UNIVERSITY OF CALIFORNIA,
IRVINE

Methods for Optimal Covariate Balance in Observational Studies for Causal Inference

DISSERTATION

submitted in partial satisfaction of the requirements
for the degree of

DOCTOR OF PHILOSOPHY

in Statistics

by

Brian Garrett Vegetabile

DISSERTATION Committee:
Professor Hal S. Stern, Chair
Professor Daniel L. Gillen
Assistant Professor Weining Shen

2018

DEDICATION

To my parents, Bob & Patti

TABLE OF CONTENTS

Page

LIST OF FIGURES **vi**

LIST OF TABLES **x**

ACKNOWLEDGMENTS **xiv**

CURRICULUM VITAE **xvi**

ABSTRACT OF THE DISSERTATION **xviii**

1 Introduction **1**

2 Background **5**

2.1 Overview of Causal Inference in Observational Studies 5

2.1.1 A Brief History of Causal Inference and the Potential Outcome Framework 6

2.1.2 The Potential Outcome Framework and Causal Estimands 10

2.1.3 Nonparametric Weighting Estimation of Treatment Effects 14

2.1.4 Methods for Estimating Propensity Scores and Balancing Scores . . . 22

2.1.5 Evaluating Covariate Balance by Weighted Distributional Measures . 24

2.2 Overview of Regression with Gaussian Processes 31

2.2.1 Binary Regression with Gaussian Processes 31

2.2.2 Kernel Functions 41

2.2.3 Examples - Binary Regression with Gaussian Processes 50

2.2.4 Multinomial Regression with Gaussian Processes 55

3 Optimally Balanced Gaussian Process Propensity Scores **59**

3.1 Introduction 59

3.2 Methodology 61

3.2.1 Estimation of Binary Propensity Scores using Gaussian Processes . . 61

3.2.2 Measuring Covariate Imbalance 64

3.2.3 Minimizing Covariate Imbalance 66

3.3 Simulations 67

3.3.1 Potential Outcomes Settings 68

3.3.2 Laplace Approximation vs. Expectation Propagation . . . 68
3.3.3 Comparative ATE Simulation Study . . . 73
3.3.4 Comparative ATT Simulation Studies . . . 81
3.4 Application to the Job Training Data of Dehejia and Wahba (1999) . . . 88
3.5 Extension to Multiple Treatment Settings . . . 93
3.6 Discussion . . . 95

4 Targeted Optimal Balancing Weights 99
4.1 Introduction . . . 99
4.2 Methodology . . . 102
4.2.1 Fixed Target Covariate Imbalance . . . 103
4.2.2 Implementation . . . 108
4.3 Simulations . . . 111
4.3.1 ATE Performance - A Univariate Example . . . 112
4.3.2 ATT Performance - A Multivariate Example . . . 119
4.3.3 ATE Performance - A Multi-Treatment Example . . . 123
4.4 Application to the Job Training Data of Dehejia and Wahba (1999) . . . 126
4.5 Discussion . . . 129

5 A Comparison Between the Methods 132
5.1 Overview . . . 133
5.2 ATE Simulation Setting of 3.3.3 . . . 133
5.3 ATT Simulation Setting of 4.3.2 . . . 138
5.4 Multi-Treatment Setting of Section 4.3.3 . . . 140
5.4.1 Comparisons of Runtimes . . . 143

6 A Discussion on Future Directions 146

Bibliography 152

A Derivations for Probit Regression with Gaussian Processes 159
A.1 Laplace Approximation . . . 159
A.1.1 Derivatives of the log Posterior Distribution . . . 160
A.1.2 Newton's Method for Finding the Posterior Mode . . . 163
A.1.3 Approximate Marginal Likelihood . . . 165
A.2 Expectation propagation . . . 167
A.2.1 Marginal Cavity Distribution . . . 168
A.2.2 Marginal Tilted Distribution . . . 169
A.2.3 Updating the Site Parameters . . . 169
A.2.4 Approximate Log Marginal Likelihood . . . 172
A.2.5 A Note on Implementation of Expectation Propagation . . . 173

B Measuring Predictability of Behavior - Entropy Rate Estimation 174
B.1 Overview of the Problem . . . 174
B.2 Modeling Behavior . . . 177
B.2.1 Modeling Behavior Using Finite-State Markov Chains . . . 177
B.2.2 Entropy Rate of a Finite Markov Chain . . . 179
B.2.3 Entropy Rate from Lempel-Ziv Compression . . . 181
B.3 Estimating the Entropy Rate . . . 183
B.3.1 Direct Estimation of Entropy Rate . . . 183
B.3.2 Sliding Window Lempel-Ziv Entropy Rate Estimation . . . 185
B.3.3 Estimation of Standard Errors via the Stationary Bootstrap . . . 186
B.4 Simulations . . . 188
B.4.1 Estimation of First-Order Markov Processes . . . 188
B.4.2 Estimating the Standard Error via the Stationary Bootstrap . . . 194
B.4.3 Estimation When m is Misspecified . . . 196
B.5 Applications . . . 200
B.5.1 Measuring Predictability of Behavior in a Rodent Study . . . 202
B.5.2 Measuring Predictability of Behavior in a Human Developmental Study 205
B.6 Discussion . . . 208

LIST OF FIGURES

Page

1.1 Relationship between "cause" and "effect" 1

2.1 Visualizing specific subpopulations of interest that have been generated using functions of the propensity score. The density for each figure was created such that $g(x) = f(x)h(x)$ with $h(x)$ defined as in Table 2.1 and the propensity score is defined as $e(x) = \Phi(x)$. The dotted line in each figure represents the density of standard normal random variable. The lower right figure in the graphic represents the relationship among the subpopulations of interest . . . 13

2.2 Visualization of the relative weights in each group under weighting schemes from Table 2.2 (Note: ATC is not shown due to its similarity to the ATT weighting). The size of each dot was created by setting the `R` plotting parameter `cex` to be the weight of the observation. The first plot shows each observation with weight one. The first row demonstrates how the ATE and ATT weighting schemes increase the weights of those who received the opposite treatment than probability would suggest. The lower panel demonstrates how the ATO, Truncated, and Matching weights all tend to focus on the areas where there is non-zero probability of being in either group. 20

2.3 Finite-sample empirical distributions of average treatment effects in subpopulations defined in Table 2.2. 21

2.4 Comparing the empirical distribution of the estimated treatment effects demonstrated in Figure 2.3 against the empirical distribution of the average treatment effects estimated using only the observed outcomes weighted to a specific subpopulation of interest 22

2.5 Visualizing the cumulative distribution function before and after weighting by the propensity score using ATE weights. The first row of the figure is the treatment assignment used to simulate the data. Notice that the figure in the upper right hand corner represents a case where there are significant positivity violations for a covariate $X \sim \mathcal{N}(0, 1)$. The second row demonstrates the lack of balance resulting from each assignment mechanism prior to propensity score weighting. The last row demonstrates the ability of the propensity score to balance covariates. Notice that with large positivity violations, it is difficult for even the true propensity score to balance covariates. 30

2.6 Visualization of the properties of the squared exponential covariance function. Left panel: Demonstration of the correlation function induced by the squared exponential kernel with inverse-length scale $\rho = 1$. For the squared exponential kernel the inverse-length scale parameterization implies that higher values correspond to more "local" functions, while values closer to zero demonstrate smoother functions. Right panel: Visualized contour plot of the correlation structure when $\rho = 1$. Demonstrates the "structure" in this covariance is completely local and therefore the model has a difficult time modeling long-range dependencies. 43
2.7 Visualization of the properties of the linear covariance function. Left panel: Demonstration of the correlation between a point x and all other points on the evaluated grid. Right panel: Visualized contour plot of the correlation structure when $\sigma_0 = 0.5$. Demonstrates that the "structure" in this covariance allows for longer range dependencies. 44
2.8 Upper Row: Draws from Gaussian processes for a grid of a univariate x, using kernel functions which are a squared exponential with $\rho = 1$ (Left) and a linear polynomial kernel with $\sigma_0 = 0.5$ (Right). Lower panel: Each simulated draw from the upper panel transformed through a probit link function. 45
2.9 Visualization of the properties of additive (upper row) and multiplicative (lower row) kernel functions. Left panel: Demonstration of the correlation between a point x and all other points on the evaluated grid. Right panel: Visualized contour plot of the correlation matrix. 48
2.10 Upper panel: Simulated draws from a Gaussian process with additive (left) and multiplicative (right) kernel functions and univariate x. Lower panel: Simulated draws transformed through the probit link function. 49
2.11 Left Panel: Sample draw of observed probabilities using a predictor that was drawn from a Gaussian process with a second-order polynomial kernel. Right Panel: Observed labels simulated based upon the observed probabilities in the left panel. 50
2.12 Comparing the point-wise 95% credible intervals among the methods. The left panel compares the credible intervals based on the MCMC samples and the Laplace approximation (LA) to the posterior distribution. The dotted lines represent the LA and the solid dark area is the MCMC credible intervals. The right panel is similar but comparing MCMC against expectation propagation. 52
2.13 Comparing the posterior distribution from MCMC sampling to the Laplace and expectation propagation approximations. The posterior distribution appears to be skewed, while both approximation methods utilize a Gaussian approximation. Note that the expectation propagation approximation places most its mass on the "long-tail", while the Laplace approximation is centered closer to the mode. 53

2.14 Left panel: Log-marginal likelihood curves based upon each approximation method. Both methods attain their maxima at approximately the same point. Right panel: Posterior expectations of the probability function for each point compared among the methods. 54

3.1 Visualization of Potential Outcome Response Functions in the Treated and Control Groups. The left figure is a setting where the potential outcomes are linearly related to the continuous covariate X_1; the center is where the potential outcomes are related to the continuous covariate, but the treatment effect is not constant; the furthest right is a more extreme example of non-linear treatment effects. 69
3.2 Comparing mean balance metrics for the standardized difference in the means. Each distribution represents the absolute value of the standardized difference in the means based upon ATE weights and each adjustment method. 70
3.3 Comparing the simulation treatment effect, i.e. $\tau_{sim} = \frac{1}{N}\sum_{i=1}^{N}\left(Y_i^{Z=1} - Y_i^{Z=0}\right)$, against estimates based upon weighting using the optimally balanced Gaussian process propensity score estimates using EP and LA. Each row is a different potential outcome setting. Overall there is agreement between the weighted estimated treatment effects and τ_{sim}. 71
3.4 Comparing runtime between the LA and EP approximations across 1000 simulations for estimating propensity scores based on 500 observations. Results were collected on a 2013 27" iMac with a 3.4 GHz Intel Quad-Core i5 processor and 16 GB of memory. 73
3.5 Visualization of propensity score simulation settings. Parameter values are described in Table 3.3. 74
3.6 Visualization of covariate distributions for estimating the ATT. 81

4.1 Visualization of propensity score functions considered for estimating the ATE; each dotted line represents a different propensity score function listed in the legend. The density of the covariate is overlaid in the graphic to demonstrate where the values of the propensity score are relevant. 113
4.2 Conditional distribution of the covariate given treatment assignment, demonstrating covariate imbalances after treatment assignment. Let $\phi(x)$ be the probability density function of standard normal random variable, each conditional density is proportional to $\phi(x) * \Pr(Z = z|X = x)$ which is visualized above. The solid line is the original distribution of the covariate. 113
4.3 Visualization of the sample mean of various metrics across 100 simulations as a function of the effective sample size penalty parameter λ_z. The upper panel of figures contains covariate imbalance measures (similar to those in Table 4.2) and the effective sample size in each group. The lower panel is the absolute bias, empirical standard error, and empirical mean squared error as a function of λ_z across the potential outcome settings. Setting λ to approximately one appears to provide adequate performance across the potential outcome settings. 119

5.1 Comparison of runtime and data set size relationship between the optimally balanced Gaussian process propensity score and the targeted optimal balancing weights methodology. Each point within each line is based upon estimating the results across 10 instances, each based on a similar data set. 144
5.2 Comparison of runtime between the optimally balanced Gaussian process propensity score and the targeted optimal balancing weights methodology, based on a similar data set where the timing results have been averaged across 10 instances. 145

LIST OF TABLES

Page

2.1 Subpopulations as defined in Li et al. (2017) Table 1 13
2.2 Subpopulations as defined in Li et al. (2017) Table 1 and the corresponding weights to estimate treatment effects for the specific subpopulation of interest. 19
2.3 Simulation results demonstrating covariate balance measures for a simple binary treatment assignment mechanism where $Pr(Z = 1|X = x) = \Phi(\alpha x)$. Each entry is the average measure across 1000 simulated data sets, each containing 500 observations. 29
2.4 Table of common kernel functions for $x, x' \in \Re^d$ (unless otherwise specified). In the above table $d(x, x')$ is defined to be $d(x, x') = ||x - x'|| = \sqrt{(x - x')^T (x - x')}$. In many kernels in the above the parameter ℓ is referred to as the length-scale of the kernel, while an alternative parameterization is typically the *inverse-length scale* defined to be $\rho = 1/\ell$. 46

3.1 Models for potential outcomes 68
3.2 Simulation results comparing the relative effectiveness between EP and LA approximations for estimating propensity score weights used to estimate treatment effects. The columns represent the mean bias, mean absolute bias, empirical standard error of the estimates, and finally the mean squared error of the estimates. Each grouped section of rows contains the results for one of the potential outcome settings in Table 3.1, Within each group section, the results based upon no weighting adjustment, adjustment using weights based on the true propensity score, and lastly weights based upon estimates using EP and LA are compared. 72
3.3 Parameter settings used to create propensity score functions 74
3.4 Models to be compared for estimating the ATE under various simulation settings and the R package used to estimate them. The first two methods (no adjustment and adjustment by the true propensity score) provide baseline performance measures. The next six methods are nonparametric methods for estimating the propensity score. The package `gpbalancer` can be found at `https://github.com/bvegetabile/gpbalancer`. The last four rows are parametric methods for estimating the propensity score. Note that all parametric models are misspecified in the sense that they are missing terms to handle the fact that $\alpha_1 \neq 1$ and $\alpha_2 \neq 0$ in the data-generating model. . . 76

3.5 Simulation results for demonstrating the performance in balancing covariates. The columns provide the proportion of the 1000 simulations which were declared to be mean balanced, i.e., $|\Delta_d| < \delta$ for all d and for various thresholds of δ 77
3.6 Simulation results for estimating the ATE in the non-GLM setting where the true propensity score was linear and included interaction terms 78
3.7 Simulation results for estimating the ATE in the non-GLM setting where the true propensity score was a second-order polynomial. 79
3.8 Models to be compared for estimating the ATT under various simulation settings and the R package used to estimate them. 82
3.9 Simulation results for demonstrating the performance in balancing covariates. The columns provide the proportion of the 1000 simulations which were declared to be mean balanced, i.e., $|\Delta_d| < \delta$ for all d and for various thresholds of δ 83
3.10 Simulation results for estimating the ATT grouped vertically by potential outcome setting. 84
3.11 Models for potential outcomes. The bold portion of the equation in the potential outcome in the treated group is what differs between the two potential outcomes 86
3.12 Simulation results for demonstrating the performance in balancing covariates. The columns provide the proportion of the 1000 simulations which were declared to be mean balanced, i.e., $|\Delta_d| < \delta$ for all d and for various thresholds of δ 87
3.13 Simulation results for estimating the ATT grouped vertically by potential outcome setting. 88
3.14 Summaries of experimental data from Dehejia and Wahba (1999). 90
3.15 Summaries of observational data from Dehejia and Wahba (1999). The first grouping are the data from the Panel Study on Income Dynamics (PSID) and the second group are data from the Current Population Survey-Social Security File (CPS). Comparing with Table 3.14, there are clearly discrepancies between the experimental and observational data. 91
3.16 Estimates of the ATT and summaries of covariate imbalance using data from Dehejia and Wahba (1999). The upper table of values describe covariate imbalance prior to adjustment for each data set and the middle set of values are the covariate imbalance after adjustment. The last three rows are estimates of the ATT. The second to last row are estimates of the ATT from Dehejia and Wahba (1999) (Table 3, column 8). The final row are estimates of the ATT using weighted least squares and the optimally balanced Gaussian process propensity score. 92
3.17 Comparisons of computational runtime across the methods considered in Section 3. The runtimes are averaged across 5 simulated data sets using the data-generating procedure of Section 3.3.3 where the true propensity score was a linear function with interaction terms. 97

4.1 Models for Potential Outcomes 112
4.2 Relative performance of each method for balancing covariates and maintaining an acceptable effective sample size across increasingly unbalanced propensity score settings. 117
4.3 Comparison of performance of methods for the ATE by propensity score model and outcome setting. 118
4.4 Models for Potential Outcomes. The bold portion of the equation in the potential outcome model for the treated group is what differs between the two potential outcome models. 121
4.5 Comparison of average covariate imbalance after weighting adjustment across 1000 simulations. Covariate imbalance in the above table is defined using Equation (4.12) with the designated number of moments defined by the column headers, $\lambda_z = 0$ for all z, and μ_d set to be the sample means in the treated subgroup to define appropriate targets. 122
4.6 Comparison of performance of methods for estimating the ATT by potential outcome setting. 123
4.7 Comparison of performance of methods for the average response in the population in a multi-treatment setting 125
4.8 Summaries of experimental and observational data from Dehejia and Wahba (1999). The first row is the number of observations from that data source and the next rows provide summaries of the mean and standard deviations for that specific category. 127
4.9 Summaries of covariate imbalance measures within each data set: the weighted standardized difference in the means and the logarithm of the ratio of the sample variances. The upper portion of the table are the metrics before weighting and the lower portion of the table are metrics after weighting using the targeted optimal balancing weights. 129
4.10 Estimates of the ATT for the Dehejia and Wahba (1999) data. The first row represents the experimental results of the National Supported Work Demonstration. The next rows are estimates using data from the observational source, the Current Population Survey-Social Security File (CPS). The first grouping of columns demonstrates estimates prior to adjustment. The next grouping of columns are the results from Dehejia and Wahba (1999) (Table 3, column 8) which were estimates of the ATT using weighted least squares where "treatment observations weighted as 1, and control observations weighted by the number of times they are matched to a treatment observation". The next grouping are the previous results from Chapter 3. The final group of columns are estimates using weights estimated through the targeted optimal balancing weights. The standard errors for the OBGPPS and OPTBW methods were found using a robust sandwich variance estimator. 130

5.1 Parameter settings used to create propensity score functions 134
5.2 Models for potential outcomes 134

5.3 Comparison in balancing covariates between the Optimally Balanced Gaussian Process Propensity Score methodology and the Targeted Optimal Balancing Weight methodology under the scenario of Section 3.3.3. Simulation results demonstrate the performance in balancing covariates. The columns provide the proportion of the 1000 simulations which were declared to be mean balanced, i.e., $|\Delta_d| < \delta$ for all d and for various thresholds of δ 135
5.4 Simulation results for estimating the ATE in the non-GLM setting where the true propensity score was linear and included interaction terms 136
5.5 Simulation results for estimating the ATE in the non-GLM setting where the true propensity score was a second-order polynomial. 137
5.6 Models for potential outcomes. The bold portion of the equation in the potential outcome model for the treated group is what differs between the two potential outcome models. 139
5.7 Comparison of average covariate imbalance after weighting adjustment across 1000 simulations. Covariate imbalance in the above table is defined using Equation (4.12) with the designated number of moments defined by the column headers, $\lambda_z = 0$ for all z, and μ_d set to be the sample means in the treated subgroup to define appropriate targets. 139
5.8 Performance estimating the ATT by potential outcome setting. 140
5.9 Performance estimating the average response in the population in a multi-treatment setting 142

ACKNOWLEDGMENTS

First, I thank my advisor, Hal Stern. It has been a honor working with him these years. He reached out to me four years ago to discuss working together and that conversation opened an entire world of research I was unaware of prior to this journey in life. As my career progresses, I will continue to look to him for guidance and friendship and I am sure he will remain an integral part of who I am as a statistician.

I also express my thanks to Dan Gillen. In my five years at UC Irvine, he has provided invaluable insight on everything from job prospects to music recommendations. His support and guidance are a few of the reasons I was able to make it through the program. I thank him for all he has taught me, but more importantly I thank him for his friendship.

I also thank my dissertation committee comprised of Hal Stern, Dan Gillen and Weining Shen and my advancement committee that further included Padhraic Smyth and Tallie Z. Baram. The hours spent reading and providing feedback on this dissertation helped to form it into the product it is today. I thank the researchers of the Conte Center at UC Irvine. Statistics is an interdisciplinary endeavor and I am grateful for the hard work and care the researchers have placed into that organization. I thank the many people of the Department of Statistics at UC Irvine and the professors that formed my foundational understanding of statistics. Without Professors Wes Johnson, Babak Shahbaba, Yaming Yu, among others, pushing my limits of work and understanding, this dissertation would not have been possible. My friend Alexander Vandenberg-Rodes for his instruction, friendship, and guidance during our time together at UCI; much of my knowledge of analysis and Markov chains is credited to him. I thank Michele Guindani and Jessica Utts for each always having an open door. Finally, I also thank Rosemary Busta and Lisa Stieler, without whom nothing would get accomplished on the second floor of Donald Bren Hall. This dissertation would not have been successful without the support of my family and my friends. My parents, Bob and Patti, my brother Jake, and my sister Lea; knowing that they support me in my endeavors, has made all of my adventures easier to begin. Also my friends Brett Bowen, Ross McDonald, and Chris Gonzalez who provided me with unparalleled friendship and a support structure throughout my graduate career. They provided much needed escapes from numbers and academics. I thank my friends at UC Irvine, Eric Nalisnick and Andrew Holbrook, who helped form many of my opinions of all things statistics and machine learning. I further thank my friends I met through the Department of Statistics, specifically Maricela Cruz, David Stenning, Lars Hertel, Chris Galbraith, Marija Pejcinovska, Garren Gaut, and the many others who often joined us at the pub.

Chapters 3 and 4 of this dissertation are comprised of work from two manuscripts, *Optimally Balanced Gaussian Process Propensity Scores for Estimating Treatment Effects* and *Targeted Optimal Balancing Weights*, both co-authored by Hal Stern and Dan Gillen. Appendix B is related to the manuscript *Estimating the Entropy Rate of Finite Markov Chains with Application to Behavior Studies* and I acknowledge my collaborators and co-authors who

contributed data and revisions on various versions of that manuscript: Stephanie Stout-Oswald, Jenny Molet, Elysia Davis, Tallie Z. Baram, and Hal Stern.

Research reported in this dissertation was supported by NIMH Silvio O. Conte Centers of the National Institutes of Health under award number P50 MH 096889. The content is solely the responsibility of the author and does not necessarily represent the official views of the National Institutes of Health.

CURRICULUM VITAE

Brian Garrett Vegetabile

EDUCATION

Doctor of Philosophy in Statistics **2013 – 2018**
University of California, Irvine *Irvine, California*

Master of Science in Statistics **2013 – 2015**
University of California, Irvine *Irvine, California*

Bachelor of Science in Aerospace Engineering **2005 – 2009**
Pennsylvania State University *University Park, Pennsylvania*

RESEARCH EXPERIENCE

Graduate Research Assistant **2014–2018**
Conte Center @ UC Irvine *Irvine, California*

TEACHING EXPERIENCE

Statistics Bootcamp for Incoming Graduate Students **2016 – 2017**
Department of Statistics *University of California, Irvine*

Teaching Assistant **2013 – 2014**
Department of Statistics *University of California, Irvine*

PROFESSIONAL EXPERIENCE

Statistical Analyst Intern **2014 – 2014**
Northrop Grumman – *Chandra X-Ray Observatory* *Cambridge, Massachusetts*

Satellite Systems Engineer, Mission Planning **2011 – 2013**
Northrop Grumman – *Chandra X-Ray Observatory* *Cambridge, Massachusetts*

Systems Engineer, Autonomy & Fault Management **2009 – 2011**
Northrop Grumman – *Defense Weather Satellite System* *Redondo Beach, California*

Systems Engineer Intern, Autonomy & Fault Management **2008 – 2008**
Northrop Grumman – *Defense Weather Satellite System* *Redondo Beach, California*

REFEREED JOURNAL PUBLICATIONS

Davis, E.P., Stout, S.A., Molet, J., **Vegetabile, B.G.**, Glynn L.M., Sandman, C.A., Heins, K., Stern, H., Baram, T.Z. (2017) **Exposure to unpredictable maternal sensory signals influences cognitive development across species** *Proceedings of the National Academy of Sciences* 114(39), pp.10390-10395.

MANUSCRIPTS

Vegetabile, B.G., Gillen, D.L., Stern, H.S., (2018) **Targeted Optimal Balancing Weights**

Vegetabile, B.G., Gillen, D.L., Stern, H.S., (2018) **Optimally Balanced Gaussian Process Propensity Scores for Estimating Treatment Effects**

Vegetabile, B.G., Stout-Oswald, S.A., Davis, E.P., Baram, T.Z., Stern, H.S., (2018) **Estimating the Entropy Rate of Finite Markov Chains with Application to Behavior Studies**

SOFTWARE

`optbw` `https://github.com/bvegetabile/optbw`
R software for estimation of balancing weights for use in estimating causal effects

`gpbalancer` `https://github.com/bvegetabile/gpbalancer`
R software for propensity score estimation using Gaussian processes where hyperparameters are selected to minimize metrics of covariate imbalance

`ccber` `https://github.com/bvegetabile/ccber`
R software for estimating the entropy rate of finite-state Markov processes based upon video-tagged data generated in the Conte Center at UC Irvine

ABSTRACT OF THE DISSERTATION

Methods for Optimal Covariate Balance in Observational Studies for Causal Inference

By

Brian Garrett Vegetabile

Doctor of Philosophy in Statistics

University of California, Irvine, 2018

Professor Hal S. Stern, Chair

The most basic approach to causal inference measures the response of a system or population to different exposures, or treatments, and compares one or more summaries of the responses. Thinking formally about causal inference requires that we consider what the *potential outcomes* would have been under a set of alternative exposures. Though it is only possible in practice to observe a single outcome for each unit, the principles of good experimental design, including random treatment assignment, allow a valid comparison of average responses to each exposure because effects from extraneous factors are minimized. In observational settings, where exposures or treatments arise "naturally", i.e., without experimental manipulation, a common strategy for estimating causal effects is to find units that are similar based upon a set of covariates, but receiving different exposures, and then compare their outcomes. This strategy is challenging if there are many covariates. Balancing scores, a low-dimensional summary of the relevant covariate space, can facilitate causal inference for observational data in settings with many covariates. Propensity scores which measure the probability of receiving a particular exposure or treatment are one example of a balancing score. To estimate treatment effects, balancing scores are used to to group individuals from different exposure groups to compare their response levels, or functions of balancing scores are used to re-weight the sample. This thesis explores novel methods for obtaining covariate balance in observational

studies through the use of balancing scores and weighting methodology.

The dissertation begins by providing an overview of the potential outcome framework for causal inference in observational studies and required background knowledge for the methods developed. The first methodological contribution is the optimally balanced Gaussian process propensity score approach that applies a binary regression framework using Gaussian processes for estimating the propensity score. The hyperparameters of the process are selected to minimize a metric of covariate imbalance. The next methodological contribution is the development of targeted balancing weights for both binary and multi-treatment settings, where a covariate imbalance metric is created with respect to a covariate density of interest (this could be the distribution within the full population under study or within a specific subpopulation of interest) and unit weights are selected that minimize this metric, without an explicit assumption on the functional form of the weights. Each method is evaluated against competing methods from the causal inference literature through series of simulations and against a benchmark causal inference data set. The dissertation concludes with suggestions for future work. A contribution to measurement in observational studies is included as an appendix.

Chapter 1

Introduction

The advancement of knowledge and discovery through science has largely been concerned with understanding relationships between so-called "causes" and their "effects". Simplistically, this is often conveyed by creating a diagram, as in Figure 1.1, where "causes" are related to "effects" through the use of arrows (Greenland et al., 1999).

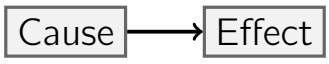

Figure 1.1: Relationship between "cause" and "effect"

There are significant distinctions related to the direction one chooses to pursue in investigating these relationships. Specifically, one may be interested in measuring the forward-direction "effects of causes" or the reverse-direction "causes of effects" (Dawid et al., 2014, 2016). Settings where the goal is to isolate "causes of effects" are largely considered to be problems of *attribution*. That is, some phenomena has been observed, and measured, and the goal is to isolate the cause, or causes, of what has been observed. Inference in this setting is hard due to the fact that there are often multiple plausible explanations (causes) for what has been observed.

The setting that may be more familiar to most, and the subject of this dissertation, is isolating the "effects of causes". In this setting, the goal is to identify a factor that is believed to be a "cause" and then to manipulate this factor into different configurations and measure the response, while holding other factors constant. There is a (potentially) measurable response in the world to each setting of this factor and "causal effects" are measured as some comparison among the responses to different configurations. This framework for manipulation and exploration is what has motivated the development of methodology in the statistics literature for assessing causal relationships, such as the "Potential Outcome Framework" of Neyman and Rubin (Splawa-Neyman et al., 1990; Rubin, 1974). This framework has been successfully used for identifying causal relationships within both experimental and observational data and is explored throughout this dissertation.

A key challenge in studying the "effects of causes" is that we can generally only apply one exposure or treatment and therefore we will only observe one of the many potential outcomes. This is due to the fact that the simple act of manipulating a factor in these systems may physically change, or alter, the system and we will no longer be able to make future observations on this system without imposing further assumptions. This motivates the use of experimentation on *samples* of units, or individuals, within some population where many different units are assigned to the various treatment levels and the average response is assessed in this population of units. The "gold standard" in these experiments is the randomization of the assignment to different configurations, or treatments, which minimizes the effect due to other factors. This simple act of randomizing assignments to configurations has been the foundation for much of causal inquiry in biological settings.

In observational studies, where the assignment to treatment is not under the control of the experimenters, there are added difficulties that must be taken into consideration. We still require that "causes" are manipulable, but they will have been "naturally" manipulated,

or assigned. A primary benefit of control over treatment assignments is that through a randomization we are assured that in very large samples a response should be observed for each level of exposure across all levels of a covariate space and further that the joint distribution of the covariates conditional on the observed treatment assignment should be similar. In observational settings we do not have this guarantee and there may be some systematic process by which units are deterministically assigned to a certain exposure levels, based upon a specific covariate value. This implies that we are never be able to observe the response for these individuals under alternative exposure levels. What we can hope for in observational settings, is that we find individuals who appear similar based upon some criteria at each level of treatment exposure so that we may compare their outcomes. We then have to make further defensible assumptions that allow us to strengthen causal claims within these settings, but ultimately we will be at the mercy of many assumptions. Our view of observational studies is through the lens of well designed experimental studies where we must learn the assignment mechanism that is being used to assign treatments.

The focus of this thesis is on understanding and developing statistical methods that allow us to compare individuals through a re-weighting of observations in samples within observational settings. The work has largely been motivated through questions arising from studies of the Conte Center on Brain Programming in Adolescent Vulnerabilities at the University in California, Irvine, a center of translational science that addresses how early-life experiences influence brain development and contribute to vulnerability to mental illnesses during adolescence and adulthood. The center is composed of various human and rodent studies to probe these questions, where in the animal studies experimental manipulation is possible and in the human studies we are restricted to observational data.

Chapters 2 through 6 of this thesis explore methods for causal inference in observational settings. Section 2.1 of Chapter 2 explores the required background information related to

the potential outcome framework, defines the assumptions required for causal inference under this setting, and culminates in defining balancing scores and the propensity score, as well as methods for evaluating them. Section 2.2 of Chapter 2 provides the requisite background for Gaussian processes and provides an overview of Bayesian categorical regression (binary and multinomial) using Gaussian processes as prior distributions on a space of functions. These sections lay the groundwork for Chapter 3, which develops a nonparametric propensity score estimation procedure within a Gaussian process regression framework. Chapter 4 develops a method for estimating balancing weights in observational settings without any assumption on the functional form of the weights and through the use of optimization techniques. Chapter 5 provides a direct comparison between the methods of Chapters 3 and 4. Chapter 6 provides a concluding discussion and suggestions for future work on various methods.

Another contribution to observational studies that is unrelated to covariate balance for causal inference is included as appendix. Appendix B provides the development of a method of measurement of predictability of maternal behavior in an observational setting. The Conte Center has focused on investigating the effect of maternal predictability and fragmentation on the cognitive and emotional outcomes of offspring during adolescence and a critical prerequisite for the Center's work are methods for quantifying predictability and fragmentation across species. This need motivated the development of a measure based on the entropy rate of a sequence for recorded maternal behaviors. This appendix outlines several methods for estimating the entropy rate of an observed sequence and compares their performance in estimating this quantity in small samples. Additionally, a bootstrap procedure is provided for estimating the variance of these estimates. As the methodology is not directly related to covariate balance in observational studies, this appendix may be read without reading the background material first.

Chapter 2

Background

2.1 Overview of Causal Inference in Observational Studies

As stated in the Introduction, we are primarily interested in assessing the "effects of causes" within observational settings. Imbens and Rubin (2015) define a randomized experiment as a study where the assignment to treatment is probabilistic and the process, or function, that is being used for this randomization is known to researchers. Extending this definition, they define an observational study as a study where the functional form of the probabilistic assignment mechanism is unknown to the researchers. In observational settings, where the randomization process is unknown to researchers, a simple idea to perform inference is to find individuals who appear similar based on a set of covariates, yet received different values of the exposure, and compare their results to assess the relationship between the "cause" and the "effect".

This "simple idea" begets many inferential challenges though. Under what conditions is the comparison causal in nature? What assumptions are necessary to perform causal inference?

How might we construct metrics for determining who within a population is similar enough so that they may be fairly compared? This section provides an overview of causal inference in observational studies, emphasizing the Potential Outcome Framework and the use of balancing scores to compare individuals. We provide a brief history of causal inference and then introduce the framework for inference that is used throughout much of this dissertation.

2.1.1 A Brief History of Causal Inference and the Potential Outcome Framework

There is a long history of causal inquiry throughout both philosophy and science. This section outlines a thread of history similar to the historical discussion of Chapter 2 of Imbens and Rubin (2015), while expanding and providing this author's views within the context of this dissertation. For a longer history, see Imbens and Rubin (2015).

The idea of a "potential outcome" was originally proposed by Jerzey Neyman in his 1923 paper *On the Application of Probability Theory to Agricultural Experiments* (Splawa-Neyman et al., 1990). The paper, published originally in Polish and translated to English in 1990, describes a setting of a field experiment to compare the yield of different crop varieties. In this setting, he describes the fact that only one variety can be grown on any one plot and discusses issues related to estimating the variance of the difference in the mean yield for two different varieties under a sampling scheme where varieties are chosen without replacement. This work provided the notation for the potential outcome framework that is common throughout much of the present causal literature, but only discussed randomization in the context of sampling from an urn without replacement. The idea of physical randomization of treatment exposures for causal effects was proposed two years later by R.A. Fisher in *Statistical Methods for Research Workers* (Fisher, 1925), which became the basis for much of experimental research.

From the 1950's through the 1970's there was an emergence of interest in understanding causal effects in observational and epidemiological settings, particularly surrounding the debate about the observed association between smoking and lung cancer. In 1965, Sir Austin Bradford Hill gave an address titled "The Environment and Disease: Association or Causation" (Hill, 1965) where he described a set of criteria for assessing causal relationships in observational data motivated by the following question regarding the relationship between *association* and *causation*,

> *There are, ..., instances in which we can reasonably answer these questions from the general body of medical knowledge. A particular, and perhaps extreme, physical environment cannot fail to be harmful; a particular chemical is known to be toxic to man and therefore suspect on the factory floor. Sometimes, alternatively, we may be able to consider what* **might** *a particular environment do to man, and then see whether such consequences are indeed to be found. But more often than not we have no such guidance, no such means of proceeding; more often than not we are dependent upon our observation and enumeration of defined events for which we then seek antecedents. In other words we see that the event B is associated with the environmental feature A, that, to take a specific example, some form of respiratory illness is associated with a dust in the environment. In what circumstances can we pass from this observed* **association** *to a verdict of* **causation***? Upon what basis should we proceed to do so?*
>
> *Sir Austin Bradford Hill*
>
> *January 14, 1965*

The set of nine criteria he laid out in his address titled, "The Environment and Disease: Association or Causation", for assessing the relationship between two phenomenon were summarized follows:

(1) Strength of the association between the events,

(2) Consistency, or repeatability, of the phenomenon,

(3) Specificity of the relationship to particular subgroups of individuals who share a common characteristic,

(4) Temporal relationship and assessing the consistency of the order of events,

(5) Biological gradient, or does the relationship exhibit some "dose-response" relationship,

(6) Plausibility given our knowledge of the events, often through biological knowledge,

(7) Coherence with the scientifically derived relationships within the world,

(8) Ability of the relationship to be supported through experimentation, or some natural experiment in the world,

(9) Analogy and support from other recognized relationships.

These criteria outline guidance for reasoning about the relationships among observed phenomenon and much of the current work in observational studies applies these concepts.

Within the statistics literature in the 1970's there was also work investigating methods in observational studies. In 1973, William Cochran and Donald Rubin (Cochran and Rubin, 1973) investigated the source of bias in observational studies. In 1974, Rubin reintroduced the concept of "potential outcomes" and the Potential Outcome Framework, in both experimental and observational studies arguing that randomization is necessary to remove bias due to unmeasured variables in observational studies (Rubin, 1974). In observational studies, he argued for what he called "subjective randomization", which he defined as being *"...a study in which there are no obviously important prior variables that systematically differ in the (exposure) trials and the (control) trials"*[1]. He then states *"Under this assumption of subjective*

[1] In the original text, E replaces exposure and C replaces control

randomization, the usual estimates and significance levels can be used as if the study had been randomized; this procedure is analogous to assuming subjective random sampling in order to make inferences about a target population. Until an obviously important variable is found that systematically differs in the E and C trials, the belief in subjective randomization is well founded.".

Another significant contribution within observational studies was the formalization of the idea of systematic randomization into an assumption titled **strong ignorability** given a set of covariates (Rosenbaum and Rubin, 1983). Rosenbaum and Rubin (1983) also define a commonly employed tool in observational studies, denoted the balancing score, of which the most impactful has been the propensity score. This paper provided a framework for conducting inference in the potential outcome framework, where if this formalized idea of systematic randomization holds, then causal effects can be estimated through not just an adjustment on a set of covariates, but through adjustment on a balancing score that often greatly reduces the dimensionality of the problem. Since this paper, there has been significant research in balancing scores and propensity scores and their application in many different settings (see Holland (1986); Dehejia and Wahba (1999); Imbens (2000); Stuart (2010); Li et al. (2017), among others).

The primary thread throughout this history has been understanding under what conditions it is possible to perform causal inference in observational settings. Sir Austin Bradford Hill's criteria grounds our scientific inquiries in observational settings and encourage us to be suspect of findings until we can further rationalize these phenomenon in the context of contemporary and historical evidence, while the potential outcome framework provides us with a mathematical justification for addressing observational studies. The next sections further investigate, in detail, the mathematical formulations of the potential outcome framework, causal estimands, weighting estimators of causal effects, and methods for assessing the assumptions within this

framework.

2.1.2 The Potential Outcome Framework and Causal Estimands

The potential outcome framework attempts to quantify the effect on an outcome Y of an assigned exposure, or factor Z, assumed to take on a finite number of values, when the factor is applied to a specific unit i. Without loss of generality, the factor Z can be assigned to any value z from an index set of possible exposures, $\{0, 1, \ldots, E\}$ (Note: there may exist a one-to-one mapping from this index set to a set of relevant treatment values that have clinical or domain significance). The framework assumes that for each unit under study there is an observable response under each exposure, i.e., there exists a $Y_i^{Z=z}$ for each z. The response vector $\mathbf{Y}_i^Z$ collects these potential outcomes,

$$\mathbf{Y}_i^Z = \begin{pmatrix} Y_i^{Z=0} \\ Y_i^{Z=1} \\ \vdots \\ Y_i^{Z=E} \end{pmatrix}. \tag{2.1}$$

Individual-level treatment effects may be defined as functions of the entries of this vector of potential outcomes, i.e., some function $\psi(\mathbf{Y}_i^Z)$. For example, consider a binary treatment indicator $Z \in \{0, 1\}$. The potential outcomes for such a treatment are $\mathbf{Y}_i^Z = \begin{pmatrix} Y_i^{Z=0} \\ Y_i^{Z=1} \end{pmatrix}$ and a common treatment effect is the simple difference between the two potential outcomes, $\psi(\mathbf{Y}_i^Z) = Y_i^{Z=1} - Y_i^{Z=0}$. An alternative treatment effect for a binary treatment regime is a ratio of the two potential outcomes, $\psi(\mathbf{Y}_i^Z) = \frac{Y_i^{Z=1}}{Y_i^{Z=0}}$. If we consider a factor with three treatment levels $Z \in \{0, 1, 2\}$ and potential outcomes $\mathbf{Y}_i^Z = \begin{pmatrix} Y_i^{Z=0} & Y_i^{Z=1} & Y_i^{Z=2} \end{pmatrix}^T$ an

example treatment effect is the vector of all pairwise differences, that is,

$$\psi(\mathbf{Y}_i^Z) = \mathbf{A}\mathbf{Y}_i^Z = \begin{pmatrix} -1 & 1 & 0 \\ -1 & 0 & 1 \\ 0 & -1 & 1 \end{pmatrix} \begin{pmatrix} Y_i^{Z=0} \\ Y_i^{Z=1} \\ Y_i^{Z=2} \end{pmatrix} = \begin{pmatrix} Y_i^{Z=1} - Y_i^{Z=0} \\ Y_i^{Z=2} - Y_i^{Z=0} \\ Y_i^{Z=2} - Y_i^{Z=1} \end{pmatrix}.$$

The function defining the treatment effect should be chosen with care to reflect the causal question of interest within the scientific setting.

Unfortunately, we are never be able to observe the full vector $\mathbf{Y}_i^Z$ for an individual. That is we can only observe one of the potential outcomes under the *exact* same conditions (time, environment, etc). This has been referred to as the *Fundamental Problem of Causal Inference* by Holland (1986) and was the motivation for Rubin's comparisons between causal inference and missing data problems (Rubin, 1974, 1978; Imbens and Rubin, 2015). The actual value observed in a study is denoted Y_i^{obs} and is defined as,

$$Y_i^{obs} = \sum_{z=0}^{E} I(Z_i = z) Y_i^{Z=z}. \tag{2.2}$$

where $I(Z_i = z)$ is an indicator function that evaluates to one if unit i received treatment z and zero otherwise. In many observational settings, such as those attempting to quantify the effects of policy implementations, the interest may not be on the individual-level treatment effects but may instead focus on the treatment effect across a population, or within specific subpopulations. In such a setting, the goal then is to understand under what conditions inference involving samples from the population can be used to create estimators that are unbiased (or asymptotically consistent) for effects defined through the potential outcome framework.

To compare individuals who are similar, we assume that a vector of measured covariates $X = x \in \Re^d$ is sufficient for describing the characteristics of individuals. A summary measure

that captures the treatment effects across individuals who are similar is the conditional average treatment effect (CATE),

$$CATE \equiv \tau(x) = E(\psi(\mathbf{Y}^Z) \mid X = x), \tag{2.3}$$

which summarizes the average treatment effect for individuals who share common values of the covariates $X = x$. Generally we are interested in the treatment effect across a population of interest. If we let $f(x)$ represent the density function for the covariates in the population of interest, then the overall average treatment effect (ATE) is obtained by marginalizing over the values of X in this population,

$$ATE_{f(x)} = E_{f(x)}\left[\tau(x)\right]. \tag{2.4}$$

If we let $g(x)$ represent the density of a specific subpopulation of X, then the treatment effect across this subpopulation could similarly be calculated as,

$$ATE_{g(x)} = E_{g(x)}\left[\tau(x)\right]. \tag{2.5}$$

Li et al. (2017) provide a general framework for defining treatment effects in subpopulations where they define a selection function $h(x)$ to be some function of the covariates and then define $g(x)$ to be the density function proportional to the function arising from the multiplication of f and h, i.e., $g(x) \propto f(x)h(x)$. In the binary treatment setting, many of the interesting $h(x)$ are functions of $e(x) = Pr(Z = 1|X = x)$, commonly referred to as the propensity score (Rosenbaum and Rubin, 1983). Functions of the propensity score define estimands that are common in the literature such as the average treatment effect in the treated subpopulation (ATT), the average treatment effect in the control subpopulation (ATC), or the average treatment effect in the overlap subpopulation (ATO), i.e. the subpopulation

comprised of individuals with covariate values which provide sufficient positive probability of being in either exposure group. Table 2.1 defines the functions $h(x)$ and their estimands as laid out in Table 1 of Li et al. (2017).

Target Subpopulation	$h(x)$	Estimand
Population	1	ATE
Treated	$e(x)$	ATT
Control	$1 - e(x)$	ATC
Overlap	$e(x)(1 - e(x))$	ATO
Truncated	$I(\delta < e(x) < 1 - \delta)$	Similar to overlap distribution, See Figure 2.1
Matching	$\min(e(x), 1 - e(x))$	Similar to overlap distribution, See Figure 2.1

Table 2.1: Subpopulations as defined in Li et al. (2017) Table 1

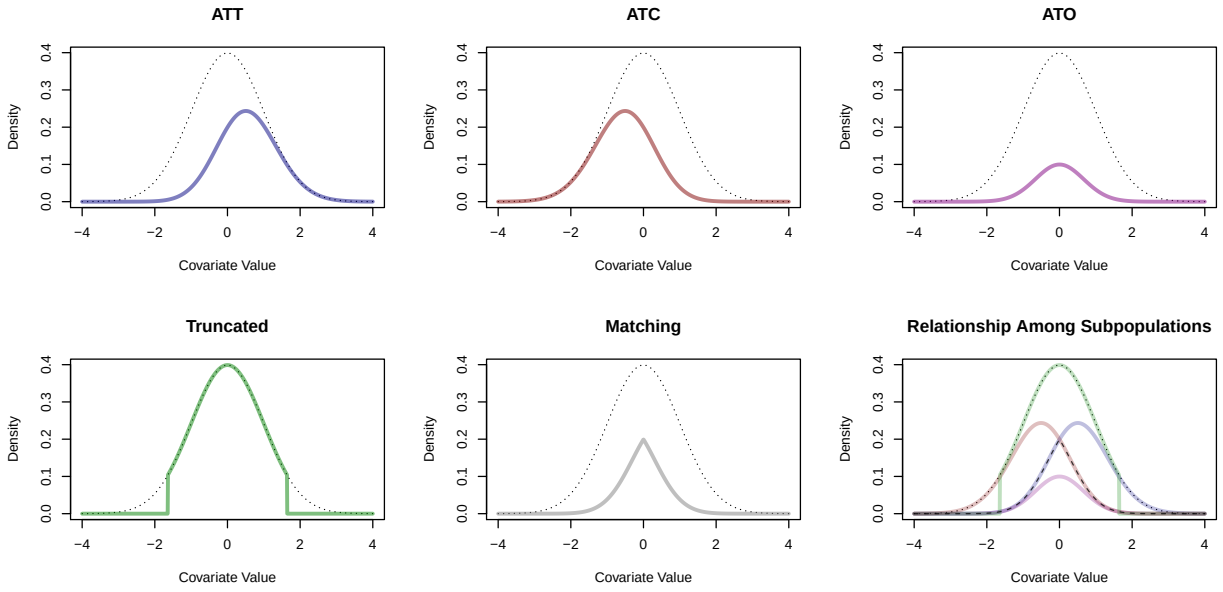

Figure 2.1: Visualizing specific subpopulations of interest that have been generated using functions of the propensity score. The density for each figure was created such that $g(x) = f(x)h(x)$ with $h(x)$ defined as in Table 2.1 and the propensity score is defined as $e(x) = \Phi(x)$. The dotted line in each figure represents the density of standard normal random variable. The lower right figure in the graphic represents the relationship among the subpopulations of interest

To visualize what these subpopulations look like, let $X \sim \mathcal{N}(0, 1)$ and consider a binary treatment setting where we assume that the probability of being assigned to the treatment exposure is a nonlinear transformation of X, i.e., $P(Z = 1|X = x) = \Phi(x)$. Here, $\Phi(\cdot)$ is

the cumulative distribution function of a standard normal random variable, or the probit link function. Figure 2.1 demonstrates unnormalized subpopulation densities as outlined in Table 2.1 where $g(x) = f(x)h(x)$. From this figure it is clear that, under this probability assignment function, each subpopulation of interest has different support. When we estimate treatment effects we must be careful to understand for who these estimates are relevant.

The framework developed in Li et al. (2017) defines treatment effects in the following way. They assume that the density of X exists with respect to base measure μ (a product of counting measure for categorical random variables and Lebesgue measure for continuous random variables) and then define the general class of estimands of the form

$$\tau_g = \frac{\int \tau(x)f(x)h(x)\mu(dx)}{\int f(x)h(x)\mu(dx)}. \tag{2.6}$$

2.1.3 Nonparametric Weighting Estimation of Treatment Effects

Section 2.1.2 introduced the potential outcome framework and defined estimands common throughout the literature. We now consider the estimation of treatment effects. For definiteness, we focus on effects that are functions of the mean outcome in the subpopulation of interest. We discuss estimators in the setting of the potential outcome framework, specifically focusing on nonparametric weighted estimators of the mean outcome in treatment group z, $E(Y^{Z=z})$, given a vector of weights $\mathbf{w}$, defined as $\bar{Y}_{\mathrm{w}}^{Z=z}$. The estimator $\bar{Y}_{\mathrm{w}}^{Z=z}$ has the form (Li et al., 2017),

$$\bar{Y}_{\mathrm{w}}^{Z=z} = \frac{\sum_{i=1}^{N} w_i I(Z_i = z) Y_i^{obs}}{\sum_{i=1}^{N} w_i I(Z_i = z)} \tag{2.7}$$

These are known as weighting estimators.

We first consider the goal of estimating the average treatment effect in the population for

a binary treatment regime through a simple difference, i.e., $\tau = E_{f(x)}[\tau(x)]$, with $\tau(x) = E(Y^{Z=1} - Y^{Z=0} \mid X = x)$. As stated at the outset, in the potential outcome framework we do not observe the $Y_i^{Z=z}$ directly, but instead observe $Y_i^{obs} = Z_i Y_i^{Z=1} + (1 - Z_i) Y_i^{Z=0}$. One would then hope that $E(Y^{obs}|Z = 1)$ - $E(Y^{obs}|Z = 0)$ may be used to estimate the treatment effect. An estimator of this quantity is,

$$\hat{\tau}_{naive} = \left(\frac{\sum_{i=1}^{N} I(Z_i = 1) Y_i^{obs}}{\sum_{i=1}^{N} I(Z_i = 1)} \right) - \left(\frac{\sum_{i=1}^{N} I(Z_i = 0) Y_i^{obs}}{\sum_{i=1}^{N} I(Z_i = 0)} \right) \tag{2.8}$$

Unfortunately, it can be shown that this naive estimator is only a consistent estimator for τ if $P(Z = 1)$ is a constant between zero and one, which cannot be assumed in observational settings, and is often only possible through the use of controlled experimentation.

To demonstrate the bias that can arise in observational studies, consider a simple example, where the potential outcomes are linear functions of a covariate X with a constant treatment effect τ (Cochran and Rubin, 1973). Let the two potential outcomes be $Y^{Z=0} = \beta_0 + \beta_1 X + \epsilon_{Z=1}$ and $Y^{Z=1} = \tau + \beta_0 + \beta_1 X + \epsilon_{Z=0}$, where $\epsilon_{Z=z} \sim \mathcal{N}(0, \sigma^2)$. Clearly,

$$E(Y^{obs}|Z = 1) - E(Y^{obs}|Z = 0) \tag{2.9}$$

$$= E([\tau + \beta_0 + \beta_1 X]|Z = 1) - E([\beta_0 + \beta_1 X]|Z = 0) \tag{2.10}$$

$$= \tau + \beta_1 [E(X|Z = 1) - E(X|Z = 0)] \tag{2.11}$$

demonstrating that estimators based on Y^{obs} will be functions of the conditional distributions of X. If $P(Z = z)$ is a constant it follows that $P(X|Z) = P(X)$ and there should be no difference between the conditional distributions of X (the motivation for randomization in experimental studies) and the naive estimator will be unbiased for τ. The difference in the expected value of the covariates may be nonzero if $P(Z = 1|X = x)$ varies at different levels of X.

Given that the naive estimator is biased in typical settings, we need assumptions that allow us to identify observational settings in which causal inference is possible. Rosenbaum and Rubin (1983) defined two assumptions that make causal inference possible, that of **strong ignorability given a set of covariates** and the **stable unit treatment value assumption** (SUTVA). Using the notation for conditional independence of Dawid (1979), an assumption of strong ignorability given a covariate vector X implies that $\{Y^{Z=z}\} \perp\!\!\!\perp Z \mid X$ and $0 < P(Z = z \mid X = x) < 1$. The first requirement is often referred to as an unconfoundedness assumption, while the second requirement is often referred to as the positivity assumption, or the overlap assumption. The unconfoundedness assumption implies that all of the covariates that were used to assign treatment exposures and that are causally related to the outcome of interest have been identified. The positivity assumption assures that we are able to find comparable units at specific levels of the covariate space. The stable unit treatment value assumption implies that there is no hidden variability in the treatment levels and that there is no interference between units. No interference implies that the application of treatment to one unit does not affect the outcomes of any other units (Rosenbaum, 2007). Mathematically, no hidden variability implies that we are able to write $Y_i^{obs} = \sum_{z=0}^{E} I(Z_i = z) Y_i^{Z=z}$. Under these assumptions, if we observe an outcome conditioned at a specific $X = x$, it follows that

$$E(Y^{obs} \mid Z = z, X = x) = E(Y^{Z=z} \mid Z = z, X = x) \tag{2.12}$$

$$= E(Y^{Z=z} \mid X = x) \tag{2.13}$$

and therefore if we marginalize across the distribution of X, it follows that

$$E_X(E(Y^{obs} \mid Z = z, X = x)) = E_X(E(Y^{Z=z} \mid X = x)) \tag{2.14}$$

$$= E(Y^{Z=z}) \tag{2.15}$$

and thus by first conditioning on X and then marginalizing across the distribution, we are

able to provide unbiased estimation of average treatment effects. In practice, methods for conditioning on X include regression modeling, subclassifying individuals in some neighborhood about X, or exactly matching on the covariate X for estimating treatment effects. See Imbens and Rubin (2015) for a thorough discussion of these methods.

In observational studies, the dimension of the covariate space is often large and therefore finding individuals that match exactly on a covariate vector may be difficult. Thus, some dimensionality reduction may be necessary for estimating treatment effects. A common method in the causal literature for reducing the dimension of X involves adjustment on a balancing score, where a balancing score is any function $b(\cdot)$ such that $Z \perp\!\!\!\perp X \mid b(X)$ (Rosenbaum and Rubin, 1983). The primary result based on balancing scores is that if strong ignorability holds given X, then there is strong ignorability given a balancing score $b(X)$ (Theorem 3 of Rosenbaum and Rubin (1983) for binary treatments and Lemma 3 of Imbens (2000) for multi-valued treatments). This implies that conditioning on a balancing score is sufficient for estimation. Note that it is trivially true that $b(X) \equiv X$ is a balancing score, but that is not helpful if the dimension of X is large.

One of the most commonly employed balancing scores is the propensity score, where the propensity score $e(x)$ is defined in a binary treatment setting as $e(x) = Pr(Z = 1|X = x)$ (Rosenbaum and Rubin, 1983) and for multi-treatment settings the generalized propensity score was put forth by Imbens (2000) to be the probability of a particular treatment received $e(z, x) = Pr(Z = z \mid X = x)$. The propensity score was further extended to continuous treatment regimes in Imai and van Dyk (2004), but the focus of this dissertation will be on binary and multi-level treatment exposures. The results described above imply that conditioning on the propensity score is sufficient for the estimation of treatment effects. This is often done through weighting by the estimated propensity score, subclassifying on the propensity score, matching on the propensity score, regression adjustment within subclasses

of propensity scores, or other methods (Imbens and Rubin, 2015, Chapters 12, 13, 17, 18).

The focus of this dissertation is on weighting estimators for treatment effects, similar to Horvitz-Thompson estimation (Horvitz and Thompson, 1952) in survey analysis. That is, we focus on estimating the average response in each group as,

$$\bar{Y}_{\mathrm{w}}^{Z=z} = \frac{\sum_{i=1}^{N} w_i(X_i) I(Z_i = z) Y_i^{obs}}{\sum_{i=1}^{N} w_i(X_i) I(Z_i = z)}, \tag{2.16}$$

$$= \frac{\sum_{i=1}^{N} w_i I(Z_i = z) Y_i^{obs}}{\sum_{i=1}^{N} w_i I(Z_i = z)}, \tag{2.17}$$

where the weights $\mathbf{w}$ are chosen to be functions of the observed covariates X, and the relevant theory. If we consider estimating the average response for group z in the population, Theorem 4 of Imbens (2000) demonstrates that,

$$E\left(\frac{I(Z = z) Y^{obs}}{Pr(Z = z|X)}\right) = E(Y^{Z=z}), \tag{2.18}$$

suggesting that if we weight observations by the inverse of the probability of receiving the treatment that they received, then we may construct consistent estimators of treatment effects.

More generally, Li et al. (2017) demonstrated that estimators of the form of Equation (2.17) are also consistent for the estimands that they had defined. Specifically, if we define weights as in Table 2.2, then we can obtain consistent estimates of treatment effects for the sub-populations of interest. This approach is explored in Chapter 3 of this dissertation.

There is nothing in Equation (2.17) that requires the weights to be functions of the propensity score. Alternative approaches have been proposed in Hainmueller (2012) and Zubizarreta (2015), which perform optimization procedures to find weights that are optimal for specific subpopulations of interest using defined criteria. This approach is explored in Chapter 4 of

Target Subpopulation	$h(X)$	Estimand	Weights
Population	1	ATE	$w_i = \frac{Z_i}{e(x_i)} + \frac{1-Z_i}{1-e(x_i)}$
Treated	$e(x)$	ATT	$w_i = Z_i + (1-Z_i)\frac{e(x_i)}{1-e(x_i)}$
Control	$1-e(x)$	ATC	$w_i = Z_i\frac{1-e(x_i)}{e(x_i)} + (1-Z_i)$
Overlap	$e(x)(1-e(x))$	ATO	$w_i = Z_i(1-e(x_i)) + (1-Z_i)e(x_i)$
Truncated	$I(\delta < e(x) < 1-\delta)$	-	$w_i = Z_i\frac{I(\delta < e(x_i) < 1-\delta)}{e(x_i)} + (1-Z_i)\frac{I(\delta < e(x_i) < 1-\delta)}{1-e(x_i)}$
Matching	$\min(e(x), 1-e(x))$	-	$w_i = Z_i\frac{\min(e(x_i), 1-e(x_i))}{e(x_i)} + (1-Z_i)\frac{\min(e(x_i), 1-e(x_i))}{1-e(x_i)}$

Table 2.2: Subpopulations as defined in Li et al. (2017) Table 1 and the corresponding weights to estimate treatment effects for the specific subpopulation of interest.

this dissertation.

2.1.3.1 Example Using Known Propensity Scores

To illustrate the ideas discussed to this point, we consider a sample of $N = 500$ observations with a single covariate $X \sim \mathcal{N}(0, 1)$ responsible for treatment assignment such that the propensity score is specified as $e(x) = \Phi(x)$, where $\Phi(\cdot)$ is the standard normal cumulative distribution function. Additionally, we consider the following potential outcomes, $Y_i^{Z=1} = X_i^2 + 2 + \epsilon_{i,Z=1}$ and $Y_i^{Z=0} = -X_i + \epsilon_{i,Z=0}$, where $\epsilon_{i,Z=z} \sim \mathcal{N}(0, 1)$. We visualize data from this generative model in Figure 2.2 with the size of the dot showing the relative weighting using the true propensity scores and based upon the weighting specifications of Table 2.2. It is clear that under some settings, the weights vary considerably.

We now demonstrate the true treatment effect for each subpopulation defined in Table 2.2. To do this, we examine the empirical distributions of weighted-averages of the difference in the potential outcomes for each subpopulation in a finite-sample setting, i.e., $\frac{\sum_{i=1}^N w_i(x_i)(Y_i^{Z=1} - Y_i^{Z=0})}{\sum_{i=1}^N w_i(x_i)}$, where the weights are chosen for specific subpopulations of interest. We simulate 1000 data sets using the generative procedure described at the beginning of this section and provide a weighted estimate of the finite-sample subpopulation treatment effect where, for each

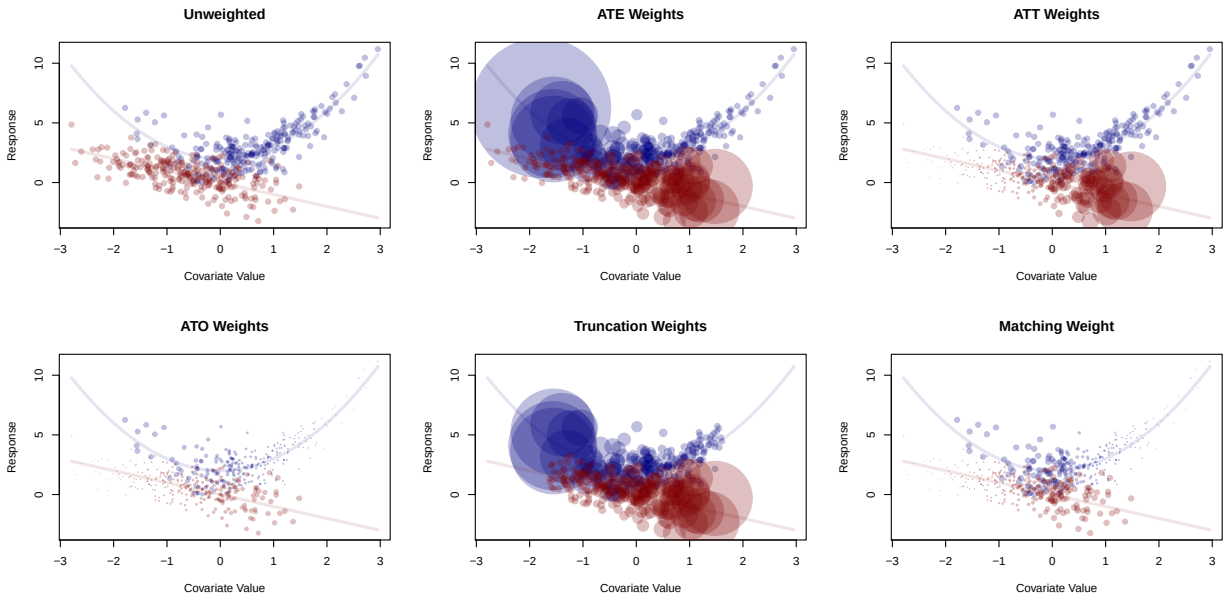

Figure 2.2: Visualization of the relative weights in each group under weighting schemes from Table 2.2 (Note: ATC is not shown due to its similarity to the ATT weighting). The size of each dot was created by setting the `R` plotting parameter `cex` to be the weight of the observation. The first plot shows each observation with weight one. The first row demonstrates how the ATE and ATT weighting schemes increase the weights of those who received the opposite treatment than probability would suggest. The lower panel demonstrates how the ATO, Truncated, and Matching weights all tend to focus on the areas where there is non-zero probability of being in either group.

subpopulation $g(x)$, the weights for each individual unit are proportional to the corresponding function $h(x_i)$ that is found in Table 2.2 for that specific subpopulation, i.e.,

$$\tau_{sim,g(x)} = \frac{\sum_{i=1}^{N} h(x_i)(Y^{Z=1} - Y^{Z=0})}{\sum_{i=1}^{N} h(x_i)}. \tag{2.19}$$

These empirical distributions are visualized in Figure 2.3, demonstrating that even in this simple setting the average treatment effects are distinct across the defined subpopulations. This suggests that understanding the estimand of interest and its support within the covariate space are an important piece of a causal analysis.

We can also compare the finite-sample empirical distributions of the average treatment effects arising from Equation (2.19) and the weighting estimators that utilize only the observed

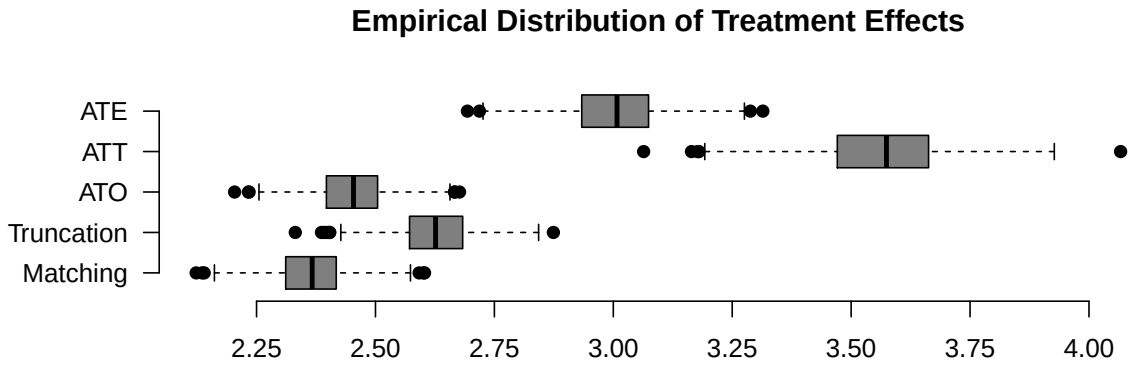

Figure 2.3: Finite-sample empirical distributions of average treatment effects in subpopulations defined in Table 2.2.

responses Y^{obs} as in Equation (2.17). To utilize Equation (2.17) the weights are defined in Table 2.2 and are computed using the true propensity scores (which are known). We can then estimate $\bar{Y}_{\text{w}}^{Z=1} - \bar{Y}_{\text{w}}^{Z=0}$ for each simulation and compare the empirical distribution of these estimates against the empirical distributions demonstrated in Figure 2.3. This is provided in Figure 2.4, From this simulation, we see that there is nice agreement between these two empirical distributions across the subpopulations of interest. In general the variability is increased under estimation using only Y^{obs} and the true propensity score (expected due to the weighting scheme). We also see that the estimates for the ATE and the ATT have much higher variability. This is tied to the fact that some observations have probability of treatment assignment close to zero or one, resulting in large weights as shown in Figure 2.2 (an implication that the positivity assumption is not perfectly valid in this scenario). The ATO, Truncation, or Matching estimators avoid this by down-weighting such observations (i.e. by focusing on the overlap distribution), or by thresholding very large weights.

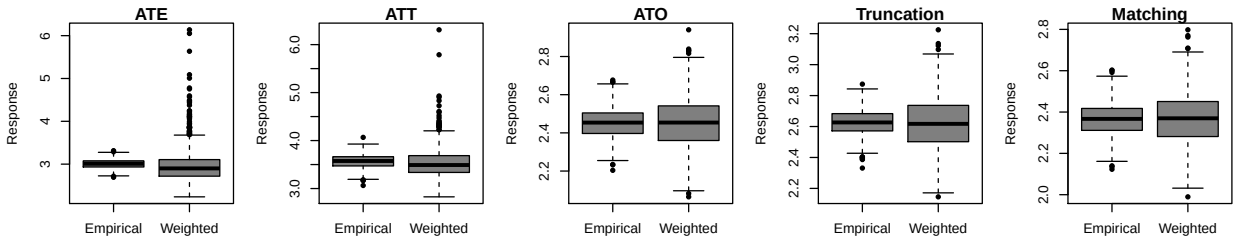

Figure 2.4: Comparing the empirical distribution of the estimated treatment effects demonstrated in Figure 2.3 against the empirical distribution of the average treatment effects estimated using only the observed outcomes weighted to a specific subpopulation of interest

2.1.4 Methods for Estimating Propensity Scores and Balancing Scores

There is an extensive literature on techniques for estimating balancing scores and propensity scores. Propensity score estimators generally approach the estimation problem from two angles: (1) nonparametric estimators of the propensity score where, under an assumption of strong ignorability, the model is unbiased for the true propensity score and an application of the resulting propensity score estimates should provide sufficient covariate balance; (2) algorithms that estimate the propensity score by "targeting" covariate balance directly, i.e., covariate balance is seen as a stopping criteria within the modeling procedure. The second strategy is endorsed by Imbens and Rubin (2015) who wrote (see Chapter 13, Section 9), "the goal... is to obtain an estimated propensity score that balances covariates..., rather than one that simply estimates the hypothetical true propensity score as accurately as possible."

The first estimation strategy is the use of nonparametric modeling procedures and modern statistical learning techniques to obtain an unbiased estimate of the propensity score. Under the assumptions of strong ignorability given X it follows that these estimates provide covariate balance and are sufficient for use within the weights defined in Table 2.2. One such method for estimating the propensity score uses Gradient Boosting Machines (McCaffrey et al., 2004), while another popular technique is the use Bayesian Additive Regression Trees (Chipman et al., 2010). This dissertation puts forward a procedure in Chapter 3 that relies on ideas from both

nonparametric estimation and a targeting of covariate balance. The idea that is explored is to model the propensity score nonparametrically using a Gaussian process methodology, but to select the hyperparameters of the process so that they are optimized with respect to a covariate imbalance metric.

An example of the second estimation strategy is outlined in Chapter 13 of Imbens and Rubin (2015) where a model building procedure is developed and covariate balance is used as a model selection criteria. They provide an algorithm that develops a binary regression model that estimates the propensity score. The first step considers a subset of pretreatment covariates X_d that are believed to be related to the assignment mechanism and plausibly related to the outcome. They then add terms in a strategic step-wise fashion, adding terms based upon a likelihood-ratio statistic, though ultimately they evaluate their model at the last step with respect to covariate balance. If the balance is inadequate, the model is augmented until covariate balance is attained. Another method that employs this strategy a bit more directly is the Covariate Balancing Propensity Score (CPBS) (Imai and Ratkovic, 2014), which requires specifying a linear model and then develops a framework to optimize the regression model coefficients with respect to covariate balance.

As stated earlier, there is an alternative to using a propensity score to achieve covariate balance that is based directly on the notion of balancing scores. This approach estimates weights for each unit directly through an optimization procedure without directly relating the weights to the treatment assignment mechanism. This is the approach taken by Hainmueller (2012) in their entropy balancing procedure, where their entropy loss is optimized under the constraint that the weighted distributions of the covariates are similar. A related approach is that of Zubizarreta (2015) which derives stable balancing weights that minimize the variability in the weights subject to constraints that ensure balanced weighted sample moments. We propose a new procedure in Chapter 4 that minimizes covariate imbalance subject to a

constraint that controls the effective sample sizes in each group.

2.1.5 Evaluating Covariate Balance by Weighted Distributional Measures

Nonparametric estimators of the form

$$\bar{Y}_{\mathrm{w}}^{Z=z} = \frac{\sum_{i=1}^{N} w_i(X_i) I(Z_i = z) Y_i^{obs}}{\sum_{i=1}^{N} w_i(X_i) I(Z_i = z)} \tag{2.20}$$

require weights $\mathbf{w} = \{w_i(X_i)\}$ that must be estimated. The previous section discussed different approaches for estimating these unit weights. Regardless of the approach, it is necessary to provide diagnostic tools that enable us to understand when causal inference is appropriate.

As the definition of a balancing score is any function $b(\cdot)$ such that $Z \perp\!\!\!\perp X \mid b(X)$, one way of assessing covariate balance is to compare functions of the distributions of the covariates X, conditional upon the assigned treatment, i,e., comparing different characteristics of the conditional distributions $Pr(X \mid Z = z)$ across the exposure groups. When there is balance,

$$Pr(X|Z = 0) \approx Pr(X|Z = 1) \approx \cdots \approx Pr(X|Z = E). \tag{2.21}$$

Commonly employed methods for these comparisons are through the use of functions of the weighted sample moments, or the weighted cumulative distribution function. These are defined formally below.

2.1.5.1 Comparing Moments of Conditional Distributions

The first method for comparing distributions is to compare functions of the empirical weighted moments in each dimension d among the groups z. Define, the p^{th} non-central moment for the d^{th} dimension of the conditional distribution of X with (sub)population density $g(x)$ given group assignment z as $M_{d,p,z} \equiv E_{g(x)}(X_d^p | Z = z)$ and similarly the p^{th} central moment for dimension d as $\mu_{d,p,z} \equiv E_{g(x)}\big((X_d - M_{d,1,z})^p | Z = z\big)$. A weighted estimator for the non-central moments is of the form,

$$\hat{M}_{d,p,z}(\mathbf{w}) = \frac{\sum_{i=1}^{N} w_i(X_i) I(Z_i = z) X_{d,i}^p}{\sum_{i=1}^{N} w_i(X_i) I(Z_i = z)}, \tag{2.22}$$

for appropriately chosen weights, and similarly for the central moments

$$\hat{\mu}_{d,p,z}(\mathbf{w}) = \frac{\sum_{i=1}^{N} w_i(X_i) I(Z_i = z)(X_{d,i} - \hat{M}_{d,1,z}(\mathbf{w}))^p}{\sum_{i=1}^{N} w_i(X_i) I(Z_i = z)} \tag{2.23}$$

where the mean is replaced with a moment-based estimator. Additionally, it is possible to compare weighted moments of interactions between two dimensions, i.e., $M_{d,p,d',q,z} \equiv E_{g(x)}(X_d^p X_{d'}^q | Z = z)$, where an appropriate weighted estimator is of the form,

$$\hat{M}_{d,p,d',q,z}(\mathbf{w}) = \frac{\sum_{i=1}^{N} w_i(X_i) I(Z_i = z) X_{d,i}^p X_{d',i}^q}{\sum_{i=1}^{N} w_i(X_i) I(Z_i = z)}. \tag{2.24}$$

This can help to capture correlations among variables. Three-way and higher-order interactions may also be inspected.

Two common dimensionless measures for evaluating covariate balance in binary treatment settings are the standardized difference in the means and the logarithm of the sample standard deviations (Imbens and Rubin, 2015). Let $\bar{X}_{d,Z=z} \equiv \hat{M}_{d,1,z}(\mathbf{w})$ and let $s^2_{d,Z=z} \equiv \hat{\mu}_{d,2,z}(\mathbf{w})$, then the standardized difference in the means and the logarithm of the ratio of the sample

standard deviations in dimension d are defined as,

$$\Delta_d = \frac{\bar{X}_{d,Z=1} - \bar{X}_{d,Z=0}}{\sqrt{\frac{s^2_{d,Z=1}+s^2_{d,Z=0}}{2}}} \quad \text{and} \quad \Gamma_d = \log\left(\frac{s_{d,Z=1}}{s_{d,Z=0}}\right) = \log s_{d,Z=1} - \log s_{d,Z=0}, \tag{2.25}$$

respectively. As both of these measures are standardized, they allow for a fair evaluation of the relative covariate imbalance among the dimensions. One issue with the above metrics is that small values of $\bar{X}_{d,Z=1} - \bar{X}_{d,Z=0}$ do not imply that $\bar{X}_{d,Z=1} \approx \bar{X}_{d,Z=0} \approx E_{g(x)}(X_d)$, implying that while the mean difference is near zero, the weighting may be such that the distribution is positioned away from the target.

An alternative approach to checking covariate balance is provided in Zubizarreta (2015), where they define a series of explicit targets, π^p_d (and similar for interactions and higher moments), and evaluate the balance between the weighted moments and the target values, i.e.,

$$\Omega_{d,p,z} = \hat{M}_{d,p,z}(\mathbf{w}) - \pi^p_d. \tag{2.26}$$

In practice these are chosen to be the sample moments of the target distribution of interest, i.e., $g(x)$. For example, if the goal is estimating the ATE, then the targets can be chosen as the unweighted sample moments from that data, $\pi^p_d = \frac{1}{N}\sum_{i=1}^N X^p_{i,d}$ and for the ATT these may take the form of the unweighted sample moments in only the treatment group, e.g., $\pi^p_d = \frac{\sum_{i=1}^N X^p_{i,d} I(Z_i=1)}{\sum_{i=1}^N I(Z_i=1)}$. These measures of covariate balance are no longer dimensionless, but can made so if the columns of X are standardized to the target distribution prior to estimating the sample moments. This would allow for relative comparisons of the balance metrics among the dimensions, as is possible for the standardized difference in means and logarithm of the ratio of the sample standard deviations.

2.1.5.2 Comparing the Weighted Empirical Cumulative Distribution Functions

Beyond the moments of the within-treatment covariate distributions, another evaluative tool is comparing the weighted cumulative distribution functions. The weighted cumulative distribution function at a point x in dimension d is,

$$\hat{F}_d(x, z, \mathbf{w}) = \frac{\sum_{i=1}^{N} w_i(X_i) I(Z_i = z) I(X_{i,d} \leq x)}{\sum_{i=1}^{N} w_i(X_i) I(Z_i = z)} \tag{2.27}$$

This metric is useful for visualizing discrepancies between the conditional distributions of the covariates across treatment groups. While not explored in this dissertation, this is also useful for constructing a weighted Kolmogrov-Smirnoff (KS) test statistic.

2.1.5.3 Effective Sample Size after Weighting

The variance of weighting estimators for treatment effects depends on the relative magnitudes of the weights. For example, if one observation within an analysis is given a very large weight while all others remain small, the effective sample size utilized in this analysis is close to one. Conversely, if all samples are given equal weight, then the effective sample size is the original size of the data N. The "effective sample size" (ESS) is a measure which captures the relative amount of data being used (McCaffrey et al., 2004),

$$ESS = \frac{(\sum_{i=1}^{N} w_i)^2}{(\sum_{i=1}^{N} w_i^2)} \tag{2.28}$$

This can also be calculated separately within each treatment group by including the indicators for treatment assignments,

$$ESS(z) = \frac{(\sum_{i=1}^{N} I(Z_i = z) w_i)^2}{(\sum_{i=1}^{N} I(Z_i = z) w_i^2)}. \tag{2.29}$$

This metric does not provide insight into the balance of the covariate distribution, but does provide information about the amount of data that is being utilized within the weighting procedure.

2.1.5.4 Example Applying Covariate Balance Measures

Consider a simple example, where $X \sim \mathcal{N}(0, 1)$ and the binary treatment assignment mechanism is $Pr(Z = 1|X = x) = \Phi(\alpha x)$, for $\alpha \in \{0, 1, 2\}$. We use the measures presented in this subsection on weighted and unweighted data to demonstrate the effect of different assignment mechanisms on mean covariate balance. The results in Table 2.3 contain the averages from 1000 simulated data sets from the specified generative mechanisms with each data set consisting of 500 observations. As can be seen, when $\alpha = 0$, as in a completely randomized experiment, there is no discrepancy between the distributions prior to weighting. Additionally, since the probability of treatment is a constant, all observations are weighted equally and we see no difference between the average unweighted effective sample size and the average weighted effective sample size across the simulated data sets.

As α increases, the unweighted difference in the means of the conditional distributions increases. To provide a weighted comparison, we weight each observation using weights as defined in Table 2.2 for estimating the ATE. By weighting in this way, we remove much of the difference in the means between the two conditional distributions. This begins to breakdown when $\alpha = 2$, where the average standardized difference in the means between the two distributions is still large (~ 0.4). This is due to the fact that when $\alpha = 2$, many of the observations have $Pr(Z = 1|X = x)$ close to zero or one. Finally, it is also demonstrated that as α increases, the effective sample size decreases, suggesting that weighting estimators of treatment effects based upon these weights would more heavily focus on a subset of observations.

	$\alpha = 0$		$\alpha = 1$		$\alpha = 2$	
Balance Measure	Unweighted	Weighted	Unweighted	Weighted	Unweighted	Weighted
$\bar{X}_{z=0}$	0.000	0.000	-0.566	-0.021	-0.716	-0.162
$\bar{X}_{z=1}$	0.001	0.001	0.564	0.022	0.712	0.167
Δ	0.000	0.000	1.375	0.063	2.043	0.417
ESS_0	249.624	249.624	250.110	124.686	249.681	78.085
ESS_1	250.376	250.376	249.890	126.544	250.319	78.566

Table 2.3: Simulation results demonstrating covariate balance measures for a simple binary treatment assignment mechanism where $Pr(Z = 1|X = x) = \Phi(\alpha x)$. Each entry is the average measure across 1000 simulated data sets, each containing 500 observations.

In addition to tabulating the balance measures, Figure 2.5 illustrates the empirical cumulative distribution function of the covariate, before and after propensity score weighting. As is apparent from the figure, the weighting "aligns" the distributions to the distribution of X, the dotted line in each figure (as expected since this was ATE weighting). We also see more evidence, in the lower right panel, that the propensity score may be inadequate for adjustment when $\alpha = 2$.

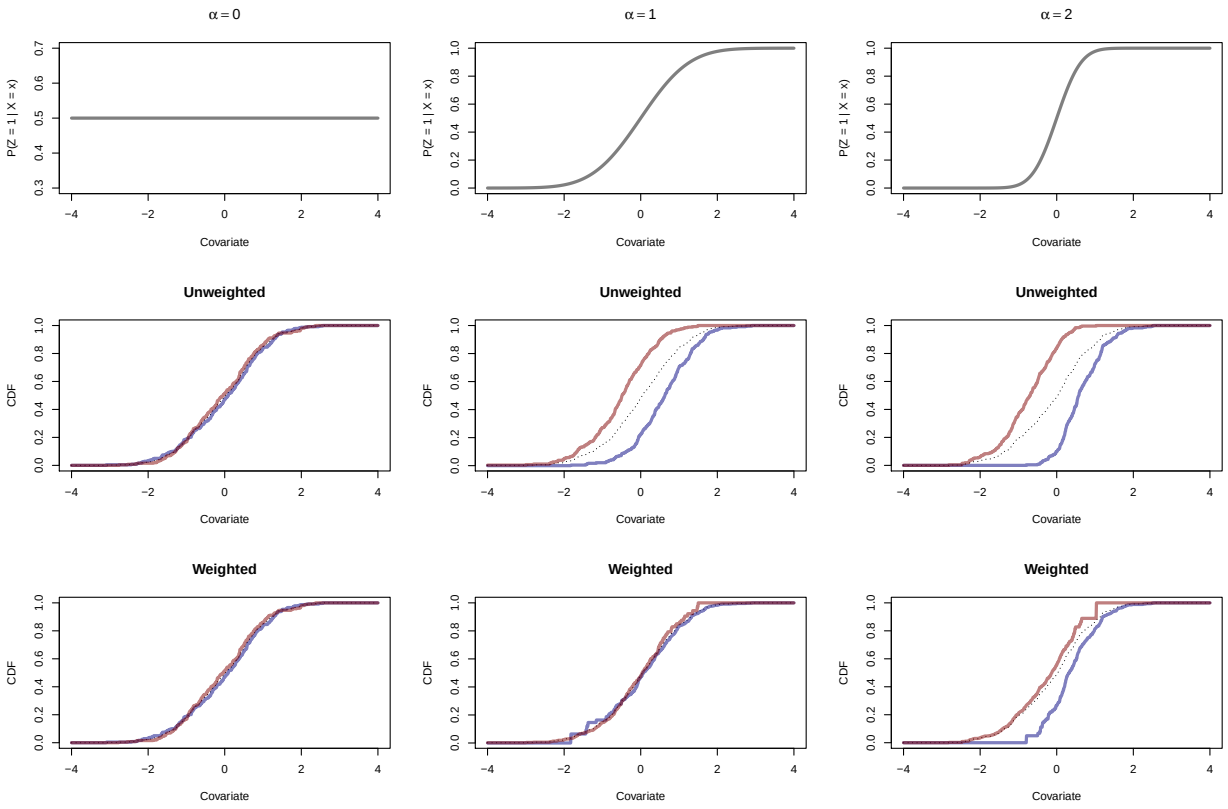

Figure 2.5: Visualizing the cumulative distribution function before and after weighting by the propensity score using ATE weights. The first row of the figure is the treatment assignment used to simulate the data. Notice that the figure in the upper right hand corner represents a case where there are significant positivity violations for a covariate $X \sim \mathcal{N}(0, 1)$. The second row demonstrates the lack of balance resulting from each assignment mechanism prior to propensity score weighting. The last row demonstrates the ability of the propensity score to balance covariates. Notice that with large positivity violations, it is difficult for even the true propensity score to balance covariates.

2.2 Overview of Regression with Gaussian Processes

One contribution of this dissertation is the development of a nonparametric method for estimating propensity scores using Gaussian processes. This section provides an overview of Gaussian process models and their use in binary and multinomial regression.

2.2.1 Binary Regression with Gaussian Processes

Regression is typically focused on modeling the relationship between the mean of an outcome variable and a set of covariates, potentially given a set of parameters. If we consider a binary random variable $Z \in \{0, 1\}$, and assume that the mean for this random variable is dependent on a transformation of a latent value that is a function of a d-dimensional vector of covariates X, i.e.,

$$P(Z = 1|X = x) = g(f(x)), \tag{2.30}$$

where $f : \Re^d \mapsto \Re$ and is known as the predictor function and $g : \Re \mapsto [0, 1]$ is known as the link function (McCullagh and Nelder, 1989). Two common choices for the transformation $g(\cdot)$ are the cumulative distribution function of a standard normal distribution (often referred to as the probit function), $g(c) = \Phi(c)$, and the inverse-logit function $g(c) = \frac{1}{1+\exp(-c)} = \frac{\exp(c)}{1+\exp(c)}$. When used in a regression setting, these transformations correspond to probit and logistic regression respectively. The choice of $g(\cdot)$ is typically motivated by the ease in scientific interpretation (inverse-logit), or by mathematical convenience (probit).

A common parametric modeling approach is the assumption that the predictor function is a linear combination of covariates and a vector of parameters θ, i.e., $f(x; \theta) = \tilde{x}^T\theta = \theta_0 + \sum_{j=1}^{d} x_j\theta_j$, where θ_0 is an offset (or intercept) term and $\tilde{x} = (1, x_1, \ldots, x_d)$. An alternative

nonparametric approach, and the one taken throughout the majority of this dissertation, is an assumption that $f(x)$ was generated from a Gaussian process.

To define a Gaussian process, first let $x_i \in \Re^d$ be an observed row vector of values, let the function $f : \Re^d \to \Re$ take as inputs x_i, and define the latent score $f_i \equiv f(x_i)$. Then, a Gaussian process (GP) is defined as a collection of random variables $\mathcal{F} = \{f_i\}$ for index set $i = 0, 1, 2, \ldots$ for which any finite subset of the collection has a joint Gaussian distribution. This definition implies that a GP can be interpreted as a distribution over *functions*. The process is characterized by the mean function,

$$m(x) = E(f(x)) \tag{2.31}$$

and a covariance function the defines the covariance between latent scores corresponding to any x and x',

$$k(x, x') = E\big((f(x) - m(x))(f(x') - m(x'))\big). \tag{2.32}$$

To denote that a single observation arose from a GP we write,

$$f(x) \sim \mathcal{GP}(m(x), k(x, \cdot))), \tag{2.33}$$

and if we collect the vectors x into a matrix $\mathbf{X}$, i.e., $\mathbf{X} = \begin{pmatrix} x_0 & x_1 & \ldots & x_n \end{pmatrix}^T$, we denote a collection of observations as arising from a GP as,

$$\mathbf{f}(\mathbf{X}) \sim \mathcal{GP}(\mathbf{m}(\mathbf{X}), \mathbf{K}(\mathbf{X})), \tag{2.34}$$

where

$$\mathbf{m}(\mathbf{X}) = \begin{pmatrix} m(x_1) \\ m(x_2) \\ \vdots \\ m(x_n) \end{pmatrix} \quad \text{and} \quad \mathbf{K}(\mathbf{X}) = \begin{pmatrix} k(x_1, x_1) & k(x_1, x_2) & \dots & k(x_1, x_n) \\ k(x_2, x_1) & k(x_2, x_2) & \dots & k(x_2, x_n) \\ \vdots & \vdots & \ddots & \vdots \\ k(x_n, x_1) & k(x_n, x_2) & \dots & k(x_n, x_n) \end{pmatrix}. \tag{2.35}$$

Often, to simplify notation, we remove the relationship of the mean and covariance on $\mathbf{X}$ and simply write,

$$\mathbf{f}|\mathbf{X} \sim \mathcal{GP}(\mathbf{m}, \mathbf{K}). \tag{2.36}$$

When we model the mean of an outcome variable using a Gaussian process, we may impose structure on the mean and covariance by specifying models for $m(x)$ and $k(x, x')$. We write $m(x; \gamma)$ as the model for the mean function with parameters γ, and we use $k(x, x'; \theta)$ to denote the model for $k(x, x')$ parameterized by θ. Correspondingly, we denote the collection of observations as arising from a GP with these parameters as follows,

$$\mathbf{f}|\mathbf{X}, \gamma, \theta \sim \mathcal{GP}(\mathbf{m}_\gamma, \mathbf{K}_\theta). \tag{2.37}$$

2.2.1.1 Bayesian Probit Regression Model

It is possible to model the binary outcome $Z|X$ in a Bayesian probit regression framework by assuming that the predictor function arose from a Gaussian process. Consider a sample

$i = 1, \ldots, N$, and let

$$Z_i | X_i = x_i \sim Bernoulli(\Phi(f(x_i)) \equiv \Phi(f_i)) \tag{2.38}$$

$$\mathbf{f} | \mathbf{X}, \theta \sim \mathcal{GP}(\mathbf{0}, \mathbf{K}_\theta) \tag{2.39}$$

where we introduce f_i as a latent predictor value and make an a priori assumption of a zero mean for the GP. We also employ an assumption that the covariance between two latent values, f_i and f_j can be modeled using a parameterized kernel function $k(x_i, x_j; \theta)$ for $i, j \in \{1, 2, \ldots, N\}$, Kernel functions and their properties are described in Section 2.2.2. Finally, we make an assumption that the hyperparameters are fixed (the model can be augmented to include prior distributions on the hyperparameters, but they are not utilized within this dissertation) and for the moment assume that they are known.

From this model specification, we can find the posterior distribution of the latent values given the observed outcomes and covariates. Here,

$$p(\mathbf{f} | \mathbf{X}, \theta, \mathbf{Z}) = \frac{p(\mathbf{Z}|\mathbf{f}) p(\mathbf{f}|\mathbf{X}, \theta)}{p(\mathbf{Z}|\mathbf{X}, \theta)} \tag{2.40}$$

$$\propto \left[\prod_{i=1}^{N} \Phi(f_i)^{Z_i} (1 - \Phi(f_i))^{1 - Z_i} \right] \det(2\pi \mathbf{K}_\theta)^{-1/2} \exp\left(-\frac{1}{2} \mathbf{f}^T \mathbf{K}_\theta^{-1} \mathbf{f} \right) \tag{2.41}$$

The above distribution is intractable and needs to be approximated. One approach is to use Markov Chain Monte Carlo (MCMC) to sample from the posterior distribution of the latent scores, but this is often computationally infeasible for Gaussian process regression due to the required inversion of the covariance matrix required at each step of the MCMC algorithm, an operation that scales as $\mathcal{O}(N^3)$. Therefore, for a Markov chain of length $C \gg N$

the MCMC algorithm requires $\mathcal{O}(CN^3)$ calculations. Alternative approximation methods to MCMC for binary regression with GPs are the Laplace approximation (LA) and the expectation propagation (EP) algorithm. These algorithms also scale as $\mathcal{O}(N^3)$, but require many fewer iterations than an MCMC algorithm and therefore are computationally preferable. They are described below.

We note that the individual likelihood components in Equation 2.41, i.e., $\Phi(f_i)^{Z_i}(1-\Phi(f_i))^{1-Z_i}$, are often mathematically inconvenient to work with. Without loss of generality, we can assume an alternative setting such that $Z_i \in \{-1, 1\}$. In this formulation, the likelihood can be written as,

$$p(\mathbf{Z}|\mathbf{f}) = \prod_{i=1}^{N} \Phi(Z_i f_i), \tag{2.42}$$

and the posterior distribution of the latent scores is as follows,

$$p(\mathbf{f}|\mathbf{X}, \theta, \mathbf{Z}) \quad \propto \quad \left[\prod_{i=1}^{N} \Phi(Z_i f_i)\right] \det\left(2\pi\mathbf{K}_\theta\right)^{-1/2} \exp\left(-\frac{1}{2}\mathbf{f}^T\mathbf{K}_\theta^{-1}\mathbf{f}\right) \tag{2.43}$$

This specification of the model does not change the intractability of the posterior distribution, but, as is seen below, is mathematically convenient for calculations in various approximation methods.

2.2.1.2 Laplace Approximation for Bayesian Probit Regression

One useful approach to inference in the Bayesian probit regression model is a Laplace approximation to the posterior distribution in Equation 2.43 (Rasmussen and Williams, 2006). This method approximates the posterior distribution of the latent values with a multivariate normal

distribution centered at the mode of the posterior distribution and with variance related to the second derivative of the log of the posterior distribution. That is,

$$\mathbf{f}|\mathbf{Z}, \mathbf{X}, \theta \dot{\sim} \mathcal{N}\left(\hat{\mathbf{f}}, \left[-\frac{\partial^2}{\partial \mathbf{f} \partial \mathbf{f}^T} \log p(\mathbf{f}|\mathbf{Z}, \mathbf{X}, \theta) \mid_{\mathrm{f}=\hat{\mathrm{f}}}\right]^{-1}\right) \tag{2.44}$$

where,

$$\hat{\mathbf{f}} = \arg\max_{\mathrm{f}} \log p(\mathbf{f}|\mathbf{Z}, \mathbf{X}, \theta) \tag{2.45}$$

$$= \arg\max_{\mathrm{f}} \left[\left(\sum_{i=1}^{n} \log \Phi(Z_i f_i)\right) - \frac{1}{2}\mathbf{f}^T \mathbf{K}_\theta \mathbf{f} + C\right] \tag{2.46}$$

which can be found using Newton's method. See Appendix A.1 for a derivation of the required derivatives and an outline of the method for probit regression with Gaussian processes. The appendix also contains the derivation of a Taylor series approximation of the log marginal likelihood that can be optimized to select hyperparameters.

The primary benefit of this method is its computational efficiency. While the approximation provides adequate point estimates of the posterior mean of $\mathbf{f}$ using the maximum of the posterior, the variance term can either over- or under-estimate the posterior variance. (See Section 2.2.3 for an example.)

2.2.1.3 Expectation Propagation for Bayesian Probit Regression

An alternative to the Laplace approximation is to approximate the posterior distribution of $\mathbf{f}$ using expectation propagation (EP), a type of message passing algorithm (Rasmussen and Williams, 2006; Gelman et al., 2014; Minka, 2001). This method has been shown to provide results that are closer to the true posterior distribution for binary regression with Gaussian processes (Kuss and Rasmussen, 2005) while still providing results much more expediently

than MCMC.

The following is a summary of the general EP algorithm as defined in Section 2.1 of Gelman et al. (2014) using notation consistent with their overview. The goal of the class of EP algorithms is to approximate a target density function $f(\theta)$ by another density $g(\theta)$. The algorithms assume a factorization such that,

$$f(\theta) \propto \prod_{k=0}^{K} f_k(\theta) \tag{2.47}$$

and in most Bayesian applications this factorization is $f(\theta) = p(\theta) \prod_{i=1}^{N} p(y_i|\theta)$ where $p(\theta)$ is a prior distribution and $\prod_{i=1}^{N} p(y_i|\theta)$ is the joint likelihood of the observed data. EP algorithms approximate $f(\theta)$ with another factorized density,

$$g(\theta) \propto \prod_{k=0}^{K} g_k(\theta) \tag{2.48}$$

where each $g_k(\theta)$, often referred to as a *site approximation*, is selected to be a member of an exponential family. For example, consider a binary probit regression model with $p(y_i|\theta) = \Phi(x_i^T\theta)^{y_i}(1-\Phi(x_i^T\theta))^{1-y_i}$ and a normal prior distribution on θ. A convenient choice for $g_k(\theta)$ is, $g_k(\theta) = \frac{1}{\sqrt{2\pi s_i^2}} \exp\left(-\frac{(\theta_i - m_i)^2}{2s_i^2}\right)$, the density of a normal distribution with parameters m_i and s_i^2, so that the normal prior distribution on θ is now conjugate for the approximate likelihood function and the posterior distribution may be found analytically.

This approximation strategy requires a method for selecting the parameters of the site approximations $g_k(\theta)$. The algorithm first creates a *cavity distribution*, defined as

$$g_{-k}(\theta) \propto \frac{g(\theta)}{g_k(\theta)} \tag{2.49}$$

and then uses the cavity distribution to construct a *tilted distribution*, defined as

$$g_{\backslash k}(\theta) \propto f_k(\theta) g_{-k}(\theta). \tag{2.50}$$

The cavity distribution is the approximation to $f(\theta)$ absent the contribution from the component $g_k(\theta)$. The tilted distribution then combines the cavity distribution with the true likelihood component $f_k(\theta)$ as another approximation to $f(\theta)$. Typically, at this point all other site parameters have been updated and the goal then is to update the site parameters at location k using the tilted distribution by minimizing some discrepancy between the tilted distribution and the approximation $g(\theta)$ with all other sites fixed. That is, the algorithm proceeds by updating the site approximations to $g_k^{new}(\theta)$,

$$g_k^{new}(\theta) = \arg\min_{g_k(\theta)} KL(f_k(\theta) g_{-k}(\theta) || g_k(\theta) g_{-k}(\theta)) \tag{2.51}$$

or some other discrepancy $D(\cdot||\cdot)$ in place of the KL-divergence. The general idea is that when this discrepancy measure is minimized the parameters of the local site approximation provide an adequate approximation to the local likelihood component k.

The formal algorithm is described in the box below,

General Expectation Propagation Algorithm (Gelman et al., 2014)

Goal: *Approximate* $f(\theta) \propto \prod_{k=0}^{K} f_k(\theta)$ *by* $g(\theta) \propto \prod_{k=0}^{K} g_k(\theta)$ *where* $g(\theta)$ *and* $g_k(\theta)$ *are from a selected exponential family*

1. Construct initial site approximations $g_k(\theta)$
2. Until all $g_k(\theta)$ have converged,
   - For $k \in \{0, 1, \ldots, K\}$ (in serial or parallel),
     (a) Compute the cavity distribution, $g_{-k}(\theta) \propto \frac{g(\theta)}{g_k,(\theta)}$
     (b) Update the site approximations

$$g_k^{new}(\theta) = \arg\min_{g_k(\theta)} D(f_k(\theta) g_{-k}(\theta) || g_k(\theta) g_{-k}(\theta))$$

We now show how EP can be applied on our GP setting where $\mathbf{f}$ replaces θ in the Gelman et al. (2014) description. For probit regression with Gaussian processes the posterior distribution of the latent values is given by Equation 2.43, which is repeated here for convenience,

$$p(\mathbf{f}|\mathbf{X}, \theta, \mathbf{Z}) \propto \det(2\pi\mathbf{K}_\theta)^{-1/2} \exp\left(-\frac{1}{2}\mathbf{f}^T\mathbf{K}_\theta^{-1}\mathbf{f}\right) \left[\prod_{i=1}^{N} \Phi(Z_i f_i)\right]. \tag{2.52}$$

For this posterior distribution, an efficient expectation propagation algorithm, as laid out in Rasmussen and Williams (2006), chooses site approximations that are unnormalized normal distributions,

$$g(f_i|\zeta_i, m_i, s_i^2) = \frac{\zeta_i}{\sqrt{2\pi s_i^2}} \exp\left(-\frac{1}{2}\frac{(f_i - m_i)^2}{s_i^2}\right), \tag{2.53}$$

so that the approximate posterior distribution is,

$$\begin{aligned} g(\mathbf{f}|\mathbf{X},\theta,\mathbf{Z}) &\propto p(\mathbf{f}|\mathbf{X},\theta) \times \prod_{i=1}^{N} g(f_i|\zeta_i, m_i, s_i^2) \\ &= \det(2\pi\mathbf{K}_\theta)^{-1/2} \exp\left(-\frac{1}{2}\mathbf{f}^T\mathbf{K}_\theta^{-1}\mathbf{f}\right)\left[\prod_{i=1}^{N}\frac{\zeta_i}{\sqrt{2\pi s_i^2}}\exp\left(-\frac{1}{2}\frac{(f_i-m_i)^2}{s_i^2}\right)\right] \\ &\propto \exp\left(-\frac{1}{2}\mathbf{f}^T\mathbf{K}_\theta^{-1}\mathbf{f}\right)\left[\prod_{i=1}^{N}\exp\left(-\frac{1}{2}\frac{(f_i-m_i)^2}{s_i^2}\right)\right] \\ &= \exp\left(-\frac{1}{2}\mathbf{f}^T\mathbf{K}_\theta^{-1}\mathbf{f}\right)\exp\left(-\frac{1}{2}(\mathbf{f}-\mathbf{m})^T\mathbf{\Sigma}^{-1}(\mathbf{f}-\mathbf{m})\right) \end{aligned}$$

where $\mathbf{m}$ is the vector of m_i values and $\mathbf{\Sigma} = diag(s_i^2)$. Using the results for normal-normal conjugate models (Rasmussen and Williams, 2006, see Appendix A.2), this implies that the approximate posterior distribution is

$$\mathbf{f}|\mathbf{X},\theta,\mathbf{Z} \sim \mathcal{N}\left((\mathbf{K}_\theta^{-1}+\mathbf{\Sigma}^{-1})^{-1}\mathbf{\Sigma}^{-1}\mathbf{m}, (\mathbf{K}_\theta^{-1}+\mathbf{\Sigma}^{-1})^{-1}\right). \tag{2.54}$$

A more detailed description of the algorithm is laid out in Appendix A.2 which provides derivations for the cavity distributions, tilted distributions, and the site approximation updates through a minimization of KL-divergence.

2.2.1.4 Selecting Hyperparameters - Maximizing a Marginal Likelihood

Up to this point the parameters θ have been assumed to be known and fixed. A common strategy for selecting these hyperparameters is to maximize the marginal likelihood of $\mathbf{Z}$, i.e.,

$$\hat{\theta} = \arg\max_{\theta} p(\mathbf{Z}|\mathbf{X},\theta) \tag{2.55}$$

$$= \arg\max_{\theta} \int p(\mathbf{Z}, \mathbf{f}|\mathbf{X}, \theta) d\mathbf{f}, \tag{2.56}$$

or a penalized form of this equation. In a Bayesian setting the penalty would correspond to a specification of a prior distribution on θ and then the maximum a-posteriori (MAP) estimate of the posterior distribution of the parameters may be found.

2.2.2 Kernel Functions

This section provides an informal overview of various kernels that are prevalent in the GP literature. Additionally it provides theorems that allow for the creation of new kernels through suitable combinations of kernels. For more thorough reviews on the general use of statistical kernel methods, see Schölkopf and Smola (2001), Hofmann et al. (2008), and, in the context of Gaussian processes, see Chapter 4 of Rasmussen and Williams (2006) and Chapter 2 of Duvenaud (2014).

A positive-definite kernel function represents the covariance between the latent function evaluated at points $x, x' \in \Re^d$,

$$k(x, x') = Cov\big(f(x), f(x')\big). \tag{2.57}$$

When evaluated on a sample of data, the *Gram matrix* (Hofmann et al., 2008, see Definition 1) is the covariance matrix that collects the pairwise evaluations of the kernel functions on

all units i and j, i.e., for a sample $x_1, x_2, \ldots, x_N$. The Gram matrix is written as,

$$\mathbf{K} = \begin{pmatrix} k(x_1, x_1) & k(x_1, x_2) & \ldots & k(x_1, x_N) \\ k(x_2, x_1) & k(x_2, x_2) & \ldots & k(x_2, x_N) \\ \vdots & \vdots & \ddots & \vdots \\ k(x_N, x_1) & k(x_N, x_2) & \ldots & k(x_N, x_N) \end{pmatrix}. \tag{2.58}$$

As this matrix must be positive definite, it follows that $\mathbf{K}_{ii} > 0$ for all i and that the matrix is symmetric. Typically kernels are functions of some parameter vector θ in addition to the observed data at two points.

Kernel functions have informally been described as a way to describe the similarity between the latent values at a point x and all other points x' (Duvenaud, 2014). As such, they can be seen as a convenient way to model the dependency between the latent values for two different observations. One of the most commonly used kernel functions is the multivariate squared exponential kernel with a common inverse-length scale,

$$k_{SE}(x, x'; \rho) = \exp\left(-\frac{\rho^2}{2}(x - x')^T(x - x')\right) \tag{2.59}$$

where ρ represents the inverse-length scale parameter. As the distance between two vectors increases, the dependency between the two vectors decreases. The left panel of Figure 2.6 visualizes the correlation between a point x and all other points x' for three choices of $x \in \Re$ (i.e., a univariate covariate). This kernel implies an assumption that units that are close in covariate space have high correlation and the definition of "close" can be controlled through the inverse-length scale parameter ρ. The right panel visualizes the induced correlation structure for a one-dimensional x evaluated on a grid from $(-3, 3)$ with $\rho = 1$ and demonstrates that this correlation is invariant to position. It is clear that this kernel may be useful for modeling "local" phenomenon and that it may not be suitable for capturing long range dependencies

within the data.

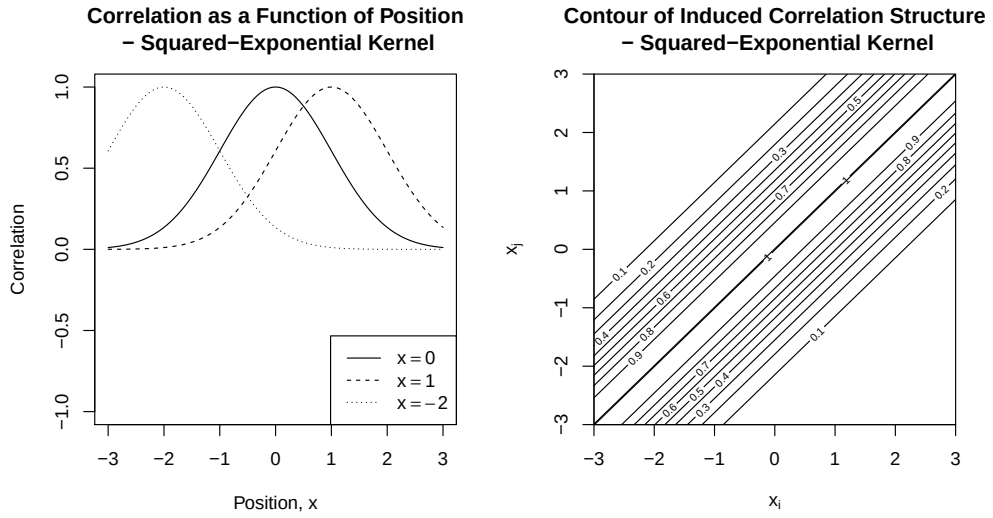

Figure 2.6: Visualization of the properties of the squared exponential covariance function. Left panel: Demonstration of the correlation function induced by the squared exponential kernel with inverse-length scale $\rho = 1$. For the squared exponential kernel the inverse-length scale parameterization implies that higher values correspond to more "local" functions, while values closer to zero demonstrate smoother functions. Right panel: Visualized contour plot of the correlation structure when $\rho = 1$. Demonstrates the "structure" in this covariance is completely local and therefore the model has a difficult time modeling long-range dependencies.

Alternatively, consider the linear kernel,

$$k_{lin}(x, x'; \sigma_0) = \sigma_0 + x^T x'. \tag{2.60}$$

This kernel is from the class of *dot product* kernels and is the GP prior that corresponds to priors on linear functions, i.e. linear regression. Figure 2.7 shows the correlation function induced by the linear kernel function between a point x and all other points x' for three choices of $x \in \Re$ (in this example σ_0 was set to 0.5). The figure also provides a contour plot demonstrating the induced correlation at all points. Comparing with Figure 2.6, we see that this kernel allows longer range dependencies and that the correlation structure changes depending on where the function is centered. Note that if $\sigma_0 \gg x^T x'$, then the kernel function

is approximately a constant and this may cause computational issues.

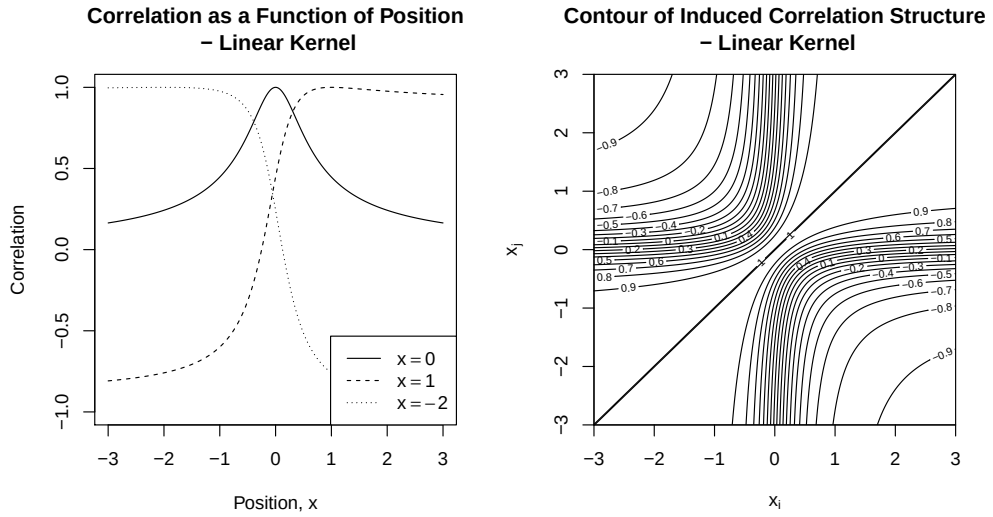

Figure 2.7: Visualization of the properties of the linear covariance function. Left panel: Demonstration of the correlation between a point x and all other points on the evaluated grid. Right panel: Visualized contour plot of the correlation structure when $\sigma_0 = 0.5$. Demonstrates that the "structure" in this covariance allows for longer range dependencies.

Finally, since realizations from a GP are distributed as a multivariate normal distribution, we can simulate from a GP by constructing the covariance matrix then sampling from a multivariate normal distribution with mean zero and covariance constructed using the defined $k(x, x')$. The upper panels of Figure 2.8 demonstrate functions drawn from a GP model with squared exponential covariance function with $\rho = 1$ (Left) and draws from a GP model with a linear covariance function with $\sigma_0 = 0.5$ (Right). Since we are ultimately interested in transformations of these functions for binary regression, the lower panel shows each function transformed through the probit link function. Functional draws using the squared exponential further demonstrate the flexibility of these models, while the linear kernel provides familiar linear predictors as with standard binary regression models.

There are many kernels that have been developed for capturing varying types of dependencies within an analysis. A sampling of common kernels is provided in Table 2.4. Not listed are

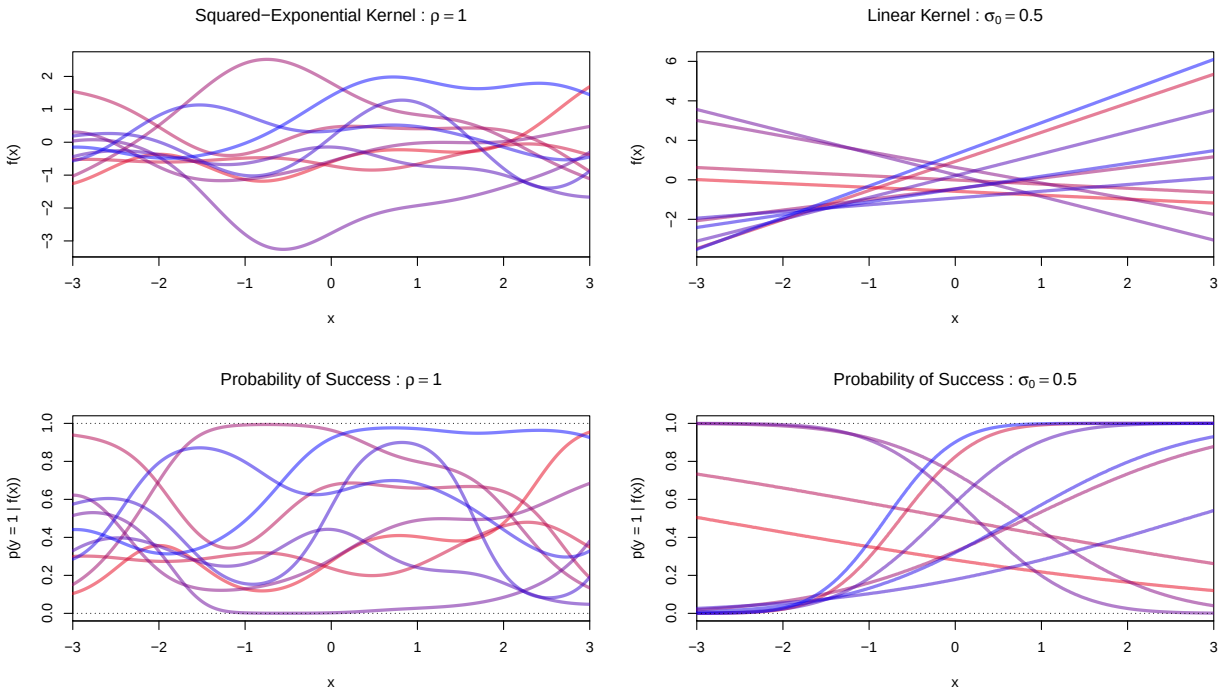

Figure 2.8: Upper Row: Draws from Gaussian processes for a grid of a univariate x, using kernel functions which are a squared exponential with $\rho = 1$ (Left) and a linear polynomial kernel with $\sigma_0 = 0.5$ (Right). Lower panel: Each simulated draw from the upper panel transformed through a probit link function.

kernels for more exotic inputs such as strings, or other mathematical objects such as graphs. Table 2.4 is not meant to be exhaustive, but merely provides a summary of common kernels. One of the convenient features of kernels is the ability to design new kernels from old kernels in order to achieve specific behavior. Consider the following useful propositions from Schölkopf and Smola (2001) and Rasmussen and Williams (2006),

Proposition 2.1 (Sums of Kernels (Schölkopf and Smola, 2001))**.** *If $k_1(x, x')$ and $k_2(x, x')$ are kernels and $\alpha_1, \alpha_2 > 0$, then $\alpha_1 k_1(x, x') + \alpha_1 k_2(x, x')$ is a valid kernel*

Proposition 2.2 (Products of Kernels (Schölkopf and Smola, 2001))**.** *If $k_1(x, x')$ and $k_2(x, x')$ are kernels, then $(k_1 k_2)(x, x') = k_1(x, x')k_2(x, x')$ is a valid kernel.*

Proposition 2.3 (Normalized Kernels (Rasmussen and Williams, 2006))**.** *If $k(x, x')$ is a kernel,*

then $\tilde{k}(x, x') = \frac{k(x,x')}{\sqrt{k(x,x)}\sqrt{k(x',x')}}$ *is a valid kernel.*

In particular, normalized kernels are useful in binary (or multinomial) regression settings to standardize a covariance matrix so that it has unit variance on the diagonal.

Kernel Name	$k(x, x'; \theta)$	**Parameters**, θ	**Notes**
White-Noise	$\delta(x, x')$	-	$\delta(\cdot,\cdot) \equiv$ Kronecker delta function
Constant	σ^2	$\sigma \in \Re^+$	
Exponential	$\exp\left(-\frac{d(x,x')}{\ell}\right)$	$\ell \in \Re^+$	
Squared Exponential	$\exp\left(-\frac{d(x,x')^2}{2\ell^2}\right)$	$\ell \in \Re^+$	
γ-Exponential	$\exp\left(-\left[\frac{d(x,x')}{\ell}\right]^{\gamma}\right)^{-\alpha}$	$\gamma \in (0, 2], \ell \in \Re^+$	
Rational Quadratic	$\left(1 + \frac{d(x,x')^2}{2\alpha\ell^2}\right)$	$\alpha, \ell \in \Re^+$	
Matern Class	$\frac{2^{1-\nu}}{\Gamma(\nu)}\left(\frac{\sqrt{2\nu}d(x,x')}{\ell}\right)^{\nu} K_\nu\left(\frac{\sqrt{2\nu}d(x,x')}{\ell}\right)$	$\nu, \ell \in \Re^+$	$K_\nu(\cdot)$ is a modified Bessel function.
Linear	$c + x^T x'$	$c \in \Re^+$	
Polynomial	$\left(c + x^T x'\right)^p$	$c \in \Re^+, p \in \mathbb{Z}$	
Periodic	$-\frac{2}{\ell^2}\sin^2\left(\frac{\pi}{p}(x - x')\right)$	$\ell \in \Re^+, p \in \Re$	One-Dimensional x

Table 2.4: Table of common kernel functions for $x, x' \in \Re^d$ (unless otherwise specified). In the above table $d(x, x')$ is defined to be $d(x, x') = ||x - x'|| = \sqrt{(x - x')^T(x - x')}$. In many kernels in the above the parameter ℓ is referred to as the length-scale of the kernel, while an alternative parameterization is typically the *inverse-length scale* defined to be $\rho = 1/\ell$.

The multiplicative property of kernels can be used to derive what is often referred to as a squared exponential kernel with *automatic relevance detection* (ARD). Consider modeling the covariance in each dimension d using a squared exponential kernel function, i.e., for each x_d, x'_d, the covariance is

$$k_{SE,d}(x_d, x'_d; \ell_d) = \exp\left(-\frac{1}{2}\frac{(x_d - x'_d)^2}{\ell_d^2}\right). \tag{2.61}$$

This kernel can be combined across all dimensions to construct the *ARD* kernel,

$$k_{ARD}(x, x'; \{\ell_d\}) = \prod_{d=1}^{D} \exp\left(-\frac{1}{2}\frac{(x_d - x'_d)^2}{\ell_d^2}\right) \tag{2.62}$$

$$= \exp\left(-\frac{1}{2}\sum_{d=1}^{D}\frac{(x_d - x'_d)^2}{\ell_d^2}\right). \tag{2.63}$$

Figure 2.9 demonstrates the potential utility of adding or multiplying kernels by demonstrating the correlation structure that can be achieved through addition or multiplication of kernels in a single dimension. The upper panels demonstrate the correlation structure of an additive squared exponential plus linear kernel function. This panel demonstrates how the local structure of the SE kernel can be added to the longer range dependency from the linear kernel. The set of lower plots in this figure demonstrates a multiplicative kernel of the same base kernels and the effect on the correlation structure. For completeness, Figure 2.10 provides function draws from a zero mean GP with covariance that is a function of a summation of a squared exponential kernel and a linear kernel (left) and a multiplication of a squared exponential kernel and a linear kernel (right).

The utility of each kernel depends on the application at hand. Chapter 13 of Schölkopf and Smola (2001) provides the theoretical foundations for designing and implementing new kernels, while Duvenaud (2014) provides an excellent overview of the appropriateness of different kernels for various modeling scenarios for Gaussian processes.

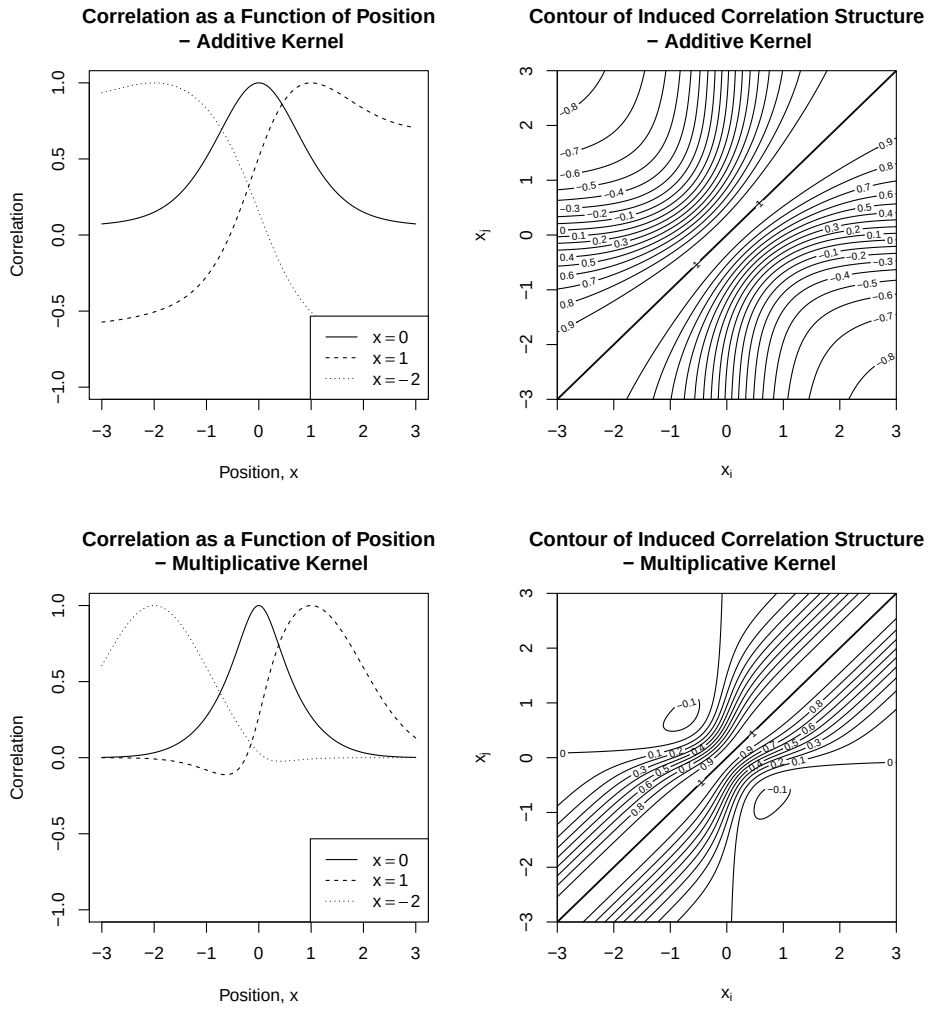

Figure 2.9: Visualization of the properties of additive (upper row) and multiplicative (lower row) kernel functions. Left panel: Demonstration of the correlation between a point x and all other points on the evaluated grid. Right panel: Visualized contour plot of the correlation matrix.

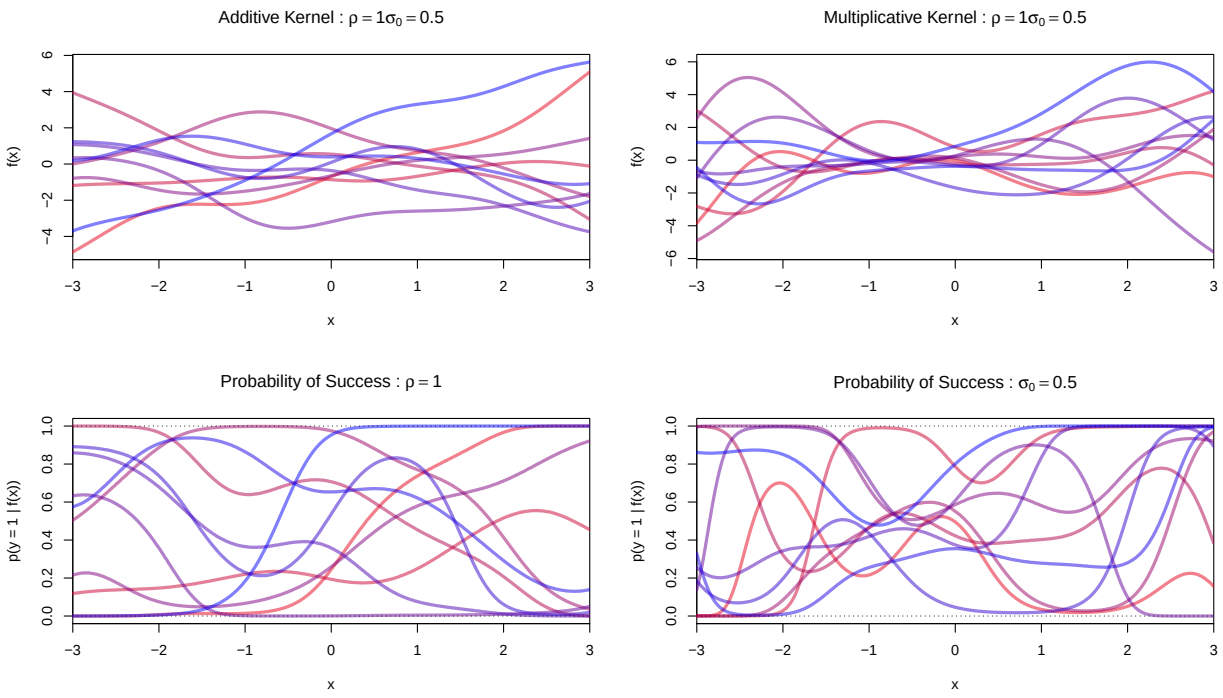

Figure 2.10: Upper panel: Simulated draws from a Gaussian process with additive (left) and multiplicative (right) kernel functions and univariate x. Lower panel: Simulated draws transformed through the probit link function.

2.2.3 Examples - Binary Regression with Gaussian Processes

To better understand the behavior of the posterior estimates from the methods developed in the previous section, we present a brief simulation study. Consider a set of 150 observed labels $\mathbf{Z}$ for cases which we observe a univariate covariate $X_i \sim \mathcal{N}(0, 1)$. Additionally, we model $P(Z_i = 1|X_i = x_i) = \Phi(f_i)$ where,

$$\mathbf{f}|\mathbf{X} \sim GP(\mathbf{0}, \mathbf{K}_\theta) \tag{2.64}$$

with

$$\mathbf{K}_{ij} = k(x_i, x_j; \theta) = (x_i x_j + \sigma_0)^2 \tag{2.65}$$

where $\sigma_0 = 2$. Values of $\Phi(f_i)$ and Z_i are plotted against X_i in Figure 2.11.

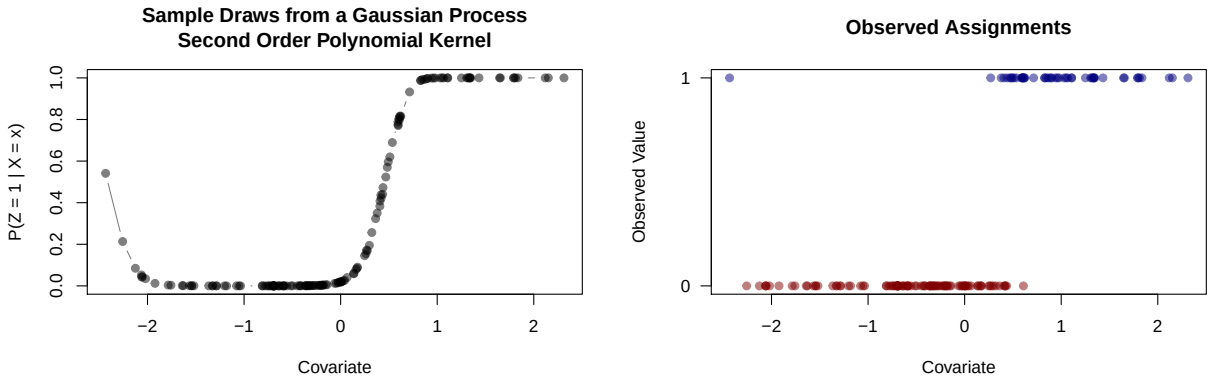

Figure 2.11: Left Panel: Sample draw of observed probabilities using a predictor that was drawn from a Gaussian process with a second-order polynomial kernel. Right Panel: Observed labels simulated based upon the observed probabilities in the left panel.

As $\mathbf{f}$ were drawn from a GP with known covariance matrix, we can find the approximate posterior distribution using Laplace approximation, expectation propagation, and as the sample size is small we can sample from the true posterior distribution using Markov Chain Monte

Carlo with fixed hyperparameters [2]. Figure 2.12 visualizes the posterior distributions and their 95% credible intervals using each method. In each figure, the dark area shape represents the 95% credible interval of the MCMC draws and the dark line is the posterior median, while the dotted lines represent the median and 95% credible intervals of the method named. Both methods of approximating the posterior distribution demonstrate nice agreement with the posterior median of the MCMC samples. The difference appears in the variance terms, where the credible intervals agree for the MCMC draws and EP, but there are small discrepancies using the Laplace approximation. Investigating this further, we look at the marginal posterior distributions of a single latent score f_i from this sample using each method, visualized in Figure 2.13. We see that the true posterior from the MCMC samples is skewed, while both approximation methods utilize a Gaussian approximation to this skewed distribution. The EP approximation places most of the mass over the "long-tail", while the Laplace approximation places the most of its mass over the mode of the true posterior. This was one of the primary findings of Kuss and Rasmussen (2005) and explains the difference in the posterior credible intervals.

Finally, we can learn the hyperparameters θ by maximizing the approximate log marginal likelihood functions for each method; these are shown in left side of Figure 2.14. We see that both methods appear to attain their maximum at the same point (the black dot located on each line). We can then plot the posterior mean estimates of the probability function along with the true values for a comparison. We see that in each instance the estimates agree well with the true probability of the class label.

[2] For 150 observations, finding the approximate posterior with Laplace approximation and expectation propagation both take under one second, while MCMC finishes 25000 sample draws in approximately 800 seconds, demonstrating the scalability issue related to MCMC for binary regression with GPs

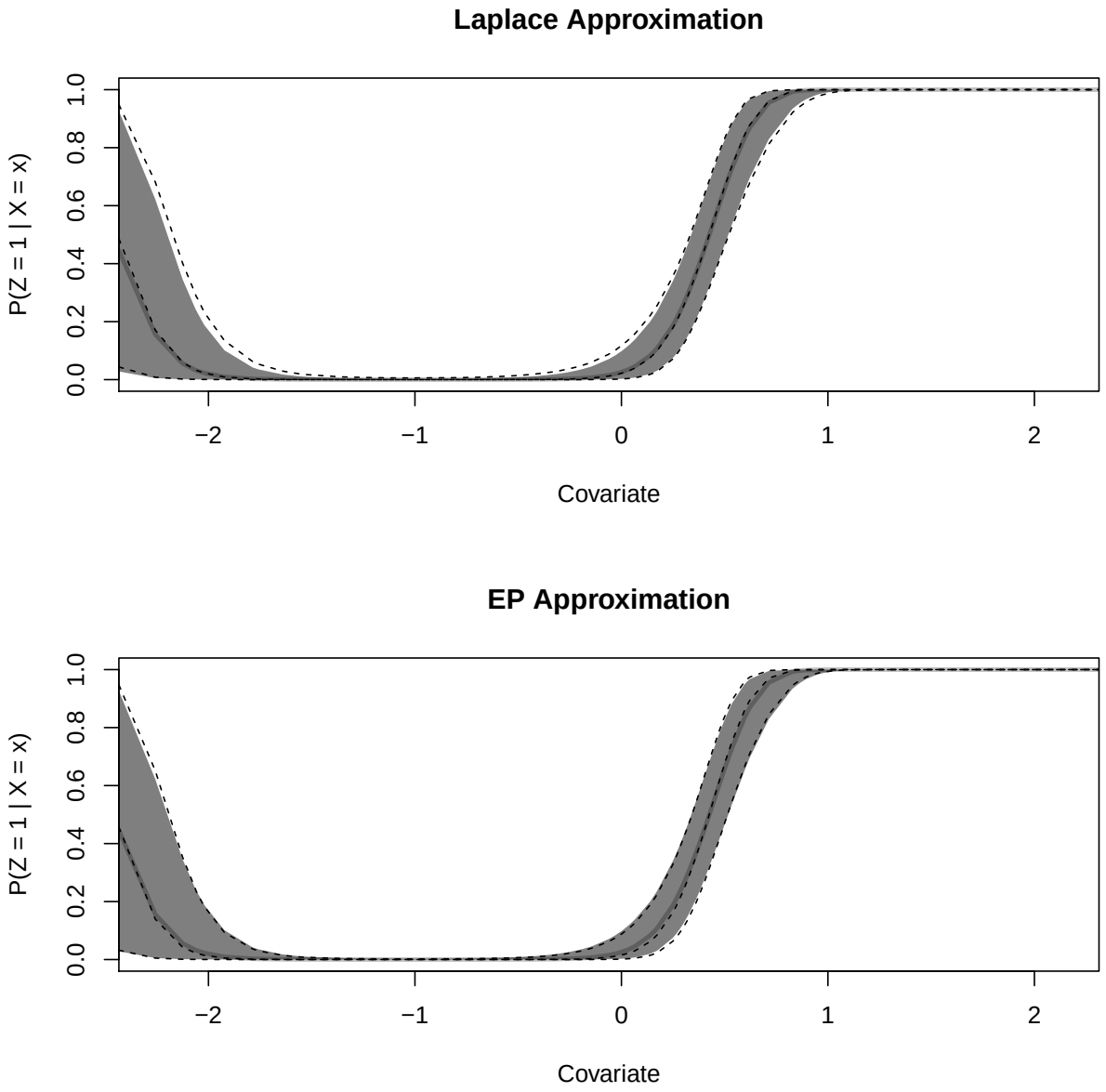

Figure 2.12: Comparing the point-wise 95% credible intervals among the methods. The left panel compares the credible intervals based on the MCMC samples and the Laplace approximation (LA) to the posterior distribution. The dotted lines represent the LA and the solid dark area is the MCMC credible intervals. The right panel is similar but comparing MCMC against expectation propagation.

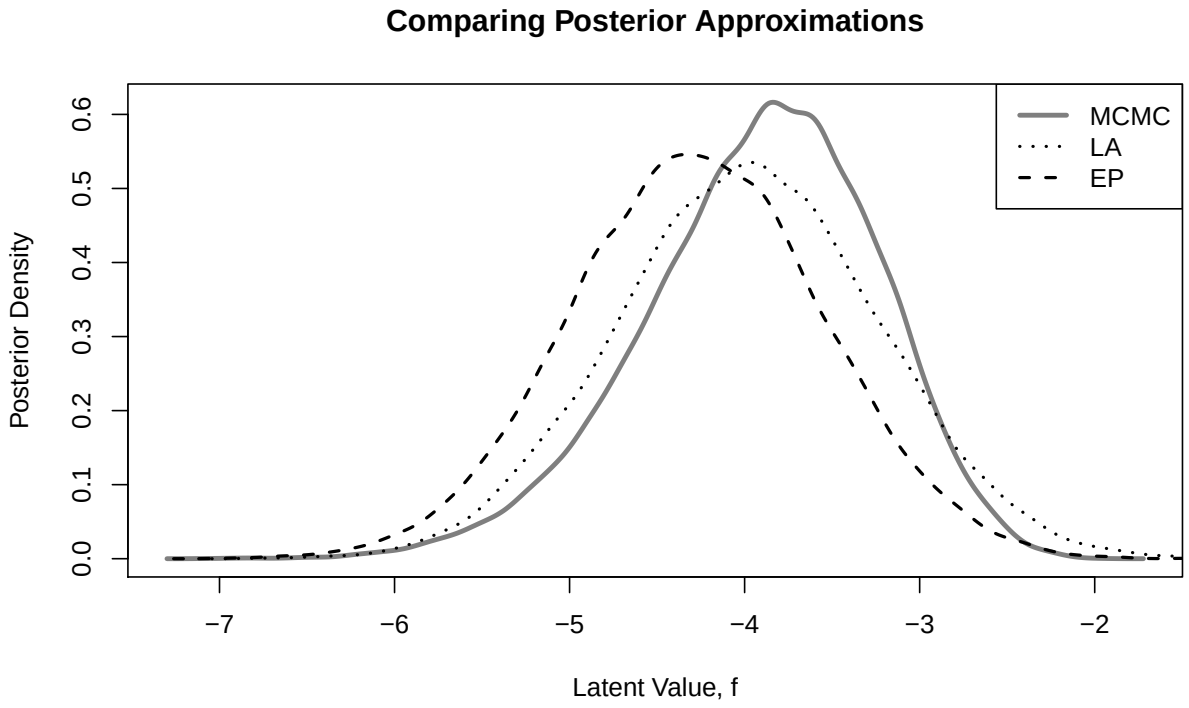

Figure 2.13: Comparing the posterior distribution from MCMC sampling to the Laplace and expectation propagation approximations. The posterior distribution appears to be skewed, while both approximation methods utilize a Gaussian approximation. Note that the expectation propagation approximation places most its mass on the "long-tail", while the Laplace approximation is centered closer to the mode.

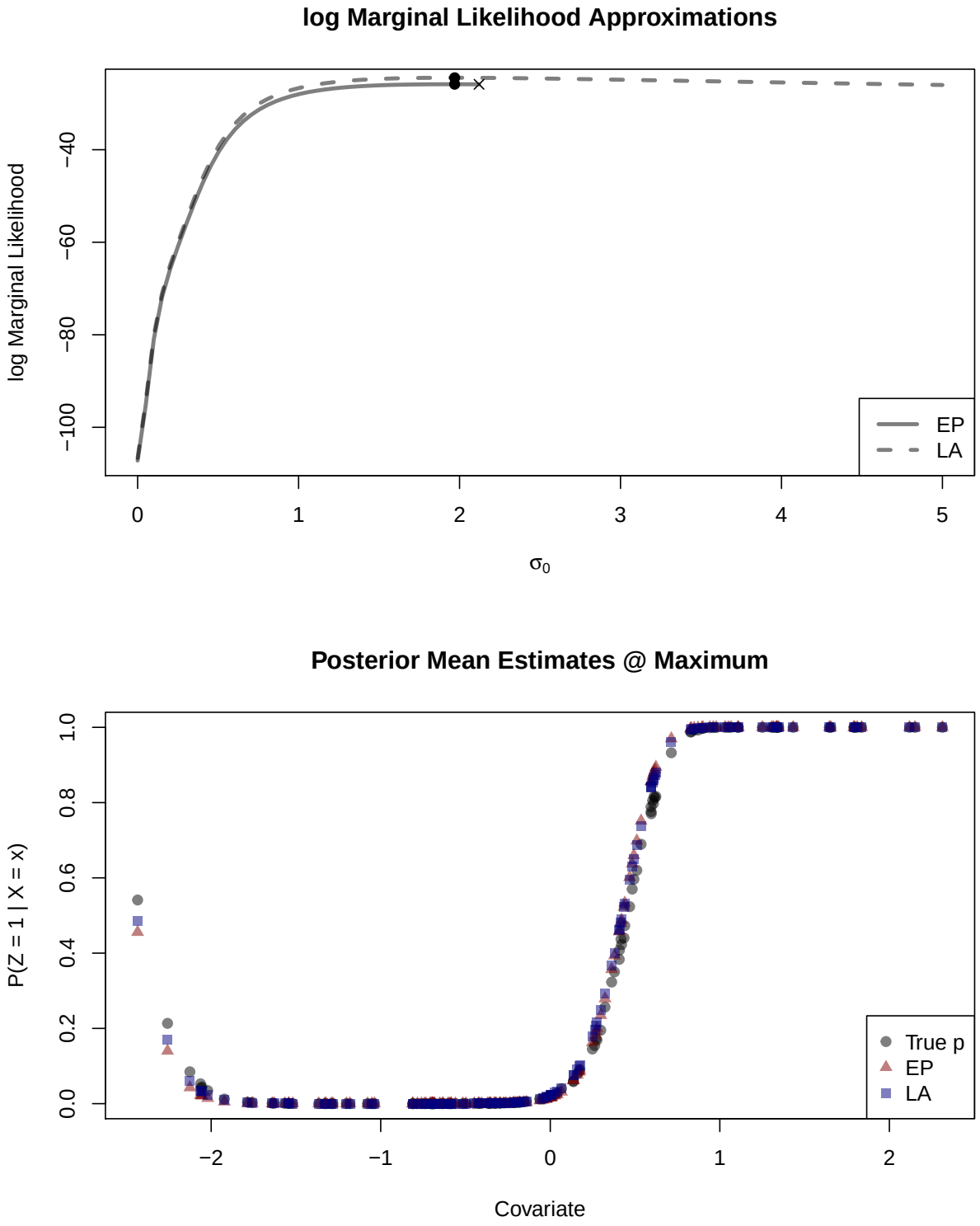

Figure 2.14: Left panel: Log-marginal likelihood curves based upon each approximation method. Both methods attain their maxima at approximately the same point. Right panel: Posterior expectations of the probability function for each point compared among the methods.

2.2.4 Multinomial Regression with Gaussian Processes

The majority of Section 2.2 has focused on modeling random variables Z with two categories; in this subsection we extend to multiple categories largely following the methods outlined in Rasmussen and Williams (2006) Section 3.5.

Consider a categorical random variable Z with C categories such that $Z = z \in \{0, 1, \ldots, C-1\}$. We make the assumption that for each category z, there exists a function $f^z : \Re^d \to \Re$ which takes as inputs $x_i \in \Re^d$ such that each $f^z(x_i) \equiv f_i^z$ is from a Gaussian process. That is, using the notation defined previously, for each category there is a latent Gaussian process such that,

$$\mathbf{f}^z|\mathbf{X} \sim \mathcal{GP}(\mathbf{m}^z, \mathbf{K}^z). \tag{2.66}$$

We also make the simplifying assumption that the functions are independent and therefore write the full process as,

$$\mathbf{f}|\mathbf{X} = \begin{pmatrix} \mathbf{f}^0 \\ \mathbf{f}^1 \\ \vdots \\ \mathbf{f}^{C-1} \end{pmatrix} \mid \mathbf{X} \sim \mathcal{GP}\left(\begin{pmatrix} \mathbf{m}^0 \\ \mathbf{m}^1 \\ \vdots \\ \mathbf{m}^{C-1} \end{pmatrix}, \begin{pmatrix} \mathbf{K}^0 & \mathbf{0} & \ldots & \mathbf{0} \\ \mathbf{0} & \mathbf{K}^1 & \ldots & \mathbf{0} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{0} & \mathbf{0} & \ldots & \mathbf{K}^{C-1} \end{pmatrix}\right) \tag{2.67}$$

These latent values can then be used to define the probability of observing each category z where,

$$P(Z_i = z|X_i = x_i) = \frac{\exp\left(f_i^z(x)\right)}{\sum_{z'} \exp\left(f_i^{z'}(x)\right)} \tag{2.68}$$

$$= \frac{\exp\left(f_i^z\right)}{\sum_{z'} \exp\left(f_i^{z'}\right)}. \tag{2.69}$$

Function 2.69 is often referred to as the softmax function (Rasmussen and Williams, 2006), and the f_i^z terms are logarithmic probabilities (McCullagh and Nelder, 1989). We note there is a formulation of this model with a link function that is similar to the probit link function (as in binary regression), but this is not explored in this dissertation.

To denote that the mean functions or covariance functions are structured via parameters, we write,

$$\mathbf{f}|\mathbf{X}, \Gamma, \Theta \sim \mathcal{GP}\left(\begin{pmatrix} \mathbf{m}^0_{\gamma_0} \\ \mathbf{m}^1_{\gamma_1} \\ \vdots \\ \mathbf{m}^{C-1}_{\gamma_{C-1}} \end{pmatrix}, \begin{pmatrix} \mathbf{K}^0_{\theta_0} & \mathbf{0} & \dots & \mathbf{0} \\ \mathbf{0} & \mathbf{K}^1_{\theta_1} & \dots & \mathbf{0} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{0} & \mathbf{0} & \dots & \mathbf{K}^{C-1}_{\theta_{C-1}} \end{pmatrix}\right) \tag{2.70}$$

$$\mathbf{f}|\mathbf{X}, \Gamma, \Theta \sim \mathcal{GP}(\mathbf{m}_\Gamma, \mathbf{K}_\Theta) \tag{2.71}$$

where $\Gamma = \{\gamma_z\}$ and $\Theta = \{\theta_z\}$ for $z \in \{0, 1, \dots, C-1\}$.

2.2.4.1 Model for Multinomial Regression with Gaussian Processes

Consider a set of observations Z_i for $i = 1, 2, \dots, N$ and a vector of covariates for each observation $X_i = x_i \in \Re^D$. We make the assumption that each Z_i given a vector of covariates X_i can be modeled using a multinomial regression framework and assume that each predictor function is from a Gaussian process. That is,

$$Z_i|X_i = x_i \sim Multinomial\left(1, \left\{p_z = \frac{\exp\left(f^z(x_i)\right)}{\sum_{z'}\exp\left(f^{z'}(x_i)\right)} \equiv \frac{\exp\left(f_i^z\right)}{\sum_{z'}\exp\left(f_i^{z'}\right)}\right\}\right) \tag{2.72}$$

$$\mathbf{f}|\mathbf{X}, \Theta \sim \mathcal{GP}(\mathbf{0}, \mathbf{K}_\Theta) \tag{2.73}$$

As in the binary setting, we make the assumption that the hyperparameters are fixed and for this section assume that they are known.

The likelihood function for this model is,

$$p(Z_i|f_i) = \prod_{i=1}^{N}\prod_{z=0}^{C-1}\left(\frac{\exp\left(f_i^z\right)}{\sum_{z'}\exp\left(f_i^{z'}\right)}\right) \tag{2.74}$$

and thus the posterior distribution of $\mathbf{f}$ given $\mathbf{Z}, \mathbf{X}, \Theta$ is as follows,

$$p(\mathbf{f}|\mathbf{Z}, \mathbf{X}, \Theta) = \frac{p(\mathbf{Z}|\mathbf{f})p(\mathbf{f}|\mathbf{X}, \Theta)}{p(\mathbf{Z}|\mathbf{X}, \Theta)} \tag{2.75}$$

$$\propto \prod_{i=1}^{N}\prod_{z=0}^{C-1}\left(\frac{\exp\left(f_i^z\right)}{\sum_{z'}\exp\left(f_i^{z'}\right)}\right)\det\left(2\pi\mathbf{K}_\Theta\right)^{-1/2}\exp\left(-\frac{1}{2}\mathbf{f}^T\mathbf{K}_\Theta^{-1}\mathbf{f}\right). \tag{2.76}$$

As in the binary setting, this is intractable and needs to be approximated.

2.2.4.2 Laplace Approximation for Multinomial Regression

We again appeal to the Laplace approximation in order to perform inference in the multinomial regression setting. As stated earlier, the method approximates the posterior distribution of the latent values with a multivariate normal distribution centered at the mode of the posterior distribution and with variance related to the second derivative of the log of the posterior distribution. That is,

$$\mathbf{f}|\mathbf{Z}, \mathbf{X}, \theta \dot{\sim} \mathcal{N}\left(\hat{\mathbf{f}}, \left[-\frac{\partial^2}{\partial\mathbf{f}\partial\mathbf{f}^T}\log p(\mathbf{f}|\mathbf{Z}, \mathbf{X}, \Theta)\mid_{\mathbf{f}=\hat{\mathbf{f}}}\right]^{-1}\right) \tag{2.77}$$

where,

$$\hat{\mathbf{f}} = \arg\max_{\mathbf{f}} \log p(\mathbf{f}|\mathbf{Z}, \mathbf{X}, \theta) \tag{2.78}$$

$$= \arg\max_{\mathbf{f}} \left[\left(\sum_{i=1}^{n} \sum_{z=0}^{C-1} \log \left\{ \frac{\exp\left(f_i^z\right)}{\sum_{z'} \exp\left(f_i^{z'}\right)} \right\} \right) - \frac{1}{2} \mathbf{f}^T \mathbf{K}_\theta \mathbf{f} + C \right] \tag{2.79}$$

and can be found using Newton's method. The implementation details can be found in Section 3.5 of Rasmussen and Williams (2006).

2.2.4.3 Alternative Approximations

Similar to the binary regression setting, there are expectation propagation algorithms (among other approximation methods) that make use of probit link functions and have been developed for multinomial regression. Expectation propagation for multinomial probit regression methods are not explored in this dissertation, but examples of such approaches may be found in Girolami and Rogers (2006); Riihimäki et al. (2013).

Chapter 3

Optimally Balanced Gaussian Process Propensity Scores

3.1 Introduction

As demonstrated in the previous chapter, the propensity score is often an important tool within a causal analysis. Introduced in Rosenbaum and Rubin (1983) in the context of study designs with binary treatment regimes, the propensity score is loosely defined as the probability of an individual or unit in a study being in the treated group given a set of pretreatment covariates. Recall from Chapter 2 that Rosenbaum and Rubin (1983) demonstrated, that under certain assumptions, adjustment on the propensity score enables unbiased estimation of the average treatment effect or the average treatment effect among the treated. In this chapter we focus on the development of a nonparametric propensity score estimation framework using Gaussian processes where the hyperparameters of the GP are selected to minimize a metric of covariate imbalance.

The utilization of a GP regression framework relies heavily on the idea that if two individuals' observed covariates are similar, then they should have a similar probability of being assigned to the treatment, or control, group. In this way, using a Gaussian process to capture the correlation among latent values is analogous to matching methods in observational studies that make use of the distance between sets of covariates to pair observations (Stuart, 2010; Rosenbaum, 2010). As shown in Chapter 2, the modeled covariance function of the GP can be parameterized by hyperparameters that affect the covariances among the latent scores. It is natural to select the hyperparameters in order to address one of the key assumptions required for causal inference in observational studies, that of covariate balance.

There is a large literature on estimating the propensity score for binary treatment regimes. A common approach is to utilize a statistical model-building procedure (e.g., logistic regression) that is iterated until a functional form of the propensity score that sufficiently "balances" covariates is obtained (Imbens and Rubin, 2015; Imai and Ratkovic, 2014). Procedures of this type are computationally fast to perform since models are often built in a step-wise fashion. They typically achieve the goal of approximately balancing covariates but, because of their restricted parametric form, may not accurately estimate treatment assignment. Alternative approaches rely on nonparametrically estimating the propensity score through modern statistical learning methods without a need to specify a functional form (Woo et al., 2008; Lee et al., 2010; McCaffrey et al., 2004). These methods are attractive because they can provide accurate estimation of the treatment assignment mechanism while requiring fewer modeling decisions to be made. However, they are often computationally demanding due to their flexibility. In addition to the methodological development, this chapter compares the effectiveness of our methodology against current methods in the literature.

3.2 Methodology

3.2.1 Estimation of Binary Propensity Scores using Gaussian Processes

We utilize the potential outcomes framework of Neyman and Rubin (Splawa-Neyman et al., 1990; Rubin, 1974) for estimating causal effects (see Section 2.1). For each sampled unit i, let $Z_i \in \{0, 1\}$ represent a binary treatment assignment, where $Z_i = 1$ and $Z_i = 0$ represent membership in a treated group and a control group, respectively, and let $X_i = x_i \in \Re^D$ be a random vector of length D representing a set of pretreatment covariates for the unit. $Y_i^{Z=1}$ is defined as the potential response for unit i under the treatment exposure and $Y_i^{Z=0}$ is the potential response for unit i under the control exposure and the observed response is $Y_i^{obs} = I(Z_i = 1)Y_i^{Z=1} + I(Z_i = 0)Y_i^{Z=0}$. Throughout this chapter we focus on treatment effects that are differences in the potential outcomes, $\tau_i \equiv Y_i^{Z=1} - Y_i^{Z=0}$, and estimating the following estimands; the ATE and ATT. We make assumptions of strong ignorability given the covariate vector X and the stable unit treatment value assumption.

We assume that the propensity score can be modeled through a probit transformation of a latent variable f_i, i.e., $e(x_i) = P(Z_i = 1|X_i = x_i) = \Phi(f_i)$, where $\Phi(\cdot)$ is the cumulative distribution function of the standard normal distribution and the dependence of f_i on X_i is left implicit to keep notation simple. For $i = 1, \ldots, N$, let

$$Z_i|X_i = x_i \sim Bernoulli(\Phi(f_i)), \tag{3.1}$$

and make the assumption that the collection of latent scores f_i arise from a Gaussian process. Let $\mathbf{X}$ represent the $N \times D$ matrix of pretreatment covariates for all units and let θ be a vector

of hyperparameters. Then, we model the latent predictor values $\mathbf{f}$ as,

$$\mathbf{f}|\mathbf{X}, \theta \sim GP(\mathbf{0}, \mathbf{K}_\theta), \tag{3.2}$$

where $\mathbf{K}_\theta$ is the covariance of the process and we make the common a priori assumption that the mean of the process is zero.

The covariance matrix $\mathbf{K}_\theta$ describes the 'similarity' between the latent scores for units within the study and is constructed using a structured kernel function, $k(x_i, x_j; \theta)$ for observations i and j with covariates X_i and X_j with hyperparameters θ. There are many specifications for kernel functions that may be chosen, as discussed in Section 2.1. Two common kernel functions for binary regression models are the squared-exponential kernel,

$$k_{se}(x_i, x_j; \rho) = \exp\left\{-\frac{\rho^2}{2}\sum_{d=1}^{D}(x_{i,d} - x_{j,d})^2\right\}, \tag{3.3}$$

and the normalized polynomial kernel,

$$k_{np}(x_i, x_j; \sigma_0, p) = \left(\frac{x_i^T x_j + \sigma_0^2}{\sqrt{x_i^T x_i + \sigma_0^2}\sqrt{x_j^T x_j + \sigma_0^2}}\right)^p. \tag{3.4}$$

To estimate the propensity score, we assume an additive structure for the covariance function using the squared exponential and the normalized polynomial kernel. That is, we construct a kernel of the form

$$k(x_i, x_j; \theta) = k_{se}(x_i, x_j; \rho) + k_{np}(x_i, x_j; \sigma_0, p), \tag{3.5}$$

where $p = p^*$ is fixed to only consider specific polynomial terms and therefore $\theta = (\rho, \sigma_0)$. The normalized polynomial kernel allows for long range dependencies of the latent scores within the data, while the squared exponential kernel captures local variability in the latent

scores.

Within this model both $\mathbf{f}$ and θ are unknown. One approach to inference is to specify a fully Bayesian model by providing a prior distribution for θ and then sampling from the posterior distribution of $\mathbf{f}, \theta|\mathbf{Z}, \mathbf{X}$ using Markov Chain Monte Carlo (MCMC). What complicates this approach is that causal applications require a specification of the propensity score that is also a balancing score, and therefore only vectors of $\mathbf{f}$ that balance covariates are helpful. Therefore in practice it is adequate to find θ such that a functional of $\mathbf{f}|\mathbf{Z}, \mathbf{X}, \theta$ balances covariates (e.g., a function of $E(\mathbf{f}|\mathbf{Z}, \mathbf{X}, \theta)$). One way to choose such a θ would be to maximize a form of the marginal likelihood with respect to θ where the latent scores have been integrated out of the joint distribution of $\mathbf{f}, \mathbf{Z}|\mathbf{X}, \theta$. This approach indirectly attempts to balance covariates by relying on the fact that an unbiased estimate of the propensity score should be balance covariates. Our proposed approach is more direct in that we model $\mathbf{f}|\mathbf{Z}, \mathbf{X}, \theta$ and select the value of θ that yields propensity score estimates, i.e., $\Phi(E(\mathbf{f}|\mathbf{Z}, \mathbf{X}, \theta))$, that minimizes a measure of covariate imbalance.

Given a value of θ, the conditional posterior distribution $\mathbf{f}|\mathbf{Z}, \mathbf{X}, \theta$ is intractable as demonstrated in the previous chapter and can be approximated using (1) Laplace approximation; (2) MCMC sampling; or (3) expectation propagation. In this chapter we utilize an efficient parallel version of expectation propagation (Van Gerven et al., 2010; Tolvanen et al., 2014) and Laplace approximations for their computational efficiency relative to MCMC, demonstrated in Section 2.2.3. Using the approximation of the distribution $\mathbf{f}|\mathbf{Z}, \mathbf{X}, \theta$, the propensity score can be estimated from its mean, $\hat{e}(x_i) = \Phi(E(f_i|\mathbf{Z}, \mathbf{X}, \theta))$. The results in this section assume that θ is already known, Section 3.2.3 describes our approach for selecting θ in order to minimize covariate imbalance.

3.2.2 Measuring Covariate Imbalance

To assess the adequacy of propensity score estimates we measure the degree to which the estimates balance the distributions of the pretreatment covariates in the treated and control groups. Here we develop a weighted measure of covariate imbalance as a function of the moments of the covariate distributions conditioned upon treatment type. Let $X_{i,d}$ be the d^{th} dimension of the covariate vector for individual i and let $\mathbf{w}^* = \{w_i^*\}$ be a vector of weights. For each dimension of the covariate space corresponding to a continuous covariate, we define

$$\bar{X}_{Z=z,d} = \frac{\sum_{i=1}^{N} w_i^* I(Z_i = z) X_{i,d}}{\sum_{i=1}^{N} w_i^* I(Z_i = z)} \quad \text{and} \quad s^2_{Z=z,d} = \frac{\sum_{i=1}^{N} w_i^* I(Z_i = z)(X_{i,d} - \bar{X}_{Z=z,d})^2}{\sum_{i=1}^{N} w_i^* I(Z_i = z)} \tag{3.6}$$

as the weighted mean and weighted sample variance for each covariate in each group, respectively (see Section 2.1 for a review). For binary covariates let $\bar{X}_{Z=z,d}$ be defined in the same way but let $s^2_{Z=z,d} = \bar{X}_{Z=z,d}(1 - \bar{X}_{Z=z,d})$. In this work, ordinal covariates are treated as continuous covariates and categorical covariates are transformed to binary covariates using dummy variables. The weights, w_i^*, are defined in two different ways, depending on the estimand of interest; for the ATE and the ATT they are defined as follows (Hirano et al., 2003; Stuart, 2010; Li et al., 2017):

$$w_i^{ATE} = \frac{Z_i}{\hat{e}(x_i)} + \frac{1 - Z_i}{1 - \hat{e}(x_i)} \qquad \text{and} \qquad w_i^{ATT} = Z_i + (1 - Z_i)\frac{\hat{e}(x_i)}{1 - \hat{e}(x_i)}. \tag{3.7}$$

The weighted means and variances defined in Equation (3.6) can be used to assess covariate balance. Chapter 2 defined the standardized difference in the means and the logarithm of

the ratio of the sample standard deviations as

$$\Delta_d = \frac{\bar{X}_{Z=1,d} - \bar{X}_{Z=0,d}}{\sqrt{(s^2_{Z=1,d} + s^2_{Z=0,d})/2}} \qquad \text{and} \qquad \Gamma_d = \log(s_{Z=1,d}) - \log(s_{Z=0,d}), \tag{3.8}$$

respectively. When the propensity score adequately balances the conditional distributions of the pretreatment covariates, both $|\Delta_d|$ and $|\Gamma_d|$ should be small. We operationalize this notion by defining $\mathcal{CB}_d$ as the covariate imbalance in dimension d, where

$$\mathcal{CB}_d = \begin{cases} |\Delta_d|^2 + |\Gamma_d|^2 & \text{if } X_d \text{ is a continuous covariate} \\ |\Delta_d|^2 & \text{if } X_d \text{ is a binary covariate} \end{cases} \tag{3.9}$$

and use a measure of overall covariate imbalance that is a sum of the covariate imbalances in each dimension,

$$\mathcal{CB} = \sum_{d=1}^{D} \mathcal{CB}_d. \tag{3.10}$$

Two choices made in the proposed definition of $\mathcal{CB}$ deserve further explanation. First, the balance metrics used for continuous and binary distributions differ. The purpose of a measure of covariate imbalance is to quantify the total difference between the moments of the distributions of the pretreatment covariates in the treated and control groups. The distribution of a binary covariate is completely defined by the first moment and therefore including a Γ_d term would have the effect of overemphasizing the first moment so that a binary covariate appears "more important" than a continuous variable. In simulation studies this negatively impacted performance. The variance terms, Γ_d, are needed for continuous covariates to reduce the bias that would arise if the distributions were not balanced with respect to the second moment. A second feature of our measure is that each term in $\mathcal{CB}_d$ is squared. The typical advice is to evaluate covariate balance by considering the absolute value of each term

(Imbens and Rubin, 2015). Squaring each term in $\mathcal{CB}_d$ penalizes solutions that allow some dimensions of imbalance to remain large while others approach zero. In simulations this was found to improve performance by ensuring that no dimensions have substantial imbalance. This is analogous to the difference between minimizing absolute error loss and minimizing squared error loss in multivariate estimation problems.

3.2.3 Minimizing Covariate Imbalance

Section 3.2.1 described our approach for estimating propensity scores given hyperparameter θ. Section 3.2.2 defined a measure of the overall covariate imbalance, $\mathcal{CB}$, which is a function of the estimated propensity scores (and hence a function of θ). We propose to find a value of θ which minimizes total covariate imbalance,

$$\theta_{opt} = \arg\min_{\theta} \mathcal{CB}(\theta). \tag{3.11}$$

The function $\mathcal{CB}(\theta)$ relies on an approximation of $E(\mathbf{f}|\mathbf{T}, \mathbf{X}, \theta)$, and therefore derivatives of $\mathcal{CB}(\theta)$ are not easily obtained. Therefore to optimize covariate imbalance we utilize a derivative-free optimization routine called "Bounded Optimization BY Quadratic Approximation (BOBYQA)" defined in Powell (2009) and efficiently implemented in the `R` package `minqa`. The algorithm is a method that optimizes a function without the specification of analytical derivatives and is a bounded optimization routine that allows for specifying regions of the parameter space that are valid. The algorithm requires starting estimates for θ. If all covariates have been standardized (i.e., mean zero and variance of one for continuous covariates and binary covariates transformed to $X_i \in \{1, -1\}$), then the initial value $\theta_{init} = (1, 1)$ has provided satisfactory performance in our simulations. Once a vector of θ has been found that minimizes covariate balance, each propensity score is estimated as $\hat{e}(x_i) = \Phi(E(f_i|\mathbf{Z}, \mathbf{X}, \theta_{opt}))$

and an estimate of a treatment effect can be found using the nonparametric weighting estimators defined in Equation 2.7, and shown below for a difference in potential outcomes:

$$\hat{\tau}^* = \left(\frac{\sum_{i=1}^{N} w_i^* I(Z_i = 1) Y_i^{obs}}{\sum_{i=1}^{N} w_i^* I(Z_i = 1)} \right) - \left(\frac{\sum_{i=1}^{N} w_i^* I(Z_i = 0) Y_i^{obs}}{\sum_{i=1}^{N} w_i^* I(Z_i = 0)} \right) \tag{3.12}$$

with $\mathbf{w}^*$ chosen for an appropriate estimand of interest, defined in Table 2.2 and repeated in equation 3.7 for the ATE and the ATT. [1]

3.3 Simulations

We provide a simulation study to investigate the performance of our method and compare its effectiveness in estimating treatment effects against other propensity score estimation methods. Section 3.3.1 describes the potential outcome settings for each simulation. Section 3.3.2 provides a simple comparison of the two approximation methods discussed in this Chapter, comparing the optimally balanced Gaussian process propensity scores found using expectation propagation and Laplace approximation. Section 3.3.3 is a comparative study that is focused on estimating the ATE. In this setting we compare results with the true propensity score (as it is defined within the simulation), as well as against other propensity score estimation methods, to assess the relative performance of our approach. Section 3.3.4 focuses on simulation results for estimating the ATT. Each section considers estimating treatment effects under three potential outcome models: (1) outcomes that are linearly related to a covariate with a constant treatment effect, (2) potential outcomes where the difference implies a non-constant treatment effect (effect modification by a covariate) that

[1] Treatment effects can be also estimated through weighted least squares. See Imbens and Rubin (2015) Section 17.8. These estimators are often referred to "doubly-robust" if either the function form of the regression function is correct, or if the propensity score is estimated correctly. The variance of the estimated treatment effect can then be found through typical "sandwich"-type variance estimators.

is a function of polynomial terms, and (3) potential outcomes where the difference implies a non-constant treatment effect but the difference is a function of exponentials. This last potential outcome setting demonstrates the general utility of the method, even though it only considers matching the first two moments of a covariate in our balancing procedure. In Section 3.3.4, an additional simulation setting is considered for estimating the ATT that contains a higher dimension covariate space and more complicated potential outcomes to further demonstrate performance.

3.3.1 Potential Outcomes Settings

The potential outcome models, which are a function of a continuous covariate X_1, are described in Table 3.1 and visualized in Figure 3.1 for a sample of X values drawn from $\mathcal{N}(0, 1)$. The error terms were simulated such that $\epsilon_{Z=z,i} \sim N(0, 0.5^2)$ for $z \in \{0, 1\}$ for each unit i. In total each simulation study consists of 1000 simulated data sets, where each simulated data set is comprised of 500 observations.

Potential Outcome Setting	Treatment Response	Control Response
1) Constant Treatment Effect	$Y^{Z=1} = X_1 + 3 + \epsilon_{Z=1}$	$Y^{Z=0} = X_1 + \epsilon_{Z=0}$
2) Effect Modification	$Y^{Z=1} = X_1^2 + 2 + \epsilon_{Z=1}$	$Y^{Z=0} = X_1 + \epsilon_{Z=0}$
3) Nonlinear Effect Modification	$Y^{Z=1} = \exp(X_1) + 4X_1 + 3 + \epsilon_{Z=1}$	$Y^{Z=0} = -X_1^2 - \exp(X_1) + \epsilon_{Z=0}$

Table 3.1: Models for potential outcomes

3.3.2 Laplace Approximation vs. Expectation Propagation

In this section, we compare the relative performance of the optimally balanced Gaussian process propensity score based on a Laplace approximation and based on expectation propagation. We consider a simple simulation where the potential outcomes are outlined in Section 3.3.1, and where the true propensity score is $e(x) = \Phi(0.75x_1)$. In each simulated data set,

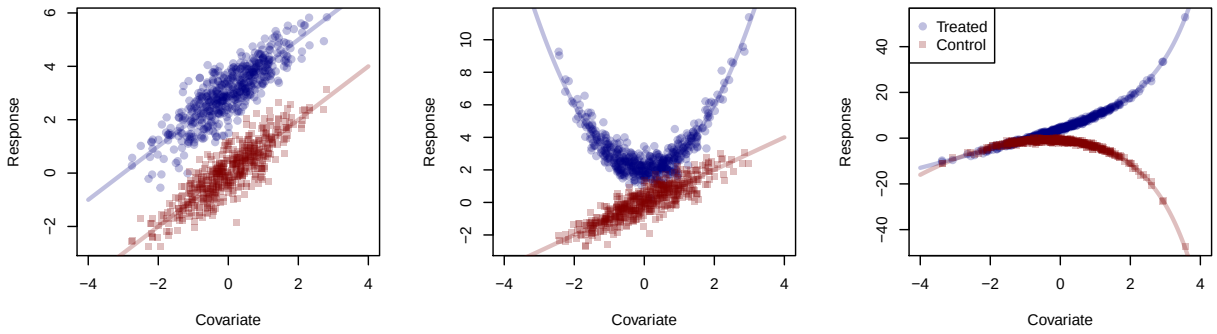

Figure 3.1: Visualization of Potential Outcome Response Functions in the Treated and Control Groups. The left figure is a setting where the potential outcomes are linearly related to the continuous covariate X_1; the center is where the potential outcomes are related to the continuous covariate, but the treatment effect is not constant; the furthest right is a more extreme example of non-linear treatment effects.

we generate $X_1 \sim \mathcal{N}(0, 1)$ and then simulate $Z|X = x \sim Bernoulli(e(x))$.

For each simulation, we estimate the propensity score using the methodology laid out in the previous sections utilizing each approximation method. The kernel that was chosen for this simulation was the additive normalized-polynomial squared-exponential kernel where the order of the polynomial was one. These propensity score estimates were then used to create weights for estimating the ATE, and then used to compare measures of covariate balance, bias, and mean squared error. To compare covariate balance, the mean standardized difference in the means was evaluated across simulations. To assess performance in estimating the ATE we compare estimates where there was no weighting adjustment on the propensity score, weighting adjustment by the true propensity score, and weighting adjustment using estimates based upon the Laplace approximation and the EP approximation. Figure 3.2 contains the distributions of the standardized difference in the means using weights based upon the true propensity score, the Laplace approximation propensity score, and finally the expectation propagation propensity score. It is demonstrated in this figure that under this setting, expectation propagation provides better covariate balance as compared with the

Laplace approximation, while both methods provide better balance than adjustment using the true propensity score weights.

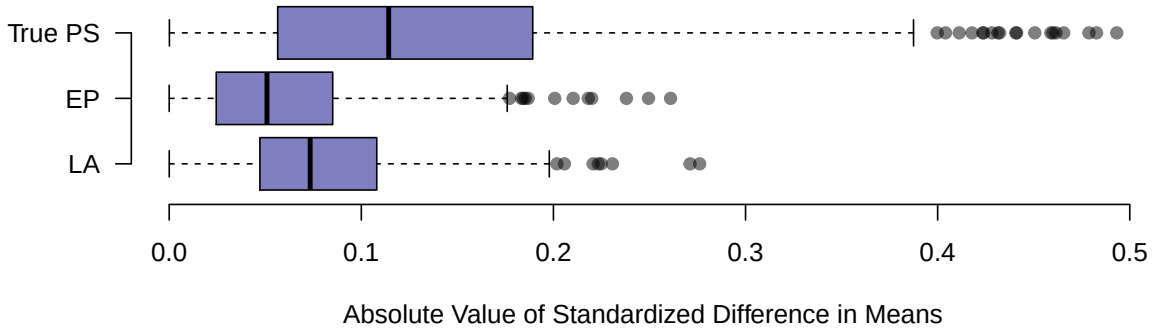

Figure 3.2: Comparing mean balance metrics for the standardized difference in the means. Each distribution represents the absolute value of the standardized difference in the means based upon ATE weights and each adjustment method.

Figure 3.3 demonstrates the relationship between the observed treatment effect within each simulation, i.e. $\tau_{sim} = \frac{1}{N}\sum_{i=1}^{N}\left(Y_i^{Z=1} - Y_i^{Z=0}\right)$, and the estimated treatment effect based upon weights from LA and from EP, for each potential outcome setting. Each row of the figure represents a potential outcome setting and each dot represents results from a simulated data set. The first column of the figure represents the relationship between τ_{sim} and estimates of the ATE based upon weights constructed from LA estimates, the second column is similar and based upon EP weights, and the final column shows the relationship between estimates based upon LA and EP. The first two columns demonstrate that the methods provides results that agree with the simulated values and that there is a high level of agreement between the two methods.

Results comparing empirical values of the mean bias, mean absolute bias, the standard error of the estimates, and the mean squared error are contained in Table 3.2. As is shown in the table, there is general agreement between the two methods for estimating treatment effects. In this setting it appears that LA performs best, but later sections show that this is not always

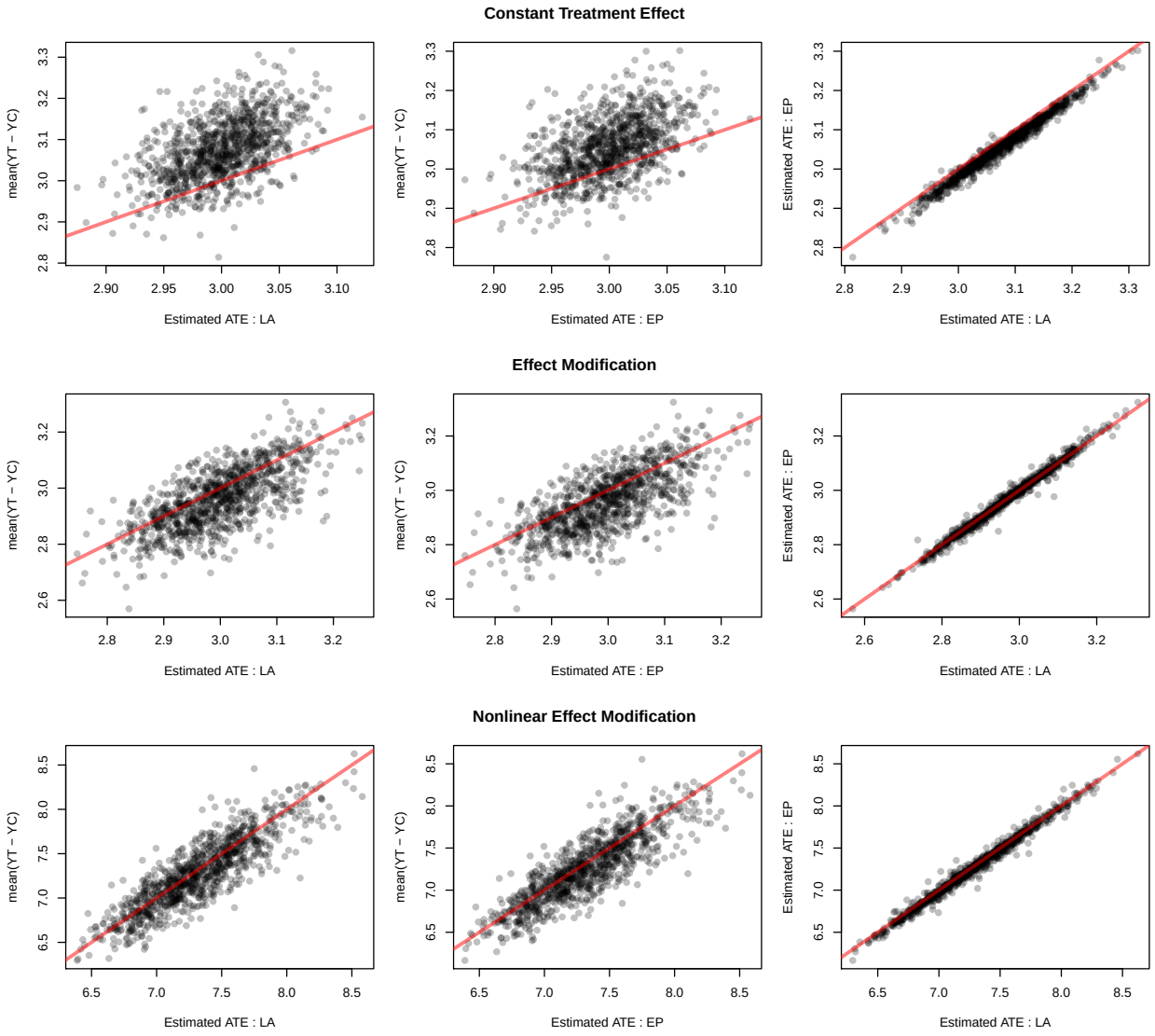

Figure 3.3: Comparing the simulation treatment effect, i.e. $\tau_{sim} = \frac{1}{N}\sum_{i=1}^{N}\left(Y_i^{Z=1} - Y_i^{Z=0}\right)$, against estimates based upon weighting using the optimally balanced Gaussian process propensity score estimates using EP and LA. Each row is a different potential outcome setting. Overall there is agreement between the weighted estimated treatment effects and τ_{sim}.

	Adjustment Method	Bias	Abs. Bias	Emp. S.E.	MSE
Const. TE	No Adjustment	0.957	0.957	0.097	0.926
	True Propensity Score	0.019	0.140	0.176	0.031
	Laplace Approximation	0.074	0.085	**0.073**	0.011
	Expectation Propagation	**0.054**	**0.073**	0.075	**0.008**
Eff. Mod.	No Adjustment	0.478	0.478	0.117	0.242
	True Propensity Score	-0.021	0.169	0.229	0.053
	Laplace Approximation	**-0.039**	**0.092**	**0.107**	**0.013**
	Expectation Propagation	-0.041	0.094	0.110	0.014
Nonlin. TE	No Adjustment	1.905	1.905	0.394	3.784
	True Propensity Score	0.007	0.590	0.800	0.640
	Laplace Approximation	**-0.042**	**0.319**	**0.391**	**0.155**
	Expectation Propagation	-0.065	0.326	0.398	0.162

Table 3.2: Simulation results comparing the relative effectiveness between EP and LA approximations for estimating propensity score weights used to estimate treatment effects. The columns represent the mean bias, mean absolute bias, empirical standard error of the estimates, and finally the mean squared error of the estimates. Each grouped section of rows contains the results for one of the potential outcome settings in Table 3.1, Within each group section, the results based upon no weighting adjustment, adjustment using weights based on the true propensity score, and lastly weights based upon estimates using EP and LA are compared.

the case.

Finally, we compare the computational runtime between the two methods, in Figure 3.4. These timing results were performed on a 2013 27" iMac with a 3.4 GHz Intel Quad-Core i5 processor and 16 GB of memory. Recall that the LA is performed in serial, while the EP method is performed in parallel. This implies that these results will vary from computer to computer, but in many settings LA may be computationally preferable.

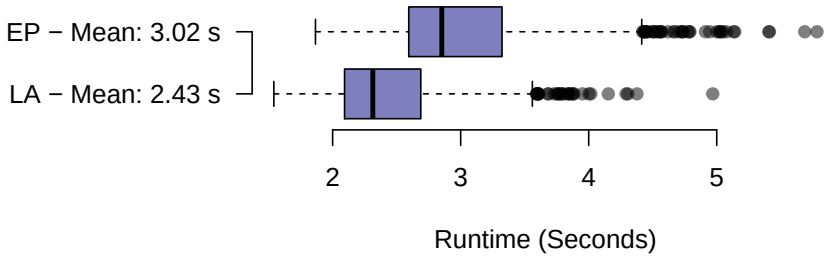

Figure 3.4: Comparing runtime between the LA and EP approximations across 1000 simulations for estimating propensity scores based on 500 observations. Results were collected on a 2013 27" iMac with a 3.4 GHz Intel Quad-Core i5 processor and 16 GB of memory.

3.3.3 Comparative ATE Simulation Study

3.3.3.1 Simulation Setting

The ATE simulation study uses data that are generated with treatment assignment determined using a true propensity score model that is a function of two covariates $X_1 = x_1$ and $X_2 = x_2$ where

$$X = \begin{pmatrix} X_1 = x_1 \\ X_2 = x_2 \end{pmatrix}, \tag{3.13}$$

and $X_1 \sim \mathcal{N}(0,1)$ and $X_2 \sim Bernoulli(0.4)$. The true propensity score is defined in terms of parameters α, β in the following manner,

$$P(Z_i = 1 | X_i = x_i) = \alpha_1 \times \Phi\left(g_j(x_i, \beta)\right) + \alpha_2.$$

In each simulation, $g_j(x_i, \beta)$ takes one of two forms: a polynomial with linear and interaction terms, $g_1(x_i, \beta) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_2$, or a second order polynomial with no interaction terms, $g_2(x_i, \beta) = \beta_0 + \beta_1 x_1 + \beta_2 x_1^2 + \beta_3 x_2$. The values α_1 and α_2 were chosen to restrict

the values of the propensity score to be in the interval $(\alpha_2, \alpha_1 + \alpha_2)$ and thereby ensure that the positivity assumption is valid. Additionally, by strategically choosing α and β values, it is possible to construct functions that are difficult to model using logistic or probit regression. Settings of the parameters for the two propensity score models are defined in Table 3.3. Figure 3.5 visualizes these functions for 500 sample draws as functions of x_1 and x_2.

Setting	Parameters, $\gamma = (\beta_0, \beta_1, \beta_2, \beta_3, \alpha_1, \alpha_2)$
Non-GLM, Linear w/ Interactions	(0.5, 4, -0.5, -3, 0.7, 0.15)
Non-GLM, Second Order	(2.5, 3, -4, -2, 0.75, 0.125)

Table 3.3: Parameter settings used to create propensity score functions

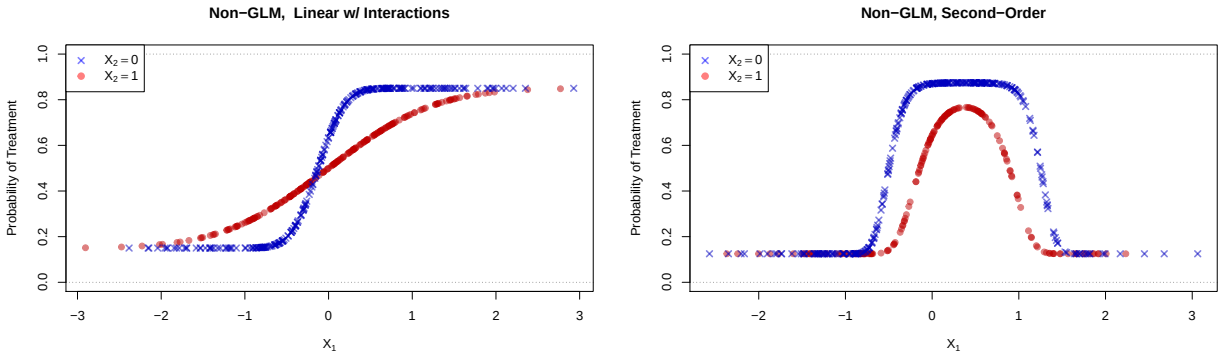

Figure 3.5: Visualization of propensity score simulation settings. Parameter values are described in Table 3.3.

Within each simulation, twelve approaches for estimating the ATE were compared. They are defined in Table 3.4. First the ATE was estimated without adjustment, i.e. $\hat{\tau} = \frac{\sum_i Z_i Y^{obs}}{\sum_i Z_i} - \frac{\sum_i (1-Z_i) Y^{obs}}{\sum_i (1-Z_i)}$, to provide baseline measures of performance. The average treatment effect was then calculated by adjustment using the true propensity score which provides performance measures that could be obtained if the true propensity score were known. Next, adjustment was performed using six nonparametric propensity score estimation methods. The first four methods are our optimally balanced Gaussian process model using the ATE weights previously defined and both the expectation propagation approximation and Laplace approximation methods. The first specification of each model utilizes the

additive kernel defined in Equation (3.5) and the second model uses a kernel such that $k_{se}(x_i, x_j; \rho) = \exp\left\{-\sum_{d=1}^{D} \rho_d^2 (x_{i,d} - x_{j,d})^2/2\right\}$ the squared exponential kernel with automatic relevance detection where each dimension has its own inverse-length scale parameter ρ_d. Additionally, we add a small value (i.e., $1e{-}6$) to the diagonal of each matrix $\mathbf{K}_\theta$ for numerical stability. The other two nonparametric approaches are a method utilizing gradient boosting machines (GBM) available in the `twang` package in `R` (McCaffrey et al., 2004) and Bayesian additive regression trees (Chipman et al., 2010; Hill, 2011) available in the `BART` package in `R`. Finally we compare the nonparametric methods against methods where the propensity score is estimated parametrically. We consider two methods: the Covariate Balancing Propensity Score, CBPS (Imai and Ratkovic, 2014) available in the `CBPS` package in `R`, and a generalized linear model (GLM) using logistic regression. For each of these parametric models we apply the model once with the correct form of $g_j(x_i, \beta)$ and once with a misspecified model as defined in Table 3.4. We note that all of the parametric models are "misspecified" in the sense that they are missing parameters to capture the effect of α_1 and α_2.

3.3.3.2 ATE Simulation Results

Results for estimating the ATE are organized into three tables. Table 3.5 summarizes results obtained with respect to covariate balance. It contains six columns and provides the proportion of the 1000 simulations which were declared to be mean balanced, i.e., $|\Delta_d| < \delta$ for all d and for various thresholds of δ. The first three columns are results using the first propensity score listed in Table 3.3 (Non-GLM, Linear w/ Interactions) and the next three columns are the second setting (Non-GLM, Second Order). Table 3.6 contains results for estimating the ATE under the potential outcome settings listed in Table 3.1 when the true propensity score was the "Non-GLM, Linear w/ Interactions" setting. The columns of this table contain the following

No.	Adjustment Weighting Method	R package
1	No Adjustment	-
2	True Propensity Score	-
3	Optimally Balanced Gaussian Process Propensity Score - EP *Normalized Polynomial + Squared Exponential, Common ρ*	gpbalancer
4	Optimally Balanced Gaussian Process Propensity Score - EP *Squared Exponential, Covariate Specific ρ_d*	gpbalancer
5	Optimally Balanced Gaussian Process Propensity Score - LA *Normalized Polynomial + Squared Exponential, Common ρ*	gpbalancer
6	Optimally Balanced Gaussian Process Propensity Score - LA *Squared Exponential, Covariate Specific ρ_d*	gpbalancer
7	Gradient Boosted Machine	twang
8	Bayesian Additive Regression Trees	BART
9	Generalized Linear Model - Correct form: $x^T\beta \equiv g_j(x,\beta)$	glm
10	Covariate Balancing Propensity Score - Correct form $x^T\beta \equiv g_j(x,\beta)$	CBPS
11	Generalized Linear Model - $x^T\beta = \beta_0 + \beta_1 x_1$	glm
12	Covariate Balancing Propensity Score - $x^T\beta = \beta_0 + \beta_1 x_1 + \beta_2 x_2$	CBPS

Table 3.4: Models to be compared for estimating the ATE under various simulation settings and the R package used to estimate them. The first two methods (no adjustment and adjustment by the true propensity score) provide baseline performance measures. The next six methods are nonparametric methods for estimating the propensity score. The package gpbalancer can be found at https://github.com/bvegetabile/gpbalancer. The last four rows are parametric methods for estimating the propensity score. Note that all parametric models are misspecified in the sense that they are missing terms to handle the fact that $\alpha_1 \neq 1$ and $\alpha_2 \neq 0$ in the data-generating model.

simulation summaries without conditioning on balance: the mean bias, the mean absolute bias, the mean reduction in bias as compared with no adjustment by the propensity score, the empirical standard error of the simulation ATE estimates, and finally the empirical mean squared error for the ATE. The table is grouped vertically by the potential outcome model which is listed at the left margin of the table. Note that 'Correct form: $x^T\beta \equiv g_j(x,\beta)$' implies the functional form used to model the propensity score is identical to the function $g_j(x,\beta)$ used in the true data-generating propensity score, i.e., $g_1(x,\beta) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_2$, or $g_2(x,\beta) = \beta_0 + \beta_1 x_1 + \beta_2 x_1^2 + \beta_3 x_2$. Finally, Table 3.7 contains results for estimating the ATE under the potential outcome settings listed in Table 3.1 when the true propensity score was the "Non-GLM, Second Order" setting and is organized similar to Table 3.6.

The first significant result demonstrated in Table 3.5 is that the optimally balanced Gaussian process propensity score methods balance covariates more effectively than other methods, across different thresholds of δ and under both propensity score settings. When the true propensity score was a linear function with interaction terms (left grouping of columns) the CBPS methods also performed well for balancing covariates, while the other nonparametric methods performed well when the true propensity score was a second-order polynomial (right grouping of columns). Neither performed well under both propensity score functions. This suggests that our method is applicable in a wider range of data-generating settings than competing methods for estimating the ATE.

	Prop. Bal. ($\lvert\Delta_d\rvert < \delta$ *for all* d)					
	Non-GLM, Linear w/ Interactions			*Non-GLM, Second Order*		
Adjustment Method	$\delta = 0.1$	$\delta = 0.15$	$\delta = 0.2$	$\delta = 0.1$	$\delta = 0.15$	$\delta = 0.2$
No Adjustment	0.000	0.000	0.000	0.000	0.000	0.000
True Propensity Score	0.380	0.662	0.843	0.338	0.584	0.804
Opt. Bal. GP PS (NPSE) - EP	**0.998**	**1.000**	**1.000**	0.969	0.996	**1.000**
Opt. Bal. GP PS (SE) - EP	0.994	0.999	**1.000**	0.973	0.996	0.999
Opt. Bal. GP PS (NPSE) - LA	0.997	0.999	**1.000**	0.972	**0.997**	**1.000**
Opt. Bal. GP PS (SE) - LA	0.989	0.998	0.999	**0.978**	**0.997**	**1.000**
GBM (`twang`)	0.071	0.450	0.812	0.665	0.939	0.995
BART	0.000	0.191	0.818	0.894	0.991	**1.000**
GLM - Correct Form	0.037	0.095	0.190	0.008	0.016	0.028
CPBS - Correct Form	0.790	0.954	0.988	0.088	0.181	0.269
GLM - Misspecified	0.050	0.178	0.349	0.002	0.007	0.032
CBPS - Misspecified	0.837	0.981	**1.000**	0.004	0.053	0.273

Table 3.5: Simulation results for demonstrating the performance in balancing covariates. The columns provide the proportion of the 1000 simulations which were declared to be mean balanced, i.e., $\lvert\Delta_d\rvert < \delta$ for all d and for various thresholds of δ

While balancing covariates is an important step in performing causal inference, the primary goal is to minimize the bias of the estimated average treatment effect through adjustment on an estimated propensity score. Tables 3.6 and 3.7 demonstrate that the optimally balanced Gaussian process propensity score provides unbiased estimation of the ATE. This demonstrates that the balance achieved through the optimization procedure is not at the expense of other properties of the estimator. Additionally, we see that the mean squared error is

	Adjustment Method	Bias	Abs. Bias	%ABR	Emp. S.E.	MSE
Constant Treatment Effect	No Adjustment	0.959	0.959	-	0.082	0.927
	True Propensity Score	0.002	0.099	89.541	0.125	0.016
	Opt. Bal. GP PS (NPSE) - EP	**0.001**	**0.025**	**97.350**	**0.032**	**0.001**
	Opt. Bal. GP PS (SE) - EP	0.008	0.027	97.190	0.034	**0.001**
	Opt. Bal. GP PS (NPSE) - LA	0.006	0.027	97.204	0.034	**0.001**
	Opt. Bal. GP PS (SE) - LA	0.011	0.029	97.009	0.036	**0.001**
	GBM (`twang`)	0.156	0.156	83.885	0.050	0.027
	BART	0.169	0.169	82.453	0.038	0.030
	GLM - Correct Form	-0.459	0.460	52.187	0.362	0.342
	CPBS - Correct Form	0.008	0.071	92.576	0.088	0.008
	GLM - Misspecified	-0.300	0.301	68.725	0.226	0.141
	CBPS - Misspecified	0.020	0.062	93.478	0.074	0.006
Effect Modification	No Adjustment	0.465	0.465	-	0.108	0.228
	True Propensity Score	0.008	0.117	72.331	0.147	0.022
	Opt. Bal. GP PS (NPSE) - EP	**-0.006**	**0.074**	82.337	**0.092**	**0.009**
	Opt. Bal. GP PS (SE) - EP	0.015	**0.074**	**82.949**	**0.092**	**0.009**
	Opt. Bal. GP PS (NPSE) - LA	**0.006**	**0.074**	82.781	**0.092**	**0.009**
	Opt. Bal. GP PS (SE) - LA	0.028	0.078	82.448	0.094	0.010
	GBM (`twang`)	0.060	0.099	78.932	0.108	0.015
	BART	0.054	0.089	81.000	0.096	0.012
	GLM - Correct Form	0.179	0.509	-13.569	0.970	0.973
	CPBS - Correct Form	0.102	0.226	50.457	0.285	0.092
	GLM - Misspecified	0.157	0.371	18.147	0.662	0.463
	CBPS - Misspecified	0.109	0.212	53.895	0.259	0.079
Nonlinear	No Adjustment	1.804	1.804	-	0.382	3.399
	True Propensity Score	-0.035	0.438	73.638	0.548	0.302
	Opt. Bal. GP PS (NPSE) - EP	-0.080	0.318	80.584	0.383	0.153
	Opt. Bal. GP PS (SE) - EP	-0.022	**0.307**	81.599	**0.379**	**0.144**
	Opt. Bal. GP PS (NPSE) - LA	-0.045	0.314	81.080	0.384	0.149
	Opt. Bal. GP PS (SE) - LA	**0.020**	0.310	**81.766**	0.385	0.148
	GBM (`twang`)	0.243	0.392	78.840	0.433	0.247
	BART	0.246	0.371	80.138	0.399	0.219
	GLM - Correct Form	0.939	1.924	-6.410	3.350	12.096
	CPBS - Correct Form	0.613	0.906	50.854	1.031	1.437
	GLM - Misspecified	0.655	1.354	25.413	2.230	5.398
	CBPS - Misspecified	0.528	0.805	56.409	0.939	1.159

Table 3.6: Simulation results for estimating the ATE in the non-GLM setting where the true propensity score was linear and included interaction terms

	Adjustment Method	Bias	Abs. Bias	%ABR	Emp. S.E.	MSE
Constant Treatment Effect	No Adjustment	0.426	0.426	-	0.086	0.189
	True Propensity Score	-0.002	0.103	74.306	0.129	0.017
	Opt. Bal. GP PS (NPSE) - EP	-0.027	0.041	89.804	0.044	0.003
	Opt. Bal. GP PS (SE) - EP	-0.019	0.037	90.793	0.044	**0.002**
	Opt. Bal. GP PS (NPSE) - LA	-0.025	0.040	89.919	0.044	0.003
	Opt. Bal. GP PS (SE) - LA	**-0.017**	**0.036**	90.896	0.044	**0.002**
	GBM (`twang`)	0.055	0.064	85.130	0.053	0.006
	BART	0.056	0.060	86.152	0.043	0.005
	GLM - Correct Form	-1.518	1.763	-334.181	1.236	3.833
	CPBS - Correct Form	-0.509	0.554	-34.163	0.578	0.592
	GLM - Misspecified	-0.028	0.038	**91.064**	**0.041**	**0.002**
	CBPS - Misspecified	0.224	0.224	47.054	0.048	0.052
Effect Modification	No Adjustment	-0.314	0.314	-	0.093	0.107
	True Propensity Score	0.004	0.137	46.306	0.175	0.030
	Opt. Bal. GP PS (NPSE) - EP	-0.040	**0.100**	**66.473**	**0.122**	**0.016**
	Opt. Bal. GP PS (SE) - EP	-0.016	0.102	64.710	0.129	0.017
	Opt. Bal. GP PS (NPSE) - LA	-0.035	0.104	64.781	0.127	0.017
	Opt. Bal. GP PS (SE) - LA	**-0.013**	0.105	63.416	0.132	0.018
	GBM (`twang`)	-0.107	0.136	56.281	0.131	0.028
	BART	-0.116	0.131	59.709	0.109	0.025
	GLM - Correct Form	4.250	4.250	-1498.183	2.623	24.933
	CPBS - Correct Form	1.498	1.498	-442.891	1.593	4.781
	GLM - Misspecified	-0.633	0.633	-117.689	0.107	0.412
	CBPS - Misspecified	-0.455	0.455	-52.641	0.089	0.215
Nonlinear	No Adjustment	1.330	1.330	-	0.342	1.887
	True Propensity Score	-0.033	0.565	51.450	0.700	0.491
	Opt. Bal. GP PS (NPSE) - EP	-0.190	0.390	64.882	0.444	0.233
	Opt. Bal. GP PS (SE) - EP	-0.127	0.367	67.576	0.437	0.207
	Opt. Bal. GP PS (NPSE) - LA	-0.171	0.387	65.320	0.448	0.230
	Opt. Bal. GP PS (SE) - LA	-0.115	0.369	67.489	0.442	0.208
	GBM (`twang`)	**0.074**	0.379	69.336	0.470	0.226
	BART	0.102	**0.355**	**71.616**	0.428	**0.193**
	GLM - Correct Form	-5.922	9.267	-648.904	9.499	125.210
	CPBS - Correct Form	-2.282	3.032	-146.321	3.565	17.903
	GLM - Misspecified	1.701	1.701	-27.738	0.505	3.148
	CBPS - Misspecified	1.475	1.475	-10.518	**0.412**	2.344

Table 3.7: Simulation results for estimating the ATE in the non-GLM setting where the true propensity score was a second-order polynomial.

often the smallest value and correspondingly provides low empirical standard errors of the estimator. In cases, where an alternative method provides better performance, the optimally balanced Gaussian process method is comparable.

The alternative nonparametric methods perform similarly to the optimally balanced Gaussian process propensity score in estimating the ATE. These methods generally obtain performance in estimating average treatment effects that are similar to the true data generating propensity score. Across both tables it appears that nonparametric models of the propensity score provide more consistent performance, for both balancing covariates and estimating the ATE, than parametric models. Specifically consider row 11 (GLM - Misspecified) of Table 3.7, the method provides very good performance for estimating the ATE under the the "Linear Related to X_1" setting, but in Table 3.5 it is demonstrated that in almost all simulations the propensity score estimates did not balance the covariates. Alternatively consider the last row of the "Effect Modification" grouping of Table 3.7 (CBPS - Misspecified), Table 3.5 demonstrates that the method provides adequate covariate balance, yet the performance in estimating the ATE under this effect modification case is worse than for all nonparametric methods (and similarly in the nonlinear case, but not as extreme). This implies that while the method may be able to balance covariates, it does not guarantee optimal bias reduction. We note that in Table 3.7, under the nonlinear potential outcome setting, that BART provided the best performance but the optimally balanced Gaussian process propensity score was comparable across many metrics.

Comparing across all ATE simulations, the optimally balanced Gaussian process propensity score provides the best performance in aggregate for both balancing the covariates and removing bias in estimating the ATE.

3.3.4 Comparative ATT Simulation Studies

3.3.4.1 ATT Simulation 1 - Setting

To compare methods for estimating the ATT, data were simulated with two covariates X_1 and X_2. In the treated group, data were simulated such that

$$X_1|Z=1 \sim \pi_1 \mathcal{N}(\mu_{1,1}=0.5, \sigma^2_{1,1}=1.5^2) + (1-\pi_1)\mathcal{N}(\mu_{1,2}=3.5, \sigma^2_{1,2}=0.5^2),$$

where $\pi_1 = 0.25$. In the control group,

$$X_1|Z=0 \sim \pi_0 \mathcal{N}(\mu_{0,1}=-1, \sigma^2_{0,1}=0.5^2) + (1-\pi_0)\mathcal{N}(\mu_{0,2}=2, \sigma^2_{0,2}=3^2),$$

where for this group $\pi_0 = 0.5$. For both groups $X_2 \sim Bernoulli(0.4)$. The density functions for X_1 in each group are shown in Figure 3.6, where the third panel is the induced probability of treatment that is found through an application of Baye's Theorem. Each simulation contained 100 samples for the treatment group and 400 samples for the control pool of individuals from these covariate densities.

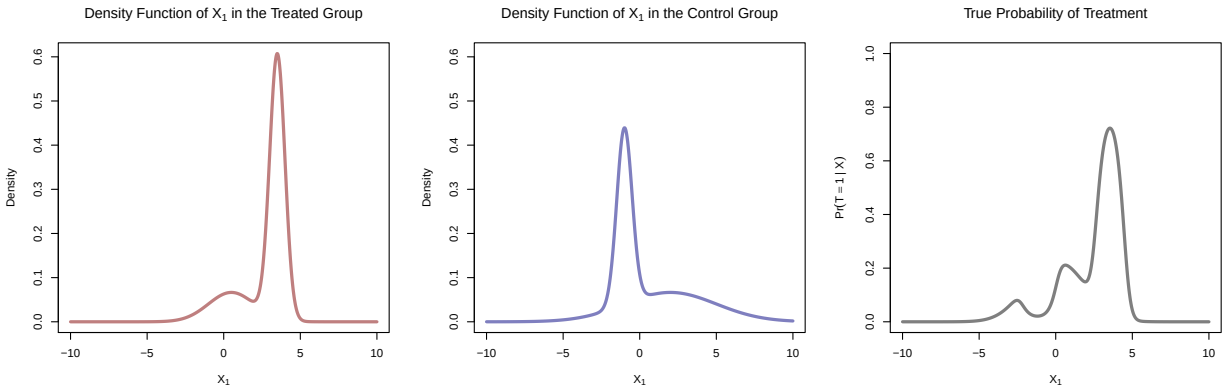

Figure 3.6: Visualization of covariate distributions for estimating the ATT.

Similar to the ATE Setting discussed in Section 3.3.3, we compare the optimally balanced

Gaussian process propensity score estimation method against other methods of estimating the propensity score. For our method, we utilize similar kernels as those defined in Section 3.3.3.1, but now we use the ATT weighting as defined in Section 3.2.2 when measuring covariate imbalance. We consider eleven methods in total for estimating the propensity score in the ATT case as outlined in Table 3.8 and described previously. In contrast to the previous section the "No Adjustment" estimator does not make sense for estimating the ATT and therefore we do not use it. Also, the "Correct form" estimators utilizing GLM and CBPS are omitted. We also note that, for the "Nonlinear Effect Modification" potential outcome setting, we scale the covariate X_1 so that it has mean zero and unit variance before calculating each potential outcome, $Y^{Z=z}$.

No.	Adjustment Weighting Method	R package
1	True Propensity Score	-
2	Optimally Balanced Gaussian Process Propensity Score - EP *Normalized Polynomial, Order 1 + Squared Exponential, Common ρ*	gpbalancer
3	Optimally Balanced Gaussian Process Propensity Score - EP *Normalized Polynomial, Order 2 + Squared Exponential, Common ρ*	gpbalancer
4	Optimally Balanced Gaussian Process Propensity Score - EP	gpbalancer
5	Optimally Balanced Gaussian Process Propensity Score - LA *Normalized Polynomial, Order 1 + Squared Exponential, Common ρ*	gpbalancer
6	Optimally Balanced Gaussian Process Propensity Score - LA *Normalized Polynomial, Order 2 + Squared Exponential, Common ρ*	gpbalancer
7	Optimally Balanced Gaussian Process Propensity Score - LA *Squared Exponential, Covariate specific ρ_d*	gpbalancer
8	Gradient Boosted Machine	twang
9	Bayesian Additive Regression Trees	BART
10	Generalized Linear Model - $x^T\beta = \beta_0 + \beta_1 x_1 + \beta_2 x_1^2 + \beta_3 x_2$	glm
11	Covariate Balancing Propensity Score - $x^T\beta = \beta_0 + \beta_1 x_1 + \beta_2 x_1^2 + \beta_3 x_2$	CBPS

Table 3.8: Models to be compared for estimating the ATT under various simulation settings and the R package used to estimate them.

3.3.4.2 ATT Simulation 1 - Results

Similar to Section 3.3.3.2 we review the results for estimating the ATT using the optimally balanced GP method and compare it against other methods of estimating the propensity score; Table 3.9 provides simulation results for assessing mean balance and Table 3.10 provides performance results for estimating the ATT. Table 3.10 is grouped vertically by potential outcome setting, similar to the previous section, and each group contains performance metrics for the simulations without conditioning on covariate balance as in Section 3.3.3.

	Prop. Bal. ($\lvert\Delta_d\rvert < \delta$ for all d)		
Adjustment Method	$\delta = 0.1$	$\delta = 0.15$	$\delta = 0.2$
True Propensity Score	0.270	0.498	0.697
Opt. Bal. GP PS (NPSE - 1) - EP	0.998	**1.000**	**1.000**
Opt. Bal. GP PS (NPSE - 2) - EP	**1.000**	**1.000**	**1.000**
Opt. Bal. GP PS (SE) - EP	0.984	**1.000**	**1.000**
Opt. Bal. GP PS (NPSE - 1) - LA	0.995	**1.000**	**1.000**
Opt. Bal. GP PS (NPSE - 2) - LA	0.998	**1.000**	**1.000**
Opt. Bal. GP PS (SE) - LA	0.955	0.999	**1.000**
GBM (`twang`)	0.703	0.936	0.991
BART	0.516	0.962	0.999
GLM	0.667	0.986	**1.000**
CBPS	**1.000**	**1.000**	**1.000**

Table 3.9: Simulation results for demonstrating the performance in balancing covariates. The columns provide the proportion of the 1000 simulations which were declared to be mean balanced, i.e., $\lvert\Delta_d\rvert < \delta$ for all d and for various thresholds of δ

Table 3.9 demonstrates that the optimally balanced Gaussian process propensity score again performs well for balancing covariates, as it was intended to do. These results also demonstrate the primary utility of the additive kernel (over using the squared exponential kernel alone), as it provides superior performance for estimating the ATT. Additionally, it is shown that by increasing the order of the polynomial (i.e., $p = 2$ for the normalized polynomial kernel) that it is possible to achieve even better balance than when we only consider linear functions. The results demonstrate that our method is comparable with CBPS, a method

	Adjustment Method	Bias	Abs. Bias	Emp. S.E.	MSE
Constant TE	True Propensity Score	0.001	0.157	0.197	0.039
	Opt. Bal. GP PS (NPSE - 1) - EP	0.005	0.040	0.051	0.003
	Opt. Bal. GP PS (NPSE - 2) - EP	**-0.001**	0.037	0.047	0.002
	Opt. Bal. GP PS (SE) - EP	0.090	0.092	0.050	0.011
	Opt. Bal. GP PS (NPSE - 1) - LA	0.063	0.068	0.053	0.007
	Opt. Bal. GP PS (NPSE - 2) - LA	0.039	0.051	0.051	0.004
	Opt. Bal. GP PS (SE) - LA	0.112	0.113	0.048	0.015
	GBM (`twang`)	0.127	0.128	0.067	0.021
	BART	0.165	0.165	0.061	0.031
	GLM	0.145	0.145	0.043	0.023
	CPBS	0.002	**0.027**	**0.034**	**0.001**
Effect Modification	True Propensity Score	-0.023	0.454	0.565	0.319
	Opt. Bal. GP PS (NPSE - 1) - EP	-0.019	**0.339**	**0.419**	**0.176**
	Opt. Bal. GP PS (NPSE - 2) - EP	-0.025	0.341	0.421	0.178
	Opt. Bal. GP PS (SE) - EP	0.065	0.345	0.423	0.183
	Opt. Bal. GP PS (NPSE - 1) - LA	0.039	0.341	0.420	0.178
	Opt. Bal. GP PS (NPSE - 2) - LA	**0.015**	0.340	0.420	0.177
	Opt. Bal. GP PS (SE) - LA	0.088	0.352	0.427	0.190
	GBM (`twang`)	0.103	0.356	0.429	0.195
	BART	0.140	0.358	0.421	0.197
	GLM	0.120	0.360	0.430	0.199
	CPBS	-0.023	0.344	0.426	0.182
Nonlinear	True Propensity Score	-0.035	0.472	0.598	0.358
	Opt. Bal. GP PS (NPSE - 1) - EP	0.080	0.534	0.669	0.453
	Opt. Bal. GP PS (NPSE - 2) - EP	**0.013**	0.531	0.668	0.447
	Opt. Bal. GP PS (SE) - EP	0.096	0.540	0.675	0.464
	Opt. Bal. GP PS (NPSE - 1) - LA	0.101	0.533	0.665	0.452
	Opt. Bal. GP PS (NPSE - 2) - LA	0.067	0.532	0.666	0.448
	Opt. Bal. GP PS (SE) - LA	0.114	0.538	0.670	0.462
	GBM (`twang`)	-0.063	**0.522**	**0.650**	**0.426**
	BART	0.028	0.532	0.670	0.450
	GLM	-0.105	0.536	0.660	0.446
	CPBS	0.039	0.533	0.667	0.447

Table 3.10: Simulation results for estimating the ATT grouped vertically by potential outcome setting.

that also optimizes with respect to covariate balance, though with parametric assumptions. Finally, in Table 3.10, we see that all methods provide similar performance and provide unbiased estimates of the ATT, but methods which optimize on covariate balance provide lower

levels of MSE and absolute bias.

3.3.4.3 ATT Simulation 2 - Setting

The results of Section 3.3.4.2 suggest that all methods tested provide comparable performance for estimating the ATT. In order to further differentiate the methods, a second simulation study is provided for estimating the ATT with a more complicated data generation procedure. Here $X = (\ X_1 \quad X_2 \quad X_3 \quad X_4 \quad X_5\ )^T$ is a five-dimensional vector of covariates, where

$$X_1 \sim \mathcal{N}(0,1) \tag{3.14}$$

$$X_2 \sim \mathcal{N}(0,1) \tag{3.15}$$

$$X_3 \sim Bernoulli(p_3 = 0.3) \tag{3.16}$$

$$X_4 \sim Bernoulli(p_4 = \Phi(x_1)) \tag{3.17}$$

$$X_5 \sim Bernoulli(p_5 = \Phi(x_2)). \tag{3.18}$$

The treatment assignment mechanism is defined as,

$$P(Z = 1|X = x) = \Phi(f(x)) \tag{3.19}$$

with

$$f(x) = 0.5x_1 + 0.25x_2 + 0.1x_1x_2x_3 + 0.05x_2x_5 + 0.025x_4 \tag{3.20}$$

Under this generative setting, we then simulate $Z \mid X = x \sim Bernoulli(\Phi(f(x)))$.

We consider two potential outcome settings defined in Table 3.11, where $\epsilon_{Z=z} \sim \mathcal{N}(0,1)$,

and define treatment effects as $\tau = E(Y^{Z=1} - Y^{Z=0} \mid Z = 1)$. We estimate the propensity score based upon the methods outlined in Table 3.8, but for the CBPS and logistic regression methods the model for the predictor function was $f(x, \beta) = \beta_0 + (\beta_1 x_1 + \beta_2 x_1^2) + (\beta_3 x_2 + \beta_4 x_2^2) + \beta_5 x_3 + \beta_6 x_4^2 + \beta_7 x_5$. For each method, treatment effects were estimated and the results are provided in the next subsection.

Potential Outcome Setting	Potential Response Functions
1) Constant Treatment Effect	$Y^{Z=1} = 5X_1^2 + X_1X_3 - 4X_2 + 50X_5 + \mathbf{10} + \epsilon_{Z=1}$ $Y^{Z=0} = 5X_1^2 + X_1X_3 - 4X_2 + 50X_5 + \epsilon_{Z=0}$
2) Effect Modification	$Y^{Z=1} = 5X_1^2 + X_1X_3 - 4X_2 + 50X_5 + \mathbf{10X}_1 - \mathbf{3X}_2^3 + \epsilon_{Z=1}$ $Y^{Z=0} = 5X_1^2 + X_1X_3 - 4X_2 + 50X_5 + \epsilon_{Z=0}$

Table 3.11: Models for potential outcomes. The bold portion of the equation in the potential outcome in the treated group is what differs between the two potential outcomes

3.3.4.4 ATT Simulation 2 - Results

Results for the second ATT simulation are organized into two tables. Table 3.12 organizes the covariate balance results for the criterion of total mean balance, i.e., $|\Delta_d| < \delta$ for all d, and for various thresholds of δ. Table 3.13 organizes results for estimating the ATT in this setting. The columns of this table contain the following simulation summaries without conditioning on balance: the mean bias, the mean absolute bias, the mean reduction in bias as compared with no adjustment by the propensity score, the empirical standard error of the simulation ATT estimates, and finally the empirical mean squared error for the ATT. The table is grouped vertically by the potential outcome which is listed at the left margin of the table.

These simulation results allow for more discrimination among the methods. Table 3.12 demonstrates once again that many of the methods are comparable for providing sufficient covariate balance when estimating the ATT and that our optimally balanced Gaussian process propensity score estimation procedure often provides the best covariate balance performance

	Prop. Bal. ($\lvert\Delta_d\rvert < \delta$ for all d)		
Adjustment Method	$\delta = 0.1$	$\delta = 0.15$	$\delta = 0.2$
True Propensity Score	0.311	0.573	0.785
Opt. Bal. GP PS (NPSE - 1) - EP	0.831	0.977	**0.998**
Opt. Bal. GP PS (NPSE - 2) - EP	0.801	0.970	0.996
Opt. Bal. GP PS (SE) - EP	0.805	**0.979**	**0.998**
Opt. Bal. GP PS (NPSE - 1) - LA	0.750	0.965	0.997
Opt. Bal. GP PS (NPSE - 2) - LA	0.712	0.945	0.992
Opt. Bal. GP PS (SE) - LA	0.735	0.958	0.996
GBM (`twang`)	0.206	0.575	0.866
BART	0.044	0.326	0.778
GLM	**0.809**	0.968	0.986
CBPS	0.770	0.976	**0.998**

Table 3.12: Simulation results for demonstrating the performance in balancing covariates. The columns provide the proportion of the 1000 simulations which were declared to be mean balanced, i.e., $|\Delta_d| < \delta$ for all d and for various thresholds of δ

among the methods. We see again that the fully nonparametric propensity score estimation procedures (those not optimized for minimizing balance) provide covariate balance that is similar to the true propensity score.

The results in Table 3.13 compare the relative performance in estimating the average treatment effect in the treated subgroup based upon mean bias, mean absolute value of bias, mean standard error of the estimator, and mean squared error of the estimator. The first result is that adjustment by the true propensity score is unbiased, yet there is a large standard error for these estimates. This explains the discrepancy between the mean bias and mean absolute bias estimates in Table 3.13. Almost every other method yields estimates that are slightly biased but with much lower standard errors. Despite this, it is clear from these results that the two best performing methods are the Covariate Balancing Propensity Score and the optimally balanced Gaussian process propensity score methods. The fully nonparameteric propensity score estimates (GBM, BART) outperform the true propensity score in terms of relative empirical MSE, but in both simulations provide subpar performance as compared with our method and the CBPS method.

	Adjustment Method	Bias	Abs. Bias	Emp. S.E.	MSE
Constant TE	True Propensity Score	0.057	2.250	2.812	7.905
	Opt. Bal. GP PS (NPSE - 1) - EP	0.647	0.839	0.884	1.198
	Opt. Bal. GP PS (NPSE - 2) - EP	0.617	0.807	0.846	1.095
	Opt. Bal. GP PS (SE) - EP	**0.499**	**0.727**	0.841	**0.955**
	Opt. Bal. GP PS (NPSE - 1) - LA	0.730	0.905	0.915	1.369
	Opt. Bal. GP PS (NPSE - 2) - LA	0.691	0.859	0.861	1.219
	Opt. Bal. GP PS (SE) - LA	0.567	0.787	0.876	1.089
	GBM (`twang`)	1.393	1.755	1.625	4.577
	BART	1.188	1.246	0.887	2.197
	GLM	0.566	1.254	1.636	2.994
	CPBS	0.778	0.801	**0.644**	1.020
Effect Modification	True Propensity Score	0.052	2.299	2.916	8.496
	Opt. Bal. GP PS (NPSE - 1) - EP	0.641	1.157	1.345	2.217
	Opt. Bal. GP PS (NPSE - 2) - EP	0.610	1.131	1.322	2.117
	Opt. Bal. GP PS (SE) - EP	**0.492**	**1.089**	1.321	**1.985**
	Opt. Bal. GP PS (NPSE - 1) - LA	0.724	1.201	1.368	2.393
	Opt. Bal. GP PS (NPSE - 2) - LA	0.685	1.160	1.329	2.235
	Opt. Bal. GP PS (SE) - LA	0.561	1.129	1.351	2.136
	GBM (`twang`)	1.387	1.903	1.922	5.612
	BART	1.182	1.427	1.356	3.232
	GLM	0.559	1.477	1.924	4.011
	CPBS	0.772	1.119	**1.202**	2.040

Table 3.13: Simulation results for estimating the ATT grouped vertically by potential outcome setting.

3.4 Application to the Job Training Data of Dehejia and Wahba (1999)

This section focuses on reproducing the results from Dehejia and Wahba (1999), which itself was a replication of earlier work by LaLonde (1986), and is often considered a benchmark data set to demonstrate the performance of propensity score estimation methods. The research goal of LaLonde (1986) was to study the extent to which observational data can be used to replicate results obtained through controlled experiments. LaLonde's starting point was an analysis assessing the effect of a job training program, the National Supported Work (NSW)

Demonstration, on post-exposure earnings. In the original study there was randomization to either a control group that did not receive training or a treatment group that did receive training. LaLonde (1986) then collected six observational data sets and attempted to use these data as pools of control individuals with which to replicate the experimental results; three from the Panel Study on Income Dynamics (PSID-1, PSID-2, PSID-3) and three from the Current Population Survey-Social Security File (CPS-1, CPS-2, CPS-3). The data sets PSID-2, PSID-3 and CPS-2, CPS-3 are subsets from PSID-1 and CPS-1 data sets, respectively, and were chosen because LaLonde believed that the subsetted individuals were more similar to the NSW experimental treatment group. Dehejia and Wahba (1999) then extended these results using propensity score estimation based upon the work of Rosenbaum and Rubin (1983) to estimate the ATT. In this section, we replicate the findings of Dehejia and Wahba (1999) using our optimally balanced Gaussian process propensity score for estimating the ATT.

The data we analyze are the subset of individuals that were utilized in Dehejia and Wahba (1999). This subset of data focused on men who were assigned to treatment after 1975 and the outcome measure was the post-intervention earnings in 1978. The data set contains the following variables for each individual: age, education in years, indicators of whether the individual was black or Hispanic, an indicator of whether the individual was married, an indicator of "no degree", and retrospective earnings in 1974 and 1975. Tables 3.14 and 3.15 provide summaries of the data. It is clear that the observational data sets are different from the experimental data indicating that adjustment is necessary to remove the effects of potential confounding variables. In particular individuals in the experimental data set were often younger, were less often married, and had significantly less earnings in 1974 and 1975; additionally there was a higher representation of black individuals within the experimental data.

The upper portion of Table 3.16 is a summary of the covariate balance when comparing each

Experimental Data

	NSW - Treated	NSW - Control
N_{obs}	185	260
Age	25.82 (7.14)	25.05 (7.04)
$Education$	10.35 (2.01)	10.09 (1.61)
$I(Black)$	0.84 (0.36)	0.83 (0.38)
$I(Hispanic)$	0.06 (0.24)	0.11 (0.31)
$I(Married)$	0.19 (0.39)	0.15 (0.36)
$I(No\ degree)$	0.71 (0.45)	0.83 (0.37)
$RE74$	2095.57 (4873.4)	2107.03 (5676.96)
$RE75$	1532.06 (3210.54)	1266.91 (3097.01)

Table 3.14: Summaries of experimental data from Dehejia and Wahba (1999).

observational data set and the NSW control group against the NSW treated group on each covariate individually and with respect to the total covariate balance as defined by Equation (3.10). This table demonstrates, on a standardized scale, the disparities between the data sets constructed using observational sources and the experimental data set. Along with Tables 3.14 and 3.15 these summaries demonstrate the lack of balance between each observational data set and the NSW treated data set, prior to controlling for the estimated propensity score. We note that PSID-2 and PSID-3 are subsets of PSID-1 and were constructed to be more similar to the NSW population. They do provide better balance but there are still large disparities between these observational data sets and the experimental data set (the same is true for CPS-2 and CPS-3 which were generated from CPS-1). To create balanced data sets we estimate the propensity score using our optimally balanced Gaussian process approach. The covariate balance resulting from adjusting using these estimates is shown in the lower portion of Table 3.16 for each of the observational data sets. Clearly, the optimally balanced Gaussian process propensity score estimates are successful in obtaining better balance between the observational data sets and the NSW treatment group.

The last three rows of Table 3.16 provide estimates of the ATT based upon the six control data sets. The second to last row contains the estimates from Dehejia and Wahba (1999)

Observational Control Data

Data set	Panel Study on Income Dynamics		
Covariate	PSID-1	PSID-2	PSID-3
N_{obs}	2490	253	128
Age	34.85 (10.44)	36.09 (12.06)	38.26 (12.84)
$Education$	12.12 (3.08)	10.77 (3.17)	10.3 (3.16)
$I(Black)$	0.25 (0.43)	0.39 (0.49)	0.45 (0.5)
$I(Hispanic)$	0.03 (0.18)	0.07 (0.25)	0.12 (0.32)
$I(Married)$	0.87 (0.34)	0.74 (0.44)	0.7 (0.46)
$I(No\ degree)$	0.31 (0.46)	0.49 (0.5)	0.51 (0.5)
$RE74$	19428.75 (13404.18)	11027.3 (10793.28)	5566.87 (7226.76)
$RE75$	19063.34 (13594.22)	7569.22 (9024.06)	2610.7 (5550.69)
Data set	Current Population Survey-Social Security File		
Covariate	CPS-1	CPS-2	CPS-3
N_{obs}	15992	2369	429
Age	33.23 (11.04)	28.25 (11.69)	28.03 (10.77)
$Education$	12.03 (2.87)	11.24 (2.58)	10.24 (2.85)
$I(Black)$	0.07 (0.26)	0.11 (0.32)	0.2 (0.4)
$I(Hispanic)$	0.07 (0.26)	0.08 (0.28)	0.14 (0.35)
$I(Married)$	0.71 (0.45)	0.46 (0.5)	0.51 (0.5)
$I(No\ degree)$	0.3 (0.46)	0.45 (0.5)	0.6 (0.49)
$RE74$	14016.8 (9569.5)	8727.96 (8965.95)	5619.24 (6780.83)
$RE75$	13650.8 (9270.11)	7397.23 (8110.5)	2466.48 (3288.16)

Table 3.15: Summaries of observational data from Dehejia and Wahba (1999). The first grouping are the data from the Panel Study on Income Dynamics (PSID) and the second group are data from the Current Population Survey-Social Security File (CPS). Comparing with Table 3.14, there are clearly discrepancies between the experimental and observational data.

based upon weighted leasts squares where the treatment observations were weighted as 1 and control observations were weighted by the number of times they were matched to a treatment observation within a matching procedure (for more information see the footnotes of Table 3 of Dehejia and Wahba (1999)), while the last row contains the weighted estimates based upon our method. From the experimental data (the third row from the bottom), the unadjusted estimated difference in post-intervention earnings between the treated and control group was \$1794, suggesting that the NSW job training program may provide a benefit for enrolled individuals who are similar to those enrolled within the study. We see from Dehejia and Wahba

Covariate Balance Prior to Propensity Score Adjustment														
	NSW		PSID-1		PSID-2		PSID-3		CPS-1		CPS-2		CPS-3	
	Δ	Γ	Δ	Γ	Δ	Γ	Δ	Γ	Δ	Γ	Δ	Γ	Δ	Γ
Age	0.11	0.01	-1.01	-0.38	-1.04	-0.53	-1.20	-0.59	-0.80	-0.44	-0.25	-0.49	-0.24	-0.41
Education	0.14	0.22	-0.68	-0.43	-0.16	-0.46	0.02	-0.46	-0.68	-0.36	-0.39	-0.25	0.04	-0.35
I(Black)	0.04	-0.04	1.48	-0.18	1.05	-0.29	0.90	-0.31	2.43	0.33	2.15	0.14	1.67	-0.10
I(Hispanic)	-0.17	-0.27	0.13	0.29	-0.03	-0.06	-0.20	-0.31	-0.05	-0.09	-0.10	-0.16	-0.28	-0.39
I(Married)	0.09	0.08	-1.84	0.14	-1.31	-0.12	-1.18	-0.16	-1.23	-0.14	-0.60	-0.24	-0.72	-0.24
I(No degree)	-0.30	0.20	0.88	-0.01	0.47	-0.10	0.42	-0.10	0.91	-0.00	0.54	-0.09	0.24	-0.08
RE74	-0.00	-0.15	-1.72	-1.01	-1.07	-0.80	-0.56	-0.39	-1.57	-0.68	-0.92	-0.61	-0.60	-0.33
RE75	0.08	0.04	-1.77	-1.44	-0.89	-1.03	-0.24	-0.55	-1.75	-1.06	-0.95	-0.93	-0.29	-0.02
Total $\mathcal{CB}$	0.37		17.55		8.36		5.47		16.91		8.89		4.57	
Covariate Balance After Propensity Score Adjustment														
	NSW		PSID-1		PSID-2		PSID-3		CPS-1		CPS-2		CPS-3	
	Δ	Γ	Δ	Γ	Δ	Γ	Δ	Γ	Δ	Γ	Δ	Γ	Δ	Γ
Age	0.11	0.01	0.07	0.07	0.04	-0.10	-0.07	-0.15	-0.02	-0.15	-0.01	-0.21	0.04	-0.31
Education	0.14	0.22	-0.00	-0.04	-0.08	-0.10	-0.15	-0.15	-0.01	-0.06	0.04	-0.11	0.00	-0.11
I(Black)	0.04	-0.04	-0.06	0.07	-0.04	0.04	-0.02	0.01	-0.04	0.04	-0.01	0.01	0.02	-0.02
I(Hispanic)	-0.17	-0.27	0.02	0.03	0.04	0.08	0.16	0.37	0.01	0.02	-0.03	-0.05	0.00	0.00
I(Married)	0.09	0.08	0.07	0.06	0.08	0.07	0.01	0.01	0.06	0.05	0.06	0.05	0.07	0.06
I(No degree)	-0.30	0.20	-0.01	0.01	0.13	-0.05	0.27	-0.08	0.00	-0.00	0.03	-0.01	0.06	-0.03
RE74	-0.00	-0.15	-0.19	0.10	-0.14	0.09	-0.08	0.16	-0.01	0.14	-0.01	0.16	0.07	0.24
RE75	0.08	0.04	-0.18	-0.10	-0.09	0.00	-0.01	-0.08	-0.03	-0.08	0.02	-0.00	0.09	0.25
Total $\mathcal{CB}$	0.37		0.12		0.10		0.35		0.06		0.09		0.26	
Estimated Average Treatment Effect in the Treated														
Unadjusted ATT Estimate	1794 (633)		-15205 (1155)		-3647 (960)		1070 (900)		-8498 (712)		-3822 (671)		-635 (657)	
Dehejia and Wahba (1999)	-		1473 (809)		1480 (808)		1549 (826)		1616 (751)		1563 (753)		662 (772)	
Propensity Score Adjusted ATT	-		892 (1002)		1708 (1025)		1803 (1075)		1273 (717)		1748 (735)		1568 (808)	

Table 3.16: Estimates of the ATT and summaries of covariate imbalance using data from Dehejia and Wahba (1999). The upper table of values describe covariate imbalance prior to adjustment for each data set and the middle set of values are the covariate imbalance after adjustment. The last three rows are estimates of the ATT. The second to last row are estimates of the ATT from Dehejia and Wahba (1999) (Table 3, column 8). The final row are estimates of the ATT using weighted least squares and the optimally balanced Gaussian process propensity score.

(1999) that this effect can be replicated using observational data sets and the constructed control groups (though the results are no longer statistically significant). Additionally we see that the optimally balanced Gaussian process propensity score adjustment method provides results consistent with those obtained in Dehejia and Wahba (1999).

3.5 Extension to Multiple Treatment Settings

The primary focus of this Chapter was the development of a Gaussian process propensity score estimation procedure for binary treatment settings, but with a few modifications these methods can be extended to multiple treatment settings. In particular, the required components are a model for the multi-treatment propensity score, $e(z, x)$ (as defined in Section 2.1.3), a loss function that captures the covariate imbalance for a given set of weights, and an algorithm to select optimal hyperparameters. As demonstrated in the overview on Gaussian processes (Section 2.2), estimation for multinomial regression where the latent predictor functions are assumed to arise from Gaussian processes is possible through a Laplace approximation to the posterior distribution, among other approximation strategies. The difficulty surrounding the modeling procedure involves selecting the covariance function $\mathbf{K}_\Theta$ to model the dependency among the latent values. In practice, simplifying assumptions can often be used. For example, a common kernel function can be chosen for modeling each latent Gaussian process and common hyperparameters assumed for covariance functions governing each process. This approach has been useful in our applications.

The primary difficulties in directly extending this method are: (1) defining an estimand of interest and the required weights for this estimand; and (2) defining balance metrics that allow for the evaluation of estimated weights. Two obvious estimands are the average treatment effect in the population (ATE, and similar to the binary treatment setting) and the average

treatment effect in a subpopulation similar to those who received treatment level z, (ATZ). If there are C treatment groups, Li et al. (2017) define the balancing weights for the ATE as,

$$w_i = \sum_{z=0}^{C-1} \frac{I(Z_i = z)}{e(z, x_i)} \tag{3.21}$$

and similarly the weights for the ATZ,

$$w_i = \sum_{z=0}^{C-1} \frac{I(Z_i = z)h(x_i)}{e(z, x_i)} \tag{3.22}$$

where $h(x_i) = e(z', x_i)$ is the propensity score function for the treatment group z' that is the target treatment group for comparing effects. To construct metrics of covariate imbalance, we appeal to ideas from Zubizarreta (2015) and Hainmueller (2012) which create metrics based on a series of moments of the target density of interest. Letting π_d^p define the p^{th} sample moment of the target distribution of interest, the covariate balance in dimension d for group z can be evaluated as the difference between the weighted sample moment and the target value, i.e.,

$$\Omega_{d,p,z} = \hat{M}_{d,p,z}(\mathbf{w}) - \pi_d^p, \tag{3.23}$$

with $\hat{M}_{d,p,z}(\mathbf{w})$ as defined in Chapter 2 (alternatively, Equation 3.23 could be altered to rely on central moments and their targets). The total covariate imbalance can then be defined as a function $g(\cdot)$ of the collection of $\Omega_{d,p,z}$, i.e.,

$$CB(\theta) = g(\Omega_{d,p,z}). \tag{3.24}$$

For example, as in the binary setting, the total imbalance could be the sum of the squares of theses discrepancies, $CB(\theta) = \sum_z \sum_d \sum_p (\Omega_{d,p,z})^2$, or potentially a weighted version of

this function. Finally, the targets π_d^p can be estimated as the unweighted sample moments in the target group, e.g., $\hat{\pi}_d^p = \frac{1}{N}\sum_{i=1}^N X_{i,d}^p$ for the ATE and $\hat{\pi}_d^p = \frac{\sum_{i=1}^N X_{i,d}^p I(Z_i=z)}{\sum_{i=1}^N I(Z_i=z)}$ for the ATZ. These covariate imbalance metrics are explored further in Chapter 4.

Finally, to estimate the propensity score in these settings, we again select θ to minimize $CB(\theta)$, i.e., $\hat{\theta} = \arg\min_\theta CB(\theta)$. This optimization can be performed with the BOBYQA algorithm, similar to the binary setting. The difficulty of the method is computational runtime. Chapter 4 develops a similar method that is computationally much faster, yet accomplishes the same goal more accurately. A comparison between the two methods is provided in Chapter 4, Section 4.3.3.

3.6 Discussion

Estimation of the propensity score is an often-used tool in causal analyses of observational data. Estimation of average treatment effects, either the ATE or the ATT, through adjustment by the estimated propensity score, provides a flexible method of allowing researchers to control for pretreatment covariates. Often though, researchers make parametric modeling assumptions when estimating the propensity score and these assumptions may not be adequate to remove bias in the estimated treatment effects due to covariate imbalance. This Chapter describes a nonparametric estimation strategy that utilizes a Gaussian process model to estimate the probability of treatment given pretreatment covariates. The hyperparameters of the Gaussian process are chosen to minimize an overall covariate imbalance metric. The potential of the proposed method was highlighted using a series of simulations and an application that replicated findings from Dehejia and Wahba (1999).

This Gaussian process propensity score method is advantageous when compared with many commonly employed propensity score estimation strategies due to the flexibility in modeling

the propensity score and the fact that what are truly needed in a causal analysis are estimates from a balancing score. This Chapter demonstrated that optimizing propensity score estimates to minimize a metric of covariate imbalance can provide better performance than other methods focused on accurately estimating the treatment assignment mechanism. This advantage was demonstrated in Section 3.3, where the methods that consistently performed best in estimating either the ATE or the ATT were those that were optimized to minimizing metrics of covariate imbalance. While the CBPS is also optimized towards this pursuit, it was demonstrated in Table 3.6 that nonparametric methods can provide comparable or better performance as it relates to the treatment effect estimation stage of a causal analysis. The Gaussian process approach provides propensity score estimates that balance covariates, while also providing the required bias reduction in estimating treatment effects, as well as lower empirical mean squared error. A secondary advantage relates to the positivity assumption in causal inference. Generalized linear models and other parametric models make an assumption that as we reach more and more extreme values of the covariate space the probability of treatment goes to either zero or one. This may not be a valid assumption and the Gaussian process propensity score method (and other nonparametric methods) allows more flexibility for modeling extreme values of the covariate space.

There are limitations associated with this approach though. The first limitation of the method is that of computational runtime. Gaussian processes are computationally challenged by the need to invert a dense covariance matrix at each step of the algorithm, a process which scales at $\mathcal{O}(N^3)$. These computational restrictions limit the maximum size of a data set that can be used to estimate the propensity score. For example consider Table 3.17 that provides runtime comparisons across the various methods under the setting of a propensity score model that is defined by a linear polynomial with interaction terms as in Section 3.3.3. We see that as the size of the data set increases, the computational burden of the GP approach increases exponentially. These simulations were run on a 2013 iMac with a quad-core 3.4 GHz Intel

Core i5 processor and 16 GB of memory.

	Runtime in Seconds		
Method	$N_{obs} = 100$	$N_{obs} = 500$	$N_{obs} = 1000$
Generalized Linear Model - Logistic Regression	<0.01	<0.01	<0.01
Optimally Balanced GP Propensity Score (NPSE) - EP	0.18	3.88	20.57
Optimally Balanced GP Propensity Score (SE) - EP	0.11	2.51	16.63
Bayesian Additive Regression Trees	1.72	2.94	5.15
Covariate Balancing Propensity Score	0.25	0.18	0.29
Gradient Boosted Machines (`twang`)	2.89	6.64	10.94

Table 3.17: Comparisons of computational runtime across the methods considered in Section 3. The runtimes are averaged across 5 simulated data sets using the data-generating procedure of Section 3.3.3 where the true propensity score was a linear function with interaction terms.

We have already made attempts to reduce this computational burden by implementing a version of the expectation propagation and Laplace approximation algorithms that provide approximations to the posterior distribution of the latent scores and are computationally efficient when compared to MCMC. There are also other considerations that may further reduce this computational burden, such as those that rely on geometric assumptions to create a sparse covariance matrix (e.g., assuming the correlation between points can be set to zero past a certain distance), reduced rank approximations to the covariance matrix, or alternative approximations of the posterior distribution. Further research will focus on reducing this computational burden and some ideas are discussed in Chapter 6. To efficiently utilize the Gaussian process propensity score method as described in this chapter it is recommended to use a system which distributes computation among many cores.

Another limitation of our Gaussian process propensity score method is that the optimization routine could be considered a black-box procedure that does not allow the user to apply important domain-specific knowledge. That is, the algorithm provides an optimal solution with respect to the defined loss function, but that does not mean it provides hyperparameters such that the imbalance of all, let alone specific, covariate dimensions go to zero. In our application setting, for example, the mean imbalance of previous earnings covariates ($RE74$ and $RE75$)

were still found to be somewhat high ($|\Delta| \sim 0.2$) after propensity score adjustment in the PSID-1 data set. This is important because previous earnings may be correlated with both an individuals inclusion into training programs, and with their future earnings. Therefore residual imbalance within important dimensions, such as previous earnings, may imply a certain level of bias cannot be removed when estimating treatment effects.

One potential adjustment to our procedure to address this would be to augment the loss function, prior to estimating treatment effects, so that lack of balance in certain dimensions receive more weight in our metric. Alternatively, as was demonstrated in Section 3, different kernel functions have slightly different properties. Choosing, or constructing a new kernel function for the application at hand may be an alternative way to incorporate domain-specific structure. This may be necessary to correctly model the propensity score in difficult situations. Finally, if neither of these solutions provide adequate covariate balance, it may be a case in which there is no function that can be used to balance covariates for the estimand of interest.

Our simulation results and application show that by using the optimally balanced Gaussian process approach to propensity score modeling, we are able to balance covariates in many settings and enable estimation of the ATE or the ATT. It is clear from this chapter that under the stable unit treatment value assumption and strong ignorability given the covariates, estimating the propensity score using Gaussian processes and optimizing parameters of the model to minimize metrics of covariate imbalance is an effective nonparametric modeling strategy which provides unbiased estimation of treatment effects.

Chapter 4

Targeted Optimal Balancing Weights

Chapter 3 investigated the performance of the optimally balanced Gaussian process propensity score methodology, and developed a metric of covariate imbalance that was only applicable in binary treatment settings. This chapter removes the structural assumption that the collection of latent scores arose from a Gaussian process and develops a loss function that is applicable to both binary and multi-treatment settings that can be optimized to select weights for analyses of causal effects.

4.1 Introduction

The previous chapters have demonstrated that covariate balance is an important characteristic in studies where the goal is an analysis comparing the average response to a set of treatment exposures. Recall that informally covariate balance holds when the joint distributions of the covariates conditional on the treatment exposure are similar. This property allows for a comparison among the responses under each exposure that is marginalized across a common distribution of covariates, i.e., average causal effects given specific populations or

subpopulations of individuals. In randomized studies where each unit has the same probability of receiving each treatment level, covariate balance is attained by design and any imbalance is related to the finite-sample properties of the randomization process. Observational studies, in comparison, do not have this guarantee of covariate balance by design, and individuals receive specific treatment levels based on their innate characteristics, such as their age, sex, genetics, etc. This can lead to differences in the characteristics among the individuals receiving different treatments. A common analysis strategy in observational studies is to find individuals, at each level of treatment exposure, who are similar based on a set of measured characteristics and compare their outcomes. As stated throughout this dissertation, implementing this strategy is challenging when even a modest number of covariates must be matched and this motivates the use of dimension reduction in order to more easily compare individuals.

As discussed previously, a common form of dimension reduction in causal analyses is through the use of balancing scores, of which the propensity score is the most common. Balancing scores can be used for adjustment in an analysis, through matching, stratification, regression adjustment, or used to re-weight the observations. The propensity score is useful in estimating causal effects insomuch as it is a balancing score. In fact, when developing a model to estimate the propensity score, targeting a set of balanced covariates has been suggested as a stopping criteria rather than attempting to obtain a propensity score model that accurately estimates the true probability of treatment (Imbens and Rubin, 2015, see Chapter 13, Section 9). This chapter develops a method where the focus is specifically on covariate balance.

Examples of propensity score estimation methods that focus on balance are the Covariate Balancing Propensity Score (CBPS) and the optimally balanced Gaussian process propensity score method developed in Chapter 3. There are also examples of methods that focus specifically on estimating the weights themselves, i.e., not estimating a propensity score, such that they balance covariates. These methods typically remove the structural assumptions on

the functional form of the weights and optimize the weights themselves without assuming that the weights are explicit functions of an estimated propensity score (Hainmueller, 2012; Zubizarreta, 2015). These methods construct a loss function that is optimized with respect to the weights; the loss function is typically subjected to constraints that are placed on balance metrics in order to control covariate imbalance. Zubizarreta (2015) establishes the approximate unbiasedness of estimators based on such weights in certain settings, such as for estimation of the average treatment effect in the population (ATE) and the average treatment effect in individuals similar to the treated subpopulation (ATT) under certain assumptions on the potential outcome response functions. This chapter develops new methodology that focuses on creating weights for estimating causal effects through the optimization of a loss function that is defined in Section 4.2. In particular, we develop a new objective function that is a function of both: (1) a discrepancy between the sample moments of the covariates in each group and a series of targets that represent the distribution over which we would like to estimate causal effects; and (2) a function of the effective sample size in each group. This loss function is minimized with respect to covariate imbalance while also penalizing collections of weights where only a few units obtain significant weight. The performance of this method is demonstrated in Section 4.3 for estimating the ATE and the ATT in both binary and multi-level treatment settings.

The chapter is organized as follows. Section 4.2 develops our proposed targeted optimal balancing weights (OPTBW) methodology including many of the implementation details for the procedure. Section 4.3 provides a series of simulations to demonstrate the performance in estimating the average treatment effect in the population (ATE), the average treatment effect in individuals similar to the treated subpopulation (ATT), and population average responses in a multi-treatment setting. Section 4.4 provides an application of the method to a common benchmark data set in the causal literature, replicating results from Dehejia and Wahba (1999). Section 4.5 provides a closing discussion.

4.2 Methodology

The methodology of this section makes use of the potential outcome framework of Neyman and Rubin (Splawa-Neyman et al., 1990; Rubin, 1974). For a comprehensive overview see Chapter 2, Section 2.1.

Consider a sample of size N. For each unit $i = 1, \ldots, N$, we observe a D-dimensional covariate vector $X_i = x_i$ and a set treatment assignment indicators $I(Z_i = z)$ for $z \in \mathcal{Z}$ that evaluates to one if the unit received the treatment exposure z and zero otherwise. We posit the existence of a potential outcome $Y_i^{Z=z}$ for each treatment level z and only observe a single response outcome Y_i^{obs} and make the stable unit treatment value assumption such that $Y_i^{obs} = \sum_{z \in \mathcal{Z}} I(Z_i = z) Y_i^{Z=z}$. Ultimately the goal is to estimate the average response under exposure z for certain subpopulations of interest using the weighting estimator

$$\bar{Y}_w^{Z=z} = \frac{\sum_{i=1}^{N} w_i I(Z_i = z) Y_i^{obs}}{\sum_{i=1}^{N} w_i I(Z_i = z)}, \tag{4.1}$$

under an assumption of strong ignorability given X. To achieve reliable estimates of the mean response in each exposure group, we must find weights $\mathbf{w} = (w_1, \ldots, w_N)^T$ such that covariate balance is achieved. We describe a method for the minimization of the imbalance between covariate moments and specified fixed targets representing moments of a distribution of interest; we refer to this as fixed target covariate imbalance minimization. The targets are either supplied by the analyst, or calculated from the data using the unweighted sample moments from subpopulations of interest.

4.2.1 Fixed Target Covariate Imbalance

The goal for fixed target covariate imbalance minimization is to find a vector of weights, $\mathbf{w} > 0$, where $\sum_{i=1}^{N} w_i I(Z_i = z) = 1$ for all z, that minimizes a function of both the total covariate imbalance, $CB(\mathbf{w})$, across the covariate space and a penalty function $\rho(\mathbf{w})$, i.e.,

$$Loss(\mathbf{w}) = CB(\mathbf{w}) + \rho(\mathbf{w}). \tag{4.2}$$

The covariate imbalance in each dimension d and for each exposure group $z \in \mathcal{Z}$ can be defined through a function of the difference between the weighted sample central moments and a series of targets $\boldsymbol{\nu}$ from a density $g(x)$,

$$\frac{\sum_{i=1}^{N} w_i I(Z_i = z)(X_{i,d} - \mu_d)^p}{\sum_{i=1}^{N} w_i I(Z_i = z)} - \nu_d^p, \tag{4.3}$$

for moments $p = 1, \ldots, P$. In Equation (4.3) μ_d is the marginal expectation of the density $g(x)$ in dimension d (i.e., $\mu_d \equiv E_{g(x)}(X_d)$), $\frac{\sum_{i=1}^{N} w_i I(Z_i=z)(X_{i,d}-\mu_d)^p}{\sum_{i=1}^{N} w_i I(Z_i=z)}$ is the p^{th} sample central moment in covariate dimension d, and ν_d^p is a specified target from the density $g(x)$ (i.e., $\nu_d^p \equiv E_{g(x)}(X_d - \mu_d)^p)$). Alternative definitions for covariate imbalance within a dimension, such as considering the raw moments themselves as opposed to the central moments, had been considered but the approach based on central moments in Equation (4.3) has worked best.

To define the total covariate imbalance, we take a weighted sum of squares of the discrepancies between the weighted moments and the targets across the dimensions d, moments p, and treatment groups z,

$$CB(\mathbf{w}) = \sum_{d=1}^{D} \sum_{p=1}^{P} \sum_{z \in \mathcal{Z}} \zeta_d^p \left(\frac{\sum_{i=1}^{N} w_i I(Z_i = z)(X_{i,d} - \mu_d)^p}{\sum_{i=1}^{N} w_i I(Z_i = z)} - \nu_d^p \right)^2. \tag{4.4}$$

The term ζ_d^p scales the contributions to the imbalance metric for moments p and dimensions d. These terms will be selected so that the covariate imbalances across dimensions and moments are given equal weight within the optimization procedure. Now, if we restrict the weights $\mathbf{w}$ to sum to one in each group, i.e., $\sum_{i=1}^{N} w_i I(Z_i = z) = 1$, the previous equation simplifies as,

$$CB(\mathbf{w}) = \sum_{d=1}^{D} \sum_{p=1}^{P} \sum_{z \in \mathcal{Z}} \zeta_d^p \left[\left(\sum_{i=1}^{N} w_i I(Z_i = z)(X_{i,d} - \mu_d)^p \right) - \nu_d^p \right]^2 . \tag{4.5}$$

Note that this function is somewhat analogous to a weighted least squares criterion.

If we were to optimize $CB(\mathbf{w})$ directly, it is easy to conceive of scenarios where a few observations have large weights while all other weights remain negligible. This is undesirable because the estimated treatment effects may exhibit large variability if they are focused heavily on a few observations. To avoid this, we construct a penalty term that is a function of the effective sample size of the weighted sample, generally defined as $ESS_{\mathrm{w}} = \frac{(\sum_{i=1}^{N} w_i)^2}{\sum_{i=1}^{N} w_i^2}$. We augment this definition in order to consider the effective sample sizes for the treatment groups,

$$ESS_{\mathrm{w}}(z) = \frac{(\sum_{i=1}^{N} w_i I(Z_i = z))^2}{\sum_{i=1}^{N} (w_i I(Z_i = z))^2}. \tag{4.6}$$

To penalize small effective sample sizes, we construct a penalty term as the inverse of the effective sample size and apply parameters λ_z to control how much we penalize solutions with respect to sample size. Again if the weights sum to one,

$$\rho(\mathbf{w}) = \sum_{z \in \mathcal{Z}} \lambda_z ESS_{\mathrm{w}}(z)^{-1} \tag{4.7}$$

$$= \sum_{z \in \mathcal{Z}} \lambda_z \Big(\sum_{i=1}^{N} (w_i I(Z_i = z))^2 \Big) \tag{4.8}$$

for $\{\lambda_z\} \in \Re^+$.

We can rewrite the objective function (Equations (4.4)) and the penalty function (Equation (4.8)) in more compact matrix notation to further elucidate the comparison to penalized least squares. Define $\mathbf{X}_c^p$ as a matrix where each column has been centered (denoted by the c) with respect to the density $g(x)$ and is defined as $X_{c,d}^p = (X_d - \mu_d)^p$. Further, define $\mathbf{X}_c = (\mathbf{X}_c^1, \mathbf{X}_c^2, \ldots, \mathbf{X}_c^P)$ as the concatenation of these centered matrices for exponents $p = 1, \ldots, P$ and let $\mathbf{D}_z = diag\big(\{I(Z_1 = z), \ldots, I(Z_N = z)\}\big)$. Equation (4.4) is rewritten more compactly as,

$$CB(\mathbf{w}) = \sum_{z \in \mathcal{Z}} \big(\mathbf{X}_c^T \mathbf{D}_z \mathbf{w} - \boldsymbol{\nu}\big)^T \Sigma \big(\mathbf{X}_c^T \mathbf{D}_z \mathbf{w} - \boldsymbol{\nu}\big) \tag{4.9}$$

where $\Sigma = diag(\{\zeta_1^1, \ldots, \zeta_D^1, \zeta_1^2, \ldots \zeta_D^2, \ldots, \zeta_1^P, \ldots, \zeta_D^P\})$. We also rewrite Equation (4.8) as,

$$\rho(\mathbf{w}) = \sum_{z \in \mathcal{Z}} \lambda_z \mathbf{w}^T \mathbf{D}_z \mathbf{w}. \tag{4.10}$$

Using these two equations the total loss function as a function of $\mathbf{w}$ is defined as

$$Loss(\mathbf{w}) = CB(\mathbf{w}) + \rho(\mathbf{w}) \tag{4.11}$$

$$= \sum_{z \in \mathcal{Z}} \left[\big(\mathbf{X}_c^T \mathbf{D}_z \mathbf{w} - \boldsymbol{\nu}\big)^T \Sigma \big(\mathbf{X}_c^T \mathbf{D}_z \mathbf{w} - \boldsymbol{\nu}\big) + \lambda_z \mathbf{w}^T \mathbf{D}_z \mathbf{w} \right]. \tag{4.12}$$

As there are constraints on the parameters $\mathbf{w}$, we define a reparameterization of $\mathbf{w}$ that allows for methods for solving an unconstrained optimization problem. We first reparameterize $\mathbf{w}$ using a set of parameters $\boldsymbol{\eta}$ such that $w_i = \exp(\eta_i)$, thus removing the lower-bound constraint. We then ensure that the weights sum to one within each group z through the following

definition,

$$w_i = \sum_{z \in \mathcal{Z}} \left(\frac{\exp(\eta_i) I(Z_i = z)}{\sum_{j=1}^{N} \exp(\eta_j) I(Z_j = z)} \right). \tag{4.13}$$

Therefore each w_i is a function of the entire vector $\boldsymbol{\eta}$ through the definition in Equation (4.13) and can be used in Equation (4.12). The goal then is to find $\boldsymbol{\eta}$ such that it minimizes the loss function,

$$\hat{\boldsymbol{\eta}} = \arg\min_{\boldsymbol{\eta}} \left[CB(\boldsymbol{\eta}) + \rho(\boldsymbol{\eta}) \right] \tag{4.14}$$

$\hat{\boldsymbol{\eta}}$ is then used in Equation 4.13 to construct the weights for use within an analysis. The derivations are included in the following subsection.

4.2.1.1 Derivatives of Loss Function

We consider the derivatives of $Loss(\boldsymbol{\eta})$ in pieces and utilize the chain rule for finding expressions for its analytical derivatives. First consider $CB(\mathbf{w})$ and expand the terms,

$$CB(\mathbf{w}) = \sum_{z \in \mathcal{Z}} \left(\mathbf{X}_c^T \mathbf{D}_z \mathbf{w} - \boldsymbol{\nu} \right)^T \Sigma \left(\mathbf{X}_c^T \mathbf{D}_z \mathbf{w} - \boldsymbol{\nu} \right) \tag{4.15}$$

$$= \sum_{z \in \mathcal{Z}} \left[\mathbf{w}^T \mathbf{D}_z \mathbf{X}_c \Sigma \mathbf{X}_c^T \mathbf{D}_z \mathbf{w} - \mathbf{w}^T \mathbf{D}_z \mathbf{X}_c \Sigma \boldsymbol{\nu} - \boldsymbol{\nu}^T \Sigma \mathbf{X}_c^T \mathbf{D}_z \mathbf{w} + \boldsymbol{\nu}^T \Sigma \boldsymbol{\nu} \right]. \tag{4.16}$$

It follows, by taking the partial derivatives with respect to $\mathbf{w}$ that,

$$\frac{\partial CB(\mathbf{w})}{\partial \mathbf{w}} = \frac{\partial}{\partial \mathbf{w}} \sum_{z \in \mathcal{Z}} \left[\mathbf{w}^T \mathbf{D}_z \mathbf{X}_c \Sigma \mathbf{X}_c^T \mathbf{D}_z \mathbf{w} - \mathbf{w}^T \mathbf{D}_z \mathbf{X}_c \Sigma \boldsymbol{\nu} - \boldsymbol{\nu}^T \Sigma \mathbf{X}_c^T \mathbf{D}_z \mathbf{w} + \boldsymbol{\nu}^T \Sigma \boldsymbol{\nu} \right] \tag{4.17}$$

$$= \sum_{z \in \mathcal{Z}} \left[2 \mathbf{D}_z \mathbf{X}_c \Sigma \mathbf{X}_c^T \mathbf{D}_z \mathbf{w} - 2 \mathbf{D}_z \mathbf{X}_c \Sigma \boldsymbol{\nu} \right] \tag{4.18}$$

$$= 2\sum_{z\in\mathcal{Z}}\left[\mathbf{D}_z\mathbf{X}_c\Sigma(\mathbf{X}_c^T\mathbf{D}_z\mathbf{w}-\boldsymbol{\nu})\right]. \tag{4.19}$$

Now considering the penalty terms,

$$penalty(\mathbf{w}) = \sum_{z\in\mathcal{Z}}\lambda_z\mathbf{w}^T\mathbf{D}_z\mathbf{w} \tag{4.20}$$

$$\tag{4.21}$$

with partial derivatives

$$\frac{\partial}{\partial\mathbf{w}}penalty(\mathbf{w}) = \sum_{z\in\mathcal{Z}}2\lambda_z\mathbf{D}_z\mathbf{w}. \tag{4.22}$$

Combining the results,

$$\frac{\partial Loss(\boldsymbol{\nu})}{\partial\mathbf{w}} = 2\sum_{z\in\mathcal{Z}}\left[\mathbf{D}_z\mathbf{X}_c\Sigma(\mathbf{X}_c^T\mathbf{D}_z\mathbf{w}-\boldsymbol{\nu})+\lambda_z\mathbf{D}_z\mathbf{w}\right]. \tag{4.23}$$

For the chain rule we also need derivatives of $\mathbf{w}$ with respect to $\boldsymbol{\eta}$,

$$\frac{\partial w_i}{\partial\eta_i} = \frac{\partial}{\partial\eta_i}\left[\exp(\eta_i)\sum_{z\in\mathcal{Z}}\left(\frac{I(Z_i=z)}{\sum_{j=1}^N\exp(\eta_j)I(Z_j=z)}\right)\right] \tag{4.24}$$

$$= \exp(\eta_i)\sum_{z\in\mathcal{Z}}\left(-\frac{I(Z_i=z)\exp(\eta_i)}{(\sum_{j=1}^N\exp(\eta_j)I(Z_j=z))^2}\right) \tag{4.25}$$

$$+\exp(\eta_i)\sum_{z\in\mathcal{Z}}\left(\frac{I(Z_i=z)}{\sum_{j=1}^N\exp(\eta_j)I(Z_j=z)}\right) \tag{4.26}$$

$$= \exp(\eta_i)\left[\sum_{z\in\mathcal{Z}}\left(\frac{I(Z_i=z)}{\sum_{j=1}^N\exp(\eta_j)I(Z_j=z)}-\frac{I(Z_i=z)\exp(\eta_i)}{(\sum_{j=1}^N\exp(\eta_j)I(Z_j=z))^2}\right)\right]. \tag{4.27}$$

4.2.2 Implementation

This section describes some of the practical issues that need to be addressed for implementing the fixed target weighting approach. Throughout this paper we focus on estimating the ATE and the ATZ (i.e. treatment effects for treatment group z) and note that when estimating the ATZ we fix the weights to be uniform for treatment group z (i.e. $w_i = (\sum_{i=1}^{N} I(Z_i = z))^{-1}$ for all i in group z).

Related to the construction of the loss function, we discuss below our method for defining targets, μ_d and ν_d^p, for specific populations and subpopulations of interest and the appropriate scaling factors ζ_d^p and λ_z that have been useful within our implementations.We also discuss the use of mixed-data types (binary, continuous, ordered categorical, etc.) within $\mathbf{X}$ and the construction of the matrix $\mathbf{X}_c$. Finally, we discuss the Broyden-Fletcher-Goldfarb-Shanno (BFGS) algorithm that has been chosen for optimizing $Loss(\boldsymbol{\eta})$. The described methodology is implemented in a software package available at https://github.com/bvegetabile/optbw.

4.2.2.1 Defining parameters: μ_d, ν_d^p, ζ_d^p, and λ_z

The factors μ_d represent the expectation of the random variable X_d with density $g(x)$, i.e. $E_{g(x)}(X_d)$. In practice, if the goal is estimating the treatment effect in the population (ATE), each μ_d is set to the sample average of the X_d,

$$\hat{\mu}_d = \frac{1}{N} \sum_{i=1}^{N} X_{i,d} \tag{4.28}$$

and if the goal is estimating the treatment in group z (denoted as the ATZ, where an example is the ATT) then it is set to be

$$\hat{\mu}_d = \frac{\sum_{i=1}^{N} I(Z_i = z) X_{i,d}}{\sum_{i=1}^{N} I(Z_i = z)}. \tag{4.29}$$

These values allow us to create the centered design matrix $\mathbf{X}_c$ (see Section 4.2.1).

Now, if the goal is estimating the ATE, the parameters ν_d^p are then constructed as the sample central moments, i.e. the means of the columns of $\mathbf{X}_c$,

$$\hat{\nu}_d^p = \frac{1}{N} \sum_{i=1}^{N} (X_{i,d} - \hat{\mu}_d)^p, \tag{4.30}$$

or if the goal is the ATZ,

$$\hat{\nu}_d^p = \frac{\sum_{i=1}^{N} I(Z_i = z)(X_{i,d} - \hat{\mu}_d)^p}{\sum_{i=1}^{N} I(Z_i = z)}. \tag{4.31}$$

Next, the parameters ζ_d^p are set to control the relative contribution to the imbalance measure of the different covariate dimensions. The variance of the columns of $\mathbf{X}_c$ vary because they correspond to different powers. This affects the relative contribution of each covariate-moment contribution to the imbalance metric. We scale by the inverse of the variance of the column to take this into account. The idea being that if each column of $\mathbf{X}_c$ has unit variance, their contribution to the imbalance metrics should be reasonably comparable. Therefore, each ζ_d^P is set as follows,

$$\zeta_d^p = \left(\frac{1}{s_{c,d}^p} \right)^2 \tag{4.32}$$

where $s_{c,d}^p$ is the sample standard deviation of the column $X_{c,d}^p$.

Finally, we discuss setting the relative contribution of the effective sample size in each group z, i.e., each term λ_z. In many other application domains, parameters similar to λ are selected through cross-validation by minimizing some out-of-sample criteria. There are challenges to implementing such a strategy for the targeted optimal balancing weights. Performing cross-validation requires weights for observations that are not in the training set and as we have not specified any structure on $\boldsymbol{\eta}$ it is difficult to find such weights. Therefore, we have found that in practice it has been sufficient to set λ_z to a small value, such as $\lambda_z = 1$ for all z when estimating the ATE, and setting $\lambda_z = 0$ and all $\lambda_{z' \neq z} = 1$ when estimating the ATZ. In practice, we recommend developing sensitivity analyses that can assess the impact of the term λ on the bias and variance of the estimated treatment effect based upon an observed set of covariates and treatment assignments.

4.2.2.2 Handling Mixed-Data Types

The matrix $\mathbf{X}$ often includes covariates of different types, i.e., some columns represent covariates that are binary, continuous, ordered categorical, or categorical random variables. Our method proceeds, by assuming that all covariates can be transformed to either a binary, or a continuous covariate. That is, we treat ordinal variables as continuous variables and transform categorical variables to binary variables using a dummy variable coding. A second consideration is that binary covariates are completely specified by their first moment and therefore it does not make sense to include higher moments of binary variables. Therefore, to counteract this behavior, we construct each matrix $\mathbf{X}_c^p$ for $p > 1$ using only columns which represent continuous covariates. In simulations, this has increased performance.

4.2.2.3 Optimization Algorithm

Finally, we require a solution $\hat{\boldsymbol{\eta}} = \arg\min_{\boldsymbol{\eta}} Loss(\boldsymbol{\eta})$. As shown above, the first order-derivatives are readily available and therefore we implement the computationally efficient BFGS algorithm. The BFGS algorithm is a line search "steepest descent" method that chooses updates at each iteration, $\boldsymbol{\eta}_{i+1}$, so that the objective function is lower at step $i+1$ than it was at step i. The algorithm chooses its step size based upon a function of analytic first-derivative functions and a rank-two approximation to the inverse Hessian matrix, and in many respects is similar to Newton's method (Nocedal and Wright, 2006). The benefit of this algorithm over Newton's method, is that it does not require an inversion of the analytic Hessian at each step, a process that computationally scales as $\mathcal{O}(N^3)$, while each step of the BFGS algorithm scales at $\mathcal{O}(N^2)$ since the inverse Hessian is constructed directly. This implies that the algorithm may scale more efficiently than Newton's method with increases in the sample size. The optimization routine is implemented in the `R` package `optim`, which has been adequate for our use cases. It is our experience that setting $\eta_i = 0$ for all i as initial values provides good performance.

For a deeper discussion of line-search methods and the BFGS method, see Nocedal and Wright (2006).

4.3 Simulations

This section describes simulations studies to demonstrate the performance of the targeted optimal balancing weights in various settings. Section 4.3.1 demonstrates the performance of the method in a simple setting, estimating the ATE given a single covariate. Section 4.3.2 investigates performance for estimating the ATT in a setting with both continuous and binary

covariates. Section 4.3.3 demonstrates the performance of the method in a multi-treatment setting for estimating population-level average responses under each treatment exposure.

4.3.1 ATE Performance - A Univariate Example

To start, we present a simple simulation where the functional form of the propensity score is known and depends on a single covariate. This allows for comparisons of the targeted optimal balancing weights approach against causal estimates based on the true known propensity score or different estimates of the propensity score (such as propensity score estimates arising from logistic regression or the CBPS). We generate data from a univariate covariate $X \sim \mathcal{N}(0, 1)$ and set the true propensity score to be $e(x) = \Phi\big(a(x + b)\big)$, where $\Phi(\cdot)$ is the cumulative distribution function of a standard normal random variable. The parameter a controls the amount of relative overlap between the distribution of the covariate conditioned upon each treatment group, i.e., larger values of a correspond to more observations with values near zero or one (thus higher values exhibit more serious violations of the positivity assumption). The parameter b controls the relative proportion of the sample sizes in each group. If b is set to be a positive number more observations exhibit propensity scores greater than 0.5 and hence a greater proportion of the sample is expected in the treated group (a similar but opposite effect if b is negative.) Treatment assignments were simulated such that $Z_i|X_i = x_i \sim Bernoulli(e(x_i))$. The probability of being assigned treatment level $Z = 1$, for three choices of $(a, b) \in \{(1, 0), (1.5, 0), (1.5, -0.5)\}$, are visualized in Figure 4.1. The covariate imbalance resulting from each propensity score function is demonstrated in Figure 4.2.

Potential Outcome Setting	Treatment Response	Control Response
1) Constant Treatment Effect	$Y^{Z=1} = X + 3 + \epsilon_{Z=1}$	$Y^{Z=0} = X + \epsilon_{Z=0}$
2) Effect Modification	$Y^{Z=1} = X^2 + 2 + \epsilon_{Z=1}$	$Y^{Z=0} = X + \epsilon_{Z=0}$
3) Nonlinear Effect Modification	$Y^{Z=1} = e^X + 4X_1 + 3 + \epsilon_{Z=1}$	$Y^{Z=0} = -X^2 - e^X + \epsilon_{Z=0}$

Table 4.1: Models for Potential Outcomes

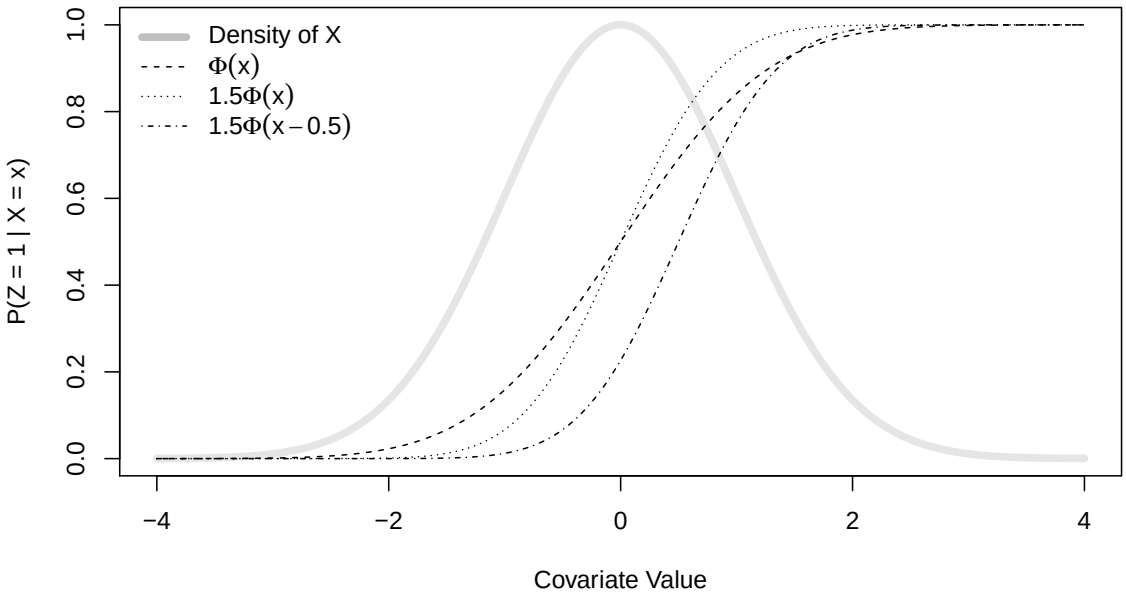

Figure 4.1: Visualization of propensity score functions considered for estimating the ATE; each dotted line represents a different propensity score function listed in the legend. The density of the covariate is overlaid in the graphic to demonstrate where the values of the propensity score are relevant.

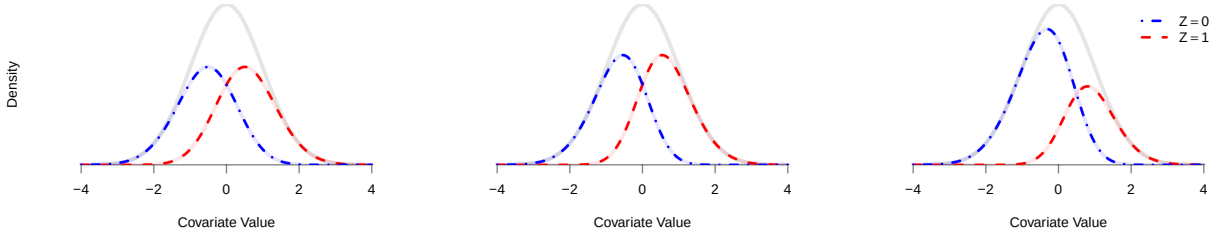

Figure 4.2: Conditional distribution of the covariate given treatment assignment, demonstrating covariate imbalances after treatment assignment. Let $\phi(x)$ be the probability density function of standard normal random variable, each conditional density is proportional to $\phi(x) * \Pr(Z = z|X = x)$ which is visualized above. The solid line is the original distribution of the covariate.

To demonstrate the performance in estimating the ATE using the nonparametric weighting estimator for the mean response in group z (see Equation (4.1)), we considered three potential outcome settings defined in Table 4.1. The first setting is labeled "Constant Treatment

Effect" and demonstrates performance estimating the ATE when the mean of the potential outcome distributions are the same linear function of X, but there is a constant difference between the functions at all points x. The second setting is labeled "Effect Modification" and considers estimating the ATE when the treatment effect is not a constant across levels of x. This setting contains a quadratic polynomial as the mean response surface for the treated outcome and a first-order polynomial for the mean response surface for the control outcome. The last setting is labeled "Nonlinear Effect Modification" and the mean response surface in each group is a function of $\exp(x)$. Each potential outcome function also includes an error term where, $\epsilon_{i,Z=z} \sim \mathcal{N}(0,1)$.

Simple comparisons of the mean response in these potential outcome settings exhibit biases that can be written as differences in the moments of the distributions of the covariates conditional on treatment assignment and therefore are settings where matching moments reduces bias. The bias in the first setting is only a function of the first moments of the conditional distributions and the bias in the second setting is a function of the first and second moments, These settings demonstrate situations where the number of moments that are required for matching are known and this information can be used in our algorithm. The final example is a setting where the bias can be written as a function of the differences in **all** unbalanced moments of the covariate distribution. This setting demonstrates the relative performance of targeted optimal balancing weights against other methods, though we only match a finite number of moments.

For each of the above data generation settings, we simulated 1000 data sets, with each data set consisting of 500 observations. In each data set, we compare the performance of seven methods for providing covariate balance and estimating the ATE. The first method compared is a "naive", or unadjusted, estimator, i.e., $\hat{\tau} = \frac{\sum_{i=1}^{N} Y_i^{obs} I(Z_i=1)}{\sum_{i=1}^{N} I(Z_i=1)} - \frac{\sum_{i=1}^{N} Y_i^{obs} I(Z_i=0)}{\sum_{i=1}^{N} I(Z_i=0)}$, which provides nominal levels of covariate imbalance and bias under each propensity score

model and potential outcome setting. The next method is adjustment by the true propensity score where, for each data set, we compute weights for estimating the ATE as defined in Table 2.2. This provides comparisons of covariate balance metrics and bias using the true data generating propensity score. The next two methods are weights based on estimates of the propensity score through logistic regression and through the covariate balancing propensity score method of Imai and Ratkovic (2014). This demonstrates settings where the true parametric model of the propensity score is known and therefore illustrates the advantage of adjustments using estimates of the propensity score, over the adjustment based upon the true propensity score. Finally, we estimate the covariate-balancing weights directly using the methodology outlined in the previous sections. Specifically, we estimate the weights using our methodology under three choices for the number of matched moments, $p = 1, 2, 3$. In comparisons of average treatment effects, the empirical distribution of the actual simulated ATE are also provided, i.e. $\tau_{sim} = \frac{1}{N}\sum_{i=1}^{N}(Y_i^{Z=1} - Y_i^{Z=0})$ for each data set. This simple simulation setting provides a comparison of our method against settings that should in theory be optimal for estimating treatment effects (adjustment by the true propensity, or when the parametric function is known and can be estimated accurately).

Under each propensity score function defined above (see Table 4.1), and for each estimation method, Table 4.2 gives results regarding covariate balance with respect to the target and the effective sample size under each weighting scheme. The results in the table are grouped together by the parameters of the propensity score function that was used to generate the data sets (vertical text at left of table). The first result of interest is that under each propensity score setting, the targeted optimal balancing weights are able to consistently achieve the lowest mean and variance imbalance metrics. This ease in balancing covariates though appears to be at the expense of effective sample size, where as the number of moments that are matched increases, the effective sample size decreases.

While balancing covariates is an important task in a causal analysis, the primary concern is the performance of the resulting weighting estimator for the treatment effect. Table 4.3 provides a comparison of the performance of the different estimation methods for the three outcome generating models. Each row gives the mean (across 1000 simulated data sets) of the bias, absolute bias and the empirical mean squared error for an estimation strategy on a given combination of propensity score and potential outcome setting. When the potential outcome included a constant treatment effect, the targeted optimal balancing weights with only the first moments matched performed the best. This is expected because the difference in the observed responses given X should only be a function of the first moment of X and therefore ensuring balancing on the mean provides the best performance. Targeted optimal balancing weights focused on balancing two moments also generally performed well. For the "Effect Modification" potential outcome model, the targeted optimal balancing weights balanced on two moments is often among the lowest in terms of performance and is comparable with CBPS or logistic regression. For the "Nonlinear Effect Modification" potential outcome model, the benefits of the targeted optimally balancing weights become more apparent. Consider the row where two moments have been balanced, here again the method provides the best performance compared against all other methods across all propensity score functions. A feature of the method, is that it is able to perform well, even when there are modest positivity violations and nonlinear response functions (as well as an imbalance in group sample sizes) as demonstrated in the setting where the propensity score model parameters were set to $(1.5, -0.5)$. As seen in the lowest right panel of the table, the targeted optimally balancing weights under all moment settings are able to provide lower bias and MSE as compared with the true propensity score and the model based estimation procedures. The results of this simulation suggest that matching the first two moments is a robust approach that should work well in application settings.

	Adjustment Method	$\lvert\bar{X}_{Z=0}-0\rvert$	$\lvert s^2_{Z=0}-1\rvert$	ESS_0	$\lvert\bar{X}_{Z=1}-0\rvert$	$\lvert s^2_{Z=1}-1\rvert$	ESS_1
$a=1, b=0$	No Adjustment	0.562	0.319	249.3	0.565	0.323	250.7
	True PS	0.125	0.183	123.1	0.118	0.180	128.3
	Logistic Regression	0.086	0.160	132.3	0.090	0.164	137.0
	CBPS	0.088	0.159	138.3	0.094	0.164	142.9
	OPTBW, $p=1$	**0.036**	0.425	**160.4**	**0.036**	0.430	**161.4**
	OPTBW, $p=2$	0.037	**0.053**	107.8	0.038	**0.053**	107.3
	OPTBW, $p=3$	**0.036**	0.058	92.7	0.037	0.059	91.5
$a=1.5, b=0$	No Adjustment	0.662	0.444	250.2	0.665	0.444	249.8
	True PS	0.194	0.246	89.6	0.189	0.249	92.2
	Logistic Regression	0.153	0.249	96.8	0.158	0.252	98.5
	CBPS	0.155	0.250	99.0	0.161	0.254	100.7
	OPTBW, $p=1$	**0.036**	0.629	**122.6**	**0.037**	0.631	**122.0**
	OPTBW, $p=2$	0.067	**0.099**	41.2	0.071	**0.102**	40.7
	OPTBW, $p=3$	0.073	0.145	31.5	0.077	0.149	31.1
$a=1.5, b=-0.5$	No Adjustment	0.460	0.375	330.3	0.899	0.500	169.7
	True PS	0.130	0.183	166.3	0.292	0.347	47.7
	Logistic Regression	0.089	0.161	166.4	0.276	0.368	54.3
	CBPS	0.086	0.159	169.6	0.281	0.368	**55.7**
	OPTBW, $p=1$	**0.035**	0.468	**230.5**	**0.061**	0.759	49.6
	OPTBW, $p=2$	0.069	**0.104**	141.4	0.200	**0.315**	19.8
	OPTBW, $p=3$	0.056	0.134	105.9	0.204	0.355	22.7

Table 4.2: Relative performance of each method for balancing covariates and maintaining an acceptable effective sample size across increasingly unbalanced propensity score settings.

4.3.1.1 Choosing λ: A Bias-Variance Relationship

Continuing with the simple univariate setting, we investigate the relationship between the choice of λ_z and the resulting bias and variance of the estimated treatment effect for the best performing number of matched moments from the previous section, i.e., setting $p = 2$. We consider the first propensity score model where $(a, b) = (1, 0)$, and estimate the treatment effect for 100 simulations, each simulation consisting of 500 observations, on a grid for $\lambda \in [0, 50]$. We calculate the mean absolute bias, the empirical standard error, and the MSE for estimating the ATE at each point λ across the simulated data sets. The results are visualized in Figure 4.3, where the upper panel of the figure demonstrates the effect of the penalty term on covariate imbalance metrics and effective sample size. As λ increases, there is a higher penalty applied to the effective sample size term and we correspondingly

		Constant			Eff. Mod.			Nonlinear		
	Empirical Distribution	**Bias**	**Abs. Bias**	**MSE**	**Bias**	**Abs. Bias**	**MSE**	**Bias**	**Abs. Bias**	**MSE**
	$\tau_{sim} = \frac{1}{N}\sum_{i=1}^{N}(Y_i^{Z=1} - Y_i^{Z=0})$	0.000	0.050	0.004	-0.003	0.078	0.010	0.000	0.302	0.143
	Adjustment Method	**Bias**	**Abs. Bias**	**MSE**	**Bias**	**Abs. Bias**	**MSE**	**Bias**	**Abs. Bias**	**MSE**
$a = 1, b = 0$	No Adjustment	1.129	1.129	1.288	0.564	0.564	0.337	2.245	2.245	5.186
	True PS	0.042	0.216	0.080	-0.040	0.241	0.138	0.037	0.784	1.454
	Logistic Regression	0.087	0.157	0.038	-0.061	0.195	0.070	-0.019	0.575	0.627
	CBPS	0.126	0.168	0.042	-0.052	0.180	0.053	0.025	0.531	0.478
	OPTBW, $p = 1$	**0.014**	**0.090**	**0.013**	-0.419	0.419	0.198	-1.019	1.021	1.166
	OPTBW, $p = 2$	0.030	0.118	0.022	**-0.001**	**0.124**	**0.025**	**-0.021**	**0.320**	**0.160**
	OPTBW, $p = 3$	0.017	0.127	0.026	-0.021	0.137	0.030	-0.140	0.327	0.167
$a = 1.5, b = 0$	No Adjustment	1.324	1.324	1.766	0.660	0.660	0.452	2.656	2.656	7.183
	True PS	0.187	0.344	0.184	-0.061	0.274	0.172	0.059	0.910	1.858
	Logistic Regression	0.253	0.290	0.114	-0.068	0.204	0.072	0.041	0.608	0.668
	CBPS	0.269	0.298	0.118	-0.062	**0.197**	0.066	0.065	0.592	0.606
	OPTBW, $p = 1$	**0.034**	**0.105**	**0.018**	-0.612	0.612	0.396	-1.443	1.443	2.185
	OPTBW, $p = 2$	0.124	0.214	0.074	**-0.008**	0.203	**0.065**	**-0.007**	**0.364**	**0.209**
	OPTBW, $p = 3$	0.143	0.255	0.108	-0.047	0.245	0.096	-0.129	0.393	0.245
$a = 1.5, b = -0.5$	No Adjustment	1.359	1.359	1.861	0.775	0.775	0.628	4.190	4.190	17.772
	True PS	0.235	0.390	0.231	**-0.189**	0.288	0.154	0.776	1.407	3.637
	Logistic Regression	0.305	0.340	0.152	-0.221	0.260	0.099	0.959	1.157	1.912
	CBPS	0.322	0.348	0.157	-0.210	**0.250**	**0.090**	0.974	1.153	1.816
	OPTBW, $p = 1$	**0.058**	**0.139**	**0.030**	-0.752	0.752	0.597	-1.047	1.048	1.202
	OPTBW, $p = 2$	0.227	0.298	0.139	-0.231	0.304	0.141	0.603	**0.747**	**0.846**
	OPTBW, $p = 3$	0.215	0.296	0.137	-0.284	0.349	0.177	**0.581**	0.770	0.907

Table 4.3: Comparison of performance of methods for the ATE by propensity score model and outcome setting.

see an increase in both the average covariate imbalance and average effective sample size. This subsequently affects the bias and empirical standard error of ATE estimates, where the bias of the estimators generally increases as a function λ and the empirical standard error generally decreases (though with a much lower slope). We note that the results for the "Effect Modification" case appear surprising in that the bias results (and subsequently empirical MSE) appear flat as a function of λ. This is due to the combination of the potential outcome functions that have been selected for this example.

These results demonstrate that the parameter λ in the effective sample size penalty plays an important role in trading off bias and variance in estimating treatment effects. We note that among the three treatment settings, it appears that a small value of λ appears to provide the best performance. In our examples, $\lambda = 1$ appears to provide consistently good performance for these treatment effect scenarios and may be an appropriate default setting in many applications. In practice, the authors suggest developing a simulation study based

upon the observed covariates and treatment assignments that can assess the sensitivity to the parameter λ.

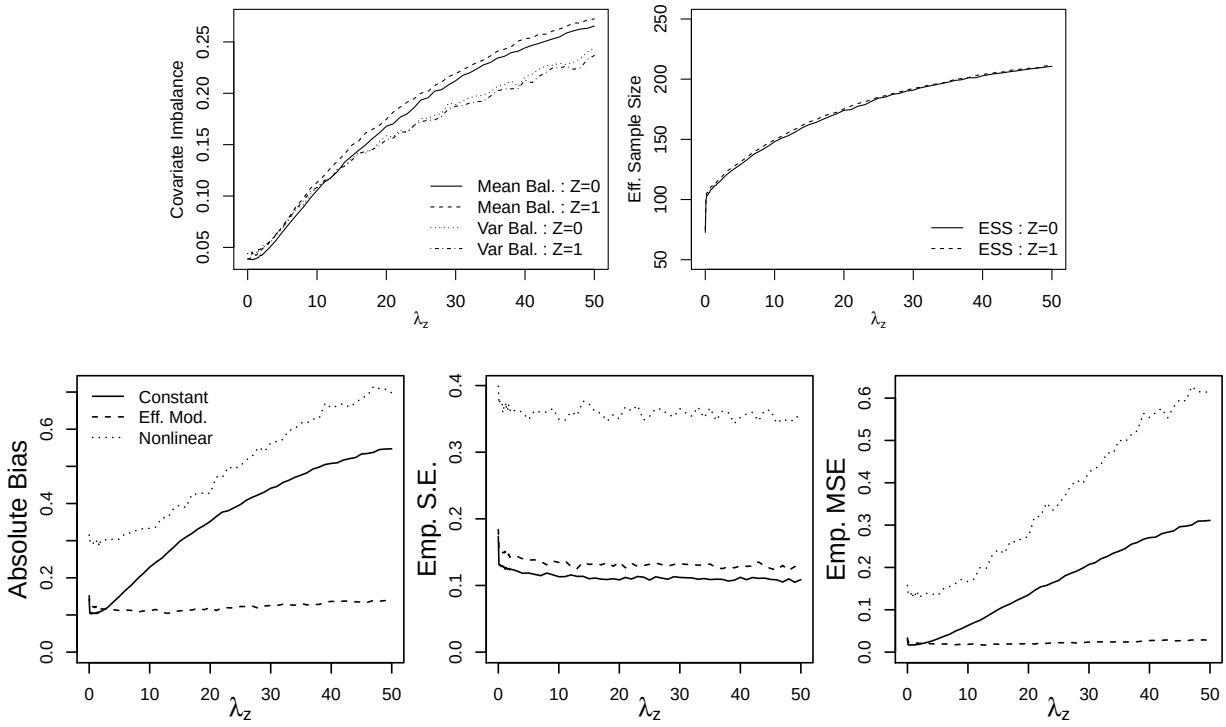

Figure 4.3: Visualization of the sample mean of various metrics across 100 simulations as a function of the effective sample size penalty parameter λ_z. The upper panel of figures contains covariate imbalance measures (similar to those in Table 4.2) and the effective sample size in each group. The lower panel is the absolute bias, empirical standard error, and empirical mean squared error as a function of λ_z across the potential outcome settings. Setting λ to approximately one appears to provide adequate performance across the potential outcome settings.

4.3.2 ATT Performance - A Multivariate Example

The results of Section 4.3.1 demonstrated the performance of targeted optimally balancing weights in estimating the ATE in a simple univariate setting. In this section, we explore the performance in estimating the ATT on 1000 simulated data sets and consider a multivariate covariate distribution consisting of both continuous and binary random variables. In this example, each $X = (X_1, X_2, X_3, X_4, X_5)^T$ is generated such that, $X_1 \sim \mathcal{N}(0, 1)$, $X_2 \sim \mathcal{N}(0, 1)$,

$X_3 \sim Bernoulli(p_3 = 0.3)$, $X_4 \sim Bernoulli(p_4 = \Phi(x_1))$ and $X_5 \sim Bernoulli(p_5 = \Phi(x_2))$. Further, the true propensity score is defined as,

$$P(Z = 1|X = x) = \Phi(f(x)) \tag{4.33}$$

where,

$$f(x) = 0.5x_1 + 0.25x_2 + 0.1x_1x_2x_3 + 0.05x_2x_5 + 0.025x_4 \tag{4.34}$$

Under this data generation setting, we then simulate $Z_i \mid X_i = x_i \sim Bernoulli(\Phi(f(x_i)))$.

In this section we consider two potential outcome settings defined in Table 4.4. The first set of potential outcomes is labeled "Constant Treatment Effect" and the two mean response functions differ by a constant at all points $X = x$. The second potential outcome is labeled "Effect Modification" and each function is similar, but the difference between the two mean functions is $\tau(x) = E(Y^{Z=1} - Y^{Z=0}|X = x) = 10x_1 - 3x_2^3$. In both settings, we assume that each $\epsilon_{i,Z=z} \sim \mathcal{N}(0, 1)$, and consider estimating the ATT. Each simulated data set consists of 500 observations.

We estimate the balancing weights using the same estimation methods outlined in the simulation of Section 4.3.1, except that the models for the CBPS and logistic regression models differ from those used in that section. For this section, the CBPS and logistic regression assumed function was $f(x, \beta) = \beta_0 + (\beta_1 x_1 + \beta_2 x_1^2) + (\beta_3 x_2 + \beta_4 x_2^2) + \beta_5 x_3 + \beta_6 x_4 + \beta_7 x_5$. Similar to the previous section, for each method and each simulation, treatment effects were estimated using ATT weights and metrics of covariate balance, bias, and mean squared error were evaluated.

Table 4.5 provides a comparison among the methods for balancing covariates. As this setting is multidimensional, we use the aggregate measure of Equation (4.9) to summarize

Potential Outcome Setting	Potential Response Functions
1) Constant Treatment Effect	$Y^{Z=1} = 5X_1^2 + X_1X_3 - 4X_2 + 50X_5 + \mathbf{10} + \epsilon_{Z=1}$
	$Y^{Z=0} = 5X_1^2 + X_1X_3 - 4X_2 + 50X_5 + \epsilon_{Z=0}$
2) Effect Modification	$Y^{Z=1} = 5X_1^2 + X_1X_3 - 4X_2 + 50X_5 + \mathbf{10X}_1 - \mathbf{3X}_2^3 + \epsilon_{Z=1}$
	$Y^{Z=0} = 5X_1^2 + X_1X_3 - 4X_2 + 50X_5 + \epsilon_{Z=0}$

Table 4.4: Models for Potential Outcomes. The bold portion of the equation in the potential outcome model for the treated group is what differs between the two potential outcome models.

the covariate imbalance across the dimensions. We compare the balance using three settings of moment conditions, i.e. $p = 1, 2, 3$, and let μ_d be the sample mean in the treated group in each simulation and calculate the appropriate target values. We also set the scale parameters to be constant for covariate dimensions and powers, $\zeta_d^p = 1$. We first evaluate the covariate balance without propensity score adjustment, i.e., by setting the weights as $w_i = \frac{I(Z_i=1)}{\sum_{i=1}^N I(Z_i=1)} + \frac{I(Z_i=0)}{\sum_{i=1}^N I(Z_i=0)}$, to provide a baseline comparison of covariate imbalance without weighting under each moment condition. This also demonstrates that the specified propensity score does indeed provide initial covariate imbalance. We then estimated balancing weights using each method and calculated the corresponding covariate imbalance. Based upon this metric, it appears that the true propensity score provides better balance than no adjustment, logistic regression and the CBPS provide superior performance than the true propensity score, and finally the targeted optimally balancing weights provide the best performance. This result is not surprising in that it is expected that estimates should provide better balance than the true propensity score and our method specifically optimizes this target function. One result that is a bit surprising is that the covariate imbalance in the final column of Table 6 (i.e., where the covariate imbalance metric incorporates the first three central moments) is 0.013 when we apply the OPTBW procedures with p = 3. We might expect to see a smaller value here as we do for the analogous entries in the first two columns. This is primarily due to the fact that we have set the scaling parameters to one for this example which ends up inflating the contribution of the third moment.

	Moments Included		
Adjustment Method	**p=1**	**p=2**	**p=3**
No Adjustment	0.706	0.739	6.024
True Propensity Score	0.045	0.178	1.385
Logistic Regression	0.010	0.050	0.550
CBPS	0.008	0.035	0.349
OPTBW, $p = 1$	$<$ 1e-3	0.057	0.298
OPTBW, $p = 2$	$<$ 1e-3	$<$ 1e-3	0.195
OPTBW, $p = 3$	$<$ 1e-3	0.001	0.013

Table 4.5: Comparison of average covariate imbalance after weighting adjustment across 1000 simulations. Covariate imbalance in the above table is defined using Equation (4.12) with the designated number of moments defined by the column headers, $\lambda_z = 0$ for all z, and μ_d set to be the sample means in the treated subgroup to define appropriate targets.

Table 4.6 organizes results for estimating the ATT under each potential outcome setting. The table is grouped into an upper section containing results for the "Constant Treatment Effect" setting, and a lower portion which contains the results for the "Effect Modification" setting, labels at left of table. The first column identifies the estimation or adjustment method and the next three columns contain estimates of the empirical distribution of estimates based upon this method. The next three columns contain performance measures for estimating the ATT, the mean bias, mean absolute bias and empirical MSE across the 1000 simulations. The first row in each section corresponds to the "true" simulated treatment effect if both potential outcomes were observed, i.e., $\tau_{sim} = \frac{\sum_{i=1}^{N} I(Z_i=1)(Y_i^{Z=1} - Y_i^{Z=0})}{\sum_{i=1}^{N} I(Z_i=1)}$. The next rows are organized similar to the previous section.

The first significant result of the table is that the estimated treatment effect results mimic the hierarchy of the covariate imbalance metrics. The true propensity score is better than no adjustment and the targeted optimal balancing weights are better than all other methods for any choice of p. The next significant finding is that the results from estimating the ATT using the targeted optimal balancing weights with $p = 2$, or $p = 3$ are similar, further supporting the recommendation that setting $p = 2$ may be a preferred default in many application settings. The last remarkable finding, is that in both potential outcome settings, the targeted optimal

		Empirical Distribution			Performance		
Constant	**Est. Treatment Effect**	**10%**	**50%**	**90%**	**Bias**	**Abs. Bias**	**MSE**
	$\tau_{sim} = \frac{\sum_{i=1}^{N} I(Z_i=1)(Y_i^{Z=1}-Y_i^{Z=0})}{\sum_{i=1}^{N} I(Z_i=1)}$	9.882	9.997	10.117	-0.002	0.072	0.008
	Adjustment Method	**10%**	**50%**	**90%**	**Bias**	**Abs. Bias**	**MSE**
	No Adjustment	11.106	14.039	16.679	3.935	4.009	20.243
	True Propensity Score	6.280	10.294	13.690	0.046	2.349	9.196
	Logistic Regression	8.765	10.613	12.287	0.545	1.240	2.544
	CBPS	10.008	10.687	11.665	0.782	0.806	1.056
	OPTBW, $p = 1$	9.401	10.387	11.384	0.396	0.702	0.772
	OPTBW, $p = 2$	9.864	10.040	10.217	**0.042**	**0.116**	**0.022**
	OPTBW, $p = 3$	9.875	10.052	10.246	0.062	0.129	0.030
Effect Modification	**Est. Treatment Effect**	**10%**	**50%**	**90%**	**Bias**	**Abs. Bias**	**MSE**
	$\tau_{sim} = \frac{\sum_{i=1}^{N} I(Z_i=1)(Y_i^{Z=1}-Y_i^{Z=0})}{\sum_{i=1}^{N} I(Z_i=1)}$	0.568	1.764	3.028	0.000	0.787	0.955
	Adjustment Method	**10%**	**50%**	**90%**	**Bias**	**Abs. Bias**	**MSE**
	No Adjustment	2.746	5.790	8.661	3.933	4.050	21.096
	True Propensity Score	-2.036	2.098	5.683	0.045	2.447	9.907
	Logistic Regression	0.112	2.386	4.568	0.544	1.479	3.534
	CBPS	1.116	2.484	4.170	0.780	1.128	2.117
	OPTBW, $p = 1$	0.559	2.166	3.888	0.393	1.073	1.828
	OPTBW, $p = 2$	0.598	1.800	3.121	**0.038**	**0.808**	**1.017**
	OPTBW, $p = 3$	0.627	1.811	3.115	0.057	**0.808**	1.019

Table 4.6: Comparison of performance of methods for estimating the ATT by potential outcome setting.

balancing weights leads to a distribution of estimated treatment effects that is similar to the true distribution of (i.e., the distribution of τ_{sim}). These results demonstrate the utility of the targeted optimal balancing weights in this higher dimensional setting and for estimating the ATT.

4.3.3 ATE Performance - A Multi-Treatment Example

The previous two sections demonstrated performance for studies with two treatments. This section demonstrates performance of the targeted optimal balancing weights in the multi-treatment setting. We consider a setting with three treatment exposures, i.e., $Z = z \in \{1, 2, 3\}$, and consider estimating the mean response in the population under each treatment exposure. The goal is estimating $E(Y^{Z=z})$, using the estimator $\bar{Y}_w^{Z=z}$ with weights constructed using inverse-probability of treatment weighting, $w_i = \sum_{z \in \mathcal{Z}} \frac{I(Z_i=z)}{P(Z_i=z|X_i=x)}$. Treatment

effects could then be constructed as relevant contrasts of these mean response functions. Our goal in this section is demonstrating performance in estimating these mean responses and we do not explicitly define specific contrasts for estimands.

The simulation in this section considers 1000 simulated data sets with the same multivariate covariate distribution used in Section 4.3.2, each consisting of 500 observations. The treatment assignment mechanism is constructed as follows. Let $f_z(x) \equiv f_z$ represent a latent function of the covariates and define $P(Z = z|X = x) \equiv p_z = \frac{\exp(f_z)}{\sum_{z \in \mathcal{Z}} \exp(f_z)}$. In this example, we set the f_z as follows,

$$f_1 = 0.75(x_1 - 0.5)^2 + 0.5x_2 + 0.15x_1x_2x_3 + 0.25x_2x_5 + 0.025x_4 - 1 \tag{4.35}$$

$$f_2 = -0.25x_1 - 0.5x_2^2 + 0.75x_4 + x_5 \tag{4.36}$$

$$f_3 = 0.5x_1x_3 + 0.5x_2x_5 - x_4 \tag{4.37}$$

We then simulate treatment assignments for each observation such that $Z_i|X_i = x \sim Multinomial(1, (p_1, p_2, p_3)^T)$. In addition, we set the potential outcome response models as follows,

$$Y^{Z=1} = 5\exp(x_1) + x_3\exp(x_2) + x_1x_4x_5 + \epsilon_{Z=1} \tag{4.38}$$

$$Y^{Z=2} = x_1x_2 + 5x_1 - 15x_1x_2x_3 - 5x_2x_5 + \epsilon_{Z=2} \tag{4.39}$$

$$Y^{Z=3} = x_2^3 + x_1 - x_4 - x_5 - 10x_3 + \epsilon_{Z=3} \tag{4.40}$$

with $\epsilon_{Z=z} \sim \mathcal{N}(0, 1)$ for all z, and finally set $Y^{obs} = \sum_{z \in \mathcal{Z}} I(Z = z)Y^{Z=z}$.

We compare the performance in estimating the mean response in each treatment group of the eight methods that have been applied throughout the previous two sections. In this setting, we utilize multinomial regression as opposed to logistic regression, using the `nnet` package available in `R`. The functional form utilized for both multinomial regression and the CBPS was

the same as the functional form in Section 4.3.2. The estimated probability of each exposure group is then used to construct weights as discussed at the outset of this section.

		Empirical Dist.			Performance		
	Adjustment Method	**10%**	**50%**	**90%**	**Bias**	**Abs. Bias**	**MSE**
Treatment Group 1	$\tau_{sim} = \frac{1}{N}\sum_{i=1}^{N} Y_i^{Z=1}$	8.241	8.850	9.536	0.000	0.406	0.265
	No Adjustment	7.885	9.409	11.085	0.574	1.099	2.072
	True Propensity Score	7.977	8.876	9.888	**0.007**	0.590	0.562
	Logistic Regression	8.122	8.859	9.618	-0.027	0.466	0.354
	CBPS	8.125	8.911	9.720	0.027	0.487	0.385
	OPTBW, $p = 1$	11.167	12.850	14.837	4.020	4.020	18.273
	OPTBW, $p = 2$	8.470	9.112	9.816	0.239	0.450	0.329
	OPTBW, $p = 3$	8.113	8.726	9.404	-0.142	**0.426**	**0.280**
Treatment Group 2	$\tau_{sim} = \frac{1}{N}\sum_{i=1}^{N} Y_i^{Z=2}$	-1.966	-1.409	-0.877	0.000	0.348	0.194
	No Adjustment	-0.406	0.148	0.780	1.584	1.584	2.733
	True Propensity Score	-2.933	-1.120	0.213	0.106	1.163	4.507
	Logistic Regression	-2.462	-0.952	0.347	0.416	1.116	4.693
	CBPS	-1.985	-0.851	0.257	0.550	0.879	1.303
	OPTBW, $p = 1$	-1.365	-0.871	-0.282	0.564	**0.599**	**0.496**
	OPTBW, $p = 2$	-2.382	-1.234	0.023	0.185	0.770	0.972
	OPTBW, $p = 3$	-2.584	-1.358	-0.047	**0.050**	0.782	0.971
Treatment Group 3	$\tau_{sim} = \frac{1}{N}\sum_{i=1}^{N} Y_i^{Z=3}$	-4.342	-3.978	-3.647	0.000	0.215	0.075
	No Adjustment	-5.262	-4.428	-3.588	-0.425	0.629	0.616
	True Propensity Score	-4.915	-3.936	-3.019	0.031	0.615	0.720
	Logistic Regression	-4.614	-3.908	-3.305	0.059	0.435	0.384
	CBPS	-4.528	-3.927	-3.377	0.046	0.373	0.222
	OPTBW, $p = 1$	-4.686	-4.127	-3.592	-0.137	0.364	0.223
	OPTBW, $p = 2$	-4.376	-3.938	-3.543	0.035	0.260	0.110
	OPTBW, $p = 3$	-4.373	-3.960	-3.595	**0.010**	**0.238**	**0.092**

Table 4.7: Comparison of performance of methods for the average response in the population in a multi-treatment setting

Table 4.7 organizes results for estimating the mean responses through weighting adjustments. Similar to the previous sections, the table is organized into distinct sections. The first (top) portion are results for estimating the mean response in treatment group 1, the second (middle) portion provides results for the mean response in treatment group 2, and so on. The columns represent the adjustment methods used, the empirical distribution of simulation estimates, the mean bias, the mean absolute bias across simulations, and the empirical mean squared error of the simulation. The rows are organized as in previous sections. The primary result of Table 4.7, is that again it appears that the best method for estimating these mean responses is our targeted optimal balancing weights method followed by the CBPS method, in terms of the supplied performance metrics. What is different from the binary simulations, is that it

appears that including third order moments improves estimates in this setting.

4.4 Application to the Job Training Data of Dehejia and Wahba (1999)

To further explore performance for estimating the ATT, we utilize a subset of data from LaLonde (1986). In particular we focus on the subset of data utilized in Dehejia and Wahba (1999) and that was used in Section 3.4, as this is a familiar benchmark data set within the causal literature. For completeness within this section, we again describe the data set and present summary statistics.

The original data set was generated as part of a controlled randomized experiment to assess the effect of the National Supported Work (NSW) Demonstration, a national job training program, on post-intervention earnings for individuals within the program. The goal of LaLonde (1986) was to evaluate the extent to which observational data sources could replicate the results generated from controlled randomization. In this pursuit, LaLonde constructed two pools of individuals from non-experimental data sources for potential use as control units to compare against the NSW treated individuals; one from the Panel Study on Income Dynamics (PSID-1) and another from the Current Population Survey-Social Security File (CPS-1). When these original control samples did not work well, subsets were developed that were more similar to the treated individuals from the experimental data (denoted PSID-2,3 and CPS-2,3).

The subset of data utilized by Dehejia and Wahba (1999) is comprised of men who were assigned to treatment after 1975 and the outcome measure was post-intervention earnings in 1978. In this analysis we focus on the CPS data only (results on the PSID data are

similar). The data set contains the following variables for each individual: age, education in years, indicators of whether the individual was black or Hispanic, an indicator of whether the individual was married, an indicator of "no degree", and retrospective earnings in 1974 and 1975. Table 4.8 provides summaries of the data. It is clear that the observational data sets are different from the experimental data indicating that adjustment is necessary to remove the effects of potential confounding variables before the CPS sample can be used as a comparison group. In particular individuals within the experimental data were younger, were less often married, and had significantly less earnings in 1974 and 1975; additionally there was a higher representation of black individuals within the experimental data.

	Experimental Data		Current Population Survey-Social Security File		
	NSW - Treated	NSW - Control	CPS-1	CPS-2	CPS-3
N_{obs}	185	260	15992	2369	429
Age	25.82 (7.14)	25.05 (7.04)	33.23 (11.04)	28.25 (11.69)	28.03 (10.77)
$Education$	10.35 (2.01)	10.09 (1.61)	12.03 (2.87)	11.24 (2.58)	10.24 (2.85)
$I(Black)$	0.84 (0.36)	0.83 (0.38)	0.07 (0.26)	0.11 (0.32)	0.2 (0.4)
$I(Hispanic)$	0.06 (0.24)	0.11 (0.31)	0.07 (0.26)	0.08 (0.28)	0.14 (0.35)
$I(Married)$	0.19 (0.39)	0.15 (0.36)	0.71 (0.45)	0.46 (0.5)	0.51 (0.5)
$I(No\ degree)$	0.71 (0.45)	0.83 (0.37)	0.3 (0.46)	0.45 (0.5)	0.6 (0.49)
$RE74$	2095.57 (4873.4)	2107.03 (5676.96)	14016.8 (9569.5)	8727.96 (8965.95)	5619.24 (6780.83)
$RE75$	1532.06 (3210.54)	1266.91 (3097.01)	13650.8 (9270.11)	7397.23 (8110.5)	2466.48 (3288.16)

Table 4.8: Summaries of experimental and observational data from Dehejia and Wahba (1999). The first row is the number of observations from that data source and the next rows provide summaries of the mean and standard deviations for that specific category.

To analyze the data, we use the targeted optimal balancing weights and, motivated by the performance in the previous section, focus on matching the first two moments. Table 4.9 presents covariate balance metrics that are common in the literature both before and after the weighting adjustment. The metrics are the weighted standardized difference in the means and the logarithm of the ratio of the sample standard deviations. Let

$$\bar{X}_d(z, \mathbf{w}) = \frac{\sum_{i=1}^{N} w_i I(Z_i = z) X_{i,d}}{\sum_{i=1}^{N} w_i I(Z_i = z)} \tag{4.41}$$

and

$$s_d^2(z, \mathbf{w}) = \frac{\sum_{i=1}^{N} w_i I(Z_i = z)(X_{i,d} - \bar{X}_d(z, \mathbf{w}))^2}{\sum_{i=1}^{N} w_i I(Z_i = z)} \tag{4.42}$$

be the weighted mean and standard deviation for covariate dimension d in group z. Then the weighted standardized difference in the means and the logarithm of the ratio of the sample standard deviations are defined as,

$$\Delta_d = \frac{\bar{X}_d(1, \mathbf{w}) - \bar{X}_d(0, \mathbf{w})}{\sqrt{\frac{s_d^2(1,\mathrm{w}) + s_d^2(0,\mathrm{w})}{2}}} \qquad \text{and} \qquad \Gamma_d = \log\left(\frac{s_d(1, \mathbf{w})}{s_d(0, \mathbf{w})}\right), \tag{4.43}$$

respectively. The upper portion of the table demonstrates the lack of covariate balance before a re-weighting of the sample and the lower portion focuses on the balancing after the weighting procedure is applied. It is clear from these results that the method is able to perform well in this applied setting.

We can also look at the estimated treatment effect in the treated group based upon this weighting scheme. Table 4.10 presents results including the unadjusted estimates using a naive estimator, the results from Dehejia and Wahba (1999) (Table 3, column 8 that estimates the ATT using weighted least squares after a matching procedure based on a subset of observations), the results of Chapter 3 based upon estimates using the optimally balanced Gaussian process propensity score, and finally our targeted optimal balancing weights. The results demonstrate that in addition to providing optimal covariate balance, the estimated treatment effects agree well with the experimental results.

Covariate Balance Prior to Weighting Adjustment								
Data set	NSW Treated		CPS-1		CPS-2		CPS-3	
Balance Measure	Δ	Γ	Δ	Γ	Δ	Γ	Δ	Γ
Age	0.108	0.013	0.797	0.437	0.252	0.494	0.242	0.412
Education	0.142	0.219	0.679	0.359	0.386	0.251	0.045	0.352
I(*Black*)	0.044	0.040	2.432	0.331	2.147	0.141	1.671	0.101
I(*Hispanic*)	0.175	0.271	0.051	0.089	0.095	0.160	0.277	0.390
I(*Married*)	0.094	0.082	1.234	0.145	0.604	0.241	0.721	0.244
I(*No degree*)	0.305	0.202	0.905	0.004	0.541	0.090	0.235	0.076
*RE*74	0.002	0.153	1.570	0.675	0.919	0.610	0.597	0.330
*RE*75	0.084	0.036	1.747	1.060	0.951	0.927	0.288	0.024
Covariate Balance After Weighting Adjustment								
	Δ	Γ	Δ	Γ	Δ	Γ	Δ	Γ
Age	-	-	0.011	0.032	0.025	0.039	0.021	0.036
Education	-	-	0.002	0.006	0.002	0.006	0.000	0.008
I(*Black*)	-	-	0.001	0.001	0.002	0.002	0.011	0.010
I(*Hispanic*)	-	-	0.003	0.005	0.001	0.002	0.001	0.001
I(*Married*)	-	-	0.002	0.002	0.006	0.005	0.007	0.006
I(*No degree*)	-	-	0.003	0.001	0.003	0.001	0.004	0.002
*RE*74	-	-	0.022	0.103	0.033	0.071	0.018	0.036
*RE*75	-	-	0.022	0.013	0.012	0.058	0.003	0.010

Table 4.9: Summaries of covariate imbalance measures within each data set: the weighted standardized difference in the means and the logarithm of the ratio of the sample variances. The upper portion of the table are the metrics before weighting and the lower portion of the table are metrics after weighting using the targeted optimal balancing weights.

4.5 Discussion

Removing, or reducing, covariate imbalances is an important first step in causal analyses of observational studies. This paper presents a new loss function to select weights for causal analyses that is a composed of a metric of covariate imbalance that is a function of the discrepancy between the weighted sample central moments of the observed data and targets representing the moments of a density of interest and the effective sample size in each group. Further this paper provides a method for selecting weights for causal analyses of varying estimands through a minimization of this covariate imbalance metric for both binary and

	No Adjustment			DW99		OBGPPS		OPTBW		
Control Dataset	Est.	S.E.	N_C	Est.	S.E.	Est.	S.E.	Est.	S.E.	ESS_C
NSW Treated	**1794**	**633**	**260**	-	-	-	-	-		
CPS-1	-8498	712	15992	1616	751	1273	717	1299	714	185
CPS-2	-3822	671	2369	1563	753	1748	735	1480	801	81
CPS-3	-635	657	429	662	772	1568	808	1567	920	59

Table 4.10: Estimates of the ATT for the Dehejia and Wahba (1999) data. The first row represents the experimental results of the National Supported Work Demonstration. The next rows are estimates using data from the observational source, the Current Population Survey-Social Security File (CPS). The first grouping of columns demonstrates estimates prior to adjustment. The next grouping of columns are the results from Dehejia and Wahba (1999) (Table 3, column 8) which were estimates of the ATT using weighted least squares where "treatment observations weighted as 1, and control observations weighted by the number of times they are matched to a treatment observation". The next grouping are the previous results from Chapter 3. The final group of columns are estimates using weights estimated through the targeted optimal balancing weights. The standard errors for the OBGPPS and OPTBW methods were found using a robust sandwich variance estimator.

multi-treatment settings. A series of simulation studies demonstrate the effectiveness of the methodology as compared to a number of methods of propensity score weighting adjustment. The method is also applied to a benchmark data set from Dehejia and Wahba (1999). The primary finding is that the proposed method is highly effective at both balancing covariates and reducing bias in causal analyses of average treatment effects.

The developed method is similar in spirit to two other methods for optimal balancing weights, namely those developed in Hainmueller (2012) and Zubizarreta (2015). The primary difference between the methodology here and those methods, is that they focus on a loss function of the weights and then constrain the loss function with respect to measures of covariate imbalance. Our method focuses specifically on minimizing covariate imbalance first, and subsequently penalizes solutions due to the effective sample size resulting from this weighting scheme. The mechanics of the optimization procedures are similar, but as the bias of the estimated treatment effect under strong ignorability given the covariates can be written as a function of the remaining covariate imbalance after weighting adjustment, our method should

be preferable in practice.

One issue that must be addressed in future work is the choice of the effective sample size penalty parameters λ_z. Our results suggest that the choice of λ_z involves a trade-off between reducing bias (low λ) and reducing variance (high λ) with $\lambda = 1$ providing good empirical performance. One area of future work would be to consider placing structural assumptions on the link function underlying our weights η_i. That is, it may make sense to assume that η_i is a linear function of the covariates, e.g. $\eta_i = x_i^T\beta$, or potentially by an assumption that the collection of η_i were generated from a Gaussian process with a specific covariance structure. By placing structure on these parameters, it would be possible to estimate the weights for new observations and a cross-validation strategy could be developed to select both optimal $\mathbf{w}$ and λ. Changing the form of η_i is not too difficult because in many cases the required changes to the derivative functions can be carried out through the chain-rule and the optimization routines simply updated.

Finally, the computational properties of the method can be explored further. In one example data set, the observed runtime was approximately 2 minutes for finding the weights for estimating the ATE on a data set of 10,000 observations[1]. This is made possible through the use of the unconstrained optimization procedure and the current software implementation of the method using the prepackaged version of the BFGS method within R. While this implementation of the method is useful for data sets of a modest size, future work will focus on optimizing these routines and exploring alternative optimization procedures to further decrease the runtime of the method so that it may be scaled to larger data sets.

[1]Runtime performance on a 2013 iMac with a 3.4 GHz Intel Core i5 (4 cores) with 16GB of memory

Chapter 5

A Comparison Between the Methods

The focus of this dissertation has been the development of methods for optimal covariate balance in observational studies in the pursuit of measuring causal effects. Chapter 3 discussed the optimally balanced Gaussian process propensity score. The focus of that method is the assumption that the propensity score can be estimated through a latent function that is assumed to arise from a Gaussian process. The hyperparameters of the process were selected in Chapter 3 to minimize a metric of covariate balance that was tailored for binary treatment settings. Chapter 4 discussed an alternative strategy that developed unit weights that minimize a metric of covariate imbalance that was a function of the differences between the weighted moments and a series of distributional targets that represent a subpopulation of interest, and a function of the resulting effective sample size of that weighting. It was demonstrated that this covariate imbalance measure could be optimized directly to select the weights, without assumptions on the functional form underlying the weights. Both methods were demonstrated to be superior, in many cases, to other comparable estimation methods that adjust for covariate imbalances (often through the estimation of propensity scores). In this chapter, we revisit a few simulation settings from the previous chapters to provide a

direct comparison between the two methods.

5.1 Overview

In this chapter, we focus on directly comparing the optimally balanced Gaussian process propensity score developed in Chapter 3 and the targeted optimal balancing weights developed in Chapter 4. The first comparison involves the simulation study of 3.3.3, evaluating the performance of the methods in estimating the ATE in a simple setting. The second comparison is the simulation setting of Section 4.3.2, estimating the ATT in a multivariate covariate setting with mixed-data types. The final section compares performance in the setting of Section 4.3.3, i.e, the performance in a multi-treatment setting for estimating average responses under different exposure types. Under each setting, we compare estimates of the targeted optimal balancing weights with $p = 1, 2, 3$ moments, against the best performing optimally balanced Gaussian process propensity score from Chapter 3.

5.2 ATE Simulation Setting of 3.3.3

This section evaluates the performance of the methods in estimating the average treatment effect in the population for the simulation setting of Section 3.3.3 (which is reviewed below). In each simulation, we consider estimating balancing weights using the optimally balanced Gaussian process propensity score using the expectation propagation approximation estimation approach with the additive normalized-first-order polynomial squared exponential kernel and then again using the targeted optimal balancing weights with $p = 1, 2, 3$.

We provide a brief review of the simulation setting for completeness, see Section 3.3.3 for a

more thorough discussion. The setting considers two propensity score functions for treatment assignment where each function is of the form,

$$P(Z_i = 1|X_i = x_i) = \alpha_1 \times \Phi\left(g_j(x_i, \beta)\right) + \alpha_2,$$

and each $g_j(x_i, \beta)$ is described in Table 5.1. The covariates $X_1 = x_1$ and $X_2 = x_2$ were simulated such that $X_1 \sim \mathcal{N}(0, 1)$ and $X_2 \sim Bernoulli(0.4)$. In addition, the performance is evaluated under three different potential outcome settings listed in Table 5.2 for each propensity score model. Each simulated data set contains 500 observations, and we construct 1000 simulated data sets.

Setting	$g_j(x_i, \beta)$	Parameters, $\gamma = (\beta_0, \beta_1, \beta_2, \beta_3, \alpha_1, \alpha_2)$
Non-GLM, Linear w/ Interactions	$\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_2$	(0.5, 4, -0.5, -3, 0.7, 0.15)
Non-GLM, Second Order	$\beta_0 + \beta_1 x_1 + \beta_2 x_1^2 + \beta_3 x_2$	(2.5, 3, -4, -2, 0.75, 0.125)

Table 5.1: Parameter settings used to create propensity score functions

Potential Outcome Setting	Treatment Response	Control Response
1) Constant Treatment Effect	$Y^{Z=1} = X_1 + 3 + \epsilon_{Z=1}$	$Y^{Z=0} = X_1 + \epsilon_{Z=0}$
2) Effect Modification	$Y^{Z=1} = X_1^2 + 2 + \epsilon_{Z=1}$	$Y^{Z=0} = X_1 + \epsilon_{Z=0}$
3) Nonlinear Effect Modification	$Y^{Z=1} = \exp(X_1) + 4X_1 + 3 + \epsilon_{Z=1}$	$Y^{Z=0} = -X_1^2 - \exp(X_1) + \epsilon_{Z=0}$

Table 5.2: Models for potential outcomes

Table 5.3 provides results related to performance in achieving covariate balance. Each column of the table reports the proportion of the simulations for which mean balance was obtained across all dimensions for various thresholds, i.e., $|\Delta_d| < \delta$ for all d and specified δ. The first result is that while the optimally balanced Gaussian process propensity score provided superior performance as compared with all methods tested in Section 3.3.3, the targeted optimal balancing weights provides even better performance under these simulation settings. Note that the propensity score functions that generated the data would be difficult to model with logistic regression due to the α parameters and were originally constructed to highlight the strengths of the optimally balanced Gaussian process methodology.

	Prop. Bal. ($\lvert\Delta_d\rvert < \delta$ *for all* d)					
	Non-GLM, Linear w/ Interactions			*Non-GLM, Second Order*		
Adjustment Method	$\delta = 0.1$	$\delta = 0.15$	$\delta = 0.2$	$\delta = 0.1$	$\delta = 0.15$	$\delta = 0.2$
No Adjustment	0.000	0.000	0.000	0.000	0.000	0.000
True Propensity Score	0.372	0.642	0.842	0.325	0.593	0.798
Opt. Bal. GP PS (NPSE) - EP	0.997	1.000	1.000	0.971	0.997	1.000
OPTBW, $p = 1$	1.000	1.000	1.000	1.000	1.000	1.000
OPTBW, $p = 2$	1.000	1.000	1.000	1.000	1.000	1.000
OPTBW, $p = 3$	1.000	1.000	1.000	1.000	1.000	1.000

Table 5.3: Comparison in balancing covariates between the Optimally Balanced Gaussian Process Propensity Score methodology and the Targeted Optimal Balancing Weight methodology under the scenario of Section 3.3.3. Simulation results demonstrate the performance in balancing covariates. The columns provide the proportion of the 1000 simulations which were declared to be mean balanced, i.e., $|\Delta_d| < \delta$ for all d and for various thresholds of δ

Tables 5.4 and 5.5 organize results related to estimating the ATE under each true propensity score function (in separate tables) and each potential outcome setting (labeled vertically at left of each table). The tables provide the following measures to evaluate the performance in estimating the ATE: the mean bias, the mean absolute bias, the empirical standard error and mean squared error across simulations. The primary finding across these results is that the targeted optimal balancing weights provide superior performance when evaluated against the optimally balanced Gaussian process propensity score (which itself provided better performance when compared with many common methods of propensity score estimation). It is demonstrated in these simulations that the optimally balanced Gaussian process propensity score often does better than the targeted balancing weights with only one moment matched, but when higher moments are included in the targeted optimal balance weighting method, it performs better.

	Adjustment Method	Bias	Abs. Bias	%ABR	Emp. S.E.	MSE
Constant	No Adjustment	0.963	0.963	0.000	0.083	0.934
	True Propensity Score	0.004	0.102	89.273	0.128	0.016
	Opt. Bal. GP PS (NPSE) - EP	**0.004**	0.026	97.329	0.032	**0.001**
	OPTBW, $p = 1$	0.011	**0.024**	**97.560**	**0.027**	**0.001**
	OPTBW, $p = 2$	0.012	**0.024**	97.465	0.028	**0.001**
	OPTBW, $p = 3$	0.017	0.027	97.200	0.029	**0.001**
Eff. Mod.	No Adjustment	0.469	0.469	0.000	0.109	0.232
	True Propensity Score	0.014	0.125	60.772	0.156	0.025
	Opt. Bal. GP PS (NPSE) - EP	-0.009	0.073	77.536	0.092	0.009
	OPTBW, $p = 1$	-0.123	0.158	54.124	0.143	0.036
	OPTBW, $p = 2$	**0.004**	**0.064**	**81.203**	**0.082**	**0.007**
	OPTBW, $p = 3$	**0.004**	**0.064**	**81.203**	**0.082**	**0.007**
Nonlinear	No Adjustment	1.843	1.843	0.000	0.380	3.541
	True Propensity Score	0.001	0.451	73.608	0.566	0.320
	Opt. Bal. GP PS (NPSE) - EP	-0.036	0.318	81.037	0.397	0.159
	OPTBW, $p = 1$	-0.231	0.420	74.329	0.462	0.267
	OPTBW, $p = 2$	0.038	0.322	81.394	0.401	0.162
	OPTBW, $p = 3$	**-0.014**	**0.295**	**82.428**	**0.370**	**0.137**

Table 5.4: Simulation results for estimating the ATE in the non-GLM setting where the true propensity score was linear and included interaction terms

	Adjustment Method	Bias	Abs. Bias	%ABR	Emp. S.E.	MSE
Constant	No Adjustment	0.421	0.421	0.000	0.089	0.185
	True Propensity Score	-0.000	0.104	73.352	0.131	0.017
	Opt. Bal. GP PS (NPSE) - EP	-0.029	0.042	89.231	0.044	0.003
	OPTBW, $p = 1$	**0.005**	**0.019**	**95.297**	**0.024**	**0.001**
	OPTBW, $p = 2$	**0.005**	0.023	94.406	0.028	**0.001**
	OPTBW, $p = 3$	0.016	0.027	93.426	0.030	**0.001**
Eff. Mod.	No Adjustment	-0.323	0.324	0.000	0.094	0.113
	True Propensity Score	-0.007	0.144	38.066	0.176	0.031
	Opt. Bal. GP PS (NPSE) - EP	-0.047	0.107	62.847	0.126	0.018
	OPTBW, $p = 1$	-0.458	0.459	-48.483	0.118	0.224
	OPTBW, $p = 2$	-0.010	**0.068**	**73.492**	**0.085**	**0.007**
	OPTBW, $p = 3$	**-0.009**	**0.068**	73.308	0.086	**0.007**
Nonlinear	No Adjustment	1.385	1.385	0.000	0.330	2.028
	True Propensity Score	0.045	0.556	55.966	0.698	0.489
	Opt. Bal. GP PS (NPSE) - EP	-0.139	0.360	69.421	0.429	0.203
	OPTBW, $p = 1$	0.714	0.730	49.777	0.409	0.676
	OPTBW, $p = 2$	**0.021**	0.303	75.301	0.382	0.146
	OPTBW, $p = 3$	0.023	**0.292**	**76.286**	**0.372**	**0.139**

Table 5.5: Simulation results for estimating the ATE in the non-GLM setting where the true propensity score was a second-order polynomial.

5.3 ATT Simulation Setting of 4.3.2

In this section, we apply the simulation setting of Section 4.3.2, to compare the performance of the methods for estimating the ATT in an example with a multivariate covariate vector comprised of mixed data types. We again briefly review the simulation setting and then provide comparative performance measures, see Section 4.3.2 for a more detailed description of the simulation setting. In Section 4.3.2, it was demonstrated that the optimally balanced GP propensity score method using the expectation propagation approximation and with the squared exponential ARD kernel performed best and therefore this is the method used as the comparative method in this section. Again, we also evaluate the performance of targeted optimal balancing weights with $p = 1, 2, 3$.

Recall that in this example, each $X = (X_1, X_2, X_3, X_4, X_5)^T$ is generated such that, $X_1 \sim \mathcal{N}(0, 1)$, $X_2 \sim \mathcal{N}(0, 1)$, $X_3 \sim Bernoulli(p_3 = 0.3)$, $X_4 \sim Bernoulli(p_4 = \Phi(x_1))$ and $X_5 \sim Bernoulli(p_5 = \Phi(x_2))$. The true propensity score is defined as,

$$P(Z = 1|X = x) = \Phi(f(x)) \tag{5.1}$$

where

$$f(x) = 0.5x_1 + 0.25x_2 + 0.1x_1x_2x_3 + 0.05x_2x_5 + 0.025x_4. \tag{5.2}$$

We then simulated $Z_i \mid X_i = x_i \sim Bernoulli(\Phi(f(x_i)))$ and then generated potential outcomes $Y_i^{Z=0}, Y_i^{Z=1}$ from two potential outcome settings labeled "Constant Treatment Effect" and "Effect Modification", listed in Table 5.6.

Table 5.7 uses the covariate imbalance metric of Section 4.2.1 to compare the performance of the methods in balancing covariates. The first result is that again the targeted optimal

Potential Outcome Setting	Potential Response Functions
1) Constant Treatment Effect	$Y^{Z=1} = 5X_1^2 + X_1X_3 - 4X_2 + 50X_5 + \mathbf{10} + \epsilon_{Z=1}$ $Y^{Z=0} = 5X_1^2 + X_1X_3 - 4X_2 + 50X_5 + \epsilon_{Z=0}$
2) Effect Modification	$Y^{Z=1} = 5X_1^2 + X_1X_3 - 4X_2 + 50X_5 + \mathbf{10X}_1 - \mathbf{3X}_2^3 + \epsilon_{Z=1}$ $Y^{Z=0} = 5X_1^2 + X_1X_3 - 4X_2 + 50X_5 + \epsilon_{Z=0}$

Table 5.6: Models for potential outcomes. The bold portion of the equation in the potential outcome model for the treated group is what differs between the two potential outcome models.

balancing weights provide superior performance in balancing covariates based upon this metric than the optimally balanced Gaussian process propensity score methodology. We note that the optimally balanced Gaussian process propensity score provides performance in balancing covariates that is similar to the Covariate Balancing Propensity Score (not included in this table) under this metric and these settings (See Section 4.3.2).

	Moments Included		
Adjustment Method	**p=1**	**p=2**	**p=3**
No Adjustment	0.706	0.739	6.024
True Propensity Score	0.045	0.178	1.385
Opt. Bal. GP PS (SE-ARD) - EP	0.008	0.034	0.355
OPTBW, $p = 1$	$<$ 1e-3	0.057	0.298
OPTBW, $p = 2$	$<$ 1e-3	$<$ 1e-3	0.195
OPTBW, $p = 3$	$<$ 1e-3	0.001	0.013

Table 5.7: Comparison of average covariate imbalance after weighting adjustment across 1000 simulations. Covariate imbalance in the above table is defined using Equation (4.12) with the designated number of moments defined by the column headers, $\lambda_z = 0$ for all z, and μ_d set to be the sample means in the treated subgroup to define appropriate targets.

Table 5.8 presents results for estimating the ATT under each potential outcome model listed vertically at the left of the table. We again compare measures of mean bias, mean absolute bias, and mean squared error in estimating the ATT for each potential outcome model. The results of this table suggest that the targeted optimal balancing weights method performs better than the optimally balanced Gaussian process propensity score method. As in the previous section, the optimally balanced GP propensity score methodology provides performance that is similar to the targeted optimal balancing weights methods that only balances first

moments.

		Empirical Distribution			Performance		
	Adjustment Method	**10%**	**50%**	**90%**	**Bias**	**Abs. Bias**	**MSE**
Constant	$\tau_{sim} = \frac{\sum_{i=1}^{N} I(Z_i=1)(Y_i^{Z=1}-Y_i^{Z=0})}{\sum_{i=1}^{N} I(Z_i=1)}$	9.882	9.997	10.117	-0.002	0.072	0.008
	No Adjustment	11.106	14.039	16.679	3.935	4.009	20.243
	True Propensity Score	6.280	10.294	13.690	0.046	2.349	9.196
	Opt. Bal. GP PS (SE-ARD) - EP	9.543	10.400	11.491	0.490	0.719	0.926
	OPTBW, $p = 1$	9.401	10.387	11.384	0.396	0.702	0.772
	OPTBW, $p = 2$	9.864	10.040	10.217	**0.042**	**0.116**	**0.022**
	OPTBW, $p = 3$	9.875	10.052	10.246	0.062	0.129	0.030
	Adjustment Method	**10%**	**50%**	**90%**	**Bias**	**Abs. Bias**	**MSE**
Effect Modification	$\tau_{sim} = \frac{\sum_{i=1}^{N} I(Z_i=1)(Y_i^{Z=1}-Y_i^{Z=0})}{\sum_{i=1}^{N} I(Z_i=1)}$	0.568	1.764	3.028	0.000	0.787	0.955
	No Adjustment	2.746	5.790	8.661	3.933	4.050	21.096
	True Propensity Score	-2.036	2.098	5.683	0.045	2.447	9.907
	Opt. Bal. GP PS (SE-ARD) - EP	0.683	2.167	4.124	0.485	1.089	2.027
	OPTBW, $p = 1$	0.559	2.166	3.888	0.393	1.073	1.828
	OPTBW, $p = 2$	0.598	1.800	3.121	**0.038**	**0.808**	**1.017**
	OPTBW, $p = 3$	0.627	1.811	3.115	0.057	**0.808**	1.019

Table 5.8: Performance estimating the ATT by potential outcome setting.

5.4 Multi-Treatment Setting of Section 4.3.3

This section investigates the comparative performance of the two methods for estimating the population-level mean response to each exposure in a multi-treatment setting. As the optimally balanced Gaussian process propensity score was not specifically designed with this setting in mind, we briefly discuss how it was implemented for this case. To generalize the optimally balanced Gaussian process propensity score approach to multi-treatment settings, we assume a multinomial regression modeling framework and assume that each predictor function can be modeled using a Gaussian process (as discussed in Section 2.2.4). We make the simplifying assumption that the covariate matrices $\mathbf{K}_{\theta_z}$ are similar with shared hyperparameters and constructed using the normalized-polynomial squared exponential kernel with common length-scale among the dimensions. To estimate the posterior distribution of

the functions, $\mathbf{f}|\mathbf{Z}, \mathbf{X}, \Theta$, we use the Laplace approximation as described in Rasmussen and Williams (2006). In order to measure covariate imbalance, we use the balance metric defined in Equation 4.9 of Section 4.2.1 and to be consistent with the binary treatment optimally balanced Gaussian process propensity score, we include only the first two moments in the balance metric. We then select the hyperparameters θ_z that minimize Equation 4.9, using the BOBYQA optimization algorithm. This provides us with a method to compare against the targeted optimal balancing weights method of Chapter 4.

The simulation of this section is the setting of Section 4.3.3, reviewed briefly here. The simulation starts with the same five-dimensional covariate distribution used in Section 4.3.2 (see also Section 5.3). The propensity score functions are defined as $P(Z = z|X = x) \equiv p_z = \frac{\exp(f_z)}{\sum_z \exp(f_z)}$, where $f_z(x) \equiv f_z$ represents a latent function of the covariates. In particular, we set the f_z as follows,

$$f_1 = 0.75(x_1 - 0.5)^2 + 0.5x_2 + 0.15x_1x_2x_3 + 0.25x_2x_5 + 0.025x_4 - 1 \tag{5.3}$$

$$f_2 = -0.25x_1 - 0.5x_2^2 + 0.75x_4 + x_5 \tag{5.4}$$

$$f_3 = 0.5x_1x_3 + 0.5x_2x_5 - x_4 \tag{5.5}$$

We simulated treatment assignments for each observations from the propensity model. Finally, we set the three potential outcome response models as follows,

$$Y^{Z=1} = 5\exp(x_1) + x_3\exp(x_2) + x_1x_4x_5 + \epsilon_{Z=1} \tag{5.6}$$

$$Y^{Z=2} = x_1x_2 + 5x_1 - 15x_1x_2x_3 - 5x_2x_5 + \epsilon_{Z=2} \tag{5.7}$$

$$Y^{Z=3} = x_2^3 + x_1 - x_4 - x_5 - 10x_3 + \epsilon_{Z=3} \tag{5.8}$$

with $\epsilon_{Z=z} \sim \mathcal{N}(0, 1)$ for all z, and finally set $Y^{obs} = \sum_z I(Z = z)Y^{Z=z}$. Results are based on 1000 simulations each consisting of 500 observations.

We compare the performance of the two approaches in estimating the mean response in each treatment group and evaluate performance based on a comparison of the empirical distribution of the estimates with the distribution (across simulations) of $\tau_{sim} = \frac{1}{N}\sum_{i=1}^{N} Y_i^{Z=z}$. We also evaluate the bias, the absolute bias, and the empirical mean squared error of the estimator. The results are contained in Table 5.9. As in the previous sections, it is demonstrated that targeted optimal balancing weights provide better performance, though the optimally balanced Gaussian process propensity score is comparable with the estimates based upon other methods previously compared in Table 4.7.

		Empirical Dist.			**Performance**		
	Adjustment Method	**10%**	**50%**	**90%**	**Bias**	**Abs. Bias**	**MSE**
Treatment Group 1	$\tau_{sim} = \frac{1}{N}\sum_{i=1}^{N} Y_i^{Z=1}$	8.241	8.850	9.536	0.000	0.406	0.265
	No Adjustment	7.885	9.409	11.085	0.574	1.099	2.072
	True Propensity Score	7.977	8.876	9.888	0.007	0.590	0.562
	Opt. Bal. GP PS (NPSE)	8.546	9.327	10.200	0.463	0.623	0.636
	OPTBW, $p = 1$	11.167	12.850	14.837	4.020	4.020	18.273
	OPTBW, $p = 2$	8.470	9.112	9.816	0.239	0.450	0.329
	OPTBW, $p = 3$	8.113	8.726	9.404	**-0.142**	**0.426**	**0.280**
Treatment Group 2	$\tau_{sim} = \frac{1}{N}\sum_{i=1}^{N} Y_i^{Z=2}$	-1.966	-1.409	-0.877	0.000	0.348	0.194
	No Adjustment	-0.406	0.148	0.780	1.584	1.584	2.733
	True Propensity Score	-2.933	-1.120	0.213	0.106	1.163	4.507
	Opt. Bal. GP PS (NPSE)	-1.446	-0.772	-0.027	0.662	0.719	0.732
	OPTBW, $p = 1$	-1.365	-0.871	-0.282	0.564	**0.599**	**0.496**
	OPTBW, $p = 2$	-2.382	-1.234	0.023	0.185	0.770	0.972
	OPTBW, $p = 3$	-2.584	-1.358	-0.047	**0.050**	0.782	0.971
Treatment Group 3	$\tau_{sim} = \frac{1}{N}\sum_{i=1}^{N} Y_i^{Z=3}$	-4.342	-3.978	-3.647	0.000	0.215	0.075
	No Adjustment	-5.262	-4.428	-3.588	-0.425	0.629	0.616
	True Propensity Score	-4.915	-3.936	-3.019	0.031	0.615	0.720
	Opt. Bal. GP PS (NPSE)	-4.403	-3.859	-3.360	0.120	0.343	0.183
	OPTBW, $p = 1$	-4.686	-4.127	-3.592	-0.137	0.364	0.223
	OPTBW, $p = 2$	-4.376	-3.938	-3.543	0.035	0.260	0.110
	OPTBW, $p = 3$	-4.373	-3.960	-3.595	**0.010**	**0.238**	**0.092**

Table 5.9: Performance estimating the average response in the population in a multi-treatment setting

5.4.1 Comparisons of Runtimes

The results of the previous sections in this Chapter demonstrate that the targeted optimal balancing weights provide superior performance in estimating average causal effects, as well as for balancing covariates, when compared with the optimally balanced Gaussian process propensity score. This section provides another important comparison regarding the computational runtime of the developed methods. The timing results of this section were calculated using a 2013 27-inch iMac with a 3.4 GHz Intel Core i5 quad-core processor and 16 GB of memory.

Figure 5.1 presents the data from this timing comparison under the simulation setting of Section 3.3.3. The propensity score that was used was the "Non-GLM, Linear w/ Interactions" propensity score and the timing results are runtimes resulting from estimating a balancing score (or balancing weights) using each of the following methods: optimally balanced Gaussian process propensity scores computed with expectation propagation and two different kernels (normalized polynomial squared exponential (NPSE) and the squared exponential (SE) with ARD) and targeted optimal balancing weights with $p = 1, 2, 3$. The timing comparisons are averaged across 10 simulated data sets for each data set size listed at the bottom of the figure. Figure 5.2 provides the same results, but the axes have been transformed to a $\log_{10}$ scale to better compare the relative increases in runtime as the data set size grows. We note that the optimally balanced Gaussian process propensity score was only considered for data sets of sizes 2000 observations or less to keep runtime reasonable, while the targeted optimal balancing weights were further estimated on data sets of up to 10000 observations.

It is clear from these figures that the targeted optimal balancing weights methodology, as currently implemented, provides better runtime performance, for each data set size. For a constraint of an anticipated runtime of 2 minutes, under this simulation setting, the optimally

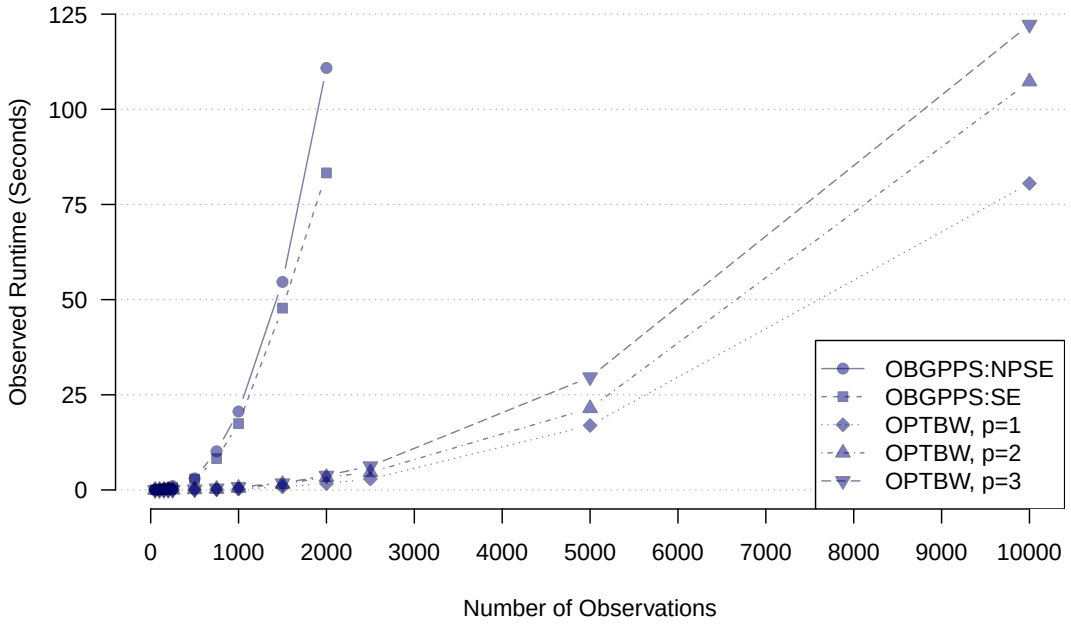

Figure 5.1: Comparison of runtime and data set size relationship between the optimally balanced Gaussian process propensity score and the targeted optimal balancing weights methodology. Each point within each line is based upon estimating the results across 10 instances, each based on a similar data set.

balanced Gaussian process propensity score can only process 2000 observations, while the targeted optimal balancing weights methodology can estimate weights on data set sizes of up to 10000 observations. Appealing to Figure 5.2, a simple extrapolation suggests that the optimally balanced GP method will take approximately 16 minutes (1000 seconds) to process 4000 observations, suggesting that the method scales poorly. The targeted optimal balancing weights appears to scale better, but may perform relatively slowly at data set sizes of size 50,000 or more.

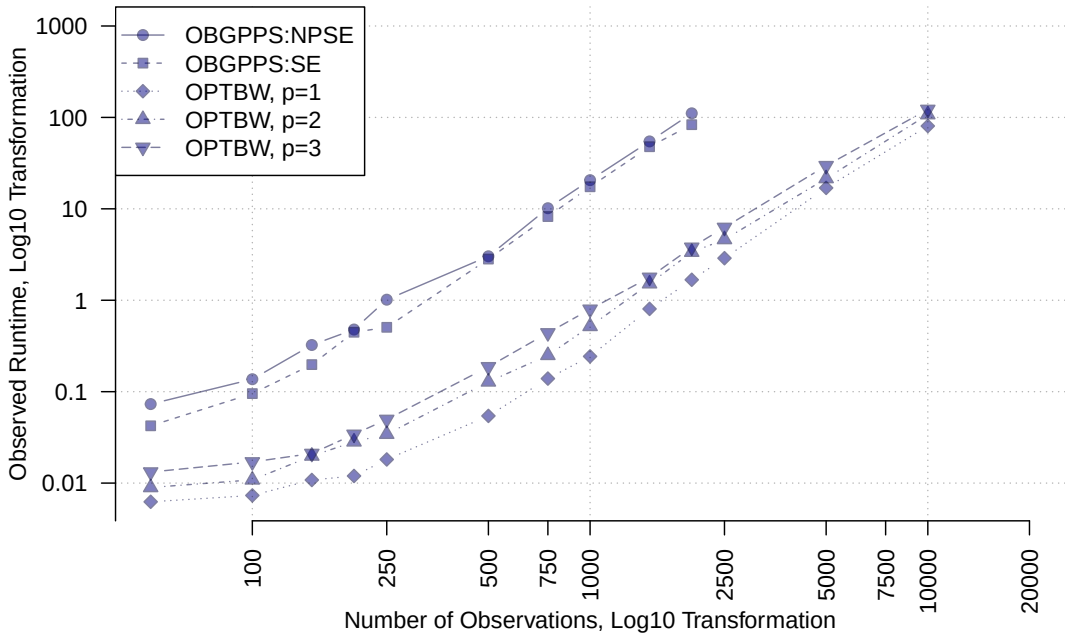

Figure 5.2: Comparison of runtime between the optimally balanced Gaussian process propensity score and the targeted optimal balancing weights methodology, based on a similar data set where the timing results have been averaged across 10 instances.

Chapter 6

A Discussion on Future Directions

This thesis has presented two contributions to the balancing and propensity score literatures for causal inference within the potential outcomes framework. In particular, the preceding chapters focused on selecting weights for weighting estimators of the mean response to varying treatments or exposures for specific subpopulations of interest. Chapter 3 introduced a method for estimating the propensity score based on a Gaussian process binary regression framework. A weighted measure of covariate imbalance was developed to assess the degree of covariate imbalance within an observational study with two treatment exposures, and then the hyperparameters of the Gaussian process were selected to minimize this balance metric. The resulting propensity score estimates can be used in weighting estimators or other approaches to causal inference. Chapter 4 focused on constructing a covariate imbalance metric that could be used with both binary and multi-level exposures. The loss function constructed in that chapter is a function of the discrepancy between the weighted sample moments observed in the data and a series of distributional targets representing the moments of a target subpopulation of interest and also incorporates a penalty term that is a function of the effective sample size resulting from that weighting. In that chapter, it was demonstrated

that this loss function could be optimized directly without an assumption on the functional form of the weights. Chapters 3 and 4 compared new approaches to a number of existing methods in simulations and in a benchmark causal inference data set. Chapter 5 compared the two new methods and demonstrated that while both methods are highly effective at balancing covariates and for use in estimating treatment effects as compared with many common forms of propensity score estimation, the targeted optimal balancing weighting methodology is best in terms of performance, and the better of the two in terms of computational runtime.

These results introduce a number of new research questions that should be explored in the future. In particular, the previous chapter demonstrated that there are issues related to the scalability of the two methods, There has been much written recently regarding "big data" and the simulation results presented would hardly be described as "large" data sets. Another issue is that a majority of this dissertation has focused on average treatment effects for populations, or for subpopulations. Estimation of average treatment effects is not the only application of propensity score methodology and performance within a doubly robust framework is worth investigating (Robins et al., 1994; Funk et al., 2011). Finally, an interesting research area considers the generalizability of studies to new populations, and we discuss how these methods could potentially be used to achieve that aim. These research ideas are discussed in further detail below,

- **Scalability**

  Chapters 3 and 4 demonstrate that the developed methods can provide superior performance for estimating treatment effects when compared with many commonly employed methods available in the literature. The issue though with many optimization-based balancing score methods is computational tractability. Chapter 5 assessed the runtimes of these methods and suggests that the primary hurdle related to their wide-spread adoption in a "big data" era will be that of scalability to larger sample sizes.

If we focus first on the Gaussian process methodology, there is a significant literature related to the scalability of Gaussian process classification methods that may be applicable to our methodology. In particular, the advances developed in Hensman et al. (2013, 2015) and Hernandez-Lobato and Hernandez-Lobato (2016) regarding scalable Gaussian process classification should be helpful. Hensman et al. (2013, 2015) develop variational approximations to Gaussian process classification and Hernandez-Lobato and Hernandez-Lobato (2016) develop a scalable EP methodology for Gaussian process classification. Each of these methods rely on a low-rank approximation to the covariance matrix through the use of what are often referred to as 'inducing points' (Quiñonero-Candela and Rasmussen, 2005). The assumption is sometimes referred to as the FITC, Fully Independent Training Conditional, approximation. These "inducing point" methods make the assumption that only a few points in the covariate space are important for estimation of the posterior distribution of all latent values and that function values at these inducing points arise from a Gaussian process. They then assume that all other functional values are independent of each other given the latent inducing points. This structure allows for a low-rank approximation of the covariance matrix that reduces the computational burden of the method. Implementing this methodology within our framework is not straight-forward due to the fact that the objective function that we optimize in our applications is a measure of covariate imbalance and not a marginal likelihood (for which these methods have been developed). The method of Hernandez-Lobato and Hernandez-Lobato (2016)is encouraging in that the derivatives of the hyperparameters of the latent process are analytical and could be incorporated into an optimization routine that focuses on covariate imbalance metrics. These methods should be investigated further.

The targeted optimal balancing weights methodology is more promising computationally, but there are still significant advances that can be made to reduce the runtime.

The optimization algorithm that is being implemented is the BFGS algorithm, which has provided stable results within our simulation settings. There are variants of this method that have been successful in much larger applications though, specifically the L-BFGS variant of this method which is a 'limited-memory' quasi-Newton optimization procedure (Nocedal and Wright, 2006). The general idea is similar to the idea above in that, to reduce computational runtime, a low-rank approximation to the inverse-Hessian is utilized that is constructed using only a limited number of previous steps in the algorithm. This L-BFGS method has been implemented in many available `R` packages, but the convergence of these algorithms has not been reliable for our loss function and the explored data set sizes. Alternative optimization strategies should be considered as they may significantly reduce the computational burden of this method in larger sample sizes.

- **Integration within Doubly Robust Methods, and Targeted Learning Methodology**

  The focus of this dissertation has been on the estimation of average treatment effects within specific populations, or subpopulations of interest, using nonparametric weighting estimation. An alternative set of methods for causal inference in observational studies consider estimating treatment effects through both an estimation of the response surface and the propensity score (Robins et al., 1994; Funk et al., 2011; Schuler and Rose, 2017). These methods have been referred to as "doubly robust", if either the propensity score model or the outcome regression model are specified correctly, then unbiased estimation of causal effects is possible (Funk et al., 2011). The methods in Chapter 3 may be directly applicable to these estimation frameworks, but the performance of the targeted balancing weights and other optimally constructed weights should be investigated in these settings as well. Regardless, there have been successes incorporating propensity score methodology and response surface methodology and the integration of the methods developed in this thesis into that literature is worthy of exploration (Dorie

et al., 2017).

- **Generalizability/Replicability of Results**

  The previous chapters have demonstrated that within a single study, with unknown treatment assignment mechanism, the methods presented in this dissertation are accurate for estimating treatment effects under the assumptions of strong ignorability and SUTVA. Often the results from one study are used to help plan another study, and in particular may be used for predicting responses in future populations. This motivates a few questions related to our methods such as: (1) Given a new set of individuals from the same population, or subpopulation, will the results be replicated? (2) Can the results from one data set be "generalized" to a sample from a new population or a new subpopulation.

  The concept of replicability is tied to the idea of prediction, and it is of interest to investigate the predictive distributions of treatment effects under these optimally balanced weighting estimators. As the weights in each method have been optimized for estimating treatment effects for this **specific** sample, it motivates the question: are there ways to estimate the weights so that they will be consistent across samples? One solution would be the development of new methods for cross-validation within a propensity score estimation procedure, such as creating a measure for optimal balance that is evaluated on out-of-sample validation data sets. The concept of optimal balance across multiple data sets may be of interest for future exploration.

  The notion of generalizability concerns how the results of a study on one subpopulation (or population) may relate to the *expected* results in a new subpopulation. This idea was somewhat explored in LaLonde (1986), where there was access to both a ground truth control set of individuals (from the randomized study) and a set of potential control individuals from a database. Li et al. (2017) defined the concept of the overlap

distribution (the distribution where there is sufficient positive probability of receiving either treatment in a binary setting) that directly relates to this concept of generalizability. Using the overlap distribution allows for a characterization of the distribution that is most comparable between the exposure groups in a study. It may be appropriate to apply this concept to assess the overlap of different data sources. Alternatively, the targeted optimal balancing weights methods allows for the specification of the moments of a target distribution and the discrepancy can be calculated between a weighted version of the data and this target distribution. It is of interest to further explore the distributional properties of the weighted covariate imbalance measure of Equation 4.9 in order to characterize when there is support within a data set for a target distribution. This would allow for a characterization of how well this sample could generalize to a new sample of individuals.

Causal inference is an ever expanding discipline that intersects with many traditional fields of study, such as, statistics, economics, medicine, computer science, and the other social sciences. Each of these fields is often tasked with analyzing data that arises from sources that are observational in nature. The work developed within this dissertation provides a significant contribution to balancing and propensity score methodology areas that should be of interest to many of these fields, as well as motivating many research topics for future investigation.

Bibliography

R. Almond, P. Deane, T. Quinlan, M. Wagner, and T. Sydorenko. A preliminary analysis of keystroke log data from a timed writing task. *ETS Research Report Series*, 2012(2): i–61, 2012. ISSN 2330-8516. doi: 10.1002/j.2333-8504.2012.tb02305.x. URL `http://dx.doi.org/10.1002/j.2333-8504.2012.tb02305.x`.

J. M. Amigó, J. Szczepański, E. Wajnryb, and M. V. Sanchez-Vives. Estimating the entropy rate of spike trains via lempel-ziv complexity. *Neural Computation*, 16(4):717–736, April 2004. ISSN 0899-7667. doi: 10.1162/089976604322860677.

T. Z. Baram, E. P. Davis, A. Obenaus, C. A. Sandman, S. L. Small, A. Solodkin, and H. Stern. Fragmentation and unpredictability of early-life experience in mental disorders. *American Journal of Psychiatry*, 169(9):907–915, 2012. doi: 10.1176/appi.ajp.2012.11091347. URL `https://doi.org/10.1176/appi.ajp.2012.11091347`. PMID: 22885631.

H. A. Chipman, E. I. George, R. E. McCulloch, et al. BART: Bayesian additive regression trees. *The Annals of Applied Statistics*, 4(1):266–298, 2010.

Z. Chu, S. Gianvecchio, H. Wang, and S. Jajodia. Who is tweeting on twitter: Human, bot, or cyborg? In *Proceedings of the 26th Annual Computer Security Applications Conference*, ACSAC '10, pages 21–30, New York, NY, USA, 2010. ACM. ISBN 978-1-4503-0133-6. doi: 10.1145/1920261.1920265. URL `http://doi.acm.org/10.1145/1920261.1920265`.

W. G. Cochran and D. B. Rubin. Controlling bias in observational studies: A review. *Sankhya: The Indian Journal of Statistics, Series A (1961-2002)*, 35(4):417–446, 1973. ISSN 0581572X. URL `http://www.jstor.org/stable/25049893`.

T. M. Cover and J. A. Thomas. *Elements of Information Theory*. Wiley-Interscience, second edition, 2006. ISBN 0-471-24195-9.

E. P. Davis, S. A. Stout, J. Molet, B. Vegetabile, L. M. Glynn, C. A. Sandman, K. Heins, H. Stern, and T. Z. Baram. Exposure to unpredictable maternal sensory signals influences cognitive development across species. *Proceedings of the National Academy of Sciences*, 114(39):10390–10395, 2017.

A. P. Dawid. Conditional independence in statistical theory. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 1–31, 1979.

A. P. Dawid, D. L. Faigman, and S. E. Fienberg. Fitting science into legal contexts: Assessing effects of causes or causes of effects? *Sociological Methods & Research*, 43(3):359–390, 2014.

A. P. Dawid, M. Musio, and S. E. Fienberg. From statistical evidence to evidence of causality. *Bayesian Analysis*, 11(3):725–752, 2016.

R. H. Dehejia and S. Wahba. Causal effects in nonexperimental studies: Reevaluating the evaluation of training programs. *Journal of the American Statistical Association*, 94(448): 1053–1062, 1999.

V. Dorie, J. Hill, U. Shalit, M. Scott, and D. Cervone. Automated versus do-it-yourself methods for causal inference: Lessons learned from a data analysis competition. *arXiv preprint arXiv:1707.02641*, 2017.

D. Duvenaud. *Automatic Model Construction with Gaussian Processes*. PhD thesis, University of Cambridge, 2014.

B. Efron and R. J. Tibshirani. *An Introduction to the Bootstrap*. CRC press, 1994.

R. A. Fisher. *Statistical Methods for Research Workers*. Oliver and Boyd, 1925.

M. J. Funk, D. Westreich, C. Wiesen, T. Strmer, M. A. Brookhart, and M. Davidian. Doubly robust estimation of causal effects. *American Journal of Epidemiology*, 173(7):761–767, 2011. doi: $10.1093/\mathrm{aje}/\mathrm{kwq439}$. URL `http://dx.doi.org/10.1093/aje/kwq439`.

Y. Gao, I. Kontoyiannis, and E. Bienenstock. Estimating the entropy of binary time series: Methodology, some theory and a simulation study. *Entropy*, 10(2):71–99, 2008.

A. Gelman, A. Vehtari, P. Jylänki, T. Sivula, D. Tran, S. Sahai, P. Blomstedt, J. P. Cunningham, D. Schiminovich, and C. Robert. Expectation propagation as a way of life: A framework for Bayesian inference on partitioned data. *ArXiv e-prints*, Dec. 2014.

M. Girolami and S. Rogers. Variational Bayesian multinomial probit regression with Gaussian process priors. *Neural Computation*, 18(8):1790–1817, 2006.

S. Greenland, J. Pearl, and J. M. Robins. Causal diagrams for epidemiologic research. *Epidemiology*, 10(1):37–48, 1999. ISSN 10443983. URL `http://www.jstor.org/stable/3702180`.

J. Hainmueller. Entropy balancing for causal effects: A multivariate reweighting method to produce balanced samples in observational studies. *Political Analysis*, 20(1):25–46, 2012.

J. Hensman, N. Fusi, and N. D. Lawrence. Gaussian processes for big data. *arXiv preprint arXiv:1309.6835*, 2013.

J. Hensman, A. Matthews, and Z. Ghahramani. Scalable variational gaussian process classification. In G. Lebanon and S. V. N. Vishwanathan, editors, *Proceedings of the Eighteenth International Conference on Artificial Intelligence and Statistics*, volume 38 of *Proceedings of Machine Learning Research*, pages 351–360, San Diego, California, USA, 09–12 May 2015. PMLR. URL http://proceedings.mlr.press/v38/hensman15.html.

D. Hernández-Lobato. *Prediction Based on Averages over Automatically Induced Learners: Ensemble Methods and Bayesian Techniques*. PhD thesis, Universidad Autónoma de Madrid, 2009.

D. Hernandez-Lobato and J. M. Hernandez-Lobato. Scalable gaussian process classification via expectation propagation. In A. Gretton and C. C. Robert, editors, *Proceedings of the 19th International Conference on Artificial Intelligence and Statistics*, volume 51 of *Proceedings of Machine Learning Research*, pages 168–176, Cadiz, Spain, 09–11 May 2016. PMLR. URL http://proceedings.mlr.press/v51/hernandez-lobato16.html.

A. B. Hill. The environment and disease: Association or causation? *Proceedings of the Royal Society of Medicine*, 58(5):295–300, 1965.

J. L. Hill. Bayesian nonparametric modeling for causal inference. *Journal of Computational and Graphical Statistics*, 20(1):217–240, 2011. doi: 10.1198/jcgs.2010.08162. URL http://dx.doi.org/10.1198/jcgs.2010.08162.

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.

T. Hofmann, B. Schölkopf, and A. J. Smola. Kernel methods in machine learning. *The Annals of Statistics*, pages 1171–1220, 2008.

P. W. Holland. Statistics and causal inference. *Journal of the American Statistical Association*, 81(396):945–960, 1986. ISSN 01621459. URL http://www.jstor.org/stable/2289064.

D. G. Horvitz and D. J. Thompson. A generalization of sampling without replacement from a finite universe. *Journal of the American Statistical Association*, 47(260):663–685, 1952. doi: 10.1080/01621459.1952.10483446. URL http://www.tandfonline.com/doi/abs/10.1080/01621459.1952.10483446.

K. Imai and M. Ratkovic. Covariate balancing propensity score. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 76(1):243–263, 2014. ISSN 1467-9868. doi: 10.1111/rssb.12027. URL http://dx.doi.org/10.1111/rssb.12027.

K. Imai and D. A. van Dyk. Causal inference with general treatment regimes. *Journal of the American Statistical Association*, 99(467):854–866, 2004. doi: 10.1198/016214504000001187. URL http://dx.doi.org/10.1198/016214504000001187.

G. Imbens. The role of the propensity score in estimating dose-response functions. *Biometrika*, 87(3):706–710, 2000. doi: 10.1093/biomet/87.3.706. URL http://dx.doi.org/10.1093/biomet/87.3.706.

G. W. Imbens and D. B. Rubin. *Causal Inference for Statistics, Social, and Biomedical Sciences: An Introduction*. Cambridge University Press, 2015. doi: 10.1017/CBO9781139025751.

S. Karlin and H. M. Taylor. *A First Course in Stochastic Processes*. Academic Press, San Diego, second edition, 1975.

S. Karlin and H. M. Taylor. *A Second Course in Stochastic Processes*. Academic Press, San Diego, 1981.

J. G. Kemeny and J. L. Snell. *Finite Markov Chains*. van Nostrand Publishing Company, Princeton, NJ, 1960.

I. Kontoyiannis, P. H. Algoet, Y. M. Suhov, and A. J. Wyner. Nonparametric entropy estimation for stationary processes and random fields, with applications to English text. *IEEE Transactions on Information Theory*, 44(3):1319–1327, May 1998. ISSN 0018-9448. doi: 10.1109/18.669425.

M. Kuss and C. E. Rasmussen. Assessing approximate inference for binary Gaussian process classification. *Journal of Machine Learning Research*, 6:1679–1704, Dec. 2005. ISSN 1532-4435. URL http://dl.acm.org/citation.cfm?id=1046920.1194901.

R. J. LaLonde. Evaluating the econometric evaluations of training programs with experimental data. *The American Economic Review*, pages 604–620, 1986.

B. K. Lee, J. Lessler, and E. A. Stuart. Improving propensity score weighting using machine learning. *Statistics in Medicine*, 29(3):337–346, Feb 2010. ISSN 1097-0258 (Electronic). doi: 10.1002/sim.3782.

A. Lempel and J. Ziv. On the complexity of finite sequences. *IEEE Transactions on Information Theory*, 22(1):75–81, Jan 1976. ISSN 0018-9448. doi: 10.1109/TIT.1976.1055501.

D. A. Levin, Y. Peres, and E. L. Wilmer. *Markov Chains and Mixing Times*. American Mathematical Society, 2009. ISBN 0-8218-4739-2.

F. Li, K. L. Morgan, and A. M. Zaslavsky. Balancing covariates via propensity score weighting. *Journal of the American Statistical Association*, pages 1–11, 2017.

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):403–425, Dec 2004. ISSN 1082-989X (Print). doi: 10.1037/1082-989X.9.4.403.

P. McCullagh and J. A. Nelder. *Generalized Linear Models*, volume 37 of *Monographs on Statistics and Applied Probability*. Chapman and Hall/CRC, Boca Raton, second edition, 1989.

J. McInerney, S. Stein, A. Rogers, and N. R. Jennings. Breaking the habit: Measuring and predicting departures from routine in individual human mobility. *Pervasive and Mobile Computing*, 9(6):808 – 822, 2013. ISSN 1574-1192. doi: http://dx.doi.org/10.1016/j.pmcj.2013.07.016. URL http://www.sciencedirect.com/science/article/pii/S1574119213000989.

T. P. Minka. Expectation propagation for approximate Bayesian inference. In *Proceedings of the Seventeenth Conference on Uncertainty in Artificial Intelligence*, pages 362–369. Morgan Kaufmann Publishers Inc., 2001.

J. Molet, K. Heins, X. Zhuo, Y. T. Mei, L. Regev, T. Z. Baram, and H. Stern. Fragmentation and high entropy of neonatal experience predict adolescent emotional outcome. *Translational Psychiatry*, 6:e702, January 2016. URL http://dx.doi.org/10.1038/tp.2015.200http://10.1038/tp.2015.200.

K. P. Murphy. *Machine Learning: A Probabilistic Perspective*. The MIT Press, 2012. ISBN 0262018020, 9780262018029.

J. Nocedal and S. J. Wright. *Numerical Optimization*. Springer, 2006. ISBN 9780387510880.

D. S. Ornstein and B. Weiss. Entropy and data compression schemes. *IEEE Transactions on Information Theory*, 39(1):78–83, Jan 1993. ISSN 0018-9448. doi: 10.1109/18.179344.

D. N. Politis and J. P. Romano. The stationary bootstrap. *Journal of the American Statistical Association*, 89(428):1303–1313, 1994. doi: 10.1080/01621459.1994.10476870. URL http://dx.doi.org/10.1080/01621459.1994.10476870.

A. Porta, S. Guzzetti, N. Montano, R. Furlan, M. Pagani, A. Malliani, and S. Cerutti. Entropy, entropy rate, and pattern classification as tools to typify complexity in short heart period variability series. *IEEE Transactions on Biomedical Engineering*, 48(11):1282–1291, Nov 2001. ISSN 0018-9294. doi: 10.1109/10.959324.

M. J. Powell. The bobyqa algorithm for bound constrained optimization without derivatives. *Cambridge NA Report NA2009/06, University of Cambridge, Cambridge*, 2009.

J. Quiñonero-Candela and C. E. Rasmussen. A unifying view of sparse approximate Gaussian process regression. *Journal of Machine Learning Research*, 6(Dec):1939–1959, 2005.

C. E. Rasmussen and C. K. Williams. *Gaussian Processes for Machine Learning*. The MIT Press, Cambridge, Massachusetts, 2006.

J. Riihimäki, P. Jylänki, and A. Vehtari. Nested expectation propagation for Gaussian process classification with a multinomial probit likelihood. *Journal of Machine Learning Research*, 14(Jan):75–109, 2013.

J. M. Robins, A. Rotnitzky, and L. P. Zhao. Estimation of regression coefficients when some regressors are not always observed. *Journal of the American Statistical Association*, 89 (427):846–866, 1994. ISSN 01621459. URL http://www.jstor.org/stable/2290910.

P. R. Rosenbaum. Interference between units in randomized experiments. *Journal of the American Statistical Association*, 102(477):191–200, 2007. doi: 10.1198/016214506000001112. URL https://doi.org/10.1198/016214506000001112.

P. R. Rosenbaum. *Design of Observational Studies*. Springer Series in Statistics. Springer Verlag, New York NY, 2010.

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. doi: 10.1093/biomet/70.1.41. URL http://biomet.oxfordjournals.org/content/70/1/41.abstract.

D. B. Rubin. Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology*, 66(5):688, 1974.

D. B. Rubin. Bayesian inference for causal effects: The role of randomization. *The Annals of Statistics*, 6(1):34–58, 1978. ISSN 00905364. URL http://www.jstor.org/stable/2958688.

B. Schölkopf and A. J. Smola. *Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond*. MIT press, 2001.

M. S. Schuler and S. Rose. Targeted maximum likelihood estimation for causal inference in observational studies. *American Journal of Epidemiology*, 185(1):65–73, 2017. doi: 10.1093/aje/kww165. URL http://dx.doi.org/10.1093/aje/kww165.

C. Shannon. A mathematical theory of communication. *The Bell System Technical Journal*, 27(3):379–423, July 1948. ISSN 0005-8580. doi: 10.1002/j.1538-7305.1948.tb01338.x.

C. Song, Z. Qu, N. Blumm, and A.-L. Barabási. Limits of predictability in human mobility. *Science*, 327(5968):1018–1021, 2010. ISSN 0036-8075. doi: 10.1126/science.1177170. URL http://science.sciencemag.org/content/327/5968/1018.

J. Splawa-Neyman, D. M. Dabrowska, and T. P. Speed. On the application of probability theory to agricultural experiments. Essay on Principles. Section 9. *Statistical Science*, 5 (4):465–472, 1990. ISSN 08834237. URL http://www.jstor.org/stable/2245382.

E. A. Stuart. Matching methods for causal inference: A review and a look forward. *Statistical Science*, 25(1):1, 2010.

V. Tolvanen, P. Jylänki, and A. Vehtari. Expectation propagation for nonstationary heteroscedastic Gaussian process regression. In *2014 IEEE International Workshop on Machine Learning for Signal Processing (MLSP)*, pages 1–6. IEEE, 2014.

M. A. Van Gerven, B. Cseke, F. P. De Lange, and T. Heskes. Efficient Bayesian multivariate fMRI analysis using a sparsifying spatio-temporal prior. *NeuroImage*, 50(1):150–161, 2010.

M.-J. Woo, J. P. Reiter, and A. F. Karr. Estimation of propensity scores using generalized additive models. *Statistics in Medicine*, 27(19):3805–3816, aug 2008. ISSN 0277-6715 (Print). doi: 10.1002/sim.3278.

A. D. Wyner and J. Ziv. Some asymptotic properties of the entropy of a stationary ergodic data source with applications to data compression. *IEEE Transactions on Information Theory*, 35(6):1250–1258, Nov 1989. ISSN 0018-9448. doi: 10.1109/18.45281.

A. D. Wyner and J. Ziv. The sliding-window Lempel-Ziv algorithm is asymptotically optimal. *Proceedings of the IEEE*, 82(6):872–877, Jun 1994. ISSN 0018-9219. doi: 10.1109/5.286191.

J. Ziv and A. Lempel. A universal algorithm for sequential data compression. *IEEE Transactions on Information Theory*, 23(3):337–343, May 1977. ISSN 0018-9448. doi: 10.1109/TIT.1977.1055714.

J. Ziv and A. Lempel. Compression of individual sequences via variable-rate coding. *IEEE Transactions on Information Theory*, 24(5):530–536, Sep 1978. ISSN 0018-9448. doi: 10.1109/TIT.1978.1055934.

J. R. Zubizarreta. Stable weights that balance covariates for estimation with incomplete outcome data. *Journal of the American Statistical Association*, 110(511):910–922, 2015. doi: 10.1080/01621459.2015.1023805. URL `https://doi.org/10.1080/01621459.2015.1023805`.

Appendix A

Derivations for Probit Regression with Gaussian Processes

A.1 Laplace Approximation

Consider a binary random variable $y \in \{-1, 1\}$ and a d-dimensional set of covariates $x \in \Re^d$. For a sample, $i = 1, \ldots, n$, let the probability of y_i given x_i be related through a latent function f, that is,

$$p(y_i|x_i) = \Phi(y_i f(x_i)) \equiv \Phi(y_i f_i) \tag{A.1}$$

Additionally, assume that $\mathbf{f}$, the collection of f_i, were generated from a Gaussian process,

$$\mathbf{f}|\mathbf{X} \sim \mathcal{GP}(\mathbf{m}, \mathbf{K}) \tag{A.2}$$

where $\mathbf{K}$ is the covariance function constructed from all observations x and x' using a kernel function $k(\cdot, \cdot)$. Under such a model and assuming that $\mathbf{K}$ has been modeled with known

hyperparameters, the posterior distribution is as follows,

$$p(\mathbf{f}|\mathbf{X},\mathbf{y}) \;=\; \frac{\left[\prod_{i=1}^{N}\Phi(y_i f_i)\right]\det(2\pi\mathbf{K})^{-1/2}\exp\left(-\frac{1}{2}(\mathbf{f}-\mathbf{m})^T\mathbf{K}^{-1}(\mathbf{f}-\mathbf{m})\right)}{p(\mathbf{y}|\mathbf{X})} \tag{A.3}$$

and is intractable. The Laplace approximation approximates $p(\mathbf{f}|\mathbf{X},\mathbf{y})$ by a normal distribution centered at the mode of $p(\mathbf{f}|\mathbf{X},\mathbf{y})$,

$$\hat{\mathbf{f}} = \arg\max_{\mathrm{f}} \; \log\left(p(\mathbf{f}|\mathbf{X},\mathbf{y})\right), \tag{A.4}$$

with variance equal to the inverse of the negative Hessian of the log posterior distribution evaluated at the mode,

$$\mathbf{A} = \left[-\frac{\partial^2}{\partial\mathbf{f}\partial\mathbf{f}^T}\log p(\mathbf{f}|\mathbf{X},\mathbf{y})\;|_{\mathrm{f}=\hat{\mathrm{f}}}\right]^{-1}. \tag{A.5}$$

Thus, the approximate posterior distribution is

$$q(\mathbf{f}|\mathbf{X},\mathbf{y}) \sim \mathcal{N}(\hat{\mathbf{f}},\mathbf{A}) \tag{A.6}$$

This section elaborates on the steps required for the Laplace approximation.

A.1.1 Derivatives of the log Posterior Distribution

We must find the gradient and the Hessian of $\log p(\mathbf{f}|\mathbf{X},\mathbf{y})$. First,

$$\log p(\mathbf{f}|\mathbf{X},\mathbf{y}) = \sum_{i=1}^{n}\log\Phi(y_i f_i) - \frac{1}{2}(\mathbf{f}-\mathbf{m})^T\mathbf{K}^{-1}(\mathbf{f}-\mathbf{m}) + C \tag{A.7}$$

where C is a constant with respect to $\mathbf{f}$. Thus,

$$\begin{aligned}\frac{\partial}{\partial \mathbf{f}} \log p(\mathbf{f}|\mathbf{X}, \mathbf{y}) &= \frac{\partial}{\partial \mathbf{f}} \sum_{i=1}^{n} \log \Phi(y_i f_i) - \frac{\partial}{\partial \mathbf{f}} \frac{1}{2} (\mathbf{f} - \mathbf{m})^T \mathbf{K}^{-1} (\mathbf{f} - \mathbf{m}) \\ &= \frac{\partial}{\partial \mathbf{f}} \sum_{i=1}^{n} \log \Phi(y_i f_i) - \mathbf{K}^{-1} (\mathbf{f} - \mathbf{m}).\end{aligned}$$

Now,

$$\begin{aligned}\frac{\partial}{\partial f_j} \sum_{i=1}^{n} \log \Phi(y_i f_i) &= \frac{1}{\Phi(y_j f_j)} \phi(y_j f_j) y_j &\text{(A.8)} \\ &= \frac{y_j \phi(f_j)}{\Phi(y_j f_j)}, &\text{(A.9)}\end{aligned}$$

where $\phi(z)$ is defined to be the standard normal pdf evaluated at z. This implies,

$$\frac{\partial}{\partial \mathbf{f}} \sum_{i=1}^{n} \log \Phi(y_i f_i) = \begin{pmatrix} \frac{y_1 \phi(f_1)}{\Phi(y_1 f_1)} \\ \frac{y_2 \phi(f_2)}{\Phi(y_2 f_2)} \\ \vdots \\ \frac{y_1 \phi(f_n)}{\Phi(y_n f_n)} \end{pmatrix} \equiv \mathbf{b}, \tag{A.10}$$

and finally,

$$\frac{\partial}{\partial \mathbf{f}} \log p(\mathbf{f}|\mathbf{X}, \mathbf{y}) = \mathbf{b} - \mathbf{K}^{-1} (\mathbf{f} - \mathbf{m}).$$

For the Hessian, we take second derivatives,

$$\frac{\partial^2}{\partial \mathbf{f} \partial \mathbf{f}^T} \log p(\mathbf{f}|\mathbf{X}, \mathbf{y}) = \frac{\partial}{\partial \mathbf{f}} \left[\mathbf{b} - \mathbf{K}^{-1}(\mathbf{f} - \mathbf{m})\right]^T \tag{A.11}$$

$$= -\mathbf{K}^{-1} + \frac{\partial}{\partial \mathbf{f}} \mathbf{b}^T \tag{A.12}$$

where,

$$\frac{\partial}{\partial \mathbf{f}} \mathbf{b}^T = \begin{pmatrix} \frac{\partial}{\partial f_1} \frac{y_1 \phi(f_1)}{\Phi(y_1 f_1)} & \frac{\partial}{\partial f_1} \frac{y_2 \phi(f_2)}{\Phi(y_2 f_2)} & \cdots & \frac{\partial}{\partial f_1} \frac{y_1 \phi(f_n)}{\Phi(y_n f_n)} \\ \frac{\partial}{\partial f_2} \frac{y_1 \phi(f_1)}{\Phi(y_1 f_1)} & \frac{\partial}{\partial f_2} \frac{y_2 \phi(f_2)}{\Phi(y_2 f_2)} & \cdots & \frac{\partial}{\partial f_2} \frac{y_1 \phi(f_n)}{\Phi(y_n f_n)} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial}{\partial f_n} \frac{y_1 \phi(f_1)}{\Phi(y_1 f_1)} & \frac{\partial}{\partial f_n} \frac{y_2 \phi(f_2)}{\Phi(y_2 f_2)} & \cdots & \frac{\partial}{\partial f_n} \frac{y_1 \phi(f_n)}{\Phi(y_n f_n)} \end{pmatrix}. \tag{A.13}$$

Each diagonal element of the matrix $\frac{\partial}{\partial \mathbf{f}} \mathbf{b}^T$ is,

$$\frac{\partial}{\partial f_i} \frac{y_i \phi(f_i)}{\Phi(y_i f_i)} = y_i \left[\frac{1}{\Phi(y_i f_i)} \frac{\partial}{\partial f_i} \phi(f_i) + \phi(f_i) \frac{\partial}{\partial f_i} \frac{1}{\Phi(y_i f_i)} \right] \tag{A.14}$$

where

$$\frac{\partial}{\partial f_i} \phi(f_i) = \frac{\partial}{\partial f_i} \sqrt{\frac{1}{2\pi}} \exp\left\{-\frac{f_i^2}{2}\right\} = -\phi(f_i) f_i \tag{A.15}$$

and

$$\frac{\partial}{\partial f_i} \frac{1}{\Phi(y_i f_i)} = -\left(\frac{1}{\Phi(y_i f_i)}\right)^2 \frac{\partial}{\partial f_i} \Phi(y_i f_i) = -\frac{y_i \phi(f_i)}{\Phi(y_i f_i)^2}, \tag{A.16}$$

implying that

$$\frac{\partial}{\partial f_i}\frac{y_i\phi(f_i)}{\Phi(y_i f_i)} = -y_i\left[\frac{f_i\phi(f_i)}{\Phi(y_i f_i)} + \frac{y_i\phi(f_i)^2}{\Phi(y_i f_i)^2}\right] \tag{A.17}$$

$$= -\left[y_i f_i \frac{\phi(f_i)}{\Phi(y_i f_i)} + \left(\frac{\phi(f_i)}{\Phi(y_i f_i)}\right)^2\right]. \tag{A.18}$$

with respect to the off-diagonal terms of $\frac{\partial}{\partial \mathbf{f}}\mathbf{b}^T$, it is clear that

$$\frac{\partial}{\partial f_j}\frac{y_i\phi(f_i)}{\Phi(y_i f_i)} = 0. \tag{A.19}$$

If we define $\mathbf{B} \equiv -\frac{\partial}{\partial \mathbf{f}}\mathbf{b}^T$, then the Hessian matrix is,

$$\frac{\partial^2}{\partial \mathbf{f}\partial \mathbf{f}^T}\log p(\mathbf{f}|\mathbf{X},\mathbf{y}) = -\mathbf{K}^{-1} - \mathbf{B} \tag{A.20}$$

A.1.2 Newton's Method for Finding the Posterior Mode

To find the posterior mode we use Newton's method, where we initialize $\mathbf{f}_{old}$ and then iterate the following until convergence,

$$\mathbf{f}_{new} = \mathbf{f}_{old} - \left[\frac{\partial^2}{\partial \mathbf{f}\partial \mathbf{f}^T}\log p(\mathbf{f}|\mathbf{X},\mathbf{y}) \mid_{\mathbf{f}=\mathbf{f}_{old}}\right]^{-1}\left[\frac{\partial}{\partial \mathbf{f}}\log p(\mathbf{f}|\mathbf{X},\mathbf{y}) \mid_{\mathbf{f}=\mathbf{f}_{old}}\right] \tag{A.21}$$

From the calculations of Section A.1.1 we can write the iteration as

$$\mathbf{f}_{new} = \mathbf{f}_{old} + (\mathbf{K}^{-1} + \mathbf{B})^{-1}(\mathbf{b} - \mathbf{K}^{-1}(\mathbf{f}_{\text{old}} - \mathbf{m})) \tag{A.22}$$

These updates will be computationally difficult to perform since there are three inverses that must be performed. We can rearrange the above as follows,

$$\begin{aligned}
\mathbf{f}_{new} &= \mathbf{f}_{old} + (\mathbf{K}^{-1} + \mathbf{B})^{-1}(\mathbf{b} - \mathbf{K}^{-1}(\mathbf{f}_{\text{old}} - \mathbf{m})) && \text{(A.23)} \\
&= (\mathbf{K}^{-1} + \mathbf{B})^{-1}(\mathbf{b} + \mathbf{K}^{-1}\mathbf{m}) + \left(\mathbf{I} - (\mathbf{K}^{-1} + \mathbf{B})^{-1}\mathbf{K}^{-1}\right)\mathbf{f}_{old} && \text{(A.24)} \\
&= (\mathbf{K}^{-1} + \mathbf{B})^{-1}(\mathbf{b} + \mathbf{K}^{-1}\mathbf{m}) + (\mathbf{K}^{-1} + \mathbf{B})^{-1}\left(\mathbf{K}^{-1} + \mathbf{B} - \mathbf{K}^{-1}\right)\mathbf{f}_{old} && \text{(A.25)} \\
&= (\mathbf{K}^{-1} + \mathbf{B})^{-1}(\mathbf{b} + \mathbf{K}^{-1}\mathbf{m}) + (\mathbf{K}^{-1} + \mathbf{B})^{-1}\mathbf{B}\mathbf{f}_{old} && \text{(A.26)} \\
&= (\mathbf{K}^{-1} + \mathbf{B})^{-1}(\mathbf{B}\mathbf{f}_{old} + \mathbf{b} + \mathbf{K}^{-1}\mathbf{m}). && \text{(A.27)}
\end{aligned}$$

Finally, we can use the matrix inversion lemma (Rasmussen and Williams, 2006, see Appendix A.3) to further reduce the computational burden,

$$\begin{aligned}
\mathbf{f}_{new} &= (\mathbf{K}^{-1} + \mathbf{B})^{-1}(\mathbf{B}\mathbf{f}_{old} + \mathbf{b} + \mathbf{K}^{-1}\mathbf{m}) && \text{(A.28)} \\
&= (\mathbf{K} - \mathbf{K}(\mathbf{B}^{-1} + \mathbf{K})^{-1}\mathbf{K})(\mathbf{B}\mathbf{f}_{old} + \mathbf{b} + \mathbf{K}^{-1}\mathbf{m}) && \text{(A.29)}
\end{aligned}$$

and lastly following the advice from Rasmussen and Williams (2006) to stabilize the inversion of $(\mathbf{B}^{-1} + \mathbf{K})$, we can factor out a $\mathbf{B}^{-1/2}$ from each term in $(\mathbf{B}^{-1} + \mathbf{K})$, such that,

$$\begin{aligned}
\mathbf{f}_{new} &= (\mathbf{K} - \mathbf{K}(\mathbf{B}^{-1} + \mathbf{K})^{-1}\mathbf{K})(\mathbf{B}\mathbf{f}_{old} + \mathbf{b} + \mathbf{K}^{-1}\mathbf{m}) && \text{(A.30)} \\
&= (\mathbf{K} - \mathbf{K}(\mathbf{B}^{-1/2}(\mathbf{I} + \mathbf{B}^{1/2}\mathbf{K}\mathbf{B}^{1/2})\mathbf{B}^{-1/2})^{-1}\mathbf{K})(\mathbf{B}\mathbf{f}_{old} + \mathbf{b} + \mathbf{K}^{-1}\mathbf{m}) && \text{(A.31)} \\
&= \mathbf{K}(\mathbf{I} - \mathbf{B}^{1/2}(\mathbf{I} + \mathbf{B}^{1/2}\mathbf{K}\mathbf{B}^{1/2})^{-1}\mathbf{B}^{1/2}\mathbf{K})(\mathbf{B}\mathbf{f}_{old} + \mathbf{b} + \mathbf{K}^{-1}\mathbf{m}) && \text{(A.32)}
\end{aligned}$$

The matrix $(\mathbf{I} + \mathbf{B}^{1/2}\mathbf{K}\mathbf{B}^{1/2})$ is better conditioned for inversion.

A.1.3 Approximate Marginal Likelihood

An approximation to the marginal likelihood is necessary in most data sets in order to select hyperparameters. The marginal likelihood is,

$$
\begin{aligned}
p(\mathbf{y}|\mathbf{X}) &= \int p(\mathbf{y}|\mathbf{f})p(\mathbf{f}|\mathbf{X})d\mathbf{f} && \text{(A.33)}\\
&= \int \exp\left(\log p(\mathbf{y}|\mathbf{f}) + \log p(\mathbf{f}|\mathbf{X})\right)d\mathbf{f} && \text{(A.34)}\\
&= \int \exp\left(g(\mathbf{f})\right)d\mathbf{f} && \text{(A.35)}
\end{aligned}
$$

To approximate $p(\mathbf{y}|\mathbf{X})$ we take a second-order Taylor series expansion of $g(\mathbf{f})$ about the mode of the posterior distribution $p(\mathbf{f}|\mathbf{X},\mathbf{y})$, previously defined as $\hat{\mathbf{f}}$,

$$
\begin{aligned}
g(\mathbf{f}) &\approx g(\hat{\mathbf{f}}) + \frac{\partial}{\partial \mathbf{f}} g(\hat{\mathbf{f}})^T(\mathbf{f}-\hat{\mathbf{f}}) + \frac{1}{2}(\mathbf{f}-\hat{\mathbf{f}})^T\left[\frac{\partial^2}{\partial \mathbf{f}\partial \mathbf{f}^T} g(\hat{\mathbf{f}})\right](\mathbf{f}-\hat{\mathbf{f}}) && \text{(A.36)}\\
&= g(\hat{\mathbf{f}}) + \frac{1}{2}(\mathbf{f}-\hat{\mathbf{f}})^T\left[\frac{\partial^2}{\partial \mathbf{f}\partial \mathbf{f}^T} g(\hat{\mathbf{f}})\right](\mathbf{f}-\hat{\mathbf{f}}) && \text{(A.37)}\\
&= g(\hat{\mathbf{f}}) - \frac{1}{2}(\mathbf{f}-\hat{\mathbf{f}})^T\mathbf{G}(\mathbf{f}-\hat{\mathbf{f}}) && \text{(A.38)}
\end{aligned}
$$

where the second line follows from the fact that $\frac{\partial}{\partial \mathbf{f}} g(\hat{\mathbf{f}})^T$ is zero at the mode and where we have defined $\mathbf{G} = -\frac{\partial^2}{\partial \mathbf{f}\partial \mathbf{f}^T} g(\hat{\mathbf{f}})$. Therefore,

$$
\begin{aligned}
p(\mathbf{y}|\mathbf{X}) &= \int p(\mathbf{y}|\mathbf{f})p(\mathbf{f}|\mathbf{X})d\mathbf{f} && \text{(A.39)}\\
&\approx \int \exp\left(g(\hat{\mathbf{f}}) - \frac{1}{2}(\mathbf{f}-\hat{\mathbf{f}})^T\mathbf{G}(\mathbf{f}-\hat{\mathbf{f}})\right)d\mathbf{f} && \text{(A.40)}\\
&= \exp\left(g(\hat{\mathbf{f}})\right)\int \exp\left(-\frac{1}{2}(\mathbf{f}-\hat{\mathbf{f}})^T\mathbf{G}(\mathbf{f}-\hat{\mathbf{f}})\right)d\mathbf{f} && \text{(A.41)}
\end{aligned}
$$

$$\approx \quad \exp\left(g(\hat{\mathbf{f}})\right) \det\left(2\pi\mathbf{G}^{-1}\right)^{1/2}, \tag{A.42}$$

and therefore,

$$\begin{aligned}
\log p(\mathbf{y}|\mathbf{X}) &\approx g(\hat{\mathbf{f}}) + \frac{1}{2}\log\det\left(2\pi\mathbf{G}^{-1}\right) && \text{(A.43)}\\
&= \left(\sum_{i=1}^{n}\log\Phi(y_i\hat{f}_i)\right) + \frac{1}{2}\log\det\left(2\pi\mathbf{G}^{-1}\right) && \text{(A.44)}\\
&\quad -\frac{1}{2}\log\det\left(2\pi\mathbf{K}\right) - \frac{1}{2}(\hat{\mathbf{f}}-\mathbf{m})^T\mathbf{K}^{-1}(\hat{\mathbf{f}}-\mathbf{m}) && \text{(A.45)}\\
&= \left(\sum_{i=1}^{n}\log\Phi(y_i\hat{f}_i)\right) - \frac{1}{2}(\hat{\mathbf{f}}-\mathbf{m})^T\mathbf{K}^{-1}(\hat{\mathbf{f}}-\mathbf{m}) && \text{(A.46)}\\
&\quad -\frac{1}{2}\log\det\left(\mathbf{KG}\right) && \text{(A.47)}\\
&= \left(\sum_{i=1}^{n}\log\Phi(y_i\hat{f}_i)\right) - \frac{1}{2}(\hat{\mathbf{f}}-\mathbf{m})^T\mathbf{K}^{-1}(\hat{\mathbf{f}}-\mathbf{m}) && \text{(A.48)}\\
&\quad -\frac{1}{2}\log\det\left(\mathbf{K}(\mathbf{B}+\mathbf{K}^{-1})\right) && \text{(A.49)}\\
&= \left(\sum_{i=1}^{n}\log\Phi(y_i\hat{f}_i)\right) - \frac{1}{2}(\hat{\mathbf{f}}-\mathbf{m})^T\mathbf{K}^{-1}(\hat{\mathbf{f}}-\mathbf{m}) && \text{(A.50)}\\
&\quad -\frac{1}{2}\log\det\left(\mathbf{KB}+\mathbf{I}_n\right)) && \text{(A.51)}
\end{aligned}$$

These results are consistent with Rasmussen and Williams (2006), Section 3.4.

A.2 Expectation propagation

Consider the binary setting of the previous section. For probit regression with Gaussian processes the posterior distribution of the latent values is,

$$p(\mathbf{f}|\mathbf{X},\theta,\mathbf{y}) \propto p(\mathbf{f}|\theta)\times\prod_{i=1}^{N}p(y_i|f_i) \tag{A.52}$$

$$= \det(2\pi\mathbf{K})^{-1/2}\exp\left(-\frac{1}{2}\mathbf{f}^T\mathbf{K}^{-1}\mathbf{f}\right)\left[\prod_{i=1}^{N}\Phi(y_i f_i)\right]. \tag{A.53}$$

Under this model, an efficient expectation propagation algorithm, as laid out in Rasmussen and Williams (2006), chooses site approximations which are unnormalized normal distributions,

$$q(f_i|Z_i,m_i,s_i^2) = \frac{Z_i}{\sqrt{2\pi s_i^2}}\exp\left(-\frac{1}{2}\frac{(f_i-m_i)^2}{s_i^2}\right), \tag{A.54}$$

so that the approximate posterior is,

$$\begin{aligned} q(\mathbf{f}|\mathbf{X},\theta,\mathbf{y}) &\propto p(\mathbf{f}|\theta)\times\prod_{i=1}^{N}q(f_i|Z_i,m_i,s_i^2) \\ &= \det(2\pi\mathbf{K})^{-1/2}\exp\left(-\frac{1}{2}\mathbf{f}^T\mathbf{K}^{-1}\mathbf{f}\right)\left[\prod_{i=1}^{N}\frac{Z_i}{\sqrt{2\pi s_i^2}}\exp\left(-\frac{1}{2}\frac{(f_i-m_i)^2}{s_i^2}\right)\right] \\ &\propto \exp\left(-\frac{1}{2}\mathbf{f}^T\mathbf{K}^{-1}\mathbf{f}\right)\left[\prod_{i=1}^{N}\exp\left(-\frac{1}{2}\frac{(f_i-m_i)^2}{s_i^2}\right)\right] \\ &= \exp\left(-\frac{1}{2}\mathbf{f}^T\mathbf{K}^{-1}\mathbf{f}\right)\exp\left(-\frac{1}{2}(\mathbf{f}-\mathbf{m})^T\mathbf{\Sigma}^{-1}(\mathbf{f}-\mathbf{m})\right) \end{aligned}$$

where $\mathbf{m}$ is the vector of m_i and $\mathbf{\Sigma} = diag(s_i^2)$. Using the results for normal-normal conjugate models (Rasmussen and Williams, 2006, see Appendix A.2, Equation A.7), this implies that

the approximating posterior distribution is,

$$\mathbf{f}|\mathbf{X},\theta,\mathbf{y} \sim \mathcal{N}\left(\mu_p \equiv (\mathbf{K}^{-1}+\boldsymbol{\Sigma}^{-1})^{-1}\boldsymbol{\Sigma}^{-1}\mathbf{m}, \boldsymbol{\Sigma}_{\mathrm{p}} \equiv (\mathbf{K}^{-1}+\boldsymbol{\Sigma}^{-1})^{-1}\right) \tag{A.55}$$

A.2.1 Marginal Cavity Distribution

Expectation propagation for Gaussian processes in binary regression focuses not on the full cavity distribution, but the *marginal* cavity distribution (Rasmussen and Williams, 2006; Gelman et al., 2014). The marginal cavity distribution is,

$$q_{-i}(f_i) \propto \int \cdots \int_{f_j \neq f_i} p(\mathbf{f}|\mathbf{X},\theta) \prod_{j \neq i} q(f_j|Z_j, m_j, s_j^2) df_{j \neq i}$$

One way to find the marginal cavity distribution is to perform this integration directly. Alternatively, and more conveniently, it is possible to calculate the marginal cavity by first finding the marginal posterior distribution of f_i and then dividing the component of $q(f_i|Z_i, m_i, s_i^2)$ from the marginal posterior distribution of f_i. Conceptually, this is the marginal posterior distribution of f_i *absent* the contributions from $q(f_i|Z_i, m_i, s_i^2)$. From, properties of normal distributions, the marginal posterior distribution for f_i is $\mathcal{N}(\mu_i^p, \sigma_{ii}^p)$ and therefore we can form the marginal cavity as follows,

$$\begin{aligned}
q_{-i}(f_i) &\propto \frac{\exp\left(-\frac{1}{2}\frac{(f_i-\mu_i^p)^2}{\sigma_{ii}^p}\right)}{\exp\left(-\frac{1}{2}\frac{(f_i-m_i)^2}{s_i^2}\right)} && \text{(A.56)}\\
&= \exp\left(-\frac{1}{2}\left[\frac{f_i^2-2\mu_i^p f_i+(\mu_i^p)^2}{\sigma_{ii}^p}-\frac{f_i^2-2m_i f_i+m_i^2}{s_i^2}\right]\right) && \text{(A.57)}\\
&\propto \exp\left(-\frac{1}{2}\left[\left(\frac{1}{\sigma_{ii}^p}-\frac{1}{s_i^2}\right)f_i^2-2\left(\frac{\mu_i^p}{\sigma_{ii}^p}-\frac{m_i}{s_i^2}\right)f_i\right]\right), && \text{(A.58)}
\end{aligned}$$

which by completing the square implies that,

$$q_{-i}(f_i) \sim \mathcal{N}(\mu_{-i}, \sigma^2_{-i}) \tag{A.59}$$

where,

$$\sigma^2_{-i} \equiv \left(\frac{1}{\sigma^p_{ii}} - \frac{1}{s_i^2}\right)^{-1} \quad \text{and} \quad \mu_{-i} \equiv \sigma^2_{-i}\left(\frac{\mu_i^p}{\sigma^p_{ii}} - \frac{m_i}{s_i^2}\right). \tag{A.60}$$

A.2.2 Marginal Tilted Distribution

The marginal tilted distribution is found by combining the true likelihood component $\Phi(y_i f_i)$ with the marginal cavity distribution. This is simply,

$$q_{\backslash i}(f_i) \propto p(y_i|f_i) q_{-i}(f_i). \tag{A.61}$$

A.2.3 Updating the Site Parameters

Using the marginal tilted distribution, we update the approximate (unnormalized) posterior distribution by minimizing the KL-divergence between the approximate marginal posterior distribution and the marginal tilted distribution. It is well known that when when the approximate distribution is from an exponential family that this amounts to matching moments between the two distributions. The unnormalized marginal posterior distribution will have the form,

$$\hat{q}(f_i) \equiv \frac{\hat{Z}_i}{\sqrt{2\pi\hat{\sigma}_i^2}} \exp\left(-\frac{1}{2}\frac{(f_i - \hat{\mu}_i)}{\hat{\sigma}_i^2}\right) \tag{A.62}$$

where $\hat{Z}_i, \hat{\mu}_i, \hat{\sigma}_i^2$ will be set by matching moments to $q_{\setminus i}(f_i) \propto p(y_i|f_i)q_{-i}(f_i)$. From results in Rasmussen and Williams (2006) Sections 3.6 & 3.9, it follows that,

$$\begin{aligned}
\hat{Z}_i &= \Phi(z_i) \qquad \text{where } z_i = \frac{y_i\mu_{-i}}{\sqrt{1+\sigma^2_{-i}}} \\
\hat{\mu}_i &= \mu_{-i} + y_i \frac{\sigma^2_{-i}}{\sqrt{1+\sigma^2_{-i}}} \frac{\phi(z_i)}{\Phi(z_i)} \\
\hat{\sigma}_i^2 &= \sigma^2_{-i} - \frac{\phi(z_i)}{\Phi(z_i)}\left(z_i + \frac{\phi(z_i)}{\Phi(z_i)}\right)\left(\frac{\sigma^2_{-i}}{\sqrt{1+\sigma^2_{-i}}}\right)^2.
\end{aligned}$$

We then need to update the parameters Z_i, m_i, s_i^2 from these results. Recall that we have that,

$$Z_i \frac{1}{\sqrt{2\pi s_i^2}} \exp\left(-\frac{1}{2}\frac{(f_i - m_i)^2}{s_i^2}\right) \equiv \frac{\hat{q}(f_i)}{q_{-i}(f_i)} \tag{A.63}$$

$$= \frac{\hat{Z}_i \frac{1}{\sqrt{2\pi\hat{\sigma}_i^2}} \exp\left(-\frac{1}{2}\frac{(f_i-\hat{\mu}_i)^2}{\hat{\sigma}_i^2}\right)}{\frac{1}{\sqrt{2\pi\sigma^2_{-i}}} \exp\left(-\frac{1}{2}\frac{(f_i-\mu_{-i})^2}{\sigma^2_{-i}}\right)}. \tag{A.64}$$

We will simplify the right hand side,

$$= \frac{\hat{Z}_i \frac{1}{\sqrt{2\pi\hat{\sigma}_i^2}} \exp\left(-\frac{1}{2}\frac{(f_i-\hat{\mu}_i)^2}{\hat{\sigma}_i^2}\right)}{\frac{1}{\sqrt{2\pi\sigma^2_{-i}}} \exp\left(-\frac{1}{2}\frac{(f_i-\mu_{-i})^2}{\sigma^2_{-i}}\right)} \tag{A.65}$$

$$= \hat{Z}_i \frac{\sigma_{-i}}{\hat{\sigma}_i} \exp\left[-\frac{1}{2}\left(\frac{(f_i-\hat{\mu}_i)^2}{\hat{\sigma}_i^2} - \frac{(f_i-\mu_{-i})^2}{\sigma^2_{-i}}\right)\right] \tag{A.66}$$

$$= \hat{Z}_i \frac{\sigma_{-i}}{\hat{\sigma}_i} \exp\left[-\frac{1}{2}\left(f_i^2\left[\frac{1}{\hat{\sigma}_i^2} - \frac{1}{\sigma^2_{-i}}\right] - 2f_i\left[\frac{\hat{\mu}_i}{\hat{\sigma}_i^2} - \frac{\mu_{-i}}{\sigma^2_{-i}}\right] + \left[\frac{\hat{\mu}_i^2}{\hat{\sigma}_i^2} - \frac{\mu^2_{-i}}{\sigma^2_{-i}}\right]\right)\right]. \tag{A.67}$$

Defining,

$$a \equiv \left[\frac{1}{\hat{\sigma}_i^2} - \frac{1}{\sigma_{-i}^2}\right] \qquad b \equiv \left[\frac{\hat{\mu}_i}{\hat{\sigma}_i^2} - \frac{\mu_{-i}}{\sigma_{-i}^2}\right] \qquad c \equiv \left[\frac{\hat{\mu}_i^2}{\hat{\sigma}_i^2} - \frac{\mu_{-i}^2}{\sigma_{-i}^2}\right] \tag{A.68}$$

and continuing from A.67,

$$= \hat{Z}_i \frac{\sigma_{-i}}{\hat{\sigma}_i} \exp\left[-\frac{1}{2}a\left(f_i^2 - 2f_i\frac{b}{a} + \frac{c}{a}\right)\right] \tag{A.69}$$

$$= \hat{Z}_i \frac{\sigma_{-i}}{\hat{\sigma}_i} \exp\left[\frac{1}{2}a\left(\frac{b^2}{a^2} - \frac{c}{a}\right)\right] \exp\left[-\frac{1}{2}a\left(f_i - \frac{b}{a}\right)^2\right] \tag{A.70}$$

$$= \hat{Z}_i \frac{\sigma_{-i}}{\hat{\sigma}_i} \exp\left[\frac{1}{2}a\left(\frac{b^2}{a^2} - \frac{c}{a}\right)\right] \sqrt{2\pi a^{-1}} \frac{1}{\sqrt{2\pi a^{-1}}} \exp\left[-\frac{1}{2}a\left(f_i - \frac{b}{a}\right)^2\right]. \tag{A.71}$$

The term $\frac{1}{\sqrt{2\pi a^{-1}}} \exp\left[-\frac{1}{2}a\left(f_i - \frac{b}{a}\right)^2\right]$ is the density function of a normal distribution function with variance a^{-1} and mean $\frac{b}{a}$. This implies that we should set the site parameters to

$$Z_i = \hat{Z}_i \frac{\sigma_{-i}}{\hat{\sigma}_i} \exp\left[\frac{1}{2}a\left(\frac{b^2}{a^2} - \frac{c}{a}\right)\right] \sqrt{2\pi a^{-1}} \tag{A.72}$$

$$m_i = \frac{b}{a} \tag{A.73}$$

$$s_i = a^{-1} \tag{A.74}$$

We note that Z_i appears in a different form than in Rasmussen and Williams (2006), but it is consistent with the multivariate form of the results from Hernández-Lobato (2009) (see Equation A.49 of that thesis).

A.2.4 Approximate Log Marginal Likelihood

As with the Laplace approximation, we require an approximation to the log marginal distribution $p(\mathbf{y}|\mathbf{X})$. Based upon our approximation above, the marginal likelihood can be approximated by,

$$p(\mathbf{y}|\mathbf{X}) = q(\mathbf{y}|\mathbf{X}) \tag{A.75}$$

$$\approx \int p(\mathbf{f}|\theta) \prod_{i=1}^{N} q(f_i|Z_i, m_i, s_i^2) d\mathbf{f} \tag{A.76}$$

$$= \left(\prod_{i=1}^{N} Z_i \right) \int \boldsymbol{\phi}(\mathbf{f}; \mathbf{0}, \mathbf{K}) \boldsymbol{\phi}(\mathbf{f}; \mathbf{m}, \boldsymbol{\Sigma}) d\mathbf{f} \tag{A.77}$$

where $\boldsymbol{\phi}(\mathbf{x}; \boldsymbol{\mu}, \mathbf{V})$ is the density function of a multivariate normal distribution for a random variable $\mathbf{X} = \mathbf{x}$ with parameters mean $\boldsymbol{\mu}$, and variance $\mathbf{V}$.

To integrate A.77, we again appeal to Rasmussen and Williams (2006) (see Appendix A.2, results A.7 and A.8), which provides relationships for the product of two multivariate normal density functions,

$$\left(\prod_{i=1}^{N} Z_i \right) \int \boldsymbol{\phi}(\mathbf{f}; \mathbf{0}, \mathbf{K}) \boldsymbol{\phi}(\mathbf{f}; \mathbf{m}, \boldsymbol{\Sigma}) d\mathbf{f} \tag{A.78}$$

$$= \left(\prod_{i=1}^{N} Z_i \right) C^{-1} \int \boldsymbol{\phi}(\mathbf{f}; (\mathbf{K}^{-1} + \boldsymbol{\Sigma}^{-1})^{-1} \boldsymbol{\Sigma}^{-1} \mathbf{m}, (\mathbf{K}^{-1} + \boldsymbol{\Sigma}^{-1})^{-1}) d\mathbf{f} \tag{A.79}$$

where

$$C^{-1} = (2\pi)^{n/2} \det(\mathbf{K} + \boldsymbol{\Sigma})^{-1/2} \exp\left(-\frac{1}{2} \mathbf{m}^T (\mathbf{K} + \boldsymbol{\Sigma})^{-1} \mathbf{m} \right). \tag{A.80}$$

Continuing,

$$\left(\prod_{i=1}^{N} Z_i\right) C^{-1} \int \boldsymbol{\phi}(\mathbf{f}; (\mathbf{K}^{-1} + \boldsymbol{\Sigma}^{-1})^{-1}\boldsymbol{\Sigma}^{-1}\mathbf{m}, (\mathbf{K}^{-1} + \boldsymbol{\Sigma}^{-1})^{-1}) d\mathbf{f} \tag{A.81}$$

$$= \left(\prod_{i=1}^{N} Z_i\right) (2\pi)^{n/2} \det(\mathbf{K} + \boldsymbol{\Sigma})^{-1/2} \exp\left(-\frac{1}{2}\mathbf{m}^T(\mathbf{K} + \boldsymbol{\Sigma})^{-1}\mathbf{m}\right). \tag{A.82}$$

Finally, from A.82, we can take the logarithm to get the log marginal likelihood,

$$\log p(\mathbf{y}|\mathbf{X}) \approx \frac{n}{2}\log(2\pi) - \frac{1}{2}\log\det(\mathbf{K} + \boldsymbol{\Sigma}) - \frac{1}{2}\mathbf{m}^T(\mathbf{K} + \boldsymbol{\Sigma})^{-1}\mathbf{m} + \sum_{i=1}^{N} \log Z_i \tag{A.83}$$

and plug in the appropriate Z_i, m_i, s_i^2 terms.

A.2.5 A Note on Implementation of Expectation Propagation

The above results were derived using the mean and variance parameters in many of the equations. Many implementations of expectation propagation operate using the natural parameters of the distributions from the exponential family such as in the implementations of expectation propagation discussed in Gelman et al. (2014). See Rasmussen and Williams (2006) for an efficient implementation of the serial version of expectation propagation using a natural parametrization for binary regression with Gaussian processes. For implementations of the parallel version of this algorithm see Van Gerven et al. (2010); Tolvanen et al. (2014).

Appendix B

Measuring Predictability of Behavior - Entropy Rate Estimation

The previous chapters have focused on inference involving treatment effects within observational studies. This appendix provides a contribution for measurement within observational studies. In particular, this appendix focuses on investigating methods for the estimation of the entropy rate of finite-state Markov chains and the performance of these methods when the number of observed states of the sequence is small, as is typical in behavioral studies.

B.1 Overview of the Problem

The behavior of an individual is often characterized by recording a series of observed actions of that individual interacting with a specified environment. Behavior can be summarized in terms of the actions that occur most frequently, the proportion of time that specific actions occur, or the variability in the types of actions performed. Recently, mathematically-defined characteristics of the patterns and rhythms in behavior have shown to be important measures

and provide a novel dimension in predicting developmental outcomes in behavioral studies (Baram et al., 2012). Davis et al. (2017) demonstrated that the patterns in the observed behaviors of mothers can have an impact on the cognitive development of their children. Their study recorded mothers' actions as they interacted with their children, modeled the series of behaviors as a discrete-state stochastic process, specifically a Markov chain, and quantified predictability of maternal behavior through the *entropy rate* of the process, a measure of the predictability of a sequence of actions. The work demonstrates the potential utility of predictability as a behavioral measure for developmental studies and suggests that such measures could be used in other behavioral and educational settings. Recent advances in technology-based educational assessments, e.g., data generated using key logging during the CBAL writing assessment (Almond et al., 2012), or behavioral time-series data created during game-based assessments provide examples where the methods of Davis et al. (2017) might apply in educational research. This appendix presents and compares approaches for estimating the entropy rate of an observed stochastic sequence with a focus on understanding the properties of these estimators for relatively short sequences which are common in behavioral studies.

Entropy is a measurement that defines the predictability of a single random variable. Entropy rate extends the concept of entropy from random variables to stochastic processes. If we consider a sequence of random variables, entropy rate quantifies the limiting behavior of the joint entropy of the random variables as the sequence length increases. For a stationary Markov chain, the entropy rate is a function of the stationary distribution and the transition matrix that defines the dependence structure of the process. Thus one method of estimating the entropy rate is to estimate the transition matrix that defines the behavior of the system and the stationary distribution that defines the limiting behavior of the system. Another approach to estimating the entropy rate is based on Lempel-Ziv compression algorithms (Ziv and Lempel, 1977). Shannon (1948) described a relationship between the compressibility

of a sequence and the joint entropy of the sequence. In theory, Lempel-Ziv compression algorithms are optimal in achieving the compression limit put forth by Shannon for Markov processes of any order (Cover and Thomas, 2006; Wyner and Ziv, 1994) and have thus been useful for estimating the entropy rate.

Entropy rate has been used in a wide range of applications to characterize the behavior of signals in the presence of noise. Specifically it has been used in measuring the predictability of human mobility (Song et al., 2010; McInerney et al., 2013), assessing the complexity of short heart period variability series (Porta et al., 2001), characterizing neural spike trains (Amigó et al., 2004), classifying differences in behavior on Twitter (Chu et al., 2010), quantifying the difference in behavioral patterns induced by distinct environments (Molet et al., 2016), and associating predictability of maternal behavior with cognitive outcomes of their infants and children (Davis et al., 2017). Gao et al. (2008) provided a thorough analysis of entropy rate estimators for binary sequences and assessed performance of these estimators for long sequences. This appendix reports on a comprehensive investigation of various entropy rate estimation techniques for finite state spaces and smaller sequence lengths typical in behavior studies.

In Section B.2, we briefly review the definitions of Markov chains, entropy, and entropy rate. Section B.3 outlines techniques for estimating the entropy rate and a bootstrap procedure for obtaining the standard error of the entropy rate estimators. Section B.4 provides a simulation study comparing the performance of the estimators under varying sequence lengths and different data-generating processes. Section B.5 applies the estimators to data from Davis et al. (2017) to demonstrate how a Markov chain can be constructed from an observed time-series to estimate the entropy rate of a behavioral sequence. Finally, Section B.6 provides a discussion of the methods, their properties, and areas where the estimators may be useful.

B.2 Modeling Behavior

B.2.1 Modeling Behavior Using Finite-State Markov Chains

We model behavior as a sequence of observable actions through time. Let $\mathcal{X} = \{X_t\}$ be a stochastic process composed of a sequence of random variables observed at time points $t \in \mathcal{T} = \{0, 1, \ldots, T\}$. Further, we define the state space $\mathcal{A} = \{\alpha_1, \alpha_2, \ldots, \alpha_\kappa\}$ to be the finite set of κ actions that an individual performs. We observe $X_t = x_t$, with $x_t \in \mathcal{A}$, for all time points t. The action observed at time t, x_t, is sometimes referred to as the state of the stochastic process. We assume that the probability distribution of the random variable X_t depends only upon previously observed actions (i.e., the future does not affect the present). Then the joint distribution can be expressed as $\Pr(X_0 = x_0, \ldots, X_T = x_T) = \Pr(X_0 = x_0) \cdot \prod_{t=1}^{T} \Pr(X_t = x_t | X_{t-1} = x_{t-1}, \ldots, X_0 = x_0)$ where each $x_t \in \mathcal{A}$. We simplify the dependence structure and assume that only the previous m observations are relevant, then

$$\begin{aligned} & \Pr(X_t = x_t | X_{t-1} = x_{t-1}, \ldots, X_0 = x_0) \\ = \; & \Pr(X_t = x_t | X_{t-1} = x_{t-1}, \ldots, X_{t-m} = x_{t-m}). \end{aligned} \tag{B.1}$$

A stochastic process with this assumption and a countable state space is known as an m^{th}-order Markov chain (see Karlin and Taylor, 1975, Chapter 1, Section 3c).

We briefly review required Markov chain theory using a first-order Markov chain (i.e., $m = 1$) and then generalize below to higher-order Markov chains. Let $P_{ij} = P(\alpha_i, \alpha_j) = \Pr(X_t = \alpha_j | X_{t-1} = \alpha_i)$ be the probability of transitioning from state α_i at time $t-1$ to state α_j at time t (we assume the transition probabilities are independent of time so the process is time-

homogeneous). Properties of the transition matrix $\mathbf{P} = \{P_{ij}\}$ are important for understanding the behavior of the process. Let μ_t represent the marginal distribution of the random variable X_t. It is easy to show that the distribution at time $t + n$ is $\mu_{t+n} = \mu_t \mathbf{P}^n$. The matrix $\mathbf{P}^n$ is referred to as the n-step transition matrix of the first-order Markov chain.

In addition to assuming the order of the process is known and that the process is time-homogeneous, we also assume that the Markov chain is irreducible and stationary. Levin et al. (2009) provide that a Markov chain is **irreducible** if for any two actions $\alpha_i, \alpha_j \in \mathcal{A}$ there exists an integer t (possibly depending on α_i and α_j) such that $\mathbf{P}^t(\alpha_i, \alpha_j) > 0$ and therefore there is no absorbing state in the observed process. The assumption of stationarity implies that the joint distributions on sets of random variables $(X_{t_1+h}, X_{t_2+h}, \ldots, X_{t_n+h})$ and $(X_{t_1}, X_{t_2}, \ldots, X_{t_n})$ are the same for all $h > 0$ and arbitrary $t_1, t_2, \ldots, t_n$ from $\mathcal{T}$ (Karlin and Taylor, 1975). For a Markov process to be a stationary process μ_t, the distribution of X_t, and μ_{t-1}, the distribution of X_{t-1}, must be the same. This common distribution is known as the stationary distribution $\pi = (\pi(\alpha_1), \pi(\alpha_2), \ldots, \pi(\alpha_k)) = (\pi_1, \pi_2, \ldots, \pi_k)$. Additionally, since $\mu_t = \mu_{t-1}\mathbf{P}$ it follows that the stationary distribution of a Markov process must satisfy $\pi = \pi\mathbf{P}$. The stationary distribution π is not guaranteed to exist for all Markov chains, but the assumption of irreducibility implies that the distribution π will both exist and be unique (see Levin et al., 2009, Proposition 1.14 and Corollary 1.17). A useful interpretation of the stationary distribution is that it describes the asymptotic proportion of time that the Markov chain will spend in any state (Levin et al., 2009). We do not make any assumption regarding the periodicity of the Markov chain.

Now we generalize these definitions to higher-order Markov processes. To do this, we first develop a first-order Markov chain *on vectors* of random variables. Let $\mathcal{Y} = \{\mathbf{Y}_t\}$ be a stochastic process composed of a set of random vectors with m elements each and a discrete time index. As an example that is relevant below, we may have, $\mathbf{Y}_t = (Y_{t,0}, Y_{t,1}, \ldots, Y_{t,m-1})^T$,

a column vector of random variables, each with state space $\mathcal{A}$. The state space of each $\mathbf{Y}_t \in \mathcal{Y}$, denoted as $\mathcal{B}$, is therefore the m-fold Cartesian product of the state spaces of the component random variables and each $\beta_i \in \mathcal{B}$ represents an ordered m-tuple of states or actions. We can construct transition matrices for the vector process, but now the matrix gives the probability of transitions between m-tuples. Definitions of stationarity and irreducibility follow as in the earlier discussions.

To relate m^{th}-order Markov chains of random variables to first-order chains of random vectors, note that we can construct a vector process from the original m^{th}-order univariate process using contiguous subsequences of random variables of the original sequence. We define $\mathbf{X}_i^m$ as the subsequence of $\{X_t\}$ starting from index i and composed of m contiguous observations, $\mathbf{X}_i^m = (X_{i+(m-1)}, X_{i+(m-2)}, \ldots, X_i)^T$ and consider the stochastic process of interest to be $\{\mathbf{X}_t^m\}$ for $t \in \{0, 1, \ldots T-(m-1)\}$. Now if we consider the probability of transitioning from $\mathbf{X}_{t-1}^m$ to $\mathbf{X}_t^m$, we find that $\Pr(\mathbf{X}_t^m = \mathbf{x}_t | \mathbf{X}_{t-1}^m = \mathbf{x}_{t-1}) = \Pr(X_{t+m-1} = x_{t+m-1} | X_{t+m-2} = x_{t+m-2}, \ldots, X_{t-1} = x_{t-1})$, where the expression follows because $\mathbf{X}_t^m$ and $\mathbf{X}_{t-1}^m$ share $m-1$ common random variables. This demonstrates that first-order transitions of vectors of the newly constructed process are equivalent to m^{th} order transitions of the original process. The advantage of this construction is that we can find an appropriate first-order transition matrix for the process of random vectors and utilize all of the definitions which we have previously outlined. It is worth noting that under this construction the transition matrix will be of dimension $(\kappa^m \times \kappa^m)$ and may pose computational issues for large m or if the cardinality of $\mathcal{A}$ is large.

B.2.2 Entropy Rate of a Finite Markov Chain

Shannon (1948) introduced the concept of entropy in the context of communication. The entropy of a discrete random variable X, taking values from the state space $\mathcal{A}$, is $H(X) =$

$-\sum_{\alpha_i \in \mathcal{A}} \Pr(X = \alpha_i) \log_2 \Pr(X = \alpha_i)$. Entropy varies between zero and $\log_2 \kappa$, where κ is the cardinality of $\mathcal{A}$. The definition of entropy is easily generalized to joint distributions of random variables, as well as to conditional distributions (for an overview see Cover and Thomas, 2006).

Cover and Thomas (2006) define the entropy rate of a stochastic process $\mathcal{X} = \{X_t\}$, $t \in \{0, 1, \ldots, T\}$ as $H(\mathcal{X}) = \lim_{T \to \infty} \frac{1}{T} H(X_0, X_1, \ldots, X_T)$ and further show that if the process is stationary this is equivalent to $H(\mathcal{X}) = \lim_{T \to \infty} H(X_T | X_{T-1}, X_{T-2}, \ldots, X_0)$ provided that the limit exists. The entropy rate of a stationary process provides a quantification of the predictability of the next observation given the history of observations which occurred before it.

We apply this definition to a stationary first-order time-homogenous Markov process, $\mathcal{X} = \{X_t\}, t \in \{0, 1, \ldots, T\}$, with finite state space $\mathcal{A} = \{\alpha_1, \alpha_2, \ldots, \alpha_\kappa\}$. Utilizing the framework outlined in Section B.2.1, we find that the entropy rate of the process is

$$
\begin{aligned}
H(\mathcal{X}) &= \lim_{T \to \infty} H(X_T | X_{T-1}, X_{T-2}, \ldots, X_0) \\
&= \lim_{T \to \infty} H(X_T | X_{T-1}) \\
&= \lim_{T \to \infty} - \left[\sum_{\alpha_i \in \mathcal{A}} \Pr(X_{T-1} = \alpha_i) \times \right. \\
&\qquad \left. \left(\sum_{\alpha_j \in \mathcal{A}} \Pr(X_T = \alpha_j | X_{T-1} = \alpha_i) \log_2 \Pr(X_T = \alpha_j | X_{T-1} = \alpha_i) \right) \right].
\end{aligned}
$$

The stationarity assumption and the assumption of a time-homogeneous Markov process implies $\Pr(X_T = \alpha_j | X_{T-1} = \alpha_i) = P(\alpha_i, \alpha_j) = P_{ij}$. Noting that $\lim_{T \to \infty} \Pr(X_{T-1} = \alpha_i) = \pi_i$ for all $\alpha_i \in \mathcal{A}$, we can write $H(\mathcal{X}) = -\sum_{\alpha_i, \alpha_j \in \mathcal{A}} \pi_i P_{ij} \log_2 P_{ij}$. The entropy rate can be seen to be a weighted average of the conditional entropy of X_T given the previous state $X_{T-1} = \alpha_i$, where the weights are given by the stationary distribution. These results generalize easily for

$m > 1$ using the approach laid out in Section B.2.1.

B.2.3 Entropy Rate from Lempel-Ziv Compression

Several authors (Amigó et al., 2004; Song et al., 2010; McInerney et al., 2013) have utilized estimators derived from properties of Lempel-Ziv compression algorithms to estimate the entropy rate. We briefly describe the idea behind Lempel-Ziv compression and its relationship to entropy.

Lempel and Ziv (1976) introduced methods to quantify the complexity of finite sequences, as well as a framework for parsing sequences for compression. They were able to relate their measure of the complexity of a sequence to the entropy rate of the source which generated that sequence. They further provided algorithms (Ziv and Lempel, 1977, 1978) for compression of a sequence that can also be utilized for estimating the entropy rate of the sequence. We use the algorithm developed in 1977, which we will refer to as the sliding-window Lempel-Ziv, or SWLZ, algorithm. The optimality of the SWLZ algorithm for achieving a compression ratio approaching the entropy of a stationary ergodic source was provided in Wyner and Ziv (1989), and investigated further in Wyner and Ziv (1994) and Ornstein and Weiss (1993).

To understand the SWLZ algorithm, consider a sequence of n observations, $x_0 x_1 \dots x_n$, as a realization from a stochastic process $\mathcal{X}$ with finite state space $\mathcal{A}$. Further define x_i^j to be the subsequence $x_i x_{i+1} \dots x_{j-1} x_j$, so that x_0^{i-1} is the history of the subsequence before observation x_i. The objective of the algorithm is to sequentially process the observations from x_0 to x_n and partition the original sequence into unique subsequences. To create these unique subsequences, consider that at some point $i \in \{1, \dots, n\}$ the previous $i-1$ observations have been parsed into subsequences. The algorithm next identifies the shortest length L such

Table B.1: Example parsing based upon the SWLZ algorithm.

Original Sequence	13131213232331313332
SWLZ Parsing	1|3|131|2|132|323|31313|332

that $x_i^{i+L-1} \not\subseteq x_0^{i-1}$, i.e. the sequence of length L starting at x_i that has not been observed before. At this point the substring x_i^{i+L-1} is unique and we add it as a new parsing and move to position x_{i+L}. This continues until the entire sequence is parsed, noting that the last subsequence may not be unique when all characters are exhausted. Table B.1 shows a unique parsing of a three state stochastic process. In this example, note that the first symbol is unique ($\{1\}$) and that the second symbol is not equal to the first ($\{3\}$), so they both are unique subsequences. When we start from the third observation we see that that $x_2 = 1$ is contained in the history $x_0x_1 = 13$ and so is $x_2x_3 = 13$, but the sequence $x_2x_3x_4 = 131$ is a unique subsequence which has not been seen before. The algorithm stores this new subsequence and the process is continued then from observation x_5.

The primary theoretical argument of the optimality of this algorithm relies on considering a *doubly-infinite* sequence, $x_{-\infty} \ldots x_{-2}x_{-1}x_0x_1 \ldots x_{\infty}$ from a stationary ergodic process $\mathcal{X}$ with $x_j \in \mathcal{A}$. For $n = 1, 2, \ldots,$ Wyner and Ziv (1989) define $L_n(\mathcal{X})$ to be the smallest integer $L > 0$ such that x_0^{L-1} does not appear as a contiguous subsequence of x_{-n}^{-1}, or $L_n(\mathcal{X}) = \arg\min_L x_0^{L-1} \not\subseteq x_{-n}^{-1}$. Note that $L_n(\mathcal{X})$ is the length of the unique parsing starting at x_0 using a history of size n by our previous definitions. Under the assumption of a Markov process, Wyner and Ziv (1989) show that the entropy rate of the source $\mathcal{X}$ is related to the length of this parsing size and the size of the history n used to create that parsing,

$$\frac{\log_2(n)}{L_n(\mathcal{X})} \rightarrow_P H(\mathcal{X}), \text{ as } n \rightarrow \infty. \tag{B.2}$$

This theoretical result is used below as the basis for an estimator of the entropy rate.

B.3 Estimating the Entropy Rate

B.3.1 Direct Estimation of Entropy Rate

One approach to estimating the entropy rate is to estimate the transition matrix and stationary distribution and then apply the formula of Section B.2.2. Directly estimating the transition matrix $\mathbf{P}$ is straightforward using the observed transitions. To estimate π, we describe two methods: 1) estimation of the stationary distribution based upon the observed proportion of time the process visits each state; 2) estimation of the stationary distribution based upon an eigenvalue decomposition of the transpose of the estimated transition matrix. Once we have estimated $\mathbf{P}$ by $\hat{\mathbf{P}}$ and then π by $\hat{\pi}$, we can estimate the entropy rate $H(\mathcal{X})$ as follows $\hat{H}(\mathcal{X}) = -\sum_i \sum_j \hat{\pi}_i \hat{P}_{ij} \log_2 \hat{P}_{ij}$.

B.3.1.1 Estimation of Transition Probabilities

The transition matrix $\mathbf{P}$ can be estimated from the observed transitions of the Markov chain. If we denote n_{ij} as the total number of transitions from i^{th} row action to the j^{th} column action, we can create a matrix $\mathbf{T}_C = \{n_{ij}\}$ of transition counts. To convert this to a valid estimator of the transition matrix, we normalize each row by its corresponding row total. Define n_{i+} to be the sum of the counts over all columns j in row i, $n_{i+} = \sum_j n_{ij}$. The empirical estimator for the transition matrix, $\hat{\mathbf{P}} = \left\{ \frac{n_{ij}}{n_{i+}} \right\}$, is the maximum likelihood estimator for $\mathbf{P}$ (Murphy, 2012, Section 17.2).

B.3.1.2 Estimation of the Stationary Distribution

Using Observed Frequencies of States. The stationary distribution of the Markov process can be estimated by considering the proportion of time each action is observed in a realization of this process. Thus an empirical estimator of π would be, $\hat{\pi}_{emp} = \left\{ \frac{n_{i+}}{n_{++}} \right\}$, where n_{i+} is as defined previously and $n_{++} = \sum_i n_{i+} = \sum_i \sum_j n_{ij}$ is the total number of transitions. It can be shown algebraically that $\hat{\pi}_{emp} \neq \hat{\pi}_{emp}\hat{\mathbf{P}}$, but the discrepancy will be small if the number of observations of the process is large.

Using an Eigendecomposition of $\hat{\mathbf{P}}^T$. An alternative is to estimate π from an eigendecomposition of $\hat{\mathbf{P}}^T$. For a stationary Markov process, the stationary distribution is a row vector and must satisfy $\pi = \pi\mathbf{P}$, or equivalently, $\pi^T = \mathbf{P}^T\pi^T$, which means that the stationary distribution, π, is the transpose of the eigenvector corresponding to an eigenvalue equal to 1 of $\mathbf{P}^T$. Because we have assumed an irreducible process there will be such an eigenvector (see Karlin and Taylor, 1981, Chapter 12, Theorem 3.1).

An estimator for the stationary distribution can be obtained by performing an eigendecomposition of the matrix $\hat{\mathbf{P}}^T$ and setting $\hat{\pi}_{eig}$ proportional to the relevant eigenvector (i.e., an eigenvector that corresponds to the eigenvalue of $\lambda = 1$). Define $\hat{x}_{\lambda=1}$ to be this eigenvector and obtain an estimate of the stationary distribution as follows, $\hat{\pi}_{eig} = \hat{x}^T_{\lambda=1} / \sum_{i=1}^k \hat{x}_{i,\lambda=1}$, where the denominator ensures that this is a valid probability distribution.

Using a Limit of $\hat{\mathbf{P}}^n$. There is another way to estimate π by taking the limit of an infinite number of "transitions" based upon an estimated transition matrix $\hat{\mathbf{P}}$. For an irreducible Markov chain, Theorem 5.1.4 of Kemeny and Snell (1960) states that $\mathbf{P}^n$ will be Cesaro-summable to a matrix Π where each row of Π is the stationary distribution π, i.e. $\lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^n P^i(\alpha_j, \cdot) = \pi$ for all $\alpha_j \in \mathcal{A}$. Therefore for a very large value of N, π can be

estimated as, $\hat{\pi}_{limit} = \frac{1}{N}\sum_{i=1}^{N}\hat{P}^i(\alpha_1, \cdot)$. This estimator achieves very slow convergence to the true π. Additionally, there are computational challenges to taking powers of matrices. Due to these computational and convergence issues, we do not provide simulation results for this estimator.

B.3.2 Sliding Window Lempel-Ziv Entropy Rate Estimation

An alternative approach to estimating the entropy rate is to rely on the theoretical results for the asymptotic optimality of the SWLZ compression algorithm for stationary ergodic sources. Recall that the sequence in Equation (B.2), $\log_2(n)/L_n(\mathcal{X})$, converges in probability to the entropy rate of a Markov process, where $L_n(\mathcal{X})$ is the length of a unique parsing of the observed sequence when using a history of size n. Kontoyiannis et al. (1998) show that a Cesàro summation of this series will also converge if the source is a stationary ergodic Markov process and that an estimator based on this relationship will be asymptotically consistent for estimating the entropy rate.

Here we consider a modified version of the SWLZ algorithm that is convenient for finding the entropy rate of a Markov chain. Defining $\Lambda_i(\mathcal{X})$ to be the length of the shortest substring x_i^{i+L-1} that does not appear as a substring in the history of i symbols, x_0^{i-1}, Theorem 1C of Kontoyiannis et al. (1998) provides that if $\mathcal{X} = \{X_i\}, i \in Z$ is a two-sided stationary ergodic Markov process with entropy $H(\mathcal{X}) > 0$, then $\lim_{n\to\infty}\frac{1}{n}\sum_{i=1}^{n}(\Lambda_i(\mathcal{X})/log_2 n) = (H(\mathcal{X}))^{-1}$ almost surely. This result becomes the basis for an estimator.

To estimate $H(\mathcal{X})$, we consider a realization of the process $\mathcal{X}$ to be the observed sequence, $x_0 x_1 \ldots x_T$. At each instance i of the observed sequence, we compute $\Lambda_i(\mathcal{X}) = \arg\min_L x_i^{i+L-1} \not\subseteq x_0^{i-1}$. We outline the process in Table B.2. To estimate the entropy rate we divide the bits required to encode a sequence length of n, $\log_2 n$, by the average size of a

Table B.2: Step i of the Sliding Window Lempel Ziv Algortihm

Original String: $x_0x_1 \ldots x_{i-1}x_ix_{i+1} \ldots x_{T-1}x_T$			
Step	History	Candidate String	$x_i^{i+j-1} \subseteq x_0^{i-1}$
i.1	$x_0x_1 \ldots x_{i-1}$	x_i	TRUE
i.2	$x_0x_1 \ldots x_{i-1}$	x_ix_{i+1}	TRUE
$\vdots$	$\vdots$	$\vdots$	$\vdots$
i.(L-1)	$x_0x_1 \ldots x_{i-1}$	$x_ix_{i+1} \ldots x_{i+L-2}$	TRUE
i.L	$x_0x_1 \ldots x_{i-1}$	$x_ix_{i+1} \ldots x_{i+L-1}$	FALSE
	Size of Unique Parsing, $\Lambda_i(\mathcal{X})$		L

unique parsing from this process. That is:

$$\hat{H}(\mathcal{X}) = \left[\frac{1}{n}\sum_{i=1}^{n}\frac{\Lambda_i(\mathcal{X})}{log_2 n}\right]^{-1} = \frac{\log_2 n}{\frac{1}{n}\sum_{i=1}^{n}\Lambda_i(\mathcal{X})} \tag{B.3}$$

This version of the estimator considers a window size, or history, which "expands" at each point i. There are alternative versions of the estimator which consider fixed window sizes. It has been our experience that the expanding window estimators perform the best for short sequence lengths (see Gao et al. (2008) for a broader comparison of those methods).

B.3.3 Estimation of Standard Errors via the Stationary Bootstrap

The previous sections provide methods for obtaining point estimates of the entropy rate of a finite Markov chain. In this section we provide a method to estimate the standard error of these point estimates. A common approach to measuring the standard error of an estimate is to use the bootstrap method (Efron and Tibshirani, 1994). This method works well when observations are independent and identically distributed, but our application is focused on dependent data. We therefore use a method called the "stationary bootstrap" of Politis and Romano (1994) which is a variant of the block bootstrap.

The block bootstrap (Efron and Tibshirani, 1994) is a common approach to creating bootstrap samples of dependent data. The simple block bootstrap begins by first partitioning a sequence of observations into blocks of equal size. To create a new bootstrap sequence, blocks are sampled with replacement from the partition and concatenated to create a sequence of approximately the same length as the original sequence. An estimate of the quantity of interest is calculated on this new sequence and this process is repeated many times. The standard error of the estimate is measured using the distribution of the point estimates across the bootstrap replicates. An alternative version of the block bootstrap considers blocks of equal size, but the blocks are allowed to overlap. Both methods are attractive in that they keep dependence between observations in the data, but are complicated by the need to choose the block size.

The stationary bootstrap is an extension of the block bootstrap which uses variable block sizes. We describe the algorithm laid out in Politis and Romano (1994) in the context of our notation. First define $C(i, l) = x_i^{i+l-1}$, where if $j \in \{i, i+1, \ldots, i+l-1\} \geq \mathcal{T}+1$ we set $j \equiv j \bmod \mathcal{T}+1$ (i.e, if a block extends past the end of the sequence, we cycle back to the beginning). Additionally define sequences of independent random variables $\mathcal{I}_1, \mathcal{I}_2, \ldots$ and $L_1, L_2, \ldots$ such that $\mathcal{I}_i$ is distributed discrete uniform on the set $\{0, 1, \ldots, \mathcal{T}\}$ and $L_i \sim Geometric(p)$ and let the random subsequence, X_k^* be defined as $X_k^* = C(\mathcal{I}_k, L_k)$. Finally define $\mathcal{X}^* = X_0^* \ldots X_{k-1}^* X_k^*$ as the concatenation of the $k+1$ random subsequences. While $\sum_{i=0}^{k} L_i \leq \mathcal{T}+1$, we draw $\mathcal{I}_{k+1}$ and L_{k+1} and set $\mathcal{X}^* = X_0^* \ldots X_k^* X_{k+1}^*$. Once the length of $\mathcal{X}^*$ is greater than $\mathcal{T}+1$ we take the first $\mathcal{T}+1$ observations and define that to be the stationary bootstrap sample $\mathcal{X}^b$. We estimate the entropy rate $H(\mathcal{X}^b)$ for each bootstrap replicate and then estimate the standard error of the bootstrap replicates.

An advantage of the method is that given the original sequence of observations, the new sequence is stationary (see Politis and Romano, 1994, Proposition 1), while the traditional

block bootstrap sequences are not. The method requires specifying the parameter p, which is analogous to choosing a block size in the traditional block bootstrap. Fortunately, the SWLZ estimate suggests a natural approach. Recall that our estimator of the entropy rate is $\log_2(n)$ divided by the average unique block length. This suggests that the average unique block length is approximately equal to $\log_2(n)/\hat{H}(\mathcal{X})$. The average block length in the stationary bootstrap is $E(L_i) = p^{-1}$. Equating these two values suggests choosing p equal to $\hat{H}(\mathcal{X})/\log_2(n)$. Section B.4.2 provides a simulation study which demonstrates the performance of this method for selecting p.

B.4 Simulations

Our research is motivated by an application of Markov chains to the study of behavior. Before exploring the application, we carry out a simulation study to illustrate the performance of the three entropy rate estimators outlined in Section B.3. We compare the methods when we have correctly specified the model (the order of the Markov process) and when the order of the Markov process is misspecified.

B.4.1 Estimation of First-Order Markov Processes

The first setting of the simulation study is stationary time-homogeneous first-order Markov processes with eight unique states. We consider three different data-generating models: a low entropy rate case, a medium entropy rate case, and a high entropy rate case. Figure B.1 provides a visualization of the transition matrix for each data-generating model with each box representing P_{ij} (scale indicated at right). The true entropy rate for each process is listed in the figure. The entropy rate for a Markov process with eight states lies between 0 and 3.

The low entropy rate transition matrix simulates a system where there is a high probability of transitions to the same state ($P_{ii} = 0.95$). The high entropy rate transition matrix is a much less organized system, designed to behave like a purely random system ($P_{ij} \approx 0.125$). The medium entropy rate transition matrix is designed to be a balance between predictability and unpredictability, with the characteristic that for some states the next state is more predictable than others.

We assess the performance of the various estimators as a function of observed sequence length. In a behavioral setting we are often interested in determining the number of observations required to obtain reliable inference and this simulation study will help inform this decision as well as provide information about the trade-offs between the estimation methods. Each Markov chain was simulated 100 times, with each simulation consisting of 10000 observations from the process. We observe $\mathcal{X}_i = x_0 x_1, \ldots x_{9999}$ for $i = 1, \ldots 100$. We applied the estimation procedures of Section B.3 to different length subsequences of each realization of the Markov chain and calculated an estimate of the entropy rate based on x_0^{j-1} for $j \in \{50, 250, 500, 1000, 5000, 10000\}$. The smaller values were chosen because behavioral applications, such as our motivating example, typically observe sequence lengths of approximately 250 to 1000, while 5000 and 10000 were chosen to explore performance for longer sequence lengths. Gao et al. (2008) provides a discussion of entropy rate estimation for sequences of much longer length.

Figure B.2 provides the results from the simulation study and is organized as follows. The top row of the figure contains results for the low entropy rate data-generating model, followed by the second row which provides results for the medium entropy rate data-generating model. The final row gives results for the high entropy rate model. The first column of the figure demonstrates the performance of directly estimating the entropy rate with an empirical estimate of the transition matrix and of the stationary distribution, while the second

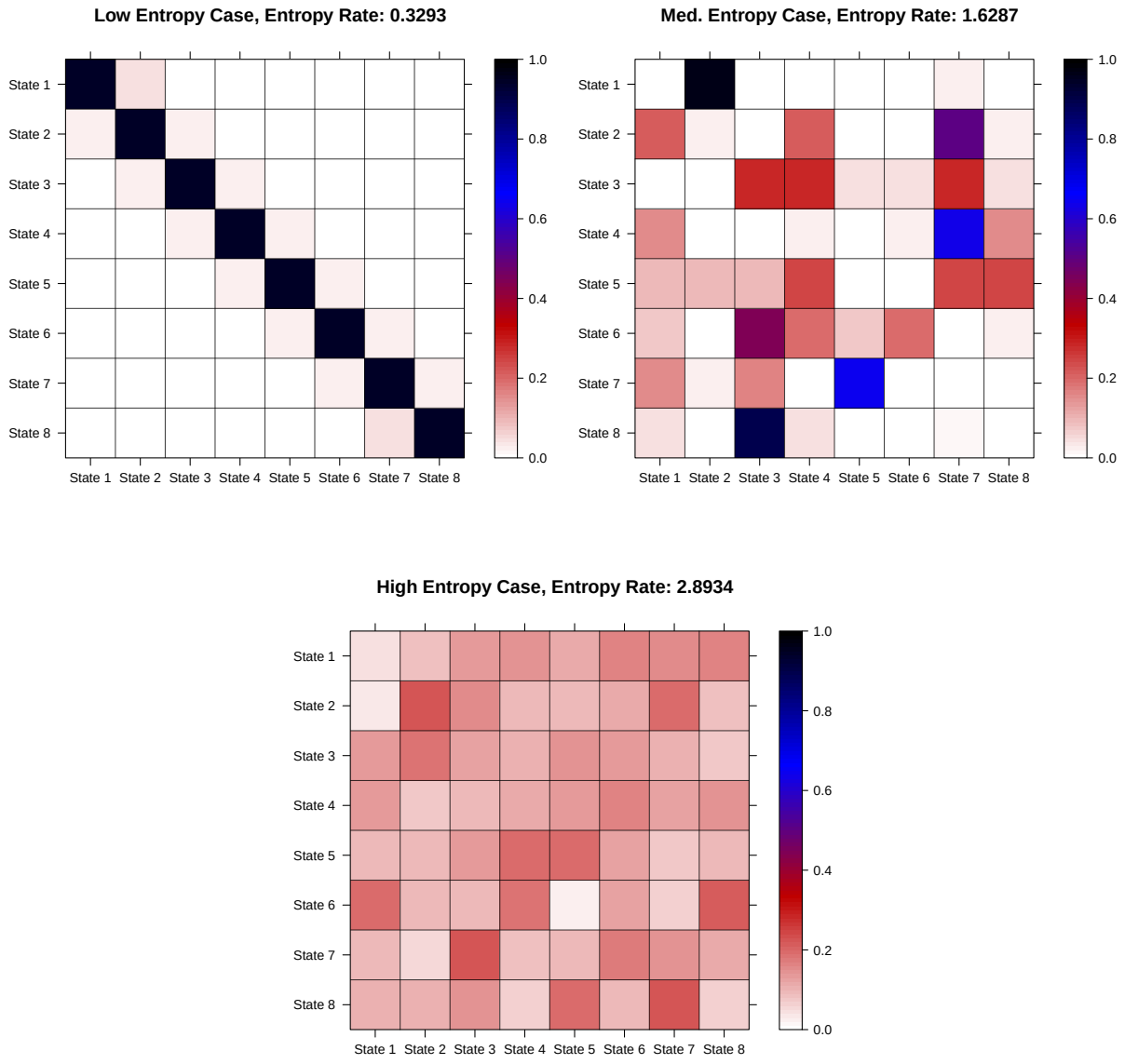

Figure B.1: Transition matrices describing three different types of behaviors: low entropy rate, intermediate entropy rate, and high entropy rate.

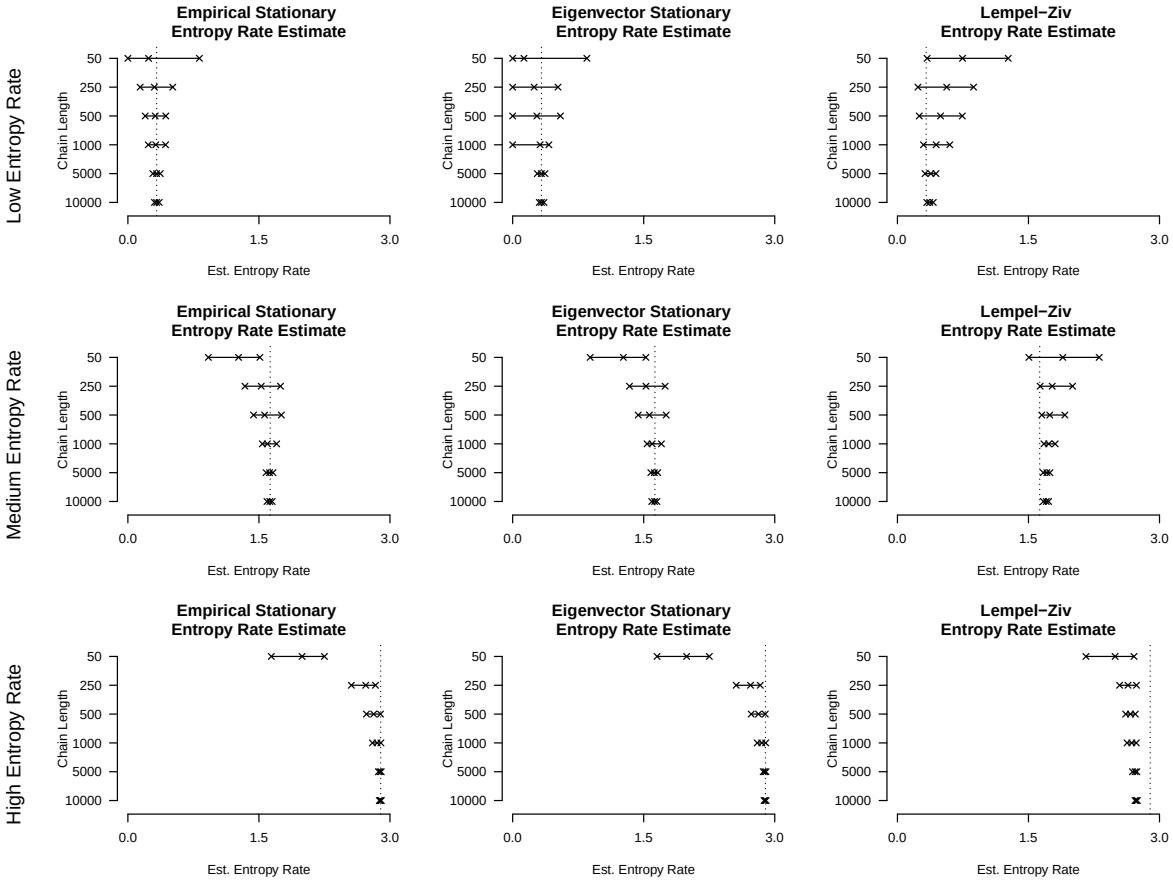

Figure B.2: Simulation results for 100 simulated Markov processes for each of the transition matrices shown in Figure B.1, ordered from top to bottom: low entropy rate, medium entropy rate, and high entropy rate. Entropy rate estimates were obtained at subsequence lengths 50, 250, 500, 1000, 5000, 10000. Each column of the figure represents a different entropy rate estimation technique.

column illustrates the performance of using an empirical estimate of the transition matrix and an eigendecomposition of the transpose of the observed transition matrix to estimate the stationary distribution of the process. The last column provides the performance of the estimator based on the SWLZ algorithm. Each line within a subplot displays the spread of the entropy rate estimates from 100 realizations of the process. The marks on the line represent the lowest entropy rate, mean entropy rate, and highest entropy rate observed across these 100 realizations. A primary result of the simulation study is that estimates from almost all realizations converge to a common value for long sequences, for all of the transition matrices.

The first row of Figure B.2 gives the performance of the three estimators in estimating the entropy rate of structured and predictable systems, i.e., low entropy rate systems. We see that results are consistent with the true entropy rate when we empirically estimate the transition matrix and the stationary distribution from the observed sequence. For this estimator, the average across the simulations is close to the true entropy rate even for relatively short sequences of length 250 and appears to be an extremely reliable estimate of the true entropy rate for sequences of length 5000 or greater. When we estimate the stationary distribution by an eigendecomposition of $\hat{\mathbf{P}}^T$ and then estimate the entropy rate for low entropy rate systems, the second plot in the first row, we notice some performance issues. In this case, we occasionally observe estimates of the entropy rate which are zero. This is a result of the transition matrix that we have chosen for this example. From Figure B.1, there is a very high probability of an observation being followed by another observation of the same state ($P_{ii} = 0.95$). It is therefore possible with shorter sequences that we will not have observed a transition out of one or more states when the sequence terminates. This creates a reducible transition matrix and the results for stationary distributions defined by eigenvalues of transition matrices are no longer valid, resulting in an entropy rate estimate of zero in our simulations. For longer sequences this approach works well. The SWLZ estimates appear to be biased high for low entropy rate systems, but approach the true entropy rate as sequence length increases. Recall

from Equation (B.2), that the theoretical results show that $\frac{\log_2(n)}{L_n(\mathcal{X})} \to_P H(\mathcal{X})$, as $n \to \infty$ and that our estimator in Equation (B.3) is $\hat{H}(\mathcal{X}) = \log_2 n / \left(\frac{1}{n}\sum_{i=1}^{n} \Lambda_i(\mathcal{X})\right)$. The numerator, $\log_2(n)$ appears because the underlying theory effectively assumes that there is a history of size n when finding the length of the unique block $\Lambda_i(\mathcal{X})$. In truth the algorithm only has at its disposal a history of size $i < n$. Due to this limited history, the entropy rate is overestimated because the lengths of unique subsequences are shorter than expected for a history of size n. As the sequence length increases, this becomes less of a concern and the estimates approach the true entropy rate.

Now consider estimation of systems which are highly unpredictable, the last row of Figure B.2. When the entropy rate is estimated using an empirical estimate of the transition matrix and stationary distribution, the mean of the estimates approach the true entropy rate from below as sequence length increases. This suggests that this method systematically underestimates the randomness of the system for short sequence lengths. Also, compared with low entropy rate systems, it takes more observations (i.e., 1000 vs. 250) for the mean of the estimates to converge to the true entropy rate. The same pattern holds when the stationary distribution is estimated by an eigendecomposition of the transpose of the empirical transition matrix, the second plot in this row. The SWLZ estimator has difficulties estimating the entropy rate in high entropy rate systems that appear similar to its difficulties for low entropy rate systems, but in the opposite direction. Instead of being biased high, the entropy rate estimate is biased below the truth. Using a similar argument as was used for low entropy rate systems, this implies that on average the unique subsequence lengths obtained are longer than expected. One potential explanation for the bias is the fact that we can only find strings of integer length. Equation (B.2) implies that we should expect unique strings of mean length of approximately 3.44 for a sequence of length 1000 to achieve $H(X) = 2.8934$. Finding unique strings of length 2 or 3 when using a history of 1000 observations is less likely than unique strings of length 4 or 5, and therefore we obtain longer than needed strings. We note additional

simulations (not shown) indicate that this bias is still present for sequences of length 50000.

Results for intermediate entropy rate systems, shown in the middle row of Figure B.2, indicate similar performance to the other rows of the figure. For both methods of direct estimation, the estimates approach the true rate from below. The entropy rate estimates based on SWLZ are biased high in the intermediate entropy rate case, similar to the low entropy rate case. Estimates based upon the SWLZ estimator are approaching the true entropy rate, but we have not observed sequences long enough to achieve convergence.

B.4.2 Estimating the Standard Error via the Stationary Bootstrap

The previous section highlighted the performance of the methods in providing a point estimate of the entropy rate of the process. In this section, we investigate the performance of the stationary bootstrap in estimating the standard error of the entropy rate estimate. The simulation setup is the same as in the previous section. The estimate of the entropy rate is used to select the parameter p for the stationary bootstrap, $p = \hat{H}(\mathcal{X})/\log_2(j)$, where j is the sequence length. One hundred bootstrap replicates of each of the 100 simulated series were constructed as outlined in Section B.3.3. The entropy rate was estimated on each of these 100 bootstrap samples and the standard error was calculated from these samples and recorded. Thus we have 100 different bootstrap standard error estimates for each simulation scenario. We use the empirical standard error of the entropy rate estimates across the 100 simulations of Section B.4.1 as a reference for assessing the performance of the bootstrap standard errors.

Figure B.3 provides the results for the standard error estimates and is organized similarly to Figure B.2. The first result from these figures is that the empirical standard error (the blue dots) agrees with the distribution of bootstrap standard errors (represented by the x's

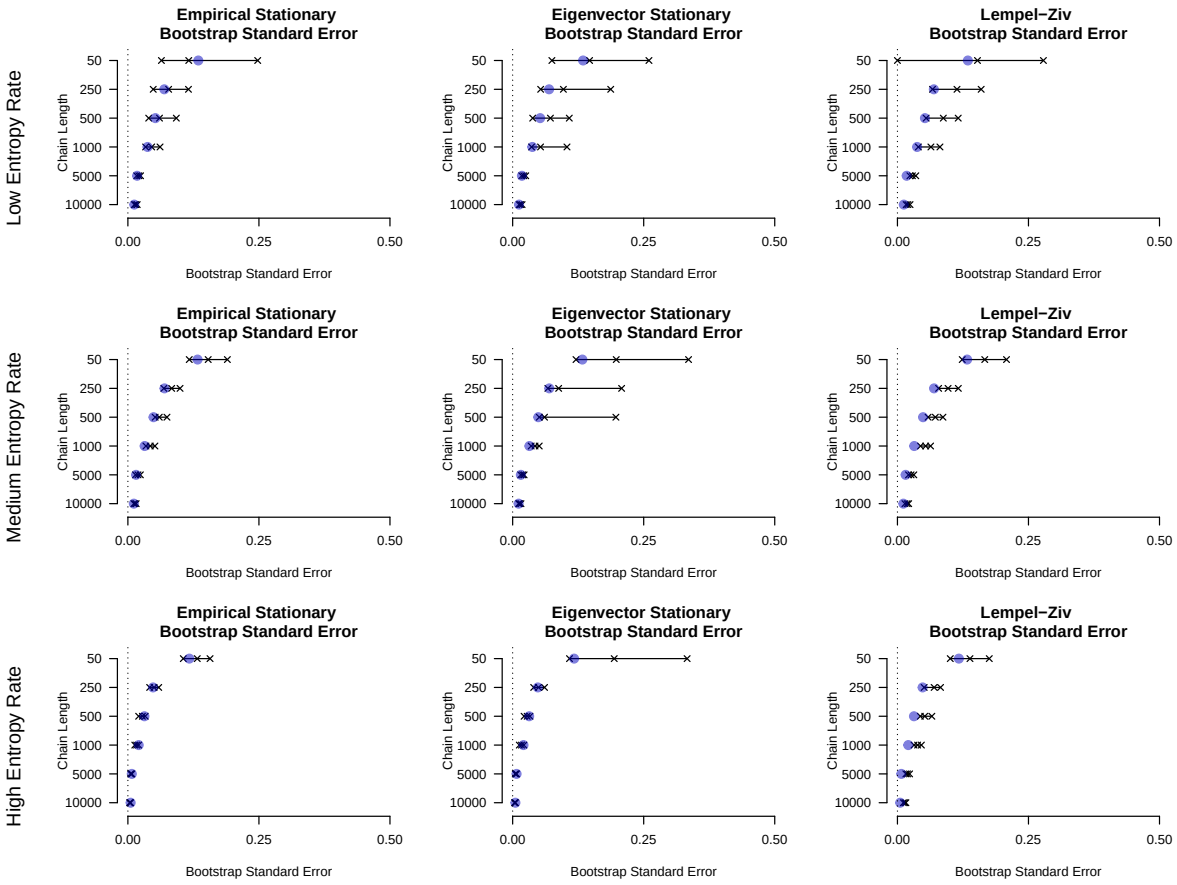

Figure B.3: Simulation results using the stationary bootstrap to estimate the standard error of the entropy rate estimate. Rows provide results for three entropy rate scenarios, ordered from top to bottom: low entropy rate, medium entropy rate, and high entropy rate. Bootstrap standard errors were obtained for each of 100 simulated Markov processes at subsequence lengths 50, 250, 500, 1000, 5000, 10000. Each column of the figure represents a different entropy rate estimation technique. In each simulation the entropy was estimated and the parameter p for the stationary bootstrap was selected as described in Section B.3.3. The blue dots represent the empirical standard error derived from the simulations of Section B.4.1.

and lines). We see from the figure that in most cases the standard errors obtained from the bootstrap procedure are greater than the empirical standard error and therefore provide conservative estimates of the standard error.

B.4.3 Estimation When m is Misspecified

Section B.4.1 suggests that direct estimation of the entropy rate works extremely well when the order of the Markov chain is specified correctly and the SWLZ estimator is less efficient. Now we consider estimating the entropy rate when the true data-generating process is a second-order Markov chain, but we do not necessarily assume this is known. In this simulation scenario, we compare the performance of four estimators of the entropy rate: direct estimation assuming (incorrectly) a first-order process, direct estimation assuming (correctly) a second-order process, direct estimation assuming a third-order process, and estimation using the SWLZ compression algorithm. This scenario highlights the advantage of the SWLZ method for estimating the entropy rate when the order of the process is unknown.

Assume that the true process is a second-order time-homogeneous Markov process on two states which we label A, B. Then for $\alpha_i, \alpha_j, \alpha_k \in \mathcal{A} = \{A, B\}$, the transition probabilities can be written,

$$\begin{aligned} & \Pr(X_t = \alpha_k | X_{t-1} = \alpha_j, X_{t-2} = \alpha_i) \\ = \; & \Pr(X_t = \alpha_k, X_{t-1} = \alpha_j | X_{t-1} = \alpha_j, X_{t-2} = \alpha_i) \\ = \; & P(\alpha_i\alpha_j, \alpha_j\alpha_k) = P_{ij,jk}. \end{aligned}$$

The transition matrix of this second-order system can be organized as follows,

$$\mathbf{P}_{(2)} = \begin{pmatrix} P(AA, AA) & P(AA, AB) & 0 & 0 \\ 0 & 0 & P(AB, BA) & P(AB, BB) \\ P(BA, AA) & P(BA, AB) & 0 & 0 \\ 0 & 0 & P(BB, BA) & P(BB, BB) \end{pmatrix},$$

where transitions that are not possible, e.g., from AA to BA, are assigned probability zero. Here the notation, $P_{(2)}$, denotes that this is a transition matrix for a second-order Markov chain and will be utilized for clarity. The row vector representing the stationary distribution of the process is

$$\pi_{(2)} = (\pi_2(AA), \pi_2(AB), \pi_2(BA), \pi_2(BB)).$$

Provided the transition matrix is irreducible, the stationary distribution exists and represents the joint distribution of consecutive observations, X_{t-1} and X_t. Now, assume the following parameterization for $\mathbf{P}_{(2)}$,

$$\mathbf{P}_{(2)} = \begin{pmatrix} 1-a & a & 0 & 0 \\ 0 & 0 & b & 1-b \\ 1-c & c & 0 & 0 \\ 0 & 0 & d & 1-d \end{pmatrix}.$$

If $a = c$ and $b = d$ then the process behavior is indistinguishable from that of a first-order Markov process. To see this note that the transition probabilities from AA and BA are identical so that only the most recent state matters. Thus differences between a and c (or b and d) control how much the process depends on the observation which occurred two time steps ago. Using this parameterization, we simulate 1000 second-order Markov chains, each sequence consisting of 1000 observations, and estimate the entropy rate of the process directly assuming orders $m = 1, 2, 3$ and additionally using the SWLZ estimator which makes no assumption on the order of the process. We consider two cases which we label as Case I and Case II: (I) $a = 0.1, c = 0.85, d = 0.2, b = 0.933$ and (II) $a = 0.52, c = 0.22, d = 0.95, b = 0.6833$. The first case highlights an extreme example when the difference between a and c (or b and d) is large and the Markov process depends almost exclusively on the

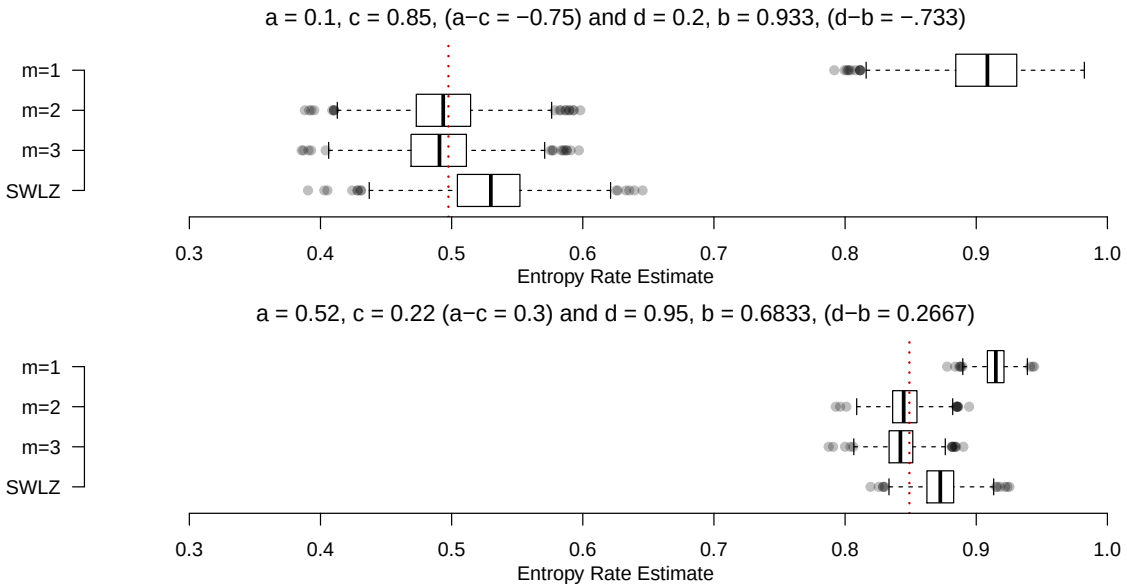

Figure B.4: Simulation results for 1000 two-state Markov chains of length 1000 in two scenarios. Results show distributions of entropy rates using direct estimation with order specified as $m = 1, 2, 3$ and with no order specification in the SWLZ method. The true entropy rate is the vertically dotted line in each figure. The upper figure contains an extreme case where a first-order process provides a poor approximation. The lower figure provides an example where the second-order process is better approximated by a first-order process. We see that misspecification results in a bias in estimating the entropy rate when we choose an m which is too small.

observation which occurred two time steps ago. The difference between the parameters a and c (or b and d) is smaller in the second case. Simulation results are in provided in Figure B.4

The simulations suggest that when the order of the Markov chain specified is too low (i.e., $m = 1$), the estimate obtained will be biased above the true entropy rate. In contrast when we provide the correct order ($m = 2$), or an order which is too large ($m = 3$), we obtain unbiased results. The SWLZ estimator accurately estimates the entropy rate of the process without requiring that the order be specified, but is biased in short sequences (as discussed in Section B.4.1).

These results also suggest that the bias from an m which is too low, increases as the difference

between the values of a and c (or b and d) increases. This relationship can actually be quantified. We can derive the first-order dependence structure that would be observed from monitoring transitions of a second-order process for a long time. This depends on a, b, c, d through the stationary distribution of the second-order process defined by $\mathbf{P}_{(2)}$. An eigendecomposition of $\mathbf{P}_{(2)}^T$ yields, $\pi_{(2)} = \psi^{-1}(d(1-c), da, da, a(1-b))$ where $\psi = a(1-b) + 2da + d(1-c)$. It can then be shown that the first-order dependence structure of the second-order process is,

$$\mathbf{P}_{(1)} = \begin{pmatrix} \frac{1-c}{(1-c)+a} & \frac{a}{(1-c)+a} \\ \frac{d}{d+(1-b)} & \frac{1-b}{d+(1-b)} \end{pmatrix}.$$

If we misspecify the order of the process, then we only observe the entries of the first-order transition matrix. We call this observed transition matrix $\mathbf{P}_{(1)}^*$ and write it as

$$\mathbf{P}_{(1)}^* = \begin{pmatrix} 1-p & p \\ q & 1-q \end{pmatrix}.$$

Then we can relate $\mathbf{P}_{(1)}$ and $\mathbf{P}_{(1)}^*$ by defining $\phi = a - c$ and $\gamma = d - b$, with the result that

$$a = p(1+\phi) \tag{B.4}$$

$$(1-c) = (1-p)(1+\phi) \tag{B.5}$$

$$d = q(1+\gamma) \tag{B.6}$$

$$(1-b) = (1-q)(1+\gamma). \tag{B.7}$$

These equations imply the following restrictions on ϕ and γ: $-1 \leq \phi \leq \min(\frac{p}{1-p}, \frac{1-p}{p})$ and $-1 \leq \gamma \leq \min(\frac{q}{1-q}, \frac{1-q}{q})$. Now we can rewrite $\mathbf{P}_{(2)}$ in terms of the observed entries of the

first-order transition matrix, p, q and the parameters ϕ, γ,

$$\mathbf{P}_{(2)} = \begin{pmatrix} (1+\phi)\left(\frac{1}{1+\phi} - p\right) & p(1+\phi) & 0 & 0 \\ 0 & 0 & (1+\gamma)\left(q - \frac{\gamma}{1+\gamma}\right) & (1-q)(1+\gamma) \\ (1-p)(1+\phi) & (1+\phi)\left(p - \frac{\phi}{1+\phi}\right) & 0 & 0 \\ 0 & 0 & q(1+\gamma) & (1+\gamma)\left(\frac{1}{1+\gamma} - q\right) \end{pmatrix}. \tag{B.8}$$

Both second-order simulations were constructed to have the same first-order behavior with $p = 0.4$ and $q = 0.75$. Equation (B.8) allows an analytical calculation of the entropy rate for any combination of ϕ and γ for fixed p and q. The contour lines of Figure B.5 give the entropy rates of the second order process when $p = 0.4$ and $q = 0.75$ for a grid of ϕ and γ. The max of this plot occurs when $\phi = \gamma = 0$ (a first-order process) and at this point $H(\mathcal{X}) = 0.915$. Additionally marked in the figure are the values of ϕ and γ for the simulations of this section. We see that for a large portion of the figure, the second-order process will have an entropy rate greater than 0.8 and therefore the bias from mistakenly using a first-order model will be small (as seen in the results for Case II given in the bottom panel of Figure B.4).

B.5 Applications

In this section we provide two studies of entropy rate estimation of finite-state Markov chains based upon data from the Conte Center at the University of California, Irvine. The first study in Section B.5.1 investigates the performance of these methods applied to data from Molet et al. (2016). The study aimed to address the impact of experimentally manipulated fragmented and unpredictable behavior (through the use of restricted bedding and nesting

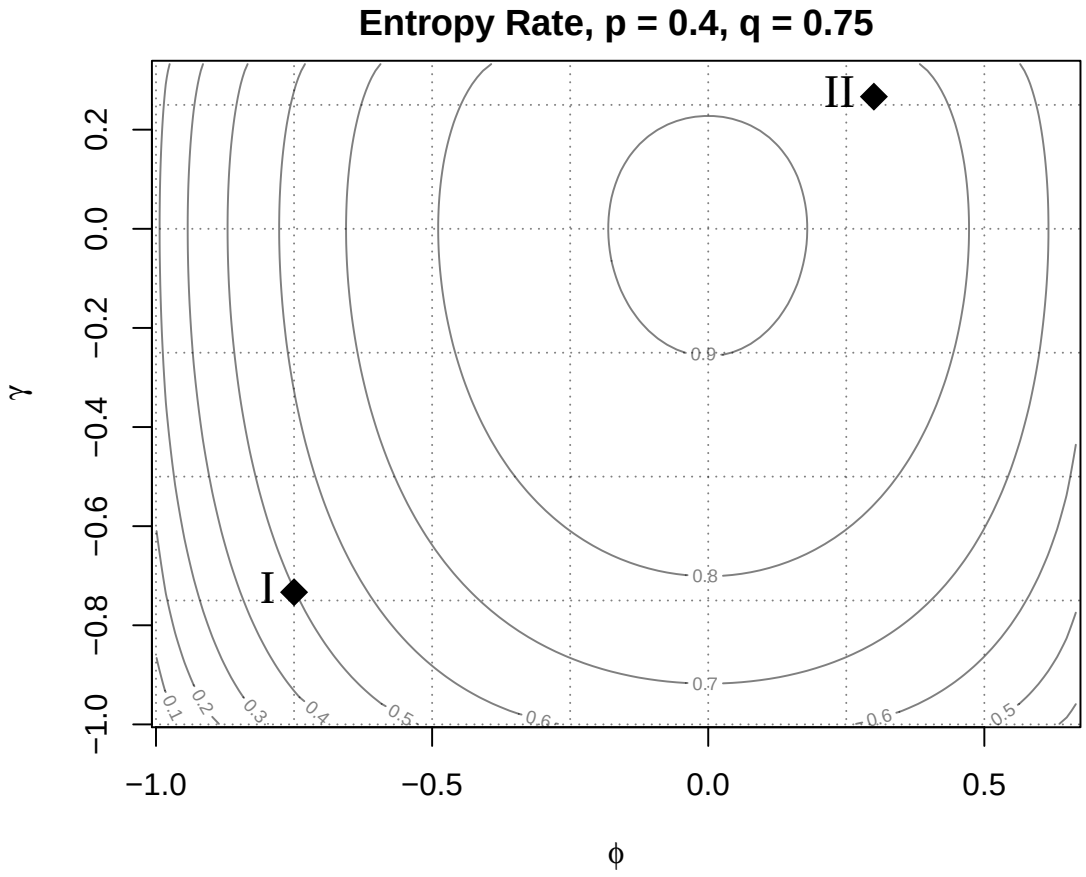

Figure B.5: True entropy rate of different second-order Markov chains which exhibit the same first-order behavior (i.e. $p = 0.4$ and $q = 0.75$). The true entropy rate is provided on a grid of points for valid values of ϕ and γ. The cases used in our simulations are identified on the figure. We find that for even reasonably large values of the tuning parameters, $|\phi| \approx 0.5$ and $-0.75 < \gamma < 0.3$, the difference between the first-order entropy estimate and the second-order entropy estimate will be moderate (as in our Case II).

materials in the cage) of rodent mothers on the emotional and cognitive outcomes of their offspring. Section B.5.2, investigates entropy rate estimation in an observational study of human behavior based upon data from Davis et al. (2017). Similar to the rodent study, the goal was to assess the impact of unpredictable maternal behavior on the cognitive outcomes of children later in life.

B.5.1 Measuring Predictability of Behavior in a Rodent Study

This section provides an application of entropy rate estimation using data that was introduced in Molet et al. (2016). The study addressed the impact of fragmented and unpredictable behavior of rodent mothers on the emotional and cognitive outcomes of their offspring. On postnatal day 2 (P2), rodent pups were randomly assigned to one of two types of rearing environments for 8 days; a normal environment or an impoverished environment (limited bedding and nesting materials). The dams (mothers) were observed twice per day for 50 minutes for eight days and sequences of their behavior were recorded. Behavior was described using a set of $\kappa = 7$ distinct actions: licking/grooming pups, carrying pups, nursing, nest building, off pups, eating, or self-grooming. On the tenth day (P10), all of the rodent pups and mothers were returned to normal environmental conditions.

The limited bedding and nesting environment led to more erratic maternal behavior and a goal was to quantify the impact of this treatment (i.e., the unpredictability of maternal care due to an adverse environment). Entropy rate was chosen as the measure which could most succinctly summarize the behavior of each dam, rather than focusing on any specific action that the rodents perform. In Molet et al. (2016), the sequence of observations for each rodent was treated as a stationary first-order Markov chain. The assumption of stationarity allowed the concatenation of the sequences from each 50 minute window together as one long Markov process. Because the analysis was focused on the predictability of the *patterns* of

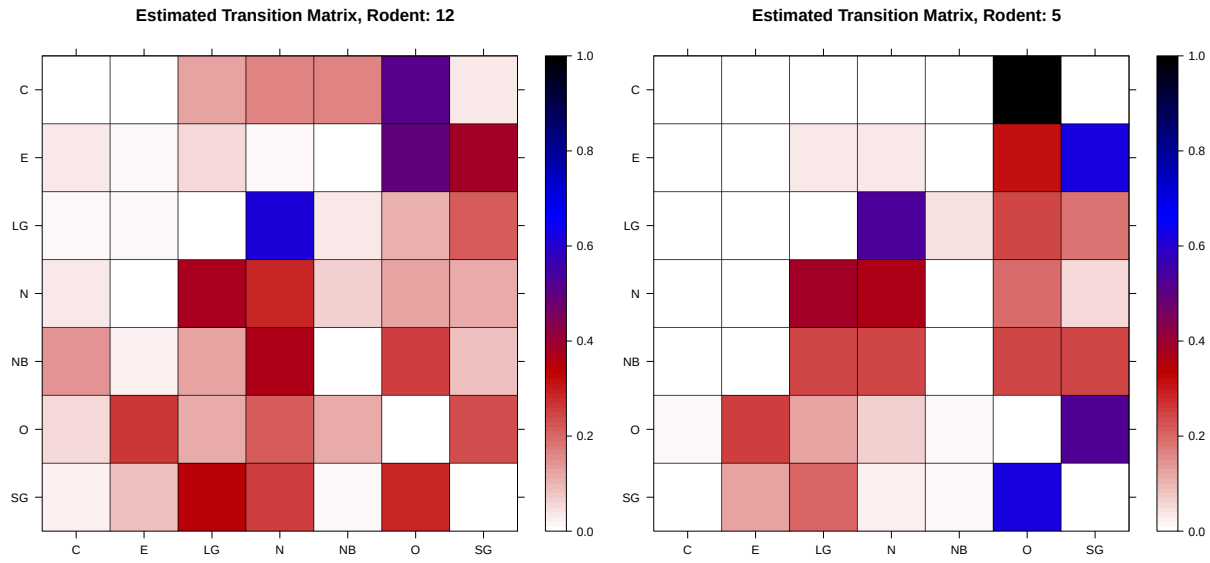

Figure B.6: Examples of observed empirical transition matrices from the behavioral application. Left: example limited bedding & nesting rodent, Right: example control rodent

actions and not the duration of actions, the sequence was treated as a discrete-time Markov chain focusing only on transitions between different maternal care behaviors. Under these assumptions, estimates of entropy rate were calculated for each rodent mother. Note that there were a total of seven possible actions which corresponds to a maximum possible entropy of $H_{\max}(\mathcal{X}) = \log_2(7) = 2.807$.

In this paper, we recreate these analyses using the estimators that were outlined in the previous sections. We obtain results that are consistent with those obtained in the original paper. Figure B.6 provides one empirical transition matrix from the limited bedding and nesting maternal group and one from the control group respectively, as an example of the data in the study.

To summarize the level of predictability, Table B.3 provides entropy rate estimates based upon treating the rodent behavior as a first-order Markov chain, a second-order chain, and by using the sliding-window Lempel-Ziv estimator not assuming any specific order for the behavior. We see that the minimum number of transitions performed by any rodent was approximately

Table B.3: Rodent Data Estimates

ID	Group	N	$\hat{H}_{SWLZ}(\mathcal{X})$	$\hat{H}^{m=1}_{emp}(\mathcal{X})$	$\hat{H}^{m=1}_{eig}(\mathcal{X})$	$\hat{H}^{m=2}_{emp}(\mathcal{X})$	$\hat{H}^{m=2}_{eig}(\mathcal{X})$
3	LBN	511	1.6956	1.8837	1.8831	1.5069	1.5060
4	LBN	404	1.6285	1.8015	1.8009	1.3307	1.3298
7	LBN	688	1.6797	1.8774	1.8770	1.4988	1.4986
8	LBN	699	1.6807	1.8403	1.8399	1.5168	1.5150
11	LBN	632	1.7916	1.9342	1.9338	1.6192	1.6171
12	LBN	886	1.8526	2.0515	2.0511	1.7462	1.7462
Group Mean			1.7215	1.8981	1.8976	1.5364	1.5354

ID	Group	N	$\hat{H}_{SWLZ}(\mathcal{X})$	$\hat{H}^{m=1}_{emp}(\mathcal{X})$	$\hat{H}^{m=1}_{eig}(\mathcal{X})$	$\hat{H}^{m=2}_{emp}(\mathcal{X})$	$\hat{H}^{m=2}_{eig}(\mathcal{X})$
1	Control	433	1.5483	1.7393	1.7395	1.3632	1.3627
2	Control	340	1.5107	1.5322	1.5319	1.2044	1.2042
5	Control	290	1.5727	1.6256	1.6242	1.2621	1.2596
6	Control	378	1.6571	1.7427	1.7417	1.3900	1.3878
9	Control	300	1.7552	1.8164	1.8155	1.3637	1.3536
10	Control	403	1.7864	1.8590	1.8591	1.4391	1.4326
Group Mean			1.6384	1.7192	1.7187	1.3371	1.3334

t-test Results - Difference in Means with Equal Variance					
Test Statistic	-1.4425	-2.9308	-2.9287	-2.986	-3.0428

300 which, by our simulation study, ensures that the estimates will be adequate for the task at hand. The series is not long enough to consider a third-order process. Molet et al. (2016) performed a t-test on the difference in mean entropy rates of the two groups which showed that experimentally introducing a stressed environment for rodent mothers leads to less predictable maternal care.

Table B.3 shows that we obtain results consistent with those of Molet et al. (2016) for first-order Markov chains. Additionally, we see that SWLZ estimation provides estimates which fall between the estimates obtained using first-order and second-order assumptions. This may imply that the behavior of the rodents is more complex than the simplifying assumption of first-order behavior and that the rodent behavior is better represented by a second-order process. The correlation between the first-order and second-order entropy rates based on the

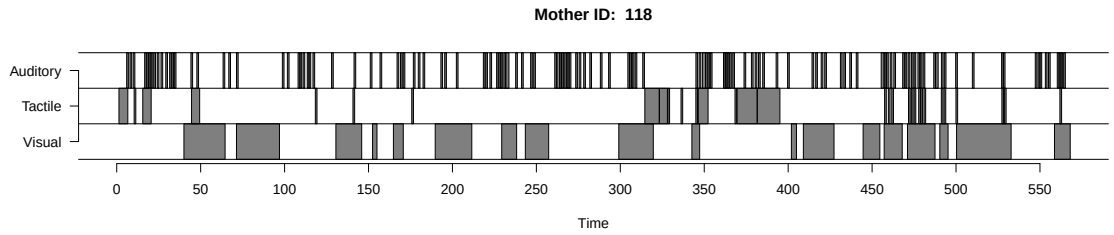

Figure B.7: Periods of auditory, visual, and tactile stimulation that were provided to a child by its mother during a ten-minute play interaction.

empirically estimated transition matrix and stationary distribution is approximately 0.937.

B.5.2 Measuring Predictability of Behavior in a Human Developmental Study

We provide a second application of entropy rate estimation using data that were introduced in Davis et al. (2017). The study focused on quantifying the association between the level of maternal predictability, as demonstrated though a 10-minute play session involving a mother and her 12-month-old child, and the cognitive function of that child one year later. Each mother-child pair was given a set of toys and the mother was instructed to play with her child. This interaction was digitally recorded and the interactions between the mother and the child were then coded by researchers to document the time and duration of specific actions or instances (e.g, mother touching the child, mother providing instructional speech to the child, etc.). These actions were then used to identify intervals during which the mother was providing auditory, tactile, or visual stimulation to the child (see the Supplemental information of Davis et al. (2017) for a more detailed description). A visualization of the intervals for each type of sensory input for one mother is provided in Figure B.7. We note that similar data might be gathered in other settings, e.g., during interactions between a teacher and a student.

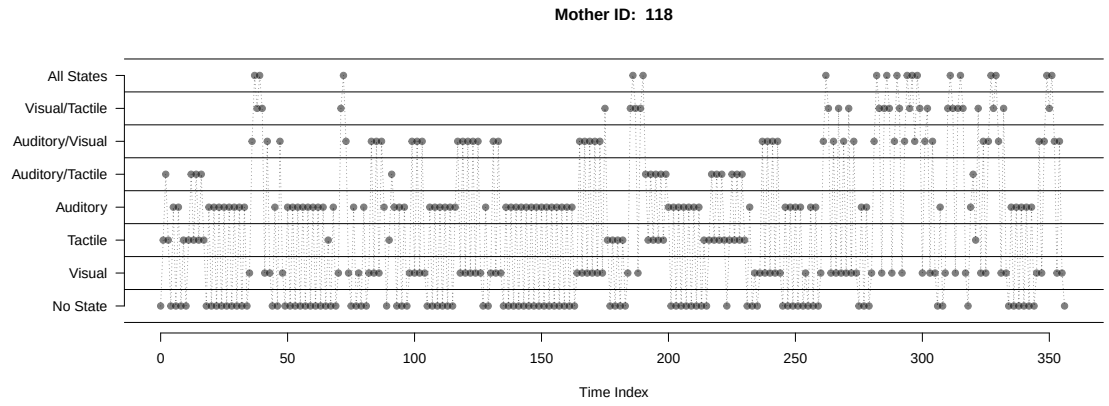

Figure B.8: Transformed time-series. In this sequence, each pair of states is mutually exclusive, so that the sequence can be modeled as a finite-state Markov chain.

The data as illustrated in Figure B.7 is not immediately amenable to being described as a finite-state Markov process, since at each time point the mother's behavior can reflect any combination of the sensory domains. Therefore, we considered each grouping of unique sensory behaviors as a discrete state (auditory, visual, tactile, auditory & tactile, auditory & visual, visual & tactile, or all domains at once), and transformed the data so that each state is now mutually-exclusive from all others. With this transformation the data shown in Figure B.8 could now be modeled using a finite-state Markov chain. We do not consider the duration of a given state, only the transitions among states.

Davis et al. (2017) modeled the data as a first-order time-homogeneous finite-state stationary Markov chain and estimated the entropy rate of the process using the methods of Section B.3.1. These measurements were then used to assess the association between maternal predictability (entropy rate) and Mental Development Index (MDI) scores of the children one year after the original interaction. It was found in the paper that higher levels of maternal predictability (lower entropy rate) at one year of age were associated with higher MDI scores indicating better cognitive performance at two years of age and that the association persisted after controlling for other variables known to affect mental development.

Here, we recreate the first-order Markov chain estimates of the entropy rate using 171 mother-child pairs from the original study, but further consider estimating the entropy rate of the sequences using the SWLZ estimator, as well as higher-order Markov chains (i.e., $m = 2$ and $m = 3$). These results are contained in Figure B.9 and summary statistics are provided in Table B.4. The left panel of the figure contains the distribution of the entropy rate estimates using each of the methods outlined previously (the stationary distributions of the Markov chains were estimated using the observed empirical distribution) and the lines show the change in each individual's estimate across methods. This figure and table highlight that the SWLZ estimates are higher than the first-order entropy rate estimates. From the simulations study in Section B.4.1, it was demonstrated that the SWLZ is biased high when compared with the true entropy rate (except in high entropy rate cases), suggesting that the first-order assumption may be appropriate in this context. The right panel demonstrates the Spearman rank correlation of the entropy rate estimates across methods. The high correlations suggest that the association between predictability and cognitive outcomes should not change if the estimation method or model assumptions are changed.

Table B.4: Summary statistics of entropy rate estimates across sequences from Davis et al. (2017). From these statistics, the SWLZ estimates consistently produce the highest entropy rate estimates across methods.

Estimation Method	Mean	Median	IQR
SWLZ	0.910	0.885	(0.751, 1.071)
$m = 1$	0.771	0.756	(0.659, 0.902)
$m = 2$	0.700	0.693	(0.602, 0.812)
$m = 3$	0.626	0.625	(0.541, 0.722)

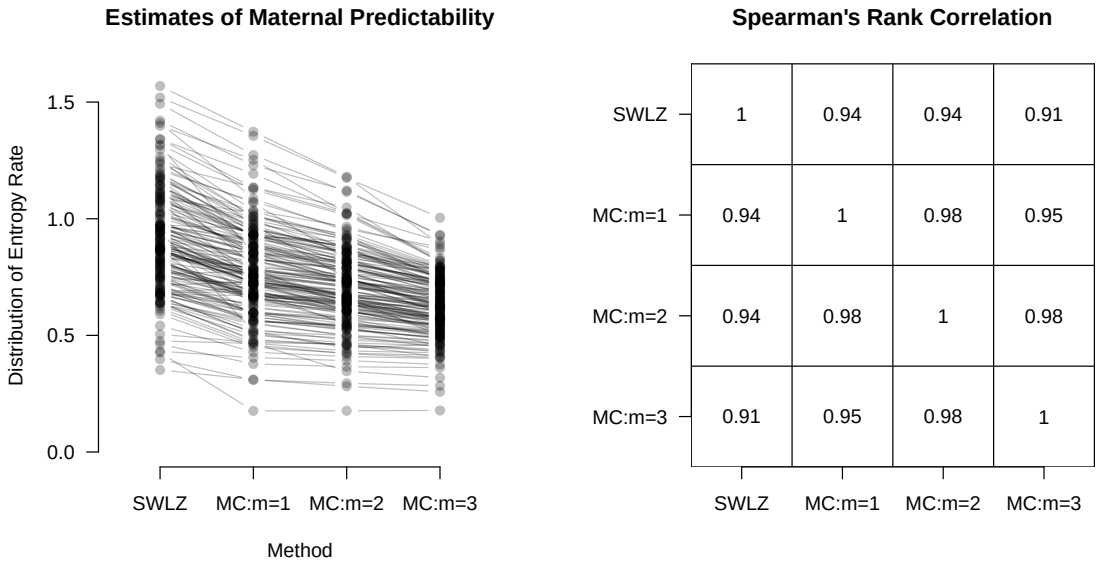

Figure B.9: Comparison of estimates of entropy rate for maternal predictability. The first figure contains the distribution of the entropy rate estimates using each method (i.e. SWLZ, Markov chain $m = 1$, Markov chain $m = 3$, Markov chain $m = 3$) with lines connecting each individual's estimate using each method. The table on the right gives the Spearman rank correlations of the entropy rate estiamtes provided by the methods.

B.6 Discussion

Behavior can be modeled as a sequence of actions performed by an individual over a given time period. The entropy rate of this sequence is a summary measure that describes the degree to which we can predict actions of the process. In this appendix we presented three different entropy rate estimators and assessed their performance as a function of the length of the observed sequences. We presented the stationary bootstrap as an approach for obtaining standard error estimates, provided a method for choosing the parameter of the stationary bootstrap, and assessed the performance of the stationary bootstrap. Additionally, it was demonstrated that entropy rate provides a measurement that quantifies predictability within behavioral studies and that such measurements can help to understand behavior within developmental and educational settings.

Estimating the entropy rate by direct estimation of the transition matrix and stationary dis-

tribution of a Markov process achieves asymptotically unbiased estimates of the true entropy rate when the order of the Markov process is known. This approach provides several summary measures of behavior. We obtain a description of the probability distribution governing transitions between states (or actions), the long term expected proportion of occurrences that each action is performed, and a measure of the predictability of actions. There are two major concerns associated with using direct methods to estimate the entropy rate. One concern is that we must observe $n \gg \kappa^m$ observations to achieve precise estimates of the transition matrix for an m^{th} order Markov process. It is often challenging to observe a sequence of behaviors long enough to achieve the required precision when m is 3 or more. A more serious concern is the requirement that the order of the Markov process is assumed known. There is always the possibility that the true data-generating process does not match the assumed order. If the true order is larger than the assumed order, then one may get misleading results.

The sliding-window Lempel-Ziv (SWLZ) estimator avoids the assumption of a specific order for the Markov chain. It is based on a data compression algorithm that only assumes that the stochastic process is both stationary and ergodic. It appears from the simulation study that estimates based upon this strategy can be slightly biased, particularly in short sequences. Section B.4.3 demonstrates the advantage of the approach in that the SWLZ method accurately estimates the true entropy rate for higher-order processes without requiring that the order be specified in advance.

The two estimation techniques may be used in concert to fully understand the predictability of an observed stochastic system. They both provide valuable information regarding predictability. The SWLZ method can be used to provide an initial estimate of the entropy rate which is assured of being close to the true entropy rate of the system and therefore may be used as a guidepost for choosing the order m of the process. Once we have an estimate of the entropy rate, we can directly estimate the entropy rate for $m = 1, 2, \ldots$ provided we

have observed enough data until the entropy rate estimate is approximately equal to that obtained from the SWLZ technique. This would allow us to further understand the process through its transition matrix and stationary distribution. Additionally, in our experience with behavioral studies, the SWLZ estimates are often highly correlated with the estimates based on a first-order Markov chain assumption and thus a simple first-order assumption may often provide useful information about the predictability of a sequence.

All of the methods applied here rely on the assumption that the process is stationary. Processes describing human behavior, like those described in our application, may not be stationary since the predictability of the individual may be context dependent. For example, an individual may be predictable when in a familiar environment, but the individual may be unpredictable if moved into a new environment. This implies that when we intend to use entropy rate to define behavioral characteristics, we should restrict our observations to specific windows of time that are long enough to estimate the entropy rate, but short enough for the assumption of stationarity to be plausible.