Dimension Reduction and Representation Learning for Improved Estimation of Causal Parameters

Sungtaek Son

A dissertation
submitted in partial fulfillment of the
requirements for the degree of
Doctor of Philosophy

University of Washington
2026

Reading Committee:
Kwun Chuen Gary Chan, Co-Chair
Eardi Lila, Co-Chair
Ting Ye

Program Authorized to Offer Degree:
Department of Biostatistics

University of Washington

Abstract

Dimension Reduction and Representation Learning
for Improved Estimation of Causal Parameters

Sungtaek Son

Co-Chairs of the Supervisory Committee:
Kwun Chuen Gary Chan
Department of Biostatistics

Eardi Lila
Department of Biostatistics

High-dimensional observational data, including electronic health records and omics measurements, create substantial challenges for causal inference. The growing complexity of such data can hinder both estimation and interpretation. This dissertation focuses on heterogeneous treatment effects, individualized treatment regimes, and mediation analysis, and develops dimension-reduction and representation-learning methods tailored to their causal targets to facilitate inference.

First, we propose a sufficient dimension-reduction method that directly targets treatment effect heterogeneity by identifying a low-dimensional linear subspace of the covariates. Combined with kernel-based covariate-balancing, this representation facilitates the estimation of optimal individualized treatment regimes through outcome-weighted learning in observational settings. Second, we develop an envelope-based dimension-reduction method for causal mediation analysis with high-dimensional mediators. The method identifies media-

tor variation associated with both the exposure and the outcome while separating variation unrelated to one or both components of the mediation pathway.

Finally, we develop a nonlinear representation-learning method for conditional average treatment effect estimation. The proposed neural-network architecture promotes information sharing among the nuisance functions and the conditional average treatment effect function through shared latent representations. The proposed framework can accommodate complex structured and unstructured covariates, such as images.

Together, these methods provide a framework for extracting causally relevant representations from complex, high-dimensional data, with the goal of improving the accuracy, interpretability, and applicability of causal inference methods in complex data settings.

Contents

Abstract **iii**

Acknowledgments **xvi**

1 Introduction **1**

2 Sufficient Dimension Reduction for CATE and ITR Estimation **5**

2.1 Introduction 5

2.2 Preliminaries 9

2.3 Method 11

2.3.1 Gradient kernel dimension reduction for the difference between potential outcomes 12

2.3.2 Kernel-based covariate balancing 15

2.3.3 Estimation of optimal ITRs 16

2.4 Theoretical guarantees 17

2.5 Simulations 20

2.5.1 Simulation settings 21

2.5.2 Results 22

2.6 Application 25

2.7 Conclusion 26

3 Envelope Models for High-Dimensional Mediation Analysis **28**

3.1 Introduction 28

3.2 Preliminaries 31

3.2.1 Review of envelope model 32

3.3 Methods . . . 33
3.3.1 Estimation for span(G), span(Γ_2), and span(Φ_2) . . . 35
3.3.2 Asymptotic properties . . . 39
3.3.3 Selecting dimensionality u . . . 42
3.4 Simulations . . . 43
3.4.1 Setting 1 . . . 45
3.4.2 Setting 2 . . . 45
3.4.3 Setting 3 . . . 45
3.4.4 Results . . . 46
3.5 Application . . . 47
3.6 Conclusion . . . 51

4 Representation Learning for Heterogeneous Treatment Effect Estimation Using Multimodal Data 52

4.1 Introduction . . . 52
4.2 Related work . . . 55
4.2.1 One-step plug-in approaches . . . 55
4.2.2 Pseudo-outcome regression approaches . . . 56
4.2.3 R-learner . . . 56
4.2.4 Multimodal causal inference . . . 57
4.3 Method . . . 58
4.3.1 Estimation with multimodality . . . 59
4.4 Experiments . . . 62
4.4.1 Tabular data - IHDP benchmark . . . 62
4.4.2 Multimodal data - ChestMNIST . . . 63
4.4.3 Settings . . . 65
4.4.4 Results . . . 66
4.5 Conclusion . . . 67

5 Discussion 69

References 71

A Chapter 3 Appendix **82**

A.1 Proof for Theorem 2.4.1 82

A.1.1 Step 1: Characterization of the dual problem 83

A.1.2 Step 2: Limiting form of the dual problem 87

A.1.3 Step 3: Identification of the convergence in probability 87

A.2 Convergence of $g_{\hat{w}}(x)$ to $g(x)$ 95

A.2.1 Convergence of $g_n(x)$ to $g(x)$ 95

A.2.2 Proof of Corollary 2.4.2 96

A.3 Proof of Theorem 2.4.3 97

A.4 Proof of Theorem 2.4.4 98

A.5 Proof of Theorem 2.4.5 100

A.6 Proof of Lemma 2.4.7 103

A.7 Proof of Proposition 1 107

A.7.1 Expressive power of an RKHS from a universal kernel 107

A.7.2 Excess risk bound 107

A.7.3 Proof of proposition 109

A.8 Equivalence of the risk and its alternative form 110

A.9 Simulation details 115

A.9.1 Performance evaluation 115

A.9.2 Estimation of ITR 115

A.9.3 Simulation settings 116

A.9.4 Setting 1 116

A.9.5 Setting 2 116

A.9.6 Setting 3 117

A.10 Additional simulation results 117

A.10.1 Results for $n = 500$ 117

A.10.2 Projection error 118

A.10.3 Computation time 121

A.11 Real world data 121

A.11.1 Estimation of ITR and evaluation of modified value function 123

B Chapter 4 Appendix **124**

B.1 Proof of Proposition 3.3.1 124

B.2 Proof of Proposition 3.3.3 124

B.3 Proof of Proposition 3.3.4 125

B.4 Proof of Proposition 3.3.6 125

B.5 Proof of Proposition 3.3.7 126

B.6 Proof of Proposition 3.3.9 127

B.6.1 Formulation of likelihood to discrepancy function 127

B.6.2 Proof of Lemma 3.3.8 129

B.6.3 Proof of Proposition 3.3.9 132

B.7 Expression for $E[Y(a, M(a'))]$ in reduced subspace 133

B.8 Derivation of the objective function 134

B.8.1 Response envelope 135

B.8.2 Predictor envelope 138

B.8.3 Derivation of the objective function 144

B.8.4 Equivalence to joint likelihood maximization 145

B.8.5 Total number of parameters 147

B.9 Additional simulation results 147

B.9.1 Results when $n = 200,\ 500,\ 1000$ 147

B.9.2 Hard threshold selection of u 150

B.9.3 Projection accuracy 150

B.10 Real data analysis 151

C Chapter 5 Appendix **156**

C.1 Proof of Proposition 4.3.1 156

List of Figures

2.1 Blue points correspond to the treatment group "+1" and red points correspond to the treatment group "-1". The sample sizes for the +1 and -1 treatment groups are 110 and 90, respectively. (a) shows the distribution of the data with respect to the covariates $\left(X^{(1)}, X^{(2)}\right)$. The magnitudes of the outcomes are represented by varying point sizes. The solid line represents the true projection line that best captures treatment effect heterogeneity, while the dashed line denotes the estimated projection line obtained from our proposed SDR approach. (b) displays the true pseudo-outcomes ($Y/\Pr(+1|X)$ and $-Y/\Pr(-1|X)$ for treatment groups $A = +1$ and $A = -1$) in the reduced subspace, where $\left(X^{(1)}, X^{(2)}\right)$ is projected onto $V = B_0^\top X$. The curve represents the heterogeneous treatment effect as a function of V. (c) shows the pseudo-outcomes projected onto $V^\perp$, the space perpendicular to that spanned by B_0. The solid line shows the treatment effect as a function of $V^\perp$. (d) shows the pseudo-outcomes as a function of $\hat{V} = \hat{B}^\top X$, where $\hat{B}$ is an estimate of B_0 obtained using the proposed SDR framework. The solid line shows the true heterogeneous treatment effect along the different levels of the dimension-reduced covariate. The dashed line shows the estimated decision rule along the different levels of the dimension-reduced covariate. The signs of these functions, which determine the true and estimated treatment assignments, are equal in the range $\hat{V} \in (-1, 1)$, where approximately 95% of the data points lie. 10

2.2 Scatterplot of the Bayes optimal treatment regimes for different values of the covariates generated under Setting 1. The red circles and blue crosses represent data points from the test dataset, corresponding to $d^*(X) = -1$ (or $d^*(V_0) = -1$) and $d^*(X) = +1$ (or $d^*(V_0) = +1$), respectively. The solid lines in the left plot display the decision boundaries in the dimension-reduced space, defined as $\left\{v = \left(v^{(1)}, v^{(2)}\right) : \tilde{f}^*(v) = 0\right\}$. 22

2.3 Simulation results for n = 1,000 from the randomized scenarios across the three different settings. Accuracy represents the proportion of correctly predicted optimal treatments. Value (%) represents the value function recovered by the estimated decision rule as a percentage of the Bayes optimal value function. The performance of the proposed method is shown under DOL-O and DOL-L: DOL-O uses the oracle $g(x)$, while DOL-L uses $g_{\hat{w}}(x)$ estimated via linear regression. MAVE-L uses interaction MAVE for dimension reduction and uses the fitted values $g_n(x)$ obtained via linear regression. 23

2.4 Simulation results for n = 1,000 from the non-randomized scenarios across the three different settings. Accuracy represents the proportion of correctly predicted optimal treatments. Value (%) represents the value function recovered by the estimated decision rule as a percentage of the Bayes optimal value function. The performance of the proposed method is shown under KCB-O and KCB-L: KCB-O uses the oracle $g(x)$, while KCB-L uses the fitted values of $g_{\hat{w}}(x)$ obtained via linear regression. MAVE-L uses interaction MAVE for dimension reduction and uses KCB weights along with the fitted values of $g_{\hat{w}}(x)$ obtained via linear regression. IPW-O uses the oracle $g(x)$, and IPW-L uses $g_{\hat{w}}(x)$ from linear regression, with propensity scores estimated via logistic regression. XGB-L uses $g_{\hat{w}}(x)$ from linear regression, with propensity scores estimated via XGBoost. NN-L uses $g_{\hat{w}}(x)$ from linear regression, with propensity scores estimated via a neural network. 24

2.5 Modified value functions over the 100 repeats. DOL represents the proposed method. AOL used the raw patient characteristics for learning ITR and fitting logistic regression to estimate IPW. QL is the ℓ_1-PLS. "Treat all" represents the average effect of TTE. P-values from paired t-tests are displayed using asterisks: **** < 0.0001; *** < 0.001; * < 0.05. 26

3.1 Causal DAGs for the association between exposure (A), outcome (Y), and the target transformation of mediator (M). The linear transformation, $G^\top M$, retains the mediation effect, and is distinguished from the immaterial projections: $\Gamma_2^\top M$, $\Phi_2^\top M$, and $D_0^\top M$. 30

3.2 An example of singular values of $\hat{\Gamma}^\top \hat{\Phi}$. Using the threshold 0.01 for the relative difference from the largest singular value chooses $u = 2$. 44

3.3 Mediator correlations for the three different settings. The left plot corresponds to Setting 1; the middle plot corresponds to Setting 2; the right plot corresponds to Setting 3. 44

3.4 The distributions of the estimated natural indirect effects from different methods. The left plot corresponds to Setting 1; the middle plot corresponds to Setting 2; the right plot corresponds to Setting 3. 46

3.5 Component-wise NIEs standardized by each corresponding mediator's standard deviation. The largest 20 standardized NIEs from each method are presented. The proteins are specified by the UniProt identifier. From the left: the proposed method with $u = 3$, OLS, classic envelope, and the proposed method with $u = 58$. 50

4.1 An illustration of the misalignment issue is shown here. The red squares indicate different clusters of four pixels. Pixels at the same coordinates across images may correspond to different anatomical structures. For example, the rightmost pixel corresponds to the shoulder in one individual (upper-left image) but to part of the lung in another (upper-right image). 53

4.2 Neural network architectures for JointRNet and RNet for estimating treatment effect heterogeneity. Each architecture admits the raw image features (X_I) and the structured covariates (X_S) as input. 60

4.3 Neural network architectures for the one-step plug-in methods. Each architecture admits the raw image features (X_I) and the structured covariates (X_S) as input. 61
4.4 Performance comparison between different learners. The distributions of the RMSE evaluated on the test set are displayed. Results from T-learner and PW- and DR-learner extensions of TARNet, DragonNet, and DR-CFR are omitted due to large errors. 63
4.5 Simulation results for different methods. JointRNet is the proposed single-step neural network approach for R-learners. The RMSEs of the estimated CATE functions over the 100 iterations are presented for balanced and imbalanced treatment assignment scenarios within Setting 1 and Setting 2. 66
4.6 MSE along increasing sample size. 67
4.7 Comparison of nuisance functions estimates between the JointRNet and the RNet. The RMSEs of the estimated $m(x)$ and $\pi(x)$ over the 100 iterations are presented for balanced and imbalanced treatment assignment scenarios within Setting 1 and Setting 2. 68

A.1 The distribution of the optimal treatment regimes according to different levels of covariates in Setting 2. The red circles and blue crosses are data points from the test dataset and represent $d^*(X) = -1$ (or $d^*(V_0) = -1$) and $d^*(X) = +1$ (or $d^*(V_0) = +1$), respectively. The solid lines on the left plot display the decision boundaries in the dimension-reduced space $\left\{v_0 = \left(v_0^{(1)}, v_0^{(2)}\right); \tilde{f}^*(v_0) = 0\right\}$. 110
A.2 Simulation results for $n = 500$ from the randomized scenarios across the three different settings. Accuracy represents the proportion of correctly predicted optimal treatments. Value (%) represents the value function recovered by the estimated decision rule as a percentage of the Bayes optimal value function. The performance of the proposed method is shown under DOL-O and DOL-L: DOL-O uses the oracle $g(x)$, while DOL uses $g_{\hat{w}}(x)$ estimated via linear regression. MAVE-L uses interaction MAVE for dimension reduction and uses the fitted values $g_n(x)$ obtained via linear regression. 118

A.3 Simulation results for $n = 500$ from the non-randomized scenarios across the three different settings. Accuracy represents the proportion of correctly predicted optimal treatments. Value (%) represents the value function recovered by the estimated decision rule as a percentage of the Bayes optimal value function. The performance of the proposed method is shown under KCB-O and KCB-L: KCB-O uses the oracle $g(x)$, while KCB-L uses the fitted values of $g_{\hat{w}}(x)$ obtained via linear regression. K-MAVE-L uses interaction MAVE for dimension reduction and uses KCB weights along with the fitted values of $g_{\hat{w}}(x)$ obtained via linear regression. IPW-O uses the oracle $g(x)$, and IPW-L uses $g_{\hat{w}}(x)$ from linear regression, with propensity scores estimated via logistic regression. XGB-L uses $g_{\hat{w}}(x)$ from linear regression, with propensity scores estimated via XGBoost. NN-L uses $g_{\hat{w}}(x)$ from linear regression, with propensity scores estimated via a neural network. 119

A.4 Simulation results that display the change in the error in the projection from $\hat{B}$ along different sample sizes represented by $\|\hat{B}\hat{B}^\top - B_0 B_0^\top\|_F$ (y-axis). From left to right, the figures correspond to the non-randomized treatment assignment scenarios of Settings 1, 2, and 3. The means and standard deviations are depicted at each sample size. 120

A.5 Comparison of projection errors of $\hat{B}$ ($\|\hat{B}\hat{B}^\top - B_0 B_0^\top\|_F$) between the proposed KCB, XGBoost (Matthew Cefalu et al. 2024), and neural network (Shang et al. 2025). From left to right, the graphs correspond to the non-randomized treatment assignment scenarios of Settings 1, 2, and 3. The means and standard deviations are depicted at each sample size. 120

A.6 Computation time (seconds) ratios for dimension reduction measured as the runtime of gKDR divided by the runtime of iMAVE, where the iMAVE runtime corresponds to a single iteration (top) or five or less iterations (bottom). Within each row, the left plot shows results for $n = 500$, and the right plot shows results for $n = 1000$. The horizontal dashed line indicates a ratio of 1. 122

B.1 The distributions of the estimated natural indirect effects when $n = 200$ from different methods. The left plot corresponds to Setting 1; the middle plot corresponds to Setting 2; the right plot corresponds to Setting 3. 148
B.2 The distributions of the estimated natural indirect effects when $n = 500$ from different methods. The left plot corresponds to Setting 1; the middle plot corresponds to Setting 2; the right plot corresponds to Setting 3. 148
B.3 The distributions of the estimated natural indirect effects from when $n = 1000$ different methods. The left plot corresponds to Setting 1; the middle plot corresponds to Setting 2; the right plot corresponds to Setting 3. 149
B.4 The distributions of the estimated NIEs obtained from different dimension selection approaches. PIE and SVD represent estimates from $\hat{u}$ that used eigengap ratio criterion. PIE0.99 and SVD0.99 represent the estimates from selecting $\hat{u}$ that correspond to the number of singular values larger than 0.99. PIE0.7 and SVD0.7 represent the estimates from selecting $\hat{u}$ that correspond to the number of singular values larger than 0.7. From the top to the bottom row, the results used $n = 100$, 200, 500, and 1,000. 149
B.5 Projection errors represented by Frobenious norms ($\|P_{\hat{G}} - P_G\|_F$) of the difference between $P_{\hat{G}}$ and P_G. The y-axis represents the mean log Frobenious error for the three simulation settings – Setting 1, 2, and 3 from left to right. The true dimensions of Γ, Φ, and G were used for the simulation. 150
B.6 Heatmap of the correlation structure of the 275 proteins relevant for inflammation and CVD. 152
B.7 Flow chart of sample selection 152
B.8 Singular values for choosing dimensionality of the reduced subspace with the subset of biomarkers. The top three singular values of $\hat{\Gamma}^\top \hat{\Phi}$ are near 1. . . . 153
B.9 The distributions of the component-wise estimates for the outcome-mediator (left) and mediator-exposure effects (right). PIE displays results for $u = 3$. . 155

List of Tables

1.1 Overview of the dissertation. 4

4.1 Summary of one-step plug-in learners for CATE estimation. The one-step plug-in estimators and their corresponding estimated quantities are presented. 56

4.2 Summary of pseudo-outcome regression learners for CATE estimation. The pseudo-outcome regression learners and their associated pseudo-outcomes are presented. The two-step extensions indicate which plug-in estimators from Table 4.1 can be used in the first stage to construct the corresponding pseudo-outcomes. 57

B.1 UniProt identifiers and corresponding protein names. 153

Acknowledgments

I would like to express my sincere gratitude to everyone who supported me throughout my five-year doctoral journey in the Department of Biostatistics.

I am especially grateful to my advisors and committee members, Dr. Kwun Chuen Gary Chan, Dr. Eardi Lila, and Dr. Ting Ye. Their insight and guidance were essential to my progress through the program and profoundly shaped my development as a researcher. Gary taught me to approach research with greater rigor, and his mentorship helped me develop a deeper and more concrete understanding of causal inference. Eardi introduced me to more complex and fascinating statistical problems, as well as the methodological tools needed to address them. His mentorship also greatly strengthened my academic communication skills. Ting provided me with a strong foundation in causal mediation analysis and offered essential insights that shaped the second project of this dissertation. I would also like to thank Dr. Abraham Flaxman for serving as the Graduate School Representative on my committee.

I am also grateful to my research supervisors for their support. Dr. Jennifer Bobb provided invaluable guidance during my first two years in the program and introduced me to important statistical challenges arising in pragmatic clinical trials. Dr. Kevin Lin introduced me to interesting research questions and ideas related to sensitivity analysis and the analysis of genetic sequencing data.

Finally, my family has been an indispensable source of support throughout this journey. My wife, Songyee, has taken every step of this journey alongside me and supported me in countless ways. I am deeply grateful to my parents-in-law, whose generous support gave me the time and energy needed to pursue my research. I am equally grateful to my parents, whose unwavering encouragement and support were vital to sustaining my life and studies in Seattle. I could not have completed this journey without them.

Chapter 1

Introduction

Modern causal inference increasingly confronts the challenges of high-dimensional data. This is particularly relevant when analyses rely on observational data sources, such as electronic health records (EHRs), which often contain a large number of variables with complex dependence structure. This challenge has been further intensified by advances in data collection technologies. For example, modern omics technologies generate complex, high-dimensional measurements, creating a need for causal inference methods capable of effectively extracting information relevant to the estimand of interest while controlling for confounding and other sources of bias.

Drawing causal conclusions from observational data is of great scientific and clinical interest, as these data reflect real-world patient populations. They also enable researchers to investigate the causal questions that are often infeasible or unethical to address through randomized controlled trials. One important causal inference problem is the estimation of the conditional average treatment effect (CATE), which characterizes how treatment effects vary with an individual's characteristics. Although sophisticated methods for CATE estimation have been developed, opportunities remain to improve their performance and broaden their applicability to increasingly complex data structures. Closely related to CATE is the problem of estimating an individualized treatment regime (ITR), which defines a mapping from patient characteristics to a treatment decision rule with the goal of maximizing the expected clinical outcome (Murphy et al. 2001; Robins 2004). Existing methods for CATE and ITR estimation often struggle to accommodate high-dimensional covariates and sam-

pling or treatment assignment biases in observational studies, thereby limiting their accuracy, interpretability, and applicability in real-world settings.

Another important class of problems arises in causal mediation analysis. Despite the vast methodological and theoretical developments in mediation analysis with single and multiple mediators (Daniel et al. 2015; Imai and Yamamoto 2013; VanderWeele and Vansteelandt 2014; VanderWeele et al. 2014), the increasing availability of high-dimensional data has introduced new challenges, as modern causal mediation analyses often involve a large number of mediators. Existing methods for handling high-dimensional mediators primarily rely on (non-) linear or other low-dimensional transformations (Chén et al. 2018; Huang and Pan 2016; Nath et al. 2023). However, these approaches often do not explicitly distinguish between material variation in the mediators that is relevant to the mediation mechanism and immaterial variation that is unrelated to the causal pathways of interest.

In this dissertation, we develop statistical methods to address the challenges posed by high-dimensional data through dimension reduction. Unlike general-purpose dimension reduction approaches that primarily preserve variation in the observed data, our methods learn lower-dimensional representations tailored to the causal estimands of interest, specifically heterogeneous treatment effects and natural indirect effects. A summary of the proposed methods is provided in Table 1.1.

Specifically, in Chapter 2, we propose a novel sufficient dimension reduction (SDR) that directly targets the treatment effect heterogeneity and identifies a low-dimensional linear subspace of the covariates capturing treatment effect heterogeneity that can be used under both randomized and non-randomized treatment assignment. Previous methods either assume that the propensity score is known or consistently estimated (Liang and Yu 2022), or impose the restrictive assumption that the subspace sufficient for modeling heterogeneous treatment effects is also sufficient for confounding adjustment (Huang and Chan 2017). To accommodate observational data, our method incorporates kernel-based covariate balancing weights (Wong and Chan 2018) in both the dimension reduction and ITR learning. This avoids potential bias arising from misspecified propensity score models and allows treatment assignment to depend on the full covariate set. The proposed method enables an accurate estimation of an optimal ITR within the outcome-weighted learning framework (Zhao et al. 2012; Zhou and Kosorok 2017). Under standard assumptions, we establish the universal

consistency of the estimated decision rules in the dimension-reduced covariate space.

In Chapter 3, we develop a novel dimension reduction approach for causal mediation analysis based on envelope models (Cook et al. 2013a; Cook, R. D. et al. 2010). Mediation occurs through components of the mediator that are associated with both the exposure and the outcome conditional on exposure. We therefore define the material mediation subspace as the intersection of the exposure-relevant and outcome-relevant mediator subspaces. Existing literature on dimension reduction (e.g., Chén et al. 2018; Huang and Pan 2016; Nath et al. 2023) generally does not distinguish this intersection from mediator directions associated with only one of these two relationships. Consequently, they may retain a larger subspace containing exposure-only and outcome-only variation that does not contribute to a complete mediation pathway. Under the linear structural equation modeling (LSEM) framework, the proposed method finds the linear subspace of the mediator that distinguishes material variation in the mediators that is relevant to both the exposure and the outcome from immaterial variation that is unrelated to one or both. By isolating the material mediation subspace, the proposed method enables more efficient estimation of the natural indirect effect. We further establish the asymptotic normality of the proposed estimator both with and without a normality assumption on the error terms.

In Chapter 4, we develop a nonlinear representation learning method for CATE estimation using neural networks trained under the R-learner loss (Nie and Wager 2021). The proposed architecture learns a CATE model by jointly learning the nuisance functions (conditional outcome expectation and propensity score) and the CATE function. This joint learning approach is motivated by the fact that these functions depend on overlapping features and can therefore benefit from shared representations. By exploiting this common structure, the proposed method aims to improve upon the conventional two-stage, cross-fitted R-learner. Moreover, it leverages the flexibility of neural networks to learn lower-dimensional representations that retain information relevant to treatment effect heterogeneity. We further extend this framework to multimodal settings in which the covariates consist of both unstructured data, such as images, and structured data. In simulation studies, we show that the proposed method effectively addresses the ultra-high dimensionality introduced by the unstructured data. This extension provides a flexible approach to estimating heterogeneous treatment effects from complex, high-dimensional, and multimodal data.

Below, we detail the aims of each of the statistical methods.

Table 1.1: Overview of the dissertation.

	Causal target	Dimension-reduction strategy
Chapter 2	CATE and optimal ITR	Sufficient dimension reduction with covariate balancing
Chapter 3	Natural indirect effect	Envelope model
Chapter 4	CATE	Nonlinear representation learning via neural networks

The remainder of this dissertation is organized as follows. Chapters 2–4 present the three methodological contributions described above. We introduce the notations and assumptions for each method development within each chapter. Chapter 5 summarizes the main findings and outlines directions for future research.

Chapter 2

Sufficient Dimension Reduction for CATE and ITR Estimation

2.1 Introduction

An individualized treatment regime (ITR) is a mapping from individual-level characteristics to a treatment rule, which aims to maximize the expected outcome and accommodate treatment effect heterogeneity across the population (Murphy et al. 2001; Robins 2004; Zhao et al. 2019). Statistical methods for identifying optimal treatments have been applied across a range of clinical areas, including inflammatory bowel disease, cancer, depression, substance abuse, and beyond (Rosthøj et al. 2006; Xie et al. 2022; Zhao et al. 2019; Zhao et al. 2009).

Broadly, there are two main approaches to learning ITRs: *indirect* and *direct* methods. Indirect methods model the conditional mean of the potential outcomes under each treatment, $\{Y(+1), Y(-1)\}$, given covariates X, that is, $\mathbb{E}[Y(+1) \mid X]$ and $\mathbb{E}[Y(-1) \mid X]$ (Murphy et al. 2001; Schulte et al. 2014; Watkins and Dayan 1992; Zhao et al. 2009). Alternatively, they may model the conditional average treatment effect, $\mathbb{E}[Y(+1) - Y(-1) \mid X]$ (Murphy 2003; Robins 2004; Schulte et al. 2014). In these approaches, the optimal treatment at a given covariate value $X = x$ is defined as the one yielding the larger expected outcome (under the convention that larger outcomes are preferable). Therefore, the problem effectively reduces to accurate estimation of treatment effect heterogeneity, and methods such as R-learning (Nie and Wager 2021) can then be used to derive optimal treatment regimes.

Direct methods, on the other hand, aim to learn the treatment rule that maximizes the expected potential outcome by formulating an optimization problem over a prespecified class of decision functions. For example, outcome-weighted learning (OWL) and its extensions formulate this problem as a weighted support vector machine (SVM) (Liu et al. 2018; Zhao et al. 2012; Zhou and Kosorok 2017; Zhou et al. 2017). While these methods can outperform indirect approaches, they remain sensitive to high-dimensional covariates (Dasgupta et al. 2019).

Sufficient dimension reduction (SDR) techniques provide a powerful tool to mitigate the curse of dimensionality and improve performance across a variety of statistical modeling tasks (Cook 2018b; Cook and Li 2002). In this work, we propose an SDR framework for estimating individualized treatment regimes. The proposed SDR aims to project high-dimensional covariates onto a lower-dimensional subspace that preserves the conditional average treatment effect, while allowing treatment assignment A to depend on the full covariate set X. We show that by targeting the subspace that retains treatment effect heterogeneity, our approach enables a more accurate estimation of optimal treatment regimes via direct methods.

The idea of reducing the dimensionality of the covariate space to facilitate learning ITRs has been explored in prior work. For example, Park et al. (2021) proposes a single-index model that linearly transforms the covariates into a one-dimensional variable. This reduced representation is then used to estimate the treatment–covariate interaction effect, from which an optimal treatment rule can be derived. However, this approach relies on the assumption that a single linear combination of covariates adequately captures treatment effect heterogeneity, which may be overly restrictive. SDR approaches for learning ITRs that relax this assumption have been proposed by Park et al. (2020) and Zhou et al. (2021). Specifically, Park et al. (2020) formulate the problem of estimating a reduced subspace of the covariates as a constrained least squares problem, modeling the interaction between treatment and the reduced covariate representation within a Reproducing Kernel Hilbert Space (RKHS). However, this approach is limited to randomized clinical trial settings, where treatment assignment is independent of the covariates. Zhou et al. (2021) focus on the setting of continuous treatments, proposing a direct and pseudo-direct learning approach. Both approaches use the Nadaraya–Watson estimator of the conditional expectation function. However, the direct approach involves a computationally intensive alternating algorithm that

iteratively updates the basis for the reduced subspace and the decision rule. The pseudo-direct method, on the other hand, relies on the additional assumption that the outcome depends only on the dimension-reduced covariates. Both approaches require solving non-convex objective functions and depend on aggressive dimension reduction to enable the use of the Nadaraya–Watson estimator.

Some SDR approaches designed for average treatment effect (ATE) estimation can also be used to learn ITRs. For example, joint SDR (Cheng et al. 2022; Huang and Chan 2017) identifies a transformation $B^\top X$ such that $\{Y(+1), Y(-1)\} \perp\!\!\!\perp X \mid B^\top X$ and $A \perp\!\!\!\perp X \mid B^\top X$. Based on this reduced representation, an ITR can be learned by estimating the conditional ATE given $B^\top X$. However, this approach relies on the strong assumption that there exists a subspace $\text{span}(B)$ such that the conditional independence holds for the potential outcomes under both regimes, i.e., $Y(+1)$ and $Y(-1)$. This assumption is relaxed in Huang and Yang (2023); however, it is still assumed that $A \perp\!\!\!\perp X \mid B^\top X$, which can also be restrictive as it assumes that the subspace sufficient for modeling treatment effect heterogeneity is also sufficient for adjusting for confounding.

Our proposed SDR approach estimates a reduced subspace for learning ITRs. Specifically, we build upon gradient kernel dimension reduction (gKDR) (Fukumizu and Leng 2014) to identify a subspace that captures the relationship between the covariates and the contrast between the two potential outcomes. This nonparametric method avoids the elliptical distribution assumption required by classical techniques such as sliced inverse regression (Li 1991). In addition, gKDR offers a computationally efficient procedure that scales well to large datasets and high-dimensional covariates, which are often challenging to deal with using other methods that also circumvent the ellipticity assumption: e.g., MAVE (Xia et al. 2002) and KDR (Fukumizu et al. 2009).

Our approach accommodates observational settings by allowing treatment assignment to depend on the full set of covariates, rather than restricting it to the same reduced subspace that captures treatment effect heterogeneity. This is achieved by incorporating covariate balancing weights into the proposed SDR framework to account for differences in covariate distributions across treatment groups. Several strategies exist for estimating such weights. The most common is inverse propensity weighting (IPW), although this can be sensitive to model misspecification. Recently, machine learning methods such as extreme gradient

boosting (XGBoost) and neural network (NN) approach (Shang et al. 2025) have been proposed to address the misspecificaiton issue. As an alternative, a growing literature proposes methods that directly estimate balancing weights in the context of ATE estimation (Chan et al. 2016; Hirshberg et al. 2019; Imai and Ratkovic 2014; Kallus and Santacatterina 2022; Wong and Chan 2018; Zubizarreta 2015). Chen et al. (2024) introduce a covariate balancing technique for learning ITRs with OWL methods. Here, we integrate the kernel-based covariate functional balancing (KCB) approach proposed by Wong and Chan (2018) in our SDR framework and develop asymptotic theory demonstrating its suitability for learning ITRs. Finally, the estimated subspace and the balancing weights are used to learn optimal treatment regimes via augmented outcome-weighted learning (AOL) (Liu et al. 2018; Zhou and Kosorok 2017). We establish theoretical guarantees for the overall procedure under mild regularity conditions.

Figure 2.1 illustrates the proposed framework. In Figure 2.1(a), we show the data, which consist of triplets of outcome Y, covariates $\left(X^{(1)}, X^{(2)}\right)$, and treatment group A. The solid line represents the linear transformation $B^\top X$ that best captures heterogeneity in treatment effect. Figure 2.1(b) shows the pseudo-outcomes, defined as $Y/\Pr(+1|X)$ and $-Y/\Pr(-1|X)$ for treatment groups $A = +1$ and $A = -1$, respectively, as a function of $V = B^\top X$. Pseudo-outcomes allow us to account for differences in covariate distributions across treatment groups. Figure 2.1(c) shows the pseudo-outcomes as a function of $V^\perp$, the subspace perpendicular to span(B). Note that there is no notable heterogeneous treatment effect when the covariates are projected onto this direction. Using the proposed SDR framework, we estimate the direction that best captures treatment effect, shown as the dashed line in Figure 2.1(a). Figure 2.1(d) depicts the pseudo-outcomes as a function of the estimated subspace and the optimal decision rule fitted using AOL. The sign of the true heterogeneous treatment effects (solid line) and the fitted decision function (dashed line) are equal in the range $\hat{V} \in (-1, 1)$, in which approximately 95% of the data points lie. Moreover, using an unsupervised dimension reduction method such as principal components analysis would provide a poor reduction for the purpose of identifying optimal ITRs, since the first principal direction is more closely aligned with $V^\perp$.

The remainder of the paper is organized as follows. Section 2.2 introduces the notation and outlines the assumptions underlying the proposed framework. In Section 2.3, we present

our methodology for SDR and describe the subsequent steps for optimal ITR estimation. Section 2.4 establishes the theoretical guarantees of the proposed approach. We present simulation results in Section 3.4. In Section 3.5, we demonstrate the utility of our method using intensive care unit sepsis patient data.

2.2 Preliminaries

We first introduce notation and the problem setup for learning ITR. Let $A \in \{-1, +1\}$ denote a binary treatment assignment, $X \in \mathcal{X} \subseteq \mathbb{R}^p$ the vector of covariates, and $Y \in \mathbb{R}$ a continuous outcome. We denote potential outcomes (Rubin 1974) by $\{Y(A) : A \in \{-1, +1\}\}$ and the observed outcome by $Y = \mathbb{1}(A = +1)Y(+1) + \mathbb{1}(A = -1)Y(-1)$, where $\mathbb{1}(\cdot)$ is the indicator function. We use standard asymptotic notation throughout: if a sequence of random variables X_n is such that $X_n = O_p(a_n)$, then X_n/a_n is bounded in probability. If $X_n = o_p(a_n)$, then X_n/a_n converges to zero in probability as $n \to \infty$. We use $X_n \to_p X$ when $X_n = X + o_p(1)$.

In practice, the observed data consist of triplets $\{(A_i, Y_i, X_i) : i \in [n]\}$, where n denotes the sample size and $[n]$ denotes the index set $\{1, 2, \ldots, n\}$. We denote the j-th component of a random vector X by $X^{(j)}$. The propensity score is defined as $\pi(A, X) = \Pr(A \mid X)$. An RKHS defined on a space $\mathcal{S}$ is denoted by $\mathcal{H}_\mathcal{S}$, and its associated norm and inner product are $\|\cdot\|_{\mathcal{H}_\mathcal{S}}$ and $\langle\cdot,\cdot\rangle_{\mathcal{H}_\mathcal{S}}$, respectively.

Our proposed methodology in Chapter 2 is based on the following assumptions:

Assumption 2.2.1 (No unmeasured confounding)**.** The potential outcomes are independent of treatment assignment given covariates: $\{Y(+1), Y(-1)\} \perp\!\!\!\perp A \mid X$.

Assumption 2.2.2 (Positivity)**.** The propensity score is strictly positive: $\pi(a, x) > 0$ for all $a \in \{-1, +1\}$ and all $x \in \mathcal{X}$.

Assumption 2.2.3 (Central mean subspace)**.** There exists a central mean subspace (CMS), $\text{span}(B_0)$, such that $\mathbb{E}[Y(+1) - Y(-1) \mid X] = \mathbb{E}[Y(+1) - Y(-1) \mid B_0^\top X]$, where $B_0 \in \mathbb{R}^{p\times u}$ with $u < p$. Moreover, $\text{span}(B_0)$ is the minimal subspace for which this equality holds (Cook and Li 2002).

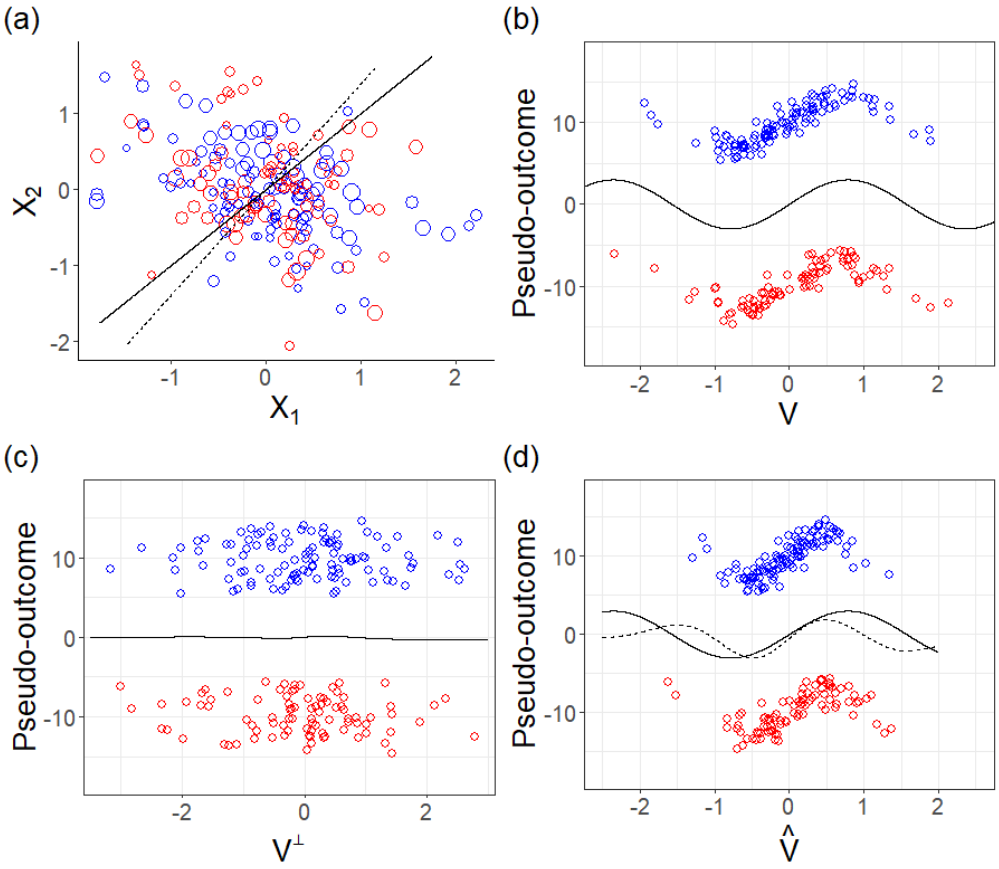

Figure 2.1: Blue points correspond to the treatment group "+1" and red points correspond to the treatment group "-1". The sample sizes for the +1 and -1 treatment groups are 110 and 90, respectively. (a) shows the distribution of the data with respect to the covariates $\left(X^{(1)}, X^{(2)}\right)$. The magnitudes of the outcomes are represented by varying point sizes. The solid line represents the true projection line that best captures treatment effect heterogeneity, while the dashed line denotes the estimated projection line obtained from our proposed SDR approach. (b) displays the true pseudo-outcomes $(Y/\Pr(+1|X)$ and $-Y/\Pr(-1|X)$ for treatment groups $A = +1$ and $A = -1$) in the reduced subspace, where $\left(X^{(1)}, X^{(2)}\right)$ is projected onto $V = B_0^\top X$. The curve represents the heterogeneous treatment effect as a function of V. (c) shows the pseudo-outcomes projected onto $V^\perp$, the space perpendicular to that spanned by B_0. The solid line shows the treatment effect as a function of $V^\perp$. (d) shows the pseudo-outcomes as a function of $\hat{V} = \hat{B}^\top X$, where $\hat{B}$ is an estimate of B_0 obtained using the proposed SDR framework. The solid line shows the true heterogeneous treatment effect along the different levels of the dimension-reduced covariate. The dashed line shows the estimated decision rule along the different levels of the dimension-reduced covariate. The signs of these functions, which determine the true and estimated treatment assignments, are equal in the range $\hat{V} \in (-1, 1)$, where approximately 95% of the data points lie.

Assumptions 2.2.1 and **2.2.2** are standard in the causal inference literature and are required to ensure identifiability. **Assumption 2.2.3** is central to the proposed methodology, which estimates a CMS such that $B_0^\top X$ captures treatment effect heterogeneity. This enables a more efficient estimation of an ITR based on the reduced covariate representation. A key feature of our approach is that we do not assume $\{Y(+1), Y(-1)\} \perp\!\!\!\perp X \mid B_0^\top X$, thereby allowing for complex relationships between the potential outcomes and the covariates, as long as their contrast depends on a lower-dimensional subspace. Also, we do not impose the assumption $A \perp\!\!\!\perp X \mid B^\top X$, thus accommodating settings where the subspace sufficient for capturing treatment effect heterogeneity may not be sufficient for adjusting for confounding.

Let $d : \mathcal{X} \to \{-1, +1\}$ denote a decision rule that defines an ITR. This can be expressed as the sign of a decision function $f : \mathcal{X} \to \mathbb{R}$ as $d(\cdot) = \text{sign} \circ f(\cdot)$. The value function, representing the expected reward under the decision rule d, is defined as the expected potential outcome under the treatment assigned by d, and is given by

$$\mathbb{E}\left[Y\{d(X)\}\right] = \mathbb{E}\left[\frac{Y}{\pi(A,X)}\mathbb{1}\{d(X) = A\}\right] = \mathbb{E}\left[\frac{Y}{\pi(A,X)}\mathbb{1}\{\text{sign} \circ f(X) = A\}\right]. \tag{2.1}$$

Without loss of generality, we assume that larger values of the outcome are more desirable.

Given $\text{span}(B_0)$, we define the dimension-reduced covariates as $V_0 = B_0^\top X \in \mathcal{V}$, where $\mathcal{V} = \{v : v = B_0^\top x, x \in \mathcal{X}\} \subseteq \mathbb{R}^u$ denotes the corresponding reduced subspace. Given an estimate $\hat{B}$ of the subspace basis with rank $\hat{u}$ selected based on an appropriate tuning procedure, the estimated low-dimensional covariates are $\hat{V} = \hat{B}^\top X \in \hat{\mathcal{V}}$, where $\hat{\mathcal{V}} = \{v : v = \hat{B}^\top x, x \in \mathcal{X}\} \subseteq \mathbb{R}^{\hat{u}}$. The decision functions in these reduced spaces are denoted by $\tilde{f}^V : \mathcal{V} \to \mathbb{R}$ and $\tilde{f}^{\hat{V}} : \hat{\mathcal{V}} \to \mathbb{R}$, respectively. With a slight abuse of notation, we use $d(\cdot)$ to denote the decision rule defined on either $\mathcal{V}$ or $\hat{\mathcal{V}}$, depending on the context.

2.3 Method

In this section, we present our proposed methodology. Section 2.3.1 introduces a gradient kernel dimension reduction for identifying a low-dimensional subspace that captures the dependence of the contrast between potential outcomes on the covariates. We begin by constructing a suitable pseudo-outcome and then develop the gKDR framework for this

variable. Section 2.3.2 describes how the estimates obtained via KCB can be integrated into our SDR procedure to account for differences in covariate distributions across treatment groups. Finally, Section 2.3.3 outlines how AOL can be applied within the identified sufficient subspace of the covariates.

2.3.1 Gradient kernel dimension reduction for the difference between potential outcomes

Construction of the pseudo-outcome

The value function (2.1) is maximized when the decision function has the same sign as $\mathbb{E}[Y(+1) - Y(-1)|X = x]$. Hence the conditional ATE characterizes the Bayes-optimal decision boundary (Zhao et al. 2012). **Assumption 3** ensures that SDR with respect to $\mathbb{E}[Y(+1) - Y(-1)|X = x]$ enables the identification of a low-dimensional subspace of the domain of the optimal decision rule. Under **Assumptions 1–3**, we can identify $\text{span}(B_0)$ by noting that $\mathbb{E}[AY/\pi(A,X) \mid B_0^\top X] = \mathbb{E}[\mathbb{E}\{AY/\pi(A,X) \mid X\} \mid B_0^\top X] = \mathbb{E}[Y(+1) - Y(-1) \mid B_0^\top X] = \mathbb{E}[Y(+1) - Y(-1) \mid X]$. Therefore, the problem reduces to finding a matrix B_0 such that $\mathbb{E}[AY/\pi(A,X) \mid X] = \mathbb{E}[AY/\pi(A,X) \mid B_0^\top X]$. To facilitate estimation, we introduce a function g defined as the projection of $\{1/\pi(A,X) - 1\}Y$ onto the linear span of X, evaluated at x:

$$g(x) = \mathbb{E}\left[x^\top (\mathbb{E}[XX^\top])^{-1} X\{1/\pi(A,X) - 1\}Y\right]. \tag{2.2}$$

Throughout the paper, we assume that $\mathbb{E}[XX^\top]$ is positive definite and $\mathbb{E}[X\{1/\pi(A,X) - 1\}Y]$ exists, so that $g(x)$ is well-defined. As motivated in Section 2.3.3, we then define the pseudo-outcome $Z = A\{Y - g(X)\}/\pi(A,X)$, which can be interpreted as a contrast between weighted residuals. Note $\mathbb{E}\left[Ag(X)/\pi(A,X) \mid X\right] = g(X)\mathbb{E}\left[A/\pi(A,X) \mid X\right] = 0$ since $\mathbb{E}\left[A/\pi(A,X) \mid X\right] = 0$. Then **Assumption 3** implies $\mathbb{E}\left[Z \mid B_0^\top X\right] = \mathbb{E}\left[\mathbb{E}[Z \mid X] \mid B_0^\top X\right] = \mathbb{E}\left[Y(+1) - Y(-1) \mid X\right]$. Next, we describe our approach for estimating $\text{span}(B_0)$, and in Section 2.3.2 we outline our procedure for estimating $g(x)$.

Gradient kernel dimension reduction

Gradient-based approaches have been widely used for developing SDR methods, including IADE (Hristache et al. 2001) and wOPG (Kang and Shin 2022). Here, we extend gKDR (Fukumizu and Leng 2014) to our setting, targeting estimation of the central mean subspace using pseudo-outcomes. Let $\mathcal{H}_\mathcal{X}$ denote the RKHS of real functions $f : \mathcal{X} \to \mathbb{R}$, equipped with a reproducing kernel $K_X : \mathbb{R}^p \times \mathbb{R}^p \to \mathbb{R}$ inducing an inner product $\langle \cdot, \cdot \rangle_{\mathcal{H}_\mathcal{X}}$. Let $P_{X,Z}$ be the joint distribution of (X, Z). Then, for any $x \in \mathcal{X}$, $\mathbb{E}[K_X(x, X)Z] = \int K_X(x, x') \int z' dP_{Z|X}(z'|x') dP_X(x') = \int K_X(x, x') \mathbb{E}[Z|X = x'] dP_X(x')$, where $P_{Z|X}$ and P_X are the distributions of $Z|X$ and X, respectively. Under the additional assumption that $\mathbb{E}[Z|X = x] \in \mathcal{H}_\mathcal{X}$, we have $\left(C_{XX}^{-1} \mathbb{E}[K_X(\cdot, X)Z]\right)(x) = \mathbb{E}[Z|X = x]$, where C_{XX}^{-1} is the inverse of the covariance operator $C_{XX} : \mathcal{H}_\mathcal{X} \to \mathcal{H}_\mathcal{X}$, which admits the integral operator representation $(C_{XX} f)(x) = \int K_X(x, x') f(x') dP_X(x')$. Then,

$$\mathbb{E}\left[Z|X = x\right] = \langle \mathbb{E}[Z|X = \cdot], K_X(\cdot, x) \rangle_{\mathcal{H}_\mathcal{X}} = \left\langle C_{XX}^{-1} \mathbb{E}[K_X(\cdot, X)Z], K_X(\cdot, x) \right\rangle_{\mathcal{H}_\mathcal{X}} \tag{2.3}$$

by the reproducing property. We use the gradient of $\mathbb{E}[Z|X = \cdot]$ to estimate B_0 by noting

$$\frac{\partial}{\partial x^{(m)}} \mathbb{E}\left[Z|X = x\right] = \sum_{a \in [u]} B_{0,ma} \frac{\partial}{\partial v^{(a)}} \mathbb{E}[Z|B_0^\top X = v], \tag{2.4}$$

where $B_{0,ma}$ is the (m, a)-th entry of B_0. Moreover, in the Appendix, we show that from Equation (2.3) and Assumption (A6) (also defined in the Appendix), we have that

$$\frac{\partial}{\partial x^{(m)}} \mathbb{E}\left[Z|X = x\right] = \left\langle C_{XX}^{-1} \mathbb{E}[K_X(\cdot, X)Z], \frac{\partial K_X(\cdot, x)}{\partial x^{(m)}} \right\rangle_{\mathcal{H}_\mathcal{X}} = \mathbb{E}\left[C_{XX}^{-1} \frac{\partial K_X(X, x)}{\partial x^{(m)}} Z \right]. \tag{2.5}$$

Equating (2.4) and (2.5) yields $\sum_{a \in [u]} B_{0,ma} \partial \mathbb{E}[Z|B_0^\top x = v] / \partial v^{(a)} = \mathbb{E}\left[C_{XX}^{-1} \partial K_X(X, x) / \partial x^{(m)} Z\right]$. Next, define the matrix $M(x) \in \mathbb{R}^{p \times p}$ with entries

$$\begin{aligned} M_{mj}(x) &= \left\{ \sum_{a \in [u]} B_{0,ma} \frac{\partial}{\partial v^{(a)}} \mathbb{E}[Z|B_0^\top X = v] \right\} \left\{ \sum_{a \in [u]} B_{0,ja} \frac{\partial}{\partial v^{(a)}} \mathbb{E}[Z|B_0^\top X = v] \right\} \\ &= \mathbb{E}\left[C_{XX}^{-1} \frac{\partial K_X(X, x)}{\partial x^{(m)}} Z \right] \cdot \mathbb{E}\left[C_{XX}^{-1} \frac{\partial K_X(X, x)}{\partial x^{(j)}} Z \right]. \end{aligned}$$

Then, $M(x) = B_0 D(x) B_0^\top$, where $D(x) \in \mathbb{R}^{u \times u}$ is a symmetric matrix with entries $D_{ab}(x) = \partial \mathbb{E}[Z|B_0^\top x = v]/\partial v^{(a)} \cdot \partial \mathbb{E}[Z|B_0^\top x = v]/\partial v^{(b)}$. Since, $\mathbb{E}[M(X)] = B_0 \mathbb{E}[D(X)] B_0^\top$, a basis for $\text{span}(B_0)$ can be obtained from the leading u eigenvectors of $\mathbb{E}[M(X)]$.

Gradient kernel dimension reduction with balancing weights

In practice, constructing the empirical counterpart of the pseudo-outcome requires estimating the inverse propensity score. While this is trivial in randomized settings, the estimates of the propensity score in observational studies can substantially affect the results of the analysis. Although the proposed SDR framework is compatible with any plug-in estimator of the propensity score, in this work, we adopt a kernel-based covariate balancing approach and integrate the resulting balancing weights $\hat{w}_i$ in our SDR and ITR frameworks. This yields the empirical pseudo-outcome $\tilde{Z}_i = \hat{w}_i A_i \{Y_i - g_{\hat{w}}(X_i)\}$, where $\hat{w}_i$ and $g_{\hat{w}}(X_i)$ are defined in Section 2.3.2. Once $\{\tilde{Z}_i;\ i \in [n]\}$ are obtained, for any $x \in \mathcal{X}$, we estimate $M(x)$ using $\hat{W}(x)$, whose (m, j)-entry is given by:

$$\hat{W}_{mj}(x) = \left\{ \frac{1}{n} \sum_{i \in [n]} (\hat{C}_{XX} + \epsilon_n I)^{-1} \frac{\partial K_X(X_i, x)}{\partial x^{(m)}} \tilde{Z}_i \right\} \left\{ \frac{1}{n} \sum_{i \in [n]} (\hat{C}_{XX} + \epsilon_n I)^{-1} \frac{\partial K_X(X_i, x)}{\partial x^{(j)}} \tilde{Z}_i \right\},$$

where I is the identity operator, $\epsilon_n I$ is a Tikhonov regularization (Nashed 1986), and $\hat{C}_{XX}$ is the empirical covariance operator such that $(\hat{C}_{XX} f)(x) = n^{-1} \sum_{i \in [n]} K_X(x, X_i) f(X_i)$.

Finally, we estimate $\mathbb{E}[M(X)]$ by the empirical average $\tilde{W}_n = n^{-1} \sum_{i \in [n]} \hat{W}(X_i)$. In Theorem 2.4.5, we show that $\tilde{W}_n$ converges in probability to $\mathbb{E}[M(X)]$, for every $x \in \mathcal{X}$. The top $\hat{u}$ eigenvectors of $\tilde{W}_n$ are used to obtain $\hat{B}$. From Corollary 2.4.6, and under appropriate identifiability conditions to resolve invariances, it follows that $\hat{B} \to_p B_0$ in Frobenius norm.

Remark 2.3.1. A limitation of gKDR is that the rank of the estimated matrix $\tilde{W}_n$, and therefore $\hat{B}$, is bounded by that of the outer-product $\tilde{Z}\tilde{Z}^\top$, where $\tilde{Z} = (\tilde{Z}_1, \cdots, \tilde{Z}_n)^\top$. The latter may be rank-deficient (Fukumizu and Leng 2014), for example, in randomized studies with binary outcomes. In such settings, the gKDR-v variant can be used (Fukumizu and Leng 2014), which partitions the data, estimates $M(x)$ within each subset, and then aggregates the results to obtain $\tilde{W}_n$. In observational studies, however, the pseudo-outcomes $\tilde{Z}_i$ typically take many distinct values, so the standard gKDR procedure is usually sufficient.

2.3.2 Kernel-based covariate balancing

We propose using the KCB of Wong and Chan (2018) to compute the balancing weights, $\{\hat{w}_i : i \in [n]\}$. These appear both in the construction of the pseudo-outcome $\{\tilde{Z}_i\}$ and the objective function for learning the optimal ITR. We illustrate KCB for the treatment group $A = +1$, but equivalent results hold for $A = -1$. KCB is motivated by the moment condition $\mathbb{E}\{m(X)\mathbb{1}(A = +1)/\pi(+1, X)\} = \mathbb{E}\{m(X)\}$, for all measurable $m : \mathcal{X} \to \mathbb{R}$, which ensures that the mean of any $m(X)$ matches the marginal mean of the potential outcome under $A = +1$. Introduce $\mathcal{H}_n = \{m \in \mathcal{H}_\mathcal{X}; \|m\|_n^2 = n^{-1}\sum_{i\in[n]} m^2(X_i) = 1\}$. The empirical and penalized counterpart of this moment condition, used to estimate $\{\hat{w}_i : A_i = +1\}$, is then

$$\min_{w\geq 1}\left[\sup_{m\in\mathcal{H}_n}\left\{\left(\frac{1}{n}\sum_{i:A_i=+1} w_i m(X_i) - \frac{1}{n}\sum_{i\in[n]} m(X_i)\right)^2 - \lambda_1\|m\|_{\mathcal{H}_\mathcal{X}}^2\right\} + \lambda_2\frac{1}{n}\sum_{i:A_i=+1} w_i^2\right]. \tag{2.6}$$

This enforces the moment condition above for all $m : \mathcal{X} \to \mathbb{R}$ in a rich function space, with appropriate regularization, thereby reducing the risk of misspecifying the covariate function used for balancing. The weights $\{\hat{w}_i : A_i = -1\}$ can be obtained similarly. A key contribution of this work is Theorem 2.4.1, which establishes that the resulting weights satisfy $n^{-1}\sum_{i:A_i=+1}\{\hat{w}_i - 1/\pi(A_i, X_i)\}^2 = o_p(1)$, under mild regularity conditions. Therefore, these weights provide a suitable notion of covariate balancing within our SDR framework and, as shown in Section 2.4, ensure universal consistency of the estimated decision rule based on the dimension-reduced covariate space. We also emphasize that the RKHS $\mathcal{H}_\mathcal{X}$ used for defining the covariate balancing weights need not be the same as the one used for gKDR. Propensity score models based on nonparametric or machine learning methods can be employed. However, the IPW may be unstable when the propensity scores are close to 0 or 1 in some regions of the covariate space.

Estimation of $g(x)$

We can estimate $g(x)$ in (2.2) consistently under correct model specification by replacing the population expectations with sample moment estimators. When $\pi(A, X)$ is known, this is achieved by using $g_n(x) = n^{-1}\sum_{i\in[n]} x^\top(n^{-1}\mathbb{X}^\top\mathbb{X})^{-1}X_i(1/\pi(A_i, X_i) - 1)Y_i$, where $x^\top(n^{-1}\mathbb{X}^\top\mathbb{X})^{-1}X_i$ is the sample analogue of $x^\top(\mathbb{E}[XX^\top])^{-1}X$, and $\mathbb{X} \in \mathbb{R}^{n\times p}$ is the design

matrix. We show $g_n(x)$ uniformly converges to $g(x)$ in probability in Appendix. When $\pi(A, X)$ is not known, we propose using $g_{\hat{w}}(x) = n^{-1} \sum_{i\in[n]} x^\top (n^{-1}\mathbb{X}^\top\mathbb{X})^{-1} X_i(\hat{w}_i - 1)Y_i$ in place of g_n. As with $g_n(x)$, we can show that $g_{\hat{w}}(x)$ converges uniformly in probability to $g(x)$ (see Corollary 2.4.2).

2.3.3 Estimation of optimal ITRs

The optimal decision function that maximizes the value function (2.1) can be characterized as the minimizer of the risk $\mathcal{R}(f) = \mathbb{E}\left[Y/\pi(A, X)\mathbb{1}\{\text{sign} \circ f(X) \neq A\}\right]$. As shown in Appendix, this can be equivalently written as

$$\begin{aligned}\mathcal{R}(f) = \mathbb{E}\left[|Y - g(X)|/\pi(A, X)\mathbb{1}\{A \cdot \text{sign}(Y - g(X)) \neq \text{sign} \circ f(X)\}\right] \\ + \mathbb{E}\left[X^\top\right]\left\{\mathbb{E}[XX^\top]\right\}^{-1}\mathbb{E}[\pi(-A, X)XY/\pi(A, X)] + \mathbb{E}[\max\{-Y + g(X), 0\}/\pi(A, X)].\end{aligned}$$

Since the last two terms do not depend on the decision function $f(X)$, minimizing $\mathcal{R}(f)$ is equivalent to minimizing $\mathbb{E}[|Y - g(X)|/\pi(A, X)\mathbb{1}\{A \cdot \text{sign}(Y - g(X)) \neq \text{sign} \circ f(X)\}]$. Moreover, given that the Bayes optimal decision rule is determined by the sign of $\mathbb{E}[Y(+1) - Y(-1)|X] = \mathbb{E}[Y(+1) - Y(-1)|B_0^\top X]$, this loss can be equivalently minimized in the reduced space using a function $\tilde{f}^V : \mathcal{V} \to \mathbb{R}$. Replacing the discontinuous and nonconvex 0–1 loss with a convex surrogate loss $\phi(\cdot)$ and using the decision rule $\tilde{f}^V$ leads to the definition of the (ϕ, g)-risk, $\mathcal{R}_{\phi,g}(\tilde{f}^V) = \mathbb{E}\left[|Y - g(X)|/\pi(A, X)\phi\Big(A \cdot \text{sign}\{Y - g(X)\} \cdot \tilde{f}^V(B_0^\top X)\Big)\right]$. Minimizing $\mathcal{R}_{\phi,g}(\tilde{f}^V)$ is then a convex optimization problem, therefore addressing the potential non-convexity that arises when negative values are observed for either $Y/\pi(A, X)$ or $\{Y - g(X)\}/\pi(A, X)$. Fisher consistency of $\phi(\cdot)$, meaning that any minimizer of the (ϕ, g)-risk is also a minimizer of the original 0–1 loss, can be shown by using classical results (Bach 2024; Bartlett et al. 2006).

Let $K_u : \mathbb{R}^u \times \mathbb{R}^u \to \mathbb{R}$ be a kernel. Using the observed (X_i, A_i, Y_i), balancing weights $\hat{w}_i$ and the estimator $g_{\hat{w}}(x)$, we define for any $\tilde{f}^{\hat{V}} : \hat{\mathcal{V}} \to \mathbb{R}$ in the RKHS induced by K_u,

$$Q\left(\tilde{f}^{\hat{V}}\right) = \frac{1}{n}\sum_{i\in[n]} \hat{w}_i \left|Y_i - g_{\hat{w}}(X_i)\right| \phi\Big(A_i \cdot \text{sign}\left\{Y_i - g_{\hat{w}}(X_i)\right\} \tilde{f}^{\hat{V}}(\hat{B}^\top X_i)\Big) + \lambda_n \alpha^\top \hat{G}\alpha,$$

where $\tilde{f}^{\hat{V}}(\hat{B}^\top x) = \sum_{i\in[n]} \alpha_i K_u(\hat{B}^\top X_i, \hat{B}^\top x) + \alpha_0$, $\hat{G} \in \mathbb{R}^{n\times n}$ is the Gram matrix with (i, j)-

th entry $K_u(\hat{B}^\top X_i, \hat{B}^\top X_j)$, $\alpha_0 \in \mathbb{R}$, and $\alpha = (\alpha_1, \alpha_2, \cdots, \alpha_n)^\top \in \mathbb{R}^n$. Finally, we estimate the decision function using $\tilde{f}_n^{\hat{V}} = \arg\min_{\tilde{f}^{\hat{V}}} Q\left(\tilde{f}^{\hat{V}}\right)$, and the corresponding decision rule is given by $\operatorname{sign} \circ \tilde{f}_n^{\hat{V}}(\hat{B}^\top x)$.

In Lemma 2.4.7, we show that $\mathcal{R}_{\phi,g}(\tilde{f}_n^{\hat{V}}) - \inf_{\tilde{f}^{\hat{V}}} \mathcal{R}_{\phi,g}(\tilde{f}^{\hat{V}}) \to_p 0$. Let $f^* : \mathcal{X} \to \mathbb{R}$ and $\tilde{f}^* : \mathcal{V} \to \mathbb{R}$ be the Bayes optimal decision functions from which the Bayes rule is given by $d^*(x) = \operatorname{sign} \circ f^*(x)$ or $d^*(B_0^\top x) = \operatorname{sign} \circ \tilde{f}^*(B_0^\top x)$. Let $\mathcal{R}^* = \mathcal{R}(f^*)$ be the corresponding optimal risk. In Proposition 2.4.8, we establish universal consistency of our estimator of the optimal decision rule, $\operatorname{sign} \circ \tilde{f}_n^{\hat{V}}(\cdot)$, that is, $\mathcal{R}(\tilde{f}_n^{\hat{V}}) \to_p \mathcal{R}^*$.

Note that indirect methods can also be applied once the covariates have been dimension-reduced. One may use Nadaraya-Watson regression to estimate the treatment effect heterogeneity and identify an optimal treatment. However, such methods often suffer in practice when the reduced subspace is not extremely low-dimensional (e.g., dimension lower than four). Dimension-reduced covariates can also be used in conjunction with more flexible learners, such as the R-learner framework (Nie and Wager 2021).

2.4 Theoretical guarantees

Here, we provide theoretical guarantees for the proposed approach to estimating optimal ITRs. The proofs are provided in Appendix A. We first show that the balancing weights $\{\hat{w}_i\}$, converge to the true IPWs in mean squared error, as stated in the following theorem.

Theorem 2.4.1. *Assume $\lambda_1 \asymp n^{-1}$, and let the KCB weights $\{\hat{w}_i : A_i = +1\}$ be the solution to the optimization problem in* (2.6)*. Then, under the regularity conditions* (A1) − (A5) *in Section A.1 of the Appendix,*

$$\frac{1}{n} \sum_{i:A_i=+1} \left\{ \hat{w}_i - \frac{1}{\pi(+1, X_i)} \right\}^2 = o_p(1).$$

By a symmetric argument, $n^{-1} \sum_{i:A_i=-1} \{\hat{w}_i - 1/\pi(-1, X_i)\}^2 = o_p(1)$ for $\{\hat{w}_i : A_i = -1\}$, hence $n^{-1} \sum_{i\in[n]} \{\hat{w}_i - 1/\pi(A_i, X_i)\}^2 = o_p(1)$. To the best of our knowledge, no existing result establishes convergence of KCB weights to the true IPW. In contrast, Wong and Chan (2018) show only that $\left| n^{-1} \sum_{i:A_i=+1} \hat{w}_i m(X_i) - n^{-1} \sum_{i\in[n]} m(X_i) \right| = O_p\left(n^{-1/2}\right)$.

Theorem 2.4.1 implies that the difference between $g_{\hat{w}}(x)$ and $g(x)$ is uniformly negligible in probability over $x \in \mathcal{X}$, as formalized in the following corollary.

Corollary 2.4.2. *Suppose $\mathcal{X}$ is compact and $\mathbb{E}[Y^2] < \infty$. Then,*

$$\sup_{x \in \mathcal{X}} |g_{\hat{w}}(x) - g(x)| = o_p(1).$$

By using Theorem 2.4.1 and Corollary 2.4.2, we can establish the consistency of our estimator for $\mathbb{E}[Zf(X)]$, where $f : \mathcal{X} \to \mathbb{R}$ is any measurable function. This is an important building block for showing the convergence of the estimator of $\mathrm{span}(B_0)$ obtained from the proposed gKDR approach. This result is formalized in the following theorem.

Theorem 2.4.3. *Let $\{\hat{w}_i; i \in [n]\}$ denote the KCB weights, and recall the pseudo-outcomes $Z = A\{Y - g(X)\}/\pi(A, X)$ and $\tilde{Z}_i = \hat{w}_i A_i\{Y_i - g_{\hat{w}}(X_i)\}$. Then, for any measurable $f : \mathcal{X} \to \mathbb{R}$ on compact $\mathcal{X}$, such that $\mathbb{E}[Z^2 f^2(X)] < \infty$ and $\mathbb{E}[f^2(X)] < \infty$, we have*

$$\frac{1}{n} \sum_{i \in [n]} \tilde{Z}_i f(X_i) \to_p \mathbb{E}[Zf(X)].$$

The following theorem establishes the consistency of the empirical (ϕ, g)-risk when the KCB weights are used. Intuitively, from Corollary 2.4.2, we expect that the number of discrepancies between the signs of $Y_i - g(X_i)$ and $Y_i - g_{\hat{w}}(X_i)$ to be controlled.

Theorem 2.4.4. *Let $\{\hat{w}_i; i \in [n]\}$ be the KCB weights. Suppose $\mathcal{X}$ is compact and that the second moment of $|Y - g(X)|\phi\big(A \cdot \mathrm{sign}\{Y - g(X)\}f(X)\big)$ is bounded. Assume that the distribution P of (X, A, Y) satisfies $|Y - g(X)|/\pi(A, X) \leq M_g$ and $|f(X)| \leq M_f$ almost everywhere, where $M_g, M_f \in \mathbb{R}$ are sufficiently large constants. Then*

$$\frac{1}{n} \sum_{i \in [n]} \hat{w}_i |Y_i - g_{\hat{w}}(X_i)| \phi\Big(A_i \cdot \mathrm{sign}\{Y_i - g_{\hat{w}}(X_i)\} f(X_i)\Big) \to_p \mathcal{R}_{\phi,g}(f).$$

To establish that $\mathrm{span}(\hat{B})$ is a consistent estimator of $\mathrm{span}(B_0)$, we first show the convergence of $\hat{W}(x)$ to $M(x)$ and that of $\tilde{W}_n$ to $\mathbb{E}[M(X)]$. We additionally make assumptions (A6)–(A10), in the Appendix, which are adapted from those in Fukumizu and Leng (2014).

Let $C_{XX}^{\xi+1}$ be the $(\xi+1)$-th power of the covariance operator C_{XX}. The following theorem formalizes these statements.

Theorem 2.4.5. *Assume* $\mathbb{E}[Y^2\{C_{XX}^{-1}\partial K(x,X)/\partial x^{(a)}\}^2]$, $\mathbb{E}[g^2(X)\{C_{XX}^{-1}\partial K(x,X)/\partial x^{(a)}\}^2] < \infty$ *for any* $x \in \mathcal{X}$ *and* $a \in [p]$, *with* p *fixed. For a constant* $\xi \geq 0$ *suppose that* $\partial K(\cdot,x)/\partial x^{(a)}$ *is in the range of* $C_{XX}^{\xi+1}$. *Then, under assumptions (A6)-(A10), for every* $x \in \mathcal{X}$,

$$\left\|\hat{W}(x) - M(x)\right\|_F = o_p(1).$$

Moreover, if $\mathbb{E}[\|M(X)\|_F^2] < \infty$ *and* $\partial K(\cdot,x)/\partial x^{(a)} = C_{XX}^{\xi+1}h_{a,x}$, *where* $h_{a,x} \in \mathcal{H}_{\mathcal{X}}$ *and* $E\|h_{a,x}\|_{\mathcal{H}_{\mathcal{X}}} < \infty$, *then*

$$\left\|\tilde{W}_n - \mathbb{E}[M(X)]\right\|_F = o_p(1).$$

Assumptions (A6)-(A10) are needed to derive the estimator for B_0. When $\partial K(x,X)/\partial x^{(a)} \in \mathcal{H}_{\mathcal{X}}$, $C_{XX}^{-1}\partial K(x,X)/\partial x^{(a)} \in \mathcal{H}_{\mathcal{X}}$ from assumption (A8). If the kernel for $\mathcal{H}_{\mathcal{X}}$ is bounded, the moment condition can be relaxed to $\mathbb{E}[Y^2]$, $\mathbb{E}[g(X)^2] < \infty$.

Theorem 2.4.5 implies the following corollary, which follows from Theorem 2 of Yu et al. (2015), a variant of the Davis–Kahan $\sin\theta$ theorem (Davis and Kahan 1970).

Corollary 2.4.6. *Let* B_0 *and* $\hat{B}$ *be matrices whose columns are the eigenvectors corresponding to the* u *largest eigenvalues of* $\mathbb{E}[M(X)]$ *and* $\tilde{W}_n$, *respectively. Suppose the non-zero eigenvalues of* $\mathbb{E}[M(X)]$ *are distinct. Then,* $\left\|\tilde{W}_n - \mathbb{E}[M(X)]\right\|_F = o_p(1)$ *implies that there exists an orthogonal matrix* $\hat{O} \in \mathbb{R}^{u\times u}$ *such that*

$$\left\|\hat{B}\hat{O} - B_0\right\|_F = o_p(1).$$

Assuming $\mathcal{X}$ is compact, then Corollary 2.4.6 implies that $\hat{O}^\top\hat{B}^\top x \to_p B_0^\top x$ for any $x \in \mathcal{X}$. When $\pi(a,x)$ is known, we can establish convergence rates in which the number of covariates p grows with the sample size (see Remark A.1 in Appendix).

For the following analysis, we make a technical assumption as in Wong and Chan (2018) and Athey et al. (2018), and assume $\hat{w}_i \leq M_w n^{1/3}$, $i \in [n]$, where M_w is a large constant. This upper bound constraint can be incorporated in the estimation of $\{\hat{w}_i : i \in [n]\}$ in (2.6).

Lemma 2.4.7. *Assume that the surrogate loss function* $\phi(\cdot)$ *is Lipschitz continuous and the kernel satisfies* $K_u(v,v) < \infty$ *for all* $v \in \mathbb{R}^u$. *Suppose* $\mathcal{X}$ *is compact, and* $\lambda_n > 0$ *with* $\lambda_n \to 0$

and $n^{1/3} \cdot \lambda_n \to \infty$ *as* $n \to \infty$. *Let* $\alpha_{0n}^{\hat{V}}$ *be the estimated intercept of* $\tilde{f}_n^{\hat{V}}$. *Additionally, assume that and the distribution* P *of* (X, A, Y) *satisfies* $|Y - g(X)|/\pi(A, X) \leq M_g$ *and* $|\sqrt{\lambda_n}\, \alpha_{0n}^{\hat{V}}| \leq M_{\alpha_0} < \infty$ *almost everywhere, and the second moment of* $\phi\Big(A \cdot \text{sign}\{Y - g(X)\}\tilde{f}^{\hat{V}}(\hat{B}^\top X)\Big)$ *is bounded. Furthermore, suppose that for a sufficiently large constant* M_w, $\hat{w}_i \leq M_w n^{1/3}$ *for all* $i \in [n]$. *Then,*

$$\mathcal{R}_{\phi,g}(\tilde{f}_n^{\hat{V}}) \to_p \inf_{\tilde{f}^{\hat{V}}} \mathcal{R}_{\phi,g}(\tilde{f}^{\hat{V}})$$

By combining this result with Lemma A.3 and Lemma A.4 in Appendix A, we establish the universal consistency of the ITR learned by the proposed method under the hinge loss. This is formalized in the following proposition.

Proposition 2.4.8. *Suppose that the RKHS of* $\tilde{f}^{\hat{V}} : \hat{\mathcal{V}} \to \mathbb{R}$ *is induced by a universal kernel, and* $\mathcal{X}$ *is a compact metric space. Let* $\phi(\cdot)$ *be the hinge loss function. Assume that the conditions stated in Lemma 2.4.7 hold. Then,*

$$\mathcal{R}(\tilde{f}_n^{\hat{V}}) \to_p \mathcal{R}^*.$$

2.5 Simulations

In this section, we evaluate the finite-sample performance of the proposed dimension-reduced outcome-weighted learning, which here we refer to as DOL, across three settings. The data were generated in both randomized and non-randomized scenarios, where $\Pr(A = +1 \mid X) = \Pr(A = -1 \mid X) = 1/2$ in the randomized setting. We compare the DOL against five methods: (AOL) the augmented outcome-weighted learning by Zhou and Kosorok (2017), which uses the full covariate set X; (MAVE) the interaction MAVE (Liang and Yu 2022); (SI) constrained single-index regression (Park et al. 2021); (QL) an ℓ_1-penalized partial least squares (ℓ_1-PLS) Q-learning method (Qian and Murphy 2011); and (RL) R-learner using Gaussian kernel ridge regression (Nie and Wager 2021).

In the non-randomized setting, we additionally compared four covariate balancing strategies for the proposed method – KCB and three methods that directly estimate IPW – logistic regression, XGBoost, and neural network (Shang et al. 2025). The same balancing weights were applied for both the gKDR and the ITR estimation.

Performance was evaluated using accuracy and the value function. Accuracy measures the proportion of the learned rules that agree with the Bayes optimal rule. In each setting, a test set of 10,000 observations, was generated. For each Monte Carlo iteration, a training set was generated, the decision rule was estimated, and performance was evaluated by using the test set (see Appendix A.9.1). The hinge loss, $\phi(x) = \max(0, 1 - x)$, was used as the surrogate loss in AOL and the proposed DOL. In both methods, the penalty term λ_n was selected from a grid of candidate values by choosing the one that yielded the largest two-fold cross-validated value function. Further details on the software implementation and estimation of ITRs are deferred to Appendix A.9.2.

In the Appendix, simulation results to show the decreasing projection errors along increasing sample size is presented in A.10.2. Furthermore, we show in Section A.10.2 that using KCB weights with gKDR yields the lower projection error than using propensity scores from XGBoost (Matthew Cefalu et al. 2024) or neural network (Shang et al. 2025). The computational advantage of gKDR is demonstrated by comparison with interaction MAVE (Liang and Yu 2022) in Appendix A.10.3.

2.5.1 Simulation settings

We considered $n = 500$ and $n = 1{,}000$, with all settings using $p = 50$. We randomly generated a latent factor $Z \in \mathbb{R}^{50}$ from a uniform distribution over the range $[-2, 2]$. This was then used to generate the observed covariates $X \in \mathbb{R}^{50}$. With the main effect, $\mu(X)$, and the treatment-covariate interaction, $\tilde{f}^V(V)$, the outcomes were generated from $N\{\mu(X) + A \cdot \tilde{f}^V(V), 1\}$. Further details of simulation settings are in Appendix A.9.3.

Setting 1 considers $V = B_0^\top X \in \mathbb{R}^2$. The decision boundaries over the original covariates X and the reduced subspace are shown in Figure 2.2. Setting 2 also considers $V = B_0^\top X \in \mathbb{R}^2$. The decision boundaries are shown in Figure A.1 of the Appendix A. This scenario introduces additional complexity due to the non-smoothness of the decision boundary. Setting 3 considers higher-dimensional reduced subspace where $V = B_0^\top X \in \mathbb{R}^4$. This setting introduces two pairs of highly correlated covariates in the reduced subspace.

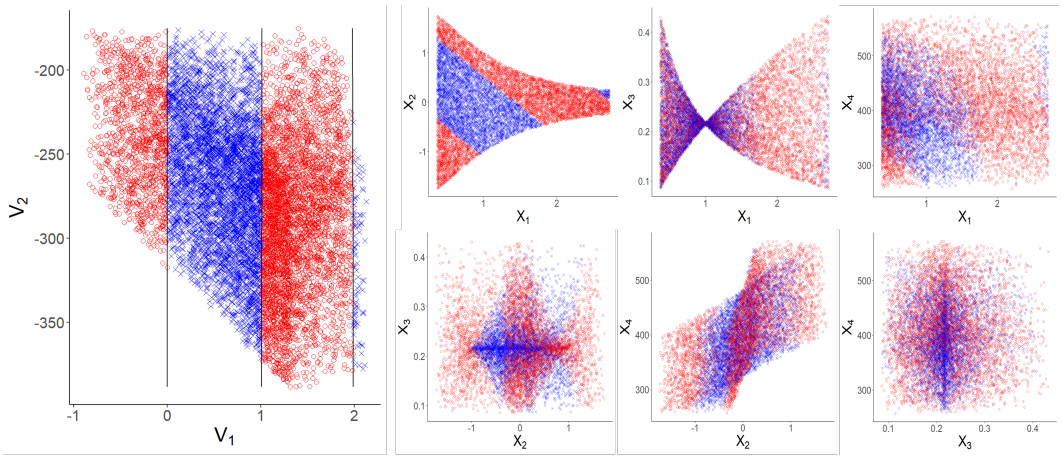

Figure 2.2: Scatterplot of the Bayes optimal treatment regimes for different values of the covariates generated under Setting 1. The red circles and blue crosses represent data points from the test dataset, corresponding to $d^*(X) = -1$ (or $d^*(V_0) = -1$) and $d^*(X) = +1$ (or $d^*(V_0) = +1$), respectively. The solid lines in the left plot display the decision boundaries in the dimension-reduced space, defined as $\left\{v = \left(v^{(1)}, v^{(2)}\right) : \tilde{f}^*(v) = 0\right\}$.

2.5.2 Results

We show the results for $n = 1{,}000$. The results for $n = 500$ are provided in Appendix A.10.1, showing similar trends. Figure 2.3 displays the results for the randomized scenario. DOL-O refers to the proposed method using the oracle $g(x)$ defined as $\mathbb{E}[\{1/\pi(A, X) - 1\}Y|X = x]$ which is an alternative formulation of $g(x)$ originally proposed to be used in the AOL by Zhou and Kosorok (2017). DOL-L uses $g_n(x)$ that estimates $g(x)$ with linear regression as described in Section 2.3.2. MAVE-L implements the proposed method by using interaction MAVE for dimension reduction and uses $g_n(x)$. AOL-O represents the estimator from AOL that uses the oracle $g(x)$. From Figure 2.3, the DOL outperforms the existing methods in Setting 2 and 3. MAVE-L performs best in Setting 1, illustrating the effectiveness of SDR for ITR estimation. But, it is outperformed by DOL in Settings 2 and 3. In particular, MAVE-L exhibits substantially higher variability in Setting 3. DOL-L performs close to DOL-O, indicating that the proposed method that uses $g_{\hat{w}}(x)$ performs close to using the true form of the oracle $g(x)$.

The results for the non-randomized scenario are in Figure 2.4. We considered two cases for the DOL: the estimator that uses the oracle $g(x)$ (KCB-O), and another where $g_{\hat{w}}(x)$ is obtained via linear regression (KCB-L). Interaction MAVE used the KCB weights and $g_{\hat{w}}(x)$

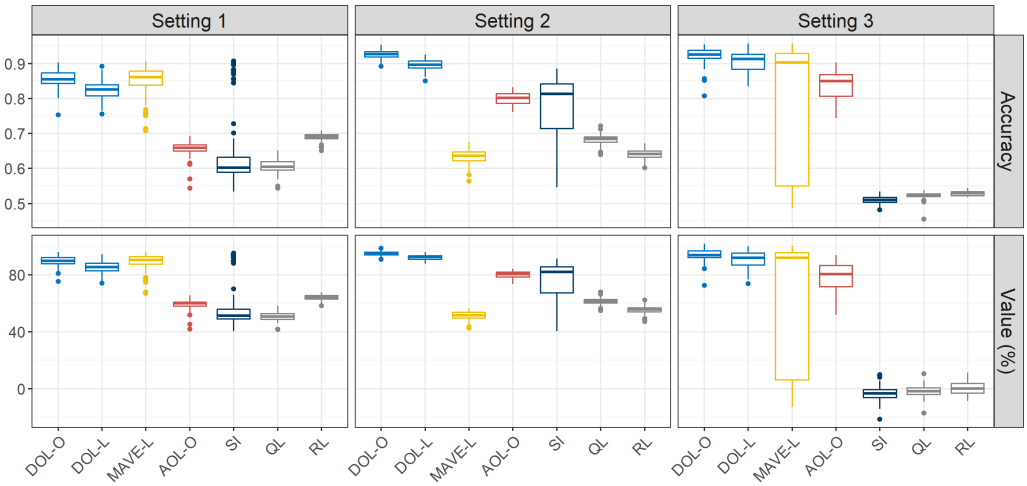

Figure 2.3: Simulation results for $n = 1{,}000$ from the randomized scenarios across the three different settings. Accuracy represents the proportion of correctly predicted optimal treatments. Value (%) represents the value function recovered by the estimated decision rule as a percentage of the Bayes optimal value function. The performance of the proposed method is shown under DOL-O and DOL-L: DOL-O uses the oracle $g(x)$, while DOL-L uses $g_{\hat{w}}(x)$ estimated via linear regression. MAVE-L uses interaction MAVE for dimension reduction and uses the fitted values $g_n(x)$ obtained via linear regression.

(K-MAVE-L). We considered logistic regression for IPW using the oracle $g(x)$ (IPW-O), and using estimated $g_{\hat{w}}(x)$ (IPW-L). The SDR and ITR were learned by using the misspecified IPW. IPW-L misspecifies $g(x)$ since it uses misspecified IPW estimator. We also considered XGBoost and a neural network (Shang et al. 2025) (XGBoost-L, NN-L). The estimated IPW was used in place of $\hat{w}_i$ in $g_{\hat{w}}(x)$ for IPW-L, XGBoost-L, and NN-L. AOL was implemented using the true propensity scores and the oracle $g(x)$ (AOL-O). AOL uses no dimension reduction, SI imposes a single-index, and IPW-O estimates B_0 with misspecified IPW.

First, we can observe that using KCB weights (KCB-O) outperforms using inverse propensity scores from a misspecified logistic regression model (IPW-O). IPW-O tends to achieve higher accuracy and value function than KCB-L on occasion which is expected as the AOL is a doubly robust method, so that using the oracle $g(x)$ suffices for universal consistency. However, it shows unstable performance and is outperformed by KCB-L in Setting 3. IPW-L shows notable declines in both accuracy and value function when the $g(x)$ is estimated by using the propensity scores estimated from a misspecified logistic regression model.

KCB-L demonstrates strong performance across all three settings and outperforms XGBoost-

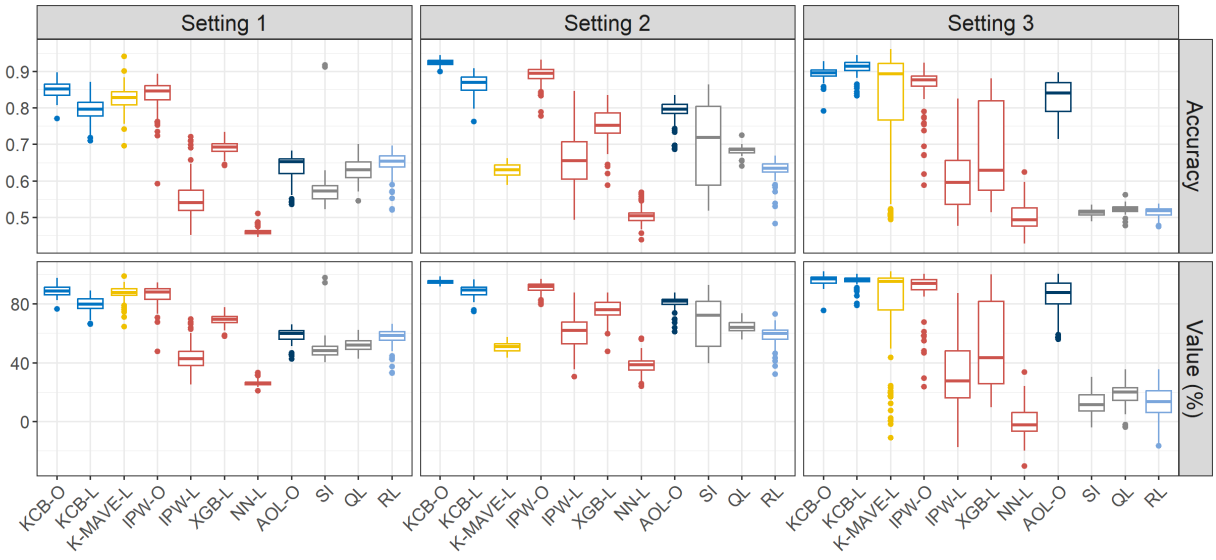

Figure 2.4: Simulation results for $n = 1{,}000$ from the non-randomized scenarios across the three different settings. Accuracy represents the proportion of correctly predicted optimal treatments. Value (%) represents the value function recovered by the estimated decision rule as a percentage of the Bayes optimal value function. The performance of the proposed method is shown under KCB-O and KCB-L: KCB-O uses the oracle $g(x)$, while KCB-L uses the fitted values of $g_{\hat{w}}(x)$ obtained via linear regression. MAVE-L uses interaction MAVE for dimension reduction and uses KCB weights along with the fitted values of $g_{\hat{w}}(x)$ obtained via linear regression. IPW-O uses the oracle $g(x)$, and IPW-L uses $g_{\hat{w}}(x)$ from linear regression, with propensity scores estimated via logistic regression. XGB-L uses $g_{\hat{w}}(x)$ from linear regression, with propensity scores estimated via XGBoost. NN-L uses $g_{\hat{w}}(x)$ from linear regression, with propensity scores estimated via a neural network.

L which shows the best performance among the methods that estimate IPW and use $g_{\hat{w}}(x)$. This suggests that KCB is an effective weighting strategy for use in non-randomized studies. Despite using $g_{\hat{w}}(x)$ which is an estimator of $g(x)$, KCB-L outperformed AOL-O, which used the true form of the oracle $g(x)$ and the true inverse propensity weights. In Setting 3, KCB-L achieves higher accuracy compared to KCB-O; however, this improvement does not translate into a much higher value function.

2.6 Application

We illustrate our proposed method by investigating ITR recommendations for using transthoracic echocardiography (TTE) in intensive care unit (ICU) sepsis patients. We retrieved data on 6,361 patients from the MIMIC-III database (Johnson et al. 2016), as described in Feng et al. (2018). In the dataset, TTE was performed on 3,262 patients (51.3%) during or within 24 hours before ICU admission, and the remaining 3,099 (48.7%) did not receive TTE. The outcome is binary, indicating 28-day survival (1 if survived, 0 if not). We used 35 baseline variables (see Appendix A.11). Missing values were imputed by using MissForest (Stekhoven and Bühlmann 2012).

To estimate an optimal ITR, we used the proposed DOL, the AOL, ℓ_1-PLS (QL), and the decision rule that prescribes TTE for every patient (Treat all), which represents the average treatment effect of TTE. We randomly sampled 1,000 subjects without replacement as a training set to learn ITR. The remaining 5,361 subjects served as a validation set by comparing the 28-day survival between the treatments (TTE or non-TTE) that match the estimated ITR and those that differ. We repeated this procedure over 100 Monte Carlo replications. At each iteration, we evaluated the modified value function (Chen et al. 2024), $\mathbb{E}[Y\{d(X)\} - Y\{-d(X)\}]$, with the ITR estimated from the training set. In our data, this indicates the improvement in the 28-day survival rate. See Appendix A.11.1 for the details on learning ITR and the evaluation of the modified value function.

The average modified value function of treating all patients with TTE is around 3.69%, and the modified value functions are substantially higher in the AOL and the DOL results, with averages of 4.95% and 6.72%, respectively (Figure 2.5). This suggests the existence of treatment effect heterogeneity and that treatment assignment based on a decision rule

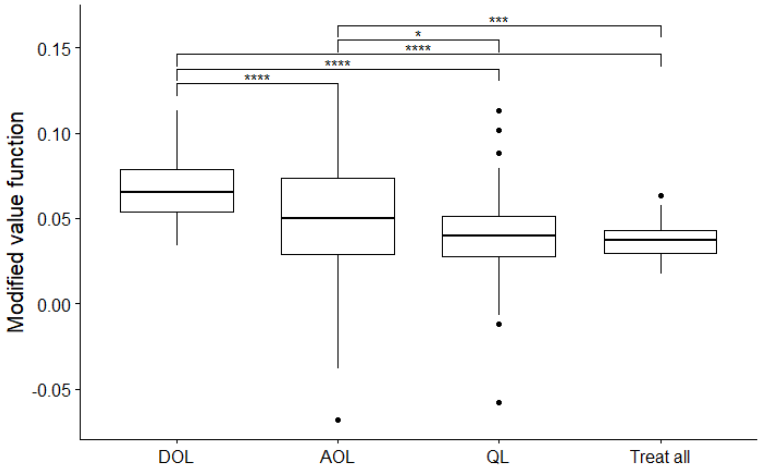

Figure 2.5: Modified value functions over the 100 repeats. DOL represents the proposed method. AOL used the raw patient characteristics for learning ITR and fitting logistic regression to estimate IPW. QL is the ℓ_1-PLS. "Treat all" represents the average effect of TTE. P-values from paired t-tests are displayed using asterisks: **** < 0.0001; *** < 0.001; * < 0.05.

that considers the individual patient characteristics could improve the survival rate. The proposed method further improves upon AOL, which demonstrates the benefit of dimension reduction combined with KCB weighting for learning ITR in observational studies.

2.7 Conclusion

We propose a novel gradient kernel dimension reduction approach that identifies a sufficient subspace for modeling heterogeneous treatment effects, enabling more accurate estimation of optimal ITRs using direct approaches. To account for differences in covariate distributions across treatment groups, we incorporate covariate balancing weights from the KCB method. This allows treatment assignment to depend on the full set of covariates. We then formulate the problem of learning an optimal ITR as a weighted SVM problem on the reduced representation of the covariates.

Theoretical guarantees are provided for the proposed procedure, which required, among other results, a new bound on the mean squared error of the balancing weights from the KCB method and a generalization of gKDR to the setting of a pseudo-outcome defined using estimated weights. In simulation studies, we demonstrated that the proposed method

is effective and exhibits excellent performance in both randomized and observational settings. The real data analysis with ICU sepsis patients corroborates the simulation results.

The proposed method involves several hyperparameters. Existing tuning strategies were effective in our numerical studies, whereas the development of a systematic tuning procedure is left for future work. The balancing weights are critical for reliable SDR and ITR. Balancing methods over even larger function classes could be considered for a future study.

Assumption 3 is crucial, but cannot be verified from the observed data. It can be assessed indirectly by comparing the predictive performance of models with and without SDR, as in the real data analysis. The proposed approach assumes that the data contain no missing values. A possible extension of this work is to adapt the proposed method, and specifically gKDR, to handle missing data without relying on imputation. Other extensions include accommodating multi-valued treatments and dynamic treatment regimes.

Chapter 3

Envelope Models for High-Dimensional Mediation Analysis

3.1 Introduction

Causal mediation analysis often involves high-dimensional mediators, particularly in areas such as genomics and imaging. Dimension reduction provides a principled approach to addressing this high dimensionality and may enable the construction of a linear lower-dimensional representation of the mediator M, which we denote as $G^\top M$, that preserves the relevant mediation effect. Specifically, such a representation aims to capture the components of the mediator that are associated with both the exposure and the outcome.

A body of prior work focuses on unsupervised linear projections, such as (sparse) PCA (Huang and Pan 2016; Zhao et al. 2020), to identify a low-dimensional representation of the mediator. Supervised dimension reduction methods have also been studied. Chén et al. (2018) has introduced a directions of mediation (DM) that reduces the mediators to orthogonal components in a way that maximizes the likelihood from the linear structural equation model (LSEM). LASSO-based methods are also widely used to find meaningful mediators. Zhao and Luo (2022) introduce pathway LASSO. Gao et al. (2019) suggest debiased LASSO to adjust for the bias. Nonlinear dimension reduction approaches have also been proposed. For example, Nath et al. (2023) uses machine learning to estimate a proper nonlinear transformation of the mediator. Bayesian methods have also been adopted

to address high-dimensional mediation analysis (Xu et al. 2026). However, these methods do not explicitly separate the material variation in the mediator that is relevant to both the exposure and the outcome from the immaterial variation that is unrelated to either, potentially reducing estimation efficiency.

Envelope model provides an effective framework for projecting either multivariate outcomes or predictors onto lower-dimensional subspaces while retaining the information relevant to estimation of the association between outcomes and predictors. Through this dimension reduction, envelope methods are known for their potential to improve estimation efficiency. Cook, R. D. et al. (2010) initially introduced the response envelope model for multivariate regression analysis which linear transforms the high-dimensional response to a lower dimension. Su and Cook (2011) introduce the partial envelope that accounts for the adjusting covariate. Cook et al. (2013a) introduced the predictor envelope model, which aims to reduce the dimensionality of the predictor while retaining the relevant association with the response.

In this chapter, we propose an envelope-based approach for estimating mediation effects within the linear structural equation modeling (LSEM) framework under normality assumptions. The proposed method identifies the subspace corresponding to the intersection of the predictor and response envelopes. Figure 3.1 illustrates the proposed framework. Our aim is to find the reducing subspace spanned by the columns of the semi-orthogonal basis matrix G such that $G^\top M$ captures the variation in M relevant to both the exposure A and the outcome Y. The remaining variation in mediator is immaterial in that it provides no additional information about the exposure or the outcome beyond the material component $G^\top M$. Envelope models have shown the potential to enhance efficiency (Cook, R. D. et al. 2010; Su and Cook 2012, 2011). Likewise, the proposed dimension reduction has the potential to substantially improve estimation efficiency by removing variation that is irrelevant to the mediation mechanism.

We show that the reducing subspace is the maximizer of the joint likelihood of outcome and mediator models in the LSEM. Moreover, we show that the projection to the reducing subspace that maximizes the likelihood is $\sqrt{n}$-consistent. Furthermore, we propose a method that uses singular value decomposition (SVD) to obtain the initial estimates of the reducing subspaces. We also verify the initial estimates are $\sqrt{n}-$consistent.

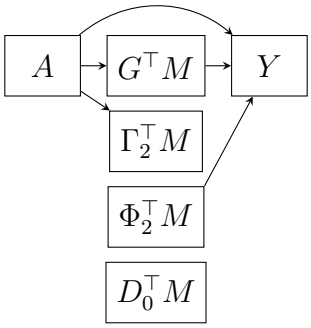

Figure 3.1: Causal DAGs for the association between exposure (A), outcome (Y), and the target transformation of mediator (M). The linear transformation, $G^\top M$, retains the mediation effect, and is distinguished from the immaterial projections: $\Gamma_2^\top M$, $\Phi_2^\top M$, and $D_0^\top M$.

An envelope space admits many equivalent basis representations. Specifically, G and GO span the same subspace for any semi-orthonormal basis matrix G and any orthogonal matrix O. This non-identifiability induces overparameterization, causing the Fisher information matrix to be singular. Shapiro (1986) introduces an asymptotic theory for overparameterized moment structure models using local regularity and generalized inverse matrices. Cook et al. (2013b), Cook, R. D. et al. (2010), and Su and Cook (2012) formulate the envelope models within the moment structural model framework to establish the asymptotic normality of the envelope estimators. Likewise, we establish the asymptotic normality of the estimators obtained from the proposed method.

The remainder of the chapter is organized as follows. We introduces the notation and assumptions underlying the proposed framework, and provide an overview of envelope model in Section 2.2. In Section 3.3, we present our proposed envelope model approach for estimating the material projection, describe the procedure for estimating the mediation effects, and establish the asymptotic properties of the resulting estimators. The simulation results are presented in Section 3.4. In Section 3.5, we analyze the Trans-Omics for Precision Medicine (TOPMed) data to demonstrate the implementation of our method. The proofs for the propositions are provided in Appendix B.

3.2 Preliminaries

Prior to presenting the proposed method, we introduce the notations and the model set up. For a random variable X, $\bar{X}$ is the sample mean, and S_X is the sample variance such that $S_X = n^{-1}\sum_{i\in[n]}(X_i - \bar{X})(X_i - \bar{X})^\top$. The covariance between two aribrary variables X_1 and X_2 is denoted by $\mathrm{Cov}(X_1, X_2)$. We use $Y|X$ to denote the conditional distribution of Y on X, and $S_{Y|X}$ is its sample variance. We use $\mathrm{span}(B)$ to denote the space spanned by the columns of B, and the projection onto $\mathrm{span}(B)$ is $P_B = B(B^\top B)^{-1}B^\top$. Furthermore, the projection onto $\mathrm{span}(B)$ with A inner product as $P_{B(A)} = B(B^\top AB)^{-1}B^\top A^\top$. We use $\|\cdot\|$ and $\|\cdot\|_F$ to denote spectral norm and Frobenius norm of a matrix, respectively, and $\|\cdot\|_2$ represents the Euclidean norm. We use 0_p or $0_{p\times q}$ to denote the zero vector of length p and $p \times q$ zero matrix, respectively. When the dimension is clear from the context, we omit the subscript for notational simplicity.

The outcome is $Y \in \mathcal{Y} \subseteq \mathbb{R}$; mediator is $M \in \mathcal{M} \subseteq \mathbb{R}^r$; exposure is $A \in \mathcal{A} \subseteq \mathbb{R}$; confounder is $X \in \mathcal{X} \in \mathbb{R}^p$. The observed data consist of $\{(A_i, Y_i, M_i, X_i) : i \in [n]\}$, where n denotes the sample size and $[n]$ denotes the index set $\{1, 2, \ldots, n\}$. Under the potential outcomes framework $\{M(a) : a \in \mathcal{A}\}$ and $\{Y(A) : A \in \mathcal{A}\}$ denote the potential mediator and the potential outcome, respectively. Furthermore, $\{Y(a, M(a)) : a \in \mathcal{A}, M(a) \in \mathcal{M}\}$ is the nested potential outcome (Pearl 2001; Robins and Greenland 1992). For the identifiability of the natural indirect and direct effects, we assume $Y(a, m) \perp\!\!\!\perp A|X$; $Y(a, m) \perp\!\!\!\perp M|(X, A)$; $M(a) \perp\!\!\!\perp A|X$; and $Y(a, m) \perp\!\!\!\perp M(a')|X$. We assume that there is no interaction between exposure and mediator. Without loss of generality, we suppose that M and A are centered to have sample mean 0. We develop our method based on the linear structural equation model (LSEM) framework introduced in Baron and Kenny (1986) under which we aim to evaluate mediation effect (Robins and Greenland 1992) from the following model set up:

$$M = \alpha_1 A + \epsilon_M, \tag{3.1a}$$

$$Y = \beta_0 + \beta_1 M + \beta_2 A + \epsilon_Y. \tag{3.1b}$$

where $\epsilon_M \sim N(0_r, \Sigma_{M|A})$ and $\epsilon_Y \sim N(0, \sigma_Y^2)$ are independent.

A subspace $\mathcal{R}$ is the reducing subspace of a matrix $\Sigma \in \mathbb{R}^{r\times r}$ when $\Sigma = P_\mathcal{R}\Sigma P_\mathcal{R} + (I_r - P_\mathcal{R})\Sigma(I_r - P_\mathcal{R})$, where I_r is the $r \times r$ identity matrix and $P_\mathcal{R}$ is the projection operator to

$\mathcal{R}$. For a semi-orthonormal matrix B, the Σ-envelope of span(B), $\mathcal{E}_\Sigma(B)$, is defined as the smallest reducing subspace of Σ that contains span(B). We use envelope models to find the reducing subspaces of $\Sigma_{M|A}$. Throughout, we focus on the scenario in which $n > r$ and r is fixed. The proposed method can be extended to $n < r$ by adopting pre-screening methods such as the envelope component screening (ECS) (Zhang and Cook 2018) or by preprocessing the mediators and retrieve $r'(< n)$ principal components. Rimal et al. (2019) showed that using the principal components performs well for predictor envelope model.

3.2.1 Review of envelope model

In this subsection, we introduce the envelope models based on the formulation in (3.1). Further details on the envelope models can be found in Cook (2018a) and Lee and Su (2020).

Response envelope

First, we use the response envelope model to perform dimension reduction on M based on (3.1a). The response envelope model aims to find the projection that distinguishes the material part of M, $P_\Gamma M$ from the immaterial part, $Q_\Gamma M$, so that (i) $\text{Cov}(P_\Gamma M, Q_\Gamma M \mid A) = 0$ and (ii) $\Gamma_0^\top M \perp\!\!\!\perp A$, where $\Gamma \in \mathbb{R}^{r\times d}$ $(d < r)$ is a semi-orthonormal basis matrix, $Q_\Gamma = I_r - P_\Gamma$, and $\text{span}(\Gamma_0) = \text{span}(\Gamma)^\perp$. Note that (i) is fulfilled when $P_\Gamma \Sigma_{M|A} Q_\Gamma = 0_{r\times r}$. Then, $\Sigma_{M|A} = (P_\Gamma + Q_\Gamma)\Sigma_{M|A}(P_\Gamma + Q_\Gamma)$, from which

$$\Sigma_{M|A} = P_\Gamma \Sigma_{M|A} P_\Gamma + Q_\Gamma \Sigma_{M|A} Q_\Gamma = \Gamma\Gamma^\top \Sigma_{M|A} \Gamma\Gamma^\top + \Gamma_0\Gamma_0^\top \Sigma_{M|A} \Gamma_0\Gamma_0^\top.$$

We have (ii) as long as $P_{\Gamma_0}\alpha_1 = 0_r$, hence $\text{span}(\alpha_1) \subseteq \text{span}(\Gamma)$ from which $\alpha_1 = \Gamma\eta$ for $\eta \in \mathbb{R}^u$. Therefore, the response envelope model is used to find a basis matrix Γ of the reducing subspace such that $\mathcal{E}_{\Sigma_{M|A}}(\text{span}(\alpha_1)) = \text{span}(\Gamma)$. From (3.1a), we obtain the low-dimensional $\Gamma^\top M$ which is the material part of M in that is associated with the exposure: $\mathbb{E}[\Gamma^\top M|A] = \Gamma^\top\Gamma\eta A = \eta A$. Note that the immaterial $\Gamma_0^\top M$ is invariant with respect to A: $\mathbb{E}[\Gamma_0^\top M|A] = \Gamma_0^\top\Gamma\eta A = 0_r$.

Predictor envelope

We use the predictor envelope model for the dimension reduction of M based on (3.1b). To obtain the material part of M that is associated with the response Y in presence of A, we first address the correlation between the mediator and the exposure. We reformulate (3.1b) by using the residual $R = M - \mathbb{E}[M|A] = M - \alpha_1 A$ as

$$\begin{aligned} Y &= \beta_0 + \beta_1(M - \alpha_1 A) + \beta_1(\alpha_1 A) + \beta_2 A + \epsilon_Y \\ &= \beta_0 + \beta_1 R + \beta_2^* A + \epsilon_Y, \end{aligned}$$

where $\beta_2^* = \beta_1\alpha_1 + \beta_2$. Then (3.1b) can be reformulated as

$$\tilde{Y} = \beta_0 + \beta_1 R + \epsilon_Y, \tag{3.2}$$

where $\tilde{Y} = Y - \beta_2^* A$. Then, the predictor envelope model finds the projection, $\Phi^\top R$ ($\Phi \in \mathbb{R}^{r\times q}$; $q \leq r$), that satisfies (iii) $\text{Cov}(Y, \Phi_0^\top R \mid \Phi^\top R) = 0$ and (iv) $\text{Cov}(\Phi_0^\top R, \Phi^\top R) = 0$, where $\Phi_0^\top R$ is the immaterial part of R and $\text{span}(\Phi_0) = \text{span}(\Phi)^\perp$. As long as $\text{span}(\beta_1^\top) \subseteq \text{span}(\Phi)$ (iii) holds from which $\text{Cov}(Y, \Phi_0^\top M|\Phi^\top M, A) = 0$. By construction, $R = \epsilon_M$ from which $\Sigma_R = \Sigma_{M|A}$. Moreover, (iv) holds as long as $\text{Cov}(\Phi^\top R, \Phi_0^\top R) = \Phi^\top\Sigma_R\Phi_0 = \Phi^\top\Sigma_{M|A}\Phi_0 = 0$, and this implies $\text{Cov}(\Phi^\top M, \Phi_0^\top M|A) = 0$. Let $Q_\Phi = I_r - P_\Phi$. Then, $\Sigma_{M|A} = (P_\Phi + Q_\Phi)\Sigma_{M|A}(P_\Phi + Q_\Phi)$, hence

$$\Sigma_{M|A} = P_\Phi\Sigma_{M|A}P_\Phi + Q_\Phi\Sigma_{M|A}Q_\Phi = \Phi\Phi^\top\Sigma_{M|A}\Phi\Phi^\top + \Phi_0\Phi_0^\top\Sigma_{M|A}\Phi_0\Phi_0^\top.$$

Therefore, the predictor envelope model is used to find a basis Φ of the reducing subspace such that $\mathcal{E}_{\Sigma_{M|A}}(\text{span}(\beta_1^\top)) = \text{span}(\Phi)$ so that $Y = \beta_0 + \xi^\top\Phi^\top M + \beta_2 A + \epsilon_Y$.

3.3 Methods

As illustrated in Figure 3.1, we aim to identify the dimension reduction $G^\top M \in \mathbb{R}^u$ that represents the material component of the mediator that is associated with both the exposure (A) and the outcome (Y). We use that $\text{span}(\Gamma)$ is composed of the two exclusive subspaces:

span(G) and span(Γ_2), where $G \in \mathbb{R}^{r\times u}$ and $\Gamma_2 \in \mathbb{R}^{r\times d_2}$ are semi-orthonormal basis matrices, and $d = u + d_2$. This implies that $\Gamma_2^\top M$ is associated with A but not with Y. Likewise, span(Φ) can be decomposed as to the union of two exclusive subspaces span(G) and span(Φ_2), where $\Phi_2 \in \mathbb{R}^{r\times q_2}$ is a semi-orthonormal basis matrix for $q = u + q_2$. This implies that $\Phi_2^\top M$ is associated with Y but not with A. Then span(G) can be retrieved as the intersection span(Γ) $\cap$ span(Φ). Since span(G), span(Γ_2), and span(Φ_2) are pairwise orthogonal reducing subspaces, the corresponding basis matrices satisfy: $G^\top\Gamma_2 = 0_{u\times d_2}$, $G^\top\Phi_2 = 0_{u\times q_2}$, and $\Phi_2^\top\Gamma_2 = 0_{r_2\times d_2}$. We can further specify the reducing subspace of $\Sigma_{M|A}$, span(D_0), where $D_0 \in \mathbb{R}^{r\times(r-u-d_2-q_2)}$ is a semi-orthonormal basis matrix such that $D_0^\top M$ is associated with neither A nor Y and D_0 is orthogonal to G, Γ_2, and Φ_2. From this, (3.1) can be reformulated as

$$M = G\eta_1 A + \Gamma_2\eta_2 A + \epsilon_M, \tag{3.3a}$$

$$Y = \beta_0 + \xi_1^\top G^\top M + \xi_2^\top \Phi_2^\top M + \beta_2 A + \epsilon_Y \tag{3.3b}$$

with the variance decomposition

$$\Sigma_{M|A} = G\Omega_G G^\top + \Gamma_2\Omega_1\Gamma_2^\top + \Phi_2\Omega_2\Phi_2^\top + D_0\Omega_0 D_0^\top, \tag{3.4}$$

where $\Omega_G = G^\top\Sigma_{M|A}G$, $\Omega_1 = \Gamma_2^\top\Sigma_{M|A}\Gamma_2$, $\Omega_2 = \Phi_2^\top\Sigma_{M|A}\Phi_2$, and $\Omega_0 = D_0^\top\Sigma_{M|A}D_0$. This formulation implies that $\Phi_0^\top M$ retains the variation in M associated with A and $\Gamma_0^\top M$ retains the variation associated with A and Y. This observation leads to the following inclusion relationship.

Proposition 3.3.1. *Let $\Gamma_0^\top M$ and $\Phi_0^\top M$ denote the immaterial components of M from the response and predictor envelope models in Section 3.2.1. Then*

$$\text{span}(\Gamma_2) \subseteq \text{span}(\Phi_0) \quad \textit{and} \quad \text{span}(\Phi_2) \subseteq \text{span}(\Gamma_0).$$

We show in Appendix B.1 that Proposition 3.3.1 implies that $G^\top M$, $\Gamma_2^\top M$, $\Phi_2^\top M$, and $D_0^\top M$ are mutually uncorrelated given A. Consequently, (3.4) represents an orthogonal decomposition of $\Sigma_{M|A}$ in which all cross-covariance terms are zero.

Remark 3.3.2. Under the nested potential outcomes framework (Pearl 2001; Robins and Greenland 1992), the potential outcome under $A = a$ and $M(a') = m$ is $Y(a, m)$. When the error terms are normal, we have the mutual independence $(G^\top M \perp\!\!\!\perp \Gamma_2^\top M \perp\!\!\!\perp \Phi_2^\top M \perp\!\!\!\perp D_0^\top M) \mid A$. Then natural indirect effect is defined based on $G^\top M$, so that

$$E[Y(a, M(a'))] - E[Y(a, M(a))] = \int E[Y|A = a, G^\top M = m_1]\{dP(m_1|A = a') - dP(m_1|A = a)\}.$$

We further illustrate the details in Appendix B.7.

3.3.1 Estimation for $\mathrm{span}(G)$, $\mathrm{span}(\Gamma_2)$, and $\mathrm{span}(\Phi_2)$

We estimate the basis matrix G over Grassmannian manifold that consists of u-dimensional subspaces of $\mathbb{R}^r$. Define $S^*_{R|\tilde{Y}} = S_{R|\tilde{Y}}\left(I_r - P_{\hat{G}(S_{R|\tilde{Y}})}\right)$; $S^{*-1}_M = S^{-1}_M\left(I - P_{\hat{G}\left(S^{-1}_M\right)}\right)$; $\check{S}_{R|\tilde{Y}} = S_{R|\tilde{Y}}\left(I_r - P_{\hat{\Phi}_2(S_{R|\tilde{Y}})}\right)$; and $\check{S}^{-1}_M = S^{-1}_M\left(I_r - P_{\hat{\Gamma}_2\left(S^{-1}_M\right)}\right)$. The estimate for the basis matrix G can be obtained by minimizing either one of the two objective functions:

$$\begin{aligned} Q_1(G, \Gamma_2, \Phi_2) = &\log\left|G^\top S^{-1}_M G\right| + \log\left|G^\top S_{R|\tilde{Y}} G\right| + \log\left|\Phi_2^\top S^{-1}_R \Phi_2\right| + \log\left|\Phi_2^\top S^*_{R|\tilde{Y}} \Phi_2\right| \\ &+ \log\left|\Gamma_2^\top S_{M|A} \Gamma_2\right| + \log\left|\Gamma_2^\top S^{*-1}_M \Gamma_2\right|; \end{aligned}$$

$$\begin{aligned} Q_2(G, \Gamma_2, \Phi_2) = &\log|\Phi_2^\top S_{R|\tilde{Y}} \Phi_2| + \log|\Phi_2^\top S^{-1}_R \Phi_2| + \log|\Gamma_2^\top S_{M|A} \Gamma_2| + \log|\Gamma_2^\top S^{-1}_M \Gamma_2| \\ &+ \log|G^\top \check{S}_{R|\tilde{Y}} G| + \log|G^\top \check{S}^{-1}_M G|. \end{aligned}$$

We derive the objective functions based on the likelihood functions from the normality of ϵ_M and ϵ_Y in Appendix B.8. When G is given, the objective function Q_2 can be decomposed to

$$L_1(\Gamma_2) = \log\left|\Gamma_2^\top S_{M|A} \Gamma_2\right| + \log\left|\Gamma_2^\top S^{*-1}_M \Gamma_2\right|; \; L_2(\Phi_2) = \log\left|\Phi_2^\top S^{-1}_R \Phi_2\right| + \log\left|\Phi_2^\top S^*_{R|\tilde{Y}} \Phi_2\right|.$$

Likewise, when $\hat{\Gamma}_2$ and $\hat{\Phi}_2$ are given, minimizing Q_1 is equivalent to the minimization of

$$L(G) = \log|G^\top \check{S}_{R|\tilde{Y}} G| + \log|G^\top \check{S}^{-1}_M G|.$$

In Section 3.3.1, we introduce the estimation of the coefficients η_1, η_2, ξ_1, and ξ_2 by using

the estimated basis matrices. Given the independence assumption of ϵ_M and ϵ_Y, the joint log-likelihood of the equation system (3.3) is

$$
\begin{aligned}
&-\frac{nr}{2}\log 2\pi - \frac{n}{2}\log|\Sigma_{M|A}| - \frac{1}{2}\sum_{i\in[n]}(M_i - G\eta_1 A_i - \Gamma_2\eta_2 A_i)^\top \Sigma_{M|A}^{-1}(M_i - G\eta_1 A_i - \Gamma_2\eta_2 A_i) \\
&-\frac{n}{2}\log 2\pi - \frac{n}{2}\log\sigma_Y^2 - \frac{1}{2}\sum_{i\in[n]}\frac{(Y_i - \beta_0 - \xi_1^\top G^\top M_i - \xi_2^\top \Phi_2^\top M_i - \beta_2 A_i)^2}{\sigma_Y^2}.
\end{aligned}
\tag{3.5}
$$

The following proposition aligns the optimization of Q_1 or Q_2 with the conventional mediation analysis that uses the joint-likelihood (3.5).

Proposition 3.3.3. *Let $(\hat{G}, \hat{\Gamma}_2, \hat{\Phi}_2)$ denote the estimates obtained by minimizing either Q_1 or Q_2. Let $(\hat{\eta}_1, \hat{\eta}_2, \hat{\xi}_1, \hat{\xi}_2)$ be the corresponding maximum likelihood estimators conditional on these estimated envelope bases. Then the resulting estimators, $(\hat{G}, \hat{\Gamma}_2, \hat{\Phi}_2, \hat{\eta}_1, \hat{\eta}_2, \hat{\xi}_1, \hat{\xi}_2)$, are the maximum likelihood estimators of the joint log-likelihood in* (3.5) *under the proposed envelope parameterization.*

Proposition 3.3.3 establishes the likelihood framework used in Section 3.3.2 to derive the asymptotic distribution of the envelope estimators. Furthermore, the minimizing Q_1 or Q_2 identifies the target reduced subspaces.

Proposition 3.3.4. *When the true covariances $\Sigma_{R|\tilde{Y}}$, Σ_M, and $\Sigma_{M|A}$, are given, the semi-orthogonal basis matrices G, Γ_2, and Φ_2 that minimize Q_1 or Q_2 satisfy*

$$\text{span}(G) = \text{span}(\Gamma)\cap\text{span}(\Phi);\ \text{span}(\Gamma_2) = \text{span}(\Gamma)\backslash\text{span}(G);\ \textit{and}\ \text{span}(\Phi_2) = \text{span}(\Phi)\backslash\text{span}(G).$$

Existing optimization strategies for envelope model estimation use objective functions involving log-determinants of quadratic forms with symmetric positive definite matrices (Zhang and Cook 2018; Zhang et al. 2023). However, the log-determinant terms in L_1, L_2, and L do not satisfy this property – they use singular positive semi-definite matrices. In addition, existing methods typically estimate a single semi-orthonormal basis matrix, while the remaining covariance parameters admit closed-form maximum likelihood estimators. In contrast, minimizing L_1, L_2, and L requires the estimation of multiple semi-orthonormal basis matrices and covariance parameters.

Computation details

In this subsection, we describe the computational approach for estimating G. By using the initial estimates of Γ_2 and Φ_2, we perform a column-wise optimization while imposing the orthogonality constraints as in the 1D algorithm (Cook and Zhang 2016).

Let $\{g_k : k \in [u], g_k \in \mathbb{R}^r\}$ be the column vectors of G so that $G = (g_1, g_2, \cdots, g_u) \in \mathbb{R}^{r\times u}$. Here we provide the procedure for minimization of L based on the initial estimates $\hat{\Gamma}_2$ and $\hat{\Phi}_2$. For $k = 0, 1, 2, \cdots, u-1$, the $k+1$-th column of $\hat{G}$ can be obtained by

$$\hat{g}_{k+1} = \text{argmin}\{\log(g_{k+1}^\top \check{S}_{R|\tilde{Y}} g_{k+1}) + \log(g_{k+1}^\top \check{S}_M^{-1} g_{k+1}) + \lambda(g_{k+1}^\top g_{k+1} - 1)\}$$

subject to

$$g_{k+1}^\top \gamma_i = 0; \forall i \in [d_2]; \quad g_{k+1}^\top \phi_i = 0; \forall i \in [q_2]; \quad g_{k+1}^\top g_i = 0; \forall i \in [k].$$

Note that the Lagrangian constraint is used to incorporate orthonormality of $\hat{G}$. We need to incorporate the orthogonality constraint within the computation. To do so, we define

$$g_{k+1}(v) = (I_r - P_{\hat{G},k} - P_{\Phi_2} - P_{\Gamma_2})v = A_k v,$$

where $P_{\hat{G},k} = \sum_{i\in[k]} \hat{g}_i \hat{g}_i^\top$. Note that $P_{\hat{G}_2,0} = 0$ and $A_k A_k = A_k$. Then

$$\begin{aligned}\hat{g}_{k+1} &= \text{argmin}_{g(v)} \left[\log\{g_{k+1}^\top(v)\check{S}_{R|\tilde{Y}} g_{k+1}(v)\} + \log\{g_{k+1}^\top(v)\check{S}_M^{-1} g_{k+1}(v)\} + \lambda\{g_{k+1}^\top(v) g_{k+1}(v) - 1\}\right] \\ &= \text{argmin}_{g(v)=A_k v}\{\log(v^\top A_k \check{S}_{R|\tilde{Y}} A_k v) + \log(v^\top A_k \check{S}_M^{-1} A_k v) + \lambda(v^\top A_k v - 1)\}.\end{aligned}$$

We still need to account for the Lagrangian constraint and that $A_k v$ is a unit vector: $\|A_k v\|_2 = 1$. These can be handled by defining $A_k v = A_k x / \|A_k x\|_2$. Then the optimization problem becomes finding x that minimizes

$$\log(x^\top A_k \check{S}_{R|\tilde{Y}} A_k x) + \log(x^\top A_k \check{S}_M^{-1} A_k x) - 2\log(x^\top A_k x).$$

From this, we obtain x by unconstrained optimization and obtain $\hat{g}_{k+1}$. The process for estimating G is described in Algorithm 1. Similar approach can be used for updating estimates

for Γ_2 and Φ_2 after estimating G.

Algorithm 1 Iteration for estimating G

Require: Φ_2 and Γ_2 given; $A_0 = 0$
  for $k = 0, 1, \cdots, u-1$ **do**
    $A_k \leftarrow I_r - \sum_{i\in[k]} \hat{g}_i \hat{g}_i^\top - P_{\Gamma_2} - P_{\Phi_2}$
    $\hat{x}_{k+1} = \text{argmin}_{x\in\mathbb{R}^u} \log(x^\top A_k \check{S}_{R|\tilde{Y}} A_k x) + \log(x^\top A_k \check{S}_M^{-1} A_k x) - 2\log(x^\top A_k x)$
    $\hat{g}_{k+1} = A_k \hat{x}_{k+1} / \|A_k \hat{x}_{k+1}\|_2$
  end for
  Return $\hat{G} = (\hat{g}_1, \hat{g}_2, \cdots, \hat{g}_u)$

Estimation of envelope parameters by SVD

In this subsection we introduce a method that uses singular value decomposition to obtain the initial estimates for Γ_2 and Φ_2. The initial estimates can be used to obtain $\check{S}_{R|\tilde{Y}}$ and $\check{S}_M^{-1}$ to minimize L with respect to G from an off-the-shelf optimization tool.

From Proposition 3.3.1, $\text{span}(\Gamma_2) \subseteq \text{span}(\Phi_0)$ and $\text{span}(\Phi_2) \subseteq \text{span}(\Gamma_0)$. This implies that $\text{span}(\Gamma_2) = \text{span}(\Gamma) \cap \text{span}(\Phi_0)$ and $\text{span}(\Phi_2) = \text{span}(\Phi) \cap \text{span}(\Gamma_0)$. By SVD, $\Gamma^\top \Phi_0 = U_1 \Lambda_1 V_1^\top$, where $U_1 \in \mathbb{R}^{d\times d_2}$ and $V_1 \in \mathbb{R}^{q_0 \times d_2}$ are orthonormal matrices, and $\Lambda_1 \in \mathbb{R}^{d_2 \times d_2}$ is a diagonal matrix whose nonzero entries are the singular values. Then $\Gamma\Gamma^\top \Phi_0 = (\Gamma U_1)\Lambda V_1^\top$ is a matrix whose first d_2 columns are the columns of Γ_2 and the rest are 0_r. Therefore, $\text{span}(\Gamma\Gamma^\top\Phi_0) = \text{span}(\Gamma U_1) = \text{span}(\Gamma_2)$. Similarly, we can use the singular value decomposition $\Gamma_0^\top \Phi = U_2 \Lambda_2 V_2^\top$ to obtain $\Gamma_0 U_2$ such that $\text{span}(\Gamma_0 U_2) = \text{span}(\Phi_2)$. Note that this satisfies the orthogonality condition $\Gamma_2^\top \Phi_2 = 0$ since $U_1^\top \Gamma^\top \Gamma_0 U_2 = 0$. Alternatively, we can use the singular value decomposition of $\Phi_0^\top \Gamma = V_1 \Lambda_1 U_1^\top$ and $\Phi^\top \Gamma_0 = V_2 \Lambda_2 U_2^\top$, and use $\Phi_0 V_1$ and ΦV_2 to obtain Γ_2 and Φ_2, respectively. This allows for the identification of basis matrices for $\text{span}(\Gamma_2)$ and $\text{span}(\Phi_2)$ as long as we use the true bases for $\text{span}(\Gamma)$ and $\text{span}(\Phi)$. Similar approach can be used to obtain G by using SVD $\Gamma^\top \Phi = U_G \Lambda_G V_G^\top$. Then G is the basis matrix for the subspace that satisfies $\text{span}(G) = \text{span}(\Gamma U_G) = \text{span}(\Phi V_G)$.

With finite samples, the SVD can use $\hat{\Gamma}$ and $\hat{\Phi}$ to obtain $\hat{\Gamma}_2$, $\hat{\Phi}_2$, and $\hat{G}$. The existing methods are provide $\sqrt{n}$-consistent estimators for Γ and Φ (Cook et al. 2013a; Cook and Zhang 2016). We establish in Proposition 3.3.5 that as long as we use the $\sqrt{n}$-consistent estimators for Γ and Φ, the SVD estimators of G, Φ_2, and Γ_2 are also $\sqrt{n}$-consistent.

Proposition 3.3.5. *Let* $\hat{\Gamma}_2 = \hat{\Gamma}U_1$ *be the estimate obtained from the SVD* $\hat{\Gamma}^\top\hat{\Phi}_0 = U_1\Lambda_1 V_1$*;* $\hat{\Phi}_2 = \hat{\Gamma}_0 U_2$ *be the estimate obtained from the SVD* $\hat{\Gamma}_0^\top\hat{\Phi} = U_2\Lambda_2 V_2$*; and* $\hat{G} = \hat{\Gamma}U_G$ *be the estimate obtained from the SVD* $\hat{\Gamma}^\top\hat{\Phi} = U_G\Lambda_G V_G$*. Then* $\|P_{\hat{\Gamma}_2} - P_{\Gamma_2}\|_F = O_p(1/\sqrt{n})$*,* $\|P_{\hat{\Phi}_2} - P_{\Phi_2}\|_F = O_p(1/\sqrt{n})$*, and* $\|P_{\hat{G}} - P_G\|_F = O_p(1/\sqrt{n})$*. The same holds when* $\hat{\Gamma}_2 = \hat{\Phi}_0 V_1$*,* $\hat{\Phi}_2 = \hat{\Phi}V_2$*, and* $\hat{G} = \hat{\Phi}V_G$*.*

However, the SVD estimates do not necessarily maximize (3.5). In Appendix B.9, we present simulation results to show that this approach may yield an inflated variability of the NIE estimates compared to the proposed method that solves for the optimization problem L_2 and L.

Estimation of NIE and NDE

From the estimated semi-orthonormal matrices $\hat{G}$ and $\hat{\Phi}_2$, we estimate for ξ_1 and ξ_2, and η_1:

$$\hat{\xi}_1 = \hat{G}^\top S_{M|A}^{-1} S_{R\tilde{Y}}; \quad \hat{\xi}_2 = \hat{\Phi}_2^\top S_{M|A}^{-1} S_{R\tilde{Y}}; \quad \hat{\eta}_1 = \hat{G}^\top S_{MA} S_A^{-1}.$$

We show that these are the maximum likelihood estimates for ξ_1, ξ_2, and η_1 in Appendix B.8. The natural indirect effect, $\beta_1\alpha_1$, becomes

$$NIE = (\xi_1^\top G^\top + \xi_2^\top \Phi_2^\top)(G\eta_1 + \Gamma_2\eta_2) = \xi_1^\top G^\top G\eta_1 + \xi_1^\top G^\top \Gamma_2\eta_2 + \xi_2^\top \Phi_2^\top G\eta_1 + \xi_2^\top \Phi_2^\top \Gamma_2\eta_2 = \xi_1^\top \eta_1.$$

This is estimated as $\widehat{NIE} = \hat{\xi}_1^\top\hat{\eta}_1$. The natural direct effect is estimated by either regressing Y on $(\hat{\xi}_1^\top\hat{G}^\top + \hat{\xi}_2^\top\hat{\Phi}_2^\top)M$ and A or subtracting the $\widehat{NIE}$ from the total effect. Note that $\hat{\Gamma}_2$ and $\hat{\eta}_2$ for the evaluation of neither mediation effects.

3.3.2 Asymptotic properties

In this subsection, we introduce the asymptotic properties of our proposed NIE estimator based on the optimization of (3.5). Let θ be the vector of envelope parameters,

$$\theta = \left(\eta_1^\top, \eta_2^\top, \xi_1^\top, \xi_2^\top, \beta_2, \mathrm{vec}^\top(G), \mathrm{vec}^\top(\Gamma_2), \mathrm{vec}^\top(\Phi_2), \mathrm{vech}^\top(\Omega_G), \mathrm{vech}^\top(\Omega_1), \mathrm{vech}^\top(\Omega_2), \mathrm{vech}^\top(\Omega_0), \sigma_Y^2\right)^\top.$$

Define the estimable function

$$h(\theta) = (h_1(\theta), h_2(\theta), h_3(\theta), h_4(\theta), h_5(\theta))^\top = \left(\alpha_1^\top, \beta_1, \beta_2, \text{vech}^\top(\Sigma_{M|A}), \sigma_Y^2\right)^\top$$
$$= \left((G\eta_1 + \Gamma_2\eta_2)^\top, (G\xi_1 + \Phi_2\xi_2)^\top, \beta_2, \text{vech}^\top(G\Omega_G G^\top + \Gamma_2\Omega_1\Gamma_2^\top + \Phi_2\Omega_2\Phi_2^\top + D_0\Omega_0 D_0^\top), \sigma_Y^2\right)^\top$$

Under (3.1), the corresponding Fisher information, $J \in \mathbb{R}^{\{2r+r(r+1)/2+2\}\times\{2r+r(r+1)/2+2\}}$, is

$$J = \begin{pmatrix} \sigma_A^2\Sigma_{M|A}^{-1} & 0 & 0 & 0 & 0 \\ 0 & (\Sigma_{M|A} + \alpha_1\alpha_1^\top\sigma_A^2)\sigma_Y^{-2} & \sigma_A^2\sigma_Y^{-2}\alpha_1 & 0 & 0 \\ 0 & \sigma_A^2\sigma_Y^{-2}\alpha_1^\top & \sigma_A^2/\sigma_Y^2 & 0 & 0 \\ 0 & 0 & 0 & 0.5E_r^\top(\Sigma_{M|A}^{-1}\otimes\Sigma_{M|A}^{-1})E_r & 0 \\ 0 & 0 & 0 & 0 & 0.5\sigma_Y^{-4} \end{pmatrix}$$

where $\otimes$ denotes the Kronecker product and $E_r \in \mathbb{R}^{r^2\times r(r+1)/2}$ is the expansion matrix and $C_r \in \mathbb{R}^{r(r+1)/2\times r^2}$ is the contraction matrix such that $E_r\text{vech}(U) = \text{vec}(U)$ and $C_r\text{vec}(U) = \text{vech}(U)$ for an $r \times r$ symmetric matrix U. Let $r_0 = r - u - d_2 - q_2$ and $U^{\otimes 2} = U \otimes U$ for an arbitrary matrix U. The Jacobian of $h(\theta)$, H, is

$$H = \begin{pmatrix} G & \Gamma_2 & 0 & 0 & 0 & \eta_1^\top\otimes I_r & \eta_2^\top\otimes I_r & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & G & \Phi_2 & 0 & \xi_1^\top\otimes I_r & 0 & \xi_2^\top\otimes I_r & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & \nabla_G h_3 & \nabla_{\Gamma_2} h_3 & \nabla_{\Phi_2} h_3 & C_r G^{\otimes 2} E_u & C_r\Gamma_2^{\otimes 2}E_{d_2} & C_r\Phi_2^{\otimes 2}E_{q_2} & C_r D_0^{\otimes 2}E_{r_0} & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \end{pmatrix}$$

where $\nabla_G h_3 = \partial h_3(\theta)/\partial\text{vec}(G)^\top$, $\nabla_{\Gamma_2} h_3 = \partial h_3(\theta)/\partial\text{vec}(\Gamma_2)^\top$, and $\nabla_{\Phi_2} h_3 = \partial h_3(\theta)/\partial\text{vec}(\Phi_2)^\top$. Define $G_0 = \begin{pmatrix}\Gamma_2 & \Phi_2 & D_0\end{pmatrix}$; $\Gamma_2^0 = \begin{pmatrix}G & \Phi_2 & D_0\end{pmatrix}$; $\Phi_2^0 = \begin{pmatrix}G & \Gamma_2 & D_0\end{pmatrix}$, and $\Omega_G^0 = G_0^\top\Sigma_{M|A}G_0$; $\Omega_1^0 = \Gamma_2^{0\top}\Sigma_{M|A}\Gamma_2^0$; $\Omega_2^0 = \Phi_2^{0\top}\Sigma_{M|A}\Phi_2^0$. Then $h_3(\theta) = \text{vech}(G\Omega_G G^\top + G_0\Omega_G^0 G_0^\top) = \text{vech}(\Gamma_2\Omega_1\Gamma_2^\top + \Gamma_2^0\Omega_1^0\Gamma_2^{0\top}) = \text{vech}(\Phi_2\Omega_2\Phi_2^\top + \Phi_2^0\Omega_2^0\Phi_2^{0\top})$. From this, we have $\nabla_G h_3 = 2C_r(G\Omega_G\otimes I_r - G\otimes G_0\Omega_G^0 G_0^\top)$, $\nabla_{\Gamma_2} h_3 = 2C_r(\Gamma_2\Omega_1\otimes I_r - \Gamma_2\otimes\Gamma_2^0\Omega_1^0\Gamma_2^{0\top})$, and $\nabla_{\Phi_2} h_3 = 2C_r(\Phi_2\Omega_2\otimes I_r - \Phi_2\otimes\Phi_2^0\Omega_2^0\Phi_2^{0\top})$.

Since H is overparamterized, the information matrix with respect to θ is singular and the usual asymptotic normality results for maximum likelihood estimators is not applicable. Instead, we utilize the fact that $h(\theta)$ is estimable and use generalized inverse to find the

asymptotic variance by using the results from Proposition 4.1 of Shapiro (1986) to establish the approximate distribution of $h(\hat{\theta})$. Let avar$(\cdot)$ denote an asymptotic variance.

Proposition 3.3.6. *Let $\tilde{\alpha}_1$ and $\tilde{\beta}_1^\top$ be the MLE estimators. Under the LSEM represented by* (3.3) *with the normal errors and known envelope dimension u, we have*

$$\sqrt{n}\left(h(\hat{\theta}) - h(\theta)\right) \to N(0, P_{H(J)}J^{-1}P_{H(J)}^\top).$$

Moreover, $J^{-1} - P_{H(J)}J^{-1}P_{H(J)}^\top \succeq 0$ and $\text{avar}(\sqrt{n}\tilde{\beta}_1\tilde{\alpha}_1) - \text{avar}(\sqrt{n}\hat{\beta}_1\hat{\alpha}_1) \geq 0$.

Proposition 3.3.6 implies that the envelope estimator of the NIE is at least as efficient as the MLE estimator. Note that $P_{H(J)}J^{-1}P_{H(J)}^\top = H(H^\top JH)^\dagger H^\top$, where $\dagger$ denotes the Moore-Penrose inverse.

We can establish the asymptotic normality of $\sqrt{n}(\hat{\theta} - \theta)$ based on Proposition 4.2 of Shapiro (1986). Since span(Γ) and span(Φ) are estimable, span(G) = span$(\Gamma) \cap$ span(Φ) is estimable. Thus, $P_G\alpha_1 = G\eta_1$ and $P_G\beta_1 = G\xi_1$ are estimable and we can establish the normal approximation of the material components of coefficients α_1 and $\beta_1^\top$.

Proposition 3.3.7. *Under the LSEM represented by* (3.3) *with the normal errors and known envelope dimension u, we have*

$$\sqrt{n}\left\{\begin{pmatrix}\hat{G}\hat{\eta}_1\\ \hat{G}\hat{\xi}_1\end{pmatrix} - \begin{pmatrix}G\eta_1\\ G\xi_1\end{pmatrix}\right\} \to N\left(0, \begin{pmatrix}G & 0 & \eta_1^\top \otimes I_r\\ 0 & G & \xi_1^\top \otimes I_r\end{pmatrix}\Pi_1\begin{pmatrix}G & 0 & \eta_1^\top \otimes I_r\\ 0 & G & \xi_1^\top \otimes I_r\end{pmatrix}^\top\right).$$

Proposition 3.3.7 can be useful for analyzing the component-wise contribution of each mediator to the NIE. For statistical inference, we can consider a similar testing procedure as in Huang and Pan (2016) to perform resampling from the asymptotic normal distribution.

Asymptotics without normality

Normal approximation also holds under non-normal errors. Let $\mathbb{W} \in \mathbb{R}^{n\times(r+1)}$ be the design matrix whose i-th row is the i-th observation of M and A: $(M_i^\top, A_i)$, and let vector $\mathbb{Y} \in \mathbb{R}^n$ be the corresponding outcome $Y_1, Y_2, \cdots, Y_n$. Let $\tilde{\alpha}_1 = S_{MA}S_A^{-1}$, $\tilde{\beta} = (\tilde{\beta}_1^\top, \hat{\beta}_2)^\top = (\mathbb{W}^\top\mathbb{W})^{-1}\mathbb{W}^\top\mathbb{Y}$, $S_{M|A}$, and $s^2_{Y|(M,A)}$ be the standard maximum likelihood estimators of $h(\theta)$.

Lemma 3.3.8. *Under the model* (3.1)*, suppose that the errors,* ϵ_Y *and* ϵ_M*, are independent and have finite fourth moments. In addition, suppose that the the maximum diagonal entry,* $\max_{i\in[n]} p_{ii}$*, of* $P_W = \mathbb{W}(\mathbb{W}^\top\mathbb{W})^{-1}\mathbb{W}^\top$ *converges to 0 as* $n \to \infty$*. Then the standard maximum likelihood estimators* $\tilde{\alpha}_1$*,* $\tilde{\beta}$*,* $\text{vech}(S_{M|A})$*, and* $s^2_{Y|(M,A)}$ *converge to a joint mean-zero normal distribution.*

In Appendix B.6, we show the explicit form of the variance, and verify that the maximization of the joint-likelihood (3.5) can be formulated as the minimum discrepancy function in Shapiro (1986). From this and Lemma 3.3.8 we can establish that the envelope estimators, $h(\hat{\theta})$, are asymptotically normally distributed even without the normality assumption, using an argument analogous to that under normal errors, while retaining their efficiency gains. Furthermore, we establish that the envelope parameters, θ, and the estimators of the material components for the NIE, $G\eta_1$ and $G\xi_1$, also approximate normal distribution without the normal errors.

Proposition 3.3.9. *Given the conditions in Lemma 3.3.8, we have that* $\sqrt{n}\left(h(\hat{\theta}) - h(\theta)\right)$ *and* $\sqrt{n}\left((\hat{G}\hat{\eta}_1 - G\eta_1)^\top, (\hat{G}\hat{\xi}_1 - G\xi_1)^\top\right)^\top$ *converge to mean zero normal distribution.*

The expressions for the asymptotic variances of $\sqrt{n}\{h(\hat{\theta})\}$ and $\sqrt{n}\left(\hat{\eta}_1^\top\hat{G}^\top, \hat{\xi}_1^\top\hat{G}^\top\right)^\top$ are given in Appendix B.6.2.

3.3.3 Selecting dimensionality u

In envelope models, there are three different approaches for selecting the dimensionality u of the reducing subspace span(G): sequential likelihood ratio test, information criterion such as the Akaike information criterion (AIC) and the Bayesian information criterion (BIC), and k-fold cross-validation. However the likelihood ratio test, AIC, and BIC are applicable when the number of parameters to estimate differs at different values of u. However, in Appendix B.8.5, we show that the number of parameters is $r^2/2 + r + d + q + 3$ for our model and does not depend on u. Furthermore, using k-fold cross-validation is challenging because u is determined jointly by the mediator and outcome models. For example, selecting u to minimize a combined prediction error requires balancing prediction performance across two distinct models. Since deriving a systematic method for selecting u is challenging, we instead propose a reasonable ad hoc approach rather than a fully rigorous selection procedure.

Singular values for choosing u

An alternative approach for tuning is to use the singular values of $\Gamma^\top\Phi$. Given the true Γ and Φ, the singular values are either zero or one, and the overlap dimension is equal to the multiplicity of the singular value of one (Figure 3.2). We use the singular values of $\hat{\Gamma}^\top\hat{\Phi}$ to select the dimension $\hat{u}$. As in Figure 3.2, the first several singular values of $\hat{\Gamma}^\top\hat{\Phi}$ are near one followed by smaller singular values.

Eigen-gap based approaches have been used in spectral clustering and signal detection (Ding and Yang 2022; Onatski 2009) where the number of clusters or signals is determined based on the maximum of eigen-gap ratios. We use a heuristic singular-value gap ratio criterion to estimate u. Let $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_p$ be the singular values of $\hat{\Gamma}^\top\hat{\Phi}$, where $p = \min(d, q)$. We can first obtain the gaps between the singular values as $d_i = \lambda_i - \lambda_{i+1}$ for $i \in [p-1]$, and obtain the gap ratios as d_i/d_{i-1}. Then we choose $\hat{u}$ as $\hat{u} = \text{argmax}_{i\in\{2,3,\cdots,p-1\}}\{d_i/d_{i-1}\}$. Since this rule may select $\hat{u}$ based on relatively small singular values, we additionally apply a threshold to exclude selected dimensions whose singular values are not sufficiently close to one. Specifically, we retain only dimensions above a preselected threshold. In our implementation, we use a threshold of 0.9. For example, if $\lambda_2 < 0.9$ while $\lambda_1 > 0.9$ then $\hat{u} = 1$. Furthermore, if the eigen-gap ratio criterion yields $\hat{u} = p - 1$, we set $\hat{u} = p$ whenever $\lambda_p > 0.9$.

3.4 Simulations

In this section, we evaluate the finite-sample performance of the mediation analysis that uses proposed dimension reduction. We study three settings that correspond to different correlation structures between mediators. Figure 3.3 shows the three correlations. We generate response, mediator, and exposure based on the association in (3.3). The variance decomposition as in (3.4) was used. We study the performance when the sample size is $n = 100,\ 200,\ 500,\ 1,000$ and $M \in \mathbb{R}^{50}$. Within each setting, we performed 50 iterations. In all settings, we considered binary exposure that takes value of $A = 1$ with probability $1/2$.

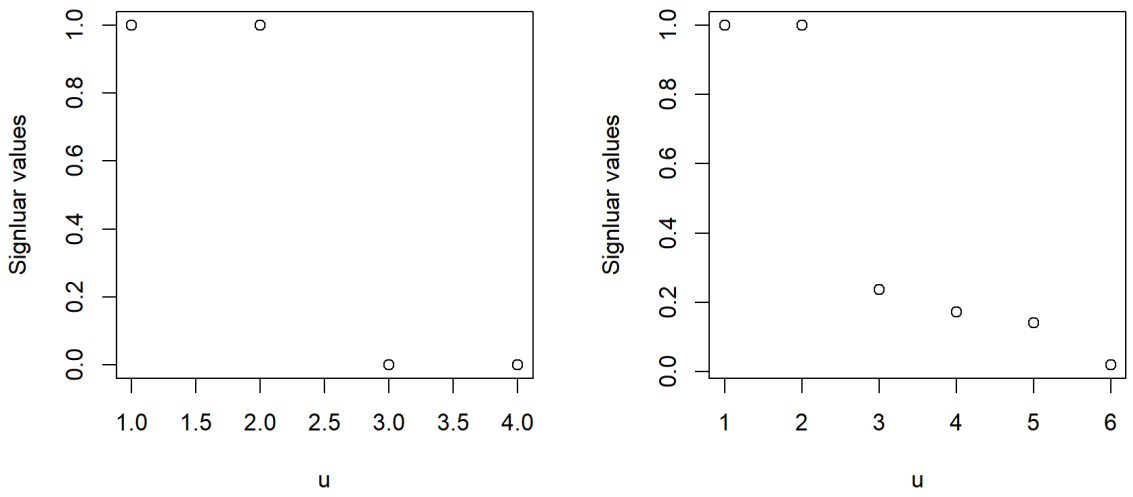

Figure 3.2: An example of singular values of $\hat{\Gamma}^\top \hat{\Phi}$. Using the threshold 0.01 for the relative difference from the largest singular value chooses $u = 2$.

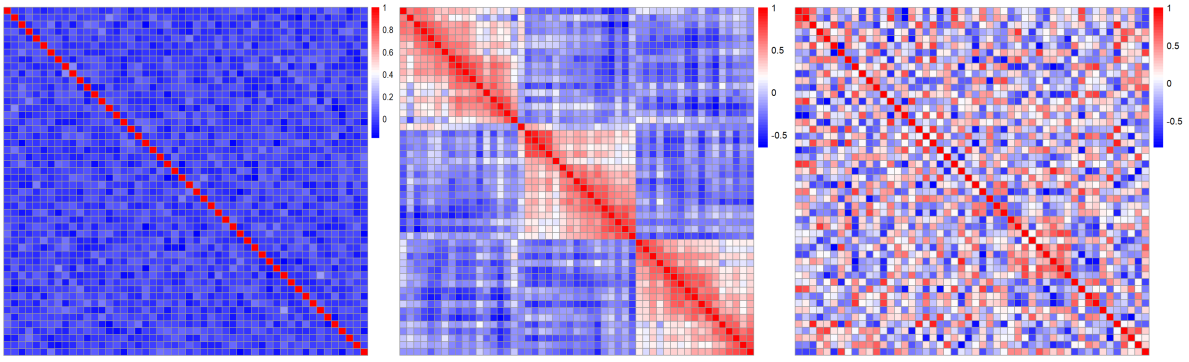

Figure 3.3: Mediator correlations for the three different settings. The left plot corresponds to Setting 1; the middle plot corresponds to Setting 2; the right plot corresponds to Setting 3.

3.4.1 Setting 1

Setting 1 considers $M \in \mathbb{R}^{50}$, $u = 2$, $d_2 = 10$, and $q_2 = 10$. The variance, $\Sigma_{M|A}$, is decomposed by $\Omega_G \in \mathbb{R}^{2\times 2}$ is the diagonal matrix whose entries take values from 0.1 to 0.2; $\Omega_1 \in \mathbb{R}^{10\times 10}$ is the diagonal matrix whose entries take values from 50 to 60; $\Omega_2 \in \mathbb{R}^{10\times 10}$ is the diagonal matrix whose entries take values from 30 to 40; and $\Omega_0 \in \mathbb{R}^{28\times 28}$ is the diagonal matrix whose entries take values from 90 to 100. The coordinate vectors $\eta_1 \in \mathbb{R}^u$ and $\eta_2 \in \mathbb{R}^{d_2}$ were generated from $Unif(1, 3)$; $\xi_1 \in \mathbb{R}^u$ was generated from $Unif(1, 3)$ and $\xi_2 \in \mathbb{R}^{q_2}$ was generated from $Unif(1, 3)$.

3.4.2 Setting 2

Setting 2 considers $M \in \mathbb{R}^{51}$, $u = 2$, $d_2 = 5$, and $q_2 = 2$. We construct $\Sigma_{M|A}$ as a block-diagonal correlation matrix, where mediators are grouped into three correlated clusters with 17 mediators in each, exhibiting an AR(1)-like structure. We first generate a 51×51 block-diagonal matrix V where each block is a 17×17 correlation matrix whose entries are $0.9^{|i-j|}$ for $i, j \in \{1, 2, \cdots, 17\}$. Then $\Omega_G \in \mathbb{R}^{2\times 2}$ is $0.1 \cdot G^\top VG$; $\Omega_1 \in \mathbb{R}^{2\times 2}$ is $200 \cdot \Gamma_2^\top V\Gamma_2$; $\Omega_2 \in \mathbb{R}^{5\times 5}$ is $100 \cdot \Phi_2^\top V\Phi_2$; and $\Omega_0 \in \mathbb{R}^{42\times 42}$ is $1000 \cdot D_0^\top VD_0$. The coordinate vectors $\eta_1 \in \mathbb{R}^u$ and $\eta_2 \in \mathbb{R}^{d_2}$ were generated from $Unif(1, 3)$; $\xi_1 \in \mathbb{R}^u$ was generated from $Unif(1, 3)$ and $\xi_2 \in \mathbb{R}^{q_2}$ was generated from $Unif(1, 3)$.

3.4.3 Setting 3

In Setting 3, we study the case when the five largest eigenvalues of $\Sigma_{M|A}$ comprise 90% of total variance. We use $M \in \mathbb{R}^{50}$, $u = 2$, $d_2 = 3$, and $q_2 = 5$. We first generate a matrix whose singular values take five 100 and forty five 0.5. By using this, we construct an $N \times 50$ matrix, where $N = 10000$. We also generate an error matrix, $E \in \mathbb{R}^N$, whose entries are from normal distribution with mean 0 and standard deviation 0.1. Then we add the two generated matrices, and retrieve the covariance, V. From this, we compute $V_G = G^\top VG$, $V_1 = \Gamma_2^\top V\Gamma_2$, $V_2 = \Phi_2^\top V\Phi_2$, and $V_0 = D_0^\top VD_0$. Then we generate $\Omega_G \in \mathbb{R}^{2\times 2}$ by setting $\Omega_G = V_G$; $\Omega_1 \in \mathbb{R}^{3\times 3}$ by setting $\Omega_1 = 2000 \cdot V_1$; $\Omega_2 \in \mathbb{R}^{5\times 5}$ by setting $\Omega_2 = 1000 \cdot V_2$; and $\Omega_0 \in \mathbb{R}^{40\times 40}$ by setting $\Omega_0 = 20000 \cdot V_0$. The coordinate vectors $\eta_1 \in \mathbb{R}^u$ and $\eta_2 \in \mathbb{R}^{d_2}$ were generated from $Unif(1, 5)$ and $Unif(2, 5)$, respectively; $\xi_1 \in \mathbb{R}^u$ was generated from

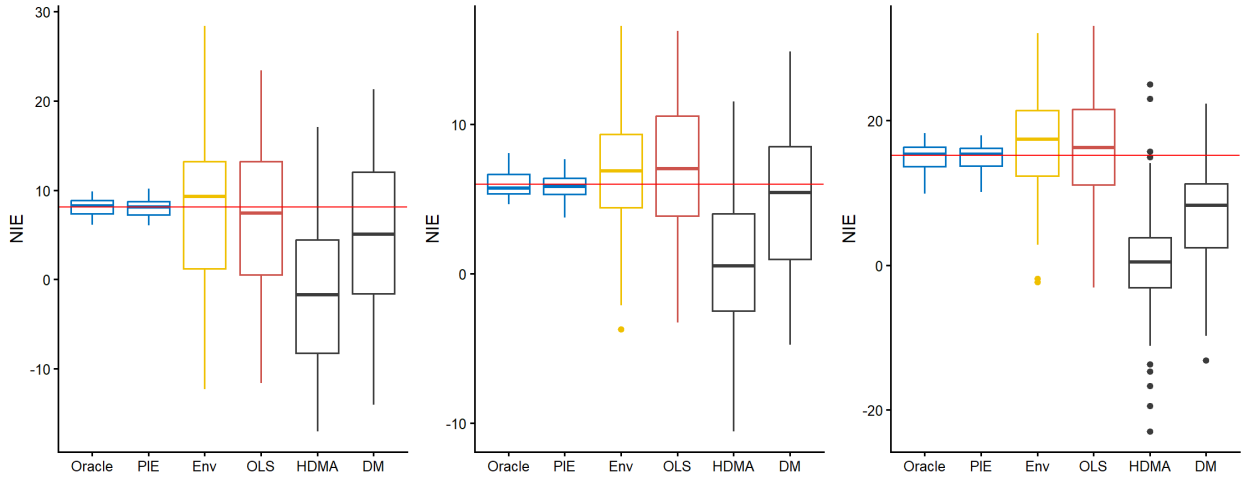

Figure 3.4: The distributions of the estimated natural indirect effects from different methods. The left plot corresponds to Setting 1; the middle plot corresponds to Setting 2; the right plot corresponds to Setting 3.

$Unif(1,5)$ and $\xi_2 \in \mathbb{R}^{q_2}$ was generated from $Unif(1,3)$.

3.4.4 Results

The simulation results for $n = 100$ are presented in Figure 3.4. More results can be found in Appendix B.9.1. OLS performs mediation analysis based on the multivariate linear regression analysis. HDMA uses the debiased-lasso method from Gao et al. (2019). DM uses the direction of mediation proposed in Chén et al. (2018). Env estimates the coefficients α and β based on the outcome and predictor envelope models, respectively, and evaluates the mediation effects based on $\hat{\Gamma}\hat{\eta}$ and $\hat{\Phi}\hat{\xi}$. Tuning for the dimension of Γ and Φ were based on the BIC criterion – we chose d and q that minimize the BIC in the response and predictor envelope models. The proposed method estimates the projection onto the intersection of the predictor and response envelopes and is referred to as PIE (Projection onto the Intersection of Envelopes). Broyden–Fletcher–Goldfarb–Shanno algorithm was used for the optimization of L_2 and L from which we retrieved $\hat{G}$, $\hat{\xi}_1$, and $\hat{\eta}_1$ to estimate NIE. Oracle represents the proposed method that uses the true Γ_2 and Φ_2, and the true dimensions u, d_2, and q_2 to perform the optimization of L and estimate $\hat{G}$.

As shown in Figure 3.4, the proposed method, PIE, yielded substantially lower variability in the estimated NIEs than the existing methods across all settings. This finding highlights

the benefit of distinguishing mediator variation that is relevant to the mediation pathway from immaterial variation. Env also separates the material and immaterial components of the mediators; however, its material subspace retains directions relevant to either the exposure-to-mediator or the mediator-to-outcome relationship. The simulation results suggest that retaining these directions may not be sufficient for precise estimation of the NIE. In contrast, PIE further targets the intersection of the subspaces relevant to these two relationships, resulting in lower variability in the estimated NIEs.

Additional Evaluation of Dimension Selection and Projection Convergence

In the Appendix B.9.2, we additionally present results using a hard-thresholding approach for selecting u to assess the robustness of the proposed method to the choice of the dimensionality. Specifically, we consider threshold values of 0.7 and 0.99 and estimate u by counting the singular values of $\hat{\Gamma}^\top \hat{\Phi}$ that exceed the corresponding threshold. Furthermore, we also investigate the performance of the SVD-based approach for estimating NIE discussed in Section 3.3.1. Figure B.4 displays the resulting NIE estimates under the different dimension-selection strategies. The results shows the robustness of the proposed method to the selection of u, whereas SVD-based approach is sensitive to the choice of u and exhibits greater variability. In addition, Appendix B.9.3 investigates how the Frobenius error $\|P_{\hat{G}} - P_G\|_F$ decreases with increasing sample size. The results indicate that the error decreases at an approximately $\sqrt{n}$-rate.

3.5 Application

We implement the proposed method onto the Trans-Omics for Precision Medicine (TOPMed) study. We study the mediation effect of Olink proteomic biomarkers in the pathway from diabetes to the kidney function measured by estimated Glomerular Filtration Rate (eGFR, mL/min/1.73m^2). The dataset consists of 5,888 adults aged 65 or older, and 28 demographic and clinical characteristics from the Cardiovascular Health Study collected during years 2-11 and year 18 following study enrollment. There are 5,201 participants in the original cohort and 687 participants in the supplemental cohort. We use the data from the original cohort. The proteomic assay results from the Olink Explore HT panel were used. The assays consist

of the total of 5,420 proteins. The proteomic biomarkers are quantified using normalized protein expression, which represents protein abundance on $\log_2$ scale.

Previous studies have reported strong interconnection between kidney function, inflammation, and cardiovascular disease (CVD) (Arnlov 2009; Ronco et al. 2008). Based on this, Carlsson et al. (2017) studied association between kidney function (eGFR) and the proteomic biomarkers linked to CVD and inflammation included in the Olink Proseek Multiplex CVD I panel. The list is often extended to include the proteins in the Olink Target 96 CVD II, CVD III, and inflammatory panels (Aylward et al. 2025; Daniels et al. 2021). For the real data application, we use 275 proteins measured on the Olink Proseek Multiplex CVD I panel and the Olink Target 96 CVD II, CVD III, and Inflammation panels. Figure B.6 shows the correlation structure between these proteins.

For the mediation analysis, we applied our proposed method using diabetes status at the baseline (year 2) as the exposure, proteomic biomarkers at year 5 as mediators, and eGFR at the year 9 follow-up as the outcome. We adjusted for baseline eGFR, age, smoking status, race, and gender as confounders. Among the 5,201 participants, we used $n = 2{,}153$ for the analysis. Figure B.7 illustrates the sample selection process. The participants had on average -7.90 (mL/min/1.73m^2) change in the eGFR from year 2 to year 9 follow-up.

The dimensions $d = 13$ and $q = 18$ were selected for estimating Γ and Φ, respectively, based on the BIC. We chose $\hat{u} = 3$ based on the singular values of $\hat{\Gamma}^\top \hat{\Phi}$ that are close to 1 (Figure B.8). Bootstrap analysis was used with 500 iterations to estimate the standard error (SE) of the mediation effect estimates. The spectral norm of the material component of the mediator was $\|\hat{\Omega}_G\| = 10.54$ while that of the immaterial component $\|\hat{\Omega}^0_G\| = 54.50$. Thus, the immaterial component of the mediator has higher variance than the material component. The NIE was -1.31 (SE $= 0.388$), the NDE was -0.84 (SE $= 0.726$), and the total effect was -2.14 (SE $= 0.787$). The classic envelope model estimated NIE by using the coefficient estimates from the response and the predictor envelope models as in Env in Section 3.4. The estimated NIE was -2.08 (SE $= 0.629$) while the NDE and total effect were -1.14 (SE $= 0.725$) and -3.00 (SE $= 0.743$), respectively. The OLS estimated NIE as -3.00 (SE $= 0.743$) with the largest standard error, NDE as -0.440 (SE $= 0.916$), and the total effect as -3.44 (SE $= 0.887$). The estimate for the mediation effects from the proposed method varies depending on the choice of the dimensionality u. The estimated NIE remained

relatively stable when moderately larger values of u were selected. For larger values of u, the estimated NIE became more variable and generally more negative (Appendix B.10). We additionally considered $d = 250$ and $q = 65$ from which we chose $u = 58$ based on the singular values of $\hat{\Gamma}^\top \hat{\Phi}$. We obtained the NIE similar to the OLS, NIE $= -2.86$ (SE $= 0.627$). NDE was -0.56 (SE $= 0.768$).

The component-wise indirect effects of individual proteins, $\{\beta_{1k}\alpha_{1k} : k \in [r]\}$, were computed from the parameter estimates to assess the contribution of each protein to the overall indirect effect. Each component-wise indirect effect was standardized by dividing it by the standard deviation of the corresponding proteomic biomarker. The twenty proteins with the largest component-wise indirect effects in magnitude, as estimated by the proposed method, were selected, and their estimated effects from each method are depicted in Figure 3.5. The corresponding protein names to each UniProt ID can be found in Table B.1.

According to the proposed method, diabetes exerts a negative indirect effect on kidney function through each of the identified proteins. In contrast, the OLS, the classic envelope model, and the proposed method with over-specified overlap dimension ($u = 58$) estimates suggest that some proteins have positive indirect effects. For example, diabetes has been shown to be associated with increased expression of endothelial-leukocyte adhesion molecule-1 (ELAM-1, P16581) (Siddiqui et al. 2023). Moreover, increased expression of ELAM-1 has been reported in the kidneys of patients with diabetic nephropathy (Hirata et al. 1998), suggesting a role for ELAM-1 in the progression of diabetic kidney disease. Therefore, a negative component-wise indirect effect for ELAM-1 is consistent with its established involvement in endothelial dysfunction and renal injury. In contrast, the other approaches estimate a positive indirect effect for ELAM-1.

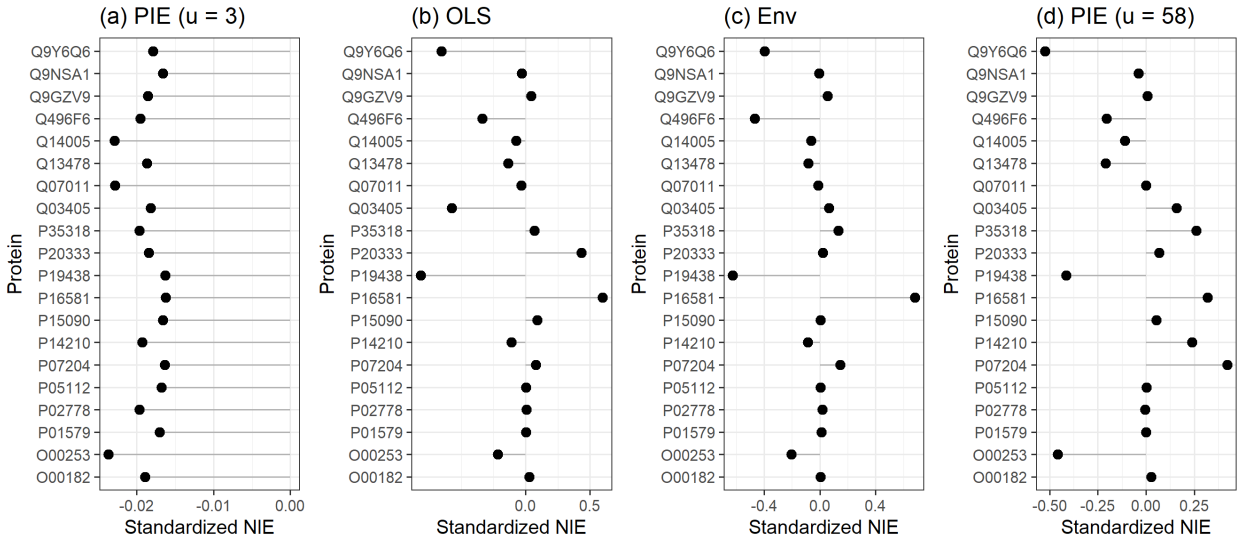

Figure 3.5: Component-wise NIEs standardized by each corresponding mediator's standard deviation. The largest 20 standardized NIEs from each method are presented. The proteins are specified by the UniProt identifier. From the left: the proposed method with $u = 3$, OLS, classic envelope, and the proposed method with $u = 58$.

Moreover, the NIE estimates from the alternative approaches tend to exhibit greater variability, with several estimates having large magnitudes. One potential contributor to this pattern is collinearity among the protein mediators, as illustrated in Figure B.6. This issue may be more evident when correlated mediators are modeled simultaneously, as in the outcome-mediator model. As shown in Figure B.9, the three alternative approaches generally produce more dispersed coefficient estimates for the outcome-mediator model than the proposed PIE method, with OLS exhibiting particularly large dispersion and several extreme estimates. In contrast, the coefficient estimates from the mediator-exposure model are more comparable between PIE and the alternative approaches. For the the classic envelope model and the PIE with $u = 58$, the greater dispersion in the outcome-mediator coefficients tend to persist as the estimated reduced subspace may not have adequately separated the material variation from the immaterial variation (Figure B.9).

3.6 Conclusion

We propose a novel envelope model approach for estimating mediation effects via dimension reduction in a high-dimensional setting. A reducing subspace of $\Sigma_{M|A}$ is identified such that the transformed mediator preserves the variation associated with both the exposure and the outcome. By restricting estimation of NIE to this material component, the proposed method removes immaterial variation and improves the accuracy and efficiency of mediation effect estimation.

We established several theoretical properties of the proposed method. In particular, under the true covariance structure, the basis matrices that optimize the objective functions span the desired reducing subspace, which captures the variation jointly associated with the exposure and the outcome. We further showed that these estimators are also the maximizers of the joint likelihood, thereby enabling the derivation of the asymptotic normality of the resulting envelope estimators with and without the normality assumption.

In simulation study, we showed that the proposed estimator has substantially lower variability than the existing methods while being unbiased. The real data analysis used TOPMed data to investigate the mediating effects of proteomic biomarkers linked to cardiovascular disease and inflammation in the pathway from diabetes status to kidney function. We also studied the mediator-specific indirect effects to find proteins that mediate larger difference in eGFR. The estimated indirect effects were consistent with the established associations of the identified proteins with both diabetes and kidney function.

Potential extension of the proposed method can adopt to various types of responses via generalized linear model as in Zhu and Su (2020). We have focused on the high-dimensional setting where the number of mediators is large but remains smaller than the sample size. This can be extended to settings with $n < r$ by applying envelope component screening (Zhang and Cook 2018), or by integrating the proposed method with the envelope regularization (Kwon and Zou 2025) and envelope-based sparse partial least squares (Zhu and Su 2020).

Chapter 4

Representation Learning for Heterogeneous Treatment Effect Estimation Using Multimodal Data

4.1 Introduction

In clinical practice, decisions are often made not only based on structured patient information such as age, sex, blood pressure, etc., but also information contained in unstructured data including medical images and clinical notes. It is therefore reasonable to expect that data-driven approaches should also be capable of incorporating multimodal data, consisting both of structured and unstructured data. Given the high dimensionality of such data, dimension reduction is often desirable.

However, linear dimension reduction may not be well suited to unstructured data, which in this chapter we focus on in the form of images. To illustrate, suppose we apply a linear dimension reduction approach, similar to that proposed in Chapter 2, to learn low-dimensional features tailored to heterogeneous treatment effect estimation. Let X_I denote the vectorized image and consider a low-dimensional representation $B^\top X_I$. Such a representation may offer some interpretability by identifying image regions that contribute strongly to the treatment effect through the weights in B. Nevertheless, applying linear dimension reduction directly to vectorized images has important limitations. Vectorization ignores spatial relationships

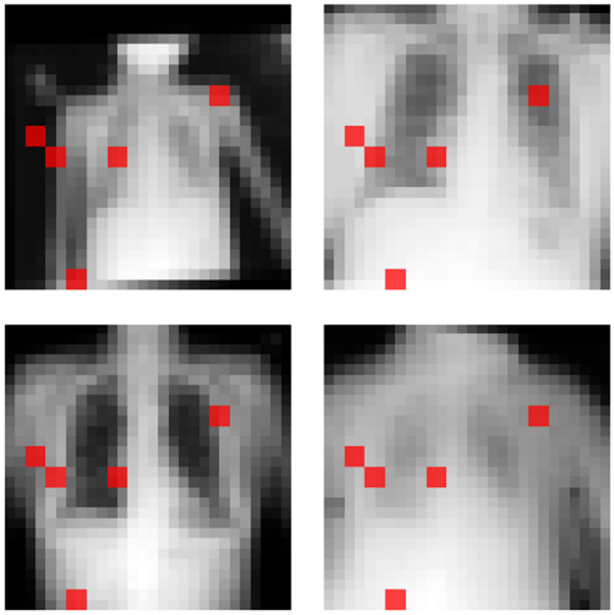

Figure 4.1: An illustration of the misalignment issue is shown here. The red squares indicate different clusters of four pixels. Pixels at the same coordinates across images may correspond to different anatomical structures. For example, the rightmost pixel corresponds to the shoulder in one individual (upper-left image) but to part of the lung in another (upper-right image).

among neighboring pixels, and corresponding features may not be aligned across individuals. For example, as shown in Figure 4.1, some pixel locations may correspond to empty regions in one chest X-ray but to anatomical structures in another, so pixels at the same coordinates do not necessarily represent the same underlying features (Figure 4.1). Therefore, image features associated with treatment effect heterogeneity are likely to exhibit nonlinear and hierarchical structure that cannot be adequately captured by a linear projection. To accommodate these characteristics, we develop a nonlinear representation-learning approach based on neural networks that preserves spatial structure while learning image representations targeted to heterogeneous treatment effect estimation.

Neural networks have gained popularity in causal inference because of their flexibility in modeling complex relationships, for instance, between confounders and outcomes. They are also a powerful tool for encoding unstructured data (Bengio et al. 2013). For example, clinical notes have been encoded using BERT (Devlin et al. 2018) and convolutional neural networks (CNN) have been used to learn representations of image data. Neural networks have

also been adopted for representation learning of structured covariates in high-dimensional settings (Clivio et al. 2024; Scholkopf et al. 2021). These recent developments motivate our use of neural networks to learn nonlinear covariate representations specifically tailored to conditional average treatment effect (CATE) estimation.

In this chapter, we denote the binary treatment as $A \in \{0, 1\}$ following the convention commonly used in the CATE literature. We adopt the no unmeasured confounder (**Assumption 2.2.1**) and positivity of propensity score (**Assumption 2.2.2**) assumptions from Chapter 2 for the identifiability of CATE.

To describe the existing methods, we classify CATE estimators other than the R-learner into two broad categories: one-step plug-in learners and pseudo-outcome regression learners. One-step plug-in learners separately estimate $\mu_1(x) = \mathbb{E}[Y(1)|X = x]$ and $\mu_0(x) = \mathbb{E}[Y(0)|X = x]$ and obtain CATE by taking their contrast, $\tau(x) = \mu_1(x) - \mu_0(x)$. On the other hand, the pseudo-outcome regression learners estimate $\tau(x)$ directly by regressing a constructed pseudo-outcome on the covariates, rather than deriving it from separately fitted models. Curth et al. (2021) shows that the one-step plug-in methods may be preferable when $\mu_1(x)$ or $\mu_0(x)$ do not follow a simple form. In particular, directly targeting $\tau(x)$ could be more effective when the response surfaces are complex but the treatment effect has a comparatively simple structure. A concrete example illustrating this limitation can be found in Nie and Wager (2021).

In this chapter, we propose JointRNet, which uses a joint R-loss function (4.1) to simultaneously estimate the nuisance functions and learn the CATE within a single end-to-end neural network rather than the two-step procedure employed by the conventional R-learner (Nie and Wager 2021). We show that the proposed joint R-loss is minimized at the oracle nuisance functions and the true heterogeneous treatment effect. However, JointRNet uses an architecture that facilitates information sharing between the nuisance and CATE function. Through extensive studies, we show that JointRNet improves on the existing two-step R-learner and performs competitively with existing CATE estimators. For this purpose, we also introduce a multimodal data-generating process that provides a benchmark for evaluating machine learning models through numerical experiments.

The rest of the chapter is organized as follows: In Section 4.2, we review related works and provide additional background. In Section 4.3, we present our proposed JointRNet

method. In Section 4.4, we describe the data-generating process used in the simulations and compare the performance of JointRNet with existing methods. In Section 4.5, we summarize our findings and provide concluding remarks.

4.2 Related work

Machine learning methods have been widely adopted for estimating treatment effect heterogeneity. Among the approaches most closely related to ours are causal forests, which provide flexible, nonparametric estimators of the CATE for structured covariates (Athey et al. 2019; Wager and Athey 2018). However, these methods do not readily generalize to unstructured covariates. Neural network-based approaches are better suited to this task and have been developed under a variety of formulations, as summarized in Table 4.1 and Table 4.2.

4.2.1 One-step plug-in approaches

Many model-agnostic CATE estimators follow a plug-in strategy: they first estimate the treatment-specific conditional mean outcome functions and then obtain the CATE from their contrast. Künzel et al. (2019) introduced T-, S-, and X-learners for estimating CATE. The T-learner is the most basic plug-in approach that separately estimates the treatment-specific outcome models using the treated and control samples and then obtains the CATE by taking the contrast between their predictions. Shalit et al. (2017) propose Treatment-Agnostic Representation Network (TARNet), which learns a representation to be used for estimating the two potential outcome models, and use their difference to estimate CATE defined as $\mu_1(x) - \mu_0(x)$. Shi et al. (2019) proposes DragonNet, a multitask architecture that uses a shared covariate representation to estimate the treatment-specific outcome models and the propensity score jointly which enables information sharing. Hassanpour and Greiner (2020) propose disentangled representations for counterfactual regression (DR-CFR), which learns separate representations for features associated with treatment assignment only, outcome only, or treatment and outcome (confounding). The representations are disentangled using an orthogonality penalty and a representation-balancing regularizer. More generally, neural network implementations of these estimators can incorporate representation learning directly into estimation (Curth and Schaar 2021).

4.2.2 Pseudo-outcome regression approaches

Pseudo-outcome regression learners often use a two step estimation procedure. In the first step, a pseudo-outcome is constructed from the estimated nuisance parameters such as $\mu_1(x)$, $\mu_0(x)$, and $\pi(x) = \Pr(A = 1|X = x)$. Then the pseudo-outcome is regressed on the covariates. One choice defines the pseudo-outcome as the contrast of inverse-propensity-weighted outcomes (PW). We refer to this approach as the PW-learner. Kennedy (2023) propose the doubly robust (DR)-learner whose pseudo-outcome is motivated by the augmented inverse propensity weighted estimator (Robins et al. 1995). Curth and Schaar (2021) propose a regression-adjusted (RA)-learner, which is a variant of X-learner, and uses a pseudo-outcome without inverse propensity weighting. This avoids the potential instability caused by propensity scores that are close to 0 or 1. The specific expressions of the pseudo-outcomes are displayed in Table 4.2.

4.2.3 R-learner

The R-learner (Nie and Wager 2021) directly estimates the CATE, $\tau(x)$, by minimizing the R-loss function, $\mathbb{E}\left[\big(Y - m(X) - \{A - \pi(X)\}\tau(X)\big)^2\right]$, which is motivated by the Robinson transformation (Robinson 1988). The estimation takes two steps. In *Step 1*, the nuisances $m(x)$ and $\pi(x)$ are estimated. In *Step 2*, $\tau(x)$ is estimated using the nuisance estimates. In both steps, flexible modern machine learning methods can be utilized.

Table 4.1: Summary of one-step plug-in learners for CATE estimation. The one-step plug-in estimators and their corresponding estimated quantities are presented.

One-step plug-in learners	
Method	**Parameters**
T-learner	$\mu_0(x), \mu_1(x)$
TARNet	$\mu_0(x), \mu_1(x)$
DragonNet	$\mu_0(x), \mu_1(x), \pi(x)$
DR-CFR	$\mu_0(x), \mu_1(x), \pi(x)$

Table 4.2: Summary of pseudo-outcome regression learners for CATE estimation. The pseudo-outcome regression learners and their associated pseudo-outcomes are presented. The two-step extensions indicate which plug-in estimators from Table 4.1 can be used in the first stage to construct the corresponding pseudo-outcomes.

Pseudo-outcome regression learners		
Method	**Pseudo-outcome**	**Two-step extensions**
Regression adjustment (RA) learner	$\widetilde{Y}_{\text{RA}} = A\{Y - \widehat{\mu}_0(X)\} + (1-A)\{\widehat{\mu}_1(X) - Y\}$	TNet+RA; TARNet+RA; DragonNet+RA; DR-CFR+RA
Propensity weighting (PW) learner	$\widetilde{Y}_{\text{PW}} = \left\{\frac{A}{\widehat{\pi}(X)} - \frac{1-A}{1-\widehat{\pi}(X)}\right\} Y$	DragonNet+PW; DR-CFR+PW
Doubly robust (DR) learner	$\widetilde{Y}_{\text{DR}} = \widehat{\mu}_1(X) - \widehat{\mu}_0(X) + \frac{A\{Y - \widehat{\mu}_1(X)\}}{\widehat{\pi}(X)} - \frac{(1-A)\{Y - \widehat{\mu}_0(X)\}}{1-\widehat{\pi}(X)}$	DragonNet+DR; DR-CFR+DR;

4.2.4 Multimodal causal inference

A growing literature incorporates unstructured data into causal inference. Veitch et al. (2020) introduce causal BERT that extends BERT (Devlin et al. 2018) to learn text embeddings for estimating ATE on the treated and the natural direct effect. Jerzak et al. (2023) introduces an approach based on Bayesian CNN for estimating treatment effect heterogeneity by using unstructured satellite image features. Their method can also be extended to incorporate structured data along with the image features. However, the approach is primarily developed for randomized designs. A related line of research considers settings in which the unstructured data encode the treatment itself rather than pretreatment effect modifiers. For example, Fong and Grimmer (2023) study the causal effects of low-dimensional latent treatment attributes embedded within high-dimensional interventions such as text. Schulte et al. (2025) use fixed representations from pretrained neural networks to adjust for non-tabular confounding in ATE estimation. They discuss CATE estimation and multimodal adjustment as potential extensions but do not develop these settings.

Liang et al. (2025) propose an approach for building a predictive model for estimating

the outcome in the presence of modalities that are missing not at random. Their method includes a data fusion step that combines text, image, and structured data while accounting for the missingness pattern. However, this approach is more centered on predicting the outcome under missingness rather than estimating a potential outcome.

4.3 Method

Let $\phi_m(x)$, $\phi_\pi(x)$, and $\phi_\tau(x)$ denote possibly nonlinear covariate representations used for modeling the conditional mean outcome, the propensity score, and the CATE, respectively. Let h_m, h_π, and h_τ be functions that map the corresponding representations to their respective quantities, i.e.,

$$m(x) = h_m\{\phi_m(x)\}, \quad \pi(x) = h_\pi\{\phi_\pi(x)\}, \quad \tau(x) = h_\tau\{\phi_\tau(x)\}.$$

For a binary treatment, the conditional mean outcome satisfies $m(x) = \mathbb{E}(Y \mid X = x) = \mu_0(x) + \pi(x)\tau(x)$. Therefore, $m(x) = h_m(x) = \mu_0(x) + h_\tau\{\phi_\tau(x)\}h_\pi\{\phi_\pi(x)\}$. This relationship shows that the conditional mean outcome depends on information relevant to the baseline outcome, propensity score, and CATE components. Motivated by this structure, we define the representation for the conditional mean outcome through the fusion $\phi_m(x) = f_{\text{fuse}}(x, \phi_\tau(x), \phi_\pi(x))$. JointRNet exploits this structure by jointly learning these components, thereby enabling information sharing across all three tasks.

Specifically, we propose a single-step implementation of R-learner that simultaneously estimates the nuisance functions and the heterogeneous treatment effect by optimizing the joint objective function:

$$\mathcal{L}(m, \pi, \tau) = \mathbb{E}\Big[\alpha_1 \mathcal{L}_Y(Y, m) + \alpha_2 \mathcal{L}_A(A, \pi) + \big(\{Y - m(X)\} - \{A - \pi(X)\}\tau(X)\big)^2\Big], \quad (4.1)$$

where $\alpha_1 > 0$ and $\alpha_2 > 0$ are tuning hyperparameters, and $\mathcal{L}_Y(Y, m)$ and $\mathcal{L}_A(A, \pi)$ are the loss functions for the conditional mean and propensity score, respectively. Typically, Fisher-consistent losses are used for $\mathcal{L}_Y$ and $\mathcal{L}_A$. That is, for P_X-almost every x, $\arg\min_{u \in \mathbb{R}} \mathbb{E}\{\mathcal{L}_Y(Y, u) \mid X = x\} = \mathbb{E}[Y \mid X = x]$ and $\arg\min_{p \in (0,1)} \mathbb{E}\{\mathcal{L}_A(A, p) \mid X = x\} = \Pr(A = 1 \mid X = x)$. For example, squared-error loss is used for the conditional mean, and binary cross-entropy loss

is used for the propensity score. We establish that the joint (unconstrained) minimizers of (4.1) are the oracle nuisances and the true CATE. The proof is provided in Appendix C.

Proposition 4.3.1. *Suppose that* $A \in \{0, 1\}$, $\mathbb{E}(Y^2) < \infty$, *and* $0 < \Pr(A = 1 \mid X = x) < 1$ *almost surely. Suppose that we use Fisher-consistent losses for* $\mathcal{L}_Y$ *and* $\mathcal{L}_A$. *Over all measurable functions for which* (4.1) *is finite, the objective is minimized by*

$$\begin{aligned} m^*(x) &= \mathbb{E}(Y \mid X = x), \\ \pi^*(x) &= \mathbb{E}(A \mid X = x) = \Pr(A = 1 \mid X = x), \\ \tau^*(x) &= \frac{\mathbb{E}[\{Y - m^*(X)\}\{A - \pi^*(X)\} \mid X = x]}{\mathbb{E}[\{A - \pi^*(X)\}^2 \mid X = x]} \\ &= \mathbb{E}(Y \mid A = 1, X = x) - \mathbb{E}(Y \mid A = 0, X = x). \end{aligned}$$

Note that the minimizing $\tau(x)$ is identical to the optimal CATE function based on the oracle nuisance functions in the two-step R-learner.

For estimation, we propose using an end-to-end neural network architecture as illustrated in Figure 4.2. The proposed architecture enables information sharing across the nuisance functions and the heterogeneous treatment effect through shared intermediate representations. Specifically, intermediate representations learned for the propensity score and the CATE are shared with the conditional mean model.

4.3.1 Estimation with multimodality

Neural networks provide a flexible framework for incorporating multimodal data, with established architectures available for learning representations from different data types. For example, CNN can extract representations from images, while transformer-based models such as BERT can encode text. The well-documented predictive performance and robustness of deep neural networks make them particularly useful for representation learning and model estimation with unstructured data (Scholkopf et al. 2021).

Figure 4.2 presents a multimodal extension of the proposed method that combines image covariates, denoted by X_I, with structured covariates, denoted by X_S. The image representation is first learned and fused with the structured covariates in the feedforward neural networks. The architecture enables information sharing across the model components, as the

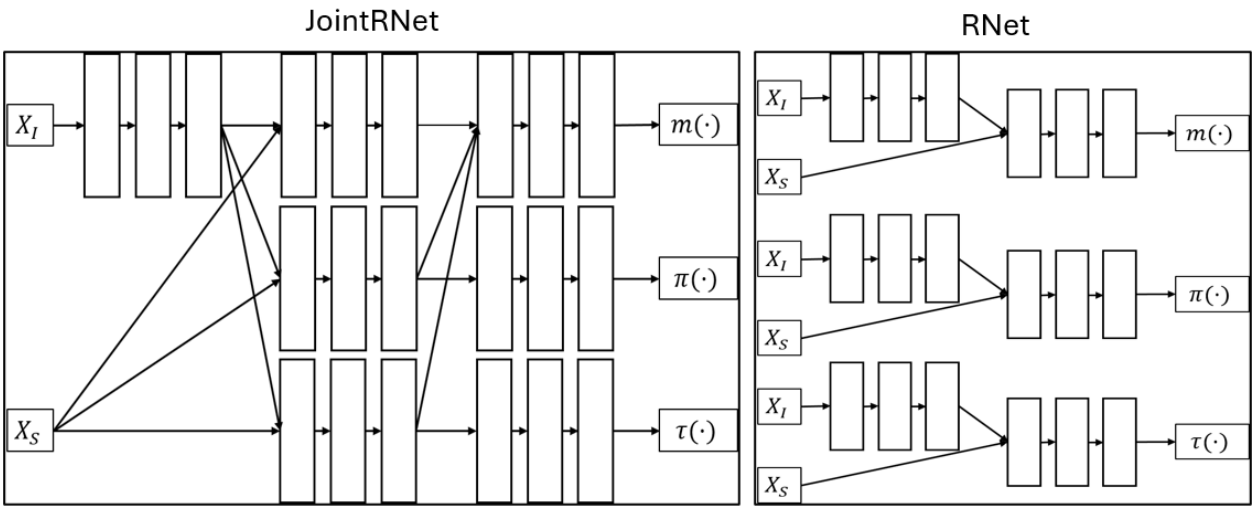

Figure 4.2: Neural network architectures for JointRNet and RNet for estimating treatment effect heterogeneity. Each architecture admits the raw image features (X_I) and the structured covariates (X_S) as input.

intermediate representations learned for the propensity score and the CATE are additionally connected to the conditional mean. The resulting representations are used to estimate the nuisance functions and the CATE.

Similar neural network architectures can also be used to implement existing CATE learners in multimodal settings. RNet implements the conventional two-stage R-learner by using separate deep learning models to obtain $\hat{m}(x)$ and $\hat{\pi}(x)$ in the first step and then uses these plug-in estimates to minimize the R-loss in the second step (Figure 4.2). As illustrated in Figure 4.2, a distinct representation of the raw image is learned within each model. The neural network implementation of the T-learner (TNet) estimates $\mu_1(x)$ and $\mu_0(x)$ using separate networks without information sharing (Figure 4.3). In contrast, other one-step plug-in methods in Figure 4.3 adopt architectures that allow varying degrees of information sharing. TARNet shares an intermediate representation between the outcome models for $\mu_1(x)$ and $\mu_0(x)$, while DragonNet additionally shares this representation with the propensity score model. DR-CFR employs a more complex architecture with multiple shared and component-specific representations.

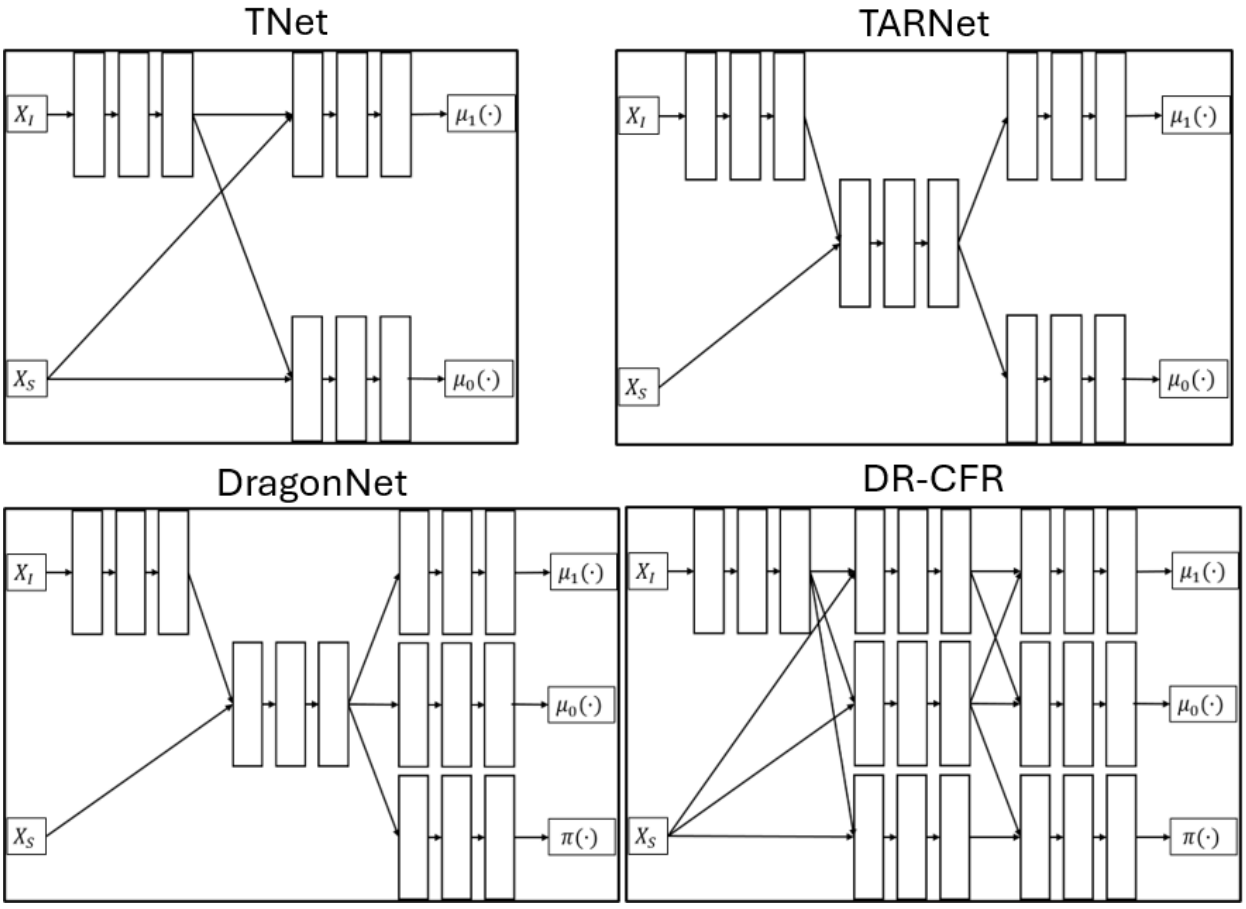

Figure 4.3: Neural network architectures for the one-step plug-in methods. Each architecture admits the raw image features (X_I) and the structured covariates (X_S) as input.

4.4 Experiments

4.4.1 Tabular data - IHDP benchmark

Prior to evaluating the proposed method on multimodal data, we first investigate the performance of different neural network architectures for estimating heterogeneous treatment effects using structured covariates. To do so, we use the semi-synthetic data generated from the Infant Health and Development Program (IHDP) data. Specifically, we use the IHDP-100 data shared by Shalit et al. (2017), which are 100 realizations of data generated based on Setting B of Hill (2011). This setting is based on the ground-truth outcome models

$$\mu_1(x) = \exp\{(x + 0.5 \cdot \mathbb{1}_{25})^\top \beta\} \quad \text{and} \quad \mu_0(x) = x^\top \beta + \omega,$$

where $x \in \mathbb{R}^{25}$, $\mathbb{1}_{25}$ is a vector of fifty ones, and $\omega \in \mathbb{R}$ take different values in each iteration to ensure that the average treatment effect on the treated is 4.

We compared the proposed method, RNet, and the learners in Table 4.1 and Table 4.2. For RNet, we used 5-fold cross-fitting to estimate the nuisance functions $m(x)$ and $\pi(x)$ by implementing separate neural networks. Then we estimate $\tau(x)$ by fitting a second-stage network that minimizes the R-loss, using the cross-fitted nuisance estimates. Each semi-synthetic dataset consists of 747 observations (139 treated and 608 control). To implement the neural network algorithms and test the estimators, we split the data into train, validation, and test sets based on the ratio of 63% ($n = 471$), 27% ($n = 201$), and 10% ($n_{test} = 75$), respectively.

The root mean squared error (RMSE) of the estimated CATE from the training set was evaluated over the 100 datasets based on the test set as

$$\sqrt{n_{test}^{-1} \sum_{i \in [n_{test}]} \{\hat{\tau}(X_i) - \tau(X_i)\}^2}.$$

Because the scale of the true CATE varies across the semi-synthetic datasets, the corresponding raw RMSE values are not directly comparable. Following Curth and Schaar (2021), we rescale the outcomes in each dataset by the standard deviation of its true CATE, thereby placing the treatment effects and RMSE values on a comparable scale. The results are pre-

sented in Figure 4.4. Because the propensity-weighted and doubly robust pseudo-outcome regression learners (Table 4.2), as well as the T-learner, exhibited large estimation errors, possibly due to unsuccessful hyperparameter tuning, their results are omitted.

JointRNet outperformed the one-step plug-in estimators considered in our study. This result differs from that of Curth et al. (2021), who found one-step plug-in approaches to perform best in the same setting and attributed their advantage to the complex structure of the treatment-specific response surfaces, $\mu_1(x)$ and $\mu_0(x)$. Consistent with their findings, however, the RNet performed relatively poorly under this data-generating process. The performance difference between JointRNet and RNet suggests that joint estimation and information sharing among the nuisance and CATE components can substantially improve the performance of the R-learner.

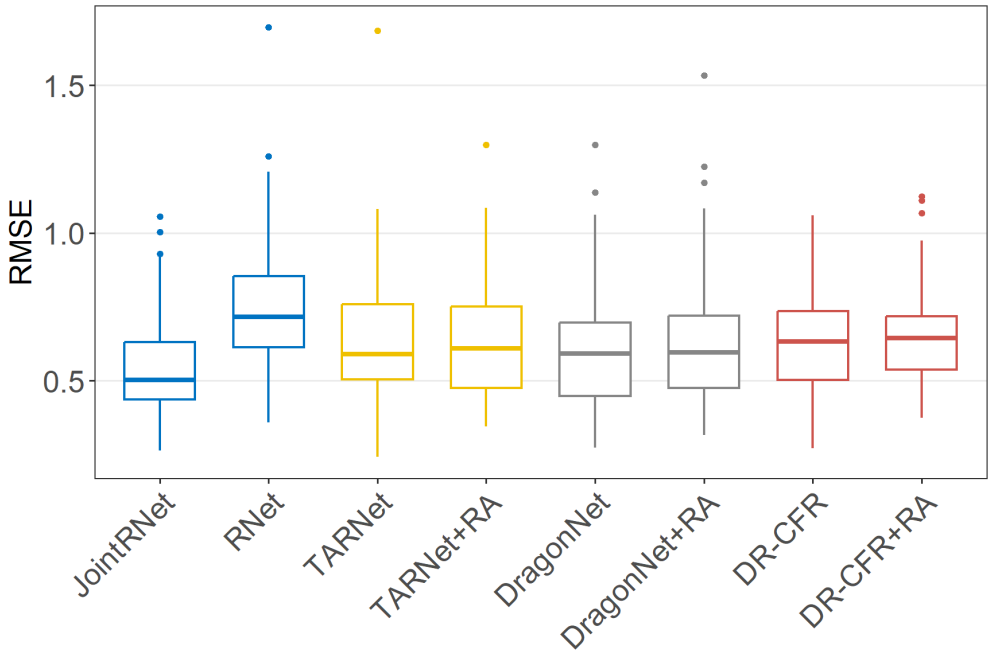

Figure 4.4: Performance comparison between different learners. The distributions of the RMSE evaluated on the test set are displayed. Results from T-learner and PW- and DR-learner extensions of TARNet, DragonNet, and DR-CFR are omitted due to large errors.

4.4.2 Multimodal data - ChestMNIST

We use ChestMNIST from the MedMNIST database (Yang et al. 2023) as the source of image covariates in simulation studies evaluating CATE estimation in the context of unstructured

covariates. These are X-ray images derived from NIH ChestX-ray14, an extension of the ChestX-ray8 dataset introduced by Wang et al. (2017), whose images have a resolution of 1024×1024 pixels. ChestMNIST contains 112,120 frontal chest X-ray images from 30,805 individuals, resized to 28×28 pixels and annotated with 14 binary indicators of thoracic conditions. The predefined data split comprises 78,468 training images, 11,219 validation images, and 22,433 test images.

In each simulation iteration, we sampled 1,000 observations without replacement from the 78,468 training observations to estimate the CATE function. We also sampled 500 observations without replacement from the 11,219 validation set for model selection and tuning. Performance was evaluated using the RMSE computed over all 22,433 test observations.

To generate the remaining variables from the image covariates in the simulation study, we use the pretrained feature representations for ChestMNIST provided by Yang et al. (2024). These representations were learned by a model trained to predict the 14 binary indicators of thoracic conditions. This gives the 512-dimensional feature representation $\Psi = \phi(X_I) \in \mathbb{R}^{512}$, where X_I denotes the raw image pixels. Then we randomly generated $B \in \mathbb{R}^{512\times 5}$ from standard normal distribution (seed = 999) to retrieve $B^\top \Psi = Z_I \in \mathbb{R}^5$. When estimating the CATE, we use the raw image X_I as the observed image covariate, while Z_I is treated as an unobserved latent representation used only to generate the simulated data.

The tabular features were generated from latent factors $Z \in \mathbb{R}^{50}$, whose components were sampled from a uniform distribution over the range $[0, 4]$. From the latent factors, the covariates $X_S \in \mathbb{R}^{50}$ are defined as $X_S^{(1)} = \exp(Z^{(1)}/2)$; $X_S^{(2)} = Z^{(2)}/\{1 + \exp(Z^{(1)})\}$; $X_S^{(3)} = (Z^{(1)} \cdot Z^{(3)}/25 + 0.6)^3$; $X_S^{(4)} = (Z^{(2)} + Z^{(4)} + 2)^2$; $X_S^{(i)} = Z^{(i)}$ for $i \in \{5, \cdots, 50\}$. The main effect was generated as

$$\mu(X) = 5 + 6Z^{(1)} + 8Z^{(2)} + 3Z^{(3)} + 5Z^{(4)} + 7Z^{(5)}.$$

The treatment effects were generated as $\beta_A^\top Z + \beta_{I,A}^\top Z_I$. For each treatment arm $A \in \{0, 1\}$, the components of coefficient β_A were independently sampled from $\{0, 0.5, 2, 2.5, 4\}$ with probabilities $\{0.6, 0.1, 0.1, 0.1, 0.1\}$, respectively. A random seed was used (seed = 998) for the initial draw, and the resulting coefficient values were held fixed across all generated datasets, so that the generated data share the same underlying coefficient. For $\beta_{I,A}$, we used

$\beta_{I,1} = (0.1, 0.2, 0.3, 0, 0.5)^\top$ and $\beta_{I,0} = (0.1, 0.2, 0, 0.4, 0)^\top$.

4.4.3 Settings

For the simulation study, we consider two different settings. Setting 1 uses the similar data generating process as in the semi-synthetic data in Setting B of Hill (2011). Let $X = (X_S, X_I)^\top$. We denote a vector of fifty ones as $\mathbb{1}_{50}$. The outcome models for the two treatment arms were

$$\mu_1(X) = 5 \cdot \mu(X) + \exp\{(\beta_1^\top Z + \beta_{I,1}^\top Z_I)/5 + 0.5\mathbb{1}_{50}\}/2; \;\; \mu_0(X) = 5 \cdot \mu(X) + (\beta_0^\top Z + \beta_{I,0}^\top Z_I)/2.$$

Then the response is generated by $Y \sim N(\mu_0(X) + A \cdot (\mu_1(X) - \mu_0(X)), 1)$. Within Setting 1, we consider two different scenarios: balanced and imbalanced treatment assignment. In the balanced scenario where the number of treated and controls are similar, the propensity score is

$$\Pr(A = 1|X) = \text{expit}\left(-Z^{(1)}/2 + Z^{(2)} + Z^{(4)} + Z^{(6)} + Z^{(8)} + Z^{(10)} + \mathbb{1}_5^\top Z_I\right).$$

The propensity scores, $\pi(x)$, range from 0.09 to 0.92 with 54% of the training observations assigned to $A = 1$. In the imbalanced scenario, treatment is assigned according to

$$\Pr(A = 1|X) = \text{expit}\left(-Z^{(1)}/2 - Z^{(2)} - Z^{(4)} - Z^{(6)} - Z^{(8)} - Z^{(10)} - \mathbb{1}_5^\top Z_I\right).$$

This gives the propensity scores, $\pi(x)$, ranging from 0.01 to 0.58, resulting in a treated proportion of 12% in the training set. In Setting 2, we consider a simpler process where

$$\mu_1(X) = \mu(X) + \beta_1^\top Z + \beta_{I,1}^\top Z_I; \;\; \mu_0(X) = \mu(X) + \beta_0^\top Z + \beta_{I,0}^\top Z_I.$$

As in Setting 1, we consider the balanced and the imbalanced scenarios with the same propensity scores.

4.4.4 Results

We compared the root mean squared errors fromm the proposed method, RNet, and the learners in Table 4.1 and Table 4.2. We used two-fold cross-fitting to estimate the nuisance functions in RNet, thereby reducing the computational burden of fitting separate CNNs for the nuisance functions and the CATE function. The simulation results are presented in Figure 4.5. Consistent with the findings in Section 4.4.1, the PW- and DR-learner variants, as well as the T-learner, produced substantially larger RMSE values and were omitted from the figure. Across all scenarios, JointRNet achieved a lower RMSE than the RNet, providing finite-sample evidence that joint estimation and information sharing improve performance. The one-step plug-in approaches and the regression adjustment (RA) learner also achieved RMSE values comparable to or lower than those of RNet; however, JointRNet outperformed these methods across the scenarios considered. These results suggest that JointRNet adapts effectively to settings in which the treatment-specific response surfaces are complex functions of the observed covariates.

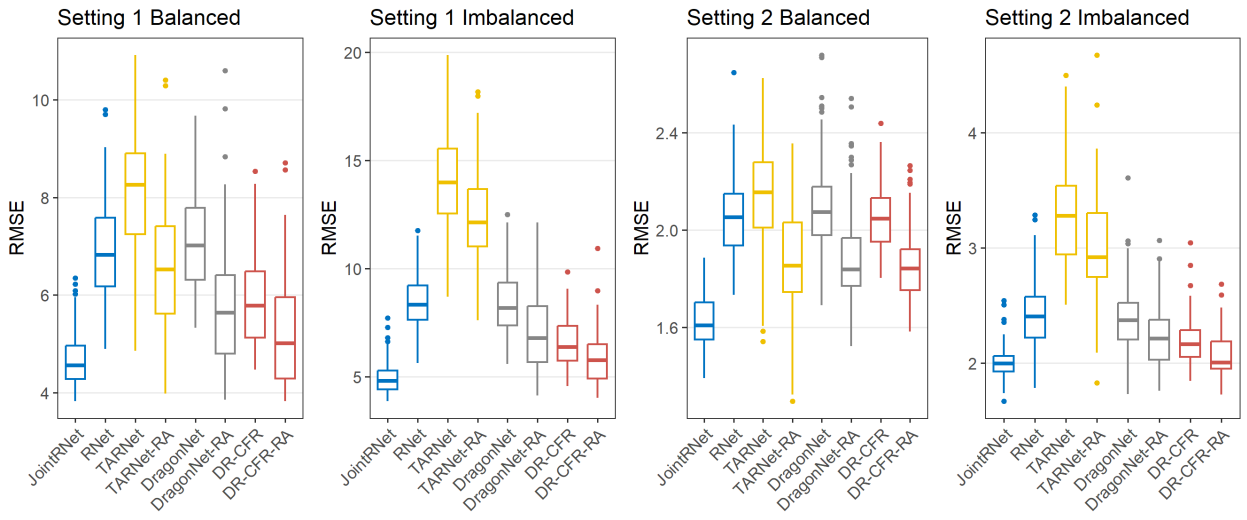

Figure 4.5: Simulation results for different methods. JointRNet is the proposed single-step neural network approach for R-learners. The RMSEs of the estimated CATE functions over the 100 iterations are presented for balanced and imbalanced treatment assignment scenarios within Setting 1 and Setting 2.

Moreover, we examined the MSE of JointRNet as the sample size increased from 500 to 10,000 across all simulation scenarios. The MSE decreased consistently with increasing sample size, providing empirical evidence of convergence (Figure 4.6). The observed rates were

comparable to those reported for existing neural-network-based CATE estimators (Curth and Schaar 2021). This suggests that JointRNet may admit similar excess-risk guarantees. Establishing formal excess-risk bounds for JointRNet remains an important direction for future research.

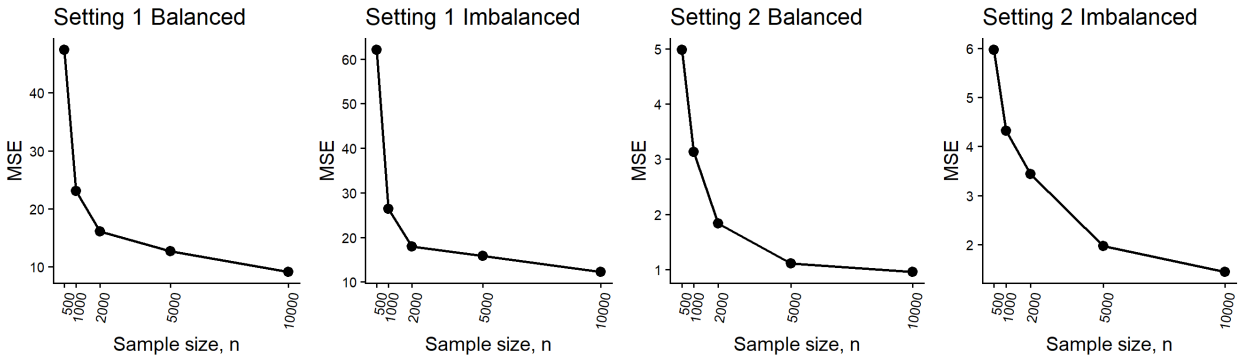

Figure 4.6: MSE along increasing sample size.

Nuisance parameters

We compared the accuracy for estimating the nuisance functions between the JointRNet and the RNet across all settings. We analyzed the RMSE of each nuisance parameter:

$$\sqrt{n_{test}^{-1} \sum_{i \in [n_{test}]} \{\hat{m}(X_i) - m(X_i)\}^2}; \quad \sqrt{n_{test}^{-1} \sum_{i \in [n_{test}]} \{\hat{\pi}(X_i) - \pi(X_i)\}^2}.$$

As shown in Figure 4.7, JointRNet yields lower estimation errors for both nuisance functions. These results suggest that its joint, single-step learning framework and shared representations also improve nuisance-function estimation. The improvement is particularly pronounced for the conditional mean function, for which JointRNet achieves substantially lower errors.

4.5 Conclusion

We propose JointRNet, a deep-learning approach for learning CATE that enables shared representations across the nuisance function and CATE components. JointRNet simultaneously estimates the nuisance functions and treatment-effect heterogeneity by minimizing a joint

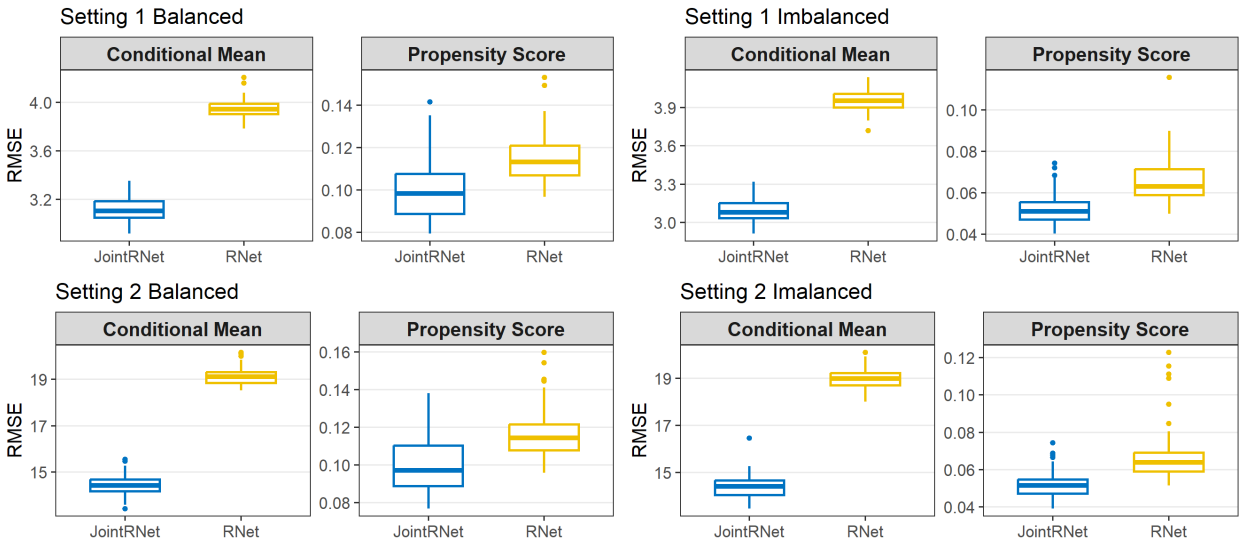

Figure 4.7: Comparison of nuisance functions estimates between the JointRNet and the RNet. The RMSEs of the estimated $m(x)$ and $\pi(x)$ over the 100 iterations are presented for balanced and imbalanced treatment assignment scenarios within Setting 1 and Setting 2.

R-learner objective. We showed that, at the population level, this objective is minimized by the true conditional mean function, propensity score, and CATE.

Through empirical experiments, we demonstrate that JointRNet accurately estimates the CATE in settings involving either tabular covariates alone or multimodal covariates comprising tabular data and unstructured images. Across all scenarios, JointRNet improved on the two-step R-learner and had smaller estimation error than the other CATE learners considered in this study, while also improving nuisance-function estimation. Furthermore, its mean squared error exhibited a desirable decreasing trend, approximately proportional to $n^{-\alpha}$, as the sample size increased. Further theoretical development is needed. In particular, future work could establish excess-risk bounds for JointRNet.

The proposed architectures integrate multimodal data using feed-forward fusion. Future studies could explore more advanced multimodal fusion strategies, such as attention-based fusion, to better capture interactions across modalities (Nagrani et al. 2021). Furthermore, the representation learned for CATE estimation, $\phi_\tau(x)$, could be incorporated into the dimension-reduced outcome-weighted learning framework as in Chapter 2 to estimate individualized treatment regimes.

Chapter 5

Discussion

This dissertation develops dimension reduction and representation learning methods for causal inference with high-dimensional covariates and mediators. The three novel methods address distinct causal estimands – optimal individualized treatment regimes, natural indirect effects, and conditional average treatment effects. They share a common principle that the representations should preserve information relevant to the causal target rather than simply capture variation in the observed data. The proposed methods implement this principle through sufficient dimension reduction, envelope models, and nonlinear representation learning.

For the linear dimension reduction methods developed in Chapter 2 and Chapter 3, we established theoretical properties supporting their use for the corresponding causal estimands. Across all three projects, simulation studies demonstrated favorable performance relative to existing approaches, while real and semi-synthetic data analyses illustrated their applicability to complex, high-dimensional settings. The results for JointRNet further demonstrate the potential of nonlinear representation learning for causal analyses involving multimodal covariates.

Specifically, in Chapter 2, we proposed a method for learning optimal individualized treatment regimes by combining the novel sufficient dimension reduction with covariate-balancing weights. We established that the risk of the proposed method converges in probability to the Bayes risk.

In Chapter 3, we developed an envelope-based dimension reduction method that identifies

the material component of the mediators within a linear structural equation model. Under the normality assumption, the material and immaterial reduced features are independent, and the natural indirect effect is determined solely by the material component. Future research could extend this framework to nonlinear structural equation models with more general error distributions, establish identification conditions for the natural indirect effect under weaker assumptions, and develop corresponding dimension-reduction methods.

Finally, in Chapter 4, we investigated neural network implementations of CATE learners with multimodal covariates and proposed JointRNet. JointRNet improved upon the original two-step, cross-fitted R-learner by jointly learning the nuisance functions and CATE through shared representations. Further theoretical work is needed to characterize its estimation error and establish excess-risk bounds. Its architecture could also be extended using more sophisticated multimodal fusion and representation-learning strategies.

Taken together, these findings demonstrate that dimension reduction in causal inference should be guided by the causal estimand and its underlying pathways. Extending this principle to nonlinear, multimodal, and other complex data settings offers a promising direction for developing interpretable and scalable methods for personalized decision-making and causal mechanism analysis.

References

Arnlov, J. (2009). "Diminished Renal Function and the Incidence of Heart Failure". In: *Current Cardiology Reviews* 5.3, pp. 223–227.

Athey, S., Imbens, G. W., and Wager, S. (2018). "Approximate Residual Balancing: Debiased Inference of Average Treatment Effects in High Dimensions". In: *Journal of the Royal Statistical Society Series B: Statistical Methodology* 80.4, pp. 597–623.

Athey, S., Tibshirani, J., and Wager, S. (2019). "Generalized random forests". In: *The Annals of Statistics* 47.2.

Aylward, R. E., Hayward, S., Chesnaye, N. C., et al. (2025). "Cardiometabolic protein expression levels and pathways associated with kidney function decline in older European adults with advanced kidney disease". In: *Clinical Kidney Journal* 18.4, sfaf079.

Bach, F. (2024). *Learning theory from first principles*. Adaptive computation and machine learning. Cambridge, Massachusetts: The MIT Press.

Baron, R. M. and Kenny, D. A. (1986). "The moderator–mediator variable distinction in social psychological research: Conceptual, strategic, and statistical considerations." In: *Journal of Personality and Social Psychology* 51.6, pp. 1173–1182.

Bartlett, P. L., Jordan, M. I., and McAuliffe, J. D. (2006). "Convexity, Classification, and Risk Bounds". In: *Journal of the American Statistical Association* 101.473, pp. 138–156.

Bengio, Y., Courville, A., and Vincent, P. (2013). "Representation Learning: A Review and New Perspectives". In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 35.8, pp. 1798–1828.

Boyd, S. P. and Vandenberghe, L. (2004). *Convex optimization*. Cambridge, UK ; New York: Cambridge University Press.

Carlsson, A. C., Ingelsson, E., Sundström, J., Jesus Carrero, J., Gustafsson, S., Feldreich, T., Stenemo, M., Larsson, A., Lind, L., and Ärnlöv, J. (2017). "Use of Proteomics To

Investigate Kidney Function Decline over 5 Years". In: *Clinical Journal of the American Society of Nephrology* 12.8, pp. 1226–1235.

Chan, K. C. G., Yam, S. C. P., and Zhang, Z. (2016). "Globally Efficient Non-Parametric Inference of Average Treatment Effects by Empirical Balancing Calibration Weighting". In: *Journal of the Royal Statistical Society Series B: Statistical Methodology* 78.3, pp. 673–700.

Chen, R., Huling, J. D., Chen, G., and Yu, M. (2024). "Robust sample weighting to facilitate individualized treatment rule learning for a target population". In: *Biometrika* 111.1, pp. 309–329.

Chén, O. Y., Crainiceanu, C., Ogburn, E. L., Caffo, B. S., Wager, T. D., and Lindquist, M. A. (2018). "High-dimensional multivariate mediation with application to neuroimaging data". In: *Biostatistics* 19.2, pp. 121–136.

Chen, Yuan, Liu, Ying, Zeng, Donglin, and Wang, Yuanjia (2020). *DTRlearn2: Statistical Learning Methods for Optimizing Dynamic Treatment Regimes*.

Cheng, D., Li, J., Liu, L., Le, T. D., Liu, J., and Yu, K. (2022). "Sufficient dimension reduction for average causal effect estimation". In: *Data Mining and Knowledge Discovery* 36.3, pp. 1174–1196.

Clivio, O., Feller, A., and Holmes, C. (2024). "Towards Representation Learning for Weighting Problems in Design-Based Causal Inference". In: *Proceedings of the Fortieth Conference on Uncertainty in Artificial Intelligence*. PMLR, pp. 856–880.

Cook, R. D., Helland, I. S., and Su, Z. (2013a). "Envelopes and Partial Least Squares Regression". In: *Journal of the Royal Statistical Society Series B: Statistical Methodology* 75.5, pp. 851–877.

— (2013b). "Envelopes and Partial Least Squares Regression". In: *Journal of the Royal Statistical Society Series B: Statistical Methodology* 75.5, pp. 851–877.

Cook, R. D. (2018a). *An Introduction to Envelopes: Dimension Reductionfor Efficient Estimation in Multivariate Statistics*. 1st ed. Wiley Series in Probability and Statistics. Wiley.

— (2018b). "Principal Components, Sufficient Dimension Reduction, and Envelopes". In: *Annual Review of Statistics and Its Application* 5.1, pp. 533–559.

Cook, R. D. and Zhang, X. (2016). "Algorithms for Envelope Estimation". In: *Journal of Computational and Graphical Statistics* 25.1, pp. 284–300.

Cook, R. and Li, B. (2002). "Dimension reduction for conditional mean in regression". In: *The Annals of Statistics* 30.2.

Cook, R. D., Li, Bing, and Chiaromonte, Franceska (2010). "Envelope models for parsimonious and efficient multivariate linear regression". In: *Statistica Sinica* 20, pp. 927–1010.

Curth, A. and Schaar, M. v. d. (2021). "Nonparametric Estimation of Heterogeneous Treatment Effects: From Theory to Learning Algorithms". In: *Proceedings of The 24th International Conference on Artificial Intelligence and Statistics*. PMLR, pp. 1810–1818.

Curth, A., Svensson, D., Weatherall, J., and Schaar, M. van der (2021). "Really Doing Great at Estimating CATE? A Critical Look at ML Benchmarking Practices in Treatment Effect Estimation". In: *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks* 1.

Daniel, R. M., De Stavola, B. L., Cousens, S. N., and Vansteelandt, S. (2015). "Causal mediation analysis with multiple mediators". In: *Biometrics* 71.1, pp. 1–14.

Daniels, J. R., Ma, J. Z., Cao, Z., et al. (2021). "Discovery of Novel Proteomic Biomarkers for the Prediction of Kidney Recovery from Dialysis-Dependent AKI Patients". In: *Kidney360* 2.11, pp. 1716–1727.

Dasgupta, S., Goldberg, Y., and Kosorok, M. R. (2019). "Feature Eliminiation in Kernel Machines in Moderately High Dimensions". In: *Annals of Statistics* 47.1, pp. 497–526.

Davis, C. and Kahan, W. M. (1970). "The Rotation of Eigenvectors by a Perturbation. III". In: *SIAM Journal on Numerical Analysis* 7.1, pp. 1–46.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2018). *BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding*. Version Number: 2.

Ding, X. and Yang, F. (2022). "Tracy-Widom Distribution for Heterogeneous Gram Matrices With Applications in Signal Detection". In: *IEEE Transactions on Information Theory* 68.10, pp. 6682–6715.

Feng, M., McSparron, J. I., Kien, D. T., Stone, D. J., Roberts, D. H., Schwartzstein, R. M., Vieillard-Baron, A., and Celi, L. A. (2018). "Transthoracic echocardiography and mortality in sepsis: analysis of the MIMIC-III database". In: *Intensive Care Medicine* 44.6, pp. 884–892.

Fong, C. and Grimmer, J. (2023). "Causal Inference with Latent Treatments". In: *American Journal of Political Science* 67.2, pp. 374–389.

Fukumizu, K., Bach, F. R., and Gretton, A. (2007). "Statistical Consistency of Kernel Canonical Correlation Analysis". In: *Journal of Machine Learning Research* 8.14, pp. 361–383.

Fukumizu, K., Bach, F. R., and Jordan, M. I. (2009). "Kernel dimension reduction in regression". In: *The Annals of Statistics* 37.4.

Fukumizu, K. and Leng, C. (2014). "Gradient-Based Kernel Dimension Reduction for Regression". In: *Journal of the American Statistical Association* 109.505, pp. 359–370.

Gao, Y., Yang, H., Fang, R., Zhang, Y., Goode, E. L., and Cui, Y. (2019). "Testing Mediation Effects in High-Dimensional Epigenetic Studies". In: *Frontiers in Genetics* 10, p. 1195.

Gretton, A., Borgwardt, K. M., Rasch, M. J., Schölkopf, B., and Smola, A. (2012). "A Kernel Two-Sample Test". In: *Journal of Machine Learning Research* 13.25, pp. 723–773.

Hable, R. (2012). "Asymptotic normality of support vector machine variants and other regularized kernel methods". In: *Journal of Multivariate Analysis* 106, pp. 92–117.

Hassanpour, N. and Greiner, R. (2020). "Learning Disentangled Representations for CounterFactual Regression". In.

Hill, J. L. (2011). "Bayesian Nonparametric Modeling for Causal Inference". In: *Journal of Computational and Graphical Statistics* 20.1, pp. 217–240.

Hirata, K., Shikata, K., Matsuda, M., Akiyama, K., Sugimoto, H., Kushiro, M., and Makino, H. (1998). "Increased expression of selectins in kidneys of patients with diabetic nephropathy". In: *Diabetologia* 41.2, pp. 185–192.

Hirshberg, D. A., Maleki, A., and Zubizarreta, J. R. (2019). *Minimax Linear Estimation of the Retargeted Mean*. Version Number: 2.

Hirshberg, D. A. and Wager, S. (2017). *Augmented Minimax Linear Estimation*. Version Number: 6.

Hristache, M., Juditsky, A., and Spokoiny, V. (2001). "Direct estimation of the index coefficient in a single-index model". In: *The Annals of Statistics* 29.3.

Huang, M.-Y. and Chan, K. C. G. (2017). "Joint sufficient dimension reduction and estimation of conditional and average treatment effects". In: *Biometrika* 104.3, pp. 583–596.

Huang, M.-Y. and Yang, S. (2023). "Robust Inference of Conditional Average Treatment Effects Using Dimension Reduction". In: *Statistica Sinica*.

Huang, Y.-T. and Pan, W.-C. (2016). "Hypothesis Test of Mediation Effect in Causal Mediation Model With High-Dimensional Continuous Mediators". In: *Biometrics* 72.2, pp. 402–413.

Imai, K. and Ratkovic, M. (2014). "Covariate Balancing Propensity Score". In: *Journal of the Royal Statistical Society Series B: Statistical Methodology* 76.1, pp. 243–263.

Imai, K. and Yamamoto, T. (2013). "Identification and Sensitivity Analysis for Multiple Causal Mechanisms: Revisiting Evidence from Framing Experiments". In: *Political Analysis* 21.2, pp. 141–171.

Jerzak, C. T., Johansson, F. D., and Daoud, A. (2023). "Image-based Treatment Effect Heterogeneity". In: *Proceedings of the Second Conference on Causal Learning and Reasoning*. PMLR, pp. 531–552.

Johnson, A. E., Pollard, T. J., Shen, L., Lehman, L.-w. H., Feng, M., Ghassemi, M., Moody, B., Szolovits, P., Anthony Celi, L., and Mark, R. G. (2016). "MIMIC-III, a freely accessible critical care database". In: *Scientific Data* 3.1, p. 160035.

Kallus, N. and Santacatterina, M. (2022). "Optimal weighting for estimating generalized average treatment effects". In: *Journal of Causal Inference* 10.1, pp. 123–140.

Kang, J. and Shin, S. J. (2022). "A Forward Approach for Sufficient Dimension Reduction in Binary Classification". In: *Journal of Machine Learning Research* 23.199, pp. 1–31.

Karatzoglou, A., Smola, A., Hornik, K., and Zeileis, A. (2004). "**kernlab** - An *S4* Package for Kernel Methods in *R*". In: *Journal of Statistical Software* 11.9.

Kennedy, E. H. (2023). "Towards optimal doubly robust estimation of heterogeneous causal effects". In: *Electronic Journal of Statistics* 17.2.

Kingma, D. P. and Ba, J. (2014). *Adam: A Method for Stochastic Optimization*. Version Number: 9.

Kosorok, M. R. (2008). *Introduction to Empirical Processes and Semiparametric Inference*. Springer Series in Statistics Ser. New York, NY: Springer.

Künzel, S. R., Sekhon, J. S., Bickel, P. J., and Yu, B. (2019). "Metalearners for estimating heterogeneous treatment effects using machine learning". In: *Proceedings of the National Academy of Sciences* 116.10, pp. 4156–4165.

Kwon, O.-R. and Zou, H. (2025). "Enhanced Response Envelope via Envelope Regularization". In: *Journal of the American Statistical Association* 120.550, pp. 859–868.

Lee, M. and Su, Z. (2020). "A Review of Envelope Models". In: *International Statistical Review* 88.3, pp. 658–676.

Li, K.-C. (1991). "Sliced Inverse Regression for Dimension Reduction". In: *Journal of the American Statistical Association* 86.414, pp. 316–327.

Liang, M. and Yu, M. (2022). "A Semiparametric Approach to Model Effect Modification". In: *Journal of the American Statistical Association* 117.538, pp. 752–764.

Liang, Z., Pan, Z., and Xiong, R. (2025). *Causal Representation Learning from Multimodal Clinical Records under Non-Random Modality Missingness*. Version Number: 1.

Liu, Y., Wang, Y., Kosorok, M. R., Zhao, Y., and Zeng, D. (2018). "Augmented outcome-weighted learning for estimating optimal dynamic treatment regimens". In: *Statistics in Medicine* 37.26, pp. 3776–3788.

Matthew Cefalu, Greg Ridgeway, Dan McCaffrey, Andrew Morral, Beth Ann Griffin, and Lane Burgette (2024). *twang: Toolkit for Weighting and Analysis of Nonequivalent Groups*.

McCaffrey, D. F., Ridgeway, G., and Morral, A. R. (2004). "Propensity score estimation with boosted regression for evaluating causal effects in observational studies". In: *Psychological Methods* 9.4, pp. 403–425.

Mendelson, S. (2017). *Extending the scope of the small-ball method*. Version Number: 2.

Murphy, S. A., Van Der Laan, M. J., Robins, J. M., and Conduct Problems Prevention Research Group (2001). "Marginal Mean Models for Dynamic Regimes". In: *Journal of the American Statistical Association* 96.456, pp. 1410–1423.

Murphy, S. A. (2003). "Optimal Dynamic Treatment Regimes". In: *Journal of the Royal Statistical Society Series B: Statistical Methodology* 65.2, pp. 331–355.

Nagrani, A., Yang, S., Arnab, A., Jansen, A., Schmid, C., and Sun, C. (2021). "Attention Bottlenecks for Multimodal Fusion". In: *Advances in Neural Information Processing Systems*. Vol. 34. Curran Associates, Inc., pp. 14200–14213.

Nashed, M. Z. (1986). "The Theory of Tikhonov Regularization for Fredholm Equations of the First Kind (C. W. Groetsch)". In: *SIAM Review* 28.1, pp. 116–118.

Nath, T., Caffo, B., Wager, T., and Lindquist, M. A. (2023). "A machine learning based approach towards high-dimensional mediation analysis". In: *NeuroImage* 268, p. 119843.

Nie, X. and Wager, S. (2021). "Quasi-oracle estimation of heterogeneous treatment effects". In: *Biometrika* 108.2, pp. 299–319.

Onatski, A. (2009). "Testing Hypotheses About the Number of Factors in Large Factor Models". In: *Econometrica* 77.5. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.3982/ECTA6964, pp. 1447–1479.

Park, H., Petkova, E., Tarpey, T., and Ogden, R. T. (2021). "A constrained single-index regression for estimating interactions between a treatment and covariates". In: *Biometrics* 77.2, pp. 506–518.

— (2020). *Sufficient Dimension Reduction for Interactions*. Version Number: 1.

Park, Hyung, Petkova, Eva, Tarpey, Thaddeus, and Ogden, R. Todd (2021). *simml: Single-Index Models with Multiple-Links*.

Pearl, J. (2001). "Direct and indirect effects". In: *Proceedings of the Seventeenth conference on Uncertainty in artificial intelligence*. UAI'01. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., pp. 411–420.

Peypouquet, J. (2015). *Convex Optimization in Normed Spaces: Theory, Methods and Examples*. SpringerBriefs in Optimization. Cham: Springer International Publishing.

Qian, M. and Murphy, S. A. (2011). "Performance guarantees for individualized treatment rules". In: *The Annals of Statistics* 39.2.

Rimal, R., Almøy, T., and Sæbø, S. (2019). "Comparison of multi-response prediction methods". In: *Chemometrics and Intelligent Laboratory Systems* 190, pp. 10–21.

Robins, J. M. and Greenland, S. (1992). "Identifiability and exchangeability for direct and indirect effects". In: *Epidemiology (Cambridge, Mass.)* 3.2, pp. 143–155.

Robins, J. M. (2004). "Optimal Structural Nested Models for Optimal Sequential Decisions". In: *Proceedings of the Second Seattle Symposium in Biostatistics*. Ed. by P. Bickel, P. Diggle, S. Fienberg, U. Gather, I. Olkin, S. Zeger, D. Y. Lin, and P. J. Heagerty. Vol. 179. Series Title: Lecture Notes in Statistics. New York, NY: Springer New York, pp. 189–326.

Robins, J. M., Rotnitzky, A., and Zhao, L. P. (1995). "Analysis of Semiparametric Regression Models for Repeated Outcomes in the Presence of Missing Data". In: *Journal of the American Statistical Association* 90.429, pp. 106–121.

Robinson, P. M. (1988). "Root-N-Consistent Semiparametric Regression". In: *Econometrica* 56.4, p. 931.

Ronco, C., Haapio, M., House, A. A., Anavekar, N., and Bellomo, R. (2008). "Cardiorenal Syndrome". In: *Journal of the American College of Cardiology* 52.19, pp. 1527–1539.

Rosthøj, S., Fullwood, C., Henderson, R., and Stewart, S. (2006). "Estimation of optimal dynamic anticoagulation regimes from observational data: a regret-based approach". In: *Statistics in Medicine* 25.24, pp. 4197–4215.

Rubin, D. B. (1974). "Estimating causal effects of treatments in randomized and nonrandomized studies." In: *Journal of Educational Psychology* 66.5, pp. 688–701.

Scholkopf, B., Locatello, F., Bauer, S., Ke, N. R., Kalchbrenner, N., Goyal, A., and Bengio, Y. (2021). "Toward Causal Representation Learning". In: *Proceedings of the IEEE* 109.5, pp. 612–634.

Schulte, P. J., Tsiatis, A. A., Laber, E. B., and Davidian, M. (2014). "Q- and A-learning Methods for Estimating Optimal Dynamic Treatment Regimes". In: *Statistical Science: A Review Journal of the Institute of Mathematical Statistics* 29.4, pp. 640–661.

Schulte, R., Rügamer, D., and Nagler, T. (2025). *Adjustment for Confounding using Pre-Trained Representations*. Version Number: 1.

Serfling, R. J. (1980). *Approximation Theorems of Mathematical Statistics*. 1st ed. Wiley Series in Probability and Statistics. Wiley.

Shalit, U., Johansson, F. D., and Sontag, D. (2017). "Estimating individual treatment effect: generalization bounds and algorithms". In: *Proceedings of the 34th International Conference on Machine Learning*. PMLR, pp. 3076–3085.

Shang, Y., Chiu, Y.-H., and Kong, L. (2025). "Robust propensity score estimation via loss function calibration". In: *Statistical Methods in Medical Research* 34.3, pp. 457–472.

Shapiro, A. (1986). "Asymptotic Theory of Overparameterized Structural Models". In: *Journal of the American Statistical Association* 81.393, pp. 142–149.

Shi, C., Blei, D. M., and Veitch, V. (2019). "Adapting neural networks for the estimation of treatment effects". In: *Proceedings of the 33rd International Conference on Neural Information Processing Systems*. 225. Red Hook, NY, USA: Curran Associates Inc., pp. 2507–2517.

Siddiqui, K., George, T. P., Mujammami, M., Isnani, A., and Alfadda, A. A. (2023). "The association of cell adhesion molecules and selectins (VCAM-1, ICAM-1, E-selectin, L-selectin, and P-selectin) with microvascular complications in patients with type 2 diabetes: A follow-up study". In: *Frontiers in Endocrinology* 14, p. 1072288.

Stekhoven, D. J. and Bühlmann, P. (2012). "MissForest—non-parametric missing value imputation for mixed-type data". In: *Bioinformatics* 28.1, pp. 112–118.

Su, Z. and Cook, R. D. (2012). "Inner envelopes: efficient estimation in multivariate linear regression". In: *Biometrika* 99.3, pp. 687–702.

— (2011). "Partial envelopes for efficient estimation in multivariate linear regression". In: *Biometrika* 98.1, pp. 133–146.

VanderWeele, T. and Vansteelandt, S. (2014). "Mediation Analysis with Multiple Mediators". In: *Epidemiologic Methods* 2.1.

VanderWeele, T. J., Vansteelandt, S., and Robins, J. M. (2014). "Effect Decomposition in the Presence of an Exposure-Induced Mediator-Outcome Confounder:" in: *Epidemiology* 25.2, pp. 300–306.

Veitch, V., Sridhar, D., and Blei, D. (2020). "Adapting Text Embeddings for Causal Inference". In: *Proceedings of the 36th Conference on Uncertainty in Artificial Intelligence (UAI)*. PMLR, pp. 919–928.

Wager, S. and Athey, S. (2018). "Estimation and Inference of Heterogeneous Treatment Effects using Random Forests". In: *Journal of the American Statistical Association* 113.523, pp. 1228–1242.

Wang, X., Peng, Y., Lu, L., Lu, Z., Bagheri, M., and Summers, R. M. (2017). "ChestX-Ray8: Hospital-Scale Chest X-Ray Database and Benchmarks on Weakly-Supervised Classification and Localization of Common Thorax Diseases". In: *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. Honolulu, HI: IEEE, pp. 3462–3471.

Watkins, C. J. C. H. and Dayan, P. (1992). "Q-learning". In: *Machine Learning* 8.3-4, pp. 279–292.

Wong, R. K. W. and Chan, K. C. G. (2018). "Kernel-based covariate functional balancing for observational studies". In: *Biometrika* 105.1, pp. 199–213.

Xia, Y., Tong, H., Li, W. K., and Zhu, L.-X. (2002). "An Adaptive Estimation of Dimension Reduction Space". In: *Journal of the Royal Statistical Society Series B: Statistical Methodology* 64.3, pp. 363–410.

Xie, S., Tarpey, T., Petkova, E., and Todd Ogden, R. (2022). "Multiple Domain and Multiple Kernel Outcome-Weighted Learning for Estimating Individualized Treatment Regimes". In: *Journal of Computational and Graphical Statistics* 31.4, pp. 1375–1383.

Xu, Y., Johnson, T. D., Heitzeg, M., and Kang, J. (2026). "Bayesian Image Mediation Analysis". In: *Journal of the American Statistical Association*, pp. 1–14.

Yang, J., Shi, R., Wei, D., Liu, Z., Zhao, L., Ke, B., Pfister, H., and Ni, B. (2024). *[MedMNIST+] 18x Standardized Datasets for 2D and 3D Biomedical Image Classification with Multiple Size Options: 28 (MNIST-Like), 64, 128, and 224.*

— (2023). "MedMNIST v2 - A large-scale lightweight benchmark for 2D and 3D biomedical image classification". In: *Scientific Data* 10.1, p. 41.

Yu, Y., Wang, T., and Samworth, R. J. (2015). "A useful variant of the Davis–Kahan theorem for statisticians". In: *Biometrika* 102.2, pp. 315–323.

Zhang, B., Tsiatis, A. A., Laber, E. B., and Davidian, M. (2012). "A Robust Method for Estimating Optimal Treatment Regimes". In: *Biometrics* 68.4, pp. 1010–1018.

Zhang, X. and Cook, D. (2018). "Fast envelope algorithms". In: *Statistica Sinica*.

Zhang, X., Deng, K., and Mai, Q. (2023). "Envelopes and principal component regression". In: *Electronic Journal of Statistics* 17.2.

Zhao, Y., Lindquist, M. A., and Caffo, B. S. (2020). "Sparse principal component based high-dimensional mediation analysis". In: *Computational Statistics & Data Analysis* 142, p. 106835.

Zhao, Y. and Luo, X. (2022). "Pathway Lasso: pathway estimation and selection with high-dimensional mediators". In: *Statistics and Its Interface* 15.1, pp. 39–50.

Zhao, Y.-Q., Laber, E. B., Ning, Y., Saha, S., and Sands, B. E. (2019). "Efficient augmentation and relaxation learning for individualized treatment rules using observational data". In: *Journal of machine learning research: JMLR* 20, p. 48.

Zhao, Y., Zeng, D., Rush, A. J., and Kosorok, M. R. (2012). "Estimating Individualized Treatment Rules Using Outcome Weighted Learning". In: *Journal of the American Statistical Association* 107.449, pp. 1106–1118.

Zhao, Y., Kosorok, M. R., and Zeng, D. (2009). "Reinforcement learning design for cancer clinical trials". In: *Statistics in Medicine* 28.26, pp. 3294–3315.

Zhou, W., Zhu, R., and Zeng, D. (2021). "A parsimonious personalized dose-finding model via dimension reduction". In: *Biometrika* 108.3, pp. 643–659.

Zhou, X. and Kosorok, M. R. (2017). *Augmented Outcome-weighted Learning for Optimal Treatment Regimes*. Version Number: 1.

Zhou, X., Mayer-Hamblett, N., Khan, U., and Kosorok, M. R. (2017). "Residual Weighted Learning for Estimating Individualized Treatment Rules". In: *Journal of the American Statistical Association* 112.517, pp. 169–187.

Zhu, G. and Su, Z. (2020). "Envelope-based sparse partial least squares". In: *The Annals of Statistics* 48.1.

Zubizarreta, J. R. (2015). "Stable Weights that Balance Covariates for Estimation With Incomplete Outcome Data". In: *Journal of the American Statistical Association* 110.511, pp. 910–922.

Appendix A

Chapter 3 Appendix

A.1 Proof for Theorem 2.4.1

For simplicity, here we use $\mathcal{H}$ in place of $\mathcal{H}_{\mathcal{X}}$ and $K : \mathcal{X} \times \mathcal{X} \mapsto \mathbb{R}$ in place of $K_X : \mathcal{X} \times \mathcal{X} \mapsto \mathbb{R}$. Furthermore, define $S_1 = \{i \in [n]; A_i = +1\}$. We first recall the regularity conditions and the statement of the theorem below.

(A1) The covariate space $\mathcal{X}$ is compact, and the covariates have a continuous density supported on $\mathcal{X}$.

(A2) The propensity score functions $\pi(+1, x)$ and $\pi(-1, x)$ are continuous in $x \in \mathcal{X}$.

(A3) The kernel $K_X(x_1, x_2)$ defining $\mathcal{H}$ is continuous, non-negative, bounded, and satisfies $K(x, x) > 0$ for all $x \in \mathcal{X}$.

(A4) The RKHS $\mathcal{H}$ is dense in the space of continuous functions on $\mathcal{X}$ with respect to the supremum norm.

(A5) The unit ball $\mathcal{H}_1 = \{m \in \mathcal{H} : \|m\|_{\mathcal{H}} \leq 1\}$ is a $\mathbb{P}$-Donsker class.

Theorem. Assume $\lambda_1 \asymp n^{-1}$, and let the KCB weights $\{\hat{w}_i : A_i = +1\}$ be the solution to the optimization problem (2.6)

$$\min_{w \succeq 1} \left[\sup_{m \in \mathcal{H}_n} \left\{ \left(\frac{1}{n} \sum_{i \in S_1} w_i m(X_i) - \frac{1}{n} \sum_{i \in [n]} m(X_i) \right)^2 - \lambda_1 \|m\|_{\mathcal{H}}^2 \right\} + \frac{\lambda_2}{n} \sum_{i \in S_1} w_i^2 \right],$$

where $\mathcal{H}_n = \{m \in \mathcal{H}; \|m\|_n^2 = n^{-1}\sum_{i\in[n]} m^2(X_i) = 1\}$. Then, under the regularity conditions (A1) – (A5),

$$\frac{1}{n}\sum_{i\in S_1}\left\{\hat{w}_i - \frac{1}{\pi(+1, X_i)}\right\}^2 \longrightarrow 0,$$

in probability, as $n \to \infty$.

Define $m_n \in \mathcal{H}_n$ as $m_n = m/\|m\|_n$ for any $m \in \mathcal{H}$ such that $m \neq 0$. Without loss of generality, assume that the Gram matrix A, derived from the kernel $K : \mathcal{X} \times \mathcal{X} \mapsto \mathbb{R}$, is positive definite. By eigendecomposition, $A = \Gamma\Lambda\Gamma^\top$. From this,

$$\|m_n\|_{\mathcal{H}}^2 = \alpha^\top A\alpha = n \cdot (n^{-1/2}\Lambda\Gamma^\top\alpha)^\top\Lambda^{-1}(n^{-1/2}\Lambda\Gamma^\top\alpha) = n \cdot \beta^\top\Lambda^{-1}\beta,$$

where α is a vector of coefficients and $\beta = n^{-1/2}\Lambda\Gamma^\top\alpha$. Note that $\beta^\top\beta = n^{-1}\alpha^\top\Gamma\Lambda^2\Gamma^\top\alpha = n^{-1}\alpha^\top A^2\alpha = \|m_n\|_n^2 = 1$. Therefore, $\|m_n\|_{\mathcal{H}}^2 \leq n/\psi_{min}$, where ψ_{min} is the minimum eigenvalue of Λ. Since $\lambda_1 \asymp n^{-1}$, it is bounded by $n^{-1}D_1 \leq \lambda_1 \leq n^{-1}D_2$ for constants $D_2 \geq D_1 > 0$. From this, $\lambda_1\|m_n\|_{\mathcal{H}}^2 \leq n^{-1}D_2 \cdot n/\psi_{min} = D_2/\psi_{min}$. Therefore, we have the upper bound as

$$\|m_n\|_{\mathcal{H}}^2 \leq \frac{D_2}{\lambda_1\psi_{min}}.$$

Let $c_1 = \sqrt{\lambda_1\psi_{min}/D_2}$ and define $m_1 = c_1 \cdot m_n$. Then we have $\|m_1\|_{\mathcal{H}}^2 \leq 1$ and $m_1 \in \mathcal{H}_1$. We present the proof for m_1, and the result can be extended to $m_n \in \mathcal{H}_n$ by scaling: $m_n = m_1/c_1$.

The proof can be written in three steps, similar to the proof of Theorem 1 of Chen et al. (2024): characterization of the dual problem; finding the limiting form of the dual problem; and showing convergence in probability to the limiting form.

A.1.1 Step 1: Characterization of the dual problem

Define an operator $M : \mathbb{R}^n \mapsto \mathcal{H}$ such that

$$Mw = \sum_{i\in S_1} w_i K(X_i, \cdot),$$

where $S_1 = \{i; A_i = +1\}$. Define $h = \frac{1}{n}\sum_{i\in[n]} K(X_i, \cdot)$. Since $m_1 \in \mathcal{H}_1$,

$$\begin{aligned}
\left\{\frac{1}{n}\sum_{i=1}^{n}(T_i w_i - 1)m_1(X_i)\right\}^2 - \lambda_1\|m_1\|_{\mathcal{H}}^2 &= \left\{\frac{1}{n}\sum_{i=1}^{n}(T_i w_i - 1)\langle m_1, K(\cdot, X_i)\rangle_{\mathcal{H}}\right\}^2 - \lambda_1\|m_1\|_{\mathcal{H}}^2 \\
&= \left\langle m_1, \frac{1}{n}\sum_{i=1}^{n}(T_i w_i - 1)K(\cdot, X_i)\right\rangle_{\mathcal{H}}^2 - \lambda_1\|m_1\|_{\mathcal{H}}^2 \\
&\leq \|m_1\|_{\mathcal{H}}^2 \cdot \left\|\frac{1}{n}\sum_{i=1}^{n}(T_i w_i - 1)K(\cdot, X_i)\right\|_{\mathcal{H}}^2 - \lambda_1\|m_1\|_{\mathcal{H}}^2 \\
&= \|m_1\|_{\mathcal{H}}^2 \left\|\frac{1}{n}\sum_{i\in S_1} w_i K(\cdot, X_i) - \frac{1}{n}\sum_{i\in[n]} K(\cdot, X_i)\right\|_{\mathcal{H}}^2 - \lambda_1\|m_1\|_{\mathcal{H}}^2 \\
&= \|m_1\|_{\mathcal{H}}^2 \cdot n^{-2}\left\|\sum_{i\in S_1} w_i K(\cdot, X_i) - n\cdot\frac{1}{n}\sum_{i\in[n]} K(\cdot, X_i)\right\|_{\mathcal{H}}^2 - \lambda_1\|m_1\|_{\mathcal{H}}^2 \\
&= \|m_1\|_{\mathcal{H}}^2 \cdot n^{-2}\|Mw - nh\|_{\mathcal{H}}^2 - \lambda_1\|m_1\|_{\mathcal{H}}^2 \\
&= \|m_1\|_{\mathcal{H}}^2 \cdot n^{-2}\{\|Mw - nh\|_{\mathcal{H}}^2 - n^2\lambda_1\}
\end{aligned}$$

where $\lambda_{1n} = n^2\lambda_1$. The inequality follows from Cauchy-Schwarz inequality. Define $s(x)$ and $r(f)$ as follows:

$$s(x) = \Lambda(x) + \frac{\lambda_2}{2}\sum_{i\in S_1} x_i^2,$$

where $\Lambda(x) = 0$ when $x \succeq 1$ and $\Lambda(x) = \infty$, otherwise, and

$$r(f) = \frac{1}{2}\|f - nh\|_{\mathcal{H}}^2 - \frac{\lambda_{1n}}{2}.$$

Then minimization of (2.6) with respect to w is equivalent to minimizing the primal $p : \mathbb{R}^n \mapsto \mathbb{R} \cup \{\infty\}$ defined as

$$p(w) = s(w) + r(Mw) = \Lambda(w) + \frac{\lambda_2}{2}\sum_{i\in S_1} w_i^2 + \frac{1}{2}\left\{\sum_{i\in S_1} w_i K(\cdot, X_i) - \sum_{i\in[n]} K(\cdot, X_i)\right\}^2 - \frac{\lambda_{1n}}{2}.$$

We can obtain the Fenchel-Rockafeller dual, $d(f)$, (Peypouquet 2015) of the primal $p(w)$:

$$d(f) = s^*(-M^*f) + r^*(f),$$

where s^* and r^* are Fenchel conjugates of s and r, respectively. Furthermore, $M^* : \mathcal{H} \mapsto \mathbb{R}^n$ is the adjoint operator of M.

Obtaining Fenchel conjugate for $s(x)$

By definition of Fenchel conjugate,

$$\begin{aligned} s^*(y) &= \sup_x \{y^\top x - s(x)\} = \sup_x \{y^\top x - \Lambda(x) - \frac{\lambda_2}{2} x^\top x\} \\ &= \sup_{x: x \succeq 1} \{y^\top x - \frac{\lambda_2}{2} x^\top x\}. \end{aligned}$$

The optimization problem can be written as

$$Q(x) = y^\top x - \frac{\lambda_2}{2} x^\top x - \gamma^\top (1 - x).$$

The vanishing gradient of Lagrangian of the Karush-Kuhn-Tucker (KKT) conditions (Boyd and Vandenberghe 2004) requires

$$\nabla_x Q(x) = y - \lambda_2 x^* + \gamma = 0$$

for $\gamma \succeq 0$, where x^* is the optimum. By solving for this equation,

$$x^* = \frac{1}{\lambda_2}(y + \gamma).$$

From this,

$$s^*(y) = \frac{1}{2\lambda_2} y^\top y - \frac{1}{2\lambda_2} \gamma^\top \gamma.$$

For each $i \in S_1$, $x_i^* = \lambda_2^{-1}(y_i + \gamma_i)$ implies $\gamma_i = \lambda_2 x_i^* - y_i$. We additionally need $\gamma_i(1 - x_i) = 0$ for complementary slackness of the KKT condition (Boyd and Vandenberghe 2004). Then

$$x_i^* = \begin{cases} 1 & \text{and } \gamma_i = \lambda_2 - y_i > 0 \\ \frac{1}{\lambda_2} y_i; \frac{1}{\lambda_2} y_i > 1 & \text{and } \gamma_i = 0. \end{cases} \tag{A.1}$$

Note that $\gamma_i = \mathbb{1}(y_i < \lambda_2)(\lambda_2 - y_i) = \max(\lambda_2 - y_i, 0)$. By definition of adjoint operator, we have $\langle M^* f, e_i \rangle_{\mathcal{H}} = \langle f, M e_i \rangle_{\mathcal{H}} = f(X_i)$, where $e_i;\ i \in [n]$ is a standard basis vector. Eventually, we have

$$s^*(-M^* f) = \frac{1}{2\lambda_2} \sum_{i \in S_1} f^2(X_i) - \frac{1}{2\lambda_2} \sum_{i \in S_1} \max\{\lambda_2 + f(X_i), 0\}.$$

Obtaining Fenchel conjugate for $r(f)$

The Fenchel conjugate of $r(f)$ can be derived similarly:

$$r^*(f) = \sup_{g \in \mathcal{H}} \left\{ \langle f, g \rangle_{\mathcal{H}} - \frac{1}{2} \|g - nh\|_{\mathcal{H}}^2 + \frac{\lambda_{1n}}{2} \right\}.$$

By using functional differentiation, the optimum g satisfies $f - g^* + nh = 0$, hence $g^* = f + nh$. By plugging this in,

$$r^*(f) = \frac{1}{2} \|f\|_{\mathcal{H}}^2 + \sum_{i \in [n]} f(X_i) + \frac{\lambda_{1n}}{2}.$$

Combining $s^*(-M^* f)$ and $r^*(f)$, we obtain the following Fenchel-Rockafeller dual as

$$d(f) = \frac{1}{2\lambda_2} \sum_{i \in S_1} f^2(X_i) + \frac{1}{2} \|f\|_{\mathcal{H}}^2 + \sum_{i \in [n]} f(X_i) - \frac{1}{2\lambda_2} \sum_{i \in S_1} \max\left[\{\lambda_2 + f(X_i)\}^2, 0 \right] + \frac{\lambda_{1n}}{2}.$$

Obtaining $\hat{w}_i$

Let $\hat{f} = \operatorname{argmin}_{f \in \mathcal{H}} d(f)$. By Theorem 3.51 of Peypouquet (2015), $d(\hat{f}) = -p(\hat{w})$, and $-M^* \hat{f} \in \partial s(\hat{w}) = \{\lambda_2 \hat{w} - \gamma; \gamma \succeq 0\}$. Therefore, for $i \in S_1$,

$$-\hat{f}(X_i) = \lambda_2 \hat{w}_i - \max\{\lambda_2 + \hat{f}(X_i), 0\},$$

By rearranging,

$$\hat{w}_i = \frac{-\hat{f}(X_i) + \max\{\lambda_2 + \hat{f}(X_i), 0\}}{\lambda_2} = \max\left\{1, -\hat{f}(X_i)/\lambda_2\right\}. \tag{A.2}$$

A.1.2 Step 2: Limiting form of the dual problem

By dividing $d(f)$ by n, we get

$$\frac{1}{n}d(f) = \frac{1}{2\lambda_2}\frac{1}{n}\sum_{i\in S_1} f^2(X_i) + \frac{1}{2n}\|f\|^2_{\mathcal{H}} + \frac{1}{n}\sum_{i\in[n]} f(X_i) - \frac{1}{2\lambda_2}\frac{1}{n}\sum_{i\in S_1}\max\{\lambda_2 + f(X_i), 0\} + \frac{\lambda_{1n}}{2n}.$$

As $n \to \infty$, this converges to

$$J(f) = \frac{1}{2\lambda_2}E\left[\pi(+1, X)f^2(X)\right] + \mathbb{E}[f(X)] - \frac{1}{2\lambda_2}E\left[\pi(+1, X)\max\{\lambda_2 + f(X), 0\}^2\right] + c_2,$$

where $c_2 \in \mathbb{R}$ is a constant from that $\lambda_{1n}/n = n\lambda_1 \asymp 1$. When $\lambda_2 + f(x) \leq 0$, the derivative of $\max\{\lambda_2 + f(x), 0\}^2$ with respect to f is 0. Otherwise, the derivative of $\max\{\lambda_2 + f(x), 0\}^2$ with respect to f is $2\max\{\lambda_2 + f(x), 0\}$. Thus, $J(f)$ is minimized with f^* such that

$$\frac{1}{\lambda_2}\pi(+1, x)f^*(x) + 1 - \frac{1}{\lambda_2}\pi(+1, x)\max\{\lambda_2 + f^*(x), 0\} = 0.$$

Then,

$$f^*(x) = -\lambda_2\frac{1}{\pi(+1, X)} + \max\{\lambda_2 + f^*(x), 0\}.$$

By plugging this in equation (A.2), we obtain

$$\begin{aligned} w_i^* &= \frac{-f^*(X_i) + \max\{\lambda_2 + f^*(x), 0\}}{\lambda_2} \\ &= \frac{1}{\lambda_2}\left\{-\left(-\lambda_2\frac{1}{\pi(+1, X_i)} + \max\{\lambda_2 + f^*(x), 0\}\right) + \max\{\lambda_2 + f^*(x), 0\}\right\} \\ &= \frac{1}{\pi(+1, X_i)}. \end{aligned}$$

A.1.3 Step 3: Identification of the convergence in probability

We show the convergence $n^{-1}\sum_{i\in S_1}(\hat{w}_i^{un} - w_i^*)^2 \to 0$ as $n \to \infty$, where $\hat{w}^{un}$ is obtained from the optimization without the constraint $w \succeq 1$. Hirshberg and Wager (2017) shows a general case of convergence of the unconstrained problem. Since $\hat{w}_i = \max(\hat{w}_i^{un}, 1)$ from (A.2), we

have the following inequality:

$$\begin{aligned}
\frac{1}{n}\sum_{i\in S_1}(\hat{w}_i^{un}-w_i^*)^2 &= \frac{1}{n}\sum_{i\in S_1:w_i^*>1}(\hat{w}_i^{un}-w_i^*)^2+\frac{1}{n}\sum_{i\in S_1:w_i^*=1}(\hat{w}_i^{un}-w_i^*)^2\\
&= \frac{1}{n}\sum_{i\in S_1:w_i^*>1;\hat{w}^{un}\geq 1}(\hat{w}_i^{un}-w_i^*)^2+\frac{1}{n}\sum_{i\in S_1:w_i^*>1;\hat{w}^{un}<1}(\hat{w}_i^{un}-w_i^*)^2\\
&\quad+\frac{1}{n}\sum_{i\in S_1:w_i^*=1;w_i^{un}>1}(\hat{w}_i^{un}-w_i^*)^2+\frac{1}{n}\sum_{i\in S_1:w_i^*=1;w_i^{un}<1}(\hat{w}_i^{un}-w_i^*)^2\\
&\geq \frac{1}{n}\sum_{i\in S_1:w_i^*>1;\hat{w}^{un}\geq 1}(\hat{w}_i^{un}-w_i^*)^2+\frac{1}{n}\sum_{i\in S_1:w_i^*>1;\hat{w}^{un}<1}(\hat{w}_i-w_i^*)^2\\
&\quad+\frac{1}{n}\sum_{i\in S_1:w_i^*=1;w_i^{un}>1}(\hat{w}_i^{un}-w_i^*)^2+\frac{1}{n}\sum_{i\in S_1:w_i^*=1;w_i^{un}<1}(\hat{w}_i-w_i^*)^2\\
&= \frac{1}{n}\sum_{i\in S_1:}(\hat{w}_i-w_i^*)^2,
\end{aligned}$$

which follows from

$$\sum_{i\in S_1:w_i^*>1;\hat{w}^{un}<1}(\hat{w}_i^{un}-w_i^*)^2\geq\sum_{i\in S_1:w_i^*>1;\hat{w}^{un}<1}(\hat{w}_i-w_i^*)^2=\sum_{i\in S_1:w_i^*>1;\hat{w}^{un}<1}(1-w_i^*)^2$$

and

$$\sum_{i\in S_1:w_i^*=1;w_i^{un}<1}(\hat{w}_i^{un}-w_i^*)^2\geq\sum_{i\in S_1:w_i^*=1;w_i^{un}<1}(\hat{w}_i-w_i^*)^2=0.$$

Without the Lagrangian constraint, we have $\hat{w}_i^{un}=-\lambda_2^{-1}\hat{f}(X_i)$ and $w_i^*=-\lambda_2^{-1}f^*(X_i)=1/\pi(+1,X_i)$. Note that $w_i^*=1/\pi(+1,X_i)$ with and without the constraint in the optimization of the limiting term of $d(f)$. Define the penalized least squares problem

$$\frac{1}{n}\sum_{i\in S_1}(w_i-w_i^*)^2+\frac{\lambda_2}{n}\|w\|_2^2.$$

This can be rewritten as follows from the dual expression:

$$\frac{1}{n}\sum_{i\in S_1}\left\{\frac{-f(X_i)}{\lambda_2}-w_i^*\right\}^2+\frac{\lambda_2}{n}\left\|\frac{-f}{\lambda_2}\right\|_{\mathcal{H}}^2=\frac{1}{n}\sum_{i\in S_1}\left\{\frac{-f(X_i)}{\lambda_2}-w_i^*\right\}^2+\frac{1}{n\cdot\lambda_2}\|f\|_{\mathcal{H}}^2$$

whose minimization is equivalent to that of

$$\tilde{d}(f) = \frac{\lambda_2}{2} \sum_{i \in S_1} \left\{ \frac{-f(X_i)}{\lambda_2} - w_i^* \right\}^2 + \frac{1}{2} \|f\|_{\mathcal{H}}^2.$$

Let $\tilde{f} = \text{argmin}_{f \in \mathcal{H}} \tilde{d}(f)$ and $\delta = f - \tilde{f}$. Then

$$\begin{aligned} 0 &= \lim_{t \to 0} \frac{\tilde{d}(\tilde{f} + t\delta) - \tilde{d}(\tilde{f})}{t} \\ &= -\sum_{i \in S_1} \frac{-\tilde{f}(X_i)}{\lambda_2} \delta(X_i) + \sum_{i \in S_1} w_i^* \delta(X_i) + \langle \tilde{f}, \delta \rangle_{\mathcal{H}} \\ &= -\sum_{i \in S_1} (\tilde{w}_i - w_i^*) \delta(X_i) + \langle \tilde{f}, \delta \rangle_{\mathcal{H}}. \end{aligned}$$

Let the Fenchel-Rockafellar dual without the Lagrangian constraint be

$$d_{un}(f) = \frac{1}{2\lambda_2} \sum_{i \in S_1} f^2(X_i) + \frac{1}{2} \|f\|_{\mathcal{H}}^2 + \sum_{i \in [n]} f(X_i) + \frac{\lambda_{1n}}{2}.$$

From this, define $R(\delta)$ as follows:

$$\begin{aligned} R(\delta) &= \frac{1}{n} d_{un}(\tilde{f} + \delta) - \frac{1}{n} d_{un}(\tilde{f}) \\ &= \frac{1}{n} \frac{1}{2\lambda_2} \sum_{i \in S_1} \delta^2(X_i) - \frac{1}{n} \sum_{i \in S_1} \frac{-\tilde{f}(X_i)}{\lambda_2} \delta(X_i) + \frac{1}{n} \sum_{i \in [n]} \delta(X_i) + \frac{1}{2n} \|f\|_{\mathcal{H}}^2 - \frac{1}{2n} \|\tilde{f}\|_{\mathcal{H}}^2 \\ &= \frac{1}{n} \frac{1}{2\lambda_2} \sum_{i \in S_1} \delta^2(X_i) - \frac{1}{n} \sum_{i \in S_1} \tilde{w}_i \delta(X_i) + \frac{1}{n} \sum_{i \in [n]} \delta(X_i) + \frac{1}{2n} \|f\|_{\mathcal{H}}^2 - \frac{1}{2n} \|\tilde{f}\|_{\mathcal{H}}^2 \\ &= \frac{1}{n} \frac{1}{2\lambda_2} \sum_{i \in S_1} \delta^2(X_i) - \frac{1}{n} \sum_{i \in S_1} \tilde{w}_i \delta(X_i) + \frac{1}{n} \sum_{i \in [n]} \delta(X_i) + \frac{1}{2n} \|f\|_{\mathcal{H}}^2 - \frac{1}{2n} \|\tilde{f}\|_{\mathcal{H}}^2 \\ &\quad + \frac{1}{n} \sum_{i \in S_1} (\tilde{w}_i - w_i^*) \delta(X_i) - \frac{1}{n} \langle \tilde{f}, \delta \rangle_{\mathcal{H}} \qquad \text{(A.3)} \\ &= \frac{1}{2\lambda_2 n} \sum_{i \in S_1} \delta^2(X_i) - \frac{1}{n} \sum_{i \in S_1} w_i^* \delta(X_i) + \frac{1}{n} \sum_{i \in [n]} \delta(X_i) + \frac{1}{2n} \|f - \tilde{f}\|_{\mathcal{H}}^2 \\ &= G(\delta) - H(\delta) + \frac{1}{2n} \|f - \tilde{f}\|_{\mathcal{H}}^2, \end{aligned}$$

where

$$G(\delta) = \frac{1}{2\lambda_2 n} \sum_{i \in S_1} \delta^2(X_i) \text{ and}$$

$$H(\delta) = \frac{1}{n} \sum_{i \in S_1} w_i^* \delta(X_i) - \frac{1}{n} \sum_{i \in [n]} \delta(X_i),$$

and the lines from (S4) are derived by subtracting $n^{-1}\left\{-\sum_{i \in S_1}(\tilde{w}_i - w_i^*)\delta(X_i) + \langle \tilde{f}, \delta \rangle_{\mathcal{H}}\right\}$. From Section A.2 of Hirshberg and Wager (2017), we can show that $\hat{w} \approx \tilde{w}$ as long as $R(\delta) \leq 0$.

By Lemma A.1 and Lemma A.2, there exists $d_n = o(n^{-1/4})$ such that for all $\epsilon > 0$ and $u \in \mathcal{H}_1$,

$$\mathbb{E}[u^2(X)] \geq d_n^2 \qquad \text{implies } n_1^{-1} \sum_{i \in S_1} u^2(X_i) \geq 2^{-1}\mathbb{E}[u^2(X)] \tag{A.4}$$

$$\mathbb{E}[u^2(X)] \leq d_n^2 \qquad \text{implies } |H(u)| \leq d_n^2 \tag{A.5}$$

with probability at least $1-\epsilon$. Define $b_n = \max\{4\lambda_2 \cdot n/n_1, 2n \cdot d_n^2\}$, where $n_1 = \sum_{i \in [n]} \mathbb{1}(A_i = +1)$. Consider the case when $\delta/b_n \in \mathcal{H}_1$ from which $\|\delta/b_n\|_{\mathcal{H}} \leq 1$, hence $\|\delta\|_{\mathcal{H}} \leq b_n$. Suppose $\mathbb{E}[\delta^2(X)] \leq b_n^2 d_n^2$, so that $\mathbb{E}[\delta^2(X)/b_n^2] \leq d_n^2$. Then by Lemma A.2, $|H(\delta/b_n)| \leq d_n^2$, and

$$|H(\delta)| = b_n \cdot |H(\delta/b_n)| \leq b_n d_n^2.$$

This leads to the following inequality:

$$R(\delta) \geq G(\delta) - |H(\delta)| \geq G(\delta) - b_n d_n^2.$$

Therefore, $R(\delta) \leq 0$ is only possible when $G(\delta) \leq b_n d_n^2$.

On the other hand, suppose $\mathbb{E}[\delta^2(X)] \geq b_n^2 d_n^2$. Let $n_1 = |S_1|$. By Lemma A.1,

$$\frac{1}{n_1} \sum_{i \in S_1} \delta^2(X_i) \geq \frac{1}{2}\mathbb{E}[\delta^2(X)] = \frac{1}{2}\beta^2,$$

with probability at least $1 - \epsilon$, where $\beta = \sqrt{\mathbb{E}[\delta^2(X)]}$. Note that $\beta \geq b_n d_n$. By rearranging,

we obtain

$$\frac{1}{2\lambda_2 n}\sum_{i\in S_1}\delta^2(X_i) \geq \frac{n_1}{4\lambda_2 n}\beta^2.$$

Define $\delta_1 = d_n \cdot \delta/\beta$. Then

$$\|\delta_1\|_{\mathcal{H}} = \left\|\frac{d_n\cdot\delta}{\beta}\right\|_{\mathcal{H}} = \frac{\|\delta\|_{\mathcal{H}} d_n}{\beta} \leq \frac{b_n\cdot d_n}{\beta} \leq 1.$$

Also, we have

$$\mathbb{E}[\delta_1^2(X)] = \mathbb{E}\left[\frac{d_n^2\cdot\delta^2(X)}{\beta^2}\right] = \frac{d_n^2\cdot\mathbb{E}[\delta^2(X)]}{\beta^2} \leq d_n^2.$$

By Lemma A.2, $|H(\delta_1)| \leq d_n^2$. Thus,

$$|H(\delta)| = \left|H\left(\frac{\beta\delta_1}{d_n}\right)\right| = \frac{\beta|H(\delta_1)|}{d_n} \leq \frac{\beta}{d_n}d_n^2 = \beta d_n = \frac{\beta^2 d_n}{\beta} \leq \frac{\beta^2}{b_n}.$$

where the last inequality comes from $1/b_n \geq d_n/\beta$. Note that

$$G(\delta) = \frac{n_1}{2\lambda_2 n}\cdot\frac{1}{n_1}\sum_{i\in S_1}\delta^2(X_i) \geq \frac{n_1}{2\lambda_2 n}\cdot\frac{1}{2}\mathbb{E}[\delta^2(X)] = \frac{n_1}{2\lambda_2 n}\cdot\frac{1}{2}\beta^2 = \frac{n_1}{4\lambda_2 n}\beta^2,$$

hence

$$R(\delta) \geq G(\delta) - |H(\delta)| \geq \frac{n_1}{4\lambda_2 n}\beta^2 - \frac{1}{b_n}\beta^2 \geq 0$$

from that $b_n \geq 4\lambda_2 n/n_1$.

Now let $\|\delta\|_{\mathcal{H}} > b_n$. Define $k = \|\delta\|_{\mathcal{H}}/(b_n) > 1$. Let $\delta_2 = \delta/k$. Then

$$\|\delta_2\|_{\mathcal{H}} = \left\|\frac{\delta}{k}\right\|_{\mathcal{H}} = \frac{1}{k}\|\delta\|_{\mathcal{H}} \leq \frac{b_n}{\|\delta\|_{\mathcal{H}}}\|\delta\|_{\mathcal{H}} \leq b_n.$$

Note that $\delta = k\delta_2$. Then

$$\begin{aligned} R(\delta) - kR(\delta_2) &= R(k\delta_2) - kR(\delta_2) \\ &= k(k-1)\frac{1}{2\lambda_2 n}\sum_{i\in S_1}\delta_2^2(X_i) - (k-k)H(\delta_2) + k(k-1)\frac{1}{2n}\|\delta_2\|_{\mathcal{H}}^2 \\ &\geq 0. \end{aligned}$$

This shows that $R(\delta) \geq kR(\delta_2) \geq R(\delta_2)$. Therefore, $R(\delta_2) \geq 0$ implies $R(\delta) \geq 0$. Note that

$\|\delta_2\|_{\mathcal{H}} \leq b_n$. When $\mathbb{E}[\delta_2^2(X)] > b_n^2 d_n^2$, $R(\delta_2) \geq G(\delta_2) - |H(\delta_2)| \geq 0$ from the similar derivation as above. Suppose $\mathbb{E}[\delta_2^2(X)] \leq b_n^2 d_n^2$. Then, also from the derivation above, $|H(\delta_2)| \leq b_n d_n^2$, and

$$R(\delta_2) \geq G(\delta_2) - |H(\delta_2)| + \frac{1}{2n}\|\delta_2\|_{\mathcal{H}}^2.$$

From this, we have the following:

$$R(\delta_2) \geq -|H(\delta_2)| + \frac{1}{2n}\|\delta_2\|_{\mathcal{H}}^2 \geq -b_n d_n^2 + \frac{1}{2n} b_n^2.$$

The lower bound is non-negative since $b_n \geq 2nd_n^2$.

Therefore, $R(\delta) \leq 0$ with probability at least $1-\epsilon$ only when $G(\delta) \leq b_n d_n^2$. Since $\hat{f} = \tilde{f} + \hat{\delta}$ minimizes $d(f)$, this inequality holds when $\hat{\delta} = \hat{f} - \tilde{f}$, from which $G(\hat{\delta}) \leq b_n d_n^2$. Then

$$\begin{aligned} G(\hat{\delta}) &= \frac{1}{2\lambda_2 n} \sum_{i \in S_1} \hat{\delta}^2(X_i) = \frac{\lambda_2}{2n} \sum_{i \in S_1} \left(\frac{\hat{f}(X_i) - \tilde{f}(X_i)}{\lambda_2} \right)^2 \\ &= \frac{\lambda_2}{2n} \sum_{i \in S_1} (\hat{w}_i^{un} - \tilde{w}_i)^2 \leq b_n d_n^2. \end{aligned}$$

Since $d_n = o(n^{-1/4})$ and $b_n = O(\sqrt{n})$, $b_n d_n^2 \to 0$ as $n \to \infty$. When n is sufficiently large, there exists a small $\eta > 0$ such that $\eta < b_n d_n^2$. From this,

$$\Pr\left(\frac{\lambda_2}{2n} \sum_{i \in S_1} (\hat{w}_i^{un} - \tilde{w}_i)^2 \leq \eta \right) \geq \Pr\left(\frac{\lambda_2}{2n} \sum_{i \in S_1} (\hat{w}_i^{un} - \tilde{w}_i)^2 \leq b_n d_n^2 \right) \geq 1 - \epsilon.$$

Therefore, $\frac{1}{n}\sum_{i \in S_1} (\hat{w}_i^{un} - \tilde{w}_i)^2 \to 0$ in probability.

From the regularity condition that $\mathcal{H}$ is universal with respect to the supremum norm, we can choose a sequence $\{u_j; j \in \mathbb{N}\} \subset \mathcal{H}$ such that $\|u_j - f^*\|_\infty \to 0$ as $j \to \infty$, where $\mathbb{N}$ denotes the set of natural numbers. Moreover, we can choose the subsequence $\{u_{jn}\}$ such that $\|u_{jn}\|_{\mathcal{H}}/\sqrt{n} = o(1)$ as $n \to \infty$. We can find the corresponding weights $v_{jn,i} = -u_{jn}(X_i)/\lambda_2$ from which

$$\begin{aligned} \max_{i \in S_1} |v_{jn,i} - w_i^*| &= \max_{i \in S_1} \left| \frac{-v_{jn}(X_i)}{\lambda_2} - \frac{-f^*(X_i)}{\lambda_2} \right| \\ &= \lambda_2^{-1} \max_{i \in S_1} | - v_{jn}(X_i) + f^*(X_i)| \to 0. \end{aligned}$$

Since $\tilde{f}$ is the minimizer of $\tilde{d}(f)$, we have

$$\begin{aligned}\frac{1}{n}\sum_{i\in S_1}(\tilde{w}_i - w_i^*)^2 &\leq \frac{2}{\lambda_2 n}\tilde{d}(\tilde{f}) \\ &= \frac{1}{n}\sum_{i\in S_1}(\tilde{w}_i - w_i^*)^2 + \frac{1}{\lambda_2 n}\|\tilde{f}\|_{\mathcal{H}}^2 \\ &\leq \frac{2}{\lambda_2 n}\tilde{d}(u_{jn}) \leq \max_{i\in S_1}(v_{jn,i} - w_i^*)^2 + \frac{1}{\lambda_2}\frac{\|u_{jn}\|_{\mathcal{H}}^2}{\sqrt{n}}.\end{aligned}$$

The last line converges to 0 as $n \to \infty$, so $\frac{1}{n}\sum_{i\in S_1}(\tilde{w}_i - w_i^*)^2 \to 0$ in probability. By triangular inequality,

$$\sqrt{\frac{1}{n}\sum_{i\in S_1}(\hat{w}_i^{un} - w_i^*)^2} \leq \sqrt{\frac{1}{n}\sum_{i\in S_1}(\hat{w}_i^{un} - \tilde{w}_i)^2} + \sqrt{\frac{1}{n}\sum_{i\in S_1}(\tilde{w}_i - w_i^*)^2}.$$

Therefore, $n^{-1}\sum_{i\in S_1}(\hat{w}_i^{un} - w_i^*)^2 \to 0$ in probability. This implies $n^{-1}\sum_{i\in S_1}(\hat{w}_i - w_i^*)^2 = o_p(1)$.

Lemma A.1. Suppose $u \in \mathcal{H}_1$ has the second moment $\mathbb{P}u^2 \geq d^2$, where $d \geq d_n = o(n^{-1/4})$. Then

$$\Pr\left(\mathbb{P}_n u^2 \geq \frac{1}{2}\mathbb{P}u^2\right) \geq 1 - 2\exp(-d_1 n d_n^2 / M_\infty^2(\mathcal{H}_1)),$$

where $M_\infty(\mathcal{H}_1) = \sup_{u\in\mathcal{H}_1}\sup_{x\in\mathcal{X}}|f(x)|$, and d_1 is a constant that depends on the multiplier $1/2$.

Proof. The proof follows from Lemma 4 and Section A.5.1 of Hirshberg and Wager (2017) that uses Corollary 3.3 of Mendelson (2017) and introduces the uniform lower bound on the ratio of the empirical and the true second moments that gives

$$\Pr\left(\mathbb{P}_n u^2 / Pu^2 \geq \xi\right) \geq 1 - 2\exp\{-d_1 n d_n^2 / M_\infty^2(\mathcal{H}_1)\},$$

where $\xi < 1$. We chose $\xi = 1/2$. □

Lemma A.2. Define an operator $H : \mathcal{H} \mapsto \mathbb{R}$ as

$$H(u) = \frac{1}{n}\sum_{i\in S_1} w_i^* u(X_i) - \frac{1}{n}\sum_{i\in[n]} u(X_i).$$

Suppose $u \in \mathcal{H}_1$ has the second moment $\mathbb{P}u^2 \leq d_n^2$, where $d_n = o(n^{-1/4})$. Then for all $\epsilon > 0$,

$$\Pr(|H(u)| \leq d_n^2) \geq 1 - \epsilon.$$

Proof. Note that $\mathbb{E}[\mathbb{1}(A = +1)w^*(X)u(X)] = E[u(X)]$. Therefore,

$$\begin{aligned} H(u) &= \frac{1}{n}\sum_{i \in S_1} w_i^* u(X_i) - \frac{1}{n}\sum_{i \in [n]} u(X_i) \\ &= \frac{1}{n}\sum_{i \in S_1} w_i^* u(X_i) - \frac{1}{n}\sum_{i \in [n]} u(X_i) - \mathbb{E}[\mathbb{1}(A = +1)w^*(X)u(X)] + \mathbb{E}[u(X)] \\ &= (\mathbb{P}_n - \mathbb{P})\{\mathbb{1}(A = +1)w^* u\} - (\mathbb{P}_n - \mathbb{P})(u). \end{aligned}$$

Since $\pi(+1, x) > 0$, $\mathbb{1}(A = +1)w^*(x)$ is bounded above. Hence the function class $\{t \cdot w^*(x); t \in \{0, 1\}\}$ is uniformly bounded and P-Donsker. Also, from the regularity condition, for all $u \in \mathcal{H}_1$,

$$|u(x)| = |\langle u, K(\cdot, x)\rangle_{\mathcal{H}}| \leq \|u\|_{\mathcal{H}}\sqrt{K(x,x)} \leq \sqrt{K(x,x)} < \infty.$$

Therefore, from Corollary 9.32 (v) of Kosorok (2008), $\{\mathbb{1}(A = +1)w^* u : u \in \mathcal{H}_1\}$ is a $\mathbb{P}$-Donsker class. From Section A.5.1 of Hirshberg and Wager (2017), there exists $d_{1n} = o(n^{-1/4})$ such that

$$\Pr\Big((\mathbb{P}_n - \mathbb{P})\{\mathbb{1}(A = +1)w^* u \leq d_{1n}^2/2\}\Big) \geq 1 - \epsilon/2.$$

Similarly, for $d_{2n} = o(n^{-1/4})$, $\Pr\{(\mathbb{P}_n - \mathbb{P})(u) \leq d_{2n}^2/2\} \geq 1 - \epsilon/2$. Define $d_n = \max(d_{1n}, d_{2n})$. By combining these two inequalities, we have

$$\Pr(|H(u)| \leq d_n^2) \geq 1 - \epsilon.$$

□

A.2 Convergence of $g_{\hat{w}}(x)$ to $g(x)$

We first demonstrate that $\sup_{x\in\mathcal{X}} |g_n(x) - g(x)| = o_p(1)$ for compact $\mathcal{X}$. Then we use this result to show that $\sup_{x\in\mathcal{X}} |g_{\hat{W}}(x) - g(x)| = o_p(1)$

A.2.1 Convergence of $g_n(x)$ to $g(x)$

We show that $g_n(x)$ uniformly converges to $g(x)$ in probability for compact $\mathcal{X}$. Let $\tilde{Y} = (\pi^{-1}(A, X) - 1)\, Y$. Then

$$\begin{aligned} |g_n(x) - g(x)| &= \left| \frac{1}{n} \sum_{i\in[n]} x^\top (n^{-1}\mathbb{X}^\top\mathbb{X})^{-1} X_i \tilde{Y}_i - \mathbb{E}[x^\top \{\mathbb{E}(XX^\top)\}^{-1} X\tilde{Y}] \right| \\ &= \left| x^\top (n^{-1}\mathbb{X}^\top\mathbb{X})^{-1} \left(n^{-1} \sum_{i\in[n]} X_i \tilde{Y}_i \right) - x^\top \{\mathbb{E}(XX^\top)\}^{-1} \mathbb{E}[X\tilde{Y}] \right| \end{aligned}$$

By the weak law of large numbers,

$$(n^{-1}\mathbb{X}^\top\mathbb{X})^{-1} = (\mathbb{E}[XX^\top] + R_n)^{-1}; \quad n^{-1} \sum_{i\in[n]} X_i \tilde{Y}_i = \mathbb{E}[X\tilde{Y}] + r_n,$$

where $R_n \to 0_{p\times p}$ and $r_n \to 0_p$ as $n \to \infty$ in probability. From this

$$(n^{-1}\mathbb{X}^\top\mathbb{X})^{-1} \left(n^{-1} \sum_{i\in[n]} X_i \tilde{Y}_i \right) \to_p \{\mathbb{E}(XX^\top)\}^{-1} \mathbb{E}[X\tilde{Y}].$$

By the Cauchy-Schwarz inequality,

$$\begin{aligned}
\sup_{x\in\mathcal{X}}|g_n(x)-g(x)| &= \sup_{x\in\mathcal{X}}\left|x^\top(n^{-1}\mathbb{X}^\top\mathbb{X})^{-1}\left(n^{-1}\sum_{i\in[n]}X_i\tilde{Y}_i\right)-x^\top\{\mathbb{E}(XX^\top)\}^{-1}\mathbb{E}[X\tilde{Y}]\right| \\
&= \sup_{x\in\mathcal{X}}\left|x^\top\left\{(n^{-1}\mathbb{X}^\top\mathbb{X})^{-1}\left(n^{-1}\sum_{i\in[n]}X_i\tilde{Y}_i\right)-\{\mathbb{E}(XX^\top)\}^{-1}\mathbb{E}[X\tilde{Y}]\right\}\right| \\
&\le \sup_{x\in\mathcal{X}}\|x\|_2\cdot\left\|(n^{-1}\mathbb{X}^\top\mathbb{X})^{-1}\left(n^{-1}\sum_{i\in[n]}X_i\tilde{Y}_i\right)-\{\mathbb{E}(XX^\top)\}^{-1}\mathbb{E}[X\tilde{Y}]\right\|_2 \\
&= \sup_{x\in\mathcal{X}}\|x\|\cdot o_p(1)=o_p(1),
\end{aligned}$$

since $\sup_{x\in\mathcal{X}}\|x\|<\infty$ follows from compactness of $\mathcal{X}$.

A.2.2 Proof of Corollary 2.4.2

For any $x\in\mathcal{X}$, by Cauchy-Schwarz inequality,

$$\begin{aligned}
\sup_{x\in\mathcal{X}}|g_{\hat{w}}(x)-g_n(x)| &= \sup_{x\in\mathcal{X}}\left|\frac{1}{n}\sum_{i\in[n]}x^\top(n^{-1}\mathbb{X}^\top\mathbb{X})^{-1}X_i(\hat{w}_i-1)Y_i\right. \\
&\quad\left.-\frac{1}{n}\sum_{i\in[n]}x^\top(n^{-1}\mathbb{X}^\top\mathbb{X})^{-1}X_i\left\{\frac{1}{\pi(A_i,X_i)}-1\right\}Y_i\right| \\
&= \sup_{x\in\mathcal{X}}\left|\frac{1}{n}\sum_{i\in[n]}x^\top(n^{-1}\mathbb{X}^\top\mathbb{X})^{-1}X_i\left\{\hat{w}_i-\frac{1}{\pi(A_i,X_i)}\right\}Y_i\right| \\
&\le \frac{1}{n}\sum_{i\in[n]}\sup_{x\in\mathcal{X}}|x^\top(n^{-1}\mathbb{X}^\top\mathbb{X})^{-1}X_i|\cdot|Y_i|\cdot\left|\hat{w}_i-\frac{1}{\pi(A_i,X_i)}\right| \\
&\le \frac{1}{n}\sum_{i\in[n]}\sup_{x\in\mathcal{X}}\|x\|_2\left\|(n^{-1}\mathbb{X}^\top\mathbb{X})^{-1}\right\|_2\cdot\|X_i\|_2\cdot|Y_i|\cdot\left|\hat{w}_i-\frac{1}{\pi(A_i,X_i)}\right| \\
&= \sup_{x\in\mathcal{X}}\|x\|_2\left\|(n^{-1}\mathbb{X}^\top\mathbb{X})^{-1}\right\|_2\cdot\max_{i\in[n]}\|X_i\|_2\cdot\sqrt{\frac{1}{n}\sum_{i\in[n]}Y_i^2}\sqrt{\frac{1}{n}\sum_{i\in[n]}\left\{\hat{w}_i-\frac{1}{\pi(A_i,X_i)}\right\}^2}.
\end{aligned}$$

Since $\mathcal{X}$ is compact $\sup_{x\in\mathcal{X}}\|x\|_2 < \infty$ and $\max_{i\in[n]}\|X_i\|_2 < \infty$. Also, by weak law of large numbers, $\left\|(n^{-1}\mathbb{X}^\top\mathbb{X})^{-1}\right\|_2 \to_p \left\|\{\mathbb{E}[XX^\top]\}^{-1}\right\|_2$. Thus

$$\sup_{x\in\mathcal{X}}\|x\|_2\cdot\left\|(n^{-1}\mathbb{X}^\top\mathbb{X})^{-1}\right\|_2\cdot\max_{i\in[n]}\|X_i\|_2 = \sup_{x\in\mathcal{X}}\|x\|_2\cdot\left\|\{\mathbb{E}[XX^\top]\}^{-1}\right\|_2\cdot\max_{i\in[n]}\|X_i\|_2 + o_p(1).$$

Hence, we can express the left hand side by $O_p(1)$. From Theorem 2.4.1,

$$\begin{aligned}\sup_{x\in\mathcal{X}}|g_{\hat{w}}(x) - g_n(x)| &\le O_p(1)\sqrt{\frac{1}{n}\sum_{i\in[n]}Y_i^2}\sqrt{\frac{1}{n}\sum_{i\in[n]}\left\{\hat{w}_i - \frac{1}{\pi(A_i,X_i)}\right\}^2}\\ &= O_p(1)\mathbb{E}[Y^2]\cdot o_p(1)\\ &= o_p(1).\end{aligned}$$

Therefore, $g_{\hat{w}}(x) \to g(x)$ over $x\in\mathcal{X}$ in probability from that $\sup_{x\in\mathcal{X}}|g_{\hat{w}}(x) - g(x)| \le \sup_{x\in\mathcal{X}}|g_{\hat{w}}(x) - g_n(x)| + \sup_{x\in\mathcal{X}}|g_n(x) - g(x)| = o_p(1)$ by triangular inequality.

A.3 Proof of Theorem 2.4.3

Let $w_i^* = 1/\pi(A_i, X_i)$. Then

$$\begin{aligned}\frac{1}{n}\sum_{i\in[n]}\tilde{Z}_i f(X_i) &= \frac{1}{n}\sum_{i\in[n]}\hat{w}_i A_i\{Y_i - g_{\hat{w}}(X_i)\}f(X_i)\\ &= \frac{1}{n}\sum_{i\in[n]}(\hat{w}_i - w_i^*)A_i\{Y_i - g_{\hat{w}}(X_i)\}f(X_i) + \frac{1}{n}\sum_{i\in[n]}w_i^* A_i\{Y_i - g_{\hat{w}}(X_i)\}f(X_i)\\ &= \frac{1}{n}\sum_{i\in[n]}(\hat{w}_i - w_i^*)A_i\{Y_i - g(X_i)\}f(X_i) + \frac{1}{n}\sum_{i\in[n]}(\hat{w}_i - w_i^*)A_i\{g(X_i) - g_{\hat{w}}(X_i)\}f(X_i)\\ &\quad + \frac{1}{n}\sum_{i\in[n]}w_i^* A_i\{g(X_i) - g_{\hat{w}}(X_i)\}f(X_i) + \frac{1}{n}\sum_{i\in[n]}w_i^* A_i\{Y_i - g(X_i)\}f(X_i)\\ &= \frac{1}{n}\sum_{i\in[n]}(\hat{w}_i - w_i^*)A_i\{Y_i - g(X_i)\}f(X_i) + \frac{1}{n}\sum_{i\in[n]}(\hat{w}_i - w_i^*)A_i\{g(X_i) - g_{\hat{w}}(X_i)\}f(X_i)\\ &\quad + \frac{1}{n}\sum_{i\in[n]}w_i^* A_i\{g(X_i) - g_{\hat{w}}(X_i)\}f(X_i) + \mathbb{E}[Zf(X)] + o_p(1),\end{aligned}$$

where the last equality comes from the weak law of large numbers. We show that the first three terms are $o_p(1)$. By the Cauchy-Schwarz inequality, the first term is bounded by

$$\begin{aligned}\sqrt{\frac{1}{n}\sum_{i\in[n]}(\hat{w}_i - w_i^*)^2}\sqrt{\frac{1}{n}\sum_{i\in[n]}\{Y_i - g(X_i)\}^2 f^2(X_i)} &= o_p(1)\sqrt{\frac{1}{n}\sum_{i\in[n]}\pi^2(A_i,X_i)w_i^{*2}\{Y_i - g(X_i)\}^2 f^2(X_i)} \\ &\leq o_p(1)\sqrt{\frac{1}{n}\sum_{i\in[n]} Z_i^2 f^2(X_i)} \\ &= o_p(1)\sqrt{\mathbb{E}[Z^2 f^2(X)] + o_p(1)} = o_p(1).\end{aligned}$$

The second term is bounded by

$$\begin{aligned}\sqrt{\frac{1}{n}\sum_{i\in[n]}(\hat{w}_i - w_i^*)^2}\sqrt{\frac{1}{n}\sum_{i\in[n]}\{g(X_i) - g_{\hat{w}}(X_i)\}^2 f^2(X_i)} &\leq o_p(1)\sup_{x\in\mathcal{X}}|g(x) - g_{\hat{w}}(x)|\sqrt{\frac{1}{n}\sum_{i\in[n]} f^2(X_i)} \\ &= o_p(1)\sqrt{\mathbb{E}[f^2(X)] + o_p(1)} = o_p(1),\end{aligned}$$

since $\sup_{x\in\mathcal{X}}|g(x) - g_{\hat{w}}(x)| = o_p(1)$ (Corollary 2.4.2). The third term is bounded by

$$\begin{aligned}\sqrt{\frac{1}{n}\sum_{i\in[n]} w_i^{*2}\{g(X_i) - g_{\hat{w}}(X_i)\}^2}\cdot\sqrt{\frac{1}{n}\sum_{i\in[n]} f^2(X_i)} &\leq \max_{i\in[n]}\{w_i^*\}\sup_{x\in[n]}|g(X_i) - g_{\hat{w}}(X_i)|\cdot\sqrt{\frac{1}{n}\sum_{i\in[n]} f^2(X_i)} \\ &= \max_{i\in[n]}\{w_i^*\}\cdot o_p(1)\sqrt{\mathbb{E}[f^2(X)] + o_p(1)} = o_p(1),\end{aligned}$$

which follows from $1/\pi(A,X) < \infty$ from **Assumption 2**. By combining these results, we have $n^{-1}\sum_{i\in[n]}\tilde{Z}_i f(X_i) = \mathbb{E}[Zf(X)] + o_p(1)$.

A.4 Proof of Theorem 2.4.4

The empirical ϕ-risk can be expressed as

$$\begin{aligned}\frac{1}{n}\sum_{i\in[n]}\hat{w}_i|Y_{g_{\hat{w}},i}|\phi\left\{A_i\cdot\operatorname{sign}(Y_{g_{\hat{w}},i})f(X_i)\right\} &= \frac{1}{n}\sum_{i\in[n]}\left\{\hat{w}_i - \frac{1}{\pi(A_i,X_i)}\right\}|Y_{g_{\hat{w}},i}|\phi\left\{A_i\cdot\operatorname{sign}(Y_{g_{\hat{w}},i})f(X_i)\right\} \\ &\quad + \frac{1}{n}\sum_{i\in[n]}\frac{|Y_i - g_{\hat{w}}(X_i)|}{\pi(A_i,X_i)}\phi\left\{A_i\cdot\operatorname{sign}(Y_{g_{\hat{w}},i})f(X_i)\right\}.\end{aligned}$$

By multiplying and dividing by $\pi(A_i, X_i)$, the first term becomes

$$\frac{1}{n}\sum_{i\in[n]}\left\{\hat{w}_i - \frac{1}{\pi(A_i,X_i)}\right\}\pi(A_i,X_i)\frac{|Y_{g_{\hat{w}},i}|}{\pi(A_i,X_i)}\phi\left\{A_i\cdot\text{sign}(Y_{g_{\hat{w}},i})f(X_i)\right\}.$$

By Cauchy inequality, the first term is bounded by

$$\sqrt{\frac{1}{n}\sum_{i\in[n]}\left(\hat{w}_i - \frac{1}{\pi(A_i,X_i)}\right)^2}\cdot\sqrt{\frac{1}{n}\sum_{i\in[n]}\pi^2(A_i,X_i)\left(\frac{|Y_i - g_{\hat{w}}(X_i)|}{\pi(A_i,X_i)}\phi\left\{A_i\cdot\text{sign}(Y_{g_{\hat{w}}})f(X_i)\right\}\right)^2}$$

$$\leq\sqrt{\frac{1}{n}\sum_{i\in[n]}\left(\hat{w}_i - \frac{1}{\pi(A_i,X_i)}\right)^2}\cdot\sqrt{\frac{1}{n}\sum_{i\in[n]}\left(\frac{|Y_i - g_{\hat{w}}(X_i)|}{\pi(A_i,X_i)}\phi\left\{A_i\cdot\text{sign}(Y_{g_{\hat{w}}})f(X_i)\right\}\right)^2},$$

where the inequality comes from $\pi(A_i, X_i) < 1$. By Corollary 2.4.2,

$$\begin{aligned}\left||Y - g_{\hat{w}}(x)| - |Y - g(x)|\right| &\leq |Y - g_{\hat{w}}(x) - \{Y - g(x)\}| \\ &= |g(x) - g_{\hat{w}}(x)| = o_p(1).\end{aligned}$$

Moreover, by **Assumption 2** and compactness of $\mathcal{X}$, $1/\pi(A, X) \leq M_\pi < \infty$ where $1/M_\pi = \inf_{a\in\mathcal{A},x\in\mathcal{X}}\pi(a, x)$. Then

$$\frac{|Y - g_{\hat{w}}(x)|}{\pi(A,X)} \leq \frac{|Y - g(x)|}{\pi(A,X)} + \frac{|g(x) - g_{\hat{w}}(x)|}{\pi(A,X)} = \frac{|Y - g(x)|}{\pi(A,X)} + o_p(1).$$

Furthermore, $\phi\left\{A_i\cdot\text{sign}(Y_{g_{\hat{w},i}})f(X_i)\right\}$ is either $\phi\left\{f(X_i)\right\}$ or $\phi\left\{-f(X_i)\right\}$. Since $\phi(\cdot)$ is a continuous function and $-M_f \leq f(x) \leq M_f$, $|\phi\left\{f(x)\right\}| \leq C_f$ for a constant $C_f \in \mathbb{R}$. Thus,

$$\frac{1}{n}\sum_{i\in[n]}\left(\frac{|Y_{g_{\hat{w}},i}|}{\pi(A_i,X_i)}\phi\left\{A_i\cdot\text{sign}(Y_{g_{\hat{w}},i})f(X_i)\right\}\right)^2 = \frac{1}{n}\sum_{i\in[n]}\left(\frac{|Y_{g,i}|}{\pi(A_i,X_i)}\phi\left\{A_i\cdot\text{sign}(Y_{g_{\hat{w}},i})f(X_i)\right\}\right)^2 + o_p(1),$$

and

$$\begin{aligned}\frac{1}{n}\sum_{i\in[n]}\left(\frac{|Y_{g,i}|}{\pi(A_i,X_i)}\phi\left\{A_i\cdot\text{sign}(Y_{g_{\hat{w}},i})f(X_i)\right\}\right)^2 &\leq \frac{M_g^2}{n}\sum_{i\in[n]}\phi^2\left\{A_i\cdot\text{sign}(Y_{g_{\hat{w}},i})f(X_i)\right\} \\ &\leq \frac{M_g^2}{n}\sum_{i\in[n]}C_f^2 = M_g^2C_f^2.\end{aligned}$$

Therefore, by Theorem 2.4.1 and Cauchy-Schwarz inequality,

$$\frac{1}{n}\sum_{i\in[n]}\left\{\hat{w}_i - \frac{1}{\pi(A_i, X_i)}\right\}|Y_i - g_{\hat{w}}(X_i)|\phi\left\{A_i \cdot \mathrm{sign}(Y_{g_{\hat{w}},i})f(X_i)\right\} = o_p(1).$$

For the convergence of the second term, we use that $|r|\cdot\phi\{\mathrm{sign}(r)\cdot f\}$ is a continuous function with respect to r. Then by Corollary 2.4.2, continuous mapping theorem, and the law of large numbers, the second term becomes

$$\frac{1}{n}\sum_{i\in[n]}\frac{|Y_i - g_{\hat{w}}(X_i)|}{\pi(A_i, X_i)}\phi\left\{A_i \cdot \mathrm{sign}(Y_{g_{\hat{w}},i})f(X_i)\right\} \to \mathbb{E}\left[\frac{|Y - g(X)|}{\pi(A, X)}\phi\{A \cdot \mathrm{sign}(Y_g)f(X)\}\right].$$

in probability. Therefore,

$$\mathbb{P}_n\hat{w}|Y - g_{\hat{w}}(X)|\phi\left\{A \cdot \mathrm{sign}(Y_{g_{\hat{w}}})f(X)\right\} \to \mathcal{R}_{\phi,g}(f)$$

in probability.

A.5 Proof of Theorem 2.4.5

The proof relies on assumptions (A6) to (A10) adopted from Fukumizu and Leng (2014): (A6) $\mathcal{H}_\mathcal{X}$ is separable; (A7) $K_X(\cdot,\cdot)$ is measurable, and $\mathbb{E}[K_X(X,X)] < \infty$, $\mathbb{E}[Z^2] < \infty$; (A8) $K_X(x,x')$ is continuously differentiable and $\partial K_X(\cdot,x)/\partial x^{(a)}$ is in the range of the covariance operator C_{XX} for $a \in [p]$; (A9) $\mathbb{E}[Z|X=\cdot] \in \mathcal{H}_\mathcal{X}$; (A10) The function $v \mapsto \mathbb{E}[Z|B_0^\top X = v]$ is differentiable for any $v \in \mathcal{V}$. These are not explicitly used in this proof. We additionally introduce assumptions (A11) and (A12) which are also adopted from Fukumizu and Leng (2014): (A11) For each $p = p_n$, there are $\xi_p \geq 0$ and $C_p \geq 0$ such that a function $h_{a,x} \in \mathcal{H}_\mathcal{X}$ satisfies

$$\frac{\partial K_X(\cdot,x)}{\partial x^{(a)}} = C_{XX}^{\xi_p+1}h_{a,x},$$

and $\|h_{a,x}\|_{\mathcal{H}_\mathcal{X}} \leq C_p$ for any $a \in [p]$ and $x \in \mathcal{X}$; (A12) With X', an independent copy of X, define $\alpha_p = \sqrt{\mathbb{E}[K_X^2(X,X)] - \mathbb{E}[K_X^2(X,X')]}$. Then $\alpha_p/\sqrt{n} \to 0$ as $n \to \infty$.

By triangular inequality,

$$\left|\hat{W}_{ij}(x) - M_{ij}(x)\right| \leq \left|\hat{W}_{ij}(x) - \hat{M}_{ij}(x)\right| + \left|\hat{M}_{ij}(x) - M_{ij}(x)\right|.$$

The first term:

$$\left|\hat{W}_{ij}(x) - \hat{M}_{ij}(x)\right|$$

$$\leq \left|\frac{1}{n}\sum_{i\in[n]}(\hat{C}_{XX} + \epsilon_n I)^{-1}\frac{\partial K_X(X_i, x)}{\partial x^{(m)}}\tilde{Z}_i\right| \cdot \left|\frac{1}{n}\sum_{i\in[n]}(\hat{C}_{XX} + \epsilon_n I)^{-1}\frac{\partial K_X(X_i, x)}{\partial x^{(j)}}(\tilde{Z}_i - Z_i)\right|$$

$$+ \left|\frac{1}{n}\sum_{i\in[n]}(\hat{C}_{XX} + \epsilon_n I)^{-1}\frac{\partial K_X(X_i, x)}{\partial x^{(m)}}(\tilde{Z}_i - Z_i)\right| \cdot \left|\frac{1}{n}\sum_{i\in[n]}(\hat{C}_{XX} + \epsilon_n I)^{-1}\frac{\partial K_X(X_i, x)}{\partial x^{(j)}}Z_i\right|.$$

Define $(\hat{C}_{XX} + \epsilon_n I)^{-1}\partial K_X(\cdot, x)/\partial x^{(a)} = \hat{f}_{a,x}(\cdot)$ and $C_{XX}^{-1}\partial K_X(\cdot, x)/\partial x^{(a)} = f_{a,x}(\cdot)$. Let $w_i^* = 1/\pi(A_i, X_i)$. Let $S_1 = \{i \in [n] : A_i = +1\}$. Then

$$\begin{aligned}
\frac{1}{n}\sum_{i\in[n]}\hat{f}_{a,x}(X_i)(\tilde{Z}_i - Z_i) &= \frac{1}{n}\sum_{i\in S_1}\hat{w}_i Y_i \hat{f}_{a,x}(X_i) - \frac{1}{n}\sum_{i\in S_1} w_i^* Y_i \hat{f}_{a,x}(X_i) \\
&\quad - \left\{\frac{1}{n}\sum_{i\notin S_1}^{n}\hat{w}_i Y_i \hat{f}_{a,x}(X_i) - \frac{1}{n}\sum_{i\notin S_1} w_i^* Y_i \hat{f}_{a,x}(X_i)\right\} \\
&\quad - \frac{1}{n}\sum_{i\in S_1}\hat{f}_{a,x}(X_i)\hat{w}_i g_{\hat{w}}(X_i) + \frac{1}{n}\sum_{i\notin S_1}\hat{f}_{a,x}(X_i)\hat{w}_i g_{\hat{w}}(X_i) \\
&= \frac{1}{n}\sum_{i\in[n]} A_i(\hat{w}_i - w_i^*) Y_i \hat{f}_{a,x}(X_i) - \frac{1}{n}\sum_{i\in[n]} A_i \hat{w}_i g_{\hat{w}}(X_i)\hat{f}_{a,x}(X_i).
\end{aligned}$$

Then the first term in the latter equality becomes

$$\begin{aligned}
\frac{1}{n}\sum_{i\in[n]} A_i(\hat{w}_i - w_i^*) Y_i \hat{f}_{a,x}(X_i) &\leq \sqrt{\frac{1}{n}\sum_{i\in[n]}(\hat{w}_i - w_i^*)^2}\sqrt{\frac{1}{n}\sum_{i\in[n]} Y_i^2 \hat{f}_{a,x}^2(X_i)} \\
&= o_p(1)\left\{\sqrt{E\left[Y^2 f_{a,x}^2(X)\right]} + O_p(1/\sqrt{n})\right\}
\end{aligned}$$

by law of large numbers (Fukumizu et al. 2007), Theorem 2.4.3 and the finite second moment assumption of the product of Y and the inverse covariance operator. Likewise, the second term, $n^{-1}\sum_{i\in[n]} A_i \hat{w}_i g_{\hat{w}}(X_i)\hat{f}_{a,x}(X_i)$, is $o_p(1)$ from the proof of Theorem 2.4.3. Similarly,

$n^{-1}\sum_{i\in[n]}\hat{f}_{a,x}(X_i)Z_i$ is $O_p(1)$ and

$$\frac{1}{n}\sum_{i\in[n]}\hat{f}_{a,x}(X_i)\tilde{Z}_i = \frac{1}{n}\sum_{i\in[n]}\hat{f}_{a,x}(X_i)(\tilde{Z}_i - Z_i) + \frac{1}{n}\sum_{i\in[n]}\hat{f}_{a,x}(X_i)Z_i = o_p(1) + O_p(1).$$

From Theorem 3 and Theorem 4 of Fukumizu and Leng 2014, $\left|\hat{M}_{ij}(x) - M_{ij}(x)\right| = O_p\left(n^{-\min\{1/3,(2\xi+1)/(4\xi+4)\}}\right)$. By combining these results,

$$\left|\hat{W}_{ij}(x) - M_{ij}(x)\right| = o_p(1) + O_p\left(n^{-\min\{1/3,(2\xi+1)/(4\xi+4)\}}\right).$$

Since the dimensionality is fixed, this rate of convergence applies to $\|\hat{W}(x) - M(x)\|_F$.

Define $\tilde{M}_n$ as $\tilde{W}_n$ so that the (i,j)-th entry is $n^{-1}\sum_{k\in[n]}\hat{M}_{ij}(X_k)$. By triangular inequality,

$$\left\|\frac{1}{n}\sum_{i\in[n]}\hat{W}(X_i) - \mathbb{E}[M(X)]\right\|_F \leq \left\|\frac{1}{n}\sum_{i\in[n]}\hat{W}(X_i) - \tilde{M}_n\right\|_F + \left\|\tilde{M}_n - \mathbb{E}[M(X)]\right\|_F.$$

From the derivation above, the first term on the right hand side of the inequality is bounded as

$$\begin{aligned}\left\|\frac{1}{n}\sum_{k\in[n]}\hat{W}(X_k) - \tilde{M}_n\right\|_F^2 &= \sum_{i\in[p]}\sum_{j\in[p]}\left\{n^{-1}\sum_{k\in[n]}\left\{\hat{W}_{ij}(X_k) - \hat{M}_{ij}(X_k)\right\}\right\}^2 \\ &\leq \sum_{i\in[p]}\sum_{j\in[p]}\left\{n^{-1}\sum_{k\in[n]}\left|\hat{W}_{ij}(X_k) - \hat{M}_{ij}(X_k)\right|\right\}^2 \\ &= p^2 o_p(1).\end{aligned}$$

Therefore, $\left\|\frac{1}{n}\sum_{k\in[n]}\hat{W}(X_k) - \tilde{M}_n\right\|_F = o_p(1)$. By Theorem 3 of Fukumizu and Leng 2014, the second term on the right hand side is

$$\left\|\tilde{M}_n - \mathbb{E}[M(X)]\right\|_F = O_p\left(n^{-\min\left\{\frac{1}{3},\frac{2\xi+1}{4\xi+4}\right\}}\right).$$

We obtain the convergence in probability by combining these results.

Remark A.1. Theorem 2.4.5 can be extended to the setting where the number of covariates

p grows with the sample size, generalizing to our setting Theorem 4 of Fukumizu and Leng (2014), under the assumption that the true propensity scores are known. Specifically, under the additional assumptions (A11) and (A12), we can establish

$$\left\|\hat{M}(x) - M(x)\right\|_F = O_p\left(p \cdot C_p^2 \left(\frac{\alpha_p^2}{n}\right)^{\min\left\{\frac{1}{3}, \frac{2\xi_p+1}{4\xi_p+4}\right\}}\right)$$

for any $x \in \mathcal{X}$, and

$$\left\|\tilde{M}_n - \mathbb{E}[M(X)]\right\|_F = O_p\left(p \cdot \frac{C_p^2}{\sqrt{n}} + p \cdot C_p^2 \left(\frac{\alpha_p^2}{n}\right)^{\min\left\{\frac{1}{3}, \frac{2\xi_p+1}{4\xi_p+4}\right\}}\right).$$

as $n \to \infty$, where α_p, C_p, and ξ_p are quantities that depend on the dimensionality p. Furthermore, Theorem 4 of Yu et al. (2015) extends this to

$$\left\|\hat{B}\hat{B}^\top - BB^\top\right\|_F = O_p\left(p^2\frac{C_p^4}{n} + p^2 C_p^4 \left(\frac{\alpha_p^2}{n}\right)^{\min\left\{\frac{2}{3}, \frac{2\xi_p+1}{2\xi_p+2}\right\}}\right).$$

However, establishing analogous results for KCB weights presents significant technical challenges.

A.6 Proof of Lemma 2.4.7

Let $Y_{g_{\hat{w}}} = Y - g_{\hat{w}}(X)$, $Y_{g_{\hat{w}},i} = Y_i - g_{\hat{w}}(X_i)$. Recall that

$$Q(\tilde{f}^{\hat{V}}) = \frac{1}{n}\sum_{i\in[n]} \hat{w}_i |Y_i - g_{\hat{w}}(X_i)| \phi\left\{A_i \cdot \operatorname{sign}(Y_{g_{\hat{w}},i}) \tilde{f}^{\hat{V}}(\hat{B}^\top X_i)\right\} + \lambda_n \alpha^\top \hat{G} \alpha.$$

Since $Q(\tilde{f}^{\hat{V}})$ is minimized when $\alpha = \hat{\alpha} \in \mathbb{R}^n$ and $\alpha_0 = \hat{\alpha}_{0n}^{\hat{V}}$, $\tilde{f}_n^{\hat{V}}(\hat{B}^\top x) = h_n^{\hat{V}}(\hat{B}^\top x) + \alpha_{0n}^{\hat{V}}$, where $h_n^{\hat{V}}(\hat{B}^\top x) = \sum_{i\in[n]} \hat{\alpha}_i K_u(\hat{B}^\top X_i, \hat{B}^\top x)$. Let $\tilde{f}^{\hat{V}}(\hat{B}^\top x) = h^{\hat{V}}(\hat{B}^\top x) + \alpha_0$, where $h^{\hat{V}}(\hat{B}^\top x) = \sum_i \alpha_i K_u(\hat{B}^\top X_i, \hat{B}^\top x)$. Then,

$$\begin{aligned}
\mathbb{P}_n \hat{w}|Y - g_{\hat{w}}(X)| \phi\left\{A \cdot \operatorname{sign}(Y_{g_{\hat{w}}}) \tilde{f}_n^{\hat{V}}(\hat{B}^\top X)\right\} &\le \mathbb{P}_n \hat{w}|Y - g_{\hat{w}}(X)| \phi\left\{A \cdot \operatorname{sign}(Y_{g_{\hat{w}}}) \tilde{f}_n^{\hat{V}}(\hat{B}^\top X)\right\} + \lambda_n \|h_n^{\hat{V}}\|^2 \\
&\le \mathbb{P}_n \hat{w}|Y - g_{\hat{w}}(X)| \phi\left\{A \cdot \operatorname{sign}(Y_{g_{\hat{w}}}) \tilde{f}^{\hat{V}}(\hat{B}^\top X)\right\} + \lambda_n \|h^{\hat{V}}\|^2,
\end{aligned}$$

where $\|h_n^{\hat{V}}\|^2 = \hat{\alpha}^\top \hat{G}\hat{\alpha}$ and $\|h^{\hat{V}}\|^2 = \alpha^\top \hat{G}\alpha$. We have $|\tilde{f}^{\hat{V}}(\hat{B}^\top x)| \leq \|\tilde{f}^{\hat{V}}\| \cdot K_u(\hat{B}^\top x, \hat{B}^\top x) < \infty$ from Cauchy-Schwarz inequality for a bounded kernel K_u. By replacing $f(X)$ by $\tilde{f}^{\hat{V}}(\hat{B}^\top X)$ in Theorem 2.4.4 and from that $\lambda_n \to 0$,

$$\mathbb{P}_n \hat{w}|Y - g_{\hat{w}}(X)|\phi\left\{A \cdot \text{sign}(Y_{g_{\hat{w}}})\tilde{f}^{\hat{V}}(\hat{B}^\top X)\right\} + \lambda_n \|h^{\hat{V}}\|^2 - \mathcal{R}_{\phi,g}(\tilde{f}^{\hat{V}}) \to 0$$

in probability. Hence, with probability 1,

$$\inf_{\tilde{f}^{\hat{V}}} \mathcal{R}_{\phi,g}(\tilde{f}^{\hat{V}}) - \mathbb{P}_n \hat{w}|Y - g_{\hat{w}}(X)|\phi\left\{A \cdot \text{sign}(Y_{g_{\hat{w}}})\tilde{f}_n^{\hat{V}}(\hat{B}^\top X)\right\}$$

$$\leq \mathcal{R}_{\phi,g}(\tilde{f}_n^{\hat{V}}) - \mathbb{P}_n \hat{w}|Y - g_{\hat{w}}(X)|\phi\left\{A \cdot \text{sign}(Y_{g_{\hat{w}}})\tilde{f}_n^{\hat{V}}(\hat{B}^\top X)\right\}.$$

The left hand side is

$$\inf_{\tilde{f}^{\hat{V}}} \mathcal{R}_{\phi,g}(\tilde{f}^{\hat{V}}) - \mathbb{P}_n \hat{w}|Y - g_{\hat{w}}(X)|\phi\left\{A \cdot \text{sign}(Y_{g_{\hat{w}}})\tilde{f}_n^{\hat{V}}(\hat{B}^\top X)\right\} + \mathcal{R}_{\phi,g}(\tilde{f}_n^{\hat{V}}) - \mathcal{R}_{\phi,g}(\tilde{f}_n^{\hat{V}})$$

$$= \inf_{\tilde{f}^{\hat{V}}} \mathcal{R}_{\phi,g}(\tilde{f}^{\hat{V}}) - \mathcal{R}_{\phi,g}(\tilde{f}_n^{\hat{V}}) + \mathcal{R}_{\phi,g}(\tilde{f}_n^{\hat{V}}) - \mathbb{P}_n \hat{w}|Y - g_{\hat{w}}(X)|\phi\left\{A \cdot \text{sign}(Y_{g_{\hat{w}}})\tilde{f}_n^{\hat{V}}(\hat{B}^\top X)\right\}.$$

Therefore, $\mathcal{R}_{\phi,g}(\tilde{f}_n^{\hat{V}}) - \inf_{\tilde{f}^{\hat{V}}} \mathcal{R}_{\phi,g}(\tilde{f}^{\hat{V}}) \to 0$ as long as

$$\mathbb{P}_n \hat{w}|Y_{g_{\hat{w}}}|\phi\left(A \cdot \text{sign}(Y_{g_{\hat{w}}})\tilde{f}_n^{\hat{V}}(\hat{B}^\top X)\right) + \lambda_n \|h_n^{\hat{V}}\|^2 - \mathcal{R}_{\phi,g}(\tilde{f}_n^{\hat{V}}) = o_p(1).$$

Since $\mathcal{R}_{\phi,g_{\hat{w}}}(\tilde{f}_n^{\hat{V}}) - \mathcal{R}_{\phi,g}(\tilde{f}_n^{\hat{V}}) = o_p(1)$ by continuous mapping theorem, we show that

$$\mathbb{P}_n \hat{w}|Y_{g_{\hat{w}}}|\phi\left(A \cdot \text{sign}(Y_{g_{\hat{w}}})\tilde{f}_n^{\hat{V}}(B_0^\top X)\right) + \lambda_n \|h_n^{\hat{V}}\|^2 - \mathcal{R}_{\phi,g_{\hat{w}}}(\tilde{f}_n^{\hat{V}}) = o_p(1).$$

For any $h^{\hat{V}}(\hat{B}^\top x) = \sum_i \alpha_i K_u(\hat{B}^\top X_i, \hat{B}^\top x)$ and $\alpha_0 \in \mathbb{R}$,

$$\mathbb{P}_n \hat{w}|Y_{g_{\hat{w}}}|\phi\left\{A \cdot \text{sign}(Y_{g_{\hat{w}}})\tilde{f}_n^{\hat{V}}(\hat{B}^\top X)\right\} + \lambda_n \|h_n^{\hat{V}}\|^2 \leq \mathbb{P}_n \hat{w}|Y_{g_{\hat{w}}}|\phi\left(A \cdot \text{sign}(Y_{g_{\hat{w}}})\{h^{\hat{V}}(\hat{B}^\top X) + \alpha_0\}\right)$$
$$+ \lambda_n \|h^{\hat{V}}\|^2,$$

hence we can choose $h^{\hat{V}} = 0$ ($\alpha_i = 0 \ \ \forall i \in \mathbb{N}$) and $\alpha_0 = 0$ so that

$$\begin{aligned}
\mathbb{P}_n \hat{w}|Y_{g_{\hat{w}}}|\phi\left(A \cdot \text{sign}(Y_{g_{\hat{w}}})\tilde{f}_n^{\hat{V}}(\hat{B}^\top X)\right) + \lambda_n \|h_n^{\hat{V}}\|^2 &\leq \phi(0)\mathbb{P}_n \hat{w}|Y_{g_{\hat{w}}}| \\
&= \phi(0)\mathbb{P}_n \hat{w}\pi(A,X)\frac{|Y - g_{\hat{w}}(X)|}{\pi(A,X)} \\
&\leq \phi(0)\mathbb{P}_n \hat{w}\frac{|Y - g_{\hat{w}}(X)|}{\pi(A,X)}
\end{aligned}$$

Define $K_i = X^\top(n^{-1}\mathbb{X}^\top\mathbb{X})^{-1}X_i(\hat{w}_i - 1) \in \mathbb{R}$. Then $|K_i| \leq M_X M_w n^{1/3}$, where $M_X < \infty$ is the bound such that $|X^\top(n^{-1}\mathbb{X}^\top\mathbb{X})^{-1}X_i| \leq M_X$. When $n^{-1}\mathbb{X}^\top\mathbb{X}$ is singular, we can use ridge-regularization and replace it with $n^{-1}\mathbb{X}^\top\mathbb{X} + \lambda I_p$, where $\lambda > 0$. We have $1/\pi(A,X) \leq M_\pi < \infty$ as in the proof of Theorem 2.4.4 by **Assumption 2** and the compactness of $\mathcal{X}$. Furthermore, since $g(x)$ is well-defined and $\mathcal{X}$ is compact, we have $|g(X)| \leq M$ almost surely for $M < \infty$. Note that

$$\begin{aligned}
\frac{|Y - g_{\hat{w}}(X)|}{\pi(A,X)} &= \frac{|Y - g(X) + g(X) - g_{\hat{w}}(X)|}{\pi(A,X)} \\
&\leq \frac{|Y - g(X)|}{\pi(A,X)} + \frac{|g(X) - g_{\hat{w}}(X)|}{\pi(A,X)} \\
&\leq M_g + \frac{|n^{-1}\sum_{i\in[n]} g(X_i) - n^{-1}\sum_{i\in[n]} X^\top(n^{-1}\mathbb{X}^\top\mathbb{X})^{-1}X_i(\hat{w}_i - 1)Y_i|}{\pi(A,X)} \\
&= M_g + \frac{|n^{-1}\sum_{i\in[n]}\{g(X_i) - K_i Y_i\}|}{\pi(A,X)} \\
&= M_g + \frac{|n^{-1}\sum_{i\in[n]}\{K_i g(X_i) - K_i Y_i + (1 - K_i)g(X_i)\}|}{\pi(A,X)} \\
&\leq M_g + \frac{n^{-1}\sum_{i\in[n]}|K_i||g(X_i) - Y_i|}{\pi(A,X)} + \frac{n^{-1}\sum_{i\in[n]}|1 - K_i||g(X_i)|}{\pi(A,X)} \\
&\leq M_g + \frac{n^{-1}\sum_{i\in[n]}|K_i||g(X_i) - Y_i|/\pi(A_i,X_i)}{\pi(A,X)} + \frac{n^{-1}\sum_{i\in[n]}|1 - K_i||g(X_i)|}{\pi(A,X)} \\
&\leq M_g + M_g M_X M_\pi M_w n^{1/3} + M M_\pi + M_X M_\pi M M_w n^{1/3} \\
&\leq \{M_g + M_g M_X M_\pi M_w + (M_X M_w + 1)M_\pi M\}n^{1/3}.
\end{aligned}$$

Define M_h as $M_h = \sqrt{\phi(0)M_w\{M_g + M_g M_X M_\pi M_w + (M_X M_w + 1)M_\pi M\}}$. Then, $\lambda_n^*\|h_n^{\hat{V}}\|^2 \leq M_h^2$, where $\lambda_n^* = n^{-2/3}\lambda_n$. From this, we have the bound $\|\sqrt{\lambda_n^*}h_n^{\hat{V}}\| \leq M_h$. Note that we have $|\sqrt{\lambda_n^*}\alpha_{0n}^{\hat{V}}| \leq |\sqrt{\lambda_n}\alpha_{0n}^{\hat{V}}| < M_{\alpha_0}$, $\lambda_n^* \to 0$, and $n\lambda_n^* \to \infty$.

Define the RKHS with bounded kernel as $\mathcal{H}_H = \{\sqrt{\lambda_n^*}h : \|\sqrt{\lambda_n^*}h\| \leq M_h\}$. Since the RKHS norm is bounded, $|\sqrt{\lambda_n^*}h_n^{\hat{V}}| \leq \bar{H}$ for an envelope function $\bar{H}$ such that $\mathbb{P}\bar{H}^2 < \infty$. Also, from the proof of Lemma A.9 of Hable (2012), $\mathcal{H}_H$ satisfies the uniform entropy bound. Thus $\mathcal{H}_H$ is $\mathbb{P}$-Donsker. Likewise, $\{\sqrt{\lambda_n^*}(h+\alpha_0) : \|\sqrt{\lambda_n^*}h\| \leq M_h, |\sqrt{\lambda_n^*}\alpha_0| \leq M_{\alpha_0}\}$ is $\mathbb{P}$-Donsker. Let C be the Lipschitz constant of $\phi(\cdot)$. Define $\phi_{\lambda_n^*}(c) = \sqrt{\lambda_n^*}\phi(c/\sqrt{\lambda_n^*})$. Note that

$$\begin{aligned}\phi_{\lambda_n^*}(c_2) - \phi_{\lambda_n^*}(c_1) &= \sqrt{\lambda_n^*}\phi(c_2/\sqrt{\lambda_n^*}) - \sqrt{\lambda_n^*}\phi(c_1/\sqrt{\lambda_n^*}) \\ &= \sqrt{\lambda_n^*}C \cdot (c_2/\sqrt{\lambda_n^*} - c_1/\sqrt{\lambda_n^*}) = C(c_2 - c_1).\end{aligned}$$

Hence, $\phi_{\lambda_n^*}(\cdot)$ is Lipschitz continuous. Also,

$$\sqrt{\lambda_n^*}\frac{|Y_{g_{\hat{w}}}|}{\pi(A,X)}\phi\left(A \cdot \text{sign}(Y_{g_{\hat{w}}})\{h(\hat{B}^\top X) + \alpha_0\}\right) = \frac{|Y_{g_{\hat{w}}}|}{\pi(A,X)}\phi_{\lambda_n^*}\left(A\sqrt{\lambda_n^*} \cdot \text{sign}(Y_{g_{\hat{w}}})\{h(\hat{B}^\top X) + \alpha_0\}\right).$$

Since $|Y - g_{\hat{w}}(X)|/\pi(A,X) \leq M_g$ and $\phi_{\lambda_n^*}(\cdot)$ is Lipschitz, it follows that

$$\left\{\sqrt{\lambda_n^*}\frac{|Y_{g_{\hat{w}}}|}{\pi(A,X)}\phi\left(A \cdot \text{sign}(Y_{g_{\hat{w}}})\{h(\hat{B}^\top X) + \alpha_0\}\right) : \|\sqrt{\lambda_n^*}h\| \leq M_h, |\sqrt{\lambda_n^*}\alpha_0| \leq M_{\alpha_0}\right\}$$

is $\mathbb{P}$-Donsker by Corollary 9.32-(iv) of Kosorok (2008). This implies that

$$\mathbb{P}_n\frac{|Y_{g_{\hat{w}}}|}{\pi(A,X)}\phi\left(A \cdot \text{sign}(Y_{g_{\hat{w}}})\tilde{f}_n^{\hat{V}}(\hat{B}^\top X)\right) - \mathcal{R}_{\phi,g_{\hat{w}}}(\tilde{f}_n^{\hat{V}}) = O_p(1/\sqrt{n\lambda_n^*}).$$

Since $n\lambda_n^* \to \infty$ and by Theorem 2.4.4 we have

$$\begin{aligned}&\mathbb{P}_n\hat{w}|Y_{g_{\hat{w}}}|\phi\left(A \cdot \text{sign}(Y_{g_{\hat{w}}})\tilde{f}_n^{\hat{V}}(\hat{B}^\top X)\right) - \mathcal{R}_{\phi,g_{\hat{w}}}(h_n^{\hat{V}} + \alpha_{0n}^{\hat{V}}) \\ &= \mathbb{P}_n\hat{w}|Y_{g_{\hat{w}}}|\phi\left(A \cdot \text{sign}(Y_{g_{\hat{w}}})\tilde{f}_n^{\hat{V}}(\hat{B}^\top X)\right) - \mathbb{P}_n\frac{|Y_{g_{\hat{w}}}|}{\pi(A,X)}\phi\left(A \cdot \text{sign}(Y_{g_{\hat{w}}})\tilde{f}_n^{\hat{V}}(\hat{B}^\top X)\right) \\ &\quad + \mathbb{P}_n\frac{|Y_{g_{\hat{w}}}|}{\pi(A,X)}\phi\left(A \cdot \text{sign}(Y_{g_{\hat{w}}})\tilde{f}_n^{\hat{V}}(\hat{B}^\top X)\right) - \mathcal{R}_{\phi,g_{\hat{w}}}(h_n^{\hat{V}} + \alpha_{0n}^{\hat{V}}) \\ &= o_p(1) + O_p\left(1/\sqrt{n\lambda_n^*}\right) = o_p(1).\end{aligned}$$

This implies that $\mathcal{R}_{\phi,g}(\tilde{f}_n^{\hat{V}}) - \inf_{\hat{f}^{\hat{V}}}\mathcal{R}_{\phi,g}(\tilde{f}^{\hat{V}}) \to 0$ in probability.

A.7 Proof of Proposition 1

We establish two lemmas on the excess risk bound and the expressive power of RKHS from a universal kernel prior to presenting the proof.

A.7.1 Expressive power of an RKHS from a universal kernel

Let $\tilde{f}_{\phi,g}^{\hat{V}*}$ be a minimizer of $\mathcal{R}_{\phi,g}(\tilde{f})$ over all continuous functions $\tilde{f} : \hat{\mathcal{V}} \mapsto \mathbb{R}$ and $\mathcal{R}_{\phi,g}^{\hat{V}*} = \mathcal{R}_{\phi,g}(\tilde{f}_{\phi,g}^{\hat{V}*})$. We present a slight modification of Lemma 2.5 of Zhou and Kosorok (2017) to establish that the minimum of $\mathcal{R}_{\phi,g}$ over RKHS functions $\tilde{f}^{\hat{V}} : \hat{\mathcal{V}} \mapsto \mathbb{R}$ from a universal kernel is equivalent to $\mathcal{R}_{\phi,g}^{\hat{V}*}$. Note that the difference from Lemma 2.5 of Zhou and Kosorok (2017) is that the decision function is defined on $\hat{\mathcal{V}}$ instead of $\mathcal{X}$. However, since $\hat{\mathcal{V}}$ consists of linear transformed variates of $\mathcal{X}$, we can apply Lusin's theorem by using the regular measure μ on $\mathcal{X}$. Then, we can approximate $\tilde{f}_{\phi,g}^{\hat{V}*}$ from a continuous function $\tilde{f}_1(\hat{B}^\top x)$ such that for all $\epsilon > 0$,

$$\mu\left\{x \in \mathcal{X}; \tilde{f}_1(\hat{B}^\top x) \neq \tilde{f}_{\phi,g}^{\hat{V}*}(\hat{B}^\top x)\right\} \leq \epsilon.$$

The proof easily follows from the proof of Lemma 2.5 of Zhou and Kosorok (2017) by replacing $f(X)$ to $\tilde{f}(\hat{B}^\top X)$.

Lemma A.3. Suppose that the covariate space $\mathcal{X} \in \mathbb{R}^p$ is a compact metric space and that the RKHS of functions $\tilde{f}^{\hat{V}} : \hat{\mathcal{V}} \mapsto \mathbb{R}$ uses the universal kernel. Suppose that ϕ is Lipschitz continuous, and $\tilde{f}_{\phi,g}^{\hat{V}*} : \hat{\mathcal{V}} \mapsto \mathbb{R}$ is measurable and bounded for any $\hat{B}$: $|\tilde{f}_{\phi,g}^{\hat{V}*}| \leq M_{\tilde{f}} < \infty$. Given any distribution P of (X, A, Y) with the bound $|Y_g|/\pi(A, X) \leq M_g < \infty$ almost everywhere with regular marginal distribution on X, we have

$$\inf_{\tilde{f}^{\hat{V}}} \mathcal{R}_{\phi,g}(\tilde{f}^{\hat{V}}) = \mathcal{R}_{\phi,g}^{\hat{V}*}.$$

A.7.2 Excess risk bound

We present a slight modification of Theorem 2.2 of Zhou and Kosorok (2017) to establish the bound of excess risk, $\mathcal{R}(\tilde{f}_n^{\hat{V}}) - \mathcal{R}^*$, by the excess (ϕ, g)-risk, $\mathcal{R}_{\phi,g}(\tilde{f}_n^{\hat{V}}) - \mathcal{R}_{\phi,g}^*$.

Prior to introducing the bound, we define $\eta_1(x)$ and $\eta_2(x)$ as follows:

$$\eta_1(x) = \mathbb{E}[\max\{Y - g(X), 0\}|X = x, A = +1] - \mathbb{E}[\max\{-(Y - g(X)), 0\}|X = x, A = -1];$$

$$\eta_2(x) = \mathbb{E}[\max\{Y - g(X), 0\}|X = x, A = -1] - \mathbb{E}[\max\{-(Y - g(X)), 0\}|X = x, A = +1].$$

By algebra, we have

$$\eta_1(x) - \eta_2(X) = \mathbb{E}[Y|X = x, A = +1] - \mathbb{E}[Y|X = x, A = -1],$$

and that

$$\mathcal{R}_{\phi,g}(f) = \mathbb{E}[\eta_1(X)\phi\{f(X)\} + \eta_2(X)\phi\{-f(X)\}],$$

where $\tilde{f}^V(B_0^\top X)$ and $\tilde{f}^{\hat{V}}(\hat{B}^\top X)$ can be used in place of f and $f(X)$. Motivated by this expression, we can define a generic conditional (ϕ, g)-risk function as

$$q_{\eta_1,\eta_2}(\alpha) = \eta_1\phi(\alpha) + \eta_2\phi(-\alpha)$$

for $\eta_1 \geq 0$, $\eta_2 \geq 0$, and $\alpha \in \mathbb{R}$. As long as $\phi(\cdot)$ is convex, $q_{\eta_1,\eta_2}(\cdot)$ is convex. Note that

$$\mathbb{E}[q_{\eta_1(X),\eta_2(X)}(f)] = \mathcal{R}_{\phi,g}(f).$$

Lemma A.4. Assume $\phi(\cdot)$ is convex, $\phi'(0)$ exists and $\phi'(0) < 0$. Suppose that for constants $C > 0$ and $s \leq 1$ such that

$$|\eta_1 - \eta_2|^s \leq C^s \left\{ q_{\eta_1,\eta_2}(0) - \min_{\alpha\in\mathbb{R}} q_{\eta_1,\eta_2}(\alpha) \right\}.$$

Then, for a decision rule $\tilde{f}^{\hat{V}} \in \mathcal{H}_{\hat{\mathcal{V}}}$,

$$\mathcal{R}(\tilde{f}^{\hat{V}}) - \mathcal{R}^* \leq C\{\mathcal{R}_{\phi,g}(\tilde{f}^{\hat{V}}) - \mathcal{R}^*_{\phi,g}\}^{1/s}.$$

Proof. The proof mostly follows from the proof of Theorem 2.2 of Zhou and Kosorok (2017) by replacing $f(X)$ with $\tilde{f}^{\hat{V}}(\hat{B}^\top X)$. For the optimal Bayes decision function $\tilde{f}^*$ such that

$d^*(B_0^\top x) = \text{sign} \circ \tilde{f}^*(x)$ and $\mathcal{R}^* = \mathcal{R}(\tilde{f}^*)$,

$$\mathbb{E}\left[\frac{Y}{\pi(A,X)}\mathbb{1}\{A \neq \text{sign} \circ \tilde{f}^{\hat{V}}(\hat{B}^\top X)\}|X = x\right] - \mathbb{E}\left[\frac{Y}{\pi(A,X)}\mathbb{1}\{A \neq \text{sign} \circ \tilde{f}^*(B_0^\top X)\} \mid X = x\right]$$

$$= \{\mathbb{E}[Y|X = x, A = +1] - \mathbb{E}[Y|X = x, A = -1]\}\left[\mathbb{1}\{\text{sign} \circ \tilde{f}^*(B_0^\top x)\} - \mathbb{1}\{\text{sign} \circ \tilde{f}^{\hat{V}}(\hat{B}^\top x)\}\right]$$

$$\leq |\eta_1(x) - \eta_2(x)| \cdot \mathbb{1}\left\{\text{sign} \circ \tilde{f}^{\hat{V}}(\hat{B}^\top x)\big(\eta_1(x) - \eta_2(x)\big) < 0\right\}.$$

Taking expectation on the left hand side of the equality leads to $\mathcal{R}(\tilde{f}^{\hat{V}}) - \mathcal{R}^*$. Let $\Delta q_{\eta_1(X),\eta_2(X)}(0) = q_{\eta_1,\eta_2}(0) - \min_{\alpha\in\mathbb{R}} q_{\eta_1,\eta_2}(\alpha)$. By taking the expectation on the right hand side of the inequality, we have

$$\begin{aligned}
\mathcal{R}(\tilde{f}^{\hat{V}}) - \mathcal{R}^* &\leq \mathbb{E}\left[|\eta_1(X) - \eta_2(X)| \cdot \mathbb{1}\left\{\text{sign} \circ \tilde{f}^{\hat{V}}(\hat{B}^\top X)\big(\eta_1(X) - \eta_2(X)\big) < 0\right\}\right] \\
&\leq \left(\mathbb{E}\left[|\eta_1(X) - \eta_2(X)|^s \cdot \mathbb{1}\left\{\text{sign} \circ \tilde{f}^{\hat{V}}(\hat{B}^\top X)\big(\eta_1(X) - \eta_2(X)\big) < 0\right\}\right]\right)^{1/s} \\
&\leq C \cdot \left(\mathbb{E}\left[\Delta q_{\eta_1(X),\eta_2(X)}(0) \cdot \mathbb{1}\left\{\text{sign} \circ \tilde{f}^{\hat{V}}(\hat{B}^\top X)\big(\eta_1(X) - \eta_2(X)\big) < 0\right\}\right]\right)^{1/s} \\
&\leq C \cdot \left(\mathbb{E}\left[\Delta q_{\eta_1(X),\eta_2(X)}(\tilde{f}^{\hat{V}}) \cdot \mathbb{1}\left\{\text{sign} \circ \tilde{f}^{\hat{V}}(\hat{B}^\top X)\big(\eta_1(X) - \eta_2(X)\big) < 0\right\}\right]\right)^{1/s} \\
&\leq C \cdot \left(\mathbb{E}\left[\Delta q_{\eta_1(X),\eta_2(X)}(\tilde{f}^{\hat{V}})\right]\right)^{1/s} = C\left(\mathcal{R}_{\phi,g}(\tilde{f}^{\hat{V}}) - \mathcal{R}^*_{\phi,g}\right)^{1/s},
\end{aligned}$$

where the second inequality follows from Jensen's inequality, and the third inequality follows from the given condition. From the given condition on $\phi(\cdot)$, $\phi(\cdot)$ is Fisher consistent. Since $\text{sign}(\tilde{f}^{\hat{V}}) \cdot (\eta_1 - \eta_2) < 0$ implies that 0 lies in between $\tilde{f}^{\hat{V}}$ and $\tilde{f}^*$, $q_{\eta_1,\eta_2}(0) \leq q_{\eta_1,\eta_2}(\tilde{f}^{\hat{V}})$. Thus, the fourth inequality follows from that $\Delta q_{\eta_1(X),\eta_2(X)}(0) \leq \Delta q_{\eta_1(X),\eta_2(X)}(\tilde{f}^{\hat{V}})$. □

In case of hinge loss, we have that $q_{\eta_1,\eta_2}(0) - \min_{\alpha\in\mathbb{R}} q_{\eta_1,\eta_2}(\alpha) = |\eta_1 - \eta_2|$. Thus, when $C = 1$ and $s = 1$ in Lemma A.4,

$$\mathcal{R}(\tilde{f}_n^{\hat{V}}) - \mathcal{R}^* \leq \mathcal{R}_{\phi,g}(\tilde{f}_n^{\hat{V}}) - \mathcal{R}^*_{\phi,g}.$$

A.7.3 Proof of proposition

Let $\lambda_n^* = n^{-1/3}\lambda_n$. It follows from the proof of Proposition 2.7 of Zhou and Kosorok (2017) that $\sqrt{\lambda_n^*}\alpha_{0n}^{\hat{V}}$ is bounded. From the proof of Lemma 2.4.7, we have that $\|\sqrt{\lambda_n^*}h_n^{\hat{V}}\| < \infty$.

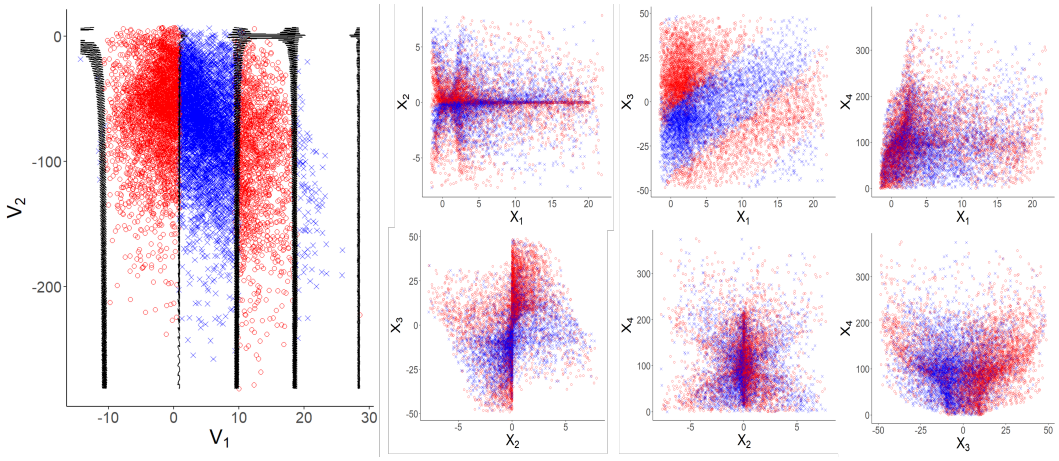

Figure A.1: The distribution of the optimal treatment regimes according to different levels of covariates in Setting 2. The red circles and blue crosses are data points from the test dataset and represent $d^*(X) = -1$ (or $d^*(V_0) = -1$) and $d^*(X) = +1$ (or $d^*(V_0) = +1$), respectively. The solid lines on the left plot display the decision boundaries in the dimension-reduced space $\left\{v_0 = \left(v_0^{(1)}, v_0^{(2)}\right); \tilde{f}^*(v_0) = 0\right\}$.

Note that $|\tilde{f}^{\hat{V}*}_{\phi,g}| \leq 1$ for hinge loss from Appendix A of Zhou and Kosorok (2017). Then by Lemma A.3, $\inf_{\tilde{f}^{\hat{V}}} \mathcal{R}_{\phi,g}(\tilde{f}^{\hat{V}}) = \mathcal{R}^{\hat{V}*}_{\phi,g}$. By combining these results and the triangular inequality,

$$\begin{aligned}\mathcal{R}_{\phi,g}(\tilde{f}_n^{\hat{V}}) - \mathcal{R}^*_{\phi,g} &\leq |\mathcal{R}_{\phi,g}(\tilde{f}_n^{\hat{V}}) - \mathcal{R}^{\hat{V}*}_{\phi,g}| + |\mathcal{R}^{\hat{V}*}_{\phi,g} - \mathcal{R}^*_{\phi,g}| \\ &\leq |\mathcal{R}_{\phi,g}(\tilde{f}_n^{\hat{V}}) - \inf_{\tilde{f}^{\hat{V}}} \mathcal{R}_{\phi,g}(\tilde{f}^{\hat{V}})| + |\inf_{\tilde{f}^{\hat{V}}} \mathcal{R}_{\phi,g}(\tilde{f}^{\hat{V}}) - \mathcal{R}^{\hat{V}*}_{\phi,g}| + |\mathcal{R}^{\hat{V}*}_{\phi,g} - \mathcal{R}^*_{\phi,g}|.\end{aligned}$$

From Lemma 2.4.7, $|\mathcal{R}_{\phi,g}(\tilde{f}_n^{\hat{V}}) - \inf_{\tilde{f}^{\hat{V}}} \mathcal{R}_{\phi,g}(\tilde{f}^{\hat{V}})| = o_p(1)$. Lemma A.3 implies $|\inf_{\tilde{f}^{\hat{V}}} \mathcal{R}_{\phi,g}(\tilde{f}^{\hat{V}}) - \mathcal{R}^{\hat{V}*}_{\phi,g}| = 0$. For a uniformly bounded continuous function $\tilde{f} : \mathbb{R}^u \mapsto \mathbb{R}$ and by Theorem Corollary 2, we have $\tilde{f}(\hat{O}^\top \hat{B}^\top x) \to \tilde{f}(B_0^\top x)$ for an orthogonal matrix $\hat{O}$ by continuous mapping theorem and Corollary 2.4.6. Thus, $\inf_{\tilde{f}|_{\hat{\mathcal{V}}}} \mathcal{R}_{\phi,g}(\tilde{f}) \to \inf_{\tilde{f}|_{\mathcal{V}}} \mathcal{R}_{\phi,g}(\tilde{f})$, from which, $|\mathcal{R}^{\hat{V}*}_{\phi,g} - \mathcal{R}^*_{\phi,g}| = o_p(1)$. When $\phi(\cdot)$ is the hinge loss, Lemma A.4 implies $\mathcal{R}(\tilde{f}_n^{\hat{V}}) - \mathcal{R}^* \leq \mathcal{R}_{\phi,g}(\tilde{f}_n^{\hat{V}}) - \mathcal{R}^*_{\phi,g}$. Therefore, we have $\mathcal{R}(\tilde{f}_n^{\hat{V}}) - \mathcal{R}^* = o_p(1)$.

A.8 Equivalence of the risk and its alternative form

In Section 2.3.3, we claimed that the minimization of $\mathcal{R}(f) = \mathbb{E}\left[Y/\pi(A,X)\mathbb{1}\{\text{sign} \circ f(X) \neq A\}\right]$ is equivalent to the minimization of $\mathbb{E}[|Y - g(X)|/\pi(A,X)\mathbb{1}\{A \cdot \text{sign}(Y - g(X)) \neq \text{sign} \circ f(X)\}]$.

Let $\tilde{m}(a,x) = \mathbb{E}[x^\top (E[XX^\top])^{-1} X\pi(-a,X)/\pi(-a,x)\mathbb{E}[Y \mid X, A=a]]$ for $a \in \{-1,+1\}$. Then $g(x)$ can be re-expressed as

$$g(x) = \pi(-1,x)\tilde{m}(+1,x) + \pi(+1,x)\tilde{m}(-1,x).$$

Note that $d(x) = \text{sign} \circ f(x)$. Define

$$\tilde{h}(x,d) = \tilde{m}(+1,x)\mathbb{1}\{d(x)=+1\} + \tilde{m}(-1,x)\mathbb{1}\{d(x)=-1\}.$$

First, we show that

$$\mathbb{E}\left[\frac{Y-\tilde{h}(X,d)}{\pi(A,X)}\mathbb{1}\{d(X)=A\} + \tilde{h}(X,d)\right] = \mathbb{E}\left[\frac{Y}{\pi(A,X)}\mathbb{1}\{d(X)=A\}\right],$$

where the left hand side is the doubly robust form of the value function introduced in Zhang et al. (2012). Note that

$$\begin{aligned}
\mathbb{E}\left[\tilde{h}(X,d) - \frac{\tilde{h}(X,d)\mathbb{1}\{d(X)=A\}}{\pi(A,X)}\right] &= \mathbb{E}\left[\tilde{h}(X,d) - \mathbb{E}\left[\frac{\tilde{h}(X,d)\mathbb{1}\{d(X)=A\}}{\pi(A,X)} \mid X\right]\right] \\
&= \mathbb{E}[\tilde{h}(X,d)] \\
&\quad - \mathbb{E}\left[\mathbb{E}\left[\frac{\tilde{h}(X,d)\mathbb{1}\{d(X)=+1\}}{\pi(+1,X)} \mid X, A=+1\right]\pi(+1,X)\right] \\
&\quad - \mathbb{E}\left[\mathbb{E}\left[\frac{\tilde{h}(X,d)\mathbb{1}\{d(X)=-1\}}{\pi(-1,X)} \mid X, A=-1\right]\pi(-1,X)\right] \\
&= \mathbb{E}[\tilde{h}(X,d)] - \mathbb{E}[\tilde{h}(X,d)\mathbb{1}\{d(X)+1\}] - \mathbb{E}[\tilde{h}(X,d)\mathbb{1}\{d(X)-1\}] \\
&= \mathbb{E}[\tilde{h}(X,d)] - \mathbb{E}[\tilde{h}(X,d)] = 0.
\end{aligned}$$

Therefore,

$$\begin{aligned}\mathbb{E}\left[\frac{Y-\tilde{h}(X,d)}{\pi(A,X)}\mathbb{1}\{d(X)=A\}+\tilde{h}(X,d)\right] &= \mathbb{E}\left[\frac{Y}{\pi(A,X)}\mathbb{1}\{d(X)=A\}\right] \\ &\quad + \mathbb{E}\left[\tilde{h}(X,d)-\frac{\tilde{h}(X,d)\mathbb{1}\{d(X)=A\}}{\pi(A,X)}\right] \\ &= \mathbb{E}\left[\frac{Y}{\pi(A,X)}\mathbb{1}\{d(X)=A\}\right].\end{aligned}$$

The integrand on the left hand side can be rearranged as follows

$$\begin{aligned}\frac{Y-\tilde{h}(X,d)}{\pi(A,X)}\mathbb{1}\{d(X)=A\}+\tilde{h}(X,d) &= \frac{Y-g(X)+g(X)-\tilde{h}(X,d)}{\pi(A,X)}\mathbb{1}\{d(X)=A\}+\tilde{h}(X,d) \\ &= \frac{Y-g(X)}{\pi(A,X)}\mathbb{1}\{d(X)=A\} \\ &\quad + \frac{g(X)-\tilde{h}(X,d)}{\pi(A,X)}\mathbb{1}\{d(X)=A\}+\tilde{h}(X,d).\end{aligned}$$

The last two terms become

$$\begin{aligned}\frac{g(X)-\tilde{h}(X,d)}{\pi(A,X)}\mathbb{1}\{d(X)=A\}+\tilde{h}(X,d) &= \frac{[\pi(-1,X)-\mathbb{1}\{d(X)=+1\}]\tilde{m}(+1,X)}{\pi(A,X)}\mathbb{1}\{d(X)=A\} \\ &\quad + \frac{[\pi(+1,X)-\mathbb{1}\{d(X)=-1\}]\tilde{m}(-1,X)}{\pi(A,X)}\mathbb{1}\{d(X)=A\} \\ &\quad + \tilde{m}(+1,X)\mathbb{1}\{d(X)=+1\}+\tilde{m}(-1,X)\mathbb{1}\{d(X)=-1\}.\end{aligned}$$

When $A = +1$,

$$\begin{aligned}
\frac{g(X) - \tilde{h}(X,d)}{\pi(+1,X)}\mathbb{1}\{d(X) = +1\} + \tilde{h}(X,d) &= \frac{[\pi(-1,X) - \mathbb{1}\{d(X) = +1\}]\tilde{m}(+1,X)}{\pi(+1,X)}\mathbb{1}\{d(X) = +1\} \\
&\quad + \frac{[\pi(+1,X) - \mathbb{1}\{d(X) = -1\}]\tilde{m}(-1,X)}{\pi(+1,X)}\mathbb{1}\{d(X) = +1\} \\
&\quad + \tilde{m}(+1,X)\mathbb{1}\{d(X) = +1\} + \tilde{m}(-1,X)\mathbb{1}\{d(X) = -1\} \\
&= \frac{\tilde{m}(+1,X)\mathbb{1}\{d(X)+1\}}{\pi(+1,X)} - \tilde{m}(+1,X)\mathbb{1}\{d(X)+1\} \\
&\quad - \frac{\tilde{m}(+1,X)\mathbb{1}\{d(X)+1\}}{\pi(+1,X)} + \tilde{m}(-1,X)\mathbb{1}\{d(X) = +1\} \\
&\quad - \frac{\tilde{m}(-1,X)\mathbb{1}\{d(X)-1\}\mathbb{1}\{d(X)+1\}}{\pi(+1,X)} \\
&\quad + \tilde{m}(+1,X)\mathbb{1}\{d(X) = +1\} + \tilde{m}(-1,X)\mathbb{1}\{d(X) = -1\} \\
&= \tilde{m}(-1,X)\mathbb{1}\{d(X) = +1\} + \tilde{m}(-1,X)\mathbb{1}\{d(X) = -1\} \\
&= \tilde{m}(-1,X).
\end{aligned}$$

Likewise, when $A = -1$,

$$\frac{g(X) - \tilde{h}(X,d)}{\pi(-1,X)}\mathbb{1}\{d(X) = -1\} + \tilde{h}(X,d) = \tilde{m}(+1,X).$$

From these two results,

$$\frac{g(X) - \tilde{h}(X,d)}{\pi(A,X)}\mathbb{1}\{d(X) = A\} + \tilde{h}(X,d) = \tilde{m}(-A,X).$$

Therefore,

$$\frac{Y - \tilde{h}(X,d)}{\pi(A,X)}\mathbb{1}\{d(X) = A\} + \tilde{h}(X,d) = \frac{Y - g(X)}{\pi(A,X)}\mathbb{1}\{d(X) = A\} + \tilde{m}(-A,X).$$

From this, maximizing the value function $\mathbb{E}[Y\{d(X)\}]$ is equivalent to the maximization of

$$\mathbb{E}\left[\frac{Y - g(X)}{\pi(A,X)}\mathbb{1}\{d(X) = A\} + \tilde{m}(-A,X)\right].$$

By using the finding in Section 2.2 of Liu et al. (2018),

$$\mathbb{E}\left[\frac{Y-g(X)}{\pi(A,X)}\mathbb{1}\{d(X)=A\}+\tilde{m}(-A,X)\right]=\mathbb{E}\left[\frac{|Y-g(X)|}{\pi(A,X)}\mathbb{1}\{A\cdot\text{sign}(Y-g(X))=d(X)\}\right] + \mathbb{E}[\tilde{m}(-A,X)]+\mathbb{E}\left[\frac{\max\{-Y+g(X),0\}}{\pi(A,X)}\right].$$

Note that

$$\begin{aligned}
\mathbb{E}[\tilde{m}(-A,X)] &= \int \mathbb{E}\left[x^\top(E[XX^\top])^{-1}X\frac{\pi(a,X)}{\pi(a,x)}\mathbb{E}[Y|X,A=-a]\right]dP_{A,X}(a,x)\\
&= \int\int \mathbb{E}\left[x^\top(E[XX^\top])^{-1}X\frac{\pi(a,X)}{\pi(a,x)}\mathbb{E}[Y|X,A=-a]\right]dP_{A|X}(a|x)dP_X(x)\\
&= \int \mathbb{E}\left[x^\top(E[XX^\top])^{-1}X\frac{\pi(+1,X)}{\pi(+1,x)}\mathbb{E}[Y|X,A=-1]\right]\pi(+1,x)dP_X(x)\\
&\quad+\int \mathbb{E}\left[x^\top(E[XX^\top])^{-1}X\frac{\pi(-1,X)}{\pi(-1,x)}\mathbb{E}[Y|X,A=+1]\right]\pi(-1,x)dP_X(x)\\
&= \int \mathbb{E}[x^\top(E[XX^\top])^{-1}X\{\pi(+1,X)\mathbb{E}[Y|X,A=-1]+\pi(-1,X)\mathbb{E}[Y|X,A=+1]\}]dP_X(x)\\
&= \int \mathbb{E}\left[x^\top(E[XX^\top])^{-1}X\mathbb{E}\left[\frac{\pi(-A,X)}{\pi(A,X)}Y\mid X\right]\right]dP_X(x)\\
&= \int \mathbb{E}\left[x^\top(E[XX^\top])^{-1}X\mathbb{E}\left(\frac{\pi(-A,X)}{\pi(A,X)}Y\right)\right]dP_X(x)\\
&= \mathbb{E}\left[X^\top\right]\left\{\mathbb{E}[XX^\top]\right\}^{-1}\mathbb{E}\left[\frac{\pi(-A,X)}{\pi(A,X)}XY\right].
\end{aligned}$$

Note that only the first term contains the decision rule. Therefore, the decision rule that maximizes the value function is equivalent to the decision rule that maximizes the first term. Equivalently, the minimizing $d(x)$ of $\mathcal{R}(f)$ is the minimizer of

$$\mathbb{E}\left[\frac{|Y-g(X)|}{\pi(A,X)}\mathbb{1}\{A\cdot\text{sign}(Y-g(X))\neq d(X)\}\right]$$

A.9 Simulation details

A.9.1 Performance evaluation

The accuracy of the fitted decision rules was evaluated using the test data as

$$n_{test}^{-1} \sum_{i=1}^{n_{test}} \mathbb{1}\left\{\hat{d}\left(X_i^{test}\right) = d^*\left(X_i^{test}\right)\right\},$$

while the value function was evaluated using the unbiased estimator of (2.1) from Qian and Murphy (2011):

$$\frac{n_{test}^{-1} \sum_{i=1}^{n_{test}} \mathbb{1}\left\{A_i^{test} = \hat{d}\left(X_i^{test}\right)\right\} Y_i^{test} / \pi\left(A_i^{test}, X_i^{test}\right)}{n_{test}^{-1} \sum_{i=1}^{n_{test}} \mathbb{1}\left\{A_i^{test} = \hat{d}\left(X_i^{test}\right)\right\} / \pi\left(A_i^{test}, X_i^{test}\right)}, \tag{A.6}$$

where the superscript *test* indicates that the observation comes from the test set. Note that $\hat{d}\left(\hat{V}_i\right)$ is evaluated in place of $\hat{d}(X_i)$ in for the proposed DOL method.

A.9.2 Estimation of ITR

The KCB weights were obtained by using the R package `ATE.ncb` (`github.com/raymondkww/ATE.ncb`). We used the R package provided in `https://github.com/muxuanliang/iMAVE` to implement interaction MAVE. IPW estimated by using XGBoost was implemented via R package `twang` (Matthew Cefalu et al. 2024). We implemented neural network estimation of propensity scores (Shang et al. 2025) by using R codes provided in `github.com/ys3298/RobustPSviaLossCalibration` and performed minimization by using Adam algorithm (Kingma and Ba 2014). Default options were used. In gKDR, the dimension u of the central mean subspace was selected using the same two-fold cross-validation procedure as for λ_n. We used $\epsilon_n = 10^{-7}$ for Setting 1 and $\epsilon_n = 10^{-5}$ for Setting 2 and 3. RKHS with Gaussian kernels were used to estimate ITRs and to implement gKDR. Median heuristic bandwidths were selected as recommended in Gretton et al. (2012). The R function `ipop` from the `kernlab` package (Karatzoglou et al. 2004) was used to implement the weighted SVM formulation implementing AOL. The SI model was fitted using `simml` and `pred.simml` from the `simml` package (Park, Hyung et al. 2021). The function `ql` from the R package `DTRlearn2` (Chen, Yuan et al. 2020) was used to fit the ℓ_1-PLS models. R-learner was imple-

mented by using `rkern` from R package `rlearner` (`https://github.com/xnie/rlearner`). The median heuristic bandwidths were used for the kernel ridge regression model and the regularization parameters were tuned by the five-fold cross-validation. The decision rule was obtained by the sign of the estimated heterogeneous treatment effect.

A.9.3 Simulation settings

A.9.4 Setting 1

The observed covariates $X \in \mathbb{R}^{50}$ are defined as $X^{(1)} = \exp(Z^{(1)}/2)$; $X^{(2)} = Z^{(2)}/\{1 + \exp(Z^{(1)})\}$;$X^{(3)} = (Z^{(1)} \cdot Z^{(3)}/25 + 0.6)^3$; $X^{(4)} = (Z^{(2)} + Z^{(4)} + 20)^2$; $X^{(i)} = Z^{(i)}$ for $i \in \{5, \cdots, 50\}$. The basis matrix B_0 consists of two columns $(0.6/a, 0.5/a, -0.2/a, \mathbf{0}_{47}^\top)^\top$ and $(0, 0.2/b, 0.5/b, 0, -0.5/b, \mathbf{0}_{46}^\top)^\top$, where $a = \sqrt{0.6^2 + 0.5^2 + (-0.2)^2}$ and $b = \sqrt{0.2^2 + 0.5^2 + (-0.5)^2}$. Therefore, the dimension-reduced covariates are $V = B_0^\top X \in \mathbb{R}^2$. The outcome values were generated from the following main effect and treatment-covariate interaction term:

$$\mu(X) = 5 + 6Z^{(1)} + 8Z^{(2)} + 3Z^{(3)} + 5Z^{(4)} + 5Z^{(5)},$$

$$\tilde{f}^V(V) = 5 \cdot \sin\left(\frac{\pi}{V^{(1)} + 1} \cdot \frac{1}{\sqrt{-V^{(2)}}}\right) + 2.5 \cdot \sin\left(\pi V^{(1)}\right) \cdot \log\left(-V^{(2)}\right),$$

from which $Y \sim N\{\mu(X) + A \cdot \tilde{f}^V(V), 1\}$. The propensity score for the non-randomized study is

$$Pr(A = +1|X) = \frac{\exp\left(Z^{(1)} - Z^{(2)} - Z^{(4)} - Z^{(6)} - Z^{(8)} - Z^{(10)}\right)}{1 + \exp\left(Z^{(1)} - Z^{(2)} - Z^{(4)} - Z^{(6)} - Z^{(8)} - Z^{(10)}\right)}.$$

A.9.5 Setting 2

We set $X \in \mathbb{R}^{50}$ by defining $X^{(1)} = \exp(Z^{(1)}+1)+Z^{(2)}$; $X^{(2)} = Z^{(2)2} \cdot Z^{(3)}$; $X^{(3)} = \sin(2Z^{(3)}) \cdot (Z^{(4)}+5)^2$; $X^{(4)} = \left(Z^{(2)3} + Z^{(4)} + 10\right) \cdot \left(Z^{(2)} + Z^{(4)3} + 10\right)$; $X^{(i)} = Z^{(i)}$ for $i \in [n]\backslash\{1, 2, 3, 4\}$. The dimension-reduced covariate $V = B_0^\top X \in \mathbb{R}^2$ is defined as in Setting 1. The outcome is generated as $N\{\mu(X) + A \cdot \tilde{f}^V(V), 1\}$, where

$$\mu(X) = 10+7Z^{(1)}+13Z^{(2)}+15Z^{(3)}+15Z^{(4)}+10Z^{(5)}+7Z^{(6)}+13Z^{(7)}+15Z^{(8)}+15Z^{(9)}+10Z^{(10)},$$

$$\tilde{f}^V(V) = 6\sin\left(V^{(1)}/3\right)\cdot\log\left(\left|V^{(2)}\right|+1\right)+4\cos\left(V^{(2)}\sqrt{|V^{(1)}|}\right)+5\tan^{-1}\left\{2\left(V^{(1)}-1\right)\log\left(\left|V_0^{(2)}\right|+1\right)\right\}.$$

For the non-randomized study setting, we define the propensity score as

$$\Pr(A=+1|X) = \frac{\exp\left(-Z^{(3)}+2Z^{(4)}-Z^{(5)}-0.5Z^{(6)}\right)}{1+\exp\left(-Z^{(3)}+2Z^{(4)}-Z^{(5)}-0.5Z^{(6)}\right)}.$$

Figure A.1 displays the distribution of the Bayes rule at different levels of covariates for Setting 2.

A.9.6 Setting 3

In Setting 3, we consider a higher dimension for the central mean subspace. We set $X \in \mathbb{R}^{50}$ by defining $X^{(2)} = \left(Z^{(2)} - 0.2Z^{(4)} + 6\right)^3$; $X^{(4)} = \exp\left(0.5Z^{(4)}\right)$; $X^{(6)} = Z^{(6)}\cdot\left\{1+\exp\left(Z^{(4)}\right)\right\}^{-1}$; $X^{(8)} = \left(Z^{(6)} + Z^{(8)} + 20\right)^2$; $X^{(i)} = Z^{(i)}$ for $i \in [n]\backslash\{2,4,6,8\}$. The dimension-reduced covariate $V = B_0^\top X \in \mathbb{R}^4$ is defined from a basis matrix that consists of $\left(1/\sqrt{2}, -1/\sqrt{2}, \mathbf{0}_{48}^\top\right)^\top$, $\left(1/\sqrt{2}, 1/\sqrt{2}, \mathbf{0}_{48}^\top\right)^\top$, $\left(0, 0, 1/\sqrt{2}, -1/\sqrt{2}, \mathbf{0}_{46}^\top\right)^\top$, and $\left(0, 0, 1/\sqrt{2}, 1/\sqrt{2}, \mathbf{0}_{46}^\top\right)^\top$. The outcome is generated as $N\{\mu(X) + A\cdot\tilde{f}^V(V), 1\}$, where

$$\mu(X) = 6Z^{(1)} + 6Z^{(2)} + 10Z^{(3)} + 10Z^{(4)} + 12Z^{(5)} + 12Z^{(6)} + 8Z^{(7)} + 8Z^{(8)} + 6Z^{(9)} + 6Z^{(10)},$$

$$\tilde{f}^V(V) = \left(V^{(3)}+10\right)\cos\left\{2\pi\log\left(V^{(1)}-2\right)\right\} + 3\sqrt{-V^{(2)}}\cdot\sin\left(0.5\pi V^{(1)}\right)/\left(\pi\sqrt{V^{(1)}}\right)$$
$$+\tan^{-1}\left\{2\log\left(-V^{(2)}\right)\right\}\sqrt{|V^{(4)}-4|} - 3.$$

For the non-randomized study setting, we define the propensity score as

$$\Pr(A=+1|X) = \frac{\exp\left(0.6Z^{(1)}+1.2Z^{(2)}+1.2Z^{(3)}-0.8Z^{(4)}-Z^{(5)}-Z^{(6)}\right)}{1+\exp\left(0.6Z^{(1)}+1.2Z^{(2)}+1.2Z^{(3)}-0.8Z^{(4)}-Z^{(5)}-Z^{(6)}\right)}.$$

A.10 Additional simulation results

A.10.1 Results for $n = 500$

The simulation results for the smaller sample size ($n = 500$) in the randomized and non-randomized scenarios are displayed in Figure A.2 and Figure A.3, respectively. With the

smaller sample size, we can observe that the performance of the proposed DOL is more pronounced than the results with $n = 1000$ when compared to AOL in all settings.

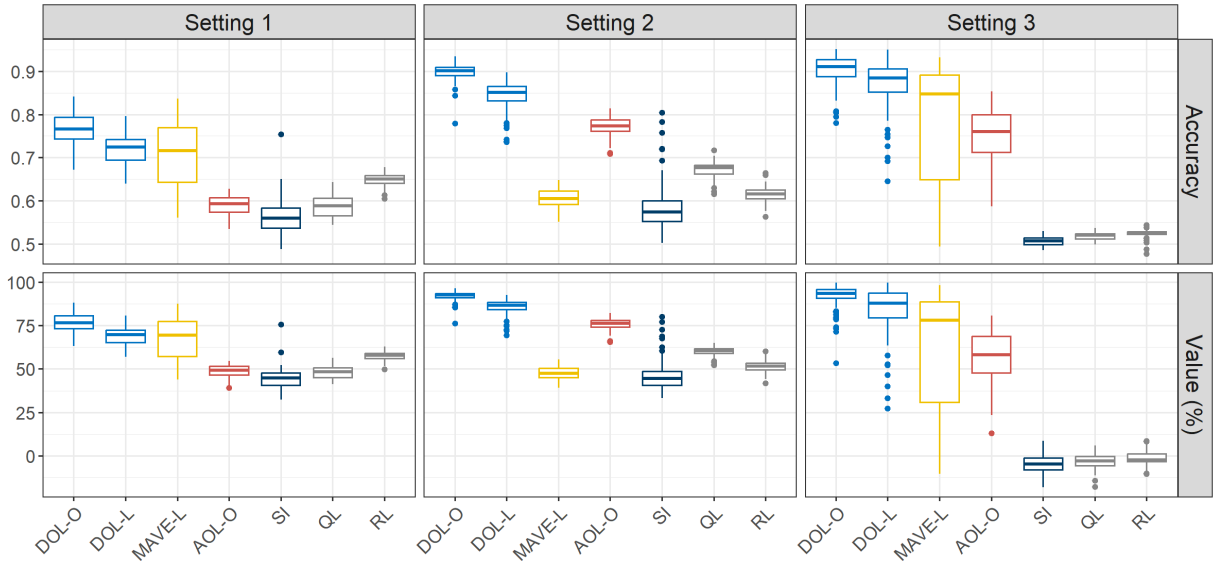

Figure A.2: Simulation results for $n = 500$ from the randomized scenarios across the three different settings. Accuracy represents the proportion of correctly predicted optimal treatments. Value (%) represents the value function recovered by the estimated decision rule as a percentage of the Bayes optimal value function. The performance of the proposed method is shown under DOL-O and DOL-L: DOL-O uses the oracle $g(x)$, while DOL uses $g_{\hat{w}}(x)$ estimated via linear regression. MAVE-L uses interaction MAVE for dimension reduction and uses the fitted values $g_n(x)$ obtained via linear regression.

A.10.2 Projection error

Here we introduce additional simulation results that show decreasing trends along the increasing sample size in the difference between the estimated projection from $\hat{B}$ and the true projection from B_0. To do so, we iterated the process of obtaining $\hat{B}$ and computing the Frobenius distance from B_0 by $\|\hat{B}\hat{B}^\top - B_0 B_0^\top\|$ at different sample sizes: 200, 500, 1,000, and 2,000. The simulation was performed on the three non-randomized settings used for comparing different approaches, where the $\hat{B}$ obtained from the pseudo outcomes that use KCB weights. The dimensions for $\hat{B}$ were fixed at the true values and the Tikhonov penalty

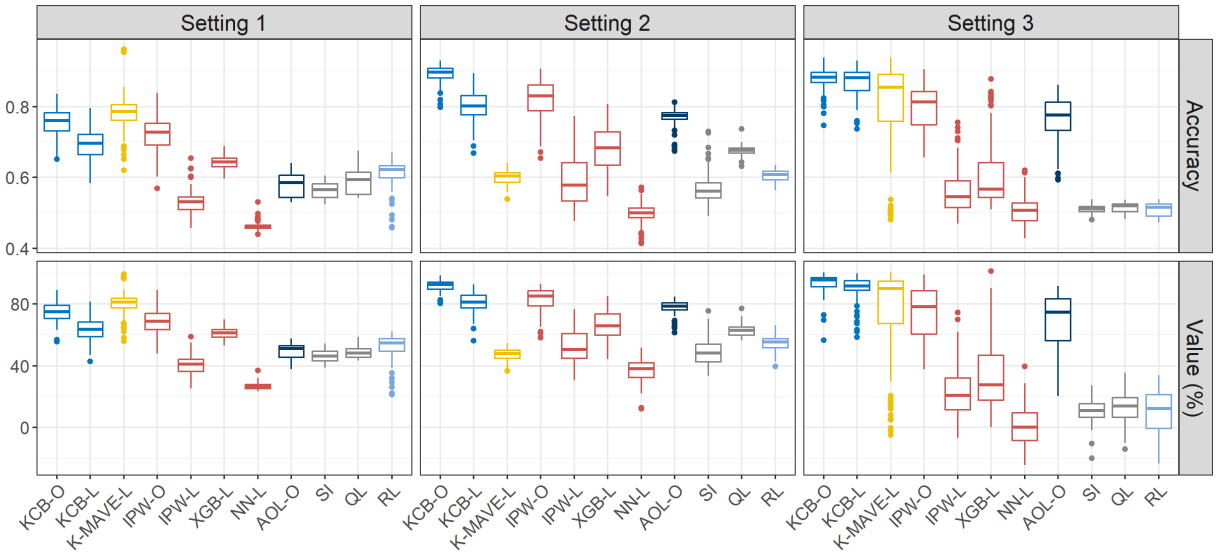

Figure A.3: Simulation results for $n = 500$ from the non-randomized scenarios across the three different settings. Accuracy represents the proportion of correctly predicted optimal treatments. Value (%) represents the value function recovered by the estimated decision rule as a percentage of the Bayes optimal value function. The performance of the proposed method is shown under KCB-O and KCB-L: KCB-O uses the oracle $g(x)$, while KCB-L uses the fitted values of $g_{\hat{w}}(x)$ obtained via linear regression. K-MAVE-L uses interaction MAVE for dimension reduction and uses KCB weights along with the fitted values of $g_{\hat{w}}(x)$ obtained via linear regression. IPW-O uses the oracle $g(x)$, and IPW-L uses $g_{\hat{w}}(x)$ from linear regression, with propensity scores estimated via logistic regression. XGB-L uses $g_{\hat{w}}(x)$ from linear regression, with propensity scores estimated via XGBoost. NN-L uses $g_{\hat{w}}(x)$ from linear regression, with propensity scores estimated via a neural network.

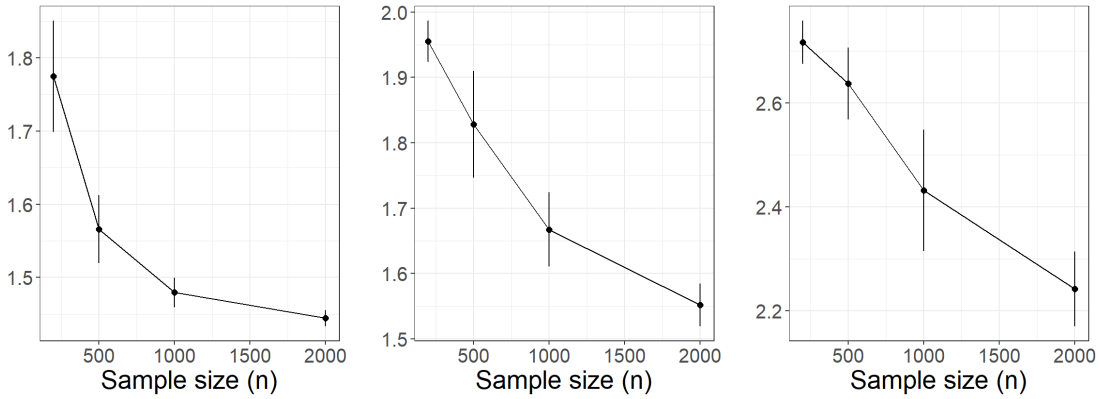

Figure A.4: Simulation results that display the change in the error in the projection from $\hat{B}$ along different sample sizes represented by $\|\hat{B}\hat{B}^\top - B_0 B_0^\top\|_F$ (y-axis). From left to right, the figures correspond to the non-randomized treatment assignment scenarios of Settings 1, 2, and 3. The means and standard deviations are depicted at each sample size.

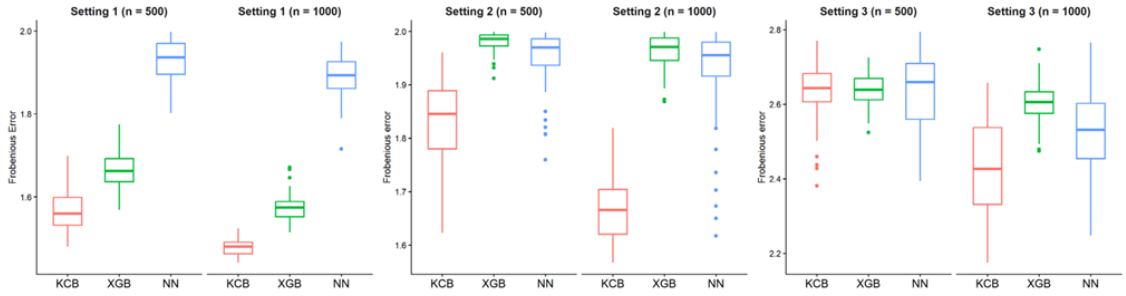

Figure A.5: Comparison of projection errors of $\hat{B}$ ($\|\hat{B}\hat{B}^\top - B_0 B_0^\top\|_F$) between the proposed KCB, XGBoost (Matthew Cefalu et al. 2024), and neural network (Shang et al. 2025). From left to right, the graphs correspond to the non-randomized treatment assignment scenarios of Settings 1, 2, and 3. The means and standard deviations are depicted at each sample size.

was fixed at 10^{-5} to focus on the association with the sample size. The results are displayed in Figure B.5. We can notice the decline in the projection errors as the sample size increases.

Comparison of projection error

We further compared the projection errors of $\hat{B}$, estimated by gKDR, between different weighting methods – KCB, XGBoost (Matthew Cefalu et al. 2024), and neural network (Shang et al. 2025). Figure A.5 shows that using KCB weights for estimating $\hat{B}$ in gKDR has the lowest error.

A.10.3 Computation time

To study the computation costs of the two dimension reduction methods, we conducted simulation study to compare the computation times. Specifically, we compared the computation times of the proposed gKDR with the interaction mean average variance estimation (iMAVE) from (Liang and Yu 2022). We implement iMAVE from the iMAVE package in `https://github.com/muxuanliang/iMAVE`. The simulation used the true dimension u of the projected covariates. Both the gKDR and the iMAVE used the rule of thumb bandwidth for the kernel bandwidth h – gKDR used the median heuristic approach to choose h to be used in Gaussian kernel and the iMAVE used $h = \{4/(u+2)\}^{1/(u+4)} \cdot n^{-1/(u+4)}$ for local kernel smoothing. We implemented iMAVE with single iteration and the maximum iteration of five. The simulation was performed for all three simulation settings in Section 5 with $n = 500$ and $n = 1000$. The results are presented in Figure A.6.

The iMAVE had larger computation times than gKDR at all three settings with the difference being multifold even when iMAVE used a single iteration. Moreover, the iMAVE tends to have greater computation cost as the dimension u of the reduced subspace increases from the Setting 3 results where $u = 4$, whereas $u = 2$ for Setting 1 and Setting 2. This is expected since iMAVE uses kernel smoothing regression in local kernel smoothing in the objective function. This potentially limits the effective implementation of the method to settings there the effective dimension reduction space is low dimensional while gKDR admits moderately high dimensional reducing subspace. On the other hand, gKDR reduces the problem to kernel matrix operations and eigenspace estimation. Furthermore iMAVE uses iteration until convergence. Hence, it takes longer time to converge with larger iterations when the reduced subspace has larger dimension. We can observe this from the bottom row of Figure A.6 that displays the computation cost comparison where the iMAVE uses up to five iterations to estimate the dimension reduction space.

A.11 Real world data

In this section, we layout the details on the 35 baseline variables: age (18-91), gender, weight, and indicator of service unit (medical ICU or surgical ICU); *severity at admission* measurements from simplified acute physiology score, the sequential organ failure assess-

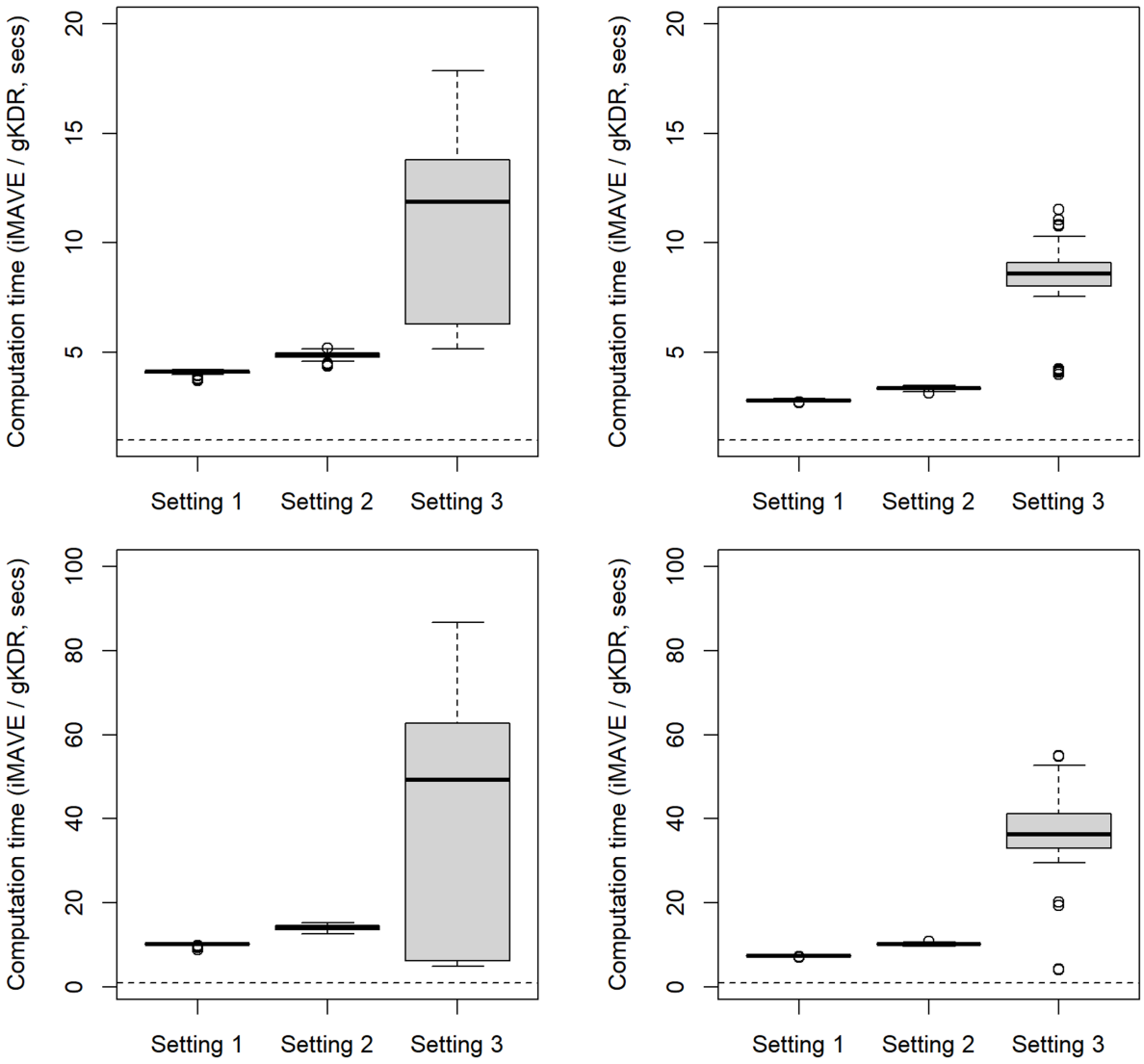

Figure A.6: Computation time (seconds) ratios for dimension reduction measured as the runtime of gKDR divided by the runtime of iMAVE, where the iMAVE runtime corresponds to a single iteration (top) or five or less iterations (bottom). Within each row, the left plot shows results for $n = 500$, and the right plot shows results for $n = 1000$. The horizontal dashed line indicates a ratio of 1.

ment, and Elixhauser comorbidity score; *comorbidity indicators*, which are congestive heart failure, atrial fibrillation, chronic renal disease, liver disease, chronic obstructive pulmonary disease (COPD), coronary artery disease (CAD), stroke, respiratory failure, and malignant tumor; *vital signs* measured from mean arterial pressure, heart rate, temperature (°F); *interventions* that use mechanical ventilation, inotropic and vasopressor agents, and/or sedative drugs within 24 hours of ICU admission; *laboratory results* that measured white blood cell, hemoglobin, sodium, potassium, bicarbonate, chloride, creatinine, pH, partial pressure of carbon dioxide (PCO2), platelet count, partial pressure of oxygen (PO2), lactate, and blood urea nitrogen.

A.11.1 Estimation of ITR and evaluation of modified value function

The gKDR was implemented by selecting the regularization parameter ϵ_n out of 8 candidate values (10^{-9}-10^{-2}) and the dimension of central mean subspace (u) out of 9 candidate dimensions (1-9) from five-fold cross-validation.

AOL was implemented by using the IPW estimated from logistic regression. In DOL, $g_{\hat{w}}(x)$ was estimated by linear regression. Likewise, $g_(x)$ was estimated by linear regression with the estimated propensity scores. The tuning parameter λ_n was chosen from five-fold cross-validation.

The modified value function is evaluated by using the validation set as

$$\frac{1}{n_{test}} \sum_{i=1}^{n_{test}} \frac{Y_i^{test}}{\hat{\pi}(A_i^{test}, X_i^{test})} \left[\mathbb{1}\left\{ A_i^{test} = \hat{d}\left(X_i^{test}\right) \right\} - \mathbb{1}\left\{ A_i^{test} = \hat{d}\left(X_i^{test}\right) \right\} \right].$$

The true propensity scores are not available to compute the modified value function with the validation set. Thus, the propensity scores were estimated from gradient boosted models (McCaffrey et al. 2004) as done in Feng et al. (2018) by using `ps` from R package `twang` (Matthew Cefalu et al. 2024), and were used in place of $\hat{\pi}(A_i^{test}, X_i^{test})$.

Appendix B

Chapter 4 Appendix

B.1 Proof of Proposition 3.3.1

Since $GG^\top + \Phi_2\Phi_2^\top + \Phi_0\Phi_0^\top = I_r$, we have $GG^\top M + \Phi_2\Phi_2^\top M + \Phi_0\Phi_0^\top M = M$. Then

$$\Gamma_2^\top GG^\top M + \Gamma_2^\top \Phi_2\Phi_2^\top M + \Gamma_2^\top \Phi_0\Phi_0^\top M = \Gamma_2^\top \Phi_0\Phi_0^\top M = \Gamma_2^\top M.$$

Since this holds for an arbitrary M, we have $\Gamma_2^\top \Phi_0\Phi_0^\top = \Gamma_2^\top$ or $\Phi_0\Phi_0^\top\Gamma_2 = \Gamma_2$, hence $\text{span}(\Gamma_2) \subseteq \text{span}(\Phi_0)$. Similarly, we have $\text{span}(\Phi_2) \subseteq \text{span}(\Gamma_0)$.

Since $\text{span}(\Gamma_2) \subseteq \text{span}(\Phi_0)$ and $\text{Cov}(\Phi^\top M, \Phi^\top M \mid A) = 0$, we have $\text{Cov}(\Gamma_2^\top M, \Phi^\top M \mid A) = 0$. Hence $\text{Cov}(\Gamma_2^\top M, G^\top M \mid A) = 0$ and $\text{Cov}(\Gamma_2^\top M, \Phi_2^\top M \mid A) = 0$. Likewise, $\text{span}(\Phi_2) \subseteq \text{span}(\Gamma_0)$ and $\text{Cov}(\Gamma_0^\top M, \Gamma^\top M \mid A) = 0$ additionally implies $\text{Cov}(\Phi_2^\top M, G^\top M \mid A) = 0$. Since $\text{span}(D_0) = \text{span}(\Phi_0) \cap \text{span}(\Gamma_0)$, we have $\text{Cov}(D_0^\top M, G^\top M \mid A) = \text{Cov}(D_0^\top M, \Gamma_2^\top M \mid A) = \text{Cov}(D_0^\top M, \Phi_2^\top M \mid A) = 0$.

B.2 Proof of Proposition 3.3.3

The proof is provided in Section B.8.4. The proof uses the notations and results from Section B.8.

B.3 Proof of Proposition 3.3.4

Define Q_0 as

$$Q_0 = \log|\Gamma^\top \Sigma_{M|A}\Gamma| + \log|\Gamma^\top \Sigma_M^{-1}\Gamma| + \log|\Phi^\top \Sigma_{R|\tilde{Y}}\Phi| + \log|\Phi^\top \Sigma_R^{-1}\Phi|.$$

Note that the objective functions Q_1 and Q_2 were derived from further analyzing Q_0. From that the column vectors of Γ and Φ consist of the bases for G, Γ_2, and Φ_2, this can be equivalently expressed as

$$\begin{aligned} Q_0 = \log\left|\begin{pmatrix} G & \Gamma_2 \end{pmatrix}^\top \Sigma_{M|A} \begin{pmatrix} G & \Gamma_2 \end{pmatrix}\right| + \log\left|\begin{pmatrix} G & \Gamma_2 \end{pmatrix}^\top \Sigma_M^{-1} \begin{pmatrix} G & \Gamma_2 \end{pmatrix}\right| \\ + \log\left|\begin{pmatrix} G & \Phi_2 \end{pmatrix}^\top \Sigma_{R|\tilde{Y}} \begin{pmatrix} G & \Phi_2 \end{pmatrix}\right| + \log\left|\begin{pmatrix} G & \Phi_2 \end{pmatrix}^\top \Sigma_R^{-1} \begin{pmatrix} G & \Phi_2 \end{pmatrix}\right|. \end{aligned}$$

The minimizing triplet (G, Γ_2, Φ_2) also minimizes Q_1 and Q_2 that uses the true covariance matrices $\Sigma_{M|A}$, $\Sigma_{R|Y}$, and Σ_M. From Proposition 6.1 of Cook (2018a), the minimum of Q_0 is $\log|\Sigma_{M|A}| + \log|\Sigma_M^{-1}|$. This is achieved when $\text{span}(G) \cup \text{span}(\Gamma_2) = \text{span}(\Gamma)$ and $\text{span}(G) \cup \text{span}(\Phi_2) = \text{span}(\Phi)$. Therefore, $\text{span}(G) \subseteq \text{span}(\Gamma)$ and $\text{span}(G) \subseteq \text{span}(\Phi)$. Furthermore, we have $\text{span}(G) \cap \text{span}(\Gamma_2) = \varnothing$ and $\text{span}(G) \cap \text{span}(\Phi_2) = \varnothing$ from the given orthogonality condition $G^\top \Gamma_2 = 0_{u,d_2}$ and $G^\top \Phi_2 = 0_{u,q_2}$. Therefore, $\text{span}(G) = \text{span}(\Gamma) \cap \text{span}(\Phi)$. Moreover, $\text{span}(\Gamma_2) = \text{span}(\Gamma) \backslash \text{span}(G)$ and $\text{span}(\Phi_2) = \text{span}(\Phi) \backslash \text{span}(G)$.

B.4 Proof of Proposition 3.3.6

The normal convergence of $\sqrt{n}\left(h(\hat{\theta}) - h(\theta)\right)$ follows from Proposition 4.1 of Shapiro (1986). Thus, we show the proof of the inequalities of the variances.

Since $P_{H(J)} J^{-1} P_{H(J)}^\top = H(H^\top JH)^\dagger H^\top$, $J^{-1} - P_{H(J)} J^{-1} P_{H(J)}^\top = J^{-1} - H(H^\top JH)^\dagger H^\top = J^{-1/2}\left\{I_r - J^{1/2}H(H^\top JH)^\dagger H^\top J^{1/2}\right\} J^{-1/2}$. Since $J^{-1/2}\left\{I_r - J^{1/2}H(H^\top JH)^\dagger H^\top J^{1/2}\right\} J^{-1/2}$ is positive semi-definite, $J^{-1} - P_{H(J)} J^{-1} P_{H(J)}^\top \succeq 0$. By delta method,

$$\sqrt{n}\left(\mathbb{I}_r h(\hat{\theta}) - \mathbb{I}_r h(\theta)\right) = \sqrt{n}\left(\begin{pmatrix} \hat{\alpha}_1 \\ \hat{\beta}_1^\top \end{pmatrix} - \begin{pmatrix} \alpha_1 \\ \beta_1^\top \end{pmatrix}\right) \to N\left(0,\ \mathbb{I}_r H(H^\top JH)^\dagger H^\top \mathbb{I}_r^\top\right)$$

where

$$\mathbb{I}_r = \begin{pmatrix} I_r & 0 & 0 & 0 & 0 \\ 0 & I_r & 0 & 0 & 0 \end{pmatrix}.$$

Then by applying delta method once more, we have

$$\sqrt{n}\left(\psi\{\mathbb{I}_r h(\hat{\theta})\} - \psi\left\{\mathbb{I}_r h(\theta)\right\}\right) = \sqrt{n}\left(\hat{\beta}_1\hat{\alpha}_1 - \beta_1\alpha_1\right)$$

$$\to N\left(0,\ \begin{pmatrix}\beta_1^\top \\ \alpha_1\end{pmatrix}^\top \mathbb{I}_r H(H^\top J H)^\dagger H^\top \mathbb{I}_r^\top \begin{pmatrix}\beta_1^\top \\ \alpha_1\end{pmatrix}\right),$$

where $\psi\left\{(\alpha_1^\top, \beta_1)^\top\right\} = \beta_1\alpha_1$. Similarly, we have

$$\sqrt{n}\left(\tilde{\beta}_1\tilde{\alpha}_1 - \beta_1\alpha_1\right) \to N\left(0,\ \begin{pmatrix}\beta_1^\top \\ \alpha_1\end{pmatrix}^\top \mathbb{I}_r J^{-1} \mathbb{I}_r^\top \begin{pmatrix}\beta_1^\top \\ \alpha_1\end{pmatrix}\right).$$

Therefore,

$$\operatorname{avar}(\sqrt{n}\tilde{\beta}_1\tilde{\alpha}_1) - \operatorname{avar}(\sqrt{n}\hat{\beta}_1\hat{\alpha}_1) = \begin{pmatrix}\beta_1^\top \\ \alpha_1\end{pmatrix}^\top \mathbb{I}_r\left(J^{-1} - H(H^\top J H)^\dagger H^\top\right)\mathbb{I}_r^\top \begin{pmatrix}\beta_1^\top \\ \alpha_1\end{pmatrix} \geq 0.$$

B.5 Proof of Proposition 3.3.7

By Proposition 4.2 of Shapiro (1986), $\sqrt{n}(\hat{\theta} - \theta)$ is asymptotically normal with one admissible covariance matrix $\Pi = (H^\top J H)^\dagger$. Let H_1 be the matrix built with the columns of H that correspond to the gradients of $h(\theta)$ with respect to η_1, ξ_1, and $\operatorname{vec}(G)$. Let H_2 be the matrix built with the remaining columns of H. Define $U_2 = J - JH_2(H_2^\top J H_2)^\dagger H_2^\top J$. Then $\sqrt{n}\left\{\left(\hat{\eta}_1^\top, \hat{\xi}_1^\top, \operatorname{vec}(\hat{G})^\top\right)^\top - \left(\eta_1^\top, \xi_1^\top, \operatorname{vec}(G)^\top\right)^\top\right\}$ converges to a mean zero normal distribution where $\Pi_1 = (H_1^\top U_2 H_1)^\dagger$ is one particular variance. Since $\operatorname{span}(\Gamma)$ and $\operatorname{span}(\Phi)$ are estimable, $\operatorname{span}(G) = \operatorname{span}(\Gamma) \cap \operatorname{span}(\Phi)$ is estimable. Thus, $P_G\alpha_1 = G\eta_1$ and $P_G\beta_1 = G\xi_1$ are estimable. Define $\psi\left\{\left(\eta_1^\top, \xi_1^\top, \operatorname{vec}(G)^\top\right)^\top\right\} = \left((G\eta_1)^\top, (G\xi_1)^\top\right)^\top$. Then the Jacobian of ψ, $\nabla\psi$ is

$$\nabla\psi = \begin{pmatrix} G & 0 & \eta_1^\top \otimes I_r \\ 0 & G & \xi_1^\top \otimes I_r \end{pmatrix}.$$

By delta method,

$$\sqrt{n}\left\{\psi\begin{pmatrix}\hat{\eta}_1\\ \hat{\xi}_1\\ \mathrm{vec}(G)\end{pmatrix}-\psi\begin{pmatrix}\eta_1\\ \xi_1\\ \mathrm{vec}(G)\end{pmatrix}\right\}=\sqrt{n}\left\{\begin{pmatrix}\hat{G}\hat{\eta}_1\\ \hat{G}\hat{\xi}_1\end{pmatrix}-\begin{pmatrix}G\eta_1\\ G\xi_1\end{pmatrix}\right\}\to N(0,\ \nabla\psi\Pi_1\nabla\psi^\top).$$

B.6 Proof of Proposition 3.3.9

We provide proofs for Lemma 3.3.8 and Proposition 3.3.9 after following the formulation of the maximum log-likelihood to the minimum discrepancy function.

B.6.1 Formulation of likelihood to discrepancy function

First, we show that the maximization of the joint-likelihood in (3.5) can be formulated into that of the moment structural model problem. Shapiro (1986) introduces the discrepancy function $F(\mathrm{x},\mathrm{x}_0)$ that is used to find the minimizing estimator $\tilde{\mathrm{x}}$ of the parameter x_0. Let $\mathrm{x}_0 = h(\theta_0)$ and $\hat{\mathrm{x}} = h(\hat{\theta})$ for $h(\theta)$ introduced in Section 3.3.2 which is at least twice differentiable with respect to θ. When $\sqrt{n}(\tilde{\mathrm{x}}-\mathrm{x}_0)$ converges to normality then, we can establish the asymptotic normality of $\sqrt{n}(h(\hat{\theta})-h(\theta_0))$ under the following regularity conditions:

- $F(\mathrm{x},\mathrm{x}_0)\geq 0$;
- $F(\mathrm{x},\mathrm{x}_0)=0$ only if $\mathrm{x}=\xi$;
- F is twice differentiable.
- There exist $C>0$, $\epsilon>0$ such that $F(\mathrm{x},\mathrm{x}_0)\geq\epsilon$ whenever $\|\mathrm{x}-\mathrm{x}_0\|\geq C$.

Define $\beta=(\beta_1,\beta_2)^\top$ and $W_i=(M_i^\top,A_i)^\top$, so that $\beta_1 M_i+\beta_2 A_i=\beta W_i$. Omitting the constant terms, (3.5) is the sum of $\ell_1(\alpha_1,\Sigma_{M|A})$ and $\ell_2(\beta,\sigma_Y^2)$, where

$$\ell_1(\alpha_1,\Sigma_{M|A})=-\frac{n}{2}\log|\Sigma_{M|A}|-\frac{1}{2}\sum_{i\in[n]}(M_i-\alpha_1 A_i)^\top\Sigma_{M|A}^{-1}(M_i-\alpha_1 A_i)$$

and

$$\ell_2(\beta,\sigma_Y^2)=-\frac{n}{2}\log\sigma_Y^2-\frac{1}{2}\sum_{i\in[n]}(Y_i-\beta W_i)^2/\sigma_Y^2.$$

The two log likelihoods can be formulated as

$$\begin{aligned}
\ell_1(\alpha_1, \Sigma_{M|A}) &= -\frac{n}{2}\log|\Sigma_{M|A}| - \frac{1}{2}\sum_{i\in[n]}(M_i - \alpha_1 A_i)^\top \Sigma_{M|A}^{-1}(M_i - \alpha_1 A_i) \\
&= -\frac{n}{2}\log|\Sigma_{M|A}| - \frac{1}{2}\sum_{i\in[n]}(M_i - \tilde{\alpha}_1 A_i + \tilde{\alpha}_1 A_1 - \alpha_1 A_i)^\top \Sigma_{M|A}^{-1}(M_i - \tilde{\alpha}_1 A_i + \tilde{\alpha}_1 A_1 - \alpha_1 A_i) \\
&= -\frac{n}{2}\log|\Sigma_{M|A}| - \frac{n}{2}\frac{1}{n}\sum_{i\in[n]}(M_i - \tilde{\alpha}_1 A_i)^\top \Sigma_{M|A}^{-1}(M_i - \tilde{\alpha}_1 A_i) \\
&\quad - \frac{1}{2}\sum_{i\in[n]}(\tilde{\alpha}_1 - \alpha_1)^\top \Sigma_{M|A}^{-1}(\tilde{\alpha}_1 - \alpha_1)A_i^2 \\
&= -\frac{n}{2}\log|\Sigma_{M|A}| - \frac{n}{2}\mathrm{tr}\left(\Sigma_{M|A}^{-1}S_{M|A}\right) - \frac{1}{2}\sum_{i\in[n]}(\tilde{\alpha}_1 - \alpha_1)^\top \Sigma_{M|A}^{-1}(\tilde{\alpha}_1 - \alpha_1)A_i^2
\end{aligned}$$

and

$$\begin{aligned}
\ell_2(\beta, \sigma_Y^2) &= -\frac{n}{2}\log\sigma_Y^2 - \frac{1}{2}\sum_{i\in[n]}(Y_i - \beta W_i)^2/\sigma_Y^2 \\
&= -\frac{n}{2}\log\sigma_Y^2 - \frac{n}{2}\frac{1}{n}\sum_{i\in[n]}(Y_i - \tilde{\beta}W_i + \tilde{\beta}W_i - \beta W_i)^2/\sigma_Y^2 \\
&= -\frac{n}{2}\log\sigma_Y^2 - \frac{n}{2\sigma_Y^2}\frac{1}{n}\sum_{i\in[n]}\left\{(Y_i - \tilde{\beta}W_i)^2 + (\tilde{\beta}W_i - \beta W_i)^2\right\} \\
&= -\frac{n}{2}\log\sigma_Y^2 - \frac{n}{2}\frac{s_{Y|W}^2}{\sigma_Y^2} - \frac{1}{2}\sum_{i\in[n]}\{(\tilde{\beta} - \beta)W_i\}^2/\sigma_Y^2
\end{aligned}$$

Let $\tilde{\mathrm{x}} = \left(\tilde{\alpha}_1^\top, \tilde{\beta}_1, \tilde{\beta}_2, \mathrm{vech}(S_{M|A})^\top, s_Y^2\right)^\top$ and $\mathrm{x}_0 = \left(\alpha_1^\top, \beta_1, \beta_2, \mathrm{vech}(\Sigma_{M|A})^\top, \sigma_Y^2\right)^\top$. Define $\ell(\alpha_1, \beta, \Sigma_{M|A}, \sigma_Y^2) = \ell_1 + \ell_2$. Let $\ell_{max} = \ell_1(\tilde{\alpha}_1, S_{M|A}) + \ell_2(\tilde{\beta}, s_{Y|W}^2)$ which is the maximum log likelihood. Then the minimum discrepancy function can be constructed as

$$F_{min}\left(\tilde{\mathrm{x}}, \mathrm{x}_0\right) = \ell_{max} - \ell(\alpha_1, \beta, \Sigma_{M|A}, \sigma_Y^2)$$

Note that $F_{min}\left(\tilde{\mathrm{x}}, \mathrm{x}_0\right) = 0$ as long as $x_0 = \tilde{\mathrm{x}}$ and it is always nonnegative. Also, since F_{min} is constructed based on the log-likelihood of multivariate normal distribution, it is at least twice differentiable. Therefore, the minimum discrepancy function satisfies the regularity conditions for the discrepancy function introduced in Shapiro (1986).

B.6.2 Proof of Lemma 3.3.8

Throughout this proof, we assume that M and A are centered. For the proof, we use Theorem B of Serfling (1980); for completeness, we restate it below.

Theorem B.6.1. *Let $\{X_i\}$ random vectors have means $\{\mu_i\}$ and covariance $\{\Sigma_i\}$, and distribution functions $\{F_i\}$. Suppose*

$$\frac{\Sigma_1 + \Sigma_2 + \cdots + \Sigma_n}{n} \to_p \Sigma,$$

and

$$\frac{1}{n}\sum_{i\in[n]} \mathbb{E}[\|X_i - \mu_i\|^2 \mathbb{1}(\|X_i - \mu_i\| > \epsilon\sqrt{n})] \to 0.$$

Then the following distributional convergence follows:

$$\frac{1}{\sqrt{n}}\sum_{i\in[n]} X_i \longrightarrow N\left(\frac{1}{\sqrt{n}}\sum_{i\in[n]}\mu_i, \Sigma\right)$$

Let $\tilde{\alpha}_1$, $\tilde{\beta}^\top$, $S_{M|A}$, and $s_Y|W^2$ be the maximum likelihood estimators based on the joint log-likelihood (3.5). We have the following differences:

$$\tilde{\alpha}_1 - \alpha_1 = \frac{1}{n}\sum_{i\in[n]} a_i\epsilon_{M,i} = \frac{1}{n}\sum_{i\in[n]} L_{1,i}$$

$$\tilde{\beta} - \beta = \frac{1}{n}\sum_{i\in[n]} w_i\epsilon_{Y,i} = \frac{1}{n}\sum_{i\in[n]} L_{2,i}$$

$$\begin{aligned}\operatorname{vech}(S_{M|A}) - \operatorname{vech}(\Sigma_{M|A}) &= \frac{1}{n}\sum_{i\in[n]} \operatorname{vech}\left(\epsilon_{M,i}\epsilon_{M,i}^\top - \Sigma_{M|A}\right) - \operatorname{vech}\left(\bar{\epsilon}_M\bar{\epsilon}_M^\top + \frac{1}{n}e_M^\top P_A e_M\right) \\ &= \frac{1}{n}\sum_{i\in[n]} L_{3,i} + o_p(1)\end{aligned}$$

$$s^2_{Y|W} - \sigma^2_Y = \frac{1}{n}\sum_{i\in[n]}(\epsilon^2_{Y,i} - \sigma^2_Y) - (\bar{\epsilon}^2_Y + e^2_Y P_W e_Y) = \frac{1}{n}\sum_{i\in[n]} L_{4,i} + o_p(1).$$

where $\mathbb{A} \in \mathbb{R}^{n\times 1}$ is a vector of A_i's, $a_i = A_i / \left(n^{-1}\sum_{i\in[n]} A_i^2\right)$, w_i is the i-th row of

$\mathbb{W}(\mathbb{W}^\top\mathbb{W}/n)^{-1}$, and $L_{3,i} = \text{vech}\left(\epsilon_{M,i}\epsilon_{M,i}^\top - \Sigma_{M|A}\right)$, $L_{4,i} = \epsilon_{Y,i}^2 - \sigma_Y^2$, $e_M \in \mathbb{R}^{n\times r}$ is the matrix whose i-th row is $\epsilon_{M,i}^\top$, and $e_Y \in \mathbb{R}^n$ is a vector of $\epsilon_{Y,i}$'s. Su and Cook (2012) show that $\text{vech}\left(\bar{\epsilon}_M\bar{\epsilon}_M^\top + \frac{1}{n}e_M^\top P_A e_M\right) = o_p(1)$, and similarly, $\bar{\epsilon}_Y^2 + e_Y^2 P_W e_Y = o_p(1)$. Furhtermore, $L_{3,i}$ and $L_{4,i}$ are independent and identically distributed with mean zero. Define $L_{3,i}^* = \text{vec}\left(\epsilon_{M,i}\epsilon_{M,i}^\top - \Sigma_{M|A}\right) = \epsilon_{M,i}\otimes\epsilon_{M,i} - \text{vec}(\Sigma_{M|A})$. Define

$$V_i = \begin{pmatrix} L_{1,i} \\ L_{2,i} \\ L_{3,i}^* \\ L_{4,i} \end{pmatrix} = \begin{pmatrix} a_i\epsilon_{M,i} \\ w_i\epsilon_{Y,i} \\ \epsilon_{M,i}\otimes\epsilon_{M,i} - \text{vec}(\Sigma_{M|A}) \\ \epsilon_{Y,i}^2 - \sigma_Y^2 \end{pmatrix}.$$

Define u_i to be a vector with n components such that the i-th component is 1 while others are 0. Note that $w_i^\top = u_i^\top\mathbb{W}(\mathbb{W}^\top\mathbb{W}/n)^{-1}$. Therefore

$$\begin{aligned}
\|V_i\|^2 &= \|L_{1,i}\|^2 + \|L_{2,i}\|^2 + \|L_{3,i}^*\|^2 + \|L_{4,i}\|^2 \\
&= a_i^2\|\epsilon_{M,i}\|^2 + \|w_i\|^2\epsilon_{Y,i}^2 + \|L_{3,i}^*\|^2 + \left(\epsilon_{Y,i}^2 - \sigma_Y^2\right)^2 \\
&= a_i^2\|\epsilon_{M,i}\|^2 + u_i^\top\mathbb{W}(\mathbb{W}^\top\mathbb{W}/n)^{-2}\mathbb{W}^\top u_i\cdot\epsilon_{Y,i}^2 + \|L_{3,i}^*\|^2 + \left(\epsilon_{Y,i}^2 - \sigma_Y^2\right)^2,
\end{aligned}$$

and $\|w_i\|^2/n = u_i^\top\mathbb{W}(\mathbb{W}^\top\mathbb{W}/n)^{-2}\mathbb{W}^\top u_i/n$. Then

$$\sum_{i\in[n]}\|w_i\|^2/n = \sum_{i\in[n]}u_i^\top\mathbb{W}(\mathbb{W}^\top\mathbb{W}/n)^{-2}\mathbb{W}^\top u_i/n = \sum_{i\in[n]}\text{tr}\left\{(\mathbb{W}^\top\mathbb{W}/n)^{-1}\mathbb{W}^\top u_i u_i^\top\mathbb{W}(\mathbb{W}^\top\mathbb{W})^{-1}\right\}$$

$$= \text{tr}\left\{(\mathbb{W}^\top\mathbb{W}/n)^{-1}\right\} \to_p \text{tr}\left(\Sigma_W^{-1}\right).$$

Likewise, $\sum_{i\in[n]} w_i w_i^\top/n \to_p \Sigma_W^{-1}$. Similarly,

$$\sum_{i\in[n]}\frac{a_i^2}{n} = \frac{1}{n}\sum_{i\in[n]}\frac{A_i^2}{\left(n^{-1}\sum_{i\in[n]}A_i^2\right)^2} \to_p \Sigma_A^{-1}.$$

For any $C > 0$,

$$\begin{aligned}
\frac{1}{n}\sum_{i\in[n]}\mathbb{E}\left[\|V_i\|^2\mathbb{1}\left(\frac{\|V_i\|}{\sqrt{n}} > C\right)\right] &= \sum_{i\in[n]}\mathbb{E}\left[\frac{\|V_i\|^2}{n}\mathbb{1}\left(\frac{\|V_i\|^2}{n} > C^2\right)\right] \\
&= \sum_{i\in[n]}\mathbb{E}\left[\left\{\frac{a_i^2}{n}\|\epsilon_{M,i}\|^2 + \frac{\|w_i\|^2}{n}\epsilon_{Y,i}^2 + \frac{\|L_{3,i}^*\|^2}{n} + \frac{L_{4,i}^2}{n}\right\}\mathbb{1}\left(\frac{\|V_i\|^2}{n} > C^2\right)\right] \\
&\le \mathbb{E}\left[\left\{\left(\mathbb{A}^\top\mathbb{A}/n\right)^{-1}\|\epsilon_{M,i}\|^2 + \mathrm{tr}\left\{\left(\mathbb{W}^\top\mathbb{W}/n\right)^{-1}\right\}\epsilon_{Y,i}^2 + \|L_{3,i}^*\|^2 + L_{4,i}^2\right\}\right. \\
&\qquad \left.\cdot\mathbb{1}\left(J_n > C^2\right)\right],
\end{aligned}$$

where

$$J_n = \|\epsilon_{M,i}\|^2\left(\mathbb{A}^\top\mathbb{A}/n\right)^{-2}\cdot\max_{i\in[n]}A_i^2/n + \epsilon_{Y,i}^2\cdot\max_{i\in[n]}\frac{\|w_i\|^2}{n} + \frac{\|L_{3,i}^*\|^2}{n} + \frac{L_{4,i}^2}{n}.$$

Since $\max i \in [n] p_{ii} \to 0$ as $n \to \infty$, we have $J_n \to 0$ as $n \to \infty$. Moreover, the finiteness of the fourth moments of the error terms imply

$$\mathbb{E}\left[\left(\mathbb{A}^\top\mathbb{A}/n\right)^{-1}\|\epsilon_{M,i}\|^2 + \mathrm{tr}\left\{\left(\mathbb{W}^\top\mathbb{W}/n\right)^{-1}\right\}\epsilon_{Y,i}^2 + \|L_{3,i}^*\|^2 + L_{4,i}^2\right] < \infty,$$

from which $n^{-1}\sum_{i\in[n]}\mathbb{E}\left[\|V_i\|^2\mathbb{1}\left(\|V_i\|/\sqrt{n} > C\right)\right] \to_p 0$. The variance of V_i, Σ_i, is

$$\begin{pmatrix}
a_i^2\Sigma_{M|A} & a_i\mathbb{E}[\epsilon_{Y,i}\epsilon_{M,i}]w_i^\top & a_i\mathbb{E}[\epsilon_{M,i}(\epsilon_{M,i}\otimes\epsilon_{M,i})^\top] & a_i\mathbb{E}[\epsilon_{Y,i}^2\epsilon_{M,i}] \\
 & w_iw_i^\top\cdot\sigma_Y^2 & w_i\mathbb{E}[\epsilon_{Y,i}(\epsilon_{M,i}\otimes\epsilon_{M,i})^\top] & w_i\mathbb{E}[\epsilon_{Y,i}^3] \\
 & & \mathbb{E}\left[\{\epsilon_{M,i}\otimes\epsilon_{M,i} - \mathrm{vec}(\Sigma_{M|A})\}^2\right] & \mathbb{E}[\epsilon_{Y,i}^2(\epsilon_{M,i}\otimes\epsilon_{M,i})] - \mathrm{vec}(\Sigma_{M|A})\sigma_Y^2 \\
 & & & \mathbb{E}[(\epsilon_{Y,i}^2 - \sigma_Y^2)^2]
\end{pmatrix},$$

where $\mathbb{E}\left[\{\epsilon_{M,i}\otimes\epsilon_{M,i} - \mathrm{vec}(\Sigma_{M|A})\}^2\right]$ is a diagonal matrix with the diagoanl entries being the marginal variances. We have following convergence of $n^{-1}\sum_{i\in[n]}\Sigma_i$ to Σ:

$$\Sigma^* = \begin{pmatrix}
\Sigma_A^{-1}\Sigma_{M|A} & 0 & 0 & 0 \\
 & \Sigma_W^{-1}\sigma_Y^2 & 0 & 0 \\
 & & \mathbb{E}\left[\{\epsilon_{M,i}\otimes\epsilon_{M,i} - \mathrm{vec}(\Sigma_{M|A})\}^2\right] & 0 \\
 & & & \mathbb{E}[(\epsilon_{Y,i}^2 - \sigma_Y^2)^2]
\end{pmatrix}$$

This follows from

- Since $n^{-1}\sum_{i\in[n]} a_i^2 \to_p \Sigma_A^{-1}$, $\sum_{i\in[n]} a_i^2\Sigma_{M|A} \to_p \Sigma_A^{-1}\Sigma_{M|A}$.
- Since $\epsilon_{Y,i} \perp\!\!\!\perp \epsilon_{M,i}$, we have $\mathbb{E}[\epsilon_{Y,i}\epsilon_{M,i}] = \mathbb{E}[\epsilon_{Y,i}]\mathbb{E}[\epsilon_{M,i}] = 0$, hence $a_i\mathbb{E}[\epsilon_{Y,i}\epsilon_{M,i}]w_i^\top = 0$. Likewise, $n^{-1}\sum_{i\in[n]} a_i\mathbb{E}[\epsilon_{Y,i}(\epsilon_{M,i}\otimes\epsilon_{M,i})^\top] = 0$ and $n^{-1}\sum_{i\in[n]} a_i\mathbb{E}[\epsilon_{Y,i}^2\epsilon_{M,i}] = 0$.
- From the independence of the error terms, we have $\mathbb{E}[\epsilon_{Y,i}^2(\epsilon_{M,i}\otimes\epsilon_{M,i})] - \mathrm{vec}(\Sigma_{M|A})\sigma_Y^2 = \mathbb{E}\left[\epsilon_{Y,i}^2\right]\mathbb{E}[\epsilon_{M,i}\otimes\epsilon_{M,i}] - \mathrm{vec}(\Sigma_{M|A})\sigma_Y^2 = 0$.
- Since $n^{-1}\sum_{i\in[n]} A_i \to_p \mathbb{E}[A] = 0$ from the mean zero assumption, we have $n^{-1}\sum_{i\in[n]} a_i = 0$. From this, $n^{-1}\sum_{i\in[n]} a_i\mathbb{E}[\epsilon_{M,i}(\epsilon_{M,i}\otimes\epsilon_{M,i})^\top] \to_p 0$.
- Since $n^{-1}\sum_{i\in[n]} w_iw_i^\top \to_p \Sigma_W^{-1}$, $n^{-1}\sum_{i\in[n]} w_iw_i^\top\sigma_Y^2 \to_p \Sigma_W^{-1}\sigma_Y^2$.
- Since $n^{-1}\sum_{i\in[n]} w_i = n^{-1}\sum_{i\in[n]} u_i^\top\mathbb{W}(\mathbb{W}^\top\mathbb{W}/n)^{-1} = n^{-1}1_n^\top\mathbb{W}(\mathbb{W}^\top\mathbb{W}/n)^{-1} \to_p 0$, we have $n^{-1}\sum_{i\in n} w_i\mathbb{E}[\epsilon_{Y,i}^3] \to_p 0$.

Consequently, by Theorem B.6.1, $n^{-1/2}\sum_{i\in[n]} V_i$ converges to mean zero normal distribution with variance Σ. Let D be the elimination matrix such that $D_r\mathrm{vec}(S) = \mathrm{vech}(S)$ for a symmetric matrix S. Since the components of $L_{3,i}^*$ are a part of $L_{3,i}$, $\sqrt{n}(\tilde{\mathrm{x}} - \mathrm{x}_0)$ converges to a mean zero normal distribution and variance

$$\Sigma = \begin{pmatrix} I_r & 0 & 0 & 0 \\ 0 & I_{r+1} & 0 & 0 \\ 0 & 0 & D_r & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix} \Sigma^* \begin{pmatrix} I_r & 0 & 0 & 0 \\ 0 & I_{r+1} & 0 & 0 \\ 0 & 0 & D_r & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix}^\top$$

for an elimination matrix $D_r \in \mathbb{R}^{r^2\times r^2}$.

B.6.3 Proof of Proposition 3.3.9

In Section B.6.1 we showed that the minimum discrepancy function F_{min} defined by the MLE $\tilde{\mathrm{x}}$ satisfies the four regularity conditions in Shapiro (1986), and $\sqrt{n}(\tilde{\mathrm{x}} - \mathrm{x}_0)$ is approximately normal from Lemma 3.3.8. Also, the function $h(\cdot)$ is twice differentiable which meets the

definition of regularity of θ in Definition 2.3 of Shapiro (1986). Therefore, by Propositon 4.1 of Shapiro (1986),

$$\sqrt{n}\{h(\hat{\theta}) - h(\theta)\} \to N(0, H(H^\top \Sigma H)^\dagger H^\top).$$

As in Section 3.3.2, we construct $U_2' = \Sigma - \Sigma H_2 (H_2^\top \Sigma H_2)^- H_2^\top \Sigma$ and $\Pi_1' = (H_1^\top U_2' H_1)^\dagger$. Then

$$\sqrt{n}\left\{ \begin{pmatrix} \hat{G}\hat{\eta}_1 \\ \hat{G}\hat{\xi}_1 \end{pmatrix} - \begin{pmatrix} G\eta_1 \\ G\xi_1 \end{pmatrix} \right\} \to N\left(0, \begin{pmatrix} G & 0 & \eta_1^\top \otimes I_r \\ 0 & G & \xi_1^\top \otimes I_r \end{pmatrix} \Pi_1' \begin{pmatrix} G & 0 & \eta_1^\top \otimes I_r \\ 0 & G & \xi_1^\top \otimes I_r \end{pmatrix}^\top \right).$$

B.7 Expression for $E[Y(a, M(a'))]$ in reduced subspace

Under the nested potential outcomes framework (Pearl 2001; Robins and Greenland 1992), the natural indirect effect (NIE) would be

$$E[Y(a, M(a))|X = x] - E[Y(a, M(a^*))|X = x]$$

$$= \int E[Y|A = a, M = m, X = x]\{dP(m|A = a, X = x) - dP(m|A = a^*, X = x)\}$$

From this, we have

$$\begin{aligned} E[Y(a, M(a'))|X = x] &= \int E[Y(a, m)|M(a') = m, X = x] dP(M(a') = m|X = x) \\ &= \int E[Y(a, m)|X = x] dP(M = m|A = a', X = x) \\ &= \int E[Y|A = a, M = m, X = x] dP(M = m|A = a', X = x). \end{aligned}$$

With X omitted for simplicity, we show that the mediation effect in $\mathbb{E}[Y(a, M(a'))|X = x]$ depends only on $G^\top M$.

$$E[Y(a, M(a'))] = \int E[Y|A = a, M = m] dP(m|A = a')$$

$$=^{(a)} \int E[Y|A = a, G^\top M = m_1, \Gamma_2^\top M = m_2, \Phi_2^\top M = m_3, D_0^\top M = m_0] dP(m_1, m_2, m_3, m_0|A = a')$$

$$=^{(b)} \iiint E[Y|A = a, G^\top M = m_1, \Phi_2^\top M = m_3] dP(m_1|A = a') dP(m_2|A = a') dP(m_3|A = a')$$

$$
\begin{aligned}
&=^{(c)} \iiint E[Y|A=a, G^\top M = m_1, \Phi_2^\top M = m_3] dP(m_1|A=a') dP(m_3) dP(m_2|A=a') \\
&= \iint E[Y|A=a, G^\top M = m_1, \Phi_2^\top M = m_3] dP(m_1|A=a') dP(m_3) \int dP(m_2|A=a') \\
&= \int \int E[Y|A=a, G^\top M = m_1, \Phi_2^\top M = m_3] dP(m_1|A=a') dP(m_3) \\
&= \int \left[\int E[Y|A=a, G^\top M = m_1, \Phi_2^\top M = m_3] dP(m_3) \right] dP(m_1|A=a') \\
&= \int E[Y|A=a, G^\top M = m_1] dP(m_1|A=a').
\end{aligned}
$$

The first three equalities arise from the following:

(a) The equality follows from that $\text{span}(G) \cup \text{span}(\Gamma_2) \cup \text{span}(\Phi_2) \cup \text{span}(D_0) = \mathbb{R}^r$, hence we are still working on the space that M lives in.

(b) The equality follow from that:

- The outcome Y is not associated with $\Gamma_2^\top M$ given A (Figure 3.1).
- In the envelope model with the normal error terms, the dimension reduction is performed to achieve

$$G^\top M \perp\!\!\!\perp \Gamma_2^\top M \perp\!\!\!\perp \Phi 2^\top M | (A).$$

- $D_0^\top M$ is associated with neither A nor Y.

(c) The outcome A is not associated with $\Phi_2^\top M$ (Figure 3.1).

The rest follows from rearranging.

B.8 Derivation of the objective function

The objective functions are derived from the likelihood functions of M, $\tilde{Y}$, and R. We use the normality assumption on the error terms ϵ_M and ϵ_Y.

B.8.1 Response envelope

We begin from building the objective function for the response envelope model based on the likelihood of M. Let $\Delta_1 = \Phi_2^\top \Omega_2 \Phi_2 + D_0^\top \Omega_0 D_0$. Suppose M has mean zero. Since $\Sigma_{M|A} = G\Omega_G G^\top + \Gamma_2 \Omega_1 \Gamma_2^\top + \Gamma_0 \Delta_1 \Gamma_0^\top$, we have

$$
\begin{aligned}
L_1 = & -\frac{nr}{2}\log(2\pi) - \frac{n}{2}\log\left|G\Omega_G G^\top + \Gamma_2\Omega_1\Gamma_2^\top + \Gamma_0\Delta_1\Gamma_0^\top\right| \\
& - \frac{1}{2}\sum_{i=1}^{n}(M_i - G\eta_1 A_i - \Gamma_2\eta_2 A_i)^\top \left(G\Omega_G G^\top + \Gamma_2\Omega_1\Gamma_2^\top + \Gamma_0\Delta_1\Gamma_0^\top\right)^{-1}(M_i - G\eta_1 A_i - \Gamma_2\eta_2 A_i) \\
= & -\frac{nr}{2}\log(2\pi) - \frac{n}{2}\log|\Omega_G| - \frac{n}{2}\log|\Omega_1| - \frac{n}{2}\log|\Delta_1| \\
& - \frac{1}{2}\sum_{i=1}^{n}(M_i - G\eta_1 A_i - \Gamma_2\eta_2 A_i)^\top \left(G\Omega_G G^\top + \Gamma_2\Omega_1\Gamma_2^\top + \Gamma_0\Delta_1\Gamma_0^\top\right)^{-1}(M_i - G\eta_1 A_i - \Gamma_2\eta_2 A_i) \\
= & -\frac{nr}{2}\log(2\pi) - \frac{n}{2}\log|\Omega_G| - \frac{n}{2}\log|\Omega_1| - \frac{n}{2}\log|\Delta_1| - \frac{n}{2}\cdot\mathrm{tr}(S_{M|A}\Sigma_{M|A}^{-1}).
\end{aligned}
$$

When G and Γ_2 are given, L_1 is maximized when $\Sigma_{M|A}$ and α_0 are $S_{M|A}$ and $\bar{M}$, respectively. Then, L_1 becomes

$$
\begin{aligned}
L_1 = & -\frac{nr}{2}\log(2\pi) \\
& - \frac{n}{2}\log|\Omega_G| - \frac{1}{2}\sum_{i=1}^{n}\left\{G^\top M_i - \eta_1 A_i\right\}^\top \Omega_G^{-1}\left\{G^\top M_i - \eta_1 A_i\right\} \\
& - \frac{n}{2}\log|\Omega_1| - \frac{1}{2}\sum_{i=1}^{n}\left\{\Gamma_2^\top M_i - \eta_2 A_i\right\}^\top \Omega_1^{-1}\left\{\Gamma_2^\top M_i - \eta_2 A_i\right\} \\
& - \frac{n}{2}\log|\Delta_1| - \frac{1}{2}\sum_{i=1}^{n} M_i\Gamma_0\Delta_1^{-1}\Gamma_0^\top M_i \\
= & -\frac{nr}{2}\log(2\pi) - \frac{nr}{2} - \frac{n}{2}\log|G^\top S_{M|A} G| - \frac{n}{2}\log|\Gamma_2^\top S_{M|A}\Gamma_2| - \frac{n}{2}\log|\Gamma_0^\top S_{M|A}\Gamma_0|
\end{aligned}
$$

Define $\Gamma_3 = \begin{pmatrix} G & \Gamma_0 \end{pmatrix}$. From this, We can express $\left|\Gamma_2^T S_{M|A} \Gamma_2\right|$ in terms of G as follows:

$$\begin{aligned}
\left|\Gamma_2^T S_{M|A} \Gamma_2\right| &= \left| \begin{pmatrix} \Gamma_3 & \Gamma_2 \end{pmatrix} \begin{pmatrix} I & \Gamma_3^T S_{M|A} \Gamma_2 \\ 0 & \Gamma_2^T S_{M|A} \Gamma_2 \end{pmatrix} \begin{pmatrix} \Gamma_3^T \\ \Gamma_2^T \end{pmatrix} \right| = |S_{M|A}| \cdot \left|\Gamma_3^T S_{M|A}^{-1} \Gamma_3\right| \\
&= |S_{M|A}| \left| \begin{pmatrix} G^T \\ \Gamma_0^T \end{pmatrix} S_{M|A}^{-1} \begin{pmatrix} G & \Gamma_0 \end{pmatrix} \right| = |S_{M|A}| \left| \begin{pmatrix} G^T S_{M|A}^{-1} G & G^T S_{M|A}^{-1} \Gamma_0 \\ \Gamma_0^T S_{M|A}^{-1} G & \Gamma_0^T S_{M|A}^{-1} \Gamma_0 \end{pmatrix} \right| \\
&= |S_{M|A}| \cdot \left|G^T S_{M|A}^{-1} G\right| \cdot \left|\Gamma_0^T S_{M|A}^{-1} \Gamma_0\right| = |S_{M|A}| \cdot \left|G^T S_{M|A}^{-1} G\right| \cdot \left|\Gamma^T S_{M|A} \Gamma\right|
\end{aligned}$$

Also, $\left|\Gamma_0^T S_{M|A} \Gamma_0\right| = \left|\Gamma^T S_M^{-1} \Gamma\right| |S_M|$. From these, the likelihood becomes

$$L_1 = const. - \frac{n}{2} \log \left|G^T S_{M|A} G\right| - \frac{n}{2} \log \left|G^T S_{M|A}^{-1} G\right| - \frac{n}{2} \log \left|\Gamma^T S_{M|A} \Gamma\right| - \frac{n}{2} \log \left|\Gamma^T S_M^{-1} \Gamma\right|,$$

where $const. = -\frac{nr}{2} \log(2\pi) - \frac{nr}{2} - \frac{n}{2} \log |S_M| - \frac{n}{2} \log \left|S_{M|A}\right|$. Note that $G^T S_{M|A} G = \Omega_G$ and $G^T S_{M|A}^{-1} G = \Omega_G^{-1}$. Therefore,

$$-\frac{n}{2} \log \left|G^T S_{M|A} G\right| - \frac{n}{2} \log \left|G^T S_{M|A}^{-1} G\right| = -\frac{n}{2} \log \left|\tilde{\Omega}_G\right| \left|\tilde{\Omega}_G^{-1}\right| = 0.$$

Therefore, we end up with the objective function for the outcome envelope model:

$$\log \left|\Gamma^T S_{M|A} \Gamma\right| + \log \left|\Gamma^T S_M^{-1} \Gamma\right|.$$

By using that Γ consists of the columns of G and Γ_2, let $\Gamma = \begin{pmatrix} G & \Gamma_2 \end{pmatrix}$ for without loss of generality. This can be retrieved by permuting Γ. Then we have

$$\begin{aligned}
\left|\Gamma^\top S_{M|A} \Gamma\right| &= \left| \begin{pmatrix} G^\top \\ \Gamma_2^\top \end{pmatrix} S_{M|A} \begin{pmatrix} G & \Gamma_2 \end{pmatrix} \right| \\
&= \left| \begin{pmatrix} G^\top S_{M|A} G & G^\top S_{M|A} \Gamma_2 \\ \Gamma_2^\top S_{M|A} G & \Gamma_2^T S_{M|A} \Gamma_2 \end{pmatrix} \right| = \left| \begin{pmatrix} G^\top S_{M|A} G & 0 \\ 0 & \Gamma_2^\top S_{M|A} \Gamma_2 \end{pmatrix} \right| \\
&= \left|G^\top S_{M|A} G\right| \left|\Gamma_2^\top S_{M|A} \Gamma_2\right|.
\end{aligned}$$

Moreover,

$$
\begin{aligned}
\left|\Gamma^\top S_M^{-1}\Gamma\right| &= \left|\begin{pmatrix} G^\top \\ \Gamma_2^\top \end{pmatrix} S_M^{-1} \begin{pmatrix} G & \Gamma_2 \end{pmatrix}\right| \\
&= \left|\begin{pmatrix} G^\top S_M^{-1} G & G^\top S_M^{-1}\Gamma_2 \\ \Gamma_2^\top S_M^{-1} G & \Gamma_2^\top S_M^{-1}\Gamma_2 \end{pmatrix}\right| \\
&= \left|G^\top S_M^{-1} G\right| \left|\Gamma_2^\top S_M^{-1}\Gamma_2 - \Gamma_2^\top S_M^{-1} G (G^\top S_M^{-1} G)^{-1} G^\top S_M^{-1}\Gamma_2\right|. \\
&= \left|G^\top S_M^{-1} G\right| \left|\Gamma_2^\top \{S_M^{-1} - S_M^{-1} G (G^\top S_M^{-1} G)^{-1} G^\top S_M^{-1}\}\Gamma_2\right| \\
&= \left|G^\top S_M^{-1} G\right| \left|\Gamma_2^\top S_M^{*-1}\Gamma_2\right|.
\end{aligned}
\tag{B.1}
$$

Therefore, the objective function with respect to (3.3a) can be constructed as

$$
\log\left|G^\top S_{M|A} G\right| + \log\left|G^\top S_M^{-1} G\right| + \log\left|\Gamma_2^\top S_{M|A}\Gamma_2\right| + \log\left|\Gamma_2^\top S_M^{*-1}\Gamma_2\right|,
$$

where $S_M^{*-1} = S_M^{-1} - S_M^{-1} G (G^\top S_M^{-1} G)^{-1} G^\top S_M^{-1} = (I_r - P_{G(S_M^{-1})})^\top S_M^{-1}$. Note that (B.1) can also be expressed as $|\Gamma_2^\top S_M^{-1}\Gamma_2||G^\top S_M^{-1} G - G^\top S_M^{-1}\Gamma_2(\Gamma_2^\top S_M^{-1}\Gamma_2)^{-1}\Gamma_2^\top S_M^{-1} G|$, so that

$$
\begin{aligned}
\left|\Gamma^\top S_M^{-1}\Gamma\right| &= |\Gamma_2^\top S_M^{-1}\Gamma_2||G^\top\{S_M^{-1} - S_M^{-1}\Gamma_2(\Gamma_2^\top S_M^{-1}\Gamma_2)^{-1}\Gamma_2^\top S_M^{-1}\}G| \\
&= |\Gamma_2^\top S_M^{-1}\Gamma_2||G^\top \check{S}_M^{-1} G|,
\end{aligned}
$$

where $\check{S}_M^{-1} = S_M^{-1} - S_M^{-1}\Gamma_2(\Gamma_2^\top S_M^{-1}\Gamma_2)^{-1}\Gamma_2^\top S_M^{-1} = (I_r - P_{\Gamma_2(S_M^{-1})})^\top S_M^{-1}$. Then the objective function for (3.3a) can be equivalently expressed as

$$
\log\left|G^\top S_{M|A} G\right| + \log\left|G^\top \check{S}_M^{-1} G\right| + \log\left|\Gamma_2^\top S_{M|A}\Gamma_2\right| + \log\left|\Gamma_2^\top S_M^{-1}\Gamma_2\right|.
$$

We show the derivation of the maximum likelihood estimators of η_1 and η_2 based on the estimated G.

Estimation of ancillary parameters: η_1 and η_2

We illustrate the process for obtaining the maximum likelihood estimator for η_1 given G. We suppose the exposure A is centered. MLE for η_2 can be obtained similarly. From that

$$-\frac{1}{2}\sum_{i\in[n]}\{\hat{G}^\top(M_i-\bar{M})-\eta_1 A_i\}^\top\Omega_G^{-1}\{G^\top(M_i-\bar{M})-\eta_1 A_i\},$$

we have

$$\begin{aligned}\ell(\eta_1) &= -\frac{1}{2}\sum_{i\in[n]}\left[(M_i-\bar{M})^\top G\Omega_G^{-1}\eta_1 A_i + A_i\eta_1^\top\Omega_G^{-1}\eta_1 A_i\right]\\ &= \sum_{i\in[n]}\left[(M_i-\bar{M})^\top G\Omega_G^{-1}\eta_1 A_i - \frac{1}{2}A_i\eta_1^\top\Omega_G^{-1}\eta_1\right]\end{aligned}$$

Then

$$\nabla_{\eta_1}\ell(\eta_1) = \sum_{i\in[n]}\left[(M_i-\bar{M})^\top G\Omega_G^{-1}A_i - \frac{1}{2}A_i\cdot 2\eta_1^\top\Omega_G^{-1}\right],$$

from which the optimal η_1 solves for

$$\sum_{i\in[n]}(M_i-\bar{M})^\top GA_i = \sum_{i\in[n]}A_i\eta_1^\top$$

Therefore

$$\hat{\eta}_1 = \left(\sum_{i\in[n]}A_i^2\right)^{-1}\left(\sum_{i\in[n]}A_iG^\top(M_i-\bar{M})\right) = G^\top S_{MA}S_A^{-1}.$$

B.8.2 Predictor envelope

Define $C=(R^\top,\tilde{Y}^\top)^\top$ and $\Delta_0=\Phi_0^\top\Sigma_{M|A}\Phi_0$. Then

$$\begin{aligned}\Sigma_{R\tilde{Y}} &= Cov(R,\beta_0+\xi_1^\top G^\top R+\xi_2^\top\Phi_2^\top R+\epsilon_Y)\\ &= \Sigma_R G\xi_1+\Sigma_R\Phi_2\xi_2\\ &= \Sigma_{M|A}(G\xi_1+\Phi_2\xi_2)\\ &= (G\Omega_G G^\top+\Phi_2\Omega_2\Phi_2^\top+\Phi_0\Delta_0\Phi_0^\top)(G\xi_1+\Phi_2\xi_2)\\ &= G\Omega_G\xi_1+\Phi_2\Omega_2\xi_2.\end{aligned}$$

From this,

$$\Sigma_C = \begin{pmatrix} \Sigma_R & \Sigma_{R\tilde{Y}} \\ \Sigma_{\tilde{Y}R} & \Sigma_{\tilde{Y}} \end{pmatrix} = \begin{pmatrix} G\Omega_G G^\top + \Phi_2\Omega_2\Phi_2^\top + \Phi_0\Delta_0\Phi_0^\top & G\Omega_G\xi_1 + \Phi_2\Omega_2\xi_2 \\ \xi_1^\top\Omega_G G^\top + \xi_2^\top\Phi_2^\top\Omega_2 & \Sigma_{\tilde{Y}} \end{pmatrix}.$$

This can be factorized as

$$\begin{pmatrix} G & \Phi_2 & \Phi_0 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} \Omega_G & 0 & 0 & \Omega_G\xi_1 \\ 0 & \Omega_2 & 0 & \Omega_2\xi_2 \\ 0 & 0 & \Delta_0 & 0 \\ \xi_1^\top\Omega_G & \xi_2^\top\Omega_2 & 0 & \Sigma_{\tilde{Y}} \end{pmatrix} \begin{pmatrix} G^\top & 0 \\ \Phi_2^\top & 0 \\ \Phi_0^\top & 0 \\ 0 & 1 \end{pmatrix} = O\Sigma_{O^\top C}O^\top,$$

where we omit the subscripts for the 0 matrices, and

$$O = \begin{pmatrix} G & \Phi_2 & \Phi_0 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix}$$

and

$$\Sigma_{O^\top C} = \begin{pmatrix} \Omega_G & 0 & 0 & \Omega_G\xi_1 \\ 0 & \Omega_2 & 0 & \Omega_2\xi_2 \\ 0 & 0 & \Delta_0 & 0 \\ \xi_1^\top\Omega_G & \xi_2^\top\Omega_2 & 0 & \Sigma_{\tilde{Y}} \end{pmatrix}.$$

The objective function can be formulated as

$$\begin{aligned} F(S_C, \Sigma_C) &= \log|\Sigma_C| + \mathrm{tr}(S_C\Sigma_C^{-1}) \\ &= \log|O\Sigma_{O^\top C}O^\top| + \mathrm{tr}(O^\top S_C O\Sigma_{O^\top C}^{-1}) \\ &= \log|\Sigma_{O^\top C}| + \mathrm{tr}(S_{O^\top C}\Sigma_{O^\top C}^{-1}). \end{aligned}$$

With G and Φ_2 fixed, the objective function is minimized when Σ_R $(= \Sigma_{M|A})$, $\Sigma_{\tilde{Y}}$, and $\Sigma_{R\tilde{Y}}$ are estimated by corresponding MLEs $S_R = S_{M|A}$, $S_{\tilde{Y}}$, and $S_{R\tilde{Y}}$, respectively. From

the model, the least squares fit for $\beta_1^\top = \Phi\xi$ is $\tilde{\beta}_1^\top = S_R^{-1} S_{R\tilde{Y}}$. Note that

$$\Phi\xi = \begin{pmatrix} G & \Phi_2 \end{pmatrix} \begin{pmatrix} \xi_1 \\ \xi_2 \end{pmatrix} = G\xi_1 + \Phi_2\xi_2.$$

Therefore, $G^T\Phi\xi = G^T(G\xi_1 + \Phi_2\xi_2) = \xi_1$. Then

$$\Omega_G\xi_1 = G^T\Sigma_R G\xi_1 = G^T\Sigma_R G G^T\Phi\xi.$$

In Section B.8.2, we introduce the derivation of MLE for ξ. By replacing $\Phi\xi$ with its MLE fit and using S_R in place of Σ_R, we get

$$\begin{aligned}
\hat{\Omega}_G\xi_1 &= (G^T S_R G) G^T S_R^{-1} S_{R\tilde{Y}} \\
&= G^T(G\hat{\Omega}_G G^\top + \Phi_2\hat{\Omega}_2\Phi_2^\top + \Phi_0\hat{\Delta}_0\Phi_0^\top) G G^\top (G\hat{\Omega}_G^{-1} G^\top + \Phi_2\hat{\Omega}_2^{-1}\Phi_2^\top + \Phi_0\hat{\Delta}_0^{-1}\Phi_0^\top) S_{R\tilde{Y}} \\
&= \hat{\Omega}_G G^\top G\hat{\Omega}_G^{-1} G^\top S_{R\tilde{Y}} = G^\top S_{R\tilde{Y}},
\end{aligned}$$

where $\hat{\Omega}_G = G^\top S_R G$, $\hat{\Omega}_2 = \Phi_2^\top S_R \Phi_2$, and $\hat{\Delta}_0 = \Phi_0^\top S_R \Phi_0$. From this, we replace $\Omega_G\xi_1$ by its MLE $G^\top S_{R\tilde{Y}}$. Similarly, the MLE for $\Phi_2\xi_2$ is $\Phi_2^\top S_{R\tilde{Y}}$. By replacing the covariance matrices by the corresponding MLEs, we can re-express $\Sigma_{O^\top C}$ as

$$S_{O^\top C} = \begin{pmatrix} G^\top S_R G & 0 & 0 & G^\top S_{R\tilde{Y}} \\ 0 & \Phi_2^\top S_R \Phi_2 & 0 & \Phi_2^\top S_{R\tilde{Y}} \\ 0 & 0 & \Phi_0^\top S_R \Phi_0 & 0 \\ S_{\tilde{Y}R} G & S_{\tilde{Y}R}\Phi_2 & 0 & S_{\tilde{Y}} \end{pmatrix}.$$

Then the objective function becomes $\log|S_{O^\top C}|$. The determinant $|S_{O^\top C}|$ can be obtained by expressing $S_{O^\top C}$ as a 2×2 block matrix

$$S_{O^\top C} = \begin{pmatrix} A & B \\ C & D \end{pmatrix},$$

where

$$A = \begin{pmatrix} G^\top S_R G & 0 & 0 \\ 0 & \Phi_2^\top S_R \Phi_2 & 0 \\ 0 & 0 & \Phi_0^\top S_R \Phi_0 \end{pmatrix}, \quad B = \begin{pmatrix} G^T S_{R\tilde{Y}} \\ \Phi_2^\top S_{R\tilde{Y}} \\ 0 \end{pmatrix}. \quad C = B^\top, \quad D = S_{\tilde{Y}}^{-1}.$$

Then

$$\begin{aligned}
|S_{O^\top C}| &= |A - BD^{-1}C| \cdot |D| \\
&= \left| \begin{pmatrix} G^\top S_R G & 0 & 0 \\ 0 & \Phi_2^\top S_R \Phi_2 & 0 \\ 0 & 0 & \Phi_0^\top S_R \Phi_0 \end{pmatrix} - \begin{pmatrix} G^\top S_{R\tilde{Y}} \\ \Phi_2^\top S_{R\tilde{Y}} \\ 0 \end{pmatrix} S_{\tilde{Y}}^{-1} \begin{pmatrix} S_{\tilde{Y}R} G & S_{\tilde{Y}R}\Phi_2 & 0 \end{pmatrix} \right| \cdot |S_{\tilde{Y}}| \\
&= \left| \begin{pmatrix} G^\top S_R G - G^\top S_{R\tilde{Y}} S_{\tilde{Y}}^{-1} S_{\tilde{Y}R} G & -G^\top S_{R\tilde{Y}} S_{\tilde{Y}}^{-1} S_{\tilde{Y}R}\Phi_2 & 0 \\ -\Phi_2^\top S_{R\tilde{Y}} S_{\tilde{Y}}^{-1} S_{\tilde{Y}R} G & \Phi_2^\top S_R \Phi_2 - \Phi_2^\top S_{R\tilde{Y}} S_{\tilde{Y}}^{-1} S_{\tilde{Y}R}\Phi_2 & 0 \\ 0 & 0 & \Phi_0^\top S_R \Phi_0 \end{pmatrix} \right| \cdot |S_{\tilde{Y}}| \\
&= \left| \begin{pmatrix} G^\top S_{R|\tilde{Y}} G & -G^\top S_{R\tilde{Y}} S_{\tilde{Y}}^{-1} S_{\tilde{Y}R}\Phi_2 & 0 \\ -\Phi_2^\top S_{R\tilde{Y}} S_{\tilde{Y}}^{-1} S_{\tilde{Y}R} G & \Phi_2^\top S_{R|\tilde{Y}} \Phi_2 & 0 \\ 0 & 0 & \Phi_0^\top S_R \Phi_0 \end{pmatrix} \right| \cdot |S_{\tilde{Y}}| \\
&= \left| \begin{pmatrix} G^\top S_{R|\tilde{Y}} G & -G^\top S_{R\tilde{Y}} S_{\tilde{Y}}^{-1} S_{\tilde{Y}R}\Phi_2 \\ -\Phi_2^\top S_{R\tilde{Y}} S_{\tilde{Y}}^{-1} S_{\tilde{Y}R} G & \Phi_2^\top S_{R|\tilde{Y}} \Phi_2 \end{pmatrix} \right| \cdot |\Phi_0^\top S_R \Phi_0| \cdot |S_{\tilde{Y}}| \qquad \text{(B.2)} \\
&= \left| G^\top S_{R|\tilde{Y}} G \right| \cdot \left| \Phi_2^\top S_{R|\tilde{Y}} \Phi_2 - \Phi_2^\top S_{R\tilde{Y}} S_{\tilde{Y}}^{-1} S_{\tilde{Y}R} G (G^\top S_{R|\tilde{Y}} G)^{-1} G^\top S_{R\tilde{Y}} S_{\tilde{Y}}^{-1} S_{\tilde{Y}R}\Phi_2 \right| \\
&\quad \times \left| \Phi^\top S_R^{-1} \Phi \right| \cdot |S_R| \cdot |S_{\tilde{Y}}| \\
&= \left| G^\top S_{R|\tilde{Y}} G \right| \cdot \left| \Phi_2^\top S_{R|\tilde{Y}} \Phi_2 - \Phi_2^\top S_{R\tilde{Y}} S_{\tilde{Y}}^{-1} S_{\tilde{Y}R} G (G^\top S_{R|\tilde{Y}} G)^{-1} G^\top S_{R\tilde{Y}} S_{\tilde{Y}}^{-1} S_{\tilde{Y}R}\Phi_2 \right| \\
&\quad \times \left| G^\top S_R^{-1} G \right| \cdot \left| \Phi_2^\top S_R^{-1} \Phi_2 \right| \cdot |S_R| \cdot |S_{\tilde{Y}}|,
\end{aligned}$$

where the last equality comes from

$$\begin{aligned}\left|\Phi^\top S_R^{-1}\Phi\right| &= \left|\begin{pmatrix} G^\top \\ \Phi_2^\top \end{pmatrix} S_R^{-1} \begin{pmatrix} G & \Phi_2 \end{pmatrix}\right| \\ &= \left|\begin{pmatrix} G^\top S_R^{-1}G & G^\top S_R^{-1}\Phi_2 \\ \Phi_2^\top S_R^{-1}G & \Phi_2^\top S_R^{-1}\Phi_2 \end{pmatrix}\right| = \left|\begin{pmatrix} G^\top S_R^{-1}G & 0 \\ 0 & \Phi_2^\top S_R^{-1}\Phi_2 \end{pmatrix}\right| \\ &= \left|G^\top S_R^{-1}G\right| \cdot \left|\Phi_2^\top S_R^{-1}\Phi_2\right|,\end{aligned}$$

since $G^\top S_R^{-1}\Phi_2 = G^\top(G\hat{\Omega}_G^{-1}G^\top + \Phi_2\hat{\Omega}_2^{-1}\Phi_2^\top + \Phi_0\hat{\Delta}_0^{-1}\Phi_0^\top)\Phi_2 = \hat{\Omega}_G^{-1}G^\top\Phi_2 = 0$. We have

$$\Phi_2^\top S_{R\tilde{Y}}S_{\tilde{Y}}^{-1}S_{\tilde{Y}R}G(G^\top S_{R|\tilde{Y}}G)^{-1}G^\top S_{R\tilde{Y}}S_{\tilde{Y}}^{-1}S_{\tilde{Y}R}\Phi_2$$

$$= \Phi_2^\top(S_R - S_R + S_{R\tilde{Y}}S_{\tilde{Y}}^{-1}S_{\tilde{Y}R})G(G^\top S_{R|\tilde{Y}}G)^{-1}G^\top(S_R - S_R + S_{R\tilde{Y}}S_{\tilde{Y}}^{-1}S_{\tilde{Y}R})\Phi_2$$

$$= \Phi_2^\top S_R G(G^\top S_{R|\tilde{Y}}G)^{-1}G^\top S_R\Phi_2 - \Phi_2^\top S_R G(G^\top S_{R|\tilde{Y}}G)^{-1}G^\top S_{R|\tilde{Y}}\Phi_2$$

$$-\Phi_2^\top S_{R|\tilde{Y}}G(G^\top S_{R|\tilde{Y}}G)^{-1}G^\top S_R\Phi_2 + \Phi_2^\top S_{R|\tilde{Y}}G(G^\top S_{R|\tilde{Y}}G)^{-1}G^\top S_{R|\tilde{Y}}\Phi_2$$

$$= \Phi_2^\top S_{R|\tilde{Y}}G(G^\top S_{R|\tilde{Y}}G)^{-1}G^\top S_{R|\tilde{Y}}\Phi_2.$$

Then, $|S_{O^\top C}|$ becomes

$$\left|G^\top S_{R|\tilde{Y}}G\right| \cdot \left|\Phi_2^\top\{S_{R|\tilde{Y}} - S_{R|\tilde{Y}}G(G^\top S_{R|\tilde{Y}}G)^{-1}G^\top S_{R|\tilde{Y}}\}\Phi_2\right| \times \left|G^\top S_R^{-1}G\right| \cdot \left|\Phi_2^\top S_R^{-1}\Phi_2\right| \cdot |S_R| \cdot |S_{\tilde{Y}}|.$$

Therefore, the objective function for G and Φ_2 with respect to (3.3b) can be constructed as

$$\log\left|G^\top S_{R|\tilde{Y}}G\right| + \log\left|\Phi_2^\top S_{R|\tilde{Y}}^* \Phi_2\right| + \log\left|G^\top S_R^{-1}G\right| + \log\left|\Phi_2^\top S_R^{-1}\Phi_2\right|,$$

where $S_{R|\tilde{Y}}^* = S_{R|\tilde{Y}} - S_{R|\tilde{Y}}G(G^\top S_{R|\tilde{Y}}G)^{-1}G^\top S_{R|\tilde{Y}} = (I_r - P_{G(S_{R|\tilde{Y}})})^\top S_{R|\tilde{Y}}$.

From that

$$\left|\begin{pmatrix} G^\top S_{R|\tilde{Y}}G & -G^\top S_{R\tilde{Y}}S_{\tilde{Y}}^{-1}S_{\tilde{Y}R}\Phi_2 \\ -\Phi_2^\top S_{R\tilde{Y}}S_{\tilde{Y}}^{-1}S_{\tilde{Y}R}G & \Phi_2^\top S_{R|\tilde{Y}}\Phi_2 \end{pmatrix}\right|$$

$$= |\Phi_2^\top S_{R|\tilde{Y}}\Phi_2| \cdot |G^\top S_{R|\tilde{Y}}G - G^\top S_{R\tilde{Y}}S_{\tilde{Y}}^{-1}S_{\tilde{Y}R}\Phi_2(\Phi_2^\top S_{R|\tilde{Y}}\Phi_2)^{-1}\Phi_2^\top S_{R\tilde{Y}}S_{\tilde{Y}}^{-1}S_{\tilde{Y}R}G|$$

$$= |\Phi_2^\top S_{R|\tilde{Y}}\Phi_2||G^\top S_{R|\tilde{Y}}G - G^\top(S_R - S_R + S_{R\tilde{Y}}S_{\tilde{Y}}^{-1}S_{\tilde{Y}R})\Phi_2(\Phi_2^\top S_{R|\tilde{Y}}\Phi_2)^{-1}\Phi_2^\top(S_R - S_R + S_{R\tilde{Y}}S_{\tilde{Y}}^{-1}S_{\tilde{Y}R})G|$$

$$= |\Phi_2^\top S_{R|\tilde{Y}} \Phi_2||G^\top S_{R|\tilde{Y}} G - G^\top S_{R|\tilde{Y}} \Phi_2 (\Phi_2^\top S_{R|\tilde{Y}} \Phi_2)^{-1} \Phi_2^\top S_{R|\tilde{Y}} G|$$

$$= |\Phi_2^\top S_{R|\tilde{Y}} \Phi_2||G^\top \{S_{R|\tilde{Y}} - S_{R|\tilde{Y}} \Phi_2 (\Phi_2^\top S_{R|\tilde{Y}} \Phi_2)^{-1} \Phi_2^\top S_{R|\tilde{Y}}\} G|,$$

the equivalent expression for (B.2) is

$$|\Phi_2^\top S_{R|\tilde{Y}} \Phi_2||G^\top \check{S}_{R|\tilde{Y}} G||G^\top S_R^{-1} G||\Phi_2^\top S_R^{-1} \Phi_2||S_R||S_{\tilde{Y}}|,$$

where $\check{S}_{R|\tilde{Y}} = S_{R|\tilde{Y}} - S_{R|\tilde{Y}} \Phi_2 (\Phi_2^\top S_{R|\tilde{Y}} \Phi_2)^{-1} \Phi_2^\top S_{R|\tilde{Y}} = (I_r - P_{\Phi_2(S_{R|\tilde{Y}})})^\top S_{R|\tilde{Y}}$. Then the alternative expression of the objective function with respect to (3.3b) can be constructed as

$$\log |\Phi_2^\top S_{R|\tilde{Y}} \Phi_2| + \log |G^\top \check{S}_{R|\tilde{Y}} G| + \log |G^\top S_R^{-1} G| + \log |\Phi_2^\top S_R^{-1} \Phi_2|.$$

We introduce the derivation of the maximum likelihood estimates of ξ_1 and ξ_2 when G and Φ_2 are given.

Estimation of ancillary parameters: ξ_1 and ξ_2

The probability density function of $(\tilde{Y}, R)$ can be expressed as the product of the two density functions:

$$f_{\tilde{Y}|R}(\tilde{y}|r) = \frac{1}{\sqrt{2\pi}} \sigma_{Y|(M,A)}^{-2} \exp\left\{ -\frac{1}{2} (\tilde{y} - \beta_0 - \xi_1^\top G^\top r - \xi_2^\top \Phi_2^\top r)^2 / \sigma_{Y|(M,A)}^2 \right\};$$

$$f_R(r) = |2\pi \Sigma_{M|A}|^{-1/2} \exp\left\{ -\frac{1}{2} (R - \mu_R)^\top \Sigma_R^{-1} (R - \mu_R) \right\}.$$

Then the log likelihood becomes

$$\begin{aligned} logL = & \log(2\pi\sigma_{Y|(M,A)}^2)^{-n/2} + \log |2\pi \Sigma_{M|A}|^{-n/2} \\ & - \frac{1}{2} \sum_{i \in [n]} \frac{(\tilde{Y}_i - \beta_0 - \xi_1^\top G^\top R_i - \xi_2^\top \Phi_2^\top R_i)^2}{\sigma_{Y|(M,A)}^2} - \frac{1}{2} \sum_{i \in [n]} R_i^\top \Sigma_R^{-1} R_i. \end{aligned}$$

Then ξ_1 is estimated by solving for

$$\nabla_{\xi_1} \ell(\xi_1) = \frac{1}{n} \sum_{i \in [n]} 2(\tilde{Y}_i - \beta_0 - \xi_1^\top G^\top R_i - \xi_2^\top \Phi_2^\top R_i)(G^\top R_i) \sigma_{Y|(M,A)}^{-2} = 0.$$

Since the MLE for β_0 is $\hat{\beta}_0 = n^{-1}\sum_{i\in[n]}\tilde{Y}_i$, the left hand side becomes

$$\frac{1}{n}\sum_{i\in[n]}(\tilde{Y}_i - \hat{\beta}_0)(G^\top R_i) - \frac{1}{n}\sum_{i\in[n]}(\xi_1^\top G^\top R_i)(G^\top R_i) - \frac{1}{n}\sum_{i\in[n]}(\xi_2^\top \Phi_2^\top R_i)(G^\top R_i) = 0.$$

By rearranging,

$$\frac{1}{n}\sum_{i\in[n]}(\xi_1^\top G^\top R_i)(R_i^\top G) = \frac{1}{n}\sum_{i\in[n]}(\tilde{Y}_i - \hat{\beta}_0)(G^\top R_i)$$

From this, the MLE $\hat{\xi}_1$ is

$$\begin{aligned}
\hat{\xi}_1 &= \left(\frac{1}{n}\sum_{i\in[n]} G^\top R_i R_i^\top G\right)^{-1}\left(\frac{1}{n}\sum_{i\in[n]}(G^\top R_i)(\tilde{Y}_i - \hat{\beta}_0)\right) \\
&= (G^\top S_R G)^{-1} G^\top S_{R\tilde{Y}} \\
&= G^\top S_R^{-1} G G^\top S_{R\tilde{Y}} \\
&= G^\top S_R^{-1} S_{R\tilde{Y}}.
\end{aligned}$$

where the last equality comes from that

$$\begin{aligned}
G^\top S_R^{-1} S_{R\tilde{Y}} &= G^\top (G G^\top S_R^{-1} G G^\top + \Phi_2 \Phi_2^\top S_R^{-1} \Phi_2 \Phi_2^\top + \Phi_0 \Phi_0^\top S_R^{-1} \Phi_0 \Phi_0^\top) S_{R\tilde{Y}} \\
&= G^\top G G^\top S_R^{-1} G G^\top S_{R\tilde{Y}} \\
&= G^\top S_R^{-1} G G^\top S_{R\tilde{Y}}
\end{aligned}$$

B.8.3 Derivation of the objective function

By combining this with the objective functions from the response envelope and predictor envelope models, we have

$$\log\left|G^\top S_{M|A} G\right| + \log\left|G^\top S_M^{-1} G\right| + \log\left|G^\top S_{R|\tilde{Y}} G\right| + \log\left|G^\top S_R^{-1} G\right|$$

$$+ \log\left|\Phi_2^\top S_R^{-1} \Phi_2\right| + \log\left|\Phi_2^\top \{S_{R|\tilde{Y}} - S_{R|Y} G (G^\top S_{R|\tilde{Y}} G)^{-1} G^\top S_{R|\tilde{Y}}\} \Phi_2\right|$$

$$+ \log\left|\Gamma_2^\top S_{M|A} \Gamma_2\right| + \log\left|\Gamma_2^\top \{S_M^{-1} - S_M^{-1} G (G^\top S_M^{-1} G)^{-1} G^\top S_M^{-1}\} \Gamma_2\right|$$

$$= \log\left|G^\top S_M^{-1} G\right| + \log\left|G^\top S_{R|\tilde{Y}} G\right|$$

$$+\log\left|\Phi_2^\top S_R^{-1}\Phi_2\right|+\log\left|\Phi_2^\top\{S_{R|\tilde{Y}}-S_{R|\tilde{Y}}G(G^\top S_{R|\tilde{Y}}G)^{-1}G^\top S_{R|\tilde{Y}}\}\Phi_2\right|$$
$$+\log\left|\Gamma_2^\top S_{M|A}\Gamma_2\right|+\log\left|\Gamma_2^\top\{S_M^{-1}-S_M^{-1}G(G^\top S_M^{-1}G)^{-1}G^\top S_M^{-1}\}\Gamma_2\right|,$$

where the equality comes from that

$$\begin{aligned}\left|G^T S_{M|A}G\right|\left|G^T S_R^{-1}G\right| &= \left|G^T S_{M|A}G\right|\left|G^T S_{M|A}^{-1}G\right| \\ &= \left|\hat{\Omega}_G\right|\cdot\left|\hat{\Omega}_G^{-1}\right| = 1.\end{aligned}$$

This is equivalent to minimizing

$$\log|\Phi_2^T S_{R|Y}\Phi_2|+\log|\Phi_2^T S_R^{-1}\Phi_2|+\log|\Gamma_2^T S_{M|A}\Gamma_2|+\log|\Gamma_2^T S_M^{-1}\Gamma_2|$$

$$+\log|G^T\{S_{R|Y}-S_{R|Y}\Phi_2(\Phi_2^T S_{R|Y}\Phi_2)^{-1}\Phi_2^T S_{R|Y}\}G|+\log|G^T\{S_M^{-1}-S_M^{-1}\Gamma_2(\Gamma_2^T S_M^{-1}\Gamma_2)^{-1}\Gamma_2^T S_M^{-1}\}G|.$$

With Φ_2 and Γ_2 fixed, we solve for this second expression, optimize with respect to G. With G fixed, we can solve for the alternative expression, and optimized over Φ_2 and Γ_2, subject to $\Phi_2^T\Gamma_2=0$.

B.8.4 Equivalence to joint likelihood maximization

Here, we provide the proof for Proposition 3.3.3. The objective functions Q_1 and Q_2 were constructed by adding the two different objective functions that correspond to the response envelope and the predictor envelope formulations for (3.3a) and (3.3b), respectively. We show that minimizing Q_1 and Q_2 is equivalent to maximizing the joint likelihood (3.5). Let $f_{M|A}(m|a)$ be the density function of the distribution of M conditional on A with the constant term omitted. First, Q_1 aims at the maximization of

$$\sum_{i\in[n]}\log f_{M|A}(M_i|A_i)=-\frac{n}{2}\log|\Sigma_{M|A}|-\frac{1}{2}\sum_{i\in[n]}(M_i-G\eta_1A_i-\Gamma_2\eta_2A_i)^\top\Sigma_{M|A}^{-1}(M_i-G\eta_1A_i-\Gamma_2\eta_2A_i).$$

Let $f_{(R,\tilde{Y})}$ be the density function of $(R,\tilde{Y})$ with the constant term omitted. Likewise, define f_R $(=f_{M|A})$, $f_{\tilde{Y}|R}$, and $f_{Y|(M,A)}$ as density functions of random variables represented by the

subscripts. Based on the normality of $(R, \tilde{Y})$, Q_2 aims at the maximization of

$$\begin{aligned}\sum_{i\in[n]} \log f_{(R,\tilde{Y})}(R_i, \tilde{Y}_i) &= \sum_{i\in[n]} \Big\{\log f_{\tilde{Y}|R}(\tilde{Y}_i|R_i) + \log f_R(R_i)\Big\} \\ &= \sum_{i\in[n]} \left\{\log f_{Y|(M,A)}(Y_i|M_i, A_i) + \log f_{M|A}(M_i|A_i)\right\} \\ &= -\frac{n}{2}\log\sigma_Y^2 - \frac{1}{2}\sum_{i\in[n]} \frac{(Y_i - \beta_0 - \xi_1^\top G^\top M_i - \xi_2^\top \Phi_2^\top M_i - \beta_2 A_i)^2}{\sigma_Y^2} \\ &\quad - \frac{n}{2}\log|\Sigma_{M|A}| - \frac{1}{2}\sum_{i\in[n]} R_i^\top \Sigma_{M|A}^{-1} R_i.\end{aligned}$$

By combining the two and replacing $\Sigma_{M|A}$ by the corresponding MLE, $S_{M|A}$, the minimization of $Q_1 + Q_2$ maximizes the following with respect to G, Φ_2, Ω_2, η_1, η_2, ξ_1, ξ_2:

$$-\frac{nr}{2}\log 2\pi - \frac{n}{2}\log|S_{M|A}| - \frac{1}{2}\sum_{i\in[n]}(M_i - G\eta_1 A_i - \Gamma_2\eta_2 A_i)^\top S_{M|A}^{-1}(M_i - G\eta_1 A_i - \Gamma_2\eta_2 A_i)$$

$$-\frac{n}{2}\log 2\pi - \frac{n}{2}\log\sigma_Y^2 - \frac{1}{2}\sum_{i\in[n]} \frac{(Y_i - \beta_0 - \xi_1^\top G^\top M_i - \xi_2^\top \Phi_2^\top M_i - \beta_2 A_i)^2}{\sigma_Y^2} - \frac{n}{2}\log|S_{M|A}| - \frac{1}{2}\sum_{i\in[n]} R_i^\top S_{M|A}^{-1} R_i.$$

When $R_i = M_i - \tilde{\alpha}_1 A_i$, where $\tilde{\alpha}_1 = S_A^{-1} S_{AM}$, we have

$$\frac{1}{2}\sum_{i\in[n]} R_i^\top S_{M|A}^{-1} R_i = \frac{n}{2}\mathrm{tr}\left(\frac{1}{n}\mathbb{R}_n^\top \mathbb{R}_n S_{M|A}^{-1}\right) = \frac{nr}{2},$$

where $\mathbb{R}_n$ is the $n \times r$ design matrix whose i-th row is $R_i^\top$. Also, $\log|S_{M|A}|$ is a constant that does not change to any envelope model estimate. Therefore the estimates $\hat{G}$, $\hat{\Phi}_2$, $\hat{\Omega}_2$, $\hat{\eta}_1$, $\hat{\eta}_2$, $\hat{\xi}_1$, $\hat{\xi}_2$ also maximize

$$-\frac{nr}{2}\log 2\pi - \frac{n}{2}\log|S_{M|A}| - \frac{1}{2}\sum_{i\in[n]}(M_i - G\eta_1 A_i - \Gamma_2\eta_2 A_i)^\top S_{M|A}^{-1}(M_i - G\eta_1 A_i - \Gamma_2\eta_2 A_i)$$

$$-\frac{n}{2}\log 2\pi - \frac{n}{2}\log\sigma_Y^2 - \frac{1}{2}\sum_{i\in[n]} \frac{(Y_i - \beta_0 - \xi_1^\top G^\top M_i - \xi_2^\top \Phi_2^\top M_i - \beta_2 A_i)^2}{\sigma_Y^2}.$$

B.8.5 Total number of parameters

From the above formulation, we can derive the number of parameters to be estimated is

$$\begin{aligned} N_u &= r + u + d_2 + 1 + u + q_2 + 1 + 1 \\ &\quad + u(r-u) + d_2(r-d_2) + q_2(r-q_2) - uq_2 - ud_2 - q_2d_2 \\ &\quad + \frac{u(u+1)}{2} + \frac{d_2(d_2+1)}{2} + \frac{q_2(q_2+1)}{2} + \frac{(r-u-d_2-q_2)(r-u-d_2-q_2+1)}{2} \\ &= \frac{r^2}{2} + r + d + q + 3, \end{aligned}$$

where $d = u+d_2$ and $q = u+q_2$. First r parameters come from estimation of α_0; u parameters come from η_1; d_2 parameters come from η_2; 1 parameter comes from β_0; u parameters come from ξ_1; q_2 parameters come from ξ_2; 1 parameter comes from β_2; 1 parameter comes from σ^2; $u(r-u)$ parameters come from estimating G from the Grassmannian manifold $\mathcal{G}(u,r)$; $d_2(r-d_2)$ parameters come from estimating Γ_2 from the Grassmannian manifold $\mathcal{G}(d_2,r)$; $q_2(r-q_2)$ parameters come from estimating Φ_2 from the Grassmannian manifold $\mathcal{G}(q_2,r)$; $-uq_2 - ud_2 - q_2d_2$ from the mutual orthogonality constraints $G^\top\Gamma_2 = 0$, $G^\top\Phi_2 = 0$, and $\Phi_2^\top\Gamma_2 = 0$; $u(u+1)/2$ parameters come from estimating of Ω_G; $d_2(d_2+1)/2$ parameters come from estimating of Ω_1; $q_2(q_2+1)/2$ parameters come from estimating of Ω_2; and $(r-u-d_2-q_2)(r-u-d_2-q_2+1)/2$ parameters come from estimating of Ω_0.

B.9 Additional simulation results

We present additional simulation results from different sample sizes and different approaches for choosing u. Moreover, we present the projection accuracy of the estimated P_G and study the rate of convergence from the simulation results.

B.9.1 Results when $n = 200,\ 500,\ 1000$

We present the simulation results with larger sample sizes. Figure B.1, B.2, and B.3 present the simulation results with $n = 200$, 500, and 1,000, respectively. We can observe that the variability from each method decreases as the sample size increases. The proposed envelope method (PIE) shows outstanding performance.

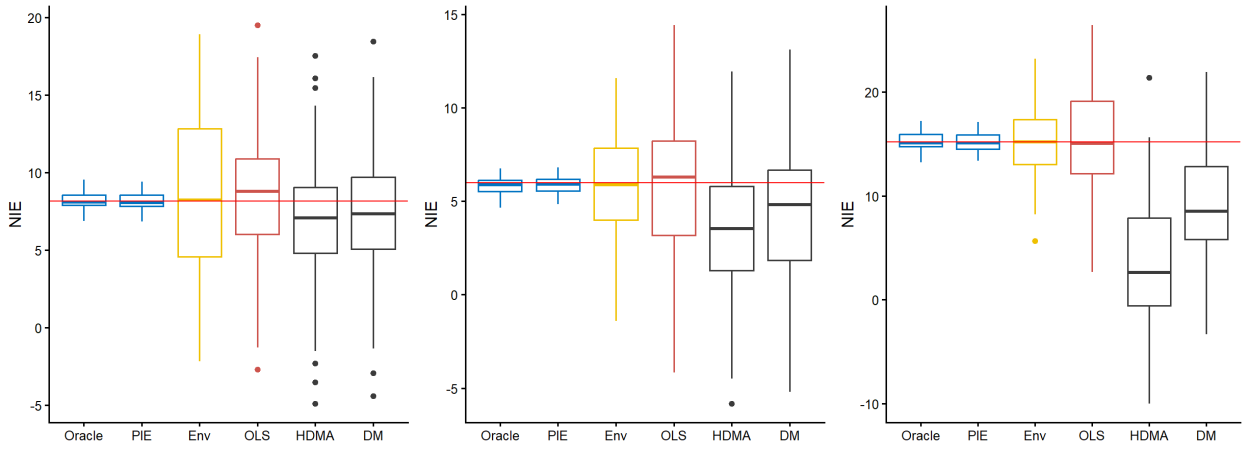

Figure B.1: The distributions of the estimated natural indirect effects when $n = 200$ from different methods. The left plot corresponds to Setting 1; the middle plot corresponds to Setting 2; the right plot corresponds to Setting 3.

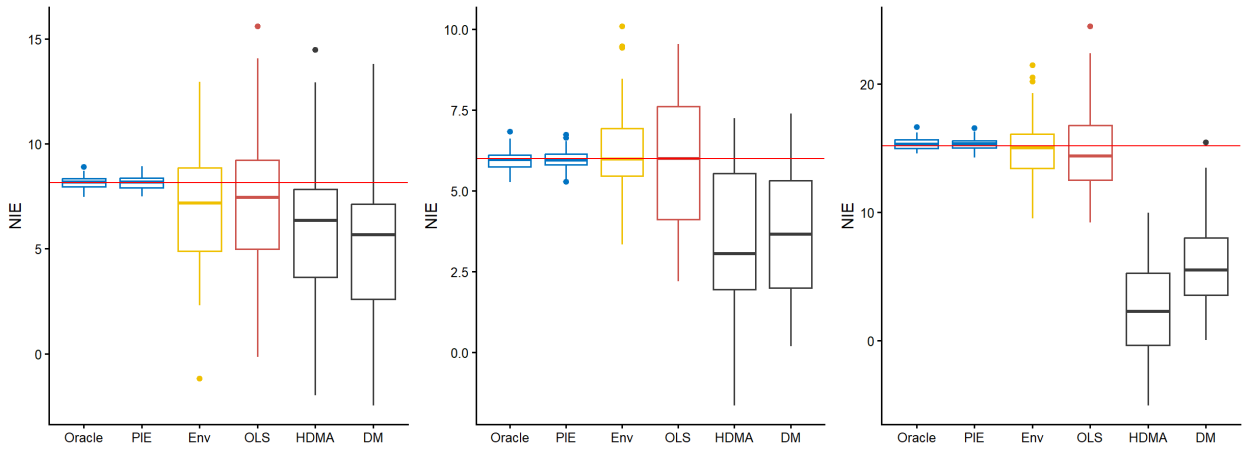

Figure B.2: The distributions of the estimated natural indirect effects when $n = 500$ from different methods. The left plot corresponds to Setting 1; the middle plot corresponds to Setting 2; the right plot corresponds to Setting 3.

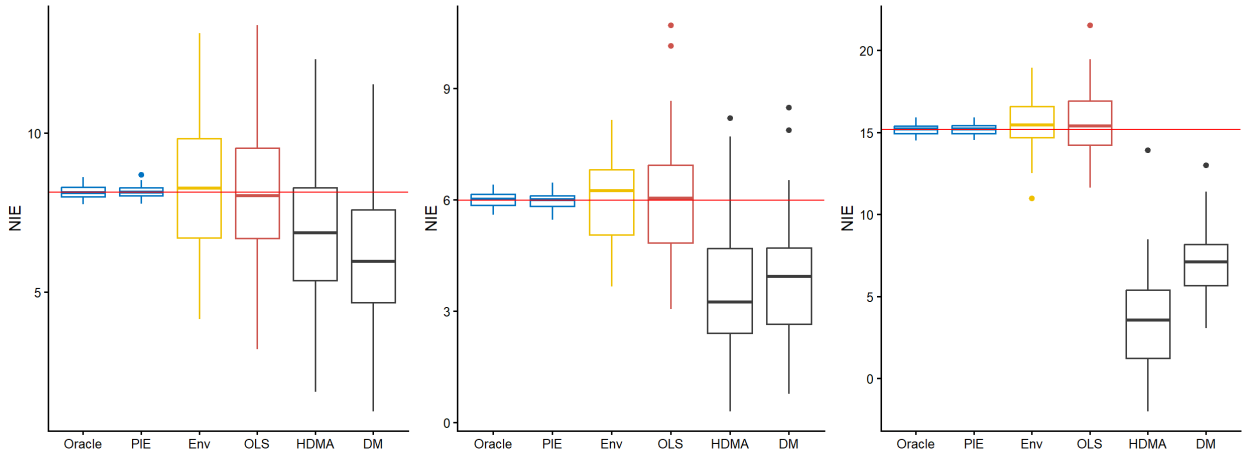

Figure B.3: The distributions of the estimated natural indirect effects from when $n = 1000$ different methods. The left plot corresponds to Setting 1; the middle plot corresponds to Setting 2; the right plot corresponds to Setting 3.

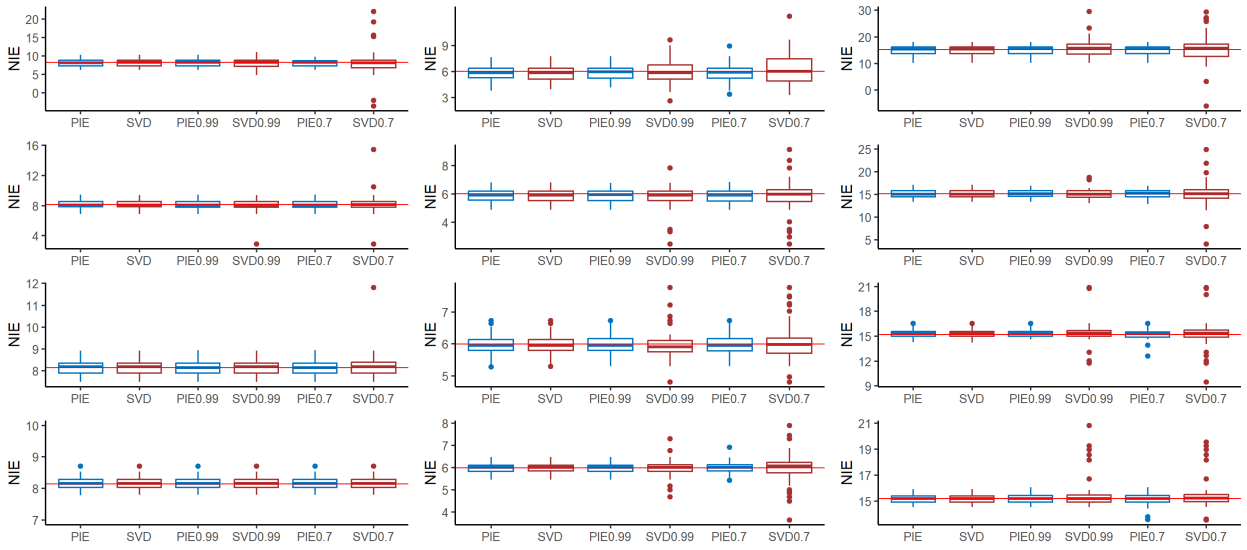

Figure B.4: The distributions of the estimated NIEs obtained from different dimension selection approaches. PIE and SVD represent estimates from $\hat{u}$ that used eigengap ratio criterion. PIE0.99 and SVD0.99 represent the estimates from selecting $\hat{u}$ that correspond to the number of singular values larger than 0.99. PIE0.7 and SVD0.7 represent the estimates from selecting $\hat{u}$ that correspond to the number of singular values larger than 0.7. From the top to the bottom row, the results used $n = 100$, 200, 500, and 1,000.

B.9.2 Hard threshold selection of u

We conducted additional simulation study to analyze the robustness of the proposed envelope methods to the selection of the tuning parameter u. To do so, we considered using hard thresholding approach where the counts of singular values larger than 0.99 and 0.7 are chosen for u. The results are displayed in Figure B.4. We can observe that the SVD method tends to be more sensitive to the dimension u as more outlying NIE estimates are observed from the SVD results – SVD0.7 and SVD0.99 in Figure B.4.

B.9.3 Projection accuracy

In this section, we present the relationship between the projection error of $P_{\hat{G}}$ estimated via the proposed method with the increasing sample size – $n = 200$, 500, 1,000, and 5,000. We conducted the Monte Carlo simulation over 50 iterations. Figure B.5 displays the mean of $\log \|P_{\hat{G}} - P_G\|_F$ at each log of sample size. We can observe that the slopes are near 1/2 which corresponds to the $\sqrt{n}$-rate of decrease.

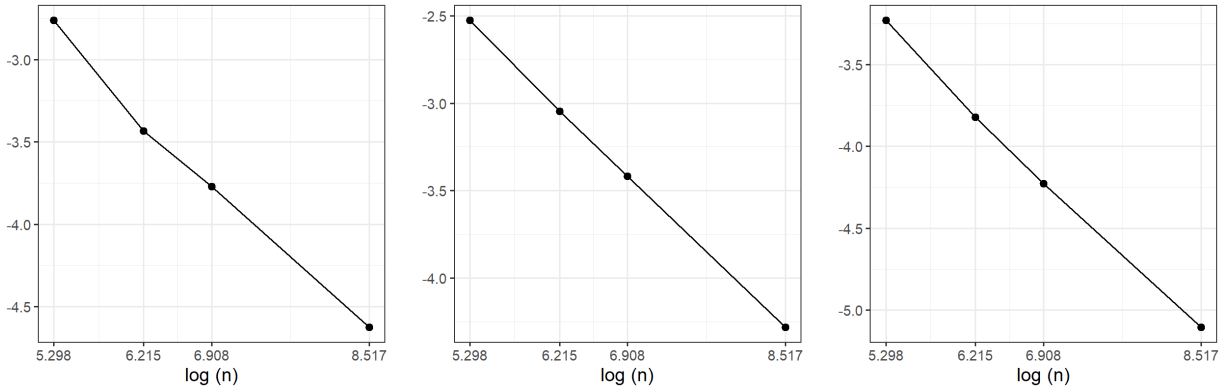

Figure B.5: Projection errors represented by Frobenious norms ($\|P_{\hat{G}} - P_G\|_F$) of the difference between $P_{\hat{G}}$ and P_G. The y-axis represents the mean log Frobenious error for the three simulation settings – Setting 1, 2, and 3 from left to right. The true dimensions of Γ, Φ, and G were used for the simulation.

B.10 Real data analysis

The correlation structure of the proteins are given in Figure B.6. There are a number of proteins that are highly correlated. Figure B.7 shows the sample selection procedure. In the original cohort, there are 5,201 individuals. We use 2,153 to implement the proposed method and perform the data analysis. Figure B.8 shows the singular values of $\hat{\Gamma}^\top\hat{\Phi}$ for choosing the dimensionality u. Figure B.9 shows the distributions of OLS estimates for the outcome-mediator and mediator-exposure models. We can observe that some of the estimates are amplified in the outcome-mediator model, partially attributable to the collinearity.

Based on the estimate $\hat{\Gamma}$ and $\hat{\Phi}$, we performed the singluar value decomposition on $\hat{\Gamma}^\top\hat{\Phi}$ to choose the dimension for the reducing subspace. We choose $u = 3$ based on the number of singular values that are close to 1 (Figure B.8). The estimated NIE remained relatively stable when moderately larger values of u were selected, yielding $\widehat{NIE} = -1.665$ (SE $= 0.406$) and $\widehat{NIE} = -1.509$ (SE $= 0.407$) for $u = 4$ and $u = 5$, respectively. However, for larger values of u, the estimated NIE became more variable and generally more negative. Specifically, for $u = 6, 7, \cdots, 10$, the estimates were $\widehat{NIE} = -2.291$ (SE $= 0.483$), -3.133 (SE $= 0.500$), -3.739(SE $= 0.585$), -2.088 (SE $= 0.513$), and -2.790 (SE $= 0.674$).

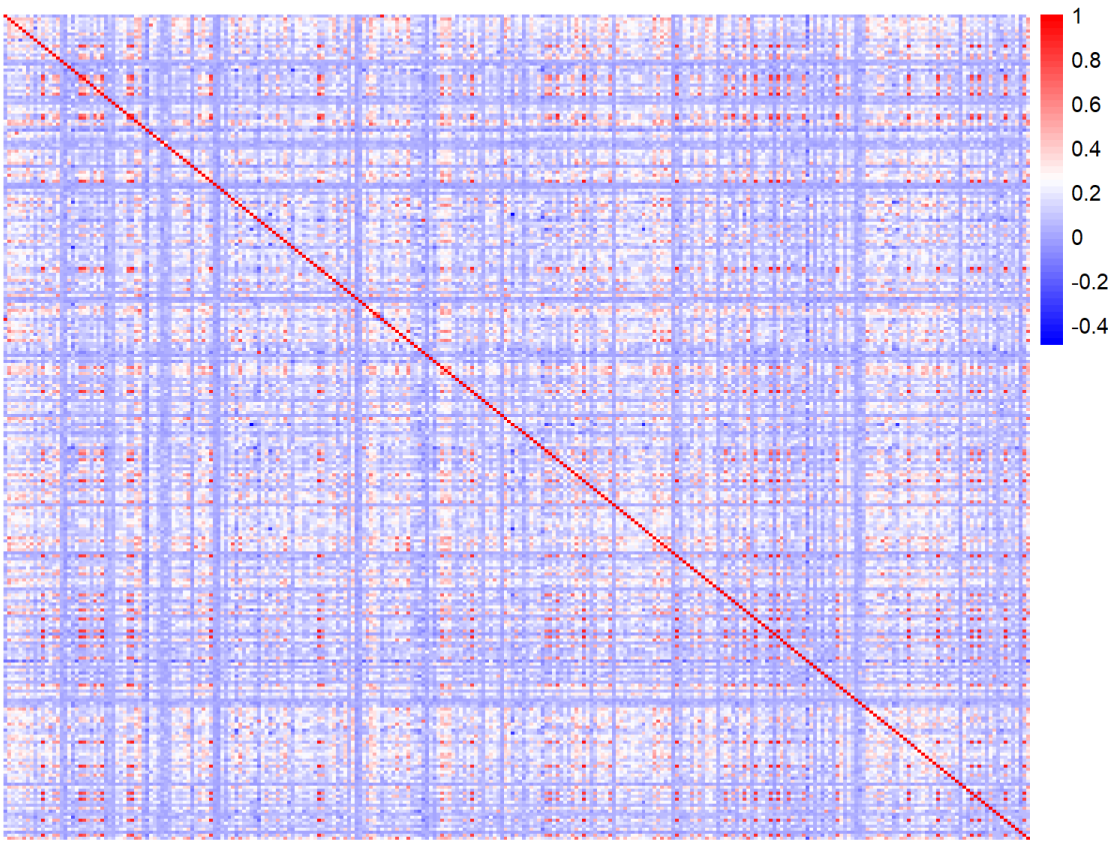

Figure B.6: Heatmap of the correlation structure of the 275 proteins relevant for inflammation and CVD.

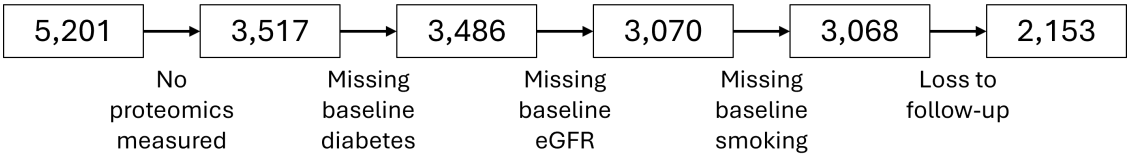

Figure B.7: Flow chart of sample selection

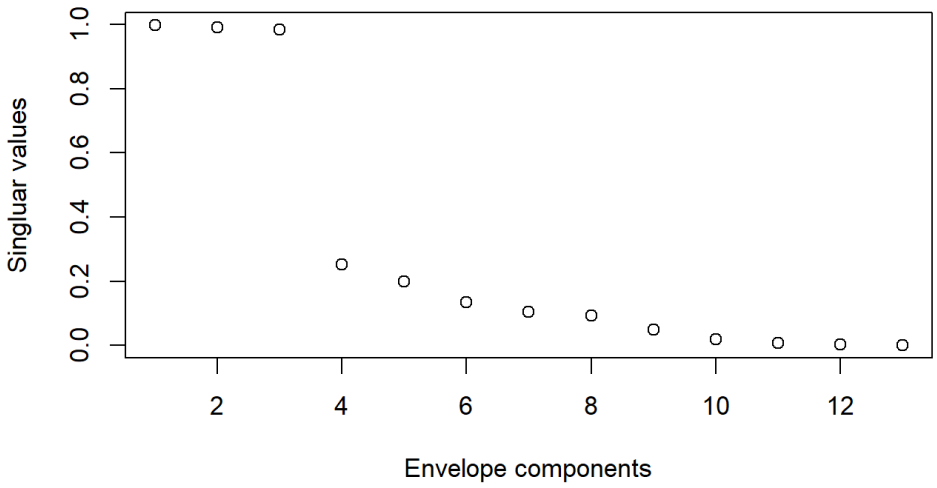

Figure B.8: Singular values for choosing dimensionality of the reduced subspace with the subset of biomarkers. The top three singular values of $\hat{\Gamma}^\top \hat{\Phi}$ are near 1.

Table B.1: UniProt identifiers and corresponding protein names.

UniProt	**Protein**
O00253	Agouti-related protein
Q14005	Interleukin-16
Q07011	TNF receptor superfamily member 9 (4-1BB/CD137)
P02778	C-X-C motif chemokine ligand 10 (IP-10)
P35318	Adrenomedullin
Q496F6	CD300e molecule
P14210	Hepatocyte growth factor
O00182	Galectin-9
Q13478	Interleukin-18 receptor 1
Q9GZV9	Fibroblast growth factor 23
P20333	TNF receptor 2 (TNFR2)

Continued on next page

Table B.1 – continued from previous page

UniProt	Protein
Q03405	Urokinase plasminogen activator receptor (uPAR)
Q9Y6Q6	RANK
P01579	Interferon gamma
P05112	Interleukin-4
P15090	Fatty acid binding protein 4
Q9NSA1	Fibroblast growth factor 21
P07204	Thrombomodulin
P19438	TNF receptor 1 (TNFR1)
P16581	E-selectin
Q96D42	Hepatitis A Virus Cellular Receptor 1 (KIM-1)
Q99988	Growth Differentiation Factor 15
Q9NQ30	Endothelial Cell-Specific Molecule 1 (Endocan)
Q02763	TEK Receptor Tyrosine Kinase (Tie2)
P01137	Transforming Growth Factor Beta 1
Q76LX8	ADAM Metallopeptidase with Thrombospondin Type 1 Motif 13
P35442	Thrombospondin 2
P49763	Placental Growth Factor
P21583	KIT Ligand (Stem Cell Factor)
P01732	CD8 Alpha Chain
P25116	Coagulation Factor II Receptor (PAR1)
P07339	Cathepsin D
P42702	Leukemia Inhibitory Factor Receptor
O43508	TNF Superfamily Member 12 (TWEAK)
Q9NZQ7	Programmed Death-Ligand 1 (PD-L1)
P09237	Matrix Metalloproteinase 7
P33241	Lymphocyte Specific Protein 1
Q16651	Prostasin
Q08334	Interleukin-10 Receptor Beta

Continued on next page

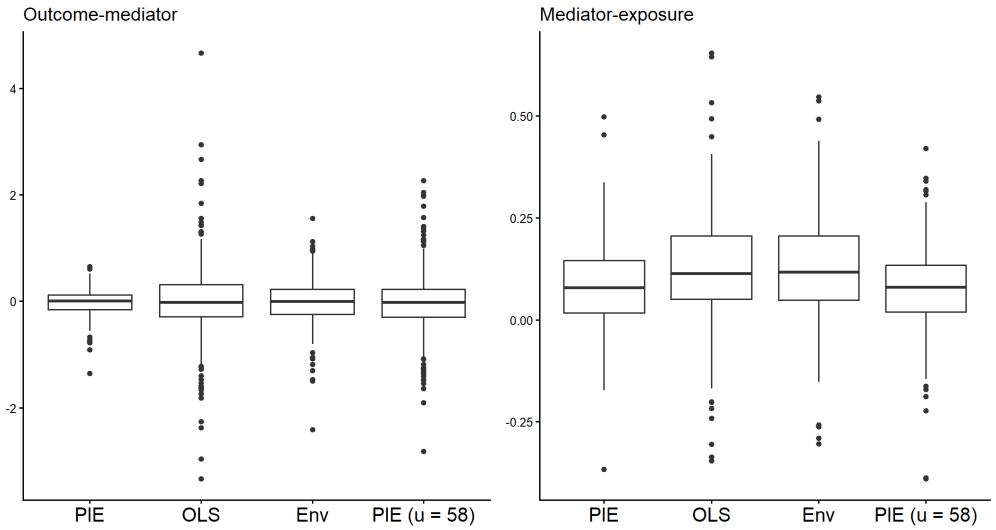

Figure B.9: The distributions of the component-wise estimates for the outcome-mediator (left) and mediator-exposure effects (right). PIE displays results for $u = 3$.

Table B.1 – continued from previous page

UniProt	**Protein**
P09603	Colony Stimulating Factor 1 (M-CSF)
O14558	Heat Shock Protein Beta-6 (HSP20)
P01730	CD4 Molecule
O00220	TRAIL Receptor 1 (DR4)

Appendix C

Chapter 5 Appendix

C.1 Proof of Proposition 4.3.1

For simplicity of the proof, we suppose $\alpha_1 = 1$ and $\alpha_2 = 1$ for simplicity. We prove the result when both $\mathcal{L}_Y$ and $\mathcal{L}_A$ are squared-error losses. This can be generalized to other non-negative losses.

Our target is to minimize

$$\mathbb{E}\left[\{Y - m(X)\}^2 + \{A - \pi(X)\}^2 + \big(\{Y - m(X)\} - \{A - \pi(X)\}\tau(X)\big)^2\right]$$

$$= \mathbb{E}\left[\mathbb{E}[\{Y - m(X)\}^2|X] + \mathbb{E}[\{A - \pi(X)\}^2|X] + \mathbb{E}[(\{Y - m(X)\} - \{A - \pi(X)\}\tau(X))^2\,|X]\right]$$

We can focus on minimizing the conditional expectations:

$$\mathbb{E}[\{Y - m(X)\}^2|X] + \mathbb{E}[\{A - \pi(X)\}^2|X] + \mathbb{E}[(\{Y - m(X)\} - \{A - \pi(X)\}\tau(X))^2\,|X]$$

The first term can be expressed as

$$\mathbb{E}[\{Y - m(X)\}^2|X] = \mathbb{E}[\{Y - \mathbb{E}[Y|X]\}^2|X] + \{\mathbb{E}[Y|X] - m(X)\}^2.$$

Likewise, the second term becomes

$$\mathbb{E}[\{A - \pi(X)\}^2|X] = \mathbb{E}[\{A - \mathbb{E}[A|X]\}^2|X] + \{\mathbb{E}[A|X] - \pi(X)\}^2.$$

Moreover, the third term can be written asd

$$\mathbb{E}\left[\Big(\{Y - m(X)\} - \{A - \pi(X)\}\tau(X)\Big)^2 | X\right]$$

$$= \mathbb{E}\left[\Big(\{Y - \mathbb{E}[Y|X]\} - \{A - \mathbb{E}[A|X]\}\tau(X)\Big)^2 | X\right] + \Big(\mathbb{E}[Y|X] - m(X) - \{\mathbb{E}[A|X] - \pi(X)\}\tau(X)\Big)^2.$$

By combining these results, we have

$$\mathbb{E}[\{Y - \mathbb{E}[Y|X]\}^2|X] + \{\mathbb{E}[Y|X] - m(X)\}^2 + \mathbb{E}[\{A - \mathbb{E}[A|X]\}^2|X] + \{\mathbb{E}[A|X] - \pi(X)\}^2$$

$$+\mathbb{E}\left[\Big(\{Y - \mathbb{E}[Y|X]\} - \{A - \mathbb{E}[A|X]\}\tau(X)\Big)^2 | X\right] + \Big(\mathbb{E}[Y|X] - m(X) - \{\mathbb{E}[A|X] - \pi(X)\}\tau(X)\Big)^2.$$

The terms in red are non-negative, and are minimzed at 0 as long as $m(X) = \mathbb{E}[Y|X]$ and $\pi(X) = \mathbb{E}[A|X]$. Therefore, the objective function is minimized by $\tau(x)$ that minimizes

$$\mathbb{E}\left[\Big(\{Y - \mathbb{E}[Y|X]\} - \{A - \mathbb{E}[A|X]\}\tau(X)\Big)^2 | X\right]$$

from which

$$\tau(X) = \frac{\mathbb{E}[\{Y - \mathbb{E}[Y|X]\} \cdot \{A - \mathbb{E}[A|X]\}|X]}{\mathbb{E}[\{A - \mathbb{E}[A|X]\}^2|X]}$$