Principled Bayesian Modeling and Statistical Learning with Non-Representative Data in Astrophysics

a thesis presented for the degree of

Doctor of Philosophy of Imperial College London

and the

Diploma of Imperial College

by

Maximilian Autenrieth

Department of Mathematics

Imperial College

180 Queen's Gate, London SW7 2BZ

August 2023

I certify that this thesis, and the research to which it refers, are the product of my own work, and that any ideas or quotations from the work of other people, published or otherwise, are fully acknowledged in accordance with the standard referencing practices of the discipline.

Signed: _____

Copyright

The copyright of this thesis rests with the author. Unless otherwise indicated, its contents are licensed under a Creative Commons Attribution-Non Commercial 4.0 International Licence (CC BY-NC). Under this licence, you may copy and redistribute the material in any medium or format. You may also create and distribute modified versions of the work. This is on the condition that: you credit the author and do not use it, or any derivative works, for a commercial purpose. When reusing or sharing this work, ensure you make the licence terms clear to others by naming the licence and linking to the licence text. Where a work has been adapted, you should indicate that the work has been changed and describe those changes. Please seek permission from the copyright holder for uses of this work that are not included in this licence or permitted under UK Copyright Law.

Thesis advisor: Professor David A. van Dyk Maximilian Autenrieth

Principled Bayesian Modeling and Statistical Learning with Non-Representative Data in Astrophysics

Abstract

This thesis tackles the fundamental issue of non-representative data in astrophysics via the development and application of methodology within the areas of statistical machine learning, Bayesian statistics and causal inference; to efficiently handle big data, to allow for probabilistic and principled parameter estimation with proper uncertainty quantification, and to deal with systematic uncertainties and biases in the data collection process. To enable (a) statistically principled, (b) scientifically justified, and (c) computationally efficient analysis of non-representative, complex astrophysical data, this thesis provides novel general-purpose statistical methodology, and statistical methodology tailored to topical scientific problems, in cosmology and high-energy astrophysics, as grouped into three related projects hereafter:

(i) We propose a simple, statistically principled, and theoretically justified general-purpose method, *StratLearn*, to improve supervised learning when the training set is not representative, a situation known as covariate shift. Building upon a well-established methodology in causal inference, we show that the effects of covariate shift can be reduced or eliminated by conditioning on propensity scores. We demonstrate that fitting learners within strata constructed on the estimated propensity scores improves upon state-of-the-art importance weighting methods on two topical scientific tasks – conditional density estimation of galaxy redshift (photo-z), and photometric supernovae type Ia (SNIa) classification.

(ii) We improve weak lensing photo-z calibration via Bayesian hierarchical modeling of full galaxy photo-z conditional density estimates obtained within *StratLearn*. We substantially improve the galaxy tomographic bin as-

signment, and obtain almost unbiased estimates of target population means within tomographic bins.

(iii) We propose a science-driven hierarchical Bayesian framework to estimate the galaxy luminosity distribution in X-rays, combining non-representative X-ray and optical surveys. Our proposed framework accounts for incompleteness bias by incorporating an X-ray incompleteness function (estimated from simulations) and an optical incompleteness function (with parameters learned from the observed data) into the model. This allows for improved recovery of the luminosity function even with high proportions of systematic incompleteness, evaluated on simulations, and applied to data from the Chandra Deep Field South (CDFS).

Professor David A. van Dyk at Imperial College London was my primary supervisor, I also worked closely, with weekly meetings, with my second supervisors Professor Roberto Trotta at the International School for Advanced Studies (SISSA), and Professor David C. Stenning at Simon Fraser University.

To my family.

List of Publications

The thesis contains research that has been accepted for publication, research projects that will be submitted for publication after submission of the thesis, as well as ongoing work. I am the first and lead author on all three projects/papers presented. As the first/lead author, I took the lead role in the development of the statistical methodology, models, and algorithms described in the papers. I implemented the code and conducted the numerical applications of the proposed methods, and I took the lead role in summarizing and interpreting the results of these analyses, as well as the writing and documentation of the results. The papers/projects are summarized via:

1) Autenrieth, M., D.A. van Dyk, R. Trotta, and D.C. Stenning. 2021. Stratified learning: a general-purpose statistical method for improved learning under covariate shift. *Statistical Analysis and Data Mining: The ASA Data Science Journal*; Status: Published.

This paper describes the statistically principled framework (*StratLearn*) for learning under covariate shift, as presented in Chapter 2. The paper (Autenrieth et al., 2023) is available via (link).

2) Autenrieth, M., B. Joachimi, R. Trotta, and D.C. Stenning, D.A. van Dyk, A. H. Wright. 2023. Improved Weak Lensing Photometric Redshift Calibration via StratLearn and Hierarchical Modeling. *Monthly Notices of the Royal Astronomical Society*; Status: To be submitted.

This paper describes the proposed weak lensing calibration methodology, as presented in Chapter 3.

3) Autenrieth, M., D. A. van Dyk, V. Kashyap, A. Zezas, A. Siemiginowska, and D.C. Stenning. 2023. Estimation of Galaxy Luminosity Distributions from Incomplete X-ray and Optical Survey Data. Status: In preparation for submission.

This paper describes the novel methodology for incompleteness corrections for $\log(N) - \log(S)$ estimation in X-ray, presented in Chapter 4.

Acknowledgments

First and foremost, I would like to express my sincere gratitude to my supervisor Prof. David A. van Dyk, and to my second supervisors Prof. Roberto Trotta and Prof. David Stenning, for providing me with the opportunity to pursue this area of research. Without their immense knowledge and enthusiasm this research would have not been possible or enjoyable as it was. Thank you for the excellent guidance and continuous support.

Besides my supervisors, I would like to thank my collaborators, Dr. Vinay Kashyap, Dr. Aneta Siemiginowska, Prof. Andreas Zezas, Dr. Angus H. Wright and Prof. Benjamin Joachimi for the many insights into the world of astrophysics, and for helping me work on numerous intricate astrophysical problems. This thesis would have not been possible without these fruitful collaborations. I would like to thank the entire CHASC collaboration for the many insightful discussions and feedback on my research through several workshops and meetings. I look forward to many further collaborations. At this point, I want to express special thanks to Prof. Richard Levine for introducing me to the world of statistics through my Masters thesis, and for fully endorsing and supporting my plan on pursuing a PhD.

During the time of my PhD, I had the opportunity to visit several research institutions with extended stays, which greatly broadened my horizon both scientifically and personally. I am very thankful for these experiences. I want to thank the International School for Advanced Studies (SISSA), with special thanks to the entire TSDS group for their great hospitality, allowing me to consider SISSA a second home institution. I would like to extend my deepest appreciation to the Harvard-Smithsonian Center for Astrophysics, and the Department of Statistics & Actuarial Science at Simon Fraser University for their kind hospitality and for the support of my research. I would like to acknowledge the Marie Skodowska-Curie RISE Grant provided by

the European Commission, for financially supporting my extended stays at Harvard-Smithsonian Center for Astrophysics, and at Simon Fraser University. I would like to thank the Roth scholarship, and the Department of Mathematics at Imperial College London for providing me with funding to complete my research, and for the continued support throughout. I would like to acknowledge the support provided by the Digital Research Alliance of Canada (alliancecan.ca) and the BC DRI Group, which in part enabled the computational work presented in this thesis.

Finally, I want to thank my friends and my entire family for always supporting me in every possible way throughout my life. None of this would have been possible without them.

Mama, Papa, Laura, Sarah, Chris und Lina, ohne euch hätte ich das und ganz viel anderes nie geschafft – Danke für Alles!

Contents

1 Introduction **1**
- 1.1 Motivation and Overview . . . 1
  - 1.1.1 Astrostatistics . . . 1
  - 1.1.2 Large and Non-Representative Data – Prospects and Challenges in Astrophysics . . . 2
- 1.2 Thesis Structure and Contributions . . . 5
- 1.3 Statistical Background . . . 8
  - 1.3.1 Bayesian Hierarchical Models . . . 8
  - 1.3.2 Markov Chain Monte Carlo Methods . . . 13
  - 1.3.3 Supervised Statistical Machine Learning . . . 19
  - 1.3.4 Causal Inference – Propensity Scores . . . 20

2 Stratified Learning: A General-Purpose Statistical Method for Improved Learning under Covariate Shift **23**
- 2.1 Introduction . . . 24
- 2.2 Preliminaries and Related Literature . . . 28
  - 2.2.1 Notation: . . . 28
  - 2.2.2 Target Risk Minimization: . . . 28
  - 2.2.3 Related Literature – Importance Weighting: . . . 29
- 2.3 A New Method: StratLearn . . . 32
  - 2.3.1 StratLearn – Methodology . . . 33
  - 2.3.2 StratLearn – Technical Details . . . 35
  - 2.3.3 StratLearn – Balance Diagnostics . . . 37
- 2.4 Numerical Demonstrations on Simulated Toy Problems . . . 39
  - 2.4.1 Simulation Study – Univariate Regression . . . 39
  - 2.4.2 *StratLearn* applied to Data from UCI Repository . . . 43
- 2.5 Classification – SNIa Identification . . . 48
- 2.6 Conditional Density Estimation – Photo-z Regression . . . 57
- 2.7 Discussion . . . 61

3 Improved Weak Lensing Photometric Redshift Calibration via StratLearn and Hierarchical Modeling **62**
- 3.1 Introduction . . . 63
- 3.2 Addressing the Non-Representative Training Data . . . 66

3.2.1 Non-Representative Spectroscopic Data and Covariate Shift . . . 66
3.2.2 Related Literature – Direct Redshift Calibration 67
3.2.3 Photometric Redshift Regression 68
3.3 StratLearn-Bayes Photometric Redshift Calibration 69
3.3.1 Photo-z Conditional Densities within *StratLearn* 69
3.3.2 Tomographic Bin Assignment 71
3.3.3 Bayesian Hierarchical Modelling of Conditional Densities 72
3.4 Estimation of the Population Distribution via Inverse-Propensity Score Weighting 78
3.5 Numerical Demonstrations 80
3.5.1 Simulation Study 80
3.5.2 Improved Bin Assignment Accuracy 86
3.5.3 Improved Population Mean Estimates Accuracy 87
3.5.4 Larger Sample Size for Weak Lensing Analysis 93
3.5.5 Population Distribution Estimates 93
3.6 Discussion 95

4 Estimation of Galaxy Luminosity Distributions from Incomplete X-ray and Optical Survey Data **97**
4.1 Introduction 98
4.1.1 Scientific Motivation and Background 98
4.1.2 Challenges and Prospects 99
4.1.3 Related Literature 105
4.1.4 Contribution 106
4.2 Source Model 109
4.3 Methodology – Principled Modelling of Multiple Survey Incompleteness 113
4.3.1 Poisson Binning – Toy Gaussian Example 114
4.3.2 Methodology – Correcting for X-ray and Optical Incompleteness via Poisson Binning 117
4.3.3 Realistic X-ray Incompleteness Function via Marginalization over Nuisance Parameters 122
4.4 Combination of Source Model and Incompleteness Corrections 125
4.5 Sampling from the Posterior Distribution – Metropolis-within-Gibbs Sampler 128
4.5.1 Technical Details on Gibbs-Type Sampler 130
4.6 Simulation Study 131
4.6.1 Simulation Study – Data Generation 131
4.6.2 Simulation Study – Model Evaluation 133
4.7 CDFS Data Analysis 136

4.7.1 CDFS: Data Description . . . 138
4.7.2 CDFS: Incompleteness Function Details . . . 141
4.7.3 CDFS: Single Pareto Model . . . 143
4.7.4 CDFS: Double Pareto Model . . . 148
4.7.5 CDFS: $\log(N) - \log(S)$ Estimation . . . 149

5 Discussion **153**
5.1 *StratLearn* – A Novel General-Purpose Method . . . 153
5.2 Photo-z Calibration . . . 154
5.3 Luminosity Distribution Estimation via Incomplete Surveys . 157
5.4 Concluding Remarks . . . 159

Appendix A Additional Results for Chapter 2 **161**
A.1 Additional Results for SNIa Identification . . . 161
A.2 StratLearn applied to original SPCC Data . . . 163
A.2.1 Covariate Balance Check on Original SPCC Data . . . 165
A.3 Additional Details and Results for Conditional Photo-z Regression . . . 169
A.3.1 Additional Details for Conditional Density Estimation under the L2-loss . . . 169
A.3.2 Additional Results for Photometric Redshift Regression 170

Appendix B Additional Results for Chapter 3 **175**
B.1 Additional Details for *StratLearn*-Bayes Model . . . 175
B.2 Additional Simulation Study Details . . . 177
B.3 Additional Simulation Results . . . 181

Appendix C Additional Results for Chapter 4 **184**
C.1 Transformation of Variables . . . 184
C.2 Additional Results for the Simulation Study and CDFS Data Analysis . . . 185

References **211**

1

Introduction

1.1 Motivation and Overview

1.1.1 Astrostatistics

Striking technological advances in space-based and earth-based instrumentation and extensive new surveys enormously increase the quantity, quality, and complexity of data available in astrophysics. A comprehensive analysis of these intricate "big-data" resources offers scientists exciting opportunities to gain a deeper understanding of the structure and evolution of astronomical sources and phenomena, and ultimately, the underlying physical mechanisms of the universe.

Scientists wishing to analyse the astronomical observations available via the enormous (present and upcoming) catalogues face major challenges, given the intricacies of the available data, and the complexity of the scientific questions. Astronomical data is collected across the electromagnetic spectrum, leading to multi-domain measurements such as spectrography, time series and imaging, which commonly leads to multi-dimensional data, characterized by heteroscedastic measurement errors. With the universe being

evolving, astronomical observations cannot be repeated under identical conditions. In addition, when making observations, we are physically limited to space measurements from earth (or from space close to earth), and we cannot travel across the universe to acquire data. We thus cannot assume to obtain observations that are representative for their population, or the universe. On the contrary, scientists have to account for systematic biases in the data acquisition when analysing astronomical data sources.

Given the intricate nature of astronomical observations and the complexity of the scientific problems, an efficient and scientifically-justified analysis of the large new data sources presents significant statistical challenges and requires the development of novel, objective and principled statistical methodology.

Astrostatistics is the scientific discipline in the interweaving of astrophysics and statistics. An aim of astrostatistical research is to provide statistically principled and scientifically justified methodology tailored to the specific astrophysical problem at hand, to allow an accurate and objective interpretation of the underlying scientific phenomena. Moreover, motivated by the complex problems and observations in astrophysics, astrostatistics allows exciting opportunities to develop novel general-purpose statistical methodology with potential applications and implications to many other scientific fields, extending the boundaries of statistical thinking.

1.1.2 Large and Non-Representative Data – Prospects and Challenges in Astrophysics

The vast amount of data available via present and upcoming large-scale sky surveys paves the way for unprecedented precision and accuracy in the estimation of astrophysical model parameters, with the potential of fundamentally changing the way we understand the universe. To provide an example, the Vera C. Rubin Observatory (formerly referred to as Large Synoptic Survey Telescope (Abell et al., 2009)), with its 10-year Legacy Survey of Space and Time (LSST), one of several upcoming large astronomical surveys, is forecast to image several billions of astronomical objects over the next 10 years, delivering around 500 petabyte of images and data products (Ivezić et al.,

2019). This will drastically increase the size of astronomical objects/data available for scientific analysis. To put this into some perspective, around 1000 supernova type Ia (a type of exploding star of high scientific interest, as detailed in Chapter 2) have been confirmed up to this date. Via LSST the number of prospective SNIa is expected to raise up to $\sim 10^5$. The availability of such large data allows minimization of statistical errors, enabling precise estimation of (astrophysical) parameters of interest.

However, with decreasing statistical errors, careful consideration of systematic errors becomes indispensable. For instance, ignoring non-representativeness of the selected (observed) data, or neglecting systematic measurement errors might lead to strongly biased parameter estimates, which would systematically mislead our understanding of the universe. We note that throughout this thesis, we refer to "non-representative" data, as the general case where observed data is not a random sample of the underlying population of interest.

While large data allows minimization of statistical errors, it does not mitigate systematic effects introduced via non-representative data. On the contrary, large data might aggravate problems arising through non-representative data. In Figure 1.1, we sketch the potential (negative) effects of non-representative data on parameter inference. In the left panel of Figure 1.1, we first sketch the 2-dimensional parameter estimate (region within the colored ovals, being some area of where the model "suspects" the true value) for three hypothetical models, given a small non-representative data set. All three model estimates have relatively large estimation uncertainties (illustrated via the large oval regions). While not centered at the "true" value (we use quotation marks to emphasize that this is a made up sketch without any meaningful true value), all model estimates include the "truth" due to the large estimation uncertainty. On the right panel, we sketch the same scenario but assuming large non-representative data for the analysis, leading to substantially more precise model estimates with lower uncertainty (small oval areas).

The sketch illustrates various problems associated to the (large) non-representative data based estimates. Due to decreased statistical errors (smaller oval

areas) the model estimates do not cover the truth, and in fact, do not overlap anymore. One might thus trust estimates that are very confident (small uncertainty) but wrong. In addition, given the results in the right panel of Figure 1.1 one might mistakenly question if the discrepancy between the various model estimates is due to some (new) physical phenomena, instead of systematic errors. In practice, different modeling approaches are usually accompanied by various model assumptions and/or rely on different data (illustrated via the different locations of the model estimates). Model misspecifications (e.g. due to simplifications) can further aggravate the effect of non-representative data (e.g. Shimodaira (2000), Karchev et al. (2023)), thus increasing the discrepancy of various model estimates. In summary, non-representatives of the observations needs careful treatment in order to obtain accurate parameter inference in small data cases, and particularly in large data scenarios.

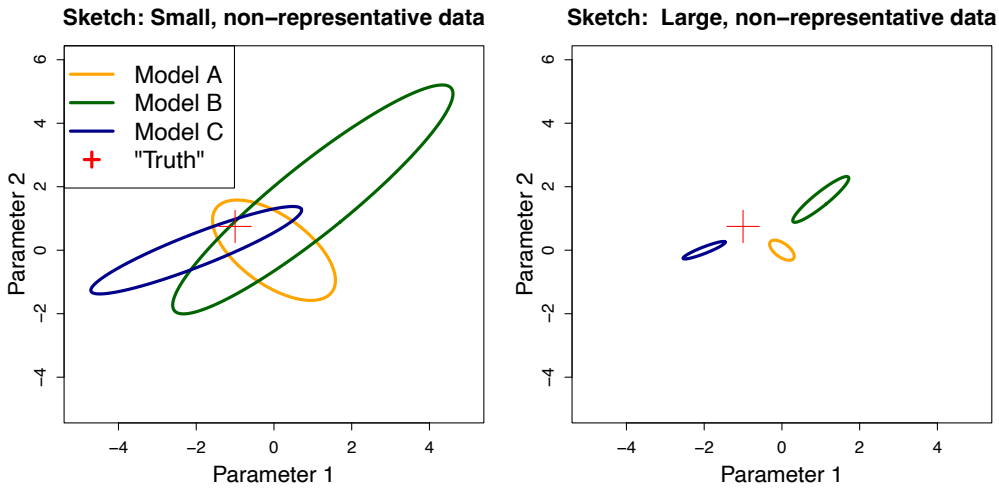


Figure 1.1: Sketch to illustrate potential effects of non-representative data on parameter inference given a small data set (left panel) vs. large data (right panel).

Non-representative data is a fundamental issue in astrophysics. There is strong selection bias in the acquisition/identification of astronomical objects. More precisely, brighter and close-by objects are more likely to be observed in astronomical surveys, commonly referred to as Malmquist bias (Malmquist, 1920). For instance, faint objects are only observed if they are in close proximity to earth, whereas very bright objects can be detected

from far distances, leading to an under-representation in the prevalence of faint sources in the observed sample compared to the underlying population. Another common source of bias is due to statistical fluctuations in measurements. More precisely, a set of objects with identical luminosities will lead to various luminosity measurements on the detector due to measurement errors (and potentially other effects). If such a scatter is close to a sensitivity limit of the detector, then only objects with measured luminosities above the sensitivity threshold will be detected, leading to a biased sample. This is a prevalent issue in astronomy and commonly known as Eddington bias (Teerikorpi, 2004). Such biases are often convoluted with additional external and detector (systematic) effects, such as weather, humidity, camera or filter imperfections, pile-up, galactic dust, etc.

The large surveys are accompanied by additional challenges. For instance, the types of astronomical objects observed via LSST are not directly identified. More precisely, most object types are obtained via automated classification based on some previously obtained training data (we provide details in Chapter 2). Again, the given training data is subject to bias and is non-representative for the majority of objects that need to be classified, potentially leading to systematic uncertainties in the identified types of the astronomical objects. Automatic classification methods need to account for the non-representativeness to obtain reliable predictions of the object types.

These, and other systematic effects need to be carefully considered and propagated through a proper statistical analysis in order to obtain accurate and precise inference on the parameters of interest.

1.2 Thesis Structure and Contributions

In this research, we tackle the fundamental issue of non-representative data in astrophysics via the development and application of methodology within the areas of statistical machine learning, Bayesian statistics and causal inference, and their combination. Statistical machine learning methods are efficient tools to handle big data. Causal inference deals with systematic uncertainties and biases in the data collection process, and to mitigate bias in parameter

estimation. Employing Bayesian methodology allows for probabilistic and principled parameter estimation with proper uncertainty quantification. To enable (i) statistically principled, (ii) scientifically justified, and (iii) computationally efficient analysis of non-representative, complex astrophysical data, this thesis provides novel general-purpose statistical methodology, and statistical methodology tailored to topical scientific problems, particularly in cosmology and high-energy astrophysics, as grouped into three related projects hereafter:

In this first project, we propose a simple, statistically principled, and theoretically justified method to improve supervised learning when the training set is not representative, a situation known as covariate shift. We build upon a well-established methodology in causal inference (Rosenbaum and Rubin, 1983), and show that the effects of covariate shift can be reduced or eliminated by conditioning on propensity scores. In practice, this is achieved by fitting learners within strata constructed by partitioning the data based on the estimated propensity scores, leading to approximately balanced covariates and much-improved target prediction. We demonstrate the effectiveness of our general-purpose method, which we name *StratLearn*, on toy examples and on two contemporary research questions in cosmology. Outperforming state-of-the-art importance weighting methods, we improve upon existing conditional density estimation of galaxy redshift (photo-z) from Sloan Data Sky Survey (SDSS) data (Sheldon et al., 2012, Izbicki et al., 2017), and we obtain the best reported AUC (0.958) on the updated "Supernovae photometric classification challenge" (Kessler et al., 2010a).

In the second project, we improve photometric redshift calibration for cosmic shear analyses. Discrepancies between cosmological parameter estimates from cosmic shear surveys (e.g., Hildebrandt et al. 2020, Troxel et al. 2018, Hikage et al. 2019) and from the recent Planck cosmic microwave background (CMB; Miville-Deschênes et al. 2020) measurements might challenge the highly successful ΛCDM model in describing the nature of the universe. To rule out systematic biases in cosmic shear survey analysis accurate redshift calibration within tomographic bins is key. In this project, we improve photo-z calibration via Bayesian hierarchical modeling of full galaxy photo-z

conditional densities. To account for systematic differences in the distribution of the spectroscopic training/source set and the photometric target set, we employ our novel methodology *StratLearn* to obtain estimates of the galaxy (object level) photo-z conditional densities. Using realistic simulations that were designed to resemble the KiDS+VIKING-450 dataset (Wright et al., 2019, 2020, Hildebrandt et al., 2020), we show that *StratLearn* estimated conditional densities allow improvement of the galaxy tomographic bin assignment. Given the improved bin assignment, our *StratLearn*-Bayesian framework leads to nearly unbiased estimates of the target population means, improving upon previously best photo-z calibration methods (Wright et al., 2020).

In the final project, we study the luminosity function, which specifies the distribution of source intensities in a population, providing information regarding the intrinsic characteristics of a class of astronomical objects and their cosmological evolution. In this project, we propose a science-driven hierarchical Bayesian framework to estimate the galaxy luminosity function in X-rays (via a broken power law model, formulated as a mixture of two Pareto distributions), combining non-representative X-ray and optical surveys. The astronomical surveys generally suffer from incompleteness in the low-intensity end, necessitating the application of incompleteness corrections generally estimated from simulations (Wright et al., 2015, Stampoulis, 2017). While this has been studied in a single band, we introduce a methodology for the application of incompleteness corrections for samples selected in two (or more) bands (e.g. combining data from optical and X-ray surveys of galaxies). We account for this incompleteness bias by incorporating an X-ray incompleteness function (estimated from simulations (Wright et al., 2015)) and an optical incompleteness function (with parameters learned from the observed data) into the model. For sources detected in X-ray and optical, we leverage the more precise optical data to estimate X-ray intensities at the location of the optical source, using a Poisson model to adjust for background contamination. The hierarchical model is fit via a Gibbs sampler. We evaluate our model on realistic simulations, mimicking data from the Chandra Deep Field Catalogue, and demonstrate unbiased recovery of the luminosity

function even with high proportions of systematic incompleteness. We apply our novel method to data from the Chandra Deep Field South (CDFS).

The remainder of this thesis is structured as follows. In Section 1.3, we provide an overview of basic statistical concepts employed and build upon in this research. In Chapter 2, we introduce our novel general-purpose method, *StratLearn*, for principled statistical learning under covariate shift. In Chapter 3, we discuss the application of *StratLearn* in combination with a hierarchical Bayesian model to improve photo-z calibration for weak lensing via full galaxy photo-z conditional density estimates. In Chapter 4, we introduce our novel framework to estimate galaxy luminosity distributions in X-rays via incomplete X-ray and optical survey data. Finally, in Chapter 5, we provide a discussion of the research conducted in this thesis, stating conclusions, limitations, and an outlook of future work. We provide additional details complementing the work presented in Chapters 2-4 via Appendices A-C at the end of this thesis (references to the respective appendices appear with prepended 'A'-'C').

1.3 Statistical Background

In this section, we provide a brief overview of the basic statistical concepts on which our research is build upon. Specifically, in Section 1.3.1, we introduce Bayesian hierarchical modeling, with Markov Chain Monte Carlo (MCMC) methods applied for Bayesian models introduced in Section 1.3.2. In Section 1.3.3, we provide a brief introduction to supervised statistical machine learning, concluding with an overview of propensity score methods in causal inference in Section 1.3.4.

1.3.1 Bayesian Hierarchical Models

Overview – Bayesian Data Analysis: Bayesian methods provide powerful and versatile tools for principled and accurate statistical inference with proper uncertainty quantification and propagation. In Bayesian data analysis inference about an unobserved parameter θ is made in terms of *probability*

statements. More precisely, we make inference about an unknown parameter θ via the posterior distribution $p(\theta|y)$, conditioning on the observed values/data y. A first step in Bayesian data analysis is to set up a *full probability model*, expressed as the joint probability distribution of the unobserved quantities θ and the observed quantities y. The joint distribution of θ and y can be written via

$$p(\theta, y) = p(y|\theta)p(\theta). \tag{1.1}$$

Specifying a full probability model can thus be divided into (i) specifying the sampling distribution $p(y|\theta)$, and (ii) specifying a prior distribution $p(\theta)$. We obtain the posterior distribution $p(\theta|y)$ via *Bayes' theorem* (the fundamental building block in Bayesian analysis), precisely

$$p(\theta|y) = \frac{p(y|\theta)p(\theta)}{p(y)} \propto p(y|\theta)p(\theta), \tag{1.2}$$

where the denominator $p(y)$ (in the astronomy literature often called the "evidence") does not depend on the parameter θ and can thus be dropped as a *normalizing* constant. The sampling distribution is commonly referred to as likelihood function $\mathcal{L}$, with $\mathcal{L}(\theta|y) = p(y|\theta)$. A fundamental principle in Bayesian statistics is that uncertainties are measured via probability distributions (Gelman et al., 1995). Via (1.2) data-appropriate models can be incorporated to accurately quantify uncertainties in scientific problems in a mathematically rigorous framework, following basic rules of probability theory. In addition, Bayesian analysis strongly facilitates the specification and interpretation of complex statistical models via the introduction of an hierarchical model framework as summarized below.

Hierarchical models: It is common in statistical applications with multiple parameters that sets of parameters are connected in some manner. In Bayesian analysis such connections can conveniently be modeled by introducing multiple model levels, referred to as a hierarchical model. For instance, the observed outcomes might be conditioned on parameters which themselves follow a common population distribution specified by additional parameters,

often referred to as hyperparameters (Gelman et al., 1995). Employing hierarchical structures has several key advantages. Non-hierarchical models mostly do not fit complex hierarchical data well. With few parameters they might not capture the full structure of the data, leading to an inaccurate fit of the data, and with too many parameters they tend to 'overfit' the given data, such that they fit the provided data 'too well', e.g. by capturing statistical noise as part of the model, which typically leads to poor generalization/predictions on new data and inaccurate parameter estimates. Hierarchical models avoid overfitting with large sets of parameters by incorporating population distributions which introduce dependencies into the parameters, thus preventing from overfitting. It is common for many hierarchical model applications that the number of model parameters is larger than the number of observed data samples.

To introduce the hierarchical model structure more formally, we split the parameter set $\theta = (\psi, \phi)$, into parameters ϕ describing the population, and ψ being considered the object-level parameters. For instance, one could consider the case of observing data $y = (y_1, \ldots, y_n)$, with n being the number of observations, and object-level parameters $\psi = (\psi_1, \ldots, \psi_n)$, where each observed outcome y_j depends on an object-level parameter ψ_j, with $j = 1, \ldots, n$, and in turn, the object-level parameters $\psi = (\psi_1, \ldots, \psi_n)$ share a common population distribution parameterized by ϕ.

The joint prior on ψ, ϕ is given via

$$p(\theta) = p(\psi, \phi) = p(\psi|\phi)p(\phi), \tag{1.3}$$

and the joint posterior distribution of ψ, ϕ given y is provided via

$$\begin{aligned} p(\psi, \phi|y) \propto & p(y|\psi, \phi)p(\psi|\phi)p(\phi) && (1.4) \\ = & p(y|\psi)p(\psi|\phi)p(\phi), && (1.5) \end{aligned}$$

where latter holds due to the conditional independence of y and ϕ given ψ, that is, the parameters (hyperparameters) ϕ affect the observed data y only through the object-level parameters ψ. Hierarchical models are thus specified

via a series of conditional distributions, breaking down the complexity of a model into several levels of simplified structures. This allows the introduction and interpretation of science-driven models via the different levels, and the combination of multiple sources of knowledge/information. For instance, in (1.4) one could model measurement errors/uncertainties on the observed data $y = (y_1, \ldots, y_n)$ via $p(y|\psi)$, given the latent object-level parameters $\psi = (\psi_1, \ldots, \psi_n)$ (e.g., considering $\psi = (\psi_1, \ldots, \psi_n)$ to be the latent, true values of $y = (y_1, \ldots, y_n)$ without errors), then $p(\psi|\phi)$ allows specification of a population distribution on the latent $\psi = (\psi_1, \ldots, \psi_n)$ without interference of any measurement errors on this second level, lastly on the third level, a prior distribution on the population distribution parameter needs to be introduced (which for instance could incorporate prior/expert knowledge about the population). We note that, as a central feature of hierarchical models, the observed data y_j can be employed to estimate (parts of) the population distribution of the ψ_j, without actually observing values for ψ_j (Gelman et al., 1995).

In many cases, we are not much interested in the distribution of each individual object-level parameter ψ_j, but we are interested in the parameters ϕ describing the population distribution. More generally, the model might include so-called nuisance parameters, that are important for the structure of the model, but there is no inferential interest regarding these parameters. In this case, the marginal posterior distribution of the parameters of interest ϕ can be obtained by integrating (marginalising) over ψ, that is

$$p(\phi|y) = p(\phi) \int p(\phi|\psi)p(y|\psi)d\psi. \tag{1.6}$$

Via the multiple hierarchy levels (e.g. the population distribution), hierarchical models pool information across objects. It is also said that the model "borrows strength", inasmuch as the posterior distribution of an object-level parameter ψ_j does not only depend on the observation y_j, but also incorporates information of the other observations via the population distribution of the object level parameters, leading to more precise object-level parameter estimates. More technically, such pooling of information is called "shrinkage",

with object-level parameter estimates being shrunk towards some population center. Such shrinkage has first been introduced via the James-Stein estimator, which has been shown to make the ordinary maximum likelihood estimator (MLE) *inadmissible* for $n > 2$ (James et al., 1961, Efron and Morris, 1973, 1975); that is, the shrinkage estimator is uniformly better in terms of mean squared error (or similar measures of discrepancy) than the MLE (which does not include population level, but only object level information) on the object-level parameter estimates. Via the multiple levels, Bayesian hierarchical models automatically induce shrinkage, as with the James-Stein estimator, however depending on the specifications of the prior distributions (Efron and Morris, 1977). Specifying "reasonable" prior distributions is thus a key part of (hierarchical) Bayesian analysis.

Prior specification: Specifying appropriate prior distributions is key in proper Bayesian analysis. Depending on the choice of priors one could interpret the Bayesian method to be either "objective" of "subjective". More precisely, one can consider different interpretations of the prior specification, which following Stenning and van Dyk (2018a) can be divided into three categories: (i) a population interpretation, with the prior distribution describing a population distribution shared by several parameters in the model (e.g., common for hierarchical model structures); (ii) interpretation of a state of knowledge, where the prior distribution incorporates subjective information, e.g. obtained via previous expert knowledge about the problem, or via ancillary scientific studies; (iii) interpretation of a lack of knowledge, e.g. corresponding to a flat prior with large variance characterizing "no prior information".

One option for an "objective" Bayesian analysis with "non-informative" prior is the implementation of *reference priors*, which under a given likelihood aim to maximize some measure of divergence between the prior and posterior distribution, e.g. to minimize information carried through the prior, or to optimize certain statistical properties of the posterior. Another choice of "non-informative" prior is using a flat (e.g. uniform) distribution that aims to minimise its influence on the statistical inference. However, such choices

of flat "non-informative" priors depend on the parametrization of the model. *Jeffreys' invariance principle* thus suggests a prior distribution that is non-informative and invariant to parameter transformations. Employing subjective priors provides a chance to incorporate expert knowledge into the model, which for instance might not directly be encoded in the data.

Particularly due to the specification of prior distributions, Bayesian methods are often labelled as subjective statistical methods. However, we note that any statistical approach is subjective, as being an idealization of the reality. Any model choice and evaluation, as well as data collection/choice, underlies the subjectivity of the statistical practitioner. We thus refer to both, "objective" and "subjective" analysis in quotation marks. In Bayesian statistics, the specification of prior distributions might add an additional level of subjectivity, with the trade-off, that after specification of the likelihood and the prior everything follows from basic probability theory (which, e.g. allows direct interpretation of probability intervals as opposed to frequentist confidence intervals). We refer to Gelman et al. (1995) and Stenning and van Dyk (2018a) for further discussion.

Bayesian hierarchical models have successfully been employed for a variety of applications in astrophysics, e.g. deriving redshift distributions of galaxies from photometric surveys (Leistedt et al., 2016), inferring cosmological parameters from cosmic shear data (Alsing et al., 2017), or from supernovae catalogues (Shariff et al., 2016), examining the solar cycle (Yu et al., 2012), or modeling globular clusters with multiple populations (Stenning et al., 2016), among many others.

1.3.2 Markov Chain Monte Carlo Methods

One challenge in Bayesian analysis is that the posterior distribution is in most cases too complex to be computed analytically. Much work has been done to numerically obtain samples from the posterior distribution instead, which allows summary statistics of the posterior distributions such as the mean, variance, percentiles, or histograms of the marginal posteriors. In many cases, with complicated multi-parameter posteriors, sampling directly

from the posterior distribution is not straightforward. In such cases Markov Chain Monte Carlo (MCMC) methods have proven to be highly beneficial, and pivotal in the development of Bayesian statistics.

Aim of MCMC methods is to produce a dependent Monte Carlo sample via a Markov chain that converges to a unique stationary distribution, coinciding with the posterior distribution. More precisely, a *Markov chain* is a sequence of random variables $\theta^{(0)}, \theta^{(1)}, \theta^{(2)}, \ldots$, with *Markov property*, that is $p(\theta^{(t)}|\theta^{(t-1)}, \theta^{(t-2)}, \ldots, \theta^{(0)}) = p(\theta^{(t)}|\theta^{(t-1)})$, which means the distribution of a value at state t only depends on the value in state $t-1$, but not on any previous values. A *stationary distribution* of a Markov chain is any distribution $f(\cdot)$ with the property

$$f(\theta^{(t)}) = \int p(\theta^{(t)}|\theta^{(t-1)}) f(\theta^{(t-1)}) d\theta^{(t-1)}. \tag{1.7}$$

While not all Markov chains have a stationary distribution, under certain conditions, one can construct a Markov chain with unique stationary distribution being the target posterior distribution. The distribution of $\theta^{(t)}$ then converges to the posterior distribution with increasing iterations t. We refer to Stenning and van Dyk (2018b) for a summary, and to Brooks et al. (2011) for additional details. For the scope of this thesis, we particularly consider two MCMC-type algorithms, the Metropolis-Hastings algorithm and the Gibbs sampler, as discussed below.

Metropolis-Hastings: The Metropolis-Hastings algorithm is a simple MCMC algorithm that produces samples which converge to the (one or higher dimensional) target posterior distribution, taking advantage of *jumping-rule distributions* (also referred to as proposal distributions or transition kernels) and the computation of rejection/acceptance rates to determine the value of subsequent states/iterations (introduced and further developed by Metropolis and Ulam 1949, Metropolis et al. 1953, Hastings 1970, Green 1995).

More precisely, let $J_t(\theta^*|\theta^{(t-1)})$ be a jumping-rule distribution that, at state t, provides a proposal θ^*, depending only on the value of the last iteration $\theta^{(t-1)}$. Given a first/initial draw $\theta^{(0)}$ from some starting distribution, the

Metropolis-Hastings algorithm follows via

1. Step: Sample θ^* from $J_t(\theta^*|\theta^{(t-1)})$.

2. Step: Compute r with

$$r = \frac{p(\theta^*|y)/J_t(\theta^*|\theta^{(t-1)})}{p(\theta^{(t-1)}|y)/J_t(\theta^{(t-1)}|\theta^*)} \tag{1.8}$$

3. Step: Set $\theta^{(t)}$ via

$$\theta^{(t)} = \begin{cases} \theta^*, & \text{with probability } \min(r, 1) \\ \theta^{(t-1)}, & \text{otherwise.} \end{cases} \tag{1.9}$$

In order to proof that the target posterior distribution is the stationary distribution of the Markov chain generated by the Metropolis-Hastings algorithm one can show that it satisfies *detailed balance* (Metropolis et al., 1953, Bhanot, 1988, Brooks et al., 2011). More precisely, a distribution $p^*(\theta)$ is said to satisfy detailed balance w.r.t. a transition probability $T(\theta_A|\theta_B)$ if

$$p^*(\theta_A)T(\theta_B|\theta_A) = p^*(\theta_B)T(\theta_A|\theta_B). \tag{1.10}$$

Detailed balance means that it is equally probable to move from a state θ_B of a Markov chain to a previous state θ_A and vice versa. Detailed balance is a sufficient (but not necessary) condition for stationarity (Bhanot, 1988, Brooks et al., 2011). Following Brooks et al. (2011), it can easily be shown that the detailed balance equation is true for the Metropolis-Hastings generated Markov chain. Assume we have some draw θ_A from the target posterior $p^*(\theta)$, and $J_t(\theta_B|\theta_A)$ is the jumping-rule distribution describing the probability of proposing θ_B at some iteration t given θ_A. In the case $\theta_A = \theta_B$ the detailed balance equation follows directly. Considering $\theta_A \neq \theta_B$ we obtain

$$p^*(\theta_A)T(\theta_B|\theta_A) = p^*(\theta_A)J_t(\theta_B|\theta_A)\min\left(\frac{p^*(\theta_B)J_t(\theta_A|\theta_B)}{p^*(\theta_A)J_t(\theta_B|\theta_A)}, 1\right) \tag{1.11}$$

$$= \min(p^*(\theta_A)J_t(\theta_B|\theta_A), p^*(\theta_B)J_t(\theta_A|\theta_B)), \tag{1.12}$$

where (1.12) follows from (1.11) by considering the two cases $p^*(\theta_A)J_t(\theta_B|\theta_A) > p^*(\theta_A)J_t(\theta_B|\theta_A)$ and $p^*(\theta_A)J_t(\theta_B|\theta_A) \leq p^*(\theta_A)J_t(\theta_B|\theta_A)$.
Finally, for $p^*(\theta_B)T(\theta_A|\theta_B)$ the expression in (1.12) can be shown analogously, which concludes the proof.

A specific version of the Metropolis-Hastings algorithm is the *Metropolis algorithm*. For the Metropolis algorithm, the jumping-rule distribution J must be symmetric, that is $J_t(\theta^*|\theta^{(t-1)}) = J_t(\theta^{(t-1)}|\theta^*)$. Typical choices for the symmetric jumping distribution are a normal distribution centered at the value of the previous iteration $\theta^{(t-1)}$, or a uniform distribution centered at $\theta^{(t-1)}$. For the Metropolis algorithm the acceptance/rejection rate is given via $r = p(\theta^*|y)/p(\theta^{(t-1)}|y)$. The more general Metropolis-Hastings algorithm is not subject to this restriction and allows non-symmetric (e.g. skewed or truncated) jumping distributions.

When constructing a Metropolis-Hastings algorithm, one aim is to obtain a possibly large *effective sample size* (ESS), which can be interpreted as the size of an *independent* sample that contains the same information about the target (posterior) distribution as the *dependent* (MCMC) sample, and is computed via

$$\text{ESS} = \frac{T}{1 + 2\sum_{k=1}^{\infty} \rho_k(\theta)}, \tag{1.13}$$

with T being the size of the (dependent) MCMC sample and $\rho_k(\theta)$ being the k-lag auto-correlation of the MCMC chain for parameter θ (e.g., Stenning and van Dyk (2018b)). In order to obtain reasonable (small) auto-correlation, leading to large ESS, two scenarios should be avoided, (i) proposals that are (almost) never accepted (e.g., due to large jumps away from the target distribution), and (ii) proposals that are (almost) never rejected (e.g., with jumps being too small). We note that sampling from (highly) skewed distributions is often challenging, e.g. leading to very slow convergence of the Metropolis-Hastings algorithm. To improve convergence, transformations of (skewed) parameters are useful, particularly if the transformed parameter distributions appear Gaussian. (More details on parameter transformations appear in Section C.1).

GIBBS SAMPLING: While direct sampling from the posterior distribution $p(\theta|y)$ is not straightforward, in some cases, sampling from the full conditional distributions $p(\theta_j|\theta_{i\neq j}, y)$ is more easily realisable (with θ_j, $j = 1, \ldots, p$, being components of θ, that is $\theta = (\theta_1, \ldots, \theta_p)$). In fact, under some conditions the full conditional distributions $p(\theta_j|\theta_{i\neq j}, y)$ completely determine the joint posterior $p(\theta|y)$ (Besag, 1974).

The Gibbs sampler (first used by Geman and Geman (1984)) separately updates the full conditional posterior distributions. More precisely, the following steps are performed

Step 1: Draw $\theta_1^{(t)}$ from $p(\theta_1|\theta_2^{(t-1)}, \ldots, \theta_p^{(t-1)}, y)$.

Step 2: Draw $\theta_2^{(t)}$ from $p(\theta_2|\theta_1^{(t)}, \theta_3^{(t-1)}, \ldots, \theta_p^{(t-1)}, y)$.

Step 3: ...

Step p: Draw $\theta_p^{(t)}$ from $p(\theta_p|\theta_1^{(t)}, \ldots, \theta_{p-1}^{(t)}, y)$.

In the Gibbs sampler, one assumes that in each step sampling from the full conditional posterior distributions is feasible, and it can be shown that the Gibbs sampler is a specific case of the Metropolis-Hastings algorithm with acceptance probability being 1 (Brooks et al., 2011). In some cases, it is not possible to directly sample from one (or several) of the full conditional densities (components) directly. These components can be updated using Metropolis-Hastings, known as the Metropolis-within-Gibbs sampler, using the terminology of Gilks et al. (1995).

In Gibbs sampling, there are mainly two approaches in selecting the order of the variables to draw from the full conditionals in each iteration, also referred to as scan order (He et al., 2016). In random scan, the ordering of the drawings (variable indices) is selected independently at random in each iteration. In systematic scan, a fixed order of variables is selected for all iterations, which often improves hardware efficiency (time needed for one iteration; He et al. 2016). Diaconis (2013) notes that in many cases the ordering strategy only has a small impact on the number of iterations needed

to obtain an accurate sample (referred to as mixing time), while He et al. (2016) demonstrates examples where the mixing time between both strategies can (either way) differ by a polynomial factor.

Many powerful extensions of the Gibbs sampler have been proposed in the literature, such as the blocked Gibbs sampler (Roberts and Sahu, 1997), which updates groups of parameters as components, or the partially collapsed Gibbs sampler, which reduces conditioning on some parameters for some components (Van Dyk and Park, 2008, Park and Van Dyk, 2009), including Metropolis-Hastings updates (Van Dyk and Jiao, 2015); among others.

Convergence diagnostics: A critical step in using MCMC methods to obtain samples from the posterior distribution is to diagnose the convergence of the MCMC chains. Precisely, it has to be examined if the MCMC chain has converged to its stationary distribution. As previously alluded, if we grow our MCMC chain $\theta_1, \theta_2, \ldots, \theta_T$ for a sufficiently large enough number of iterations T, the latter draws are approximately from the target posterior (as being the stationary distribution). This requires to discard the first part of the chain, called burn-in, of values not yet converged to the stationary distribution. In order to determine a reasonable burn-in, it is good practice to grow multiple MCMC chains with different starting values and visually examine convergence of the multiple chains, e.g. via a traceplot.

In addition, using multiple chains allows a numerical evaluation of convergence via the Gelman-Rubin $\hat{R}$ statistic (Gelman and Rubin, 1992), defined via

$$\hat{R} = \sqrt{\frac{\widehat{\text{var}}^+(\theta|y)}{W}}, \qquad \text{where } \widehat{\text{var}}^+(\theta|y) = \frac{T-1}{T}W + \frac{1}{T}B, \tag{1.14}$$

where W is the within chain variance and B is the between chain variance, with $\widehat{\text{var}}^+(\theta|y)$ being an over-estimate of the posterior variance (referring to Stenning and van Dyk (2018b) for details). Values of $\hat{R}$ close to 1 indicate reasonable convergence of the multiple chains.

1.3.3 Supervised Statistical Machine Learning

Broadly speaking, supervised statistical machine learning deals with the construction of a statistical model to predict an *output* given one or several *inputs* (Hastie et al., 2009, James et al., 2013).

More precisely, let $\mathcal{X} \subset \mathbb{R}^F$, $F > 0$, be the covariate space (the *input* space; also referred to as the feature space in some literature) and $\mathcal{Y}$ the label space (the *output* space). The labels typically consist of $K > 1$ classes, or, in multivariate regression with K dependent variables, a subset of $\mathbb{R}^K$. In a supervised learning task, let $f : \mathcal{X} \to \mathbb{R}^K$ be the training function (e.g. some statistical model), and $\ell : \mathbb{R}^K \times \mathcal{Y} \to [0, \infty)$ be the loss function comparing the output of f with the true label, y in $\mathcal{Y}$. This describes a general multivariate regression case; in a probabilistic classification task with K classes we usually have $f : \mathcal{X} \to [0, 1]^K$.

The risk function associated with our supervised learning task is $\mathcal{R}(f) := \mathbb{E}_{p(x,y)}[\ell(f(x), y)]$. The aim of supervised learning methods is to obtain accurate predictions $f(x)$, via covariates $x \in \mathcal{X}$, of the labels $y \in \mathcal{Y}$, obtained via minimization of the associated risk $\mathcal{R}(f)$. We cannot generally compute the risk $\mathcal{R}(f)$, since the exact joint distribution $p(x, y)$ is unknown. However, an approximation of the risk can be obtained by computing the empirical risk by averaging the loss on the training sample, where we have access to both, the covariates x and the labels y, e.g. via cross-validation or bootstrapping (Efron and Tibshirani, 1994, Hastie et al., 2009).

To provide some examples, a simple and commonly used loss function for regression tasks is the *squared error loss* (also referred to as L2 loss), which is the squared difference between the prediction $f(x)$ and the actual label y (evaluated for each example). Simple examples for the training function f are, for instance, linear models, or regression trees, such as random forest (Breiman, 2001). We provide further specifications of the exact choice of employed loss functions ℓ and training (model) function f in the respective parts of the thesis.

To obtain an empirical risk estimate of $\mathcal{R}(f)$, a common assumption in the

supervised learning literature is the availability of observations (x, y) drawn from $p(x, y)$. In this thesis, we deal with the scenario, that the set of available observations is not a representative sample of $p(x, y)$. More generally, we assume that observations (x', y') are available, drawn from a joint distribution $p'(x, y)$. However, given (x', y'), the aim is to minimize the risk $R(f)$ under $p(x, y)$, where $p'(x, y)$ and $p(x, y)$ are related, but different, i.e. $p'(x, y) \neq p(x, y)$. We refer to Chapter 2 for details and additional background.

1.3.4 Causal Inference – Propensity Scores

The introduction of propensity scores (Rosenbaum and Rubin, 1983) was groundbreaking in causal inference for obtaining unbiased treatment effect estimates from confounded, observational data. Rosenbaum and Rubin (1983) define the propensity score as the probability of treatment assignment given the observed covariates. In observational studies, the treatment assignment can typically be observed, but treatment is not randomly allocated. Confounding covariates, associated with both the treatment assignment and the outcome variable, can systematically bias average treatment effect estimates. Rosenbaum and Rubin (1983) show that, under certain assumptions, conditional on the propensity scores the treatment and control group have balanced covariates, eliminating or mitigating bias in treatment effect estimates.

More precisely, given a set of observed covariates X and a binary indicator Z for treatment assignment (treatment vs control), Rosenbaum and Rubin (1983) introduce the propensity score as

$$e(X) := P(Z = 1|X) \tag{1.15}$$

and define treatment allocation Z as strongly ignorable, if

$$\text{(i)} \quad (Y_1, Y_0) \perp\!\!\!\perp Z|X \qquad \text{and} \qquad \text{(ii)} \quad 0 < e(X) < 1. \tag{1.16}$$

Condition (i) means that treatment assignment Z is conditionally independent of the potential outcome $(Y_1,\ Y_0)$, given the observed covariates. The potential outcomes are the possible outcomes for an object, depending on

its treatment status, and at most one is observed, (e.g., for a treated object the observed outcome is $Y = Y_1$). In practice, condition (i) means that no confounders (covariates that are associated with the treatment and outcome) are unmeasured. Rosenbaum and Rubin (1983) show that if (1.16) holds, the propensity score is a balancing score. That is, given the propensity score, the distribution of the covariates in treatment and control are the same, i.e., $p(X|e(X), Z = 1) = p(X|e(x), Z = 0)$. Thus, conditional on the propensity score, unbiased average treatment effect estimates can be obtained, i.e., $\mathbb{E}[Y_1|e(x), Z = 1] - \mathbb{E}[Y_0|e(x), Z = 0] = \mathbb{E}[Y_1 - Y_0|e(x)]$. In practice, conditioning on the estimated (rather than true) propensity score can achieve better empirical balance as this corrects for statistical fluctuations in the sample as well (Rosenbaum and Rubin, 1983, Hirano et al., 2003).

Four main methods are used to condition on the propensity scores: inverse probability of treatment weighting (IPTW), using the propensity score for covariate adjustment, matching, and stratification on the propensity scores (Rosenbaum and Rubin, 1983, 1984). Extensive work has been done on best practice and generalization of propensity scores in causal inference, (e.g., Hirano et al., 2003, Imai and van Dyk, 2004, Lunceford and Davidian, 2004, Stuart et al., 2013, Franklin et al., 2014, Austin and Stuart, 2015, Griffin et al., 2017). Various methods have been proposed to improve the estimation of propensity scores to obtain unbiased treatment effects (e.g., McCaffrey et al., 2004, Setoguchi et al., 2008, Lee et al., 2010, Westreich et al., 2011, Mccaffrey et al., 2013, Pirracchio et al., 2014, Imai and Ratkovic, 2014, Parast et al., 2017, Griffin et al., 2017, Pirracchio and Carone, 2018, Autenrieth et al., 2021). We refer to Imbens and Rubin (2015) for an overview.

In the causal inference literature, there is an ongoing debate on the relative merits of using weighted estimators vs stratification or matching (e.g., Rubin, 2004, Lunceford and Davidian, 2004, Austin and Stuart, 2017). While, under correct specification of the propensity score model, weighting leads to consistent estimation of treatment effects, this may not hold for stratification due to potential residual confounding within strata (Lunceford and Davidian, 2004). However, the bias introduced by stratified estimators is traded with reduced variance compared to weighted estimators (Lunceford and Davidian,

2004). In addition, estimates of the propensity score are less variable than estimates of their reciprocal, or of a density ratio (to form weights). A small change in an estimated propensity score that is near zero can lead to a large difference in the computed inverse-propensity weight, causing massive variance in the estimates based on the weights (Lunceford and Davidian, 2004, Austin and Stuart, 2017).

In the astrophysics literature, propensity scores have only recently been employed (Revsbech et al., 2018), with prolific results. The approach by Revsbech et al. (2018) can be seen as a prototype of our methodology proposed in Chapter 2.

2

Stratified Learning: A General-Purpose Statistical Method for Improved Learning under Covariate Shift

Overview: We propose a simple, statistically principled, and theoretically justified method to improve supervised learning when the training set is not representative, a situation known as covariate shift. We build upon a well-established methodology in causal inference, and show that the effects of covariate shift can be reduced or eliminated by conditioning on propensity scores. In practice, this is achieved by fitting learners within strata constructed by partitioning the data based on the estimated propensity scores, leading to approximately balanced covariates and much-improved target prediction. We refer to the overall method as Stratified Learning, or *StratLearn*. We demonstrate the effectiveness of this general-purpose method on toy ex-

amples, and on two contemporary research questions in cosmology, outperforming state-of-the-art importance weighting methods. We obtain the best reported AUC (0.958) on the updated "Supernovae photometric classification challenge", and we improve upon existing conditional density estimation of galaxy redshift from Sloan Digital Sky Survey (SDSS) data.

2.1 Introduction

In supervised learning, a model is fit to a data set consisting of covariates and outcomes and used to predict the outcome variables in a second data set where only the covariates are observed. Domain shift refers to the situation where the initial data used to fit the model are systematically different from the second data set, whose outcome variables are predicted with the model. In this case, which occurs commonly in many situations, the learning model is unlikely to generalize well, leading to unreliable predictions. This work proposes a new, general method to address this problem. Following standard nomenclature, we refer to the partially observed outcome variables as labels, to the labeled data used to fit the learning model as the training or source data, and to the unlabeled data as the target data.

Domain adaptation methods aim to obtain accurate target predictions under domain shift (Pan and Yang, 2009, Quionero-Candela et al., 2009), arising in applications such as in medical imaging, where mechanical configurations may vary between medical centers (Stonnington et al., 2008, Van Opbroek et al., 2014, Kamnitsas et al., 2017); natural language processing, where annotated training data is often highly specialized and thus different from the target data (Jiang and Zhai, 2007); robotics and computer vision, where simulated and observed data is often combined to improve classification performance on the unlabeled target data (Patel et al., 2015, Hoffman et al., 2016, Tai et al., 2017, Csurka, 2017); in cyber security; epidemiology; astronomy; among others. A variety of methods designed to tackle domain shift have been proposed. Following Kouw and Loog (2019), these methods can be organized into three categories: covariate-based methods, such as subspace mappings (Fernando et al., 2013, Gong et al., 2013, Kouw et al.,

2016), finding domain-invariant spaces (Pan et al., 2010, Gong et al., 2016) and optimal transport (Courty et al., 2016, 2017); inference-based methods including minimax estimators (Liu and Ziebart, 2014, Wen et al., 2014, Chen et al., 2016) and self-learning (Bruzzone and Marconcini, 2009, Yoon et al., 2020); and sample-based approaches, with a focus on importance weighting (e.g., Shimodaira, 2000, Sugiyama et al., 2008, Bickel et al., 2009, Cortes et al., 2010), mainly in the covariate shift framework, which is our focus.

Covariate Shift: In this paper we present a new method to address covariate shift, a common case of domain shift, characterized by the fact that the conditional distribution of the labels given the predictive covariates is the same for source and target data, but the distribution of source and target covariates differ (Moreno-Torres et al., 2012). In the world of machine learning, with its high dimensional covariate spaces, the problem is both widespread and often difficult to diagnose.

Covariate shift usually occurs when the training sample is not selected at random, but is biased in terms of certain covariates. For instance, brighter astronomical objects are more likely to be better observed and therefore selected in the training set (Lima et al., 2008, Gieseke et al., 2010, Vilalta et al., 2013, Freeman et al., 2017, Revsbech et al., 2018, Allam Jr et al., 2018, Möller and de Boissière, 2020). Selection bias has been widely studied in the statistical literature (Heckman, 1979, Little and Rubin, 2019), e.g., when estimating treatment effects via observational studies, where the treatment assignment is often not random but is biased with regards to certain covariates. The transfer of causal inference techniques to domain adaptation has received more attention in recent years – both fields share the goal of obtaining accurate estimators under distribution shift (Magliacane et al., 2018, Rojas-Carulla et al., 2018, Hassanpour and Greiner, 2019). We build on this work in the present chapter, via the transfer of the propensity score framework (Rosenbaum and Rubin, 1983) from causal inference to the covariate shift setting.

Propensity scores have found wide application to related areas, such as classification with imbalanced classes (Rivera et al., 2014), fairness-aware machine

learning (Calders et al., 2013), learning from positive and unlabeled data (Bekker et al., 2019), unbiased learning-to-rank (Joachims et al., 2017), or the evaluation of recommender systems (Schnabel et al., 2016), among others. In the covariate shift framework, estimated propensity scores are used only implicitly for importance weighting (e.g., Zadrozny, 2004, Kanamori et al., 2009) analogously to IPTW, and for matching to obtain validation data (Chan et al., 2020). There has been no effort, however, to transfer the general methodology.

Contribution: We propose stratified learning (*StratLearn*), a simple and statistically principled framework to improve learning under covariate shift, based on propensity score stratification. On the theory side, we show that conditioning on estimated propensity scores eliminates the effects of covariate shift. In practice, we partition (stratify) the (source and target) data into subgroups based on the estimated propensity scores, improving covariate balance within strata. We show that supervised learning models can then be optimized within source strata without further adjustment for covariate shift, leading to reduced bias in the predictions for each stratum. *StratLearn* is general-purpose, meaning it is in principle applicable to any supervised regression or classification task under any model. We provide theoretical evidence for the effectiveness of *StratLearn*, and demonstrate it in a range of low- and high-dimensional applications. We show that the principled transfer of the propensity score methodology from causal inference to the covariate shift framework allows the statistical learning community to employ hard-won practical advice from causal inference, e.g., balance diagnostics, propensity score model assessment/selection, etc. (e.g., Rosenbaum and Rubin, 1984, Imai and van Dyk, 2004, Austin et al., 2007, Pirracchio et al., 2014). We stratify to condition on propensity scores instead of using importance weighting to avoid the massive variance sometimes associated with the latter (Rubin, 2001, 2004, Williamson et al., 2012). *StratLearn* (stratification) does not use individually estimated propensity score values except to form strata, leading to a more robust method (Rubin, 2004), as demonstrated in our numerical studies.

This chapter is organized as follows. In Section 2.2, we formally introduce our notation, and the target risk minimization task. In Section 2.2.3, we summarize related literature on covariate shift, particularly on importance weighting methods. We develop our new methodology in Section 2.3. In Section 2.4, we illustrate our method on simulated toy examples. In Section 2.5, we apply *StratLearn* to an astronomical problem that has attracted broad interest in recent years, namely Type Ia supernovae (SNIa) classification (Kessler et al., 2010b, Boone, 2019). Improving upon Revsbech et al. (2018), *StratLearn* obtains the best reported AUC[1] (0.958) on the updated "Supernovae photometric classification challenge" (SPCC) (Kessler et al., 2010a). In Section 2.6, we improve upon non-parametric full conditional density estimation of galaxy photometric redshift (so-called photo-z estimation) (Izbicki et al., 2017), a key quantity in cosmology. We conclude with a short summary of the chapter in Section 2.7.

Appendix A provides complementing results. Precisely, Section A.1 provides additional numerical results for the SNIa classification presented in Section 2.5. Section A.2 provides numerical evidence on a variation of the SPCC data (Kessler et al., 2010b) presented in Section 2.5. Finally, Section A.3 provides complementing results for the photo-z regression presented in Section 2.6.

StratLearn is computationally efficient, easy to implement and readily adaptable to various applications. Our investigations show that *StratLearn* is competitive with state-of-the-art importance weighting methods in lower dimensions and greatly beneficial for higher-dimensional applications.

[1]The AUC is the area under the Receiver Operator Characteristic (ROC) curve, obtained by plotting classifier efficiency against the false positive rate for different classification thresholds (between $[0, 1]$).

2.2 Preliminaries and Related Literature

2.2.1 Notation:

Adapting a model trained on unrepresentative source data to accurately predict the labels of the target data requires information about how the distributions of the source and target data differ. Let $\mathcal{X} \subset \mathbb{R}^F$, $F > 0$, be the covariate space that is observed for all data and let $\mathcal{Y}$ be the label space, observed only for the source data. Different domains are defined as different joint distributions $p(x, y)$ over the same joint covariate-label space $\mathcal{X} \times \mathcal{Y}$ (Kouw and Loog, 2019). Let $\boldsymbol{D}_S = \{(x_S^{(i)}, y_S^{(i)})\}_{i=1}^{n_s}$ denote the source data, a sample of size n_s from the joint distribution p_S, and let $\boldsymbol{D}_T = \{x_T^{(i)}\}_{i=1}^{n_t}$ denote the target data, an unlabelled sample of size n_t from the distribution p_T. To avoid the trivial case, we assume that $p_S(x, y) \neq p_T(x, y)$, which means the joint distribution of the covariates x and the outcome labels y differs for the source and target data. For ease of notation, we implicitly condition on an indicator variable S, with $p_S(x, y) := p(x, y|s = 1)$ representing source data (analogously $p_T(x, y) := p(x, y|s = 0)$ for target data).

Covariate shift refers to the case where the conditional distribution of the labels given the covariates is the same for all data, but the marginal distribution of the covariates differs between the source and target data. In this notation, covariate shift corresponds to $p_S(y|x) = p_T(y|x)$ but $p_S(x) \neq p_T(x)$. Thus, if an object in the source data has the same set of covariates as an object in the target data, then the conditional distribution of the outcomes of both objects is the same (i.e., $p_S(y|x) = p_T(y|x)$). However, the distribution of covariates for objects in the source data is systematically different to the distribution of covariates for objects in the target data (i.e., $p_S(x) \neq p_T(x)$).

2.2.2 Target Risk Minimization:

In our supervised learning task, let $f : \mathcal{X} \to \mathbb{R}^K$ be the training function, and $\ell : \mathbb{R}^K \times \mathcal{Y} \to [0, \infty)$ be the loss function comparing the output of f

with the true label, y in $\mathcal{Y}$.[2] In the covariate shift setting, the objective is to minimize the target risk $\mathcal{R}_T(f) := \mathbb{E}_{p_T(x,y)}[\ell(f(x), y)]$, via the labelled source data D_S and unlabelled target data D_T, assuming that $p_S(x, y) \neq p_T(x, y)$, with $p_S(y|x) = p_T(y|x)$ but $p_S(x) \neq p_T(x)$. More precisely, our task is to train a model function f that minimizes $\mathcal{R}_T(f)$, being able to compute only the source loss $\ell(f(x_S), y_S)$, but not the target loss $\ell(f(x_T), y_T)$, since y_T is unavailable in practice. Section 2.2.3 reviews importance weighting methods to minimize the target risk under covariate shift.

2.2.3 Related Literature – Importance Weighting:

Overview: In an influential work, Shimodaira (2000) proposes a weighted maximum likelihood estimation (MWLE) and shows that this MWLE converges in probability to the minimizer of the target risk under covariate shift. Following Shimodaira (2000), assuming that the support of $p_T(x)$ is contained in $p_S(x)$, the expected loss (risk) w.r.t. $p_T(x, y)$ equals that w.r.t. $p_S(x, y)$ with weights $w(x) := p_T(x)/p_S(x)$ for the loss incurred by each x,

$$\mathbb{E}_{p_T(x,y)}\left[\ell(f(x), y)\right] = \mathbb{E}_{p_S(x,y)}\left[w(x)\ell(f(x), y)\right]. \tag{2.1}$$

In short, the target risk can be minimized by weighting the source domain loss by a ratio of the densities of target and source domain covariates. The importance weights $w(x)$ are paramount in the covariate shift literature and several approaches optimize the estimation of the weights. One approach estimates the densities $p_T(x)$ and $p_S(x)$ separately (Shimodaira, 2000), e.g., through kernel density estimators (Sugiyama and Müller, 2005, Yu and Szepesvári, 2012, Baktashmotlagh et al., 2014). Others estimate the density ratio directly, e.g., via Kernel-Mean-Matching (Huang et al., 2007, Wen et al., 2015), Kullback-Leibler importance estimation (KLIEP) (Sugiyama et al., 2008, Tsuboi et al., 2009), and variations of unconstrained least-squares importance fitting (uLSIF) (Kanamori et al., 2009, 2012). Given $w(x)$, Sugiyama

[2] This describes a general multivariate regression case with K dependent variables. In a probabilistic classification task we usually have $f : \mathcal{X} \to [0, 1]^K$, with K being the number of classes.

et al. (2007) propose importance weighted cross-validation (IWCV) and show that in theory this can deliver an almost unbiased estimate of the target risk. Zadrozny (2004) links covariate shift with selection bias, and shows that the target risk can be minimised by importance sampling of source domain data, employing the inverse probability of source set assignment for importance weights. This allows any probabilistic classifier to be used to obtain the weights, e.g., logistic regression (Bickel and Scheffer, 2007).

Comparison methods: To benchmark the performance of our proposed *StratLearn* framework against existing literature, we numerically compare *StratLearn* with importance weighting (in Sections 2.4, 2.5 and 2.6), considering a range of well-established importance weight estimators. More precisely, we implement

- KLIEP – Kullback-Leibler importance estimation procedure (Sugiyama et al., 2008).
- uLSIF – Unconstrained least-squares importance fitting (Kanamori et al., 2009).[3]
- NN – Several versions of the nearest-neighbor importance weight estimator (Lima et al., 2008, Kremer et al., 2015, Loog, 2012), varying the number of neighbors.
- IPS – Importance weight estimation through probabilistic classification of source set assignment (Zadrozny, 2004, Kanamori et al., 2009).

Below we detail how the various importance weight estimation methods can be employed for different importance weighting strategies (e.g., IWCV, importance sampling, weighted least square regression (WLS), and loss weighting). Following Kanamori et al. (2009) and Bickel et al. (2009), by Bayes Theorem,

$$w(x) = \frac{p_T(x)}{p_S(x)} = \frac{p(s=1)}{p(s=0)} \frac{p(s=0|x)}{p(s=1|x)} \approx \frac{n_S}{n_T} \left(\frac{1}{p(s=1|x)} - 1 \right). \tag{2.2}$$

[3]KLIEP and uLSIF were implemented with the original author's public domain MATLAB code (link).

The importance-weighting-based estimators, KLIEP, uLSIF and NN directly estimate $w(x)$. IPS obtains an estimate of $w(x)$ by plugging in the estimated inverse probability of source set assignment[4] into the right-hand side of (2.2), which allows any probabilistic classifier to be used to obtain the weights, e.g., logistic regression (Bickel and Scheffer, 2007). The estimated weights in (2.2) can then be used for weighted empirical risk minimization (following Formula (2.1); e.g., through WLS in Section 2.4, through IWCV in Section 2.5, and through (2.19) in Section 2.6).

To implement importance sampling (such as in Section 2.5), we employ the framework proposed by Zadrozny (2004). Specifically, Zadrozny (2004) shows that for any distribution $\mathcal{D}$ with feature-label-selection space $\mathcal{X} \times \mathcal{Y} \times \mathcal{S}$ and (x, y, s) examples drawn from $\mathcal{D}$, for all classifiers f and loss functions $\ell = \ell(f(x), y)$, if we assume that $P(s = 1|x, y) = P(s = 1|x) \neq 0$, then

$$\mathbb{E}_{\mathcal{D}}\left[\ell(f(x), y)\right] = \mathbb{E}_{\hat{\mathcal{D}}}\left[\ell(f(x), y)|s = 1\right], \tag{2.3}$$

$$\text{with } \hat{\mathcal{D}}(x, y, s) := \frac{P(s = 1)}{P(s = 1|x)}\mathcal{D}(x, y, s). \tag{2.4}$$

The target risk can thus be minimised by sampling from $\hat{\mathcal{D}}$, e.g., by importance sampling of source domain data. Formula (2.4) allows us to directly use the inverse probability of source set assignment (IPS). Moreover, rearranging the terms in the formulation of $w(x)$ in the right hand side of (2.2) yields

$$p(s = 0)w(x) + p(s = 1) = \frac{p(s = 1)}{p(s = 1|x)}. \tag{2.5}$$

Then, substituting into (2.4),

$$\hat{\mathcal{D}}(x, y, s) := \frac{P(s = 1)}{P(s = 1|x)}\mathcal{D}(x, y, s) \tag{2.6}$$

$$= \big(w(x)p(s = 0) + p(s = 1)\big)\mathcal{D}(x, y, s). \tag{2.7}$$

Thus, following Formula (2.3) (Zadrozny, 2004), IPS can be used for impor-

[4]We note that the inverse probability of source set assignment is equivalent with the inverse propensity score, we thus call this estimator IPS. In this way, propensity scores have implicitly been used for importance weighting, however, without identification or transfer of the general propensity score methodology (as proposed in our work).

tance sampling via (2.6), and the direct-weight estimators KLIEP, uLSIF and NN can be used for importance sampling via (2.7).

In our numerical examples, we incorporate the estimated weights as in the corresponding benchmark publication (if existent). Specifically,
in Sections 2.4.1 and 2.4.2 the estimated weights are used for importance weighted least squares regression (Shimodaira, 2000). In Section 2.6, we incorporate the estimated weights as suggested in Izbicki et al. (2017). Following Izbicki et al. (2017), the estimated weights are used for loss weighting as in (2.1). In Section 2.5, importance weighting has not previously been applied. We implement IWCV, importance sampling, and a combination of both, to demonstrate the advantage of *StratLearn* with respect to either.

Issues of importance weighting: Although importance weighting in theory enables minimization of the target risk, there are challenges. Cortes et al. (2008) theoretically analyze the effect of errors in the weight estimation on the outcome hypothesis of the learning algorithm. Based on a measure of domain dissimilarity (e.g., Rényi divergence), Cortes et al. (2010) show that weighting leads to high generalization upper error bounds, making predictions unreliable, especially with large importance weights. In addition, Reddi et al. (2015) points out that while weighting can reduce bias, it can also greatly increase variance. Unfortunately, with increasing covariate space dimension, the variance of the importance-weighted empirical risk estimates may increase sharply (Izbicki et al., 2014, Stojanov et al., 2019). This can be partly tackled by dimensionality reduction methods (Sugiyama et al., 2010, 2011, Stojanov et al., 2019); see Kouw and Loog (2019) for a detailed discussion. We address these variance concerns via propensity scores.

2.3 A New Method: StratLearn

In this section, we introduce our novel *StratLearn* methodology for principled learning under covariate shift via the transfer of the propensity score framework from causal inference to the covariate shift setting. In Section 2.3.1, we provide theoretical justification that conditioning on propensity scores elimi-

nates covariate shift, and we devise how propensity scores can be employed in practice to mitigate the effects of covariate shift via propensity score stratification. The flowchart in Figure 2.1 outlines the *StratLearn* framework, with technical details described in Section 2.3.2. In Section 2.3.3, we describe covariate balance diagnostics which can readily be transferred from the causal inference framework to assess the propensity score model, and we introduce outcome balance diagnostics, employing additional structure in the covariate shift setting.

2.3.1 StratLearn – Methodology

In the covariate shift framework, we define the propensity score to be the probability that object i is in the source data, given its observed covariates, i.e.,

$$e(x_i) := P(s_i = 1|x_i), \text{ with } 0 < e(x_i) < 1. \tag{2.8}$$

Proposition 1 (Learning conditional on the propensity score)**.** *If $p_S(x, y)$ and $p_T(x, y)$ satisfy the covariate shift definition and $0 < e(x) < 1$, then it holds that*

$$p_T(x, y|e(x)) = p_S(x, y|e(x)). \tag{2.9}$$

That is, conditional on $e(x)$ the joint source and target distributions are the same, eliminating covariate shift. It follows, for any loss function $\ell = \ell(f(x), y)$,

$$E_{p_T(x,y|e(x))}[\ell(f(x), y)] = E_{p_S(x,y|e(x))}[\ell(f(x), y)]. \tag{2.10}$$

Verification of Proposition 1: For Proposition 1, we start from [Theorem 1] in Rosenbaum and Rubin (1983), which proves that the propensity score $e(x)$, with $0 < e(x) < 1$, is a balancing score, which for our case means

$$x \perp\!\!\!\perp s|e(x), \tag{2.11}$$

or equivalently $p_S(x|e(x)) = p_T(x|e(x))$. That is, conditional on the propensity score, the covariates in the source and target data have the same distribution. It follows that

$$\begin{aligned} p_S(x, y|e(x)) :=& p(x, y|e(x), s = 1) \\ =& p(y|x, e(x), s = 1)p(x|e(x), s = 1) \quad (2.12) \\ =& p(y|x, e(x), s = 0)p(x|e(x), s = 1) \quad (2.13) \\ =& p(y|x, e(x), s = 0)p(x|e(x), s = 0) \quad (2.14) \\ =& p(x, y|e(x), s = 0) \\ =:& p_T(x, y|e(x)). \end{aligned}$$

Step (2.13) follows from the covariate shift assumption that $p(y|x, s = 1) = p(y|x, s = 0)$, with $e(x)$ as a function of x not changing the conditional distributions. Step (2.14) follows from the balancing property of the propensity score (2.11). Thus, conditional on the propensity score, the source and target data have the same joint distribution. Equation (2.10) follows directly.

Note that the condition of Proposition 1, that is $0 < e(x) < 1$, is no stronger than the conditions required for (2.1). The support of $p_T(x)$ being contained in $p_S(x)$ implies $0 < e(x)$, and $e(x) = 1$ implies $p_T(x) = 0$, in which case the importance weight $w(x) = 0$, which is equivalent to discarding the sample.

With Proposition 1, we extend the basic causal inference theory to use propensity scores in the covariate shift framework. Conditioning on estimated propensity scores enables statistically principled minimization of the target risk based on source data. According to Proposition 1, if we were to condition on any single value of the propensity score, the distribution of x and y in the source and target domains would be identical and we could minimize their target risk using the source data alone. Because sample sizes with identical propensity scores are too small in practice for model fitting, we employ an approximation.

StratLearn takes advantage of Proposition 1 via propensity score stratification; source data $\boldsymbol{D}_S$ and target data $\boldsymbol{D}_T$ are divided into k non-overlapping subgroups (strata) based on quantiles of the estimated propensity scores.

More precisely, letting q_j be the j'th k-quantile of $\{e(x_i) : x_i \in (x_S \cup x_T)\}$, for $j \in 1, \ldots, k$, we divide D_S and D_T into

$$\begin{aligned} D_{Sj}^{(k)} &= \{(x, y) \in D_S : q_{k-j} < e(x) \leq q_{k-j+1}\} \quad \text{and} \\ D_{Tj}^{(k)} &= \{x \in D_T : q_{k-j} < e(x) \leq q_{k-j+1}\}, \end{aligned} \tag{2.15}$$

where $q_0 = 0$ and $q_k = 1$. By Proposition 1, within strata,

$$p_{T_j}(y, x) \approx p_{S_j}(y, x), \text{ for } j \in 1, \ldots, k, \tag{2.16}$$

where S_j indicates conditioning on assignment to the j'th source stratum (analogously for target T_j). It follows that for $j \in 1, \ldots, k$,

$$\mathbb{E}_{p_{T_j}(x,y)}[\ell(f(x), y)] \approx \mathbb{E}_{p_{S_j}(x,y)}[\ell(f(x), y)]. \tag{2.17}$$

Thus, we can minimize the target risk within strata by minimizing the source risk within strata. In this way, we reduce the covariate shift problem to non-overlapping subgroups where the source and target domain are approximately the same, which in principle allows us to fit any supervised learner to D_{S_j} to predict the target objects in D_{T_j}. Figure 2.1 presents a flow chart illustrating the steps of our proposed *StratLearn* methodology.

2.3.2 StratLearn – Technical Details

In general, any probabilistic classifier could be used to estimate propensity scores (e.g., Pirracchio et al. (2014)). Logistic regression is commonly used in causal inference, and we adopt it for the majority of applications in this paper. In practice, the covariate shift assumption, $p_S(y|x) = p_T(y|x)$, requires there be no unobserved confounding covariates. To meet this requirement, we include all potential confounders as main effects. An estimate of the propensity scores in (2.8) is then obtained by probabilistic classification of source assignment, given source data x_S and target data x_T.

Using the estimated propensity scores, the source and target data are grouped into strata, following (2.15). We use $k = 5$ strata based on empirical evidence

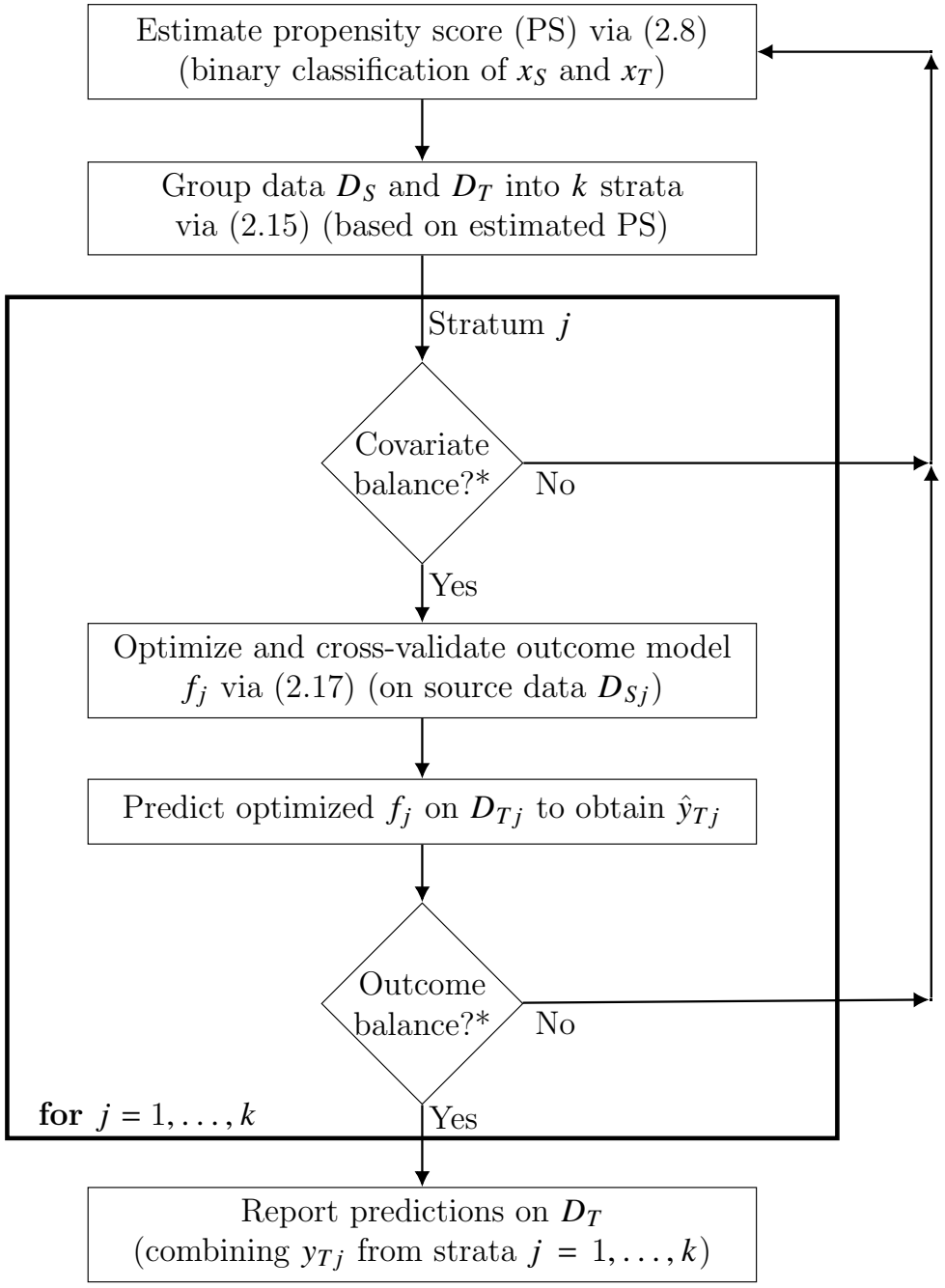


Figure 2.1: *StratLearn* flow chart. (*Covariate balance and outcome balance is assessed as described in Section 2.3.3, with a numerical example given in Section 2.5.)

provided by Cochran (1968), showing that sub-grouping into five strata is enough to remove at least 90% of the bias for many continuous distributions (Rosenbaum and Rubin, 1983). Given the stratified data, we fit a model f_j to source data $\boldsymbol{D}_{S_j}$, and predict the respective target samples in $\boldsymbol{D}_{T_j}$, for $j \in 1, \ldots, k$. Model hyperparameters for f_j can be selected through empirical risk minimization on source data $\boldsymbol{D}_{S_j}$, for instance through cross-validation on $\boldsymbol{D}_{S_j}$. The model functions f_j are trained independently and can be computed in parallel to reduce computational time. If the source distribution does not cover the target distribution well enough, some of the strata may contain too little source data to reliably train the model. In this case, we add source data from one or more adjacent strata to avoid highly variable predictions. There is a bias-variance trade-off here in that this reduction in variance requires a relaxation of the approximation in (2.16), which inevitably increases bias somewhat. Although a general and precise criterion for combining the strata is elusive (more complex models require more data and data sets of the same size may be more or less informative for the same model), we illustrate the combination of source strata in Section 2.5 (and Section 2.4.2), where one or more source strata have insufficient data.

2.3.3 StratLearn – Balance Diagnostics

A key advantage of propensity scores derived in causal inference is their covariate balancing score property (Rosenbaum and Rubin, 1983), that is, $p_S(x|e(x)) = p_T(x|e(x))$. In causal inference, this property is used to verify the propensity score model and/or the choice of covariates, x, e.g., by checking that x has the same within-strata distribution in the treatment and control groups. Employing the balancing property in the derivation of Proposition 1 allows us to take advantage of such diagnostic tools in our framework. We refer to the large literature on this (e.g., Rosenbaum and Rubin, 1984, Imai and van Dyk, 2004, Austin et al., 2007, Austin, 2011), and provide an example of such a balance check in Section 2.5 (and in Sections A.2 and 2.4.2).

In Remark 1, we detail how additional structure in the covariate shift setting

can be exploited to justify a corollary model diagnostic.

Remark 1. *In the propensity score framework of causal inference (Rosenbaum and Rubin, 1983) we have potential outcomes Y_0 and Y_1. In the covariate shift framework, the potential outcomes are identical ($Y_0 \equiv Y_1$). That is, there is no "treatment effect" from being assigned to the source or target set, though only the source data are observed ($Y_1 \equiv Y$). Now, given the propensity score $e(x)$, with $0 < e(x) < 1$, and the covariate shift condition $p(y|x, s = 1) = p(y|x, s = 0)$, source data assignment is 'strongly ignorable' (using the terminology of Rosenbaum and Rubin (1983)). It follows through Theorem 4 in Rosenbaum and Rubin (1983) that, conditional on the propensity score, source and target outcome are the same in expectation.*

In cases where labels are observed for (part of) the target group we can use Remark 1 as a model diagnostic. Although in practice the labels are mostly unobserved in the target group, they are available in our real-world scientific/experimental settings described in Sections 2.5 and 2.6. In Section 2.5 (as well as Sections 2.4.2, A.2, and A.3), we use Remark 1 to demonstrate a reduction of within-strata covariate shift (i.e., by conditioning on the propensity score).

We further demonstrate the possibility of similarly using predicted labels instead of actual labels as a model diagnostic. While the actual target labels y_T are usually not available in real-world data applications, the distribution of the model predicted outcome labels ($\tilde{y} = f(x)$) can be evaluated for source $f(x_S)$ and target $f(x_T)$. With f being a measurable function of the covariates x, and by employing the balancing property of propensity scores, it holds $p_S(f(x)|e(x)) = p_T(f(x)|e(x))$. Consequently, a discrepancy between the distributions of predicted source outcome $f(x_S)$ and predicted target outcome $f(x_T)$ is an indication of residual (covariate) shift in the source and target distribution.[5] An advantage of assessing the balance in the predicted

[5]In practice, one has to ensure that f is not overfitted on the available training (source) data sample, a standard check in supervised learning, which can readily be assessed via standard validation tools, such as cross-validation or bootstrapping of source data.

outcome $f(x)$ (in addition to covariate balance) is that $f(x)$ is designed to approximate the outcome y. Thus, a discrepancy of $f(x_S)$ and $f(x_T)$ indicates an imbalance in strongly predictive covariates, a straightforward sign for remaining, (likely) concerning confounding. We demonstrate the application of balance diagnostics via predicted labels in Section 2.5.

2.4 Numerical Demonstrations on Simulated Toy Problems

This section provides numerical evaluation of *StratLearn* with a comparison to state-of-the-art importance weighting methods on two toy simulation examples. We note that these toy examples serve mainly for illustration purposes. More precisely, in Section 2.4.1, we implement *StratLearn* on a well-known univariate regression example from Shimodaira (2000). In Section 2.4.2, we evaluate *StratLearn* under violation of the covariate shift assumption on data from the UCI repository (Dua and Graff, 2017).

2.4.1 Simulation Study – Univariate Regression

Univariate toy regression example – Linear model fit: As an illustration, we apply *StratLearn* to a simple, well known univariate toy simulation from Shimodaira (2000). Source data are simulated from $X_S \sim N(\mu_S = 0.5, \tau_S = 0.5)$ and target data from $X_T \sim N(\mu_T, \tau_T)$, independently. We investigate three values of the target parameters: Setting (i) $\mu_T = 0, \tau_T = 0.3$; (ii) $\mu_T = 0, \tau_T = 0.5$; and (iii) $\mu_T = 0.2, \tau_T = 0.5$. Setting (i) replicates Shimodaira (2000), and settings (ii) and (iii) vary the covariate shift mechanism. In all three settings, the outcome is simulated as $y = -x + x^3 + \epsilon$, with $\epsilon \sim N(0, 0.3^2)$. We independently sample X_S and X_T, each of size 100, along with the source outcome (blue circles in Figure 2.2), with the aim of predicting the target outcome (black triangles). Following Shimodaira (2000), we fit a misspecified (linear) model.

To compare the performance of *StratLearn* to ordinary least squares (OLS) and weighted least squares (WLS), employing importance weights, following Shimodaira (2000), we evaluate the target predictive performance of each

method on $m = 1000$ replicate data sets. Weights for WLS are estimated using the methods described in Section 2.2.3. We also compute the weights analytically, since the density of the target and source data are known. We assess performance with the mean squared error (MSE) of the target data predictions in each of the $m = 1000$ data sets as summarized in the boxplots in Figure 2.3.

We implement *StratLearn* following Section 2.3, by splitting the data into five equally sized strata based on logistic-regression estimated propensity scores, and fitting OLS on the source data within each stratum to predict the respective target outcomes. To ensure sufficient source data in the last stratum, source data from stratum 4 was added to fit the model for stratum 5 in cases where replicate data had fewer than 10 source samples in stratum 5.[6] A sample fit of *StratLearn* is shown in Figure 2.2 for settings (i)-(iii) (blue lines), illustrating that *StratLearn* approximates the true data generation model (grey line) well.

In setting (i), *StratLearn* performs comparably to the WLS methods, as shown by overlapping boxplot ranges in Figure 2.3, strongly improving upon OLS. In setting (ii), and especially in setting (iii), *StratLearn* led to the lowest MSE, substantially improving upon all WLS methods and OLS. The results demonstrate robustness of *StratLearn* with respect to the underlying covariate shift mechanism.

We emphasize that the simulations in this section serve as a toy example, where the one-dimensional covariate space allows a clear visualization of how *StratLearn* compares to weighted methods. As pointed out by Shimodaira (2000), fitting a correctly specified model (polynomial with degree $d = 3$) leads to a good approximation of the true model function, dispensing with the need for further covariate adjustment. However, in more realistic (high-dimensional) data setups, model misspecification is more likely, and the fitted model must not be overly complex to avoid high variability. Nevertheless, in the next paragraph, we show that allowing more complexity, by fitting a quadratic (instead of linear) model, which might better approximate the

[6] Using a minimum threshold of 5 source samples led to very similar model performance, as shown in Table 2.1, demonstrating robustness with respect to the choice of this threshold.

target data with settings (ii) and (iii), leads to similar results, underlining the advantage of *StratLearn*.

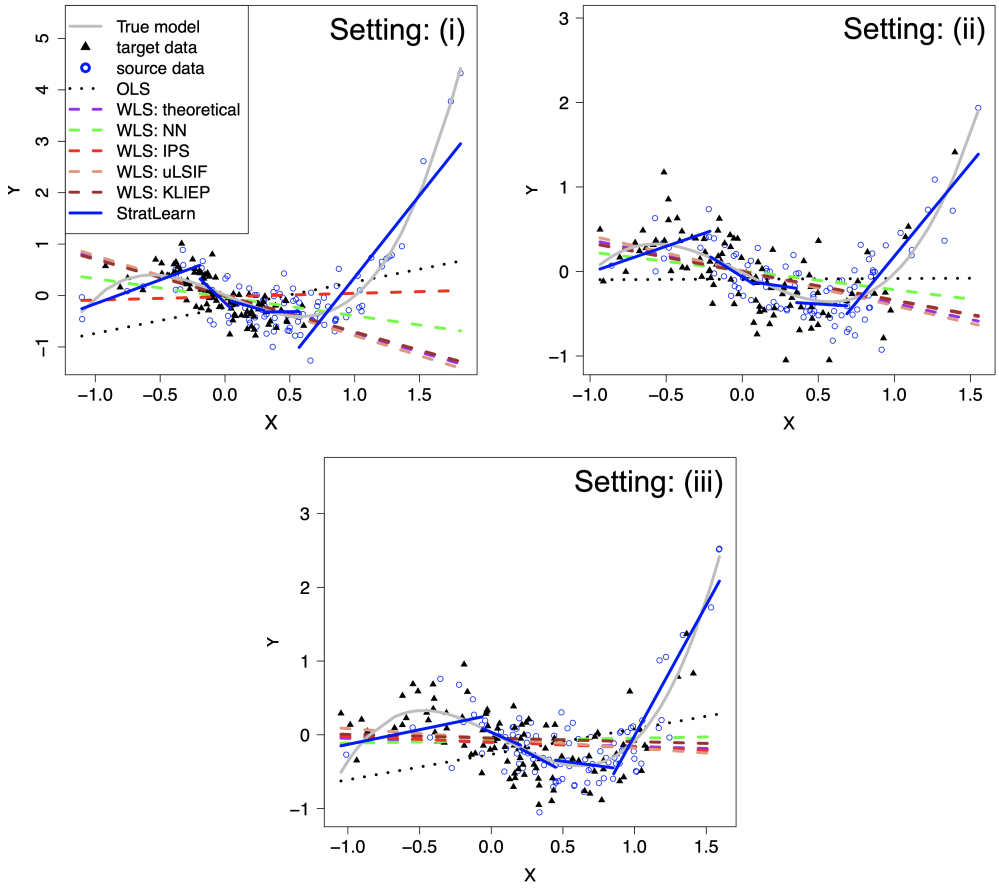


Figure 2.2: Representative linear model fit for each of the target parameter settings (i)-(iii), as discussed in Section 2.4.1.

Univariate toy regression Example – Quadratic model fit: As pointed out by Shimodaira (2000), fitting a correctly specified model (polynomial with degree $d = 3$) leads to a good approximation of the true model function, dispensing with the need for further covariate adjustment. However, in more realistic (high-dimensional) data setups, model misspecification is more likely, and the fitted model has to be not overly complex to avoid high variability. Although it may seem that *StratLearn* has an advantage for this particular example due to its flexibility in model choice, we note that the target data in setting (i) can be reasonably well approximated by

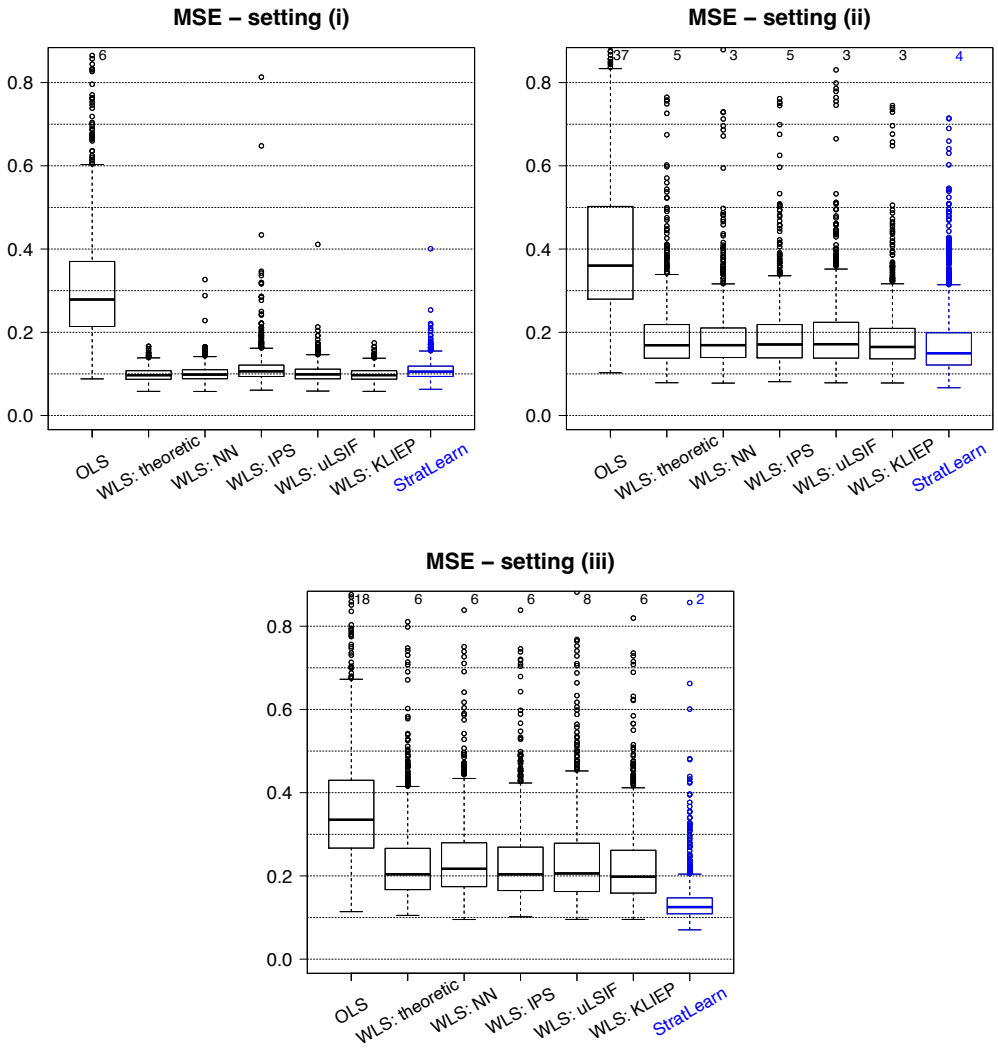


Figure 2.3: Boxplot of the target MSE in the univariate regression example with linear model fit (Section 2.4.1), obtained by using *StratLearn*, OLS, and the WLS methods on $m = 1000$ Monte Carlo simulations; settings (i) to (iii) from left to right. The small numbers at the top are the number of outliers beyond the limit of the graph. Mean and standard error of the MSE are reported in Table 2.1.

a linear model, and in settings (ii) and (iii), with more target data around $x \in (1.0, 1.5)$ where y is strongly increasing, a quadratic model (polynomial with degree $d = 2$) agrees reasonably well with the target data. Thus, *StratLearn* is not unduly advantaged relative to the non-adjusted (OLS) fit and the weighted (WLS) fits. Figure 2.5 presents boxplots of the target MSE

Table 2.1: Mean and standard error (se) of the $m = 1000$ target MSE's presented in Figures 2.3 (settings (i) to (iii) from left to right). *StratLearn (min = 10)* refers to values presented in Figure 2.3, where source data from stratum 4 was added to fit the model for stratum 5 in cases where replicate data had fewer than 10 source samples in stratum 5, to ensure that there is enough source data in the last stratum. We also add the *StratLearn* results obtained with a threshold of minimum 5 samples in source stratum 5 as a combination rule (*StratLearn (min = 5)*). Best model performances are highlighted in bold.

	Setting (i):		Setting (ii):		Setting (iii):	
Method \ MSE	Mean	(se)	Mean	(se)	Mean	(se)
OLS	0.312	(0.148)	0.415	(0.208)	0.370	(0.173)
WLS: theoretical	**0.098**	(0.016)	0.196	(0.109)	0.239	(0.131)
WLS: NN	0.101	(0.020)	0.189	(0.088)	0.244	(0.123)
WLS: IPS	0.115	(0.045)	0.196	(0.108)	0.236	(0.130)
WLS: uLSIF	0.102	(0.022)	0.197	(0.101)	0.243	(0.140)
WLS: KLIEP	**0.098**	(0.016)	0.188	(0.095)	0.229	(0.126)
StratLearn (min = 10)	0.108	(0.023)	**0.181**	(0.108)	**0.140**	(0.076)
StratLearn (min = 5)	0.113	(0.040)	**0.178**	(0.143)	**0.141**	(0.077)

for $m = 1000$ replicate data sets, fitting a quadratic model ($d = 2$) instead of a linear model ($d = 1$) as presented in the previous paragraph. Figure 2.5 demonstrates a strong advantage of *StratLearn* over the non-adjusted (OLS) fit, and over the weighted (WLS) fits, when fitting a quadratic model in settings (ii) and (iii), strengthening our previous findings that *StratLearn* leads to better covariate shift adjustment than the WLS methods in these settings. Representative quadratic model fits are displayed in Figure 2.4 for each target parameter setting (i) - (iii) (based on the same simulated data examples as presented in Figure 2.2, for comparison). Figure 2.4 illustrates that, while the WLS methods fail to fit the target data (especially on the right side of the curve), *StratLearn* well approximates the true data generation model (grey line) with quadratic model fits in each source stratum.

2.4.2 *StratLearn* applied to Data from UCI Repository

In this section, we demonstrate the performance of *StratLearn* under violation of the covariate shift assumption. We apply *StratLearn* on two datasets from the publicly available UCI repository (Dua and Graff, 2017): the "wine quality data" (Cortez et al., 2009) and the "Parkinson Telemonitoring data"

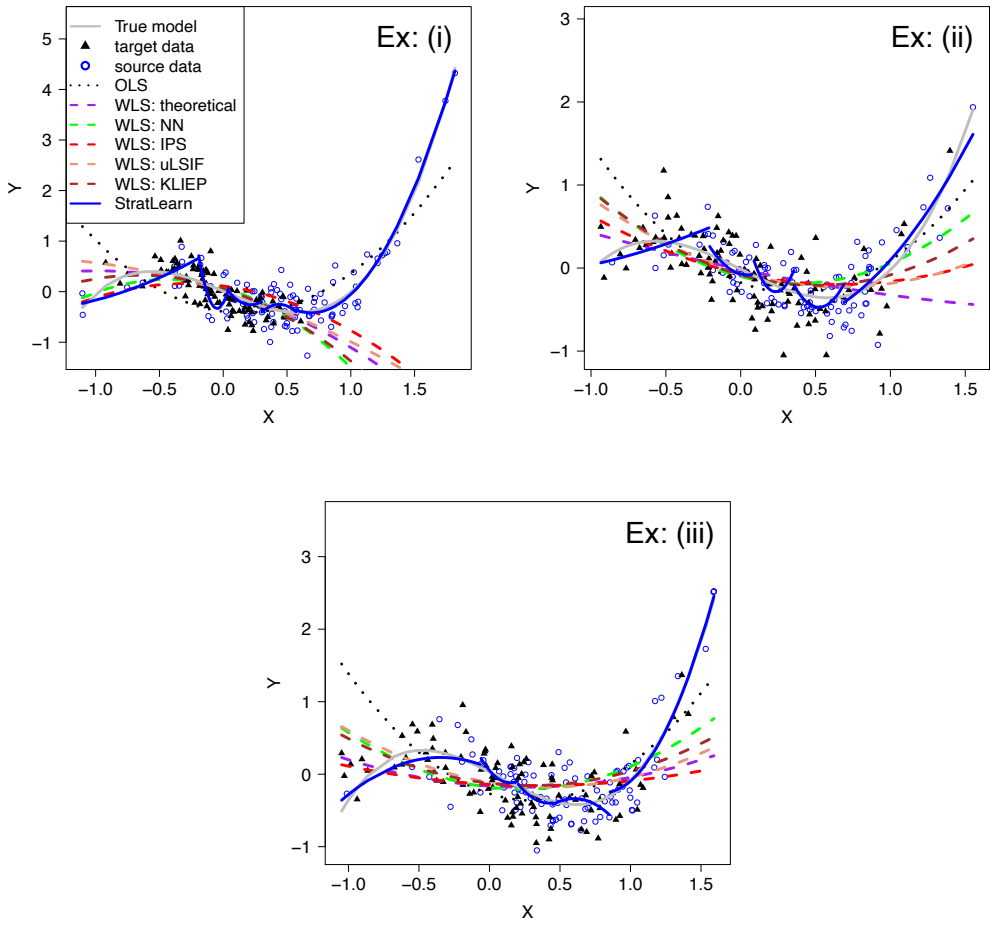


Figure 2.4: Representative quadratic model fit (Section 2.4.1) for each of the target parameter settings (i)-(iii) (Ex: (i) - (iii)). For comparison, the same simulated data set examples as presented in Figure 2.2 are displayed for all three settings.

(Little et al., 2008). These datasets have previously been used to explore regression problems in the presence of covariate shift (Chen et al., 2016).

Wine Quality Data: The data set contains 6497 samples of which 4898 are of type "white wine", used as source data $\boldsymbol{D}_S$, and 1599 samples are of type "red wine", used as the target data $\boldsymbol{D}_T$. The output is a wine "quality score" (between 0 and 10), which we aim to predict for the target data (red wine), given the scores of the source data (white wine). There are 11 physicochemical, predictive covariates. We refer to Cortez et al. (2009) for

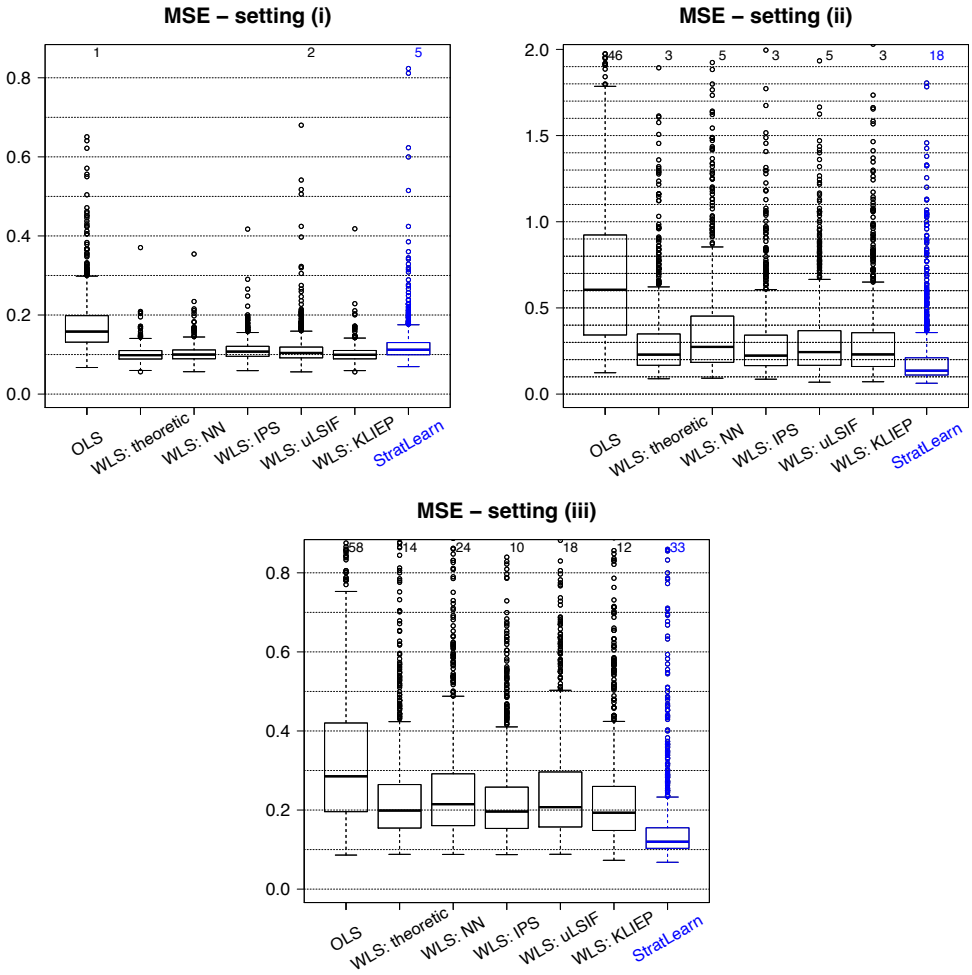


Figure 2.5: Boxplot of the target MSE in the univariate example (Section 2.4.1), fitting a quadratic model (polynomial with degree $d = 2$) instead of a linear model ($d = 1$), such as presented in Figure 2.3. The quadratic models are fitted on the same $m = 1000$ Monte Carlo simulations (settings (i) to (iii) from left to right) as presented in Figure 2.3.

detailed information.

Parkinson Data: The data comprises 5875 instances and 26 attributes. Similar to Chen et al. (2016), we take 17 of the attributes to be predictive covariates, with the aim of predicting the UPDRS Parkinson's disease symptom score (output). The source data D_S contains all instances with age below 60, leading to a source subset of size $|D_S| = 1877$. Instances with age above (inclusively) 60 and below 70 are used as target samples D_T, with $|D_T| = 2127$. As in Chen et al. (2016), instances with age above 70 are not

considered. We refer to Little et al. (2008) for further information on the data.[7]

Implementation and results on UCI data: Following Chen et al. (2016), our objective is to predict the wine "quality score" and the Parkinson's "UPDRS" score using linear regression models, specifically OLS. We assume that the source and target splitting covariates (wine type and age) are not observed for model fitting. Note that, because these are confounding covariates, this violates the covariate shift assumption. All other available covariates are used to estimate the propensity scores and the importance weights, and as predictor variables for the outcome models. To improve covariate balance on the wine quality data, we use gradient boosting machines to estimate the propensity scores, commonly used in the propensity score causal inference literature (McCaffrey et al., 2004, Mccaffrey et al., 2013, Autenrieth et al., 2021). Covariate balance is assessed as described in Section 2.3.3. On the wine quality data, we assess the covariate balance in stratum 4, which is the only stratum with both source and target samples (Table 2.2). Using logistic regression, the average (SD) absolute standardized mean difference in the observed covariate set is 0.667 (0.436), and the average (SD) of the Kolmogorov-Smirnov test statistics is 0.388 (0.154). Estimating the propensity scores using gradient boosting machines increases balance in stratum 4, by reducing the average (SD) absolute standardized mean difference to 0.561 (0.292), and the average (SD) Kolmogorov-Smirnov test statistics to 0.288 (0.109). While the balance improvement is marginal to moderate, it is consistent for both considered balance measures. In addition, there is negligible overhead in computational cost using gradient boosting machines vs. logistic regression in this example, which motivates the application of the gradient boosting machine propensity score estimator. Additional propensity score estimators are discussed in the literature and could be employed to further improve balance if required (Pirracchio et al., 2014, Parast et al., 2017, Autenrieth et al., 2021). On the Parkinson data set, logistic regression is used to estimate the propensity scores because it leads to better

[7]The data used from the UCI repository (Dua and Graff, 2017) is publicly available (UCI Wine quality data (link), Parkinsons Telemonitoring Data (link)).

covariate balance in the strata than using gradient boosting machines.

Table 2.2 presents the composition of the *StratLearn* strata for the Wine and Parkinson data, both conditioned on the estimated propensity scores. For the wine data set, all target samples are in strata 4 and 5, and only stratum 4 contains both source and target samples. Following our *StratLearn* strategy, we fit OLS on $\boldsymbol{D}_{S_4}$ to predict $\boldsymbol{D}_{T_4}$ and $\boldsymbol{D}_{T_5}$. In the Parkinson data set there is enough source data in each strata to fit OLS separately on each source stratum $\boldsymbol{D}_{S_j}$ to predict the data in the respective target stratum $\boldsymbol{D}_{T_j}$, $j \in \{1, \ldots, 5\}$.

Table 2.3 presents the MSE of the target predictions, comparing *StratLearn* with (non-adjusted) OLS ('Biased') and WLS, using the methods described in Section 2.2.3 to estimate the weights. Applying *StratLearn* results in a target MSE of 0.715 for the Wine data and 114.97 for the Parkinson data, both improving upon the Biased OLS fit which yields a target MSE of 1.024 (Wine) and 130.88 (Parkinson). *StratLearn* further provides better results than WLS:uLSIF, WLS:KLIEP and WLS:NN on both data sets; only WLS:IPS leads to slightly better results than *StratLearn*.

We also compare *StratLearn* with the method that Chen et al. (2016) develops and illustrates on these datasets. Specifically, Chen et al. (2016) proposes a "robust bias-aware regression" based on the Kullback-Leibler divergence. This proposal is tailored specifically to regression and to loss functions that account for uncertainty in the prediction (e.g., empirical logloss), whereas *StratLearn* is entirely general but can be successfully applied to this task. Following Chen et al. (2016), we compute the target empirical logloss as a performance measure. On the Wine data, *StratLearn* yields an empirical logloss of 1.271, which exactly matches the reported performance of the method proposed in Chen et al. (2016). On the Parkinson data we could not replicate the exact data split as reported in Chen et al. (2016), and thus a meaningful comparison of empirical logloss to the values reported by Chen et al. (2016) is not possible.

In summary, in these illustrative regression examples *StratLearn* demonstrates certain robustness against violation of the covariate shift assumption,

Table 2.2: Composition of the five *StratLearn* strata for the UCI wine and UCI parkinson data. The number of samples/subjects in source and target stratum, as well as the mean outcome ("quality score" and "UPDRS score") are presented.

		UCI Wine data	UCI Parkinson data
Stratum	Set	# samples (Mean "quality")	# subjects (Mean "UPDRS")
1	Source	1299 (5.98)	627 (22.00)
	Target	0 (0.00)	174 (29.15)
2	Source	1300 (5.92)	486 (26.36)
	Target	0 (0.00)	315 (25.21)
3	Source	1300 (5.93)	314 (25.53)
	Target	0 (0.00)	487 (24.86)
4	Source	999 (5.63)	269 (27.38)
	Target	301 (5.49)	532 (28.15)
5	Source	0 (0.00)	181 (27.06)
	Target	1298 (5.67)	619 (30.00)
All	Source	4898 (5.88)	1877 (24.98)
	Target	1599 (5.64)	2127 (27.58)

Table 2.3: MSE of target predictions on UCI Wine and Parkinson data, based on ordinary least squares regression (OLS), various importance weighted least squares regression methods (WLS), and our proposed *StratLearn* method.

Method \ Data	UCI wine data	UCI Parkinson data
OLS (Biased)	1.024	130.88
WLS:uLSIF	2.363	120.81
WLS:KLIEP	3.968	116.72
WLS:NN	2.377	117.47
WLS:IPS	0.660	112.80
StratLearn	0.715	114.97

improving upon the (non-adjusted) Biased fit and performing comparably with the best importance weighting approach.

2.5 Classification – SNIa Identification

Objective: Type Ia supernovae (SNIa) are invaluable for the study of the accelerated expansion history of the universe (e.g., Riess et al., 1998, Perlmutter et al., 1999). SNIa are exploding stars that can be seen at cosmological distances (billions of light years away), occurring in a particular physical

scenario which causes their intrinsic luminosity to correlate with observable properties of their light curve (apparent brightness at Earth as a function of time). This "standardizable candle" property of SNIa makes it possible to measure their distance, which in turn depends on parameters describing the physical contents of the universe.

To take advantage of this, reliable identification of SNIa based on photometric light curve (LC) data is a major challenge in modern observational cosmology. Photometric LC data are easily collectable, consisting of measurements of an astronomical object's apparent brightness (i.e. flux), filtered through different passbands (wavelength ranges), at a sequence of time points (as illustrated in Figure 2.6). Only a small subset of the objects are labelled via expensive and time-consuming spectroscopical observations, which enable SNIa identification thanks to its characteristic spectral lines. The labeled source data, $\boldsymbol{D}_S$, are therefore not representative of the photometric target data, $\boldsymbol{D}_T$, as the selection of spectroscopic source samples is biased towards brighter and bluer objects. The automatic classification of unlabelled objects, based on biased spectroscopically confirmed source data, is the subject of much research, including public classification challenges (Kessler et al., 2010b, 2019).

Leading SNIa classification approaches are based on data augmentation; they sample synthetic objects from Gaussian process (GP) fits of the LCs to overcome covariate shift (Revsbech et al., 2018, Boone, 2019). The method of Revsbech et al. (2018) can be viewed as a prototype of *StratLearn*, as it augments the source data separately in strata based on the estimated propensity scores. However, to optimize data augmentation within strata, Revsbech et al. (2018) requires a sub-sample of labeled target data that is unavailable in practice. While effective in this particular case, GP data augmentation is not an option in most covariate shift tasks. We show that *StratLearn* makes augmentation unnecessary. We use target prediction AUC to compare performance to published results.

Data and Preprocessing: We use data from the updated "Supernova photometric classification challenge" (SPCC) (Kessler et al., 2010a), contain-

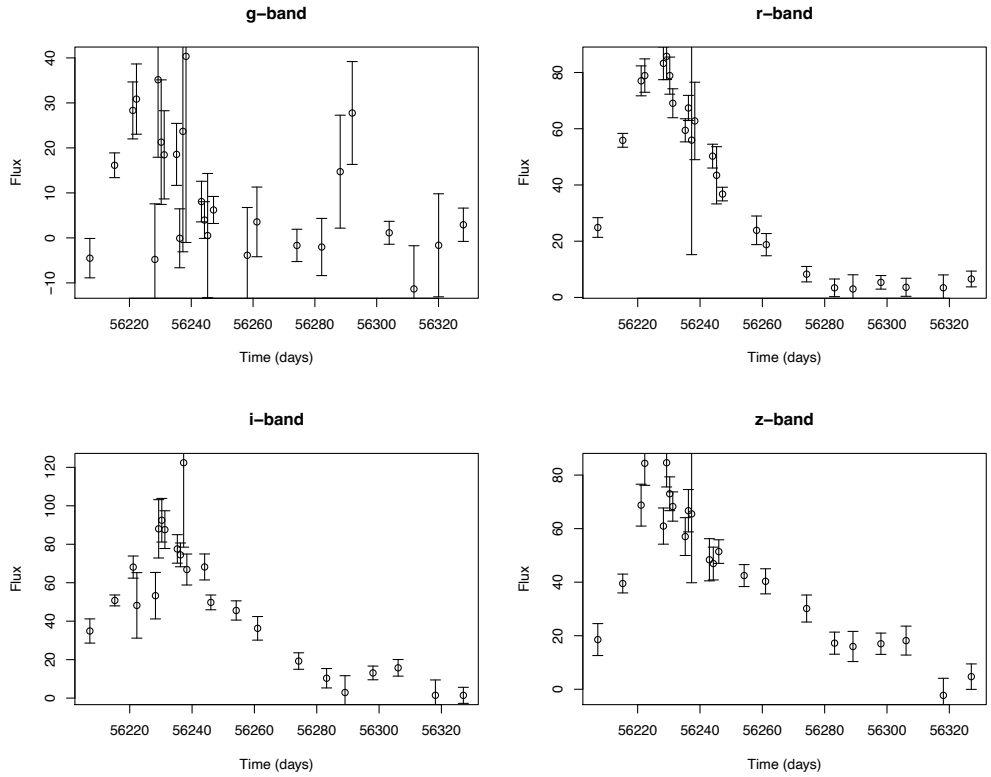


Figure 2.6: Example of photometric LC data, including 1σ error bars, for a typical SNIa (specifically, SN2475 from the updated Kessler et al. (2010a) simulated SPCC data).

ing a total of 21,318 simulated SNIa and of other types (Ib, Ic and II). For each supernova (SN), LCs are given in four color bands, $\{g, r, i, z\}$. The data include a source set $\boldsymbol{D}_S$ of 1102 spectroscopically confirmed SNe with known types and 20,216 SNe with unknown types (target set $\boldsymbol{D}_T$). 51% of the source objects are SNIa, while only 23% of the target date are SNIa, a consequence of the strong covariate shift in the data.

We follow the approach in Revsbech et al. (2018), which was applied to an earlier release of the SPCC data (Kessler et al., 2010b, discussed in Section A.2), to extract a set of covariates from the LC data that can be used for classification. First, a GP with a squared exponential kernel is used to model the LCs. Then, a diffusion map (Coifman and Lafon, 2006) (as used in Richards et al. (2012)) is applied, resulting in a vector of 100 similarity measures between the LCs that we use as predictor variables. Combining these

with redshift (a measure of cosmological distance, defined in Section 2.6) and a measure of overall brightness, we obtain 102 predictive covariates.[8]

RESULTS: To evaluate the impact of covariate shift on classification, we first consider a 'biased fit' by training a random forest classifier (as in Revsbech et al. (2018)) on the source covariates ignoring covariate shift, resulting in an AUC of 0.902 on the target data (black ROC curve in Figure 2.7). Next, we obtain a 'gold standard' benchmark by randomly selecting 1102 objects from target data as a representative source set. The same classification procedure with the unbiased 'gold standard' training data (unavailable in practice) yields an AUC of 0.972 on the remaining 19,114 target objects.

Given the biased source data, *StratLearn* is implemented as described in Section 2.3, including all 102 covariates in the logistic propensity score estimation model. After stratification, a random forest classifier is trained and optimized on source strata $\boldsymbol{D}_{S_1}$ and $\boldsymbol{D}_{S_2}$ separately to predict samples in target strata $\boldsymbol{D}_{T_1}$ and $\boldsymbol{D}_{T_2}$. We use repeated 10-fold cross validation with a large hyperparameter grid to minimize the empirical risk of (2.17) within each strata, employing log-loss[9] as our loss-function.[10] Source strata $\boldsymbol{D}_{S_j}$ for $j \in \{3,4,5\}$ have a small sample size, (13,7,4) respectively. Thus, source strata $\boldsymbol{D}_{S_j}$ for $j \in \{3,4,5\}$ are merged with $\boldsymbol{D}_{S_2}$ to train the random forest

[8]The original and pre-processed SPCC data can be found here: SPCC data link.

[9]The log-loss (also referred to as cross-entropy loss) compares the output of a classification $f(x) \in [0,1]$ with the true output y for an observation (x,y) via $\ell_{\text{logloss}}(f(x),y) := -(y\log(f(x)) + (1-y)\log(1-f(x)))$.

[10]We use the "ranger" random forest implementation in `caret` (Kuhn, 2008) for SNe classification in this section. Repeated 10-fold cross-validation (with five repetitions) is implemented with a large hyperparameter grid, containing four different numbers of covariates to possibly split at in each node (3,5,8,10); five values for minimum node size (3,5,7,10,20); and two different splitting rules ("gini" and "extratrees"). To optimize the selected hyperparameter setting, we compute the average log-loss via cross-validation on source data. Using *StratLearn*, cross-validation for hyperparameter selection (optimizing the average log-loss) is done on each source group separately. We employ the log-loss to compute the empirical risk by evaluating the loss function $l(f(x),y)$ on each sample separately as described in Section 2.2. The AUC (used to assess target predictive performance in this example) does not include the evaluation of a loss function that compares the true label y and prediction $f(x)$ for each sample x separately. However, to demonstrate robustness w.r.t. the choice of hyperparameter optimization, we also compute the AUC on source data (on each source group separately) via cross-validation, which leads to the same hyperparameter setting as using the log-loss.

to predict $\boldsymbol{D}_{T_j}$ for $j \in \{3, 4, 5\}$. With *StratLearn*, we obtain an AUC of 0.958 on the target data (blue ROC curve in Figure 2.7), very near the optimal 'gold standard' benchmark.

Figure 2.7 compares *StratLearn* to importance sampling methods designed to adjust for covariate shift. For importance sampling, the bootstrapped samples in the random forest fit were resampled with probabilities proportional to the estimated importance weights (see Section 2.2.3). NN and IPS led to the best importance weighted classifier (AUC = 0.923, 0.921) – an improvement over the biased fit, but substantially lower than *StratLearn*. AUC standard errors (see Figure 2.7) are small relative to the large performance improvement of *StratLearn*. KLIEP failed to fit importance weights and is thus not included in the results. We also implemented IWCV using the same hyperparameter grid as for *StratLearn*, and a combination of IWCV and importance sampling, which both led to lower AUC than the ones reported in Figure 2.7 (see Section A.1).

Previous state-of-the-art methods report an AUC of 0.855 (Lochner et al., 2016) using boosted decision trees, 0.939 (Pasquet et al., 2019) using a framework of an autoencoder and a convolutional neural network and 0.94 (Revsbech et al., 2018) using LC augmentation and target data leakage, all lower than *StratLearn*.

Balance Assessment on updated SPCC data: To illustrate the balancing property of propensity scores (see Section 2.3.3) and its effect on predictive target performance, we assess the covariate balance in the updated SPCC data within strata conditional on the estimated propensity scores, by means of two commonly used balance measures: absolute standardized mean differences (SMD) and the Kolmogorov-Smirnov test statistics (KS-stats) (Austin, 2011, Austin and Stuart, 2015).

Figure 2.8 provides a detailed covariate balance comparison, by plotting the "raw" SMD against the *StratLearn* SMD in stratum 1 (black) for each covariate. We remove two outliers (redshift and brightness) with very large "raw" SMD (1.1 and 1.7), because including them in the Figure makes it

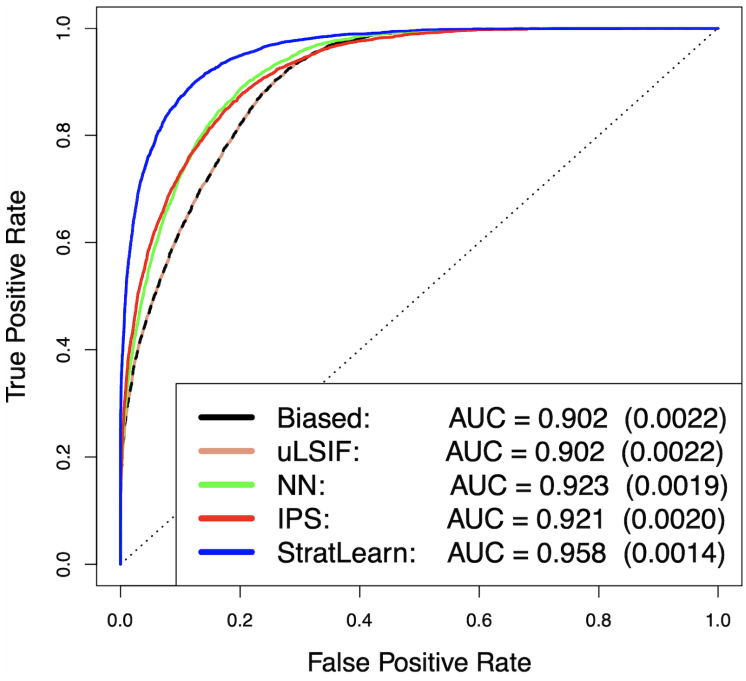


Figure 2.7: Comparison of ROC curves for SNIa classification using the updated SPCC data. Here, Biased and uLSIF are identical. Bootstrap AUC standard errors (from 400 bootstrap samples) are given in parentheses.

more difficult to illustrate the balance of the bulk of the covariates; both are well balanced in stratum 1 using *StratLearn* (SMD equals 0.12 and 0.17). Points below the diagonal line are better balanced in the stratum than in the "raw" non-stratified data. This is the case for the vast majority (71%) of black points in Figure 2.8, illustrating the balance improvement achieved with *StratLearn*.

Figure 2.8 also plots (red) the SMD achieved by STACCATO (Revsbech et al., 2018), which uses two covariates (redshift and brightness, as opposed to the 102 used by *StratLearn*) in the logistic regression to estimate propensity scores. While STACCATO improves the balance of the majority (69%) of the covariates, most (66%) black (*StratLearn*) SMD have smaller vertical values, indicating better balance than STACCATO (red).

On average across the 102 covariates, *StratLearn* improves covariate balance compared to the "raw" non-stratified data measured by SMD by ~70% in

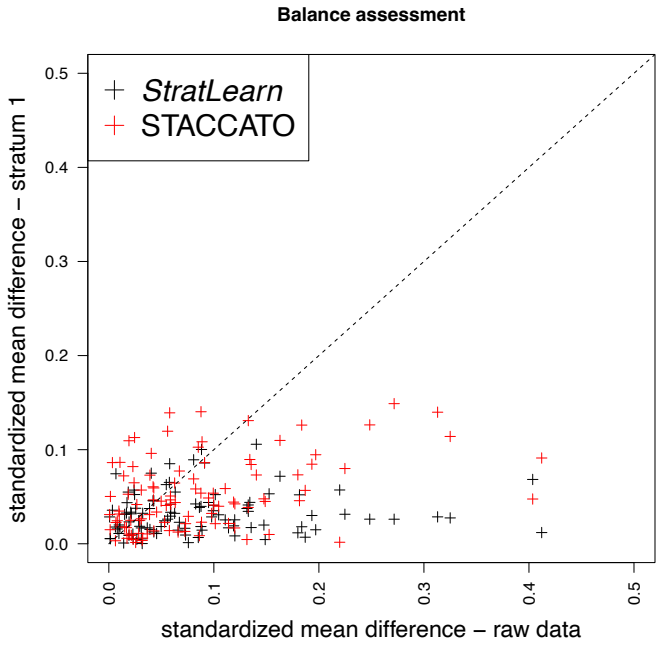


Figure 2.8: Absolute standardized mean differences between source and target data of stratum 1 plotted against "raw" data absolute standardized mean differences for *StratLearn* and STACCATO.

stratum 1 and ~10% in stratum 2 (KS-stats: ~70% in stratum 1 and ~30% in stratum 2)[11]. It further improves upon STACCATO by ~36% in stratum 1 and ~46% in stratum 2 using SMD (KS-stats: ~24% in stratum 1 and ~36% in stratum 2). The remaining strata contain too few source data to assess covariate balance. Details are provided in Table A.2.

The improved covariate balance (reduced covariate shift) directly translates into improved predictive performance. STACCATO (including data augmentation and target data leakage, mostly unavailable in practice) yields a target AUC of 0.94, whereas with *StratLearn* we obtain a target AUC of 0.958 (without data augmentation and no target data leakage) – a substantial improvement resulting from the improved covariate balance by accounting for potentially confounding covariates. In general, we note that balance is particularly important for covariates that are strongly predictive for the outcome.

[11]Percentages are calculated by taking the ratio of the average SMD (average KS-stats) of all 102 covariates.

Table 2.4: Strata composition on the updated SPCC data (Section 2.5), applying STACCATO (left) and *StratLearn* (right). The number of SNe, as well as the number and proportion of SNIa are presented in source and target stratum 1 and 2. For conciseness, we present the combined strata 3 to 5, containing too little source data for meaningful comparison of the SNIa proportions.

		STACCATO			*StratLearn*		
Stratum	Set	Number of SNe	Number of SNIa	Prop. of SNIa	Number of SNe	Number of SNIa	Prop. of SNIa
1	Source	924	414	**0.45**	958	518	**0.54**
	Target	3340	1125	**0.34**	3306	1790	**0.54**
2	Source	153	125	**0.82**	120	28	**0.23**
	Target	4111	973	**0.24**	4144	927	**0.22**
3 to 5	Source	25	19	0.76	24	12	0.5
	Target	12,765	2431	0.19	12,766	1812	0.14

Domain-specific expertise might be necessary to identify such covariates in the individual cases in practice. In Section 2.4.2, we demonstrate how covariate balance can be improved by adjusting the propensity score model.

Table 2.4 presents the composition of the five *StratLearn* strata. Recall that according to Remark 1, conditional on the propensity score the marginal distributions of source and target outcome are the same in expectation. Table 2.4 shows that the proportion of SNIa in the source and target data (which in this case can be computed from knowledge of the true target labels in the simulation) align well for *StratLearn* in the first two strata, indicating the expected reduction in covariate shift. The source sample sizes in strata 3-5 are quite small, rendering meaningful comparison of the SNIa proportions impossible. In Strata 1 and 2, however, *StratLearn* achieves much better balance than either STACCATO or the raw (un-stratified) data (51% SNIa in source, 23% SNIa in target).

In Table 2.5, we demonstrate how predicted outcomes can be employed for balance diagnostics by assessing the predicted proportions of SNIa within strata obtained by STACCATO and by *StratLearn*. For both methods, we compute the predicted outcomes by classifying objects to be SNIa if the (random forest) predicted SNIa probabilities are above 0.5. While STACCATO leads to a strong discrepancy between predicted SNIa proportions in the

Table 2.5: Outcome balance diagnostics via predicted labels on the updated SPCC data (Section 2.5), applying STACCATO (left) and *StratLearn* (right). The number and proportion of predicted SNIa are presented in source and target stratum 1 and 2. P-values are computed via Fisher's exact test of independence between predicted SNIa target and source proportions within strata.

		STACCATO (predicted)			*StratLearn* (predicted)		
Stratum	Set	Number of SNIa	Prop. of SNIa	p-value	Number of SNIa	Prop. of SNIa	p-value
1	Source	414	**0.45**	8.4e-11	518	**0.54**	0.284
	Target	1106	**0.33**		1853	**0.56**	
2	Source	125	**0.82**	2.8e-13	28	**0.23**	0.749
	Target	2166	**0.53**		1040	**0.25**	

first two strata (indicating remaining confounding), *StratLearn* leads to well-matched predicted SNIa proportions. We further quantify the discrepancy by performing a two-sided Fisher's exact test of independence, with the null hypothesis that there is no association of source/target set assignment and predicted SNIa proportion. Comparing different propensity score models, a higher p-value is an indicator for better balance in the predicted outcomes, and should thus be desirable. *StratLearn* leads to much higher p-values than STACCATO (failing to reject the null hypothesis for reasonable significance levels), which implies much weaker relation between source/target assignment and predicted outcomes.

In this particular example, with *StratLearn*, we fail to reject the null hypothesis for most significance levels. This may not always be the case (e.g., Section A.2). We recall that the strategy of conditioning on propensity scores via stratification leads to subgroups with similar (not identical) propensity scores, and thus to similar (not identical) joint distributions within strata (this is the approximation in (2.16)). This in turn might lead to differences in the distributions of the covariates and the (predicted) outcomes, even if we could condition on the true propensity scores. We thus employ the p-values of (predicted) outcomes as an additional tool to asses, and primarily to compare, propensity score models to detect and reduce confounding of highly predictive and thus most relevant covariates.

2.6 Conditional Density Estimation – Photo-z Regression

Objective: The wavelength of light from extragalactic objects is stretched because of the expansion of the universe – a phenomenon called 'redshift'. This fractional shift towards the red end of the spectrum is denoted by z. A precise measurement of redshift allows cosmologists to estimate distances to astronomical sources, and its accurate quantification is essential for cosmological inference (e.g., redshift is a key component of the Big Bang theory). Because of instrumental limitations, redshift can be precisely measured only for a small fraction of the $\sim 10^7$ galaxies observed to date (set to grow to $\sim 10^9$ within a decade). These source data are subject to covariate shift relative to the set of galaxies with unknown redshift (target). Izbicki et al. (2017) employed importance weighting to adjust for covariate shift in x, a set of observed photometric magnitudes (a logarithmic measure of passband-filtered brightness), when estimating z. They obtain a non-parametric estimate of the full conditional density, $f(z|x)$, to quantify predictive uncertainty of redshift estimates. Proper quantification of predictive uncertainties is crucial to avoid systematic errors in the scientific downstream analysis (Izbicki et al., 2017, Sheldon et al., 2012). Using the same setup and conditional density estimation models (hist-NN, ker-NN, Series and Comb, detailed in Izbicki et al. (2017)),[12] we show that *StratLearn* leads to better overall predictive performance than importance weighting.

Assuming that source and target data follow the same distribution, under the L^2–loss, conditional density estimators typically aim to minimize the *generalized* risk (generalized in that the underlying loss can be negative):

$$\hat{R}_S(\hat{f}) = \frac{1}{n_S}\sum_{k=1}^{n_S}\int \hat{f}^2(z|x_S^{(k)})dz - 2\frac{1}{n_S}\sum_{k=1}^{n_S}\hat{f}(z_S^{(k)}|x_S^{(k)}). \tag{2.18}$$

Following the arguments in Izbicki et al. (2017), we provide derivations of the generalized risk presented in (2.18) in Section A.3.1. To provide an intuition for the *generalized* risk in (2.18), the second part of (2.18) computes

[12] For the computation of the conditional density estimators we used code by Izbicki et al. (2017) (link).

the values of the conditional density estimates at the true spectroscopic redshift (known for the source data), which is optimized if the true redshift is at (or close to) the mode of the conditional density estimate $\hat{f}(z_i|x_i)$ with low uncertainty otherwise (the Dirac delta distribution at the true redshift value is the optimal limiting case). On the contrary, the first part of (2.18), which integrates the squared conditional density estimates over the redshift range (without information of the true redshift), is minimized for flat conditional density estimates, penalizing very certain predictions. Thus, overall, estimates that are very certain, but fail to cover the truth, lead to a high risk.

Izbicki et al. (2017) propose to adjust for covariate shift by adapting (2.18), via optimizing weighted versions of the conditional density estimators (Izbicki et al., 2017, Sections 5.1-5.3) with respect to an importance weighted *generalized* risk:

$$\hat{R}_{S1}(\hat{f}) = \frac{1}{n_T}\sum_{k=1}^{n_T}\int \hat{f}^2(z|x_T^{(k)})dz - 2\frac{1}{n_S}\sum_{k=1}^{n_S}\hat{f}(z_S^{(k)}|x_S^{(k)})\hat{w}(x_S^{(k)}), \tag{2.19}$$

where the weights, $\hat{w}(x_S) = p_T(x)/p_S(x)$, are estimated using the methods described in Section 2.2.3. As their best performing model for $f(z|x)$, Izbicki et al. (2017) propose an average of importance weighted ker-NN and Series,

$$\hat{f}^{\alpha}(z|x) = \sum_{k=1}^{p}\alpha_k\hat{f}_k(z|x), \text{ with constraints (i): } \alpha_i \geq 0, \text{ and (ii): } \sum_{k=1}^{p}\alpha_k = 1, \tag{2.20}$$

referred to as 'Comb' (i.e., combination), where $p = 2$ and α_i is optimized to minimize (2.19).

With *StratLearn*, we optimize the unweighted conditional density estimators (hist-NN, ker-NN, Series) by minimizing (2.18) in each source stratum separately (accounting for covariate shift following Proposition 1). We also propose a *StratLearn* version of Comb by optimizing (2.20) on each source stratum separately (via the generalized risk in (2.19) with $w(x) \equiv 1$), including ker-NN and Series (each optimized via *StratLearn* beforehand). *StratLearn*

and the other methods are compared with a 'Biased' (unweighted) method that simply optimizes (2.18). We abbreviate the combination of each method (*StratLearn*, Biased, and each of the weighting methods in Section 2.2.3) with the models for $f(z|x)$ (hist-NN, ker-NN, Series and Comb) as $\text{Method}_{\text{Model}}$.

DATA: We use the same data as Izbicki et al. (2017), consisting of 467,710 galaxies from Sheldon et al. (2012)[13], each with spectroscopic redshift z (measured with negligible error), and five photometric covariates x. As in Izbicki et al. (2017), we use the r-band magnitude and the four colors (differences of magnitude values in adjacent photometric bands) as our covariates. We denote this spectroscopic source sample by $\boldsymbol{D}_S$. To simulate realistic covariate shift, we follow Izbicki et al. (2017): starting from $\boldsymbol{D}_S$, we use rejection sampling to simulate a photometric, unrepresentative target sample $\boldsymbol{D}_T$, with the prescription $p(s = 0|x) = f_{B(13,4)}(x_{(r)})/\max_{x_{(r)}} f_{B(13,4)}(x_{(r)})$, where $x_{(r)}$ is the r-band magnitude and $f_{B(13,4)}$ is a beta density with parameters (13,4). Additionally, we investigated adding $k \in \{10, 50\}$ i.i.d. standard normal covariates as potential predictors to the 5 photometric covariates. This simulates the realistic case where additional potentially confounding covariates are present. For comparability, we follow Izbicki et al. (2017) and use $|\boldsymbol{D}_S^{\text{train}}| = 2800$ galaxies randomly sampled from $\boldsymbol{D}_S$ as training data, plus a validation set of $|\boldsymbol{D}_S^{\text{val}}| = 1200$ galaxies. We assess the performance of each $\text{Method}_{\text{Model}}$ pair using a random subset of $\boldsymbol{D}_T$, i.e., $|\boldsymbol{D}_T^{\text{test}}| = 6000$.

RESULTS: For evaluation of $\hat{f}$ under each $\text{Method}_{\text{Model}}$ pair, we use the (in our simulation) known target redshifts, z_T, to compute the target risk, $\hat{R}_T(\hat{f})$, via a non-weighted version of (2.19) with $x_S^{(k)}$ and $y_S^{(k)}$ replaced by $x_T^{(k)}$ and $y_T^{(k)}$, given by:

$$\hat{R}_T(\hat{f}) = \frac{1}{n_T}\sum_{k=1}^{n_T}\int \hat{f}^2(z|x_T^{(k)})dz - 2\frac{1}{n_T}\sum_{k=1}^{n_T}\hat{f}(z_T^{(k)}|x_T^{(k)}). \tag{2.21}$$

[13]The data in Section 2.6 was used by permission of the original authors of Sheldon et al. (2012).

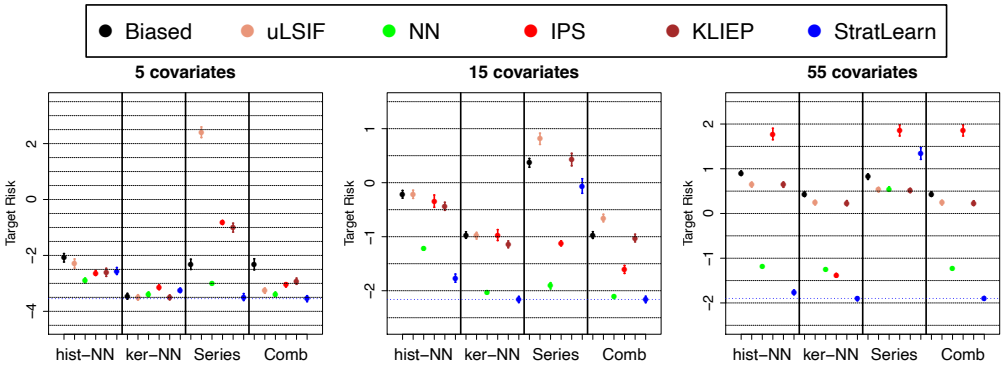


Figure 2.9: Target risk ($\hat{R}_T$) of the four photo-z estimation models under each method (different colors), using different sets of predictors. Bars give the mean ± 2 bootstrap standard errors (from 400 bootstrap samples).

Figure 2.9 compares the resulting target risk $\hat{R}_T$ across models and covariate sets, showing that $StratLearn_{Comb}$ gives the best performance in all three covariate setups.

For small covariate space dimension (Figure 2.9, left panel), $StratLearn_{Comb}$ improves upon $StratLearn_{ker\text{-}NN}$ and $StratLearn_{Series}$, optimizing the source risk in each stratum separately and combining their predictions. In the presence of potential additional confounding covariates (Figure 2.9, middle and right panels), the performance of the Series estimator degrades strongly under most methods. In these cases, $StratLearn_{Comb}$ exploits the higher performance of $StratLearn_{ker\text{-}NN}$. In contrast, for the non-adjusted (Biased) and importance weighted methods (e.g. IPS), the combination of approaches (Comb) does not necessarily lead to improved performance (e.g., IPS_{Comb} exhibits a higher target risk than $\text{IPS}_{\text{ker-NN}}$ on its own (Figure 2.9, right panel)), indicating that the optimization in (2.20) fails due to remaining covariate shift in the data. More precisely, the weighted empirical *source* risk minimization ((2.19), as a form of (2.1)) does not lead to *target* risk minimization in these situations. In general, the improvement of *StratLearn* relative to weighting methods increases with the dimensionality of the covariate space, leading to a more robust regime.

2.7 Discussion

In this chapter, we provide a novel framework for learning under covariate shift, theoretically justified and based on principles derived in the propensity score causal inference framework (Rosenbaum and Rubin, 1983). Our approach is computationally efficient, easy to implement and readily adaptable to various data applications. We show that *StratLearn* outperforms a range of state-of-the-art importance weighting methods both on toy covariate shift examples and on contemporary research questions in cosmology, especially in the presence of a high dimensional covariate space. Our framework is entirely general and versatile, as demonstrated with examples of regression, conditional density estimation and classification. We believe it may become a powerful alternative to importance weighting, with a myriad of possible extensions and applications. We provide an extensive discussion of *StratLearn*, with limitations and prospective future work in Section 5.1. In the next section, we demonstrate the broad utility of *StratLearn* by employing it for photo-z calibration for weak lensing.

3

Improved Weak Lensing Photometric Redshift Calibration via StratLearn and Hierarchical Modeling

Overview: Discrepancies between cosmological parameter estimates from cosmic shear surveys and from recent Planck cosmic microwave background measurements challenge the ability of the highly successful ΛCDM model to describe the nature of the universe. To rule out systematic biases in cosmic shear survey analysis, accurate redshift calibration within tomographic bins is key. In this paper, we improve photo-z calibration via Bayesian hierarchical modeling of full galaxy photo-z conditional densities, by employing *StratLearn*, a recently developed statistical methodology, which accounts for systematic differences in the distribution of the spectroscopic training/source set and the photometric target set. Using realistic simulations that were designed to resemble the KiDS+VIKING-450 dataset, we show that *StratLearn*-estimated conditional densities allow improvement of

the galaxy tomographic bin assignment. Given the improved bin assignment, our *StratLearn*-Bayesian framework leads to nearly unbiased estimates of the target population means, improving upon previously best photo-z calibration methods.

3.1 Introduction

Cosmological parameter estimation from the cosmic microwave background (CMB; Miville-Deschênes et al. 2020) and from tomographic cosmic shear measurements (Hildebrandt et al., 2020, Troxel et al., 2018, Hikage et al., 2019) lead to discrepancies in the estimated clustering strength of dark matter. Such systematic discrepancies challenge the highly successful dark energy and cold dark matter paradigm (ΛCDM) in describing the true nature of the Universe. Of course, such a claim needs critical and detailed consideration of the surveys and analysis steps performed by the various collaborations to rule out systematic biases in the different procedures, which might explain the discrepancies. Extensive explorations of various survey, model, and analysis modifications have been performed (Hildebrandt et al., 2020), demonstrating that a modification of the redshift calibration approach might explain the observed discrepancy with the *Planck* CMB results. Comparing the methodologies applied by the Dark Energy Survey (DES; Troxel et al. 2018), Hyper-Suprime Camera (HSC; Hikage et al. 2019), and the Kilo-Degree Survey (KiDS; Hildebrandt et al. 2020) collaborations, the redshift calibration approach constitutes the most fundamental difference (Hildebrandt et al., 2020, Wright et al., 2020). To investigate discrepancies to the *Planck* CMB estimates, careful consideration of the redshift calibration method used in cosmic shear tomography is thus fundamental.

In cosmic shear tomography, galaxies are assigned to (pre-defined) tomographic redshift bins (Wright et al., 2020, Porqueres et al., 2021) based on an estimate of their photometric redshift (photo-z). If the estimated population redshift distribution (within a tomographic bin) differs systematically from the (non-observable) true redshift distribution, the parameter estimates from cosmic shear tomography might be systematically biased. For instance,

if the estimated redshift distribution is systematically lower than it is in reality, then observed gravitational distortions are attributed to an overly dense and too highly clustered matter distribution. It is thus essential to obtain accurate redshift distribution estimates. In particular, it is crucial to obtain an unbiased estimate of the first moment (mean) of the true underlying redshift population distribution (within a tomographic bin), in order to avoid such systematic biases in the analysis (Amara and Refregier, 2008).

Several redshift calibration methods have been investigated to improve cosmic shear tomography. Wright et al. (2019) group these approaches into three categories: (i) cross-correlation with reference galaxy samples that have precise and accurate redshifts (Schneider et al., 2006, Newman, 2008, McQuinn and White, 2013, Morrison et al., 2017); (ii) stacking of individual galaxy redshift distributions (Hildebrandt et al., 2012, Hoyle et al., 2018, Tanaka et al., 2018); and (iii) direct redshift calibration (Lima et al., 2008, Hildebrandt et al., 2016, 2020, Buchs et al., 2019, Wright et al., 2020). The idea of direct redshift calibration is to reweight the distribution of spectroscopic redshift, obtained only for a small and non-representative subsample of the data (source/training data), to match the distribution of the photometric target data. In recent work, Wright et al. (2020) develop a direct redshift calibration method based on self organising maps (SOM; Kohonen 1982), outperforming previously proposed methods on comprehensive simulations designed to realistically resemble the KIDS+Viking-450 dataset (Wright et al., 2019, 2020, Hildebrandt et al., 2020). They obtain the tomographic bin assignment via a Bayesian-Photometric-Redshift estimate (Benitez, 2000), further denoted as z_B, calculated for each galaxy. While improving on other direct redshift calibration methods, the SOM method proposed in Wright et al. (2020) still leads to potentially concerning bias in some tomographic bins, and only alleviates this bias by introducing systematic quality cuts to the data, which increases error bars on the fitted parameters in the scientific analysis and might lead to systematic errors.

In this paper, we propose a different strategy to improve redshift calibration, based on galaxy (object level) conditional photo-z density estimates. More precisely, we employ *StratLearn*, that allows principled photo-z con-

ditional density estimation under non-representative source/training data. *StratLearn* alleviates (or bypasses) the problem of non-representative source/-training data (identified as covariate shift), by subgrouping the data into strata based on estimated propensity scores. In the context of this chapter, the propensity score is the probability of a galaxy being assigned to the spectroscopic training/source set given the observed covariates (photometric magnitudes/colors). In Chapter 2 of this thesis, we demonstrate that fitting conditional density estimators within strata, constructed by partitioning the data based on the estimated propensity scores, strongly improves full conditional photo-z density estimates under-non representative source/training data. Here, we show that the *StratLearn* conditional densities[1] can be used directly to improve the tomographic bin assignment, by assigning each galaxy to the tomographic bin with its highest conditional probability. In a second step, we construct a Bayesian hierarchical framework to model summaries of the object level conditional densities (within tomographic bins), leading to nearly unbiased estimates of the target population means. We evaluate our novel *StratLearn*-Bayes approach on comprehensive simulations constructed by Wright et al. (2020), demonstrating a substantial reduction of bias compared to the previously proposed SOM calibration method (Wright et al., 2020).

The remainder of the paper is structured as follows. In Section 3.2.1, we specify notation and we formally introduce the underlying covariate shift scenario, arising through the non-representativeness of the spectroscopic training/source data. In Section 3.2.2, we summarize the direct redshift calibration method, as introduced in Lima et al. (2008) and recently improved in Wright et al. (2020). We then briefly introduce the supervised learning task with a focus on conditional density estimation under the covariate shift scenario. In Section 3.3, we formally introduce our approach. In Section 3.3.1, we provide detailed instructions on estimating conditional densities under covariate shift within *StratLearn*. We then specify how these conditional

[1]We note that the conditional density estimators themselves are not part of the *StratLearn* methodology, and have been proposed elsewhere (Izbicki et al., 2017, Izbicki and Lee, 2016). For conciseness, we loosely refer to the conditional density estimates obtained within the *StratLearn* framework as '*StratLearn* conditional densities'.

densities can be used to improve galaxy tomographic bin assignment (Section 3.3.2). In Section 3.3.3, we demonstrate how summaries of the estimated conditional densities can be employed in a Bayesian hierarchical framework to accurately estimate the redshift population means (per tomographic bin). In Section 3.4, we provide proof of concept how inverse propensity score weighting (inverse-PS) can be employed to estimate the redshift population shapes (per tomographic bin). In Section 3.5, we present numerical evaluation of our newly proposed method. We first introduce the simulation study used to evaluate our proposed *StratLearn*-Bayes method in Section 3.5.1. We then evaluate our new bin assignment with a comparison to the previously used z_B bin assignment (Section 3.5.2). We present our redshift calibration results in Section 3.5.3, and illustrate preliminary inverse-PS results in Section 3.5.5. Finally, in Section 3.6, we conclude with a short discussion of our findings.

3.2 Addressing the Non-Representative Training Data

3.2.1 Non-Representative Spectroscopic Data and Covariate Shift

Let z_i be the true spectroscopic redshift of galaxy i, and x_i be the vector of its observed photometric magnitudes/colors (the exact choice of covariates is described in Section 3.5.1). In a cosmic shear analysis, we have access to a relatively small set of galaxies with measured spectroscopic redshift, which for our purposes can be considered to be the true redshift. We denote this spectroscopic set as source (or training) data $\boldsymbol{D}_S = \{(x_S^{(i)}, z_S^{(i)})\}_{i=1}^{n_s}$, with n_s galaxies drawn from the joint distribution $p_S(x, z)$. The so-called "photo-z estimation problem" (Hildebrandt et al., 2010, Izbicki et al., 2017, Freeman et al., 2017, Jones and Heavens, 2019, Dey et al., 2023) is to find an estimate for the true redshift of a much larger set of galaxies, for which only photometric data x_T are available, but not spectroscopically measured redshifts, z_T. We denote this photometric set as our target data $\boldsymbol{D}_T = \{x_T^{(i)}\}_{i=1}^{n_t}$, with n_t unlabelled samples from the joint distribution $p_T(x, z)$ (with $n_T \gg n_S$). The problem is compounded by the fact that $p_S(x, z) \neq p_T(x, z)$, i.e., the spectro-

scopic source and photometric target distributions differ systematically due to selection effects in the acquisition of spectroscopy based on characteristics of the photometric magnitudes, leading to $p_S(x) \neq p_T(x)$. However, the conditional distribution of redshift z given the magnitudes x are the same in spectroscopic source and photometric target data, i.e. $p_S(z|x) = p_T(z|x)$. This situation, in which $p_S(z|x) = p_T(z|x)$ but $p_S(x) \neq p_T(x)$, is identified as *covariate shift*, as defined in the previous chapter. As illustrated in Chapter 2, if covariate shift is ignored, this leads to poor performance of machine learning or other statistical methods that aim at learning the relationship between covariates and true redshift, since the training set is not representative of the test set, implying lack of generalization.

3.2.2 Related Literature – Direct Redshift Calibration

Since in the covariate shift scenario $p_S(z|x) = p_T(z|x)$ but $p_S(x) \neq p_T(x)$, it generally follows that the redshift distribution of the spectroscopic set differs from that of the target, $p_S(z) \neq p_T(z)$. Direct redshift calibration methods reweight the spectroscopic redshift sample to match the photometric redshift distribution (Lima et al., 2008).

More precisely, with x denoting set of covariates, and z being the spectroscopic redshift, under the covariate shift scenario, it holds that

$$p_T(z, x) = p_T(z|x)p_T(x) \tag{3.1}$$
$$= p_S(z|x)p_T(x) \tag{3.2}$$
$$= p_S(z, x)\frac{p_T(x)}{p_S(x)} \tag{3.3}$$

That means one can express the joint target distribution by reweighting the joint source distribution. Precisely, $p_T(z, x) = \omega(x)p_S(z, x)$, with weights $\omega(x) = p_T(x)/p_S(x)$. In practice, one can reweight galaxies in the spectroscopic source set (with weights $\omega(x)$) to obtain a representative sample of the joint target distribution. In principle, looking at the marginal sample of z in the weighted joint distribution thus provides us a consistent estimate of the target redshift distribution $p_T(z)$.

Accurate estimation of the weights $\omega(x)$ is key for direct redshift calibration methods. Lima et al. (2008) and Hildebrandt et al. (2020) implement a k-nearest-neighbor (kNN) method for weight estimation. Wright et al. (2020) demonstrate improvement over the kNN method by computing the weights via an SOM method, a form of unsupervised neural network which maps the high-dimensional covariate space to a lower-dimensional grid.

Unfortunately, weighting methods typically entail high variance, particularly if there is a small number of objects with very large weights. In addition, finding a suitable set of weights ω is not trivial, but key for direct calibration methods. Noisy, inaccurate weights might lead to potentially severe bias and strongly increased variance.

3.2.3 Photometric Redshift Regression

Instead of reweighting, our approach uses the labelled spectroscopic source data as a training set to fit supervised full conditional density models. Our trained models then deliver a non-parametric estimate of the full conditional redshift (photo-z) distribution for each galaxy in the photometric target data (conditional on its observed covariates), $\hat{f}(z|x)$. If source and target data follow the same distribution, under the L^2-loss, conditional density estimators typically aim to minimize the generalized risk $\hat{R}_S(\hat{f})$ provided in (2.18).

In the presence of covariate shift, however, to obtain accurate target estimates it is necessary to minimize the target risk $\hat{R}_T(\hat{f})$ as given via (2.21), which is obtained by replacing all source samples in (2.18) with target samples, where typically $\hat{R}_S(\hat{f}) \neq \hat{R}_T(\hat{f})$. The challenge is to minimize $\hat{R}_T(\hat{f})$ without access to the target true redshift z_T. In the next section, we briefly recapitulate how we employ *StratLearn*, as introduced in Chapter 2 in this thesis, to minimize $\hat{R}_T(\hat{f})$ via source data.

3.3 StratLearn-Bayes Photometric Redshift Calibration

3.3.1 Photo-z Conditional Densities within *StratLearn*

In short, *StratLearn* allows target risk minimization by subgrouping the source and target data into strata based on estimated propensity scores. Within strata, the joint distribution of target data and source data are approximately the same, and target risk can thus be minimized via source risk minimization. In the following, we provide a more detailed description of our procedure.

In the context of this chapter, the propensity score $e(x_i)$, as defined in (2.8), is the probability of a galaxy i being in the spectroscopic source data, given its observed photometric covariates (colors/magnitudes) x_i. In practice, we obtain an estimate $\hat{e}(x_i)$ of the propensity score via binary, probabilistic classification of source and target data using a logistic regression model with all the photometric magnitudes/colors as independent variables (main effects) and source/target set assignment as the binary dependent variable.[2] As suggested in Chapter 2, we then subgroup (stratify) the source and target data into $k = 5$ groups, based on the quintiles of the estimated propensity score distribution $\hat{e}(x)$.

Then, within strata, by Proposition 1 (in Chapter 2), it holds that

$$p_{T_j}(z, x) \approx p_{S_j}(z, x), \text{ for } j \in 1, \ldots, k, \tag{3.4}$$

where S_j indicates conditioning on assignment to the j'th source stratum (analogously for target T_j). It directly follows that $\hat{R}_{Tj}(\hat{f}) \approx \hat{R}_{Sj}(\hat{f})$ within strata $j \in 1, \ldots, k$. Thus, we can minimize the target risk $\hat{R}_{Tj}(\hat{f})$ within strata by minimizing the source risk $\hat{R}_{Sj}(\hat{f})$ within strata.

[2] To provide an intuition for propensity score estimation, for example, if we had only one photometric band, and e.g., 10 galaxies with the same magnitude with 2 galaxies being in the spectroscopic source sample and the remaining 8 galaxies being in the photometric target sample, then the estimated propensity score for said galaxy would be 1/5. Of course, our photometric set has more than one dimension and continuous covariate values, which motivates the use of binary probabilistic classification models (e.g. logistic regression) to discriminate source and target data assignment.

Given the strata conditional on the estimated propensity score, we can now fit any supervised model on the spectroscopic source data within each stratum and predict on its respective photometric target stratum. As suggested in Chapter 2, within each strata, we employ a weighted average (convex combination) of two conditional density estimators: *ker-NN* (Izbicki et al., 2017) and *Series* (Izbicki and Lee, 2016). The kernel nearest neighbor estimator (ker-NN) computes the conditional density of an object via a kernel smoothed histogram of the redshift of its nearest neighbors. The spectral series estimator (Series) estimates the conditional densities by adapting to an intrinsic dimension of the data, as a lower-dimensional subspace of the covariate space, based on data-dependent eigenfunctions of a kernel-based operator (Izbicki and Lee, 2016). We refer to Izbicki et al. (2017) and Izbicki and Lee (2016) for additional details on these estimators. We combine both estimators with the rationale that, due to their differing working mechanisms, each of them performs better in a different regime, leading to a more robust combined estimator.[3]

Firstly, we individually optimize the conditional density estimators (ker-NN, Series), minimizing (2.18) in each source stratum separately. Our final *StratLearn* conditional density estimate is then obtained by combining the ker-NN and Series conditional density estimates $\hat{f}_{\text{ker-NN}}(z|x)$ and $\hat{f}_{\text{Series}}(z|x)$ via optimizing

$$\hat{f}(z|x) = (1-\alpha_1)\hat{f}_{\text{Series}}(z|x) + \alpha_1 \hat{f}_{\text{ker-NN}}(z|x), \tag{3.5}$$

constraining, $0 \leq \alpha_1 \leq 1$. The parameter α_1 is optimized to minimize the generalized risk

$$\hat{R}_{S2}(\hat{f}) = \frac{1}{n_T}\sum_{k=1}^{n_T}\int \hat{f}^2(z|x_T^{(k)})dz - 2\frac{1}{n_S}\sum_{k=1}^{n_S}\hat{f}(z_S^{(k)}|x_S^{(k)}). \tag{3.6}$$

We note that (3.6) displays the risk presented in (2.19), setting $w(x) \equiv 1$. The risk in (3.6) only differs from (2.18) by integrating over the photometric

[3]We refer the interested reader to the ensemble learning literature for further background (Wolpert, 1992, Van der Laan et al., 2007, Naimi and Balzer, 2018).

sample x_T instead of the spectroscopic x_S, which does not require any target redshift z_T. Finally, (3.5) provides us with a galaxy-per-galaxy full conditional density redshift estimate $\hat{f}(z|x)$. Some illustrative examples of the resulting galaxy conditional density estimates are shown in Figure 3.1. In Section 3.5.1, we provide additional details on computation and parameter optimization of the conditional density estimators.

3.3.2 Tomographic Bin Assignment

For cosmic shear analysis, the photometric galaxies are assigned to tomographic bins, preferably based on an accurate estimate of their redshift. Wright et al. (2020) (and others) use the posterior mode of the z_B estimate (the Bayesian-Photometric-Redshift estimate; Benitez 2000) to assign the photometric galaxies to five non-overlapping tomographic redshift bins: $(0.1, 0.3], (0.3, 0.5], (0.5, 0.7], (0.7, 0.9], (0.9, 1.2]$; we denote these ranges as bins 1 through 5, respectively. Galaxies with z_B estimates outside of the five bin ranges ($z_B \leq 0.1$ and $z_B > 1.2$) are discarded.

Our approach provides galaxy-per-galaxy estimates of photo-z. Instead of relying on z_B, we can therefore evaluate our conditional density estimates $\hat{f}(z_i|x_i)$ to provide an alternative tomographic bin assignment for each galaxy i. A natural choice is to assign each galaxy to the bin which contains the highest conditional probability: let $b(i)$ be the bin assignment of galaxy i, then

$$b(i) = \text{argmax}_m \int_{B(m)} \hat{f}(z_i|x_i)dz, \quad m = 1, \ldots, 5, l, r, \tag{3.7}$$

with $B(i)$, $i = 1, \ldots, 5$ specifying the five tomographic bin redshift ranges. $B(l) := \{z|z \leq 0.1\}$ and $B(r) := \{z|z > 1.2\}$ specify two end bins for galaxies outside of the bin ranges, which are not used in the analysis. Figure 3.1 shows examples of conditional density estimates, plotted on the tomographic redshift bin grid. The assigned bins $b(i)$ with highest conditional probabilities are shaded.

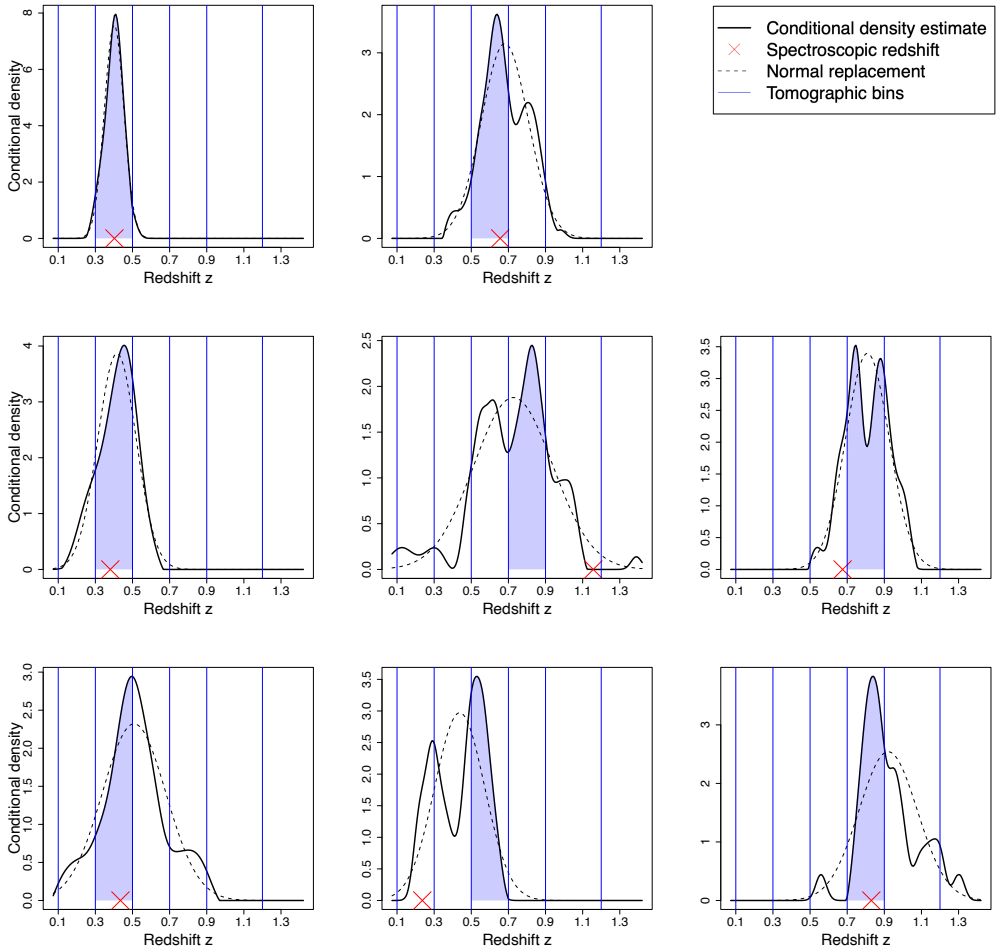


Figure 3.1: Examples of conditional density estimates $\hat{f}(z_i|x_i)$, for galaxies in the photometric target samples, illustrated on the tomographic bin grid. The *StratLearn* assigned bin (the one containing the highest conditional probability) is shaded in blue. The true spectroscopic redshift is shown by the red cross. A fraction of 30% appear to have a normal shape like in the top left example, but many conditional densities can be skewed and multimodal. Normal distributions with the same means and variances as the conditional density examples $\hat{f}(z_i|x_i)$ are added as dashed lines (as discussed in Section 3.3.3.)

3.3.3 Bayesian Hierarchical Modelling of Conditional Densities

In this section, we detail our Bayesian hierarchical framework for accurate estimation of the redshift population mean within each tomographic bin,

given the object-level conditional density estimates. Employing a hierarchical Bayesian framework allows us to model the conditional density estimates in a statistically principled framework, with optimal shrinkage on the object-level photo-z estimates, presumably leading to more precise population estimates.

Our target is to accurately estimate the population mean μ_b of z_i ($i = 1, \ldots, n_{Tb}$, with n_{Tb} being the number of galaxies within tomographic bin b.), which has been shown to be most crucial in the calibration of photo-z populations (Amara and Refregier, 2008). To this aim, we model the redshift population within each bin as a normal distribution – a convenient choice that facilitates the introduction of a hierarchical Bayesian framework. Specifically, on the redshift population level, we model

$$\text{Population Level:} \quad z_i|\mu_b, \sigma_b \overset{\text{indep.}}{\sim} N(\mu_b, \sigma_b^2), \tag{3.8}$$

with σ_b^2 being the redshift population variance. On the object (galaxy) level, $\hat{f}(z_i|x_i)$ is an estimate of the conditional density $p(z_i|x_i)$. Via Bayes theorem, the conditional density $p(z_i|x_i)$ can be expressed via

$$p(z_i|x_i) \propto p(x_i|z_i)p(z_i). \tag{3.9}$$

By assuming a flat prior on z_i (e.g., a uniform prior within the spectroscopic redshift range), we have $p(z_i) \propto 1$. Then, (3.9) simplifies to

$$p(z_i|x_i) \propto p(x_i|z_i). \tag{3.10}$$

We can now formulate the joint posterior distribution $p(z_1, \ldots, z_{n_{Tb}}, \mu_b, \sigma_b|X)$, with $X := \{x_i\}_{i=1}^{n_{Tb}}$, via

$$\begin{aligned}
p(z_1, \ldots, z_{n_{Tb}}, \mu_b, \sigma_b|X) \propto & p(X|z_1, \ldots, z_{n_{Tb}}, \mu_b, \sigma_b) \\
& p(z_1, \ldots, z_{n_{Tb}}|\mu_b, \sigma_b)p(\mu_b, \sigma_b) && (3.11)\\
= & p(\mu_b, \sigma_b) \prod_i p(x_i|z_i)p(z_i|\mu_b, \sigma_b) && (3.12)\\
\propto & p(\mu_b, \sigma_b) \prod_i p(z_i|x_i)p(z_i|\mu_b, \sigma_b), && (3.13)
\end{aligned}$$

where (3.11) to (3.12) holds due to the independence in (3.8) and the conditional independence $x_i \perp\!\!\!\perp (\{z_j\}_{j\neq i}, \mu_b, \sigma_b)|z_i$. That is, given the redshift z_i for an object i, its distribution of the photometry x_i does not depend on other observed redshifts, nor on the parameters describing the population of redshift. (3.12) to (3.13) follows from (3.10). Since we are not interested in the object-level redshifts z_i themselves, but rather in an accurate estimate of the population-level mean, μ_b, we integrate over the galaxies' redshifts to obtain the marginal posterior

$$p(\mu_b, \sigma_b|X) \propto p(\mu_b, \sigma_b) \prod_i \int p(z_i|x_i) p(z_i|\mu_b, \sigma_b) dz_i. \tag{3.14}$$

Gaussian replacement of conditional densities: Instead of using directly the conditional density estimates $\hat{f}(z_i|x_i)$ derived from the data, we simplify the problem further by modelling $p(z_i|x_i)$ as normal distributions:

$$z_i|x_i \overset{\text{indep.}}{\sim} N(\hat{\zeta}_i, \hat{\tau}_i^2). \tag{3.15}$$

The replacement of the conditional density estimates $\hat{f}(z_i|x_i)$ via normal distributions has two main reasons. Firstly, by modeling both the population- and object-level distributions as Gaussians, the resulting Bayesian posterior distribution for the population mean can be calculated analytically. This allows us to scale our model to the large photometric data at hand. Secondly, and more subtly, because the redshift of the spectroscopic data has a limited range, the conditional densities of photo-z are truncated, and have no (or little) mass in z-regions with no (or little) spectroscopic samples, leading to conditional densities skewed away from the redshift boundaries. The Gaussian replacements, being symmetric by construction, appear to adjust

for these boundary effects.[4]

The estimate of the object-level mean $\hat{\zeta}_i$ is simply the mean of the conditional density estimate, $\hat{f}(z_i|x_i)$, while the object-level Gaussian variance $\hat{\tau}_i^2$ is obtained by computing the variance of $\hat{f}(z_i|x_i)$.[5] More precisely, by considering $\hat{f}(z_i|x_i)$ as a histogram evaluated on K bins, we have

$$\hat{\tau}_i^2 = \frac{1}{\sum_k \hat{f}_{m_k}(z_i|x_i)} \sum_k \hat{f}_{m_k}(z_i|x_i)(m_k - \hat{\zeta}_i)^2, \tag{3.16}$$

where $k = 1, \ldots, K$ specifies the histogram bin with location m_k, and $\hat{f}_{m_k}$ is the value of the conditional density (histogram) at location m_k.[6] The summary statistics $\hat{\zeta}_i$ and $\hat{\tau}_i^2$ are thus considered to be observed quantities, summarized by $\hat{X}_b := \{\hat{\zeta}_i, \hat{\tau}_i\}_{i=1}^{n_{Tb}}$.

Then (3.14) can be written as

$$p(\mu_b, \sigma_b|\hat{X}_b) \propto p(\mu_b, \sigma_b) \prod_i \int N(z_i|\hat{\zeta}_i, \hat{\tau}_i^2) N(z_i|\mu_b, \sigma_b^2) dz_i, \tag{3.17}$$

where $N(t|\theta, \phi^2)$ is the probability density function (pdf) of a normal distribution with mean θ and variance ϕ^2, evaluated at location t.

The integral in (3.17) can now analytically be solved (see Section B.1 in

[4] In our simulations, we also investigated a hierarchical model based on the conditional densities $\hat{f}(z_i|x_i)$ directly (without Normal replacement). In this case, we obtained the posterior distributions via MCMC sampling using only a small fraction of the target data due to computational limitations. Surprisingly, (given the results on the subset) the Normal-Normal model led to better estimates of the population means than using the conditional density estimates directly. This effect was consistently apparent for various simulation variations, e.g. using various subsets of the (non-weighted) full photometric sample (from the blue dashed distribution in Figure 3.2), and for different subsets of the weighted full photometric sample (from the green dashed distribution in Figure 3.2), demonstrating some robustness w.r.t. the underlying simulation. Further analysis might be helpful to study the effect of the Normal replacement in more detail in future work.

[5] With (3.15) we implicitly assume that $p(z_i|x_i, \hat{\zeta}_i, \hat{\tau}_i^2) = p(z_i|\hat{\zeta}_i, \hat{\tau}_i^2)$, that is, given $\hat{\zeta}_i$ and $\hat{\tau}_i^2$ the galaxy redshift distribution is independent of its photometric magnitudes x. This is a simplifying assumption, mainly to allow analytical calculation of the Bayesian posterior distribution as detailed in Section 3.3.3.

[6] We note that by bins k (with locations $m(k)$), we refer to the density (histogram) bins and not the tomographic redshift bins.

the Appendix for theoretical justifications) to obtain the (joint) marginal posterior density as

$$p(\mu_b, \sigma_b|\hat{X}_b) \propto p(\mu_b, \sigma_b) \prod_i N(\hat{\zeta}_i|\mu_b, \hat{\tau}_i^2 + \sigma_b^2). \tag{3.18}$$

We can now compute the posterior distribution of μ_b given $\sigma_b, \hat{X}_b$ (with uniform conditional prior density $p(\mu_b|\sigma_b)$) via

$$\mu_b|\sigma_b, \hat{X}_b \sim N(\tilde{\mu}_b, V_{\mu_b}), \tag{3.19}$$

with

$$\tilde{\mu}_b = \frac{\sum_i \frac{1}{\hat{\tau}_i^2+\sigma_b^2}\hat{\zeta}_i}{\sum_i \frac{1}{\hat{\tau}_i^2+\sigma_b^2}} \quad \text{and } V_{\mu_b}^{-1} = \sum_i \frac{1}{\hat{\tau}_i^2 + \sigma_b^2}, \tag{3.20}$$

with $V_{\mu_b}^{-1}$ being the total precision.

Since we are not interested in uncertainty on σ_b, we choose an empirical Bayesian approach by setting σ_b as a fixed value estimated from the data. An obvious estimate for σ_b is the MAP estimate of the marginal posterior $p(\sigma_b|\hat{X}_b)$ (shown in the Appendix (B.6)), which can be obtained numerically via sampling from $p(\sigma_b|\hat{X}_b)$. However, in our simulations, we found that the MAP estimate strongly and consistently underestimates σ_b. We thus choose a different strategy to estimate σ_b from the data as detailed in the next paragraph.

Finally, given an estimate of σ_b, an estimate of μ_b can be obtained analytically via (3.20), with $\tilde{\mu}_b$ being the MAP estimate of μ_b.

Population variance estimation via stacking of conditional densities: Here we demonstrate our strategy of estimating the population variance σ_b^2 via a "stacked estimate" of the marginal redshift population distribution $p_b(z)$ of galaxies within tomographic bin b. More precisely, we obtain an estimate $\hat{p}_b^{\text{stack}}(z)$ of $p_b(z)$ via averaging (stacking) over the conditional

densities within bin, that is,

$$\hat{p}_b^{\text{stack}}(z) = \frac{1}{n_{Tb}} \sum_j \hat{f}(z_j|\tilde{x}_j). \tag{3.21}$$

with $\tilde{x}_j, j = 1, \ldots, n_{Tb}$, being the photometric magnitudes of the observed galaxies within tomographic bin b. (The stacked estimator presented in (3.21) is justified later in this paragraph.)

An estimate for the redshift population variance σ_b^2 can then be obtained by calculating the variance of $\hat{p}_b^{\text{stack}}(z)$, via

$$\hat{\sigma}_b^2 = \frac{1}{\sum_k \hat{p}_{b,m(k)}^{\text{stack}}(z)} \sum_k \hat{p}_{b,m(k)}^{\text{stack}}(z)(m(k) - \hat{\mu}_b^{\text{stack}})^2, \tag{3.22}$$

where $k = 1, \ldots, K$ specifies the (density/histogram) bin with location m_k, $\hat{p}_{b,m(k)}^{\text{stack}}$ is the value of the stacked density (histogram) $\hat{p}_b^{\text{stack}}$ of bin b at location $m(k)$, and $\hat{\mu_b}^{\text{stack}}$ is the mean of $\hat{p}_b^{\text{stack}}(z)$. An estimate of the population standard deviation σ_b is then obtained by simply taking the square-root of (3.22).

We compute two different versions of (3.22). Firstly, we compute (3.22) via stacking over $\hat{f}(z_i|x_i)$, the galaxy conditional density estimates obtained via *StratLearn*. We denote this method as option *StratLearn*-Bayes (A). Secondly, we substitute the conditional density estimates $\hat{f}(z_i|x_i)$ with the normal replacements described in (3.15). We denote this option as *StratLearn*-Bayes (B).

We now justify the stacked estimator of the marginal redshift population distribution $p_b(z)$ in (3.21). Precisely, we can express $p_b(z)$ via

$$p_b(z) = \int p_b(z|x) p_b(x) dx, \tag{3.23}$$

with x being photometric magnitudes/colors, $p_b(x)$ being the distribution of the covariates (magnitudes/colors) of bin b, and $p_b(z|x)$ being the conditional distribution of redshift z given covariates x of bin b. By assuming the set

of photometric magnitudes/colors to be finite, and assuming that the conditional distribution of z given x is the same for all bins (i.e., $p_b(z|x) = p(z|x)$), we obtain

$$p_b(z) = \sum_i p(z|x = x_i) p_b(x = x_i). \tag{3.24}$$

Via *StratLearn*, we obtain an estimate of the conditional densities $p(z_i|x_i) \approx f(z_i|x_i)$. Since $z_i \overset{\text{iid}}{\sim} p(z)$, it holds that $p(z|x = x_i) = p(z_i|x = x_i) \approx \hat{f}(z_i|x_i)$. Further, we can approximate $p_b(x = x_i)$ by counting occurrences of x_i in the sample of observed magnitudes/colors within tomographic bin b. Alternatively, averaging over all estimated conditional densities $\hat{f}(z_i|x_i)$ of galaxies in bin b directly incorporates these occurrence frequencies, leading to the estimator in (3.21).

3.4 Estimation of the Population Distribution via Inverse-Propensity Score Weighting

In Section 3.3.3, we develop a hierarchical Bayesian framework to accurately estimate the photometric redshift population means (as demonstrated in the numerical evaluations in Section 3.5) within tomographic bins. While there is a longstanding argument that the population mean is most crucial for photo-z calibration (e.g., Amara and Refregier 2008), an accurate estimate of the redshift population shapes is of additional importance for the weak lensing analysis. Of course, by modelling the redshift populations per tomographic bin as Gaussians, the aim of our hierarchical Bayesian framework is not to obtain realistic distribution shapes. However, in this section we demonstrate that our method allows accurate population shape estimates quasi as a byproduct. More precisely, we demonstrate that propensity score estimates can be utilized for direct redshift calibration.

As described in Section 3.2.2, direct redshift calibration methods depend on the estimation of weights $\omega(x) = p_T(x)/p_S(x)$, used to reweight the spetroscopic sample (with true redshifts) to obtain an estimate of the redshift distribution of the photometric sample (per tomographic bin). The weights

$\omega(x)$ can also be expressed via

$$\omega(x) = \frac{p_T(x)}{p_S(x)} = \frac{p(s=1)}{p(s=0)} \frac{p(s=0|x)}{p(s=1|x)} \propto \left(\frac{1}{p(s=1|x)} - 1 \right). \tag{3.25}$$

We can thus obtain an estimate of the weights $w(x)$ by employing the inverse of the propensity score (inverse-PS), via the right-hand side of (3.25).

As a proof of concept, to demonstrate the ability of accurate population shape estimation using inverse-PS via direct redshift calibration, we apply the following procedure. Firstly, for each galaxy in the photometric and spectroscopic set, we obtain a *StratLearn* conditional density estimate $\hat{f}(z|x)$ as described in Section 3.3.1. Based on this estimate, each galaxy (in the photometric and spectroscopic set) is assigned to its respective tomographic bin, following the *StratLearn* binning strategy described in Section 3.3.2.[7]

For each tomographic bin, we now estimate a set of propensity scores (employing logistic regression as detailed in Section 3.5.1), based on the covariates of the spectroscopic source galaxies and photometric target galaxies in the respective tomographic bin. For each tomographic bin, we now reweight the binned spectroscopic data using the respective estimated inverse-PS weights (obtained via 3.25). Per tomographic bin, we then obtain an estimate of the photometric redshift distribution by considering the marginal sample of z in the weighted joint distribution. As a proof of concept, we provide a numerical demonstration of the inverse-PS direct calibration method in Section 3.5.5. In general, inverse-PS direct redshift calibration might suffer from the same large variance possibility as other direct redshift calibration weighting schemes (as discussed in Section 3.2.2). By estimating the weights via propensity scores, we consider the transfer and application of covariate balance diagnostics tools (as advised in the causal inference propensity score

[7]We note that the binning of the spectroscopic set (as previously performed by Wright et al. (2020)) is needed when performing direct redshift calibration of the binned photometric set, since the photometric bin assignment is based on (summaries) of the conditional density estimates (as described in Section 3.3.2). The conditional density estimates implicitly incorporate information of source redshift (through the fitting process described in Section 3.3.1). To prevent unmeasured confounding (information encoded in the spectroscopic redshift, but not in the magnitudes/colors) the same selection function has to be applied for source and target data.

literature, e.g. Austin and Stuart 2017, 2015 and Imai and van Dyk 2004, described in Section 2.3.3), to carefully assess/diagnose the inverse-PS direct calibration fit as part of future work.

3.5 Numerical Demonstrations

3.5.1 Simulation Study

We explore the performance of our new framework using a comprehensive set of realistic simulations introduced in Wright et al. (2020). The simulations aim to mimic the KiDS+VIKING-450 dataset, presented in Wright et al. (2019) and Hildebrandt et al. (2020), starting from the MICE2 simulation (Fosalba et al., 2015, Crocce et al., 2015, Carretero et al., 2015, Hoffmann et al., 2015) and based on a framework provided in Van Den Busch et al. (2020). In the following, we provide a summary of the simulated data employed in our study, and refer to Wright et al. (2020) for further details.

Photometric Survey: The simulated photometric data $\boldsymbol{D}_{\boldsymbol{T}}^{\boldsymbol{u}}$ (the superscript $\boldsymbol{u}$ refers to "unweighted", below we describe a pre-processing step to produce a shear-measurement weighted photometric sample as employed in the downstream scientific analysis) consists of $\sim 21 \times 10^6$ galaxies, for each of which a simulated measurement of its position, lensing convergence, morphological information, and model magnitudes in the $ugriZYJHK_s$-bands is provided. Magnitudes include photometric noise, realistic to KiDS survey data (Wright et al., 2020). A large proportion of galaxies ($\sim$17%) has a flux error greater than or equal to the flux measurement in at least one band. Figure B.2 in the Appendix illustrates the full pattern of such cases, which generally are considered as non-detections. The flux measurement of such non-detections was removed prior to our analysis and only placeholder/indicator values were available to indicate these non-detection cases. Due to missing information on the flux measurements, we treat these cases as "missing data". (The last two paragraphs of this section detail our processing of such "missing data" cases for the propensity score estimation and

the conditional density estimation). If further information of the flux measurements of these non-detections (e.g. a full Bayesian posterior, or upper limits) were available, such information could be incorporated to infer more realistic values/constraints of the non-detections. Another effective approach is to set the flux of such non-detections to zero (e.g., Benitez 2000, Jones and Heavens 2018). We note that our proposed method could technically incorporate any of such treatments of non-detections, and additional information and consideration of the non-detection cases might potentially improve our method in future work.

While the spectroscopic redshift z is of course unavailable for galaxies in the photometric set, a Bayesian redshift estimate z_B is provided, which was used in previous work (e.g., Wright et al. 2020) to apportion galaxies to the tomographic bins.

Finally, a cosmic shear-measurement weight $\hat{w}_i$ is provided for each galaxy, which relates to the quality of its shear measurement, and which filters though to the cosmological analysis via the weighted sample $\hat{w}_i \boldsymbol{D}^u_{Ti}$, which is the one employed for cosmological parameter estimation. Thus, for redshift calibration, the mean of the weighted distribution $p_T \propto \hat{w} p^u_T$ (within tomographic bin) is the target estimand of interest. As a preprocessing step, to obtain a sample $\boldsymbol{D}_T$ of the weighted distribution $p_T \propto \hat{w} p^u_T$, we resample the data $\boldsymbol{D}^u_T$ proportionally to the cosmic shear-measurement weights $\hat{w}$. In this way, we obtain our target sample $\boldsymbol{D}_T$, with $|\boldsymbol{D}_T| = 12.48 \times 10^6$ galaxies. Figure 3.2 in the Appendix demonstrates that there is a negligible difference in targeting the shear-measurement resampled distribution (obtained via the above preprocessing step) and targeting the shear-measurement weighted distribution (as done in Wright et al. 2020). Our results on the resampled data thus hold without loss of generality.

Spectroscopic Survey: A much smaller spectroscopic source data set $\boldsymbol{D}_S$, with $|\boldsymbol{D}_S| = 21,537$, galaxies is provided, composed of data from three surveys: zCOSMOS (9930 galaxies); DEEP2 (6919 galaxies); and VVDS (4688 galaxies), altogether spanning a redshift range $0.07 \leq z \leq 1.43$. The spectroscopic redshift distributions of the three surveys are illustrated in

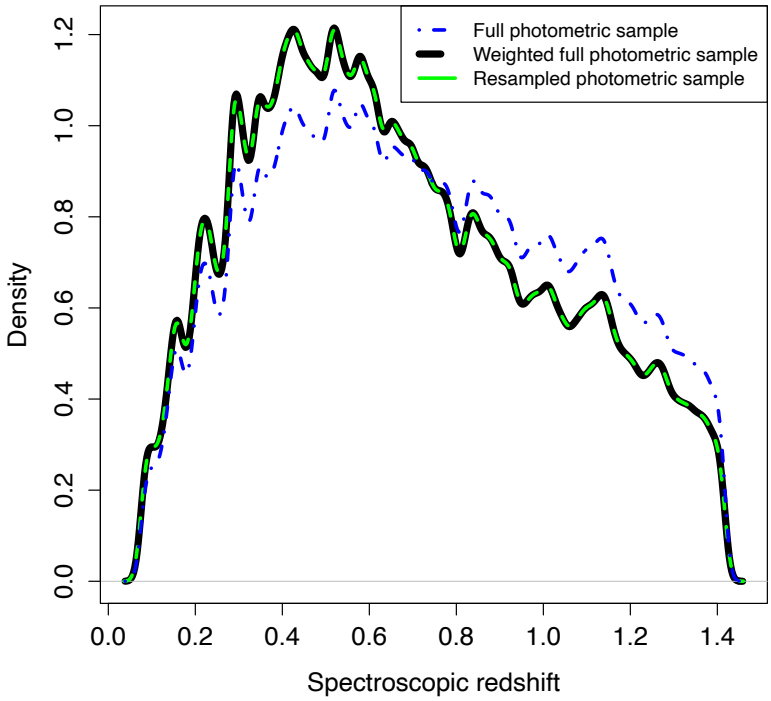


Figure 3.2: Spectroscopic (true) redshift distributions of the photometric survey (not known in practice), with and without incorporation of shear-measurements weights. The blue dashed line shows the redshift distribution of the full photometric sample D_T^u (before resampling). The black line shows the redshift density of the full photometric density weighted by the shear-measurment weights $\hat{w}$, and the green line illustrates the redshift density of the resampled sample D_T. The black and green line perfectly match, with a mean difference of $\sim 10^{-4}$, illustrating that there is negligible difference of targeting the resampled distribution (as in this study) and targeting the weighted distribution (as in Wright et al., 2020). A histogram of the cosmic shear-measurement weights $\hat{w}$ is provided in Figure B.1.

Figure 3.3. The spectroscopic source set is not a representative sample of the photometric target distribution (selection effects are described in Wright et al. 2020). Figure 3.4 illustrates the density of the spectroscopic source redshift distribution (red), and the (true) redshift distribution of the photometric simulated target data (blue; not available in practice). While both

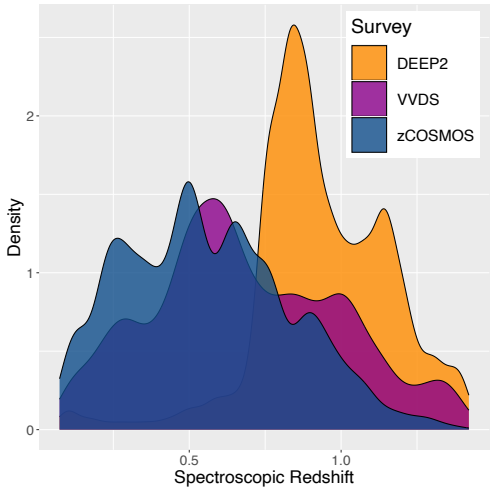


Figure 3.3: Spectroscopic redshift distributions of the three spectroscopic surveys used as source data.

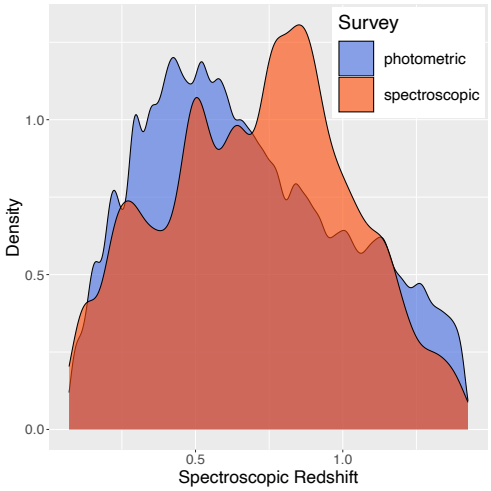


Figure 3.4: Spectroscopic (true) redshift distributions of the photometric simulated target data D_T (blue, not available in practice) compared with the spectroscopic source data (red).

distributions cover the same redshift range, the difference in densities is immediately apparent, an effect of the underlying covariate shift. For each galaxy in the spectroscopic set, the same set of covariates as provided for galaxies in the photometric set is available. In addition, an accurate spectroscopic measurement of the true redshift z is available for each galaxy, with measurement error that is negligible for our purposes.

To account for sample variance, 100 independent spectroscopic catalogues (with above described specifications) are provided, from 100 independent fields (lines-of-sights, abbreviated as LoS). The fields of the three spectroscopic surveys are independent of each other across the 100 LoS.

CHOICE OF COVARIATES AND MISSING DATA: To obtain the conditional density estimates for all objects in the photometric target data, we choose as covariates the r-band magnitude and the 8 colors: $u-g$, $g-r$, $r-i$, $i-Z$, $Z-Y$, $Y-J$, $J-H$, $H-Ks$), a setup previously adopted, e.g., by Wright et al. (2020)

or Izbicki et al. (2017). We note that using colors instead of magnitudes does not worsen the "missing data" pattern, as illustrated in Figures B.3 and B.2. As a preprocessing step, all covariates are scaled to have mean zero and standard deviation one. In the *StratLearn* framework, the missing data pattern has to be taken into account in the propensity score estimation step, and in the computation of the conditional density estimators within strata. For propensity score estimation, we use mean imputation of the 9 covariates to fill the missing values; we also add 9 binary indicator variables as dependent variables (main effects) to our logistic regression propensity score model, which describe the missingness of each covariate. Figure B.4 in the Appendix illustrates the distributions of the estimated propensity scores for source and target data. The support of the target propensity score distribution is well covered by the support of the source propensity score distribution, demonstrating the availability of source galaxies that match the covariate space of the target galaxies.

Computational Details: The computation of the conditional density estimators (*ker-NN* and *Series*) requires the calculation of Euclidean distances between the covariate vectors of each galaxy. In the missing data cases, we compute the pairwise distances of two galaxies only based on the covariate values with complete measurements (no missingness) for both galaxies (as previously conducted by Wright et al. (2020)). The large size of the photometric target set causes additional computational challenges: for prediction of the conditional densities on the target data, distance matrices between photometric target set and spectroscopic source set are required, which is computationally not feasible for the entire target set at once. We thus process the prediction on the photometric target set in batches of 60,000 target samples. Table 3.1 shows the strata composition of spectroscopic source and photometric target data for one random batch, illustrating that there is enough spectroscopic source/training data in each stratum to fit the conditional density estimators within strata separately. While there is a slight discrepancy between the average redshift in source (0.71) and target data (0.68) overall, most strata have well-balanced redshift means between source and target, an indicator of reduced covariate shift after the propensity score

Table 3.1: Composition of the five *StratLearn* strata. The number of galaxies and the average spectroscopic (true) redshift is presented in each source and target stratum. (Composition of one random batch of 60K photometric samples is shown for illustration).

Stratum	Set	#galaxies	Mean z
1	Source	6091	0.74
	Target	10217	0.74
2	Source	5036	0.77
	Target	11271	0.74
3	Source	4351	0.72
	Target	11957	0.72
4	Source	3668	0.65
	Target	12639	0.66
5	Source	2391	0.58
	Target	13916	0.57
All	Source	21537	0.71
	Target	60000	0.68

stratification, as justified in Chapter 2. Other batches demonstrate a similar pattern. To reduce the computational burden, for each LoS, we use a fixed set of hyperparameters for prediction of the conditional density estimators on all target batches. Individually for each LoS, we obtained the fixed hyperparameter set by optimizing (2.18) and (3.6) within each stratum separately, given one initial strata composition, with a randomly selected target batch.[8] Using batches of photometric target data has the advantage that distance matrices can be stored in memory, and predictions can be processed in parallel on several batches.[9]

[8]Optimization of (2.18) was performed by splitting the source data within each strata in a training and validation set (one half each). The parameters which led to the best predictive performance on the source validation sets (per strata) were then selected for each conditional density estimator (ker-NN and Series) separately. The final optimization in (3.6) was then performed on the same source strata validation sets, using the optimized ker-NN and Series source validation set predictions. The hyperparameters obtained for the five strata and for all 100 LoS are illustrated in the Appendix, Figures B.5 and B.6.

[9]We performed all computations on a CPU cluster employing up to ~150 CPU simultaneously.

3.5.2 Improved Bin Assignment Accuracy

In this section, we evaluate the accuracy of our new tomographic bin assignment, obtained via *StratLearn*-based conditional density estimates, as described in Section 3.3.2, and illustrated in Figure 3.1.

In Table 3.2, we compare our bin assignment with the standard practice of using z_B for the assignment, across five different classification performance metrics, demonstrating improvement in all of them. On average across the 100 LoS, our *StratLearn* binning assigns the photometric target galaxies to the correct tomographic bin in 62.2% of the cases (considering the five tomographic bins, and both edge bins separately). This is a substantial improvement over the z_B binning, with an accuracy of 52.5%. *StratLearn* improves both, the sensitivity (true positive rate) and the specificity (true negative rate) of tomographic bin assignment compared to z_B, thus also leading to an improvement of the balanced accuracy and Cohen's kappa, which take into account the imbalance of class (bin) proportions.[10] We note that the standard deviations of all performance measures across the 100 LoS is relatively low (Table 3.2), which demonstrates that the improvement is consistent throughout the 100 LoS (the z_B assignment is the same for all 100 LoS).

Table 3.2: Tomographic bin assignment performance evaluated over 100 LoS, comparing the *StratLearn* bin assignment (following Section 3.3.2) and the z_B bin assignment. The average (sd) of the performance metrics computed for each of the 100 LoS is reported for *StratLearn*. Using z_B, the bin assignment is consistently the same for all 100 LoS.

	StratLearn	z_B
Performance metric	mean (sd)	mean (sd)
Accuracy	0.622 (0.003)	0.526 (-)
Balanced Accuracy	0.718 (0.003)	0.706 (-)
Sensitivity	0.502 (0.006)	0.493 (-)
Specificity	0.934 (0.001)	0.918 (-)
Kappa	0.439 (0.006)	0.415 (-)

[10]The balanced accuracy is defined as (specificity + sensitivity)/2. The Cohen's Kappa measures the relative performance of the classifier with the performance of a random guess (based on the class frequency). Both metrics take on values between 0 and 1 (with 1 being a perfect classifier), and both metrics are useful performance measures for classification tasks with imbalanced class proportions.

Figures 3.5a and 3.5b show the confusion matrices of tomographic bin assignment using *StratLearn* and z_B (averaged over the 100 LoS). The confusion matrices demonstrate that *StratLearn* improves the bin assignment across all five tomographic bins (top five diagonal values), and most substantially in the second bin (with $z \in (0.3, 0.5]$), the one with the largest fraction of galaxies (21.7%). In this bin, *StratLearn* improves over the z_B bin assignment by more than 55%.

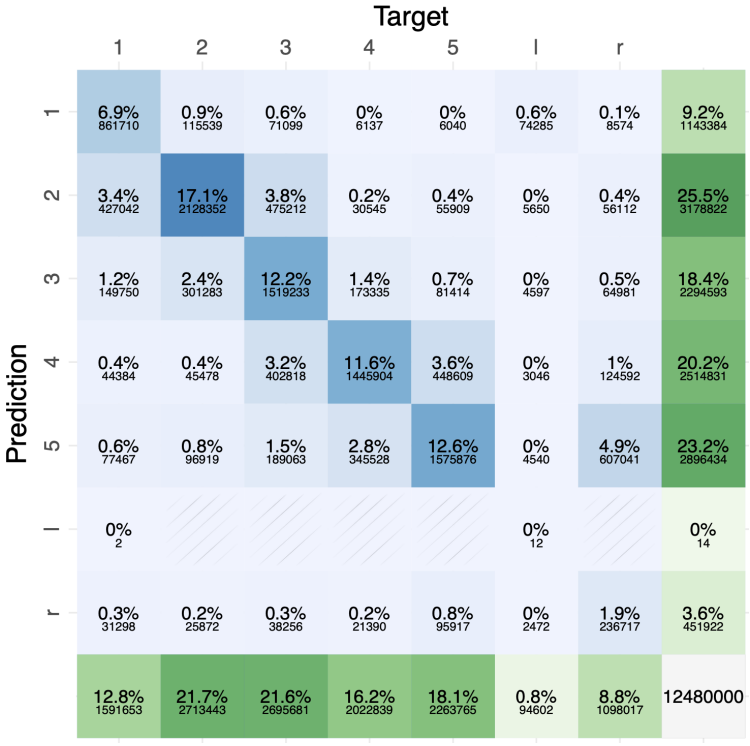
The heatmap in Figure 3.6 provides a visual comparison of the confusion matrices in Figures 3.5a and 3.5b. Green squares in Figure 3.6 correspond to an improvement of *StratLearn* over z_B, while pink squares correspond to better performance of the z_B assignment. The heatmap is computed by subtracting the diagonal values of Figure 3.5b, the bin assignment accuracy of z_B, from the diagonal values of Figure 3.5a, the bin assignment accuracy of *StratLearn*; on the off-diagonal, the sign is reversed, so that green (positive) values denote *StratLearn* improvement everywhere. The improvement of *StratLearn* is particularly strong in the top five diagonal squares, the five tomographic bins, which are of highest interest for the scientific analysis.

Figure 3.7 illustrates the changes in bin assignment of *StratLearn* vs. z_B, showing a moderate to strong disagreement of *StratLearn* and z_B in most of the bins. The reassignment of galaxies, and especially the improved bin assignment accuracy of *StratLearn*, might thus substantially improve cosmological results – this will be subject of a future, dedicated study.

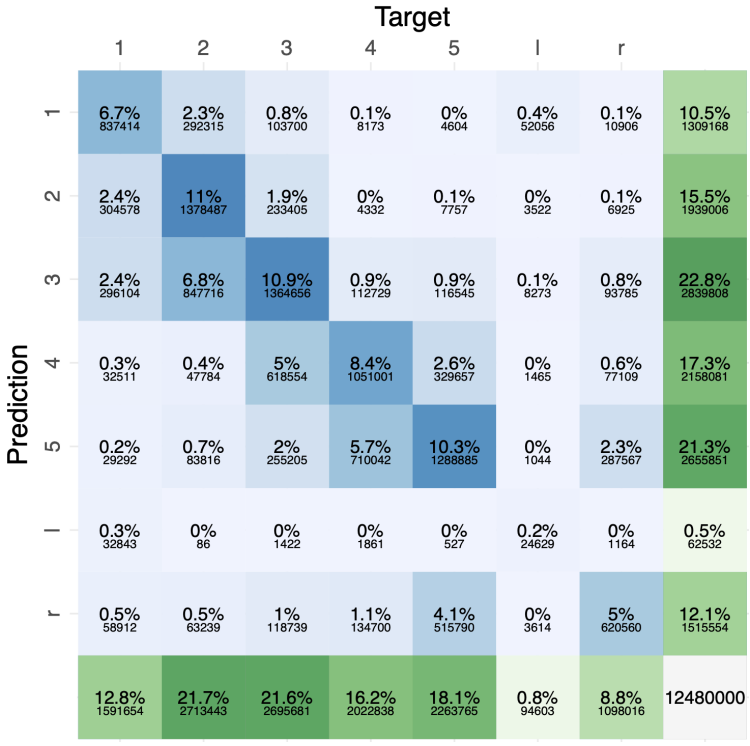
3.5.3 Improved Population Mean Estimates Accuracy

The main purpose of this study is to obtain accurate estimates of the (true) redshift population means within tomographic bins. The foremost criteria to evaluate redshift calibration methods (Wright et al., 2020) is the mean discrepancy

$$\mathbb{E}[\hat{\mu}_b - \mu_b^{\text{true}}] \simeq \frac{1}{L}\sum_{l=1}^{L}(\hat{\mu}_{b,l} - \mu_{b,l}^{\text{true}}), \quad (= \widehat{\text{bias}}_b) \tag{3.26}$$



(a) StratLearn tomographic bin assignment.



(b) z_B tomographic bin assignment.

Figure 3.5: (a) Confusion matrix of the *StratLearn* tomographic bin assignment (averaged over 100 LoS). (b) Confusion matrix of the z_B tomographic bin assignment.



Figure 3.6: Heatmap of confusion matrix (accuracy) differences between *StratLearn* and z_B. On the diagonal, the difference of *StratLearn* - z_B accuracy's is shown. Off-axis, the difference of z_B - *StratLearn* is shown. Thus higher values (in green) illustrate that *StratLearn* performs better than using the z_B estimate.

where $L = 100$ is the number of LoS, $\hat{\mu}_{b,l}$ the estimated mean redshift and $\mu_{b,l}^{\text{true}}$ the true redshift mean for LoS l for galaxies assigned to tomographic bin b. We note that (3.26) is not a bias in a strict statistical sense, since the true redshift mean (within tomographic bin) varies across lines of sights. However, being consistent with the notation of previous studies, we will loosely refer to (3.26) as "bias". In addition to (3.26), we are interested in the standard

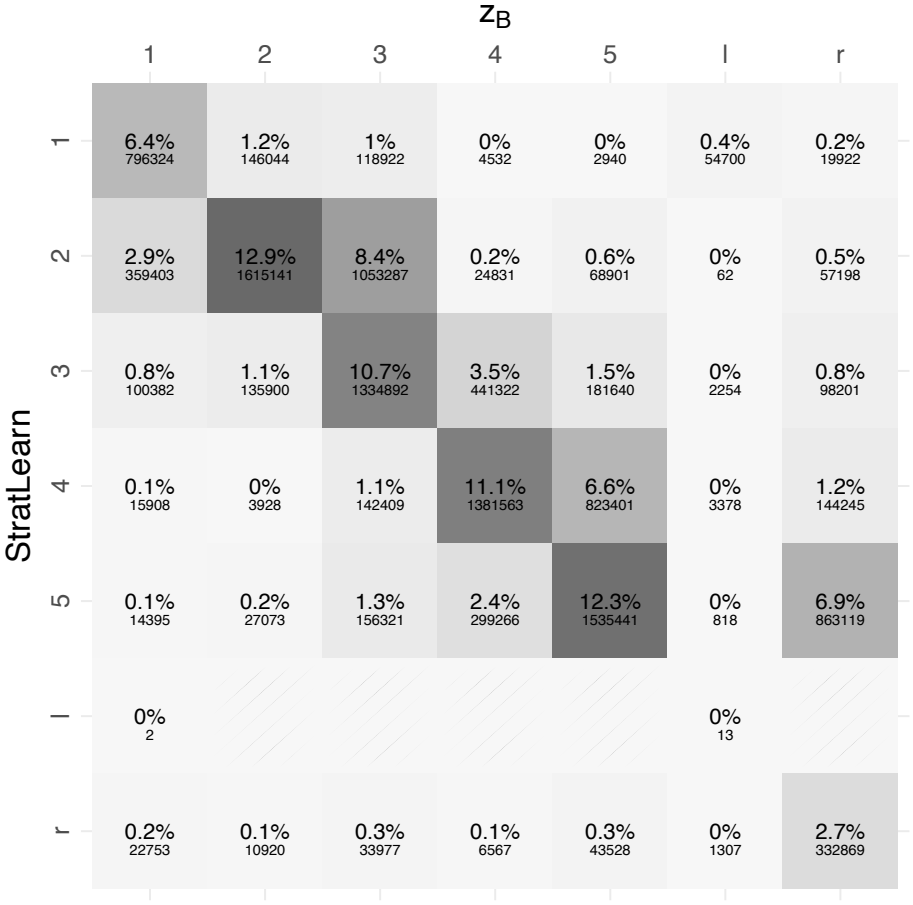


Figure 3.7: Changes in bin assignment using *StratLearn* vs z_B binning, averaged across the 100 LoS.

deviation, SD, of the mean differences across the 100 LoS:

$$\mathrm{SD}(\hat{\mu}_b - \mu_b^{\mathrm{true}}) = \sqrt{\frac{\sum_{l=1}^{L}(\mu_{b,l}^{\mathrm{diff}} - \widehat{\mathrm{bias}}_b)^2}{L-1}} \tag{3.27}$$

with $\mu_{b,l}^{\mathrm{diff}} = \hat{\mu}_{b,l} - \mu_{b,l}^{\mathrm{true}}$, for $l = 1, \ldots, L$ and $b = 1, \ldots, 5$.

Table 3.3 presents the bias results obtained for our novel *StratLearn*-Bayes method, with a comparison to the previously best performing method, the SOM. Based on our *StratLearn* binning, our method options *StratLearn*-Bayes (A) and (B) lead to an average absolute bias of 0.0053 and 0.0052 across the five tomographic bins, an improvement of $\sim 40\%$ w.r.t the SOM

method with z_B binning, which leads to an average absolute bias of 0.0085. We further note that the SOM (with z_B binning) method requires systematic quality cuts (further also denoted as "gold" sample, using the terminology of Wright et al. (2020)), which reduce the data size for the scientific analysis (we return to this point below).

We also apply the SOM calibration method using our new *StratLearn* binning, applying quality cuts as described in Wright et al. (2020), which leads to an increase of bias (0.105 absolute average bias) compared to SOM with the z_B binning (0.0085 absolute average bias). Using our *StratLearn*-Bayes model on the z_B bin assignment also leads to an increase in bias to 0.0141 and 0.0131 (on absolute average across the five bins). Such a reduction in performance could in fact be expected: our *StratLearn*-Bayes model is based on the modelling of the object level (galaxy) conditional density estimates, but with the z_B binning systematic errors are introduced in the assignment of galaxies per tomographic bin. Our *StratLearn*-Bayes framework is not designed for correction of such external errors (biases), which lead to systematic shifts of the population mean estimates. For instance, if the z_B galaxy bin assignment is correlated with the variance of *StratLearn* galaxy conditional density estimates, then the (tomographic bin) population mean estimate in (3.20) can be systematically shifted. We thus advise against the combination of *StratLearn*-Bayes based on z_B binning, and advocate for the use of *StratLearn*-Bayes via the *StratLearn*-based binning, which leads to the best performance.

Table 3.3: Mean discrepancy (bias) computed over 100 lines of sight for different calibration methods, and different bin assignment strategies. We abbreviate the *StratLearn* bin assignment as *SL*. (gold) denote the quality cuts applied to the data according to Wright et al. (2020). The Galaxies column shows the total number of galaxies (in millions) available in the five tomographic bins.

	Binning	Galaxies [M]	Bin 1	Bin 2	Bin 3	Bin 4	Bin 5	|Average|
StratLearn-Bayes (A)	*SL*	12.02	0.0123	-0.0076	0.0053	-0.0010	0.0001	0.0053
StratLearn-Bayes (B)	*SL*	12.02	0.0095	-0.0092	0.0047	-0.0013	0.0012	0.0052
SOM	*SL* (gold)	11.48	-0.0084	0.0022	0.0156	0.0117	0.0148	0.0105
StratLearn-Bayes (A)	z_B	10.90	0.0259	0.0127	0.0084	0.0003	-0.0231	0.0141
StratLearn-Bayes (B)	z_B	10.90	0.0228	0.0117	0.0071	-0.0002	-0.0236	0.0131
SOM	z_B (gold)	10.17	-0.0005	0.0036	0.0135	0.0147	-0.0102	0.0085

Table 3.4 shows the standard deviation (SD) population scatters from the

100 LoS. *StratLearn*-Bayes with *StratLearn* binning leads to slightly increased standard deviations of 0.0066 (option (A)) and 0.0067 (option (B)) on average across the five tomographic bins, compared to the SOM method based on z_B binning with an average of 0.0051. The results in Table 3.4 indicate that the standard deviation results are related to the binning strategy, rather than to the calibration method. Using SOM calibration on the *StratLearn* binning (with gold quality cuts) leads to comparable increase in SD of 0.0066 on average across the five bins. On the contrary, using the *StratLearn*-Bayes model applied on the z_B binning leads to a decrease in SD to an average of 0.0048 and 0.0047, even lower than applying SOM on the z_B binning.

In general, we note that the similarity in results of our *StratLearn*-Bayes options (A) and (B) demonstrate robustness with respect to the computation of the population variance (as described in Section 3.3.3). Both methods outperform the previously best method (SOM on z_B binning). Given a slight improvement of bias reduction, we propose the application of *StratLearn*-Bayes (B) as our best method.

Finally, our proposed method, *StratLearn*-Bayes (B), leads to a maximum bias within tomographic bins of $\Delta\langle z\rangle = 0.0095 \pm 0.0089$ (in bin 1). In contrast, using the previously best calibration method, SOM based on z_B binning, leads to maximum biases of 0.0135 ± 0.0052 and 0.0147 ± 0.0040 (in bin 3 and 4), both exceeding the potentially critical bias value of $\Delta\langle z\rangle > 0.01$. In addition, SOM based on z_B binning requires systematic quality cuts (gold quality cuts), which are not necessary for our methodology.

Table 3.4: As in Table 3.3, but showing standard deviation (SD) computed over 100 lines of sight for the different calibration methods, and different bin assignment strategies.

	Binning	Bin 1	Bin 2	Bin 3	Bin 4	Bin 5	|Average|
StratLearn-Bayes (A)	*SL*	0.0087	0.0065	0.0046	0.0052	0.0082	0.0066
StratLearn-Bayes (B)	*SL*	0.0089	0.0065	0.0045	0.0052	0.0082	0.0067
SOM	*SL* (gold)	0.0085	0.0064	0.0059	0.0058	0.0066	0.0066
StratLearn-Bayes (A)	z_B	0.0055	0.0048	0.0049	0.0037	0.0051	0.0048
StratLearn-Bayes (B)	z_B	0.0052	0.0048	0.0047	0.0036	0.0051	0.0047
SOM	z_B (gold)	0.0055	0.0061	0.0052	0.0040	0.0049	0.0051

3.5.4 Larger Sample Size for Weak Lensing Analysis

In Table 3.5, we show the absolute numbers of galaxies obtained via the different tomographic bin assignment strategies and quality cuts (gold). In bins 2,4, and 5, the number of galaxies is higher when using *StratLearn* binning compared to z_B. In bins 1 and 3, the number of galaxies is slightly higher using the z_B binning. Overall, due to the improved binning accuracy of *StratLearn*, there is an approximately 10% increase in the number of available galaxies for science (over all bins) when using *StratLearn* for tomography instead of z_B. Table 3.5 also provides the number of galaxies within bin after applying the gold selection, as introduced by Wright et al. (2020). We note that the gold selection cut is not needed when applying our *StratLearn*-Bayes approach, while it is a necessary step to obtain the SOM results (Wright et al., 2020). Thus, compared to the previously best combination of bin assignment and calibration method (Wright et al., 2020), our *StratLearn*-Bayes approach leads to an increase of galaxies available for science of ~18%.

Table 3.5: Sample sizes (in millions) within tomographic bins obtained via different bin assignment strategies and quality cuts (mean and standard deviation computed over 100 LoS). With (gold), we refer to the quality cuts as described in Wright et al. (2020).

		Bin 1 $(0.1, 0.3]$	Bin 2 $(0.3, 0.5]$	Bin 3 $(0.5, 0.7]$	Bin 4 $(0.7, 0.9]$	Bin 5 $(0.9, 1.2]$	Total $(0.1, 1.2]$
StratLearn	mean	1.14	3.18	2.29	2.51	2.90	12.03
	(sd)	(0.112)	(0.207)	(0.189)	(0.204)	(0.220)	
StratLearn (gold)	mean	1.06	2.94	2.21	2.51	2.75	11.48
	(sd)	(0.089)	(0.158)	(0.171)	(0.205)	(0.223)	
z_B	mean	1.31	1.94	2.84	2.16	2.66	10.90
	(sd)	(-)	(-)	(-)	(-)	(-)	
z_B (gold)	mean	1.15	1.91	2.36	2.10	2.65	10.17
	(sd)	(0.034)	(0.007)	(0.071)	(0.047)	(0.003)	

3.5.5 Population Distribution Estimates

Figure 3.8 illustrates the redshift population distributions of our *StratLearn*-Bayes framework per tomographic bin, with option (A) in blue, and option (B) in green. The true redshift distributions per tomographic bin (not known in practice) are shown with dashed lines. As expected, the Gaussian popula-

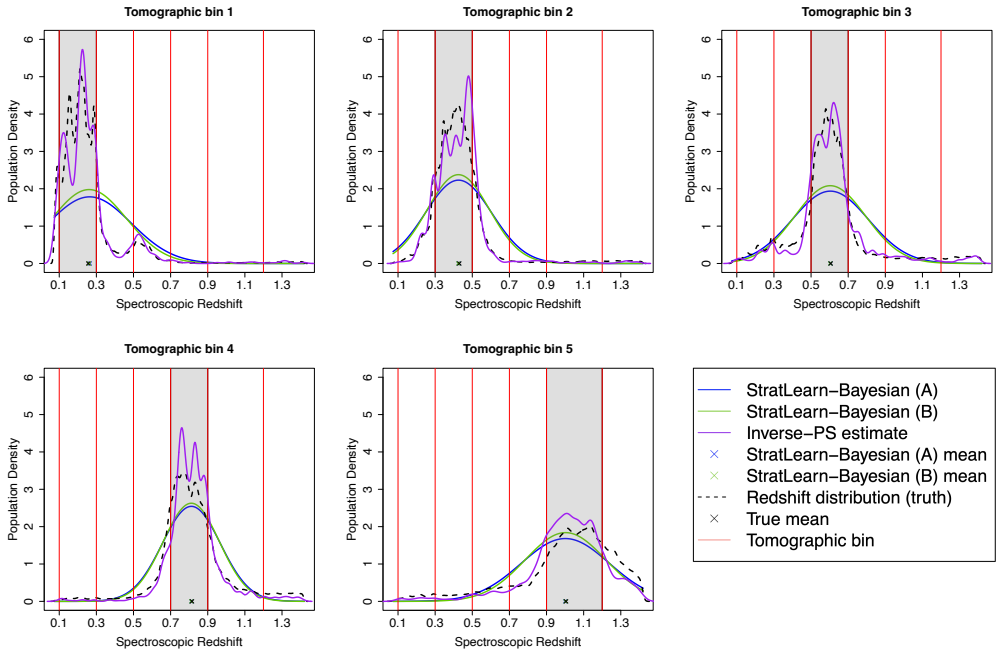


Figure 3.8: Redshift population distribution (estimates) per tomographic bin. The Figure shows the Gaussian population distributions of our *StratLearn*-Bayes frameworks with option (A) in blue, and option (B) in green. The true redshift distributions within tomographic bin (not known in practice) are shown with dashed lines. The mean estimates of the *StratLearn*-Bayes method and the true population means are added, being hardly distinguishable due to the small (desired) mean estimation errors. The inverse-PS redshift population distribution (estimates) per tomographic bin are illustrated in purple. For illustration purposes, results are shown for one LoS. (Other LoS exhibit a very similar pattern).

tion distributions do not cover the actual shapes of the true redshift populations well. We again emphasize that this was never the goal of our proposed models. We also illustrate the true population means, as well as the population mean estimates of our models on the x-axis. The true population means and our estimates are almost identical and visually barely distinguishable, a desired result, illustrating the accuracy of our population mean estimates.

While the population mean estimates are most crucial for photo-z calibration (e.g., Amara and Refregier 2008), in the weak lensing analysis, an an accurate estimate of the redshift population shapes is of additional importance. In Figure 3.8, we also illustrate the inverse-PS weighted redshift distributions per tomographic bin (in purple), obtained via the procedure described

in Section 3.4. The purple inverse-PS weighted distributions exhibit a similar shape as the true redshift population distributions within bin (dashed lines), recovering reasonably well the true photometric population distribution shapes.

To combine best of both worlds, in future work, we plan to combine our *StratLearn*-Bayes method for nearly unbiased population mean estimates, with the inverse-PS method to accurately capture the shape of the photometric redshift distributions. This will require a comprehensive assessment of the inverse-PS method, as well as the development of a principled combination strategy. We note that inverse-PS weighting is much more sensitive to inaccuracies in the estimated propensity scores than the *StratLearn* framework (based on stratification). Employing inverse-PS weighting will thus require careful considerations of the propensity score model and the estimated propensity score distributions.

3.6 Discussion

In this chapter, we propose a novel framework, *StratLearn*-Bayes, for improved photo-z calibration for weak lensing. We demonstrate that our *StratLearn* methodology, as proposed in Chapter 2, can successfully be employed to improve galaxy tomographic bin assignment via galaxy-per-galaxy photo-z conditional density estimates. Dispensing with the need of quality cuts (as applied by Wright et al. (2020)), the improved bin assignment leads to an increase of ~18% more galaxies available for the scientific analysis.

Constructing a hierarchical Bayesian model on summaries of the conditional densities, *StratLearn*-Bayes, then enables almost unbiased estimates of redshift population means per tomographic bin, the key quantity for photo-z calibration (Amara and Refregier, 2008, Wright et al., 2020). Our proposed method leads to substantially less bias than the previously best calibration method (SOM; Wright et al. 2020) evaluated on comprehensive simulations resembling the KiDS+VIKING-450 dataset (Wright et al., 2019, Hildebrandt et al., 2020), accompanied by a slight increase in uncertainty. In addition to our hierarchical Bayesian method for accurate estimation of the population

means, we provide proof of concept of how propensity scores can be used (via inverse-PS weighting) to obtain accurate redshift population shape estimates per tomographic bin.

We provide a detailed discussion of our method with prospective future work and science applications/implications in Section 5.2.

4

Estimation of Galaxy Luminosity Distributions from Incomplete X-ray and Optical Survey Data

Overview: The luminosity function specifies the distribution of source intensities in a population, providing information regarding the intrinsic characteristics of a class of astronomical objects and their cosmological evolution. In this project, we propose a science-driven hierarchical Bayesian framework to estimate the galaxy luminosity function in X-rays (via a broken power law model, formulated as a mixture of two Pareto distributions), combining non-representative X-ray and optical surveys. The astronomical surveys generally suffer from incompleteness in the low-intensity end, necessitating the application of incompleteness corrections generally estimated from simulations. While this is standard practice for samples selected in a single band, we introduce a methodology for the application of incompleteness corrections for samples selected in two (or more) bands (e.g. X-ray observations

of galaxies detected in optical surveys). We account for this incompleteness bias by incorporating an X-ray incompleteness function (estimated from simulations) and an optical incompleteness function (with parameters learned from the observed data) into the model. We avoid detection in the X-ray band for sources detected in the optical band by estimating the source intensities at the locations of the optical sources, using a Poisson model to adjust for background contamination. For the undetected sources in neither band or X-ray sources that are not detected in the optical band we use the corresponding incompleteness functions. The hierarchical model is fit via a Gibbs sampler. We evaluate our model on realistic simulations, mimicking data from the Chandra Deep Field Catalogue, and demonstrate improved recovery of the luminosity function even under high proportions of systematic incompleteness. We apply our novel method to data from the Chandra Deep Field South (CDFS).

4.1 Introduction

4.1.1 Scientific Motivation and Background

Estimating the distribution of source luminosities in a population has been a long-standing problem in astrophysics. Population luminosity distributions are commonly estimated by modelling the distribution of the flux. The flux, also referred to as observed luminosity or apparent brightness, is a measure of the power (amount of energy) radiated from an astronomical source (e.g. a galaxy or a star) over a fixed amount of time and area. Flux measurements allow astronomers to compare the relative amount of energy radiated from astronomical objects with different characteristics (e.g. object size or age). Modelling the distribution of the flux of different populations of astronomical sources provides information about the geometry and structure of the universe, e.g. about stellar evolution, allowing to test theoretical assumptions about the universe.

Historically, the distribution of fluxes in a population is characterized by counting the number N of sources with flux greater than a given source

flux S, also denoted as $N(> S)$. Astronomers are particularly interested in the $\log(N(> S)) - \log(S)$ relationship (in short $\log(N) - \log(S)$). In the literature, the $\log(N) - \log(S)$ relationship is commonly assumed to be linear, or piecewise-linear (Udaltsova, 2014, Stampoulis, 2017).

We are interested in estimating the galaxy luminosity distribution of a population in X-rays. X-ray is a type of high-energy electromagnetic radiation with wavelengths shorter than visible or UV light and typically longer than Gamma ray wavelengths. X-rays are emitted by a variety of astronomical objects, including galaxy clusters, stars, binary stars, supernova remnants, neuron stars and black holes (X-ray binaries). Since X-rays are mostly absorbed by the earth's atmosphere, X-ray surveys are commonly obtained via space based telescopes, such as the Chandra X-ray Observatory. The Chandra X-ray Observatory records the arrival time, the energy and the 2D sky coordinates of photons that arrive at the count-based detectors. Instead of recording the flux S, the cumulated flux is recorded as the number of photon counts. Given the source emitted photon counts one can compute the source intensity λ, which is the number of counts from a source reaching the observer (detector) for a certain exposure time and source area.

The arrival and measurement of photon counts underlies Poisson-like variability. Detector based uncertainties and biases, as well as external effects such as background contamination make the estimation of source luminosities (source intensities and flux) and consequentially the estimation of luminosity distributions via the $\log(N) - \log(S)$ a statistically challenging scientific problem. Below we outline major challenges when estimating the $\log(N) - \log(S)$ and prospects of how these challenges can be tackled in this and future work.

4.1.2 Challenges and Prospects

Some of the central challenges in the estimation of source luminosities from X-ray data is background contamination and incompleteness, as illustrated on a real data example in this section. Given the location of a source, the recorded number of counts in the source area is the sum of background and source counts. In a statistically naive approach one could simply ex-

tract the estimated number of background counts (obtained from areas expected to contain no sources) from the counts recorded in the source region to obtain an estimate of the source intensity (after accounting for exposure time and source area). While this leads to a poor estimate of the source intensities with possibly negative values, and with the emphasis that we do not espouse this approach, for illustration purposes, we present the $\log(N) - \log(S)$ curve obtained via this naive approach in Figure 4.1 (brown dots). More precisely, in Figure 4.1 we compute the $\log(N) - \log(S)$ via the empirical distribution function of the naive source intensity estimates, with $N(> S)_{\text{X-ray}} = (1 - \text{eCDF}(\hat{\lambda}^n_{\text{X-ray}}))N_{\text{X-ray}}$, with $N_{\text{X-ray}}$ being the number of detected X-ray sources in a pre-defined subregion of the Chandra Deep Field South (CDFS; details appear in Section 4.7), $\hat{\lambda}^n_{\text{X-ray}}$ being the source intensity estimates obtained via the described naive background subtraction, and eCDF being the abbreviation for empirical Cumulative Distribution Function. We note that we refer to $\log(N(> S)) - \log(S)$ as it is common in the literature, but actually compute $\log(N(> \lambda)) - \log(\lambda)$ based on the source intensities λ instead of the flux S. The source intensities λ (with units (count/s/cm^2)) differ from the flux (with units (energy/s/cm^2)) by some flux-to-count conversion factor $\kappa(S)$, which in turn depends on the flux S, that is $\lambda = S\kappa(S)$. Thus, $\log(\lambda) = \log(S) + \log(\kappa(S))$. The location and slope of the actual $\log(N(> S)) - \log(S)$ curve (obtained on the flux S instead of λ) will thus be altered by accounting for the flux-to-count conversion factor $\kappa(S)$ in a final step. (For consistency, all $\log(N(> S)) - \log(S)$ curves presented in this chapter are computed based on the source intensities λ instead of the actual flux S, implicitly assuming $\kappa(S) \equiv 1$.)

The illustrated X-ray $\log(N) - \log(S)$ curve in Figure 4.1 (brown dots) underlies the strong assumption that the galaxies detected in the X-ray survey reflect a complete, representative sample of the galaxy population in the defined subregion of the sky. However, this is unfortunately not the case, for various reasons. Firstly, X-ray source detection methods (e.g., Freeman et al. (2002)) apply a certain threshold to consider a pixel to belong to a source or to the background. To avoid background fluctuations being considered as potential sources, this threshold cannot be too low, which in turn might lead

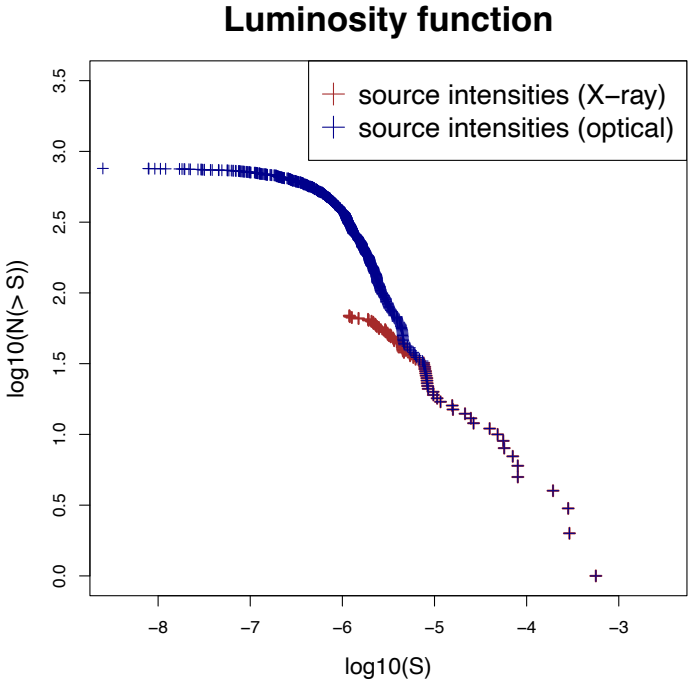


Figure 4.1: Illustration of a naive $\log(N) - \log(S)$ estimation via eCDF of source intensities obtained by simply subtracting the expected background counts from the source region. The estimated $\log(N) - \log(S)$ curves are shown for sources detected in X-ray (in brown), and for sources detected in an optical catalogue (in blue). In both cases the source samples are assumed to be complete, leading to an apparent discrepancy in the $\log(N) - \log(S)$ estimate, illustrating the strong necessity of incompleteness corrections.

to non-detections of faint sources. Secondly, the Poisson-like variability of measured source counts might lead to non-detections of faint sources if the fluctuation is close to a telescope specific detection limit. More precisely, a set of sources with identical source intensity will lead to differing numbers of observed counts on the detector due to measurement errors (and potentially other effects). If such a scatter is close to a sensitivity limit of the detector, then only sources with counts above the sensitivity threshold will be detected, leading to a biased sample and to bias in the inferred luminosity. This is a prevalent issue in astronomy and commonly known as Eddington bias (Teerikorpi, 2004). Thirdly, brighter and closer objects are more likely to be observed in many astronomical surveys (e.g., due detection sensitivity limits of the telescopes), leading to non-representative survey data, commonly referred to as Malmquist bias (Malmquist, 1920). These listed issues of systematic effects on the source detection probability are additionally convoluted with telescope and survey specific properties and effects, such as the exposure times, or the location of the source on the detectors etc. Overall, especially fainter and far away objects are more likely to be non-detections, indicating that the missing data mechanism is *non-ignorable* (Little and Rubin, 2019). Inferring parameters from the incomplete sample might thus lead to strongly biased and non-reliable estimates.

A possibility of probing the described incompleteness in X-ray source detections is the examination of additional surveys, e.g. using optical, radio, or previous X-ray surveys, to locate the positions of sources in the population. More precisely, here we extract the locations of sources detected in an optical catalogue (Arnouts et al. (2001); details appear in Section 4.7) that are located in our previously defined CDFS subregion. The location of sources detected in the optical catalogue can now be transferred to the X-ray survey, with the respective source areas defined by the X-ray point spread function (PSF) at the respective source locations (details appear in Section 4.7). Figure 4.2 shows a patch of the CDFS subregion with X-ray detected sources illustrated in yellow circles and sources detected in the optical catalogue (we also refer to these sources as *optical sources*) are added in blue circles. Figure 4.2 illustrates that the number of optical detected sources N_{optical} (in

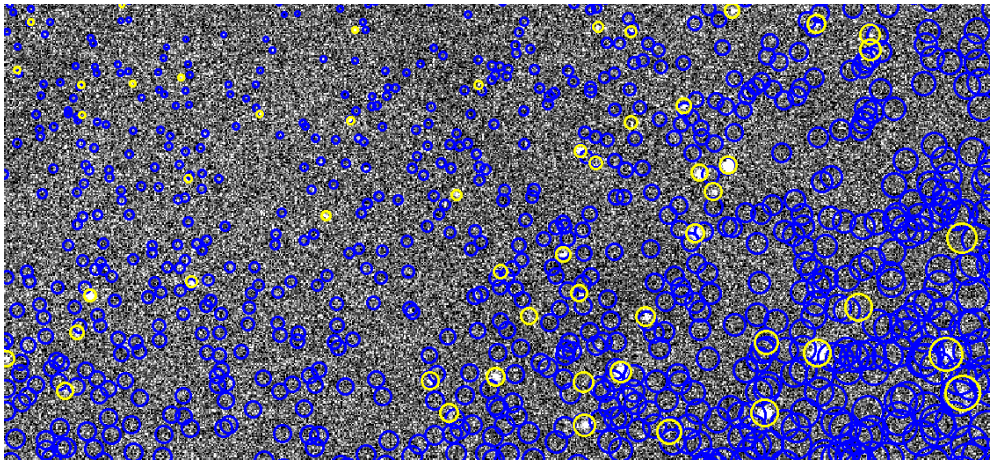


Figure 4.2: Patch of the CDFS subregion. Source areas of sources detected in the X-ray catalogue are shown in yellow circles, optical sources are illustrated in blue. Source areas (circular apertures around the source locations) are calculated via the X-ray PSF. Photon counts are shown in white. White points outside of the source areas are considered to be background counts.

blue) is much larger than the number of X-ray detected sources $N_{\text{X-ray}}$ (in yellow).

Given the locations and areas of the optical sources (in blue), we can now extract the respective X-ray counts and we can obtain a naive estimate of the source intensities. We again emphasize that we obtain the locations of the optical sources from optical catalogues, but then extract the source counts from the X-ray survey at the given source locations (with source areas obtained via the X-ray PSF). Figure 4.1 shows the $\log(N) - \log(S)$ obtained from the optical sources (blue dots) with $N(> S)_{\text{optical}} = (1 - \text{eCDF}(\hat{\lambda}^n_{\text{optical}}))N_{\text{optical}}$, with N_{optical} being the number of optical sources in our CDFS subregion (for simplicity here we only consider optical sources without overlap to any other optical source), and $\hat{\lambda}^n_{\text{optical}}$ being the source intensity estimates for optical sources obtained via the described naive background subtraction. Firstly, we note that around 50% of the naive source intensity estimates $\hat{\lambda}^n_{\text{optical}}$ are negative,[1] confirming the need of more sophisticated models for source intensity estimation. Secondly, and more interestingly, even though the naive

[1]For illustration of the naive $\log(N) - \log(S)$ estimates in Figure 4.1, we disregard any sources with negative naive source intensity estimates, with $\log(\hat{\lambda}^n_{\text{optical}})$ being undefined for $\hat{\lambda}^n_{\text{optical}} \leq 0$.

source intensity estimates might not be very accurate, the strong discrepancy between the $\log(N) - \log(S)$ obtained via the X-ray sources and the optical sources in Figure 4.1 are an apparent illustration and indicator for incompleteness of the X-ray detected sources. More precisely, the optical sources reach much farther into the low-intensity end of the population (negative source intensity estimates are disregarded in Figure 4.1).

Figure 4.2 evince many further challenges when estimating the $\log(N) - \log(S)$, particularly when accounting for the survey incompleteness. Matching the optical sources with X-ray sources, Figure 4.2 illustrates a large proportion of optical sources (in blue) without X-ray counterpart (in yellow). That is the optical source areas do not overlap with any X-ray source center. Most of the X-ray sources have an optical counterpart, that is, an optical source area overlaps with the center of the X-ray source. However, there are some X-ray sources without any recognisable optical counterpart. Figure 4.3 provides a simplified sketch illustrating the different source detection cases. We consider a source to be detected in both (X-ray and optical) if the source area of the optical detected source overlaps the center of the X-ray detected source. For illustration purposes, in Figure 4.3, we add undetected sources, neither detected in X-ray nor in the optical catalogue. We have to assume that there is a (large) proportion of sources that are neither detected in X-ray nor in the optical catalogue. Obviously, such cases are not visible in the CDFS patch in Figure 4.2. Finally, there is a large proportion of overlapping optical sources. Considering overlapping sources is a major challenge, since the observed counts in the source region are a combination of the various overlapping sources, which needs to be disentangled for accurate estimation of the source-wise intensities.[2]

In the last paragraphs, we have discussed some of the main topical challenges in $\log(N) - \log(S)$ estimation, (partly) tackled in our work. There

[2]We note that there is no overlap between detected X-ray sources, due to technical limitations of X-ray detection algorithms. The optical catalogue has finer resolutions and detects sources in closer proximity. However, in our case the source areas (for X-ray and optical sources, as illustrated in Figure 4.2) are calculated via the X-ray PSF, which is much larger than the optical PSF). This leads to the illustrated overlap between optical sources.

are additional, more detailed/technical issues, such as accurate matching of X-ray and optical sources, accuracy of source locations, incorporation of realistic PSFs, exposure and off-axis corrections, etc., which will be addressed throughout this chapter, or discussed as future work in Section 5.3. In the next section, we summarize previous approaches in the literature to estimate the $\log(N) - \log(S)$. In Section 4.1.4, we provide an overview of our strategy of how optical data can be leveraged to estimate the $\log(N) - \log(S)$ with principled incompleteness corrections.

4.1.3 Related Literature

Estimating the $\log(N) - \log(S)$ has a long-standing interest in astronomy with various proposed approaches. Some approaches - by no means exhaustive - are summarized here for an overview.

As illustrated in the previous section, careful incompleteness corrections and considerations of various roots of uncertainties are essential for reliable, unbiased $\log(N) - \log(S)$ estimation for galaxy luminosity distributions. Unfortunately, the majority of previous approaches estimating the $\log(N) - \log(S)$ assume complete surveys. Murdoch et al. (1973) and Crawford et al. (1970) model a linear $\log(N) - \log(S)$ relationship via maximum-likelihood (ML) estimation of the slope parameter. They consider measurement errors of the source counts via normal approximations. Maccacaro et al. (1982) and Schmitt and Maccacaro (1986) extend this framework, modeling a linear relationship, but replacing the Gaussian approximations with more realistic Poisson errors. Georgakakis et al. (2008) extends the linear model to a piecewise linear $\log(N) - \log(S)$ relationship, also considering Poisson errors. Wong et al. (2014) proposes a more flexible approach, by estimating the number and locations of different linear pieces. They replace the common ML approach by a more elaborated Expectation-Maximisation algorithm. Zezas et al. (2007) considers incorporation of uncertainty on the flux-to-count conversion parameters into their ML framework.

Udaltsova (2014) provides one of the first approaches to account for incompleteness when estimating the $\log(N) - \log(S)$. They propose a hierarchical

Bayesian framework to account for measurement errors and for source incompleteness. More precisely, they model the joint distribution of the complete data (the observed sources and missing sources) and the model parameters, and then marginalize over the missing sources to obtain the posterior distribution of the model parameters. Stampoulis (2017) extends upon the work of Udaltsova (2014) in several directions. They incorporate a more realistic and survey specific incompleteness function, provided by Wright et al. (2015), a key ingredient for unbiased inference on the $\log(N)-\log(S)$ parameters. They correct their model for mathematical inaccuracies in posterior derivations of Udaltsova (2014). In addition, they extend the model with several nuisance parameters leading to a more realistic joint distribution and modelling of the missing sources. Our research is in the line of work of Stampoulis (2017). We substantially extend upon the approach described by Stampoulis (2017) in several ways, as summarized below.

4.1.4 Contribution

In this project, we develop a novel Bayesian hierarchical framework which allows statistically principled combination of incomplete X-ray and incomplete optical survey data to estimate the galaxy luminosity distribution in X-rays. A major issue when estimating galaxy luminosity distributions is that low-intensity sources are most prevalent in the distribution but least likely to be detected in X-ray. Sources detected in optical catalogues tend to reach much farther into the low-intensity end of the luminosity distribution, as indicated in Figure 4.1. We leverage the optical data by combining X-ray detected and optical detected sources to obtain a combined source sample with less incompleteness in the low-intensity end. We account for the incompleteness in both surveys by incorporating an X-ray incompleteness function, known from previous simulations (Wright et al., 2015), and by modeling an optical incompleteness function, with parameters learned from the data. Figure 4.4 provides a sketch illustrating the X-ray incompleteness function (in red) and the incompleteness function of the combined optical and X-ray data (in blue) as a function of the source intensity λ, illustrating the (expected) improved detection probability of the combined sample for low-intensity sources.

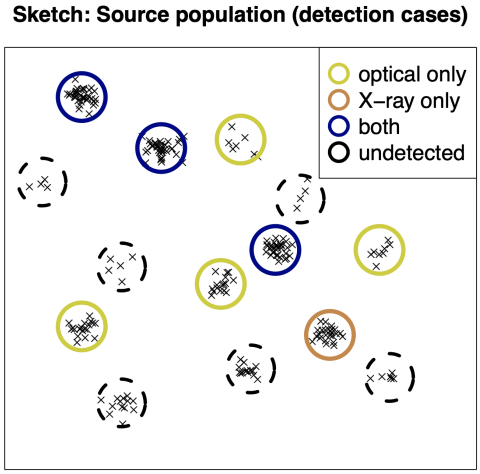


Figure 4.3: Sketch to illustrate the source population with the different source detection (and non-detection) cases in the X-ray and optical survey.

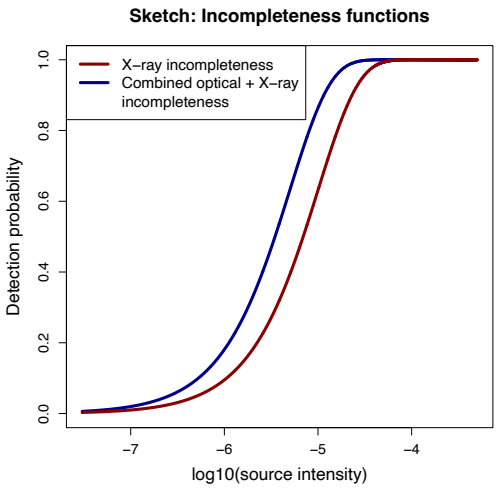


Figure 4.4: Sketch to illustrate possible incompleteness functions of the X-ray survey (red) and the combined X-ray and optical data (in blue). In the optical survey additional low-intensity sources are detected.

Given that a source is detected in the X-ray survey, the determination of its exact location is not straightforward in X-ray and accompanied by uncertainty. The determination of source locations in optical surveys is much more accurate than the determination of source locations in X-ray surveys. For sources detected in the X-ray and optical survey (and for sources detected in the optical survey only), we thus extract the X-ray counts at the source locations determined in the optical survey. By including latent Poisson variables for background and source counts for each source area, our hierarchical Bayesian framework allows principled modeling of the background contamination. As for the population distribution of source intensities, we consider the cases of a double Pareto and a single Pareto distribution, leading to a piece-wise linear or linear $\log(N) - \log(S)$ relationship. We first evaluate our novel framework on simulated data, introduced to resemble characteristics of the CDFS data. Based on the slightly more sophisticated double Pareto population distribution, we show that our model allows accurate estimation of the luminosity function. We then apply our model to data from the Chandra Deep Field Survey, discussing a single and double Pareto model fit.

To provide some more technical summary of our contribution, in our novel framework, we account for the incompleteness of the X-ray and optical survey by binning the data based on the (latent) source intensity λ. Given a fine grid on λ, within each bin, we can count the occurrences of X-ray and optical observed sources. Within each bin, these occurrences can then be modelled as Poisson distributions, with the Poisson rates being the product of the respective incompleteness function (X-ray or optical), population distribution (e.g. the double Pareto), the expected sample size, and the bin width, evaluated at the given λ. This strategy allows us to infer the parameters describing the optical incompleteness function as well as the population distribution parameters of interest. In Figure 4.4, we sketch the incompleteness function being dependent only on the source intensity λ, which is a simplified illustration. In reality, the X-ray incompleteness function provided by Wright et al. (2015) is dependent on additional nuisance parameters (e.g., the off-axis angle of the source, the telescope effective area, and the background rate). In our model, we marginalise over these nuisance parameters to obtain an X-ray incompleteness function that only depends on the source intensities λ. We develop a Metropolis-Hastings within Gibbs sampler to fit the hierarchical model, updating the source intensities and the background rate, as well as the population distribution and incompleteness parameters separately. Our novel model framework does not require the modeling and marginalization over missing sources as previously applied by Udaltsova (2014) and Stampoulis (2017) (for sources missing in the X-ray catalogue). To the best of our knowledge, this is the first principled approach that combines X-ray detected and optical detected sources to estimate the galaxy luminosity distribution in X-rays, while carefully accounting for incompleteness in the respective surveys.

The remainder of this chapter is organized as follows. In Section 4.2, we introduce the hierarchical Bayesian source model to estimate the source intensities and background rate, as well as the (single and double) Pareto source population distribution (assuming a complete X-ray survey). Section 4.2 mostly follows the work of Stampoulis (2017) and Udaltsova (2014), while we model the source intensities instead of the flux, which slightly changes

the framework. The remainder of the chapter is entirely new and original work developed in our research. In Section 4.3, we withdraw the assumption of completeness, introducing the combination of incomplete X-ray and optical surveys. In Section 4.3.1, we first introduce a toy example to illustrate and validate our binning strategy. In Section 4.3.2, we then introduce our binning strategy for the combination of X-ray and optical surveys, assuming the latent source intensities (as estimated in Section 4.2) to be known. In Section 4.3.3, we describe the X-ray incompleteness function as presented by Wright et al. (2015) and our marginalization over the nuisance parameters. In Section 4.4, we then present our full hierarchical model combining the source model presented in Section 4.2 and the incompleteness corrections introduced in Section 4.3, by assuming the realistic case of source intensities to be latent (unknown) and the observed sample (X-ray and optical) being incomplete. In Section 4.5, we introduce our Metropolis-within-Gibbs sampler to fit the hierarchical model. We evaluate our novel framework via simulations in Section 4.6. Finally, in Section 4.7, we apply our novel framework to estimate the $\log(N) - \log(S)$ on CDFS data with optical sources from Arnouts et al. (2001).

4.2 Source Model

In this section, we introduce the hierarchical source model to fit the luminosity function assuming complete X-ray survey data. Suppose that for an observation time of $\mathcal{T}$ seconds, we count Y_i photons in the X-ray field in source region i, for $i = 1, \ldots, N_{\text{obs}}$, and N_{obs} being the number of observed sources. We define the source region i as a circular area centered at the suspected source location with radius determined by the X-ray PSF.[3] We denote as a_i (in pixels) the area of the source region. In addition, we observe counts X from a background area A (in pixels) which does not contain any

[3]For simplicity, here we assume the PSF to follow a 2D normal distribution with i.i.d components. To expect around 99% of the photons to fall within the source region, we set the radius of the source area to be 3σ of the PSF, with σ being determined by the CDFS provided PSF map. We note that our model is flexible with regards to the PSF, and a more realistic PSF can easily be incorporated.

observed sources. Due to background contamination, the observed photon counts Y_i in source region i are the sum of the background counts $\mathcal{B}_i$ and the counts from the source $\mathcal{S}_i$, i.e.,

$$Y_i = \mathcal{S}_i + \mathcal{B}_i, \tag{4.1}$$

where $\mathcal{S}_i$ and $\mathcal{B}_i$ are independent. We emphasize that $\mathcal{S}_i$ and $\mathcal{B}_i$ are latent variables which are not observed. The arrival of photons at the detector can be modeled as a Poisson process, that is, Y_i and X are independent Poisson random variables. Further, the counts of photons originating from the source $\mathcal{S}_i$ and the counts from the background $\mathcal{B}_i$ (at the source region a_i) can be modeled as independent Poisson random variables. With the aim of estimating the source intensities λ_i, we now specify our model in more detail.

In addition to the exposure time $\mathcal{T}$ (in seconds s) and source intensity, λ_i (count/s/cm^2), the expected count from source i in its source region is affected by two known constants: (i) the proportion, r_i, of photons from the source that are expected to fall in the source region as determined by the PSF, and (ii) the telescope effective area, e_i (cm^2), at the source location, which characterizes the efficiency of the telescope. For simplicity, we can assume that all photons from the source fall within the source region and we thus set $r_i = 1$, for all $i = 1, \ldots, N_{\text{obs}}$. Then,

$$\mathcal{S}_i \big| \lambda_i \overset{\text{indep}}{\sim} \text{Poisson}(e_i \lambda_i \mathcal{T}). \tag{4.2}$$

We assume that the background does not systematically differ at different parts of the considered sky region. We thus model the background rate ξ (count/s/pixel) to be the same across different source regions. The observed photon count in the background can then be modeled by

$$X \big| \xi \sim \text{Poisson}(A \xi \mathcal{T}). \tag{4.3}$$

Given the background rate ξ, the background count $\mathcal{B}_i$ in source region i is

then modeled as

$$\mathcal{B}_i \big| \xi \overset{\text{indep}}{\sim} \text{Poisson}(a_i \xi \mathcal{T}). \tag{4.4}$$

As a sum of Poisson variables, the observed count in region i, $Y_i = \mathcal{S}_i + \mathcal{B}_i$, thus also follows a Poisson distribution, precisely,

$$Y_i \big| (\lambda_i, \xi) \overset{\text{indep}}{\sim} \text{Poisson}\left((a_i \xi + e_i \lambda_i)\mathcal{T}\right). \tag{4.5}$$

Given the data $\mathbf{D} = \{a_i, Y_i, e_i, X\}_{i=1}^n$, we can now specify the likelihood function $L(\xi, \boldsymbol{\lambda} \big| \mathbf{D}) = p(D|\xi, \boldsymbol{\lambda})$ for the unknown parameters $(\xi, \boldsymbol{\lambda})$, with $\boldsymbol{\lambda} = (\lambda_1, \ldots, \lambda_{\text{Nobs}})$, via

$$L(\xi, \boldsymbol{\lambda} \big| \mathbf{D}) = \exp(-A\mathcal{T}\xi) \frac{(A\mathcal{T}\xi)^X}{X!} \prod_{i=1}^{n} \exp[-(a_i \xi + e_i \lambda_i)\mathcal{T}] \frac{[(a_i \xi + e_i \lambda_i)\mathcal{T}]^{Y_i}}{Y_i!} \tag{4.6}$$

We specify the prior distribution on ξ as $\text{Gamma}(a_\xi, b_\xi)$.[4] Below we detail the choice of the source population (prior) distribution on $\boldsymbol{\lambda}$.

Population distribution of source intensities: There has been extensive work on determining the best fit population distribution for source luminosities (in our case expressed as source intensities) (e.g., Wong et al., 2014, Udaltsova, 2014, Stampoulis, 2017). Udaltsova (2014) shows that a linear $\log(N) - \log(S)$ relation corresponds to a Pareto population distribution (single power-law), and more generally, a piece-wise linear $\log(N) - \log(S)$ relationship with m pieces corresponds to an m-component broken power-law population model (formulated as a mixture of truncated Pareto distributions and one (untruncated) Pareto distribution (Lemma 2, Udaltsova, 2014)).

$$p(\lambda|\theta_1, \ldots, \theta_m, \tau_1, \ldots, \tau_m) = \sum_{j=1}^{m} \left[\prod_{k=1}^{j-1} \left(\frac{\tau_{k+1}}{\tau_k} \right)^{-\theta_k} \right] \left(\frac{\theta_j}{\tau_j} \right) \left(\frac{\lambda_i}{\tau_j} \right)^{-(\theta_j+1)} 1_{\{\tau_j \le \lambda < \tau_{j+1}\}}, \tag{4.7}$$

[4] The probability density function (pdf) of the gamma distribution is given by $p(x|\alpha, \beta) = (\beta^\alpha / \Gamma(\alpha)) x^{(\alpha-1)} \exp(-\beta x)$, with shape parameter α and rate parameter β. The mean is given by $\mu = \alpha/\beta$ and the variance by $\theta = \alpha/\beta^2$.

with knots τ_j and power-law slopes θ_j, with $j = 1, \ldots, m$. As common in the literature, we are particularly interested in a linear $\log(N) - \log(S)$, and a two-component piece-wise linear $\log(N) - \log(S)$ relationship.

Linear $\log(N) - \log(S)$**:** Following Udaltsova (2014), to obtain a linear $\log(N) - \log(S)$ relationship, source intensities above a threshold τ_1 follow a (single) Pareto distribution, that is, $\lambda_i \overset{\text{iid}}{\sim}$ single-Pareto(τ_1, θ_1), for $i = 1, \ldots, N_{\text{obs}}$, with (single) Pareto density f_{SP} via

$$f_{\text{SP}}(\lambda_i; \tau_1, \theta_1) = \theta_1 \tau_1^{\theta_1} \lambda_i^{-(\theta_1+1)}, \quad \lambda_1 > \tau_1, \quad \theta_1 > 0. \tag{4.8}$$

The parameter τ_1 describes the population source intensity minimum, describing the theoretical luminosity limit at which we can still observe a source.

Two-component piece-wise linear $\log(N) - \log(S)$**:** For $m = 2$, the source intensities above the threshold τ_1 follow a double-Pareto distribution. That is, $\lambda_i \overset{\text{iid}}{\sim}$ double-Pareto$(\tau_1, \theta_1, \tau_2, \theta_2)$, for $i = 1, \ldots, N_{\text{obs}}$, with double-Pareto density f_{DP} via

$$\begin{aligned} f_{\text{DP}}(\lambda_i; \theta_1, \theta_2, \tau_1, \tau_2) = &\left(\frac{\theta_1}{\tau_1}\right)\left(\frac{\lambda_i}{\tau_1}\right)^{-(\theta_1+1)} 1_{\{\tau_1 \le \lambda < \tau_2\}} \\ &+ \left(\frac{\tau_2}{\tau_1}\right)^{-\theta_1}\left(\frac{\theta_2}{\tau_2}\right)\left(\frac{\lambda_i}{\tau_2}\right)^{-(\theta_2+1)} 1_{\{\tau_2 \le \lambda < \infty\}}, \end{aligned} \tag{4.9}$$

where τ_1 again describes the population minimum source intensity, τ_2 describes the breakpoint, θ_1 and θ_2 describe the first and second power law slope.

In the following sections, to keep the description of our model flexible in the choice of the population distribution, we employ a general notation $f_P(\lambda; \boldsymbol{\theta})$ that refers to both, the (single) Pareto population distribution with $\boldsymbol{\theta} := \{\theta_1, \tau_1\}$, and the double-Pareto population distribution with $\boldsymbol{\theta} := \{\theta_1, \theta_2, \tau_1, \tau_2\}$.

As for prior distributions on the population parameters, we follow Udaltsova (2014) and Stampoulis (2017) and choose Gamma priors for τ_1, and for

the m power law slopes $(\theta_1, \ldots, \theta_m)$, i.e., $\tau_1 \sim \text{Gamma}(a_\tau, b_\tau)$ and $\theta_j \sim \text{Gamma}(a_j, b_j)$, $j = 1, \ldots m$ (in our case $m \in \{1, 2\}$). The prior on τ_2 needs some more consideration to preserve the increasing order of $\tau_1 < \tau_2$. We choose a truncated Normal prior distribution on the conditional distribution of τ_2 given τ_1, i.e., $\tau_2|\tau_1 \sim \text{truncated-Normal}(\mu_\tau, \sigma_\tau^2, a_0 = \tau_1, b_0 = \infty)$.[5][6]

Full posterior distribution (assuming a complete X-ray sample): Given the data $\mathbf{D} = \{a_i, Y_i, e_i, X\}_{i=1}^n$, we can now derive the posterior distributions of the parameters $\xi, \boldsymbol{\lambda}, \boldsymbol{\theta}$, via

$$p(\xi, \boldsymbol{\lambda}, \boldsymbol{\theta}|\boldsymbol{D}) = p(\boldsymbol{D}|\xi, \boldsymbol{\lambda})p(\xi, \boldsymbol{\lambda}, \boldsymbol{\theta}) \tag{4.10}$$

$$= L(\xi, \boldsymbol{\lambda}|\boldsymbol{D})p(\boldsymbol{\lambda}|\boldsymbol{\theta})p(\boldsymbol{\theta})p(\xi), \tag{4.11}$$

which holds, by assuming that the background rate ξ is a priori independent of the source intensities $\boldsymbol{\lambda}$, and the parameters describing the source intensity population distribution $\boldsymbol{\theta}$, and due to the conditional independence of $\boldsymbol{D}$ and $\boldsymbol{\theta}$ given $\boldsymbol{\lambda}$. That is the observed data $\boldsymbol{D}$ depends on the parameters describing the source population only through the latent source intensities $\boldsymbol{\lambda}$.

4.3 Methodology – Principled Modelling of Multiple Survey Incompleteness

In the previous section, we detailed a hierarchical Bayesian approach for estimation of the source intensities $\boldsymbol{\lambda}$, the background rate ξ, and the population distribution parameters $\boldsymbol{\theta}$, assuming a complete sample of X-ray sources. In Section 4.3.2, we drop the unrealistic assumption of completeness. More precisely, we consider the scenario described in Section 4.1.2, with a sample of

[5] A truncated-Normal parameterized by $(\mu_\tau, \sigma_\tau^2, a_0, b_0)$ is obtained by setting the pdf of a normal distribution with mean μ_τ and variance σ_τ^2 to 0 for all values below a_0 and for all values above b_0 (re-normalising the non-truncated part of the density). With $f_{\text{truncNorm}}(x; \mu_\tau, \sigma_\tau^2, a_0, b_0)$, we refer to the truncated-Normal pdf.

[6] We note that the prior that we choose for τ_2 differs from the prior adopted by Stampoulis (2017), as we found the interpretation of the truncated-Normal parameters more straightforward and simpler to modify.

sources detected in an X-ray and/or optical catalogue, with both catalogues being incomplete. In Section 4.3.2, we assume the source intensities $\boldsymbol{\lambda}$ of the observed sources (as estimated in Section 4.2) to be known.

In our novel framework in Section 4.3.2, we employ a discrete binning based on the source intensity λ, and Poisson modeling of the observed sources within λ-bin – we also refer to this strategy as "Poisson binning". In the following Section 4.3.1, we first introduce a simplified toy example to illustrate the Poisson binning strategy, before developing our incompleteness correction strategy in Section 4.3.2.

4.3.1 Poisson Binning – Toy Gaussian Example

In this section, we consider N_T to be the number of observed sources with $N_T \sim \text{Pois}(\psi_T)$, where ψ_T is the expected (unknown) number of sources in the population. For simplicity, we assume the toy source intensity population distribution $p(\Lambda)$ being Gaussian, with

$$\Lambda_i \overset{iid}{\sim} N(\mu_T, \sigma_T), \quad i = 1, \ldots, N_T, \tag{4.12}$$

where Λ_i are the observed toy source intensities, with population mean μ_T and population standard derivation σ_T. We employ the notation $\boldsymbol{\Lambda} = (\Lambda_1, \ldots, \Lambda_{N_T})$. (We note that, for clarity, in this toy example we use slightly different notation to distinguish it from sections directly dealing with the scientific problem.)

Likelihood function without binning: In this toy example it is straightforward to express the likelihood function of the model parameters μ_T, σ_T, and ψ_T, given the data (observed source intensities) $\boldsymbol{\Lambda}$. Precisely, the (unbinned) likelihood function L_u can be formulated as

$$L_u(\mu_T, \sigma_T, \psi_T | \boldsymbol{\Lambda}) = \left[\prod_i^{N_T} f_N(\Lambda_i; \mu_T, \sigma_T) \right] \frac{\psi_T^{N_T} e^{-\psi_T}}{N_T!}, \tag{4.13}$$

with $f_N(\Lambda_i; \mu_T, \sigma_T)$ being the normal density with mean μ_T and standard

deviation σ_T, evaluated at location Λ_i. The log-likelihood then follows (up to a constant) via

$$\ell_u(\mu_T, \sigma_T, \psi_T | \mathbf{\Lambda}) = -\psi_T + \sum_i^{N_t} \log(f_N(\Lambda_i; \mu_T, \sigma_T)) + N_T \log(\lambda). \tag{4.14}$$

Likelihood function with Poisson binning: We now illustrate the Poisson binning strategy. Precisely, let $k = 1, \ldots, K$ be bins on Λ, with $\Lambda^{(k)}$ being the position of bin k and $\Delta_{\Lambda^{(k)}}$ being the size of bin k. Now, let $m_\Lambda^{(k)}$ be the number of sources with source intensity falling within bin k. We then have

$$m_\Lambda^{(k)} \sim \text{Pois}(f_N(\Lambda^{(k)}; \mu_T, \sigma_T)\Delta_{\Lambda^{(k)}}\psi_T). \tag{4.15}$$

Given (4.15), we can then define the binned likelihood function $L_b(\mu_T, \sigma_T, \psi_T | \mathbf{\Lambda})$ via

$$L_b(\mu_T, \sigma_T, \psi_T | m_\Lambda^{(k)}) = \prod_k \frac{e^{-f_N(\Lambda^{(k)}; \mu_T, \sigma_T)\Delta_{\Lambda^{(k)}}\psi_T}[f_N(\Lambda^{(k)}; \mu_T, \sigma_T)\Delta_{\Lambda^{(k)}}\psi_T]^{m_\Lambda^{(k)}}}{m_\Lambda^{(k)}!}. \tag{4.16}$$

With $\prod_k \psi_T^{m_\Lambda^{(k)}} = \psi_T^{N_T}$, (4.16) can be simplified to

$$\begin{aligned} L_b(\mu_T, \sigma_T, \psi_T | m_\Lambda^{(k)}) \propto\ & e^{-\psi_T \sum_k f_N(\Lambda^{(k)}; \mu_T, \sigma_T)\Delta_{\Lambda^{(k)}}} \\ & \left[\prod_k [f_N(\Lambda^{(k)}; \mu_T, \sigma_T)\Delta_{\Lambda^{(k)}}]^{m_\Lambda^{(k)}}\right] \psi_T^{N_T} \end{aligned} \tag{4.17}$$

Up to a constant, the binned log-likelihood $\ell_b(\mu_T, \sigma_T, \psi_T | m_\Lambda^{(k)})$ can then be written as

$$\begin{aligned} \ell_b(\mu_T, \sigma_T, \psi_T | m_\Lambda^{(k)}) =& -\psi_T \sum_k f_N(\Lambda^{(k)}; \mu_T, \sigma_T)\Delta_{\Lambda^{(k)}} \\ &+ \sum_k m_\Lambda^{(k)} \log(f_N(\Lambda^{(k)}; \mu_T, \sigma_T)) \\ &+ \sum_k m_\Lambda^{(k)} \log(\Delta_{\Lambda^{(k)}}) + N_T \log(\psi_T). \end{aligned} \tag{4.18}$$

If the bin range covers the support of the population distribution f_N (up to the distribution tails with negligible mass), and with a sufficiently fine bin grid (small bin sizes), we can approximate $\sum_k f_N(\Lambda^{(k)}; \mu_T, \sigma_T)\Delta_{\Lambda^{(k)}} \approx 1$. With $\sum_k m_\Lambda^{(k)} \log(\Delta_{\Lambda^{(k)}})$ being a constant term, (4.18) simplifies to

$$\ell_b(\mu_T, \sigma_T, \psi_T | m_\Lambda^{(k)}) = -\psi_T + \sum_k m_\Lambda^{(k)} \log(f_N(\Lambda^{(k)}; \mu_T, \sigma_T)) + N_T \log(\psi_T) \tag{4.19}$$

Finally, if we make the bin grid to be fine enough, s.t. $m_\Lambda^{(k)} \in \{0, 1\}$, we can express (4.20) via

$$\ell_b(\mu_T, \sigma_T, \psi_T | m_\Lambda^{(k)}) \approx -\psi_T + \sum_i^N \log(f_N(\Lambda_i; \mu_T, \sigma_T)) + N_T \log(\psi_T). \tag{4.20}$$

such that the locations $\Lambda^{(k)}$ of the bins with observed source, that is $m_\Lambda^{(k)} = 1$, are approximately (or even exactly) at the observed source intensity Λ_i within the bin. More precisely, if the bin size $\Delta_{\Lambda^{(k)}}$ converges towards zero ($\Delta_{\Lambda^{(k)}} \to 0, \forall k = 1, \ldots, K$), then ℓ_b converges towards ℓ_u. That is, if we use a sufficiently fine grid on Λ then the (unbinned) log-likelihood in (4.14) can be well approximated via the binned log-likelihood in (4.19) (via (4.20)). We now illustrate the approximation of ℓ_u via ℓ_b in a simple numerical example.

Numerical toy example for Poisson binning: Let $\psi_T = 100$ be the expected population size, $\mu_T = 1$ be the population mean, and $\sigma_T = 0.2$ be the population standard deviation. Via $N_T \sim \text{Poisson}(\psi_T = 100)$, we obtain $N_T = 91$ sources in this specific example. In the next step, we sample Λ_i, $i = 1, \ldots, N_T = 91$, i.i.d. source intensities from (4.12).

For illustration purposes, we assume the true values of σ_T, N_T, and $\mathbf{\Lambda}$ to be known, and calculate the log-likelihood of the unbinned model via (4.14), and for binned model via (4.19) for a sequence of values for the mean μ_T. In this example, we bin the data equidistantly, s.t. $m_\Lambda^{(k)} \in \{0, 1\}, \forall k = 1, \ldots, K$, leading to $K = 40,960$ bins. Figure 4.5 compares the log-likelihoods (up to a constant) of the binned and unbinned model, showing that ℓ_b (in orange)

well approximates ℓ_u (in black) in this numerical example, both leading to the maximum log-likelihood at the true population mean $\mu_T = 1$.

In the next section we show that the binning strategy provides a convenient framework to account for incompleteness in our scientific problem.

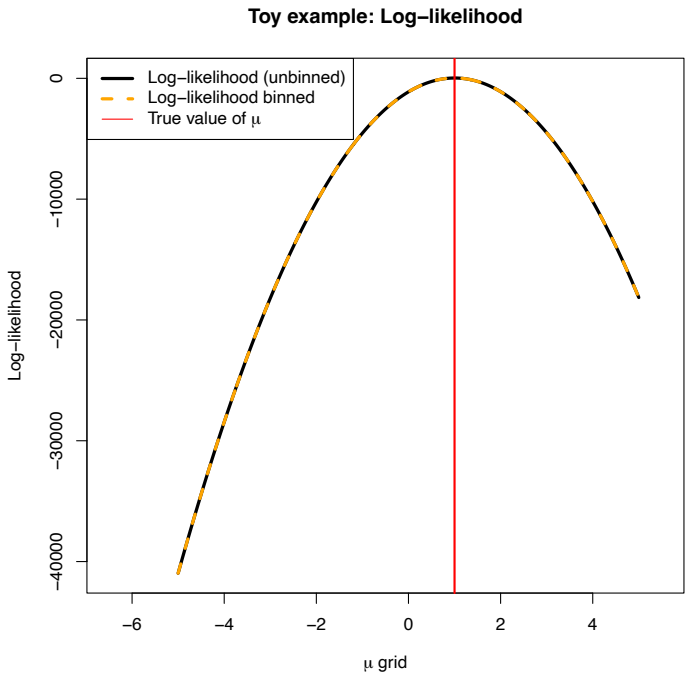


Figure 4.5: Comparison of the log-likelihood (up to a constant) in our toy simulation example, computed via the binned model (4.19) (in orange) vs. the unbinned model (4.14) (in black). The figure illustrates that the binned log-likelihood ℓ_b well approximates the unbinned log-likelihood ℓ_u, both leading to the maximum log-likelihood at the true value $\mu = 1$ (shown in red).

4.3.2 Methodology – Correcting for X-ray and Optical Incompleteness via Poisson Binning

We now consider the realistic case of an incomplete sample as described in Section 4.1.2. More formally, we assume that $N_{\text{obs}} < N_{\text{total}}$, with N_{obs} being the number of observed sources and N_{total} being the total number of sources in the population, with $N_{\text{total}} \sim \text{Pois}(\psi)$, and ψ being the expected population

size. Let λ_i be the source intensity for source i, with $i = 1, \ldots, N_{\text{total}}$, and $\lambda_i \overset{\text{iid}}{\sim} \text{Pareto}(\boldsymbol{\theta})$, with Pareto density $f_P(\lambda_i; \boldsymbol{\theta})$ (recall that f_P is a placeholder for the single and double Pareto population density). In this section, we assume that the source intensities for *observed* sources can be measured with negligible error (we will withdraw this assumption in Section 4.4).

As described in Section 4.1.2, we assume that our sample contains sources detected in an X-ray and/or optical survey, with both surveys being incomplete. More precisely, the set of detected (observed) sources from both surveys is overlapping, though both surveys contain sources non-detected in the other, and we have to assume that there are sources non-detected in either of the surveys (as sketched in Figure 4.3).

To describe these cases, we define an indicator variable Z_i, with

$$Z_i = \begin{cases} 2 & \text{if source } i \text{ is observed in X-ray} \qquad (\text{w/p } g(\lambda_i)) \\ 1 & i \text{ is observed in optical data only} \qquad (\text{w/p } h(\lambda_i|\phi)) \\ 0 & i \text{ is not observed} \qquad (\text{w/p } 1 - g(\lambda_i) - h(\lambda_i|\phi)), \end{cases} \tag{4.21}$$

with $g(\lambda_i) = P(Z_i = 2|\lambda_i)$ describing the probability of a source i being detected in the X-ray survey[7] given its source intensity λ_i. We assume $g(\cdot)$ to be a known function. Further, $h(\lambda_i|\phi) = P(Z_i = 1|\lambda_i, \phi)$ describes the probability of a source i with source intensity λ_i being detected in the optical catalogue only, with ϕ being a free model parameter. A potential form of the X-ray incompleteness function $g(\cdot)$ is sketched in Figure 4.4 (red curve), with the sum of X-ray and optical incompleteness functions, $g(\cdot) + h(\cdot|\phi)$, sketched in blue.

Of course, cases $Z_i \in \{1, 2\}$ can be observed, but not the case $Z_i = 0$. For the

[7]Note that with $Z = 2$ we generally consider sources that are detected in X-ray and optical, as well as sources that are detected in X-ray only. Assuming that optical matches (sources detected in X-ray but not in optical) are missing at random, one can also discard the X-ray sources without optical match, and only consider X-ray detected sources with optical match as case $Z = 2$ (e.g., as conducted and further detailed in Section 4.7).

observed cases, we formulate the joint distributions of λ and Z via

$$p(\lambda, Z = 2|\boldsymbol{\theta}) = f_P(\lambda;\boldsymbol{\theta})g(\lambda) \tag{4.22}$$

$$p(\lambda, Z = 1|\boldsymbol{\theta}, \phi) = f_P(\lambda;\boldsymbol{\theta})h(\lambda|\phi) \tag{4.23}$$

Poisson binning with incompleteness functions: We now bin the data based on λ. More precisely, let $k = 1, \dots K$ be bins on λ, with bin position $\lambda^{(k)}$ and bin size $\Delta_{\lambda^{(k)}}$. Then, within bin k we have

$X_1^{(k)} = \#$ of objects in bin k with $Z = 1$, with

$$X_1^{(k)} \sim \text{Pois}(f_P(\lambda^{(k)};\boldsymbol{\theta})\Delta_{\lambda^{(k)}}h(\lambda^{(k)}|\phi)\psi) \tag{4.24}$$

$X_2^{(k)} = \#$ of objects in bin k with $Z = 2$, with

$$X_2^{(k)} \sim \text{Pois}(f_P(\lambda^{(k)};\boldsymbol{\theta})\Delta_{\lambda^{(k)}}g(\lambda^{(k)})\psi). \tag{4.25}$$

That is, the number of objects observed in bin k and detected in X-ray (4.25) can be modelled via a Poisson distribution, with Poisson mean (rate parameter) being the product of the population density f_P, evaluated at the bin location $\lambda^{(k)}$, the bin width $\Delta_{\lambda^{(k)}}$, the population size ϕ, and the incompleteness function $g(\cdot)$ evaluated at the bin location $\lambda^{(k)}$. (Analogously for the optical case in (4.24), replacing $g(\cdot)$ by $h(\cdot|\phi)$). We note that (4.24) and (4.25) requires the incompleteness functions $g(\cdot)$ and $h(\cdot|\phi)$ to be non-zero for the entire support of λ, to ensure that the Poisson rate parameter is well-defined in $(0, \infty)$.

To express the likelihood function, we assume $X_1^{(k)}, X_2^{(k)} \perp\!\!\!\perp X_1^{(l)}, X_2^{(l)}|\boldsymbol{\theta}, \phi, \psi$, for all $k \neq l$, with $k, l = 1, \dots, K$. That is, the number of observed sources in a bin k (X-ray or optical only) is conditionally independent of the number of observed sources within another bin l given the parameters $(\boldsymbol{\theta}, \phi, \psi)$. We further assume that $X_1^{(k)} \perp\!\!\!\perp X_2^{(k)}|\lambda^{(k)}$, that is, given the bin location $\lambda^{(k)}$ the number of sources in detected *only* in the optical catalog (within bin k) does not depend on the number of X-ray detected sources in this bin. We further note that $N_{\text{X-ray}} = \sum_k X_2^{(k)}$ is the number of observed sources in the

X-ray catalogue, and $N_{\text{opt}} = \sum_k X_1^{(k)}$ is the number of sources in the optical catalogue only[8], with $N_{\text{X-ray}} + N_{\text{opt}} = N_{\text{obs}} < N_{\text{total}}$.

Given the binned data $X_1^{(k)}$ and $X_2^{(k)}$, for $k = 1, \ldots, K$, we can express our likelihood function $L(\boldsymbol{\theta}, \phi, \psi \big| \{X_1^{(k)}\}_{k=1}^K, \{X_2^{(k)}\}_{k=1}^K) = p(\{X_1^{(k)}\}_{k=1}^K, \{X_2^{(k)}\}_{k=1}^K | \boldsymbol{\theta}, \phi, \psi)$ via

$$
\begin{aligned}
&L(\boldsymbol{\theta}, \phi, \psi \big| \{X_1^{(k)}\}_{k=1}^K, \{X_2^{(k)}\}_{k=1}^K) \\
&\quad = \prod_k \Bigg[\frac{e^{-f_P(\lambda^{(k)};\boldsymbol{\theta})\Delta_{\lambda^{(k)}} h(\lambda^{(k)}|\phi)\psi} \left[f_P(\lambda^{(k)};\boldsymbol{\theta})\Delta_{\lambda^{(k)}} h(\lambda^{(k)}|\phi)\psi \right]^{X_1^{(k)}}}{X_1^{(k)}!} \\
&\qquad \times \frac{e^{-f_P(\lambda^{(k)};\boldsymbol{\theta})\Delta_{\lambda^{(k)}} g(\lambda^{(k)})\psi} \left[f_P(\lambda^{(k)};\boldsymbol{\theta})\Delta_{\lambda^{(k)}} g(\lambda^{(k)})\psi \right]^{X_2^{(k)}}}{X_2^{(k)}!} \Bigg],
\end{aligned}
\tag{4.26}
$$

which (dropping the normalizing constant $\frac{1}{\prod_k X_1^{(k)} X_2^{(k)}}$) can be simplified to[9]

$$
\begin{aligned}
&L(\boldsymbol{\theta}, \phi, \psi \big| \{X_1^{(k)}\}_{k=1}^K, \{X_2^{(k)}\}_{k=1}^K) \\
&\qquad \propto e^{-\psi \sum f_P(\lambda^{(k)};\boldsymbol{\theta})\Delta_{\lambda^{(k)}} h(\lambda^{(k)}|\phi)} \left[\prod_k \left[f_P(\lambda^{(k)};\boldsymbol{\theta})\Delta_{\lambda^{(k)}} h(\lambda^{(k)}|\phi) \right]^{X_1^{(k)}} \right] \psi^{N_{\text{opt}}} \\
&\qquad \times e^{-\psi \sum f_P(\lambda^{(k)};\boldsymbol{\theta})\Delta_{\lambda^{(k)}} g(\lambda^{(k)})} \left[\prod_k \left[f_P(\lambda^{(k)};\boldsymbol{\theta})\Delta_{\lambda^{(k)}} g(\lambda^{(k)}) \right]^{X_2^{(k)}} \right] \psi^{N_{\text{X-ray}}}.
\end{aligned}
\tag{4.27}
$$

[8]Note that in Section 4.1.2 we denote with N_{optical} the number of sources that are detected in the optical catalogue (which includes optical sources with an X-ray match), here with N_{opt} we refer to the number of optical sources without X-ray match (sources detected only on the optical catalogue).

[9]Note that with operator $(\times)$ we express ordinary multiplication, that is $a \times b = ab$.

We can then express the log-likelihood via

$$
\begin{aligned}
&\ell(\boldsymbol{\theta}, \phi, \psi \big| \{X_1^{(k)}\}_{k=1}^K, \{X_2^{(k)}\}_{k=1}^K) \\
&\quad = -\psi \sum_k f_P(\lambda^{(k)}; \boldsymbol{\theta}) \Delta_{\lambda^{(k)}} h(\lambda^{(k)} | \phi) \\
&\qquad + \sum_{\{k: X_1^{(k)} > 0\}} X_1^{(k)} \log(f_P(\lambda^{(k)}; \boldsymbol{\theta})) + \sum_{\{k: X_1^{(k)} > 01\}} X_1^{(k)} \log(h(\lambda^{(k)} | \phi)) \\
&\qquad + \sum_{\{k: X_1^{(k)} > 0\}} X_1^{(k)} \log(\Delta_{\lambda^{(k)}}) + N_{\text{opt}} \log(\psi) \\
&\qquad - \psi \sum_k f_P(\lambda^{(k)}; \boldsymbol{\theta}) \Delta_{\lambda^{(k)}} g(\lambda^{(k)}) \\
&\qquad + \sum_{\{k: X_2^{(k)} > 0\}} X_2^{(k)} \log(f_P(\lambda^{(k)}; \boldsymbol{\theta})) + \sum_{\{k: X_2^{(k)} > 0\}} X_2^{(k)} \log(g(\lambda^{(k)})) \\
&\qquad + \sum_{\{k: X_2^{(k)} > 0\}} X_2^{(k)} \log(\Delta_{\lambda^{(k)}}) + N_{\text{X-ray}} \log(\psi). \qquad (4.28)
\end{aligned}
$$

By removing constant terms, this can be simplified to

$$
\begin{aligned}
&\ell(\boldsymbol{\theta}, \phi, \psi \big| \{X_1^{(k)}\}_{k=1}^K, \{X_2^{(k)}\}_{k=1}^K) \\
&\quad = -\psi \sum_k f_P(\lambda^{(k)}; \boldsymbol{\theta}) \Delta_{\lambda^{(k)}} \left[h(\lambda^{(k)} | \phi) + g(\lambda^{(k)}) \right] \\
&\qquad + \sum_{\{k: X_1^{(k)} > 0\}} X_1^{(k)} \log(f_P(\lambda^{(k)}; \boldsymbol{\theta})) + \sum_{\{k: X_2^{(k)} > 0\}} X_2^{(k)} \log(f_P(\lambda^{(k)}; \boldsymbol{\theta})) \\
&\qquad + \sum_{\{k: X_1^{(k)} > 0\}} X_1^{(k)} \log(h(\lambda^{(k)} | \phi)) + \sum_{\{k: X_2^{(k)} > 0\}} X_2^{(k)} \log(g(\lambda^{(k)})) \\
&\qquad + N_{\text{obs}} \log(\psi) \qquad (4.29)
\end{aligned}
$$

We note that with $\Delta_{\lambda^{(k)}} \to 0$, it follows that $X_1^{(k)}, X_2^{(k)} \to \{0, 1\}$, and with sufficiently small bins, it follows that $\lambda^{(k)} \to \lambda_i$ (in cases with $X_1^{(k)} = 1$ or $X_2^{(k)} = 1$). That is, the location of the bins (with observed sources) approximates the location of observed source intensities λ_i. That is, with a sufficiently fine bin grid, it holds that the binned likelihood approximates the likelihood without binning (as illustrated in the previous section via a toy example), that is $L(\boldsymbol{\theta}, \phi, \psi \big| \{X_1^{(k)}\}_{k=1}^K, \{X_2^{(k)}\}_{k=1}^K) \approx L(\boldsymbol{\theta}, \phi, \psi \big| \boldsymbol{\lambda}, \boldsymbol{Z}_{\text{obs}})$ (or

equivalently $p(\{X_1^{(k)}\}_{k=1}^K, \{X_2^{(k)}\}_{k=1}^K | \boldsymbol{\theta}, \phi, \psi) \approx p(\boldsymbol{\lambda}, \boldsymbol{Z}_{\rm obs} | \boldsymbol{\theta}, \phi, \psi))$, with $\boldsymbol{Z}_{\rm obs} = (Z_1, \ldots, Z_{N_{\rm obs}})$ being a vector, with $Z_i \in \{1, 2\}$, $i = 1, \ldots, N_{\rm obs}$ describing the survey case of observed source i.

Assuming the source intensities of observed sources to be known, we now have a model to account for the incompleteness of the observed sample (in X-ray and optical). As shown in Section 4.2, the source intensities of observed sources are latent variables and need to be estimated form the observed source counts. In Section 4.4, we thus carry out our final model step by combining the source model from Section 4.2 with the incompleteness corrections developed in this section. In an intermediate step, in Section 4.3.3 we drop some simplifying assumptions made so far regarding the X-ray incompleteness function, and describe the incorporation of a realistic X-ray incompleteness function into our model.

4.3.3 Realistic X-ray Incompleteness Function via Marginalization over Nuisance Parameters

Realistic X-ray incompleteness function depending on nuisance parameters: In the previous sections, we conveniently express the X-ray incompleteness function $g(\cdot)$ as a function of the source intensities λ (e.g., in (4.21) and (4.22)). Unfortunately, in reality this is not the case. More precisely, in our model we employ the realistic X-ray incompleteness function provided by Wright et al. (2015), which is based on extensive simulations, and is specific for Chandra X-ray survey data. Wright et al. (2015) expresses the X-ray incompleteness function being dependent on additional parameters. More precisely, the incompleteness function in Wright et al. (2015) depends on the background count per pixel ξ_T (in count/pixel), with $\xi_T = \mathcal{T}\xi$, and the (nuisance) parameters L and w_0, where L is the off-axis angle, which describes the distance (of a detected source location) to the centre of the field of view, and w_0 is a threshold that determines the sensitivity of the X-ray source detection method, wavdetect (Freeman et al., 2002).

In addition, Wright et al. (2015) expresses the incompleteness function in terms of source intensity C (in counts), instead of source intensity λ (in

count/s/cm^2), leading to following form

$$P(Z = 2|C, \xi_T, L, w_0) = 1 - \exp\left(\frac{-C^{\nu_1(\xi_T, L, w_0)}}{10^{\nu_2(\xi_T, L, w_0)}}\right) \tag{4.30}$$

where values for $\nu_1(\xi_T, L, w_0)$ and $\nu_2(\xi_T, L, w_0)$ are provided by Wright et al. (2015) as tabular data for various parameter combinations (as result of an MCMC sampler).[10]

To obtain an incompleteness function that depends on λ, we now employ the relation $C_i = \lambda_i \mathcal{T} e_i$, with effective area e_i at the location of the source i, and the constant exposure time $\mathcal{T}$. Additionally, employing $\xi_T = \mathcal{T}\xi$, we can now express the incompleteness function in (4.30) via

$$P(Z = 2|\lambda, \xi, L, w_0, e, \mathcal{T}) = 1 - \exp\left(\frac{-(\lambda \mathcal{T} e)^{\nu_1(\mathcal{T}\xi, L, w_0)}}{10^{\nu_2(\mathcal{T}\xi, L, w_0)}}\right) \tag{4.31}$$

Marginalization over nuisance parameters: We now show that marginalization over the nuisance parameters allows us to express the realistic incompleteness function provided by Wright et al. (2015) solely in terms of the source intensities λ, as required for our model described in Section 4.3.2. First, when detecting the X-ray source locations (via wavdetect), we fix the source detection threshold to $w_0 = 10^{-6}$, and we keep w_0 as a fixed parameter throughout our analysis. In addition, the total exposure time $\mathcal{T}$ is a fixed constant in our analysis. We can thus fix these two parameters in (4.31), and we express the incompleteness function via the remaining nuisance parameters L, e and ξ. We can now formulate the joint distribution of the X-ray source intensities λ and remaining nuisance parameters (L, e, ξ) via

$$p(\lambda, L, e, \xi, Z = 2|\boldsymbol{\theta}) = p(\lambda, L, e, \xi|\boldsymbol{\theta}) P(Z = 2|\lambda, L, e, \xi), \tag{4.32}$$

with $Z = 2$ indicating observation in X-ray, and conditioning everything on the Pareto parameters $\boldsymbol{\theta}$, with the incompleteness function being dependent

[10] We note that Wright et al. (2015) uses the notation $\lambda_1(\cdot)$ and $\lambda_2(\cdot)$, which we replaced by $\nu_1(\cdot)$ and $\nu_2(\cdot)$ to avoid confusion and notation overlap with our source intensities λ.

on $\boldsymbol{\theta}$ only through the given λ. We can now simplify (4.32) by integrating out the nuisance parameters, leading to

$$\int p(\lambda, L, e, \xi, Z = 2|\boldsymbol{\theta}) d(L, e, \xi) = \int p(\lambda, L, e, \xi|\boldsymbol{\theta}) P(Z = 2|\lambda, L, e, \xi) d(L, e, \xi) \tag{4.33}$$

By assuming λ, L, ξ and e all being independent, and assuming L, ξ and e being independent of the source population parameters $\boldsymbol{\theta}$, we obtain

$$p(\lambda, Z = 2|\boldsymbol{\theta}) = f_P(\lambda|\boldsymbol{\theta}) \int p(L)p(e)p(\xi)P(Z = 2|\lambda, L, e, \xi) dL, de, d\xi, \tag{4.34}$$

where $f_P(\lambda; \boldsymbol{\theta}) = p(\lambda|\boldsymbol{\theta})$ is the Pareto source intensity population density. The left hand side (LHS) of (4.34) follows from the LHS of (4.33) by integrating out (L, e, ξ). With the LHS of (4.22) and (4.34) being the same, we can now simply exchange the function $g(\lambda)$ in (4.22) with the function

$$g_{\text{int}}(\lambda) = \int p(L)p(e)p(\xi)P(Z = 2|\lambda, L, e, \xi) dL, de, d\xi \tag{4.35}$$

to account for the contribution of additional nuisance parameters in the X-ray incompleteness function. We detail the computation of g_{int} with values specific to our CDFS data analysis in Section 4.7.2.

4.4 Combination of Source Model and Incompleteness Corrections

In this final model step, we combine the source model presented in Section 4.2 with the incompleteness corrections developed in Section 4.3. More precisely, we consider the realistic case that source intensities λ_i are not directly observed, but can be estimated via the observed counts Y_i in source regions i and the observed background counts X (as described in Section 4.2). We further assume survey incompleteness, with the observed sources being detected either in X-ray and/or in an optical catalogue (as described in Section 4.3. The aim of this section is to develop a model that allows joint estimation of the source intensities λ_i and the source population distribution by rigorously accounting for (X-ray and optical) survey incompleteness.

Using the same notation as in Section 4.2 and 4.3, we are interested in the posterior distribution $p(\xi, \boldsymbol{\lambda}, \boldsymbol{\theta}, \phi, \psi | \boldsymbol{D}, \boldsymbol{Z}_{\rm obs})$, where $\mathbf{D} = \{a_i, Y_i, e_i, X, A, \mathcal{T}\}_{i=1}^{N_{\rm obs}}$, as described in Section 4.2. $\boldsymbol{Z}_{\rm obs} = (Z_1, \ldots, Z_{N_{\rm obs}})$ is a vector, with $Z_i \in \{1, 2\}$, describing the survey case of observed source i, and $\boldsymbol{\lambda} = (\lambda_1, \ldots \lambda_{N_{\rm obs}})$ is a vector with the latent source intensity λ_i for all observed sources, $i = 1, \ldots, N_{\rm obs}$. The posterior distribution $p(\xi, \boldsymbol{\lambda}, \boldsymbol{\theta}, \phi, \psi | \boldsymbol{D}, \boldsymbol{Z}_{\rm obs})$ can then be expressed via

$$
\begin{aligned}
p(\xi, \boldsymbol{\lambda}, \boldsymbol{\theta}, \phi, \psi | \boldsymbol{D}, \boldsymbol{Z}_{\rm obs}) =& p(\boldsymbol{D} | \xi, \boldsymbol{\lambda}, \boldsymbol{\theta}, \phi, \psi, \boldsymbol{Z}_{\rm obs}) p(\xi, \boldsymbol{\lambda}, \boldsymbol{\theta}, \phi, \psi | \boldsymbol{Z}_{\rm obs}) && (4.36)\\
=& p(\boldsymbol{D} | \xi, \boldsymbol{\lambda}, \boldsymbol{\theta}, \phi, \psi, \boldsymbol{Z}_{\rm obs}) \frac{p(\boldsymbol{\lambda}, \boldsymbol{\theta}, \phi, \psi, \boldsymbol{Z}_{\rm obs})}{p(\boldsymbol{Z}_{\rm obs})} p(\xi) && (4.37)\\
\propto& p(\boldsymbol{D} | \xi, \boldsymbol{\lambda}, \boldsymbol{\theta}, \phi, \psi, \boldsymbol{Z}_{\rm obs}) p(\boldsymbol{\lambda}, \boldsymbol{Z}_{\rm obs} | \boldsymbol{\theta}, \phi, \psi) p(\boldsymbol{\theta}, \phi, \psi) p(\xi) && (4.38)\\
=& p(\boldsymbol{D} | \xi, \boldsymbol{\lambda}) p(\boldsymbol{\lambda}, \boldsymbol{Z}_{\rm obs} | \boldsymbol{\theta}, \phi, \psi) p(\boldsymbol{\theta}) p(\phi) p(\psi) p(\xi) && (4.39)\\
\approx& p(\boldsymbol{D} | \xi, \boldsymbol{\lambda}) p(\{X_1^{(k)}\}_{k=1}^K, \{X_2^{(k)}\}_{k=1}^K | \boldsymbol{\theta}, \phi, \psi) \\
& \times p(\boldsymbol{\theta}) p(\phi) p(\psi) p(\xi), && (4.40)
\end{aligned}
$$

where the second line holds since ξ is a priori independent of the other pa-

rameters and of $\boldsymbol{Z}_{\text{obs}}$. In the third line we drop the constant $p(\boldsymbol{Z}_{\text{obs}})$. From (4.38) to (4.39), we assume the population size ψ, the parameters $\boldsymbol{\theta}$, describing the Pareto population distribution, and the parameter ϕ, describing the optical incompleteness function, to be a priori independent. We further assume that given the source intensities $\boldsymbol{\lambda} = (\lambda_1, \ldots \lambda_{N_{\text{obs}}})$, the distribution of the observed data is independent of the source population parameters $\boldsymbol{\theta}$, the optical incompleteness function parameter ϕ, the the population size ψ, and the observed cases $\boldsymbol{Z}_{\text{obs}}$. The final step, (4.39) to (4.40), follows as described in Section 4.3.

Finally, the formulation of the posterior distribution in (4.40) allows straightforward combination of the source model and the incompleteness corrections, with $p(\boldsymbol{D}|\xi, \boldsymbol{\lambda})$ being formalized in Section 4.2 via (4.6), and $p(\{X_1^{(k)}\}_{k=1}^K, \{X_2^{(k)}\}_{k=1}^K|\boldsymbol{\theta}, \phi, \psi)$ being given in Section 4.3 via (4.27), by treating $X_1^{(k)}$ and $X_2^{(k)}$ as latent parameters (both dependent on the latent $\boldsymbol{\lambda}$) instead of observed data. With prior distributions for $\boldsymbol{\theta}$ and for ξ specified in Section 4.2, to obtain a proper posterior distribution, (4.40) eventually requires the specification of proper prior distributions on the population size ψ and on the optical incompleteness function parameter ϕ. We choose Gamma priors on both parameters, i.e., $\psi \sim \text{Gamma}(a_\psi, b_\psi)$, and $\phi \sim \text{Gamma}(a_\phi, b_\phi)$. Table C.1 in the Appendix provides a summary of the various prior choices and prior parameters used in Sections 4.6 and 4.7.

Joint posterior distribution for the double Pareto population model: Specifying the source population distribution being the double Pareto distribution with density f_{DP} as specified in Section 4.2, the joint posterior distribution of $(\xi, \boldsymbol{\lambda}, \theta_1, \theta_2, \tau_1, \tau_2, \phi, \psi)$ given the data $\mathbf{D}$ =

$\{a_i, Y_i, e_i, X, A, \mathcal{T}\}_{i=1}^{N_{\text{obs}}}$ and $\boldsymbol{Z}_{\text{obs}}$, is then obtained via

$$
\begin{aligned}
&p(\xi,\boldsymbol{\lambda},\theta_1,\theta_2,\tau_1,\tau_2,\phi,\psi|\boldsymbol{D},\boldsymbol{Z}_{\text{obs}}) \\
&\quad\propto p(\boldsymbol{D}|\xi,\boldsymbol{\lambda})p(\{X_1^{(k)}\}_{k=1}^K,\{X_2^{(k)}\}_{k=1}^K|\theta_1,\theta_2,\tau_1,\tau_2,\phi,\psi) \\
&\qquad\times p(\theta_1)p(\theta_2)p(\tau_1)p(\tau_2|\tau_1)p(\phi)p(\psi)p(\xi) \qquad (4.41)\\
&\quad\propto \exp(-A\mathcal{T}\xi)\frac{(A\mathcal{T}\xi)^X}{X!}\prod_{i=1}^n \exp[-(a_i\xi+e_i\lambda_i)\mathcal{T}]\frac{[(a_i\xi+e_i\lambda_i)\mathcal{T}]^{Y_i}}{Y_i!} \\
&\qquad\times e^{-\psi\sum f_{\text{DP}}(\lambda^{(k)};\theta_1,\theta_2,\tau_1,\tau_2)\Delta_{\lambda^{(k)}}h(\lambda^{(k)}|\phi)} \\
&\qquad\times\left[\prod_k\left[f_{\text{DP}}(\lambda^{(k)};\theta_1,\theta_2,\tau_1,\tau_2)\Delta_{\lambda^{(k)}}h(\lambda^{(k)}|\phi)\right]^{X_1^{(k)}}\right]\psi^{N_{\text{opt}}} \\
&\qquad\times e^{-\psi\sum f_{\text{DP}}(\lambda^{(k)};\theta_1,\theta_2,\tau_1,\tau_2)\Delta_{\lambda^{(k)}}g_{\text{int}}(\lambda^{(k)})} \\
&\qquad\times\left[\prod_k\left[f_{\text{DP}}(\lambda^{(k)};\theta_1,\theta_2,\tau_1,\tau_2)\Delta_{\lambda^{(k)}}g_{\text{int}}(\lambda^{(k)})\right]^{X_2^{(k)}}\right]\psi^{N_{\text{X-ray}}} \\
&\qquad\times\theta_1^{(a_1-1)}\exp(-b_1\theta_1)\times\theta_2^{(a_2-1)}\exp(-b_2\theta_2) \\
&\qquad\times\tau_1^{(a_\tau-1)}\exp(-b_\tau\tau_1)\times 1_{\{\tau_1\le\tau_2\}}\frac{1}{\sigma_\tau}\exp\left(-\frac{1}{2}\left(\frac{(\tau_2-\mu_\tau)}{\sigma_\tau}\right)\right) \\
&\qquad\times\phi^{(a_\phi-1)}\exp(-b_\phi\phi)\times\psi^{(a_\psi-1)}\exp(-b_\psi\psi)\times\xi^{(a_\xi-1)}\exp(-b_\xi\xi) \qquad (4.42)
\end{aligned}
$$

For ease of notation, we refer via f_{DP} to the double-Pareto density as defined in (4.9), and via $g_{\text{int}}(\cdot)$ to the marginalized X-ray incompleteness function as given by (4.35).

The full posterior distribution for a (single) Pareto source population distribution follows analogously, by replacing f_{DP} via f_{SP} (as defined in (4.8)) and by removing the prior distributions $p(\theta_2)$ and $p(\tau_2|\tau_1)$. Below we develop a Gibbs-type sampler to obtain posterior draws from (4.42).

4.5 Sampling from the Posterior Distribution – Metropolis-within-Gibbs Sampler

In this section, we develop a Metropolis-within-Gibbs sampler to obtain draws from the posterior distribution detailed in the previous section. Precisely, we want to sample from the joint posterior distribution

$$\begin{aligned} p(\xi, \boldsymbol{\lambda}, \theta_1, \theta_2, \tau_1, \tau_2, \phi, \psi | \boldsymbol{D}, \boldsymbol{Z}_{\text{obs}}) \\ \propto & p(\boldsymbol{D}|\xi, \boldsymbol{\lambda}) p(\{X_1^{(k)}\}_{k=1}^K, \{X_2^{(k)}\}_{k=1}^K | \theta_1, \theta_2, \tau_1, \tau_2, \phi, \psi) \\ & \times p(\theta_1) p(\theta_2) p(\tau_1) p(\tau_2|\tau_1) p(\phi) p(\psi) p(\xi), \end{aligned} \tag{4.43}$$

which is formulated in detail in (4.42). Our Metropolis-within-Gibbs sampling scheme can be summarized via the following.

Metropolis-within-Gibbs-sampler: For number of iterations $t = 1, \ldots, T$:

- Sample $\boldsymbol{\lambda}$ component-wise for the observed sources $i = 1, \ldots, N_{\text{obs}}$ via:

$$\begin{aligned} p(\lambda_i|\cdot) \propto f_{\text{Pois}}(Y_i; (a_i\xi + e_i\lambda_i)\mathcal{T}) f_{\text{DP}}(\lambda_i; \theta_1, \theta_2, \tau 1, \tau_2) \\ [g(\lambda_i) 1_{\{Z_i=2\}} + h(\lambda_i|\phi) 1_{\{Z_i=1\}}] \end{aligned} \tag{4.44}$$

  where $f_{\text{Pois}}(x; \nu_{\text{Pois}})$ denotes the probability mass function of a Poisson random variable with rate parameter ν_{Pois} evaluated at position x.

- Sample ξ as:

$$p(\xi) \propto f_{\text{Gamma}}(\xi; a_\xi, b_\xi) f_{\text{Pois}}(X; A\xi\mathcal{T}) \prod_{i=1}^{N_{\text{obs}}} f_{\text{Pois}}(Y_i; (a_i\xi + e_i\lambda_i)\mathcal{T}) \tag{4.45}$$

  where $f_{\text{Gamma}}(x; \alpha, \beta)$ denotes the pdf of a Gamma distributed r.v. with shape parameter α and rate parameter β evaluated at position x.

- Sample ψ as:

$$p(\psi) \propto f_{\text{Gamma}}(\psi; a_\psi, b_\psi)\psi^{N_{\text{obs}}} \times \exp\left(-\psi \sum_k f_{\text{DP}}(\lambda^{(k)}; \theta_1, \theta_2, \tau_1, \tau_2)\Delta_{\lambda^{(k)}}[g(\lambda_i) + h(\lambda^{(k)}|\phi)]\right) \tag{4.46}$$

- Sample θ_1 as:

$$p(\theta_1) \propto f_{\text{Gamma}}(\theta_1; a_1, b_1) \times \exp\left(-\psi \sum f_{\text{DP}}(\lambda^{(k)}; \theta_1, \theta_2, \tau_1, \tau_2)\Delta_{\lambda^{(k)}}[g(\lambda_i) + h(\lambda^{(k)}|\phi)]\right) \times \prod_k \left[f_{\text{DP}}(\lambda^{(k)}; \theta_1, \theta_2, \tau_1, \tau_2)\right]^{(X_1^{(k)}+X_2^{(k)})} \tag{4.47}$$

- Sample θ_2, τ_1, and τ_2 following (4.47), but replacing the prior distribution in the first line of (4.47), $f_{\text{Gamma}}(\theta_1; a_1, b_1))$, by the prior distribution of the respective parameter. That is, $f_{\text{Gamma}}(\theta_2; a_2, b_2)$ for θ_2; $f_{\text{Gamma}}(\tau_1; a_\tau, b_\tau)$ for τ_1; and $f_{\text{truncNorm}}(\tau 2; \mu_\tau, \sigma_\tau^2, a_0 = \tau_1, b_0 = \infty)$ for τ_2.

- Sample ϕ as:

$$p(\phi|\cdot) \propto f_{\text{Gamma}}(\phi; a_\phi, b_\phi) \times \exp\left(-\psi \sum_k f_{\text{DP}}(\lambda^{(k)}; \theta_1, \theta_2, \tau_1, \tau_2)\Delta_{\lambda^{(k)}} h(\lambda^{(k)}|\phi)\right) \prod_k h(\lambda^{(k)|\phi)})^{X_1^{(k)}} \tag{4.48}$$

We note that the Metropolis-within-Gibbs sampler for the model with (single) Pareto population distribution follows analogously, by just removing the sampling steps for θ_2 and for τ_2, and by replacing $f_{\text{DP}}(\cdot; \theta_1, \theta_2, \tau_1, \tau_2)$ by $f_{\text{SP}}(\cdot; \theta_1, \tau_1)$.

4.5.1 Technical Details on Gibbs-Type Sampler

Jumping distributions and log-transformations: To avoid sampling from overly skewed distributions, we apply log-transformations to the parameters τ_2 and λ_i, $i = 1 \ldots, N_{\text{obs}}$, with Jacobian adjustments following Section C.1. The remaining parameters are sampled in ordinary scale.

We choose symmetric Gaussian jumping distributions for the parameters $\xi, \psi, \theta_1, \theta_2$, and τ_2, being centered at the respective values of the previous iterations, applying ordinary Metropolis updates. For τ_1 and for λ_i, in each iteration $t = 1, \ldots, T$ of the sampler one has to ensure that $\lambda_i^{(t)} > \tau_1^{(t)}$, for all $i = 1, \ldots, N_{\text{obs}}$. We thus choose a truncated normal jumping distributions for $\lambda_i^{(t)}$ with mode at $\lambda_i^{(t-1)}$ and with lower truncation at $\tau_1^{(t-1)}$. Then, τ_1 is updated via a truncated-Normal with mode at $\tau_1^{(t-1)}$ and with upper truncation at $\min(\lambda_i^{(t)})$ with $i = 1, \ldots, N_{\text{obs}}$, and with lower truncation at 0. Finally, ϕ is updated via a truncated normal distribution with lower truncation at 0. Given a non-symmetric (truncated-Normal) jumping distribution, we apply Metropolis-Hastings updates following Section 1.3.2, to correct for bias resulting from the non-symmetry of the jumping distribution.

Starting values: To obtain posterior samples from (4.42) as described in Section 4.5, we need to define starting values for the MCMC chains for each parameter. For λ_i, with $i \in 1, \ldots, N_{\text{obs}}$, we obtain starting values via

$$\hat{\lambda}_i^{\text{start}} = \frac{Y_i - X_i}{e_i \mathcal{T}}.$$

If $\hat{\lambda}_i^{\text{start}} < 0$, $i = 1, \ldots, N_{\text{obs}}$, we sample $\hat{\lambda}_i^{\text{start}} \sim$ Log-uniform(min($|\lambda_j|$),-max($|\lambda_j|$)), with $j = 1, \ldots, N_{\text{obs}}$, and $j \neq i$. We obtain a starting value for ξ via

$$\hat{\xi}^{\text{start}} = \frac{X}{A\mathcal{T}} \tag{4.49}$$

For all other parameters, namely $\theta_1, \theta_2, \tau_1\tau_2, \psi$ and ϕ, we obtain starting values by sampling from the respective prior distribution. Being the population minimum, for τ_1, the starting value has to be smaller than $\min(\hat{\lambda}_i^{\text{start}})$. We thus sample from the τ_1 prior until this condition is met, to accept the sampled value as starting value for τ_1.

4.6 Simulation Study

In this section, we evaluate our proposed model on simulated data. In Section 4.6.1, we detail our data generation process, aiming to mimic characteristics of the CDFS survey. In Section 4.6.2, we evaluate our model on the simulated data. We note that here we simulate and evaluate our framework under the (slightly more complicated) double-Pareto population case.

4.6.1 Simulation Study – Data Generation

To resemble realistic CDFS X-ray data, we design our simulation study based on previously collected CDFS data employed by Stampoulis (2017), and we incorporate results by Stampoulis (2017) to set our data generating parameters.[11] Table 4.1 provides a summary of the true parameter values used in our simulation. For the two power law slopes, we use the mean posterior values reported by Stampoulis (2017) for their double-Pareto model, that is $\theta_1^s = 0.759$ and $\theta_2^s = 1.52$. We use the superscript 's' to indicate that these are values used for our simulations. For the two Power-law knots we use values $\tau_1^s = 0.91\text{e}{-7}$ and $\tau_2^s = 5.07\text{e}{-2}$. We note that these values are an order of 10^{10} and 10^{13} higher than the values reported in Stampoulis (2017), which is due to the different scales of flux (as used in Stampoulis (2017)) and source intensities (used in our study).

[11] We note that we base our simulations on the previously collected CDFS data used by Stampoulis (2017), since we performed the simulation study before collecting the data employed in our data analysis described in Section 4.7. We find our simulation study to be sufficiently related to the data used in Section 4.7. Conducting additional simulation studies that precisely mimic the data described Section 4.7 is consideration of future work.

Table 4.1: True parameter values employed to generate our simulated data, with short parameter descriptions.

	Simulation value	Description
ψ	1000	Expected population size
θ_1	0.759	First Power law slope
θ_2	1.52	Second Power law slope
τ_1	0.91e−7	Population minimum
τ_2	5.07e−2	Breakpoint
ϕ	1.00e−5	Optical incompleteness parameter
ξ	2.01e−7	Background rate

We generate the data via the following procedure. First, we obtain the total population size via sampling $N_{\text{total}} \sim \text{Poisson}(\psi^s = 1000)$, obtaining a value of $N^s_{\text{total}} = 980$ in this particular example. We then generate our population of source intensities by sampling λ^s_i, with $i = 1, \ldots, N^s_{\text{total}} = 980$, source intensities iid from a double-Pareto distribution with parameters as specified in Table 4.1. Given the population of sources intensities, we now generate our incomplete, X-ray and optically, observed sample. To not make the simulations overly complicated, we define an X-ray incompleteness function g_{sim} in the same functional form as (4.31), but without dependence on nuisance parameters. More precisely, we define $g_{\text{sim}}(\cdot)$ via

$$g_{\text{sim}}(\lambda) = 1 - \exp\left(-\frac{\lambda}{10^{-5}}\right) \tag{4.50}$$

Given g_{sim} we then define the optical incompleteness function $h_{\text{sim}}(\cdot|\phi^s)$ via

$$h_{\text{sim}}(\lambda|\phi) = (1 - g_{\text{sim}}(\lambda)) \left(1 - \exp\left(-\frac{\lambda}{\phi^s}\right)\right) \tag{4.51}$$

(Figure 4.4 in Section 4.1.2 illustrates $g_{\text{sim}}(\cdot)$ (in red) and $(g_{\text{sim}}(\cdot) + h_{\text{sim}}(\cdot|\phi))$ (in blue).) Given a source intensities, λ^s_i with $i = 1, \ldots, N^s_{\text{total}}$, for each source in the population, we now sample X-ray survey assignment with probabilities $g_{\text{sim}}(\lambda_i)$, $i = 1, \ldots, N^s_{\text{total}}$. In our example, $N^s_{\text{X-ray}} = 90$ sources are sampled to be observed in X-ray. For the remaining sources in the population we sample optical survey assignment with probabilities $h_{\text{sim}}(\cdot|\phi^s)$. $N^s_{\text{opt}} = 34$ sources are assigned to the optical survey this way, leading to $N^s_{\text{obs}} = 124$ observed

sources. The remaining $N_{\text{missing}} = 856$ sources are considered unobserved, leading to a high proportion of incompleteness.

Given the source intensities of the observed sources, we can now generate the observed data. We set the exposure time $\mathcal{T}^s$ to the value reported by Stampoulis (2017), i.e., $\mathcal{T}^s = 1960631$ (in seconds). We obtain realistic source areas a_i^s and effective areas e_i^s, $i = 1, \ldots, N_{\text{obs}}^s$, by sampling with replacement from the reported values of source and effective areas in the CDFS data employed by Stampoulis (2017). We set the background rate via $\xi^s = X'/(A'\mathcal{T})$, where X' is the reported background count and A' the reported background area in Stampoulis (2017), leading to $\xi^s = 2.01\text{e}{-7}$. Given ξ^s we obtain the observed background count X^s via $X^s \sim \text{Poisson}(\xi^s A^s \mathcal{T}^s)$ with background area $A^s = \sum_{i=1}^{N_{\text{obs}}} a_i^s$ (we choose the background area to be the same size as the sum of source areas), leading to $X^s = 6098$. Finally, we simulate the observed counts Y_i^s at source region i via $Y_i^s \overset{\text{iid}}{\sim} \text{Poisson}((a_i^s \xi^s + e_i^s \lambda_i^s)\mathcal{T}^s)$, $i = 1, \ldots, N_{\text{obs}}^s$.

4.6.2 Simulation Study – Model Evaluation

Model details and sampler diagnostics: We now fit our full model as detailed in Section 4.4 via the Metropolis-within-Gibbs sampler introduced in Section 4.5. Table C.1 in the Appendix provides a summary of all prior distributions and the chosen prior hyperparameters.

We run $M_{\text{sim}} = 5$ chains with different starting values. We run each chain for $T = 80,000$ iterations, and we use a burn-in of $T_{\text{Burn}} = 20,000$. The remaining 60,000 draws of each chain are collected as posterior draws. Table 4.6.2 presents sampling diagnostics of the multiple chains via the Gelman-Rubin $\hat{R}$, and the effective sample size for each parameter. The Gelman-Rubin $\hat{R}$ is very close to 1 for all parameters, which indicates that no convergence issues are detected. Figure C.1 provides traceplots of the chains after burn-in (in log-scale) visually indicating convergence of the 5 chains to the same stationary distribution. Due to moderately high auto-correlation, the effective sample size for parameters ψ (877) and τ_1 (338) is markedly lower than for the remaining parameters. However, we consider the obtained effective sample

sizes being large enough for the purposes of our study.

Table 4.2: Posterior sampling diagnostics for simulation study parameters (Section 4.6.2).

	ψ^s	θ_1^s	θ_2^s	τ_1^s	τ_2^s	ϕ^s	ξ
Gelman-Rubin $\hat{R}$	1.02	1.01	1.00	1.05	1.00	1.00	1.00
Effective size	877	3359	13964	338	38573	26045	29991

Simulation results: Figure 4.6 illustrates the marginal distributions (in $\log_{10}$-scale) of the main parameters in the simulation study. We are particularly interested in the (double) pareto population distribution parameters, i.e., τ_1, τ_2, θ_1, and θ_2. We also consider the remaining parameters ψ, ϕ and ξ as parameters of interest (for reasons of space, we do not illustrate posterior distributions for the latent source intensities λ_i). We refer to all parameters except for the λ_i's as 'main parameters'. Figure 4.6 illustrates that the four population parameter marginal posterior distributions well recover the true simulation values (shown in red). The marginal posteriors for population size ψ and background rate ξ are well-centered at or close to the true values. The marginal posterior for the optical incompleteness function is slightly shifted from the true value. Given the good recovery of the remaining parameters, we do not consider this shift of much concern at this level.[12] Table 4.3 presents summary statistics of the main parameter marginal posteriors in original scale.

In Figure 4.7, we illustrate pairwise marginal posterior contour-and scatter-plots plotted in $\log_{10}$-scale. There is an apparent (negative) correlation in the posterior of τ_1 and ψ. This correlation is expected. If the estimate for the population minimum τ_1 goes far into the low-intensity region the estimate for the expected population size ψ is strongly increased via the incompleteness corrections in the low-intensity end. If τ_1 is large, the number of expected

[12]We note that the sample size of sources observed as 'optical only' is relatively small in our simulations $N^s_{\text{opt}} = 34$, and of much larger proportion in the CDFS analysis in Section 4.7. We constructed the simulations before having access to the X-ray and optical surveys considered in the CDFS data analysis section. Additional simulations with larger proportions of optical data are considerations of future work. Repeating the simulations for frequentist evaluations of the posterior estimates is also part of future work.

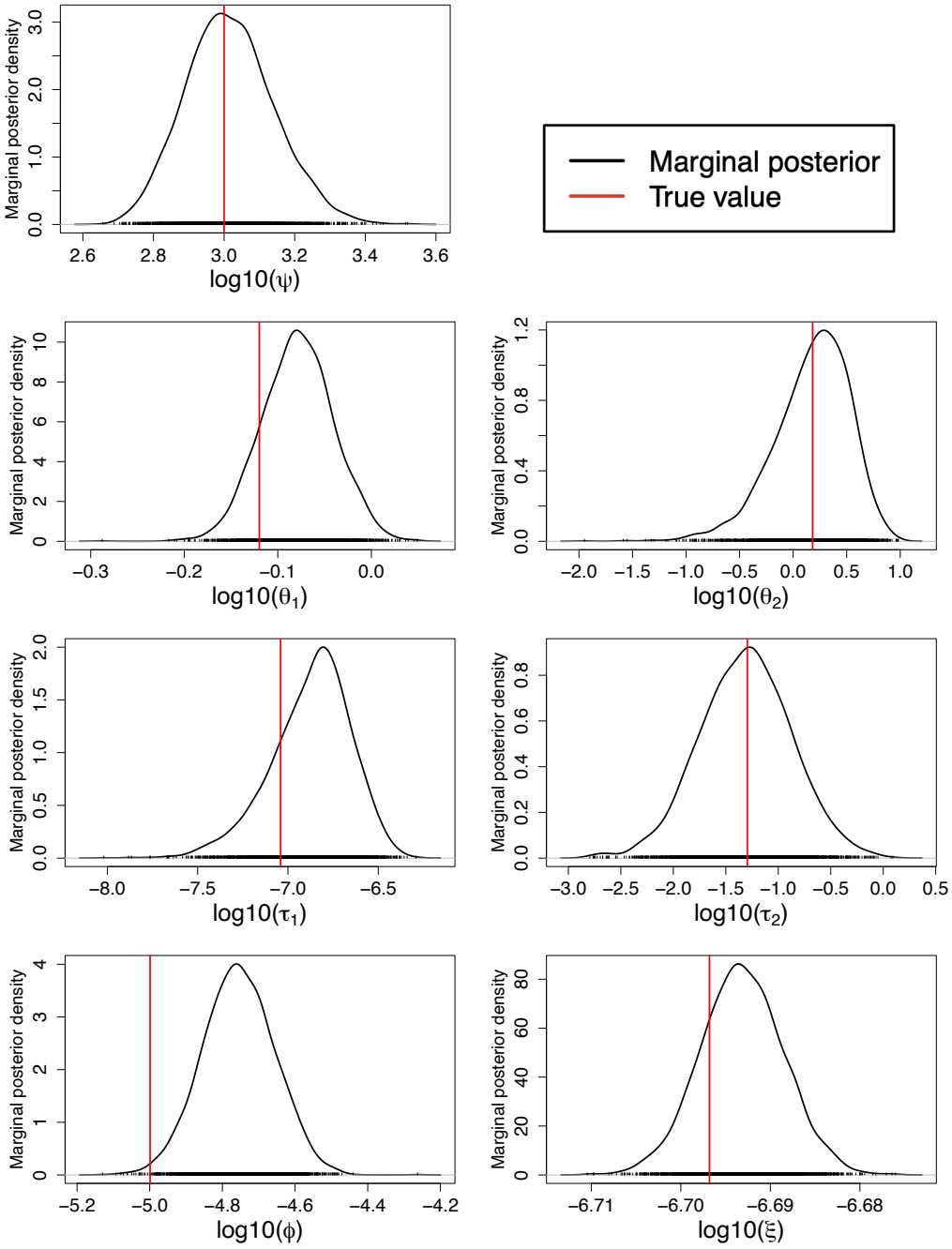


Figure 4.6: Marginal posterior distributions for the main parameters in the simulation study (Section 4.6.2) illustrated in $\log_{10}$-scale. The location of the true values used to simulate the data are shown in red.

Table 4.3: Summary statistics of the marginal posteriors for the main parameters in the simulation study (Section 4.6.2).

	Mean	Median	SD	2.5%	97.5%	True value
ψ	1082	1019	336	612	1916	1000
θ_1	0.84	0.83	0.07	0.7	0.98	0.759
θ_2	1.97	1.66	1.39	0.24	5.51	1.52
τ_1	1.5e−07	1.42e−07	7.04e−08	4.24e−08	3.14e−07	0.91e−7
τ_2	8.19e−02	5.03e−02	1.04e−01	7.18e−03	3.46e−01	5.07e−2
ϕ	1.8e−05	1.75e−05	4.22e−06	1.12e−05	2.76e−05	1.00e−5
ξ	2.03e−07	2.03e−07	2.18e−09	1.98e−07	2.07e−07	2.01e−7

unobserved sources is much lower, leading to lower expected population size ψ. In addition, there is visible correlation between τ_1 and θ_1. If τ_1 is small, then the first power law slope is flattened, given by smaller values of θ_1. In contrast, if τ_1 goes farther into the data, then the (negative) slope increase, given by larger values if θ_1. The remaining parameters do not show any visible correlations or noticeable patterns.

In summary, the simulation results demonstrate that our posterior distributions cover the true simulation parameters well, even under high proportions of systematic incompleteness. The Gibbs-type-sampler developed in Section 4.5 shows good convergence results and allows us to obtain samples (approximately) from the posterior distribution, with large effective sample sizes for most parameters. Given the promising simulation results, we now apply our model on real data CDFS data, presented in Section 4.7.

4.7 CDFS Data Analysis

In this section, we apply our novel model for incompleteness corrections to real X-ray data from the Chandra Deep Field South (CDFS), with optical source locations from Arnouts et al. (2001). We provide details regarding our data acquisition in Section 4.7.1. We provide details on the incompleteness functions employed for our analysis in Section 4.7.2, and then apply the single Pareto model, with results discussed in Section 4.7.3, and the double Pareto model, results discussed in Section 4.7.4 to the CDFS data. We discuss the

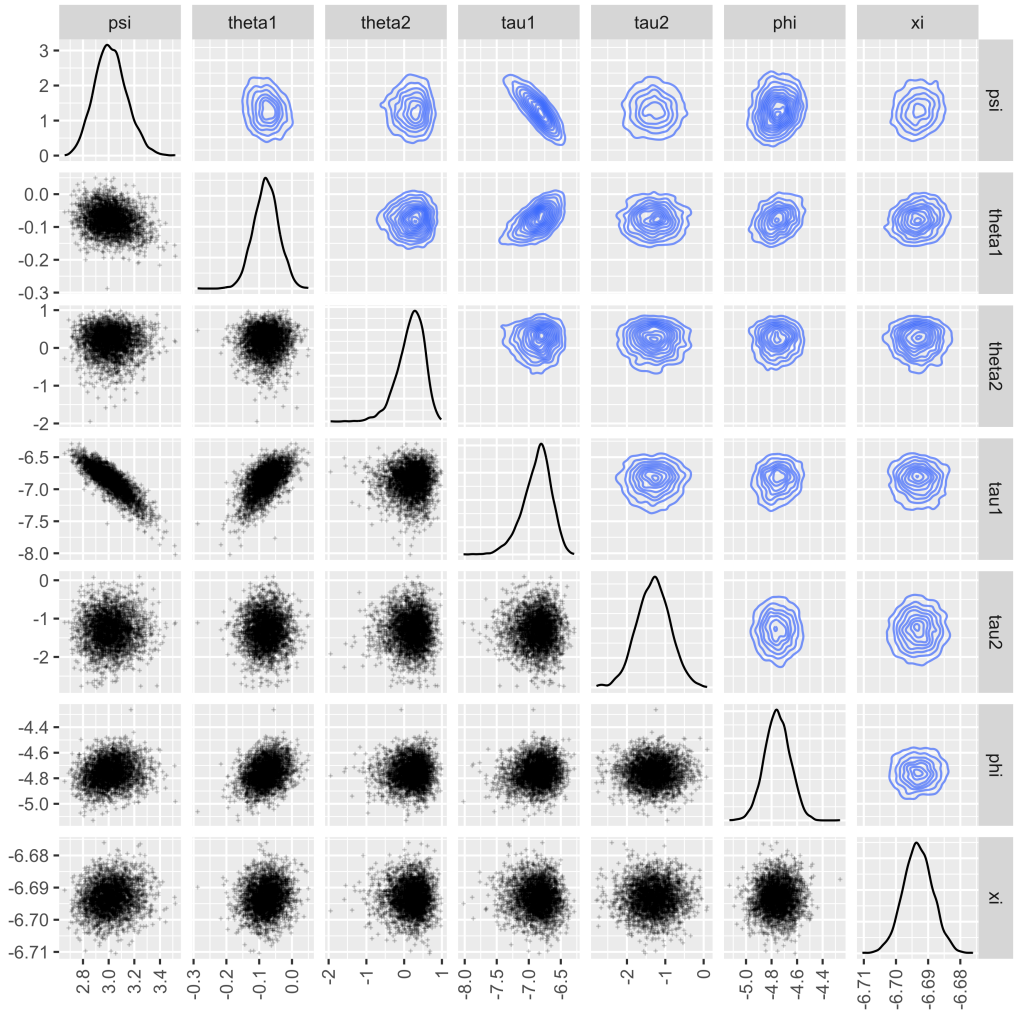


Figure 4.7: Pairwise marginal posterior distributions for the main parameters in the simulation study (Section 4.6.2), illustrated in $\log_{10}$-scale. (We note that left upper y-axis labels present the scale of the marginal density for ψ. In all other cases, the ticks labels present the parameter coordinates of the scatterplots.)

obtained $\log(N) - \log(S)$ fit in Section 4.7.5.

4.7.1 CDFS: Data Description

Chandra Deep Field South (CDFS): The Chandra Deep Field South (CDFS) was obtained via the Chandra X-ray Observatory (Weisskopf et al., 2002), a NASA operated space telescope with very high sensitivity to faint X-ray sources, which pointed at the same area of the sky for around 44 days to obtain the CDFS. In this research, we employ the merged 2000, 2007 and 2010 CDFS dataset.[13] The merged CDFS data has a total exposure time of $\mathcal{T} = 3.8 \times 10^6$ seconds. Figure 4.8 illustrates the CDFS event file for the merged dataset, with X-ray counts being illustrated in white. For the purposes of our analysis, we consider a subregion of the CDFS dataset within the red circular area illustrated in Figure 4.8.[14] The considered subregion (cicular area in Figure 4.8) has a total area of 1,168,052 pixel. (For illustration purposes, Figure 4.2 shows a smaller patch (zoomed in) within our circular area, to illustrate extracted source regions in (blue and yellow) as described below.)

To detect sources in the CDFS X-ray catalogue, we use the wavdetect source detection algorithm (Freeman et al., 2002) with a detection threshold of $w_0 = 10^{-6}$ to avoid background fluctuations being falsely detected as X-ray sources. We obtain a total of 225 detected X-rays sources in our subregion. Some of the detected X-ray source regions are illustrated in the patch shown in Figure 4.2, via yellow circles.

Optical catalogue: We now add data from the optical survey. More precisely, we consider optical sources from Arnouts et al. (2001). They describe a total of 45,912 optical sources detected in the v-band.[15] A total of

[13]The merged CDFS data can be found here link. The merged CDFS Dataset Identifier is: ADS/Sa.CXO#Contrib/ChandraDeepFieldS.

[14]We consider a circular area around the center of the CDFS observations, given at RA= $03:32:28.2$ and DEC = $-27:48:36.0$ (FK5). The radius of our considered circular area is 5 arcmin.

[15]The complete optical data catalogue, as described in Arnouts et al. (2001), can be accessed here: link

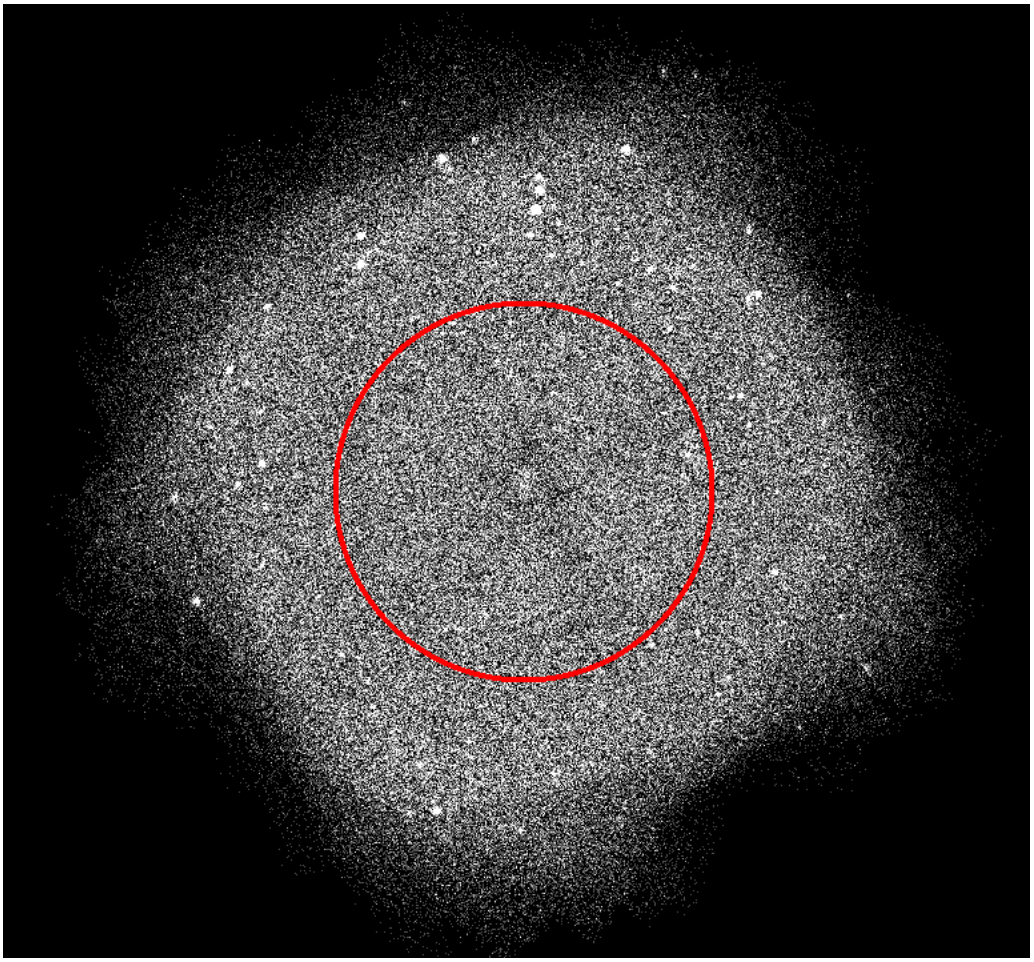


Figure 4.8: Illustration of the CDFS event file for the merged (2000,2007, and 2010) CDFS data as described in Section 4.7.1. For the purpose of our analysis, we consider the area within the red circle as our subregion, located at the center of the CDFS field.

4027 of these optical sources are located in our subregion of interest. We extract the exact location (not the area) of the 4027 optical sources located in our subregion. In the next paragraph, we describe the combination (matching) of the optical detected and X-ray detected sources and the calculation of the source areas.

Combining Optical detected sources with X-ray detected sources: We are now provided with source locations of sources detected in X-ray (via wavdetect), and of sources detected in optical (via the catalogue provided by Arnouts et al. (2001)). Since some of these source locations (optical and X-ray) might describe the same source, we have to match the optical catalogue with the X-ray catalogue. We first calculate the source areas a_i (in pixel) for each source. For all detected sources (X-ray and optical), we obtain the source areas via the X-ray point spread function (PSF) map. We employ a CDFS provided PSF map, assuming the PSF to follow a 2D normal distribution with i.i.d components (for each pixel). To expect around 99% of the photons to fall within the source region, we set the radius of the source area to be $3\sigma_{\text{PSF}}$ of the PSF, with σ_{PSF} being determined by the PSF map value at the respective source location. (We note that our model is flexible with regards to the PSF choice, and a more realistic PSF can easily be incorporated.)

Before matching the X-ray and optical catalogue, we have to perform an additional pre-processing step. In its current form, our model does not allow overlap between sources, however, there is a large proportion of optical sources with overlapping source areas. Thus, we remove all optical sources that overlap with at least one other optical source. We also remove all X-ray sources that is connected (overlaps) with overlapping optical sources.

Finally, given the source areas a_i for (non-overlapping) optical and X-ray sources (all based on the X-ray psf map), we consider an X-ray and optical detected source to be the same, if the optical source area overlaps with the center of the X-ray detected source. This leads to $N_{\text{X-ray}} = 70$ sources that are detected in both, the X-ray and optical catalogue, and $N_{\text{opt}} = 1524$ sources being detected in the optical catalogue only, resulting in a total of

$N_{\text{opt}} = 1594$ observed sources. For all these sources, we consider the source location to be at the reported location of the respective optical source, due to higher accuracy in the optical source location.[16]

To obtain the data needed for our model, for $i = 1, \dots, N_{\text{Nobs}}$, with $N_{\text{Nobs}} = N_{\text{X-ray}} + N_{\text{opt}}$, we now extract the source counts Y_i at the respective source locations of sources i in the source regions a_i. Further, we define an indicator variable Z_i, with $Z_i = 2$, if source i is detected in X-ray (and optical), and $Z_i = 1$, if source i is detected in the optical catalogue only. We obtain the effective area e_i for source i via the CDFS provided exposure map value at the source location.[17]. Finally, we extract a background area $A = 962,795.9$ (in pixel), which we consider to be the entire area of our subregion, without the detected source areas.[18] We obtain a total background count of $X = 1,654,328$.

4.7.2 CDFS: Incompleteness Function Details

In order to obtain the marginalized incompleteness function $g_{\text{int}}(\lambda)$ as derived in (4.35), we have to introduce distributions on the nuisance parameters L, e, and ξ. The realistic CDFS incompleteness function introduced by Wright et al. (2015), as presented in (4.30), is obtained via MCMC simulations, and

[16] We note that for the analysis in this section, we withdraw X-ray sources without optical match. The uncertainty on the exact location of X-ray sources is much higher than the uncertainty on the optical detected source locations. By assuming that optical matches to X-ray sources are missing at random, we do not introduce any systematic effects by removing the X-ray sources without match, but we avoid the large uncertainty on the source location for these sources, which could substantially affect the number of source counts reported for the respective source area.

[17] We note that, in previous work by Stampoulis (2017), a scaling factor was applied to the exposure map to obtain the effective areas as the (scaled) average exposure map value at the source region. Since we had no access to the scaling factor employed by Stampoulis (2017) to the time of this study, we scaled the exposure map values by a factor of 10^{19}, to obtain effective area values that are in a similar range as the values employed by Stampoulis (2017). This scaling factor can easily be adjusted in future studies. we note that changing the scaling factor might change the scale of source intensities λ_i via (4.2). This might lead to a shift of the $\log(N) - \log(S)$ curve on the x-axis (a similar shift as described for the count-to-flux rate described in Section 4.1)

[18] The background area might contain sources that are not detected. In a secondary analysis step, we consider accounting for the thus added background counts in future work.

values for the incompleteness function by Wright et al. (2015) are reported as tabular data for L and ξ.

More precisely, Wright et al. (2015) report values for $\xi_{\mathrm{T}} \in \{0.00072, 0.0030, 0.012, 0.048, 0.192, 0.768\}$, and $L \in \{0, 1, 2, 3, 4, 5\}$. We a priori assume the background rates (in count/pixel)[19] to be same likely. By assuming that the population of sources is equally distributed across the sky, we obtain the prior probabilities for $p(L = j)$, $j \in \{0, 1, 2, 3, 4, 5\}$ by looking at the areas of each off-axis angle proportional to the entire subregion. We thus obtain $p(L = j) \in \{0.004, 0.034, 0.094, 0.185, 0.306, 0.377\}$ ordered for $j \in \{0, 1, 2, 3, 4, 5\}$. Finally, we model $e \sim N(\mu_e = 1.85, \sigma_e = 0.1)$, which we found to cover the distributions of effective areas well.

Given the finite grid on L and ξ values, we can simplify the integral in (4.32) via

$$\int p(L)p(e)p(\xi)P(Z = 2|\lambda, L, e, \xi)dL, d\xi, de \tag{4.52}$$

$$= \int \sum_{L} \sum_{\xi} p(L)p(e)p(\xi)P(Z = 2|\lambda, L, e, \xi)de \tag{4.53}$$

The blue curve in Figure 4.9 shows the marginalized X-ray incompleteness function $g_{\mathrm{int}}(\lambda)$ evaluated and illustrated on a fine grid of λ values.

Finally, for the optical incompleteness function h_{opt} we choose a similar functional form as for the X-ray incompleteness function. More precisely, we model h_{opt} via

$$h_{\mathrm{opt}}(\lambda|\phi) = (1 - g_{\mathrm{int}}(\lambda)) \left(1 - \exp\left(-\frac{\lambda}{\phi}\right)\right), \tag{4.54}$$

with ϕ being a free model parameter. The first part of the function (the dependence on the X-ray incompleteness; $(1 - g_{\mathrm{int}}(\lambda))$), is added to ensure that the sum of optical and X-ray incompleteness function is between zero and one, i.e., $0 \leq g_{\mathrm{int}}(\lambda) + h_{\mathrm{opt}}(\lambda|\phi) \leq 1$, for any λ. (The blue curve in

[19] We recall that $\xi_T = \xi\mathcal{T}$. Using the scale (in count/pixel/s), the grid values are $\xi \in \{1.89\mathrm{e}{-10}, 7.89\mathrm{e}{-10}, 3.16\mathrm{e}{-09}, 1.26\mathrm{e}{-08}, 5.05\mathrm{e}{-08}, 2.02\mathrm{e}{-07}\}$.

Figure 4.9 illustrates the optical incompleteness function $h_{\text{opt}}(\lambda|\hat{\phi}_{DP})$ with $\hat{\phi}_{DP}$ being the mean marginal posterior of ϕ in the double Pareto model, with detailed results described in Section 4.7.4).

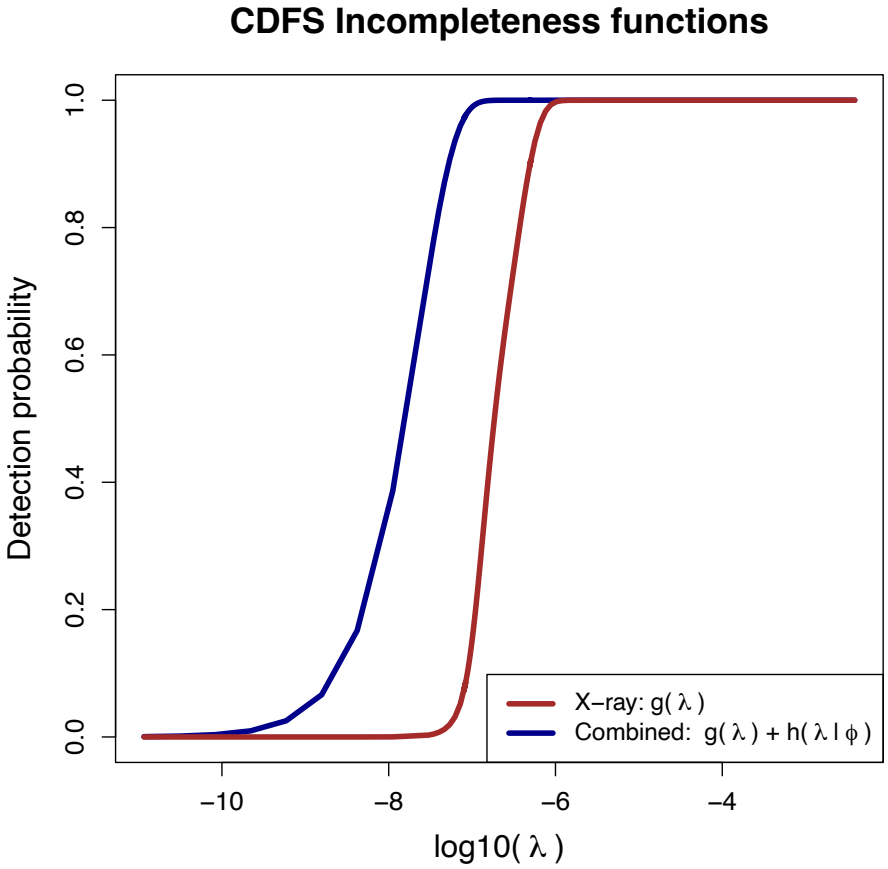


Figure 4.9: Incompleteness functions for the CDFS data analysis. In the figure legend, $g(\lambda)$ refers to the marginalized X-ray incompleteness function $g_{\text{int}}(\lambda)$, calculated via (4.52). For illustration purposes, for the optical incompleteness function $h_{\text{opt}}(\lambda|\phi)$, we set the free model parameter ϕ in this figure to the mean posterior value obtained for ϕ in the double Pareto model cases, i.e., $\hat{\phi}_{DP} = 2.29\text{e}{-08}$, as discussed in Section 4.7.4.

4.7.3 CDFS: Single Pareto Model

Single Pareto – Model details and sampler diagnostics: We fit our model as described in Section 4.5, using the single Pareto population f_{SP}.

We run $M_{\text{SP}} = 10$ chains of our Metropolis-within-Gibbs sampler to obtain posterior draws. We again run each chain for $T = 80,000$ iterations, use a burn-in of the first $T_{\text{Burn}} = 20,000$ iterations, and combine the remaining $60,000$ samples of each chain for posterior inference. Table 4.4 provides sampling diagnostics, with Gelman-Rubin $\hat{R}$ being close to 1 for all parameters. Traceplots of the chains for all main parameters are illustrated in Figure C.2, indicating convergence of the chain to the stationary distribution (the posterior distribution). Auto-correlations are somewhat higher in this real data case compared to the simulation study, but the effective sample sizes (as shown in Table 4.4) are again reasonably large for posterior inference in this case.

Table 4.4: Posterior sampling diagnostics for parameters CDFS single Pareto (Section 4.7.3).

	ψ	θ_1	τ_1	ϕ	ξ
Gelman-Rubin $\hat{R}$	1.01	1.05	1.04	1.04	1.00
Effective size	854	407	318	1001	21138

Posterior distributions for single Pareto model: In Figure 4.10, we illustrate the marginal posterior distributions (in $\log_{10}$-scale) obtained for the main parameters of the single Pareto model in the CDFS data analysis. In $\log_{10}$-scale, all five parameters show ordinary bell shaped marginal posterior distributions. Summary statistics of the marginal posterior distributions are presented in Table 4.5. Notably the marginal posterior distribution of the expected sample size ψ indicates a large proportion of incompleteness, with a mean marginal posterior of $\hat{\psi}_{\text{SP}} = 108,323$ the expected population size is much larger than the total number of observed sources $N_{\text{opt}} = 1594$.

We further note that the marginal posterior distribution that we obtain for the power law slope θ_1, with a posterior mean of $\hat{\theta}_{\text{SP}} = 0.52$, is systematically lower than the one reported by Stampoulis (2017), with a mean posterior of 0.963. We note that Stampoulis (2017) considers 358 X-ray sources detected in the CDFS field (based on a wider subregion). They further do not consider any optical sources, and therefore do not exclude any overlapping sources.[20]

[20]We recall that overlap of sources only occurs by adding sources detected in the optical catalogue.

Table 4.5: Summary statistics (in original scale) of the marginal posteriors for the main parameters in the CDFS data analysis via the single Pareto population model (Section 4.7.3).

	Mean	Median	SD	2.5%	97.5%
ψ	108,323	106.947	21,642	70,631	153,945
θ_1	0.52	0.52	0.02	0.47	0.56
τ_1	2.27e−12	2.02e−12	1.25e−12	5.47e−13	5.26e−12
ϕ	2.32e−08	2.3e−08	5.42e−09	1.31e−08	3.44e−08
ξ	4.51e−07	4.51e−07	3.38e−10	4.51e−07	4.52e−07

Model differences and data differences could thus explain the effect. We note that a comparison between our estimate of the population minimum value τ_1 and the value reported by Stampoulis (2017) is not meaningful, due to the different scales of flux S (considered by Stampoulis (2017)) and source intensities λ (considered in our study). An appropriate count-to-flux scaling factor will be applied in future work to provide direct comparisons.

We further note, that our posterior mean for the background rate $\hat{\xi}_{\text{SP}} = 4.51\text{e}{-07}$, leads to a value of $\xi_T = \hat{\xi}_{\text{SP}}\mathcal{T} = 1.71$ (in count/pixel) which slightly exceeds the largest tabulated value by Wright et al. (2015), which is 0.768. This is mainly due to much longer exposure times of the CDFS data in our study ($\mathcal{T} = 3.8 \times 10^6$ seconds, vs. 1.96×10^6 seconds considered by Wright et al. (2015)). In future work, we will carefully extend the grid values of the X-ray incompleteness function to account for this difference.

Figure 4.11 illustrates the pairwise marginal posterior distributions. As already observed in our simulation study, there is a correlation between population minimum τ_1 and power-law slope θ_1. Lower values for τ_1 correspond to a flatter power law slope θ_1. In addition, a lower population minimum τ_1 leads to larger expected population size ψ. Further, there are moderate correlations visible between the optical incompleteness parameter ϕ and θ_1, τ_1.

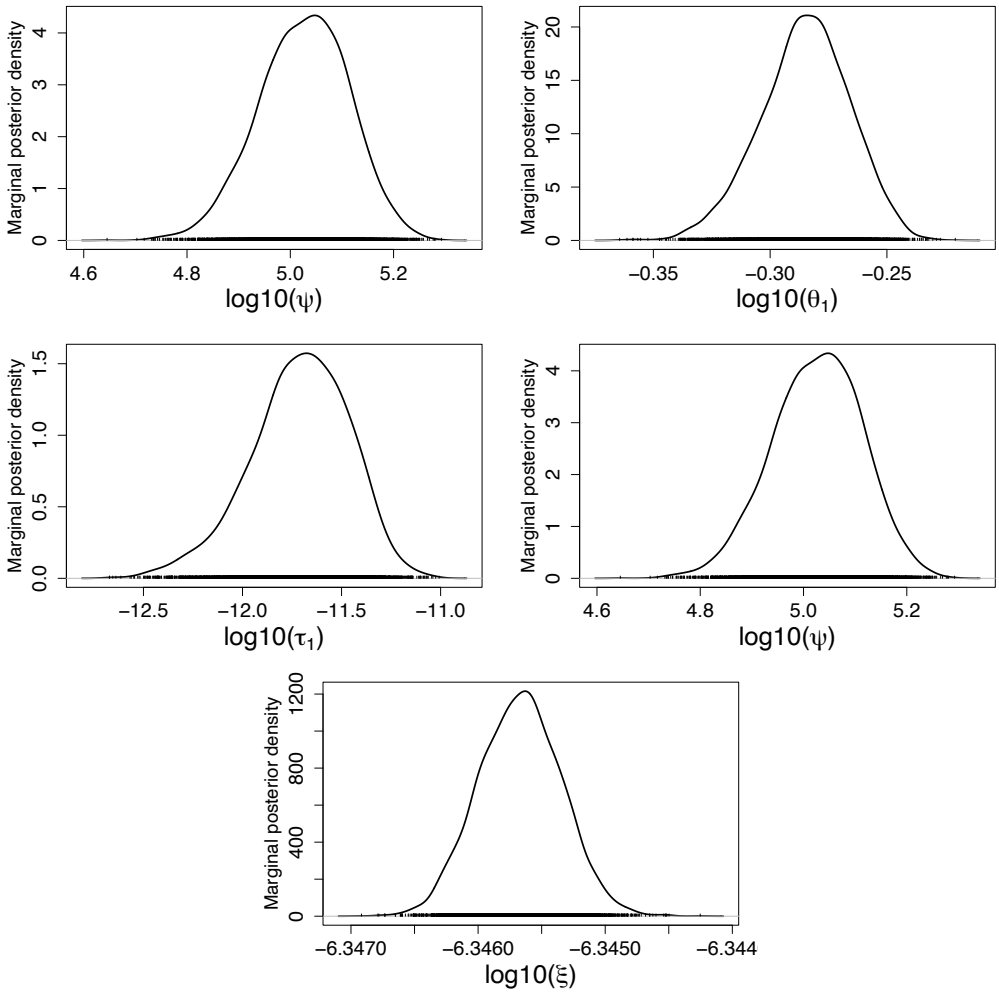


Figure 4.10: Marginal posterior distributions for the main parameters in CDFS data analysis via the single Pareto model (Section 4.7.3), illustrated in $\log_{10}$-scale.

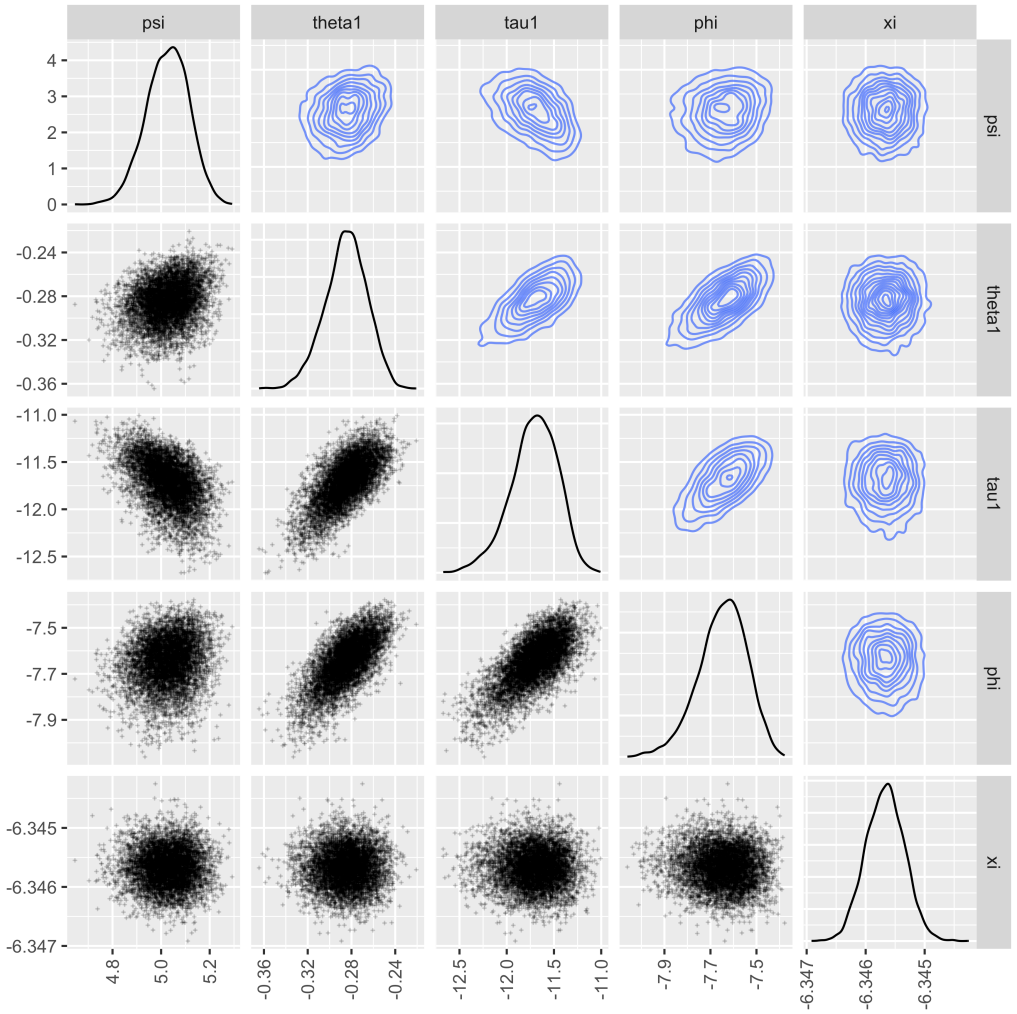


Figure 4.11: Pairwise marginal posterior distributions for the main parameters in the CDFS data analysis via the single Pareto model (Section 4.7.3), illustrated in $\log_{10}$-scale. (We note that left upper y-axis labels present the scale of the marginal density for ψ. In all other cases, the ticks labels present the parameter coordinates of the scatterplots.)

4.7.4 CDFS: Double Pareto Model

Double Pareto – Model details and chain diagnostics: In this paragraph we fit our model via the double Pareto population f_{DP}, running $M_{\text{DP}} = 10$ chains of our Metropolis-within-Gibbs sampler to obtain posterior draws, again for $T = 80,000$ iterations with a burn-in of the first $T_{\text{Burn}} = 20,000$ iterations. Numerical chain diagnostics via the Gelman-Rubin $\hat{R}$ (shown in Table 4.6), and visual inspection (via traceplots; Figure C.3) indicate convergence of the chains to the posterior distribution. Effective sample sizes are reasonably large for posterior inference.

Table 4.6: Posterior sampling diagnostics for parameters CDFS double Pareto (Section 4.7.4).

	ψ	θ_1	θ_2	τ_1	τ_2	ϕ	ξ
Gelman-Rubin $\hat{R}$	1.04	1.04	1.00	1.08	1.00	1.04	1.00
Effective size	926	399	19310	276	67014	1032	20286

Posterior distributions for double Pareto model: Compared to the single Pareto model, the double Pareto model includes two additional parameters, the breakpoint τ_2, and the second power law slope θ_2. We first compare the parameter estimates of the single and double Pareto model for the remaining parameters that are included in both models. Table 4.7 presents summary statistics of the marginal posterior distributions obtained via the double Pareto model, and Figure 4.13 illustrates the pairwise marginal distributions. Notably, the posterior distributions of the parameters $\tau_1, \theta_1, \psi, \phi$ and ξ are almost identical for the single and double Pareto fit. In turn, the posterior distributions of the additional double Pareto parameters τ_2 and θ_2 have very large uncertainties, as illustrated via the marginal posterior distributions (shown in original scale) in Figure 4.12. In fact, the marginal posterior distribution of θ_2 shows high probabilities in regions that have high probability of θ_1. This indicates that there is little evidence for a breakpoint and a second power law slope in this case. We note that such a result can be expected for the CDFS data at hand. Being a *deep field* survey with very long-lasting exposure time, for the CDFS survey, a region of the sky with (almost) no bright sources was selected to facilitate the detection of faint

sources. Breakpoints are generally required for such very bright sources.

Table 4.7: Summary statistics of the marginal posteriors for the main parameters in the CDFS data analysis via the double Pareto population model (Section 4.7.4).

	Mean	Median	SD	2.5%	97.5%
ψ	110,704	109,192	21,405	72,660	156,631
θ_1	0.52	0.52	0.02	0.47	0.56
θ_2	2.14	1.83	1.62	0.12	6.1
τ_1	2.19e−12	1.93e−12	1.26e−12	5.21e−13	5.25e−12
τ_2	5.53e−04	5.31e−04	2.37e−04	1.65e−04	1.13e−03
ϕ	2.29e−08	2.27e−08	5.46e−09	1.30e−08	3.42e−08
ξ	4.51e−07	4.51e−07	3.37e−10	4.51e−07	4.52e−07

Notably, the marginal posterior distribution for τ_2 exhibits multiple modes, as illustrated in Figure 4.12. There is a small rise in the marginal posterior distribution of τ_2 close to zero. This small mode close to zero leads to the highly skewed, long left-tailed posterior distributions of τ_2 in $\log_{10}$-scale, as shown in Figure 4.13 (5^{th} row and column).

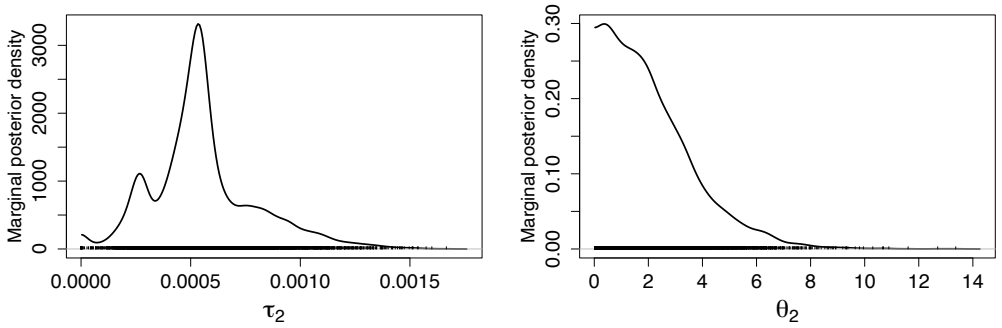


Figure 4.12: Marginal posterior distributions for τ_2 and θ_2 (in original scale) for the double Pareto model (Section 4.7.4). The marginal distributions of the remaining parameters are very similar to the single Pareto case, illustrated in Figure C.4 in the Appendix.

4.7.5 CDFS: $\log(N) - \log(S)$ Estimation

We finally compute the $\log(N) - \log(S)$ curve based on the posterior distributions obtained in the previous sections. Figure 4.14 presents the $\log(N) - \log(S)$ obtained via the single Pareto (blue dashed line) and the double Pareto

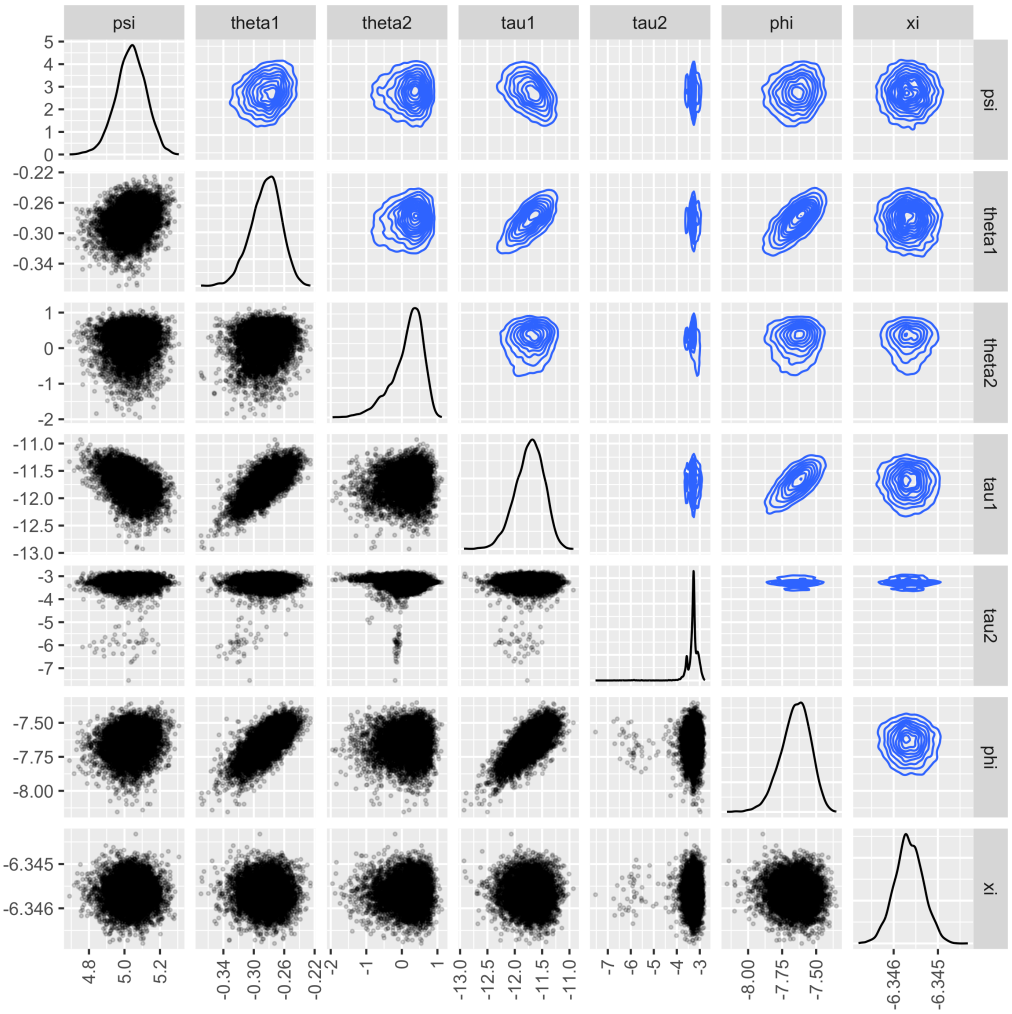


Figure 4.13: Pairwise marginal posterior distributions for the main parameters in the CDFS data analysis via the double Pareto model (Section 4.7.4), illustrated in $\log_{10}$-scale. (We note that left upper y-axis labels present the scale of the marginal density for ψ. In all other cases, the ticks labels present the parameter coordinates of the scatterplots.)

(orange solid line) model. More precisely, we compute the single Pareto $\log(N) - \log(S)$ via $N(> S) = (1 - F_{\mathrm{SP}}(\lambda; \tau_1, \theta_1))\psi$ and the double Pareto via $N(> S) = (1 - F_{\mathrm{DP}}(\lambda; \tau_1, \tau_2, \theta_1, \theta_2))\psi$, where F_{SP} denotes the CDF of the single Pareto distribution, and F_{DP} the CDF of the double Pareto distribution. In Figure 4.14, we use the mean of the marginal posterior distributions presented in Tables 4.5 and 4.7 as values for the population parameters and for ψ to compute $N(> S)$.

As expected from the results presented in Sections 4.7.3 and 4.7.4, the $\log(N) - \log(S)$ for the single Pareto and for the double Pareto model are almost identical. There is a visible difference between the two $\log(N) - \log(S)$ curves after the breakpoint τ_2 of the double Pareto model. However, given that at the location of the breakpoint τ_2, we have $\log_{10}(N > S) \approx 0.7$, and $10^{0.7} \approx 5$, the are around 5 sources expected in the single Pareto model that are brighter than in the double Pareto model after this breakpoint. For illustration purposes, we add the naive $\log(N) - \log(S)$ estimate as described in Section 4.1 (and previously illustrated in Figure 4.1). Of course, the naive estimator and our $\log(N) - \log(S)$ estimate lead to very different results. For instance, the naive estimator does not account for incompleteness, leading to a much lower population sizes (illustrated via lower values on the y-axis).

Lastly, Figure 4.9 illustrates the combined incompleteness function of X-ray and optical (in blue), using the posterior mean value for ψ (obtained via the double Pareto model) to compute $h_{\mathrm{opt}}(\lambda|\phi)$. Figure 4.9 illustrates that combining X-ray detected sources with sources detected in an optical catalogue leads to much improved completeness for lower source intensities. This underlines the advantage of combining the sources detected in X-ray and optical surveys, while carefully accounting for the incompleteness in both surveys. We provide an extended discussion and summary of the findings, as well as prospective future work in Section 5.3.

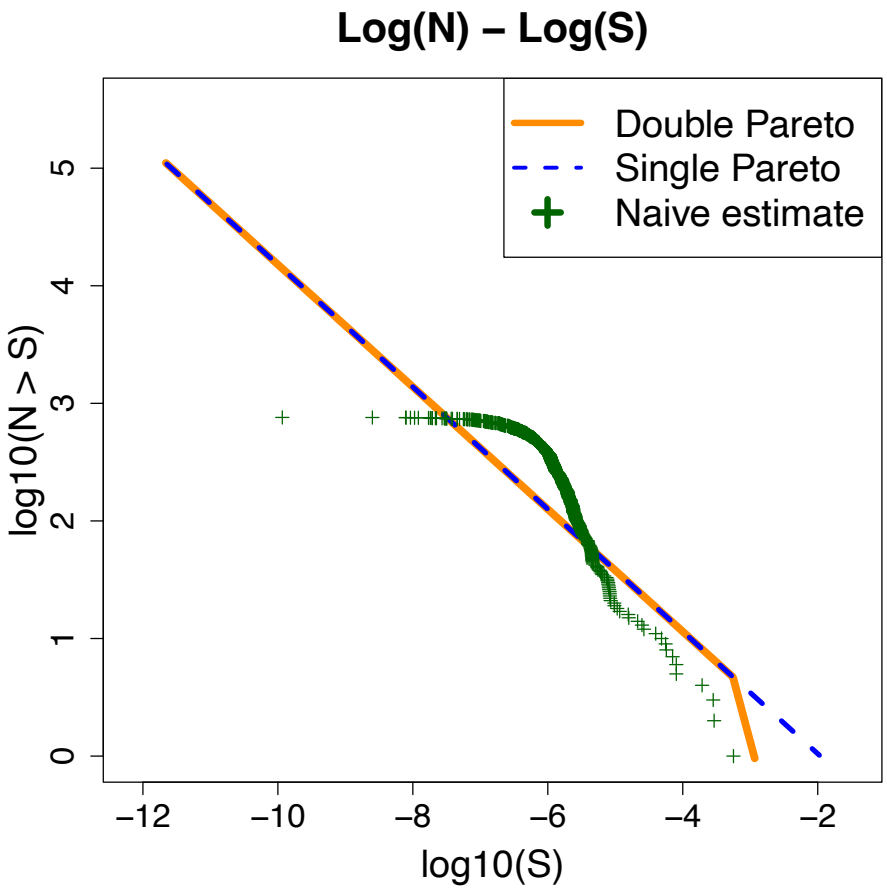


Figure 4.14: $\log(N) - \log(S)$ curve for the CDFS data analysis via the single (blue dashed line) and double (orange solid line) Pareto model. Posterior means are used from each model to obtain the presented $\log(N) - \log(S)$ curves. The green 'curve' is obtained via the naive estimate described in Section 4.1.

5

Discussion

In this last chapter, we provide a summary and discussion of the research conducted in this thesis. We discuss Chapters 2 to 4 separately, summarizing the key points of the projects, state limitations and outline prospective future work. We conclude with some final remarks in Section 5.4.

5.1 *StratLearn* – A Novel General-Purpose Method

In Chapter 2, we provide a simple, though statistically principled and theoretically justified method for learning under covariate shift conditions. We show that *StratLearn* outperforms a range of state-of-the-art importance weighting methods on toy covariate shift examples, and on two contemporary research questions in cosmology especially in the presence of a high-dimensional covariate space. The assumption of covariate shift is rather strong, requiring that there are no unmeasured confounding covariates – something that cannot be guaranteed in general. In a toy example, however, we demonstrate a certain robustness of our method against violation of this assumption. Further work is necessary to assess more fully the performance of *StratLearn* when this assumption is only approximately fulfilled. We emphasize that the

covariate shift framework is best justified in the presence of a large number of covariates mitigating the risk of unmeasured confounders – in which case it is critical to adopt a method that, like *StratLearn*, can robustly handle many covariates. Our framework is entirely general and versatile, as illustrated with examples of regression, conditional density estimation and classification. Notably, our numerical demonstrations illustrate the advantage of using only a subset of the source data when formulating predictions for individual objects in the target, where the subset is chosen for its similarity to the target data in question (through stratification). This is a markedly different strategy to the widespread practice of including all possible available observations when fitting learning models.

The novelty of our approach is grounded in the transfer of the well-established causal inference propensity score framework (Rosenbaum and Rubin, 1983) to the domain adaptation/covariate shift setting, by demonstrating that a method established to obtain unbiased treatment effect estimates can be adapted to optimize the target risk of a supervised learner under covariate shift. In future work, this extension gives a chance to transfer hard won practical advice from causal inference (e.g., balance diagnostics, estimation of propensity scores, choice of included covariates (Rosenbaum and Rubin, 1984, Austin et al., 2007, Pirracchio et al., 2014), etc.) to the covariate shift framework. We will also explore the possibility of taking advantage of Proposition 1 through a matching approach (Imbens and Rubin, 2015), which could prove more sensitive to the underlying propensity score distribution. We believe *StratLearn* may become a powerful alternative to importance weighting, with a myriad of possible extensions and applications.

5.2 Photo-z Calibration

In Chapter 3, we propose a novel method that improves photometric redshift calibration for weak lensing. The central planck of our approach is the estimation of galaxy-per-galaxy photo-z conditional densities within a Bayesian hierarchical model, coupled with the *StratLearn* framework. The computation of galaxy-level conditional density estimates allows us to in-

troduce an alternative tomographic binning strategy to the usual photo-z point estimate based binning (e.g. z_B-based binning). We presented a hierarchical Bayesian framework, (*StratLearn*-Bayes), to model summaries of the conditional density estimates to obtain nearly unbiased photometric redshift population mean estimates within tomographic bins.

We evaluated our method on a comprehensive and realistic simulation study (Wright et al., 2020, Van Den Busch et al., 2020) mimicking the KiDS+VIKING-450 dataset (Wright et al., 2019, Hildebrandt et al., 2020) with realistic photometric noise and spectroscopic incompleteness. The results of our study can be summarized in three points:

1. Our *StratLearn* conditional density-based tomographic binning strategy substantially improves upon the z_B tomographic bin assignment, with an overall binning accuracy of ~62.2% using *StratLearn*, compared to ~52.6% using z_B.

2. Our *StratLearn*-Bayes model leads to the lowest bias in the estimation of tomographic redshift population means. On average across the five tomographic bins, our proposed *StratLearn*-Bayes method leads to an absolute bias of 0.0052, a substantial improvement over the previously best calibration method, SOM with z_B binning (Wright et al., 2020), of 0.0085 average absolute bias. The strong reduction of bias is accompanied by a slightly increase in uncertainty, leading to an average standard deviation of 0.0067, compared to 0.0051 (Wright et al., 2020). Using our *StratLearn*-Bayes framework, we find a maximum bias of $\Delta\langle z\rangle = 0.0095 \pm 0.0089$, slightly below the potentially critical bias value of $\Delta\langle z\rangle > 0.01$, compared to SOM based on z_B binning, which leads to maximum biases of 0.0135 ± 0.0052 and 0.0147 ± 0.0040.

3. While the previously best calibration method, SOM based on z_B binning, requires systematic quality cuts to define a gold standard sample (Wright et al., 2020), our method does not require any cuts of the photometric sample. Thus, together with the improved tomographic bin assignment, our *StratLearn*-Bayes framework delivers an increase of ~18% in the galaxies available for the cosmic shear analysis.

It has long been understood that bias in the first moment (mean) estimate of the photometric redshift population distributions (within tomographic bin) is the pre-dominant source of systematic errors in cosmic shear (e.g, Amara and Refregier 2008, Wright et al. 2020). We show that exploiting the properties of a hierarchical Bayesian framework to model the photo-z conditional densities allows nearly unbiased recovery of the population means.

In this work, we modeled the redshift populations (within tomographic bins) as Gaussian distributions, which is not a realistic shape for the redshift population distributions. However, an accurate estimation of the population redshift distribution could also be incorporated in our framework: we provide a proof of concept demonstration in Section 3.5.5, where we employ inverse propensity score (inverse-PS) weighting of the spectroscopic data to recover the redshift population distributions within bins. To combine the best of both worlds, in a future work we will extend our *StratLearn*-Bayes method for nearly unbiased population mean estimates to incorporate the inverse-PS method to accurately capture the shape of the photometric redshift distributions. This will require a comprehensive assessment of the inverse-PS method, as well as the development of a principled combination strategy.

Finally, we believe that our improved tomographic binning assignment, the reduction of population mean bias within tomographic bin, and the increase in the number of galaxies available for cosmic shear analysis will have a substantial impact on the eventual scientific results and cosmological parameter inference. Re-analysing the KiDS+VIKING-450 data with our improved calibration method might lead to more precise and more accurate constraints on cosmological parameter estimates, particularly on S_8, the clustering strength of (predominantly dark) matter. We further believe that our proposed method might provide a powerful tool to improve the analysis of present and upcoming cosmic shear analysis.

5.3 Luminosity Distribution Estimation via Incomplete Surveys

In chapter 4, we develop a novel method to combine incomplete (X-ray and optical) surveys for estimation of the galaxy luminosity distribution in X-rays. Accurate modeling of the luminosity distribution allows astronomers to answer questions about the geometry and structure of the universe. A major complication in the estimation of galaxy luminosity distributions in X-ray is large incompleteness in the low-intensity end of the distribution. Sources with low-intensities are least likely to be detected, but most prevalent in the population. Ignoring the non-representativeness of the observed sources would lead to large inferential bias, and a potential misunderstanding of the underlying structure of the universe.

In this project, we demonstrate that optical survey data can be leveraged to obtain a much larger and more complete observed sample for low-intensity sources. We carefully account for the incompleteness in X-ray detected sources by incorporating a known incompleteness function, obtained via previous simulations (Wright et al., 2015). The incompleteness of the optically detected sources is modelled via a second (optical) incompleteness function with parameters fitted within the model. Evaluating our model on simulations realistic for the CDFS data, we show that we can accurately recover the luminosity function even under high proportions of incompleteness. Our simulations as well as the CFDS data application show a moderate correlation between the population minimum parameter τ_1 and the expected population size ψ, as well as the first power-law slope θ_1. Underestimating τ_1 would lead to a strong inflation of the expected population size and a flattening of the first power-law slope. In our simulations, all three parameters are well recovered by our model, providing confidence for our model fit on the CDFS data. We consider additional analysis steps regarding the reliability of the τ_1 population minimum estimate in future work.

In our analysis of the CDFS data, the single Pareto population model and the double Pareto population model lead to almost identical results, i.e., there is not much evidence for a power-law breakpoint τ_2. The CDFS survey was

constructed to observe faint sources via long-lasting exposure times. To reduce interference, a field of the sky without occurrences of very bright X-ray sources was selected. Breakpoints are usually required for such (very) bright sources. We will apply and evaluate our model on additional X-ray surveys to investigate the presence of one (or additional) power-law breakpoints (e.g., data from Chandra Cygnus OB2 Legacy Survey (Wright et al., 2014)). The hierarchical structure of our model allows relatively straightforward extensions to models with additional breakpoints, e.g. a triple Pareto model, as briefly discussed by Stampoulis (2017).

In addition, there might be occurrences of sources that are detected in an optical catalogue, having optical emissions, but no X-ray emissions (being X-ray *dark*). Such sources would have zero probabilities of being detected in X-ray. To investigate such cases, a zero-inflated component could be added to the model, allowing source intensities to be zero. Such a component might effect the estimate for the population minimum τ_1 of sources that have X-ray emissions. We will consider a zero-inflated model option in future work.

Another major line of future work is the consideration of overlapping sources. Modelling each segment of overlap between sources would allow he incorporation of overlapping sources into the model, which would strongly increase the available data size, and might thus substantially improve model precision. A downside of modelling the segments of overlap is the increasing model complexity and large number of added parameters, which might make model fitting more challenging. There are additional technical challenges and extensions considered in future work, such as including more realistic (individual) point spread functions for each source to obtain more accurate source areas, add a realistic count-to-flux rate (potentially with errors) to obtain the $\log(N) - \log(S)$ based on the source flux S, adjust the exposure map scaling, and account for background contamination via non-detected sources. In addition, our model could readily be extended to include additional surveys e.g. radio surveys, or other optical surveys. A combined incompleteness function, including all additional sources, could directly be fitted in the current form of our model. Our model could also be extended to allow individual incompleteness functions for any additional survey.

We note that the intricacy of the scientific problem makes the design of realistic simulations that cover all data aspects challenging. We are confident that the simulations provided in Chapter 4 realistically cover major aspects of the CDFS data analysis. Implementing additional simulations (e.g. considering various types of misspecifications) could provide further evidence for our proposed model, and will be part of future work.

5.4 Concluding Remarks

This thesis considers several cases of non-representative data in astrophysics. We highlight that disregarding non-representativeness of the observed data with respect to its underlying population might lead to very inaccurate statistical models, non-reliable statistical inference, and potentially to an erroneous understanding of the Universe. We emphasize that the (current and upcoming) availability of large data does not alleviate, but might aggravate systematic errors arising through non-representativeness of observed data.

In our research, we provide novel general-purpose methodology for statistically principled learning from non-representative data, presented in Chapter 2. We believe that our novel framework will have a myriad of applications to account for non-representativeness in astrophysics, as well as in other scientific fields. We extend our method proposed in Chapter 2 to develop a general methodology for a topical scientific issue, the calibration of photometric redshift for weak lensing, as presented in Chapter 3. We believe that our proposed method might mitigate inferential biases and discrepancies in present and future scientific weak lensing analyses. In chapter 4, we design a highly principled, scientifically justified statistical model tailored to an intricate scientific problem, the estimation of X-ray galaxy luminosity distributions under strong incompleteness.

Non-representativeness of the observed data is often difficult to detect and diagnose, but prevalent and an underlying issue in many scientific problems. In this thesis, we demonstrate several examples of non-representative data, and provide solutions of how non-representativeness can be tackled depending on the level of background and detail available, ranging from readily applica-

ble general-purpose methodology to highly tailored statistical and scientific analysis. The broad range of research conducted in this thesis allows several exciting avenues for further statistical and scientific explorations, with the overall goal of mitigating bias in our scientific understanding.

A

Additional Results for Chapter 2

A.1 Additional Results for SNIa Identification

This Section provides additional results and details complementing Section 2.5, which describes the classification of supernovae (SNe) into Type Ia (SNIa) and non-SNIa based on photometric light curve data, given a strongly biased source data set from Kessler et al. (2010a).

Comparison of Importance Weighting Methods: In addition to importance sampling (as presented in Figure 2.7), we investigate two further importance weighting approaches to obtain a comprehensive performance comparison of importance weighting and *StratLearn*. More precisely, we implement IWCV, and a combination of IWCV and importance sampling. Table A.1 presents the resulting target AUC for all implemented weighting methods. In summary, none of the importance weighting methods are competitive with *StratLearn*, which obtained a target AUC of 0.958. In fact, using IWCV alone leads to a decrease in performance for NN (0.897) and

Table A.1: AUC results on the updated SPCC (photometric) target data (Kessler et al., 2010a), using various importance weighting approaches to adjust for covariate shift.

	IWCV	importance sampling	IWCV + importance sampling
uLSIF	0.906	0.902	0.906
NN	0.897	0.923	0.914
IPS	0.897	0.921	0.924

Table A.2: Average (SD) of SMD and KS-stats computed on the observed covariates from the updated SPCC data (diffusion map coordinates, redshift and brightness) on "raw" data, and in strata built by *StratLearn* and STACCATO, respectively. Strata 3-5 are not displayed due to a shortage of source data. Smaller values indicate better balance.

	"raw" data	STACCATO		*StratLearn*	
		Stratum 1	Stratum 2	Stratum 1	Stratum 2
SMD	0.114 (0.202)	0.053 (0.039)	0.189 (0.195)	0.034 (0.028)	0.103 (0.161)
KS-stats	0.187 (0.084)	0.074 (0.05)	0.206 (0.093)	0.056 (0.048)	0.131 (0.08)

IPS (0.897) relative to the Biased fit (0.902). In order to reweight the loss of each object separately when implementing IWCV, we optimize with respect to the log-loss instead of AUC. We do not expect this to be responsible for the decreased target performance. For example, using log-loss as the hyper-parameter selection criterion for *StratLearn* in each source group (instead of AUC) leads to the same target AUC of 0.958. Importance sampling is the only weighting method for which we present the associated ROC curve in Figure 2.7, because it has the highest average AUC.

Covariate Balance on Updated SPCC Data: Table A.2 presents the covariate balance on the updated SPCC data before stratification ("raw" data), and after stratification via STACCATO and *StratLearn*, respectively.

On the non-stratified "raw" data, we measure an average (SD) SMD of 0.114 (0.202) across the 102 covariates, with an average (SD) KS-stats of 0.187 (0.084). *StratLearn* leads to increased balance within strata, strongly reducing the average (SD) SMD to 0.034 (0.028) in stratum 1; KS-stats: 0.056 (0.048); and to an average (SD) SMD of 0.103 (0.161); KS-stats: 0.131 (0.08); in stratum 2. (The remaining strata contain too few source data to assess covariate balance.)

On average (SD) across all covariates, STACCATO led to an SMD of 0.053 (0.039); KS-stats: 0.074 (0.05), in stratum 1, and an average (SD) SMD of 0.189 (0.195), KS-stats: 0.206 (0.093), in stratum 2, much higher than *StratLearn*.

A.2 StratLearn applied to original SPCC Data

This section provides additional numerical evidence for *StratLearn*, performing SNIa classification on the original "Supernova photometric classification challenge" (SPCC) data set provided by (Kessler et al., 2010b), an earlier version of the SPCC data described in Section 2.5. The application of *StratLearn* on the original SPCC data is mainly for comparison with previous literature that use this data set.

Data: The earlier version of the SPCC data features a total of 17,330 simulated SNe of type Ia (SNIa), Ib, Ic and II. The data set is divided into a source (training) set $\boldsymbol{D}_S$ of 1217 spectroscopically confirmed SNe with known types, and (target set) $\boldsymbol{D}_T$ of 16,113 SNe with unknown types and only photometric information. The simulation used to generate the original SPCC data suffered from a bug that made classification easier, thus leading to the updated SPCC data used in Section 2.5. As in Section 2.5, we applied GP fitting and diffusion maps (Revsbech et al., 2018, Richards et al., 2012) to obtain a set of 102 predictive covariates; 100 covariates from the diffusion map, plus redshift (defined in Section 2.6) and a measure of the objects' brightness (Revsbech et al., 2018).

Results: Table A.3 presents the composition of the five strata obtained by conditioning on the estimated propensity scores, exhibiting a similar pattern as the strata composition on the updated SPCC data (Table 2.4), though with a higher proportion of SNIa in the first stratum and even less source data in strata $3-5$. We thus follow the same strategy as described in Section 2.5. After stratification, a random forest classifier is trained and optimized on source strata $\boldsymbol{D}_{S_1}$ and $\boldsymbol{D}_{S_2}$ separately to classify SNe in target strata $\boldsymbol{D}_{T_1}$

Table A.3: Composition of the five strata on the original SPCC data (Kessler et al., 2010b). The number of SNe, as well as the number and proportion of SNIa are presented in each source and target stratum.

Stratum	Set	Number of SNe	Number of SNIa	Prop. of SNIa
1	Source	996	794	0.80
	Target	2470	1759	0.71
2	Source	210	56	0.27
	Target	3256	1010	0.31
3	Source	9	0	0
	Target	3457	385	0.11
4	Source	2	1	0.50
	Target	3464	258	0.07
5	Source	0	0	NA
	Target	3466	180	0.05

and $\boldsymbol{D}_{T_2}$, respectively. Source strata $\boldsymbol{D}_{S_j}$ for $j \in \{3, 4, 5\}$ are merged with $\boldsymbol{D}_{S_2}$ to train the random forest to predict $\boldsymbol{D}_{T_j}$ for $j \in \{3, 4, 5\}$. Repeated 10-fold cross validation (with the hyperparameter grid described in Section A.1) is performed to minimize the empirical risk of (2.17) within each group.

With *StratLearn* we obtain an AUC of 0.977 on the target data of the original SPCC data (blue ROC curve in Figure A.1), very near the optimal 'gold standard' benchmark, which is 0.981 on this data set (obtained on a set with 1217 objects randomly selected from the target data as a representative source set, which would not exhibit covariate shift and would not be available in practice). The AUC obtained with *StratLearn* is also larger than the previous best published AUC for this data, AUC = 0.961 (Revsbech et al., 2018) with STACCATO, which employs a data augmentation strategy and target data leakage. Note that the predictive target performance on the original SPCC data (Kessler et al., 2010b) is generally higher than on the updated version (Kessler et al., 2010a), due to the aforementioned bug that made classification easier.

Most of the importance weighting methods (Table A.4) lead to slight improvements over the (non-adjusted) Biased model fit. The best performing weighting approaches (using estimated NN or IPS weights for importance

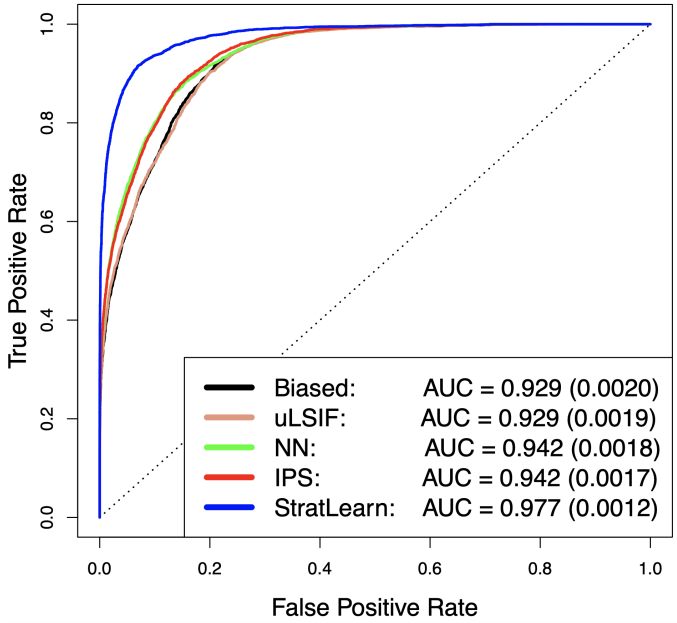


Figure A.1: Comparison of ROC curves for SNIa classification using the original SPCC data (Kessler et al., 2010b). Bootstrap AUC standard errors (from 400 bootstrap samples) are presented in parenthesis.

Table A.4: AUC results on the original SPCC (photometric) target data (Kessler et al., 2010b), using various importance weighting approaches to adjust for covariate shift.

	IWCV	importance sampling	IWCV + importance sampling
uLSIF	0.934	0.929	0.936
NN	0.934	0.942	0.931
IPS	0.934	0.942	0.931

sampling) obtain a target AUC of 0.942, substantially lower than *StratLearn* (AUC: 0.977). Importance sampling is the only weighting method for which we present the associated ROC curve in Figure A.1, because it has the highest average AUC.

A.2.1 Covariate Balance Check on Original SPCC Data

In this section, we illustrate the balancing property of propensity scores by assessing the covariate balance in the original SPCC data within strata conditional on the estimated propensity scores, such as described in the last

paragraph of Section 2.3. In particular, we show that such balance assessment could be helpful to assess the suitability of choice of covariates and/or the propensity score model.

As mentioned in Section 2.5, the method proposed by Revsbech et al. (2018) (STACCATO) can be viewed as a prototype of *StratLearn*, as it augments the spectroscopic source data separately in strata based on the estimated propensity scores. STACCATO employs two covariates ("redshift" and measure of brightness) as main effects in a logistic regression model to estimate the propensity scores. In contrast, *StratLearn* includes all 100 diffusion map covariates (Section 2.5), in addition to "redshift" and "brightness", as main effects in a logistic regression model to estimate the propensity scores. We assess the balance achieved through stratification by means of two commonly used balance measures: absolute standardized mean differences and the Kolmogorov-Smirnov test statistics (Austin, 2011, Mccaffrey et al., 2013), computed for each covariate within strata. Table A.5 shows that, on average across observed covariates, *StratLearn* leads to substantially reduced absolute standardized mean differences and Kolmogorov-Smirnov test statistics in strata 1 and 2, compared to STACCATO. We only report balance measures in strata 1 and 2 because of a shortage of source data in the remaining strata. Thus, building the strata by conditioning on the *StratLearn* estimated propensity scores leads to improved covariate balance (i.e., reduced covariate shift) within strata, compared to using the STACCATO estimated propensity scores. Notably, comparing the predictive performance, STACCATO (without augmentation) yields a target AUC of 0.942, whereas with *StratLearn* we obtain an AUC of 0.977 on the target data – a substantial improvement resulting from the improved covariate balance by accounting for potentially confounding covariates.

For illustration purposes, in Figure A.2 we display a more detailed balance check, comparing the absolute standardized mean differences computed for each covariate in stratum 1 (obtained through *StratLearn* (black) and STACCATO (red)) with the absolute standardized mean differences of the covariates without stratification. In general, points below the diagonal line indicate better balance in the stratum compared to the balance on the 'raw'

Table A.5: Average (SD) of SMD and KS-stats computed on the observed covariates from the original SPCC data (diffusion map coordinates, redshift and brightness) on "raw" data, and in strata built by *StratLearn* and STACCATO, respectively. Strata 3-5 are not displayed due to a shortage of source data. Smaller values indicate better balance.

	"raw" data	*StratLearn*		STACCATO	
		Stratum 1	Stratum 2	Stratum 1	Stratum 2
SMD	0.126 (0.184)	0.031 (0.026)	0.066 (0.121)	0.057 (0.055)	0.132 (0.127)
KS-stats	0.186 (0.073)	0.053 (0.019)	0.095 (0.053)	0.084 (0.027)	0.163 (0.066)

non-stratified data. Even though Revsbech et al. (2018) (STACCATO) just include redshift and brightness in the propensity score estimation, both approaches balance the majority of the covariates. However, the black (*StratLearn*) standardized mean differences have visibly smaller values in the y-direction, indicating better balance compared to the approach in Revsbech et al. (2018) (red). Note that for clarity of Figure A.2, we removed one outlier (brightness) with a raw standardized mean difference of 1.5, which was very well balanced with a standardized mean difference of around 0.09 in stratum 1 by both approaches. Assessing the balance of individual covariates as illustrated in Figure A.2 might be particularly useful to evaluate (and improve) the resulting balance of covariates important for outcome prediction. Some domain-specific expertise might be necessary to identify such covariates. In Section 2.4.2, we show how covariate balance assessment can be employed to evaluate the propensity score model choice.

Table A.6 presents balance comparisons between STACCATO and *StratLearn* via predicted outcome labels (as described in Section 2.3.3). *StratLearn* leads to much better balance between predicted outcome proportions within strata 1 and 2, and thus to much higher p-values than STACCATO, implying a much weaker relation between source/target assignment and predicted outcomes. The comparison of predicted outcome proportions and p-values (between *StratLearn* and STACCATO) provide a straightforward indication to employ the *StratLearn* propensity score model for stratification, which in turn leads to much improved outcome model predictive performance of SNIa classification.

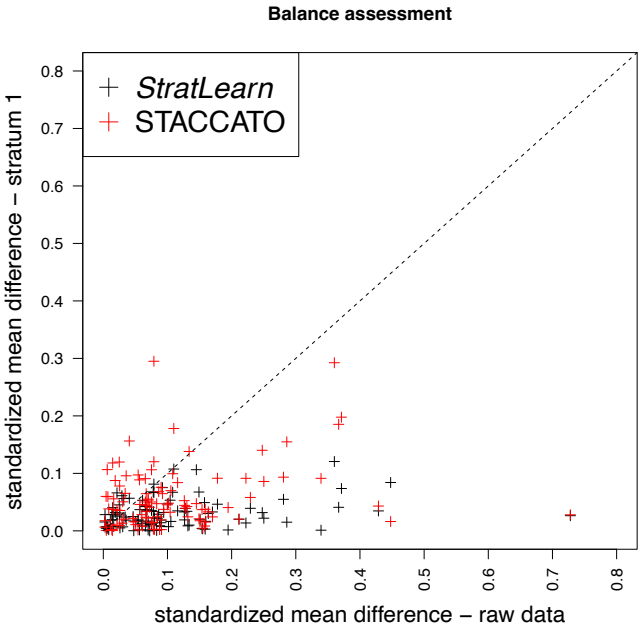


Figure A.2: Absolute standardized mean differences between source and target data of stratum 1 plotted against raw data absolute standardized mean differences for both propensity score (PS) estimation approaches (*StratLearn* and STACCATO). For visual clarity we left out one covariate which would appear at the coordinates (1.5,0.09).

Table A.6: Outcome balance diagnostics via predicted labels on the original SPCC data (Section A.2), applying STACCATO (left) and *StratLearn* (right). The number and proportion of predicted SNIa are presented in source and target strata 1 and 2. P-values are computed via Fisher's exact test of independence between predicted SNIa target and source proportions within strata.

		STACCATO (predicted)			*StratLearn* (predicted)		
Stratum	Set	Number of SNIa	Prop. of SNIa	p-value	Number of SNIa	Prop. of SNIa	p-value
1	Source	652	**0.69**	5.0e-21	794	**0.80**	0.013
	Target	1290	**0.51**		1872	**0.76**	
2	Source	181	**0.74**	6.9e-23	56	**0.27**	0.016
	Target	1336	**0.41**		1132	**0.35**	

A.3 Additional Details and Results for Conditional Photo-z Regression

In this section, we complement the results presented in Section 2.6. In section A.3.1, we provide derivations for the generalized risk in (2.21). In Section A.3.2, we provide additional numerical results.

A.3.1 Additional Details for Conditional Density Estimation under the L2-loss

In this section, following Izbicki et al. (2017), we provide derivations of the *generalized* risk under the L^2-loss given in (2.21). Following Izbicki et al. (2017), we start with the risk based on the general L^2-loss via

$$R_S(\hat{f}) = \iint (\hat{f}(z|x) - f(z|x))^2 dP_S(x)dz, \tag{A.1}$$

with $\hat{f}(z|x)$ being the full conditional density estimate of redshift z given the covariates at point x, $f(z|x)$ being the true conditional density of z given x, and $P_S(x)$ being the distribution of the source covariates. In extended form,

(A.1) can be written as

$$R_S(\hat{f}) = \iint \hat{f}^2(z|x)dP_S(x)dz - 2\iint \hat{f}(z|x)f(z|x)dP_S(x)dz \tag{A.2}$$

$$+ \underbrace{\iint f^2(z|x)dP_S(x)dz}_{=\text{ constant } C}, \tag{A.3}$$

which up to the constant C is equal to

$$R_S(\hat{f}) = \iint \hat{f}^2(z|x)dP_S(x)dz - 2\iint \hat{f}(z|x)dP_S(x,z). \tag{A.4}$$

From (A.2) to (A.4), the equality $dP_L(x,z) = f(z|x)dP_L(x)dz$ (via Radon-Nikodym derivative) is employed. Given the labelled source samples (x_S, z_S), we can get an estimate of (A.4) via

$$\hat{R}_S(\hat{f}) = \frac{1}{n_S}\sum_{k=1}^{n_S}\int \hat{f}^2(z|x_S^{(k)})dz - 2\frac{1}{n_S}\sum_{k=1}^{n_S}\hat{f}(z_S^{(k)}|x_S^{(k)}), \tag{A.5}$$

as presented in (2.21).

A.3.2 Additional Results for Photometric Redshift Regression

In the context of the photometric redshift regression example in Section 2.6, we present additional numerical results that demonstrate the robustness of *StratLearn* to varying degrees of covariate shift. Specifically, we reproduce the results of Section 2.6, but under two additional covariate shift scenarios.

Additional Covariate Shift Setups: We use the same data setup as in Section 2.6, but change the rejection sampling used to simulate an unrepresentative target D_T from D_S, to

- Weak covariate shift:

$$p(s=0|x) = f_{B(9,4)}(x_{(r)}) / \max_{x_{(r)}} f_{B(9,4)}(x_{(r)}) \tag{A.6}$$

- Strong covariate shift:

$$p(s=0|x) = f_{B(18,4)}(x_{(r)}) / \max_{x_{(r)}} f_{B(18,4)}(x_{(r)}), \tag{A.7}$$

where $x_{(r)}$ is the r-band magnitude and $f_{B(9,4)}$ and $f_{B(18,4)}$ are beta densities with parameters (9,4) and (18,4), respectively. Except for the adjusted degree of covariate shift, simulations are performed as described in Section 2.6.

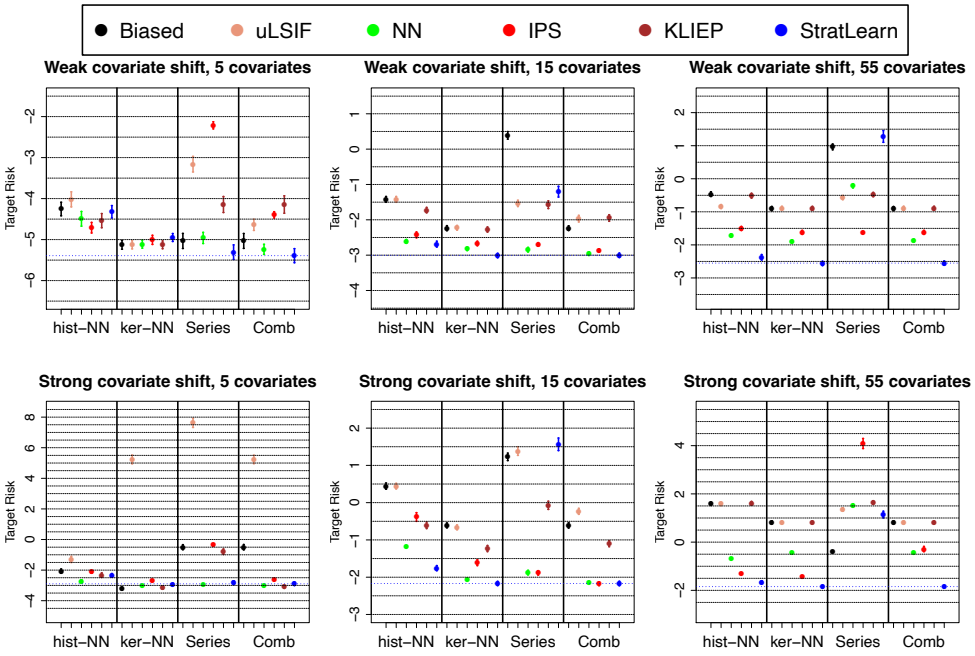


Figure A.3: Target risk ($\hat{R}_T$) of the four photo-z estimation models under each method (different colors), using different sets of predictors. Bars give the mean ± 2 bootstrap standard errors (from 400 bootstrap samples). Top row: Weak covariate shift (following (A.6)); Bottom row: Strong covariate shift (following (A.7)).

Summary – Conditional Density Estimation: Figure A.3 compares the resulting target risk $\hat{R}_T(\hat{f})$ across models and covariate sets, for the weak

Table A.7: Composition of the five *StratLearn* strata for the medium covariate shift scenario (as described in Section 2.6) on SDSS data, using estimated propensity scores with different sets of predictors. The number of galaxies in source and target stratum, as well as the mean outcome (redshift z) are presented.

Stratum	Set	5 covariates #galaxies (Mean z)	15 covariates #galaxies (Mean z)	55 covariates #galaxies (Mean z)
1	Source	1631 (0.09)	1583 (0.09)	1620 (0.09)
	Target	7 (0.07)	9 (0.07)	7 (0.07)
2	Source	1500 (0.13)	1515 (0.13)	1546 (0.13)
	Target	112 (0.11)	113 (0.13)	98 (0.11)
3	Source	618 (0.29)	641 (0.29)	594 (0.30)
	Target	1481 (0.33)	1499 (0.33)	1480 (0.33)
4	Source	116 (0.42)	114 (0.40)	108 (0.40)
	Target	2196 (0.39)	2215 (0.38)	2258 (0.38)
5	Source	135 (0.48)	147 (0.46)	132 (0.47)
	Target	2204 (0.48)	2164 (0.48)	2157 (0.48)
All	Source	4000 (0.16)	4000 (0.16)	4000 (0.16)
	Target	6000 (0.40)	6000 (0.40)	6000 (0.40)

Table A.8: Composition of the five *StratLearn* strata for the weak covariate shift (A.6) and strong covariate shift (A.7) scenarios on SDSS data, using a set of five predictors for propensity score estimation. The number of galaxies in source and target stratum, as well as the mean outcome (redshift z) are presented.

Stratum	Set	weak covariate shift, 5 covariates #galaxies (Mean z)	strong covariate shift, 5 covariates #galaxies (Mean z)
1	Source	1468 (0.09)	1603 (0.07)
	Target	145 (0.09)	0 (NaN)
2	Source	1220 (0.13)	1579 (0.10)
	Target	602 (0.11)	4 (0.10)
3	Source	740 (0.20)	637 (0.24)
	Target	1275 (0.21)	1494 (0.28)
4	Source	301 (0.37)	73 (0.36)
	Target	1958 (0.37)	2297 (0.33)
5	Source	271 (0.47)	108 (0.41)
	Target	2020 (0.44)	2205 (0.37)
All	Source	4000 (0.17)	4000 (0.13)
	Target	6000 (0.33)	6000 (0.34)

covariate shift (top row) and the strong covariate shift scenario (bottom row). These results, combined with those from the medium covariate shift scenario in Figure 2.9, reinforce the advantage of using *StratLearn*, especially for higher dimensional covariate sets. In Figures 2.9 and A.3, $\textit{StratLearn}_{Series}$ performs well in the setup with the fewest covariates (left column). In the

weak covariate shift scenario, this leads to an additional boost in performance of $\textit{StratLearn}_{Comb}$ (which combines $\textit{StratLearn}_{ker\text{-}NN}$ and $\textit{StratLearn}_{Series}$, following (2.20)). In the setups with higher dimensional covariates, the Series methods (using either *StratLearn* or the weighting methods) exhibit relatively poor performance overall. In these cases, the $\textit{StratLearn}_{Comb}$ predictions rely solely on $\textit{StratLearn}_{ker\text{-}NN}$, which again demonstrates the successful empirical risk minimization of (2.17), inasmuch as $\textit{StratLearn}_{Comb}$ automatically selects the better *target* model (i.e., $\textit{StratLearn}_{ker\text{-}NN}$) in each stratum, based on the respective empirical *source* risk estimates in each stratum. Overall, IPS and NN weighting are most competitive with *StratLearn*, in addition to KLIEP in the low dimensional, strong covariate shift setup. Note that the performances of IPS_{Comb}, $\text{KLIEP}_{\text{Comb}}$ and $\text{uLSIF}_{\text{Comb}}$ are degraded in some cases (e.g., Figure A.3, upper left panel) with inclusion of the poorly performing Series predictions, illustrating that the weighted empirical *source* risk minimization ((2.19), as a form of (2.1)) does not lead to *target* risk minimization in these situations. With increasing number of covariates, none of the weighted models yield results that are competitive with *StratLearn*.

Table A.7 presents the composition of the five *StratLearn* strata for the medium covariate shift scenario (described in Section 2.6) for each of the three sets of covariates. Each source stratum contains a sufficient sample size to fit the conditional density estimators for prediction on the respective target stratum. The overall redshift averages in the source (0.11) and target data (0.28) are very unequal, a consequence of the covariate shift. Within strata, the redshift averages are well aligned between source and target, indicating improved balance after stratification. Notably, the composition of the strata in the high-dimensional covariate setup is similar to the that of the lower-dimensional setups, with well-balanced strata. This is an indication of the robustness of *StratLearn* with respect to high dimensional (noisy) sets of covariates. Table A.8 compares the five *StratLearn* strata for the weak and strong covariate shift scenarios, again with well-aligned average redshift within each stratum. Note that for the strong covariate shift scenario there are almost no galaxies in the first two target strata. Thus, effectively only data in source stratum 3 to 5 are used for target prediction in the respective

strata. These examples illustrate the advantage of using only a small subset of the source data when formulating predictions for individuals in the target, where the subset is chosen for its similarity to the target data in question. This is a markedly different strategy to the widespread practice of including all possible available data when fitting machine learning models.

B

Additional Results for Chapter 3

B.1 Additional Details for *StratLearn*-Bayes Model

This section provides the theoretical justification of the posterior derivations in Section 3.3.3.

Deriving (3.18) from (3.17) is obtained by integrating over the product of normal densities in (3.17), which can analytically be done via completing the squares. Precisely, we want to show that

$$\int_{\mathbb{R}} N(z_i|\hat{\zeta}_i, \hat{\tau}_i^2) N(z_i|\mu, \sigma^2) dz_i = N(\hat{\zeta}_i|\mu, \hat{\tau}_i^2 + \sigma^2), \tag{B.1}$$

by completing the squares we obtain,

$$\int_{\mathbb{R}} N(z_i|\hat{\zeta}_i, \hat{\tau}_i^2) N(z_i|\mu, \sigma^2) dz_i \tag{B.2}$$

$$= \int_{\mathbb{R}} \frac{1}{\sqrt{2\pi\hat{\tau}_i^2}} \exp\left(-\frac{1}{2}\left(\frac{\hat{\zeta}_i - z_i}{\hat{\tau}_i}\right)^2\right) \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2}\left(\frac{z_i - \mu}{\sigma}\right)^2\right) dz_i$$

$$= \frac{1}{2\pi\hat{\tau}_i\sigma} \int_{\mathbb{R}} \exp\left(- \underbrace{\left(\frac{\sigma^2 + \hat{\tau}_i^2}{2\hat{\tau}_i^2\sigma^2}\right)}_{\text{a}} z_i^2 + \underbrace{\left(\frac{\hat{\zeta}_i\sigma^2 + \mu\hat{\tau}_i^2}{\hat{\tau}_i^2\sigma^2}\right)}_{\text{b}} z_i - \underbrace{\left(\frac{\mu^2\hat{\tau}_i^2 + \hat{\zeta}_i^2\sigma^2}{2\hat{\tau}_i^2\sigma^2}\right)}_{\text{c}} \right) dz_i$$

$$= \frac{1}{2\pi\hat{\tau}_i\sigma} \exp(-c) \int_{\mathbb{R}} \exp\left(-az_i^2 + bz_i\right) dz_i$$

$$= \frac{1}{2\pi\hat{\tau}_i\sigma} \exp(-c) \exp\left(\frac{b^2}{4a}\right)\left(\frac{\pi}{a}\right) \underbrace{\int_{\mathbb{R}} \frac{1}{\sqrt{2\pi\frac{1}{2a}}} \exp\left(-\frac{1}{2}\left(\frac{z_i - \frac{b}{2a}}{\sqrt{1/2a}}\right)^2\right) dz_i}_{=1}$$

$$= \frac{1}{\sqrt{2\pi(\sigma^2 + \hat{\tau}_i^2)}} \exp\left(\frac{-\mu^2\hat{\tau}_i^2 - \hat{\zeta}_i^2\sigma^2}{2\hat{\tau}_i^2\sigma^2} + \frac{(\hat{\zeta}_i\sigma^2 + \mu\hat{\tau}_i^2)^2}{2\hat{\tau}_i^2\sigma^2(\sigma^2 + \hat{\tau}_i^2)}\right)$$

$$= \frac{1}{\sqrt{2\pi(\sigma^2 + \hat{\tau}_i^2)}} \exp\left(-\frac{1}{2}\left(\frac{\hat{\zeta}_i - \mu}{\sqrt{\sigma^2 + \hat{\tau}_i^2}}\right)^2\right)$$

$$= N(\hat{\zeta}_i|\mu, \hat{\tau}_i^2 + \sigma^2), \tag{B.3}$$

which concludes (B.1). Alternatively, as a simpler argument, flipping the z_i and $\hat{\zeta}_i$ in (3.17) constitutes the standard normal-normal hierarchical model with Gaussian measurement errors on latent z_i with Gaussian population. The analytical derivation of this model is a standard result in Bayesian statistics (e.g., Gelman et al. 1995, page 116), demonstrating that

$$\int_{\mathbb{R}} N(\hat{\zeta}_i|z_i, \hat{\tau}_i^2) N(z_i|\mu, \sigma^2) dz_i = N(\hat{\zeta}_i|\mu, \hat{\tau}_i^2 + \sigma^2). \tag{B.4}$$

Mathematically, the densities $N(\hat{\zeta}_i|z_i, \hat{\tau}_i^2)$ and $N(z_i|\hat{\zeta}_i, \hat{\tau}_i^2)$ are identical, due

to the symmetry of the normal distribution. It thus directly follows that

$$\int_{\mathbb{R}} N(z_i|\hat{\zeta}_i, \hat{\tau}_i^2) N(z_i|\mu, \sigma^2) dz_i = N(\hat{\zeta}_i|\mu, \hat{\tau}_i^2 + \sigma^2), \tag{B.5}$$

which also concludes (3.18) from (3.17).

With a uniform conditional prior density $p(\mu_b|\sigma_b)$, the Gaussian conditional posterior of μ_b given $\sigma_b, \hat{X}_b$ in (3.19) then follows directly as another standard result (e.g., Gelman et al. 1995, page 117). More precisely, (3.18) is a product of Gaussian densities, which yields a Gaussian density (the log-posterior is quadratic in μ_b). The parameters of the Gaussian conditional posterior in (3.19) are obtained by considering the $\hat{\zeta}_i$ as independent estimates of μ_b with variances $(\hat{\tau}_i^2 + \sigma_b^2)$ (Gelman et al., 1995, page 117).

Posterior distribution of σ: In our hierarchical Bayesian model, the marginal posterior distribution of the population variance σ_b can be obtained via

$$p(\sigma_b|y) \propto p(\sigma_b) V_\mu^{1/2} \prod_{j=1}^{J} (\tau_j^2 + \sigma_b^2)^{-1/2} \exp\left(-\frac{(\hat{\zeta}_j - \tilde{\mu}_b)^2}{2(\hat{\tau}_j^2 + \sigma_b^2)}\right) \tag{B.6}$$

with $\tilde{\mu}_b$ and V_{μ_b} as defined in (3.20). Choosing a uniform prior on σ_b, $p(\sigma_b) \propto 1$, makes (B.6) a proper marginal posterior density (see, e.g., Gelman et al. 1995, page 117).

B.2 Additional Simulation Study Details

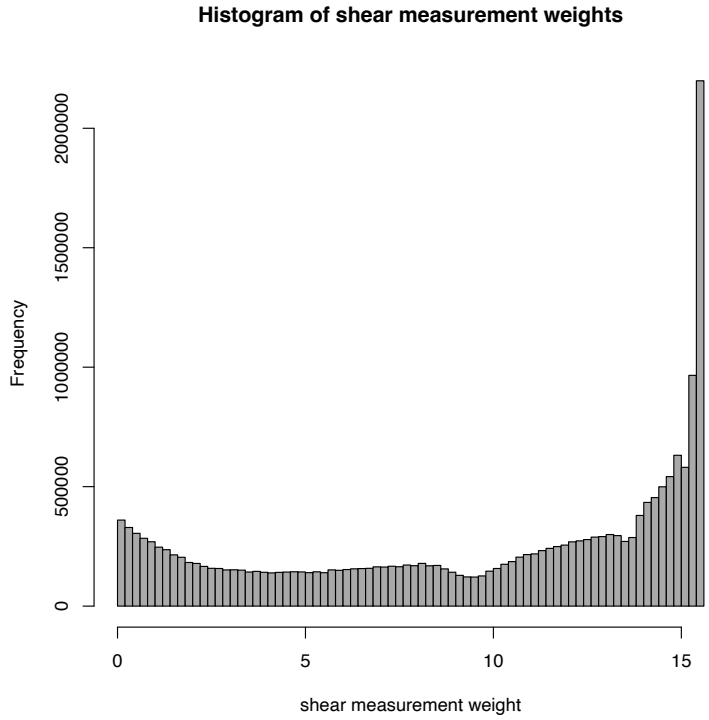


Figure B.1: Histogram of the shear-measurement weights $\hat{w}$ for galaxies in the photometric survey.

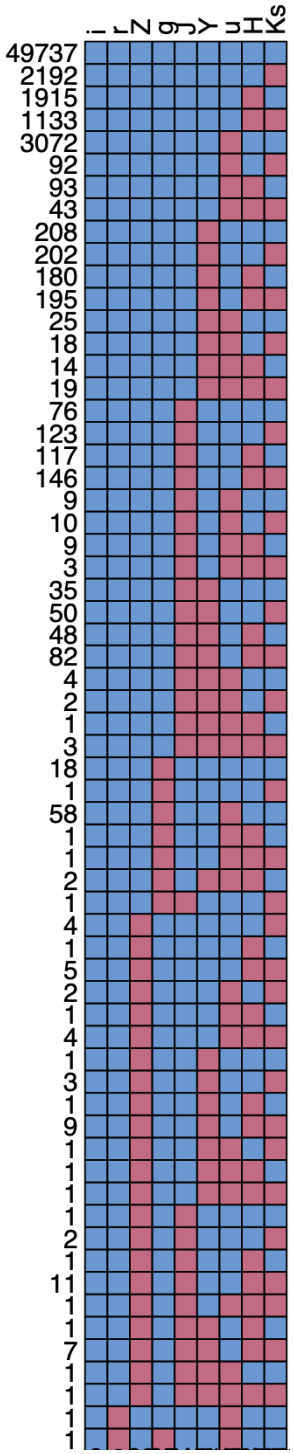


Figure B.2: "Missing data" pattern for magnitudes of a random subsample of 60,000 galaxies from the photometric survey.

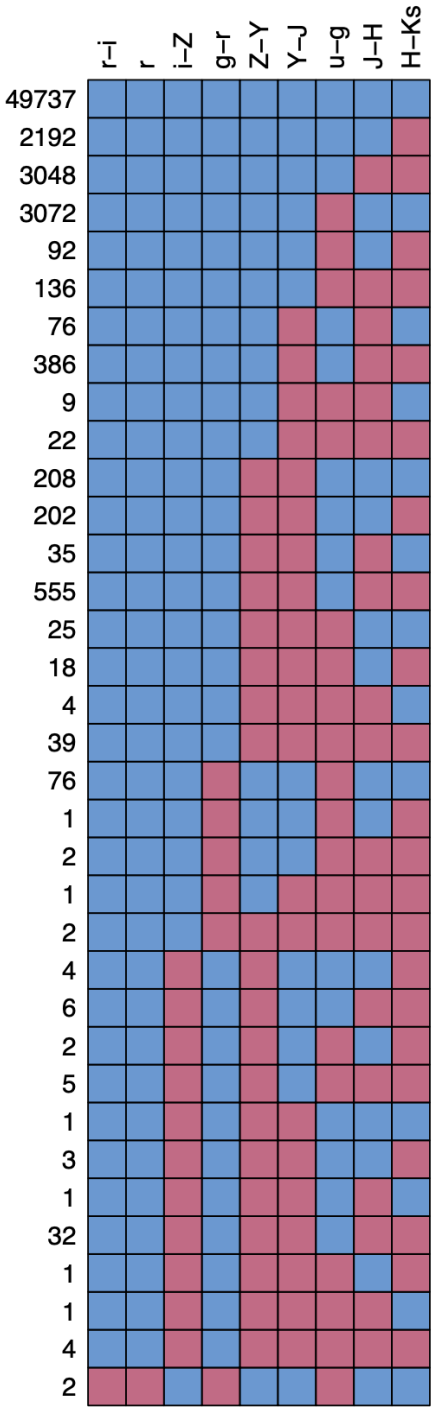


Figure B.3: "Missing data" pattern for colors of a random subsample of 60,000 galaxies from the photometric survey.

B.3 Additional Simulation Results

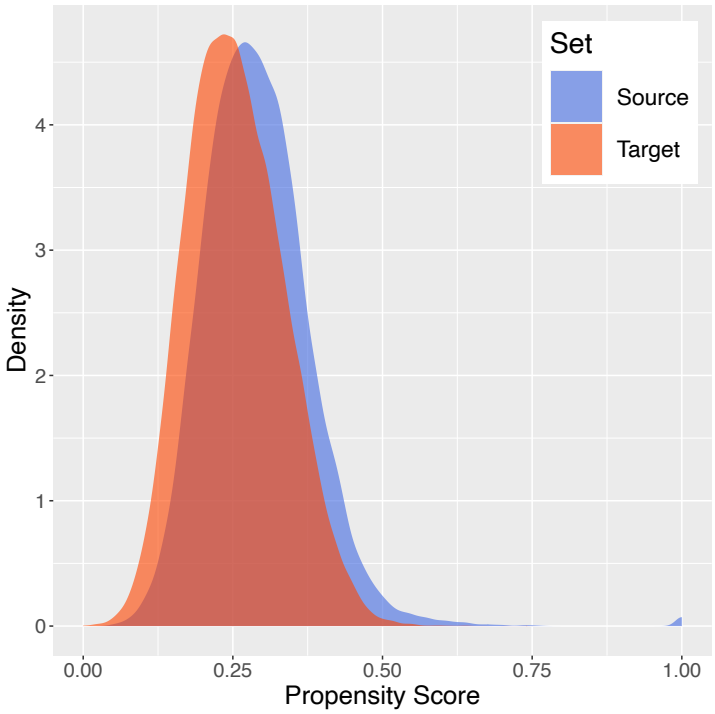
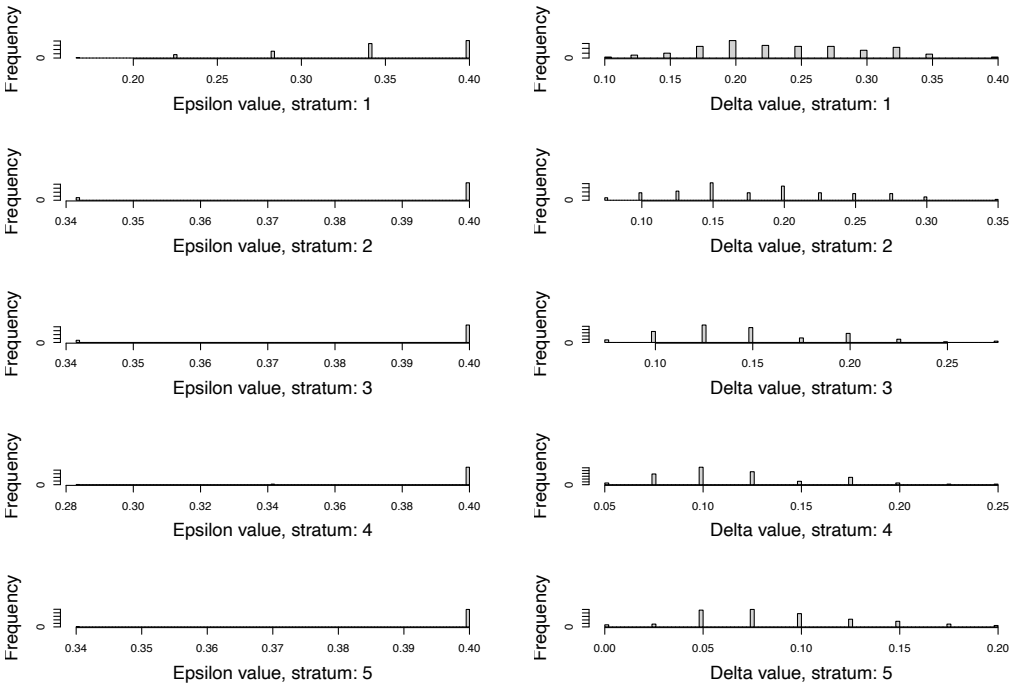
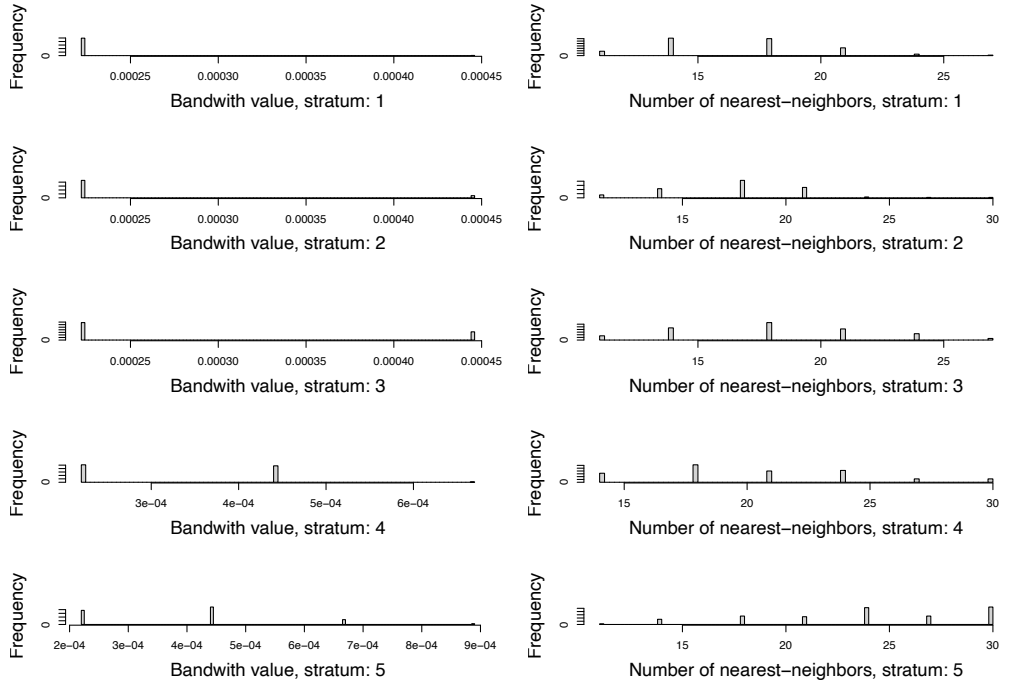


Figure B.4: Propensity score distributions of source and target data in Section 3.5, illustrated for one random LoS, and based on one random batch of 60,000 target samples, and the entire source sample.

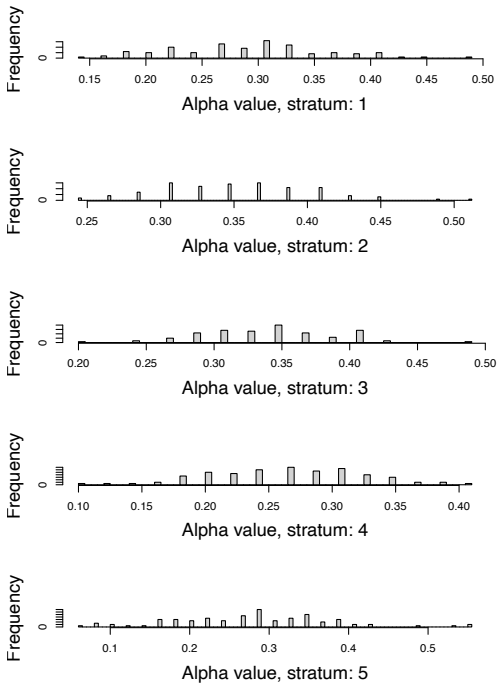


(a) Epsilon (Series estimator) **(b)** Delta (Series estimator)



(c) Bandwidth (ker-NN estimator) **(d)** Number nearest-neighbors (ker-NN)

Figure B.5: Histograms of optimized hyperparameters for the conditional density estimators. The histograms illustrate the optimized hyperparameters found for each of the 100 LoS (in each stratum separately).



(a) α_1 (Comb estimator)

Figure B.6: Histogram of the optimized α_1 hyperparameter for the Comb estimator (in each stratum, values across all 100 LoS).

C

Additional Results for Chapter 4

C.1 Transformation of Variables

Here we review the transformation of random variables, a basic statistical concept with useful applications in construction of MCMC samplers (Gelman et al., 1995). Let u be a random variable/vector with $u \in S \subseteq \mathbb{R}^n$ and let f be the density function of u. Further, let y be a random variable (vector) with $v \in T \subseteq \mathbb{R}^n$ and density function g, with transformation

$$v = r(u), \quad r : S \to T, \tag{C.1}$$

and r being an injective, differentiable function. It then holds that,

$$g(v) = f(u) \left| \det \left(\frac{du}{dv} \right) \right|. \tag{C.2}$$

The scale factor $\det \left(\frac{du}{dv} \right)$ is also called the *Jacobian* adjustment.

Table C.1: Prior distributions, and prior distribution parameters chosen for the various model runs, i.e., the double Pareto model on simulated data (Section 4.6.2), and the double-Pareto model on the CDFS data (Section 4.7.4). For the single Pareto model on the CDFS (Section 4.7.3), the same prior parameters are used as for the double Pareto model (without θ_2 and τ_2).

	Simulation study	CDFS data analysis
Prior	Prior params.	Prior. params
$\psi \sim \text{Gamma}(a_\psi, b_\psi)$	$a_\psi = 5, b_\psi = 0.005$	$a_\psi = 12, b_\psi = 0.0003$
$\theta_1 \sim \text{Gamma}(a_1, b_1)$	$a_1 = 2, b_1 = 1$	$a_1 = 2, b_1 = 1$
$\theta_2 \sim \text{Gamma}(a_2, b_2)$	$a_2 = 2, b_2 = 1$	$a_2 = 2, b_2 = 1$
$\tau_1 \sim \text{Gamma}(a_\tau, b_\tau)$	$a_\tau = 1.05, b_\tau = 5\text{e}6$	$a_\tau = 1.05, b_\tau = 5\text{e}6$
$\tau_2 \vert \tau_1 \sim \text{trunc-Norm}(\mu_\tau, \sigma_\tau^2,$	$\mu_\tau = \max(\hat{\lambda}_i^{\text{start}})$,	$\mu_\tau = \max(\hat{\lambda}_i^{\text{start}})$,
$a_0 = \tau_1, b_0 = \infty)$	$\sigma_\tau = (\max(\hat{\lambda}_i^{\text{start}}) - \min(\hat{\lambda}_i^{\text{start}}))/2$	$\sigma_\tau = (\max(\hat{\lambda}_i^{\text{start}}) - \min(\hat{\lambda}_i^{\text{start}}))/2$
$\phi \sim \text{Gamma}(a_\phi, b_\phi)$	$a_\phi = 1.1, b_\phi = 1e3$	$a_\phi = 1.05, b_\phi = 4e5$
$\xi \sim \text{Gamma}(a_\xi, b_\xi)$	$a_\xi = 1.75, b_\xi = 3e5$	$a_\xi = 1.02, b_\xi = 2e6$

In the univariate case, (C.2) can be written as

$$g(v) = f(u) \left| \frac{du}{dv} \right|, \tag{C.3}$$

with $u = r^{-1}(v)$ being the inverse function of r.

Log-transformation: Particularly for positive and skewed parameters, performing a parameter transformation via the logarithm is often useful. More precisely, we define $v = \log(u)$, then following (C.3), we obtain $g(v) = f(u)u$.

C.2 Additional Results for the Simulation Study and CDFS Data Analysis

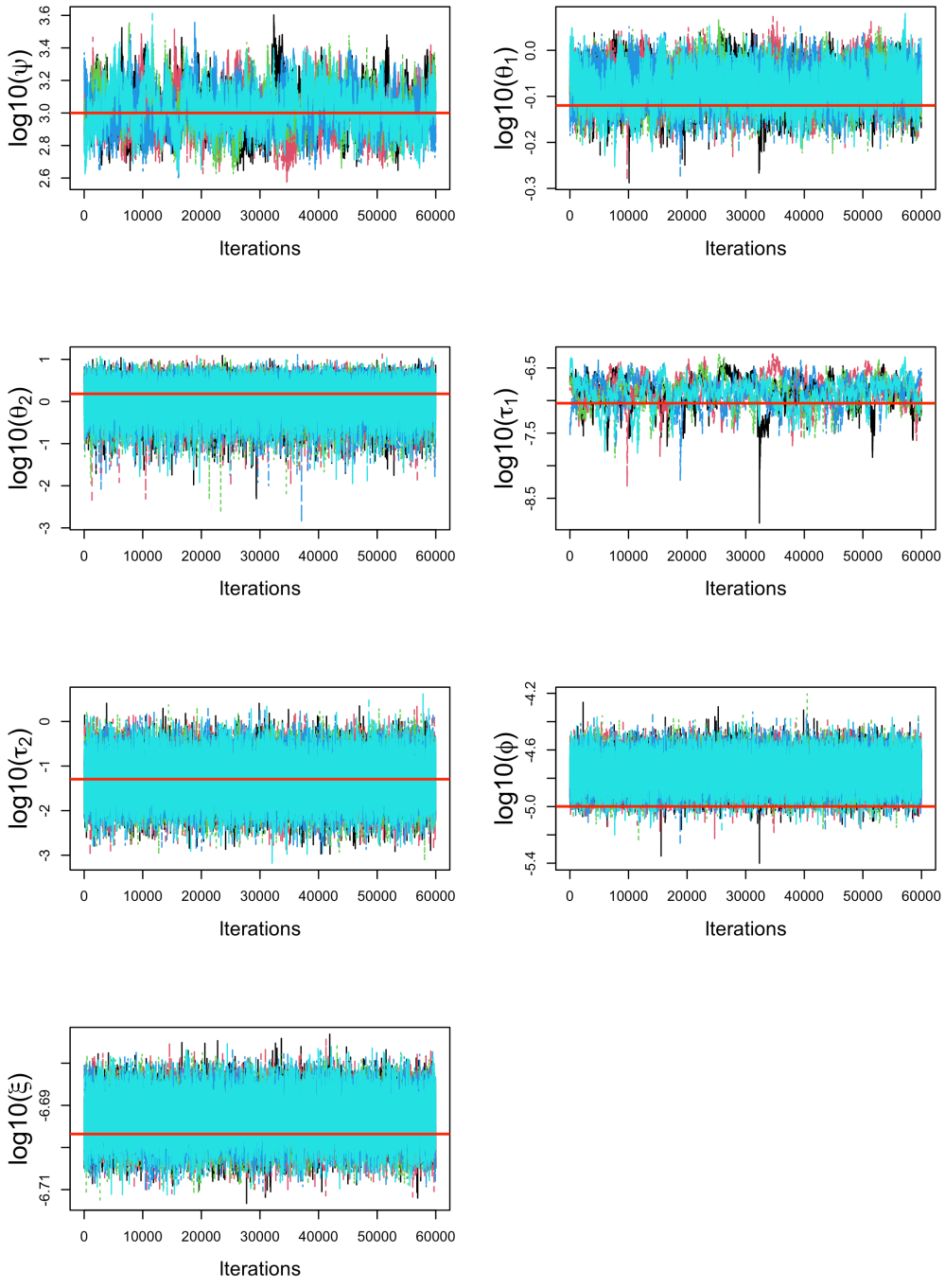


Figure C.1: Traceplots of $M_{\text{sim}} = 5$ chains of our Metropolis-within-Gibbs type sampler applied to the simulation study in Section 4.6.2. All chains seem converged to the stationary distribution, and no convergence issues are visually detected.

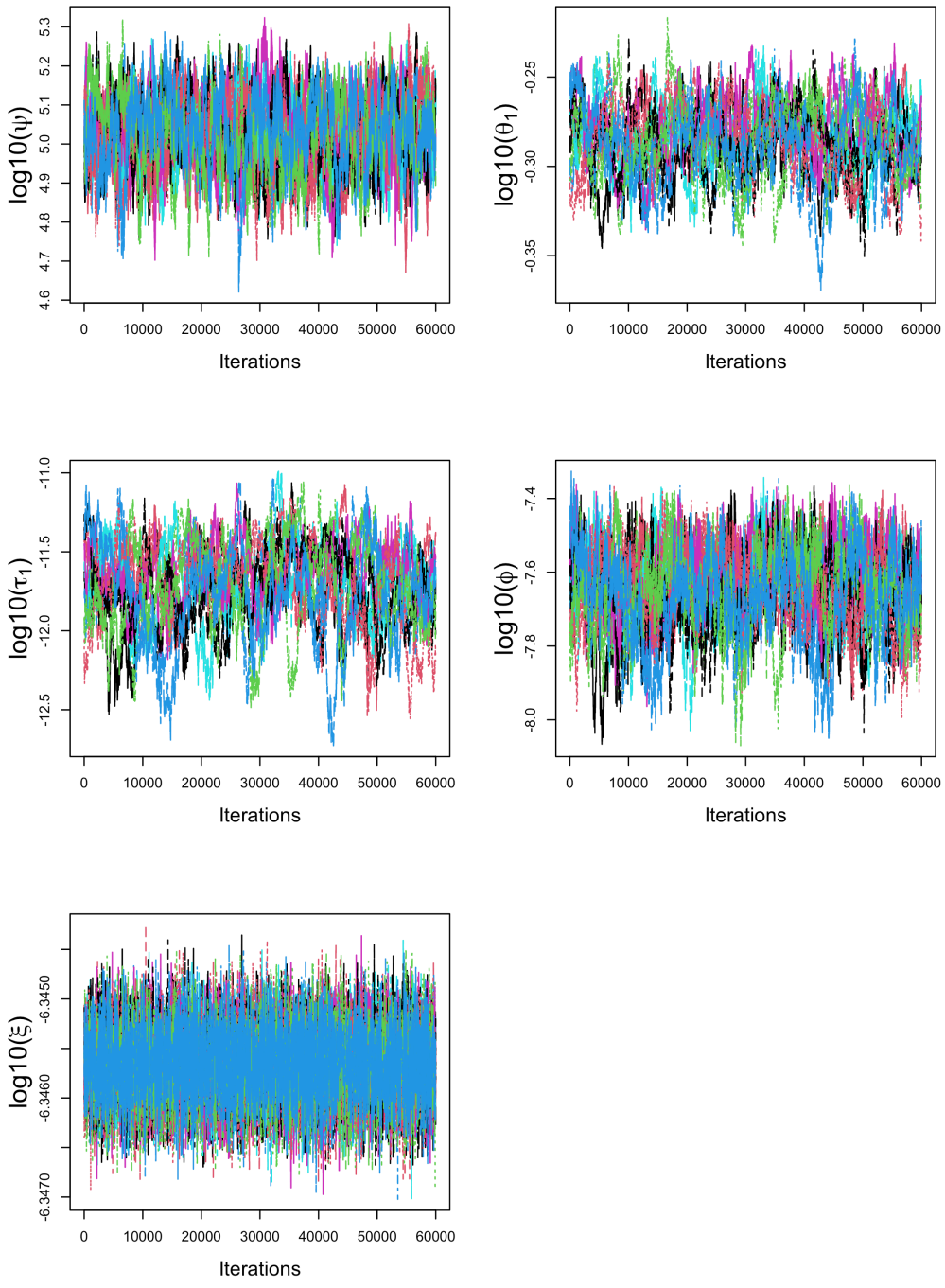


Figure C.2: Traceplots of $M_{\text{SP}} = 10$ chains of our Metropolis-within-Gibbs type sampler applied to the CDFS data in Section 4.7.3, under the single Pareto model. All chains seem converged to the stationary distribution, and no convergence issues are visually detected.

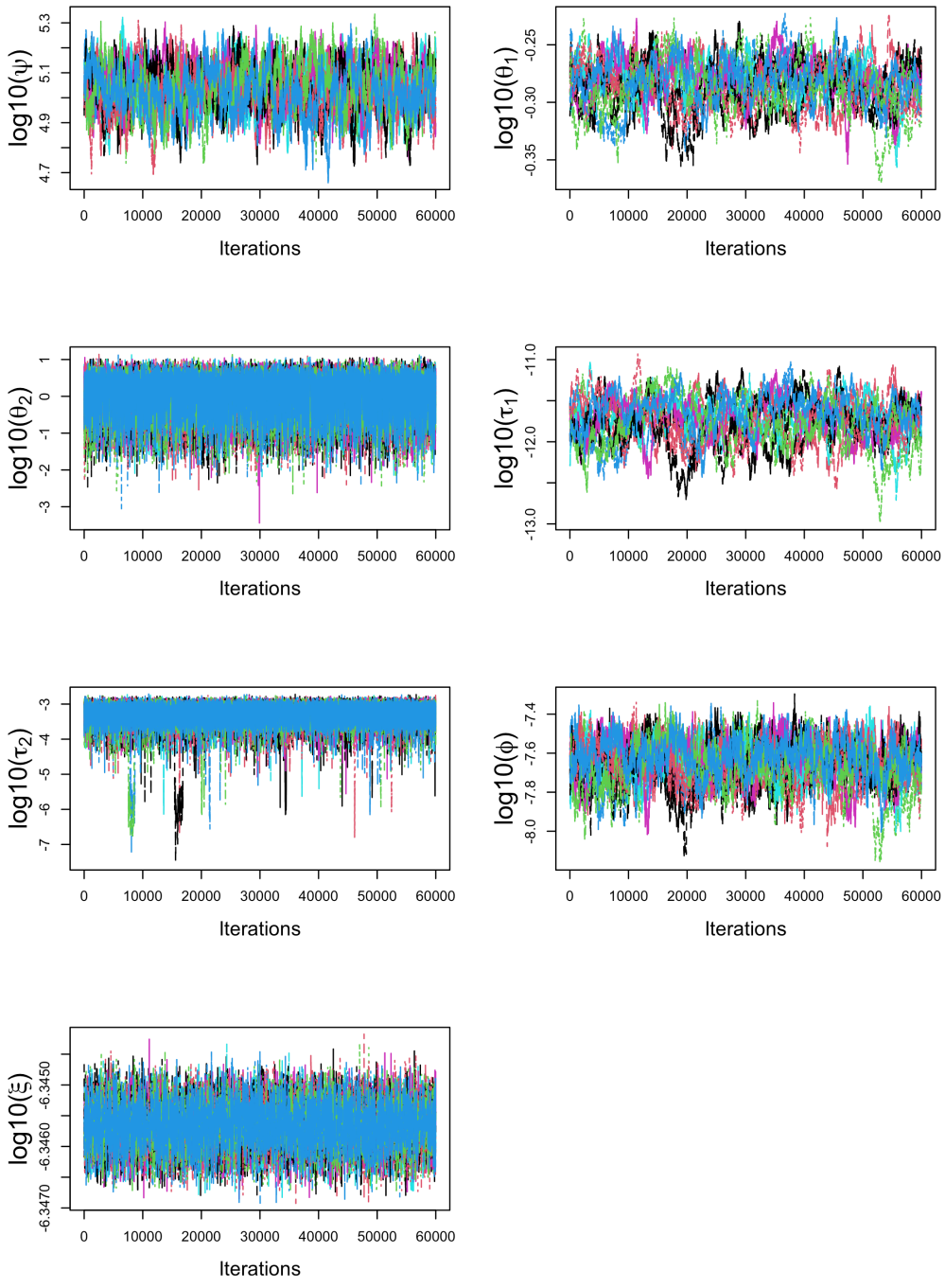


Figure C.3: Traceplots of $M_{\text{DP}} = 10$ chains of our Metropolis-within-Gibbs type sampler applied to the CDFS data in Section 4.7.4, under the double Pareto model. All chains seem converged to the stationary distribution, and no convergence issues are visually detected.

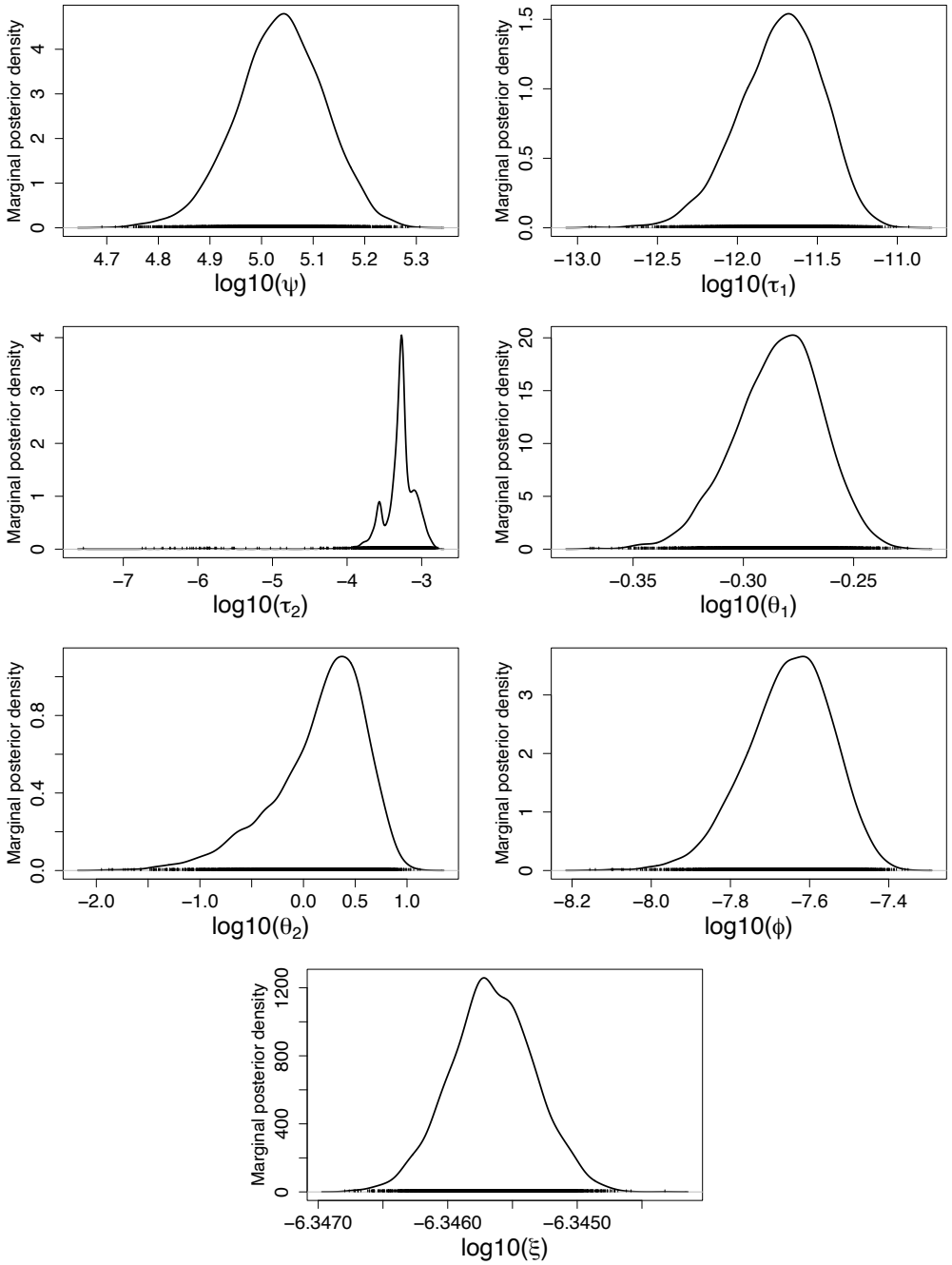


Figure C.4: Marginal posterior distributions for the main parameters in the CDFS analysis via the double Pareto model (Section 4.7.4), illustrated in $\log_{10}$-scale.

References

P. A. Abell, J. Allison, S. F. Anderson, J. R. Andrew, J. R. P. Angel, L. Armus, D. Arnett, S. Asztalos, T. S. Axelrod, S. Bailey, et al. Lsst science book, version 2.0. *arXiv preprint arXiv:0912.0201*, 2009.

T. Allam Jr, A. Bahmanyar, R. Biswas, M. Dai, L. Galbany, R. Hložek, E. E. Ishida, S. W. Jha, D. O. Jones, R. Kessler, et al. The photometric lsst astronomical time-series classification challenge (plasticc): Data set. *arXiv preprint arXiv:1810.00001*, 2018.

J. Alsing, A. Heavens, and A. H. Jaffe. Cosmological parameters, shear maps and power spectra from cfhtlens using bayesian hierarchical inference. *Monthly Notices of the Royal Astronomical Society*, 466(3):3272–3292, 2017.

A. Amara and A. Refregier. Systematic bias in cosmic shear: extending the fisher matrix. *Monthly Notices of the Royal Astronomical Society*, 391(1): 228–236, 2008.

S. Arnouts, B. Vandame, C. Benoist, M. Groenewegen, L. Da Costa, M. Schirmer, R. Mignani, R. Slijkhuis, E. Hatziminaoglou, R. Hook, et al. Eso imaging survey-deep public survey: Multi-color optical data for the chandra deep field south. *Astronomy & Astrophysics*, 379(2):740–754, 2001.

P. C. Austin. An introduction to propensity score methods for reducing the effects of confounding in observational studies. *Multivariate behavioral research*, 46(3):399–424, 2011.

P. C. Austin and E. A. Stuart. Moving towards best practice when using inverse probability of treatment weighting (iptw) using the propensity score

to estimate causal treatment effects in observational studies. *Statistics in medicine*, 34(28):3661–3679, 2015.

P. C. Austin and E. A. Stuart. The performance of inverse probability of treatment weighting and full matching on the propensity score in the presence of model misspecification when estimating the effect of treatment on survival outcomes. *Statistical methods in medical research*, 26(4):1654–1670, 2017.

P. C. Austin, P. Grootendorst, and G. M. Anderson. A comparison of the ability of different propensity score models to balance measured variables between treated and untreated subjects: a monte carlo study. *Statistics in medicine*, 26(4):734–753, 2007.

M. Autenrieth, R. A. Levine, J. Fan, M. A. Guarcello, et al. Stacked ensemble learning for propensity score methods in observational studies. *Journal of Educational Data Mining*, 13(1):24–189, 2021.

M. Autenrieth, D. A. van Dyk, R. Trotta, and D. C. Stenning. Stratified learning: A general-purpose statistical method for improved learning under covariate shift. *Stat.Anal. Data Min.: ASA Data Sci. J.*, 2023. URL https://doi.org/10.1002/sam.11643.

M. Baktashmotlagh, M. T. Harandi, B. C. Lovell, and M. Salzmann. Domain adaptation on the statistical manifold. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2481–2488, 2014.

J. Bekker, P. Robberechts, and J. Davis. Beyond the selected completely at random assumption for learning from positive and unlabeled data. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 71–85. Springer, 2019.

N. Benitez. Bayesian photometric redshift estimation. *The Astrophysical Journal*, 536(2):571, 2000.

J. Besag. Spatial interaction and the statistical analysis of lattice systems. *Journal of the Royal Statistical Society: Series B (Methodological)*, 36(2): 192–225, 1974.

G. Bhanot. The metropolis algorithm. *Reports on Progress in Physics*, 51 (3):429, 1988.

S. Bickel and T. Scheffer. Dirichlet-enhanced spam filtering based on biased samples. In *Advances in neural information processing systems*, pages 161–168, 2007.

S. Bickel, M. Brückner, and T. Scheffer. Discriminative learning under covariate shift. *Journal of Machine Learning Research*, 10(Sep):2137–2155, 2009.

K. Boone. Avocado: Photometric classification of astronomical transients with gaussian process augmentation. *The Astronomical Journal*, 158(6): 257, 2019.

L. Breiman. Random forests. *Machine learning*, 45:5–32, 2001.

S. Brooks, A. Gelman, G. Jones, and X.-L. Meng. *Handbook of markov chain monte carlo*. CRC press, 2011.

L. Bruzzone and M. Marconcini. Domain adaptation problems: A dasvm classification technique and a circular validation strategy. *IEEE transactions on pattern analysis and machine intelligence*, 32(5):770–787, 2009.

R. Buchs, C. Davis, D. Gruen, J. DeRose, A. Alarcon, G. Bernstein, C. Sánchez, J. Myles, A. Roodman, S. Allen, et al. Phenotypic redshifts with self-organizing maps: A novel method to characterize redshift distributions of source galaxies for weak lensing. *Monthly Notices of the Royal Astronomical Society*, 489(1):820–841, 2019.

T. Calders, A. Karim, F. Kamiran, W. Ali, and X. Zhang. Controlling attribute effect in linear regression. In *2013 IEEE 13th international conference on data mining*, pages 71–80. IEEE, 2013.

J. Carretero, F. Castander, E. Gaztañaga, M. Crocce, and P. Fosalba. An algorithm to build mock galaxy catalogues using mice simulations. *Monthly Notices of the Royal Astronomical Society*, 447(1):646–670, 2015.

A. Chan, A. Alaa, Z. Qian, and M. Van Der Schaar. Unlabelled data improves bayesian uncertainty calibration under covariate shift. In *International Conference on Machine Learning*, pages 1392–1402. PMLR, 2020.

X. Chen, M. Monfort, A. Liu, and B. D. Ziebart. Robust covariate shift regression. In *Artificial Intelligence and Statistics*, pages 1270–1279. PMLR, 2016.

W. G. Cochran. The effectiveness of adjustment by subclassification in removing bias in observational studies. *Biometrics*, pages 295–313, 1968.

R. R. Coifman and S. Lafon. Diffusion maps. *Applied and computational harmonic analysis*, 21(1):5–30, 2006.

C. Cortes, M. Mohri, M. Riley, and A. Rostamizadeh. Sample selection bias correction theory. In *International conference on algorithmic learning theory*, pages 38–53. Springer, 2008.

C. Cortes, Y. Mansour, and M. Mohri. Learning bounds for importance weighting. In *Advances in neural information processing systems*, pages 442–450, 2010.

P. Cortez, A. Cerdeira, F. Almeida, T. Matos, and J. Reis. Modeling wine preferences by data mining from physicochemical properties. *Decision support systems*, 47(4):547–553, 2009.

N. Courty, R. Flamary, D. Tuia, and A. Rakotomamonjy. Optimal transport for domain adaptation. *IEEE transactions on pattern analysis and machine intelligence*, 39(9):1853–1865, 2016.

N. Courty, R. Flamary, A. Habrard, and A. Rakotomamonjy. Joint distribution optimal transportation for domain adaptation. *arXiv preprint arXiv:1705.08848*, 2017.

D. F. Crawford, D. L. Jauncey, and H. S. Murdoch. Maximum-likelihood estimation of the slope from number-flux counts of radio sources. *Astrophysical Journal, vol. 162, p. 405*, 162:405, 1970.

M. Crocce, F. Castander, E. Gaztañaga, P. Fosalba, and J. Carretero. The mice grand challenge lightcone simulation–ii. halo and galaxy catalogues. *Monthly Notices of the Royal Astronomical Society*, 453(2):1513–1530, 2015.

G. Csurka. A comprehensive survey on domain adaptation for visual applications. *Domain adaptation in computer vision applications*, pages 1–35, 2017.

B. Dey, D. Zhao, J. A. Newman, B. H. Andrews, R. Izbicki, and A. B. Lee. Conditionally calibrated predictive distributions by probability-probability map: Application to galaxy redshift estimation and probabilistic forecasting, 2023.

P. Diaconis. Some things we've learned (about Markov chain Monte Carlo). *Bernoulli*, 19(4):1294 – 1305, 2013. doi: 10.3150/12-BEJSP09. URL https://doi.org/10.3150/12-BEJSP09.

D. Dua and C. Graff. UCI machine learning repository, 2017. URL http://archive.ics.uci.edu/ml.

B. Efron and C. Morris. Stein's estimation rule and its competitors—an empirical bayes approach. *Journal of the American Statistical Association*, 68(341):117–130, 1973.

B. Efron and C. Morris. Data analysis using stein's estimator and its generalizations. *Journal of the American Statistical Association*, 70(350):311–319, 1975.

B. Efron and C. Morris. Stein's paradox in statistics. *Scientific American*, 236(5):119–127, 1977.

B. Efron and R. J. Tibshirani. *An introduction to the bootstrap*. CRC press, 1994.

B. Fernando, A. Habrard, M. Sebban, and T. Tuytelaars. Unsupervised visual domain adaptation using subspace alignment. In *Proceedings of the IEEE international conference on computer vision*, pages 2960–2967, 2013.

P. Fosalba, M. Crocce, E. Gaztañaga, and F. Castander. The mice grand challenge lightcone simulation–i. dark matter clustering. *Monthly Notices of the Royal Astronomical Society*, 448(4):2987–3000, 2015.

J. M. Franklin, J. A. Rassen, D. Ackermann, D. B. Bartels, and S. Schneeweiss. Metrics for covariate balance in cohort studies of causal effects. *Statistics in medicine*, 33(10):1685–1699, 2014.

P. E. Freeman, V. Kashyap, R. Rosner, and D. Q. Lamb. A wavelet-based algorithm for the spatial analysis of poisson data. *The Astrophysical Journal Supplement Series*, 138(1):185, 2002.

P. E. Freeman, R. Izbicki, and A. B. Lee. A unified framework for constructing, tuning and assessing photometric redshift density estimates in a selection bias setting. *Monthly Notices of the Royal Astronomical Society*, 468(4):4556–4565, 2017.

A. Gelman and D. B. Rubin. Inference from iterative simulation using multiple sequences. *Statistical science*, 7(4):457–472, 1992.

A. Gelman, J. B. Carlin, H. S. Stern, and D. B. Rubin. *Bayesian data analysis*. Chapman and Hall/CRC, 1995.

S. Geman and D. Geman. Stochastic relaxation, gibbs distributions, and the bayesian restoration of images. *IEEE Transactions on pattern analysis and machine intelligence*, (6):721–741, 1984.

A. Georgakakis, K. Nandra, E. Laird, J. Aird, and M. Trichas. A new method for determining the sensitivity of x-ray imaging observations and the x-ray number counts. *Monthly Notices of the Royal Astronomical Society*, 388 (3):1205–1213, 2008.

F. Gieseke, K. L. Polsterer, A. Thom, P. Zinn, D. Bomanns, R.-J. Dettmar, O. Kramer, and J. Vahrenhold. Detecting quasars in large-scale astronom-

ical surveys. In *2010 Ninth International Conference on Machine Learning and Applications*, pages 352–357. IEEE, 2010.

W. R. Gilks, N. G. Best, and K. K. Tan. Adaptive rejection metropolis sampling within gibbs sampling. *Journal of the Royal Statistical Society Series C: Applied Statistics*, 44(4):455–472, 1995.

B. Gong, K. Grauman, and F. Sha. Reshaping visual datasets for domain adaptation. *Advances in Neural Information Processing Systems*, 26:1286–1294, 2013.

M. Gong, K. Zhang, T. Liu, D. Tao, C. Glymour, and B. Schölkopf. Domain adaptation with conditional transferable components. In *International conference on machine learning*, pages 2839–2848. PMLR, 2016.

P. J. Green. Reversible jump markov chain monte carlo computation and bayesian model determination. *Biometrika*, 82(4):711–732, 1995.

B. A. Griffin, D. F. McCaffrey, D. Almirall, L. F. Burgette, and C. M. Setodji. Chasing balance and other recommendations for improving nonparametric propensity score models. *Journal of causal inference*, 5(2), 2017.

N. Hassanpour and R. Greiner. Counterfactual regression with importance sampling weights. In *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, IJCAI-19*, pages 5880–5887, 2019.

T. Hastie, R. Tibshirani, J. H. Friedman, and J. H. Friedman. *The elements of statistical learning: data mining, inference, and prediction*, volume 2. Springer, 2009.

W. K. Hastings. Monte carlo sampling methods using markov chains and their applications. 1970.

B. D. He, C. M. De Sa, I. Mitliagkas, and C. Ré. Scan order in gibbs sampling: Models in which it matters and bounds on how much. *Advances in neural information processing systems*, 29, 2016.

J. J. Heckman. Sample selection bias as a specification error. *Econometrica: Journal of the econometric society*, pages 153–161, 1979.

C. Hikage, M. Oguri, T. Hamana, S. More, R. Mandelbaum, M. Takada, F. Köhlinger, H. Miyatake, A. J. Nishizawa, H. Aihara, et al. Cosmology from cosmic shear power spectra with subaru hyper suprime-cam first-year data. *Publications of the Astronomical Society of Japan*, 71(2):43, 2019.

H. Hildebrandt, S. Arnouts, P. Capak, L. Moustakas, C. Wolf, F. B. Abdalla, R. Assef, M. Banerji, N. Benítez, G. Brammer, et al. Phat: Photo-z accuracy testing. *Astronomy & Astrophysics*, 523:A31, 2010.

H. Hildebrandt, T. Erben, K. Kuijken, L. van Waerbeke, C. Heymans, J. Coupon, J. Benjamin, C. Bonnett, L. Fu, H. Hoekstra, et al. Cfhtlens: improving the quality of photometric redshifts with precision photometry. *Monthly Notices of the Royal Astronomical Society*, 421(3):2355–2367, 2012.

H. Hildebrandt, A. Choi, C. Heymans, C. Blake, T. Erben, L. Miller, R. Nakajima, L. van Waerbeke, M. Viola, A. Buddendiek, et al. Rcslens: the red cluster sequence lensing survey. *Monthly Notices of the Royal Astronomical Society*, 463(1):635–654, 2016.

H. Hildebrandt, F. Köhlinger, J. Van den Busch, B. Joachimi, C. Heymans, A. Kannawadi, A. Wright, M. Asgari, C. Blake, H. Hoekstra, et al. Kids+ viking-450: Cosmic shear tomography with optical and infrared data. *Astronomy & Astrophysics*, 633:A69, 2020.

K. Hirano, G. W. Imbens, and G. Ridder. Efficient estimation of average treatment effects using the estimated propensity score. *Econometrica*, 71 (4):1161–1189, 2003.

J. Hoffman, D. Wang, F. Yu, and T. Darrell. Fcns in the wild: Pixel-level adversarial and constraint-based adaptation. *arXiv preprint arXiv:1612.02649*, 2016.

K. Hoffmann, J. Bel, E. Gaztanaga, M. Crocce, P. Fosalba, and F. J. Castander. Measuring the growth of matter fluctuations with third-order galaxy correlations. *Monthly Notices of the Royal Astronomical Society*, 447(2):1724–1745, 2015.

B. Hoyle, D. Gruen, G. Bernstein, M. Rau, J. De Vicente, W. Hartley, E. Gaztanaga, J. DeRose, M. Troxel, C. Davis, et al. Dark energy survey year 1 results: redshift distributions of the weak-lensing source galaxies. *Monthly Notices of the Royal Astronomical Society*, 478(1):592–610, 2018.

J. Huang, A. Gretton, K. Borgwardt, B. Schölkopf, and A. J. Smola. Correcting sample selection bias by unlabeled data. In *Advances in neural information processing systems*, pages 601–608, 2007.

K. Imai and M. Ratkovic. Covariate balancing propensity score. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 76(1): 243–263, 2014.

K. Imai and D. A. van Dyk. Causal inference with general treatment regimes: Generalizing the propensity score. *Journal of the American Statistical Association*, 99(467):854–866, 2004.

G. W. Imbens and D. B. Rubin. *Causal inference in statistics, social, and biomedical sciences*. Cambridge University Press, 2015.

Ž. Ivezić, S. M. Kahn, J. A. Tyson, B. Abel, E. Acosta, R. Allsman, D. Alonso, Y. AlSayyad, S. F. Anderson, J. Andrew, and et al. LSST: From Science Drivers to Reference Design and Anticipated Data Products. , 873:111, Mar. 2019. doi: 10.3847/1538-4357/ab042c.

R. Izbicki and A. B. Lee. Nonparametric conditional density estimation in a high-dimensional regression setting. *Journal of Computational and Graphical Statistics*, 25(4):1297–1316, 2016.

R. Izbicki, A. Lee, and C. Schafer. High-dimensional density ratio estimation with extensions to approximate likelihood computation. In *Artificial Intelligence and Statistics*, pages 420–429. PMLR, 2014.

R. Izbicki, A. B. Lee, P. E. Freeman, et al. Photo-z estimation: An example of nonparametric conditional density estimation under selection bias. *The Annals of Applied Statistics*, 11(2):698–724, 2017.

G. James, D. Witten, T. Hastie, R. Tibshirani, et al. *An introduction to statistical learning*, volume 112. Springer, 2013.

W. James, C. Stein, and J. Neyman. Proceedings of the fourth berkeley symposium on mathematical statistics and probability. 1961.

J. Jiang and C. Zhai. Instance weighting for domain adaptation in nlp. ACL, 2007.

T. Joachims, A. Swaminathan, and T. Schnabel. Unbiased learning-to-rank with biased feedback. In *Proceedings of the Tenth ACM International Conference on Web Search and Data Mining*, pages 781–789. ACM, 2017.

D. M. Jones and A. F. Heavens. Bayesian photometric redshifts of blended sources. *Monthly Notices of the Royal Astronomical Society*, 483(2):2487–2505, 12 2018. ISSN 0035-8711. doi: 10.1093/mnras/sty3279. URL https://doi.org/10.1093/mnras/sty3279.

D. M. Jones and A. F. Heavens. Gaussian mixture models for blended photometric redshifts. *Monthly Notices of the Royal Astronomical Society*, 490(3):3966–3986, sep 2019. doi: 10.1093/mnras/stz2687. URL https://doi.org/10.1093%2Fmnras%2Fstz2687.

K. Kamnitsas, C. Baumgartner, C. Ledig, V. Newcombe, J. Simpson, A. Kane, D. Menon, A. Nori, A. Criminisi, D. Rueckert, et al. Unsupervised domain adaptation in brain lesion segmentation with adversarial networks. In *International conference on information processing in medical imaging*, pages 597–609. Springer, 2017.

T. Kanamori, S. Hido, and M. Sugiyama. A least-squares approach to direct importance estimation. *Journal of Machine Learning Research*, 10(Jul): 1391–1445, 2009.

T. Kanamori, T. Suzuki, and M. Sugiyama. Statistical analysis of kernel-based least-squares density-ratio estimation. *Machine Learning*, 86(3):335–367, 2012.

K. Karchev, R. Trotta, and C. Weniger. Sicret: Supernova ia cosmology with truncated marginal neural ratio estimation. *Monthly Notices of the Royal Astronomical Society*, 520(1):1056–1072, 2023.

R. Kessler, B. Bassett, P. Belov, V. Bhatnagar, H. Campbell, A. Conley, J. A. Frieman, A. Glazov, S. González-Gaitán, R. Hlozek, et al. Results from the supernova photometric classification challenge. *Publications of the Astronomical Society of the Pacific*, 122(898):1415, 2010a.

R. Kessler, A. Conley, S. Jha, and S. Kuhlmann. Supernova photometric classification challenge. *arXiv preprint arXiv:1001.5210*, 2010b.

R. Kessler, G. Narayan, A. Avelino, E. Bachelet, R. Biswas, P. Brown, D. Chernoff, A. Connolly, M. Dai, S. Daniel, et al. Models and simulations for the photometric lsst astronomical time series classification challenge (plasticc). *Publications of the Astronomical Society of the Pacific*, 131 (1003):094501, 2019.

T. Kohonen. Self-organized formation of topologically correct feature maps. *Biological cybernetics*, 43(1):59–69, 1982.

W. M. Kouw and M. Loog. A review of domain adaptation without target labels. *IEEE transactions on pattern analysis and machine intelligence*, 2019.

W. M. Kouw, L. J. Van Der Maaten, J. H. Krijthe, and M. Loog. Feature-level domain adaptation. *The Journal of Machine Learning Research*, 17 (1):5943–5974, 2016.

J. Kremer, F. Gieseke, K. S. Pedersen, and C. Igel. Nearest neighbor density ratio estimation for large-scale applications in astronomy. *Astronomy and Computing*, 12:67–72, 2015.

M. Kuhn. Building predictive models in r using the caret package. *Journal of Statistical Software, Articles*, 28(5):1–26, 2008. ISSN 1548-7660. doi: 10.18637/jss.v028.i05. URL https://www.jstatsoft.org/v028/i05.

B. K. Lee, J. Lessler, and E. A. Stuart. Improving propensity score weighting using machine learning. *Statistics in medicine*, 29(3):337–346, 2010.

B. Leistedt, D. J. Mortlock, and H. V. Peiris. Hierarchical bayesian inference of galaxy redshift distributions from photometric surveys. *Monthly Notices of the Royal Astronomical Society*, 460(4):4258–4267, 2016.

M. Lima, C. E. Cunha, H. Oyaizu, J. Frieman, H. Lin, and E. S. Sheldon. Estimating the redshift distribution of photometric galaxy samples. *Monthly Notices of the Royal Astronomical Society*, 390(1):118–130, 2008.

M. Little, P. McSharry, E. Hunter, J. Spielman, and L. Ramig. Suitability of dysphonia measurements for telemonitoring of parkinson's disease. *Nature Precedings*, pages 1–1, 2008.

R. J. Little and D. B. Rubin. *Statistical analysis with missing data*, volume 793. John Wiley & Sons, 2019.

A. Liu and B. D. Ziebart. Robust classification under sample selection bias. In *NIPS*, pages 37–45, 2014.

M. Lochner, J. D. McEwen, H. V. Peiris, O. Lahav, and M. K. Winter. Photometric supernova classification with machine learning. *The Astrophysical Journal Supplement Series*, 225(2):31, 2016.

M. Loog. Nearest neighbor-based importance weighting. In *2012 IEEE International Workshop on Machine Learning for Signal Processing*, pages 1–6. IEEE, 2012.

J. K. Lunceford and M. Davidian. Stratification and weighting via the propensity score in estimation of causal treatment effects: a comparative study. *Statistics in medicine*, 23(19):2937–2960, 2004.

T. Maccacaro, E. Feigelson, M. Fener, R. Giacconi, I. Gioia, R. Griffiths, S. Murray, G. Zamorani, J. Stocke, and J. Liebert. A medium sensitivity x-ray survey using the einstein observatory-the log n-log s relation for extragalactic x-ray sources. *Astrophysical Journal, Part 1, vol. 253, Feb. 15, 1982, p. 504-511.*, 253:504–511, 1982.

S. Magliacane, T. van Ommen, T. Claassen, S. Bongers, P. Versteeg, and J. M. Mooij. Domain adaptation by using causal inference to predict invariant conditional distributions. In *Advances in Neural Information Processing Systems*, pages 10846–10856, 2018.

G. K. Malmquist. A study of the stars of spectral type a. *Meddelanden fran Lunds Astronomiska Observatorium Serie II*, 22:3–69, 1920.

D. Mccaffrey, B. A. Griffin, D. Almirall, M. Slaughter, R. Ramchand, and L. F Burgette. A tutorial on propensity score estimation for multiple treatments using generalized boosted models. 32, 08 2013.

D. F. McCaffrey, G. Ridgeway, and A. R. Morral. Propensity score estimation with boosted regression for evaluating causal effects in observational studies. *Psychological methods*, 9(4):403, 2004.

M. McQuinn and M. White. On using angular cross-correlations to determine source redshift distributions. *Monthly Notices of the Royal Astronomical Society*, 433(4):2857–2883, 2013.

N. Metropolis and S. Ulam. The monte carlo method. *Journal of the American statistical association*, 44(247):335–341, 1949.

N. Metropolis, A. W. Rosenbluth, M. N. Rosenbluth, A. H. Teller, and E. Teller. Equation of state calculations by fast computing machines. *The journal of chemical physics*, 21(6):1087–1092, 1953.

M. Miville-Deschênes, V. Pettorino, M. Bucher, J. Delabrouille, K. Ganga, M. Le Jeune, G. Patanchon, C. Rosset, G. Roudier, Y. Fantaye, et al. Planck 2018 results: Vi. cosmological parameters. *Astronomy and Astrophysics*, 641:A6–A6, 2020.

A. Möller and T. de Boissière. Supernnova: an open-source framework for bayesian, neural network-based supernova classification. *Monthly Notices of the Royal Astronomical Society*, 491(3):4277–4293, 2020.

J. G. Moreno-Torres, T. Raeder, R. Alaiz-RodríGuez, N. V. Chawla, and F. Herrera. A unifying view on dataset shift in classification. *Pattern Recognition*, 45(1):521–530, 2012.

C. B. Morrison, H. Hildebrandt, S. J. Schmidt, I. K. Baldry, M. Bilicki, A. Choi, T. Erben, and P. Schneider. The-wizz: clustering redshift estimation for everyone. *Monthly Notices of the Royal Astronomical Society*, 467(3):3576–3589, 2017.

H. S. Murdoch, D. F. Crawford, and D. L. Jauncey. Maximum-likelihood estimation of the number-flux distribution of radio sources in the presence of noise and confusion. *Astrophysical Journal, Vol. 183, pp. 1-14 (1973)*, 183:1–14, 1973.

A. I. Naimi and L. B. Balzer. Stacked generalization: an introduction to super learning. *European journal of epidemiology*, 33:459–464, 2018.

J. A. Newman. Calibrating redshift distributions beyond spectroscopic limits with cross-correlations. *The Astrophysical Journal*, 684(1):88, 2008.

S. J. Pan and Q. Yang. A survey on transfer learning. *IEEE Transactions on knowledge and data engineering*, 22(10):1345–1359, 2009.

S. J. Pan, I. W. Tsang, J. T. Kwok, and Q. Yang. Domain adaptation via transfer component analysis. *IEEE Transactions on Neural Networks*, 22 (2):199–210, 2010.

L. Parast, D. F. McCaffrey, L. F. Burgette, F. H. de la Guardia, D. Golinelli, J. N. V. Miles, and B. A. Griffin. Optimizing variance-bias trade-off in the twang package for estimation of propensity scores. *Health Services and Outcomes Research Methodology*, 17(3):175–197, Dec 2017. ISSN 1572-9400. doi: 10.1007/s10742-016-0168-2. URL https://doi.org/10.1007/s10742-016-0168-2.

T. Park and D. A. Van Dyk. Partially collapsed gibbs samplers: Illustrations and applications. *Journal of Computational and Graphical Statistics*, 18 (2):283–305, 2009.

J. Pasquet, J. Pasquet, M. Chaumont, and D. Fouchez. Pelican: deep architecture for the light curve analysis. *Astronomy & Astrophysics*, 627:A21, 2019.

V. M. Patel, R. Gopalan, R. Li, and R. Chellappa. Visual domain adaptation: A survey of recent advances. *IEEE signal processing magazine*, 32(3):53–69, 2015.

S. Perlmutter, G. Aldering, G. Goldhaber, R. Knop, P. Nugent, P. Castro, S. Deustua, S. Fabbro, A. Goobar, D. Groom, et al. Measurements of ω and λ from 42 high-redshift supernovae. *The Astrophysical Journal*, 517 (2):565, 1999.

R. Pirracchio and M. Carone. The balance super learner: A robust adaptation of the super learner to improve estimation of the average treatment effect in the treated based on propensity score matching. *Statistical Methods in Medical Research*, 27(8):2504–2518, 2018. doi: 10.1177/0962280216682055. URL https://doi.org/10.1177/0962280216682055. PMID: 28339317.

R. Pirracchio, M. L. Petersen, and M. van der Laan. Improving propensity score estimators' robustness to model misspecification using super learner. *American journal of epidemiology*, 181(2):108–119, 2014.

N. Porqueres, A. Heavens, D. Mortlock, and G. Lavaux. Lifting weak lensing degeneracies with a field-based likelihood. *Monthly Notices of the Royal Astronomical Society*, 509(3):3194–3202, nov 2021. doi: 10.1093/mnras/stab3234. URL https://doi.org/10.1093%2Fmnras%2Fstab3234.

J. Quionero-Candela, M. Sugiyama, A. Schwaighofer, and N. D. Lawrence. *Dataset shift in machine learning*. The MIT Press, 2009.

S. Reddi, B. Poczos, and A. Smola. Doubly robust covariate shift correction. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 29, 2015.

E. A. Revsbech, R. Trotta, and D. A. van Dyk. Staccato: a novel solution to supernova photometric classification with biased training sets. *Monthly Notices of the Royal Astronomical Society*, 473(3):3969–3986, 2018.

J. W. Richards, D. Homrighausen, P. E. Freeman, C. M. Schafer, and D. Poznanski. Semi-supervised learning for photometric supernova classifica-

tion. *Monthly Notices of the Royal Astronomical Society*, 419(2):1121–1135, 2012.

A. G. Riess, A. V. Filippenko, P. Challis, A. Clocchiatti, A. Diercks, P. M. Garnavich, R. L. Gilliland, C. J. Hogan, S. Jha, R. P. Kirshner, et al. Observational evidence from supernovae for an accelerating universe and a cosmological constant. *The Astronomical Journal*, 116(3):1009, 1998.

W. A. Rivera, A. Goel, and J. P. Kincaid. Oups: a combined approach using smote and propensity score matching. In *2014 13th International Conference on Machine Learning and Applications*, pages 424–427. IEEE, 2014.

G. O. Roberts and S. K. Sahu. Updating schemes, correlation structure, blocking and parameterization for the gibbs sampler. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 59(2):291–317, 1997.

M. Rojas-Carulla, B. Schölkopf, R. Turner, and J. Peters. Invariant models for causal transfer learning. *The Journal of Machine Learning Research*, 19(1):1309–1342, 2018.

P. R. Rosenbaum and D. B. Rubin. The central role of the propensity score in observational studies for causal effects. *Biometrika*, 70(1):41–55, 1983.

P. R. Rosenbaum and D. B. Rubin. Reducing bias in observational studies using subclassification on the propensity score. *Journal of the American statistical Association*, 79(387):516–524, 1984.

D. B. Rubin. Using propensity scores to help design observational studies: application to the tobacco litigation. *Health Services and Outcomes Research Methodology*, 2:169–188, 2001.

D. B. Rubin. On principles for modeling propensity scores in medical research. *Pharmacoepidemiology and drug safety*, 13(12):855–857, 2004.

J. H. Schmitt and T. Maccacaro. Number-counts slope estimation in the presence of poisson noise. *Astrophysical Journal, Part 1 (ISSN 0004-637X), vol. 310, Nov. 1, 1986, p. 334-342.*, 310:334–342, 1986.

T. Schnabel, A. Swaminathan, A. Singh, N. Chandak, and T. Joachims. Recommendations as treatments: Debiasing learning and evaluation. In *international conference on machine learning*, pages 1670–1679. PMLR, 2016.

M. Schneider, L. Knox, H. Zhan, and A. Connolly. Using galaxy two-point correlation functions to determine the redshift distributions of galaxies binned by photometric redshift. *The Astrophysical Journal*, 651(1):14, 2006.

S. Setoguchi, S. Schneeweiss, M. A. Brookhart, R. J. Glynn, and E. F. Cook. Evaluating uses of data mining techniques in propensity score estimation: a simulation study. *Pharmacoepidemiology and drug safety*, 17(6):546–555, 2008.

H. Shariff, X. Jiao, R. Trotta, and D. A. van Dyk. Bahamas: new analysis of type ia supernovae reveals inconsistencies with standard cosmology. *The Astrophysical Journal*, 827(1):1, 2016.

E. S. Sheldon, C. E. Cunha, R. Mandelbaum, J. Brinkmann, and B. A. Weaver. Photometric redshift probability distributions for galaxies in the sdss dr8. *The Astrophysical Journal Supplement Series*, 201(2):32, 2012.

H. Shimodaira. Improving predictive inference under covariate shift by weighting the log-likelihood function. *Journal of statistical planning and inference*, 90(2):227–244, 2000.

V. Stampoulis. *Bayesian estimation of luminosity distributions and model based classification of astrophysical sources*. PhD thesis, Imperial College London, 2017.

D. Stenning, R. Wagner-Kaiser, E. Robinson, D. Van Dyk, T. Von Hippel, A. Sarajedini, and N. Stein. Bayesian analysis of two stellar populations in galactic globular clusters. i. statistical and computational methods. *The Astrophysical Journal*, 826(1):41, 2016.

D. C. Stenning and D. A. van Dyk. Bayesian statistical methods for astronomy part i: Foundations. *Statistics for Astrophysics: Bayesian Methodology*, pages 11–28, 2018a.

D. C. Stenning and D. A. van Dyk. Bayesian statistical methods for astronomy. part 2. markov chain monte carlo. *Statistics for Astrophysics: Bayesian Methodology*, pages 29–58, 2018b.

P. Stojanov, M. Gong, J. Carbonell, and K. Zhang. Low-dimensional density ratio estimation for covariate shift correction. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pages 3449–3458. PMLR, 2019.

C. M. Stonnington, G. Tan, S. Klöppel, C. Chu, B. Draganski, C. R. Jack Jr, K. Chen, J. Ashburner, and R. S. Frackowiak. Interpreting scan data acquired from multiple scanners: a study with alzheimer's disease. *Neuroimage*, 39(3):1180–1185, 2008.

E. A. Stuart, B. K. Lee, and F. P. Leacy. Prognostic score–based balance measures can be a useful diagnostic for propensity score methods in comparative effectiveness research. *Journal of clinical epidemiology*, 66(8): S84–S90, 2013.

M. Sugiyama and K.-R. Müller. Input-dependent estimation of generalization error under covariate shift. *Statistics & Risk Modeling*, 23(4):249–279, 2005.

M. Sugiyama, M. Krauledat, and K.-R. Müller. Covariate shift adaptation by importance weighted cross validation. *Journal of Machine Learning Research*, 8(May):985–1005, 2007.

M. Sugiyama, S. Nakajima, H. Kashima, P. V. Buenau, and M. Kawanabe. Direct importance estimation with model selection and its application to covariate shift adaptation. In *Advances in neural information processing systems*, pages 1433–1440, 2008.

M. Sugiyama, M. F. Oberwolfach, M. Kawanabe, and P. L. Chui. Dimensionality reduction for density ratio estimation in high-dimensional spaces. *Neural Networks*, 23(1):44–59, 2010.

M. Sugiyama, M. Yamada, P. Von Buenau, T. Suzuki, T. Kanamori, and M. Kawanabe. Direct density-ratio estimation with dimensionality reduction via least-squares hetero-distributional subspace search. *Neural Networks*, 24(2):183–198, 2011.

L. Tai, G. Paolo, and M. Liu. Virtual-to-real deep reinforcement learning: Continuous control of mobile robots for mapless navigation. In *2017 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 31–36. IEEE, 2017.

M. Tanaka, J. Coupon, B.-C. Hsieh, S. Mineo, A. J. Nishizawa, J. Speagle, H. Furusawa, S. Miyazaki, and H. Murayama. Photometric redshifts for hyper suprime-cam subaru strategic program data release 1. *Publications of the Astronomical Society of Japan*, 70(SP1):S9, 2018.

P. Teerikorpi. Influence of a generalized eddington bias on galaxy counts. *Astronomy & Astrophysics*, 424(1):73–78, 2004.

M. A. Troxel, N. MacCrann, J. Zuntz, T. Eifler, E. Krause, S. Dodelson, D. Gruen, J. Blazek, O. Friedrich, S. Samuroff, et al. Dark energy survey year 1 results: Cosmological constraints from cosmic shear. *Physical Review D*, 98(4):043528, 2018.

Y. Tsuboi, H. Kashima, S. Hido, S. Bickel, and M. Sugiyama. Direct density ratio estimation for large-scale covariate shift adaptation. *Journal of Information Processing*, 17:138–155, 2009.

I. S. Udaltsova. *The Universe at Your Fingertips: Bayesian Modeling and Computation in Problems of Observational Cosmology*. University of California, Davis, 2014.

J. Van Den Busch, H. Hildebrandt, A. Wright, C. Morrison, C. Blake, B. Joachimi, T. Erben, C. Heymans, K. Kuijken, and E. Taylor. Test-

ing kids cross-correlation redshifts with simulations. *Astronomy & Astrophysics*, 642:A200, 2020.

M. J. Van der Laan, E. C. Polley, and A. E. Hubbard. Super learner. *Statistical applications in genetics and molecular biology*, 6(1), 2007.

D. A. Van Dyk and X. Jiao. Metropolis-hastings within partially collapsed gibbs samplers. *Journal of Computational and Graphical Statistics*, 24(2): 301–327, 2015.

D. A. Van Dyk and T. Park. Partially collapsed gibbs samplers: Theory and methods. *Journal of the American Statistical Association*, 103(482): 790–796, 2008.

A. Van Opbroek, M. A. Ikram, M. W. Vernooij, and M. De Bruijne. Transfer learning improves supervised image segmentation across imaging protocols. *IEEE transactions on medical imaging*, 34(5):1018–1030, 2014.

R. Vilalta, K. D. Gupta, and L. Macri. A machine learning approach to cepheid variable star classification using data alignment and maximum likelihood. *Astronomy and Computing*, 2:46–53, 2013.

M. Weisskopf, B. Brinkman, C. Canizares, G. Garmire, S. Murray, and L. Van Speybroeck. An overview of the performance and scientific results from the chandra x-ray observatory. *Publications of the Astronomical Society of the Pacific*, 114(791):1, 2002.

J. Wen, C.-N. Yu, and R. Greiner. Robust learning under uncertain test distributions: Relating covariate shift to model misspecification. In *International Conference on Machine Learning*, pages 631–639. PMLR, 2014.

J. Wen, R. Greiner, and D. Schuurmans. Correcting covariate shift with the frank-wolfe algorithm. In *Twenty-Fourth International Joint Conference on Artificial Intelligence*, 2015.

D. Westreich, S. R. Cole, M. J. Funk, M. A. Brookhart, and T. Stürmer. The role of the c-statistic in variable selection for propensity score models. *Pharmacoepidemiology and drug safety*, 20(3):317–320, 2011.

E. Williamson, R. Morley, A. Lucas, and J. Carpenter. Propensity scores: from naive enthusiasm to intuitive understanding. *Statistical methods in medical research*, 21(3):273–293, 2012.

D. H. Wolpert. Stacked generalization. *Neural networks*, 5(2):241–259, 1992.

R. K. Wong, P. Baines, A. Aue, T. C. Lee, and V. L. Kashyap. Automatic estimation of flux distributions of astrophysical source populations. 2014.

A. H. Wright, H. Hildebrandt, K. Kuijken, T. Erben, R. Blake, H. Buddelmeijer, A. Choi, N. Cross, J. T. de Jong, A. Edge, et al. Kids+ viking-450: A new combined optical and near-infrared dataset for cosmology and astrophysics. *Astronomy & Astrophysics*, 632:A34, 2019.

A. H. Wright, H. Hildebrandt, J. L. Van den Busch, and C. Heymans. Photometric redshift calibration with self-organising maps. *Astronomy & Astrophysics*, 637:A100, 2020.

N. J. Wright, J. J. Drake, M. G. Guarcello, T. L. Aldcroft, V. L. Kashyap, F. Damiani, J. DePasquale, and A. Fruscione. The chandra cygnus ob2 legacy survey: Design and x-ray point source catalog, 2014.

N. J. Wright, J. J. Drake, M. G. Guarcello, V. L. Kashyap, and A. Zezas. Simulating the sensitivity to stellar point sources of chandra x-ray observations. *arXiv preprint arXiv:1511.03943*, 2015.

J. Yoon, Y. Zhang, J. Jordon, and M. van der Schaar. Vime: Extending the success of self-and semi-supervised learning to tabular domain. *Advances in Neural Information Processing Systems*, 33, 2020.

Y. Yu and C. Szepesvári. Analysis of kernel mean matching under covariate shift. *arXiv preprint arXiv:1206.4650*, 2012.

Y. Yu, D. A. van Dyk, V. L. Kashyap, and C. A. Young. A bayesian analysis of the correlations among sunspot cycles. *Solar Physics*, 281:847–862, 2012.

B. Zadrozny. Learning and evaluating classifiers under sample selection bias. In *Proceedings of the twenty-first international conference on Machine learning*, page 114. ACM, 2004.

A. Zezas, G. Fabbiano, A. Baldi, F. Schweizer, A. King, A. Rots, and T. Ponman. Chandra monitoring observations of the antennae galaxies. ii. x-ray luminosity functions. *The Astrophysical Journal*, 661(1):135, 2007.