

Exact bounds of Spearman’s footrule in the presence of missing data with applications to independence testing

Yijin Zeng¹, Niall Adams¹ and Dean Bodenham¹ 

¹Imperial College London, e-mail: yijin.zeng20@imperial.ac.uk; n.adams@imperial.ac.uk; dean.bodenham@imperial.ac.uk

Abstract: This work studies exact bounds of Spearman’s footrule in the presence of missing data for two n -dimensional distinct real-valued vectors X and Y . The lower bound is obtained by sequentially constructing imputations of the partially observed vectors, each with a non-increasing value of Spearman’s footrule. The upper bound is found by first 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. Algorithms with computational complexities $O(n^2)$ and $O(n^3)$ are provided for computing these tight lower and upper bounds, respectively. As an application, we propose a novel two-sample independence testing method for data with missing values. Improving on all existing approaches, our method controls the Type I error under arbitrary missingness. Simulation results demonstrate that our method has good power, typically when the proportion of pairs containing missing data is below 15%. We illustrate our method on real-world data by comparing the dependence of yearly precipitation levels in the United Kingdom, Germany and the United States.

MSC2020 subject classifications: Primary 62H20, 62D10; secondary 62G10.

Keywords and phrases: Spearman’s footrule, rank correlation, missing data, independence testing.

Received August 2025.

1. Introduction

We study exact bounds of Spearman’s footrule between two partially observed vectors $X, Y \in \mathbb{R}^n$ with distinct values. Spearman’s footrule is a measure of statistical association based on ranks, introduced by Spearman in [43] as a robust alternative to Spearman’s rank correlation. This statistic is defined between two real-valued vectors X and Y and calculates the sum of absolute differences between the ranks of corresponding components in X and Y . In this work we consider the data to have distinct values and the rank vectors to contain no ties, as is the case for continuous-valued data. The case where ties are allowed is the subject of future work.

Among rank-based measures of association, Spearman’s rank correlation and Kendall’s τ coefficient [22] are more well-known. However, Spearman’s footrule has experienced a resurgence of interest, perhaps due to [13], where the asymptotic normality of Spearman’s footrule is established under the assumption of independence between X and Y . Because of its simplicity and robustness to outliers, Spearman’s footrule has found applications in various fields, including information retrieval [16, 24], web search [3, 4], rank aggregation [15, 25], genomics [23], welfare studies [33] and management science [32, 7].

In practice, it frequently occurs that a subset of the data are missing, posing a challenge to the direct computation of Spearman’s footrule. A common setting with missing data is called

top- k lists [16, 25, 15, 4], also known as right-censored rankings [34], censored rankings [37] and progressive censoring [41]. This setting assumes that all missing data are implicitly larger than any observed data, resulting in only the data with smaller ranks being observed. Under this setting, and additionally assume that for any pair of X and Y , at least one component in X or Y is observed, Spearman's footrule is demonstrated to remain constant, regardless of the values of missing data [16].

In this paper, we study the exact bounds of Spearman's footrule without the previously described assumption. To the best of our knowledge, the issue of defining Spearman's footrule under conditions of general missingness has received limited attention, with the exception of the work in [2], where the authors define Spearman's footrule with missing data using the conditional expectation given all observed values. Other related work is discussed in Section 1.3. However, these approaches all make different assumptions, either imposing constraints on the missing values or assuming specific patterns of missingness.

While one might either ignore all pairs with at least one missing observation or impute any missing data using the observed values, we show that these practices will result in a biased estimate of Spearman's footrule except in special cases such as when the missing data are missing completely at random. It appears intractable to give an unbiased estimate of Spearman's footrule based on observed values only under general cases of missingness. Our contribution fills this gap by establishing exact bounds for Spearman's footrule without relying on restrictive assumptions about the nature of the missing data.

A related issue to measuring the statistical association of X and Y is testing for independence between the two quantities. In the presence of missing data, there is no default approach to testing for independence. When the missingness mechanisms are known, for example if the data are missing completely at random, the practice of ignoring the missing data may be justified and more sophisticated approaches may be viable, e.g. [31]. However, in most real-world data analysis situations, such information is unavailable, and such practices may not necessarily be valid. Perhaps the most common practice is to simply ignore all the pairs with missing data and conduct the independence testing only for completely observed pairs, as noted in [35, 1]. The "risk" of ignoring pairs with missing values is considered by some authors to be inconsequential when the proportion of missing data is less than 5% [38, 20] or 10% [6, 6], but others argue against such heuristics [29]. We show that even in certain cases where 5% to 10% of the data is missing, ignoring or imputing these missing values is perilous.

As an application, we explore the use of our new bounds for Spearman's footrule for independence testing in the presence of missing data. Spearman's footrule is a natural statistic for testing independence [33, 17, 28, 10], and has greater statistical power than Spearman's rank correlation and Kendall's τ under certain alternatives [17, 28, 10].

The key idea of our testing method is to provide tight bounds of all possible p -values given the observed values. For a pre-specified significance level α , the null hypothesis is rejected only when the upper bound of the p -values, and hence all possible p -values, is less than or equal to α . When at least one p -value is not significant, then the null hypothesis fails to be rejected.

One motivation for this approach is that it allows us to have confidence in any significant result, because the significant result was obtained without making any assumption about the missing data. In other words, if a significant result is obtained, then even if the missing data could have been observed or an imputation method had been employed, the result would still

have been significant. While this is a conservative method of testing, the derived bounds for Spearman's footrule are tight, and so the results cannot be improved while taking into account all possible values for the missing data. The idea of employing bounds to consider all possible outcomes was previously used in [21] to take into account the effect of missing covariates and outcomes in clinical trials and observational studies.

This approach of taking all possible p -values into accounts could also be interpreted as a sensitivity analysis [46, 18, 11, 42] that considers all possible assumptions on the missing data. As we mentioned before, if a significant result is obtained by our method, then this result is consistent among all possible missing data assumptions, since every possible imputation would lead to a significant result. If, however, our method fails 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 -values 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 method answers whether different approaches to handling missing data lead to different conclusions [46].

We demonstrate the usefulness of the proposed method through various numerical simulations. The proposed method has good statistical power, typically when less than 15% of data are missing, while all other methods fail to control the Type I error when the data are missing informatively. The utility of the proposed method is further illustrated on a real-world case study examining the dependency of yearly precipitation in the United Kingdom and France.

1.1. Contributions

We provide exact bounds of Spearman's footrule in the presence of missing data for two univariate vectors of distinct values. The bounds are tight, meaning they cannot be improved using the observed data only. Let X and Y be n -dimensional distinct real-valued vectors, which may only be partially observed. If the missing values of X and Y are replaced with real values, we obtain *imputations* X_* and Y_* which will be more formally defined in Section 2.1. Theorem 2.11 provides conditions under which X_* and Y_* achieve the lower bound of Spearman's footrule, while Theorem 2.22 deals with the upper bound.

While it is possible to find X_* and Y_* that give the lower and upper bounds simply by enumerating all possible ranks of missing data, this strategy is computationally infeasible for even moderate sample sizes. Suppose X and Y are of length n and have m_X and m_Y missing values, respectively. There are $(n!)^2 / ((n - m_X)!(n - m_Y)!)$ possible permutations of the ranks. Consequently, when $n = 100$, $m_X = m_Y = 10$, the number of possible permutations of ranks is approximately 3.95×10^{39} . The second contribution of this paper is to provide efficient algorithms for computing our exact lower and upper bounds of Spearman's footrule.

To find X_* and Y_* for the lower bound, we propose Algorithm 2, which requires computation complexity of $O(n^2)$. However, finding the upper bound is more challenging. An algorithm with computational complexity $O(n^3)$ is provided in Section E.4 of the Supplementary Material [49] for finding the upper bound. More precisely, Lemma 2.23 states that the algorithm requires computational complexity $O(n \log n + \min\{m_1, m_2\}^2 + m_1 m_2 m_3)$, where m_1, m_2, m_3 are the number of pairs in (X, Y) such that only components in X are missing, only components in Y are missing, and both components in X, Y are missing, respectively.

Our work offers insight to anyone using Spearman's footrule with missing data. Furthermore, given the close link between Spearman's footrule and Kendall's τ , our results lead to non-trivial bounds for Kendall's τ with missing data.

We explore the use of the bounds for Spearman's footrule for independence testing in the presence of missing data. The core idea is to give the bounds of all possible p -values given the observed values. Unlike all other approaches for testing independence with missing data, our test is capable of controlling the Type I error regardless of the values of missing data and the missingness mechanisms. Extensive numerical simulations are conducted and confirm our method is particularly useful when data are missing not at random.

1.2. Missing data mechanisms

We briefly describe the three missing data scenarios outlined in [26]: missing completely at random (MCAR), missing at random (MAR) and missing not at random (MNAR). Consider a vector of univariate real-value samples $z = (z_1, z_2, \dots, z_N)$. Let $\iota = (\iota_1, \dots, \iota_N)$ be an indicator function of z such that ι_i taking value 1 if z_i is missing and 0 if z_i is observed. The core idea of [26] is to admit ι as a probabilistic phenomenon. Let z be a realized value of a random variable Z . Let $f_\theta(\iota|z)$ denotes the probability of $I = \iota$ given $Z = z$, where θ denotes any unknown parameters of the distribution. Then, the missingness mechanism is MCAR if $f_\theta(\iota|z) = f_\theta(\iota|\tilde{z}), \forall \iota, z, \tilde{z}$. In such cases, the missingness mechanism I is independent of the value of samples. Denote z' as a sub-vector of z including all observed samples in z , i.e. including all z_i such that $\iota_i = 0$. Then, the missingness mechanism is MAR if $f_\theta(\iota|z) = f_\theta(\iota|\tilde{z}), \forall \iota, z, \tilde{z}$ such that $z' = \tilde{z}'$. In such cases, the missingness mechanism I is independent of the values of missing samples. If the missingness mechanisms is neither MCAR nor MAR, it is MNAR.

1.3. Related work

The contributions of this paper are providing exact bounds of Spearman's footrule in the presence of missing data and proposing a new independence testing method with missing data based on these bounds. A special missing data scenario often discussed in relation to Spearman's footrule is the so-called "top- k lists" situation, as explored in [16, 25, 15, 34, 37, 41], where all missing data are assumed larger than any observed data and only smaller ranks can be observed. Our study of Spearman's footrule with missing data makes no assumption about the missing data except that they are distinct values.

To the best of our knowledge, only [2] considers the same general missing pattern as we do, where Spearman's footrule is defined as the conditional mean given all observed data. Another closely-related work [30] assumes all data in X, Y are distinct and establishes the exact lower bound of Spearman's footrule under the missing case where all pairs must either both be observed or missing, denoted *Missing Case III* below. The bounds found in this special case are then extended to the case where observations in X and Y are potentially tied [27, 44]. Additional missing data scenarios are studied in [8], where the missing data are categorized into partial, interval, and bucket cases, with potential for ties or ranking within a defined range. Their findings indicate that computing the bounds of Spearman's footrule in "partial

missing" cases is NP-hard, leaving the computational feasibility of other scenarios open. In this work, we assume all data are distinct, and values are not tied.

Another line of research closely related to our work is independence testing in the presence of missing data. Spearman's rank correlation and Kendall's τ coefficient in the presence of missing data are defined using conditional expectation given all observed values in [1]. Under the assumption that the incomplete rank vectors are uniformly distributed over all possible permutations, the distributions of these two statistics are then derived and used for hypothesis testing. The test results are analyzed empirically in [9]. Using the same approach as in [1], Spearman's footrule has been applied for trend testing in the presence of missing data [2]. The independence testing problem in the case where data are missing at random has been considered in [35, 31].

Beyond the above test statistics adapted for independence testing with missing data, broader methodologies for handling missing data are available. These include case deletion (i.e. only using the fully observed data) [40], single imputation [39], multiple imputation [38], and the expectation maximization algorithm [12]. All methods rely on assumptions about the nature of the missing data such as missing completely at random or missing at random. When data are missing not at random, knowledge of the specific missingness mechanisms are required [39]. For comprehensive overviews of these strategies, [26, 39, 14, 5] are recommended. Our unique contribution makes no assumption about the missing data except that it assumes all components of a vector are distinct. Our testing method offers a quantifiable measure of the impact of missing data on testing outcomes, through identifying bounds on p -values.

2. Exact bounds of Spearman's footrule in the presence of missing data

2.1. Notation and definitions

For any $n \in \mathbb{N}$, let $[n]$ denote the set $\{1, \dots, n\}$. 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 $X = (X(1), \dots, X(n))$.

We will only consider vectors with distinct real values, i.e. there are no ties. 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 $X \in \mathbb{R}^n$ and all components of X are distinct.

If $X \in \mathcal{V}^n$, given a set of indices $U \subset [n]$, we shall consider $X_* \in \mathcal{V}^n$ to be an *imputation* of X when $X_*(i) = X(i)$ for all $i \in [n] \setminus U$, and the *imputed values* are $X_*(u)$ 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 $(X(l))_{l \in S} = (X(1), X(3), X(4))$.

We define the rank of a component using the indicator function. For a statement A , we denote $I(A)$ as a indicator function such that if the statement A is correct, $I(A) = 1$, otherwise $I(A) = 0$. Subsequently, for any n -dimensional vector X of distinct real numbers, the rank of any component $X(i)$ in X is defined as

$$R(X(i), X) = \sum_{l=1}^n I(X(l) \leq X(i)).$$

Using this notation, Spearman's footrule for any two n -dimensional real-valued, distinct vectors X, Y is defined as

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

If a vector contains missing values, we will refer to it as *partially observed*, while a vector with no missing values is *fully observed*. Our objective is to establish the exact bounds of Spearman's footrule with partially observed vectors X and Y , where the components of X and Y can be missing in any configuration. 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 2.1. 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.

$$(X_2, X_3) = (X(\sigma(1)), \dots, X(\sigma(n))), \quad (Y_2, Y_3) = (Y(\sigma(1)), \dots, Y(\sigma(n))).$$

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*.

The proof can be found in Section A.1 in the Supplementary Material. For illustrating the result, we provide one example in Table 1, where the data under the 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 that is 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

Table 1

(Left) Example data under the general missing case, with missing values denoted by *, where we do not make any assumption about the missingness pattern. (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

	<i>a</i>	<i>b</i>	<i>c</i>	<i>d</i>	<i>e</i>	<i>f</i>	<i>g</i>	<i>h</i>		<i>a</i>	<i>d</i>	<i>g</i>	<i>e</i>	<i>c</i>	<i>f</i>	<i>b</i>	<i>h</i>
<i>X</i>	5.2	*	4.6	2.1	*	3.3	1.8	*	<i>X</i>	5.2	1.8	6	*	4.6	3.3	*	*
<i>Y</i>	*	*	1.2	*	3.6	2.4	*	*	<i>Y</i>	*	*	*	3.6	1.2	2.4	*	*

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.

$$R(X, X) = (R(X(1), X), \dots, R(X(n), X)).$$

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.

2.2. Lower bound of Spearman's footrule

In what follows, we first provide the exact lower bounds of Spearman's footrule. In Section 2.2.1 and 2.2.2, the problem for establishing exact lower bounds under Missing Case I and II are discussed, respectively. Algorithms 1 and 2 provide computationally efficient methods for computing these bounds. The lower bound under Missing Case III is discussed in [30], although it not shown there that the bound is tight. In Proposition 2.10, we show this bound is tight. Finally, by synthesizing the results developed for Missing Case I, II and III, we provide the lower bound in the general case. The discussion of the upper bound in Section 2.3 follows the same structure. The proof of all results can be found in the Supplementary Material.

2.2.1. 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.

The following proposition lays the groundwork for future results. It investigates the conditions for taking the minimum Spearman's footrule when one component $X(u)$ in X is replaced by a new value $X_*(u)$.

Proposition 2.2. Suppose $X, Y \in \mathcal{V}^n$ and let $X_* \in \mathcal{V}^n$ be an imputation of X for an index $u \in [n]$. If we choose $X_*(u)$ such that $R(X_*(u), X_*) = R(Y(u), Y)$, then

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

Moreover, for any other imputation $X' \in \mathcal{V}^n$ of X for index u , $D(X_*, Y) \leq D(X', Y)$.

This proposition demonstrates that if one component of X is changed, Spearman's footrule will be minimized 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 value 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 2.2, 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 2.3. *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*

$$X_*(i) = X(i) \text{ for } i \in [n] \setminus U, \text{ and } R(X_*(i), X_*) = R(Y(i), Y) \text{ for } i \in U. \quad (1)$$

The proof for Lemma 2.3 in the Supplementary Material 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 $u_1 = \operatorname{argmin}\{R(Y(u), Y) \mid u \in U\}$. 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 1 below formally describes the procedure of the proof of Lemma 2.3. Note that in Algorithm 1 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 $D(X_*, Y)$.

Algorithm 1 An Algorithm for Computing the Exact Lower Bound of Spearman's Footrule Under Missing Case I

Require: $X, Y \in \mathcal{V}^n$, where X may be partially observed and Y is fully observed.

Ensure: Minimum possible Spearman's footrule distance between X and Y .

```

1: Rank all observed components in  $X$  and  $Y$ .
2: while  $U \neq \emptyset$  do
3:   Identify  $u_1 = \operatorname{argmin}\{R(Y(i), Y) \mid i \in U\}$ .
4:   Let  $R(X(u_1), (X(i))_{i \in \{[n] \setminus U\} \cup \{u_1\}}) = R(Y(u_1), Y)$ .
5:   for  $i \in [n] \setminus U$  do
6:     Denote  $r_i = R(X(i), (X(i))_{i \in [n] \setminus U})$ 
7:     Update  $R(X(i), (X(j))_{j \in \{[n] \setminus U\} \cup \{u_1\}}) \leftarrow r_i + I(r_i \geq R(Y(u_1), Y))$ .
8:   end for
9:   Update  $U \leftarrow U \setminus \{u_1\}$ .
10: end while
11: Output  $D(X, Y)$ .
```

Remark 1. The computational complexity of Algorithm 1 is analyzed as follows. Ranking all observed components in X and Y in line 1 requires $O(n \log n)$ steps. Inside the while loop, identifying u_1 in line 3 requires the computational complexity $O(n)$, and updating ranks of observed components between line 5 and line 8 requires the computational complexity $O(n)$. Denote $m_1 = |U|$. The while loop run m_1 times. Since each iteration is $O(n)$, the computational complexity of loop is $O(m_1 n)$. Overall, the computational complexity for Algorithm 1 is $O(n \log n + m_1 n)$.

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, we have the following lemma which will be useful in proving later results.

Lemma 2.4. *Suppose $X, Y \in \mathcal{V}^n$ and $U \subset [n]$ is a subset of indices. Consider any $X_1, X_2 \in \mathcal{V}^n$ satisfying the conditions in (1). Then $D(X_1, Y) = D(X_2, Y)$.*

We are now ready to state the main result for *Missing Case I*, which is an extension of Proposition 2.2.

Theorem 2.5. *Suppose $X, Y \in \mathcal{V}^n$ and $U \subset [n]$ is a subset of indices. Consider any $X_* \in \mathcal{V}^n$ satisfying the conditions in (1). Then*

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

Moreover, for any $X' \in \mathcal{V}^n$ such that $X'(i) = X(i)$ for $i \in [n] \setminus U$, we have

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

Theorem 2.5 provides the conditions under which the imputation X_* will minimise Spearman's footrule under *Missing Case I*. Moreover, Lemma 2.3 shows that X_* exists and so this lower bound is tight.

The strategy for proving Theorem 2.5 is to sequentially construct imputations $X_1, X_2, X_3 \in \mathcal{V}^n$ of X for the set of indices U . As before, define $u_1 = \operatorname{argmin}\{R(Y(u), Y) \mid u \in U\}$.

The first vector X_1 is an imputation of X for indices $U \setminus \{u_1\}$ such that the ranks of X_1 and Y are equal for the set $U \setminus \{u_1\}$, i.e. $R((X_1(u))_{u \in U \setminus \{u_1\}}, X_1) = R((Y(u))_{u \in U \setminus \{u_1\}}, Y)$. The second vector X_2 imputes the u_1 component, so that its rank is equal to that of its paired component, i.e. $R(X_2(u_1), X_2) = R(Y(u_1), Y)$. However, it may now be the case that the ranks of the components of X_2 for the indices in $U \setminus \{u_1\}$ may not equal the ranks of their paired components in Y , i.e. it may be that $R((X_2(u))_{u \in U \setminus \{u_1\}}, X_1) \neq R((Y(u))_{u \in U \setminus \{u_1\}}, Y)$. Therefore, the third vector X_3 is constructed as an imputation of X_2 for the indices $U \setminus \{u_1\}$, so that the ranks of its imputed components are equal to the ranks of its paired components in Y , and $R((X_3(u))_{u \in U}, X_3) = R((Y(u))_{u \in U}, Y)$.

It is then demonstrated in the proof of Theorem 2.5 that Spearman's footrule decreases sequentially as follows:

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

Since X_3 satisfies the conditions in (1), if we use Algorithm 1 to construct X_* which also satisfies the conditions in (1), by Lemma 2.4 we must have $D(X_3, Y) = D(X_*, Y)$.

Theorem 2.5 reveals an interesting property of Spearman's footrule. In order to minimize Spearman's footrule in the presence of missing data, or in other words, minimize Spearman's footrule for the pair $X, Y \in \mathcal{V}^n$ by imputing values for $(X(u))_{u \in U}$ for a subset of indices $U \subset [n]$, one simply needs to construct an imputation X_* that satisfies the conditions in (1).

This property, however, does not hold for minimizing Spearman's rank correlation ρ and Kendall's τ coefficient defined for $X, Y \in \mathcal{V}^n$ as

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

Table 2

(Left) Example ranked data under Missing Case I to compare Spearman’s footrule D , Spearman’s rank correlation ρ and Kendall’s τ coefficients, where the $*$ indicates that the rank of object d for vector X is missing. (Right) Values of these statistics for selected imputations X_* of X , when the rank of d is imputed as either 1 or 4. Note that Spearman’s footrule is minimized when the rank of $d = 4$, while Spearman’s ρ and Kendall’s τ are minimized when $d = 1$. For more details, see the complete table where d is imputed as 1, 2, . . . , 8 provided in Table 3.1 in the Supplementary material

	a	b	c	d	e	f	g	h	X_*	$D(X_*, Y)$	$\rho(X_*, Y)$	$\tau(X_*, Y)$
$R(X, X)$	7	3	6	*	2	5	4	1	$d = 1$	26	122	19
$R(Y, Y)$	1	2	3	4	5	6	7	8	$d = 4$	24	128	20

$$\tau(X, Y) = \sum_{i>j} \{I(X_i > X_j) I(Y_i < Y_j) + I(X_i < X_j) I(Y_i > Y_j)\}, \tag{3}$$

respectively. Note that $\rho(X, Y)$ and $\tau(X, Y)$ in Equations (2) and (3) are unscaled versions of the coefficients that measure the amount of discordance. Equation (6) provides the more familiar versions of the coefficients scaled to $[-1, 1]$.

As an example, consider the data shown in left-hand part of Table 2, where objects a, b, c, d, e, f, g, h have been given two different sets of ranks, denoted by X and Y , but the rank of object d for X is missing.

A natural question to ask is what rank of d will maximize the “agreement” between the two referees X and Y . For any of the three rank correlation statistics, maximizing the agreement is equivalent to minimizing the statistic, since a perfect agreement would mean equal ranks and a zero-valued statistic. Theorem 2.5 concludes that minimizing Spearman’s footrule requires the rank of d to be 4, to match the rank given by Y . However, as shown in Table 3.1 in Supplementary Material, Spearman’s rank correlation ρ and Kendall’s τ coefficient are both minimized when the rank of d is 1; the right-hand part of Table 2 provides partial results.

This special property of Spearman’s footrule is further explored and extended in the context of Missing Case II, as discussed in Theorem 2.8.

2.2.2. 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 A. $X_*, Y_* \in \mathcal{V}^n$ 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 B. $X_*, Y_* \in \mathcal{V}^n$ are such, for indices $U, V \subset [n]$,

$$R(X_*(i), X_*) = R(Y_*(i), Y_*), i \in U \cup V.$$

Condition A ensures that X_* and Y_* are imputations of X and Y satisfying Missing Case II. Condition B 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 2.8 that the minimum possible value of Spearman's footrule between X and Y equals $D(X_*, Y_*)$ when X_* and Y_* satisfy both Conditions A and B. While finding X_* and Y_* that satisfy Condition A is straightforward, it is less obvious that we can always find imputations that satisfy both Conditions A and B. Hence, we first prove the existence of such imputations.

Proposition 2.6. *Suppose $X, Y \in \mathcal{V}^n$ and disjoint $U, V \subset [n]$ are such that $U \cup V \neq \emptyset$. Then, there exist imputations $X_*, Y_* \in \mathcal{V}^n$ satisfying both Conditions A and B.*

The proof of Proposition 2.6 in the Supplementary Material follows a similar idea to the proof of Lemma 2.3. The construction of X_* and Y_* is summarised in Algorithm 2 below. Table 3 provides two examples illustrating this result.

Algorithm 2 An algorithm for computing the exact lower bound of Spearman's footrule under *Missing Case II*

Require: $X, Y \in \mathcal{V}^n$, where both vectors may be partially observed under *Missing Case II*.

Ensure: Minimum possible Spearman's footrule distance between X and Y .

```

1: Rank all observed components in  $X, Y$ .
2: while  $U \neq \emptyset$  and  $V \neq \emptyset$  do
3:   Let  $u_1 = \operatorname{argmin} \left\{ R \left( X(i), (X(j))_{j \in [n] \setminus U} \right) \mid i \in V \right\}$ .
4:   Let  $u_2 = \operatorname{argmin} \left\{ R \left( Y(i), (Y(j))_{j \in [n] \setminus V} \right) \mid i \in U \right\}$ .
5:   if  $R \left( X(u_1), (X(j))_{j \in [n] \setminus U} \right) \leq R \left( Y(u_2), (Y(j))_{j \in [n] \setminus V} \right)$  then
6:     Let  $R \left( Y(u_1), (Y(j))_{j \in \{[n] \setminus V\} \cup \{u_1\}} \right) = R \left( X(u_1), (X(j))_{j \in [n] \setminus U} \right)$ 
7:     for  $i \in [n] \setminus V$  do
8:       Denote  $R \left( Y(i), (Y(j))_{j \in [n] \setminus V} \right) = s_i$ .
9:       Update  $R \left( Y(i), (Y(j))_{j \in \{[n] \setminus V\} \cup \{u_1\}} \right) \leftarrow s_i + I \left( s_i \geq R \left( X(u_1), (X(j))_{j \in [n] \setminus U} \right) \right)$ .
10:    end for
11:    Update  $V \leftarrow V \setminus \{u_1\}$ .
12:   else
13:     Let  $R \left( X(u_2), (X(j))_{j \in \{[n] \setminus U\} \cup \{u_2\}} \right) = R \left( Y(u_2), (Y(j))_{j \in [n] \setminus V} \right)$ 
14:     for  $i \in [n] \setminus U$  do
15:       Denote  $R \left( X(i), (X(j))_{j \in [n] \setminus U} \right) = r_i$ .
16:       Update  $R \left( X(i), (X(j))_{j \in \{[n] \setminus U\} \cup \{u_2\}} \right) \leftarrow r_i + I \left( r_i \geq R \left( Y(u_2), (Y(j))_{j \in [n] \setminus V} \right) \right)$ .
17:     end for
18:     Update  $U \leftarrow U \setminus \{u_2\}$ .
19:   end if
20: end while
21: if  $U \neq \emptyset$  or  $V \neq \emptyset$  then
22:   Run Algorithm 1 when  $U \neq \emptyset$ . Reverse  $X$  and  $Y$  and run Algorithm 1 when  $V \neq \emptyset$ .
23: end if
```

Remark 2. The computational complexity of Algorithm 2 is analyzed as follows. Ranking all observed components in X and Y in line 3 takes $O(n \log n)$ steps. Inside the while loop, the computational complexity for finding the minimum ranks in line 3 and 4 is $O(n)$, and for updating ranks of observed components from line 7 to line 10, and line 14 to line 17 is $O(n)$. Let $|U| = m_1$, and $|V| = m_2$. The loop can run up to $m_1 + m_2 - 1$ times. Since each iteration is $O(n)$, the computational complexity for the while loop is $O(n(m_1 + m_2))$. The computational complexity for Algorithm 1, called in line 22, after ranking all observed components is $O(nm_1)$ when $U \neq \emptyset$ and $O(nm_2)$ when $V \neq \emptyset$, according to Remark 1. Therefore, the total computation complexity for Algorithm 2 is $O(n \log n + n(m_1 + m_2))$.

Table 3

(Left) Values of (X, Y) under Missing Case II, and an example of two pairs of 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 2.6 guarantees the existence of such imputations, while Algorithm 2 provides a method for constructing such imputations. Proposition 2.7 shows that all such imputations yield the same value of Spearman’s footrule. Theorem 2.8 further demonstrates that these imputations minimize Spearman’s footrule

Values	<i>a</i>	<i>b</i>	<i>c</i>	<i>d</i>	<i>e</i>	<i>f</i>	<i>g</i>	<i>h</i>	Ranks	<i>a</i>	<i>b</i>	<i>c</i>	<i>d</i>	<i>e</i>	<i>f</i>	<i>g</i>	<i>h</i>
<i>X</i>	6.3	2.4	5.5	*	*	4.7	3.8	1.8	<i>R</i> (<i>X</i> , <i>X</i>)	6	2	5	*	*	4	3	1
<i>Y</i>	*	*	1.1	2.4	3.5	4.2	5.8	6.7	<i>R</i> (<i>Y</i> , <i>Y</i>)	*	*	1	2	3	4	5	6
<i>X</i> ₁	6.3	2.4	5.5	2.9	3.1	4.7	3.8	1.8	<i>R</i> (<i>X</i> ₁ , <i>X</i> ₁)	8	2	7	<u>3</u>	<u>4</u>	6	5	1
<i>Y</i> ₁	8.2	2.1	1.1	2.4	3.5	4.2	5.8	6.7	<i>R</i> (<i>Y</i> ₁ , <i>Y</i> ₁)	<u>8</u>	<u>2</u>	1	3	4	5	6	7
<i>X</i> ₂	6.3	2.4	5.5	2.0	2.3	4.7	3.8	1.8	<i>R</i> (<i>X</i> ₂ , <i>X</i> ₂)	8	4	7	<u>2</u>	<u>3</u>	6	5	1
<i>Y</i> ₂	8.2	4.1	1.1	2.4	3.5	4.2	5.8	6.7	<i>R</i> (<i>Y</i> ₂ , <i>Y</i> ₂)	<u>8</u>	<u>4</u>	1	2	3	5	6	7

The following result shows that all imputations X_* and Y_* satisfying both Conditions A and B yield the same value of Spearman’s footrule.

Proposition 2.7. Suppose $X, Y \in \mathcal{V}^n$ and disjoint $U, V \subset [n]$ are such that $U \cup V \neq \emptyset$. If X_1, Y_1 and X_2, Y_2 are pairs satisfying Conditions A and B, then $D(X_1, Y_1) = D(X_2, Y_2)$.

Table 3 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 A and B, 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 2.4 can be proved in a relatively straightforward approach, the proof of Proposition 2.7 is much more challenging. For Missing Case I, the condition in (1) explicitly provides the ranks of all components in U for both X_1 and X_2 . On the other hand, Condition B 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 $R(Y_1(i), Y_1) = R(Y_2(i), Y_2)$.

Furthermore, the following result shows that minimum possible Spearman’s footrule is achieved when imputations X_* and Y_* satisfy both Conditions A and B.

Theorem 2.8. Suppose $X, Y \in \mathcal{V}^n$ and disjoint $U, V \subset [n]$ are such that $U \cup V \neq \emptyset$. If $X_*, Y_* \in \mathcal{V}^n$ are imputations satisfying both Conditions A and B, then

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

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

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

Theorem 2.8 provides the conditions under which the imputation X_*, Y_* will minimise Spearman’s footrule under Missing Case II. Moreover, Proposition 2.6 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 3. This example demonstrates how Proposition 2.6, Algorithm 2, Proposition 2.7, and Theorem 2.8 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.

2.2.3. 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 [30].

Theorem 2.9 ([30]). *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 2.10. *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,*

- (i) $R(X(i), X) = R(Y(i), Y)$, for all $i \in W$,
- (ii) $\min_{j \in W} X(j) > \max_{j \in [n] \setminus W} X(j)$,
- (iii) $\min_{j \in W} Y(j) > \max_{j \in [n] \setminus W} Y(j)$,

then $D(X, Y) = D(X', Y')$.

2.2.4. General missingness case

We are now ready to provide the lower bound of Spearman's footrule in the general case, where the missing pattern of data is a combination of *Missing Cases I, II*, and *III*. The previous results can be combined to give the following result. We are now ready to provide the lower bound of Spearman's footrule in the general case, where we do not make any assumptions of missing patterns. The previous results for *Missing Cases I, II*, and *III* will be combined together to prove the following result.

Theorem 2.11. *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*

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

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,

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

Table 4

(Left) Example ranked data under the general missingness case, where we do not make any assumptions about the missingness pattern and * indicates a missing value. (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 2.11. This imputation therefore gives a lower bound of Spearman’s footrule

	<i>a</i>	<i>b</i>	<i>c</i>	<i>d</i>	<i>e</i>	<i>f</i>	<i>g</i>	<i>h</i>		<i>a</i>	<i>b</i>	<i>c</i>	<i>d</i>	<i>e</i>	<i>f</i>	<i>g</i>	<i>h</i>
$R(X, X)$	5	1	4	*	*	3	2	*	$R(X_*, X_*)$	7	1	6	<u>3</u>	<u>4</u>	5	2	<u>8</u>
$R(Y, Y)$	*	*	1	2	3	4	5	*	$R(Y_*, Y_*)$	<u>7</u>	<u>1</u>	2	3	4	5	6	<u>8</u>

It is straightforward to show the existence of such X_*, Y_* in Theorem 2.11 by applying Proposition 2.6. Hence the lower bounds for Spearman’s footrule in the general missingness case are tight.

For the imputations X_* and Y_* in Theorem 2.11, Proposition 2.10 gives the equation $D(X_*, Y_*) = D((X_*(i))_{i \in [n] \setminus W}, (Y_*(i))_{i \in [n] \setminus W})$, 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 missingness case can be determined using Algorithm 2 after ignoring all components with indices in W . This algorithm has complexity $O(n \log n + n(m_1 + m_2))$, as discussed in Remark 2.

Table 4 provides a simple example where X and Y are missing under the general missingness case. The imputations X_*, Y_* satisfy the three conditions required in Theorem 2.11. Hence, the imputations X_*, Y_* will give the minimum value of Spearman’s footrule across all possible imputations.

2.3. Upper bound of Spearman’s footrule

This section provides the exact upper bound of Spearman’s footrule in the presence of missing data. We consider *Missing Cases I, II, and III* separately and then combine our results to obtain the upper bound in the general missing case.

2.3.1. 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 is complementary to Proposition 2.2 which considers the lower bound for *Missing Case I*.

Proposition 2.12. Suppose $X, Y \in \mathcal{V}^n$ and let $X_1, X_2 \in \mathcal{V}^n$ be imputations of X for an index $u \in [n]$ such that $R(X_1(u_1), X_1) = 1$ and $R(X_2(u_1), X_2) = n$. Then

$$D(X, Y) \leq \max\{D(X_1, Y), D(X_2, Y)\},$$

and for any imputation $X' \in \mathcal{V}^n$ of X for index u ,

$$D(X', Y) \leq \max\{D(X_1, Y), D(X_2, Y)\}.$$

When a single component of X is missing, Proposition 2.12 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 .

It is shown in the proof that in the special case where $R(Y(u_1), Y) = 1$, X_2 maximizes Spearman's footrule, and in the special case where $R(Y(u_1), Y) = n$, X_1 maximizes Spearman's footrule.

To extend Proposition 2.12 to cases when multiple components of X may be missing, we first make the following definition.

Definition 2.13. Let $U \subset T \subset [n]$ be a subset of indices. Define

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

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 state the following result.

Proposition 2.14. 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 , we have $D(X', Y) \leq D(X_*, Y)$.

When the components of X with indices in $U \subset [n]$ are missing, Proposition 2.14 allows us to obtain the maximum possible Spearman's footrule by considering rank vectors of all possible imputations X_* of X in the set $\mathcal{Z}(U, [n], n)$.

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 2.14 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. We start by introducing the following lemma.

Lemma 2.15. Suppose $X, Y \in \mathcal{V}^n$ and for $m_1 \leq n$ assume $U = \{1, \dots, m_1\}$ and $Y(1) < \dots < Y(m_1)$. Let X' be any imputation of X for indices U . Then, 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 $X_*(1) > \dots > X_*(m_1)$. Then, $D(X', Y) \leq D(X_*, Y)$.

Consider any imputation X' of X for indices U , Lemma 2.15 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 2.16. Suppose $Y \in \mathcal{V}^n$ and let $U \subset T \subset [n]$. Define

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

Then, combining Proposition 2.14 and Lemma 2.15, the following result will allow us to efficiently determine an imputation X_* that maximises Spearman's footrule.

Theorem 2.17. 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 , we have $D(X', Y) \leq D(X_*, Y)$.

Compared with Proposition 2.14, Theorem 2.17 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 3. 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 $O(n \log n)$. 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 $O(nm_1 \log n)$. 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 2.5 provides an efficient algorithm for computing the maximum Spearman's footrule under *Missing Case I* with computational cost $O(m_1 + n \log n)$.

2.3.2. 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.

The following result is obtained by applying Theorem 2.17 under *Missing Case I* twice.

Theorem 2.18. Suppose $X, Y \in \mathcal{V}^n$ and $U, V \subset [n]$. 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 $D(X', Y') \leq D(X_*, Y_*)$.

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 3 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 $(m_1 + 1)(m_2 + 1)$.

2.3.3. 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 Proposition 2.12 repeatedly.

Proposition 2.19. *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*

$$\begin{aligned} R(X_1(u_1), X_1) &= 1, \text{ and } R(Y_1(u_1), Y_1) = n, \\ R(X_2(u_1), X_2) &= n, \text{ and } R(Y_2(u_1), Y_2) = 1. \end{aligned}$$

Then, $D(X, Y) \leq \max\{D(X_1, Y_1), D(X_2, Y_2)\}$.

When one pair $(X(u), Y(u))$ is missing, Proposition 2.19 gives similar results as in Proposition 2.12, offering two possibilities for determining the maximum possible value of Spearman's footrule.

Furthermore, to achieve the maximum possible Spearman's footrule, Proposition 2.19 requires the paired ranks of the imputations for the index u to be at opposite ends of the sequence $1, \dots, n$. 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 2.20. Let $W \subset [n]$ be a subset of indices. Define $\mathcal{Y}(W, n)$ as

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

Using Definition 2.20, we can state the following result.

Theorem 2.21. *Let $X, Y \in \mathcal{V}^n$ 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 $D(X', Y') \leq D(X_*, Y_*)$.*

Theorem 2.21 establishes the existence of imputations $(X_*, Y_*) \in \mathcal{Y}(W, n)$ of X and Y that will maximize Spearman's footrule $D(X, Y)$. In order to find this pair (X_*, Y_*) , we could enumerate all possible imputations. If $X_*, Y_* \in \mathbb{R}^n$, there would be uncountably 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 2.20. Therefore, according to Definition 2.13 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 $r = \sum_{i \in W} I(X_*(i) < (X_*(l))_{l \in [n] \setminus W})$, 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 $r \in \{1, 2, \dots, m_3 - 1\}$. 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 $1, \dots, r, n - (m_3 - r) + 1, \dots, n$, and then the components of Y_* in W are arranged as $n, \dots, n + 1 - r, m_3 - r, 1$.

In the preceding paragraph, we considered $r \in \{1, 2, \dots, m_3 - 1\}$, 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 4 in the Supplementary Material.

2.3.4. 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 missing pattern. The previous results for *Missing Cases I, II, and III* can be combined to give the following result.

Theorem 2.22. *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 $D(X', Y') \leq D(X_*, Y_*)$.*

We note that since the imputations X_*, Y_* exist, the upper bounds for Spearman's footrule are tight.

Remark 4. 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 analyzed 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 2.20, and that X_* and Y_* satisfy the conditions of Definition 2.16 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 2.21, 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 3.

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 $(m_1 + 1)(m_2 + 1)(m_3 + 1)$.

Directly computing all $(m_1 + 1)(m_2 + 1)(m_3 + 1)$ values of Spearman's footrule requires computational complexity $O(m_1 m_2 m_3 n \log n)$ since computing Spearman's footrule once is $O(n \log n)$. However, there is a more efficient approach to computing the upper bound, as the following lemma shows.

Lemma 2.23. *Suppose $X, Y \in \mathcal{V}^n$, which can be partially observed. Denote the number of pairs in (X, Y) such that only components in X are missing, only components in Y are missing, and both components in X, Y are missing, as m_1, m_2 , and m_3 , respectively. One can then compute the tight upper bound of Spearman's footrule in $O(n \log n + \min\{m_1, m_2\}^2 + m_1 m_2 m_3)$.*

The tight lower bound for Spearman's footrule can be computed according to Algorithm 2 with computational time $O(n \log n + n(m_1 + m_2))$, as discussed following Theorem 2.11. An algorithm for computing the tight upper bound of Spearman's footrule is proposed in Section E.4 in the Supplementary Material with computational time $O(n \log n + \min\{m_1, m_2\}^2 + m_1 m_2 m_3)$. See more details in Remark 5 in the Supplementary Material.

Remark 5. If m_1, m_2 and m_3 are all proportional to n , then this means that finding the tight upper bound is $O(n^3)$. While this may be computationally expensive for large n , without any of the results 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 2.22, we include a simple example in Table 5. While one could naively consider all $(8 \times 7 \times 6)^2 = 112,896$ 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.

Table 5

(Left) Example ranked data under the general missingness case, where * indicates a missing value. (Right) Imputed rank vectors (X_*, Y_*) that maximize Spearman’s footrule. The imputed components are underlined. There are 112,896 possible imputations according to Remark 5, but by using Theorem 2.22 we only need to consider 18 of these imputations to find the maximum, as noted in Remark 4

	<i>a</i>	<i>b</i>	<i>c</i>	<i>d</i>	<i>e</i>	<i>f</i>	<i>g</i>	<i>h</i>		<i>a</i>	<i>b</i>	<i>c</i>	<i>d</i>	<i>e</i>	<i>f</i>	<i>g</i>	<i>h</i>
$R(X, X)$	5	1	4	*	*	3	2	*	$R(X_*, X_*)$	6	2	5	<u>8</u>	<u>7</u>	4	3	<u>1</u>
$R(Y, Y)$	*	*	1	2	3	4	5	*	$R(Y_*, Y_*)$	<u>1</u>	<u>7</u>	2	3	4	5	6	<u>8</u>

2.4. Bounds for Kendall’s τ

We have considered efficient methods for constructing tight bounds of Spearman’s footrule in the presence of missing data under the general missingness case. A natural question is whether parallel results exist for other rank-based association measurements such as Spearman’s rank correlation ρ and Kendall’s τ .

It is possible to derive tight bounds for Kendall’s τ under certain missingness patterns, such as *Missing Case III*, as demonstrated by [30]. However, deriving efficient methods for computing the tight bounds under the general missingness case is more challenging. There do not appear to be tight bounds for Spearman’s rank correlation ρ under any specific or general missingness pattern.

While the bounds of Spearman’s footrule above will not provide tight bounds for Kendall’s τ , we can derive non-trivial bounds that may have some utility when data are partially observed.

Let $X, Y \in \mathcal{V}^n$, and define Kendall’s τ coefficient according to (3). The following result is proved in Theorem 2 in [13], which shows that the Kendall’s τ coefficient and Spearman’s footrule D differ by at most a constant factor.

Theorem 2.24 ([13]). *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 us denote the exact lower and upper bounds of Spearman’s footrule between X, Y as $D_{\min}(X, Y)$ and $D_{\max}(X, Y)$, respectively, which can be computed according to Theorem 2.11 and Theorem 2.22, respectively. Since $D_{\min}(X, Y) \leq D(X, Y) \leq D_{\max}(X, Y)$, Theorem 2.24 gives

$$D_{\min}(X, Y)/2 \leq \tau(X, Y) \leq D_{\max}(X, Y). \tag{4}$$

In Section 2.6, these bounds of the Kendall’s τ coefficient given in (4) are investigated empirically using numerical simulations.

2.5. An efficient algorithm for computing the upper bounds

This section provides an efficient algorithm for calculating exact upper bounds of Spearman’s footrule in the presence of missing data. We consider *Missing Case I* in this section, i.e. missing components are either all in X or all in Y . The algorithms for *Missing Case II, III* and the general missing cases can be constructed similarly, and are provided in Section 2 of the Supplementary Material.

Without loss of generality, let us assume X is partially observed while Y is fully observed. Denote the indices of all missing components in X as U and denote $m_1 = |U|$. Further, assume $U = \{1, \dots, m_1\}$ and the components of Y with indices in U are such that $Y(1) < \dots < Y(m_1)$. If this is not the case, simply relabel the relevant components in X and Y .

Under *Missing Case I*, Theorem 2.17 allows us to obtain the maximum possible Spearman's footrule by considering imputations X_* of X for indices U in $\mathcal{X}(Y; U, [n], n)$. Remark 3 further demonstrates that it is adequate to only consider $m_1 + 1$ imputations $X_* \in \mathcal{X}(Y; U, [n], n)$, each with a distinct value of $r = \sum_{i \in U} I(X_*(i) < \min(X_*(l))_{l \in [n] \setminus U})$, i.e. the number of components of X_* in U that are smaller than all components of X_* with indices in $[n] \setminus U$, where $r \in \{0, \dots, m_1\}$. Let us denote the $m_1 + 1$ imputations as $X_*^{(0)}, \dots, X_*^{(m_1)}$, corresponding to the imputations with $r = 0, \dots, m_1$. Then the maximum possible value of Spearman's footrule between partially observed X and Y is $\max \{D(X_*^{(0)}, Y), \dots, D(X_*^{(m_1)}, Y)\}$.

The computational complexity of naively computing all $m_1 + 1$ values in the above set $\{D(X_*^{(0)}, Y), \dots, D(X_*^{(m_1)}, Y)\}$ is $O((m_1 + 1)n \log n)$, since computing the value of Spearman's footrule once is $O(n \log n)$.

However, we notice that 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. To construct an algorithm for efficiently calculating exact upper bounds of Spearman's footrule under *Missing Case I* in $O(n \log n)$, we provide the following result

Proposition 2.25. *Suppose $X, Y \in \mathcal{V}^n$ and for $m_1 \in [n]$, let $U = \{1, \dots, m_1\} \subset [n]$ be a subset of indices. Suppose $Y(1) < \dots < Y(m_1)$, and let $X_*^{(r)} \in \mathcal{X}(Y; U, [n], n)$ be an imputation of X for indices in U such that*

$$\sum_{i \in U} I(X_*^{(r)}(i) < \min(X_*^{(r)}(l))_{l \in [n] \setminus U}) = r.$$

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 $r \in \{0, \dots, m_1 - 1\}$, denote $s_r = \sum_{i \in [n] \setminus U} I(d_i \leq r)$. Then, for any $r \in \{0, \dots, m_1 - 1\}$, we have

$$D(X_*^{(r+1)}, Y) = D(X_*^{(r)}, Y) + 2s_r - n + m_1 + C^{(r)},$$

where $C^{(r)} = |r + 1 - R(Y(m_1 - r), Y)| - |n - m_1 + r + 1 - R(Y(m_1 - r), Y)|$.

Proposition 2.25 shows that the maximum possible values of Spearman's footrule can be found in linear time after ranking all observed components in X and Y , as shown in Algorithm 3.

Remark 6. The computational complexity of Algorithm 3 is analyzed as follows. Ranking and relabeling all observed components in X and Y in line 3 requires $O(n \log n)$ steps. Using these rankings, in line 5 computing each d_i is $O(1)$, and so overall line 5 is $O(n - m_1)$. In line 6, $s_0, \dots, s_{m_1-1}$ can be computed collectively in $O(n - m_1)$ steps. See Algorithm 1 in the Supplementary Material. is $O(n + m_1) = O(n)$. In line 7, each iteration of the for loop is $O(1)$, and since the loop runs m_1 times, the computational complexity for the loop is $O(m_1)$. Line 8 and 9 takes $O(1)$ and $O(m_1)$ steps, respectively. Therefore, the overall computational complexity for Algorithm 3 is $O(n \log n)$.

Algorithm 3 An efficient algorithm for computing exact upper bounds of Spearman's footrule under Missing Case I

Require: $X, Y \in \mathcal{V}^n$, where either X or Y is fully observed, while the other is partially observed.

Ensure: 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 $(X(1), \dots, X(m_1))$ are unobserved and $Y(1) < \dots < Y(m_1)$.
- 5: Let $U = \{1, 2, \dots, m_1\}$. For $i \in [n] \setminus U$, let $d_i = R(Y(i), Y) - R(X(i), (X(l))_{l \in [n] \setminus U})$.
- 6: Computing $s_r = \sum_{i \in [n] \setminus U} I(d_i \leq r)$ for any $r = \{0, \dots, m_1 - 1\}$.
- 7: For any $r \in \{0, \dots, m_1 - 1\}$, compute

$$C^{(r)} = |r + 1 - R(Y(m_1 - r), Y)| - |n - m_1 + r + 1 - R(Y(m_1 - r), Y)|.$$

- 8: Initialize $D_0 = \sum_{i=1}^{m_1} |(n - i + 1) - R(Y(i), (Y(l))_{l \in [n]})| + \sum_{i \in [n] \setminus U} |d_i|$.
 - 9: For $r \in \{0, \dots, m_1 - 1\}$, compute $D_{r+1} = D_r + 2s_r - n + m_1 + C^{(r)}$.
 - 10: Return $\max\{D_0, \dots, D_{m_1}\}$.
-

2.6. Simulation results for the bounds

This section provides numerical simulations that evaluate the performance of the exact bounds of Spearman's footrule, and the bounds of Kendall's τ coefficient provided in (4), between partially observed $X, Y \in \mathcal{V}^n$, with different proportions of the missing pairs.

Additionally, 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. When the data are missing not at random (MNAR), ignoring the missing data can lead to biased estimates of the correlation coefficients, as shown in Figure 2.

We first scale all the rank correlation coefficients between $[-1, 1]$. For scaling Spearman's footrule, we define

$$D_{\text{Scale}}(X, Y) = 1 - 3D(X, Y)/(n^2 - 1). \quad (5)$$

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 $[-0.5\{1 + 3/(n^2 - 1)\}, 1]$ [22]. For scaling Spearman's ρ and Kendall's τ , define

$$\begin{aligned} \rho_{\text{Scale}}(X, Y) &= 1 - \rho(X, Y)/\{n(n^2 - 1)\}, \\ \tau_{\text{Scale}}(X, Y) &= 1 - 4\tau(X, Y)/\{n(n - 1)\}, \end{aligned} \quad (6)$$

respectively, where $\tau(X, Y)$ and $\rho(X, Y)$ are given in (2) and (3), respectively. It can be shown that both $\rho_{\text{Scale}}(X, Y)$ and $\tau_{\text{Scale}}(X, Y)$ are in the range $[-1, 1]$ [22].

2.6.1. The bounds when data are missing completely at random

In the first simulation, 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 uncorrelated. The sample size for X and Y is $n = 100$. Then, for each given proportion $s \in \{0, 0.05, \dots, 0.30\}$ a subset of indices T in $\{1, 2, \dots, n\}$ of size $\lfloor s \cdot n \rfloor$ is

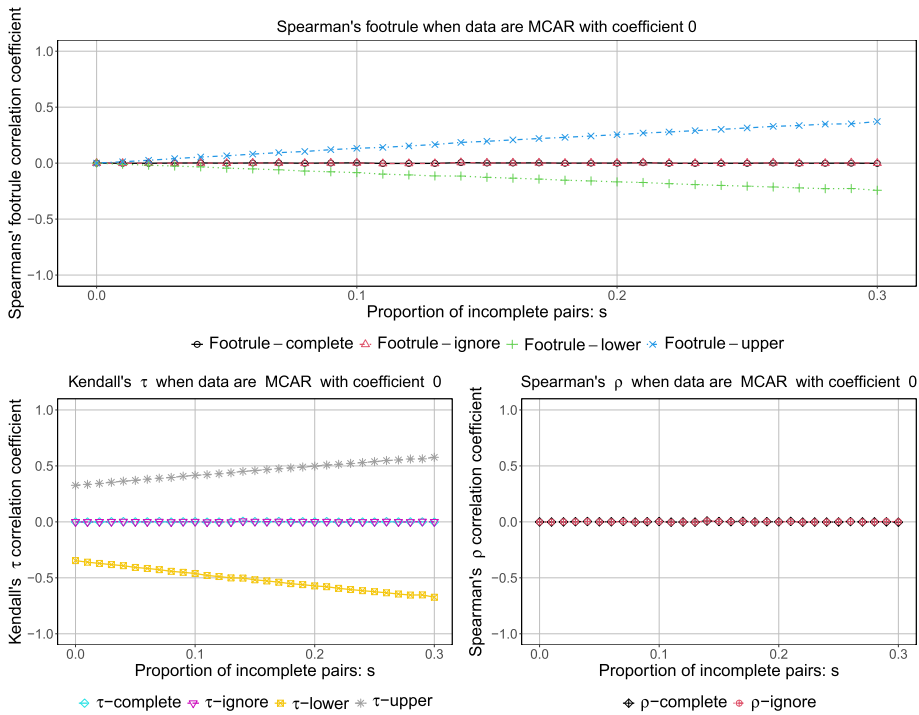


Fig 1. Rank correlation coefficients when data are missing completely at random (MCAR). X, Y are generated such that $(X, Y) \sim \text{iid } N(0, I_2)$. The methods are described in Table 6. The sample size for X and Y is $n = 100$. Results represent the average of 1000 Monte Carlo simulations.

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*.

Subsequently, the bounds of Spearman's footrule for partially observed X and Y will be computed using Theorems 2.11 and 2.22, while the bounds for Kendall's τ for partially observed data will be computed using (4). 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 1 presents the results, which are an average over 1000 Monte Carlo simulations.

Figure 1 shows that when the (uncorrelated) 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*, 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

Table 6

Description of the rank correlation coefficients or their bounds, often in the presence of missing data, shown in Figure 1 – 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.

−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.

2.6.2. 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 the Supplementary Material. Here, the index set $T \subset \{1, \dots, n\}$ of the missing components is chosen now depends on the values of X and Y . Let $q = \sum_{i=1}^n I(X(i)Y(i) > 0)$ 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 :

$$p(i \in T \mid C = \lfloor s \cdot n \rfloor) = \begin{cases} \min\{1, sn/q\}, & \text{if } X(i)Y(i) > 0, \\ \max\{0, (sn - q)/(n - q)\}, & \text{otherwise,} \end{cases} \tag{7}$$

for any given s . If $sn < q$, then $\min\{1, sn/q\} < 1$ 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 2 shows that when the data are MNAR with the missingness mechanism specified in (7), then *Footrule-ignore*, *τ -ignore* and *ρ -ignore* produce biased estimates of their rank correlation 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 1. In Section G.2.4 of the Supplementary Material we repeat this experiment using a different MNAR missing mechanism, with similar results.

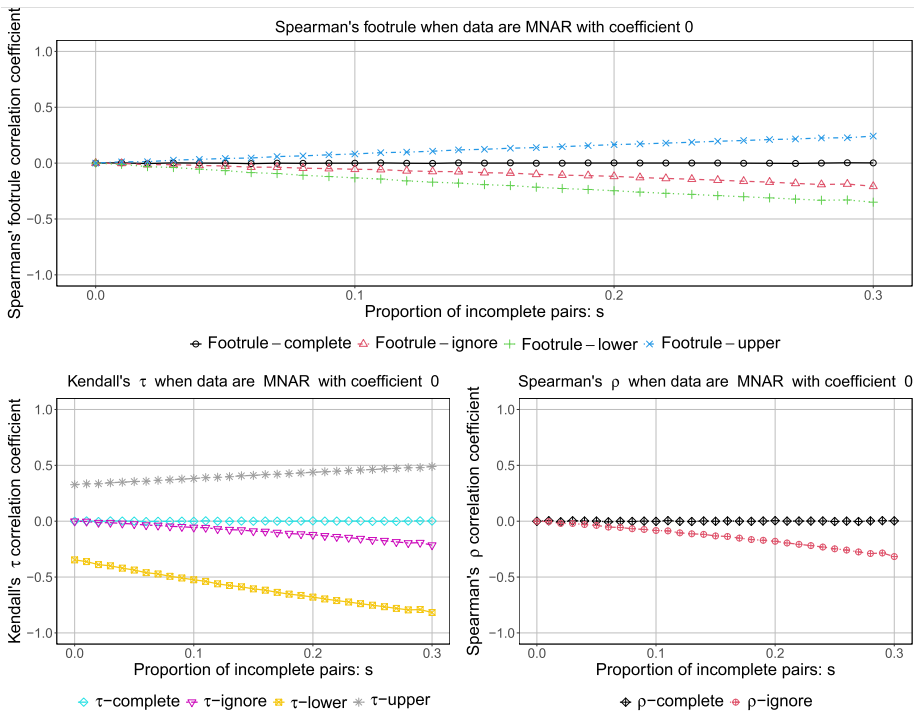


Fig 2. Rank correlation coefficients when data are missing completely at random (MNAR). X, Y are generated such that $(X, Y) \sim_{iid} N(0, I_2)$. The methods are described in Table 6. The sample size for X and Y is $n = 100$. Results represent the average of 1000 Monte Carlo simulations.

2.7. Asymptotic behaviour of the bounds

An advantage of our approach is that the bounds are tight without requiring any assumptions about the missing data. Although Algorithm 2 above and Algorithm 5 in Section E.4 of the Supplementary Material provide a way to compute $D_{\min}(X, Y)$ and $D_{\max}(X, Y)$ for $X, Y \in \mathcal{V}^n$ in $O(n^3)$, there is no closed-form solution for either of these expressions. While this makes it challenging to investigate the asymptotic behaviour of $D_{\min}(X, Y)$ and $D_{\max}(X, Y)$ analytically, we can investigate their asymptotic behaviour empirically for an increasing sample size n and a fixed proportion of missing data s .

Figures 31-36 in Section G.3 of the Supplementary Material appear to show that $D_{\min}(X, Y)$ and $D_{\max}(X, Y)$ converge or are fixed for different proportions of missing values $s \in \{0.05, 0.1\}$ and different correlation coefficients.

More importantly, empirical results in the next section show that the power of the proposed method, which depends on these bounds, converges to 1 as the sample size n increases, all while controlling the Type I error; see Figure 6 and similar figures in Section G.2.2 of the Supplementary Material.

3. Independence testing with missing data

We now explore the use of the bounds of Spearman's footrule derived above for independence testing in the presence of missing data. The core idea is to use the bounds for the statistic

Table 7

Description of the independence testing methods, often in the presence of missing data, shown in Figure 3 – 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 [1].
Alvo and Cabilio's τ	Using p -value of estimate of Kendall's τ coefficient, from [1].

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

to obtain bounds for the p -value, and then to reject the null hypothesis when *all possible* p -values will be significant.

More specifically, our method uses the bounds of the Spearman's footrule statistic to compute bounds for the possible p -values, given the missing data. In other words, if the missing data could have been fully observed, the resultant p -value would be within the computed bounds. Then, for a given significance level α , if the bounds show that all possible p -values are smaller than α , the null hypothesis is rejected. On the other hand, if at least one p -value is not significant, then the null hypothesis fails to be rejected. Our motivation for this approach is discussed in the introduction.

Below we will first show how our method bounds the p -values of Spearman's footrule in the presence of missing data using the bounds of the Spearman's footrule statistic. Then we will perform numerical simulations for investigating the Type I error and statistical power of our method. The unique contribution of our method is that, unlike all other existing methods, it provides an independence testing approach that controls the Type I error regardless of the missing data mechanism. The only assumption we make about the data is that the values are all distinct in order to avoid dealing with the issue of ties.

3.1. Bounds of p -values in the presence of missing data

For deriving the bounds of the p -value of Spearman's footrule with missing data, we begin by studying the distribution of Spearman's footrule statistic under the null hypothesis that X and Y are generated by independent continuous random variables *without* missing data.

A well-known result [13, Theorem 1] shows that under the null hypothesis, the distribution of Spearman's footrule statistic for n pairs of observations approximately follows a normal distribution with mean and variance equal to $n^2/3$ and $2n^3/45$, respectively. A well-known result [13, Theorem 1] shows that under the null hypothesis, the distribution of Spearman's footrule statistic for n pairs of observations approximately follows a normal distribution.

Exact values for mean and variance of this normal distribution can be computed to be $(n+1)(n-1)/3$ and $(n+1)(2n^2+7)/45$, respectively [47]. The accuracy of the normal approximation is investigated in [36], showing that when the sample size is at least $n = 40$, this 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 .

Suppose the sample size n is large enough for employing normal approximation, e.g. $n \geq 40$. Denote F_n as the cumulative distribution function of a normal distribution with mean equal to $(n+1)(n-1)/3$ and variance equal to $(n+1)(2n^2+7)/45$. The p -value of Spearman's footrule is then calculated as

$$p(D(X, Y)) = 2 \cdot \min\{F_n(D(X, Y)), 1 - F_n(D(X, Y))\}. \quad (8)$$

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

Proposition 3.1. *Suppose $X, Y \in \mathcal{V}^n$ are partially observed. Assume n is sufficiently large. Let $D_{\min}$ and $D_{\max}$ be the minimum and maximum possible values of Spearman's footrule between X and Y , respectively. Denote $p_1 = p(D_{\min})$ and $p_2 = p(D_{\max})$, where $p(\cdot)$ is defined in (8). Furthermore, define $p_{\min} = \min\{p_1, p_2\}$ and define*

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

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

Remark 7. In Proposition 3.1, “ n is sufficiently large” is used to allow the use of normal approximation for computing the p -value of Spearman's footrule.

Remark 8. The bounds of p -values of $D(X, Y)$ are tight when $(D_{\min} - \mu)(D_{\max} - \mu) \geq 0$, in other words, when the scaled Spearman's footrule are always positive, or always negative. When $(D_{\min} - \mu)(D_{\max} - \mu) < 0$, the scaled Spearman's footrule can be either positive or negative, depending on the values of imputations of the missing data. The maximum possible p -value is bounded by 1 in this case, and the test does not reject the null hypothesis.

We now describe our method for independence testing in the presence of missing data using Proposition 3.1. Given a pre-specified significance level α , our method will reject the null hypothesis if $p_{\max} \leq \alpha$, since this implies that all possible p -values are less than α and will be significant. Otherwise, our method will fail to reject the null hypothesis since $p_{\max} > \alpha$ implies that there exists imputations of the data with insignificant testing results.

Remark 9. Our proposed method is conservative by construction, since it only declares a significant result when the maximum possible p -value, considered over all possible imputations, is significant. When the proportion of missing data increases, the number of possible imputations for these missing data necessarily increases, and so the upper and lower bounds of Spearman's footrule increase and decrease, respectively. This is shown in Figures 1 and 2, which considers simulations for data that are missing completely at random and missing not at random.

However, our method controls the Type I error regardless of the values of the missing data. Since our method rejects the null hypothesis only when $p(D(X, Y))$ is significant, the probability of our method making a Type I error is no larger than the probability of making a Type I error when data are fully observed, which is approximately equal to α .

In next section, we perform various numerical simulations highlighting the peril of not taking missing data into account when data are missing not at random, and demonstrating the usefulness of our method.

3.2. Simulation results for independence testing

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

For comparison with the proposed method, we also consider 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 [39].

Two independence testing methods that take missing data into account are also considered. These methods [1] 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.

3.2.1. As the proportion of missing data increases

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 1, as described in Section 2.6.1, 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 the covariance matrix is now $\Sigma = \begin{pmatrix} 1 & \gamma \\ \gamma & 1 \end{pmatrix}$, with covariance coefficient $\gamma = 0.5$. The Supplementary Material 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 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 in Section 2.6.2.

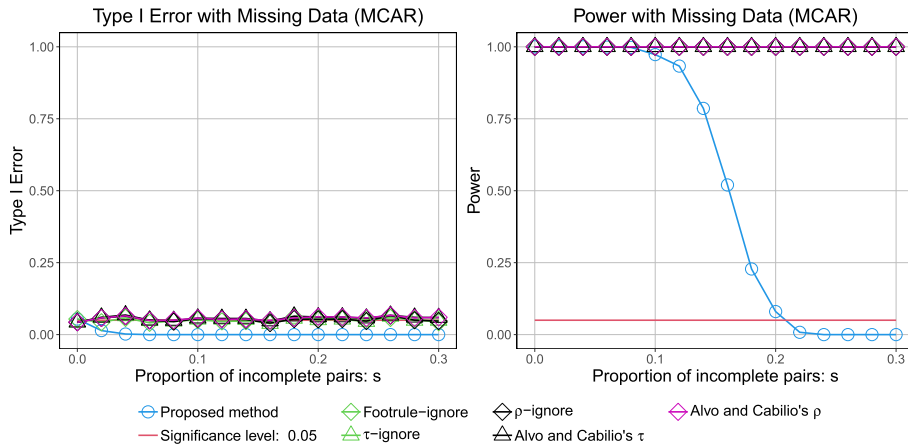


Fig 3. Statistical Type I error and power of the proposed method, as the proportion of missing data s increases, where $s \in \{0.00, 0.02, \dots, 0.30\}$. Other methods that ignore missing data, and Alvo and Cabilio's ρ and Alvo and Cabilio's τ methods are also considered. These methods are described in Table 7. The data is missing completely at random (MCAR). (Left) Type I error: $(X, Y) \stackrel{iid}{\sim} N(0, I_2)$; (Right) Power: $(X, Y) \stackrel{iid}{\sim} N(0, \Sigma)$, 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.

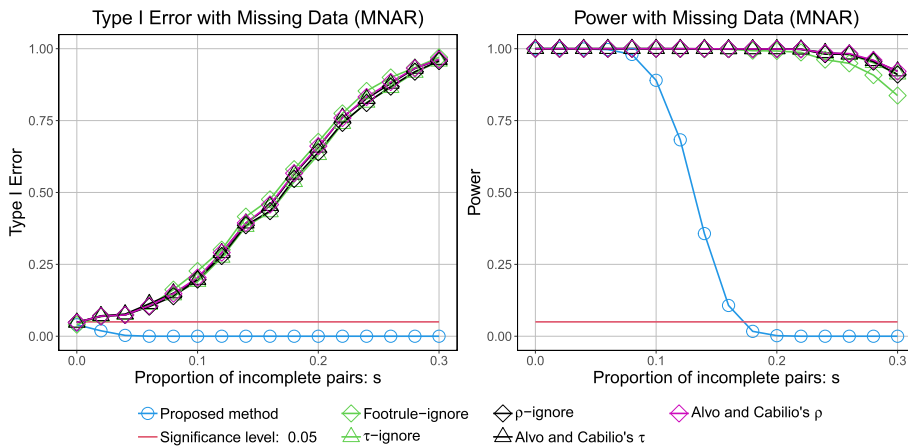


Fig 4. Statistical Type I error and power of the proposed method as the proportion of missing data s increases, where $s \in \{0.00, 0.02, \dots, 0.30\}$. Other methods that ignore missing data, and Alvo and Cabilio's ρ and Alvo and Cabilio's τ methods are also considered. These methods are described in Table 7. The data is missing completely at random (MNAR). (Left) Type I error: $(X, Y) \stackrel{iid}{\sim} N(0, I_2)$; (Right) Power: $(X, Y) \stackrel{iid}{\sim} N(0, \Sigma)$, 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 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

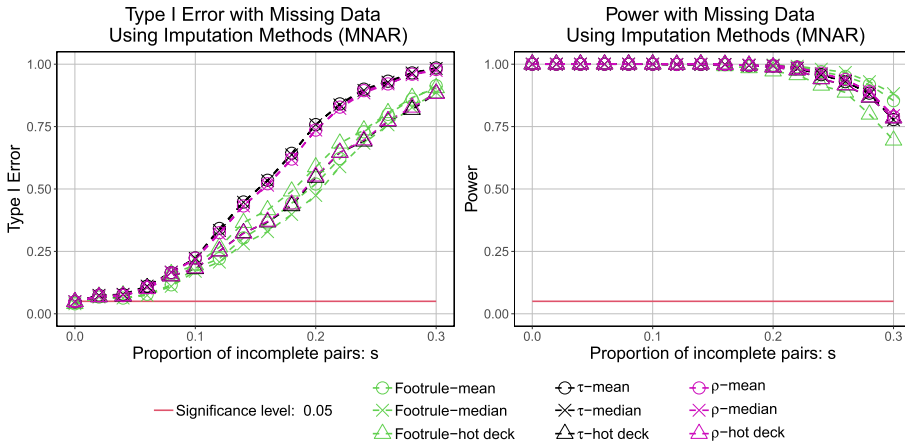


Fig 5. Statistical Type I error and power of imputation methods as the proportion of missing data s increases, where $s \in \{0.00, 0.02, \dots, 0.30\}$. The imputation methods impute missing data using either mean, median or randomly selected (hot deck) values of observed data. These methods are described in Table 7. The data is missing completely at random (MNAR). (Left) Type I error: $(X, Y) \stackrel{iid}{\sim} N(0, I_2)$; (Right) Power: $(X, Y) \stackrel{iid}{\sim} N(0, \Sigma)$, 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.

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%.

Remark 10. It is notable that the power of the proposed method decreases in Figures 3 and 4 as the proportion of missing data exceeds 10%. This is related to the conservative approach of the proposed method, as discussed in Remark 9. However, in these experiments, this is also due to the value of the covariance coefficient $\gamma = 0.5$ and the sample size $n = 200$. If we increase the covariance to $\gamma = 0.8$ or $\gamma = 0.9$, similar experiments in Section G.2.1 of the Supplementary Material show that the power only decreases after the proportion of missing data exceeds 20%. Furthermore, if we increase the sample size further, the power also increases.

3.2.2. The performance of imputation methods

We also investigate how imputation methods perform in the experiment above, when data are missing not at random. We consider using mean imputation, median imputation and hot deck imputation [39] for Spearman's footrule, Kendall's τ and Spearman's ρ . The results are shown in Figure 5, which demonstrates that these imputation methods also fail to control the Type I error rate when at least 4% of the data is incomplete.

3.2.3. As the sample size increases

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 6 shows that the proposed method

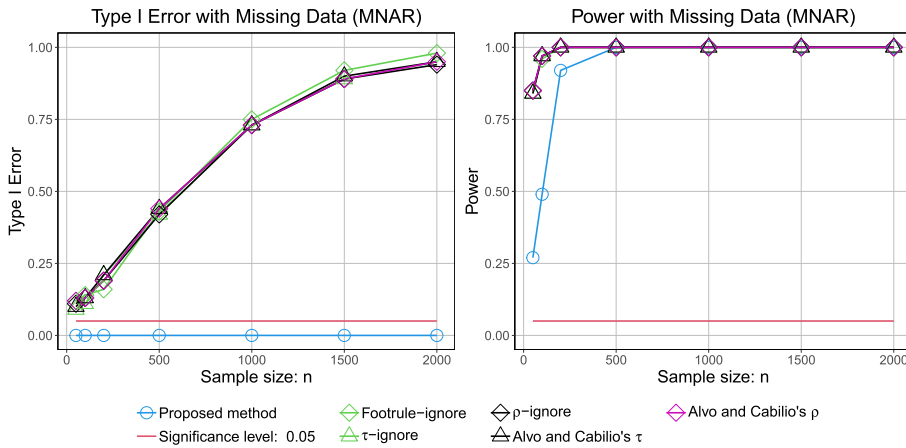


Fig 6. Statistical Type I error and power of the proposed method as the sample size n increases, where $n \in \{40, 70, 100, 200, 300, 500\}$. Other methods that ignore missing data, and Alvo and Cabilio's ρ and Alvo and Cabilio's τ methods are also considered. These methods are described in Table 7. The data is missing completely at random (MNAR). (Left) Type I error: $(X, Y) \stackrel{iid}{\sim} N(0, I_2)$; (Right) Power: $(X, Y) \stackrel{iid}{\sim} N(0, \Sigma)$, 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.

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.

3.2.4. An approximate method using random imputations

We recall that our proposed method considers all possible imputations and is therefore based on tight bounds. One may wonder how an approximation of this approach would perform, where one only considers a subset of all possible imputations. This approximate approach is considered and carefully described Section G.2.6 in the Supplementary Material. While this approach can control the Type I error in certain cases, it is heavily dependent on the number of imputations relative to the sample size and the number of missing values, and can fail to control the Type I error if an insufficient number of random imputations is used. For example, when $n = 500$ and 10% of the data are missing, then using 2000 imputations may suffice to control the Type I error. However, for a similar experiment but with $n = 750$, even 10,000 imputations are insufficient to control the Type I error; see Figure 30 of the Supplementary Material.

On the other hand, our proposed method is the exact version of this approach, and always controls the Type I error without needing to be concerned if a suitable number of imputations has been computed or not.

3.2.5. Other simulation results

More simulation results with different coefficients, sample sizes, proportions of missing data, and significance levels are provided in Section G in the Supplementary Material. In particular,

Section G.2.4 considers another missing not at random (MNAR) missingness mechanism that is different to the one specified in Equation (7). Experiments when the data are Gamma-distributed, rather than Normal-distributed are included in Section G.2.5. The results for these simulations are similar to the results shown in this section.

3.3. Application to precipitation data

To further demonstrate the usefulness of 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 and Germany from 1901 to 2024. The data are sourced from [19], 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 8. 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, respectively, with corresponding p -values of 7.42×10^{-1} , 7.66×10^{-1} , and 7.71×10^{-1} , respectively. All the results are summarized in Table 9.

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 analyzing the associations of the precipitation data between all three countries with missing data, we apply all the test methods listed in Table 7. The estimated coefficients and testing results for each method are summarized in Table 9.

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

Table 8

Precipitation per year in the United States, the United Kingdom, and France 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

Table 9
Estimated coefficients and testing results for analyzing 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.	<i>p</i> -value	est. coef.	<i>p</i> -value	est. coef.	<i>p</i> -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 7.

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 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 emphasizes 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 [2.15×10^{-6} , 1].

Acknowledgments

The authors would like to thank the anonymous referees, an Associate Editor and the Editor for their constructive comments that improved the quality of this paper.

Funding

Yijin Zeng is funded by a Roth Studentship from the Department of Mathematics, Imperial College London and the EPSRC CDT in Statistics and Machine Learning.

Supplementary Material

Supplementary Material Exact Bounds of Spearman's footrule in the Presence of Missing Data with Applications to Independence Testing

(doi: [10.1214/25-EJS2479SUPP](https://doi.org/10.1214/25-EJS2479SUPP); .pdf). Proofs and additional simulation results.

Code.

All analyses were performed using the R programming language [45, v4.4.0]. The R package `bosfr` provides an implementation for the proposed method and is available online [48].

References

- [1] ALVO, M. and CABILIO, P. (1995). Rank correlation methods for missing data. *The Canadian Journal of Statistics* **23** 345–358. [MR1378932](#)
- [2] ALVO, M. and CHARBONNEAU, M. (1997). The use of Spearman's footrule in testing for trend when the data are incomplete. *Communications in Statistics-Simulation and Computation* **26** 193–213. [MR1436281](#)
- [3] BAR-ILAN, J. (2005). Comparing rankings of search results on the web. *Information Processing & Management* **41** 1511–1519.
- [4] BAR-ILAN, J., LEVENE, M. and MAT-HASSAN, M. (2006). Methods for evaluating dynamic changes in search engine rankings: a case study. *Journal of Documentation* **62** 708–729.
- [5] BARALDI, A. N. and ENDERS, C. K. (2010). An introduction to modern missing data analyses. *Journal of School Psychology* **48** 5–37.
- [6] BENNETT, D. A. (2001). How can I deal with missing data in my study? *Australian and New Zealand Journal of Public Health* **25** 464–469.
- [7] BENNETT, V. M. (2020). Changes in persistence of performance over time. *Strategic Management Journal* **41** 1745–1769.
- [8] BRANDENBURG, F. J., GLEISSNER, A. and HOFMEIER, A. (2013). The nearest neighbor Spearman footrule distance for bucket, interval, and partial orders. *Journal of Combinatorial Optimization* **26** 310–332. [MR3071874](#)
- [9] CABILIO, P. and TILLEY, J. (1999). Power calculations for tests of trend with missing observations. *Environmetrics* **10** 803–816.
- [10] CHEN, C., XU, W., ZHANG, W., ZHU, H. and DAI, J. (2023). Asymptotic properties of Spearman's footrule and Gini's gamma in bivariate normal model. *Journal of the Franklin Institute* **360** 9812–9843. [MR4620708](#)
- [11] COOK, T. and ZEA, R. (2020). Missing data and sensitivity analysis for binary data with implications for sample size and power of randomized clinical trials. *Statistics in Medicine* **39** 192–204. [MR4041900](#)

- [12] DEMPSTER, A. P., LAIRD, N. M. and RUBIN, D. B. (1977). Maximum Likelihood from Incomplete Data via the *EM* Algorithm. *Journal of the Royal Statistical Society: Series B (Methodological)* **39** 1–22. [MR0501537](#)
- [13] DIACONIS, P. and GRAHAM, R. L. (1977). Spearman's Footrule as a Measure of Disarray. *Journal of the Royal Statistical Society: Series B (Methodological)* **39** 262–268. [MR0652736](#)
- [14] DONG, Y. and PENG, C. Y. J. (2013). Principled missing data methods for researchers. *SpringerPlus* **2** 1–7.
- [15] DWORK, C., KUMAR, R., NAOR, M. and SIVAKUMAR, D. (2001). Rank Aggregation Methods for the Web. In *Proceedings of the 10th International Conference on World Wide Web* 613–622.
- [16] FAGIN, R., KUMAR, R. and SIVAKUMAR, D. (2003). Comparing Top k Lists. *SIAM Journal on Discrete Mathematics* **17** 134–160. [MR2033311](#)
- [17] GENEST, C., NEŠLEHOVÁ, J. and BEN GHORBAL, N. (2010). Spearman's footrule and Gini's gamma: a review with complements. *Journal of Nonparametric Statistics* **22** 937–954. [MR2738876](#)
- [18] GOLDBERG, S. B., BOLT, D. M. and DAVIDSON, R. J. (2021). 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** e26749.
- [19] HARRIS, I., OSBORN, T. J., JONES, P. and LISTER, D. (2020). Version 4 of the CRU TS monthly high-resolution gridded multivariate climate dataset. *Scientific data* **7** 109.
- [20] HEYMANS, M. W. and TWISK, J. W. R. (2022). Handling missing data in clinical research. *Journal of Clinical Epidemiology* **151** 185–188.
- [21] HOROWITZ, J. L. and MANSKI, C. F. (2000). Nonparametric analysis of randomized experiments with missing covariate and outcome data. *Journal of the American Statistical Association* **95** 77–84. [MR1803142](#)
- [22] KENDALL, M. G. (1948). *Rank correlation methods.*, 4 ed. Griffin, London. [MR1527612](#)
- [23] KIM, B. S., RHA, S. Y., CHO, G. B. and CHUNG, H. C. (2004). Spearman's footrule as a measure of cDNA microarray reproducibility. *Genomics* **84** 441–448.
- [24] KUMAR, R. and VASSILVITSKII, S. (2010). Generalized Distances between Rankings. In *Proceedings of the 19th International Conference on World Wide Web* 571–580.
- [25] LIN, S. (2010). Rank aggregation methods. *Wiley Interdisciplinary Reviews: Computational Statistics* **2** 555–570.
- [26] LITTLE, R. J. A. and RUBIN, D. B. (2019). *Statistical Analysis with Missing Data* **793**, 3 ed. John Wiley & Sons, Hoboken. [MR1925014](#)
- [27] LOUKAS, S. and PAPAIOANNOU, T. (1991). Rank correlation inequalities with ties and missing data. *Statistics & Probability Letters* **11** 53–56. [MR1093418](#)
- [28] LUIGI CONTI, P. and NIKITIN, Y. (1999). 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.
- [29] MADLEY-DOWD, P., HUGHES, R., TILLING, K. and HERON, J. (2019). The proportion of missing data should not be used to guide decisions on multiple imputation. *Journal of Clinical Epidemiology* **110** 63–73.
- [30] PAPAIOANNOU, T. and LOUKAS, S. (1984). Inequalities on Rank Correlation with Missing Data. *Journal of the Royal Statistical Society: Series B (Methodological)* **46** 68–71. [MR0745217](#)

- [31] PARZEN, M., LIPSITZ, S., METTERS, R. and FITZMAURICE, G. (2010). Correlation when data are missing. *Journal of the Operational Research Society* **61** 1049–1056.
- [32] POWELL, T. C. and REINHARDT, I. (2010). Rank friction: an ordinal approach to persistent profitability. *Strategic Management Journal* **31** 1244–1255.
- [33] PÉREZ, A. and PRIETO-ALAIZ, M. (2016). Measuring the Dependence Among Dimensions of Welfare: A Study Based on Spearman's Footrule and Gini's Gamma. *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems* **24** 87–105. [MR3589787](#)
- [34] QUADE, D. and SALAMA, I. A. (2006). Concordance of Complete or Right-Censored Rankings Based on Spearman's Footrule. *Communications in Statistics-Theory and Methods* **35** 1059–1069. [MR2256263](#)
- [35] RAYKOV, T., SCHNEIDER, B. C., MARCOULIDES, G. A. and LICHTENBERG, P. A. (2014). Examining Measure Correlations with Incomplete Data Sets. *Structural Equation Modeling: A Multidisciplinary Journal* **21** 318–324. [MR3260466](#)
- [36] SALAMA, I. A. and QUADE, D. (1990). A note on spearman's footrule. *Communications in Statistics-Simulation and Computation* **19** 591–601. [MR1075464](#)
- [37] SALAMA, I. A. and QUADE, D. (2004). Agreement Among Censored Rankings using Spearman's Footrule. *Communications in Statistics-Theory and Methods* **33** 1837–1850. [MR2065177](#)
- [38] SCHAFER, J. L. (1999). Multiple imputation: a primer. *Statistical Methods in Medical Research* **8** 3–15.
- [39] SCHAFER, J. L. and GRAHAM, J. W. (2002). Missing Data: Our View of the State of the Art. *Psychological Methods* **7** 147–177.
- [40] SCHEFFER, J. (2002). Dealing with missing data. *Research Letters in the Information and Mathematical Sciences* **3** 153–160.
- [41] SEN, P., SALAMA, I. and QUADE, D. (2003). Spearman's footrule under progressive censoring. *Journal of Nonparametric Statistics* **15** 53–60. [MR1958959](#)
- [42] SMUK, M. (2015). Missing data methodology: sensitivity analysis after multiple imputation, PhD thesis, London School of Hygiene & Tropical Medicine.
- [43] SPEARMAN, C. (1906). Footrule for measuring correlation. *British Journal of Psychology* **2** 89.
- [44] SPEEVAK, T. (2017). Inequalities for sums of squares of reranked differences involving ties and missing data. *Communications in Statistics-Theory and Methods* **46** 8419–8429. [MR3680766](#)
- [45] R CORE TEAM (2021). R: A Language and Environment for Statistical Computing R Foundation for Statistical Computing, Vienna, Austria.
- [46] THABANE, L., MBUAGBAW, L., ZHANG, S., SAMAAN, Z., MARCUCCI, M., YE, C., THABANE, M., GIANREGORIO, L., DENNIS, B. and KOSA, D. E. A. (2013). A tutorial on sensitivity analyses in clinical trials: the what, why, when and how. *BMC medical research methodology* **13** 1–12.
- [47] URY, H. K. and KLEINECKE, D. C. (1979). Tables of the distribution of Spearman's footrule. *Journal of the Royal Statistical Society Series C: Applied Statistics* **28** 271–275.
- [48] ZENG, Y. (2025). bosfr: Computes Exact Bounds of Spearman's Footrule with Missing Data R package version 0.1.0.

- [49] ZENG, Y., ADAMS, N. and BODENHAM, D. (2025). Supplement to “Exact bounds of Spearman’s footrule in the presence of missing data with applications to independence testing”. DOI: <https://doi.org/10.1214/25-EJS2479SUPP>