Shrinkage Methods for High-Dimensional Regression and Mediation Models

by

Jonathan A. Boss

A dissertation submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
(Biostatistics)
in the University of Michigan
2023

Doctoral Committee:

Professor Jian Kang, Co-Chair
Professor Bhramar Mukherjee, Co-Chair
Professor John Meeker
Associate Professor Sung Kyun Park
Professor Xiang Zhou

Jonathan A. Boss
bossjona@umich.edu
ORCID iD: 0000-0001-5285-3038

ACKNOWLEDGMENTS

I began my time at the University of Michigan in the summer of 2015 as part of the inaugural Big Data Summer Institute (BDSI) cohort. At the time I was an undergraduate student at Michigan State University, laser-focused on completing a master's degree in statistics and starting a career as a master's-level statistician. I will never forget when I walked into the 2015 BDSI welcome dinner and Bhramar was at the entrance personally greeting each student by name. Through that action, I knew that Bhramar was committed to her role as an educator and that commitment showed throughout the duration of the program. As a result of BDSI, I decided to apply to the MS program in the Department of Biostatistics and I entered the program in the Fall of 2016 as both an MS student and a Graduate Student Research Assistant (GSRA) under Bhramar's supervision. It took some time for me to adjust to life as a GSRA, but towards the end of the first year I felt more confident in my ability to be an effective biostatistical collaborator. As the Fall 2017 semester began, Bhramar and I then had the "I think you should do a PhD" conversation, and for the first time since Middle School, I decided to deviate from my terminal MS plan. I entered the PhD program in the Fall of 2018.

The following 5.5 years as a PhD student have been a long, winding journey. There are seemingly incalculable numbers of people who contributed in many different capacities to make this dissertation possible. First, I would like to thank Bhramar and Jian, my two co-advisors. Bhramar and Jian have both invested so much in me and the ideas presented in the dissertation. Without their supervision and guidance over the years, this document would not exist. I would also like to recognize collaborators that made substantial intellectual contributions to the work presented in this dissertation including Amber Cathey, Jyotishka Datta, Kelly Ferguson, Wei Hao, Mike Kleinsasser, John Meeker, Sung Kyun Park, Xing Wang, Barrett Welch, and Xiang Zhou. In addition to my co-authors, I would also like to thank my labmates in both the Mukherjee and Kang labs. Their comradery and conversations enriched my PhD experience and the collaborative environment in both labs fostered environments that were positive for everyone's intellectual growth. I would like to thank Stephen Goutman and Eva Feldman for financially supporting me as a GSRA for my entire time at the University of Michigan and involving me in many interesting projects at the intersection of amyotrophic lateral sclerosis and the environment. There were many graduate students in my cohort that I have positive interactions with and I would particularly like to recog-

nize Emily Roberts, Steven Salerno, and Kim Hirschhorn for being excellent friends and working with me on coursework, publications, and qualifying exam studying.

There are also many people in my personal life that I would like to recognize. First, is my girlfriend Cecilia, who has weathered the highs and lows of my final 2.5 years of dissertation research with me. Words feel inadequate to express how much I love and care for Cecilia and how much Cecilia has supported me, even when things have been extremely difficult and uncertain. I would also like to thank my family, including my Dad, my sister Elizabeth, my niece Annie, Grandma and Grandpa Boss, Grandma and Grandpa McBride, and Lisa for all of the phone calls, encouragement, and positive thinking. I am very fortunate to have all of you in my corner.

I also want to thank all the unsung heroes that contributed to my successes over the last 5.5 years. My high school English teacher Brian Bishop, whose courses and personalized feedback dramatically improved my writing skills. Jeanne Wald, a professor in the Michigan State University Department of Mathematics, for leading the Advanced Track program, which gave me rigorous training in mathematics. Jim Stapleton, a professor emeritus in the Michigan State University Department of Statistics before his passing in 2022, who provided excellent professional guidance and mentorship throughout my undergraduate training and supported my pursuit of advanced study in statistics. My good friend Lauren Lee, whose conversations throughout the COVID-19 pandemic gave me consistent socialization and support, at a time when I was vulnerable to social isolation. My late cat Tiger, who was a great companion throughout the early stages of my PhD, provided me with a lot of joy, and watched Detroit Lions and Michigan State football games with me.

Lastly, I would like to thank my Mom for all the love and support she showed me, even when she was struggling with her health. I know that Mom would be proud of me.

TABLE OF CONTENTS

LIST OF FIGURES

FIGURE

LIST OF TABLES

TABLE

LIST OF APPENDICES

ABSTRACT

With advances in high throughput assaying technology, environmental health scientists are increasingly interested in characterizing the joint effects of a set of exogenous environmental exposures ($\boldsymbol{A}$) coupled with endogenous omics data on a health outcome ($\boldsymbol{Y}$). The omics data is often treated as a high-dimensional mediator ($\boldsymbol{M}$), representing potential intermediary pathways through which $\boldsymbol{A}$ yields $\boldsymbol{Y}$. This allows for a deeper understanding of how exposure mixtures impact health outcomes and what endogenous biological mechanisms underlie mixture effects. In this dissertation, we develop association and mediation models that are specifically tailored to the structure of correlated environmental exposure mixtures and high-dimensional omics data.

In the first project, we focus on the problem of regression coefficient estimation in multi-pollutant models where areas of high collinearity in the exposure space are contained within known covariate groupings called exposure classes. To assuage variance inflation induced by correlated exposures, we propose the group inverse-gamma gamma (GIGG) prior, a heavy-tailed prior that can trade-off between local and group shrinkage in a data-adaptive fashion. Compared to a benchmark shrinkage method like horseshoe regression, GIGG regression reduces mean-squared error by at least 30% across a range of correlation structures and within-group signal densities. We apply GIGG regression to data from the National Health and Nutrition Examination Survey, identifying a toxic effect of metal mixtures on gamma-glutamyl transferase.

For a widely studied environmental exposure, there is likely literature establishing statistical and biological significance of the total exposure effect (TE), defined as the effect of $\boldsymbol{A}$ on $\boldsymbol{Y}$ given a set of confounders $\boldsymbol{C}$. In the second project, we show that leveraging external summary-level information on the TE can improve estimation efficiency of the mediation effects for linear mediation models. Moreover, the efficiency gain depends on the partial R^2 between the $(\boldsymbol{Y} \mid \boldsymbol{M}, \boldsymbol{A}, \boldsymbol{C})$ and $(\boldsymbol{Y} \mid \boldsymbol{A}, \boldsymbol{C})$ models, with smaller (larger) values benefiting direct (indirect) effect estimation. We then robustify our base estimation procedure (Mediation with External Summary Statistic information or MESSI) to incongenial external information. In the highly congenial simulation scenarios, we observe relative efficiency gains for mediation effect estimation of up to 40%. We illustrate our methodology using data from the Puerto Rico Testsite for Exploring Contamination Threats (PROTECT), where Cytochrome p450 lipid metabolites are hypothesized to mediate the effect of phthalate exposure on gestational age at delivery. External information on the TE comes from a

recently published pooled analysis of 16 studies.

The third project considers the problem of estimating mediation effects with respect to exposure mixtures. We develop a method called the mediation mixture map (MedMix), which combines ideas from mediation analysis and latent factor modeling to simultaneously estimate mediation effects corresponding to changes in individual exposures and latent sources of exposure variation. In some simulation settings, MedMix leads to a substantial reduction in root mean-squared error for estimating the mixture mediation effect (approximately 30%) and better quantifies model uncertainty compared to a naïve two-step estimator. We apply MedMix to PROTECT and identify a common source of variation corresponding to mono(carboxynonyl) (MCNP), mono(carboxyoctyl) (MCOP), and mono(3-carboxypropyl) (MCPP) phthalate exposure that is associated with shorter gestational age at delivery (1.13 day decrease per interquartile range increase in the latent mixture; 95% Credible Interval (CI): 0.01, 2.23) and smaller head circumference z-score (0.15 standard deviations smaller head circumference per interquartile range increase in the latent mixture; 95% CI: 0.03, 0.28).

CHAPTER 1

Introduction

1.1 Modeling Exposure Mixtures in Environmental Health

As individuals navigate their daily lives, chemical exposures are omnipresent in their lived environments. Chemical exposure events occur through many different modes, such as breathing polluted air, ingesting contaminated food and water, and physically contacting the skin with contaminated items. Given the ubiquity of chemical exposures in both personal and professional settings, individuals experience a litany of exposure events throughout their life, almost all of which are complex combinations of many individual chemicals. For example, an individual that buys and consumes an agricultural product from the grocery store is simultaneously exposed to pesticides that are used to grow the constituent components of the product, phthalates that are used in manufacturing plastic packaging materials, and polycyclic aromatic hydrocarbons (PAHs) from transporting the agricultural product to the store. This phenomenon is known as exposure to a chemical mixture, where the mixture is comprised of pesticide, phthalate, and PAH exposures. Since individual chemical exposures rarely occur in isolation, conceptual models of exposure are viewed through the lens of exposure mixtures rather than individual chemical exposures (Dominici et al., 2010; Braun et al., 2016; Weisskopf et al., 2018; Bind, 2019).

A fundamental challenge in the field of environmental health is to understand how and why exposure mixtures influence human health and well-being (Wild, 2005). Although environmental health research has traditionally focused on single-exposure models, there has been a substantial push in the last 20 years to model multiple exposures simultaneously, to more accurately reflect how chemical exposures occur in humans (Dominici et al., 2010; Braun et al., 2016; Weisskopf et al., 2018; Bind, 2019). Despite all of the progress that has been made to date, the development of statistical methods that model exposure mixtures remains an important strategic priority for the National Institute for Environmental Health Sciences (NIEHS) (Bind, 2019; Joubert et al., 2022). This dissertation focuses on two broad classes of research questions within the field of multipollutant modeling. The first set of questions is interested in associating exposure mixtures with

health outcomes, while the second set of questions focuses on the endogenous biological mechanisms underlying mixture effects. The goal of this dissertation is to develop several regression and mediation models specially designed to support the increasing number of exogenous exposures and endogenous omics data that are collected in modern environmental health studies.

1.2 Statistical Methods for Analyzing Exposure Mixtures

1.2.1 Existing Literature

When a primary study aim is to identify exposures that are associated with health outcomes of interest, there are many ways to proceed. One proposed analytical framework with roots in the applied environmental health literature is the Exposome-Wide Association Study (EWAS). Despite the substantial structural differences between genetic and chemical exposure data, EWAS applies the analytical framework of Genome-Wide Association Studies (GWAS) to chemical exposure data without making any sizable modifications (Patel et al., 2010). In turn, this has led to a slew of applied analyses that fit outcome-exposure models one-at-a-time, followed by false discovery rate correction on the resultant p-values (Patel et al., 2010; Lind et al., 2013; Hall et al., 2014; McGinnis et al., 2016; Zhuang et al., 2018; Chung et al., 2019). The appeal of this framework, when translated to the field of environmental health, is its simplicity. Many environmental health researchers are familiar with regression techniques for modeling one exposure and the EWAS framework reduces the problem of analyzing multi-exposure data to that of fitting many single exposure models. However, from a statistical perspective, there are several major concerns with GWAS-inspired EWAS as proposed by Patel et al. (2010). One major concern is confounding due to co-exposure, which arises when a proper subset of simultaneously occurring co-exposures are truly associated with the health outcome under study (Braun et al., 2016; Weisskopf et al., 2018). Chemical exposure data is intrinsically correlated and, consequently, strong marginal associations between chemical exposures that are not truly associated with the outcome may disappear after conditioning on a truly associated co-exposure. Therefore, if a truly associated chemical exposure exists, EWAS will generally identify false positives corresponding to truly unassociated co-exposures. Moreover, given the multivariate nature of exposure to chemical mixtures, there are questions regarding the interpretability of effect estimates that are unconditional on other co-exposures (Braun et al., 2016; Weisskopf et al., 2018). Addressing confounding due to co-exposure by jointly modeling chemical exposures is primarily focused on reducing bias and improving the interpretability of effect estimates, however jointly modeling chemical exposures has an additional benefit with regard to variance reduction. Namely, analytical frameworks which model exposures separately will likely experience efficiency losses compared to statistical models which model chemical exposures

jointly, provided that they appropriately account for the exposure correlation structure. Therefore, from a statistical and philosophical perspective, our recommendation is to proceed by modeling chemical exposures jointly, not separately.

When jointly modeling multiple chemical exposures, one of the first considerations is the exposure correlation structure, as there are often collinearity issues present in chemical exposure data. Therefore, linear regression models associating multiple chemical exposures with a health outcome often require penalization of the regression coefficients to reduce the impact of variance inflation. The most common forms of penalization are group lasso, adaptive elastic net, and sparse group lasso (Yuan and Lin, 2006; Zou and Zhang, 2009; Simon et al., 2013). Group lasso and sparse group lasso are typically used when the chemical exposures can be naturally grouped based on chemical structure, toxicological profile, or pharmacokinetics, and adaptive elastic net is used when there is no easily a priori identifiable exposure groupings. There are also penalized regression methods for additive and pairwise interaction models to account for potential nonlinearity in the exposure-response surface (Herring, 2010; Wei et al., 2020; Boss et al., 2021; Bai et al., 2022). An alternative to penalized regression approaches are exposure index regression models, which parameterize the additive contributions of chemical exposures as a weighted sum of discretized chemical exposures (Carrico et al., 2015; Czarnota et al., 2015; Keil et al., 2020). Re-parameterizing the regression coefficients as weights that sum to one naturally facilitates hypothesis testing with respect to the entire exposure index. A widely applied exposure index regression model within this family of methods is weighted quantile sum (WQS) regression, which has a major limitation of enforcing directional homogeneity of the regression coefficients within the weighted exposure index (Carrico et al., 2015; Czarnota et al., 2015). There are extensions of the WQS regression framework that do not make such strong assumptions on the directionality of regression coefficient estimates and allow for pairwise interaction terms in the mean structure, such as quantile g-computation (Keil et al., 2020). However, the lack of a penalty function in existing exposure-index regression model formulations can lead to variance inflation and numerical instability in the presence of highly collinear exposure data (Boss et al., 2018).

While both penalized regression and exposure index methods have been extended to more flexible mean structures through approaches such as hierarchical integrative group lasso (HiGLASSO) (Boss et al., 2021) and Bayesian multiple index models (McGee et al., 2023), there are several other regression approaches that do not fall within these two classes of methods. Bayesian kernel machine regression (BKMR) is a widely applied semiparametric model which associates a unidimensional nonparametric function of the chemical exposures with a health outcome adjusted for confounders (Bobb et al., 2015). The nonparametric function of chemical exposures in BKMR is modeled using a kernel function, however, in practice, the *bkmr* R package specifically implements BKMR with a Gaussian kernel function. Wilson et al. (2022) extended BKMR to a setting

with time-varying exposure to identify windows of susceptibility corresponding to particulate matter exposure throughout pregnancy in association with gestational age standardized birth weight. Tree-based machine learning methods such as classification and regression trees (CART) (Sun et al., 2013) and Bayesian additive regression trees (BART) have been used in the context of risk score construction (Sun et al., 2013; Park et al., 2017) and identification of windows of susceptibility (Mork and Wilson, 2023). Gaussian process regression (Ferrari and Dunson, 2020) and ensemble learning techniques such as SuperLearner (Park et al., 2017) have also been applied to chemical exposure data.

The final category of methods used to analyze chemical mixtures are dimension reduction techniques, which aim to summarize exposure to chemical mixtures using relatively few data adaptively defined variables. The first subclass of methods within the data reduction techniques are clustering methods, which aim to group individuals based on their overall exposure profile (Molitor et al., 2010; Zanobetti et al., 2014; Gibson et al., 2019). The clusters can then be used in place of the chemical exposure measurements in the outcome model. Clustering methods are particularly useful if there are subpopulations within the study population that have distinct exposure profiles, but can be challenging to interpret if no such subpopulations exist. A second subclass of methods are methods based on principal components analysis (PCA), which aims to explain the maximal amount of variability in the exposure space through a reduced set of uncorrelated variables, called principal components (Gibson et al., 2019). Traditional PCA is an unsupervised method, in that it only takes into account the correlation structure of the chemical exposures, however variants of PCA that incorporate information on the outcome (Sun et al., 2013) and quantify rare, extreme exposure events (Gibson et al., 2022) have been applied to chemical exposure data. The resulting principal components can then be used in place of the original chemical exposure data when fitting the outcome model. The final subclass of methods are latent factor models which represent the chemical exposure data as a low-dimensional encoding of common sources of variation (Gibson et al., 2019). The estimated latent factors frequently replace the chemical exposures in the outcome model, however there are fully Bayesian model specifications which combine the outcome and latent factor models into one modeling framework (Ferrari and Dunson, 2021). Dimension reduction techniques such as PCA and latent factor models tend to coincide with blocks of correlated chemical exposures, however there is no guarantee that the principal components or latent factors accurately summarize interpretable exposure events.

If the goal of an analysis is risk prediction and stratification in internal or external study populations, then the environmental risk score (ERS) is a tool that helps address those aims (Park et al., 2014). The ERS is a scalar score that quantifies the risk of an individual experiencing a health outcome given their exposure profile (Park et al., 2014). Provided that predicted outcome values given a set of chemical exposure concentrations are calculable, the finalized outcome model used

to calculate ERSs does not need to have a particular structure (Park et al., 2017; Boss et al., 2018). The workflow for risk stratification via ERS in an internal study population is to split the internal data into training and testing sets, fit the desired risk prediction model on the training set, and calculate the predicted outcome values on the test dataset. The ERSs in the test dataset can then be divided into quantiles and associated with the outcome. An important caveat is that larger values of the ERS do not necessarily coincide with higher overall exposure; ERSs characterize risk, not the magnitude of exposure. ERSs that are equitably transportable across study populations, can subsequently be used by environmental health researchers to assess risk corresponding to exposure profiles observed in their cohorts. Moreover, ERSs can theoretically be used in a clinical capacity to inform personalized screening protocols and preventative treatments for health conditions that an individual is at high risk of developing based on their exposure profile, although this is not currently standard medical practice.

1.2.2 Our Contributions

In Chapter 2, we focus on the problem of jointly regressing many environmental chemicals against a continuous health outcome of interest, when the chemical exposures have an a priori known grouping structure. Moreover, we consider a common setting in chemical mixtures data where collinearity issues in the exposure space are contained within exposure groups. In this scenario, unbiased estimators of model parameters suffer from substantial variance inflation and numerical instability, a problem that is exacerbated as the mean specification becomes more flexible. Moreover, existing penalization techniques such as adaptive elastic net, group lasso, and sparse group lasso, either do not explicitly incorporate information regarding the known grouping structure, do not allow for sparse signals within group, or overregularize large signals (Carvalho et al., 2010; Bhadra et al., 2016). To attain a more optimal bias variance tradeoff for regression coefficient estimation in linear multi-pollutant models, we introduce the group inverse-gamma gamma (GIGG) prior, a heavy-tailed, group shrinkage prior that facilitates data-adaptive multivariate shrinkage within exposure groups. We show the connections between GIGG shrinkage and popular shrinkage priors such as the horseshoe prior and the Laplacian prior and prove posterior consistency and posterior concentration results for regression coefficients in linear models and mean parameters in sparse normal means models. From a computational perspective, the full conditional distributions of all model parameters can be derived in closed-form, leading to straightforward posterior computation. When compared with other Bayesian group regularization priors via simulation, GIGG results in consistently low mean-squared error across a wide range of correlation structures and within-group signal densities. We apply GIGG regression to data from the 2003-2004 National Health and Nutrition Examination Survey (NHANES) cycle for jointly associating metals, ph-

thalates, polybrominated diphenyl ethers (PBDEs), polycyclic aromatic hydrocarbons (PAHs), and organochlorine pesticides with gamma-glutamyl transferase (GGT), a marker of liver functionality.

1.3 Single Exposure Mediation Analysis

1.3.1 Existing Literature

For the second set of questions, this dissertation concerns itself with mediation hypotheses, whereby exogenous exposures impact a health outcome, partially through endogenous pathways that are captured within the omics data. Mediation models in environmental health aim to decompose the effect of exogenous exposures or exogenous exposure mixtures on a health outcome, called the total effect (TE), into two components (Robins and Greenland, 1992; Pearl, 2001; VanderWeele and Vansteelandt, 2014; VanderWeele, 2015). The first component is the indirect effect (IE), which quantifies how much of the TE is exclusively explained through changes in the endogenous biological pathways being studied. The second component is the direct effect (DE), which quantifies how much of the TE is explained through other unmeasured biological mechanisms. An indirect effect that explains a large portion of the TE provides evidence that the endogenous pathways under study partially explain how elevated levels of exposure lead to health outcomes.

Common applied practice for multivariate mediation analyses in environmental health is to first fit individual mediation models for each exposure-mediator pair and then consider several methods specifically designed for high-dimensional mediation analysis with a single exposure or summary measure of exposure (Aung et al., 2020; Blum et al., 2020). Methods that fit individual mediation models for each exposure-mediator pair have seen success in high-dimensional genomics applications (Zeng et al., 2021; Liu et al., 2022), however they are less suitable for environmental health research considering the dimensionality and multivariate flavor of chemical exposure data. For a comprehensive review and comparison of various statistical tests used to evaluate the significance of the IE in a single exposure-mediator pair model see Du et al. (2022). Existing methods for high-dimensional mediation analysis with a single exposure include shrinkage-based methods which directly penalize regression coefficients in the mediation model (Zhang et al., 2016; Gao et al., 2019; Wang et al., 2019; Song et al., 2020; Zhang, 2022), methods that orthogonalize the mediator space so that separate mediation models can be fit on each transformed mediator (Huang and Pan, 2016; Chén et al., 2018; Zhao et al., 2020), mediation models with Gaussian mixture model (GMM) priors (Song et al., 2021a,b), and shrinkage methods for targeted penalization of the IE (Zhou et al., 2020; Song et al., 2021b; Zhao and Luo, 2022). For a comprehensive comparison of methods for high-dimensional mediation analysis with a single exposure see Clark-Boucher et al. (2023).

1.3.2 Our Contributions

In Chapter 3, we consider the problem of single-exposure mediation analysis when summary-level information on the TE is available from an external study or meta-analysis. If researchers are interested in mediation hypotheses, then there is likely existing literature establishing the statistical and biological significance of the TE. Moreover, due to budget and time constraints, studies that collect measures of exogenous exposure and endogenous omics data frequently have sample sizes that are notably smaller than external studies of the TE. Considering these two points, it is natural to leverage externally available information on the TE as a way to improve estimation efficiency of internal mediation model parameters. We show that leveraging external summary-level information on the TE through a constrained optimization improves estimation efficiency of the IE and DE for mediation models with continuous outcomes and mediators. Moreover, the efficiency gain depends on the partial R^2 between the outcome model conditional on both mediators and exposure and the outcome model conditional only on exposure, with smaller values benefiting DE estimation and larger values benefiting IE estimation. We provide a modified version of the constrained estimation procedure that is robust to differences in the TE between external and internal study populations, which we call Mediation with External Summary Statistic information (MESSI). The MESSI framework allows data adaptive shrinkage towards the external information if the TEs in the internal and external populations agree. Our motivating example comes from the Puerto Rico Testsite for Exploring Contamination Threats (PROTECT), a prospective birth cohort study which aims to understand how endocrine-disrupting chemicals negatively impact birth and developmental outcomes. The PROTECT investigators are interested in assessing a mediation hypothesis where elevated phthalate exposure leads to shorter gestational age at delivery through the Cytochrome p450 inflammatory pathway. The internal sample size of the study is approximately 450, however a pooled study of 16 US based cohorts recently published TE estimates corresponding to multiple phthalate metabolites with an approximate sample size of 5000 after omitting PROTECT (Welch et al., 2022).

1.4 Mediation of the Effect of Exposure Mixtures

1.4.1 Existing Literature

A handful of penalization-based methods have been generalized to multi-exposure, high-dimensional mediation problems (Wang et al., 2019; Zhou et al., 2020; Zhang, 2022; Zhao et al., 2022). All of these methods consider a linear mediation model with some form of penalization applied to the regression coefficients or a function of the regression coefficients. In particular, Wang et al. (2019) considered a difference-of-coefficient Laplacian prior, Zhou et al. (2020) pro-

posed a debiased lasso based estimator to estimate and obtain inference for the global IE, Zhang (2022) proposed a linear mixed effects outcome model with an adaptive lasso penalty, and Zhao et al. (2022) considered a multi-exposure extension of the pathway lasso estimator (Zhao and Luo, 2022). If standard unconfoundedness assumptions for causal mediation analysis hold, expressions for the IE, DE, and TE in the multi-exposure linear mediation setting are a function of counterfactual differences in the exposure vector (Zhang, 2022), meaning that these methods can estimate IE, DE, and TEs for a priori specified counterfactual exposure levels. However, the exposure sources underlying chemical exposure data are unknown, making it difficult to ensure that the a priori specified counterfactual exposure levels coincide with interpretable exposure phenomena. To date, we are only aware of two approaches that estimate mediation effects corresponding to functions of chemical exposures (Aung et al., 2020; Devick et al., 2022). Aung et al. (2020) constructed an ERS and treated the ERS as the exposure in high-dimensional mediation models. Devick et al. (2022) devloped a semiparametric mediation model where the exposure is a unidimensional, nonparametric function of individual chemical exposures and effect modifiers. A common theme with Aung et al. (2020) and Devick et al. (2022) is that an outcome or mediation model is used to estimate the exposure mixture. However, this dissertation argues that *exposure mixtures are inherently defined by co-exposure* and thus should explicitly consider the correlation structure of the exposures.

1.4.2 Our Contributions

In Chapter 4, we consider the problem of characterizing mediation effects corresponding to data adaptively estimated exposure mixtures. The primary challenge with quantifying mediation due to exposure mixtures is that information regarding co-exposure is unobserved. We develop a method called the mediation mixture map (MedMix), which combines ideas from mediation analysis and latent factor modeling to simultaneously estimate mediation effects corresponding to changes in individual exposures and latent sources of exposure variation. The primary assumptions underlying MedMix is that highly correlated exposures are more likely to be related in some capacity and that the exposure correlation structure can be appropriately summarized using a latent factor model. In simulation, MedMix more accurately identifies the true number of latent exposure mixtures and better quantifies uncertainty compared to the naive two-step approach of fitting a latent factor model and then subsequently plugging the estimated latent factors into a mediation model. We revisit PROTECT and apply MedMix to multi-exposure data from PROTECT, where individual chemical data on 40 exposures across four different exposures classes, metals, phthalates, phenols and parabens, and PAHs, has been collected. Of primary interest is estimating mediation effects corresponding to the TE of data adaptively estimated exposure mixtures on birth outcomes such as final gestational age and head circumference z-score through the Cytochrome p450 and

Lipoxygenase pathways. In order to maximize available sample size corresponding to the implied TE model from MedMix, we also implement a TE only version of MedMix and show that the TE only version is able to identify an important subset of phthalate metabolites associated with head circumference z-score and gestational age at delivery.

CHAPTER 2

Group Inverse-Gamma Gamma Shrinkage for Sparse Linear Models with Block-Correlated Regressors

2.1 Introduction

Regression with grouped regressors is a common problem in many biomedical applications. Some examples include metabolomics data, where metabolites are grouped by subpathway membership, neuroimaging data, where adjacent voxels are spatially grouped, and environmental exposure data, where exposures are grouped by chemical structure, toxicological profile, and pharmacokinetics (see Figure 2.1). In such cases, leveraging relevant grouping information to construct grouped multivariate shrinkage profiles may help achieve additional variance reduction beyond comparable methods that ignore the grouping structure. The methodological focus of this chapter will be on grouped multivariate regularization in a continuous shrinkage prior framework.

Ever since the publication of the horseshoe prior (Carvalho et al., 2009, 2010), there has been an explosion of continuous shrinkage priors designed for sparse estimation problems, notably normal-gamma shrinkage (Brown and Griffin, 2010), generalized double Pareto shrinkage (Armagan et al., 2013a), Dirichlet–Laplace shrinkage (Bhattacharya et al., 2015), horseshoe+ shrinkage (Bhadra et al., 2017), and normal beta prime (NBP) shrinkage (Bai and Ghosh, 2019; Cadonna et al., 2020), among others. These priors have become increasingly popular for sparse regression problems because of their good theoretical and empirical properties, in addition to their scale mixture representation, which facilitates straightforward and efficient posterior simulation algorithms. The general recipe for constructing a continuous shrinkage prior with good estimation and prediction properties is substantial mass at the origin, to sufficiently shrink null coefficients towards zero, and regularly-varying tails, to avoid overregularizing non-null coefficients (Bhadra et al., 2016). Surveying the continuous shrinkage prior literature on regression with known grouping structure, there are many papers which discuss Bayesian group lasso and its applications (Kyung et al., 2010; Li

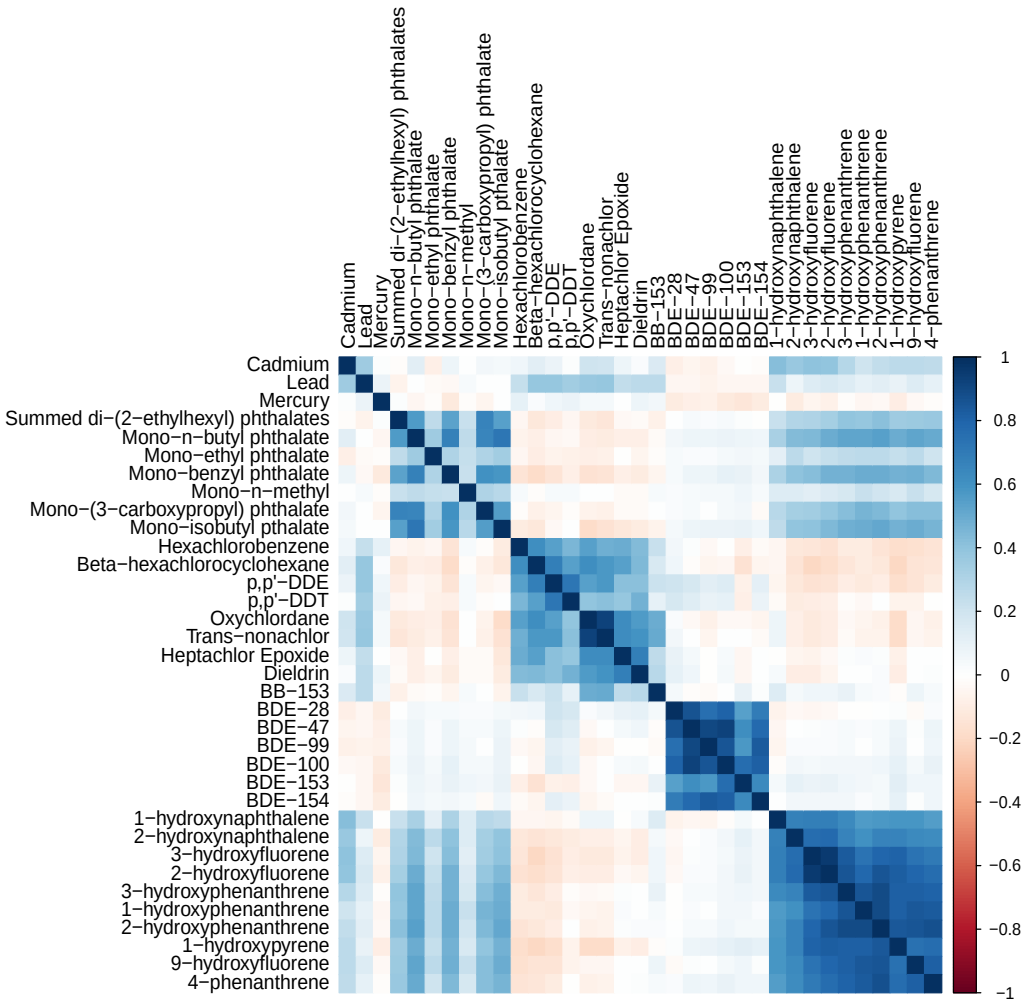


Figure 2.1: Pairwise Spearman correlation plot between metals, phthalates, organochlorine pesticides, polybrominated diphenyl ethers, and polycyclic aromatic hydrocarbons from the 2003-2004 National Health and Nutrition Examination Survey ($n = 990$).

et al., 2015; Xu and Ghosh, 2015; Hefley et al., 2017; Kang et al., 2019) and several papers which propose extensions to Bayesian sparse group lasso (Xu and Ghosh, 2015), Bayesian group bridge regularization (Mallick and Yi, 2017), and the Normal Exponential Gamma prior with grouping structure (Rockova and Lesaffre, 2014). Xu et al. (2016) introduced the, so called, group horseshoe prior with an emphasis on prediction in Bayesian generalized additive models. However, the group horseshoe prior does not reduce to the horseshoe prior for a group of size one, meaning that the group horseshoe prior, as proposed by Xu et al. (2016), is not a direct generalization of the horseshoe prior. Wei et al. (2020) developed a multivariate Dirichlet-Laplace prior for use

in Bayesian additive models with first order interactions. Intuitively, the multivariate Dirichlet-Laplace prior can be thought of as treating the corresponding basis expansion for each regressor as a group. Lastly, although not specifically framed as a grouped multivariate shrinkage prior, Som et al. (2015) proposed a block hyper-g shrinkage prior where the blocks are defined by areas of high collinearity in the regressor space, as in our data example.

Bayesian group lasso-style shrinkage is not generally preferred as a default method for estimation problems, as the Laplacian prior has neither an infinite spike at zero nor regularly-varying tails (Polson and Scott, 2011; Castillo et al., 2015; Bhadra et al., 2016). The multivariate Dirichlet-Laplace prior and the block hyper-g prior, are group/block sparse priors and, therefore, are not designed for problems that require shrinkage at both a group-level and an individual-level. The group horseshoe prior of Xu et al. (2016) has the desired origin and tail behavior marginally, however no hyperparameter in the prior controls how dependent the shrinkage is within a group. Thus, this prior implicitly assumes that the degree of dependence induced by grouped multivariate shrinkage only depends on group size. This assumption is inadequate when we *a priori* believe that, irrespective of group size, some groups have more heterogeneous effect sizes than others. Moreover, this assumption does not avail the opportunity to learn how dependent the shrinkage should be in a data adaptive manner, which is an intrinsic feature in some application areas. For example, in modeling multiple pollutants, this is a relevant consideration as some exposure classes have more homogeneous toxicological profiles than others (Ferguson et al., 2014).

To address these limitations, we propose the group inverse-gamma gamma (GIGG) prior, which extends the horseshoe and normal beta prime (NBP) priors to incorporate grouping structures. The GIGG prior introduces a group level shrinkage parameter, in addition to the usual global and local shrinkage parameters, such that the induced prior on the product of the group and local shrinkage parameters yields the desired marginal shrinkage profile. This allows the user to control the trade-off between group-level and individual-level shrinkage, leading to relatively low estimation error irrespective of the signal density and the degree of multicollinearity within each group. Additionally, the GIGG prior is constructed such that all parameters have closed-form full conditional distributions, implying that techniques to scale horseshoe regression to large sample sizes and high-dimensional regressor spaces are also applicable to GIGG regression (Bhattacharya et al., 2016; Terenin et al., 2019; Johndrow et al., 2020). Theoretically, we establish posterior consistency and posterior concentration results for regression coefficients with grouping structure in linear regression models and mean parameters with grouping structure in sparse normal means models with respect to several GIGG hyperparameters and correlation structures. To our knowledge, we are the first to apply existing theoretical frameworks for posterior consistency in the sparse linear regression model (Armagan et al., 2013b) and posterior concentration in the sparse normal means model (Datta and Ghosh, 2013) to a non-exchangeable prior, which will be useful for future evaluations

of other non-exchangeable priors.

The structure of the chapter is as follows. We start with an intuitive explanation of the GIGG prior in Section 2.2, succeeded by some theoretical results in Section 2.3. After the methodological and theoretical discussion, we outline computational details, including hyperparameter estimation via marginal maximum likelihood estimation (MMLE) (Section 2.4). In Section 2.5, we conduct a simulation study to empirically verify that the intuition and theory developed in Sections 2.2 and 2.3 hold for linear regression models with group-correlated regressors. We then apply GIGG regression to data from the 2003-2004 National Health and Nutrition Examination Survey (NHANES) to identify toxicants and metals associated with a biomarker of liver function (Section 2.6) and conclude with a discussion (Section 2.7).

2.2 Methods

Throughout the chapter, $N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ denotes a multivariate normal distribution with mean parameter $\boldsymbol{\mu}$ and variance-covariance matrix $\boldsymbol{\Sigma}$, $G(a, b)$ denotes a gamma distribution with shape parameter a and rate parameter b, and $IG(a, b)$ denotes an inverse-gamma distribution with shape parameter a and scale parameter b. Additionally, we will use $\pi(\cdot)$ as general notation for a prior probability measure and $\pi(\cdot \mid \boldsymbol{y})$ as general notation for a posterior probability measure.

2.2.1 Group Inverse-Gamma Gamma (GIGG) Prior

Consider a Bayesian sparse linear regression model

$$\begin{gathered}[\boldsymbol{y}|\boldsymbol{\alpha}, \boldsymbol{\beta}, \sigma^2] \sim N\left(\boldsymbol{C}\boldsymbol{\alpha} + \sum_{g=1}^{G} \boldsymbol{X}_g\boldsymbol{\beta}_g, \sigma^2\boldsymbol{I}_n\right) \\ \pi(\boldsymbol{\alpha}) \propto 1, \;\; [\boldsymbol{\beta} \mid \sigma^2] \sim \pi(\boldsymbol{\beta} \mid \sigma^2), \;\; \pi(\sigma^2) \propto \sigma^{-2},\end{gathered} \tag{2.2.1}$$

where $g = 1, \ldots, G$ indexes the groups, $\boldsymbol{y}$ is an $n \times 1$ vector of centered continuous responses, $\boldsymbol{C}$ is a matrix of adjustment covariates, $\boldsymbol{X}_g$ is an $n \times p_g$ matrix of standardized regressors in the g-th group, $\boldsymbol{\beta}_g = (\beta_{g1}, \ldots, \beta_{gp_g})^\top$ is a $p_g \times 1$ vector of regression coefficients corresponding to the g-th group, $\boldsymbol{\beta} = (\boldsymbol{\beta}_1^\top, \ldots, \boldsymbol{\beta}_G^\top)^\top$ is a $p \times 1$ vector of regression coefficients to employ shrinkage on, and $\boldsymbol{I}_n$ is an $n \times n$ identity matrix. We assume the model is sparse in the sense that many of the entries in $\boldsymbol{\beta}$ are zero. The group inverse-gamma gamma (GIGG) prior is defined as

$$[\beta_{gj}|\tau^2, \gamma_g^2, \lambda_{gj}^2] \sim N(0, \tau^2\gamma_g^2\lambda_{gj}^2)$$

$$[\gamma_g^2|a_g] \sim G(a_g, 1), \;\; [\lambda_{gj}^2|b_g] \sim IG(b_g, 1), \;\; [\tau^2, \sigma^2] \sim \pi(\tau^2, \sigma^2),$$

where $j = 1, \ldots, p_g$ indexes the regressors within the g-th group. In this chapter, we assign $\tau \mid \sigma \sim C^+(0, \sigma)$ and $\pi(\sigma^2) \propto \sigma^{-2}$, where $C^+(0, \sigma)$ is a half-Cauchy distribution (Polson and Scott, 2011). Alternatively, we may also express the prior on $\boldsymbol{\beta}$ as a vector, $[\boldsymbol{\beta}|\tau^2, \boldsymbol{\Gamma}, \boldsymbol{\Lambda}] \sim N(0, \tau^2\boldsymbol{\Gamma}\boldsymbol{\Lambda})$, where $\boldsymbol{\Lambda} = \text{diag}(\lambda^2_{11}, ..., \lambda^2_{Gp_G})$ and $\boldsymbol{\Gamma} = \text{diag}(\gamma^2_1, ..., \gamma^2_1, \gamma^2_2, ..., \gamma^2_2, ..., \gamma^2_G, ..., \gamma^2_G)$ such that γ^2_g is repeated p_g times along the diagonal of $\boldsymbol{\Gamma}$. In the GIGG prior specification, the priors on the group shrinkage parameter, γ^2_g, and local shrinkage parameter, λ^2_{gj}, are selected such that the induced prior on the product is a beta prime prior, $\gamma^2_g\lambda^2_{gj} \sim \beta'(a_g, b_g)$ (see Supplementary Material A.1 for distributional definitions). Since the group shrinkage parameter is shared by all p_g observations in the g-th group, assigning a beta prime prior on the product ensures normal beta prime shrinkage marginally while the shrinkage is dependent within-group. One point that deserves further clarification is the assignment of the gamma and inverse-gamma priors to the group and local shrinkage parameters, respectively, when either configuration would yield a beta prime prior in the product. The rationale behind this choice is that the inverse-gamma prior is heavier-tailed than the gamma prior, thereby preventing overregularization of large, non-null coefficients.

Setting $a_g = b_g = 1/2$ for all g yields a special case of the GIGG prior which we will call the group horseshoe prior

$$[\beta_{gj}|\tau^2, \gamma^2_g, \lambda^2_{gj}] \sim N(0, \tau^2\gamma^2_g\lambda^2_{gj})$$

$$\gamma^2_g \sim G(1/2, 1), \;\; \lambda^2_{gj} \sim IG(1/2, 1), \;\; [\tau^2, \sigma^2] \sim \pi(\tau^2, \sigma^2)$$

For a group horseshoe prior with a group of size one, the group shrinkage parameter becomes a local shrinkage parameter. That is, for a group g' of size one,

$$[\beta_{g'1}|\tau^2, \gamma^2_{g'}, \lambda^2_{g'1}] \sim N(0, \tau^2\gamma^2_{g'}\lambda^2_{g'1})$$

$$\gamma^2_{g'} \sim G(1/2, 1), \;\; \lambda^2_{g'1} \sim IG(1/2, 1), \;\; [\tau^2, \sigma^2] \sim \pi(\tau^2, \sigma^2)$$

can be re-indexed as

$$[\beta_{g'1}|\tau^2, \gamma^2_{g'1}, \lambda^2_{g'1}] \sim N(0, \tau^2\gamma^2_{g'1}\lambda^2_{g'1})$$

$$\gamma^2_{g'1} \sim G(1/2, 1), \;\; \lambda^2_{g'1} \sim IG(1/2, 1), \;\; [\tau^2, \sigma^2] \sim \pi(\tau^2, \sigma^2)$$

which is equivalent to the horseshoe prior

$$[\beta_{g'1}|\tau^2, \eta^2_{g'1}] \sim N(0, \tau^2\eta^2_{g'1}), \;\; \eta_{g'1} \sim C^+(0, 1), \;\; [\tau^2, \sigma^2] \sim \pi(\tau^2, \sigma^2).$$

It is important to note that this is different from the group horseshoe prior specification described in Xu et al. (2016). The prior in Xu et al. (2016) assigns independent $\beta'(1/2, 1/2)$ priors on both the

group and local shrinkage parameters, meaning that the implied prior on the product of the group and local shrinkage parameters is the product of two independent $\beta'(1/2, 1/2)$ random variables. In our construction, the product of the group and local shrinkage parameters is itself $\beta'(1/2, 1/2)$. Consequently, our group horseshoe prior specification has horseshoe regularization marginally, while the group horseshoe prior in Xu et al. (2016) does not. To more clearly distinguish between the two group horseshoe priors, we will refer to the prior in Xu et al. (2016) as the group horseshoe+ prior and the GIGG prior with $a_g = b_g = 1/2$ as the group horseshoe prior for the remainder of the chapter.

2.2.2 Marginal Prior Properties

When discussing a proposed shrinkage prior on $\boldsymbol{\beta}$, there are two key features of the marginal prior that need to be investigated. The first is the behavior in a tight neighborhood around zero and the second is the rate at which the prior decays in the extremes. For $\tau^2 = 1$ fixed, Bai and Ghosh (2019) showed that the marginal prior $\pi(\beta_{gj} \mid \tau^2, a_g, b_g)$ has a pole at 0 if and only if $0 < a_g \leq 1/2$, with the pole at zero becoming stronger the closer a_g is to zero. Therefore, one should select $a_g \in (0, 1/2]$ for sparse estimation problems to sufficiently shrink null coefficients towards zero. To clarify the tail behavior we need to introduce the notion of a regularly varying function (Bingham et al., 1989): A positive, measurable function f is said to be regularly varying at ∞ with index $\omega \in \mathbb{R}$ if $\lim_{x\to\infty} f(tx)/f(x) = t^{\omega}$, for all $t > 0$.

Theorem 2.2.1. *Let $\mathcal{B}(a_g, b_g)$ denote the beta function evaluated at a_g and b_g and $\Gamma(b_g + 1/2)$ denote the gamma function evaluated at $b_g + 1/2$. The tails of the marginal prior probability density function of β_{gj} decay at the following rate,*

$$\lim_{\beta_{gj}\to\infty} \frac{\pi(\beta_{gj} \mid \tau^2, a_g, b_g)}{r(\beta_{gj}, \tau^2, a_g, b_g)} = 1$$

$$r(\beta_{gj}, \tau^2, a_g, b_g) = \frac{(2\tau^2)^{b_g}\Gamma(b_g + 1/2)}{\sqrt{\pi}\mathcal{B}(a_g, b_g)} |\beta_{gj}|^{-(1+2b_g)} \left(\frac{\beta_{gj}^2/\tau^2}{1 + \beta_{gj}^2/\tau^2}\right)^{a_g}.$$

Consequently, the index of regular variation is $\omega = -1 - 2b_g$.

Proof. See the Supplementary Material A.2.

The concept of regular variation has been extensively discussed in the context of Bayesian robustness and noninformative inference (Dawid, 1973; O'Hagan, 1979; Andrade and O'Hagan, 2006), with the latter being recently elaborated on in the context of global-local shrinkage priors (Bhadra et al., 2016). When the index $\omega < 0$, regular variation essentially states that the tail of the function decays at a polynomial rate and is therefore considered heavy-tailed. Some examples of

priors with regularly varying tails include the student's t prior and the horseshoe prior. Conversely, commonly used priors such as the normal prior and the Laplace prior do not have regularly-varying tails. As a consequence of having exponentially decaying tails, Bayesian linear regression with independent normal priors and Bayesian lasso are prone to overregularizing large signals and are not flexible enough to facilitate conflict resolution between discordant likelihood and prior information (Andrade and O'Hagan, 2006; Polson and Scott, 2011). Theorem 2.2.1 shows that for any pair of hyperparameters a_g and b_g, the marginal GIGG prior has regularly varying tails and, furthermore, that b_g controls the rate at which the tails decay.

2.2.3 Connection to Bayesian LASSO

As pointed out by a reviewer, an interesting connection between the GIGG prior and Bayesian LASSO-type priors can be seen from integrating out the group shrinkage parameter

$$\pi\big(\boldsymbol{\beta}_g \mid \tau^2, \boldsymbol{\lambda}_g^2, a_g\big) = \int_0^\infty \pi\big(\boldsymbol{\beta}_g \mid \tau^2, \gamma_g^2, \boldsymbol{\lambda}_g^2\big)\pi(\gamma_g^2 \mid a_g) d\gamma_g^2$$

$$= \frac{2}{\Gamma(a_g)(2\pi)^{p_g/2}|\tau^2\boldsymbol{\Lambda}_g|^{1/2}} \left(\sqrt{\frac{1}{2\tau^2}\boldsymbol{\beta}_g\boldsymbol{\Lambda}_g^{-1}\boldsymbol{\beta}_g}\right)^{a_g - p_g/2} K_{a_g - p_g/2}\left(\sqrt{\frac{2}{\tau^2}\boldsymbol{\beta}_g\boldsymbol{\Lambda}_g^{-1}\boldsymbol{\beta}_g}\right),$$

where $\boldsymbol{\Lambda}_g = \text{diag}(\boldsymbol{\lambda}_g^2) = \text{diag}(\lambda_{g1}^2, \ldots, \lambda_{gp_g}^2)$ and $K_\zeta(\cdot)$ denotes the modified Bessel function of the second kind with parameter ζ. If $a_g = 1$, then we see that $[\boldsymbol{\beta}_g \mid \tau^2, \boldsymbol{\lambda}_g^2] \sim \mathcal{ML}(\mathbf{0}, \tau^2\boldsymbol{\Lambda}_g)$, has a multivariate-Laplace prior with location parameter $\mathbf{0}$ and scale parameter $\tau^2\boldsymbol{\Lambda}_g$. Recall that for the multivariate Laplace distribution, a diagonal scale does not correspond to independence. Therefore, when $a_g = 1$, we can interpret the GIGG prior as a mixture of multivariate-Laplace priors with the mixing distribution equal to independent inverse-gamma distributions for each λ_{gj}^2. Moreover, mixing over the local shrinkage parameters implies that the GIGG prior with $a_g = 1$ is a heavy-tailed version of the multivariate-Laplace prior. To connect this result with Bayesian LASSO, we use identity 10.2.17 in Abramowitz and Stegun (1964) and conclude that if $a_g = 1$ and $p_g = 1$, then

$$\pi(\beta_{g1} \mid \tau^2, \lambda_{g1}^2) = \frac{1}{\sqrt{2\tau^2\lambda_{g1}^2}} \exp\left(-\sqrt{\frac{2}{\tau^2\lambda_{g1}^2}}|\beta_{g1}|\right), \;\; \beta_{g1} \in (-\infty, \infty).$$

That is, for a group of size one with $a_g = 1$, the GIGG prior can be interpreted as a mixture of Laplace priors, explicitly connecting the GIGG prior with Bayesian LASSO.

2.2.4 Sparse Normal Means

To further elucidate the shrinkage profile of the GIGG prior, we will focus on a special case of the sparse linear regression model called the sparse normal means model ($\boldsymbol{X} = \boldsymbol{I}_n$ and $\boldsymbol{C}$ empty). In the global-local shrinkage prior literature, it is conventional to work with the sparse normal means problem for analytical tractability, even when the ultimate goal is regression (Rockova and Lesaffre, 2014; Bhattacharya et al., 2015), as the posterior mean has a convenient representation, $E[\beta_{gj} \mid y_{gj}, \tau^2, \sigma^2] = (1 - E[\kappa_{gj} \mid y_{gj}, \tau^2, \sigma^2])y_{gj}$. Here, $\kappa_{gj} = \sigma^2/(\sigma^2 + \tau^2\gamma_g^2\lambda_{gj}^2)$ is called a shrinkage factor, because it quantifies how much the posterior mean is shrunk relative to the maximum likelihood estimator y_{gj}. Calculating the joint prior distribution for the shrinkage factors in the g-th group, $\boldsymbol{\kappa}_g = (\kappa_{g1}, ..., \kappa_{gp_g})^\top$, we have

$\pi\big(\boldsymbol{\kappa}_g \mid \tau^2, \sigma^2, a_g, b_g\big) =$

$$\frac{\Gamma(a_g + p_g b_g)}{\Gamma(a_g)\big(\Gamma(b_g)\big)^{p_g}}\left(\frac{\tau^2}{\sigma^2}\right)^{p_g b_g}\left(1 + \frac{\tau^2}{\sigma^2}\sum_{j=1}^{p_g}\frac{\kappa_{gj}}{1-\kappa_{gj}}\right)^{-(a_g + p_g b_g)}\left(\prod_{j=1}^{p_g}\kappa_{gj}^{b_g - 1}(1-\kappa_{gj})^{-(b_g+1)}\right),$$

where $0 < \kappa_{gj} < 1$ for all $1 \leq j \leq p_g$. Evaluating the prior distribution of $\boldsymbol{\kappa}_g$, we see that the joint density multiplicatively factorizes into "dependent" and "independent" parts where the degree of dependence is governed by the $\sum_{j=1}^{p_g}\kappa_{gj}/(1-\kappa_{gj})$ term. That is, as $a_g + p_g b_g$ goes to zero, the regularization is highly individualistic, whereas if $a_g + p_g b_g$ moves away from zero, then the shrinkage becomes more dependent within the g-th group.

Although the dependence between the shrinkage factors in the g-th group is controlled by $a_g + p_g b_g$, we can use the marginal prior properties to better understand the primary roles of a_g and b_g. From Section 2.2.2, we know that $a_g \in (0, 1/2]$ should be used for sparse estimation problems, because the pole at the origin of the marginal prior on β_{gj} only arises if $a_g \in (0, 1/2]$. Since a_g is heavily restricted in the range of values it can take for sparse estimation problems, then $a_g + p_g b_g$ is primarily determined by the choice of b_g. Even setting the restriction on a_g for sparse estimation problems aside, if we interpret $a_g + p_g b_g$ as a weighted sum of hyperparameters, b_g is given more weight than a_g for groups larger than size one, with the weights becoming increasingly disproportionate as group size increases. Therefore, upon simultaneous inspection of the joint prior on the shrinkage factors and the marginal prior properties for the prior on β_{gj}, b_g offers more control over the dependence of the multivariate shrinkage and a_g offers more control over the strength of the approximate thresholding effect near zero, although these roles are not mutually exclusive. To illustrate this point, Figure 2.2 visualizes the marginal posterior mean of β_{g1} for a group of size two as a function of a_g, b_g, y_{g1}, and y_{g2}. When a_g and b_g are close to zero then the thresholding effect on the marginal posterior mean of β_{g1} hardly depends on the value of y_{g2}, indicating highly individualistic shrinkage. This corroborates our intuition from looking at the

joint posterior distribution of the shrinkage weights within the same group. The second major observation is that as b_g moves away from zero, the marginal posterior mean of β_{g1} becomes increasingly more dependent on the value of y_{g2}. In particular, if we look at the case when $a_g = 0.05$ and $b_g = 2$, we see that when $y_{g2} = 0$ the thresholding effect on β_{g1} is much stronger when compared to $y_{g2} = 10$. The last major observation is that as a_g moves towards zero, the thresholding effect becomes stronger, which coincides with a stronger pole at zero in the marginal prior on β_{g1}.

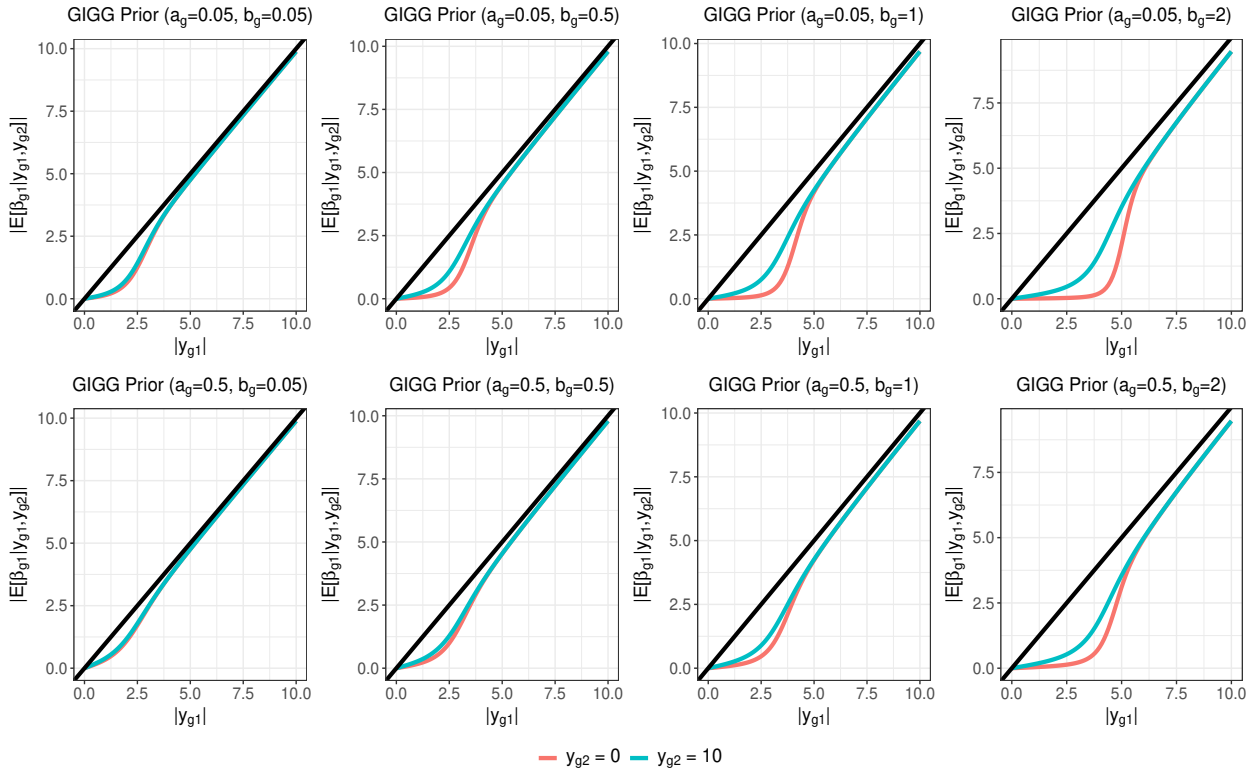


Figure 2.2: Marginal posterior mean of β_{g1} for a group with two observations as a_g, b_g, y_{g1}, and y_{g2} vary. Here, $\tau^2 = 0.2$ and $\sigma^2 = 1$ are fixed.

2.3 Theoretical Properties

In this section, we first prove posterior consistency (Section 2.3.1) and we then consider posterior concentration properties of GIGG shrinkage across a range of different settings (Section 2.3.2).

2.3.1 Posterior Consistency

Let $\boldsymbol{X}_n = [\boldsymbol{X}_1, \ldots, \boldsymbol{X}_{G_n}]$ and $\mathcal{H}_n = \{\boldsymbol{a}, \boldsymbol{b}\}$ denote the collection of hyperparameters where $\boldsymbol{a} = \{a_1, ..., a_{G_n}\}$ and $\boldsymbol{b} = \{b_1, ..., b_{G_n}\}$. Here, the subscript n in G_n refers to the fact that the number of groups in the regressor space is growing as a function of the sample size. Furthermore, let $\mathcal{A}_n = \{(g, j) : \beta^0_{gj} \neq 0\}$ denote the true active set with cardinality $|\mathcal{A}_n|$. Then, Theorem 2.3.1 states that the posterior distribution of $\boldsymbol{\beta}_n$ under the GIGG prior is consistent a posteriori for the true $\boldsymbol{\beta}^0_n$. Similarly, we add a subscript n to $\boldsymbol{\beta}^0_n$ and $\boldsymbol{\beta}_n$ to indicate that the total number of regressors, p_n, is growing as function of sample size. In the theoretical analysis of our method, letting the number of regressors grow as a function of sample size allows us to consider cases where the number of variables included in the model grows with increasing sample size, in addition to cases where the number of variables does not change as a function of sample size (Ghosal, 1999; Armagan et al., 2013b).

Theorem 2.3.1. *Suppose that $p_n = o(n)$, $L_n = \sup_{(g,j)} |\beta^0_{gj}| < \infty$, where β^0_{gj} indicates the true j-th regression coefficient in the g-th group, $0 < \lim_{n\to\infty} \inf \mathcal{H}_n \leq \lim_{n\to\infty} \sup \mathcal{H}_n < \infty$, and $|\mathcal{A}_n| = o(n/\log(n))$. Further, suppose that the smallest and largest singular values of $\boldsymbol{X}_n$, denoted by $\theta_{n,min}(\boldsymbol{X}_n)$ and $\theta_{n,max}(\boldsymbol{X}_n)$, satisfy $0 < \liminf_{n\to\infty} \theta_{n,min}(\boldsymbol{X}_n)/\sqrt{n} \leq \limsup_{n\to\infty} \theta_{n,max}(\boldsymbol{X}_n)/\sqrt{n} < \infty$. Then for any $\epsilon > 0$,*

$$\pi_n(\boldsymbol{\beta}_n : \|\boldsymbol{\beta}_n - \boldsymbol{\beta}^0_n\|_2 < \epsilon \mid \boldsymbol{y}_n, \mathcal{H}_n, \tau^2_n, \sigma^2) \to 1$$

almost surely as $n \to \infty$ provided that $\tau^2_n = C/(p_n n^\rho \log(n))$ for some $\rho, C \in (0, \infty)$.

Proof. See the Supplementary Material A.3.

Of note, the only restriction placed on the values of the hyperparameters in Theorem 2.3.1 is that they do not converge to the boundary of the hyperparameter space as $n \to \infty$.

Remark 2.3.1. *Theorem 2.3.1 is a generalization of Theorem 5 in Armagan et al. (2013b) which proved posterior consistency for the NBP prior when $b_g \in (1, \infty)$. Restricting $b_g \in (1, \infty)$ was done to utilize an argument which required the existence of the second moment of β_{gj}, but does not cover special cases of particular interest such as the horseshoe prior. Therefore, our result extends the existing posterior consistency result from Armagan et al. (2013b) to a more general collection of hyperparameter values with potential grouping structure.*

Remark 2.3.2. *Although Song and Liang (2017) provide an existing theoretical framework for posterior consistency in high-dimensional linear regression when $\log(p_n) = o(n)$, this result cannot be directly applied because the GIGG prior is non-exchangeable.*

2.3.2 Concentration Properties of Shrinkage Parameters

In this subsection, we consider posterior concentration properties corresponding to GIGG shrinkage in different settings, which describe the behavior of the posterior distribution for fixed n. These concentration properties are important to show for new group global-local shrinkage priors, as Datta and Ghosh (2013) showed that such concentration properties were important for the horseshoe prior. We will consider results for general low-dimensional linear regression models when possible, however, for certain componentwise results, we need to focus on the sparse normal means setting. Separate subsection headers are available to distinguish between the results for linear regression models and the results that are only applicable to sparse normal means models.

2.3.2.1 Linear Regression

First, we partially extend the posterior concentration theoretical framework for the sparse normal means model to a low-dimensional linear regression ($p < n$) model with general correlation structure. Going forward, we will drop the subscript n from the notation introduced in the statement of Theorem 2.3.1 to clarify that the subsequent theoretical results hold for fixed p.

Theorem 2.3.2. *Fix $\epsilon \in (0,1)$, p, and n, such that $p < n$. Further, suppose that the smallest and largest singular values of $\boldsymbol{X}^\top\boldsymbol{X}$, denoted by $\theta_{min}(\boldsymbol{X}^\top\boldsymbol{X})$ and $\theta_{max}(\boldsymbol{X}^\top\boldsymbol{X})$, satisfy $0 < \theta_{min}(\boldsymbol{X}^\top\boldsymbol{X}) \leq \theta_{max}(\boldsymbol{X}^\top\boldsymbol{X}) < \infty$. The full conditional posterior mean corresponding to the GIGG prior is,*

$$E[\boldsymbol{\beta} \mid \boldsymbol{y}, \sigma^2, \tau^2, \boldsymbol{\Gamma}, \boldsymbol{\Lambda}] = \left(\boldsymbol{I}_p + (\boldsymbol{X}^\top\boldsymbol{X})^{-1}\frac{\sigma^2}{\tau^2}(\boldsymbol{\Gamma}\boldsymbol{\Lambda})^{-1}\right)^{-1}\hat{\boldsymbol{\beta}}^{OLS}, \;\; \hat{\boldsymbol{\beta}}^{OLS} = (\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{X}^\top\boldsymbol{y}.$$

Then the inequality,

$$\left\|\hat{\boldsymbol{\beta}}^{OLS} - E[\boldsymbol{\beta} \mid \boldsymbol{y}, \sigma^2, \tau^2, \boldsymbol{\Gamma}, \boldsymbol{\Lambda}]\right\|_2 \geq \left(\frac{1}{1 + \theta_{max}(\boldsymbol{X}^\top\boldsymbol{X})\sigma^{-2}\tau^2 \max_{(g,j)} \gamma_g^2\lambda_{gj}^2}\right)\left\|\hat{\boldsymbol{\beta}}^{OLS}\right\|_2,$$

holds and we have the following results:

a)

$$\pi\left(\frac{1}{1 + \theta_{max}(\boldsymbol{X}^\top\boldsymbol{X})\sigma^{-2}\tau^2 \max_{(g,j)} \gamma_g^2\lambda_{gj}^2} \geq \epsilon \;\middle|\; \boldsymbol{y}, \mathcal{H}, \tau^2, \sigma^2\right) \to 1 \;\; \textit{as } \tau^2 \to 0.$$

b)

$$\pi\left(\left\|\hat{\boldsymbol{\beta}}^{OLS} - E[\boldsymbol{\beta} \mid \boldsymbol{y}, \tau^2, \boldsymbol{\Gamma}, \boldsymbol{\Lambda}, \sigma^2]\right\|_2 \geq \epsilon\left\|\hat{\boldsymbol{\beta}}^{OLS}\right\|_2 \;\middle|\; \boldsymbol{y}, \mathcal{H}, \tau^2, \sigma^2\right) \to 1 \;\; \textit{as } \tau^2 \to 0.$$

Proof. See the Supplementary Material A.4.

Theorem 2.3.2 states that, irrespective of the correlation structure, $\tau^2 \to 0$ sufficiently shrinks the posterior mean towards zero. The argument used in the proof of Theorem 2.3.2 can be applied to a litany of other continuous shrinkage priors for which existing posterior concentration results are limited to the sparse normal means model. To supplement these results, we consider the case where we have block diagonal correlation structure, with the blocks defined by the groups, as in Figure 2.1.

Corollary 2.3.1. *Suppose that the regressors in* $\boldsymbol{X}$ *satisfy* $\boldsymbol{X}_g^\top \boldsymbol{X}_{g'} = \boldsymbol{0}$ *for all* $g \neq g'$*, where* $\boldsymbol{0}$ *denotes a* $p_g \times p_{g'}$ *matrix of zeros. If* τ^2*,* σ^2*, and* $a_g \in (0,1)$ *are fixed, then there exists a constant*

$$\epsilon_g(\tau^2, \sigma^2) = \frac{\sigma^2}{\sigma^2 + \theta_{max}(\boldsymbol{X}_g^\top \boldsymbol{X}_g)\tau^2},$$

such that for all $\delta \in (0, \epsilon_g(\tau^2, \sigma^2))$

$$\pi\left(\left\|\hat{\boldsymbol{\beta}}_g^{OLS} - E[\boldsymbol{\beta}_g \mid \boldsymbol{y}, \tau^2, \gamma_g^2, \lambda_{g1}^2, \ldots, \lambda_{gp_g}^2, \sigma^2]\right\|_2 \geq \delta \left\|\hat{\boldsymbol{\beta}}_g^{OLS}\right\|_2 \,\middle|\, \boldsymbol{y}, a_g, b_g, \tau^2, \sigma^2\right) \to 1$$

as $b_g \to \infty$*.*

Proof. See the Supplementary Material A.5.

The conclusion of Corollary 2.3.1 is that if the hyperparameter $b_g \to \infty$ then there is at least some amount of shrinkage relative to the ordinary least squares estimator in the g-th group. If τ^2/σ^2 is close to zero, then $\epsilon(\tau^2, \sigma^2) \approx 1$, implying shrinkage of the posterior mean towards zero. Therefore, we can interpret the case when $b_g \to \infty$ and τ^2/σ^2 close to zero as shrinkage of the entire g-th group towards zero.

2.3.2.2 Sparse Normal Means

Although we would ideally consider additional posterior concentration results within the context of a linear regression model, there is not an analytically tractable analog of componentwise shrinkage factors for a general design matrix without any orthogonality. Therefore, we will proceed by considering posterior concentration results within the sparse normal means framework, to make precise statements regarding componentwise shrinkage, as opposed to shrinkage of the entire L_2-norm.

One question that arises is whether the dependence induced between the β_{gj}'s by γ_g^2 will overly dominate the individual-level shrinkage. As an example, one can conceptualize a case where a group has only one signal, which is overly shrunk by virtue of being grouped with an overwhelming majority of null means. An alternative situation that could occur is a case where few null means

are grouped with many non-null means, leading to insufficient shrinkage of the null means toward zero. These two scenarios are described by Som et al. (2016) as the *Conditional Lindley's Paradox* and *Essentially Least Squares Estimation*, respectively. Theorem 2.3.3a states that if the gl-th observation is sufficiently large then there will be minimal shrinkage on y_{gl}. This guarantees that group shrinkage will not overly dominate individual shrinkage if the observation is large. Conversely, Theorem 2.3.3b states that if the global shrinkage parameter converges to zero, then the GIGG prior will sufficiently shrink the y_{gl}'s toward zero. Let $\boldsymbol{y}_g = (y_{g1}, \ldots, y_{gp_g})^\top$.

Theorem 2.3.3. *Suppose that* $p_g \in \{2, 3, 4, \ldots\}$.

a) Fix $\psi, \delta \in (0, 1)$. *Then there exists a function* $h(p_g, \tau^2, \sigma^2, a_g, b_g, \psi, \delta)$ *such that*

$$\pi(\kappa_{gl} > \psi \mid \boldsymbol{y}_g, \tau^2, \sigma^2, a_g, b_g) \leq \exp\left(-\frac{\psi(1-\delta)}{2\sigma^2} y_{gl}^2 + \frac{\psi\delta}{2\sigma^2} \sum_{j \neq l} y_{gj}^2\right) h(p_g, \tau^2, \sigma^2, a_g, b_g, \psi, \delta).$$

Consequently, if $|y_{gl}| \to \infty$, *then* $\pi(\kappa_{gl} \leq \psi \mid \boldsymbol{y}_g, \tau^2, \sigma^2, a_g, b_g) \to 1$.

b) Fix $\epsilon \in (0, 1)$. *Then there exists a function* $h(p_g, \sigma^2, \boldsymbol{y}_g, a_g, b_g, \epsilon)$ *such that,*

$$\pi(\kappa_{gl} < \epsilon \mid \boldsymbol{y}_g, \tau^2, \sigma^2, a_g, b_g) \leq \left(\frac{\tau^2}{\sigma^2}\right)^{p_g/2+b_g} \left(\min\left(1, \frac{\tau^2}{\sigma^2}\right)\right)^{-p_g/2} h(p_g, \sigma^2, \boldsymbol{y}_g, a_g, b_g, \epsilon).$$

Consequently, $\pi(\kappa_{gl} \geq \epsilon \mid \boldsymbol{y}_g, \tau^2, \sigma^2, a_g, b_g) \to 1$ *as* $\tau^2 \to 0$.

Proof. See the Supplementary Material A.6 and A.7.

The theoretical statements outlined in Theorem 2.3.3 were originally discussed for the horseshoe prior (Datta and Ghosh, 2013), but have also been used in the context of several other continuous shrinkage priors (Datta and Dunson, 2016; Bhadra et al., 2017; Bai and Ghosh, 2019), dynamic trend filtering (Kowal et al., 2019), and small area estimation (Tang et al., 2018). We also note that Theorem 2.3.3 does not restrict the range of values a_g and b_g can take, meaning that Theorem 2.3.3 applies to a more general class of hyperparameter values than those considered in Bai and Ghosh (2019).

2.4 Computation

2.4.1 Gibbs Sampler

The full conditional updates corresponding to model (2.2.1), where $\boldsymbol{\beta}$ is endowed with a GIGG prior, are enumerated in Supplementary Material A.8. Following Polson and Scott (2011), we assign a half-Cauchy prior scaled by the residual error standard deviation $\tau \mid \sigma \sim C^{+}(0, \sigma)$ and use a prevalent data augmentation trick,

$$[\tau^2 \mid \nu] \sim IG(1/2, 1/\nu), \;\; [\nu \mid \sigma^2] \sim IG(1/2, 1/\sigma^2),$$

to obtain closed form full conditional updates for τ^2 and σ^2 (Makalic and Schmidt, 2016). There are two major computational bottlenecks for the proposed algorithm. The first is the full conditional update of $\boldsymbol{\beta}$,

$$[\boldsymbol{\beta} \mid \cdot] \sim N\bigg(\boldsymbol{Q}^{-1}\frac{1}{\sigma^2}\boldsymbol{X}^{\top}\Big(\boldsymbol{y} - \boldsymbol{C}\boldsymbol{\alpha}\Big), \boldsymbol{Q}^{-1}\bigg), \;\; \boldsymbol{Q} = \frac{1}{\sigma^2}\boldsymbol{X}^{\top}\boldsymbol{X} + \frac{1}{\tau^2}\boldsymbol{\Gamma}^{-1}\boldsymbol{\Lambda}^{-1}.$$

The second occurs when there are a multitude of group and local parameters that need to be drawn at each iteration of the Gibbs sampler, which is often the case in "large p" scenarios. Rather than naïvely sampling from the full conditional distributions there are several strategies to achieve faster posterior computation:

- Draw $\boldsymbol{v} \sim N\big(\sigma^{-2}\boldsymbol{X}^{\top}(\boldsymbol{y} - \boldsymbol{C}\boldsymbol{\alpha}), \boldsymbol{Q}\big)$, and then solve $\boldsymbol{Q}\boldsymbol{\beta} = \boldsymbol{v}$, rather than explicitly calculating $\boldsymbol{Q}^{-1}$.
- For "small n, large p" problems, the Woodbury identity can be utilized so that the full conditional update of $\boldsymbol{\beta}$ scales linearly in p (Bhattacharya et al., 2016).

2.4.2 Hyperparameter Selection

If the modeler wants to remain relatively agnostic to the choice of hyperparameters, one can use Marginal Maximum Likelihood Estimation (MMLE) (Casella, 2001), an empirical-Bayes approach executed iteratively within the Gibbs sampler. The $(l+1)$th update is

$$a_g^{(l+1)} = \psi_0^{-1}\bigg(E_{a_g^{(l)}}\big[\log(\gamma_g^2) \mid \boldsymbol{y}\big]\bigg), \;\; b_g^{(l+1)} = \psi_0^{-1}\bigg(-\frac{1}{p_g}\sum_{j=1}^{p_g} E_{b_g^{(l)}}\big[\log(\lambda_{gj}^2) \mid \boldsymbol{y}\big]\bigg),$$

where $\psi_0(\cdot)$ is the digamma function and the expectation terms can be estimated through standard Monte Carlo methods. The iterative procedure terminates when

$$\sum_{g=1}^{G} \left(a_g^{(l+1)} - a_g^{(l)}\right)^2 + \sum_{g=1}^{G} \left(b_g^{(l+1)} - b_g^{(l)}\right)^2$$

is less than some prespecified error tolerance. However, in our experience it is preferred to fix $a_g = 1/n$ for all g and use MMLE to estimate the b_g hyperparameters. The first reason is that a_g controls the strength of the thresholding effect and choosing a_g close to zero guarantees strong shrinkage of null coefficients towards zero. The second reason is that only estimating one hyperparameter per group is more feasible than estimating two hyperparameters per group, particularly when the number of groups is large. Since b_g primarily controls how dependent the shrinkage is within-group, it is more important to focus estimation on the b_g hyperparameters. We do recognize that setting $a_g = 1/n$ violates a condition in Theorem 2.3.1 where the infimum of the set of hyperparameters cannot converge to zero as $n \to \infty$. However, for practical purposes, this approach provides an automatic way to set a_g while also yielding similar results to a_g close to zero and fixed as a function of the sample size, such as $a_g = 1/100$.

Although MMLE is useful for problems where the number of groups, G, is small relative to the sample size, the estimates for the a_g's and b_g's will become increasingly variable in high-dimensional settings where the number of groups is large. There may also be low-dimensional settings where the user wants to incorporate explicit prior knowledge about the nature of the within-group signal density. In such cases, it may be preferred to fix hyperparameter values in accordance with subject matter expertise. As with the modified MMLE approach, we recommend setting $a_g = 1/n$ for all g. To fix b_g we recommend a useful heuristic whereby local, group, and global shrinkage parameters are simulated from the GIGG prior. Using the simulated shrinkage parameters, shrinkage factors can be constructed and the correlation between shrinkage factors within the same group can be empirically calculated. Selecting the hyperparameter b_g is then equivalent to selecting how dependent the shrinkage is within-group, a more easily understandable concept.

Another alternative in high-dimensional cases is to set $a_g = a$ and $b_g = b$ for all $g = 1, \ldots, G$ and $a, b > 0$. While this strategy loses the flexibility of customizing shrinkage for each group, it is at least capable of estimating a global tradeoff between group and local shrinkage in a manner that is more feasible for a MMLE procedure to reliably estimate. The corresponding MMLE updates for this procedure are

$$a^{(l+1)} = \psi_0^{-1}\left(\frac{1}{G}\sum_{g=1}^{G} E_{a^{(l)}}\left[\log(\gamma_g^2) \mid \boldsymbol{y}\right]\right), \;\; b^{(l+1)} = \psi_0^{-1}\left(-\frac{1}{p}\sum_{g=1}^{G}\sum_{j=1}^{p_g} E_{b^{(l)}}\left[\log(\lambda_{gj}^2) \mid \boldsymbol{y}\right]\right).$$

Implementations of GIGG regression with fixed hyperparameters and hyperparameters estimated via MMLE are available in the *gigg* R package on the Comprehensive R Archive Network (CRAN).

2.5 Simulations

2.5.1 Generative Model

The data generative mechanism is linear regression model (2.2.1), where $\boldsymbol{C}$ includes the intercept term and five adjustment covariates drawn from independent standard normal distributions, $\boldsymbol{\alpha} = (0, 1, 1, 1, 1, 1)^\top$, and $\boldsymbol{X}$ is drawn from a multivariate normal distribution with mean $\mathbf{0}$ and covariance matrix $\boldsymbol{\Sigma_X}$. $\boldsymbol{\Sigma_X}$ is determined such that the regressors have unit variance and block-diagonal exchangeable correlation structure. Pairwise correlations within each group are $\rho = 0.8$ for the high correlation simulation settings or $\rho = 0.6$ for the medium correlation simulation settings. For all simulation settings, the pairwise correlations across groups are 0.2 and the residual error variance, σ^2, is fixed such that $\boldsymbol{\beta}^\top \boldsymbol{\Sigma_X} \boldsymbol{\beta} / (\boldsymbol{\beta}^\top \boldsymbol{\Sigma_X} \boldsymbol{\beta} + \sigma^2) = 0.7$.

Label	Group Sizes	Correlation	Signal Type	Signal Details
C10H	10,10,10,10,10	0.8	Concentrated	Signal concentrated in one of the regressors in all five groups
D10H	10,10,10,10,10	0.8	Distributed	Signal distributed across all regressors within the first group
C10M	10,10,10,10,10	0.6	Concentrated	Signal concentrated in one of the regressors in all five groups
D10M	10,10,10,10,10	0.6	Distributed	Signal distributed across all regressors within the first group
C5	5,5,5,5,5,5,5,5,5,5	0.8	Concentrated	Signal concentrated in one regressor for five out of ten groups
D5	5,5,5,5,5,5,5,5,5,5	0.8	Distributed	Signal distributed across all regressors within the first two groups
C25	25,25	0.8	Concentrated	Signal concentrated in three regressors in the first group and two regressors in the second group
D25	25,25	0.8	Distributed	Signal distributed across first ten regressors within the first group
CL	30,10,5,3,2	0.8	Concentrated, Large Groups	Signal concentrated in one regressor in the group of size 30 and one regressor in the group of size 10
DL	30,10,5,3,2	0.8	Distributed, Large Groups	Signal distributed across all regressors within the group of size 30
CS	30,10,5,3,2	0.8	Concentrated, Small Groups	Signal concentrated in one regressor in the group of size 3 and one regressor in the group of size 2
DS	30,10,5,3,2	0.8	Distributed, Small Groups	Signal distributed across all regressors within the groups of size 5, 3, and 2

Table 2.1: Fixed coefficient simulation settings where $n = 500$ and $p = 50$. The label column refers to the name of the simulation setting that will be used throughout the rest of the simulation section. The group sizes column shows the sizes of all the groups within each simulation setting. The correlation column lists the pairwise correlations between regressors in the same group. The signal type and signal details columns explain how the signal is distributed among regressors within the active groups.

The first set of simulation settings will be called the fixed coefficient simulation settings, where $n = 500$ and $p = 50$ (see Table 2.1 for simulation setting details). In the context of this simulation study, a concentrated signal qualitatively refers to a simulation setting where the signal is contained within few regressors in a group and a distributed signal qualitatively refers to a simulation setting

where the signal is shared across many regressors within the same group. The purpose of the fixed coefficient simulation settings with equally sized groups is to ascertain which methods perform well when the within-group signal is sparse or dense, and whether or not the performance depends on group size or strength of the within-group regressor correlations. The purpose of the fixed coefficient simulation settings with groups of different sizes is to determine if the performance depends on whether concentrated or distributed signals are contained within groups of large or small size. Here, the groups of size 30 and 10 are considered the large groups and the groups of size 5, 3, and 2 are considered the small groups.

Beyond the fixed regression coefficient simulation settings, we also consider *random coefficient* simulations in the high correlation setting, where for each simulation iteration a random regression coefficient vector is generated. Here, we have a low-dimensional simulation setting with $n = 500$ and $p = 50$, as well as a high-dimensional simulation setting with $n = 200$ and $p = 500$. All groups in both random coefficient simulation settings contain 10 regressors. To construct a regression coefficient vector, we start by randomly selecting either a concentrated or distributed signal for the first group with even probability to guarantee that each simulation iteration will have at least one true signal. The concentrated and distributed signal magnitudes are selected such that the contribution to $\boldsymbol{\beta}^\top \boldsymbol{\Sigma}_{\boldsymbol{X}} \boldsymbol{\beta}$ is equal, namely the distributed signal is $\beta_{gj} = 0.25$ for $j = 1, ..., 10$ and the concentrated signal is $\beta_{g1} = 5.125$ and $\beta_{gj} = 0$ for $j = 2, ..., 10$. For the remaining groups, we randomly select a concentrated signal with probability 0.2, a distributed signal with probability 0.2, and no signal with probability 0.6. The goal of the random coefficient simulation settings is to show that, averaged across many combinations of regression coefficient vectors comprised of sparse within-group signals, dense within-group signals, and inactive groups, GIGG regression results in low mean-squared error.

2.5.2 Competing Methods and Evaluation Metrics

Estimation properties will be evaluated based on empirical mean-squared error (MSE), stratified by null and non-null coefficients, across 5000 replicates. That is,

$$\widehat{\text{MSE}} = \frac{1}{5000} \sum_{r=1}^{5000} \left(\hat{\boldsymbol{\beta}}^r - \boldsymbol{\beta} \right)^\top \left(\hat{\boldsymbol{\beta}}^r - \boldsymbol{\beta} \right),$$

where $\hat{\boldsymbol{\beta}}^r$ is the estimate of $\boldsymbol{\beta}$ from simulated dataset r. 5000 was selected so that the MSEs listed in the simulation results section are relatively precise. In the random coefficient simulations, calculating the MSE corresponds to an integrated mean-squared error (IMSE) metric averaged across the generative distribution of the regression coefficient vectors. For the fixed coefficient simulations we will consider several special cases of the GIGG prior with fixed hyperparameters,

namely all possible combinations of $a_g \in \{1/n, 1/2\}$ and $b_g \in \{1/n, 1/2, 1\}$. That way, we can check whether the intuition gleaned from Figure 2.2 empirically translates to the regression setting. We will also consider the GIGG prior when the hyperparameters $a_g = 1/n$ are fixed and b_g are estimated via MMLE.

The list of competing methods include Ordinary Least Squares (OLS), Horseshoe regression, Group Horseshoe+ regression (Xu et al., 2016), Spike-and-Slab Lasso (Rockova and George, 2018), Bayesian Group Lasso with Spike-and-Slab Priors (BGL-SS) (Xu and Ghosh, 2015), and Bayesian Sparse Group Selection with Spike-and-Slab Priors (BSGS-SS) (Xu and Ghosh, 2015). As a reminder, to avoid confusion with the group horseshoe prior proposed in this chapter, we will refer to the group horseshoe prior from Xu et al. (2016) as the group horseshoe+ prior. We will use the posterior mean estimator for all Bayesian methods, with the exception of Spike-and-Slab Lasso, BGL-SS, and BSGS-SS. BGL-SS and BSGS-SS will use the posterior median estimator and Spike-and-Slab Lasso will use the posterior mode estimator. Most methods requiring Markov chain Monte Carlo (MCMC) sampling have 10000 burn-in draws, followed by 10000 posterior draws with no thinning. Some exceptions are BGL-SS and BSGS-SS which have 1000 burn-in draws and 2000 posterior draws with no thinning, due to the relatively slower posterior sampling algorithms. Another exception is group horseshoe+ regression in the high-dimensional random coefficient simulation, which required 100000 burn-in draws to consistently converge.

2.5.3 Simulation Results

Table 2.2 presents the MSE for simulation settings C10H and D10H and Supplementary Tables A.1, A.2, and A.3 list the MSEs for the C10M, D10M, C5, D5, C25, and D25 simulation settings. Because the results for C10H and D10H are similar to C10M, D10M, C5, D5, C25, and D25, we will only focus our discussion around the C10H and D10H simulation settings. The first noteworthy observation is that group horseshoe regression has a uniformly lower MSE than both OLS and horseshoe regression for both null and non-null estimation, although the discrepancy between horseshoe and OLS is much larger than the difference between group horseshoe and horseshoe, particularly for the null coefficients. For GIGG regression with fixed hyperparameters, the top performer is GIGG regression with $b_g = 1/n$ when the signal is concentrated within-group (Null MSE = 0.11, Non-Null MSE = 0.30) and $a_g = 1/n, b_g = 1$ when the signal is distributed within-group (MSE = 1.46), exactly as Figure 2.2 suggests. However, if the user sets $b_g = 1$ when the signal is concentrated (Null MSE = 0.53, Non-Null MSE = 0.49) or $b_g = 1/n$ when the signal is distributed (Null MSE = 0.04, Non-Null MSE = 3.60), then the "incorrect" prior information results in notably worse MSE compared to the "correct" prior information. That being said, $b_g = 1/2$ appears to be a middle ground where the performance for both concentrated and distributed simulation settings

$\rho = 0.8$	Concentrated			Distributed		
Method	Null	Non-Null	Overall	Null	Non-Null	Overall
Ordinary Least Squares	3.74	0.41	4.16	8.09	2.03	10.12
Horseshoe	0.51	0.41	0.92	0.85	2.14	2.99
GIGG ($a_g = 1/n, b_g = 1/n$)	0.11	0.30	**0.40**	0.04	3.60	3.63
GIGG ($a_g = 1/2, b_g = 1/n$)	0.11	0.30	**0.41**	0.04	3.56	3.59
GIGG ($a_g = 1/n, b_g = 1/2$)	0.29	0.39	0.67	0.03	1.57	**1.61**
*GIGG ($a_g = 1/2, b_g = 1/2$)	0.33	0.40	0.72	0.24	1.70	1.94
GIGG ($a_g = 1/n, b_g = 1$)	0.53	0.49	1.03	0.03	1.43	**1.46**
GIGG ($a_g = 1/2, b_g = 1$)	0.58	0.49	1.07	0.26	1.43	1.69
GIGG (MMLE)	0.23	0.36	**0.59**	0.04	1.36	**1.40**
Group Horseshoe+	0.30	0.39	0.70	0.08	1.64	1.73
Spike-and-Slab Lasso	0.15	0.33	**0.48**	0.21	4.27	4.49
BGL-SS	2.02	0.80	2.82	0.04	1.31	**1.34**
BSGS-SS	0.23	0.42	0.65	0.04	1.84	1.88

Table 2.2: Mean-squared errors (MSE) for simulation settings C10H and D10H in Table 2.1 ($n = 500, p = 50$) with high pairwise correlations ($\rho = 0.8$). Bolded cells indicate the four methods with the lowest overall MSE. Four methods are highlighted to emphasize that GIGG (MMLE) is the best method with respect to MSE for both concentrated and distributed signals aside from methods that only perform well for one of the two settings. *GIGG ($a_g = 1/2$ and $b_g = 1/2$) is equivalent to group horseshoe regression.

is generally good.

Examining the performance of the competing methods, we note that Spike-and-Slab Lasso does very well for the concentrated signal setting (MSE = 0.48), but struggles when the signal is distributed (MSE = 4.49). Conversely, BGL-SS does poorly when the signal is concentrated (MSE = 2.82), but has good performance when the signal is distributed (MSE = 1.34). Group horseshoe+ regression and BSGS-SS have relatively low MSE across the low-dimensional simulation settings, however, GIGG with MMLE almost always outperforms both methods in the low-dimensional cases with respect to overall MSE. The improved performance for GIGG regression with MMLE over a method like group horseshoe+ regression is precisely because GIGG regression with MMLE is able to data-adaptively control the dependence of the grouped multivariate shrinkage. Group horseshoe+ regression cannot directly control within-group dependence because there are no hyperparameters in the prior specification.

Table 2.3 shows the MSE results for the CL and DL simulation settings and Table 2.4 lists the MSE results for the CS and DS simulation settings. As with the other fixed coefficient simulation settings, the same general conclusions hold. Whether or not a concentrated signal is contained in large or small groups, GIGG with MMLE and GIGG with fixed hyperparameters where $b_g = 1/n$

$\rho = 0.8$	Concentrated			Distributed		
Method	Null	Non-Null	Overall	Null	Non-Null	Overall
Ordinary Least Squares	2.02	0.08	2.11	1.58	2.85	4.43
Horseshoe	0.19	0.06	0.25	0.16	1.04	1.20
GIGG ($a_g = 1/n, b_g = 1/n$)	0.02	0.04	**0.07**	0.01	1.92	1.93
GIGG ($a_g = 1/2, b_g = 1/n$)	0.03	0.07	**0.10**	0.01	1.88	1.89
GIGG ($a_g = 1/n, b_g = 1/2$)	0.06	0.05	0.10	0.01	0.99	**1.00**
*GIGG ($a_g = 1/2, b_g = 1/2$)	0.06	0.05	0.11	0.05	0.99	1.04
GIGG ($a_g = 1/n, b_g = 1$)	0.12	0.06	0.19	0.01	0.85	**0.86**
GIGG ($a_g = 1/2, b_g = 1$)	0.13	0.06	0.19	0.05	0.83	**0.88**
GIGG (MMLE)	0.03	0.04	**0.07**	0.01	0.80	**0.81**
Group Horseshoe+	0.06	0.05	0.11	0.04	1.00	1.03
Spike-and-Slab Lasso	0.04	0.03	**0.07**	0.08	3.29	3.36
BGL-SS	1.26	0.22	1.48	0.02	1.36	1.38
BSGS-SS	0.06	0.06	0.12	0.01	1.30	1.31

Table 2.3: Mean-squared errors (MSE) for simulation settings CL and DL in Table 2.1 ($n = 500, p = 50$) with high pairwise correlations ($\rho = 0.8$). Bolded cells indicate the four methods with the lowest overall MSE. Four methods are highlighted to emphasize that GIGG (MMLE) is the best method with respect to MSE for both concentrated and distributed signals aside from methods that only perform well for one of the two settings. *GIGG ($a_g = 1/2$ and $b_g = 1/2$) is equivalent to group horseshoe regression.

have some of the lowest overall MSEs across all methods. Whether or not a distributed signal is contained in large or small groups, GIGG with MMLE and GIGG with fixed hyperparameters where $b_g = 1$ have some of the lowest overall MSEs. Spike-and-Slab LASSO performed well in the concentrated simulation settings, but BGL-SS only performed well in the distributed setting when the groups containing the true signals were small. Overall, it does not appear that group size and signal distribution within the groups fundamentally change the performance of GIGG with MMLE or GIGG with fixed hyperparameters, within the scope of the data generative parameters that we explored.

Next, Table 2.5 summarizes the b_g hyperparameter estimates across 5000 simulation iterations for all high correlation simulation settings with $n = 500$, $p = 50$, and $G = 5$. For simulation setting C10H we see that median b_g hyperparameter estimate for groups 1-5 goes from 0.52 in group 1, which contains the smallest signal, to 0.27 and 0.26 for the largest signals. That is, as the concentrated signal becomes stronger, the median b_g estimate starts moving towards zero. Conversely, for all simulation settings with distributed signals, we generally observe that groups with either all signals or all noise regressors tend to result in b_g estimates that are greater than one regardless of how large the groups containing the distributed signals are. For the CL simulation setting, we

$\rho = 0.8$	Concentrated			Distributed		
Method	Null	Non-Null	Overall	Null	Non-Null	Overall
Ordinary Least Squares	2.05	0.06	2.11	2.07	0.39	2.46
Horseshoe	0.19	0.03	0.22	0.37	0.52	0.89
GIGG ($a_g = 1/n, b_g = 1/n$)	0.02	0.03	**0.04**	0.02	1.06	1.08
GIGG ($a_g = 1/2, b_g = 1/n$)	0.02	0.04	**0.06**	0.02	1.06	1.08
GIGG ($a_g = 1/n, b_g = 1/2$)	0.04	0.04	0.08	0.00	0.37	0.37
*GIGG ($a_g = 1/2, b_g = 1/2$)	0.05	0.03	0.08	0.04	0.36	0.40
GIGG ($a_g = 1/n, b_g = 1$)	0.06	0.05	0.11	0.00	0.32	**0.33**
GIGG ($a_g = 1/2, b_g = 1$)	0.09	0.04	0.13	0.04	0.32	**0.36**
GIGG (MMLE)	0.02	0.03	**0.05**	0.00	0.32	**0.32**
Group Horseshoe+	0.04	0.03	0.07	0.09	0.43	0.52
Spike-and-Slab Lasso	0.04	0.02	**0.06**	0.09	1.49	1.58
BGL-SS	0.08	0.06	0.13	0.00	0.28	**0.28**
BSGS-SS	0.02	0.03	0.06	0.00	0.45	0.45

Table 2.4: Mean-squared errors (MSE) for simulation settings CS and DS in Table 2.1 ($n = 500, p = 50$) with high pairwise correlations ($\rho = 0.8$). Bolded cells indicate the four methods with the lowest overall MSE. Four methods are highlighted to emphasize that GIGG (MMLE) is the best method with respect to MSE for both concentrated and distributed signals aside from methods that only perform well for one of the two settings. *GIGG ($a_g = 1/2$ and $b_g = 1/2$) is equivalent to group horseshoe regression.

observe that the general trends for concentrated and distributed signals hold, namely that group 1 and group 2, which contain the concentrated signals, have median b_g estimates between 0.2 and 0.3, and groups 3-5, which are null groups, have median b_g estimates greater than one. However, the CS simulation setting is a little more interesting. Group 4 and group 5 in the CS simulation setting are the active groups with concentrated signals and we see that the median b_g estimates are 0.16 and 0.14, respectively. That is, small groups with concentrated signals seem to result in b_g hyperparameter estimates that are even closer to zero compared with larger groups with concentrated signals. Moveover, groups 1-3 in the CS simulation setting are all null groups, and they show a general trend of the median b_g hyperparameter estimates getting smaller, the larger the group is. Specifically, the group of size 30 in the CS simulation setting has a median b_g estimate of 0.66 and the group of size 5 in the CS simulation setting has a median b_g estimate of 0.92. Finally, it is important to mention that b_g is capped at four in our implementation to facilitate numerical stability of the MMLE procedure. For group 5 in the CL simulation setting, b_g was set to four in 131 out of 5000 simulation iterations. For group 4 in the DL simulation setting, b_g was set to four in 350 out of 5000 simulation iterations. For group 5 in the DL simulation setting, b_g was set to four 1195 times out of 5000 simulation iterations.

Label	Group Sizes	Group 1	Group 2	Group 3	Group 4	Group 5
C10H	10,10,10,10,10	0.52 (0.32-0.84)	0.35 (0.25-0.71)	0.29 (0.23-0.55)	0.27 (0.22-0.43)	0.26 (0.22-0.44)
D10H	10,10,10,10,10	1.84 (1.11-2.59)	1.19 (0.75-1.80)	1.19 (0.74-1.81)	1.19 (0.74-1.81)	1.19 (0.71-1.81)
CL	30,10,5,3,2	0.28 (0.24-0.40)	0.22 (0.19-0.30)	1.30 (0.69-2.69)	1.51 (0.68-3.41)	1.73 (0.72-4.00*)
DL	30,10,5,3,2	1.96 (1.04-2.91)	1.60 (0.83-2.45)	2.09 (0.94-3.51)	2.59 (1.04-4.00*)	3.13 (1.03-4.00*)
CS	30,10,5,3,2	0.66 (0.54-0.88)	0.79 (0.57-1.22)	0.92 (0.60-1.54)	0.16 (0.13-0.23)	0.14 (0.12-0.21)
DS	30,10,5,3,2	0.95 (0.64-1.47)	1.31 (0.75-2.33)	2.07 (1.21-2.97)	1.51 (0.72-2.14)	1.03 (0.25-1.46)

Table 2.5: Median (2.5% Quantile - 97.5% Quantile) b_g estimates for GIGG regression with MMLE in all fixed regression coefficient, high correlation simulations settings with $n = 500$, $p = 50$ and $G = 5$ with 5000 replicates. See Table 2.1 for the simulation setting details. Here, large groups correspond to groups of size 30 and 10 and small groups correspond to groups of size 5, 3, and 2. *b_g is capped at four to facilitate numerical stability of the MMLE procedure.

Lastly, we consider the IMSE for the random coefficient simulation settings presented in Table 2.6. As with the fixed regression coefficient simulations, group horseshoe (Null IMSE = 0.39) and group horseshoe+ regression (Null ISME = 0.36) lead to a substantial improvement in IMSE compared to horseshoe regression in the low-dimensional simulation setting. However, in the low-dimensional simulation setting, we also observe that the additional flexibility of GIGG regression with MMLE to self-adapt to different types of within-group signal distributions results in noticeable improvements in IMSE for the null coefficients (Null IMSE = 0.21). Spike-and-Slab Lasso and BGL-SS struggle in the random coefficient simulation scenario because they tend to only work well when the signal is concentrated or distributed, respectively, leading to unfavorable average performance. The high-dimensional simulation setting shows that GIGG regression with MMLE does not perform as well as group horseshoe regression, group horseshoe+ regression, and BSGS-SS, likely due to the fact that there is limited sample size to estimate many more group-specific b_g hyperparameters. Note that BSGS-SS has very low Null MSE (Null MSE = 2.22), but very high Non-Null MSE (Non-Null MSE = 254.02) compared with many of the GIGG regression methods. Being a spike-and-slab based method, BSGS-SS has an inherent advantage over continuous shrinkage methods in estimating the null counterpart of sparse parameters because it shrinks coefficients to exact zero. Moreover, because BSGS-SS is based off of sparse group lasso, it shrinks all coefficients much more strongly toward zero than GIGG regression methods. GIGG regression with fixed hyperparameters $a_g = 1/2$ and $b_g = 1$ has the best performance of the continuous shrinkage prior methods, likely because averaging a signal across highly correlated regressors in a high-dimensional setting is preferable to assigning the entire signal to one regressor, with respect to a squared error loss function. The high-dimensional simulations indicate that better strategies to determine hyperparameter values for high-dimensional regression problems could result in improved estimation properties.

	Low-Dimensional			High-Dimensional		
Method	Null	Non-Null	Overall	Null	Non-Null	Overall
Ordinary Least Squares	8.84	3.38	12.21	-	-	-
Horseshoe	0.70	1.18	1.88	86.04	215.36	301.40
GIGG ($a_g = 1/n, b_g = 1/n$)	0.09	1.79	1.88	131.32	252.04	383.36
GIGG ($a_g = 1/2, b_g = 1/n$)	0.10	1.83	1.93	128.61	250.49	379.10
GIGG ($a_g = 1/n, b_g = 1/2$)	0.33	1.15	**1.47**	80.61	213.24	**293.85**
*GIGG ($a_g = 1/2, b_g = 1/2$)	0.39	1.13	**1.52**	60.73	207.55	**268.29**
GIGG ($a_g = 1/n, b_g = 1$)	0.69	1.11	1.79	90.12	210.61	300.74
GIGG ($a_g = 1/2, b_g = 1$)	0.75	1.11	1.85	53.66	203.86	**257.53**
GIGG (MMLE)	0.21	1.06	**1.27**	93.11	220.90	314.01
Group Horseshoe+	0.36	1.14	**1.49**	82.91	213.06	295.97
Spike-and-Slab Lasso	0.16	3.65	3.81	159.02	344.82	503.84
BGL-SS	2.84	2.44	5.28	1918.84	678.36	2597.19
BSGS-SS	0.36	1.45	1.81	2.22	254.02	**256.25**

Table 2.6: Integrated mean-squared errors (IMSE) for the random regression coefficient simulation settings with high pairwise correlations ($\rho = 0.8$). The low-dimensional simulation setting has $n = 500$ and $p = 50$ and the high-dimensional simulation setting has $n = 200$ and $p = 500$. Bolded cells indicate the four methods with the lowest overall IMSE. *GIGG ($a_g = 1/2$ and $b_g = 1/2$) is equivalent to group horseshoe regression.

2.6 Data Example

The National Health and Nutrition Examination Survey (NHANES) is a collection of studies conducted by the National Center for Health Statistics with the overarching goal of evaluating the health and nutritional status of the United States' populace. Data collection consists of a written survey and physical examination which records demographic, socioeconomic, dietary, and health-related information, including physiological measurements and laboratory tests. We will specifically apply GIGG regression to a subset of 990 adults from NHANES 2003-2004 with 35 measured contaminants across five exposure classes: metals (3 exposures), phthalates (7 exposures), organochlorine pesticides (8 exposures), polybrominated diphenyl ethers (PBDEs) (7 exposures), and polycyclic aromatic hydrocarbons (PAHs) (10 exposures). Figure 2.1 illustrates the block diagonal correlation structure of these exposures, where areas of high correlation are mostly contained within exposure class. Gamma glutamyl transferase (GGT), an enzymatic marker of liver functionality, is the outcome of interest. GGT and all environmental exposures were log-transformed to remove right skewness and then subsequently standardized. The final model was adjusted for age, sex, body mass index, poverty-to-income ratio, ethnicity, and urinary creatinine.

Figure 2.3 presents the estimated percent change in GGT corresponding to a twofold change in

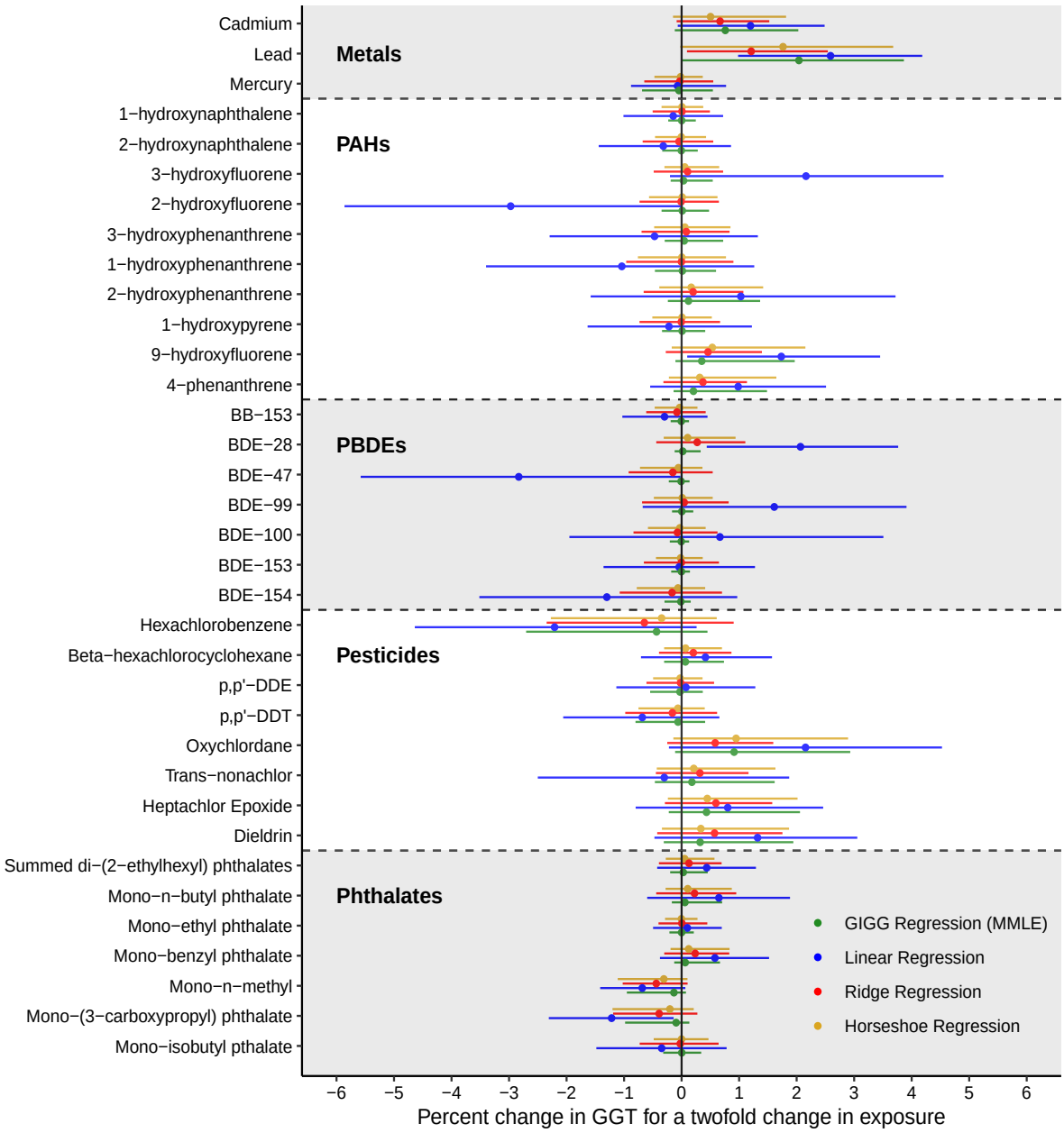


Figure 2.3: Estimated associations between environmental toxicants (metals, phthalates, pesticides, PBDEs, and PAHs) and gamma glutamyl transferase (GGT) from NHANES 2003-2004 ($n = 990$).

each environmental exposure and their associated 95% credible intervals for methods commonly used in multipollutant modeling. Bayesian linear regression with noninformative priors and ridge regression were implemented in R Stan using four chains with no thinning, each with 1000 burn-in draws and 1000 posterior draws. Horseshoe regression and GIGG regression with MMLE used 10000 burn-in samples, followed by 10000 posterior draws with a thinning interval of five. As

with the simulation section, GIGG regression with MMLE refers to an implementation of GIGG which fixes $a_g = 1/n$ for all g and then uses MMLE to estimate the b_g hyperparameters. Convergence of the MCMC chains was evaluated using Gelman-Rubin's potential scale reduction factor (PSRF) (Gelman and Rubin, 1992). All methods had a PSRF of $1.00 - 1.01$ for the regression coefficients, indicating that all MCMC chains converged. For GIGG regression with MMLE, the median effective sample size for the β_{gj}'s was 7309 with an interquartile range (IQR) of 3634-9266 and the effective sample size for σ^2 was 10000. For the shrinkage parameters, the local shrinkage parameters had a median effective sample size of 9541 with an IQR of 7819-10000, the group shrinkage parameters had a median effective sample size of 1790 with an IQR of 1707-2271, and the global shrinkage parameter had an effective sample size of 983.

Figure 2.3 compares GIGG regression with Bayesian linear regression (non-informative priors), ridge regression, and horseshoe regression. GIGG is generally more efficient than the other methods, having narrower credible intervals, because GIGG better deals with multicollinearity and homogeneous within-group effect sizes. When there is little multicollinearity and heterogeneous within-group effect sizes, GIGG has similar efficiency to the horseshoe. Further, GIGG allows for different shrinkage on coefficients, unlike ridge regression which overshrinks large coefficients. In detail, the median credible interval length for GIGG regression with MMLE is 21.0% shorter for the PAHs, 63.2% shorter for the PBDEs, and 22.5% shorter for the phthalates compared to horseshoe regression, which are all exposures classes with high pairwise correlations and common estimated effect sizes. However, the metals exposure class, which has weak pairwise correlations and heterogeneous estimated effect sizes, results in a median credible interval length of 0.31 for GIGG regression with MMLE and 0.28 for horseshoe regression. Ridge regression estimates that a twofold change in lead exposure is associated with 1.21% higher GGT (95% CI: 0.09, 2.54), while horseshoe regression estimates 1.76% higher GGT (95% CI: -0.02, 3.68) and GIGG regression with MMLE estimates 2.04% higher GGT (95% CI: 0.01, 3.87). From a computational perspective, GIGG regression with MMLE generated a median effective sample size of 559.6 per second for the β_{gj}'s, compared to a median effective sample size of 791.1 per second for horseshoe regression.

Supplementary Figure A.1 provides a focused comparison of the various group shrinkage methods from the simulation study. GIGG is generally more efficient than the other continuous shrinkage prior methods, having narrower credible intervals than group horseshoe and group horseshoe+. As with the results in Figure 2.3, these efficiency gains are attributable to GIGG with MMLE better handling multicollinearity and homogeneous within-group effect sizes. In detail, GIGG regression with MMLE, group horseshoe regression, and group horseshoe+ regression all have very similar performance in terms of point estimation. The 95% credible interval for lead covers zero for group horseshoe+ regression (1.82% higher GGT; 95% CI: -0.02, 3.70), while the 95% credible interval

for lead does not cover zero for group horseshoe regression (1.88% higher; 95% CI: 0.01, 3.72) and GIGG regression with MMLE (2.04% higher GGT; 95% CI: 0.01, 3.87). For the PAHs, GIGG regression with MMLE has a 26.1% shorter median credible interval length than group horseshoe regression and a 25.7% shorter median credible interval length than group horseshoe+ regression. For the PBDEs, GIGG regression with MMLE has a 57.7% shorter median credible interval length than group horseshoe regression and a 60.6% shorter median credible interval length than group horseshoe+ regression. Differences in credible interval interval length for the metals and pesticides among GIGG regression with MMLE, group horseshoe regression, and group horseshoe+ regression were much smaller. The posterior median estimator corresponding to BGL-SS selected both the metals and pesticides groups, despite the fact that no other method identified any pesticides based on 95% credible intervals covering zero or posterior inclusion probabilities being larger than 0.5. BSGS-SS selected lead and cadmium, while the 95% credible intervals for GIGG regression with MMLE and group horseshoe regression only identified lead.

2.7 Discussion

The principal methodological contribution of this chapter is to construct a continuous shrinkage prior that improves regression coefficient estimation in the presence of grouped regressors. GIGG regression flexibly controls the relative contributions of individual and group shrinkage to improve regression coefficient estimation, resulting in a relative IMSE reduction of 32.4% compared to horseshoe regression. One of the main limitations of GIGG regression is that regressor groupings must be explicitly specified and regressor groupings may not overlap. Additionally, although the GIGG prior can be imposed on regression coefficients in Bayesian generalized linear models, a theoretical evaluation of the shrinkage properties for non-normal outcome data would be necessary to determine if the GIGG prior is appropriate for such models.

There are several considerations for deciding between a spike-and-slab based bi-level selection method and a grouped multivariate shrinkage prior based method. The first how large the dataset is. Computationally, it is much slower to sample from the posterior corresponding to BSGS-SS than it is to sample from the posterior distribution corresponding to GIGG regression. Therefore, in higher dimensional problems, sampling the posterior corresponding to BSGS-SS may be computationally prohibitive. The second consideration is with regard to the tradeoff between group and local shrinkage. BSGS-SS only has two hyperparameters, so fixing those hyperparameters defines a group-local tradeoff *for all* groups. The GIGG prior is different in that a unique a_g and b_g for each group allows group-local tradeoffs *for each* group. If the expectation is for some groups to have concentrated signals and for others to have distributed signals, then GIGG regression with MMLE is better able to tailor the shrinkage corresponding to each group. The third is how important vari-

able selection is. There are several techniques to define selection for continuous shrinkage priors, however spike-and-slab based methods define variable selection much more naturally through posterior inclusion probabilities. Therefore, if selection is a primary goal, then a spike-and-slab based method like BSGS-SS might be preferred.

One limitation of the GIGG prior is that the groups must be specified a priori, leading to potential scenarios where grouping information may be uninformative or misleading. However, the ability of GIGG shrinkage to adapt to concentrated and distributed signals within the a priori defined groups affords some robustness to misspecification of the grouping structure. Namely, in cases where the grouping information is uninformative, then GIGG regression with MMLE will still estimate an appropriate level of dependence within the multivariate shrinkage profile, likely resulting in less dependent shrinkage within group if the effect sizes are heterogeneous. If the grouping information is misleading, GIGG regression with MMLE will likely estimate less dependent multivariate shrinkage within group, such that the impact of other regressors in the group is lessened. That is, the flexibility of the dependence in the multivariate shrinkage profiles provides some protection against improper grouping information.

The analysis of multiple pollutant data and chemical mixtures is a key thrust of the National Institute of Environmental Health Sciences, and the GIGG prior provides a useful framework for achieving variance reduction in the presence of group-correlated exposures, characterizing uncertainties in point estimates, and constructing policy relevant metrics, like summary risk scores, in a principled way. However, the generality of the GIGG prior coupled with the relative ease of computation means that, despite its motivation coming from environmental epidemiology, the GIGG prior is applicable to many other areas. For example, in neuroimaging studies, scalar-on-image regression (Kang et al., 2018) has been widely used to study the association between brain activity and clinical outcomes of interest. The whole brain can be partitioned into a set of exclusive regions according to brain functions and anatomical structures. Within the same region, the brain imaging biomarkers tend to be more correlated and have similar effects on the outcome variable. The GIGG prior can be extended for scalar-on-image regression and it has great potential to improve estimating the effects of imaging biomarkers by incorporating brain region information.

In this chapter, our focus was sparse estimation, but it is also natural to inquire about uncertainty quantification and variable selection. Based on our simulations, the conclusions of van der Pas et al. (2017) are relevant for the GIGG prior when $0 < a_g \leq 1/2$, but a comprehensive study needs to be carried out. There is no consensus way of defining variable selection for continuous shrinkage priors, however there are several approaches to determine a final active set, including credible intervals covering zero (van der Pas et al., 2017), decoupling shrinkage and selection (DSS) (Hahn and Carvalho, 2015), and penalized credible regions (Zhang and Bondell, 2018). For horseshoe-style shrinkage, variable selection defined through credible intervals covering zero is

highly conservative, but works well if one wants to limit the number of false discoveries. The penalized credible region approach searches for the sparsest model that falls within the $100\times(1-\alpha)\%$ joint elliptical credible region, while DSS constructs an adaptive lasso-style objective function with the goal of sparsifying the posterior mean such that most of the predictive variability is still explained. Since the DSS construction is framed from a prediction perspective, this approach may not be ideal for regression coefficient estimation problems in the presence of correlated regressors. Another crucial point to make is that if one is interested in selection, the posterior mode estimator for the horseshoe prior will result in exact zero estimates, and an approximate algorithm for calculating the joint posterior mode was developed in Bhadra et al. (2019) using the horseshoe-like prior. Therefore, one could conceptualize an extension of the expectation-maximization algorithm developed by Bhadra et al. (2019) using a "GIGG-like" prior. A second option, from a variable selection perspective rather than a model selection perspective, is to ascertain whether or not the marginal posterior modes equal zero, based on the posterior draws (Liu and Ghosh, 2020). Here, a posterior mode equal to zero refers to a regressor that is not selected and a posterior mode not equal to zero refers to a regressor that is selected. Further work is needed to juxtapose the behavior of all of these different methods for selection and develop novel algorithms for calculating the marginal and joint posterior modes.

CHAPTER 3

Mediation Analysis with External Information on the Total Effect

3.1 Introduction

Mediation analysis is an important tool in epidemiology to elucidate the intermediary pathways by which an exposure affects an outcome (Baron and Kenny, 1986; Robins and Greenland, 1992; Pearl, 2001; VanderWeele, 2015; Song et al., 2020). In mediation analysis, the total effect (TE) characterizes the effect of the exposure on the outcome and is additively decomposed into the natural direct effect (NDE) and the natural indirect effect (NIE). The NDE and NIE quantify how well measures of the intermediary pathways, called mediators, explain the TE. The logical progression of mediation analysis is generally sequential, where researchers first establish that the exposure is causally related to the outcome, and then hypothesize mechanisms that may explain the causal relationship. Consequently, researchers frequently consider mediation hypotheses only if there is a well-established literature showing statistical and biological significance of the TE. The objective of this chapter is to integrate available external summary-level information on the TE into mediation models, thereby improving NDE and NIE estimation for mediation analyses with individual-level omics data on a limited number of participants.

The motivating example comes from the Puerto Rico Testsite for Exploring Contamination Threats (PROTECT), a prospective birth cohort study in Puerto Rico. Preterm births, defined as gestational age at delivery of less than 37 weeks, coupled with their downstream health complications, are a large concern for the Puerto Rican health care system. One widely studied risk factor for preterm deliveries is elevated exposure to a class of endocrine disrupting chemicals called phthalates (Ferguson et al., 2014; Welch et al., 2022). The goal of the present study is to test whether metabolites corresponding to the inflammatory pathway Cytochrome p450 ($\boldsymbol{M}$) mediate the relationship between phthalate exposure ($\boldsymbol{A}$) and gestational age at delivery ($\boldsymbol{Y}$) adjusted for confounders ($\boldsymbol{C}$). The sample size in PROTECT with exposure and mediator data is approximately

450. However, a study by Welch et al. (2022), which pools data corresponding to the TE of phthalates on birth outcomes across 16 studies, has an approximate sample size of 5000 (after omitting PROTECT). The goal of this chapter is to utilize the external summary-level information on the TE from the Welch et al. (2022) pooled study to improve estimation efficiency of the NDE and NIE in PROTECT.

There is no existing work explicitly incorporating external summary-level information on the TE into an internal mediation model, however there is related work on nested internal and external models in the data integration literature (Chatterjee et al., 2016; Cheng et al., 2018; Estes et al., 2018; Cheng et al., 2019; Gu et al., 2019; Han and Lawless, 2019; Gu et al., 2021; Zhai and Han, 2022). Specifically, these papers consider the situation where an external model or prediction algorithm is fit on a set of predictors and the resulting summary-level statistics or predictions are then used to inform an internal model that contains a proper superset of those predictors. External information on the TE can partially be framed in a similar manner where the external information comes from the TE model, $\boldsymbol{Y} \mid \boldsymbol{A}, \boldsymbol{C}$, and the model of interest is $\boldsymbol{Y} \mid \boldsymbol{M}, \boldsymbol{A}, \boldsymbol{C}$. However, the key difference for mediation models is that mediation models are specified from $\boldsymbol{M} \mid \boldsymbol{A}, \boldsymbol{C}$ and $\boldsymbol{Y} \mid \boldsymbol{M}, \boldsymbol{A}, \boldsymbol{C}$ models. Therefore, it is important to understand how information on $\boldsymbol{Y} \mid \boldsymbol{A}, \boldsymbol{C}$ informs parameter estimation corresponding to $\boldsymbol{Y} \mid \boldsymbol{M}, \boldsymbol{A}, \boldsymbol{C}$ and $\boldsymbol{M} \mid \boldsymbol{A}, \boldsymbol{C}$ simultaneously.

Our work has several new aspects. First, we develop a method to integrate external summary-level information on the TE into an internal mediation model through constrained maximum likelihood estimation. Second, we show that, for a continuous outcome and continuous candidate mediators, the constrained estimator is asymptotically more efficient than the unconstrained estimator for estimating the NDE and, provided that the outcome-mediator association conditional on exposure is non-zero, the NIE. More specifically, the magnitude of the asymptotic relative efficiency gains for estimating the NDE and NIE are both functions of the partial R^2 between the $\boldsymbol{Y} \mid \boldsymbol{A}, \boldsymbol{C}$ and $\boldsymbol{Y} \mid \boldsymbol{M}, \boldsymbol{A}, \boldsymbol{C}$ models. Third, we robustify this mediation framework to violations of transportability assumptions by introducing a mediation model where the internal TE parameter is modeled as a random effect to deal with potential incongeniality of the external and internal TE estimates. The random effect treatment of the internal TE parameter facilitates Empirical-Bayes style shrinkage which data-adaptively shrinks more strongly towards the external TE estimate if the internal and external populations appear to have *similar* TEs (Morris, 1983; Mukherjee and Chatterjee, 2008). Lastly, we provide corroborative evidence in PROTECT to the conclusions of Aung et al. (2020), which found a significant indirect effect of phthalate exposure on gestational age at delivery through the Cytochrome p450 pathway in the LIFECODES prospective birth cohort with participants from the greater Boston area. The two cohorts are very different in demographics, socioeconomic profile, behavior and lifestyle factors, thus this replicated finding may offer a genuine biological insight. To our knowledge, this is the first methodological work that combines

ideas from data integration and mediation analysis.

The structure of the chapter is as follows. In Section 3.2 we explicitly define the problem, discuss methods for estimating model parameters in a linear mediation model with and without external information. We derive estimators of the NDE and NIE corresponding to each method. In Section 3.3, we compare asymptotic results corresponding to the NDE and NIE estimators defined in Section 3.2 and discuss robustness to violations of transportability. In Section 3.4 we empirically substantiate the findings from Section 3.3 with a comprehensive simulation study. In Section 3.5 we apply this methodology to the PROTECT mediation analysis. Section 3.6 offers a brief concluding discussion.

3.2 Methods

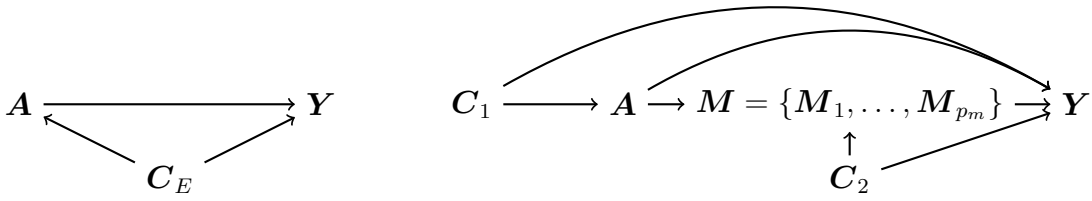
3.2.1 Notation and Model Specifications

We consider a mediation analysis setting where a collection of continuous candidate mediators is hypothesized to mediate the association between a single exposure and a continuous health outcome (see Figure 3.1). For the internal study, we assume that we have individual-level data on n observations. For observation i ($i = 1, \ldots, n$), let Y_i denote the outcome, $\boldsymbol{M}_{i\cdot}^\top = (M_{i1}, \ldots, M_{ip_m})^\top$ denote a collection of p_m candidate mediators, A_i denote the exposure with $E[A_i] = 0$, and $\boldsymbol{C}_{i\cdot}^\top$ denote a collection of p_c confounders and adjustment covariates plus the intercept term. To be clear on the notation, $\boldsymbol{M}_{i\cdot}^\top$ is a $p_m \times 1$ column vector, M_{ij} is the realization of the j-th mediator for observation i, and $\boldsymbol{C}_{i\cdot}^\top$ is a $p_c \times 1$ column vector. We do not distinguish between confounders of the outcome-exposure relationship and the outcome-mediator relationship, as in Figure 3.1; we assume that $\boldsymbol{C}_{i\cdot}^\top$ contains all confounders for both relationships. In our presentation, we also use matrix notation, namely $\boldsymbol{Y} = (Y_1, \ldots, Y_n)^\top$ is the $n \times 1$ column vector containing the observed outcomes, $\boldsymbol{M} = (\boldsymbol{M}_{1\cdot}^\top, \ldots, \boldsymbol{M}_{n\cdot}^\top)^\top$ is the $n \times p_m$ design matrix of observed mediator values, $\boldsymbol{A} = (A_1, \ldots, A_n)^\top$ is the $n \times 1$ column vector containing the observed exposures, and $\boldsymbol{C} = (\boldsymbol{C}_{1\cdot}^\top, \ldots, \boldsymbol{C}_{n\cdot}^\top)^\top$ represents the $n \times p_c$ matrix of observed confounders, plus an intercept term. The true generative model for the internal data is

$$[Y_i \mid \boldsymbol{M}_{i\cdot}, A_i, \boldsymbol{C}_{i\cdot}] \sim N\Big(\boldsymbol{M}_{i\cdot}\boldsymbol{\beta}_m + A_i\beta_a + \boldsymbol{C}_{i\cdot}\boldsymbol{\beta}_c, \sigma_e^2\Big), \tag{3.2.1}$$

$$[\boldsymbol{M}_{i\cdot}^\top \mid \boldsymbol{A}_i, \boldsymbol{C}_i] \sim N\Big(A_i\boldsymbol{\alpha}_a + \boldsymbol{\alpha}_c\boldsymbol{C}_i^\top, \boldsymbol{\Sigma}_m\Big), \;\; i = 1, \ldots, n. \tag{3.2.2}$$

We refer to (3.2.1) as the outcome model and (3.2.2) as the mediator model. Note that integrating



(a) External Total Effect Model

(b) Internal Mediation Model

Figure 3.1: Directed Acyclic Graph (DAG) for internal and external mediation models where $\boldsymbol{A}$ denotes the exposure, $\boldsymbol{M}$ denotes the collection of p_m candidate mediators, $\boldsymbol{Y}$ denotes the outcome, $\boldsymbol{C}_1$ denotes the outcome-exposure confounders, and $\boldsymbol{C}_2$ denotes the outcome-mediator confounders conditional on exposure. Throughout the chapter, $\boldsymbol{C} \supseteq \boldsymbol{C}_1 \cup \boldsymbol{C}_2$. Here, $\boldsymbol{C}_E$ will likely have some overlap with $\boldsymbol{C}_1$, but the total effect model adjustment sets for the external and internal models are not assumed to be the same.

out the mediators in (3.2.1) and (3.2.2) yields:

$$[Y_i \mid A_i, \boldsymbol{C}_{i\cdot}] \sim N\Big(A_i\theta_a^I + \boldsymbol{C}_{i\cdot}\boldsymbol{\theta}_c, \sigma_t^2\Big), \;\; i = 1, \ldots, n, \tag{3.2.3}$$

where $\theta_a^I = \beta_a + \boldsymbol{\alpha}_a^\top \boldsymbol{\beta}_m$, $\boldsymbol{\theta}_c = \boldsymbol{\beta}_c + \boldsymbol{\alpha}_c^\top \boldsymbol{\beta}_m$, and $\sigma_t^2 = \sigma_e^2 + \boldsymbol{\beta}_m^\top \boldsymbol{\Sigma}_m \boldsymbol{\beta}_m$. We refer to (3.2.3) as the internal TE model.

For the external study, we assume that we have summary-level information on the TE, θ_a^I, in the form of a point estimate $\widehat{\theta}_a^E$ and an associated measure of uncertainty $\widehat{\text{Var}}(\widehat{\theta}_a^E)$ based on a sample size of n_E ($n \ll n_E$). Furthermore, we assume that we do not have access to the individual-level data from the external data source. The main objective of this chapter is to leverage the available external summary-level information, $\widehat{\theta}_a^E$ and $\widehat{\text{Var}}(\widehat{\theta}_a^E)$, to improve estimation of the NDE and the NIE in the internal study.

3.2.2 Identification of Causal Effects

In the potential outcomes framework, $\boldsymbol{M}_{i\cdot}^\top(a)$ is the counterfactual value of the mediator vector had the exposure been equal to a and $Y_i(a, \boldsymbol{m})$ is the counterfactual outcome had the exposure been equal to a and had the candidate mediator vector been equal to $\boldsymbol{m}$. Combining these two counterfactual quantities, $Y_i(a, \boldsymbol{M}_{i\cdot}^\top(a))$ is the potential outcome for exposure level a and the TE, which quantifies how the exposure marginally impacts the counterfactual outcome in the internal population, is defined as $Y_i(a, \boldsymbol{M}_{i\cdot}^\top(a)) - Y_i(a^*, \boldsymbol{M}_{i\cdot}^\top(a^*))$, where the exposure changes from the

reference level a^* to a. The NDE and NIE are obtained by decomposing the TE as follows:

$$\begin{aligned}\text{TE}_i &= Y_i(a, \boldsymbol{M}_{i\cdot}^\top(a)) - Y_i(a^*, \boldsymbol{M}_{i\cdot}^\top(a^*)) \\ &= Y_i(a, \boldsymbol{M}_{i\cdot}^\top(a)) - Y_i(a, \boldsymbol{M}_{i\cdot}^\top(a^*)) + Y_i(a, \boldsymbol{M}_{i\cdot}^\top(a^*)) - Y_i(a^*, \boldsymbol{M}_{i\cdot}^\top(a^*)) \\ &= \text{NIE}_i + \text{NDE}_i.\end{aligned}$$

The NDE quantifies how the potential outcome changes as a function of the exposure level subject to identical realizations of the reference level mediator values. Conversely, the NIE quantifies how the potential outcome changes as a function of the counterfactual mediators subject to identical realizations of the comparison exposure value. The conditional independence assumptions required for identification of the average NDE and NIE from observed data are: (i) $Y_i(a, \boldsymbol{m}) \perp A_i \mid \boldsymbol{C}_i$, (ii) $Y_i(a, \boldsymbol{m}) \perp \boldsymbol{M}_{i\cdot}^\top \mid \{A_i, \boldsymbol{C}_i\}$, (iii) $\boldsymbol{M}_{i\cdot}^\top(a) \perp A_i \mid \boldsymbol{C}_i$, and (iv) $Y_i(a, \boldsymbol{m}) \perp \boldsymbol{M}_{i\cdot}^\top(a^*) \mid \boldsymbol{C}_i$ (for a detailed exposition see Song et al. (2020)). We will assume that (i)-(iv) hold for the internal study. Under these assumptions:

$$\begin{aligned}\text{NDE} &= E[Y_i(a, \boldsymbol{M}_{i\cdot}^\top(a^*)) - Y_i(a^*, \boldsymbol{M}_{i\cdot}^\top(a^*)) \mid \boldsymbol{C}_i] = \beta_a(a - a^*) \\ \text{NIE} &= E[Y_i(a, \boldsymbol{M}_{i\cdot}^\top(a)) - Y_i(a, \boldsymbol{M}_{i\cdot}^\top(a^*)) \mid \boldsymbol{C}_i] = \boldsymbol{\alpha}_a^\top \boldsymbol{\beta}_m(a - a^*) \\ \text{TE} &= E[Y_i(a, \boldsymbol{M}_{i\cdot}^\top(a)) - Y_i(a^*, \boldsymbol{M}_{i\cdot}^\top(a^*)) \mid \boldsymbol{C}_i] = (\beta_a + \boldsymbol{\alpha}_a^\top \boldsymbol{\beta}_m)(a - a^*) = \theta_a^I(a - a^*)\end{aligned}$$

For observation i' in the external study ($i' = 1, \ldots, n_E$), we define $Y_{E,i'}$ as the observed outcome, $A_{E,i'}$ as the observed exposure, and $\boldsymbol{C}_{E,i'\cdot}$ as the observed confounder vector, which may or may not be the same as the set of confounders in the internal study. To incorporate external summary-level information on the TE, certain methods we present in this chapter require the following transportability condition:

$$\begin{aligned}&E[Y_i \mid A_i = a, \boldsymbol{C}_{i\cdot} = \boldsymbol{c}] - E[Y_i \mid A_i = a^*, \boldsymbol{C}_{i\cdot} = \boldsymbol{c}] \\ &\qquad = E[Y_{E,i'} \mid A_{E,i'} = a, \boldsymbol{C}_{E,i'\cdot} = \boldsymbol{c}_E] - E[Y_{E,i'} \mid A_{E,i'} = a^*, \boldsymbol{C}_{E,i'\cdot} = \boldsymbol{c}_E] \quad (3.2.4)\end{aligned}$$

for all possible realizations of a and a^*. Transportability condition (3.2.4) in our context ensures that $\theta_a^I = \theta_a^E$, where θ_a^E is the true TE in the external population.

3.2.3 Maximum Likelihood Estimation without External Information

To establish an inferential baseline, consider a mediation analysis that ignores available external summary-level information on the TE. That is, the model specification is (3.2.1) and (3.2.2), which we call the unconstrained model. The maximum likelihood estimator (MLE) with respect to model

specification (3.2.1) and (3.2.2) is defined as

$$\underset{\boldsymbol{\alpha}_a,\boldsymbol{\alpha}_c,\boldsymbol{\Sigma}_m,\beta_a,\boldsymbol{\beta}_m,\boldsymbol{\beta}_c,\sigma_e^2}{\arg\min}\left\{\frac{np_m}{2}\log(|\boldsymbol{\Sigma}_m|)+\frac{1}{2}\sum_{i=1}^{n}\Big(\boldsymbol{M}_{i\cdot}^\top-A_i\boldsymbol{\alpha}_a-\boldsymbol{\alpha}_c\boldsymbol{C}_{i\cdot}^\top\Big)^\top\boldsymbol{\Sigma}_m^{-1}\Big(\boldsymbol{M}_{i\cdot}^\top-A_i\boldsymbol{\alpha}_a-\boldsymbol{\alpha}_c\boldsymbol{C}_{i\cdot}^\top\Big)\right.$$
$$\left.+\frac{n}{2}\log(\sigma_e^2)+\frac{1}{2\sigma_e^2}\Big(\boldsymbol{Y}-\boldsymbol{A}\beta_a-\boldsymbol{M}\boldsymbol{\beta}_m-\boldsymbol{C}\boldsymbol{\beta}_c\Big)^\top\Big(\boldsymbol{Y}-\boldsymbol{A}\beta_a-\boldsymbol{M}\boldsymbol{\beta}_m-\boldsymbol{C}\boldsymbol{\beta}_c\Big)\right\} \quad (3.2.5)$$

The MLE is denoted as $(\widehat{\boldsymbol{\alpha}}_a^U,\widehat{\boldsymbol{\alpha}}_c^U,\widehat{\boldsymbol{\Sigma}}_m^U,\widehat{\beta}_a^U,\widehat{\boldsymbol{\beta}}_m^U,\widehat{\boldsymbol{\beta}}_c^U,\{\widehat{\sigma}_e^U\}^2)$, where $\widehat{\boldsymbol{\alpha}}_a^U$ is the MLE of $\boldsymbol{\alpha}_a$, $\widehat{\beta}_a^U$ is the MLE of β_a, and $\widehat{\boldsymbol{\beta}}_m^U$ is the MLE of $\boldsymbol{\beta}_m$. For the unconstrained model, the MLEs have closed-form expressions (see Supplementary Material B.1). Going forward, $\widehat{\text{NDE}}^U=\widehat{\beta}_a^U$ and $\widehat{\text{NIE}}^U=\{\widehat{\boldsymbol{\alpha}}_a^U\}^\top\widehat{\boldsymbol{\beta}}_m^U$ are called the unconstrained estimators of the NDE and NIE, respectively.

3.2.4 Maximum Likelihood Estimation with Congenial External Information

Next, assume that there is an external point estimate, $\widehat{\theta}_a^E$, such that $\widehat{\theta}_a^E$ is a consistent estimate of θ_a^I. This corresponds to the situation where transportability assumption (3.2.4) holds. When $\widehat{\theta}_a^E$ is a consistent estimate of θ_a^I, we say that the external information is congenial with the internal study population. Consider the optimization problem:

$$\underset{\boldsymbol{\alpha}_a,\boldsymbol{\alpha}_c,\boldsymbol{\Sigma}_m,\boldsymbol{\beta}_m,\boldsymbol{\beta}_c\sigma_e^2}{\arg\min}\left\{\frac{np_m}{2}\log(|\boldsymbol{\Sigma}_m|)+\frac{1}{2}\sum_{i=1}^{n}\Big(\boldsymbol{M}_{i\cdot}^\top-A_i\boldsymbol{\alpha}_a-\boldsymbol{\alpha}_c\boldsymbol{C}_{i\cdot}^\top\Big)^\top\boldsymbol{\Sigma}_m^{-1}\Big(\boldsymbol{M}_{i\cdot}^\top-A_i\boldsymbol{\alpha}_a-\boldsymbol{\alpha}_c\boldsymbol{C}_{i\cdot}^\top\Big)\right.$$
$$\left.+\frac{n}{2}\log(\sigma_e^2)+\frac{1}{2\sigma_e^2}\Big(\boldsymbol{Y}-\boldsymbol{A}\beta_a^E-\boldsymbol{M}\boldsymbol{\beta}_m-\boldsymbol{C}\boldsymbol{\beta}_c\Big)^\top\Big(\boldsymbol{Y}-\boldsymbol{A}\beta_a^E-\boldsymbol{M}\boldsymbol{\beta}_m-\boldsymbol{C}\boldsymbol{\beta}_c\Big)\right\} \quad (3.2.6)$$

where $\beta_a^E=\widehat{\theta}_a^E-\boldsymbol{\alpha}_a^\top\boldsymbol{\beta}_m$. Alternatively, (3.2.6) can be viewed as a minimization over the negative log-likelihood corresponding to the following model specification, which we call the hard constraint model:

$$\big[Y_i\mid\boldsymbol{M}_{i\cdot},A_i,\boldsymbol{C}_{i\cdot}\big]\sim N\big(\boldsymbol{M}_{i\cdot}\boldsymbol{\beta}_m+A_i\{\widehat{\theta}_a^E-\boldsymbol{\alpha}_a^\top\boldsymbol{\beta}_m\}+\boldsymbol{C}_{i\cdot}\boldsymbol{\beta}_c,\sigma_e^2\big)$$
$$\big[\boldsymbol{M}_{i\cdot}^\top\mid A_i,\boldsymbol{C}_{i\cdot}\big]\sim N(A_i\boldsymbol{\alpha}_a+\boldsymbol{\alpha}_c\boldsymbol{C}_{i\cdot}^\top,\boldsymbol{\Sigma}_m),\ \ i=1,\ldots,n.$$

We denote the optimizer of (3.2.6) as $(\widehat{\boldsymbol{\alpha}}_a^H,\widehat{\boldsymbol{\alpha}}_c^H,\widehat{\boldsymbol{\Sigma}}_m^H,\widehat{\boldsymbol{\beta}}_m^H,\widehat{\boldsymbol{\beta}}_c^H,\{\widehat{\sigma}_e^H\}^2)$, where $\widehat{\boldsymbol{\alpha}}_a^H$ is the estimator of $\boldsymbol{\alpha}_a$ and $\widehat{\boldsymbol{\beta}}_m^H$ is the estimator of $\boldsymbol{\beta}_m$. The purpose of (3.2.6) is to impose a hard constraint on TE estimation so that the estimated TE is always equal to $\{\widehat{\beta}_a^E\}^H+\{\widehat{\boldsymbol{\alpha}}_a^H\}^\top\widehat{\boldsymbol{\beta}}_m^H=\widehat{\theta}_a^E$. Cyclical coordinate descent is used to compute the optimizer of (3.2.6) (see Supplementary Material B.1).

The approach described in this section represents the other extreme compared to the uncon-

strained estimator, where the estimated TE is forced to be equal to $\widehat{\theta}_a^E$, showing exact and extreme faith in the external information. Moreover, the hard constraint on the estimated TE induces information sharing between the internal mediator and outcome models through the $\boldsymbol{\alpha}_a^\top \boldsymbol{\beta}_m$ term in the mean function of the outcome model. Going forward, we refer to $\widehat{\text{NDE}}^H = \widehat{\theta}_a^E - \left\{\widehat{\boldsymbol{\alpha}}_a^H\right\}^\top \widehat{\boldsymbol{\beta}}_m^H$ and $\widehat{\text{NIE}}^H = \left\{\widehat{\boldsymbol{\alpha}}_a^H\right\}^\top \widehat{\boldsymbol{\beta}}_m^H$ as the hard constraint estimators of the NDE and NIE, respectively.

3.2.5 Robust Soft Constraint Estimator

The final method considers the case where $\widehat{\theta}_a^E$ may or may not be a consistent estimate of θ_a^I. That is, the validity of transportability assumption (3.2.4) is unknown. When $\widehat{\theta}_a^E$ is an inconsistent estimate of θ_a^I, we say that the external information is incongenial with the internal study. Transportability assumption (3.2.4) may be violated for a variety of reasons, including fundamentally different $\boldsymbol{Y} \mid \boldsymbol{A}, \boldsymbol{C}$ distributions in the external and internal populations, unmeasured confounding in the external TE model, and differing adjustment sets between the external and internal TE models. To address potential violations of (3.2.4), we treat the internal TE parameter as a random effect, $\widetilde{\theta}_a^I$, and define a random effect mediation model, which we call the soft constraint model:

$$
\begin{aligned}
[Y_i \mid \boldsymbol{M}_{i\cdot}, A_i, \boldsymbol{C}_{i\cdot}, \widetilde{\theta}_a^I] &\sim N\Big(\boldsymbol{M}_{i\cdot}\boldsymbol{\beta}_m + A_i\{\widetilde{\theta}_a^I - \boldsymbol{\alpha}_a^\top \boldsymbol{\beta}_m\} + \boldsymbol{C}_{i\cdot}\boldsymbol{\beta}_c, \sigma_e^2\Big) \\
[\boldsymbol{M}_{i\cdot}^\top \mid A_i, \boldsymbol{C}_i] &\sim N\Big(A_i\boldsymbol{\alpha}_a + \boldsymbol{\alpha}_c\boldsymbol{C}_i^\top, \boldsymbol{\Sigma}_m\Big), \;\; i = 1, \ldots, n \\
\widetilde{\theta}_a^I &\sim N\Big(\widehat{\theta}_a^E, s^2\widehat{\text{Var}}(\widehat{\theta}_a^E)\Big)
\end{aligned}
$$

It is important to clarify that the soft constraint model is a working model and the true generative model of the internal data remains (3.2.1) and (3.2.2). The advantage of a random effects formulation is that it allows for shrinkage towards the external information without imposing inflexible hard constraints on the estimated TE. After integrating out $\widetilde{\theta}_a^I$ the soft constraint likelihood function becomes

$$
L(\boldsymbol{\alpha}_a, \boldsymbol{\alpha}_c, \boldsymbol{\Sigma}_M, \boldsymbol{\beta}_m, \boldsymbol{\beta}_c, \sigma_e^2 \mid \boldsymbol{Y}, \boldsymbol{M}, \boldsymbol{A}, \boldsymbol{C}) = \int_{-\infty}^{\infty} \pi(\boldsymbol{Y} \mid \boldsymbol{M}, \boldsymbol{A}, \boldsymbol{C}, \widetilde{\theta}_a^I)\pi(\boldsymbol{M} \mid \boldsymbol{A}, \boldsymbol{C})\pi(\widetilde{\theta}_a^I)d\widetilde{\theta}_a^I,
$$

where π is general notation for a probability density function. The maximum likelihood estimators, defined by

$$
\underset{\boldsymbol{\alpha}_a, \boldsymbol{\alpha}_c, \boldsymbol{\Sigma}_m, \boldsymbol{\beta}_m, \boldsymbol{\beta}_c, \sigma_e^2}{\arg\max} L(\boldsymbol{\alpha}_a, \boldsymbol{\alpha}_c, \boldsymbol{\Sigma}_M, \boldsymbol{\beta}_m, \boldsymbol{\beta}_c, \sigma_e^2 \mid \boldsymbol{Y}, \boldsymbol{M}, \boldsymbol{A}, \boldsymbol{C}), \tag{3.2.7}
$$

are denoted as $(\widehat{\boldsymbol{\alpha}}_a^S, \widehat{\boldsymbol{\alpha}}_c^S, \widehat{\boldsymbol{\Sigma}}_M^S, \widehat{\boldsymbol{\beta}}_m^S, \widehat{\boldsymbol{\beta}}_c^S, \left\{\widehat{\sigma}_e^S\right\}^2)$ and the soft constraint estimator of the NIE is $\widehat{\text{NIE}}^S = \{\widehat{\boldsymbol{\alpha}}_a^S\}^\top \widehat{\boldsymbol{\beta}}_m^S$. The soft constraint estimator of the TE is the posterior mean estimator corresponding to the posterior distribution $\pi(\widetilde{\theta}_a^I \mid \boldsymbol{Y}, \boldsymbol{M}, \boldsymbol{A}, \boldsymbol{C})$, with the maximum likelihood

estimators, $\widehat{\boldsymbol{\alpha}}_a^S$, $\widehat{\boldsymbol{\alpha}}_c^S$, $\widehat{\boldsymbol{\Sigma}}_M^S$, $\widehat{\boldsymbol{\beta}}_m^S$, $\widehat{\boldsymbol{\beta}}_c^S$, and $\left\{\widehat{\sigma}_e^S\right\}^2$, substituted in for their corresponding true parameter values (Verbeke and Molenberghs, 2000). The resultant soft constraint estimator for the NDE is the difference between the soft constraint TE and NIE estimators:

$$\widehat{\text{NDE}}^S = \left[\frac{\boldsymbol{A}^\top \boldsymbol{A}}{\left\{\widehat{\sigma}_e^S\right\}^2} + \frac{1}{s^2\widehat{Var}(\widehat{\theta}_a^E)}\right]^{-1}\left[\frac{\boldsymbol{A}^\top(\boldsymbol{Y} - \boldsymbol{C}\widehat{\boldsymbol{\beta}}_c^S - \boldsymbol{M}\widehat{\boldsymbol{\beta}}_m^S)}{\left\{\widehat{\sigma}_e^S\right\}^2} + \frac{(\widehat{\theta}_a^E - \{\widehat{\boldsymbol{\alpha}}_a^S\}^\top\widehat{\boldsymbol{\beta}}_m^S)}{s^2\widehat{Var}(\widehat{\theta}_a^E)}\right]$$

For the soft constraint model, the Expectation-Maximization (EM) algorithm is used to solve (3.2.7), where $\boldsymbol{Y}$, $\boldsymbol{M}$, $\boldsymbol{A}$, and $\boldsymbol{C}$ are treated as the observed data and $\widetilde{\theta}_a^I$ is treated as the unobserved latent data (Dempster et al., 1977). See Supplementary Material B.1 for details on the EM algorithm implementation.

3.3 Asymptotic Efficiency Results

The goals of this section are to understand the efficiency gain attributable to incorporating congenial external information on the TE and to provide commentary on dealing with potentially incongenial external information. Here, $\sigma_a^2 = \text{Var}(A_i \mid \boldsymbol{C}_{i\cdot})$, which is obtained by regressing out the confounders from the exposure using a linear regression model.

3.3.1 Asymptotic Distributions of the Unconstrained and Hard Constraint Estimators

Theorem 3.3.1. *The joint asymptotic distribution of $\widehat{\boldsymbol{\alpha}}_a^U$, $\widehat{\beta}_a^U$, and $\widehat{\boldsymbol{\beta}}_m^U$ is,*

$$\sqrt{n}\begin{pmatrix}\widehat{\boldsymbol{\alpha}}_a^U - \boldsymbol{\alpha}_a \\ \widehat{\beta}_a^U - \beta_a \\ \widehat{\boldsymbol{\beta}}_m^U - \boldsymbol{\beta}_m\end{pmatrix} \to_d N\left(\mathbf{0}, \left\{\mathcal{I}^U(\boldsymbol{\alpha}_a, \beta_a, \boldsymbol{\beta}_m)\right\}^{-1}\right)$$

$$\left\{\mathcal{I}^U(\boldsymbol{\alpha}_a, \beta_a, \boldsymbol{\beta}_m)\right\}^{-1} = \begin{pmatrix}\frac{1}{\sigma_a^2}\boldsymbol{\Sigma}_m & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \frac{\sigma_e^2}{\sigma_a^2}\left(1 + \sigma_a^2\boldsymbol{\alpha}_a^\top\boldsymbol{\Sigma}_m^{-1}\boldsymbol{\alpha}_a\right) & -\sigma_e^2\boldsymbol{\alpha}_a^\top\boldsymbol{\Sigma}_m^{-1} \\ \mathbf{0} & -\sigma_e^2\boldsymbol{\Sigma}_m^{-1}\boldsymbol{\alpha}_a & \sigma_e^2\boldsymbol{\Sigma}_m^{-1}\end{pmatrix}$$

Let NDE $= \beta_a$, NIE $= \boldsymbol{\alpha}_a^\top\boldsymbol{\beta}_m$, TE $=$ NDE $+$ NIE, and $\widehat{TE}^U = \widehat{NDE}^U + \widehat{NIE}^U$. Then,

$$\sqrt{n}(\widehat{NDE}^U - NDE) \to_d N\left(0, \frac{\sigma_e^2}{\sigma_a^2} + \sigma_e^2\boldsymbol{\alpha}_a^\top\boldsymbol{\Sigma}_m^{-1}\boldsymbol{\alpha}_a\right),$$

$$\sqrt{n}(\widehat{TE}^{U} - TE) \to_d N\left(0, \frac{1}{\sigma_a^2}\left\{\sigma_e^2 + \boldsymbol{\beta}_m^\top \boldsymbol{\Sigma}_m \boldsymbol{\beta}_m\right\}\right),$$

and, provided that $\boldsymbol{\alpha}_a \neq \mathbf{0}$ *or* $\boldsymbol{\beta}_m \neq \mathbf{0}$*,*

$$\sqrt{n}(\widehat{NIE}^{U} - NIE) \to_d N\left(0, \frac{1}{\sigma_a^2}\boldsymbol{\beta}_m^\top \boldsymbol{\Sigma}_m \boldsymbol{\beta}_m + \sigma_e^2 \boldsymbol{\alpha}_a^\top \boldsymbol{\Sigma}_m^{-1} \boldsymbol{\alpha}_a\right).$$

Proof. See Supplementary Material B.2 for details.

Theorem 3.3.2. *Suppose that* $\widehat{\theta}_a^E \to_p \theta_a^I$ *as* $n_E \to \infty$*. Then,*

$$\sqrt{n}\begin{pmatrix} \widehat{\boldsymbol{\alpha}}_a^H - \boldsymbol{\alpha}_a \\ \widehat{\boldsymbol{\beta}}_m^H - \boldsymbol{\beta}_m \end{pmatrix} \to_d N\left(\mathbf{0}, \left\{\mathcal{I}^H(\boldsymbol{\alpha}_a, \boldsymbol{\beta}_m)\right\}^{-1}\right)$$

$$\left\{\mathcal{I}^H(\boldsymbol{\alpha}_a, \boldsymbol{\beta}_m)\right\}^{-1} = \begin{pmatrix} \frac{1}{\sigma_a^2}\left(\boldsymbol{\Sigma}_m^{-1} + \frac{1}{\sigma_e^2}\boldsymbol{\beta}_m\boldsymbol{\beta}_m^\top\right)^{-1} & \mathbf{0} \\ \mathbf{0} & \sigma_e^2 \boldsymbol{\Sigma}_m^{-1} \end{pmatrix}$$

Let $NDE = \theta_a^I - \boldsymbol{\alpha}_a^\top \boldsymbol{\beta}_m$*,* $NIE = \boldsymbol{\alpha}_a^\top \boldsymbol{\beta}_m$*,* $TE = \theta_a^I$*, and* $\widehat{TE}^H = \widehat{\theta}_a^E$*. Then, provided that* $\boldsymbol{\alpha}_a \neq \mathbf{0}$ *or* $\boldsymbol{\beta}_m \neq \mathbf{0}$*,*

$$\sqrt{n}(\widehat{NDE}^H - NDE) \to_d N\left(0, \frac{\sigma_e^2}{\sigma_a^2} R^2_{\boldsymbol{M}|\boldsymbol{A},\boldsymbol{C}} + \sigma_e^2 \boldsymbol{\alpha}_a^\top \boldsymbol{\Sigma}_m^{-1} \boldsymbol{\alpha}_a\right)$$

$$\sqrt{n}(\widehat{NIE}^H - NIE) \to_d N\left(0, \frac{1}{\sigma_a^2}\boldsymbol{\beta}_m^\top \boldsymbol{\Sigma}_m \boldsymbol{\beta}_m \left\{1 - R^2_{\boldsymbol{M}|\boldsymbol{A},\boldsymbol{C}}\right\} + \sigma_e^2 \boldsymbol{\alpha}_a^\top \boldsymbol{\Sigma}_m^{-1} \boldsymbol{\alpha}_a\right),$$

where

$$R^2_{\boldsymbol{M}|\boldsymbol{A},\boldsymbol{C}} = \frac{\boldsymbol{\beta}_m^\top \boldsymbol{\Sigma}_m \boldsymbol{\beta}_m}{\sigma_e^2 + \boldsymbol{\beta}_m^\top \boldsymbol{\Sigma}_m \boldsymbol{\beta}_m}$$

Proof. See Supplementary Material B.2 for details.

Theorems 3.3.1 and 3.3.2 show the asymptotic distributions of the unconstrained and hard constraint estimators, respectively. The inverted information matrices clarify that efficiency gains for NIE estimation exclusively come from improved estimation of $\boldsymbol{\alpha}_a$. When the TE and outcome models are framed as nested regression models, this result is consistent with Gu et al. (2019), which showed that leveraging external summary-level information from a reduced model only results in efficiency gains for the regression coefficients corresponding to regressors in common between the two models. Theorems 3.3.1 and 3.3.2 also show that $\boldsymbol{\beta}_m = \mathbf{0}$ implies no efficiency gain for NIE estimation. When $\boldsymbol{\beta}_m \neq 0$, the absolute efficiency gain corresponding to the NDE and NIE is completely dependent on the quantity, $R^2_{\boldsymbol{M}|\boldsymbol{A},\boldsymbol{C}}$, the partial R^2 between the TE and outcome models. If $R^2_{\boldsymbol{M}|\boldsymbol{A},\boldsymbol{C}} \approx 1$, then the inclusion of candidate mediators substantially improves model fit and consequently there are large gains for NIE estimation. Conversely, if $R^2_{\boldsymbol{M}|\boldsymbol{A},\boldsymbol{C}} \approx 0$,

then inclusion of candidate mediators do not improve model fit compared to the TE model and consequently there are large gains for NDE estimation. An intuitive understanding of the latter point comes when $R^2_{\boldsymbol{M}|\boldsymbol{A},\boldsymbol{C}} = 0$, which implies that $\boldsymbol{\beta}_m = \mathbf{0}$ and TE = NDE. That is, the smaller $||\boldsymbol{\beta}_m||_2$ is, the closer the TE is to the NDE, and the external information on the TE becomes increasingly more relevant for NDE estimation. With respect to NDE and NIE estimation, a small value of $\sigma_e^2 \boldsymbol{\alpha}_a^\top \boldsymbol{\Sigma}_m^{-1} \boldsymbol{\alpha}_a$, a scaling of the signal-to-noise ratio in the mediator model, leads to larger relative efficiency gains.

3.3.2 Asymptotic Distribution when $\boldsymbol{\alpha}_a = \boldsymbol{\beta}_m = \mathbf{0}$

Theorem 3.3.3 clarifies the asymptotic distribution of the unconstrained and hard constraint estimators when $\boldsymbol{\alpha}_a = \boldsymbol{\beta}_m = \mathbf{0}$. There are no efficiency gains for NIE estimation in this setting because $\boldsymbol{\beta}_m = \mathbf{0}$.

Theorem 3.3.3. *Suppose that $\boldsymbol{\alpha}_a = \boldsymbol{\beta}_m = \mathbf{0}$ and that there are estimators $\widehat{\boldsymbol{\alpha}}_a$ and $\widehat{\boldsymbol{\beta}}_m$ of $\boldsymbol{\alpha}_a$ and $\boldsymbol{\beta}_m$, respectively, which satisfy:*

$$\sqrt{n}\begin{pmatrix} \widehat{\boldsymbol{\alpha}}_a - \boldsymbol{\alpha}_a \\ \widehat{\boldsymbol{\beta}}_m - \boldsymbol{\beta}_m \end{pmatrix} \to_d N\left(\begin{pmatrix} \mathbf{0} \\ \mathbf{0} \end{pmatrix}, \begin{pmatrix} \sigma_a^{-2}\boldsymbol{\Sigma}_m & \mathbf{0} \\ \mathbf{0} & \sigma_e^2 \boldsymbol{\Sigma}_m^{-1} \end{pmatrix} \right)$$

Then,

$$n\Big(\widehat{\boldsymbol{\alpha}}_a^\top \widehat{\boldsymbol{\beta}}_m - \boldsymbol{\alpha}_a^\top \boldsymbol{\beta}_m\Big) \to_d \frac{1}{2}\sqrt{\frac{\sigma_e^2}{\sigma_a^2}}\big(\xi_1 - \xi_2\big),$$

where ξ_1 and ξ_2 are independent $\chi^2_{p_m}$ random variables.

Proof. See Supplementary Material B.2 for details.

Theorems 3.3.1, 3.3.2, and 3.3.3 can be used to inform the construction of hypothesis tests and confidence intervals for the NIE and NDE. However, when determining the reference distribution, how much to weight the $\boldsymbol{\alpha}_a = \boldsymbol{\beta}_m = \mathbf{0}$ case relative to the $\boldsymbol{\alpha}_a \neq \mathbf{0}$ or $\boldsymbol{\beta}_m \neq \mathbf{0}$ case is unknown. One straightforward workaround is to check whether or not $\boldsymbol{\alpha}_a = \boldsymbol{\beta}_m = \mathbf{0}$ by using a Wald test with the reference distribution given by

$$n\begin{pmatrix} \widehat{\boldsymbol{\alpha}}_a^\top & \widehat{\boldsymbol{\beta}}_m^\top \end{pmatrix} \begin{pmatrix} \sigma_a^{-2}\boldsymbol{\Sigma}_m & \mathbf{0} \\ \mathbf{0} & \sigma_e^2 \boldsymbol{\Sigma}_m^{-1} \end{pmatrix}^{-1} \begin{pmatrix} \widehat{\boldsymbol{\alpha}}_a \\ \widehat{\boldsymbol{\beta}}_m \end{pmatrix} \to_d \chi^2_{2p_m}$$

and then use the appropriate asymptotic result to construct confidence intervals.

3.3.3 Robustness to Incongenial External Information

Theorem 3.3.2 focuses on settings where transportability condition (3.2.4) holds, however there are likely many instances where fundamental differences across internal and external study populations lead to violations of (3.2.4). The desire to account for such cases motivates an estimator that is robust to departures from (3.2.4), but is still more efficient than the unconstrained estimator when (3.2.4) is satisfied. In this context, robustness refers to the fact that the estimator is as asymptotically efficient as the unconstrained estimator when (3.2.4) does not hold. Theorem 3.3.4 establishes that the soft constraint estimator is as efficient or more efficient than the unconstrained estimator with respect to NIE estimation when (3.2.4) does hold.

Theorem 3.3.4. *Suppose that* $ns^2\widehat{Var}(\widehat{\theta}_a^E) \to_p \tau_a^2$*, where* $\tau_a^2 \in (0, \infty)$*, and* $\widehat{\theta}_a^E \to_p \theta_a^I$ *as* $n \to \infty$ *and* $n_E \to \infty$*. Then,*

$$\sqrt{n}\begin{pmatrix} \widehat{\boldsymbol{\alpha}}_a^S - \boldsymbol{\alpha}_a \\ \widehat{\boldsymbol{\beta}}_m^S - \boldsymbol{\beta}_m \end{pmatrix} \to_d N\left(\mathbf{0}, \left\{\mathcal{I}^S(\boldsymbol{\alpha}_a, \boldsymbol{\beta}_m)\right\}^{-1}\right)$$

$$\left\{\mathcal{I}^S(\boldsymbol{\alpha}_a, \boldsymbol{\beta}_m)\right\}^{-1} = \begin{pmatrix} \frac{1}{\sigma_a^2}\left(\boldsymbol{\Sigma}_m^{-1} + \frac{1}{\tau_a^2}\left[\frac{\sigma_a^2}{\sigma_e^2} + \frac{1}{\tau_a^2}\right]^{-1}\frac{1}{\sigma_e^2}\boldsymbol{\beta}_m\boldsymbol{\beta}_m^\top\right)^{-1} & \mathbf{0} \\ \mathbf{0} & \sigma_e^2\boldsymbol{\Sigma}_m^{-1} \end{pmatrix}$$

Let NIE $= \boldsymbol{\alpha}_a^\top\boldsymbol{\beta}_m$*. Then, provided that* $\boldsymbol{\alpha}_a \neq \mathbf{0}$ *or* $\boldsymbol{\beta}_m \neq \mathbf{0}$*,*

$$\sqrt{n}(\widehat{NIE}^S - NIE) \to_d N\left(0, \frac{1}{\sigma_a^2}\boldsymbol{\beta}_m^\top\boldsymbol{\Sigma}_m\boldsymbol{\beta}_m\left[1 + \frac{1}{\tau_a^2}\left(\frac{\sigma_a^2}{\sigma_e^2} + \frac{1}{\tau_a^2}\right)^{-1}\frac{1}{\sigma_e^2}\boldsymbol{\beta}_m^\top\boldsymbol{\Sigma}_m\boldsymbol{\beta}_m\right]^{-1} + \sigma_e^2\boldsymbol{\alpha}_a^\top\boldsymbol{\Sigma}_m^{-1}\boldsymbol{\alpha}_a\right)$$

Proof. See Supplementary Material B.2 for details.

Theorem 3.3.4 provides the asymptotic distribution of the soft constraint estimator of the NIE. The asymptotic variance-covariance matrix converges to the asymptotic variance-covariance matrix of the unconstrained estimator if $s^2 \to \infty$ and the asymptotic variance-covariance matrix of the hard constraint estimator if $s^2 \to 0$. Additionally, the conclusions of Theorem 3.3.3 hold for the soft constraint estimator when $\boldsymbol{\alpha}_a = \boldsymbol{\beta}_m = \mathbf{0}$. Inference corresponding to the soft constraint NDE estimator is more challenging because there is not an easily derivable asymptotic distribution. In this chapter, for interval estimation, we use quantile-based confidence intervals via the parametric bootstrap (Efron, 1982).

Although Theorem 3.3.4 is derived for a fixed value of s^2, s^2 can be data-adaptively estimated to robustify model parameter estimation from incongenial external information. We obtain a data adaptive estimator for s^2 following an empirical-Bayes argument (Morris, 1983;

Mukherjee and Chatterjee, 2008), where the MLE of the TE, denoted by $\widehat{\theta}_a^I$, has the conditional distribution $[\widehat{\theta}_a^I \mid \widetilde{\theta}_a^I] \sim N\big(\widetilde{\theta}_a^I, \text{Var}(\widehat{\theta}_a^I)\big)$, coupled with the random effect distribution $\widetilde{\theta}_a^I \sim N\Big(\widehat{\theta}_a^E, s^2\widehat{\text{Var}}(\widehat{\theta}_a^E)\Big)$. Maximizing the marginal likelihood after integrating out $\widetilde{\theta}_a^I$ yields $\widehat{s}^2 = \big\{\widehat{\text{Var}}(\widehat{\theta}_a^E)\big\}^{-1} \max\{0, (\widehat{\theta}_a^I - \widehat{\theta}_a^E)^2 - \widehat{\text{Var}}(\widehat{\theta}_a^I)\}$. That is, $(\widehat{\theta}_a^I - \widehat{\theta}_a^E)^2 \leq \widehat{\text{Var}}(\widehat{\theta}_a^I)$ corresponds to the hard constraint model and $(\widehat{\theta}_a^I - \widehat{\theta}_a^E)^2 > \widehat{\text{Var}}(\widehat{\theta}_a^I)$ corresponds to the soft constraint model. However, from a practical perspective, we recommend that if $\widehat{s}^2 = 0$, then set $\widehat{s}^2$ equal to a small value close to zero and use the soft constraint algorithm; this helps to avoid coverage issues for the parametric bootstrap confidence intervals of the NDE.

3.4 Simulations

3.4.1 Generative Model

The purpose of the simulation section is to empirically show estimation properties of the unconstrained, hard constraint, and soft constraint estimators. We consider one set of simulation scenarios where $\theta_a^I = \theta_a^E$ and two sets of simulation scenarios where $\theta_a^I \neq \theta_a^E$. The generative model for the internal data in all simulation settings is (3.2.1) and (3.2.2).

For the $\theta_a^I = \theta_a^E$ simulation scenarios, which we refer to as the congenial simulation scenarios, the parameters are set as follows: $p_m = 50$, $p_c = 5$, and $(A_i, \boldsymbol{C}_i^\top)^\top \sim MVN(\boldsymbol{0}, \boldsymbol{\Omega})$. Here, $\boldsymbol{\Omega}$ has an exchangeable correlation structure with the correlation parameter $\rho = 0.2$ and variance parameters equal to one. We consider two values of the internal sample size, $n = 200$ and $n = 2000$, with external sample sizes 10, 100, and 1000 times greater than the internal sample sizes. For the regression coefficient parameters in the mediator model, we fix $\boldsymbol{\alpha}_c$ so that it is a matrix of 0.1's and set $\boldsymbol{\alpha}_a = (0.6, \ldots, 0.6, 0, \ldots, 0)^\top$ where 0.6 and 0 are repeated 10 and 40 times, respectively. The error variance-covariance matrix in the mediator model, $\boldsymbol{\Sigma}_m$, has a block exchangeable correlation structure with correlation parameters within blocks set to 0.3 and the correlation parameters across blocks set to 0.2. The error variance parameters in $\boldsymbol{\Sigma}_m$ are determined by pre-specified values of $R^2_{\boldsymbol{A}|\boldsymbol{C}}$, where $R^2_{\boldsymbol{A}|\boldsymbol{C}} = (\boldsymbol{\alpha}_a)_1^2/\{\boldsymbol{\Sigma}_{m,jj} + (\boldsymbol{\alpha}_a)_1^2\}$ and $\boldsymbol{\Sigma}_{m,jj}$ is the j-th entry along the diagonal of $\boldsymbol{\Sigma}_m$. We consider two options: $R^2_{\boldsymbol{A}|\boldsymbol{C}} = 0.05$ and $R^2_{\boldsymbol{A}|\boldsymbol{C}} = 0.2$. For the regression coefficient parameters in the outcome model, $\boldsymbol{\beta}_c = (0.1, \ldots, 0.1)^\top$, $\boldsymbol{\beta}_m = (0.1, \ldots, 0.1, 0, \ldots, 0, 0.1, \ldots, 0.1, 0, \ldots, 0)^\top$, where the 0.1's are repeated 5 times and the 0's are repeated 5 and 35 times, respectively, and $\beta_a = \theta_a^I - \boldsymbol{\alpha}_a^\top \boldsymbol{\beta}_m$. Here, $\theta_a^I = 1$. The error variance in the outcome model, σ_e^2, is determined based on $R^2_{\boldsymbol{M}|\boldsymbol{A},\boldsymbol{C}}$. In this case, we consider three options: $R^2_{\boldsymbol{M}|\boldsymbol{A},\boldsymbol{C}} = 0.2$, $R^2_{\boldsymbol{M}|\boldsymbol{A},\boldsymbol{C}} = 0.5$, and $R^2_{\boldsymbol{M}|\boldsymbol{A},\boldsymbol{C}} = 0.8$. The purpose of varying $R^2_{\boldsymbol{A}|\boldsymbol{C}}$ and $R^2_{\boldsymbol{M}|\boldsymbol{A},\boldsymbol{C}}$ is because our asymptotic variance results from Section 3.3 suggest that these quantities govern the asymptotic relative efficiency gains for NDE and NIE estimation. To generate external datasets

from the external TE model, we use the same parameter values as the internal TE model but with different sample sizes of $n_E = 10n$, $n_E = 100n$, and $n_E = 1000n$. The simulated external TE estimate is then obtained by calculating the MLE on the simulated external data, $\widehat{\theta}_a^E$ and $\widehat{\text{Var}}(\widehat{\theta}_a^E)$.

For the $\theta_a^I \neq \theta_a^E$ simulation scenarios, we consider the same simulation parameters as the congenial external model simulation settings, with one exception; we either set $\theta_a^E = 2$ or randomly generate an external TE parameter such that $\theta_a^E \sim N(1, 0.1)$. The former scenario considers a case where the external population is notably different from the internal population, potentially due to transportability violations. The latter scenario considers the average performance across a distribution of possible population-level external TE realizations, including both congenial and incongenial settings with the internal TE. The goal of these simulations is to determine which methods are robust to discordant external and internal targets for the TE. We refer to the scenarios with $\theta_a^E = 2$ as incongenial settings and those with $\theta_a^E \sim N(1, 0.1)$ as random settings.

3.4.2 Comparison Methods and Evaluation Metrics

We consider three estimators in all simulation scenarios: the unconstrained estimator, the soft constraint estimator with $\widehat{s}^2 = \left\{\widehat{\text{Var}}(\widehat{\theta}_a^E)\right\}^{-1} \max\{0, (\widehat{\theta}_a^I - \widehat{\theta}_a^E)^2 - \widehat{\text{Var}}(\widehat{\theta}_a^I)\}$, and the hard constraint estimator. For brevity, we refer to the soft constraint estimator as the soft constraint empirical-Bayes (EB) estimator. As a benchmark for the maximal possible efficiency gain attainable from leveraging external information on the TE, we also consider the hard constraint estimator with the true θ_a^I enforced as the hard constraint on the TE, although this is not implementable in practice. We evaluate these estimators based on their relative root mean-squared error (RMSE) for NDE and NIE estimation compared with their unconstrained equivalents. Note that in the random simulation settings this is a root integrated mean-squared error (RIMSE) metric, where the mean-squared error is integrated over the generative distribution of θ_a^E. Moreover, we evaluate the empirical coverage probabilities corresponding to 95% asymptotic confidence intervals. Since $\boldsymbol{\alpha}_a \neq 0$ and $\boldsymbol{\beta}_m \neq 0$, we do not use the asymptotic results from Section 3.3.2 to construct confidence intervals. All RMSE and RIMSE estimates are based on 2000 simulation replicates.

3.4.3 Results

Figures 3.2 and 3.3 show the relative RMSE for NDE and NIE estimation in the congenial simulation settings where $n = 200$. In general, the hard and soft constraint estimators demonstrate smaller RMSE than the unconstrained estimator, with the hard constraint estimator mostly having the best performance. More specifically, larger relative efficiency gains for NDE estimation occur when $R^2_{\boldsymbol{A}|\boldsymbol{C}}$ and $R^2_{\boldsymbol{M}|\boldsymbol{A},\boldsymbol{C}}$ are small. For example, when $n_E = 20000$, $R^2_{\boldsymbol{A}|\boldsymbol{C}} = 0.05$, and $R^2_{\boldsymbol{M}|\boldsymbol{A},\boldsymbol{C}} = 0.2$, the RMSE of the unconstrained estimator is 31.4% higher than that of the hard

constraint estimator and 16.1% higher than that of the soft constraint estimator. However, when $n_E = 20000$, $R^2_{\boldsymbol{A}|\boldsymbol{C}} = 0.2$, and $R^2_{\boldsymbol{M}|\boldsymbol{A},\boldsymbol{C}} = 0.8$, the RMSE of the unconstrained estimator is only 3.7% higher than that of the hard constraint estimator and 2.3% higher than that of the soft constraint estimator. Conversely, larger relative efficiency gains for NIE estimation occur when $R^2_{\boldsymbol{A}|\boldsymbol{C}}$ is small and $R^2_{\boldsymbol{M}|\boldsymbol{A},\boldsymbol{C}}$ is large. As an example, when $n_E = 20000$, $R^2_{\boldsymbol{A}|\boldsymbol{C}} = 0.05$, and $R^2_{\boldsymbol{M}|\boldsymbol{A},\boldsymbol{C}} = 0.8$, the RMSE of the unconstrained estimator is 69.5% higher than that of the hard constraint estimator and 35.2% higher than that of the soft constraint estimator. However, when $n_E = 20000$, $R^2_{\boldsymbol{A}|\boldsymbol{C}} = 0.2$, and $R^2_{\boldsymbol{M}|\boldsymbol{A},\boldsymbol{C}} = 0.2$, the RMSE of the unconstrained estimator is only 0.3% lower than that of the hard constraint estimator and 0.4% higher than that of the soft constraint estimator. Therefore, these findings empirically corroborate the conclusions of Theorems 3.3.1, 3.3.2, and 3.3.4. Additionally, Figures 3.2 and 3.3 show that as n_E increases for a fixed value of n, the hard constraint estimator approaches the upper bound of the achievable relative RMSE as measured by the dashed horizontal line at hard constraint (oracle). For the $n = 2000$ congenial simulations, trends for the relative RMSE are the same as the trends for the $n = 200$ congenial simulation scenarios (see Supplementary Figures B.1 and B.2).

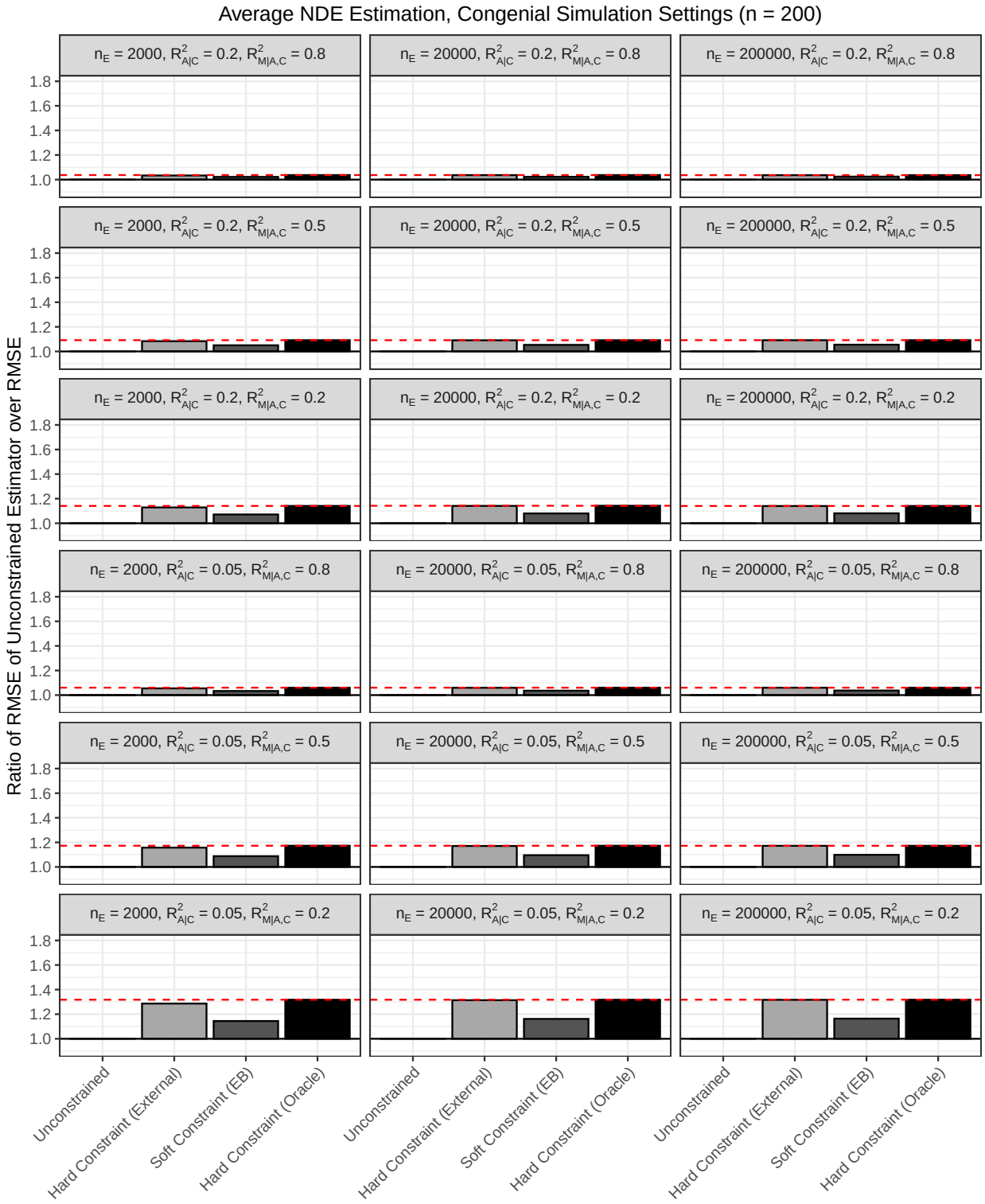


Figure 3.2: Relative root mean-squared error (RMSE) corresponding to Natural Direct Effect (NDE) estimation for the congenial simulation scenarios ($n = 200$). The red, horizontal dashed line indicates the upper bound on the possible gain in estimation efficiency, as determined by the hard constraint estimator with the oracle constraint.

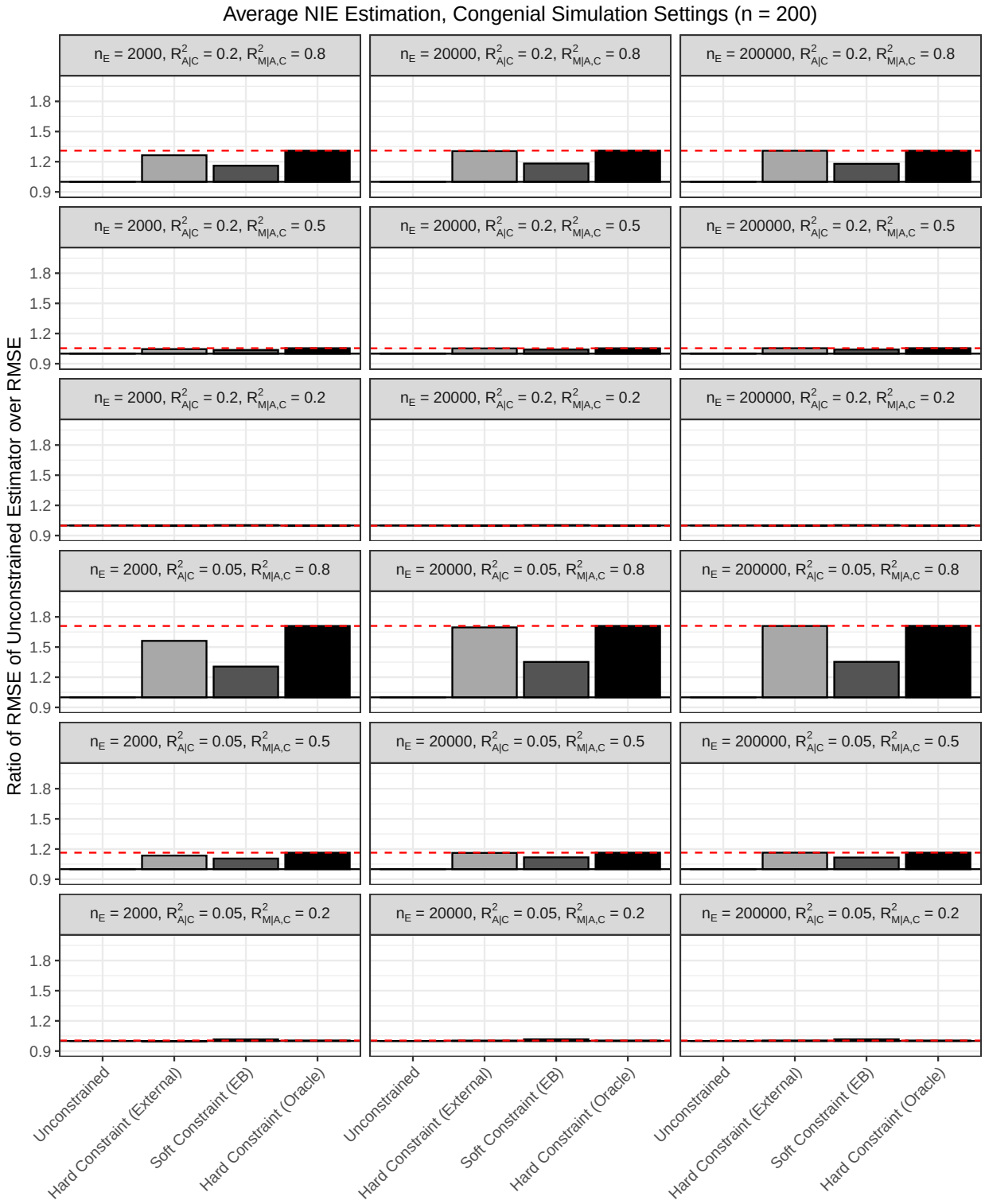


Figure 3.3: Relative root mean-squared error (RMSE) corresponding to Natural Indirect Effect (NIE) estimation for the congenial simulation scenarios ($n = 200$). The red, horizontal dashed line indicates the upper bound on the possible gain in estimation efficiency, as determined by the hard constraint estimator with the oracle constraint.

Supplementary Figures B.3 and B.4 show the empirical coverage probabilities of the 95% asymptotic confidence intervals in the $n = 2000$ congenial simulation settings. For the NDE, all methods have approximately 95% empirical coverage probabilities across all simulation settings. For the NIE, all methods achieve nominal coverage rates when $R^2_{\boldsymbol{M}|\boldsymbol{A},\boldsymbol{C}}$ is small, although the hard and soft constraint methods tend to be slightly anti-conservative when $R^2_{\boldsymbol{M}|\boldsymbol{A},\boldsymbol{C}}$ is large and $R^2_{\boldsymbol{A}|\boldsymbol{C}}$ is small. Supplementary Figures B.5 and B.6 show the empirical coverage probabilities of the 95% asymptotic confidence intervals in the $n = 200$ congenial simulation settings. For the NDE, the asymptotic normality based confidence intervals for the unconstrained and hard constraint estimators exhibit some degree of undercoverage, suggesting that a larger internal sample size is needed for the asymptotic confidence intervals derived from Theorems 3.3.1 and 3.3.2 to achieve nominal coverage rates. Conversely, the parametric bootstrap-based confidence intervals for NDE inference in the soft constraint method result in nominal coverage rates. For the NIE, all methods generally exhibit slight undercoverage in all settings, except for the unconstrained method when $R^2_{\boldsymbol{A}|\boldsymbol{C}} = 0.05$.

Figures 3.4 and 3.5 show the relative RMSE for NDE and NIE estimation in the $n = 200$ incongenial simulation settings. The unconstrained estimator has a 34.8% - 62.4% lower RMSE for NDE estimation and a 53.6% - 92.7% lower RMSE for NIE estimation compared to the hard constraint estimator. For example, when $n_E = 20000$, $R^2_{\boldsymbol{A}|\boldsymbol{C}} = 0.2$, and $R^2_{\boldsymbol{M}|\boldsymbol{A},\boldsymbol{C}} = 0.5$, the RMSE of the unconstrained estimator is 60.5% lower and 86.2% lower than that of the hard constraint estimator for NDE and NIE estimation, respectively. However, the soft constraint estimator has nearly identical RMSE to the unconstrained estimator, indicating no loss in estimation efficiency; when $n_E = 20000$, $R^2_{\boldsymbol{A}|\boldsymbol{C}} = 0.2$, and $R^2_{\boldsymbol{M}|\boldsymbol{A},\boldsymbol{C}} = 0.5$, the RMSE of the unconstrained estimator is 0.2% lower and 0.3% lower than that of the soft constraint for NDE and NIE estimation, respectively. Hence, the soft constraint (EB) estimator recovers the performance of the unconstrained estimator when the external information is incongenial. Moreover, for the random simulation settings, the conclusions are similar to the incongenial simulations settings, although the trends are less extreme because $\theta^E_a \sim N(1, 0.1)$ is almost always closer to $\theta^I_a = 1$ than $\theta^E_a = 2$ (see Supplementary Figures B.7 and B.8). With respect to coverage, both the soft constraint (EB) and unconstrained methods achieve the nominal coverage rate when $n = 2000$ (see Supplementary Figures B.9, B.10, B.11, and B.12). This suggests that coverage for the soft constraint (EB) confidence intervals is also robust to incongenial external information.

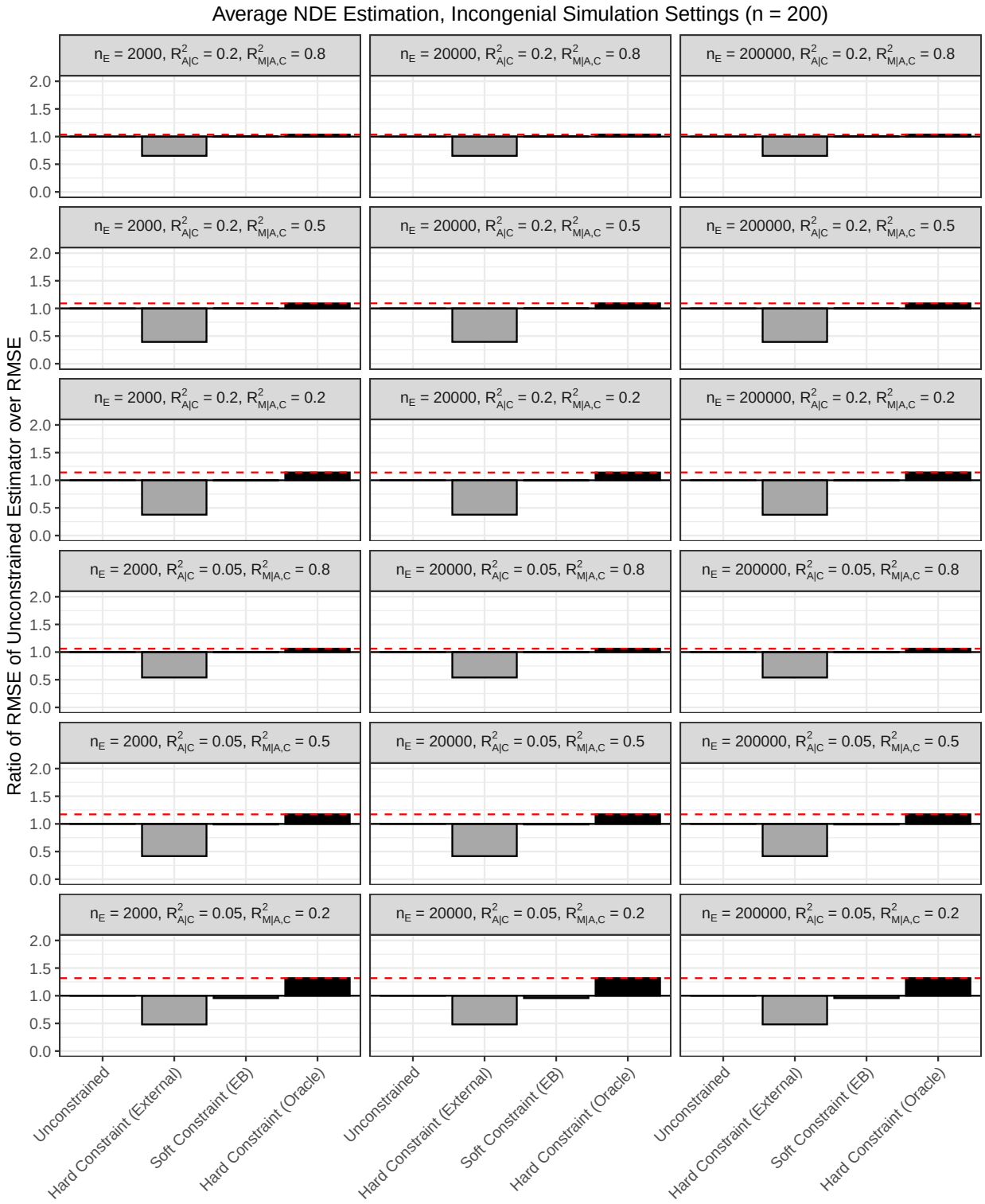


Figure 3.4: Relative root mean-squared error (RMSE) corresponding to Natural Direct Effect (NDE) estimation for the incongenial simulation scenarios ($n = 200$). The red, horizontal dashed line indicates the upper bound on the possible gain in estimation efficiency, as determined by the hard constraint estimator with the oracle constraint.

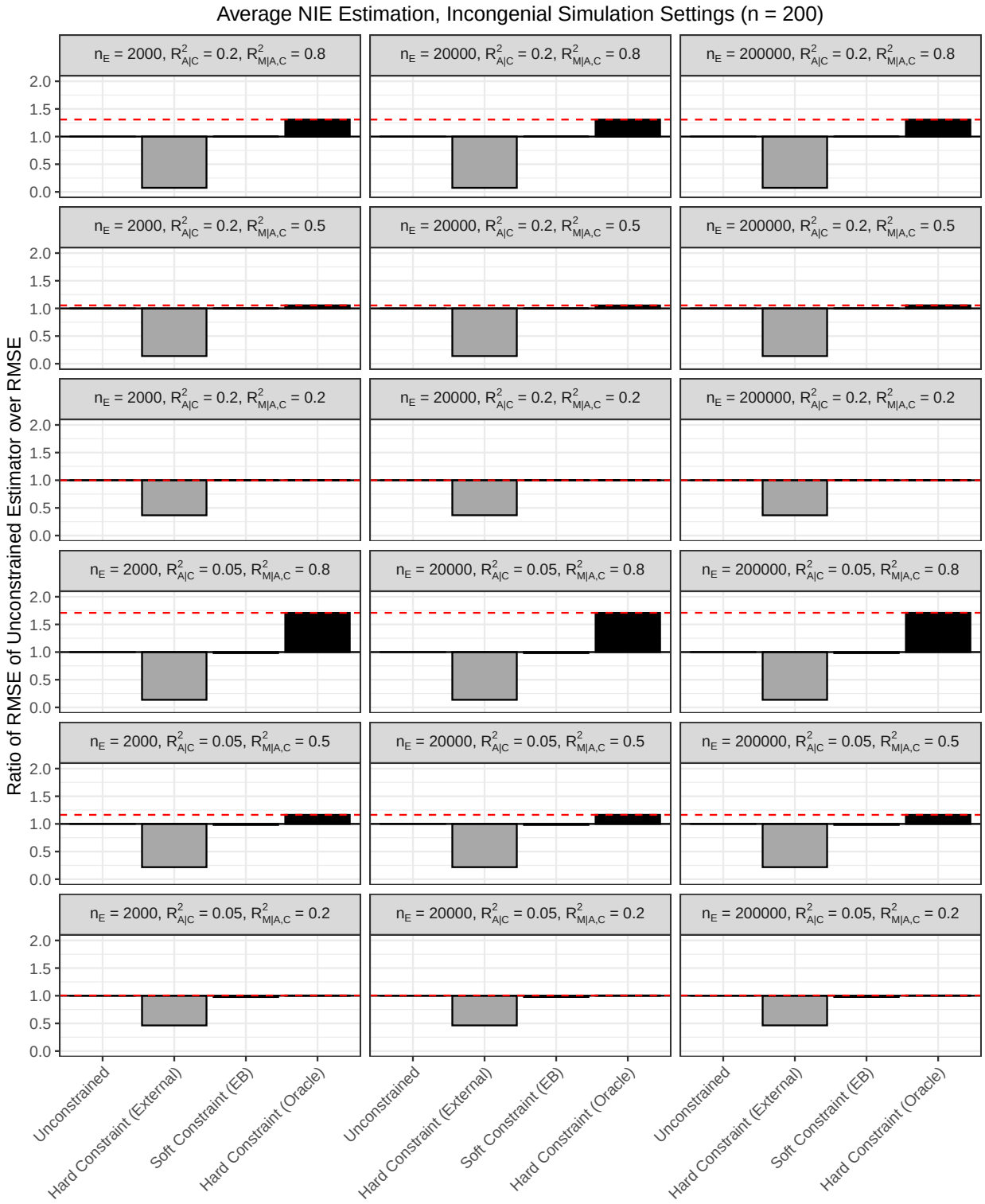


Figure 3.5: Relative root mean-squared error (RMSE) corresponding to Natural Indirect Effect (NIE) estimation for the incongenial simulation scenarios ($n = 200$). The red, horizontal dashed line indicates the upper bound on the possible gain in estimation efficiency, as determined by the hard constraint estimator with the oracle constraint.

3.5 Data Example

PROTECT is a prospective birth cohort study in Puerto Rico that aims to better understand how environmental chemical exposures adversely impact birth outcomes. Women are followed-up at three visits throughout pregnancy, with visit 1 taking place at a median of 18 weeks, visit 2 at a median of 22 weeks, and visit 3 at a median of 26 weeks. In the proposed analysis, gestational age at delivery is the outcome of interest and urinary phthalate metabolites at visits 1 and 2 are the exposure of interest. Phthalates are used to make plastics more durable and flexible and exposure in humans usually occurs through ingestion of contaminated food and water, the use of personal care products, and physical contact with household items such as polyvinyl flooring and shower curtains (Ferguson et al., 2014; Boss et al., 2018). Numerous studies in the United States have shown that higher exposure to phthalates is significantly associated with shorter gestational age at delivery (Welch et al., 2022).

Though current research is sparse, there is evidence of altered Cytochrome p450 metabolites among women that had spontaneous preterm deliveries compared to those who had full-term deliveries (Aung et al., 2019; Borkowski et al., 2020). There is also evidence that cytochrome p450 partially mediates the effect of a phthalate risk score on gestational age at delivery (Aung et al., 2020). In PROTECT, 18 Cytochrome p450 metabolites are measured at the third visit. Therefore, the proposed analysis investigates a mediation hypothesis, where the effect of log-transformed, specific-gravity adjusted phthalate metabolites at the first and second visits on gestational age at delivery is mediated by log-transformed Cytochrome p450 metabolites at the third visit, adjusted for maternal age, education, and pre-pregnancy body mass index. The phthalate metabolites of interest in this analysis are Monobutyl phthalate (MBP), Monobenzyl phthalate (MBzP), and Monoisobutyl phthalate (MiBP), which are selected based on their significant TEs as reported in eTable 13 in Welch et al. (2022). The external summary-level information on the TE is obtained by re-generating the eTable 13 models from Welch et al. (2022) excluding the PROTECT study. Depending on the visit number and phthalate metabolite, the external sample size ranges between 4890 and 4944 and the internal sample size ranges between 445 and 456 (see Supplementary Table B.1 for descriptive statistics).

Figure 3.6 presents the results of the mediation analyses using the unconstrained, soft constraint (EB), and hard constraint methods. MBzP is the only phthalate metabolite for which at least one method identifies a significant NIE. Interestingly, both the unconstrained and soft constraint methods decompose the MBzP TE in such a way that the estimated TE, NDE, and NIE are all negative, implying that Cytochrome p450 metabolites may partially explain the mechanism by which MBzP exposure shortens gestational age at delivery. For example, according to the unconstrained and soft constraint (EB) methods, the Cytochrome p450 metabolites are estimated to mediate 48.2% and

58.8% of the relationship between MBzP and gestational age at delivery at visit 2, respectively. The hard constraint interval lengths for the TE, NDE, and NIE are uniformly narrower than the soft constraint (EB) interval lengths, which in turn are uniformly narrower than the unconstrained interval lengths. For example, the unconstrained method for MiBP at visit 2 has interval lengths of 0.479 for the TE, 0.465 for the NDE, and 0.166 for the NIE, the soft constraint (EB) method yields interval lengths of 0.352 for the TE, 0.382 for the NDE, and 0.161 for the NIE, and the hard constraint method yields interval lengths of 0 for the TE, 0.159 for the NDE, and 0.159 for NIE. Larger reductions in the interval lengths are observed for the NDE compared to the NIE because $\widehat{R}^2_{\boldsymbol{M}|\boldsymbol{A},\boldsymbol{C}}$ ranges between 0.08 and 0.10, as estimated by the unconstrained method. Also note that the hard constraint method has a TE interval length of 0 because the hard constraint model guarantees $\widehat{\text{TE}}^H = \widehat{\theta}^E_a$. Our results provide corroborating evidence to the findings of the LIFECODES study, which also identified a significant NIE associated with Cytochrome p450 metabolites (Aung et al., 2020).

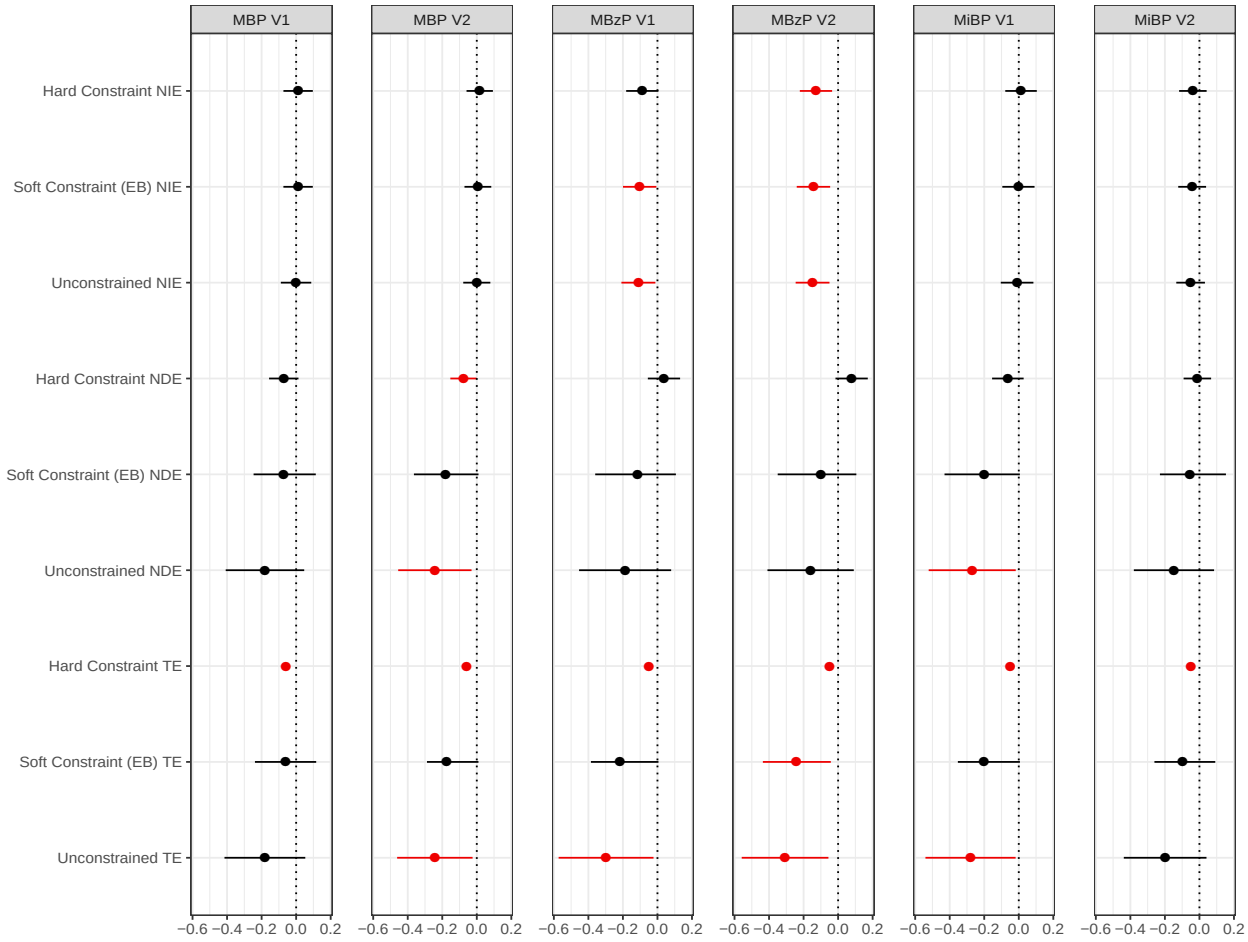


Figure 3.6: Results for the PROTECT mediation analysis, with the unconstrained, soft constraint, and hard constraint methods. Recall that MBP, MBzP, and MiBP are individual phthalate metbolites and V1 and V2 indicate that the results correspond to visit one and visit two, respectively. External sample size for MBP is 4944 and the external sample sizes for MBzP and MiBP are 4890. The internal sample sizes slightly differ for each phthalate metabolite and visit pair, however they all range between 445 and 456. All models are adjusted for maternal age, education, and maternal pre-pregnancy body mass index.

3.6 Discussion

In this chapter, we show that external summary-level information on the TE can be used to improve NDE and NIE estimation in an internal mediation analysis (see Supplementary Table B.2 for a summary of the different methods presented in this chapter). When the signal-to-noise ratio in the mediator model is low, large $R^2_{\boldsymbol{M}|\boldsymbol{A},\boldsymbol{C}}$ results in substantially more efficient NIE estimation and small $R^2_{\boldsymbol{M}|\boldsymbol{A},\boldsymbol{C}}$ results in substantially more efficient NDE estimation. Smaller values of $R^2_{\boldsymbol{M}|\boldsymbol{A},\boldsymbol{C}}$ and the signal-to-noise ratio are more common in practice, so we generally expect see

large improvements for NDE estimation. Furthermore, when the TEs in the internal and external populations differ, we can then employ EB estimation strategies to robustify shrinkage towards the external TE estimate. Implementations of the unconstrained, hard constraint, and soft constraint versions of MESSI are available in the *messi* R package on the Comprehensive R Archive Network (CRAN).

When deciding whether or not to borrow external summary-level information on the TE, there are several considerations that practitioners must weigh:

- *Internal sample size relative to both the number of mediators and the external sample size*. If the internal sample size is small, say $n < 200$, and the number of mediators is large relative to the internal sample size, say $n/p_m < 5$, then prioritizing variance reduction is important and leveraging external summary-level information on the TE may be highly beneficial. However, when the internal sample size is large, say $n > 1000$, and the number of mediators is small relative to the internal sample size, say $n/p_m > 25$, then introducing external information into an internal mediation model at best results in incremental efficiency gains and at worst introduces substantial bias into the estimation procedure. If $n < 1000$, then a general rule of thumb is to consider using external summary-level information if $5n < n_E$ and $n/p_m < 50$. If $n > 1000$, then a general rule of thumb is to consider using external summary-level information if $10n < n_E$ and $n/p_m < 20$.

- *The level of congeniality displayed between the internal and external data sources*. If $(\widehat{\theta}_a^I - \widehat{\theta}_a^E)^2 \leq \widehat{\text{Var}}(\widehat{\theta}_a^I)$, then the soft constraint estimator with EB determined s^2 corresponds to the hard constraint estimator, indicating a high level of congeniality. Therefore, when $(\widehat{\theta}_a^I - \widehat{\theta}_a^E)^2 \leq \widehat{\text{Var}}(\widehat{\theta}_a^I)$ we generally recommend using external summary-level information on the TE. If $\widehat{\theta}_a^I$ and $\widehat{\theta}_a^E$ appear to be quite different, then a practitioner can test $\theta_a^I = \theta_a^E$ with a Wald test, and use the result to guide their decision on whether to proceed with external summary-level information in the internal mediation model.

- *The overall goal of the mediation analysis and how that relates to* $R^2_{\boldsymbol{M}|\boldsymbol{A},\boldsymbol{C}}$. If the goal of the analysis is to assess whether or not there is a non-zero NIE, then leveraging external summary-level information on the TE only improves NIE estimation in the high congeniality setting when $R^2_{\boldsymbol{M}|\boldsymbol{A},\boldsymbol{C}}$ is large. Therefore, if improving the estimation efficiency of the NDE and TE in a high congeniality setting is not an important objective, then a low value of $R^2_{\boldsymbol{M}|\boldsymbol{A},\boldsymbol{C}}$, say $R^2_{\boldsymbol{M}|\boldsymbol{A},\boldsymbol{C}} < 0.4$, minimally benefits NIE estimation, defeating the purpose of leveraging external summary-level information on the total effect. However, if improving TE and NDE estimation is also an important objective of leveraging external information in an internal mediation analysis, then it is advantageous to leverage external summary-level information on the TE in a highly congenial setting, regardless of $R^2_{\boldsymbol{M}|\boldsymbol{A},\boldsymbol{C}}$.

One practical challenge that arises in applied settings is leveraging external summary-level information on the TE from multiple external cohorts, when the TE estimates have not been meta-analyzed and perhaps should not be pooled together. One approach to modify the soft constraint estimator is to alter the specification of the random effects distribution, using ideas from Bayesian dynamic borrowing (Viele et al., 2014; Kaplan et al., 2023). Suppose that we have e external summary-level estimates $\widehat{\theta}_a^{E(1)}, \ldots, \widehat{\theta}_a^{E(e)}$ and their corresponding variance estimates $\widehat{\text{Var}}\left(\widehat{\theta}_a^{E(1)}\right), \ldots, \widehat{\text{Var}}\left(\widehat{\theta}_a^{E(e)}\right)$. Then we can model the external estimates as follows

$$\widehat{\theta}_a^{E(1)} \sim N\left(\theta_a^{E(1)}, \widehat{\text{Var}}\left(\widehat{\theta}_a^{E(1)}\right)\right), \ldots, \widehat{\theta}_a^{E(e)} \sim N\left(\theta_a^{E(e)}, \widehat{\text{Var}}\left(\widehat{\theta}_a^{E(e)}\right)\right)$$

with a random effects distribution for the internal total effect parameter and external total effect parameters, $\widetilde{\theta}_a^I, \theta_a^{E(1)}, \ldots, \theta_a^{E(e)} \sim N(\theta_a^E, V_a^E)$, where $\theta_a^E \in (-\infty, \infty)$ and $V_a^E > 0$. The common θ_a^E and V_a^E allows for information sharing between the internal and external total effect parameters, while still allowing for heterogeneity in the external total effect parameters.

One major limitation is the generalizability of our results when mediators and outcomes are non-continuous or when the internal mediation model is misspecified. When the outcome or mediators are non-continuous or there is misspecification of the mean structure of the mediator and outcome models, such as when exposure-mediator interaction is present, then the expression for the TE becomes a function of the confounders and the exposure level, making leveraging external estimates less straightforward. More work is needed to build a general framework for incorporating external information on the TE into a broader class of mediation models.

One additional technical challenge that we did not fully address in this chapter is how to handle the $\boldsymbol{\alpha}_a = \boldsymbol{\beta}_m = \mathbf{0}$ case when constructing asymptotic confidence intervals for the NDE and NIE. While we recommended using a Wald test to test whether the null hypothesis $\boldsymbol{\alpha}_a = \boldsymbol{\beta}_m = \mathbf{0}$ holds as a workaround, there are likely more principled ways to construct the appropriate asymptotic reference distribution for confidence interval construction based on mixture distributions. The major challenge is that the relative weight to assign the asymptotic reference distributions from Theorems 3.3.1, 3.3.2, and 3.3.4 compared to the reference distribution from Theorem 3.3.3 is unknown, and therefore needs to be estimated (Liu et al., 2022). There is existing work in the mediation analysis literature which discusses this issue in the context of large-scale causal mediation effect identification, namely through the construction of the Divide-Aggregate Composite-Null (DACT) test, however this solves the problem by running many single-mediator tests to estimate the probability of each of the three cases in the composite null rather than trying to estimate the probability that $\boldsymbol{\alpha}_a = \boldsymbol{\beta}_m = \mathbf{0}$ (Liu et al., 2022). In our simulations we assumed that $\boldsymbol{\alpha}_a \neq \mathbf{0}$ or $\boldsymbol{\beta}_m \neq \mathbf{0}$ and directly used the asymptotic normality results from Theorems 3.3.1, 3.3.2, and 3.3.4 as a way to check our theoretical results. In the data example, we used the Wald test to test $\boldsymbol{\alpha}_a = \boldsymbol{\beta}_m = \mathbf{0}$ versus $\boldsymbol{\alpha}_a \neq \mathbf{0}$ or $\boldsymbol{\beta}_m \neq \mathbf{0}$ for all methods, all of which rejected the null hypothesis at the 0.05

level, and therefore directly used Theorems 3.3.1, 3.3.2, and 3.3.4 to construct confidence intervals. Since the main aim of the chapter is to comment on the relative efficiency gains attributable to leveraging the external summary-level information on the TE, we leave this topic as future work.

CHAPTER 4

Mediation Mixture Map

4.1 Introduction

Understanding the mechanisms by which exposure mixtures, defined as simultaneous exposure to multiple chemicals, impact human health is a priority for the field of environmental health (Braun et al., 2016; Bind, 2019). Furthermore, due to advances in highthroughput assaying technologies, environmental health studies are increasingly able to collect high-dimensional exposure data in conjunction with endogenous biomarker data, greatly facilitating discovery-based, high-dimensional mediation analyses (Aung et al., 2020). When characterizing mediation effects in environmental health studies, it is often desirable to estimate mediation effects attributable to both individual exposures and chemical mixtures, as an amplified total effect (TE) or proportion mediated for a chemical mixture relative to its constituent components is relevant for establishing biological mechanisms. From a regulatory perspective, one can also conceptualize and implement interventions that are either targeted towards individual exposures or chemical mixtures. For example, the Environmental Protection Agency provides regulatory standards for acceptable levels of ambient exposure to ground-level ozone, particulate matter, carbon monoxide, lead, sulfur dioxide, and nitrogen dioxide (https://www.epa.gov/naaqs), and also runs the Superfund program to remediate contaminated land due to hazardous waste dumps from manufacturing facilities, processing plants, landfills, and mining sites (https://www.epa.gov/superfund). The former can be thought of as interventions on individual exposures and the latter can be thought of as interventions on exposure mixtures through regulating sources of co-emission. Therefore, it is desirable for a modern environmental health mediation framework to simultaneously discuss mediation effects corresponding to single exposures and mediation effects corresponding to underlying exposure mixtures.

As with Chapter 3, the motivating example for Chapter 4 comes from PROTECT, where multi-exposure data on metals, phthalates, phenols, parabens, and polycyclic aromatic hydrocarbons (PAHs) has been collected on a subset of participants. Examples of exposure sources for phthalates,

phenols, and parabens include contact with items such as plastic food and drink packaging, vinyl flooring, and personal care products and examples of exposure sources for PAHs are consumption of grilled foods, cigarette smoke, and vehicle exhaust. Metals exposures occur from many different exposure sources, notably through the consumption of contaminated food and drinks and exposure to industrial manufacturing processes. The PROTECT investigators are interested in understanding the biological mechanisms that explain how exposure to mixtures of chemicals lead to adverse birth outcomes, such as preterm deliveries and small head circumference at delivery. The endogenous pathways under study are the lipoxygenase and cytochrome p450 pathways, both of which have shown suggestive or significant indirect effects in the LIFECODES prospective birth cohort for the relationship between a phthalate risk score and final gestational age (Aung et al., 2020).

One major challenge with developing a mediation framework for chemical mixtures is that information regarding exposure co-occurrence is unobserved. Instead, collecting and assaying biological samples allows us to observe individual chemical concentrations that may comprise relevant chemical mixtures that are latent. There is an emerging body of literature on causal inference with latent variables, with much attention paid to topics such as causal effect identification with unmeasured confounding (Glymour and Spirtes, 1988; Louizos et al., 2017; Wang and Blei, 2019; Mathur and VanderWeele, 2022), causal representation learning (Silva et al., 2006; Frot et al., 2019; Schölkopf et al., 2021; Ahuja et al., 2021), and causal effect identification in the presence of measurement error or causally inert proxy variables (Bollen, 1989; Kuroki and Pearl, 2014; Louizos et al., 2017; Miao et al., 2018). However, mediation of exposure mixture effects, does not fall neatly within these frameworks. In our conceptual formulation the chemical exposures that comprise the chemical mixtures are not considered causally inactive proxies of the latent chemical mixtures. Rather, the latent chemical mixtures can be interpreted as a low-dimensional encoding of the individual chemical exposure data. VanderWeele and Vansteelandt (2022) distinguish between causal models where the byproducts of a latent process are either causally active or inactive, and develop a statistical test to distinguish between the two cases. That being said, both causal models discussed in VanderWeele and Vansteelandt (2022) establish temporal ordering from the latent to the resultant proxies, while, in our case, the observed chemical exposures and the latent chemical mixtures occur simultaneously. That is, the observed individual chemical data and the latent chemical mixture data quantify the same exposure phenomenon, and can intuitively be thought of as different ways of summarizing the same exposure information. In this chapter, we conceptualize exposure relatedness through the statistical correlation structure of the observed exposure data, intuiting that highly correlated exposures are more likely to be linked in some way than uncorrelated exposures.

Some popular approaches for discovery-based high-dimensional mediation analyses with multiple exposures simply reduce the problem to that of a single exposure variable either by fitting

many mediation models one exposure-mediator pair at a time (Liu et al., 2022; Du et al., 2022), fitting separate high-dimensional mediation models for each exposure (Clark-Boucher et al., 2023), or fitting a high-dimensional mediation model with a summary measure, such as an environmental risk score (ERS) (Park et al., 2014) as the exposure variable (Aung et al., 2020). There is some recent methodological work on jointly modeling multiple exposures within the context of high-dimensional linear mediation models (Wang et al., 2019; Zhou et al., 2020; Zhao et al., 2022; Zhang, 2022). Zhou et al. (2020) proposed a debiased LASSO approach called HILMA that specifically focuses on the problem of global indirect effect estimation and inference when the number of mediators exceeds the sample size. Other work in the high-dimensional mediation analysis literature focuses on frequentist and Bayesian regularization techniques including a difference-of-coefficient Laplacian penalty (Wang et al., 2019), scaled adaptive LASSO (Zhang, 2022), and Pathway LASSO (Zhao et al., 2022). However, all of these methods are not targeted towards outcome and mediator adaptive learning of exposure mixtures. Devick et al. (2022) proposed a BKMR-based causal mediation framework for a data adaptively estimated non-parametric function of exposures and effect modifiers, however this framework estimates the transformed exposure mixture exclusively through a mediation model, rather than explicitly incorporating information on co-exposure through the exposure correlation structure.

In this chapter, we propose a new method called the mediation mixture map (MedMix) for discovery-based high dimensional mediation analyses in environmental health. MedMix works by combining a linear mediation model with a latent factor model, with the specific goal of estimating global mediation effects corresponding to data adaptively estimated exposure mixtures. The joint model derives outcome-guided latent mixtures that depend not just on exposure correlation but on outcome-exposure relationship. Consequently, the mediation mixture map allows us to estimate mediation effects with respect to changes in the individual exposures and changes in the latent variables that summarize common sources of variation in the exposure space. Ideally, the estimated latent constructs correspond to important exogenous chemical mixtures, however that interpretation is not guaranteed. We show that MedMix more accurately identifies the correct number of latent factors and has more appropriate uncertainty quantification for estimating mediation effects compared with a naive two-step approach that first fits a latent factor model and then plugs the estimated latent factors into a mediation model. In addition, motivated by the pathway LASSO estimator, we construct a Bayesian shrinkage framework that can directly assign a global-local shrinkage prior to the product of regression coefficients. However, similar to Wang et al. (2019), we conclude that direct shrinkage on the product leads to overly aggressive shrinkage of the indirect effect, and therefore recommend independently regularizing regression coefficients in the outcome and mediator models. For the PROTECT data example, we do not find evidence of a significant indirect effect corresponding to the cytochrome p450 and lipoxygenase pathways, how-

ever we do observe larger total effects for mixtures of phthalate metabolites compared to individual phthalate metabolites.

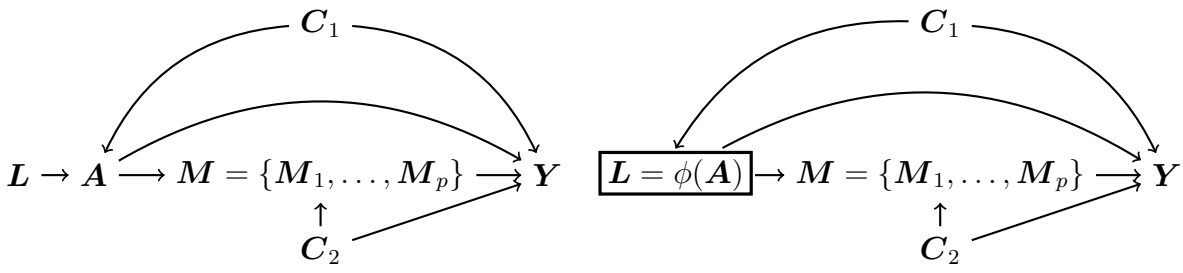
The structure of the chapter is as follows. In Section 4.2 we discuss the mediation mixture map model, the product shrinkage prior specification, and computation. In Section 4.3 we evaluate the performance of the mediation mixture map estimator compared with the naive two-step estimator in terms of estimation and inference for mediation model parameters. In Section 4.4 we discuss the PROTECT multi-exposure mediation analysis and in Section 4.5 we discuss high-level conclusions, avenues for future work, and limitations of the mediation mixture map.

4.2 Methods

In this chapter, the observed data includes individual chemical exposure concentrations ($\boldsymbol{A}$), endogenous biological processes ($\boldsymbol{M}$), the health outcome of interest ($\boldsymbol{Y}$), confounders of the outcome-exposure association ($\boldsymbol{C_1}$), and confounders of the outcome-mediator association conditional on exposure ($\boldsymbol{C}_2$). The chemical exposures, characterized by $\boldsymbol{A}$ are assumed to cause the health outcome of interest ($\boldsymbol{Y}$), partially through endogenous biological processes ($\boldsymbol{M}$). $\boldsymbol{L}$ represents exposure to chemical mixtures, and is considered a low-dimensional representation of the information in $\boldsymbol{A}$. However, in practice, this interpretation of $\boldsymbol{L}$ may be obscured by other common sources of variation such as spatiotemporal variation and endogenous biological processes that breakdown parent compounds. The objective of this chapter is to develop a high-dimensional mediation analysis framework that characterizes mediation effects attributable to changes with respect to both $\boldsymbol{A}$ and $\boldsymbol{L}$. The directed acyclic graph (DAG) considered for this chapter is depicted in Figure 4.1b, where $\boldsymbol{L}$ represents exposure to chemical mixtures and $\boldsymbol{A}$ represents the specific concentrations of individual chemicals that comprise $\boldsymbol{L}$. The box around $\boldsymbol{L}$ and $\boldsymbol{A}$ in Figure 4.1 indicates that both $\boldsymbol{A}$ and $\boldsymbol{L}$ quantify the same exposure phenomenon, and can intuitively be thought of as different ways of summarizing the same exposure information. There is no assumption regarding temporal ordering or causality that the latent source leads to the exposure vector.

4.2.1 Notation and Model Specification

For Chapter 4, Y_i denotes a continuous health outcome for observation i, $\boldsymbol{M}_{i\cdot}^\top$ denotes a $p \times 1$ vector of mediators, and $\boldsymbol{C}_{i\cdot}^\top$ denotes a $s \times 1$ vector of $s-1$ confounders, plus a one for the intercept term ($i = 1, \ldots, n$). Suppose the p mediators can be partitioned into G groups and the gth group has p_g mediators ($g = 1, \ldots, G$) such that $\sum_{g=1}^{G} p_g = p$. Let $\boldsymbol{M}_{i\cdot}^\top = (\boldsymbol{M}_{i1}^\top, \ldots, \boldsymbol{M}_{iG}^\top)^\top$, where $\boldsymbol{M}_{ig}^\top$ is a $p_g \times 1$ vector of mediators in the g-th group ($g = 1, \ldots, G$). For observation i, the individual chemical exposure data is denoted by $\boldsymbol{A}_{i\cdot}^\top$, a $q \times 1$ vector, and the latent exposure



(a) Statistical Latent Factor Model

(b) Mediation Mixture Map

Figure 4.1: Directed Acyclic Graphs comparing the Statistical Latent Factor Model discussed in VanderWeele and Vansteelandt (2022) and the Mediation Mixture Map. Here, $\boldsymbol{C}_1$ and $\boldsymbol{C}_2$ are confounders, $\boldsymbol{L}$ denotes the latent chemical mixtures, $\boldsymbol{A}$ denotes the individual chemical exposures, $\boldsymbol{M}$ denotes a set of p mediators, and $\boldsymbol{Y}$ denotes the outcome. The function $\phi(\cdot)$ represents the fact that $\boldsymbol{L}$ can be represented as stochastic transformation of $\boldsymbol{A}$.

mixture data is denoted by $\boldsymbol{L}_{i\cdot}^\top$, a $r \times 1$ vector such that $r \ll q$. That is, the mediation mixture map assumes that there is a latent, low-dimensional representation of the observed chemical exposures. The exposures in $\boldsymbol{A}_{i\cdot}^\top$ are standardized so that they have mean zero and variance one.

In this chapter, we assume that the true data generation model is:

$$\begin{aligned}
\left[Y_i \mid \boldsymbol{M}_{i\cdot}, \boldsymbol{L}_{i\cdot}, \boldsymbol{C}_{i\cdot}\right] &\sim N\left(\boldsymbol{M}_{i\cdot}\boldsymbol{\beta}_m + \boldsymbol{L}_{i\cdot}\boldsymbol{\beta}_a + \boldsymbol{C}_{i\cdot}\boldsymbol{\beta}_c, \sigma_e^2\right) \\
\left[\boldsymbol{M}_{i\cdot}^\top \mid \boldsymbol{L}_{i\cdot}^\top, \boldsymbol{C}_{i\cdot}^\top\right] &\sim N\left(\boldsymbol{\alpha}_a\boldsymbol{L}_{i\cdot}^\top + \boldsymbol{\alpha}_c\boldsymbol{C}_{i\cdot}^\top, \boldsymbol{\Sigma}_m\right) \\
\left[\boldsymbol{A}_{i\cdot}^\top \mid \boldsymbol{L}_{i\cdot}^\top\right] &\sim N\left(\boldsymbol{\Psi}\boldsymbol{L}_{i\cdot}^\top, \boldsymbol{\Sigma}_L\right) \\
\boldsymbol{L}_{i\cdot}^\top &\sim N(\boldsymbol{0}, \boldsymbol{I}_r)
\end{aligned} \tag{4.2.1}$$

where $\boldsymbol{\alpha}_a$ is a $p \times r$ matrix, $\boldsymbol{\alpha}_c$ is a $p \times s$ matrix, $\boldsymbol{\beta}_m$ is a $p \times 1$ vector, $\boldsymbol{\beta}_a$ is a $r \times 1$ vector, $\boldsymbol{\beta}_c$ is a $s \times 1$ vector, σ_e^2 is a scalar, $\boldsymbol{\Sigma}_m$ is a $p \times p$ symmetric, positive-definite matrix, $\boldsymbol{\Psi}$ is a $q \times r$ loadings matrix, $\boldsymbol{\Sigma}_L = \text{diag}(\sigma_{L1}^2, \ldots, \sigma_{Lq}^2)$ is a $q \times q$ diagonal matrix, and $\boldsymbol{I}_r$ is an $r \times r$ identity matrix. For the purposes of modeling $\boldsymbol{\Sigma}_m$, we assume an unstructured correlation structure between mediators that are in the same group and zero correlation between mediators in different groups. The regression coefficients $\boldsymbol{\alpha}_a$ and $\boldsymbol{\beta}_a$ refer to the effect of the latent sources of exposure variation on the mediators conditional on confounders and the effect of the latent sources of exposure variation on the outcome conditional on the mediators and confounders, respectively. Additionally, $\boldsymbol{\beta}_m$ indicates the effect of the mediators on the outcome conditional on latent sources of exposure variation and confounders. The entries in the loadings matrix $\boldsymbol{\Psi}$ map the common sources of exposure variation to the observed exposures. For example, large values of $\boldsymbol{\Psi}_{k'\cdot}\boldsymbol{\Psi}_{k'\cdot}^\top$ relative to $\sigma_{Lk'}^2$, where $\boldsymbol{\Psi}_{k'\cdot}$ is the k'-th row of $\boldsymbol{\Psi}$, indicate that the r latent sources of exposure variation explain a large proportion of variation in the k'-th exposure. This fact is seen from the expression for the marginal

variance of $\boldsymbol{A}_{i\cdot}$, $\text{Var}(\boldsymbol{A}_{i\cdot}) = \boldsymbol{\Psi}\boldsymbol{\Psi}^\top + \boldsymbol{\Sigma}_L$.

Model (4.2.1) combines the Baron and Kenny mediation model (Baron and Kenny, 1986) with respect to latent sources of exposure variation, defined by the conditional distributions $[Y_i \mid \boldsymbol{M}_{i\cdot}, \boldsymbol{L}_{i\cdot}, \boldsymbol{C}_{i\cdot}]$ and $\left[\boldsymbol{M}_{i\cdot}^\top \mid \boldsymbol{L}_{i\cdot}^\top, \boldsymbol{C}_{i\cdot}^\top\right]$, and a latent factor model, defined by the joint distribution $\left[\boldsymbol{A}_{i\cdot}^\top, \boldsymbol{L}_{i\cdot}^\top\right]$. This is distinctly different than the naive two-step approach which sequentially fits a latent factor model to the exposure data and then fits a mediation or total effect model with the estimated latent factors as the exposure (Gibson et al., 2019). Combining the mediation and latent factor components together allows the mediation model to inform the estimation of parameters and latent factors in the latent factor model and accounts for uncertainty in estimation of the latent factors. Model (4.2.1) makes two implicit assumptions about the joint distribution $[Y_i, \boldsymbol{M}_{i\cdot}, \boldsymbol{A}_{i\cdot}, \boldsymbol{L}_{i\cdot} \mid \boldsymbol{C}_{i\cdot}]$, namely that $\{Y_i, \boldsymbol{M}_{i\cdot}\} \perp\!\!\!\perp \boldsymbol{A}_{i\cdot} \mid \{\boldsymbol{L}_{i\cdot}, \boldsymbol{C}_{i\cdot}\}$ and $\{\boldsymbol{A}_{i\cdot}, \boldsymbol{L}_{i\cdot}\} \perp\!\!\!\perp \boldsymbol{C}_{i\cdot}$. The first assumption essentially states that the latent low-dimensional representation $\boldsymbol{L}_{i\cdot}$ is sufficient for characterizing the joint conditional distribution $[Y_i, \boldsymbol{M}_{i\cdot} \mid \boldsymbol{A}_{i\cdot}, \boldsymbol{C}_{i\cdot}]$ and the second assumptions states that confounders in the mediation model do not influence the characterization of the low-dimensional latent exposure space. We will refer to Model (4.2.1) as the mediation mixture map. Estimates of mediation effects with respect to changes in the observed exposure from the mediation mixture map will have a causal interpretation if the four unconfoundness assumptions listed in (VanderWeele, 2015) or (Song et al., 2020) hold. Estimates of mediation effects with respect to changes in the latent exposures from the mediation mixture map do not have a causal interpretation.

The specification of $\left[\boldsymbol{A}_{i\cdot}^\top \mid \boldsymbol{L}_{i\cdot}^\top\right]$ in Model (4.2.1) explains how the latent exposure space is mapped to the observed exposure space. However, for the purposes of interpreting mediation effects with respect to latent sources of exposure variation, it is helpful to understand how information on the observed exposure space maps back to the latent exposure space. Under model specification (4.2.1),

$$\left[\boldsymbol{L}_{i\cdot}^\top \mid \boldsymbol{A}_{i\cdot}^\top\right] \sim N\left(\left(\boldsymbol{I}_r + \boldsymbol{\Psi}^\top\boldsymbol{\Sigma}_L^{-1}\boldsymbol{\Psi}\right)^{-1}\boldsymbol{\Psi}^\top\boldsymbol{\Sigma}_L^{-1}\boldsymbol{A}_{i\cdot}^\top, \left(\boldsymbol{I}_r + \boldsymbol{\Psi}^\top\boldsymbol{\Sigma}_L^{-1}\boldsymbol{\Psi}\right)^{-1}\right).$$

Therefore, the stochastic map $\phi(\cdot)$ which maps between realizations of the observed exposure values and unobserved realizations in the latent exposure space is

$$\begin{aligned}\phi^{-1}(\boldsymbol{L}_{i\cdot}) &= \boldsymbol{\Psi}\boldsymbol{L}_{i\cdot}^\top + \boldsymbol{\epsilon}_L, \quad \boldsymbol{\epsilon}_L \sim N(\boldsymbol{0}, \boldsymbol{\Sigma}_L)\\ \phi(\boldsymbol{A}_{i\cdot}) &= \boldsymbol{P}_{\boldsymbol{\Psi},\boldsymbol{\Sigma}_L}\boldsymbol{A}_{i\cdot}^\top + \boldsymbol{\epsilon}_{L|A}, \quad \boldsymbol{\epsilon}_{L|A} \sim N(\boldsymbol{0}, \boldsymbol{\Sigma}_{L|A})\end{aligned}$$

where $\boldsymbol{P}_{\boldsymbol{\Psi},\boldsymbol{\Sigma}_L} = \left(\boldsymbol{I}_r + \boldsymbol{\Psi}^\top\boldsymbol{\Sigma}_L^{-1}\boldsymbol{\Psi}\right)^{-1}\boldsymbol{\Psi}^\top\boldsymbol{\Sigma}_L^{-1}$ is a projection matrix and $\boldsymbol{\Sigma}_{L|A} = \Big(\boldsymbol{I}_r +$

$\boldsymbol{\Psi}^\top \boldsymbol{\Sigma}_L^{-1} \boldsymbol{\Psi}\Big)^{-1}$. The mean specification of the implied mediation model with respect to the observed exposures is then

$$
\begin{aligned}
E[Y_i \mid \boldsymbol{M}_{i\cdot}, \boldsymbol{A}_{i\cdot}, \boldsymbol{C}_{i\cdot}] &= \boldsymbol{M}_{i\cdot}\boldsymbol{\beta}_m + \boldsymbol{A}_{i\cdot}\boldsymbol{P}_{\boldsymbol{\Psi},\boldsymbol{\Sigma}_L}^\top \boldsymbol{\beta}_a + \boldsymbol{C}_{i\cdot}\boldsymbol{\beta}_c \\
E\big[\boldsymbol{M}_{i\cdot}^\top \mid \boldsymbol{A}_{i\cdot}^\top, \boldsymbol{C}_{i\cdot}^\top\big] &= \boldsymbol{\alpha}_a \boldsymbol{P}_{\boldsymbol{\Psi},\boldsymbol{\Sigma}_L} \boldsymbol{A}_{i\cdot}^\top + \boldsymbol{\alpha}_c \boldsymbol{C}_{i\cdot}^\top
\end{aligned}
$$

where $\boldsymbol{P}_{\boldsymbol{\Psi},\boldsymbol{\Sigma}_L}^\top \boldsymbol{\beta}_a$ is the effect of the observed exposures on the outcome given mediators and confounders and $\boldsymbol{\alpha}_a \boldsymbol{P}_{\boldsymbol{\Psi},\boldsymbol{\Sigma}_L}$ is the effect of the observed exposures on the mediators given confounders. Note that the mediation mixture map model specification implies a linear mediation model with respect to the individual chemical exposures.

The parameters of interest in the mediation mixture map are the indirect effect (IE), the direct effect (DE), and the total effect (TE) corresponding to changes in the individual exposure levels

$$
\begin{aligned}
\text{DE}^A(\boldsymbol{a}, \boldsymbol{a}^*) &= (\boldsymbol{a} - \boldsymbol{a}^*)^\top \boldsymbol{P}_{\boldsymbol{\Psi},\boldsymbol{\Sigma}_L}^\top \boldsymbol{\beta}_a \\
\text{IE}^A(\boldsymbol{a}, \boldsymbol{a}^*) &= (\boldsymbol{a} - \boldsymbol{a}^*)^\top \boldsymbol{P}_{\boldsymbol{\Psi},\boldsymbol{\Sigma}_L}^\top \boldsymbol{\alpha}_a^\top \boldsymbol{\beta}_m \\
\text{TE}^A(\boldsymbol{a}, \boldsymbol{a}^*) &= (\boldsymbol{a} - \boldsymbol{a}^*)^\top \boldsymbol{P}_{\boldsymbol{\Psi},\boldsymbol{\Sigma}_L}^\top \big(\boldsymbol{\beta}_a + \boldsymbol{\alpha}_a^\top \boldsymbol{\beta}_m\big)
\end{aligned}
$$

and the mediation effects corresponding to changes in the latent sources of variation in the exposure space

$$
\begin{aligned}
\text{DE}^L(\boldsymbol{l}, \boldsymbol{l}^*) &= (\boldsymbol{l} - \boldsymbol{l}^*)^\top \boldsymbol{\beta}_a \\
\text{IE}^L(\boldsymbol{l}, \boldsymbol{l}^*) &= (\boldsymbol{l} - \boldsymbol{l}^*)^\top \boldsymbol{\alpha}_a^\top \boldsymbol{\beta}_m \\
\text{TE}^L(\boldsymbol{l}, \boldsymbol{l}^*) &= (\boldsymbol{l} - \boldsymbol{l}^*)^\top \big(\boldsymbol{\beta}_a + \boldsymbol{\alpha}_a^\top \boldsymbol{\beta}_m\big),
\end{aligned}
$$

In these expressions, $\boldsymbol{a}$ and $\boldsymbol{l}$ represent the comparative exposure levels and $\boldsymbol{a}^*$ and $\boldsymbol{l}^*$ represent the reference exposure levels. Under several no unmeasured confounding assumptions, the stable unit treatment value assumption, the mediation effects have causal interpretations (VanderWeele, 2015; Song et al., 2020; Zhang, 2022).

Expressions for the mediation effects clarify several points. The first is that the DE is always a function of $\boldsymbol{\beta}_a$ regardless of whether the change occurs on the observed exposure scale or the latent exposure scale. Similarly, the IE is always a function of $\boldsymbol{\alpha}_a$ and $\boldsymbol{\beta}_m$ regardless of whether the change occurs on the observed exposure scale or the latent exposure scale. The primary difference between expressions on the latent exposure scale versus the observed exposure scale is that the observed exposure scale applies the projection matrix $\boldsymbol{P}_{\boldsymbol{\Psi},\boldsymbol{\Sigma}_L}^\top$ to $\boldsymbol{a} - \boldsymbol{a}^*$. From an inferential perspective, $\text{DE}^A(\boldsymbol{a}, \boldsymbol{a}^*)$ and $\text{DE}^L(\boldsymbol{l}, \boldsymbol{l}^*)$, $\text{IE}^A(\boldsymbol{a}, \boldsymbol{a}^*)$ and $\text{IE}^L(\boldsymbol{l}, \boldsymbol{l}^*)$, and $\text{TE}^A(\boldsymbol{a}, \boldsymbol{a}^*)$ and $\text{TE}^L(\boldsymbol{l}, \boldsymbol{l}^*)$ are linked together in the sense that, on average, a difference $\boldsymbol{a} - \boldsymbol{a}^*$ corresponds to a difference

on the latent scale $\boldsymbol{l} - \boldsymbol{l}^* = \boldsymbol{P}_{\boldsymbol{\Psi},\boldsymbol{\Sigma}_L}(\boldsymbol{a} - \boldsymbol{a}^*)$ for any $\boldsymbol{a}, \boldsymbol{a}^* \in \mathbb{R}^q$. Similarly, on average, inferential statements regarding a difference $\boldsymbol{l} - \boldsymbol{l}^*$ correspond to a difference on the observed scale $\boldsymbol{a} - \boldsymbol{a}^* = \boldsymbol{\Psi}(\boldsymbol{l} - \boldsymbol{l}^*)$ for any $\boldsymbol{l}, \boldsymbol{l}^* \in \mathbb{R}^r$. However, when $\boldsymbol{l} - \boldsymbol{l}^* = \boldsymbol{P}_{\boldsymbol{\Psi},\boldsymbol{\Sigma}_L}(\boldsymbol{a} - \boldsymbol{a}^*)$, hypothesis tests that test whether $\text{DE}^L(\boldsymbol{l}, \boldsymbol{l}^*) = 0$, $\text{IE}^L(\boldsymbol{l}, \boldsymbol{l}^*) = 0$, or $\text{TE}^L(\boldsymbol{l}, \boldsymbol{l}^*) = 0$ are not necessarily equivalent to hypothesis tests that test $\text{DE}^A(\boldsymbol{a}, \boldsymbol{a}^*) = 0$, $\text{IE}^A(\boldsymbol{a}, \boldsymbol{a}^*) = 0$, or $\text{TE}^A(\boldsymbol{a}, \boldsymbol{a}^*) = 0$, respectively. This is primarily because testing $\text{IE}^A(\boldsymbol{a}, \boldsymbol{a}^*) = 0$ involves additional estimation uncertainty with respect to $\boldsymbol{P}_{\boldsymbol{\Psi},\boldsymbol{\Sigma}_L}$. In a fully Bayesian paradigm, inferential conclusions are not necessarily identical between mediation effects on the observed exposure scale and the latent exposure scale either. Namely, $\boldsymbol{P}_{\boldsymbol{\Psi},\boldsymbol{\Sigma}_L}$is a random quantity within a fully Bayesian modeling framework, and therefore $\boldsymbol{a} - \boldsymbol{a}^*$ stochastically corresponds to a range of $\boldsymbol{l} - \boldsymbol{l}^*$ values rather than one specific value of $\boldsymbol{l} - \boldsymbol{l}^*$. Consequently, we recommend using $\text{DE}^A(\boldsymbol{a}, \boldsymbol{a}^*)$, $\text{IE}^A(\boldsymbol{a}, \boldsymbol{a}^*)$, and $\text{TE}^A(\boldsymbol{a}, \boldsymbol{a}^*)$ if inferential statements with respect to the observed exposure space are of interest and $\text{DE}^L(\boldsymbol{l}, \boldsymbol{l}^*)$, $\text{IE}^L(\boldsymbol{l}, \boldsymbol{l}^*)$, and $\text{TE}^L(\boldsymbol{l}, \boldsymbol{l}^*)$ if inferential statements with respect to the latent exposure space are of interest.

It important to clarify that, because the mediation model component of the mediation mixture map can be expressed with respect to $\boldsymbol{L}$ or with respect to $\boldsymbol{A}$, then is also possible to obtain estimation and inference with respect to $\boldsymbol{\alpha}_a$, $\boldsymbol{\beta}_a$, $\boldsymbol{\beta}_m$, $\boldsymbol{\alpha}_a\boldsymbol{P}_{\boldsymbol{\Psi},\boldsymbol{\Sigma}_L}$, and $\boldsymbol{P}^\top_{\boldsymbol{\Psi},\boldsymbol{\Sigma}_L}\boldsymbol{\beta}_a$. Consequently, associations between exposures on the observed and latent scale, mediators, the outcome, and functions thereof can all be examined.

Remark. When analyzing the PROTECT data we will also consider a total effect only version of model (4.2.1)

$$\begin{aligned}
\big[Y_i \mid \boldsymbol{L}_{i\cdot}, \boldsymbol{C}_{i\cdot}\big] &\sim N\big(\boldsymbol{L}_{i\cdot}\boldsymbol{\theta}_a + \boldsymbol{C}_{i\cdot}\boldsymbol{\theta}_c, \sigma_t^2\big) \\
\big[\boldsymbol{A}_{i\cdot}^\top \mid \boldsymbol{L}_{i\cdot}^\top\big] &\sim N\big(\boldsymbol{\Psi}\boldsymbol{L}_{i\cdot}^\top, \boldsymbol{\Sigma}_L\big) \\
\boldsymbol{L}_{i\cdot}^\top &\sim N(\boldsymbol{0}, \boldsymbol{I}_r)
\end{aligned} \tag{4.2.2}$$

where $\boldsymbol{\theta}_a$ is a $r \times 1$ vector and $\boldsymbol{\theta}_c$ is a $s \times 1$ vector. In Model (4.2.2), $\boldsymbol{\theta}_a$ is the effect of the latent sources of exposure variation on the outcome given confounders. As with Model (4.2.1), the implied total effect model with respect to observed exposures is $E[Y_i \mid \boldsymbol{A}_{i\cdot}, \boldsymbol{C}_{i\cdot}] = \boldsymbol{A}_{i\cdot}\boldsymbol{P}^\top_{\boldsymbol{\Psi},\boldsymbol{\Sigma}_L}\boldsymbol{\theta}_a + \boldsymbol{C}_{i\cdot}\boldsymbol{\theta}_c$. Consequently, $\boldsymbol{P}^\top_{\boldsymbol{\Psi},\boldsymbol{\Sigma}_L}\boldsymbol{\theta}_a$ is the effect of the observed exposures on the outcome given confounders. Model (4.2.2) is called the total effect mixture map. In principle, $\text{TE}^L(\boldsymbol{l}, \boldsymbol{l}^*)$ and $\text{TE}^A(\boldsymbol{a}, \boldsymbol{a}^*)$ are connected to Model (4.2.2) through $\boldsymbol{\theta}_a = \boldsymbol{\beta}_a + \boldsymbol{\alpha}_a^\top\boldsymbol{\beta}_m$. However, in practice, the estimated latent factors $\widehat{\boldsymbol{L}}_{i\cdot}$ are not necessarily the same for Model (4.2.1) and Model (4.2.2) because Model (4.2.1) uses information from the mediators to estimate $\boldsymbol{L}_{i\cdot}$.

4.2.2 Invariance of Mediation Effects to Rotations

Next, we consider the the issue of identifiabilty of $\boldsymbol{\Psi}$, which is a known issue in latent factor modeling (Geweke and Zhou, 1996). More specifically, a latent factor model cannot identify $\boldsymbol{\Psi}$ up to a rotation because $\boldsymbol{L}_{i\cdot} \sim N(0, \boldsymbol{I}_r)$ and $\boldsymbol{L}_{i\cdot}\boldsymbol{P} \sim N(0, \boldsymbol{I}_r)$ for all $r \times r$ matricies $\boldsymbol{P}$ that satisfy $\boldsymbol{P}\boldsymbol{P}^\top = \boldsymbol{I}_r$. One popular identifiability condition is to encode structural zeros in the elements of $\boldsymbol{\Psi}$, namely $\boldsymbol{\Psi}_{k'k} = 0$ for $k' < k$ and $\boldsymbol{\Psi}_{k'k} > 0$ for $k' = k$ where $\boldsymbol{\Psi}_{k'k}$ is the entry of $\boldsymbol{\Psi}$ in the k'-th row and k-th column (Geweke and Zhou, 1996). However, recent applications of Bayesian latent factor models to environmental exposure data impose identifiability constraints through algorithmic post-processing of posterior draws rather than a priori specified structural zeros in $\boldsymbol{\Psi}$, particularly when parameters of interest are invariant to rotations of $\boldsymbol{\Psi}$ (Ferrari and Dunson, 2021).

Theorem 4.2.1. *Consider a rotational reparameterization of Model* (4.2.1)

$$\begin{aligned}
\left[Y_i \mid \boldsymbol{M}_{i\cdot}, \boldsymbol{L}^*_{i\cdot}, \boldsymbol{C}_{i\cdot}\right] &\sim N\left(\boldsymbol{M}_{i\cdot}\boldsymbol{\beta}_m + \boldsymbol{L}^*_{i\cdot}\boldsymbol{\beta}^*_a + \boldsymbol{C}_{i\cdot}\boldsymbol{\beta}_c, \sigma^2_e\right) \\
\left[\boldsymbol{M}^\top_{i\cdot} \mid \boldsymbol{L}^{*\top}_{i\cdot}, \boldsymbol{C}^\top_{i\cdot}\right] &\sim N\left(\boldsymbol{\alpha}^*_a\boldsymbol{L}^{*\top}_{i\cdot} + \boldsymbol{\alpha}_c\boldsymbol{C}^\top_{i\cdot}, \boldsymbol{\Sigma}_m\right) \\
\left[\boldsymbol{A}^\top_{i\cdot} \mid \boldsymbol{L}^{*\top}_{i\cdot}\right] &\sim N\left(\boldsymbol{\Psi}^*\boldsymbol{L}^{*\top}_{i\cdot}, \boldsymbol{\Sigma}_L\right) \\
\boldsymbol{L}^{*\top}_{i\cdot} &\sim N(\boldsymbol{0}, \boldsymbol{I}_r),
\end{aligned}$$

where $\boldsymbol{\Psi}^* = \boldsymbol{\Psi}\boldsymbol{P}$, $\boldsymbol{L}^*_{i\cdot} = \boldsymbol{L}_{i\cdot}\boldsymbol{P}$, $\boldsymbol{\beta}^*_a = \boldsymbol{P}^\top\boldsymbol{\beta}_a$, $\boldsymbol{\alpha}^*_a = \boldsymbol{\alpha}_a\boldsymbol{P}$, *and* $\boldsymbol{P}$ *is an* $r \times r$ *matrix that satisfies* $\boldsymbol{P}\boldsymbol{P}^\top = \boldsymbol{I}_r$. *Then,*

$$\begin{aligned}
DE^A(\boldsymbol{a}, \boldsymbol{a}^*) &= (\boldsymbol{a} - \boldsymbol{a}^*)^\top \boldsymbol{P}^\top_{\boldsymbol{\Psi},\boldsymbol{\Sigma}_L}\boldsymbol{\beta}_a \\
IE^A(\boldsymbol{a}, \boldsymbol{a}^*) &= (\boldsymbol{a} - \boldsymbol{a}^*)^\top \boldsymbol{P}^\top_{\boldsymbol{\Psi},\boldsymbol{\Sigma}_L}\boldsymbol{\alpha}^\top_a\boldsymbol{\beta}_m
\end{aligned}$$

Proof. See Supplementary Material C.1.

Therefore, as with several other Bayesian applications of latent factor models, inference on $\text{DE}^A(\boldsymbol{a}, \boldsymbol{a}^*)$, $\text{IE}^A(\boldsymbol{a}, \boldsymbol{a}^*)$, and $\text{TE}^A(\boldsymbol{a}, \boldsymbol{a}^*)$ does not require identifiability assumptions on the loadings matrix $\boldsymbol{\Psi}$ (Bhattacharya and Dunson, 2011; Ferrari and Dunson, 2021). However, estimation and inference corresponding to $\text{DE}^L(\boldsymbol{l}, \boldsymbol{l}^*)$, $\text{IE}^L(\boldsymbol{l}, \boldsymbol{l}^*)$, $\text{TE}^L(\boldsymbol{l}, \boldsymbol{l}^*)$, $\boldsymbol{\Psi}$, and $\boldsymbol{L}$, necessitates a post-processing step to resolve rotational ambiguity. By rotational ambiguity, we are referring to the fact that $\boldsymbol{\alpha}_a$, $\boldsymbol{\beta}_a$, $\boldsymbol{\Psi}$, and $\boldsymbol{L}$ are not identified in Model (4.2.1) up to a rotation $\boldsymbol{P}$. Rather than imposing identifiability constraints on $\boldsymbol{\Psi}$ at the modeling stage, we use the two-step post-processing algorithm proposed in Poworoznek et al. (2021), which obtains the Varimax rotation (Kaiser, 1958) to maximize the variance of the squared loadings in each factor, applies the Varimax rotation to the posterior draws of $\boldsymbol{\alpha}_a$, $\boldsymbol{\beta}_a$, $\boldsymbol{\Psi}$, and $\boldsymbol{L}$ and then corrects sign and label switching across posterior

samples by matching the columns and signs of the varimax rotated posterior draws to a reference varimax rotated posterior draw. The varimax rotation allows for a parsimonious estimate of $\boldsymbol{\Psi}$, boosting the interpretability of $\text{DE}^L(\boldsymbol{l}, \boldsymbol{l}^*)$, $\text{IE}^L(\boldsymbol{l}, \boldsymbol{l}^*)$, and $\text{TE}^L(\boldsymbol{l}, \boldsymbol{l}^*)$.

Because $\text{DE}^A(\boldsymbol{a}, \boldsymbol{a}^*)$, $\text{IE}^A(\boldsymbol{a}, \boldsymbol{a}^*)$, and $\text{TE}^A(\boldsymbol{a}, \boldsymbol{a}^*)$ are invariant to rotations of $\boldsymbol{\Psi}$, then the causal interpretability of $\text{DE}^A(\boldsymbol{a}, \boldsymbol{a}^*)$, $\text{IE}^A(\boldsymbol{a}, \boldsymbol{a}^*)$, and $\text{TE}^A(\boldsymbol{a}, \boldsymbol{a}^*)$ is guaranteed provided that the conditions for causal identifiability hold (VanderWeele, 2015; Song et al., 2020; Zhang, 2022) and model (4.2.1) is correctly specified. However, the causal interpretability of $\text{DE}^L(\boldsymbol{l}, \boldsymbol{l}^*)$, $\text{IE}^L(\boldsymbol{l}, \boldsymbol{l}^*)$, and $\text{TE}^L(\boldsymbol{l}, \boldsymbol{l}^*)$ is dependent on the implicit identifiability conditions imposed on $\boldsymbol{\Psi}$ through the Varimax rotation post-processing step (Ahuja et al., 2021). Therefore, one should be cautious about interpreting $\text{DE}^L(\boldsymbol{l}, \boldsymbol{l}^*)$, $\text{IE}^L(\boldsymbol{l}, \boldsymbol{l}^*)$, and $\text{TE}^L(\boldsymbol{l}, \boldsymbol{l}^*)$ causally, even if all conditions for causal identifiability hold.

4.2.3 Prior Specification

When the endogenous biomechanism under investigation is quantified through a collection of high-dimensional omics data, regularization is useful strategy to help estimate global mediation effects. However, when the primary target parameter of interest, namely the IE, involves the product of regression coefficients from outcome and mediator models, some have argued that directly regularizing the product of regression coefficients is preferred over independent regularization of $\boldsymbol{\beta}_m$ and $\boldsymbol{\alpha}_a$ (Song et al., 2021b; Zhao et al., 2022; Zhao and Luo, 2022). Others have argued that targeted shrinkage of the product is too aggressive and tends to over-regularize the IE (Wang et al., 2019). To date, methods for targeted shrinkage on the product either involve thresholding the product of regression coefficients (Song et al., 2021b) or deploying a lasso-style penalty on the product of regression coefficients (Zhao et al., 2022; Zhao and Luo, 2022). We propose a new Bayesian shrinkage prior framework for targeted shrinkage of the product, which allows any shrinkage prior to be applied to the product of regression coefficients provided that the desired prior can be expressed as a scale mixture of normals.

Before, discussing the exact prior specification we first introduce some additional notation to clarify the mediator grouping structure in $\boldsymbol{\beta}_m$ and $\boldsymbol{\alpha}_a$. We use the following notation $\boldsymbol{\beta}_m = \left((\boldsymbol{\beta}_m)_{1\cdot}^\top, \ldots, (\boldsymbol{\beta}_m)_{G\cdot}^\top\right)^\top$ where $(\boldsymbol{\beta}_m)_{g\cdot}$ is a $p_g \times 1$ subvector corresponding to the g-th mediator group ($g = 1, \ldots, G$). Furthermore, $(\boldsymbol{\beta}_m)_{g\cdot} = \left((\boldsymbol{\beta}_m)_{g1}, \ldots, (\boldsymbol{\beta}_m)_{gp_g}\right)^\top$ where $(\boldsymbol{\beta}_m)_{gj}$ is a scalar corresponding to the j-th mediator in the g-th mediator group ($j = 1, \ldots, p_g$). Similarly, for $\boldsymbol{\alpha}_a$, $\boldsymbol{\alpha}_a = (\boldsymbol{\alpha}_a^1, \ldots, \boldsymbol{\alpha}_a^r)$, where $\boldsymbol{\alpha}_a^k$ is a $p \times 1$ vector of associations between the k-th latent exposure mixture in $\boldsymbol{L}_{i\cdot}$ and p mediators. $\boldsymbol{\alpha}_a^k$ can be further broken down such that $\boldsymbol{\alpha}_a^k = \left((\boldsymbol{\alpha}_a^k)_{1\cdot}^\top, \ldots, (\boldsymbol{\alpha}_a^k)_{G\cdot}^\top\right)^\top$, where $(\boldsymbol{\alpha}_a^k)_{g\cdot}$ is a $p_g \times 1$ subvector associating the k-th latent exposure mixture in $\boldsymbol{L}_{i\cdot}$ with the p_g mediators in the g-th group. Here, $(\boldsymbol{\alpha}_a^k)_{gj}$ is the scalar association

between the k-th exposure of $\boldsymbol{L}_{i\cdot}^{\top}$ and the j-th mediator in the g-th mediator group.

Consider the following prior specification for $\boldsymbol{\beta}_m$ and $\boldsymbol{\alpha}_a$ in model (4.2.1)

$$\begin{aligned}\left[(\boldsymbol{\beta}_m)_{gj} \mid \tau_\beta^2, \gamma_{\beta g}^2, \lambda_{\beta gj}^2\right] &\sim N\left(0, \tau_\beta^2\gamma_{\beta g}^2\lambda_{\beta gj}^2\right)\\ \left[(\boldsymbol{\alpha}_a^k)_{gj} \mid (\boldsymbol{\beta}_m)_{gj}, \tau_{\alpha l}^2, \gamma_{\alpha g}^2, \lambda_{\alpha gj}^2\right] &\sim N\left(0, \tau_{\alpha k}^2\gamma_{\alpha g}^2\lambda_{\alpha gj}^2\left[f\{(\boldsymbol{\beta}_m)_{gj}\}\right]^2\right)\end{aligned}$$

In words, this prior specification applies global-local shrinkage on $\boldsymbol{\beta}_m$ and adaptive global-local shrinkage on $\boldsymbol{\alpha}_a$ (Kundu et al., 2021). Under this prior specification, we can show that

$$\left[\frac{(\boldsymbol{\alpha}_a^k)_{gj}}{f\{(\boldsymbol{\beta}_m)_{gj}\}} \;\middle|\; \tau_{\alpha k}^2, \gamma_{\alpha g}^2, \lambda_{\alpha gj}^2\right] \sim N(0, \tau_{\alpha k}^2\gamma_{\alpha g}^2\lambda_{\alpha gj}^2)$$

Therefore, if $f\{(\boldsymbol{\beta}_m)_{gj}\} = 1/(\boldsymbol{\beta}_m)_{gj}$, then

$$\left[(\boldsymbol{\alpha}_a^k)_{gj}(\boldsymbol{\beta}_m)_{gj} \mid \tau_{\alpha k}^2, \gamma_{\alpha g}^2, \lambda_{\alpha gj}^2\right] \sim N(0, \tau_{\alpha k}^2\gamma_{\alpha g}^2\lambda_{\alpha gj}^2)$$

Consequently, we call the following prior the product shrinkage prior

$$\begin{aligned}\left[(\boldsymbol{\beta}_m)_{gj} \mid \tau_\beta^2, \gamma_{\beta g}^2, \lambda_{\beta gj}^2\right] &\sim N\left(0, \tau_\beta^2\gamma_{\beta g}^2\lambda_{\beta gj}^2\right)\\ \left[(\boldsymbol{\alpha}_a^k)_{gj} \mid (\boldsymbol{\beta}_m)_{gj}, \tau_{\alpha l}^2, \gamma_{\alpha g}^2, \lambda_{\alpha gj}^2\right] &\sim N\left(0, \frac{\tau_{\alpha k}^2\gamma_{\alpha g}^2\lambda_{\alpha gj}^2}{(\boldsymbol{\beta}_m)_{gj}^2}\right)\end{aligned}$$

Note that this prior specification also motivates a family of two-step frequentist penalized regression estimators for targeted shrinkage of the product, where one can first fit a penalized regression model for the outcome and then subsequently fit an adaptively weighted penalized regression model for the mediators. This two-step frequentist analog has a similar flavor to penalized regression methods such as outcome-adaptive lasso (Shortreed and Ertefaie, 2017), with the key difference being that the adaptive weights in outcome-adaptive lasso reduce the shrinkage corresponding to large regression coefficients in the outcome model. Conversely, the adaptive weights in product shrinkage results in stronger shrinkage on $(\boldsymbol{\alpha}_a^k)_{gj}$ corresponding to larger $|(\boldsymbol{\beta}_m)_{gj}|$.

In this chapter, we also consider the independent shrinkage prior which corresponds to $f\{(\boldsymbol{\beta}_m)_{gj}\} = 1$

$$\begin{aligned}\left[(\boldsymbol{\beta}_m)_{gj} \mid \tau_\beta^2, \gamma_{\beta g}^2, \lambda_{\beta gj}^2\right] &\sim N\left(0, \tau_\beta^2\gamma_{\beta g}^2\lambda_{\beta gj}^2\right)\\ \left[(\boldsymbol{\alpha}_a^k)_{gj} \mid \tau_{\alpha l}^2, \gamma_{\alpha g}^2, \lambda_{\alpha gj}^2\right] &\sim N\left(0, \tau_{\alpha k}^2\gamma_{\alpha g}^2\lambda_{\alpha gj}^2\right)\end{aligned}$$

For both the product and independent shrinkage priors we specifically use GIGG shrinkage by

specifying the following hyperpriors for the global, group, and local shrinkage parameters

$$\gamma^2_{\beta g} \sim G(a_\beta, 1),\ \ \lambda^2_{\beta gj} \sim IG(b_\beta, 1),\ \ j = 1, \ldots, p_g$$

$$\left[\tau_\beta \mid \sigma_e\right] \sim C^+(0, \sigma_e),\ \ \pi(\sigma^2_e) \propto \sigma_e^{-2}$$

$$\tau_{\alpha k} \sim C^+(0, \eta),\ \ \eta^2 \sim IG(0.001, 0.001),\ \ \gamma^2_{\alpha g} \sim G(a_\alpha, 1),\ \ \lambda^2_{\alpha gj} \sim IG(b_\alpha, 1)$$

For the remaining model parameters in the mediation model component of model (4.2.1) we specify noninformative priors for the remaining regression coefficients, $\pi(\boldsymbol{\beta}_a) \propto 1$, $\pi(\boldsymbol{\beta}_c) \propto 1$, and $\pi(\boldsymbol{\alpha}_c) \propto 1$, and a block inverse-Wishart prior for the error variance-covariance matrix in the mediator model,

$$\boldsymbol{\Sigma}_m = \begin{pmatrix} (\boldsymbol{\Sigma}_m)_1 & \mathbf{0} & \ldots & \mathbf{0} \\ \mathbf{0} & \ddots & & \vdots \\ \vdots & & \ddots & \mathbf{0} \\ \mathbf{0} & \ldots & \mathbf{0} & (\boldsymbol{\Sigma}_m)_G \end{pmatrix}$$

where $(\boldsymbol{\Sigma}_m)_g \sim \mathcal{W}^{-1}(\boldsymbol{I}_{p_g}, p_g)$. For the latent factor model component of model (4.2.1) we assign a Jeffery's prior to the diagonal entries in $\boldsymbol{\Sigma}_L$ and a multiplicative gamma process prior to $\boldsymbol{\Psi}$ (Bhattacharya and Dunson, 2011):

$$\pi(\sigma^2_{Lk'}) \propto \sigma^{-2}_{Lk'},\ \ k' = 1, \ldots, q$$

$$\boldsymbol{\Psi}_{k'k} \mid \phi_{k'k}, \tau_k \sim N\left(0, \phi^{-1}_{k'k}\tau_k^{-1}\right),\ \ \phi_{k'k} \sim G(a_\phi/2, a_\phi/2),\ \ \tau_k = \prod_{m=1}^{k} \delta_m$$

$$\delta_1 \sim G(a_{\delta_1}, 1),\ \ \delta_m \sim G(a_{\delta_m}, 1),\ \ m \geq 2$$

The hyperparameters in the multiplicative gamma process prior are $a_\phi = 3$, $a_{\delta_1} = 2.1$, and $a_{\delta_m} = 3.1$, which are the default values for the Bayesian latent factor model in the infinitefactor R package. With these prior specifications, we now have a fully Bayesian specification of the mediation mixture map.

4.2.4 Computation

In this section, matrix notation is used to derive the posterior sampling algorithm. $\boldsymbol{Y} = (Y_1, \ldots, Y_n)^\top$ is the $n \times 1$ column vector containing the observed outcomes, $\boldsymbol{M}_{\cdot g} = (\boldsymbol{M}_{1g}^\top, \ldots, \boldsymbol{M}_{ng}^\top)^\top$ is the $n \times p_g$ matrix of mediators in group g, $\boldsymbol{M} = (\boldsymbol{M}_{\cdot 1}, \ldots, \boldsymbol{M}_{\cdot G})$ is the $n \times p$ matrix of continuous mediators, $\boldsymbol{A} = (\boldsymbol{A}_{1\cdot}^\top, \ldots, \boldsymbol{A}_{n\cdot}^\top)^\top$ is the $n \times q$ matrix of continuous chemical exposures, $\boldsymbol{L} = (\boldsymbol{L}_{1\cdot}^\top, \ldots, \boldsymbol{L}_{n\cdot}^\top)^\top$ is the unobserved $n \times r$ matrix of continuous latent

common sources of exposure variation, and $\boldsymbol{C} = (\boldsymbol{C}_{1\cdot}^\top, \ldots, \boldsymbol{C}_{n\cdot}^\top)^\top$ represents the $n \times s$ matrix of confounders and intercept term.

To simulate the posterior distribution for a fixed number of latent factors, we use a Gibbs sampling algorithm. A list of all full conditional distributions not discussed in this section can be found in Supplementary Material C.2. One subtlety with the Gibbs sampling algorithm for the product shrinkage prior is that the full conditional update of $\boldsymbol{\beta}_m$ does not have a closed-form and therefore requires a Metropolis-Hastings step within the Gibbs sampling algorithm. To circumvent adding a Metropolis-Hastings step we simultaneously sample $\boldsymbol{\alpha}_a$ and $\boldsymbol{\beta}_m$, by drawing from the joint full conditional distribution $\pi(\boldsymbol{\beta}_m, \boldsymbol{\alpha}_a \mid \cdot)$. The joint full conditional can be decomposed such that $\pi(\boldsymbol{\beta}_m, \boldsymbol{\alpha}_a \mid \cdot) = \pi(\boldsymbol{\beta}_m \mid \cdot)\pi(\boldsymbol{\alpha}_a \mid \boldsymbol{\beta}_m, \cdot)$, where $\pi(\boldsymbol{\beta}_m \mid \cdot)$ and $\pi(\boldsymbol{\alpha}_a \mid \boldsymbol{\beta}_m, \cdot)$ are both normal distributions. That is, unconditional on $\boldsymbol{\alpha}_a$,

$$[\boldsymbol{\beta}_m \mid \cdot] \sim N\left(\left[\frac{1}{\sigma_e^2}\boldsymbol{M}^\top\boldsymbol{M} + \frac{1}{\tau_\beta^2}\boldsymbol{\Gamma}_\beta^{-1}\boldsymbol{\Lambda}_\beta^{-1}\right]^{-1}\frac{1}{\sigma_e^2}\boldsymbol{M}^\top(\boldsymbol{Y} - \boldsymbol{L}\boldsymbol{\beta}_a - \boldsymbol{C}\boldsymbol{\beta}_c), \left[\frac{1}{\sigma_e^2}\boldsymbol{M}^\top\boldsymbol{M} + \frac{1}{\tau_\beta^2}\boldsymbol{\Gamma}_\beta^{-1}\boldsymbol{\Lambda}_\beta^{-1}\right]^{-1}\right)$$

$\boldsymbol{\Gamma}_\beta = \text{diag}(\gamma_{\beta 1}^2, \ldots, \gamma_{\beta 1}^2, \gamma_{\beta 2}^2, \ldots, \gamma_{\beta 2}^2, \ldots, \gamma_{\beta G}^2, \ldots, \gamma_{\beta G}^2)$ is a $p \times p$ matrix where $\gamma_{\beta g}^2$ is repeated along the diagonal p_g times and $\boldsymbol{\Lambda}_\beta = \text{diag}(\lambda_{\beta 11}^2, \ldots, \lambda_{\beta G p_G}^2)$ is a $p \times p$ matrix. Conditional on $\boldsymbol{\beta}_m$,

$$[\text{vec}(\boldsymbol{\alpha}_a^\top) \mid \boldsymbol{\beta}_m, \cdot] \sim N\Bigg(\left[\left(\boldsymbol{\Sigma}_m^{-1} \otimes \boldsymbol{L}^\top\boldsymbol{L}\right) + \left(\boldsymbol{\Gamma}_\alpha\boldsymbol{\Lambda}_\alpha\boldsymbol{F}_\beta \otimes \boldsymbol{T}_\alpha\right)^{-1}\right]^{-1}\left(\boldsymbol{\Sigma}_m^{-1} \otimes \boldsymbol{L}^\top\right)\text{vec}(\boldsymbol{M} - \boldsymbol{C}\boldsymbol{\alpha}_c^\top),$$
$$\left[\left(\boldsymbol{\Sigma}_m^{-1} \otimes \boldsymbol{L}^\top\boldsymbol{L}\right) + \left(\boldsymbol{\Gamma}_\alpha\boldsymbol{\Lambda}_\alpha\boldsymbol{F}_\beta \otimes \boldsymbol{T}_\alpha\right)^{-1}\right]^{-1}\Bigg)$$

where $\boldsymbol{T}_\alpha = \text{diag}(\tau_{\alpha 1}^2, \ldots, \tau_{\alpha r}^2)$ is an $r \times r$ matrix, $\boldsymbol{\Gamma}_\alpha = \text{diag}(\gamma_{\alpha 1}^2, \ldots, \gamma_{\alpha 1}^2, \gamma_{\alpha 2}^2, \ldots, \gamma_{\alpha 2}^2, \ldots, \gamma_{\alpha G}^2, \ldots, \gamma_{\alpha G}^2)$ is a $p \times p$ matrix where $\gamma_{\alpha g}^2$ is repeated along the diagonal p_g times, $\boldsymbol{\Lambda}_\alpha = \text{diag}(\lambda_{\alpha 11}^2, \ldots, \lambda_{\alpha G p_G}^2)$ is a $p \times p$ matrix, and $\boldsymbol{F}_\beta = \text{diag}[\{f((\boldsymbol{\beta}_m)_{11})\}^2, \ldots, \{f((\boldsymbol{\beta}_m)_{G p_G})\}^2]$ is a $p \times p$ matrix. The vectorization operation in this case refers to stacking the columns of a $\boldsymbol{\alpha}_a^\top$ to create a long column vector.

For our mediation mixture map implementation we set $a_\alpha = a_\beta = 0.5$, and use Marginal Maximum Likelihood Estimation (MMLE) to estimate b_α and b_β (Casella, 2001). The $(l+1)$th MMLE updates are

$$b_\alpha^{(l+1)} = \psi_0^{-1}\left(-\frac{1}{p}\sum_{g=1}^{G}\sum_{j=1}^{p_g} E_{b_\alpha^{(l)}}\left[\log(\lambda_{\alpha gj}^2) \mid \boldsymbol{Y}, \boldsymbol{M}, \boldsymbol{A}, \boldsymbol{L}, \boldsymbol{C}\right]\right)$$
$$b_\beta^{(l+1)} = \psi_0^{-1}\left(-\frac{1}{p}\sum_{g=1}^{G}\sum_{j=1}^{p_g} E_{b_\beta^{(l)}}\left[\log(\lambda_{\beta gj}^2) \mid \boldsymbol{Y}, \boldsymbol{M}, \boldsymbol{A}, \boldsymbol{L}, \boldsymbol{C}\right]\right),$$

where $\psi_0(\cdot)$ is the digamma function. Once posterior samples are obtained, the Varimax-based post-processing algorithm described in Poworoznek et al. (2021) is deployed to resolve rotational ambiguity, sign-switching, and label-switching with respect to model parameters $\boldsymbol{\alpha}_a$, $\boldsymbol{\beta}_a$, $\boldsymbol{\Psi}$, and $\boldsymbol{L}$. The $\text{DE}^A(\boldsymbol{a}, \boldsymbol{a}^*)$, $\text{IE}^A(\boldsymbol{a}, \boldsymbol{a}^*)$, and $\text{TE}^A(\boldsymbol{a}, \boldsymbol{a}^*)$ estimators are the posterior mean estimators, denoted by $\widehat{\text{DE}}^A(\boldsymbol{a}, \boldsymbol{a}^*)$, $\widehat{\text{IE}}^A(\boldsymbol{a}, \boldsymbol{a}^*)$, and $\widehat{\text{TE}}^A(\boldsymbol{a}, \boldsymbol{a}^*)$. Similarly, the $\text{DE}^L(\boldsymbol{l}, \boldsymbol{l}^*)$, $\text{IE}^L(\boldsymbol{l}, \boldsymbol{l}^*)$, and $\text{TE}^L(\boldsymbol{l}, \boldsymbol{l}^*)$ estimators are the posterior mean estimators, denoted by $\widehat{\text{DE}}^L(\boldsymbol{l}, \boldsymbol{l}^*)$, $\widehat{\text{IE}}^L(\boldsymbol{l}, \boldsymbol{l}^*)$, and $\widehat{\text{TE}}^L(\boldsymbol{l}, \boldsymbol{l}^*)$.

When the number of latent factors requires estimation, we specify a grid of possible values for the number of exposure mixtures and use the Bayesian Information Criterion (BIC) as a model selection tool to determine the number of exposure mixtures (Schwarz, 1978). For the purposes of calculating BIC, we do not count $\boldsymbol{L}$ towards the number of model parameters in (4.2.1). For mixtures analyses, it is recommended to consider a grid of values that contain some values greater than or equal to the number of exposure classes represented in the chemical exposure data.

4.3 Simulations

4.3.1 Generative Models and Simulation Design

The primary goals of the simulation study are to empirically show that the mediation mixture map improves estimation of the true number of latent factors relative to the naive two-step estimator and, conditional on the true number of latent factors, jointly modeling the mediation and latent factor models more appropriately quantifies uncertainty corresponding to DE, IE, and TE estimation. We consider two different sample sizes in the simulation study, $n = 400$ and $n = 2000$. A sample size of $n = 400$ is generally more realistic in applied practice, however the $n = 2000$ simulation settings are useful for evaluating the large sample performance of the mediation mixture map and naive two-step estimators. For both sample sizes, we consider settings with $p = 50$ mediators, $q = 40$ chemical exposures, $s = 5$ confounders, and $r = 4$ true latent chemical mixtures. The generative model for the simulated data is (4.2.1), where $\boldsymbol{C}_{i\cdot}^\top \sim N(\boldsymbol{0}_s, \boldsymbol{I}_s)$. The loadings matrix in the latent factor model is set to

$$\boldsymbol{\Psi} = \begin{pmatrix} 0.1 & 0 & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots \\ 0.1 & 0 & 0 & 0 \\ 0 & 0.1 & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots \\ 0 & 0.1 & 0 & 0 \\ 0 & 0 & 0.1 & 0 \\ \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & 0.1 & 0 \\ 0 & 0 & 0 & 0.1 \\ \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & 0.1 \end{pmatrix}$$

where the 0.1 loadings are repeated 10 times, and the error variances in the latent factor model are specified such that the R^2s for all chemical exposures in the latent factor model is equal to either 0.5 or 0.1,

$$\boldsymbol{\Sigma}_L = \frac{\left(1 - R_L^2\right)}{R_L^2} \text{diag}\left(\boldsymbol{\Psi}_{1\cdot}\boldsymbol{\Psi}_{1\cdot}^\top, \ldots, \boldsymbol{\Psi}_{q\cdot}\boldsymbol{\Psi}_{q\cdot}^\top\right), \quad R_L^2 = 0.5 \text{ or } R_L^2 = 0.1.$$

In practice, 0.5 is a more realistic value for R_L^2 to take.

For the mediation model, we consider several choices of $\boldsymbol{\alpha}_a$ and $\boldsymbol{\beta}_m$. The first specification of $\boldsymbol{\alpha}_a$ and $\boldsymbol{\beta}_m$ is a setting with a zero IE

$$\boldsymbol{\alpha}_a = \begin{pmatrix} 0 & 0 & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & 0 \\ 0.1 & 0.1 & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots \\ 0.1 & 0.1 & 0 & 0 \end{pmatrix}, \quad \boldsymbol{\beta}_m = \begin{pmatrix} 0.2 \\ \vdots \\ 0.2 \\ 0 \\ \vdots \\ 0 \end{pmatrix}$$

where the 0.1, 0.2, and 0 coefficients are repeated 25 times. We refer to this setting as the IE zero setting. The other specifications of $\boldsymbol{\alpha}_a$ and $\boldsymbol{\beta}_m$ are cases where the IE is nonzero, but the signal density varies. That is, the second specification corresponds to dense contributions to the global IE

$$\boldsymbol{\alpha}_a = \begin{pmatrix} 0.15 & 0.15 & 0.15 & 0.15 \\ \vdots & \vdots & \vdots & \vdots \\ 0.15 & 0.15 & 0.15 & 0.15 \\ 0 & 0 & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & 0 \end{pmatrix}, \quad \boldsymbol{\beta}_m = \begin{pmatrix} \frac{2}{15} \\ \vdots \\ \frac{2}{15} \\ 0 \\ \vdots \\ 0 \end{pmatrix}$$

where the 0.15 coefficients are repeated 10 times, the $2/15$ coefficients are repeated 10 times, and the 0 coefficients are repeated 30 times. We refer to this setting as the IE dense setting. The third specification corresponds to sparse contributions to the global IE

$$\boldsymbol{\alpha}_a = \begin{pmatrix} \frac{1}{3} & \frac{1}{3} & \frac{1}{3} & \frac{1}{3} \\ \frac{1}{3} & \frac{1}{3} & \frac{1}{3} & \frac{1}{3} \\ 0 & 0 & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & 0 \end{pmatrix}, \quad \boldsymbol{\beta}_m = \begin{pmatrix} 0.3 \\ 0.3 \\ 0 \\ \vdots \\ 0 \end{pmatrix}$$

where the 0 coefficients are repeated 48 times. We call this setting the IE sparse setting. Note that these choices of $\boldsymbol{\alpha}_a$ and $\boldsymbol{\beta}_m$ result in $\boldsymbol{\alpha}_a^\top \boldsymbol{\beta}_m = (0.2, 0.2, 0.2, 0.2)^\top$ for the IE dense and the IE sparse settings. The DE is set to $\boldsymbol{\beta}_a = (1,1,1,1)^\top - \boldsymbol{\alpha}_a^\top \boldsymbol{\beta}_m$ so that the TE is equal to $\boldsymbol{\beta}_a + \boldsymbol{\alpha}_a^\top \boldsymbol{\beta}_m = (1,1,1,1)^\top$. The regression coefficient parameters corresponding to the confounders are

$$\boldsymbol{\alpha}_c = \begin{pmatrix} 0.1 & \dots & 0.1 \\ \vdots & & \vdots \\ 0.1 & \dots & 0.1 \end{pmatrix}, \quad \boldsymbol{\beta}_c = \begin{pmatrix} 0.1 \\ \vdots \\ 0.1 \end{pmatrix}$$

For $\boldsymbol{\Sigma}_m$, the correlation structure is block diagonal where the blocks are defined by groups of five mediators. The correlation structure within blocks is exchangable, with correlations equal to 0.3. The variances are determined by setting the R^2 squared of the mediator model to 0.1

$$\frac{\left(1 - R_M^2\right)}{R_M^2} \begin{pmatrix} \boldsymbol{\alpha}_a & \boldsymbol{\alpha}_c \end{pmatrix} \begin{pmatrix} \boldsymbol{\alpha}_a & \boldsymbol{\alpha}_c \end{pmatrix}^\top, \quad R_M^2 = 0.1.$$

The error variance in the outcome model is determined by setting the adjusted R^2 of the outcome model to 0.3 or 0.6

$$\sigma_e^2 = \frac{\left(1 - R_O^2\right)}{R_O^2} \boldsymbol{\beta}^\top \boldsymbol{V} \boldsymbol{\beta}, \quad R_O^2 = 1 - \frac{(n - p - r - s)}{n - 1}(1 - \text{Adjusted } R_O^2),$$

$$\text{Adjusted } R^2_O = 0.3 \text{ or Adjusted } R^2_O = 0.6$$

where $\boldsymbol{\beta}^\top = (\boldsymbol{\beta}_m^\top, \boldsymbol{\beta}_a^\top, \boldsymbol{\beta}_c^\top)^\top$ and $\boldsymbol{V}$ is the variance-covariance matrix of $\boldsymbol{M}$, $\boldsymbol{L}$, and $\boldsymbol{C}$. Setting the Adjusted R^2_O to 0.3 is more realistic in practice, however $R^2_O = 0.6$ is useful for understanding the properties of methods when the signal-to-noise-ratio is larger.

For each combination of sample size, $n = 400$ and $n = 2000$, and choice of $\boldsymbol{\alpha}_a$ and $\boldsymbol{\beta}_m$, we consider two different simulation types. The first type assumes that we do not know the true number of latent factors and the second type assumes that we do know the true number of latent factors. For the simulations that assume we do not know the true number of latent factors, we will evaluate the one-step and two-step estimators in terms of their ability to correctly estimate the true number of latent factors. Given the correct number of latent factors we will evaluate estimation and inference corresponding to $\text{DE}^L(\boldsymbol{l}, \boldsymbol{l}^*)$, $\text{IE}^L(\boldsymbol{l}, \boldsymbol{l}^*)$, and $\text{TE}^L(\boldsymbol{l}, \boldsymbol{l}^*)$ for $(\boldsymbol{l} - \boldsymbol{l}^*) = (1, 0, 0, 0)^\top$, $(\boldsymbol{l} - \boldsymbol{l}^*) = (0, 1, 0, 0)^\top$, $(\boldsymbol{l} - \boldsymbol{l}^*) = (0, 0, 1, 0)^\top$, and $(\boldsymbol{l} - \boldsymbol{l}^*) = (0, 0, 0, 1)^\top$ using bias of the posterior mean estimator, root mean-squared error (RMSE), posterior standard deviation (SD), and coverage probabilities for 95% credible intervals. All results for each simulation type and setting corresponds to 1000 simulation replications.

4.3.2 Simulation Results

When the true number of latent factors is known, Table 4.1 shows the simulation results corresponding to the $n = 400$ IE zero setting where $R^2_L = 0.5$ and Adjusted $R^2_O = 0.3$. In general we observe that across all combinations of shrinkage priors and mediation parameters the one-step estimator always has larger posterior SD than the two-step estimator. This is expected because the one-step estimator accounts for uncertainty in $\boldsymbol{\Psi}$, $\boldsymbol{L}$, and $\boldsymbol{\Sigma}_L$, while the two-step estimator does not. The empirical bias for the one-step and two-step estimator is comparable, however it is worth noting that bias is very small for IE estimation regardless of the method used. Small bias for IE estimation is due to the fact that the independent and product shrinkage priors shrink the product terms toward zero, which in the IE zero setting happens to be the true value of the IE. Because the empirical bias is comparable between the one-step and two-step methods and the posterior SD is slightly larger for the one-step estimator, then the RMSE ends up being larger for the one-step estimator. Despite the RMSE being larger for the one-step estimator, we see that empirical coverage probabilities for the DE and TE for the one-step estimator range between 0.93-0.95, while empirical coverage probabilities for the DE and TE for the two-step estimator range between 0.92-0.94. Therefore, the two-step estimator slightly undercovers with respect to the nominal coverage rate. In contrast, the empirical coverage probabilities for the IE are 1 for both one-step and two-step estimators. Looking at Supplementary Table C.1, which provides the simulation results for the $n = 2000$ IE zero setting where $R^2_L = 0.5$ and Adjusted $R^2_O = 0.3$, we see nearly identical results

to the $n = 400$ IE zero setting, but with biases for DE and TE estimation attenuating towards zero.

Tables 4.2 and 4.3 show the simulation results corresponding to the $n = 400$ IE dense and IE sparse settings where $R^2_L = 0.5$ and Adjusted $R^2_O = 0.3$, respectively. For both settings we observe substantial differences in performance between independent shrinkage and product shrinkage, with product shrinkage resulting in larger absolute bias than independent shrinkage. For example, in the IE dense setting, biases for one-step IE estimation tend to be around -0.17 and biases for two-step IE estimation tend to be around -0.20 for product shrinkage, while biases for one-step IE estimation tend to be around -0.10 and biases for two-step IE estimation tend to be around -0.16 for independent shrinkage. Recall that the true IE value in this case is 0.20, implying that product shrinkage is much more aggressively shrinking the IE towards zero than independent shrinkage. For the one-step and two-step estimators, we see that empirical biases in the IE sparse setting are comparable, however, in the IE dense setting, the one-step estimator has notably lower bias than the two-step estimator. Biased estimation as a result of assigning shrinkage priors to $\boldsymbol{\beta}_m$ and $\boldsymbol{\alpha}_a$ leads to lower than nominal coverage for both independent shrinkage and product shrinkage, with empirical coverage being higher for independent shrinkage. As with the IE zero simulation settings, we also observe that the one-step estimator has larger posterior SDs across the board, compared with the two-step estimator. That being said, for independent shrinkage, the RMSE is lower for the one-step estimator due to the one-step estimator being less biased. This is particularly pronounced in the IE dense simulation setting with independent shrinkage where the one-step estimator has an RMSE of 0.11 for each choice of $\boldsymbol{l} - \boldsymbol{l}^*$ and the two-step estimator has an RMSE of 0.16 for each choice of $\boldsymbol{l} - \boldsymbol{l}^*$. Therefore, we conclude that combining the mediation and latent factor models together particularly benefits the setting where the contributions to the global IE are dense. Supplementary Tables C.2 and C.3 show the simulation results corresponding to the $n = 2000$ IE dense and IE sparse settings when $R^2_L = 0.5$ and Adjusted $R^2_O = 0.3$, respectively. The primary difference between the $n = 400$ and $n = 2000$ settings, is that the bias attenuates towards zero across the board when the sample size is larger, resulting in higher empirical coverage probabilities.

			One-Step				Two-Step			
Shrinkage	**Parameter**	$(\boldsymbol{l}-\boldsymbol{l}^*)^\top$	**Bias**	**RMSE**	**Posterior SD**	**Coverage**	**Bias**	**RMSE**	**Posterior SD**	**Coverage**
Independent	$\text{DE}^L(\boldsymbol{l},\boldsymbol{l}^*)$	(1,0,0,0)	0.04	0.21	0.20	0.93	0.03	0.20	0.19	0.92
Independent	$\text{DE}^L(\boldsymbol{l},\boldsymbol{l}^*)$	(0,1,0,0)	0.06	0.20	0.20	0.95	0.05	0.20	0.19	0.94
Independent	$\text{DE}^L(\boldsymbol{l},\boldsymbol{l}^*)$	(0,0,1,0)	0.07	0.20	0.20	0.94	0.07	0.20	0.18	0.93
Independent	$\text{DE}^L(\boldsymbol{l},\boldsymbol{l}^*)$	(0,0,0,1)	0.06	0.20	0.20	0.94	0.05	0.19	0.18	0.93
Product	$\text{DE}^L(\boldsymbol{l},\boldsymbol{l}^*)$	(1,0,0,0)	0.04	0.21	0.20	0.93	0.03	0.20	0.19	0.93
Product	$\text{DE}^L(\boldsymbol{l},\boldsymbol{l}^*)$	(0,1,0,0)	0.05	0.20	0.20	0.95	0.05	0.20	0.19	0.94
Product	$\text{DE}^L(\boldsymbol{l},\boldsymbol{l}^*)$	(0,0,1,0)	0.07	0.20	0.20	0.94	0.07	0.20	0.18	0.93
Product	$\text{DE}^L(\boldsymbol{l},\boldsymbol{l}^*)$	(0,0,0,1)	0.06	0.20	0.19	0.94	0.05	0.19	0.18	0.93
Independent	$\text{IE}^L(\boldsymbol{l},\boldsymbol{l}^*)$	(1,0,0,0)	0.00	0.02	0.04	1.00	0.00	0.02	0.04	1.00
Independent	$\text{IE}^L(\boldsymbol{l},\boldsymbol{l}^*)$	(0,1,0,0)	0.00	0.02	0.05	1.00	0.00	0.02	0.04	1.00
Independent	$\text{IE}^L(\boldsymbol{l},\boldsymbol{l}^*)$	(0,0,1,0)	0.00	0.00	0.02	1.00	0.00	0.00	0.01	1.00
Independent	$\text{IE}^L(\boldsymbol{l},\boldsymbol{l}^*)$	(0,0,0,1)	0.00	0.00	0.02	1.00	0.00	0.00	0.01	1.00
Product	$\text{IE}^L(\boldsymbol{l},\boldsymbol{l}^*)$	(1,0,0,0)	0.00	0.01	0.01	1.00	0.00	0.00	0.00	1.00
Product	$\text{IE}^L(\boldsymbol{l},\boldsymbol{l}^*)$	(0,1,0,0)	0.00	0.01	0.01	1.00	0.00	0.00	0.00	1.00
Product	$\text{IE}^L(\boldsymbol{l},\boldsymbol{l}^*)$	(0,0,1,0)	0.00	0.00	0.00	1.00	0.00	0.00	0.00	1.00
Product	$\text{IE}^L(\boldsymbol{l},\boldsymbol{l}^*)$	(0,0,0,1)	0.00	0.00	0.00	1.00	0.00	0.00	0.00	1.00
Independent	$\text{TE}^L(\boldsymbol{l},\boldsymbol{l}^*)$	(1,0,0,0)	0.04	0.21	0.20	0.93	0.03	0.20	0.19	0.92
Independent	$\text{TE}^L(\boldsymbol{l},\boldsymbol{l}^*)$	(0,1,0,0)	0.06	0.20	0.20	0.94	0.05	0.20	0.19	0.94
Independent	$\text{TE}^L(\boldsymbol{l},\boldsymbol{l}^*)$	(0,0,1,0)	0.07	0.20	0.20	0.94	0.07	0.20	0.18	0.93
Independent	$\text{TE}^L(\boldsymbol{l},\boldsymbol{l}^*)$	(0,0,0,1)	0.06	0.20	0.20	0.94	0.05	0.19	0.18	0.93
Product	$\text{TE}^L(\boldsymbol{l},\boldsymbol{l}^*)$	(1,0,0,0)	0.04	0.21	0.20	0.93	0.03	0.20	0.19	0.92
Product	$\text{TE}^L(\boldsymbol{l},\boldsymbol{l}^*)$	(0,1,0,0)	0.05	0.20	0.20	0.95	0.05	0.20	0.19	0.94
Product	$\text{TE}^L(\boldsymbol{l},\boldsymbol{l}^*)$	(0,0,1,0)	0.07	0.20	0.20	0.94	0.07	0.20	0.18	0.93
Product	$\text{TE}^L(\boldsymbol{l},\boldsymbol{l}^*)$	(0,0,0,1)	0.06	0.20	0.19	0.94	0.05	0.19	0.18	0.93

Table 4.1: Simulation results for the $n = 400$ IE zero simulation setting for the one-step and two-step estimators of $\text{DE}^L(\boldsymbol{l},\boldsymbol{l}^*)$, $\text{IE}^L(\boldsymbol{l},\boldsymbol{l}^*)$, and $\text{TE}^L(\boldsymbol{l},\boldsymbol{l}^*)$ when $R^2_L = 0.5$ and Adjusted $R^2_O = 0.3$.

			One-Step				Two-Step			
Shrinkage	**Parameter**	$(\boldsymbol{l}-\boldsymbol{l}^*)^\top$	**Bias**	**RMSE**	**Posterior SD**	**Coverage**	**Bias**	**RMSE**	**Posterior SD**	**Coverage**
Independent	$\text{DE}^L(\boldsymbol{l},\boldsymbol{l}^*)$	(1,0,0,0)	0.11	0.21	0.17	0.88	0.08	0.19	0.16	0.89
Independent	$\text{DE}^L(\boldsymbol{l},\boldsymbol{l}^*)$	(0,1,0,0)	0.13	0.21	0.17	0.86	0.10	0.19	0.16	0.89
Independent	$\text{DE}^L(\boldsymbol{l},\boldsymbol{l}^*)$	(0,0,1,0)	0.13	0.21	0.17	0.88	0.10	0.19	0.16	0.90
Independent	$\text{DE}^L(\boldsymbol{l},\boldsymbol{l}^*)$	(0,0,0,1)	0.12	0.21	0.17	0.88	0.09	0.19	0.16	0.90
Product	$\text{DE}^L(\boldsymbol{l},\boldsymbol{l}^*)$	(1,0,0,0)	0.10	0.20	0.17	0.89	0.08	0.19	0.16	0.89
Product	$\text{DE}^L(\boldsymbol{l},\boldsymbol{l}^*)$	(0,1,0,0)	0.11	0.20	0.17	0.89	0.09	0.19	0.16	0.89
Product	$\text{DE}^L(\boldsymbol{l},\boldsymbol{l}^*)$	(0,0,1,0)	0.12	0.20	0.17	0.90	0.10	0.19	0.16	0.90
Product	$\text{DE}^L(\boldsymbol{l},\boldsymbol{l}^*)$	(0,0,0,1)	0.10	0.20	0.17	0.90	0.09	0.19	0.16	0.90
Independent	$\text{IE}^L(\boldsymbol{l},\boldsymbol{l}^*)$	(1,0,0,0)	-0.10	0.11	0.05	0.54	-0.16	0.16	0.04	0.17
Independent	$\text{IE}^L(\boldsymbol{l},\boldsymbol{l}^*)$	(0,1,0,0)	-0.10	0.11	0.05	0.54	-0.16	0.16	0.04	0.18
Independent	$\text{IE}^L(\boldsymbol{l},\boldsymbol{l}^*)$	(0,0,1,0)	-0.09	0.11	0.05	0.55	-0.16	0.16	0.04	0.19
Independent	$\text{IE}^L(\boldsymbol{l},\boldsymbol{l}^*)$	(0,0,0,1)	-0.10	0.11	0.05	0.55	-0.16	0.16	0.04	0.19
Product	$\text{IE}^L(\boldsymbol{l},\boldsymbol{l}^*)$	(1,0,0,0)	-0.17	0.17	0.02	0.16	-0.20	0.20	0.00	0.00
Product	$\text{IE}^L(\boldsymbol{l},\boldsymbol{l}^*)$	(0,1,0,0)	-0.17	0.17	0.02	0.16	-0.20	0.20	0.00	0.00
Product	$\text{IE}^L(\boldsymbol{l},\boldsymbol{l}^*)$	(0,0,1,0)	-0.16	0.17	0.02	0.16	-0.20	0.20	0.00	0.00
Product	$\text{IE}^L(\boldsymbol{l},\boldsymbol{l}^*)$	(0,0,0,1)	-0.17	0.17	0.02	0.17	-0.20	0.20	0.00	0.00
Independent	$\text{TE}^L(\boldsymbol{l},\boldsymbol{l}^*)$	(1,0,0,0)	0.02	0.18	0.17	0.93	-0.08	0.19	0.16	0.90
Independent	$\text{TE}^L(\boldsymbol{l},\boldsymbol{l}^*)$	(0,1,0,0)	0.03	0.18	0.17	0.94	-0.06	0.18	0.16	0.92
Independent	$\text{TE}^L(\boldsymbol{l},\boldsymbol{l}^*)$	(0,0,1,0)	0.04	0.17	0.17	0.95	-0.05	0.18	0.16	0.92
Independent	$\text{TE}^L(\boldsymbol{l},\boldsymbol{l}^*)$	(0,0,0,1)	0.03	0.17	0.17	0.95	-0.07	0.18	0.16	0.91
Product	$\text{TE}^L(\boldsymbol{l},\boldsymbol{l}^*)$	(1,0,0,0)	-0.07	0.20	0.17	0.90	-0.12	0.21	0.16	0.86
Product	$\text{TE}^L(\boldsymbol{l},\boldsymbol{l}^*)$	(0,1,0,0)	-0.06	0.19	0.17	0.92	-0.11	0.20	0.16	0.90
Product	$\text{TE}^L(\boldsymbol{l},\boldsymbol{l}^*)$	(0,0,1,0)	-0.05	0.18	0.17	0.92	-0.10	0.19	0.16	0.90
Product	$\text{TE}^L(\boldsymbol{l},\boldsymbol{l}^*)$	(0,0,0,1)	-0.06	0.19	0.17	0.92	-0.11	0.20	0.16	0.89

Table 4.2: Simulation results for the $n = 400$ IE dense simulation setting for the one-step and two-step estimators of $\text{DE}^L(\boldsymbol{l},\boldsymbol{l}^*)$, $\text{IE}^L(\boldsymbol{l},\boldsymbol{l}^*)$, and $\text{TE}^L(\boldsymbol{l},\boldsymbol{l}^*)$ when $R^2_L = 0.5$ and Adjusted $R^2_O = 0.3$.

			One-Step				Two-Step			
Shrinkage	**Parameter**	$(\boldsymbol{l}-\boldsymbol{l}^*)^\top$	**Bias**	**RMSE**	**Posterior SD**	**Coverage**	**Bias**	**RMSE**	**Posterior SD**	**Coverage**
Independent	$\text{DE}^L(\boldsymbol{l},\boldsymbol{l}^*)$	(1,0,0,0)	0.06	0.19	0.17	0.92	0.04	0.18	0.16	0.91
Independent	$\text{DE}^L(\boldsymbol{l},\boldsymbol{l}^*)$	(0,1,0,0)	0.07	0.19	0.17	0.92	0.06	0.18	0.16	0.92
Independent	$\text{DE}^L(\boldsymbol{l},\boldsymbol{l}^*)$	(0,0,1,0)	0.08	0.18	0.17	0.93	0.06	0.18	0.16	0.94
Independent	$\text{DE}^L(\boldsymbol{l},\boldsymbol{l}^*)$	(0,0,0,1)	0.07	0.18	0.17	0.94	0.05	0.17	0.16	0.94
Product	$\text{DE}^L(\boldsymbol{l},\boldsymbol{l}^*)$	(1,0,0,0)	0.06	0.19	0.17	0.92	0.04	0.18	0.16	0.91
Product	$\text{DE}^L(\boldsymbol{l},\boldsymbol{l}^*)$	(0,1,0,0)	0.07	0.18	0.17	0.92	0.06	0.18	0.16	0.92
Product	$\text{DE}^L(\boldsymbol{l},\boldsymbol{l}^*)$	(0,0,1,0)	0.07	0.18	0.17	0.94	0.06	0.18	0.16	0.94
Product	$\text{DE}^L(\boldsymbol{l},\boldsymbol{l}^*)$	(0,0,0,1)	0.06	0.18	0.17	0.94	0.05	0.17	0.16	0.93
Independent	$\text{IE}^L(\boldsymbol{l},\boldsymbol{l}^*)$	(1,0,0,0)	-0.04	0.07	0.06	0.88	-0.04	0.08	0.06	0.87
Independent	$\text{IE}^L(\boldsymbol{l},\boldsymbol{l}^*)$	(0,1,0,0)	-0.04	0.07	0.06	0.89	-0.04	0.07	0.06	0.89
Independent	$\text{IE}^L(\boldsymbol{l},\boldsymbol{l}^*)$	(0,0,1,0)	-0.04	0.07	0.06	0.89	-0.04	0.08	0.06	0.88
Independent	$\text{IE}^L(\boldsymbol{l},\boldsymbol{l}^*)$	(0,0,0,1)	-0.04	0.07	0.06	0.87	-0.04	0.08	0.06	0.87
Product	$\text{IE}^L(\boldsymbol{l},\boldsymbol{l}^*)$	(1,0,0,0)	-0.11	0.14	0.04	0.54	-0.13	0.15	0.04	0.43
Product	$\text{IE}^L(\boldsymbol{l},\boldsymbol{l}^*)$	(0,1,0,0)	-0.10	0.14	0.04	0.55	-0.13	0.15	0.04	0.43
Product	$\text{IE}^L(\boldsymbol{l},\boldsymbol{l}^*)$	(0,0,1,0)	-0.10	0.14	0.04	0.56	-0.12	0.15	0.04	0.43
Product	$\text{IE}^L(\boldsymbol{l},\boldsymbol{l}^*)$	(0,0,0,1)	-0.11	0.14	0.04	0.55	-0.13	0.15	0.04	0.42
Independent	$\text{TE}^L(\boldsymbol{l},\boldsymbol{l}^*)$	(1,0,0,0)	0.02	0.18	0.18	0.94	0.00	0.18	0.17	0.92
Independent	$\text{TE}^L(\boldsymbol{l},\boldsymbol{l}^*)$	(0,1,0,0)	0.04	0.18	0.18	0.94	0.02	0.18	0.17	0.93
Independent	$\text{TE}^L(\boldsymbol{l},\boldsymbol{l}^*)$	(0,0,1,0)	0.04	0.18	0.18	0.95	0.02	0.17	0.17	0.94
Independent	$\text{TE}^L(\boldsymbol{l},\boldsymbol{l}^*)$	(0,0,0,1)	0.03	0.17	0.18	0.95	0.01	0.17	0.17	0.94
Product	$\text{TE}^L(\boldsymbol{l},\boldsymbol{l}^*)$	(1,0,0,0)	-0.05	0.20	0.18	0.92	-0.09	0.21	0.17	0.88
Product	$\text{TE}^L(\boldsymbol{l},\boldsymbol{l}^*)$	(0,1,0,0)	-0.04	0.19	0.18	0.92	-0.07	0.20	0.17	0.90
Product	$\text{TE}^L(\boldsymbol{l},\boldsymbol{l}^*)$	(0,0,1,0)	-0.03	0.18	0.18	0.94	-0.06	0.19	0.17	0.91
Product	$\text{TE}^L(\boldsymbol{l},\boldsymbol{l}^*)$	(0,0,0,1)	-0.04	0.18	0.18	0.94	-0.08	0.19	0.17	0.89

Table 4.3: Simulation results for the $n = 400$ IE sparse simulation setting for the one-step and two-step estimators of $\text{DE}^L(\boldsymbol{l},\boldsymbol{l}^*)$, $\text{IE}^L(\boldsymbol{l},\boldsymbol{l}^*)$, and $\text{TE}^L(\boldsymbol{l},\boldsymbol{l}^*)$ when $R_L^2 = 0.5$ and Adjusted $R_O^2 = 0.3$.

Simulation results for the $n = 400$ settings where $R_L^2 = 0.5$ and Adjusted $R_O^2 = 0.6$ are listed in Tables C.4, C.5, and C.6. Simulation results for the $n = 400$ settings where $R_L^2 = 0.1$ and Adjusted $R_O^2 = 0.6$ are listed in Tables C.7, C.8, and C.9. Simulation results for the $n = 400$ settings where $R_L^2 = 0.1$ and Adjusted $R_O^2 = 0.3$ are listed in Tables C.10, C.11, and C.12. In general, conclusions for the simulation settings with $R_L^2 = 0.5$ and Adjusted $R_O^2 = 0.6$ were comparable to the conclusions for the simulation settings with $R_L^2 = 0.5$ and Adjusted $R_O^2 = 0.3$. However, there are some differences between the $R_L^2 = 0.5$ simulation settings and the $R_L^2 = 0.1$ simulation settings. The primary difference is that, when $R_L^2 = 0.1$, there is more uncertainty in estimating the latent factors and loadings matrix. Consequently, the RMSEs corresponding to the one-step estimator with independent shrinkage are generally lower than the RMSEs corresponding to the two-step estimator with independent shrinkage, despite the fact that the average posterior standard deviation is generally higher for the one-step estimator. For example, in the $n = 400$ IE

sparse setting where $R^2_L = 0.1$ and Adjusted $R^2_O = 0.6$, the RMSE corresponding to TE estimation is 0.20 for the one-step estimator and 0.23 for the two-step estimator. This occurs despite the fact that the average posterior standard deviation for the one-step estimator is 0.19 and the average posterior standard deviation for the two-step estimator is 0.17. Therefore, outcome and mediator adaptive estimation of the underlying latent factors is particularly helpful for estimating mediation parameters when R^2_L is small.

Table 4.4 shows the results of estimating the number of latent factors via BIC when $R^2_L = 0.5$ and Adjusted $R^2_O = 0.3$. Note that regardless of the shrinkage prior, the one-step estimator for $n = 400$ almost always identifies the correct number of latent factors $r = 4$. However, the two-step estimator regardless of sample size tends to overestimate the number of latent factors. Namely, when $n = 400$, 22.7% of simulation replications result in the two-step estimator identifying four latent factors, while 55.0% of simulation replications result in the two-step estimator identifying five latent factors. When $n = 2000$, both the one-step estimator correctly estimates the true number of latent factors approximately 60% of the time, while the two-step estimator correctly estimates the true number of latent factors 1.7% of the time. These phenomena are observed for all other simulation settings (see Tables C.13 and C.14). Therefore, imparting information from the mediation model into the latent factor model helps stabilize estimation of the true number of latent factors. It is important to note, that although there is systematic overestimation of the true number of latent factors for the one-step and two-step estimators in the $n = 2000$ case, the extraneous latent factors tend to correspond to columns in the loading matrix close to the zero vector. That is, the loadings corresponding to the extraneous factors tend to be shrunk to zero by the multiplicative gamma process shrinkage prior.

	Estimated Number of Latent Factors			
Method	**4***	**5**	**6**	**7**
$n = 400$				
One-Step Independent	999 (99.9%)	1 (0.1%)	0 (0.0%)	0 (0.0%)
One-Step Product	989 (98.9%)	10 (1.0%)	0 (0.0%)	1 (0.1%)
Two-Step	227 (22.7%)	550 (55.0%)	204 (20.4%)	19 (1.9%)
$n = 2000$				
One-Step Independent	596 (59.6%)	377 (37.7%)	27 (2.7%)	0 (0.0%)
One-Step Product	599 (59.9%)	372 (37.2%)	29 (2.9%)	0 (0.0%)
Two-Step	17 (1.7%)	310 (31.0%)	378 (37.8%)	295 (29.5%)

Table 4.4: Results for estimating of the number of latent factors under the IE dense simulation settings when $R^2_L = 0.5$ and Adjusted $R^2_O = 0.3$. The numbers in the table indicate the number of times out of 1000 simulation replications each number of latent factors was selected via BIC. *The true number of latent factors is $r = 4$.

4.4 Data Example

4.4.1 Data Processing

The PROTECT multi-exposure analysis in this chapter considers the TE of chemical exposures across four exposure classes, metals, phthalates, phenols and parabens, and PAHs, on two birth outcomes, gestational age at delivery and head circumference z-score. Gestational age at delivery measures the duration of pregnancy, with less than 37 weeks defined as a preterm birth. Head circumference z-score is a relative measure of fetal growth comparing the measured head circumference at delivery to a reference distribution of head circumference measurements for newborns with identical gestational age at delivery. The primary goal of the PROTECT multi-exposure analysis is discovery-based high-dimensional mediation analysis, where the intermediary pathways of interest are quantified through cytochrome p450 and lipoxygenase metabolites. Therefore, the workflow of the proposed analysis is first to investigate the total effects corresponding to mixtures of exposures and the proceed to mediation analyses corresponding to slightly more targeted mediation hypotheses.

Before discussing the models, we will first discuss data cleaning and merging steps. For this analysis we specifically consider chemical exposures measured at visit 1 with less than 50% below the respective limits of detection (LOD). Values below the LOD are handled through LOD$/\sqrt{2}$

substitution. After removal of exposures with too many observations below the LOD, we then remove mono-hydroxyisobutyl phthalate (MHiBP), mono-3-hydroxybutyl phthalate (MHBP), and Bisphenol S, due to having substantially more missingness than the other phthalate, paraben, and phenol metabolites. The resulting exposure dataset contains 40 total exposure, of which 11 are phthalate metabolites, 8 are phenol and paraben metabolites, 7 are PAHs, and 14 are metals. Since all chemical exposures are measured in urine samples, we next perform specific gravity correction for all 40 chemical exposures prior to modeling (Kuiper et al., 2021). It is very important to adjust for specific gravity prior to modeling, rather than including it as an adjustment covariate in the mediation model, so that the latent factor model does not capture exposure correlation structure due to differences in urinary dilution across participants. Exposures are then log-transformed and standardized prior to model fitting. In PROTECT the cytochrome p450 and lipoxygenase metabolites are measured at visit 3 and we remove lipoxygenase metabolites 5(s),14(r)-Lipoxin B4, Leukotriene B3, Leukotriene C4 methyl ester, 9(s)-HETE, (±)19-HETE due to having substantially more missingness than the other cytochrome p450 and lipoxygenase metabolites. This leaves us with 32 total metabolites, 18 of which correspond to the cytochrome p450 pathway and 14 of which correspond to the lipoxygenase pathway. All cytochrome p450 and lipoxygenase metabolites are log-transformed and standardized prior to model fitting.

For the total effect mixture map we consider gestational age at delivery and head circumference z-score as the two outcomes of interest. All models are adjusted for maternal age and education, and the head circumference z-score models are additionally adjusted for pre-pregnancy body mass index. Subjects that had a missing value for any of the $q = 40$ exposures, $p = 32$ mediators, gestational age at delivery, head circumference z-score, and the adjustment covariates were removed from the total effect only analysis. The total sample size for the total effect only models is $n = 478$. Coupling the results of the total effect analysis with existing literature establishing phthalate exposure as a risk factor for preterm birth (Ferguson et al., 2014; Welch et al., 2022), the mediation mixture map models will be restricted to the phthalate exposure class to maximize available sample size. We remove subjects that had a missing value for any of the $q = 11$ phthalate metabolites and $p = 32$ mediators, but use the maximal available sample size for gestational age at delivery and head circumference z-score and their corresponding adjustment sets. Therefore, the phthalate only gestational age at delivery mediation mixture map models have a sample size of $n = 466$ participants and the phthalate only head circumference z-score mediation mixture map models have a sample size of $n = 241$ participants. The number of latent factors in the total effect mixture map and the mediation mixture map are chosen by optimal BIC.

In addition to the mediation mixture map, we also implement several other approaches to high-dimensional mediation analysis that are frequently used in practice. The first is based on HDMT, a method that fits individual mediation models, tests for significance of the indirect effect using

HDMT for each exposure-mediator pair, and applies a Benjamini-Hochberg correction step on the HDMT p-values to correct for multiple testing (Dai et al., 2022). The second competing method is a method for single exposure high-dimensional mediation analysis called HDMA (Gao et al., 2019), which applies sure independence screening (SIS) with respect to the $\boldsymbol{Y} \mid \boldsymbol{M}, \boldsymbol{A}, \boldsymbol{C}$ model and then fits a high-dimensional mediation model with a de-biased lasso penalty restricted to the mediators that were retained after screening. The third competing method is the naive two-step version of the mediation mixture map, which fits a linear latent factor model and subsequently plugs the estimated latent factors into a mediation model. For the naive two-step version of the mediation mixture model, we will also use optimal BIC to estimate the number of latent factors. It is very important to note that the point estimates reported from HDMT, HDMA, and the mediation mixture map are not directly comparable because the mediation mixture map models all cytochrome p450 and lipoxygenase metabolites simultaneously, HDMA fits the mediation model only on the subset of cytochrome p450 and lipoxygenase metabolites that pass the sure independence screening step, and HDMT considers a single mediator at a time. The purpose of including HDMA and HDMT is to see if established methods in the high-dimensional mediation analysis literature yield comparable conclusions to the conclusions of the mediation mixture map.

4.4.2 Total Effect Only Results

Figure 4.2 shows the estimated loadings matrices corresponding to the one-step mediation mixture map with final gestational age as the outcome, the one-step mediation mixture map with head circumference z-score as the outcome, and the naive two-step version of the mediation mixture map. The columns in the estimated loadings matrices are ordered based on the proportion of variance explained from left to right. The naive two-step estimator, identifies $r = 6$ latent factors, with 13.0% of the variability explained by the first latent factor, 9.7% of the variability explained by the second latent factor, 7.7% of the variability explained by the third latent factor, 5.6% of the variability explained by the fourth latent factor, 4.7% of the variability explained by the fifth latent factor, and 4.6% of the variability explained by the sixth latent factor. The first latent factor corresponds to phthalate exposure, especially monobutyl phthalate (MBP), monobenzyl phthalate (MBzP), mono(2-ethyl-5-carboxypentyl) phthalate (MECPP), mono(2-ethyl-5-hydroxyhexyl) phthalate (MEHHP), mono(2-ethylhexyl) phthalate (MEHP), mono(2-ethyl-5-oxohexyl) phthalate (MEOHP), and monoisobutyl phthalate (MiBP). The second and third latent factors correspond to PAH and metals exposure, respectively. The fourth latent factor also corresponds to phthalate exposure, namely mono(carboxynonyl) phthalate (MCNP), mono(carboxyoctyl) phthalate (MCOP), and mono(3-carboxypropyl) phthalate (MCPP). The fifth and sixth latent factors correspond to dichlorophenol and paraben exposure, respectively. The estimated loadings matrix for the medi-

ation mixture map with head circumference as the outcome is very similar to the naive two-step loadings estimate, with the only difference being that the one-step mediation mixture map estimates one less latent factor, omitting the latent factor corresponding to paraben exposure. For the one-step mediation mixture map estimator with gestational age at delivery as the outcome, breaks the phthalates into three latent factors, with one phthalate factor corresponding to MECPP, MEHHP, MEHP, and MEOHP exposure, the second phthalate latent factor corresponding to MBP, MBzP, and MiBP exposure, and the third phthalate latent factor corresponding to MCNP, MCOP, and MCPP exposure. As a result of breaking the phthalates down into three latent factors rather than two latent factors, the latent factor corresponding to PAH exposure ends up having the highest proportion of variability explained with 9.3% explained.

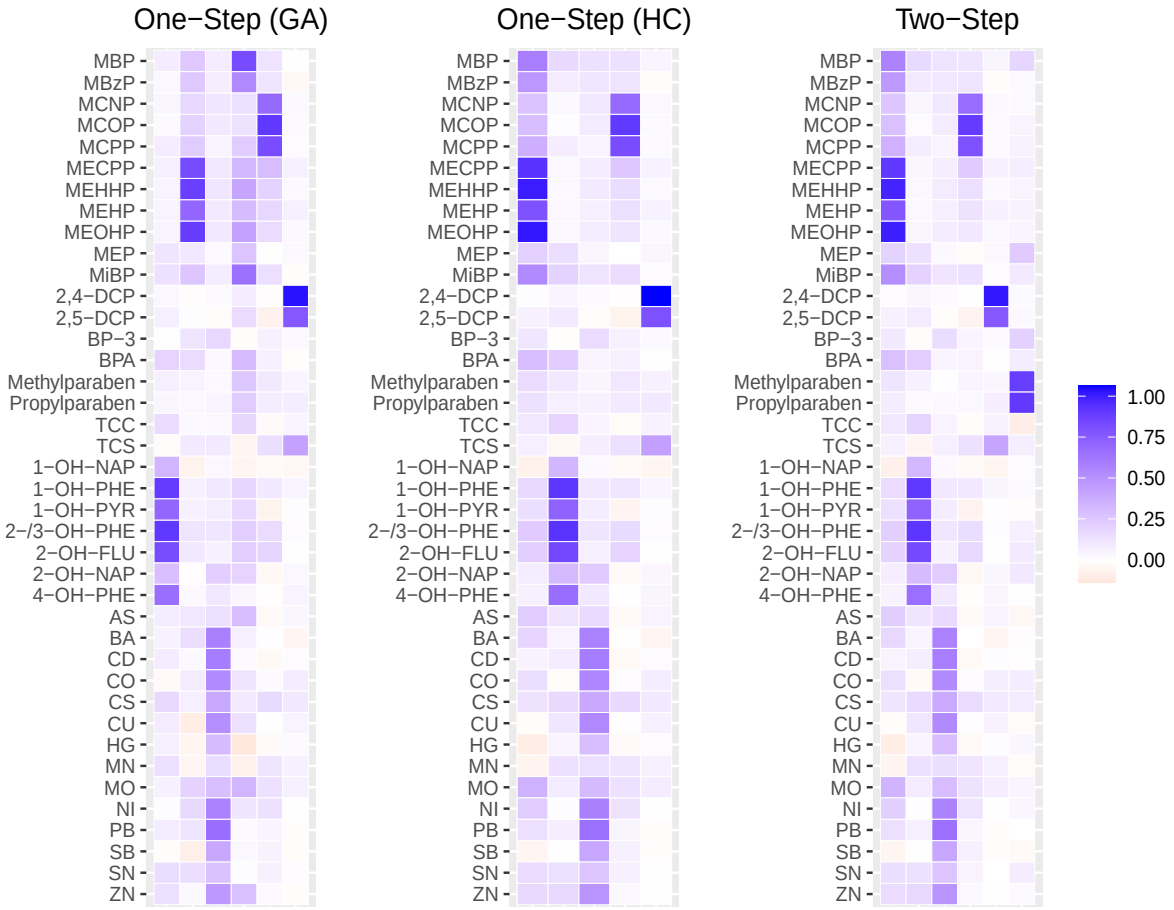


Figure 4.2: Estimated loadings matricies corresponding to the total effect mixture model for final gestational age (GA) and head circumference z-score (HC) ($n = 478$). Final gestational age models are adjusted for maternal age and educational attainment. Head circumference z-score models are adjusted for maternal age, educational attainment, and pre-pregnancy body mass index.

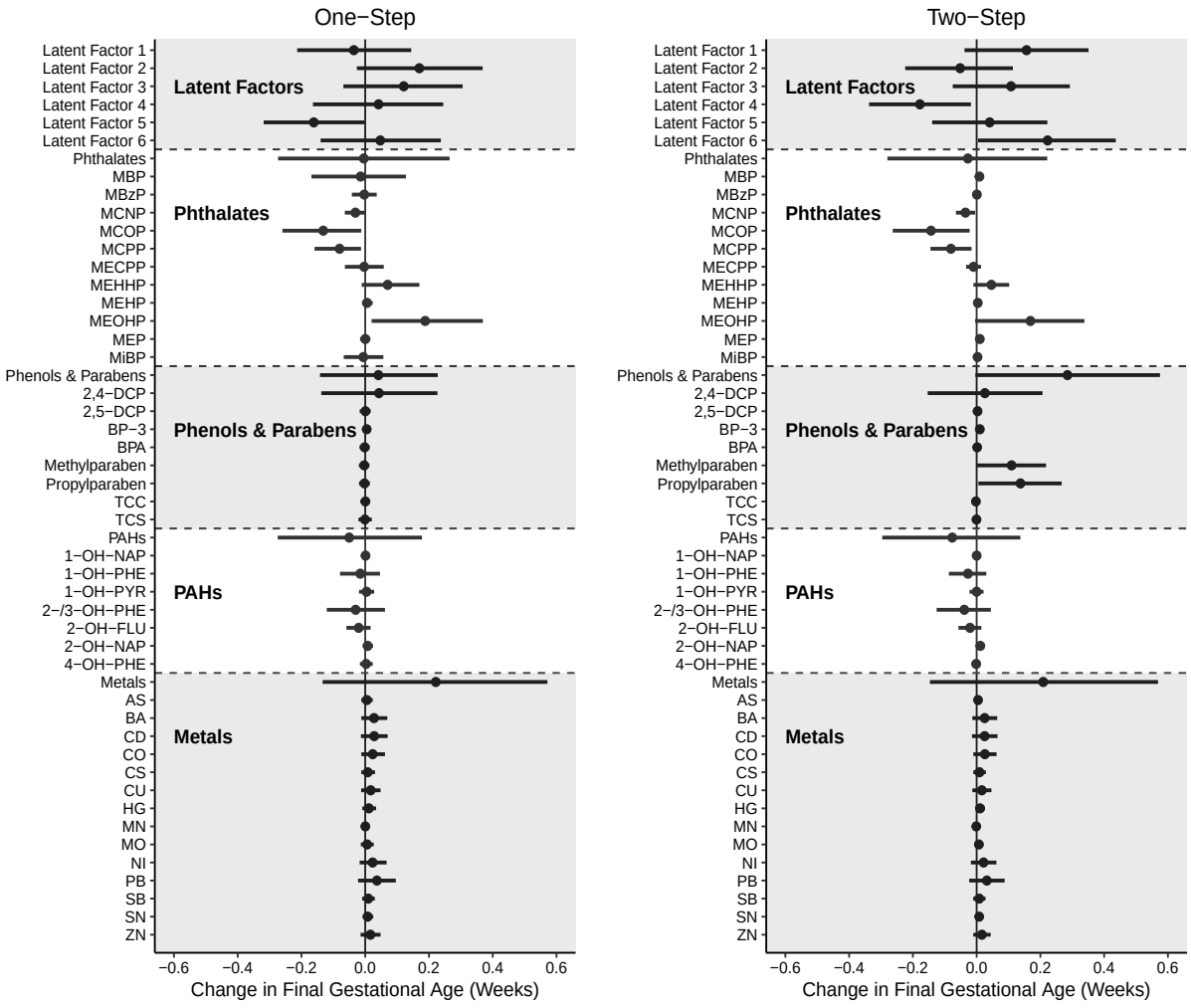


Figure 4.3: Estimated total effect corresponding to final gestational age from the total effect mediation mixture map corresponding to an interquartile range (IQR) change in exposure ($n = 478$). For individual chemical exposures, TE estimates correspond to an IQR change on the log-scale. For exposure classes, TE estimates correspond to simultaneous IQR changes on the log-scale for all exposures contained within the exposure class. For the latent factors, TE estimates correspond to IQR changes in the corresponding latent factor variable. For interpretations of the latent factors see Figure 4.2. Final gestational age models are adjusted for maternal age and educational attainment. Head circumference z-score models are adjusted for maternal age, educational attainment, and pre-pregnancy body mass index.

Figure 4.3 shows the TE estimates and corresponding 95% credible intervals for the one-step and two-step mediation mixture map estimators with gestational age at delivery as the outcome. Sifting through the latent factors we see that latent factor 5 for the one-step estimator and latent factor 4 for the two-step estimator, both corresponding to MCNP, MCOP, and MCPP exposure, have 95% credible intervals that do not cover zero. More specifically, an IQR change in the latent fac-

tors corresponding to MCNP, MCOP, and MCPP exposure is associated with a 1.13 day decrease in gestational age at delivery (95% CI: 0.01, 2.23) for the one-step estimator and is associated with a 1.25 day decrease in gestational age at delivery (95% CI: 0.13, 2.37) for the two-step estimator adjusted for maternal age and educational attainment. For the two-step mediation mixture map, we also observe that an IQR change in the latent factor corresponding to paraben exposure is associated with a 1.56 day increase in gestational age at delivery (95%: 0.03, 3.05) adjusted for maternal age and educational attainment. This association is likely only observed for the two-step estimator, because no latent factor corresponding to the one-step estimator captures exposure to methylparaben and propylparaben. Of the four exposure classes, the most intriguing is the phthalate exposure class, where effects corresponding to IQR changes in MCNP, MCOP, MCPP, and MECPP appear to cancel out when looking at the TE corresponding to an IQR change in all phthalates. Looking at MCNP, MCOP, and MCPP for the one-step estimator, one IQR increase in MCNP exposure is associated with 0.22 days shorter gestational age at delivery (95% CI: 0.00, 0.45), one IQR increase in MCOP exposure is associated with 0.92 days shorter gestational age at delivery (95% CI: 0.08, 1.82), and one IQR increase in MCPP exposure is associated with 0.56 days shorter gestational age at delivery (95% CI: 0.09, 1.11). We observe similar associations for the two-step estimator. Note that the estimated TEs corresponding to MCNP, MCOP, and MCPP exposure are smaller than the estimated TE corresponding to the latent factor that summarizes MCNP, MCOP, and MCPP exposure. This is in-line with a common environmental health hypothesis that contributions from several chemicals simultaneously results in a larger effect compared to the effects corresponding to individual chemicals. There is no evidence of a significant TE that is negatively associated with gestational age at delivery outside of the phthalate exposure class.

Figure 4.4 shows the TE estimates and corresponding 95% credible intervals for the one-step and two-step mediation mixture map estimators with head circumference z-score as the outcome. Both the one-step and two-step estimators identify that an IQR change in the latent factor corresponding to MCNP, MCOP, and MCPP exposure is negatively associated with head circumference z-score, Namely, one IQR increase in the latent factor corresponding to MCNP, MCOP, and MCPP is associated with 0.15 standard deviations smaller head circumference (95% CI: 0.03, 0.28) for the one-step estimator and is associated with 0.15 standard deviations smaller head circumference (95% CI: 0.02, 0.27) for the two-step estimator. As with the results for gestational age at delivery, the effect sizes for the TEs corresponding to one IQR change in MCNP, MCOP, and MCPP are smaller than the effect size corresponding to the latent factor that summarizes exposure to all three phthalate metabolites. However, one difference between the results for head circumference z-score and the results for final gestational age is that the phthalates have negative or approximately zero TEs across the entire exposure class. Consequently, an IQR change in all 11 phthalate metabolites simultaneously, as estimated by the one-step method, is associated with 0.24 standard deviations

smaller head circumference (95% CI: 0.05, 0.44). Therefore, the TE corresponding to phthalate exposure and head circumference adheres to the principle of cumulative burden, where elevated to exposure to increasingly many phthalate metabolites results in increasingly large effects on head circumference.

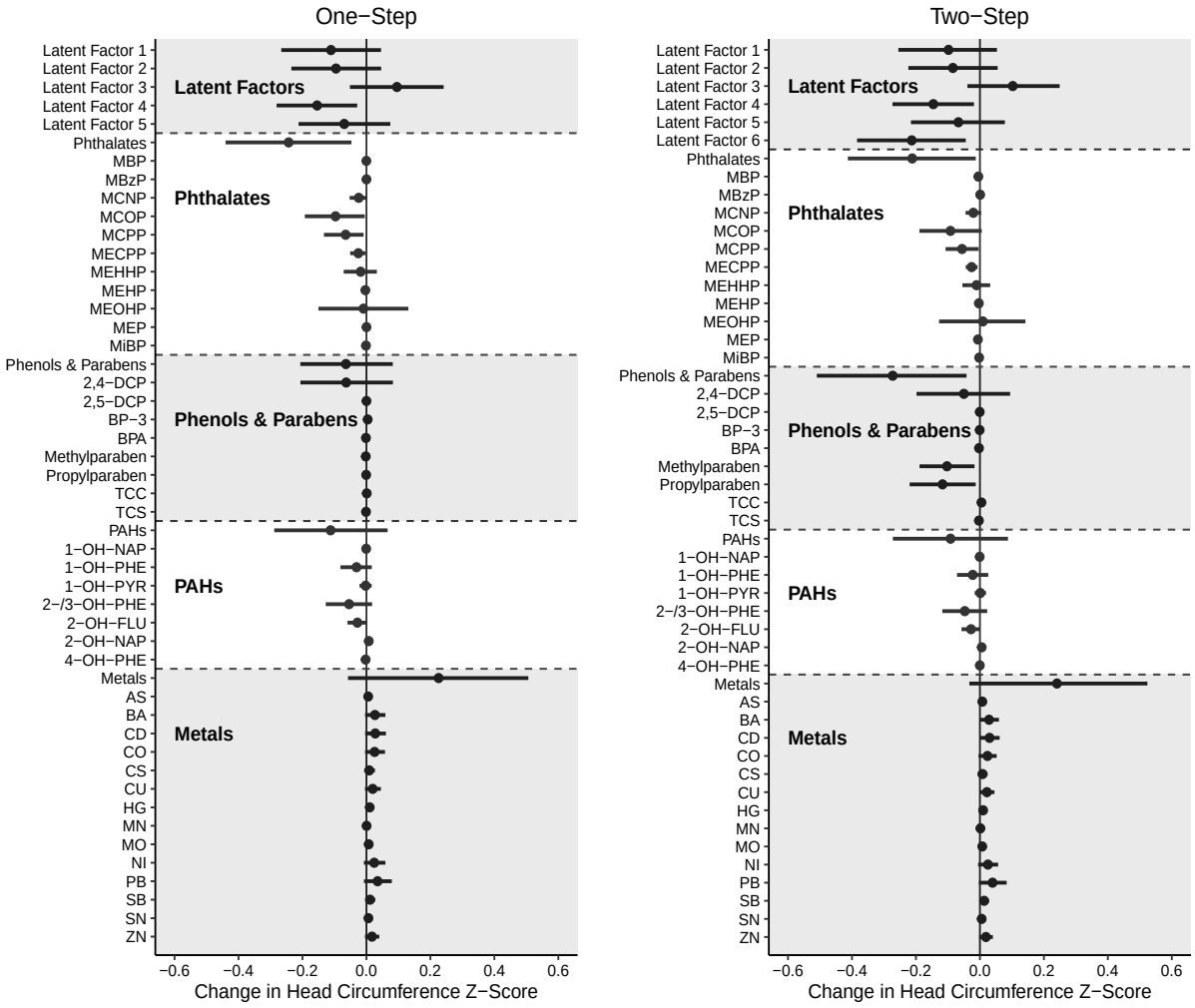


Figure 4.4: Estimated total effect corresponding to head circumference z-score from the total effect mediation mixture map corresponding to an interquartile range (IQR) change in exposure ($n = 478$). For individual chemical exposures, TE estimates correspond to an IQR change on the log-scale. For exposure classes, TE estimates correspond to simultaneous IQR changes on the log-scale for all exposures contained within the exposure class. For the latent factors, TE estimates correspond to IQR changes in the corresponding latent factor variable. For interpretations of the latent factors see Figure 4.2. Final gestational age models are adjusted for maternal age and educational attainment. Head circumference z-score models are adjusted for maternal age, educational attainment, and pre-pregnancy body mass index.

4.4.3 Mediation Analysis Results

Given that the TE analysis results pinpoint phthalate exposure as being particularly associated with gestational age at delivery and head circumference z-score, and given that the literature on gestational age at delivery has identified phthalate exposure as a risk factor for preterm births (Ferguson et al., 2014; Welch et al., 2022), we restrict our focus to the phthalate exposure class with gestational age at delivery as the outcome to maximize the available sample size for the mediation mixture map. Table 4.5 displays the phthalate-eicosanoid pairs from HDMT-based high-dimensional mediation analysis, that had negative DEs, IEs, and TEs and a q-value of less than 0.1. The most noteworthy phthtalate metabolites were MCNP and MCPP, which had 12 and 10 eicosanoids, respectively, for which there was a significant indirect effects after multiple testing correction. For MCNP, half of the eicosanoid metabolites correspond to the cytochrome p450 pathway, with percent mediated ranging between 40%-95% and the other half of the eicosanoid metabolites correspond to the lipoxygenase pathway, with percent mediated ranging between 25%-75%. Of the 10 eicosanoid metabolites with significant indirect effects for MCPP, 9 also had a significant indirect effect for MCNP, but with percent mediated ranging between 30%-70% for the cytochrome p450 pathway and 35-50% for the lipoxygenase pathway. In addition to MCNP and MCPP, there were four eicosanoid metabolites with significant indirect effects corresponding to MECPP and two eicosanoid metabolites with significant indirect effects corresponding to MBzP.

Phthalate Metabolite	**TE**	**Eicosanoid Pathway**	**Eicosanoid Metabolites**	**% Mediated**
MCNP	-0.11	Cytochrome p450	18-HETE, 14(15)-EET, 17-HETE	40-95%
			5,6-DHET, 16-HETE, 9S-HODE	
		Lipoxygenase	12-oxoETE, 13-oxoODE, 15-oxoETE	25-75%
			5-oxoETE, LTB4, 13S-HODE, LTE4	
MCPP	-0.15	Cytochrome p450	18-HETE, 14(15)-EET, 17-HETE	30-70%
			5,6-DHET, CAA, 16-HETE	
		Lipoxygenase	12-oxoETE, 13-oxoODE, 15-oxoETE, LTB4	35-50%
MECPP	-0.06	Cytochrome p450	14(15)-EET, 17-HETE, 5,6-DHET, 16-HETE	55-95%
		Lipoxygenase	LTB4	76%
MBzP	-0.18	Cytochrome p450	16-HETE	18%
		Lipoxygenase	13-oxoODE	23%

Table 4.5: Results for HDMT-based single exposure single mediator analysis with final gestational age as the outcome ($n = 466$). Estimated total effects in the table correspond to changes in gestational age at delivery (weeks) for an IQR increase in exposure. Phthalate metabolite and eicosanoid metabolite pairs listed in the Table are results where there was significant indirect effect, ascertained by having a q-value less than 0.1, a negative total effect, and a negative indirect effect.

Table 4.6 shows the results for HDMA-based high dimensional mediation analysis corresponding to phthalate metabolites that have negative estimated DEs, IEs, and TEs. We observe that the SIS-step in HDMA screens the number of potential mediators down from $q = 32$ to three eicosanoid metabolites. Of the metabolites identified by the HDMA-based high dimensional mediation analysis, MCPP and MBzP, were also identified by HDMT-based high dimensional mediation analysis. HDMA estimates that one IQR change in MBzP exposure is associated with 1.26 days shorter gestational age at delivery adjusted for maternal age and education, however none of the three eicosanoids selected after SIS have significant product terms contributing to the IE. Conversely, eicosanoid metabolites retained after SIS corresponding to MCPP, CAA, 9(10)-EpoME, and 15-oxoETE, have significant product terms contributing to the IE. More specifically, one IQR change in MCPP exposure is associated with 1.05 days shorter gestational age at delivery adjusted for maternal age and education, with an estimated 97% of the total effect mediated through CAA, 9(10)-EpoME, and 15-oxoETE.

Exposure	**# Post-SIS Mediators**	**IE**	**TE**	**% IE**	**Significant Eicosanoids**
MBP	3	-0.03	-0.14	21	None
MBzP	3	-0.07	-0.18	38	None
MCPP	3	-0.14	-0.15	97	CAA, 9(10)-EpoME, 15-oxoETE
MiBP	3	-0.06	-0.16	36	None

Table 4.6: Results for the HDMA-based multi-mediator, single exposure analysis with final gestational age as the outcome ($n = 466$). Estimated IEs and TEs in the table correspond to changes in gestational age at delivery (weeks) for an IQR increase in exposure. The results in the table are restricted to the exposures which result in a negative total effect (TE), a negative direct effect (DE), and a negative indirect effect (IE) for gestational age at delivery. The significant eicosanoids column corresponds to metabolites with significant product terms contributing the IE.

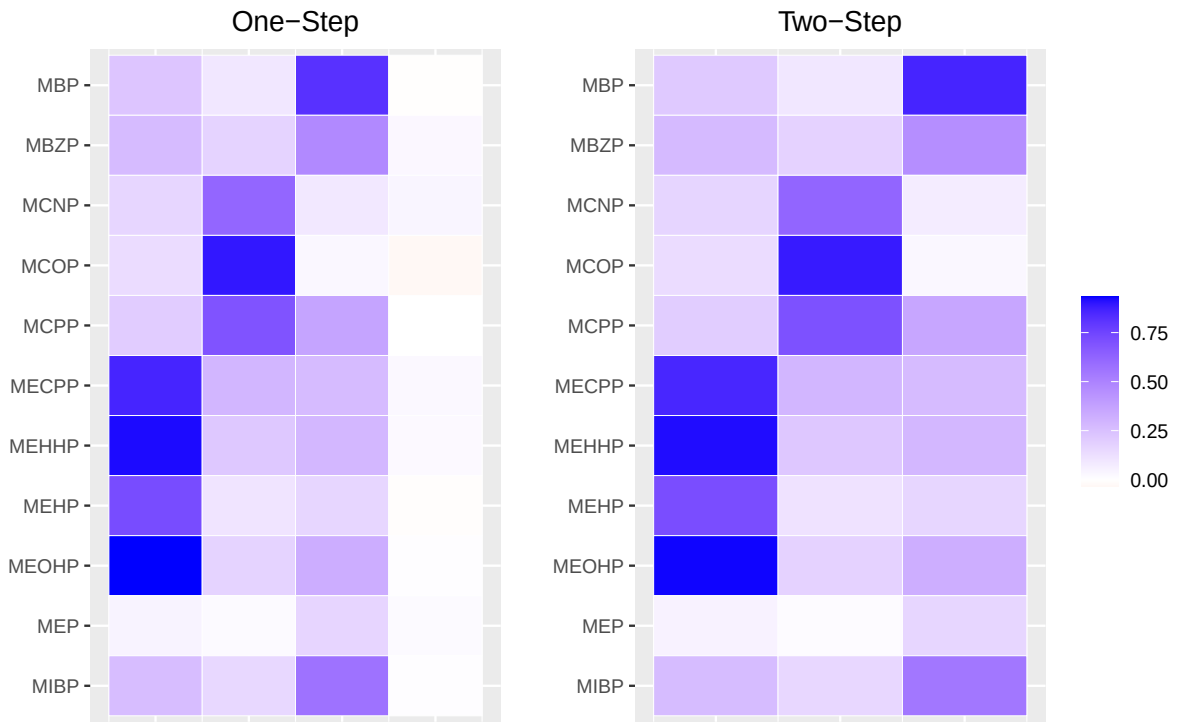


Figure 4.5: Estimated loadings matricies corresponding to the mediation mixture model with independent shrinkage for gestational age at delivery ($n = 466$). Gestational age at delivery models are adjusted for maternal age and educational attainment.

Figure 4.5 shows the estimated loadings matrices corresponding to the one-step and two-step mediation mixture maps with independent shrinkage and gestational age at delivery as the outcome. When restricting to phthalate metabolites only, we see that the phthalate metabolites are divided into three latent factors in a similar manner to the total effect only models. Namely, for the one-step estimator, the first latent factor corresponding to MECPP, MEHHP, MEHP, and MEOHP explains 29.7% of the variability, the second latent factor corresponding to MCNP, MCOP, and MCPP exposure explains 17.3% of the variability, and the third latent factor corresponding to MBP, MBzP, and MiBP explains 15.3% of the variability. Note that this is much higher than the cumulative proportion of variability explained, than the total effect analysis that considers all exposure classes. Table 4.7 shows the mediation effects attributable to exposure mixtures as estimated by the mediation mixture map with independent shrinkage and gestational age at delivery as the outcome. For the one-step mediation mixture map, we observe that a simultaneous IQR change in all phthalates corresponds to 1.54 days shorter gestational age at delivery on average (95% CI: -0.84, 3.92). In contrast, the two-step mediation mixture map estimates 0.42 days shorter gestational age at delivery on average (95% CI: -1.82, 2.73) for an IQR change in all phthalate metabolites.

Looking at the IEs, the one-step estimator estimates a percent mediated of 88% corresponding to the latent factor summarizing exposure to MCNP, MCOP, and MCPP, although an IQR change in the latent factor summarizing exposure to MCNP, MCOP, and MCPP has a modest TE of 0.14 days. The two-step estimator identifies a significant TE corresponding to an IQR change in the latent factor summarizing exposure to MBP, MBzP, and MiBP, but the estimated IE in this case is 0. However, the two-step estimator does not estimate a negative TE corresponding to latent factor summarizing MCNP, MCOP, and MCPP exposure. Given these results, it appears that including information on the mediation model to inform estimation of latent factors helps estimate the directionality of the TE estimate and a percent mediated that is more consistent with the HDMT-based and HDMA-based high-dimensional mediation analyses.

Exposure	**IE**	**IE 95% CI**	**TE**	**TE 95% CI**	**% Mediated**	**Largest Weights**
One-Step						
Latent Factor 1	-0.01	(-0.04, 0.03)	0.06	(-0.18, 0.29)	-	MECPP, MEHHP, MEHP, MEOHP
Latent Factor 2	-0.02	(-0.18, 0.12)	-0.02	(-0.24, 0.20)	88%	MCNP, MCOP, MCPP
Latent Factor 3	0.01	(-0.04, 0.06)	-0.27	(-0.52, 0.00)	-	MBP, MBzP, MiBP
All Phthalates	-0.02	(-0.19, 0.13)	-0.22	(-0.56, 0.12)	8%	
Two-Step						
Latent Factor 1	0.00	(-0.01, 0.01)	0.05	(-0.17, 0.28)	-	MECPP, MEHHP, MEHP, MEOHP
Latent Factor 2	-0.04	(-0.14, 0.02)	0.11	(-0.13, 0.34)	-	MCNP, MCOP, MCPP
Latent Factor 3	0.00	(-0.01, 0.02)	-0.23	(-0.44, -0.03)	0%	MBP, MBzP, MiBP
All Phthalates	-0.04	(-0.14, 0.02)	-0.06	(-0.39, 0.26)	56%	

Table 4.7: One-step and two-step mediation mixture map results with independent shrinkage corresponding to gestational age at delivery ($n = 466$). Estimated IEs and TEs in the table correspond to changes in gestational age at delivery (weeks) for an IQR increase in exposure.

Given the differing mediation results between HDMT, HDMA, and the mediation mixture map, we further evaluate the posterior distribution of $\boldsymbol{\alpha}_a$ and $\boldsymbol{\beta}_m$ to better understand why the mediation mixture map does not identify significant indirect effects. Supplementary Table C.15 shows the posterior mean and 95% credible intervals corresponding to entries in $\boldsymbol{\alpha}_a$ and Supplementary Table C.16 shows the posterior mean and 95% credible intervals corresponding to entries in $\boldsymbol{\beta}_m$. We observe that the latent factor summarizing MCNP, MCOPP, and MCPP exposure is significantly negatively associated with 23 of the 32 eicosanoids and significantly positively associated with 4 of the 32 eicosanoids. However, all entries in $\boldsymbol{\beta}_m$ have 95% posterior credible intervals that cover zero, indicating that the lack of significance of the indirect effect is specifically due to the lack of significance corresponding to $\boldsymbol{\beta}_m$. To directly compare with HDMA, we fit the two-step mediation mixture map with independent shrinkage priors, restricted to only MCPP and the three eicosanoids

identified by HDMA. The estimated total effect for this analysis is -0.14 (95% CI: -0.31, 0.05), the estimated indirect effect is -0.10 (95% CI: -0.16, -0.04), and the estimated percent mediated is 72%. Therefore, the mediation mixture map restricted to a single exposure coincides with the HDMA results, indicating that there is a fundamental difference between mediation results from a multi-exposure mediation analysis and mediation results from a method that screens mediators based on the $\boldsymbol{Y} \mid \boldsymbol{M}, \boldsymbol{L}, \boldsymbol{C}$ model and subsequently models exposures independently of one another.

Supplementary Figure C.2 and Supplementary Table C.17 shows the estimated loadings matrices and mixture mediation results corresponding to product shrinkage with final gestational age as the outcome. Both the one-step and two-step estimators identify significant TEs corresponding to a latent factor that summarizes exposure to MBP, MBzP, and MiBP. The one-step estimator also identifies an estimated 57% mediated corresponding to an IQR change in the latent factor summarizing MCNP, MCOP, and MCPP exposure, with a TE is equal to -0.07 days. As with the simulations, we generally observe the phenomenon of product shrinkage regularizing parameters that comprise the IE more aggressively than the independent shrinkage prior.

4.4.4 Conclusions

Across many of the results presented in the PROTECT multi-exposure mediation analysis, higher exposure to the MCNP, MCOP, and MCPP grouping of phthalate metabolites is consistently and significantly associated with shorter gestational age at delivery and smaller head circumference z-score. Moreover, with respect to the TE, we generally observe increasingly adverse TE estimates as simultaneous exposure to increasingly many phthalates occurs. This provides evidence for conceptual models of increased cumulative risk as an individual is exposed to increasingly many phthalate metabolites.

4.5 Discussion

In this chapter, we discussed a framework for mediation analysis with exposure mixtures through latent factors which summarize the correlation structure of the individual chemical exposures. In this work we are interpreting the latent factors as common sources of exposure variation, given that highly correlated exposures are both more likely to co-occur and tend to load onto the same loading after the varimax rotation is applied to identify the loadings matrix. The principal methodological innovation is building a mediation framework that can simultaneously estimate mediation effects corresponding to changes in the individual exposure space and the latent exposure mixtures space. We showed that modeling the mediation model and latent factor model jointly, as in the one-step mediation mixture map, leads to larger posterior standard deviations than a naive two-step imple-

mentation of the mediation mixture map, as a result of accounting for uncertainty in estimating the latent factors. Moreover, the one-step mediation mixture map is less biased in the dense mediation setting, indicating that jointly modeling the mediation model and latent factor model may be particularly important when the indirect effect is distributed across many intermediaries. In the PROTECT multi-exposure mediation analysis, we show that the mediation mixture map can estimate a low-dimensional representation of the exposure space and helps identify important mixtures contributing to adverse birth outcomes, such as gestational age at delivery and head circumference.

The results of the PROTECT data analysis show that conclusions can be quite different between mediation models that jointly model exposures and mediation models that separately model exposures. If the total effect with respect to a single exposure is subject to confounding due to co-exposure then total effect estimation will be biased and, consequently, estimates of the direct and indirect effects will also be biased. Although we cannot directly assess co-exposure using the chemical exposure data, we do see that pairwise correlations between the phthalate metabolites are moderate to high within exposure class (see Supplementary Figure C.1). Therefore, it is likely that the multi-exposure mediation analysis is more accurately estimating the direct, indirect, and total effects, and the large percent mediated estimates from the single exposure mediation analyses should be interpreted with skepticism. In PROTECT, the mediation mixture map does appear to quantify latent sources of exposure variation that operate on different metabolic pathways. For instance, when evaluating the the posterior distribution of $\boldsymbol{\alpha}_a$ we observe that the latent factor corresponding to MCNP, MCOP, and MCPP is strongly associated with eicoanoids in the cytochrome p450 and lipoxygenase pathways, while the other latent factors are not. This indicates that the mediation mixture map can detect relationships between specific types of phthalate metabolites and inflammatory pathways, even if those pathways do not ultimately end up being significantly associated with the outcome in the $\boldsymbol{Y} \mid \boldsymbol{M}, \boldsymbol{L}, \boldsymbol{C}$ model.

In this chapter, we introduced a relatively simple prior specification that elucidates targeted shrinkage of the product. In both the simulations and the data example, product shrinkage tended to apply extremely aggressive shrinkage towards zero, making it very difficult to correctly estimate non-zero indirect effects. This has also been observed in other work on shrinkage priors for high-dimensional mediation models (Wang et al., 2019). As previously noted, the primary reason for overregularization is due to the fact that sparsity in the product occurs when one or the other terms in the product are zero. This is mathematically encoded in the product shrinkage prior by either shrinking $(\boldsymbol{\beta}_m)_{gj}$ to zero or by shrinking $(\boldsymbol{\alpha}_a^k)_{gj}$ to zero when $|(\boldsymbol{\beta}_m)_{gj}|$ is large. This formulation is in direct contrast to regularization techniques discussed in the high-dimensional confounder adjustment literature, such as the outcome-adaptive lasso (OAL) and Bayesian adjustment for confounding (BAC) (Wang et al., 2012; Shortreed and Ertefaie, 2017), which when translated to our setting advocate for $(\boldsymbol{\alpha}_a^k)_{gj}$ to be less aggressively shrunk when $|(\boldsymbol{\beta}_m)_{gj}|$ is large. Therefore, a

promising future direction for shrinkage in high-dimensional mediation models is to modify our product shrinkage prior framework by setting $f\{(\boldsymbol{\beta}_m)_{gj}\} = (\boldsymbol{\beta}_m)_{gj}$, which corresponds to targeted shrinkage on the ratio $(\boldsymbol{\alpha}_a^k)_{gj}/(\boldsymbol{\beta}_m)_{gj}$. This prior specification is consistent with ideas from OAL and BAC in that all three shrink $(\boldsymbol{\alpha}_a^k)_{gj}$ less aggressively when $|(\boldsymbol{\beta}_m)_{gj}|$ is large. Therefore, we recommend against using targeted shrinkage of the product if estimation of a non-zero global IE is of interest.

Although we handled the dimensionality of the mediator space using shrinkage priors, there are other approaches for modeling the mediators. One idea is to model the mediators using dimension reduction techniques (Crainiceanu et al., 2011; Chén et al., 2018). If a low-dimensional representation of the mediator space summarizes all information contained in $\boldsymbol{M}$, then the mediation effect estimates corresponding to a linear mediation model are the same regardless of whether the observed mediators or transformed mediators are used. Moreover, if the information in the observed mediator space can be well-summarized with a handful of latent factors, then shrinkage priors are not necessary, allowing us to circumvent issues related to overregularization. One straightforward modification of the mediation mixture map is to perform dimension reduction with respect to the mediator space using a latent factor model (Bhattacharya and Dunson, 2011). However, in practice, the low-dimensional representation of the mediator space characterized by a latent factor model may not sufficiently summarize the information contained in $\boldsymbol{M}$ and can therefore can be thought of as an approximation method; the quality of the approximation depends on how much of the variance in $\boldsymbol{M}$ the latent factors explain. Further work is needed to determine whether or not dimension reduction is a more effective tool for modeling high-dimensional mediators in the context of the mediation mixture map.

While this chapter introduces a more model-driven correlative framework for mediation analysis with exposure mixtures, there are many limitations and challenges with extending the mediation mixture map to a comprehensive causal mediation framework. In this associative work we estimated exposure mixtures based on the correlation structure of individual chemical exposures, however measurements of chemical exposures from biological samples do not contain sufficient information to infer intervenable exogenous exposure sources. There are also conceptual questions about trying to use outcome and mediator data to infer latent exposure sources, as, in theory, the latent exogenous exposure sources should not depend on the outcome and mediators. Thus, the interpretation of the estimated latent factors from MedMix is limited to statements of association rather than causality. That being said, the idea underlying our associative mediation framework can in principle be generalizable to more complicated settings. That is, we can consider a generalized

version of the mediation mixture map

$$
\begin{aligned}
[Y_i, \boldsymbol{M}_{i\cdot} \mid \boldsymbol{L}_{i\cdot}, \boldsymbol{C}_{i\cdot}] &\sim \mathcal{M}_{\mathcal{P}_M}(y, \boldsymbol{m} \mid \boldsymbol{l}, \boldsymbol{c}) \\
[\boldsymbol{A}_{i\cdot} \mid \boldsymbol{L}_{i\cdot}] &\sim \mathcal{L}_{\mathcal{P}_L}(\boldsymbol{a} \mid \boldsymbol{l}),
\end{aligned} \tag{4.5.1}
$$

for $i = 1, \ldots, n$ where $\mathcal{M}_{\mathcal{P}_M}(y, \boldsymbol{m} \mid \boldsymbol{l}, \boldsymbol{c})$ is a mediation model that depends on $\boldsymbol{L}_{i\cdot}$, $\boldsymbol{C}_{i\cdot}$, and parameters $\mathcal{P}_M$ and $\mathcal{L}_{\mathcal{P}_L}(\boldsymbol{a} \mid \boldsymbol{l})$ is a latent factor model that depends on $\boldsymbol{L}_{i\cdot}$ and parameters $\mathcal{P}_L$. While model (4.5.1) is much more flexible it comes with the tradeoff of potentially being more difficult to define and interpret mediation effects with respect to latent exposure mixtures. We leave this topic for future work.

CHAPTER 5

Discussion

5.1 Handling Collinearity Among Chemical Exposures

In Chapter 2, we proposed the GIGG prior, as a tool for addressing variance inflation in linear regression models. The GIGG prior construction assigns multivariate shrinkage to regression coefficients within the same exposure class. The prior correlation structure is responsible for achieving additional variance reduction beyond what is achievable using independent shrinkage priors. Moreover, the GIGG prior has two hyperparameters per exposure group, leading to customizable shrinkage for groups with dense or sparse signals. Consequently, GIGG regression results in a relative IMSE reduction of 32.4% compared to horseshoe regression, a gold standard independent shrinkage prior. Moreover, the computational burden of sampling from the posterior distribution when the regression coefficients are endowed with the GIGG prior is comparable to the computational burden of sampling from the posterior distribution when the regression coefficients are endowed with a horseshoe prior.

Extending GIGG regression to a generalized additive modeling framework is relatively straightforward, as there are several papers with general strategies for extending the family of global-local shrinkage priors to additive models (Xu et al., 2016; Wei et al., 2020). Namely, Xu et al. (2016) and Wei et al. (2020) both use basis expansion approaches to express nonlinear functions of each exposure as linear combinations of basis functions, and then assigns the same local shrinkage parameter to basis coefficients corresponding to the same chemical exposure. However, extensions of GIGG regression become less straightforward when considering models with increasingly flexible mean structures, such as models involving high-order interaction terms. For example, in the context of pairwise interaction models, strong heredity assumptions are generally imposed on shrinkage estimators to ensure that pairwise interactions are only considered if the corresponding main effects are present in the fitted model (Boss et al., 2021). The purpose of strong heredity assumptions is to facilitate interpretability of the interaction terms, but they often result in increasingly complex penalty functions and shrinkage priors. A second, and much larger challenge, is that more flexible

mean structures are exponentially more difficult to reliably estimate when there is collinearity in the exposure space. Consider a linear multi-pollutant model with two exposures E_1 and E_2 such that the Pearson correlation between E_1 and E_2 is equal to 0.9. If the true generative model only involves first-order terms then we only need to distinguish between E_1 and E_2, but if the true generative model involves second-order terms then we need to additionally distinguish between E_1^2, E_2^2, and E_1E_2. That is, as the true generative model becomes increasingly higher order, the ability to estimate the mean structure in the presence of correlated chemical exposure becomes more difficult. While there are methodological developments for modeling exposure mixtures with increasingly flexible mean structures (Bobb et al., 2015; Ferrari and Dunson, 2020; McGee et al., 2023), explicit commentary on how to fit models these models to high-dimensional, highly correlated chemical exposure data remains a seminal methodological challenge.

5.2 Leveraging External Information on the Total Effect

In Chapter 3, we argued that for many mediation analyses in environmental health, external summary-level information on the total effect is available from prior studies of the total effect. Moreover, we showed that leveraging external summary-level information on the total effect can improve estimation efficiency for IE, DE, and TE estimation through a constraint on the parameter space, provided that the outcome-mediator association given exposure is non-zero. Given that environmental health studies with protocols for collecting omics data generally have limited sample size, it is often useful to leverage externally available information when the true TEs for the internal and external study populations are the same. We developed the MESSI framework to protect against biased estimation of mediation effects when the internal and external study populations have different TEs, while still providing efficiency gains for estimating mediation parameters when the TEs in the internal and external study populations are the same.

In principle, the logical progression of establishing the biological and statistical significance of the total effect prior to investigating mediation hypotheses means that a substantially more general framework for estimating mediation effects with external summary-level information on the total effect can, in theory, be developed. However, developing a general framework that considers a more general class of mediation models, such as mediation models with exposure-mediator interaction terms or mediation models with non-continuous outcome and mediator data, are more challenging to work with. The main difficulty is that non-linear mediation models and non-linear data types coincide with total effect expressions that are functions of confounder and counterfactual exposure levels (VanderWeele, 2015). In practice, marginal effect estimates of exposures on the outcome are generally not reported as a functions of confounder and counterfactual exposure levels, and therefore are not necessarily interpretable as external total effect estimates. Moreover,

marginal effect estimates of the exposure on the outcome may not be logically consistent with the specification of the internal mediation model. For example, in the case of a mediation model with a binary outcome, a logistic regression model for $[\boldsymbol{Y} \mid \boldsymbol{M}, \boldsymbol{A}, \boldsymbol{C}]$ does not necessarily coincide with a logistic regression model for $[\boldsymbol{Y} \mid \boldsymbol{A}, \boldsymbol{C}]$ after integrating out $\boldsymbol{M}$ (Cheng et al., 2019). Therefore, it is important to first establish how information from a marginal outcome-exposure association translates to information on the total effect. Another future direction is to generate predictions from the external marginal model and use those predictions to inform model fitting in an internal mediation analysis (Gu et al., 2019). This approach may circumvent the need to establish explicit relationships between the external marginal outcome-exposure association, the external total effect, and the internal total effect, and consequently may be more straightforward for mediation analyses with complex data structures and limited information on the external models.

Extending the MESSI framework to multi-exposure mediation models also poses practical limitations on external summary-level information that can be used, even in the case of a linear mediation model, continuous outcome data, and continuous mediator data. One difficulty is that studies rarely have perfect overlap in the chemical exposures they consider. However, even if multiple studies consider the same sets of chemical exposures, different laboratory protocols, equipment, and environments may differentially impact the quality of exposure assessment for the same set of chemical exposures. Statistically, this manifests as differential measurement error mechanisms, and greatly impacts the feasibility of modeling chemical exposures with highly variable measurements or low detection frequencies. Therefore, a pragmatic framework for incorporating summary-level information from an external mixtures analysis must allow the set of chemical exposures to differ between the internal and external studies. The problem can be formalized as leveraging external summary-level information on the total effect model, $[\boldsymbol{Y} \mid \boldsymbol{A}_E, \boldsymbol{C}_E]$, where $\boldsymbol{A}_E$ are the set of chemical exposures in the external model and $\boldsymbol{C}_E$ are the set of confounders in the external model, to improve estimation of mediation parameters in the internal mediation model $[\boldsymbol{Y}, \boldsymbol{M} \mid \boldsymbol{A}_I, \boldsymbol{C}_I]$, where $\boldsymbol{A}_I$ are the set of chemical exposures in the internal model and $\boldsymbol{C}_I$ are the set of confounders in the internal model.

5.3 Causal Inference with Chemical Mixtures

In Chapter 4, we developed the mediation mixture map, which combines ideas from latent factor modeling and mediation analysis. The latent factor model component summarizes common sources of variation amongst the measured exposures, with the hope that those common sources of variation comprise relevant exposure mixtures. However, there is no guarantee that the estimated latent factors correspond to true exposure events, as they could characterize other common sources of variation such as spatiotemporal variation or pharmacokinetics. Moreover, outcome and media-

tor adaptive learning of latent sources of exposure variation does not admit a causal interpretation, as exposure sources should, in theory, not depend on the health outcome or intermediary pathways of interest. In the PROTECT example, we saw that the mediation mixture map with a varimax rotation post-processing step results in estimated latent factors corresponding to blocks of highly correlated chemical exposures. The total effects corresponding to these data adaptively estimated latent factors tend to be stronger than the total effects for the individual components that comprise the latent factors, giving us some assurance that the latent factors are likely summarizing some notion of cumulative exposure burden across multiple chemicals.

5.3.1 Causal Framework for Estimating the Effect of Exposure Mixtures

Building a causal framework for estimating the effect of exposure mixtures on a health outcome requires a significant amount of attention, given the complexity of the problem. Existing definitions of exposure mixtures frequently define an exposure mixture as a multivariate vector of individual chemical exposures (Devick et al., 2022), however it is crucial to distinguish how the individual chemical exposure concentrations are measured. If the individual chemical exposure concentrations are obtained from biological samples, then the multivariate vector of exposure concentrations corresponds to an endogenous chemical mixture that is circulating within specific biological systems. From an exposure source perspective, there is also the notion of exogenous chemical mixtures that circulate in the environment and enter the body through ingestion, respiration, or physical contact. From a causal intervention perspective, an exogenous chemical mixture is a quantity that regulatory policies can directly target, either by imposing caps on emissions of particular chemicals or engaging in remediation activities to clean up contaminated sites. However, causal interventions with respect to endogenous chemical mixtures occur at the individual level, generally through treatments for acute exposure events, such as chelation therapy for lead poisoning. That is, interventions on exogenous exposure mixtures try to limit chemical exposures *before* they enter the body and interventions on endogenous chemical mixtures try to reduce chemical exposure concentrations *after* they enter the body.

When designing a causal framework for estimating the effect of exposure mixtures, we need to be mindful of what the intervention is. If the intervention of interest is regulatory in nature, then either exogenous exposure data should be directly collected or endogenous exposure data should be leveraged to infer the exogenous exposure data. The later is itself a very challenging problem, as endogenous exposure data corresponds to the accumulation of many different exposure events that happen over time, and therefore may not clearly map to specific exogenous exposure events. In the development of the mediation mixture map, we used a latent factor model to attempt to define outcome-adaptive exposure mixtures, but the gold standard for estimating mediation effects

corresponding to exogenous exposure mixtures will likely need to involve exposure assessment at the exposure source rather than exposures derived from biological samples. This motivates the use of study designs which obtain samples of the lived environment, such as drinking and bathing water, measures of indoor and outdoor air quality, regularly consumed food products, and soil near the residence, to better understand chemical profiles of common exposure sources. Moreover, study designs that integrate information on common exposure sources with endogenous exposure measurements from biological samples may be helpful in understanding how exogenous exposure mixtures map to endogenous exposure mixtures.

Beyond the need to distinguish between exogenous exposure mixtures and endogenous exposure mixtures, there is also a discussion required regarding causal interventions targeting single chemical exposures. Regulators can certainly enact policies targeted at curbing exposure to specific chemicals and it is also biologically plausible that the causal effect of an endogenous exposure mixture is driven by one particular chemical. Therefore, a causal framework for exposure mixtures not only needs to distinguish between exogenous and endogenous exposure mixtures, but it also needs to be able to estimate causal effects corresponding to individual chemicals as well. The mediation mixture map is an attempt at defining a mediation framework when there is only data available on the endogenous exposure mixture, however there are concerns regarding the causal identifiability of the latent factors and loadings matrix. A more detailed investigation and development of a general causal inference framework that harmonizes individual chemical exposures with exogenous and endogenous exposure mixtures is critically important to ensure that there is a unified causal framework for environmental health research.

5.3.2 Unmeasured Confounding due to Co-Exposures

An additional complication to developing a causal framework for exposure mixtures is that assaying technology is not at the stage where thousands of chemicals across a wide spectrum of exposure classes can be practically measured. As a result, environmental health studies tend to restrict their focus to a handful of exposure classes, which may or may not correspond to real-world exposure events. Section 1 introduced the notion of confounding via co-exposure, as a central motivation for transitioning to joint models of chemical exposures. However, given the scope restriction due to technological and financial limitations, there are likely many other unmeasured co-exposures for each of the measured chemical exposures. Moreover, of the measured chemical exposures, a sizable portion often need to be discarded due to measurement error issues, making them functionally unmeasured for the purposes of statistical modeling. Therefore, a comprehensive causal framework for analyzing exposure mixtures must address unmeasured confounding via co-exposure. A promising future direction is to consider modifications of approaches like the deconfounder which

account for unmeasured confounding provided that the unmeasured confounders are the root causes for multiple exposures (Wang and Blei, 2019). Approaches based on ideas from the deconfounder may be particularly suitable for environmental health research because of the rich correlation structure in the exposure space; chemical exposures within the same exposure class tend to co-occur, meaning that unmeasured co-exposures likely correspond to multiple measured exposures.

5.3.3 Incorporating Measurement Error

This dissertation primarily focuses on statistical methods under the implicit assumption of no exposure measurement error, however it is well known that measurement error in chemical exposure assessment is a ubiquitous challenge. Measurement error can occur randomly, as there is inherent variability in exposure ascertainment, or systematically, where there is some underlying mechanism governing how errors are generated. Some examples of systematic exposure measurement errors include imperfect calibration of measurement instruments, sensitivity of measurement instruments to detect the presence of a chemical, and temporal degradation of stored biosamples. The statistical consequence of ignoring random measurement error is often attenuation of effect estimates towards the null (Chesher, 1991), but systematic forms of measurement error can also lead to biases away from the null.

One common form of systemic exposure measurement error in chemical exposure assessment occurs when the measurement instrument is not sensitive enough to detect chemical concentrations below a particular threshold, called the limit of detection (LOD) (Boss et al., 2019). Functionally, the distribution of observed chemical concentrations is left censored, where values below the LOD range between a concentration of zero and a concentration equal to the LOD. Outcome modeling when independent variables are subject to detection limits is handled either by specifying a measurement error model that models the left censoring process (Lynn, 2001; Nie et al., 2010; May et al., 2011; Kong and Nan, 2016; Chen et al., 2022) or multiply imputing concentrations below the LOD (Lynn, 2001; Lubin et al., 2004; Chen et al., 2011; Arunajadai and Rauh, 2012; Wang and Feng, 2012; Bernhardt et al., 2015; Atem et al., 2017; Ding et al., 2018; Boss et al., 2019). In current practice, the gold standard measurement error models combine the desired outcome model with a semiparametric accelerated failure time model to account the measurement error process defined by LODs (Kong and Nan, 2016; Chen et al., 2022). Imputation approaches based on accelerated failure time models have also been developed to impute values below the LOD (Atem et al., 2017; Ding et al., 2018). A causal framework for modeling multiple exposures would ideally have the flexibility to handle observed exposures that are subject to a left censoring process defined by their respective LODs.

5.3.4 Final Thoughts

Sections 5.3.1, 5.3.2, and 5.3.3 illustrate the complexity of defining a comprehensive causal framework for mixtures analyses. On one hand we need to distinguish between chemical exposures at the source and inside the human body. If a study obtains exposure measurements from biological samples, that leaves three sets of latent variables to infer: the exogenous exposure mixtures that lead to the observed endogenous chemical exposures in biological samples, the true endogenous exposure concentrations themselves, and the unmeasured co-exposures that may distort estimated mixture effects if not accounted for. This is incredibly ambitious, possibly unrealistic, but a substantial breakthrough on this front would, in our view, be a seminal contribution to the field of environmental statistics.

5.4 Extensions to Exposomics

This dissertation concerns itself with joint models of chemical exposures and endogenous biomarkers, however there is also interest in statistical models associating the exposome, defined as the totality of all exposures that an individual experiences throughout their life, with health outcomes of interest (Wild, 2005). When discussing the exposome, exogenous exposures are defined broadly, encompassing chemical exposures, infectious agents, the lived environment, and lifestyle factors, among others (Sillé et al., 2023). Given a large and varied collection of exposures — or, alternatively, omics structures thought to contain exposure signatures, such as the metabolome — one can jointly associate the measured exposome with an outcome of interest and evaluate intermediary pathways by which the measured exposome may lead to an outcome of interest.

GIGG regression is generally adaptable to settings with large collections of exposures, however there are a few challenges that require further discussion. The first difficulty arises in the context of untargeted identification of chemical exposures or exposure signatures from biological samples. GIGG regression requires a priori specification of the exposure groups, which is not easily defined in the context of untargeted chemical exposure data. In general, our recommendation is to define groups of size one, corresponding to exposures or exposure signatures that are unknown or have no a priori known grouping structure. Another path forward is to consider a methodological extension of GIGG regression that treats the grouping structure as an unknown variable and models the group assignment of exposures or exposure signatures using a dirichlet process prior (Ferguson, 1973) or clustering procedure (Witten et al., 2014). The other major challenge is related to computation and hyperparameter estimation, as exposome-wide analyses are inherently high-dimensional. As stated in Section 2.4, the primary computational bottleneck for a naive Gibbs sampling algorithm is obtaining full conditional draws of $\boldsymbol{\beta}$. For "small n, large p" problems, the Woodbury identity can

be utilized so that the full conditional update of $\boldsymbol{\beta}$ scales linearly in p (Bhattacharya et al., 2016). If n and p are both large, say an order of magnitude of 10,000 each, there are several recently developed approximation approaches, the former of which exploits the ability of the horseshoe prior to shrink $\tau^2\lambda_{gj}^2$ close to zero (Johndrow et al., 2020) while the latter uses a conjugate gradient algorithm to find an approximate solution to $\boldsymbol{Q}\boldsymbol{\beta} = \boldsymbol{v}$ (Nishimura and Suchard, 2022). Parallelization can also be used within the Gibbs sampler to simultaneously update the shrinkage parameters corresponding to each group (Terenin et al., 2019). With respect to hyperparameter estimation, it is also unlikely that MMLE with group-specific hyperparameters will perform as well as it does for the multi-pollutant models considered in Chapter 2. It is therefore our recommendation to use MMLE with a common set of GIGG hyperparameters shared across all groups for exposome-wide analyses.

To utilize MESSI in exposome-wide analyses, there are some requisite methodological extensions that must be developed. The first methodological generalization is to develop a multi-exposure version of MESSI, the challenges of which are articulated in Section 5.2. If there is external summary-level information on the total effect which coincides with the desired internal mediation analysis, then the next methodological extension is to extend MESSI to handle high-dimensional mediators. This is particularly important because a common intermediary pathway of interest in exposome-wide analyses are epigenetic mechanisms such as DNA methylation (Song et al., 2020; Sillé et al., 2023). Intuitively, one can penalize the regression coefficient vectors $\boldsymbol{\alpha}_a$ and $\boldsymbol{\beta}_m$, although there is not a consensus in the high-dimensional mediation literature on the specific penalty functions to assign to $\boldsymbol{\alpha}_a$ and $\boldsymbol{\beta}_m$. Some of our preliminary theoretical results suggest that, when $\widehat{\theta}_a^E \rightarrow_p \theta_a^I$ and the number of mediators are fixed, Theorem 3.3.1 and Theorem 3.3.2 generalize to a version of MESSI with adaptive LASSO penalties on $\boldsymbol{\alpha}_a$ and $\boldsymbol{\beta}_m$. Further work is required to develop code and evaluate the performance of MESSI with adaptive lasso penalties in high-dimensional settings.

Conceptually, the Total Effect Mixture Map and the Mediation Mixture Map are well-suited for exposome-wide analyses, provided that the exposure data can be appropriately modeled with a latent factor model. Specifically, the dimension reduction attributable to latent factor modeling helps deal with concerns regarding the high-dimensional nature of the observed exposures. Moreover, the Mediation Mixture Map model specification endows $\boldsymbol{\alpha}_a$ and $\boldsymbol{\beta}_m$ with shrinkage priors to simultaneously handle high-dimensional mediator spaces. Methodological variants of the Mediation Mixture Map that use dimension reduction techniques on the mediator space as an alternative to shrinkage priors, should also be appropriate for mediation analyses with high-dimensional mediator spaces. However, one challenge for the Mediation Mixture Map are settings where the exposures have distributions that are not in accordance with modeling assumptions in the latent factor model. As a concrete example, Goutman et al. (2023) summarized occupational history

variables through the number of job-years worked in standardized occupational categories, resulting in zero-inflated, right-skewed exposure variables. If one were to perform an exposome-wide analysis that considers occupational history, the zero-inflated and right-skewed nature of the occupational exposure variables clearly violates the normality assumption on $\boldsymbol{A} \mid \boldsymbol{L}$. In such cases, the latent factor component in the Mediation Mixture Map needs to be modified to more flexibly accommodate different types of exposure distributions. Future work is needed to determine which types of exposure data should be modeled using a latent factor model, and the latent factor model component of the Mediation Mixture Map needs to be generalized to accommodate exposure data of different types and structures.

APPENDIX A

Supplement for Chapter 2

A.1 Distributions used in Chapter 2

Beta Prime Distribution:

$$X \sim \beta'(a,b) \implies f_X(x) = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} x^{a-1}(1+x)^{-a-b}, \quad x > 0.$$

Gamma Distribution:

$$X \sim G(a,b) \implies f_X(x) = \frac{b^a}{\Gamma(a)} x^{a-1} \exp(-bx), \quad x > 0.$$

Generalized Inverse Gaussian Distribution:

$$X \sim GIG(\lambda,\psi,\chi) \implies f_X(x) = \frac{(\psi/\chi)^{\lambda/2}}{2K_\lambda(\sqrt{\psi\chi})} x^{\lambda-1} \exp\left(-\frac{1}{2}\left(\frac{\chi}{x}+\psi x\right)\right), \quad x > 0,$$

where $K_\lambda(\cdot)$ is the modified Bessel function of the third kind with index λ (Hörmann and Leydold, 2014).
Half-Cauchy Distribution:

$$X \sim C^+(0,\sigma) \implies f_X(x) = \frac{2}{\pi\sigma(1+x^2/\sigma^2)}, \quad x > 0.$$

Inverse-Gamma Distribution:

$$X \sim IG(a,b) \implies f_X(x) = \frac{b^a}{\Gamma(a)} x^{-a-1} \exp\left(-\frac{b}{x}\right), \quad x > 0.$$

A.2 Proof of Theorem 2.2.1

For shorthand, we will use:

$$r(x) \sim s(x) := \lim_{x\to\infty} \frac{r(x)}{s(x)} = 1.$$

Define

$$L(u) = \frac{1}{(\tau^2)^\lambda \mathcal{B}(a_g, b_g)} \left(\frac{u/\tau^2}{1+u/\tau^2} \right)^{a_g}, \quad \mathcal{B}(a_g, b_g) = \frac{\Gamma(a_g)\Gamma(b_g)}{\Gamma(a_g+b_g)},$$

which is slowly-varying function, i.e.,

$$\lim_{u\to\infty} \frac{L(tu)}{L(u)} = 1,$$

for all $t > 0$. Moreover, let

$$\pi(\beta_{gj} \mid \tau^2, a_g, b_g) = \int_0^\infty (2\pi u)^{-1/2} \exp\left(-\frac{\beta_{gj}^2}{2u} \right) f(u \mid \tau^2, a_g, b_g) du$$

denote the normal variance mixture probability density function and let $f(u \mid \tau^2, a_g, b_g)$ denote the scaled $\beta'(a_g, b_g)$ mixing density function with fixed scale parameter τ^2. Then

$$\lim_{u\to\infty} \frac{f(u \mid \tau^2, a_g, b_g)}{\exp(-\psi_+ u) u^{\lambda-1} L(u)}$$

$$= \lim_{u\to\infty} \frac{(\tau^2 \mathcal{B}(a_g, b_g))^{-1} (u/\tau^2)^{a_g-1} (1+u/\tau^2)^{-(a_g+b_g)}}{\exp(-\psi_+ u) u^{\lambda-1} L(u)} = \lim_{u\to\infty} \frac{(u/\tau^2)^{-1} (1+u/\tau^2)^{-b_g}}{\exp(-\psi_+ u)(u/\tau^2)^{\lambda-1}}$$

$$= \lim_{u\to\infty} \exp(\psi_+ u)(u/\tau^2)^{-\lambda} (1+u/\tau^2)^{-b_g},$$

where $\psi_+ = \sup\{w \in \mathbb{R} : \phi(w) < \infty\}$ and

$$\phi(w) = \frac{1}{\mathcal{B}(a_g, b_g)} \int_0^\infty \exp(wu) \frac{1}{\tau^2} \left(\frac{u}{\tau^2} \right)^{a_g-1} \left(1 + \frac{u}{\tau^2} \right)^{-(a_g+b_g)} du.$$

Note that $\psi_+ = 0$. Fix $\lambda = -b_g$. Then,

$$\lim_{u\to\infty} \exp(\psi_+ u)(u/\tau^2)^{-\lambda} (1+u/\tau^2)^{-b_g} = \lim_{u\to\infty} \left(\frac{u/\tau^2}{1+u/\tau^2} \right)^{b_g} = 1.$$

By Theorem 6.1 in Barndorff-Nielsen et al. (1982) we conclude that

$$\pi(\beta_{gj} \mid \tau^2, a_g, b_g) \sim (2\pi)^{-1/2} 2^{b_g+1/2} \Gamma(b_g + 1/2) |\beta_{gj}|^{-(1+2b_g)} L(\beta_{gj}^2).$$

To get the index of regular variation, we note the following straightforward lemma:

Lemma 1. *Suppose that r and s are two positive, measurable functions such that $r(x) \sim s(x)$ and s is regularly varying with index $\omega \in \mathbb{R}$. Then r is regularly varying with index ω.*

Proof.

$$\lim_{x\to\infty} \frac{r(tx)}{r(x)} = \lim_{x\to\infty} \frac{r(tx)}{r(x)} \frac{s(tx)}{s(tx)} \frac{s(x)}{s(x)} = \lim_{x\to\infty} \left(\frac{r(tx)/s(tx)}{r(x)/s(x)} \right) \frac{s(tx)}{s(x)} = \lim_{x\to\infty} \frac{s(tx)}{s(x)} = t^\omega.$$

Despite its simplicity, Lemma 1 is of great practical use, particularly if the function whose tail behavior we are interested in does not have a closed form. When working with global-local mixture priors we often do not have closed

form marginal prior distributions for $\boldsymbol{\beta}$ and it is usually easier to construct and work with a closed form function that has asymptotically equivalent tail behavior. Since the index of regular variation of

$$(2\pi)^{-1/2}2^{b_g+1/2}\Gamma(b_g+1/2)|\beta_{gj}|^{-(1+2b_g)}L(\beta_{gj}^2)$$

is $\omega = -1-2b_g$, then by Lemma 1 the index of regular variation of $\pi(\beta_{gj} \mid \tau^2, a_g, b_g)$ is also $\omega = -1-2b_g$.

A.3 Proof of Theorem 2.3.1

Let $\mathcal{G}_n = \{\gamma_1^2, ..., \gamma_{G_n}^2\}$ denote the collection of group shrinkage parameters and let p_g indicate the number of regressors in the g-th group. Define the following sets: $\mathcal{A}_g = \{j : \beta_{gj}^0 \neq 0\}$, $\mathcal{A}_g^c = \{j : \beta_{gj}^0 = 0\}$, $\mathcal{A}_n = \{(g,j) : \beta_{gj}^0 \neq 0\}$, $\mathcal{A}_n^c = \{(g,j) : \beta_{gj}^0 = 0\}$. In words, $\mathcal{A}_g$ is the active set for the g-th group, $\mathcal{A}_g^c$ is the non-active set for the g-th group, $\mathcal{A}_n$ is the active set across all groups, and $\mathcal{A}_n^c$ is the non-active set across all groups.

$$\pi\big(\boldsymbol{\beta}_n \mid \mathcal{G}_n, \mathcal{H}_n, \tau_n^2\big) = \prod_{g=1}^{G_n}\prod_{j=1}^{p_g} \frac{\Gamma(b_g + 1/2)}{\Gamma(b_g)\sqrt{2\pi\tau_n^2\gamma_g^2}}\bigg(1 + \frac{\beta_{gj}^2}{2\tau_n^2\gamma_g^2}\bigg)^{-(b_g+1/2)}$$

Then, we see that

$$\pi\bigg(\boldsymbol{\beta}_n : \|\boldsymbol{\beta}_n - \boldsymbol{\beta}_n^0\|_2 < \frac{\Delta}{n^{\rho/2}} \;\bigg|\; \mathcal{G}_n, \mathcal{H}_n, \tau_n^2\bigg)$$

$$\geq \prod_{g=1}^{G_n}\bigg[\prod_{j\in\mathcal{A}_g}\pi\bigg(|\beta_{gj} - \beta_{gj}^0| < \frac{\Delta}{\sqrt{p_n}n^{\rho/2}} \;\bigg|\; \mathcal{G}_n, \mathcal{H}_n, \tau_n^2\bigg) \times \prod_{j\in\mathcal{A}_g^c}\pi\bigg(|\beta_{gj}| < \frac{\Delta}{\sqrt{p_n}n^{\rho/2}} \;\bigg|\; \mathcal{G}_n, \mathcal{H}_n, \tau_n^2\bigg)\bigg].$$

Continuing

$$\pi\bigg(|\beta_{gj} - \beta_{gj}^0| < \frac{\Delta}{\sqrt{p_n}n^{\rho/2}} \;\bigg|\; \mathcal{G}_n, \mathcal{H}_n, \tau_n^2\bigg)$$

$$= \int_{\beta_{gj}^0 - \frac{\Delta}{\sqrt{p_n}n^{\rho/2}}}^{\beta_{gj}^0 + \frac{\Delta}{\sqrt{p_n}n^{\rho/2}}} \frac{\Gamma(b_g + 1/2)}{\Gamma(b_g)\sqrt{2\pi\tau_n^2\gamma_g^2}}\bigg(1 + \frac{\beta_{gj}^2}{2\tau_n^2\gamma_g^2}\bigg)^{-(b_g+1/2)} d\beta_{gj}$$

$$\geq \frac{2\Delta\Gamma(b_g + 1/2)}{\Gamma(b_g)\sqrt{2\pi\tau_n^2\gamma_g^2}\sqrt{p_n}n^{\rho/2}}\bigg(1 + \frac{(L_n + \frac{\Delta}{\sqrt{p_n}n^{\rho/2}})^2}{2\tau_n^2\gamma_g^2}\bigg)^{-(b_g+1/2)}$$

and

$$\pi\bigg(|\beta_{gj}| < \frac{\Delta}{\sqrt{p_n}n^{\rho/2}} \;\bigg|\; \mathcal{G}_n, \mathcal{H}_n, \tau_n^2\bigg)$$

$$= \int_{-\frac{\Delta}{\sqrt{p_n}n^{\rho/2}}}^{\frac{\Delta}{\sqrt{p_n}n^{\rho/2}}} \frac{\Gamma(b_g + 1/2)}{\Gamma(b_g)\sqrt{2\pi\tau_n^2\gamma_g^2}}\bigg(1 + \frac{\beta_{gj}^2}{2\tau_n^2\gamma_g^2}\bigg)^{-(b_g+1/2)} d\beta_{gj}$$

$$= 2\int_0^{\frac{\Delta}{\sqrt{p_n}n^{\rho/2}}} \frac{\Gamma(b_g + 1/2)}{\Gamma(b_g)\sqrt{2\pi\tau_n^2\gamma_g^2}}\bigg(1 + \frac{\beta_{gj}^2}{2\tau_n^2\gamma_g^2}\bigg)^{-(b_g+1/2)} d\beta_{gj}$$

$$\geq 2\int_0^{\frac{\Delta}{\sqrt{p_n}n^{\rho/2}}} \frac{\Gamma(b_g + 1/2)}{\Gamma(b_g)\sqrt{2\pi\tau_n^2\gamma_g^2}}\exp\bigg(-\frac{\beta_{gj}(b_g + 1/2)}{\sqrt{2\tau_n^2\gamma_g^2}}\bigg) d\beta_{gj}$$

$$= \frac{2\Gamma(b_g + 1/2)}{\Gamma(b_g)\sqrt{2\pi\tau_n^2\gamma_g^2}}\bigg(\frac{\sqrt{2\tau_n^2\gamma_g^2}}{b_g + 1/2}\bigg)\bigg[1 - \exp\bigg(-\frac{\Delta(b_g + 1/2)}{\sqrt{2\tau_n^2\gamma_g^2 p_n n^\rho}}\bigg)\bigg]$$

$$= \frac{2\Gamma(b_g + 1/2)}{\Gamma(b_g)\sqrt{\pi}(b_g + 1/2)}\bigg[1 - \exp\bigg(-\frac{\Delta(b_g + 1/2)}{\sqrt{2\tau_n^2\gamma_g^2 p_n n^\rho}}\bigg)\bigg].$$

Note that we have the above inequality because $(1+x)^{-1} \geq \exp(-\sqrt{x})$ for all $x \geq 0$. Therefore,

$$\prod_{g=1}^{G_n}\left[\prod_{j\in\mathcal{A}_g}\pi\left(|\beta_{gj}-\beta_{gj}^0|<\frac{\Delta}{\sqrt{p_n}n^{\rho/2}}\,\middle|\,\mathcal{G}_n,\mathcal{H}_n,\tau_n^2\right)\times\prod_{j\in\mathcal{A}_g^c}\pi\left(|\beta_{gj}|<\frac{\Delta}{\sqrt{p_n}n^{\rho/2}}\,\middle|\,\mathcal{G}_n,\mathcal{H}_n,\tau_n^2\right)\right]$$

$$\geq\prod_{g=1}^{G_n}\left(\frac{2\Delta\Gamma(b_g+1/2)}{\Gamma(b_g)\sqrt{2\pi\tau_n^2\gamma_g^2}\sqrt{p_n}n^{\rho/2}}\right)^{|\mathcal{A}_g|}\left(1+\frac{(L_n+\frac{\Delta}{\sqrt{p_n}n^{\rho/2}})^2}{2\tau_n^2\gamma_g^2}\right)^{-|\mathcal{A}_g|(b_g+1/2)}$$

$$\times\left(\frac{2\Gamma(b_g+1/2)}{\Gamma(b_g)\sqrt{\pi}(b_g+1/2)}\right)^{|\mathcal{A}_g^c|}\left[1-\exp\left(-\frac{\Delta(b_g+1/2)}{\sqrt{2\tau_n^2\gamma_g^2p_nn^\rho}}\right)\right]^{|\mathcal{A}_g^c|}.$$

Substituting in $\tau_n^2 = C/(p_n n^\rho \log(n))$ and taking the negative logarithm of the final expression yields

$$-\sum_{g=1}^{G_n}\left[|\mathcal{A}_g|\log\left(\frac{2\Delta\Gamma(b_g+1/2)\sqrt{\log(n)}}{\Gamma(b_g)\sqrt{2C\pi\gamma_g^2}}\right)\right.$$

$$-|\mathcal{A}_g|(b_g+1/2)\log\left(1+\frac{p_nn^\rho\log(n)(L_n+\frac{\Delta}{\sqrt{p_n}n^{\rho/2}})^2}{2C\gamma_g^2}\right)$$

$$\left.+|\mathcal{A}_g^c|\log\left(\frac{2\Gamma(b_g+1/2)}{\Gamma(b_g)\sqrt{\pi}(b_g+1/2)}\right)+|\mathcal{A}_g^c|\log\left(1-\exp\left(-\frac{\Delta(b_g+1/2)\sqrt{\log(n)}}{\sqrt{2C\gamma_g^2}}\right)\right)\right]$$

$$=\sum_{g=1}^{G_n}\left[-|\mathcal{A}_g|\log\left(\frac{2\Delta\Gamma(b_g+1/2)\sqrt{\log(n)}}{\Gamma(b_g)\sqrt{2C\pi\gamma_g^2}}\right)\right.$$

$$+|\mathcal{A}_g|(b_g+1/2)\log\left(1+\frac{p_nn^\rho\log(n)(L_n+\frac{\Delta}{\sqrt{p_n}n^{\rho/2}})^2}{2C\gamma_g^2}\right)$$

$$\left.-|\mathcal{A}_g^c|\log\left(\frac{2\Gamma(b_g+1/2)}{\Gamma(b_g)\sqrt{\pi}(b_g+1/2)}\right)-|\mathcal{A}_g^c|\log\left(1-\exp\left(-\frac{\Delta(b_g+1/2)\sqrt{\log(n)}}{\sqrt{2C\gamma_g^2}}\right)\right)\right].$$

Let

$$T_{n1}=\inf_{g\in\{1,\ldots,G_n\}}\log\left(\frac{2\Delta\Gamma(b_g+1/2)}{\Gamma(b_g)\sqrt{2C\pi\gamma_g^2}}\right),$$

$$T_{n2}=\inf_{g\in\{1,\ldots,G_n\}}\log\left(\frac{2\Gamma(b_g+1/2)}{\Gamma(b_g)\sqrt{\pi}(b_g+1/2)}\right),$$

$$T_{n3}=\inf_{g\in\{1,\ldots,G_n\}}\frac{\Delta(b_g+1/2)}{\sqrt{2C\gamma_g^2}},$$

$$\gamma_{n,min}^2=\inf_{g\in\{1,\ldots,G_n\}}\gamma_g^2,\ \text{ and }\ b_{n,max}=\sup_{g\in\{1,\ldots,G_n\}}b_g.$$

Then,

$$\sum_{g=1}^{G_n}\Bigg[-|\mathcal{A}_g|\log\left(\frac{2\Delta\Gamma(b_g+1/2)\sqrt{\log(n)}}{\Gamma(b_g)\sqrt{2C\pi\gamma_g^2}}\right)$$

$$+|\mathcal{A}_g|(b_g+1/2)\log\left(1+\frac{p_n n^{\rho}\log(n)(L_n+\frac{\Delta}{\sqrt{p_n}n^{\rho/2}})^2}{2C\gamma_g^2}\right)$$

$$-|\mathcal{A}_g^c|\log\left(\frac{2\Gamma(b_g+1/2)}{\Gamma(b_g)\sqrt{\pi}(b_g+1/2)}\right)-|\mathcal{A}_g^c|\log\left(1-\exp\left(-\frac{\Delta(b_g+1/2)\sqrt{\log(n)}}{\sqrt{2C\gamma_g^2}}\right)\right)\Bigg]$$

$$\leq\sum_{g=1}^{G_n}\Bigg[-\frac{|\mathcal{A}_g|}{2}\log\big(\log(n)\big)-|\mathcal{A}_g|T_{n1}+|\mathcal{A}_g|(b_{n,max}+1/2)\log\left(1+\frac{p_n n^{\rho}\log(n)(L_n+\frac{\Delta}{\sqrt{p_n}n^{\rho/2}})^2}{2C\gamma_{n,min}^2}\right)$$

$$-|\mathcal{A}_g^c|T_{n2}-|\mathcal{A}_g^c|\log\Big(1-\exp\Big(T_{n3}\sqrt{\log(n)}\Big)\Big)\Bigg]$$

$$=-\frac{|\mathcal{A}_n|}{2}\log\big(\log(n)\big)-|\mathcal{A}_n|T_{n1}+|\mathcal{A}_n|(b_{n,max}+1/2)\log\left(1+\frac{p_n n^{\rho}\log(n)(L_n+\frac{\Delta}{\sqrt{p_n}n^{\rho/2}})^2}{2C\gamma_{n,min}^2}\right)$$

$$-|\mathcal{A}_n^c|T_{n2}-|\mathcal{A}_n^c|\log\Big(1-\exp\Big(-T_{n3}\sqrt{\log(n)}\Big)\Big).$$

Note that the above expression is dominated by the

$$|\mathcal{A}_n|(b_{n,max}+1/2)\log\left(1+\frac{p_n n^{\rho}\log(n)(L_n+\frac{\Delta}{\sqrt{p_n}n^{\rho/2}})^2}{2C\gamma_{n,min}^2}\right)$$

term. If $|\mathcal{A}_n|=o(n/\log(n))$, then by *Theorem 1* in Armagan et al. (2013b), we obtain posterior consistency conditional on $\mathcal{G}_n$, i.e., $\forall\epsilon>0$

$$\pi\big(\boldsymbol{\beta}_n:\|\boldsymbol{\beta}_n-\boldsymbol{\beta}_n^0\|_2<\epsilon\;\big|\;\boldsymbol{y}_n,\mathcal{G}_n,\mathcal{H}_n,\tau_n^2,\sigma^2\big)\to 1$$

almost surely. Lastly, by combining the law of iterated expectations with the dominated convergence theorem, we obtain

$$\pi\big(\boldsymbol{\beta}_n:\|\boldsymbol{\beta}_n-\boldsymbol{\beta}_n^0\|_2<\epsilon\;\big|\;\boldsymbol{y}_n,\mathcal{H}_n,\tau_n^2,\sigma^2\big)$$

$$=E_{\mathcal{G}_n|\boldsymbol{y}_n,\mathcal{H}_n,\tau_n^2,\sigma^2}\Big[\pi\big(\boldsymbol{\beta}_n:\|\boldsymbol{\beta}_n-\boldsymbol{\beta}_n^0\|_2<\epsilon\;\big|\;\boldsymbol{y}_n,\mathcal{G}_n,\mathcal{H}_n,\tau_n^2,\sigma^2\big)\Big]$$

$$\to E_{\mathcal{G}_n|\boldsymbol{y}_n,\mathcal{H}_n,\tau_n^2,\sigma^2}[1]=1.$$

A.4 Proof of Theorem 2.3.2

Let $\boldsymbol{M} = \boldsymbol{I}_p - (\boldsymbol{I}_p + (\boldsymbol{X}^\top \boldsymbol{X})^{-1}\boldsymbol{D})^{-1} = (\boldsymbol{I}_p + \boldsymbol{D}^{-1}\boldsymbol{X}^\top \boldsymbol{X})^{-1}$ where $\boldsymbol{D}^{-1} = \sigma^{-2}\tau^2\boldsymbol{\Gamma}\boldsymbol{\Lambda}$. For brevity, we will denote the full conditional posterior mean as

$$E[\boldsymbol{\beta} \mid \cdot] = E[\boldsymbol{\beta} \mid \boldsymbol{y}, \sigma^2, \tau^2, \boldsymbol{\Gamma}, \boldsymbol{\Lambda}]$$

The Rayleigh quotient inequality yields,

$$\left\|\hat{\boldsymbol{\beta}}^{OLS} - E[\boldsymbol{\beta} \mid \cdot]\right\|_2^2 \geq e_{min}(\boldsymbol{M}^\top \boldsymbol{M})\left\|\hat{\boldsymbol{\beta}}^{OLS}\right\|_2^2 = \big(\theta_{min}(\boldsymbol{M})\big)^2\left\|\hat{\boldsymbol{\beta}}^{OLS}\right\|_2^2,$$

where $e_{min}(\boldsymbol{M}^\top \boldsymbol{M})$ denotes the minimum eigenvalue of $\boldsymbol{M}^\top \boldsymbol{M}$ and $\theta_{min}(\boldsymbol{M})$ denotes the minimum singular value of $\boldsymbol{M}$. Then, we have

$$\big(\theta_{min}(\boldsymbol{M})\big)^2\left\|\hat{\boldsymbol{\beta}}^{OLS}\right\|_2^2 = \frac{1}{\big(\theta_{max}(\boldsymbol{I}_p + \boldsymbol{D}^{-1}\boldsymbol{X}^\top \boldsymbol{X})\big)^2}\left\|\hat{\boldsymbol{\beta}}^{OLS}\right\|_2^2 = \frac{1}{\|\boldsymbol{I}_p + \boldsymbol{D}^{-1}\boldsymbol{X}^\top \boldsymbol{X}\|_{\mathcal{O}}^2}\left\|\hat{\boldsymbol{\beta}}^{OLS}\right\|_2^2,$$

where $\|\cdot\|_{\mathcal{O}}$ is the operator 2-norm and $\|\cdot\|_2$ is the usual L_2-norm. The operator 2-norm is sub-additive and sub-multiplicative, implying that

$$\begin{aligned}
\frac{1}{\|\boldsymbol{I}_p + \boldsymbol{D}^{-1}\boldsymbol{X}^\top \boldsymbol{X}\|_{\mathcal{O}}^2}\left\|\hat{\boldsymbol{\beta}}^{OLS}\right\|_2^2 &\geq \frac{1}{\big(\|\boldsymbol{I}_p\|_{\mathcal{O}} + \|\boldsymbol{D}^{-1}\|_{\mathcal{O}}\|\boldsymbol{X}^\top \boldsymbol{X}\|_{\mathcal{O}}\big)^2}\left\|\hat{\boldsymbol{\beta}}^{OLS}\right\|_2^2 \\
&= \left(\frac{1}{1 + \theta_{max}(\boldsymbol{X}^\top \boldsymbol{X})\sigma^{-2}\tau^2 \max_{(g,j)} \gamma_g^2\lambda_{gj}^2}\right)^2\left\|\hat{\boldsymbol{\beta}}^{OLS}\right\|_2^2
\end{aligned}$$

First, note that

$$\begin{aligned}
\pi\left(\frac{1}{1 + c\sigma^{-2}\tau^2 \max_{(g,j)} \gamma_g^2\lambda_{gj}^2} < \epsilon \;\middle|\; \boldsymbol{y}, \mathcal{H}, \tau^2, \sigma^2\right) &\leq \pi\left(\bigcup_{g=1}^{G}\bigcup_{j=1}^{p_g}\left\{\frac{1}{1 + c\sigma^{-2}\tau^2\gamma_g^2\lambda_{gj}^2} < \epsilon\right\} \;\middle|\; \boldsymbol{y}, \mathcal{H}, \tau^2, \sigma^2\right) \\
&\leq \sum_{g=1}^{G}\sum_{j=1}^{p_g}\pi\left(\frac{1}{1 + c\sigma^{-2}\tau^2\gamma_g^2\lambda_{gj}^2} < \epsilon \;\middle|\; \boldsymbol{y}, \mathcal{H}, \tau^2, \sigma^2\right),
\end{aligned}$$

where $c = \theta_{max}(\boldsymbol{X}^\top \boldsymbol{X})$. Define

$$K_{gj} = \frac{1}{1 + c\sigma^{-2}\tau^2\gamma_g^2\lambda_{gj}^2}$$

and for notational simplicity let $\mathcal{K} = \mathcal{K}_{-g} \cup \{K_{gl}\}$ where $\mathcal{K}_{-g} = \{K_{g'1} : g' \neq g\}$. Also, let $\mathcal{L} = \cup_{g=1}^{G}\mathcal{L}_g$ denote the collection of all local shrinkage parameters where $\mathcal{L}_g = \{\lambda_{gj}^2 : 1 \leq j \leq p_g\}$. Then,

$$\pi\big(K_{gl} < \epsilon \;\big|\; \boldsymbol{y}, \mathcal{H}, \tau^2, \sigma^2\big) = \frac{1}{\pi(\boldsymbol{y} \mid \mathcal{H}, \tau^2, \sigma^2)}\int_0^\epsilon \pi(\boldsymbol{y} \mid K_{gl}, \mathcal{H}, \tau^2, \sigma^2)\pi(K_{gl} \mid \mathcal{H}, \tau^2, \sigma^2)dK_{gl}$$

$$= \frac{1}{\pi(\boldsymbol{y} \mid \mathcal{H}, \tau^2, \sigma^2)}\int_0^\epsilon \int_{(0,1)^{p-1}} \int_{(0,\infty)^p} \pi(\boldsymbol{y} \mid \mathcal{K}, \mathcal{L}, \mathcal{H}, \tau^2, \sigma^2)\pi(\mathcal{K} \mid \mathcal{L}, \mathcal{H}, \tau^2, \sigma^2)\pi(\mathcal{L} \mid \mathcal{H}, \tau^2, \sigma^2)d\mathcal{L}d\mathcal{K}_{-g}dK_{gl},$$

where $(0,1)^{p-1}$ indicates a $p-1$ dimensional hypercube on $(0,1)$ and $(0,\infty)^p = (0,\infty) \times \cdots \times (0,\infty)$ p times. Looking at the individual components we first observe that $\pi(\boldsymbol{y} \mid \mathcal{K}, \mathcal{L}, \mathcal{H}, \tau^2, \sigma^2)$ is just a reparameterized version of

$\pi(\boldsymbol{y} \mid \boldsymbol{\Gamma}, \boldsymbol{\Lambda}, \mathcal{H}, \tau^2, \sigma^2)$, where

$$[\boldsymbol{y} \mid \boldsymbol{\Gamma}, \boldsymbol{\Lambda}, \mathcal{H}, \tau^2, \sigma^2] \sim N(\boldsymbol{0}, \sigma^2 \boldsymbol{I}_n + \tau^2 \boldsymbol{X}\boldsymbol{\Gamma}\boldsymbol{\Lambda}\boldsymbol{X}^\top).$$

Note that $\tau^2 \boldsymbol{X}\boldsymbol{\Gamma}\boldsymbol{\Lambda}\boldsymbol{X}^\top$ is a symmetric, positive definite matrix and therefore has positive, real eigenvalues. Thus the determinant of $\sigma^2 \boldsymbol{I}_n + \tau^2 \boldsymbol{X}\boldsymbol{\Gamma}\boldsymbol{\Lambda}\boldsymbol{X}^\top$ satisfies $|\sigma^2 \boldsymbol{I}_n + \tau^2 \boldsymbol{X}\boldsymbol{\Gamma}\boldsymbol{\Lambda}\boldsymbol{X}^\top| \geq \left(\sigma^2\right)^n$ and we get that

$$\pi(\boldsymbol{y} \mid \boldsymbol{\Gamma}, \boldsymbol{\Lambda}, \mathcal{H}, \tau^2, \sigma^2) \leq (2\pi\sigma^2)^{-n/2}.$$

Therefore, we conclude that,

$$\pi(\boldsymbol{y} \mid \mathcal{K}, \mathcal{L}, \mathcal{H}, \tau^2, \sigma^2) \leq (2\pi\sigma^2)^{-n/2}.$$

One helpful observation is that

$$\begin{aligned}
&\pi(\mathcal{K} \mid \mathcal{L}, \mathcal{H}, \tau^2, \sigma^2)\pi(\mathcal{L} \mid \mathcal{H}, \tau^2, \sigma^2) \\
&\qquad = \pi(K_{gl} \mid \lambda_{gl}^2, \mathcal{H}, \tau^2, \sigma^2)\pi(\lambda_{gl}^2 \mid \mathcal{H}, \tau^2, \sigma^2) \prod_{g' \neq g} \pi(K_{g'1} \mid \mathcal{L}_{g'}, \mathcal{H}, \tau^2, \sigma^2)\pi(\mathcal{L}_{g'} \mid \mathcal{H}, \tau^2, \sigma^2)
\end{aligned}$$

Thus, we get a simplified upper bound

$$\begin{aligned}
&\pi\big(K_{gl} < \epsilon \,\big|\, \boldsymbol{y}, \mathcal{H}, \tau^2, \sigma^2\big) \\
&\qquad \leq \frac{1}{\pi(\boldsymbol{y} \mid \mathcal{H}, \tau^2, \sigma^2)} \int_0^\epsilon \int_0^\infty (2\pi\sigma^2)^{-n/2} \pi(K_{gl} \mid \lambda_{gl}^2, \mathcal{H}, \tau^2, \sigma^2)\pi(\lambda_{gl}^2 \mid \mathcal{H}, \tau^2, \sigma^2) d\lambda_{gl}^2 dK_{gl}.
\end{aligned}$$

Note that because we are conditioning on the local shrinkage parameter then calculating $\pi(K_{gl} \mid \lambda_{gl}^2, \mathcal{H}, \tau^2, \sigma^2)$ is a single variable transformation problem, where

$$\pi(K_{gl} \mid \lambda_{gl}^2, \mathcal{H}, \tau^2, \sigma^2) = \frac{1}{\Gamma(a_g)} \left(\frac{\sigma^2}{c\tau^2\lambda_{gl}^2} \right)^{a_g} (1 - K_{gl})^{a_g - 1} (K_{gl})^{-(1+a_g)} \exp\left(- \frac{\sigma^2(1 - K_{gl})}{c\tau^2\lambda_{gl}^2 K_{gl}} \right).$$

Moreover,

$$\pi(\lambda_{gl}^2 \mid \mathcal{H}, \tau^2, \sigma^2) = \frac{1}{\Gamma(b_g)} (\lambda_{gl}^2)^{-b_g - 1} \exp\left(- \frac{1}{\lambda_{gl}^2} \right),$$

is simply the prior on the local shrinkage parameters. Therefore,

$$
\begin{aligned}
&\pi\big(K_{gl} < \epsilon \,\big|\, \boldsymbol{y}, \mathcal{H}, \tau^2, \sigma^2\big) \\
&\quad\leq \frac{1}{(2\pi\sigma^2)^{n/2}\pi(\boldsymbol{y} \mid \mathcal{H}, \tau^2, \sigma^2)} \int_0^\epsilon \int_0^\infty \pi(K_{gl} \mid \lambda_{gl}^2, \mathcal{H}, \tau^2, \sigma^2)\pi(\lambda_{gl}^2 \mid \mathcal{H}, \tau^2, \sigma^2) d\lambda_{gl}^2 dK_{gl} \\
&\quad= \frac{1}{(2\pi\sigma^2)^{n/2}\Gamma(a_g)\Gamma(b_g)\pi(\boldsymbol{y} \mid \mathcal{H}, \tau^2, \sigma^2)} \left(\frac{\sigma^2}{c\tau^2}\right)^{a_g} \times \\
&\quad\int_0^\epsilon (1-K_{gl})^{a_g-1}(K_{gl})^{-(1+a_g)} \int_0^\infty \big(\lambda_{gl}^2\big)^{-(a_g+b_g)-1} \exp\bigg(-\bigg(1+\frac{\sigma^2(1-K_{gl})}{c\tau^2 K_{gl}}\bigg)\Big/\lambda_{gl}^2\bigg) d\lambda_{gl}^2 dK_{gl} \\
&\quad= \frac{\Gamma(a_g+b_g)}{(2\pi\sigma^2)^{n/2}\Gamma(a_g)\Gamma(b_g)\pi(\boldsymbol{y} \mid \mathcal{H}, \tau^2, \sigma^2)} \left(\frac{\sigma^2}{c\tau^2}\right)^{a_g} \times \\
&\quad\int_0^\epsilon (1-K_{gl})^{a_g-1}(K_{gl})^{-(1+a_g)} \left(1+\frac{\sigma^2(1-K_{gl})}{c\tau^2 K_{gl}}\right)^{-(a_g+b_g)} dK_{gl} \\
&\quad\leq \frac{\Gamma(a_g+b_g)}{(2\pi\sigma^2)^{n/2}\Gamma(a_g)\Gamma(b_g)\pi(\boldsymbol{y} \mid \mathcal{H}, \tau^2, \sigma^2)} \left(\frac{\sigma^2}{c\tau^2}\right)^{a_g} \\
&\quad\int_0^\epsilon (1-K_{gl})^{a_g-1}(K_{gl})^{-(1+a_g)} \left(\frac{c\tau^2 K_{gl}}{c\tau^2 K_{gl}+\sigma^2(1-\epsilon)}\right)^{(a_g+b_g)} dK_{gl} \\
&\quad\leq \frac{\Gamma(a_g+b_g)(1-\epsilon)^{-(a_g+b_g)}}{(2\pi\sigma^2)^{n/2}\Gamma(a_g)\Gamma(b_g)\pi(\boldsymbol{y} \mid \mathcal{H}, \tau^2, \sigma^2)} \left(\frac{c\tau^2}{\sigma^2}\right)^{b_g} \int_0^\epsilon (1-K_{gl})^{a_g-1}(K_{gl})^{b_g-1} dK_{gl} \\
&\quad\leq \frac{\Gamma(a_g+b_g)\epsilon^{b_g}(1-\epsilon)^{-(a_g+b_g)}}{(2\pi\sigma^2)^{n/2}\Gamma(a_g)b_g\Gamma(b_g)\pi(\boldsymbol{y} \mid \mathcal{H}, \tau^2, \sigma^2)} \left(\frac{c\tau^2}{\sigma^2}\right)^{b_g} \max\{1, (1-\epsilon)^{a_g-1}\}.
\end{aligned}
$$

Lastly, we see that

$$
\lim_{\tau^2\to 0} \pi(\boldsymbol{y} \mid \mathcal{H}, \tau^2, \sigma^2) = \lim_{\tau^2\to 0} \int_{(0,\infty)^G} \int_{(0,\infty)^p} \pi(\boldsymbol{y} \mid \mathcal{G}, \mathcal{L}, \mathcal{H}, \tau^2, \sigma^2)\pi(\mathcal{G} \mid \mathcal{H}, \tau^2, \sigma^2)\pi(\mathcal{L} \mid \mathcal{H}, \tau^2, \sigma^2) d\mathcal{L} d\mathcal{G}.
$$

Since $\pi(\boldsymbol{y} \mid \mathcal{G}, \mathcal{L}, \mathcal{H}, \tau^2, \sigma^2) = \pi(\boldsymbol{y} \mid \boldsymbol{\Gamma}, \boldsymbol{\Lambda}, \mathcal{H}, \tau^2, \sigma^2) \leq (2\pi\sigma^2)^{-n/2}$, then

$$
\int_{(0,\infty)^G} \int_{(0,\infty)^p} \pi(\boldsymbol{y} \mid \mathcal{G}, \mathcal{L}, \mathcal{H}, \tau^2, \sigma^2)\pi(\mathcal{G} \mid \mathcal{H}, \tau^2, \sigma^2)\pi(\mathcal{L} \mid \mathcal{H}, \tau^2, \sigma^2) d\mathcal{L} d\mathcal{G} \leq (2\pi\sigma^2)^{-n/2},
$$

and by the dominated convergence theorem we have that

$$
\lim_{\tau^2\to 0} \pi(\boldsymbol{y} \mid \mathcal{H}, \tau^2, \sigma^2) = \frac{1}{(2\pi\sigma^2)^{n/2}} \exp\bigg(-\frac{1}{2\sigma^2}\boldsymbol{y}^\top \boldsymbol{y}\bigg).
$$

So, we conclude that

$$
\begin{aligned}
\pi\bigg(\frac{1}{1+c\sigma^{-2}\tau^2 \max_{(g,j)} \gamma_g^2 \lambda_{gj}^2} < \epsilon \,\bigg|\, \boldsymbol{y}, \mathcal{H}, \tau^2, \sigma^2\bigg) &\leq \sum_{g=1}^G \sum_{j=1}^{p_g} \pi\big(K_{gj} < \epsilon \,\big|\, \boldsymbol{y}, \mathcal{H}, \tau^2, \sigma^2\big) \\
&\leq \sum_{g=1}^G \frac{p_g \Gamma(a_g+b_g)\epsilon^{b_g}(1-\epsilon)^{-(a_g+b_g)}}{(2\pi\sigma^2)^{n/2}\Gamma(a_g)b_g\Gamma(b_g)\pi(\boldsymbol{y} \mid \mathcal{H}, \tau^2, \sigma^2)} \left(\frac{c\tau^2}{\sigma^2}\right)^{b_g} \max\{1, (1-\epsilon)^{a_g-1}\},
\end{aligned}
$$

where the upper bound goes to zero as $\tau^2 \to 0$. Therefore, we have shown that for fixed $\epsilon \in (0, 1)$, that

$$\pi\left(\frac{1}{1 + \theta_{max}(\boldsymbol{X}^\top \boldsymbol{X})\sigma^{-2}\tau^2 \max_{(g,j)} \gamma_g^2 \lambda_{gj}^2} \geq \epsilon \;\middle|\; \boldsymbol{y}, \mathcal{H}, \tau^2, \sigma^2\right) \to 1$$

as $\tau^2 \to 0$. Since,

$$\left\|\hat{\boldsymbol{\beta}}^{OLS} - E[\boldsymbol{\beta} \mid \cdot]\right\|_2^2 \geq \left(\frac{1}{1 + \theta_{max}(\boldsymbol{X}^\top \boldsymbol{X})\sigma^{-2}\tau^2 \max_{(g,j)} \gamma_g^2 \lambda_{gj}^2}\right)^2 \left\|\hat{\boldsymbol{\beta}}^{OLS}\right\|_2^2,$$

then we must also have that

$$\pi\left(\left\|\hat{\boldsymbol{\beta}}^{OLS} - E[\boldsymbol{\beta} \mid \cdot]\right\|_2 \geq \epsilon \left\|\hat{\boldsymbol{\beta}}^{OLS}\right\|_2 \;\middle|\; \boldsymbol{y}, \mathcal{H}, \tau^2, \sigma^2\right) \to 1$$

as $\tau^2 \to 0$.

A.5 Proof of Corollary 2.3.1

For brevity, we will use the following notation

$$E[\boldsymbol{\beta}_g \mid \cdot] = E[\boldsymbol{\beta}_g \mid \boldsymbol{y}, \tau^2, \gamma_g^2, \lambda_{g1}^2, \ldots, \lambda_{gp_g}^2, \sigma^2].$$

If $\boldsymbol{X}_g^\top \boldsymbol{X}_{g'} = \boldsymbol{0}$ for all $g \neq g'$, then we have

$$E[\boldsymbol{\beta}_g \mid \cdot] = \left(\boldsymbol{I}_{p_g} + (\boldsymbol{X}_g^\top \boldsymbol{X}_g)^{-1} \frac{\sigma^2 \gamma_g^2}{\tau^2} \boldsymbol{\Lambda}_g^{-1}\right)^{-1} \hat{\boldsymbol{\beta}}_g^{OLS}, \;\; \hat{\boldsymbol{\beta}}_g^{OLS} = (\boldsymbol{X}_g^\top \boldsymbol{X}_g)^{-1} \boldsymbol{X}_g^\top \boldsymbol{y},$$

where $\boldsymbol{\Lambda}_g = \text{diag}\big(\lambda_{g1}^2, \ldots, \lambda_{gp_g}^2\big)$. Following a similar argument to the proof of Theorem 3.3, we arrive at

$$\left\|\hat{\boldsymbol{\beta}}_g^{OLS} - E[\boldsymbol{\beta}_g \mid \cdot]\right\|_2^2 \geq \left(\frac{1}{1 + \theta_{max}(\boldsymbol{X}_g^\top \boldsymbol{X}_g)\sigma^{-2}\tau^2\gamma_g^2 \max_j \lambda_{gj}^2}\right)^2 \left\|\hat{\boldsymbol{\beta}}_g^{OLS}\right\|_2^2,$$

Also, from the proof of Theorem 3.3, we have that

$$\pi\left(\frac{1}{1 + c\sigma^{-2}\tau^2\gamma_g^2 \max_j \lambda_{gj}^2} < \epsilon \;\middle|\; \boldsymbol{y}, \mathcal{H}, \tau^2, \sigma^2\right) \leq \sum_{j=1}^{p_g} \pi\big(K_{gj} < \epsilon \mid \boldsymbol{y}, \mathcal{H}, \tau^2, \sigma^2\big)$$

$$\leq \frac{p_g \Gamma(a_g + b_g)\epsilon^{b_g}(1-\epsilon)^{-(a_g+b_g)}}{(2\pi\sigma^2)^{n/2}\Gamma(a_g)b_g\Gamma(b_g)\pi(\boldsymbol{y} \mid \mathcal{H}, \tau^2, \sigma^2)} \left(\frac{c\tau^2}{\sigma^2}\right)^{b_g} \max\{1, (1-\epsilon)^{a_g - 1}\},$$

If, $a_g \in (0, 1)$ and

$$\frac{c\epsilon\tau^2}{\sigma^2(1-\epsilon)} < 1$$

then

$$\frac{p_g \Gamma(a_g + b_g)\epsilon^{b_g}(1-\epsilon)^{-(a_g+b_g)}}{(2\pi\sigma^2)^{n/2}\Gamma(a_g)b_g\Gamma(b_g)\pi(\boldsymbol{y} \mid \mathcal{H}, \tau^2, \sigma^2)} \left(\frac{c\tau^2}{\sigma^2}\right)^{b_g} \max\{1, (1-\epsilon)^{a_g - 1}\} \to 0$$

as $b_g \to \infty$. Since,

$$\left\|\hat{\boldsymbol{\beta}}_g^{OLS} - E[\boldsymbol{\beta}_g \mid \cdot]\right\|_2^2 \geq \left(\frac{1}{1 + \theta_{max}(\boldsymbol{X}_g^\top \boldsymbol{X}_g)\sigma^{-2}\tau^2\gamma_g^2 \max_j \lambda_{gj}^2}\right)^2 \left\|\hat{\boldsymbol{\beta}}_g^{OLS}\right\|_2^2,$$

then we must also have that for all $\delta \in (0, \sigma^2/(\sigma^2 + \theta_{max}(\boldsymbol{X}_g^\top \boldsymbol{X}_g)\tau^2))$

$$\pi\left(\left\|\hat{\boldsymbol{\beta}}_g^{OLS} - E[\boldsymbol{\beta}_g \mid \cdot]\right\|_2 \geq \delta \left\|\hat{\boldsymbol{\beta}}_g^{OLS}\right\|_2 \;\middle|\; \boldsymbol{y}, \mathcal{H}, \tau^2, \sigma^2\right) \to 1$$

as $b_g \to \infty$.

A.6 Proof of Theorem 2.3.3a

The posterior distribution of the shrinkage weights in the g-th group are given by

$$\pi\big(\boldsymbol{\kappa}_g \mid \boldsymbol{y}_g, \tau^2, \sigma^2, a_g, b_g\big) \propto \left(1+\frac{\tau^2}{\sigma^2}\sum_{j=1}^{p_g}\frac{\kappa_{gj}}{1-\kappa_{gj}}\right)^{-(a_g+p_g b_g)} \left(\prod_{j=1}^{p_g}\kappa_{gj}^{b_g-1/2}(1-\kappa_{gj})^{-(b_g+1)}\exp\left(-\frac{y_{gj}^2}{2\sigma^2}\kappa_{gj}\right)\right),$$

where $\boldsymbol{\kappa}_g = (\kappa_{g1}, ..., \kappa_{gp_g})$, $0 < \kappa_{gj} < 1$ for all $1 \leq j \leq p_g$, and $\boldsymbol{y}_g = (y_{g1}, ..., y_{gp_g})^\top$.

$$\pi(\kappa_{gl} > \psi \mid \boldsymbol{y}_g, \tau^2, \sigma^2, a_g, b_g) = \frac{A_{gl}}{B_g},$$

where

$$A_{gl} = \int_\psi^1\int_0^1\cdots\int_0^1 \left(1+\frac{\tau^2}{\sigma^2}\sum_{j=1}^{p_g}\frac{\kappa_{gj}}{1-\kappa_{gj}}\right)^{-(a_g+p_g b_g)}$$

$$\times\left(\prod_{j=1}^{p_g}\kappa_{gj}^{b_g-1/2}(1-\kappa_{gj})^{-(b_g+1)}\exp\left(-\frac{y_{gj}^2}{2\sigma^2}\kappa_{gj}\right)\right)d\kappa_{g1}\cdots d\kappa_{g,l-1}d\kappa_{g,l+1}\cdots d\kappa_{gp_g}d\kappa_{gl}$$

and

$$B_g = \int_0^1\cdots\int_0^1 \left(1+\frac{\tau^2}{\sigma^2}\sum_{j=1}^{p_g}\frac{\kappa_{gj}}{1-\kappa_{gj}}\right)^{-(a_g+p_g b_g)} \left(\prod_{j=1}^{p_g}\kappa_{gj}^{b_g-1/2}(1-\kappa_{gj})^{-(b_g+1)}\exp\left(-\frac{y_{gj}^2}{2\sigma^2}\kappa_{gj}\right)\right)d\kappa_{g1}\cdots d\kappa_{gp_g}.$$

Note that

$$\left(1+\frac{\tau^2}{\sigma^2}\sum_{j=1}^{p_g}\frac{\kappa_{gj}}{1-\kappa_{gj}}\right)^{-(a_g+p_g b_g)}$$

$$= \left(1+\frac{\tau^2}{\sigma^2}\sum_{j=1}^{p_g}\frac{\kappa_{gj}}{1-\kappa_{gj}}\right)^{-(a_g/p_g+b_g)}\left(1+\frac{\tau^2}{\sigma^2}\sum_{j=1}^{p_g}\frac{\kappa_{gj}}{1-\kappa_{gj}}\right)^{-(p_g-1)(a_g/p_g+b_g)}$$

$$\leq \left(1+\frac{\tau^2}{\sigma^2}\frac{\kappa_{gl}}{1-\kappa_{gl}}\right)^{-(a_g/p_g+b_g)}\left(1+\frac{\tau^2}{\sigma^2}\sum_{j\neq l}\frac{\kappa_{gj}}{1-\kappa_{gj}}\right)^{-(p_g-1)(a_g/p_g+b_g)}$$

Then,

$$A_{gl} \leq \left(\int_0^1\cdots\int_0^1 \left(1+\frac{\tau^2}{\sigma^2}\sum_{j\neq l}\frac{\kappa_{gj}}{1-\kappa_{gj}}\right)^{-(p_g-1)(a_g/p_g+b_g)}\prod_{j\neq l}\kappa_{gj}^{b_g-1/2}(1-\kappa_{gj})^{-(b_g+1)}\exp\left(-\frac{y_{gj}^2}{2\sigma^2}\kappa_{gj}\right)d\kappa_{gj}\right)\times$$

$$\left(\int_\psi^1\left(1+\frac{\tau^2}{\sigma^2}\frac{\kappa_{gl}}{1-\kappa_{gl}}\right)^{-(a_g/p_g+b_g)}\kappa_{gl}^{b_g-1/2}(1-\kappa_{gl})^{-(b_g+1)}\exp\left(-\frac{y_{gl}^2}{2\sigma^2}\kappa_{gl}\right)d\kappa_{gl}\right)$$

$$\leq \left(\int_0^1\cdots\int_0^1\left(1+\frac{\tau^2}{\sigma^2}\sum_{j\neq l}\frac{\kappa_{gj}}{1-\kappa_{gj}}\right)^{-(p_g-1)(a_g/p_g+b_g)}\prod_{j\neq l}\kappa_{gj}^{b_g-1/2}(1-\kappa_{gj})^{-(b_g+1)}d\kappa_{gj}\right)\times$$

$$\exp\left(-\frac{\psi}{2\sigma^2}y_{gl}^2\right)\left(\int_\psi^1\left(1+\frac{\tau^2}{\sigma^2}\frac{\kappa_{gl}}{1-\kappa_{gl}}\right)^{-(a_g/p_g+b_g)}\kappa_{gl}^{b_g-1/2}(1-\kappa_{gl})^{-(b_g+1)}d\kappa_{gl}\right)$$

$$\leq \left(\int_0^1 \cdots \int_0^1 \left(1 + \frac{\tau^2}{\sigma^2} \sum_{j \neq l} \frac{\kappa_{gj}}{1-\kappa_{gj}} \right)^{-(a_g^* + (p_g-1)b_g)} \prod_{j \neq l} \kappa_{gj}^{b_g - 1} (1-\kappa_{gj})^{-(b_g+1)} d\kappa_{gj} \right) \times$$

$$\exp\left(-\frac{\psi}{2\sigma^2} y_{gl}^2 \right) \left(\int_\psi^1 \left(1 - \left(1 - \frac{\tau^2}{\sigma^2} \right) \kappa_{gl} \right)^{-(a_g/p_g + b_g)} \kappa_{gl}^{b_g - 1/2} (1-\kappa_{gl})^{a_g/p_g - 1} d\kappa_{gl} \right)$$

$$\leq \left(\left(\frac{\tau^2}{\sigma^2} \right)^{-(p_g-1)b_g} \frac{\Gamma(a_g^*)(\Gamma(b_g))^{p_g-1}}{\Gamma(a_g^* + (p_g-1)b_g)} \right) \left(\min\left(1, \frac{\tau^2}{\sigma^2}\right) \right)^{-(a_g/p_g + b_g)}$$

$$\times \exp\left(-\frac{\psi}{2\sigma^2} y_{gl}^2 \right) \int_\psi^1 \kappa_{gl}^{b_g - 1/2} (1-\kappa_{gl})^{a_g/p_g - 1} d\kappa_{gl}$$

$$\leq \left(\left(\frac{\tau^2}{\sigma^2} \right)^{-(p_g-1)b_g} \frac{\Gamma(a_g^*)(\Gamma(b_g))^{p_g-1}}{\Gamma(a_g^* + (p_g-1)b_g)} \right) \left(\min\left(1, \frac{\tau^2}{\sigma^2}\right) \right)^{-(a_g/p_g + b_g)} \max\left(1, \psi^{b_g - 1/2}\right)$$

$$\times \exp\left(-\frac{\psi}{2\sigma^2} y_{gl}^2 \right) \int_\psi^1 (1-\kappa_{gl})^{a_g/p_g - 1} d\kappa_{gl}$$

$$= \left(\left(\frac{\tau^2}{\sigma^2} \right)^{-(p_g-1)b_g} \frac{\Gamma(a_g^*)(\Gamma(b_g))^{p_g-1}}{\Gamma(a_g^* + (p_g-1)b_g)} \right) \left(\min\left(1, \frac{\tau^2}{\sigma^2}\right) \right)^{-(a_g/p_g + b_g)} \max\left(1, \psi^{b_g - 1/2}\right)$$

$$\times \exp\left(-\frac{\psi}{2\sigma^2} y_{gl}^2 \right) \frac{p_g}{a_g} (1-\psi)^{a_g/p_g},$$

where $a_g^* = (p_g - 1)a_g/p_g$. We can simplify the integral four lines above based on the prior distribution of the shrinkage weights for a group of size $p_g - 1$.

Next, let $\delta \in (0, 1)$ be fixed constant. Then,

$$B_g \geq \int_0^{\psi\delta} \cdots \int_0^{\psi\delta} \left(1 + \frac{\tau^2}{\sigma^2} \sum_{j=1}^{p_g} \frac{\kappa_{gj}}{1-\kappa_{gj}} \right)^{-(a_g + p_g b_g)} \left(\prod_{j=1}^{p_g} \kappa_{gj}^{b_g - 1/2} (1-\kappa_{gj})^{-(b_g+1)} \exp\left(-\frac{y_{gj}^2}{2\sigma^2} \kappa_{gj} \right) d\kappa_{gj} \right)$$

$$\geq \left(1 + \frac{p_g \tau^2}{\sigma^2} \frac{\psi\delta}{1-\psi\delta} \right)^{-(a_g + p_g b_g)} \prod_{j=1}^{p_g} \int_0^{\psi\delta} \kappa_{gj}^{b_g - 1/2} (1-\kappa_{gj})^{-(b_g+1)} \exp\left(-\frac{y_{gj}^2}{2\sigma^2} \kappa_{gj} \right) d\kappa_{gj}$$

$$\geq \left(1 + \frac{p_g \tau^2}{\sigma^2} \frac{\psi\delta}{1-\psi\delta} \right)^{-(a_g + p_g b_g)} \exp\left(-\frac{\psi\delta}{2\sigma^2} \sum_{j=1}^{p_g} y_{gj}^2 \right) \prod_{j=1}^{p_g} \int_0^{\psi\delta} \kappa_{gj}^{b_g - 1/2} (1-\kappa_{gj})^{-(b_g+1)} d\kappa_{gj}$$

$$\geq \left(1 + \frac{p_g \tau^2}{\sigma^2} \frac{\psi\delta}{1-\psi\delta} \right)^{-(a_g + p_g b_g)} \exp\left(-\frac{\psi\delta}{2\sigma^2} \sum_{j=1}^{p_g} y_{gj}^2 \right) \prod_{j=1}^{p_g} \int_0^{\psi\delta} \kappa_{gj}^{b_g - 1/2} d\kappa_{gj}$$

$$= \left(1 + \frac{p_g \tau^2}{\sigma^2} \frac{\psi\delta}{1-\psi\delta} \right)^{-(a_g + p_g b_g)} \exp\left(-\frac{\psi\delta}{2\sigma^2} \sum_{j=1}^{p_g} y_{gj}^2 \right) (b_g + 1/2)^{-p_g} (\psi\delta)^{p_g(b_g + 1/2)}.$$

Therefore,

$$\frac{A_{gl}}{B_g} \leq \frac{f(p_g, \tau^2, \sigma^2, a_g, b_g, \psi)}{g(p_g, \tau^2, \sigma^2, a_g, b_g, \psi, \delta)} \exp\left(\frac{\psi\delta}{2\sigma^2} \sum_{j \neq l} y_{gj}^2 \right) \exp\left(-\frac{\psi(1-\delta)}{2\sigma^2} y_{gl}^2 \right),$$

where

$$f(p_g, \tau^2, \sigma^2, a_g, b_g, \psi) = \left(\left(\frac{\tau^2}{\sigma^2} \right)^{-(p_g - 1)b_g} \frac{\Gamma(a_g^*)(\Gamma(b_g))^{p_g - 1}}{\Gamma(a_g^* + (p_g - 1)b_g)} \right)$$

$$\times \left(\min \left(1, \frac{\tau^2}{\sigma^2} \right) \right)^{-(a_g/p_g + b_g)} \max \left(1, \psi^{b_g - 1/2} \right) \frac{p_g}{a_g} (1 - \psi)^{a_g/p_g}$$

and

$$g(p_g, \tau^2, \sigma^2, a_g, b_g, \psi, \delta) = \left(1 + \frac{p_g \tau^2}{\sigma^2} \frac{\psi\delta}{1 - \psi\delta} \right)^{-(a_g + p_g b_g)} (b_g + 1/2)^{-p_g} (\psi\delta)^{p_g(b_g + 1/2)}.$$

If we take the limit of this upper bound as $|y_{gl}| \to \infty$, then we see that $\pi(\kappa_{gl} > \psi \mid \boldsymbol{y}_g, \tau^2, \sigma^2, a_g, b_g) \to 0$. This concludes the proof.

A.7 Proof of Theorem 2.3.3b

The posterior distribution of the shrinkage weights in the g-th group are given by

$$\pi\big(\boldsymbol{\kappa}_g \mid \boldsymbol{y}_g, \tau^2, \sigma^2, a_g, b_g\big) \propto \left(1+\frac{\tau^2}{\sigma^2}\sum_{j=1}^{p_g}\frac{\kappa_{gj}}{1-\kappa_{gj}}\right)^{-(a_g+p_g b_g)} \left(\prod_{j=1}^{p_g}\kappa_{gj}^{b_g-1/2}(1-\kappa_{gj})^{-(b_g+1)}\exp\left(-\frac{y_{gj}^2}{2\sigma^2}\kappa_{gj}\right)\right),$$

where $\boldsymbol{\kappa}_g = (\kappa_{g1}, ..., \kappa_{gp_g})$, $0 < \kappa_{gj} < 1$ for all $1 \leq j \leq p_g$, and $\boldsymbol{y}_g = (y_{g1}, ..., y_{gp_g})^\top$.

$$\pi(\kappa_{gl} < \epsilon \mid \boldsymbol{y}_g, \tau^2, \sigma^2, a_g, b_g) = \frac{A_{gl}}{B_g},$$

where

$$A_{gl} = \int_0^\epsilon \int_0^1 \cdots \int_0^1 \left(1+\frac{\tau^2}{\sigma^2}\sum_{j=1}^{p_g}\frac{\kappa_{gj}}{1-\kappa_{gj}}\right)^{-(a_g+p_g b_g)}$$

$$\times\left(\prod_{j=1}^{p_g}\kappa_{gj}^{b_g-1/2}(1-\kappa_{gj})^{-(b_g+1)}\exp\left(-\frac{y_{gj}^2}{2\sigma^2}\kappa_{gj}\right)\right)d\kappa_{g1}\cdots d\kappa_{g,l-1}d\kappa_{g,l+1}\cdots d\kappa_{gp_g}d\kappa_{gl}$$

and

$$B_g = \int_0^1 \cdots \int_0^1 \left(1+\frac{\tau^2}{\sigma^2}\sum_{j=1}^{p_g}\frac{\kappa_{gj}}{1-\kappa_{gj}}\right)^{-(a_g+p_g b_g)} \left(\prod_{j=1}^{p_g}\kappa_{gj}^{b_g-1/2}(1-\kappa_{gj})^{-(b_g+1)}\exp\left(-\frac{y_{gj}^2}{2\sigma^2}\kappa_{gj}\right)\right)d\kappa_{g1}\cdots d\kappa_{gp_g}.$$

Note that

$$\left(1+\frac{\tau^2}{\sigma^2}\sum_{j=1}^{p_g}\frac{\kappa_{gj}}{1-\kappa_{gj}}\right)^{-(a_g+p_g b_g)}$$

$$= \left(1+\frac{\tau^2}{\sigma^2}\sum_{j=1}^{p_g}\frac{\kappa_{gj}}{1-\kappa_{gj}}\right)^{-(a_g/p_g+b_g)} \left(1+\frac{\tau^2}{\sigma^2}\sum_{j=1}^{p_g}\frac{\kappa_{gj}}{1-\kappa_{gj}}\right)^{-(p_g-1)(a_g/p_g+b_g)}$$

$$\leq \left(1+\frac{\tau^2}{\sigma^2}\frac{\kappa_{gl}}{1-\kappa_{gl}}\right)^{-(a_g/p_g+b_g)} \left(1+\frac{\tau^2}{\sigma^2}\sum_{j\neq l}\frac{\kappa_{gj}}{1-\kappa_{gj}}\right)^{-(p_g-1)(a_g/p_g+b_g)}$$

Then,

$$A_{gl} \leq \left(\int_0^1 \cdots \int_0^1 \left(1+\frac{\tau^2}{\sigma^2}\sum_{j\neq l}\frac{\kappa_{gj}}{1-\kappa_{gj}}\right)^{-(p_g-1)(a_g/p_g+b_g)} \prod_{j\neq l}\kappa_{gj}^{b_g-1/2}(1-\kappa_{gj})^{-(b_g+1)}\exp\left(-\frac{y_{gj}^2}{2\sigma^2}\kappa_{gj}\right)d\kappa_{gj}\right)\times$$

$$\left(\int_0^\epsilon \left(1+\frac{\tau^2}{\sigma^2}\frac{\kappa_{gl}}{1-\kappa_{gl}}\right)^{-(a_g/p_g+b_g)} \kappa_{gl}^{b_g-1/2}(1-\kappa_{gl})^{-(b_g+1)}\exp\left(-\frac{y_{gl}^2}{2\sigma^2}\kappa_{gl}\right)d\kappa_{gl}\right)$$

$$\leq \left(\int_0^1 \cdots \int_0^1 \left(1+\frac{\tau^2}{\sigma^2}\sum_{j\neq l}\frac{\kappa_{gj}}{1-\kappa_{gj}}\right)^{-(p_g-1)(a_g/p_g+b_g)} \prod_{j\neq l}\kappa_{gj}^{b_g-1/2}(1-\kappa_{gj})^{-(b_g+1)}d\kappa_{gj}\right)\times$$

$$\left((1-\epsilon)^{-(b_g+1)}\int_0^\epsilon \kappa_{gl}^{b_g-1/2}d\kappa_{gl}\right)$$

$$\leq \frac{\epsilon^{b_g+1/2}}{(b_g+1/2)(1-\epsilon)^{b_g+1}} \int_0^1 \cdots \int_0^1 \left(1+\frac{\tau^2}{\sigma^2}\sum_{j\neq l}\frac{\kappa_{gj}}{1-\kappa_{gj}}\right)^{-(a_g^*+(p_g-1)b_g)} \prod_{j\neq l} \kappa_{gj}^{b_g-1}\kappa_{gj}^{1/2}(1-\kappa_{gj})^{-(b_g+1)} d\kappa_{gj}$$

$$\leq \frac{\epsilon^{b_g+1/2}}{(b_g+1/2)(1-\epsilon)^{b_g+1}} \int_0^1 \cdots \int_0^1 \left(1+\frac{\tau^2}{\sigma^2}\sum_{j\neq l}\frac{\kappa_{gj}}{1-\kappa_{gj}}\right)^{-(a_g^*+(p_g-1)b_g)} \prod_{j\neq l} \kappa_{gj}^{b_g-1}(1-\kappa_{gj})^{-(b_g+1)} d\kappa_{gj}$$

$$= \left(\frac{\epsilon^{b_g+1/2}}{(b_g+1/2)(1-\epsilon)^{b_g+1}}\right)\left(\left(\frac{\tau^2}{\sigma^2}\right)^{-(p_g-1)b_g} \frac{\Gamma(a_g^*)(\Gamma(b_g))^{p_g-1}}{\Gamma(a_g^*+(p_g-1)b_g)}\right)$$

where $a_g^* = (p_g-1)a_g/p_g$. We have the last equality based on the prior distribution of the shrinkage weights for a group of size $p_g - 1$. Next,

$$B_g \geq \exp\left(-\frac{1}{2\sigma^2}\sum_{j=1}^{p_g} y_{gj}^2\right) \int_0^1 \cdots \int_0^1 \left(1+\frac{\tau^2}{\sigma^2}\sum_{j=1}^{p_g}\frac{\kappa_{gj}}{1-\kappa_{gj}}\right)^{-(a_g+p_g b_g)} \prod_{j=1}^{p_g} \kappa_{gj}^{b_g-1/2}(1-\kappa_{gj})^{-(b_g+1)} d\kappa_{gj}$$

$$= \exp\left(-\frac{1}{2\sigma^2}\sum_{j=1}^{p_g} y_{gj}^2\right) \int_0^1 \cdots \int_0^1 \left(1+\frac{\tau^2}{\sigma^2}\sum_{j=1}^{p_g}\frac{\kappa_{gj}}{1-\kappa_{gj}}\right)^{-(a_g+p_g b_g^*)} \left(1+\frac{\tau^2}{\sigma^2}\sum_{j=1}^{p_g}\frac{\kappa_{gj}}{1-\kappa_{gj}}\right)^{p_g/2}$$

$$\times \prod_{j=1}^{p_g} \kappa_{gj}^{b_g^*-1}(1-\kappa_{gj})^{-(b_g^*+1)}(1-\kappa_{gj})^{1/2} d\kappa_{gj}$$

$$\geq \exp\left(-\frac{1}{2\sigma^2}\sum_{j=1}^{p_g} y_{gj}^2\right) \int_0^1 \cdots \int_0^1 \left(1+\frac{\tau^2}{\sigma^2}\sum_{j=1}^{p_g}\frac{\kappa_{gj}}{1-\kappa_{gj}}\right)^{p_g/2} \left(\prod_{j=1}^{p_g}(1-\kappa_{gj})^{1/2}\right) \left(1+\frac{\tau^2}{\sigma^2}\sum_{j=1}^{p_g}\frac{\kappa_{gj}}{1-\kappa_{gj}}\right)^{-(a_g+p_g b_g^*)}$$

$$\times \prod_{j=1}^{p_g} \kappa_{gj}^{b_g^*-1}(1-\kappa_{gj})^{-(b_g^*+1)} d\kappa_{gj}$$

$$\geq \exp\left(-\frac{1}{2\sigma^2}\sum_{j=1}^{p_g} y_{gj}^2\right) \int_0^1 \cdots \int_0^1 \left(\prod_{j=1}^{p_g}\left(1-\kappa_{gj}+\frac{\tau^2}{\sigma^2}\kappa_{gj}\right)\right)^{1/2} \left(1+\frac{\tau^2}{\sigma^2}\sum_{j=1}^{p_g}\frac{\kappa_{gj}}{1-\kappa_{gj}}\right)^{-(a_g+p_g b_g^*)}$$

$$\times \prod_{j=1}^{p_g} \kappa_{gj}^{b_g^*-1}(1-\kappa_{gj})^{-(b_g^*+1)} d\kappa_{gj}$$

$$\geq \exp\left(-\frac{1}{2\sigma^2}\sum_{j=1}^{p_g} y_{gj}^2\right)\left(\min\left(1,\frac{\tau^2}{\sigma^2}\right)\right)^{p_g/2} \int_0^1 \cdots \int_0^1 \left(1+\frac{\tau^2}{\sigma^2}\sum_{j=1}^{p_g}\frac{\kappa_{gj}}{1-\kappa_{gj}}\right)^{-(a_g+p_g b_g^*)}$$

$$\times \prod_{j=1}^{p_g} \kappa_{gj}^{b_g^*-1}(1-\kappa_{gj})^{-(b_g^*+1)} d\kappa_{gj}$$

$$= \exp\left(-\frac{1}{2\sigma^2}\sum_{j=1}^{p_g} y_{gj}^2 \right)\left(\min\left(1, \frac{\tau^2}{\sigma^2}\right)\right)^{p_g/2}\left(\left(\frac{\tau^2}{\sigma^2}\right)^{-p_g b_g^*} \frac{\Gamma(a_g)(\Gamma(b_g^*))^{p_g}}{\Gamma(a_g + p_g b_g^*)} \right),$$

where $b_g^* = b_g + 1/2$. Similarly, we have the last equality based on the prior distribution of the shrinkage weights for a group of size p_g. Therefore,

$$\frac{A_{gl}}{B_g} \leq \exp\left(\frac{1}{2\sigma^2}\sum_{j=1}^{p_g} y_{gj}^2 \right) \frac{\epsilon^{b_g+1/2}}{(b_g+1/2)(1-\epsilon)^{b_g+1}} \left(\frac{\tau^2}{\sigma^2}\right)^{p_g/2+b_g} \left(\min\left(1, \frac{\tau^2}{\sigma^2}\right)\right)^{-p_g/2} \times$$

$$\frac{\Gamma(a_g + p_g b_g^*)\Gamma(a_g^*)(\Gamma(b_g))^{p_g-1}}{\Gamma(a_g^* + (p_g-1)b_g)\Gamma(a_g)(\Gamma(b_g^*))^{p_g}}.$$

If we take the limit of this expression as $\tau \to 0$, then we see that $\pi(\kappa_{gl} < \epsilon \mid \boldsymbol{y}_g, \tau^2, \sigma^2, a_g, b_g) \to 0$. This concludes the proof.

A.8 Full Conditional Distributions for Gibbs Sampler

The full conditional distributions for all model parameters are

$$[\boldsymbol{\alpha} \mid \cdot] \sim N\Big(\big(\boldsymbol{C}^\top \boldsymbol{C}\big)^{-1} \boldsymbol{C}^\top (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}), \sigma^2 \big(\boldsymbol{C}^\top \boldsymbol{C}\big)^{-1} \Big)$$

$$[\boldsymbol{\beta} \mid \cdot] \sim N\left(\boldsymbol{Q}^{-1} \frac{1}{\sigma^2} \boldsymbol{X}^\top \Big(\boldsymbol{y} - \boldsymbol{C}\boldsymbol{\alpha}\Big), \boldsymbol{Q}^{-1} \right), \;\; \boldsymbol{Q} = \frac{1}{\sigma^2} \boldsymbol{X}^\top \boldsymbol{X} + \frac{1}{\tau^2} \boldsymbol{\Gamma}^{-1} \boldsymbol{\Lambda}^{-1}$$

$$[\tau^2 \mid \cdot] \sim IG\left(\frac{p+1}{2}, \frac{1}{2} \boldsymbol{\beta}^\top \boldsymbol{\Gamma}^{-1} \boldsymbol{\Lambda}^{-1} \boldsymbol{\beta} + \frac{1}{\nu} \right), \;\; [\nu \mid \cdot] \sim IG\left(1, \frac{1}{\tau^2} + \frac{1}{\sigma^2} \right)$$

$$[\sigma^2 \mid \cdot] \sim IG\left(\frac{n+1}{2}, \frac{1}{2} \big(\boldsymbol{y} - \boldsymbol{C}\boldsymbol{\alpha} - \boldsymbol{X}\boldsymbol{\beta}\big)^\top \big(\boldsymbol{y} - \boldsymbol{C}\boldsymbol{\alpha} - \boldsymbol{X}\boldsymbol{\beta}\big) + \frac{1}{\nu} \right)$$

$$[\lambda_{gj}^2 \mid \cdot] \sim IG\left(b_g + \frac{1}{2}, 1 + \frac{\beta_{gj}^2}{2\tau^2 \gamma_g^2} \right), \;\; [\gamma_g^{-2} \mid \cdot] \sim GIG\left(\frac{p_g}{2} - a_g, \frac{1}{\tau^2} \sum_{j=1}^{p_g} \frac{\beta_{gj}^2}{\lambda_{gj}^2}, 2 \right),$$

where GIG refers to the generalized inverse Gaussian distribution (Hörmann and Leydold, 2014). See the Distributional Definition section on the first page of this supplement for details on the probability density function parameterizations.

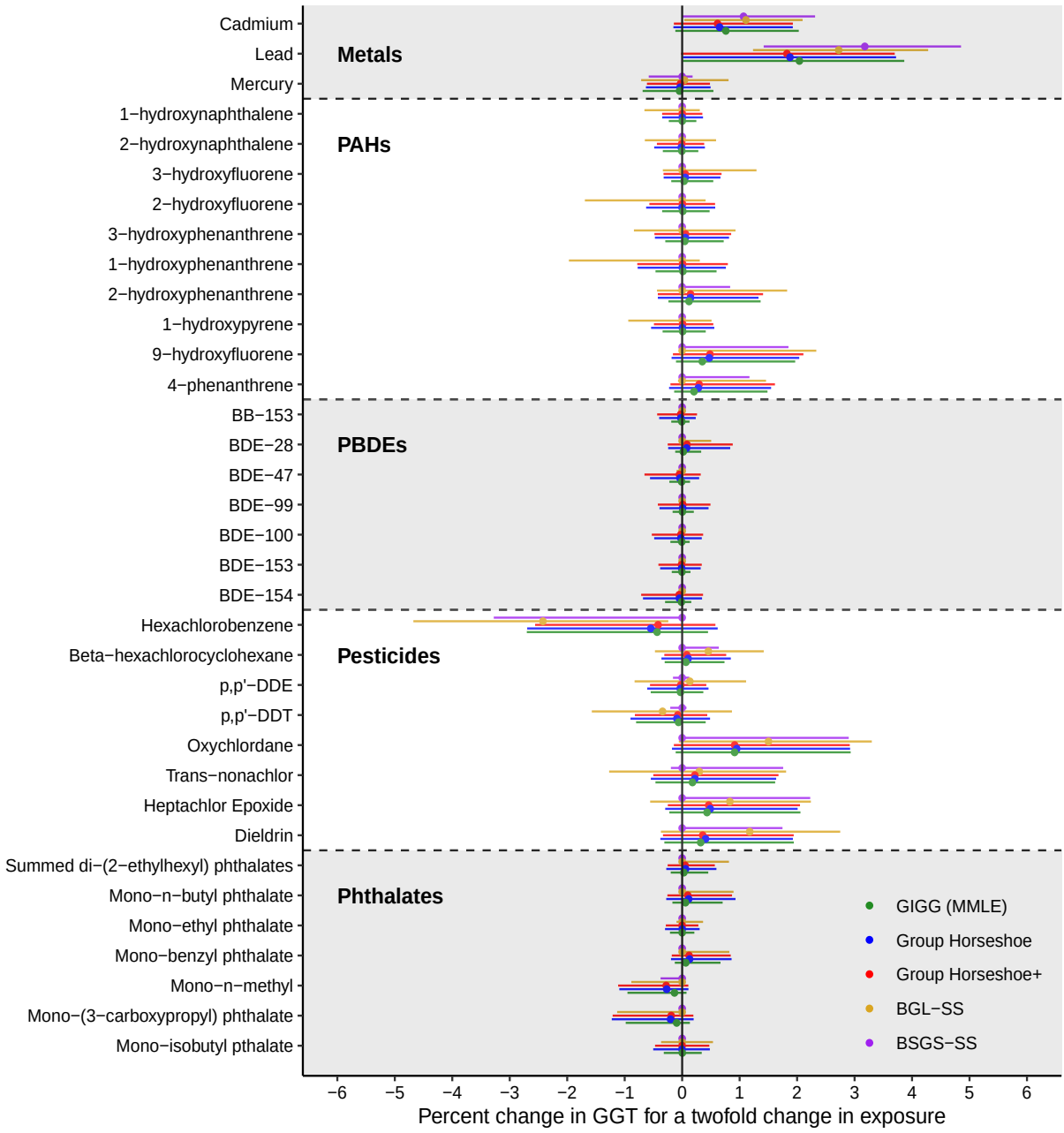


Figure A.1: Associations between environmental toxicants (metals, phthalates, pesticides, PBDEs, and PAHs) and gamma glutamyl transferase (GGT) from NHANES 2003-2004 ($n = 990$).

$\rho = 0.6$	Concentrated			Distributed		
Method	Null	Non-Null	Overall	Null	Non-Null	Overall
Ordinary Least Squares	1.88	0.21	2.09	3.20	0.78	3.99
Horseshoe	0.29	0.21	0.50	0.52	0.93	1.45
GIGG ($a_g = 1/n, b_g = 1/n$)	0.05	0.19	**0.24**	0.04	1.52	1.56
GIGG ($a_g = 1/2, b_g = 1/n$)	0.05	0.20	**0.26**	0.04	1.50	1.53
GIGG ($a_g = 1/n, b_g = 1/2$)	0.15	0.22	0.37	0.03	0.69	**0.71**
*GIGG ($a_g = 1/2, b_g = 1/2$)	0.18	0.21	0.39	0.16	0.73	0.88
GIGG ($a_g = 1/n, b_g = 1$)	0.29	0.26	0.56	0.03	0.65	**0.67**
GIGG ($a_g = 1/2, b_g = 1$)	0.33	0.25	0.57	0.16	0.65	0.82
GIGG (MMLE)	0.11	0.21	0.32	0.03	0.62	**0.65**
Group Horseshoe+	0.17	0.21	0.37	0.07	0.71	0.77
Spike-and-Slab Lasso	0.03	0.25	**0.28**	0.01	2.18	2.19
BGL-SS	1.25	0.42	1.67	0.01	0.61	**0.62**
BSGS-SS	0.10	0.22	**0.32**	0.02	0.81	0.82

Table A.1: Mean-squared errors (MSE) for simulation settings C10M and D10M in Table 1 ($n = 500, p = 50$) with medium pairwise correlations ($\rho = 0.6$). Bolded cells indicate the four methods with the lowest overall MSE. *GIGG regression with $a_g = 1/2$ and $b_g = 1/2$ is equivalent to group horseshoe regression.

$\rho = 0.8$	Concentrated			Distributed		
Method	Null	Non-Null	Overall	Null	Non-Null	Overall
Ordinary Least Squares	5.68	0.63	6.31	8.62	2.14	10.76
Horseshoe	0.83	0.75	1.58	1.45	2.03	3.48
GIGG ($a_g = 1/n, b_g = 1/n$)	0.18	0.77	**0.95**	0.59	3.42	4.01
GIGG ($a_g = 1/2, b_g = 1/n$)	0.19	0.77	**0.95**	0.59	3.37	3.97
GIGG ($a_g = 1/n, b_g = 1/2$)	0.44	0.79	**1.23**	0.85	1.88	2.73
*GIGG ($a_g = 1/2, b_g = 1/2$)	0.53	0.78	1.31	0.93	1.88	2.80
GIGG ($a_g = 1/n, b_g = 1$)	0.80	0.86	1.66	1.04	1.58	**2.62**
GIGG ($a_g = 1/2, b_g = 1$)	0.89	0.85	1.74	1.12	1.58	**2.70**
GIGG (MMLE)	0.42	0.79	**1.21**	1.05	1.63	**2.68**
Group Horseshoe+	0.50	0.78	1.28	0.93	1.97	2.90
Spike-and-Slab Lasso	0.63	0.79	1.42	1.22	4.06	5.28
BGL-SS	3.07	1.00	4.07	1.76	1.28	3.04
BSGS-SS	0.49	0.79	1.27	0.67	2.02	**2.70**

Table A.2: Mean-squared errors (MSE) for simulation settings C25 and D25 in Table 1 ($n = 500, p = 50$) with high pairwise correlations ($\rho = 0.8$). Bolded cells indicate the four methods with the lowest overall MSE. Four methods are highlighted to emphasize that GIGG (MMLE) is the best method with respect to MSE for both concentrated and distributed signals aside from methods that only perform well for one of the two settings. *GIGG regression with $a_g = 1/2$ and $b_g = 1/2$ is equivalent to group horseshoe regression.

$\rho = 0.8$	Concentrated			Distributed		
Method	Null	Non-Null	Overall	Null	Non-Null	Overall
Ordinary Least Squares	3.37	0.37	3.73	4.91	1.23	6.14
Horseshoe	0.36	0.29	0.65	0.46	1.50	1.96
GIGG ($a_g = 1/n, b_g = 1/n$)	0.08	0.23	**0.30**	0.03	2.63	2.66
GIGG ($a_g = 1/2, b_g = 1/n$)	0.08	0.25	**0.33**	0.03	2.55	2.58
GIGG ($a_g = 1/n, b_g = 1/2$)	0.26	0.30	0.57	0.05	1.04	**1.09**
*GIGG ($a_g = 1/2, b_g = 1/2$)	0.23	0.29	0.52	0.21	1.23	1.44
GIGG ($a_g = 1/n, b_g = 1$)	0.45	0.38	0.83	0.05	0.91	**0.96**
GIGG ($a_g = 1/2, b_g = 1$)	0.44	0.36	0.80	0.27	0.96	1.22
GIGG (MMLE)	0.18	0.27	**0.44**	0.05	0.88	**0.93**
Group Horseshoe+	0.22	0.28	0.50	0.08	1.16	1.24
Spike-and-Slab Lasso	0.08	0.28	**0.37**	0.04	3.40	3.44
BGL-SS	1.01	0.50	1.51	0.05	0.80	**0.85**
BSGS-SS	0.18	0.32	0.50	0.04	1.23	1.27

Table A.3: Mean-squared errors (MSE) for simulation settings C5 and D5 in Table 1 ($n = 500, p = 50$) with high pairwise correlations ($\rho = 0.8$). Bolded cells indicate the four methods with the lowest overall MSE. Four methods are highlighted to emphasize that GIGG (MMLE) is the best method with respect to MSE for both concentrated and distributed signals aside from methods that only perform well for one of the two settings. *GIGG regression with $a_g = 1/2$ and $b_g = 1/2$ is equivalent to group horseshoe regression.

APPENDIX B

Supplement for Chapter 3

B.1 Optimization Algorithms

Unconstrained (Closed-Form Optimization)

Suppose that $\boldsymbol{M}_{i\cdot}^\top \sim N(A_i\boldsymbol{\alpha}_a + \boldsymbol{\alpha}_c\boldsymbol{C}_{i\cdot}^\top, \boldsymbol{\Sigma}_m)$ and $Y_i \sim N(\boldsymbol{M}_{i\cdot}\boldsymbol{\beta}_m + A_i\beta_a + \boldsymbol{C}_{i\cdot}\boldsymbol{\beta}_c, \sigma_e^2)$ for $i = 1, \ldots, n$. Let $\boldsymbol{Z} = (\boldsymbol{A}, \boldsymbol{C}, \boldsymbol{M})$. Then,

$$\widehat{\boldsymbol{\beta}} = \left(\boldsymbol{Z}^\top\boldsymbol{Z}\right)^{-1}\boldsymbol{Z}^\top\boldsymbol{Y}$$

$$\widehat{\sigma}_e^2 = \frac{1}{n}(\boldsymbol{Y} - \boldsymbol{Z}\widehat{\boldsymbol{\beta}})^\top(\boldsymbol{Y} - \boldsymbol{Z}\widehat{\boldsymbol{\beta}})$$

Let

$$\boldsymbol{X} = (\boldsymbol{A}, \boldsymbol{C}), \;\; \boldsymbol{\alpha} = \begin{pmatrix}\boldsymbol{\alpha}_a^\top \\ \boldsymbol{\alpha}_c^\top\end{pmatrix}.$$

Then,

$$\widehat{\boldsymbol{\alpha}} = \left(\boldsymbol{X}^\top\boldsymbol{X}\right)^{-1}\boldsymbol{X}^\top\boldsymbol{M}$$

$$\widehat{\boldsymbol{\Sigma}}_m = \frac{1}{n}\sum_{i=1}^{n}\left(\boldsymbol{M}_{i\cdot}^\top - \widehat{\boldsymbol{\alpha}}^\top\boldsymbol{X}_{i\cdot}^\top\right)\left(\boldsymbol{M}_{i\cdot}^\top - \widehat{\boldsymbol{\alpha}}^\top\boldsymbol{X}_{i\cdot}^\top\right)^\top$$

Hard Constraint (Cyclical Coordinate Descent)

Suppose that $\boldsymbol{M}_{i\cdot}^\top \sim N(A_i\boldsymbol{\alpha}_a + \boldsymbol{\alpha}_c\boldsymbol{C}_{i\cdot}^\top, \boldsymbol{\Sigma}_m)$ and $Y_i \sim N(\boldsymbol{M}_{i\cdot}\boldsymbol{\beta}_m + A_i(\widehat{\theta}_a^E - \boldsymbol{\alpha}_a^\top\boldsymbol{\beta}_m) + \boldsymbol{C}_{i\cdot}\boldsymbol{\beta}_c, \sigma_e^2)$ for $i = 1, \ldots, n$. Then the cyclical coordinate descent updates are as follows:

$$\widetilde{\boldsymbol{\alpha}}_c \leftarrow \left(\boldsymbol{C}^\top\boldsymbol{C}\right)^{-1}\boldsymbol{C}^\top\left(\boldsymbol{M} - \boldsymbol{A}\widetilde{\boldsymbol{\alpha}}_a^\top\right)$$

$$\widetilde{\boldsymbol{\alpha}}_a \leftarrow \frac{1}{\boldsymbol{A}^\top\boldsymbol{A}}\left(\widetilde{\boldsymbol{\Sigma}}_m^{-1} + \frac{1}{\widetilde{\sigma}_e^2}\widetilde{\boldsymbol{\beta}}_m\widetilde{\boldsymbol{\beta}}_m^\top\right)^{-1}\sum_{i=1}^{n}\left[A_i\widetilde{\boldsymbol{\Sigma}}_m^{-1}\left(\boldsymbol{M}_{i\cdot}^\top - \widetilde{\boldsymbol{\alpha}}_c^\top\boldsymbol{C}_{i\cdot}^\top\right) - \frac{1}{\widetilde{\sigma}_e^2}A_i\left(Y_i - \widehat{\theta}_a^E A_i - \boldsymbol{M}_{i\cdot}\widetilde{\boldsymbol{\beta}}_m - \boldsymbol{C}_{i\cdot}\widetilde{\boldsymbol{\beta}}_c\right)\widetilde{\boldsymbol{\beta}}_m\right]$$

$$\widetilde{\boldsymbol{\Sigma}}_m \leftarrow \frac{1}{n}\sum_{i=1}^{n}\left(\boldsymbol{M}_{i\cdot}^\top - \widetilde{\boldsymbol{\alpha}}^\top\boldsymbol{X}_{i\cdot}^\top\right)\left(\boldsymbol{M}_{i\cdot}^\top - \widetilde{\boldsymbol{\alpha}}^\top\boldsymbol{X}_{i\cdot}^\top\right)^\top, \;\; \boldsymbol{X} = (\boldsymbol{A}, \boldsymbol{C}), \;\; \boldsymbol{\alpha} = \begin{pmatrix}\boldsymbol{\alpha}_a^\top \\ \boldsymbol{\alpha}_c^\top\end{pmatrix}$$

$$\widetilde{\boldsymbol{\beta}}_c \leftarrow \left(\boldsymbol{C}^\top \boldsymbol{C}\right)^{-1} \boldsymbol{C}^\top \left(\boldsymbol{Y} - \{\widehat{\theta}_a^E - \widetilde{\boldsymbol{\alpha}}_a^\top \widetilde{\boldsymbol{\beta}}_m\}\boldsymbol{A} - \boldsymbol{M}\widetilde{\boldsymbol{\beta}}_m\right)$$

$$\widetilde{\boldsymbol{\beta}}_m \leftarrow \left(\sum_{i=1}^n \left(\boldsymbol{M}_{i\cdot}^\top - A_i\widetilde{\boldsymbol{\alpha}}_a\right)\left(\boldsymbol{M}_{i\cdot}^\top - A_i\widetilde{\boldsymbol{\alpha}}_a\right)^\top\right)^{-1} \sum_{i=1}^n \left(Y_i - A_i\widehat{\theta}_a^E - \boldsymbol{C}_{i\cdot}\widetilde{\boldsymbol{\beta}}_c\right)\left(\boldsymbol{M}_{i\cdot}^\top - A_i\widetilde{\boldsymbol{\alpha}}_a\right)$$

$$\widetilde{\sigma}_e^2 \leftarrow \frac{1}{n}\sum_{i=1}^n \left(Y_i - \{\widehat{\theta}_a^E - \widetilde{\boldsymbol{\alpha}}_a^\top \widetilde{\boldsymbol{\beta}}_m\}A_i - \boldsymbol{M}_{i\cdot}\widetilde{\boldsymbol{\beta}}_m - \boldsymbol{C}_{i\cdot}\widetilde{\boldsymbol{\beta}}_c\right)^2$$

Note that the cyclic coordinate descent algorithm is similar to Gibbs sampling, in that you use the always use the most recently updated values for all as you update each parameter. The initial values for the algorithm come from the unconstrained optimization algorithm.

Expectation-Maximization (EM) Algorithm (Soft Constraint)

Within the EM algorithm framework we are treating $\boldsymbol{Y}$, $\boldsymbol{M}$, $\boldsymbol{A}$, and $\boldsymbol{C}$ as our observed data and $\widetilde{\theta}_a^I$ as the unobserved latent data. Let $\mathcal{P} = \{\boldsymbol{\alpha}_a, \boldsymbol{\alpha}_c, \boldsymbol{\Sigma}_M, \boldsymbol{\beta}_m, \boldsymbol{\beta}_c, \sigma_e^2\}$. That is, we wish to maximize the marginal likelihood:

$$L(\mathcal{P} \mid \boldsymbol{Y}, \boldsymbol{M}, \boldsymbol{A}, \boldsymbol{C}) = \int_{-\infty}^{\infty} \pi(\boldsymbol{Y} \mid \boldsymbol{M}, \boldsymbol{A}, \boldsymbol{C}, \widetilde{\theta}_a^I)\pi(\boldsymbol{M} \mid \boldsymbol{A}, \boldsymbol{C})\pi(\widetilde{\theta}_a^I)d\widetilde{\theta}_a^I$$

Moreover, we have that

$$\pi(\widetilde{\theta}_a^I \mid \boldsymbol{Y}, \boldsymbol{M}, \boldsymbol{A}, \boldsymbol{C}) \propto \pi(\boldsymbol{Y} \mid \boldsymbol{M}, \boldsymbol{A}, \boldsymbol{C}, \widetilde{\theta}_a^I)\pi(\widetilde{\theta}_a^I)$$

which implies that

$$[\widetilde{\theta}_a^I \mid \boldsymbol{Y}, \boldsymbol{M}, \boldsymbol{A}, \boldsymbol{C}] \sim N\left(\left[\frac{\boldsymbol{A}^\top\boldsymbol{A}}{\sigma_e^2} + \frac{1}{s^2\widehat{\text{Var}}(\widehat{\theta}_a^E)}\right]^{-1}\left[\frac{\boldsymbol{A}^\top\boldsymbol{Y}^*}{\sigma_e^2} + \frac{\widehat{\theta}_a^E}{s^2\widehat{\text{Var}}(\widehat{\theta}_a^E)}\right], \left[\frac{\boldsymbol{A}^\top\boldsymbol{A}}{\sigma_e^2} + \frac{1}{s^2\widehat{\text{Var}}(\widehat{\theta}_a^E)}\right]^{-1}\right)$$

where $\boldsymbol{Y}^* = \boldsymbol{Y} - \boldsymbol{M}\boldsymbol{\beta}_m - \boldsymbol{C}\boldsymbol{\beta}_c + \boldsymbol{A}\boldsymbol{\alpha}_a^\top\boldsymbol{\beta}_m$. The complete data log-likelihood is:

$$l(\mathcal{P} \mid \boldsymbol{Y}, \boldsymbol{M}, \boldsymbol{A}, \boldsymbol{C}, \widetilde{\theta}_a^I) = -\frac{np_m}{2}\log(2\pi) - \frac{n}{2}\log(|\boldsymbol{\Sigma}_m|)$$

$$-\frac{1}{2}\sum_{i=1}^n (\boldsymbol{M}_{i\cdot}^\top - \boldsymbol{\alpha}_c\boldsymbol{C}_{i\cdot}^\top - A_i\boldsymbol{\alpha}_a)^\top \boldsymbol{\Sigma}_m^{-1}(\boldsymbol{M}_{i\cdot}^\top - \boldsymbol{\alpha}_c\boldsymbol{C}_{i\cdot}^\top - A_i\boldsymbol{\alpha}_a)$$

$$-\frac{n}{2}\log(2\pi) - \frac{n}{2}\log(\sigma_e^2) - \frac{1}{2\sigma_e^2}\sum_{i=1}^n (Y_i - \boldsymbol{C}_{i\cdot}\boldsymbol{\beta}_c - \boldsymbol{M}_{i\cdot}\boldsymbol{\beta}_m - A_i\{\widetilde{\theta}_a^I - \boldsymbol{\alpha}_a^\top\boldsymbol{\beta}_m\})^2$$

$$-\frac{1}{2}\log\left(2\pi s^2\widehat{\text{Var}}(\widehat{\theta}_a^E)\right) - \frac{1}{2s^2\widehat{\text{Var}}(\widehat{\theta}_a^E)}(\widetilde{\theta}_a^I - \widehat{\theta}_a^E)^2$$

The E-Step is therefore:

$$E_{\widetilde{\theta}_a^I \mid \boldsymbol{Y}, \boldsymbol{M}, \boldsymbol{A}, \boldsymbol{C}, \mathcal{P}^{(t)}}\left[l(\mathcal{P} \mid \boldsymbol{Y}, \boldsymbol{M}, \boldsymbol{A}, \boldsymbol{C}, \widetilde{\theta}_a^I)\right] = -\frac{np_m}{2}\log(2\pi) - \frac{n}{2}\log(|\boldsymbol{\Sigma}_m|)$$

$$-\frac{1}{2}\sum_{i=1}^n (\boldsymbol{M}_{i\cdot}^\top - \boldsymbol{\alpha}_c\boldsymbol{C}_{i\cdot}^\top - A_i\boldsymbol{\alpha}_a)^\top \boldsymbol{\Sigma}_m^{-1}(\boldsymbol{M}_{i\cdot}^\top - \boldsymbol{\alpha}_c\boldsymbol{C}_{i\cdot}^\top - A_i\boldsymbol{\alpha}_a)$$

$$-\frac{n}{2}\log(2\pi)-\frac{n}{2}\log(\sigma_e^2)-\frac{1}{2\sigma_e^2}\sum_{i=1}^{n}E_{\widetilde{\theta}_a^I|\boldsymbol{Y},\boldsymbol{M},\boldsymbol{A},\boldsymbol{C},\mathcal{P}^{(t)}}\left[(Y_i-\boldsymbol{C}_{i\cdot}\boldsymbol{\beta}_c-\boldsymbol{M}_{i\cdot}\boldsymbol{\beta}_m-A_i\{\widetilde{\theta}_a^I-\boldsymbol{\alpha}_a^\top\boldsymbol{\beta}_m\})^2\right]$$

$$-\frac{1}{2}\log\left(2\pi s^2\widehat{\text{Var}}(\widehat{\theta}_a^E)\right)-\frac{1}{2s^2\widehat{\text{Var}}(\widehat{\theta}_a^E)}E_{\widetilde{\theta}_a^I|\boldsymbol{Y},\boldsymbol{M},\boldsymbol{A},\boldsymbol{C},\mathcal{P}^{(t)}}\left[(\widetilde{\theta}_a^I-\widehat{\theta}_a^E)^2\right]$$

$$E_{\widetilde{\theta}_a^I|\boldsymbol{Y},\boldsymbol{M},\boldsymbol{A},\boldsymbol{C},\mathcal{P}^{(t)}}\left[l(\mathcal{P}\mid\boldsymbol{Y},\boldsymbol{M},\boldsymbol{A},\boldsymbol{C},\widetilde{\theta}_a^I)\right]=-\frac{np_m}{2}\log(2\pi)-\frac{n}{2}\log(|\boldsymbol{\Sigma}_m|)$$

$$-\frac{1}{2}\sum_{i=1}^{n}(\boldsymbol{M}_{i\cdot}^\top-\boldsymbol{\alpha}_c\boldsymbol{C}_{i\cdot}^\top-A_i\boldsymbol{\alpha}_a)^\top\boldsymbol{\Sigma}_m^{-1}(\boldsymbol{M}_{i\cdot}^\top-\boldsymbol{\alpha}_c\boldsymbol{C}_{i\cdot}^\top-A_i\boldsymbol{\alpha}_a)$$

$$-\frac{n}{2}\log(2\pi\sigma_e^2)-\frac{1}{2}\log\left(2\pi s^2\widehat{\text{Var}}(\widehat{\theta}_a^E)\right)$$

$$-\frac{1}{2s^2\widehat{\text{Var}}(\widehat{\theta}_a^E)}E_{\widetilde{\theta}_a^I|\boldsymbol{Y},\boldsymbol{M},\boldsymbol{A},\boldsymbol{C},\mathcal{P}^{(t)}}\left[\{\widetilde{\theta}_a^I\}^2\right]+\frac{\widehat{\theta}_a^E}{s^2\widehat{\text{Var}}(\widehat{\theta}_a^E)}E_{\widetilde{\theta}_a^I|\boldsymbol{Y},\boldsymbol{M},\boldsymbol{A},\boldsymbol{C},\mathcal{P}^{(t)}}\left[\widetilde{\theta}_a^I\right]-\frac{(\widehat{\theta}_a^E)^2}{2s^2\widehat{\text{Var}}(\widehat{\theta}_a^E)}$$

$$-\frac{1}{2\sigma_e^2}\left(\boldsymbol{Y}-\boldsymbol{M}\boldsymbol{\beta}_m-\boldsymbol{C}\boldsymbol{\beta}_c+\boldsymbol{A}\boldsymbol{\alpha}_a^\top\boldsymbol{\beta}_m\right)^\top\left(\boldsymbol{Y}-\boldsymbol{M}\boldsymbol{\beta}_m-\boldsymbol{C}\boldsymbol{\beta}_c+\boldsymbol{A}\boldsymbol{\alpha}_a^\top\boldsymbol{\beta}_m\right)$$

$$+\frac{1}{\sigma_e^2}E_{\widetilde{\theta}_a^I|\boldsymbol{Y},\boldsymbol{M},\boldsymbol{A},\boldsymbol{C},\mathcal{P}^{(t)}}\left[\widetilde{\theta}_a^I\right]\boldsymbol{A}^\top\left(\boldsymbol{Y}-\boldsymbol{M}\boldsymbol{\beta}_m-\boldsymbol{C}\boldsymbol{\beta}_c+\boldsymbol{A}\boldsymbol{\alpha}_a^\top\boldsymbol{\beta}_m\right)-\frac{1}{2\sigma_e^2}E_{\widetilde{\theta}_a^I|\boldsymbol{Y},\boldsymbol{M},\boldsymbol{A},\boldsymbol{C},\mathcal{P}^{(t)}}\left[\{\widetilde{\theta}_a^I\}^2\right]\boldsymbol{A}^\top\boldsymbol{A}$$

Note that we can calculate

$$E_{\widetilde{\theta}_a^I|\boldsymbol{Y},\boldsymbol{M},\boldsymbol{A},\boldsymbol{C},\mathcal{P}^{(t)}}\left[\widetilde{\theta}_a^I\right]\text{ and }E_{\widetilde{\theta}_a^I|\boldsymbol{Y},\boldsymbol{M},\boldsymbol{A},\boldsymbol{C},\mathcal{P}^{(t)}}\left[\{\widetilde{\theta}_a^I\}^2\right]$$

from

$$[\widetilde{\theta}_a^I\mid\boldsymbol{Y},\boldsymbol{M},\boldsymbol{A},\boldsymbol{C}]\sim N\left(\left[\frac{\boldsymbol{A}^\top\boldsymbol{A}}{\sigma_e^2}+\frac{1}{s^2\widehat{\text{Var}}(\widehat{\theta}_a^E)}\right]^{-1}\left[\frac{\boldsymbol{A}^\top\boldsymbol{Y}^*}{\sigma_e^2}+\frac{\widehat{\theta}_a^E}{s^2\widehat{\text{Var}}(\widehat{\theta}_a^E)}\right],\left[\frac{\boldsymbol{A}^\top\boldsymbol{A}}{\sigma_e^2}+\frac{1}{s^2\widehat{\text{Var}}(\widehat{\theta}_a^E)}\right]^{-1}\right)$$

For the M-Step we will use a cyclical coordinate descent algorithm with the following updates:

$$\widetilde{\boldsymbol{\alpha}}_c\leftarrow\left(\boldsymbol{C}^\top\boldsymbol{C}\right)^{-1}\boldsymbol{C}^\top\left(\boldsymbol{M}-\boldsymbol{A}\widetilde{\boldsymbol{\alpha}}_a^\top\right)$$

$$\widetilde{\boldsymbol{\alpha}}_a\leftarrow\frac{1}{\boldsymbol{A}^\top\boldsymbol{A}}\left(\widetilde{\boldsymbol{\Sigma}}_m^{-1}+\frac{1}{\widetilde{\sigma}_e^2}\widetilde{\boldsymbol{\beta}}_m\widetilde{\boldsymbol{\beta}}_m^\top\right)^{-1}\left[-\frac{1}{\widetilde{\sigma}_e^2}\boldsymbol{A}^\top\left(\boldsymbol{Y}-\boldsymbol{M}\widetilde{\boldsymbol{\beta}}_m-\boldsymbol{C}\widetilde{\boldsymbol{\beta}}_c-E_{\widetilde{\theta}_a^I|\boldsymbol{Y},\boldsymbol{M},\boldsymbol{A},\boldsymbol{C},\mathcal{P}^{(t)}}\left[\widetilde{\theta}_a^I\right]\boldsymbol{A}\right)\widetilde{\boldsymbol{\beta}}_m+\sum_{i=1}^{n}A_i\widetilde{\boldsymbol{\Sigma}}_m^{-1}\left(\boldsymbol{M}_{i\cdot}^\top-\widetilde{\boldsymbol{\alpha}}_c^\top\boldsymbol{C}_{i\cdot}^\top\right)\right]$$

$$\widetilde{\boldsymbol{\Sigma}}_m\leftarrow\frac{1}{n}\sum_{i=1}^{n}\left(\boldsymbol{M}_{i\cdot}^\top-\widetilde{\boldsymbol{\alpha}}^\top\boldsymbol{X}_{i\cdot}^\top\right)\left(\boldsymbol{M}_{i\cdot}^\top-\widetilde{\boldsymbol{\alpha}}^\top\boldsymbol{X}_{i\cdot}^\top\right)^\top,\ \ \boldsymbol{X}=(\boldsymbol{A},\boldsymbol{C}),\ \ \boldsymbol{\alpha}=\begin{pmatrix}\boldsymbol{\alpha}_a^\top\\\boldsymbol{\alpha}_c^\top\end{pmatrix}$$

$$\widetilde{\boldsymbol{\beta}}_c\leftarrow\left(\boldsymbol{C}^\top\boldsymbol{C}\right)^{-1}\boldsymbol{C}^\top\left(\boldsymbol{Y}-\{E_{\widetilde{\theta}_a^I|\boldsymbol{Y},\boldsymbol{M},\boldsymbol{A},\boldsymbol{C},\mathcal{P}^{(t)}}\left[\widetilde{\theta}_a^I\right]-\widetilde{\boldsymbol{\alpha}}_a^\top\widetilde{\boldsymbol{\beta}}_m\}\boldsymbol{A}-\boldsymbol{M}\widetilde{\boldsymbol{\beta}}_m\right)$$

$$\widetilde{\boldsymbol{\beta}}_m\leftarrow\left[\left(\boldsymbol{M}-\boldsymbol{A}\widetilde{\boldsymbol{\alpha}}_a^\top\right)^\top\left(\boldsymbol{M}-\boldsymbol{A}\widetilde{\boldsymbol{\alpha}}_a^\top\right)\right]^{-1}\left(\boldsymbol{M}-\boldsymbol{A}\widetilde{\boldsymbol{\alpha}}_a^\top\right)^\top\left(\boldsymbol{Y}-\boldsymbol{C}\widetilde{\boldsymbol{\beta}}_c-\boldsymbol{A}E_{\widetilde{\theta}_a^I|\boldsymbol{Y},\boldsymbol{M},\boldsymbol{A},\boldsymbol{C},\mathcal{P}^{(t)}}\left[\widetilde{\theta}_a^I\right]\right)$$

$$\widetilde{\sigma}_e^2\leftarrow\frac{1}{n}\left(\boldsymbol{Y}-\boldsymbol{M}\widetilde{\boldsymbol{\beta}}_m-\boldsymbol{C}\widetilde{\boldsymbol{\beta}}_c+\boldsymbol{A}\widetilde{\boldsymbol{\alpha}}_a^\top\widetilde{\boldsymbol{\beta}}_m\right)^\top\left(\boldsymbol{Y}-\boldsymbol{M}\widetilde{\boldsymbol{\beta}}_m-\boldsymbol{C}\widetilde{\boldsymbol{\beta}}_c+\boldsymbol{A}\widetilde{\boldsymbol{\alpha}}_a^\top\widetilde{\boldsymbol{\beta}}_m\right)$$

$$-\frac{2}{n}E_{\widetilde{\theta}_a^I|\boldsymbol{Y},\boldsymbol{M},\boldsymbol{A},\boldsymbol{C},\mathcal{P}^{(t)}}\left[\widetilde{\theta}_a^I\right]\boldsymbol{A}^\top\left(\boldsymbol{Y}-\boldsymbol{M}\widetilde{\boldsymbol{\beta}}_m-\boldsymbol{C}\widetilde{\boldsymbol{\beta}}_c+\boldsymbol{A}\widetilde{\boldsymbol{\alpha}}_a^\top\widetilde{\boldsymbol{\beta}}_m\right)$$

$$+\frac{\boldsymbol{A}^\top\boldsymbol{A}}{n}E_{\widetilde{\theta}_a^I|\boldsymbol{Y},\boldsymbol{M},\boldsymbol{A},\boldsymbol{C},\mathcal{P}^{(t)}}\left[\{\widetilde{\theta}_a^I\}^2\right]$$

B.2 Project 2 Proofs

Proof of Theorem 1

Suppose that $[\boldsymbol{M}_{i\cdot}^\top \mid A_i] \sim N(\boldsymbol{\alpha}_c + A_i\boldsymbol{\alpha}_a, \boldsymbol{\Sigma}_m)$ and $[Y_i \mid \boldsymbol{M}_{i\cdot}, A_i] \sim N(\beta_c + \boldsymbol{M}_{i\cdot}\boldsymbol{\beta}_m + A_i\beta_a, \sigma_e^2)$ is the true generative model where $i = 1, \ldots, n$. That is, without loss of generality, we are assuming that there are no confounders, but that there are intercept terms in both the outcome and mediator models. This implies that $\sigma_a^2 = \text{Var}(A_i)$. Moreover, assume that $E[A_i] = 0$. Let $\mathcal{D} = \{\boldsymbol{Y}, \boldsymbol{M}, \boldsymbol{A}, \boldsymbol{C}\}$ denote the data. Then the log-likelihood is:

$$l(\boldsymbol{\alpha}_c, \boldsymbol{\alpha}_a, \beta_c, \boldsymbol{\beta}_m, \beta_a \mid \mathcal{D}) = -\frac{np_m}{2}\log(2\pi) - \frac{n}{2}\log(|\boldsymbol{\Sigma}_m|) - \frac{1}{2}\sum_{i=1}^{n}(\boldsymbol{M}_{i\cdot}^\top - \boldsymbol{\alpha}_c - A_i\boldsymbol{\alpha}_a)^\top\boldsymbol{\Sigma}_m^{-1}(\boldsymbol{M}_{i\cdot}^\top - \boldsymbol{\alpha}_c - A_i\boldsymbol{\alpha}_a)$$

$$-\frac{n}{2}\log(2\pi) - \frac{n}{2}\log(\sigma_e^2) - \frac{1}{2\sigma_e^2}\sum_{i=1}^{n}(Y_i - \beta_c - \boldsymbol{M}_{i\cdot}\boldsymbol{\beta}_m - A_i\beta_a)^2$$

The first order derivatives are:

$$\frac{\partial l}{\partial \boldsymbol{\alpha}_a} = \sum_{i=1}^{n} A_i\boldsymbol{\Sigma}_m^{-1}\left[\boldsymbol{M}_{i\cdot}^\top - \boldsymbol{\alpha}_c - A_i\boldsymbol{\alpha}_a\right]$$

$$\frac{\partial l}{\partial \boldsymbol{\alpha}_c} = \sum_{i=1}^{n} \boldsymbol{\Sigma}_m^{-1}\left[\boldsymbol{M}_{i\cdot}^\top - \boldsymbol{\alpha}_c - A_i\boldsymbol{\alpha}_a\right]$$

$$\frac{\partial l}{\partial \boldsymbol{\Sigma}_m^{-1}} = \frac{n}{2}\boldsymbol{\Sigma}_m - \frac{1}{2}\sum_{i=1}^{n}\left[\boldsymbol{M}_{i\cdot}^\top - \boldsymbol{\alpha}_c - A_i\boldsymbol{\alpha}_a\right]\left[\boldsymbol{M}_{i\cdot}^\top - \boldsymbol{\alpha}_c - A_i\boldsymbol{\alpha}_a\right]^\top$$

$$\frac{\partial l}{\partial \boldsymbol{\beta}_m} = \frac{1}{\sigma_e^2}\sum_{i=1}^{n}\boldsymbol{M}_{i\cdot}^\top\left[Y_i - \beta_c - \boldsymbol{M}_{i\cdot}\boldsymbol{\beta}_m - A_i\beta_a\right]$$

$$\frac{\partial l}{\partial \beta_a} = \frac{1}{\sigma_e^2}\sum_{i=1}^{n}A_i\left[Y_i - \beta_c - \boldsymbol{M}_{i\cdot}\boldsymbol{\beta}_m - A_i\beta_a\right]$$

$$\frac{\partial l}{\partial \beta_c} = \frac{1}{\sigma_e^2}\sum_{i=1}^{n}\left[Y_i - \beta_c - \boldsymbol{M}_{i\cdot}\boldsymbol{\beta}_m - A_i\beta_a\right]$$

$$\frac{\partial l}{\partial \sigma_e^2} = -\frac{n}{2\sigma_e^2} + \frac{1}{2\sigma_e^4}\sum_{i=1}^{n}\left[Y_i - \beta_c - \boldsymbol{M}_{i\cdot}\boldsymbol{\beta}_m - A_i\beta_a\right]^2$$

The second order derivatives are:

$$\frac{\partial^2 l}{\partial \boldsymbol{\alpha}_a \partial \boldsymbol{\alpha}_a^\top} = -\boldsymbol{\Sigma}_m^{-1}\sum_{i=1}^{n}A_i^2 \implies -\frac{1}{n}E\left[\frac{\partial^2 l}{\partial \boldsymbol{\alpha}_a \partial \boldsymbol{\alpha}_a^\top}\right] \to \sigma_a^2\boldsymbol{\Sigma}_m^{-1}$$

$$\frac{\partial^2 l}{\partial \boldsymbol{\alpha}_c \partial \boldsymbol{\alpha}_c^\top} = -n\boldsymbol{\Sigma}_m^{-1} \implies -\frac{1}{n}E\left[\frac{\partial^2 l}{\partial \boldsymbol{\alpha}_c \partial \boldsymbol{\alpha}_c^\top}\right] = \boldsymbol{\Sigma}_m^{-1}$$

$$\frac{\partial^2 l}{\partial \boldsymbol{\alpha}_a \partial \boldsymbol{\alpha}_c^\top} = -\boldsymbol{\Sigma}_m^{-1}\sum_{i=1}^{n}A_i \implies -\frac{1}{n}E\left[\frac{\partial^2 l}{\partial \boldsymbol{\alpha}_a \partial \boldsymbol{\alpha}_c^\top}\right] \to \mathbf{0}$$

$$\frac{\partial^2 l}{\partial \boldsymbol{\alpha}_c \partial \boldsymbol{\alpha}_a^\top} = -\boldsymbol{\Sigma}_m^{-1} \sum_{i=1}^{n} A_i \implies -\frac{1}{n} E\left[\frac{\partial^2 l}{\partial \boldsymbol{\alpha}_c \partial \boldsymbol{\alpha}_a^\top}\right] \to \mathbf{0}$$

$$\frac{\partial^2 l}{\partial \boldsymbol{\beta}_m \partial \boldsymbol{\beta}_m^\top} = -\frac{1}{\sigma_e^2} \sum_{i=1}^{n} \boldsymbol{M}_{i\cdot}^\top \boldsymbol{M}_{i\cdot} \implies -\frac{1}{n} E\left[\frac{\partial^2 l}{\partial \boldsymbol{\beta}_m \partial \boldsymbol{\beta}_m^\top}\right] \to \frac{1}{\sigma_e^2}(\boldsymbol{\Sigma}_m + \boldsymbol{\alpha}_c \boldsymbol{\alpha}_c^\top) + \frac{\sigma_a^2}{\sigma_e^2} \boldsymbol{\alpha}_a \boldsymbol{\alpha}_a^\top$$

$$\frac{\partial^2 l}{\partial \beta_a^2} = -\frac{1}{\sigma_e^2} \sum_{i=1}^{n} A_i^2 \implies -\frac{1}{n} E\left[\frac{\partial^2 l}{\partial \beta_a^2}\right] \to \frac{\sigma_a^2}{\sigma_e^2}$$

$$\frac{\partial^2 l}{\partial \beta_c^2} = -\frac{n}{\sigma_e^2} \implies -\frac{1}{n} E\left[\frac{\partial^2 l}{\partial \beta_c^2}\right] = \frac{1}{\sigma_e^2}$$

$$\frac{\partial^2 l}{\partial \boldsymbol{\beta}_m \partial \beta_a} = -\frac{1}{\sigma_e^2} \sum_{i=1}^{n} A_i \boldsymbol{M}_{i\cdot}^\top \implies -\frac{1}{n} E\left[\frac{\partial^2 l}{\partial \boldsymbol{\beta}_m \partial \beta_a}\right] \to \frac{\sigma_a^2}{\sigma_e^2} \boldsymbol{\alpha}_a$$

$$\frac{\partial^2 l}{\partial \beta_a \partial \boldsymbol{\beta}_m^\top} = -\frac{1}{\sigma_e^2} \sum_{i=1}^{n} A_i \boldsymbol{M}_{i\cdot} \implies -\frac{1}{n} E\left[\frac{\partial^2 l}{\partial \beta_a \partial \boldsymbol{\beta}_m^\top}\right] \to \frac{\sigma_a^2}{\sigma_e^2} \boldsymbol{\alpha}_a^\top$$

$$\frac{\partial^2 l}{\partial \boldsymbol{\beta}_m \partial \beta_c} = -\frac{1}{\sigma_e^2} \sum_{i=1}^{n} \boldsymbol{M}_{i\cdot}^\top \implies -\frac{1}{n} E\left[\frac{\partial^2 l}{\partial \boldsymbol{\beta}_m \partial \beta_c}\right] \to \frac{1}{\sigma_e^2} \boldsymbol{\alpha}_c$$

$$\frac{\partial^2 l}{\partial \beta_c \partial \boldsymbol{\beta}_m^\top} = -\frac{1}{\sigma_e^2} \sum_{i=1}^{n} \boldsymbol{M}_{i\cdot} \implies -\frac{1}{n} E\left[\frac{\partial^2 l}{\partial \beta_c \partial \boldsymbol{\beta}_m^\top}\right] \to \frac{1}{\sigma_e^2} \boldsymbol{\alpha}_c^\top$$

$$\frac{\partial^2 l}{\partial \beta_a \partial \beta_c} = -\frac{1}{\sigma_e^2} \sum_{i=1}^{n} A_i \implies -\frac{1}{n} E\left[\frac{\partial^2 l}{\partial \beta_a \partial \beta_c}\right] \to 0$$

Note that:

$$-\frac{1}{n} E\left[\frac{\partial^2 l}{\partial \boldsymbol{\alpha}_a \partial (\boldsymbol{\Sigma}_m^{-1})_{kl}}\right] \to \mathbf{0}, \quad 1 \le k, l \le p_m$$

$$-\frac{1}{n} E\left[\frac{\partial^2 l}{\partial \boldsymbol{\alpha}_c \partial (\boldsymbol{\Sigma}_m^{-1})_{kl}}\right] \to \mathbf{0}, \quad 1 \le k, l \le p_m$$

$$-\frac{1}{n} E\left[\frac{\partial^2 l}{\partial \beta_a \partial \sigma_e^2}\right] = 0$$

$$-\frac{1}{n} E\left[\frac{\partial^2 l}{\partial \beta_c \partial \sigma_e^2}\right] = 0$$

$$-\frac{1}{n} E\left[\frac{\partial^2 l}{\partial \boldsymbol{\beta}_m \partial \sigma_e^2}\right] = \mathbf{0}$$

Because the outcome and mediator models are independent of one another, then we can consider the Fisher information matrix for the outcome model:

$$\mathcal{I}^U(\beta_c, \beta_a, \boldsymbol{\beta}_m) = \begin{pmatrix} \frac{1}{\sigma_e^2} & 0 & \frac{1}{\sigma_e^2}\boldsymbol{\alpha}_c^\top \\ 0 & \frac{\sigma_a^2}{\sigma_e^2} & \frac{\sigma_a^2}{\sigma_e^2}\boldsymbol{\alpha}_a^\top \\ \frac{1}{\sigma_e^2}\boldsymbol{\alpha}_c & \frac{\sigma_a^2}{\sigma_e^2}\boldsymbol{\alpha}_a & \frac{1}{\sigma_e^2}(\boldsymbol{\Sigma}_m + \boldsymbol{\alpha}_c\boldsymbol{\alpha}_c^\top) + \frac{\sigma_a^2}{\sigma_e^2}\boldsymbol{\alpha}_a\boldsymbol{\alpha}_a^\top \end{pmatrix}$$

Using the block inversion formula for 2×2 block matricies, we can see that:

$$\Big(\mathcal{I}^U(\beta_a, \boldsymbol{\beta}_m)\Big)^{-1} = \begin{pmatrix} \frac{\sigma_a^2}{\sigma_e^2} & \frac{\sigma_a^2}{\sigma_e^2}\boldsymbol{\alpha}_a^\top \\ \frac{\sigma_a^2}{\sigma_e^2}\boldsymbol{\alpha}_a & \frac{1}{\sigma_e^2}\boldsymbol{\Sigma}_m + \frac{\sigma_a^2}{\sigma_e^2}\boldsymbol{\alpha}_a\boldsymbol{\alpha}_a^\top \end{pmatrix}^{-1}$$

Again, using the block inversion formula for 2×2 block matricies, we can see that:

$$\Big(\mathcal{I}^U(\beta_a, \boldsymbol{\beta}_m)\Big)^{-1} = \begin{pmatrix} \frac{\sigma_e^2}{\sigma_a^2}\big(1 + \sigma_a^2\boldsymbol{\alpha}_a^\top\boldsymbol{\Sigma}_m^{-1}\boldsymbol{\alpha}_a\big) & -\sigma_e^2\boldsymbol{\alpha}_a^\top\boldsymbol{\Sigma}_m^{-1} \\ -\sigma_e^2\boldsymbol{\Sigma}_m^{-1}\boldsymbol{\alpha}_a & \sigma_e^2\boldsymbol{\Sigma}_m^{-1} \end{pmatrix}$$

Therefore, conclude that:

$$\Big(\mathcal{I}^U(\boldsymbol{\alpha}_a, \beta_a, \boldsymbol{\beta}_m)\Big)^{-1} = \begin{pmatrix} \frac{1}{\sigma_a^2}\boldsymbol{\Sigma}_m & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \frac{\sigma_e^2}{\sigma_a^2}\big(1 + \sigma_a^2\boldsymbol{\alpha}_a^\top\boldsymbol{\Sigma}_m^{-1}\boldsymbol{\alpha}_a\big) & -\sigma_e^2\boldsymbol{\alpha}_a^\top\boldsymbol{\Sigma}_m^{-1} \\ \mathbf{0} & -\sigma_e^2\boldsymbol{\Sigma}_m^{-1}\boldsymbol{\alpha}_a & \sigma_e^2\boldsymbol{\Sigma}_m^{-1} \end{pmatrix}$$

The asymptotic distribution for the unconstrained estimator of the NIE can then be computed using the multivariate delta method provided that $\boldsymbol{\alpha}_a \neq \mathbf{0}$ or $\boldsymbol{\beta}_m \neq \mathbf{0}$.

To extend this result to the model with confounders

$$[\boldsymbol{Y} \mid \boldsymbol{M}, \boldsymbol{A}, \boldsymbol{C}] \sim N\Big(\boldsymbol{M}\boldsymbol{\beta}_m + \boldsymbol{A}\beta_a + \boldsymbol{C}\boldsymbol{\beta}_c, \sigma_e^2\boldsymbol{I}\Big),$$
$$[\boldsymbol{M}_{i\cdot}^\top \mid \boldsymbol{A}_i, \boldsymbol{C}_i] \sim N\Big(A_i\boldsymbol{\alpha}_a + \boldsymbol{\alpha}_c\boldsymbol{C}_i^\top, \boldsymbol{\Sigma}_m\Big), \;\; i = 1, \ldots, n.$$

we note that the model with confounders is equivalent to a model with no confounders after the confounders are regressed out from the outcome, mediators, and exposure. That is, we can use the residuals from linear regression models $[\boldsymbol{Y} \mid \boldsymbol{C}]$, $[\boldsymbol{M} \mid \boldsymbol{C}]$, and $[\boldsymbol{A} \mid \boldsymbol{C}]$ as the new outcome, mediators, and exposure and run the unconstrained model. Therefore, the only difference in the asymptotic normality expression is that the exposure are now the residuals from the $[\boldsymbol{A} \mid \boldsymbol{C}]$ linear regression model, implying that $\sigma_a^2 = \text{Var}(A_i \mid \boldsymbol{C}_{i\cdot})$.

Proof of Theorem 2

Suppose that $[\boldsymbol{M}_{i\cdot}^\top \mid A_i] \sim N(\boldsymbol{\alpha}_c + A_i\boldsymbol{\alpha}_a, \boldsymbol{\Sigma}_m)$ and $[Y_i \mid \boldsymbol{M}_{i\cdot}, A_i] \sim N(\beta_c + \boldsymbol{M}_{i\cdot}\boldsymbol{\beta}_m + A_i\beta_a, \sigma_e^2)$ is the true generative model where $i = 1, \ldots, n$. That is, without loss of generality, we are assuming that there are no confounders, but that there are intercept terms in both the outcome and mediator models. This implies that $\sigma_a^2 = \text{Var}(A_i)$. Moreover, assume that $E[A_i] = 0$. Let $\mathcal{D} = \{\boldsymbol{Y}, \boldsymbol{M}, \boldsymbol{A}, \boldsymbol{C}\}$ denote the data. Suppose that we fit the hard constraint model to our data. Then the log-likelihood is:

$$l(\boldsymbol{\alpha}_c, \boldsymbol{\alpha}_a, \boldsymbol{\Sigma}_m, \boldsymbol{\beta}_m, \beta_c, \sigma_e^2 \mid \mathcal{D}) = -\frac{np_m}{2}\log(2\pi) - \frac{n}{2}\log(|\boldsymbol{\Sigma}_m|) - \frac{1}{2}\sum_{i=1}^{n}(\boldsymbol{M}_{i\cdot}^\top - \boldsymbol{\alpha}_c - A_i\boldsymbol{\alpha}_a)^\top \boldsymbol{\Sigma}_m^{-1}(\boldsymbol{M}_{i\cdot}^\top - \boldsymbol{\alpha}_c - A_i\boldsymbol{\alpha}_a)$$

$$-\frac{n}{2}\log(2\pi) - \frac{n}{2}\log(\sigma_e^2) - \frac{1}{2\sigma_e^2}\sum_{i=1}^{n}(Y_i - \beta_c - \boldsymbol{M}_{i\cdot}\boldsymbol{\beta}_m - A_i\{\widehat{\theta}_a^E - \boldsymbol{\alpha}_a^\top\boldsymbol{\beta}_m\})^2$$

The first order derivatives are:

$$\frac{\partial l}{\partial \boldsymbol{\alpha}_a} = \sum_{i=1}^{n} A_i\boldsymbol{\Sigma}_m^{-1}\big[\boldsymbol{M}_{i\cdot}^\top - \boldsymbol{\alpha}_c - A_i\boldsymbol{\alpha}_a\big] - \frac{1}{\sigma_e^2}\sum_{i=1}^{n} A_i\boldsymbol{\beta}_m\big[Y_i - \beta_c - A_i(\widehat{\theta}_a^E - \boldsymbol{\alpha}_a^\top\boldsymbol{\beta}_m) - \boldsymbol{M}_{i\cdot}\boldsymbol{\beta}_m\big]$$

$$\frac{\partial l}{\partial \boldsymbol{\alpha}_c} = \sum_{i=1}^{n} \boldsymbol{\Sigma}_m^{-1}\big[\boldsymbol{M}_{i\cdot}^\top - \boldsymbol{\alpha}_c - A_i\boldsymbol{\alpha}_a\big]$$

$$\frac{\partial l}{\partial \boldsymbol{\Sigma}_m^{-1}} = \frac{n}{2}\boldsymbol{\Sigma}_m - \frac{1}{2}\sum_{i=1}^{n}\big[\boldsymbol{M}_{i\cdot}^\top - \boldsymbol{\alpha}_c - A_i\boldsymbol{\alpha}_a\big]\big[\boldsymbol{M}_{i\cdot}^\top - \boldsymbol{\alpha}_c - A_i\boldsymbol{\alpha}_a\big]^\top$$

$$\frac{\partial l}{\partial \boldsymbol{\beta}_m} = \frac{1}{\sigma_e^2}\sum_{i=1}^{n}\big[Y_i - \beta_c - A_i(\widehat{\theta}_a^E - \boldsymbol{\alpha}_a^\top\boldsymbol{\beta}_m) - \boldsymbol{M}_{i\cdot}\boldsymbol{\beta}_m\big]\big[\boldsymbol{M}_{i\cdot}^\top - A_i\boldsymbol{\alpha}_a\big]$$

$$\frac{\partial l}{\partial \beta_c} = \frac{1}{\sigma_e^2}\sum_{i=1}^{n}\big[Y_i - \beta_c - A_i(\widehat{\theta}_a^E - \boldsymbol{\alpha}_a^\top\boldsymbol{\beta}_m) - \boldsymbol{M}_{i\cdot}\boldsymbol{\beta}_m\big]$$

$$\frac{\partial l}{\partial \sigma_e^2} = -\frac{n}{2\sigma_e^2} + \frac{1}{2\sigma_e^4}\sum_{i=1}^{n}\big[Y_i - \beta_c - A_i(\widehat{\theta}_a^E - \boldsymbol{\alpha}_a^\top\boldsymbol{\beta}_m) - \boldsymbol{M}_{i\cdot}\boldsymbol{\beta}_m\big]^2$$

The second order derivatives are:

$$\frac{\partial^2 l}{\partial \boldsymbol{\alpha}_a \partial \boldsymbol{\alpha}_a^\top} = -\boldsymbol{\Sigma}_m^{-1}\sum_{i=1}^{n} A_i^2 - \frac{1}{\sigma_e^2}\boldsymbol{\beta}_m\boldsymbol{\beta}_m^\top\sum_{i=1}^{n} A_i^2 \implies -\frac{1}{n}E\left[\frac{\partial^2 l}{\partial \boldsymbol{\alpha}_a \partial \boldsymbol{\alpha}_a^\top}\right] \to \sigma_a^2\boldsymbol{\Sigma}_m^{-1} + \frac{\sigma_a^2}{\sigma_e^2}\boldsymbol{\beta}_m\boldsymbol{\beta}_m^\top$$

$$\frac{\partial^2 l}{\partial \boldsymbol{\alpha}_c \partial \boldsymbol{\alpha}_c^\top} = -n\boldsymbol{\Sigma}_m^{-1} \implies -\frac{1}{n}E\left[\frac{\partial^2 l}{\partial \boldsymbol{\alpha}_c \partial \boldsymbol{\alpha}_c^\top}\right] = \boldsymbol{\Sigma}_m^{-1}$$

$$\frac{\partial^2 l}{\partial \boldsymbol{\alpha}_a \partial \boldsymbol{\alpha}_c^\top} = -\boldsymbol{\Sigma}_m^{-1}\sum_{i=1}^{n} A_i \implies -\frac{1}{n}E\left[\frac{\partial^2 l}{\partial \boldsymbol{\alpha}_a \partial \boldsymbol{\alpha}_c^\top}\right] \to \mathbf{0}$$

$$\frac{\partial^2 l}{\partial \boldsymbol{\alpha}_c \partial \boldsymbol{\alpha}_a^\top} = -\boldsymbol{\Sigma}_m^{-1}\sum_{i=1}^{n} A_i \implies -\frac{1}{n}E\left[\frac{\partial^2 l}{\partial \boldsymbol{\alpha}_c \partial \boldsymbol{\alpha}_a^\top}\right] \to \mathbf{0}$$

$$\frac{\partial^2 l}{\partial \boldsymbol{\beta}_m \partial \boldsymbol{\beta}_m^\top} = -\frac{1}{\sigma_e^2}\sum_{i=1}^{n}\left[\boldsymbol{M}_{i\cdot}^\top - A_i\boldsymbol{\alpha}_a\right]\left[\boldsymbol{M}_{i\cdot} - A_i\boldsymbol{\alpha}_a^\top\right] \implies -\frac{1}{n}E\left[\frac{\partial^2 l}{\partial \boldsymbol{\beta}_m \partial \boldsymbol{\beta}_m^\top}\right] \to \frac{1}{\sigma_e^2}(\boldsymbol{\Sigma}_m + \boldsymbol{\alpha}_c\boldsymbol{\alpha}_c^\top)$$

$$\frac{\partial^2 l}{\partial \beta_c^2} = -\frac{n}{\sigma_e^2} \implies -\frac{1}{n}E\left[\frac{\partial^2 l}{\partial \beta_c^2}\right] = \frac{1}{\sigma_e^2}$$

$$\frac{\partial^2 l}{\partial \boldsymbol{\beta}_m \partial \beta_c} = -\frac{1}{\sigma_e^2}\sum_{i=1}^{n}\left[\boldsymbol{M}_{i\cdot}^\top - A_i\boldsymbol{\alpha}_a\right] \implies -\frac{1}{n}E\left[\frac{\partial^2 l}{\partial \boldsymbol{\beta}_m \partial \beta_c}\right] \to \frac{1}{\sigma_e^2}\boldsymbol{\alpha}_c$$

$$\frac{\partial^2 l}{\partial \beta_c \partial \boldsymbol{\beta}_m^\top} = -\frac{1}{\sigma_e^2}\sum_{i=1}^{n}\left[\boldsymbol{M}_{i\cdot} - A_i\boldsymbol{\alpha}_a^\top\right] \implies -\frac{1}{n}E\left[\frac{\partial^2 l}{\partial \beta_c \partial \boldsymbol{\beta}_m^\top}\right] \to \frac{1}{\sigma_e^2}\boldsymbol{\alpha}_c^\top$$

$$\frac{\partial^2 l}{\partial \boldsymbol{\alpha}_a \partial \boldsymbol{\beta}_m^\top} = -\frac{1}{\sigma_e^2}\sum_{i=1}^{n}\left[A_i(Y_i - \beta_c - A_i(\widehat{\theta}_a^E - \boldsymbol{\alpha}_a^\top\boldsymbol{\beta}_m) - \boldsymbol{M}_{i\cdot}\boldsymbol{\beta}_m)\boldsymbol{I} + A_i\boldsymbol{\beta}_m(A_i\boldsymbol{\alpha}_a^\top - \boldsymbol{M}_{i\cdot})\right]$$

$$\implies -\frac{1}{n}E\left[\frac{\partial^2 l}{\partial \boldsymbol{\alpha}_a \partial \boldsymbol{\beta}_m^\top}\right] \to \mathbf{0}$$

$$\frac{\partial^2 l}{\partial \boldsymbol{\beta}_m \partial \boldsymbol{\alpha}_a^\top} = -\frac{1}{\sigma_e^2}\sum_{i=1}^{n}\left[A_i(Y_i - \beta_c - A_i(\widehat{\theta}_a^E - \boldsymbol{\alpha}_a^\top\boldsymbol{\beta}_m) - \boldsymbol{M}_{i\cdot}\boldsymbol{\beta}_m)\boldsymbol{I} + A_i(A_i\boldsymbol{\alpha}_a - \boldsymbol{M}_{i\cdot}^\top)\boldsymbol{\beta}_m^\top\right]$$

$$\implies -\frac{1}{n}E\left[\frac{\partial^2 l}{\partial \boldsymbol{\beta}_m \partial \boldsymbol{\alpha}_a^\top}\right] \to \mathbf{0}$$

Note that:

$$-\frac{1}{n}E\left[\frac{\partial^2 l}{\partial \boldsymbol{\alpha}_a \partial (\boldsymbol{\Sigma}_m^{-1})_{kl}}\right] \to \mathbf{0}, \quad 1 \leq k, l \leq p_m$$

$$-\frac{1}{n}E\left[\frac{\partial^2 l}{\partial \boldsymbol{\alpha}_c \partial (\boldsymbol{\Sigma}_m^{-1})_{kl}}\right] \to \mathbf{0}, \quad 1 \leq k, l \leq p_m$$

$$-\frac{1}{n}E\left[\frac{\partial^2 l}{\partial \boldsymbol{\alpha}_a \partial \sigma_e^2}\right] = \mathbf{0}$$

$$-\frac{1}{n}E\left[\frac{\partial^2 l}{\partial \beta_c \partial \sigma_e^2}\right] = 0$$

$$-\frac{1}{n}E\left[\frac{\partial^2 l}{\partial \boldsymbol{\beta}_m \partial \sigma_e^2}\right] = \mathbf{0}$$

Using the block inversion formula for 2×2 block matricies, we can see that:

$$\left(\mathcal{I}^H(\boldsymbol{\alpha}_a, \boldsymbol{\beta}_m)\right)^{-1} = \begin{pmatrix} \frac{1}{\sigma_a^2}\left(\boldsymbol{\Sigma}_m^{-1} + \frac{1}{\sigma_e^2}\boldsymbol{\beta}_m\boldsymbol{\beta}_m^\top\right)^{-1} & \mathbf{0} \\ \mathbf{0} & \sigma_e^2\boldsymbol{\Sigma}_m^{-1} \end{pmatrix}$$

The asymptotic distribution for the hard constraint estimator of the NIE can then be computed using the multivariate delta method provided that $\boldsymbol{\alpha}_a \neq \mathbf{0}$ or $\boldsymbol{\beta}_m \neq \mathbf{0}$. Because the hard constraint estimator of the NDE is just a location shift of the hard constraint estimator of the NIE, then provided that $\sqrt{n}(\widehat{\theta}_a^E - \theta_a^I) \to_p 0$, Slutsky's Theorem tells us

that the asymptotic distribution for the hard constraint estimator of the NDE is the same as the asymptotic distribution for the hard constraint estimator of the NIE.

To extend this result to the model with confounders we use the same trick that was used in the proof of Theorem 1, namely that the model with confounders is equivalent to a model with no confounders after the confounders are regressed out from the outcome, mediators, and exposure via linear regression models. Therefore, we again see that $\sigma_a^2 = \text{Var}(A_i \mid \boldsymbol{C}_{i\cdot})$, when confounders are present.

Proof of Theorem 3

We have that

$$\sqrt{n}\begin{pmatrix}\widehat{\boldsymbol{\alpha}}_a - \boldsymbol{\alpha}_a \\ \widehat{\boldsymbol{\beta}}_m - \boldsymbol{\beta}_m\end{pmatrix} \rightarrow_d N\left(\begin{pmatrix}\mathbf{0} \\ \mathbf{0}\end{pmatrix}, \begin{pmatrix}\sigma_a^{-2}\boldsymbol{\Sigma}_m & \mathbf{0} \\ \mathbf{0} & \sigma_e^2\boldsymbol{\Sigma}_m^{-1}\end{pmatrix}\right)$$

Using the second order Taylor expansion around $\boldsymbol{\alpha}_a = \mathbf{0}$ and $\boldsymbol{\beta}_m = \mathbf{0}$, we can show that

$$n\left(\widehat{\boldsymbol{\alpha}}_a^\top\widehat{\boldsymbol{\beta}}_m - \boldsymbol{\alpha}_a^\top\boldsymbol{\beta}_m\right) = \frac{n}{2}\begin{pmatrix}\widehat{\boldsymbol{\alpha}}_a^\top - \boldsymbol{\alpha}_a^\top & \widehat{\boldsymbol{\beta}}_m^\top - \boldsymbol{\beta}_m^\top\end{pmatrix}\begin{pmatrix}\mathbf{0} & \boldsymbol{I} \\ \boldsymbol{I} & \mathbf{0}\end{pmatrix}\begin{pmatrix}\widehat{\boldsymbol{\alpha}}_a - \boldsymbol{\alpha}_a \\ \widehat{\boldsymbol{\beta}}_m - \boldsymbol{\beta}_m\end{pmatrix}.$$

In this expression 0 is a $p_m \times p_m$ matrix of zeros and I is a $p_m \times p_m$ identity matrix. Because this is quadratic form of an asymptotically normal random vector (where the matrix in the quadratic form is a symmetric matrix), then we may apply the continuous mapping theorem for convergence in distribution. That is, we just need to work with the asymptotic distribution when determining the asymptotic distribution of the quadratic form. Define

$$\boldsymbol{Z} = \sqrt{n}\begin{pmatrix}\sigma_a^{-2}\boldsymbol{\Sigma}_m & \mathbf{0} \\ \mathbf{0} & \sigma_e^2\boldsymbol{\Sigma}_m^{-1}\end{pmatrix}^{-1/2}\begin{pmatrix}\widehat{\boldsymbol{\alpha}}_a - \boldsymbol{\alpha}_a \\ \widehat{\boldsymbol{\beta}}_m - \boldsymbol{\beta}_m\end{pmatrix},$$

where

$$\begin{pmatrix}\sigma_a^{-2}\boldsymbol{\Sigma}_m & \mathbf{0} \\ \mathbf{0} & \sigma_e^2\boldsymbol{\Sigma}_m^{-1}\end{pmatrix}^{-1/2},$$

is the inverse of the matrix square root of

$$\begin{pmatrix}\sigma_a^{-2}\boldsymbol{\Sigma}_m & \mathbf{0} \\ \mathbf{0} & \sigma_e^2\boldsymbol{\Sigma}_m^{-1}\end{pmatrix}$$

Note that the matrix square root always exists for symmetric, positive definite matrices. Then we can write

$$\frac{n}{2}\begin{pmatrix}\widehat{\boldsymbol{\alpha}}_a^\top - \boldsymbol{\alpha}_a^\top & \widehat{\boldsymbol{\beta}}_m^\top - \boldsymbol{\beta}_m^\top\end{pmatrix}\begin{pmatrix}\mathbf{0} & \boldsymbol{I} \\ \boldsymbol{I} & \mathbf{0}\end{pmatrix}\begin{pmatrix}\widehat{\boldsymbol{\alpha}}_a - \boldsymbol{\alpha}_a \\ \widehat{\boldsymbol{\beta}}_m - \boldsymbol{\beta}_m\end{pmatrix}$$

$$= \frac{1}{2}\boldsymbol{Z}^\top\begin{pmatrix}\sigma_a^{-2}\boldsymbol{\Sigma}_m & \mathbf{0} \\ \mathbf{0} & \sigma_e^2\boldsymbol{\Sigma}_m^{-1}\end{pmatrix}^{1/2}\begin{pmatrix}\mathbf{0} & \boldsymbol{I} \\ \boldsymbol{I} & \mathbf{0}\end{pmatrix}\begin{pmatrix}\sigma_a^{-2}\boldsymbol{\Sigma}_m & \mathbf{0} \\ \mathbf{0} & \sigma_e^2\boldsymbol{\Sigma}_m^{-1}\end{pmatrix}^{1/2}\boldsymbol{Z}$$

Because

$$\begin{pmatrix}\sigma_a^{-2}\boldsymbol{\Sigma}_m & \mathbf{0} \\ \mathbf{0} & \sigma_e^2\boldsymbol{\Sigma}_m^{-1}\end{pmatrix}^{1/2}\begin{pmatrix}\mathbf{0} & \boldsymbol{I} \\ \boldsymbol{I} & \mathbf{0}\end{pmatrix}\begin{pmatrix}\sigma_a^{-2}\boldsymbol{\Sigma}_m & \mathbf{0} \\ \mathbf{0} & \sigma_e^2\boldsymbol{\Sigma}_m^{-1}\end{pmatrix}^{1/2} = \begin{pmatrix}\mathbf{0} & \frac{\sqrt{\sigma_e^2}}{\sqrt{\sigma_a^2}}\boldsymbol{I} \\ \frac{\sqrt{\sigma_e^2}}{\sqrt{\sigma_a^2}}\boldsymbol{I} & \mathbf{0}\end{pmatrix}$$

is a symmetric matrix, then we can obtain an eigendecomposition

$$\begin{pmatrix}\mathbf{0} & \frac{\sqrt{\sigma_e^2}}{\sqrt{\sigma_a^2}}\boldsymbol{I} \\ \frac{\sqrt{\sigma_e^2}}{\sqrt{\sigma_a^2}}\boldsymbol{I} & \mathbf{0}\end{pmatrix} = \boldsymbol{P}^\top\boldsymbol{\Lambda}\boldsymbol{P},$$

where $\boldsymbol{P}$ is an orthogonal matrix and $\boldsymbol{\Lambda}$ is a diagonal matrix. Then the expression becomes

$$\frac{1}{2}(\boldsymbol{P}\boldsymbol{Z})^\top\boldsymbol{\Lambda}(\boldsymbol{P}\boldsymbol{Z})$$

Note that here

$$\boldsymbol{Z} \to_d N(\boldsymbol{0}, \boldsymbol{I})$$

and because $\boldsymbol{P}$ is an orthogonal transformation of $\boldsymbol{Z}$, then

$$\boldsymbol{P}\boldsymbol{Z} \to_d N(\boldsymbol{0}, \boldsymbol{I})$$

Therefore, when $\boldsymbol{\alpha}_a = \boldsymbol{\beta}_m = \boldsymbol{0}$, we conclude that

$$n\Big(\widehat{\boldsymbol{\alpha}}_a^\top \widehat{\boldsymbol{\beta}}_m - \boldsymbol{\alpha}_a^\top \boldsymbol{\beta}_m\Big) \to_d \frac{1}{2} \sum_{j=1}^{2p_m} \lambda_j \chi_1^2,$$

where the λ_j are the eigenvalues of

$$\sqrt{\frac{\sigma_e^2}{\sigma_a^2}} \begin{pmatrix} \boldsymbol{0} & \boldsymbol{I} \\ \boldsymbol{I} & \boldsymbol{0} \end{pmatrix},$$

which has p_m eigenvalues equal to $\sqrt{\sigma_e^2/\sigma_a^2}$ and p_m eigenvalues equal to $-\sqrt{\sigma_e^2/\sigma_a^2}$. Therefore,

$$\frac{1}{2} \sum_{j=1}^{2p_m} \lambda_j \chi_1^2 = \frac{1}{2} \sqrt{\frac{\sigma_e^2}{\sigma_a^2}} \Big(\chi_{p_m}^2 - \chi_{p_m}^2\Big),$$

since the χ_1^2 random variables are independent.

Proof of Theorem 4

Suppose that $[\boldsymbol{M}_{i\cdot}^\top \mid A_i] \sim N(A_i\boldsymbol{\alpha}_a, \boldsymbol{\Sigma}_m)$ and $[Y_i \mid \boldsymbol{M}_{i\cdot}, A_i] \sim N(\boldsymbol{M}_{i\cdot}\boldsymbol{\beta}_m + A_i\beta_a, \sigma_e^2)$ is the true generative model where $i = 1, \ldots, n$. That is, without loss of generality, we are assuming that there are no confounders. This implies that $\sigma_a^2 = \text{Var}(A_i)$. Moreover, assume that $E[A_i] = 0$. Let $\mathcal{D} = \{\boldsymbol{Y}, \boldsymbol{M}, \boldsymbol{A}, \boldsymbol{C}\}$ denote the data. Suppose that we fit the soft constraint model to our data with a fixed value s^2. Then the likelihood is:

$$L(\boldsymbol{\alpha}_a, \boldsymbol{\Sigma}_M, \boldsymbol{\beta}_m, \sigma_e^2 \mid \boldsymbol{Y}, \boldsymbol{M}, \boldsymbol{A}) = \int_{-\infty}^{\infty} \pi(\boldsymbol{Y} \mid \boldsymbol{M}, \boldsymbol{A}, \widetilde{\theta}_a^I)\pi(\boldsymbol{M} \mid \boldsymbol{A})\pi(\widetilde{\theta}_a^I)d\widetilde{\theta}_a^I$$

$$= \frac{\pi(\boldsymbol{M} \mid \boldsymbol{A})}{(2\pi\sigma_e^2)^{n/2}\sqrt{2\pi s^2\widehat{\text{Var}}(\widehat{\theta}_a^E)}} \exp\left[-\frac{1}{2\sigma_e^2}\big(\boldsymbol{Y} - \boldsymbol{M}\boldsymbol{\beta}_m + \boldsymbol{\alpha}_a^\top\boldsymbol{\beta}_m\boldsymbol{A}\big)^\top\big(\boldsymbol{Y} - \boldsymbol{M}\boldsymbol{\beta}_m + \boldsymbol{\alpha}_a^\top\boldsymbol{\beta}_m\boldsymbol{A}\big)\right]$$

$$\times \exp\left[-\frac{(\widehat{\theta}_a^E)^2}{2s^2\widehat{\text{Var}}(\widehat{\theta}_a^E)}\right] \exp\left[\frac{1}{2}\left(\frac{\boldsymbol{A}^\top\boldsymbol{A}}{\sigma_e^2} + \frac{1}{s^2\widehat{\text{Var}}(\widehat{\theta}_a^E)}\right)^{-1}\left(\frac{\boldsymbol{A}^\top\big(\boldsymbol{Y} - \boldsymbol{M}\boldsymbol{\beta}_m + \boldsymbol{\alpha}_a^\top\boldsymbol{\beta}_m\boldsymbol{A}\big)}{\sigma_e^2} + \frac{\widehat{\theta}_a^E}{s^2\widehat{\text{Var}}(\widehat{\theta}_a^E)}\right)^2\right]$$

$$\times\sqrt{2\pi\left(\frac{\boldsymbol{A}^\top\boldsymbol{A}}{\sigma_e^2} + \frac{1}{s^2\widehat{\text{Var}}(\widehat{\theta}_a^E)}\right)^{-1}}$$

Therefore, the log-likelihood function is:

$$l(\boldsymbol{\alpha}_a, \boldsymbol{\Sigma}_M, \boldsymbol{\beta}_m, \sigma_e^2 \mid \boldsymbol{Y}, \boldsymbol{M}, \boldsymbol{A}) = -\frac{np_m}{2} - \frac{n}{2}\log|\boldsymbol{\Sigma}_m| - \frac{1}{2}\sum_{i=1}^{n}\big(\boldsymbol{M}_{i\cdot}^\top - A_i\boldsymbol{\alpha}_a\big)^\top\boldsymbol{\Sigma}_m^{-1}\big(\boldsymbol{M}_{i\cdot}^\top - A_i\boldsymbol{\alpha}_a\big)$$

$$-\frac{n}{2}\log\big(2\pi\sigma_e^2\big) - \frac{1}{2}\log\big(2\pi s^2\big) - \frac{1}{2\sigma_e^2}\big(\boldsymbol{Y} - \boldsymbol{M}\boldsymbol{\beta}_m + \boldsymbol{\alpha}_a^\top\boldsymbol{\beta}_m\boldsymbol{A}\big)^\top\big(\boldsymbol{Y} - \boldsymbol{M}\boldsymbol{\beta}_m + \boldsymbol{\alpha}_a^\top\boldsymbol{\beta}_m\boldsymbol{A}\big) - \frac{(\widehat{\theta}_a^E)^2}{2s^2\widehat{\text{Var}}(\widehat{\theta}_a^E)}$$

$$+\frac{1}{2}\left(\frac{\boldsymbol{A}^\top\boldsymbol{A}}{\sigma_e^2} + \frac{1}{s^2\widehat{\text{Var}}(\widehat{\theta}_a^E)}\right)^{-1}\left(\frac{\boldsymbol{A}^\top\big(\boldsymbol{Y} - \boldsymbol{M}\boldsymbol{\beta}_m + \boldsymbol{\alpha}_a^\top\boldsymbol{\beta}_m\boldsymbol{A}\big)}{\sigma_e^2} + \frac{\widehat{\theta}_a^E}{s^2\widehat{\text{Var}}(\widehat{\theta}_a^E)}\right)^2$$

$$+\frac{1}{2}\log(2\pi) - \frac{1}{2}\log\left(\frac{\boldsymbol{A}^\top\boldsymbol{A}}{\sigma_e^2} + \frac{1}{s^2\widehat{\text{Var}}(\widehat{\theta}_a^E)}\right)$$

The first order derivatives are:

$$\frac{\partial l}{\partial\boldsymbol{\alpha}_a} = \sum_{i=1}^{n} A_i\boldsymbol{\Sigma}_m^{-1}\big(\boldsymbol{M}_{i\cdot}^\top - A_i\boldsymbol{\alpha}_a\big) - \frac{1}{\sigma_e^2}\boldsymbol{A}^\top\big(\boldsymbol{Y} - \boldsymbol{M}\boldsymbol{\beta}_m + \boldsymbol{\alpha}_a^\top\boldsymbol{\beta}_m\boldsymbol{A}\big)\boldsymbol{\beta}_m$$

$$+\left(\frac{\boldsymbol{A}^\top\boldsymbol{A}}{\sigma_e^2} + \frac{1}{s^2\widehat{\text{Var}}(\widehat{\theta}_a^E)}\right)^{-1}\left(\frac{\boldsymbol{A}^\top\big(\boldsymbol{Y} - \boldsymbol{M}\boldsymbol{\beta}_m + \boldsymbol{\alpha}_a^\top\boldsymbol{\beta}_m\boldsymbol{A}\big)}{\sigma_e^2} + \frac{\widehat{\theta}_a^E}{s^2\widehat{\text{Var}}(\widehat{\theta}_a^E)}\right)\frac{\boldsymbol{A}^\top\boldsymbol{A}}{\sigma_e^2}\boldsymbol{\beta}_m$$

$$\frac{\partial l}{\partial\boldsymbol{\beta}_m} = \frac{1}{\sigma_e^2}\big(\boldsymbol{M} - \boldsymbol{A}\boldsymbol{\alpha}_a^\top\big)^\top\big(\boldsymbol{Y} - \boldsymbol{M}\boldsymbol{\beta}_m + \boldsymbol{\alpha}_a^\top\boldsymbol{\beta}_m\boldsymbol{A}\big)$$

$$-\left(\frac{\boldsymbol{A}^\top\boldsymbol{A}}{\sigma_e^2} + \frac{1}{s^2\widehat{\text{Var}}(\widehat{\theta}_a^E)}\right)^{-1}\left(\frac{\boldsymbol{A}^\top\big(\boldsymbol{Y} - \boldsymbol{M}\boldsymbol{\beta}_m + \boldsymbol{\alpha}_a^\top\boldsymbol{\beta}_m\boldsymbol{A}\big)}{\sigma_e^2} + \frac{\widehat{\theta}_a^E}{s^2\widehat{\text{Var}}(\widehat{\theta}_a^E)}\right)\frac{\big(\boldsymbol{M} - \boldsymbol{A}\boldsymbol{\alpha}_a^\top\big)^\top\boldsymbol{A}}{\sigma_e^2}$$

The second order derivatives are:

$$\frac{\partial^2 l}{\partial\boldsymbol{\alpha}_a\partial\boldsymbol{\alpha}_a^\top} = -\boldsymbol{A}^\top\boldsymbol{A}\left[\boldsymbol{\Sigma}_m^{-1} + \frac{1}{\sigma_e^2}\frac{1}{s^2\widehat{\text{Var}}(\widehat{\theta}_a^E)}\left(\frac{\boldsymbol{A}^\top\boldsymbol{A}}{\sigma_e^2} + \frac{1}{s^2\widehat{\text{Var}}(\widehat{\theta}_a^E)}\right)^{-1}\boldsymbol{\beta}_m\boldsymbol{\beta}_m^\top\right]$$

$$\implies -\frac{1}{n}E\left[\frac{\partial^2 l}{\partial \boldsymbol{\alpha}_a \partial \boldsymbol{\alpha}_a^\top}\right] \to \sigma_a^2\left[\boldsymbol{\Sigma}_m^{-1} + \frac{1}{\tau_a^2}\left(\frac{\sigma_a^2}{\sigma_e^2} + \frac{1}{\tau_a^2}\right)^{-1}\frac{1}{\sigma_e^2}\boldsymbol{\beta}_m\boldsymbol{\beta}_m^\top\right]$$

$$\frac{\partial^2 l}{\partial \boldsymbol{\alpha}_a \partial \boldsymbol{\beta}_m^\top} = \frac{\boldsymbol{A}^\top\boldsymbol{A}}{\sigma_e^2}\left(\frac{\boldsymbol{A}^\top\boldsymbol{A}}{\sigma_e^2} + \frac{1}{s^2\widehat{\text{Var}}(\widehat{\theta}_a^E)}\right)^{-1}\frac{\widehat{\theta}_a^E}{s^2\widehat{\text{Var}}(\widehat{\theta}_a^E)}\boldsymbol{I}$$

$$+\frac{1}{s^2\widehat{\text{Var}}(\widehat{\theta}_a^E)}\left(\frac{\boldsymbol{A}^\top\boldsymbol{A}}{\sigma_e^2} + \frac{1}{s^2\widehat{\text{Var}}(\widehat{\theta}_a^E)}\right)^{-1}\left[\frac{1}{\sigma_e^2}\boldsymbol{\beta}_m\boldsymbol{A}^\top\left(\boldsymbol{M} - \boldsymbol{A}\boldsymbol{\alpha}_a^\top\right) - \frac{1}{\sigma_e^2}\boldsymbol{A}^\top\left(\boldsymbol{Y} - \boldsymbol{M}\boldsymbol{\beta}_m + \boldsymbol{\alpha}_a^\top\boldsymbol{\beta}_m\boldsymbol{A}\right)\boldsymbol{I}\right]$$

$$\implies -\frac{1}{n}E\left[\frac{\partial^2 l}{\partial \boldsymbol{\alpha}_a \partial \boldsymbol{\beta}_m^\top}\right] = -\frac{1}{n}E\left[\frac{\partial^2 l}{\partial \boldsymbol{\beta}_m \partial \boldsymbol{\alpha}_a^\top}\right] \to \boldsymbol{0}$$

$$\frac{\partial^2 l}{\partial \boldsymbol{\beta}_m \partial \boldsymbol{\beta}_m^\top} = -\frac{1}{\sigma_e^2}\left(\boldsymbol{M}-\boldsymbol{A}\boldsymbol{\alpha}_a^\top\right)^\top\left(\boldsymbol{M}-\boldsymbol{A}\boldsymbol{\alpha}_a^\top\right) - \left(\frac{\boldsymbol{A}^\top\boldsymbol{A}}{\sigma_e^2} + \frac{1}{s^2\widehat{\text{Var}}(\widehat{\theta}_a^E)}\right)^{-1}\frac{1}{\sigma_e^2}\left(\boldsymbol{M}-\boldsymbol{A}\boldsymbol{\alpha}_a^\top\right)^\top\boldsymbol{A}\boldsymbol{A}^\top\left(\boldsymbol{M}-\boldsymbol{A}\boldsymbol{\alpha}_a^\top\right)^\top\frac{1}{\sigma_e^2}$$

$$\implies -\frac{1}{n}E\left[\frac{\partial^2 l}{\partial \boldsymbol{\beta}_m \partial \boldsymbol{\beta}_m^\top}\right] \to \frac{1}{\sigma_e^2}\boldsymbol{\Sigma}_m$$

Note that:

$$-\frac{1}{n}E\left[\frac{\partial^2 l}{\partial \boldsymbol{\alpha}_a \partial (\boldsymbol{\Sigma}_m^{-1})_{kl}}\right] \to \boldsymbol{0}, \;\; 1 \leq k, l \leq p_m$$

$$-\frac{1}{n}E\left[\frac{\partial^2 l}{\partial \boldsymbol{\beta}_m \partial (\boldsymbol{\Sigma}_m^{-1})}\right] = \boldsymbol{0}, \;\; 1 \leq k, l \leq p_m$$

$$-\frac{1}{n}E\left[\frac{\partial^2 l}{\partial \boldsymbol{\alpha}_a \partial (\sigma_e^2)}\right] = \boldsymbol{0}$$

$$-\frac{1}{n}E\left[\frac{\partial^2 l}{\partial \boldsymbol{\beta}_m \partial (\sigma_e^2)}\right] = \boldsymbol{0}$$

Then we conclude that

$$\left\{\mathcal{I}^S(\boldsymbol{\alpha}_a, \boldsymbol{\beta}_m)\right\}^{-1} = \begin{pmatrix} \frac{1}{\sigma_a^2}\left(\boldsymbol{\Sigma}_m^{-1} + \frac{1}{\tau_a^2}\left[\frac{\sigma_a^2}{\sigma_e^2} + \frac{1}{\tau_a^2}\right]^{-1}\frac{1}{\sigma_e^2}\boldsymbol{\beta}_m\boldsymbol{\beta}_m^\top\right)^{-1} & \boldsymbol{0} \\ \boldsymbol{0} & \sigma_e^2\boldsymbol{\Sigma}_m^{-1} \end{pmatrix}$$

The asymptotic distribution for the soft constraint estimator of the NIE can then be computed using the multivariate delta method provided that $\boldsymbol{\alpha}_a \neq \boldsymbol{0}$ or $\boldsymbol{\beta}_m \neq \boldsymbol{0}$.

To extend this result to the model with confounders we use the same trick that was used in the proof of Theorem 1, namely that the model with confounders is equivalent to a model with no confounders after the confounders are regressed out from the outcome, mediators, and exposure via linear regression models. Therefore, we again see that $\sigma_a^2 = \text{Var}(A_i \mid \boldsymbol{C}_{i\cdot})$, when confounders are present.

B.3 Supporting Figures and Tables

Visit	Covariate	Total	Preterm	Full-term	P-Value
1	Pre-Pregnancy BMI (kg/m^2)	25.9 (5.9)	26.4 (7.3)	25.9 (5.7)	0.607
	Maternal Age (years)	26.9 (5.5)	26.4 (6.0)	27.0 (5.4)	0.528
	Education				0.225
	GED/Equivalent or Less	97 (21.6)	16 (30.2)	81 (20.5)	
	Some College	154 (34.3)	18 (34.0)	136 (34.3)	
	Bachelor's Degree or Higher	198 (44.1)	19 (35.8)	179 (45.2)	
2	Pre-Pregnancy BMI (kg/m^2)	26.1 (6.0)	26.5 (7.2)	26.0 (5.8)	0.678
	Maternal Age (years)	26.8 (5.6)	26.3 (6.1)	26.9 (5.5)	0.525
	Education				0.634
	GED/Equivalent or Less	97 (21.3)	13 (25.0)	84 (20.8)	
	Some College	157 (34.4)	19 (36.5)	138 (34.2)	
	Bachelor's Degree or Higher	202 (44.3)	20 (38.5)	182 (45.0)	

Table B.1: Descriptive Statistics for subset of the PROTECT Cohort with at least one of MBP, MiBP, and MBzP measured at visit X and eicosanoid measures at visit 3. Sample size at visit 1 is 449 total participants (396 full-term deliveries and 53 preterm deliveries). Sample size at visit 2 is 456 total participants (404 full-term deliveries and 52 preterm deliveries). P-values corresponding to differences between preterm and full-term deliveries for continuous and categorical variables come from t-tests and chi-squared tests, respectively.

Method	Internal TE Model	External TE Model	Notes on Bias and Estimation Efficiency
Unconstrained	Correctly Specified	None	Estimators are unbiased, but they are also the least efficient when θ_a^I and θ_a^E are close. Use when θ_a^I and θ_a^E are known to be different.
Hard Constraint	Correctly Specified	Equation (4) Must Hold	Estimators may be asymptotically biased if $\theta_a^I \neq \theta_a^E$. Most asymptotically efficient method when $\theta_a^I = \theta_a^E$. Only consider using if $(\widehat{\theta}_a^E - \widehat{\theta}_a^I)^2 \leq \widehat{\text{Var}}(\widehat{\theta}_a^I)$.
Soft Constraint (EB)	Correctly Specified	None	Estimators are minimally biased regardless of how close θ_a^I and θ_a^E are. Asymptotically more efficient than the unconstrained estimators when θ_a^I and θ_a^E are close. As efficient as the unconstrained estimators when θ_a^I and θ_a^E are substantially different. Less efficient than the hard constraint estimators when θ_a^E and θ_a^I are close. Use this method as the default method.

Table B.2: Summary of the methods presented in the paper and when to use each. TE, total effect.

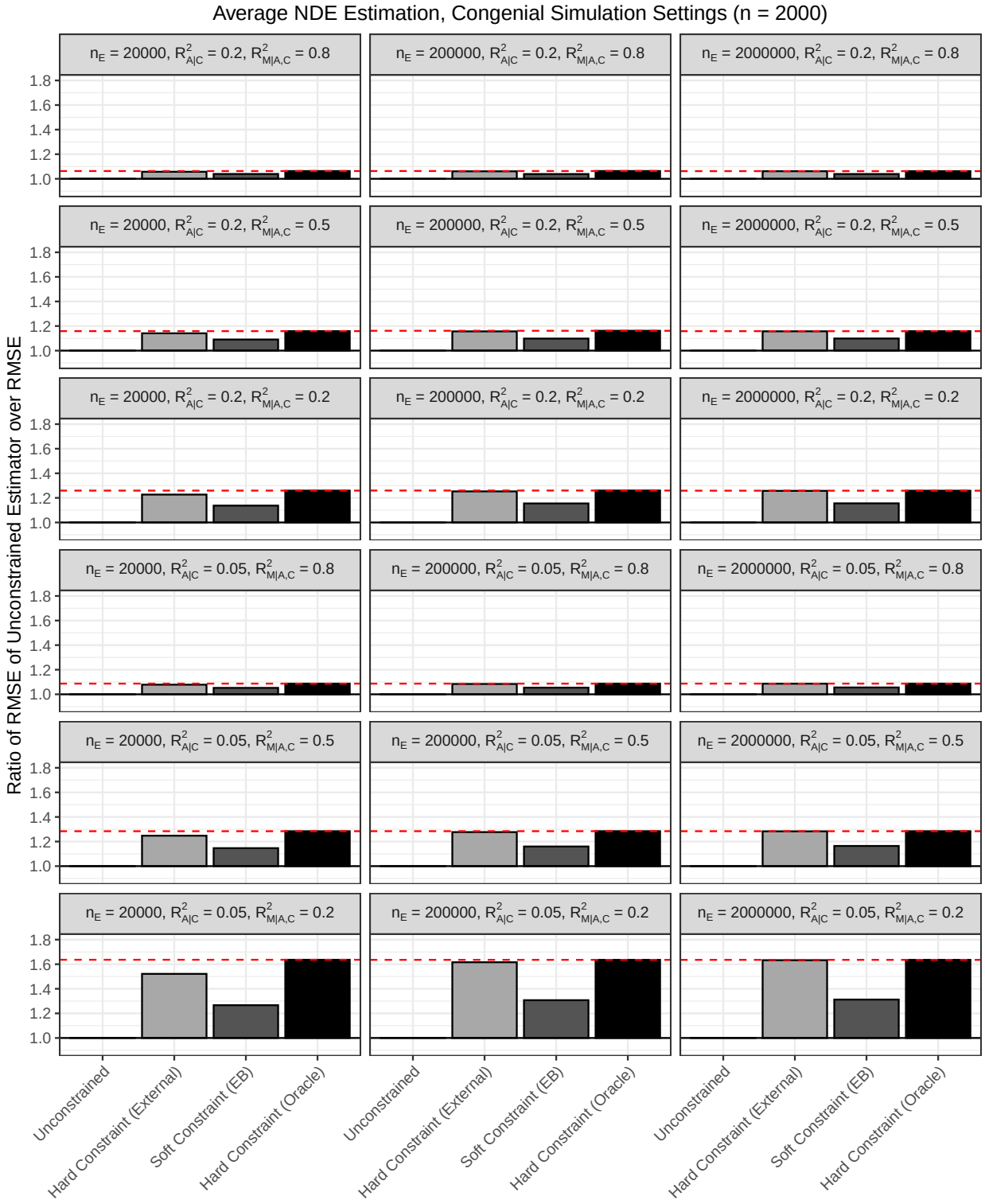


Figure B.1: Relative root mean-squared error (RMSE) corresponding to Natural Direct Effect (NDE) estimation for the congenial simulation scenarios ($n = 2000$). The red, horizontal dashed line indicates the upper bound on the possible gain in estimation efficiency, as determined by the hard constraint estimator with the oracle constraint.

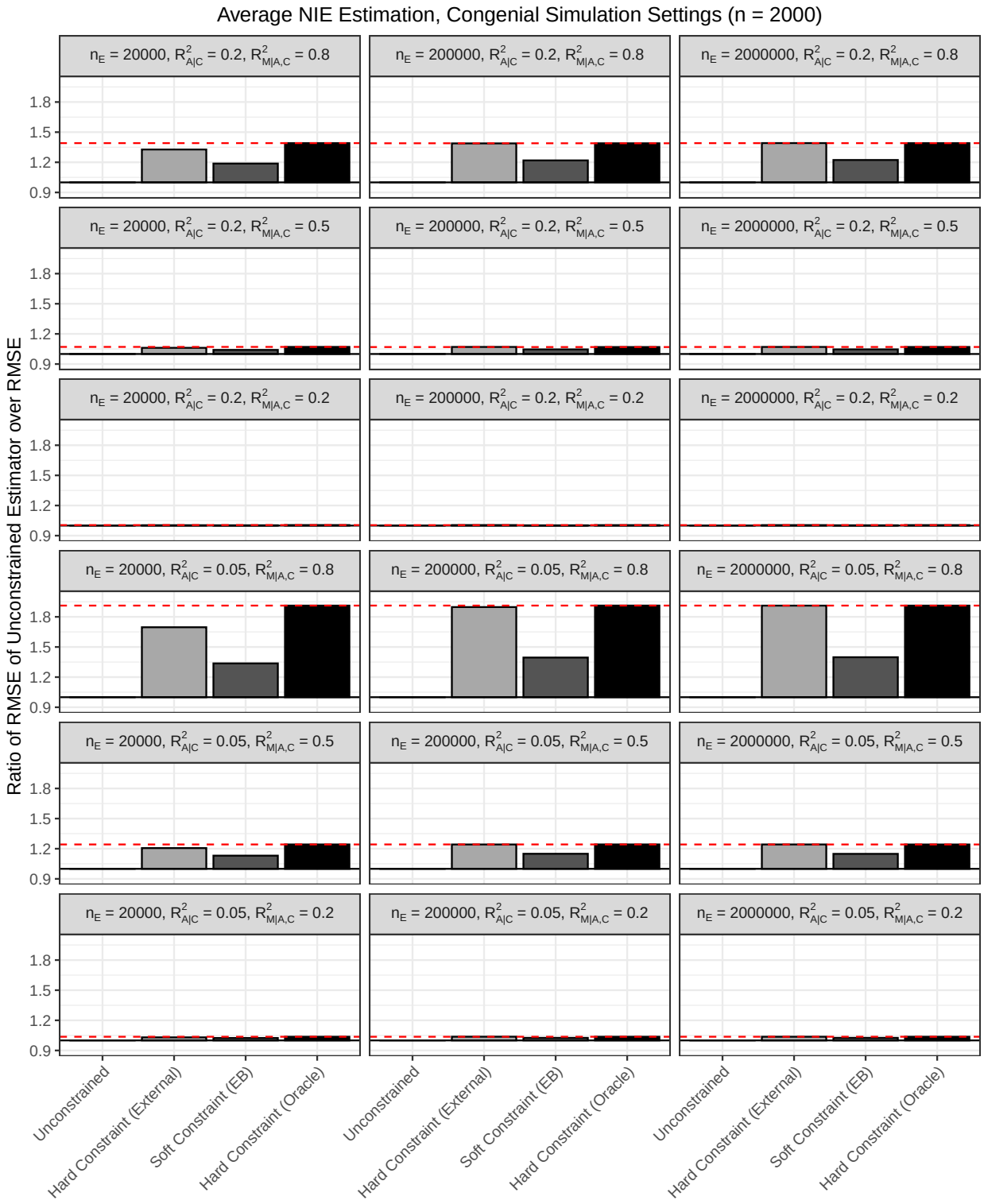


Figure B.2: Relative root mean-squared error (RMSE) corresponding to Natural Indirect Effect (NIE) estimation for the congenial simulation scenarios ($n = 2000$). The red, horizontal dashed line indicates the upper bound on the possible gain in estimation efficiency, as determined by the hard constraint estimator with the oracle constraint.

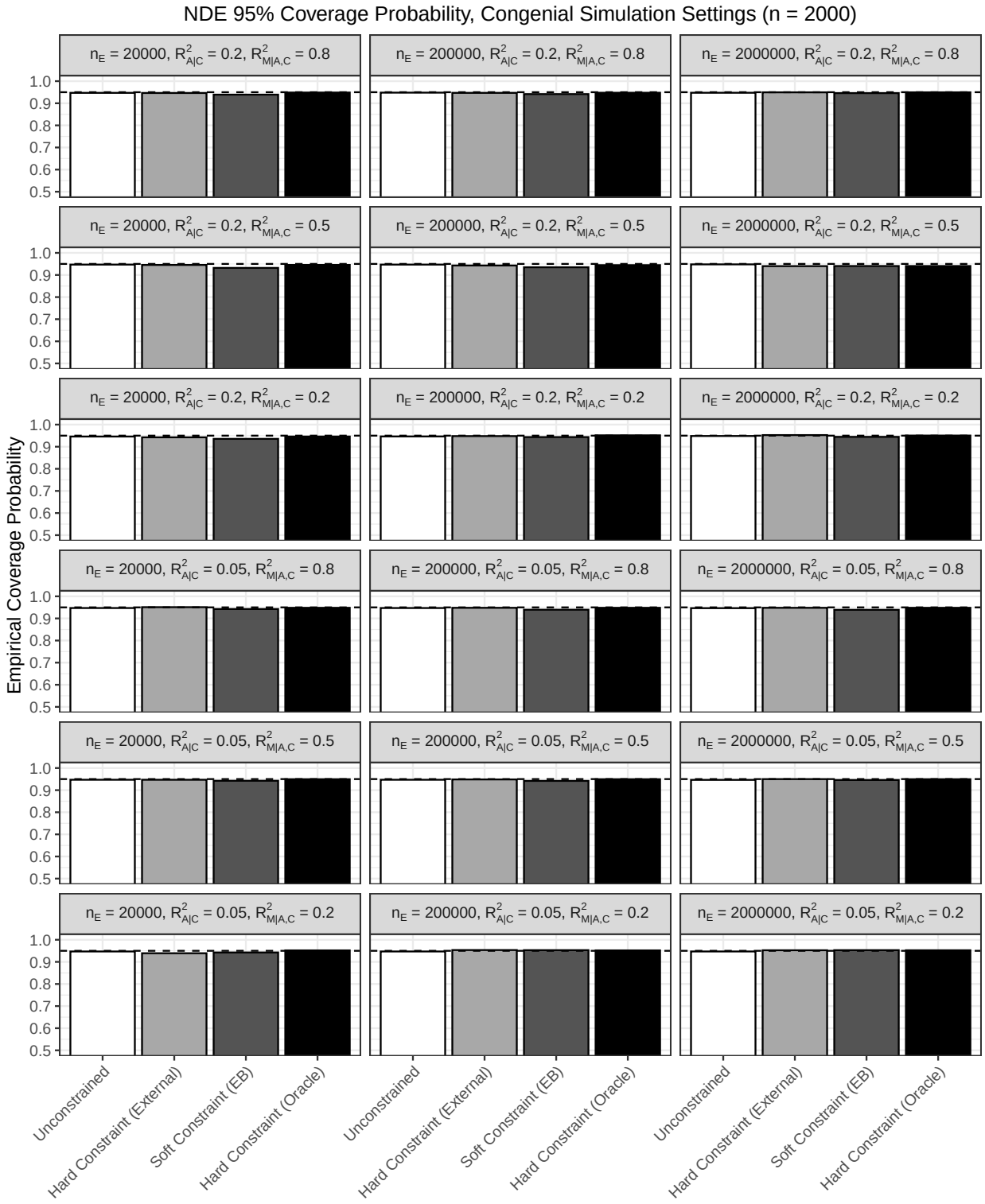


Figure B.3: Empirical coverage probability corresponding to Natural Direct Effect (NDE) estimation for the congenial simulation scenarios ($n = 2000$). The horizontal dashed line indicates the nominal coverage rate of 0.95.

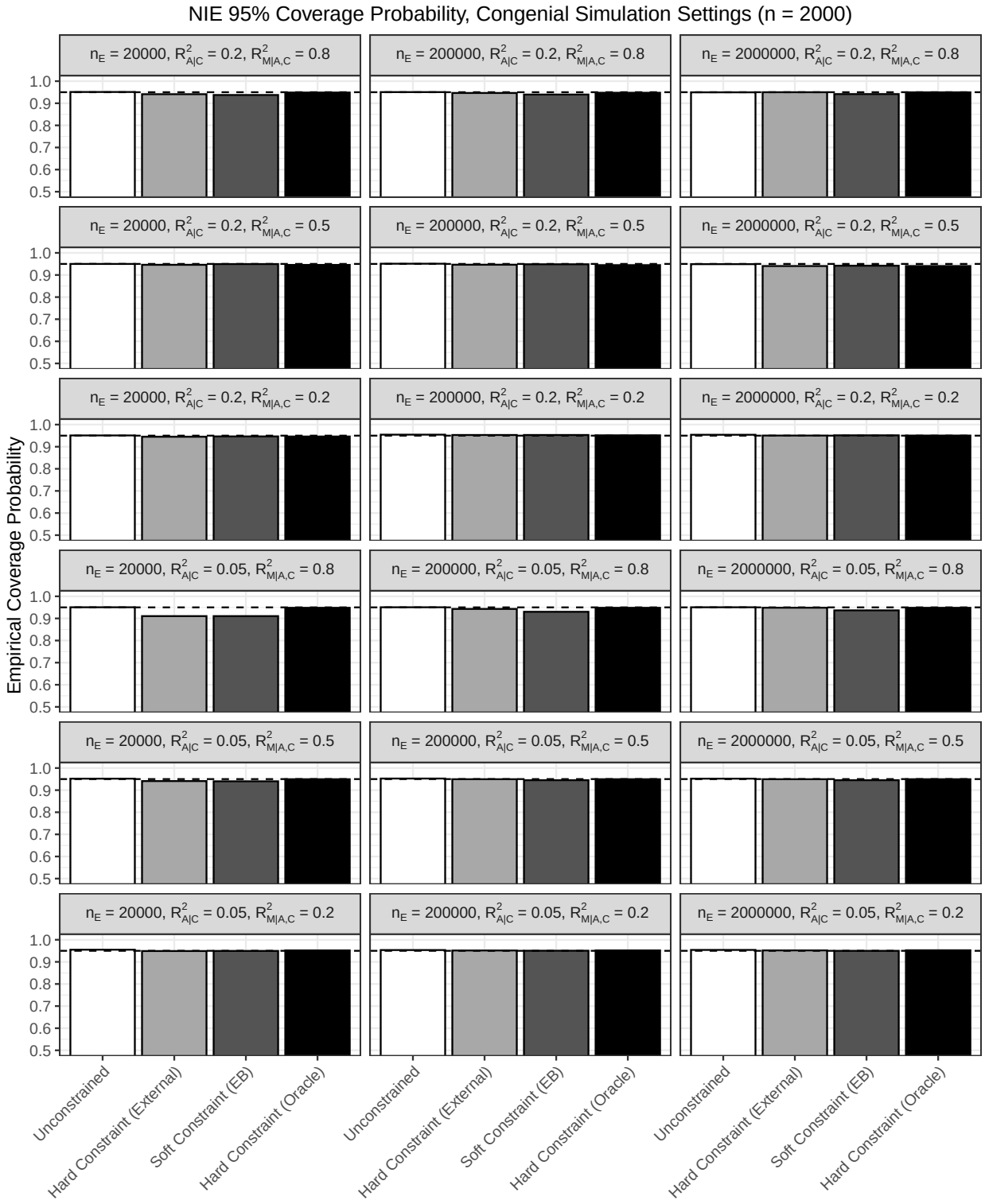


Figure B.4: Empirical coverage probability corresponding to Natural Indirect Effect (NIE) estimation for the congenial simulation scenarios ($n = 2000$). The horizontal dashed line indicates the nominal coverage rate of 0.95.

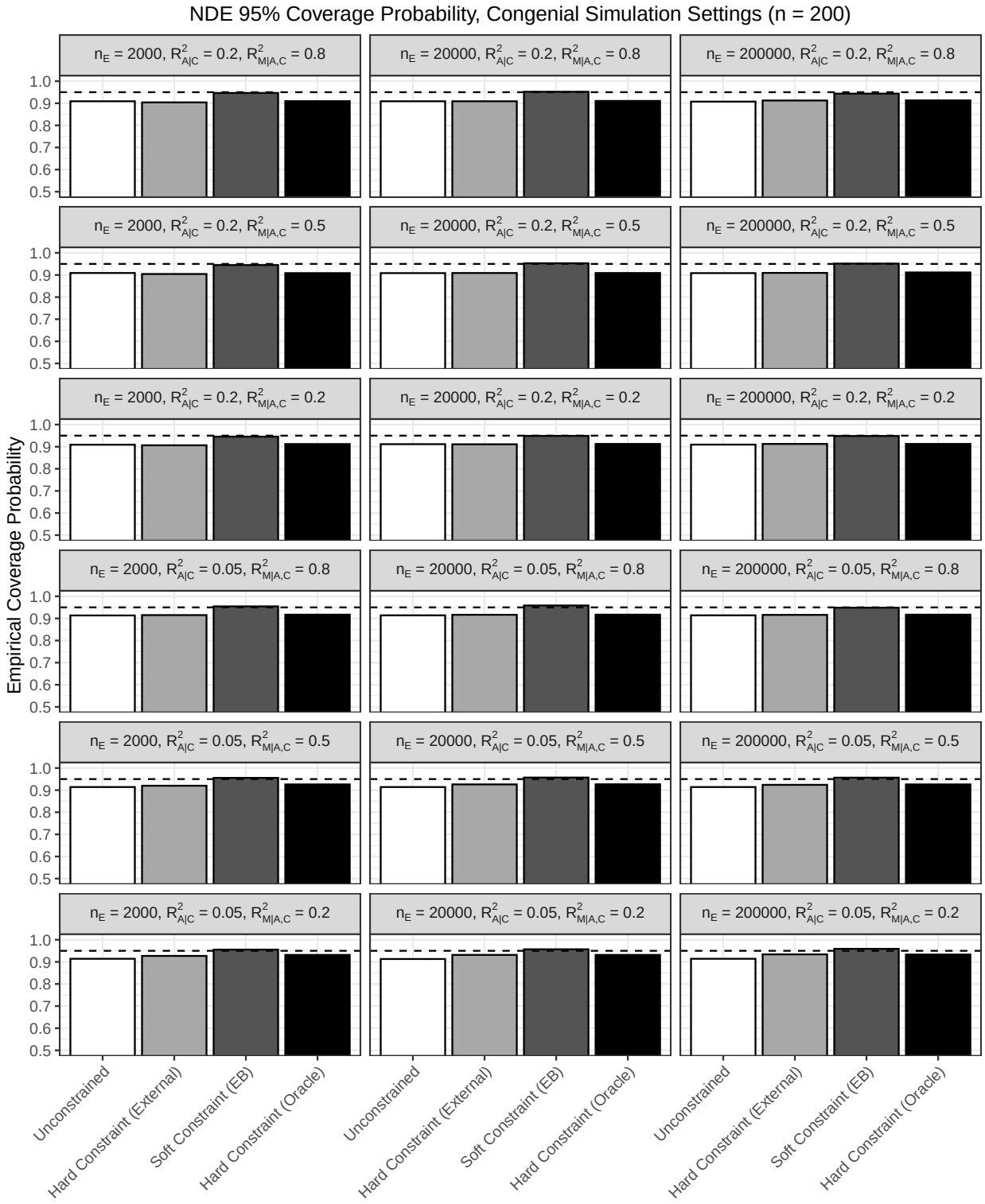


Figure B.5: Empirical coverage probability corresponding to Natural Direct Effect (NDE) estimation for the congenial simulation scenarios ($n = 200$). The horizontal dashed line indicates the nominal coverage rate of 0.95.

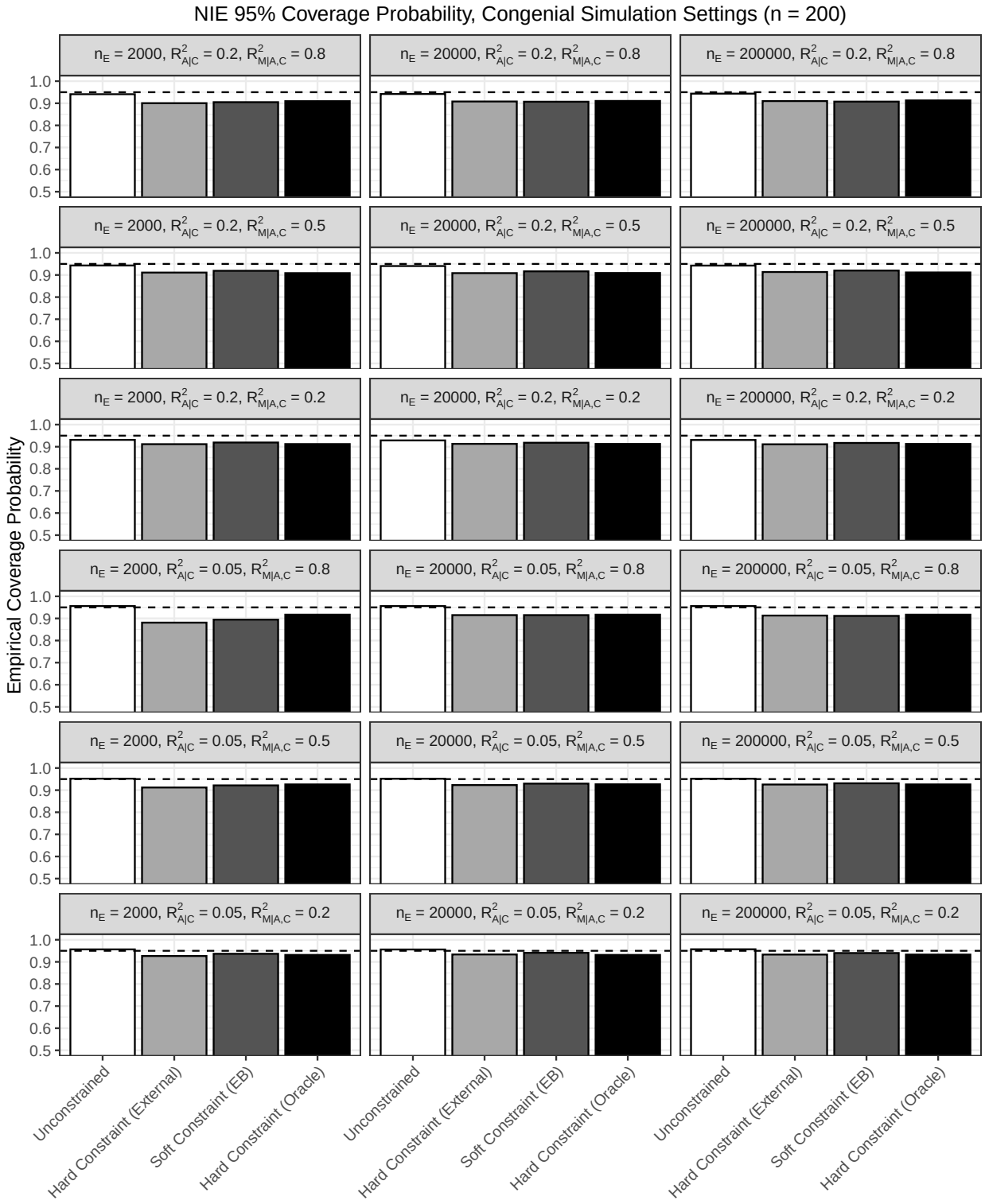


Figure B.6: Empirical coverage probability corresponding to Natural Indirect Effect (NIE) estimation for the congenial simulation scenarios ($n = 200$). The horizontal dashed line indicates the nominal coverage rate of 0.95.

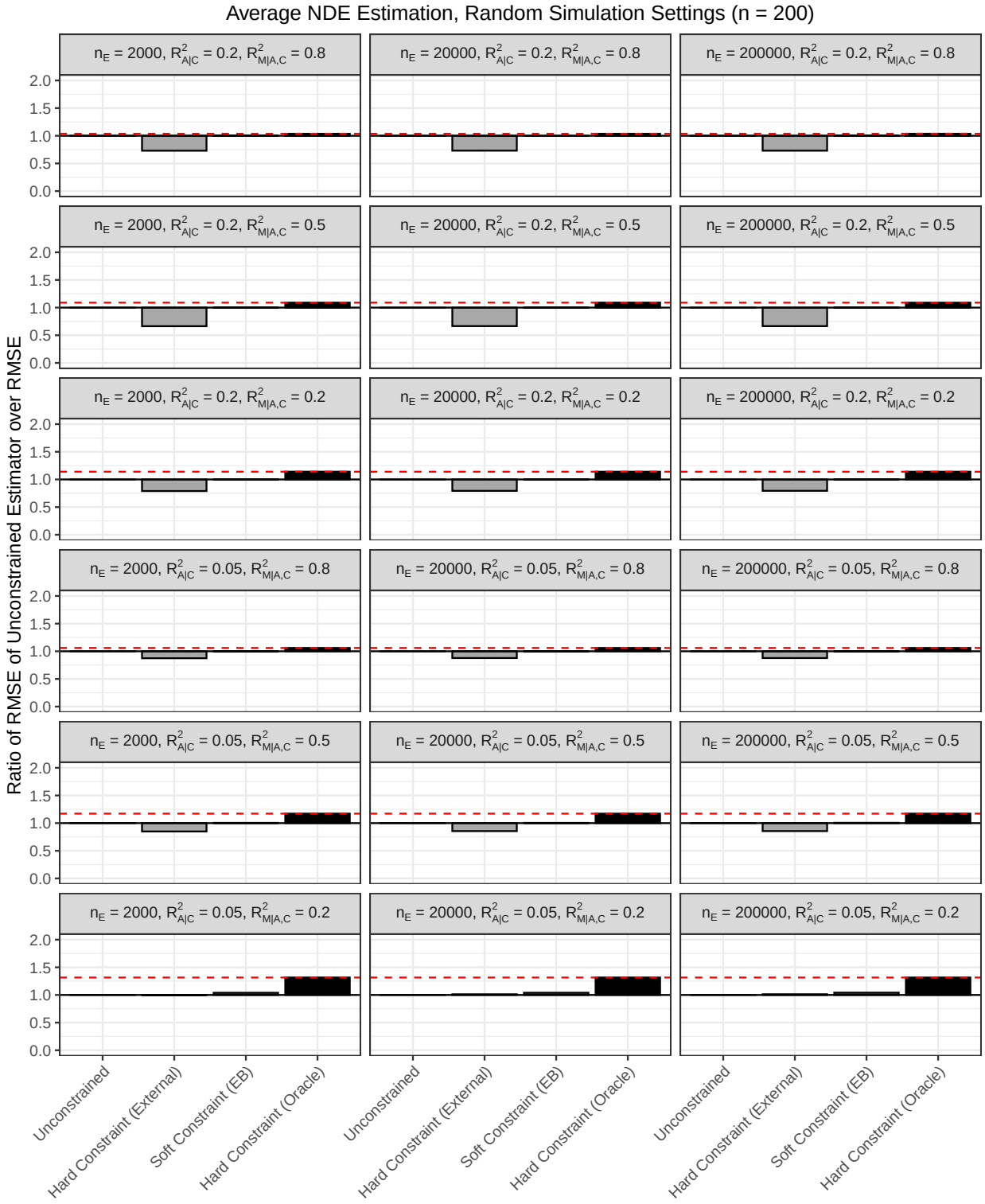


Figure B.7: Relative root mean-squared error (RMSE) corresponding to Natural Direct Effect (NDE) estimation for the random simulation scenarios ($n = 200$). The red, horizontal dashed line indicates the upper bound on the possible gain in estimation efficiency, as determined by the hard constraint estimator with the oracle constraint.

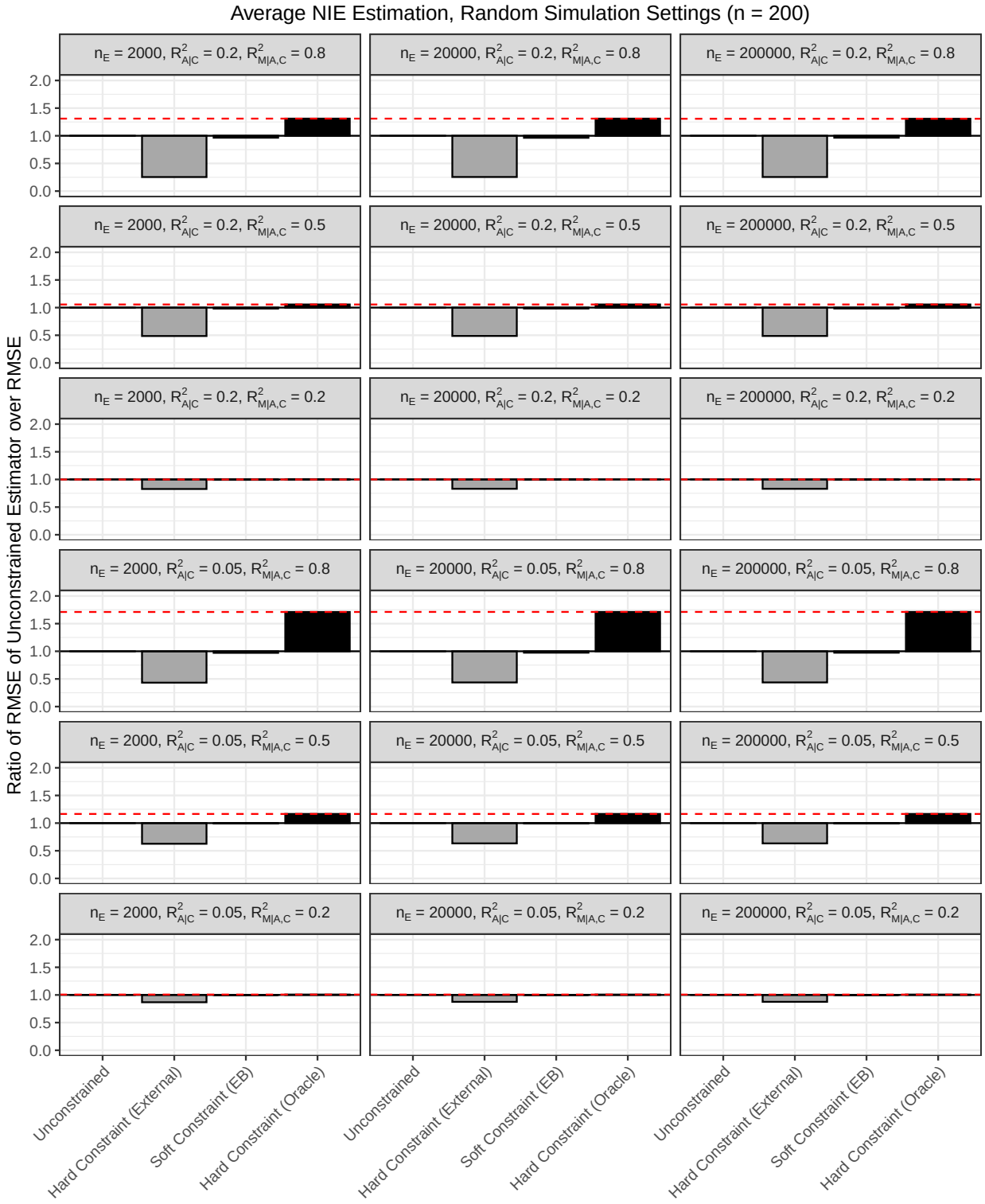


Figure B.8: Relative root mean-squared error (RMSE) corresponding to Natural Indirect Effect (NIE) estimation for the random simulation scenarios ($n = 200$). The red, horizontal dashed line indicates the upper bound on the possible gain in estimation efficiency, as determined by the hard constraint estimator with the oracle constraint.

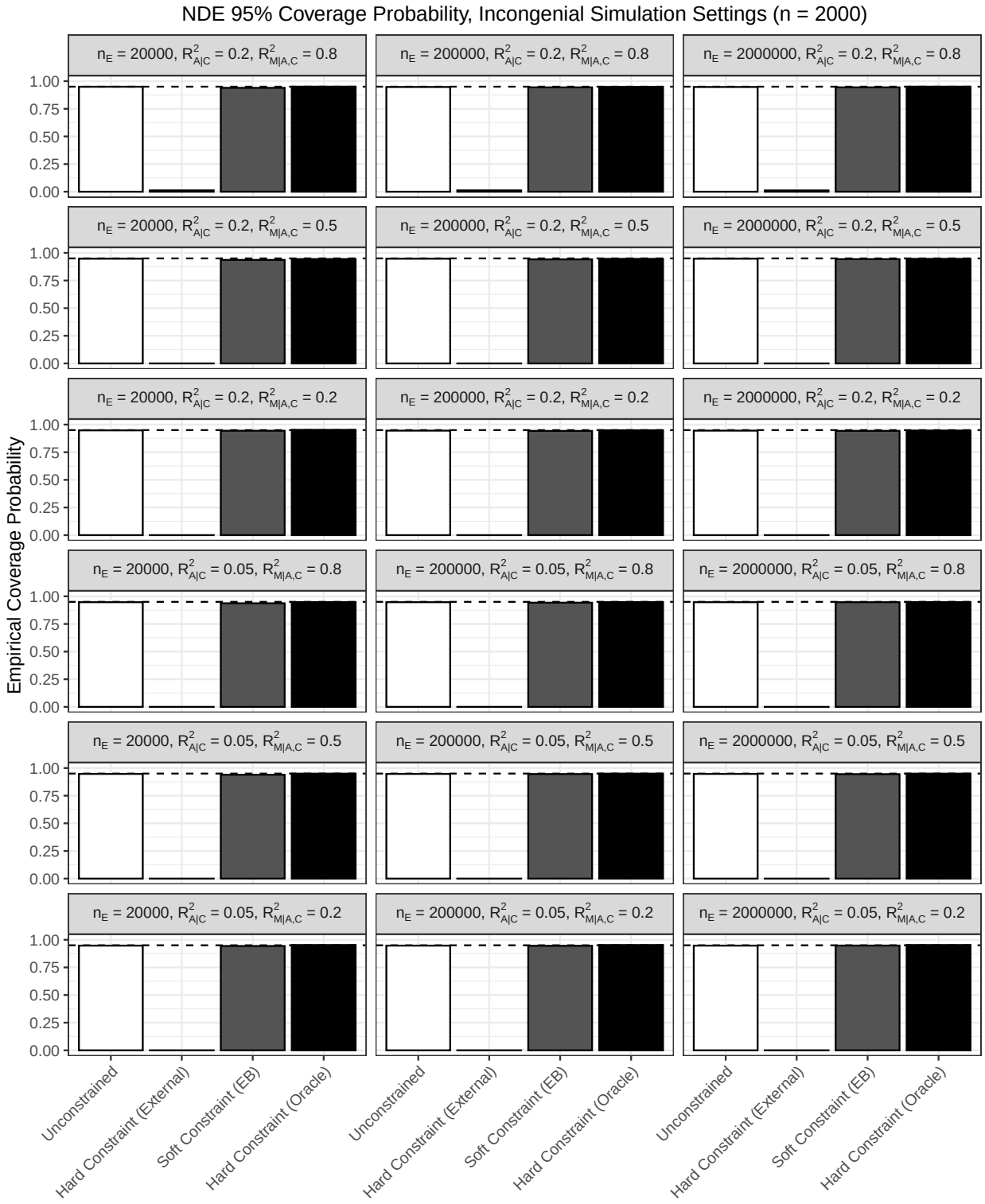


Figure B.9: Empirical coverage probability corresponding to Natural Direct Effect (NDE) estimation for the incongenial simulation scenarios ($n = 2000$). The horizontal dashed line indicates the nominal coverage rate of 0.95.

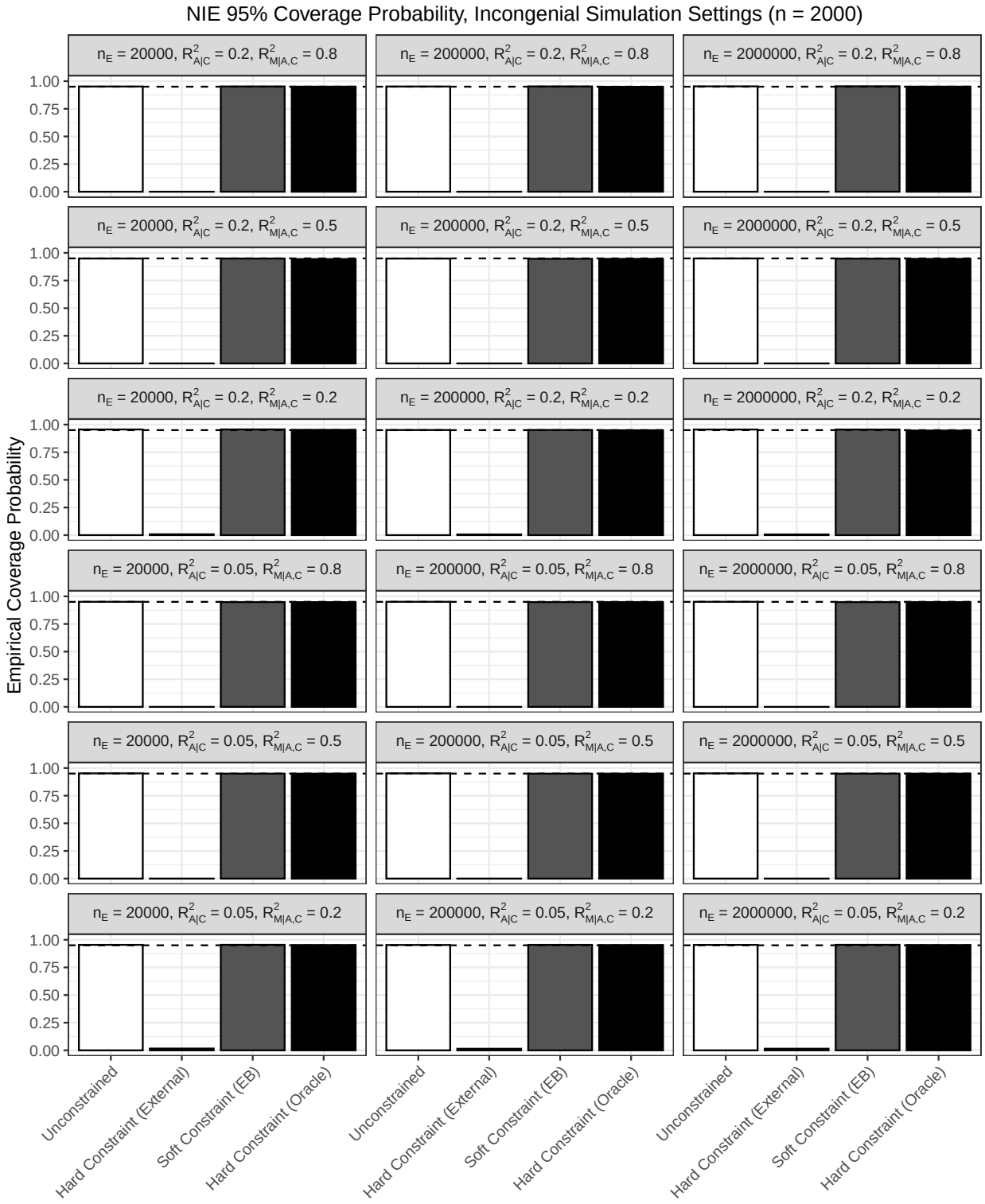


Figure B.10: Empirical coverage probability corresponding to Natural Indirect Effect (NIE) estimation for the incongenial simulation scenarios ($n = 2000$). The horizontal dashed line indicates the nominal coverage rate of 0.95.

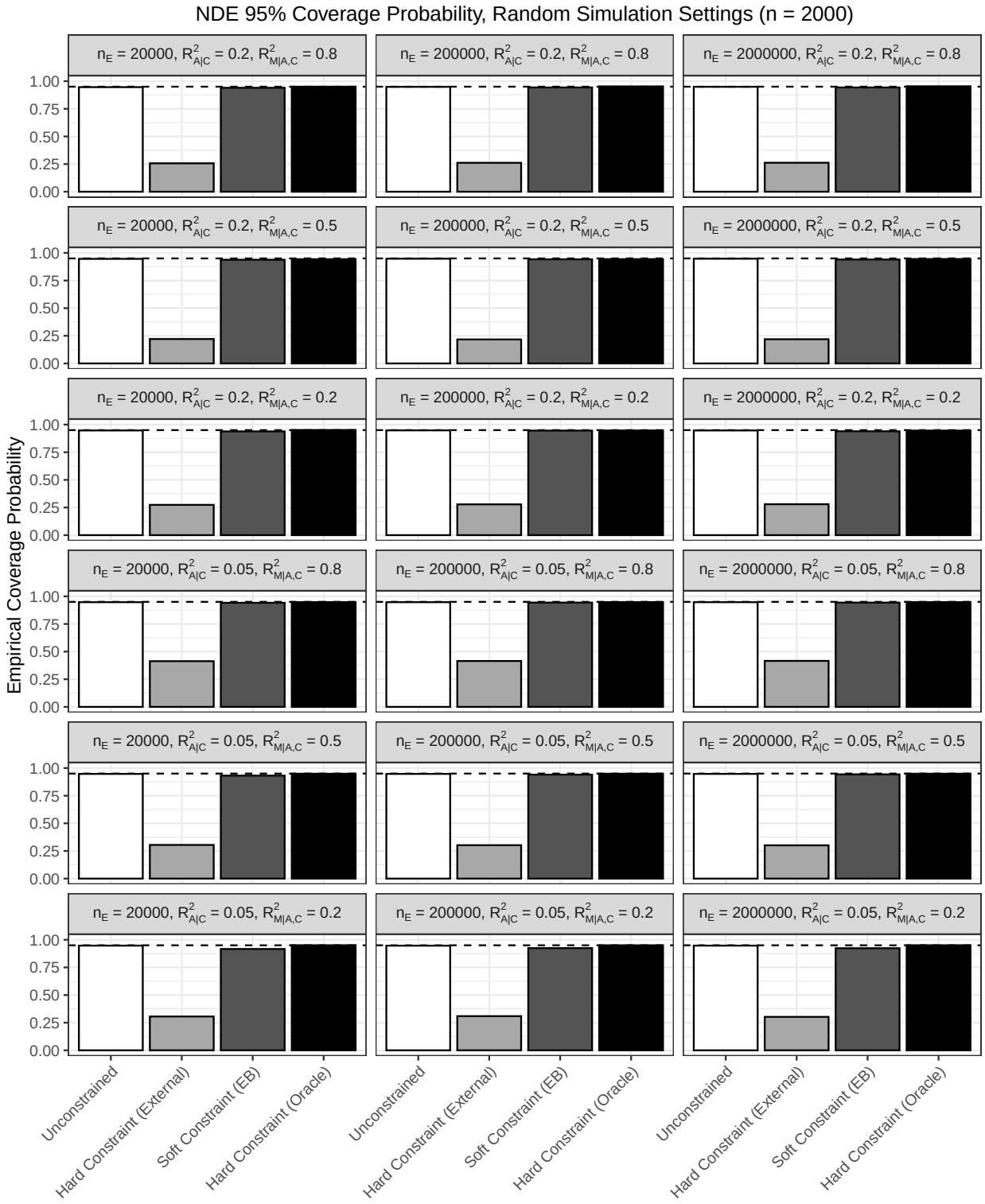


Figure B.11: Empirical coverage probability corresponding to Natural Direct Effect (NDE) estimation for the random simulation scenarios ($n = 2000$). The horizontal dashed line indicates the nominal coverage rate of 0.95.

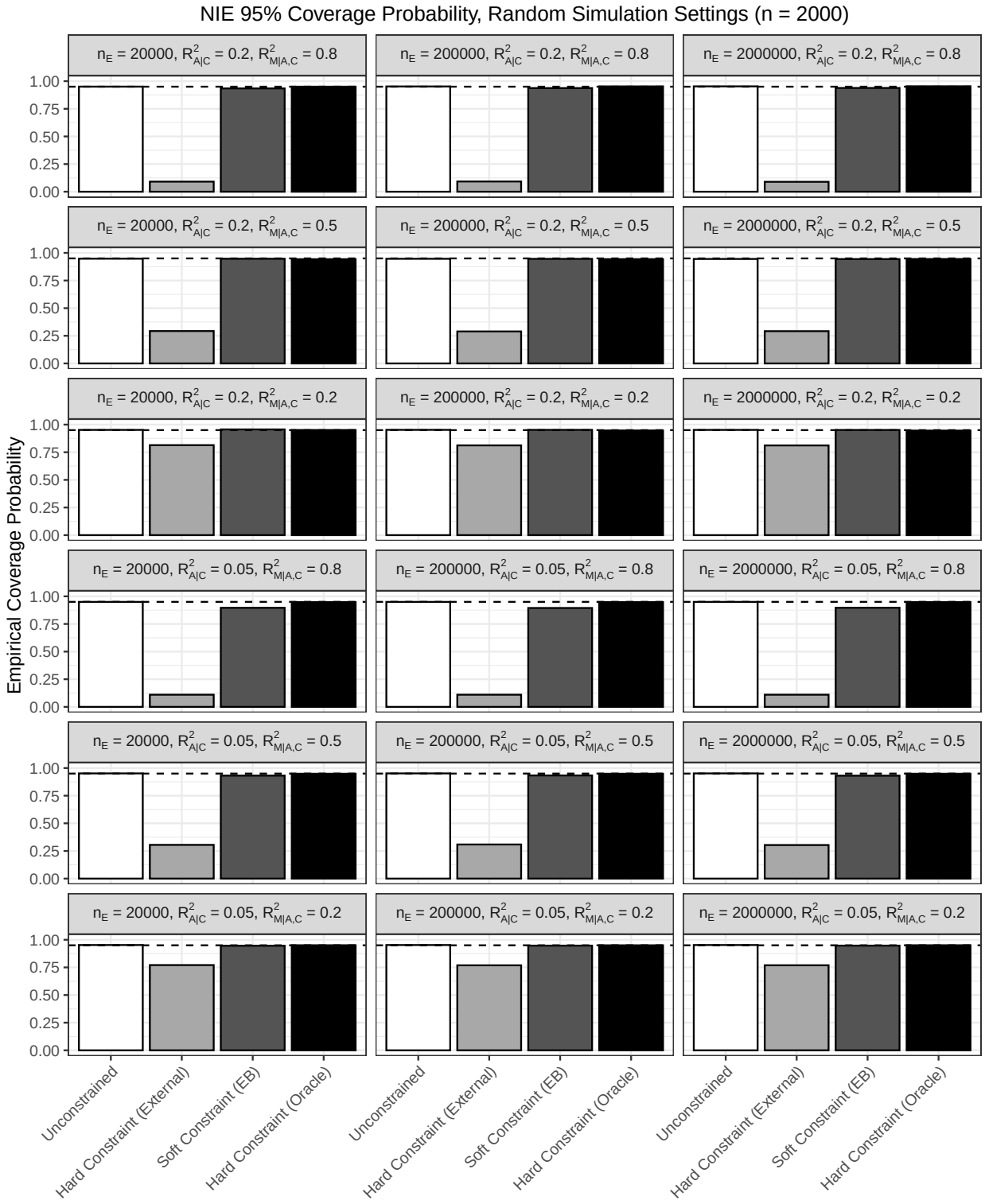


Figure B.12: Empirical coverage probability corresponding to Natural Indirect Effect (NIE) estimation for the random simulation scenarios ($n = 2000$). The horizontal dashed line indicates the nominal coverage rate of 0.95.

APPENDIX C

Supplement for Chapter 4

C.1 Proof of Rotational Invariance

Consider a rotational reparameterization of Model

$$\begin{aligned}
\big[Y_i \mid \boldsymbol{M}_{i\cdot}, \boldsymbol{L}_{i\cdot}, \boldsymbol{C}_{i\cdot}\big] &\sim N\big(\boldsymbol{M}_{i\cdot}\boldsymbol{\beta}_m + \boldsymbol{L}_{i\cdot}\boldsymbol{P}\boldsymbol{P}^\top\boldsymbol{\beta}_a + \boldsymbol{C}_{i\cdot}\boldsymbol{\beta}_c, \sigma_e^2\big) \\
\big[\boldsymbol{M}_{i\cdot}^\top \mid \boldsymbol{L}_{i\cdot}^\top, \boldsymbol{C}_{i\cdot}^\top\big] &\sim N\big(\boldsymbol{\alpha}_a\boldsymbol{P}\boldsymbol{P}^\top\boldsymbol{L}_{i\cdot}^\top + \boldsymbol{\alpha}_c\boldsymbol{C}_{i\cdot}^\top, \boldsymbol{\Sigma}_m\big) \\
\big[\boldsymbol{A}_{i\cdot}^\top \mid \boldsymbol{L}_{i\cdot}^\top\big] &\sim N\big(\boldsymbol{\Psi}\boldsymbol{P}\boldsymbol{P}^\top\boldsymbol{L}_{i\cdot}^\top, \boldsymbol{\Sigma}_L\big) \\
\boldsymbol{L}_{i\cdot}^\top &\sim N(\boldsymbol{0}, \boldsymbol{I}),
\end{aligned}$$

where $\boldsymbol{P}$ is an $r \times r$ matrix that satisfies $\boldsymbol{P}\boldsymbol{P}^\top = \boldsymbol{I}$. Then,

$$\begin{aligned}
\mathrm{DE}^A(\boldsymbol{a}, \boldsymbol{a}^*) &= (\boldsymbol{a} - \boldsymbol{a}^*)^\top\boldsymbol{\Sigma}_L^{-1}\boldsymbol{\Psi}\boldsymbol{P}\big(\boldsymbol{I} + \boldsymbol{P}^\top\boldsymbol{\Psi}^\top\boldsymbol{\Sigma}_L^{-1}\boldsymbol{\Psi}\boldsymbol{P}\big)^{-1}\boldsymbol{P}^\top\boldsymbol{\beta}_a \\
&= (\boldsymbol{a} - \boldsymbol{a}^*)^\top\boldsymbol{\Sigma}_L^{-1}\boldsymbol{\Psi}\boldsymbol{P}\big(\boldsymbol{P}^\top\boldsymbol{P} + \boldsymbol{P}^\top\boldsymbol{\Psi}^\top\boldsymbol{\Sigma}_L^{-1}\boldsymbol{\Psi}\boldsymbol{P}\big)^{-1}\boldsymbol{P}^\top\boldsymbol{\beta}_a \\
&= (\boldsymbol{a} - \boldsymbol{a}^*)^\top\boldsymbol{\Sigma}_L^{-1}\boldsymbol{\Psi}\boldsymbol{P}\boldsymbol{P}^{-1}\big(\boldsymbol{I} + \boldsymbol{\Psi}^\top\boldsymbol{\Sigma}_L^{-1}\boldsymbol{\Psi}\big)^{-1}(\boldsymbol{P}^\top)^{-1}\boldsymbol{P}^\top\boldsymbol{\beta}_a \\
&= (\boldsymbol{a} - \boldsymbol{a}^*)^\top\boldsymbol{\Sigma}_L^{-1}\boldsymbol{\Psi}\big(\boldsymbol{I} + \boldsymbol{\Psi}^\top\boldsymbol{\Sigma}_L^{-1}\boldsymbol{\Psi}\big)^{-1}\boldsymbol{\beta}_a
\end{aligned}$$

$$\begin{aligned}
\mathrm{IE}^A(\boldsymbol{a}, \boldsymbol{a}^*) &= (\boldsymbol{a} - \boldsymbol{a}^*)^\top\boldsymbol{\Sigma}_L^{-1}\boldsymbol{\Psi}\boldsymbol{P}\big(\boldsymbol{I} + \boldsymbol{P}^\top\boldsymbol{\Psi}^\top\boldsymbol{\Sigma}_L^{-1}\boldsymbol{\Psi}\boldsymbol{P}\big)^{-1}\boldsymbol{P}^\top\boldsymbol{\alpha}_a^\top\boldsymbol{\beta}_m \\
&= (\boldsymbol{a} - \boldsymbol{a}^*)^\top\boldsymbol{\Sigma}_L^{-1}\boldsymbol{\Psi}\boldsymbol{P}\big(\boldsymbol{P}^\top\boldsymbol{P} + \boldsymbol{P}^\top\boldsymbol{\Psi}^\top\boldsymbol{\Sigma}_L^{-1}\boldsymbol{\Psi}\boldsymbol{P}\big)^{-1}\boldsymbol{P}^\top\boldsymbol{\alpha}_a^\top\boldsymbol{\beta}_m \\
&= (\boldsymbol{a} - \boldsymbol{a}^*)^\top\boldsymbol{\Sigma}_L^{-1}\boldsymbol{\Psi}\boldsymbol{P}\boldsymbol{P}^{-1}\big(\boldsymbol{I} + \boldsymbol{\Psi}^\top\boldsymbol{\Sigma}_L^{-1}\boldsymbol{\Psi}\big)^{-1}(\boldsymbol{P}^\top)^{-1}\boldsymbol{P}^\top\boldsymbol{\alpha}_a^\top\boldsymbol{\beta}_m \\
&= (\boldsymbol{a} - \boldsymbol{a}^*)^\top\boldsymbol{\Sigma}_L^{-1}\boldsymbol{\Psi}\big(\boldsymbol{I} + \boldsymbol{\Psi}^\top\boldsymbol{\Sigma}_L^{-1}\boldsymbol{\Psi}\big)^{-1}\boldsymbol{\alpha}_a^\top\boldsymbol{\beta}_m
\end{aligned}$$

C.2 Full Conditional Updates

$$[\sigma_e^2 \mid \cdot] \sim IG\bigg(\frac{n+1}{2}, \frac{1}{2}(\boldsymbol{Y}-\boldsymbol{M}\boldsymbol{\beta}_m-\boldsymbol{L}\boldsymbol{\beta}_a-\boldsymbol{C}\boldsymbol{\beta}_c)^\top(\boldsymbol{Y}-\boldsymbol{M}\boldsymbol{\beta}_m-\boldsymbol{L}\boldsymbol{\beta}_a-\boldsymbol{C}\boldsymbol{\beta}_c)+\frac{1}{\nu_\beta}\bigg)$$

$$[\boldsymbol{\beta}_c \mid \cdot] \sim N\bigg(\big(\boldsymbol{C}^\top\boldsymbol{C}\big)^{-1}\boldsymbol{C}^\top(\boldsymbol{Y}-\boldsymbol{M}\boldsymbol{\beta}_m-\boldsymbol{L}\boldsymbol{\beta}_a), \sigma_e^2\big(\boldsymbol{C}^\top\boldsymbol{C}\big)^{-1}\bigg)$$

$$[\boldsymbol{\beta}_a \mid \cdot] \sim N\bigg(\big(\boldsymbol{L}^\top\boldsymbol{L}\big)^{-1}\boldsymbol{L}^\top(\boldsymbol{Y}-\boldsymbol{M}\boldsymbol{\beta}_m-\boldsymbol{C}\boldsymbol{\beta}_c), \sigma_e^2\big(\boldsymbol{L}^\top\boldsymbol{L}\big)^{-1}\bigg)$$

$$[\tau_\beta^2 \mid \cdot] \sim IG\bigg(\frac{p+1}{2}, \frac{1}{2}\boldsymbol{\beta}_m^\top\boldsymbol{\Gamma}_\beta^{-1}\boldsymbol{\Lambda}_\beta^{-1}\boldsymbol{\beta}_m+\frac{1}{\nu_\beta}\bigg),$$

$$[\nu_\beta \mid \cdot] \sim IG\bigg(1, \frac{1}{\tau_\beta^2}+\frac{1}{\sigma_e^2}\bigg)$$

$$[\lambda_{\beta gj}^2 \mid \cdot] \sim IG\bigg(b_\beta+\frac{1}{2}, 1+\frac{(\boldsymbol{\beta}_m)_{gj}^2}{2\tau_\beta^2\gamma_{\beta g}^2}\bigg)$$

$$[\gamma_{\beta g}^{-2} \mid \cdot] \sim GIG\bigg(\frac{p_g}{2}-a_\beta, \frac{1}{\tau_\beta^2}\sum_{j=1}^{p_g}\frac{(\boldsymbol{\beta}_m)_{gj}^2}{\lambda_{\beta gj}^2}, 2\bigg)$$

For $d = 1, \ldots, s$,

$$[\boldsymbol{\alpha}_c^d \mid \cdot] \sim N\bigg(\bigg(\sum_{i=1}^n C_{id}^2\bigg)^{-1}\sum_{i=1}^n C_{id}\big(\boldsymbol{M}_{i\cdot}^\top-\boldsymbol{\alpha}_a\boldsymbol{L}_{i\cdot}^\top-\boldsymbol{\alpha}_c^{-d}\boldsymbol{C}_{i,-d}^\top\big), \bigg(\sum_{i=1}^n C_{id}^2\bigg)^{-1}\boldsymbol{\Sigma}_m\bigg),$$

where $\boldsymbol{\alpha}_c^d$ is a $s \times 1$ vector referring to the d-th column of $\boldsymbol{\alpha}_c$, $\boldsymbol{\alpha}_c^{-d}$ is $\boldsymbol{\alpha}_c$ with the d-th column removed, and $\boldsymbol{C}_{i,-d}$ refers to the i-th row of $\boldsymbol{C}$ with the d-th column removed.

$$[\tau_{\alpha k}^2 \mid \cdot] \sim IG\bigg(\frac{p+1}{2}, \frac{1}{2}\sum_{g=1}^G\sum_{j=1}^{p_g}\frac{(\boldsymbol{\alpha}_a^k)_{gj}^2}{\gamma_{\alpha g}^2\lambda_{\alpha gj}^2\{f((\boldsymbol{\beta}_m)_{gj})\}^2}+\frac{1}{\nu_{\alpha k}}\bigg), \quad k = 1, \ldots, r$$

$$[\nu_{\alpha k} \mid \cdot] \sim IG\bigg(1, \frac{1}{\eta^2}+\frac{1}{\tau_{\alpha k}^2}\bigg), \quad k = 1, \ldots, r$$

$$[\eta^2 \mid \cdot] \sim IG\bigg(0.001+\frac{r}{2}, 0.001+\sum_{k=1}^r\frac{1}{\nu_{\alpha k}}\bigg)$$

$$[\gamma_{\alpha g}^{-2} \mid \cdot] \sim GIG\bigg(\frac{p_g r}{2}-a_\alpha, \sum_{k=1}^r\sum_{j=1}^{p_g}\frac{(\boldsymbol{\alpha}_a^k)_{gj}^2}{\tau_{\alpha k}^2\lambda_{\alpha gj}^2\{f((\boldsymbol{\beta}_m)_{gj})\}^2}, 2\bigg)$$

$$[\lambda_{\alpha gj}^2 \mid \cdot] \sim IG\bigg(b_\alpha+\frac{r}{2}, 1+\frac{1}{2}\sum_{k=1}^r\frac{(\boldsymbol{\alpha}_a^k)_{gj}^2}{\tau_{\alpha k}^2\gamma_{\alpha g}^2\{f((\boldsymbol{\beta}_m)_{gj})\}^2}\bigg)$$

$$[(\boldsymbol{\Sigma}_m)_g \mid \cdot] \sim \mathcal{W}^{-1}\bigg(\boldsymbol{I}_{p_g\times p_g}+\sum_{i=1}^n\Big(\boldsymbol{M}_{ig}^\top-(\boldsymbol{\alpha}_a)_{g\cdot}\boldsymbol{L}_{i\cdot}^\top-(\boldsymbol{\alpha}_c)_{g\cdot}\boldsymbol{C}_{i\cdot}^\top\Big)\Big(\boldsymbol{M}_{ig}^\top-(\boldsymbol{\alpha}_a)_{g\cdot}\boldsymbol{L}_{i\cdot}^\top-(\boldsymbol{\alpha}_c)_{g\cdot}\boldsymbol{C}_{i\cdot}^\top\Big)^\top, n+p_g\bigg),$$

where $(\boldsymbol{\alpha}_a)_{g\cdot}$ and $(\boldsymbol{\alpha}_c)_{g\cdot}$ are the submatricies of $\boldsymbol{\alpha}_a$ and $\boldsymbol{\alpha}_c$ corresponding to the mediators in the g-th mediator group.

$$[\boldsymbol{L}_{i\cdot}^\top \mid \cdot] \sim N\Bigg(\Bigg[\frac{1}{\sigma_e^2}\boldsymbol{\beta}_a\boldsymbol{\beta}_a^\top + \boldsymbol{\alpha}_a^\top\boldsymbol{\Sigma}_m^{-1}\boldsymbol{\alpha}_a + \boldsymbol{\Psi}^\top\boldsymbol{\Sigma}_L^{-1}\boldsymbol{\Psi} + \boldsymbol{I}\Bigg]^{-1}\Bigg[\frac{\boldsymbol{\beta}_a}{\sigma_e^2}(Y_i - \boldsymbol{M}_{i\cdot}\boldsymbol{\beta}_m - \boldsymbol{C}_{i\cdot}\boldsymbol{\beta}_c) + \boldsymbol{\alpha}_a^\top\boldsymbol{\Sigma}_m^{-1}(\boldsymbol{M}_{i\cdot}^\top - \boldsymbol{\alpha}_c\boldsymbol{C}_{i\cdot}^\top) + \boldsymbol{\Psi}^\top\boldsymbol{\Sigma}_L^{-1}\boldsymbol{A}_{i\cdot}^\top\Bigg], \Bigg[\frac{1}{\sigma_e^2}\boldsymbol{\beta}_a\boldsymbol{\beta}_a^\top + \boldsymbol{\alpha}_a^\top\boldsymbol{\Sigma}_m^{-1}\boldsymbol{\alpha}_a + \boldsymbol{\Psi}^\top\boldsymbol{\Sigma}_L^{-1}\boldsymbol{\Psi} + \boldsymbol{I}\Bigg]^{-1}\Bigg)$$

$$[\sigma_{Lk'}^2 \mid \cdot] \sim IG\Bigg(\frac{n}{2}, \frac{1}{2}\sum_{i=1}^{n}\big(A_{ik'} - \boldsymbol{\Psi}_{k'\cdot}\boldsymbol{L}_{i\cdot}^\top\big)^2\Bigg), \quad k' = 1, \ldots, q,$$

where $\boldsymbol{\Psi}_{k'\cdot}$ is the k'-th row of $\boldsymbol{\Psi}$.

$$[\boldsymbol{\Psi}_{k'\cdot}^\top \mid \cdot] \sim N\Bigg(\Bigg\{\boldsymbol{D}_{k'}^{-1} + \sigma_{Lk'}^{-2}\boldsymbol{L}^\top\boldsymbol{L}\Bigg\}^{-1}\boldsymbol{L}^\top\sigma_{Lk'}^{-2}\boldsymbol{A}_{\cdot k'}, \Bigg\{\boldsymbol{D}_{k'}^{-1} + \sigma_{Lk'}^{-2}\boldsymbol{L}^\top\boldsymbol{L}\Bigg\}^{-1}\Bigg),$$

where $\boldsymbol{D}_{k'}^{-1} = \text{diag}(\phi_{k'1}\tau_1, \ldots, \phi_{k'r}\tau_r)$, $k' = 1, \ldots, q$, and $\boldsymbol{A}_{\cdot k'}$ is the k'-th column of $\boldsymbol{A}$ which is an $n \times 1$ vector.

$$[\phi_{k'k} \mid \cdot] \sim G\Bigg(\frac{3}{2}, \frac{3 + \tau_k\boldsymbol{\Psi}_{k'k}^2}{2}\Bigg)$$

$$[\delta_1 \mid \cdot] \sim G\Bigg(2.1 + \frac{qr}{2}, 1 + \frac{1}{2}\Big(\sum_{k'=1}^{q}\sum_{k=2}^{r}\phi_{k'k}\boldsymbol{\Psi}_{k'k}^2\Big(\prod_{m=2}^{k}\delta_m\Big) + \sum_{k'=1}^{q}\phi_{k'1}\boldsymbol{\Psi}_{k'1}^2\Big)\Bigg)$$

$$[\delta_m \mid \cdot] \sim G\Bigg(3.1 + \frac{q}{2}(r - m + 1), 1 + \frac{1}{2}\sum_{k'=1}^{q}\sum_{k=m}^{r}\phi_{k'k}\boldsymbol{\Psi}_{k'k}^2\Big(\prod_{l=1, l\neq m}^{k}\delta_l\Big)\Bigg),$$

where $m = 2, \ldots, r$.

C.3 Supplementary Simulation Results

			One-Step				Two-Step			
Shrinkage	**Parameter**	$(\boldsymbol{l}-\boldsymbol{l}^*)^\top$	**Bias**	**RMSE**	**Posterior SD**	**Coverage**	**Bias**	**RMSE**	**Posterior SD**	**Coverage**
Independent	$\text{DE}^L(\boldsymbol{l},\boldsymbol{l}^*)$	(1,0,0,0)	0.01	0.10	0.10	0.96	0.00	0.09	0.10	0.96
Independent	$\text{DE}^L(\boldsymbol{l},\boldsymbol{l}^*)$	(0,1,0,0)	0.00	0.10	0.10	0.95	-0.01	0.10	0.10	0.94
Independent	$\text{DE}^L(\boldsymbol{l},\boldsymbol{l}^*)$	(0,0,1,0)	0.01	0.09	0.09	0.95	0.00	0.09	0.09	0.94
Independent	$\text{DE}^L(\boldsymbol{l},\boldsymbol{l}^*)$	(0,0,0,1)	0.02	0.10	0.09	0.95	0.02	0.10	0.09	0.94
Product	$\text{DE}^L(\boldsymbol{l},\boldsymbol{l}^*)$	(1,0,0,0)	0.01	0.10	0.10	0.96	0.00	0.09	0.10	0.96
Product	$\text{DE}^L(\boldsymbol{l},\boldsymbol{l}^*)$	(0,1,0,0)	0.00	0.10	0.10	0.95	-0.01	0.10	0.10	0.94
Product	$\text{DE}^L(\boldsymbol{l},\boldsymbol{l}^*)$	(0,0,1,0)	0.01	0.09	0.09	0.95	0.00	0.09	0.09	0.94
Product	$\text{DE}^L(\boldsymbol{l},\boldsymbol{l}^*)$	(0,0,0,1)	0.02	0.10	0.09	0.94	0.02	0.10	0.09	0.94
Independent	$\text{IE}^L(\boldsymbol{l},\boldsymbol{l}^*)$	(1,0,0,0)	0.00	0.02	0.03	0.99	0.02	0.03	0.03	0.98
Independent	$\text{IE}^L(\boldsymbol{l},\boldsymbol{l}^*)$	(0,1,0,0)	0.00	0.02	0.03	0.99	0.02	0.03	0.03	0.97
Independent	$\text{IE}^L(\boldsymbol{l},\boldsymbol{l}^*)$	(0,0,1,0)	0.00	0.00	0.00	1.00	0.00	0.00	0.00	1.00
Independent	$\text{IE}^L(\boldsymbol{l},\boldsymbol{l}^*)$	(0,0,0,1)	0.00	0.00	0.00	1.00	0.00	0.00	0.00	1.00
Product	$\text{IE}^L(\boldsymbol{l},\boldsymbol{l}^*)$	(1,0,0,0)	0.00	0.02	0.03	0.99	0.01	0.02	0.02	0.97
Product	$\text{IE}^L(\boldsymbol{l},\boldsymbol{l}^*)$	(0,1,0,0)	0.00	0.02	0.03	0.99	0.01	0.02	0.02	0.96
Product	$\text{IE}^L(\boldsymbol{l},\boldsymbol{l}^*)$	(0,0,1,0)	0.00	0.00	0.00	1.00	0.00	0.00	0.00	1.00
Product	$\text{IE}^L(\boldsymbol{l},\boldsymbol{l}^*)$	(0,0,0,1)	0.00	0.00	0.00	1.00	0.00	0.00	0.00	1.00
Independent	$\text{TE}^L(\boldsymbol{l},\boldsymbol{l}^*)$	(1,0,0,0)	0.01	0.09	0.10	0.96	0.01	0.09	0.09	0.95
Independent	$\text{TE}^L(\boldsymbol{l},\boldsymbol{l}^*)$	(0,1,0,0)	0.00	0.10	0.10	0.94	0.00	0.10	0.09	0.95
Independent	$\text{TE}^L(\boldsymbol{l},\boldsymbol{l}^*)$	(0,0,1,0)	0.01	0.09	0.09	0.95	0.00	0.09	0.09	0.94
Independent	$\text{TE}^L(\boldsymbol{l},\boldsymbol{l}^*)$	(0,0,0,1)	0.02	0.10	0.09	0.95	0.02	0.10	0.09	0.94
Product	$\text{TE}^L(\boldsymbol{l},\boldsymbol{l}^*)$	(1,0,0,0)	0.01	0.09	0.10	0.96	0.01	0.09	0.09	0.95
Product	$\text{TE}^L(\boldsymbol{l},\boldsymbol{l}^*)$	(0,1,0,0)	0.00	0.10	0.10	0.94	0.00	0.10	0.09	0.94
Product	$\text{TE}^L(\boldsymbol{l},\boldsymbol{l}^*)$	(0,0,1,0)	0.01	0.09	0.09	0.95	0.00	0.09	0.09	0.94
Product	$\text{TE}^L(\boldsymbol{l},\boldsymbol{l}^*)$	(0,0,0,1)	0.02	0.10	0.09	0.94	0.02	0.10	0.09	0.94

Table C.1: Simulation results for the $n = 2000$ IE zero simulation setting for the one-step and two-step estimators of $\text{DE}^L(\boldsymbol{l},\boldsymbol{l}^*)$, $\text{IE}^L(\boldsymbol{l},\boldsymbol{l}^*)$, and $\text{TE}^L(\boldsymbol{l},\boldsymbol{l}^*)$ when $R^2_L = 0.5$ and Adjusted $R^2_O = 0.3$.

			One-Step				Two-Step			
Shrinkage	**Parameter**	$(\boldsymbol{l}-\boldsymbol{l}^*)^\top$	**Bias**	**RMSE**	**Posterior SD**	**Coverage**	**Bias**	**RMSE**	**Posterior SD**	**Coverage**
Independent	$\text{DE}^L(\boldsymbol{l},\boldsymbol{l}^*)$	(1,0,0,0)	0.04	0.09	0.08	0.93	0.01	0.08	0.08	0.95
Independent	$\text{DE}^L(\boldsymbol{l},\boldsymbol{l}^*)$	(0,1,0,0)	0.03	0.09	0.08	0.93	0.00	0.08	0.08	0.94
Independent	$\text{DE}^L(\boldsymbol{l},\boldsymbol{l}^*)$	(0,0,1,0)	0.03	0.09	0.08	0.93	0.00	0.08	0.08	0.94
Independent	$\text{DE}^L(\boldsymbol{l},\boldsymbol{l}^*)$	(0,0,0,1)	0.04	0.09	0.08	0.92	0.01	0.08	0.08	0.93
Product	$\text{DE}^L(\boldsymbol{l},\boldsymbol{l}^*)$	(1,0,0,0)	0.04	0.09	0.08	0.93	0.01	0.08	0.08	0.95
Product	$\text{DE}^L(\boldsymbol{l},\boldsymbol{l}^*)$	(0,1,0,0)	0.03	0.09	0.08	0.93	0.00	0.08	0.08	0.93
Product	$\text{DE}^L(\boldsymbol{l},\boldsymbol{l}^*)$	(0,0,1,0)	0.03	0.09	0.08	0.93	0.00	0.08	0.08	0.94
Product	$\text{DE}^L(\boldsymbol{l},\boldsymbol{l}^*)$	(0,0,0,1)	0.04	0.09	0.08	0.92	0.01	0.08	0.08	0.93
Independent	$\text{IE}^L(\boldsymbol{l},\boldsymbol{l}^*)$	(1,0,0,0)	-0.03	0.04	0.03	0.76	-0.02	0.03	0.03	0.89
Independent	$\text{IE}^L(\boldsymbol{l},\boldsymbol{l}^*)$	(0,1,0,0)	-0.03	0.04	0.03	0.77	-0.02	0.03	0.03	0.88
Independent	$\text{IE}^L(\boldsymbol{l},\boldsymbol{l}^*)$	(0,0,1,0)	-0.03	0.04	0.03	0.79	-0.01	0.03	0.03	0.91
Independent	$\text{IE}^L(\boldsymbol{l},\boldsymbol{l}^*)$	(0,0,0,1)	-0.03	0.04	0.03	0.78	-0.01	0.03	0.03	0.90
Product	$\text{IE}^L(\boldsymbol{l},\boldsymbol{l}^*)$	(1,0,0,0)	-0.03	0.04	0.03	0.74	-0.02	0.03	0.03	0.85
Product	$\text{IE}^L(\boldsymbol{l},\boldsymbol{l}^*)$	(0,1,0,0)	-0.03	0.04	0.03	0.76	-0.02	0.04	0.03	0.86
Product	$\text{IE}^L(\boldsymbol{l},\boldsymbol{l}^*)$	(0,0,1,0)	-0.03	0.04	0.03	0.78	-0.02	0.03	0.03	0.88
Product	$\text{IE}^L(\boldsymbol{l},\boldsymbol{l}^*)$	(0,0,0,1)	-0.03	0.04	0.03	0.76	-0.02	0.03	0.03	0.87
Independent	$\text{TE}^L(\boldsymbol{l},\boldsymbol{l}^*)$	(1,0,0,0)	0.01	0.08	0.08	0.95	-0.01	0.08	0.08	0.95
Independent	$\text{TE}^L(\boldsymbol{l},\boldsymbol{l}^*)$	(0,1,0,0)	0.00	0.08	0.08	0.94	-0.01	0.08	0.08	0.93
Independent	$\text{TE}^L(\boldsymbol{l},\boldsymbol{l}^*)$	(0,0,1,0)	0.00	0.08	0.08	0.94	-0.01	0.08	0.08	0.93
Independent	$\text{TE}^L(\boldsymbol{l},\boldsymbol{l}^*)$	(0,0,0,1)	0.01	0.08	0.08	0.94	0.00	0.08	0.08	0.94
Product	$\text{TE}^L(\boldsymbol{l},\boldsymbol{l}^*)$	(1,0,0,0)	0.01	0.08	0.08	0.96	-0.01	0.08	0.08	0.94
Product	$\text{TE}^L(\boldsymbol{l},\boldsymbol{l}^*)$	(0,1,0,0)	0.00	0.08	0.08	0.94	-0.02	0.08	0.08	0.93
Product	$\text{TE}^L(\boldsymbol{l},\boldsymbol{l}^*)$	(0,0,1,0)	0.00	0.08	0.08	0.94	-0.01	0.08	0.08	0.94
Product	$\text{TE}^L(\boldsymbol{l},\boldsymbol{l}^*)$	(0,0,0,1)	0.01	0.08	0.08	0.94	-0.01	0.08	0.08	0.93

Table C.2: Simulation results for the $n = 2000$ IE dense simulation setting for the one-step and two-step estimators of $\text{DE}^L(\boldsymbol{l},\boldsymbol{l}^*)$, $\text{IE}^L(\boldsymbol{l},\boldsymbol{l}^*)$, and $\text{TE}^L(\boldsymbol{l},\boldsymbol{l}^*)$ when $R^2_L = 0.5$ and Adjusted $R^2_O = 0.3$.

			One-Step				Two-Step			
Shrinkage	**Parameter**	$(\boldsymbol{l}-\boldsymbol{l}^*)^\top$	**Bias**	**RMSE**	**Posterior SD**	**Coverage**	**Bias**	**RMSE**	**Posterior SD**	**Coverage**
Independent	$\text{DE}^L(\boldsymbol{l},\boldsymbol{l}^*)$	(1,0,0,0)	0.02	0.08	0.08	0.96	0.00	0.08	0.08	0.96
Independent	$\text{DE}^L(\boldsymbol{l},\boldsymbol{l}^*)$	(0,1,0,0)	0.01	0.08	0.08	0.95	0.00	0.08	0.08	0.94
Independent	$\text{DE}^L(\boldsymbol{l},\boldsymbol{l}^*)$	(0,0,1,0)	0.01	0.08	0.08	0.95	0.00	0.08	0.08	0.95
Independent	$\text{DE}^L(\boldsymbol{l},\boldsymbol{l}^*)$	(0,0,0,1)	0.02	0.09	0.08	0.94	0.00	0.08	0.08	0.94
Product	$\text{DE}^L(\boldsymbol{l},\boldsymbol{l}^*)$	(1,0,0,0)	0.02	0.08	0.08	0.96	0.00	0.08	0.08	0.96
Product	$\text{DE}^L(\boldsymbol{l},\boldsymbol{l}^*)$	(0,1,0,0)	0.01	0.08	0.08	0.95	0.00	0.08	0.08	0.94
Product	$\text{DE}^L(\boldsymbol{l},\boldsymbol{l}^*)$	(0,0,1,0)	0.01	0.08	0.08	0.95	0.00	0.08	0.08	0.94
Product	$\text{DE}^L(\boldsymbol{l},\boldsymbol{l}^*)$	(0,0,0,1)	0.02	0.09	0.08	0.94	0.00	0.08	0.08	0.94
Independent	$\text{IE}^L(\boldsymbol{l},\boldsymbol{l}^*)$	(1,0,0,0)	-0.01	0.03	0.03	0.94	0.00	0.03	0.03	0.95
Independent	$\text{IE}^L(\boldsymbol{l},\boldsymbol{l}^*)$	(0,1,0,0)	-0.01	0.03	0.03	0.93	0.00	0.03	0.03	0.94
Independent	$\text{IE}^L(\boldsymbol{l},\boldsymbol{l}^*)$	(0,0,1,0)	-0.01	0.03	0.03	0.94	0.01	0.03	0.03	0.97
Independent	$\text{IE}^L(\boldsymbol{l},\boldsymbol{l}^*)$	(0,0,0,1)	-0.01	0.03	0.03	0.93	0.00	0.03	0.03	0.95
Product	$\text{IE}^L(\boldsymbol{l},\boldsymbol{l}^*)$	(1,0,0,0)	-0.01	0.03	0.03	0.93	0.00	0.03	0.03	0.94
Product	$\text{IE}^L(\boldsymbol{l},\boldsymbol{l}^*)$	(0,1,0,0)	-0.01	0.03	0.03	0.93	0.00	0.03	0.03	0.94
Product	$\text{IE}^L(\boldsymbol{l},\boldsymbol{l}^*)$	(0,0,1,0)	-0.01	0.03	0.03	0.94	0.01	0.03	0.03	0.96
Product	$\text{IE}^L(\boldsymbol{l},\boldsymbol{l}^*)$	(0,0,0,1)	-0.01	0.03	0.03	0.93	0.00	0.03	0.03	0.94
Independent	$\text{TE}^L(\boldsymbol{l},\boldsymbol{l}^*)$	(1,0,0,0)	0.01	0.08	0.09	0.96	0.00	0.08	0.08	0.95
Independent	$\text{TE}^L(\boldsymbol{l},\boldsymbol{l}^*)$	(0,1,0,0)	0.00	0.09	0.09	0.94	0.00	0.09	0.08	0.93
Independent	$\text{TE}^L(\boldsymbol{l},\boldsymbol{l}^*)$	(0,0,1,0)	0.00	0.09	0.09	0.95	0.00	0.09	0.08	0.94
Independent	$\text{TE}^L(\boldsymbol{l},\boldsymbol{l}^*)$	(0,0,0,1)	0.01	0.09	0.09	0.94	0.01	0.09	0.08	0.93
Product	$\text{TE}^L(\boldsymbol{l},\boldsymbol{l}^*)$	(1,0,0,0)	0.01	0.08	0.09	0.96	0.00	0.08	0.08	0.94
Product	$\text{TE}^L(\boldsymbol{l},\boldsymbol{l}^*)$	(0,1,0,0)	0.00	0.09	0.09	0.94	0.00	0.09	0.08	0.93
Product	$\text{TE}^L(\boldsymbol{l},\boldsymbol{l}^*)$	(0,0,1,0)	0.00	0.09	0.09	0.95	0.00	0.09	0.08	0.94
Product	$\text{TE}^L(\boldsymbol{l},\boldsymbol{l}^*)$	(0,0,0,1)	0.01	0.09	0.09	0.94	0.01	0.09	0.08	0.93

Table C.3: Simulation results for the $n = 2000$ IE sparse simulation setting for the one-step and two-step estimators of $\text{DE}^L(\boldsymbol{l},\boldsymbol{l}^*)$, $\text{IE}^L(\boldsymbol{l},\boldsymbol{l}^*)$, and $\text{TE}^L(\boldsymbol{l},\boldsymbol{l}^*)$ when $R^2_L = 0.5$ and Adjusted $R^2_O = 0.3$.

			One-Step				Two-Step			
Shrinkage	**Parameter**	$(\boldsymbol{l}-\boldsymbol{l}^*)^\top$	**Bias**	**RMSE**	**Posterior SD**	**Coverage**	**Bias**	**RMSE**	**Posterior SD**	**Coverage**
Independent	$\text{DE}^L(\boldsymbol{l},\boldsymbol{l}^*)$	(1,0,0,0)	0.04	0.14	0.13	0.93	0.03	0.13	0.12	0.91
Independent	$\text{DE}^L(\boldsymbol{l},\boldsymbol{l}^*)$	(0,1,0,0)	0.06	0.14	0.13	0.93	0.04	0.13	0.12	0.92
Independent	$\text{DE}^L(\boldsymbol{l},\boldsymbol{l}^*)$	(0,0,1,0)	0.07	0.14	0.13	0.91	0.06	0.14	0.11	0.90
Independent	$\text{DE}^L(\boldsymbol{l},\boldsymbol{l}^*)$	(0,0,0,1)	0.06	0.14	0.13	0.93	0.05	0.13	0.11	0.90
Product	$\text{DE}^L(\boldsymbol{l},\boldsymbol{l}^*)$	(1,0,0,0)	0.04	0.14	0.13	0.93	0.03	0.13	0.12	0.91
Product	$\text{DE}^L(\boldsymbol{l},\boldsymbol{l}^*)$	(0,1,0,0)	0.05	0.14	0.13	0.94	0.04	0.13	0.12	0.92
Product	$\text{DE}^L(\boldsymbol{l},\boldsymbol{l}^*)$	(0,0,1,0)	0.07	0.14	0.13	0.92	0.06	0.14	0.11	0.89
Product	$\text{DE}^L(\boldsymbol{l},\boldsymbol{l}^*)$	(0,0,0,1)	0.06	0.14	0.13	0.93	0.05	0.13	0.11	0.90
Independent	$\text{IE}^L(\boldsymbol{l},\boldsymbol{l}^*)$	(1,0,0,0)	0.00	0.02	0.03	1.00	0.01	0.02	0.03	1.00
Independent	$\text{IE}^L(\boldsymbol{l},\boldsymbol{l}^*)$	(0,1,0,0)	0.00	0.02	0.03	1.00	0.01	0.02	0.03	1.00
Independent	$\text{IE}^L(\boldsymbol{l},\boldsymbol{l}^*)$	(0,0,1,0)	0.00	0.00	0.01	1.00	0.00	0.00	0.01	1.00
Independent	$\text{IE}^L(\boldsymbol{l},\boldsymbol{l}^*)$	(0,0,0,1)	0.00	0.00	0.01	1.00	0.00	0.00	0.01	1.00
Product	$\text{IE}^L(\boldsymbol{l},\boldsymbol{l}^*)$	(1,0,0,0)	0.00	0.01	0.01	1.00	0.00	0.00	0.00	1.00
Product	$\text{IE}^L(\boldsymbol{l},\boldsymbol{l}^*)$	(0,1,0,0)	0.00	0.01	0.01	1.00	0.00	0.00	0.01	1.00
Product	$\text{IE}^L(\boldsymbol{l},\boldsymbol{l}^*)$	(0,0,1,0)	0.00	0.00	0.00	1.00	0.00	0.00	0.00	1.00
Product	$\text{IE}^L(\boldsymbol{l},\boldsymbol{l}^*)$	(0,0,0,1)	0.00	0.00	0.00	1.00	0.00	0.00	0.00	1.00
Independent	$\text{TE}^L(\boldsymbol{l},\boldsymbol{l}^*)$	(1,0,0,0)	0.04	0.14	0.13	0.93	0.04	0.14	0.12	0.90
Independent	$\text{TE}^L(\boldsymbol{l},\boldsymbol{l}^*)$	(0,1,0,0)	0.06	0.14	0.13	0.93	0.05	0.14	0.12	0.91
Independent	$\text{TE}^L(\boldsymbol{l},\boldsymbol{l}^*)$	(0,0,1,0)	0.07	0.14	0.13	0.92	0.06	0.14	0.11	0.90
Independent	$\text{TE}^L(\boldsymbol{l},\boldsymbol{l}^*)$	(0,0,0,1)	0.06	0.14	0.13	0.93	0.05	0.13	0.11	0.90
Product	$\text{TE}^L(\boldsymbol{l},\boldsymbol{l}^*)$	(1,0,0,0)	0.04	0.14	0.13	0.93	0.03	0.13	0.12	0.91
Product	$\text{TE}^L(\boldsymbol{l},\boldsymbol{l}^*)$	(0,1,0,0)	0.05	0.14	0.13	0.94	0.04	0.13	0.12	0.91
Product	$\text{TE}^L(\boldsymbol{l},\boldsymbol{l}^*)$	(0,0,1,0)	0.07	0.14	0.13	0.92	0.06	0.14	0.11	0.90
Product	$\text{TE}^L(\boldsymbol{l},\boldsymbol{l}^*)$	(0,0,0,1)	0.06	0.14	0.13	0.93	0.05	0.13	0.11	0.90

Table C.4: Simulation results for the $n = 400$ IE zero simulation setting for the one-step and two-step estimators of $\text{DE}^L(\boldsymbol{l},\boldsymbol{l}^*)$, $\text{IE}^L(\boldsymbol{l},\boldsymbol{l}^*)$, and $\text{TE}^L(\boldsymbol{l},\boldsymbol{l}^*)$ when $R^2_L = 0.5$ and Adjusted $R^2_O = 0.6$.

			One-Step				Two-Step			
Shrinkage	**Parameter**	$(\boldsymbol{l}-\boldsymbol{l}^*)^\top$	**Bias**	**RMSE**	**Posterior SD**	**Coverage**	**Bias**	**RMSE**	**Posterior SD**	**Coverage**
Independent	$\text{DE}^L(\boldsymbol{l},\boldsymbol{l}^*)$	(1,0,0,0)	0.05	0.12	0.11	0.92	0.03	0.12	0.10	0.90
Independent	$\text{DE}^L(\boldsymbol{l},\boldsymbol{l}^*)$	(0,1,0,0)	0.06	0.12	0.11	0.92	0.04	0.12	0.10	0.91
Independent	$\text{DE}^L(\boldsymbol{l},\boldsymbol{l}^*)$	(0,0,1,0)	0.06	0.12	0.11	0.92	0.04	0.11	0.10	0.91
Independent	$\text{DE}^L(\boldsymbol{l},\boldsymbol{l}^*)$	(0,0,0,1)	0.05	0.12	0.11	0.92	0.04	0.11	0.10	0.91
Product	$\text{DE}^L(\boldsymbol{l},\boldsymbol{l}^*)$	(1,0,0,0)	0.05	0.12	0.11	0.92	0.03	0.12	0.10	0.90
Product	$\text{DE}^L(\boldsymbol{l},\boldsymbol{l}^*)$	(0,1,0,0)	0.06	0.12	0.11	0.92	0.04	0.12	0.10	0.91
Product	$\text{DE}^L(\boldsymbol{l},\boldsymbol{l}^*)$	(0,0,1,0)	0.06	0.12	0.11	0.92	0.04	0.11	0.10	0.91
Product	$\text{DE}^L(\boldsymbol{l},\boldsymbol{l}^*)$	(0,0,0,1)	0.05	0.12	0.11	0.92	0.04	0.11	0.10	0.91
Independent	$\text{IE}^L(\boldsymbol{l},\boldsymbol{l}^*)$	(1,0,0,0)	-0.03	0.06	0.05	0.91	-0.03	0.07	0.06	0.89
Independent	$\text{IE}^L(\boldsymbol{l},\boldsymbol{l}^*)$	(0,1,0,0)	-0.02	0.06	0.05	0.93	-0.03	0.07	0.06	0.92
Independent	$\text{IE}^L(\boldsymbol{l},\boldsymbol{l}^*)$	(0,0,1,0)	-0.02	0.06	0.05	0.91	-0.03	0.07	0.06	0.90
Independent	$\text{IE}^L(\boldsymbol{l},\boldsymbol{l}^*)$	(0,0,0,1)	-0.03	0.06	0.05	0.91	-0.03	0.07	0.06	0.89
Product	$\text{IE}^L(\boldsymbol{l},\boldsymbol{l}^*)$	(1,0,0,0)	-0.04	0.08	0.05	0.81	-0.06	0.11	0.05	0.69
Product	$\text{IE}^L(\boldsymbol{l},\boldsymbol{l}^*)$	(0,1,0,0)	-0.04	0.08	0.05	0.82	-0.06	0.11	0.05	0.70
Product	$\text{IE}^L(\boldsymbol{l},\boldsymbol{l}^*)$	(0,0,1,0)	-0.04	0.08	0.05	0.82	-0.06	0.11	0.05	0.69
Product	$\text{IE}^L(\boldsymbol{l},\boldsymbol{l}^*)$	(0,0,0,1)	-0.04	0.08	0.05	0.82	-0.06	0.11	0.05	0.68
Independent	$\text{TE}^L(\boldsymbol{l},\boldsymbol{l}^*)$	(1,0,0,0)	0.02	0.13	0.12	0.95	0.00	0.13	0.11	0.91
Independent	$\text{TE}^L(\boldsymbol{l},\boldsymbol{l}^*)$	(0,1,0,0)	0.03	0.13	0.12	0.94	0.01	0.13	0.11	0.91
Independent	$\text{TE}^L(\boldsymbol{l},\boldsymbol{l}^*)$	(0,0,1,0)	0.04	0.12	0.12	0.94	0.02	0.12	0.11	0.91
Independent	$\text{TE}^L(\boldsymbol{l},\boldsymbol{l}^*)$	(0,0,0,1)	0.03	0.12	0.12	0.95	0.01	0.13	0.11	0.91
Product	$\text{TE}^L(\boldsymbol{l},\boldsymbol{l}^*)$	(1,0,0,0)	0.01	0.14	0.12	0.93	-0.03	0.15	0.11	0.84
Product	$\text{TE}^L(\boldsymbol{l},\boldsymbol{l}^*)$	(0,1,0,0)	0.02	0.14	0.12	0.91	-0.02	0.15	0.11	0.84
Product	$\text{TE}^L(\boldsymbol{l},\boldsymbol{l}^*)$	(0,0,1,0)	0.02	0.13	0.12	0.92	-0.02	0.14	0.11	0.86
Product	$\text{TE}^L(\boldsymbol{l},\boldsymbol{l}^*)$	(0,0,0,1)	0.01	0.13	0.12	0.92	-0.02	0.14	0.11	0.84

Table C.5: Simulation results for the $n = 400$ IE sparse simulation setting for the one-step and two-step estimators of $\text{DE}^L(\boldsymbol{l},\boldsymbol{l}^*)$, $\text{IE}^L(\boldsymbol{l},\boldsymbol{l}^*)$, and $\text{TE}^L(\boldsymbol{l},\boldsymbol{l}^*)$ when $R^2_L = 0.5$ and Adjusted $R^2_O = 0.6$.

			One-Step				Two-Step			
Shrinkage	**Parameter**	$(\boldsymbol{l}-\boldsymbol{l}^*)^\top$	**Bias**	**RMSE**	**Posterior SD**	**Coverage**	**Bias**	**RMSE**	**Posterior SD**	**Coverage**
Independent	$\text{DE}^L(\boldsymbol{l},\boldsymbol{l}^*)$	(1,0,0,0)	0.07	0.14	0.11	0.89	0.05	0.12	0.10	0.89
Independent	$\text{DE}^L(\boldsymbol{l},\boldsymbol{l}^*)$	(0,1,0,0)	0.08	0.14	0.11	0.87	0.05	0.12	0.10	0.88
Independent	$\text{DE}^L(\boldsymbol{l},\boldsymbol{l}^*)$	(0,0,1,0)	0.09	0.14	0.11	0.87	0.06	0.12	0.10	0.88
Independent	$\text{DE}^L(\boldsymbol{l},\boldsymbol{l}^*)$	(0,0,0,1)	0.08	0.13	0.11	0.88	0.05	0.12	0.10	0.89
Product	$\text{DE}^L(\boldsymbol{l},\boldsymbol{l}^*)$	(1,0,0,0)	0.06	0.13	0.11	0.90	0.05	0.12	0.10	0.89
Product	$\text{DE}^L(\boldsymbol{l},\boldsymbol{l}^*)$	(0,1,0,0)	0.07	0.13	0.11	0.88	0.05	0.12	0.10	0.88
Product	$\text{DE}^L(\boldsymbol{l},\boldsymbol{l}^*)$	(0,0,1,0)	0.08	0.13	0.11	0.88	0.06	0.12	0.10	0.88
Product	$\text{DE}^L(\boldsymbol{l},\boldsymbol{l}^*)$	(0,0,0,1)	0.07	0.13	0.11	0.90	0.05	0.12	0.10	0.89
Independent	$\text{IE}^L(\boldsymbol{l},\boldsymbol{l}^*)$	(1,0,0,0)	-0.06	0.08	0.04	0.66	-0.15	0.15	0.04	0.22
Independent	$\text{IE}^L(\boldsymbol{l},\boldsymbol{l}^*)$	(0,1,0,0)	-0.06	0.08	0.04	0.68	-0.15	0.15	0.04	0.24
Independent	$\text{IE}^L(\boldsymbol{l},\boldsymbol{l}^*)$	(0,0,1,0)	-0.06	0.08	0.04	0.69	-0.14	0.15	0.04	0.23
Independent	$\text{IE}^L(\boldsymbol{l},\boldsymbol{l}^*)$	(0,0,0,1)	-0.06	0.08	0.04	0.66	-0.15	0.15	0.04	0.22
Product	$\text{IE}^L(\boldsymbol{l},\boldsymbol{l}^*)$	(1,0,0,0)	-0.13	0.15	0.02	0.34	-0.20	0.20	0.00	0.00
Product	$\text{IE}^L(\boldsymbol{l},\boldsymbol{l}^*)$	(0,1,0,0)	-0.13	0.15	0.02	0.33	-0.20	0.20	0.00	0.00
Product	$\text{IE}^L(\boldsymbol{l},\boldsymbol{l}^*)$	(0,0,1,0)	-0.13	0.15	0.02	0.34	-0.20	0.20	0.00	0.00
Product	$\text{IE}^L(\boldsymbol{l},\boldsymbol{l}^*)$	(0,0,0,1)	-0.13	0.15	0.02	0.33	-0.20	0.20	0.00	0.00
Independent	$\text{TE}^L(\boldsymbol{l},\boldsymbol{l}^*)$	(1,0,0,0)	0.01	0.12	0.11	0.93	-0.10	0.16	0.10	0.78
Independent	$\text{TE}^L(\boldsymbol{l},\boldsymbol{l}^*)$	(0,1,0,0)	0.02	0.12	0.11	0.94	-0.09	0.15	0.10	0.81
Independent	$\text{TE}^L(\boldsymbol{l},\boldsymbol{l}^*)$	(0,0,1,0)	0.02	0.12	0.12	0.94	-0.09	0.15	0.10	0.82
Independent	$\text{TE}^L(\boldsymbol{l},\boldsymbol{l}^*)$	(0,0,0,1)	0.01	0.12	0.11	0.95	-0.09	0.15	0.10	0.80
Product	$\text{TE}^L(\boldsymbol{l},\boldsymbol{l}^*)$	(1,0,0,0)	-0.07	0.16	0.11	0.81	-0.15	0.19	0.10	0.62
Product	$\text{TE}^L(\boldsymbol{l},\boldsymbol{l}^*)$	(0,1,0,0)	-0.06	0.15	0.11	0.83	-0.15	0.18	0.10	0.65
Product	$\text{TE}^L(\boldsymbol{l},\boldsymbol{l}^*)$	(0,0,1,0)	-0.06	0.15	0.11	0.84	-0.14	0.18	0.10	0.69
Product	$\text{TE}^L(\boldsymbol{l},\boldsymbol{l}^*)$	(0,0,0,1)	-0.07	0.15	0.11	0.83	-0.15	0.18	0.10	0.65

Table C.6: Simulation results for the $n = 400$ IE dense simulation setting for the one-step and two-step estimators of $\text{DE}^L(\boldsymbol{l},\boldsymbol{l}^*)$, $\text{IE}^L(\boldsymbol{l},\boldsymbol{l}^*)$, and $\text{TE}^L(\boldsymbol{l},\boldsymbol{l}^*)$ when $R^2_L = 0.5$ and Adjusted $R^2_O = 0.6$.

			One-Step				Two-Step			
Shrinkage	**Parameter**	$(\boldsymbol{l}-\boldsymbol{l}^*)^\top$	**Bias**	**RMSE**	**Posterior SD**	**Coverage**	**Bias**	**RMSE**	**Posterior SD**	**Coverage**
Independent	$\text{DE}^L(\boldsymbol{l},\boldsymbol{l}^*)$	(1,0,0,0)	-0.04	0.19	0.20	0.96	-0.09	0.21	0.18	0.92
Independent	$\text{DE}^L(\boldsymbol{l},\boldsymbol{l}^*)$	(0,1,0,0)	-0.03	0.18	0.20	0.96	-0.07	0.20	0.18	0.92
Independent	$\text{DE}^L(\boldsymbol{l},\boldsymbol{l}^*)$	(0,0,1,0)	0.04	0.19	0.20	0.96	0.03	0.19	0.18	0.94
Independent	$\text{DE}^L(\boldsymbol{l},\boldsymbol{l}^*)$	(0,0,0,1)	0.03	0.18	0.20	0.96	0.01	0.18	0.18	0.94
Product	$\text{DE}^L(\boldsymbol{l},\boldsymbol{l}^*)$	(1,0,0,0)	-0.08	0.20	0.20	0.95	-0.09	0.21	0.18	0.93
Product	$\text{DE}^L(\boldsymbol{l},\boldsymbol{l}^*)$	(0,1,0,0)	-0.06	0.19	0.20	0.96	-0.07	0.20	0.18	0.93
Product	$\text{DE}^L(\boldsymbol{l},\boldsymbol{l}^*)$	(0,0,1,0)	0.03	0.19	0.20	0.96	0.03	0.19	0.18	0.94
Product	$\text{DE}^L(\boldsymbol{l},\boldsymbol{l}^*)$	(0,0,0,1)	0.02	0.18	0.20	0.97	0.01	0.18	0.18	0.93
Independent	$\text{IE}^L(\boldsymbol{l},\boldsymbol{l}^*)$	(1,0,0,0)	0.00	0.01	0.03	1.00	0.01	0.02	0.03	1.00
Independent	$\text{IE}^L(\boldsymbol{l},\boldsymbol{l}^*)$	(0,1,0,0)	0.00	0.01	0.03	1.00	0.01	0.02	0.03	1.00
Independent	$\text{IE}^L(\boldsymbol{l},\boldsymbol{l}^*)$	(0,0,1,0)	0.00	0.01	0.02	1.00	0.00	0.00	0.02	1.00
Independent	$\text{IE}^L(\boldsymbol{l},\boldsymbol{l}^*)$	(0,0,0,1)	0.00	0.01	0.02	1.00	0.00	0.00	0.02	1.00
Product	$\text{IE}^L(\boldsymbol{l},\boldsymbol{l}^*)$	(1,0,0,0)	0.00	0.00	0.00	1.00	0.00	0.00	0.00	1.00
Product	$\text{IE}^L(\boldsymbol{l},\boldsymbol{l}^*)$	(0,1,0,0)	0.00	0.00	0.00	1.00	0.00	0.00	0.00	1.00
Product	$\text{IE}^L(\boldsymbol{l},\boldsymbol{l}^*)$	(0,0,1,0)	0.00	0.00	0.00	1.00	0.00	0.00	0.00	1.00
Product	$\text{IE}^L(\boldsymbol{l},\boldsymbol{l}^*)$	(0,0,0,1)	0.00	0.00	0.00	1.00	0.00	0.00	0.00	1.00
Independent	$\text{TE}^L(\boldsymbol{l},\boldsymbol{l}^*)$	(1,0,0,0)	-0.04	0.19	0.20	0.96	-0.08	0.20	0.18	0.93
Independent	$\text{TE}^L(\boldsymbol{l},\boldsymbol{l}^*)$	(0,1,0,0)	-0.03	0.18	0.20	0.96	-0.06	0.19	0.18	0.92
Independent	$\text{TE}^L(\boldsymbol{l},\boldsymbol{l}^*)$	(0,0,1,0)	0.03	0.19	0.20	0.96	0.03	0.19	0.18	0.94
Independent	$\text{TE}^L(\boldsymbol{l},\boldsymbol{l}^*)$	(0,0,0,1)	0.02	0.18	0.20	0.97	0.01	0.18	0.18	0.94
Product	$\text{TE}^L(\boldsymbol{l},\boldsymbol{l}^*)$	(1,0,0,0)	-0.08	0.20	0.20	0.95	-0.09	0.21	0.18	0.93
Product	$\text{TE}^L(\boldsymbol{l},\boldsymbol{l}^*)$	(0,1,0,0)	-0.06	0.19	0.20	0.96	-0.07	0.20	0.18	0.93
Product	$\text{TE}^L(\boldsymbol{l},\boldsymbol{l}^*)$	(0,0,1,0)	0.03	0.19	0.20	0.96	0.03	0.19	0.18	0.94
Product	$\text{TE}^L(\boldsymbol{l},\boldsymbol{l}^*)$	(0,0,0,1)	0.02	0.18	0.20	0.97	0.01	0.18	0.18	0.93

Table C.7: Simulation results for the $n = 400$ IE zero simulation setting for the one-step and two-step estimators of $\text{DE}^L(\boldsymbol{l},\boldsymbol{l}^*)$, $\text{IE}^L(\boldsymbol{l},\boldsymbol{l}^*)$, and $\text{TE}^L(\boldsymbol{l},\boldsymbol{l}^*)$ when $R^2_L = 0.1$ and Adjusted $R^2_O = 0.6$.

			One-Step				Two-Step			
Shrinkage	**Parameter**	$(\boldsymbol{l}-\boldsymbol{l}^*)^\top$	**Bias**	**RMSE**	**Posterior SD**	**Coverage**	**Bias**	**RMSE**	**Posterior SD**	**Coverage**
Independent	$\text{DE}^L(\boldsymbol{l},\boldsymbol{l}^*)$	(1,0,0,0)	0.01	0.17	0.17	0.95	-0.06	0.17	0.15	0.92
Independent	$\text{DE}^L(\boldsymbol{l},\boldsymbol{l}^*)$	(0,1,0,0)	0.02	0.17	0.17	0.95	-0.04	0.16	0.15	0.92
Independent	$\text{DE}^L(\boldsymbol{l},\boldsymbol{l}^*)$	(0,0,1,0)	0.03	0.17	0.17	0.95	-0.04	0.16	0.15	0.94
Independent	$\text{DE}^L(\boldsymbol{l},\boldsymbol{l}^*)$	(0,0,0,1)	0.02	0.17	0.17	0.96	-0.05	0.16	0.15	0.93
Product	$\text{DE}^L(\boldsymbol{l},\boldsymbol{l}^*)$	(1,0,0,0)	-0.04	0.16	0.17	0.96	-0.06	0.17	0.15	0.93
Product	$\text{DE}^L(\boldsymbol{l},\boldsymbol{l}^*)$	(0,1,0,0)	-0.02	0.15	0.17	0.97	-0.05	0.16	0.15	0.92
Product	$\text{DE}^L(\boldsymbol{l},\boldsymbol{l}^*)$	(0,0,1,0)	-0.02	0.16	0.16	0.96	-0.04	0.16	0.15	0.94
Product	$\text{DE}^L(\boldsymbol{l},\boldsymbol{l}^*)$	(0,0,0,1)	-0.03	0.15	0.17	0.96	-0.05	0.16	0.15	0.93
Independent	$\text{IE}^L(\boldsymbol{l},\boldsymbol{l}^*)$	(1,0,0,0)	-0.08	0.11	0.07	0.76	-0.10	0.14	0.07	0.64
Independent	$\text{IE}^L(\boldsymbol{l},\boldsymbol{l}^*)$	(0,1,0,0)	-0.08	0.11	0.07	0.77	-0.10	0.14	0.07	0.64
Independent	$\text{IE}^L(\boldsymbol{l},\boldsymbol{l}^*)$	(0,0,1,0)	-0.08	0.11	0.07	0.78	-0.10	0.14	0.08	0.64
Independent	$\text{IE}^L(\boldsymbol{l},\boldsymbol{l}^*)$	(0,0,0,1)	-0.08	0.11	0.07	0.78	-0.10	0.14	0.07	0.64
Product	$\text{IE}^L(\boldsymbol{l},\boldsymbol{l}^*)$	(1,0,0,0)	-0.17	0.18	0.02	0.20	-0.18	0.19	0.02	0.14
Product	$\text{IE}^L(\boldsymbol{l},\boldsymbol{l}^*)$	(0,1,0,0)	-0.17	0.18	0.02	0.20	-0.18	0.19	0.02	0.13
Product	$\text{IE}^L(\boldsymbol{l},\boldsymbol{l}^*)$	(0,0,1,0)	-0.17	0.18	0.02	0.19	-0.18	0.19	0.02	0.12
Product	$\text{IE}^L(\boldsymbol{l},\boldsymbol{l}^*)$	(0,0,0,1)	-0.17	0.18	0.02	0.20	-0.18	0.19	0.02	0.14
Independent	$\text{TE}^L(\boldsymbol{l},\boldsymbol{l}^*)$	(1,0,0,0)	-0.07	0.20	0.19	0.92	-0.16	0.24	0.17	0.80
Independent	$\text{TE}^L(\boldsymbol{l},\boldsymbol{l}^*)$	(0,1,0,0)	-0.05	0.20	0.19	0.94	-0.14	0.23	0.17	0.82
Independent	$\text{TE}^L(\boldsymbol{l},\boldsymbol{l}^*)$	(0,0,1,0)	-0.05	0.20	0.19	0.93	-0.14	0.23	0.17	0.83
Independent	$\text{TE}^L(\boldsymbol{l},\boldsymbol{l}^*)$	(0,0,0,1)	-0.06	0.20	0.19	0.95	-0.15	0.23	0.17	0.83
Product	$\text{TE}^L(\boldsymbol{l},\boldsymbol{l}^*)$	(1,0,0,0)	-0.20	0.27	0.17	0.74	-0.23	0.29	0.15	0.62
Product	$\text{TE}^L(\boldsymbol{l},\boldsymbol{l}^*)$	(0,1,0,0)	-0.19	0.26	0.17	0.78	-0.22	0.28	0.15	0.66
Product	$\text{TE}^L(\boldsymbol{l},\boldsymbol{l}^*)$	(0,0,1,0)	-0.19	0.26	0.17	0.78	-0.22	0.28	0.15	0.67
Product	$\text{TE}^L(\boldsymbol{l},\boldsymbol{l}^*)$	(0,0,0,1)	-0.19	0.26	0.17	0.78	-0.23	0.28	0.15	0.64

Table C.8: Simulation results for the $n = 400$ IE sparse simulation setting for the one-step and two-step estimators of $\text{DE}^L(\boldsymbol{l},\boldsymbol{l}^*)$, $\text{IE}^L(\boldsymbol{l},\boldsymbol{l}^*)$, and $\text{TE}^L(\boldsymbol{l},\boldsymbol{l}^*)$ when $R^2_L = 0.1$ and Adjusted $R^2_O = 0.6$.

			One-Step				Two-Step			
Shrinkage	**Parameter**	$(\boldsymbol{l}-\boldsymbol{l}^*)^\top$	**Bias**	**RMSE**	**Posterior SD**	**Coverage**	**Bias**	**RMSE**	**Posterior SD**	**Coverage**
Independent	$\text{DE}^L(\boldsymbol{l},\boldsymbol{l}^*)$	(1,0,0,0)	0.02	0.18	0.17	0.95	-0.08	0.18	0.15	0.90
Independent	$\text{DE}^L(\boldsymbol{l},\boldsymbol{l}^*)$	(0,1,0,0)	0.03	0.18	0.17	0.94	-0.07	0.17	0.15	0.91
Independent	$\text{DE}^L(\boldsymbol{l},\boldsymbol{l}^*)$	(0,0,1,0)	0.03	0.19	0.17	0.93	-0.07	0.17	0.15	0.92
Independent	$\text{DE}^L(\boldsymbol{l},\boldsymbol{l}^*)$	(0,0,0,1)	0.03	0.18	0.17	0.94	-0.08	0.16	0.15	0.91
Product	$\text{DE}^L(\boldsymbol{l},\boldsymbol{l}^*)$	(1,0,0,0)	-0.07	0.17	0.16	0.94	-0.08	0.18	0.15	0.90
Product	$\text{DE}^L(\boldsymbol{l},\boldsymbol{l}^*)$	(0,1,0,0)	-0.06	0.16	0.16	0.94	-0.07	0.17	0.15	0.91
Product	$\text{DE}^L(\boldsymbol{l},\boldsymbol{l}^*)$	(0,0,1,0)	-0.06	0.16	0.16	0.95	-0.07	0.17	0.15	0.91
Product	$\text{DE}^L(\boldsymbol{l},\boldsymbol{l}^*)$	(0,0,0,1)	-0.06	0.16	0.16	0.94	-0.08	0.16	0.15	0.92
Independent	$\text{IE}^L(\boldsymbol{l},\boldsymbol{l}^*)$	(1,0,0,0)	-0.13	0.14	0.05	0.39	-0.18	0.18	0.03	0.09
Independent	$\text{IE}^L(\boldsymbol{l},\boldsymbol{l}^*)$	(0,1,0,0)	-0.13	0.14	0.05	0.40	-0.18	0.18	0.03	0.08
Independent	$\text{IE}^L(\boldsymbol{l},\boldsymbol{l}^*)$	(0,0,1,0)	-0.13	0.14	0.05	0.42	-0.18	0.18	0.03	0.09
Independent	$\text{IE}^L(\boldsymbol{l},\boldsymbol{l}^*)$	(0,0,0,1)	-0.13	0.14	0.05	0.42	-0.18	0.18	0.03	0.07
Product	$\text{IE}^L(\boldsymbol{l},\boldsymbol{l}^*)$	(1,0,0,0)	-0.20	0.20	0.00	0.02	-0.20	0.20	0.00	0.00
Product	$\text{IE}^L(\boldsymbol{l},\boldsymbol{l}^*)$	(0,1,0,0)	-0.20	0.20	0.00	0.02	-0.20	0.20	0.00	0.00
Product	$\text{IE}^L(\boldsymbol{l},\boldsymbol{l}^*)$	(0,0,1,0)	-0.20	0.20	0.00	0.03	-0.20	0.20	0.00	0.00
Product	$\text{IE}^L(\boldsymbol{l},\boldsymbol{l}^*)$	(0,0,0,1)	-0.20	0.20	0.00	0.02	-0.20	0.20	0.00	0.00
Independent	$\text{TE}^L(\boldsymbol{l},\boldsymbol{l}^*)$	(1,0,0,0)	-0.11	0.23	0.18	0.85	-0.26	0.31	0.15	0.56
Independent	$\text{TE}^L(\boldsymbol{l},\boldsymbol{l}^*)$	(0,1,0,0)	-0.10	0.22	0.18	0.87	-0.25	0.30	0.15	0.60
Independent	$\text{TE}^L(\boldsymbol{l},\boldsymbol{l}^*)$	(0,0,1,0)	-0.09	0.23	0.18	0.87	-0.25	0.29	0.15	0.61
Independent	$\text{TE}^L(\boldsymbol{l},\boldsymbol{l}^*)$	(0,0,0,1)	-0.10	0.23	0.18	0.88	-0.26	0.30	0.15	0.57
Product	$\text{TE}^L(\boldsymbol{l},\boldsymbol{l}^*)$	(1,0,0,0)	-0.27	0.31	0.16	0.58	-0.28	0.32	0.15	0.50
Product	$\text{TE}^L(\boldsymbol{l},\boldsymbol{l}^*)$	(0,1,0,0)	-0.26	0.30	0.16	0.63	-0.27	0.31	0.15	0.54
Product	$\text{TE}^L(\boldsymbol{l},\boldsymbol{l}^*)$	(0,0,1,0)	-0.25	0.30	0.16	0.65	-0.27	0.31	0.15	0.55
Product	$\text{TE}^L(\boldsymbol{l},\boldsymbol{l}^*)$	(0,0,0,1)	-0.26	0.30	0.16	0.62	-0.28	0.31	0.15	0.51

Table C.9: Simulation results for the $n = 400$ IE dense simulation setting for the one-step and two-step estimators of $\text{DE}^L(\boldsymbol{l},\boldsymbol{l}^*)$, $\text{IE}^L(\boldsymbol{l},\boldsymbol{l}^*)$, and $\text{TE}^L(\boldsymbol{l},\boldsymbol{l}^*)$ when $R_L^2 = 0.1$ and Adjusted $R_O^2 = 0.6$.

			One-Step				Two-Step			
Shrinkage	**Parameter**	$(\boldsymbol{l}-\boldsymbol{l}^*)^\top$	**Bias**	**RMSE**	**Posterior SD**	**Coverage**	**Bias**	**RMSE**	**Posterior SD**	**Coverage**
Independent	$\text{DE}^L(\boldsymbol{l},\boldsymbol{l}^*)$	(1,0,0,0)	-0.05	0.28	0.29	0.95	-0.07	0.29	0.27	0.94
Independent	$\text{DE}^L(\boldsymbol{l},\boldsymbol{l}^*)$	(0,1,0,0)	-0.03	0.27	0.29	0.97	-0.05	0.27	0.27	0.94
Independent	$\text{DE}^L(\boldsymbol{l},\boldsymbol{l}^*)$	(0,0,1,0)	0.04	0.27	0.28	0.96	0.03	0.27	0.26	0.95
Independent	$\text{DE}^L(\boldsymbol{l},\boldsymbol{l}^*)$	(0,0,0,1)	0.02	0.26	0.28	0.96	0.01	0.26	0.26	0.95
Product	$\text{DE}^L(\boldsymbol{l},\boldsymbol{l}^*)$	(1,0,0,0)	-0.07	0.28	0.28	0.95	-0.07	0.29	0.27	0.94
Product	$\text{DE}^L(\boldsymbol{l},\boldsymbol{l}^*)$	(0,1,0,0)	-0.05	0.27	0.28	0.97	-0.05	0.27	0.27	0.95
Product	$\text{DE}^L(\boldsymbol{l},\boldsymbol{l}^*)$	(0,0,1,0)	0.03	0.27	0.28	0.96	0.03	0.27	0.26	0.94
Product	$\text{DE}^L(\boldsymbol{l},\boldsymbol{l}^*)$	(0,0,0,1)	0.02	0.26	0.28	0.96	0.01	0.26	0.26	0.95
Independent	$\text{IE}^L(\boldsymbol{l},\boldsymbol{l}^*)$	(1,0,0,0)	0.00	0.01	0.04	1.00	0.01	0.01	0.03	1.00
Independent	$\text{IE}^L(\boldsymbol{l},\boldsymbol{l}^*)$	(0,1,0,0)	0.00	0.01	0.04	1.00	0.01	0.01	0.03	1.00
Independent	$\text{IE}^L(\boldsymbol{l},\boldsymbol{l}^*)$	(0,0,1,0)	0.00	0.01	0.02	1.00	0.00	0.00	0.02	1.00
Independent	$\text{IE}^L(\boldsymbol{l},\boldsymbol{l}^*)$	(0,0,0,1)	0.00	0.01	0.02	1.00	0.00	0.00	0.02	1.00
Product	$\text{IE}^L(\boldsymbol{l},\boldsymbol{l}^*)$	(1,0,0,0)	0.00	0.00	0.00	1.00	0.00	0.00	0.00	1.00
Product	$\text{IE}^L(\boldsymbol{l},\boldsymbol{l}^*)$	(0,1,0,0)	0.00	0.00	0.00	1.00	0.00	0.00	0.00	1.00
Product	$\text{IE}^L(\boldsymbol{l},\boldsymbol{l}^*)$	(0,0,1,0)	0.00	0.00	0.00	1.00	0.00	0.00	0.00	1.00
Product	$\text{IE}^L(\boldsymbol{l},\boldsymbol{l}^*)$	(0,0,0,1)	0.00	0.00	0.00	1.00	0.00	0.00	0.00	1.00
Independent	$\text{TE}^L(\boldsymbol{l},\boldsymbol{l}^*)$	(1,0,0,0)	-0.04	0.28	0.28	0.95	-0.06	0.29	0.27	0.94
Independent	$\text{TE}^L(\boldsymbol{l},\boldsymbol{l}^*)$	(0,1,0,0)	-0.02	0.26	0.28	0.96	-0.04	0.27	0.27	0.95
Independent	$\text{TE}^L(\boldsymbol{l},\boldsymbol{l}^*)$	(0,0,1,0)	0.03	0.27	0.28	0.96	0.03	0.27	0.26	0.95
Independent	$\text{TE}^L(\boldsymbol{l},\boldsymbol{l}^*)$	(0,0,0,1)	0.02	0.26	0.28	0.96	0.01	0.26	0.26	0.95
Product	$\text{TE}^L(\boldsymbol{l},\boldsymbol{l}^*)$	(1,0,0,0)	-0.07	0.28	0.28	0.95	-0.07	0.29	0.27	0.94
Product	$\text{TE}^L(\boldsymbol{l},\boldsymbol{l}^*)$	(0,1,0,0)	-0.05	0.27	0.28	0.97	-0.05	0.27	0.27	0.95
Product	$\text{TE}^L(\boldsymbol{l},\boldsymbol{l}^*)$	(0,0,1,0)	0.03	0.27	0.28	0.96	0.03	0.27	0.26	0.94
Product	$\text{TE}^L(\boldsymbol{l},\boldsymbol{l}^*)$	(0,0,0,1)	0.02	0.26	0.28	0.96	0.01	0.26	0.26	0.95

Table C.10: Simulation results for the $n = 400$ IE zero simulation setting for the one-step and two-step estimators of $\text{DE}^L(\boldsymbol{l},\boldsymbol{l}^*)$, $\text{IE}^L(\boldsymbol{l},\boldsymbol{l}^*)$, and $\text{TE}^L(\boldsymbol{l},\boldsymbol{l}^*)$ when $R^2_L = 0.1$ and Adjusted $R^2_O = 0.3$.

			One-Step				Two-Step			
Shrinkage	**Parameter**	$(\boldsymbol{l}-\boldsymbol{l}^*)^\top$	**Bias**	**RMSE**	**Posterior SD**	**Coverage**	**Bias**	**RMSE**	**Posterior SD**	**Coverage**
Independent	$\text{DE}^L(\boldsymbol{l},\boldsymbol{l}^*)$	(1,0,0,0)	0.03	0.25	0.25	0.95	-0.05	0.24	0.22	0.94
Independent	$\text{DE}^L(\boldsymbol{l},\boldsymbol{l}^*)$	(0,1,0,0)	0.04	0.24	0.25	0.95	-0.03	0.23	0.23	0.94
Independent	$\text{DE}^L(\boldsymbol{l},\boldsymbol{l}^*)$	(0,0,1,0)	0.05	0.25	0.25	0.95	-0.02	0.23	0.22	0.94
Independent	$\text{DE}^L(\boldsymbol{l},\boldsymbol{l}^*)$	(0,0,0,1)	0.04	0.24	0.25	0.95	-0.04	0.23	0.22	0.95
Product	$\text{DE}^L(\boldsymbol{l},\boldsymbol{l}^*)$	(1,0,0,0)	-0.04	0.24	0.24	0.95	-0.05	0.24	0.22	0.94
Product	$\text{DE}^L(\boldsymbol{l},\boldsymbol{l}^*)$	(0,1,0,0)	-0.03	0.23	0.24	0.96	-0.03	0.23	0.23	0.94
Product	$\text{DE}^L(\boldsymbol{l},\boldsymbol{l}^*)$	(0,0,1,0)	-0.02	0.23	0.24	0.96	-0.03	0.23	0.22	0.94
Product	$\text{DE}^L(\boldsymbol{l},\boldsymbol{l}^*)$	(0,0,0,1)	-0.03	0.22	0.24	0.96	-0.04	0.22	0.22	0.95
Independent	$\text{IE}^L(\boldsymbol{l},\boldsymbol{l}^*)$	(1,0,0,0)	-0.09	0.12	0.07	0.72	-0.11	0.14	0.07	0.63
Independent	$\text{IE}^L(\boldsymbol{l},\boldsymbol{l}^*)$	(0,1,0,0)	-0.09	0.12	0.07	0.72	-0.10	0.14	0.07	0.64
Independent	$\text{IE}^L(\boldsymbol{l},\boldsymbol{l}^*)$	(0,0,1,0)	-0.09	0.11	0.07	0.73	-0.10	0.14	0.07	0.64
Independent	$\text{IE}^L(\boldsymbol{l},\boldsymbol{l}^*)$	(0,0,0,1)	-0.09	0.12	0.07	0.72	-0.10	0.14	0.07	0.63
Product	$\text{IE}^L(\boldsymbol{l},\boldsymbol{l}^*)$	(1,0,0,0)	-0.19	0.20	0.01	0.06	-0.19	0.19	0.01	0.10
Product	$\text{IE}^L(\boldsymbol{l},\boldsymbol{l}^*)$	(0,1,0,0)	-0.19	0.20	0.01	0.05	-0.19	0.19	0.01	0.10
Product	$\text{IE}^L(\boldsymbol{l},\boldsymbol{l}^*)$	(0,0,1,0)	-0.19	0.20	0.01	0.05	-0.19	0.19	0.01	0.09
Product	$\text{IE}^L(\boldsymbol{l},\boldsymbol{l}^*)$	(0,0,0,1)	-0.19	0.20	0.01	0.05	-0.19	0.19	0.01	0.10
Independent	$\text{TE}^L(\boldsymbol{l},\boldsymbol{l}^*)$	(1,0,0,0)	-0.07	0.27	0.26	0.93	-0.15	0.30	0.24	0.86
Independent	$\text{TE}^L(\boldsymbol{l},\boldsymbol{l}^*)$	(0,1,0,0)	-0.05	0.26	0.26	0.95	-0.14	0.28	0.24	0.89
Independent	$\text{TE}^L(\boldsymbol{l},\boldsymbol{l}^*)$	(0,0,1,0)	-0.04	0.26	0.26	0.95	-0.13	0.28	0.24	0.91
Independent	$\text{TE}^L(\boldsymbol{l},\boldsymbol{l}^*)$	(0,0,0,1)	-0.05	0.26	0.26	0.95	-0.14	0.28	0.24	0.90
Product	$\text{TE}^L(\boldsymbol{l},\boldsymbol{l}^*)$	(1,0,0,0)	-0.24	0.34	0.24	0.82	-0.23	0.34	0.23	0.79
Product	$\text{TE}^L(\boldsymbol{l},\boldsymbol{l}^*)$	(0,1,0,0)	-0.22	0.32	0.24	0.84	-0.22	0.32	0.23	0.83
Product	$\text{TE}^L(\boldsymbol{l},\boldsymbol{l}^*)$	(0,0,1,0)	-0.21	0.31	0.24	0.86	-0.21	0.31	0.23	0.84
Product	$\text{TE}^L(\boldsymbol{l},\boldsymbol{l}^*)$	(0,0,0,1)	-0.23	0.32	0.24	0.86	-0.22	0.32	0.23	0.84

Table C.11: Simulation results for the $n = 400$ IE sparse simulation setting for the one-step and two-step estimators of $\text{DE}^L(\boldsymbol{l},\boldsymbol{l}^*)$, $\text{IE}^L(\boldsymbol{l},\boldsymbol{l}^*)$, and $\text{TE}^L(\boldsymbol{l},\boldsymbol{l}^*)$ when $R^2_L = 0.1$ and Adjusted $R^2_O = 0.3$.

			One-Step				Two-Step			
Shrinkage	**Parameter**	$(\boldsymbol{l}-\boldsymbol{l}^*)^\top$	**Bias**	**RMSE**	**Posterior SD**	**Coverage**	**Bias**	**RMSE**	**Posterior SD**	**Coverage**
Independent	$\text{DE}^L(\boldsymbol{l},\boldsymbol{l}^*)$	(1,0,0,0)	0.04	0.26	0.24	0.95	-0.06	0.24	0.22	0.93
Independent	$\text{DE}^L(\boldsymbol{l},\boldsymbol{l}^*)$	(0,1,0,0)	0.06	0.25	0.25	0.94	-0.04	0.23	0.22	0.94
Independent	$\text{DE}^L(\boldsymbol{l},\boldsymbol{l}^*)$	(0,0,1,0)	0.07	0.26	0.24	0.94	-0.04	0.23	0.22	0.95
Independent	$\text{DE}^L(\boldsymbol{l},\boldsymbol{l}^*)$	(0,0,0,1)	0.06	0.25	0.24	0.94	-0.05	0.22	0.22	0.94
Product	$\text{DE}^L(\boldsymbol{l},\boldsymbol{l}^*)$	(1,0,0,0)	-0.05	0.24	0.23	0.94	-0.06	0.24	0.22	0.93
Product	$\text{DE}^L(\boldsymbol{l},\boldsymbol{l}^*)$	(0,1,0,0)	-0.04	0.23	0.23	0.96	-0.04	0.23	0.22	0.94
Product	$\text{DE}^L(\boldsymbol{l},\boldsymbol{l}^*)$	(0,0,1,0)	-0.03	0.23	0.23	0.96	-0.04	0.23	0.22	0.94
Product	$\text{DE}^L(\boldsymbol{l},\boldsymbol{l}^*)$	(0,0,0,1)	-0.04	0.22	0.23	0.96	-0.05	0.22	0.22	0.94
Independent	$\text{IE}^L(\boldsymbol{l},\boldsymbol{l}^*)$	(1,0,0,0)	-0.14	0.15	0.05	0.35	-0.18	0.18	0.03	0.07
Independent	$\text{IE}^L(\boldsymbol{l},\boldsymbol{l}^*)$	(0,1,0,0)	-0.14	0.15	0.05	0.34	-0.18	0.18	0.03	0.08
Independent	$\text{IE}^L(\boldsymbol{l},\boldsymbol{l}^*)$	(0,0,1,0)	-0.14	0.15	0.05	0.36	-0.18	0.18	0.03	0.09
Independent	$\text{IE}^L(\boldsymbol{l},\boldsymbol{l}^*)$	(0,0,0,1)	-0.14	0.15	0.05	0.36	-0.18	0.18	0.03	0.07
Product	$\text{IE}^L(\boldsymbol{l},\boldsymbol{l}^*)$	(1,0,0,0)	-0.20	0.20	0.00	0.01	-0.20	0.20	0.00	0.00
Product	$\text{IE}^L(\boldsymbol{l},\boldsymbol{l}^*)$	(0,1,0,0)	-0.20	0.20	0.00	0.00	-0.20	0.20	0.00	0.00
Product	$\text{IE}^L(\boldsymbol{l},\boldsymbol{l}^*)$	(0,0,1,0)	-0.20	0.20	0.00	0.01	-0.20	0.20	0.00	0.00
Product	$\text{IE}^L(\boldsymbol{l},\boldsymbol{l}^*)$	(0,0,0,1)	-0.20	0.20	0.00	0.01	-0.20	0.20	0.00	0.00
Independent	$\text{TE}^L(\boldsymbol{l},\boldsymbol{l}^*)$	(1,0,0,0)	-0.10	0.28	0.25	0.91	-0.24	0.34	0.22	0.77
Independent	$\text{TE}^L(\boldsymbol{l},\boldsymbol{l}^*)$	(0,1,0,0)	-0.09	0.27	0.25	0.92	-0.22	0.32	0.22	0.82
Independent	$\text{TE}^L(\boldsymbol{l},\boldsymbol{l}^*)$	(0,0,1,0)	-0.07	0.27	0.25	0.93	-0.22	0.31	0.22	0.83
Independent	$\text{TE}^L(\boldsymbol{l},\boldsymbol{l}^*)$	(0,0,0,1)	-0.09	0.27	0.25	0.93	-0.23	0.32	0.22	0.82
Product	$\text{TE}^L(\boldsymbol{l},\boldsymbol{l}^*)$	(1,0,0,0)	-0.25	0.34	0.23	0.79	-0.26	0.35	0.22	0.75
Product	$\text{TE}^L(\boldsymbol{l},\boldsymbol{l}^*)$	(0,1,0,0)	-0.24	0.33	0.23	0.82	-0.24	0.33	0.22	0.80
Product	$\text{TE}^L(\boldsymbol{l},\boldsymbol{l}^*)$	(0,0,1,0)	-0.23	0.32	0.23	0.83	-0.24	0.33	0.22	0.81
Product	$\text{TE}^L(\boldsymbol{l},\boldsymbol{l}^*)$	(0,0,0,1)	-0.24	0.33	0.23	0.83	-0.25	0.33	0.22	0.80

Table C.12: Simulation results for the $n = 400$ IE dense simulation setting for the one-step and two-step estimators of $\text{DE}^L(\boldsymbol{l},\boldsymbol{l}^*)$, $\text{IE}^L(\boldsymbol{l},\boldsymbol{l}^*)$, and $\text{TE}^L(\boldsymbol{l},\boldsymbol{l}^*)$ when $R^2_L = 0.1$ and Adjusted $R^2_O = 0.3$.

	Estimated Number of Latent Factors			
Method	**4***	**5**	**6**	**7**
$n = 400$				
One-Step Independent	999 (99.9%)	1 (0.1%)	0 (0.0%)	0 (0.0%)
One-Step Product	999 (99.9%)	1 (0.1%)	0 (0.0%)	0 (0.0%)
Two-Step	236 (23.6%)	543 (54.3%)	205 (20.5%)	16 (1.6%)
$n = 2000$				
One-Step Independent	603 (60.3%)	365 (36.5%)	31 (3.1%)	1 (0.1%)
One-Step Product	623 (62.3%)	347 (34.7%)	29 (2.9%)	1 (0.1%)
Two-Step	13 (1.3%)	311 (31.1%)	406 (40.6%)	270 (27.0%)

Table C.13: Results for estimating of the number of latent factors under the IE zero simulation settings when $R_L^2 = 0.5$ and Adjusted $R_O^2 = 0.3$. The numbers in the table indicate the number of times out of 1000 simulation replications each number of latent factors was selected via BIC. *The true number of latent factors is $r = 4$.

	Estimated Number of Latent Factors			
Method	**4***	**5**	**6**	**7**
$n = 400$				
One-Step Independent	999 (99.9%)	1 (0.1%)	0 (0.0%)	0 (0.0%)
One-Step Product	998 (99.8%)	2 (0.2%)	0 (0.0%)	0 (0.0%)
Two-Step	232 (23.2%)	565 (56.5%)	187 (18.7%)	16 (1.6%)
$n = 2000$				
One-Step Independent	594 (59.4%)	374 (37.4%)	30 (3.0%)	2 (0.2%)
One-Step Product	613 (61.3%)	359 (35.9%)	26 (2.6%)	2 (0.2%)
Two-Step	15 (1.5%)	331 (33.1%)	399 (39.9%)	255 (25.5%)

Table C.14: Results for estimating of the number of latent factors under the IE sparse simulation settings when $R_L^2 = 0.5$ and Adjusted $R_O^2 = 0.3$. The numbers in the table indicate the number of times out of 1000 simulation replications each number of latent factors was selected via BIC. *The true number of latent factors is $r = 4$.

C.4 Supplementary Results for Data Example

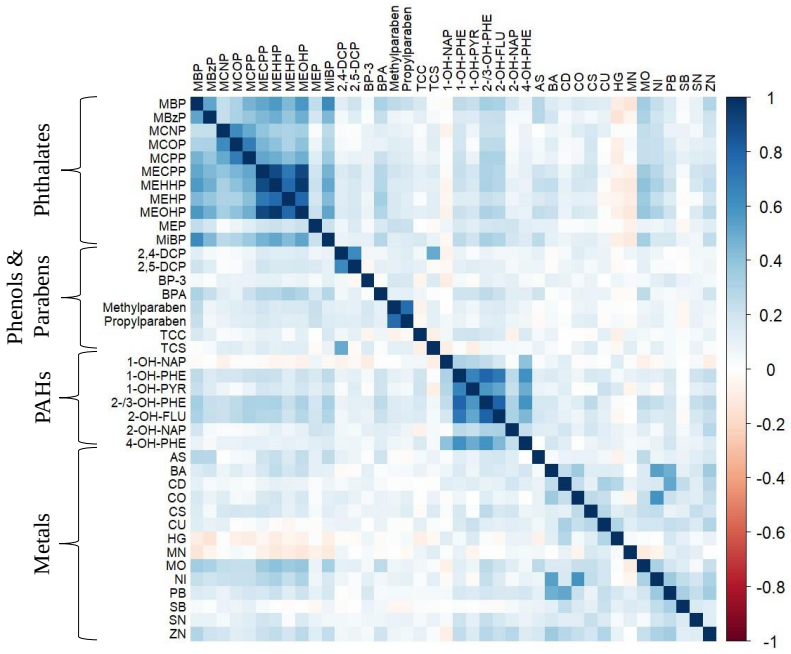


Figure C.1: Pairwise spearman rank correlations for specific-gravity corrected chemical concentrations in PROTECT corresponding to the total effect only mediation mixture map models ($n = 478$).

Eicosanoid	Pathway	Latent Factor 1	Latent Factor 2	Latent Factor 3	Latent Factor 4
11,12-DHET	Cytochrome p450	0.01 (-0.02, 0.04)	-0.03 (-0.09, 0.04)	-0.01 (-0.06, 0.03)	-0.07 (-0.23, 0.20)
11(12)-EET	Cytochrome p450	0.02 (-0.01, 0.07)	0.05 (-0.02, 0.13)	-0.03 (-0.10, 0.01)	-0.02 (-0.14, 0.11)
11-HETE	Cytochrome p450	0.01 (-0.08, 0.10)	**-0.19 (-0.29, -0.09)**	-0.03 (-0.17, 0.14)	-0.38 (-0.79, 0.77)
12,13-DiHOME	Cytochrome p450	0.00 (-0.05, 0.06)	**-0.12 (-0.20, -0.05)**	-0.01 (-0.09, 0.09)	-0.22 (-0.50, 0.47)
12(13)-EpoME	Cytochrome p450	0.01 (-0.04, 0.07)	**0.14 (0.08, 0.22)**	-0.01 (-0.10, 0.08)	0.21 (-0.45, 0.48)
14(15)-EET	Cytochrome p450	-0.01 (-0.09, 0.07)	**-0.22 (-0.32, -0.13)**	-0.01 (-0.13, 0.15)	-0.35 (-0.75, 0.72)
16-HETE	Cytochrome p450	-0.01 (-0.09, 0.08)	**-0.24 (-0.34, -0.14)**	-0.01 (-0.14, 0.16)	-0.37 (-0.78, 0.75)
17-HETE	Cytochrome p450	0.00 (-0.09, 0.08)	**-0.24 (-0.34, -0.14)**	-0.01 (-0.15, 0.17)	-0.39 (-0.82, 0.79)
18-HETE	Cytochrome p450	0.01 (-0.08, 0.10)	**-0.22 (-0.32, -0.12)**	-0.02 (-0.15, 0.15)	-0.38 (-0.79, 0.77)
CAA	Cytochrome p450	0.00 (-0.10, 0.10)	**-0.27 (-0.38, -0.16)**	-0.02 (-0.18, 0.19)	-0.46 (-0.94, 0.91)
20-HETE	Cytochrome p450	-0.01 (-0.05, 0.03)	0.05 (-0.02, 0.11)	0.02 (-0.04, 0.08)	0.12 (-0.28, 0.31)
5,6-DHET	Cytochrome p450	0.00 (-0.10, 0.09)	**-0.25 (-0.35, -0.15)**	-0.01 (-0.17, 0.18)	-0.43 (-0.89, 0.86)
5(6)-EET	Cytochrome p450	0.01 (-0.04, 0.05)	**-0.07 (-0.14, 0.00)**	-0.01 (-0.09, 0.06)	-0.15 (-0.38, 0.35)
8,9-DHET	Cytochrome p450	-0.01 (-0.09, 0.07)	**-0.21 (-0.30, -0.12)**	0.00 (-0.12, 0.15)	-0.33 (-0.70, 0.68)
8(9)-EET	Cytochrome p450	0.02 (-0.01, 0.07)	**0.08 (0.01, 0.16)**	-0.03 (-0.09, 0.01)	0.04 (-0.14, 0.17)
9,10-DiHOME	Cytochrome p450	-0.01 (-0.05, 0.02)	**-0.09 (-0.16, -0.03)**	0.01 (-0.04, 0.08)	-0.09 (-0.28, 0.25)
9(10)-EpoME	Cytochrome p450	-0.01 (-0.05, 0.03)	**-0.11 (-0.19, -0.05)**	0.01 (-0.06, 0.08)	-0.15 (-0.38, 0.35)
9S-HODE	Cytochrome p450	-0.01 (-0.08, 0.06)	**-0.18 (-0.27, -0.10)**	0.00 (-0.10, 0.14)	-0.29 (-0.64, 0.61)
LTB4	Lipoxygenase	0.03 (-0.03, 0.09)	**0.20 (0.11, 0.30)**	-0.04 (-0.15, 0.06)	0.21 (-0.45, 0.48)
LTD4	Lipoxygenase	0.00 (-0.05, 0.05)	**-0.11 (-0.18, -0.04)**	0.00 (-0.08, 0.08)	-0.17 (-0.41, 0.38)
LTE4	Lipoxygenase	0.02 (-0.01, 0.08)	0.06 (-0.01, 0.15)	-0.04 (-0.11, 0.01)	0.01 (-0.10, 0.12)
RVD1	Lipoxygenase	0.01 (-0.03, 0.06)	**0.10 (0.04, 0.18)**	-0.01 (-0.09, 0.05)	0.14 (-0.32, 0.34)
RVD2	Lipoxygenase	0.00 (-0.06, 0.05)	**-0.12 (-0.19, -0.05)**	0.00 (-0.08, 0.09)	-0.20 (-0.46, 0.43)
12-HETE	Lipoxygenase	-0.01 (-0.04, 0.02)	0.00 (-0.07, 0.06)	0.01 (-0.03, 0.06)	0.03 (-0.12, 0.15)
12-oxoETE	Lipoxygenase	-0.02 (-0.10, 0.06)	**-0.24 (-0.34, -0.15)**	0.01 (-0.12, 0.17)	-0.35 (-0.74, 0.71)
13-oxoODE	Lipoxygenase	0.00 (-0.07, 0.06)	**-0.17 (-0.25, -0.09)**	0.00 (-0.11, 0.12)	-0.26 (-0.58, 0.55)
13S-HODE	Lipoxygenase	-0.02 (-0.09, 0.05)	**-0.20 (-0.30, -0.11)**	0.02 (-0.09, 0.14)	-0.26 (-0.57, 0.54)
15-HETE	Lipoxygenase	0.00 (-0.05, 0.05)	**-0.12 (-0.19, -0.05)**	0.00 (-0.08, 0.09)	-0.20 (-0.47, 0.44)
15-oxoETE	Lipoxygenase	-0.01 (-0.08, 0.06)	**-0.17 (-0.26, -0.09)**	-0.01 (-0.11, 0.12)	-0.28 (-0.62, 0.59)
5-HETE	Lipoxygenase	0.01 (-0.07, 0.09)	**-0.17 (-0.27, -0.08)**	-0.03 (-0.16, 0.13)	-0.35 (-0.74, 0.72)
5-oxoETE	Lipoxygenase	0.01 (-0.07, 0.10)	**-0.17 (-0.26, -0.07)**	-0.03 (-0.17, 0.12)	-0.35 (-0.73, 0.71)
8-HETE	Lipoxygenase	0.03 (-0.06, 0.12)	**-0.14 (-0.24, -0.03)**	-0.05 (-0.19, 0.12)	-0.37 (-0.77, 0.74)

Table C.15: One-step mediation mixture map results for $\boldsymbol{\alpha}_a$ with independent shrinkage corresponding to gestational age at delivery ($n = 466$). Estimates in the table are Posterior Mean (95% Credible Interval). Bolded entries correspond to credible intervals that do not cover zero. Latent Factor 2 summarizes exposure to MCNP, MCOP, and MCPP.

Eicosanoid	Pathway	Posterior Mean (95% CI)
11,12-DHET	Cytochrome p450	-0.05 (-0.25, 0.03)
11(12)-EET	Cytochrome p450	-0.02 (-0.16, 0.06)
11-HETE	Cytochrome p450	-0.02 (-0.19, 0.08)
12,13-DiHOME	Cytochrome p450	0.02 (-0.07, 0.14)
12(13)-EpoME	Cytochrome p450	0.02 (-0.07, 0.20)
14(15)-EET	Cytochrome p450	0.01 (-0.10, 0.12)
16-HETE	Cytochrome p450	-0.02 (-0.17, 0.07)
17-HETE	Cytochrome p450	-0.01 (-0.15, 0.09)
18-HETE	Cytochrome p450	0.02 (-0.07, 0.18)
CAA	Cytochrome p450	0.05 (-0.06, 0.40)
20-HETE	Cytochrome p450	-0.01 (-0.11, 0.08)
5,6-DHET	Cytochrome p450	-0.03 (-0.23, 0.07)
5(6)-EET	Cytochrome p450	-0.01 (-0.13, 0.07)
8,9-DHET	Cytochrome p450	-0.02 (-0.19, 0.06)
8(9)-EET	Cytochrome p450	0.00 (-0.11, 0.10)
9,10-DiHOME	Cytochrome p450	-0.01 (-0.11, 0.09)
9(10)-EpoME	Cytochrome p450	-0.03 (-0.21, 0.06)
9S-HODE	Cytochrome p450	0.02 (-0.07, 0.17)
LTB4	Lipoxygenase	-0.01 (-0.14, 0.08)
LTD4	Lipoxygenase	-0.03 (-0.17, 0.06)
LTE4	Lipoxygenase	-0.05 (-0.24, 0.04)
RVD1	Lipoxygenase	0.00 (-0.09, 0.10)
RVD2	Lipoxygenase	0.01 (-0.09, 0.11)
12-HETE	Lipoxygenase	-0.03 (-0.19, 0.05)
12-oxoETE	Lipoxygenase	0.02 (-0.08, 0.16)
13-oxoODE	Lipoxygenase	0.04 (-0.05, 0.24)
13S-HODE	Lipoxygenase	0.02 (-0.06, 0.16)
15-HETE	Lipoxygenase	0.00 (-0.10, 0.12)
15-oxoETE	Lipoxygenase	0.05 (-0.05, 0.27)
5-HETE	Lipoxygenase	-0.01 (-0.13, 0.10)
5-oxoETE	Lipoxygenase	0.01 (-0.10, 0.13)
8-HETE	Lipoxygenase	-0.04 (-0.23, 0.06)

Table C.16: One-step mediation mixture map results for $\boldsymbol{\beta}_m$ with independent shrinkage corresponding to gestational age at delivery ($n = 466$). Estimates in the table are Posterior Mean (95% Credible Interval). Bolded entries correspond to credible intervals that do not cover zero.

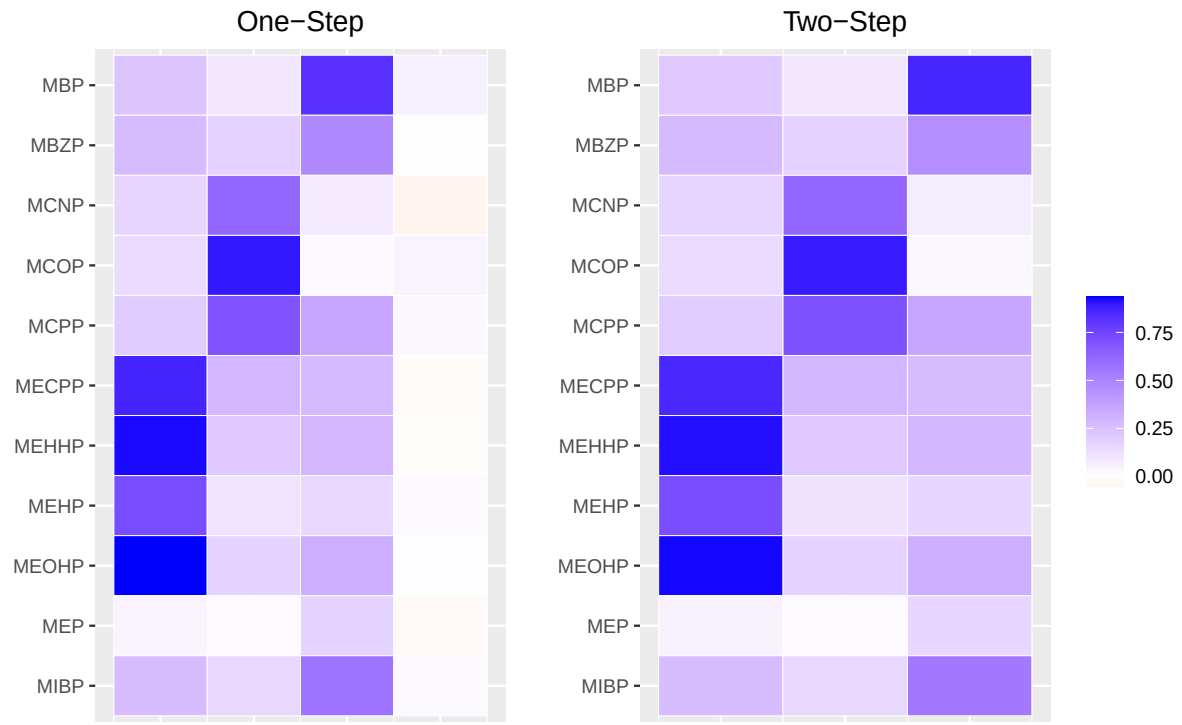


Figure C.2: Estimated loadings matricies corresponding to the mediation mixture model with product shrinkage for gestational age at delivery ($n = 466$). Gestational age at delivery models are adjusted for maternal age and educational attainment.

Exposure	**IE**	**IE 95% CI**	**TE**	**TE 95% CI**	**% Mediated**	**Largest Weights**
One-Step						
Latent Factor 1	0.00	(-0.03, 0.03)	0.06	(-0.18, 0.29)	1%	MECPP, MEHHP, MEHP, MEOHP
Latent Factor 2	-0.01	(-0.16, 0.12)	-0.01	(-0.24, 0.22)	57%	MCNP, MCOP, MCPP
Latent Factor 3	0.00	(-0.04, 0.04)	-0.28	(-0.54, -0.03)	0%	MBP, MBzP, MiBP
All Phthalates	-0.01	(-0.17, 0.14)	-0.23	(-0.56, 0.12)	4%	
Two-Step						
Latent Factor 1	0.00	(0.00, 0.00)	0.05	(-0.16, 0.27)	0%	MECPP, MEHHP, MEHP, MEOHP
Latent Factor 2	0.00	(-0.03, 0.01)	0.14	(-0.08, 0.36)	-	MCNP, MCOP, MCPP
Latent Factor 3	0.00	(0.00, 0.00)	-0.24	(-0.45, -0.03)	0%	MBP, MBzP, MiBP
All Phthalates	0.00	(-0.03, 0.01)	-0.03	(-0.35, 0.29)	9%	

Table C.17: One-step and two-step mediation mixture map results with product shrinkage corresponding to gestational age at delivery ($n = 466$). Estimated IEs and TEs in the table correspond to changes in gestational age at delivery (weeks) for an IQR increase in exposure.

BIBLIOGRAPHY

Abramowitz, M. and Stegun, I. A. (1964). *Handbook of mathematical functions with formulas, graphs, and mathematical tables*, volume 55. US Government printing office.

Ahuja, K., Hartford, J., and Bengio, Y. (2021). "Properties from Mechanisms: An Equivariance Perspective on Identifiable Representation Learning." *arXiv Preprint*, 1–24.

Andrade, J. A. A. and O'Hagan, A. (2006). "Bayesian Robustness Modeling Using Regularly Varying Distributions." *Bayesian Analysis*, 1(1): 169–188.

Armagan, A., Dunson, D. B., and Lee, J. (2013a). "Generalized Double Pareto Shrinkage." *Statistica Sinica*, 23(1): 119–143.

Armagan, A., Dunson, D. B., Lee, J., Bajwa, W. U., and Strawn, N. (2013b). "Posterior consistency in linear models under shrinkage priors." *Biometrika*, 100(4): 1011–1018.

Arunajadai, S. G. and Rauh, V. A. (2012). "Handling covariates subject to limits of detection in regression." *Environmental and Ecological Statistics*, 19: 369–391.

Atem, F. D., Qian, J., Maye, J. E., Johnson, K. A., and Betensky, R. A. (2017). "Linear regression with a randomly censored covariate: application to an Alzheimer's study." *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, 66(2): 313–328.

Aung, M. T., Song, Y., Ferguson, K. K., Cantonwine, D. E., Zeng, L., McElrath, T. F., Pennathur, S., Meeker, J. D., and Mukherjee, B. (2020). "Application of an analytical framework for multivariate mediation analysis of environmental data." *Nature Communications*, 11: 5624.

Aung, M. T., Yu, Y., Ferguson, K. K., Cantonwine, D. E., Zeng, L., McElrath, T. F., Pennathur, S., Mukherjee, B., and Meeker, J. D. (2019). "Prediction and associations of preterm birth and its subtypes with eicosanoid enzymatic pathways and inflammatory markers." *Scientific Reports*, 9: 17049.

Bai, R. and Ghosh, M. (2019). "Large-scale multiple hypothesis testing with the normal-beta prime prior." *Statistics*, 53(6): 1210–1233.

Bai, R., Moran, G. E., Antonelli, J. L., Chen, Y., and Boland, M. R. (2022). "Spike-and-Slab Group Lassos for Grouped Regression and Sparse Generalized Additive Models." *Journal of the American Statistical Association*, 117(537): 184–197.

Barndorff-Nielsen, O., Kent, J., and Sørensen, M. (1982). "Normal Variance-Mean Mixtures and z Distributions." *International Statistical Review*, 50(2): 145–159.

Baron, R. M. and Kenny, D. A. (1986). "The Moderator-Mediator Variable Distinction in Social Psychological Research: Conceptual, Strategic, and Statistical Considerations." *Journal of Personality and Social Psychology*, 51(6): 1173–1182.

Bernhardt, P. W., Wang, H. J., and Zhang, D. (2015). "Statistical Methods for Generalized Linear Models with Covariates Subject to Detection Limits." *Statistics in Biosciences*, 7(1): 68–89.

Bhadra, A., Datta, J., Polson, N. G., and Willard, B. (2016). "Default Bayesian analysis with global-local shrinkage priors." *Biometrika*, 103(4): 955–969.

— (2017). "The Horseshoe+ Estimator of Ultra-Sparse Signals." *Bayesian Analysis*, 12(4): 1105–1131.

Bhadra, A., Datta, J., Polson, N. G., and Willard, B. T. (2019). "The Horseshoe-Like Regularization for Feature Subset Selection." *Sankhya B*.

Bhattacharya, A., Chakraborty, A., and Mallick, B. K. (2016). "Fast sampling with Gaussian scale mixture priors in high-dimensional regression." *Biometrika*, 103(4): 985–991.

Bhattacharya, A. and Dunson, D. B. (2011). "Sparse Bayesian infinite factor models." *Biometrika*, 98(2): 291–306.

Bhattacharya, A., Pati, D., Pillai, N. S., and B., D. D. (2015). "Dirichlet–Laplace Priors for Optimal Shrinkage." *Journal of the American Statistical Association*, 110(512): 1479–1490.

Bind, M.-A. (2019). "Causal Modeling in Environmental Health." *Annual Review of Public Health*, 40: 23–43.

Bingham, N. H., Goldie, C. M., and Teugels, J. L. (1989). *Regular Variation, vol. 27 of Encyclopedia of Mathematics and its Applications*. Cambridge, UK: Cambridge University Press.

Blum, M. G. B., Valeri, L., François, O., Cadiou, S., Siroux, V., Lepeule, J., and Slama, R. (2020). "Challenges Raised by Mediation Analysis in a High-Dimension Setting." *Environmental Health Perspectives*, 128(5): 55001.

Bobb, J. F., Valeri, L., Claus Henn, B., Christiani, D. C., Wright, R. O., Mazumdar, M., Godleski, J. J., and Coull, B. A. (2015). "Bayesian kernel machine regression for estimating the health effects of multi-pollutant mixtures." *Biostatistics*, 16(3): 493–508.

Bollen, K. A. (1989). *Structural Equations with Latent Variables*. New York, NY: John Wiley & Sons, Inc.

Borkowski, K., Newman, J. W., Aghaeepour, N., Mayo, J. A., Blazenović, I., Fiehn, O., Stevenson, D. K., Shaw, G. M., and Carmichael, S. L. (2020). "Mid-gestation serum lipidomic profile associations with spontaneous preterm birth are influenced by body mass index." *PLoS One*, 15(11): e0239115.

Boss, J., Mukherjee, B., Ferguson, K. K., Aker, A., Alshawabkeh, A. N., Cordero, J. F., Meeker, J. D., and Kim, S. (2019). "Estimating outcome-exposure associations when exposure biomarker detection limits vary across batches." *Epidemiology*, 30(5): 746–755.

Boss, J., Rix, A., Chen, Y.-H., Narisetty, N. N., Wu, Z., Ferguson, K. K., McElrath, T. F., Meeker, J. D., and Mukherjee, B. (2021). "A hierarchical integrative group least absolute shrinkage and selection operator for analyzing environmental mixtures." *Environmetrics*.

Boss, J. B., Zhai, J. Z., Aung, M. T., Ferguson, K. K., Johns, L. E., McElrath, T. F., Meeker, J. D., and Mukherjee, B. (2018). "Associations between mixtures of urinary phthalate metabolites with gestational age at delivery: a time to event analysis using summative phthalate risk scores." *Environmental Health*, 17: 56.

Braun, J. M., Gennings, C., Hauser, R., and Webster, T. F. (2016). "What Can Epidemiological Studies Tell Us about the Impact of Chemical Mixtures on Human Health?" *Environmental Health Perspectives*, 124(1): A6–A9.

Brown, P. J. and Griffin, J. E. (2010). "Inference with normal-gamma prior distributions in regression problems." *Bayesian Analysis*, 5(1): 171–188.

Cadonna, A., Frühwirth-Schnatter, S., and Knaus, P. (2020). "Triple the Gamma—A Unifying Shrinkage Prior for Variance and Variable Selection in Sparse State Space and TVP Models." *Econometrics*, 8(2): 20.

Carrico, C., Gennings, C., Wheeler, D. C., and Factor-Litvak, P. (2015). "Characterization of Weighted Quantile Sum Regression for Highly Correlated Data in a Risk Analysis Setting." *Journal of Agricultural, Biological, and Environmental Statistics*, 20(100-120): 100–120.

Carvalho, C. M., Polson, N. G., and Scott, J. G. (2009). "Handling Sparsity via the Horseshoe." *Proceedings of the Twelth International Conference on Artificial Intelligence and Statistics, PMLR*, 5: 73–80.

— (2010). "The horseshoe estimator for sparse signals." *Biometrika*, 97(2): 465–480.

Casella, G. (2001). "Empirical Bayes Gibbs sampling." *Biostatistics*, 2(4): 485–500.

Castillo, I., Schmidt-Hieber, J., and van der Vaart, A. (2015). "Bayesian linear regression with sparse priors." *Annals of Statistics*, 43(5): 1986–2018.

Chatterjee, N., Chen, Y.-H., Maas, P., and Carroll, R. J. (2016). "Constrained Maximum Likelihood Estimation for Model Calibration Using Summary-Level Information From External Big Data Sources." *Journal of the American Statistical Association*, 111(513): 107–117.

Chen, H., Quandt, S. A., Grzywacz, J. G., and Arcury, T. A. (2011). "A Distribution-Based Multiple Imputation Method for Handling Bivariate Pesticide Data with Values below the Limit of Detection." *Environmental Health Perspectives*, 119(3): 351–356.

Chen, L.-W., Fine, J. P., Bair, E., Ritter, V. S., McElrath, T. F., Cantonwine, D. E., Meeker, J. D., Ferguson, K. K., and Zhao, S. (2022). "Semiparametric analysis of a generalized linear model with multiple covariates subject to detection limits." *Statistics in Medicine*, 41: 4791–4808.

Cheng, W., Taylor, J. M. G., Gu, T., Tomlins, S. A., and Mukherjee, B. (2019). "Informing a risk prediction model for binary outcomes with external coefficient information." *Journal of the Royal Statistical Society, Series C (Applied Statistics)*, 68(1): 121–139.

Cheng, W., Taylor, J. M. G., Vokonas, P. S., Park, S. K., and Mukherjee, B. (2018). "Improving estimation and prediction in linear regression incorporating external information from an established reduced model." *Statistics in Medicine*, 37: 1515–1530.

Chesher, A. (1991). "The effect of measurement error." *Biometrika*, 78(3): 451–462.

Chung, M. K., Buck Louis, G. M., Kannan, K., and Patel, C. J. (2019). "Exposome-wide association study of semen quality: Systematic discovery of endocrine disrupting chemical biomarkers in fertility require large sample sizes." *Environment International*, 125: 505–514.

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." *Biostatistics*, 19(2): 121–136.

Clark-Boucher, D., Zhou, X., Du, J., Liu, Y., Needham, B. L., Smith, J. A., and Mukherjee, B. (2023). "Methods for Mediation Analysis with High-Dimensional DNA Methylation Data: Possible Choices and Comparison." *medRxiv Preprint*, 1–27.

Crainiceanu, C. M., Caffo, B. S., Luo, S., Zipunnikov, V. M., and Punjabi, N. M. (2011). "Population Value Decomposition, a Framework for the Analysis of Image Populations." *Journal of the American Statistical Association*, 106(495): 775–790.

Czarnota, J., Gennings, C., and Wheeler, D. C. (2015). "Assessment of weighted quantile sum regression for modeling chemical mixtures and cancer risk." *Cancer Informatics*, 14(Suppl 2): 159–71.

Dai, J. Y., Stanford, J. L., and LeBlanc, M. (2022). "A Multiple-Testing Procedure for High-Dimensional Mediation Hypotheses." *Journal of the American Statistical Association*, 117(537): 198–213.

Datta, J. and Dunson, D. B. (2016). "Bayesian inference on quasi-sparse count data." *Biometrika*, 103(4): 971–983.

Datta, J. and Ghosh, J. K. (2013). "Asymptotic Properties of Bayes Risk for the Horseshoe Prior." *Bayesian Analysis*, 8(1): 111–132.

Dawid, A. P. (1973). "Posterior Expectations for Large Observations." *Biometrika*, 60(3): 664–667.

Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977). "Maximum Likelihood from Incomplete Data via the EM Algorithm." *Journal of the Royal Statistical Society, Series B (Methodological)*, 39(1): 1–38.

Devick, K. L., Bobb, J. F., Mazumdar, M., Claus Henn, B., Bellinger, D. C., Christiani, D. C., Wright, R. O., Williams, P. L., Coull, B. A., and Valeri, L. (2022). "Bayesian kernel machine regression-causal mediation analysis." *Statistics in Medicine*, 41(5): 860–876.

Ding, Y., Kong, S., Kang, S., and Chen, W. (2018). "A semiparametric imputation approach for regression with censored covariate with application to an AMD progression study." *Statistics in Medicine*, 37(23): 3293–3308.

Dominici, F., Peng, R. D., Barr, C. D., and Bell, M. L. (2010). "Protecting Human Health from Air Pollution: Shifting from a Single-Pollutant to a Multi-pollutant Approach." *Epidemiology*, 21(2): 187–194.

Du, J., Zhou, X., Clark-Boucher, D., Hao, W., Liu, Y., Smith, J. A., and Mukherjee, B. (2022). "Methods for large-scale single mediator hypothesis testing: Possible choices and comparisons." *Genetic Epidemiology*, 47(2): 167–184.

Efron, B. (1982). *The Jackknife, the Bootstrap and Other Resampling Plans*. Society for Industrial and Applied Mathematics.

Estes, J. P., Taylor, J. M. G., and Mukherjee, B. (2018). "Empirical Bayes Estimation and Prediction Using Summary-Level Information From External Big Data Sources Adjusting for Violations of Transportability." *Statistics in Biosciences*, 10: 568–586.

Ferguson, K. K., McElrath, T. F., and Meeker, J. D. (2014). "Environmental Phthalate Exposure and Preterm Birth." *JAMA Pediatrics*, 168(1): 61–67.

Ferguson, T. S. (1973). "A Bayesian Analysis of Some Nonparametric Problems." *Annals of Statistics*, 1(2): 209–230.

Ferrari, F. and Dunson, D. B. (2020). "Identifying main effects and interactions among exposures using Gaussian processes." *Annals of Applied Statistics*, 14(4): 1743–1758.

— (2021). "Bayesian Factor Analysis for Inference on Interactions." *Journal of the American Statistical Association*, 116(535): 1521–1532.

Frot, B., Nandy, P., and Maathuis, M. H. (2019). "Robust causal structure learning with some hidden variables." *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 81(3): 459–487.

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

Gelman, A. and Rubin, D. B. (1992). "Inference from Iterative Simulation Using Multiple Sequences." *Statistical Science*, 7(4): 457–472.

Geweke, J. and Zhou, G. (1996). "Measuring the Pricing Error of the Arbitrage Pricing Theory." *The Review of Financial Studies*, 9(2): 557–587.

Ghosal, S. (1999). "Asymptotic Normality of Posterior Distributions in High-Dimensional Linear Models." *Bernoulli*, 5(2): 15–331.

Gibson, E. A., Nunez, Y., Abuawad, A., Zota, A. R., Renzetti, S., Devick, K. L., Gennings, C., Goldsmith, J., Coull, B. A., and Kioumourtzoglou, M.-A. (2019). "An overview of methods to address distinct research questions on environmental mixtures: an application to persistent organic pollutants and leukocyte telomere length." *Environmental Health*, 18: 76.

Gibson, E. A., Zhang, J., Yan, J., Chillrud, L., Benavides, J., Nunez, Y., Herbstman, J. B., Goldsmith, J., Wright, J., and Kioumourtzoglou, M.-A. (2022). "Principal Component Pursuit for Pattern Identification in Environmental Mixtures." *Environmental Health Perspectives*, 130(11): 117008.

Glymour, C. and Spirtes, P. (1988). "Latent variables, causal models and overidentifying constraints." *Journal of Econometrics*, 39(1-2): 175–198.

Goutman, S. A., Boss, J., Godwin, C., Mukherjee, B., Feldman, E. L., and Batterman, S. A. (2023). "Occupational history associates with ALS survival and onset segment." *Amyotrophic Lateral Sclerosis and Frontotemporal Degeneration*, 3-4: 219–229.

Gu, T., Taylor, J. M. G., Cheng, W., and Mukherjee, B. (2019). "Synthetic data method to incorporate external information into a current study." *The Canadian Journal of Statistics*, 47(4): 580–603.

Gu, T., Taylor, J. M. G., and Mukherjee, B. (2021). "A meta-inference framework to integrate multiple external models into a current study." *Biostatistics*, kxab017.

Hahn, P. R. and Carvalho, C. M. (2015). "Decoupling Shrinkage and Selection in Bayesian Linear Models: A Posterior Summary Perspective." *Journal of the American Statistical Association*, 110(509): 435–448.

Hall, M. A., Dudek, S. M., Goodloe, R., Crawford, D. C., Pendergrass, S. A., Peissig, P., Brilliant, M., McCarty, C. A., and Ritchie, M. D. (2014). "Environment-wide association study (EWAS) for type 2 diabetes in the Marshfield Personalized Medicine Research Project Biobank." *Pacific Symposium on Biocomputing*, 200–211.

Han, P. and Lawless, J. F. (2019). "Empirical Likelihood Estimation Using Auxiliary Summary Information with Different Covariate Distributions." *Statistica Sinica*, 29(3): 1321–1342.

Hefley, T. J., Hooten, M. B., Hanks, E. M., Russell, R. E., and Walsh, D. P. (2017). "The Bayesian Group Lasso for Confounded Spatial Data." *Journal of Agricultural, Biological, and Environmental Statistics*, 22(1): 42–59.

Herring, A. H. (2010). "Nonparametric Bayes Shrinkage for Assessing Exposures to Mixtures Subject to Limits of Detection." *Epidemiology*, 21(4): S71–S76.

Hörmann, W. and Leydold, J. (2014). "Generating generalized inverse Gaussian random variates." *Statistics and Computing*, 24: 547–557.

Huang, Y.-T. and Pan, W.-C. (2016). "Hypothesis Test of Mediation Effect in Causal Mediation Model with High-Dimensional Continuous Mediators." *Biometrics*, 72(2): 402–413.

Johndrow, J. E., Orenstein, P., and Bhattacharya, A. (2020). "Scalable Approximate MCMC Algorithms for the Horseshoe Prior." *Journal of Machine Learning Research*, 21: 1–61.

Joubert, B. R., Kioumourtzoglou, M.-A., Chamberlain, T., Chen, H. Y., Gennings, C., Turyk, M. E., Miranda, M. L., Webster, T. F., Ensor, K. B., Dunson, D. B., and Coull, B. A. (2022). "Powering Research through Innovative Methods for Mixtures in Epidemiology (PRIME) Program: Novel and Expanded Statistical Methods." *International Journal of Environmental Research and Public Health*, 19(3): 1378.

Kaiser, H. F. (1958). "The varimax criterion for analytic rotation in factor analysis." *Psychometrika*, 23: 187–200.

Kang, J., Reich, B. J., and Staicu, A.-M. (2018). "Scalar-on-image regression via the soft-thresholded Gaussian process." *Biometrika*, 105(1): 165–184.

Kang, K., Song, X., Hu, X. J., and Zhu, H. (2019). "Bayesian adaptive group lasso with semiparametric hidden Markov models." *Statistics in Medicine*, 38(9): 1634–1650.

Kaplan, K., Chen, J., Yavuz, S., and Lyu, W. (2023). "Bayesian Dynamic Borrowing of Historical Information with Applications to the Analysis of Large-Scale Assessments." *Psychometrika*, 88(1): 1–30.

Keil, A. P., Buckley, J. P., O'Brien, K. M., Ferguson, K. K., Zhao, S., and White, A. J. (2020). "A Quantile-Based g-Computation Approach to Addressing the Effects of Exposure Mixtures." *Environmental Health Perspectives*, 128(4): 47004.

Kong, S. and Nan, B. (2016). "Semiparametric approach to regression with a covariate subject to a detection limit." *Biometrika*, 103(1): 161–174.

Kowal, D. R., Matteson, D. S., and Ruppert, D. (2019). "Dynamic shrinkage processes." *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 81(4): 781–804.

Kuiper, J. R., O'Brien, K. M., Ferguson, K. K., and Buckley, J. P. (2021). "Urinary specific gravity measures in the U.S. population: Implications for the adjustment of non-persistent chemical urinary biomarker data." *Environment International*, 156: 106656.

Kundu, D., Mitra, R., and Gaskins, J. T. (2021). "Bayesian variable selection for multioutcome models through shared shrinkage." *Scandinavian Journal of Statistics*, 48: 295–320.

Kuroki, M. and Pearl, J. (2014). "Measurement bias and effect restoration in causal inference." *Biometrika*, 101(2): 423–437.

Kyung, M., Gill, J., Ghosh, M., and Casella, G. (2010). "Penalized Regression, Standard Errors, and Bayesian Lassos." *Bayesian Analysis*, 5(2): 369–412.

Li, J., Wang, Z., Li, R., and Wu, R. (2015). "Bayesian Group Lasso for Nonparametric Varying-Coefficient Models with Application to Functional Genome-Wide Association Studies." *The Annals of Applied Statistics*, 9(2): 640–664.

Lind, P. M., Risérus, U., Salihovic, S., van Bavel, B., and Lind, L. (2013). "An environmental wide association study (EWAS) approach to the metabolic syndrome." *Environment International*, 55: 1–8.

Liu, B. and Ghosh, S. K. (2020). "On empirical estimation of mode based on weakly dependent samples." *Computational Statistics & Data Analysis*, 152: 107046.

Liu, Z., Shen, J., Barfield, R., Schwartz, J., Baccarelli, A. A., and Lin, X. (2022). "Large-Scale Hypothesis Testing for Causal Mediation Effects with Applications in Genome-wide Epigenetic Studies." *Journal of the American Statistical Association*, 117(537): 67–81.

Louizos, C., Shalit, U., Mooij, J. M., Sontag, D., Zemel, R., and Welling, M. (2017). "Causal Effect Inference with Deep Latent-Variable Models." *Advances in Neural Information Processing Systems 30 (NIPS 2017)*, 6449–6459.

Lubin, J. H., Colt, J. S., Camann, D., Davis, S., Cerhan, J. R., Severson, R. K., Bernstein, L., and Hartge, P. (2004). "Epidemiologic Evaluation of Measurement Data in the Presence of Detection Limits." *Environmental Health Perspectives*, 112(17): 1691–1696.

Lynn, H. S. (2001). "Maximum likelihood inference for left-censored HIV RNA data." *Statistics in Medicine*, 20(1): 33–45.

Makalic, E. and Schmidt, D. F. (2016). "A Simple Sampler for the Horseshoe Estimator." *IEEE Signal Processing Letters*, 23(1): 179–182.

Mallick, H. and Yi, N. (2017). "Bayesian group bridge for bi-level variable selection." *Computational Statistics & Data Analysis*, 110: 115–133.

Mathur, M. B. and VanderWeele, T. J. (2022). "Methods to Address Confounding and Other Biases in Meta-Analyses: Review and Recommendations." *Annual Review of Public Health*, 43: 19–35.

May, R. C., Ibrahim, J. G., and Chu, H. (2011). "Maximum likelihood estimation in generalized linear models with multiple covariates subject to detection limits." *Statistics in Medicine*, 30(20): 2551–2561.

McGee, G., Wilson, A., Webster, T. F., and Coull, B. A. (2023). "Bayesian multiple index models for environmental mixtures." *Biometrics*, 79(1): 462–474.

McGinnis, D. P., Brownstein, J. S., and Patel, C. J. (2016). "Environment-Wide Association Study of Blood Pressure in the National Health and Nutrition Examination Survey (1999–2012)." *Scientific Reports*, 6: 30373.

Miao, W., Geng, Z., and Tchetgen Tchetgen, E. J. (2018). "Identifying causal effects with proxy variables of an unmeasured confounder." *Biometrika*, 105(4): 987–993.

Molitor, J., Papathomas, M., Jerrett, M., and Richardson, S. (2010). "Bayesian profile regression with an application to the National survey of children's health." *Biostatistics*, 11(3): 484–498.

Mork, D. and Wilson, A. (2023). "Estimating perinatal critical windows of susceptibility to environmental mixtures via structured Bayesian regression tree pairs." *Biometrics*, 79(1): 449–461.

Morris, C. N. (1983). "Parametric Empirical Bayes Inference: Theory and Applications." *Journal of the American Statistical Association*, 78(381): 47–55.

Mukherjee, B. and Chatterjee, N. (2008). "Exploiting Gene-Environment Independence for Analysis of Case–Control Studies: An Empirical Bayes-Type Shrinkage Estimator to Trade-Off between Bias and Efficiency." *Biometrics*, 64(3): 685–694.

Nie, L., Chu, H., Liu, C., Cole, S. R., Vexler, A., and Schisterman, E. F. (2010). "Linear Regression with an Independent Variable Subject to a Detection Limit." *Epidemiology*, 21(Suppl 4): S17–S24.

Nishimura, A. and Suchard, M. A. (2022). "Prior-Preconditioned Conjugate Gradient Method for Accelerated Gibbs Sampling in "Large n, Large p" Bayesian Sparse Regression." *Journal of the American Statistical Association*.

O'Hagan, A. (1979). "On Outlier Rejection Phenomena in Bayes Inference." *Journal of the Royal Statistical Society: Series B (Methodological)*, 41(3): 358–367.

Park, S. K., Yao, Y., Meeker, J. D., Harlow, S. D., and Mukherjee, B. (2014). "Environmental risk score as a new tool to examine multi-pollutants in epidemiologic research: an example from the NHANES study using serum lipid levels." *PLoS One*, 9(6): e98632.

Park, S. K., Zhao, Z., and Mukherjee, B. (2017). "Construction of environmental risk score beyond standard linear models using machine learning methods: application to metal mixtures, oxidative stress and cardiovascular disease in NHANES." *Environmental Health*, 16(1): 102.

Patel, C. J., Bhattacharya, J., and Butte, A. J. (2010). "An Environment-Wide Association Study (EWAS) on Type 2 Diabetes Mellitus." *PLoS One*, 5(5): e10746.

Pearl, J. (2001). "Direct and Indirect Effects." *Proceedings of the Seventeenth conference on Uncertainty in artificial intelligence*, 411–420.

Polson, N. G. and Scott, J. G. (2011). "Shrink Globally, Act Locally: Sparse Bayesian Regularization and Prediction." In Bernardo, J. M., Bayarri, M. J., Berger, J. O., Dawid, A. P., Heckerman, D., Smith, A. F. M., and West, M. (eds.), *Bayesian Statistics 9*, chapter 17. Oxford, United Kingdom: Oxford University Press.

Poworoznek, E., Ferrari, F., and Dunson, D. B. (2021). "Efficiently resolving rotational ambiguity in Bayesian matrix sampling with matching." *arXiv Preprint*, 1–27.

Robins, J. M. and Greenland, S. (1992). "Identifiability and Exchangeability for Direct and Indirect Effects." *Epidemiology*, 3(2): 143–155.

Rockova, V. and George, E. I. (2018). "The Spike-and-Slab LASSO." *Journal of the American Statistical Association*, 113(521): 431–444.

Rockova, V. and Lesaffre, E. (2014). "Incorporating grouping information in Bayesian variable selection with applications in genomics." *Bayesian Analysis*, 9(1): 221–258.

Schwarz, G. (1978). "Estimating the Dimension of a Model." *Annals of Statistics*, 6(2): 461–464.

Schölkopf, B., Locatello, F., Bauer, S., Ke, N. R., Kalchbrenner, N., Goyal, A., and Bengio, Y. (2021). "Toward Causal Representation Learning." *Proceedings of the IEEE*, 109(5): 612–634.

Shortreed, S. M. and Ertefaie, A. (2017). "Outcome-Adaptive Lasso: Variable Selection for Causal Inference." *Biometrics*, 73: 1111–1122.

Sillé, F. C. M., Karakitsios, S., Kleensang, A., Koehler, K., Maertens, A., Miller, G. W., Prasse, C., Quiros-Alcala, L., Ramachandran, G., Rappaport, S. M., Rule, A. M., Sarigiannis, D., Smirnova, L., and Hartung, T. (2023). "The exposome - a new approach for risk assessment." *ALTEX*, 37(1): 3–23.

Silva, R., Scheines, R., Glymour, C., and Spirtes, P. (2006). "Learning the Structure of Linear Latent Variable Models." *Journal of Machine Learning Research*, 7: 191–246.

Simon, N., Friedman, J., Hastie, T., and Tibshirani, R. (2013). "A Sparse-Group Lasso." *Journal of Computational and Graphical Statistics*, 22(2): 231–245.

Som, A., Hans, C., and MacEachern, S. N. (2015). "Block Hyper-g Priors in Bayesian Regression." *arXiv*.

Som, A., Hans, C. M., and MacEachern, S. M. (2016). "A conditional Lindley paradox in Bayesian linear models." *Biometrika*, 103(4): 993–999.

Song, Q. and Liang, F. (2017). "Nearly optimal Bayesian Shrinkage for High Dimensional Regression." *arXiv Preprint*.

Song, Y., Zhou, X., Kang, J., Aung, M. T., Zhang, M., Zhao, W., Needham, B. L., Kardia, S. L. R., Liu, Y., Meeker, J. D., Smith, J. A., and Mukherjee, B. (2021a). "Bayesian hierarchical models for high-dimensional mediation analysis with coordinated selection of correlated mediators." *Statistics in Medicine*.

— (2021b). "Bayesian sparse mediation analysis with targeted penalization of natural indirect effects." *Journal of the Royal Statistical Society: Series C (Applied Statistics)*.

Song, Y., Zhou, X., Zhang, M., Zhao, W., Liu, Y., Kardia, S. L. R., Diez Roux, A. V., Needham, B. L., Smith, J. A., and Mukherjee, B. (2020). "Bayesian shrinkage estimation of high dimensional causal mediation effects in omics studies." *Biometrics*, 76(3): 700–710.

Sun, Z., Tao, Y., Li, S., Ferguson, K. K., Meeker, J. D., Park, S. K., Batterman, S. A., and Mukherjee, B. (2013). "Statistical strategies for constructing health risk models with multiple pollutants and their interactions: possible choices and comparisons." *Environmental Health*, 12: 85.

Tang, X., Ghosh, M., Ha, N. S., and Sedransk, J. (2018). "Modeling Random Effects Using Global–Local Shrinkage Priors in Small Area Estimation." *Journal of the American Statistical Association*, 113(524): 1476–1489.

Terenin, A., Dong, S., and Draper, D. (2019). "GPU-accelerated Gibbs sampling: a case study of the Horseshoe Probit model." *Statistics and Computing*, 29(2): 301–310.

van der Pas, S., Szabó, B., and van der Vaart, A. (2017). "Uncertainty Quantification for the Horseshoe (with Discussion)." *Bayesian Analysis*, 12(4): 1221–1274.

VanderWeele, T. (2015). *Explanation in Causal Inference : Methods for Mediation and Interaction*. Oxford University Press, Incorporated.

VanderWeele, T. J. and Vansteelandt, S. (2014). "Mediation Analysis with Multiple Mediators." *Epidemiologic Methods*, 2(1): 95–115.

— (2022). "A statistical test to reject the structural interpretation of a latent factor model." *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 84(5): 2032–2054.

Verbeke, G. and Molenberghs, G. (2000). *Linear Mixed Models for Longitudinal Data*. Springer-Verlag New York, Incorporated.

Viele, K., Berry, S., Neuenschwander, B., Amzal, B., Chen, F., Enas, N., Hobbs, B., Ibrahim, J. G., Kinnersley, N., Lindborg, S., Micallef, S., Roychoudhury, S., and Thompson, L. (2014). "Use of historical control data for assessing treatment effects in clinical trials." *Pharmaceutical Statistics*, 13(1): 41–54.

Wang, C., Parmigiani, G., and Dominici, F. (2012). "Bayesian Effect Estimation Accounting for Adjustment Uncertainty." *Biometrics*, 68(3): 661–671.

Wang, H. J. and Feng, X. (2012). "Multiple Imputation for M-Regression With Censored Covariates." *Journal of the American Statistical Association*, 107(497): 194–204.

Wang, Y. and Blei, D. M. (2019). "The Blessings of Multiple Causes." *Journal of the American Statistical Association*, 114(528): 1574–1596.

Wang, Y.-B., Chen, Z., Goldstein, J. M., Buck Louis, G. M., and Gilman, S. E. (2019). "A Bayesian regularized mediation analysis with multiple exposures." *Statistics in Medicine*, 38(5): 828–843.

Wei, R., Reich, B. J., Hoppin, J. A., and Ghosal, S. (2020). "Sparse Bayesian Additive Nonparametric Regression with Application to Health Effects of Pesticides Mixtures." *Statistica Sinica*, 30: 55–79.

Weisskopf, M. G., Seals, R. M., and Webster, T. F. (2018). "Bias Amplification in Epidemiologic Analysis of Exposure to Mixtures." *Environmental Health Perspectives*, 126(4): 047003.

Welch, B. M., Keil, A. P., Buckley, J. P., Calafat, A. M., Christenbury, K. E., Engel, S. M., O'Brien, K. M., Rosen, E. M., James-Todd, T., Zota, A. R., Ferguson, K. K., the Pooled Phthalate Exposure, and Group, P. B. S. (2022). "Associations Between Prenatal Urinary Biomarkers of Phthalate Exposure and Preterm Birth A Pooled Study of 16 US Cohorts." *JAMA Pediatrics*, 176(9): 895–905.

Wild, C. P. (2005). "Complementing the genome with an "exposome": the outstanding challenge of environmental exposure measurement in molecular epidemiology." *Cancer Epidemiology, Biomarkers, & Prevention*, 14(8): 1847–1850.

Wilson, A., Leon Hsu, H.-H., Mathilda Chiu, Y.-H., Wright, R. O., Wright, R. J., and Coull, B. A. (2022). "Kernel machine and distributed lag models for assessing windows of susceptibility to environmental mixtures in children's health studies." *Annals of Applied Statistics*, 16(2): 1090–1110.

Witten, D. M., Shojaie, A., and Zhang, F. (2014). "The Cluster Elastic Net for High-Dimensional Regression With Unknown Variable Grouping." *1*, 56(Technometrics): 112–122.

Xu, X. and Ghosh, M. (2015). "Bayesian Variable Selection and Estimation for Group Lasso." *Bayesian Analysis*, 10(4): 909–936.

Xu, Z., Schmidt, D. F., Makalic, E., Qian, G., and Hopper, J. L. (2016). "Bayesian Grouped Horseshoe Regression with Application to Additive Models." In on Artificial Intelligence 2016, A. J. C. (ed.), *AI 2016: Advances in Artificial Intelligence*, chapter 3. Hobart, Australia: Springer.

Yuan, M. and Lin, Y. (2006). "Model selection and estimation in regression with grouped variables." *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 68(1): 49–67.

Zanobetti, A., Austin, E., Coull, B. A., Schwartz, J., and Koutrakis, P. (2014). "Health effects of multi-pollutant profiles." *Environment International*, 71: 13–19.

Zeng, P., Shao, Z., and Zhou, X. (2021). "Statistical methods for mediation analysis in the era of high-throughput genomics: Current successes and future challenges." *Computational and Structural Biotechnology Journal*, 19: 3209–3224.

Zhai, Y. and Han, P. (2022). "Data Integration with Oracle Use of External Information from Heterogeneous Populations." *Journal of Computational and Graphical Statistics*, 31(4): 1001–1012.

Zhang, H., Zheng, Y., Zhang, Z., Gao, T., Joyce, B., Yoon, G., Zhang, W., Schwartz, J., Just, A., Colicino, E., Vokonas, P., Zhao, L., Lv, J., Baccarelli, A., Hou, L., and Liu, L. (2016). "Estimating and testing high-dimensional mediation effects in epigenetic studies." *Bioinformatics*, 32(20): 3150–3154.

Zhang, Q. (2022). "High-Dimensional Mediation Analysis with Applications to Causal Gene Identification." *Statistics in Biosciences*, 14: 432–451.

Zhang, Y. and Bondell, H. D. (2018). "Variable Selection via Penalized Credible Regions with Dirichlet–Laplace Global-Local Shrinkage Priors." *Bayesian Analysis*, 13(3): 823–844.

Zhao, Y., Li, L., and Alzheimer's Disease Neuroimaging Initiative (2022). "Multimodal data integration via mediation analysis with high-dimensional exposures and mediators." *Human Brain Mapping*, 43(8): 2519–2533.

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

Zhao, Y. and Luo, X. (2022). "Pathway Lasso: Pathway Estimation and Selection with High-Dimensional Mediators." *Statistics and Its Interface*, 15(1): 39–50.

Zhou, R. R., Wang, L., and Zhao, D. (2020). "Estimation and inference for the indirect effect in high-dimensional linear mediation models." *Biometrika*, 107(3): 573–589.

Zhuang, X., Ni, A., Liao, L., Guo, Y., Dai, W., Jiang, Y., Zhou, H., Hu, X., Du, Z., Wang, X., and Liao, X. (2018). "Environment-wide association study to identify novel factors associated with peripheral arterial disease: Evidence from the National Health and Nutrition Examination Survey (1999-2004)." *Atherosclerosis*, 269: 172–177.

Zou, H. and Zhang, H. H. (2009). "On the adaptive elastic-net with a diverging number of parameters." *Annals of Statistics*, 37(4): 1733–1751.