An Approach to Adjust for Confounding in the Presence of Multivariate Exposures

Ander Wilson*
Department of Biostatistics, Harvard T. H. Chan School of Public Health
Corwin M. Zigler
Department of Biostatistics, Harvard T. H. Chan School of Public Health
Chirag J. Patel
Department of Biomedical Informatics, Harvard Medical School
and
Francesca Dominici
Department of Biostatistics, Harvard T. H. Chan School of Public Health

October 9, 2015

Abstract

There is increasing interest in the health effects of mixtures of environmental agents. Several large environment-wide association studies (EWAS) have estimated the association between agents and health outcomes. However, these studies have only adjusted for a small set of confounding variables. When the exposure is multivariate and may include interactions and the number of potential confounding variables is large ($p \approx n$), it is challenging to adjust for confounding and maintain model

*The authors gratefully acknowledge *NIH T32 ES007142; NCI P01 CA134294; AHRQ HS021991; NIH R21 ES022585-01; NIH R01 ES019560; NIH R21 ES020152; NIH R21 ES024012; NIH R01 GM111339; NIH R01 ES024332; NIH R35 CA197449; NIH/NIEHS K99 ES023504; NIH/NIEHS R21 ES025052; NIH Common Fund U54 HG007963*; and *PhRMA Foundation.* This publication was made possible by USEPA grant *RD-83479801.* Its contents are solely the responsibility of the grantee and do not necessarily represent the official views of the USEPA. Further, USEPA does not endorse the purchase of any commercial products or services mentioned in the publication.

parsimony. Here, we demonstrate that approaches for confounder adjustment when estimating the effect of a single agent are inadequate for a multivariate exposure. Then we develop a new approach rooted in the ideas of Bayesian model averaging to adjust for confounding in the multivariate exposure setting. We introduce an informative prior that assigns likely confounders a higher probability of being included into the model. Our approach can be formulated as a penalized likelihood that leads to a simple tuning approach. Through a simulation study we demonstrate that the proposed approach identifies parsimonious models that are fully adjusted for observed confounding. We apply the method to an EWAS study using National Heath and Nutrition Examination Survey to estimate the effect of mixtures of nutrients and pesticides on lipid levels.

Keywords: Bayesian model averaging, Confounding, Exposome, Model uncertainty, Multiple exposures, Multivariate exposure effects

1 Introduction

With the rapidly increasing availability of personal, environmental exposure data, there is growing interest in studying the multitude of exposures, often referred to as the *exposome*, that may influence complex diseases (Wild, 2005; Louis and Sundaram, 2012). Recent studies have screened large numbers of environmental agents for associations with health outcomes and biological endpoints (Patel et al., 2013; Patel and Ioannidis, 2014; Wilson et al., 2014c). Increasingly, research is focusing on estimating the health effects associated with simultaneous exposure to multiple agents, rather than estimating the effects associated with exposure to a single agent while controlling for others. For example, recent research has estimated the effect of multiple air pollutants and temperature on health outcomes (Bobb et al., 2014; Wilson et al., 2014b), the effect of multiple chemical components of particulate matter air pollution on health outcomes (Bell et al., 2009; Peng et al., 2009; Zanobetti et al., 2014; Chung et al., 2015; Kioumourtzoglou et al., 2015), and the effect of mixtures of nutrients on atherosclerosis in mice (Verschuren et al., 2011).

Estimates of the effect of exposure to one or more agents can be sensitive to the choice of confounding variables. Furthermore, we often lack prior knowledge of which confounders should be included in a regression model to estimate and unconfounded exposure effect. As such, several recent methods have been proposed to adjust for confounding when the goal is estimation of the health effects associated with exposure to a single agent. However, in the more realistic context of estimating the health effects associated with a multivariate exposure that includes multiple agents and interactions between agents (e.g. exposure $\boldsymbol{Z} = (X_1, X_2, \ldots, X_m, X_1X_2, \ldots, X_{m-1}X_m)$), the literature lacks both a rigorous definition of what constitutes a confounder and methods to adjust for confounding (Ioannidis, 2008; Patel et al., 2015).

In epidemiology, adjusting for confounding in the context of exposure to a single agent (X_1) is often addressed with sensitivity analysis or exposure modeling. Sensitivity analysis shows how the estimated health effect of exposure to a single agent varies over a range of models including different sets of potential confounders (Dominici et al., 2004; Peng et al., 2006). Exposure modeling can identify potential confounding variables that should be included in the outcome model by identifying covariates associated with the exposure (Greenland, 2008). The outcome model has the outcome as the dependent variable and the exposure agent of interests and all potential confounders as independent variables.

When estimating the effect of exposure to a single agent, joint analyses of the exposure and outcome models may provide an effective approach to identify covariates that are associated with both the exposure and the outcome and therefore are potential confounders. Recent approaches have formalized the framework in model based confounder selection methods for a single continuous agent. Crainiceanu et al. (2008) developed visualization tools and theoretical results for this approach. Bayesian adjustment for confounding (BAC; Wang et al., 2012a, 2015) uses an exposure model, an outcome model, and a joint approach for variable selection and model averaging to adjust for confounding. Bayesian penalized credible region confounder selection (BayesPen; Wilson and Reich, 2014) uses a decision-theoretic approach to find the sparsest solution that contains all important covariates and all confounding variables. Wilson and Reich (2014) also proposed an extension of BayesPen for models with multiple additive exposures. Approaches to select and adjust for confounding in the context of a single binary exposure or treatment have been proposed by Zigler and Dominici (2014). However, all these approaches which rely on exposure modeling have limitations in the context of a multivariate, continuous exposure. First, exposure modeling requires the specification of an exposure model for each single agent which is not feasible when the number of agents is large. Second, none of these methods address

confounder adjustment when interest is in a more general multivariate exposure $\boldsymbol{Z}$ that includes interactions between agents.

There is an added level of complexity in confounder adjustment when estimating the effect of a multivariate exposure on an outcome in comparison to estimating the effect of exposure to a single agent (X). When the focus is estimation of health effects associated with simultaneous exposure to multiple agents and their interactions, to properly identify and adjust for confounding we must define and treat the exposure as a multivariate vector $\boldsymbol{Z}$. As we will demonstrate, the vector of confounding variables of the effect of $\boldsymbol{Z}$ on the outcome cannot be obtained by taking the union of all confounders of the effect of each single agent ($X_1, \ldots, X_m$) and the outcome. To our knowledge there are no formal methods that explicitly identify and adjust for confounding bias in non-additive multivariate exposure models.

This paper makes several contributions. First, we show that existing methods to adjust for confounding when estimating the effect of a single agent (X) on an outcome do not fully address confounding when estimating the effect of a multivariate exposure with interaction terms ($\boldsymbol{Z}$). Second, we develop a new approach to adjust for confounding when the focus is estimating the effect of a multivariate exposure ($\boldsymbol{Z}$) on a continuous health outcome (Y) and there is a large number of potential confounders to choose from. Our goal with this model is to simultaneously eliminate confounding bias and estimate the effect of a multivariate exposure of interest under a parsimonious model that is robust to model misspecification. To balance parsimony and confounder adjustment, we provide an easy to implement tuning method that does not require computationally expensive fitting of models at multiple candidate values. Third, we apply the proposed approach to data from the National Health and Nutrition Examination Survey (NHANES) and estimate the effect of 132 agents grouped into 24 multivariate exposures representing clusters of related

nutrients and persistent pesticides on lipid levels. Our proposed approach identifies statistically significant effects that were not evident from either a model without any confounder adjustment or from a less parsimonious model that included all available covariates.

The proposed method is based on a Bayesian model averaging approach (BMA; Raftery et al., 1997) with a prior probability of including each potential confounder in the outcome model specifically designed to adjust for confounding in the context of a multivariate exposure. The prior has two desirable properties: 1) covariates that are linearly associated with both the exposure vector $\boldsymbol{Z}$ and the outcome Y have high probability of being included in the outcome model and 2) using an easily implemented tuning approach, covariates that are associated with the exposure $\boldsymbol{Z}$ only but not with the outcome (often called instrumental variables) are not forced into the model. The approach can also be formulated as a penalized likelihood problem which sheds perspective on how this approach compares to other methods and gives intuitive guidance on tuning the strength of the prior on covariate inclusion. Our proposed method is computationally simple even when the multivariate exposure becomes larger (e.g. 50 or 100 including interactions). We also develop and make available a `R` package called `ACPME`.

2 Confounding and Model Selection

2.1 Terminology of the multivariate exposure effect

We begin by defining the terminology of the multivariate exposure problem. In this paper, we are interested in estimating the change in a health outcome associated with a change in a multivariate exposure while controlling for a potentially high dimensional vector of covariates. We formalize the problem in terms of the *agents* we are studying, the *inter-*

vention we want to infer about, and the *multivariate exposure* vector that is of scientific interest.

The agents are the singular components that comprise the multivariate exposure. As an example throughout the paper, and part of the data analysis, we consider the effect of two agents–the persistent pesticides 1-phenanthrene and 3-fluoranthene–on the outcome high-density lipoprotein-cholesterol (HDL). More generally, we denote the vector of m agents as $\boldsymbol{X} = (X_1, X_2, \ldots, X_m)$. In the data analysis presented in Section 5 we consider $m = 132$ agents, clustered into 24 groups, each group consists of between 1 and 22 related agents.

Rather than hypothesizing a scenario where individuals are exposed to one agent at the time, we consider the more realistic scenario where individuals are exposed to multiple agents simultaneously. We define the multivariate exposure as a vector $\boldsymbol{Z} = z(\boldsymbol{X})$ of length $r \geq m$. In the pesticide example $\boldsymbol{Z} = (X_1, X_2, X_1X_2)$, where X_1=1-phenanthrene and X_2=3-fluoranthene.

The intervention, denoted by $\boldsymbol{\delta}$, is a pre-specified and known change in the vector of agents from $\boldsymbol{X} = \boldsymbol{x}$ to $\boldsymbol{X} = \boldsymbol{x} + \boldsymbol{\delta}$. We assume that the intervention $\boldsymbol{\delta}$ affects the outcome through a change in the multivariate exposure from $z(\boldsymbol{x})$ to $z(\boldsymbol{x}+\boldsymbol{\delta})$. The fixed and known value of $\boldsymbol{\delta}$ could be a real intervention that has changed the concentration of the agents such a dietary change which is likely to affect exposure to several nutrients and pesticide simultaneously (Bradman et al., 2015) or reduction in air pollution mixture as a result of interventions that occurred prior the Olympics in Bejing, China (Dominici and Mittleman, 2012; Rich et al., 2012). Alternatively, the intervention may be a hypothetical change which we want to infer about (Bobb et al., 2013; Wilson et al., 2014b; Snowden et al., 2015).

The exposure effect $\Delta_{\boldsymbol{x},\boldsymbol{\delta}}$ is the change in outcome Y associated with the intervention, conditional on the vector of the potential confounders $\boldsymbol{C}$,

$$\Delta_{\boldsymbol{x},\boldsymbol{\delta}} = E\left(Y|\boldsymbol{X} = \boldsymbol{x} + \boldsymbol{\delta}, \boldsymbol{C}\right) - E\left(Y|\boldsymbol{X} = \boldsymbol{x}, \boldsymbol{C}\right), \tag{1}$$

where the effect of the intervention acts on Y through the change in the multivariate exposure $\boldsymbol{Z}$. As such, (1) can be equivalently expressed as $\Delta_{\boldsymbol{x},\boldsymbol{\delta}} = E\left[Y|\boldsymbol{Z} = z(\boldsymbol{x}+\boldsymbol{\delta}), \boldsymbol{C}\right] - E\left[Y|\boldsymbol{Z} = z(\boldsymbol{x}), \boldsymbol{C}\right]$. The estimand $\Delta_{\boldsymbol{x},\boldsymbol{\delta}}$ can be interpreted as the effect of the intervention (i.e., the induced change in multivariate exposure) provided that $\boldsymbol{C}$ contain all relevant confounders.

The exposure effect is commonly estimated with a parametric regression model and confounder adjustment becomes a confounder selection problem. In this paper we focus on the linear model

$$Y_i = \boldsymbol{\beta}_0 + \boldsymbol{Z}_i^T\boldsymbol{\beta} + \sum_{j=1}^{k} C_{ji}\eta_j + \epsilon_i, \tag{2}$$

where $\boldsymbol{\beta}$ is a vector of unknown regression coefficients for the exposure of dimension $r \geq m$, $\boldsymbol{\eta} = (\eta_1, \ldots, \eta_k)^T$ are unknown regression coefficients for the k covariates, and ϵ_i are iid $\mathrm{N}(0, \sigma^2)$ for individuals $i = 1, \ldots, n$.

Under model (2), the exposure effect is $\Delta_{\boldsymbol{x},\boldsymbol{\delta}} = \left[z(\boldsymbol{x}+\boldsymbol{\delta}) - z(\boldsymbol{x})\right]^T\boldsymbol{\beta}$. For the nutrient example $\Delta_{\boldsymbol{x},\boldsymbol{\delta}}$ is defined as $\Delta_{\boldsymbol{x},\boldsymbol{\delta}} = \beta_1\delta_1 + \beta_2\delta_2 + \beta_3(x_1\delta_2 + x_2\delta_1 + \delta_1\delta_2)$. For notational simplicity we will write $\Delta_{\boldsymbol{x},\boldsymbol{\delta}}$ as $\boldsymbol{d}^T\boldsymbol{\beta}$ where $\boldsymbol{d}$ is a known quantity determined by the function of agents that comprises the exposure in (2) and the shift in agents corresponding to the intervention.

For exposure to a single agent, Wang et al. (2012a) define the minimal set of confounders $\boldsymbol{C}^*$ to be the subset of the measured covariates $\boldsymbol{C}$ that are associated with both the agent X and the outcome Y. In the multivariate exposure setting we extend the definition of the minimal set to be all covariates that are associated with the multivariate exposure $\boldsymbol{Z}$ and outcome Y. When the regression model (2) includes $\boldsymbol{C}^*$ or any larger set of $\boldsymbol{C}$ that includes the minimal set $\boldsymbol{C}^*$ then $\widehat{\Delta}_{\boldsymbol{x},\boldsymbol{\delta}}$ is unconfounded and $\Delta_{\boldsymbol{x},\boldsymbol{\delta}}$ can be interpreted as the effect of $\boldsymbol{Z}$ on Y. In this paper, we identify a prior distribution that assign high prior probability of inclusion to the observed covariates that are included in $\boldsymbol{C}^*$.

2.2 Exposure to a single agent versus multivariate exposure

In this section we demonstrate that we cannot identify the confounders of the effect of a multivariate exposure $\boldsymbol{Z}$ on outcome Y by taking the union of the confounders of the effect of exposure to each single agent (e.g. X_1 and X_2) on Y.

A covariate can be a confounder of the effect of $\boldsymbol{Z}$ on Y but not a confounder the effect of any single agent (e.g. X_1 or X_2) on Y, if that covariate is correlated with a function of agents (e.g. the interaction X_1X_2) but it is not correlated with X_1 nor with X_2. More generally, this can occur when the covariate is balanced across levels of each individual agent but not balanced across levels of combinations of agents.

More specifically, consider the pesticide example used in Section 2.1 where we are interested in estimating the health effect of an intervention on 1-phenanthrene and 3-fluoranthene (X_1 and X_2, respectively) on an outcome HDL-cholesterol (Y) adjusted by the potential confounder C_1 which we now define as and indicator of any previous cardiovascular event. In the NHANES data, the variables X_1 and X_2 are balanced between subjects with and without a previous cardiovascular event (standardized mean difference equal to -0.12 and 0.05). However the variable X_1X_2 is not, and has a standardized mean difference of -0.86. Hence, a situation may arise in data analysis where a potential confounder is only associated with the interaction between two agents but not with either main effect.

Confounder adjustment methods that rely on exposure modeling (e.g. Crainiceanu et al., 2008; Wang et al., 2012a; Wilson and Reich, 2014; Wang et al., 2015) do not address this issue. These methods identify potential confounders by specify an exposure model for each agent (one for X_1 and one for X_2), where the single agent is the dependent variable and all the potential confounders are the independent variables. Hence, potential confounders associated with the interaction terms only, but not with the main effect, will not be identified as true confounders in these single agent exposure models.

To demonstrate this phenomenon, we construct a simple hypothetical example where we can calculate the closed-form confounding bias. We assume there are two agents $\boldsymbol{X} = (X_1, X_2)$, a multivariate exposure that includes an interaction $\boldsymbol{Z} = (X_1, X_2, X_1X_2)$, and a single covariate C_1 which we now assume is continuous. We assume that C_1 and X_1 are both distributed independent $\mathrm{N}(0, 1)$. As is common in environmental epidemiology X_1 and C_1 might affect the level of the other exposure X_2. We assume X_2 is $\mathrm{N}(X_1 + X_1C_1, 1)$. Finally, the outcome Y is distributed $\mathrm{N}(X_1 + X_2 + X_1X_2 + C_1, 1)$. In this case C_1 is linearly associated with X_1X_2 (covariance 1) and with Y, but not with X_1 or X_2 (because C_1 and X_1 are independent and centered on 0). We also assume that we are interested in estimating $\Delta_{\mathbf{0},\mathbf{1}}$, here defined as the effect associated with a change from no exposure to one unit of both agents (X_1 and X_2).

Because C_1 is not linearly associated with X_1 or X_2, any approach that relies on exposure modeling of X_1 and X_2 with two separate exposure models (e.g. Wang et al., 2012a; Wilson and Reich, 2014; Wang et al., 2015) will not identify C_1 as a confounder and, therefore, will fail to identify the necessary confounders of the effect of $\boldsymbol{Z}$ on Y. Under model (2), the true $\Delta_{\mathbf{0},\mathbf{1}} = 3$. However, the model without C_1 has $E[\widehat{\Delta}_{\mathbf{0},\mathbf{1}}|\boldsymbol{X}] = 3.17$ and is therefore biased whereas the $E[\widehat{\Delta}_{\mathbf{0},\mathbf{1}}|\boldsymbol{X}, C_1] = 3$ and is unbiased. Hence, the union of confounders identified by the collection of single agent exposure models does not include the full set of confounders for the association between the multivariate exposure ($\boldsymbol{Z}$) and the outcome (Y). In summary, new methods are needed for confounder adjustment in the multiple exposure setting that adapt to the multivariate exposure effect being estimated.

3 Model

3.1 Approach

We begin by introducing the "augmented" linear regression model as used by Raftery et al. (1997). We define a vector of unknown parameters $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_k)^T$ where $\alpha_j \in \{0, 1\}$ indicates whether covariate C_j is included into the model (3) defined below:

$$Y_i = \boldsymbol{\beta}_0 + \boldsymbol{Z}_i^T \boldsymbol{\beta} + \sum_{j=1}^{k} C_{ji} \alpha_j \eta_j + \epsilon_i. \tag{3}$$

Conditionally on a pre-specified model indexed by $\boldsymbol{\alpha}$ which identifies the pre-specified set of covariate $\{C_j : \alpha_j = 1\}$ that are included in (3), we can estimate the posterior distribution of the exposure effect $\Pr[\Delta_{\boldsymbol{x},\boldsymbol{\delta}} | \boldsymbol{Z}, \boldsymbol{C}, \boldsymbol{Y}, \boldsymbol{\alpha}]$ by using standard methods for Bayesian linear regression. To account for model uncertainty about the specification of confounders we estimate the posterior distribution of $\Delta_{\boldsymbol{x},\boldsymbol{\delta}}$ as

$$\Pr(\Delta_{\boldsymbol{x},\boldsymbol{\delta}} | \boldsymbol{Z}, \boldsymbol{C}, \boldsymbol{Y}) = \sum_{\boldsymbol{\alpha}} \Pr(\Delta_{\boldsymbol{x},\boldsymbol{\delta}} | \boldsymbol{Z}, \boldsymbol{C}, \boldsymbol{Y}, \boldsymbol{\alpha}) \Pr(\boldsymbol{\alpha} | \boldsymbol{Z}, \boldsymbol{C}, \boldsymbol{Y}), \tag{4}$$

where $\Pr(\boldsymbol{\alpha} | \boldsymbol{Z}, \boldsymbol{C}, \boldsymbol{Y}) \propto \Pr(\boldsymbol{Y} | \boldsymbol{Z}, \boldsymbol{C}, \boldsymbol{\alpha}) \Pr(\boldsymbol{\alpha})$ (Raftery et al., 1997; Hoeting et al., 1999).

In practice the prior distribution $\Pr(\boldsymbol{\alpha})$ is commonly assumed to be non-informative $\Pr(\alpha_j = 1) = 1 - \Pr(\alpha_j = 0) = 0.5$ and $j = 1, \ldots, p$. Wang et al. (2012a) showed that applying BMA with this non-informative prior on $\boldsymbol{\alpha}$ will select covariates that will be associated with Y only and, therefore, will minimize prediction error for Y but the selection of these covariates will not minimize the confounding bias for the estimated exposure effect $\widehat{\Delta}_{\boldsymbol{x},\boldsymbol{\delta}}$.

Wang et al. (2012a) demonstrated that the confounding bias can be eliminated by averaging only over the subset of models that contain the minimal set of confounders $\boldsymbol{C}^*$

and that the interpretation of $\Delta_{\boldsymbol{x},\boldsymbol{\delta}}$ remains the same for all models that include $\boldsymbol{C}^*$. We denote the minimal model as $\boldsymbol{\alpha}^*$, which denotes the vector of indicator variables that identify $\boldsymbol{C}^*$. In this section, we show how to effectively specify an informative prior for $\boldsymbol{\alpha}$ so that high prior probability is assigned to models that contain the minimal model $\boldsymbol{\alpha}^*$ in the context of a multivariate exposure $\boldsymbol{Z}$. Under this informative prior specification for $\boldsymbol{\alpha}$, the posterior distribution of $\Delta_{\boldsymbol{x},\boldsymbol{\delta}}$ as defined in (4) will be a weighted average of models $\boldsymbol{\alpha}$ that include the minimal model $\boldsymbol{\alpha}^*$ and therefore minimize confounding bias.

3.2 Construction of prior to adjust for confounding

In this section we introduce an informative prior on $\boldsymbol{\alpha}$, so that we can estimate $\Delta_{\boldsymbol{x},\boldsymbol{\delta}}$ by averaging across models that include the minimal set of confounders $\boldsymbol{C}^*$. We also compare the proposed prior with alternative choices in Supplemental Section 1.1. We start by calculating the confounding bias in $\widehat{\Delta}_{\boldsymbol{x},\boldsymbol{\delta}}$ that would occur if we would fit model (3) without any adjustment, that is, under $\boldsymbol{\alpha} = \mathbf{0}$.

We assume that Y is centered and omit the intercept from the model. We define as confounding bias the difference between the posterior mean of $\boldsymbol{d}^T\boldsymbol{\beta}$ under a model that includes all observed covariates ($\boldsymbol{\alpha} = \mathbf{1}$) and posterior mean of $\boldsymbol{d}^T\boldsymbol{\beta}$ under a model that does not include any observed covariates ($\boldsymbol{\alpha} = \mathbf{0}$). Under a flat prior on $\boldsymbol{\beta}$ and $\boldsymbol{\eta}$ and inverse-gamma prior on σ^2 (with fixed hyper-parameters), the confounding bias of $\widehat{\Delta}_{\boldsymbol{x},\boldsymbol{\delta}} = \boldsymbol{d}^T\widehat{\boldsymbol{\beta}}$ is

$$\begin{aligned} &E\left(\widehat{\Delta}_{\boldsymbol{x},\boldsymbol{\delta}}|\boldsymbol{Y},\boldsymbol{Z},\boldsymbol{C},\boldsymbol{\alpha}=\mathbf{1}\right) - E\left(\widehat{\Delta}_{\boldsymbol{x},\boldsymbol{\delta}}|\boldsymbol{Y},\boldsymbol{Z},\boldsymbol{C},\boldsymbol{\alpha}=\mathbf{0}\right) \\ &\qquad = \boldsymbol{d}^T\left(\boldsymbol{Z}^T\boldsymbol{Z}\right)^{-1}\boldsymbol{Z}^T\boldsymbol{C}\left(\boldsymbol{C}^T\boldsymbol{P}_Z^{\perp}\boldsymbol{C}\right)^{-1}\boldsymbol{C}^T\boldsymbol{P}_Z^{\perp}\boldsymbol{Y}, \end{aligned} \tag{5}$$

where $\boldsymbol{P}_Z = \boldsymbol{Z}(\boldsymbol{Z}^T\boldsymbol{Z})^{-1}\boldsymbol{Z}^T$ is the perpendicular projection onto the column space of $\boldsymbol{Z}$ and $\boldsymbol{P}_Z^{\perp} = \boldsymbol{I} - \boldsymbol{P}_Z$. The full details of (5) are in the Supplemental Section 1.2. The confounding

bias in (5) is zero when $\left(\boldsymbol{Z}^T\boldsymbol{Z}\right)^{-1}\boldsymbol{Z}^T\boldsymbol{C} = \boldsymbol{0}$ or $\boldsymbol{C}^T\boldsymbol{P}_Z^{\perp}\boldsymbol{Y} = \boldsymbol{0}$ and, therefore, is zero when $\boldsymbol{C}$ is independent of either $\boldsymbol{Z}$ or Y.

Under the assumption that the shift in exposure induced by the intervention is in the column space of $\boldsymbol{Z}$, we can write $\boldsymbol{d} = \boldsymbol{a}\boldsymbol{Z}^T$ for some vector $\boldsymbol{a}$, and the confounding bias defined in (5) is equal to 0 if the following is equal to 0:

$$\begin{aligned}&\left\|E\left(\boldsymbol{Z}^T\boldsymbol{\beta}|\boldsymbol{Y},\boldsymbol{Z},\boldsymbol{C},\boldsymbol{\alpha}=\boldsymbol{0}\right) - E\left(\boldsymbol{Z}^T\boldsymbol{\beta}|\boldsymbol{Y},\boldsymbol{Z},\boldsymbol{C},\boldsymbol{\alpha}=\boldsymbol{1}\right)\right\|_2^2 \\ &\qquad = \sum_{l=1}^{r}\left[\sqrt{\zeta_l}\boldsymbol{q}_l^T\left(\boldsymbol{C}^T\boldsymbol{P}_Z^{\perp}\boldsymbol{C}\right)^{-1}\boldsymbol{C}^T\boldsymbol{P}_Z^{\perp}\boldsymbol{Y}\right]^2, \qquad (6)\end{aligned}$$

where $\boldsymbol{C}^T\boldsymbol{P}_Z\boldsymbol{C} = \sum_{l=1}^{k}\zeta_l\boldsymbol{q}_l\boldsymbol{q}_l^T$ is the spectral decomposition (see the Supplemental Section 1.3 for details on this calculation). The simplifying assumption that $\boldsymbol{d} = \boldsymbol{a}\boldsymbol{Z}^T$ is satisfied for many real and hypothetical interventions (Rich et al., 2012; Bobb et al., 2013; Wilson et al., 2014b; Bradman et al., 2015); however, we have found that the proposed method is still highly effective when this assumption is not met as shown by our simulation.

We construct our informative prior for $\boldsymbol{\alpha}$ by 1) isolating the part of (6) that is a function only of $\boldsymbol{Z}$ and $\boldsymbol{C}$, so as not to have the prior depend on the outcome Y, and 2) assigning higher prior inclusion probabilities to the covariates that are linearly dependent with $\boldsymbol{Z}$ and therefore contribute most to the confounding bias expression in (6). Specifically, we assume that α_j has *a priori* a Bernouilli distribution with mean $\pi_j(\boldsymbol{Z},\boldsymbol{C},\lambda) = \text{Logit}^{-1}\left\{\lambda\omega_j(\boldsymbol{Z},\boldsymbol{C})\right\}$, where $\boldsymbol{\omega}\left(\boldsymbol{Z},\boldsymbol{C}\right) = \sum_{l=1}^{r}\left[\sqrt{\zeta_l}\boldsymbol{q}_l^T\left(\boldsymbol{C}^T\boldsymbol{P}_Z^{\perp}\boldsymbol{C}\right)^{-1}\right]^2$ is a k-vector taken from (6) and $\lambda \geq 0$ is a tuning parameter. Hence, $\lambda\omega_j\left(\boldsymbol{Z},\boldsymbol{C}\right)$ is the log-odds of including C_j in the regression model (3).

This prior specification has the following desirable properties. If C_j is linearly independent of $\boldsymbol{Z}$ then $\omega_j\left(\boldsymbol{Z},\boldsymbol{C}\right) = 0$ and the prior probability of including C_j is $\Pr(\alpha_j = 1|\boldsymbol{Z},\boldsymbol{C},\lambda) = 0.5$. If C_j is linearly dependent with $\boldsymbol{Z}$ then $\omega_j\left(\boldsymbol{Z},\boldsymbol{C}\right) > 0$, and $\Pr(\alpha_j = 1|\boldsymbol{Z},\boldsymbol{C},\lambda) \in [0.5, 1)$ is proportional to the strength of this association. When $\lambda > 0$ and

C_j is linearly dependent with $\boldsymbol{Z}$ then the inequality is strict, $\Pr(\alpha_j = 1|\boldsymbol{Z}, \boldsymbol{C}, \lambda) \in (0.5, 1)$. The stronger the association between C_j and $\boldsymbol{Z}$, the larger the potential confounding bias that could result from omitting C_j, and the larger $\omega_j(\boldsymbol{Z}, \boldsymbol{C})$. Therefore, a covariate C_j that has high potential to cause confounding bias if omitted is assigned higher prior probability of inclusion than a covariate that has relatively low potential to cause confounding bias if omitted.

Another key requirement in building our informative prior for $\boldsymbol{\alpha}$ is to be able to exclude covariates that are associated only with $\boldsymbol{Z}$ and are independent from Y. We address this issue with the tuning parameter λ which controls the strength of the prior. When $\lambda \to \infty$ we assume that any covariate that is correlated with $\boldsymbol{Z}$ (even weakly) is forced into the model regardless of whether it is associated with Y (an undesirable property for a prior). In other words, in this situation (large λ) we prioritize confounding adjustment maximally over model parsimony. On the other hand, when $\lambda = 0$ we assume that the prior on $\boldsymbol{\alpha}$ is flat, $\pi_j(\boldsymbol{Z}, \boldsymbol{C}, 0) = 0.5 \ \forall j$ regardless the amount of correlation between $\boldsymbol{Z}$ and $\boldsymbol{C}$, and the prior provides no improvement to confounder adjustment. In this situation ($\lambda = 0$) we prioritize model parsimony maximally over confounding adjustment, as is the case with most variable selection approaches (e.g. Tibshirani, 1996). Hence, the structure of the prior comes from $\omega_j(\boldsymbol{Z}, \boldsymbol{C})$ and the strength of the prior from λ. We discuss choices for λ in Section 3.4 and recommend a tuning method that balances confounder adjustment and model parsimony.

We complete the Bayesian specification for the normal linear model following that of Raftery et al. (1997) and provide details in Web Supplement Section 1.4. The R package ACPME (tinyurl.com/nhe3ft4) provides software to use this method and reproduce the simulated data.

3.3 Relation to a penalized likelihood approach

The proposed approach can also be formulated as a maximization of a penalized likelihood where the penalty of the likelihood will reflect the two properties of the prior distribution $\Pr(\boldsymbol{\alpha}|\boldsymbol{Z}, \boldsymbol{C}, \lambda)$ as defined above (minimizing confounding bias and maximizing model parsimony). Importantly, this formulation will guide the selection of the tuning parameter λ.

We approximate the posterior model probability of $\Pr(\boldsymbol{\alpha}|\boldsymbol{Y}, \boldsymbol{X}, \boldsymbol{C}) \propto \Pr(\boldsymbol{Y}|\boldsymbol{Z}, \boldsymbol{C}, \boldsymbol{\alpha}) \times \Pr(\boldsymbol{\alpha}|\boldsymbol{Z}, \boldsymbol{C}, \lambda)$ using the Bayesian information criterion (BIC; Schwarz, 1978) as

$$\text{BIC}(\boldsymbol{\alpha}) = -2 \times ll(\boldsymbol{Y}, \boldsymbol{X}, \boldsymbol{C}; \boldsymbol{\alpha}, \widehat{\boldsymbol{\beta}}_{\boldsymbol{\alpha}}, \widehat{\boldsymbol{\eta}}_{\boldsymbol{\alpha}}, \widehat{\sigma}^2_{\boldsymbol{\alpha}}) - 2\lambda \sum_{j=1}^{k} \alpha_j \omega_j (\boldsymbol{Z}, \boldsymbol{C}) + \log(n) \sum_{j=1}^{k} \alpha_j. \quad (7)$$

The three terms in (7) are, from left to right, 1) the negative profile log-likelihood of the normal linear regression model from (2) which is minimized when covariates predictive of Y are included into the regression model, 2) the negative prior log-odds of covariate inclusion defined in Section 3.1, and 3) the BIC sparsity penalty. Hence, the proposed method optimizes model fit while balancing confounder adjustment and model parsimony.

The second and third terms of (7) can be combined and viewed as a single penalty on the log-likelihood. Hence, the penalty for including covariate j is $\alpha_j [\log(n) - 2\lambda\omega_j (\boldsymbol{Z}, \boldsymbol{C})]$. The tuning parameter λ controls the balance between confounder adjustment (large λ) and model parsimony (small λ). A unique feature of this penalty is that it can be either positive, acting as a penalty, or negative, acting as an incentive for covariate inclusion.

3.4 Choice of tuning parameter

We are interested in finding a value of λ that balances confounder adjustment with parsimony. The BIC approximation of the posterior probability defined in (7) provides a

theoretical basis to select λ such that we can achieve balance between parsimony and confounder adjustment. Specifically, we balance the competing interests of parsimony and confounder adjustment by choosing λ so that the penalty $\alpha_j \left[\log(n) - 2\lambda\omega_j \left(\boldsymbol{Z}, \boldsymbol{C}\right)\right]$ is on average zero. Setting $0 = k^{-1} \sum_{j=1}^{k} \alpha_j \left[\log(n) - 2\lambda\omega_j \left(\boldsymbol{Z}, \boldsymbol{C}\right)\right]$ yields the balancing penalty

$$\lambda^* = \frac{k \log(n)}{2 \sum_{j=1}^{k} \omega_j \left(\boldsymbol{Z}, \boldsymbol{C}\right)}. \tag{8}$$

In practice, $\lambda = \lambda^*$ can be seen as a benchmark choice. By taking $\lambda > \lambda^*$ we include more covariates and take a more conservative approach to confounder adjustment at the expense of less parsimony. Alternatively $\lambda < \lambda^*$ prioritizes sparsity over confounder adjustment.

4 Simulation

4.1 Simulation overview

We evaluate the performance of the proposed method with a simulation study using two scenarios. The two scenarios demonstrate the ability of the newly proposed method to properly adjust for confounding in the simple context of only 2 agents and their interaction and in the more realistic and complex situation of an exposure of up to 10 agents and all the 45 possible pairwise interactions.

4.2 Simulation scenario one

The first simulation scenario includes two agents and their interaction, $\boldsymbol{Z} = (X_1, X_2, X_1X_2)^T$, plus 100 covariates $\boldsymbol{C}$ with $n \in \{200, 500\}$ observations. We have simulated the data so that: $C_1, \ldots, C_{15}$, are associated with Y and X_1 and/or X_2 (confounders associated with

exposure main effects); $C_{16}, \ldots, C_{20}$ are correlated with Y and X_1X_2 (confounders associated with the interaction X_1X_2 only); $C_{21}, \ldots, C_{30}$ are associated with Y but not with $\boldsymbol{Z}$ (predictors of Y); $C_{31}, \ldots, C_{35}$ are associated with $\boldsymbol{Z}$ but not Y and should not be included in the model; and $C_{36}, \ldots, C_{100}$ are independent of both Y and $\boldsymbol{Z}$ (noise). Hence, the true model contains only covariates $C_1, \ldots, C_{30}$ and the minimal set $\boldsymbol{C}^*$ is $\{C_1, \ldots, C_{20}\}$. For each sample size we simulate 1000 data sets as follows:

$$\begin{aligned}
C_{ji} &\sim \mathrm{N}(0,1) \quad \text{for } j = 1, \ldots, 100 \\
X_{1i} &\sim \mathrm{N}\left(11^{-1/2} \sum_{j=1}^{10} \boldsymbol{C}_{ji}, 11^{-1/2}\right) \\
X_{2i} &\sim \mathrm{N}\left(21^{-1/2} \left[\sum_{j=6}^{15} \boldsymbol{C}_{ji} + X_{1i} \sum_{j=16}^{50} \boldsymbol{C}_{ji} + \sum_{j=31}^{35} \boldsymbol{C}_{ji}\right], 21^{-1/2}\right) \\
Y_i &\sim \mathrm{N}\left(\boldsymbol{Z}\boldsymbol{\beta} + \sum_{j=1}^{30} \eta_j \boldsymbol{C}_{ji}, 1\right). \qquad (9)
\end{aligned}$$

For X_{1i} and X_{2i} the regression coefficients $11^{-1/2}$ and $21^{-1/2}$ are chosen so that both agents have variance 1. The regression coefficients $\beta_1, \beta_2, \beta_3$ and $\{\eta_j\}_{j=1}^{30}$, are simulated as independent Uniform$(0.2, 0.5)$. Supplemental Figure 2 illustrates the correlation structure in the data.

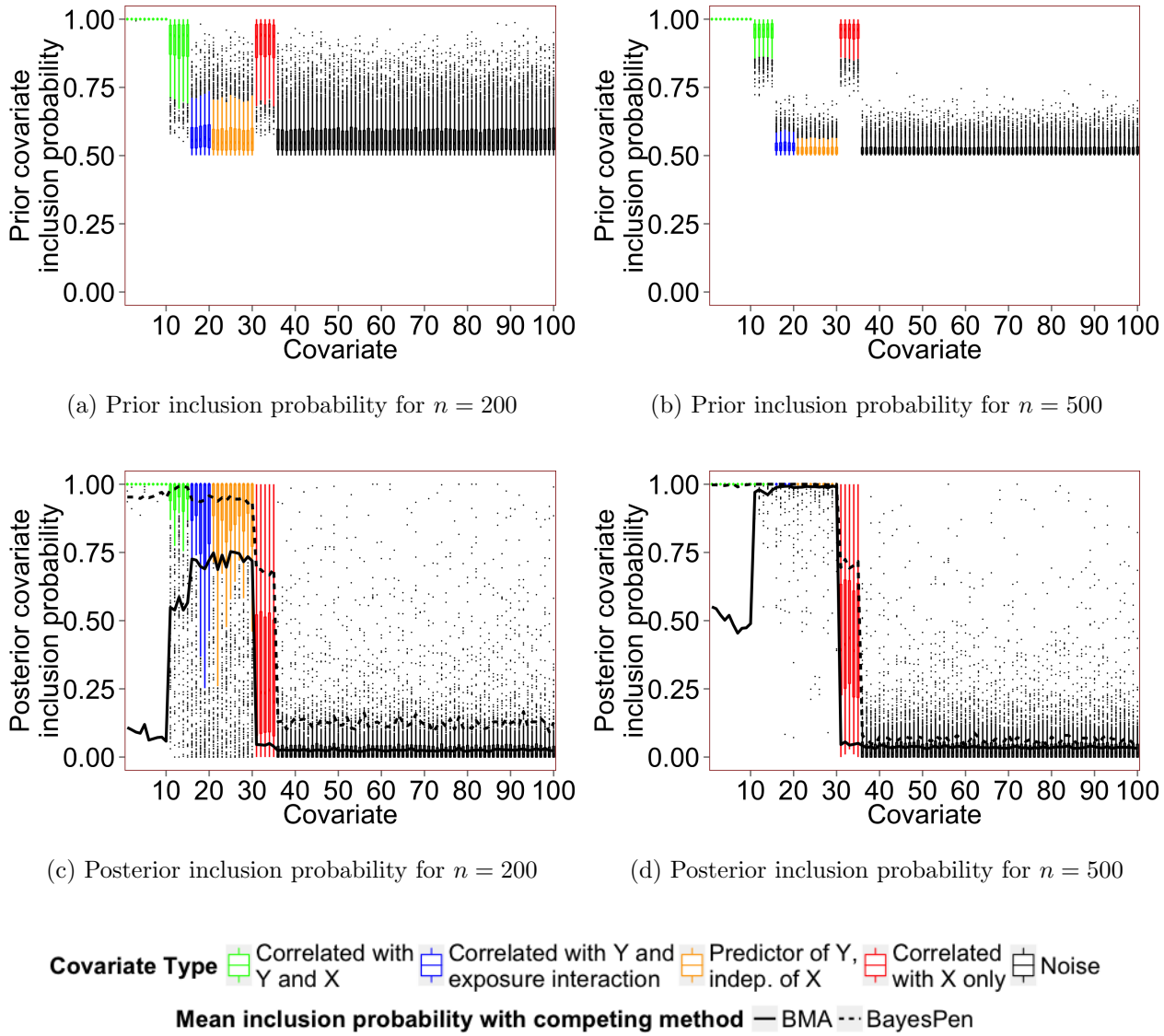
We compare the proposed method, henceforth ACPME (adjustment for confounding in the presence of multivariate exposures), to five alternatives: BMA (equivalent to ACPME with $\lambda = 0$, that is $\Pr(\alpha_j = 1 | \boldsymbol{Z}, \boldsymbol{C}, \lambda) = 0.5 \ \forall j$); Bayesian penalized credible regions (BayesPen) implemented with the Wilson et al. (2014a) `R` package and using one exposure model for each agent but not for the interactions; the full Bayesian regression model that includes as potential confounders all 100 measured covariates; the true Bayesian regression model that controls for covariates 1 to 30 only; and the unadjusted Bayesian regression model that regresses the outcome on the multivariate exposure $\boldsymbol{Z}$ with no covariate adjust-

ment. For ACPME we use $\lambda = \lambda^*$ as described in Section 3.4. We estimate the effect of a simultaneous change in both agents from 0 to 1, $\Delta_{\mathbf{0},\mathbf{1}} = \beta_1 + \beta_2 + \beta_3$.

Figure 1 shows the prior (panels 1a and 1b) and posterior (panels 1c and 1d) covariate inclusion probabilities with ACPME. The posterior inclusion rates for minimal set of confounder (covariates 1 to 20) under the ACPME approach are all near one. While ACPME identifies the covariates that are only associated with the exposure and not the outcome (covariates 31 to 35) as potential confounders and assigns high prior probability, these covariates have average posterior inclusion probability of less than 0.5. This demonstrates that setting the tuning parameter to λ^* as defined in Section 3.4 balances the goals of parsimony and confounder adjustment and does not force covariates into the model that are only associated with the exposure and not the outcome.

For comparison, Figures 1c and 1d show the mean posterior inclusion probability with BMA and proportion of times each covariate is selected with BayesPen. BMA fails to adjust for important confounders with high probability. BayesPen selects true confounders with high probability but has a higher average rate of including covariates that are independent of the outcome (covariates 31 to 100) and should not be included in the model, relative to ACPME.

Table 1 show results for estimating the exposure effect $\Delta_{\mathbf{0},\mathbf{1}}$. ACPME has lower root mean square error (RMSE) compared to all the alternatives except for the true model at both sample sizes. In addition, ACPME has credible interval coverage near or at the nominal level indicating the correct posterior inference is being made. BMA with a flat prior provides a biased estimate of $\Delta_{\mathbf{0},\mathbf{1}}$ and has larger RMSE indicating that important confounders are omitted and highlighting the effectiveness of the ACPME informative prior. BayesPen had lower RMSE than the full model but higher than ACPME indicating the importance of treating the exposure as multivariate and accounting for interactions. ACPME



(a) Prior inclusion probability for $n = 200$

(b) Prior inclusion probability for $n = 500$

(c) Posterior inclusion probability for $n = 200$

(d) Posterior inclusion probability for $n = 500$

Figure 1: Prior (top) and posterior (bottom) inclusion probabilities for each covariate in simulation scenario one for our proposed method. The box plots show the distribution of prior and posterior probabilities across 1000 simulated data sets. For comparison, the lines in panels 1c and 1d show the average posterior inclusion probability for BMA (solid line) and the proportion of times each covariate was selected into the model with BayesPen (dashed line).

is additionally benefitted by accounting for uncertainty in the model choice. The unadjusted model performs very poorly and is incapable of making proper inference on the exposure effect, indicating the extent of confounding in this design.

4.3 Simulation with large number of agents

We conduct a second simulation to evaluate the performance of ACPME relative to alternative methods in the context of a high-dimensional multivariate exposure for two sample sizes ($n = 200$ and $n = 500$). We generate a new data set for several specifications of $\boldsymbol{Z}$, where we allow the number of agents to vary from $m = 2, \ldots, 10$ and include all pairwise interactions so that $\boldsymbol{Z}$ has $m + m(m-1)/2$ columns (ranging from 3 to 55 including main effects and interactions).

For each sample size we have simulated 100 covariates so that: $C_1, \ldots, C_{15}$ are confounders associated with at least one of the agents $X_1, \ldots, X_m$ and Y; $C_{16}, \ldots, C_{25}$ are confounders associated with Y and at least one of the interactions between agents; $C_{26}, \ldots, C_{30}$ are predictors of Y and are independent of $\boldsymbol{Z}$; and $C_{31}, \ldots, C_{100}$ are independent of both outcome and exposure. Hence, the true model includes covariates $C_1, \ldots, C_{30}$ and the minimal set of confounders is $C_1, \ldots, C_{25}$. The specifics of the generating method are as

Table 1: Simulation results for simulation scenario 1. The first four columns show the mean bias, mean RMSE, mean posterior SD or SE, and 95% interval coverage rate. The right most columns show statistics for covariate inclusion–the true inclusion rate defined as the mean probability that the true confounders and predictors of the outcome (covariates 1 to 30) are included into the regression model and the false selection rate defined as the mean probability that covariates independent of the outcome are included in the model (covariates 31 to 100). Covariates are consider included if they have posterior inclusion probability exceeding 0.5.

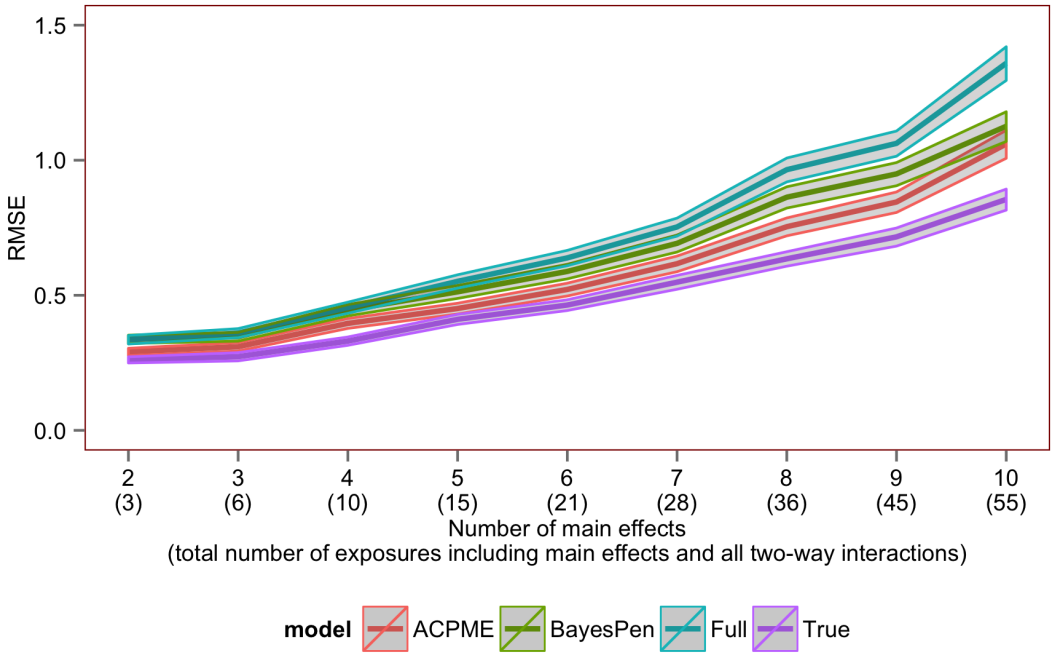
Method	Bias	RMSE	Mean SD / SE	95% Int. Coverage	True Inc. Rate	False Sel. Rate
$n = 200$						
ACPME	0.07	0.34	0.34	0.94	0.89	0.06
BayesPen	0.11	0.42	0.28	0.78	0.96	0.17
BMA	1.23	1.25	0.17	0.00	0.48	0.03
Unadjusted	1.65	1.66	0.21	0.00	0.00	0.00
Full	0.00	0.43	0.44	0.95	1.00	1.00
True	0.00	0.28	0.29	0.96	1.00	0.00
$n = 500$						
ACPME	0.02	0.18	0.19	0.96	1.00	0.07
BayesPen	0.03	0.21	0.18	0.91	1.00	0.11
BMA	0.66	0.78	0.15	0.24	0.84	0.04
Unadjusted	1.64	1.65	0.13	0.00	0.00	0.00
Full	0.02	0.21	0.21	0.96	1.00	1.00
True	0.01	0.17	0.17	0.95	1.00	0.00

follows:

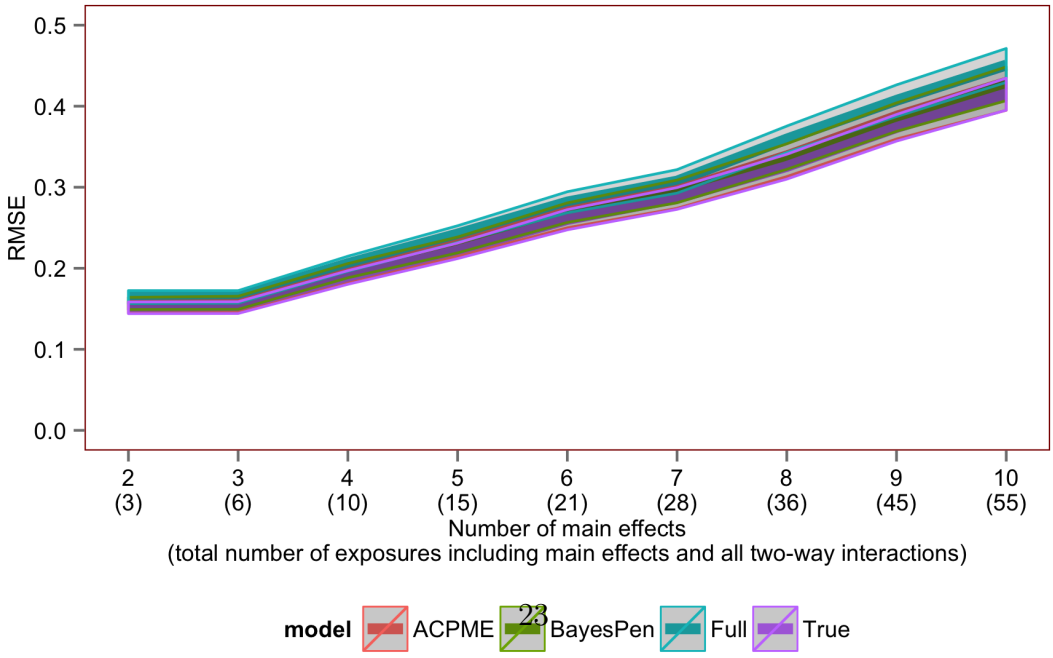
$$
\begin{aligned}
C_{ji} &\sim \mathrm{N}(0,1) \quad \text{for } j = 1, \ldots, 100 \\
h_j &\sim \mathrm{Cat}(1, \ldots, m_1) \quad \text{for } j = 1, \ldots, 10 \\
h_j &\sim \mathrm{Cat}(1, \ldots, m) \quad \text{for } j = 11, \ldots, 25 \\
q_j &\sim \mathrm{Cat}(m_1 + 1, \ldots, m_2) \quad \text{for } j = 1, \ldots, 15 \\
X^*_{ki} &\sim \mathrm{N}\left(\sum_{j=1}^{10} C_{ji} \mathbb{1}\{h_j = k\} X_{q_j i} + \sum_{j=11}^{25} C_{ji} \mathbb{1}\{h_j = k\}, 1 \right) \\
Y_i &\sim \mathrm{N}\left(\boldsymbol{Z}\boldsymbol{\beta} + \sum_{j=1}^{30} \eta_j \boldsymbol{C}_{ji}, 1 \right), \qquad (10)
\end{aligned}
$$

where $\mathrm{Cat}(1, \ldots, l)$ indicates a categorical random variable that takes integer values $1, \ldots, l$ with equal probability $1/l$, $m_1 = m/2$ rounded down to the nearest integer, $m_2 = m - m_1$, and, X_{ki} is X^*_{ki} scaled to have variance 1 and $\boldsymbol{Z}$ includes all $\boldsymbol{X}$ and all parities interactions. The regression coefficients $\{\beta_l\}_{l=1}^{r}$ and $\{\eta_j\}_{j=1}^{30}$, are independent Uniform$(0.2, 0.5)$. Supplemental Figure 3 illustrates the correlation structure in the data.

Figure 2 shows the RMSE for estimating the exposure effect of a one unit increase in each agent, $\Delta_{\mathbf{0},\mathbf{1}}$. At the smaller sample size ($n = 200$), ACPME has lower RMSE than BayesPen and both these two approaches had lower RMSE than the full model. The RMSE of ACPME improves relative to the alternative approaches as the dimension of the exposure increases. At $n = 500$, ACPME has near identical RMSE to the true model and there are no statistically significance differences between the methods. This demonstrates that the ACPME informative prior captures information about confounding covariates and the choice of tuning parameter λ^* is appropriate even when the multivariate exposure is large relative to sample size.



(a) $n = 200$



(b) $n = 500$

Figure 2: RMSE of the estimated exposure effect for simulation scenario two for $n = 200$ (panel 2a) and $n = 500$ (panel 2b). The x-axis shows the number of agents and in parentheses the dimension of the multivariate exposure including all main effects and two-way interactions. In all cases there are 100 additional covariates in the model of which 30 are true confounders or and predictors of the outcome that should be included in the model.

5 Data Analysis

5.1 Analysis and data overview

We apply ACPME to the NHANES data. Patel et al. (2012) provide a detailed description of the NHANES data. We briefly recap here and note important differences needed for the multivariate analysis. The data combines laboratory and questioner data from the 1999-2000, 2001-2002, 2003-2004, and 2005-2006 surveys. Each survey is a non-overlapping sample that is representative of the general US populations. The data include measurements of blood serum and urine biomarkers of 132 nutrients and persistent pesticides and three outcomes denoting three types of lipid levels that might be affected by these multivariate exposures. The three outcomes are: 1) low-density lipoprotein-cholesterol (LDL), 2) high-density lipoprotein-cholesterol (HDL), and 3) triglyceride. Patel et al. (2012) previously screened these 132 agents independently for their marginal associations with each of the three outcomes described above in an environment-wide association study (EWAS). The EWAS study controlled for individual covariates (such as age, gender, and BMI) but not for exposure to the other nutrients and persistent pesticides. However, as noted by Patel and Ioannidis (2014) multiple agents are often highly correlated and confounding may underlie many of the strong correlations observed in EWAS studies. Hence, it is important to account for confounding in a multivariate exposure setting.

In the present analysis, we group the 132 agents into 24 mutually exclusive exposure groups of related agents, defined by Patel et al. (2012), that may effect the same biological pathways. We are interested in the effect of exposure to all agents and to all their pairwise interactions within each group on each of the three outcomes (LDL, HDL, and triglyceride), adjusted for confounding by agents in other groups and individual level covariates. Hence, for each group we define the multivariate exposure as the main effect of each agent and all

pairwise interactions. The intervention is a one unit increase in each scaled agent within the group, equivalent to a one standard deviation increase of the agents on log scale. We conduct separate analyses to estimate the exposure effect of each of the 24 groups on the 3 outcomes.

Different exposures were measured for different subsets of sample frame. As such, the number number of persons with available measurements for lipid levels, nutrient, and persistent pesticide exposures ranged by exposure group. Because we are interested in a joint analysis of multiple agents we limit our analysis to a subset of individuals with complete data on multiple agents as described in supplemental materials Section 3.1. Table 2 shows the 24 exposure groups, the sample size (n ranging from 158 to 1370), number of agents in each group (m ranging from 1 to 22), and the number of covariates included as potential confounders (k ranging from 22 to 92). In addition to the agents in the other groups that are measured in that subsample, the potential confounders include: nine body measurements (weight; standing height; body mass index; upper leg length; maximal calf circumference; waist circumference; thigh circumference; triceps skinfold; and subscapular skinfold) and 13 demographic and socioeconomic status variables (age; age squared; poverty to income ratio; indicator for any heard disease; indicator of at least one chronic disease; indicators for race/ethnicity: black, Mexican-American, other hispanic, other race/ethnicity; indicator for female; indicators for SES tertile; indicators for education: less than high school, high school, or more than high school). In the last column of Figure 2 we shows the ratio of p divided by n, which ranges from 0.77 to 0.03. Supplemental Figure 4 shows which groups are included as covariates for the analysis of other exposures.

Our approach of identifying subsets of individuals with data on multiple exposures may results in selection bias. We discuss this in the supplement and estimates that have been identified as potentially biased due to our selection approach are presented as faded in the

Table 2: Summary of the NHANES data by exposure group. The table shows the total number of subjects with complete observations (n); the number of agents (m); the dimension of the multivariate exposure including the main effect of each agent and each two-way interaction (r); the total number of potential confounders including the main effect of agents in other groups (k); the total number of independent variables including the multivariate exposure, potential confounders, and an intercept ($p = r + k + 1$); and the p/n ratio.

Exposure Group	Sample Size (n)	# Agents (m)	Dim($\boldsymbol{Z}$) (r)	# Potential Confounders (k)	# Indep. Variables (p)	p to n Ratio (p/n)
volatile compounds	179	10	55	82	138	0.77
pcbs	558	22	253	59	313	0.56
phenols	179	3	6	92	99	0.55
dioxins	201	5	15	83	99	0.49
furans - dibenzofuran	201	5	15	83	99	0.49
pest. - pyrethyroid	201	1	1	87	89	0.44
diakyl	225	6	21	76	98	0.44
pest. - phenols	225	4	9	78	88	0.39
pest. - chloroacetanilide	225	1	1	81	83	0.37
pest. - organophosphate	225	1	1	81	83	0.37
heavy metals	444	13	91	48	140	0.32
phthalates	387	11	66	51	118	0.30
pest. - organochlorine	288	7	28	53	82	0.28
nutrients - minerals	158	2	3	37	41	0.26
polyflourochemicals	444	10	55	51	107	0.24
hydrocarbons	292	9	45	22	68	0.23
phytoestrogens	432	6	21	49	71	0.16
nutrients - vitamin C	444	1	1	60	62	0.14
nutrients - carotenoid	1370	5	15	33	49	0.04
nutrients - vitamin A	1370	3	6	35	42	0.03
nutrients - vitamin B	1370	3	6	35	42	0.03
nutrients - vitamin E	1370	2	3	36	40	0.03
cotinine	1370	1	1	37	39	0.03
nutrients - vitamin D	1370	1	1	37	39	0.03

analysis figures and are not discussed in the results.

5.2 Comparison of estimates to the full and unadjusted models

We estimate the multivariate exposure effect with ACPME, the full model including all k potential confounders, and the unadjusted model where we regress the lipid levels on the multivariate exposures corresponding to each of the 24 groups without any additional covariate adjustment. Figure 3 presents the point estimates and 95% posterior intervals.

To highlight the advantage of using ACPME we focus on exposure effect estimates for volatile compounds. In this case there are $n = 179$ individuals, $k = 82$ potential confounders, $m = 10$ agents and all one way interactions, hence $p = 138$ independent variables. Please note that the volatile compounds group is "interesting" because the p to n ratio is 0.77. When $n >> p$ then the full model works fine but when $p \approx n$ there is a need for a more parsimonious model. The effect of volatile compounds on HDL and triglyceride both change sign with confounder adjustment using ACPME and with the full model relative to the unadjusted model. In addition, using ACPME to select a more parsimonious model resulted about a 30% decrease in posterior standard deviation compared to the full model for all three outcomes as shown in Figure 4. Hence, for volatile compounds ACPME resulted in lower variance estimates of the exposure effect that are fully adjusted for confounding.

In general, the ACPME point estimates are similar to the full model. This suggests that all important confounders are included in ACPME. While ACPME and the full model produce similar posterior means, the unadjusted model, which does not adjust for any confounding, produces notably different posterior means in several cases suggesting that the unadjusted estimates are confounded. The ACPME estimates had smaller posterior variance on average than estimates from the full model due to the decreased model size as

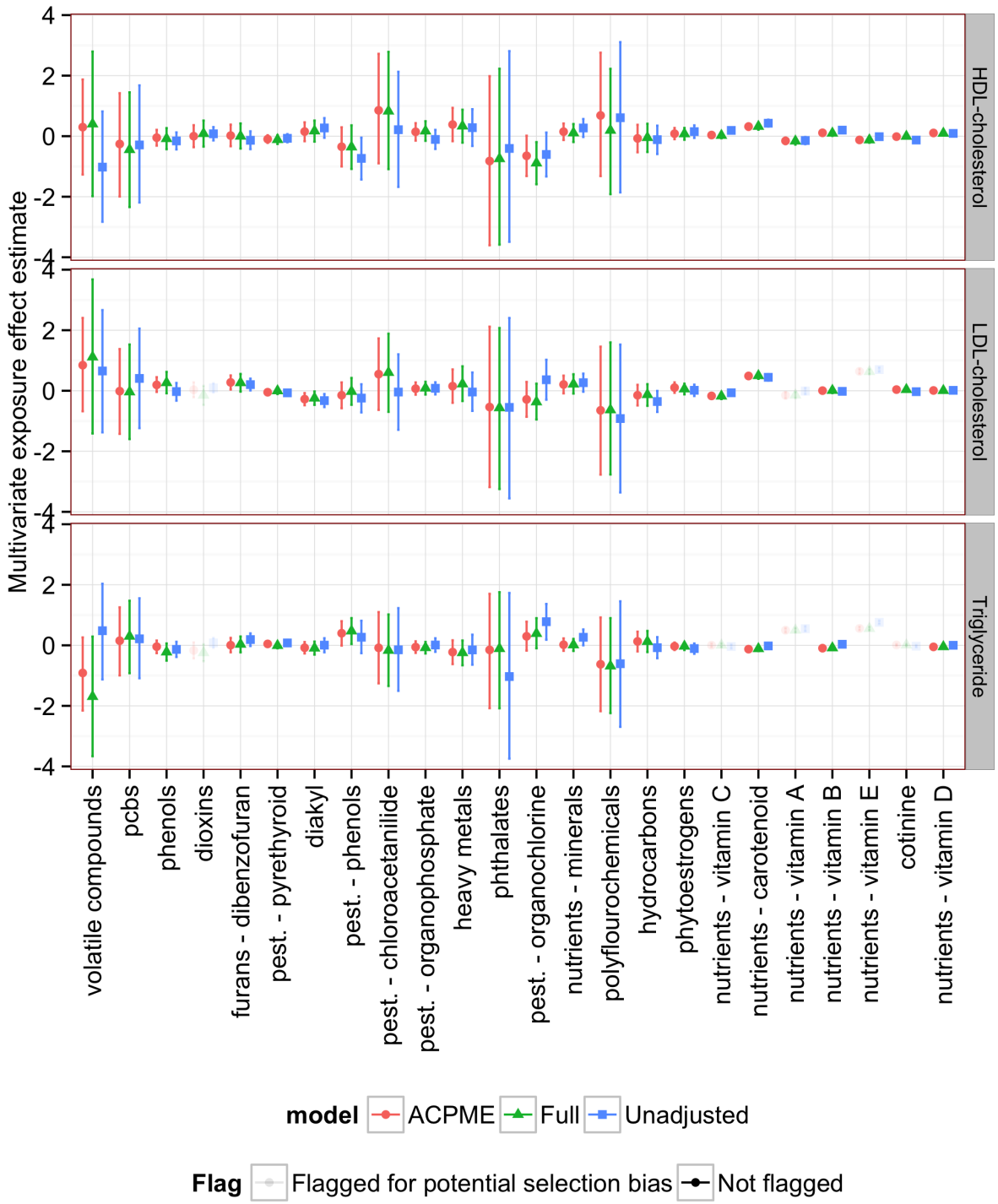


Figure 3: Point estimates and 95% credible intervals of the association between the multivariate exposure $\boldsymbol{Z}$ and each of the 3 outcome adjusted by the exposure to the other 23 groups and baseline covariates. Results are reported under the ACPME model, the full model ($\boldsymbol{\alpha} = \mathbf{1}$), and the unadjusted model (model that includes $\boldsymbol{Z}$ as the only independent variables). Faded estimates were flagged for potentially selection bias.

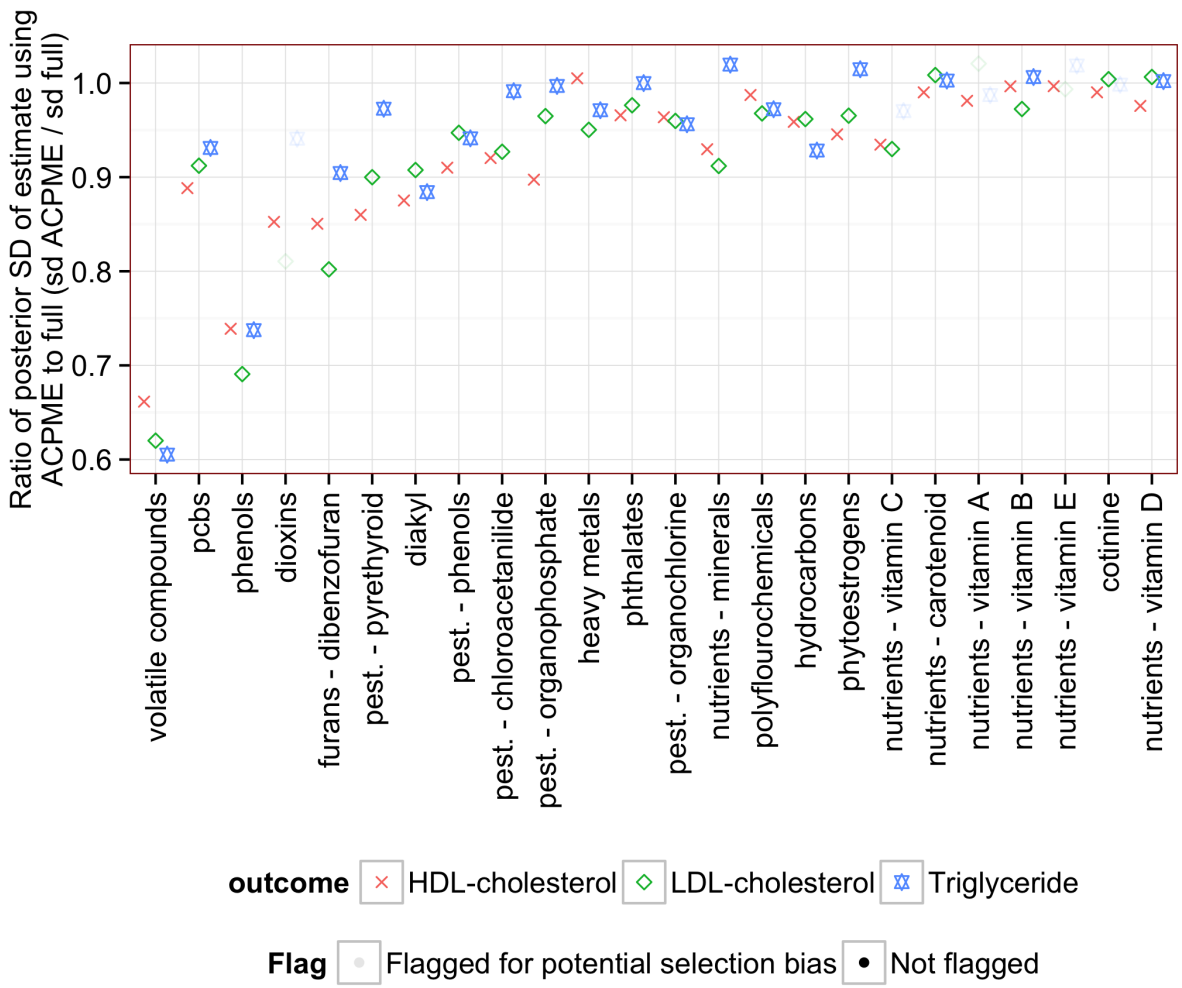


Figure 4: Ratio of the posterior standard deviation of exposure effect estimate under the ACPME model compared to the full model. Results are showed ranked by groups with the largest (left) to the smallest (right) p/n. In general, the estimates from ACPME have lower SD because of the reduced model size. Faded estimates were flagged for potentially selection bias.

seen in Figure 4. Hence, ACPME appears to fully adjust for observed confounding in these data while decreasing posterior variance.

5.3 Multivariate exposures associated with lipid levels

In this section we look at the effect of confounder adjustment on the significance level of the multivariate exposure effect. Figure 5 presents the posterior probability that each exposure group results in a change (either positive or negative) in lipid level on negative log scale. We define the posterior probability of a change as the highest probability symmetric credible interval that does not contain zero. The horizontal lines indicate 0.05 and 0.01 significance levels after Bonferroni adjustment for multiple comparisons.

Using the unadjusted model, four groups (carotenoid, vitamin B, vitamin C, and cottoning) had a statistically significant effect on HDL levels and one group (carotenoid) has a statistically significant effect on LDL levels. However, only two–the effect of carotenoid on HDL and LDL levels–remain significant after confounder adjustment with either ACPME or the fully adjusted model. Figure 3 shows the point estimates for the other three groups all shrink toward zero when covariates are included in the model. Hence, three of the five are likely false discovers as a result of confounding.

There are two cases where the posterior probability of a significant exposure effect substantially increased in the more parsimonious ACPME model compared to the full model. Vitamin C was significant for LDL at the 0.01 level using ACPME but only at the 0.05 level with the full model and not at all with the unadjusted model. This highlights a new result not reported in Patel et al. (2012) where the analysis did not control for other exposures. Figure 3 shows that vitamin C is negatively associated with LDL levels and the posterior mean shifts away from 0 after confounder adjustment, while figure Figure 4 show a small decrease in posterior deviation using the more parsimonious ACPME com-

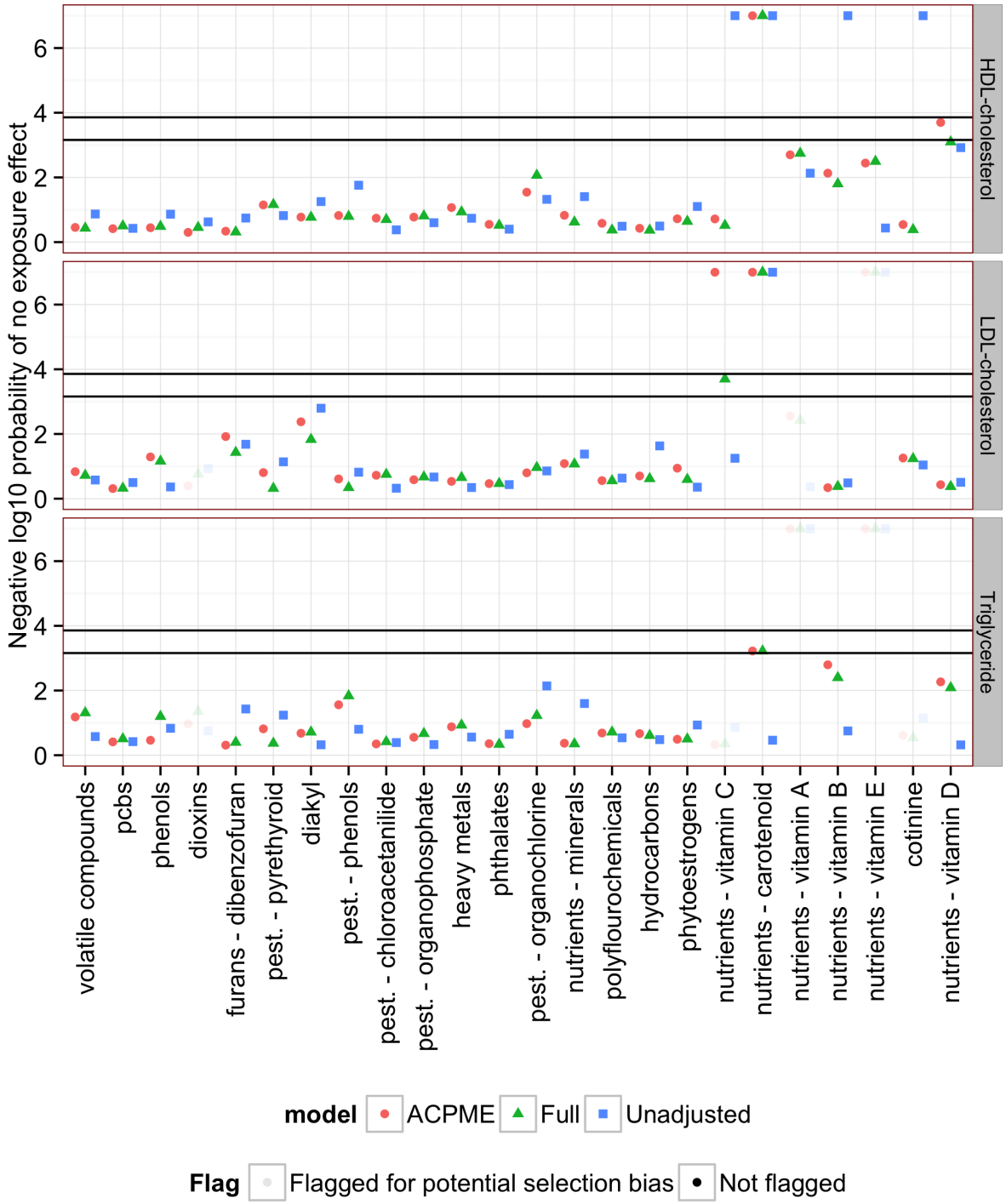


Figure 5: Posterior probability that the multivariate exposure effect is non-zero for each exposure group, outcome, and method on -$\log_{10}$ scale. The black horizontal lines indicate the 0.05 and 0.01 significance level for the multivariate exposure effect. The probability levels are adjusted using Bonferroni corrections for 72 tests, the total number of tests calculated with each method. Faded estimates were flagged for potentially selection bias.

pared the full model. ACPME identified several agents as important confounders: two E vitamins (α-tocopherol, γ-tocopherol), two carotenoids (combined lutein/zeaxanthin and thans-lycophene), and Folate. In addition to these agents body mass index, standing height, and weight were also included in the model with high posterior probability. Finally. the effect of vitamin D on HLD levels is significant at the 0.05 level only with ACPME. Combined lutein/zeaxanthin and γ-tocopherol were identified as important adjusted to adjust for confounding in this case as well.

6 Discussion

In this paper we address the problem of confounder adjustment when the goal is estimating the effect of a multivariate exposure comprised of several agents and their two-way interactions in presence of a large set of potential confounders. We differentiate the problem of confounder adjustment for a multivariate exposure from the more straightforward situation of confounding adjustment in the context of exposure to a single agent. To our knowledge, the proposed method is the first to explicitly address confounding of a multivariate exposure that includes interactions.

The approach presented here relies on the BMA framework and adjusts for observed confounding with a carefully specified informative prior on covariate inclusion. Using the proposed method, the model includes the minimal set of confounders of the association between $\boldsymbol{Z}$ and Y with high posterior probability. Thus, the model yields an unconfounded estimate of the effect of the intervention. At the same time it allows covariates that are associated with the exposure $\boldsymbol{Z}$ only or with neither $\boldsymbol{Z}$ nor Y to be excluded from the regression model. As shown by our simulation study the proposed method yields a parsimonious model that is fully adjusted for confounding and performs better compared to

recently developed alternatives for confounding adjustment in the context of a multivariate exposure.

We analyze the effect of exposure to 132 nutrient and persistent pesticides grouped into 24 groups on lipid levels using the proposed approach. The analysis provided fully adjusted estimates of the exposure effects under more parsimonious models resulting in smaller average standard errors, especially when p is close to n. As a result, we identified two groups of significant exposures that were not evident without the proposed method: the effects of vitamin C on LDL and vitamin D on HDL. We also identified multivariate exposures that are significant in their association with the outcome absent confounding adjustment but lose statistical significance when the estimate was adjusted for the exposure to the other 23 groups and baseline and demographic covariates

The proposed method uses information from the relationship between the potential confounders ($\boldsymbol{C}$) and the multivariate exposure ($\boldsymbol{Z}$) to construct the prior. Therefore the ACPME prior does not depend on the outcome and indeed builds on other previous work that uses the covariate space to construct a prior. Perhaps most well known is Zellner's g-prior where the covariance structure of the prior comes from the covariates but the strength of the prior is user-specified (Zellner, 1986). In our case, the relative ordering of the prior inclusion probabilities on each covariate (C_j) is determined by the covariate space ($\boldsymbol{Z}$ and $\boldsymbol{C}$) while the strength of the prior is defined by the user through the tuning parameter (λ).

The proposed method offers several advantages over previously developed confounder adjustment methods. First, our approach scales as the number of agents increases and can be applied when we are interested in including into $\boldsymbol{Z}$ higher level interactions. Previous confounder adjustment approaches have relied on exposure modeling (Wang et al., 2012a; Wilson and Reich, 2014; Wang et al., 2015). Our proposed method does not use and exposure model. This provides an advantage when the number of agents is large. Take, for

example, the PCB exposure group in our data analysis where the number of single agents in this group is $m = 22$. Using the multiple exposure model presented by Wilson and Reich (2014) would require the specification of 22 separate exposure models for each of the 22 PCB agents and it would still miss identifying the confounders of any of 231 possible two-ways interactions between any pair of the 22 agents. In contrast, the approach proposed here naturally scales with the number of agents and automatically adapts to adjust for confounding when we consider the multitude of interactions between the 22 agents.

A second advantage of the proposed method is that it explicitly addresses confounding of interaction terms. This has not been previously addressed in the literature. We described both real data and hypothetical scenarios where a potential confounder is associated with the interaction between individual agents but not associated with any of the individual agents. In addition to adjusting for confounding when the multivariate exposure includes interactions between agents the proposed method could be applied in other cases, for example, basis expansions of one or more agents or quadratic terms.

Another advantage is the specific guidance provided for tuning the strength of the prior on covariate inclusion. We give a choice of tuning parameter (λ^*) that can be easily calculated *a priori*, balances model parsimony and confounder adjustment, and performed well in both the simulation and data analysis. In contrast, the method of Wang et al. (2012a) showed sensitivity to choice of tuning parameters but did not give specific tuning guidance as discussed by Vansteelandt (2012) and Wang et al. (2012b). Wilson and Reich (2014) provide an approach to tuning in the single agent case that relies refitting several models and post-hoc selecting the best choice for tuning parameter.

A fourth advantage of the proposed method is that it allows covariates that are associated with the exposure but not the outcome to be dropped from the model. In contrast, Wang et al. (2012a) explicitly forces covariates associated with the exposure to be included

in the outcome model when the tuning parameter approaches infinity (which is the case that receives the most focus in that work). These covariates, sometimes called instrumental variables, should not be included in the analysis as they can increase the variance of the estimated exposure effect.

With increased availability of high-dimensional personal exposure data (called exposome) there is growing interest in understanding the effect of the exposome on complex diseases. However, there is often uncertainty as to which covariates to include in the model to eliminate confounding bias. The proposed method fills a methodological gap to adjust for confounding when estimating multivariate exposure effects. This method is a valuable tool that can reliably be used in EWAS to estimate health effects of mixtures while simultaneously allowing a rigorous adjustment for confounding and guarantee model parsimony.

SUPPLEMENTARY MATERIAL

Title: Supplement to An Approach to Adjust for Confounding in the Presence of Multivariate Exposures. (pdf)

References

Bell, M. L., Ebisu, K., Peng, R. D., Samet, J. M., and Dominici, F. (2009). Hospital admissions and chemical composition of fine particle air pollution. *American Journal of Respiratory and Critical Care Medicine*, 179(12):1115–1120.

Bobb, J. F., Dominici, F., and Peng, R. D. (2013). Reduced hierarchical models with application to estimating health effects of simultaneous exposure to multiple pollutants. *Journal of the Royal Statistical Society: Series C*, 62(3):451–472.

Bobb, J. F., Valeri, L., Claus Henn, B., Christiani, D. C., Wright, R. O., Mazumdar, M., Godleski, J. J., and Coull, B. A. (2014). Bayesian kernel machine regression for estimating the health effects of multi-pollutant mixtures. *Biostatistics*, pages 1–16.

Bradman, A., Quirós-Alcalá, L., Castorina, R., Aguilar Schall, R., Camacho, J., Holland, N. T., Barr, D. B., and Eskenazi, B. (2015). Effect of Organic Diet Intervention on Pesticide Exposures in Young Children Living in Low-Income Urban and Agricultural Communities. *Environmental Health Perspectives*, 123.

Chung, Y., Dominici, F., Wang, Y., Coull, B. A., and Bell, M. L. (2015). Associations between Long-Term Exposure to Chemical Constituents of Fine Particulate Matter (PM2.5) and Mortality in Medicare Enrollees in the Eastern United States. *Environmental Health Perspectives*, 123(5):467–474.

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

Dominici, F., McDermott, A., and Hastie, T. J. (2004). Improved semiparametric time series models of air pollution and mortality. *Journal of the American Statistical Association*, 99(468):938–948.

Dominici, F. and Mittleman, M. A. (2012). China's Air Quality Dilemma: Reconciling Economic Growth With Environmental Protection. *JAMA*, 307(19):2100–2102.

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

Hoeting, J. A., Madigan, D., Raftery, A. E., and Volinsky, C. T. (1999). Bayesian Model Averaging: A Tutorial. *Statistical Science*, 14(4):382–417.

Ioannidis, J. P. A. (2008). Why most discovered true associations are inflated. *Epidemiology*, 19(5):640–648.

Kioumourtzoglou, M.-A., Austin, E., Koutrakis, P., Dominici, F., Schwartz, J., and Zanobetti, A. (2015). PM2.5 and Survival Among Older Adults. *Epidemiology*, 26(3):321–327.

Louis, G. M. B. and Sundaram, R. (2012). Exposome: Time for transformative research. *Statistics in Medicine*, 31(22):2569–2575.

Patel, C. J., Burford, B., and Ioannidis, J. P. (2015). Assessment of vibration of effects due to model specification can demonstrate the instability of observational associations. *Journal of Clinical Epidemiology*, 68(9):1046–1058.

Patel, C. J., Cullen, M. R., Ioannidis, J. P. A., and Butte, A. J. (2012). Systematic evaluation of environmental factors: Persistent pollutants and nutrients correlated with serum lipid levels. *International Journal of Epidemiology*, 41(3):828–843.

Patel, C. J. and Ioannidis, J. P. A. (2014). Studying the elusive environment in large scale. *JAMA*, 311(21):2173–4.

Patel, C. J., Rehkopf, D. H., Leppert, J. T., Bortz, W. M., Cullen, M. R., Chertow, G. M., and Ioannidis, J. P. A. (2013). Systematic evaluation of environmental and behavioural factors associated with all-cause mortality in the united states national health and nutrition examination survey. *International Journal of Epidemiology*, 42(6):1795–1810.

Peng, R. D., Bell, M. L., Geyh, A. S., McDermott, A., Zeger, S. L., Samet, J. M., and Dominici, F. (2009). Emergency admissions for cardiovascular and respiratory diseases

and the chemical composition of fine particle air pollution. *Environmental Health Perspectives*, 117(6):957–963.

Peng, R. D., Dominici, F., and Louis, T. A. (2006). Model choice in time series studies of air pollution and mortality. *Journal of the Royal Statistical Society: Series A*, 169(2):179–203.

Raftery, A. E., Madigan, D., and Hoeting, J. A. (1997). Bayesian model averaging for linear regression models. *Journal of the American Statistical Association*, 92(437):179.

Rich, D. Q., Kipen, H. M., Huang, W., Wang, G., Wang, Y., Zhu, P., Ohman-Strickland, P., Hu, M., Philipp, C., Diehl, S. R., Lu, S.-E., Tong, J., Gong, J., Thomas, D., Zhu, T., and Zhang, J. J. (2012). Association Between Changes in Air Pollution Levels During the Beijing Olympics and Biomarkers of Inflammation and Thrombosis in Healthy Young Adults. *JAMA*, 307(19):2068.

Schwarz, G. (1978). Estimating the dimension of a model. *The Annals of Statistics*, 6(2):461–464.

Snowden, J. M., Reid, C. E., and Tager, I. B. (2015). Framing Air Pollution Epidemiology in Terms of Population Interventions, with Applications to Multipollutant Modeling. *Epidemiology*, 26(2):271–279.

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

Vansteelandt, S. (2012). Discussions. *Biometrics*, 68(3):175–678.

Verschuren, L., Wielinga, P. Y., van Duyvenvoorde, W., Tijani, S., Toet, K., van Ommen, B., Kooistra, T., and Kleemann, R. (2011). A dietary mixture containing fish oil, resver-

atrol, lycopene, catechins, and vitamins E and C reduces atherosclerosis in transgenic mice. *The Journal of Nutrition*, 141(5):863–869.

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

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

Wang, C., Parmigiani, G., and Dominici, F. (2012b). Rejoinder: Bayesian Effect Estimation Accounting for Adjustment Uncertainty. *Biometrics*, 68(3):680–686.

Wild, C. P. (2005). Complementing the genome with an "exposome": The outstanding challenge of environmental exposure measurement in molecular epidemiology. *Cancer Epidemiology Biomarkers and Prevention*, 14(8):1847–1850.

Wilson, A., Bondell, H. D., and Reich, B. J. (2014a). Package 'BayesPen'.

Wilson, A., Rappold, A. G., Neas, L. M., and Reich, B. J. (2014b). Modeling the effect of temperature on ozone-related mortality. *The Annals of Applied Statistics*, 8(3):1728–1749.

Wilson, A. and Reich, B. J. (2014). Confounder selection via penalized credible regions. *Biometrics*, 70(4):852–861.

Wilson, A., Reif, D. M., and Reich, B. J. (2014c). Hierarchical dose-response modeling for high-throughput toxicity screening of environmental chemicals. *Biometrics*, 70(1):237–46.

Zanobetti, A., Austin, E., Coull, B. A., Schwartz, J., and Koutrakis, P. (2014). Health effects of multi-pollutant profiles. *Environment International*, 71:13–19.

Zellner, A. (1986). On assessing prior distributions and Bayesian regression analysis with g-prior distributions. *Bayesian inference and decision techniques: Essays in Honor of Bruno De Finetti*, 6:233–243.

Zigler, C. M. and Dominici, F. (2014). Uncertainty in Propensity Score Estimation: Bayesian Methods for Variable Selection and Model-Averaged Causal Effects. *Journal of the American Statistical Association*, 109(505):95–107.