \mathrm{I}(i \le m_2 - r_2)(n - i + 1) + \mathrm{I}(i > m_2 - r_2)(m_2 - i + 1)$.

6: For any $i \in \{1, \ldots, m_1\}$, let $p_i^{(0)} = \mathrm{I}(i \le m_1)(n - i + 1) + \mathrm{I}(i > m_1)(m_1 - i + 1)$.

7: Run Algorithm 4 for computing $s_i = \sum_{l\in[n]\setminus(U\cup V)} \mathrm{I}(d_l \le i)$ for any $i \in \{-m_2, \ldots, m_1 - 1\}$.

8: Run Algorithm 4 for computing

$$R^{(0,r_2)} = \sum_{i\in U} \mathrm{I}\left(-p_{i-m_2}^{(0)} + R(Y(i), (Y(l))_{l\in[n]\setminus V}) + 1 \le -r_2\right), \text{ for any } r_2 \in \{0, \ldots, m_2\}.$$

9: For any given $r_2 \in \{0, \ldots, m_2\}$, run Algorithm 4 for computing

$$S^{(r_1,r_2)} = \sum_{i\in V} \mathrm{I}\left(q_i^{(r_2)} - R(X(i), (X(l))_{l\in[n]\setminus U}) \le r_1\right), \text{ for any } r_1 \in \{0, \ldots, m_1\}.$$

10: Initialise

$$\begin{aligned} D_{0,0} = &\sum_{i\in V} \left|R(X(i), (X(l))_{l\in[n]\setminus U}) + r_1 - (n - i + 1)\right| \\ &+ \sum_{i\in U} \left|(n - i + m_2 + 1) - R(Y(i), (Y(l))_{l\in[n]\setminus V}) - r_2\right| + \sum_{i\in[n]\setminus(U\cup V)} |r_1 - r_2 - d_i|. \end{aligned}$$

11: **for** $r_2 = 0, \ldots, m_2$ **do**

12: **for** $r_1 = 0, \ldots, m_1 - 1$ **do**

13: Compute

$$\begin{aligned} C_1^{(r_1,r_2)} = &|r_1 + 1 - R(Y(m_2 + m_1 - r_1), (Y(l))_{l\in[n]\setminus V}) - r_2| \\ &- |n - m_1 + r_1 + 1 - R(Y(m_2 + m_1 - r_1), (Y(l))_{l\in[n]\setminus V}) - r_2|. \end{aligned}$$

14: Compute $D_{r_1+1,r_2} = D_{r_1,r_2} + 2s_{r_1-r_2} + n' + 2S^{(r_1,r_2)} + m_2 + C_1^{(r_1,r_2)}$.

15: **end for**

16: **if** $r_2 < m_2$ **then**

17: Let $r_1 = 0$ and compute

$$\begin{aligned} C_2^{(r_1,r_2)} = &|r_2 + 1 - R(X(m_2 - r_2), (X(l))_{l\in[n]\setminus U}) - r_1| \\ &- \left|n - m_2 + r_2 + 1 - R(X(m_2 - r_2), (X(l))_{l\in[n]\setminus U}) - r_1\right|. \end{aligned}$$

18: Compute $D_{r_1,r_2+1} = D_{r_1,r_2} - 2s_{r_1-r_2-1} + n' - 2R^{(r_1,r_2)} + m_1 + C_2^{(r_1,r_2)}$.

19: **end if**

20: **end for**

21: Return $\max\{D_{0,0}, \ldots, D_{m_1,m_2}\}$.

then we have $Y \in \mathcal{Z}(W,[n],n)$, $\sum_{i\in W} \mathrm{I}\left(Y(i) < \min(Y(l))_{l\in[n]\setminus W}\right) = m_3 - r$, and

$$\begin{aligned}\sum_{i\in W} |R(X(i),X) - R(Y(i),Y)| = &\sum_{i=1}^{m_3} \left(\mathrm{I}\left(i \le m_3 - r\right)|n+1-2i|\right) \\ &+ \sum_{i=1}^{m_3} \left(\mathrm{I}\left(i > m_3 - r\right)|2m_3 - 2i + 1 - n|\right).\end{aligned} \tag{C.188}$$

Proof. Without loss of generality, let us assume (after relabelling) $Y(n-m_3+1) < \ldots < Y(n)$.

Since $(X,Y) \in \mathcal{Y}(W,n)$, then we have $X \in \mathcal{Z}(W,[n],n)$, and

$$R(X(i),X) + R(Y(i),Y) = n+1, \text{ for any } i \in W. \tag{C.189}$$

Then, for any $i,j \in W$, we have

$$\text{if } R(X(i),X) > R(X(j),X), \text{ then } R(Y(i),Y) < R(Y(j),Y).$$

Thus, we have

$$X(i) > X(j) \implies Y(i) > Y(j), \text{ for any } i,j \in W.$$

Hence, we have shown that $X \in \mathcal{X}(Y;W,[n],n)$. Then, according to Lemma C.5.1, for any $i \in \{1,\ldots,m_3\}$, we have

$$R(X_*{}^{(r)}(n-m_3+i), X_*^{(r)}) = \begin{cases} n-i+1, & \text{if } r = 0, \\ q^{(r)}, & \text{if } m_3 > r > 0, \\ m_3 - i + 1, & \text{if } r = m_3, \end{cases} \tag{C.190}$$

where $q^{(r)} = \mathrm{I}\left(n - m_3 + i \le n - r\right)(n-i+1) + \mathrm{I}\left(n - m_3 + i > n - r\right)(m_3 - i + 1)$.

In the following, we prove that the following three statements:

$$\begin{aligned}&\text{(i)} : Y \in \mathcal{Z}(W,[n],n), \\ &\text{(ii)} : \sum_{i\in W} \mathrm{I}\left(Y(i) < \min(Y(l))_{l\in[n]\setminus W}\right) = m_3 - r, \\ &\text{(iii)} : \text{ equation (C.188) holds,}\end{aligned}$$

are true when $r = 0$, $m_3 > r > 0$ and $r = m_3$, separately.

Suppose $r = 0$. Then according to (C.190), we have

$$R(X(n-m_3+i),X) = n - i + 1, \text{ for any } i \in \{1,\ldots,m_3\}.$$

According to (C.189), we further have

$$R(Y(n-m_3+i),Y) = i, \text{ for any } i \in \{1,\ldots,m_3\}.$$

Since the m_3 components of W takes the ranks from 1 to m_3, then we have

$$Y(i) < \min(Y(l))_{l\in[n]\setminus W}, \text{ for any } i \in W.$$

Hence, we have $Y \in \mathcal{Z}(W,[n],n)$ and $\sum_{i\in W} \mathrm{I}\left(Y(i) < \min(Y(l))_{l\in[n]\setminus W}\right) = m_3 = m_3 - r$. Further, for any $i \in \{1,\ldots,m_3\}$, we have

$$R(X(n-m_3+i),X) - R(Y(n-m_3+i),Y) = n-i+1-i = n+1-2i,$$

which gives

$$\sum_{i\in W} |R(X(i),X) - R(Y(i),Y)| = \sum_{i=1}^{m_3} |n+1-2i|.$$

This proves our result when $r = 0$.

Suppose $m_3 > r > 0$. Then according to (C.190), we have

$$\begin{aligned} R(X(n-m_3+i),X) &= n-i+1, \text{ for any } i \in \{1,\ldots,m_3-r\}, \\ \text{and } R(X(n-m_3+i),X) &= m_3-i+1, \text{ for any } i \in \{m_3-r+1,\ldots,m_3\}. \end{aligned}$$

Then according to (C.189), we have

$$\begin{aligned} R(Y(n-m_3+i),Y) &= i, \text{ for any } i \in \{1,\ldots,m_3-r\}, \\ \text{and } R(Y(n-m_3+i),Y) &= n-m_3+i, \text{ for any } i \in \{m_3-r+1,\ldots,m_3\}, \end{aligned}$$

Since the m_3 components of W takes either the ranks from 1 to m_3-r, or the ranks from $n-r+1$ to n, then we have

$$Y(i) < \min(Y(l))_{l\in[n]\setminus W}, \text{ or } Y(i) > \min(Y(l))_{l\in[n]\setminus W}, \text{ for any } i \in W.$$

Hence, we have $Y \in \mathcal{Z}(W,[n],n)$ and $\sum_{i\in W} \mathrm{I}\left(Y(i) < \min(Y(l))_{l\in[n]\setminus W}\right) = m_3 - r$. Further, for any $i \in \{1,\ldots,m_3-r\}$, we have

$$R(X(n-m_3+i),X) - R(Y(n-m_3+i),Y) = n+1-2i,$$

and for any $i \in \{m_3-r+1,\ldots,m_3\}$, we have

$$R(X(n-m_3+i),X) - R(Y(n-m_3+i),Y) = 2m_3-2i+1-n.$$

Hence, we have

$$\begin{aligned} &\sum_{i\in W} |R(X(i),X) - R(Y(i),Y)| \\ &= \sum_{i=1}^{m_3-r} |n+1-2i| + \sum_{i=m_3-r+1}^{m_3} |2m_3-2i+1-n|. \end{aligned}$$

This proves our result when $m_3 > r > 0$.

Suppose $r = m_3$. Then according to (C.190), we have

$$R(X(n - m_3 + i), X) = m_3 - i + 1, \text{ for any } i \in \{1, \ldots, m_3\}.$$

Then according to (C.189), we have

$$R(Y(n - m_3 + i), Y) = n - m_3 + i, \text{ for any } i \in \{1, \ldots, m_3\}.$$

Since the m_3 components of W takes the ranks from $n - m_3 + 1$ to n, then we have

$$Y(i) > \max(Y(l))_{l \in [n] \setminus W}, \text{ for any } i \in W.$$

Hence, we have $Y \in \mathcal{Z}(W, [n], n)$ and $\sum_{i \in W} \mathrm{I}\left(Y(i) < \min(Y_*(l))_{l \in [n] \setminus W}\right) = m_3 - r$. Further, for any $i \in \{1, \ldots, m_3\}$, we have

$$R(X(n - m_3 + i), X) - R(Y(n - m_3 + i), Y) = 2m_3 - 2i + 1 - n,$$

which gives

$$\sum_{i \in W} (R(X(i), X) - R(Y(i), Y)) = \sum_{i=1}^{m_3} |2m_3 - 2i + 1 - n|.$$

This proves our result when $r = m_3$, and completes our proof. □

Proposition C.5.13. Suppose $X, Y \in \mathcal{V}^n$ and for $0 < m_3 < n$, let $W = \{n - m_3 + 1, \ldots, n\} \subset [n]$ be a subset of indices. Let $(X_*^{(r)}, Y_*^{(r)}) \in \mathcal{Y}(W, n)$ be imputations of X and Y for indices in W such that $\sum_{i \in W} \mathrm{I}\left(X_*{}^{(r)}(i) < \min(X_*{}^{(r)}(l))_{l \in [n] \setminus W}\right) = r$. For any $i \in [n] \setminus W$, denote $d_i = R(Y(i), (Y(j))_{j \in [n] \setminus W}) - R(X(i), (X(j))_{j \in [n] \setminus W})$. Then, for any $r \in \{0, \ldots, m_3\}$, we have

$$D(X_*^{(r)}, Y_*^{(r)}) = \sum_{i=1}^{n - m_3} |d_i + m_3 - 2r| + g^{(r)},$$

where $g^{(r)} = \sum_{i=1}^{m_3} \left(\mathrm{I}\left(i \leq m_3 - r\right) |n + 1 - 2i|\right) + \sum_{i=1}^{m_3} \left(\mathrm{I}\left(i > m_3 - r\right) |2m_3 - 2i + 1 - n|\right)$.

Proof. To start, according to the definition of Spearman's footrule, we have

$$\begin{aligned} D(X_*^{(r)}, Y_*^{(r)}) = & \sum_{i \in W} \left| R(Y_*^{(r)}(i), Y_*^{(r)}) - R(X_*{}^{(r)}(i), X_*^{(r)}) \right| \\ & + \sum_{i \in [n] \setminus W} \left| R(Y_*^{(r)}(i), Y_*^{(r)}) - R(X_*{}^{(r)}(i), X_*^{(r)}) \right|. \end{aligned}$$

Then, in order to prove our results, it is sufficient to show that

$$\sum_{i \in W} \left| R(Y_*^{(r)}(i), Y_*^{(r)}) - R(X_*{}^{(r)}(i), X_*^{(r)}) \right| = g^{(r)},$$

and

$$\sum_{i\in[n]\setminus W}\left|R(Y_*^{(r)}(i),Y_*^{(r)})-R({X_*}^{(r)}(i),X_*^{(r)})\right|=\sum_{i=1}^{n-m_3}|d_i+m_3-2r|.$$

Notice that

$$\sum_{i\in W}\left|R(Y_*^{(r)}(i),Y_*^{(r)})-R({X_*}^{(r)}(i),X_*^{(r)})\right|=g^{(r)}$$

is true according to Lemma C.5.12. Hence, we only need to show that

$$\sum_{i\in[n]\setminus W}\left|R(Y_*^{(r)}(i),Y_*^{(r)})-R({X_*}^{(r)}(i),X_*^{(r)})\right|=\sum_{i=1}^{n-m_3}|d_i+m_3-2r|.$$

Since $({X_*}^{(r)},Y_*^{(r)})\in\mathcal{Y}(W,n)$, then we have ${X_*}^{(r)}\in\mathcal{Z}(W,[n],n)$. Notice that

$$\sum_{i\in W}\mathrm{I}\left({X_*}^{(r)}(i)<\min({X_*}^{(r)}(l))_{l\in[n]\setminus W}\right)=r,$$

then according to Lemma C.5.2, we have

$$R({X_*}^{(r)}(i),X_*^{(r)})=R(X(i),(X(l))_{l\in[n]\setminus W})+r,\text{ for any }i\in[n]\setminus W.$$

Next, according to Lemma C.5.12, we have $Y_*^{(r)}\in\mathcal{Z}(W,[n],n)$, and

$$\sum_{i\in W}\mathrm{I}\left(Y_*^{(r)}(i)<\min(Y_*^{(r)}(l))_{l\in[n]\setminus W}\right)=m_3-r.$$

Then, according to Lemma C.5.2, we have

$$R(Y_*^{(r)}(i),Y_*^{(r)})=R(Y(i),(Y(l))_{l\in[n]\setminus W})+m_3-r,\text{ for any }i\in[n]\setminus W.$$

Hence, we have

$$\begin{aligned}&\sum_{i\in[n]\setminus W}\left|R(Y_*^{(r)}(i),Y_*^{(r)})-R({X_*}^{(r)}(i),X_*^{(r)})\right|\\&=\sum_{i\in[n]\setminus W}\left|R(Y(i),(Y(l))_{l\in[n]\setminus W})-R(X(i),(X(l))_{l\in[n]\setminus W})+m_3-2r\right|\\&=\sum_{i=1}^{n-m_3}|d_i+m_3-2r|.\end{aligned}$$

This completes our proof. □

Lemma C.5.14. Suppose $n\in\mathbb{N}$ such that $n>2$, and $m_3\in[n]$. For any $r\in\{0,\ldots,m_3\}$, define

$$g^{(r)}=\sum_{i=1}^{m_3}\left(\mathrm{I}\left(i\le m_3-r\right)|n+1-2i|\right)+\sum_{i=1}^{m_3}\left(\mathrm{I}\left(i>m_3-r\right)|2m_3-2i+1-n|\right).$$

Then for any $r \in \{0, \ldots, m_3 - 1\}$, we have

$$g^{(r+1)} - g^{(r)} = |2r + 1 - n| - |n + 1 - 2(m_3 - r)|.$$

Proof. We prove our results when $r = 0$, $m_3 > r > 0$, and $r = m_3$, separately.

First, suppose $r = 0$. Then we have

$$g^{(r)} = \sum_{i=1}^{m_3} |n + 1 - 2i|$$

and

$$g^{(r+1)} = \sum_{i=1}^{m_3-1} |n + 1 - 2i| + |1 - n|.$$

Hence, we have

$$g^{(r+1)} - g^{(r)} = |1 - n| - |n + 1 - 2m_3|.$$

This proves our results when $r = 0$.

Next, suppose $m_3 - 1 > r > 0$. Then we have

$$g^{(r)} = \sum_{i=1}^{m_3-r} |n + 1 - 2i| + \sum_{i=m_3-r+1}^{m_3} |2m_3 - 2i + 1 - n|,$$

and

$$g^{(r+1)} = \sum_{i=1}^{m_3-r-1} |n + 1 - 2i| + \sum_{i=m_3-r}^{m_3} |2m_3 - 2i + 1 - n|.$$

Hence, we have

$$\begin{aligned} g^{(r+1)} - g^{(r)} &= |2m_3 - 2(m_3 - r) + 1 - n| - |n + 1 - 2(m_3 - r)| \\ &= |2r + 1 - n| - |n + 1 - 2(m_3 - r)|. \end{aligned}$$

This proves our results when $m_3 - 1 > r > 0$.

Finally, suppose $r = m_3 - 1$. Then we have

$$g^{(r)} = |n - 1| + \sum_{i=2}^{m_3} |2m_3 - 2i + 1 - n|,$$

and

$$g^{(r+1)} = \sum_{i=1}^{m_3} |2m_3 - 2i + 1 - n|.$$

Hence, we have

$$g^{(r+1)} - g^{(r)} = |2m_3 - 1 - n| - |n - 1|.$$

This proves our results when $r = m_3 - 1$, and completes our proof. □

Lemma C.5.15. Suppose $X = (X(u_1), \ldots, X(u_n))$ is a vector of integers, where $U = \{u_1, \ldots, u_n\}$ is a set of indices such that $u_1 < \ldots < u_n$. Then, we have

$$\sum_{i \in U} |X(i) + 2| - \sum_{i \in U} |X(i)| = 2 \sum_{i \in U} \mathrm{I}\left(X(i) \geq 0\right) + 2 \sum_{i \in U} \mathrm{I}\left(X(i) + 1 \geq 0\right) - 2n.$$

Proof. To start, for any $i \in U$, we have

$$\begin{aligned} |X(i) + 2| - |X(i)| &= 2, \text{ if } X(i) \geq 0, \\ |X(i) + 2| - |X(i)| &= 0, \text{ if } X(i) = -1. \end{aligned}$$

If, however $X(i) < -1$, then since X is a vector of integers, we have $X(i) + 2 \leq 0$. Hence, we have

$$|X(i) + 2| - |X(i)| = -2, \text{ if } X(i) < -1.$$

Therefore, we have

$$\begin{aligned} \sum_{i \in U} |X(i) + 2| - \sum_{i \in U} |X(i)| &= \sum_{i \in U} \left(|X(i) + 2| - |X(i) - Y(i)|\right) \\ &= 2 \sum_{i \in U} \mathrm{I}\left(X(i) \geq 0\right) - 2 \sum_{i \in U} \mathrm{I}\left(X(i) < -1\right) \\ &= 2 \sum_{i \in U} \mathrm{I}\left(X(i) \geq 0\right) - 2 \left(n - \sum_{i \in U} \mathrm{I}\left(X(i) \geq -1\right)\right) \\ &= 2 \sum_{i \in U} \mathrm{I}\left(X(i) \geq 0\right) + 2 \sum_{i \in U} \mathrm{I}\left(X(i) + 1 \geq 0\right) - 2n. \end{aligned}$$

This completes our proof. □

Proposition C.5.16. Suppose $X, Y \in \mathcal{V}^n$ and for $0 < m_3 < n$, let $W = \{n - m_3 + 1, \ldots, n\} \subset [n]$ be a subset of indices. Let $({X_*}^{(r)}, Y_*^{(r)}) \in \mathcal{Y}(W, n)$ be imputations of X and Y for indices in W such that $\sum_{i \in W} \mathrm{I}\left({X_*}^{(r)}(i) < \min({X_*}^{(r)}(l))_{l \in [n] \setminus W}\right) = r$. For any $i \in [n] \setminus W$, let $d_i = R(Y(i), (Y(j))_{j \in [n] \setminus W}) - R(X(i), (X(j))_{j \in [n] \setminus W})$. For any $i \in \{-m_3, \ldots, m_3 - 1\}$, let $s_i = \sum_{l \in [n] \setminus W} \mathrm{I}\left(d_l \leq i\right)$. Then, for any $r \in \{0, \ldots, m_3 - 1\}$, we have

$$\begin{aligned} D(X_*^{(r+1)}, Y_*^{(r+1)}) = D(X_*^{(r)}, Y_*^{(r)}) &+ 2(s_{2r+1-m_3} + s_{2r-m_3} - n + m_3) \\ &+ |2r + 1 - n| - |n + 1 - 2(m_3 - r)|. \end{aligned}$$

Proof. To start, according to Proposition C.5.13, we have

$$D(X_*^{(r)}, Y_*^{(r)}) = \sum_{i=1}^{n-m_3} |d_i + m_3 - 2r| + g^{(r)},$$
$$\text{and } D(X_*^{(r+1)}, Y_*^{(r+1)}) = \sum_{i=1}^{n-m_3} |d_i + m_3 - 2r - 2| + g^{(r+1)}.$$

Hence,

$$\begin{aligned} &D(X_*^{(r+1)}, Y_*^{(r+1)}) - D(X_*^{(r)}, Y_*^{(r)}) \\ &= \sum_{i=1}^{n-m_3} |d_i + m_3 - 2r - 2| - \sum_{i=1}^{n-m_3} |d_i + m_3 - 2r| + g^{(r+1)} - g^{(r)}. \end{aligned}$$

Therefore, in order to prove our result, it is sufficient to show that

$$g^{(r+1)} - g^{(r)} = |2r + 1 - n| - |n + 1 - 2(m_3 - r)|,$$

and

$$\sum_{i=1}^{n-m_3} |d_i + m_3 - 2r - 2| - \sum_{i=1}^{n-m_3} |d_i + m_3 - 2r| = 2(s_{2r-m_3} + s_{2r-m_3+1} - n - m_3).$$

Notice that according to Lemma C.5.14,

$$g^{(r+1)} - g^{(r)} = |2r + 1 - n| - |n + 1 - 2(m_3 - r)|$$

is true. Hence we only need to show

$$\sum_{i=1}^{n-m_3} |d_i + m_3 - 2r - 2| - \sum_{i=1}^{n-m_3} |d_i + m_3 - 2r| = 2(s_{2r+1-m_3} + s_{2r-m_3} - n + m_3)$$

is true.

According to Lemma C.5.15, we have

$$\begin{aligned} &\sum_{i=1}^{n-m_3} |d_i + m_3 - 2r - 2| - \sum_{i=1}^{n-m_3} |d_i + m_3 - 2r| \\ &= 2(n - m_3) - 2\sum_{i=1}^{n-m_3} \mathrm{I}\left(d_i + m_3 - 2r - 2 \geq 0\right) - 2\sum_{i=1}^{n-m_3} \mathrm{I}\left(d_i + m_3 - 2r - 2 + 1 \geq 0\right) \\ &= 2(n - m_3) - 2\left(n - m_3 - \sum_{i=1}^{n-m_3} \mathrm{I}\left(d_i + m_3 - 2r - 2 < 0\right)\right) \\ &- 2\left(n - m_3 - \sum_{i=1}^{n-m_3} \mathrm{I}\left(d_i + m_3 - 2r - 1 < 0\right)\right) \\ &= 2\left(\sum_{i=1}^{n-m_3} \mathrm{I}\left(d_i < 2r + 2 - m_3\right) + \sum_{i=1}^{n-m_3} \mathrm{I}\left(d_i < 2r + 1 - m_3\right) - n + m_3\right). \end{aligned}$$

Since r, m_3 are integers and for any $i \in \{1, \ldots, n - m_3\}$, d_i is also an integer, then we have

$$\begin{aligned} &\mathrm{I}\left(d_i < 2r + 2 - m_3\right) = \mathrm{I}\left(d_i \leq 2r + 1 - m_3\right) \\ \text{and } &\mathrm{I}\left(d_i < 2r + 1 - m_3\right) = \mathrm{I}\left(d_i \leq 2r - m_3\right). \end{aligned}$$

Hence, we have

$$\sum_{i=1}^{n-m_3} |d_i + m_3 - 2r - 2| - \sum_{i=1}^{n-m_3} |d_i + m_3 - 2r| = 2(s_{2r+1-m_3} + s_{2r-m_3} - n + m_3).$$

This completes our proof. □

Algorithm 7 An efficient algorithm for computing exact upper bounds of Spearman's Footrule under Missing Case III

Input: $X, Y \in \mathcal{V}^n$, where X and Y might be partially observed, and for any pair (X_i, Y_i), where $i \in [n]$, the two values are either both observed, or both missing.

Output: Maximum possible Spearman's footrule distance between X and Y.

1: Denote m_3 as the number of components such that both values $(X(i), Y(i))$ are missing. If $m_3 = n$, then return $\sum_{i=1}^{n} |i - (n - i + 1)|$.

2: Rank all observed data in X and Y. Relabel X and Y such that $(X(n-m_3+1), Y(n-m_3+1)), \ldots, (X(n), Y(n))$ are missing.

3: Let $W = \{n - m_3 + 1, \ldots, n\}$, and for any $i \in [n] \setminus W$, let

$$d_i = R(Y(i), (Y(l))_{l \in [n] \setminus V}) - R(X(i), (X(l))_{l \in [n] \setminus U}).$$

4: Run Algorithm 4 for computing $s_i = \sum_{l \in [n] \setminus W} \mathrm{I}\left(d_l \leq i\right)$ for any $i \in \{-m_3, \ldots, m_3 - 1\}$.

5: Initialise $D_0 = \sum_{i=1}^{m_3} |n + 1 - 2i| + \sum_{i \in [n] \setminus W} |d_i|$.

6: **for** $r = 0, \ldots, m_3 - 1$ **do**

7: Compute

$$\begin{aligned} D_{r+1} = D_r &+ 2(s_{2r+1-m_3} + s_{2r-m_3} - n + m_3) \\ &+ |2r + 1 - n| - |n + 1 - 2(m_3 - r)|. \end{aligned}$$

8: **end for**

9: Return $\max\{D_0, \ldots, D_{m_3}\}$.

Remark C.5.17. A few comments of Algorithm 7 are made below:

In Algorithm 7, the initialisation D_0 computed in line 5 equals to $D(X_*^{(0)}, Y_*^{(0)})$ defined in Proposition C.5.13. Spearman's footrule D_{r+1}, and D_r equal to $D(X_*^{(r+1)}, Y_*^{(r)})$, $D(X_*^{(r)}, Y_*^{(r)})$, respectively, and are updated according to Proposition C.5.16.

By computing $\{D_0, \ldots, D_{m_3}\}$, Algorithm 7 finds all possible Spearman's footrule values $D\left(X_*, Y_*\right)$, where $(X_*, Y_*) \in \mathcal{Y}(W, n)$ are imputations of X and Y for indices W. See discussions following Theorem 2.20 in the main paper for explanations. Hence, according to Theorem C.4.18, the algorithm guarantees to find the maximum possible Spearman's footrule between X and Y.

The computational complexity of Algorithm 7 is analysed as follows. Ranking and relabelling all observed components in X and Y in line 2 requires $\mathcal{O}(n \log n)$ steps. Using these rankings, in line 3 computing each d_i is $\mathcal{O}(1)$, and so overall line 3 is $\mathcal{O}(n - m_3)$. According to Remark C.5.7, the computational complexity of running Algorithm 4 in line 4 is $\mathcal{O}(n + m_3)$. Line 5 requires

$\mathcal{O}(n)$ steps. In line 7, each iteration of the for loop is $\mathcal{O}(1)$, and since the loop runs m_3 times, the computational complexity for the loop is $\mathcal{O}(m_3)$. Therefore, the overall computational complexity for Algorithm 7 is $\mathcal{O}(n \log n)$.

C.5.4. Algorithm for tight upper bound under general missingness case

This subsection provides an efficient algorithm for computing upper bounds of Spearman's footrule under general missingness case.

We first prove the following result.

Lemma C.5.18. Suppose $X, Y \in \mathcal{V}^n$ and for $0 < m_1, m_2, m_3 < n$ such that $m_1+m_2+m_3 < n$, let $V = \{1, \ldots, m_2\}$, $U = \{m_2+1, \ldots, m_2+m_1\}$ and $W = \{n-m_3+1, \ldots, n\}$ be subsets of indices. Suppose $X(1) < \ldots < X(m_2)$, and $Y(m_2+1) < \ldots < Y(m_2+m_1)$. Let $(X_*{}^{(r_1,r_3)}, Y_*^{(r_2,r_3)}) \in (\mathcal{X}(Y; U, [n] \setminus W, n), \mathcal{X}(X; V, [n] \setminus W, n)) \cap \mathcal{Y}(W, n)$ be imputations of X, Y for indices U, V such that

$$\begin{aligned} &\sum_{i \in U} \mathrm{I}\left(X_*{}^{(r_1,r_3)}(i) < \min(X_*{}^{(r_1,r_3)}(l))_{l \in ([n]\setminus W)\setminus U}\right) = r_1, \\ &\sum_{i \in V} \mathrm{I}\left(Y_*^{(r_2,r_3)}(i) < \min(Y_*^{(r_2,r_3)}(l))_{l \in ([n]\setminus W)\setminus V}\right) = r_2, \\ \text{and } &\sum_{i \in W} \mathrm{I}\left(X_*{}^{(r_1,r_3)}(i) < \min(X_*{}^{(r_1,r_3)}(l))_{l \in [n]\setminus W}\right) = r_3. \end{aligned}$$

Then, for any $i \in ([n] \setminus W) \setminus U$, we have

$$R(X_*{}^{(r_1,r_3)}(i), X_*^{(r_1,r_3)}) = r_3 + r_1 + R(X(i), (X(l))_{l \in ([n]\setminus W)\setminus (U \cup V)}). \tag{C.191}$$

For any $i \in ([n] \setminus W) \setminus V$, we have

$$R(Y_*^{(r_2,r_3)}(i), Y_*^{(r_2,r_3)}) = m_3 - r_3 + r_2 + R(Y(i), (Y(l))_{l \in ([n]\setminus W)\setminus (U \cup V)}). \tag{C.192}$$

For any $r_1 \in \{0, \ldots, m_1\}$, $r_2 \in \{0, \ldots, m_2\}$, $r_3 \in \{0, \ldots, m_3\}$, we have

$$R(X_*{}^{(r_1,r_3)}(i+m_2), X_*^{(r_1,r_3)}) = p_i^{(r_1)} + r_3, \text{ for any } i \in \{1, \ldots, m_1\}, \tag{C.193}$$

$$\text{and } R(Y_*^{(r_2,r_3)}(i), Y_*^{(r_2,r_3)}) = q_i^{(r_2)} + m_3 - r_3, \text{ for any } i \in \{1, \ldots, m_2\}, \tag{C.194}$$

where $p_i^{(r_1)} = \mathrm{I}(i \le m_1 - r_1)(n-m_3-i+1) + \mathrm{I}(i > m_1 - r_1)(m_1-i+1)$, and $q_i^{(r_2)} = \mathrm{I}(i \le m_2 - r_2)(n-m_3-i+1) + \mathrm{I}(i > m_2 - r_2)(m_2-i+1)$.

Proof. First, we show (C.191) is true, since $(X_*{}^{(r_1,r_3)}, Y_*^{(r_2,r_3)}) \in \mathcal{Y}(W, n)$, then, according to the definition of $\mathcal{Y}(W, n)$, we have $X_*{}^{(r_1,r_3)} \in \mathcal{Z}(W, [n], n)$. Notice that, according to the definition of imputations, we have $X_*{}^{(r_1,r_3)}$ is an imputation of itself for indices W. Since $\sum_{i \in W} \mathrm{I}\left(X_*{}^{(r_1,r_3)}(i) < \min(X_*{}^{(r_1,r_3)}(l))_{l \in [n]\setminus W}\right) = r_3$, then according to Lemma C.5.2, for any $i \in [n] \setminus W$, we have

$$R(X_*{}^{(r_1,r_3)}(i), X_*^{(r_1,r_3)}) = R(X_*{}^{(r_1,r_3)}(i), (X_*^{(r_1,r_3)}(l))_{l \in [n]\setminus W}) + r_3. \tag{C.195}$$

Further, since ${X_*}^{(r_1,r_3)} \in \mathcal{X}(Y; U, [n] \setminus W, n)$, according to the definition of $\mathcal{X}(Y; U, [n] \setminus W, n)$, we also have ${X_*}^{(r_1,r_3)} \in \mathcal{Z}(U, [n] \setminus W, n)$. According to the definition of $\mathcal{Z}(U, [n] \setminus W, n)$, for any $i \in U$, we have

$$ {X_*}^{(r_1,r_3)}(i) > \max({X_*}^{(r_1,r_3)}(l))_{l\in[n]\setminus W}, \text{ or } {X_*}^{(r_1,r_3)}(i) < \min({X_*}^{(r_1,r_3)}(l))_{l\in[n]\setminus W}. $$

Hence, we also have $({X_*}^{(r_1,r_3)}(l))_{l\in[n]\setminus W} \in \mathcal{Z}(U, [n] \setminus W, n - m_3)$. Since $({X_*}^{(r_1,r_3)}(l))_{l\in[n]\setminus W}$ is an imputation of itself for indices U, and we have

$$ \sum_{i\in U} \mathrm{I}\left({X_*}^{(r_1,r_3)}(i) < \min({X_*}^{(r_1,r_3)}(l))_{l\in([n]\setminus W)\setminus U} \right) = r_1, $$

then according to Lemma C.5.2, for any $i \in ([n] \setminus W) \setminus U$, we have

$$ \begin{aligned} &R({X_*}^{(r_1,r_3)}(i), (X_*^{(r_1,r_3)}(l))_{l\in[n]\setminus W}) \\ &= R({X_*}^{(r_1,r_3)}(i), (X_*^{(r_1,r_3)}(l))_{l\in([n]\setminus W)\setminus U}) + r_1. \end{aligned} \tag{C.196} $$

Since $X_*^{(r_1,r_3)}$ is an imputation of X for indices $U \cup W$, for any $i \in ([n] \setminus W) \setminus U$, we have

$$ R({X_*}^{(r_1,r_3)}(i), (X_*^{(r_1,r_3)}(l))_{l\in([n]\setminus W)\setminus U}) = R(X(i), (X(l))_{l\in([n]\setminus W)\setminus U}). \tag{C.197} $$

Combining (C.195), (C.196), and (C.197), for any $i \in ([n] \setminus W) \setminus U$, we have

$$ R({X_*}^{(r_1,r_3)}(i), X_*^{(r_1,r_3)}) = r_3 + r_1 + R(X(i), (X(l))_{l\in([n]\setminus W)\setminus(U\cup V)}). $$

This proves (C.191).

Similarly, we can show (C.192) is true. Since $({X_*}^{(r_1,r_3)}, Y_*^{(r_2,r_3)}) \in \mathcal{Y}(W, n)$, then according to Lemma C.5.12, we have $Y_*^{(r_2,r_3)} \in \mathcal{Z}(W, [n], n)$, and

$$ \sum_{i\in W} \mathrm{I}\left(Y_*^{(r_2,r_3)}(i) < \min(Y_*^{(r_2,r_3)}(l))_{l\in[n]\setminus W} \right) = m_3 - r_3. $$

Notice that, according to the definition of imputations, $Y_*^{(r_2,r_3)}$ is an imputation of itself for indices W. Hence, according to Lemma C.5.2, for any $i \in [n] \setminus W$, we have

$$ R(Y_*^{(r_2,r_3)}(i), Y_*^{(r_2,r_3)}) = R(Y_*^{(r_2,r_3)}(i), (Y_*^{(r_2,r_3)}(l))_{l\in[n]\setminus W}) + m_3 - r_3. \tag{C.198} $$

Further, since $Y_*^{(r_2,r_3)} \in \mathcal{X}(X; V, [n] \setminus W, n)$, we have $Y_*^{(r_2,r_3)} \in \mathcal{Z}(V, [n] \setminus W, n)$. Then according to the definition of $\mathcal{Z}(V, [n] \setminus W, n)$, we have

$$ Y_*^{(r_2,r_3)}(i) > \max(Y_*^{(r_2,r_3)}(l))_{l\in[n]\setminus W}, \text{ or } Y_*^{(r_2,r_3)}(i) < \min(Y_*^{(r_2,r_3)}(l))_{l\in[n]\setminus W}, \text{ for } i \in V. $$

Hence, we also have $(Y_*^{(r_2,r_3)}(l))_{l\in[n]\setminus W} \in \mathcal{Z}(V, [n] \setminus W, n - m_3)$. Since $(Y_*^{(r_2,r_3)}(l))_{l\in[n]\setminus W}$ is an imputation of itself for indices V, and we have

$$ \sum_{i\in V} \mathrm{I}\left(Y_*^{(r_2,r_3)}(i) < \min(Y_*^{(r_2,r_3)}(l))_{l\in([n]\setminus W)\setminus V} \right) = r_2. $$

then according to Lemma C.5.2, for any $i \in ([n] \setminus W) \setminus V$, we have

$$\begin{aligned} &R(Y_*^{(r_2,r_3)}(i), (Y_*^{(r_2,r_3)}(l))_{l\in[n]\setminus W}) \\ &= R(Y_*^{(r_2,r_3)}(i), (Y_*^{(r_2,r_3)}(l))_{l\in([n]\setminus W)\setminus V}) + r_2. \end{aligned} \tag{C.199}$$

Since $Y_*^{(r_2,r_3)}$ is an imputation of Y for indices $V \cup W$, for any $i \in ([n] \setminus W) \setminus V$, we have

$$R(Y_*^{(r_2,r_3)}(i), (Y_*^{(r_2,r_3)}(l))_{l\in([n]\setminus W)\setminus V}) = R(Y(i), (Y(l))_{l\in([n]\setminus W)\setminus V}). \tag{C.200}$$

Combining (C.198), (C.199), and (C.200), for any $i \in ([n] \setminus W) \setminus V$, we have

$$R(Y_*^{(r_2,r_3)}(i), Y_*^{(r_2,r_3)}) = m_3 - r_3 + r_2 + R(Y(i), (Y(l))_{l\in([n]\setminus W)\setminus(U\cup V)}).$$

This proves (C.192) is true.

Next, we show (C.193) is true, since ${X_*}^{(r_1,r_3)} \in \mathcal{X}(Y; U, [n] \setminus W, n)$, then we have

$${X_*}^{(r_1,r_3)}(i) > \max({X_*}^{(r_1,r_3)}(l))_{l\in[n]\setminus W}, \text{ or } {X_*}^{(r_1,r_3)}(i) < \min({X_*}^{(r_1,r_3)}(l))_{l\in[n]\setminus W}, \text{ for any } i \in U,$$
$$\text{and } {X_*}^{(r_1,r_3)}(i) > {X_*}^{(r_1,r_3)}(j), \text{ if } Y(i) < Y(j), \text{ for any } i, j \in U.$$

Hence, we have $({X_*}^{(r_1,r_3)}(l))_{l\in[n]\setminus W} \in \mathcal{X}((Y(l))_{l\in[n]\setminus W}; U, [n] \setminus W, n - m_3)$. Since

$$\sum_{i\in U} \mathrm{I}\left({X_*}^{(r_1,r_3)}(i) < \min({X_*}^{(r_1,r_3)}(l))_{l\in([n]\setminus W)\setminus U}\right) = r_1,$$

then according to Lemma C.5.1, for any $i \in \{1, \ldots, m_1\}$, we have

$$R({X_*}^{(r_1,r_3)}(m_2 + i), ({X_*}^{(r_1,r_3)}(l))_{l\in[n]\setminus W}) = \begin{cases} n - m_3 - i + 1, & \text{if } r_1 = 0, \\ p_i^{(r_1)}, & \text{if } m_1 > r > 0, \\ m_1 - i + 1, & \text{if } r_1 = m_2, \end{cases}$$

where $p_i^{(r_1)} = \mathrm{I}(m_2 + i \le m_2 + m_1 - r_1)(n - m_3 - i + 1) + \mathrm{I}(m_2 + i > m_2 + m_1 - r_1)(m_1 - i + 1) = \mathrm{I}(i \le m_1 - r_1)(n - m_3 - i + 1) + \mathrm{I}(i > m_1 - r_1)(m_1 - i + 1)$.

Notice that $R({X_*}^{(r_1,r_3)}(i), ({X_*}^{(r_1,r_3)}(l))_{l\in[n]\setminus W}) = p_i^{(r_1)}$ is still true when $r_1 = 0$, or $r_1 = m_2$. Combining this result with (C.195), we have

$$R({X_*}^{(r_1,r_3)}(i), X_*^{(r_1,r_3)}) = p_i^{(r_1)} + r_3, \text{ for any } i \in U.$$

This completes our proof for (C.193).

Next, we show (C.193) is true, since $Y_*^{(r_2,r_3)} \in \mathcal{X}(X; V, [n] \setminus W, n)$, then we have

$$Y_*^{(r_2,r_3)}(i) > \max(Y_*^{(r_2,r_3)}(l))_{l\in[n]\setminus W}, \text{ or } Y_*^{(r_2,r_3)}(i) < \min(Y_*^{(r_2,r_3)}(l))_{l\in[n]\setminus W}, \text{ for any } i \in V,$$
$$\text{and } Y_*^{(r_2,r_3)}(i) > Y_*^{(r_2,r_3)}(j), \text{ if } X(i) < X(j), \text{ for any } i, j \in V.$$

Hence, we have $(Y_*^{(r_1,r_3)}(l))_{l\in[n]\setminus W} \in \mathcal{X}((X(l))_{l\in[n]\setminus W}; V, [n]\setminus W, n-m_3)$. Since

$$\sum_{i\in V} \mathrm{I}\left(Y_*^{(r_2,r_3)}(i) < \min(Y_*^{(r_2,r_3)}(l))_{l\in([n]\setminus W)\setminus V}\right) = r_2,$$

then according to Lemma C.5.1, for any $i \in \{1,\ldots,m_2\}$, we have

$$\begin{aligned}
&R(Y_*^{(r_2,r_3)}(i), (Y_*^{(r_2,r_3)}(l))_{l\in[n]\setminus W}) \\
&= \begin{cases} n-m_3-i+1, & \text{if } r_2 = 0, \\ \mathrm{I}\left(i \le m_2 - r_2\right)(n-m_3-i+1) + \mathrm{I}\left(i > m_2-r_2\right)(m_2-i+1), & \text{if } m_2 > r_2 > 0, \\ m_2-i+1, & \text{if } r_2 = m_2 \end{cases} \\
&= \mathrm{I}\left(i \le m_2 - r_2\right)(n-m_3-i+1) + \mathrm{I}\left(i > m_2-r_2\right)(m_2-i+1) \\
&= p_i^{(r_2)}.
\end{aligned}$$

Combining this result with (C.198), we have

$$R(Y_*^{(r_2,r_3)}(i), Y_*^{(r_2,r_3)}) = p_i^{(r_2)} + m_3 - r_3.$$

This proves (C.194) is true, and completes our proof. □

Using the results of Lemma C.5.18, we can show the following result:

Proposition C.5.19. Suppose $X, Y \in \mathcal{V}^n$ and for $0 < m_1, m_2, m_3 < n$ such that $m_1 + m_2 + m_3 < n$, let $V = \{1,\ldots,m_2\}$, $U = \{m_2+1,\ldots,m_2+m_1\}$ and $W = \{n-m_3+1,\ldots,n\}$ be subsets of indices. Suppose $X(1) < \ldots < X(m_2)$, and $Y(m_2+1) < \ldots < Y(m_2+m_1)$. Let $(X_*^{(r_1,r_3)}, Y_*^{(r_2,r_3)}) \in (\mathcal{X}(Y; U, [n]\setminus W, n), \mathcal{X}(X; V, [n]\setminus W, n)) \cap \mathcal{Y}(W, n)$ be imputations of X, Y for indices $U\cup W, V\cup W$ such that

$$\begin{aligned}
&\sum_{i\in U} \mathrm{I}\left(X_*^{(r_1,r_3)}(i) < \min(X_*^{(r_1,r_3)}(l))_{l\in([n]\setminus W)\setminus U}\right) = r_1, \\
&\sum_{i\in V} \mathrm{I}\left(Y_*^{(r_2,r_3)}(i) < \min(Y_*^{(r_2,r_3)}(l))_{l\in([n]\setminus W)\setminus V}\right) = r_2, \\
\text{and } &\sum_{i\in W} \mathrm{I}\left(X_*^{(r_1,r_3)}(i) < \min(X_*^{(r_1,r_3)}(l))_{l\in[n]\setminus W}\right) = r_3.
\end{aligned}$$

Denote $d_i = R(Y(i), (Y(l))_{l\in([n]\setminus W)\setminus V}) - R(X(i), (X(l))_{l\in([n]\setminus W)\setminus U})$ for any $i \in ([n]\setminus W)\setminus(U\cup V)$. Then for any $r_1 \in \{0,\ldots,m_1\}$, $r_2 \in \{0,\ldots,m_2\}$, and $r_2 \in \{0,\ldots,m_3\}$, we have

$$\begin{aligned}
D\left(X_*^{(r_1,r_3)}, Y_*^{(r_2,r_3)}\right) = &\sum_{i\in V} \left|R(X(i), (X(l))_{l\in([n]\setminus W)\setminus U}) + r_1 - q_i^{(r_2)} + 2r_3 - m_3\right| \\
&+ \sum_{i\in U} \left|p_{i-m_2}^{(r_1)} - R(Y(i), (Y(l))_{l\in([n]\setminus W)\setminus V}) - r_2 + 2r_3 - m_3\right| \\
&+ \sum_{i\in([n]\setminus W)\setminus(U\cup V)} |r_1 - d_i - r_2 + 2r_3 - m_3| + g^{(r_3)},
\end{aligned}$$

where $p_i^{(r_1)} = \mathrm{I}\left(i \le m_1 - r_1\right)(n-m_3-i+1) + \mathrm{I}\left(i > m_1 - r_1\right)(m_1-i+1)$, $q_i^{(r_2)} = \mathrm{I}\left(i \le m_2 - r_2\right)(n-$

$m_3 - i + 1) + \mathrm{I}\left(i > m_2 - r_2\right)(m_2 - i + 1)$, and

$$g^{(r_3)} = \sum_{i=1}^{m_3} \left(\mathrm{I}\left(i \le m_3 - r_3\right)|n + 1 - 2i|\right) + \sum_{i=1}^{m_3} \left(\mathrm{I}\left(i > m_3 - r_3\right)|2m_3 - 2i + 1 - n|\right).$$

Proof. To start, according to Lemma C.5.18, we have

$$R(X_*{}^{(r_1,r_3)}(i), X_*^{(r_1,r_3)}) = r_3 + r_1 + R(X(i), (X(l))_{l\in([n]\setminus W)\setminus(U\cup V)}),$$

for any $i \in V$ and

$$R(Y_*^{(r_2,r_3)}(i), Y_*^{(r_2,r_3)}) = q_i^{(r_2)} + m_3 - r_3, \text{ for any } i \in \{1, \ldots, m_2\}.$$

Hence, we have

$$\begin{aligned}
&\sum_{i\in V} \left|R(X_*{}^{(r_1,r_3)}(i), X_*^{(r_1,r_3)}) - R(Y_*^{(r_2,r_3)}(i), Y_*^{(r_2,r_3)})\right| \\
&= \sum_{i\in V} \left|R(X(i), (X(l))_{l\in([n]\setminus W)\setminus U}) + r_1 - q_i^{(r_2)} + 2r_3 - m_3\right|.
\end{aligned} \tag{C.201}$$

Next, according to Lemma C.5.18, we also have

$$R(Y_*^{(r_2,r_3)}(i), Y_*^{(r_2,r_3)}) = m_3 - r_3 + r_2 + R(Y(i), (Y(l))_{l\in([n]\setminus W)\setminus(U\cup V)}),$$

for any $i \in U$ and

$$\begin{aligned}
&R(X_*{}^{(r_1,r_3)}(i + m_2), X_*^{(r_1,r_3)}) = p_i^{(r_1)} + r_3, \text{ for any } i \in \{1, \ldots, m_1\}. \\
\Longrightarrow\ &R(X_*{}^{(r_1,r_3)}(i), X_*^{(r_1,r_3)}) = p_{i-m_2}^{(r_1)} + r_3, \text{ for any } i \in \{m_2 + 1, \ldots, m_2 + m_1\}.
\end{aligned}$$

Hence,

$$\begin{aligned}
&\sum_{i\in U} \left|R(X_*{}^{(r_1,r_3)}(i), X_*^{(r_1,r_3)}) - R(Y_*^{(r_2,r_3)}(i), Y_*^{(r_2,r_3)})\right| \\
&= \sum_{i\in U} \left|p_{i-m_2}^{(r_1)} + 2r_3 - m_3 - r_2 - R(Y(i), (Y(l))_{l\in([n]\setminus W)\setminus(U\cup V)})\right|.
\end{aligned} \tag{C.202}$$

Next, according to Lemma C.5.18, for any $i \in ([n] \setminus W) \setminus (U \cup V)$, we have

$$\begin{aligned}
&R(X_*{}^{(r_1,r_3)}(i), X_*^{(r_1,r_3)}) = r_3 + r_1 + R(X(i), (X(l))_{l\in([n]\setminus W)\setminus(U\cup V)}), \\
\text{and } &R(Y_*^{(r_2,r_3)}(i), Y_*^{(r_2,r_3)}) = m_3 - r_3 + r_2 + R(Y(i), (Y(l))_{l\in([n]\setminus W)\setminus(U\cup V)}).
\end{aligned}$$

Hence,

$$\begin{aligned}
&\sum_{i\in([n]\setminus W)\setminus(U\cup V)} \left|R(X_*{}^{(r_1,r_3)}(i), X_*^{(r_1,r_3)}) - R(Y_*^{(r_2,r_3)}(i), Y_*^{(r_2,r_3)})\right| \\
&= \sum_{i\in([n]\setminus W)\setminus(U\cup V)} |2r_3 + r_1 - m_3 - r_2 - d_i|.
\end{aligned} \tag{C.203}$$

Now, since $({X_*}^{(r_1,r_3)}, Y_*^{(r_2,r_3)}) \in \mathcal{Y}(W,n)$ and

$$\sum_{i\in W} \mathrm{I}\left({X_*}^{(r_1,r_3)}(i) < \min({X_*}^{(r_1,r_3)}(l))_{l\in[n]\setminus W}\right) = r_3,$$

then according to Lemma C.5.12, we have

$$\sum_{i\in W} \left| R({X_*}^{(r_1,r_3)}(i), X_*^{(r_1,r_3)}) - R(Y_*^{(r_2,r_3)}(i), Y_*^{(r_2,r_3)}) \right| = g^{(r_3)}. \tag{C.204}$$

Finally, combining (C.201), (C.202), (C.203) and (C.204), we have

$$\begin{aligned}
&D\left(X_*^{(r_1,r_3)}, Y_*^{(r_2,r_3)}\right)\\
&= \sum_{i=1}^{n} \left| R({X_*}^{(r_1,r_3)}(i), X_*^{(r_1,r_3)}) - R(Y_*^{(r_2,r_3)}(i), Y_*^{(r_2,r_3)}) \right|\\
&= \sum_{i\in V} \left| R({X_*}^{(r_1,r_3)}(i), X_*^{(r_1,r_3)}) - R(Y_*^{(r_2,r_3)}(i), Y_*^{(r_2,r_3)}) \right|\\
&+ \sum_{i\in U} \left| R({X_*}^{(r_1,r_3)}(i), X_*^{(r_1,r_3)}) - R(Y_*^{(r_2,r_3)}(i), Y_*^{(r_2,r_3)}) \right|\\
&+ \sum_{i\in([n]\setminus W)\setminus(U\cup V)} \left| R({X_*}^{(r_1,r_3)}(i), X_*^{(r_1,r_3)}) - R(Y_*^{(r_2,r_3)}(i), Y_*^{(r_2,r_3)}) \right|\\
&+ \sum_{i\in W} \left| R({X_*}^{(r_1,r_3)}(i), X_*^{(r_1,r_3)}) - R(Y_*^{(r_2,r_3)}(i), Y_*^{(r_2,r_3)}) \right|\\
&= \sum_{i\in V} \left| R(X(i), (X(l))_{l\in([n]\setminus W)\setminus U}) + r_1 - q_i^{(r_2)} + 2r_3 - m_3 \right|\\
&+ \sum_{i\in U} \left| p_{i-m_2}^{(r_1)} - R(Y(i), (Y(l))_{l\in([n]\setminus W)\setminus V}) - r_2 + 2r_3 - m_3 \right|\\
&+ \sum_{i\in([n]\setminus W)\setminus(U\cup V)} |r_1 - d_i - r_2 + 2r_3 - m_3| + g^{(r_3)}.
\end{aligned}$$

This completes our proof. □

Finally, we show the following result:

Proposition C.5.20. Suppose $X, Y \in \mathcal{V}^n$ and for $0 < m_1, m_2, m_3 < n$ such that $m_1 + m_2 + m_3 < n$, let $V = \{1, \ldots, m_2\}$, $U = \{m_2 + 1, \ldots, m_2 + m_1\}$ and $W = \{n - m_3 + 1, \ldots, n\}$ be subsets of indices. Suppose $X(1) < \ldots < X(m_2)$, and $Y(m_2 + 1) < \ldots < Y(m_2 + m_1)$. Let $({X_*}^{(r_1,r_3)}, Y_*^{(r_2,r_3)}) \in (\mathcal{X}(Y; U, [n] \setminus W, n), \mathcal{X}(X; V, [n] \setminus W, n)) \cap \mathcal{Y}(W, n)$ be imputations of X, Y for indices $U \cup W$, and $V \cup W$, respectively, such that

$$\begin{aligned}
&\sum_{i\in U} \mathrm{I}\left({X_*}^{(r_1,r_3)}(i) < \min({X_*}^{(r_1,r_3)}(l))_{l\in([n]\setminus W)\setminus U}\right) = r_1,\\
&\sum_{i\in V} \mathrm{I}\left(Y_*^{(r_2,r_3)}(i) < \min(Y_*^{(r_2,r_3)}(l))_{l\in([n]\setminus W)\setminus V}\right) = r_2,\\
\text{and } &\sum_{i\in W} \mathrm{I}\left({X_*}^{(r_1,r_3)}(i) < \min({X_*}^{(r_1,r_3)}(l))_{l\in[n]\setminus W}\right) = r_3.
\end{aligned}$$

For any $r_1 \in \{0,\dots,m_1\}$, $r_2 \in \{0,\dots,m_2\}$, and $r_3 \in \{0,\dots,m_3\}$, denote

$$S^{(r_1,r_2,r_3)} = \sum_{i\in V} \mathrm{I}\left(R(X(i),(X(l))_{l\in([n]\setminus W)\setminus U}) + r_1 - q_i^{(r_2)} + 2r_3 - m_3 \geq 0\right),$$
$$R^{(r_1,r_2,r_3)} = \sum_{i\in U} \mathrm{I}\left(p_{i-m_2}^{(r_1)} - R(Y(i),(Y(l))_{l\in([n]\setminus W)\setminus V}) - r_2 - 1 + 2r_3 - m_3 \geq 0\right),$$

where $p_i^{(r_1)} = \mathrm{I}\left(i \leq m_1 - r_1\right)(n-m_3-i+1) + \mathrm{I}\left(i > m_1 - r_1\right)(m_1-i+1)$, $q_i^{(r_2)} = \mathrm{I}\left(i \leq m_2 - r_2\right)(n - m_3 - i + 1) + \mathrm{I}\left(i > m_2 - r_2\right)(m_2 - i + 1)$. For any $i \in \{-m_1 - 1 - 2m_3, \dots, m_2\}$, let $s_i = \sum_{l\in([n]\setminus W)\setminus(U\cup V)} \mathrm{I}\left(d_i \leq i\right)$, and denote $n' = n - m_1 - m_2 - m_3$. Then, for any $r_1 \in \{0,\dots,m_1-1\}$, $r_2 \in \{0,\dots,m_2\}$, and $r_3 \in \{0,\dots,m_3\}$, we have

$$\begin{aligned} D(X_*^{(r_1+1,r_3)}, Y_*^{(r_2,r_3)}) &= D(X_*^{(r_1,r_3)}, Y_*^{(r_2,r_3)}) \\ &\quad + 2S^{(r_1,r_2,r_3)} - 2s_{m_3+r_2-r_1-2r_3-1} + C_1^{(r_1,r_2,r_3)}, \end{aligned} \tag{C.205}$$

for any $r_1 \in \{0,\dots,m_1\}$, $r_2 \in \{0,\dots,m_2-1\}$, and $r_3 \in \{0,\dots,m_3\}$, we have

$$\begin{aligned} D(X_*^{(r_1,r_3)}, Y_*^{(r_2+1,r_3)}) &= D(X_*^{(r_1,r_3)}, Y_*^{(r_2,r_3)}) \\ &\quad - 2R^{(r_1,r_2,r_3)} + 2s_{m_3+r_2-r_1-2r_3} + C_2^{(r_1,r_2,r_3)}. \end{aligned} \tag{C.206}$$

and for any $r_1 \in \{0,\dots,m_1\}$, $r_2 \in \{0,\dots,m_2\}$, and $r_3 \in \{0,\dots,m_3-1\}$, we have

$$\begin{aligned} &D\left(X_*^{(r_1,r_3+1)}, Y_*^{(r_2,r_3+1)}\right) = D\left(X_*^{(r_1,r_3)}, Y_*^{(r_2,r_3)}\right) + |2r_3 + 1 - n| \\ &- |n + 1 - 2(m_3 - r_3)| + 2S^{(r_1,r_2,r_3)} + 2S^{(r_1+1,r_2,r_3)} - 2m_2 \\ &+ 2R^{(r_1,r_2-1,r_3)} + 2R^{(r_1,r_2-2,r_3)} - 2m_1 \\ &+ 2s_{2r_3+r_1-r_2-m_3} + 2s_{2r_3+r_1-r_2-m_3+1} - 2n'. \end{aligned} \tag{C.207}$$

where for any $r_1 \in \{0,\dots,m_1\}$, $r_2 \in \{0,\dots,m_2\}$, and $r_3 \in \{0,\dots,m_3\}$,

$$\begin{aligned} C_1^{(r_1,r_2,r_3)} &= -m_2 + n' \\ &+ \left|r_1 + 1 - R(Y(m_1 + m_2 - r_1),(Y(l))_{l\in([n]\setminus W)\setminus V}) - r_2 + 2r_3 - m_3\right| \\ &- \left|n - m_3 - m_1 + r_1 + 1 - R(Y(m_1 + m_2 - r_1),(Y(l))_{l\in([n]\setminus W)\setminus V}) - r_2 + 2r_3 - m_3\right|, \end{aligned}$$

and

$$\begin{aligned} C_2^{(r_1,r_2,r_3)} &= m_1 - n' \\ &+ \left|r_2 + 1 - R(X(m_2 - r),(X(l))_{l\in([n]\setminus W)\setminus U}) - r_1 - 2r_3 + m_3\right| \\ &- \left|n - m_3 - m_2 + r_2 + 1 - R(X(m_2 - r),(X(l))_{l\in([n]\setminus W)\setminus U}) - r_1 - 2r_3 + m_3\right|. \end{aligned}$$

Proof. First, we prove (C.205) is true. According to Proposition C.5.19, we have

$$\begin{aligned}
&D\left(X_*^{(r_1,r_3)}, Y_*^{(r_2,r_3)}\right)\\
&=\sum_{i\in V}\left|R(X(i),(X(l))_{l\in([n]\setminus W)\setminus U})+r_1-q_i^{(r_2)}+2r_3-m_3\right|\\
&+\sum_{i\in U}\left|p_{i-m_2}^{(r_1)}-R(Y(i),(Y(l))_{l\in([n]\setminus W)\setminus V})-r_2+2r_3-m_3\right|\\
&+\sum_{i\in([n]\setminus W)\setminus(U\cup V)}|r_1-d_i-r_2+2r_3-m_3|+g^{(r_3)},
\end{aligned}$$

and

$$\begin{aligned}
&D\left(X_*^{(r_1+1,r_3)}, Y_*^{(r_2,r_3)}\right)\\
&=\sum_{i\in V}\left|R(X(i),(X(l))_{l\in([n]\setminus W)\setminus U})+r_1+1-q_i^{(r_2)}+2r_3-m_3\right|\\
&+\sum_{i\in U}\left|p_{i-m_2}^{(r_1+1)}-R(Y(i),(Y(l))_{l\in([n]\setminus W)\setminus V})-r_2+2r_3-m_3\right|\\
&+\sum_{i\in([n]\setminus W)\setminus(U\cup V)}|r_1+1-d_i-r_2+2r_3-m_3|+g^{(r_3)},
\end{aligned}$$

Hence, we have

$$\begin{aligned}
&D\left(X_*^{(r_1+1,r_3)}, Y_*^{(r_2,r_3)}\right)-D\left(X_*^{(r_1,r_3)}, Y_*^{(r_2,r_3)}\right)\\
&=\sum_{i\in V}\left|R(X(i),(X(l))_{l\in([n]\setminus W)\setminus U})+r_1+1-q_i^{(r_2)}+2r_3-m_3\right|\\
&-\sum_{i\in V}\left|R(X(i),(X(l))_{l\in([n]\setminus W)\setminus U})+r_1-q_i^{(r_2)}+2r_3-m_3\right|\\
&+\sum_{i\in U}\left|p_{i-m_2}^{(r_1+1)}-R(Y(i),(Y(l))_{l\in([n]\setminus W)\setminus V})-r_2+2r_3-m_3\right|\\
&-\sum_{i\in U}\left|p_{i-m_2}^{(r_1)}-R(Y(i),(Y(l))_{l\in([n]\setminus W)\setminus V})-r_2+2r_3-m_3\right|\\
&+\sum_{i\in([n]\setminus W)\setminus(U\cup V)}|r_1+1-d_i-r_2+2r_3-m_3|\\
&-\sum_{i\in([n]\setminus W)\setminus(U\cup V)}|r_1-d_i-r_2+2r_3-m_3|.
\end{aligned}\tag{C.208}$$

Notice that, according to Lemma C.5.4, we have

$$\begin{aligned}
&\sum_{i\in V}\left|R(X(i),(X(l))_{l\in([n]\setminus W)\setminus U})+r_1+1-q_i^{(r_2)}+2r_3-m_3\right|\\
&-\sum_{i\in V}\left|R(X(i),(X(l))_{l\in([n]\setminus W)\setminus U})+r_1-q_i^{(r_2)}+2r_3-m_3\right|\\
&=2\sum_{i\in V}\mathrm{I}\left(R(X(i),(X(l))_{l\in([n]\setminus W)\setminus U})+r_1-q_i^{(r_2)}+2r_3-m_3\geq 0\right)-m_2,\\
&=2S^{(r_1,r_2,r_3)}-m_2,
\end{aligned}\tag{C.209}$$

and

$$\begin{aligned}
&\sum_{i\in([n]\setminus W)\setminus(U\cup V)} |d_i + r_1 + 1 - r_2 + 2r_3 - m_3| - \sum_{i\in([n]\setminus W)\setminus(U\cup V)} |d_i + r_1 - r_2 + 2r_3 - m_3| \\
&= 2 \sum_{i\in([n]\setminus W)\setminus(U\cup V)} \mathrm{I}\left(d_i + r_1 - r_2 + 2r_3 - m_3 \geq 0\right) - (n - m_1 - m_2 - m_3) \\
&= 2 \sum_{i\in([n]\setminus W)\setminus(U\cup V)} \mathrm{I}\left(d_i \geq m_3 + r_2 - r_1 - 2r_3\right) - n' \\
&= 2n' - 2 \sum_{i\in([n]\setminus W)\setminus(U\cup V)} \mathrm{I}\left(d_i < m_3 + r_2 - r_1 - 2r_3\right) - n'.
\end{aligned}$$

Since m_3, r_1, r_2, r_3 are integers, and for any $i \in ([n] \setminus W) \setminus (U \cup V)$, d_i is an integer, we have

$$\mathrm{I}\left(d_i < m_3 + r_2 - r_1 - 2r_3\right) = \mathrm{I}\left(d_i \leq m_3 + r_2 - r_1 - 2r_3 - 1\right).$$

Thus, we have

$$\begin{aligned}
&\sum_{i\in([n]\setminus W)\setminus(U\cup V)} |d_i + r_1 + 1 - r_2 + 2r_3 - m_3| \\
&- \sum_{i\in([n]\setminus W)\setminus(U\cup V)} |d_i + r_1 - r_2 + 2r_3 - m_3| \\
&= n' - 2 \sum_{i\in([n]\setminus W)\setminus(U\cup V)} \mathrm{I}\left(d_i \leq m_3 + r_2 - r_1 - 2r_3 - 1\right) \\
&= n' - 2s_{m_3+r_2-r_1-2r_3-1}.
\end{aligned} \tag{C.210}$$

Next, notice that

$$\begin{aligned}
&\sum_{i\in U} \left| p_{i-m_2}^{(r_1+1)} - R(Y(i), (Y(l))_{l\in([n]\setminus W)\setminus V}) - r_2 + 2r_3 - m_3 \right| \\
&- \sum_{i\in U} \left| p_{i-m_2}^{(r_1)} - R(Y(i), (Y(l))_{l\in([n]\setminus W)\setminus V}) - r_2 + 2r_3 - m_3 \right| \\
=& \sum_{i=m_2+1}^{m_2+m_1} \left| p_{i-m_2}^{(r_1+1)} - R(Y(i), (Y(l))_{l\in([n]\setminus W)\setminus V}) - r_2 + 2r_3 - m_3 \right| \\
-& \sum_{i=m_2+1}^{m_2+m_1} \left| p_{i-m_2}^{(r_1)} - R(Y(i), (Y(l))_{l\in([n]\setminus W)\setminus V}) - r_2 + 2r_3 - m_3 \right| \\
=& \sum_{i=1}^{m_1} \left| p_{i}^{(r_1+1)} - R(Y(m_2+i), (Y(l))_{l\in([n]\setminus W)\setminus V}) - r_2 + 2r_3 - m_3 \right| \\
-& \sum_{i=1}^{m_1} \left| p_{i}^{(r_1)} - R(Y(m_2+i), (Y(l))_{l\in([n]\setminus W)\setminus V}) - r_2 + 2r_3 - m_3 \right|,
\end{aligned}$$

where $p_i^{(r_1)} = \mathrm{I}\left(i \leq m_1 - r_1\right)(n - m_3 - i + 1) + \mathrm{I}\left(i > m_1 - r_1\right)(m_1 - i + 1)$ for any $i \in \{1, \ldots, m_1\}$

and $r_1 \in \{0, \ldots, m_1\}$. Then, according to Lemma C.5.5, we have

$$\begin{aligned}
&\sum_{i=1}^{m_1} \left| p_i^{(r_1+1)} - R(Y(m_2+i), (Y(l))_{l\in([n]\setminus W)\setminus V}) - r_2 + 2r_3 - m_3 \right| \\
-&\sum_{i=1}^{m_1} \left| p_i^{(r_1)} - R(Y(m_2+i), (Y(l))_{l\in([n]\setminus W)\setminus V}) - r_2 + 2r_3 - m_3 \right| \\
&= \left| r_1 + 1 - R(Y(m_1+m_2-r_1), (Y(l))_{l\in([n]\setminus W)\setminus V}) - r_2 + 2r_3 - m_3 \right| \\
&- \left| n - m_3 - m_1 + r_1 + 1 - R(Y(m_1+m_2-r_1), (Y(l))_{l\in([n]\setminus W)\setminus V}) - r_2 + 2r_3 - m_3 \right| \\
&= C_1^{(r_1,r_2,r_3)} + m_2 - n'.
\end{aligned}$$

Hence, we have

$$\begin{aligned}
&\sum_{i\in U} \left| p_{i-m_2}^{(r_1+1)} - R(Y(i), (Y(l))_{l\in([n]\setminus W)\setminus V}) - r_2 + 2r_3 - m_3 \right| \\
&- \sum_{i\in U} \left| p_{i-m_2}^{(r_1)} - R(Y(i), (Y(l))_{l\in([n]\setminus W)\setminus V}) - r_2 + 2r_3 - m_3 \right| \\
&= C_1^{(r_1,r_2,r_3)} + m_2 - n'.
\end{aligned} \tag{C.211}$$

Put (C.209), (C.210), and (C.211) back into (C.208), we obtain

$$\begin{aligned}
&D(X_*^{(r_1+1,r_3)}, Y_*^{(r_2,r_3)}) - D(X_*^{(r_1,r_3)}, Y_*^{(r_2,r_3)}) \\
&= 2S^{(r_1,r_2,r_3)} - m_2 + n' - 2t_{m_3+r_2-r_1-2r_3-1} + C_1^{(r_1,r_2,r_3)} + m_2 - n' \\
&= 2S^{(r_1,r_2,r_3)} - 2s_{m_3+r_2-r_1-2r_3-1} + C_1^{(r_1,r_2,r_3)}.
\end{aligned}$$

This proves (C.205) is true.

Similarly, we can show (C.206) is true. According to Proposition C.5.19, we have

$$\begin{aligned}
&D\left(X_*^{(r_1,r_3)}, Y_*^{(r_2,r_3)}\right) \\
&= \sum_{i\in V} \left| R(X(i), (X(l))_{l\in([n]\setminus W)\setminus U}) + r_1 - q_i^{(r_2)} + 2r_3 - m_3 \right| \\
&+ \sum_{i\in U} \left| p_{i-m_2}^{(r_1)} - R(Y(i), (Y(l))_{l\in([n]\setminus W)\setminus V}) - r_2 + 2r_3 - m_3 \right| \\
&+ \sum_{i\in([n]\setminus W)\setminus(U\cup V)} |r_1 - d_i - r_2 + 2r_3 - m_3| + q^{(r_3)},
\end{aligned}$$

and

$$\begin{aligned}
&D\left(X_*^{(r_1,r_3)}, Y_*^{(r_2+1,r_3)}\right) \\
&= \sum_{i\in V} \left| R(X(i), (X(l))_{l\in([n]\setminus W)\setminus U}) + r_1 - q_i^{(r_2+1)} + 2r_3 - m_3 \right| \\
&+ \sum_{i\in U} \left| p_{i-m_2}^{(r_1)} - R(Y(i), (Y(l))_{l\in([n]\setminus W)\setminus V}) - r_2 - 1 + 2r_3 - m_3 \right| \\
&+ \sum_{i\in([n]\setminus W)\setminus(U\cup V)} |r_1 - d_i - r_2 - 1 + 2r_3 - m_3| + q^{(r_3)},
\end{aligned}$$

Hence, we have

$$
\begin{aligned}
&D\left(X_*^{(r_1,r_3)}, Y_*^{(r_2+1,r_3)}\right) - D\left(X_*^{(r_1,r_3)}, Y_*^{(r_2,r_3)}\right) \\
&= \sum_{i\in V}\left|R(X(i),(X(l))_{l\in([n]\setminus W)\setminus U}) + r_1 - q_i^{(r_2+1)} + 2r_3 - m_3\right| \\
&- \sum_{i\in V}\left|R(X(i),(X(l))_{l\in([n]\setminus W)\setminus U}) + r_1 - q_i^{(r_2)} + 2r_3 - m_3\right| \\
&+ \sum_{i\in U}\left|p_{i-m_2}^{(r_1)} - R(Y(i),(Y(l))_{l\in([n]\setminus W)\setminus V}) - r_2 - 1 + 2r_3 - m_3\right| \\
&- \sum_{i\in U}\left|p_{i-m_2}^{(r_1)} - R(Y(i),(Y(l))_{l\in([n]\setminus W)\setminus V}) - r_2 + 2r_3 - m_3\right| \\
&+ \sum_{i\in([n]\setminus W)\setminus(U\cup V)} |r_1 - d_i - r_2 - 1 + 2r_3 - m_3| \\
&- \sum_{i\in([n]\setminus W)\setminus(U\cup V)} |r_1 - d_i - r_2 + 2r_3 - m_3| .
\end{aligned}
\tag{C.212}
$$

Notice that, according to Lemma C.5.4, we have

$$
\begin{aligned}
&\sum_{i\in U}\left|p_{i-m_2}^{(r_1)} - R(Y(i),(Y(l))_{l\in([n]\setminus W)\setminus V}) - r_2 - 1 + 2r_3 - m_3\right| \\
&- \sum_{i\in U}\left|p_{i-m_2}^{(r_1)} - R(Y(i),(Y(l))_{l\in([n]\setminus W)\setminus V}) - r_2 + 2r_3 - m_3\right| \\
&= -2\sum_{i\in U}\mathrm{I}\left(p_{i-m_2}^{(r_1)} - R(Y(i),(Y(l))_{l\in([n]\setminus W)\setminus V}) - r_2 - 1 + 2r_3 - m_3 \geq 0\right) + m_1, \\
&= -2R^{(r_1,r_2,r_3)} + m_1,
\end{aligned}
\tag{C.213}
$$

and

$$
\begin{aligned}
&\sum_{i\in([n]\setminus W)\setminus(U\cup V)} |d_i + r_1 - r_2 - 1 + 2r_3 - m_3| - \sum_{i\in([n]\setminus W)\setminus(U\cup V)} |d_i + r_1 - r_2 + 2r_3 - m_3| \\
&= -2\sum_{i\in([n]\setminus W)\setminus(U\cup V)} \mathrm{I}\left(d_i + r_1 - r_2 - 1 + 2r_3 - m_3 \geq 0\right) + (n - m_1 - m_2 - m_3) \\
&= -2\sum_{i\in([n]\setminus W)\setminus(U\cup V)} \mathrm{I}\left(d_i \geq m_3 + r_2 - r_1 - 2r_3 + 1\right) + n' \\
&= -2\left(n' - \sum_{i\in([n]\setminus W)\setminus(U\cup V)} \mathrm{I}\left(d_i < m_3 + r_2 - r_1 - 2r_3 + 1\right)\right) + n'.
\end{aligned}
$$

Since m_3, r_1, r_2, r_3 are integers, and for any $i \in ([n] \setminus W) \setminus (U \cup V)$, d_i is an integer, we have

$$
\mathrm{I}\left(d_i < m_3 + r_2 - r_1 - 2r_3 + 1\right) = \mathrm{I}\left(d_i \leq m_3 + r_2 - r_1 - 2r_3\right).
$$

Hence, we have

$$\begin{aligned}
&\sum_{i\in([n]\setminus W)\setminus(U\cup V)} |d_i + r_1 - r_2 - 1 + 2r_3 - m_3| \\
&- \sum_{i\in([n]\setminus W)\setminus(U\cup V)} |d_i + r_1 - r_2 + 2r_3 - m_3| \\
&= 2s_{m_3+r_2-r_1-2r_3} - n'.
\end{aligned} \tag{C.214}$$

Next, notice that

$$\begin{aligned}
&\sum_{i\in V} \left| R(X(i), (X(l))_{l\in([n]\setminus W)\setminus U}) + r_1 - q_i^{(r_2+1)} + 2r_3 - m_3 \right| \\
&- \sum_{i\in V} \left| R(X(i), (X(l))_{l\in([n]\setminus W)\setminus U}) + r_1 - q_i^{(r_2)} + 2r_3 - m_3 \right| \\
&= \sum_{i\in V} \left| q_i^{(r_2+1)} - R(X(i), (X(l))_{l\in([n]\setminus W)\setminus U}) - r_1 - 2r_3 + m_3 \right| \\
&- \sum_{i\in V} \left| q_i^{(r_2)} - R(X(i), (X(l))_{l\in([n]\setminus W)\setminus U}) - r_1 - 2r_3 + m_3 \right|.
\end{aligned}$$

Since $q_i^{(r_2)} = \mathrm{I}\left(i \le m_2 - r_2\right)(n - m_3 - i + 1) + \mathrm{I}\left(i > m_2 - r_2\right)(m_2 - i + 1)$, then according to Lemma C.5.5, we have

$$\begin{aligned}
&\sum_{i\in V} \left| q_i^{(r_2+1)} - R(X(i), (X(l))_{l\in([n]\setminus W)\setminus U}) - r_1 - 2r_3 + m_3 \right| \\
&- \sum_{i\in V} \left| q_i^{(r_2)} - R(X(i), (X(l))_{l\in([n]\setminus W)\setminus U}) - r_1 - 2r_3 + m_3 \right| \\
&= \left| r_2 + 1 - R(X(m_2 - r), (X(l))_{l\in([n]\setminus W)\setminus U}) - r_1 - 2r_3 + m_3 \right| \\
&- \left| n - m_3 - m_2 + r_2 + 1 - R(X(m_2 - r), (X(l))_{l\in([n]\setminus W)\setminus U}) - r_1 - 2r_3 + m_3 \right| \\
&= C_2^{(r_1,r_2,r_3)} - m_1 + n'.
\end{aligned}$$

Hence, we have

$$\begin{aligned}
&\sum_{i\in V} \left| R(X(i), (X(l))_{l\in([n]\setminus W)\setminus U}) + r_1 - q_i^{(r_2+1)} + 2r_3 - m_3 \right| \\
&- \sum_{i\in V} \left| R(X(i), (X(l))_{l\in([n]\setminus W)\setminus U}) + r_1 - q_i^{(r_2)} + 2r_3 - m_3 \right| \\
&= C_2^{(r_1,r_2,r_3)} - m_1 + n'.
\end{aligned} \tag{C.215}$$

Put (C.213), (C.214), and (C.215) back into (C.212), we obtain

$$\begin{aligned}
&D(X_*^{(r_1,r_3)}, Y_*^{(r_2+1,r_3)}) - D(X_*^{(r_1,r_3)}, Y_*^{(r_2,r_3)}) \\
&= -2R^{(r_1,r_2,r_3)} + m_1 + 2t_{m_3+r_2-r_1-2r_3} - n' + C_2^{(r_1,r_2,r_3)} - m_1 + n' \\
&= -2R^{(r_1,r_2,r_3)} + 2s_{m_3+r_2-r_1-2r_3} + C_2^{(r_1,r_2,r_3)}.
\end{aligned}$$

This proves (C.206) is true.

Finally, we show (C.207) is true. According to Proposition C.5.19, we have

$$\begin{aligned}
&D\left(X_*^{(r_1,r_3)}, Y_*^{(r_2,r_3)}\right)\\
&=\sum_{i\in V}\left|R(X(i),(X(l))_{l\in([n]\setminus W)\setminus U})+r_1-q_i^{(r_2)}+2r_3-m_3\right|\\
&+\sum_{i\in U}\left|p_{i-m_2}^{(r_1)}-R(Y(i),(Y(l))_{l\in([n]\setminus W)\setminus V})-r_2+2r_3-m_3\right|\\
&+\sum_{i\in([n]\setminus W)\setminus(U\cup V)}|r_1-d_i-r_2+2r_3-m_3|+g^{(r_3)},
\end{aligned}$$

and

$$\begin{aligned}
&D\left(X_*^{(r_1,r_3+1)}, Y_*^{(r_2,r_3+1)}\right)\\
&=\sum_{i\in V}\left|R(X(i),(X(l))_{l\in([n]\setminus W)\setminus U})+r_1-q_i^{(r_2)}+2r_3+2-m_3\right|\\
&+\sum_{i\in U}\left|p_{i-m_2}^{(r_1)}-R(Y(i),(Y(l))_{l\in([n]\setminus W)\setminus V})-r_2+2r_3+2-m_3\right|\\
&+\sum_{i\in([n]\setminus W)\setminus(U\cup V)}|r_1-d_i-r_2+2r_3+2-m_3|+g^{(r_3+1)},
\end{aligned}$$

Hence, we have

$$\begin{aligned}
&D\left(X_*^{(r_1,r_3+1)}, Y_*^{(r_2,r_3+1)}\right)-D\left(X_*^{(r_1,r_3)}, Y_*^{(r_2,r_3)}\right)\\
&=\sum_{i\in V}\left|R(X(i),(X(l))_{l\in([n]\setminus W)\setminus U})+r_1-q_i^{(r_2)}+2r_3+2-m_3\right|\\
&-\sum_{i\in V}\left|R(X(i),(X(l))_{l\in([n]\setminus W)\setminus U})+r_1-q_i^{(r_2)}+2r_3-m_3\right|\\
&+\sum_{i\in U}\left|p_{i-m_2}^{(r_1)}-R(Y(i),(Y(l))_{l\in([n]\setminus W)\setminus V})-r_2+2r_3+2-m_3\right|\\
&-\sum_{i\in U}\left|p_{i-m_2}^{(r_1)}-R(Y(i),(Y(l))_{l\in([n]\setminus W)\setminus V})-r_2+2r_3-m_3\right|\\
&+\sum_{i\in([n]\setminus W)\setminus(U\cup V)}|r_1-d_i-r_2+2r_3+2-m_3|\\
&-\sum_{i\in([n]\setminus W)\setminus(U\cup V)}|r_1-d_i-r_2+2r_3-m_3|+g^{(r_3+1)}-g^{(r_3)}.
\end{aligned}\tag{C.216}$$

According to Lemma C.5.14, we have

$$g^{(r_3+1)}-g^{(r_3)}=|2r_3+1-n|-|n+1-2(m_3-r_3)|.\tag{C.217}$$

According to Lemma C.5.15, we have

$$\begin{aligned}
&\sum_{i\in V}\left|R(X(i),(X(l))_{l\in([n]\setminus W)\setminus U})+r_1-q_i^{(r_2)}+2r_3+2-m_3\right|\\
&-\sum_{i\in V}\left|R(X(i),(X(l))_{l\in([n]\setminus W)\setminus U})+r_1-q_i^{(r_2)}+2r_3-m_3\right|\\
&=2\sum_{i\in V}\mathrm{I}\left(r_1+2r_3+R(X(i),(X(l))_{l\in([n]\setminus W)\setminus(U\cup V)})-q_i^{(r_2)}-m_3\geq 0\right)\\
&+2\sum_{i\in V}\mathrm{I}\left(r_1+2r_3+R(X(i),(X(l))_{l\in([n]\setminus W)\setminus(U\cup V)})-q_i^{(r_2)}-m_3+1\geq 0\right)-2m_2\\
&=2S^{(r_1,r_2,r_3)}+2S^{(r_1+1,r_2,r_3)}-2m_2.
\end{aligned}\tag{C.218}$$

and

$$\begin{aligned}
&\sum_{i\in U}\left|p_{i-m_2}^{(r_1)}-R(Y(i),(Y(l))_{l\in([n]\setminus W)\setminus V})-r_2+2r_3+2-m_3\right|\\
&-\sum_{i\in U}\left|p_{i-m_2}^{(r_1)}-R(Y(i),(Y(l))_{l\in([n]\setminus W)\setminus V})-r_2+2r_3-m_3\right|\\
&=2\sum_{i\in U}\mathrm{I}\left(p_{i-m_2}^{(r_1)}+2r_3-m_3-r_2-R(Y(i),(Y(l))_{l\in([n]\setminus W)\setminus(U\cup V)})\geq 0\right)\\
&+2\sum_{i\in U}\mathrm{I}\left(p_{i-m_2}^{(r_1)}+2r_3-m_3-r_2-R(Y(i),(Y(l))_{l\in([n]\setminus W)\setminus(U\cup V)})+1\geq 0\right)-2m_1\\
&=2R^{(r_1,r_2-1,r_3)}+2R^{(r_1,r_2-2,r_3)}-2m_1,
\end{aligned}\tag{C.219}$$

and

$$\begin{aligned}
&\sum_{i\in([n]\setminus W)\setminus(U\cup V)}|r_1-d_i-r_2+2r_3+2-m_3|\\
&-\sum_{i\in([n]\setminus W)\setminus(U\cup V)}|r_1-d_i-r_2+2r_3-m_3|\\
&=2\sum_{i\in([n]\setminus W)\setminus(V\cup U)}\mathrm{I}\left(r_1-d_i-r_2+2r_3-m_3\geq 0\right)\\
&+2\sum_{i\in([n]\setminus W)\setminus(V\cup U)}\mathrm{I}\left(r_1-d_i-r_2+2r_3-m_3+1\geq 0\right)-2(n-m_1-m_2-m_3)\\
&=2\sum_{i\in([n]\setminus W)\setminus(V\cup U)}\mathrm{I}\left(d_i\leq 2r_3+r_1-r_2-m_3\right)\\
&+2\sum_{i\in([n]\setminus W)\setminus(V\cup U)}\mathrm{I}\left(d_i\leq 2r_3+r_1-r_2-m_3+1\right)-2n'\\
&=2s_{2r_3+r_1-r_2-m_3}+2s_{2r_3+r_1-r_2-m_3+1}-2n'.
\end{aligned}\tag{C.220}$$

Put (C.217), (C.218), (C.219) and (C.220) back into (C.216), we obtain

$$\begin{aligned}
&D\left(X_*^{(r_1,r_3+1)},Y_*^{(r_2,r_3+1)}\right)-D\left(X_*^{(r_1,r_3)},Y_*^{(r_2,r_3)}\right)\\
&=|2r_3+1-n|-|n+1-2(m_3-r_3)|+2S^{(r_1,r_2,r_3)}+2S^{(r_1+1,r_2,r_3)}-2m_2\\
&+2R^{(r_1,r_2-1,r_3)}+2R^{(r_1,r_2-2,r_3)}-2m_1+2s_{2r_3+r_1-r_2-m_3}+2s_{2r_3+r_1-r_2-m_3+1}-2n'.
\end{aligned}$$

This proves (C.207), and completes our proof. □

Remark C.5.21. A few comments of Algorithm 8 are made below:

In Algorithm 8, the initialisation $D_{0,0,0}$ computed in line 15 equals to $D(X_*^{(0,0)}, Y_*^{(0,0)})$ defined in Proposition C.5.19. Spearman's footrule D_{r_1+1,r_2,r_3}, D_{r_1,r_2+1,r_3}, D_{r_1,r_2,r_3+1}, and D_{r_1,r_2,r_3} equal to $D(X_*^{(r_1+1,r_3)}, Y_*^{(r_2,r_3)})$, $D(X_*^{(r_1,r_3)}, Y_*^{(r_2+1,r_3)})$, $D(X_*^{(r_1,r_3+1)}, Y_*^{(r_2,r_3+1)})$, and $D(X_*^{(r_1,r_3)}, Y_*^{(r_2,r_3)})$, respectively, and are updated according to Proposition C.5.20.

By computing $\{D_{0,0,0}, \ldots, D_{m_1,m_2,m_3}\}$, Algorithm 8 finds all possible Spearman's footrule values $D(X_*, Y_*)$, where $(X_*, Y_*) \in (\mathcal{X}(Y; U, [n]\backslash W, n), \mathcal{X}(X; V, [n]\backslash W, n)) \cap \mathcal{Y}(W, n)$ are imputations X and Y for indices $U \cup W$ and $V \cup W$, respectively. See discussions following Theorem 2.21 in the main paper for explanations. Hence, according to Theorem C.4.19, the algorithm guarantees to find the maximum possible Spearman's footrule between X and Y.

The computational complexity of Algorithm 8 is analysed as follows. Ranking and relabelling all observed components in X and Y in line 3 requires $\mathcal{O}(n \log n)$ steps. Using these ranks, in line 5 computing each d_i is $\mathcal{O}(1)$, and so overall line 5 is $\mathcal{O}(n - m_1 - m_2 - m_3)$. Line 6 and line 7 takes $\mathcal{O}(m_2^2)$ and $\mathcal{O}(m_1)$ steps, respectively.

According to Remark C.5.7, the computational complexity of running Algorithm 4 in line 8 and line 9 is $\mathcal{O}(m_1 + m_2 + m_3)$, and $\mathcal{O}(m_2 + m_3)$, respectively. Line 9 takes $\mathcal{O}(m_2 m_3)$ steps. In line 12, each iteration of running Algorithm 4 within the for loop is $\mathcal{O}(m_1 + m_3)$. Line 13 takes $\mathcal{O}(m_1 m_3)$. Since the loop runs $m_2 + 1$ times, the computational complexity for the loop is $\mathcal{O}(m_1 m_2 m_3)$.

Line 15 requires $\mathcal{O}(n)$ steps using the ranks of observed components. Line 19, line 20, line 24, and line 25 all take $\mathcal{O}(1)$ steps. Since the loop in line 18 runs m_1 times, the computational complexity for the loop is $\mathcal{O}(m_1)$. Since line 29 takes $\mathcal{O}(1)$ steps, each iteration within the loop in line 17 takes $\mathcal{O}(m_1 + 1)$ steps. Since the loop in line 17 runs m_2 times, the computational complexity for this loop is $\mathcal{O}(m_1 m_2)$. Finally, the loop in line 16 runs m_3 times. Hence, the computational complexity for this loop is $\mathcal{O}(m_1 m_2 m_3)$.

Therefore, the overall computational complexity for Algorithm 6 is $\mathcal{O}(n \log n + m_2^2 + m_1 m_2 m_3)$. Note that, however, if $m_1 < m_2$, we can swap X and Y. Hence, the computational complexity for find the tight upper bounds of Spearman's footrule under the general missingness case can be reduced to $\mathcal{O}(n \log n + \min\{m_1, m_2\}^2 + m_1 m_2 m_3)$.

Algorithm 8 An efficient algorithm for computing exact upper bounds of Spearman's Footrule under general missingness case.

Input: $X, Y \in \mathcal{V}^n$, where X and Y might be partially observed.

Output: Maximum possible Spearman's footrule distance between X and Y.

Denote m_1 as the number of pairs $(X(i), Y(i))$ such that $Y(i)$ is observed while $X(i)$ is missing. Denote m_2 as the number of pairs $(X(i), Y(i))$ such that $X(i)$ is observed while $Y(i)$ is missing. Denote m_3 as the number of pairs such that both $(X(i), Y(i))$ are missing.

If $m_1 + m_2 = 0$, then run Algorithm 7. However, if $m_3 = 0$, then run Algorithm 6. If $m_1 + m_2 + m_3 = n$, then return $\sum_{i=1}^{n} |i - (n - i + 1)|$.

Rank all observed data in X and Y. Relabel X and Y such that $Y(1), \ldots, Y(m_2)$ are missing, while $X(1), \ldots, X(m_2)$ are observed, and $X(1) < \ldots, X(m_2)$. Let $X(m_2+1), \ldots, X(m_2+m_2)$ be missing, while $Y(m_2+1), \ldots, Y(m_2+m_1)$ are observed and $Y(m_2+1) < \ldots < Y(m_2+m_1)$. Let $(X(n - m_3 + 1), Y(n - m_3 + 1)), \ldots, (X(n), Y(n))$ be missing.

Denote $V = \{1, \ldots, m_2\}$, $U = \{m_2 + 1, \ldots, m_2 + m_1\}$, and $W = \{n - m_3 + 1, \ldots, n\}$.

For $i \in ([n] \setminus W) \setminus (U \cup V)$, let

$$d_i = R(Y(i), (Y(l))_{l \in ([n] \setminus W) \setminus V}) - R(X(i), (X(l))_{l \in ([n] \setminus W) \setminus U}).$$

For any $i \in \{1, \ldots, m_2\}$ and $r_2 \in \{0, \ldots, m_2\}$, let $q_i^{(r_2)} = \mathrm{I}\left(i \leq m_2 - r_2\right)(n - m_3 - i + 1) + \mathrm{I}\left(i > m_2 - r_2\right)(m_2 - i + 1)$.

For any $i \in \{1, \ldots, m_1\}$, let $p_i^{(0)} = \mathrm{I}\left(i \leq m_1\right)(n - m_3 - i + 1) + \mathrm{I}\left(i > m_1\right)(m_1 - i + 1)$.

Run Algorithm 4 computing $s_i = \sum_{l \in ([n] \setminus W) \setminus (U \cup V)} \mathrm{I}\left(d_l \leq i\right)$ for any $i \in \{-m_1 - m_2 - m_3 - 1, \ldots, m_1 + m_2 + m_3 + 1\}$.

Run Algorithm 4 computing

$$R^{(k)} = \sum_{i \in U} \mathrm{I}\left(R(Y(i), (Y(l))_{l \in ([n] \setminus W) \setminus V}) - p_{i-m_2}^{(0)} \leq k\right),$$

for any $k \in \{-m_2 - m_3 - 1, \ldots, m_3 - 1\}$.

For any $r_2 \in \{0, \ldots, m_2\}, r_3 \in \{0, \ldots, m_3\}$, let

$$R^{(0, r_2, r_3)} = R^{(k)}, \text{ if } -r_2 - 1 + 2r_3 - m_3 = k.$$

for $r_2 \in \{0, \ldots, m_2\}$ **do**

    Run Algorithm 4 computing

$$S^{(k)} = \sum_{i \in V} \mathrm{I}\left(q_i^{(r_2)} - R(X(i), (X(l))_{l \in ([n] \setminus W) \setminus U}) \leq k\right),$$

    for any $k \in \{-m_3, \ldots, m_1 + m_3\}$.

    For any $r_1 \in \{0, \ldots, m_1\}, r_3 \in \{0, \ldots, m_3\}$, let

$$S^{(r_1, r_2, r_3)} = S^{(k)}, \text{ if } r_1 + 2r_3 - m_3 = k.$$

end for

15: Initialise

$$D_{0,0,0} = \sum_{i=1}^{m_2} \left|n + i - 1 - R(X(i), (X(l))_{l\in([n]\setminus W)\setminus U})\right| + \sum_{i=1}^{m_1} \left|n - m_3 + i - 1 - R(Y(m_2 + i), (Y(l))_{l\in([n]\setminus W)\setminus U})\right| + \sum_{i=1}^{m_3} |n + 1 - 2i| + \sum_{i\in([n]\setminus W)\setminus(U\cup V)} |d_i + m_3|.$$

16: **for** $r_3 \in \{0, \dots, m_3\}$ **do**
17: **for** $r_2 \in \{0, \dots, m_2\}$ **do**
18: **for** $r_1 \in \{0, \dots, m_1\}$ **do**
19: Denote $c_1 = r_1 + 1 - r_2 + 2r_3 - m_3$, and compute

$$C_1^{(r_1,r_2,r_3)} = -m_2 + n' + \left|c_1 - R(Y(m_1 + m_2 - r_1), (Y(l))_{l\in([n]\setminus W)\setminus V})\right| - \left|c_1 + n - m_3 - m_1 - R(Y(m_1 + m_2 - r_1), (Y(l))_{l\in([n]\setminus W)\setminus V})\right|.$$

20: Compute

$$D_{r_1+1,r_2,r_3} = D_{r_1,r_2,r_3} + 2S^{(r_1,r_2,r_3)} - 2s_{m_3+r_2-r_1-2r_3-1} + C_1^{(r_1,r_2,r_3)}.$$

21: **end for**
22: Let
23: **if** $r_2 < m_2$ **then**
24: Let $r_1 = 0$. Denote $c_2 = r_2 + 1 - r_1 - 2r_3 + m_3$, and compute

$$C_2^{(r_1,r_2,r_3)} = m_1 - n' + \left|c_2 - R(X(m_2 - r), (X(l))_{l\in([n]\setminus W)\setminus U})\right| - \left|c_2 + n - m_3 - m_2 - R(X(m_2 - r), (X(l))_{l\in([n]\setminus W)\setminus U})\right|.$$

25: Compute

$$D_{r_1,r_2+1,r_3} = D_{r_1,r_2,r_3} - 2R^{(r_1,r_2,r_3)} + 2s_{m_3+r_2-r_1-2r_3} + C_2^{(r_1,r_2,r_3)}.$$

26: **end if**
27: **end for**
28: **if** $r_3 < m_3$ **then**
29: Let $r_1 = r_2 = 0$ and compute

$$D_{r_1,r_2,r_3+1} = D_{r_1,r_2,r_3} + |2r_3 + 1 - n| - |n + 1 - 2(m_3 - r_3)| + 2S^{(r_1,r_2,r_3)} + 2S^{(r_1+1,r_2,r_3)} - 2m_2 + 2R^{(r_1,r_2-1,r_3)} + 2R^{(r_1,r_2-2,r_3)} - 2m_1 + 2s_{2r_3+r_1-r_2-m_3} + 2s_{2r_3+r_1-r_2-m_3+1} - 2n'.$$

30: **end if**
31: **end for**
32: Return $\max\{D_{0,0,0}, \dots, D_{m_1,m_2,m_3}\}$.

D. Proofs of results in Chapter 6

This chapter provides proofs for the results used in Chapter 6.

D.1. Bounds of the Wilcoxon-Mann-Whitney statistic with ties

We start by proving Proposition 6.1.3, which extends Theorem 3.1.3 to the case where ties are allowed, and the data are defined in a support Ω.

Proposition D.1.1. Suppose that $X = \{x_1, \ldots, x_n\}$ and $Y = \{y_1, \ldots, y_m\}$ are samples of observations, which need not necessarily be distinct, from a space $\Omega \subset \mathbb{R}$. Suppose that $X' \subset X$ and $Y' \subset Y$ are sub-multisets with sizes n' and m', respectively, which are observed. Denote $Z' = X' \cup Y'$. Then the Wilcoxon-Mann-Whitney statistic $W(X, Y)$ is bounded as follows:

$$\min_{Z \setminus Z'} W(X,Y) = W(X',Y') + \left(\left|\partial Y'_a\right|(n-n') + \left|\partial X'_b\right|(m-m')\right)/2,$$
$$\max_{Z \setminus Z'} W(X,Y) = W(X',Y') + (mn - m'n') - \left(\left|\partial X'_a\right|(m-m') + \left|\partial Y'_b\right|(n-n')\right)/2.$$

where

$$\left|\partial X'_a\right| = \begin{cases} \sum_{x \in X'} \mathrm{I}(x = \min\Omega) & \text{if } \min\Omega \text{ exists,} \\ 0 & \text{otherwise,} \end{cases}$$
$$\left|\partial X'_b\right| = \begin{cases} \sum_{x \in X'} \mathrm{I}(x = \max\Omega) & \text{if } \max\Omega \text{ exists,} \\ 0 & \text{otherwise,} \end{cases}$$
$$\left|\partial Y'_a\right| = \begin{cases} \sum_{y \in Y'} \mathrm{I}(y = \min\Omega) & \text{if } \min\Omega \text{ exists,} \\ 0 & \text{otherwise,} \end{cases}$$
$$\left|\partial Y'_b\right| = \begin{cases} \sum_{y \in Y'} \mathrm{I}(y = \max\Omega) & \text{if } \max\Omega \text{ exists,} \\ 0 & \text{otherwise.} \end{cases}$$

Proof. Note that we only need to prove

$$\min_{Z \setminus Z'} R_{X|Z} = R_{X'|Z'} + (n-n')(n+n'+1)/2 + |\partial Y'_a|(n-n')/2 + |\partial X'_b|(m-m')/2,$$
$$\max_{Z \setminus Z'} R_{X|Z} = R_{X'|Z'} + \{n(n+2m+1) - n'(n'+2m'+1)\}/2 - |\partial X'_a|(m-m')/2 - |\partial Y'_b|(n-n')/2.$$

Since if these are true, then Proposition D.1.1 can be proved following the same approach used in Theorem 3.1.3.

To start, let us denote $X \setminus X' = X^*$ and $Y \setminus Y' = Y^*$. In other words, X^* is a multiset including all unknown elements in X and Y^* is a multiset including all unknown elements in Y.

Then, the rank sum of X in Z can be decomposed:

$$R_{X|Z} = \sum_{x \in X} R(x, Z) = \sum_{x \in X'} R(x, Z) + \sum_{x \in X^*} R(x, Z) = R_{X'|Z} + R_{X^*|Z}. \tag{D.1}$$

We consider $R_{X'|Z}$ and $R_{X^*|Z}$ separately. First, notice that

$$\begin{aligned} R_{X'|Z} &= \sum_{x \in X'} R(x, Z) \\ &= \sum_{x \in X'} \left(\sum_{i=1}^{n} \mathrm{I}\left(z_i < x\right) + \left(\sum_{i=1}^{n} \mathrm{I}\left(z_i = x\right) + 1 \right) / 2 \right) \\ &= \sum_{x \in X'} \sum_{i=1}^{n} \mathrm{I}\left(z_i < x\right) + \sum_{x \in X'} \sum_{i=1}^{n} \mathrm{I}\left(z_i = x\right) / 2 + n'/2. \end{aligned}$$

Since $Z = X \cup Y' \cup Y^*$, we have

$$\sum_{i=1}^{n} \mathrm{I}\left(z_i < x\right) = \sum_{z \in X} \mathrm{I}\left(z < x\right) + \sum_{z \in Y'} \mathrm{I}\left(z < x\right) + \sum_{z \in Y_*} \mathrm{I}\left(z < x\right),$$

and

$$\sum_{i=1}^{n} \mathrm{I}\left(z_i = x\right) = \sum_{z \in X} \mathrm{I}\left(z = x\right) + \sum_{z \in Y'} \mathrm{I}\left(z = x\right) + \sum_{z \in Y_*} \mathrm{I}\left(z = x\right).$$

Therefore, we have

$$\begin{aligned} R_{X'|Z} = n'/2 &+ \underbrace{\sum_{x \in X'} \sum_{z \in X} \mathrm{I}\left(z < x\right)}_{:= C_1} + \underbrace{\sum_{x \in X'} \sum_{z \in Y'} \mathrm{I}\left(z < x\right)}_{:= C_2} + \underbrace{\sum_{x \in X'} \sum_{z \in Y_*} \mathrm{I}\left(z < x\right)}_{:= C_3} \\ &+ \underbrace{\sum_{x \in X'} \sum_{z \in X} \mathrm{I}\left(z = x\right) / 2}_{:= C_4} + \underbrace{\sum_{x \in X'} \sum_{z \in Y'} \mathrm{I}\left(z = x\right) / 2}_{:= C_5} + \underbrace{\sum_{x \in X'} \sum_{z \in Y_*} \mathrm{I}\left(z = x\right) / 2}_{:= C_6}. \end{aligned}$$

Similarly, we can show that

$$\begin{aligned} R_{X_*|Z} = (n - n')/2 &+ \underbrace{\sum_{x \in X_*} \sum_{z \in X} \mathrm{I}\left(z < x\right)}_{:= D_1} + \underbrace{\sum_{x \in X_*} \sum_{z \in Y'} \mathrm{I}\left(z < x\right)}_{:= D_2} + \underbrace{\sum_{x \in X_*} \sum_{z \in Y_*} \mathrm{I}\left(z < x\right)}_{:= D_3} \\ &+ \underbrace{\sum_{x \in X_*} \sum_{z \in X} \mathrm{I}\left(z = x\right) / 2}_{:= D_4} + \underbrace{\sum_{x \in X_*} \sum_{z \in Y'} \mathrm{I}\left(z = x\right) / 2}_{:= D_5} + \underbrace{\sum_{x \in X_*} \sum_{z \in Y_*} \mathrm{I}\left(z = x\right) / 2}_{:= D_6}. \end{aligned}$$

Hence, we have

$$R_{X|Z} = C_1 + C_2 + \cdots + C_6 + D_1 + D_2 + \cdots + D_6 + n/2 \tag{D.2}$$

Notice that

$$\begin{aligned}
&C_1 + C_4 + D_1 + D_4 \\
&= \sum_{x\in X'}\sum_{z\in X} \mathrm{I}\left(z < x\right) + \sum_{x\in X'}\sum_{z\in X} \mathrm{I}\left(z = x\right)/2 + \sum_{x\in X_*}\sum_{z\in X} \mathrm{I}\left(z < x\right) + \sum_{x\in X_*}\sum_{z\in X} \mathrm{I}\left(z = x\right)/2 \\
&= \sum_{x\in X}\sum_{z\in X} \mathrm{I}\left(z < x\right) + \sum_{x\in X}\sum_{z\in X} \mathrm{I}\left(z = x\right)/2 \\
&= \sum_{x\in X}\sum_{z\in X} \mathrm{I}\left(z \le x\right) - n/2 \\
&= (1+n)n/2 - n/2 = n^2/2.
\end{aligned}$$

Also, we have

$$\begin{aligned}
R_{X'|Z'} &= \sum_{x\in X'} R(x, Z') = \sum_{x\in X'} \left(\sum_{z\in Z'} \mathrm{I}\left(z < x\right) + \left(\sum_{z\in Z'} \mathrm{I}\left(z = x\right) + 1 \right)/2 \right) \\
&= \sum_{x\in X'}\sum_{z\in Z'} \mathrm{I}\left(z < x\right) + \sum_{x\in X'}\sum_{z\in Z'} \mathrm{I}\left(z = x\right)/2 + n'/2 \\
&= \sum_{x\in X'}\sum_{z\in X'} \mathrm{I}\left(z < x\right) + \sum_{x\in X'}\sum_{z\in Y'} \mathrm{I}\left(z < x\right) + \sum_{x\in X'}\sum_{z\in X'} \mathrm{I}\left(z = x\right)/2 + \sum_{x\in X'}\sum_{z\in Y'} \mathrm{I}\left(z = x\right)/2 + n'/2 \\
&= C_2 + C_5 + \sum_{x\in X'}\sum_{z\in X'} \mathrm{I}\left(z < x\right) + \sum_{x\in X'}\sum_{z\in X'} \mathrm{I}\left(z = x\right)/2 + n'/2 \\
&= C_2 + C_5 + \sum_{x\in X'} \left(\sum_{z\in X'} \mathrm{I}\left(z < x\right) + \left(\sum_{z\in X'} \mathrm{I}\left(z = x\right) + 1 \right)/2 \right) \\
&= C_2 + C_5 + \sum_{x\in X'} R(x, X') = (1+n')n'/2 + C_2 + C_5.
\end{aligned}$$

Thus,

$$C_2 + C_5 = R_{X'|Z'} - (1+n')n'/2.$$

Now, put the above results back into Equation (D.2),

$$\begin{aligned}
R_{X|Z} &= R_{X'|Z'} - (1+n')n'/2 + n^2/2 + n/2 + C_3 + C_6 + D_2 + D_3 + D_5 + D_6 \\
&= R_{X'|Z'} + (n-n')(n+n'+1)/2 + C_3 + C_6 + D_2 + D_3 + D_5 + D_6
\end{aligned} \tag{D.3}$$

Below we consider $C_3 + C_6$, $D_2 + D_5$, $D_3 + D_6$ separately. Notice that

$$C_3 + C_6 = \sum_{x\in X'}\sum_{z\in Y_*} \mathrm{I}\left(z < x\right) + \sum_{x\in X'}\sum_{z\in Y_*} \mathrm{I}\left(z = x\right)/2$$

Suppose $\max \Omega$ does not exist. In other words, $z \in Y_*$ can be arbitrarily large. Then we have

$$C_3 + C_6 \ge 0,$$

where the "=" holds when $z > \max X'$ for any $z \in Y_*$. However, suppose $\max \Omega$ exist. Denote

$|\partial X_b'| = \sum_{x \in X'} \mathrm{I}\,(x = \max \Omega)$. Then we have

$$C_3 + C_6 \geq |\partial X_b'|(m - m')/2,$$

where the "=" holds when $\min Y^* = \max Y^* = \max \Omega$, i.e. all $z \in Y^*$ equals to $\max \Omega$. Overall, we have

$$C_3 + C_6 \geq |\partial X_b'|(m - m')/2,$$

where

$$\left|\partial X_b'\right| = \begin{cases} \sum\limits_{x \in X'} \mathrm{I}\,(x = \max \Omega) & \text{if } \max \Omega \text{ exists,} \\ 0 & \text{otherwise.} \end{cases}$$

Similarly, for $D_2 + D_5$, we have

$$D_2 + D_5 = \sum_{x \in X_*} \sum_{z \in Y'} \mathrm{I}\,(z < x) + \sum_{x \in X_*} \sum_{z \in Y'} \mathrm{I}\,(z = x)\,/2.$$

Suppose $\min \Omega$ does not exist. In other words, $z \in X_*$ can be arbitrarily small. Then

$$D_2 + D_5 \geq 0,$$

where the " = " holds when $x < \min Y'$ for any $x \in X_*$. However, suppose $\min \Omega$ exists. Denote $|\partial Y_a'| = \sum_{y \in Y'} \mathrm{I}\,(y = \min \Omega)$. Then we have

$$D_2 + D_5 \geq |\partial Y_a'|(n - n')/2,$$

where the "=" holds when all x in X equals $\min \Omega$, i.e.

$$\min X^* = \max X^* = \min \Omega.$$

Overall, we have

$$D_2 + D_5 \geq |\partial Y_a'|(n - n')/2,$$

where

$$\left|\partial Y_a'\right| = \begin{cases} \sum\limits_{y \in Y'} \mathrm{I}\,(y = \min \Omega) & \text{if } \min \Omega \text{ exists,} \\ 0 & \text{otherwise,} \end{cases}$$

For $D_3 + D_6$, we have

$$D_3 + D_6 = \sum_{x \in X_*} \sum_{z \in Y_*} \mathrm{I}\,(z < x) + \sum_{x \in X_*} \sum_{z \in Y_*} \mathrm{I}\,(z = x)\,/2.$$

Then,

$$D_3 + D_6 \geq 0$$

where the "= " holds when $x < z$ for any $x \in X_*$ and $z \in Y_*$.

Put the above results for $C_3 + C_6$, $D_2 + D_5$, $D_3 + D_6$ back into Equation D.3, we have

$$R_{X|Z} \geq R_{X'|Z'} + (n - n')(n + n' + 1)/2 + |\partial X'_b|(m - m')/2 + |\partial Y'_a|(n - n')/2,$$

where the "= " holds when the missing values in $x \in X_*$ such that $x = \min \Omega$ if $\min \Omega$ exists, otherwise $x < \min Y'$; the missing values in $y \in Y_*$ such that $y = \max \Omega$ if $\max \Omega$ exists, otherwise $y > \max X'$. Hence, we conclude the result for the lower bound.

For the upper bound, note that we have

$$\min_{Z \setminus Z'} R_{Y|Z} = R_{Y'|Z'} + \underbrace{(m - m')(m + m' + 1)/2}_{:=E_1} + |\partial X'_a|(m - m')/2 + |\partial Y'_b|(n - n')/2.$$

Since

$$R_{Z|Z} = R_{X|Z} + R_{Y|Z} = \underbrace{(1 + n + m)(n + m)/2}_{:=E_2}$$

is a constant, therefore when $R_{Y|Z}$ takes its minimum among all possible Z^*, $R_{X|Z}$ takes its maximum, i.e.

$$\begin{aligned}
&\max_{Z \setminus Z'} R_{X|Z} = R_{Z|Z} - \min_{Z \setminus Z'} R_{Y|Z} \\
&= E_2 - R_{Y'|Z'} - E_1 - |\partial X'_a|(m - m')/2 - |\partial Y'_b|(n - n')/2.
\end{aligned}$$

Also, notice that

$$\begin{aligned}
R_{Y'|Z'} + R_{X'|Z'} &= \underbrace{(1 + n' + m')(n' + m')/2}_{:=E_3} \\
\implies R_{Y'|Z'} &= E_3 - R_{X'|Z'}
\end{aligned}$$

Hence, we have

$$\begin{aligned}
\max_{Z \setminus Z'} R_{X|Z} &= E_2 - E_3 + R_{X'|Z'} - E_1 - |\partial X'_a|(m - m')/2 - |\partial Y'_b|(n - n')/2 \\
&= R_{X'|Z'} + (E_2 - E_3 - E_1) - |\partial X'_a|(m - m')/2 - |\partial Y'_b|(n - n')/2.
\end{aligned}$$

Recall in Equation A.5, we have shown that

$$E_2 - E_3 - E_1 = \{n(n + 2m + 1) - n'(n' + 2m' + 1)\}/2.$$

Hence, we conclude

$$\begin{aligned}&\max_{Z\setminus Z'} R_{X|Z}\\&= R_{X'|Z'} + (\{n(n+2m+1) - n'(n'+2m'+1)\}/2) - |\partial X'_a|(m-m')/2 - |\partial Y'_b|(n-n')/2,\end{aligned}$$

which completes our proof. □

D.2. Issue of controlling p-value with ties

We now prove Proposition 6.1.6, which shows that the bounds of the p-value constructed according to Proposition 3.2.4 might not always control the p-value when data are tied.

Proposition D.2.1. Suppose that $X = \{x_1, \ldots, x_n\}$ and $Y = \{y_1, \ldots, y_m\}$ are samples of real-valued observations, which need not necessarily be distinct. Suppose that $X' \subset X$ and $Y' \subset Y$ are sub-multisets with sizes n' and m', respectively, which are known. Define $Z = X \cup Y$ and $Z' = X' \cup Y'$ and suppose that $Z \setminus Z'$ is unknown. Suppose X_1, Y_1 and X_2, Y_2 are such that $X' \subset X_i$ and $Y' \subset Y_i$ for $i = 1, 2$ and

$$W(X_1, Y_1) = \min_{Z\setminus Z'} W(X, Y), \qquad W(X_2, Y_2) = \max_{Z\setminus Z'} W(X, Y).$$

Then, among candidates for X, Y, it is possible X_3, Y_3 exist such that $X' \subset X_3$ and $Y' \subset Y_3$ and

$$W(X_1, Y_1) \leq W(X_3, Y_3) \leq W(X_2, Y_2),$$

but moreover, after defining $\mu = nm/2$, the standardised statistics are such that

$$(W(X_3, Y_3) - \mu)/\sigma(X_3, Y_3) \notin [(W(X_1, Y_1) - \mu)/\sigma(X_1, Y_1), (W(X_2, Y_2) - \mu)/\sigma(X_2, Y_2)],$$

and consequently $p_3 \notin [p_{\min}(X, Y), p_{\max}(X, Y)]$, where $p_{\min}(X, Y)$, and $p_{\max}(X, Y)$ are defined according to Proposition 3.2.4 and p_3 is the p-values for $W(X_3, Y_3)$.

Proof. We provide an example. Let us consider the case $X = \{1, 2, 3, 2, 2, 1, 1\}$, $Y = \{3, 3, 3, 3, 3, 3, y_7\}$ and the support is $\{1, 2, 3, 4\}$, i.e. y_7 is a real number taking from $\{1, 2, 3, 4\}$. Suppose $X' = X$, $Y' = Y \setminus \{y_7\}$, and denote $Z' = X' \cup Y'$. According to the definition of ranks with ties, we can verify that $R(1, Z') = 2, R(2, Z') = 5$ and $R(3, Z') = 10$. Subsequently,

$$\begin{aligned}W(X', Y') &= R_{X'|Z'} - n'(n'+1)/2\\&= 3R(1, Z') + 3R(2, Z') + R(3, Z') - 7 \cdot 8/2\\&= 6 + 15 + 10 - 28 = 3.\end{aligned}$$

Now, we are going to obtain the minimum and maximum possible $W(X, Y)$ using the result of Proposition 6.1.3. Since the support $\Omega = \{1, 2, 3, 4\}$, we have $|\partial X'_a| = 3, |\partial Y'_a| = 0, |\partial X'_b| =$

$0, |\partial Y_b'| = 0$. Also, we have $n - n' = 0$ and $m - m' = 1$. Therefore, according to Proposition 6.1.3

$$\min_{Z \setminus Z'} W(X, Y) = W(X', Y') + 0 = 3,$$
$$\max_{Z \setminus Z'} W(X, Y) = W(X', Y') + (49 - 42) - 3/2 = 8.5.$$

Let us denote $X_1 = X$, $Y_1 = Y' \cup \{4\}$ and $X_2 = X$, $Y_2 = Y' \cup \{1\}$. That is, y_7 takes 1 and 4 for Y_1, and Y_2, separately. Following the above procedures as we obtain $W(X', Y')$, we can verify that

$$W(X_1, Y_1) = \min_{Z \setminus Z'} W(X, Y) = 3, \ W(X_2, Y_2) = \max_{Z \setminus Z'} W(X, Y) = 8.5.$$

Recall that the variance of $W(X, Y)$ with ties is defined as

$$\sigma^2(X, Y) = nm(n + m + 1)/12 - nm\{12(n + m)(n + m - 1)\}^{-1} \sum_{i=1}^{3} (d_i^3 - d_i),$$

and $nm = 49, n + m + 1 = 15, n + m - 1 = 13$. Hence, we obtain

$$\begin{aligned}
\sigma^2(X_1, Y_1) &= 49 \cdot 15/12 - 49/(12 \cdot 14 \cdot 13)[(3^3 - 3) + (3^3 - 3) + (7^3 - 7)], \\
&= 735/12 - 49/2184 \cdot 384 \\
&= (133770 - 18816)/2184 \\
&= 114954/2184,
\end{aligned}$$

and

$$\begin{aligned}
\sigma^2(X_2, Y_2) &= 49 \cdot 15/12 - \frac{49}{12 \cdot 14 \cdot 13}[(4^3 - 4) + (3^3 - 3) + (7^3 - 7)] \\
&= 735/12 - 49/2184 \cdot 420 \\
&= (133770 - 20580)/2184 \\
&= 113190/2184,
\end{aligned}$$

respectively. For both cases, we have $\mu = nm/2 = 49/2$. Thus, it further follows

$$\begin{aligned}
(W(X_1, Y_1) - \mu)/\sigma(X_1, Y_1) &= (3 - 49/2)/(114954/2184)^{1/2} \\
&= -43/\{2(114954/2184)^{1/2}\} \\
(W(X_2, Y_2) - \mu)/\sigma(X_2, Y_2) &= (8.5 - 49/2)/(735/12 - 49/2184 \cdot 384)^{1/2} \\
&= -16/(113190/2184)^{1/2}.
\end{aligned}$$

Define $X_3 = X$, $Y_3 = Y' \cup \{3\}$. We can then verify that $W(X_3, Y_3) = 3.5$, which implies

$$\min_{Z \setminus Z'} W(X, Y) = W(X_1, Y_1) < W(X_3, Y_3) < W(X_2, Y_2) = \max_{Z \setminus Z'} W(X, Y).$$

For the variance, we have

$$\begin{aligned}\sigma^2(X_3, Y_3) &= 49 \cdot 15/12 - 49/(12 \cdot 14 \cdot 13)[(3^3 - 3) + (3^3 - 3) + (8^3 - 8)] \\ &= 735/12 - 49/2184 \cdot 552 \\ &= (133770 - 27084)/2184 \\ &= 106686/2184.\end{aligned}$$

Thus,

$$\begin{aligned}(W(X_3, Y_3) - \mu)/\sigma(X_3, Y_3) &= (3.5 - 49/2)/(735/12 - 49/2184 \cdot 552)^{1/2} \\ &= -21/(106686/2184)^{1/2}.\end{aligned}$$

Notice that

$$\begin{aligned}\{(W(X_1, Y_1) - \mu)/\sigma(X_1, Y_1)\}^2 &= 43^2/\{4(114954/2184)\} \\ &= 43 \cdot 43 \cdot 2184/(4 \cdot 114954) \\ &= 1009554/114954, \\ \{(W(X_2, Y_2) - \mu)/\sigma(X_2, Y_2)\}^2 &= 16^2/(113190/2184) \\ &= 16 \cdot 16 \cdot 2184/113190 \\ &= 559104/113190 \\ \{(W(X_3, Y_3) - \mu)/\sigma(X_3, Y_3)\}^2 &= 21^2/(106686/2184) \\ &= 21 \cdot 21 \cdot 2184/106686 \\ &= 963144/106686.\end{aligned}$$

Now, we compare the above three numbers using the concept of cross-multiplication. Since

$$963144 \cdot 114954 = 110717255376 > 106686 \cdot 114954 = 12263982444,$$

we have

$$\{(W(X_3, Y_3) - \mu)/\sigma(X_3, Y_3)\}^2 > \{(W(X_1, Y_1) - \mu)/\sigma(X_1, Y_1)\}^2.$$

Similarly, since

$$1009554 \cdot 113190 = 114271417260 > 559104 \cdot 114954 = 64271241216,$$

we have

$$\{(W(X_1, Y_1) - \mu)/\sigma(X_1, Y_1)\}^2 > \{(W(X_2, Y_2) - \mu)/\sigma(X_2, Y_2)\}^2.$$

Thus, it follows

$$\begin{aligned}\{(W(X_3,Y_3)-\mu)/\sigma(X_3,Y_3)\}^2 &> \{(W(X_1,Y_1)-\mu)/\sigma(X_1,Y_1)\}^2 \\ &> \{(W(X_2,Y_2)-\mu)/\sigma(X_2,Y_2)\}^2.\end{aligned}$$

Note that $(W(X_3,Y_3)-\mu)/\sigma(X_3,Y_3)$, $(W(X_1,Y_1)-\mu)/\sigma(X_1,Y_1)$ and $(W(X_2,Y_2)-\mu)/\sigma(X_2,Y_2)$ are all negative real numbers. Hence,

$$\begin{aligned}(W(X_3,Y_3)-\mu)/\sigma(X_3,Y_3) &< (W(X_1,Y_1)-\mu)/\sigma(X_1,Y_1) \\ &< (W(X_2,Y_2)-\mu)/\sigma(X_2,Y_2).\end{aligned}$$

Therefore,

$$(W(X_3,Y_3)-\mu)/\sigma(X_3,Y_3) \notin \left[(W(X_1,Y_1)-\mu)/\sigma(X_1,Y_1), (W(X_2,Y_2)-\mu)/\sigma(X_2,Y_2)\right].$$

Further, since $(W(X_3,Y_3)-\mu)/\sigma(X_3,Y_3)$, $(W(X_1,Y_1)-\mu)/\sigma(X_1,Y_1)$ and $(W(X_2,Y_2)-\mu)/\sigma(X_2,Y_2)$ all follow standard normal distribution and the cumulative distribution function Φ is monotonic, we have

$$\begin{aligned}2\Phi((W(X_3,Y_3)-\mu)/\sigma(X_3,Y_3)) &\leq 2\Phi((W(X_1,Y_1)-\mu)/\sigma(X_1,Y_1)) \\ &\leq 2\Phi((W(X_2,Y_2)-\mu)/\sigma(X_2,Y_2)).\end{aligned}$$

That is, $p_3 < p_1 < p_2$, which follows $p_3 < \min\{p_1,p_2\} = p_{\min}(X,Y)$. Thus, we conclude our result. □

D.3. Bounds of variance

We now bound the variance of the Wilcoxon-Mann-Whitney test statistic when the data can be tied.

Before proving the result Proposition D.3.2 in the main, we introduce the following Lemma.

Lemma D.3.1. Let $f(x) = x^3 - x$, where $x \in \mathbb{R}$. Suppose $a_1,\ldots,a_k$ and $b_1,\ldots,b_k$ are both non-negative integer sequence such that $a_i \leq b_i$ for any $i = 1,\ldots,k$. Denote

$$a_{\max} = \max\{a_1,\ldots,a_k\}$$

and define z such as

$$z := a_{\max} + \sum_{i=1}^{k}(b_i - a_i). \tag{D.4}$$

Then, we have

$$f(z) - f(a_{\max}) \geq \sum_{i=1}^{k}\left(f(b_i) - f(a_i)\right). \tag{D.5}$$

Proof. To start, we notice that the equation (D.4) can be written as

$$z = a_{\max} + b_j - a_j + \sum_{i=1, i\neq j}^{k} (b_i - a_i),$$

where j is any number in $\{1, \ldots, k\}$. Since $b_i \geq a_i$ for any $i = 1, \ldots, k$, for any given $j \in \{1, \ldots, k\}$, we have

$$\sum_{i=1, i\neq j}^{k} (b_i - a_i) \geq 0,$$

which gives us

$$z \geq a_{\max} + b_j - a_j.$$

Recall that $a_{\max}$ is defined as the maximum number in $\{a_1, \ldots, a_k\}$. This follows

$$z \geq b_j. \tag{D.6}$$

Subsequently, according to the definition of $f(x)$, we have

$$\begin{aligned}\sum_{i=1}^{k} (f(b_i) - f(a_i)) &= \sum_{i=1}^{k} (b_i^3 - b_i - a_i^3 + a_i) \\ &= \sum_{i=1}^{k} (b_i^3 - a_i^3) - \sum_{i=1}^{k} (b_i - a_i).\end{aligned}$$

Using the definition of $a_{\max}$ in equation (D.4), we further have

$$\begin{aligned}\sum_{i=1}^{k} (f(b_i) - f(a_i)) &= \sum_{i=1}^{k} (b_i^3 - a_i^3) - (z - a_{\max}) \\ &= \sum_{i=1}^{k} \{(b_i - a_i)(b_i^2 + a_i b_i + a_i^2)\} - (z - a_{\max}).\end{aligned}$$

Subsequently, the inequality (D.6) gives us

$$\sum_{i=1}^{k} (f(b_i) - f(a_i)) \sum_{i=1}^{k} (b_i - a_i)(z^2 + z a_{\max} + a_{\max}^2) - (z - a_{\max}).$$

Using the definition of $a_{\max}$ in equation (D.4) again, it follows

$$\begin{aligned}\sum_{i=1}^{k} (f(b_i) - f(a_i)) &= (z - a_{\max})(z^2 + z a_{\max} + a_{\max}^2) - (z - a_{\max}) \\ &= z^3 - a_{\max}^3 - (z - a_{\max}) \\ &= z^3 - z - a_{\max}^3 + a_{\max} \\ &= f(z) - f(a_{\max})\end{aligned}$$

Thus, we conclude our result. □

Applying Lemma D.3.1, we can now bound the variance using the following result.

Proposition D.3.2. Suppose that $X = \{x_1, \ldots, x_n\}$ and $Y = \{y_1, \ldots, y_m\}$ are samples of observations, which need not necessarily be distinct, from a space $\Omega \in \mathbb{R}$. Suppose that X' and Y' are sub-multisets with sizes n' and m', respectively, which are known and that $X' \cup Y'$ contains e' distinct values with multiplicities $d'_1, \ldots, d'_{e'}$ and $d'_1 \leq \ldots \leq d'_{e'}$. Define

$$\sigma^2_{\max}(X,Y) = mn(n+m+1)/12 - mn\{12(n+m)(n+m-1)\}^{-1} \sum_{i=1}^{e'} \{(d'_i)^3 - d'_i\}$$
$$\sigma^2_{\min}(X,Y) = \sigma^2_{\max}(X,Y) - mn\{12(n+m)(n+m-1)\}^{-1}\{d^3_{\max} - d_{\max} - (d'_{e'})^3 + d'_{e'}\},$$

where $d_{\max} = d'_{e'} + n + m - n' - m'$, then $\sigma^2(X,Y)$, the variance of Wilcoxon-Mann-Whitney statistic $W(X,Y)$, is bounded by $\sigma^2_{\min}(X,Y) \leq \sigma^2(X,Y) \leq \sigma^2_{\max}(X,Y)$.

Proof. To start, let us assume that $X \cup Y$ take e distinct values from $\{s_1, s_2, \cdots, s_e\}$ and the number of values in $X \cup Y$ equals to s_i is d_i. Without loss of generality, assume that $X' \cup Y'$ take values from $\{s_1, s_2, \cdots, s_{e'}\}$ such that the numbers of values in $X' \cup Y'$ equals to s_i is d'_i and $d'_1 \leq \ldots \leq d'_{e'}$. Define $d'_{e'+1}, \ldots, d'_e = 0$.

We start by proving $\sigma^2_{\max}(X,Y) \geq \sigma^2(X,Y)$. Notice that

$$\sigma^2(X,Y) = mn(n+m+1)/12 - mn\{12(n+m)(n+m-1)\}^{-1} \sum_{i=1}^{e} \{(d_i)^3 - d_i\},$$

which follows

$$\begin{aligned}
\sigma^2_{\max}(X,Y) - \sigma^2(X,Y) &= mn\{12(n+m)(n+m-1)\}^{-1} \sum_{i=1}^{e} \{(d_i)^3 - d_i\} \\
&\quad - mn\{12(n+m)(n+m-1)\}^{-1} \sum_{i=1}^{e'} \{(d'_i)^3 - d'_i\} \\
&= mn\{12(n+m)(n+m-1)\}^{-1} \sum_{i=1}^{e} [\{(d_i)^3 - d_i\} - \{(d'_i)^3 - d'_i\}].
\end{aligned}$$

Define

$$f(x) = x^3 - x, \text{ where } x \in \mathbb{R}.$$

We then have

$$\sigma^2_{\max}(X,Y) - \sigma^2(X,Y) = mn\{12(n+m)(n+m-1)\}^{-1} \sum_{i=1}^{e} (f(d_i) - f(d'_i)).$$

Since d_i is defined as the number of values in $X \cup Y$ taking s_i while d'_i is defined as the number of values in $X' \cup Y'$ taking s_i, we have

$$d'_i \leq d_i, i = 1, 2, \cdots, e.$$

Notice that $f(0) = f(1) < f(2) < \ldots$, which implies $f(d_i) - f(d'_i) \geq 0$ for any $i = 1, \ldots, e$. It follows

$$\sigma^2_{\max}(X, Y) - \sigma^2(X, Y) \geq 0.$$

Thus, we have proved $\sigma^2_{\max}(X, Y) \geq \sigma^2(X, Y)$.

We now prove $\sigma^2_{\min}(X, Y) \leq \sigma^2(X, Y)$. To start, notice that

$$\begin{aligned}
\sigma^2(X, Y) - \sigma^2_{\min} &= mn\{12(n+m)(n+m-1)\}^{-1} \sum_{i=1}^{e'} \{(d'_i)^3 - d'_i\} \\
&\quad + mn\{12(n+m)(n+m-1)\}^{-1} \{d^3_{\max} - d_{\max} - (d'_{e'})^3 + d'_{e'}\} \\
&\quad - mn\{12(n+m)(n+m-1)\}^{-1} \sum_{i=1}^{e} \{(d_i)^3 - d_i\} \\
&= mn\{12(n+m)(n+m-1)\}^{-1} \left[\sum_{i=1}^{e'} f(d'_i) + f(d_{\max}) - f(d'_{e'}) - \sum_{i=1}^{e} f(d_i) \right].
\end{aligned}$$

Since we define $d'_{e'+1}, \ldots, d'_e = 0$ and $f(0) = 0$, we have $\sum_{i=1}^{e'} f(d'_i) = \sum_{i=1}^{e} f(d'_i)$. Thus,

$$\sigma^2(X, Y) - \sigma^2_{\min} = mn\{12(n+m)(n+m-1)\}^{-1} \left\{ f(d_{\max}) - f(d'_{e'}) + \sum_{i=1}^{e} (f(d'_i) - f(d_i)) \right\}$$

From the above Lemma S1, we have

$$f(d_{\max}) - f(d'_{e'}) \geq \sum_{i=1}^{e} (f(d_i) - f(d'_i)).$$

To see this, consider $d'_1, d'_2, \cdots d'_e$ to be $a_1, a_2, \cdots, a_k$ in Lemma 1, $d_1, d_2, \cdots d_e$ to be $b_1, b_2, \cdots, b_k$. Then, we have $a_i \leq b_i$ for any $i = 1, \ldots, k$. Recall that $d'_{e'}$ is defined as the maximum over all $d'_1, d'_2, \cdots d'_e$ and $d_{\max} = d'_{e'} + n + m - n' - m' = d'_{e'} + \sum_{i=1}^{e} (d_i - d'_i)$.

Thus, we have shown

$$\sigma^2(X, Y) - \sigma^2_{\min} \geq 0,$$

which completes our proof.

□

D.4. Bounds of the p-value

We now prove the bounds of the p-value, using the results of Proposition 6.1.3 and Proposition 6.1.8.

Corollary D.4.1. Suppose that $X = \{x_1, \ldots, x_n\}$ and $Y = \{y_1, \ldots, y_m\}$ are samples of observations, which need not necessarily be distinct, from a space $\Omega \subset \mathbb{R}$. Suppose that $X' \subset X$ and $Y' \subset Y$ are sub-multisets with sizes n' and m', respectively, which are known, and compute $|\partial X'_a|$, $|\partial X'_b|$, $|\partial Y'_a|$ and $|\partial Y'_b|$ from Proposition 6.1.3 and $\sigma^2_{\max}(X, Y)$ from Proposition 6.1.8, and

define

$$
\begin{aligned}
&W_{\min}(X,Y) = W(X',Y') + \left|\partial Y'_a\right|(n-n')/2 + \left|\partial X'_b\right|(m-m')/2, \quad \mu = nm/2,\\
&W_{\max}(X,Y) = W(X',Y') + (mn - m'n') - \left|\partial X'_a\right|(m-m') + \left|\partial Y'_b\right|(n-n'),\\
&p_1 = 2\Phi(-\left|W_{\min}(X,Y)-\mu\right|/\sigma_{\min}(X,Y)),\ p_2 = 2\Phi(-\left|W_{\max}(X,Y)-\mu\right|/\sigma_{\max}(X,Y)),\\
&p_3 = 2\Phi(-\left|W_{\max}(X,Y)-\mu\right|/\sigma_{\min}(X,Y)),\ p_4 = 2\Phi(-\left|W_{\min}(X,Y)-\mu\right|/\sigma_{\max}(X,Y)).
\end{aligned}
$$

Then, the the p-value of $W(X,Y)$ is bounded as follows

$$
\begin{cases}
p_1 \le p \le p_2, & \text{if } W_{\min}(X,Y)-\mu < 0 \text{ and } W_{\max}(X,Y)-\mu < 0,\\
\min\{p_1,p_3\} \le p \le 1, & \text{if } W_{\min}(X,Y)-\mu < 0 \text{ and } W_{\max}(X,Y)-\mu \ge 0,\\
p_3 \le p \le p_4, & \text{if } W_{\min}(X,Y)-\mu \ge 0 \text{ and } W_{\max}(X,Y)-\mu \ge 0.
\end{cases} \tag{D.7}
$$

Proof. Recall that $W(X,Y)$ is bounded as such

$$
W_{\min}(X,Y) \le W(X,Y) \le W_{\max}(X,Y), \tag{D.8}
$$

where $W_{\min}(X,Y)$ and $W_{\max}(X,Y)$ are defined in Proposition 6.1.3 and $\sigma(X,Y)$ is bounded as such

$$
\sigma_{\min}(X,Y) \le \sigma(X,Y) \le \sigma_{\max}(X,Y), \tag{D.9}
$$

where $\sigma_{\min}$ and $\sigma_{\max}$ are defined in Proposition 6.1.8.

Suppose now $W_{\min}(X,Y)-\mu < 0$ and $W_{\max}(X,Y)-\mu < 0$. Then according to Inequalities (D.8), we also have $W(X,Y) < 0$. Then,

$$
\sigma_{\max}(X,Y) \ge \sigma(X,Y) \implies \sigma_{\max}(X,Y)(W(X,Y)-\mu) \le \sigma(X,Y)(W(X,Y)-\mu).
$$

Notice that

$$
\sigma(X,Y)(W(X,Y)-\mu) \le \sigma(X,Y)(W_{\max}(X,Y)-\mu).
$$

Then we have

$$
\begin{aligned}
&\sigma_{\max}(X,Y)(W(X,Y)-\mu) \le \sigma(X,Y)(W_{\max}(X,Y)-\mu)\\
\implies &\frac{W(X,Y)-\mu}{\sigma(X,Y)} \le \frac{W_{\max}(X,Y)-\mu}{\sigma_{\max}(X,Y)}.
\end{aligned} \tag{D.10}
$$

Similarly, still assume that $W_{\min}(X,Y)-\mu < 0$, and $W_{\max}(X,Y)-\mu < 0$, we have

$$
\sigma(X,Y) \ge \sigma_{\min}(X,Y) \implies \sigma(X,Y)(W(X,Y)-\mu) \le \sigma_{\min}(X,Y)(W(X,Y)-\mu).
$$

Notice that

$$
\sigma(X,Y)(W_{\min}(X,Y)-\mu) \le \sigma(X,Y)(W(X,Y)-\mu).
$$

Hence, we have

$$\begin{aligned} &\sigma(X,Y)(W_{\min}(X,Y)-\mu) \le \sigma_{\min}(X,Y)(W(X,Y)-\mu) \\ \Longrightarrow &\frac{W_{\min}(X,Y)-\mu}{\sigma_{\min}(X,Y)} \le \frac{W(X,Y)-\mu}{\sigma(X,Y)} \end{aligned}$$

Combining the above result with Inequality (D.10), we have

$$\begin{aligned} &\frac{W_{\min}(X,Y)-\mu}{\sigma_{\min}(X,Y)} \le \frac{W(X,Y)-\mu}{\sigma(X,Y)} \le \frac{W_{\max}(X,Y)-\mu}{\sigma_{\max}(X,Y)} \\ \Longrightarrow &p_1 \le p \le p_2. \end{aligned}$$

Hence, we prove our results when $W_{\min}(X,Y)-\mu<0$, and $W_{\max}(X,Y)-\mu<0$.

Suppose now $W_{\min}(X,Y)-\mu<0$ while $W_{\max}(X,Y)-\mu \ge 0$. Then, if $W(X,Y)-\mu<0$, we have

$$\sigma_{\min}(X,Y)(W(X,Y)-\mu) \ge \sigma(X,Y)(W(X,Y)-\mu) \ge \sigma(X,Y)(W_{\min}(X,Y)-\mu).$$

Hence,

$$\begin{aligned} &\frac{W_{\min}(X,Y)-\mu}{\sigma_{\min}} \le \frac{W(X,Y)-\mu}{\sigma(X,Y)}, \\ \Longrightarrow &p \ge p_1. \end{aligned}$$

Similarly, if $W(X,Y)-\mu \ge 0$, then

$$\sigma_{\min}(X,Y)(W(X,Y)-\mu) \le \sigma_{\min}(X,Y)(W_{\max}(X,Y)-\mu) \le \sigma(X,Y)(W_{\max}(X,Y)-\mu).$$

Hence,

$$\begin{aligned} &\frac{W(X,Y)-\mu}{\sigma(X,Y)} \le \frac{W_{\max}(X,Y)-\mu}{\sigma_{\min}} \\ \Longrightarrow &p \ge p_3. \end{aligned}$$

Therefore, we have $p \ge \min\{p_1, p_3\}$ whether $W(X,Y)-\mu<0$ or $W(X,Y)-\mu \ge 0$ holds. By definition of p-values, we have $p \le 1$. Hence, we conclude

$$\min\{p_1, p_3\} \le p \le 1,$$

which proves our results when $W_{\min}(X,Y)-\mu<0$ and $W_{\max}(X,Y)-\mu \ge 0$.

Finally, suppose now $W_{\min}(X,Y)-\mu \ge 0$ and $W_{\max}(X,Y)-\mu \ge 0$. Then, we have $W(X,Y) \ge 0$, which follows

$$\begin{aligned} &\frac{W_{\min}(X,Y)-\mu}{\sigma_{\max}(X,Y)} \le \frac{W(X,Y)-\mu}{\sigma(X,Y)} \le \frac{W_{\max}(X,Y)-\mu}{\sigma_{\min}(X,Y)} \\ \Longrightarrow &p_3 \le p \le p_4. \end{aligned}$$

This proves our results when $W_{\min}(X,Y)-\mu \geq 0$ and $W_{\max}(X,Y)-\mu \geq 0$ and completes our proof. □

D.5. Bounds of the Wilcoxon-Mann-Whitney statistic as a random variable

Proposition D.5.1. Let $X=\{X_1,\ldots,X_n\}$ and $Y=\{Y_1,\ldots,Y_m\}$ be continuous random variables. Suppose $X'=\{X'_1,\ldots,X'_{n'}\}$ is any non-empty proper subset of X. Let us order $X_1,\ldots,X_n$ from small to large as $X_{(1)},\ldots,X_{(n)}$ and denote

$$X_*=X\setminus\{X_{(n'+1)},\ldots,X_{(n)}\},\ X_*=X\setminus\{X_{(1)},\ldots,X_{(n-n')}\}.$$

Subsequently, we have

$$W(X_*,Y)\leq W(X',Y)\leq W(X_*,Y).$$

Proof. To start, denote $X\setminus X'=\tilde{X}=\{\tilde{X}_1,\ldots,\tilde{X}_{n-n'}\}$. Let us order random variables in $\tilde{X}$ from small to large as $\tilde{X}_{(1)},\ldots,\tilde{X}_{(n-n')}$. Since $\tilde{X}\subset X$, for any $i\in\{1,\ldots,n-n'\}$,

$$\tilde{X}_{(i)}\geq X_{(i)},\ \text{and}\ \tilde{X}_{(i)}\leq X_{(i+n')}.$$

Subsequently,

$$\sum_{i=n'+1}^{n}\sum_{j=1}^{m}\mathrm{I}\left(X_{(i)}\geq Y_j\right)\geq\sum_{i=1}^{n-n'}\sum_{j=1}^{m}\mathrm{I}\left(\tilde{X}_{(i)}\geq Y_j\right)\geq\sum_{i=1}^{n-n'}\sum_{j=1}^{m}\mathrm{I}\left(X_{(i)}\geq Y_j\right).$$

Notice that

$$\sum_{i=1}^{n-n'}\sum_{j=1}^{m}\mathrm{I}\left(\tilde{X}_{(i)}\geq Y_j\right)=\sum_{i=1}^{n-n'}\sum_{j=1}^{m}\mathrm{I}\left(\tilde{X}_{i}\geq Y_j\right).$$

Hence,

$$\sum_{i=n'+1}^{n}\sum_{j=1}^{m}\mathrm{I}\left(X_{(i)}\geq Y_j\right)\geq\sum_{i=1}^{n-n'}\sum_{j=1}^{m}\mathrm{I}\left(\tilde{X}_{i}\geq Y_j\right)\geq\sum_{i=1}^{n-n'}\sum_{j=1}^{m}\mathrm{I}\left(X_{(i)}\geq Y_j\right).$$

According to the definition of Wilcoxon-Mann-Whitney U test statistic,

$$\begin{aligned}W(X',Y)&=\sum_{i=1}^{n'}\sum_{j=1}^{m}\mathrm{I}\left(X'_i\geq Y_j\right)\\&=\sum_{i=1}^{n}\sum_{j=1}^{m}\mathrm{I}\left(X_i\geq Y_j\right)-\sum_{i=1}^{n-n'}\sum_{j=1}^{m}\mathrm{I}\left(\tilde{X}_{i}\geq Y_j\right).\end{aligned}$$

Hence,

$$W(X', Y) \leq \sum_{i=1}^{n} \sum_{j=1}^{m} \mathrm{I}\left(X_i \geq Y_j\right) - \sum_{i=1}^{n-n'} \sum_{j=1}^{m} \mathrm{I}\left(X_{(i)} \geq Y_j\right) = W(X_*, Y),$$

and

$$W(X', Y) \geq \sum_{i=1}^{n} \sum_{j=1}^{m} \mathrm{I}\left(X_i \geq Y_j\right) - \sum_{i=n'+1}^{n} \sum_{j=1}^{m} \mathrm{I}\left(X_{(i)} \geq Y_j\right) = W(X_*, Y),$$

which complete our proof. □

We have similar result for Y when $Y' = \{Y'_1, \ldots, Y'_{m'}\}$ is any non-empty proper subset of Y.

Proposition D.5.2. Let $X = \{X_1, \ldots, X_n\}$ and $Y = \{Y_1, \ldots, Y_m\}$ be continuous random variables. Suppose $Y' = \{Y'_1, \ldots, Y'_{m'}\}$ is any non-empty proper subset of Y. Let us order $Y_1, \ldots, Y_m$ from small to large as $Y_{(1)}, \ldots, Y_{(m)}$ and denote

$$Y_* = Y \setminus \{Y_{(m'+1)}, \ldots, X_{(m)}\},\ Y_* = Y \setminus \{Y_{(1)}, \ldots, Y_{(m-m')}\}.$$

Subsequently, we have

$$W(X, Y_*) \leq W(X, Y') \leq W(X, Y_*).$$

Proof. To start, denote $Y \setminus Y' = \tilde{Y} = \{\tilde{Y}_1, \ldots, \tilde{Y}_{m-m'}\}$. Let us order random variables in $\tilde{Y}$ from small to large as $\tilde{Y}_{(1)}, \ldots, \tilde{Y}_{(m-m')}$. Since $\tilde{Y} \subset Y$, for any $i \in \{1, \ldots, m - m'\}$,

$$\tilde{Y}_{(i)} \geq Y_{(i)}, \text{ and } \tilde{Y}_{(i)} \leq Y_{(i+m')}.$$

Subsequently,

$$\sum_{i=1}^{n} \sum_{j=1}^{m-m'} \mathrm{I}\left(X_i \geq Y_{(j)}\right) \geq \sum_{i=1}^{n} \sum_{j=1}^{m-m'} \mathrm{I}\left(X_i \geq \tilde{Y}_{(j)}\right) \geq \sum_{i=1}^{n} \sum_{j=m'+1}^{m} \mathrm{I}\left(X_i \geq Y_{(j)}\right).$$

Notice that

$$\sum_{i=1}^{n} \sum_{j=1}^{m-m'} \mathrm{I}\left(X_i \geq \tilde{Y}_{(j)}\right) = \sum_{i=1}^{n} \sum_{j=1}^{m-m'} \mathrm{I}\left(X_i \geq \tilde{Y}_j\right).$$

Hence,

$$\sum_{i=1}^{n} \sum_{j=1}^{m-m'} \mathrm{I}\left(X_i \geq Y_{(j)}\right) \geq \sum_{i=1}^{n} \sum_{j=1}^{m-m'} \mathrm{I}\left(X_i \geq \tilde{Y}_j\right) \geq \sum_{i=1}^{n} \sum_{j=m'+1}^{m} \mathrm{I}\left(X_i \geq Y_{(j)}\right),$$

According to the definition of Wilcoxon-Mann-Whitney U test statistic,

$$\begin{aligned} W(X,Y') &= \sum_{i=1}^{n}\sum_{j=1}^{m'} \mathrm{I}\left(X_i \geq Y'_j\right) \\ &= \sum_{i=1}^{n}\sum_{j=1}^{m} \mathrm{I}\left(X_i \geq Y_j\right) - \sum_{i=1}^{n}\sum_{j=1}^{m-m'} \mathrm{I}\left(X_i \geq \tilde{Y}_j\right). \end{aligned}$$

Hence,

$$W(X,Y') \leq \sum_{i=1}^{n}\sum_{j=1}^{m} \mathrm{I}\left(X_i \geq Y_j\right) - \sum_{i=1}^{n}\sum_{j=m'+1}^{m} \mathrm{I}\left(X_i \geq Y_{(j)}\right) = W(X,Y_*),$$

and

$$W(X,Y') \geq \sum_{i=1}^{n}\sum_{j=1}^{m} \mathrm{I}\left(X_i \geq Y_j\right) - \sum_{i=1}^{n}\sum_{j=1}^{m-m'} \mathrm{I}\left(X_i \geq Y_{(j)}\right) = W(X,Y_*).$$

which complete our proof. □

Subsequently, by combining Proposition D.5.1 and D.5.2 together, we have the following theorem.

Theorem D.5.3. Let $X = \{X_1, \ldots, X_n\}$ and $Y = \{Y_1, \ldots, Y_m\}$ be continuous random variables. Suppose $X' = \{X'_1, \ldots, X'_{n'}\}$ is any non-empty proper subset of X and $Y' = \{Y'_1, \ldots, Y'_{m'}\}$ is any non-empty proper subset of Y. Let us order $X_1, \ldots, X_n$ from small to large as $X(1), \ldots, X(n)$ and denote

$$X_* = X \setminus \{X(n'+1), \ldots, X(n)\},\ X^* = X \setminus \{X(1), \ldots, X(n-n')\}.$$

Similarly, let us order $Y_1, \ldots, Y_m$ from small to large as $Y(1), \ldots, Y(m)$ and denote

$$Y_* = Y \setminus \{Y(m'+1), \ldots, Y(m)\},\ Y^* = Y \setminus \{Y(1), \ldots, Y(m-m')\}.$$

Then, we have

$$W(X_*, Y^*) \leq W(X', Y') \leq W(X^*, Y_*).$$

Proof. Applying Proposition D.5.1, it follows

$$W(X_*, Y') \leq W(X', Y') \leq W(X_*, Y').$$

Applying Proposition D.5.2 to $W(X_*, Y')$ and $W(X_*, Y')$ separately,

$$\begin{aligned} W(X_*, Y_*) \leq W(X_*, Y') \leq W(X_*, Y_*), \\ W(X_*, Y_*) \leq W(X_*, Y') \leq W(X_*, Y_*). \end{aligned}$$

Hence,

$$W(X_*, Y_*) \leq W(X', Y') \leq W(X_*, Y_*),$$

which proves our result. □

D.6. p-value of the variant method

This section proves the Proposition 6.2.3 in the main.

Proposition D.6.1. Suppose $X = \{X_1, \ldots, X_n\}$ and $Y = \{Y_1, \ldots, Y_m\}$ are independent continuous random variables following the same distribution, $X' = \{X'_1, \ldots, X'_{n'}\}$ is a non-empty subset of X and $Y' = \{Y'_1, \ldots, Y'_{m'}\}$ is a non-empty subset of Y. Let $\{W_i(P_*, Q^*)\}_{i=1}^B$ and $\{W_i(P^*, Q_*)\}_{i=1}^B$ be independent random variables generated according to Algorithm 1. Define

$$p_1 = \frac{1 + \sum_{i=1}^B \mathrm{I}\left(W_i(P_*, Q^*) \leq W(X', Y')\right)}{B+1},$$
$$\text{and } p_2 = \frac{1 + \sum_{i=1}^B \mathrm{I}\left(W_i(P^*, Q_*) \geq W(X', Y')\right)}{B+1}.$$

Then, we have $\Pr(p_1 \leq \alpha/2) \leq \alpha/2$ and $\Pr(p_2 \leq \alpha/2) \leq \alpha/2$ for a given significance level $\alpha \in (0, 1)$. Therefore, defining the p-value of the test statistic $W(X', Y')$ as

$$p = \min\{2\min\{p_1, p_2\}, 1\},$$

we further have $\Pr(p \leq \alpha) \leq \alpha$.

Proof. We first show that $\Pr(p_1 \leq \alpha/2) \leq \alpha/2$, and $\Pr(p_2 > 1 - \alpha/2) \leq \alpha/2$ can then be proven similarly.

Define $W(X_*, Y^*)$ and $W(X^*, Y_*)$ according to Theorem 6.2.1, and denote

$$p_1^* = \frac{1 + \sum_{i=1}^B \mathrm{I}\left(W_i(P_*, Q^*) \leq W(X_*, Y^*)\right)}{B+1}.$$

According to Theorem 6.2.1, we have $W(X_*, Y^*) \leq W(X', Y')$. Hence,

$$p_1^* \leq p_1 \implies \Pr(p_1 < \alpha/2) \leq \Pr(p_1^* < \alpha/2).$$

Therefore, for proving $\Pr(p_1 \leq \alpha/2) \leq \alpha/2$, it is sufficient to show that

$$\Pr(p_1^* \leq \alpha/2) \leq \alpha/2.$$

Note that

$$\begin{aligned} p_1^* \leq \frac{\alpha}{2} &\iff \frac{1 + \sum_{i=1}^B \mathrm{I}\left(W_i(P_*, Q^*) \leq W(X_*, Y^*)\right)}{B+1} \leq \frac{\alpha}{2} \\ &\iff \sum_{i=1}^B \mathrm{I}\left(W_i(P_*, Q^*) \leq W(X_*, Y^*)\right) \leq \frac{\alpha(B+1)}{2} - 1. \end{aligned}$$

If $\alpha(B+1)/2-1<0$, then we have

$$\Pr\left(p_1^* \leq \frac{\alpha}{2}\right) = 0 < \frac{\alpha}{2},$$

which proves the result directly. Hence, we consider the case when $\alpha(B+1)/2-1 \geq 0$. Let k be the maximum integer smaller or equal to $(\alpha(B+1))/2-1$. Then, we further have

$$p_1^* \leq \alpha/2 \iff \sum_{i=1}^{B} \mathrm{I}\left(W_i(P_*, Q^*) \leq W(X_*, Y^*)\right) \leq k.$$

Suppose we can show that

$$\Pr\left(\sum_{i=1}^{B} \mathrm{I}\left(W_i(P_*, Q^*) \leq W(X_*, Y^*)\right) \leq k\right) \leq \frac{k+1}{B+1}, \tag{D.11}$$

then since

$$\frac{k+1}{B+1} \leq \frac{1}{B+1}\left(\frac{\alpha(B+1)}{2}-1\right) + \frac{1}{B+1} = \frac{\alpha}{2},$$

we would have

$$\Pr\left(p_1^* \leq \alpha/2\right) = \Pr\left(\sum_{i=1}^{B} \mathrm{I}\left(W_i(P_*, Q^*) \leq W(X_*, Y^*)\right) \leq k\right) \leq \frac{k+1}{B+1} \leq \frac{\alpha}{2}.$$

Therefore, for completing our proving $\Pr\left(p_1 < \alpha/2\right) \leq \alpha/2$, it is then sufficient to show that Inequality (D.11) holds. Note that $\{W_i(P_*, Q^*)\}_{i=1}^{B}$ and $W(X_*, Y^*)$ are independent and identically distributed. The Inequality (D.11) can be shown straightforwardly if $\{W_i(P_*, Q^*)\}_{i=1}^{B}$ and $W(X_*, Y^*)$ are all distinct, i.e. without ties, by applying the fact that $\{W_i(P_*, Q^*)\}_{i=1}^{B}$ and $W(X_*, Y^*)$ are exchangeable.

However, since $\{W_i(P_*, Q^*)\}_{i=1}^{B}$ and $W(X_*, Y^*)$ are discrete random variables, the probability that the ties exist is non-zero. For proving Inequality (D.11), we therefore introduce random noises for breaking the ties first. Let

$$\hat{W}_i = W_i(P_*, Q^*) + U_i, \text{ for any } i \in \{1, \ldots, B\}, \text{ and } \hat{W}(X_*, Y^*) = (X_*, Y^*) + U,$$

where $U, U_1, \ldots, U_B$ are independent identically distributed random variable following standard uniform distribution between $[0, 1]$.

Note that $\{\hat{W}_i(P_*, Q^*)\}_{i=1}^{B}$ and $\hat{W}(X_*, Y^*)$ are still independent and identically distributed, but now with probability of ties being 0. Therefore, by applying the changeability of $\{\hat{W}_i(P_*, Q^*)\}_{i=1}^{B}$ and $\hat{W}(X_*, Y^*)$, we have

$$\Pr\left(\sum_{i=1}^{B} \mathrm{I}\left(\hat{W}_i(P_*, Q^*) \leq \hat{W}(X_*, Y^*)\right) \leq k\right) \leq \frac{k+1}{B+1}.$$

Note that

$$\hat{W}_i(P_*, Q^*) \leq \hat{W}(X_*, Y^*) \implies W_i(P_*, Q^*) \leq W(X_*, Y^*),$$

since $U, U_1, \ldots, U_B \in [0, 1]$. Hence

$$\sum_{i=1}^{B} \mathrm{I}\left(W_i(P_*, Q^*) \leq W(X_*, Y^*)\right) \geq \sum_{i=1}^{B} \mathrm{I}\left(\hat{W}_i(P_*, Q^*) \leq \hat{W}(X_*, Y^*)\right),$$

which further implies that

$$\begin{aligned}\Pr\left(\sum_{i=1}^{B} \mathrm{I}\left(W_i(P_*, Q^*) \leq W(X_*, Y^*)\right) \leq k\right) &\leq \Pr\left(\sum_{i=1}^{B} \mathrm{I}\left(\hat{W}_i(P_*, Q^*) \leq \hat{W}(X_*, Y^*)\right) \leq k\right) \\ &\leq \frac{k+1}{B+1}.\end{aligned}$$

Hence, we have shown Inequality (D.11). This finish our proof for $\Pr(p_1 \leq \alpha/2) \leq \alpha/2$.

The inequality $\Pr(p_2 \leq \alpha/2) \leq \alpha/2$ can be proved similarly. Thus, it is omitted here.

Now we show that $\Pr(p \leq \alpha) \leq \alpha$, where

$$p = \min\{2\min\{p_1, p_2\}, 1\}.$$

If $2\min\{p_1, p_2\} \geq 1$, then we have $p = 1$. Since $\alpha < 1$, we have

$$\Pr(p \leq \alpha) = 0 \leq \alpha.$$

However, suppose $2\min\{p_1, p_2\} < 1$. We then have $p = 2\min\{p_1, p_2\}$. Further

$$\begin{aligned}\Pr(p \leq \alpha) &= \Pr(2\min\{p_1, p_2\} \leq \alpha) \\ &= \Pr\left(p_1 \leq \frac{\alpha}{2} \text{ or } p_2 \geq \frac{\alpha}{2}\right) \\ &\leq \Pr\left(p_1 \leq \frac{\alpha}{2}\right) + \Pr\left(p_2 \geq \frac{\alpha}{2}\right) \\ &\leq \alpha.\end{aligned}$$

This completes our proofs. □

E. Additional simulation results

E.1. Additional simulation results for Chapter 3

This section provides additional simulation results for Chapter 3.

E.1.1. Additional results for data missing completely at random

In this section, we consider the data to be missing completely at random (MCAR), but with different sample sizes n, m and mean value shift as considered in the experiment for Figure 3.1. The results for this section are shown in Figure E.1 and Figure E.2. Figure E.1 considers larger sample sizes $n = m = 500$, and shows similar results to Figure 3.1, although the power of the proposed method in Figure E.1 appears to be larger.

Figure E.2 shows the increased power of the proposed method by considering the first sample consists of observations from $\mathrm{N}(0, 1)$, but the second sample consists of observations from $\mathrm{N}(2, 1)$. The proposed method still has good statistical power when 20% of the data are missing.

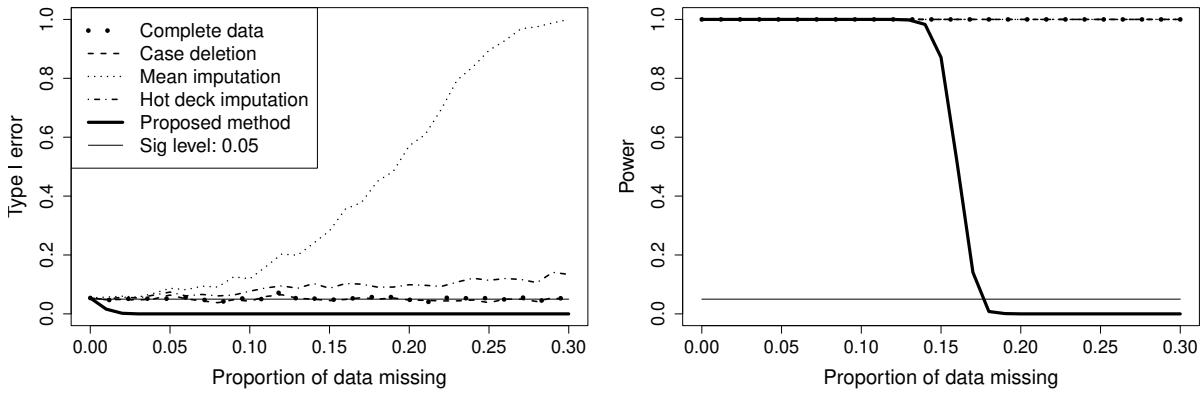


Figure E.1.: The Type I error and statistical power of the proposed method and the standard Wilcoxon-Mann-Whitney test after the missing data is either known or has been imputed or ignored as the proportion of missing data increases. The data is missing completely at random (MCAR). (Left) Type I error: $\mathrm{N}(0,1)$ vs $\mathrm{N}(0,1)$; (Right) Power: $\mathrm{N}(0,1)$ vs $\mathrm{N}(1,1)$. For both figures, a significance threshold of $\alpha = 0.05$ has been used and the total sample sizes are $n = 500$, $m = 500$, and 1000 trials were used.

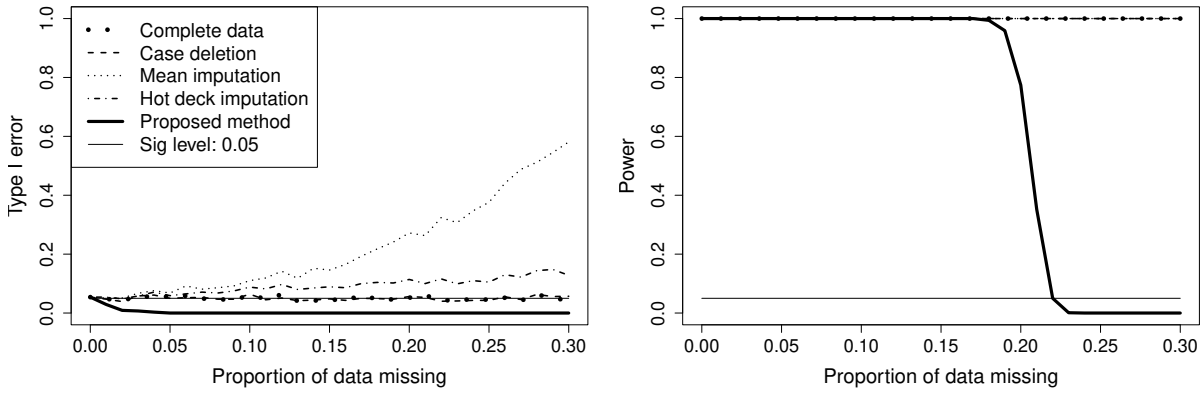


Figure E.2.: The Type I error and statistical power of the proposed method and the standard Wilcoxon-Mann-Whitney test after the missing data is either known or has been imputed or ignored as the proportion of missing data increases. The data is missing completely at random (MCAR). (Left) Type I error: $\mathrm{N}(0,1)$ vs $\mathrm{N}(0,1)$; (Right) Power: $\mathrm{N}(0,1)$ vs $\mathrm{N}(2,1)$. For both figures, a significance threshold of $\alpha = 0.05$ has been used and the total sample sizes are $n = 100$, $m = 100$, and 1000 trials were used.

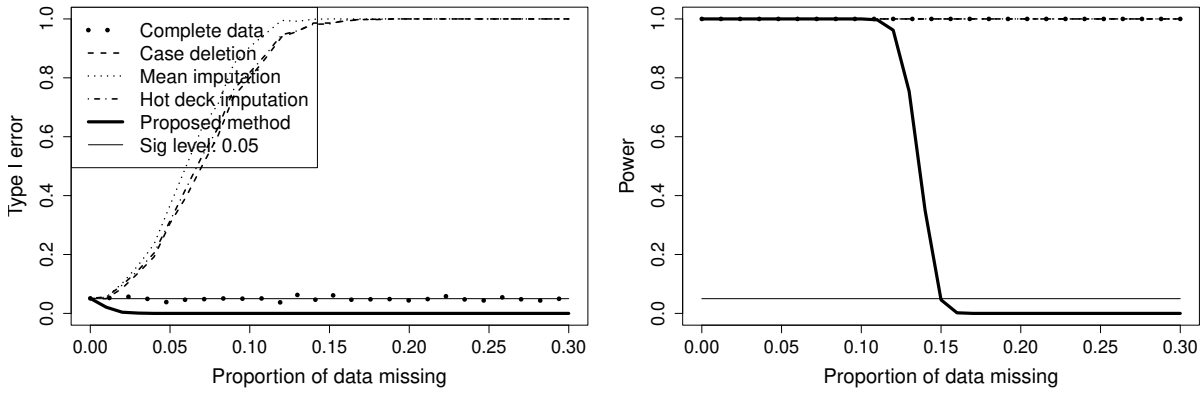


Figure E.3.: The Type I error and statistical power of the proposed method and the standard Wilcoxon-Mann-Whitney test after the missing data is either known or has been imputed or ignored as the proportion of missing data increases. The data is missing not at random (MNAR), according to the mechanism in Equation (3.10), and (3.11). (Left) Type I error: $\mathrm{N}(0,1)$ vs $\mathrm{N}(0,1)$; (Right) Power: $\mathrm{N}(0,1)$ vs $\mathrm{N}(1,1)$. For both figures, a significance threshold of $\alpha = 0.05$ has been used and the total sample sizes are $n = 500$, $m = 500$, and 1000 trials were used.

E.1.2. Additional results for data missing not at random

This section considers the data to be missing not at random (MNAR) following the same missingness mechanism used in Figure 3.2, but with different sample sizes n, m and mean value shift. The results in this section are shown in Figure E.3 and Figure E.4. Figure E.3 considers larger sample size $n, m = 500$ rather than $n = m = 100$. It shows that the Type I error of case deletion, mean imputation and hot deck imputation increases as the sample size increases. When the proportion of missing data is around 15%, the Type I error of the three methods is close to 1, while the proposed method controls the Type I error. Additionally, results in Figure E.3 also suggests the power of the proposed method increases with larger sample sizes.

Figure E.4 considers the case where the first sample consists of observations from $\mathrm{N}(0,1)$, but the second sample consists of observations from $\mathrm{N}(2,1)$. The results are similar to Figure E.3 , but the power of the proposed method is greater.

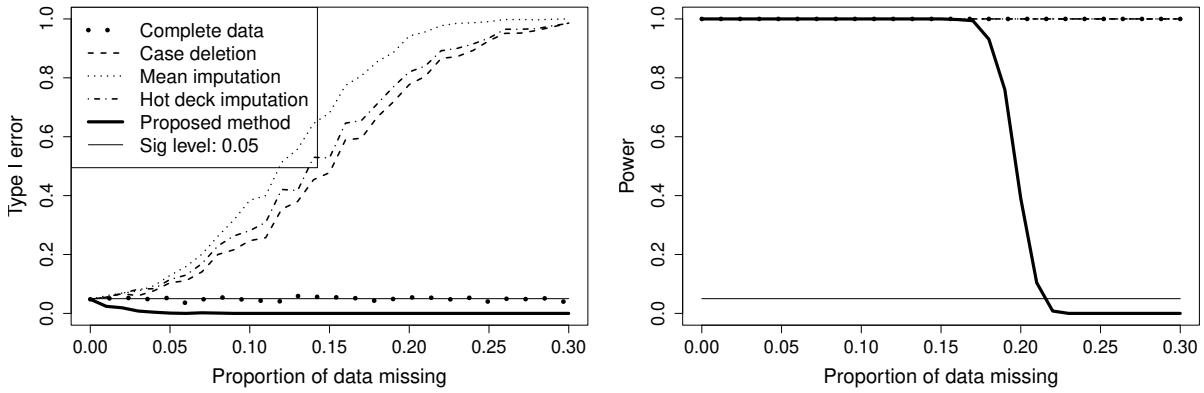


Figure E.4.: The Type I error and statistical power of the proposed method and the standard Wilcoxon-Mann-Whitney test after the missing data is either known or has been imputed or ignored as the proportion of missing data increases. The data is missing not at random (MNAR), according to the mechanism in Equation (3.10), and (3.11). (Left) Type I error: $\mathrm{N}(0,1)$ vs $\mathrm{N}(0,1)$; (Right) Power: $\mathrm{N}(0,1)$ vs $\mathrm{N}(2,1)$. For both figures, a significance threshold of $\alpha = 0.05$ has been used and the total sample sizes are $n = 100$, $m = 100$, and 1000 trials were used.

E.1.3. Gamma-distributed data

This section presents Figure E.5 and Figure E.6, which are similar to Figure 3.1 and Figure 3.2 in the main results, but the data are now Gamma-distributed, rather than Normal-distributed.

Similar to Figure 3.1, Figure E.5 also considers the data to be missing completely at random (MCAR). However, the data in Figure E.5 follow the Gamma distribution, rather than the normal distribution. Specifically, the data in sample X are generated from Gamma(1,1), while the data in sample Y are generated in Gamma(1,1) for evaluating the Type I error, and Gamma(4,2) for evaluating power. The distributions Gamma(1,1) and Gamma(4,2) are chosen to match the mean value shift and variance settings used in the normal case in Figure 3.1: Gamma(1,1) has mean 1, and variance 1, and Gamma(4,2) has mean 2, and variance 1. The results of Figure E.5 are similar to the results of Figure 3.1.

Figure E.6 considers the data to be missing not at random (MNAR). The data in sample X are generated from Gamma(1,1), while the data in sample Y are generated in Gamma(1,1) for evaluating the Type I error, and Gamma(4,2) for evaluating power.

The missingness mechanisms of X and Y are similar to Equation (E.1) and Equation (E.2), with slight modifications:

$$\Pr\left(x_i \text{ is missing}\right) = \begin{cases} \min(1, sn/\sum_{j=1}^{n} \mathrm{I}\left(x_j < 1\right)), & \text{if } x_i < 1, \\ \max(0, sn/\sum_{j=1}^{n} \mathrm{I}\left(x_j < 1\right) - 1), & \text{otherwise,} \end{cases} \tag{E.1}$$

and

$$\Pr\left(y_i \text{ is missing}\right) = \begin{cases} \min(1, sm/\sum_{j=1}^{m} \mathrm{I}\left(y_j > 1\right)), & \text{if } y_i > 1, \\ \max(0, sm/\sum_{j=1}^{m} \mathrm{I}\left(y_j > 1\right) - 1), & \text{otherwise,} \end{cases} \tag{E.2}$$

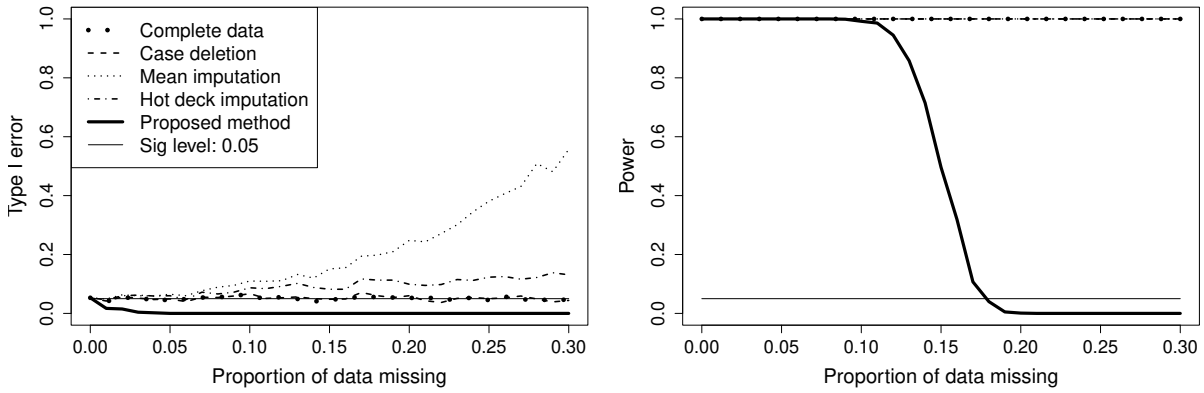


Figure E.5.: The Type I error and statistical power of the proposed method and the standard Wilcoxon-Mann-Whitney test after the missing data is either known or has been imputed or ignored as the proportion of missing data increases. The data is missing completely at random (MCAR). (Left) Type I error: Gamma$(1, 1)$ vs Gamma$(1, 1)$; (Right) Power: Gamma$(1, 1)$ vs Gamma$(4, 2)$. For both figures, a significance threshold of $\alpha = 0.05$ has been used and the total sample sizes are $n = 500$, $m = 500$, and 1000 trials were used.

The missingness mechanism is modified because all values from Gamma distributions are positive. Hence, the original missingness mechanism specified by Equation (3.10) and Equation (3.11) will randomly select values in X and Y to be missing, and the missingness case is equivalent to MCAR.

To make data MNAR, we consider missingness mechanism specified in Equation (E.1) and (E.2). Equation (E.1) and (E.2) are similar to Equation (3.10) and (3.11), but they select values in X and Y according to whether values is larger or smaller than 1, rather than 0. The results in Figure E.6 are also similar to the results in Figure 3.2.

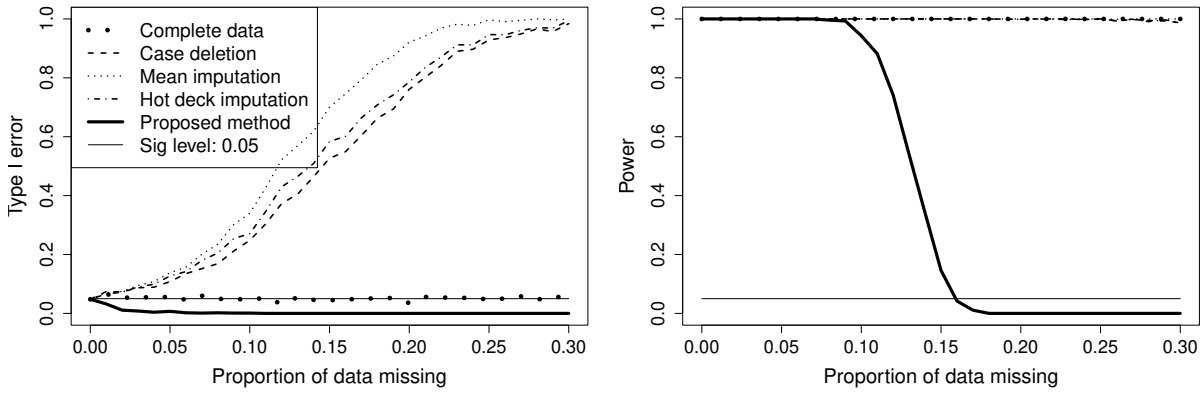


Figure E.6.: The Type I error and statistical power of the proposed method and the standard Wilcoxon-Mann-Whitney test after the missing data is either known or has been imputed or ignored as the proportion of missing data increases. The data is missing not at random (MNAR), according to the mechanism in Equation (E.1), and (E.2). (Left) Type I error: Gamma$(1, 1)$ vs Gamma$(1, 1)$; (Right) Power: Gamma$(1, 1)$ vs Gamma$(4, 2)$. For both figures, a significance threshold of $\alpha = 0.05$ has been used and the total sample sizes are $n = 100$, $m = 100$, and 1000 trials were used.

E.1.4. A different missingness mechanism

This section presents Figure E.7, which considers a similar case to Figure 3.2 when data are missing not at random (MNAR), but Figure E.7 applies a different missingness mechanism. The data in X are randomly selected to be missing, while the data in Y are missing according to

$$\Pr\left(y_i \text{ is missing}\right) = \frac{y_i + |\min Y|}{\sum_{j=1}^{m} y_j + m|\min Y|}. \tag{E.3}$$

In other words, the probability of a value in Y to be missing is proportional to its value after adding $|\min Y|$. The value $|\min Y|$ is added to each value in Y so that all values in Y have positive probability of being selected – except for $\min Y$ if $\min Y \leq 0$.

The results of Figure E.7 show that the Type I error of case deletion, mean imputation and hot deck imputation is smaller than that of Figure 3.2, although it still exceed the significance level α.

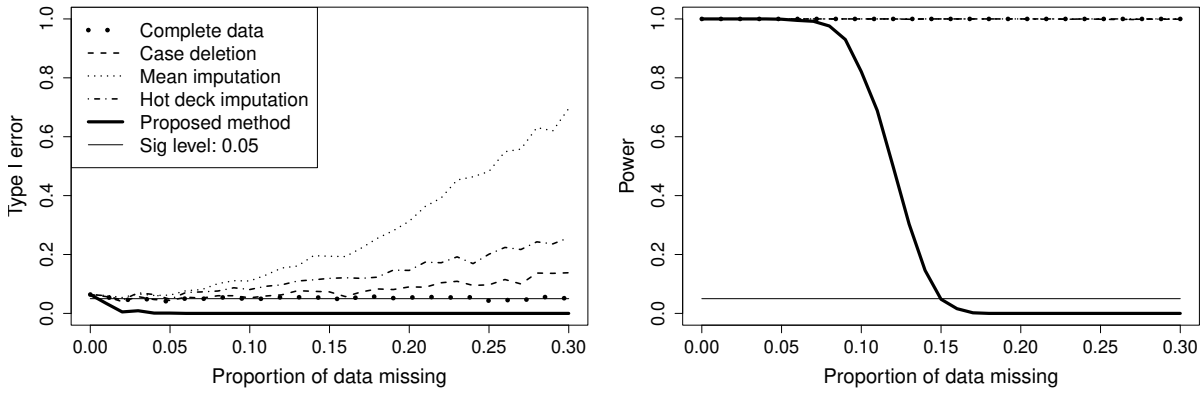


Figure E.7.: The Type I error and statistical power of the proposed method and the standard Wilcoxon-Mann-Whitney test after the missing data is either known or has been imputed or ignored as the proportion of missing data increases. The data are missing not at random (MNAR). The data in X are randomly selected to be missing, while the data in Y are missing according to (E.3). (Left) Type I error: $\mathrm{N}(0,1)$ vs $\mathrm{N}(0,1)$; (Right) Power: $\mathrm{N}(0,1)$ vs $\mathrm{N}(1,1)$. For both figures, a significance threshold of $\alpha = 0.05$ has been used and the total sample sizes are $n = 100$, $m = 100$, and 1000 trials were used.

E.2. Additional simulation results for Chapter 4

This section provides additional simulation results for Chapter 4.

E.2.1. Scale testing for data missing completely at random

This section considers similar experiments as that for Figure 4.1 when data are missing completely at random (MCAR), but with different sample sizes and scale parameters. The results are shown in Figure E.8 and Figure E.9. Figure E.8 considers larger sample sizes $n = m = 500$, and shows similar results to Figure 4.1 with larger power.

Figure E.9 shows the increased power of the proposed method by considering the first sample consists of observations from $\mathrm{N}(0,1)$, but the second sample consists of observations from $\mathrm{N}(0,\sigma^2)$ with $\sigma = 5$, rather than 3 used in Figure 4.1.

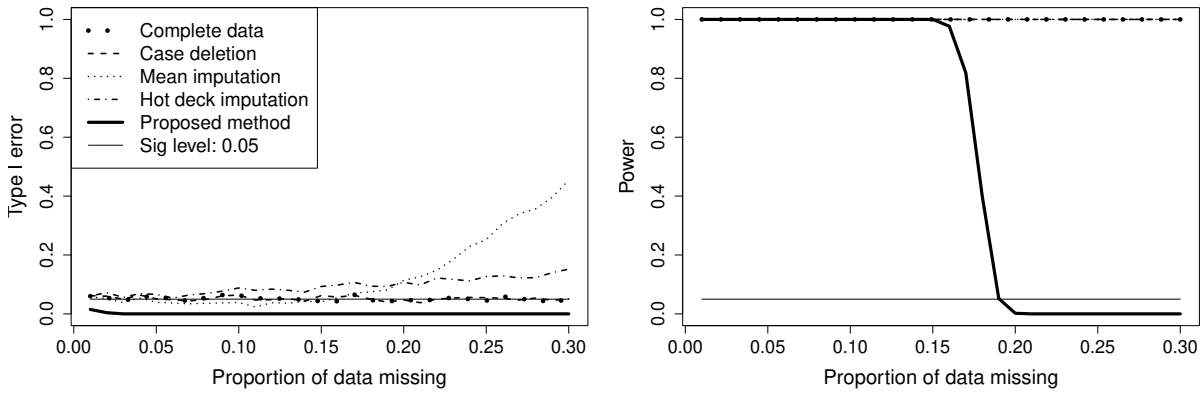


Figure E.8.: The Type I error and statistical power of the proposed method and the standard Ansari-Bradley test after the missing data is either known or has been imputed or ignored as the proportion of missing data increases. The data is missing completely at random (MCAR). (Left) Type I error: $\mathrm{N}(0,1)$ vs $\mathrm{N}(0,1)$; (Right) Power: $\mathrm{N}(0,1)$ vs $\mathrm{N}(0,\sigma^2)$, with the scale parameter $\sigma = 3$. For both figures, a significance threshold of $\alpha = 0.05$ has been used and the total sample sizes are $n = 500$, $m = 500$, and 1000 trials were used.

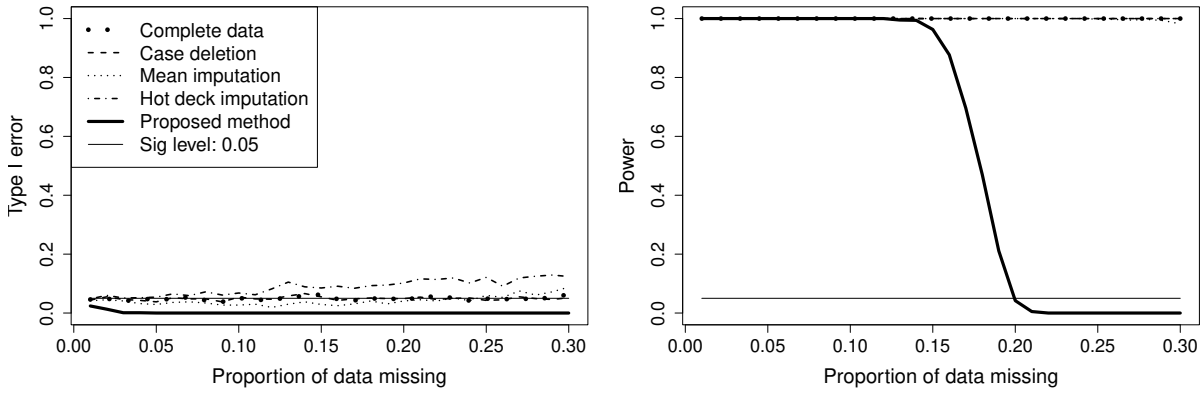


Figure E.9.: The Type I error and statistical power of the proposed method and the standard Ansari-Bradley test after the missing data is either known or has been imputed or ignored as the proportion of missing data increases. The data is missing completely at random (MCAR). (Left) Type I error: $\mathrm{N}(0,1)$ vs $\mathrm{N}(0,1)$; (Right) Power: $\mathrm{N}(0,1)$ vs $\mathrm{N}(0,\sigma^2)$, with the scale parameter $\sigma = 5$. For both figures, a significance threshold of $\alpha = 0.05$ has been used and the total sample sizes are $n = 100$, $m = 100$, and 1000 trials were used.

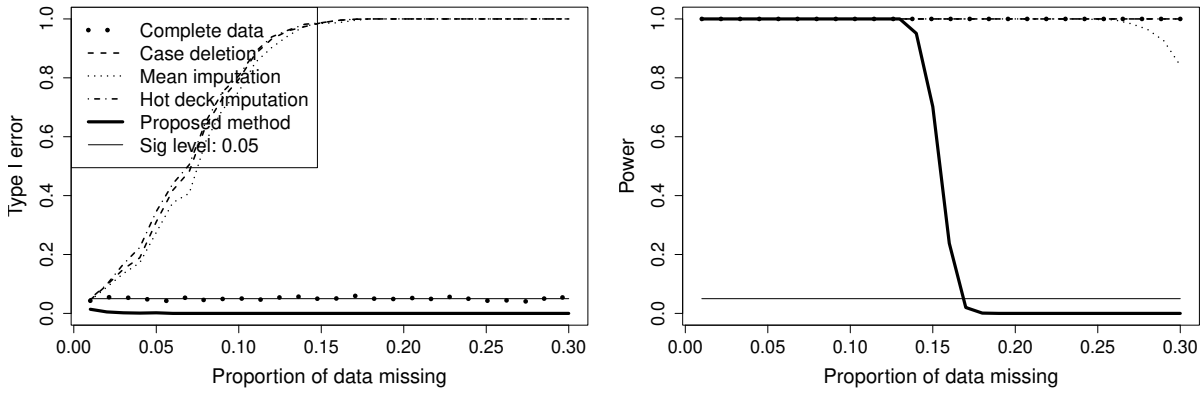


Figure E.10.: The Type I error and statistical power of the proposed method and the standard Ansari-Bradley test after the missing data is either known or has been imputed or ignored as the proportion of missing data increases. The data is missing not at random (MNAR), according to the mechanism in Equation (4.3), and (4.4). (Left) Type I error: $\mathrm{N}(0,1)$ vs $\mathrm{N}(0,1)$; (Right) Power: $\mathrm{N}(0,1)$ vs $\mathrm{N}(0,\sigma^2)$, with scale parameter $\sigma = 3$. For both figures, a significance threshold of $\alpha = 0.05$ has been used and the total sample sizes are $n = 500$, $m = 500$, and 1000 trials were used.

E.2.2. Scale testing for data missing not at random

This section considers experiments similar to Figure 4.2 when data are missing not at random (MNAR), but with different sample sizes and scale parameters. The results are shown in Figure E.10 and E.11.

Figure E.10 considers different sample sizes $n = m = 500$, while Figure E.11 considers the first sample consists of observations from $\mathrm{N}(0,1)$, but the second sample consists of observations from $\mathrm{N}(0,\sigma^2)$ with $\sigma = 5$, rather than 3 used in Figure 4.2.

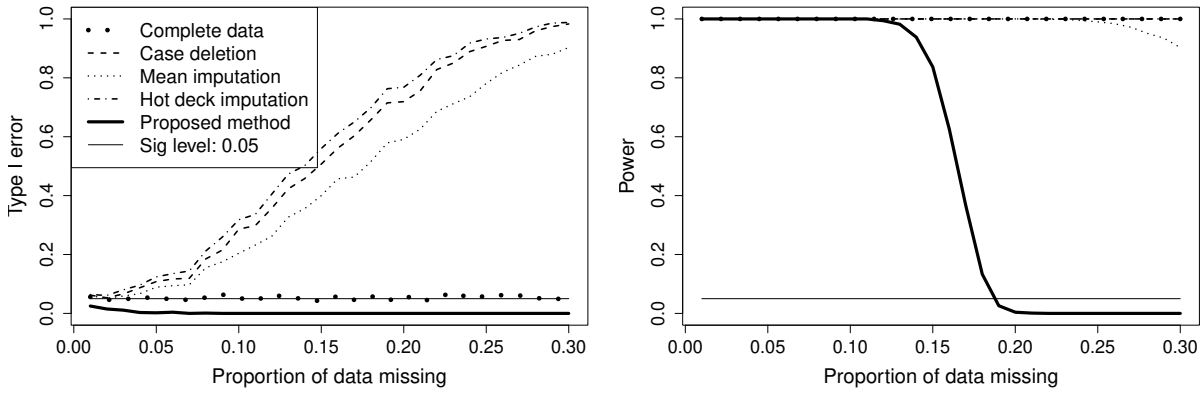


Figure E.11.: The Type I error and statistical power of the proposed method and the standard Ansari-Bradley test after the missing data is either known or has been imputed or ignored as the proportion of missing data increases. The data is missing not at random (MNAR), according to the mechanism in Equation (4.3), and (4.4). (Left) Type I error: $\mathrm{N}(0,1)$ vs $\mathrm{N}(0,1)$; (Right) Power: $\mathrm{N}(0,1)$ vs $\mathrm{N}(0,\sigma^2)$, with scale parameter $\sigma = 5$. For both figures, a significance threshold of $\alpha = 0.05$ has been used and the total sample sizes are $n = 100$, $m = 100$, and 1000 trials were used.

E.2.3. Scale testing for Gamma-distributed data

This section presents Figure E.12 and Figure E.13, which are similar to Figure 4.1 and Figure 4.2 in the main results, but the data are now Gamma-distributed, rather than Normal-distributed.

Similar to Figure 4.1, Figure E.12 also considers the data to be missing completely at random (MCAR). However, the data in Figure E.12 follow the Gamma distribution, rather than the normal distribution. Specifically, the data in sample X are generated from Gamma(1,1), while the data in sample Y are generated in Gamma(1,1) for evaluating the Type I error, and Gamma(9,0.11) for evaluating power. The distributions Gamma(1,1) and Gamma(9,0.11) are chosen so that the two distributions have roughly the same mean value but different variances: Gamma(1,1) has mean 1, and variance 1; Gamma(9,0.11) has mean 0.99, and variance approximately 0.1.

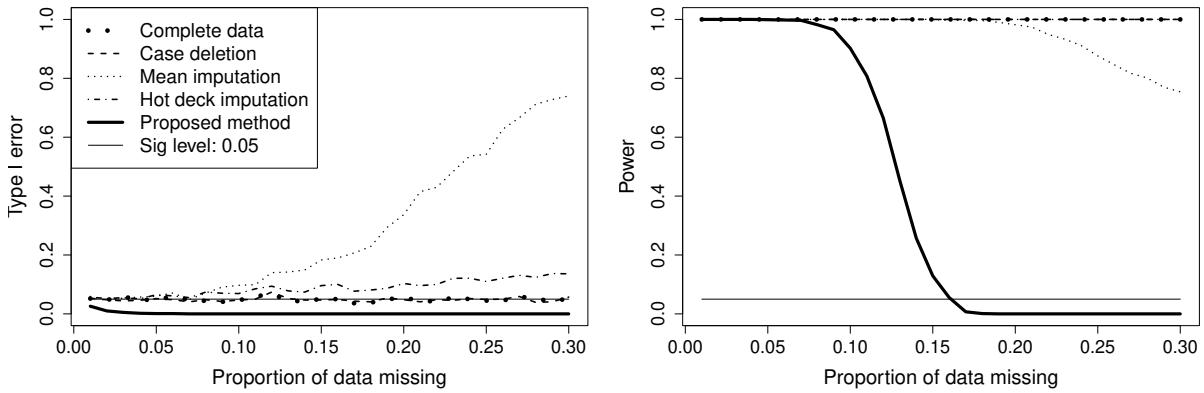


Figure E.12.: The Type I error and statistical power of the proposed method and the standard Ansari-Bradley test after the missing data is either known or has been imputed or ignored as the proportion of missing data increases. The data is missing completely at random (MCAR). (Left) Type I error: Gamma$(1, 1)$ vs Gamma$(1, 1)$; (Right) Power: Gamma$(1, 1)$ vs Gamma$(9, 0.11)$. For both figures, a significance threshold of $\alpha = 0.05$ has been used and the total sample sizes are $n = 100$, $m = 100$, and 1000 trials were used.

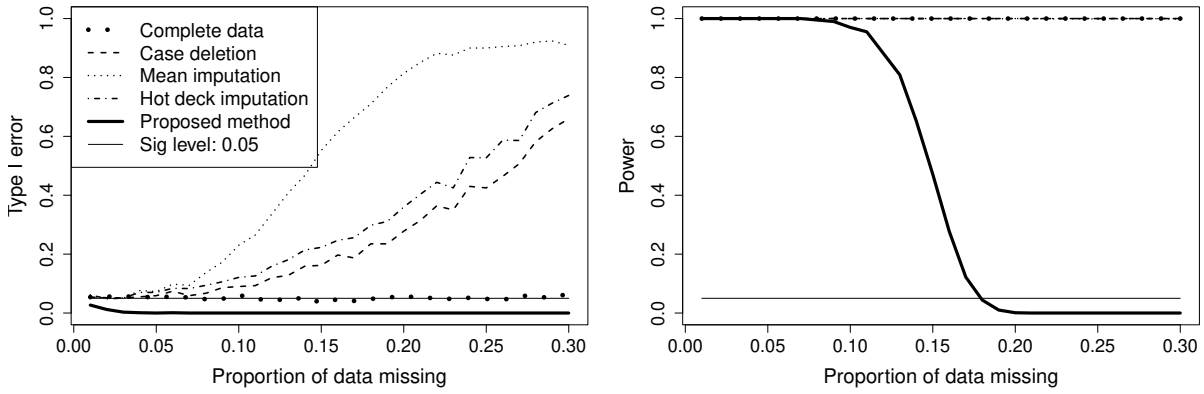


Figure E.13.: The Type I error and statistical power of the proposed method and the standard Ansari-Bradley test after the missing data is either known or has been imputed or ignored as the proportion of missing data increases. The data is missing not at random (MNAR), according to the mechanism in Equation (4.3), and (4.4). (Left) Type I error: Gamma$(1, 1)$ vs Gamma$(1, 1)$; (Right) Power: Gamma$(1, 1)$ vs Gamma$(9, 0.11)$. For both figures, a significance threshold of $\alpha = 0.05$ has been used and the total sample sizes are $n = 100$, $m = 100$, and 1000 trials were used.

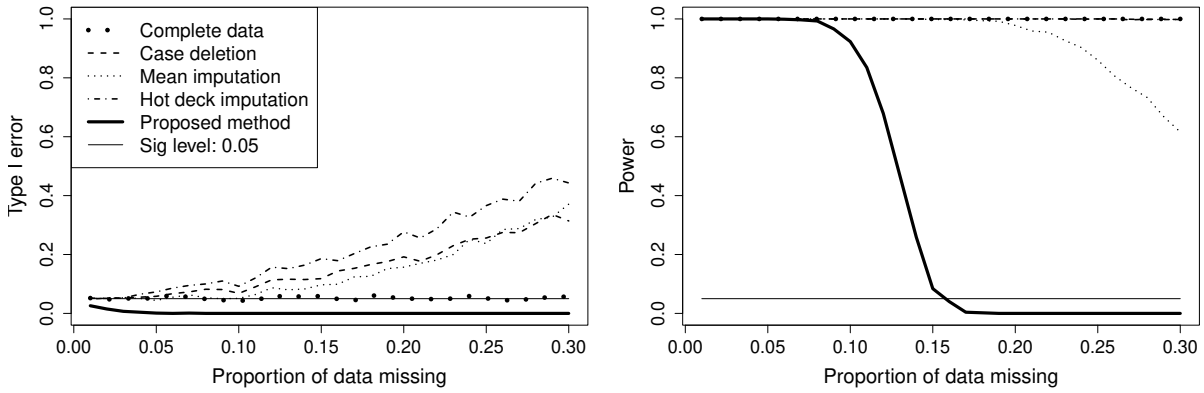


Figure E.14.: The Type I error and statistical power of the proposed method and the standard Ansari-Bradley test after the missing data is either known or has been imputed or ignored as the proportion of missing data increases. The data are missing not at random (MNAR). The data in X are randomly selected to be missing, while the data in Y are missing according to (E.4). (Left) Type I error: $\mathrm{N}(0,1)$ vs $\mathrm{N}(0,1)$; (Right) Power: $\mathrm{N}(0,1)$ vs $\mathrm{N}(0,\sigma^2)$, with scale parameter $\sigma = 3$. For both figures, a significance threshold of $\alpha = 0.05$ has been used and the total sample sizes are $n = 100$, $m = 100$, and 1000 trials were used.

E.2.4. Scale testing for a different missingness mechanism

Figure E.14 considers a similar case to Figure 4.2 when data are missing not at random (MNAR), but with a different missingness mechanism. For Figure E.14, the data in X are randomly selected to be missing, while the data in Y are missing according to

$$\Pr\left(y_i \text{ is missing}\right) = \frac{|y_i|}{\sum_{j=1}^{m} |y_j|}. \tag{E.4}$$

In other words, the probability of a value in Y to be missing is proportional to its absolute value after adding $|\min Y|$.

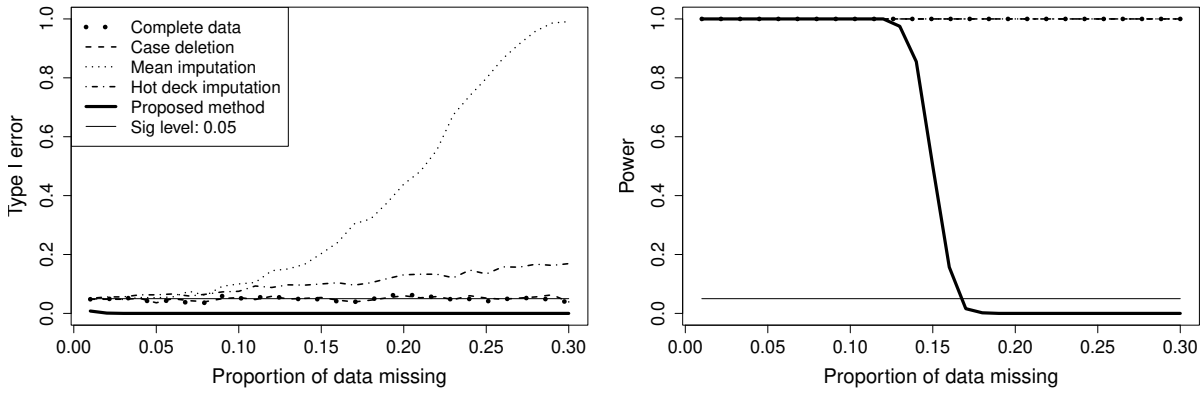


Figure E.15.: The Type I error and statistical power of the proposed location-scale testing method, and the combination of the Ansari-Bradley test and the Wilcoxon-Mann-Whitney test using the Holm-Bonferroni method after the missing data is either known or has been imputed or ignored as the proportion of missing data increases. The data are missing completely at random (MCAR). (Left) Type I error: $\mathrm{N}(0,1)$ vs $\mathrm{N}(0,1)$; (Right) Power: $\mathrm{N}(0,1)$ vs $\mathrm{N}(\mu,\sigma^2)$, with the location parameter $\mu = 1$, and the scale parameter $\sigma = 3$. For both figures, a significance threshold of $\alpha = 0.05$ has been used and the total sample sizes are $n = 500$, $m = 500$, and 1000 trials were used.

E.2.5. Location-scale testing for data missing completely at random

This section considers similar experiments as that for Figure 4.7 when data are missing completely at random (MCAR), but with different sample sizes and scale parameters. The results are shown in Figure E.15 to Figure E.17. Figure E.15 considers larger sample sizes $n = m = 500$, and shows similar results to Figure 4.7 with larger power.

Figure E.16 shows power of the proposed method by considering the first sample consists of observations from $\mathrm{N}(0,1)$, but the second sample consists of observations from $\mathrm{N}(\mu,\sigma^2)$ with different combinations of mean μ and scale σ.

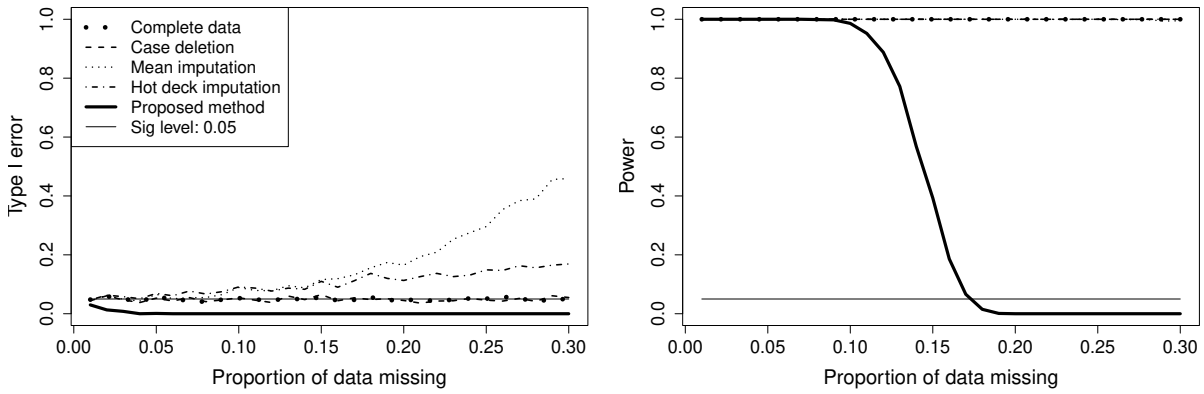


Figure E.16.: The Type I error and statistical power of the proposed location-scale testing method, and the combination of the Ansari-Bradley test and the Wilcoxon-Mann-Whitney test using the Holm-Bonferroni method after the missing data is either known or has been imputed or ignored as the proportion of missing data increases. The data are missing completely at random (MCAR). (Left) Type I error: $\mathrm{N}(0,1)$ vs $\mathrm{N}(0,1)$; (Right) Power: $\mathrm{N}(0,1)$ vs $\mathrm{N}(\mu,\sigma^2)$, with the location parameter $\mu = 1$, and the scale parameter $\sigma = 4$. For both figures, a significance threshold of $\alpha = 0.05$ has been used and the total sample sizes are $n = 100$, $m = 100$, and 1000 trials were used.

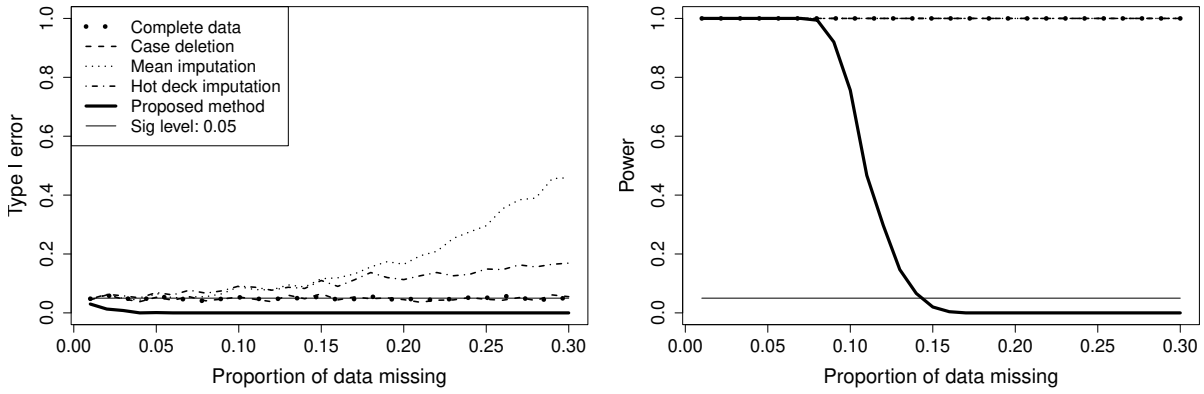


Figure E.17.: The Type I error and statistical power of the proposed location-scale testing method, and the combination of the Ansari-Bradley test and the Wilcoxon-Mann-Whitney test using the Holm-Bonferroni method after the missing data is either known or has been imputed or ignored as the proportion of missing data increases. The data are missing completely at random (MCAR). (Left) Type I error: $\mathrm{N}(0,1)$ vs $\mathrm{N}(0,1)$; (Right) Power: $\mathrm{N}(0,1)$ vs $\mathrm{N}(\mu,\sigma^2)$, with the location parameter $\mu = 2$, and the scale parameter $\sigma = 3$. For both figures, a significance threshold of $\alpha = 0.05$ has been used and the total sample sizes are $n = 100$, $m = 100$, and 1000 trials were used.

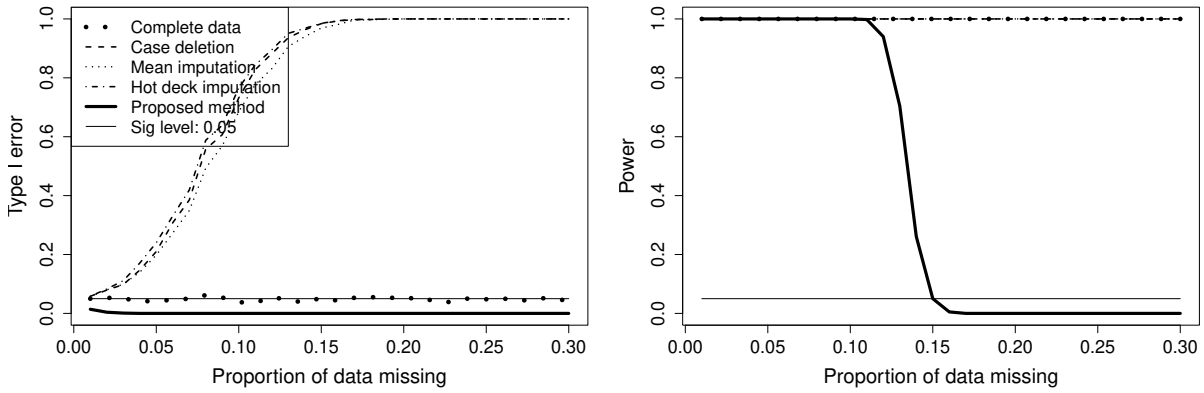


Figure E.18.: The Type I error and statistical power of the proposed location-scale testing method and the combination of the Ansari-Bradley test and the Wilcoxon-Mann-Whitney test using the Holm-Bonferroni method after the missing data is either known or has been imputed or ignored as the proportion of missing data increases. The data are missing not at random (MNAR), according to the mechanism in Equation (4.3), and (4.4). (Left) Type I error: $\mathrm{N}(0,1)$ vs $\mathrm{N}(0,1)$; (Right) Power: $\mathrm{N}(0,1)$ vs $\mathrm{N}(\mu,\sigma)$, with the location parameter $\mu = 1$ and the scale parameter $\sigma = 3$. For both figures, a significance threshold of $\alpha = 0.05$ has been used and the total sample sizes are $n = 500$, $m = 500$, and 1000 trials were used.

E.2.6. Location-scale testing for data missing not at random

This section considers experiments similar to Figure 4.8 when data are missing not at random (MNAR), but with different sample sizes and scale parameters. The results are shown in Figure E.18 to E.20.

Figure E.18 considers different sample sizes $n = m = 500$, while Figure E.19 considers the first sample consists of observations from $\mathrm{N}(0,1)$, but the second sample consists of observations from $\mathrm{N}(\mu,\sigma^2)$ with different combinations of mean μ and scale σ.

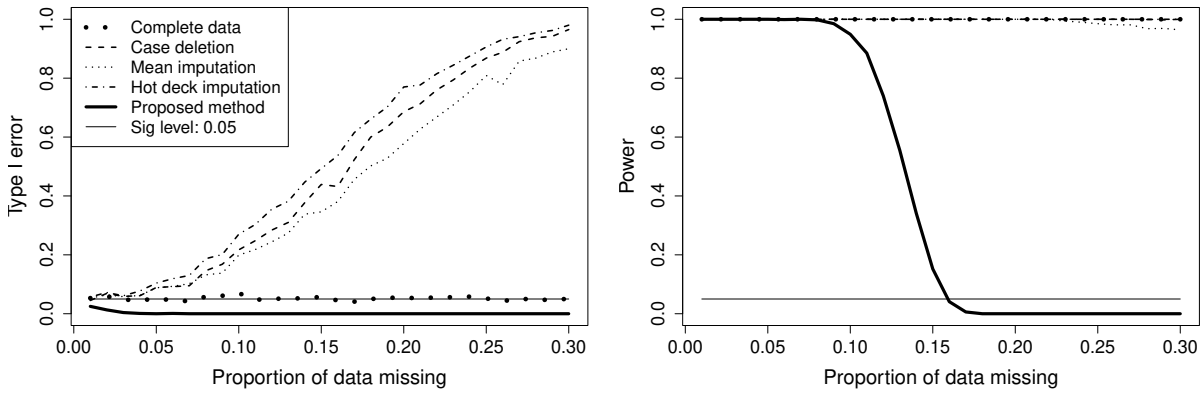


Figure E.19.: The Type I error and statistical power of the proposed location-scale testing method and the combination of the Ansari-Bradley test and the Wilcoxon-Mann-Whitney test using the Holm-Bonferroni method after the missing data is either known or has been imputed or ignored as the proportion of missing data increases. The data are missing not at random (MNAR), according to the mechanism in Equation (4.3), and (4.4). (Left) Type I error: $\mathrm{N}(0,1)$ vs $\mathrm{N}(0,1)$; (Right) Power: $\mathrm{N}(0,1)$ vs $\mathrm{N}(\mu,\sigma)$, with the location parameter $\mu = 1$ and the scale parameter $\sigma = 4$. For both figures, a significance threshold of $\alpha = 0.05$ has been used and the total sample sizes are $n = 100$, $m = 100$, and 1000 trials were used.

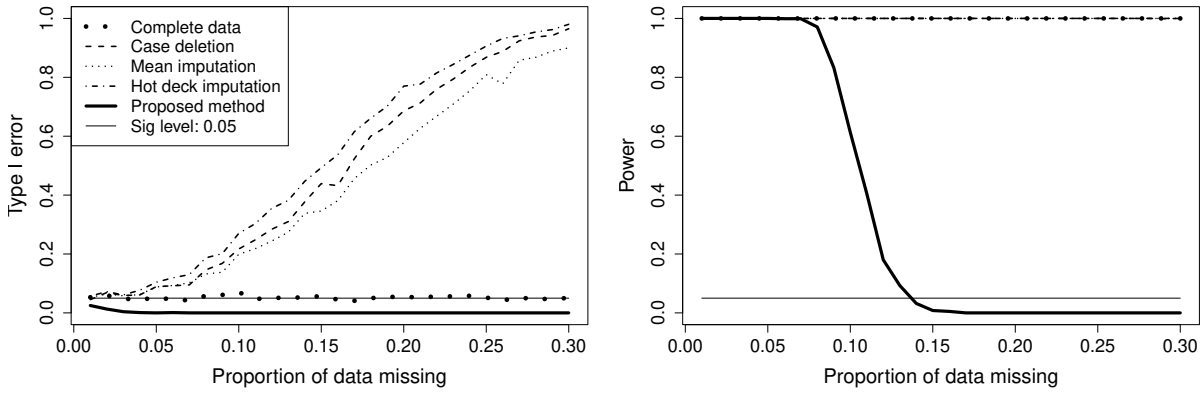


Figure E.20.: The Type I error and statistical power of the proposed location-scale testing method and the combination of the Ansari-Bradley test and the Wilcoxon-Mann-Whitney test using the Holm-Bonferroni method after the missing data is either known or has been imputed or ignored as the proportion of missing data increases. The data are missing not at random (MNAR), according to the mechanism in Equation (4.3), and (4.4). (Left) Type I error: $\mathrm{N}(0,1)$ vs $\mathrm{N}(0,1)$; (Right) Power: $\mathrm{N}(0,1)$ vs $\mathrm{N}(\mu,\sigma)$, with the location parameter $\mu = 2$ and the scale parameter $\sigma = 3$. For both figures, a significance threshold of $\alpha = 0.05$ has been used and the total sample sizes are $n = 100$, $m = 100$, and 1000 trials were used.

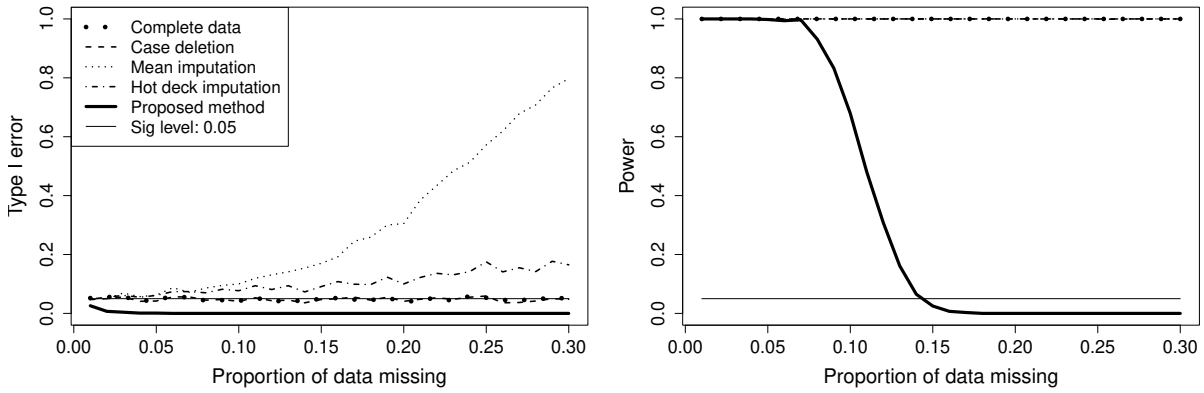


Figure E.21.: The Type I error and statistical power of the proposed location-scale testing method, and the combination of the Ansari-Bradley test and the Wilcoxon-Mann-Whitney test using the Holm-Bonferroni method after the missing data is either known or has been imputed or ignored as the proportion of missing data increases. The data are missing completely at random (MCAR). (Left) Type I error: Gamma(1,1) vs Gamma(1,1); (Right) Power: Gamma(1,1) vs Gamma(5,0.1). For both figures, a significance threshold of $\alpha = 0.05$ has been used and the total sample sizes are $n = 100$, $m = 100$, and 1000 trials were used.

E.2.7. Location-scale testing for Gamma-distributed data

This section presents Figure E.21 and Figure E.22, which are similar to Figure 4.7 and Figure 4.8 in the main results, but the data are now Gamma-distributed, rather than Normal-distributed.

Similar to Figure 4.7, Figure E.21 also considers the data to be missing completely at random (MCAR). However, the data in Figure E.21 follow the Gamma distribution, rather than the normal distribution. Specifically, the data in sample X are generated from Gamma(1,1), while the data in sample Y are generated in Gamma(1,1) for evaluating the Type I error, and Gamma(5,0.1) for evaluating power. The distributions Gamma(1,1) and Gamma(5,0.1) are chosen so that the two distributions have different mean and variance: Gamma(1,1) has mean 1, and variance 1; Gamma(5,0.1) has mean 0.5, and variance 0.5.

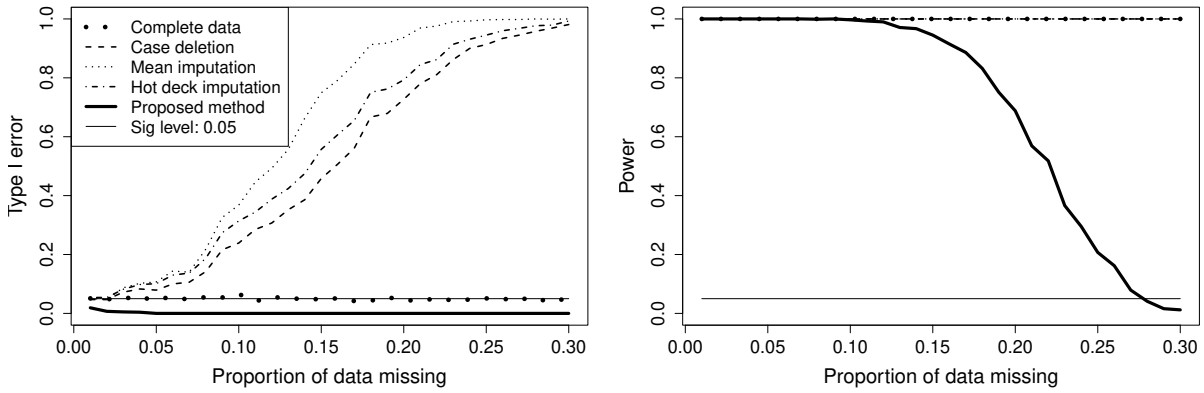


Figure E.22.: The Type I error and statistical power of the proposed location-scale testing method, and the combination of the Ansari-Bradley test and the Wilcoxon-Mann-Whitney test using the Holm-Bonferroni method after the missing data is either known or has been imputed or ignored as the proportion of missing data increases. The data are missing not at random (MNAR), according to the mechanism in Equation (4.3), and (4.4). (Left) Type I error: Gamma(1,1) vs Gamma(1,1); (Right) Power: Gamma(1,1) vs Gamma(5,0.1). For both figures, a significance threshold of $\alpha = 0.05$ has been used and the total sample sizes are $n = 100$, $m = 100$, and 1000 trials were used.

E.2.8. Location-scale testing for a different missingness mechanism

Figure E.23 considers a similar case to Figure 4.8 when data are missing not at random (MNAR), but with a different missingness mechanism. For Figure E.23, the data in X are randomly selected to be missing, while the data in Y are missing according to Equation E.4, i.e., the probability of a value in Y to be missing is proportional to its absolute value after adding $|\min Y|$.

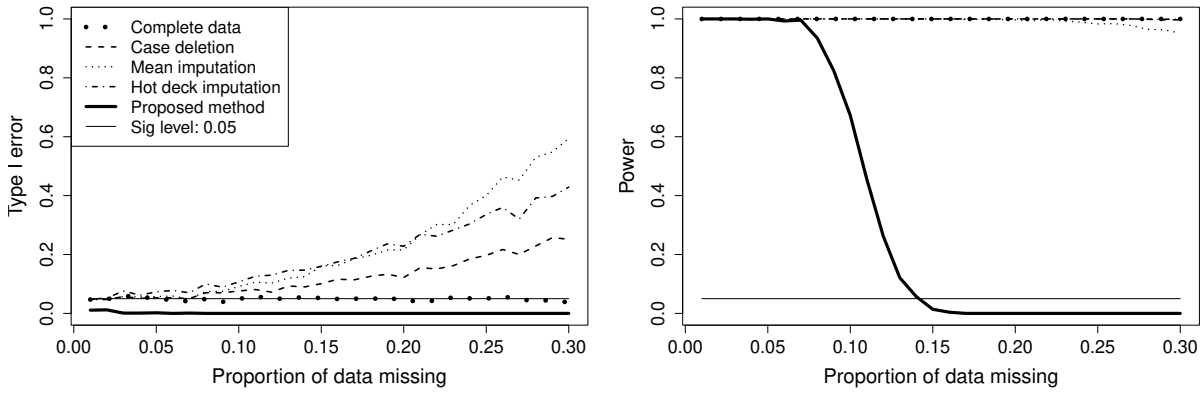


Figure E.23.: The Type I error and statistical power of the proposed location-scale testing method and the combination of the Ansari-Bradley test and the Wilcoxon-Mann-Whitney test using the Holm-Bonferroni method after the missing data is either known or has been imputed or ignored as the proportion of missing data increases. The data are missing not at random (MNAR). The data in X are randomly selected to be missing, while the data in Y are missing according to (E.4). (Left) Type I error: $\mathrm{N}(0,1)$ vs $\mathrm{N}(0,1)$; (Right) Power: $\mathrm{N}(0,1)$ vs $\mathrm{N}(\mu,\sigma^2)$, with $\mu = 1$ and $\sigma = 3$. For both figures, a significance threshold of $\alpha = 0.05$ has been used and the total sample sizes are $n = 100$, $m = 100$, and 1000 trials were used.

E.3. Additional simulation results for Chapter 5

This section provides additional simulation results for Chapter 5.

E.3.1. Bounds with missing data

We first perform simulations following the similar approach as for the figures in Section 5.4 in the main text, with different correlation coefficients $\gamma \in \{-0.5, 0.5\}$. The results are shown in Figure E.24–E.27.

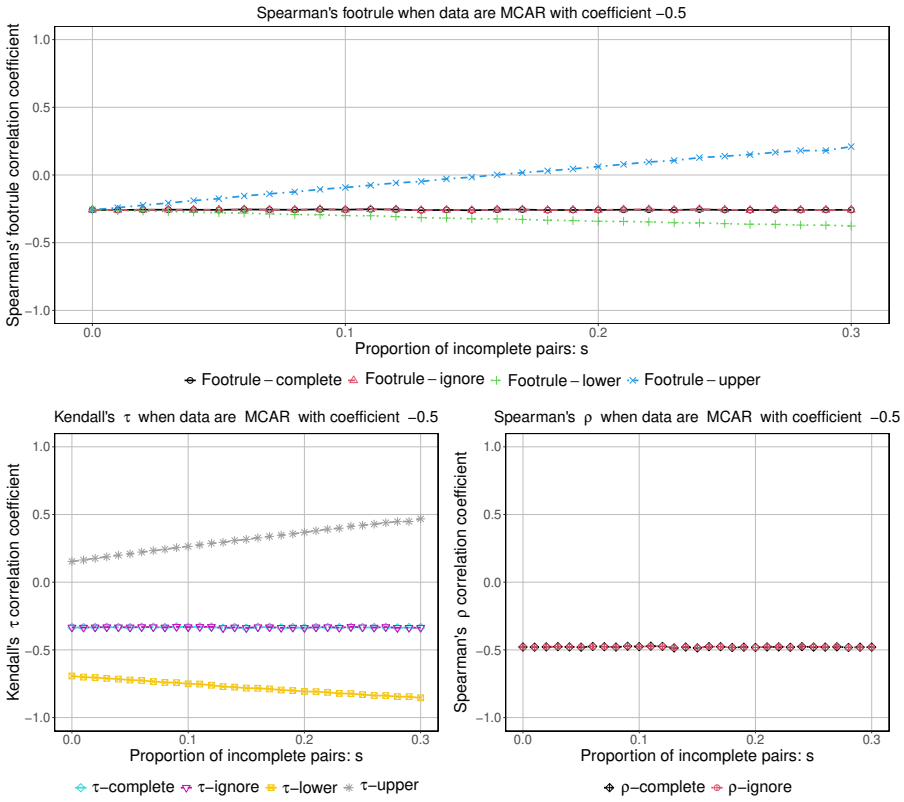


Figure E.24.: Rank correlation coefficients when data are missing completely at random (MCAR). The subset of indices T containing all components with at least one missing value is selected randomly. For each $i \in T$, there is probability $1/3$ such that $X(i)$ is missing but $Y(i)$ is observed, probability $1/3$ such that $Y(i)$ is missing but $X(i)$ is observed, and probability $1/3$ such that both $X(i)$ and $Y(i)$ are missing. X, Y are generated such that $(X, Y) \overset{\text{i.i.d.}}{\sim} \mathrm{N}(0, \Sigma)$, where $\Sigma = \begin{pmatrix} 1 & \gamma \\ \gamma & 1 \end{pmatrix}$, with covariance coefficient $\gamma = -0.5$. The methods are described in Table 5.3. The sample size for X and Y is $n = 100$. Results represent the average of 1000 Monte Carlo simulations.

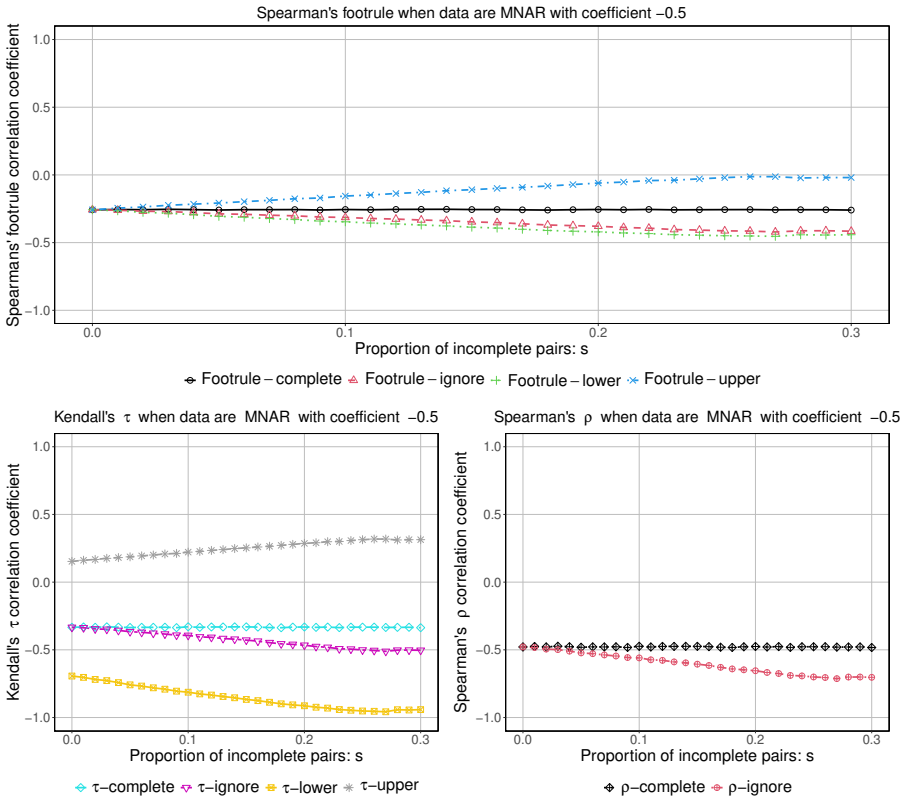


Figure E.25.: Rank correlation coefficients when data are missing not at random (MNAR). The subset of indices T containing all components with at least one missing value is selected according to Equation (5.4). For each $i \in T$, there is probability $1/3$ such that $X(i)$ is missing but $Y(i)$ is observed, probability $1/3$ such that $Y(i)$ is missing but $X(i)$ is observed, and probability $1/3$ such that both $X(i)$ and $Y(i)$ are missing. X, Y are generated such that $(X, Y) \overset{\text{i.i.d.}}{\sim} \mathrm{N}(0, \Sigma)$, where $\Sigma = \begin{pmatrix} 1 & \gamma \\ \gamma & 1 \end{pmatrix}$, with covariance coefficient $\gamma = -0.5$. The methods are described in Table 5.3. The sample size for X and Y is $n = 100$. Results represent the average of 1000 Monte Carlo simulations.

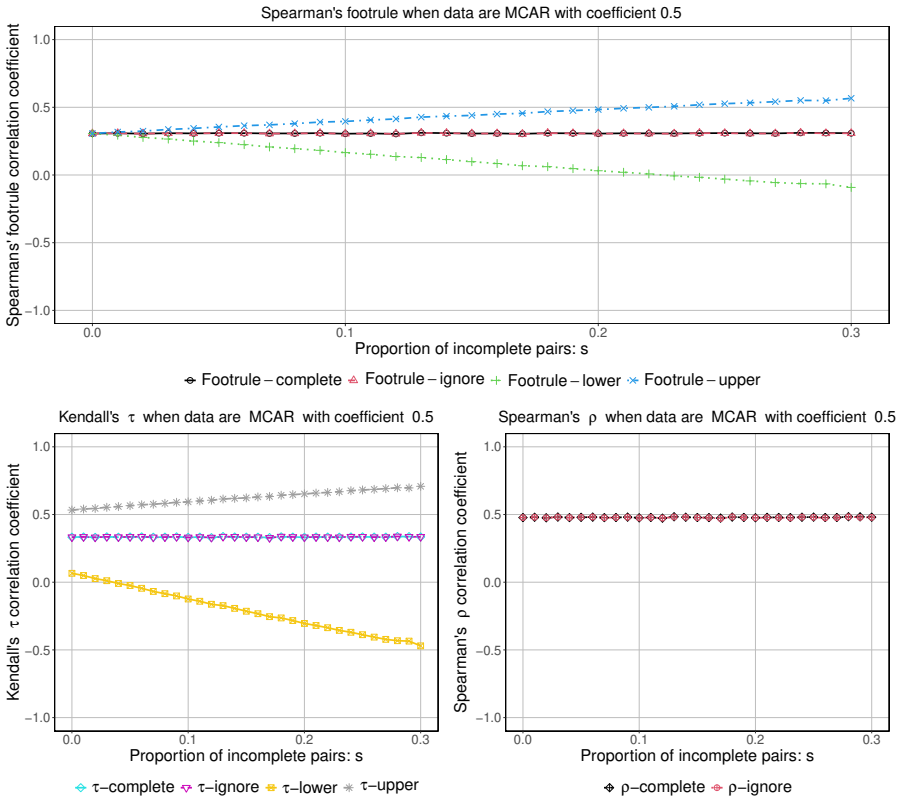


Figure E.26.: Rank correlation coefficients when data are missing completely at random (MCAR). The subset of indices T containing all components with at least one missing value is selected randomly. For each $i \in T$, there is probability $1/3$ such that $X(i)$ is missing but $Y(i)$ is observed, probability $1/3$ such that $Y(i)$ is missing but $X(i)$ is observed, and probability $1/3$ such that both $X(i)$ and $Y(i)$ are missing. X, Y are generated such that $(X, Y) \overset{\text{i.i.d.}}{\sim} \mathrm{N}(0, \Sigma)$, where $\Sigma = \begin{pmatrix} 1 & \gamma \\ \gamma & 1 \end{pmatrix}$, with covariance coefficient $\gamma = 0.5$. The methods are described in Table 5.3. The sample size for X and Y is $n = 100$. Results represent the average of 1000 Monte Carlo simulations.

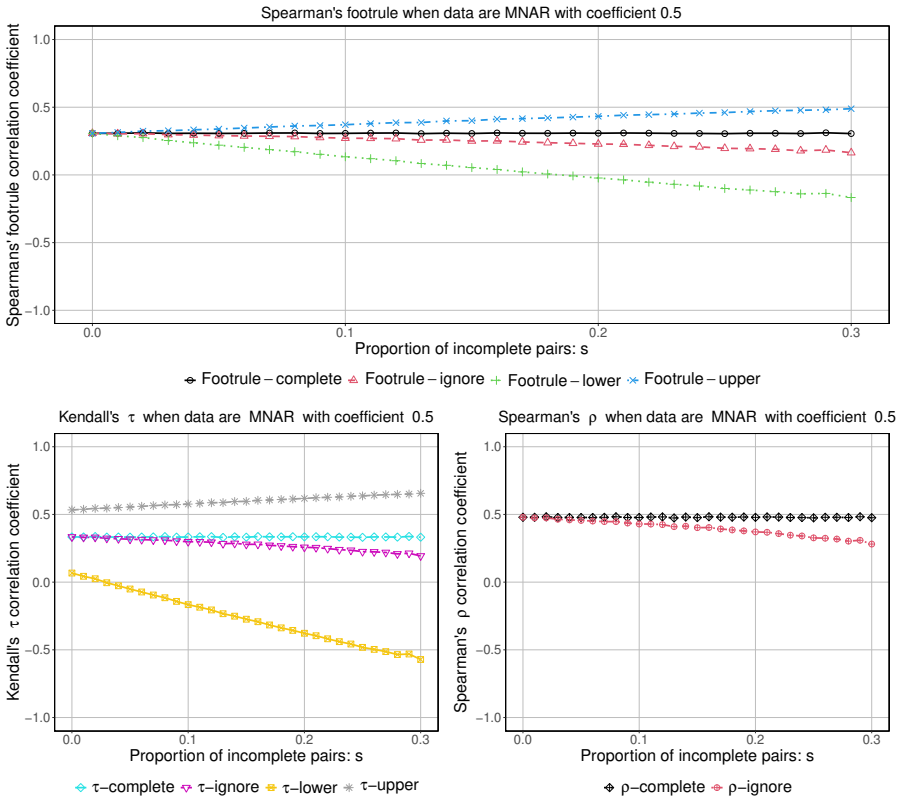


Figure E.27.: Rank correlation coefficients when data are missing not at random (MNAR). The subset of indices T containing all components with at least one missing value is selected according to Equation (5.4). For each $i \in T$, there is probability $1/3$ such that $X(i)$ is missing but $Y(i)$ is observed, probability $1/3$ such that $Y(i)$ is missing but $X(i)$ is observed, and probability $1/3$ such that both $X(i)$ and $Y(i)$ are missing. X, Y are generated such that $(X, Y) \overset{\text{i.i.d.}}{\sim} \mathrm{N}(0, \Sigma)$, where $\Sigma = \begin{pmatrix} 1 & \gamma \\ \gamma & 1 \end{pmatrix}$, with covariance coefficient $\gamma = 0.5$. The methods are described in Table 5.3. The sample size for X and Y is $n = 100$. Results represent the average of 1000 Monte Carlo simulations.

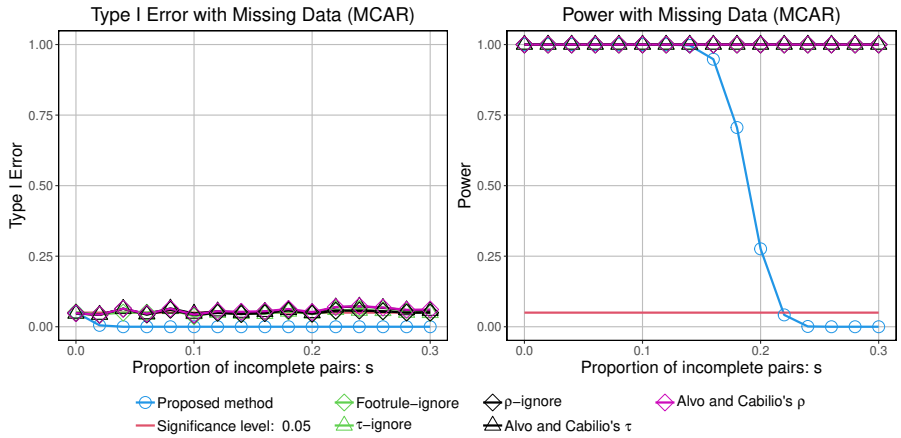


Figure E.28.: Statistical Type I error and power of the proposed method, methods that ignore missing data or use complete data, and Alvo and Cabilio's ρ and Alvo and Cabilio's τ methods, as the proportion of missing data s increases, where $s \in \{0.00, 0.02, \ldots, 0.30\}$. These methods are described in Table 5.4. The data is missing completely at random (MCAR). The subset of indices T containing all components with at least one missing value is selected randomly. For each $i \in T$, there is probability $1/3$ such that $X(i)$ is missing but $Y(i)$ is observed, probability $1/3$ such that $Y(i)$ is missing but $X(i)$ is observed, and probability $1/3$ such that both $X(i)$ and $Y(i)$ are missing. (Left) Type I error: $(X, Y) \overset{\text{i.i.d.}}{\sim} \mathrm{N}(0, I_2)$; (Right) Power: $(X, Y) \overset{\text{i.i.d.}}{\sim} \mathrm{N}(0, \Sigma)$, where $\Sigma = \begin{pmatrix} 1 & \gamma \\ \gamma & 1 \end{pmatrix}$, with covariance coefficient $\gamma = 0.5$. For both figures, a significance level $\alpha = 0.05$ is used and the sample size $n = 500$. The results in the figures are average of 1000 trials.

E.3.2. As the proportion of missing data varies

We perform simulations following the same approach as in Section 3.2.1 in the main paper when the proportion of missing data s increases, with different correlation coefficients γ, sample size n, or significance level α. The results are shown in Figure E.28–E.38.

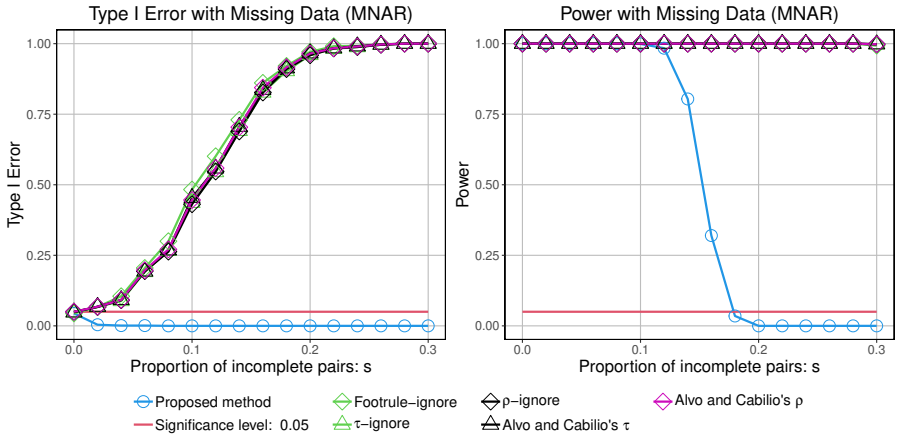


Figure E.29.: Statistical Type I error and power of the proposed method, methods that ignore missing data or use complete data, and Alvo and Cabilio's ρ and Alvo and Cabilio's τ methods, as the proportion of missing data s increases, where $s \in \{0.00, 0.02, \ldots, 0.30\}$. These methods are described in Table 5.4. The data is missing not at random (MNAR). The subset of indices T containing all components with at least one missing value is selected according to Equation (5.4). For each $i \in T$, there is probability $1/3$ such that $X(i)$ is missing but $Y(i)$ is observed, probability $1/3$ such that $Y(i)$ is missing but $X(i)$ is observed, and probability $1/3$ such that both $X(i)$ and $Y(i)$ are missing. (Left) Type I error: $(X, Y) \overset{\text{i.i.d.}}{\sim} \mathrm{N}(0, I_2)$; (Right) Power: $(X, Y) \overset{\text{i.i.d.}}{\sim} \mathrm{N}(0, \Sigma)$, where $\Sigma = \begin{pmatrix} 1 & \gamma \\ \gamma & 1 \end{pmatrix}$, with covariance coefficient $\gamma = 0.5$. For both figures, a significance level $\alpha = 0.05$ is used and the sample size $n = 500$. The results in the figures are average of 1000 trials.

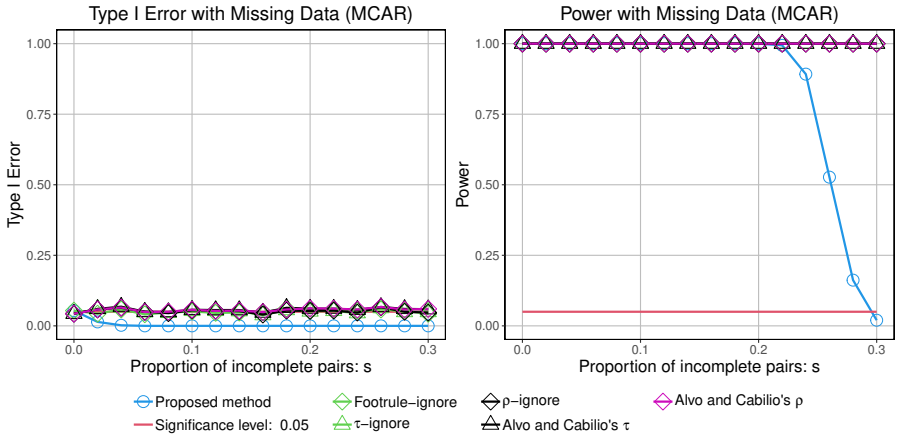


Figure E.30.: Statistical Type I error and power of the proposed method, methods that ignore missing data or use complete data, and Alvo and Cabilio's ρ and Alvo and Cabilio's τ methods, as the proportion of missing data s increases, where $s \in \{0.00, 0.02, \ldots, 0.30\}$. These methods are described in Table 5.4. The data is missing completely at random (MCAR). The subset of indices T containing all components with at least one missing value is selected randomly. For each $i \in T$, there is probability $1/3$ such that $X(i)$ is missing but $Y(i)$ is observed, probability $1/3$ such that $Y(i)$ is missing but $X(i)$ is observed, and probability $1/3$ such that both $X(i)$ and $Y(i)$ are missing. (Left) Type I error: $(X, Y) \overset{\text{i.i.d.}}{\sim} \mathrm{N}(0, I_2)$; (Right) Power: $(X, Y) \overset{\text{i.i.d.}}{\sim} \mathrm{N}(0, \Sigma)$, where $\Sigma = \begin{pmatrix} 1 & \gamma \\ \gamma & 1 \end{pmatrix}$, with covariance coefficient $\gamma = 0.8$. For both figures, a significance level $\alpha = 0.05$ is used and the sample size $n = 200$. The results in the figures are average of 1000 trials.

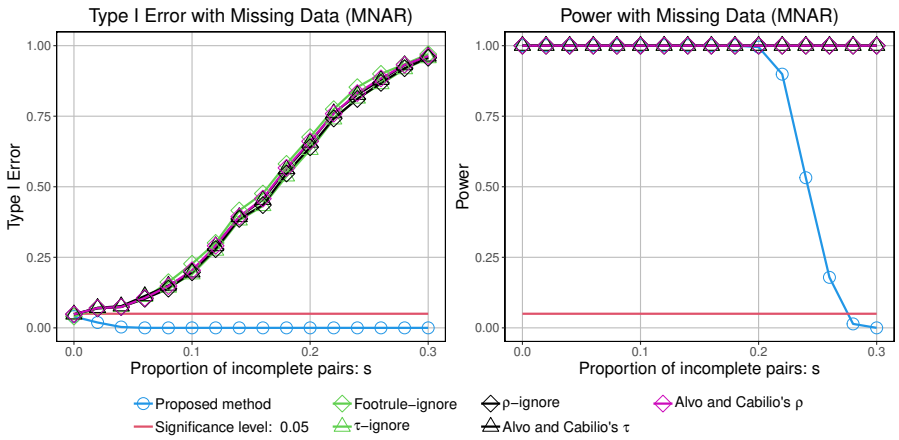


Figure E.31.: Statistical Type I error and power of the proposed method, methods that ignore missing data or use complete data, and Alvo and Cabilio's ρ and Alvo and Cabilio's τ methods, as the proportion of missing data s increases, where $s \in \{0.00, 0.02, \ldots, 0.30\}$. These methods are described in Table 5.4. The data is missing not at random (MNAR). The subset of indices T containing all components with at least one missing value is selected according to Equation (5.4). For each $i \in T$, there is probability $1/3$ such that $X(i)$ is missing but $Y(i)$ is observed, probability $1/3$ such that $Y(i)$ is missing but $X(i)$ is observed, and probability $1/3$ such that both $X(i)$ and $Y(i)$ are missing. (Left) Type I error: $(X, Y) \overset{\text{i.i.d.}}{\sim} \mathrm{N}(0, I_2)$; (Right) Power: $(X, Y) \overset{\text{i.i.d.}}{\sim} \mathrm{N}(0, \Sigma)$, where $\Sigma = \begin{pmatrix} 1 & \gamma \\ \gamma & 1 \end{pmatrix}$, with covariance coefficient $\gamma = 0.8$. For both figures, a significance level $\alpha = 0.05$ is used and the sample size $n = 200$. The results in the figures are average of 1000 trials.

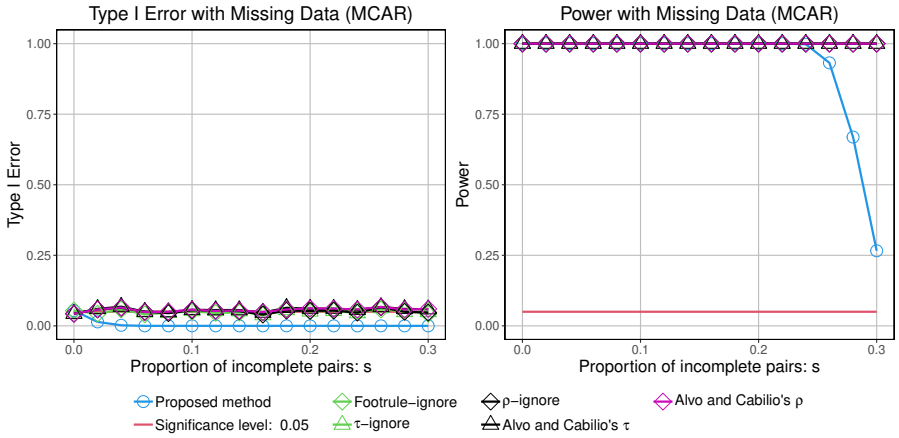


Figure E.32.: Statistical Type I error and power of the proposed method, methods that ignore missing data or use complete data, and Alvo and Cabilio's ρ and Alvo and Cabilio's τ methods, as the proportion of missing data s increases, where $s \in \{0.00, 0.02, \ldots, 0.30\}$. These methods are described in Table 5.4. The data is missing completely at random (MCAR). The subset of indices T containing all components with at least one missing value is selected randomly. For each $i \in T$, there is probability $1/3$ such that $X(i)$ is missing but $Y(i)$ is observed, probability $1/3$ such that $Y(i)$ is missing but $X(i)$ is observed, and probability $1/3$ such that both $X(i)$ and $Y(i)$ are missing. (Left) Type I error: $(X, Y) \overset{\text{i.i.d.}}{\sim} \mathrm{N}(0, I_2)$; (Right) Power: $(X, Y) \overset{\text{i.i.d.}}{\sim} \mathrm{N}(0, \Sigma)$, where $\Sigma = \begin{pmatrix} 1 & \gamma \\ \gamma & 1 \end{pmatrix}$, with covariance coefficient $\gamma = 0.8$. For both figures, a significance level $\alpha = 0.05$ is used and the sample size $n = 200$. The results in the figures are average of 1000 trials.

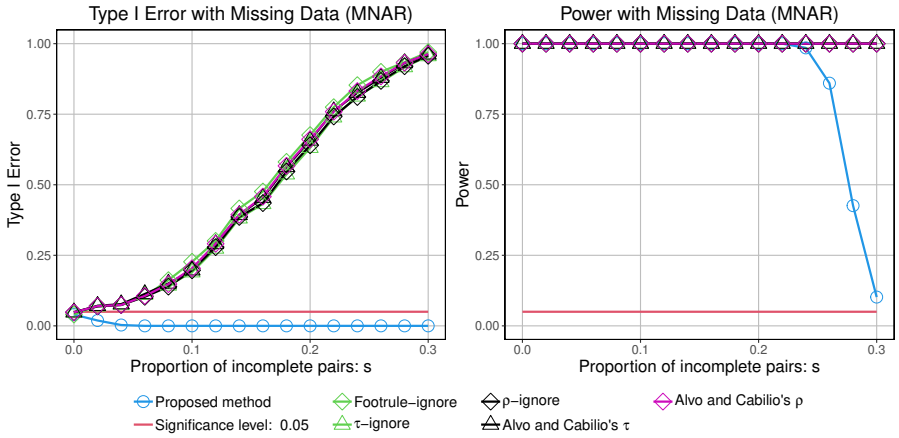


Figure E.33.: Statistical Type I error and power of the proposed method, methods that ignore missing data or use complete data, and Alvo and Cabilio's ρ and Alvo and Cabilio's τ methods, as the proportion of missing data s increases, where $s \in \{0.00, 0.02, \ldots, 0.30\}$. These methods are described in Table 5.4. The data is missing not at random (MNAR). The subset of indices T containing all components with at least one missing value is selected according to Equation (5.4). For each $i \in T$, there is probability $1/3$ such that $X(i)$ is missing but $Y(i)$ is observed, probability $1/3$ such that $Y(i)$ is missing but $X(i)$ is observed, and probability $1/3$ such that both $X(i)$ and $Y(i)$ are missing. (Left) Type I error: $(X, Y) \overset{\text{i.i.d.}}{\sim} \mathrm{N}(0, I_2)$; (Right) Power: $(X, Y) \overset{\text{i.i.d.}}{\sim} \mathrm{N}(0, \Sigma)$, where $\Sigma = \begin{pmatrix} 1 & \gamma \\ \gamma & 1 \end{pmatrix}$, with covariance coefficient $\gamma = 0.9$. For both figures, a significance level $\alpha = 0.05$ is used and the sample size $n = 200$. The results in the figures are average of 1000 trials.

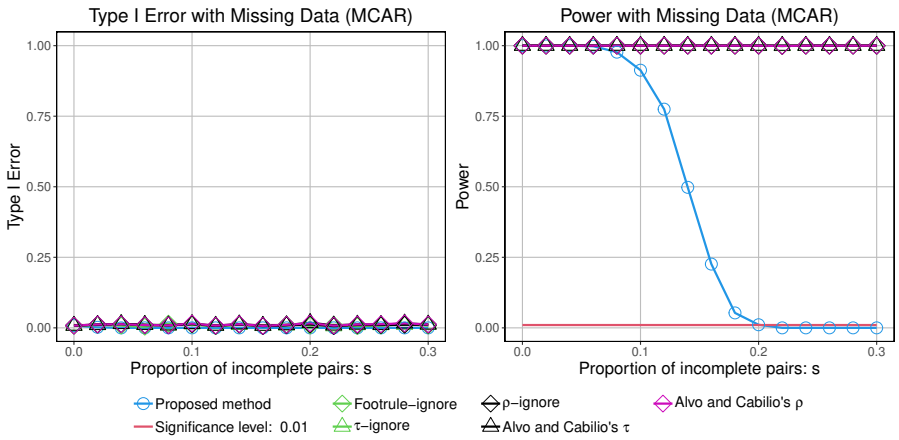


Figure E.34.: Statistical Type I error and power of the proposed method, methods that ignore missing data or use complete data, and Alvo and Cabilio's ρ and Alvo and Cabilio's τ methods, as the proportion of missing data s increases, where $s \in \{0.00, 0.02, \ldots, 0.30\}$. These methods are described in Table 5.4. The data is missing completely at random (MCAR). The subset of indices T containing all components with at least one missing value is selected randomly. For each $i \in T$, there is probability $1/3$ such that $X(i)$ is missing but $Y(i)$ is observed, probability $1/3$ such that $Y(i)$ is missing but $X(i)$ is observed, and probability $1/3$ such that both $X(i)$ and $Y(i)$ are missing. (Left) Type I error: $(X, Y) \overset{\text{i.i.d.}}{\sim} \mathrm{N}(0, I_2)$; (Right) Power: $(X, Y) \overset{\text{i.i.d.}}{\sim} \mathrm{N}(0, \Sigma)$, where $\Sigma = \begin{pmatrix} 1 & \gamma \\ \gamma & 1 \end{pmatrix}$, with covariance coefficient $\gamma = 0.5$. For both figures, a significance level $\alpha = 0.05$ is used and the sample size $n = 200$. The results in the figures are average of 1000 trials.

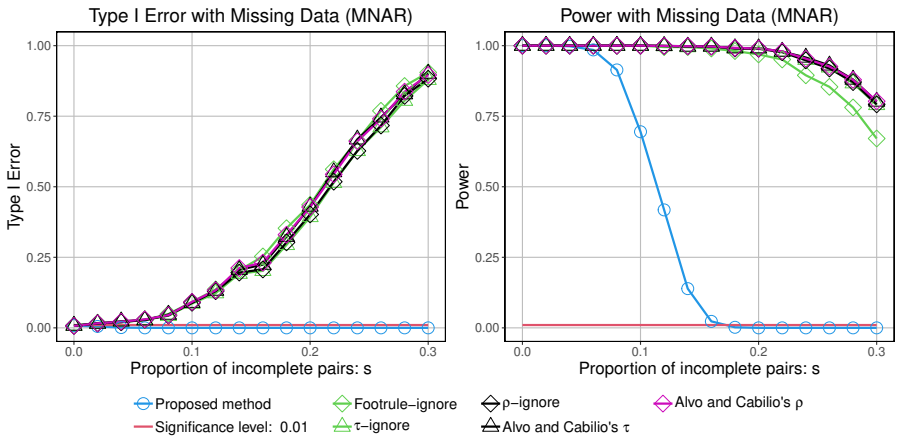


Figure E.35.: Statistical Type I error and power of the proposed method, methods that ignore missing data or use complete data, and Alvo and Cabilio's ρ and Alvo and Cabilio's τ methods, as the proportion of missing data s increases, where $s \in \{0.00, 0.02, \ldots, 0.30\}$. These methods are described in Table 5.4. The data is missing not at random (MNAR). The subset of indices T containing all components with at least one missing value is selected according to Equation (5.4). For each $i \in T$, there is probability $1/3$ such that $X(i)$ is missing but $Y(i)$ is observed, probability $1/3$ such that $Y(i)$ is missing but $X(i)$ is observed, and probability $1/3$ such that both $X(i)$ and $Y(i)$ are missing. (Left) Type I error: $(X, Y) \overset{\text{i.i.d.}}{\sim} \mathrm{N}(0, I_2)$; (Right) Power: $(X, Y) \overset{\text{i.i.d.}}{\sim} \mathrm{N}(0, \Sigma)$, where $\Sigma = \begin{pmatrix} 1 & \gamma \\ \gamma & 1 \end{pmatrix}$, with covariance coefficient $\gamma = 0.5$. For both figures, a significance level $\alpha = 0.05$ is used and the sample size $n = 200$. The results in the figures are average of 1000 trials.

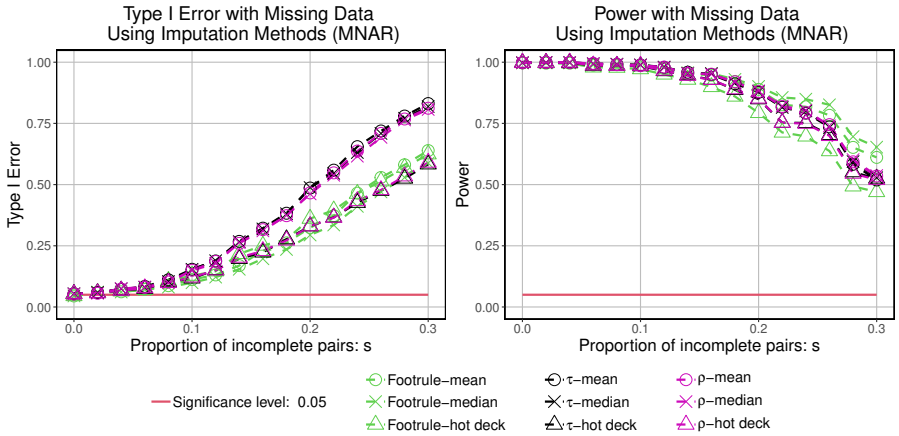


Figure E.36.: Statistical Type I error and power of imputation methods as the proportion of missing data s increases, where $s \in \{0.00, 0.02, \ldots, 0.30\}$. The imputation methods impute missing data using either mean, median or randomly selected (hot deck) values of observed data. These methods are described in Table 5.4. The data is missing not at random (MNAR). The subset of indices T containing all components with at least one missing value is selected according to Equation (5.4). For each $i \in T$, there is probability $1/3$ such that $X(i)$ is missing but $Y(i)$ is observed, probability $1/3$ such that $Y(i)$ is missing but $X(i)$ is observed, and probability $1/3$ such that both $X(i)$ and $Y(i)$ are missing. (Left) Type I error: $(X, Y) \overset{\text{i.i.d.}}{\sim} \mathrm{N}(0, I_2)$; (Right) Power: $(X, Y) \overset{\text{i.i.d.}}{\sim} \mathrm{N}(0, \Sigma)$, where $\Sigma = \begin{pmatrix} 1 & \gamma \\ \gamma & 1 \end{pmatrix}$, with covariance coefficient $\gamma = 0.5$. For both figures, a significance level $\alpha = 0.05$ is used and the sample size $n = 100$. The results in the figures are average of 1000 trials.

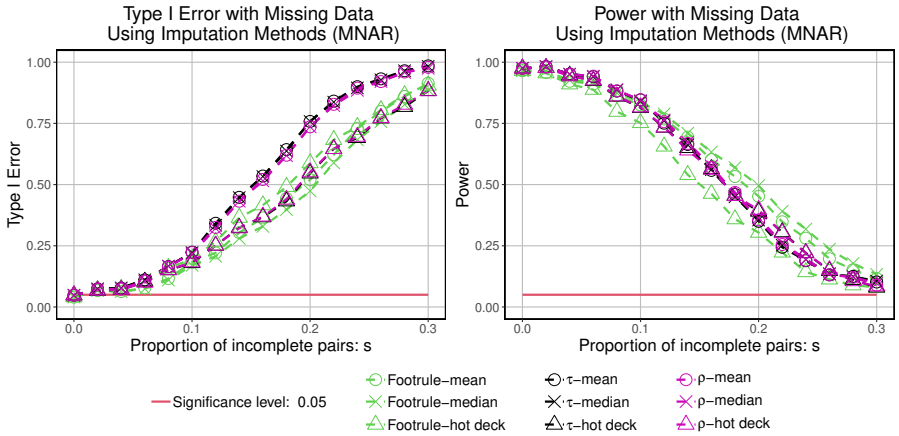


Figure E.37.: Statistical Type I error and power of imputation methods as the proportion of missing data s increases, where $s \in \{0.00, 0.02, \ldots, 0.30\}$. The imputation methods impute missing data using either mean, median or randomly selected (hot deck) values of observed data. These methods are described in Table 5.4. The data is missing not at random (MNAR). The subset of indices T containing all components with at least one missing value is selected according to Equation (5.4). For each $i \in T$, there is probability $1/3$ such that $X(i)$ is missing but $Y(i)$ is observed, probability $1/3$ such that $Y(i)$ is missing but $X(i)$ is observed, and probability $1/3$ such that both $X(i)$ and $Y(i)$ are missing. (Left) Type I error: $(X, Y) \overset{\text{i.i.d.}}{\sim} \mathrm{N}(0, I_2)$; (Right) Power: $(X, Y) \overset{\text{i.i.d.}}{\sim} \mathrm{N}(0, \Sigma)$, where $\Sigma = \begin{pmatrix} 1 & \gamma \\ \gamma & 1 \end{pmatrix}$, with covariance coefficient $\gamma = 0.3$. For both figures, a significance level $\alpha = 0.05$ is used and the sample size $n = 200$. The results in the figures are average of 1000 trials.

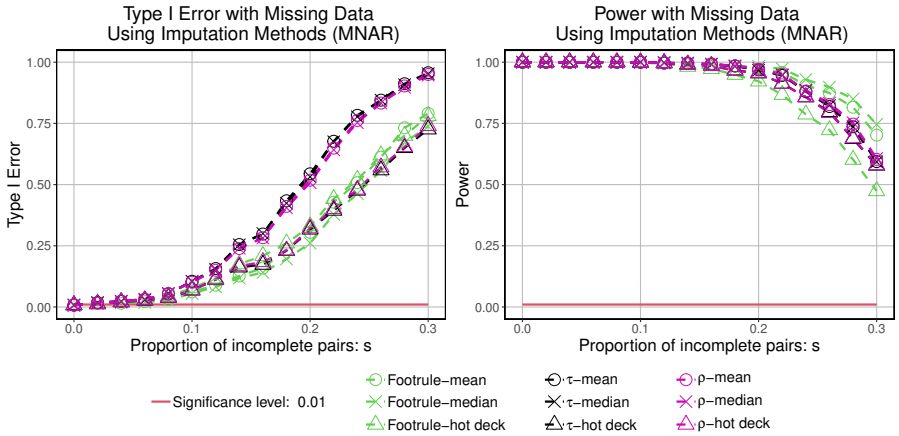


Figure E.38.: Statistical Type I error and power of imputation methods as the proportion of missing data s increases, where $s \in \{0.00, 0.02, \ldots, 0.30\}$. The imputation methods impute missing data using either mean, median or randomly selected (hot deck) values of observed data. These methods are described in Table 5.4. The data is missing not at random (MNAR). The subset of indices T containing all components with at least one missing value is selected according to Equation (5.4). For each $i \in T$, there is probability $1/3$ such that $X(i)$ is missing but $Y(i)$ is observed, probability $1/3$ such that $Y(i)$ is missing but $X(i)$ is observed, and probability $1/3$ such that both $X(i)$ and $Y(i)$ are missing. (Left) Type I error: $(X, Y) \overset{\text{i.i.d.}}{\sim} \mathrm{N}(0, I_2)$; (Right) Power: $(X, Y) \overset{\text{i.i.d.}}{\sim} \mathrm{N}(0, \Sigma)$, where $\Sigma = \begin{pmatrix} 1 & \gamma \\ \gamma & 1 \end{pmatrix}$, with covariance coefficient $\gamma = 0.5$. For both figures, a significance level $\alpha = 0.01$ is used and the sample size $n = 200$. The results in the figures are average of 1000 trials.

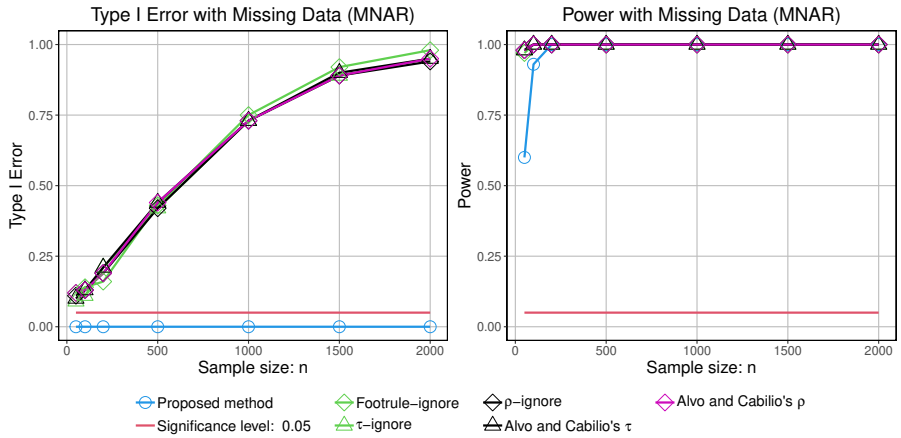


Figure E.39.: Statistical Type I error and power of the proposed method, methods that ignore missing data or use complete data, and Alvo and Cabilio's ρ and Alvo and Cabilio's τ methods, as the sample size n increases, where $n \in \{40, 70, 100, 200, 300, 500\}$. These methods are described in Table 5.4. The data is missing not at random (MNAR). The subset of indices T containing all components with at least one missing value is selected according to Equation (5.4). For each $i \in T$, there is probability $1/3$ such that $X(i)$ is missing but $Y(i)$ is observed, probability $1/3$ such that $Y(i)$ is missing but $X(i)$ is observed, and probability $1/3$ such that both $X(i)$ and $Y(i)$ are missing. (Left) Type I error: $(X, Y) \overset{\text{i.i.d.}}{\sim} \mathrm{N}(0, I_2)$; (Right) Power: $(X, Y) \overset{\text{i.i.d.}}{\sim} \mathrm{N}(0, \Sigma)$, where $\Sigma = \begin{pmatrix} 1 & \gamma \\ \gamma & 1 \end{pmatrix}$, with covariance coefficient $\gamma = 0.6$. For both figures, a significance level $\alpha = 0.05$ is used and the proportion of missing pairs $s = 0.1$. The results in the figures are average of 1000 trials.

E.3.3. As the sample size varies

We perform simulations following the same approach as described in Section 3.2.2 in the main paper when the sample size n increases, with potentially different correlation coefficients γ, proportion of missing data s, and significance level α. The results are shown in Figure E.39 – E.43.

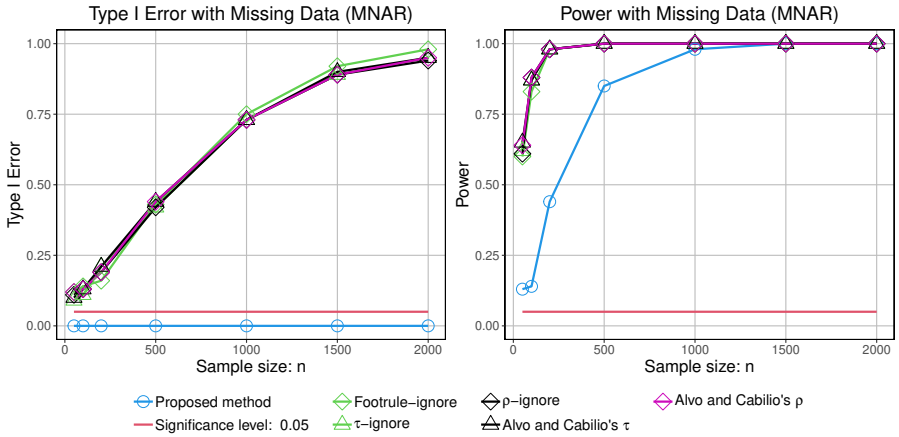


Figure E.40.: Statistical Type I error and power of the proposed method, methods that ignore missing data or use complete data, and Alvo and Cabilio's ρ and Alvo and Cabilio's τ methods, as the sample size n increases, where $n \in \{40, 70, 100, 200, 300, 500\}$. These methods are described in Table 5.4. The data is missing not at random (MNAR). The subset of indices T containing all components with at least one missing value is selected according to Equation (5.4). For each $i \in T$, there is probability $1/3$ such that $X(i)$ is missing but $Y(i)$ is observed, probability $1/3$ such that $Y(i)$ is missing but $X(i)$ is observed, and probability $1/3$ such that both $X(i)$ and $Y(i)$ are missing. (Left) Type I error: $(X, Y) \overset{\text{i.i.d.}}{\sim} \mathrm{N}(0, I_2)$; (Right) Power: $(X, Y) \overset{\text{i.i.d.}}{\sim} \mathrm{N}(0, \Sigma)$, where $\Sigma = \begin{pmatrix} 1 & \gamma \\ \gamma & 1 \end{pmatrix}$, with covariance coefficient $\gamma = 0.4$. For both figures, a significance level $\alpha = 0.05$ is used and the proportion of missing pairs $s = 0.1$. The results in the figures are average of 1000 trials.

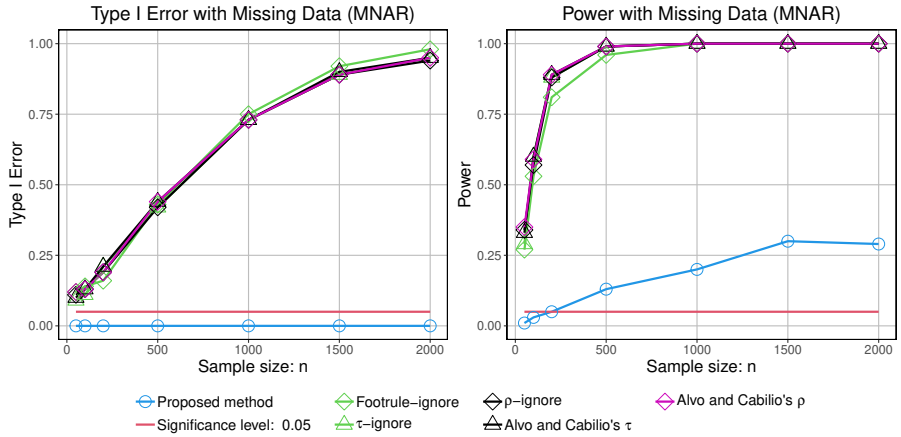


Figure E.41.: Statistical Type I error and power of the proposed method, methods that ignore missing data or use complete data, and Alvo and Cabilio's ρ and Alvo and Cabilio's τ methods, as the sample size n increases, where $n \in \{40, 70, 100, 200, 300, 500\}$. These methods are described in Table 5.4. The data is missing not at random (MNAR). The subset of indices T containing all components with at least one missing value is selected according to Equation (5.4). For each $i \in T$, there is probability $1/3$ such that $X(i)$ is missing but $Y(i)$ is observed, probability $1/3$ such that $Y(i)$ is missing but $X(i)$ is observed, and probability $1/3$ such that both $X(i)$ and $Y(i)$ are missing. (Left) Type I error: $(X, Y) \overset{\text{i.i.d.}}{\sim} \mathrm{N}(0, I_2)$; (Right) Power: $(X, Y) \overset{\text{i.i.d.}}{\sim} \mathrm{N}(0, \Sigma)$, where $\Sigma = \begin{pmatrix} 1 & \gamma \\ \gamma & 1 \end{pmatrix}$, with covariance coefficient $\gamma = 0.3$. For both figures, a significance level $\alpha = 0.05$ is used and the proportion of missing pairs $s = 0.1$. The results in the figures are average of 1000 trials.

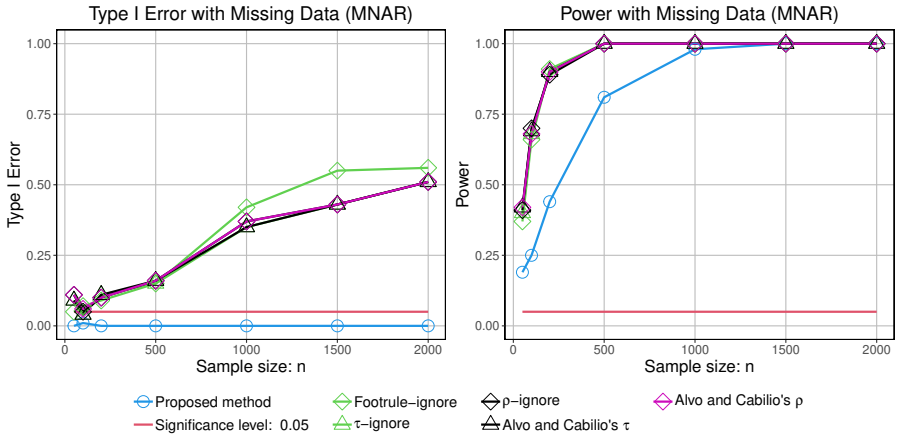


Figure E.42.: Statistical Type I error and power of the proposed method, methods that ignore missing data or use complete data, and Alvo and Cabilio's ρ and Alvo and Cabilio's τ methods, as the sample size n increases, where $n \in \{40, 70, 100, 200, 300, 500\}$. These methods are described in Table 5.4. The data is missing not at random (MNAR). The subset of indices T containing all components with at least one missing value is selected according to Equation (5.4). For each $i \in T$, there is probability $1/3$ such that $X(i)$ is missing but $Y(i)$ is observed, probability $1/3$ such that $Y(i)$ is missing but $X(i)$ is observed, and probability $1/3$ such that both $X(i)$ and $Y(i)$ are missing. (Left) Type I error: $(X, Y) \overset{\text{i.i.d.}}{\sim} \mathrm{N}(0, I_2)$; (Right) Power: $(X, Y) \overset{\text{i.i.d.}}{\sim} \mathrm{N}(0, \Sigma)$, where $\Sigma = \begin{pmatrix} 1 & \gamma \\ \gamma & 1 \end{pmatrix}$, with covariance coefficient $\gamma = 0.3$. For both figures, a significance level $\alpha = 0.05$ is used and the proportion of missing pairs $s = 0.06$. The results in the figures are average of 1000 trials.

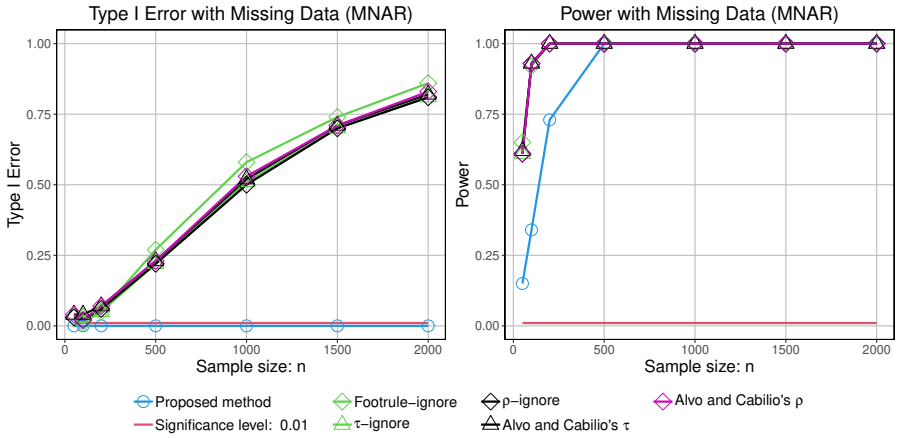


Figure E.43.: Statistical Type I error and power of the proposed method, methods that ignore missing data or use complete data, and Alvo and Cabilio's ρ and Alvo and Cabilio's τ methods, as the sample size n increases, where $n \in \{40, 70, 100, 200, 300, 500\}$. These methods are described in Table 5.4. The data is missing not at random (MNAR). The subset of indices T containing all components with at least one missing value is selected according to Equation (5.4). For each $i \in T$, there is probability $1/3$ such that $X(i)$ is missing but $Y(i)$ is observed, probability $1/3$ such that $Y(i)$ is missing but $X(i)$ is observed, and probability $1/3$ such that both $X(i)$ and $Y(i)$ are missing. (Left) Type I error: $(X, Y) \overset{\text{i.i.d.}}{\sim} \mathrm{N}(0, I_2)$; (Right) Power: $(X, Y) \overset{\text{i.i.d.}}{\sim} \mathrm{N}(0, \Sigma)$, where $\Sigma = \begin{pmatrix} 1 & \gamma \\ \gamma & 1 \end{pmatrix}$, with covariance coefficient $\gamma = 0.5$. For both figures, a significant level $\alpha = 0.01$ is used and the proportion of missing pairs $s = 0.1$. The results in the figures are average of 1000 trials.

E.3.4. A different missingness mechanism

Now we perform simulations following the same was as in Section 3.2 in the main paper when data are missing not at random (MNAR), with a different MNAR missingness mechanism described below.

The way that the index set $T \subset \{1, \ldots, n\}$ of the missing components is chosen depends on the values of X and Y. Let

$$q = \sum_{i=1}^{n} \mathrm{I}\left(|R(X(i), X) - R(Y(i), Y)| < n/2\right)$$

be the number of pairs of components in X and Y such that

$$|R(X(i), X) - R(Y(i), Y)| < n/2.$$

Each index $i \in [n]$ is selected to be in the set T of indices of missing components according to the following probability conditional on $C = |T|$, the size of the set T:

$$\begin{aligned} &p\left(i \in T \,|\, C = \lfloor s \cdot n \rfloor\right) \\ &= \begin{cases} \min\left\{1, sn/q\right\}, & \text{if } |R(X(i), X) - R(Y(i), Y)| < n/2, \\ \max\{0, (sn - q)/(n - q)\}, & \text{otherwise,} \end{cases} \end{aligned} \tag{E.5}$$

for any given s. The results are shown in Figure E.44–E.45.

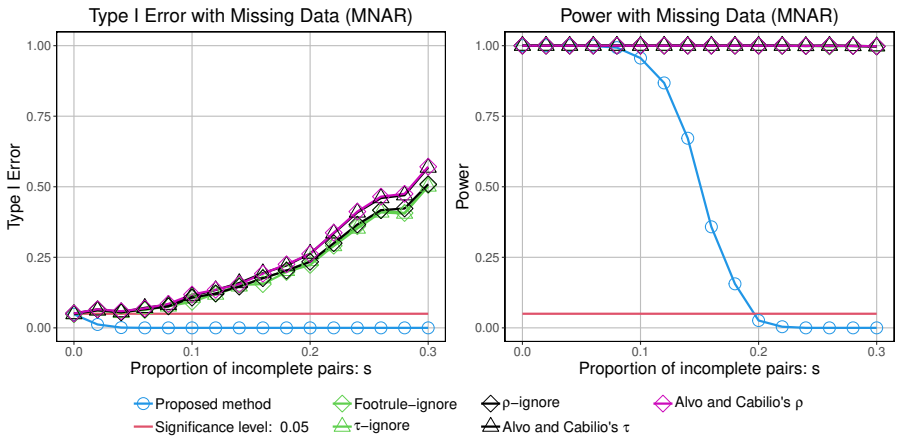


Figure E.44.: Statistical Type I error and power of the proposed method, methods that ignore missing data or use complete data, and Alvo and Cabilio's ρ and Alvo and Cabilio's τ methods, as the proportion of missing data s increases, where $s \in \{0.00, 0.02, \ldots, 0.30\}$. These methods are described in Table 5.4. The data is missing not at random (MNAR). The subset of indices T containing all components with at least one missing value is selected according to Equation (E.5). For each $i \in T$, there is probability $1/3$ such that $X(i)$ is missing but $Y(i)$ is observed, probability $1/3$ such that $Y(i)$ is missing but $X(i)$ is observed, and probability $1/3$ such that both $X(i)$ and $Y(i)$ are missing. (Left) Type I error: $(X, Y) \overset{\text{i.i.d.}}{\sim} \mathrm{N}(0, I_2)$; (Right) Power: $(X, Y) \overset{\text{i.i.d.}}{\sim} \mathrm{N}(0, \Sigma)$, where $\Sigma = \begin{pmatrix} 1 & \gamma \\ \gamma & 1 \end{pmatrix}$, with covariance coefficient $\gamma = 0.5$. For both figures, a significance level $\alpha = 0.05$ is used and the sample size $n = 200$. The results in the figures are average of 1000 trials.

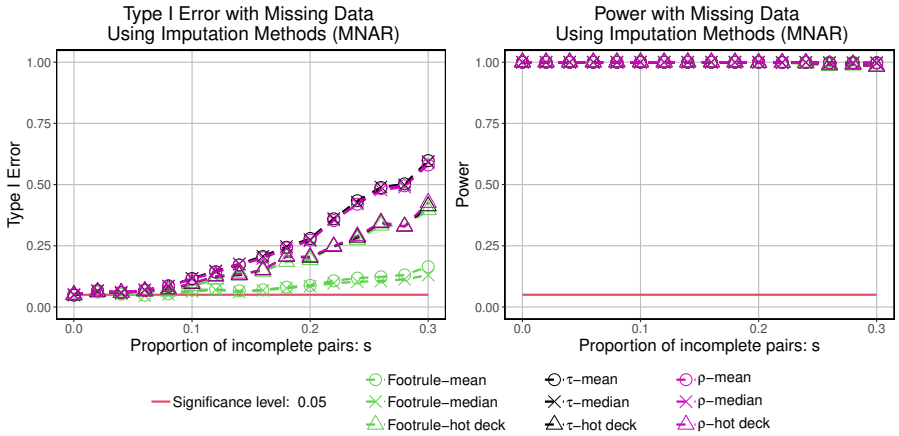


Figure E.45.: Statistical Type I error and power of imputation methods as the proportion of missing data s increases, where $s \in \{0.00, 0.02, \ldots, 0.30\}$. The imputation methods impute missing data using either mean, median or randomly selected (hot deck) values of observed data. These methods are described in Table 5.4. The data is missing completely at random (MNAR). The subset of indices T containing all components with at least one missing value is selected according to Equation (E.5). For each $i \in T$, there is probability $1/3$ such that $X(i)$ is missing but $Y(i)$ is observed, probability $1/3$ such that $Y(i)$ is missing but $X(i)$ is observed, and probability $1/3$ such that both $X(i)$ and $Y(i)$ are missing. (Left) Type I error: $(X, Y) \overset{\text{i.i.d.}}{\sim} \mathrm{N}(0, I_2)$; (Right) Power: $(X, Y) \overset{\text{i.i.d.}}{\sim} \mathrm{N}(0, \Sigma)$, where $\Sigma = \begin{pmatrix} 1 & \gamma \\ \gamma & 1 \end{pmatrix}$, with covariance coefficient $\gamma = 0.5$. For both figures, a significance level $\alpha = 0.05$ is used and the sample size $n = 200$. The results in the figures are average of 1000 trials.

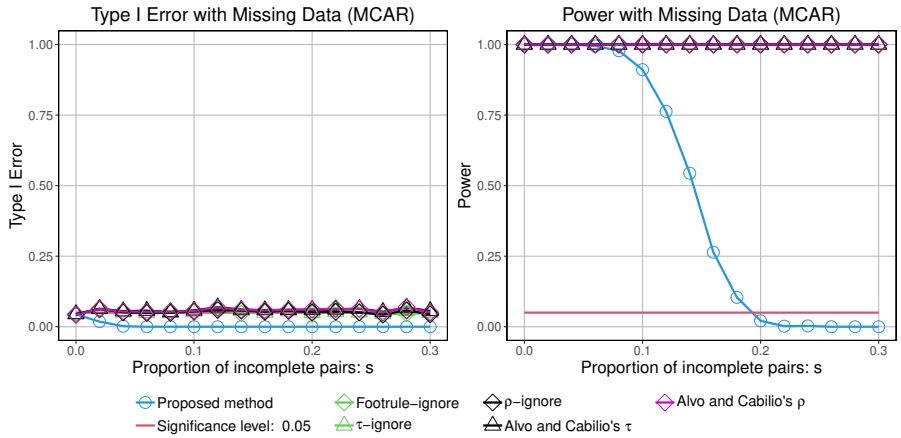


Figure E.46.: Statistical Type I error and power of the proposed method, methods that ignore missing data, and Alvo and Cabilio's ρ and Alvo and Cabilio's τ methods, as the proportion of missing data s increases, where $s \in \{0.00, 0.02, \ldots, 0.30\}$. These methods are described in Table 5.4. The data is missing completely at random (MCAR). The subset of indices T containing all components with at least one missing value is selected randomly. For each $i \in T$, there is probability $1/3$ such that $X(i)$ is missing but $Y(i)$ is observed, probability $1/3$ such that $Y(i)$ is missing but $X(i)$ is observed, and probability $1/3$ such that both $X(i)$ and $Y(i)$ are missing. (Left) Type I error: $(X, Y) \overset{\text{i.i.d.}}{\sim} Gamma(1, 1)$; (Right) Power: $Y(i) = X(i)/2 + \epsilon_i$, where $X_i, \epsilon_i \overset{\text{i.i.d.}}{\sim} Gamma(1, 1)$ for any $i \in \{1, \ldots, n\}$. For both figures, a significance level $\alpha = 0.05$ is used and the sample size $n = 200$. The results in the figures are average of 1000 trials.

E.3.5. Gamma-distributed data

In this section, we consider the cases when data are following Gamma distributions, rather than normal distributions. The proposed method, methods that ignore or impute the missing data show similar results as before. The results can be seen in Figure E.46 and E.47.

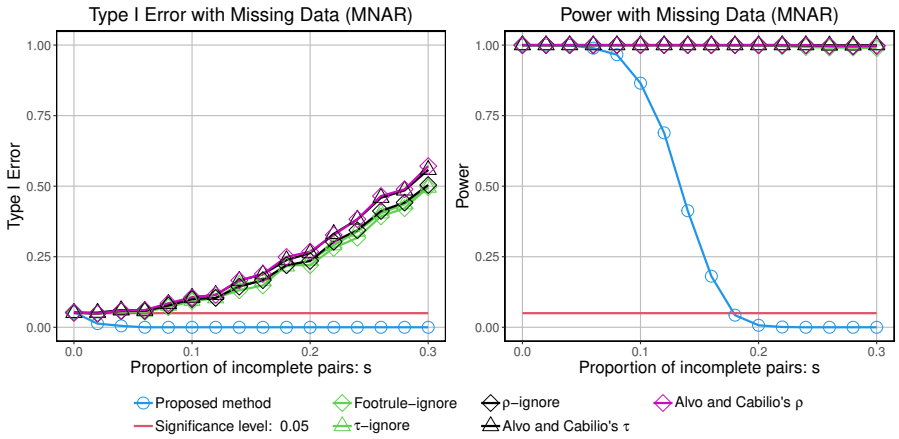


Figure E.47.: Statistical Type I error and power of the proposed method, methods that ignore missing data, and Alvo and Cabilio's ρ and Alvo and Cabilio's τ methods, as the proportion of missing data s increases, where $s \in \{0.00, 0.02, \ldots, 0.30\}$. These methods are described in Table 5.4. The data is missing not at random (MNAR). The subset of indices T containing all components with at least one missing value is selected according to Equation (E.5). For each $i \in T$, there is probability $1/3$ such that $X(i)$ is missing but $Y(i)$ is observed, probability $1/3$ such that $Y(i)$ is missing but $X(i)$ is observed, and probability $1/3$ such that both $X(i)$ and $Y(i)$ are missing. (Left) Type I error: $(X, Y) \overset{\text{i.i.d.}}{\sim} Gamma(1, 1)$; (Right) Power: $Y(i) = X(i)/2 + \epsilon_i$, where $X_i, \epsilon_i \overset{\text{i.i.d.}}{\sim} Gamma(1, 1)$ for any $i \in \{1, \ldots, n\}$. For both figures, a significance level $\alpha = 0.05$ is used and the sample size $n = 200$. The results in the figures are average of 1000 trials.

Bibliography

D. W. Alling. Early Decision in the Wilcoxon Two-Sample Test. *Journal of the American Statistical Association*, 58(303):713–720, 1963.

M. Alvo and P. Cabilio. Rank Correlation Methods for Missing Data. *The Canadian Journal of Statistics*, 23:345–358, 1995.

M. Alvo and M. Charbonneau. The use of spearman's footrule in testing for trend when the data are incomplete. *Communications in Statistics-Simulation and Computation*, 26(1):193–213, 1997.

I. L. Amerise and A. Tarsitano. Correction methods for ties in rank correlations. *Journal of Applied Statistics*, 42(12):2584–2596, 2015.

A. R. Ansari and R. A. Bradley. Rank-Sum Tests for Dispersions. *The Annals of Mathematical Statistics*, pages 1174–1189, 1960.

A. Azzalini. A Class of Distributions Which Includes the Normal Ones. *Scandinavian Journal of Statistics*, pages 171–178, 1985.

G. L. Bakris, R. Agarwal, J. C. Chan, M. E. Cooper, R. T. Gansevoort, H. Haller, G. Remuzzi, P. Rossing, R. E. Schmieder, Christina C. Nowack, K. Peter, J. Amer, P. Alexander, K. Nina, M. R. Luis, and Mineralocorticoid Receptor Antagonist Tolerability Study–Diabetic Nephropathy (ARTS-DN) Study Group. Effect of Finerenone on Albuminuria in Patients With Diabetic Nephropathy: A Randomized Clinical Trial. *Journal of the American Medical Association*, 61 (2):524–531, 2015.

A. N. Baraldi and C. K. Enders. An introduction to modern missing data analyses. *Journal of School Psychology*, 48(1):5–37, 2010.

Bayer. A randomized, double-blind, placebo-controlled, multi-center study to assess the safety and efficacy of different oral doses of BAY94-8862 in subjects with type 2 diabetes mellitus and the clinical diagnosis of diabetic nephropathy [dataset]. Vivli. https://doi.org/10.25934/PR00008549, 2022.

D. A. Bennett. How can I deal with missing data in my study? *Australian and New Zealand Journal of Public Health*, 25:464–469, 2001.

T. B. Berrett. Nonparametric tests of missing completely at random. *Statistical Machine Learning for High Dimensional*, page 9, 2024.

K. J. Berry and J. P. W. Mielke. Spearman's Footrule as a Measure of Agreement. *Psychological Reports*, 80(3):839–846, 1997.

A. Bordino and T. B. Berrett. Tests of missing completely at random based on sample covariance matrices. *The Annals of Statistics*, 53(5):2204–2229, 2025.

G. W. Brown and A. M. Mood. On median tests for linear hypotheses. In *Proceedings of the Second Berkeley Symposium on Mathematical Statistics and Probability*, volume 2, pages 159–167, 1951.

J. Y. Campbell, A. W. Lo, A. C. MacKinlay, and R. F. Whitelaw. The Econometrics of Financial Markets. *Macroeconomic Dynamics*, 2(4):559–562, 1998.

C. Chen, W. Xu, W. Zhang, H. Zhu, and J. Dai. Asymptotic properties of spearman's footrule and gini's gamma in bivariate normal model. *Journal of the Franklin Institute*, 360:9812–9843, 2023.

D. Chen, F. Song, and L. Feng. Rank Based Tests for High Dimensional White Noise. *arXiv preprint arXiv:2204.08402*, 2022a.

W. Chen, A. Tripp, and J. M. Hernández-Lobato. Meta-learning Adaptive Deep Kernel Gaussian Processes for Molecular Property Prediction. *arXiv preprint arXiv:2205.02708*, 2022b.

Y. K. Cheung. Exact two-sample inference with missing data. *Biometrics*, 61(2):524–531, 2005.

T. Cook and R. Zea. Missing data and sensitivity analysis for binary data with implications for sample size and power of randomized clinical trials. *Statistics in Medicine*, 39(2):192–204, 2020a.

T. Cook and R. Zea. Missing data and sensitivity analysis for binary data with implications for sample size and power of randomized clinical trials. *Statistics in Medicine*, 39(2):192–204, 2020b.

K. E. Corey, E. Kane, C. Munroe, L. L. Barlow, H. Zheng, and R. T. Chung. Hepatitis c virus infection and its clearance alter circulating lipids: implications for long-term follow-up. *Hepatology*, 50(4):1030–1037, 2009.

O. Cucconi. Un nuovo test non parametrico per il confronto fra due gruppi di valori campionari. *Giornale degli Economisti e Annali di Economia*, 27(3/4):225–248, 1968.

A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum Likelihood from Incomplete Data via the EM Algorithm. *Journal of the Royal Statistical Society: Series B (Methodological)*, 39(1): 1–22, 1977.

B. L. Van der Waerden. Order tests for the two-sample problem and their power. In *Indagationes Mathematicae (Proceedings)*, volume 55, pages 453–458, 1952.

P. Diaconis and R. L. Graham. Spearman's Footrule as a Measure of Disarray. *Journal of the Royal Statistical Society: Series B (Methodological)*, 39:262–268, 1977.

Y. Dong and C.-Y. J. Peng. Principled missing data methods for researchers. *SpringerPlus*, 2 (1):1–17, 2013.

C. Dwork, R. Kumar, M. Naor, and D. Sivakumar. Rank Aggregation Methods for the Web. In *Proceedings of the 10th international conference on World Wide Web*, pages 613–622, 2001.

R. Fagin, R. Kumar, and D. Sivakumar. Comparing Top k Lists. *SIAM Journal on Discrete Mathematics*, 17(1):134–160, 2003.

G. Fahoome and S. S. Sawilowsky. Review of Twenty Nonparametric Statistics and Their Large Sample Approximations. *Journal of Modern Applied Statistical Methods*, 1, 2000.

A. M. Feldman, M. R. Bristow, W. W. Parmley, P. E. Carson, C. J. Pepine, E. M. Gilbert, J. E. Strobeck, g. h. Hendrix, E. R. Powers, R. P. Bain, B.G. White, and the Vesnarinone Study Group. Effects of Vesnarinone on Morbidity and Mortality in Patients with Heart Failure. *New England Journal of Medicine*, 329(3):149–155, 1993.

R. A. Fisher. Statistical Methods for Research Workers. In *Breakthroughs in statistics: Methodology and distribution*, pages 66–70. Springer, 1970.

E. A. Gehan. A generalized two-sample Wilcoxon test for doubly censored data. *Biometrika*, 52 (3/4):650–653, 1965.

C. Genest, J. Nešlehová, and G. N. Ben. Spearman's footrule and gini's gamma: a review with complements. *Journal of Nonparametric Statistics*, 22:937–954, 2010.

S. B. Goldberg, D. M. Bolt, and R. J. Davidson. Data missing not at random in mobile health research: Assessment of the problem and a case for sensitivity analyses. *Journal of Medical Internet Research*, 23(6):e26749, 2021.

A. Gretton, K. M. Borgwardt, M. J. Rasch, B. Schölkopf, and A. Smola. A kernel two-sample test. *The Journal of Machine Learning Research*, 13(1):723–773, 2012.

D. Gyamfi, E. O. Awuah, and S. Owusu. Lipid Metabolism: An Overview. *The molecular nutrition of fats*, pages 17–32, 2019.

I. Hacking. *Logic of Statistical Inference*. Cambridge University Press, 2016.

M. Halperin and J. Ware. Early Decision in a Censored Wilcoxon Two-Sample Test for Accumulating Survival Data. *Journal of the American Statistical Association*, 69(346):414–422, 1974.

I. Harris, T. J. Osborn, P. Jones, and D. Lister. Version 4 of the cru ts monthly high-resolution gridded multivariate climate dataset. *Scientific Data*, 7(1):109, 2020.

M. W. Heymans and J. W. R. Twisk. Handling missing data in clinical research. *Journal of Clinical Epidemiology*, 151:185–188, 2022.

W. Hoeffding. A non-parametric test of independence. In *The Collected Works of Wassily Hoeffding*, pages 214–226. Springer, 1994.

S. Holm. A Simple Sequentially Rejective Multiple Test Procedure. *Scandinavian journal of statistics*, pages 65–70, 1979.

M. G. Kendall. *Rank correlation methods*. Griffn, London, 4 edition, 1948.

V. Khullar and R. J. Firpi. Hepatitis C cirrhosis: New perspectives for diagnosis and treatment. *World Journal of Hepatology*, 7(14):1843, 2015.

J. M. Lachin. Worst-Rank Score Analysis with Informatively Missing Observations in Clinical Trials. *Controlled Clinical Trials*, 20(5):408–422, 1999.

Y. J. Lee. A Two-Sample Nonparametric Test with Missing Observations. *American Journal of Mathematical and Management Sciences*, 17(1-2):187–200, 1997.

E. L. Lehmann and H. J. D'Abrera. *Nonparametrics: statistical methods based on ranks*. Springer, 1975.

E. L. Lehmann and J. P. Romano. *Testing Statistical Hypotheses*, volume 3. Springer, 1986.

Y. Lepage. A Combination of Wilcoxon's and Ansari-Bradley's Statistics. *Biometrika*, 58(1): 213–217, 1971.

R. Lichtinghagen, F. Klawonn, and G. F. Hoffmann. HCV data. UCI Machine Learning Repository, 2020. DOI: https://doi.org/10.24432/C5D612.

S. Lin. Rank aggregation methods. *Wiley Interdisciplinary Reviews: Computational Statistics*, 2 (5):555–570, 2010.

J. Liou, H. Mani, and J. Yen. Viral Hepatitis, Cholesterol Metabolism, and Cholesterol-Lowering Natural Compounds. *International Journal of Molecular Sciences*, 23(7):3897, 2022.

R. Little and L. Yau. Intent-to-treat analysis for longitudinal studies with drop-outs. *Biometrics*, pages 1324–1333, 1996.

R. J. A. Little. A test of missing completely at random for multivariate data with missing values. *Journal of the American statistical Association*, 83(404):1198–1202, 1988.

R. J. A. Little and D. B. Rubin. *Statistical Analysis with Missing Data*, volume 793. John Wiley & Sons, Hoboken, 3 edition, 2019.

S. Loukas and T. Papaioannou. Rank correlation inequalities with ties and missing data. *Statistics & Probability Letters*, 11:53–56, 1991.

C. P. Luigi and Y. Nikitin. Asymptotic efficiency of independence tests based on gini's rank association coefficient, spearman's footrule and their generalizations. *Communications in Statistics-Theory and Methods*, 28:453–465, 1999.

H. B. Mann and D. R. Whitney. On a Test of Whether one of Two Random Variables is Stochastically Larger than the Other. *The Annals of Mathematical Statistics*, pages 50–60, 1947.

K. B. Margulies, A. F. Hernandez, M. M. Redfield, M. M. Givertz, G. H. Oliveira, R. Cole, D. L. Mann, D. J. Whellan, M. S. Kiernan, G. M. Felker, S. E. McNulty, K. J. Anstrom, M. R. B. E.

Shah, and T. P. Cappola. Effects of Liraglutide on Clinical Stability Among Patients With Advanced Heart Failure and Reduced Ejection Fraction: A Randomized Clinical Trial. *Journal of the American Medical Association*, 316(5):500–508, 2016.

M. Marozzi. Nonparametric Simultaneous Tests for Location and Scale Testing: A Comparison of Several Methods. *Communications in Statistics-Simulation and Computation*, 42(6):1298–1317, 2013.

M. B. Mathur. The m-value: A simple sensitivity analysis for bias due to missing data in treatment effect estimates. *American Journal of Epidemiology*, 192(4):612–620, 2023.

R. A. Matsouaka and R. A. Betensky. Power and Sample Size Calculations for the Wilcoxon–Mann–Whitney Test in the Presence of Missing Observations due to Death. *Statisics in Medicine*, 34(3):406–431, 2015.

A. M. Mood. On the asymptotic efficiency of certain nonparametric two-sample tests. *The annals of mathematical statistics*, pages 514–522, 1954.

H. Murakami and M. Neuhäuser. A two-sample nonparametric test for one-sided location-scale alternative. *Journal of Applied Statistics*, 52(3):715–743, 2025.

M. Neuhäuser. An adaptive location-scale test. *Biometrical Journal*, 43(7):809–819, 2001.

J. Neyman and E. S. Pearson. On the problem of the most efficient tests of statistical hypotheses. *Philosophical Transactions of the Royal Society of London. Series A, Containing Papers of a Mathematical or Physical Character*, 231(694-706):289–337, 1933.

NIDDK. Albuminuria: Albumin in the urine. https://www.niddk.nih.gov/health-information/kidney-disease/chronic-kidney-disease-ckd/tests-diagnosis/albuminuria-albumin-urine, 2016. National Institute of Diabetes and Digestive and Kidney Diseases, Accessed: 2024-01-01.

T. Papaioannou and S. Loukas. Inequalities on Rank Correlation with Missing Data. *Journal of the Royal Statistical Society: Series B (Methodological)*, 46:68–71, 1984.

D. Quade and I. A. Salama. Concordance of Complete or Right-Censored Rankings Based on Spearman's Footrule. *Communications in Statistics-Theory and Methods*, 35:1059–1069, 2006.

R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2024. URL `https://www.R-project.org/`.

D. B. Rubin. Inference and Missing Data. *Biometrika*, 63(3):581–592, 1976.

D. B. Rubin. Multiple imputation after 18+ years. *Journal of the American statistical Association*, 91(434):473–489, 1996.

D. B. Rubin. Discussion on Multiple Imputation. *International Statistical Review*, 71(3):619–625, 2003.

D. B. Rubin and N. Schenker. Multiple Imputation for Interval Estimation From Simple Random Samples With Ignorable Nonresponse. *Journal of the American Statistical Association*, 81(394): 366–374, 1986.

I. Salama and D. Quade. Computing the distribution of spearman's footrule in o (nˆ 4) time. *Journal of Statistical Computation and Simulation*, 72(11):895–898, 2002.

I. A. Salama and D. Quade. A note on spearman's footrule. *Communications in Statistics-Simulation and Computation*, 19(2):591–601, 1990.

I. A. Salama and D. Quade. Agreement Among Censored Rankings using Spearman's Footrule. *Communications in Statistics-Theory and Methods*, 33:1837–1850, 2004.

J. L. Schafer. Multiple imputation: a primer. *Statistical Methods in Medical Research*, 8(1):3–15, 1999.

J. L. Schafer and J. W. Graham. Missing data: Our view of the state of the art. *Psychological Methods*, 7(2):147, 2002.

G. G. Schwartz, A. G. Olsson, M. D. Ezekowitz, P. Ganz, M. F. Oliver, D. Waters, A. Zeiher, B. R. Chaitman, S. Leslie, T. Stern, and Myocardial Ischemia Reduction with Aggressive Cholesterol Lowering (MIRACL) Study Investigators. Effects of Atorvastatin on Early Recurrent Ischemic Events in Acute Coronary Syndromes: The MIRACL Study: A Randomized Controlled Trial. *Journal of the American Medical Association*, 285(13):1711–1718, 2001.

P. Sen, I. Salama, and D. Quade. Spearman's footrule under progressive censoring. *Journal of Nonparametric Statistics*, 15:53–60, 2003.

D. A. Sheridan, I. T. Shawa, E. L. Thomas, D. J. Felmlee, S. H. Bridge, D. Neely, J. F. Cobbold, E. Holmes, M. F. Bassendine, and S. D. Taylor-Robinson. Infection with the hepatitis c virus causes viral genotype-specific differences in cholesterol metabolism and hepatic steatosis. *Scientific Reports*, 12(1):5562, 2022.

D. Siroker and P. Koomen. *A / B Testing: The Most Powerful Way to Turn Clicks Into Customers*. John Wiley & Sons, 2015.

C. Spearman. Footrule for measuring correlation. *British Journal of Psychology*, 2(1):89, 1906.

T. Speevak. Inequalities for sums of squares of reranked differences involving ties and missing data. *Communications in Statistics-Theory and Methods*, 46:8419–8429, 2017.

A. Staudt, J. Freyer-Adam, T. Ittermann, C. Meyer, G. Bischof, U. John, and S. Baumann. Sensitivity analyses for data missing at random versus missing not at random using latent growth modelling: a practical guide for randomised controlled trials. *BMC Medical Research Methodology*, 22(1):250, 2022.

S. A. Stouffer, E. A. Suchman, L. C. DeVinney, S. A. Star, and J. R. M. Williams. The american soldier: Adjustment during army life.(studies in social psychology in world war ii), vol. 1. 1949.

F. Strack, L. L. Martin, and S. Stepper. Inhibiting and facilitating conditions of the human smile: a nonobtrusive test of the facial feedback hypothesis. *Journal of Personality and Social Psychology*, 54(5):768, 1988.

L. Thabane, L. Mbuagbaw, S. Zhang, Z. Samaan, M. Marcucci, C. Ye, M. Thabane, L. Giangregorio, B. Dennis, D. Kosa, B. D. Victoria, D. Rejane, F. Vincent, B. Monica, L. Juneyoung, W. George, and H. G. Charles. A tutorial on sensitivity analyses in clinical trials: the what, why, when and how. *BMC Medical Research Methodology*, 13:1–12, 2013.

H. K. Ury and D. C. Kleinecke. Tables of the Distribution of Spearman's Footrule. *Journal of the Royal Statistical Society Series C: Applied Statistics*, 28(3):271–275, 1979.

Vivli. The Vivli platform. https://vivli.org/, 2023. Accessed: 2023-12-01.

I. R. White, P. Royston, and A. M. Wood. Multiple imputation using chained equations: issues and guidance for practice. *Statistics in Medicine*, 30(4):377–399, 2011.

F. Wilcoxon. Individual Comparisons by Ranking Methods. In *Breakthroughs in statistics*, pages 196–202. Springer, 1992.

M. Yağanoğlu. Hepatitis c virus data analysis and prediction using machine learning. *Data & Knowledge Engineering*, 142:102087, 2022.

Y. Zeng. *abwm: Performs Ansari-Bradley Test with Missing Data*, 2025a. URL `https://CRAN.R-project.org/package=abwm`. R package version 0.1.0.

Y. Zeng. *bosfr: Computes Exact Bounds of Spearman's Footrule with Missing Data*, 2025b. URL `https://CRAN.R-project.org/package=bosfr`. R package version 0.1.0.

Y. Zeng, N. M. Adams, and D. A. Bodenham. On two-sample testing for data with arbitrarily missing values. *arXiv preprint arXiv:2403.15327*, 2024a.

Y. Zeng, B. Dean, and N. M. Adams. *wmwm: Performs Wilcoxon-Mann-Whitney Test with Missing Data*, 2024b. URL `https://CRAN.R-project.org/package=wmwm`. R package version 1.0.0.

Y. Zeng, N. M. Adams, and D. A. Bodenham. Mmd two-sample testing in the presence of arbitrarily missing data. *Transactions on Machine Learning Research*, 2025a. ISSN 2835-8856.

Y. Zeng, N. M. Adams, and D. A. Bodenham. Exact bounds of spearman's footrule in the presence of missing data with applications to independence testing. *Electron. J. Statist.*, 19 (2):6103–6140, 2025b.

Y. Zeng, N. M. Adams, and D. A. Bodenham. Scale two-sample testing with arbitrarily missing data. *arXiv preprint arXiv:2509.20332*, 2025c.

Q. Zhang, C. Xu, and T. Yang. Variability of water resource in the yellow river basin of past 50 years, china. *Water resources management*, 23(6):1157–1170, 2009.

Y. Zhou, K. Xu, L. Zhu, and R. Li. Rank-based indices for testing independence between two high-dimensional vectors. *Annals of Statistics*, 52(1):184, 2024.

Y. Zhu, J. Lu, Z. Yu, D. Li, M. Wu, P. Shen, H. Lin, J. Wang, and K. Chen. Intra-individual variability of total cholesterol is associated with cardiovascular disease mortality: A cohort study. *Nutrition, Metabolism and Cardiovascular Diseases*, 29(11):1205–1213, 2019.