Methods for Confounding Adjustment and High-Dimensional Environmental Exposures

Joshua P. Keller

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

Doctor of Philosophy

University of Washington

2016

Reading Committee:

Kenneth Rice, Chair

Adam Szpiro, Chair

Mathias Drton

Noah Simon

Program Authorized to Offer Degree:
Department of Biostatistics

University of Washington

Abstract

Methods for Confounding Adjustment and High-Dimensional Environmental Exposures

Joshua P. Keller

Co-Chairs of the Supervisory Committee:
Associate Professor Kenneth Rice
Department of Biostatistics

Associate Professor Adam Szpiro
Department of Biostatistics

Environmental exposures have complex multivariate relationships with one another and with geographic, anthropogenic, social, and physiological factors. This dissertation comprises methods for addressing the confounding and high-dimensional challenges of environmental exposures in cohort studies. We consider three different settings for improving statistical inference about associations between exposures and health effects using these multivariate relationships. First we present a method for clustering multi-pollutant observations in the context of an air pollution epidemiology cohort, where exposure must be predicted at subject locations. We then present a method for shrinkage estimation, with particular focus on small sample benefit in the presence of many confounders. Third, we present methods for adjusting for unmeasured spatial confounding in analyses with environmental exposures. We apply each method to analyses of cardiovascular outcomes in a cohort study.

TABLE OF CONTENTS

Page

LIST OF FIGURES

Figure Number Page

LIST OF TABLES

Table Number Page

ACKNOWLEDGMENTS

This dissertation would not have been possible without the support of many people. I first would like to thank my co-advisors Ken Rice and Adam Szpiro. Their insights, patience, and mentorship has been instrumental to my development as a biostatistician and played a key role in the research contained here. I am grateful to committee members Mathias Drton, Noah Simon, and Edmund Seto for their willingness to share feedback and ideas that improved this research.

I have benefited greatly from ongoing collaborations with the faculty and staff of the Environmental Biostatistics and Air Pollution Epidemiology working groups. Lianne Sheppard, Paul Sampson, Joel Kaufman, Sverre Vedal, Tim Larson, Casey Olives, Sun-Young Kim, and Amanda Gassett have all helped to make my research productive and enjoyable. Thank you in particular to Lianne, for supporting my training through the Biostatistics, Epidemiologic, and Bioinformatic Training in Environmental Health Training Grant (NIH/NIEHS award number T32ES015459) and taking an active role in my professional development.

Lastly, I owe many thanks to the Biostatistics Entering Class of 2011, whose friendship and camaraderie got me through the years of coursework and qualifying exams.

DEDICATION

to my parents, for their unwavering support and enthusiasm over the years
and to Whitney, who makes life a wonderful adventure worth exploring

Chapter 1

INTRODUCTION

Environmental exposures such as air pollution vary across space due a variety of geographic, meteorological, and anthropogenic factors. To accurately estimate associations between health outcomes and environmental exposures in cohort studies, we must account for the complex relationships between exposures and other factors. This dissertation comprises three projects that address inference from environmental exposures. In each project, we trade off competing goals to improve our overall ability to estimate health associations.

In Chapter 2, we consider the problem of estimating the association between multi-pollutant exposures and health outcomes in air pollution cohort studies. Multi-dimensional pollutant measurements are typically made at regulatory monitoring locations, but epidemiological analyses require low-dimensional predictions at unmonitored cohort locations. We propose a method, which we call *predictive k-means*, which adapts the standard k-means clustering procedure to use prediction covariates when selecting cluster centers. By incorporating prediction covariates into the selection of cluster centers, we trade off between the representativeness of cluster centers and prediction error at cohort locations. This leads to higher power for detecting effect modification by cluster. We apply this method to an analysis of particulate matter components and blood pressure in the NIEHS Sister Study.

Chapter 3 considers the problem of estimating a parameter of interest in a linear regression model in settings with limited sample sizes and many confounders. From a causal inference perspective, optimal inference is well understood in observational studies with unlimited data, by adjusting for appropriate confounders, but in the small samples often found in epidemiology the optimal approach is less clear. Shrinkage estimators such as the LASSO and Ridge regression can reduce mean-squared error by trading some amounts of bias for

reductions in variance. However, when inference is the goal, there are no standard methods for choosing the penalty parameter that governs this tradeoff. We propose selecting the penalty parameters for these estimators by minimizing bias and variance in future similar datasets drawn from a posterior predictive distribution. We demonstrate this method using subclinical measures of cardiovascular health and smoking status.

The final chapter considers the challenge of adjusting for residual spatial confounding in regression models with estimated spatial exposures. This is motivated by air pollution epidemiology, where pollutant concentrations are frequently correlated with socioeconomic factors that also impact health. We present approaches for addressing this confounding through a pre-adjustment procedure with different spatial basis functions. We explore the role of spatial confounding and its adjustment in an analysis of fine particular matter in the NIEHS Sister Study.

Chapter 2

DIMENSION REDUCTION FOR SPATIALLY-MISALIGNED COHORT DATA

2.1 Introduction

Cohort studies provide a valuable platform for investigating health effects of long-term air pollution exposure by leveraging fine-scale spatial contrasts in exposure between subjects (Künzli et al. 2001; Dominici et al. 2003; Wilson et al. 2005). These studies facilitate a level of precision in exposure assignment that is not available in traditional analyses based upon aggregated data from administrative districts. However, cohort-specific exposure monitoring is rarely done at more than a small subset of subject locations for a short period of time, if at all (Cohen et al. 2009). Instead, pollutant concentrations measured at locations in regulatory monitoring networks, not at cohort locations, are used. This *spatial misalignment* between monitor and subject locations is often addressed through a two-stage modeling approach. First, an exposure prediction model is developed using the regulatory monitoring data, and predictions are made at cohort subject locations (e.g., Brauer et al. 2003; Keller et al. 2015). These predicted exposures are then used in regression analyses, where their association with health outcomes is estimated (e.g., Adar et al. 2010).

Fine particulate matter (particles with aerodynamic diameter less than 2.5 μm; $\mathrm{PM}_{2.5}$) is a mixture of many components whose chemical composition varies widely due to sources, meteorology, and other factors (Bell et al. 2007). Variations in $\mathrm{PM}_{2.5}$ composition can modify the association between total $\mathrm{PM}_{2.5}$ mass and health effects (Brook et al. 2010; Franklin et al. 2008; Zanobetti et al. 2009), and analysis that distinguishes between different component profiles can improve our understanding of exposures' health effects (Brauer 2010).

Multi-pollutant exposures such as $\mathrm{PM}_{2.5}$ component concentrations present challenges to the two-stage modeling approach for addressing spatial misalignment. Multi-dimensional prediction requires ignoring correlation between pollutants or making strong assumptions about correlation structure that may be difficult to verify with limited monitoring data. Interpreting coefficient estimates for simultaneous exposures to multiple pollutants presents challenges of generalizability. Reducing the dimension of a multi-pollutant exposure prior to prediction provides an attractive means to address these challenges in prediction and interpretation. Dimension reduction methods simplify the complex structure of multi-pollutant exposures by reducing them to a smaller set of low-dimensional observations that retain most of the characteristics of the original data but that can be predicted more reliably.

Clustering methods are a class of dimension reduction methods that partition multi-pollutant observations into a pre-specified number of clusters. For multi-dimensional observations of $PM_{2.5}$ components, this amounts to assigning each observation to a representative component profile. Oakes et al. (2014) highlight clustering as a promising approach for understanding multi-pollutant health effects. The 'k-means' algorithm is a popular clustering method that identifies clusters that minimize the distance between each observation and the center of its assigned cluster. Recent work has applied clustering methods (including k-means) to time series of $PM_{2.5}$ observations to find groups of days with similar component profiles in daily averages (Austin et al. 2012) and groups of locations with similar profiles in long-term averages (Austin et al. 2013). These clusters were used for analyzing exposures by city (Kioumourtzoglou et al. 2015), but have not been used for cohort subject locations. For cohort studies with spatially misaligned monitoring data, the lack of monitoring observations means we cannot directly cluster cohort locations using component profiles. One option is then to use k-means to cluster monitoring data and to subsequently predict cluster membership at subject locations. However, this can work poorly when membership in the clusters identified by k-means is not predictable using available geographic covariates. Modifying the k-means procedure to account for the covariates used in the subsequent prediction model provides a promising approach for efficient prediction of cluster membership at subject locations.

In this chapter, we present a method for clustering multi-pollutant exposures in the context of cohort studies with spatially misaligned data and apply it to an analysis of $\text{PM}_{2.5}$ component exposure in a national cohort. Section 2.2 presents the motivating analysis of total $\text{PM}_{2.5}$ and systolic blood pressure in the Sister Study cohort. Section 2.3 describes an approach for clustering multi-pollutant data in a cohort study using a combination of existing methods. In Section 2.4, we introduce our new method for defining clusters that improves predictive accuracy at cohort locations. Section 2.5 details simulations illustrating this method, and in Section 2.6 we apply the method to the Sister Study cohort. We conclude in Section 2.7 with a discussion.

2.2 $\text{PM}_{2.5}$ and SBP in the Sister Study

The National Institute of Environmental Health Sciences (NIEHS) Sister Study cohort comprises 50,884 women with a sister with breast cancer from across the United States enrolled between 2003 and 2009. In a cross-sectional analysis of the Sister Study cohort, Chan et al. (2015) found that a difference of $10\mu g/m^3$ in annual average $\text{PM}_{2.5}$ was associated with 1.4 mmHg higher systolic blood pressure (SBP) [95% CI: 0.6, 2.3; $p < 0.001$]. Chan et al. (2015) used predictions of 2006 annual average ambient $\text{PM}_{2.5}$ exposures from a universal kriging (UK) model fit to monitoring data from the EPA Air Quality System (AQS) (Sampson et al. 2013). The UK model has two components: a regression on geographic covariates for the mean combined with spatial smoothing via a Gaussian Process. The geographic covariates included measures of land-cover, road network characteristics, vegetative index, population density, and distance to various geographic features, which Sampson et al. (2013) reduced in dimension using partial least squares. An exponential covariance structure was used for smoothing in the Gaussian Process.

During baseline home visits for the Sister Study, blood pressure measurements were taken, along with anthropometric measurements and phlebotomy. Residential history of subjects is available for assigning long-term exposures based upon participant locations. In their health model, Chan et al. (2015) performed linear regression of SBP on $\text{PM}_{2.5}$, adjusting for age, race, socioeconomic status (household income, education, marital status, working more than 20 hours per week outside the home, perceived stress score, and socioeconomic status Z-score), rural-urban continuum code, geographic location (via spatial regression splines), cardiovascular risk factors (BMI, waist-hip-ratio, smoking status, alcohol use, history of diabetes and hypercholesterolemia), and blood pressure medication use.

In order to better understand how the observed $\mathrm{PM}_{2.5}$ effect varies by $\mathrm{PM}_{2.5}$ composition, we will re-analyze the Sister Study cohort in Section 2.6 to investigate whether the association between $\mathrm{PM}_{2.5}$ and SBP is modified by clustering subjects using component profiles of $\mathrm{PM}_{2.5}$.

2.3 Clustering Spatially Misaligned Data

In this section we consider clustering $\mathrm{PM}_{2.5}$ component observations into K component profiles, in the presence of spatial misalignment between the monitor and subject locations, by combining existing methods for unsupervised clustering and spatial prediction.

Ideally we would like to observe the matrix $\boldsymbol{X} \in \mathbb{R}^{n \times p}$ of annual average mass fractions at n cohort locations for p components of $\mathrm{PM}_{2.5}$, which we refer to as component species. However, we can only observe the matrix $\boldsymbol{X}^* \in \mathbb{R}^{n^* \times p}$ of annual average mass fractions at n^* AQS monitoring locations. (Throughout this chapter we use an asterisk to denote values at monitor locations, while values without an asterisk correspond to cohort locations). Geographic covariates such as distance to primary roadways and land use categorizations are available at both monitoring and cohort locations. Let $\boldsymbol{R}^* \in \mathbb{R}^{n^* \times d}$ and $\boldsymbol{R} \in \mathbb{R}^{n \times d}$ be matrices containing values of d geographic covariates (which may include spatial splines) at monitoring and cohort locations, respectively. Let $\boldsymbol{U}^* \in \mathbb{R}^{n^* \times K}$ denote an assignment matrix for monitoring locations, with each row having a 1 in a single entry and zeros in all other entries. If $U^*_{ik} = 1$, observation i is assigned to cluster k. Denote by $\mathcal{U}$ the set of matrices of this form.

For a two-stage exposure-health analysis, we first cluster the mass fraction observations to reduce dimension and identify representative component profiles. Then only cluster labels (assignments), not full exposure vectors, need to be predicted at cohort locations. The procedure can be broken down into the following steps:

Step 1: Cluster monitoring data

(a) Create cluster centers $\boldsymbol{M} = \begin{bmatrix} \boldsymbol{\mu}_1 & \cdots & \boldsymbol{\mu}_K \end{bmatrix}$ from the monitoring data $\boldsymbol{X}^*$.

(b) Make cluster assignments $\boldsymbol{U}^*$ at monitor locations s^* by assigning each location to the cluster with the closest center.

Step 2: Predict cluster membership

(a) Train a classification model for predicting cluster assignments using covariates $\boldsymbol{R}^*$ and cluster assignments $\boldsymbol{U}^*$ at monitoring locations.

(b) Predict cluster assignments $\widehat{\boldsymbol{U}}$ at cohort locations using this classification model and covariates $\boldsymbol{R}$.

Cluster assignments from Step 2(b) can be used as effect modifiers of the association between health outcomes and total $\mathrm{PM}_{2.5}$ mass, which we assume has already been predicted at subject locations. By separating the procedure into two steps (clustering and prediction), we allow for flexibility in the choice of a prediction model, recognizing that different methods may perform better in certain scenarios.

In the following subsections we describe the procedure in more detail. In Section 2.4 we present an alternative to k-means clustering for Step 1(a), which leads to improved performance in Step 2 and increased power to detect effect modification in a health analysis.

2.3.1 Step 1: Clustering Monitoring Data

The widely-used k-means algorithm provides a straightforward way to simultaneously define cluster centers for the mass fraction data (Step 1(a)) and make cluster assignments at monitor locations (Step 1(b)). The k-means solution is a reduction, indexed by the assignment matrix $\boldsymbol{U}^*$, of multivariate data ($\boldsymbol{X}^*$) into K clusters, each identified by its center (or representative vector) $\boldsymbol{\mu}_k$, that minimizes the within-cluster Sum-of-Squares (wSS^*):

$$wSS^* = \frac{1}{n^*} \left|\left| \boldsymbol{X}^* - \boldsymbol{U}^* \boldsymbol{M}^\mathsf{T} \right|\right|_F^2 , \tag{2.1}$$

where $\boldsymbol{M} = \begin{bmatrix} \boldsymbol{\mu}_1 & \cdots & \boldsymbol{\mu}_K \end{bmatrix}$. The center for the kth cluster is the mean of the vectors assigned to that cluster: $\boldsymbol{\mu}_k = \frac{1}{N_k^*} \sum_{i:U_{ik}^*=1} \boldsymbol{x}_i^*$, where $N_k^* = \sum_{i=1}^{n^*} U_{ik}^*$ is the number of observations in the kth cluster. Implementations of the k-means algorithm, often that of Hartigan and Wong (1979), exist in many statistical packages, which makes this approach easy to implement using existing software.

2.3.2 Step 2: Predicting Cluster Membership

The classification model chosen for Step 2 can be any multi-class prediction method. Here we focus on multinomial logistic regression although we also consider other methods such as support vector machines (SVMs) in the simulations and particulate matter analysis.

For multinomial logistic regression, let $Z_i \in \{1, \ldots, K\}$ denote the assignment of observation i to one of K classes (here, clusters from Step 1). The multinomial logistic regression model postulates that

$$\begin{aligned} \log \frac{P(Z_i = k)}{P(Z_i = K)} &= \boldsymbol{r}_i^\mathsf{T} \boldsymbol{\gamma}_k \quad \text{for } k = 1, \ldots K-1, \\ P(Z_i = K) &= 1 - \sum_{i=1}^{K-1} P(Z_i = k), \end{aligned} \tag{2.2}$$

where $\boldsymbol{\Gamma} = (\boldsymbol{\gamma}_1, \ldots, \boldsymbol{\gamma}_{K-1})$ is a matrix of regression coefficients and $\boldsymbol{r}_i^\mathsf{T}$ is a row of $\boldsymbol{R}$. The system (2.2) defines a generalized linear model, and maximum likelihood estimates of $\boldsymbol{\Gamma}$ can be computed using a standard iteratively-reweighted least squares algorithm. Rewriting (2.2) as the softmax function

$$P(Z_i = k; \boldsymbol{\Gamma}, \boldsymbol{r}_i) = \frac{\exp(\boldsymbol{r}_i^\mathsf{T} \boldsymbol{\gamma}_k)}{1 + \sum_{k'=1}^{K-1} \exp(\boldsymbol{r}_i^\mathsf{T} \boldsymbol{\gamma}_{k'})} \tag{2.3}$$

and plugging in the maximum likelihood estimates $\widehat{\boldsymbol{\Gamma}} = (\hat{\boldsymbol{\gamma}}_1, \ldots, \hat{\boldsymbol{\gamma}}_{K-1})$ yields classification probabilities for each observation. The matrix $\hat{\boldsymbol{U}}$ of predicted cluster membership is created by assigning each observation to the cluster with the largest classification probability:

$$\hat{u}_{ik} = \begin{cases} 1 & \text{if } P(Z_i = k; \hat{\boldsymbol{\Gamma}}, \boldsymbol{r}_i) > P(Z_i = k'; \hat{\boldsymbol{\Gamma}}, \boldsymbol{r}_i) \text{ for } k' \neq k, \\ 0 & \text{otherwise.} \end{cases} \tag{2.4}$$

2.3.3 Evaluating Predictive Error

The performance of the clustering procedure can be evaluated by the mean-squared prediction error (*MSPE*) across cohort locations, $MSPE = \frac{1}{n}||\boldsymbol{X} - \widehat{\boldsymbol{U}}\boldsymbol{M}^\mathsf{T}||_F^2$, which gives the sum of squared distances between observations $\boldsymbol{X}$ and the centers of the clusters to which each observation is predicted to belong ($\widehat{\boldsymbol{U}}\boldsymbol{M}^\mathsf{T}$). *MSPE* can be broken down into two components: representativeness of the cluster centers and accuracy of predicted cluster membership.

Similar to representativeness at monitor locations, which is quantified by wSS^* as defined in (2.1), cluster representativeness at cohort locations is computed as $wSS = \frac{1}{n}||\boldsymbol{X} - \boldsymbol{U}\boldsymbol{M}^\mathsf{T}||_F^2$. The matrix $\boldsymbol{U} = \arg\min_{\widetilde{\boldsymbol{U}} \in \mathcal{U}} \left|\left|\boldsymbol{X} - \widetilde{\boldsymbol{U}}\boldsymbol{M}\right|\right|_F^2$ contains assignment to the nearest cluster (which may not be the cluster to which a location was predicted to belong).

The accuracy of predicted cluster membership is quantified using two metrics, classification accuracy (*Acc*) and mean-squared misclassification error (*MSME*). Classification accuracy is the proportion of locations correctly classified: $Acc = \frac{1}{n}\sum_{k=1}^{K}\sum_{i:\boldsymbol{U}_{ik}=1} \mathbb{1}(\widehat{\boldsymbol{U}}_{ik} = 1)$. The straightforward interpretation of *Acc* makes it an attractive metric. However, *Acc* does not account for the magnitude of misclassification. *MSME* provides this information, by averaging the squared distances between the closest cluster centers $\boldsymbol{U}\boldsymbol{M}^\mathsf{T}$ and the predicted cluster centers $\widehat{\boldsymbol{U}}\boldsymbol{M}^\mathsf{T}$. That is, $MSME = \frac{1}{n}\left|\left|\boldsymbol{U}\boldsymbol{M}^\mathsf{T} - \widehat{\boldsymbol{U}}\boldsymbol{M}^\mathsf{T}\right|\right|_F^2$.

All of these measures require knowing the (typically unavailable) cohort observations $\boldsymbol{X}$, but in applications can be estimated via cross-validation. Because wSS and *MSME* are on the same scale, we can directly compare them to assess the tradeoff between representativeness and prediction accuracy, analogous to trading off between bias and variance, respectively, to achieve lower mean squared error in parameter estimation.

2.4 Covariate-adaptive Clustering of Spatially Misaligned Data

The k-means algorithm clusters multi-pollutant observations at monitored locations, but does not account for the need to predict cluster membership at cohort locations (Step 2), which is required for spatially misaligned data. There is no reason to expect that membership in clusters identified by k-means using pollutant observations at monitoring locations will be accurately predicted at subject locations using geographic covariates. If cluster membership cannot be predicted well at subject locations, then the identified clusters are of little use for epidemiological analysis.

To address this problem, we propose incorporating the geographic covariates that will be used for predicting cluster membership into the procedure for defining cluster centers. We first use a soft-assignment procedure, described in Section 2.4.1, that yields cluster centers. We then make hard assignments to clusters by minimizing the distance between observations and their assigned cluster center, in the same manner as k-means. We refer to this clustering procedure as *predictive k-means.*

2.4.1 Defining Cluster Centers for Predictive k-means

Let Z_i^* be a latent random variable that takes on values $k = 1, \dots, K$ and represents cluster membership. We relate this variable to the covariates $\boldsymbol{r}_i^*$ via a multinomial logistic regression model. Let $q_k(\boldsymbol{r}_i^*, \boldsymbol{\Gamma})$ denote $P(Z_i^* = k; \boldsymbol{\Gamma}, \boldsymbol{r}_i^*)$, with the latter defined as the softmax function in (2.3). Conditional on the value of Z_i^*, assume that the observation $\boldsymbol{x}_i^*$ is normally distributed as $(\boldsymbol{x}_i^* | Z_i^* = k) \sim \mathcal{N}_p(\boldsymbol{\mu}_k, \sigma^2 \boldsymbol{I})$. This model implies the following log-likelihood function:

$$\begin{aligned}\ell(\boldsymbol{\Gamma}, \boldsymbol{M}, \sigma^2 | \boldsymbol{X}^*; \boldsymbol{R}^*) = \sum_{i=1}^{n} \log \Bigg(\sum_{k=1}^{K} & q_k(\boldsymbol{r}_i^*, \boldsymbol{\Gamma}) (2\pi\sigma^2)^{-p/2} \\ & \times \exp\left(-\frac{1}{2\sigma^2} ||\boldsymbol{x}_i^* - \boldsymbol{\mu}_k||^2 \right) \Bigg). \end{aligned} \tag{2.5}$$

The log-likelihood in (2.5) corresponds to a one-level *mixture of experts* problem (Jordan and Jacobs 1994). Mixture of experts models use a set classification models (the 'experts') that are combined via a 'gating' network that uses soft assignment to select between experts. By incorporating hierarchical levels of gating networks, mixture of experts models can be quite flexible. Following the approach of Jordan and Jacobs (1994), we solve (2.5) using the EM algorithm with iterative updates to $\hat{\boldsymbol{\mu}}_k$, $\hat{\sigma}^2$, and $\hat{\boldsymbol{\Gamma}}$. Details of the algorithm are provided in Appendix A.1.

Using this approach, the cluster centers $\boldsymbol{\mu}_k$ (columns of $\boldsymbol{M}$) depend upon the covariates $\boldsymbol{R}^*$ via a multinomial logistic regression model for cluster assignment. The incorporation of prediction covariates into the cluster centers improves the accuracy of predicting cluster membership at cohort locations.

The parameter estimates $\hat{\boldsymbol{\Gamma}}$ provide 'working' cluster assignments for monitor locations. This suggests an alternative approach for prediction in which the cluster membership at cohort locations is predicted using $q_k(\boldsymbol{r}_i, \hat{\boldsymbol{\Gamma}})$ instead of building a separate classification model (Step 2). Such an approach, however, does not use optimal assignments (conditional on identified cluster centers) at monitor locations. Furthermore, we wish to avoid the assumption that a single parametric latent variable mixture model is a good model for the complicated processes that generate the particulate matter under study. In the simulations and $\text{PM}_{2.5}$ analysis, we compare this approach to multinomial logistic regression and classification using an SVM.

2.4.2 The Role of the Variance

The parameter σ^2 implicitly controls the tradeoff between representativeness and predictive accuracy. As $\sigma^2 \to 0$, the optimization problem of maximizing the log-likelihood (2.5) reduces to the k-means optimization problem, assuming all q_k are non-zero (Bishop 2006, Chap. 9). For predictive k-means, we restrict σ^2 to be positive, but small values of σ^2 allow for increased representativeness (smaller wSS) while larger values of σ^2 allow for improved predictive accuracy (smaller $MSPE$ and $MSME$) at the cost of decreasing representativeness.

Here we estimate σ^2 using maximum likelihood, as described in Section 2.4.1. An alternative approach is to select σ^2 using cross-validation (CV). The predictive k-means procedure (selection of cluster centers, assignment of monitors to clusters, fitting of classification model, and prediction of cluster membership) could be repeated on CV data sets for various fixed values of σ^2, and then the value of σ^2 that yielded the smallest cross-validated value of $MSPE$ selected for use in the primary analysis. However, this can be computationally impractical in situations where CV is already being used for model selection. For that reason, we do not select σ^2 by CV in the analysis of $\mathrm{PM}_{2.5}$ components in Section 2.6, but we provide an example of this approach in the simulations.

2.5 Simulations

We conducted two sets of simulations to demonstrate the clustering approaches presented here. The first set illustrates the differences between the clusters from predictive k-means and standard k-means procedures in a two-dimensional setting that allows for easy visualization of the centers. The second set demonstrates the methods in a higher-dimensional setting and includes a simulated health analysis to elucidate benefits in power achieved by using clusters from predictive k-means.

2.5.1 Two-dimensional Exposures

For the first simulation set, we consider two-dimensional exposures (X_1, X_2) and three independent covariates (R_1, R_2, and W). Only $R_1 \sim N(0,1)$ and $R_2 \sim N(0,1)$ are observed, while $W \sim Bernoulli(0.5)$ is unobserved. The covariates determine membership in one of four underlying clusters (denoted by $Z \in \{1,2,3,4\}$), constructed so that two clusters cannot be distinguished using the observed covariates:

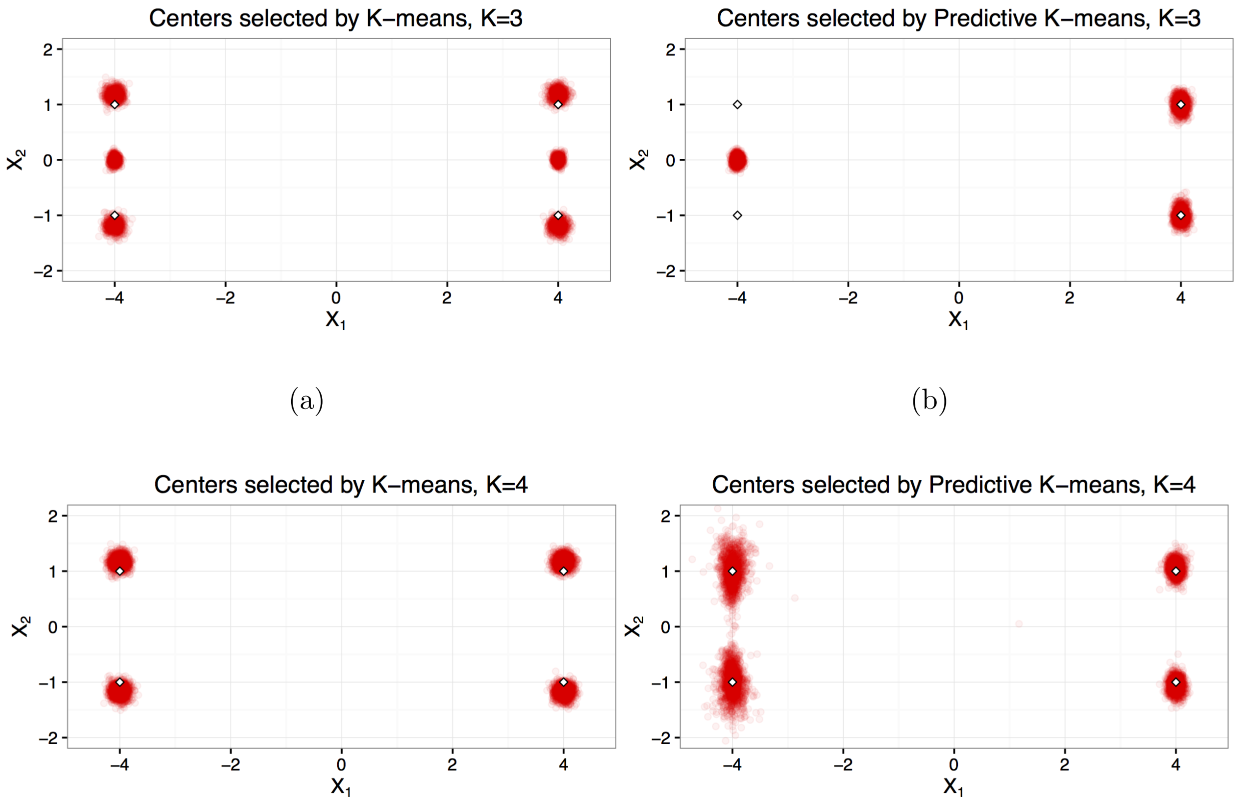
$$Z = \begin{cases} 1 & \text{if } R_1 < 0 \text{ and } W = 1, \text{ for all } R_2, \\ 2 & \text{if } R_1 < 0 \text{ and } W = 0, \text{ for all } R_2, \\ 3 & \text{if } R_1 > 0 \text{ and } R_2 > 0, \text{ for all } W, \\ 4 & \text{if } R_1 > 0 \text{ and } R_2 < 0 \text{ for all } W. \end{cases}$$

Conditional on cluster membership, the exposures X_1 and X_2 are normally distributed: $(X_1|Z = k) \sim N(\mu_{k1}, 1)$ and $(X_2|Z = k) \sim N(\mu_{k2}, 1)$, where $\boldsymbol{\mu}_1 = (-4, 1)$, $\boldsymbol{\mu}_2 = (-4, -1)$, $\boldsymbol{\mu}_3 = (4, 1)$, and $\boldsymbol{\mu}_4 = (4, -1)$. By design, observations from clusters 1 and 2 cannot be distinguished using the observed covariates available for prediction.

For a set of 1000 replications, each with a sample size of $n = 1000$, cluster centers were identified using the k-means and predictive k-means procedures described in Sections 2.3 and 2.4. The iterative optimization algorithms for both methods are not guaranteed to find global optima (Jordan and Jacobs 1994), so 50 different starting values were used for optimization. For each replication, a second sample of 1000 observations was drawn from the same data generating mechanism and underlying cluster membership at these test locations predicted using multinomial logistic regression with covariates R_1 and R_2. This simulation was done twice, once identifying $K = 3$ clusters and once identifying $K = 4$ clusters.

When $K = 3$, we are selecting a number of clusters fewer than the number in the data generating model. This scenario is plausible in applications when the underlying data generating mechanism is not fully known. We see in Figure 2.1a that k-means correctly identified two cluster centers (either $\boldsymbol{\mu}_1$ and $\boldsymbol{\mu}_2$ or $\boldsymbol{\mu}_3$ and $\boldsymbol{\mu}_4$) and would estimate the center of the third cluster as approximately $(4, 0)$ or $(-4, 0)$, respectively. Because k-means does not incorporate R_1 or R_2 into the cluster centers, the estimated centers are evenly split between these two possibilities. On the other hand, Figure 2.1b shows that the predictive k-means procedure estimated centers in approximately the same location for all replications: $(4, 1)$, $(4, -1)$, and $(-4, 0)$. The first two clusters correspond to $\boldsymbol{\mu}_3$ and $\boldsymbol{\mu}_4$, while the third center estimated by predictive k-means is directly between $\boldsymbol{\mu}_1$ and $\boldsymbol{\mu}_2$, which are indistinguishable by the prediction covariates R_1 and R_2.

Measures of representativeness and predictive accuracy are reported in Table 2.1. The classification accuracy of k-means is 0.83, and predictive k-means improves upon this by eight percentage points (0.91). While wSS is less than 1% higher for predictive k-means than for regular k-means, misclassification error ($MSME$) drops by more than 50% (0.54 for predictive k-means compared to 1.13 for regular k-means).



(a)

(b)

(c)

(d)

Figure 2.1: Cluster centers from Simulation 1. Figures (a) and (c) are the centers identified by regular k-means when $K = 3$ and $K = 4$, respectively. Figures (b) and (d) are the centers identified by predictive k-means when $K = 3$ and $K = 4$, respectively. Each point in the clouds is a cluster center from a single replication; the outlined diamonds denote the latent cluster centers.

Table 2.1: Performance measures from Simulation 1, for $K = 3$ and $K = 4$ when using informative covariates (R_1 and R_2) and uninformative covariates (White Noise) to predict cluster membership. PKM, predictive k-means. KM, k-means. The measures of performance (*MSPE*, *wSS*, *MSME*, and *Acc*) are described in Section 3.3 of the main text.

K	Prediction Covariates	Clustering Method	*MSPE*	*wSS*	*MSME*	*Acc*
3	R_1 and R_2	KM	3.31	2.33	1.13	0.83
		PKM	2.75	2.34	0.54	0.91
	White Noise	KM	35.1	2.33	32.8	0.50
		PKM	35.0	2.36	32.5	0.50
4	R_1 and R_2	KM	3.42	1.65	2.02	0.66
		PKM	3.12	1.69	1.53	0.68
	White Noise	KM	36.3	1.65	34.7	0.25
		PKM	35.9	1.68	34.0	0.26

When $K = 4$, we are selecting the same number of clusters as in the data generating mechanism. In this scenario, predictive k-means also provides measurable improvement in predictive accuracy, as *MSME* drops by almost 25% (from 2.02 to 1.53) with little loss in representativeness (*wSS* increases by 2%). Predictive k-means achieves this tradeoff by selecting centers corresponding to clusters 1 and 2 (Figure 2.1d) that are closer to one another than the centers identified by k-means (Figure 2.1c). This reduces prediction error when cluster membership is incorrectly predicted.

These simulations demonstrate how when informative covariates (R_1, R_2) are allowed to influence cluster centers, we can get substantial improvements in predictive accuracy with little loss in representativeness. This simulation was repeated using uninformative covariates (i.i.d. $N(0, 1)$ random variables independent of all other covariates and the outcome) in the predictive k-means procedure and to predict cluster membership in the test set. The results from this simulation, also presented in Table 2.1, show that predictive k-means performs essentially the same as k-means when the covariates do not provide useful information.

2.5.2 Multi-pollutant Spatial Exposures

For the second set of simulations, we simulated long-term average observations for $p = 15$ pollutants at 7,333 AQS monitor locations throughout the contiguous United States. We first assigned each location to belong to one of three latent spatial clusters and one of three latent non-spatial clusters, with membership denoted by $A_i \in \{1, 2, 3\}$ and $B_i \in \{1, 2, 3\}$, respectively. To assign A_i, a correlated spatial surface was simulated according to the model $\boldsymbol{z} \sim N(\widetilde{\boldsymbol{x}}^L + \widetilde{\boldsymbol{y}}^L, 0.25\boldsymbol{V})$, where $\widetilde{x}_i^L$ and $\widetilde{y}_i^L$ are normalized versions of the Lambert coordinates x_i^L and y_i^L. The matrix $\boldsymbol{V}$ has exponential covariance structure: $V_{ij} = \exp(-||(x_i^L, y_i^L) - (x_j^L, y_j^L)||_2/400)$. This surface was partitioned into tertiles to give the values A_i. Membership in the non-spatial clusters (B_i) was assigned using i.i.d. draws from a uniform distribution.

Conditional on latent cluster membership, the pollutant observations $\boldsymbol{x}_i$ at each location were simulated from a log-normal distribution:

$$(x_{ij}|A_i = k, B_i = k') \sim LN\left(\log(4 + a_{jk} + b_{jk'}) - 0.125, 0.25\right)$$

for $j = 1, \ldots, p$. The component means $\mathrm{E}[x_{ij}|A_i = k, B_i = k'] = 4 + a_{jk} + b_{jk}$ are combinations of coefficients determined from spatial and non-spatial cluster memberships. Each a_{jk} (for $j = 1, \ldots, p$ and $k = 1, \ldots, K$) is an independent observation from the normal distribution $N(0, \sigma_A^2)$. Similarly, $b_{jk} \sim N(0, \sigma_B^2)$. We considered two settings for (σ_A^2, σ_B^2): $(1, 2)$, which induces greater separation between clusters in the non-spatial partition than between clusters in the spatial partition, and $(2.5, 0.5)$, which results in greater separation between clusters in the spatial partition. In the latter scenario we expect both k-means and predictive k-means to find similar cluster centers, since the greatest between-cluster separation is among clusters that depend upon spatial covariates. The component concentrations were converted to mass fractions by dividing by total particulate matter, i.e. $\tilde{x}_{ij} = x_{ij}/PM_i$, where $PM_i = \sum_{j=1}^{p} x_{ij}$.

For each of 500 replications, 200 locations were randomly selected as 'monitors' and the remaining locations served as 'cohort' locations. Cluster centers were estimated from the mass fractions $\tilde{x}_{ij}$ at 'monitor' locations via regular k-means and predictive k-means, using a matrix of thin-plate regression splines (TPRS) with 15 degrees of freedom (df) as $\boldsymbol{R}^*$. We present results for estimating the mixture model variance parameter σ^2 via maximum-likelihood and via CV. Cluster membership at 'cohort' locations was predicted using multinomial logistic regression (MLR), an SVM with radial kernel, and the working coefficients from the mixture of experts model.

Predicted cluster assignments were then used as interaction variables in a linear regression analysis of the association between SBP and PM. Blood pressure measurements for each 'cohort' location were simulated as $y_i = 115 + \sum_{j=1}^{p} x_{ij}\beta_j + \epsilon_i$, where $\epsilon_i \sim N(0, \sigma^2_{SBP})$. The values of β_j were chosen so that the variability in the SBP–PM association was the same among the latent spatial and non-spatial clusters. For each set of predicted cluster assignments $\widehat{\boldsymbol{U}}$, we fit the linear model

$$\mathrm{E}\left[y_i | PM_i, \widehat{U}_i\right] = \beta_0 + \beta_{02} I_{\widehat{U}_i=2} + \beta_{03} I_{\widehat{U}_i=3} + \beta_1 PM_i + \beta_{12} PM_i I_{\widehat{U}_i=2} + \beta_{13} PM_i I_{\widehat{U}_i=3}.$$

A Wald test of the null hypothesis $H_0 : \beta_{12} = \beta_{13} = 0$ was performed to determine whether there were between-cluster differences in the association between SBP and PM.

When $(\sigma^2_A, \sigma^2_B) = (1, 2)$, overall prediction error was lowest for predictive k-means with σ^2 selected by CV and MLR used as the classifier ($MSPE = 15.02$). Misclassification error ($MSME$) was more than 50% smaller for predictive k-means compared to regular k-means (1.72 compared to 4.18) and classification accuracy was 15 percentage points higher (see Table 2.2). The clusters identified by predictive k-means were only slightly less representative (wSS of 13.57 and 13.69) than those identified by k-means (13.38).

The power for detecting a between-cluster difference in health effect (at the $\alpha = 0.05$ level) is plotted in Figure 2.2 for varying values of σ_{SBP}. In the setting $(\sigma^2_A, \sigma^2_B) = (1, 2)$, all three prediction methods gave similar results for predictive k-means with σ^2 selected by maximum likelihood, while MLR performed best for clusters from regular k-means and

Table 2.2: Measures of representativeness (wSS) and predictive accuracy ($MSPE$, $MSME$, Acc) for Simulation 2. Results provided for clusters identified by k-means and predictive k-means, using either maximum likelihood (EM) or cross-validation (CV) for selecting σ^2. Predictions were made using multinomial logistic regression (MLR), support vector machines (SVM), and the working coefficients from the Mixture of Experts algorithm (ME-Working).

(σ_A^2, σ_B^2)	Clustering Method	Prediction Method	$MSPE$	wSS	$MSME$	Acc
$(1, 2)$	k-means	MLR	16.80	13.38	4.18	0.45
		SVM	17.09	13.38	4.43	0.42
	Predictive k-means with $\hat{\sigma}^2$ selected by EM	MLR	15.43	13.57	2.45	0.58
		SVM	15.75	13.57	2.47	0.54
		ME-Working	15.68	13.57	2.63	0.55
	Predictive k-means with $\hat{\sigma}^2$ selected by CV	MLR	15.03	13.69	1.72	0.60
		SVM	15.27	13.69	1.86	0.57
		ME-Working	15.17	13.69	1.81	0.58
$(2.5, 0.5)$	k-means	MLR	14.12	12.97	1.92	0.75
		SVM	14.31	12.97	2.07	0.72
	Predictive k-means with $\hat{\sigma}^2$ selected by EM	MLR	13.79	12.90	1.46	0.79
		SVM	13.89	12.90	1.54	0.78
		ME-Working	14.01	12.90	1.65	0.76
	Predictive k-means with $\hat{\sigma}^2$ selected by CV	MLR	13.75	12.90	1.37	0.80
		SVM	13.88	12.90	1.48	0.78
		ME-Working	13.85	12.90	1.44	0.78

predictive k-means with σ^2 chosen by CV. The highest power was obtained by predictive k-means with σ^2 selected by CV (0.76 at $\sigma_{SBP} = 4$), followed by predictive k-means with σ^2 chosen by maximum likelihood (0.60) and regular k-means (0.42). When true (oracle) cluster assignments were used, the power from regular k-means clusters (0.90) exceeded that from predictive k-means clusters with σ^2 chosen by maximum likelihood (0.78). This demonstrates that the benefits in power for predictive k-means are due to the improved predictive accuracy, despite the slight loss in representativeness.

When $(\sigma_A^2, \sigma_B^2) = (2.5, 0.5)$, representativeness was essentially the same for both methods (12.97 for k-means, 12.90 for predictive k-means for both approaches to selecting σ^2). Although overall prediction error was only slightly smaller for predictive k-means, prediction accuracy was 4 percentage points higher and misclassification error approximately 25% lower for predictive k-means (1.46 and 1.37 versus 1.92). The power for detecting effect modification was essentially the same for all clustering and classification methods, with the exception of low power when the SVM approach and the mixture of experts working coefficients were used for predictive k-means with σ^2 chosen by CV (see Figure 2). These results show that predictive k-means and k-means have comparable performance in settings where they are identifying similar cluster centers.

2.6 PM$_{2.5}$ components and NIEHS Sister Study

To expand upon the analysis of Chan et al. (2015), we investigated the relationship between SBP and long-term exposure to PM$_{2.5}$, grouping subjects by predicted membership in clusters with different component profiles. Our analysis included 47,206 cohort subjects with complete covariate information.

We obtained data for 130 AQS monitoring locations that in 2010 measured mass concentration for twenty-two PM$_{2.5}$ component species (elemental carbon [EC], organic carbon [OC], NO_3^-, SO_4^{2-}, Al, As, Br, Cd, Ca, Co, Cr, Cu, Fe, K, Mn, Na, S, Si, Se, Ni, V, and Zn) in addition to measurements of PM$_{2.5}$ mass made in accordance with Federal Reference Methods. Annual averages were computed by averaging all available daily observations from each

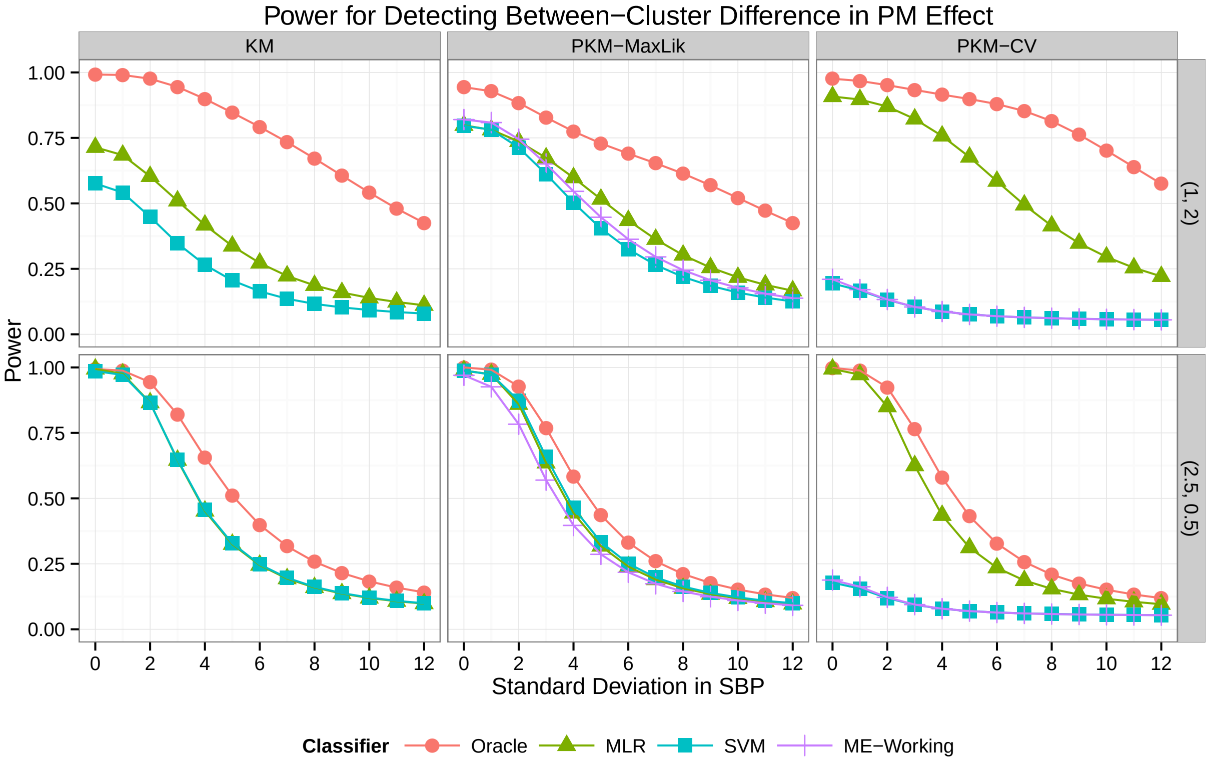


Figure 2.2: Power for detecting a between-cluster difference in SBP-PM association at the $\alpha = 0.05$ level in Simulation 2. Clusters identified by k-means (KM) and predictive k-means with σ^2 chosen by maximum likelihood (PKM-MaxLik) or cross-validation (PKM-CV). Cluster membership was predicted using multinomial logistic regression (MLR), SVM, working coefficients from the mixture of experts model (ME-Working), or oracle assignment using true exposure values. The rows correspond to $(\sigma_A^2, \sigma_B^2) = (1, 2)$ and $(\sigma_A^2, \sigma_B^2) = (2.5, 0.5)$.

monitoring location having at least 41 measurements in the calendar year with a maximum gap of 45 days between observations. We converted mass concentrations to mass fractions by dividing the annual average of each species at a monitoring location by the annual average $PM_{2.5}$ concentration at that location. To make the distribution of mass fractions within each component more symmetric, we log-transformed the mass fractions.

We applied the predictive k-means method to this monitoring data, selecting the number of clusters and the covariates by 10-fold cross-validation. Because of the limited number of observations ($n^* = 130$), we only investigated $K \leq 10$. We computed the first three scores from a principal component analysis (PCA) of a collection of more than 200 geographic covariates. We considered models with either 2 or 3 PCA scores and TPRS with either 5 or 10 df, with the same covariates used for determining cluster centers and in the classification model. The smallest cross-validation $MSPE$ was for the model with $K = 8$ clusters and a combination of 2 PCA scores and 10 df TPRS as the covariates. Appendix Table A.1 provides CV performance metrics for various models with $K = 8$ and Table A.2 provides metrics for other choices of K. A support vector machine (SVM) was used as the classification model, because it resulted in better cross-validated predictive accuracy ($MSPE = 18.33$) than multinomial logistic regression (21.28) or using the working coefficients from the mixture-of-experts model (24.33). For comparison, we applied regular k-means to the component data using the same prediction covariates. Cross-validated prediction accuracy was slightly better for predictive k-means than regular k-means, with the former trading a small reduction in representativeness for notable improvements in misclassification error and classification accuracy (see Table A.1).

The cluster centers identified by predictive k-means are plotted in Figure 2.3 and a map of assigned membership for monitors is provided in Figure 2.4a. Many of the monitor locations in the Midwest and Mid-Atlantic regions were assigned to Cluster 1 ($n^* = 32$), which has above-average mass fractions of SO_4^{2-} and NO_3^- , suggestive of high ambient ammonia levels from agricultural emissions favoring particulate over gaseous NO_3^- (U.S. EPA 2003). Cluster 2 ($n^* = 26$) included monitors from New England, the southeastern coast, and parts of the

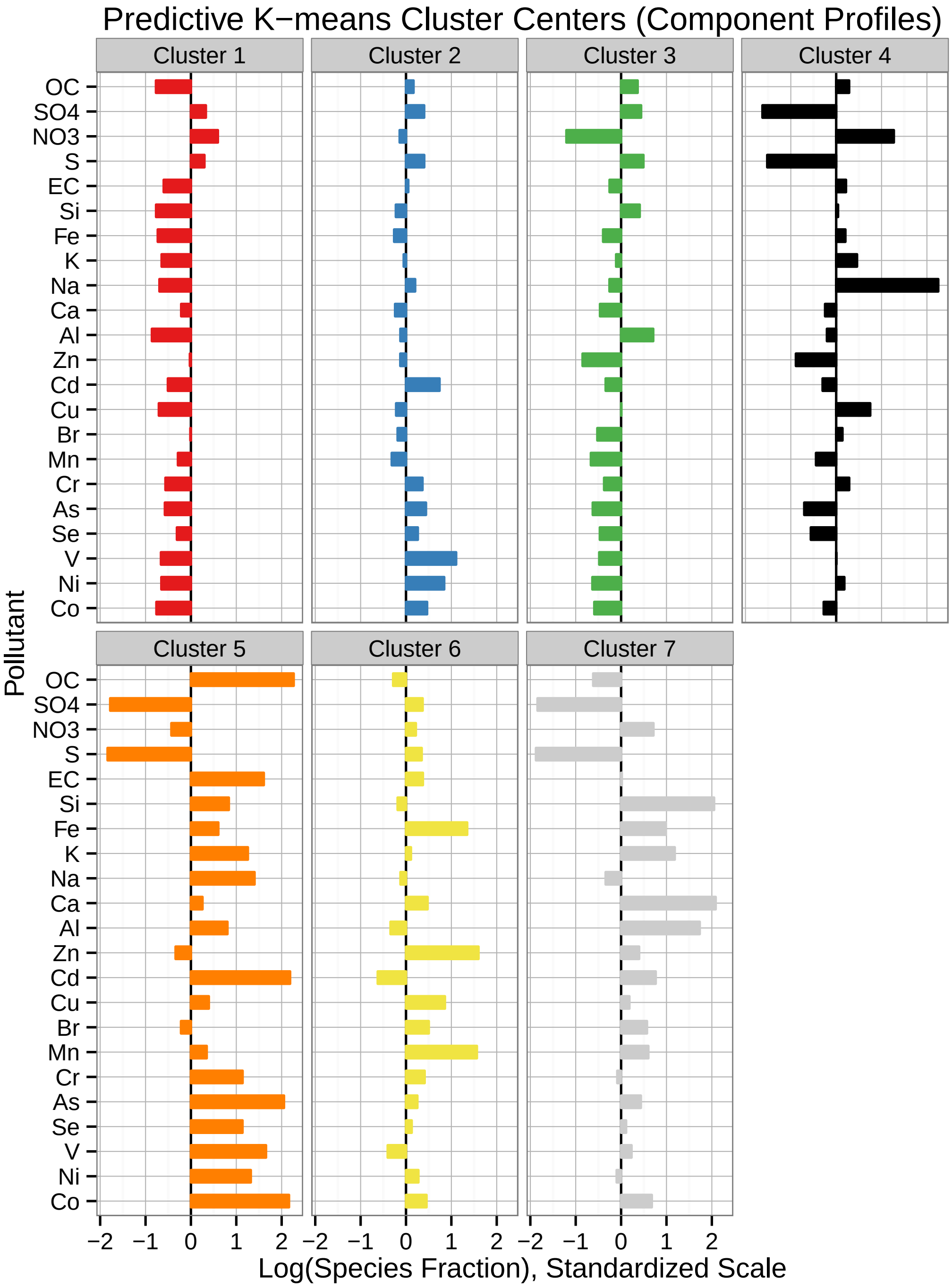
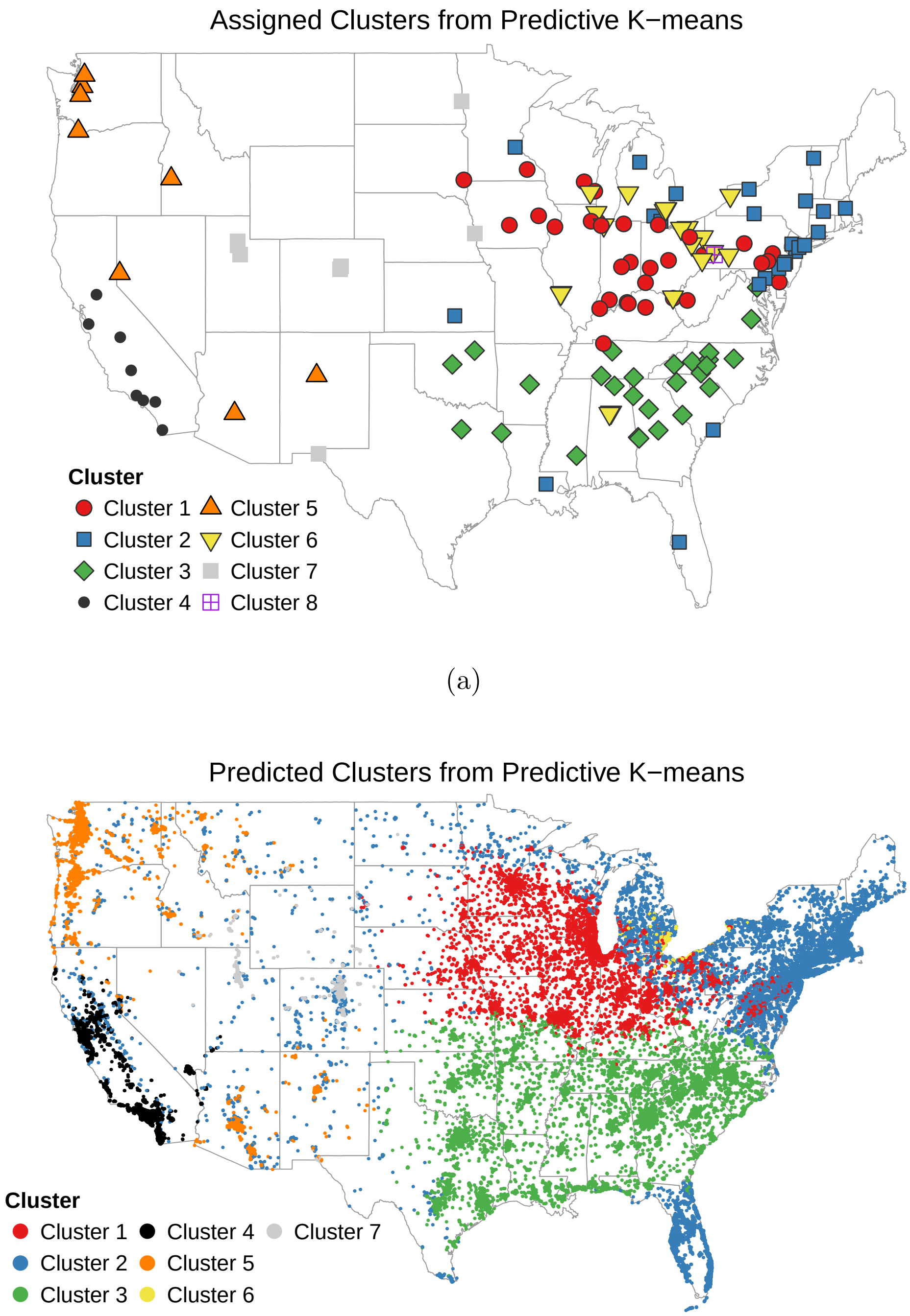


Figure 2.3: Cluster centers identified by predictive k-means in the 2010 annual average $PM_{2.5}$ component data. Species mass fractions were log transformed and then standardized, so values shown represent relative composition. Components are ordered by decreasing mass concentration.

upper Midwest, and had higher fractions of Cd, V and Ni, which are associated with ship emissions (Thurston et al. 2013) and residual oil burning in New York City (Peltier et al. 2009). Monitors in the Southeast were mostly assigned to Cluster 3 ($n^* = 27$) and had a component profile notable for its relatively low fraction of particulate nitrate (NO_3^-) relative to sulfate (SO_4^{2-}), a pattern that has previously been attributed to high amounts of acidic sulfate and low levels of ammonia in the region (Blanchard and Hidy 2003). The California monitors were grouped into Cluster 4 ($n^* = 8$), which also had low sulfur fractions and large fractions of sodium and nitrate particles, likely from marine aerosols and agricultural emissions, respectively. Cluster 5 ($n^* = 8$) included monitors from the Pacific Northwest and Southwest, with high fractions of almost all pollutants except sulfate. Cluster 6 ($n^* = 20$) had high fractions of Fe, Zn, and Mn, which are indicative of emissions from steel furnaces and other metal processing (Thurston et al. 2013), and the monitors assigned to this cluster were all near industrial plants of some kind. Cluster 7 ($n^* = 8$) had high fractions of the crustal elements Si, Ca, K and Al, indicative of the surface soil composition in the Western U.S. (Shacklette and Boerngen 1984). The eigth cluster was a single site outside of Pittsburgh, PA, which has been previously noted for non-attainment of air quality standards due to nearby industrial sources (U.S. EPA 2006).

Predicted assignments to the predictive k-means clusters at Sister Study cohort locations are mapped in Figure 2.4b. Predicted membership at cohort locations tended to follow the same general spatial patterns as monitor assignments, with some differences in the Mountain West and Mid-Atlantic regions. A majority of subjects were predicted to belong to Cluster 1 ($n = 12,828$), Cluster 2 ($n = 13,926$), or Cluster 3 ($n = 9,915$).

Using a linear model for SBP with the same confounders as Chan et al. (2015) (see Section 2.2), we estimated the association between SBP and long-term $PM_{2.5}$ exposure, stratifying exposure by cluster. We used predictions of 2010 annual average $PM_{2.5}$ concentrations from a universal kriging model following the same approach as Sampson et al. (2013). The association coefficient estimates are provided in Table 2.3. The estimated difference in SBP associated with a 10 $\mu g/m^3$ difference in $PM_{2.5}$ overall (without clustering) was 1.81 mmHg,



(a)

(b)

Figure 2.4: (a) Assigned predictive *k*-means cluster membership at AQS monitor locations. (b) Predicted cluster membership at Sister Study cohort locations (jittered to protect confidentiality).

which is higher than, but still contained within the confidence interval for, the estimate obtained by Chan et al. (2015) for 2006 annual average exposure. When estimating cluster-specific associations, Cluster 1 had a much stronger association (4.37 mmHg higher SBP for each 10 $\mu g/m^3$ difference in $\text{PM}_{2.5}$, 95% Confidence Interval [CI]: 2.38, 6.35) than the estimate that pools all subjects together. The point estimates for Clusters 3 and 4 were also higher (2.91 and 3.51, respectively) than the unclustered estimate. Although the point estimates for Clusters 5 and 6 were large (3.07 and 5.60, respectively), their confidence intervals were quite wide and included 0. In Clusters 2 and 7, there was no evidence of an association between $\text{PM}_{2.5}$ and SBP. A Wald test for effect modification showed that the differences between clusters were statistically significant ($p = 0.020$). As a sensitivity analysis, we explored adjusting for finer scale spatial variation and allowing the coefficients for the covariates in the health model to vary by $\text{PM}_{2.5}$ cluster assignment, however this did not substantively change the results (data not shown).

For comparison, we used regular k-means to cluster the $\text{PM}_{2.5}$ component data and predicted cluster membership at cohort locations using the same prediction covariates. The clustering results and association estimates are provided in Appendix A.4. The estimates for Clusters 1, 3, and 4 were attenuated compared to the predictive k-means analysis.

2.7 Discussion

We have presented a novel approach for clustering multivariate environmental exposures and predicting cluster assignments in cohort studies of health outcomes. The motivating application is air pollution epidemiology, where multi-pollutant exposure data are available from regulatory monitoring networks, but these monitors do not measure exposure at cohort locations. We first demonstrated how dimension reduction could be performed through the existing method of k-means clustering followed by spatial prediction. However, the clusters identified by k-means may not be predictable at subject locations, which makes them of limited use for epidemiological analysis. To address this, we introduced the predictive k-means method, which incorporates prediction covariates into the estimation of cluster centers.

Table 2.3: Estimated difference in SBP (in mmHg) associated with a 10 $\mu g/m^3$ difference in annual ambient $\text{PM}_{2.5}$ exposure. Cohort is partitioned by membership in clusters from predictive k-means.

Exposure	n	Est.	95% CI	p-value
Overall $\text{PM}_{2.5}$	47,206	1.81	(0.74, 2.88)	<0.001
$\text{PM}_{2.5}$ by Cluster				0.015[a]
Cluster 1	12,828	4.37	(2.38, 6.35)	0.000016
Cluster 2	13,926	0.77	(−1.19, 2.74)	0.44
Cluster 3	9,915	2.91	(0.19, 5.62)	0.036
Cluster 4	4,033	3.51	(0.68, 6.34)	0.015
Cluster 5	4,057	3.07	(−1.07, 7.21)	0.15
Cluster 6	1,029	5.60	(−0.71, 11.9)	0.08
Cluster 7	1,418	−2.11	(−6.55, 2.33)	0.35

[a] p-value for a Wald test for a difference between cluster coefficient estimates.

Through simulations, we demonstrated that clusters from predictive k-means provide substantial gains in prediction accuracy compared to the k-means approach. The simulations did not provide strong evidence to favor one of the three classification approaches compared (multinomial logistic regression, working coefficients from the mixture of experts model, and an SVM), however the SVM clearly outperformed the alternatives in the analysis of the $\text{PM}_{2.5}$ component data. In addition to improved predictive accuracy, the simulations demonstrated that predictive k-means clusters yield higher power for detecting effect modification by cluster membership.

As with any cluster analysis, the choice of the number of clusters is important. In our analysis of the $\text{PM}_{2.5}$ component data, we chose $K = 8$ based upon a cross-validation analysis. We restricted the candidate choices to $K \leq 10$ due to the need to have enough monitors assigned to each cluster so that a prediction model could be developed. The results of Simulation 1 suggest that the benefits of predictive k-means remain even when the chosen number of clusters does not match the underlying data generation mechanism. A potential extension of this method that we could explore in future work is to allow the cluster variance parameter (σ^2) to vary between clusters rather than assuming a constant value for the entire dataset.

A challenge for the predictive k-means approach is adequately accounting for uncertainty in cluster assignments in the health model. If one considers cluster assignment conditional on a fixed set of cluster centers, then the problem is an extension of multi-pollutant exposure prediction and could be addressed by extending the measurement error approaches of Bergen and Szpiro (2015). Accounting for uncertainty in predicted cluster assignment at the same time as determining the cluster centers is more difficult. Even for a fixed K, choosing different covariates for the predictive k-means model can result in different clusters, which makes interpretation of the clusters across models unclear. A direction for addressing this problem is the post-selection inference approaches of Berk et al. (2010) and Lee et al. (2016).

We found a significant association in the NIEHS Sister Study between SBP and 2010 long-term ambient $\mathrm{PM}_{2.5}$ exposure that was higher than previous estimates based upon 2006 exposure when ignoring $\mathrm{PM}_{2.5}$ composition (Chan et al. 2015). Although all baseline measurements on Sister Study participants were complete prior to 2010, we used 2010 measurements due to changes in the collection of $\mathrm{PM}_{2.5}$ speciation data during prior years. Using clusters identified by predictive k-means, we found that this association varied significantly by $\mathrm{PM}_{2.5}$ composition and was strongest among subjects predicted to belong to Clusters 1 and 3, which included most subjects living in the Midwest and Southeast. These results are consistent with the findings of Thurston et al. (2013), who found that $\mathrm{PM}_{2.5}$ exposure dominated by secondary aerosols were significantly associated with mortality. The strength of the estimated effects in clusters with component profiles notable for secondary aerosols may be due in part to the available speciation data, since the relatively small number of monitors means that the component data, and the clusters derived from them, capture regional variation better than small scale (within-city and near-source) variability.

By incorporating covariate information into cluster centers, the predictive k-means procedure performs dimension reduction appropriate for spatially-misaligned data. This method provides a useful tool for understanding how differences in exposure composition are associated with health effects.

2.8 References

Adar, S. D., R. Klein, B. E. K. Klein, et al. 2010. “Air pollution and the microvasculature: a cross-sectional assessment of in vivo retinal images in the population-based multi-ethnic study of atherosclerosis (MESA).” *PLoS medicine* 7 (11): e1000372.

Austin, E., B. Coull, D. Thomas, and P. Koutrakis. 2012. “A framework for identifying distinct multipollutant profiles in air pollution data.” *Environment International* 45:112–21.

Austin, E., B. A. Coull, A. Zanobetti, and P. Koutrakis. 2013. “A framework to spatially cluster air pollution monitoring sites in US based on the PM2.5 composition.” *Environment International* 59:244–54.

Bell, M. L., F. Dominici, K. Ebisu, S. L. Zeger, and J. M. Samet. 2007. “Spatial and temporal variation in PM2.5 chemical composition in the United States for health effects studies.” *Environmental Health Perspectives* 115 (7): 989–95.

Bergen, S., and A. A. Szpiro. 2015. “Mitigating the impact of measurement error when using penalized regression to model exposure in two-stage air pollution epidemiology studies”. *Environmental and Ecological Statistics* 22:601–631.

Berk, R., L. Brown, and L. Zhao. 2010. “Statistical inference after model selection”. *Journal of Quantitative Criminology* 26 (2): 217–236.

Bishop, C. 2006. *Pattern Recognition and Machine Learning.* Springer.

Blanchard, C. L., and G. M. Hidy. 2003. “Effects of changes in sulfate, ammonia, and nitric acid on particulate nitrate concentrations in the southeastern United States.” *Journal of the Air & Waste Management Association* 53:283–290.

Brauer, M. 2010. “How much, how long, what, and where: air pollution exposure assessment for epidemiologic studies of respiratory disease.” *Proceedings of the American Thoracic Society* 7 (2): 111–5.

Brauer, M., G. Hoek, P. van Vliet, et al. 2003. “Estimating long-term average particulate air pollution concentrations: application of traffic indicators and geographic information systems.” *Epidemiology* 14 (2): 228–39.

Brook, R. D., S. Rajagopalan, C. A. Pope, et al. 2010. “Particulate matter air pollution and cardiovascular disease: An update to the scientific statement from the American Heart Association.” *Circulation* 121 (21): 2331–78.

Chan, S. H., V. C. van Hee, S. Bergen, A. A. Szpiro, L. A. DeRoo, S. J. London, J. D. Marshall, J. D. Kaufman, and D. P. Sandler. 2015. “Long-term air pollution exposure and blood pressure in the Sister Study”. *Environmental Health Perspectives* 123 (10): 951–958.

Cohen, M. A., S. D. Adar, R. W. Allen, et al. 2009. “Approach to estimating participant pollutant exposures in the Multi-Ethnic Study of Atherosclerosis and Air Pollution (MESA Air).” *Environmental science & technology* 43 (13): 4687–93.

Dominici, F., L. Sheppard, and M. Clyde. 2003. “Health effects of air pollution: A statistical review”. *International Statistical Review* 71 (2): 243–276.

Franklin, M., P. Koutrakis, and J. Schwartz. 2008. “The role of particle composition on the association between PM2.5 and mortality”. *Epidemiology* 19 (5): 680–689.

Hartigan, J., and M. Wong. 1979. “Algorithm AS 136: A k-means clustering algorithm”. *Applied Statistics* 28 (1): 100–108.

Jordan, M., and R. Jacobs. 1994. “Hierarchical mixtures of experts and the EM algorithm”. *Neural computation* 6 (2): 181–214.

Keller, J. P., C. Olivers, S.-Y. Kim, L. Sheppard, P. D. Sampson, A. A. Szpiro, A. P. Oron, J. Lindström, S. Vedal, and J. D. Kaufman. 2015. “A Unified Spatiotemporal Modeling Approach for Predicting Concentrations of Multiple Air Pollutants in the Multi-Ethnic Study of Atherosclerosis and Air Pollution”. *Environmental Health Perspectives* 123 (4): 301–309.

Kioumourtzoglou, M.-A., E. Austin, P. Koutrakis, F. Dominici, J. Schwartz, and A. Zanobetti. 2015. “PM2.5 and survival among older adults: Effect modification by particulate composition”. *Epidemiology* 26 (3): 321–327.

Künzli, N, S Medina, and R Kaiser. 2001. “Assessment of deaths attributable to air pollution: should we use risk estimates based on time series or on cohort studies?” *American Journal of Epidemiology* 153 (11): 1050–1055.

Lee, J. D., D. L. Sun, Y. Sun, and J. E. Taylor. 2016. “Exact post-selection inference, with application to the lasso”. *Annals of Statistics* 44 (3): 907–927. arXiv: 1311.6238.

Oakes, M., L. Baxter, and T. C. Long. 2014. “Evaluating the application of multipollutant exposure metrics in air pollution health studies”. *Environment International* 69:90–99.

Peltier, R. E., S.-I. Hsu, R. Lall, and M. Lippmann. 2009. “Residual oil combustion: a major source of airborne nickel in New York City.” *Journal of exposure science & environmental epidemiology* 19 (6): 603–612.

Sampson, P. D., M. Richards, A. A. Szpiro, S. Bergen, L. Sheppard, T. V. Larson, and J. D. Kaufman. 2013. “A regionalized national universal kriging model using Partial Least Squares regression for estimating annual PM2.5 concentrations in epidemiology”. *Atmospheric Environment* 75:383–392.

Shacklette, H. T., and J. Boerngen. 1984. *Element concentrations in soils and other surficial materials of the conterminous United States.* Tech. rep.

Thurston, G. D., K. Ito, R. Lall, et al. 2013. “NPACT Study 4. Mortality and Long-Term Exposure to PM2.5 and Its Components in the American Cancer Society’s Cancer Prevention Study II Cohort”. In *National Particle Component Toxicity (NPACT) Initiative: Integrated Epidemiologic and Toxicologic Studies of the Health Effects of Particulate Matter Components. Research Report 177.* Boston, MA: Health Effects Institute.

U.S. EPA. 2006. “Chapter 4 : Air Quality Impacts”. In *Regulatory Impact Analysis, 2006 National Ambient Air Quality Standards for Particle Pollution.* Research Triangle Park, NC, USA.

— . 2003. *Compilation of Existing Studies on Source Apportionment for PM2.5.* Tech. rep. Washington, D.C.: Office of Air Quality Planning and Standards.

Wilson, J. G., S. Kingham, J. Pearce, and A. P. Sturman. 2005. “A review of intraurban variations in particulate air pollution: Implications for epidemiological research”. *Atmospheric Environment* 39 (34): 6444–6462.

Zanobetti, A., M. Franklin, P. Koutrakis, and J. Schwartz. 2009. “Fine particulate air pollution and its components in association with cause-specific emergency admissions.” *Environmental Health* 8:58.

Chapter 3

SELECTING SHRINKAGE PENALTY PARAMETERS FOR ESTIMATION

3.1 Introduction

In regression analyses, the accuracy of inference about an association between an exposure and outcome depends critically on the choice of adjustment variables. When clear information about causal relationships is available, it is now well-understood how this choice of adjustment impacts the large-sample consistency of the estimate (Hernan and Robins 2016). However, when we have small sample sizes and a large number of potential confounders, an approach that includes all variables may work poorly for parameter estimation. Consider the linear regression model

$$\boldsymbol{y} = \beta \boldsymbol{x} + \gamma_1 \boldsymbol{z}_1 + \cdots + \gamma_p \boldsymbol{z}_p + \boldsymbol{\epsilon}, \tag{3.1}$$

where $\boldsymbol{y} = \begin{bmatrix} y_1 & \ldots y_n \end{bmatrix}^T$ denotes our outcome of interest, $\boldsymbol{x} = \begin{bmatrix} x_1 & \ldots x_n \end{bmatrix}^T$ the exposure of interest, $\boldsymbol{Z} = \begin{bmatrix} \boldsymbol{z}_1 & \ldots & \boldsymbol{z}_p \end{bmatrix} \in \mathbb{R}^{n \times p}$ potential confounders determined from prior knowledge (e.g., following the approach of Pearl 2009), $\boldsymbol{\epsilon}$ mean zero variability, and n the sample size. Estimates of β are usually evaluated by their mean squared error (MSE), the sum of their squared bias and variance:

$$MSE(\hat{\beta}, \beta) = \mathrm{E}\left[\left(\hat{\beta} - \beta\right)^2\right] = \mathrm{E}\left[\left(\hat{\beta} - \mathrm{E}\left[\hat{\beta}\right]\right)^2\right] + \mathrm{Var}\left(\hat{\beta}\right)$$

Among linear unbiased estimators of the parameter vector $\begin{bmatrix} \beta & \gamma_1 & \ldots & \gamma_p \end{bmatrix}$, the fully-adjusted ordinary least squares (OLS) estimator $\hat{\beta}_{OLS} = (\boldsymbol{X}^T \boldsymbol{X})^{-1} \boldsymbol{X}^T \boldsymbol{y}$ (where $\boldsymbol{X} = \begin{bmatrix} \boldsymbol{x} & \boldsymbol{Z} \end{bmatrix}$) has minimum variance, and thus minimum MSE. However, strong correlations among the confounders may cause $\hat{\beta}_{OLS}$ to have high variance in small sample sizes and thus higher MSE than other estimators, which can out-perform $\hat{\beta}_{OLS}$ by having lower variances that compensate for their small but non-zero bias.

To pick a set of adjustment covariates when using OLS estimators, we can use model selection approaches. A wide variety of model selection methods are used in the epidemiology literature, including step-wise procedures, screening variables using significance tests or changes-in-estimate, and choosing models based upon information criteria (Greenland 2008; Greenland and Pearce 2015; Weng et al. 2009; Walter and Tiemeier 2009). Many of these procedures do not directly address questions of inference and can work poorly in practice (Greenland et al. 2016; Greenland and Pearce 2015). Moreover, model selection steps are often ignored when making inference in the final model, which can result in precision being over-stated.

3.1.1 Bayesian Estimators

When taking a Bayesian approach, inference on parameters is based on the posterior distribution, which is informed by both prior information about the parameters and the observed data. A Bayesian analogue to using model selection as a means for considering different parameter estimates is model averaging (MA), which allows simultaneous consideration of different subsets of confounders from a larger underlying model (Raftery et al. 1997). MA uses posterior model probabilities to average estimates from models with different confounders. But while MA is useful for prediction, it can perform poorly for effect estimation and the interpretation of averaged coefficients from different models can be difficult since MA averages together effect estimates with different causal interpretations (Crainiceanu et al. 2008). Bayesian Adjustment for Confounding (BAC) was developed to address these concerns about model averaging (Wang et al. 2012; Lefebvre et al. 2014a; Wang et al. 2015). Similar to MA, BAC averages estimates from a set of outcome models according to their posterior probability. BAC also fits a set of 'treatment' models (regression models with the exposure as the dependent variable) and links the inclusion of covariates in the treatment model to their inclusion in the outcome model in an effort to obtain an unconfounded estimate. This means that variables that are uninformative for the outcome y may be unnecessarily included in the health model due to their strong correlation with x.

3.1.2 Shrinkage Estimators

Shrinkage methods can provide improvements in MSE by trading off bias in the estimator for reductions in its variance. Here we consider shrinkage estimators based on ridge regression (Hoerl and Kennard 1970) and the LASSO (Tibshirani 1996). These estimators are solutions to the penalized optimization problem

$$(\hat{\beta}, \hat{\boldsymbol{\gamma}}) = \underset{(b,\boldsymbol{d})}{\arg\min} \left|\left|\boldsymbol{y} - \boldsymbol{x}b - \boldsymbol{Z}\boldsymbol{d}\right|\right|_2^2 + \lambda \sum_j |d_j|^\nu, \tag{3.2}$$

for $b \in \mathbb{R}$, $\boldsymbol{d} \in \mathbb{R}^p$ and where $\nu = 1$ for the LASSO and $\nu = 2$ for ridge regression. The second summation term penalizes large values of the coefficients d_j, so that the coefficients that solve (3.2) are 'shrunk' towards zero. The penalty parameter λ controls the amount of shrinkage. An important difference between the LASSO and ridge regression is that the former may shrink some coefficients to zero, and in this way perform a form of variable selection as well as estimation. Ridge regression may shrink coefficients very near to zero, but almost surely no ridge coefficeints are exactly zero and so no version of model selection occurs.

Unlike typical uses of ridge and LASSO, we do not penalize the coefficient of $\boldsymbol{x}$, which is the parameter of interest. By penalizing only the coefficients of the confounders $z_1, \ldots, z_p$, we select estimators of β that lie between the OLS estimate from the fully adjusted model (equivalent to $\lambda = 0$) and the OLS estimate from a completely unadjusted model (equivalent to $\lambda = \infty$). The ridge regression estimator can be written in closed form (see Appendix Section B.1.2) and while a closed-form expression for the LASSO estimator does not exist, it can be computed efficiently using, for instance, coordinate descent (Friedman et al. 2010).

Despite the potential improvements in finite-sample performance provided by shrinkage-estimator approaches, they are not widely used in epidemiological practice (Walter and

Tiemeier 2009), although a recent application of note is to use shrinkage estimators as an alternative to high-dimensional propensity score models (Franklin et al. 2015).

3.1.3 Penalty Parameter Selection

In order to apply these shrinkage estimators to a particular dataset, we must select a value for the penalty parameter λ. Selection of λ is frequently done by cross-validation (CV) for both ridge regression and LASSO (Bühlmann and Geer 2011). Although straightforward to implement with existing software such as R's glmnet package (Friedman et al. 2010), selection of λ by CV does not address the goal of parameter estimation, since it minimizes the error of observations around their conditional mean, and not the error in estimating β. Chichignoud et al. (2014) provide a recent alternative for choosing the LASSO penalty λ to minimize sup norm estimation error $||\hat{\boldsymbol{\beta}}_\lambda - \boldsymbol{\beta}||_\infty$, although their procedure still requires choosing a parameter that governs the magnitude of this error. For ridge regression, a variety of rules for directly selecting λ have been proposed (e.g. Golub et al. 1979; Khalaf and Shukur 2005; Wong and Chiu 2015; Draper and Nostrand 1979), although most are approximations to cross-validation and minimize error in predicting future outcomes y. This motivates the development of a new approach to selecting λ that addresses the primary goal of estimation of β.

3.2 Methods

We present a method for selecting the penalty parameter λ so that shrinkage estimators can be used for inference, with the goal that the shrinkage estimator will outperform the fully-adjusted OLS estimator in terms of MSE. To do this, we minimize a combination of the shrinkage estimator's bias and variance. In practice, the true model parameters are unknown and so the bias and variance of potential estimators cannot be calculated directly. We use the term 'future Bias' (fBias) and 'future Variance' (fVar) to refer to bias and variance computed on future similar datasets, which we will obtain as samples from the posterior predictive distribution. Methods exist for using the posterior predictive distribution for model-checking (Gelman et al. 1996) and model selection (Gelfand and Ghosh 1998; Hahn and Carvalho 2015) but these have not directly targeted estimation of a specific coefficient.

To select λ using Bayesian methods, we first specify the following hierarchical model with weakly informative priors:

$$\begin{aligned} \boldsymbol{y} &\sim N(\boldsymbol{x}\beta + \boldsymbol{Z}\boldsymbol{\gamma}, \sigma^2) \\ (\beta, \boldsymbol{\gamma})|\sigma^2 &\sim N\left(\boldsymbol{0}, \sigma^2\boldsymbol{V}_0\right) \\ \sigma^{-2} &\sim Gamma(a_0, b_0) \end{aligned}$$

In the simulations, we fix the hyperparameter values as $\boldsymbol{V}_0 = v_0\boldsymbol{I}_{p+1}$, where $v_0 = 1000$, and $(a_0, b_0) = (1, 5 \times 10^{-5})$, where the shape-rate parameterization of the gamma distribution is used. This choice of prior allows for simple forms of the posterior distributions due to conjugacy and allows the observed data to have large influence on the posterior. When relevant data are available, more informative priors could be introduced. Using the notation $\boldsymbol{X} = \begin{bmatrix}\boldsymbol{x} & \boldsymbol{Z}\end{bmatrix}$ for compactness, the posterior distributions corresponding to these priors are:

$$\begin{aligned} (\beta, \boldsymbol{\gamma})|\sigma^2; \boldsymbol{y}, \boldsymbol{X} &\sim N\left(\boldsymbol{V}\boldsymbol{X}^T\boldsymbol{y}, \sigma^2\boldsymbol{V}\right) && (3.3) \\ \sigma^{-2}|\boldsymbol{y}, \boldsymbol{X} &\sim Gamma\left(a_0 + n/2, b_0 + \frac{1}{2}(\boldsymbol{y}^T\boldsymbol{y} - \boldsymbol{y}^T\boldsymbol{X}\boldsymbol{V}\boldsymbol{X}^T\boldsymbol{y})\right), \end{aligned}$$

where $\boldsymbol{V} = (\boldsymbol{X}^T\boldsymbol{X} + \boldsymbol{V}_0^{-1})^{-1}$. Let $p(\boldsymbol{y}^*|\boldsymbol{y}, \boldsymbol{X}) = \int p(\beta, \boldsymbol{\gamma}|\sigma^2; \boldsymbol{y}, \boldsymbol{X})p(\sigma^2|\boldsymbol{y}, \boldsymbol{X})\pi(\beta, \boldsymbol{\gamma}, \sigma^2)d\beta\, d\boldsymbol{\gamma}\, d\,\sigma$ be the posterior predictive distribution for new observations $\boldsymbol{y}^*$ conditional on the observed data $(\boldsymbol{y}, \boldsymbol{X})$. We define fBias and fVar as

$$\begin{aligned} \text{fBias}(\hat{\beta}_\lambda; \beta, \boldsymbol{\gamma}, \sigma^2) &= \text{E}_{\boldsymbol{y}^*|\beta,\boldsymbol{\gamma},\sigma^2}\left(\hat{\beta}_\lambda\right) - \beta \\ \text{fVar}(\hat{\beta}_\lambda; \beta, \boldsymbol{\gamma}, \sigma^2) &= \text{Var}_{\boldsymbol{y}^*|\beta,\boldsymbol{\gamma},\sigma^2}\left(\hat{\beta}_\lambda\right). \end{aligned}$$

While a natural combination of fBias and fVar is their sum fMSE $=$ fBias^2 $+$ fVar, here we consider the maximum of fVar and fBias^2 as our loss function:

$$\begin{aligned} L(\lambda; \beta, \boldsymbol{\gamma}, \sigma^2) &= \max\left\{\text{fBias}(\hat{\beta}_\lambda; \beta, \boldsymbol{\gamma}, \sigma^2)^2, \text{fVar}(\hat{\beta}_\lambda; \beta, \boldsymbol{\gamma}, \sigma^2)\right\} \\ &= \text{fMBV}\left(\hat{\beta}_\lambda(\boldsymbol{y}^*); \beta, \boldsymbol{\gamma}, \sigma^2\right). && (3.4) \end{aligned}$$

Minimizing fMBV targets settings in which fBias^2 and fVar are equal and yields greater shrinkage than targeting fMSE. In the simulations below we present a comparison of the two loss functions, and demonstrate when each method performs best. The optimal value of λ, denoted by $\hat{\lambda}$, is chosen to be the value that minimizes the posterior risk:

$$\hat{\lambda} = \mathrm{E}_{\beta,\boldsymbol{\gamma},\sigma^2|\boldsymbol{y},\boldsymbol{X}}[fMBV(\hat{\beta}_\lambda(\boldsymbol{y}^*))].$$

We then evaluate the shrinkage estimator on the original data $(\boldsymbol{y}, \boldsymbol{X})$ using $\hat{\lambda}$ to obtain an estimate of β.

We refer to this estimator as Ridge-fMBV or LASSO-fMBV, depending upon which shrinkage approach is chosen. The value of $\hat{\lambda}$ and the amount of shrinkage is different for the two methods. The Ridge-fMBV and LASSO-fMBV estimators can be computed using an R package we have developed. Section B.1.1 of the Appendix provides the technical details of the algorithm for computing $\hat{\lambda}$. While the approach is conceptually the same for ridge and LASSO, one important computational difference is that the closed-form expression for the ridge estimator permits direct analytic computation of fMBV, while the LASSO requires simulating additional datasets from the posterior predictive distribution.

To conduct inference about β using the Ridge-fMBV and LASSO-fMBV estimators, we can compute both standard errors and confidence intervals for $\hat{\beta}_{\hat{\lambda}}$. For a standard error estimate that incorporates the variability of selecting $\hat{\lambda}$, we must consider how variable $\widehat{\beta}_{\hat{\lambda}}$ would be in replicate datasets. We again use the posterior predictive distribution to approximate future replicate datasets. For each $(\tilde{\beta}_j, \tilde{\boldsymbol{\gamma}}_j, \tilde{\sigma}^2_j)$ in the posterior distribution, we find the $\tilde{\lambda}_j$ that minimizes fMBV$(\hat{\beta}_\lambda, \tilde{\beta}_j)$ (note that here we do this for each vector in the posterior, rather than averaging over all vectors in the posterior as we do in computing the posterior risk). We then compute the expected value of the estimator $\widehat{\beta}_{\tilde{\lambda}_j}$ that corresponds to $\tilde{\lambda}_j$. The variance of these estimators across the posterior sample provides an estimate of the variance of $\widehat{\beta}_{\hat{\lambda}}$. For the LASSO, we follow the analogous procedure.

Because the shrinkage estimator is inherently biased, a confidence interval in standard form that is symmetric around the point estimate will not have correct coverage. To obtain an interval with coverage closer to the nominal rate, we 'invert the test' by ruling out values of β that are not likely to have led to the observed estimate. The algorithm for this procedure is provided in Appendix B.1.4.

The Ridge-fMBV and LASSO-fMBV estimators are designed to be advantageous in the context of small to moderate sample sizes. However, we can establish the large sample consistency of the estimators using existing theoretical results for shrinkage estimators. As the sample size increases, the optimal penalty approaches zero, meaning that the Ridge-fMBV and LASSO-fMBV approach $\hat{\beta}_{OLS}$ (which is asymptotically efficient). Additional discussion is provided in Appendix B.4.

3.3 Simulations

3.3.1 Simulation Setup

To demonstrate the behavior and benefit of the proposed Ridge-fMBV and LASSO-fMBV estimators, we present five simulations. We compare to multiple competing estimators: the OLS estimator with no adjustment for potential confounders (the 'unadjusted' estimator), the fully-adjusted OLS estimator, the OLS estimators corresponding to the models selected by minimizing Akaike Information Criteria (AIC; Akaike (1973)) and Bayesian Information Criteria (BIC; Schwarz (1978)) among all possible subsets of confounder combinations, using the LASSO with λ chosen by 10-fold CV (Hastie et al. 2009, Chap. 7), using ridge regression with λ chosen by 10-fold CV, using ridge regression with λ chosen by generalized cross-validation (GCV) (Golub et al. 1979), MA, and BAC. For selection of λ by CV, we do not penalize the coefficient of x to match the approach of the fMBV estimators. For BAC, we use the approximate procedure available in the BACprior R package (Lefebvre et al. 2014b). We also included a comparison with an 'oracle' ridge estimator that uses the (in practice unknown) true value of β to choose the value of λ that minimizes fMBV. Additionally, we present results for the Ridge-fMSE and LASSO-fMSE estimators, which use fMSE as the loss function in place of fMBV and are computed in the analogous manner.

For Simulations 1 through 3, data were generated by the linear model (3.1) with normal errors and $p = 6$ confounders. Each observation in the fixed design matrix was a single draw from a multivariate normal distribution: $(x, z_1, z_2, z_3, z_4, z_5, z_6)^T \sim N(0, \boldsymbol{W})$. The correlation matrix $\boldsymbol{W}$ was structured to have three pairs of correlated confounders (see Appendix Section B.2). The confounder effects were varied between simulations to reflect different scenarios:

- Simulation 1: Weak confounding effects, ranging from 0.05 to 0.15.
- Simulation 2: Strong confounding effects, ranging from 0.1 to 0.5.
- Simulation 3: Weak confounding effects, all equal to 0.05.

We then consider a more complex situation with $p = 12$ putative confounders for Simulations 4 and 5. The correlation matrix $\boldsymbol{W}$ was structured to have five blocks of correlated confounders. The confounder effects were again varied between simulations:

- Simulation 4: Moderate confounding effects, ranging in magnitude from 0.05 to 0.2, and null effects for two variables.
- Simulation 5: Weak confounding effects, ranging in magnitude from 0.05 to 0.1, but with no null effects.

In all of the primary simulations, the parameter of interest was set at $\beta = 1$, the sample size was $n = 100$, and the error variance is fixed at $\sigma^2 = 1$. For each simulation setup, the MSE of the estimators was evaluated using 1,000 replicate datasets. For computing $\hat{\lambda}$, we used a posterior sample of size 2,000 and (for LASSO) 500 draws from the posterior predictive distribution.

3.3.2 Results

The MSE and bias of the estimators for each simulation are provided in Table 3.1. In Simulation 1, the Ridge-fMBV and LASSO-fBMV estimators had the smallest MSE among all (non-oracle) estimators (MSE=0.060 and 0.059, respectively), providing substantial reduction in MSE compared to the fully-adjusted estimator (MSE=0.088). They performed markedly better than the Ridge-fMSE and LASSO-fMSE estimators (which had MSEs of 0.072 and 0.075, respectively), although the latter also out-performed the fully-adjusted estimator. Despite the similar MSE, the Ridge-fMBV estimator had almost half the bias of the LASSO-fMBV estimator (0.07 and 0.13, respectively). Selecting λ by CV for both LASSO and Ridge resulted in estimators almost identical to the unadjusted estimate. The AIC and BIC approaches, which do not penalize coefficients but may not include all variables in the model, do slightly better (MSE=0.079 and 0.085, respectively) than full adjustment, but do not achieve as small an MSE as the shrinkage estimators. BAC achieves similar MSE to the fully-adjusted model. Figure 3.1 presents the MSE for each estimator in Simulation 1 in terms of squared bias and variance. This highlights the tradeoff of bias for variance that Ridge-fMBV and LASSO-fMBV make to achieve lower MSE.

In Figure 3.1 we see that in Simulation 1, some points on the oracle MSE curve (solid black line) have lower MSE than Ridge-fMBV. This is reflected in Table 1, where the oracle ridge estimator has lower MSE than all other estimators. The difference between the Ridge-fMBV choice of penalty and the oracle penalty can be explained in part by the (frequentist) uncertainty with which we know the underlying parameter values, and hence how well we estimate the bias and variance in replicate experiments. Figure 3.2 shows the squared bias and estimated variance for adjusted and unadjusted estimators for a subset of the datasets in Simulation 1. For each dataset, we construct a curve in this Variance-Bias^2 space corresponding to the range of possible ridge estimators (by varying λ). These curves connect the points corresponding to the adjusted and unadjusted estimates, which are the limiting cases of ridge regression. The non-zero bias for each individual fully-adjusted estimate means that the point of smallest MSE on each curve tends to lie further to the right (i.e. smaller λ, meaning higher variance and lower bias) than the oracle minimum MSE on the black curve.

Within a particular dataset, there is additional uncertainty about fBias^2 and fVar with respect to the posterior distribution. Figure 3.3 illustrates this uncertainty, by giving for a

Table 3.1: MSE and bias for estimators of β. Simulations 1 through 3 have the same design matrix containing 6 confounders. Simulations 4 and 5 have a different design matrix containing 12 confounders. For all simulations, $n = 100$ and $\beta = 1$. See text for difference in confounder effects.

Method	MSE ($\times 10^{-1}$)					Bias				
Simulation	1	2	3	4	5	1	2	3	4	5
Unadjusted	0.92	8.76	0.32	2.03	0.76	0.29	0.93	0.15	0.44	0.26
Fully Adjusted	0.88	0.88	0.88	0.99	0.99	−0.02	−0.02	−0.02	−0.003	−0.003
All subsets AIC	0.79	1.24	0.52	0.97	0.78	0.13	0.12	0.08	0.05	0.17
All subsets BIC	0.85	2.04	0.36	0.94	0.77	0.24	0.33	0.13	0.13	0.24
MA	0.92	7.50	0.32	1.33	0.75	0.29	0.85	0.15	0.13	0.26
BAC	0.88	0.88	0.88	0.93	0.85	−0.02	−0.02	−0.02	0.03	0.03
LASSO-fMBV	0.59	1.14	0.41	0.67	0.62	0.13	0.19	0.06	0.10	0.12
LASSO-fMSE	0.75	1.02	0.60	0.81	0.77	0.05	0.03	0.03	0.05	0.07
LASSO-CV	0.92	5.99	0.32	1.27	0.75	0.29	0.75	0.15	0.33	0.26
Ridge-fMBV	0.60	1.02	0.50	0.75	0.66	0.07	0.16	0.03	0.10	0.08
Ridge-fMSE	0.72	0.97	0.62	0.86	0.79	0.04	0.05	0.06	0.06	0.05
Ridge-CV	0.92	7.00	0.32	1.49	0.75	0.29	0.82	0.15	0.37	0.26
Ridge-GCV	0.87	0.90	0.66	0.79	0.86	0.08	0.08	0.05	0.05	0.04
Ridge-Oracle	0.50	0.81	0.27	0.61	0.50	−0.13	−0.02	−0.21	−0.10	−0.14

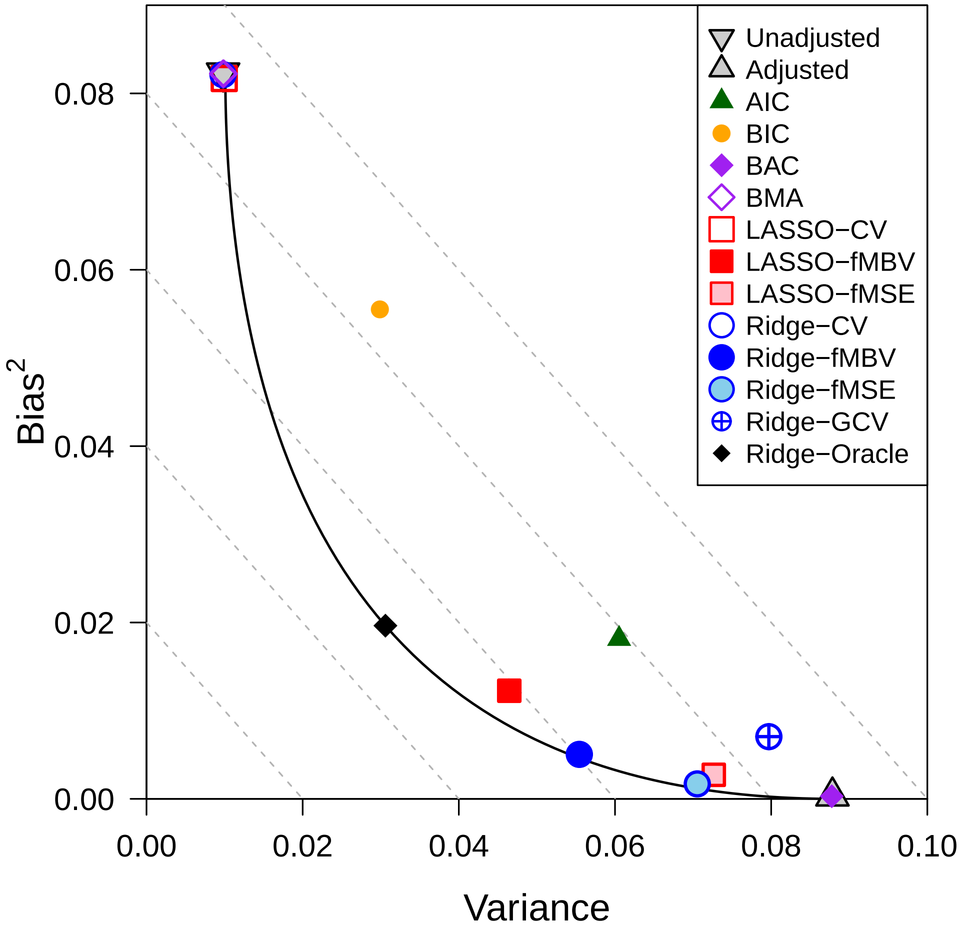


Figure 3.1: Variance and squared bias for estimators in Simulation 1. The dashed lines represent contours of equal MSE. The solid black curve represents the theoretical MSE for ridge estimators, for varying values of λ.

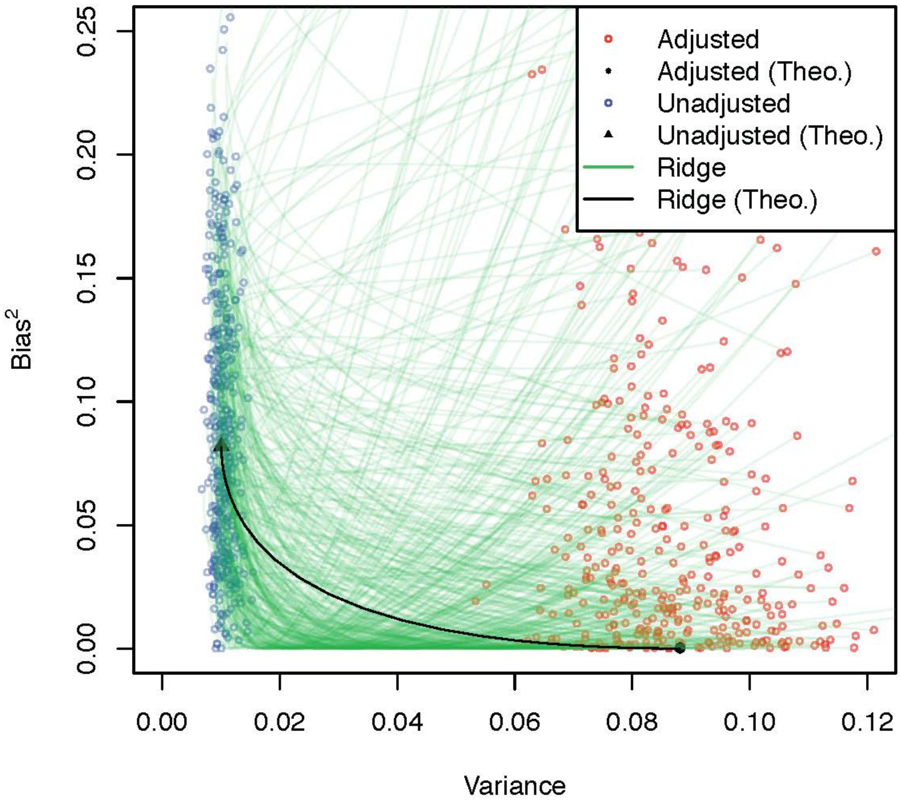


Figure 3.2: Squared bias and estimated variance of the unadjusted and adjusted estimates of β for 400 of the simulated datasets in Simulation 1. The green curves show the paths taken by ridge estimators, when varying λ. The black points and lines show the theoretical values.

single dataset the posterior distribution of fBias^2 and fVar for the adjusted and unadjusted estimates, and ridge estimators with different λ. The distributions are mostly symmetric with respect to fVar but are skewed towards higher values of fBias^2 (other than the adjusted estimate, which is unbiased relative to the posterior mean). The impact of this skewness on the posterior distributions of fMBV and fMSE can be seen in Figure 3.4. Because fMBV targets the setting when fBias^2 and fVar are equal, it favors greater shrinkage. On the other hand, fMSE targets the setting when the sum of fVar and fBias^2 is smallest, which occurs here when the fBias^2 is much smaller than fVar and thus yields less shrinkage. Figure 3.5 compares the optimal penalty parameters selected under each loss function. As a sensitivity analysis, Table B.1 provides the MSE for the Ridge-fMSE and Ridge-fMBV estimators for sample sizes up to $n = 10,000$. For large enough n, we see that the MSE of Ridge-fMSE eventually falls below that of Ridge-fMBV. In these settings with large n, the additional shrinkage provided by the choice of fMBV loss, which helped reduce MSE in small samples, yields greater bias and higher MSE.

In Simulation 2, the MSE for the Ridge-fMBV estimator is 0.102 and for LASSO-fMBV is 0.114, which are both worse than the fully adjusted estimator's MSE (0.088). This is expected; the large effects in this simulation mean that shrinking the coefficients introduces non-trivial bias. Notably, choosing λ by CV in this setting gives MSEs of 0.700 for ridge and 0.599 for LASSO, both far worse than selecting λ by minimizing fMBV. The fMBV estimators also outperform using OLS estimators for models selected by AIC or BIC. Here, the LASSO-fMSE and Ridge-fMSE estimators outperform their fMBV counterparts, with MSEs of 0.102 and 0.097, respectively, and much smaller bias (0.03 and 0.05 compared to 0.19 and 0.15, respectively). This is consistent with less shrinkage being favorable when the confounding effects are large. Figure B.1 in the Appendix provides a plot of the variance and squared bias of each approach.

In Simulation 3, the small effect sizes, compared to the first two settings, mean that the confounding bias in the unadjusted estimator is relatively small (0.15). The MSEs for the Ridge-fMBV and LASSO-fMBV (0.050 and 0.041, respectively) are higher than the

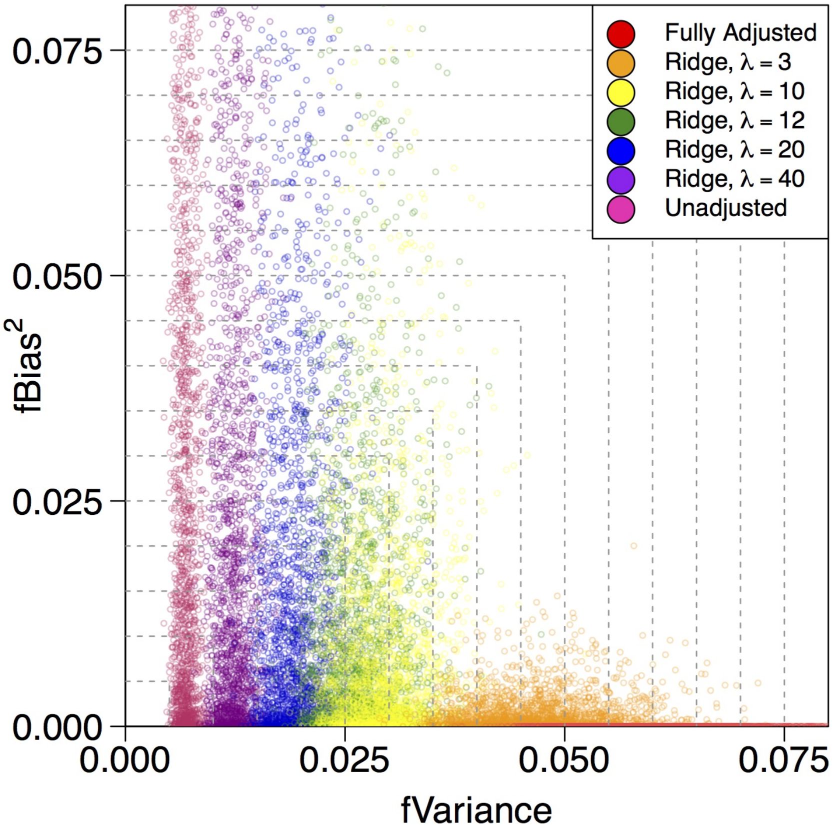
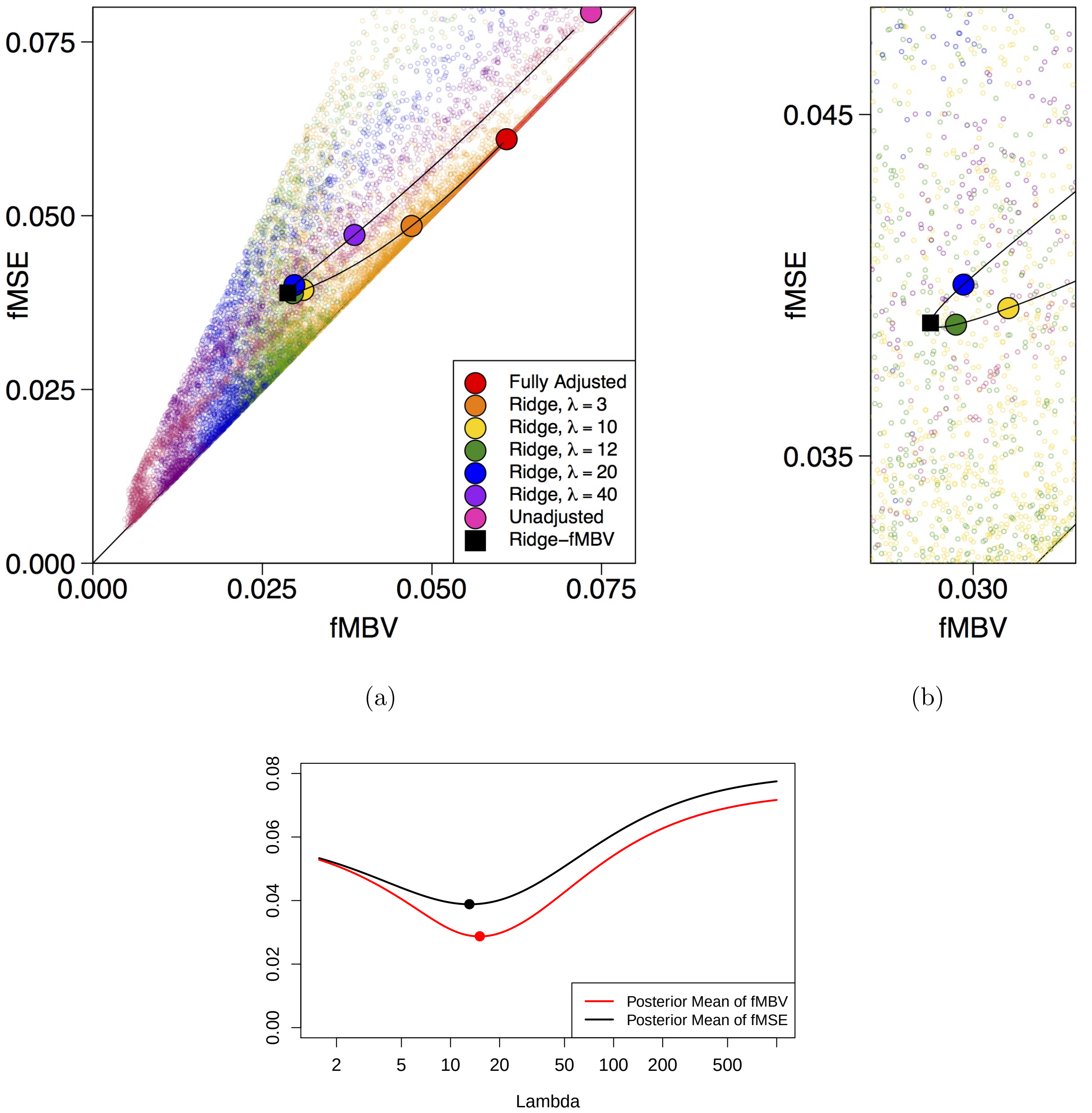


Figure 3.3: Posterior distribution of the squared bias and variance for estimators from a single data set in Simulation 1. Clouds of points represent samples from the posterior distribution of fVar and fBias^2 for particular values of λ.



(a)

(b)

(c)

Figure 3.4: Posterior distributions of the fMBV and fMSE from a single data set in Simulation 1. Clouds of colored points in (a) represent the posterior distributions of fMBV and fMSE for particular values of λ. The large circles indicate the posterior mean of fMBV and fMSE for each λ value. Sub-figure (b) shows a zoomed-in portion of the full figure in (a). Sub-figure (c) shows the posterior means of fMBV and fMSE as a function of λ.

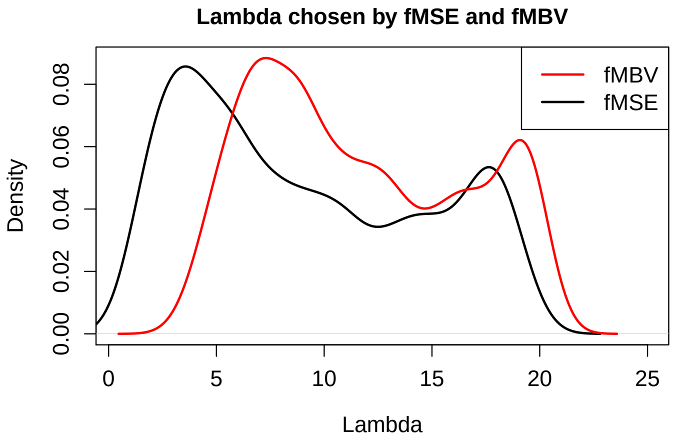


Figure 3.5: Density plot of λ values selected using fMBV and fMSE loss in Simulation 1.

unadjusted estimate (MSE=0.032), but are still a more than 20% reduction compared to the adjusted estimate (MSE=0.088). In this setting where shrinkage improves performance, the estimators based upon fMSE loss perform markedly worse than those based upon fMBV loss. The LASSO-CV and Ridge-CV estimates select a large amount of penalization and achieve the same results as the unadjusted estimate (MSE= 0.032). Figure B.2 shows the squared bias and variance results graphically.

Table 3.2 provides the true and estimated standard errors for Ridge-fMBV, LASSO-fMBV, and their fMSE counterparts in each Simulation. For all, the estimated standard error tends to be slightly below the true value, although this difference disappears for larger samples (see Appendix Table B.2). Nominal 95% confidence intervals achieve correct coverage, although they are slightly wider than the corresponding confidence intervals for the fully-adjusted OLS estimators (Table 3.3).

In Simulations 4 and 5, the LASSO-fMBV and Ridge-fMBV estimators outperform all of the competing estimators in terms of MSE (see Table 3.1 and Figure 3.6). In Simulation 4, the MSE of LASSO-fMBV is 0.067, 33% lower than that of the fully-adjusted OLS estimate (0.099), while Ridge-fMBV has a relative reduction of 25% in MSE (0.075). The AIC,

Table 3.2: Standard error estimates for the LASSO and Ridge estimators in the simulations.

		Simulation				
Estimator		1	2	3	4	5
LASSO-fMBV	Estimated SE	0.212	0.264	0.201	0.221	0.206
	True SE	0.216	0.275	0.194	0.241	0.220
LASSO-fMSE	Estimated SE	0.262	0.300	0.252	0.261	0.254
	True SE	0.270	0.318	0.244	0.279	0.270
Ridge-fMBV	Estimated SE	0.214	0.264	0.204	224	0.210
	True SE	0.236	0.279	0.222	0.256	0.246
Ridge-fMSE	Estimated SE	0.253	0.296	0.241	0.260	0.249
	True SE	0.266	0.308	0.248	0.288	0.277

Table 3.3: Coverage rate and average width for nominal 95% Confidence Intervals for the fully-adjusted OLS estimator and the Ridge-fMBV estimator.

Simulation	1	2	3	4	5
Coverage					
Fully-Adjusted OLS	0.946	0.946	0.946	0.939	0.939
Ridge-fMBV	0.946	0.945	0.946	0.931	0.932
Width					
Fully-adjusted OLS	1.179	1.179	1.179	1.207	1.207
Ridge-fMBV	1.199	1.202	1.200	1.152	1.152

BIC, and CV approaches all have higher MSE than the fully-adjusted OLS estimator. In Simulation 5, like Simulation 3, the small γ values mean that the confounding bias is small and the unadjusted estimator achieves lower MSE (0.076) than the fully-adjusted estimator (0.099). But both are outperformed by the Ridge-fMBV and LASSO-fMBV estimators, which have MSE of 0.066 and 0.062, respectively. In these two settings, the coverage of the confidence intervals is slightly below the nominal level (0.931), although similar to coverage from confidence intervals for the fully-adjusted estimator (0.939).

3.4 Cardiovascular Outcomes in MESA

To demonstrate the estimators using data from an epidemiological study, we present an analysis of the association between smoking status and carotid intima-media thickness (cIMT) in the Multi-Ethnic Study of Atherosclerosis (MESA). The diverse MESA cohort comprises adults from six U.S. metropolitan areas, aged 45 to 84 and free of clinical cardiovascular disease at study entry (Bild et al. 2002). MESA was designed to study subclinical cardiovascular disease, and measurements of cIMT were made at baseline. MESA is an attractive cohort for this analysis because it includes multiple ethnic sub-cohorts of different sizes, which allow us to compare our estimators in different sample sizes. Lefebvre et al. (2014a) recently analyzed cIMT[1] in MESA to demonstrate the Bayesian Adjustment for Confounding (BAC) method, and we use the same data for our comparison.

3.4.1 Analysis

Following Lefebvre et al. (2014a), we consider the effect of having ever been a smoker (>100 cigarettes in lifetime) on baseline cIMT, measured by ultrasonography. We perform separate analyses on two sub-cohorts of the MESA cohort: Caucasians under the age of 65 and Chinese-Americans under the age of 65. Measured variables at baseline that we consider as

[1] Although described by Lefebvre et al. (2014a) as common carotid artery intima-media thickness, the measures reported by those authors correspond to internal carotid artery intima-media thickness, which is what we use for this analysis.

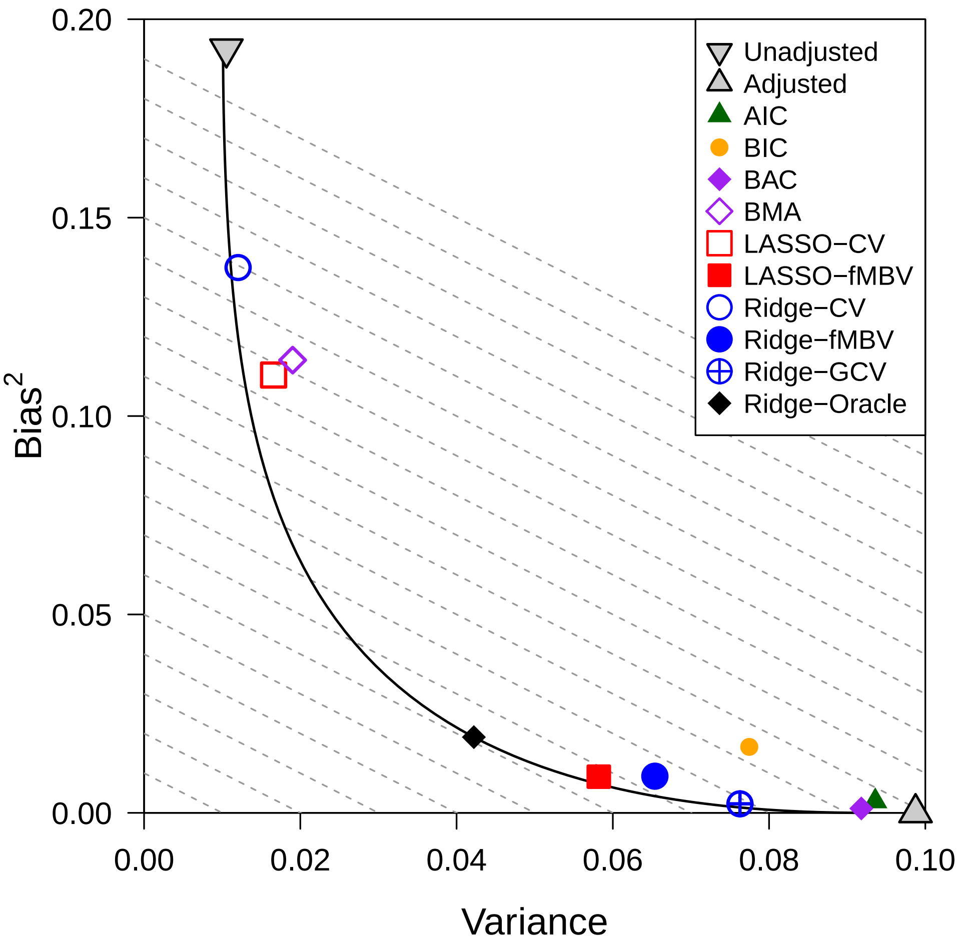


Figure 3.6: Variance and squared bias for estimators in Simulation 4. The dashed lines represent contours of equal MSE. The solid black curve represents the theoretical MSE for ridge estimators, as λ varies.

potential confounders include age, sex, body mass index, physical activity, cholesterol levels (total and high-density lipoprotein), triglycerides, inflammatory marker levels (interleukin 6 [IL6], C-reactive protein [CRP]), diabetes, use of diabetes and lipid lowering medications, hemostatic marker (fibrinogen) levels, alcohol consumption, education, and income. Table 3.4 provides summary statistics for the two sub-cohorts, stratified by smoking status.

We compare the Ridge-fMBV and LASSO-fMBV estimators to the fully-adjusted and unadjusted linear regression results, the BAC and MA estimates reported by Lefebvre et al. (2014a), and the posterior mean from a standard Bayesian linear regression analysis with no shrinkage method applied. For selecting λ by minimizing fMBV, we use the weakly informative priors of Raftery et al. (1997), which were also used for the BAC analysis of Lefebvre et al. (2014a). Specifically, the regression coefficients have a mean-zero prior with block-diagonal variance structure. For the exposure $\boldsymbol{x}$ and the continuous and binary covariates $\boldsymbol{z}_1, \ldots, \boldsymbol{z}_{p'}$, the prior variances are independent with values $\phi^2 s_x^{-2}, \phi^2 s_1^{-2}, \ldots, \phi^2 s_{p'}^{-2}$, where s_x^2 and s_j^2 are the sample variances of $\boldsymbol{x}$ and $\boldsymbol{z}_j$, respectively. For categorical variables $\boldsymbol{z}_{p'+1}, \ldots, \boldsymbol{z}_p$ with $c_j > 2$ classes, the prior variances are $c_j \times c_j$ blocks equal to $\phi^2(\boldsymbol{z}_j^T \boldsymbol{z}_j/n)^{-1}$, where $\tilde{\boldsymbol{z}}_j$ is the centered $n \times c_j$ design matrix for dummy-variable coding of $\boldsymbol{z}_j$. The hyperparameters are set to $a_0 = 2.58/2$, $b_0 = (2.58)(0.28)/2$, and $\phi = 2.85$. For computing the posterior expectation of fMBV we drew samples of size 3,000. The categorical variables were coded according to a sum-to-zero constraint and alcohol consumption was log-transformed. The design matrix was standardized prior to applying ridge and LASSO, but the reported estimates are back-transformed to the original scale.

3.4.2 Results

Of the 1,378 participants in the Caucasian sub-cohort with complete covariate information, 774 were smokers and 604 were non-smokers (see Table 3.4). The estimated association between smoking status and cIMT is reported in Table 3.5 for each approach. The unadjusted difference in cIMT between smokers and non-smokers is 88.78μm (standard error [SE]: 25.04). The fully-adjusted estimate is 48.61μm (SE: 24.48). The Ridge-fMBV and LASSO-fMBV

Table 3.4: Descriptive statistics for the Causcasian and Chinese-American sub-cohorts of MESA. Values are mean (sd) or count (%).

	Caucasian Sub-cohort		Chinese-American Sub-cohort	
	Ever Smoker	Non-smoker	Ever Smoker	Non-smoker
	(n=774)	(n=604)	(n=93)	(n=343)
cIMT (um)	993.0 (472.7)	904.2 (445.9)	839.78 (335.3)	763.3 (326.7)
Age (years)	54.86 (5.56)	54.39 (5.65)	54.61 (5.26)	54.45 (5.85)
Male	373 (48.2)	275 (45.5)	84 (90.3)	126 (50.4)
Body mass index (kg/m2)	28.0 (5.48)	28.1 (5.55)	24.8 (3.09)	24.0 (3.27)
Physical Activity (Metabolic Equivalent of Task-hours/week)	2440 (2625)	2697 (2882)	1609 (1827)	1792 (2241)
Total Cholesterol (mg/dl)	196.0 (36.4)	198.8 (36.6)	194.6 (30.9)	194.6 (31.5)
High-density Lipoprotein Cholesterol (mg/dl)	52.20 (16.27)	51.92 (14.98)	44.49 (9.50)	50.39 (13.74)
Triglycerides (mg/dl)	135.7 (91.4)	135.7 (118.5)	156.5 (86.1)	141.8 (85.7)
Interleukin-6 (pg/ml)	1.40 (1.25)	1.23 (1.13)	1.06 (0.81)	1.06 (1.01)
C-reactive Protein (mg/l)	3.25 (4.39)	3.42 (6.10)	1.66 (2.89)	1.89 (4.42)
Fibrinogen (mg/dl)	323.4 (68.1)	325.6 (66.9)	311.5 (54.5)	325.1 (60.0)
Diabetes	38 (4.9)	25 (4.1)	11 (11.8)	31 (9.0)
Use of antidiabetic medications	27 (3.5)	19 (3.1)	8 (8.6)	21 (6.1)
Use of lipid lowering medications	115 (14.9)	84 (13.9)	10 (10.8)	26 (7.6)
Alcohol consumption (drinks/week)	7.21 (10.81)	3.83 (7.87)	1.52 (3.59)	0.57 (2.54)
Education completed				
Less than high school	26 (3.4)	11 (1.8)	16 (17.2)	61 (17.8)
High school	375 (48.4)	198 (32.8)	33 (35.5)	128 (37.3)
College	172 (22.2)	166 (27.5)	25 (26.9)	81 (23.6)
Graduate School	201 (26.0)	229 (37.9)	19 (20.4)	73 (21.3)
Income				
<$25,000	82 (10.6)	51 (8.4)	29 (31.2)	123 (35.9)
$25,000 – $50,000	179 (23.1)	139 (23.0)	27 (29.0)	84 (24.5)
$50,000 – $100,000	296 (38.2)	214 (35.4)	23 (24.7)	84 (24.5)
>$100,000	217 (28.0)	200 (33.1)	14 (15.1)	52 (15.2)

Table 3.5: Estimated difference in cIMT (in μm) between smokers and non-smokers in the Caucasian sub-cohort.

Method	Estimate	Standard Error	95% CI
Unadjusted	88.78	25.04	(39.7, 137.9)
Fully-Adjusted	48.61	24.48	(0.58, 96.6)
Posterior Mean	48.60	24.31	(0.91, 96.2)
BAC (Lefebvre *et al.*)	49.66	24.45	(1.74, 97.6)
Ridge-fMBV	64.19	24.23	(2.01, 97.6)
LASSO-fMBV	68.63	24.19	(1.81, 98.7)

estimates of 64.19μm and 68.63 μm, respectively, were between the unadjusted and adjusted estimates. With the large size of the Caucasian cohort, the shrinkage estimators have only a slight reduction in estimated standard error (24.19 for LASSO-fMBV and 24.23 for Ridge-fMBV). The posterior mean estimate was nearly identical to the fully-adjusted estimate, which is unsurprising given the large sample size and uninformative priors. Figure 3.7 shows the relative values of the coefficients for the fMBV estimators.

The LASSO-fMBV estimate shrunk the coefficients for physical activity, CRP, fibrinogen, diabetes medication, college education, income above $25,000, and alcohol use to zero (see Figure 3.7). This contrasts with the most probable model under BAC, which left out BMI, triglycerides, CRP, fibrinogen, diabetes (diagnosis and medication use), and income (see Table 3.6). Examination of the correlation between the potential confounders, cIMT, and smoking status gives some insight into this different selection of confounders, between the methods. Alcohol use is correlated with smoking status ($r = 0.22$) but not with cIMT ($r = 0.05$). Because BAC forces all variables in the exposure model to be included in the outcome model, the correlation between alcohol use and smoking status leads to its inclusion in the most probable BAC model for cIMT. Our shrinkage estimator approach instead only

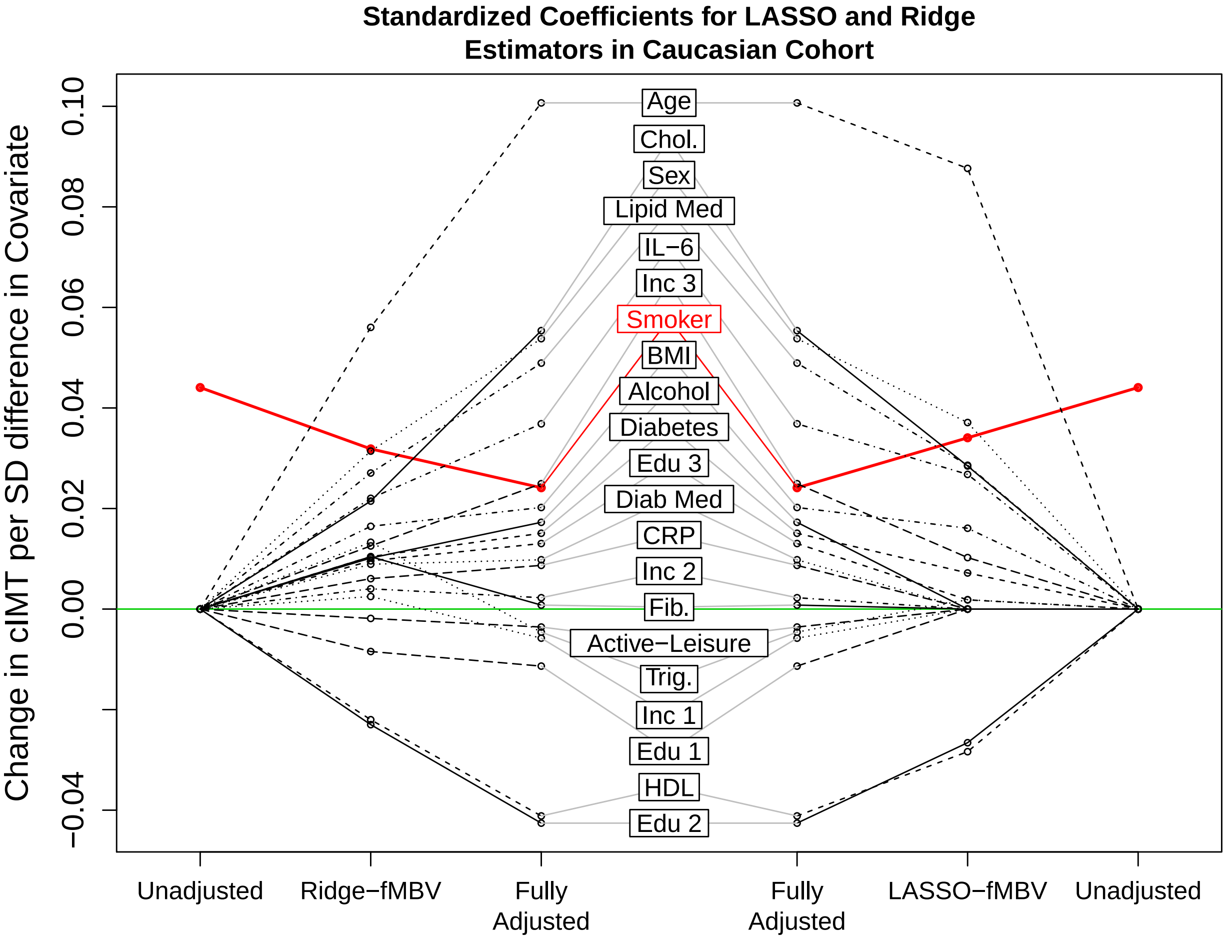


Figure 3.7: Coefficient estimates from the analysis of cIMT in the Caucasian sub-cohort of MESA. Estimates are plotted on standardized scale.

considers a model with cIMT as the outcome, and the relatively weak correlation between alcohol use and cIMT means that it is one of the first variables removed; it does not appear to have the large effect size required for it to introduce a large confounding bias. Conversely, both BMI and triglyceride levels are slightly correlated with cIMT ($r = 0.14$ and $r = 0.13$) but not with smoking status ($r = 0.00$ for both). While the most probable BAC model included neither, both had non-zero estimated coefficients in the LASSO-fMBV estimate. A full correlation matrix for the potential confounders, cIMT, and smoking status is provided in Appendix Table B.3.

In the Chinese-American subcohort, 93 of the 436 participants were smokers. The crude smoking effect on cIMT was 76.40 μm (SE: 38.41), while the fully adjusted estimate provided no evidence of an association: -13.66 μm (SE: 40.69) (see Table 3.7). The Ridge-fMBV and LASSO-fMBV estimates were small but positive (12.77 and 26.12, respectively) and had smaller standard errors than the fully-adjusted results (36.35 and 39.57, respectively). Figure 3.8 shows the standardized coefficients. Notably, diabetes has a larger impact on cIMT in this cohort compared to the Caucasian cohort, and the direct effect of age is less than in the Caucasian dataset.

As a sensitivity analysis, we re-analyzed the Chinese-American cohort, using the posterior distribution from the Caucasian analysis as the prior distribution for selecting $\hat{\lambda}$. Because the smoking effect in Caucasians was strongly positive, this sensitivity analysis yielded estimates that were larger than analyzing the Chinese-American data alone: 28.65μm for Ridge-fMBV and 43.94μm for LASSO-fMBV. As a second sensitivity analysis, we multiplied the prior variance by four, to represent skepticism about the similarity of underlying effects between these two populations, and the results were similar.

3.5 Discussion

The selection of penalty parameters is a key step in the application of shrinkage methods for effect estimation, and here we have presented a principled approach to this problem. Our method, based upon minimizing the bias and variance of the estimator in future similar

Table 3.6: Variables included in the LASSO-fMBV estimator and the top three BAC models for the Caucasian subcohort. Correlation with the outcome (cIMT) and exposure (Smoking Status) are also provided. Variable inclusion in the BAC models is based upon the results of Lefebvre et al. (2014a).

								Lipid		Diab.			
Model	Age	Sex	BMI	Chol.	HDL	Trig.	IL6	Med.	Diabetes	Med.	Income	Education	Alcohol
LASSO-fMBV	✓	✓	✓	✓	✓	✓	✓	✓	✓		✓	✓	
BAC-1	✓	✓		✓	✓		✓	✓				✓	✓
BAC-2	✓	✓		✓	✓		✓	✓	✓			✓	✓
BAC-3	✓	✓		✓	✓		✓	✓		✓		✓	✓
Correlation:													
cIMT	0.25	0.14	0.14	0.08	−0.14	0.13	0.15	0.13	0.10	0.08	0.05-0.09	−0.13 - 0.08	0.05
Smoking	0.04	0.03	0.00	−0.04	0.01	0.00	0.07	0.01	0.02	0.01	0.04-0.06	−0.13 - 0.13	0.22

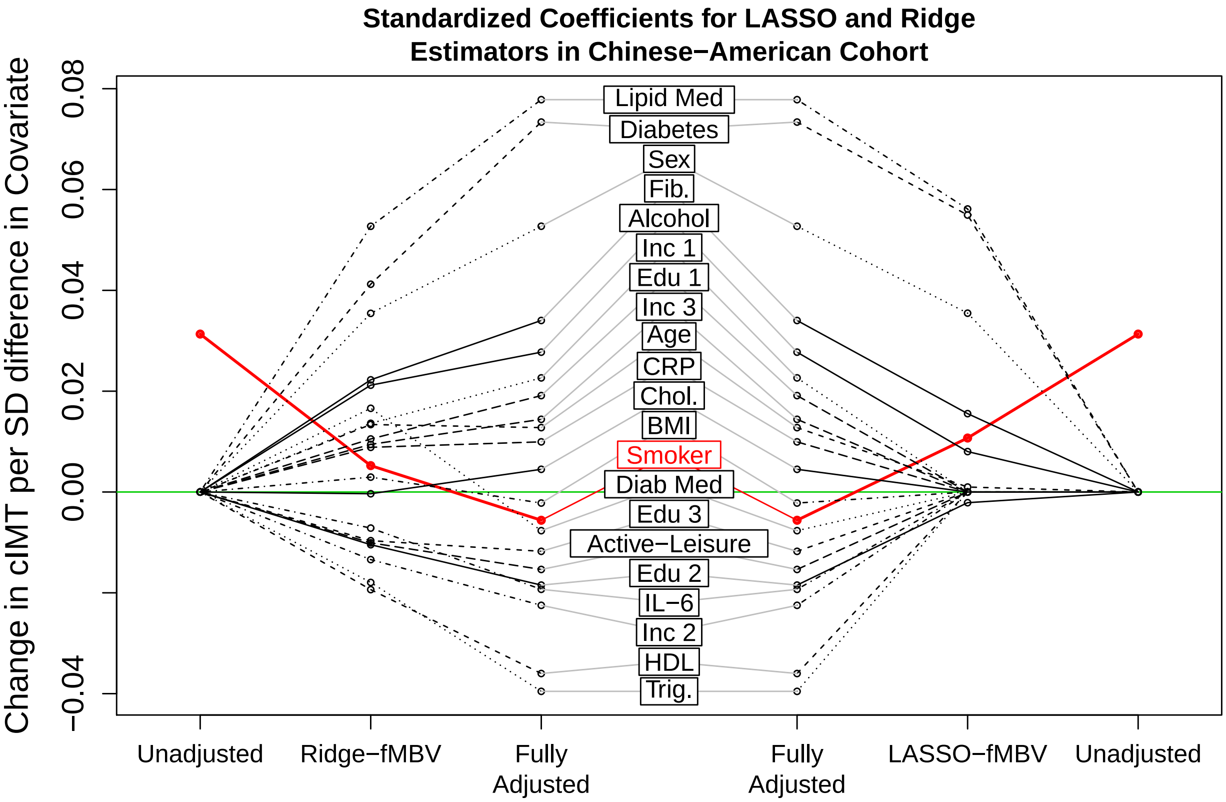


Figure 3.8: Coefficient estimates from the analysis of cIMT in the Chinese-American subcohort of MESA. Estimates are displayed on the standardized scale.

Table 3.7: Estimated difference in cIMT (in μm) between smokers and non-smokers in the Chinese-American sub-cohort.

Model	Estimate	Standard Error	95% CI
Unadjusted	76.40	38.41	(0.90, 151.9)
Fully-Adjusted	−13.66	40.69	(−93.6, 66.3)
Posterior Mean	−13.56	39.97	(−91.6, 64.9)
BAC (Lefebvre *et al.*)	2.60	40.10	(−76.0, 81.2)
Ridge-fMBV	12.77	36.35	(−91.6, 65.9)
LASSO-fMBV	26.12	39.57	(−79.7, 59.7)

datasets, directly addresses performance in estimating β. Through simulations, we demonstrated that when there are many correlated putative confounders that have small (possibly null) effects and the sample size is small, choosing λ by minimizing fMBV reduces the error in estimation of β. In most of the scenarios considered here, using fMBV as a loss function outperformed using fMSE, although the latter did better in settings with larger sample sizes.

This procedure should not used in place of applying *a priori* scientific knowledge. If there are known confounders with strong effects, those variables can be included in the regression model without penalty. But this method can be used to improve estimation in the context of less well understood variable relationships.

The estimate we obtain is inherently biased, however the reduced variability makes it useful for inference in a data analysis with a small sample size. Furthermore, our estimator provides information about variable importance through examination of which variables either drop out of or are heavily penalized in the final model. Although model selection is not our primary goal, this information can be useful for further exploration of causal relationships.

These estimators can be computed simply using an R package we have developed. However, the LASSO version can be slow to compute due to the large number of repeated optimizations required, especially for computing confidence intervals. Nonetheless, this approach can be easily parallelized and remains computationally simpler than methods such as MCMC. As an alternative approach, we explored an importance-sampling type algorithm to expedite this procedure. While passable in some scenarios, we found the high variability of the sampling weights made the estimates unstable, leading to poorer performance despite computational savings.

The choice between the ridge and LASSO estimators can be tailored to each problem. The LASSO estimator provides variable selection in addition to a point estimate, which may be beneficial in determining which putative confounders should be included in future models. The estimates produced by both methods for the parameter of interest tend to be similar, and so the computation simplicity of ridge may be preferred if conclusions will be based on this parameter alone.

In the MESA analysis, our estimators gave point estimates that were notably higher than the fully-adjusted estimates and the BAC estimate. The standard errors of our estimators were slightly less in the Caucasian cohort, but the difference was much smaller than the change in point estimate. In the smaller Chinese-American cohort, the standard error of the Ridge-fMBV estimator showed a notable decrease of about 10%, but this did not substantively affect inference. Compared to the simulations, the correlation of the MESA variables was quite weak, which is likely the reason for the only slight reductions in standard errors.

Overall, the estimators we present for applying shrinkage estimators to the problem of effect estimation provide important benefits for small-sample regression analysis.

3.6 References

Akaike, H. 1973. "Information theory and an extension of the maximum likelihood principle". In *Second International Symposium on Information Theory*, ed. by B. Petrov and F. Csaki, 267–281. Budapest: Akademiai Kiado.

Bild, D. E., D. A. Bluemke, G. L. Burke, et al. 2002. "Multi-Ethnic Study of Atherosclerosis: Objectives and design". *American Journal of Epidemiology* 156 (9): 871–881.

Bühlmann, P., and S. van de Geer. 2011. *Statistics for High-Dimensional Data*. New York: Springer.

Chichignoud, M., J. Lederer, and M. J. Wainwright. 2014. "Tuning Lasso for sup-norm optimality": 1–15. arXiv: `arXiv:1410.0247v1`.

Crainiceanu, C. M., F. Dominici, and G. Parmigiani. 2008. "Adjustment uncertainty in effect estimation". *Biometrika* 95 (3): 635–651.

Draper, N. R., and R. C. V. Nostrand. 1979. "Ridge Regression and James-Stein Estimation: Review and Comments". *Technometrics* 21 (4): 451–466.

Franklin, J. M., W. Eddings, R. J. Glynn, and S. Schneeweiss. 2015. "Regularized Regression Versus the High-Dimensional Propensity Score for Confounding Adjustment in Secondary Database Analyses". *American Journal of Epidemiology* 182 (7): 651–659.

Friedman, J., T. Hastie, and R. Tibshirani. 2010. "Regularization Paths for Generalized Linear Models via Coordinate Descent". *Journal of Statistical SoftwareR* 30 (1): 1–22.

Gelfand, A. E., and S. K. Ghosh. 1998. "Model choice: A minimum posterior predictive loss approach". *Biometrika* 85 (1): 1–11.

Gelman, A., X. Meng, and H. Stern. 1996. "Posterior predictive assessment of model fitness via realized discrepancies". *Statistica Sinica* 6:733–807.

Golub, G. H., M. Heath, and G. Wahba. 1979. "Generalized Cross-Validation as a Method for Choosing a Good Ridge Parameter". *Technometrics* 21 (2): 215–223.

Greenland, S. 2008. “Invited commentary: Variable selection versus shrinkage in the control of multiple confounders”. *American Journal of Epidemiology* 167 (5): 523–529.

Greenland, S., and N. Pearce. 2015. “Statistical Foundations for Model-Based Adjustments”. *Annual Review of Public Health* 36 (1): 89–108.

Greenland, S., R. Daniel, and N. Pearce. 2016. “Outcome modelling strategies in epidemiology : traditional methods and basic alternatives”. *International journal of epidemiology*, no. April: in press.

Hahn, P. R., and C. M. Carvalho. 2015. “Decoupling Shrinkage and Selection in Bayesian Linear Models: A Posterior Summary Perspective”. *Journal of the American Statistical Association* 110 (509): 435–448.

Hastie, T., R. Tibshirani, and J. Friedman. 2009. *The Elements of Statistical Learning.* 2nd. New York, NY: Springer.

Hernan, M., and J. Robins. 2016. *Causal Inference.* Boca Raton: Chapman & Hall/CRC.

Hoerl, A. E., and R. W. Kennard. 1970. “Ridge Regression: Biased Estimation for Nonorthogonal Problems”. *Technometrics* 12 (1): 55–67.

Khalaf, G., and G. Shukur. 2005. “Choosing Ridge Parameter for Regression Problems”. *Communications in Statistics - Theory and Methods* 34 (5): 1177–1182.

Lefebvre, G., J. a. Delaney, and R. L. McClelland. 2014a. “Extending the Bayesian Adjustment for Confounding algorithm to binary treatment covariates to estimate the effect of smoking on carotid intima-media thickness: the Multi-Ethnic Study of Atherosclerosis”. *Statistics in Medicine* 33 (16): 2797–2813.

Lefebvre, G., J. Atherton, and D. Talbot. 2014b. “The effect of the prior distribution in the Bayesian Adjustment for Confounding algorithm”. *Computational Statistics and Data Analysis* 70:227–240.

Pearl, J. 2009. *Causality: models, reasoning, and inference.* 2nd. New York: Cambridge University Press.

Raftery, A., D Madigan, and J. Hoeting. 1997. “Bayesian Model Averaging for Linear Regression Models”. *Journal of the American Statistical Association* 92 (437): 179–191.

Schwarz, G. 1978. “Estimating the Dimension of a Model”. *The Annals of Statistics* 6 (2): 461–464.

Tibshirani, R. 1996. “Regression shrinkage and selection via the lasso”. *Journal of the Royal Statistical Society: Series B (Methodological)* 58 (1): 267–288.

Walter, S., and H. Tiemeier. 2009. “Variable selection: Current practice in epidemiological studies”. *European Journal of Epidemiology* 24 (12): 733–736.

Wang, C., F. Dominici, G. Parmigiani, and C. M. Zigler. 2015. “Accounting for uncertainty in confounder and effect modifier selection when estimating average causal effects in generalized linear models”. *Biometrics* 71 (3): 654–665.

Wang, C., G. Parmigiani, and F. Dominici. 2012. “Bayesian effect estimation accounting for adjustment uncertainty.” *Biometrics* 68 (3): 661–71.

Weng, H. Y., Y. H. Hsueh, L. L. M. Messam, and I. Hertz-Picciotto. 2009. “Methods of covariate selection: Directed acyclic graphs and the change-in-estimate procedure”. *American Journal of Epidemiology* 169 (10): 1182–1190.

Wong, K. Y., and S. N. Chiu. 2015. “An iterative approach to minimize the mean squared error in ridge regression”. *Computational Statistics.*

Chapter 4

ADJUSTING FOR UNMEASURED SPATIAL CONFOUNDING

4.1 Introduction

In environmental epidemiology, we are often interested in making inference about the relationships between exposures and health outcomes that are both spatially referenced. Exposures frequently have spatial structure due to underlying ecological processes and human activities, while health outcomes can have spatial structure derived from both the exposure of interest and other, possibly unrelated, factors. For example, cardiovascular mortality and other health outcomes that are of interest in air pollution studies are associated with socioeconomic status (Diez Roux et al. 2001), which can have geographic structure and be difficult to quantify (Gee and Payne-Sturges 2004; Jerrett et al. 2004). These additional sources of spatial variability in the health outcome can induce systematic variation beyond that due to exposure and can confound the association of interest. When not measured directly (or measured incompletely), these factors can cause *unmeasured spatial confounding*, which is the inability to distinguish the effect of a spatially-varying exposure from residual spatial variation in a health outcome, resulting in biased point estimates and standard errors (Paciorek 2010).

4.1.1 Background

Early discussion of spatial confounding in the literature includes Clayton et al. (1993), who described the ‘confounding effect due to location’ in regression models for ecological studies, which they attributed to unmeasured confounding factors. Clayton et al. (1993) advocated for the inclusion of a spatially-correlated error term in hierarchical models for spatial data (Clayton and Kaldor 1987; Besag et al. 1991) and claimed this would account for confounding bias but might result in conservative inference. Since then, the approach of adding spatially structured error terms has frequently been used in spatial models for areal data (Wakefield 2007). The assumed nature of confounding is often not explicitly stated, but has been attributed to an unmeasured random source (Hanks et al. 2015). In these settings, Reich et al. (2006) demonstrate that correlation between the exposure of interest and the spatial random effect can dramatically change the estimate of interest when the random effect is included. To address this, Hodges and Reich (2010) suggest adding a spatial random effect that is orthogonal to the exposure. This attributes all spatial variability in the outcome that is in the space spanned by the exposure surface to the exposure. They call this *restricted spatial regression*, and related methods have been developed to extend it and improve its performance within areal data settings (Hodges and Reich 2010; Hughes and Haran 2013; Hanks et al. 2015).

For point-referenced data, Paciorek (2010) provided one of the first rigorous discussions of spatial confounding and the effectiveness of different efforts to adjust for the bias it can induce. An important finding of Paciorek (2010) is that the spatial scales of variability in the exposure and outcome are critical. In the examples of Paciorek (2010), scale was indexed by a spatial range parameter in a Matern covariance function (Cressie 1993). Reductions in bias can be obtained when the scale of unconfounded variability in the exposure is smaller than the scale of confounded variability.

Understanding the spatial scale of associations and confounding has important implications for the manner in which we conduct analyses. For example, many air pollution epidemiology studies use data from existing monitoring networks, which often have sparse spatial coverage. This means that exposure prediction models can typically identify large or regional scale variability in the exposure surface, but not smaller scale variability. For example, this was explored by Brochu et al. (2011) in their analysis of the spatial scale of association between particulate matter and measures of socioeconomic status, where they adjusted for ‘regional’, ‘subregional’, and all but metropolitan variability.

Rather than assume confounding arises due to random variaion, one can consider spatial confounding as due to a fixed, but unmeasured, spatial surface. This has particular relevance in air pollution cohort studies of long-term exposures, when the long-term ambient concentration can be viewed as fixed (Szpiro and Paciorek 2013). If we were able to perfectly measure the confounding variables, then we could include them in the regression model and obtain an unbiased estimator for the association of interest. But we because we cannot measure them, our estimate may be biased due to spatial confounding. In the fixed surface setting, spatial confounding can also be thought of as due to approximate *concurvity*, which is multicollinearity for spline basis functions (Hastie and Tibshirani 1990; Ramsay et al. 2003b; Ramsay et al. 2003a).

Spatial data under the fixed surface assumption are usually approached using non-Bayesian methods. Approaches to dealing with spatial confounding have typically followed the work done in the time series literature. Time series studies have a fundamentally different sampling scheme based upon repeated measurements over time, but temporal associations between exposure and health outcomes can be confounded by unmeasured temporal variability in a manner analogous to unmeasured spatial confounding (Dominici et al. 2002). A standard approach to dealing with temporal confounding is the addition of smoothing splines or other flexible temporal basis function (Burnett and Krewski 1991; Schwartz 1994; Dominici et al. 2003). The spline or basis adjustment is typically added to the health model, but can also be used to first 'pre-whiten' the exposure. Under certain conditions, these approaches have been shown to yield unbiased inference in time series studies (Dominici et al. 2004) and for short-term exposures in cohort studies (Szpiro et al. 2014).

The spatial analogue to this time series approach to account for unmeasured confounding from a fixed surface is to include spatially-structured splines or basis functions. A typical approach is to include thin-plate regression splines (TPRS; Wood (2003)) in the health model. In brief, TPRS are low-rank approximations to thin-plate smoothing splines indexed by a degrees of freedom (*df*) parameter. We discuss TPRS in greater detail in Section 4.2.5, but here we mention some important characteristics of their use. While flexible and easily implemented via the `mgcv` package in R (Wood 2003), they have some drawbacks: (1) TPRS depend directly upon the location of the observed data, in contrast to knot-based methods. This has the drawback of implicitly creating a different surface for confounding adjustment depending upon which subject locations are selected for an analysis. While advantageous for prediction and smoothing, this can be a major problem in the context of parameter inference, since the basis for spatial adjustment varies by the sampled subjects. (2) There is no natural mapping between *df* and the spatial scale of the data, meaning that the extent of the adjustment is difficult to interpret. (3) Because the scientific interpretation of *df* is unclear, there is no clear method for choosing an appropriate extent of adjustment.

4.1.2 Motivating Example

As a motivating example, we consider an analysis of systolic blood pressure and fine particulate matter ($\text{PM}_{2.5}$) in the NIEHS Sister Study (Chan et al. 2015). As previously described in Section 2.2, Chan et al. (2015) found that long-term exposure to ambient fine particulate matter ($\text{PM}_{2.5}$) was associated with higher systolic blood pressure (SBP). Particulate matter is known to be correlated with differences in socioeconomic status (Brochu et al. 2011; Jerrett et al. 2004), and to obtain accurate estimates of the SBP-PM association we want to limit confounding by socioeconomic status. To account for spatial confounding from unmeasured regional differences in socio-economic and health patterns, Chan et al. (2015) included TPRS with 10 *df* in their model.

Here we consider a re-analysis of this cohort using $\text{PM}_{2.5}$ exposures at grid locations, rather than at subject residences, to accommodate the methods we describe below. We use predictions of the 2006 annual average ambient concentration from the universal kriging model of Sampson et al. (2013), made on a national 25km by 25km grid. For each Sister Study subject, we assign exposure based upon the closest grid cell center. Using the same measured confounders as Chan et al. (2015) and the analysis from Chapter 2, but no spatial smoothing, we find that a difference of 10 $\mu g/m^3$ in $\text{PM}_{2.5}$ is associated with 0.30 mmHg higher SBP (95% CI: -0.11, 0.70). However, if we add TPRS with 10 *df*, as was done by Chan et al. (2015), then the estimated difference in SBP is 1.48 mmHg (95% CI: 0.82, 2.14). For other amounts of spatial adjustment, as indexed by df, the estimate varies and eventually fluctuates around zero (see Figure 4.1). This strongly suggests that some form of spatial confounding is present. But it is not clear on what scale the confounding is occurring, how much adjustment should be done, and in what way adjustment should be done.

4.1.3 Objectives

In this chapter, we have three main objectives: (1) to describe methods for quantifying the scale and extent of spatial confounding in situations that may arise in epidemiological contexts, (2) to develop methods for adjusting for spatial confounding that are *tunable* to a desired spatial scale, and (3) to apply these methods to the NIEHS Sister Study analysis. In Section 4.2 we introduce notation for describing spatial scale and follow in Section 4.3 with methods for adjusting for spatial confounding. In Sections 4.4 and 4.5 we present several simulations comparing these approaches and discuss their effectiveness. We discuss their application to the motivating study in Section 4.6 and conclude with a discussion in Section 4.7.

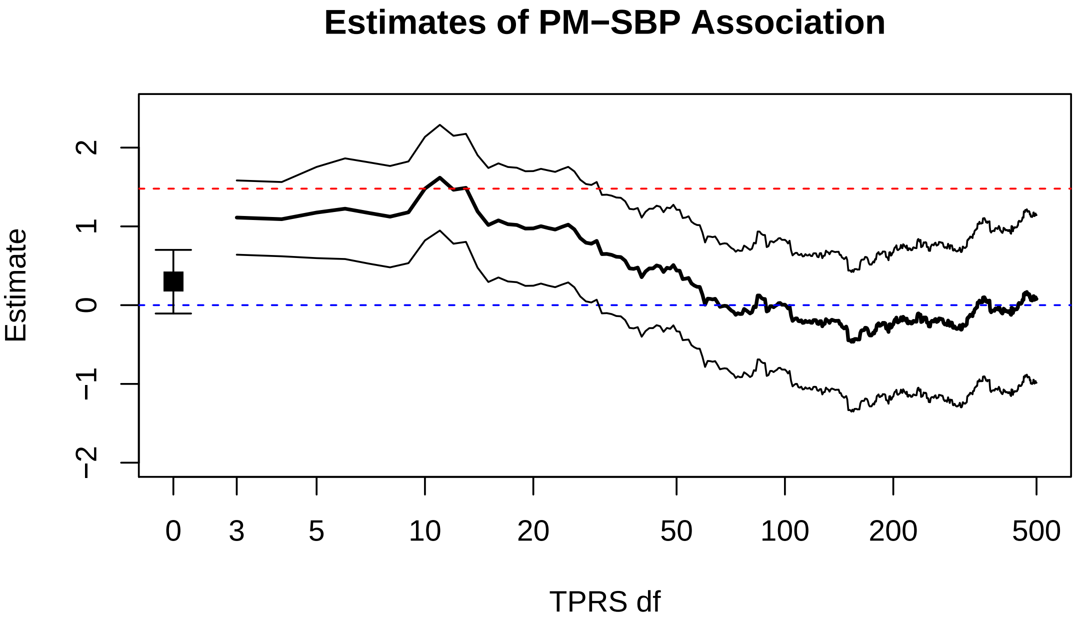


Figure 4.1: Estimated difference in SBP (in mmHg) for a difference of 10 $\mu g/m^3$ in annual average ambient $\text{PM}_{2.5}$, when including TPRS with varying values of df. The square point on the left is the estimate without spatial adjustment, and has error bars indicating a 95% confidence interval. The thick black curve represents estimates for different choices of df, with the thin black curves representing point-wise confidence intervals. The horizontal dashed red line indicates the association estimate at 10 df, which was the extent of adjustment in the primary model of Chan et al. (2015). The dashed blue line is at zero.

4.2 Spatial Basis Functions and Scales

4.2.1 Notation

We assume that the health outcomes y_i arise from the model

$$y_i = \beta_0 + \beta x(s_i) + f(s_i) + \epsilon_i, \tag{4.1}$$

where s_i denotes the location of subject i and $\epsilon_i \overset{iid}{\sim} N(0, \sigma_\epsilon^2)$. The target of inference is the parameter $\beta \in \mathbb{R}$, which represents the association between exposure $x(s_i)$ and the outcome. Unmeasured spatial variability in y_i comes from $f(s_i)$, which we assume to be fixed and unknown.

We assume that the spatial exposure $x(s_i) = g(s_i) + \eta(s_i)$ is a combination of a spatially smooth surface g and small-scale variability η, which could be due to measurement error, for example (Tiefelsdorf and Griffith 2007). We further assume the exposure surface has been observed. The methods that follow can be applied to predicted exposures, but we make no formal accounting for measurement error that may be induced by the use of predicted exposures. Measurement error, like confounding, may cause bias in the point estimate, but the confounding we consider here may occur even in the absence of measurement error.

4.2.2 Spatial Basis Functions

Variation at particular spatial scales may be described using a chosen set of spatial basis functions (or surfaces). Orthogonal basis functions are a particularly useful class of spatial basis, because they can be easily partitioned into (typically hierarchical) variation at different scales. Let $\{h_j(\cdot)\}$, $j = 1, 2, \dots$ be a set of orthogonal basis functions, ordered such that h_j has finer-scale variation than $h_{j'}$, for $j > j'$. (In this notation, we do not explicitly indicate ties, when different basis functions have the same scale of variability, but they do not pose a significant problem for this approach.) Following Dominici et al. (2004) and Szpiro et al. (2014), we assume that $f(s_i)$ and $g(s_i)$ can be decomposed into m_1 and m_2 of these basis functions:

$$f(s_i) = \sum_{j=1}^{m_1} h_j(s_i)\gamma_j \qquad \text{and} \qquad g(s_i) = \sum_{j=1}^{m_2} h_j(s_i)\psi_j.$$

For an arbitrary exposure surface, the precise form of h_j that generated the data may not be known, although we will in various settings choose a particular basis. In such cases, any variation in x that cannot be represented by the chosen basis h_j is subsumed into η.

4.2.3 Fourier basis

One choice for the h_j are Fourier basis functions, which can represent any continuous function as a (possibly infinite) sum of sinusoidal functions. For finite discrete data, assume the data exist on a regular $M \times N$ grid, which we index by (u, v) for $u = 0, \dots, M-1$ and $v = 0, \dots, N-1$. Then the value of the function $g(u, v)$ can be written as

$$\begin{aligned} g(u,v) &= \frac{1}{\sqrt{MN}} \sum_{m=0}^{M-1} \sum_{n=0}^{N-1} G(m,n) e^{i2\pi(mu/M + nv/N)} \\ &= \frac{1}{\sqrt{MN}} \sum_{m=0}^{M-1} \sum_{n=0}^{N-1} G(m,n) \left(\cos\left[2\pi\left(\frac{mu}{M} + \frac{nv}{N}\right)\right] + i \sin\left[2\pi\left(\frac{mu}{M} + \frac{nv}{N}\right)\right]\right) \end{aligned}$$

where m and n are spectral coordinates and $G(m, n)$ are coefficients from the Discrete Fourier Transform (DFT) of $g(u, v)$. Although the decomposition involves complex coefficients, $g(u, v)$ is real-valued.

The spectral coordinates (m, n) provide a hierarchical ordering of the sinusoidal basis functions by frequency ω. Allowing for ties as necesary, we can order the spatial basis functions by *effective frequency* :

$$\omega_{(m,n)} = \frac{1}{\tau}\sqrt{\left(\frac{m}{M}\right)^2 + \left(\frac{n}{N}\right)^2}, \tag{4.2}$$

where $1/\tau$ is the sampling frequency (so τ is the sampling interval length) (Burger and Burge 2009, p.164). Because each basis function is oriented to a particular angle, there may be multiple basis functions with the same frequency ω but different orientation. Additionally, aliasing between frequencies prevents us from identifying frequencies higher than the Nyquist frequency of $\omega = 2/M$. If the sampling interval is $\tau = 1/M$ and $M = N$ (i.e the data are on the unit square), then $\omega_{(m,n)} = \sqrt{m^2 + n^2}$. This frequency corresponds to a period of $1/\omega = 1/\sqrt{m^2 + n^2}$, which provides an index of spatial scale and thus a distance of spatial variability.

4.2.4 Wavelet Basis

A second choice of orthogonal spatial basis functions h_j are wavelets. Like Fourier basis functions, wavelets are localized in frequency, but they are also localized in space (Nason 2008). This means that wavelets can compactly describe variation at different frequencies in different areas of a spatial domain. Wavelet basis functions are indexed by *level*, based upon successive halving of the spatial domain. Within each level L (and thus, within each frequency 2^L and spatial scale 2^{-L}), there are multiple wavelet basis functions with different orientations and spatial positions.

For finite data on a discrete grid, the Discrete Wavelet Transform (DWT) maps any surface to a set of wavelet coefficients. There are many different sets of wavelets, and here we consider the widely used Daubechies wavelets, which are quite smooth (Daubechies 1988). Unlike the DFT, the DWT requires that the grid have length equal to a power of two. Data on a non-square grid can be embedded within a larger grid with dyadic dimension to apply the DWT. For an accessible introduction to wavelets, we refer the reader to Nason (2008) and its accompanying R package `wavethresh`, which provides extensive wavelet capabilities in R.

It is important to note that although a Fourier frequency and wavelet level may correspond to the same spatial scale (e.g. $\omega = 16$ and $L = 4$, which both correspond to a spatial scale of $1/16 = 2^{-4} = 0.0625$ on the unit grid), the nature of variation described by the basis functions is different. This is because the two sets of basis functions have distinct shapes, and thus span different sets of spatial surfaces.

4.2.5 Thin-plate regression splines

Thin-plate regression splines (TPRS) are low-rank smoothing splines that approximate thin-plate splines (TPS) (Wood 2003). TPRS achieve computational benefits over full-rank TPS by using a tuncated eigenbasis to approximate the distance matrix used in calculation of TPS basis functions. TPRS are based upon the spatial locations of observed points, eliminating the need for knot-selection.

When used for semi-parametric smoothing of a spatial surface, TPRS are typically penalized using a GCV criterion. But for problems of inference, the splines can be unpenalized and the degrees of freedom (df), which controls the dimension of the basis, can be fixed. This yields a set of basis functions $t_1(s; \boldsymbol{s}, df), t_2(s; \boldsymbol{s}, df), \ldots, t_{df}(s; \boldsymbol{s}, df)$. Here the subscript $j = 1, 2, \ldots, df$ indexes each basis function, and we have explicitly indicated the dependence upon all spatial locations and the total number of basis functions created. The first two basis functions are linear in the first two spatial coordinates (i.e. x and y directions). Notably, the t_j are not orthogonal to one another. Each individual basis function depends upon the total degrees of freedom in the model and so they are not nested, i.e. $t_j(s; \boldsymbol{s}, k) \neq t_j(s; \boldsymbol{s}, k')$ for $k, k' \geq j$ such that $k \neq k'$. But the space they span is hierachical, i.e. $span(\{t_1(\cdot; \boldsymbol{s}, k), \ldots, t_k(\cdot; \boldsymbol{s}, k)\}) \subseteq span(\{t_1(\cdot; \boldsymbol{s}, k'), \ldots, t_k(\cdot; \boldsymbol{s}, k')\})$ for $k' \geq k$, which allows for comparisons of increasing amounts of adjustment.

4.3 Adjustment Approaches

In this section, we describe two different strategies for spatial confounding adjustment: making adjustment in the health model and pre-adjusting the exposure surface.

4.3.1 Adjustment in Health Model

Including basis functions in the health model is a straightforward way to adjust for spatial structure and represents a direct extension of time series methods. For TPRS, including basis functions in the health model amounts to fitting a generalized additive model for y_i

and can be easily accomplished using the `mgcv` package in R (Wood 2003).

For wavelets and Fourier bases, the spatial basis functions must be explicitly constructed and included in the health model. Wand and Ormerod (2011) provide an excellent discussion of how penalized wavelets can be used in a manner analogous to penalized splines. They give an overview of how this can be implemented in R, using `wavethresh`. However, Wand and Ormerod (2011) only considered 1-dimensional functions. For two-dimensional functions, the number of necessary basis functions grows rapidly (for L levels, $\sum_{\ell=0}^{L-1} 3 \times 4^L$ basis functions are needed), making such an approach computationally impractical for grids of even moderate size. The number of basis function grows even faster when allowing for translations of the basis functions. For Fourier basis, adjusting in the health model also requires an exceptionally large number of basis functions. Although in principle any finite function can be written as a linear combination of Fourier basis functions, the large number of possible rotations for each frequency make the number of explicit basis functions needed impractical.

4.3.2 Pre-adjusting Exposure

A second approach to confounding adjustment is to first 'pre-adjust' the exposure $\boldsymbol{x}$. This achieves unconfounded health estimates by eliminating variability in the exposure that is correlated with the confounding surface. This is a form of detrending for spatial data, where the goal is to remove the trend of the unmeasured spatial confounder. Spatial filtering has been applied in various forms throughout the spatial literature, although typically as means to explicitly account for correlated errors and not confounding, directly (e.g. Haining 1991; Tiefelsdorf and Griffith 2007).

For some choice of orthogonal basis functions h_j and $m' < m_2$, we can decompose the exposure as (Szpiro et al. 2014):

$$\begin{aligned} x(s_i) &= \sum_{j=1}^{m_2} h_j(s_i)\psi_j + \eta(s_i) \\ &= \left(\sum_{j=1}^{m'} h_j(s_i)\psi_j\right) + \left(\sum_{j=m'+1}^{m_2} h_j(s_i)\psi_j + \eta(s_i)\right) \\ &= x_{m'}(s_i) + \eta_{m'}(s_i). \end{aligned}$$

We then fit the model

$$y_i = \beta_0 + \eta_{m'}\beta_x + x_{m'}\tilde{\beta}_x + \epsilon'_i \tag{4.3}$$

where $\epsilon'_i = \epsilon_i + f(s_i)$ and $\beta_x, \tilde{\beta}_x \in \mathbb{R}$. Due to the orthogonality of the h_j, the $\tilde{\beta}_x$ term does not confound the estimate of β_x, but we include it to adequately model the variation in y. Forcing $f(s_i)$ into the error term creates correlation between the ϵ'_i, but this can be accounted for using 'sandwich' standard error estimates. The extent of bias adjustment depends upon the relationship between m_1, m_2, and m'. If $m' \geq m_1$, then straightforward computation shows that $\hat{\beta}_x$ is unbiased. If $m_1 > m'$, then the estimate is biased. However, the bias may still be small, if γ_j is small relative to ψ_j, for all $j = m' + 1, \ldots, m_2$.

Pre-adjustment using TPRS

Pre-adjustment using TPRS is done by first choosing the desired value of *df* and constructing the TPRS basis functions from the locations of the exposure measurements, which are not required to be on a grid. The exposure is then partitioned into the space spanned by these basis functions and its orthogonal complement (since TPRS basis functions are not orthogonal) to yield $\boldsymbol{x}_{m'}$ and $\boldsymbol{\eta}_{m'}$, respectively.

Frequency filtering as exposure pre-adjustment

A major advantage of the pre-adjustment approach is that it does not require the direct computation of a large number of Fourier basis functions. Instead, the exposure can be transformed to frequency space using the DFT and the pre-adjustment done there using a high-pass filter. This approach to pre-adjusting $x(u, v)$ can be outlined as follows.

1. Compute the Fourier decomposition $\mathcal{X}(m, n)$ of $x(u, v)$.

2. Create a high-pass filter at frequency ω_0 as $H_{\omega_0}(m, n) = \begin{cases} 1 & \text{if } \omega_{(m,n)} > \omega_0 \\ 0 & \text{if } \omega_{(m,n)} \leq \omega_0 \end{cases}$, where $\omega_{(m,n)}$ is the effective frequency (4.2).

3. Multiply $\mathcal{X}(m, n)$ and $H_{\omega_0}(m, n)$ elementwise to obtain $\mathcal{X}_{\omega_0}(m, n)$.

4. Apply the inverse Fourier transform to $\mathcal{X}_{\omega_0}(m, n)$ to get $x_{\omega_0}(u, v)$, which serves as the pre-adjusted exposure $\boldsymbol{\eta}_{m'}$.

The choice of ω_0 allows for this adjustment to be tuned to a particular spatial scale.

Wavelet thresholding as exposure pre-adjustment

For a wavelet basis, the coefficient thresholding approach is similar to the high-pass filter approach used for the Fourier basis. De-correlation of the exposure and outcome via wavelets was presented in an ecological context by Carl and Kühn (2008). However, they applied the thresholding to the exposure and the outcome jointly and performed regression on the wavelet coefficients, which is not practical for the large grids and when we have other confounders in the model.

To pre-adjust $x(u, v)$ using wavelets, we:

1. Compute the DWT of $x(u, v)$.

2. Threshold all coefficients at levels $\ell = 0, \ldots, L$, where $L \leq log_2(M)$ and $M \times M$ is the grid size.

3. Apply the inverse wavelet transform to get the pre-adjusted exposure $x_L(u, v)$.

Although this procedure is described as a global threshold of all wavelet coefficients up to a particular level, the extent of thresholding could be varied across the exposure domain if desired.

4.4 Simulations

We conduct several simulations to compare the adjustment approaches proposed in Section 4.3. The data for the simulations are created on a 512×512 grid on the unit square $[0,1) \times [0,1)$. Each point on this grid is indexed by the coordinates (u, v). For each simulation, we construct a fixed exposure surface $X(u, v)$ and a fixed confounder surface $Z_1(u, v)$ on this grid. Together these determine the outcome value

$$Y(u,v) = X(u,v)\beta + Z_1(u,v)\gamma_1 + \epsilon(u,v),$$

where $\epsilon(u,v) \overset{iid}{\sim} N(0,1)$. The definitions of X and Z_1 differ by simulation and are described on a case-by-case basis below. In each simulation, we consider 1000 replications, where n observation locations (selected from the grid points) and the ϵ values are resampled in each replication. For a given choice of X, Z_1, and the model parameters $(\beta, \boldsymbol{\gamma})$, we compare the unadjusted estimator and fully-adjusted (oracle) estimator to four sets of estimators: adjustment in the health model using TPRS, pre-adjustment of exposure using TPRS, pre-adjustment using Fourier basis via a high pass filter, and pre-adjustment of exposure via wavelet thresholding. An outline of the simulation algorithm is provided in Appendix C.1.

4.4.1 Simulation 1: Confounder that is periodic and TPRS

Setup

The first simulation set was designed to compare the effectiveness of the three approaches when the confounders are various combinations of periodic surfaces and thin-plate regression splines. We first created three surfaces:

$$
\begin{aligned}
G_1(u,v) &= \sum_{k=1}^{50} \xi_k t_k(s; \boldsymbol{s}, 50) \\
G_2(u,v) &= \cos(24\pi u + 10\pi v) + \cos(10\pi u + 24\pi v) + \cos(2\pi x) \\
G_X(u,v) &\sim GP(0.5, 0.1).
\end{aligned}
$$

where we use $GP(\phi, \nu)$ to denote a Gaussian Process with zero mean and Matern covariance with range ϕ and differentiability parameter ν. The coefficients ξ_k are drawn from a Uniform$(-2, 2)$ distribution. A single, fixed realization of G_X is used for the simulation. The three surfaces are plotted in Figure 4.2. The surfaces G_1 and G_2 are used to construct the confounder surface Z_1 as

$$Z_1(u,v) = \alpha_1 G_1(u,v) + \alpha_2 G_2(u,v). \tag{4.4}$$

The parameters α_1 and α_2, which were varied in the simulation, control the relative the contribution of TPRS (G_1) and sinusoidal functions (G_2) to the confounder surface. The exposure surface was created as a linear combination of this confounder and G_X as

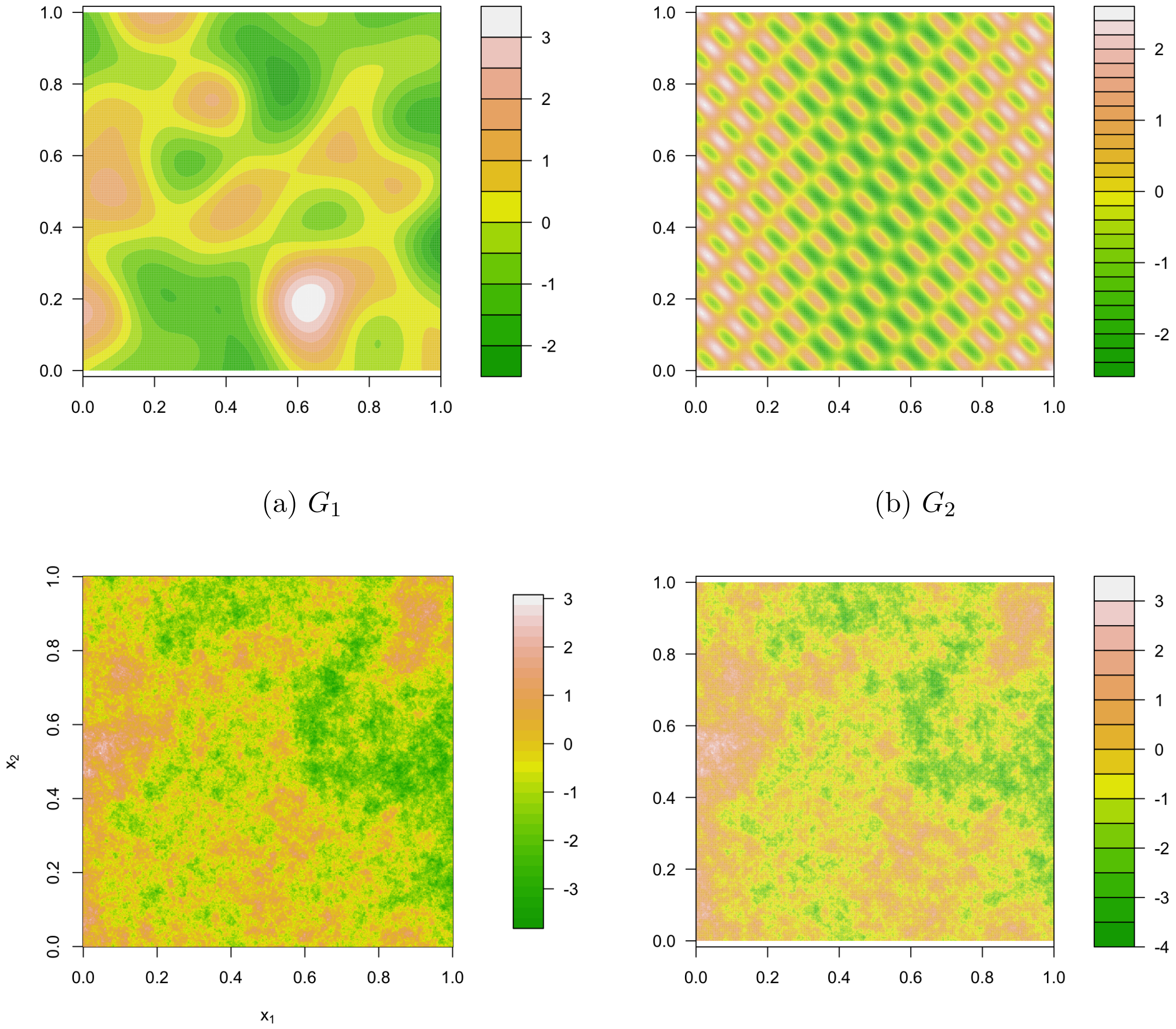
$$X(u,v) = 0.2 Z_1(u,v) + G_X(u,v). \tag{4.5}$$

Lastly, the outcome mean was constructed as

$$\mathrm{E}[Y(u,v)|X, Z_1] = \mu_Y(u,v) = X(u,v) + 0.5 Z_1(u,v). \tag{4.6}$$

Although G_X is generated independently of G_1 and G_2, any particular realization of G_X may be correlated with G_1 and G_2. Because this simulation combines confounding from these two sources, we wish to treat them symmetrically. Therefore we chose a realization of G_X that had low correlation with G_1 and G_2: $\mathrm{Cor}(G_X, G_1) = 0.12$ and $\mathrm{Cor}(G_X, G_2) = 0.04$.

By construction, the power spectrum of $G_2(u,v)$ is limited to the effective frequencies $\omega = 1$ and $\omega = 13$. Figure 4.3a, which plots the power spectrum of X and Z_1 as a function of ω, shows why we expect the high-pass filtering pre-adjustment approach to work when



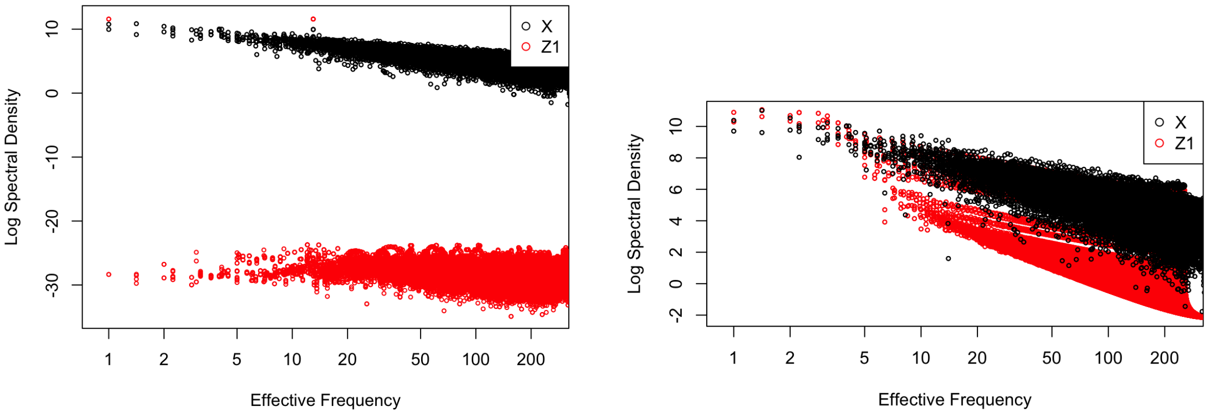
(a) G_1

(b) G_2

(c) G_X

(d) X

Figure 4.2: Surfaces from Simulation 1. The surface shown for X corresponds to when $\alpha_1 = \alpha_2 = 1$.



(a)

(b)

Figure 4.3: Log power spectrum for X and Z_1 in Simulation 1. In (a), (α_1, α_2) equals $(0, 1)$, while in (b) it equals (1, 0).

confounding is due to G_2 alone. Once the energy spike at $\omega = 13$ has been removed, there is much greater (spectral) power at higher frequencies for X compared to Z_1, and so the remaining variability in X is unconfounded.

Results

The results for applying each of the adjustment approaches in Simulation 1 are provided in Figure 4.4, for three different combinations of α_1 and α_2 when $n = 2,000$.

We first consider the results for $\alpha_1 = 1$ and $\alpha_2 = 0$, when the confounding is due entirely to G_1, which is a linear combination of the first 50 TPRS basis surfaces. In this setting, the average unadjusted estimate is 1.2. In the top-left panel of Figure 4.4, we see that adjusting for TPRS with *df* > 50 in the health models eliminates bias in the estimate. This is as expected; the confounder has been completely adjusted for by eliminating all variation in the exposure surface that can be projected onto the linear space of the confounder. The reduction in bias is not, however, monotonic. The bias when adjusting using *df* $= 7$ is greater than when adjusting using *df* $= 5$. The pre-adjustment approach using TPRS performs similarly to adjusting using TPRS in the health model. The high-pass filter approach and wavelet thresholding approach both eliminate much, but not all, of the bias (see first column of Figure 4.4). This is somewhat surprising, given that the confounder is not an explicit function of periodic sinusoidal surfaces. However, we see from Figure 4.3b that the power spectrum for the confounder Z_1 under this scenario remains lower than that for the exposure, thus the spectral filtering approach reduces confounded variation in the exposure as ω increases.

For the opposite setting when $\alpha_1 = 0$ and $\alpha_2 = 1$, the confounder Z_1 is entirely composed of sinusoidal functions and the average value of the unadjusted estimate is 1.16. As we expect, the high-pass filtering approach can completely remove the confounding bias for $\omega >= 13$, which corresponds to a spatial scales smaller than 0.077 (see bottom-middle panel of Figure 4.4). However, for values of ω below this threshold, the amount of bias increases with ω. This behavior can be explained by the power spectra in Figure 4.3a. For $\omega < 13$, most of the spectral energy removed by pre-adjustment using Fourier basis comes from variability that is not confounded by Z, thus increasing the correlation between the pre-adjusted X and the confounder Z, yielding increased bias. The wavelet thresholding approach, while using a different basis, is also able to eliminate most of the confounding bias for $L \geq 4$. This corresponds to a spatial scale of 0.063 (relative to the unit square spatial domain) and corresponds to $\omega \geq 16$. Like the high-pass filter based upon the Fourier basis, pre-adjustment using wavelets increases bias up the threshold value of $L = 3$. In this setting, the TPRS approaches perform poorly, with increasing bias for *df* ≥ 5.

Lastly, we consider the setting when the confounder is a combination of the G_1 and G_2 (that is, when $\alpha_1 = \alpha_2 = 1$). In this scenario, the bias in the unadjusted estimator is 0.34 (Table 4.1). Exposure pre-adjustment using TPRS is unable to remove much of the confounding bias in this scenario for any value of *df*. At *df* $= 50$, the TPRS methods removes only one-third of the bias (average estimate 1.21), but for higher values of *df* bias increases.

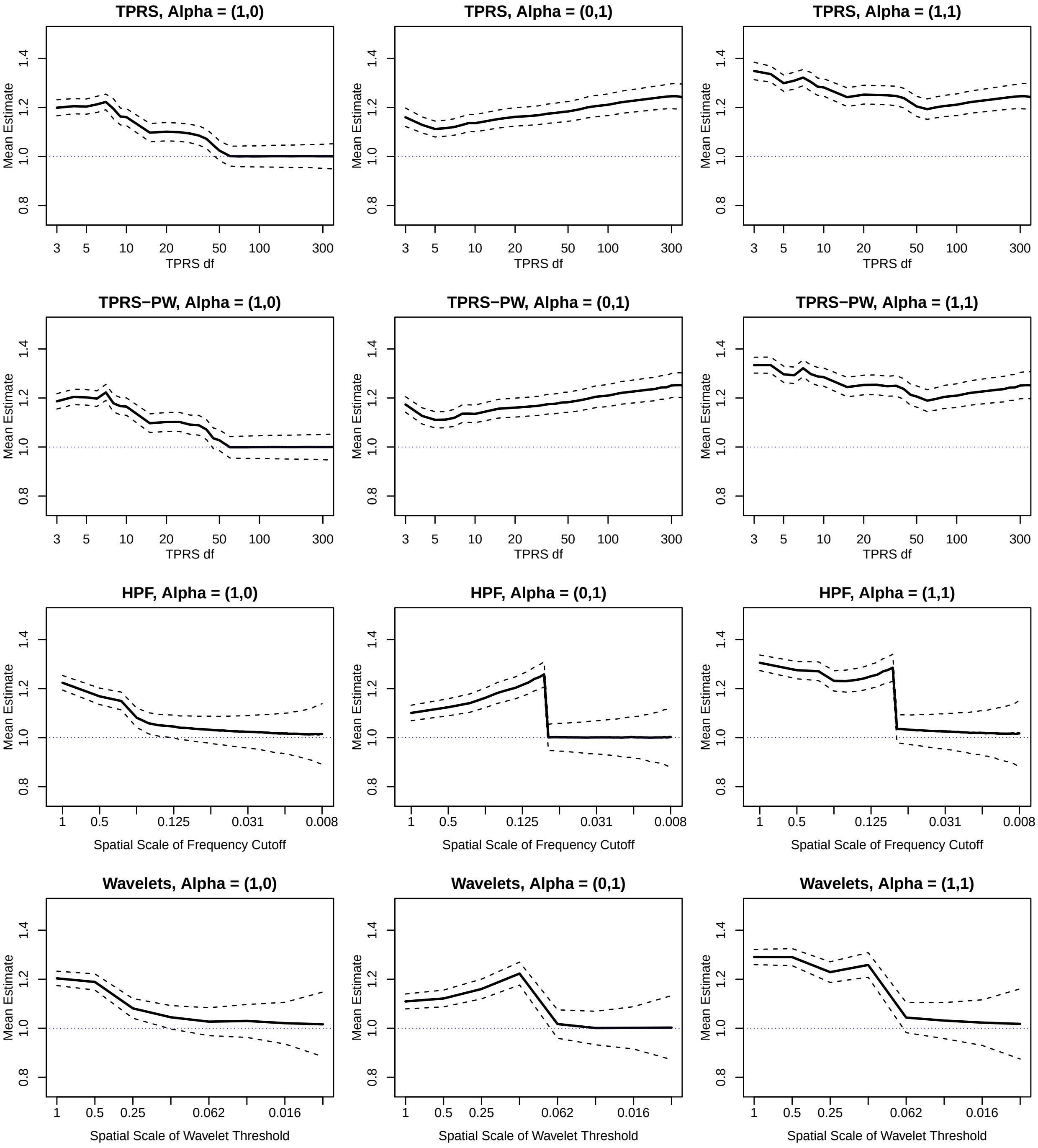


Figure 4.4: Estimates of β_1 from Simulation 1, when $n = 2,000$. The dashed lines indicates $+/-$ one standard error of the mean.

In contrast, both the high-pass filter approach and the wavelet thresholding approaches are able to eliminate most of the confounding (the bias is approximately 0.02 for $\omega \geq 32$ and $L \geq 6$), although this reduction is not monotonic. The behavior of each method under this scenario is essentially a composite of the behavior under the prior two settings. It is notable how well the frequency filtering and wavelet thresholding approaches do, even when the confounder is not entirely periodic.

Table 4.1 provides the mean-squared error (MSE), standard error estimates, and coverage rates for the estimates in the $\alpha_1 = \alpha_2 = 1$ setting. For any fixed choice of adjustment, the standard error estimates are accurate, but the coverage is highly dependent upon the bias of each estimator. In this scenario, all of the estimators have lower MSE than the unadjusted estimator that does not account for spatial confounding. As the extent of adjustment increases, we see that the standard errors of the estimator increase, due to reduced exposure variability. This demonstrates an inherent tradeoff that must be made when adjusting for spatial confounding: addition of spatial basis functions will increase the variability of an estimator.

4.4.2 Simulation 2

Setup

Simulation 2 was designed with a smoother exposure surface. This was achieved by having the Gaussian Process component of X be generated with a larger spatial range and smoothness parameter (i.e. $G_X = GP(2, 1)$). This new surface is plotted in Figure 4.5c. This allows for more pronounced confounding, because there is less small-scale variation in the exposure.

The TPRS surface here was constructed in the same manner as Simulation 1, but with different coefficients. Specifically,

$$G_3(u, v) = \sum_{k=1}^{50} \xi_k t_k(s; \boldsymbol{s}, 50)$$

for ξ_k drawn randomly from a Unif$(-1, 1)$ distribution. In contrast to Simulation 1, the

Table 4.1: Performance measures for different estimators in Simulation 1 for $(\alpha_1, \alpha_2) = (1, 1)$ setting. The TPRS values shown correspond to pre-adjustment of exposure. 'HPF' stands for high-pass filter, the pre-adjustment approach using Fourier basis functions. 'Wave' refers to pre-adjustment using wavelet thresholding. 'SE' denotes standard error, and reported coverage is for a nominal 95% confidence interval.

	$\hat{\beta}$	Bias($\hat{\beta}$)	$SE(\hat{\beta})$	MSE($\hat{\beta}$)	$\widehat{SE}(\hat{\beta})$	Coverage
Unadjusted	1.34	0.341	0.014	0.1167	0.013	0.000
Adjusted	1.00	0.000	0.012	0.0002	0.012	0.952
HPF: $\omega = 1$	1.31	0.306	0.014	0.0937	0.014	0.000
HPF: $\omega = 2$	1.28	0.276	0.016	0.0763	0.016	0.000
HPF: $\omega = 4$	1.23	0.233	0.018	0.0546	0.018	0.000
HPF: $\omega = 12$	1.28	0.285	0.024	0.0817	0.023	0.000
HPF: $\omega = 13$	1.03	0.035	0.025	0.0018	0.025	0.724
HPF: $\omega = 16$	1.03	0.031	0.026	0.0017	0.027	0.794
HPF: $\omega = 32$	1.02	0.024	0.032	0.0016	0.032	0.891
HPF: $\omega = 64$	1.02	0.018	0.040	0.0019	0.041	0.937
HPF: $\omega = 128$	1.01	0.014	0.057	0.0034	0.058	0.948
TPRS: $df = 3$	1.33	0.335	0.014	0.1121	0.014	0.000
TPRS: $df = 5$	1.30	0.297	0.015	0.0885	0.015	0.000
TPRS: $df = 50$	1.21	0.207	0.019	0.0431	0.019	0.000
TPRS: $df = 300$	1.25	0.251	0.023	0.0635	0.023	0.000
Wave: $L = 0$	1.29	0.291	0.014	0.0851	0.013	0.000
Wave: $L = 1$	1.29	0.290	0.015	0.0846	0.015	0.000
Wave: $L = 2$	1.23	0.231	0.018	0.0535	0.018	0.000
Wave: $L = 4$	1.04	0.043	0.027	0.0026	0.027	0.650
Wave: $L = 6$	1.02	0.022	0.040	0.0021	0.042	0.937
Wave: $L = 7$	1.01	0.014	0.061	0.0040	0.062	0.946

periodic surface for this simulation has lower frequency components:

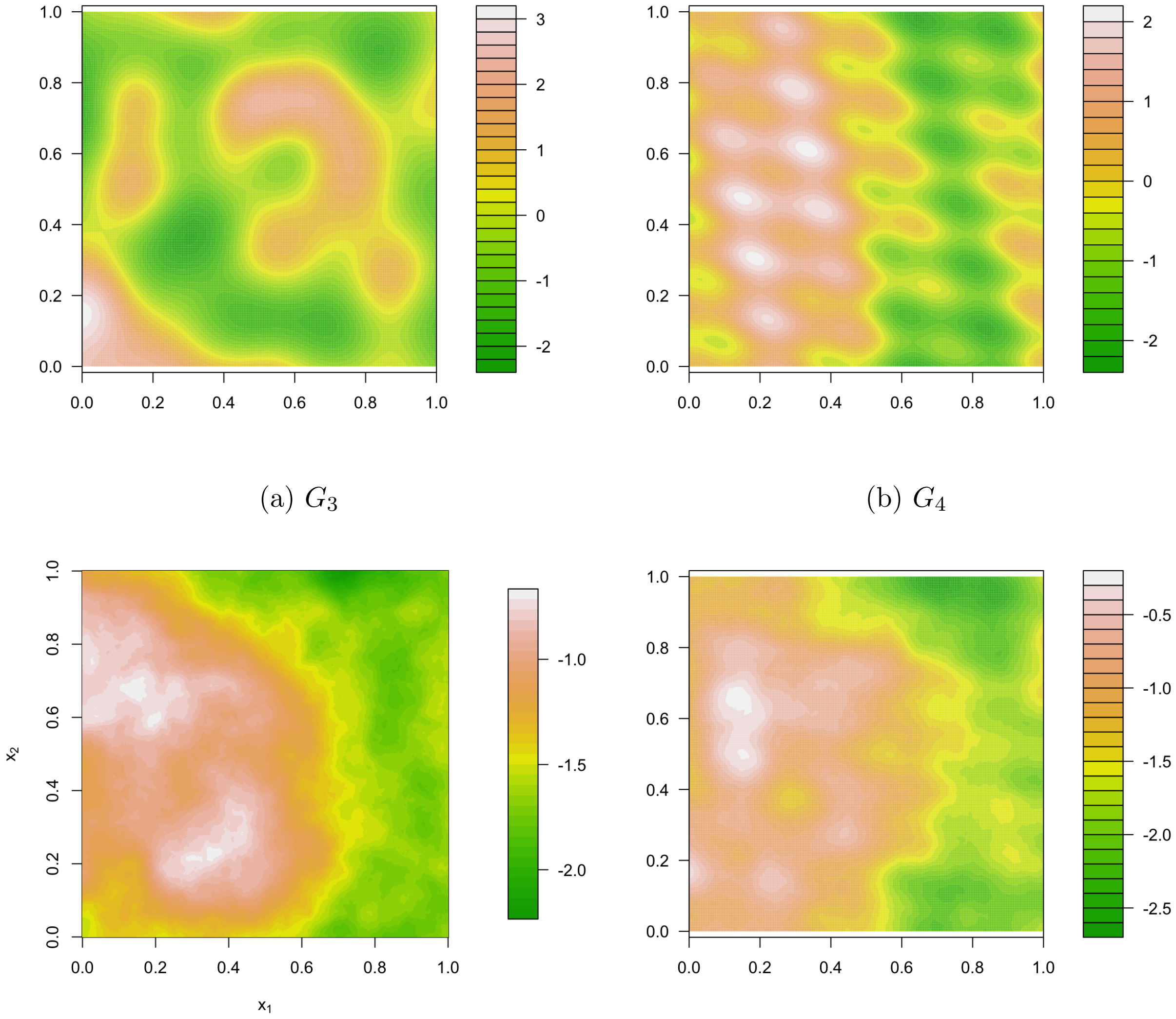
$$G_4(x,y) = \frac{1}{2}\cos(8\pi x + 2\pi y) + \frac{1}{2}\cos(2\pi x + 12\pi y) + \frac{3}{2}\sin(2\pi x) + \sin(\pi y).$$

G_3 and G_4, which are plotted in Figures 4.5a and 4.5b, were combined to form Z_1 in the same manner as G_1 and G_2 in (4.4). The lower frequency variation in G_4 means that there is a relatively high correlation of 0.74 between G_X and G_4. This is an example of the approximate concurvity that can result between two fixed surfaces. In contrast, the correlation between G_3 and G_X is -0.1.

Results

The results for applying each of the adjustment methods in Simulation 2 are shown in Figure 4.6, for $n = 2,000$. As expected, we see once again that when the adjustment approach matches the confounding surface, essentially all bias is removed. The first column of Figure 4.6 corresponds to the scenario when the confounder is only a linear combination of TPRS ($Z_1 = G_3$) and the average value of the unadjusted estimator is 1.59. In that case, the estimators that adjust for TPRS are unbiased for $df \geq 50$, although adjustment with fewer *df* results in higher bias than the unadjusted estimate. Neither the Fourier approach nor the wavelet thresholding approach can adequately account for this confounding surface, and a large bias (about 0.95) remains for all scales of adjustment. The large confidence bounds for high values of ω and L reflect the small remaining variability in the exposure after extensive pre-adjustment. The converse behavior is observed in the setting $(\alpha_1, \alpha_2) = (0, 1)$ (so that $Z_1 = G_4$), when both the Fourier approach and wavelet thresholding eliminate bias from confounding while adjustment via TPRS does not. A notable difference from Simulation 1, however, is that the TPRS approach does start to have reduced bias for very large (≥ 200) values of *df*. This suggests that TPRS may in some cases eventually be able to adequately model the exposure surface even when it is not generated by TPRS bases. However, the standard error of the estimator for such high *df* is quite large.

In the setting where the confounder is comprised of both the TPRS and periodic compo-



(a) G_3

(b) G_4

(c) G_X

(d) X

Figure 4.5: Surfaces in Simulation 2.

nents (i.e., $\alpha_1 = \alpha_2 = 1$), we see the Fourier and wavelet approaches perform poorly, and the TPRS remains biased for all but very high values of adjustment. Table 4.2 provides a summary of the estimators for specific choices of adjustment scale (*df*, ω, or L) in this setting. Although none of the methods are able to remove all of the confounding bias, adjustment at a small enough scale (i.e. large enough values of *df*, ω, and L) yields estimates that are less biased than the unadjusted estimate. This results in such high variance, though, that the adjusted estimators' MSE are in many cases worse than that of the unadjusted estimator.

4.4.3 Simulation 3

Setup

Simulation 3 was designed to complement Simulation 2 by having a higher-frequency periodic component, G_2 from Simulation 1, and a simpler TPRS component, $G_5 = \sum_{k=11}^{15} \xi_k t_k(s; \boldsymbol{s}, 15)$, with $\boldsymbol{\xi} = (0.6, 1, 0.5, 0.4, 0.4)$ (see Figure 4.7a). These were combined using α_1 and α_2 in the manner of (4.4). Figure 4.7b shows the surface X when $\alpha_1 = \alpha_2 = 1$. The low correlations between G_X and the confounder surfaces ($\text{Cor}(G_X, G_5) = -0.06$ and $\text{Cor}(G_X, G_2) = -0.03$) mean that the bias in the unadjusted estimators is similar for the $(\alpha_1, \alpha_2) = (1, 0)$ and (0,1) settings: 0.61 and 0.58, respectively.

Results

The results for applying each of the adjustment methods are shown in Figure 4.8. Appendix Table C.3 provides the results for specific choices of adjustment (*df*, ω, or L) in the $(\alpha_1, \alpha_2) = (1, 1)$ setting. Here we see again that method can essentially eliminate bias when it matches the spatial basis of confounder, but is unable to eliminate all bias when it does not match. However, we also see a reversed trend compared to Simulation 2: the high-pass filtering method comes much closer to eliminating bias than the TPRS does under the mis-matched scenario. In fact, the TPRS approach continues to have higher bias as *df* is increased, leading to much worse performance (MSE ≥ 4) than the unadjusted estimator (MSE=0.96).

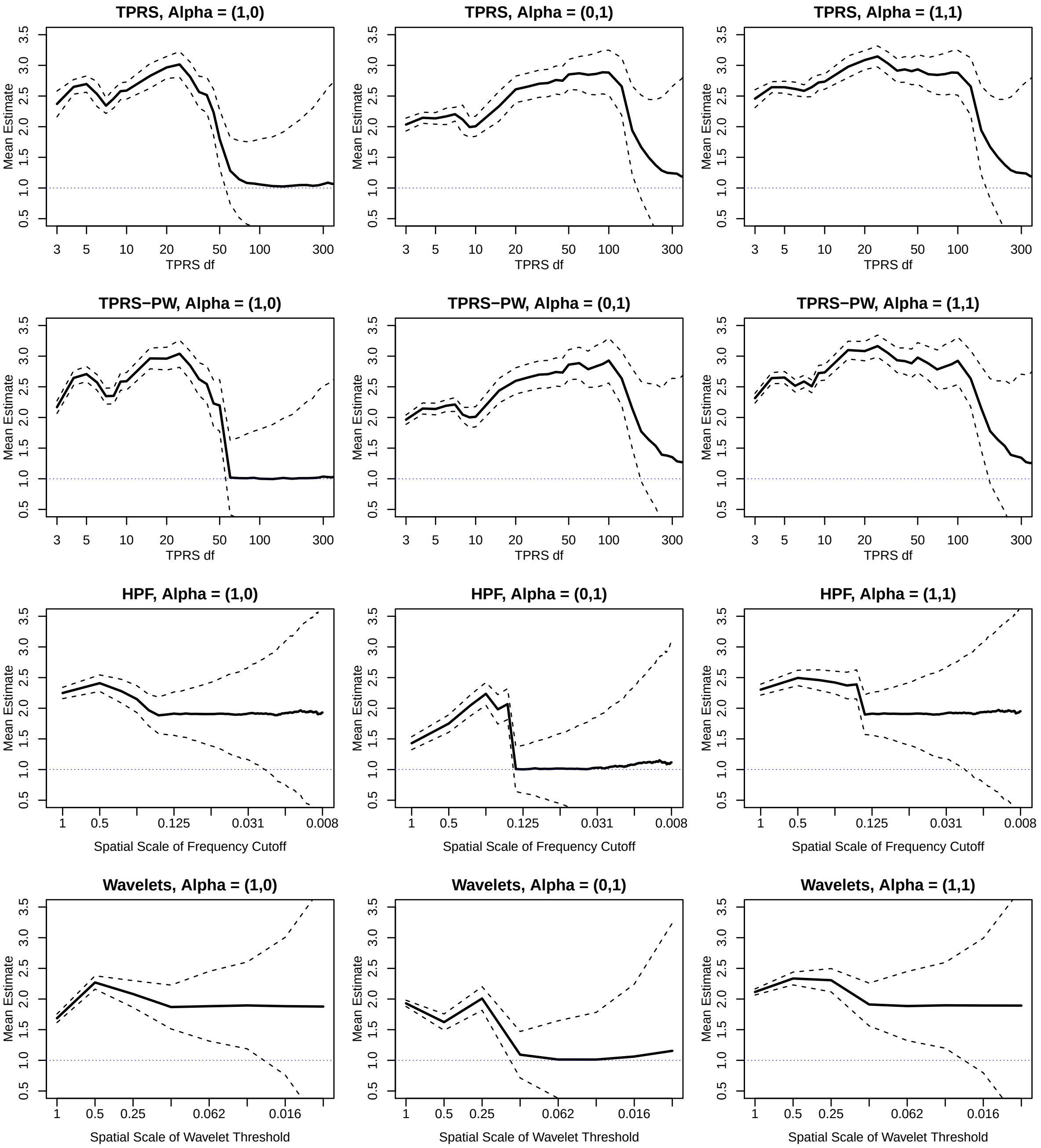


Figure 4.6: Estimates of β_1 from Simulation 2.

Table 4.2: Performance measures for different estimators in Simulation 2 for $(\alpha_1, \alpha_2) = (1,$ 1) setting. See Table 4.1 for an explanation of abbreviations.

	$\hat{\beta}$	Bias($\hat{\beta}$)	$SE(\hat{\beta})$	MSE($\hat{\beta}$)	$\widehat{SE}(\hat{\beta})$	Coverage
Unadjusted	2.06	1.056	0.020	1.1162	0.020	0.000
Adjusted	1.00	0.000	0.034	0.0011	0.034	0.950
HPF: $\omega = 1$	2.31	1.307	0.040	1.7098	0.039	0.000
HPF: $\omega = 2$	2.50	1.503	0.055	2.2608	0.056	0.000
HPF: $\omega = 4$	2.43	1.427	0.081	2.0426	0.084	0.000
HPF: $\omega = 16$	1.93	0.927	0.224	0.9102	0.225	0.017
HPF: $\omega = 32$	1.95	0.945	0.331	1.0034	0.321	0.176
HPF: $\omega = 128$	1.98	0.976	0.724	1.4767	0.722	0.720
TPRS: $df = 3$	2.32	1.315	0.037	1.7317	0.036	0.000
TPRS: $df = 5$	2.65	1.652	0.046	2.7318	0.045	0.000
TPRS: $df = 20$	3.09	2.091	0.072	4.3764	0.071	0.000
TPRS: $df = 50$	2.98	1.978	0.112	3.9252	0.112	0.000
TPRS: $df = 100$	2.92	1.924	0.174	3.7336	0.175	0.000
TPRS: $df = 200$	1.60	0.605	0.425	0.5455	0.429	0.718
TPRS: $df = 300$	1.28	0.278	0.602	0.4386	0.606	0.932
Wave: $L = 0$	2.12	1.117	0.022	1.2474	0.022	0.000
Wave: $L = 1$	2.34	1.341	0.047	1.8009	0.046	0.000
Wave: $L = 2$	2.32	1.316	0.082	1.7385	0.083	0.000
Wave: $L = 4$	1.91	0.907	0.246	0.8835	0.245	0.050
Wave: $L = 7$	1.95	0.952	0.824	1.5848	0.810	0.774

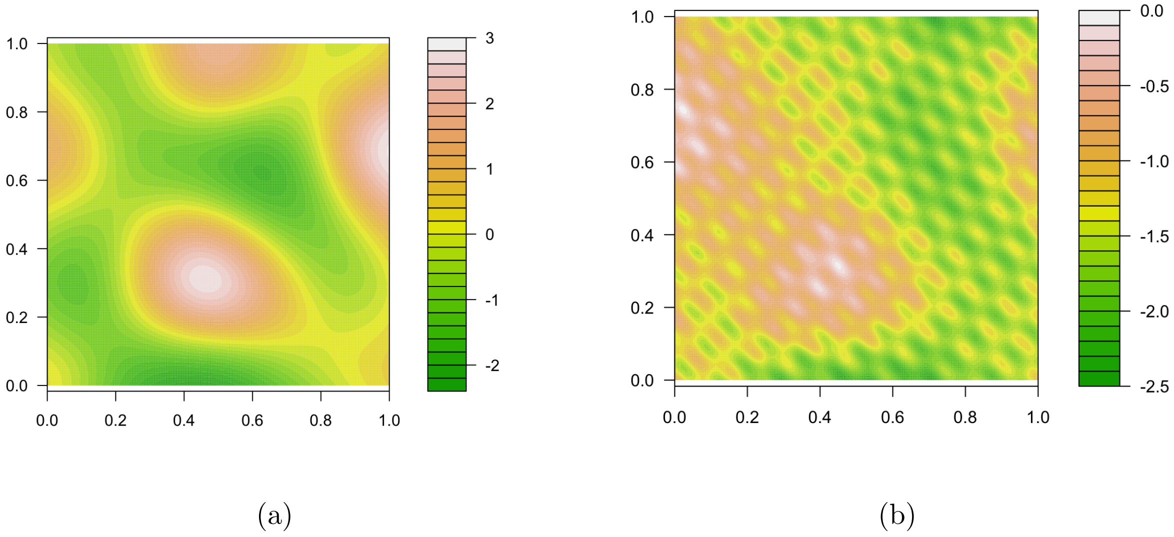


Figure 4.7: Surfaces (a) G_5 and (b) X (when $\alpha_1 = \alpha_2 = 1$) in Simulation 3.

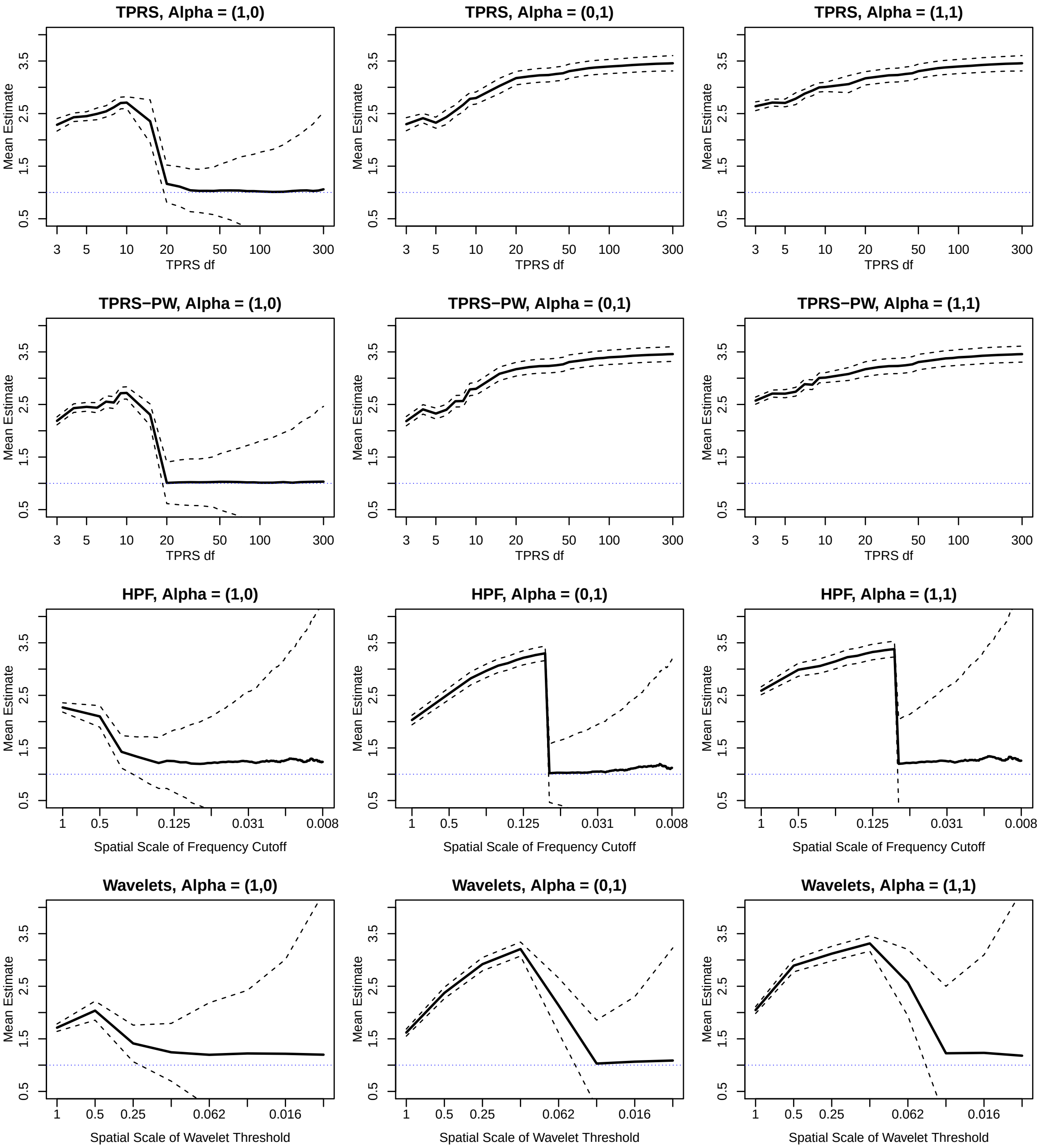


Figure 4.8: Estimates of β_1 from Simulation 3.

4.5 Combining Approaches to Pre-Adjustment

In the simulations of Section 4.4, we identified scenarios under which each pre-adjustment approach performed well and each performed poorly. The latter generally occurred when the spatial basis for of adjustment did not match, and could not well approximate, the spatial basis functions used to create the confounding surface. This leads us to consider combining pre-adjustment approaches. In this section, we present results from Simulations 2 and 3, now sequentially applying the pre-adjustment approaches.

4.5.1 Simulation 2: Combining Methods

We now return to Simulation 2, but compare the estimators that combine pre-adjustment approaches. For the setting when $\alpha_1 = \alpha_2 = 1$, (the confounder is a combination of TPRS and periodic surfaces), the results are shown in Figure 4.9. Based upon the results from applying a single method, one might expect that if we use the high-pass filter approach and TPRS adjustment, then all bias should be eliminated. However, Figure 4.9a shows that that is not necessarily the case. We see that for $df \leq 50$, higher ω adjustment results in the same or increased bias. Only for $df > 100$ does the bias start to go away. Similarly, bias increases for increasing df from 10 to 50 while ω is ≥ 32. While adding pre-adjustment to low frequency Fourier basis functions to exposure already pre-adjusted by TPRS does not change the bias by much, the converse is not true. This is notably apparent in the far left of Figure 4.9a, where when $\omega \geq 8$ and the exposure is also pre-adjusted with $df = 3$ TPRS, bias increases.

It is important to note that order of pre-adjustment matters. The estimates obtained by first filtering the exposure using Fourier bases followed by TPRS (Figure 4.9b) are different from the patterns seen when the exposure is first filtered by TPRS and Fourier bases (Figure 4.9a).

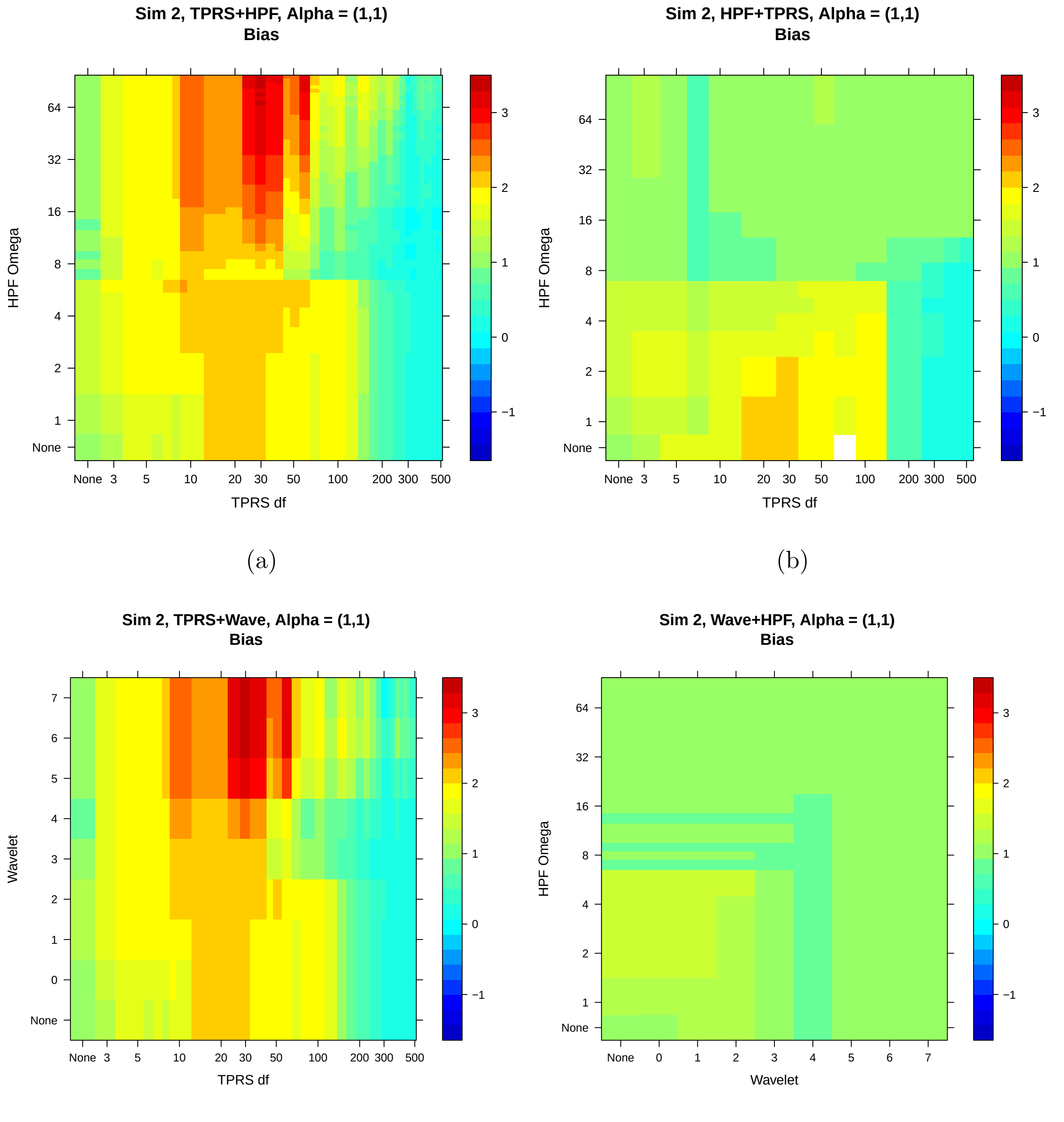
Combining wavelet thresholding and the high-pass filter does not change bias much (see Figure 4.9d), which is expected since they are each capturing periodic variability of similar scales, albeit of different forms.

When the confounder is entirely comprised of TPRS surfaces ($\alpha_1 = 1$ and $\alpha_2 = 0$) and we pre-adjust using TPRS and either Fourier or wavelet basis functions, the bias is largely unchanged relative to pre-adjustment by TPRS alone (see Figure 4.10). However, we see once again that order appears to matter importantly, with a high-pass filter approach that is followed by TPRS performing poorly (Figure 4.10b), even when the correct *df* is chosen.

When the confounder is entirely sinusoidal (i.e. $\alpha_1 = 0$ and $\alpha_2 = 1$), but we use multiple methods to pre-adjust, then additional bias is introduced. Figures 4.11a and 4.11b show that even if the correct frequency is chosen for the frequency filter approach, the addition of TPRS filtering (whether before or after the high-pass filter) can introduce bias. However, when Fourier basis functions and wavelets are both used to pre-adjust exposure, no additional bias is introduced (Figure 4.11d).

4.5.2 Simulation 3: Combining Methods

We now return to Simulation 3. In Simulation 3, we observed that when using a single pre-adjustment method, the high-pass filter and wavelet thresholding did much better than TPRS for the combined confounder surface. We find once again that combining approaches does not necessarily reduce bias, and in some cases can dramatically increase it. Figures C.4 and C.5 in the Appendix show this behavior in the $(\alpha_1, \alpha_2) = (1, 0)$ and (0, 1) settings. In Figure C.4a, we see that adding a small amount of TPRS adjustment (that is, TPRS adjustment with low *df*) to an exposure already pre-adjusted using a high-pass filter of Fourier basis markedly increases the bias.



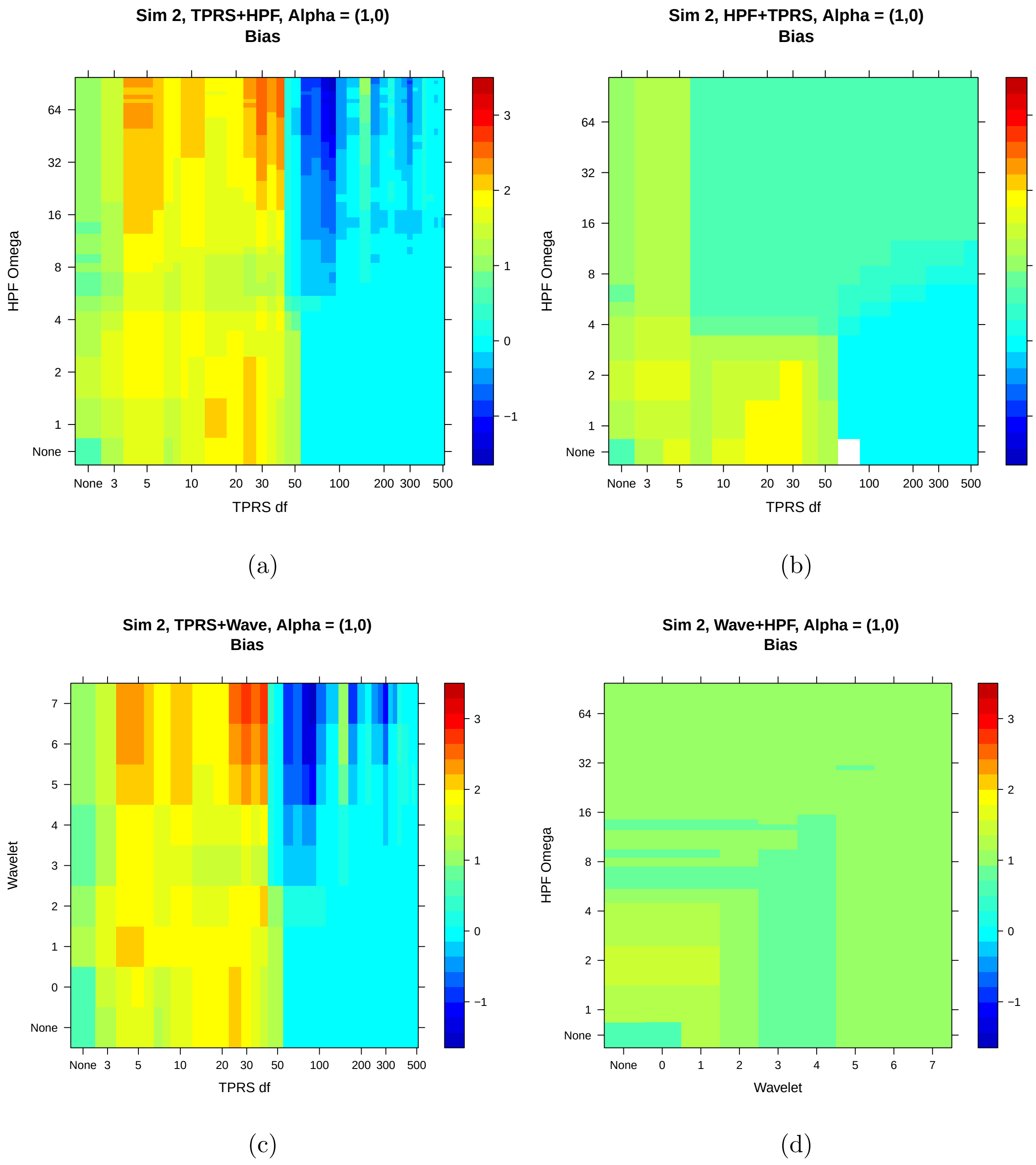
(a)

(b)

(c)

(d)

Figure 4.9: Estimates of β_1 from combining pre-adjustment approaches in Simulation 2, when $(\alpha_1, \alpha_2) = (1, 1)$. The order of pre-adjustment is given in each subfigure title. 'HPF' refers to a high-pass filter, which is pre-adjustment using Fourier basis functions. 'Wave' refers to wavelet thresholding.



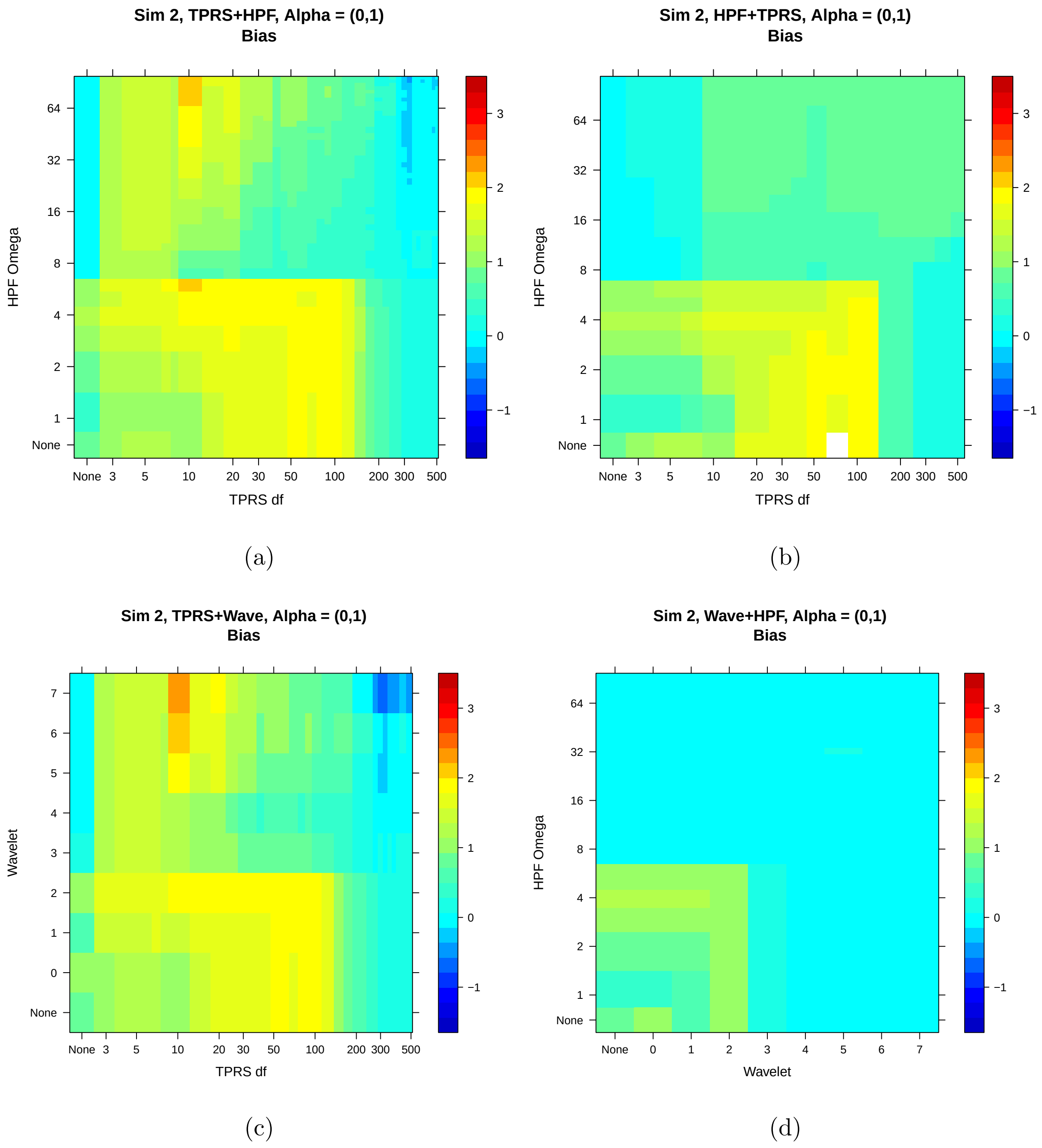
(a)

(b)

(c)

(d)

Figure 4.10: Estimates of β_1 from combining pre-adjustment approaches in Simulation 2, when $(\alpha_1, \alpha_2) = (1, 0)$.



(a)

(b)

(c)

(d)

Figure 4.11: Estimates of β_1 from combining pre-adjustment approaches in Simulation 2, when $(\alpha_1, \alpha_2) = (0, 1)$.

4.6 Sister Study Application

To better understand the nature of spatial confounding present in the Sister Study example, we apply the three pre-adjustment methods to the analysis of SBP and $\mathrm{PM}_{2.5}$.

Predictions from the $\mathrm{PM}_{2.5}$ exposure model of Sampson et al. (2013) are only defined over the continental U.S. To apply the pre-adjustment approaches with Fourier and wavelet basis functions, we must embed the gridded locations within a larger rectangular and square grid, respectively. We do this by adding points with exposure concentration of zero. This results in a grid of size 184 × 115 for the Fourier approach and 256 × 256 for the Wavelet approach. We present results for two version of pre-adjustment by TPRS: one using TPRS defined from only points with exposure predictions (i.e the irregularly shaped grid that covers the contiguous U.S.) and the other using TPRS defined from the points in the 184 × 115 grid from the Fourier analysis.

As an approximate means of comparing adjustment scales, we compare the residual variance of the exposure surface for these different pre-adjustment approaches. From Figure 4.12, we see pre-adjustment by Fourier basis covers the largest range of reduction in exposure variability, while adjustment by TPRS covers a lower, and narrower, range than wavelets. Exposure pre-adjustment at the $\omega = 40$, $L = 6$ and $df = 80$ scales all correspond to approximately the same residual variability in the exposure surface (0.74, 0.72, and 0.72, respectively).

The estimated associations when pre-adjusting $\mathrm{PM}_{2.5}$ by Fourier and wavelet basis functions are plotted in Figure 4.13. We see similar trends in the estimates from both approaches, with adjustment beyond $\omega = 3$ and $L = 1$ (spatial scales of 2,300 km and 1,5333 km, respectively) doing little to change the estimate. The point estimate for $\omega = 3$ is 1.15 mmHg per 10 $\mu g/m^3$ difference in $\mathrm{PM}_{2.5}$ and for $L = 1$ it is 1.51 mmHg. This stable behavior beyond these amounts of adjustment could be because the confounding bias has been completely removed, as in Simulation 1, with $(\alpha_1, \alpha_2) = (0, 1)$. Or it could be that the remaining confounding bias is due to spatial variability that cannot be well represented by periodic functions, as in Simulation 2.

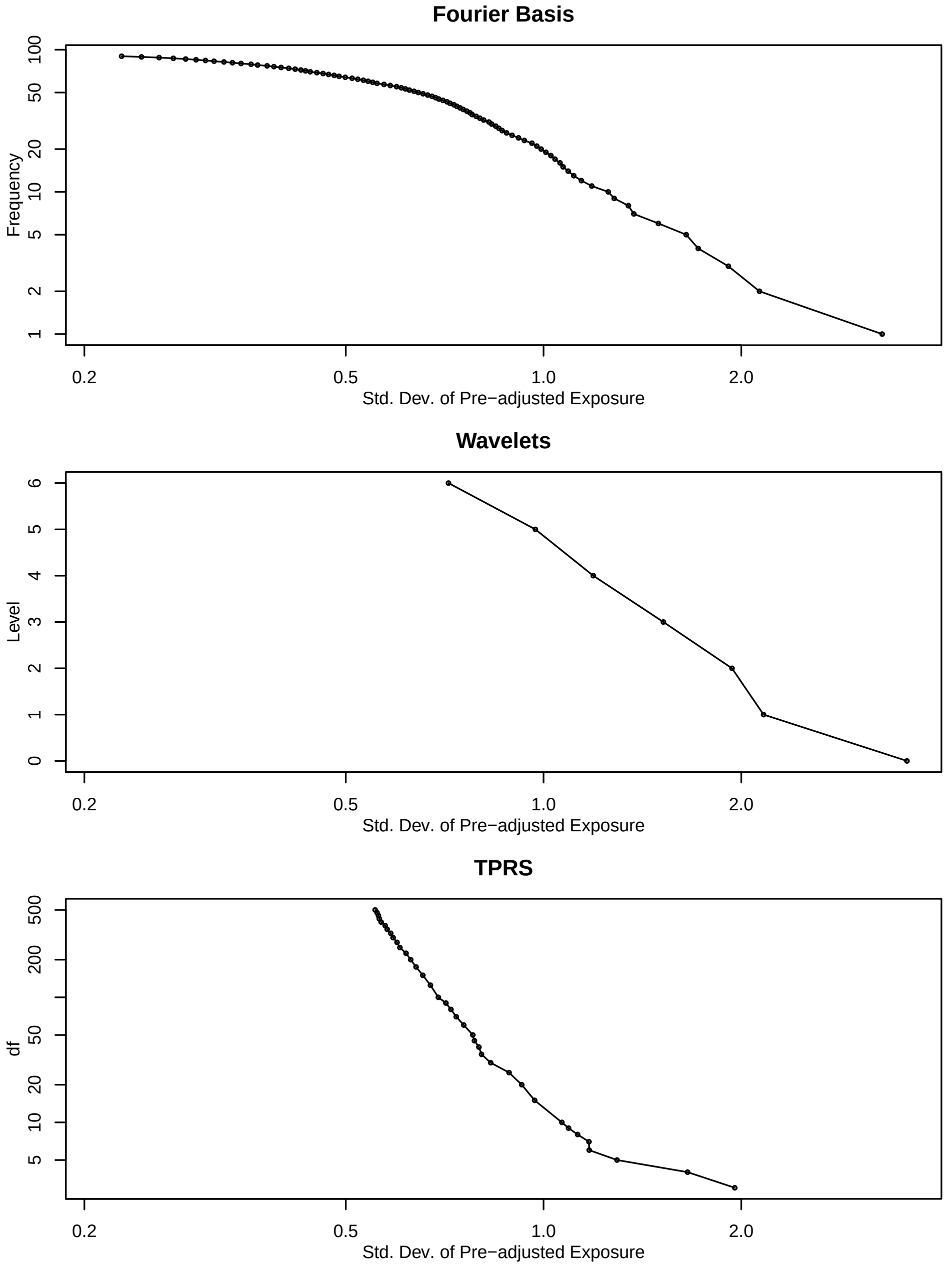


Figure 4.12: Variability remaining in the $PM_{2.5}$ exposure surface after pre-adjustment by each method. The TPRS values correspond to the basis functions derived from the irregularly-shaped grid.

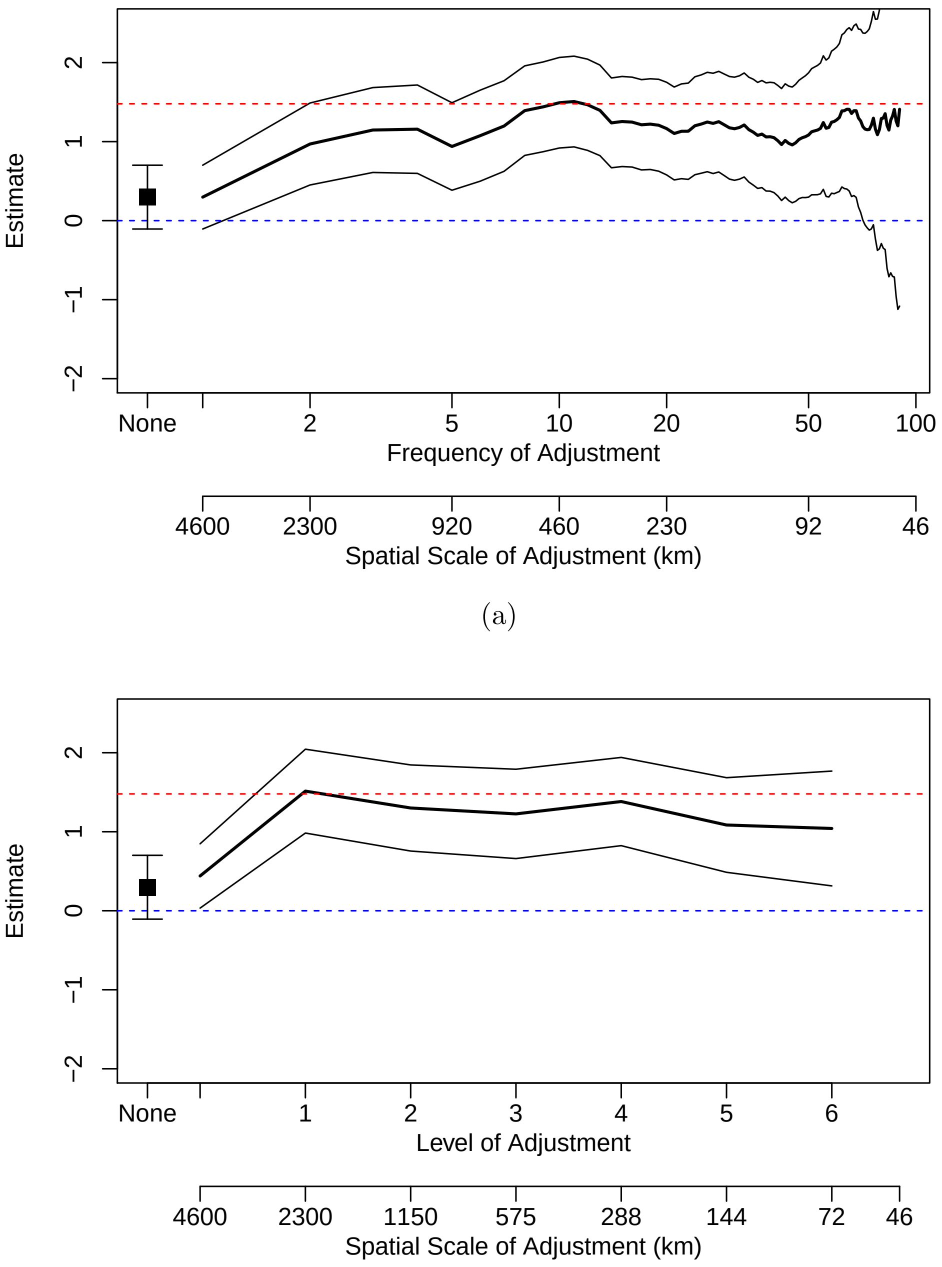
The results for adjusting for TPRS in the health model are provided in Figure 4.1. There, the estimated association was greatest when adjustment was made for TPRS with 12 *df*. The value corresponding to 10 *df* was contained within the confidence intervals for the estimates with 3 to 30 *df*. For $df \geq 50$, the estimated association is essentially zero. An important difference between the health model adjustment with TPRS and the pre-adjustment is the role of additional confounders in the model, which have may spatial structure.

When pre-adjusting by TPRS, we see a similar trend, with estimates fluctuating near 1.40 for *df* between 5 and 80 (see Figure 4.14a). Beyond $df = 80$, the estimate drops to around 0.80 mmHg. The TPRS pre-adjustment that used basis functions defined on the rectangular grid has a similar pattern, but approach zero for large amounts of adjustment. Compared to the adjustment in the health model, the pre-adjustment approach results in standard errors that are 10-25% smaller, because they take advantage of exposure information at locations without subjects.

An important limitation to the pre-adjustment approaches is that they treat each of the grid locations equally. The exposure is made orthogonal to spatial basis functions evaluated at all locations. If subjects are sampled uniformly across the grid, then on average the subjects exposures' will be 'stochastically orthogonal' to the spatial basis desired (Szpiro et al. 2014). But when subjects are sampled non-uniformly across space and there are grid points without assigned subjects, then the exposures may not be fully orthogonal to the spatial basis functions as desired.

4.7 Discussion

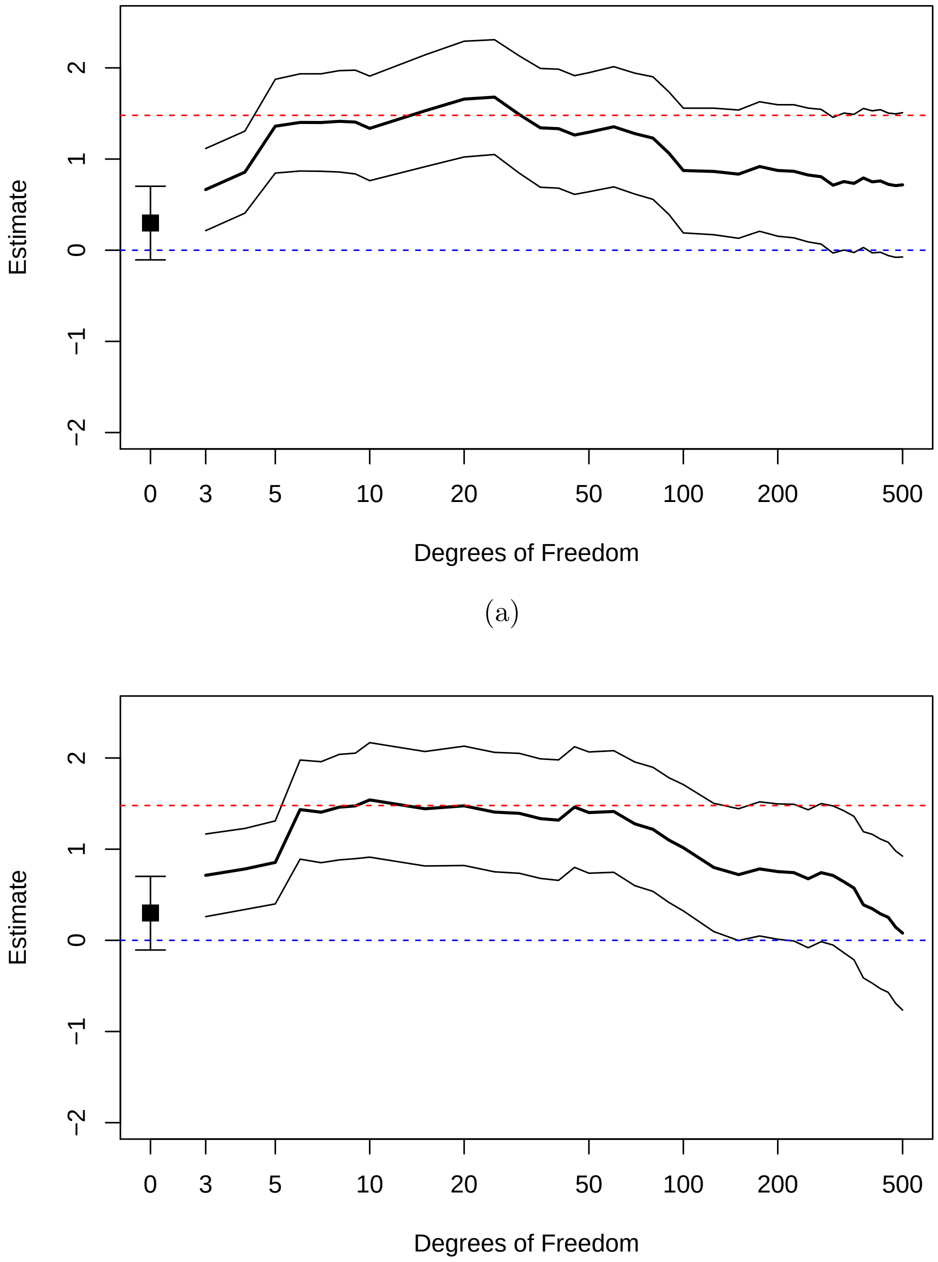
We have examined different approaches to adjusting for unmeasured spatial confounding. These approaches are motivated as extensions of time series methods for temporal confounding. In particular, we have explored alternatives to the common practice of using TPRS as a smoother in health models, by pre-adjusting exposure using TPRS, a Fourier basis, and wavelets. Each choice of spatial basis indexes variability in different forms and on different scales.



(a)

(b)

Figure 4.13: Estimated association between SBP and $PM_{2.5}$ for different amounts of pre-adjustment using (a) Fourier basis functions and (b) wavelet thresholding.



(a)

(b)

Figure 4.14: Estimated association between SBP and $PM_{2.5}$ for different amounts of pre-adjustment using TPRS defined on (a) grid locations across U.S. and (b) a rectangular grid.

We demonstrated how just using TPRS alone can perform quite poorly in models, and provides no interpretable spatial scale of adjustment. Yet we also showed that under other scenarios the high-pass (Fourier) filter approach and wavelet thresholding approaches can work just as poorly. The best solution may be to use some combination of these approaches, although even that is not guaranteed to eliminate bias.

One of the key results from our simulations is that the addition of spatial basis functions, or an equivalent pre-adjustment procedure, does not necessarily reduce bias in the point estimate. In fact, it can increase it substantially. Even in the simple setting of Simulation 1, when the confounder was a periodic surface with only two frequencies of variation, performing a high-pass filter for $\omega < 13$ gave progressively higher bias (Figure 4.4) until the confounder was fully adjusted for. In the same setting, pre-adjustment with TPRS at first decreased bias, but later increased it. This demonstrates how challenging it can be to guide a choice of spatial confounding adjustment based upon changes in a point estimate, since attempting to account for spatial confounding by adding spatial basis functions can *increase* confounding bias. These results are similar to the behavior described by Hodges and Reich (2010) for area data.

An even more surprising conclusion is that the combination of methods does not necessarily reduce confounding bias, even when each the adjustment bases match the basis functions of the confounding surface. In Simulation 3, we saw that adding a high-pass filter to an exposure that has already been completely unconfounded by appropriate TPRS pre-adjustment can increase bias (e.g., when $df > 15$ in Figure C.4a). This also held when the confounder was periodic and the order of adjustment was switched (Figure C.5a). This shows that unlike the conclusions of Paciorek (2010), we do not find that merely having finer scale variability in the exposure is sufficient to allow for the elimination of bias. In addition to scale, the choice of basis function is critically important.

Pre-adjusting the exposure by wavelets is limited severely by the requirement of a square grid to apply the DWT. The thresholding at dyadic levels leads to limited available scales for adjustment. However, the thresholding could be done non-uniformly, thresholding certain spatial regions to higher or lower levels than the remainder of the exposure surface. This flexibility means that the pre-adjustment could be tuned in accordance with known structure in the exposure surface. Extensions of wavelet methods to non-regular data in two dimensions exists (Jansen et al. 2009), however the analogue of wavelet pre-adjustment in those settings is not clear.

The development of an automated procedure for choosing the extent of adjustment, e.g., by minimizing estimated MSE, is attractive in principle. However, given the difficulties in removing confounding bias even when the adjustment basis is known to match the confounding surface, it is not clear if such an automated procedure would be of any practical use.

Our recommendation is that the extent of adjustment should be selected *a priori*. With the Fourier and wavelet approaches, this can be done in concert with a selection of spatial scale, but with TPRS is must be done in terms of *df*. We recommend this because, without precise knowledge of the unmeasured confounding surface, there is no guarantee that any amount of adjustment will yield an unbiased estimator. One could explore a sequence of scales as a sensitivity analysis, but choosing a 'knee' in a plot is not guaranteed to give good performance and can be greatly effected by variability in a particular dataset. Choosing the largest amount of pre-adjustment possible is one strategy, but this results in estimates that are highly variable and may in fact not have the smallest bias possible for a chosen basis.

In applications such as large scale air pollution epidemiology, we are often most concern with broad, regional-scale spatial confounding. The methods we have presented provide multiple approaches to adjusting for such confounding in a flexible manner.

4.8 References

Besag, J., J. York, and A Mollié. 1991. “Bayesian image restoration with two applications in spatial statistics”. *Annals of the Institute of Statistical Mathematics* 43 (1): 1–20.

Brochu, P. J., J. D. Yanosky, C. J. Paciorek, J. Schwartz, J. T. Chen, R. F. Herrick, and H. H. Suh. 2011. “Particulate air pollution and socioeconomic position in rural and urban areas of the Northeastern United States”. *American Journal of Public Health* 101 (SUPPL. 1): 224–230.

Burger, W., and J. M. Burge. 2009. *Principles of Digital Image Processing.* Springer.

Burnett, R., and D. Krewski. 1991. “Air pollution effects on hospital admission rates : A random effects modeling approach”. 22 (4): 441–458.

Carl, G., and I. Kühn. 2008. “Analyzing spatial ecological data using linear regression and wavelet analysis”. *Stochastic Environmental Research and Risk Assessment* 22 (3): 315–324.

Chan, S. H., V. C. van Hee, S. Bergen, A. A. Szpiro, L. A. DeRoo, S. J. London, J. D. Marshall, J. D. Kaufman, and D. P. Sandler. 2015. “Long-term air pollution exposure and blood pressure in the Sister Study”. *Environmental Health Perspectives* 123 (10): 951–958.

Clayton, D. G., L Bernardinelli, and C Montomoli. 1993. “Spatial correlation in ecological analysis.” *International journal of epidemiology* 22 (6): 1193–1202.

Clayton, D., and J. Kaldor. 1987. “Empirical Bayes estimates of age-standardized relative risks for use in disease mapping.” *Biometrics* 43 (3): 671–681.

Cressie, N. 1993. *Statistics for Spatial Data.* Revised Ed. Hoboken, NJ: John Wiley & Sons.

Daubechies, I. 1988. “Orthonormal Bases of Compactly Supported Wavelets II. Variations on a Theme”. *SIAM Journal on Mathematical Analysis* 24 (2): 499–519.

Diez Roux, A. V., S. S. Merkin, D Arnett, L Chambless, M Massing, F. J. Nieto, P Sorlie, M Szklo, H. A. Tyroler, and R. L. Watson. 2001. "Neighborhood of residence and incidence of coronary heart disease". *N Engl J Med* 345 (2): 99–106.

Dominici, F, A McDermott, S. L. Zeger, and J. M. Samet. 2002. "On the use of generalized additive models in time-series studies of air pollution and health". *American Journal of Epidemiology* 156 (3): 193–203.

Dominici, F., A. Mcdermott, S. L. Zeger, and J. M. Samet. 2003. "Airborne particulate matter and mortality: timescale effects in four US cities." *American Journal of Epidemiology* 157 (12): 1055–65.

Dominici, F., A. McDermott, and T. J. Hastie. 2004. "Improved Semiparametric Time Series Models of Air Pollution and Mortality". *Journal of the American Statistical Association* 99 (468): 938–948.

Gee, G. C., and D. C. Payne-Sturges. 2004. "Environmental health disparities: A framework integrating psychosocial and environmental concepts". *Environmental Health Perspectives* 112 (17): 1645–1653.

Haining, R. 1991. "Bivariate Correlation with Spatial Data". *Geographical Analysis* 23 (3): 210–227.

Hanks, E. M., E. M. Schliep, M. B. Hooten, and J. a. Hoeting. 2015. "Restricted spatial regression in practice: geostatistical models, confounding, and robustness under model misspecification". *Environmetrics.*

Hastie, T. J., and R. Tibshirani. 1990. *Generalized Additive Models.* New York: Chapman & Hall.

Hodges, J. S., and B. J. Reich. 2010. "Adding Spatially-Correlated Errors Can Mess Up the Fixed Effect You Love". *The American Statistician* 64 (4): 325–334.

Hughes, J., and M. Haran. 2013. “Dimension reduction and alleviation of confounding for spatial generalized linear mixed models”. *Journal of the Royal Statistical Society. Series B: Statistical Methodology* 75:139–159. arXiv: `1011.6649`.

Jansen, M., G. P. Nason, and B. W. Silverman. 2009. “Multiscale methods for data on graphs and irregular multidimensional situations”. *Journal of the Royal Statistical Society. Series B: Statistical Methodology* 71 (1): 97–125.

Jerrett, M, R. T. Burnett, J Brook, P Kanaroglou, C Giovis, N Finkelstein, and B Hutchison. 2004. “Do socioeconomic characteristics modify the short term association between air pollution and mortality? Evidence from a zonal time series in Hamilton, Canada.” *Journal of epidemiology and community health* 58 (1): 31–40.

Nason, G. P. 2008. *Wavelet Methods in Statistics with R*. New York: Springer.

Paciorek, C. J. 2010. “The importance of scale for spatial-confounding bias and precision of spatial regression estimators.” *Statistical Science* 25 (1): 107–125.

Ramsay, T., R. Burnett, and D. Krewski. 2003a. “Exploring bias in a generalized additive model for spatial air pollution data”. *Environmental Health Perspectives* 111 (10): 1283–1288.

Ramsay, T. O., R. T. Burnett, and D. Krewski. 2003b. “The Effect of Concurvity in Generalized Additive Models Linking Mortality to Ambient Particulate Matter”. *Epidemiology* 14 (1): 18–23.

Reich, B. J., J. S. Hodges, and V. Zadnik. 2006. “Effects of residual smoothing on the posterior of the fixed effects in disease-mapping models”. *Biometrics* 62 (December): 1197–1206.

Sampson, P. D., M. Richards, A. A. Szpiro, S. Bergen, L. Sheppard, T. V. Larson, and J. D. Kaufman. 2013. “A regionalized national universal kriging model using Partial Least Squares regression for estimating annual PM2.5 concentrations in epidemiology”. *Atmospheric Environment* 75:383–392.

Schwartz, J. 1994. “Nonparametric smoothing in the analysis of air pollution and respiratory illness”. *Canadian journal of statistics* 22 (4): 471–487.

Szpiro, A. A., and C. J. Paciorek. 2013. “Measurement error in two-stage analyses, with application to air pollution epidemiology”. *Environmetrics* 24:501–517.

Szpiro, A. A., L. Sheppard, S. D. Adar, and J. D. Kaufman. 2014. “Estimating acute air pollution health effects from cohort study data”. *Biometrics* 70 (1): 164–174.

Tiefelsdorf, M., and D. a. Griffith. 2007. “Semiparametric filtering of spatial autocorrelation: The eigenvector approach”. *Environment and Planning A* 39 (5): 1193–1221.

Wakefield, J. 2007. “Disease mapping and spatial regression with count data”. *Biostatistics* 8 (2): 158–183.

Wand, M. P., and J. T. Ormerod. 2011. “Penalized wavelets: Embedding wavelets into semiparametric regression”. *Electronic Journal of Statistics* 5:1654–1717.

Wood, S. N. 2003. “Thin plate regression splines”. *Journal of the Royal Statistical Society: Series B* 65 (1): 95–114.

Appendix A

APPENDIX FOR CHAPTER 2

A.1 EM Algorithm for Predictive K-means

Here we outline the steps of the EM algorithm used to optimize the log-likelihood function for predictive k-means. For ease of notation, we omit asterisks (*), which were used in the main text to signify data from monitoring locations (as opposed to cohort locations). We closely follow the procedure described in Jordan and Jacobs (1994).

The log-likehood for the observed data is given by equation (5) in the main text, which we repeat here:

$$\ell\left(\boldsymbol{\Gamma}, \boldsymbol{M}, \sigma^2 \middle| \boldsymbol{X}; \boldsymbol{R}\right) = \sum_{i=1}^{n} \log\left(\sum_{k=1}^{K} q_k(\boldsymbol{r}_i, \boldsymbol{\Gamma}) \times \mathcal{N}_p\left(\boldsymbol{x}_i | \boldsymbol{\mu}_k, \sigma^2 \boldsymbol{I}\right)\right),$$

where $\boldsymbol{M} = \begin{bmatrix} \boldsymbol{\mu}_1 & \cdots & \boldsymbol{\mu}_K \end{bmatrix}$ and $\mathcal{N}_p(\cdot|\boldsymbol{m}, \boldsymbol{V})$ is the density function for a p-dimensional multivariate normal distribution with mean $\boldsymbol{m}$ and variance $\boldsymbol{V}$. Recall that $q_k(\boldsymbol{r}_i, \boldsymbol{\Gamma}) = \Pr(z_{ik} = 1; \boldsymbol{\Gamma}, \boldsymbol{r}_i)$ is the softmax function (equation (3) in the main text), where we introduce the indicator variables

$$z_{ik} = \begin{cases} 1 & \text{if } Z_i = k \\ 0 & \text{otherwise.} \end{cases}$$

The complete data log-likelihood function, when observing z_{ik}, is:

$$\ell_c\left(\boldsymbol{\Gamma}, \boldsymbol{M}, \sigma^2 \middle| \boldsymbol{X}, \boldsymbol{z}; \boldsymbol{R}\right) = \sum_{i=1}^{n} \sum_{k=1}^{K} z_{ik} \left[\log q_k(\boldsymbol{r}_i, \boldsymbol{\Gamma}) + \log \mathcal{N}_p\left(\boldsymbol{x}_i | \boldsymbol{\mu}_k, \sigma^2 \boldsymbol{I}\right)\right]$$

The $(j+1)$th E-step of the EM algorithm amounts to finding

$$\begin{aligned}
&\mathrm{E}\Big[\ell_c(\boldsymbol{\Gamma}, \boldsymbol{M}, \sigma^2|\boldsymbol{X}, \boldsymbol{z}; \boldsymbol{R})\Big|\boldsymbol{\Gamma}^{(j)}, \boldsymbol{M}^{(j)}, \sigma^{2(j)}, \boldsymbol{X}; \boldsymbol{R}\Big] \\
&\qquad\qquad = \mathrm{E}\Bigg[\sum_{i=1}^{n}\sum_{k=1}^{K} z_{ik}\Big[\log q_k(\boldsymbol{r}_i, \boldsymbol{\Gamma}) \\
&\qquad\qquad\qquad + \log \mathcal{N}_p(\boldsymbol{x}_i|\boldsymbol{\mu}_k, \sigma^2\boldsymbol{I})\Big]\Bigg|\boldsymbol{\Gamma}^{(j)}, \boldsymbol{M}^{(j)}, \sigma^{2(j)}, \boldsymbol{X}; \boldsymbol{R}\Bigg] \\
&\qquad\qquad = \sum_{i=1}^{n}\sum_{k=1}^{K} \mathrm{E}\left[z_{ik}\Big|\boldsymbol{\Gamma}^{(j)}, \boldsymbol{M}^{(j)}, \sigma^{2(j)}, \boldsymbol{X}; \boldsymbol{R}\right] \\
&\qquad\qquad\qquad \times \left[\log q_k(\boldsymbol{r}_i, \boldsymbol{\Gamma}) + \log \mathcal{N}_p\left(\boldsymbol{x}_i\middle|\boldsymbol{\mu}_k, \sigma^2\boldsymbol{I}\right)\right].
\end{aligned}$$

To do this, we need only compute

$$\begin{aligned}
h_{ik}^{(j+1)} &:= \mathrm{E}\left[z_{ik}\Big|\boldsymbol{\Gamma}^{(j)}, \boldsymbol{M}^{(j)}, \sigma^{2(j)}, \boldsymbol{X}; \boldsymbol{R}\right] \\
&= P\left(z_{ik}\Big|\boldsymbol{\Gamma}^{(j)}, \boldsymbol{M}^{(j)}, \sigma^{2(j)}, \boldsymbol{X}; \boldsymbol{R}\right) \\
&= \frac{\mathcal{N}_p\Big(\boldsymbol{x}_i\Big|\boldsymbol{\mu}_k^{(j)}, \sigma^{2(j)}\boldsymbol{I}\Big)q_k\left(\boldsymbol{r}_i, \boldsymbol{\Gamma}^{(j)}\right)}{\sum_{k'=1}^{K}\mathcal{N}_p\Big(\boldsymbol{x}_i\Big|\boldsymbol{\mu}_{k'}^{(j)}, \sigma^{2(j)}\boldsymbol{I}\Big)q_{k'}\left(\boldsymbol{r}_i, \boldsymbol{\Gamma}^{(j)}\right)}.
\end{aligned}$$

This is followed by the $(j+1)$th M-step, in which we find

$$\begin{aligned}
&\Big(\boldsymbol{\Gamma}^{(j+1)}, \boldsymbol{M}^{(j+1)}, \sigma^{2(j+1)}\Big) \\
&\qquad = \arg\max \mathrm{E}\left[\ell_c(\boldsymbol{\Gamma}, \boldsymbol{M}, \sigma^2|\boldsymbol{X}, \boldsymbol{z}; \boldsymbol{R})\middle|\boldsymbol{\Gamma}^{(j)}, \boldsymbol{M}^{(j)}, \sigma^{2(j)}, \boldsymbol{X}; \boldsymbol{R}\right].
\end{aligned}$$

This can be done in two separate steps, by computing

$$\left(\boldsymbol{M}^{(j+1)}, \sigma^{2(j+1)}\right) = \underset{\boldsymbol{M},\sigma^2}{\arg\max}\sum_{i=1}^{n}\sum_{k=1}^{K} h_{ik}^{(j+1)} \log \mathcal{N}_p\left(\boldsymbol{x}_i|\boldsymbol{\mu}_k, \sigma^2 I\right), \tag{A.1}$$

and

$$\boldsymbol{\Gamma}^{(j+1)} = \underset{\boldsymbol{\Gamma}}{\arg\max}\sum_{i=1}^{n}\sum_{k=1}^{K} h_{ik}^{(j+1)} \log q_k(\boldsymbol{r}_i, \boldsymbol{\Gamma}). \tag{A.2}$$

The optimization problem in (A.2) is equivalent to solving a multinomial logistic regression problem and can be accomplished using an iteratively reweighted least squares algorithm. To solve the optimization in (A.1), note that for any value of σ^2,

$$\begin{aligned}\boldsymbol{\mu}_k^{(j+1)} &= \arg\max_{\boldsymbol{\mu}_k} \sum_{i=1}^{n} h_{ik}^{(j+1)} \log \mathcal{N}_p(\boldsymbol{x}_i | \boldsymbol{\mu}_k, \sigma^2 \boldsymbol{I}) \\ &= \arg\max_{\boldsymbol{\mu}_k} \sum_{\ell=1}^{p} \sum_{i=1}^{n} -h_{ik}^{(j+1)} (x_{i\ell} - (\boldsymbol{\mu}_k)_\ell)^2.\end{aligned}$$

Each entry of $\boldsymbol{\mu}_k^{(j+1)}$ can thus be computed as a weighted average:

$$\left(\boldsymbol{\mu}_k^{(j+1)}\right)_\ell = \arg\max_{m} \sum_{i=1}^{n} -h_{ik}^{(j+1)} (x_{i\ell} - m)^2 = \frac{\displaystyle\sum_{i=1}^{n} h_{ik}^{(j+1)} x_{i\ell}}{\displaystyle\sum_{i=1}^{n} h_{ik}^{(j+1)}}.$$

Lastly, $\sigma^{2(j+1)}$ is computed as

$$\begin{aligned}\sigma^{2(j+1)} &= \arg\max_{\sigma^2} \sum_{i=1}^{n} \sum_{k=1}^{K} h_{ik}^{(j+1)} \log \mathcal{N}_p \left(\boldsymbol{x}_i \middle| \boldsymbol{\mu}_k^{(j+1)}, \sigma^2 \boldsymbol{I} \right) \\ &= \frac{\displaystyle\sum_{i=1}^{n} \sum_{k=1}^{K} h_{ik}^{(j+1)} \left(\boldsymbol{x}_i - \boldsymbol{\mu}_k^{(j+1)} \right)^{\mathsf{T}} \left(\boldsymbol{x}_i - \boldsymbol{\mu}_k^{(j+1)} \right)}{\displaystyle\sum_{i=1}^{n} \sum_{k=1}^{K} p h_{ik}^{(j+1)}}. \qquad \text{(A.3)}\end{aligned}$$

A.2 Cross-validation results for PM$_{2.5}$ data

This section provides additional results from the 10-fold cross-validation of the analysis of the PM$_{2.5}$ component data. Among the models considered, the model with lowest cross-validated *MSPE* had $K = 8$ clusters, 2 PCA componnts and TPRS with 10 df as the covariates, and an SVM as the classifier. Table A.1 provides the complete set of cross-validation metrics for this model, for both predictive k-means and regular k-means and all three prediction methods.

Table A.1: Measures of clustering performance from 10-fold cross-validation of the PM$_{2.5}$ component data when $K = 8$ and the covariates are 2 PCA components and TPRS with 10 df. The measures of performance (*MSPE*, *wSS*, *MSME*, and *Acc*) are described in Section 3.3 of the main text.

Clustering Method	Prediction Method	*MSPE*	*wSS*	*MSME*	*Acc*
k-means	Multinom Logit	20.10	14.46	8.96	0.67
	SVM	18.95	14.46	7.06	0.68
Predictive k-means	Multinom Logit	21.28	14.88	9.90	0.62
with $\hat{\sigma}^2$ selected by EM	SVM	18.33	14.88	5.97	0.70
	ME-Working	24.33	14.88	13.00	0.61

For comparing performance between different numbers of clusters, we present in Table A.2 the best-performing (in terms of cross-validated *MSPE*) model for each choice of K between 4 and 10.

Table A.2: Cross-validation results for the best-performing (in terms of *MSPE*) predictive k-means model for different numbers of clusters K. The measures of performance (*MSPE*, *wSS*, *MSME*, and *Acc*) are described in Section 3.3 of the main text.

K	Covariate Model	Prediction Method	*MSPE*	*wSS*	*MSME*	*Acc*
4	2 PCA, 10 df TPRS	SVM	20.19	17.03	5.19	0.75
5	2 PCA, 10 df TPRS	SVM	19.49	15.69	7.49	0.68
6	3 PCA, 5 df TPRS	SVM	19.52	15.52	6.33	0.74
7	3 PCA, 5 df TPRS	SVM	19.45	15.17	6.94	0.72
8	2 PCA, 10 df TPRS	SVM	18.33	14.88	5.97	0.70
9	2 PCA, 10 df TPRS	SVM	19.50	14.40	8.27	0.60
10	2 PCA, 10 df TPRS	SVM	18.69	13.92	7.43	0.62

A.3 Sensitivity Analysis: Health Results for Different K

This section presents the health analysis results for varying numbers of clusters K. See Table A.2 for a summary of the covariates in each model. We present a map of predicted cluster membership for Sister Study locations and a table of health effect estimates for each cluster. Throughout this section, we label the clusters from 1 to K for simplicity, but these labels do not necessarily have any relation to Clusters 1 through 8 described in the main text. For $K \geq 6$, one cluster comprised a single monitor and no health effects were estimated for that cluster.

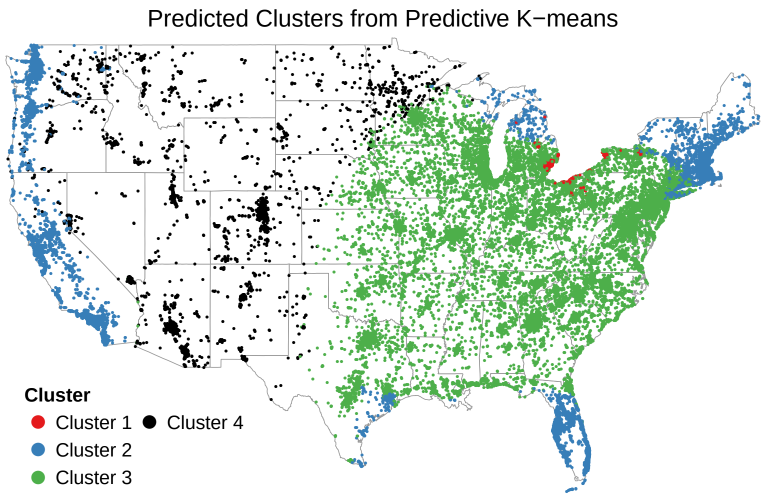


Figure A.1: Predicted cluster membership at Sister Study cohort locations for $K = 4$ (jittered to protect confidentiality).

Table A.3: Estimated difference in SBP (in mmHg) associated with a 10 $\mu g/m^3$ difference in annual ambient $\text{PM}_{2.5}$ exposure, $K = 4$.

Exposure	n	Est.	95% CI	p-value
$\text{PM}_{2.5}$ by Cluster				0.007
Cluster 1	1,093	6.83	(1.62, 12.0)	0.01
Cluster 2	13,447	1.54	(−0.37, 3.45)	0.11
Cluster 3	27,877	3.30	(1.84, 4.75)	0.00001
Cluster 4	4,789	−0.82	(−3.23, 1.60)	0.51

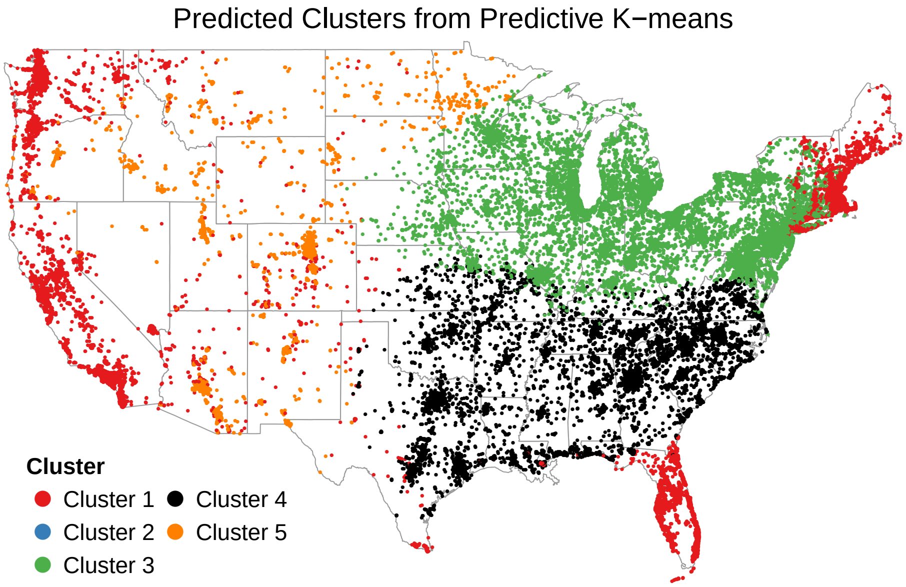


Figure A.2: Predicted cluster membership at Sister Study cohort locations for $K = 5$ (jittered to protect confidentiality).

Table A.4: Estimated difference in SBP (in mmHg) associated with a 10 $\mu g/m^3$ difference in annual ambient $\mathrm{PM}_{2.5}$ exposure, $K = 5$.

Exposure	n	Est.	95% CI	p-value
$\mathrm{PM}_{2.5}$ by Cluster				0.018
Cluster 1	13,477	1.24	(−0.53, 3.02)	0.17
Cluster 2	10	−0.29	(−63.0, 57.3)	0.93
Cluster 3	19,961	3.37	(1.90, 4.85)	<0.0001
Cluster 4	9,951	2.78	(0.11, 5.46)	0.04
Cluster 5	3,807	−1.79	(−4.58, 0.99)	0.21

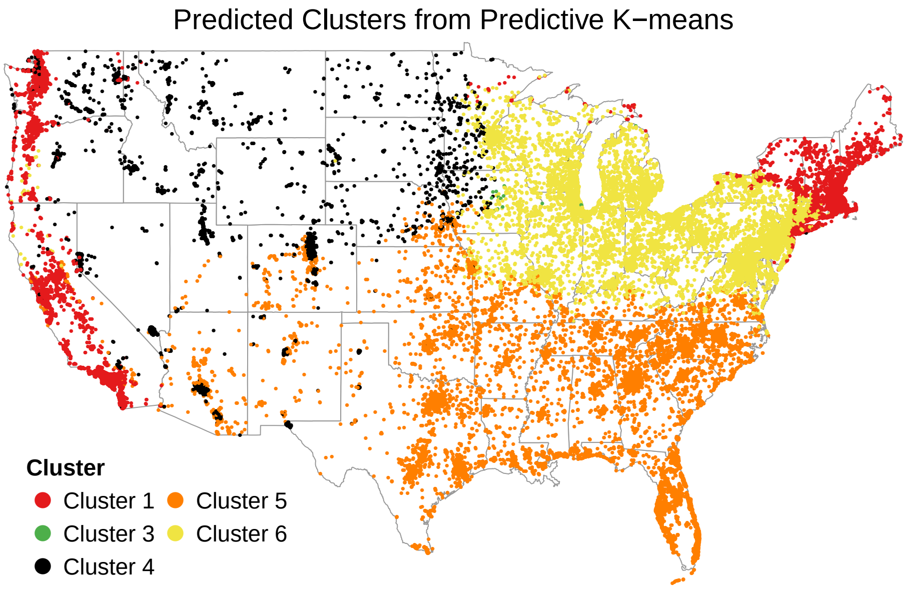


Figure A.3: Predicted cluster membership at Sister Study cohort locations for $K = 6$ (jittered to protect confidentiality).

Table A.5: Estimated difference in SBP (in mmHg) associated with a 10 $\mu g/m^3$ difference in annual ambient $\mathrm{PM}_{2.5}$ exposure, $K = 6$.

Exposure	n	Est.	95% CI	p-value
$\mathrm{PM}_{2.5}$ by Cluster				0.12
Cluster 1	9,894	3.44	(1.22, 5.66)	0.002
Cluster 2	–	–	–	–
Cluster 3	51	−0.24	(−24.6, 24.1)	0.98
Cluster 4	4,618	−0.81	(−3.26, 1.64)	0.52
Cluster 5	14,554	2.12	(0.29, 3.95)	0.02
Cluster 6	18,089	2.69	(1.16, 4.23)	0.0006

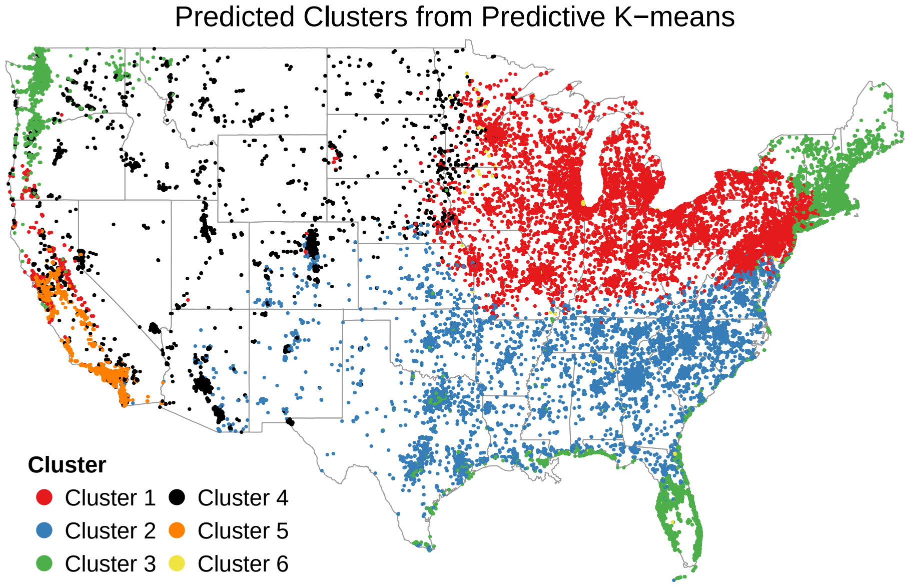


Figure A.4: Predicted cluster membership at Sister Study cohort locations for $K = 7$ (jittered to protect confidentiality).

Table A.6: Estimated difference in SBP (in mmHg) associated with a 10 $\mu g/m^3$ difference in annual ambient PM$_{2.5}$ exposure, $K = 7$.

Exposure	n	Estimate	95% CI	p-value
PM$_{2.5}$ by Cluster				0.015
Cluster 1	18,928	2.31	(0.84, 3.79)	0.002
Cluster 2	9,609	3.30	(1.19, 5.41)	0.002
Cluster 3	9,933	−1.01	(−3.61, 1.58)	0.44
Cluster 4	5,061	−0.46	(−2.77, 1.86)	0.70
Cluster 5	3,359	3.07	(−0.62, 6.76)	0.10
Cluster 6	316	4.58	(−2.21, 11.37)	0.19
Cluster 7	–	–	–	–

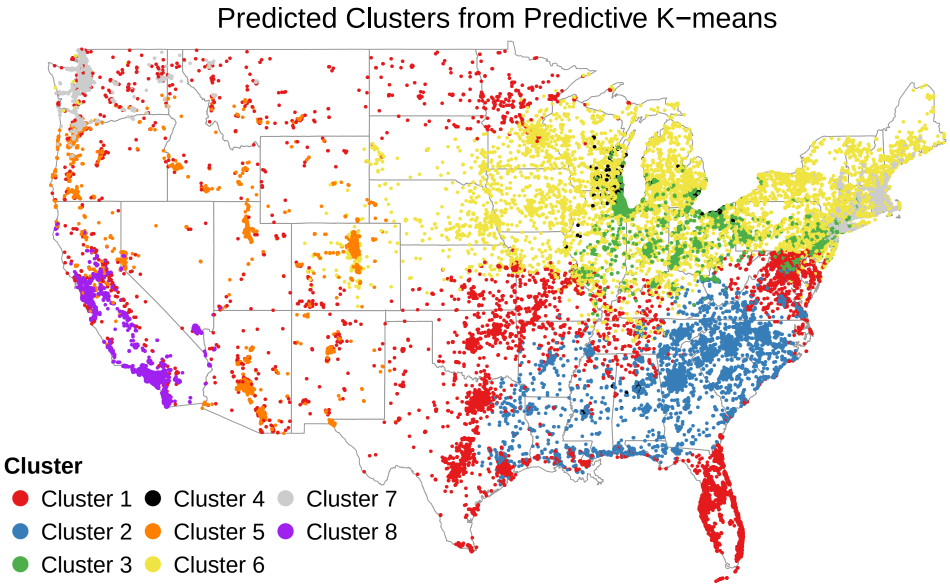


Figure A.5: Predicted cluster membership at Sister Study cohort locations for $K = 9$ (jittered to protect confidentiality).

Table A.7: Estimated difference in SBP (in mmHg) associated with a 10 $\mu g/m^3$ difference in annual ambient $\text{PM}_{2.5}$ exposure, $K = 9$.

Exposure	n	Estimate	95% CI	p-value
$\text{PM}_{2.5}$ by Cluster				0.024
Cluster 1	9,574	−0.78	(−3.20, 1.64)	0.53
Cluster 2	6,097	2.89	(−0.38, 6.16))	0.08
Cluster 3	9,122	1.05	(−1.03, 3.13)	0.32
Cluster 4	489	6.39	(−2.57, 15.4)	0.16
Cluster 5	3,681	−2.50	(−5.44, 0.44)	0.095
Cluster 6	9,837	1.30	(−0.60, 3.21)	0.18
Cluster 7	4,555	3.95	(−0.87, 8.77)	0.11
Cluster 8	3,851	3.82	(0.96, 6.69)	0.009
Cluster 9	–	–	–	

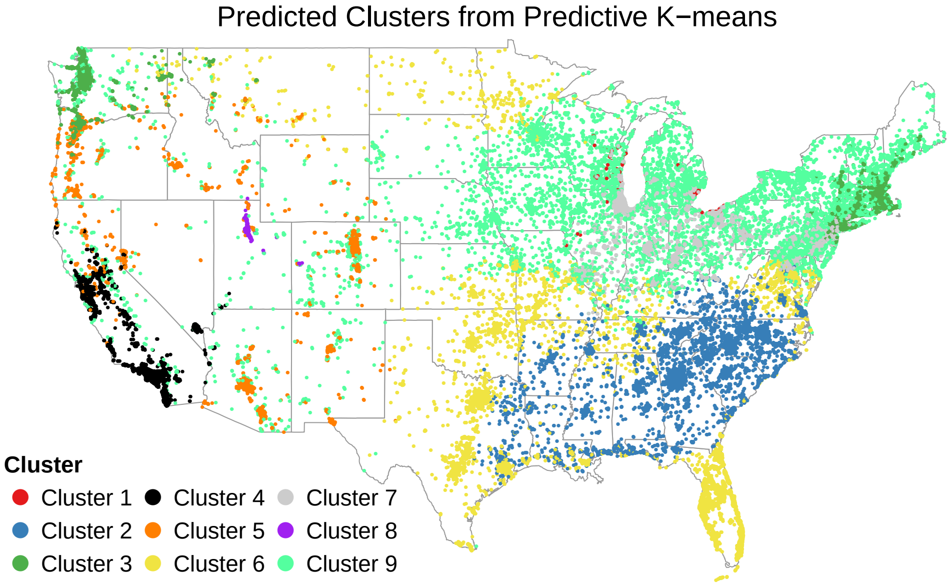


Figure A.6: Predicted cluster membership at Sister Study cohort locations for $K = 10$ (jittered to protect confidentiality).

Table A.8: Estimated difference in SBP (in mmHg) associated with a 10 $\mu g/m^3$ difference in annual ambient $\text{PM}_{2.5}$ exposure, $K = 10$.

Exposure	n	Estimate	95% CI	p-value
$\text{PM}_{2.5}$ by Cluster				0.069
Cluster 1	280	3.16	(−0.73, 16.1)	0.63
Cluster 2	6,379	3.98	(0.71, 7.24)	0.017
Cluster 3	4,477	3.70	(−1.21, 8.60)	0.14
Cluster 4	3,870	3.45	(0.61, 6.28)	0.017
Cluster 5	3,363	−0.51	(−4.03, 3.02)	0.78
Cluster 6	8,318	−1.51	(−4.48, 1.46)	0.32
Cluster 7	9,666	1.12	(−0.91, 3.15)	0.28
Cluster 8	260	17.7	(0.91, 34.5)	0.039
Cluster 9	10,593	1.17	(−0.69, 3.02)	0.22

A.4 Analysis using k-means

The cluster centers identified by k-means and a map of monitor assignments are provided in Figures A.7 and A.8. A table providing the health associations in all k-means clusters is provided in Table A.9.

Table A.9: Estimated difference in SBP (in mmHg) associated with a 10 $\mu g/m^3$ difference in annual ambient $\text{PM}_{2.5}$ exposure. Cohort is partitioned by predicted membership in clusters from regular k-means, with 2 PCA scores and 10 df TPRS as the covariates for the SVM prediction model.

Exposure	n	Est.	95% CI	p-value
$\mathbf{PM}_{2.5}$ by Cluster				0.0139[a]
Cluster 1	13,441	2.98	(1.07, 4.89)	0.002
Cluster 2	15,798	0.51	(−1.30, 2.32)	0.51
Cluster 3	9,121	2.64	(−0.24, 5.52)	0.072
Cluster 4	3,767	3.41	(0.43, 6.38)	0.025
Cluster 5	3,599	4.35	(−0.24, 8.94)	0.063
Cluster 6	267	9.56	(−1.76, 20.9)	0.10
Cluster 7	1,213	−4.69	(−9.73, 0.35)	0.068

[a] p-value for a Wald test for a difference between cluster coefficient estimates.

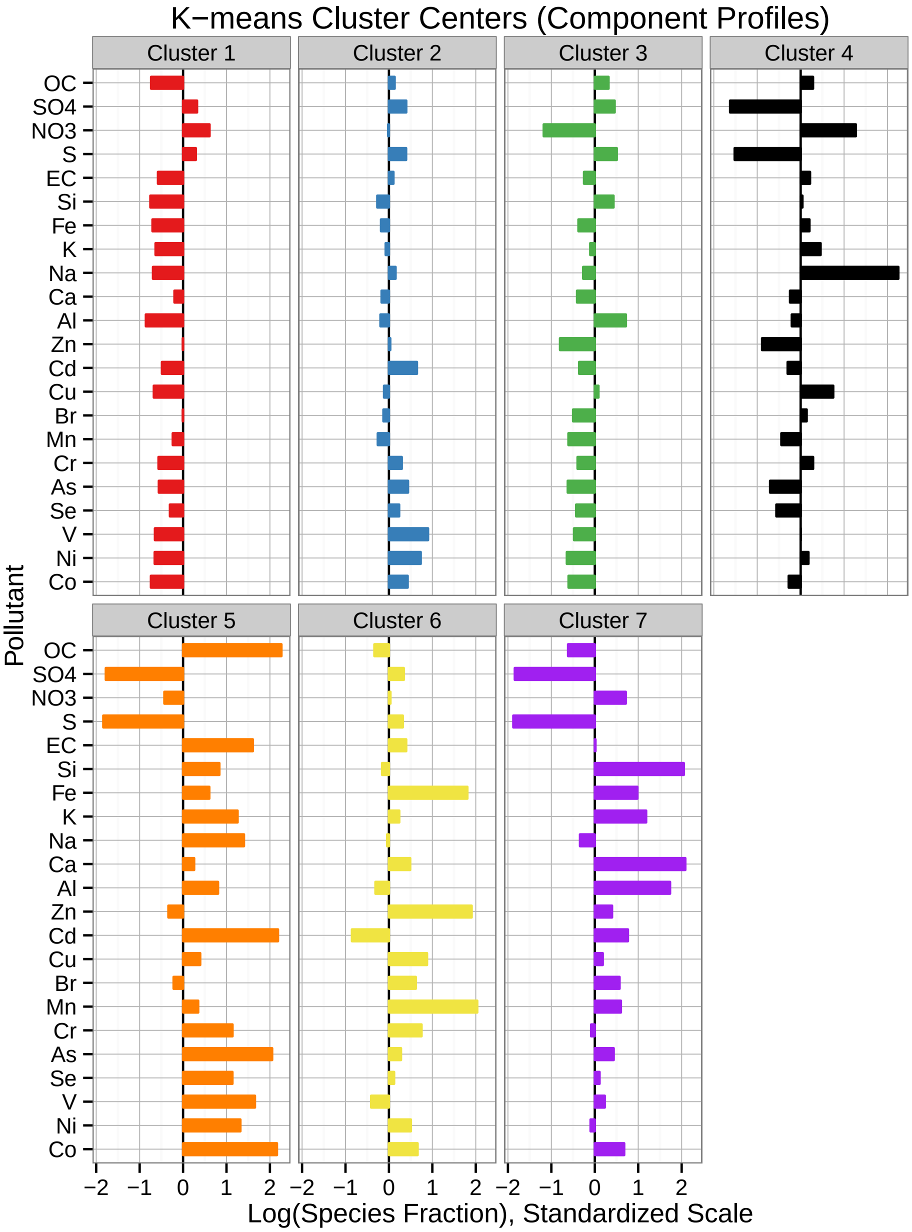
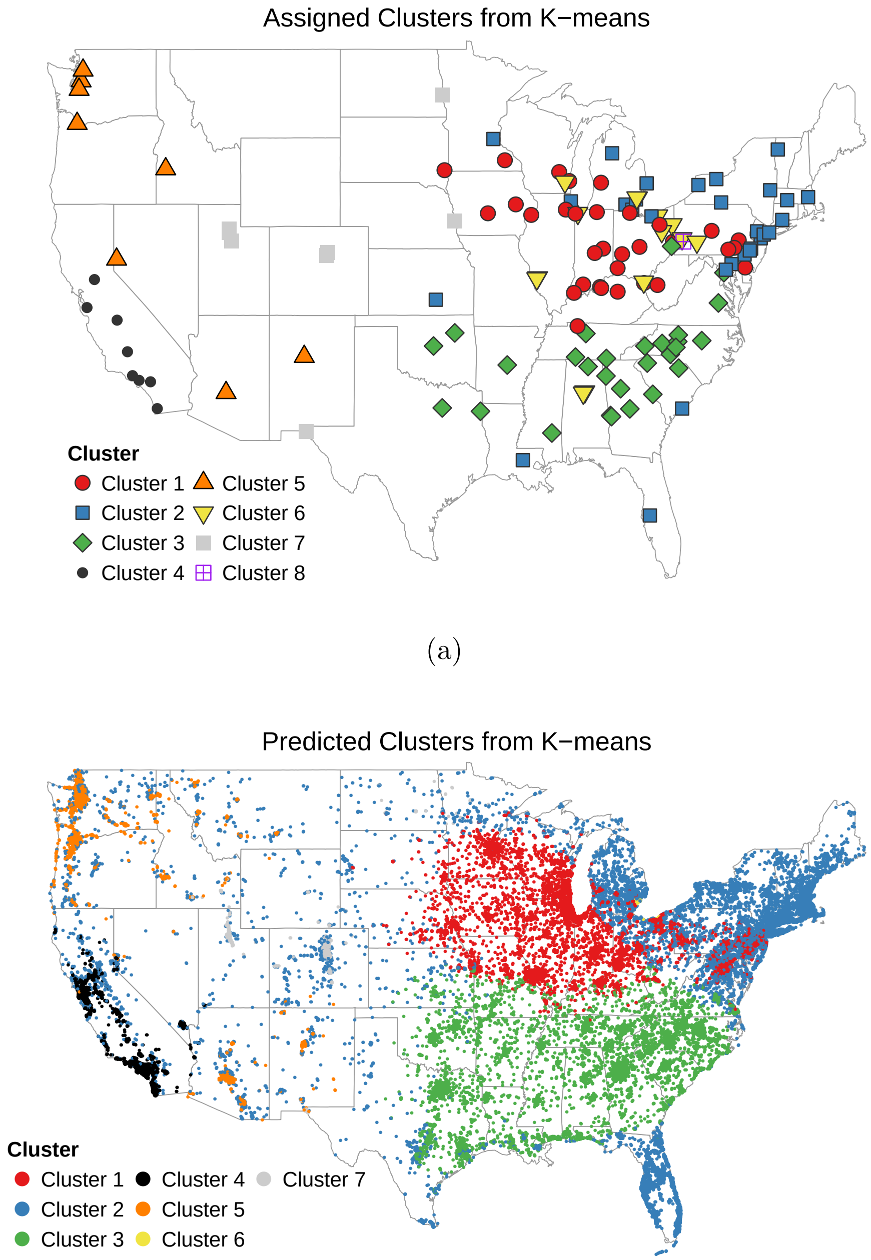


Figure A.7: Cluster centers identified by k-means, when $K = 8$. A singleton cluster is not shown. Values are on the standardized (mean zero and unit variance) scale. Components are ordered by decreasing mass concentration.



(a)

(b)

Figure A.8: (a) Assigned *k*-means cluster membership at AQS monitor locations. (b) Predicted cluster membership at Sister Study cohort locations (jittered to protect confidentiality).

Appendix B

APPENDIX FOR CHAPTER 3

B.1 Method Details

B.1.1 Outline and Implementation

To find the value of λ that minimizes the posterior expectation of the loss (3.4), we use numerical integration for computing posterior moments and search over a grid of candidate λ values. The steps of this procedure are:

1. Draw a sample of size M from the posterior distributions of $\boldsymbol{\beta}$ and σ^2 as given in (3.3).

2. For each vector $(\tilde{\beta}_j, \tilde{\boldsymbol{\gamma}}_j, \tilde{\sigma}^2_j)$ in the posterior sample $(j = 1, \ldots, M)$, compute fMBV $\left(\hat{\beta}_\lambda; \tilde{\beta}_j, \tilde{\boldsymbol{\gamma}}_j, \tilde{\sigma}^2_j\right)$ for a sequence of λ values. This step is described in more detail for ridge regression in Section B.1.2 and for the LASSO in Section B.1.3.

3. Approximate the posterior risk $\mathrm{E}_{\beta,\boldsymbol{\gamma},\sigma^2|\boldsymbol{y}}\left[\mathrm{fMBV}\left(\hat{\beta}_\lambda; \beta, \boldsymbol{\gamma}, \sigma^2\right)\right]$ by the summation $\frac{1}{M}\sum_{j=1}^{M} \mathrm{fMBV}\left(\hat{\beta}_\lambda; \tilde{\beta}_j, \tilde{\boldsymbol{\gamma}}_j, \tilde{\sigma}^2_j\right)$.

4. Compute the optimal penalty parameter $\hat{\lambda} = \arg\min_\lambda \mathrm{E}_{\beta,\boldsymbol{\gamma},\sigma^2|\boldsymbol{y}}\left[\mathrm{fMBV}\left(\hat{\beta}_\lambda, \beta, \boldsymbol{\gamma}, \sigma^2\right)\right]$.

5. Compute the estimate $\hat{\beta}_{\hat{\lambda}}(\boldsymbol{y})$ from the original data.

When implementing this procedure, we use at least 100 candidate λ values. The maximum of this range is chosen to correspond to include large enough penalty so that the all coefficients are shrunk essentially to zero (this can be achieved exactly for LASSO and within some suitably small tolerance for ridge). The minimum of the range is chosen to correspond to a strictly positive number small enough to effectively result in no penalization.

B.1.2 fMBV for Ridge Regression

The closed form for the ridge estimator, $\hat{\boldsymbol{\beta}}_{\lambda}^{Ridge} = (\boldsymbol{X}^{\top}\boldsymbol{X} + \Lambda)^{-1}\boldsymbol{X}^{\top}\boldsymbol{y}$, where Λ is a diagonal matrix with entries $(0, \lambda, \ldots, \lambda)$, permits exact computation of fMBV in Step 2 of this procedure. By not penalizing the coefficient of $\boldsymbol{x}$, we are fitting a special case of the adaptive ridge estimator, which allows for different penalties λ_i on each element of $\boldsymbol{\beta} = (\beta, \boldsymbol{\gamma})$ (Brown and Zidek 1980). The bias of $\hat{\boldsymbol{\beta}}_{\lambda}^{Ridge}$ is the vector $\text{Bias}(\hat{\boldsymbol{\beta}}_{\lambda}^{Ridge}, \boldsymbol{\beta}) = -(\boldsymbol{X}^{\top}\boldsymbol{X} + \Lambda)^{-1}\Lambda\boldsymbol{\beta}$ and the covariance matrix is

$$\text{Var}(\hat{\boldsymbol{\beta}}_{\lambda}^{Ridge}) = \sigma^2(\boldsymbol{X}^{\top}\boldsymbol{X} + \Lambda)^{-1}\boldsymbol{X}^{\top}\boldsymbol{X}(\boldsymbol{X}^{\top}\boldsymbol{X} + \Lambda)^{-1}. \tag{B.1}$$

We compute fMBV $\left(\hat{\beta}_{\lambda}; \tilde{\beta}_j, \tilde{\boldsymbol{\gamma}}_j, \tilde{\sigma}_j^2\right)$ in Step 2 above by plugging $(\tilde{\beta}_j, \tilde{\boldsymbol{\gamma}}_j, \tilde{\sigma}_j^2)$ into

$$\begin{aligned} \text{fMBV}(\hat{\beta}_{\lambda}^{Ridge}; \boldsymbol{\beta}, \sigma^2) = \max \Big\{ & (\boldsymbol{e}_1^{\top}(\boldsymbol{X}^{\top}\boldsymbol{X} + \Lambda)^{-1}\Lambda\boldsymbol{\beta})^2, \\ & \sigma^2\boldsymbol{e}_1^{\top}(\boldsymbol{X}^{\top}\boldsymbol{X} + \Lambda)^{-1}\boldsymbol{X}^{\top}\boldsymbol{X}(\boldsymbol{X}^{\top}\boldsymbol{X} + \Lambda)^{-1}\boldsymbol{e}_1 \Big\}, \end{aligned}$$

where $\boldsymbol{e}_1$ is a vector with first element 1 and zero in all other elements.

B.1.3 fMBV for LASSO

The lack of closed-form expression for the LASSO requires that we use numerical approximation to compute fMBV $\left(\hat{\beta}_{\lambda}; \tilde{\beta}_j, \tilde{\boldsymbol{\gamma}}_j, \tilde{\sigma}_j^2\right)$ in Step 2 of our procedure. For each $(\tilde{\beta}_j, \tilde{\boldsymbol{\gamma}}_j, \tilde{\sigma}_j^2)$, we simulate B replicates $\boldsymbol{y}_{\tilde{\beta}}^*$ according to the model $\boldsymbol{y}^* = \tilde{\beta}_j\boldsymbol{x} + \boldsymbol{Z}\tilde{\boldsymbol{\gamma}}_j + \boldsymbol{\epsilon}^j$, where $\epsilon_i^j \sim N(0, \tilde{\sigma}_j^2)$. We compute the LASSO estimator for a sequence of λ values to give a set of estimators $\hat{\beta}_{\lambda}^{LASSO}(\boldsymbol{y}_{\tilde{\beta}}^*)$. fMBV is approximated using the empirical bias and variance:

$$\text{fMBV}\left(\hat{\beta}_{\lambda}^{LASSO}, \tilde{\boldsymbol{\beta}}_j, \tilde{\sigma}_j^2\right) \approx \max\left\{\left(\tilde{\beta}_j - \frac{1}{B}\sum\hat{\beta}_{\lambda}^{LASSO}\right)^2, \frac{1}{B}\sum\left(\hat{\beta}_{\lambda}^{LASSO} - \frac{1}{B}\sum\hat{\beta}_{\lambda}^{LASSO}\right)^2\right\}$$

This is then averaged over the posterior sample to get the posterior risk and we proceed with Steps 4 and 5.

B.1.4 Confidence Intervals

To construct confidence intervals, we use an 'invert the test' approach. For a dataset $(y, \boldsymbol{X})$, we do the following:

1. Select a value β^0 to test.

2. For each of $j = 1, \ldots, B^*$ replications, do the following:

   (a) Generate $\boldsymbol{\theta}_j^* = (\boldsymbol{\gamma}_j^*, \sigma_j^{2*})$ from a slice of the posterior distribution $\pi(\boldsymbol{\theta}|\beta = \beta^0; \boldsymbol{y}, \boldsymbol{X}) = \pi(\boldsymbol{\gamma}|\sigma^2, \beta = \beta^0; \boldsymbol{y}, \boldsymbol{X})\pi(\sigma^2|\beta = \beta^0; \boldsymbol{y}, \boldsymbol{X})$. For our conjugate setting, the conditional distribution for $\boldsymbol{\gamma}$ is:

   $$\pi(\boldsymbol{\gamma}|\sigma^2, \beta = \beta^0; \boldsymbol{y}, \boldsymbol{X}) = N\left(\boldsymbol{m}_2 + \boldsymbol{v}_{21}/v_{11}(\beta^0 - m_1), \sigma^2(\boldsymbol{V}_{22} - \boldsymbol{v}_{21}\boldsymbol{v}_{12}/v_{11})\right)$$

   where $\boldsymbol{m}$ and $\boldsymbol{v}$ are the posterior mean and variance defined in (3.3). The conditional posterior $\pi(\sigma^2|\beta = \beta^0; \boldsymbol{y}, \boldsymbol{X})$ is the same as the unconditional posterior $\pi(\sigma^2|\boldsymbol{y}, \boldsymbol{X})$ as defined in (3.3).

   (b) Sample n draws $\boldsymbol{\epsilon}_j^*$ from $N(0, \sigma_j^{2*})$.

   (c) Compute $\boldsymbol{y}_j^* = \boldsymbol{x}\beta^0 + \boldsymbol{Z}\boldsymbol{\gamma}_j^* + \boldsymbol{\epsilon}_j^*$

   (d) Compute $\hat{\beta}_{\tilde{\lambda}}^j = \hat{\beta}_{\tilde{\lambda}}(\boldsymbol{y}_j^*)$ via the procedure outlined in Section B.1.1. (The posterior distributions from our original dataset are not used in this step. Rather the computations are based upon the uninformative priors originally used)

3. If the original $\hat{\beta}$ is outside the range of the $(\alpha/2, 1 - \alpha/2)$ percentiles of the collection of estimators $\{\hat{\beta}_{\tilde{\lambda}}^j\}$, then reject β^0 at the α level.

4. Repeating this procedure for multiple choices of β^0 to compute the bounds of a confidence interval.

B.2 Simulation Setup

In Simulations 1 through 3, the fixed design matrix was a single draw from a multivariate normal distribution: $(x, z_1, z_2, z_3, z_4, z_5, z_6)^T \sim N(0, \boldsymbol{W})$. The correlation matrix $\boldsymbol{W}$ was structured to have three pairs of correlated confounders:

$$\boldsymbol{W} = \begin{bmatrix} 1 & 0.5 & 0.5 & 0.5 & 0.5 & 0.5 & 0.5 \\ 0.5 & 1 & 0.7 & - & - & - & - \\ 0.5 & 0.7 & 1 & - & - & - & - \\ 0.5 & - & - & 1 & 0.7 & - & - \\ 0.5 & - & - & 0.7 & 1 & - & - \\ 0.5 & - & - & - & - & 1 & 0.7 \\ 0.5 & - & - & - & - & 0.7 & 1 \end{bmatrix} \tag{B.2}$$

Here '$-$' denotes 0. The confounder effects were:

- Simulation 1: (0.05, 0.05, 0.1, 0.1, 0.15, 0.15)
- Simulation 2: (0.5, 0.5, 0.3, 0.3, 0.1, 0.1)
- Simulation 3: (0.05, 0.05, 0.05, 0.05, 0.05, 0.05)

In Simulations 4 and 5, the fixed design was again a draw from a multivariate normal distribution, with correlation:

$$\boldsymbol{W} = \begin{bmatrix}
1.00 & 0.30 & 0.30 & 0.40 & 0.40 & 0.40 & 0.40 & 0.40 & 0.40 & 0.50 & 0.50 & 0.50 & 0.50 \\
0.30 & 1.00 & 0.60 & 0.05 & 0.05 & 0.05 & 0.05 & 0.05 & 0.05 & 0.05 & 0.05 & 0.05 & 0.05 \\
0.30 & 0.60 & 1.00 & 0.05 & 0.05 & 0.05 & 0.05 & 0.05 & 0.05 & 0.05 & 0.05 & 0.05 & 0.05 \\
0.40 & 0.05 & 0.05 & 1.00 & 0.60 & 0.05 & 0.05 & 0.05 & 0.05 & 0.05 & 0.05 & 0.05 & 0.05 \\
0.40 & 0.05 & 0.05 & 0.60 & 1.00 & 0.05 & 0.05 & 0.05 & 0.05 & 0.05 & 0.05 & 0.05 & 0.05 \\
0.40 & 0.05 & 0.05 & 0.05 & 0.05 & 1.00 & 0.60 & 0.05 & 0.05 & 0.05 & 0.05 & 0.05 & 0.05 \\
0.40 & 0.05 & 0.05 & 0.05 & 0.05 & 0.60 & 1.00 & 0.05 & 0.05 & 0.05 & 0.05 & 0.05 & 0.05 \\
0.40 & 0.05 & 0.05 & 0.05 & 0.05 & 0.05 & 0.05 & 1.00 & 0.60 & 0.05 & 0.05 & 0.05 & 0.05 \\
0.40 & 0.05 & 0.05 & 0.05 & 0.05 & 0.05 & 0.05 & 0.60 & 1.00 & 0.05 & 0.05 & 0.05 & 0.05 \\
0.50 & 0.05 & 0.05 & 0.05 & 0.05 & 0.05 & 0.05 & 0.05 & 0.05 & 1.00 & 0.50 & 0.50 & 0.50 \\
0.50 & 0.05 & 0.05 & 0.05 & 0.05 & 0.05 & 0.05 & 0.05 & 0.05 & 0.50 & 1.00 & 0.50 & 0.50 \\
0.50 & 0.05 & 0.05 & 0.05 & 0.05 & 0.05 & 0.05 & 0.05 & 0.05 & 0.50 & 0.50 & 1.00 & 0.50 \\
0.50 & 0.05 & 0.05 & 0.05 & 0.05 & 0.05 & 0.05 & 0.05 & 0.05 & 0.50 & 0.50 & 0.50 & 1.00
\end{bmatrix}$$

The confounder effects were:

- Simulation 4: (0.15, 0.1, -0.1, 0, 0, 0.5, 0.5, 0.1, 0.1, 0.2, -0.05)

- Simulation 5: (0.1, 0.1, 0.1, 0.1, 0.05, 0.05, 0.05, 0.05, 0.05, 0.05, 0.05, 0.05)

B.3 Additional Simulation Results

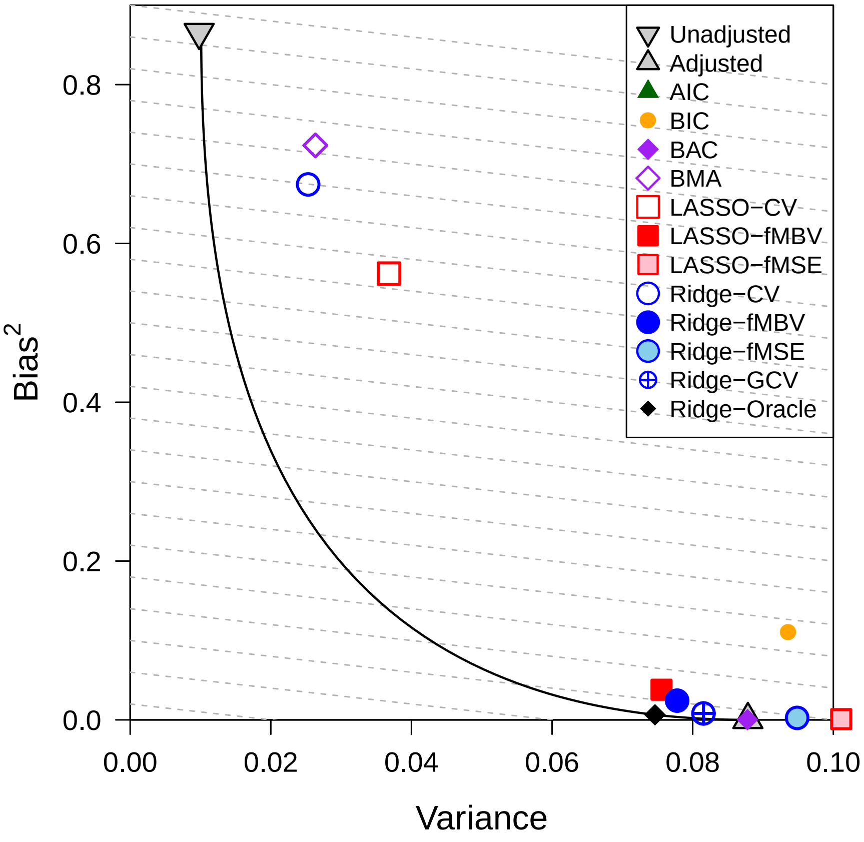


Figure B.1: Variance and squared bias for stimators in Simulation 2. The dashed lines represent contours of equal MSE. The solid black curve represents the theoretical MSE for ridge estimators, for varying values of λ.

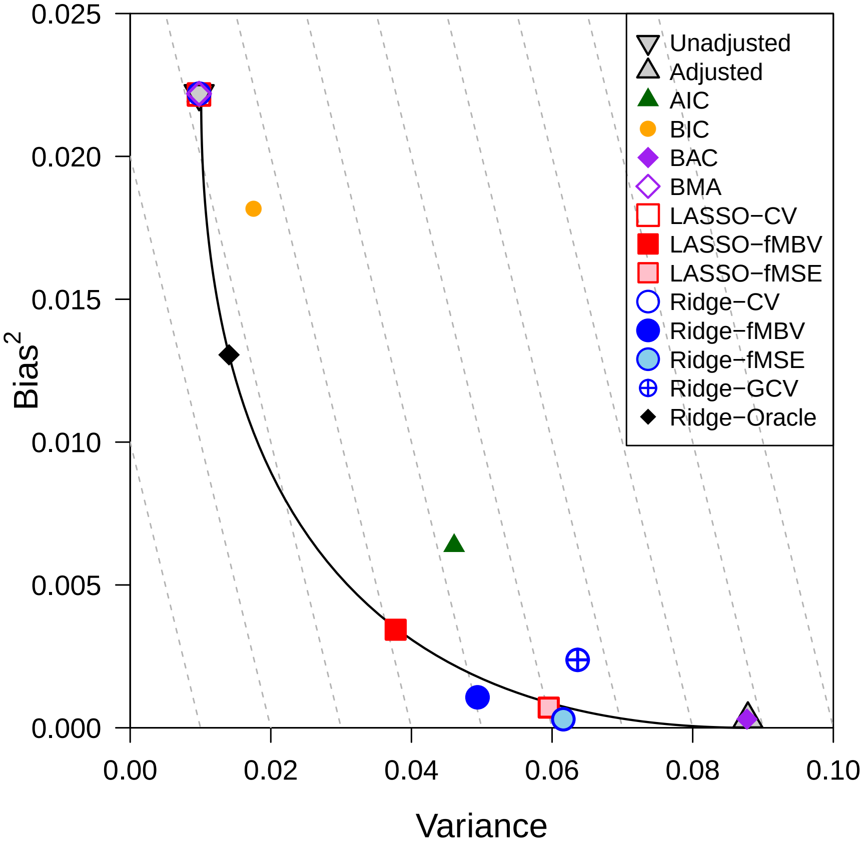
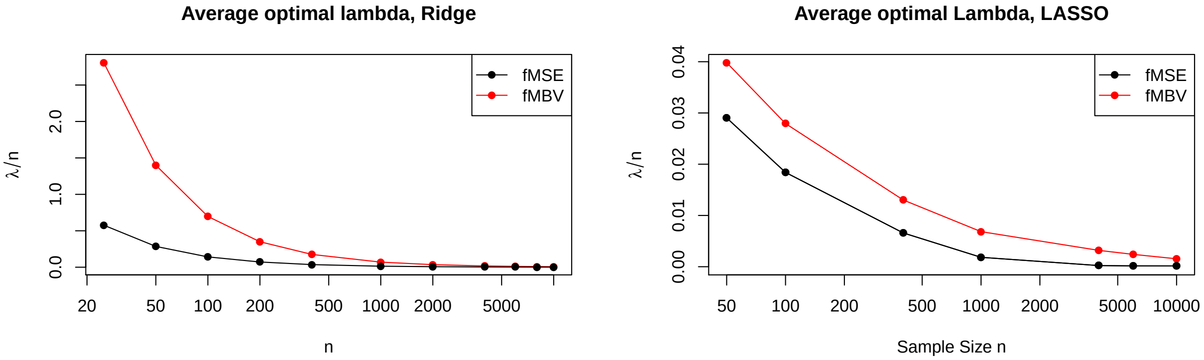


Figure B.2: Variance and squared bias for stimators in Simulation 3. The dashed lines represent contours of equal MSE. The solid black curve represents the theoretical MSE for ridge estimators, for varying values of λ.



(a) Ridge-fMBV

(b) LASSO-fMBV

Figure B.3: Average value of λ/n for the Ridge and LASSO estimators in Simulation 1, by sample size.

B.4 Large Sample Behavior

In settings with a finite number of confounders, Knight and Fu (2000) established the consistency of the Ridge and LASSO estimators if $\lambda = o(n)$ when the number of confounders p is fixed. Their results assume all elements of the parameter vector are penalized, but extend readily to the present setting where the coefficient of x is not penalized. To satisfy the $\lambda = o(n)$ assumption, one could choose a bounded range of λ values over which to maximize, with the upper bound of that range $o(n)$. Figure B.3 shows that in extensions of Simulation 1, λ/n approaches 0 for large sample sizes, providing empirical support for the theoretical result in the current context.

Notably, $\hat{\lambda}$ selected according to fMSE appears to be going to zero faster than $\hat{\lambda}$ selected by fMBV. We see this further in Table B.1, where the MSE for Ridge-fMSE approaches that of the fully-adjusted estimator for large n, while the MSE for Ridge-fMBV is still somewhat higher than the MSE for the fully-adjusted estimator for large n.

Table B.1: MSE for different estimators in Simulations 1 through 3, for varying sample sizes. All values $\times 10^{-1}$.

	n					
Method	50	100	400	1000	6000	10000
Simulation 1						
Unadjusted	0.936	0.921	0.892	0.878	0.861	0.890
Fully Adjusted	1.570	0.882	0.210	0.088	0.014	0.008
Ridge-fMSE	1.178	0.722	0.229	0.101	0.014	0.008
Ridge-fMBV	0.973	0.605	0.219	0.111	0.021	0.014
Simulation 2						
Unadjusted	5.543	8.756	9.074	8.607	8.312	8.300
Fully Adjusted	1.570	0.882	0.210	0.088	0.014	0.008
Ridge-fMSE	1.645	0.974	0.222	0.090	0.014	0.008
Ridge-fMBV	1.526	1.021	0.320	0.143	0.024	0.015
Simulation 3						
Unadjusted	0.358	0.320	0.256	0.236	0.224	0.226
Fully Adjusted	1.570	0.882	0.210	0.088	0.014	0.008
Ridge-fMSE	1.086	0.619	0.176	0.089	0.015	0.009
Ridge-fMBV	0.881	0.505	0.152	0.080	0.018	0.012

Table B.2: Standard error estimates for different estimators in Simulations 1 through 3, for varying sample sizes.

		n					
Estimator		50	100	400	1000	6000	10000
Simulation 1							
Ridge-fMBV	Estimated SE	0.267	0.214	0.124	0.086	0.037	0.029
	True SE	0.305	0.236	0.128	0.088	0.036	0.028
Ridge-fMSE	Estimated SE	0.315	0.253	0.145	0.096	0.038	0.029
	True SE	0.341	0.266	0.146	0.098	0.037	0.029
Simulation 2							
Ridge-fMBV	Estimated SE	0.308	0.264	0.144	0.091	0.037	0.029
	True SE	0.353	0.279	0.142	0.093	0.037	0.029
Ridge-fMSE	Estimated SE	0.356	0.296	0.150	0.093	0.037	0.029
	True SE	0.398	0.308	0.148	0.095	0.037	0.029
Simulation 3							
Ridge-fMBV	Estimated SE	0.259	0.204	0.110	0.074	0.036	0.028
	True SE	0.295	0.222	0.115	0.080	0.035	0.028
Ridge-fMSE	Estimated SE	0.307	0.241	0.130	0.087	0.039	0.030
	True SE	0.329	0.248	0.129	0.091	0.038	0.030

Brown, P. J., and J. V. Zidek. 1980. “Adaptive multvariate ridge regression”. *Annals of Statistics* 8 (1): 64–74.

Knight, K., and W. Fu. 2000. “Asymptotics for LASSO-type Estimators”. *The Annals of Statistics* 28 (5): 1356–1378.

Table B.3: Correlation between measured variables in the MESA Caucasian sub-cohort.

						Phys.								Diab.	Lipid	Education			Income ($K)			
	cIMT	Smoker	Age	Sex	BMI	Act.	Chol.	HDL	Trig.	IL-6	CRP	Fib.	Diab.	Med.	Med	H.S.	Col	Grad	25-50	50-100	100+	Alcohol
cIMT	1.00	0.10	0.25	0.14	0.14	-0.02	0.08	-0.14	0.13	0.14	0.07	0.11	0.10	0.08	0.13	-0.06	-0.12	0.08	0.06	0.05	0.09	0.05
Smoker	0.10	1.00	0.04	0.03	-0.00	-0.05	-0.04	0.01	0.00	0.07	-0.02	-0.02	0.02	0.01	0.01	-0.07	-0.13	0.13	0.04	0.05	0.06	0.21
Age	0.25	0.04	1.00	0.00	0.02	0.05	0.03	0.08	0.05	0.12	0.09	0.14	0.06	0.05	0.14	-0.08	-0.05	0.00	-0.03	-0.04	-0.01	-0.00
gender1	0.14	0.03	0.00	1.00	0.03	0.02	-0.11	-0.45	0.07	-0.04	-0.21	-0.11	0.06	0.04	0.08	0.04	0.12	-0.14	-0.12	-0.05	-0.10	0.30
bmi1c	0.14	-0.00	0.02	0.03	1.00	-0.13	0.05	-0.31	0.24	0.28	0.25	0.31	0.15	0.12	0.08	-0.04	-0.06	0.10	0.10	0.07	0.07	-0.11
Phys. Act.	-0.02	-0.05	0.05	0.02	-0.13	1.00	-0.02	0.05	-0.03	-0.09	-0.02	-0.07	-0.01	-0.02	0.01	0.00	0.02	-0.04	-0.07	-0.10	-0.08	-0.01
chol1	0.08	-0.04	0.03	-0.11	0.05	-0.02	1.00	0.14	0.41	-0.04	0.06	0.17	-0.04	-0.08	-0.18	-0.06	-0.05	0.06	0.07	0.01	0.07	-0.03
hdl1	-0.14	0.01	0.08	-0.45	-0.31	0.05	0.14	1.00	-0.34	-0.10	0.01	-0.10	-0.10	-0.07	-0.10	0.01	0.01	-0.07	-0.06	-0.07	-0.06	0.07
trig1	0.13	0.00	0.05	0.07	0.24	-0.03	0.41	-0.34	1.00	0.07	0.13	0.14	0.17	0.04	0.07	-0.04	-0.10	0.10	0.11	0.04	0.08	-0.06
il61	0.14	0.07	0.12	-0.04	0.28	-0.09	-0.04	-0.10	0.07	1.00	0.46	0.35	0.04	0.03	-0.00	-0.15	-0.06	0.02	0.11	0.06	0.11	-0.05
crp1	0.07	-0.02	0.09	-0.21	0.25	-0.02	0.06	0.01	0.13	0.46	1.00	0.41	0.02	0.01	-0.05	-0.08	-0.07	0.06	0.09	0.05	0.05	-0.11
fib1	0.11	-0.02	0.14	-0.11	0.31	-0.07	0.17	-0.10	0.14	0.35	0.41	1.00	0.07	0.07	0.05	-0.09	-0.06	0.07	0.10	0.04	0.07	-0.14
diabetes	0.10	0.02	0.06	0.06	0.15	-0.01	-0.04	-0.10	0.17	0.04	0.02	0.07	1.00	0.83	0.10	0.00	-0.02	0.04	0.06	0.04	0.02	-0.06
diabhx1	0.08	0.01	0.05	0.04	0.12	-0.02	-0.08	-0.07	0.04	0.03	0.01	0.07	0.83	1.00	0.08	-0.00	-0.00	0.04	0.06	0.07	0.04	-0.04
lipid1c	0.13	0.01	0.14	0.08	0.08	0.01	-0.18	-0.10	0.07	-0.00	-0.05	0.05	0.10	0.08	1.00	-0.04	0.04	0.02	-0.06	-0.05	-0.08	-0.01
educ1	-0.06	-0.07	-0.08	0.04	-0.04	0.00	-0.06	0.01	-0.04	-0.15	-0.08	-0.09	0.00	-0.00	-0.04	1.00	-0.15	-0.23	-0.10	-0.03	-0.10	0.03
educ2	-0.12	-0.13	-0.05	0.12	-0.06	0.02	-0.05	0.01	-0.10	-0.06	-0.07	-0.06	-0.02	-0.00	0.04	-0.15	1.00	-0.31	-0.28	-0.19	-0.31	0.04
educ3	0.08	0.13	0.00	-0.14	0.10	-0.04	0.06	-0.07	0.10	0.02	0.06	0.07	0.04	0.04	0.02	-0.23	-0.31	1.00	0.25	0.15	0.21	-0.11
inc1	0.06	0.04	-0.03	-0.12	0.10	-0.07	0.07	-0.06	0.11	0.11	0.09	0.10	0.06	0.06	-0.06	-0.10	-0.28	0.25	1.00	0.52	0.66	-0.14
inc2	0.05	0.05	-0.04	-0.05	0.07	-0.10	0.01	-0.07	0.04	0.06	0.05	0.04	0.04	0.07	-0.05	-0.03	-0.19	0.15	0.52	1.00	0.65	-0.11
inc3	0.09	0.06	-0.01	-0.10	0.07	-0.08	0.07	-0.06	0.08	0.11	0.05	0.07	0.02	0.04	-0.08	-0.10	-0.31	0.21	0.66	0.65	1.00	-0.12
logalc	0.05	0.21	-0.00	0.30	-0.11	-0.01	-0.03	0.07	-0.06	-0.05	-0.11	-0.14	-0.06	-0.04	-0.01	0.03	0.04	-0.11	-0.14	-0.11	-0.12	1.00

Appendix C

APPENDIX FOR CHAPTER 4

C.1 Simulation Outline

The following outline describes the procedure followed in the simulations in Chapter 4.

1. Create the surfaces Z_1, X, and μ_Y as described in the text for each simulation.

2. Filter X using a high-pass filter thresholded at ‘effective frequency’ $\omega = 1, \ldots, 128$ as described in Section 4.3.2. Denote each of these new exposures $X_{HP(\omega)}$.

3. Apply wavelet thresholding to X at levels $\ell = 0, \ldots, L$ for $L = 0, \ldots, 7$, as described in Section 4.3.2. Denote each of these new exposures $X_{Wave(L)}$.

4. Pre-adjust X using TPRS. First define TPRS across the entire domain (all locations). Project X onto the space defined by the first *df* TPRS, by fitting a linear regression model with mean $\hat{X}_{df} = \sum_{j=1}^{df} t_j(s; \boldsymbol{s}, df)\tilde{\alpha}_j$. Define $X_{TPRSpw(df)} = X - \hat{X}_{df}$.

5. For $b = 1, \ldots, B$ replications:

   (a) Select n locations $s_1, \ldots, s_n$ from the grid to use as ‘subjects’, and extract the relevant values $Z_1(s_i)$, $X(s_i)$, $X_{HP(\omega)}(s_i)$, $X_{Wave(L)}(s_i)$, $X_{TPRSpw(df)}(s_i)$, and $\mu_Y(s_i)$.

   (b) Create the surface $Y_b = \mu_Y + \boldsymbol{\epsilon}$, for $\epsilon \overset{iid}{\sim} N(0, 1)$.

   (c) Create a matrix of thin-plate regression splines (TPRS) using the n subject locations s_i for various fixed degrees of freedom, indexed by *df*. Call these matrices of spline values $\mathbf{S}^{df}$ (we suppress their dependence upon the collection of locations $\{s_i\}$ for notational convenience).

(d) Fit regression models with the following mean structure:

$$\begin{aligned}
\mathrm{E}[Y_i|X, Z_1] &= \beta_0 + \beta_1 X + \beta_2 Z_1 \\
\mathrm{E}[Y_i|X] &= \beta_0 + \beta_1^{Unadj} X \\
\mathrm{E}[Y_i|X_{HP(\omega)}] &= \beta_0 + \beta_1^{HP(\omega)} X_{HP(\omega)} + \tilde{\beta}_1^{HP(\omega)}(X - X_{HP(\omega)}) \\
\mathrm{E}[Y_i|X_{Wave(L)}] &= \beta_0 + \beta_1^{Wave(L)} X_{Wave(L)} + \tilde{\beta}_1^{Wave(L)}(X - X_{Wave(L)}) \\
\mathrm{E}[Y_i|X, \mathbf{S}^{df}] &= \beta_0 + \beta_1^{TPRS(df)} X + \mathbf{S}^{df} \\
\mathrm{E}[Y_i|X_{TPRSpw(df)}] &= \beta_0 + \beta_1^{TPRSpw(df)} X_{TPRSpw(df)} + \tilde{\beta}_1^{TPRSpw(df)}(X - X_{TPRSpw(df)})
\end{aligned}$$

The $\tilde{\beta}$'s are not used for estimating associations, but rather included to fully model the variation in Y. In each model, sandwich estimates of the covariance matrix are used to compute standard errors.

6. Compare the values of $\hat{\beta}_1$'s to the true value ($\beta_1 = 1$).

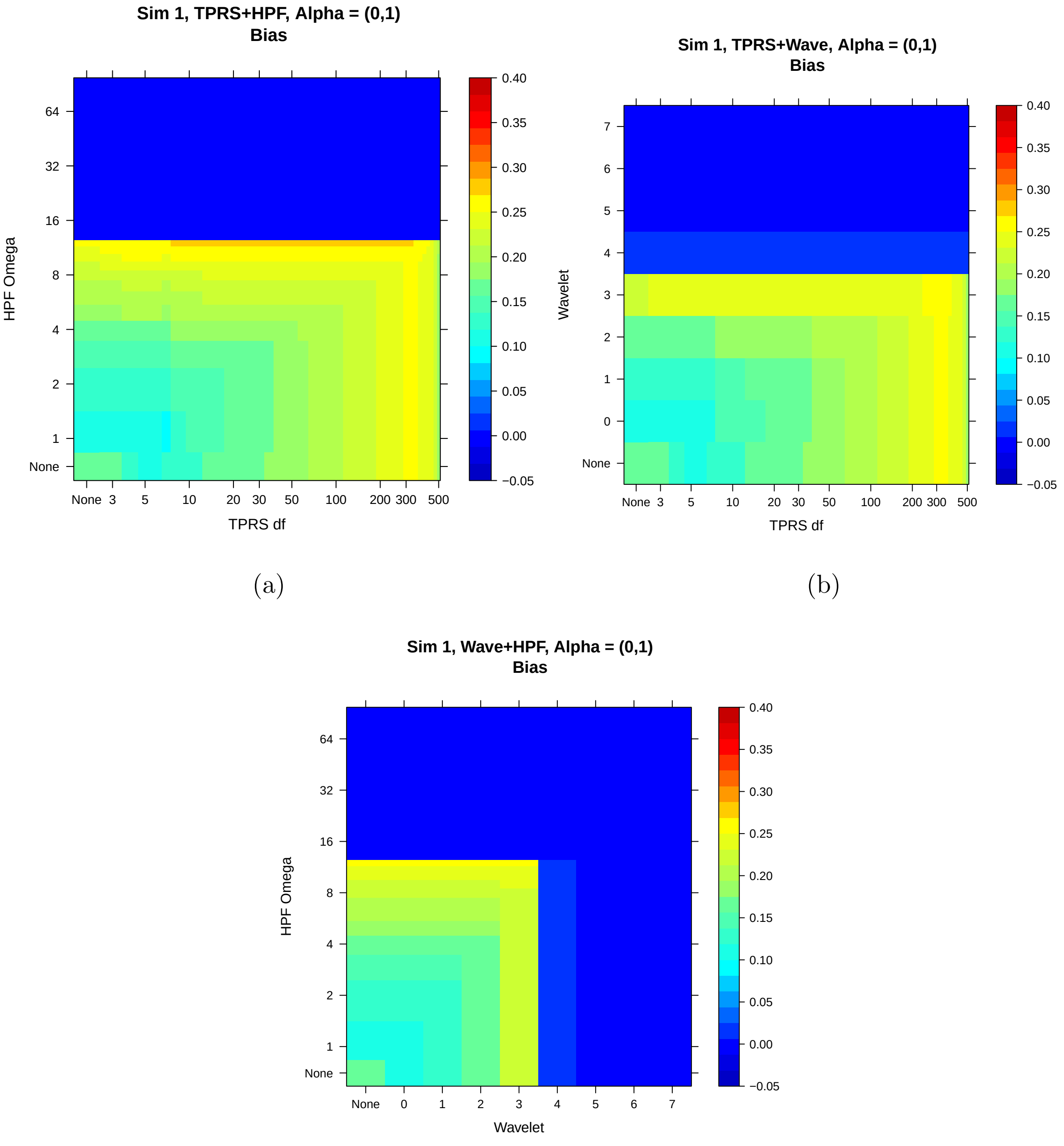
C.2 Additional Results for Simulation 1

Table C.1: Performance measures for different estimators in Simulation 1 for $(\alpha_1, \alpha_2) = (1, 0)$ setting. See Table 4.1 for an explanation of abbreviations.

	$\hat{\beta}$	Bias($\hat{\beta}$)	$SE(\hat{\beta})$	MSE($\hat{\beta}$)	$\widehat{SE}(\hat{\beta})$	Coverage
Unadjusted	1.20	0.199	0.013	0.0399	0.013	0.000
Adjusted	1.00	0.000	0.012	0.0002	0.012	0.952
HPF: $\omega = 1$	1.22	0.225	0.013	0.0506	0.013	0.000
HPF: $\omega = 2$	1.17	0.169	0.015	0.0287	0.015	0.000
HPF: $\omega = 4$	1.08	0.083	0.017	0.0071	0.017	0.000
HPF: $\omega = 12$	1.04	0.036	0.023	0.0018	0.023	0.635
HPF: $\omega = 13$	1.03	0.035	0.023	0.0017	0.023	0.681
HPF: $\omega = 16$	1.03	0.031	0.024	0.0016	0.025	0.753
HPF: $\omega = 32$	1.02	0.024	0.029	0.0014	0.030	0.882
HPF: $\omega = 64$	1.02	0.019	0.036	0.0017	0.038	0.938
HPF: $\omega = 128$	1.02	0.015	0.053	0.0030	0.053	0.947
TPRS: $df = 3$	1.19	0.187	0.014	0.0351	0.014	0.000
TPRS: $df = 5$	1.20	0.203	0.014	0.0415	0.014	0.000
TPRS: $df = 50$	1.03	0.029	0.019	0.0012	0.019	0.671
TPRS: $df = 300$	1.00	0.000	0.022	0.0005	0.023	0.953
Wave: $L = 0$	1.20	0.204	0.013	0.0420	0.013	0.000
Wave: $L = 1$	1.19	0.189	0.015	0.0359	0.014	0.000
Wave: $L = 2$	1.08	0.082	0.018	0.0071	0.017	0.001
Wave: $L = 4$	1.03	0.027	0.025	0.0014	0.025	0.812
Wave: $L = 6$	1.02	0.022	0.037	0.0019	0.039	0.932
Wave: $L = 7$	1.01	0.015	0.057	0.0035	0.057	0.942

Table C.2: Performance measures for different estimators in Simulation 1 for $(\alpha_1, \alpha_2) = (0, 1)$ setting. See Table 4.1 for an explanation of abbreviations.

	$\hat{\beta}$	Bias($\hat{\beta}$)	$SE(\hat{\beta})$	MSE($\hat{\beta}$)	$\widehat{SE}(\hat{\beta})$	Coverage
Unadjusted	1.16	0.162	0.013	0.0264	0.013	0.000
Adjusted	1.00	0.000	0.012	0.0002	0.012	0.951
HPF: $\omega = 1$	1.10	0.102	0.014	0.0106	0.014	0.000
HPF: $\omega = 2$	1.13	0.125	0.015	0.0159	0.015	0.000
HPF: $\omega = 4$	1.16	0.164	0.017	0.0271	0.018	0.000
HPF: $\omega = 12$	1.26	0.258	0.022	0.0669	0.022	0.000
HPF: $\omega = 13$	1.00	0.000	0.024	0.0006	0.024	0.950
HPF: $\omega = 16$	1.00	0.000	0.025	0.0006	0.025	0.945
HPF: $\omega = 32$	1.00	0.000	0.030	0.0009	0.030	0.942
HPF: $\omega = 64$	1.00	0.001	0.037	0.0014	0.038	0.959
HPF: $\omega = 128$	1.00	-0.001	0.052	0.0027	0.054	0.956
TPRS: $df = 3$	1.17	0.175	0.014	0.0307	0.014	0.000
TPRS: $df = 5$	1.11	0.112	0.014	0.0127	0.014	0.000
TPRS: $df = 50$	1.18	0.183	0.018	0.0339	0.018	0.000
TPRS: $df = 300$	1.25	0.251	0.021	0.0634	0.022	0.000
Wave: $L = 0$	1.11	0.111	0.013	0.0124	0.013	0.000
Wave: $L = 1$	1.12	0.122	0.015	0.0151	0.015	0.000
Wave: $L = 2$	1.16	0.162	0.017	0.0264	0.017	0.000
Wave: $L = 4$	1.02	0.016	0.026	0.0009	0.026	0.901
Wave: $L = 6$	1.00	0.000	0.037	0.0014	0.039	0.959
Wave: $L = 7$	1.00	-0.000	0.056	0.0032	0.058	0.948

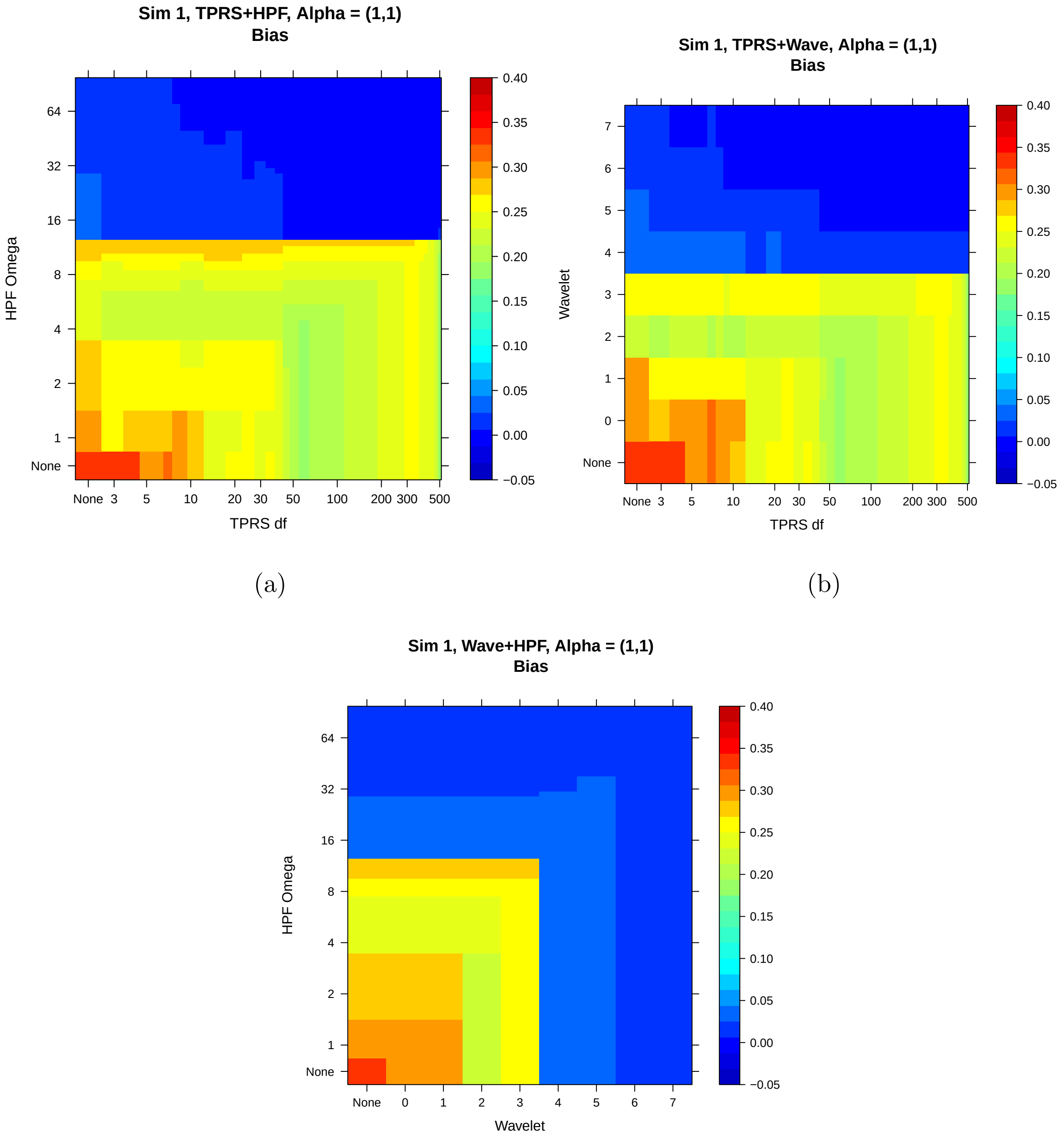


(a)

(b)

(c)

Figure C.1: Estimates of β_1 from combining pre-adjustment approaches in Simulation 1, when $(\alpha_1, \alpha_2) = (0, 1)$.

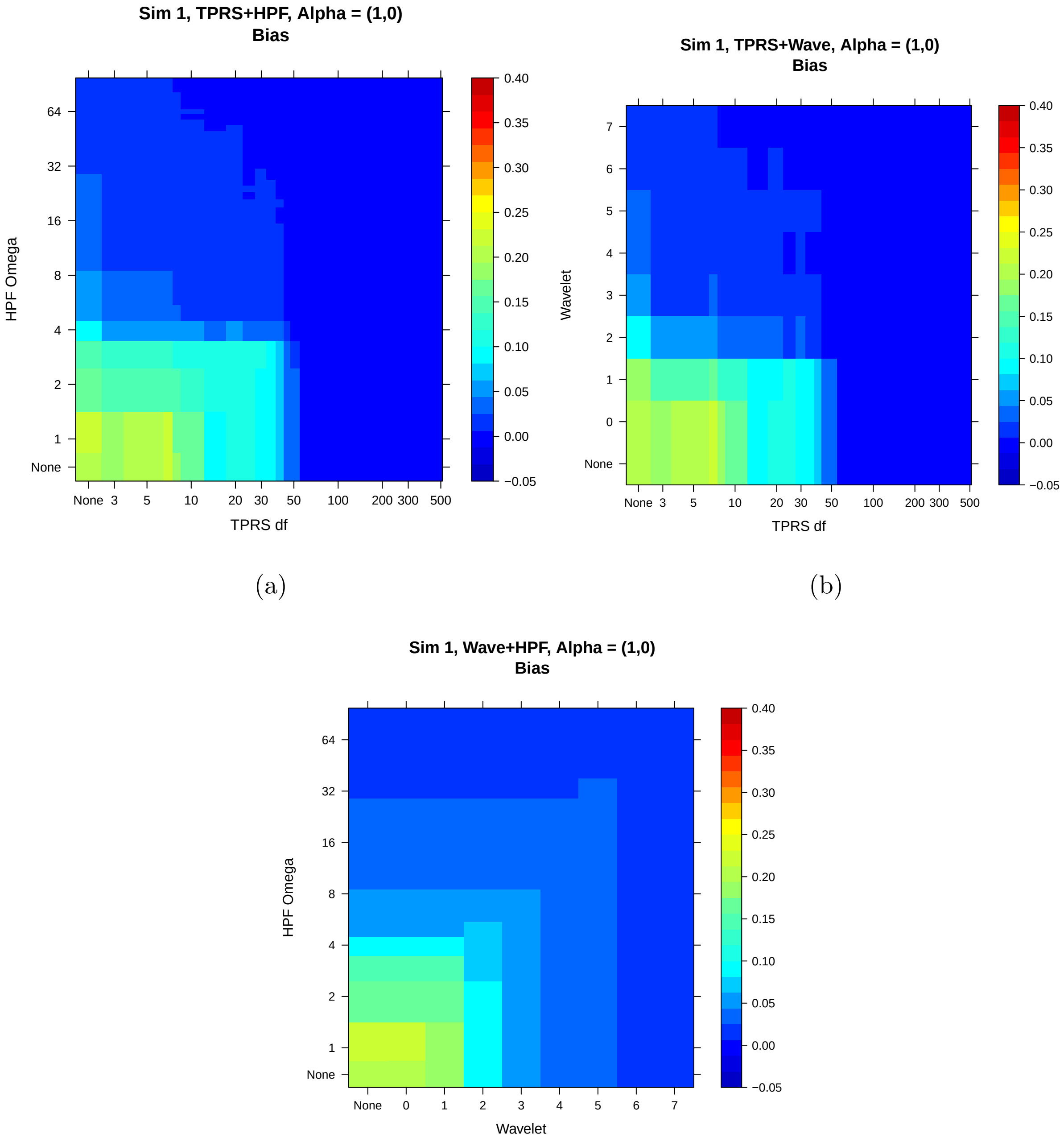


(a)

(b)

(c)

Figure C.2: Estimates of β_1 from combining pre-adjustment approaches in Simulation 1, when $(\alpha_1, \alpha_2) = (1, 1)$.



(a)

(b)

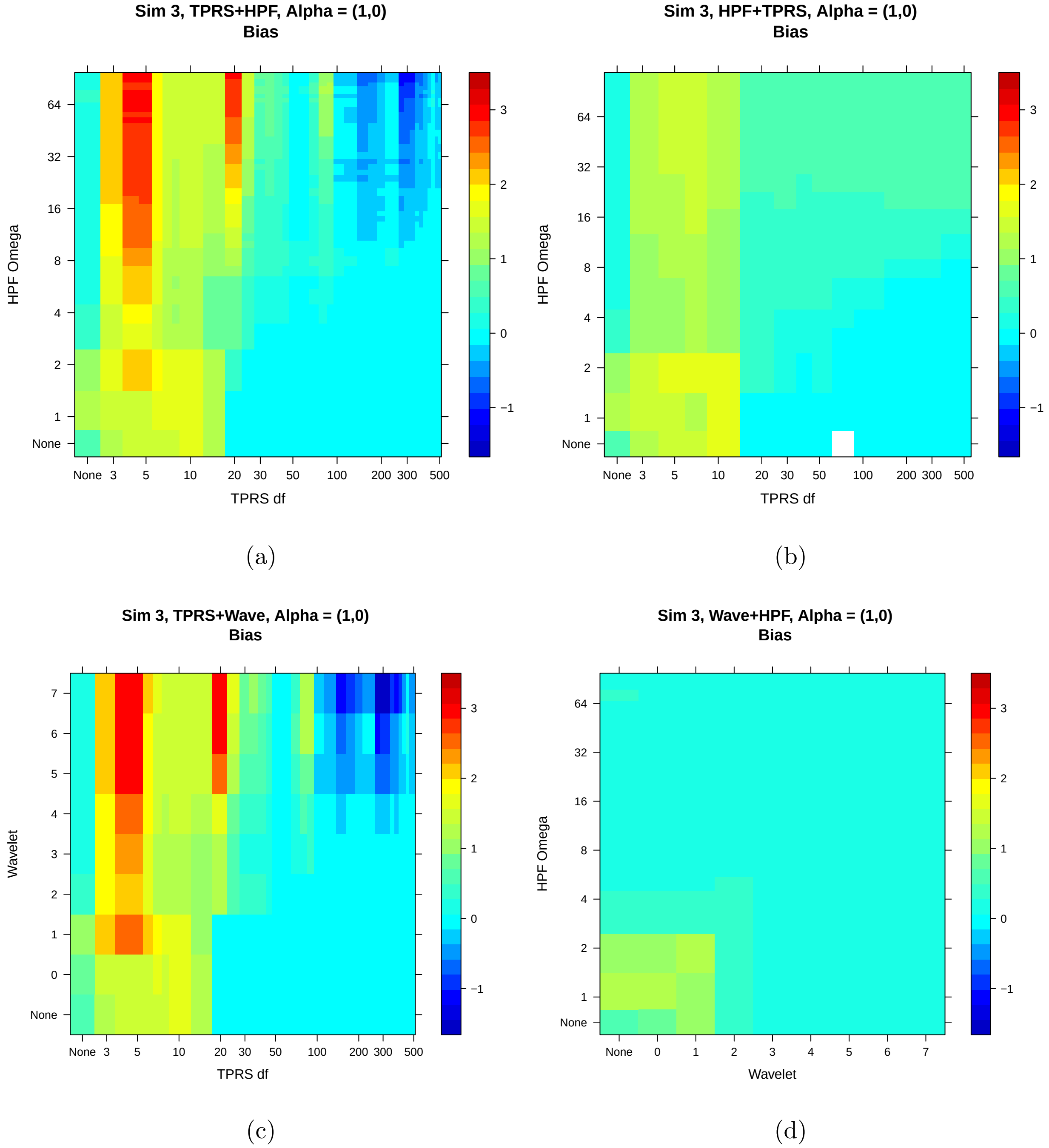
(c)

Figure C.3: Estimates of β_1 from combining pre-adjustment approaches in Simulation 1, when $(\alpha_1, \alpha_2) = (1, 0)$.

C.3 Additional Results for Simulation 3

Table C.3: Performance measures for different estimators in Simulation 3 for $(\alpha_1, \alpha_2) = (1, 1)$ setting. See Table 4.1 for an explanation of abbreviations.

	$\hat{\beta}$	Bias($\hat{\beta}$)	$SE(\hat{\beta})$	MSE($\hat{\beta}$)	$\widehat{SE}(\hat{\beta})$	Coverage
Unadjusted	1.98	0.982	0.026	0.9646	0.027	0.000
Adjusted	1.00	0.000	0.030	0.0009	0.030	0.945
HPF: $\omega = 1$	2.59	1.593	0.034	2.5402	0.034	0.000
HPF: $\omega = 2$	2.99	1.990	0.055	3.9621	0.056	0.000
HPF: $\omega = 4$	3.15	2.149	0.062	4.6209	0.063	0.000
HPF: $\omega = 16$	1.15	0.145	0.420	0.1975	0.413	0.928
HPF: $\omega = 64$	1.16	0.159	0.886	0.8095	0.891	0.952
HPF: $\omega = 128$	1.14	0.135	1.460	2.1479	1.401	0.941
TPRS: $df = 3$	2.58	1.577	0.031	2.4887	0.031	0.000
TPRS: $df = 5$	2.71	1.710	0.034	2.9237	0.034	0.000
TPRS: $df = 20$	3.17	2.173	0.063	4.7270	0.064	0.000
TPRS: $df = 50$	3.31	2.309	0.065	5.3340	0.066	0.000
TPRS: $df = 100$	3.40	2.399	0.067	5.7614	0.067	0.000
TPRS: $df = 300$	3.46	2.461	0.068	6.0599	0.068	0.000
Wave: $L = 0$	2.05	1.050	0.028	1.1025	0.028	0.000
Wave: $L = 1$	2.89	1.894	0.054	3.5897	0.053	0.000
Wave: $L = 2$	3.12	2.122	0.062	4.5082	0.063	0.000
Wave: $L = 4$	2.54	1.537	0.294	2.4487	0.287	0.001
Wave: $L = 6$	1.15	0.155	0.815	0.6876	0.810	0.946
Wave: $L = 7$	1.12	0.122	1.433	2.0653	1.417	0.948



Sim 3, TPRS+HPF, Alpha = (1,0)
Bias

(a)

Sim 3, HPF+TPRS, Alpha = (1,0)
Bias

(b)

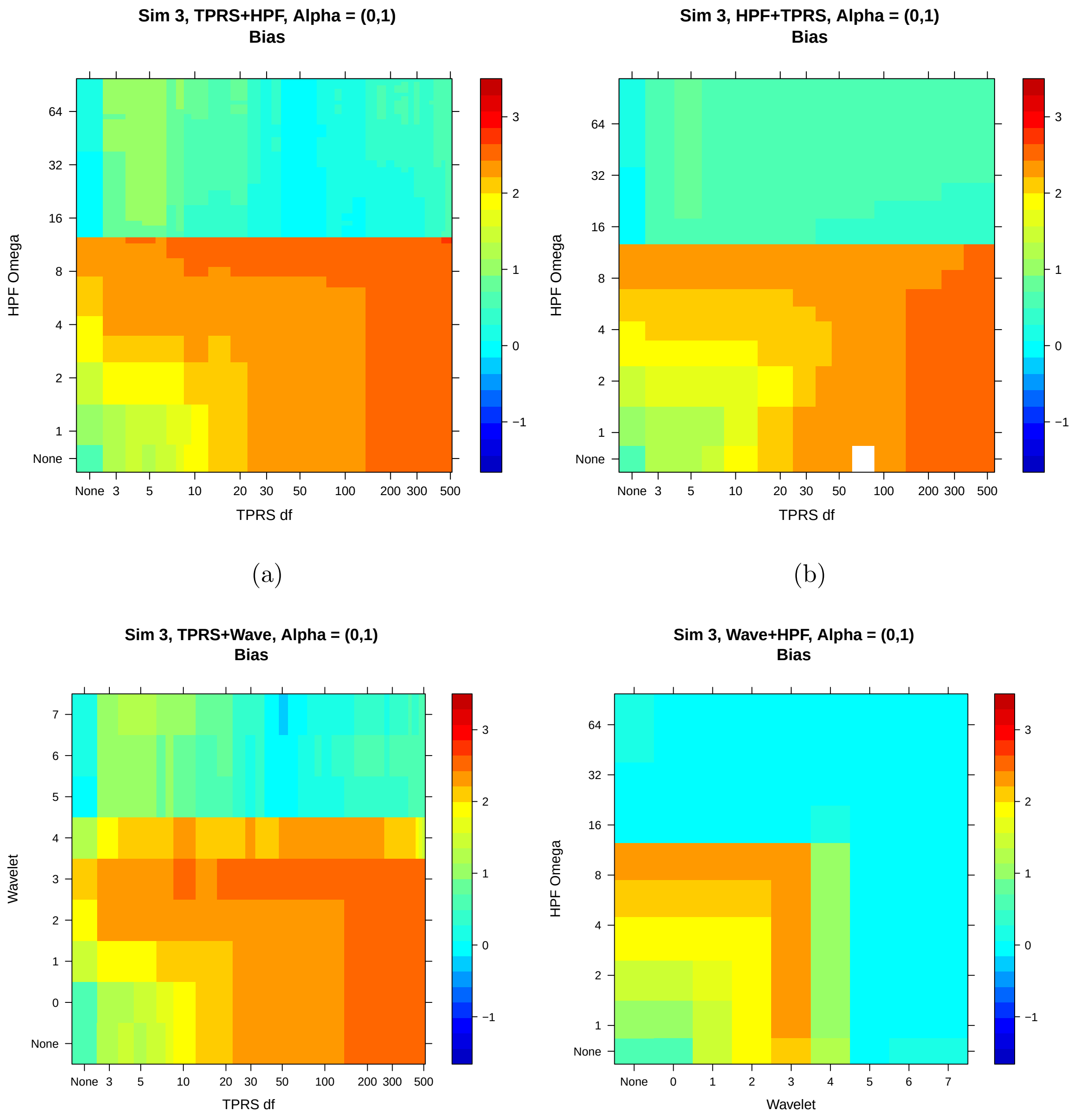
Sim 3, TPRS+Wave, Alpha = (1,0)
Bias

(c)

Sim 3, Wave+HPF, Alpha = (1,0)
Bias

(d)

Figure C.4: Estimates of β_1 from combining pre-adjustment approaches in Simulation 3, when $(\alpha_1, \alpha_2) = (1, 0)$.



(a)

(b)

(c)

(d)

Figure C.5: Estimates of β_1 from combining pre-adjustment approaches in Simulation 3, when $(\alpha_1, \alpha_2) = (0, 1)$.