Robust Distributed Lag Models with Multiple Pollutants using Data Adaptive Shrinkage

by

Yin-Hsiu Chen

A dissertation submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
Biostatistics
in the University of Michigan
2017

Doctoral Committee:

Professor Bhramar Mukherjee, Chair
Assistant Professor Sara Dubowsky Adar
Associate Professor Veronica Berrocal
Assistant Professor Xiaoquan William Wen

Yin-Hsiu Chen

yinhsiuc@umich.edu

ORCID iD: 0000-0001-9172-946X

A C K N O W L E D G M E N T S

I would like to gratefully and sincerely thank my dissertation advisor, Bhramar Mukherjee, for her guidance, patience, and enduring encouragement. Her mentorship is paramount and I believe it is influential to not only my doctoral journey but also my entire career. I would also like to thank my committee members, Veronica Berrocal, Xiaoquan William Wen, and Sara Dubowsky Adar for their guidance and great support. I also would like to thank my wife, Hsin-Ying Lin, for her relentless support during this journey. Every individual who has helped me and inspired me, just let me say thank you. I am truly grateful.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ABSTRACT

There is growing interest in investigating the short-term delayed lag effects of environmental pollutants (e.g. air particulate matter and ozone) on a health outcome of interest measured at a certain time (e.g. daily mortality counts). Previous studies have shown that not only the current level of the exposure but exposure levels up to past few days may be associated with health event/outcome measured on current day. Distributed lag model (DLM) has been used in environmental epidemiology to characterize the lag structure of exposure effects. These models assume that the coefficients corresponding to exposures at different lags follow a given function of the lags. Under mis-specification of this function, DLM can lead to seriously biased estimates. In this dissertation, we first explore different methods to make the traditional DLM more robust. We then extend the single pollutant DLM to multi-pollutant scenarios. We illustrate the proposed methods using air pollution data from the National Morbidity, Mortality and Air Pollution Study (NMMAPS) and a dataset from Brigham and Women's Hospital (BWH) prospective birth cohort study.

In the first project, we propose three classes of shrinkage methods to combine an unconstrained DLM estimator and a constrained DLM estimator and achieve a balance between robustness and efficiency. The three classes of methods can be broadly described as (1) empirical Bayes-type shrinkage, (2) hierarchical Bayes, and (3) generalized ridge regression. A two-step double shrinkage approach that enforces the effect estimates approach zero at larger lags is also considered. A simulation study shows that all four approaches are effective in trading off between bias and variance to attain lower mean squared error with the two-step approaches having edge over others.

In the second project, we extend DLM to two-pollutant scenarios and focus on characterizing pollutant-by-pollutant interaction. We first consider to model the interaction surface by assuming the underlying basis functions are tensor products of the basis functions that generate the main-effect distributed lag functions. We also extend Tukey's one-degree-of-freedom interaction model to two-dimensional DLM context as a parsimonious way to model the interaction surface between two pollutants. Data adaptive approaches to allow departure from the specified Tukey's structure are also considered. A simulation study

shows that shrinkage approach Bayesian constrained DLM has the best average performance in terms of relative efficiency.

In the third project, we extend DLM to a truly multi-dimensional space and focus on identifying important pollutants and pairwise interactions associated with a health outcome. Penalization-based approaches that induce sparsity in solution are considered. We propose a Hierarchical integrative Group LASSO (HiGLASSO) approach to perform variable selection at a group level while maintaining strong heredity constraints. Empirically, HiGLASSO identifies the correct set of important variables more frequently than other approaches. Theoretically, we show that HiGLASSO enjoys Oracle properties including selection and estimation consistency.

CHAPTER 1

Introduction

Particulate matter (PM) and ozone (O_3) are two major pollutants affecting the air quality in the United States and throughout the world according to United States Environmental Protection Agency (USEPA). Multiple studies have shown that increased risk of adverse health outcomes are associated with exposure to air pollutants [Bell et al., 2004a, Ezzati et al., 2002]. Air pollution has both short-term and long-term effects. Short-term effects include asthma, respiratory infections, and emphysema [Le Tertre et al., 2002, Spix et al., 1998, Katsouyanni et al., 1995, Touloumi et al., 1994]. Long-term effects include chronic obstructive pulmonary disease, heart disease, and impaired brain function [Berglund and Abbey, 1996, Brunekreef and Holgate, 2002, Miller et al., 2007]. Particularly, the association between acute exposure to air particulate matter and mortality or morbidity has been extensively studied. Many such studies have indicated that the short-term lagged effects of air particulate matter are likely to be present [Samet et al., 2000, Schwartz, 2000, Braga et al., 2001, Zanobetti et al., 2003]. In other words, the current measure of health outcome such as mortality count on a given day is not only affected by the current/same-day measure of the exposure but also its lagged measures within a time window preceding the health event. It is crucial to account for such lagged/delayed effects in statistical modeling. Estimating the coefficients of lagged effects collectively in a traditional regression setting is difficult because the exposure values are serially correlated. Distributed lag model (DLM) is a common solution employed in environmental epidemiology for estimating short-term effects of environmental exposures. This dissertation considers the framework of DLM and employs shrinkage methods to make DLM more robust and efficient.

DLM was originally proposed by Almon [1965] in the econometrics literature. The

fundamental assumption underlying DLM is that the regression coefficients follow a certain structure that implies constraint(s) on the coefficients as a function of the lags. DLM can be viewed as a special type of varying coefficient models [Hastie and Tibshirani, 1993] and they serve as a general solution to circumvent the collinearity problem in serially measured exposure data. At the same time, the effect coefficients can be estimated with greater precision due to reduction in number of parameters. Common constraints include a polynomial and a natural cubic spline [Hastie and Tibshirani, 1993]. Recently, some variations of DLM have been developed to capture the distributed lag function more flexibly. Generalized additive distributed lag models [Zanobetti et al., 2000] flexibly quantify the distributed lag function using splines [Hastie and Tibshirani, 1990]. Distributed lag nonlinear models [Gasparrini et al., 2010] were developed to simultaneously capture the nonlinear exposure-response association and nonlinear distributed lag function. Bayesian DLM [Welty et al., 2009] was proposed to constrain the shape of the distributed lag function through the structural specification of the prior covariance matrix. While a myriad of DLM extensions have been proposed, they largely focus on associating the increased risk of adverse health outcomes with exposure to a single air pollutant. Since we are simultaneously exposed to a complex mixture of multiple chemicals/exposures, it is important to develop statistical methods that extend DLM to the scenario with two or more pollutants.

Air pollution policies worldwide are typically based on the scientific evidence regarding the health impact of each pollutant separately [Dominici et al., 2010]. For example, as of today, National Ambient Air Quality Standards (NAAQS) are still established on the basis of six individual criteria pollutants - carbon monoxide (CO), lead (Pb), nitrogen dioxide (NO_2), ozone (O_3), particle matter (PM), and sulfur dioxide (SO_2). Regulators are aware that the ambient levels of the criteria pollutants are related and the harm to human health from each pollutant potentially depends on the levels of other pollutants (i.e. possible existence of interactions between pollutants). However, the ability to design and implement multi-pollutant policies is limited due to the scarcity of scientific results on how air pollution mixtures jointly affect human health. If the synergistic effects associated with simultaneous exposure to multiple pollutants can be estimated in a more reliable fashion, the air pollution standards can be established based upon combined levels of multiple pol-

lutants, such as the Air Quality Index (AQI) given by USEPA. There is a growing need for statistical approaches addressing health risks due to multiple pollutants and their potential interactions in a chemical mixture.

Some of the past multi-pollutant studies report the health effect of one pollutant adjusted for the exposure to other pollutants [Bell et al., 2007, Rojas-Martinez et al., 2007]. Nonetheless, additive models are most likely inadequate. Some of the major elements of associating multiple pollutants to a health outcome in a time-series design are: (1) to identify the key pollutants that are strongly associated with the outcome of interest; (2) to consider potential interaction effects between pollutants; (3) to handle the collinearity between multiple pollutants; and (4) to account for serially correlated exposure measurements. Multiple regression models with main effects for each pollutant and interaction terms for each pair of pollutants as predictors have already been suggested [Mauderly et al., 2010]. However, it is well-known that the statistical power for detecting an interaction effect is low and effect estimation becomes highly unstable when two or more pollutants are highly correlated [Farrar and Glauber, 1967]. Tree-based regression approaches such as classification and regression tree (CART) are useful to account for higher-order and nonlinear interactions [Hu et al., 2008]. The interpretation of effect estimates is difficult in these types of model-free, data driven approaches. Deletion/substitution/addition (DSA) algorithm [Sinisi and van der Laan, 2004] allows users to specify the constraints on polynomial function form of exposure and order of interaction. Nevertheless, statistical inference is again not reliable when predictors are highly correlated [Dominici et al., 2008]. Some modern variable selection and dimension reduction techniques can be useful in multi-pollutant settings. Penalized regression methods such as least absolute shrinkage and selection operator (LASSO) [Tibshirani, 1996] can be employed to identify a small subset of individual predictors that are highly associated with the outcome. Principal component analysis (PCA) have been used to investigate the synergistic effects of multiple pollutants in several studies [Burnett et al., 2001, Qian et al., 2004, Arif and Shah, 2007]. However, none of the existing multi-pollutant approaches tries to capture the lagged effects and their possible interactions over a biologically meaningful time period.

In a time-series setting, jointly considering the temporal dynamics of the current

and past exposure in association with the outcome is crucial. Few attempts have been made to incorporate the prior knowledge about the distributed lag function with two or more pollutants. Roberts [Roberts, 2004] investigated the interaction between daily particulate air pollution and daily mean temperature in Cook County, Illinois by stratifying the lagged effect of particulate air pollution on mortality by temperature. High degree DLMs extend basic DLMs to incorporate higher-order interactions between lagged predictors [Heaton and Peng, 2014] but the extension is still restricted to single pollutants. The bivariate DLM [Muggeo, 2007] is by far the only attempt to extend one-pollutant DLM to two-pollutant scenarios. In [Muggeo, 2007], bivariate DLM is used to model the joint effect of temperature and air particular matter with aerodynamic diameter less than 10 micrometers (PM_{10}) main effect in the same way as parametric DLM with two separate sets of basis functions. Tensor products of the two were employed to characterize the DL surface for temperature-PM_{10} interaction.

In this dissertation proposal, we propose to extend and modify single pollutant DLM to the situation with two or more pollutants. In Chapter II, we first provide an overview of DLMs and their variations in modeling the association between a time-series measured health outcome and a single time-series measured air pollutant. We then introduce three robust DLMs that shrink an unconstrained DLM estimator toward a model-dependent constrained DLM estimator using data-adaptive shrinkage. The three approaches are empirical Bayes (EB), hierarchical Bayes (HB), and generalized ridge regression (GRR). In Chapter III, we introduce a variance component score test (VCST) to test a given constrained DLM against an unconstrained alternative. The test is motivated by GRR and serves as a more powerful alternative to standard likelihood ratio test. In Chapter IV, we extend DLM to two-pollutant scenarios and examine different strategies to model pairwise interactions with consideration of lag structure. We propose an unified approach that combines DLM and Tukey's one degree of freedom model [Tukey, 1949]. The corresponding shrinkage estimator [Ko et al., 2014] that is robust to misspecification of interaction structure is also considered. In addition, we propose a Bayesian constrained DLM (BCDLM) approach to characterize the joint effect of two pollutants and their interactions. In Chapter V, we focus on penalization-based approaches to assess the joint effects of multiple pollutants

on a health outcome. We propose a new algorithm called Hierarchical integrative Group LASSO (HiGLASSO) to perform variable selection at group level while maintaining the strong heredity constraints (inclusion of interaction only in presence of main effects). The approach is quite general and can potentially be applied to a wide spectrum of problems when a sparse solution is desired to identify interactions among a correlated or grouped set of predictors. Across all chapters, we compare our method with various existing alternatives via extensive simulation studies. In addition, we illustrate the proposed methods by using the data from the National Morbidity, Mortality, and Air Pollution Study (NMMAPS). In Chapter V, we also illustrate HiGLASSO using a data set from Brigham and Women's Hospital (BWH) prospective pregnancy/birth cohort study that collects biological samples and detailed clinical data to identify effects of mixtures.

Overall, the dissertation provides a statistical framework for the assessment of health effects of multiple environmental exposures, with the scientific goal of understanding the interplay between a complex mixture of air pollutants after incorporating potential lagged effects. These methods can potentially be useful in areas outside environmental epidemiology. We hope that our work will lead to further research in other applications that involve characterizing the joint effect of a set of correlated predictors and their interactions.

CHAPTER 2

Robust Distributed Lag Models using Data Adaptive Shrinkage

2.1 Introduction

In environmental epidemiology, investigators are often interested in estimating the effects of air pollution levels on counts of some health events (e.g. mortality and cardiovascular events). Sometimes the effects are not limited to the concurrent time periods but delayed in time. A number of early studies suggest that multi-day average pollution levels are more predictive of health event counts than a single-day pollution measure [Schwartz and Dockery, 1992, Schwartz, 1994]. More recent time series studies found that models with just single-day pollution measures might underestimate the occurrence of health events associated with air pollution [Schwartz, 2000, Roberts, 2005]. Modeling each single lagged effect in separate models is not desirable and it is difficult to synthesize the results across different models. The most straightforward approach to jointly consider the temporal dynamics is to use a generalized linear model (GLM) with current health event count as the outcome and with current and past air pollution levels as covariates in the same regression model. However, this simple but naive modeling entails two problems. First, a large number of parameters needs to be estimated, resulting in loss of power due to large degrees of freedom (df), especially when the sample size is small and the maximal number of lags (L) is large. Second, the serial autocorrelation between lagged pollution levels is often high. Thus, the lagged effect estimates, though consistent for the true effects in large samples, could have inflated variance, and the sign of the effect estimates could be reversed

in small samples [Farrar and Glauber, 1967].

Polynomial DLMs [Almon, 1965], originally proposed in econometrics, assume that the unknown lag coefficients lie on a polynomial function of the lag with known degree. More generally, a constrained DLM imposes a pre-specified structure to constrain the lag coefficients as a function of the lags. They serve as a general solution to circumvent the collinearity problem and estimate effect coefficients with greater precision. Beyond polynomial constraints, several other functional forms [Corradi, 1977, Hastie and Tibshirani, 1993] have been used. The choice of the DL function often relies on prior knowledge about the effects of exposure on health events. Thus, a linear DL function may be appropriate for uniformly decreasing lagged effects and a quadratic DL function may be appropriate for short delays in health effects after exposure. Such explicit prior knowledge may not be available in many studies. Even with some degree of knowledge about the shape of the DL functions, the parsimonious structure may omit some detailed characteristics of the lag course, but lead to increased precision due to the reduced number of parameters to be estimated [Zanobetti et al., 2000, 2002]. In addition, in examining multiple exposure-disease pairs, it is difficult to assess each DL function in detail on a case-by-case basis.

As a potential solution, one could expand and enrich the class of DL functions, but that would defeat the purpose of reducing the number of parameters to be estimated. Recently, some variations of constrained DLMs have been proposed to capture the DL function more flexibly. Generalized additive distributed lag models (GADLM) [Zanobetti et al., 2000] use splines to represent the DL function. Muggeo [2008] proposed a flexible segmented break point model with doubly penalized B-splines. Distributed lag nonlinear models (DLNMs) [Gasparrini et al., 2010] were developed to simultaneously model the nonlinear exposure-response dependencies and nonlinear DL function. Bayesian DLM (BDLM) [Welty et al., 2009] has been proposed to incorporate prior knowledge about the shape of the DL function through specification of the prior covariance matrix. BDLM has been extended to Bayesian hierarchical DLM by adding another layer of hierarchy in order to account for regional heterogeneity [Peng et al., 2009]. Obermeier et al. [2015] introduced a flexible DLM where the lag effects are smoothed via a difference penalty and the last lag coefficient is shrunk towards 0 via a ridge penalty.

Using smoothing techniques to flexibly model the distributed lag function has been extensively discussed in the literature. In the single stage shrinkage methods we propose an alternative framework to achieve the desired bias-variance tradeoff. Our proposed method shrinks the unconstrained DLM estimator toward a model-dependent constrained DLM estimator in a data-adaptive way. The underlying objective is to retain the flexibility of unconstrained DLM for enhanced robustness and retain precision advantages by shrinking toward a parsimonious constrained DLM. The resulting shrinkage estimators are robust to misspecification of the working distributed lag function. The first approach is to perform component-wise shrinkage by combining the two estimators using an EB type of weighting [Mukherjee and Chatterjee, 2008, Chen et al., 2009]. The second approach is a new HB approach. The third approach is GRR. The idea is the same as traditional ridge regression except that the unconstrained DLM estimators are shrunk toward the constrained DLM estimator rather than shrinkage towards the null. The amount of shrinkage is controlled by a tuning parameter chosen via a criterion such as corrected Akaike information criterion (AICC) [Hurvich et al., 1998] and generalized cross-validation (GCV) [Golub et al., 1979]. The three shrinkage methods provide a general framework to shrink one estimator toward its constrained counterpart in a data-adaptive manner. We also consider a two-stage shrinkage approach where a hyperprior is introduced to penalize the estimates obtained from any of the shrinkage approaches to ensure that the estimated DL function smoothly goes to zero at larger lags, akin to BDLM. The two-stage methods allow misspecification of the maximal number of lags L, thus ensuring robustness with respect to the choice of L, another user-defined tuning parameter in constructing a standard DLM.

In addition to introducing different shrinkage approaches to robustly model the distributed lag function, a major contribution of the chapter is to establish the correspondence between a transformation matrix used in DLM with a constraint matrix that helps to define the nonnull shrinkage targets driving the specification of corresponding priors and penalties. In Section 2.2, we first give an overview of DLM and their variations, including Almon polynomial DLM [Welty et al., 2009], GADLM [Zanobetti et al., 2000], DLNM [Gasparrini et al., 2010], and BDLM [Welty et al., 2009]. The definitions of the transformation matrix and constraint matrix and the details of the correspondence along with our shrink-

age approaches will be introduced in Section 2.3. In Section 2.4, we conduct an extensive simulation study to compare the proposed approaches to existing alternatives. In section 2.5, we illustrate our methods by analyzing data from NMMAPS to explore association between a set of ambient pollutants and counts of overall mortality, cardiovascular mortality, and deaths due to respiratory events in Chicago, Illinois, from 1987 to 2000. Section 2.6 contains concluding remarks.

2.2 Distributed Lag Models (DLM)

DLMs are used to model the current values of a dependent variable based on both the current values of an explanatory variable and the lagged values of this explanatory variable for time series data. We use the following notation throughout the chapter. Let x_t denote the exposure measured at time t, such as ambient air pollution level, y_t denote the response measured at time t, such as daily mortality count, and $\boldsymbol{z}_t$ denote the covariates at time t, such as temperature and humidity. Let T be the length of the time series. We consider the GLM $g[E(y_t|x_t, x_{t-1}, ..., x_{t-L}, \boldsymbol{z}_t)] = \alpha_0 + \boldsymbol{z}_t^\top \boldsymbol{\alpha}_1 + \sum_{\ell=0}^{L} \beta_\ell x_{t-\ell}$ where α_0 is the intercept, $\boldsymbol{\alpha}_1$ represents the effect of covariates, L is the pre-determined maximum number of lags, and $\boldsymbol{\beta} = (\beta_0, \beta_1, ..., \beta_L)^\top$ is the vector of lagged effects. We consider the log-linear Poisson model as follows.

$$y_t \sim \text{Poisson}(\mu_t)$$

$$\log \mu_t = \alpha_0 + \boldsymbol{z}_t^\top \boldsymbol{\alpha}_1 + \sum_{\ell=0}^{L} \beta_\ell x_{t-\ell}$$

The goal is to estimate the lag effect coefficients. For simplicity and without loss of generality, we leave out intercept and covariates in later presentation. Distributed lag function describes the relationship between the coefficient of the lagged exposure (i.e. $\boldsymbol{\beta} = (\beta_0, \beta_1, ..., \beta_L)^\top$) and the lag (i.e. $\ell = 0, ..., L$). Different DLM impose different constraints on the temporally dynamic relationship between $\boldsymbol{\beta}$ and ℓ. One common difficulty to all DLM is the choice of maximum lag. We assume that the lag effect diminishes to zero after a certain lag and the choice of L is large enough to cover all the lags with nonzero effects. Only finite distributed lag models are considered.

2.2.1 Unconstrained Distributed Lag Models

Unconstrained distributed lag models impose no constraints on the shape of the distributed lag function. Any pattern of the $L+1$ lag coefficients can be estimated. The most direct approach to estimate the coefficients is through unconstrained maximum likelihood estimation (MLE). Let $\boldsymbol{X}_t = (x_t, x_{t-1}, ..., x_{t-L})^\top$ and let $\ell(\boldsymbol{\beta})$ denote the likelihood function. The unconstrained GLM estimator $\hat{\boldsymbol{\beta}}$ can be written as

$$\hat{\boldsymbol{\beta}}_{UDLM} = \arg\max_{\boldsymbol{\beta}} \ell(\boldsymbol{\beta}) = \arg\max_{\boldsymbol{\beta}} \sum_{t=1}^{\top} [y_t \boldsymbol{\beta}^\top \boldsymbol{X}_t - e^{\boldsymbol{\beta}^\top \boldsymbol{X}_t} - \log(y_t!)].$$

The estimation is simple and the interpretation is straightforward. However, unconstrained distributed lag models entail two problems. The first one is that the serially measured exposures can be highly collinear. The multi-collinearity would lead to unreliable coefficient estimates with inflated variance. The sequence of estimated lag coefficients might bounce around and the signs could be switched in small samples [Farrar and Glauber, 1967]. The second problem is that a large number of parameters is to be estimated. The df can be depleted quickly and result in loss of power, especially when the sample size is small and the maximal number of lags (L) is large. The problem would be magnified if two or more correlated pollutants are included in a regression model. Constrained DLM serve as a remedy to the two problems.

2.2.2 Almon Polynomial Distributed Lag Models

Polynomial distributed lag models were first explored by Almon [Almon, 1965]. They impose smoothness on the coefficients by restricting the lag coefficients to lie on a polynomial function. If it is assumed that the lag coefficients lie on a polynomial of degree d ($d < L+1$),

$$\beta_\ell = \sum_{j=0}^{d} \theta_j \ell^j \tag{2.1}$$

where $\boldsymbol{\theta} = (\theta_0, \theta_1, ..., \theta_d)^\top$ are the $d+1$ free parameters to be estimated in the lower-dimensional space. The construction reduce the number of parameters to be estimated from

$L+1$ to $d+1$. Through matrix representation, Equation 2.1 can be rewritten as

$$\boldsymbol{\beta} = \boldsymbol{C}\boldsymbol{\theta}$$

where $\boldsymbol{C}$ is a $(L+1) \times (d+1)$ matrix such that the (i,j) element is $(i-1)^{j-1}$, that is

$$\boldsymbol{C} = \begin{bmatrix} 0^0 & 0^1 & \cdots & 0^d \\ 1^0 & 1^1 & \cdots & 1^d \\ \vdots & \vdots & \vdots & \vdots \\ L^0 & L^1 & \cdots & L^d \end{bmatrix}_{(L+1)\times(d+1)} .$$

We can define

$$\boldsymbol{Q}_t = \boldsymbol{C}^\top \boldsymbol{X}_t.$$

Now $\boldsymbol{Q}_t$ is a $(d+1) \times 1$ vector representing the transformed independent variables to be regressed on corresponding to parameters $\boldsymbol{\theta}_{(d+1)\times 1}$ in the constrained space. The constrained log-likelihood function for Almon polynomial DLM estimator can be written as

$$\ell(\boldsymbol{\theta}) = \sum_{t=1}^{T} [y_t \boldsymbol{\theta}^\top \boldsymbol{Q}_t - e^{\boldsymbol{\theta}^\top \boldsymbol{Q_t}} - \log(y_t!)]$$

If we let $\hat{\boldsymbol{\theta}} = \arg\max_{\boldsymbol{\theta}} \ell(\boldsymbol{\theta})$, the implied estimated lagged effects from Almon polynomial DLM can be expressed as

$$\hat{\boldsymbol{\beta}}_{CDLM} = \boldsymbol{C}\hat{\boldsymbol{\theta}}$$

and the variance estimates are

$$V(\hat{\boldsymbol{\beta}}) = \boldsymbol{C} V(\hat{\boldsymbol{\theta}}) \boldsymbol{C}^\top.$$

The restrictions implied by a polynomial distributed lag model can always be tested against the higher-degree polynomial distributed lag model or unrestricted polynomial lag model. The lower the order of the polynomial, the smoother the lag distribution is. In other words, lower order of polynomial distributed lag model assumes that the effects of adjacent lag

coefficients are more similar.

Note that the construction of $\boldsymbol{C}$ is not unique. The above construction corresponds to choosing 1, ℓ, ..., ℓ^d as the $d+1$ basis functions that generate the class of functions that $d-$degree polynomial distributed lag function can lie in. Alternatively, any $(d+1)$ basis functions that are obtained via a full-rank linear transformation from the above $d+1$ basis functions would lead to identical lag effect estimates. The full generalization of constructing $\boldsymbol{C}$ will be detailed later.

2.2.3 Generalized Additive Distributed Lag Models

Zanobetti et al. [2000] proposed GADLMs. The approach is motivated by the effect of "mortality displacement" in environmental epidemiology. Mortality displacement [Schimmel and Murawski, 1976] is the occurrence that high air pollution levels advance the deaths of frail individuals by several days. The effect of particular matter on mortality may take effect with some retard. The distributed lag function may be zero or positive at early lags and then decrease and become negative (i.e. rebound effect) at larger lags [Zanobetti et al., 2000, 2002]. Polynomial DLM may not be sufficient to capture this more "localized" structure. Generalize additive DLM are more flexible to model the lag effect of the exposure of interest. As the name suggests, they combine generalized additive models (GAM) [Hastie and Tibshirani, 1990] and DLM.

Zanobetti proposed to model the distributed lag function as a regression spline function of ℓ as follows:

$$\beta_\ell = \sum_{j=0}^{d} \theta_j \ell^j + \sum_{k=1}^{K} \theta_{\kappa k} (\ell - \kappa_k)_+^d$$

where $\kappa_1, ..., \kappa_K$ is a set of K knot positions between 0 and L. Note that β_ℓ becomes a piecewise dth degree polynomial in ℓ with K internal knots connecting the pieces. Now there are $K+d+1$ parameters to estimate (i.e. $\boldsymbol{\theta} = (\theta_0, ..., \theta_d, \theta_{\kappa 1}, ..., \theta_{\kappa K})^\top$). If we expand the basis matrix $\boldsymbol{C}$ as

$$\boldsymbol{C} = \begin{bmatrix} 0^0 & 0^1 & \cdots & 0^d & (0-\kappa_1)_+^p & \cdots & (0-\kappa_K)_+^p \\ 1^0 & 1^1 & \cdots & 1^d & (1-\kappa_1)_+^p & \cdots & (1-\kappa_K)_+^p \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ L^0 & L^1 & \cdots & L^d & (L-\kappa_1)_+^p & \cdots & (L-\kappa_K)_+^p \end{bmatrix}_{(L+1)\times(K+d+1)},$$

the estimation and inference can be conveniently followed as in Section 2.2.2.

In general, if we assume that $B_1(\cdot)$, $B_2(\cdot)$, ..., $B_p(\cdot)$ are p known basis functions that generate the class of distributed lag function that $\boldsymbol{\beta}$ can lie in, the corresponding transformation matrix $\boldsymbol{C}$ is a $(L+1) \times p$ matrix where the (i, j) element is $B_j(i-1)$. Again, the DLM solution is invariant up to a full-rank linear transformation of $\boldsymbol{C}$.

2.2.4 Distributed Lag Nonlinear Models (DLNM)

Zanobetti et al. [2000] developed a unified framework that model the exposure-response dependencies and the lag effects simultaneously with an additional lag dimension. Inherently, a 3-dimensional space with exposure, lag, and response as the three axes is considered. Both exposure-response relationship and the distributed lag function can be flexibly model in a nonparametric fashion. A cross-basis, defined as a bi-dimensional space of functions, is used to describe the shape of the relationship along the exposure and its lag effects.

Consider Poisson regression

$$y_t \sim \text{Poisson}(\mu_t)$$

$$\log \mu_t = f(\boldsymbol{X}_t; \boldsymbol{\beta})$$

with an unknown smooth function f. Suppose v_x basis functions that span the space of functions of which we believe that f lies in are chosen. Let $b_{t\cdot}$ be a $1 \times v_x$ row vector obtained by applying v_x basis functions to x_t. Similarly, v_ℓ basis functions are chosen for distributed lag function. Define a $T \times v_x \times (L+1)$ array $\dot{\boldsymbol{B}}$ representing the lagged occur-

rences of each of the basis variables of exposure. The (i, j, k) element of $\dot{\boldsymbol{B}}$ is $b_{i-(k-1),j}$, the $(k-1)$ lagged exposure measure at time i evaluated at the jth basis function. DLNM can be specified by

$$f(\boldsymbol{X}_t; \boldsymbol{\beta}) = \sum_{j=1}^{v_x} \sum_{k=1}^{v_\ell} b_{tj\cdot}^\top c_{\cdot k}^* \theta_{jk} = \boldsymbol{w}_{t\cdot}^\top \boldsymbol{\theta}$$

where $b_{tj\cdot}$ is the $(L+1) \times 1$ vector sliced in the first two dimensions at t and j, respectively, and $\boldsymbol{w}_{t\cdot}$ is obtained by applying the $v_x \cdot v_\ell$ cross-basis functions to x_t. Note that $\boldsymbol{w}_{t\cdot}$ is analogous to $\boldsymbol{Q}_t$ introduced in Section 2.2.2. Let $\boldsymbol{C}^*$ be a $(L+1) \times v_\ell$ transformation matrix obtained by applying v_ℓ basis functions to lag vector $(0, ..., L)^\top$. The cross-basis is presented as a tensor product. Define

$$\dot{\boldsymbol{A}} = (\mathbf{1}^\top \otimes \dot{\boldsymbol{B}}) \odot (\mathbf{1} \otimes P_{1,3}(\boldsymbol{C}^*) \otimes \mathbf{1}^\top)$$

where $\mathbf{1}$ is a vector of ones with proper dimensions, $\otimes$ is the Kronecker product, $\odot$ is the Hadamard product, and $P_{i,j}$ defined as the operator that permutes the index i and index j of an array. The final matrix of cross-basis functions $\boldsymbol{W}$ can be obtained by summing along the third dimension of the $T \times (v_x \cdot v_\ell) \times (L+1)$ array $\dot{\boldsymbol{A}}$.

Interpreting the results of DLNM with nonlinear dependencies is difficult. One solution is to present the response surface on a 3-dimensional plot. Alternatively, one can fix the exposure level at a suitable value and show the relationship between response and lag, or examine the exposure-response relationship at a certain lag. Given a vector of $\boldsymbol{x}^P$ with m exposure values used for prediction, the corresponding $m \times v_x \times (L+1)$ array $\dot{\boldsymbol{B}}^P$ and the final array $\dot{\boldsymbol{A}}^P$ can be derived following the above procedures. The predicted effects of the m exposure levels at lag ℓ are given as

$$\dot{\boldsymbol{A}}_{\cdot\cdot\ell}^P \hat{\boldsymbol{\theta}}$$

and the estimated variances are

$$\text{diag}[\dot{\boldsymbol{A}}_{\cdot\cdot\ell}^P \hat{\text{Var}}(\hat{\boldsymbol{\theta}}) \dot{\boldsymbol{A}}_{\cdot\cdot\ell}^{P\top}].$$

Note that the formulation in the paper is different from the generalized framework that we described at the end of Section 2.2.3. If we define $\boldsymbol{B}_t$ as a $(L+1) \times v_x$ matrix such that the ith row is $b^{\top}_{(t-i+1),.}$, we can express $\boldsymbol{w}_{t.}$ as

$$\boldsymbol{w}_{t.} = (\boldsymbol{I} \otimes \boldsymbol{C}^{*\top})\text{vec}(\boldsymbol{B}_t)$$

where vec$(\cdot)$ is the vectorization function and $\boldsymbol{I}$ is a $v_x \times v_x$ identity matrix. Now the transformation matrix $\boldsymbol{C}$ is simply $(\boldsymbol{I} \otimes \boldsymbol{C}^{*\top})$, the set of predictors in the original parameter space at time t is vec$(\boldsymbol{B}_t)$, and the set of predictors in the transformed parameter space at time t is $\boldsymbol{w}_{t.}$.

2.2.5 Bayesian Distributed Lag Models (BDLM)

The DLM introduced so far characterize the distributed lag function using a particular parametric form or a constant degree of smoothness. Welty et al. [2009] proposed to incorporate the prior knowledge about the shape of the distributed lag function through a structural specification of the prior covariance matrix. One advantage of this approach is that the degree of smoothness of the distributed lag function can be estimated from the data. The formulation of BDLM is relevant when the lagged effects of an exposure are unknown at the first few lags and they taper off with increased lag. In other words, coefficients at earlier lags are less unconstrained. The full hierarchical specifications are:

$$\boldsymbol{Y}|\boldsymbol{\beta} \sim \text{Poisson}(e^{\boldsymbol{X}\boldsymbol{\beta}})$$

$$\boldsymbol{\beta}|\boldsymbol{\omega}, \sigma^2 \sim \mathcal{N}(0, \sigma^2\boldsymbol{\Omega}(\boldsymbol{\omega}))$$

$$\boldsymbol{\Omega}(\boldsymbol{\omega}) = \boldsymbol{V}(\omega_1)\boldsymbol{W}(\omega_2)\boldsymbol{V}(\omega_1)$$

$$\sigma^2 = 10 \cdot \text{Var}(\hat{\beta}_0)$$

where $\boldsymbol{V}(\omega) = \text{diag}[1, \exp(\omega), \exp(2\omega), ..., \exp(L\omega)]$, $\boldsymbol{W}(\omega_2)$ is the correlation matrix derived from the covariance matrix $\boldsymbol{V}(\omega_2)\boldsymbol{V}(\omega_2)^{\top} + \{\boldsymbol{I}_{L+1} - \boldsymbol{V}(\omega_2)\}\mathbf{1}_{L+1}\mathbf{1}^{\top}_{L+1}\{\boldsymbol{I}_{L+1} - \boldsymbol{V}(\omega_2)\}^{\top}$, $\boldsymbol{I}_{L+1}$ is a $(L+1) \times (L+1)$ identity matrix, $\mathbf{1}_{L+1}$ is a $(L+1) \times 1$ vector of ones,

and $\hat{\beta}_0$ is the estimated coefficient for lag 0 from unconstrained GLM. $\boldsymbol{\Omega}$ is constructed in this way so that (i) the coefficients smoothly approach zero with increasing lag and (ii) coefficients at smaller lags are less constrained (larger variance). The posterior distribution can be computed through Gibbs sampling or other Markov chain Monte Carlo methods [Carlin and Louis, 1997].

2.3 Robust Distributed Lag Models

We first consider the log-linear Poisson model:

$$y_t \sim \text{Poisson}(\mu_t)$$
$$\log \mu_t = \alpha_0 + \boldsymbol{z}_t^\top \boldsymbol{\alpha}_1 + \sum_{\ell=0}^{L} \beta_\ell x_{t-\ell}$$

The goal is to estimate the lagged effect coefficients $\{\beta_\ell\}$. For simplicity and without loss of generality, we leave out intercept and covariates in subsequent presentation. A straightforward approach to estimate the coefficients is through unconstrained MLE. Let $\boldsymbol{X_t} = (x_t, x_{t-1}, ..., x_{t-L})^\top$. The unconstrained DLM estimator $\hat{\boldsymbol{\beta}}_{UDLM}$ can be written as

$$\hat{\boldsymbol{\beta}}_{UDLM} = \arg\max_{\boldsymbol{\beta}} \ell_u(\boldsymbol{\beta}) = \arg\max_{\boldsymbol{\beta}} \sum_{t=1}^{T} [y_t \boldsymbol{\beta}^\top \boldsymbol{X_t} - e^{\boldsymbol{\beta}^\top \boldsymbol{X_t}} - \log(y_t!)]. \tag{2.2}$$

Constrained DLM imposes structure on $\boldsymbol{\beta}$ by assuming β_ℓ is a known function of ℓ for $\ell = 0, \cdots, L$. We assume that $B_1(\cdot)$, $\cdots$, $B_p(\cdot)$ are the p basis functions that generate the class of functions in which $\boldsymbol{\beta}$ lies. A transformation matrix $\boldsymbol{C}$ [Gasparrini et al., 2010] is defined as a $(L+1) \times p$ matrix where the element $(\ell+1, j)$ is the j^{th} basis function $B_j(\cdot)$ measured at ℓ (i.e. $B_j(\ell)$). For instance, a $p-1$ degree polynomial DLM requires the specification of p basis functions. If a linear constraint is implemented, one possible choice of basis functions is $B_1(\ell) = 1$ and $B_2(\ell) = \ell$ and the corresponding $\boldsymbol{C}$ becomes a $(L+1)\times 2$ matrix with all 1's in the first column and $0, 1, ..., L$ in the second column. We can define $\boldsymbol{W}_t = \boldsymbol{C}^\top \boldsymbol{X}_t$ where $\boldsymbol{W}_t$ is a $p \times 1$ vector representing the transformed independent variables in the model, with corresponding coefficients $\boldsymbol{\theta}_{p\times 1}$ in a lower-dimensional space

to be regressed on. The constrained DLM estimator is $\hat{\boldsymbol{\beta}}_{CDLM} = \boldsymbol{C}\hat{\boldsymbol{\theta}}$ where

$$\hat{\boldsymbol{\theta}} = \arg\max_{\boldsymbol{\theta}} \ell_c(\boldsymbol{\theta}) = \arg\max_{\boldsymbol{\theta}} \sum_{t=1}^{T} [y_t \boldsymbol{\theta}^\top \boldsymbol{W_t} - e^{\boldsymbol{\theta}^\top \boldsymbol{W_t}} - \log(y_t!)]. \tag{2.3}$$

and the variance of $\hat{\boldsymbol{\beta}}_{CDLM}$ is given by $V(\hat{\boldsymbol{\beta}}_{CDLM}) = \boldsymbol{C}V(\hat{\boldsymbol{\theta}})\boldsymbol{C}^\top$.

Note that the choice of basis functions for constructing $\boldsymbol{C}$ is unique only up to a full-rank linear transformation. In section 2.3.2 to section 2.3.4, we will introduce different approaches to shrink $\hat{\boldsymbol{\beta}}_{UDLM}$ toward $\hat{\boldsymbol{\beta}}_{CDLM}$ in a data-adaptive manner. All the methods introduced in this section are summarized in Table 2.3.

2.3.1 Connection Between the Transformation Matrix C and the Constraint Matrix R

We establish the connection between a given transformation matrix $\boldsymbol{C}$ and its corresponding constraint matrix $\boldsymbol{R}$ (as introduced below) that helps us generalize the proposed methods to a wider class of DLMs. The notion of the constraint matrix $\boldsymbol{R}$ originates from the "smoothness prior" introduced by Shiller [1973].

Consider a $(L+1) \times p$ transformation matrix $\boldsymbol{C}$. Specifying p basis functions underlying a DL function results in p unconstrained parameters $\boldsymbol{\theta}$ to be estimated as in (2.3). Equivalently, it can be formulated as $L+1$ parameters in $\boldsymbol{\beta}$ to be estimated with $L+1-p$ constraints on $\boldsymbol{\beta}$, obtained by maximizing (2.2) subject to the constraints. The constraints can be represented by $\boldsymbol{R\beta} = \boldsymbol{0}$ where $\boldsymbol{R}$ is the $(L+1-p) \times (L+1)$ constraint matrix. The basis functions in $\boldsymbol{C}$ span the solution space of $\boldsymbol{R\beta} = \boldsymbol{0}$, thus $\boldsymbol{C}$ and $\boldsymbol{R}$ have a direct correspondence. Define $\boldsymbol{C}_e$ as a $(L+1) \times (L+1)$ matrix $[\boldsymbol{C}\ \boldsymbol{0}_{(L+1)\times(L+1-p)}]$ where $\boldsymbol{0}_{(L+1)\times(L+1-p)}$ is a $(L+1) \times (L+1-p)$ matrix with zero entries. Applying singular value decomposition (SVD) $\boldsymbol{C}_e^\top = \boldsymbol{U}_C \boldsymbol{D}_C \boldsymbol{V}_C^\top$ where $\boldsymbol{U}_C$ is the $(L+1) \times (L+1)$ unitary matrix with left-singular column vectors, $\boldsymbol{V}_C$ is the $(L+1) \times (L+1)$ unitary matrix with right-singular column vector, and $\boldsymbol{D}_C$ is a $(L+1) \times (L+1)$ diagonal matrix with singular values of $\boldsymbol{C}_e^\top$ along the diagonal, the $(L+1-p) \times (L+1)$ constraint matrix $\boldsymbol{R}$ can be obtained as the last $(L+1-p)$ rows of $\boldsymbol{V}_C^\top$. More detailed description of the connection

between $\boldsymbol{R}$ and $\boldsymbol{C}$ is provided in the Appendix 2.7.1. We summarize two important results that are going to be used in the subsequent development.

Result 1: $\hat{\boldsymbol{\beta}}_{CDLM} = \boldsymbol{C}\hat{\boldsymbol{\theta}}$, where $\hat{\boldsymbol{\theta}}$ is as given in (2.3), is equivalent to the maximizer of the likelihood function in (2.2) subject to the constraint $\boldsymbol{R}\boldsymbol{\beta} = \mathbf{0}$, where $\boldsymbol{R}$ is as defined above.

Result 2: The lag coefficients of polynomial DLMs, spline-based DLMs with known knot locations, or using any other basis functions can all be represented by $\boldsymbol{\beta} = \boldsymbol{C}\boldsymbol{\theta}$ where $\boldsymbol{C}$ is a suitably defined $(L+1) \times p$ transformation matrix and $\boldsymbol{\theta}$ is a vector of unconstrained parameters in $\mathbb{R}^p$. Therefore, the constrained DLM solutions can alternatively be defined as an element belonging to the null space of the corresponding constraint matrix $\boldsymbol{R}$.

Remark 1: Throughout we use polynomial DLM as our shrinkage target in this chapter but Results 1 and 2 suggest that the methods are generalizable to other more flexible DLMs.

2.3.2 Empirical Bayes-Type Shrinkage Estimator

The simplest way to combine two estimators is taking the weighted average of the two with some reasonable data-adaptive choices for the weights. Mukherjee and Chatterjee [2008] and Chen et al. [2009] proposed an Empirical Bayes type estimator to shrink a model-free estimator toward a model-based estimator. For our context, we consider the following EB-type estimator

$$\hat{\boldsymbol{\beta}}_{EB} = \hat{\boldsymbol{\beta}}_{UDLM} + \boldsymbol{K}(\hat{\boldsymbol{\beta}}_{CDLM} - \hat{\boldsymbol{\beta}}_{UDLM}) \tag{2.4}$$

with $\boldsymbol{K} = (\hat{\boldsymbol{V}} \circ \boldsymbol{I}_{L+1})[(\hat{\boldsymbol{V}} + \hat{\boldsymbol{\psi}}\hat{\boldsymbol{\psi}}^\top) \circ \boldsymbol{I}_{L+1}]^{-1}$. $\hat{\boldsymbol{V}}$ is the estimated variance-covariance matrix of $\hat{\boldsymbol{\beta}}_{UDLM}$, $\hat{\boldsymbol{\psi}} = \hat{\boldsymbol{\beta}}_{CDLM} - \hat{\boldsymbol{\beta}}_{UDLM}$, $\boldsymbol{I}_{L+1}$ is a $(L+1) \times (L+1)$ identity matrix, and $\circ$ is the Hadamard product. The shrinkage factor can be represented by $\boldsymbol{K} = \text{diag}[k_1, ..., k_{L+1}]$ with $k_i = v_i/(v_i + \hat{\psi}_i^2)$ where $\hat{\psi}_i^2$ is the i^{th} diagonal component of $\hat{\boldsymbol{\psi}}\hat{\boldsymbol{\psi}}^\top$, and v_i is the i^{th} diagonal element of $\hat{\boldsymbol{V}}$ for $i = 1, \cdots, L+1$. An alternative choice for defining the weights is to consider the estimated variance-covariance matrix of $\hat{\boldsymbol{\psi}}$ instead of $\hat{\boldsymbol{\psi}}\hat{\boldsymbol{\psi}}^\top$ in (2.4). The expression and derivation of the variance-covariance estimate of $\hat{\boldsymbol{\psi}}$ are given in the Appendix 2.7.2. From now on, we will denote the EB estimator in (2.4) as EB1 and the EB estimator that replaces $\hat{\boldsymbol{\psi}}\hat{\boldsymbol{\psi}}^\top$ with $\hat{\text{Cov}}(\hat{\boldsymbol{\psi}})$ in (2.4) as EB2.

The shrinkage factor assesses how close the assumed working DL function in CDLM is

to the pattern observed in the data. At one extreme, $\boldsymbol{K} = \boldsymbol{I}$ yields $\hat{\boldsymbol{\beta}}_{EB} = \hat{\boldsymbol{\beta}}_{CDLM}$. At the other extreme, $\boldsymbol{K} = \boldsymbol{0}$ yields $\hat{\boldsymbol{\beta}}_{EB} = \hat{\boldsymbol{\beta}}_{UDLM}$. When the working DL function in CDLM is not correctly specified, $\hat{\boldsymbol{\beta}}_{EB}$ is asymptotically equivalent to $\hat{\boldsymbol{\beta}}_{UDLM}$ and therefore $\hat{\boldsymbol{\beta}}_{EB}$ is consistent. Let $\boldsymbol{\Sigma}$ be the asymptotic variance-covariance matrix of $(\hat{\boldsymbol{\beta}}_{UDLM}^{\top}, \hat{\boldsymbol{\beta}}_{CDLM}^{\top})^{\top}$. Since $\hat{\boldsymbol{\beta}}_{EB}$ is a function of $\hat{\boldsymbol{\beta}}_{UDLM}$ and $\hat{\boldsymbol{\beta}}_{CDLM}$, the asymptotic variance-covariance of $\hat{\boldsymbol{\beta}}_{EB}$ can be expressed in the form of $\boldsymbol{G}\boldsymbol{\Sigma}\boldsymbol{G}^{\top}$ by using Taylor expansion where the exact expression of $\boldsymbol{G}$ is provided in the Appendix 2.7.2. The limiting distribution of $\sqrt{N}(\hat{\boldsymbol{\beta}}_{EB} - \boldsymbol{\beta})$ is not a normal distribution as expected for most model averaged estimators [Claeskens and Carroll, 2007]. However, Chen et al. [2009] showed that the normal approximation works well and is acceptable in practice.

2.3.3 Hierarchical Bayes Model

We propose a HB approach that sets up a nonnull shrinkage target through specification of the prior mean. The formulation of the prior rests on the "smoothness" prior [Shiller, 1973] that smooths over the lag curve by specifying a certain degree of order differences of $\boldsymbol{\beta}$ to follow a zero-mean normal distribution. For ease of presentation, we focus on polynomial DLM below. The prior structure can be represented by

$$\boldsymbol{R}_{p-1}\boldsymbol{\beta} \sim \mathcal{N}(\boldsymbol{0}, \sigma_{\pi}^{2}\boldsymbol{I}_{L-p+1}),$$

where $\boldsymbol{R}_{p-1}$ is a $(L-p+1)\times(L+1)$ constraint matrix for the $(p-1)^{th}$ degree smoothness prior that uses the p-degree order differences of $\boldsymbol{\beta}$ while σ_{π}^{2} is the prior variance. The element (i,j) of $\boldsymbol{R}_{p-1}$ is $(-1)^{(j-i)}\binom{p}{j-i}$ for $j = i, ..., i+p$ and 0 elsewhere. The shrinkage target implied by the prior specification lie in the space spanned by the solution of $\boldsymbol{R}_{p-1}\boldsymbol{\beta} = \boldsymbol{0}$ (i.e. $\sum_{j=0}^{p}(-1)^{j}\binom{p}{j}\beta_{\ell+j} = 0$ for $\ell = 0, 1, ..., L-p+1$). We have shown that the maximizer of the objective function in (2.2) subject to the constraint $\boldsymbol{R}_{p-1}\boldsymbol{\beta} = \boldsymbol{0}$ coincides with the $(p-1)$-degree polynomial DLM estimator. In other words, the smoothness approach is indeed shrinking $\hat{\boldsymbol{\beta}}_{UDLM}$ toward $\hat{\boldsymbol{\beta}}_{CDLM}$. The proof is provided in the Appendix 2.7.3. Without loss of generality, hereafter we denote $\boldsymbol{R}$ as the constraint matrix with M rows where $M < L+1$ is the number of constraints.

Define a $T \times (L+1)$ design matrix $\boldsymbol{X} = (\boldsymbol{X}_1, \cdots, \boldsymbol{X}_T)^\top$ and and an outcome vector $\boldsymbol{Y} = (y_1, \cdots, y_T)^\top$ of length T. In order to allow uncertainty on the variance component σ_π^2, we specify the full HB model as:

$$\begin{aligned}
\boldsymbol{Y}|\boldsymbol{\beta} &\sim \text{Poisson}(e^{\boldsymbol{X}\boldsymbol{\beta}}) \\
\boldsymbol{R}\boldsymbol{\beta}|\sigma_\pi^2 &\sim \mathcal{N}(\boldsymbol{0}, \sigma_\pi^2 \boldsymbol{I}_M) \\
\sigma_\pi^2 &\sim IG(a_\pi, b_\pi),
\end{aligned}$$

where a_π and b_π are hyper-prior parameters of the Inverse-Gamma (IG) distribution. The full conditional distributions of σ_π^2 and $\boldsymbol{\beta}$ are provided in Appendix 2.7.4. The marginal posterior density of $\boldsymbol{\beta}$ is not available in closed form. We use Metropolis Hastings algorithm within a Gibbs sampler to approximate the posterior distribution and obtain the HB estimator $\hat{\beta}_{HB}$ as the posterior mean.

The connection between Bayesian modelling and penalized likelihood approach by viewing prior as penalty is well-known. The dual problem of the HB model is to minimize

$$\ell_p(\boldsymbol{\beta}) = -\sum_{t=1}^{T} [y_t \boldsymbol{\beta}^\top \boldsymbol{X_t} - e^{\boldsymbol{\beta}^\top \boldsymbol{X_t}} - \log(y_t!)] + \lambda \boldsymbol{\beta}^\top \boldsymbol{R}^\top \boldsymbol{R} \boldsymbol{\beta}$$

where $\boldsymbol{R}$ is defined previously and λ is the tuning parameter. We can use the Newton-Raphson algorithm [Gill et al., 1981] to obtain GRR estimator $\hat{\boldsymbol{\beta}}_{GRR}$ by minimizing $\ell_p(\boldsymbol{\beta})$ given λ. GCV [Golub et al., 1979] and AICC [Hurvich et al., 1998] are two common criteria that can be used to choose the tuning parameter λ. Using the results demonstrated in the previous section, we can assure that $\hat{\boldsymbol{\beta}}_{GRR} \to \hat{\boldsymbol{\beta}}_{CDLM}$ as $\lambda \to \infty$ and $\hat{\boldsymbol{\beta}}_{GRR} \to \hat{\boldsymbol{\beta}}_{UDLM}$ as $\lambda \to 0$. The GRR model and HB model are similar and the major difference is in how the amount of shrinkage is determined. It has been shown that the asymptotic variance of $\hat{\boldsymbol{\beta}}_{GRR}$ is a monotonic decreasing function of λ, the asymptotic bias of $\hat{\boldsymbol{\beta}}_{GRR}$ is a monotonic increasing function of λ, and the asymptotic mean square errors (MSE) of $\hat{\boldsymbol{\beta}}_{GRR}$ is lower than the asymptotic MSE of $\hat{\boldsymbol{\beta}}_{UDLM}$. The proofs are provided in the Appendix 2.7.5. The described asymptotic properties assume that the tuning parameter λ is fixed. Choosing λ from data would induce another layer of uncertainty in $\hat{\boldsymbol{\beta}}_{GRR}$ and the derived variance formula may underestimate its true variance. To address this issue, we

compare the proposed variance estimator with the empirical variance of the estimates in our simulation study in section 2.4.

2.3.4 Two-stage Shrinkage

The Bayesian distributed lag model (BDLM) proposed by Welty et al. [2009] smooths over the lagged effects $\boldsymbol{\beta}$. They construct the prior variance-covariance matrix on $\boldsymbol{\beta}$ in a way to ensure $\text{Var}(\beta_\ell) \to 0$ and $\text{Cor}(\beta_{\ell-1}, \beta_\ell) \to 1$ as ℓ increases. The following hierarchy is specified:

$$\boldsymbol{Y}|\boldsymbol{\beta} \sim \text{Poisson}(e^{\boldsymbol{X}\boldsymbol{\beta}})$$
$$\boldsymbol{\beta}|\boldsymbol{\omega}, \sigma^2 \sim \mathcal{N}(0, \sigma^2\boldsymbol{\Omega}(\boldsymbol{\omega})), \quad \boldsymbol{\Omega}(\boldsymbol{\omega}) = \boldsymbol{V}(\omega_1)\boldsymbol{W}(\omega_2)\boldsymbol{V}(\omega_1), \quad \sigma^2 = 10 \cdot \text{Var}(\hat{\beta}_0),$$

where $\boldsymbol{V}(\omega) = \text{diag}[1, \exp(\omega), \exp(2\omega), ..., \exp(L\omega)]$, $\boldsymbol{W}(\omega_2) = \boldsymbol{V}(\omega_2)\boldsymbol{V}(\omega_2)^\top + \{\boldsymbol{I}_{L+1} - \boldsymbol{V}(\omega_2)\}\mathbf{1}_{L+1}\mathbf{1}_{L+1}^\top\{\boldsymbol{I}_{L+1} - \boldsymbol{V}(\omega_2)\}^\top$, $\boldsymbol{I}_{L+1}$ is the $(L+1) \times (L+1)$ identity matrix, $\mathbf{1}_{L+1}$ is a $(L+1) \times 1$ vector of ones, and $\hat{\beta}_0$ is the estimated coefficient for lag 0 from unconstrained DLM. Rather than setting fixed values for $\boldsymbol{\omega} = (\omega_1, \omega_2)^\top$, Welty et al. [2009] lets $\boldsymbol{\omega}$ follow a discrete uniform distribution on $\mathbb{R}^2$ and the posterior distribution of $\boldsymbol{\beta}$ can be obtained accordingly.

We consider a two-stage shrinkage approach to ensure the additional property that the estimated DL coefficients from one of the above shrinkage approaches smoothly go to zero at larger lags. In the first stage, we shrink $\hat{\boldsymbol{\beta}}_{UDLM}$ toward $\hat{\boldsymbol{\beta}}_{CDLM}$ through one of the shrinkage approaches introduced in section 2.3.2-2.3.3. In the second stage, we specify the hyperprior on the variance-covariance matrix on $\boldsymbol{\beta}$ that constrains the coefficients at larger lags to approach zero similar to BDLM. Without loss of generality, we consider the EB-type estimator $\hat{\boldsymbol{\beta}}_{EB}$ as the shrinkage estimator from the first stage. The full specification of the two-stage shrinkage model, with $\boldsymbol{G}$ and $\boldsymbol{\Sigma}$ defined in section 2.3.2, is given by:

$$\hat{\boldsymbol{\beta}}_{EB}|\boldsymbol{\beta} \sim \mathcal{N}(\boldsymbol{\beta}, \boldsymbol{G}\boldsymbol{\Sigma}\boldsymbol{G}^\top)$$
$$\boldsymbol{\beta}|\boldsymbol{\omega}, \sigma^2 \sim \mathcal{N}(0, \sigma^2\boldsymbol{\Omega}(\boldsymbol{\omega})), \quad \boldsymbol{\Omega}(\boldsymbol{\omega}) = \boldsymbol{V}(\omega_1)\boldsymbol{W}(\omega_2)\boldsymbol{V}(\omega_1), \quad \sigma^2 \sim IG(a_0, b_0),$$

where $\boldsymbol{V}(\omega)$ and $\boldsymbol{W}(\omega)$ are as defined in section 2.3.3. The full conditional distributions of

$\boldsymbol{\beta}$, σ^2, and $\boldsymbol{\omega} = (\omega_1, \omega_2)^\top$ are provided in Appendix 2.7.4. The joint posterior distribution can be obtained via a Gibbs sampling technique and the two-stage shrinkage estimate $\hat{\boldsymbol{\beta}}_{TSB}$ can be obtained accordingly.

The analogue of the previous two-stage shrinkage approach is the two-stage hyper-penalized approach. Again, the estimator from the first stage can be any one of the shrinkage estimators introduced previously. We take $\hat{\boldsymbol{\beta}}_{EB}$ as the shrinkage estimator obtained in the first stage as before. A penalized objective function is constructed in the second stage to penalize the departure from $\text{Var}(\beta_\ell) \to 0$ and $\text{Cor}(\beta_{\ell-1}, \beta_\ell) \to 1$ as ℓ increases. The two-stage hyper-penalized estimator is given by

$$\hat{\boldsymbol{\beta}}_{TSP} = \arg\min_{\boldsymbol{\beta}} \ell_{TSP}(\boldsymbol{\beta}) = \arg\min_{\boldsymbol{\beta}} [(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}_{EB})^\top (\boldsymbol{G}\boldsymbol{\Sigma}\boldsymbol{G}^\top)^{-1} (\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}_{EB}) + \lambda \boldsymbol{\beta}^\top \boldsymbol{\Omega}(\boldsymbol{\omega})^{-1} \boldsymbol{\beta}],$$

where λ is the tuning parameter. We select λ based on cross-validation. For $\boldsymbol{\omega}$, we search through a grid of possible values of $\boldsymbol{\omega}$ and choose the values that minimize the above criterion. When $\hat{\boldsymbol{\beta}}_{GRR}$ is chosen as the shrinkage estimator from the first stage, a similar framework can be followed.

2.4 Simulation Study

2.4.1 Simulation 1: Comparison of Single-Step Shrinkage Approaches

We conducted a simulation study to compare the estimation properties of UDLM, CDLM, GADLM, BDLM, and the three shrinkage approaches introduced in sections 2.3.2 and 2.3.3 under a time-series setting. All together, we considered eight different smoothing methods: UDLM, CDLM, EB1, EB2, GRR (with tuning parameter selected via AICC), GADLM, BDLM, and HB. Among these, UDLM, CDLM, BDLM, and GADLM are existing alternatives. A cubic spline with four equally spaced internal knots is applied for GADLM. The prior on $\boldsymbol{\omega} = (\omega_1, \omega_2)^\top$ for BDLM was set to be a discrete uniform distribution over the equally spaced sequence of length 50 ranging from -0.2 to -0.004 in both dimensions. The hyperprior on the variance for HB was set to be weakly informative [Gelman et al., 2008], with both inverse gamma prior parameters set to 0.001.

2.4.1.1 Simulation Settings

We first generated an exposure series of length 200 with mean 0 and first order auto-correlation equal to 0.6 from the model $x_t = 0.6x_{t-1} + \epsilon_t$ where $\epsilon_t \sim$ i.i.d $N(0, 1)$ for $t = 1, ..., 200$. Following the structure of Welty et al. [2009], we simulated the outcome series $\boldsymbol{Y}$ as continuous rather than count data for simplicity. The continuous $\boldsymbol{Y}$ can represent the logarithm transformation of the counts and the normal approximation is applied for modeling purposes. We set $L = 10$ and generated the outcome series $\boldsymbol{Y}$ from the model $y_t = \sum_{\ell=0}^{10} \beta_\ell x_{t-\ell} + \epsilon_t$ where $\boldsymbol{\beta} = (\beta_0, ..., \beta_{10})^\top$ denote the true coefficients and $\epsilon_t \sim$ i.i.d $N(0, 0.25)$ for $t = 1, ..., 200$. The error variance was determined to control the signal-to-noise ratio.

Four sets of true $\boldsymbol{\beta}$s were considered and different specifications of the working DL function in CDLM were used. Shrinkage between UDLM and the CDLM constructed based on the working DL was performed for all shrinkage methods including EB1, EB2, GRR, and HB. The three combinations of true coefficients and specified working DL function reflect the first three scenarios of interest for comparing various methods: (1) the working DL function completely matches true DL function, (2) the working DL function moderately departs from the true DL function, and (3) the working DL function is very different from the true DL structure. Scenario 4 is created to reflect a realistic situation when one is exploring association between multiple pollutants (e.g. O_3, CO, SO_2, NO, PM_{10}) and various outcomes (e.g. mortality, cardiovascular events, hospital admission). Each exposure-outcome pair may have a different DL structure and it is not feasible to examine each structure in depth. We consider a setting where data are generated from one of the five underlying true DL functions, including (a) constant, (b) linear, (c) cubic, (d) cubic-like smooth function with slight departure, and (e) oscillating, is used to generate data with 20% frequency each while the working DL function is a cubic polynomial. The summary parameter configurations corresponding to the four scenarios is provided in Table 2.4. We generated 1000 data sets for each scenario to evaluate the estimation performance.

2.4.1.2 Evaluation Metrics

To compare the estimation performance of the eight methods, we used two sets of metrics. The first set of metrics measures the estimation properties of $\hat{\boldsymbol{\beta}}$ as a vector. They are (i) squared bias, (ii) variance, (iii) relative efficiency with respect to UDLM, and (iv) the mean Euclidean distance to the true coefficient. The second set of metrics measures the estimation properties of the total effect (i.e. $\sum_{j=0}^{10} \beta_j$). The metrics are (i) squared bias, (ii) variance, and (iii) relative efficiency with respect to UDLM. The relative efficiency is the ratio of the MSE of UDLM estimates to the MSE of the estimate under each method. The expressions of the metrics used for comparison are summarized in Table 2.5.

2.4.1.3 Simulation Results

The simulation results for the estimated lagged coefficient vector ($\hat{\boldsymbol{\beta}}$) are summarized in the upper part of Table 2.1. As we observe, in scenario 1 when the working DL function completely matches the true DL function, CDLM is nearly unbiased with lowest variance and MSE across all the methods as expected. The relative efficiency is 8.43. Nonetheless, GRR, HB, and GADLM with relative efficiency ranging from 4.52 to 5.38 perform reasonably well and are superior to EB1, EB2, and BDLM with relative efficiency ranging from 1.68 to 1.99. In scenario 2 when the working DL function moderately departs from the true DL function, CDLM is more efficient than UDLM, with the loss from the bias compensated for by a large reduction in variance. CDLM has relative efficiency equal to 2.26 and the relative efficiencies of the shrinkage methods range from 1.56 to 4.22. GRR and HB outperform CDLM and UDLM in terms of relative efficiency and mean distance whereas EB1 and EB2 are less efficient than CDLM. BDLM is approximately as efficient as CDLM, and the mean distances are similar. When the working DL function is very different from the true DL structure as depicted in scenario 3, CDLM and GADLM are the least efficient with relative efficiency around 0.70 since the large squared bias contributes to the MSE despite the low variance. All the shrinkage methods and BDLM outperform both UDLM and CDLM in terms of efficiency and mean distance in this scenario. In scenario 4, we can observe that GRR (2.09) and HB (2.22) have higher relative efficiency compared

to other methods as well as stable performances across different individual lag structures. This simulation scenario illustrates that the shrinkage methods can be useful in improving robustness as well as retaining reasonable precision when encountering uncertainty in real-world analysis. Overall, GRR and HB have the best average performance across various lag structures (scenario 4), as well as reasonable efficiency under a given lag structure (scenarios 1-3). For example, GRR has relative efficiency of 5.38, 3.54, 1.15 and 2.09 and HB has relative efficiency of 4.52, 4.22, 1.37 and 2.26 across simulation scenarios 1-4. Based on the simulation results, HB and GRR have robust performance.

The simulation results for the estimated total effect ($\sum_{l=0}^{10} \hat{\beta}_\ell$) are summarized in the lower part of Table 2.1. As we can see in scenarios 1 and 2, all the methods yield nearly unbiased estimates for total effect and the variances are at a similar level except for EB1 and EB2. In scenario 3, when the true DL is non-smooth, the total effects estimated from EB1, EB2, GADLM, and BDLM are slightly biased. In terms of relative efficiency, GRR, GADLM, BDLM, and HB are approximately as efficient as UDLM for estimating the total effect. Overall, the biases of the total effect estimates are minimal and the variances of the total effect estimates are similar across the board with slightly higher values for EB1 and EB2.

2.4.2 Simulation 2: Comparison of Two-stage Shrinkage Approaches

Our second simulation study was designed to investigate the effect of the two-stage shrinkage when the number of maximum lag is allowed to be much larger than the truth. We considered seven methods - EB1, TSB with EB1 from the first stage (TSB-EB1), TSP with EB1 from the first stage (TSP-EB1), GRR, TSB with GRR from the first stage (TSB-GRR), TSP with GRR from the first stage (TSP-GRR), and BDLM. For BDLM and TSB, the prior on $\boldsymbol{\omega} = (\omega_1, \omega_2)^\top$ was set to be a discrete uniform distribution over the equally spaced sequence of length 50 ranging from -0.2 to -0.004 in both dimensions. For TSP, $\boldsymbol{\omega}$ was chosen as the minimizer of the hyper-penalized criterion. The tuning parameters in TSP-EB1 and TSP-GRR were selected based on 5-fold cross-validation. The working DL function in CDLM was specified as a cubic polynomial throughout.

2.4.2.1 Simulation Settings

We generated exposure series in the same way as the first simulation study. $L = 15$ was chosen and the true lagged coefficients beyond lag 7 are all set equal to 0. We generated the outcome series $\boldsymbol{Y}$ from the model $y_t = \sum_{\ell=0}^{15} \beta_\ell x_{t-\ell} + \epsilon_t$ with true coefficients $\boldsymbol{\beta} = (0.07, 0.135, 0.2, 0.210, 0.18, 0.125, 0.06, 0.02, 0, 0, 0, 0, 0, 0, 0, 0)^\top$ and $\epsilon_t \sim$ i.i.d $N(0, 0.25)$ for $t = 1, ..., 200$. We generated 1000 data sets to evaluate the estimation performance.

2.4.2.2 Evaluation Metrics

We evaluated the estimation properties of the seven methods based on the same four metrics used in the first simulation scenario. The two-stage shrinkage methods can potentially alleviate the problem of having nonzero coefficient estimates at larger lags when the number of maximum lags is large. Let MAV denote the mean absolute value of the coefficient estimates for the lags with the true coefficients equal to 0 (i.e. MAV = $\frac{1}{8000} \sum_{i=1}^{1000} \sum_{j=8}^{15} |\hat{\beta}_{ij}|$). We examine the MAVs of the seven methods to assess their performance when the maximum number of lags L is misspecified.

2.4.2.3 Simulation Results

The results are presented in Table 2.2. Overall, the two-stage approaches are effective in increasing efficiency when L is misspecified. Both TSB and TSP further reduce MSE and reduce the mean distance compared to the shrinkage estimator obtained in the first stage. Specifically, compared to EB1 (1.83), TSB-EB1 (1.95) and TSP-EB1 (1.98) have higher efficiencies while all three are less efficient than BDLM (3.42); in contrast, TSB-GRR (10.47) and TSP-GRR (10.13) have higher efficiencies compared to GRR (6.54). The efficiency gain from the second-stage shrinkage is limited for EB1 while the gain is considerable for GRR.

The MAVs of the seven methods being compared are 0.047, 0.040, 0.029, 0.025, 0.012, 0.012, 0.019, respectively. The reduction from 0.047 to 0.040 and 0.029, corresponding to 15% and 37% reduction in MAV, suggests the usefulness of imposing a second-stage

shrinkage on EB1 to mitigate the "tail" problem. Similarly, a second stage shrinkage on GRR aids in reducing the MAVs from 0.025 to 0.012 and 0.012, equivalent to 49% and 50% reduction in MAV. In this setting, a two-stage shrinkage approach with GRR in the first stage (TSB-GRR) performs the best with respect to relative efficiency, mean distance to the true coefficients, and MAV.

Remark 2: We conducted an analysis to evaluate whether ignoring the uncertainty from choosing the tuning parameter λ in GRR would underestimate the variance of the cumulative effects which are one of the primary quantities of interest in our context. We considered the empirical variance of the 1000 cumulative estimates up to lag ℓ from 1000 repetitions as the reference (i.e. $\frac{1}{1000}\sum_{i=1}^{1000}(\sum_{j=0}^{\ell}\hat{\beta}_{ij} - \sum_{j=0}^{\ell}\bar{\beta}_j)^2$ for $\ell = 0, ..., 10$). We computed the average of the 1000 estimated variances of the cumulative lag coefficients from the 1000 repetitions (i.e. $\frac{1}{1000}\sum_{i=1}^{1000}\hat{\text{Var}}(\sum_{j=0}^{\ell}\hat{\beta}_{ij})$ for $\ell = 0, ..., 10$) as a percentage of the reference. The results are presented in Table 2.6. We observe that the asymptotic variances are slightly smaller on average than the empirical variances. The percentages range from 0.83 to 1.02 across simulations, indicating no more than 10% underestimation of the standard errors. The findings are in line with the coverage properties of confidence intervals of GAMs using penalized regression splines studied by Marra and Wood [2012]. To ensure the validity of comparison across different methods, we will consider bootstrapping to obtain standard error estimates for GRR and TSP-GRR in the analysis of NMMAPS Data.

2.5 Application to NMMAPS Data

We first explore the association of (1) daily PM_{10}, (2) daily O_3, and (3) daily SO_2 with (1) daily non-accidental mortality counts, (2) daily cardiovascular mortality counts, and (3) daily respiratory mortality counts in Chicago, Illionis for the period between 1987 and 2000 using part of the NMMAPS data via UDLM, CDLM, and HB. A cubic polynomial working DL function was applied for CDLM and is set as the shrinkage target for all shrinkage methods. We then applied eight of the methods (UDLM, CDLM, EB1, GRR, BDLM, HB, TSB-GRR, TSP-GRR) included in the simulation study to investigate the association of PM_{10} and O_3 with mortality counts and compare and contrast the two distributed lag

analyses. A 4-degree polynomial working DL function was applied. The NMMAPS data contain daily mortality, air pollution, and weather data collected across 108 metropolitan areas in the United States from 1987 to 2000. Further details with respect to NMMAPS data are available at http://www.ihapss.jhsph.edu/data/NMMAPS/.

Zanobetti et al. [2002] have shown that it is unlikely that lags beyond two weeks would have substantial influence on associations between short-term exposures to pollution and mortality; rather, inclusion of lags beyond two weeks might confound the estimation of lagged effects. We consider lags up to $L = 14$ for PM_{10}, O_3, and SO_2. Let x_{tk}, y_{tk}, and $\boldsymbol{z}_{tk}$ denote exposure level, outcome count, and vector of time-varying covariates, measured on day t for age group k in Chicago with $t = 1, ..., 5114$ and $k = 1, 2, 3$, respectively. The three age categories are greater or equal to 75 years old, between 65 and 74 years old, and less than 65 years old. The three exposures were shared across the three age groups (i.e. $x_{tk} \equiv x_t$) and the vector of covariates $\boldsymbol{z}_{tk}$ is specified in the same way as in previous analysis by Dominici et al. [2005]. The same set of covariates is considered in the models for all exposures. We assume that the mortality count in Chicago on day t for each of the age group k is a Poisson random variable Y_{tk} with mean μ_{tk} such that

$$\begin{aligned}
\log(\mu_{tk}) &= \boldsymbol{X}_t^\top \boldsymbol{\beta} + \boldsymbol{z}_{tk}^\top \boldsymbol{\alpha} \\
&= \boldsymbol{X}_t^\top \boldsymbol{\beta} + \alpha_0 + \sum_{j=1}^{2} \alpha_{1j} \mathrm{I}(k = j) + \sum_{j=1}^{6} \alpha_{2j} \mathrm{I}(\mathrm{dow}_t = j) + \mathrm{ns}(\mathrm{temp}_t; 6 \text{ df}, \boldsymbol{\alpha}_3) \\
&\quad + \mathrm{ns}(\overline{\mathrm{temp}}_t^{(3)}; 6 \text{ df}, \boldsymbol{\alpha}_4) + \mathrm{ns}(\mathrm{dptp}_t; 3 \text{ df}, \boldsymbol{\alpha}_5) + \mathrm{ns}(\overline{\mathrm{dptp}}_t^{(3)}; 3 \text{ df}, \boldsymbol{\alpha}_6) \\
&\quad + \mathrm{ns}(t; 98 \text{ df}, \boldsymbol{\alpha}_7) + \mathrm{ns}(t; 14 \text{ df}, \boldsymbol{\alpha}_8)\mathrm{I}(k = 1) + \mathrm{ns}(t; 14 \text{ df}, \boldsymbol{\alpha}_9)\mathrm{I}(k = 2),
\end{aligned}$$

where $\boldsymbol{X}_t = (x_t, ..., x_{t-14})^\top$, $\boldsymbol{\beta} = (\beta_0, ..., \beta_{14})^\top$, $\mathrm{I}(\cdot)$ is the indicator function,$\mathrm{ns}(\cdot)$ denotes the natural spline with specified degrees of freedom (df) and $\boldsymbol{\alpha}_i$ represents the spline coefficients for $i = 3, ..., 9$. Predictors dow_t, temp_t, $\overline{\mathrm{temp}}_t$, dptp_t, and $\overline{\mathrm{dptp}}_t$ represent the day of week, current day's temperature, average of the previous 3 days' temperatures, current day's dewpoint temperature, and the average of the previous 3 days' dewpoint temperatures for day t. The indicator variables allow different baseline mortality rates within each age group and within each day of week. The smooth term for time (t) is to adjust for long-term

trends and seasonality and 98 df corresponds to 7 df per year over the 14-year horizon. The last two product terms separate smooth functions of time with 2 df per year for each age group contrast. The primary goal is to estimate the lagged coefficients $\boldsymbol{\beta}$ while $\boldsymbol{\alpha}$ is the set of covariate related parameters.

The mean concentrations (standard deviations in parenthesis) of PM_{10}, O_3, and SO_2 are 37.06 (19.25) μg/m^3, 19.14 (10.20) ppb, and 6.24 (2.95) ppb, respectively. The average daily non-accidental mortality count, daily cardiovascular mortality count, and daily respiratory mortality count are 38.47 (15.89), 16.97 (10.63), and 3.06 (2.73), respectively. We present the results of exploratory analysis in Figure 2.1. Along the columns, we can see that the estimated DL functions for cardiovascular deaths are similar to the estimates for total mortality while the estimated DL functions for respiratory deaths are less informative across different exposures. The finding suggests that cardiovascular death is the leading composite of mortality in association with PM_{10}, O_3, and SO_2. Along the rows, we can see that the fitted DL functions of PM_{10} and SO_2 are similar in that they increase at early lags, decrease at mid-range lags, and increase back to 0 line at late lags. The trend suggests the delayed effects of PM_{10} and SO_2 and the phenomenon of mortality displacement [Zanobetti et al., 2002, Zanobetti and Schwartz, 2008]. On the other hand, the fitted DL functions of O_3 peak at earlier lags and decrease toward 0 at large lags suggesting the acute effects of O_3 compared to PM_{10} and SO_2. Departure of HB fit from the CDLM fit for PM_{10} indicates that better bias-variance tradeoff can be achieved using shrinkage while the consistency between the CDLM fits and HB fits for O_3 and SO_2 suggest that the CDLM fits are adequate and the HB approach data-adaptively aligns with CDLM in these situations.

Partial autocorrelation function (PACF) plots of PM_{10} and O_3 are presented in the Figure 2.5. One can notice the slower decay and stronger autocorrelation in O_3 time series than in PM_{10} time series. Figure 2.2 compares the estimated DL functions obtained from the eight methods for the association between PM_{10} and O_3 and mortality in Chicago from 1987 to 2000. The stronger autocorrelation of O_3 time series corresponds to the more variable UDLM estimates. In addition, PM_{10} demonstrates the strongest positive effects at lag 2-3, whereas O_3 starts to demonstrate a positive effect at lag 0 itself. This observation suggests an earlier onset of the short-term ozone effect on mortality in Chicago during the study

period.

2.5.1 Estimation of Lag Coefficients

With respect to PM_{10}, the strongest association occurs at lag 3 for UDLM, EB1, GRR, BDLM, and TSP-GRR and at lag 2 for CDLM, HB, and TSB-GRR. The interquartile range of PM_{10} is $21.49\mu g/m^3$. The quantity $100[\exp(21.49\beta_\ell) - 1]$ represents the percentage change in daily mortality with an interquartile range (IQR) increase in PM_{10} at lag ℓ. The estimated percentage increases in mortality associated with a $21.49\mu g/m^3$ increase in PM_{10} at lag 3 are 0.65%, 0.56%, 0.44%, 0.54%, and 0.37% for UDLM, EB1, GRR, BDLM, and TSP-GRR, respectively. All of the 95% confidence/credible intervals (CIs) do not contain zero suggesting that PM_{10} at lag 3 is significantly associated with daily mortality. Although all other methods shrink and smooth the DL function and result in attenuated lagged effect estimates, the standard error estimates are smaller as well. From the left panel of Figure 2.2, we can observe that the estimated DL function obtained by HB and GRR for PM_{10} is closer to the UDLM fit than the CDLM fit indicating that CDLM might have led to over-smoothing the DL function. Consequently, the effects at lags 2 and 3 are much less evident for CDLM compared to UDLM, GRR, and HB due to potential underestimation of the effects. In this example, shrinkage methods are certainly preferred since CDLM is potentially underestimating the effects by misspecifying the DL function.

In contrast, the strongest association unequivocally occurs at lag 2 across all eight methods for O_3. The IQR of O_3 is 14.65 ppb. The quantity $100[\exp(14.65\beta_\ell) - 1]$ represents the percentage change in daily mortality with an IQR increase in O_3 at lag ℓ. The estimated percentage increases in mortality associated with a 14.65 ppb increase in O_3 at lag 2 range from 0.59% to 1.19% across the eight methods. All of the 95% CIs do not cover zero indicating that O_3 at lag 2 is significantly associated with daily mortality in Chicago from 1987 to 2000. The peak at earlier lags for O_3 indicates an earlier window of susceptibility and a more acute effect on mortality compared to PM_{10}. The estimated DL function of GRR/HB is more similar to the CDLM fit in this case. The two examples also illustrate the data adaptive feature of GRR/HB. In a given situation, one will not know which DL structure

is the best and GRR/HB can be taken as a default choice that will automatically adapt the fit. The estimated lagged effects with 95% CIs obtained for PM_{10} and O_3 are tabulated in the Table 2.4. The BDLM and two-stage approaches enable us to misspecify L or allow a maximal large L. Figure 2.3 shows analysis of NMMAPS data with $L = 28$ and one can note the attractive feature of the two-step approaches, providing accurate and efficient estimates at smaller lags, and shrinking the coefficients at larger lags to zero.

2.5.2 Estimation of Cumulative Lag Coefficients

Table 2.7 summarizes the estimated cumulative lagged effects of PM_{10} and O_3 on mortality up to lag 3, lag 7, and lag 14, respectively, with an IQR increase in exposure level. The corresponding graphical representation is shown in Figure 2.4. An interquartile ($21.49\mu g/m^3$) increase in PM_{10} in each of lag 0 to lag 3 is associated with an increase in relative risk of mortality ranging from 0.48% to 0.75% across different methods. The 95% CIs with lower bound close to 0 suggest plausible positive association. However, the estimated cumulative lagged effects up to lag 7 range from 0.13% to 0.41% across the eight methods with all the 95% CIs containing 0. The drop between lag 3 and lag 7 suggests the phenomenon of mortality displacement that has been noted in previous studies [Zanobetti et al., 2002]. The deaths of frail individuals would occur several days after the high air pollution level episode resulting in the DL function to be positive at early lags and decrease and then become negative at larger lags. The estimates of the total effect (up to lag 14) from all eight methods are similar, ranging from -0.87% to -0.43%. The finding is consistent with results from the simulation study. The proposed shrinkage methods are capable of capturing the trend of the DL functions (i.e. effects at each individual lag) more precisely than other methods, whereas the total effect estimates and their standard errors are usually similar across methods. From Figure 2.2, we can also observe that the two-stage shrinkage methods TSB-GRR and TSP-GRR shrink the tail of the estimated DL function towards 0. A interquartile(14.65 ppb) increase in O_3 in each of lag 0 to lag 3 is associated with an increase in relative risk of mortality ranging from 1.81% to 2.07% across different methods. All the 95 % CIs are above 0 indicating the positive short-term effects of ozone on mortality in Chicago. The

slightly larger cumulative effects up to lag 7 compared to the cumulative effects up to lag 3 suggests the tapering positive ozone effect on mortality between lag 3 and lag 7. In addition, the slightly smaller cumulative effects up to lag 14 compared to the cumulative effects up to lag 7 suggests the "harvesting" effects [Zanobetti and Schwartz, 2008].

2.6 Discussion

In this chapter, we first reviewed unconstrained DLMs and constrained DLMs for modeling the lagged effects of air pollution levels on a health outcome in a time-series setting. The unconstrained DLM estimator is robust because it imposes no constraint on the DL function, whereas the constrained DLM estimator is efficient due to parsimony. We introduced three classes of statistical approaches to combine the two estimators in order to achieve bias-variance tradeoff. The commonality is that the amount of shrinkage is determined in a data-adaptive manner. The resulting shrinkage estimators are found to be more robust to deviation of the working DL function in CDLM from the true DL function. They are more efficient than a vanilla unconstrained DLM estimator across the board. Our simulation results indicate that GRR and HB perform well in terms of estimation accuracy across different simulation scenarios. GADLM is competitive when the true DL function is smooth but it leads to seriously biased estimates when the true DL function is non-smooth (simulation setting 3). In contrast to spline-based DLMs and BDLM, our shrinkage approaches leverage the efficiency gain from the parsimonious parametrization of the working DL function in CDLM.

Based on the simulation results, we recommend GRR and HB as the preferred methods of choice. With massive data sets or multiple exposure-outcome pairs to explore, if computational cost is of concern, GRR is computationally less expensive than HB. To help understand the differences in relative computing times, Table 2.12 presents the computational time for analyzing the NMMAPS data by each method. Moreover, existing methods like CDLM require the DL function be carefully selected on a case-by-case basis. Practitioners may not have the resources to conduct such in depth exploration of the lag structure when an agnostic association analysis is carried out with multiple outcome-exposure combina-

tions. Use of shrinkage methods can be viewed as a way to automate this process and avoid selection of a parametric structure for each individual analysis, as in simulation Scenario 4 and NMMAPS analysis. The proposed shrinkage methods are robust to misspecification of the working DL function and can be used to conduct agnostic discovery searches in an automatic and efficient fashion.

One of the key components for setting up the smoothness prior in HB and the penalty term in GRR is the configuration of the constraint matrix $\boldsymbol{R}$. It induces a nonnull shrinkage target in both approaches. We established the connection between $\boldsymbol{R}$ and the transformation matrix $\boldsymbol{C}$ in DLM framework. This correspondence is a major contribution of the paper. There are two implications of this connection. First, $\boldsymbol{R}$ can be conveniently obtained as long as $\boldsymbol{C}$, that transforms the constrained parameters in the original space to the parameters in a lower-dimensional unconstrained space, is available. Second, one can explicitly determine the constraint(s) between adjacent lag coefficients by integrating subject-matter knowledge about the shape and smoothness regarding the DL function and define the corresponding $\boldsymbol{C}$ or $\boldsymbol{R}$, thus the framework is flexible.

Unconstrained DLMs, constrained DLMs, and the other one-stage shrinkage methods do not guarantee that the coefficients at larger lags approach zero. Two-stage shrinkage methods are useful in remedying this problem. However, the computation time needed is longer as taking into account the uncertainty at both stages concurrently requires some resampling technique such as bootstrapping. Overall, the choice of the methods has less influence on the estimated cumulative effects, as observed in the simulation study and the NMMAPS analysis. Nevertheless, the shrinkage methods are useful in characterizing the DL functions in a more precise manner by recognizing the possible bias in the CDLM specification. Precisely identifying the window of susceptibility to a disease event in association with air pollution would facilitate environmental scientists to understand the pathway of environmental factors to disease risk and possible interaction between different exposures.

These methods can potentially be extended to areas outside environmental epidemiology. The notion of combining a model-free estimator and a model-based estimator is attractive in real-world situations when no single estimator is universally optimal and it is difficult to examine the validity of the underlying assumptions needed for a model-based

estimator. We hope that our work will lead to further research in other applications.

2.7 Appendix

2.7.1 Connection Between C and R beyond Polynomial DLM

Denote

$\boldsymbol{C}$: a $(L+1) \times p$ transformation matrix

$\boldsymbol{R}$: a $(L+1-p) \times (L+1)$ constraint matrix

$\boldsymbol{C}_e$: a $(L+1) \times (L+1)$ matrix $[\boldsymbol{C}\ \mathbf{0}_{(L+1)\times(L+1-p)}]$ where $\mathbf{0}_{(L+1)\times(L+1-p)}$ is a $(L+1) \times (L+1-p)$ zero matrix

$\boldsymbol{R}_e$: a $(L+1) \times (L+1)$ matrix $\begin{bmatrix} \boldsymbol{R} \\ \mathbf{0}_{p\times(L+1)} \end{bmatrix}$ where $\mathbf{0}_{p\times(L+1)}$ is a $p \times (L+1)$ zero matrix

(1) $\boldsymbol{R} \to \boldsymbol{C}$

The p basis functions corresponding to the p columns of $\boldsymbol{C}$ span the solution space of $\boldsymbol{R\beta} = 0$ (or $\boldsymbol{R}_e\boldsymbol{\beta} = 0$). $\boldsymbol{C}$ can be obtained by applying SVD on $\boldsymbol{R}_e$ (i.e. $\boldsymbol{R}_e = \boldsymbol{U}_R\boldsymbol{D}_R\boldsymbol{V}_R^\top$). The last p columns of $\boldsymbol{V}_R$ is one choice of $\boldsymbol{C}$.

(2) $\boldsymbol{C} \to \boldsymbol{R}$

$\boldsymbol{\beta} = \boldsymbol{C\eta}$ and $\boldsymbol{R\beta} = 0$ so we have $\boldsymbol{RC\eta} = 0$. Deriving $\boldsymbol{R}$ from $\boldsymbol{C}$ is equivalent of solving $\boldsymbol{C}^\top\boldsymbol{R}^\top = 0$ (or $\boldsymbol{C}_e^\top\boldsymbol{R}^\top = 0$). $\boldsymbol{R}$ can be obtained by applying SVD on $\boldsymbol{C}_e^\top$ (i.e. $\boldsymbol{C}_e^\top = \boldsymbol{U}_C\boldsymbol{D}_C\boldsymbol{V}_C^\top$). The last $(L+1-p)$ rows of $\boldsymbol{V}_C^\top$ is one choice of $\boldsymbol{R}$.

Remark: DLM solution is invariant to row operations on $\boldsymbol{R}$. For example, consider a piecewise linear distributed lag function with $L = 6$ and only internal knot at 3. With basis functions 1, ℓ, and $(\ell - 3)_+$, $\boldsymbol{C}$ is given by

$$\boldsymbol{C} = \begin{bmatrix} 1 & 0 & 0 \\ 1 & 1 & 0 \\ 1 & 2 & 0 \\ 1 & 3 & 0 \\ 1 & 4 & 1 \\ 1 & 5 & 2 \\ 1 & 6 & 3 \end{bmatrix}.$$

Following the above procedure, $\boldsymbol{R}$ can be obtained as

$$\boldsymbol{R} = \begin{bmatrix} 0.000 & -0.346 & 0.693 & -0.030 & -0.255 & -0.439 & 0.377 \\ 0.000 & -0.218 & 0.436 & -0.290 & -0.238 & 0.691 & -0.381 \\ 0.000 & -0.090 & 0.179 & -0.549 & 0.779 & -0.180 & -0.139 \\ -0.560 & 0.727 & 0.226 & -0.275 & -0.157 & -0.039 & 0.079 \end{bmatrix}.$$

Through row operations, we can obtain

$$\boldsymbol{R} = \begin{bmatrix} 1 & -2 & 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & -2 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & -2 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 & -2 & 1 \end{bmatrix}$$

as suggested. The solution of $\boldsymbol{R}\boldsymbol{\beta} = \boldsymbol{0}$ is a piecewise linear function with internal knot at 3.

2.7.2 Asymptotic Results for the Empirical Bayes estimator

We first derive the variance-covariance expression of $\hat{\boldsymbol{\psi}} = \hat{\boldsymbol{\beta}}_{UDLM} - \hat{\boldsymbol{\beta}}_{CDLM}$ and then obtain the asymptotic theory of $\hat{\boldsymbol{\beta}}_{EB1}$. Let $S_U^{(t)}(\boldsymbol{\beta})$ denote the first-order derivative of the unconstrained DLM likelihood for time t (i.e. $(y_t - e^{-\boldsymbol{X}_t^\top \boldsymbol{\beta}})\boldsymbol{X}_t$) $S_C^{(t)}(\boldsymbol{\theta})$ denote the first-order derivative of the constrained DLM likelihood for time t (i.e. $(y_t - e^{-\boldsymbol{Z}_t^\top \boldsymbol{\theta}})\boldsymbol{Z}_t$), and let $H_U(\boldsymbol{\beta})$ and $H_C(\boldsymbol{\theta})$ denote the Hessian matrices from the two models, respectively. Let $\boldsymbol{\beta}_0$

denote the true vector of lagged coefficients. By Taylor expansion,

$$
\begin{aligned}
&\ell^{'}(\hat{\boldsymbol{\beta}}_{UDLM}) = \ell^{'}(\boldsymbol{\beta}_0) + \ell^{''}(\boldsymbol{\beta}_0)(\hat{\boldsymbol{\beta}}_{UDLM} - \boldsymbol{\beta}_0) + o_p(|\hat{\boldsymbol{\beta}}_{UDLM} - \boldsymbol{\beta}_0|) \\
\Rightarrow &\hat{\boldsymbol{\beta}}_{UDLM} - \boldsymbol{\beta}_0 = [-\ell^{''}(\boldsymbol{\beta}_0) + o_p(1)]^{-1}\ell^{'}(\boldsymbol{\beta}_0) \\
\Rightarrow &\text{Cov}(\hat{\boldsymbol{\beta}}_{UDLM}) = [-\ell^{''}(\boldsymbol{\beta}_0)]^{-1}\text{Cov}(\ell^{'}(\boldsymbol{\beta}_0))[-\ell^{''}(\boldsymbol{\beta}_0)]^{-1} + o_p(1) \\
\Rightarrow &[-\ell^{''}(\boldsymbol{\beta}_0)]^{-1}\text{Cov}(\ell^{'}(\boldsymbol{\beta}_0))[-\ell^{''}(\boldsymbol{\beta}_0)]^{-1} \xrightarrow{P} \text{Cov}(\hat{\boldsymbol{\beta}}_{UDLM})
\end{aligned}
$$

Since $\hat{\boldsymbol{\beta}}_{UDLM} \to \boldsymbol{\beta}_0$, $-H_U^{-1}(\hat{\boldsymbol{\beta}}_{UDLM})^{-1} \xrightarrow{P} [-\ell^{''}(\boldsymbol{\beta}_0)]^{-1}$. Also, $\text{Cov}(\ell^{'}(\boldsymbol{\beta}_0))$ can be consistently estimated by empirical variance $\sum_{t=1}^{\top} S_U^{(t)}(\hat{\boldsymbol{\beta}}_{UDLM})S_U^{(t)}(\hat{\boldsymbol{\beta}}_{UDLM})^{\top}$. So,

$$
\text{Cov}(\hat{\boldsymbol{\beta}}) = H_U^{-1}(\hat{\boldsymbol{\beta}})[\sum_{t=1}^{\top} S_U^{(t)}(\hat{\boldsymbol{\beta}})S_U^{(t)}(\hat{\boldsymbol{\beta}})^{\top}]H_U^{-1}(\hat{\boldsymbol{\beta}})
$$

Similarly,

$$
\begin{aligned}
\text{Cov}(\hat{\boldsymbol{\theta}}) &= H_C^{-1}(\hat{\boldsymbol{\theta}})[\sum_{t=1}^{\top} S_C^{(t)}(\hat{\boldsymbol{\theta}})S_C^{(t)}(\hat{\boldsymbol{\theta}})^{\top}]H_C^{-1}(\hat{\boldsymbol{\theta}}) \\
\text{Cov}(\hat{\boldsymbol{\beta}}_{UDLM}, \hat{\boldsymbol{\theta}}) &= H_U^{-1}(\hat{\boldsymbol{\beta}}_{UDLM})[\sum_{t=1}^{\top} S_U^{(t)}(\hat{\boldsymbol{\beta}}_{UDLM})S_C^{(t)}(\hat{\boldsymbol{\theta}})^{\top}]H_C^{-1}(\hat{\boldsymbol{\theta}})
\end{aligned}
$$

Therefore,

$$
\begin{aligned}
\text{Cov}(\hat{\boldsymbol{\psi}}) &= \text{Cov}(\hat{\boldsymbol{\beta}}_{UDLM} - \hat{\boldsymbol{\beta}}_{CDLM}) \\
&= \text{Cov}(\hat{\boldsymbol{\beta}}_{UDLM} - \boldsymbol{C}\hat{\boldsymbol{\theta}}) \\
&= \text{Cov}(\hat{\boldsymbol{\beta}}_{UDLM}) - 2\text{Cov}(\hat{\boldsymbol{\beta}}_{UDLM}, \hat{\boldsymbol{\theta}})\boldsymbol{C}^{\top} + \boldsymbol{C}\text{Cov}(\hat{\boldsymbol{\theta}})\boldsymbol{C}^{\top} \\
&= H_U^{-1}(\hat{\boldsymbol{\beta}}_{UDLM})[\sum_{t=1}^{\top} S_U^{(t)}(\hat{\boldsymbol{\beta}}_{UDLM})S_U^{(t)}(\hat{\boldsymbol{\beta}}_{UDLM})^{\top}]H_U^{-1}(\hat{\boldsymbol{\beta}}_{UDLM}) \\
&- 2H_U^{-1}(\hat{\boldsymbol{\beta}}_{UDLM})[\sum_{t=1}^{\top} S_U^{(t)}(\hat{\boldsymbol{\beta}}_{UDLM})S_C^{(t)}(\hat{\boldsymbol{\theta}})^{\top}]H_C^{-1}(\hat{\boldsymbol{\theta}})\boldsymbol{C}^{\top} \\
&+ \boldsymbol{C}H_C^{-1}(\hat{\boldsymbol{\theta}})[\sum_{t=1}^{\top} S_C^{(t)}(\hat{\boldsymbol{\theta}})S_C^{(t)}(\hat{\boldsymbol{\theta}})^{\top}]H_C^{-1}(\hat{\boldsymbol{\theta}})\boldsymbol{C}^{\top}
\end{aligned}
$$

Let $\boldsymbol{\beta}_{CDLM}$, $\boldsymbol{\beta}_{EB1}$, and $\boldsymbol{\psi}$ be the asymptotic limit of constrained DLM estimator, EB1

estimator, and bias, respectively. We have

$$\boldsymbol{\beta}_{EB1} = \boldsymbol{\beta}_0 + \boldsymbol{K}\boldsymbol{\psi}$$

where $\boldsymbol{K} = (\boldsymbol{V} \circ \boldsymbol{I}_{L+1})[(\boldsymbol{V} + \boldsymbol{\psi}\boldsymbol{\psi}^\top) \circ \boldsymbol{I}_{L+1}]^{-1}$. When $\boldsymbol{\beta}_0 \neq \boldsymbol{\beta}_{CDLM}$ ($\boldsymbol{\psi} \neq \boldsymbol{0}$), we can use first-order Taylor expansion of $\hat{\boldsymbol{\beta}}_{EB1}$ at $(\boldsymbol{\beta}_0^\top, \boldsymbol{\beta}_{CDLM}^\top)^\top$ and the fact that $\boldsymbol{V} = O_p(N^{-1})$ to obtain

$$\sqrt{N}(\hat{\boldsymbol{\beta}}_{EB1} - \boldsymbol{\beta}_{EB1}) = \boldsymbol{G} \times \sqrt{N}[(\hat{\boldsymbol{\beta}}_{UDLM}^\top, \hat{\boldsymbol{\beta}}_{CDLM}^\top)^\top - (\boldsymbol{\beta}_0^\top, \boldsymbol{\beta}_{CDLM}^\top)^\top] + o_p(1)$$

where

$$\boldsymbol{G} = [\text{diag`}\{\frac{v_j(v_j - \psi_j^2)}{(v_j + \psi_j^2)^2}\}, \boldsymbol{I}_{L+1} - \text{diag}\{\frac{v_j(v_j - \psi_j^2)}{(v_j + \psi_j^2)^2}\}]$$

where v_js are the diagonal elements of $\boldsymbol{V}$ and ψ_js are the elements of $\boldsymbol{\psi}$. Thus, $\hat{\boldsymbol{\beta}}_{EB1}$ is $\sqrt{N}$-consistent and asymptotically normal when $\boldsymbol{\beta}_0 \neq \boldsymbol{\beta}_{CDLM}$. Let $\hat{\boldsymbol{\Sigma}}$ denote the estimated variance-covariance matrix of $(\hat{\boldsymbol{\beta}}_{UDLM}^\top, \hat{\boldsymbol{\beta}}_{CDLM}^\top)^\top$. With plug-in estimate of $\boldsymbol{G}$, the asymptotic variance of $\hat{\boldsymbol{\beta}}_{EB1}$ can be estimated as $\hat{\boldsymbol{G}}\hat{\boldsymbol{\Sigma}}\hat{\boldsymbol{G}}^\top$.

2.7.3 Equivalence of $(p-1)$-degree Polynomial DLM Estimator and GRR/HB Shrinkage Target Corresponding to R_{p-1}

The general form of $(p-1)$-degree polynomial distributed lag function is $\beta(\ell) = \sum_{i=0}^{p-1} a_i \ell^i = \boldsymbol{C}\boldsymbol{\theta}$ for $\ell = 0, \cdots, L$ where $\boldsymbol{C}$ is a $(L+1) \times p$ transformation matrix with element $(\ell+1, j)$ equal to $\ell^{(j-1)}$ and $\boldsymbol{\theta} = (a_1, \cdots, a_p)^\top$. Let $\boldsymbol{R}_{p-1}$ be the $(p-1)$-degree polynomial constraint matrix. We first show that $\boldsymbol{R}_{p-1}\boldsymbol{C} = \boldsymbol{0}$.

The corresponding constraints constructed in $\boldsymbol{R}_{p-1}$ are $\sum_{j=0}^{p}(-1)^j \binom{p}{j} \beta(\ell + j) = 0$ for $\ell = 0, ..., L-p$. Showing $\boldsymbol{R}_{p-1}\boldsymbol{C} = \boldsymbol{0}$ is the same as showing $\sum_{j=0}^{p}(-1)^j \binom{p}{j} [\sum_{i=1}^{p} a_i(\ell +$

$j)^{i-1}] = 0$ for $\ell = 0, ..., L - p$.

$$\begin{aligned}
&\sum_{j=0}^{p}(-1)^j\binom{p}{j}[\sum_{i=1}^{p}a_i(\ell+j)^{i-1}] \\
=&\sum_{i=1}^{p}\sum_{j=0}^{p}(-1)^j\binom{p}{j}a_i(\ell+j)^{i-1} \\
=&\sum_{i=1}^{p}a_i[\sum_{j=0}^{p}(-1)^j\binom{p}{j}(\ell+j)^{i-1}]
\end{aligned}$$

It is sufficient to show that $\sum_{j=0}^{p}(-1)^j\binom{p}{j}(\ell+j)^{i-1} = 0$ for $\ell = 0, ..., L-p$ and $1 \leq i \leq p$. It is well-known that each polynomial can be uniquely expressed as a linear combination of binomial coefficients. $\sum_{j=0}^{p}(-1)^j\binom{p}{j}(\ell+j)^{i-1} = 0$ corresponds to the binomial coefficient involved $(p-1)$-degree term of the characteristic polynomial $(\ell+j)^{i-1}$. We know that i is at most p so the coefficients of all the terms of degree larger than $p-1$ must be zero so we have $\sum_{j=0}^{p}(-1)^j\binom{p}{j}(\ell+j)^{i-1} = 0$ for $\ell = 0, ..., L-p$ and $1 \leq i \leq p$. Therefore, $\boldsymbol{R}_{p-1}\boldsymbol{C} = \boldsymbol{0}$.

The shrinkage target of GRR/HB estimator corresponding to $\boldsymbol{R}_{p-1}$ is the maximizer of likelihood function in (1) of main text subject to the constraint $\boldsymbol{R}_{p-1}\boldsymbol{\beta} = \boldsymbol{0}$. Let $\hat{\boldsymbol{\beta}}$ denote the GRR/HB estimator. Since $\hat{\boldsymbol{\beta}}$ conforms to the constraint, we have $\boldsymbol{R}_{p-1}\hat{\boldsymbol{\beta}} = \boldsymbol{0}$ and $\hat{\boldsymbol{\beta}}$ is an element in the kernel of $\boldsymbol{R}_{p-1}$. From above, we have $\boldsymbol{R}_{p-1}\boldsymbol{C} = \boldsymbol{0}$ and we know that the p columns of $\boldsymbol{C}$ are linearly independent. Thus, the kernel of $\boldsymbol{R}_{p-1}$ is spanned by the p columns of $\boldsymbol{C}$. Subsequently, every element in the kernel can be expressed as $\boldsymbol{C\theta}$ so $\hat{\boldsymbol{\beta}}$ must be in the form of $\boldsymbol{C\theta}$. Therefore, the maximizing the likelihood function in (1) in terms of $\boldsymbol{\beta}$ subject to the constraint $\boldsymbol{R}_{p-1}\boldsymbol{\beta} = \boldsymbol{0}$ is equivalent to maximizing the likelihood function in (2) in terms of $\boldsymbol{\theta}$ without any constraint. We then conclude that $(p-1)$-degree polynomial DLM estimator and GRR/HB shrinkage target corresponding to $\boldsymbol{R}_{p-1}$ are equivalent.

2.7.4 Conditional Distributions of HB Estimator and Two-stage Shrinkage Estimator

The full conditional distributions of σ_π^2 and $\boldsymbol{\beta}$ for HB estimator are given by

$$f(\sigma_\pi^2|\boldsymbol{\beta},\boldsymbol{Y}) \propto IG(a_\pi + M/2, b_\pi + \boldsymbol{\beta}^\top \boldsymbol{R}^\top \boldsymbol{R}\boldsymbol{\beta}/2)$$
$$f(\boldsymbol{\beta}|\sigma_\pi^2,\boldsymbol{Y}) \propto \prod_{t=1}^{\top}[\exp(y_t \boldsymbol{X}_t^\top \boldsymbol{\beta} - e^{\boldsymbol{X}_t^\top \boldsymbol{\beta}})] \cdot \exp(-\frac{\boldsymbol{\beta}^\top \boldsymbol{R}^\top \boldsymbol{R}\boldsymbol{\beta}}{2\sigma_\pi^2}).$$

For two-stage shrinkage approach, if we let $\boldsymbol{\omega} = (\omega_1, \omega_2)^\top$ have a discrete uniform prior distribution, the full conditional distributions of $\boldsymbol{\beta}, \sigma^2$, and $\boldsymbol{\omega}$ are given by:

$$f(\boldsymbol{\beta}|\hat{\boldsymbol{\beta}}_{EB},\boldsymbol{\omega},\sigma^2) \sim N([1/\sigma^2\boldsymbol{\Omega}(\boldsymbol{\omega})^{-1} + (\boldsymbol{G}\boldsymbol{\Sigma}\boldsymbol{G}^\top)^{-1}]^{-1}(\boldsymbol{G}\boldsymbol{\Sigma}\boldsymbol{G}^\top)^{-1}\hat{\boldsymbol{\beta}}_{EB}, [1/\sigma^2\boldsymbol{\Omega}(\boldsymbol{\omega})^{-1} + (\boldsymbol{G}\boldsymbol{\Sigma}\boldsymbol{G})^{-1}]^{-1})$$
$$p(\boldsymbol{\omega}|\hat{\boldsymbol{\beta}}_{EB},\boldsymbol{\beta},\sigma^2) = \frac{|\boldsymbol{\Omega}(\boldsymbol{\omega})|^{-1/2}\exp[-\frac{1}{2\sigma^2}\boldsymbol{\beta}^\top\boldsymbol{\Omega}(\boldsymbol{\omega})^{-1}\boldsymbol{\beta}]}{\sum_{\boldsymbol{\omega}^*}|\boldsymbol{\Omega}(\boldsymbol{\omega}^*)|^{-1/2}\exp[-\frac{1}{2\sigma^2}\boldsymbol{\beta}^\top\boldsymbol{\Omega}(\boldsymbol{\omega}^*)^{-1}\boldsymbol{\beta}]}$$
$$f(\sigma^2|\hat{\boldsymbol{\beta}}_{EB},\boldsymbol{\beta},\boldsymbol{\omega}) \sim IG(a_0 + (L+1)/2, b_0 + \boldsymbol{\beta}^T\boldsymbol{\Omega}(\boldsymbol{\omega})^{-1}\boldsymbol{\beta}/2).$$

2.7.5 Analytical Results for the GRR Estimator

GRR estimator $\hat{\boldsymbol{\beta}}_{GRR}$ is given by

$$\hat{\boldsymbol{\beta}}_{GRR} = \arg\min_{\boldsymbol{\beta}}[-\sum_{t=1}^{T}[y_t\boldsymbol{\beta}^\top\boldsymbol{X_t} - e^{\boldsymbol{\beta}^\top\boldsymbol{X_t}} - \log(y_t!)] + \lambda\boldsymbol{\beta}^\top\boldsymbol{R}^\top\boldsymbol{R}\boldsymbol{\beta}]$$

and its asymptotic MSE $E[(\hat{\boldsymbol{\beta}}_{GRR}-\boldsymbol{\beta})^\top(\hat{\boldsymbol{\beta}}_{GRR}-\boldsymbol{\beta})]$ can be decomposed into $f_1(\lambda)+f_2(\lambda)$ where

$$f_1(\lambda) = E[(\hat{\boldsymbol{\beta}}_{UDLM}-\boldsymbol{\beta})^\top\boldsymbol{H}^\top\boldsymbol{H}(\hat{\boldsymbol{\beta}}_{UDLM}-\boldsymbol{\beta})] = \text{trace}[(\boldsymbol{X}^\top\boldsymbol{W}\boldsymbol{X})(\boldsymbol{X}^\top\boldsymbol{W}\boldsymbol{X}+\lambda\boldsymbol{R}^\top\boldsymbol{R})^{-2}]$$

$$f_2(\lambda) = (\boldsymbol{H}\boldsymbol{\beta}-\boldsymbol{\beta})^\top(\boldsymbol{H}\boldsymbol{\beta}-\boldsymbol{\beta}) = \lambda^2\boldsymbol{\beta}^\top(\boldsymbol{R}^\top\boldsymbol{R})(\boldsymbol{X}^\top\boldsymbol{W}\boldsymbol{X}+\lambda\boldsymbol{R}^\top\boldsymbol{R})^{-2}(\boldsymbol{R}^\top\boldsymbol{R})\boldsymbol{\beta}.$$

We first show that $f_1(\lambda)$ is monotonic decreasing and $f_2(\lambda)$ is monotonic increasing and then we show that $f_1(\lambda) + f_2(\lambda)$ is convex.

$$
\begin{aligned}
& df_1(\lambda) \\
=& d\text{trace}[(\boldsymbol{X}^\top \hat{\boldsymbol{W}} \boldsymbol{X})(\boldsymbol{X}^\top \hat{\boldsymbol{W}} \boldsymbol{X} + \lambda \boldsymbol{R}^\top \boldsymbol{R})^{-2}] \\
=& \text{trace}[2(\boldsymbol{X}^\top \hat{\boldsymbol{W}} \boldsymbol{X} + \lambda \boldsymbol{R}^\top \boldsymbol{R})^{-1}(\boldsymbol{X}^\top \hat{\boldsymbol{W}} \boldsymbol{X}) d(\boldsymbol{X}^\top \hat{\boldsymbol{W}} \boldsymbol{X} + \lambda \boldsymbol{R}^\top \boldsymbol{R})^{-1}] \\
=& \text{trace}\{-2(\boldsymbol{X}^\top \hat{\boldsymbol{W}} \boldsymbol{X} + \lambda \boldsymbol{R}^\top \boldsymbol{R})^{-1}(\boldsymbol{X}^\top \hat{\boldsymbol{W}} \boldsymbol{X})(\boldsymbol{X}^\top \hat{\boldsymbol{W}} \boldsymbol{X} + \lambda \boldsymbol{R}^\top \boldsymbol{R})^{-1}[d(\boldsymbol{X}^\top \hat{\boldsymbol{W}} \boldsymbol{X} + \lambda \boldsymbol{R}^\top \boldsymbol{R})] \\
& \quad (\boldsymbol{X}^\top \hat{\boldsymbol{W}} \boldsymbol{X})(\boldsymbol{X}^\top \hat{\boldsymbol{W}} \boldsymbol{X} + \lambda \boldsymbol{R}^\top \boldsymbol{R})^{-1}\} \\
=& \text{trace}[-2\boldsymbol{R}^\top \boldsymbol{R}(\boldsymbol{X}^\top \hat{\boldsymbol{W}} \boldsymbol{X})(\boldsymbol{X}^\top \hat{\boldsymbol{W}} \boldsymbol{X} + \lambda \boldsymbol{R}^\top \boldsymbol{R})^{-3}] d\lambda
\end{aligned}
$$

Since $\boldsymbol{X}^\top \hat{\boldsymbol{W}} \boldsymbol{X}$ and $\boldsymbol{R}^\top \boldsymbol{R}$ are positive definite and $\lambda > 0$, $\boldsymbol{X}^\top \hat{\boldsymbol{W}} \boldsymbol{X} + \lambda \boldsymbol{R}^\top \boldsymbol{R}$ is positive definite (Weyl's inequality). It follows that $2\boldsymbol{R}^\top \boldsymbol{R}(\boldsymbol{X}^\top \hat{\boldsymbol{W}} \boldsymbol{X})(\boldsymbol{X}^\top \hat{\boldsymbol{W}} \boldsymbol{X} + \lambda \boldsymbol{R}^\top \boldsymbol{R})^{-3}$ is positive definite and $\text{trace}[-2\boldsymbol{R}^\top \boldsymbol{R}(\boldsymbol{X}^\top \hat{\boldsymbol{W}} \boldsymbol{X})(\boldsymbol{X}^\top \hat{\boldsymbol{W}} \boldsymbol{X} + \lambda \boldsymbol{R}^\top \boldsymbol{R})^{-3}] < 0$. Therefore, we have shown that $f_1(\lambda)$ is monotonic decreasing ($f_1^{'}(\lambda) < 0 \; \forall \lambda > 0$).

Assume $\lambda_2 > \lambda_1 > 0$,

$$
\begin{aligned}
& f_2(\lambda_2) - f_2(\lambda_1) \\
=& \lambda_2^2 \boldsymbol{\beta}^\top (\boldsymbol{R}^\top \boldsymbol{R})(\boldsymbol{X}^\top \hat{\boldsymbol{W}} \boldsymbol{X} + \lambda_2 \boldsymbol{R}^\top \boldsymbol{R})^{-2} (\boldsymbol{R}^\top \boldsymbol{R}) \boldsymbol{\beta} - \lambda_1^2 \boldsymbol{\beta}^\top (\boldsymbol{R}^\top \boldsymbol{R})(\boldsymbol{X}^\top \hat{\boldsymbol{W}} \boldsymbol{X} + \lambda_1 \boldsymbol{R}^\top \boldsymbol{R})^{-2} (\boldsymbol{R}^\top \boldsymbol{R}) \boldsymbol{\beta} \\
=& \boldsymbol{\beta}^\top (\boldsymbol{R}^\top \boldsymbol{R})[(\frac{1}{\lambda_2} \boldsymbol{X}^\top \hat{\boldsymbol{W}} \boldsymbol{X} + \boldsymbol{R}^\top \boldsymbol{R})^{-2} - (\frac{1}{\lambda_1} \boldsymbol{X}^\top \hat{\boldsymbol{W}} \boldsymbol{X} + \boldsymbol{R}^\top \boldsymbol{R})^{-2}](\boldsymbol{R}^\top \boldsymbol{R}) \boldsymbol{\beta} \\
=& \boldsymbol{\beta}^\top (\boldsymbol{R}^\top \boldsymbol{R})(\frac{1}{\lambda_2} \boldsymbol{X}^\top \hat{\boldsymbol{W}} \boldsymbol{X} + \boldsymbol{R}^\top \boldsymbol{R})^{-2} [(\frac{1}{\lambda_1} \boldsymbol{X}^\top \hat{\boldsymbol{W}} \boldsymbol{X} + \boldsymbol{R}^\top \boldsymbol{R})^2 - (\frac{1}{\lambda_2} \boldsymbol{X}^\top \hat{\boldsymbol{W}} \boldsymbol{X} + \boldsymbol{R}^\top \boldsymbol{R})^2] \\
& (\frac{1}{\lambda_1} \boldsymbol{X}^\top \hat{\boldsymbol{W}} \boldsymbol{X} + \boldsymbol{R}^\top \boldsymbol{R})^{-2} (\boldsymbol{R}^\top \boldsymbol{R}) \boldsymbol{\beta} \\
=& \boldsymbol{\beta}^\top (\boldsymbol{R}^\top \boldsymbol{R})(\frac{1}{\lambda_2} \boldsymbol{X}^\top \hat{\boldsymbol{W}} \boldsymbol{X} + \boldsymbol{R}^\top \boldsymbol{R})^{-2} [(\frac{1}{\lambda_1^2} - \frac{1}{\lambda_2^2})(\boldsymbol{X}^\top \hat{\boldsymbol{W}} \boldsymbol{X})^2 + (\frac{1}{\lambda_1} - \frac{1}{\lambda_2})(\boldsymbol{X}^\top \hat{\boldsymbol{W}} \boldsymbol{X})(\boldsymbol{R}^\top \boldsymbol{R}))] \\
& (\frac{1}{\lambda_1} \boldsymbol{X}^\top \hat{\boldsymbol{W}} \boldsymbol{X} + \boldsymbol{R}^\top \boldsymbol{R})^{-2} (\boldsymbol{R}^\top \boldsymbol{R}) \boldsymbol{\beta} \\
=& \text{trace}\{\boldsymbol{\beta} \boldsymbol{\beta}^\top (\boldsymbol{R}^\top \boldsymbol{R})^2 (\frac{1}{\lambda_2} \boldsymbol{X}^\top \hat{\boldsymbol{W}} \boldsymbol{X} + \boldsymbol{R}^\top \boldsymbol{R})^{-2} [(\frac{1}{\lambda_1^2} - \frac{1}{\lambda_2^2})(\boldsymbol{X}^\top \hat{\boldsymbol{W}} \boldsymbol{X})^2 \\
& + (\frac{1}{\lambda_1} - \frac{1}{\lambda_2})(\boldsymbol{X}^\top \hat{\boldsymbol{W}} \boldsymbol{X})(\boldsymbol{R}^\top \boldsymbol{R}))] (\frac{1}{\lambda_1} \boldsymbol{X}^\top \hat{\boldsymbol{W}} \boldsymbol{X} + \boldsymbol{R}^\top \boldsymbol{R})^{-2}\} \\
=& \text{trace}(\boldsymbol{A}) \\
=& \sum_{\ell=1}^{L+1} \alpha_\ell
\end{aligned}
$$

where γ_ℓ is the ℓ^{th} eigenvalue of $\boldsymbol{B}$. Since $\boldsymbol{\beta}\boldsymbol{\beta}^\top$, $\boldsymbol{X}^\top \hat{\boldsymbol{W}} \boldsymbol{X}$, and $\boldsymbol{R}^\top \boldsymbol{R}$ are positive definite and $\lambda_2 > \lambda_1$, all of the terms that $\boldsymbol{A}$ is composed of are positive definite and so is $\boldsymbol{A}$. Hence, $f_2(\lambda_2) - f_2(\lambda_1) = \sum_{\ell=1}^{L+1} \alpha_\ell > 0$. Therefore, we have shown that $f_2(\lambda)$ is monotonic increasing.

$$
\begin{aligned}
& f_2^{'}(\lambda) \\
=& 2\lambda \boldsymbol{\beta}^\top (\boldsymbol{R}^\top \boldsymbol{R})(\boldsymbol{X}^\top \hat{\boldsymbol{W}} \boldsymbol{X} + \lambda \boldsymbol{R}^\top \boldsymbol{R})^{-2} (\boldsymbol{R}^\top \boldsymbol{R}) \boldsymbol{\beta} - \lambda^2 \text{trace}[2(\boldsymbol{R}^\top \boldsymbol{R})^3 \boldsymbol{\beta} \boldsymbol{\beta}^\top (\boldsymbol{X}^\top \hat{\boldsymbol{W}} \boldsymbol{X} + \lambda \boldsymbol{R}^\top \boldsymbol{R})^{-3}]
\end{aligned}
$$

$$
\begin{aligned}
& f_1''(\lambda) + f_2''(\lambda) \\
=& \frac{d\text{trace}[-2\boldsymbol{R}^\top\boldsymbol{R}(\boldsymbol{X}^\top\hat{\boldsymbol{W}}\boldsymbol{X})(\boldsymbol{X}^\top\hat{\boldsymbol{W}}\boldsymbol{X}+\lambda\boldsymbol{R}^\top\boldsymbol{R})^{-3}]}{d\lambda} \\
& + 2\lambda\boldsymbol{\beta}^\top(\boldsymbol{R}^\top\boldsymbol{R})(\boldsymbol{X}^\top\hat{\boldsymbol{W}}\boldsymbol{X}+\lambda\boldsymbol{R}^\top\boldsymbol{R})^{-2}(\boldsymbol{R}^\top\boldsymbol{R})\boldsymbol{\beta} \\
& + 2\lambda\frac{d[\boldsymbol{\beta}^\top(\boldsymbol{R}^\top\boldsymbol{R})(\boldsymbol{X}^\top\hat{\boldsymbol{W}}\boldsymbol{X}+\lambda\boldsymbol{R}^\top\boldsymbol{R})^{-2}(\boldsymbol{R}^\top\boldsymbol{R})\boldsymbol{\beta}]}{d\lambda} \\
& - 2\lambda\text{trace}[2(\boldsymbol{R}^\top\boldsymbol{R})^3\boldsymbol{\beta}\boldsymbol{\beta}^\top(\boldsymbol{X}^\top\hat{\boldsymbol{W}}\boldsymbol{X}+\lambda\boldsymbol{R}^\top\boldsymbol{R})^{-3}] \\
& - \lambda^2\frac{d[2(\boldsymbol{R}^\top\boldsymbol{R})^3\boldsymbol{\beta}\boldsymbol{\beta}^\top(\boldsymbol{X}^\top\hat{\boldsymbol{W}}\boldsymbol{X}+\lambda\boldsymbol{R}^\top\boldsymbol{R})^{-3}]}{d\lambda} \\
=& \text{trace}[6(\boldsymbol{R}^\top\boldsymbol{R})^2(\boldsymbol{X}^\top\hat{\boldsymbol{W}}\boldsymbol{X})(\boldsymbol{X}^\top\hat{\boldsymbol{W}}\boldsymbol{X}+\lambda\boldsymbol{R}^\top\boldsymbol{R})^{-4}] \\
& + 2\lambda\boldsymbol{\beta}^\top(\boldsymbol{R}^\top\boldsymbol{R})(\boldsymbol{X}^\top\hat{\boldsymbol{W}}\boldsymbol{X}+\lambda\boldsymbol{R}^\top\boldsymbol{R})^{-2}(\boldsymbol{R}^\top\boldsymbol{R})\boldsymbol{\beta} \\
& + 2\lambda\text{trace}[2(\boldsymbol{R}^\top\boldsymbol{R})^3\boldsymbol{\beta}\boldsymbol{\beta}^\top(\boldsymbol{X}^\top\hat{\boldsymbol{W}}\boldsymbol{X}+\lambda\boldsymbol{R}^\top\boldsymbol{R})^{-3}] \\
& - 2\lambda\text{trace}[2(\boldsymbol{R}^\top\boldsymbol{R})^3\boldsymbol{\beta}\boldsymbol{\beta}^\top(\boldsymbol{X}^\top\hat{\boldsymbol{W}}\boldsymbol{X}+\lambda\boldsymbol{R}^\top\boldsymbol{R})^{-3}] \\
& + \lambda^2\text{trace}[6(\boldsymbol{R}^\top\boldsymbol{R})^4\boldsymbol{\beta}\boldsymbol{\beta}^\top(\boldsymbol{X}^\top\hat{\boldsymbol{W}}\boldsymbol{X}+\lambda\boldsymbol{R}^\top\boldsymbol{R})^{-4}] \\
=& \text{trace}[6(\boldsymbol{R}^\top\boldsymbol{R})^2(\boldsymbol{X}^\top\hat{\boldsymbol{W}}\boldsymbol{X})(\boldsymbol{X}^\top\hat{\boldsymbol{W}}\boldsymbol{X}+\lambda\boldsymbol{R}^\top\boldsymbol{R})^{-4}] \\
& + 2\lambda\boldsymbol{\beta}^\top(\boldsymbol{R}^\top\boldsymbol{R})(\boldsymbol{X}^\top\hat{\boldsymbol{W}}\boldsymbol{X}+\lambda\boldsymbol{R}^\top\boldsymbol{R})^{-2}(\boldsymbol{R}^\top\boldsymbol{R})\boldsymbol{\beta} \\
& + \lambda^2\text{trace}[6(\boldsymbol{R}^\top\boldsymbol{R})^4\boldsymbol{\beta}\boldsymbol{\beta}^\top(\boldsymbol{X}^\top\hat{\boldsymbol{W}}\boldsymbol{X}+\lambda\boldsymbol{R}^\top\boldsymbol{R})^{-4}] > 0
\end{aligned}
$$

Therefore, we have shown that $f_1(\lambda) + f_2(\lambda)$ is convex.

Table 2.1: Squared bias (in the unit of 10^{-3}), variance (in the unit of 10^{-3}), relative efficiency measured with respect to the variance of the UDLM estimator, and distance. Distances are the average Euclidean distance between the vector of lag coefficient estimates and the vector of the true coefficients (i.e. $||\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}||_2$). Results for distributed lag (DL) function estimation (upper) and results for total effect estimation (lower) are averaged across 1000 simulation repetitions. Best performers in each row are in bold.

DL Function Estimation		UDLM	CDLM	EB1	EB2	GRR	GADLM	BDLM	HB
	Squared Bias	0.02	**0.00**	0.01	0.01	**0.00**	**0.00**	0.51	**0.00**
(1) Working DL Function*	Variance	19.49	**2.31**	9.80	10.56	3.62	4.15	11.13	4.32
Completely Matches	Relative Efficiency	1.00	**8.43**	1.99	1.85	5.38	4.70	1.68	4.52
True DL Function	Distance	0.14	**0.05**	0.09	0.10	0.05	0.06	0.11	0.06
	Squared Bias	**0.02**	7.53	0.74	0.62	1.02	1.21	0.57	0.96
(2) Working DL Function*	Variance	20.03	**1.36**	11.64	12.20	4.64	5.50	8.02	3.79
Moderately Departs from	Relative Efficiency	1.00	2.26	1.62	1.56	3.54	2.99	2.33	**4.22**
True DL Function	Distance	0.14	0.09	0.11	0.11	**0.07**	0.08	0.09	**0.07**
	Squared Bias	**0.02**	27.59	1.68	1.41	7.27	17.68	6.29	6.15
(3) Non-smooth True	Variance	20.23	**1.36**	15.50	15.95	10.38	9.62	8.95	8.65
DL Function	Relative Efficiency	1.00	0.70	1.18	1.17	1.15	0.72	1.33	1.37
	Distance	0.14	0.17	0.13	0.13	0.13	0.16	0.12	**0.12**
(4) Multiple True	Squared Bias	**0.02**	1.19	0.17	0.15	0.40	0.36	0.34	0.26
DL Functions	Relative Efficiency	1.00	1.54	1.53	1.42	2.09	1.79	1.77	**2.26**
Total Effect Estimation		UDLM	CDLM	EB1	EB2	GRR	GADLM	BDLM	HB
(1) Working DL Function*	Squared Bias	0.01	**0.00**	0.02	0.02	0.01	0.00	0.19	0.01
Completely Matches	Variance	3.31	3.26	3.74	3.76	3.29	3.35	**3.20**	3.31
True DL Function	Relative Efficiency	1.00	**1.02**	0.88	0.88	1.01	0.99	0.98	1.00
(2) Working DL Function*	Squared Bias	**0.01**	0.05	0.03	0.02	0.01	0.01	0.01	0.01
Moderately Departs from	Variance	3.29	3.26	4.43	4.43	3.24	**3.15**	3.18	3.25
True DL Function	Relative Efficiency	1.00	1.00	0.74	0.74	1.01	**1.04**	1.03	1.01
	Squared Bias	**0.00**	0.00	0.04	0.03	0.00	0.02	0.04	0.00
(3) Non-smooth True	Variance	3.04	2.99	3.55	3.53	3.00	3.08	**2.90**	3.01
DL Function	Relative Efficiency	1.00	1.02	0.85	0.85	1.02	0.99	**1.04**	1.01

*The working distributed lag (DL) function in CDLM for CDLM, EB1, EB2, GRR, and HB.

Table 2.2: Squared bias (in the unit of 10^{-3}), variance (in the unit of 10^{-3}), relative efficiency measured with respect to the variance of UDLM estimator, and distance of the vector of the distributed lag coefficient estimates obtained from seven statistical methods under the scenario that maximum lag (ℓ) is excessively specified. Distances are the average Euclidean distance between the vector of lag coefficient estimates and the vector of the true coefficients (i.e. $||\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}||_2$) across 1000 simulation repetitions. Best performers in each row are in bold.

	EB1	TSB-EB1	TSP-EB1	GRR	TSB-GRR	TSP-GRR	BDLM
Squared Bias	2.22	1.62	1.22	0.83	2.13	1.43	**0.68**
Variance	66.16	62.39	61.77	18.26	**9.79**	10.91	35.80
Efficiency	1.83	1.95	1.98	6.54	**10.47**	10.13	3.42
Distance	0.25	0.24	0.24	0.13	**0.10**	**0.10**	0.18

Table 2.3: Summaries to various distributed lag model (DLM) estimators.

Methods	One-stage	Two-stage	Descriptions
UDLM (Unconstrained Distributed Lag Model)	×		Distributed lag function estimated without constraints
CDLM (Constrained Distributed Lag Model)	×		Distributed lag function generated from a set of parametric functions
EB1/EB2 (Empirical Bayes)	×		Multi-variate weighted average of UDLM and CDLM
GRR (Generalized Ridge Regression)	×		UDLM with penalization on departure of CDLM
HB (Hierarchical Bayes)	×		Hierarchical Bayesian model with shrinkage toward CDLM
BDLM (Bayesian Distributed Lag Model)	×		Smoothing distributed lag function via prior variance-covariance specifications
GADLM (Generalized Additive DLM)	×		Generalized additive distributed lag model with regression splines
TSB-EB1 (Two-stage Bayesian)		×	EB1 estimator in stage 1 and BDLM-type Bayesian model in stage 2
TSP-EB1 (Two-stage Penalized)		×	EB1 estimator in stage 1 and BDLM-type penalization in stage 2
TSB-GRR (Two-stage Bayesian)		×	GRR estimator in stage 1 and BDLM-type Bayesian model in stage 2
TSP-GRR (Two-stage Penalized)		×	GRR estimator in stage 1 and BDLM-type penalization in stage 2

Table 2.4: Summary of the three simulation scenarios for comparing UDLM, CDLM, EB1, EB2, GRR, GADLM, BDLM, and HB in simulation study 1.

Scenario	Working DL Function*	True Distributed Lag Coefficients
(1) Working DL Function Completely Matches True DL Function	Cubic	$\beta_j = (j^3 - 17j^2 + 70j)/400$ for $j = 0, ..., 10$
(2) Working DL Function Moderately Departs from True DL Function	Quadratic	Slight Departure from $\beta_j = (-0.7j^2 + 2.3j + 50.8)/400$ for $j = 0, ..., 10$
(3) Non-smooth True DL Function	Quadratic	Oscillating between 0.02 and 0.18
(4) Mixture of 5 True DL Function	Cubic	(a) $\beta_j = 0$ for $j = 0, \cdots, 10$ (b) $\beta_j = 0.014(10 - j)$ for $j = 0, \cdots, 10$ (c) Same as (1) (d) Same as (2) (e) Same as (3)

*The working distributed lag (DL) function in CDLM for CDLM, EB1, EB2, GRR, and HB.

Table 2.5: Metrics used for evaluating the estimation precision in simulation study 1.

Metric	Lag Effects Vector ($\boldsymbol{\beta}$)	Total Effect ($\sum_{j=0}^{10} \beta_j$)
Squared bias	$(\hat{\bar{\boldsymbol{\beta}}} - \boldsymbol{\beta})^\top (\hat{\bar{\boldsymbol{\beta}}} - \boldsymbol{\beta})$	$[\sum_{j=0}^{10} (\hat{\bar{\beta}}_j - \beta_j)]^2$
Variance	$\text{trace}[\frac{1}{1000} \sum_{i=1}^{1000} (\hat{\boldsymbol{\beta}}_i - \hat{\bar{\boldsymbol{\beta}}})(\hat{\boldsymbol{\beta}}_i - \hat{\bar{\boldsymbol{\beta}}})^\top]$	$\frac{1}{1000} \sum_{i=1}^{1000} (\sum_{j=0}^{10} \hat{\beta}_{ij} - \sum_{j=0}^{10} \hat{\bar{\beta}}_j)^2$
Relative Efficiency[1]	$\frac{\sum_{i=1}^{1000} \|\|\hat{\boldsymbol{\beta}}_i^{UDLM} - \boldsymbol{\beta}\|\|_2^2}{\sum_{i=1}^{1000} \|\|\hat{\boldsymbol{\beta}}_i - \boldsymbol{\beta}\|\|_2^2}$	$\frac{\sum_{i=1}^{1000} (\sum_{j=0}^{10} \hat{\beta}_{ij}^{UDLM} - \sum_{j=0}^{10} \beta_j)^2}{\sum_{i=1}^{1000} (\sum_{j=0}^{10} \hat{\beta}_{ij} - \sum_{j=0}^{10} \beta_j)^2}$
Distance[2]	$\frac{1}{1000} \sum_{i=1}^{1000} \|\|\hat{\boldsymbol{\beta}}_i - \boldsymbol{\beta}\|\|_2$	-

[1]Relative efficiency with respect to UDLM in terms of mean squared errors (MSE)
[2]Mean distance to the true coefficient vector $\boldsymbol{\beta}$
*$\hat{\boldsymbol{\beta}}_i = (\hat{\beta}_{i0}, ..., \hat{\beta}_{i10})^\top$: the estimated lag coefficients from the i^{th} data set for a particular method
**$\hat{\bar{\boldsymbol{\beta}}} = \frac{1}{1000} \sum_{i=1}^{1000} \hat{\boldsymbol{\beta}}_i$ and $\hat{\bar{\beta}}_j = \frac{1}{1000} \sum_{i=1}^{1000} \hat{\beta}_{ij}$ for $j = 0, ..., 10$.

Table 2.6: Average of the 1000 estimated variances as a percentage of the empirical variance of the 1000 estimates from 1000 repetitions for the 11 cumulative lag coefficient estimates based on GRR across the three scenarios in simulation study 1.

Percentage	Scenario 1	Scenario 2	Scenario 3
Lag 0	0.91	0.83	0.96
Lag 1	0.97	0.88	0.93
Lag 2	0.93	0.96	0.94
Lag 3	0.92	0.92	1.02
Lag 4	0.92	0.90	0.95
Lag 5	0.93	0.99	0.89
Lag 6	0.93	0.97	0.91
Lag 7	0.95	0.95	0.93
Lag 8	0.97	0.95	0.95
Lag 9	0.95	0.95	0.91
Lag 10	0.96	0.98	0.95

Table 2.7: Estimated mean and 95% confidence/credible interval of the cumulative lagged effect (% change in mortality count) up to 3, 7, and 14 days of PM_{10} (upper) and O_3 (lower) on mortality with an interquartile range increase in exposure level (PM_{10}: 21.49$\mu g/m^3$, O_3: 14.65 ppb) in Chicago, Illinois from 1987 to 2000 based on the data from the National Morbidity, Mortality, and Air Pollution Study (NMMAPS) under eight estimation methods.

PM_{10}	Up to Lag 3 (95% CI[1])	Up to Lag 7 (95% CI[1])	Up to Lag 14 (95% CI[1])
UDLM	0.75 (-0.01, 1.52)	0.32 (-0.71, 1.36)	-0.75 (-2.24, 0.75)
CDLM	0.51 (-0.22, 1.23)	0.40 (-0.61, 1.40)	-0.87 (-2.36, 0.62)
EB1	0.81 (0.07, 1.56)	0.41 (-0.54, 1.37)	-0.71 (-1.98, 0.56)
GRR	0.67 (-0.06, 1.40)	0.21 (-0.80, 1.22)	-0.74 (-2.23, 0.75)
BDLM	0.57 (-0.24, 1.43)	-0.01 (-1.23, 1.16)	-1.05 (-2.76, 0.69)
HB	0.63 (0.14, 1.12)	0.26 (-0.41, 0.94)	-0.72 (-1.60, 0.15)
HB2-GRR	0.48 (-0.21, 1.18)	0.14 (-0.84, 1.12)	-0.43 (-1.91, 1.05)
HP-GRR	0.97 (0.27, 1.67)	0.48 (-0.45, 1.41)	-0.57 (-1.78, 0.64)
O_3	Up to Lag 3 (95% CI[1])	Up to Lag 7 (95% CI[1])	Up to Lag 14 (95% CI[1])
UDLM	2.04 (0.98, 3.11)	2.63 (1.31, 3.98)	2.25 (0.53, 4.01)
CDLM	2.03 (1.07, 3.00)	2.52 (1.28, 3.77)	2.10 (0.38, 3.85)
EB1	2.09 (0.82, 3.39)	2.59 (0.97, 4.24)	2.19 (0.11, 4.32)
GRR	2.08 (1.10, 3.07)	2.59 (1.33, 3.88)	2.21 (0.48, 3.97)
BDLM	1.91 (0.93, 2.90)	2.32 (1.11, 3.56)	2.26 (0.64, 3.91)
HB	2.12 (1.18, 3.07)	2.63 (1.41, 3.87)	2.23 (0.61, 3.88)
HB2-GRR	1.94 (1.01, 2.89)	2.30 (1.10, 3.52)	2.12 (0.46, 3.80)
HP-GRR	1.83 (0.97, 2.70)	2.18 (1.06, 3.31)	2.12 (0.53, 3.73)

[1]CI refers to confidence interval for UDLM, CDLM, EB1, GRR, and HP-GRR and refers to credible interval for BDLM, HB, and HB2-GRR.

Table 2.8: Estimated mean and 95% confidence/credible intervals (in parenthesis) for the lag effects (% change in mortality count) of an interquartile range increase of PM_{10} ($21.49\mu g/m^3$) on mortality in Chicago, Illinois from 1987 to 2000 based on the data from the National Morbidity, Mortality, and Air Pollution Study (NMMAPS) under eight estimation methods.

	UDLM	CDLM	EB1	GRR	BDLM	HB	HB2-GRR	HP-GRR
Lag 0	-0.15	-0.13	-0.13	-0.17	-0.13	-0.15	-0.13	-0.15
	(-0.61, 0.31)	(-0.51, 0.26)	(-0.59, 0.34)	(-0.61, 0.27)	(-0.57, 0.31)	(-0.58, 0.28)	(-0.55, 0.29)	(-0.55, 0.25)
Lag 1	0.04	0.15	0.13	0.03	0.05	-0.02	0.10	0.04
	(-0.40, 0.48)	(-0.06, 0.35)	(-0.18, 0.43)	(-0.26, 0.33)	(-0.37, 0.46)	(-0.34, 0.31)	(-0.17, 0.37)	(-0.23, 0.31)
Lag 2	0.22	0.25	0.25	0.37	0.21	0.37	0.27	0.33
	(-0.22, 0.65)	(0.06, 0.44)	(-0.23, 0.73)	(0.14, 0.60)	(-0.19, 0.60)	(0.11, 0.63)	(0.05, 0.48)	(0.12, 0.55)
Lag 3	0.65	0.23	0.56	0.44	0.54	0.48	0.25	0.37
	(0.22, 1.08)	(0.05, 0.42)	(0.36, 0.77)	(0.22, 0.66)	(0.16, 0.92)	(0.21, 0.75)	(0.05, 0.45)	(0.18, 0.57)
Lag 4	0.22	0.15	0.15	0.19	0.13	0.21	0.07	0.12
	(-0.20, 0.64)	(-0.01, 0.31)	(-0.17, 0.48)	(-0.01, 0.40)	(-0.22, 0.49)	(-0.02, 0.43)	(-0.10, 0.24)	(-0.06, 0.29)
Lag 5	-0.27	0.03	-0.17	-0.13	-0.23	-0.17	-0.11	-0.15
	(-0.69, 0.15)	(-0.11, 0.17)	(-0.32, -0.01)	(-0.32, 0.05)	(-0.55, 0.08)	(-0.39, 0.06)	(-0.26, 0.04)	(-0.31, 0.00)
Lag 6	-0.41	-0.09	-0.31	-0.29	-0.27	-0.31	-0.17	-0.23
	(-0.83, 0.02)	(-0.24, 0.05)	(-0.46, -0.15)	(-0.47, -0.10)	(-0.56, 0.03)	(-0.52, -0.10)	(-0.32, -0.02)	(-0.37, -0.10)
Lag 7	0.03	-0.19	-0.08	-0.23	-0.11	-0.21	-0.14	-0.19
	(-0.40, 0.45)	(-0.34, -0.05)	(-0.24, 0.07)	(-0.42, -0.05)	(-0.36, 0.15)	(-0.42, 0.00)	(-0.28, 0.00)	(-0.32, -0.07)
Lag 8	-0.25	-0.26	-0.26	-0.12	-0.15	-0.10	-0.10	-0.16
	(-0.70, 0.20)	(-0.40, -0.11)	(-0.71, 0.20)	(-0.31, 0.06)	(-0.38, 0.07)	(-0.31, 0.11)	(-0.22, 0.02)	(-0.27, -0.04)
Lag 9	-0.03	-0.28	-0.14	-0.08	-0.13	-0.09	-0.09	-0.14
	(-0.48, 0.43)	(-0.41, -0.14)	(-0.28, 0.00)	(-0.27, 0.10)	(-0.33, 0.07)	(-0.32, 0.13)	(-0.20, 0.02)	(-0.25, -0.04)
Lag 10	-0.27	-0.25	-0.25	-0.14	-0.14	-0.15	-0.09	-0.14
	(-0.73, 0.18)	(-0.40, -0.11)	(-0.75, 0.24)	(-0.34, 0.06)	(-0.32, 0.04)	(-0.38, 0.07)	(-0.20, 0.02)	(-0.23, -0.04)
Lag 11	-0.12	-0.20	-0.19	-0.24	-0.13	-0.24	-0.08	-0.13
	(-0.58, 0.33)	(-0.37, -0.04)	(-0.55, 0.17)	(-0.44, -0.04)	(-0.29, 0.04)	(-0.48, 0.01)	(-0.18, 0.02)	(-0.21, -0.04)
Lag 12	-0.33	-0.13	-0.22	-0.31	-0.12	-0.31	-0.07	-0.12
	(-0.79, 0.12)	(-0.30, 0.04)	(-0.39, -0.04)	(-0.52, -0.10)	(-0.28, 0.03)	(-0.55, -0.06)	(-0.17, 0.02)	(-0.19, -0.04)
Lag 13	-0.25	-0.08	-0.14	-0.23	-0.10	-0.24	-0.07	-0.10
	(-0.70, 0.21)	(-0.25, 0.09)	(-0.33, 0.06)	(-0.50, 0.04)	(-0.24, 0.03)	(-0.55, 0.07)	(-0.16, 0.02)	(-0.17, -0.03)
Lag 14	0.18	-0.07	0.07	0.18	-0.08	0.18	-0.06	-0.09
	(-0.24, 0.60)	(-0.38, 0.25)	(-0.27, 0.42)	(-0.21, 0.56)	(-0.21, 0.06)	(-0.22, 0.58)	(-0.14, 0.02)	(-0.15, -0.03)

Table 2.9: Estimated mean and 95% confidence/credible intervals (in parenthesis) for the lag effects (% change in mortality count) of an interquartile range increase of O_3 (14.65 ppb) on mortality in Chicago, Illinois from 1987 to 2000 based on the data from the National Morbidity, Mortality, and Air Pollution Study (NMMAPS) under eight estimation methods.

	UDLM	CDLM	EB1	GRR	BDLM	HB	HB2-GRR	HP-GRR
Lag 0	0.50	0.33	0.36	0.32	0.45	0.37	0.32	0.30
	(-0.21, 1.20)	(-0.19, 0.85)	(-0.26, 0.98)	(-0.33, 0.98)	(-0.22, 1.11)	(-0.31, 1.00)	(-0.29, 0.94)	(-0.26, 0.85)
Lag 1	0.12	0.57	0.29	0.53	0.18	0.50	0.46	0.48
	(-0.59, 0.83)	(0.29, 0.85)	(-0.08, 0.67)	(0.06, 1.00)	(-0.47, 0.83)	(0.00, 1.09)	(0.01, 0.91)	(0.08, 0.88)
Lag 2	1.20	0.61	1.04	0.68	1.00	0.69	0.63	0.59
	(0.49, 1.91)	(0.35, 0.86)	(0.70, 1.38)	(0.34, 1.01)	(0.38, 1.62)	(0.26, 1.00)	(0.31, 0.95)	(0.30, 0.89)
Lag 3	0.22	0.51	0.39	0.54	0.27	0.54	0.52	0.45
	(-0.48, 0.92)	(0.27, 0.75)	(0.09, 0.70)	(0.21, 0.86)	(-0.29, 0.84)	(0.17, 0.90)	(0.22, 0.82)	(0.18, 0.72)
Lag 4	0.36	0.35	0.35	0.31	0.25	0.30	0.27	0.24
	(-0.34, 1.06)	(0.15, 0.55)	(-0.39, 1.10)	(0.01, 0.60)	(-0.24, 0.75)	(0.01, 0.70)	(0.01, 0.54)	(0.02, 0.46)
Lag 5	0.20	0.18	0.18	0.18	0.13	0.16	0.09	0.11
	(-0.49, 0.90)	(0.00, 0.36)	(-0.62, 0.98)	(-0.11, 0.46)	(-0.30, 0.56)	(-0.14, 0.53)	(-0.16, 0.33)	(-0.08, 0.29)
Lag 6	0.02	0.03	0.03	0.10	0.03	0.10	0.01	0.02
	(-0.68, 0.71)	(-0.16, 0.22)	(-0.79, 0.85)	(-0.19, 0.38)	(-0.34, 0.40)	(-0.28, 0.37)	(-0.22, 0.24)	(-0.14, 0.19)
Lag 7	0.01	-0.08	-0.07	-0.07	0.00	-0.05	-0.02	-0.03
	(-0.68, 0.70)	(-0.27, 0.12)	(-0.74, 0.60)	(-0.35, 0.21)	(-0.33, 0.33)	(-0.45, 0.21)	(-0.23, 0.19)	(-0.18, 0.12)
Lag 8	-0.08	-0.12	-0.12	-0.27	-0.04	-0.27	-0.06	-0.04
	(-0.77, 0.62)	(-0.31, 0.07)	(-0.82, 0.59)	(-0.55, 0.01)	(-0.35, 0.27)	(-0.55, 0.09)	(-0.26, 0.15)	(-0.17, 0.10)
Lag 9	-0.57	-0.11	-0.40	-0.22	-0.07	-0.26	-0.07	-0.02
	(-1.26, 0.12)	(-0.29, 0.07)	(-0.60, -0.21)	(-0.51, 0.06)	(-0.40, 0.26)	(-0.46, 0.20)	(-0.26, 0.11)	(-0.13, 0.10)
Lag 10	0.18	-0.05	0.02	0.12	0.03	0.12	-0.02	0.00
	(-0.51, 0.88)	(-0.24, 0.14)	(-0.26, 0.29)	(-0.17, 0.41)	(-0.20, 0.26)	(-0.23, 0.46)	(-0.19, 0.14)	(-0.11, 0.11)
Lag 11	0.57	0.02	0.41	0.32	0.05	0.36	0.02	0.00
	(-0.13, 1.27)	(-0.21, 0.24)	(0.15, 0.67)	(0.01, 0.64)	(-0.20, 0.30)	(-0.15, 0.59)	(-0.14, 0.17)	(-0.09, 0.10)
Lag 12	-0.19	0.05	-0.03	-0.03	-0.01	-0.02	-0.01	0.00
	(-0.88, 0.50)	(-0.18, 0.28)	(-0.36, 0.31)	(-0.34, 0.28)	(-0.20, 0.18)	(-0.38, 0.31)	(-0.15, 0.14)	(-0.09, 0.08)
Lag 13	-0.55	0.00	-0.39	-0.54	-0.03	-0.57	-0.03	0.00
	(-1.23, 0.14)	(-0.22, 0.23)	(-0.64, -0.14)	(-0.97, -0.10)	(-0.22, 0.17)	(-0.89, 0.14)	(-0.18, 0.12)	(-0.08, 0.08)
Lag 14	0.28	-0.20	0.13	0.24	0.01	0.26	-0.01	0.00
	(-0.34, 0.90)	(-0.64, 0.25)	(-0.36, 0.63)	(-0.33, 0.81)	(-0.14, 0.15)	(-0.43, 0.73)	(-0.14, 0.12)	(-0.07, 0.07)

Table 2.10: Estimated mean and 95% confidence intervals (in parenthesis) for the cumulative lag effect (% change in mortality count) of an interquartile range increase of PM_{10} ($21.49\mu g/m^3$) across lags on mortality in Chicago, Illinois from 1987 to 2000 based on the data from the National Morbidity, Mortality, and Air Pollution Study (NMMAPS) under eight estimation methods.

	UDLM	CDLM	EB1	GRR	BDLM	HB	HB2-GRR	HP-GRR
Lag 0	-0.15	-0.13	-0.13	-0.17	-0.18	-0.25	-0.13	-0.15
	(-0.61, 0.31)	(-0.51, 0.26)	(-0.51, 0.25)	(-0.61, 0.27)	(-0.66, 0.24)	(-0.59, 0.08)	(-0.55, 0.29)	(-0.53, 0.23)
Lag 1	-0.12	0.02	0.00	-0.14	-0.15	-0.09	-0.03	0.01
	(-0.71, 0.48)	(-0.53, 0.58)	(-0.42, 0.42)	(-0.70, 0.42)	(-0.78, 0.41)	(-0.48, 0.31)	(-0.56, 0.51)	(-0.40, 0.43)
Lag 2	0.10	0.27	0.25	0.23	0.17	0.29	0.24	0.36
	(-0.58, 0.78)	(-0.37, 0.92)	(-0.39, 0.88)	(-0.42, 0.89)	(-0.60, 0.91)	(-0.16, 0.73)	(-0.38, 0.86)	(-0.25, 0.96)
Lag 3	0.75	0.51	0.81	0.67	0.57	0.63	0.48	0.97
	(-0.01, 1.52)	(-0.22, 1.23)	(0.07, 1.56)	(-0.06, 1.40)	(-0.24, 1.43)	(0.14, 1.12)	(-0.21, 1.18)	(0.27, 1.67)
Lag 4	0.97	0.65	0.97	0.86	0.71	0.78	0.55	1.12
	(0.13, 1.81)	(-0.14, 1.45)	(0.10, 1.83)	(0.06, 1.67)	(-0.18, 1.69)	(0.23, 1.32)	(-0.22, 1.32)	(0.30, 1.94)
Lag 5	0.70	0.68	0.80	0.73	0.55	0.70	0.44	0.94
	(-0.20, 1.61)	(-0.19, 1.55)	(-0.11, 1.71)	(-0.15, 1.61)	(-0.45, 1.59)	(0.11, 1.30)	(-0.39, 1.28)	(0.06, 1.82)
Lag 6	0.30	0.59	0.50	0.44	0.24	0.49	0.28	0.60
	(-0.67, 1.27)	(-0.35, 1.52)	(-0.44, 1.44)	(-0.50, 1.39)	(-0.84, 1.35)	(-0.15, 1.13)	(-0.63, 1.18)	(-0.31, 1.51)
Lag 7	0.32	0.40	0.41	0.21	-0.01	0.26	0.14	0.48
	(-0.71, 1.36)	(-0.61, 1.40)	(-0.54, 1.37)	(-0.80, 1.22)	(-1.23, 1.16)	(-0.41, 0.94)	(-0.84, 1.12)	(-0.45, 1.41)
Lag 8	0.07	0.14	0.16	0.09	-0.15	0.09	0.04	0.24
	(-1.03, 1.18)	(-0.93, 1.21)	(-0.93, 1.25)	(-0.99, 1.16)	(-1.39, 1.13)	(-0.61, 0.79)	(-1.02, 1.09)	(-0.78, 1.26)
Lag 9	0.05	-0.13	0.02	0.00	-0.27	-0.03	-0.06	0.12
	(-1.12, 1.22)	(-1.27, 1.01)	(-1.10, 1.13)	(-1.14, 1.15)	(-1.55, 1.13)	(-0.76, 0.70)	(-1.19, 1.07)	(-0.92, 1.16)
Lag 10	-0.22	-0.39	-0.24	-0.14	-0.40	-0.16	-0.15	-0.05
	(-1.46, 1.01)	(-1.59, 0.82)	(-1.54, 1.07)	(-1.35, 1.08)	(-1.79, 1.05)	(-0.92, 0.60)	(-1.35, 1.06)	(-1.15, 1.05)
Lag 11	-0.35	-0.59	-0.43	-0.38	-0.64	-0.37	-0.23	-0.19
	(-1.65, 0.95)	(-1.86, 0.68)	(-1.70, 0.84)	(-1.66, 0.91)	(-2.09, 0.91)	(-1.17, 0.42)	(-1.51, 1.05)	(-1.31, 0.92)
Lag 12	-0.68	-0.72	-0.65	-0.69	-0.97	-0.67	-0.30	-0.35
	(-2.05, 0.68)	(-2.06, 0.62)	(-1.96, 0.66)	(-2.03, 0.66)	(-2.51, 0.65)	(-1.49, 0.16)	(-1.65, 1.05)	(-1.49, 0.79)
Lag 13	-0.93	-0.80	-0.78	-0.92	-1.23	-0.91	-0.37	-0.48
	(-2.35, 0.49)	(-2.21, 0.61)	(-2.10, 0.53)	(-2.33, 0.50)	(-2.85, 0.45)	(-1.77, -0.04)	(-1.78, 1.05)	(-1.66, 0.69)
Lag 14	-0.75	-0.87	-0.71	-0.74	-1.05	-0.72	-0.43	-0.57
	(-2.24, 0.75)	(-2.36, 0.62)	(-1.98, 0.56)	(-2.23, 0.75)	(-2.76, 0.69)	(-1.60, 0.15)	(-1.91, 1.05)	(-1.78, 0.64)

Table 2.11: Estimated mean and 95% confidence intervals (in parenthesis) for the cumulative lag effect (% change in mortality count) of an interquartile range increase of O_3 (14.65 ppb) across lags on mortality in Chicago, Illinois from 1987 to 2000 based on the data from the National Morbidity, Mortality, and Air Pollution Study (NMMAPS) under eight estimation methods.

	UDLM	CDLM	EB1	GRR	BDLM	HB	HB2-GRR	HP-GRR
Lag 0	0.50	0.33	0.36	0.32	0.45	0.37	0.32	0.30
	(-0.21, 1.20)	(-0.19, 0.85)	(-0.26, 0.98)	(-0.33, 0.98)	(-0.22, 1.11)	(-0.28, 1.03)	(-0.29, 0.94)	(-0.26, 0.85)
Lag 1	0.61	0.90	0.65	0.85	0.63	0.87	0.78	0.78
	(-0.19, 1.42)	(0.16, 1.64)	(-0.27, 1.58)	(0.10, 1.61)	(-0.15, 1.40)	(0.11, 1.64)	(0.06, 1.51)	(0.10, 1.45)
Lag 2	1.82	1.51	1.70	1.53	1.63	1.57	1.42	1.38
	(0.87, 2.77)	(0.65, 2.38)	(0.59, 2.82)	(0.66, 2.41)	(0.74, 2.53)	(0.73, 2.42)	(0.58, 2.27)	(0.60, 2.16)
Lag 3	2.04	2.03	2.09	2.08	1.91	2.12	1.94	1.83
	(0.98, 3.11)	(1.07, 3.00)	(0.82, 3.39)	(1.10, 3.07)	(0.93, 2.90)	(1.18, 3.07)	(1.01, 2.89)	(0.97, 2.70)
Lag 4	2.40	2.39	2.45	2.39	2.17	2.42	2.22	2.08
	(1.27, 3.55)	(1.34, 3.44)	(0.86, 4.07)	(1.31, 3.48)	(1.12, 3.23)	(1.39, 3.47)	(1.19, 3.26)	(1.14, 3.02)
Lag 5	2.61	2.57	2.63	2.57	2.29	2.59	2.31	2.18
	(1.41, 3.83)	(1.45, 3.69)	(1.13, 4.16)	(1.42, 3.73)	(1.18, 3.42)	(1.47, 3.71)	(1.22, 3.41)	(1.19, 3.19)
Lag 6	2.63	2.59	2.66	2.67	2.32	2.69	2.32	2.21
	(1.36, 3.91)	(1.42, 3.78)	(1.01, 4.34)	(1.46, 3.89)	(1.16, 3.50)	(1.51, 3.88)	(1.17, 3.48)	(1.15, 3.27)
Lag 7	2.63	2.52	2.59	2.59	2.32	2.63	2.30	2.18
	(1.31, 3.98)	(1.28, 3.77)	(0.97, 4.24)	(1.33, 3.88)	(1.11, 3.56)	(1.41, 3.87)	(1.10, 3.52)	(1.06, 3.31)
Lag 8	2.55	2.39	2.47	2.32	2.28	2.35	2.25	2.14
	(1.18, 3.95)	(1.08, 3.72)	(0.84, 4.11)	(1.00, 3.66)	(1.02, 3.57)	(1.07, 3.65)	(0.99, 3.52)	(0.95, 3.34)
Lag 9	1.97	2.28	2.05	2.09	2.21	2.09	2.17	2.12
	(0.54, 3.41)	(0.91, 3.68)	(0.38, 3.75)	(0.71, 3.49)	(0.87, 3.57)	(0.74, 3.45)	(0.85, 3.51)	(0.86, 3.40)
Lag 10	2.15	2.23	2.07	2.22	2.24	2.21	2.15	2.13
	(0.66, 3.67)	(0.80, 3.69)	(0.32, 3.85)	(0.77, 3.68)	(0.85, 3.66)	(0.80, 3.64)	(0.77, 3.55)	(0.79, 3.48)
Lag 11	2.73	2.25	2.48	2.55	2.29	2.57	2.17	2.13
	(1.17, 4.32)	(0.75, 3.77)	(0.63, 4.37)	(1.02, 4.09)	(0.84, 3.76)	(1.10, 4.06)	(0.71, 3.64)	(0.73, 3.55)
Lag 12	2.53	2.30	2.46	2.52	2.28	2.55	2.16	2.13
	(0.91, 4.18)	(0.73, 3.89)	(0.45, 4.50)	(0.93, 4.13)	(0.77, 3.81)	(1.04, 4.09)	(0.64, 3.70)	(0.66, 3.62)
Lag 13	1.97	2.31	2.06	1.97	2.26	1.96	2.13	2.12
	(0.30, 3.67)	(0.67, 3.97)	(-0.02, 4.18)	(0.31, 3.65)	(0.69, 3.85)	(0.39, 3.57)	(0.54, 3.74)	(0.59, 3.68)
Lag 14	2.25	2.10	2.19	2.21	2.26	2.23	2.12	2.12
	(0.53, 4.01)	(0.38, 3.85)	(0.11, 4.32)	(0.48, 3.97)	(0.64, 3.91)	(0.61, 3.88)	(0.46, 3.80)	(0.53, 3.73)

Table 2.12: Computation times of applying eight estimation methods to National Morbidity, Mortality, and Air Pollution Study (NMMAPS) data on an Intel i7-2600 CPU with a single 3.4GHz core.

Methods	Time
UDLM	1.7 seconds
CDLM	1.6 seconds
EB1	5.8 seconds
GRR[1]	63.7 seconds
BDLM[2]	5.4 seconds
HB[3]	1.1 hours
HB2-GRR[1,4]	13.1 hours
HP-GRR[1,4]	14.1 hours

[1] Tuning parameter is chosen from a grid of 100 equally-spaced values
[2] Asymptotic normality of the Poisson likelihood is applied
[3] Gibbs sampler is based on 10000 iterations
[4] Standard error estimates are based on 1000 bootstrap samples

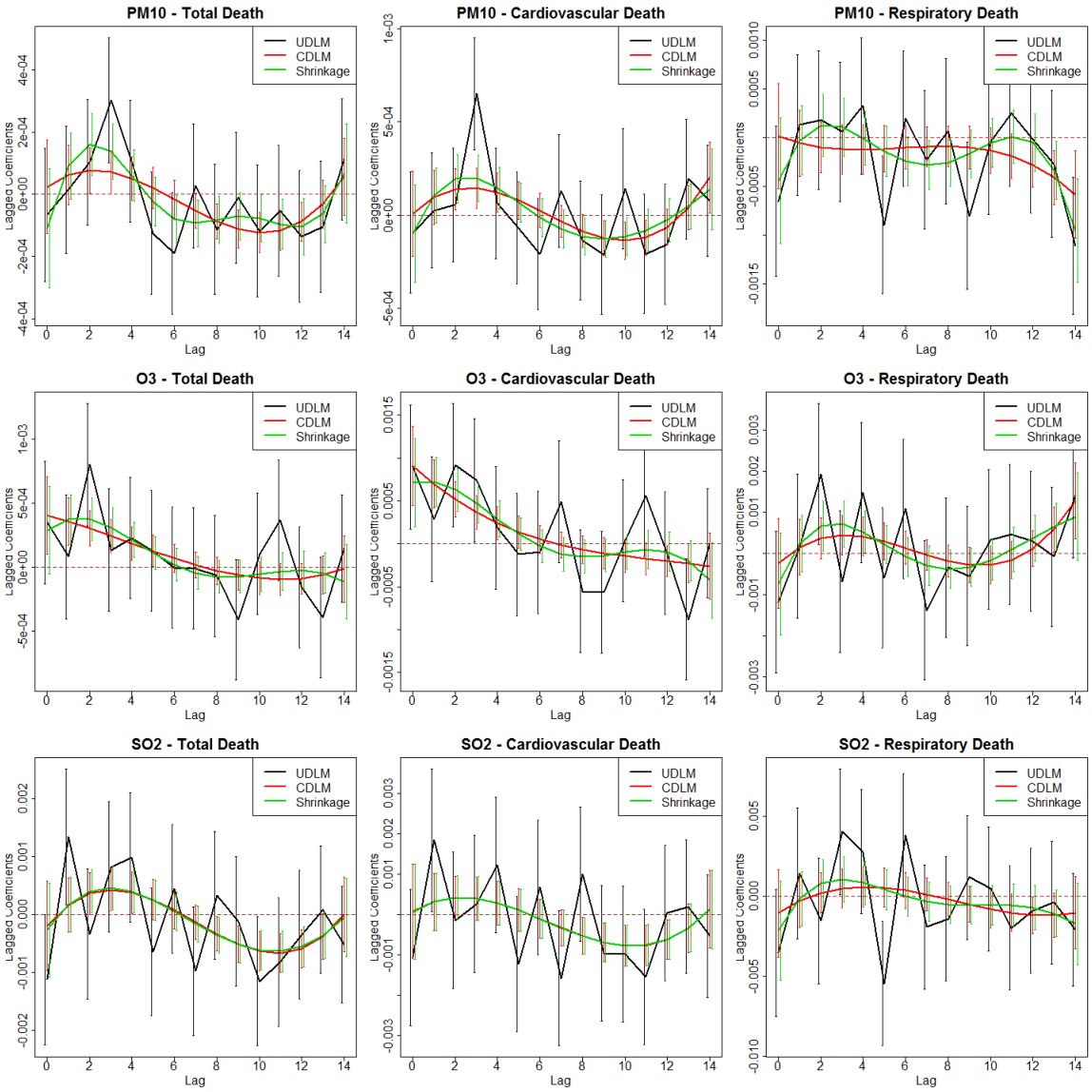

Figure 2.1: Estimated distributed lag functions up to 14 days for PM_{10}, O_3, and SO_2 on total mortality, cardiovascular mortality, and respiratory mortality with 95% confidence/credible interval at each lag in Chicago, Illinois from 1987 to 2000 based on the National Morbidity, Mortality, and Air Pollution Study (NMMAPS) data.

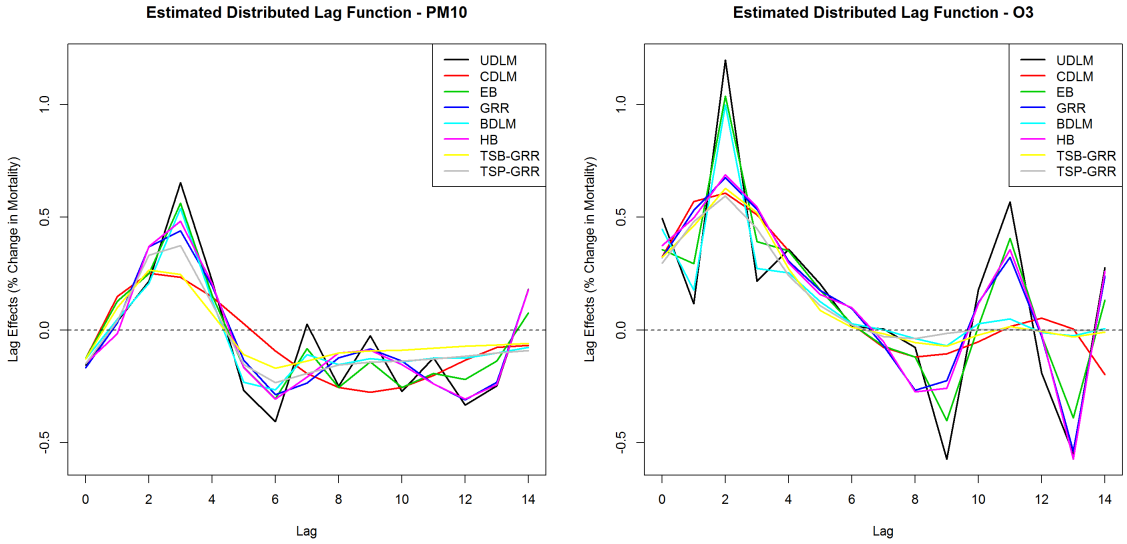

Figure 2.2: Estimated distributed lag functions up to 14 days for PM_{10} (left) and O_3 (right) on mortality in Chicago, Illinois from 1987 to 2000 based on the data from the National Morbidity, Mortality, and Air Pollution Study (NMMAPS) under eight estimation methods. The lag effects are presented as the percentage change in mortality with an interquartile range increase in the exposure level (PM_{10}: 21.49$\mu g/m^3$, O_3: 14.65 ppb).

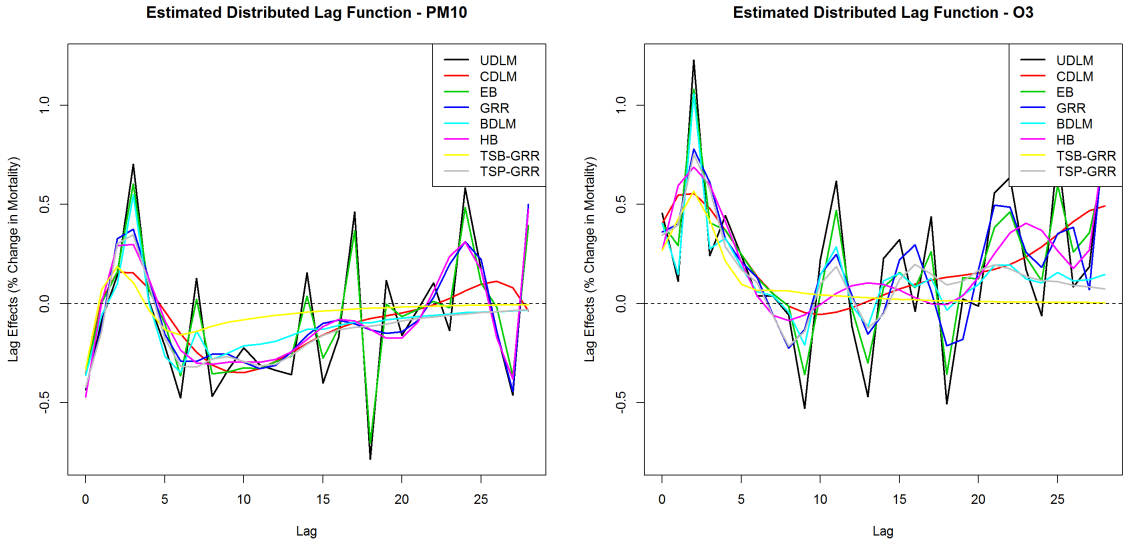

Figure 2.3: Estimated distributed lag functions up to 28 days for PM_{10} (left) and O_3 (right) on mortality in Chicago, Illinois from 1987 to 2000 based on the data from the National Morbidity, Mortality, and Air Pollution Study (NMMAPS) under eight estimation methods. The lag effects are presented as the percentage change in mortality with an interquartile range increase in the exposure level (PM_{10}: 21.49$\mu g/m^3$, O_3: 14.65 ppb).

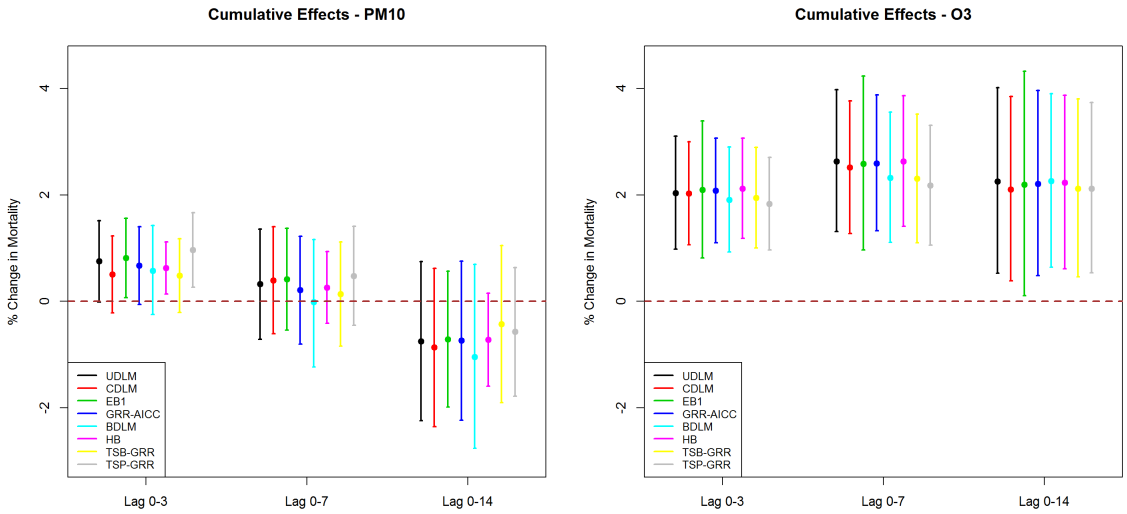

Figure 2.4: Estimated mean and 95% confidence/credible interval of the cumulative lagged effect (% change in mortality count) up to 3, 7, and 14 days of PM_{10} (left) and O_3 (right) on mortality with an interquartile range increase in exposure level (PM_{10}: $21.49 \mu g/m^3$, O_3: 14.65 ppb) in Chicago, Illinois from 1987 to 2000 based on the data from the National Morbidity, Mortality, and Air Pollution Study (NMMAPS) under eight estimation methods.

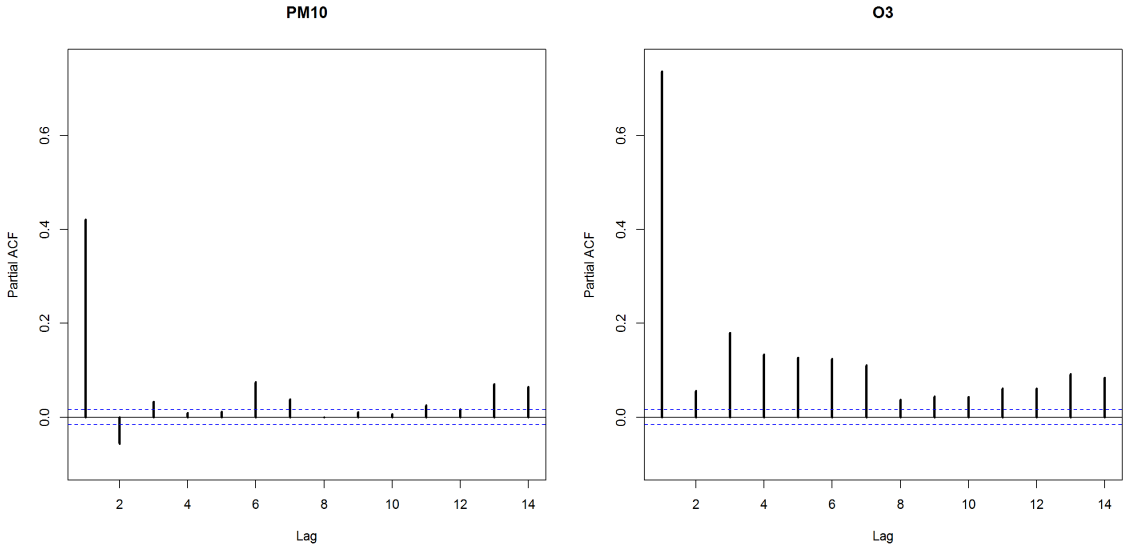

Figure 2.5: Partial autocorrelation function (PACF) plots for daily measurements of PM_{10} (left) and O_3 (right) in Chicago, Illinois from 1987 to 2000 based on the National Morbidity, Mortality and Air Pollution Study (NMMAPS) data.

CHAPTER 3

A New Variance Component Score Test for Testing Distributed Lag Functions

3.1 Introduction

DLMs, first introduced in the econometrics literature, are often used to model the current value of a response variable at time t in association with both the current value and the lagged values of an independent variable in a time series analysis. For example, environmental epidemiologists model the current day mortality counts in association with daily air pollution related exposure, such as PM_{10}, up to several days prior to the event day [Schwartz, 2000, Welty et al., 2009]. Economists study the long-term effects of macroeconomic variables on stock returns using distributed lag models [Majid and Yusof, 2009, Hsu, 2015]. Unconstrained DLMs entail the potential problem of multicollinearity among the various lagged values of the independent variable and the number of parameters to be estimated can be large. Constrained DLMs assume some functional relationship between lag coefficients and lag indices (in the form of a distributed lag function) and serve as a potential solution to the problem. Common constraints include a polynomial [Almon, 1965], a spline [Corradi, 1977], and a natural cubic spline [Hastie and Tibshirani, 1993]. There exists extensive literature on characterization of the distributed lag function and inference, assuming the constraints are correctly specified. Very few strategies exist for testing given distributed lag (DL) constraints. In this chapter, we propose a new simple and efficient framework for testing a constrained DLM against an unconstrained DLM. In Section 3.2, we briefly introduce DLMs and present the proposed VCST procedure. In Section 3.3, we

conduct a simulation study to compare the statistical power of the standard likelihood ratio test (LRT) and VCST and illustrate both of the approaches using the NMMAPS data. We conclude with discussions in Section 3.4.

3.2 Method

Let x_t denote the independent variable measured at time t, y_t denote the response variable measured at time t, $\boldsymbol{z}_t$ denote the other covariates obtained at time t, T be the length of the time series, and L be the pre-determined maximum number of lags. Without loss of generality, we leave out intercept and covariates in the rest of the presentation and consider the generalized linear model $\eta_t = g[\mu_t] = g[E(y_t|x_t, x_{t-1}, ..., x_{t-L})] = \boldsymbol{X}_t^\top \boldsymbol{\beta}$ where $\boldsymbol{\beta} = (\beta_0, \beta_1, ..., \beta_L)^\top$ is the vector of the lag effects, $\boldsymbol{X_t} = (x_t, x_{t-1}, ..., x_{t-L})^\top$, η_t is the canonical (natural) parameter, $g(\cdot)$ is the link function, y_t is a random variable generated from a distribution $\mathcal{F}$ in canonical exponential family with probability density

$$f(y_t) = \exp\{\frac{y_t\eta_t - b(\eta_t)}{a(\phi)} + c(y_t; \phi)\}, \tag{3.1}$$

ϕ is the dispersion parameter, and $a(\cdot)$, $b(\cdot)$, and $c(\cdot)$ are known functions. It is well-known that the exponential family possess the properties of $\mu_t = b^{'}(\boldsymbol{X}_t^\top \boldsymbol{\beta}) = g^{-1}(\boldsymbol{X}_t^\top \boldsymbol{\beta})$ and $V(y_t) = b^{''}(\boldsymbol{X}_t^\top \boldsymbol{\beta})a(\phi) = \nu(\boldsymbol{X}_t^\top \boldsymbol{\beta})a(\phi)$ where $\nu(\cdot)$ is the variance function.

3.2.1 Constrained DLM

Constrained DLM imposes a pre-specified structure to constrain the lag coefficients to be a smooth function of the lags (i.e. $\beta_\ell = f(\ell)$ for $\ell = 0, \cdots, L$). Denote the p basis functions that generate the class of functions in which $\boldsymbol{\beta}$ can lie as $B_1(\cdot), \cdots, B_p(\cdot)$. The

transformation matrix $\boldsymbol{C}$ as defined by Gasparrini et al. [2010] is given by

$$\boldsymbol{C} = \begin{bmatrix} B_1(0) & B_2(0) & \cdots & B_p(0) \\ B_1(1) & B_2(1) & \cdots & B_p(1) \\ \vdots & \vdots & \vdots & \vdots \\ B_1(L) & B_2(L) & \cdots & B_p(L) \end{bmatrix}_{(L+1)\times p}$$

The constrained DLM estimator can be expressed in the form of $\boldsymbol{\beta} = \boldsymbol{C\theta}$ where $\boldsymbol{\theta}$ is a vector of p free parameters to be estimated in $\mathbb{R}^p$. The maximum likelihood estimate of $\boldsymbol{\theta}$ is obtained as

$$\hat{\boldsymbol{\theta}} = \arg\max_{\boldsymbol{\theta}} \sum_{t=1}^{T} \Big[y_t \boldsymbol{X}_t^\top \boldsymbol{C\theta} - b(\boldsymbol{X}_t^\top \boldsymbol{C\theta}) \Big].$$

The estimation of $\boldsymbol{\theta}$ does not involve the dispersion parameter ϕ. The constrained DLM estimator is given by $\hat{\boldsymbol{\beta}}_{CDLM} = \boldsymbol{C}\hat{\boldsymbol{\theta}}$. The $\boldsymbol{C}$ corresponding to an unconstrained DLM is a $(L+1) \times (L+1)$ identity matrix.

3.2.2 Hypothesis Testing

$\hat{\boldsymbol{\beta}}_{CDLM}$ can be alternatively obtained by maximizing log-likelihood with respect to $\boldsymbol{\beta}$ subject to $\boldsymbol{R\beta} = \boldsymbol{0}$ where $\boldsymbol{R}$ is a $(L+1-p) \times (L+1)$ constraint matrix corresponding to the transformation matrix $\boldsymbol{C}$ [Chen et al., 2017]. The basis functions in $\boldsymbol{C}$ span the solution space of $\boldsymbol{R\beta} = \boldsymbol{0}$. $\boldsymbol{R}$ can be obtained from $\boldsymbol{C}$ via the following procedure. Define $\boldsymbol{C}_e$ as a $(L+1)\times(L+1)$ matrix $[\boldsymbol{C}\ \boldsymbol{0}_{(L+1)\times(L+1-p)}]$ where $\boldsymbol{0}_{(L+1)\times(L+1-p)}$ is a $(L+1)\times(L+1-p)$ matrix with zero entries. Applying SVD $\boldsymbol{C}_e^\top = \boldsymbol{U}_C \boldsymbol{D}_C \boldsymbol{V}_C^\top$ where $\boldsymbol{U}_C$ is the $(L+1)\times(L+1)$ unitary matrix with left-singular column vectors, $\boldsymbol{V}_C$ is the $(L+1)\times(L+1)$ unitary matrix with right-singular column vectors, and $\boldsymbol{D}_C$ is a $(L+1) \times (L+1)$ diagonal matrix with singular values of $\boldsymbol{C}_e^\top$ along the diagonal, the constraint matrix $\boldsymbol{R}$ can then be obtained as the last $(L+1-p)$ rows of $\boldsymbol{V}_C^\top$.

Testing a particular DLM structure against an unconstrained alternative can now be formulated as testing $H_0 : \boldsymbol{R\beta} = \boldsymbol{0}$ against $H_1 : \boldsymbol{R\beta} \neq \boldsymbol{0}$. A standard likelihood ratio test (LRT), Wald test, and score test can be conducted and the test statistics asymptotically follow a χ^2 distribution with $L+1-p$ degrees of freedom and large sample inference

can be obtained. We propose a VCST approach to this problem. Consider a generalized ridge regression estimator [Chen et al., 2017] that minimizes the penalized negative log-likelihood function

$$\ell_p(\boldsymbol{\beta}) = \Big[S(\boldsymbol{X}\boldsymbol{\beta}) - \boldsymbol{Y}^\top \boldsymbol{X}\boldsymbol{\beta}\Big] + \lambda \boldsymbol{\beta}^\top \boldsymbol{R}^\top \boldsymbol{R}\boldsymbol{\beta} \tag{3.2}$$

where $\boldsymbol{Y} = (y_1, ..., y_T)^\top$, $\boldsymbol{X}$ is a $T \times (L+1)$ matrix with $\boldsymbol{X}_t^\top$ as the t-th row for $t = 1, \cdots, T$, $S(\cdot)$ is the $\mathbb{R}^T \to \mathbb{R}^1$ cumulant function such that $S(\boldsymbol{a}) = \sum_{t=1}^{\top} b(a_t)$ with $\boldsymbol{a} = (a_1, \cdots, a_T)^\top$, and λ is the tuning parameter. We can rewrite

$$\boldsymbol{R}^\top \boldsymbol{R} = \boldsymbol{U}\boldsymbol{D}\boldsymbol{U}^\top \tag{3.3}$$

where $\boldsymbol{U}$ is a $(L+1) \times (L+1)$ matrix with orthogonal columns and $\boldsymbol{D}$ is a diagonal matrix with the eigenvalues of $\boldsymbol{R}^\top \boldsymbol{R}$ using SVD. Since $\boldsymbol{R}^\top \boldsymbol{R}$ is not of full rank and has rank $L+1-p$, we can write $\boldsymbol{D} = \begin{bmatrix} \boldsymbol{D}_1 & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{0} \end{bmatrix}$ where $\boldsymbol{D}_1$ is a $(L+1-p) \times (L+1-p)$ diagonal matrix of full rank. Let $\boldsymbol{U} = [\boldsymbol{U}_1\ \boldsymbol{U}_2]$ where $\boldsymbol{U}_1$ is a $(L+1) \times (L+1-p)$ matrix with columns of singular vectors corresponding to nonzero eigenvalues in $\boldsymbol{D}$ and $\boldsymbol{U}_2$ is a $(L+1) \times p$ matrix with columns of singular vectors corresponding to the eigenvalues of 0. One can rewrite the penalized negative log-likelihood as

$$\ell_p(\boldsymbol{\beta}) = \Big[S(\boldsymbol{X}^*\boldsymbol{\beta}^* + \boldsymbol{Z}^*\boldsymbol{u}) - \boldsymbol{Y}^\top(\boldsymbol{X}^*\boldsymbol{\beta}^* + \boldsymbol{Z}^*\boldsymbol{u})\Big] + \lambda \boldsymbol{u}^\top \boldsymbol{u} \tag{3.4}$$

where $\boldsymbol{X}^* = \boldsymbol{X}\boldsymbol{U}_2$, $\boldsymbol{\beta}^* = \boldsymbol{U}_2^\top \boldsymbol{\beta}$, $\boldsymbol{Z}^* = \boldsymbol{X}\boldsymbol{U}_1\boldsymbol{D}_1^{-1/2}$, and $\boldsymbol{u} = \boldsymbol{D}_1^{1/2}\boldsymbol{U}_1^\top \boldsymbol{\beta}$. We rewrite (3.2) as (3.4) to divide $\boldsymbol{X}$ into two parts, namely $\boldsymbol{X}^*$ and $\boldsymbol{Z}^*$. The first part $\boldsymbol{X}^*$ represents the orthogonal projection of $\boldsymbol{X}$ onto the space (say, $\mathcal{W}$) spanned by the p basis functions. The second part $\boldsymbol{Z}^*$ contains the transformed variables in the orthogonal complement of $\mathcal{W}$. The notion is to transform $\boldsymbol{\beta}$ and separate the component representing the specified DLM ($\boldsymbol{\beta}^*$) from the component representing the departure from the DLM ($\boldsymbol{u}$). Testing $H_0 : \boldsymbol{R}\boldsymbol{\beta} = \boldsymbol{0}$ against $H_1 : \boldsymbol{R}\boldsymbol{\beta} \neq \boldsymbol{0}$ is the same as testing whether $\boldsymbol{u}$ varies significantly from $\boldsymbol{0}$ or not.

The expressions in (3.4) can be viewed as the joint log-likelihood of $f(\boldsymbol{Y}, \boldsymbol{u}|\boldsymbol{X}^*) = f(\boldsymbol{Y}|\boldsymbol{u}, \boldsymbol{X}^*)f(\boldsymbol{u}|\boldsymbol{X}^*) = f(\boldsymbol{Y}|\boldsymbol{u}, \boldsymbol{X}^*)f(\boldsymbol{u})$ in a typical generalized linear mixed model of the form

$$\boldsymbol{Y}|\boldsymbol{u}, \boldsymbol{X}^* \sim \mathcal{F}(\boldsymbol{X}^*\boldsymbol{\beta}^* + \boldsymbol{Z}^*\boldsymbol{u}) \text{ , } \boldsymbol{u} \sim N(\boldsymbol{0}, \sigma_u^2) \tag{3.5}$$

where $\mathcal{F}$ is the canonical distribution belonging to the exponential family defined in (3.1) and σ_u^2 is the random effect variance, inversely proportional to λ in (3.4), with $\boldsymbol{u}$ independent of $\boldsymbol{X}^*$ as a result of the definition of $\boldsymbol{u}$ and $\boldsymbol{X}^*$ by performing the SVD as (3.3).

As λ in (3.4) becomes larger, $\boldsymbol{R\beta}$ is forced to approach $\boldsymbol{0}$. Since $\sigma_u^2 \propto \frac{1}{\lambda}$ and $\sigma_u^2 \to 0$ as $\lambda \to \infty$, testing $H_0 : \boldsymbol{R\beta} = \boldsymbol{0}$ against $H_1 : \boldsymbol{R\beta} \neq \boldsymbol{0}$ is equivalent to testing $H_0 : \sigma_u^2 = 0$ against $H_1 : \sigma_u^2 > 0$. Let $\mu_t^{(H_0)} = E_{H_0}(y_t)$, the expected value of y_t under the null and let $\boldsymbol{\mu}^{(H_0)} = (\mu_1^{(H_0)}, \cdots, \mu_T^{(H_0)})^\top$, $\boldsymbol{\Delta} = \text{diag}[g'(\mu_t)]$, and $\boldsymbol{V} = \text{diag}\{a(\phi)\nu(\mu_t)[g'(\mu_t)]^2\}$. The VCST statistic [Lin, 1997, Zhang and Lin, 2003] is given by

$$Q = (\boldsymbol{Y} - \hat{\boldsymbol{\mu}}^{(H_0)})^\top \hat{\boldsymbol{\Delta}} \hat{\boldsymbol{V}}^{-1} \boldsymbol{Z}^* \boldsymbol{Z}^{*\top} \hat{\boldsymbol{V}}^{-1} \hat{\boldsymbol{\Delta}} (\boldsymbol{Y} - \hat{\boldsymbol{\mu}}^{(H_0)}). \tag{3.6}$$

where $\hat{\boldsymbol{\mu}}^{(H_0)}$, $\hat{\boldsymbol{\Delta}}$, and $\hat{\boldsymbol{V}}$ are restricted maximum likelihood (REML) estimates of $\boldsymbol{\mu}^{(H_0)}$, $\boldsymbol{\Delta}$, and $\boldsymbol{V}$, respectively, under H_0. The test statistic Q asymptotically follows a mixture $\sum_{i=1}^{L+1-p} \gamma_i \chi_{1,i}^2$ where $\gamma_1, ..., \gamma_{L+1-p}$ are the eigenvalues of $\hat{\boldsymbol{V}}^{-1/2} \boldsymbol{Z}^{*\top} \boldsymbol{Z}^* \hat{\boldsymbol{V}}^{-1/2}$ and $\chi_{1,i}^2$ are independent χ_1^2 random variables. The Davies exact method [Davies, 1980a] based on inverting the characteristic function can be used to calculate the distribution of any quadratic form in normal random variables. The details are provided in the Appendix 3.5.2. We make use of the method to obtain the threshold of a significance level and p-value corresponding to Q via `R` package `CompQuadForm` for simulation and data analysis in later sections.

Remark: The VCST approach can be applied for testing $H_0 : \boldsymbol{R}_0\boldsymbol{\beta} = \boldsymbol{0}$ against $H_1 : \boldsymbol{R}_1\boldsymbol{\beta} = \boldsymbol{0}$ where the DLM represented by $\boldsymbol{R}_0$ is nested within the DLM represented by $\boldsymbol{R}_1$. Let $\boldsymbol{C}_0$ and $\boldsymbol{C}_1$ denote the $(L+1)\times p_0$ transformation matrix and $(L+1)\times p_1$ transformation matrix corresponding to $\boldsymbol{R}_0$ and $\boldsymbol{R}_1$, respectively. The penalized negative log-likelihood

function analogous to (3.2) is given by

$$\ell_p(\boldsymbol{\theta}) = \left[S(\boldsymbol{X}\boldsymbol{C}_1\boldsymbol{\theta}) - \boldsymbol{Y}^\top \boldsymbol{X}\boldsymbol{C}_1\boldsymbol{\theta}\right] + \lambda\boldsymbol{\theta}^\top \boldsymbol{C}_1^\top \boldsymbol{R}_0^\top \boldsymbol{R}_0 \boldsymbol{C}_1\boldsymbol{\theta}. \tag{3.7}$$

Similarly, (3.7) can be rewritten as the joint log-likelihood function of a generalized linear mixed model and the hypothesis testing can be conducted subsequently. However, the power gain from VCST as opposed to LRT is less in this situation because the degrees of freedom of LRT, namely $p_1 - p_0$, is generally small. For example, for testing a cubic DLM versus a linear DLM, we will have $p_1 - p_0 = 2$. We considered a simulation setting to illustrate the potential power gain when testing a nested DLM against a more general one for small to moderate $p_1 - p_0$. The results are provided in Appendix 3.5.2. One can observe that the LRT and VCST have similar power when the difference in the degree of the two DLMs is small and there is modest power gain by using VCST when this difference is large.

3.3 Results

3.3.1 Simulation

We conducted a simulation study to compare the power of the standard LRT and VCST for testing $H_0 : \boldsymbol{R}\boldsymbol{\beta} = \mathbf{0}$ against $H_1 : \boldsymbol{R}\boldsymbol{\beta} \neq \mathbf{0}$. We generated a predictor series of length 2000 with mean 0 and first order autocorrelation $\rho = 0.6$ from the model $x_t = \rho x_{t-1} + \epsilon_t$ where $\epsilon_t \sim$ i.i.d $N(0, 1)$ for $t = 1, ..., 2000$ for the independent variable. We then set $L = 30$ and generated the outcome series $\boldsymbol{Y}$ from the model $y_t = \delta \sum_{\ell=0}^{L} \beta_\ell x_{t-\ell} + \epsilon_t$ where $\boldsymbol{\beta} = (\beta_0, ..., \beta_L)^\top$ denote the true coefficients and $\epsilon_t \sim$ i.i.d $N(0, 1)$ for $t = 1, ..., 2000$. δ is used to control the signal-to-noise ratio. We generated 1000 data sets and calculated the empirical power at level $\alpha = 0.05$ as the observed proportion of times that a test yields a p-value less than α.

We first examined the type I error rate and power of the two testing procedures across different combinations of true DLMs and fitted DLMs. The results with moderate signal-to-noise level ($\delta = 1$) are presented in Table 3.1 and the results with stronger ($\delta = 1.25$) and

weaker ($\delta = 0.8$) signal-to-noise levels are provided in Tables 3.3 and 3.4. When the true DLM coincides with the fitted DLM, the numbers in italics are estimated type I error rates. When the true DLM is different from the fitted DLM, the numbers are estimated powers. As we can observe, both LRT and VCST properly maintain the type I error rate at level $\alpha = 0.05$. In addition, VCST is more powerful than LRT irrespective of the signal-to-noise level across all combinations of true DLM and fitted DLM. We also examined the gain in statistical power to detect the departure from a DLM using VCST as opposed to using LRT with different autocorrelation levels for predictor series ρ and different maximum number of lags L. The left panel of Figure 3.1 displays the power curves of VCST against LRT with three different autocorrelation levels. The three curves are all above the 45° line indicating that VCST is more powerful than LRT across the board. Moreover, the higher the correlation in the predictor series is, the more advantageous using VCST is. The justification is that LRT treats L correlated predictors as L independent variables and it does not make use of the serial correlation when constructing the test statistic, whereas VCST orthogonalizes the design matrix by removing the correlation and increases power through testing on a single variance component by estimating the true underlying degrees of freedom (df). The right panel of Figure 3.1 exhibits the power curves of VCST against LRT with three different maximum numbers of lags L. Again, the curves are all above the 45° line illustrating that VCST is more powerful than LRT across different L. We can also notice that the larger the L, the more power gain from using VCST. The explanation is that large L leads to loss of power at the expense of more degrees of freedom for LRT. The gain in power by using VCST via estimating the effective df is more appreciable when L is large and the correlation in x_t is strong.

3.3.2 Application to NMMAPS Data

We illustrated LRT and VCST using NMMAPS data. The details describing NMMAPS data are available at `http://www.ihapss.jhsph.edu/data/NMMAPS/`. We follow Welty et al. [2009] for the choice of $L = 14$ and the specification of covariates to associate daily particular matter, with aerodynamic diameter less than 10 micrometers (PM_{10}),

with (1) cardiovascular death count and (2) daily non-accidental mortality counts. We consider polynomial DLMs from degrees 1 to 5 for the lagged effects of exposure on each health outcome. We conducted hypothesis testing to evaluate whether the specified DLMs depart from the underlying unconstrained DLM. Since the outcomes are counts, we specify the first term of the penalized log-likelihood function (3.2) as the negative log-likelihood of Poisson distribution with mean $e^{\boldsymbol{X}\boldsymbol{\beta}}$. Similar testing procedure as described in Section 3.2.2 follows.

Table 3.2 presents the p-values obtained from LRT and VCST. In general, the two testing procedures yield similar results for each combination of a specified DLM and a health outcome, with the p-values from VCST slightly smaller than those obtained from LRT. We resist to interpret that smaller p-values indicate higher power although the smaller p-values from VCST provide stronger evidence against the null. For this example, both tests suggest that none of the DL functions considered is adequate for modeling cardiovascular death in association with PM_{10}. For mortality, it is suggested that a 5-degree polynomial is more appropriate to characterize the DL function than the lower-order polynomial.

The estimated distributed lag functions are displayed in Figure 3.2. The interquartile range of PM_{10} is 21.49μg/m^3. The quantity $100[\exp(21.49\sum_{j=0}^{\ell}\beta_j) - 1]$ represents to the percentage change in daily cardiovascular death/non-accidental mortality associated with an IQR increase in PM_{10} (21.49μg/m^3) across lag 0-ℓ. For cardiovascular death, the estimated cumulative lag effects up to lag 4, lag 7, and lag 14 are 1.61% (95%CI: [0.34%, 2.87%]), 1.27% (95%CI: [-0.30%, 2.84%]), and 0.53% (95%CI: [-1.74%, 2.80%]) based on the unconstrained DLM. For non-accidental mortality, the estimated cumulative lag effects up to lag4, lag 7, and lag 14 are 0.69% (95%CI: [-0.11%, 1.49%]), 0.27% (95%CI: [-0.74%, 1.27%]), and -0.80% (95%CI: [-2.29%, 0.69%]) based on the 5-degree polynomial DLM.

3.4 Discussion

The simulation study indicates that VCST procedure is more powerful than a standard LRT for testing a constrained DLM against an unconstrained alternative across different

scenarios. The power gain from VCST over LRT is greater when the predictor series autocorrelation is larger and when the maximal lag is larger. One caveat is that the VCST framework only applies to the constrained DLMs for which the estimator can be written in the form of $\boldsymbol{C\theta}$, so that the corresponding constraint matrix $\boldsymbol{R}$ can be obtained and the hypothesis testing can be constructed as $H_0 : \boldsymbol{R\beta} = \mathbf{0}$ against $H_1 : \boldsymbol{R\beta} \neq \mathbf{0}$. Most commonly used DLMs fall into this category. Bayesian DLM [Welty et al., 2009] is an exception since all the lag effect coefficients are related through the specification of variance-covariance matrix rather than the mean. Although we illustrate the testing procedure using a data set in environmental epidemiology, the new test is potentially useful for time series analysis in economics, finance, ecology, and a wide range of applications.

3.5 Appendix

3.5.1 Davies Exact Method

The test statistic Q for VCST asymptotically follows a mixture $\sum_{i=1}^{L+1-p} \gamma_i X_i$ where $\gamma_1, ..., \gamma_{L+1-p}$ are the eigenvalues of $\hat{\boldsymbol{V}}^{-1/2} \boldsymbol{Z}^{*\top} \boldsymbol{Z}^* \hat{\boldsymbol{V}}^{-1/2}$ and X_is are independent χ_1^2 random variables. The characteristic function of Q is

$$\psi(\mu) = E(e^{i\mu Q}) = \frac{1}{\prod_{j=1}^{L+1-p}(1 - 2i\mu\lambda_j)^{\frac{1}{2}}}$$

and the numerical inversion of the characteristic function based on Poisson formula can be used to derive the distribution function

$$F(x) \sim \sum_{\nu=-\infty}^{\infty} \frac{sin(\delta\nu x)}{\pi\nu}\psi(\delta\nu) \approx \sum_{\nu=-N}^{N} \frac{sin(\delta\nu x)}{\pi\nu}\psi(\delta\nu)$$

with sufficiently small δ. The calculation of truncation point N is discussed in Bohman [1970] and the bounds of integration error are detailed in Davies [1980b].

3.5.2 Simulation on Testing $H_0 : R_0\beta = 0$ against $H_1 : R_1\beta = 0$

We conducted a simulation study to compare the power of the standard LRT and VCST for testing $H_0 : \boldsymbol{R}_0\boldsymbol{\beta} = \mathbf{0}$ against $H_1 : \boldsymbol{R}_1\boldsymbol{\beta} \neq \mathbf{0}$ where $\boldsymbol{R}_0$ corresponds to quadratic DLM and $\boldsymbol{R}_1$ corresponds to a higher-degree polynomial DLM. We set $L = 30$ and generated data from the DLM under the alternative. Figure 3.3 displays the power curves of VCST against 2DF LRT. When the degrees of freedom under the null and under the alternative (i.e. $p_1 - p_0$) only differ by 2, the curve almost coincides with the 45 degree line indicating that the two tests have similar power in detecting the departure from a quadratic DLM. When the difference in degrees of freedom increases, VCST becomes more powerful compared to LRT. The example demonstrates that VCST can be applied to test a nested DLM against a more general one but it is not ideal in the situations where the difference between the number of parameters under the alternative and the number of parameters under the null is small and a LRT with low DF is sufficiently powerful.

Table 3.1: Empirical type I error and power of likelihood ratio test (LRT) and variance component score test (VCST) for testing a constrained DLMs against an unconstrained alternative based on 1000 repetitions when signal-to-noise level is moderate ($\delta = 1$) with significance level 0.05.

Fitted DLM / True DLM	Linear		Quadratic		Cubic	
	LRT	VCST	LRT	VCST	LRT	VCST
Linear	*0.057*	*0.040*	-	-	-	-
Quadratic	0.234	0.584	*0.038*	*0.039*	-	-
Cubic	0.282	0.611	0.276	0.612	*0.045*	*0.047*
Unstructured	0.317	0.640	0.312	0.641	0.319	0.572

*Italicized numbers are type I error rates at level 0.05.

Table 3.2: P-values obtained from likelihood ratio test (LRT) (with $L + 1 - p$ degrees of freedom) and variance component score test (VCST) for testing a specified distributed lag model (DLM) against an unconstrained DLM in association of daily PM_{10} measurements with cardiovascular death and non-accidental mortality in Chicago, Illinois from 1987 to 2000 using the National Mortality, Morbidity, and Air Pollution Study (NMMAPS) data where the maximum number of lags L is fixed at 14 days and p denote the number of basis functions of a DLM.

DLM	CVD[1]		Mortality	
	LRT	VCST	LRT	VCST
Linear ($p = 2$)	0.002	0.002	0.009	0.005
Quadratic ($p = 3$)	0.003	0.007	0.017	0.015
Cubic ($p = 4$)	0.020	0.019	0.055	0.013
4DF Polynomial ($p = 5$)	0.040	0.032	0.084	0.019
5DF Polynomial ($p = 6$)	0.038	0.033	0.286	0.150

[1]CVD: cardiovascular deaths

Table 3.3: Empirical type I error and power of likelihood ratio test (LRT) and variance component score test (VCST) for testing a constrained distributed lag model (DLM) against an unconstrained alternative based on 1000 repetitions when signal-to-noise level is weak ($\delta = 0.8$) with significance level 0.05.

Fitted DLM / True DLM	Linear		Quadratic		Cubic	
	LRT	VCST	LRT	VCST	LRT	VCST
Linear	*0.051*	*0.046*	-	-	-	-
Quadratic	0.153	0.359	*0.054*	*0.037*	-	-
Cubic	0.156	0.384	0.136	0.370	*0.054*	*0.044*
Unstructured	0.172	0.393	0.163	0.380	0.185	0.337

*Italicized numbers are type I error rates at level 0.05.

Table 3.4: Empirical type I error and power of likelihood ratio test (LRT) and variance component score test (VCST) for testing a constrained distributed lag model (DLM) against an unconstrained alternative based on 1000 repetitions when signal-to-noise level is strong ($\delta = 1.25$) with significance level 0.05.

Fitted DLM / True DLM	Linear		Quadratic		Cubic	
	LRT	VCST	LRT	VCST	LRT	VCST
Linear	*0.047*	*0.051*	-	-	-	-
Quadratic	0.413	0.837	*0.049*	*0.043*	-	-
Cubic	0.448	0.813	0.442	0.831	*0.059*	*0.055*
Unstructured	0.493	0.851	0.504	0.849	0.498	0.793

*Italicized numbers are type I error rates at level 0.05.

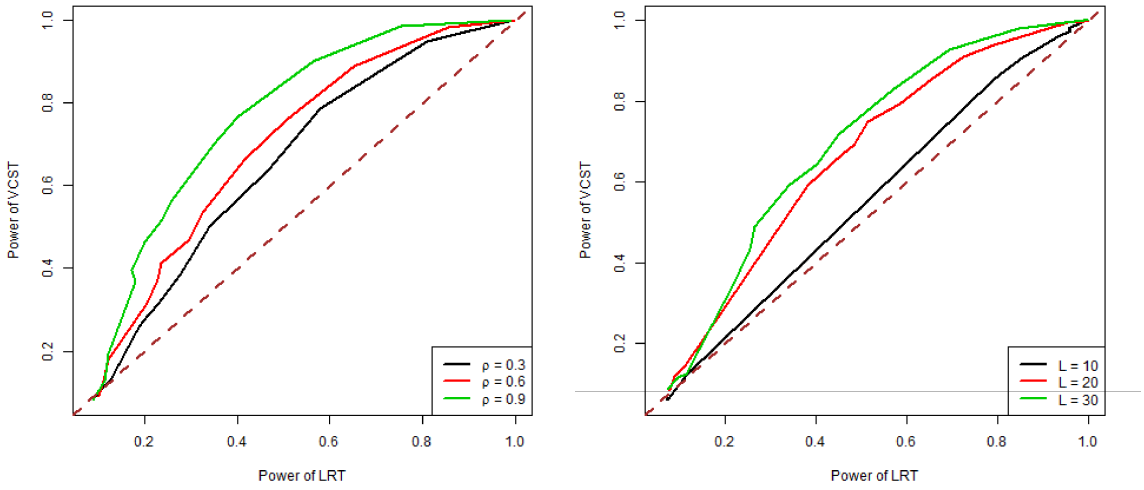

Figure 3.1: Plots of the power of variance component score test (VCST) against the power of likelihood ratio test (LRT) for testing a constrained distributed lag model (DLM) against an unconstrained alternative with three different first order autocorrelation levels (ρ) for predictor series (left panel) and three different maximum number of lags (L) for predictor series (right panel) based on 1000 repetitions.

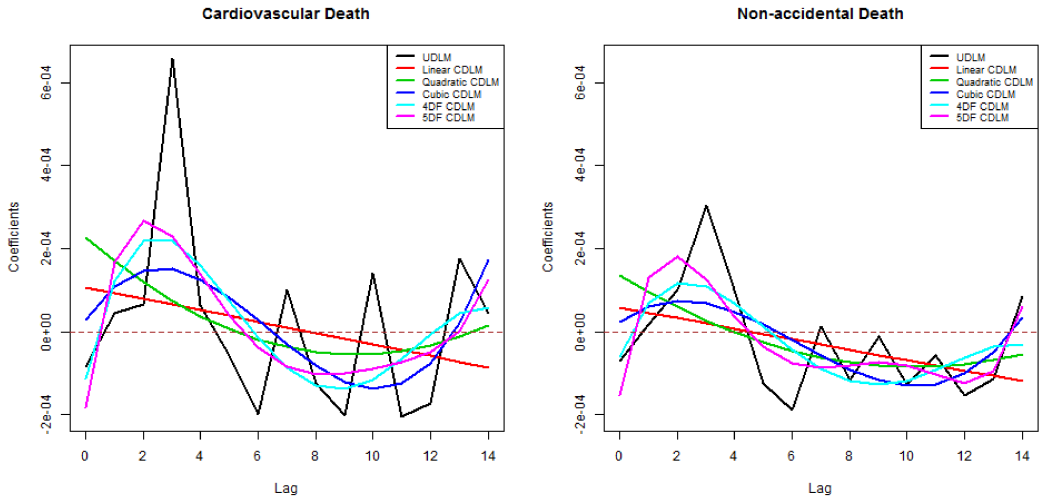

Figure 3.2: Plots of estimated distributed lag functions for cardiovascular death (left panel) and non-accidental mortality (right panel) in association with PM_{10} in Chicago, Illinois from 1987 to 2000 using the National Mortality, Morbidity, and Air Pollution Study (NMMAPS) data.

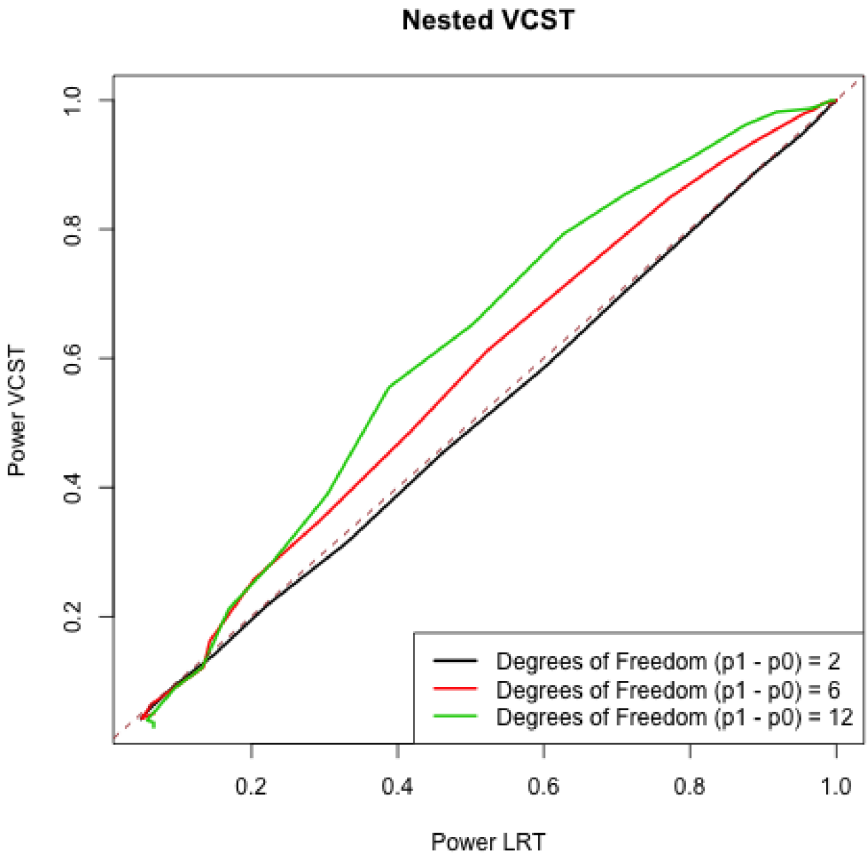

Figure 3.3: Plots of the power of variance component score test (VCST) against the power of likelihood ratio test (LRT) for testing a quadratic distributed lag model (DLM) ($p_0 = 2$) against a higher-degree polynomial ($p_1 = 4, 8, 14$) distributed lag model (DLM) based on 1000 repetitions.

CHAPTER 4

Distributed Lag Models with Two Pollutants

4.1 Introduction

The association between air pollution and adverse health outcomes has been an important public health concern and a topic of extensive research in environmental epidemiology [Pope and Dockery, 2006, Brook et al., 2010]. While long-term studies focus on estimating the effects of exposure to air pollution by following cohorts over years to decades [Pope, 2007], short-term studies focus on examining the relationship between daily counts of events related to mortality and morbidity in a geographically referenced population and ambient exposure levels. The short-term or acute effects of air pollution exposure on health outcomes, such as mortality and cardiovascular events, have been widely studied [Pope et al., 1995, Katsouyanni et al., 1997, Bell et al., 2004b, Pope and Dockery, 2006]. Dominici et al. [2006] estimated the risks of cardiovascular and respiratory hospital admissions associated with short-term exposure to fine particulate air pollution in 204 U.S urban counties. However, most studies so far have considered adverse health effects of exposure to a single pollutant [Dominici et al., 2010]. When ambient data are available on multiple pollutants, it is standard practice to analyze their effects one at a time by fitting multiple single pollutant models. For instance, Zeka and Schwartz [2004] assessed the individual effects of multiple pollutants using the NMMAPS data. However, the health burden from simultaneous exposure to multiple pollutants may differ from the sum of individual effects. A multi-pollutant approach that considers the joint effects of chemical mixtures of exposures is likely to yield more accurate assessment of health risk [Billionnet et al., 2012,

Coull et al., 2015]. One pollutant may modify the effects of other pollutants and the mode of action can be synergistic or antagonistic [Mauderly, 1993, Greenland, 1993]. It is often desirable to consider the interaction effects between two pollutants in a joint model.

A variety of approaches have been proposed to estimate the health effects of multiple pollutants [Sun et al., 2013]. The most straightforward approach is a multiple regression model with a main effect for each pollutant and a two-way cross-product linear interaction term for each pair of pollutants [Dominici et al., 2010]. Penalized regression methods such as LASSO [Tibshirani, 1996] and elastic net [Zou and Hastie, 2005] can be employed to identify a small subset of individual pollutants and interaction terms that are most notably associated with the outcome. In several studies, PCA have been used as a dimension reduction tool prior to multi-pollutant modeling [Arif and Shah, 2007, Qian et al., 2004]. Tree-based approaches such as CART are useful to account for higher-order and nonlinear interactions [Hu et al., 2008]. The DSA algorithm [Sinisi and van der Laan, 2004] allows users to specify the constraints on polynomial function form of exposure and the order of interaction. Bobb et al. [2013] used reduced hierarchical models to estimate health effects of simultaneous exposure to multiple pollutants by allowing for nonlinear associations of each of the pollutants and their interaction via natural splines. In a health effects analysis of mixtures, Bayesian kernel machine regression (BKMR) [Bobb et al., 2014] was developed to flexibly estimate the exposure-response relationship and facilitate inference on the strength of the association between individual pollutants and health outcomes. These dimension reduction or variable selection techniques typically consider cross-sectional data measured at a single time point. Very few methods so far consider the problem of capturing the lagged effect of two pollutants and their potential interactions over a biologically meaningful time period. Recent time-series studies reported that models with only single-day pollution measures might underestimate risk when there is a cumulative effect of air pollution over a time window preceding a health event [Schwartz, 2000, Roberts, 2005].

DLMs are a class of models often used to simultaneously include lagged measures of concentration levels of an ambient air pollutant. Parametric DLM assumes that the lag effect coefficients lie on a function of the lags, such as lower-degree polynomials [Almon, 1965] or a spline [Corradi, 1977]. Generalized additive DLM [Zanobetti et al., 2000] uses

penalized regression splines [Marx and Eilers, 1998] to represent the DL function in a more flexible manner. BDLM [Welty et al., 2009] was proposed to incorporate prior knowledge about the DL function through specification of the prior variance-covariance matrix of lag coefficients. Most of the discussion regarding DLM has been in the context of associating a health event time series with an exposure time series corresponding to a single pollutant. Extensions to higher dimensions include bivariate DLM [Muggeo, 2007] (BiDLM) and high degree DLM (HDDLM) [Heaton and Peng, 2014]. BiDLM was proposed to analyze the joint effect of temperature and PM_{10} on mortality in Sicily, South Italy. The temperature main effect and PM_{10} main effect were modeled in the same way as parametric DLM with two separate sets of basis functions. Tensor products of the two were employed to characterize the DL surface for temperature-PM_{10} interaction. BiDLM is the only previous method that has been proposed to account for lag effects in two-dimensional settings. The HDDLM paper extended DLM to incorporate higher-order interactions between lagged predictors corresponding to a single exposure, using a Gaussian process prior on lag coefficients to account for predictor collinearity and as a dimension reduction tool. However, this approach still estimates the cumulative lagged effects of a single pollutant. The goal of this chapter is to propose DLM with two pollutants that characterize interaction between the two exposure time series on a health outcome time series in a meaningful manner. We try to borrow ideas from the classical interaction analysis literature to ask the scientifically relevant question whether the two distributed lag coefficients/profiles for pollutant 1 will be significantly different when pollutant 2 is fixed at the lowest quartile versus when it is fixed at the highest quartile. Our main tools are dimension reduction and shrinkage to capture the interacting lag profiles in a parsimonious manner.

Tukey's one degree-of-freedom test for non-additivity [Tukey, 1949] is a parsimonious approach to model the interaction term as a scaled product of its corresponding main effects. Tukey's model implicitly assumes that interaction term can only be present when both the main effects are present. Tukey's single parameter form of interaction has recently been adopted for testing gene-environment interaction and gene-gene interaction to achieve higher statistical power [Chatterjee et al., 2006, Maity et al., 2009, Wang et al., 2015]. Ko et al. [2014] proposed to model gene-environment interaction using a shrinkage

estimator that combines the estimates from Tukey's model and the estimates from the saturated interaction model. The rationale is to simultaneously preserve the robustness when the underlying truth departs from Tukey's interaction structure and gain efficiency from the parsimony of Tukey's model when the model is plausible. One can conceptualize a Tukey type interaction structure for DLMs where the main effects are described by DLMs and the interaction is a scaled product of the main effects. In this chapter, we extend Tukey's model to DLMs where the interaction is parameterized as a scaled product of two DLM main effects. We will consider estimation and inference under such an extension in both frequentist and Bayesian framework.

In addition to the Tukey structure DLMs, we also propose a Bayesian constrained DLM (BCDLM) approach to characterize the joint effect of two pollutants. We use a set of B-spline [Beatty and Barsky, 1987] basis functions to model the DL function of each exposure. The tensor-product of the two basis sets are used to model the DL surface of the interaction between two exposures. Hierarchical structure of the BCDLM defined via hyperprior specification on the DL coefficients enables shrinkage and avoids overfitting. Instead of shrinking all main effects and interaction effects toward zero, we set a pre-specified parametric DLM as the shrinkage target in this approach. BCDLM is able to strike a desirable bias-variance tradeoff in a data-adaptive way with a fit that lies in between a fully flexible fit and a constrained parametric DLM fit.

The rest of the chapter is organized as follows. In Section 4.2, we first review the existing methods and their variations, including (1) unconstrained DLM (UDLM), (2) bivariate DLM (BiDLM), and (3) two-dimensional HDDLM (BiHDDLM). We then introduce the proposed new methods (1) Tukey's DLM (TDLM), (2) Bayesian Tukey's DLM (BTDLM), and (3) Bayesian constrained DLM (BCDLM). Our parameters of interest are the marginal effects of one pollutant when the other pollutant level is fixed at certain values, after accounting for potential interactions. Mathematically, this composite parameter can be represented as a function of main effects and interaction parameters. In Section 4.3, we conduct a simulation study to evaluate the operating characteristics of the six different methods and come up with a recommendation and guideline for practitioners. In Section 4.4, we illustrate the methods by analyzing data from the NMMAPS to estimate the lagged

effects of air PM_{10} and O_3 concentration on mortality in Chicago, Illinois from 1987 to 2000. We conclude with a discussion in Section 4.5.

There are several novel features of this chapter. The first is to extend DLM to handle two pollutants. We attempt to characterize the changes in DL function corresponding to one exposure when the other is fixed at different values. Extending the well-known Tukey's model for interaction to DLM is another innovation. Finally, beyond the class of Tukey's model, using data adaptive shrinkage to allow for an unconstrained interaction model to shrink towards a parametric DLM structure is a new contribution to the literature. More broadly, beyond air pollution epidemiology, the chapter posits new ideas for thinking about interaction structures between a pair of time series predictors with potential lagged effects on an outcome time series.

4.2 Methods

Let x_{1t} denote the first exposure measured at time t (e.g. PM_{10}), x_{2t} denote the second exposure measured at time t (e.g. O_3), y_t denote the response measured at time t (e.g. daily mortality count), and $\boldsymbol{z}_t$ denote the vector of covariates at time t, such as temperature and humidity, in addition to a constant 1 corresponding to the intercept parameter. Let T be the length of the time series, L_1 and L_2 be the maximum number of lags considered for the first and second exposure, respectively. In addition, we denote $\boldsymbol{X}_{1t} = (x_{1t}, \cdots, x_{1,t-L_1})^\top$, $\boldsymbol{X}_{2t} = (x_{2t}, \cdots, x_{2,t-L_2})^\top$, and $\boldsymbol{X}_{It} = \boldsymbol{X}_{1t} \otimes \boldsymbol{X}_{2t}$ where $\otimes$ is the Kronecker product and the $(L_1+1)(L_2+1)$ elements in $\boldsymbol{X}_{It}$ refer to the two-way interaction terms between the two exposures. The log-linear Poisson DLM with all pairwise interactions between lagged measurements of the two exposures is described as

$$y_t \sim \text{Poisson}(\mu_t) \tag{4.1}$$

$$\begin{aligned} \log(\mu_t) &= \boldsymbol{z}_t^\top \boldsymbol{\alpha} + \boldsymbol{X}_{1t}^\top \boldsymbol{\beta}_1 + \boldsymbol{X}_{2t}^\top \boldsymbol{\beta}_2 + \boldsymbol{X}_{It}^\top \boldsymbol{\gamma} \\ &= \boldsymbol{z}_t^\top \boldsymbol{\alpha} + \sum_{i=0}^{L_1} x_{1,t-i}\beta_{1i} + \sum_{j=0}^{L_2} x_{2,t-j}\beta_{2j} + \sum_{i=0}^{L_1}\sum_{j=0}^{L_2} \gamma_{ij} x_{1,t-i} x_{2,t-j} \end{aligned} \tag{4.2}$$

where $\boldsymbol{\alpha}$ represents the effect of covariates, $\boldsymbol{\beta}_1 = (\beta_{10}, \cdots, \beta_{1L_1})^\top$ is the (L_1+1)-vector of lagged main effects of the first exposure, $\boldsymbol{\beta}_2 = (\beta_{20}, \cdots, \beta_{2L_2})^\top$ is the (L_2+1)-vector of lagged main effects of the second exposure, and $\boldsymbol{\gamma} = \text{vec}(\boldsymbol{\Gamma}) = (\gamma_{00}, \gamma_{01}, \cdots, \gamma_{L_1L_2})^\top$ such that $\boldsymbol{\Gamma}$ is the $(L_1+1) \times (L_2+1)$ matrix of interaction effects. Our primary goal is to estimate main effects $\boldsymbol{\beta}_1$ and $\boldsymbol{\beta}_2$ and interaction effects $\boldsymbol{\gamma}$. For simplicity, we leave out $\boldsymbol{z}_t^\top \boldsymbol{\alpha}$ in subsequent presentation.

Remark: (4.1) and (4.2) model the conditional mean response at a time point t given the current and past measurements of the two exposures. Nonnull interaction effect in (4.2) implies that the lagged effects of the first exposure depend on the level of the second exposure, and vice versa. It is noted that the interaction effects are not symmetric in (4.2), namely $\gamma_{ij} \neq \gamma_{ji}$ for $i \neq j$. Naturally, the quantity of interest is the marginal effect of one exposure at a certain lag, given the other exposure fixed at a certain level such as median or a specified quantile. Algebraically, if we fix the second exposure at x_2^* across all lags, the marginal lag effects of the first exposure at lag i can be written as $\beta_{1i}^* = \beta_{1i} + x_2^* \sum_{j=0}^{L_2} \gamma_{ij}$ for $i = 0, \cdots, L_1$. The vector representation is

$$\boldsymbol{\beta}_1^m(x_2^*) = \boldsymbol{\beta}_1 + x_2^* \cdot \boldsymbol{\Gamma}\mathbf{1} \tag{4.3}$$

where $\mathbf{1}$ is a vector of 1s. Similarly, if we fix the first exposure at x_1^*, the marginal lag effects of the second exposure at lag j can be written as $\beta_{2j}^* = \beta_{2j} + x_1^* \sum_{i=0}^{L_1} \gamma_{ij}$ for $j = 0, \cdots, L_2$ with vector representation

$$\boldsymbol{\beta}_2^m(x_1^*) = \boldsymbol{\beta}_2 + x_1^* \cdot \boldsymbol{\Gamma}^\top \mathbf{1}.$$

Throughout the rest of this chapter, we will summarize the estimates of $\boldsymbol{\beta}_1$, $\boldsymbol{\beta}_2$, and $\boldsymbol{\gamma} = \text{vec}(\boldsymbol{\Gamma})$ based on the above expressions and interpret the marginal lagged effects of one exposure when the other exposure is fixed across all lags.

4.2.1 Existing Methods

4.2.1.1 Unconstrained Distributed Lag Model (UDLM)

As the name suggests, UDLM does not impose any constraints on coefficients $\boldsymbol{\psi} = (\boldsymbol{\beta}_1^\top, \boldsymbol{\beta}_2^\top, \boldsymbol{\gamma}^\top)^\top$ in (4.2). The UDLM coefficients can be simply estimated via MLE.

$$\hat{\boldsymbol{\psi}}_{UDLM} = \arg\max_{\boldsymbol{\psi}} \sum_{t=1}^{T} [y_t \boldsymbol{X}_t^\top \boldsymbol{\psi} - e^{\boldsymbol{X}_t^\top \boldsymbol{\psi}} - \log(y_t!)],$$

where $\boldsymbol{X}_t = (\boldsymbol{X}_{1t}^\top, \boldsymbol{X}_{2t}^\top, \boldsymbol{X}_{It}^\top)^\top$. Standard frequentist inference based on large sample theory of MLEs can be drawn subsequently. However, due to the collinearity between serially measured exposure levels and the large number of parameters (i.e. $L_1 + L_2 + 2$ main effect terms and $(L_1 + 1)(L_2 + 1)$ interaction terms), the lagged effect estimates could be less efficient with inflated variance [Farrar and Glauber, 1967] and the estimated DL functions could be highly variable.

4.2.1.2 Bivariate Distributed Lag Model (BiDLM)

Parametric DLM imposes a smooth structure on lagged effect coefficients by assuming each lag coefficient to be a linear combination of known basis functions measured at its lag index. BiDLM extends this configuration to two-dimensional scenarios. Assume $B_{11}(\cdot), \cdots, B_{1p_1}(\cdot)$ are the p_1 basis functions applied to $\boldsymbol{\beta}_1$ and $B_{21}(\cdot), \cdots, B_{2p_2}(\cdot)$ are the p_2 basis functions applied to $\boldsymbol{\beta}_2$. Main effect coefficients are assumed to be of the form $\beta_{1i} = \sum_{m=1}^{p_1} B_{1m}(i)\theta_{1m}$ for $i = 0, \cdots, L_1$ and $\beta_{2j} = \sum_{n=1}^{p_2} B_{2n}(j)\theta_{2n}$ for $j = 0, \cdots, L_2$ where $\{\beta_{1i}\}$ and $\{\beta_{2j}\}$ are elements of $\boldsymbol{\beta}_1$ and $\boldsymbol{\beta}_2$, respectively, and $\{\theta_{1m}\}$ and $\{\theta_{2n}\}$ are free parameters to be estimated. In order to smooth the interaction surface, Muggeo [2007] utilizes tensor products of marginal basis functions [Dierckx, 1995, De Boor et al., 1978]. The element corresponding to the interaction between $x_{1,t-\ell_1}$ and $x_{2,t-\ell_2}$ can be expressed as $\gamma_{ij} = \sum_{m=1}^{p_1} \sum_{n=1}^{p_2} B_{1m}(i) B_{2n}(j) \theta_{Imn}$.

Define $\boldsymbol{C}_1$ as a $(L_1 + 1) \times p_1$ transformation matrix [Gasparrini et al., 2010] where the element $(i+1, m)$ is $B_{1m}(i)$ and similarly define $\boldsymbol{C}_2$ as a $(L_2+1) \times p_2$ transformation matrix where the element $(j + 1, n)$ is $B_{2n}(j)$. Denote $\boldsymbol{\theta}_1 = (\theta_{11}, \cdots, \theta_{1p_1})$, $\boldsymbol{\theta}_2 = (\theta_{21}, \cdots, \theta_{2p_2})$,

and $\boldsymbol{\theta}_I = (\theta_{I11}, \theta_{I12}, \cdots, \theta_{Ip_1p_2})$. The BiDLM coefficients can be written in terms of the free parameters to be estimated as

$$\boldsymbol{\beta}_1 = \boldsymbol{C}_1\boldsymbol{\theta}_1, \boldsymbol{\beta}_2 = \boldsymbol{C}_2\boldsymbol{\theta}_2, \boldsymbol{\gamma} = (\boldsymbol{C}_1 \otimes \boldsymbol{C}_2)\boldsymbol{\theta}_I \tag{4.4}$$

where $\otimes$ is the Kronecker product. The free parameters $\boldsymbol{\theta}_1$, $\boldsymbol{\theta}_2$, and $\boldsymbol{\theta}_I$ can be obtained by maximizing the likelihood function

$$\sum_{t=1}^{T}[y_t[\boldsymbol{W}_{1t}^\top\boldsymbol{\theta}_1 + \boldsymbol{W}_{2t}^\top\boldsymbol{\theta}_2 + \boldsymbol{W}_{It}^\top\boldsymbol{\theta}_I]^\top - e^{\boldsymbol{W}_{1t}^\top\boldsymbol{\theta}_1+\boldsymbol{W}_{2t}^\top\boldsymbol{\theta}_2+\boldsymbol{W}_{It}^\top\boldsymbol{\theta}_I} - \log(y_t!)]$$

where $\boldsymbol{W}_{1t} = \boldsymbol{C}_1^\top\boldsymbol{X}_{1t}$, $\boldsymbol{W}_{2t} = \boldsymbol{C}_2^\top\boldsymbol{X}_{2t}$, and $\boldsymbol{W}_{It} = (\boldsymbol{C}_1 \otimes \boldsymbol{C}_2)^\top\boldsymbol{X}_{It}$. Let $\boldsymbol{\Theta} = (\boldsymbol{\theta}_1^\top, \boldsymbol{\theta}_2^\top, \boldsymbol{\theta}_I^\top)^\top$, a vector of length $p_1 + p_2 + p_1p_2$ and $\boldsymbol{C} = \text{diag}[\boldsymbol{C}_1, \boldsymbol{C}_2, \boldsymbol{C}_1 \otimes \boldsymbol{C}_2]$, a $[(L_1 + 1) + (L_2 + 1) + (L_1 + L_2 + 2)] \times [p_1 + p_2 + p_1p_2]$ matrix. The BiDLM estimator can be written as $\hat{\boldsymbol{\psi}}_{BiDLM} = \boldsymbol{C}\hat{\boldsymbol{\Theta}}$ and $\text{Cov}(\hat{\boldsymbol{\psi}}_{BiDLM}) = \boldsymbol{C}\text{Cov}(\hat{\boldsymbol{\Theta}})\boldsymbol{C}^\top$.

4.2.1.3 Two-dimensional High Degree Distributed Lag Model (BiHDDLM)

HDDLM [Heaton and Peng, 2014] was originally proposed to incorporate higher-order interactions between lagged predictors in single-pollutant settings. We modify the underlying structure of HDDLM to accommodate two-pollutant scenarios, considering up to two-way interactions with the underlying model exactly stated in (4.2). The modeling strategy is to construct a predictive process prior from the assumed Gaussian process prior on the lag coefficients as a dimension reduction tool to handle the collinearity between time-series exposure measurements. Moreover, a conditioning technique is incorporated to ensure that the lagged coefficients at larger lags smoothly approach 0.

An important step to specify the predictive process is to choose the pseudo knot locations for $\boldsymbol{\beta}_1$, $\boldsymbol{\beta}_2$, and $\boldsymbol{\gamma}$ in (4.2) to approximate the parent process. Consider knot vectors $\boldsymbol{\beta}_1^* = \{\beta_{1\ell_i^*}\}$, $\boldsymbol{\beta}_2^* = \{\beta_{2\ell_j^*}\}$, and $\boldsymbol{\gamma}^* = \{\gamma_{\ell_k^*}\}$ where $\ell_i^* \in \mathbb{R}^1$, $\ell_j^* \in \mathbb{R}^1$, and $\ell_k^* \in \mathbb{R}^2$ are artificially chosen internal knot locations for $\boldsymbol{\beta}_1$, $\boldsymbol{\beta}_2$, and $\boldsymbol{\gamma}$, respectively and $R_1 := \dim(\boldsymbol{\beta}_1^*) < L_1 + 1$, $R_2 := \dim(\boldsymbol{\beta}_2^*) < L_2 + 1$, $R_I := \dim(\boldsymbol{\gamma}^*) < (L_1 + 1)(L_2 + 1)$. $\boldsymbol{\beta}_1$, $\boldsymbol{\beta}_2$, and $\boldsymbol{\gamma}$ can be mapped from $\boldsymbol{\beta}_1^*$, $\boldsymbol{\beta}_2^*$, and $\boldsymbol{\gamma}^*$ using the predictive process interpolator [Banerjee et al.,

2008] and the number of parameters reduces from $(L_1+1)+(L_2+1)+(L_1+1)(L_2+1)$ to $R_1+R_2+R_I$.

Distributed lag functions in $\mathbb{R}^1$ and the distributed lag surface in $\mathbb{R}^2$ are subject to constraints and lag effects should decrease to zero as the lag time increases. Consider two large maximum numbers of lag $M_1 > L_1$ and $M_2 > L_2$ and corresponding expanded vectors of coefficients $\boldsymbol{\beta}_1^{(e)} = (\boldsymbol{\beta}_1^\top, \boldsymbol{\beta}_1^{+\top})^\top$, $\boldsymbol{\beta}_2^{(e)} = (\boldsymbol{\beta}_2^\top, \boldsymbol{\beta}_2^{+\top})^\top$, and $\boldsymbol{\gamma}^{(e)} = (\boldsymbol{\gamma}^\top, \boldsymbol{\gamma}^{+\top})^\top$ where $\boldsymbol{\beta}_1^+$, $\boldsymbol{\beta}_2^+$, and $\boldsymbol{\gamma}^+$ are the additional lag coefficients with lengths $M_1 - L_1 - 1$, $M_2 - L_2 - 1$, and $M_1M_2 - (L_1+1)(L_2+1)$, respectively. Conditioning on L_1 and L_2, the distributions for $\boldsymbol{\beta}_1$, $\boldsymbol{\beta}_2$, and $\boldsymbol{\gamma}$ reduce to finding the conditional distribution of $[\boldsymbol{\beta}_1|\boldsymbol{\beta}_1^+]$, $[\boldsymbol{\beta}_2|\boldsymbol{\beta}_2^+]$, and $[\boldsymbol{\gamma}|\boldsymbol{\gamma}^+]$.

Now we combine the predictive process in conjunction with conditioning on the expanded vectors. Let each of $\boldsymbol{\beta}_1^*|\boldsymbol{\beta}_1^+$, $\boldsymbol{\beta}_2^*|\boldsymbol{\beta}_2^+$, and $\boldsymbol{\gamma}^*|\boldsymbol{\gamma}^+$ follows a zero-mean Gaussian process with isotropic Matérn covariance function. The corresponding prior specifications are

$$\boldsymbol{\beta}_1^*|\boldsymbol{\beta}_1^+ = \mathbf{0}, \sigma_1^2, \nu_1, \psi_1 \sim N(\mathbf{0}, \{\sigma_1^2 \mathcal{M}_{\nu_1}(||\ell_i - \ell_{i'}||; \psi_1)\}_{i,i'})$$

$$\boldsymbol{\beta}_2^*|\boldsymbol{\beta}_2^+ = \mathbf{0}, \sigma_2^2, \nu_2, \psi_2 \sim N(\mathbf{0}, \{\sigma_2^2 \mathcal{M}_{\nu_2}(||\ell_j - \ell_{j'}||; \psi_2)\}_{j,j'})$$

$$\boldsymbol{\gamma}^*|\boldsymbol{\gamma}^+ = \mathbf{0}, \sigma_I^2, \nu_I, \psi_I \sim N(\mathbf{0}, \{\sigma_I^2 \mathcal{M}_{\nu_I}(||\ell_k - \ell_{k'}||; \psi_I)\}_{k,k'})$$

where $\mathcal{M}_\nu(||\cdot||;\psi)$ is the Matérn correlation function with smoothness parameter ν and decay parameter ψ. By conditioning on the additional lag coefficients equal to zero, BiHD-DLM ensures that the lag coefficients decrease to zero as the lag increases. Details of the procedure to construct the predictive process interpolator are presented in Appendix 4.6.1.

The smoothness parameter ν in $\mathcal{M}_\nu$ controls the smoothness of DL functions and DL surface. Gneiting et al. [2012] indicates that estimating the smoothness parameter ν is notoriously difficult. Following Heaton and Peng [2014], we fix $\nu_1 = \nu_2 = \nu_I = 3$ to allow for the resulting distributed lag curves for main effects and the distributed lag surface for interaction effect to be twice differentiable. Zhang [2004] Heaton and Peng Heaton and Peng [2014] showed that weakly consistent estimators for ψ_1, ψ_2, and ψ_I do not exist, implying that the decay parameters can be fixed *a priori* without sacrificing flexibility as long as non-informative priors are specified for σ_1^2, σ_2^2, and σ_I^2 [Du et al., 2009, Zhang

and Wang, 2010]. We chose $\psi_1 = \psi_2 = \psi_I = 0.6$ based on the guidelines provided by Heaton and Peng [2014] and Datta et al. [2015]. To complete the model specification, we assume a vague prior on each of σ_1^2, σ_2^2, and σ_I^2 as an inverse gamma distribution with shape parameter equal to 2 and scale parameter equal to 1. Posterior draws are obtained using well-established Markov chain Monte Carlo (MCMC) techniques [Gamerman and Lopes, 2006].

4.2.2 Proposed Methods

4.2.2.1 Tukey's Distributed Lag Model (TDLM)

The underlying motivation for Tukey's model for interaction is through a latent variable framework [Chatterjee et al., 2006]. Suppose we define a surrogate variable for each exposure that aggregates the temporal lagged effect of the exposure through weighted sum at time t. Namely,

$$s_{1t} = \sum_{i=0}^{L_1} w_{1i} x_{1,t-i}, s_{2t} = \sum_{j=0}^{L_2} w_{2j} x_{2,t-j}. \tag{4.5}$$

If we assume that the association between y_t, $\boldsymbol{X}_{1t}$ and $\boldsymbol{X}_{2t}$ is through the interaction model

$$\log(E[y_t]) = \mu_0 + \mu_1 s_{1t} + \mu_2 s_{2t} + \mu_I s_{1t} s_{2t}. \tag{4.6}$$

Substituting (4.5) in (4.6), we can obtain

$$\begin{aligned}
\log(E[y_t]) &= \mu_0 + \sum_{i=0}^{L_1} \mu_1 w_{1i} x_{1,t-i} + \sum_{j=0}^{L_2} \mu_2 w_{2j} x_{2,t-j} + \sum_{i=0}^{L_1} \sum_{j=0}^{L_2} \mu_I w_{1i} w_{2j} x_{1,t-i} x_{2,t-j} \\
&= \mu_0 + \sum_{i=0}^{L_1} \beta_{1i} x_{1,t-i} + \sum_{j=0}^{L_2} \beta_{2j} x_{2,t-j} + \sum_{i=0}^{L_1} \sum_{j=0}^{L_2} \gamma_{ij} x_{1,t-i} x_{2,t-j}
\end{aligned}$$

where $\beta_{1i} = \mu_1 w_{1i}$, $\beta_{2j} = \mu_2 w_{2j}$, and $\gamma_{ij} = \mu_I w_{1i} w_{2j}$. Note that we can express the interaction coefficient as $\gamma_{ij} = \beta_{1i}\beta_{2j}(\frac{\mu_I}{\mu_1 \mu_2})$, a scaled product of the corresponding main-effect coefficients. This motivates the use of Tukey's style interaction in our context. Estimating the lagged effects is the same as estimating the relative weights to combine the exposure lagged measurements into a summary surrogate variable. To extend the classical

Tukey's interaction structure to DLMs, we now assume that the main effects are specified in the same way as BiDLM with constrained parameterization such that $\boldsymbol{\beta}_1 = \boldsymbol{C}_1\boldsymbol{\theta}_1$ and $\boldsymbol{\beta}_2 = \boldsymbol{C}_2\boldsymbol{\theta}_2$ as in (4.4). In matrix form, the interaction coefficients can be expressed under Tukey's model as

$$\boldsymbol{\gamma} = \eta \cdot (\boldsymbol{\beta}_1 \otimes \boldsymbol{\beta}_2) = (\boldsymbol{C}_1 \otimes \boldsymbol{C}_2)[\eta(\boldsymbol{\theta}_1 \otimes \boldsymbol{\theta}_2)].$$

Note that the interaction structure corresponding to TDLM is a special case of BiDLM with $\boldsymbol{\theta}_I = \eta(\boldsymbol{\theta}_1 \otimes \boldsymbol{\theta}_2)$. The number of parameters used for modeling the interaction effect reduces from p_1p_2 to 1. The free parameters $\boldsymbol{\theta}_1$, $\boldsymbol{\theta}_2$, and η can be estimated by maximizing the likelihood function

$$\sum_{t=1}^{T}\{y_t[\boldsymbol{W}_{1t}^\top\boldsymbol{\theta}_1+\boldsymbol{W}_{2t}^\top\boldsymbol{\theta}_2+\eta\cdot\boldsymbol{W}_{It}^\top(\boldsymbol{\theta}_1\otimes\boldsymbol{\theta}_2)]-e^{\boldsymbol{W}_{1t}^\top\boldsymbol{\theta}_1+\boldsymbol{W}_{2t}^\top\boldsymbol{\theta}_2+\eta\cdot\boldsymbol{W}_{It}^\top(\boldsymbol{\theta}_1\otimes\boldsymbol{\theta}_2)}-\log(y_t!)\}. \quad (4.7)$$

TDLM is a nonlinear regression model where the objective function (4.7) involves product of the parameters. Linear approximation using first-order Taylor series expansion can be applied for parameter estimation and statistical inference [Bates and Watts, 1988]. However, empirically, we found that the approximation accuracy using first order approximation is poor and the asymptotic variance is far from empirical variance based on resampling [Efron, 1981]. We therefore consider an iterative approach for estimation. We first (a) fix $\boldsymbol{\theta}_1, \boldsymbol{\theta}_2$ and estimate η, (b) fix $\boldsymbol{\theta}_2, \eta$ and estimate $\boldsymbol{\theta}_1$, and then (c) fix $\boldsymbol{\theta}_1, \eta$ and estimate $\boldsymbol{\theta}_2$ until the solution converges (details provided in Appendix 4.6.2). The stopping criteria is when the percentage change in the value of likelihood (4.7) is smaller than a pre-specified margin (e.g. 10^{-6}). Since the value of the objective function decreases at each step, the solution is guaranteed to converge. We recognize that the likelihood function (4.7) is non-convex in terms of the parameters $\boldsymbol{\beta}_1$, $\boldsymbol{\beta}_2$, and η so the convergence to a global maximum is not guaranteed by this iterative procedure. However, in our numerical studies, when the main effects are bounded away from zero, the choice of various initial values did not affect the final parameter estimates. When at least one of the main effects are close to the null value, the parameter η is not identifiable and estimation instability occurs in these cases.

In the actual application context, one would expect that the interest in two pollutant DLM will originate only after observing significance in the marginal DLM models for each single pollutant and thus assuming at least one lag coefficient is nonzero for each pollutant is a reasonable assumption. For statistical inference, we consider a standard vanilla bootstrap by resampling observations with replacement to obtain bootstrap standard errors and confidence intervals.

4.2.2.2 Bayesian Tukey's Distributed Lag Model (BTDLM)

Under a Bayesian formulation of Tukeys DLM, the main effects are parametrically specified in the same way as in (4.4). The interaction effects are modelled in the spirit of TDLM. The distinction from the presentation in the previous section is that BTDLM allows departure from Tukey's interaction structure in a data-adaptive way. Instead of assuming that each interaction term is a scaled product of the corresponding main effects, BTDLM assumes that the scalar parameter can vary across different interaction terms through the following prior specification

$$\boldsymbol{\gamma} = \boldsymbol{\eta} \odot (\boldsymbol{\beta}_1 \otimes \boldsymbol{\beta}_2), \boldsymbol{\eta} \sim N(\mathbf{0}, \sigma^2 \boldsymbol{\Sigma}(\omega))$$

where $\boldsymbol{\eta} = (\eta_{00}, \eta_{01}, \cdots, \eta_{L_1 L_2})^\top$ is the vector of scalars, $\odot$ is the operator denoting element-wise multiplication, σ^2 is the common variance, and $\boldsymbol{\Sigma}$ is the correlation matrix parameterized by a single parameter $\omega > 0$. The correlation between η_{ij} and $\eta_{i^*j^*}$ is given by $\omega^{\sqrt{(i-i^*)^2+(j-j^*)^2}}$. The prior on $\boldsymbol{\eta}$ relaxes the strict specification of Tukey's interaction structure. The amount of departure from Tukey's model is controlled by the parameter ω. At one extreme, when $\omega = 0$, no structure is imposed on the interaction effects. The interaction coefficients are simply a reparametrization of the UDLM coefficients in (4.2). At the other extreme when $\omega = 1$, the model degenerates to TDLM and enforces the interaction coefficients to follow the Tukey's structure. When ω approaches 1, the correlation between neighboring coefficients tends to be larger, resulting in a smoother interaction surface.

To complete the model specifications, we assign $\boldsymbol{\theta}_1 \sim N(\mathbf{0}, 100^2 \boldsymbol{I})$ and $\boldsymbol{\theta}_2 \sim N(\mathbf{0}, 100^2 \boldsymbol{I})$ as vague priors for main effect coefficients. We assume a non-informative prior on variance

parameter [Gelman et al., 2006] $\sigma^2 \sim IG(a = 0.001, b = 0.001)$ where a and b are the shape and scale parameters of Inverse-Gamma (IG) distribution for common variance. To alleviate computational burden, we let ω have a discrete uniform prior rather than a continuous one. The marginal posterior density of $\boldsymbol{\beta}_1$, $\boldsymbol{\beta}_2$, and $\boldsymbol{\gamma}$ is not available in closed form. We use Metropolis Hastings algorithm [Hastings, 1970] within a Gibbs sampler [Geman and Geman, 1984] to approximate the posterior distribution and obtain the BTDLM estimator as the posterior mean and the corresponding highest posterior density (HPD) interval [Box and Tiao, 2011] as the corresponding credible interval. The full conditional distributions are presented in Appendix 4.6.3.

4.2.2.3 Bayesian Constrained Distributed Lag Model (BCDLM)

BiDLM is a fully parametric model. The dimension reduction from $(L_1 + 1) + (L_2 + 1) + (L_1 + 1)(L_2 + 1)$ parameters to $p_1 + p_2 + p_1 p_2$ parameters results in efficiency gain in estimation. However, the benefit can be counterbalanced by potential bias when the underlying structure for DL functions/surface is misspecified. There are various ways to allow departure from BiDLM and achieve bias-variance tradeoff. For example, the robust distributed lag models proposed by Chen et al. [2017]. We propose a BCDLM to shrink UDLM estimates (identical to BiDLM estimates with a full-rank transformation) in a smooth manner toward a pre-specified BiDLM.

Let $B^+_{11}(\cdot), \cdots, B^+_{1,L_1+1}(\cdot)$ be $L_1 + 1$ basis functions for the first exposure. For example, the B-spline basis functions of degree 3 (cubic) with intercept and $L_1 - 3$ equispaced internal knots positioned between 0 and L_1. Note that the basis functions describe the nonlinearity in DL function and exposure effect at each lag is still assumed to be linear. Let $\boldsymbol{T}_1$ be the corresponding $(L_1 + 1) \times (L_1 + 1)$ transformation matrix. Let $\boldsymbol{T}_2$ denote the square transformation matrix with dimension $(L_2 + 1) \times (L_2 + 1)$, constructed in a similar manner for the second exposure, and let the transformation matrix for the interaction parameter be $\boldsymbol{T}_I = (\boldsymbol{T}_1 \otimes \boldsymbol{T}_2)$ with dimension $(L_1 + 1)(L_2 + 1) \times (L_1 + 1)(L_2 + 1)$. If we applied the transformation operators $\boldsymbol{T}_1$, $\boldsymbol{T}_2$, and $\boldsymbol{T}_I$ to BiDLM, the resulting estimator would be identical to UDLM estimator since full-rank transformation on regression coefficients does not change the model fit. However, if we imposed shrinkage on regression coefficients

using $L2$ penalty, BiDLM estimator and UDLM estimator would be different since the shrinkage is employed in different parameter spaces. UDLM estimator can be viewed as choosing $B^+_{1m}(i) = I(m = i + 1)$ for $m = 1, \cdots, L_1 + 1$ and $B^+_{2n}(j) = I(n = j + 1)$ for $n = 1, \cdots, L_2 + 1$, where $I(\cdot)$ is an indicator function, corresponding to $\boldsymbol{T}_1 = \boldsymbol{I}$ and $\boldsymbol{T}_2 = \boldsymbol{I}$. Although the two sets of estimates share the same shrinkage target (i.e. zero line), the solution paths are different. If the basis functions selected for $\boldsymbol{T}_1$ and $\boldsymbol{T}_2$ are smooth, BiDLM with shrinkage leads to smooth estimates, whereas UDLM with shrinkage does not lead to smooth estimates.

Instead of shrinking the model coefficients toward 0, we consider shrinking them to a nonnull target, determined by the transformation matrices $\boldsymbol{C}_1$, $\boldsymbol{C}_2$, and $\boldsymbol{C}_I = (\boldsymbol{C}_1 \otimes \boldsymbol{C}_2)$ for BiDLM defined in (4.4). Without loss of generality, we only describe how to construct the nonnull shrinkage target for the first exposure. We first separate $\boldsymbol{T}_1$ into two parts - $\boldsymbol{C}_1$ and $\boldsymbol{C}_1^{\mathsf{c}}$ where $\boldsymbol{C}_1^\top \boldsymbol{C}_1^{\mathsf{c}} = \boldsymbol{0}$. We make use of this orthogonal decomposition to obtain $\boldsymbol{C}_1^{\mathsf{c}}$ the columns of which span the complementary column space of $\boldsymbol{C}_1$. $\boldsymbol{C}_1$ and $\boldsymbol{C}_1^{\mathsf{c}}$ defines the decomposition of transformation corresponding to shrinkage toward a pre-specified target and shrinkage toward 0, respectively. The orthogonal projection of $\boldsymbol{T}_1$ onto the complementary column space of $\boldsymbol{C}_1$ is given by

$$\boldsymbol{P}_1 = [\boldsymbol{I} - \boldsymbol{C}_1(\boldsymbol{C}_1^\top \boldsymbol{C}_1)^{-1}\boldsymbol{C}_1^\top]\boldsymbol{T}_1. \tag{4.8}$$

Using singular value decomposition (SVD), we can write

$$\boldsymbol{P}_1 = \boldsymbol{U}_1 \boldsymbol{D}_1 \boldsymbol{V}_1^\top$$

where $\boldsymbol{U}_1$ contains the columns of left-singular vectors, $\boldsymbol{D}_1$ is a diagonal matrix with eigenvalues of $\boldsymbol{P}_1$, and $\boldsymbol{V}_1$ contains the columns of right-singular vectors. Since the rank of $\boldsymbol{P}_1$ is $L_1 + 1 - p_1$, we can write $\boldsymbol{U}_1 = [\boldsymbol{U}_{11}\ \boldsymbol{U}_{12}]$ where $\boldsymbol{U}_{11}$ is a $(L_1 + 1) \times (L_1 + 1 - p_1)$ matrix with columns of singular vectors corresponding to nonzero eigenvalues in $\boldsymbol{D}_1$ and $\boldsymbol{U}_{12}$ is a $(L_1 + 1) \times p_1$ matrix with columns of singular vectors corresponding to the eigenvalues of 0. We consider $\boldsymbol{C}_1^{\mathsf{c}} = \boldsymbol{U}_{11}$. It is easy to show that $\boldsymbol{C}_1^\top \boldsymbol{C}_1^{\mathsf{c}} = \boldsymbol{0}$ and the p_1 columns of $\boldsymbol{C}_1$ and the $L_1 + 1 - p_1$ columns of $\boldsymbol{C}_1^{\mathsf{c}}$ span the entire $\mathbb{R}^{L_1+1}$. In other words, shrinkage through

the columns of $\boldsymbol{C}_1^{\mathsf{c}}$ defines BiDLM estimate as the shrinkage target. The complementary matrices $\boldsymbol{C}_2^{\mathsf{c}}$ and $\boldsymbol{C}_I^{\mathsf{c}}$ for the second exposure and interaction can be constructed using $\boldsymbol{C}_2$, $\boldsymbol{T}_2$ and $\boldsymbol{C}_I$, $\boldsymbol{T}_I$, respectively, in a similar way.

The likelihood corresponding to the above specification is given by

$$\boldsymbol{Y}|\boldsymbol{\beta}_1, \boldsymbol{\beta}_2, \boldsymbol{\gamma} \sim \text{Poisson}(e^{\boldsymbol{X}_1\boldsymbol{\beta}_1+\boldsymbol{X}_2\boldsymbol{\beta}_2+\boldsymbol{X}_I\boldsymbol{\gamma}})$$

where $\boldsymbol{Y} = (y_1, \cdots, y_T)^\top$, $\boldsymbol{X}_1 = (\boldsymbol{X}_{11}, \cdots, \boldsymbol{X}_{1T})^\top$, $\boldsymbol{X}_2 = (\boldsymbol{X}_{21}, \cdots, \boldsymbol{X}_{2T})^\top$, and $\boldsymbol{X}_I = (\boldsymbol{X}_{I1}, \cdots, \boldsymbol{X}_{IT})^\top$. The prior specifications corresponding to BCDLM parameters are

$$\boldsymbol{\beta}_1 = \boldsymbol{C}_1\boldsymbol{\theta}_1 + \boldsymbol{C}_1^{\mathsf{c}}\boldsymbol{\theta}_1^{\mathsf{c}},\ \boldsymbol{\beta}_2 = \boldsymbol{C}_2\boldsymbol{\theta}_2 + \boldsymbol{C}_2^{\mathsf{c}}\boldsymbol{\theta}_2^{\mathsf{c}},\ \boldsymbol{\gamma} = \boldsymbol{C}_I\boldsymbol{\theta}_I + \boldsymbol{C}_I^{\mathsf{c}}\boldsymbol{\theta}_I^{\mathsf{c}}$$

$$\boldsymbol{\theta}_1 \sim N(\boldsymbol{0}, 100^2\boldsymbol{I}),\ \boldsymbol{\theta}_2 \sim N(\boldsymbol{0}, 100^2\boldsymbol{I}),\ \boldsymbol{\theta}_I \sim N(\boldsymbol{0}, 100^2\boldsymbol{I})$$

$$\boldsymbol{\theta}_1^{\mathsf{c}} \sim N(\boldsymbol{0}, \sigma_1^2\boldsymbol{I}),\ \boldsymbol{\theta}_2^{\mathsf{c}} \sim N(\boldsymbol{0}, \sigma_2^2\boldsymbol{I}),\ \boldsymbol{\theta}_I^{\mathsf{c}} \sim N(\boldsymbol{0}, \sigma_I^2\boldsymbol{I})$$

where $\boldsymbol{\theta}_1$, $\boldsymbol{\theta}_2$, and $\boldsymbol{\theta}_I$ are the coefficients without shrinkage and $\boldsymbol{\theta}_1^{\mathsf{c}}$, $\boldsymbol{\theta}_2^{\mathsf{c}}$, and $\boldsymbol{\theta}_I^{\mathsf{c}}$ are the coefficients to be shrunk toward 0. In other words, $\boldsymbol{\beta}_1$, $\boldsymbol{\beta}_2$, and $\boldsymbol{\gamma}$, are shrunk toward $\boldsymbol{C}_1\boldsymbol{\theta}_1$, $\boldsymbol{C}_2\boldsymbol{\theta}_2$, and $\boldsymbol{C}_I\boldsymbol{\theta}_I$, respectively. To complete the model specification, we assign hyper-priors as

$$\sigma_1^2 \sim IG(a_0, b_0),\ \sigma_2^2 \sim IG(a_0, b_0),\ \sigma_I^2 \sim IG(a_0, b_0).$$

We fix $a_0 = b_0 = 0.001$ to assume an noninformative hyper-prior [Gelman et al., 2006]. Metropolis Hastings algorithm Hastings [1970] within a Gibbs sampler [Geman and Geman, 1984] can alternatively be used to approximate the posterior distribution of the model parameters. The full conditional distributions are provided in Appendix 4.6.4. The hyper-priors of BCDLM can alternatively be viewed as penalty terms in penalized likelihood. The dual representation is provided in Appendix 4.6.5.

Remark: The hyper-priors of BCDLM can be viewed as the penalty in penalized likeli-

hood. The dual representation of BCDLM in frequentist framework is to minimize

$$-\sum_{t=1}^{\top}\{y_t[\boldsymbol{X}_{1t}^{\top}(\boldsymbol{C}_1\boldsymbol{\theta}_1+\boldsymbol{C}_1^{\mathsf{c}}\boldsymbol{\theta}_1^{\mathsf{c}})+\boldsymbol{X}_{2t}^{\top}(\boldsymbol{C}_2\boldsymbol{\theta}_2+\boldsymbol{C}_2^{\mathsf{c}}\boldsymbol{\theta}_2^{\mathsf{c}})+\boldsymbol{X}_{It}^{\top}(\boldsymbol{C}_I\boldsymbol{\theta}_I+\boldsymbol{C}_I^{\mathsf{c}}\boldsymbol{\theta}_I^{\mathsf{c}})]$$
$$-e^{\boldsymbol{X}_{1t}^{\top}(\boldsymbol{C}_1\boldsymbol{\theta}_1+\boldsymbol{C}_1^{\mathsf{c}}\boldsymbol{\theta}_1^{\mathsf{c}})+\boldsymbol{X}_{2t}^{\top}(\boldsymbol{C}_2\boldsymbol{\theta}_2+\boldsymbol{C}_2^{\mathsf{c}}\boldsymbol{\theta}_2^{\mathsf{c}})+\boldsymbol{X}_{It}^{\top}(\boldsymbol{C}_I\boldsymbol{\theta}_I+\boldsymbol{C}_I^{\mathsf{c}}\boldsymbol{\theta}_I^{\mathsf{c}})}-\log(y_t!)\}$$
$$+\lambda_1\boldsymbol{\theta}_1^{\mathsf{c}\top}\boldsymbol{\theta}_1^{\mathsf{c}}+\lambda_2\boldsymbol{\theta}_2^{\mathsf{c}\top}\boldsymbol{\theta}_2^{\mathsf{c}}+\lambda_I\boldsymbol{\theta}_I^{\mathsf{c}\top}\boldsymbol{\theta}_I^{\mathsf{c}}$$

where λ_1, λ_2, and λ_I are tuning parameters that control the amount of shrinkage. When λ_1, λ_2, $\lambda_I \to 0$, the DL coefficients estimates approach $\hat{\boldsymbol{\Psi}}_{UDLM}$ as full-rank transformation of regression coefficients does not change the model fit of unpenalized likelihood. When λ_1, λ_2, $\lambda_I \to \infty$, the DL coefficients estimates approach $\hat{\boldsymbol{\Psi}}_{BiDLM}$ as $\boldsymbol{\theta}_1^{\mathsf{c}}$, $\boldsymbol{\theta}_2^{\mathsf{c}}$, and $\boldsymbol{\theta}_I^{\mathsf{c}}$ are all shrunk to zero.

The asymptotic MSE of $\hat{\boldsymbol{\Psi}}_{BCDLM}$ can be decomposed into the sum of squared bias

$$\boldsymbol{\Psi}^{\top}(\boldsymbol{W}^{\top}\hat{\boldsymbol{\Omega}}\boldsymbol{W}+\boldsymbol{S})^{-2}\boldsymbol{\Psi}$$

and variance

$$\sum_{j=1}^{p_1}\frac{1}{k_j}+\sum_{j=p_1+1}^{L_1+1}\frac{k_j}{(k_j+\lambda_1)^2}+\sum_{j=L_1+2}^{d_1}\frac{1}{k_j}+\sum_{j=d_1+1}^{d_2}\frac{k_j}{(k_j+\lambda_2)^2}+\sum_{j=d_2+1}^{d_3}\frac{1}{k_j}+\sum_{j=d_3+1}^{d_4}\frac{k_j}{(k_j+\lambda_I)^2}$$

where $\boldsymbol{W}=\left[\boldsymbol{X}_1\boldsymbol{C}_1|\boldsymbol{X}_1\boldsymbol{C}_1^{\mathsf{c}}|\boldsymbol{X}_2\boldsymbol{C}_2|\boldsymbol{X}_2\boldsymbol{C}_2^{\mathsf{c}}|\boldsymbol{X}_I\boldsymbol{C}_I|\boldsymbol{X}_I\boldsymbol{C}_I^{\mathsf{c}}\right]$, $\hat{\boldsymbol{\Omega}}$ is a diagonal matrix with the mean value of $\boldsymbol{Y}$ along the diagonal,
$\boldsymbol{S}=\text{diag}[\boldsymbol{0}_{p_1},\lambda_1\boldsymbol{1}_{L_1+1},\boldsymbol{0}_{p_2},\lambda_2\boldsymbol{1}_{L_2+1},\boldsymbol{0}_{p_1p_2},\lambda_I\boldsymbol{I}_{(L_1+1)(L_2+1)}]$, k_j is the j^{th} eigenvalue of $\boldsymbol{W}^{\top}\hat{\boldsymbol{\Omega}}\boldsymbol{W}$, and $d_1=L_1+p_2+1$, $d_2=L_1+L_2+2$, $d_3=L_1+L_2+p_1p_2+2$, and $d_4=(L_1+1)+(L_2+1)+(L_1+1)(L_2+1)$, respectively. Through some algebra, it has been shown that the squared bias is a monotonically increasing function and the variance is a monotonically decreasing function of the tuning parameters. Along with the convexity of the asymptotic MSE, there always exist λ_1, λ_2, $\lambda_I > 0$ that achieve greater bias-variance tradeoff.

4.3 Simulation Study

We conducted a simulation study to compare the estimation performance of the six methods introduced in Section 4.2 under different settings. We evaluate the estimation precision in terms of bias and MSE of the marginal lagged effects $\boldsymbol{\beta}_1^m(x_2^*)$ introduced in (4.3) based on 1000 repetitions for each simulation setting. We implemented the three frequentist methods UDLM, BiDLM, and TDLM using the built-in R function `glm`. We implemented the three Bayesian methods BiHDDLM, BTDLM, and BCDLM by calling the software Just Another Gibbs Sampler (JAGS) using R package `rjags` [Lunn et al., 2009]. The average computation times for 1000 data sets under each method are provided in Table 4.3 and all simulations were performed in R version 3.3.1.

4.3.1 Simulation Settings

We generated two separate exposure time series ($i = 1, 2$) of length 1000 with mean 3 and first-order autocorrelation equal to 0.5 from the model $x_{it} = 0.5x_{it-1} + \epsilon_{it}$ where $\epsilon_{it} \sim$ i.i.d $N(0, 0.75)$ for $i = 1, 2$ and $t = 1, \cdots, 1000$. We set $L_1 = L_2 = 9$ and generated the outcome y_t from a Poisson distribution with mean $\exp(\beta_0 + \boldsymbol{X}_{1t}^\top\boldsymbol{\beta}_1 + \boldsymbol{X}_{2t}^\top\boldsymbol{\beta}_2 + \boldsymbol{X}_{It}^\top\boldsymbol{\gamma})$ for $t = 1, \cdots, 1000$ where $\boldsymbol{X}_{1t}$, $\boldsymbol{X}_{2t}$, and $\boldsymbol{X}_{It}$ are as defined in Section 4.2. Let $\beta_0 = 3$ and consider two DL functions for main-effect coefficients $\boldsymbol{\beta}_1$ and $\boldsymbol{\beta}_2$ - (a) cubic and (b) function with departure from cubic (see Appendix 4.6.5). We consider five different underlying true interaction structures for $\boldsymbol{\gamma}$:

- (1) No interaction: $\gamma_{ij} = 0$ for $i = 0, \cdots, L_1$ and $j = 0, \cdots, L_2$
- (2) Tukey's style interaction: interaction effects are scaled product of their corresponding main effects - $\boldsymbol{\gamma} \propto (\boldsymbol{\beta}_1 \otimes \boldsymbol{\beta}_2)$
- (3) Kronecker product interaction: basis functions for interaction effects are tensor product of main-effect basis functions - $\boldsymbol{\gamma} \propto (\boldsymbol{C}_1 \otimes \boldsymbol{C}_2)\boldsymbol{\theta}_I$
- (4) Sparse interaction: only the interaction terms between exposure 1 at lag 1-2 and exposure 2 at lag 1-2 are nonzero

$$\begin{cases} \gamma_{ij} = 0.7, & i = 1, 2 \text{ and } j = 1, 2 \\ \gamma_{ij} = 0, & \text{otherwise} \end{cases}$$

- (5) Unstructured interaction

The exact specifications of the two main-effect coefficients and the five interaction-effect coefficients, including the unstructured interaction, are available in Appendix 4.6.5. In total, nine simulation scenarios, including all combinations of the two main-effect coefficients (a-b) and five interaction-effect coefficients (1-5) except the combination of (b) and (3), are considered. Excluding the combination of (b) and (3) is due to that Kronecker product interaction cannot be constructed when the corresponding main effects are not fully parametric as their underlying basis functions are undefined.

4.3.2 Evaluation Metrics

The marginal lagged effects of the first exposure defined in (4.3) is a function of the second exposure. The effects depend on the level at which the second exposure is fixed. One way to eliminate the effect of the second exposure is to integrate it out. We consider to use finite Riemann sum to numerically approximate the integral given by

$$\boldsymbol{\beta}_1^* = \int \boldsymbol{\beta}_1^*(x_2) dx_2 \approx \frac{1}{S} \sum_{s=1}^{S} \boldsymbol{\beta}_1^*(x_2^{(q_{(s-0.5)/S})})$$

where $x_2^{(q_{(s-0.5)/S})}$ is the $(s-0.5)/S$-th quantile of x_2. The empirical bias and empirical relative efficiency of the above quantity with $S = 20$ are used to summarize the simulation results across different scenarios. The squared bias is computed as

$$(\hat{\bar{\boldsymbol{\beta}}}_1^* - \boldsymbol{\beta}_1^*)^\top (\hat{\bar{\boldsymbol{\beta}}}_1^* - \boldsymbol{\beta}_1^*)$$

where $\hat{\bar{\boldsymbol{\beta}}}_1^*$ is the average of the estimates from 1000 repetitions. The empirical MSE is computed as

$$\frac{1}{1000} \sum_{j=1}^{1000} ||\hat{\boldsymbol{\beta}}_{1j}^* - \boldsymbol{\beta}_1^*||_2^2.$$

The relative efficiency is expressed with respect to the MSE of UDLM estimate, namely the MSE of UDLM divided by the MSE of a certain method. We emphasize that the efficiency is defined defined through MSE rather than variance in this chapter. Because of the symmetry between x_1 and x_2, we only present the results for the marginal lagged effects of x_1.

4.3.3 Simulation Results

The results for the situation with main effects generated from a cubic DL function are summarized in Table 4.1. As we can observe in scenario (1), all methods are more efficient than UDLM with relative efficiency ranging from 6.27 to 19.24 since the non-UDLM methods model the main effects and interaction effects in a parsimonious fashion, parametrically or nonparametrically. The empirical squared bias is minimal for UDLM (0.02), BiDLM (0.00) and BCDLM (0.00) and is moderately small for TDLM (0.19), BiHDDLM (0.45), and BTDLM (0.13). No interaction is a special case of Tukey's model with $\eta = 0$. Because TDLM correctly specify the main effects and interaction effects with a smaller number of parameters, it achieves the highest efficiency (19.24). In scenario (2) where the nonnull interaction effects are of Tukey's form, all methods have similar, though slightly smaller, relative efficiency in comparison with scenario (1) except BiHDDLM (0.78), ranging from 5.76 to 18.66. Again, TDLM has the highest relative efficiency as expected. The loss in efficiency for BiHDDLM is largely due to the biased estimates for interaction effects. BiHDDLM only assumes that the interaction surface is smooth without any particular structure. Tukey's style interaction does not guarantee the smoothness of the interaction surface and BiHDDLM fails to capture the structure. Scenario (3) represents the situation where the true interaction structure departs from Tukey's form. As we can see that now TDLM (3.45) is less efficient than BiDLM (6.68) because of the bias induced in estimating the interaction surface. However, TDLM is still more efficient than UDLM (1.00), BiHDDLM (1.10), and BTDLM (2.77) because the gain from using a single parameter for modeling the interaction effect only partly offsetted by the imposed bias (1.05). BiDLM correctly specifies both main effects and interaction effects in this scenario and it attains the highest

efficiency.

Both scenario (1) and scenario (2) are special cases of scenario (3). Even though BiDLM is less efficient than TDLM in the first two scenarios as expected, it still maintains a decent level of efficiency. Across scenarios (1)-(3), we note that the squared bias and relative efficiency of BTDLM always fall between BiDLM and TDLM, suggesting that BTDLM successfully performs shrinkage between the two models and achieves a better average performance. In addition, we can observe that the BCDLM (relative efficiency = 6.27, 5.76, 6.17) is slightly less efficient than BiDLM (relative efficiency = 6.82, 6.14, 6.68) across the three scenarios. The gap is due to the flexibility of BCDLM that accounts for possible departure from Kronecker product type of interaction structure. Scenarios (4) and (5) are the situations where UDLM is the only method that can unbiasedly estimate the interaction surface. As expected, both BiDLM and TDLM suffer from serious bias and the efficiency gains from dimension reduction diminished substantially. The relative efficiency ranges from 0.05 to 0.07. The class of interaction surfaces that BiDLM and TDLM can described is restricted and is distant from sparse and unstructured interaction structures. Note that all the methods jointly estimate the main effects and interaction effects and thus misspecifying the interaction effects could possibly distort the estimation of main effects as well as they are not orthogonal. BiHDDLM has squared bias 13.32 and 20.25 for scenarios (4) and (5) and is less biased than BiDLM and TDLM. On the other hand, it is much less efficient than UDLM. Both BTDLM (1.71, 1.88) and BCDLM (2.80, 2.70) are more efficient than UDLM by allowing departure from a specified interaction structure. BCDLM is less biased and more efficient than BTDLM across the two scenarios. The possible explanation is that BTDLM does shrinkage between Tukey's style interaction and Kronecker product interaction whereas BCDLM allows the most general case to be completely unstructured. Across all the scenarios when the main-effects are correctly specified, BCDLM has the best average performance in terms of estimation efficiency.

We summarize the results where main effects deviate from a cubic DL function in Table 4.2. As we can see that both BiDLM and TDLM are seriously biased, largely due to misspecification of the main-effect terms. These two methods are the least efficient. If we contrast scenarios (1) and scenario (2), we can see that the squared bias hugely increase for

the two, indicating that misspecification of main effect not only influences the estimation accuracy of main-effect DL function, but also the interaction DL surface. The reason is that TDLM explicitly connects the interaction coefficients with main-effect coefficients through the single scalar η and BiDLM implicitly connects the two sets of coefficients through specifying the transformation matrix of interaction as the Kronecker product of the transformation matrices of two exposures (i.e. $\boldsymbol{C}_1 \otimes \boldsymbol{C}_2$). BiHDDLM and BTDLM are biased across the board as well, with squared bias ranging from 8.17 and 132.07 and from 7.39 to 35.50, respectively. They are more efficient than UDLM only in the scenario where there is no interaction. BCDLM is slightly biased across different scenarios with squared bias ranging from 0.09 to 0.52. The BCDLM leads to gains in efficiency with reduced bias, especially in the scenario where there is no interaction. The relative efficiencies are 3.25, 1.35, 1.78, and 1.34 across the four scenarios. Summarizing the results in Tables 4.1 and 4.2, it is clear that the BCDLM approach has desirable MSE properties across the scenarios, offering a robust and efficient solution to this problem. The two tables summarize the simulation results to assess the estimation precision of the quantity provided in (4.3.2). We also provide the results based on the marginal lagged effects of one exposure when the other exposure is fixed at median in 4.3. The results and findings are similar.

4.4 Application

4.4.1 Data Overview and Modeling

We apply the six methods compared in the simulation section to the NMMAPS data. We jointly model daily time series of (1) PM_{10} and (2) O_3 in association to all-cause non-accidental mortality counts in Chicago, Illinois for the period between 1987 and 2000. The data contain daily mortality, air pollution, and weather data collected across 109 metropolitan areas in the United States from 1987 to 2000. Further details with respect to data assembly are available at `http://www.ihapss.jhsph.edu/data/NMMAPS/`. Previous single city studies found that the largest effects are present in the first seven lags [Schwartz, 2000, Goodman et al., 2004]. In addition, Zanobetti et al. [2000] indicated that

it is unlikely that lags beyond two weeks would have substantial influence on associations between short-term exposures to pollution and mortality. We therefore set $L_1 = L_2 = 14$ for PM_{10} and O_3, respectively.

Previous studies showed that it is crucial to account for meteorologic variables as potential confounders such as weather and seasonality in the analysis of air pollution effects [Peng et al., 2006, Welty and Zeger, 2005]. We specify the adjustment covariates in the same way as Dominici et al. [2005]. Let x_{1tk}, x_{2tk}, y_{tk}, and $\boldsymbol{z}_{tk}$ denote PM_{10} level, O_3 level, mortality count, and vector of time-varying covariates, measured on day t for age group k in Chicago for $t = 1, ..., 5114$ and $k = 1, 2, 3$, respectively. The three age categories are greater or equal to 75 years old, between 65 and 74 years old, and less than 65 years old. PM_{10} and O_3 were shared exposures across the three age groups so we have $x_{tk} \equiv x_t$. We assume that the mortality count in Chicago on day t for each of the age group k is a Poisson random variable Y_{tk} with mean μ_{tk} such that

$$\begin{aligned}
\log(\mu_{tk}) &= \boldsymbol{X}_{1t}^\top \boldsymbol{\beta}_1 + \boldsymbol{X}_{2t}^\top \boldsymbol{\beta}_2 + \boldsymbol{X}_{It}^\top \boldsymbol{\gamma} + \boldsymbol{z}_{tk}^\top \boldsymbol{\alpha} \\
&= \boldsymbol{X}_{1t}^\top \boldsymbol{\beta}_1 + \boldsymbol{X}_{2t}^\top \boldsymbol{\beta}_2 + \boldsymbol{X}_{It}^\top \boldsymbol{\gamma} + \alpha_0 + \sum_{j=1}^{2} \alpha_{1j} \mathrm{I}(k = j) \\
&\quad + \sum_{j=1}^{6} \alpha_{2j} \mathrm{I}(\mathrm{dow}_t = j) + \mathrm{ns}(\mathrm{temp}_t; 6 \text{ df}, \boldsymbol{\alpha}_3) \\
&\quad + \mathrm{ns}(\overline{\mathrm{temp}}_t^{(3)}; 6 \text{ df}, \boldsymbol{\alpha}_4) + \mathrm{ns}(\mathrm{dptp}_t; 3 \text{ df}, \boldsymbol{\alpha}_5) + \mathrm{ns}(\overline{\mathrm{dptp}}_t^{(3)}; 3 \text{ df}, \boldsymbol{\alpha}_6) \\
&\quad + \mathrm{ns}(t; 98 \text{ df}, \boldsymbol{\alpha}_7) + \mathrm{ns}(t; 14 \text{ df}, \boldsymbol{\alpha}_8) \mathrm{I}(k = 1) + \mathrm{ns}(t; 14 \text{ df}, \boldsymbol{\alpha}_9) \mathrm{I}(k = 2)
\end{aligned}$$

where $\boldsymbol{X}_{1t} = (x_{1t}, ..., x_{1,t-14})^\top$, $\boldsymbol{X}_{2t} = (x_{2t}, ..., x_{2,t-14})^\top$, $\boldsymbol{X}_{It} = \boldsymbol{X}_{1t} \otimes \boldsymbol{X}_{2t}$, $\mathrm{I}(\cdot)$ is the indicator function, and $\mathrm{ns}(\cdot)$ denotes the natural spline with specified df. Predictors dow_t, temp_t, $\overline{\mathrm{temp}}_t$, dptp_t, and $\overline{\mathrm{dptp}}_t$ represent the day of week, current day's temperature, adjusted average lag 1-3 temperature, current day's dewpoint temperature, and adjusted average lag 1-3 dewpoint temperatures for day t. The indicator variables allow different baseline mortality rates within each age group and within each day of week. The smooth term for time (t) is to adjust for long-term trends and seasonality and 98 df corresponds to 7 df per year over the 14-year horizon. The last two product terms separate smooth functions

of time with 2 df per year for each age group contrast. The primary goal is to estimate the coefficients $\boldsymbol{\beta}_1$, $\boldsymbol{\beta}_2$, and $\boldsymbol{\gamma}$ and $\boldsymbol{\alpha}$ is the set of covariate parameters. A 4-degree polynomial DL function is applied to both $\boldsymbol{\beta}_1$ and $\boldsymbol{\beta}_2$ for CDLM, TDLM, BTDLM, and BCDLM. The computation times are provided in Table 4.4.

The mean concentrations (standard deviations in parentheses) of PM_{10} and O_3 are 37.06 (19.25) $\mu g/m^3$ and 19.14 (10.20) ppb, respectively. The first quartile ($Q1$), second quartile ($Q2$), and third quartile ($Q3$) of PM_{10} are 24.29 $\mu g/m^3$, 34.25 $\mu g/m^3$, and 45.78 $\mu g/m^3$ and the three quartiles of O_3 are 13.51 ppb, 20.53 ppb, and 27.92 ppb, respectively. We observed minimal skewness in PM_{10} and O_3 measurements and no extreme values need further investigation. PACF indicates that the autocorrelation PM_{10} and O_3 have autocorrelation 0.42 and 0.74, respectively, at lag 1. O_3 time series displays strong correlation and slower decay than PM_{10}. The correlation on both time series suggests that some smoothing on DL coefficients is needed. The average daily non-accidental mortality count is 38.47 with standard deviation 15.89.

4.4.2 Estimating Marginal Distributed Lag Function

The quantity $100\{\exp[10(\beta_{1i} + x_2^* \sum_{n=0}^{L_2} \gamma_{in})]\}$ represents the percentage change in daily mortality with 10 $\mu g/m^3$ increase in PM_{10} at lag i when O_3 is at x_2^* ppb. Similarly, the quantity $100\{\exp[10(\beta_{2j} + x_1^* \sum_{m=0}^{L_1} \gamma_{mj})]\}$ represents the percentage change in daily mortality with 10 ppb increase in O_3 at lag j when PM_{10} is at x_1^* $\mu g/m^3$. We present the marginal lagged effects of PM_{10} and O_3 when the other pollutant is fixed at $Q1$, $Q2$, and $Q3$, respectively in Figures 4.1 and 4.2. Each panel in Figure 4.1 presents the marginal DL functions of PM_{10} when O_3 is fixed at $Q1$, $Q2$, and $Q3$ for one of the six methods. Likewise, each panel in Figure 4.2 presents the marginal DL functions of O_3 when PM_{10} is fixed at $Q1$, $Q2$, and $Q3$. If we look across the panels in Figure 4.1, we can observe that the fits of the three shrinkage methods BiHDDLM, BTDLM, and BCDLM are closer to UDLM fit than the CDLM fit and TDLM fit, suggesting that CDLM and TDLM might over-smooth the DL function. When O_3 is at $Q2$ and $Q3$, the over-smoothing of CDLM and TDLM results in underestimation of the PM_{10} peak effect at lag 3. For instance, the estimated percentage

increases in mortality associated with a $10\mu g/m^3$ increase in PM_{10} at lag 3 when O_3 fixed at $Q3$ are 0.59%, 0.41%, 0.31%, 0.26%, 0.18%, and 0.06% for UDLM, BTDLM, BCDLM, BiHDDLM, CDLM, and TDLM. The lower bounds of 95% confidence/credible intervals for the former four methods are appreciably above zero and the lower bounds of CDLM and TDLM are closer to zero. In this situation, shrinkage methods are more desirable since CDLM and TDLM misspecify the DL function and potentially underestimate the lag effects. In contrast, when we look across the panles in Figure 4.2, all methods except UDLM yield similar DL functions, indicating that potential misspecification by using CDLM and TDLM is minimal. We observe that the DL function peaks at lag 0 with PM_{10} fixed at Q1 and Q2 and peaks at lag 2 with PM_{10} fixed at Q3. The earlier peak for O_3 compared to PM_{10} suggests a more acute effect with an earlier window of susceptibility. We also observe that the UDLM fits of PM_{10} fluctuate more drastically than the UDLM fits of O_3. This reflects the stronger collinearity of O_3 time series and smoothing the DL function is certainly needed and preferred in this case.

We can observe that some of the estimated lagged effects are negative at larger lags for PM_{10}. The phenomenon is noted as mortality displacement [Zanobetti et al., 2000] and had been discovered in previous studies. Mortality displacement, also referred as harvesting effect [Zanobetti et al., 2002], is the temporal shift of mortality rate. Usually higher mortality rate due to the deaths of frail individuals a couple of days after the high air pollution episode is followed by compensatory reduction in mortality rate due to the decrease in frail individuals. The finding of possible mortality displacement is consistent with previous studies.

4.4.3 Assessing Interaction Effects

Within each panel of Figures 4.1 and 4.2, we notice that the estimated DL functions of one pollutant vary with the level of the other pollutant, indicating PM_{10} might moderate O_3 effect and vice versa. For UDLM, CDLM, and TDLM, we conducted likelihood ratio test to inspect whether the interaction effects are significantly different from 0. The p-values are 1.65×10^{-11} (DF = 225), 5.33×10^{-9} (DF = 25), and $< 10^{-4}$ (DF = 1), respectively.

The precision of the p-vale of TDLM is only up to 10^{-4} due to finite bootstrap samples. For three shrinkage methods BiHDDLM, BTDLM, and BCDLM, we computed the difference in deviance information criterion (DIC) [Spiegelhalter et al., 2002] between the model without and the model with interaction as a measure of model complexity and model fit. The DIC differences are 41.62, 25.56, and 68.35, respectively. It is difficult to determine a clear threshold of DIC difference for model selection [Plummer, 2008]. However, models with smaller DIC are generally preferred when DIC differences are greater than 10. Coupled with the p-values obtained from the frequentist approaches, we conclude that the interaction between PM_{10} and O_3 is evident.

From Figure 4.1 and Figure 4.2, we can see that generally the $Q3$ curves are above $Q2$ curves and $Q1$ curves suggesting that PM_{10} and O_3 have synergistic effects on each other. In other words, PM_{10} presents higher effect on mortality when O_3 is at a higher level, and vice versa. Furthermore, we observe that the gaps between the curves of the three quartiles diminish beyond lag 6 across the board. The interaction between PM_{10} and O_3 occurs at early lags. We added a dotted blue curve in each panel for the estimated DL function from single-pollutant analysis (model with PM_{10} alone or O_3 alone). The interposed curves represent the "average" DL effects if we disregard the interaction effect between the two pollutants.

4.4.4 Estimating Total Effects

As marginal lagged effects, total effects of PM_{10} and O_3 vary with the level of the other pollutant. The quantity $100\{\exp[10\sum_{i=0}^{L_1}(\beta_{1i} + x_2^*\sum_{n=0}^{L_2}\gamma_{in})]\}$ is the total effect of PM_{10} when O_3 is fixed at x_2^*. Similarly, the quantity $100\{\exp[10\sum_{j=0}^{L_2}(\beta_{2j} + x_1^*\sum_{m=0}^{L_1}\gamma_{mj})]\}$ is the total effect of O_3 when PM_{10} is fixed at x_1^*. Generally, the shrinkage methods have narrow confidence intervals than frequentist methods, suggesting that bias-variance trade-off and higher efficiency are achieved. The total effect of PM_{10} range from -1.66% to -0.97%, from -0.93% to -0.31%, and from 0.01% to 0.40% across different methods when O_3 is fixed at $Q1$, $Q2$, and $Q3$, respectively. The total effect of O_3 range from -1.63% to -1.28%, from -0.63% to -0.08%, and from 0.46% to 1.39% across different methods when

PM_{10} is fixed at the three quartiles, respectively. As we can see that the total effects of PM_{10} (O_3) are larger when O_3 (PM_{10}) is fixed at a higher level. Once again, PM_{10} and O_3 present synergistic effects on each other.

4.5 Discussion

In analyzing NMMAPS data, we demonstrated the importance of accounting for interaction between PM_{10} and O_3 time series in modeling the joint pollution effect on mortality. Two major pieces of evidence support the existence of pollutant-pollutant interaction - (1) marginal DL function of one pollutant varies when the level of the other pollutant changes, and (2) small p-values from frequentist approaches and large DICs from Bayesian approaches suggesting evidence in favor of $PM_{10}\times$ O_3 interaction. This adds to the evidence in support of plausible synergism involving O_3 that has been established in previous studies [Mauderly and Samet, 2009]. Development of two-pollutant DLM is key to our analysis.

In this chapter, we presented six different strategies to model lagged effects of two pollutants in a joint model. We reviewed two existing frequentist methods UDLM and BiDLM and we adapted HDDLM to two-dimensional situation (i.e. BiHDDLM). We proposed frequentist TDLM using Tukey's interaction structure, its Bayesian version, and a Bayesian approach to perform shrinkage between UDLM and BiDLM. There are two major novelties. We adopted Tukey's one-degree-of-freedom interaction structure to parsimoniously model two-way interaction. The estimation is efficient and the interaction testing is powerful. We also introduced the Bayesian version of TDLM (i.e. BTDLM) and the Bayesian version of BiDLM (i.e. BCDLM). The Bayesian models allow departure from the prespecified structure of DL function/surface and are robust to misspecification. They are data-adaptive and able to achieve bias-variance trade-off.

There are some limitations for the six approaches. UDLM is unbiased but potentially less efficient, especially when the autocorrelation between serial pollution measurement is large, especially when the autocorrelation between serial pollution measurement is high. BiDLM imposes some structure to constrain the lag coefficients as a function of the lags

and reduces the number of free parameters in the model. It can potentially achieve greater estimation precision. Nonetheless, when the DL structure is misspecified, the model-dependent BiDLM estimator can be seriously biased. Tukey's type interaction is mostly used for hypothesis testing rather than estimation in previous research. There are a few drawbacks that hinder the use of Tukey's model in effect estimation. Expressing interaction effects as a scaled product of its corresponding main effects implies that the interaction effects can be nonzero only when the main effects are nonzero. This feature adheres to the statistical principle that higher-order terms are considered only when their lower-order terms are present in ths model, whereas the lack of identifiability with respect to the scaled parameter in Tukey's model when main effects are not present makes the inference invalid. In addition, Tukey's model is not invariant to location shifts. Different centering schemes lead to different estimates of scaled parameter η and no universal remedy exists. Computationally, the non-convexity of the parameter space does not guarantee global optimum can be achieved. Also, the approximation of the maximum likelihood function using second-order Taylor expansion in terms of the model parameters is not precise. Therefore, analytically obtaining the estimates and their standard errors is forfeited. It is necessary to make use of iterative procedure and resampling approach.

One limitation of BiHDDLM is that it does not fully integrate the prior knowledge of the DL structure and the estimation efficiency is partly offset by its flexibility. It only assumes the smoothness of DL functions and DL surface. In addition, BiHDDLM implicitly assumes that the two pollutants are symmetrically modelled. The symmetry specification implies that $\text{Cor}(\gamma_{ij}, \gamma_{i+d,j}) = \text{Cor}(\gamma_{ij}, \gamma_{i,j+d})$, which may not be true in practice. In addition, predictive process interpolator tends to oversmooth DL functions and DL surface, resulting in biased estimates. Nevertheless, BiHDDLM assumes that the lagged effects at larger lags approach to 0. It is difficult for other methods to incorporate this feature. The hierarchical Bayesian model BCDLM is robust to misspecification of DL structure. The data-adaptive shrinkage can be regarded as an automatic procedure to attain a balance between a more general model UDLM and a more constrained model CDLM. The full-rank transformation on UDLM imposes smoothness on the shrinkage path and the *a priori* knowledge about the DL structure can be incorporated. It is noted that BCDLM can be

extended for exploring higher-order interaction and multiple-pollutant scenarios.

The two-pollutant DLM approaches introduced in this chapter can be directly extended to multi-pollutant situations where up to two-way interaction is considered. If one would like to consider higher-order interaction and/or nonlinear interaction, extension from tree-based approach such as CART and BKMR can be promising. In some occasions, choosing important pollutants among multiple candidates that are associated with a health outcome is the primary goal. LASSO and its variations are useful in dealing with variable selection in multi-pollutant situations.

In real-world situations, it is usually difficult to validate the underlying assumptions of a model-based estimator. The notion of data-adaptive shrinkage is attractive when no single estimator is universally optimal. When facing uncertainty, robust models such as BCDLM that possesses better average performance are more desirable. BCDLM can potentially be extended to areas outside environmental epidemiology. We hope our work will lead to more attempts in developing two-dimensional and multi-dimensional DLM in the future.

4.6 Appendix

4.6.1 Predictive Process Interpolator for Two-dimensional High Degree Distributed Lag Model (BiHDDLM)

Without loss of generality, we describe the procedures to construct the predictive process interpolator for $\boldsymbol{\beta}_1$. The interpolators for $\boldsymbol{\beta}_2$ and $\boldsymbol{\gamma}$ can be obtained in a similar manner.

1. Obtain the joint variance-covariance matrix of $(\boldsymbol{\beta}_1^\top, \boldsymbol{\beta}_1^{*\top}, \boldsymbol{\beta}_1^{+^\top})^\top$ using the Gaussian process prior.

2. Obtain the conditional variance-covariance matrix of $\boldsymbol{\beta}_1, \boldsymbol{\beta}_1^* | \boldsymbol{\beta}_1^+ = \mathbf{0}$ using the properties of multivariate Gaussian distribution and conditional distribution. Namely,

$$\mathbb{V}\text{ar}(\boldsymbol{\beta}_1, \boldsymbol{\beta}_1^* | \boldsymbol{\beta}_1^+) = (\boldsymbol{R}_{11} - \boldsymbol{R}_{12}\boldsymbol{R}_{22}^{-1}\boldsymbol{R}_{12}^\top)$$

where $\boldsymbol{R}_{11} = \mathbb{V}\text{ar}((\boldsymbol{\beta}_1^\top, \boldsymbol{\beta}_1^{*\top})^\top)$, $\boldsymbol{R}_{22} = \mathbb{V}\text{ar}(\boldsymbol{\beta}_1^+)$, and $\boldsymbol{R}_{12} = \mathbb{C}\text{ov}((\boldsymbol{\beta}_1^\top, \boldsymbol{\beta}_1^{*\top})^\top, \boldsymbol{\beta}_1^\top)$.

3. Compute the predictive process basis matrix as

$$\boldsymbol{B}_1 = \mathbb{C}\text{ov}(\boldsymbol{\beta}_1, \boldsymbol{\beta}_1^* | \boldsymbol{\beta}_1^+ = \mathbf{0}) \mathbb{V}\text{ar}^{-1}(\boldsymbol{\beta}_1^* | \boldsymbol{\beta}_1^+ = \mathbf{0}).$$

Bayesian computation is performed based on $\boldsymbol{\beta}_1^*$ using MCMC and the coefficient estimates of $\boldsymbol{\beta}_1$ can be mapped through

$$\boldsymbol{\beta}_1 = \boldsymbol{B}_1 \boldsymbol{\beta}_1^*.$$

4.6.2 Iterative Algorithm for Tukey's Distributed Lag Model (TDLM)

Define Vec$(\cdot)$ as an operator that transforms a $q \times r$ matrix $\boldsymbol{A}$ to a column vector $(a_{11}, \cdots, a_{1r}, \cdots, a_{qr})^T$ of length qr and Vec$^{(-1)}(\cdot)$ is the reversed operator. Let $Q(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2, \eta)$ be the likelihood function of TDLM valued at $\boldsymbol{\theta}_1, \boldsymbol{\theta}_2, \eta$, namely

$$Q(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2, \eta) = \sum_{t=1}^{T} [y_t [\boldsymbol{W}_{1t}^\top \boldsymbol{\theta}_1 + \boldsymbol{W}_{2t}^\top \boldsymbol{\theta}_2 + \boldsymbol{W}_{It}^\top \boldsymbol{\theta}_I]^\top - e^{\boldsymbol{W}_{1t}^\top \boldsymbol{\theta}_1 + \boldsymbol{W}_{2t}^\top \boldsymbol{\theta}_2 + \boldsymbol{W}_{It}^\top \boldsymbol{\theta}_I} - \log(y_t!)]$$

The iterative algorithm for fitting a Tukey's distributed lag model proceeds as follows:
Step 1: Initialize $\hat{\boldsymbol{\theta}}_1^{(0)}$, $\hat{\boldsymbol{\theta}}_2^{(0)}$, and $\hat{\eta}^{(0)}$

Step 2: Given $\hat{\boldsymbol{\theta}}_2^{(m-1)}$ and $\hat{\eta}^{(m-1)}$, update $\hat{\boldsymbol{\theta}}_1^{(m)}$.

Let $W_{0t}^* = \boldsymbol{W}_{2t}^\top \hat{\boldsymbol{\theta}}_2^{(m-1)}$ and $\boldsymbol{W}_{1t}^* = \boldsymbol{W}_{1t} + \hat{\eta}^{(m-1)} \text{Vec}^{-1}(\boldsymbol{W}_{It}) \hat{\boldsymbol{\theta}}_2^{(m-1)}$.

$$\hat{\boldsymbol{\theta}}_1^{(m)} = \underset{\boldsymbol{\theta}_1}{\text{argmin}} \sum_{t=1}^{T} [y_t [W_{0t}^* + \boldsymbol{W}_{1t}^{*\top} \boldsymbol{\theta}_1] - e^{W_{0t} + \boldsymbol{W}_{1t}^{*\top} \boldsymbol{\theta}_1} - \log(y_t!)]$$

Step 3: Given $\hat{\boldsymbol{\theta}}_1^{(m)}$ and $\hat{\eta}^{(m-1)}$, update $\hat{\boldsymbol{\theta}}_2^{(m)}$.

Let $W_{0t}^* = \boldsymbol{W}_{1t}^\top \hat{\boldsymbol{\theta}}_1^{(m)}$ and $\boldsymbol{W}_{2t}^* = \boldsymbol{W}_{2t} + \hat{\eta}^{(m-1)} \text{Vec}^{-1}(\boldsymbol{W}_{It}^T) \hat{\boldsymbol{\theta}}_1^{(m)}$.

$$\hat{\boldsymbol{\theta}}_2^{(m)} = \underset{\boldsymbol{\theta}_2}{\text{argmin}} \sum_{t=1}^{T} [y_t [W_{0t}^* + \boldsymbol{W}_{2t}^{*\top} \boldsymbol{\theta}_1] - e^{W_{0t} + \boldsymbol{W}_{2t}^{*\top} \boldsymbol{\theta}_2} - \log(y_t!)]$$

Step 4: Given $\hat{\boldsymbol{\theta}}_1^{(m)}$ and $\hat{\boldsymbol{\theta}}_2^{(m)}$, update $\hat{\eta}^{(m)}$.

Let $W_{0t}^* = \boldsymbol{W}_{1t}^\top \hat{\boldsymbol{\theta}}_1^{(m)} + \boldsymbol{W}_{2t}^\top \hat{\boldsymbol{\theta}}_2^{(m)}$ and $W_{It}^* = \boldsymbol{W}_{It}^\top (\hat{\boldsymbol{\theta}}_1^{(m)} \otimes \hat{\boldsymbol{\theta}}_2^{(m)})$.

$$\hat{\eta}^{(m)} = \underset{\eta}{\operatorname{argmin}} \sum_{t=1}^{T} [y_t [W_{0t}^* + \eta W_{It}^*] - e^{W_{0t}^* + \eta W_{It}^*} - \log(y_t!)]$$

Step 5: Compute relative increase in likelihood

$$\Delta^{(m)} = \frac{Q(\hat{\boldsymbol{\theta}}_1^{(m)}, \hat{\boldsymbol{\theta}}_2^{(m)}, \hat{\eta}^{(m)})}{Q(\hat{\boldsymbol{\theta}}_1^{(m-1)}, \hat{\boldsymbol{\theta}}_2^{(m-1)}, \hat{\eta}^{(m-1)})} - 1$$

Stop the algorithm if $\Delta^{(m)}$ is smaller than the pre-specified value. Otherwise, repeat steps 2-4 until convergence.

4.6.3 Full Conditional Distribution of Bayesian Tukey's Distributed Lag Model (BTDLM)

With constraints $\boldsymbol{\beta}_1 = \boldsymbol{C}_1\boldsymbol{\theta}_1, \boldsymbol{\beta}_2 = \boldsymbol{C}_2\boldsymbol{\theta}_2$, and $\boldsymbol{\gamma} = \boldsymbol{\eta} \odot (\boldsymbol{\beta}_1 \otimes \boldsymbol{\beta}_2)$, we have $f(\boldsymbol{Y}|\boldsymbol{\theta}_1, \boldsymbol{\theta}_2, \boldsymbol{\eta}) =$ $\exp\{\boldsymbol{Y}^T[\boldsymbol{X}_1\boldsymbol{C}_1\boldsymbol{\theta}_1 + \boldsymbol{X}_2\boldsymbol{C}_2\boldsymbol{\theta}_2 + \boldsymbol{X}_I(\boldsymbol{\eta} \odot ((\boldsymbol{C}_1 \otimes \boldsymbol{C}_2)(\boldsymbol{\theta}_1 \otimes \boldsymbol{\theta}_2)))]$

$$- e^{\boldsymbol{X}_1\boldsymbol{C}_1\boldsymbol{\theta}_1 + \boldsymbol{X}_2\boldsymbol{C}_2\boldsymbol{\theta}_2 + \boldsymbol{X}_I(\boldsymbol{\eta} \odot ((\boldsymbol{C}_1 \otimes \boldsymbol{C}_2)(\boldsymbol{\theta}_1 \otimes \boldsymbol{\theta}_2)))}\}$$

The full conditional distributions of $\boldsymbol{\theta}_1$, $\boldsymbol{\theta}_2$, $\boldsymbol{\eta}$, σ^2 and ω are

$$f(\boldsymbol{\theta}_1|\boldsymbol{\theta}_2, \boldsymbol{\eta}, \sigma^2, \omega, \boldsymbol{Y}) \propto f(\boldsymbol{Y}|\boldsymbol{\theta}_1, \boldsymbol{\theta}_2, \boldsymbol{\eta}) \cdot \exp(-\frac{\boldsymbol{\theta}_1^T\boldsymbol{\theta}_1}{2 \cdot 100^2})$$

$$f(\boldsymbol{\theta}_2|\boldsymbol{\theta}_1, \boldsymbol{\eta}, \sigma^2, \omega, \boldsymbol{Y}) \propto f(\boldsymbol{Y}|\boldsymbol{\theta}_1, \boldsymbol{\theta}_2, \boldsymbol{\eta}) \cdot \exp(-\frac{\boldsymbol{\theta}_2^T\boldsymbol{\theta}_2}{2 \cdot 100^2})$$

$$f(\boldsymbol{\eta}|\boldsymbol{\theta}_1, \boldsymbol{\theta}_2, \sigma^2, \omega, \boldsymbol{Y}) \propto f(\boldsymbol{Y}|\boldsymbol{\theta}_1, \boldsymbol{\theta}_2, \boldsymbol{\eta}) \cdot \exp(-\frac{\boldsymbol{\eta}^T\boldsymbol{\Sigma}(\omega)\boldsymbol{\eta}}{2\sigma^2})$$

$$f(\sigma^2|\boldsymbol{\theta}_1, \boldsymbol{\theta}_2, \boldsymbol{\eta}, \omega, \boldsymbol{Y}) \sim IG(a + (L_1+1)(L_2+1)/2, b + \boldsymbol{\eta}^T\boldsymbol{\Sigma}^{-1}(\omega)\boldsymbol{\eta}/2)$$

$$p(\omega|\boldsymbol{\theta}_1, \boldsymbol{\theta}_2, \boldsymbol{\eta}, \sigma^2, \boldsymbol{Y}) = \frac{|\sigma^2\boldsymbol{\Sigma}(\omega)|^{-1/2}\exp(-\frac{1}{2\sigma^2}\boldsymbol{\eta}^T\boldsymbol{\Sigma}^{-1}(\omega)\boldsymbol{\eta})}{\sum_{\omega^*}|\sigma^2\boldsymbol{\Sigma}(\omega^*)|^{-1/2}\exp(-\frac{1}{2\sigma^2}\boldsymbol{\eta}^T\boldsymbol{\Sigma}^{-1}(\omega^*)\boldsymbol{\eta})}$$

4.6.4 Full Conditional Distribution of Bayesian Constrained Distributed Lag Model (BCDLM)

With constraints $\boldsymbol{\beta}_1 = \boldsymbol{C}_1\boldsymbol{\theta}_1 + \boldsymbol{C}_1^{\mathsf{c}}\boldsymbol{\theta}_1^{\mathsf{c}}$, $\boldsymbol{\beta}_2 = \boldsymbol{C}_2\boldsymbol{\theta}_2 + \boldsymbol{C}_2^{\mathsf{c}}\boldsymbol{\theta}_2^{\mathsf{c}}$, and $\boldsymbol{\gamma} = \boldsymbol{C}_I\boldsymbol{\theta}_I + \boldsymbol{C}_I^{\mathsf{c}}\boldsymbol{\theta}_I^{\mathsf{c}}$, we have
$f(\boldsymbol{Y}|\boldsymbol{\theta}_1, \boldsymbol{\theta}_2, \boldsymbol{\theta}_I, \boldsymbol{\theta}_1^{\mathsf{c}}, \boldsymbol{\theta}_2^{\mathsf{c}}, \boldsymbol{\theta}_I^{\mathsf{c}})$
$= \exp\{\boldsymbol{Y}^T[\boldsymbol{X}_1\boldsymbol{C}_1\boldsymbol{\theta}_1 + \boldsymbol{X}_2\boldsymbol{C}_2\boldsymbol{\theta}_2 + \boldsymbol{X}_I\boldsymbol{C}_I\boldsymbol{\theta}_I + \boldsymbol{X}_1\boldsymbol{C}_1^{\mathsf{c}}\boldsymbol{\theta}_1^{\mathsf{c}} + \boldsymbol{X}_2\boldsymbol{C}_2^{\mathsf{c}}\boldsymbol{\theta}_2^{\mathsf{c}} + \boldsymbol{X}_I\boldsymbol{C}_I^{\mathsf{c}}\boldsymbol{\theta}_I^{\mathsf{c}}]$
$-e^{\boldsymbol{X}_1\boldsymbol{C}_1\boldsymbol{\theta}_1+\boldsymbol{X}_2\boldsymbol{C}_2\boldsymbol{\theta}_2+\boldsymbol{X}_I\boldsymbol{C}_I\boldsymbol{\theta}_I+\boldsymbol{X}_1\boldsymbol{C}_1^{\mathsf{c}}\boldsymbol{\theta}_1^{\mathsf{c}}+\boldsymbol{X}_2\boldsymbol{C}_2^{\mathsf{c}}\boldsymbol{\theta}_2^{\mathsf{c}}+\boldsymbol{X}_I\boldsymbol{C}_I^{\mathsf{c}}\boldsymbol{\theta}_I^{\mathsf{c}}}$. Without loss of generality, we only present the conditional distribution of $\boldsymbol{\theta}_1$, $\boldsymbol{C}_1^{\mathsf{c}}$, and σ_1^2. The conditional distributions of other parameters can be constructed by symmetry.

$$f(\boldsymbol{\theta}_1|\boldsymbol{\theta}_2, \boldsymbol{\theta}_I, \boldsymbol{\theta}_1^{\mathsf{c}}, \boldsymbol{\theta}_2^{\mathsf{c}}, \boldsymbol{\theta}_I^{\mathsf{c}}, \sigma_1^2, \sigma_2^2, \sigma_I^2, \boldsymbol{Y}) \propto f(\boldsymbol{Y}|\boldsymbol{\theta}_1, \boldsymbol{\theta}_2, \boldsymbol{\theta}_I, \boldsymbol{\theta}_1^{\mathsf{c}}, \boldsymbol{\theta}_2^{\mathsf{c}}, \boldsymbol{\theta}_I^{\mathsf{c}}) \cdot \exp(-\frac{\boldsymbol{\theta}_1^\top\boldsymbol{\theta}_1}{2 \cdot 100^2})$$

$$f(\boldsymbol{\theta}_1^{\mathsf{c}}|\boldsymbol{\theta}_1, \boldsymbol{\theta}_2, \boldsymbol{\theta}_I, \boldsymbol{\theta}_2^{\mathsf{c}}, \boldsymbol{\theta}_I^{\mathsf{c}}, \sigma_1^2, \sigma_2^2, \sigma_I^2, \boldsymbol{Y}) \propto f(\boldsymbol{Y}|\boldsymbol{\theta}_1, \boldsymbol{\theta}_2, \boldsymbol{\theta}_I, \boldsymbol{\theta}_1^{\mathsf{c}}, \boldsymbol{\theta}_2^{\mathsf{c}}, \boldsymbol{\theta}_I^{\mathsf{c}}) \cdot \exp(-\frac{\boldsymbol{\theta}_1^{\mathsf{c}\top}\boldsymbol{\theta}_1^{\mathsf{c}}}{2\sigma_1^2})$$

$$f(\sigma_1^2|\boldsymbol{\theta}_1, \boldsymbol{\theta}_2, \boldsymbol{\theta}_I, \boldsymbol{\theta}_1^{\mathsf{c}}, \boldsymbol{\theta}_2^{\mathsf{c}}, \boldsymbol{\theta}_I^{\mathsf{c}}, \sigma_2^2, \sigma_I^2, \boldsymbol{Y}) \sim IG(a_0 + (L_1 - p_1 + 1)/2, b_0 + \boldsymbol{\theta}_1^{\mathsf{c}\top}\boldsymbol{\theta}_1^{\mathsf{c}}/2)$$

4.6.5 Simulation Settings

Main effect coefficients $\boldsymbol{\beta}_1$ and $\boldsymbol{\beta}_2$:

- (a) Cubic:

$$\beta_{ij} = 0.0014[j(j-7)(j-9) + 8] \text{ for } i = 1, 2 \text{ and } j = 0, \cdots, 9$$

- (b) Function with departure from cubic:

$$\boldsymbol{\beta}_1 = \boldsymbol{\beta}_2 = (0.25, 0.2, 0, -0.1, -0.15, 0, 0.2, 0.1, 0, -0.05)^\top$$

Interaction coefficients $\boldsymbol{\gamma}$:

- (1) No interaction, (2) Tukey's style interaction, (3) Kronecker product interaction,

and (4) Sparse interaction are as defined in the main text

- (5) Unstructured interaction:

$$v = (0.2, 0.6, 1, 0.5, 0, 0.1, 0.7, 0.9, 0.2, 0.4)^\top$$

$$\boldsymbol{\gamma} = v \otimes v$$

Table 4.1: Empirical squared bias and empirical relative efficiency (measured with respect to the mean squared error of UDLM estimate) of marginal lagged effects across six different 2-dimensional distributed lag models based on 1000 simulation iterations. The lagged effects of the both exposures are generated from the same cubic DL function.

Interaction Structure	Metric	UDLM	BiDLM	TDLM	BiHDDLM	BTDLM	BCDLM
(1) No Interaction	Squared Bias	0.02	0.00	0.19	0.45	0.13	0.00
	Relative Efficiency	1.00	6.82	19.24	12.39	8.09	6.27
(2) Tukey's Structure	Squared Bias	0.01	0.00	0.01	5.49	0.01	0.00
	Relative Efficiency	1.00	6.14	18.66	0.78	6.71	5.76
(3) Kronecker Product	Squared Bias	0.02	0.00	1.05	3.66	0.90	0.00
	Relative Efficiency	1.00	6.68	3.45	1.10	2.77	6.17
(4) Sparse	Squared Bias	0.00	66.22	67.14	13.32	1.43	0.08
	Relative Efficiency	1.00	0.07	0.07	0.34	1.71	2.80
(5) Unstructured	Squared Bias	0.00	93.08	93.98	20.25	1.08	0.09
	Relative Efficiency	1.00	0.05	0.05	0.23	1.88	2.70

Table 4.2: Empirical squared bias and empirical relative efficiency (measured with respect to the mean squared error of UDLM estimate) of marginal lagged effects across six different 2-dimensional distributed lag models based on 1000 simulation iterations. The lagged effects of the both exposures are generated from the same cubic-like DL function (moderate departure from a cubic function).

Interaction Structure	Metric	UDLM	BiDLM	TDLM	BiHDDLM	BTDLM	BCDLM
(1) No Interaction	Squared Bias	0.02	69.51	70.03	8.17	7.39	0.10
	Relative Efficiency	1.00	0.24	0.25	1.78	1.59	3.25
(2) Tukey's Structure	Squared Bias	0.01	990.83	1023.84	132.07	35.50	0.09
	Relative Efficiency	1.00	0.00	0.00	0.02	0.05	1.35
(4) Sparse	Squared Bias	0.01	210.32	215.94	33.15	10.80	0.52
	Relative Efficiency	1.00	0.02	0.02	0.14	0.35	1.78
(5) Unstructured	Squared Bias	0.01	989.93	1019.06	131.61	31.83	0.10
	Relative Efficiency	1.00	0.00	0.00	0.02	0.04	1.34

Table 4.3: Average computation times of applying six two-pollutant distributed lag models on an Intel i7-2600 CPU with a single 3.4GHz core in one simulation scenario with length of time series $T = 1000$, maximum number of lag of the first pollutant $L_1 = 9$, and maximum number of lag of the second pollutant $L_2 = 9$ based on 1000 repetitions.

Methods	Time
UDLM	0.07 seconds
BiDLM	0.01 seconds
TDLM	0.33 seconds
BiHDDLM[1]	2.6 mins
BTDLM[1]	8.9 mins
BCDLM[1]	8.8 mins

[1] Gibbs sampler is based on 20000 burn-ins and 10000 posterior draws with thinning interval equal to 10

Table 4.4: Computation times of applying six two-pollutant distributed lag models on an Intel i7-2600 CPU with a single 3.4GHz core to the National Morbidity and Mortality Air Pollution Study (NMMAPS) to estimate the lagged effects of air particulate matter with aerodynamic diameter less than 10 micrometers (PM_{10}) and ozone (O_3) concentration on mortality in Chicago, Illinois from 1987 to 2000.

Methods	Time
UDLM	9.2 seconds
BiDLM	3.2 seconds
TDLM[1]	190 mins
BiHDDLM[2]	38 mins
BTDLM[2]	79 mins
BCDLM[2]	82 mins

[1] Standard error estimate is based on 1000 bootstrap samples

[2] Gibbs sampler is based on 20000 burn-ins and 10000 posterior draws with thinning interval equal to 10

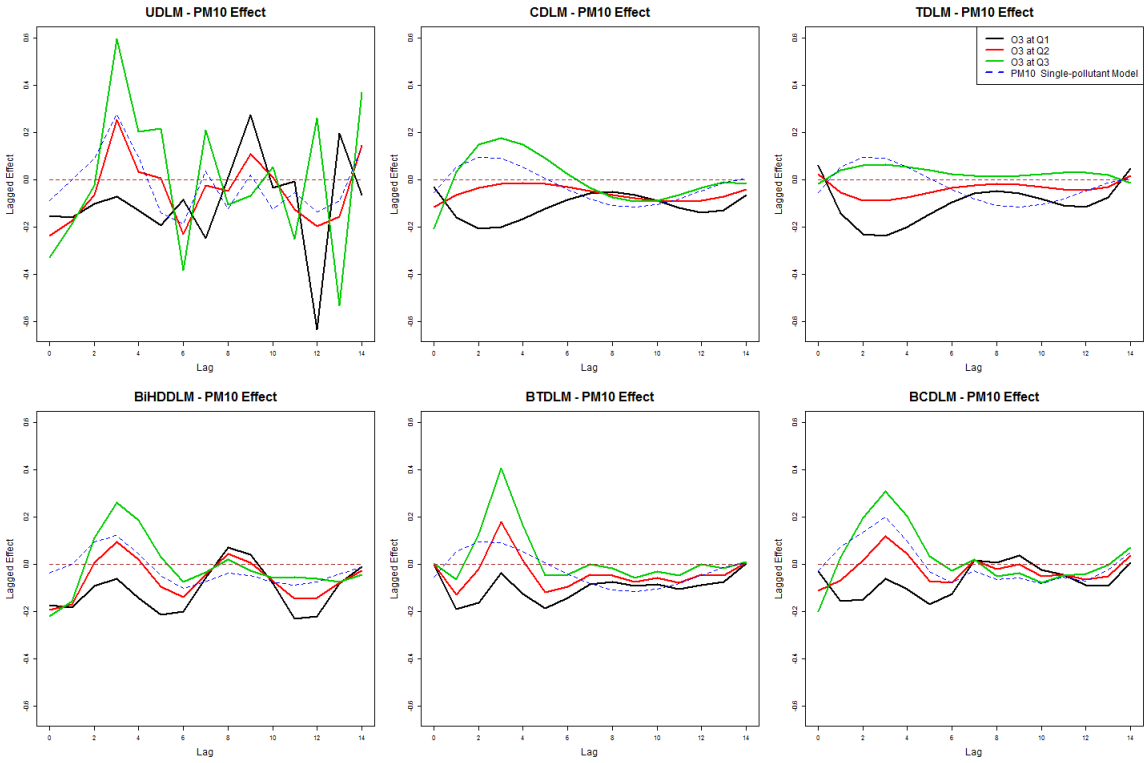

Figure 4.1: Estimated distributed lag functions up to 14 days for PM_{10} on mortality in Chicago, Illinois from 1987 to 2000 based on the data from the National Morbidity, Mortality, and Air Pollution Study (NMMAPS) under six estimation methods when O_3 is fixed at first quartile (black), second quartile (red), and third quartile (green) in a joint model and when O_3 is disregarded in a single-pollutant model for PM_{10} (blue). The lag effects are presented as the percentage change in mortality with an 10 $\mu g/m^3$ increase in PM_{10}. The six estimation methods are unconstrained distributed lag model (UDLM), bivariate distributed lag model (BiDLM), two-dimensional high degree distributed lag models (BiHDDLM), Tukey's distributed lag model (TDLM), Bayesian Tukey's distributed lag model (BTDLM), Bayesian constrained distributed lag model (BCDLM).

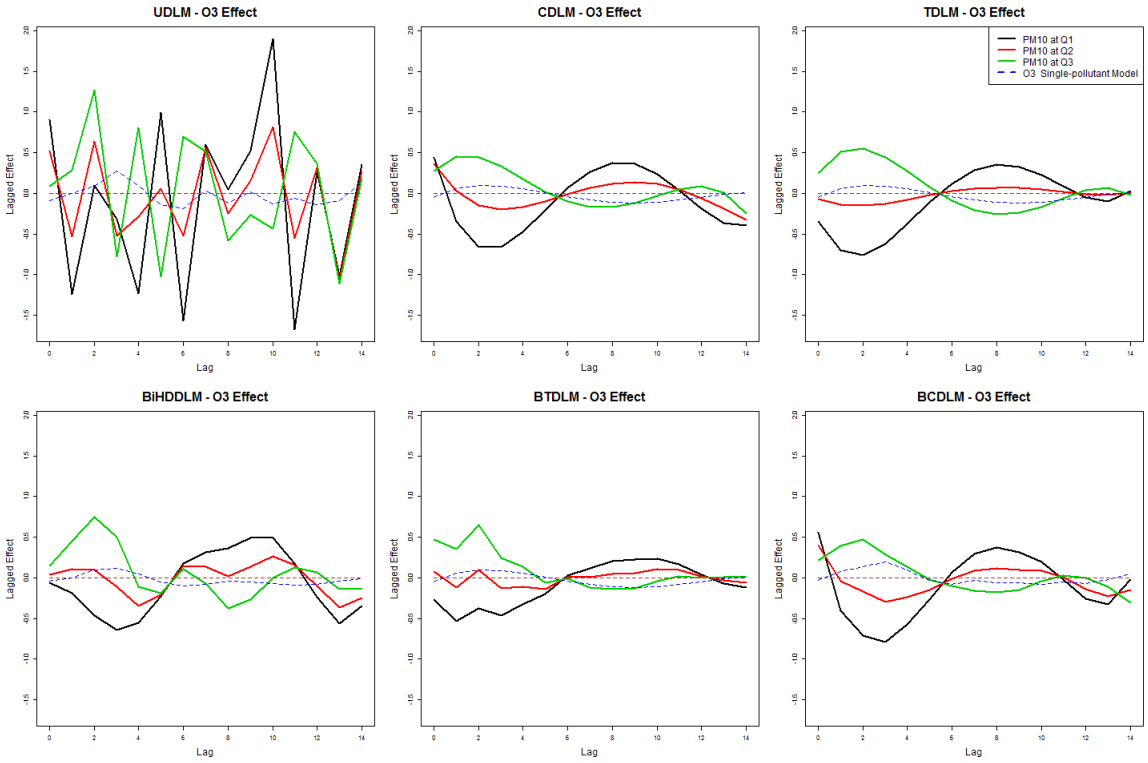

Figure 4.2: Estimated distributed lag functions up to 14 days for O_3 on mortality in Chicago, Illinois from 1987 to 2000 based on the data from the National Morbidity, Mortality, and Air Pollution Study (NMMAPS) under six estimation methods when PM_{10} is fixed at first quartile (black), second quartile (red), and third quartile (green) in a joint model and when PM_{10} is disregarded in a single-pollutant model for O_3 (blue). The lag effects are presented as the percentage change in mortality with an 10 ppb increase in O_3. The six estimation methods are unconstrained distributed lag model (UDLM), bivariate distributed lag model (BiDLM), two-dimensional high degree distributed lag models (BiHDDLM), Tukey's distributed lag model (TDLM), Bayesian Tukey's distributed lag model (BTDLM), Bayesian constrained distributed lag model (BCDLM).

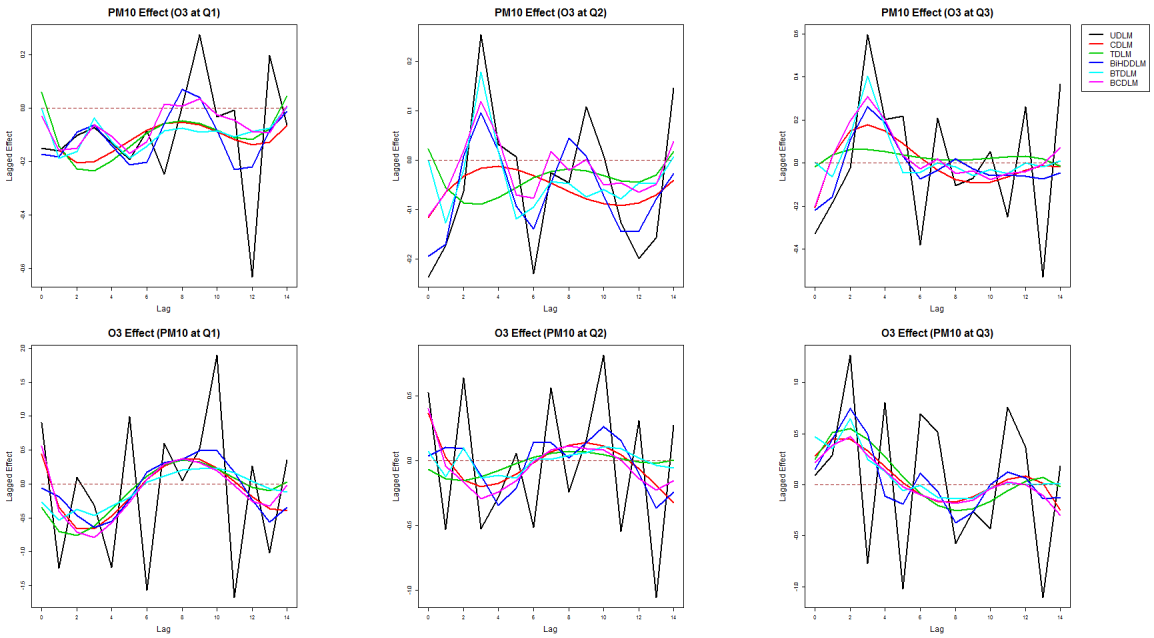

Figure 4.3: Estimated distributed lag functions up to 14 days for PM_{10} (upper) and O_3 (lower) on mortality in Chicago, Illinois from 1987 to 2000 based on the data from the National Morbidity, Mortality, and Air Pollution Study (NMMAPS) under six estimation methods when the other exposure is fixed at first quartile (left), second quartile (middle), and third quartile (right) in a joint model. The lag effects are presented as the percentage change in mortality with an 10 $\mu g/m^3$ increase in PM_{10} or 10 ppb increase in O_3. The six estimation methods are unconstrained distributed lag model (UDLM), bivariate distributed lag model (BiDLM), two-dimensional high degree distributed lag models (BiHDDLM), Tukey's distributed lag model (TDLM), Bayesian Tukey's distributed lag model (BTDLM), Bayesian constrained distributed lag model (BCDLM).

CHAPTER 5

Hierarchical Integrative Group LASSO

5.1 Introduction

A natural extension of single-pollutant DLM and two-pollutant DLM from previous chapters is to consider multiple-pollutant DLM. A variety of approaches have been proposed to estimate the health effects of multiple pollutants [Billionnet et al., 2012, Sun et al., 2013]. The most straightforward approach is a multiple regression model with a main effect for each pollutant and a two-way cross-product linear interaction term for each pair of pollutants [Dominici et al., 2010]. Penalized regression methods such as LASSO [Tibshirani, 1996] and elastic-net [Zou and Hastie, 2005] can be employed to identify a small subset of individual pollutants and interaction terms that are most notably associated with the outcome. In several studies, PCA have been used as a dimension reduction tool prior to multi-pollutant modeling [Arif and Shah, 2007, Qian et al., 2004].

Tree-based approaches such as CART are useful to account for higher-order and nonlinear interactions [Hu et al., 2008]. The DSA algorithm [Sinisi and van der Laan, 2004] allows users to specify the constraints on polynomial function form of exposure and the order of interaction. Bobb et al. [2013] used reduced hierarchical models to estimate health effects of simultaneous exposure to multiple pollutants by modeling nonlinear associations of main effects of the pollutants and their interactions via natural splines. In a health effects analysis of mixtures, BKMR [Bobb et al., 2014] was developed to avoid spline basis specifications and estimate the exposure-response relationship and facilitate inference on the strength of the association between individual pollutants and health outcomes.

However, these dimension reduction or variable selection techniques typically consider cross-sectional data. Very few methods so far directly address the lagged effect of multiple pollutants and their potential interactions over time.

A direct approach to incorporate the temporal dynamics of lags from multiple pollutants is a multiple regression model with all lagged measurements from multiple pollutants and their pairwise product terms as predictors. The model is certainly not optimal due to large number of parameters. The effect estimates are often not available by using traditional methods when the number of predictors is greater than sample size. Even with a larger sample size, the effect estimates can be unstable due to collinearity and statistical power for detecting main effect and/or interaction effect will typically be very low. Dimension reduction techniques such as a parametric DLM can be employed to help with the curse of dimensionality, but the number of predictors relative to sample size can still be large. A natural approach to tackle the problem is to induce sparsity in estimation. Penalized regression methods such as LASSO [Tibshirani, 1996] is a popular solution for variable selection that identifies a small subset of predictors highly associated with the outcome through selection and shrinkage.

The general structure of the statistical problem that we are solving is how to perform variable selection and screening in the model accounting for pairwise interaction effects with S groups of predictors given as

$$\boldsymbol{y} = \sum_{j=1}^{S} \boldsymbol{X}_j \boldsymbol{\beta}_j + \sum_{k<l} \boldsymbol{X}_{kl} \boldsymbol{\gamma}_{kl} + \boldsymbol{\epsilon} \tag{5.1}$$

where $\boldsymbol{X}_j$ is the $n \times p_j$ design matrix for group j with corresponding coefficient vector $\boldsymbol{\beta}_j$ of length p_j for $j = 1, \cdots, S$ and $\boldsymbol{X}_{kl}$ be the $n \times (p_k p_l)$ design matrix for two-way interaction between group k and group l where $\boldsymbol{\gamma}_{kl}$ is the corresponding $(p_k p_l)-$vector of interaction coefficients for $1 \leq k < l \leq S$, and $\boldsymbol{\epsilon}$ is the error vector following a multivariate standard Gaussian distribution. The group configuration can be defined through (a) set of basis functions representing nonlinearity or distributed lag structure, (b) multiple serially measurements from the same variable, (c) dummy variables representing multiple levels of categorical variables, or (d) natural grouping based on domain knowledge. We will give

examples later in this chapter. In addition, a sparse solution under strong heredity imposed at a group level is considered. Namely, an interaction term can be nonzero only when its corresponding main effect terms are nonzero.

We explain how a multi-pollutant DLM with consideration of lagged effects fits into the structure presented in (5.1) by expanding an individual pollutant into a group of variables containing its current and past measurements. We consider the scenario with S serially measured pollutants in association with a serially measured continuous health outcome for time $t = 1, \cdots, T$. Assume that the maximum number of lags for pollutant s is L_s with $s = 1, \cdots, S$. Let x_{ts} denote the measurement of pollutant s at time t. Let $\boldsymbol{x}_{ts} = (x_{ts}, \cdots, x_{t-L_s,s})^\top$ be the $L_s + 1-$vector of lagged measurements at time t for pollutant s. We denote $\boldsymbol{X}_s$ as a $T \times (L_s + 1)$ matrix with $\boldsymbol{x}_{ts}^\top$ on row t for $t = 1, \cdots, T$. Let $\boldsymbol{X}_{ss'}$ be a $T \times (L_s + 1)(L_{s'} + 1)$ matrix with $(\boldsymbol{x}_{ts}^\top \otimes \boldsymbol{x}_{ts'}^\top)$ on row t, representing inter-pollutant interactions. Also, let $\boldsymbol{y} = (y_1, \cdot, y_T)^\top$ denote the vector of outcome. The saturated model with S pollutants each with L_s lags and pairwise interaction is written as

$$E(\boldsymbol{y}) = \sum_{s=1}^{S} \boldsymbol{X}_s \boldsymbol{\beta}_s + \sum_{s<s'} \boldsymbol{X}_{ss'} \boldsymbol{\gamma}_{ss'} \tag{5.2}$$

where $\boldsymbol{\beta}_s$ is a $(L_s + 1)$-vector of lagged coefficients for pollutant s and $\boldsymbol{\gamma}_{ss'}$ is a $(L_s + 1)(L_{s'} + 1)$-vector of coefficients for interaction between pollutant s and pollutant s'. The number of parameters to be estimated is $\sum_{s=1}^{S}(L_s + 1) + \sum_{s \neq s'}(L_s + 1)(L_{s'} + 1)$. In assessing the pollutants' short-term effects, the pollutants are measured daily and number of lags typically to be considered is between 7 and 14 days, corresponding to $L_s = 8$ or 15. The dimensions of $\boldsymbol{\beta}_s$s and $\boldsymbol{\gamma}_{ss'}$s can thus be large and some form of dimension reduction is needed.

Recall the transformation matrix constructed for DLM introduced in Section 2.2.3. Assume $\boldsymbol{C}_s$ is a $(L_s + 1) \times d_s$ transformation matrix applied to $\boldsymbol{\beta}_s$ for $s = 1, \cdots, S$. In addition, let $\boldsymbol{C}_{ss'} = \boldsymbol{C}_s \otimes \boldsymbol{C}_{s'}$ be a $(L_s + 1)(L_{s'} + 1) \times d_s d_{s'}$ transformation matrix corresponding to $\boldsymbol{\gamma}_{ss'}$ following the framework of BiDLM [Muggeo, 2007] in Chapter IV. We

have the following equations

$$\boldsymbol{\beta}_s = \boldsymbol{C}_s\boldsymbol{\theta}_s \text{ for } s = 1, \cdots, S$$

$$\boldsymbol{\gamma}_{ss'} = \boldsymbol{C}_{ss'}\boldsymbol{\theta}_{ss'} \text{ for } 1 \leq s < s' \leq S.$$

where $\boldsymbol{\theta}_s$s and $\boldsymbol{\theta}_{ss'}$s are vectors of parameters in a lower dimensional subspace. Replacing two sets of equations into (5.2), we have

$$\begin{aligned} E(\boldsymbol{y}) &= \sum_{s=1}^{S} \boldsymbol{X}_s\boldsymbol{C}_s\boldsymbol{\theta}_s + \sum_{s<s'} \boldsymbol{X}_{ss'}\boldsymbol{C}_{ss'}\boldsymbol{\theta}_{ss'} \\ &= \sum_{s=1}^{S} \boldsymbol{W}_s\boldsymbol{\theta}_s + \sum_{s<s'} \boldsymbol{W}_{ss'}\boldsymbol{\theta}_{ss'} \end{aligned} \tag{5.3}$$

where $\boldsymbol{W}_s = \boldsymbol{X}_s\boldsymbol{C}_s$ for $s = 1, \cdots, S$ and $\boldsymbol{W}_{ss'} = \boldsymbol{X}_{ss'}\boldsymbol{C}_{ss'}$ for $1 \leq s < s' \leq S$. $\boldsymbol{W}_s$s and $\boldsymbol{W}_{ss'}$s can be viewed as new design matrices for main effects and interaction effects, respectively. The size for group s is now d_s for $s = 1, \cdots, S$ and the number of parameters is reduced to $\sum_{s=1}^{S} d_s + \sum_{s \neq s'} d_s d_{s'}$. Now we can see that the mean model in (5.3) and (5.1) are of the same form and a variable selection technique that can be applied for selecting main and interaction effects in (5.1) applied to (5.3).

Motivated by the need for a multi-pollutant DLM incorporating sparsity, we propose a HiGLASSO approach to perform variable selection at a group level in (5.1) while maintaining the strong heredity constraint [Bien et al., 2013]. Strong heredity enforces a model to include interaction only when both its corresponding main effects are present in the model. Analogous to adaptive group LASSO [Wang and Leng, 2008], weights are attached to the penalty terms to guarantee selection consistency and estimation consistency under certain conditions. Selection consistency implies that null variables converge to zero in probability as n goes to infinity. Estimation consistency implies that the difference between estimated coefficients and the true coefficients converges to zero in probability as n goes to infinity. Weights, assumed as functions of model coefficients, are integratively estimated with model coefficients in a one-stage framework.

The rest of this chapter is organized as follows. In Section 5.2, we give an overview of

various existing variable selection approaches that are potential choices for interaction selection and screening. In Section 5.3, we present the newly proposed HiGLASSO method. In Section 5.4, we compare and contrast various potential methods via an extensive simulation study in terms of their selection performance across different methods. In Section 5.5, we illustrate the proposed method by using NMMAPS dataset. In Section 5.6, we apply the proposed method to a dataset from Brigham and Women's Hospital (BWH) prospective pregnancy/birth cohort study that collects biological samples and detailed clinical data to identify important complex mixtures of chemical compounds and their possible interactions that are associated with biomarkers of oxidative stress. We will conclude with a discussion in Section 5.7.

5.2 LASSO-type Variable Selection Approaches

Two major classes of methods for variable selection methods are forward (stepwise) selection methods and penalization-based methods. Forward selection approaches provide useful alternatives to penalization-based approaches due to their scalability and interpretability. However, we focus on penalization-based approaches in this chapter. For forward selection approaches without interaction effects, we refer to Boos et al. [2009], Wasserman and Roeder [2009], and Luo and Ghosal [2015].

5.2.1 Variable Selection without Group Structure

Consider the usual regression setting with n observations and p predictors. Let $\boldsymbol{X}$ be the $n \times p$ design matrix, $\boldsymbol{y}$ be a $n-$vector of responses variables, and $\boldsymbol{\beta} = (\beta_1, ..., \beta_p)^\top$ be the $p-$vector of coefficients. The first-order model containing only linear main-effects without any interactions is given as

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$$

where $\boldsymbol{\epsilon}$ is the error vector following a multivariate standard Gaussian distribution. Throughout this chapter, let $||\boldsymbol{\mu}||_q$ denote the L_q-norm of a length r vector $\boldsymbol{\mu} = (\mu_1, \cdots, \mu_r)^\top$, defined as $(\sum_{i=1}^{r} ||\mu_i||^q)^{1/q}$, for $0 < q < \infty$.

Ridge regression [Hoerl and Kennard, 1970] with L_2-norm penalty is the first proposed penalization-based approach. However, it is a remedial measure to alleviate multicollinearity among regression correlated set of predictor variables in a model and does not shrink the regression coefficients to exact zero. On the other hand, LASSO [Tibshirani, 1996] utilizes L_1-norm constraint to regularize the parameter vector. The LASSO estimate is given by

$$\hat{\boldsymbol{\beta}}^{\text{LASSO}} = \arg\min_{\boldsymbol{\beta}} \Big\{ ||\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}||_2^2 + \lambda ||\boldsymbol{\beta}||_1 \Big\}.$$

Because of the nature of the constraint, some of the coefficients are shrunk exactly to zero. Least angle regression [Efron et al., 2004], gradient descent algorithm, and shooting method [Fu, 1998] can be used to obtain the coefficient path with varying λ. In signal processing literature, LASSO is also known as basis pursuit. [Chen et al., 2001]. It has been shown that there exist certain scenarios where the LASSO is biased in estimation and inconsistent for variable selection. Some recent approaches are proposed to construct a de-bias LASSO estimator [Javanmard and Montanari, 2015, Tian et al., 2015].

Zou [2006] introduced adaptive LASSO where different coefficients are penalized with different weights. The adaptive LASSO estimates are given by

$$\hat{\boldsymbol{\beta}}^{\text{aLASSO}} = \arg\min_{\boldsymbol{\beta}} \Big\{ ||\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}||_2^2 + \lambda \sum_{j=1}^{p} \hat{w}_j |\beta_j| \Big\}$$

where $\hat{w}_j$ is the pre-specified weight assigned to $|\beta_j|$. It has been shown that when weights are chosen as $\hat{w}_j = |\hat{\beta}_j|^{-\gamma}$ with $\gamma > 0$ where $\hat{\boldsymbol{\beta}} = (\hat{\beta}_1, \cdots, \hat{\beta}_p)^\top$s is an $\sqrt{n}$-consistent estimate of $\boldsymbol{\beta}$, adaptive LASSO has theoretical properties - (1) consistency in variable selection, (2) consistency in estimation, and (3) asymptotic normality. Better subset regression using the nonnegative garrote [Breiman, 1995] constrains the number of nonzero coefficients to achieve subset selection and the estimates are given by L_0-norm penalty

$$\hat{\boldsymbol{\beta}}^{\text{NNG}} = \arg\min_{\boldsymbol{\beta}} \Big\{ ||\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}||_2^2 + \lambda ||\boldsymbol{\beta}||_0 \Big\}$$

where $\lambda||\boldsymbol{\beta}||_0 = \lambda \sum_{j=1}^{p} I(\beta_j \neq 0)$ and $I(\cdot)$ is an indicator function. Zou [2006] showed that nonnegative garrote is closely related to a special case of adaptive LASSO with $\gamma = 1$

with additional sign constraints and nonnegative garrote can be viewed as an integrative LASSO.

5.2.2 Variable Selection with Group Structure

The methods described in Section 5.2.1 treat each predictor variable separately and are designed to select variables individually. However, there are situations where it is desirable to choose variables in a grouped manner. Group LASSO [Yuan and Lin, 2006] was proposed to behave like LASSO at a group level. Suppose that p predictors are divided into S groups and p_j is the size of group j for $j = 1, ..., S$. Let $\boldsymbol{X}_j$ and $\boldsymbol{\beta}_j$ denote the $n \times p_j$ design matrix and coefficient vector of length p_j corresponding to group j, respectively. A model with group of predictors without any interactions is given as

$$\boldsymbol{y} = \sum_{j=1}^{S} \boldsymbol{X}_j \boldsymbol{\beta}_j + \boldsymbol{\epsilon}$$

The objective function of group LASSO is

$$\hat{\boldsymbol{\beta}}^{\text{gLASSO}} = \arg\min_{\boldsymbol{\beta}_j} \Big\{ ||\boldsymbol{y} - \sum_{j=1}^{S} \boldsymbol{X}_j \boldsymbol{\beta}_j||_2^2 + \lambda \sum_{j=1}^{S} \sqrt{p_j} ||\boldsymbol{\beta}_j||_2 \Big\}.$$

The L_2 norm penalty induced at a group level ensures that all the coefficients within the same group retain or drop out of a model simultaneously. In other words, variable selection is operated at a group level. Group LARS [Efron et al., 2004], a group version of nonnegative garrote [Breiman, 1995], and block gradient descent can be applied to obtain the solution path. Lin and Zhang introduced COmponent Selection and Smoothing Operator (COSSO) [Lin et al., 2006], a penalized least squares method with the penalty function being the sum of component norms rather than the sum of squared component norms, to achieve group-level sparsity. An algorithm that iterates between the smoothing spline and the nonnegative garrote is considered. Group LASSO and COSSO yield sparsity at a group level but not within a group. If sparsity at both a group level and an individual level is

desired, the sparse group LASSO [Friedman et al., 2010] with objective function

$$\hat{\boldsymbol{\beta}}^{\text{sgLASSO}} = \arg\min_{\boldsymbol{\beta}_j} \Big\{ ||\boldsymbol{y} - \sum_{j=1}^{S} \boldsymbol{X}_j \boldsymbol{\beta}_j||_2^2 + \lambda_1 \sum_{j=1}^{S} \sqrt{p_j} ||\boldsymbol{\beta}_j||_2 + \lambda_2 ||\boldsymbol{\beta}||_1 \Big\}.$$

may be considered.

The above variable selection approaches at a group level are not fully efficient and the resulting variables selected could be inconsistent, due to the same amount of shrinkage applied to each group of regression coefficients, under certain conditions. Adaptive group LASSO [Wang and Leng, 2008] was proposed in the spirit of adaptive LASSO to serve as a remedy. The objective function of adaptive group LASSO is given by

$$\hat{\boldsymbol{\beta}}^{\text{agLASSO}} = \arg\min_{\boldsymbol{\beta}_j} \Big\{ ||\boldsymbol{y} - \sum_{j=1}^{S} \boldsymbol{X}_j \boldsymbol{\beta}_j||_2^2 + \lambda \sum_{j=1}^{S} \hat{w}_j ||\boldsymbol{\beta}_j||_2 \Big\}.$$

A common choice of weights is $\hat{w}_j = ||\hat{\boldsymbol{\beta}}_j||_2^{-\gamma}$ with $\gamma > 0$ for $j = 1, \cdots, S$. The proposed adaptive group LASSO has been shown to achieve selection consistency and estimation consistency.

5.2.3 Variable Selection Models for Interaction Identification with Heredity Assumption

If interactions could be selected under no heredity constraint, the approaches introduced in the previous two sections can be conveniently extended by including the interaction terms as a product of two variables in the feature space. Consider the p-predictor setting in Section 5.2.1. Let $\boldsymbol{x}_j$ be the $n-$vector for predictor j with coefficient β_j, for $j = 1, \cdots, p$. Let γ_{kl} be the coefficient of interaction between $\boldsymbol{x}_k$ and $\boldsymbol{x}_l$. Conventionally, interaction models are studied under strong or weak heredity constraints [Hamada and Wu, 1992, Chipman, 1996, Nelder, 1977, Peixoto, 1987], defined as follows.

- Strong Heredity: If an interaction term is included in the model, both of its corresponding main effects must be present in the model as well. That is, if $\gamma_{kl} \neq 0$, then $\beta_k \neq 0$ and $\beta_l \neq 0$.

- Weak Heredity: If an interaction term is included in the model, either of its corresponding main effects must be present in the model as well. That is, if $\gamma_{kl} \neq 0$, then $\beta_k \neq 0$ or $\beta_l \neq 0$.

The heredity property is also known as "marginality" and "being hierarchically well-formulated." Some statisticians argue that models violating heredity constraints are not sensible and heredity constraints make model interpretation easier [McCullagh, 1984] and improve statistical power [Cox, 1984]. Constraints to enforce heredity have been incorporated into traditional step-wise selection approaches [Wu et al., 2010, Bickel et al., 2010, Crews et al., 2011]. Recently, Hao and Zhang [2014] proposed iFORM that allows linear interactions to appear in the model only when the main effects have already been selected. Narisetty et al. [2017] proposed Selection of Non-linear Interactions by a Forward stepwise method (SNIF) as an extension to account for potential non-linear interactions. Again, we focus on penalization-based approaches instead of selection approaches in the rest of this section.

A generic second-order model accounting for pairwise interaction effects with individual predictors is given as

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{X}_I\boldsymbol{\gamma} + \boldsymbol{\epsilon}$$

where $\boldsymbol{X}_I = [\boldsymbol{x}_1 \odot \boldsymbol{x}_2, \cdots, \boldsymbol{x}_{p-1} \odot \boldsymbol{x}_p]$ denotes the $n \times [p(p-1)/2]$ design matrix for interaction and $\boldsymbol{\gamma} = (\gamma_{12}, \cdots, \gamma_{p-1,p})^\top$. Yuan et al. [2009] modified the nonnegative garrote algorithm by adding linear inequality constraints to enforce heredity. Choi et al. [2010] proposed a non-convex strong heredity interaction model (SHIM) by reparametrizing the interaction coefficients. The interaction coefficients are expressed as scaled products of their corresponding main effect terms, namely $\gamma_{ij} = \eta_{ij}\beta_i\beta_j$ for $1 \leq i < j \leq p$ where η_{ij}s are scalar parameters to be penalized. The strong heredity is therefore automatically enforced. Penalization is imposed on the derived scalars for interaction, rather than the interaction coefficients at the original scale. An iterative algorithm between LASSO and group LASSO is applied to estimate the model coefficients. The hierNet approach [Bien et al., 2013], a LASSO for hierarchical interactions, imposes a set of convex constraints to LASSO to accommodate the hierarchical restrictions. Specifically, it minimizes the fol-

lowing object function

$$\frac{1}{2}||\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta} - \sum_{k=1}^{p}\sum_{l=1}^{p}(\boldsymbol{x}_k \cdot \boldsymbol{x}_l)\gamma_{kl}||_2^2 + \lambda_1 \sum_{j=1}^{p}|\beta_j| + \frac{1}{2}\lambda_2 \sum_{k=1}^{p}\sum_{l=1}^{p}|\gamma_{kl}|$$

subject to the constraints $\gamma_{kl} = \gamma_{lk}\ \forall\ 1 \leq k, l \leq p$ and $\sum_{l=1}^{p}|\gamma_{kl}| \leq |\beta_k|\ \forall k = 1, \cdots, p$. The first constraint assumes symmetry and the second constraint induces strong heredity. The convex relaxation of the non-convex constraints is employed by separating the positive and negative parts of each β_j. An Alternating Direction Method of Multipliers (ADMM) algorithm [Boyd et al., 2011] can be used to solve the constrained optimization problems. A FrAmework for Modeling Interactions with a convex penaLtY (FAMILY) [Haris et al., 2016] was recently proposed to generalize LASSO using main effects only, LASSO using main effects and all pairwise interactions, and hierNet. It can be formulated as a convex optimization problem and can also be solved using an efficient ADMM algorithm.

Learning Interactions via Hierarchical Group-Lasso (GLinternet) [Lim and Hastie, 2013] utilizes overlapped group LASSO penalty to enforce the strong heredity. The objective function is given by

$$\frac{1}{2}||\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta} - \sum_{k=1}^{p}\sum_{l=1}^{p}[\boldsymbol{x}_k, \boldsymbol{x}_l, \boldsymbol{x}_k \cdot \boldsymbol{x}_l]\boldsymbol{\gamma}_{kl}^*||_2^2 + \lambda_1||\boldsymbol{\beta}||_1 + \frac{1}{2}\lambda_2 \sum_{k=1}^{p}\sum_{l=1}^{p}||\boldsymbol{\gamma}_{kl}^*||_2$$

where each $\boldsymbol{\gamma}_{kl}^*$ is a three dimensional vector with the first two elements corresponding to main effects and the third element corresponding to interaction effect. Note that the main effects appear twice in the above function, creating an overlap in the penalty term. Strong heredity constraints are naturally respected because group LASSO penalties enforce all the variables within the same group to be selected or not selected. In other words, whenever an interaction is estimated to be nonzero, both its associated main effects are also included in the model since all three terms are bundled in the same group. A Fast Iterative Soft Thresholding Algorithm (FISTA) can be adapted to solve the GLinternet optimization problem [Beck and Teboulle, 2009]. A similar approach that employs composite absolute penalties [Zhao et al., 2009] was also proposed for selection under hierarchical constraints.

Following the notation in Section 5.2.2, a second-order model accounting for pairwise

interaction effects with groups of predictors is given as

$$\boldsymbol{y} = \sum_{j=1}^{S} \boldsymbol{X}_j \boldsymbol{\beta}_j + \sum_{k=1}^{p-1} \sum_{l=k+1}^{p} \boldsymbol{X}_{kl} \boldsymbol{\gamma}_{kl} + \boldsymbol{\epsilon}$$

where $\boldsymbol{X}_{kl}$ is the $n \times (p_k p_l)$ design matrix for two-way interaction between group k and group l and $\boldsymbol{\gamma}_{kl}$ be the corresponding $(p_k p_l)-$vector of interaction coefficients. Variable selection using Adaptive Nonlinear Interaction Structures in High dimensions (VANISH) [Radchenko and James, 2010] is by far the only penalization-based method capable of performing variable selection with both main effects and interaction effects at a group level while maintaining the strong heredity constraints. The group structure was originally designed to allow for nonlinear effects but it can potentially accommodate other situations where groups are defined otherwise. VANISH optimizes a penalized objective function as

$$\frac{1}{2}||\boldsymbol{y} - \sum_{j=1}^{S} \boldsymbol{X}_j \boldsymbol{\beta}_j - \sum_{k=1}^{p-1} \sum_{l=k+1}^{p} \boldsymbol{X}_{kl} \boldsymbol{\gamma}_{kl}||_2^2 + \lambda_1 \sum_{j=1}^{p} (||\boldsymbol{\beta}_j||_2^2 + \sum_{k>j} ||\boldsymbol{\gamma}_{kj}||_2^2 + \sum_{l>j} ||\boldsymbol{\gamma}_{jl}||_2^2)^{1/2}$$
$$+ \lambda_2 \sum_{k=1}^{p-1} \sum_{l=k+1}^{p} ||\boldsymbol{\gamma}_{kl}||_2.$$

By construction, $\boldsymbol{\beta}_j$s and $\boldsymbol{\gamma}_{kl}$s are bundled together in the first penalty term so main effect coefficients and interaction coefficients are all zero or all nonzero, based on the property of group LASSO penalty. An accelerated algorithm that incorporates block gradient descent and involves a single sweep through all variables can be applied to reduce computational burden.

5.3 Hierarchical Integrative Group LASSO (HiGLASSO)

We propose a HiGLASSO framework to solve the variable selection problem at a group level with two-way interaction using integrative LASSO, while maintaining the strong heredity constraints. The framework serves as the first penalization-based approach that incorporates integrative weights in group variable selection models of interaction with heredity. It has theoretical advantages of achieving selection consistency and estimation

consistency. A quadratic approximation for penalty function is considered to circumvent the non-convex problem and reduces computational burden.

5.3.1 Major Features of HiGLASSO

HiGLASSO has four major properties.

(1) Induces sparsity for variable selection (LASSO) :
As all the methods introduced in Section 5.2 , HiGLASSO induces sparsity by including penalty terms in least squared objective function so it is capable of performing variable selection akin to other LASSO-based approaches.

(2) Maintains group structure (Group):
HiGLASSO has the capability to select predictive variables in a grouped manner. The group structure can be defined based on the context of application. For example, it could be

(a) set of basis functions representing nonlinear relationships. For example, using cubic splines without intercept, a single continuous variable x can be expanded to a group of three variables [x, x^2, x^3]. Without loss of generality, we consider the group structure to be nonlinear expansion of continuous variables via some pre-specified basis functions in our presentation. Particularly, with p basis functions $B_1(\cdot), \cdots, B_p(\cdot)$, a continuous variable x can be expanded to a group of p variables $B_1(x), \cdots, B_p(x)$.

(b) multiple serially measurements from the same variable. For example, a serially measured pollutant x_t can be expanded to a group of $(L + 1)$ lagged variables $[x_t, \cdots, x_{t-L}]$ where x_t is the pollutant measured at time t.

(c) dummy variables representing different levels of categorical variables. For example, a 3-level categorical variable x can be expanded into a group of two dummy variables $[I(x = 1), I(x = 2)]$ where $x = 0$, 1, or 2.

(d) natural grouping based on domain knowledge. For example, exposure markers sharing the same metabolic pathway may be classified in the same group.

HiGLASSO framework is very general and can be applied to broader context where two-way interaction is considered among a set of predictors and selection is to be conducted at a group level. All the variables within the same group are selected or not selected. The sparsity is induced at a group level, not individual level.

(3) Imposes strong heredity on two-way interaction (Hierarchical):
When conducting statistical analysis, a standard practice is to consider the higher-order terms only when the corresponding lower-order terms are present in the model. It is well-known that not properly adjusting for nonlinear main effects might result in spurious detection of interaction effects [Bauer and Cai, 2009, Cornelis et al., 2012, Mukherjee et al., 2012, He et al., 2016], since the higher-order terms for main effects and interaction terms between two predictor are not easily differentiable when the signal-to-noise level is low. Therefore, it is desirable to impose strong heredity constraints in an interaction model allowing for nonlinear main effects. HiGLASSO reparametrizes each interaction coefficient as a scaled product of its corresponding main effect coefficients, in the spirit of SHIM [Choi et al., 2010]. Penalization is operated on the derived scalar parameters rather than the original interaction coefficients. The strong heredity is thus maintained through the underlying construction.

(4) Incorporates weights (Integrative):
LASSO penalizes each variable in the same magnitude. Likewise, group LASSO penalizes each group of variables in a similar manner. It has been shown that the same tuning parameter λ (amount of penalization) for each predictor/group without assessing their relative importance may degrade the estimation efficiency and affect the selection consistency [Leng et al., 2006, Zou, 2006, Yuan and Lin, 2007]. Adaptive LASSO [Zou, 2006] and adaptive group LASSO [Wang and Leng, 2008] serve as potential remedies. A separate penalty factor can be assigned to each predictor/group based on initial estimates. Typically, a two-stage approach is employed.

In the first stage, the reciprocal of the absolute values of the ordinary least squared (OLS) estimates are obtained as adaptive weights. In the second stage, the adaptive weights are substituted into the penalty term and the penalized least squared function can be minimized in terms of the model coefficients. However, this approach has two major limitations. When $p > n$, OLS does not work. One workaround is to recourse to the estimate obtained from "unadaptive" version of LASSO or group LASSO, following adaptive elastic-net [Zou and Zhang, 2009]. Furthermore, the estimation consistency only holds when the adaptive weights are inversely proportional $\sqrt{n}$-consistent estimates of $\boldsymbol{\beta}$ to a positive power. Since we consider groups of variables and their two-way interaction, the number of effective predictors can potentially be larger than the sample size. It is difficult to obtain a consistent estimate of main effects and interaction effects in a high-dimensional situation. We therefore consider estimating the weights and model parameters simultaneously in HiGLASSO.

None of the existing variable selection methods targeted towards interaction possesses all four features simultaneously. On one hand, GLinternet [Lim and Hastie, 2013] and VANISH [Radchenko and James, 2010] have the first three features but it is difficult to incorporate the fourth since these two methods have a penalty term that involves both main effect terms and interactions. Shrinking interaction terms and their corresponding main effect terms by the same penalty factors is not optimal. On the other hand, SHIM [Choi et al., 2010] only considers linear main effects and linear interaction effects and it possess features (1), (3), and (4). The problem with imposing group structure on SHIM is that obtaining adaptive weights following Breiman [1995] and Zou [2006] might not be feasible anymore because of high dimensionality. HiGLASSO is developed to overcome the barriers.

5.3.2 Developing the HiGLASSO Framework

Consider the regression setting where there are n subjects and S groups of predictors with p_j as the size of group j for $j = 1, \cdots, S$. Let x_{ijk} denote the k^{th} predictor in group j for subject i. Let $\boldsymbol{x}_{ij} = (x_{ij1}, \cdots, x_{ijp_j})^\top$ be the vector of group j for subject i and let $\boldsymbol{X}_j$ be

the $n \times p_j$ design matrix for group j with $\boldsymbol{x}_{ij}^\top$ on row i. Let $\boldsymbol{X}_{jj'}$ denote the $n \times (p_j p_{j'})$ design matrix for two-way interaction between group j and group j' with $(\boldsymbol{x}_{ij} \otimes \boldsymbol{x}_{ij'})^\top$ on row i, for $1 \leq j < j' \leq S$. Let y_i be the continuous outcome for subject i and let $\boldsymbol{y} = (y_1, \cdots, y_n)^\top$ be the outcome vector of length n. Without loss of generality, we center all variables and leave out intercept and covariates in the subsequent presentation. Also, we will consider the group structure defined as a set of basis functions representing nonlinear relationships.

We consider a linear regression model with main effect terms and all possible two-way interaction terms, that is

$$E(\boldsymbol{y}) = \sum_{j=1}^{S} \boldsymbol{X}_j \boldsymbol{\beta}_j + \sum_{j<j'} \boldsymbol{X}_{jj'} \boldsymbol{\gamma}_{jj'} \tag{5.4}$$

where $\boldsymbol{\beta}_j$ is the p_j-vector of main-effect coefficients for group j and $\boldsymbol{\gamma}_{jj'}$ is the $(p_j p_{j'})-$vector of coefficients for cross-product interaction between group j and group j'. We emphasize that we do not consider inter-group interaction and only consider inter-group interaction. The quadratic terms are not included either. In total, there are $\sum_{j=1}^{S} p_j$ main-effect terms and $\sum_{k=1}^{S-1} \sum_{l=k+1}^{S} p_k p_l$ interaction terms. The dimension grows linearly with group size p_j and grows quadratically with number of groups S. It can be large with moderate S and p_j. For example, if $S = 10$ and $p_j = 4\ \forall j$, there will be 40 main effect terms and 720 interaction terms. In many occasions, the OLS solution is not available. The goal is to incorporate a weight function to perform variable selection at a group level while maintaining the hereditary constraints.

In order to enforce heredity constraints, we rewrite (5.4) as

$$E(\boldsymbol{y}) = \sum_{j=1}^{S} \boldsymbol{X}_j \boldsymbol{\beta}_j + \sum_{j<j'} \boldsymbol{X}_{jj'} [\boldsymbol{\eta}_{jj'} \odot (\boldsymbol{\beta}_j \otimes \boldsymbol{\beta}_{j'})]$$

by reparametrizing $\boldsymbol{\gamma}_{jj'} = \boldsymbol{\eta}_{jj'} \odot (\boldsymbol{\beta}_j \otimes \boldsymbol{\beta}_{j'})$ for $1 \leq j < j' \leq S$ where $\boldsymbol{\eta}_{jj'}$ is a $(p_j p_{j'})-$vector of scalars for interaction between group j and group j'. Each interaction coefficient is written as the product of a scalar and its corresponding two main effect coefficients. Whenever $\boldsymbol{\beta}_j = \mathbf{0}$ and/or $\boldsymbol{\beta}_{j'} = \mathbf{0}$, $\boldsymbol{\gamma}_{jj'} = \mathbf{0}$. If $\boldsymbol{\gamma}_{jj'} \neq \mathbf{0}$, it implies that $\boldsymbol{\eta}_{jj'} \neq \mathbf{0}$,

$\boldsymbol{\beta}_j \neq \mathbf{0}$, and $\boldsymbol{\beta}_{j'} \neq \mathbf{0}$. Therefore, the heredity constraints are maintained.

For the purpose of variable selection with heredity constraints, we impose penalization on $\boldsymbol{\beta}_j$s and $\boldsymbol{\eta}_{jj'}$s rather than $\boldsymbol{\beta}_j$s and $\boldsymbol{\gamma}_{jj'}$s. We consider the penalized least squares criterion given by

$$\min_{\boldsymbol{\beta}_j, \boldsymbol{\eta}_{jj'}} \frac{1}{2}||\boldsymbol{y} - \sum_{j=1}^{S} \boldsymbol{X}_j\boldsymbol{\beta}_j - \sum_{j<j'} \boldsymbol{X}_{jj'}[\boldsymbol{\eta}_{jj'} \odot (\boldsymbol{\beta}_j \otimes \boldsymbol{\beta}_{j'})]||_2^2 + \lambda_1 \sum_{j=1}^{S} ||\boldsymbol{\beta}_j||_2 + \lambda_2 \sum_{j<j'} ||\boldsymbol{\eta}_{jj'}||_2. \tag{5.5}$$

λ_1 and λ_2 controls the amount of shrinkage at main-effect levels and interaction levels, respectively. When both $\boldsymbol{\beta}_j$ and $\boldsymbol{\beta}_{j'}$ are not equal to zero, the model has the flexibility of selecting main effects only or both main effects and interaction effects, with different λ_2 values.

Adaptive ideas have been widely used in previous literature [Zou, 2006, Wang et al., 2007, Zhang and Lu, 2007]. Penalizing different parameters differently can improve estimation efficiency. We apply the adaptive idea to improve criterion (5.5) as

$$\min_{\boldsymbol{\beta}_j, \boldsymbol{\eta}_{jj'}} \frac{1}{2}||\boldsymbol{y} - \sum_{j=1}^{S} \boldsymbol{X}_j\boldsymbol{\beta}_j - \sum_{j<j'} \boldsymbol{X}_{jj'}[\boldsymbol{\eta}_{jj'} \odot (\boldsymbol{\beta}_j \otimes \boldsymbol{\beta}_{j'})]||_2^2 + \lambda_1 \sum_{j=1}^{S} w_j||\boldsymbol{\beta}_j||_2 + \lambda_2 \sum_{j<j'} w_{jj'}||\boldsymbol{\eta}_{jj'}||_2. \tag{5.6}$$

where w_js and $w_{jj'}$s are pre-specified weights. The idea is to lightly penalize the coefficients with stronger effects and to heavily penalize the coefficients with weaker effects. Hence, the resulting coefficient estimates from criterion (5.6) are more coherent.

Following Breiman [1995], Zou [2006], and Wang and Leng [2008], the weights can be obtained using OLS estimates. Accounting for reparametrization, the weights can be computed as

$$w_j = \frac{1}{||\hat{\boldsymbol{\beta}}_j^{OLS}||_2^2}, w_{jj'} = \frac{1}{||\hat{\boldsymbol{\gamma}}_{jj'}^{OLS} \oslash (\hat{\boldsymbol{\beta}}_j^{OLS} \otimes \hat{\boldsymbol{\beta}}_{j'}^{OLS})||_2^2}$$

where $\hat{\boldsymbol{\beta}}_j^{OLS}$s and $\hat{\boldsymbol{\gamma}}_{jj'}^{OLS}$s are OLS estimates and $\oslash$ is defined as element-wise division operator for vectors. However, the number of effective predictors (i.e. $\sum_{j=1}^{S} p_j + \sum_{k=1}^{S-1} \sum_{l=k+1}^{S} p_k p_l$) can be larger than sample size n and the OLS estimates are not available. One alternative is to replace OLS estimates with ridge regression estimates [Choi et al., 2010]. The issue becomes the selection of tuning parameter for ridge regression. In addition, inconsistency of

ridge regression estimates prohibit an adaptive-LASSO-type approach to relish estimation consistency. We therefore consider an integrative approach to estimate weights and model parameters concurrently.

HiGLASSO is useful in identifying nonlinear main effects and nonlinear interaction effects under heredity constraints. If identifying the composition of nonlinearity is the additional aim, HiGLASSO incorporating inter-group sparsity may be considered. Criterion (5.6) can be extended as

$$\min_{\boldsymbol{\beta}_j, \boldsymbol{\eta}_{jj'}} \frac{1}{2}||\boldsymbol{y} - \sum_{j=1}^{S} \boldsymbol{X}_j\boldsymbol{\beta}_j - \sum_{j<j'} \boldsymbol{X}_{jj'}[\boldsymbol{\eta}_{jj'} \odot (\boldsymbol{\beta}_j \otimes \boldsymbol{\beta}_{j'})]||_2^2$$
$$+\lambda_1 \sum_{j=1}^{S} w_j||\boldsymbol{\beta}_j||_2 + \lambda_2 \sum_{j<j'} w_{jj'}||\boldsymbol{\eta}_{jj'}||_2 + \lambda_3 \sum_{j=1}^{S} ||\boldsymbol{\beta}_j||_1 + \lambda_4 \sum_{j<j'} ||\boldsymbol{\eta}_{jj'}||_1 \quad .$$

The two additional penalty terms induce sparsity at individual level in the spirit of sparse group LASSO [Friedman et al., 2010]. For instance, if a certain variable only has linear relationship with the response, all the variables corresponding to nonlinear effects within the group can be shrunk toward zero. However, introduction of two additional tuning parameters massively increase computational burden. Also, it is notoriously difficult to consider integrative weights for the two additional terms. This extension is beyond the scope of this chapter and we will focus on the HiGLASSO framework including only group-level selection and interaction in the rest of presentation.

5.3.3 Integrative Weight Function Approximation

To estimate weights and model parameters simultaneously, the first step is to specify the functional form of the weight functions. We consider weight functions based on the extreme values of each group, namely,

$$w_j \equiv \exp[-\frac{||\boldsymbol{\beta}_j||_\infty}{\sigma}] \text{ for } j = 1, \cdots, S \tag{5.7}$$

$$w_{jj'} \equiv \exp[-\frac{||\boldsymbol{\eta}_{jj'}||_\infty}{\sigma}] \text{ for } 1 \leq j < j' \leq S$$

where $||\boldsymbol{\mu}||_\infty$ is the L_∞ norm of $\boldsymbol{\mu}$ defined as the largest absolute element of $\boldsymbol{\mu}$ and σ is a pre-determined scale parameter. In most practical situation, fixing $\sigma = 1$ is sufficient. The weights are constructed in a way that a exponentially decayed weight in terms of the extremum of a group is assigned to the group. Some other common choices are L_0 norm, L_1 norm, and L_2 norm. The rationale of choosing L_∞ norm is that in many biological applications more emphasis is placed on largest effect within a group rather than 'average' effect [Pan and Zhao, 2016]. For example, in modeling the lagged effects of air pollution in association with some health outcomes, each group of variables may represent the lagged measurements of the same pollutant over time. The effect may peak at a certain point or linger over the entire period. The value of the largest effect across different time points is more relevant in the context of identifying important pollutants. In addition, since we do not impose sparsity within each group, the extremum of a group is more indicative than other 'averaging' quantities for assessing the effect size of the group.

The first term of criterion (5.6) involves the product of $\boldsymbol{\beta}_j$s and $\boldsymbol{\eta}_{jj'}$s. Thus, a $(S+1)-$step iterative approach to cycle through $\boldsymbol{\beta}_1, \cdots, \boldsymbol{\beta}_S$, and $\boldsymbol{\eta}_{jj'}$s until convergence is considered. We first describe in details how to optimize the criterion in terms of $\boldsymbol{\beta}_j$ given $\hat{\boldsymbol{\beta}}_{j'}$s with $j' \neq j$ and $\hat{\boldsymbol{\eta}}_{jj'}$s from previous steps in Section 5.3.3.1. How to obtain $\hat{\boldsymbol{\eta}}_{jj'}$ estimates given $\hat{\boldsymbol{\beta}}_j$s can be conducted in a similar fashion and will be outlined in Section 5.3.3.2.

5.3.3.1 Update Main-effect Coefficients

By substituting the specified weight function (5.7) into (5.6), given $\hat{\boldsymbol{\beta}}_{j'}$s with $j' \neq j$ and $\hat{\boldsymbol{\eta}}_{jj'}$s, the objective function can be expressed as

$$\min_{\boldsymbol{\beta}_j} \frac{1}{2}||\tilde{\boldsymbol{y}} - \tilde{\boldsymbol{X}}_j\boldsymbol{\beta}_j||_2^2 + \lambda_1 \exp[-\frac{||\boldsymbol{\beta}_j||_\infty}{\sigma}]||\boldsymbol{\beta}_j||_2 \tag{5.8}$$

such that

$$\tilde{\boldsymbol{y}} = \boldsymbol{y} - \sum_{k\neq j} \boldsymbol{X}_k\hat{\boldsymbol{\beta}}_k - \sum_{k,l\neq j} \boldsymbol{X}_{kl}[\hat{\boldsymbol{\eta}}_{kl} \odot (\hat{\boldsymbol{\beta}}_k \otimes \hat{\boldsymbol{\beta}}_l)]$$

$$\tilde{\boldsymbol{X}}_j = \boldsymbol{X}_j + \sum_{k<j} \boldsymbol{X}_{kj} \cdot \text{diag}(\hat{\boldsymbol{\eta}}_{kj})(\hat{\boldsymbol{\beta}}_k \otimes \boldsymbol{I}_{p_j}) + \sum_{l>j} \boldsymbol{X}_{jl} \cdot \text{diag}(\hat{\boldsymbol{\eta}}_{jl})(\boldsymbol{I}_{p_j} \otimes \hat{\boldsymbol{\beta}}_l)$$

where $\boldsymbol{I}_{p_j}$ is an identity matrix of dimension p_j. $\tilde{\boldsymbol{X}}_j$ and $\tilde{\boldsymbol{y}}$ represent the working design matrix and working response vector at current step. Technically, the minimization problem in (5.8) can be solved directly using gradient descent algorithm or Newton-Raphson algorithm [Gill et al., 1981].

Observing that the first term in (5.8) is in quadratic form of $\boldsymbol{\beta}_j$, a direct application of local quadratic approximation (LQA) proposed by Fan and Li [2001] can be used to solve the minimization problem with closed form. Let $\textbf{Pen}_1(\boldsymbol{\beta}_j)$ denote the penalty term in (5.8). If we apply LQA approach, $\textbf{Pen}_1(\boldsymbol{\beta}_j)$ can be approximated as

$$\textbf{Pen}_1(\boldsymbol{\beta}_j) \approx \textbf{Pen}_1(\hat{\boldsymbol{\beta}}_j^{(m)}) + \frac{1}{2}\sum_{k=1}^{p_j} d_{jk}^{(m)}[\beta_{jk}^2 - (\hat{\beta}_{jk}^{(m)})^2]$$

where β_{jk} is the k^{th} element of $\boldsymbol{\beta}_j$, $\hat{\boldsymbol{\beta}}_j^{(m)}$ is the estimate of $\boldsymbol{\beta}_j$ from m^{th} iteration, and d_{jk} is defined through $\frac{\partial \textbf{Pen}_1(\boldsymbol{\beta}_j)}{\partial \beta_{jk}} = d_{jk}\beta_{jk}$. By calculating the derivative of $\textbf{Pen}_1(\boldsymbol{\beta}_j)$, we have

$$d_{jk} = \begin{cases} \exp[-\frac{||\boldsymbol{\beta}_j||_\infty}{\sigma}](||\boldsymbol{\beta}_j||_2)^{-1} & \text{if } |\beta_{jk}| \neq ||\boldsymbol{\beta}_j||_\infty \\ \exp[-\frac{||\boldsymbol{\beta}_j||_\infty}{\sigma}](||\boldsymbol{\beta}_j||_2)^{-1} - \exp[-\frac{||\boldsymbol{\beta}_j||_\infty}{\sigma}]||\boldsymbol{\beta}_j||_2(|\beta_{jk}|\sigma)^{-1} & \text{if } |\beta_{jk}| = ||\boldsymbol{\beta}_j||_\infty. \end{cases} \tag{5.9}$$

The problem with LQA approximation is that when $|\beta_{jk}| = ||\boldsymbol{\beta}_j||_\infty$, d_{jk} might be negative and there is no guarantee that the approximated $\textbf{Pen}_j(\boldsymbol{\beta}_j)$ to be convex still.

Pan and Zhao [2016] proposed generalized local quadratic approximation (GLQA) to operate convex quadratic approximation for a penalty function. Let $\mathcal{P}_1(\boldsymbol{\beta}_j)$ denote GLQA of $\textbf{Pen}_1(\boldsymbol{\beta}_j)$. Three preferred properties of $\mathcal{P}_1(\boldsymbol{\beta}_j)$ are

1. $\mathcal{P}_1(\boldsymbol{\beta}_j)$ is convex
2. $\mathcal{P}_1(\hat{\boldsymbol{\beta}}_j^{(m)}) = \textbf{Pen}_1(\hat{\boldsymbol{\beta}}_j^{(m)})$
3. $\frac{\partial \mathcal{P}_1(\boldsymbol{\beta}_j)}{\partial \beta_{jk}}|_{\beta_{jk}=\hat{\beta}_{jk}^{(m)}} = \frac{\partial \textbf{Pen}_1(\boldsymbol{\beta}_j)}{\partial \beta_{jk}}|_{\beta_{jk}=\hat{\beta}_{jk}^{(m)}} \; \forall\, k.$

Essentially, $\mathcal{P}_1(\boldsymbol{\beta}_j)$ has to be a convex quadratic function of $\boldsymbol{\beta}_j$ and the functional value up

to first derivative measured at $\hat{\boldsymbol{\beta}}_j^{(m)}$ should retain. A natural choice is of the form

$$\mathcal{P}_1(\boldsymbol{\beta}_j) = \mathbf{Pen}_1(\hat{\boldsymbol{\beta}}_j^{(m)}) + \frac{1}{2}\sum_{k=1}^{p_j} |d_{jk}^{(m)}|[(\beta_{jk}^2 + c_1)^2 + c_2].$$

c_1 and c_2 can be solved using the second and third conditions as above. The resulting choice of $\mathcal{P}_1(\boldsymbol{\beta}_j)$ becomes

$$\mathcal{P}_1(\boldsymbol{\beta}_j) = \mathbf{Pen}_1(\hat{\boldsymbol{\beta}}_j^{(m)}) + \frac{1}{2}\sum_{k=1}^{p_j} |d_{jk}^{(m)}|[(\beta_{jk}^2 - (1 - \frac{d_{jk}^{(m)}}{|d_{jk}^{(m)}|})\hat{\beta}_{jk}^{(m)})^2 - (\hat{\beta}_{jk}^{(m)})^2].$$

If we rewrite the $\mathcal{P}_1(\boldsymbol{\beta}_j)$ in matrix form, (5.8) can be approximated as

$$\frac{1}{2}||\tilde{\boldsymbol{y}} - \tilde{\boldsymbol{X}}_j\boldsymbol{\beta}_j||_2^2 + \frac{1}{2}\lambda_1\boldsymbol{\beta}_j^\top \boldsymbol{D}_j^{(m)}\boldsymbol{\beta}_j - \lambda_1 \boldsymbol{c}^{(m)\top}\boldsymbol{\beta}_j + \text{Constant}$$

where $\boldsymbol{D}_j^{(m)}$ =diag$[(d_{j1}^{(m)}, \cdots, d_{jp_j}^{(m)})]$ and $\boldsymbol{c}^{(m)} = ((|d_{j1}^{(m)}| - d_{j1}^{(m)})\hat{\beta}_{j1}^{(m)}, \cdots, (|d_{jp_j}^{(m)}| - d_{jp_j}^{(m)})\hat{\beta}_{jp_j}^{(m)})^\top$. $\boldsymbol{\beta}_j$ can be updated in closed-form as

$$\hat{\boldsymbol{\beta}}_j = (\tilde{\boldsymbol{X}}_j^\top \tilde{\boldsymbol{X}}_j + \lambda_1 \boldsymbol{D}_j^{(m)})^{-1}(\tilde{\boldsymbol{X}}_j^\top \tilde{\boldsymbol{y}} + \lambda_1 \cdot \boldsymbol{c}^{(m)}). \tag{5.10}$$

5.3.3.2 Update Interaction Scalars

By substituting the specified weight function (5.3.3) into (5.6), given $\hat{\boldsymbol{\beta}}_j$s, the objective function can be expressed as

$$\min_{\boldsymbol{\beta}_j} \frac{1}{2}||\tilde{\boldsymbol{y}} - \sum_{j<j'} \tilde{\boldsymbol{X}}_{jj'}\boldsymbol{\eta}_{jj'}||_2^2 + \lambda_2 \sum_{j<j'} \exp[-\frac{||\boldsymbol{\eta}_{jj'}||_\infty}{\sigma}]||\boldsymbol{\eta}_{jj'}||_2 \tag{5.11}$$

where

$$\tilde{\boldsymbol{y}} = \boldsymbol{y} - \sum_{k=1}^{S} \boldsymbol{X}_k\hat{\boldsymbol{\beta}}_k$$

$$\tilde{\boldsymbol{X}}_{jj'} = \boldsymbol{X}_{jj'}\text{diag}[(\hat{\boldsymbol{\beta}}_j \otimes \hat{\boldsymbol{\beta}}_{j'})] \text{ for } 1 \leq j < j' \leq S.$$

Let $\mathbf{Pen}_2(\boldsymbol{\eta}_{jj'})$ denote the individual penalty term in (5.11) and let $\mathcal{P}_2(\boldsymbol{\beta}_{jj'})$ denote

GLQA of $\mathbf{Pen}_2(\boldsymbol{\eta}_{jj'})$. We have

$$\mathcal{P}_2(\boldsymbol{\eta}_{jj'}) = \mathbf{Pen}_1(\hat{\boldsymbol{\eta}}_{jj'}^{(m)}) + \frac{1}{2}\sum_{k=1}^{p_j p_{j'}} |d_{jj'k}^{(m)}|[(\eta_{jj'k}^2 - (1 - \frac{d_{jj'k}^{(m)}}{|d_{jj'k}^{(m)}|})\hat{\eta}_{jj'k}^{(m)})^2 - (\hat{\eta}_{jj'k}^{(m)})^2]$$

where $\eta_{jj'k}$ is the k^{th} element of $(p_j p_{j'})-$vector of $\boldsymbol{\eta}_{jj'}$ and $d_{jj'k}$ is similarly defined through $\frac{\partial \mathbf{Pen}_2(\boldsymbol{\eta}_{jj'})}{\partial \eta_{jj'k}} = d_{jj'k}\eta_{jj'k}$ as (5.9). (5.11) can be approximated as

$$\frac{1}{2}||\tilde{\boldsymbol{y}} - \tilde{\boldsymbol{X}}\boldsymbol{\eta}||_2^2 + \frac{1}{2}\lambda_2 \boldsymbol{\eta}^\top \boldsymbol{D}^{(m)}\boldsymbol{\eta} - \lambda_2 \boldsymbol{C}^{(m)\top}\boldsymbol{\eta} + \text{Constant}$$

where $\tilde{\boldsymbol{X}} = [\tilde{\boldsymbol{X}}_{12}, \cdots, \tilde{\boldsymbol{X}}_{S-1,S}]$, $\boldsymbol{\eta} = (\boldsymbol{\eta}_{12}^\top, \cdots, \boldsymbol{\eta}_{S-1,S}^\top)^\top$, $\boldsymbol{D}^{(m)} = \text{diag}[(d_{121}^{(m)}, \cdots, d_{12(p_1p_2)}^{(m)}, \cdots, d_{(S-1)S(p_{S-1}p_S)}^{(m)}]$, and $\boldsymbol{C}^{(m)}$ is a $[S(S-1)/2] \times [\sum_{j<j'} p_j p_{j'}]$ block diagonal matrix such that the block corresponding to the interaction between group j and group j' is defined as a row vector of length $p_j p_{j'}$ with k^{th} element $(|d_{jj'k}^{(m)}| - d_{jj'k}^{(m)})\hat{\eta}_{jj'k}^{(m)}$. $\boldsymbol{\eta}_{jj'}$s can then be updated in closed form as

$$\hat{\boldsymbol{\eta}} = (\tilde{\boldsymbol{X}}^\top \tilde{\boldsymbol{X}} + \lambda_2 \boldsymbol{D}^{(m)})^{-1}(\tilde{\boldsymbol{X}}^\top \tilde{\boldsymbol{y}} + \lambda_2 \cdot \boldsymbol{C}^{(m)}). \tag{5.12}$$

5.3.4 Algorithm

In previous section, we provided mathematical details about how to update individual $\boldsymbol{\beta}_j$ and all $\boldsymbol{\eta}_{jj'}$s at each iteration. Here we describe the full algorithm for estimating $\boldsymbol{\beta}_j$s and $\boldsymbol{\eta}_{jj'}$s i HiGLASSO. Since the least squares criterion involves the product of $\boldsymbol{\beta}_j$ and $\boldsymbol{\eta}_{jj'}$, an iterative approach is employed. We first fix $\boldsymbol{\eta}_{jj'}$ to estimate $\boldsymbol{\beta}_j$, then fix $\boldsymbol{\beta}_j$ to estimate $\boldsymbol{\eta}_{jj'}$, and iterate the two steps until converge. The entire algorithm proceeds as follows:

Step 1: Orthogonalize main-effect design matrices $\boldsymbol{X}_j$ for $j = 1, \cdots, S$ and interaction design matrices $\boldsymbol{X}_{jj'}$ for $1 \le j < j' \le S$ and center response vector $\boldsymbol{y}$.

Step 2: Initial $\hat{\boldsymbol{\beta}}_j^{(0)}$ for $j = 1, \cdots, S$ and $\hat{\boldsymbol{\eta}}_{jj'}^{(0)}$ for $1 \le j < j' \le S$. Set $m = 1$.

Step 3: For each j in $1, \cdots, S$, $\hat{\boldsymbol{\beta}}_j^{(m)}$ is updated via closed-form formula in (5.10), given $\hat{\boldsymbol{\eta}}_{kj}^{(m-1)}$s and $\hat{\boldsymbol{\beta}}_k^{(m)}$s for $k < j$, and $\hat{\boldsymbol{\eta}}_{jl}^{(m-1)}$s $\hat{\boldsymbol{\beta}}_l^{(m-1)}$s for $l > j$. Backtracking line

search algorithm is followed to guarantee that $\hat{\boldsymbol{\beta}}_j^{(m)}$ leads to a lower value of the objective function (5.8) than $\hat{\boldsymbol{\beta}}_j^{(m)}$.

Step 4: Given $\hat{\boldsymbol{\beta}}_j^{(m)}$s for $j = 1, \cdots, S$, $\hat{\boldsymbol{\eta}}_{jj'}^{(m)}$s are updated via closed-form formula in (5.12). Backtracking line search algorithm is followed to guarantee that $\hat{\boldsymbol{\eta}}_{jj'}^{(m)}$s leads to a lower value of the objective function (5.11) than $\hat{\boldsymbol{\eta}}_{jj'}^{(m-1)}$s.

Step 5: Stop if the percentage change of penalized likelihood value is less than a pre-specified margin δ, namely

$$\frac{P_n^{(m-1)} - P_n^{(m)}}{P_n^{(m-1)}} < \delta.$$

where $P_n^{(m)}$ is the value of (5.6) evaluated at $\hat{\boldsymbol{\beta}}_j^{(m)}$s and $\hat{\boldsymbol{\eta}}_{jj'}^{(m)}$s and $\hat{\boldsymbol{\eta}}_{jj'}^{(m)}$s. Otherwise, let $m = m + 1$ and repeat **Step** 3 and **Step** 4.

A common choice of $\hat{\boldsymbol{\beta}}_j^{(0)}$s and $\hat{\boldsymbol{\eta}}_{jj'}^{(0)}$s is group LASSO estimator. Since we utilize surrogate penalties to approximate the original penalties, there is no guarantee that each of the $S+1$ updates decreases the value of penalized least squares criterion. In addition, GLQA is more accurate in the neighborhood of previous update. We employ backtracking line search algorithm [Dennis Jr and Schnabel, 1996] to ensure that the penalized least squares criterion monotonically decreases throughout the entire procedure. Backtracking line search method to determine the maximum amount to move along a given search direction based on the Armijo-Goldstein condition [Armijo, 1966]. At each update, take the quantity obtained from the closed-form formula (5.10) or (5.12) as a candidate. If the candidate does not result to a decrease of the objective function, iteratively shrink the step size between the previous update and the candidate quantity until a decrease of the objective function is observed. At each step throughout the algorithm, now the value of the objective function decreases so the solution is guaranteed to converge.

5.3.5 Asymptotic Properties

The theoretical properties of HiGLASSO are introduced in this section. For the purpose of presentation, let $\boldsymbol{\Theta}$ denote the vector of all coefficients, including main-effect coefficients and interaction coefficients in original scale. Namely, $\boldsymbol{\Theta} = (\boldsymbol{\beta}^\top, \boldsymbol{\eta}^\top)^\top$ where

$\boldsymbol{\beta} = (\boldsymbol{\beta}_1^\top, \cdots, \boldsymbol{\beta}_S^\top)^\top$ and $\boldsymbol{\eta} = (\boldsymbol{\eta}_{12}^\top, \cdots, \boldsymbol{\eta}_{S-1,S}^\top)^\top$. Without loss of generality, we rearrange the group indices to facilitate the proofs. Let first $s_0 \leq S$ groups of predictors have nonzero main effects. Suppose there are i_0 nonzero two-way interaction terms out of at most $s_0(s_0-1)/2$ possible pairs under strong heredity constraints and let $\mathcal{I}$ denote the set of (j, j') pairs with nonzero interaction effects. Denote $\boldsymbol{\beta}^{(1)} = (\boldsymbol{\beta}_1^\top, \cdots, \boldsymbol{\beta}_{s_0}^\top)^\top$ and $\boldsymbol{\beta}^{(0)} = (\boldsymbol{\beta}_{s_0+1}^\top, \cdots, \boldsymbol{\beta}_S^\top)^\top$. Likewise, let $\boldsymbol{\gamma}^{(1)}$ be a vector concatenating all nonzero interaction vectors and let $\boldsymbol{\gamma}^{(0)}$ be a vector concatenating all irrelevant interaction vectors. Let $a_n = \max(\lambda_1, \lambda_2)$. The subscript n in a_n reflects the fact that the tuning parameter λ_1 and λ_2 depend on sample size. Similarly, $\sigma \equiv \sigma_n$.

Theorem 1: Assume $a_n n^{-1/2} \to_p 0$ and $\sigma_n \to 0$ as $n \to \infty$. Under the regular conditions (A)-(D) of Andersen and Gill [1982], there exists a local minimizer such that $||\hat{\boldsymbol{\Theta}} - \boldsymbol{\Theta}||_2 = O_p(n^{-1/2})$.

Proof: We first rewrite penalized least squares function (5.6) as

$$Q(\boldsymbol{\Theta}) = -l(\boldsymbol{\Theta}) + \lambda_1 \sum_{j=1}^{S} w_j(\boldsymbol{\beta}_j)||\boldsymbol{\beta}_j||_2 + \lambda_2 \sum_{j<j'} w_{jj'}(\boldsymbol{\eta}_{jj'})||\boldsymbol{\eta}_{jj'}||_2$$

where $l(\boldsymbol{\Theta})$ denotes the log-likelihood of $\boldsymbol{\Theta}$, corresponding to the least square term. $\mathbb{X} = [\boldsymbol{X}_1, \cdots, \boldsymbol{X}_S, \boldsymbol{X}_{12}, \cdots, \boldsymbol{X}_{S-1,S}]$. Let $\nabla l(\boldsymbol{\Theta}) = \frac{\partial}{\partial \boldsymbol{\Theta}} l(\boldsymbol{\Theta})$ and $\nabla^2 l(\boldsymbol{\Theta}) = \frac{\partial^2}{\partial \boldsymbol{\Theta} \partial \boldsymbol{\Theta}^\top} l(\boldsymbol{\Theta})$. we have $\nabla l(\boldsymbol{\Theta}) = \frac{\partial}{\partial \boldsymbol{\Theta}} l(\boldsymbol{\Theta}) = O_p(\sqrt{n})$ and $\nabla^2 l(\boldsymbol{\Theta}) - I(\boldsymbol{\Theta}) = O_p(n)$ where $I(\boldsymbol{\Theta}) = E[\nabla l(\boldsymbol{\Theta}) \nabla l(\boldsymbol{\Theta})^\top]$. Define $\boldsymbol{\mu} = (\boldsymbol{\mu}_1^\top, \cdots, \boldsymbol{\mu}_S^\top, \boldsymbol{\mu}_{12}^\top, \cdots, \boldsymbol{\mu}_{S-1,S}^\top)^\top$ as a vector of length

$\sum_{j=1}^{S} p_j + \sum_{k=1}^{S-1} \sum_{l=k+1}^{S} p_k p_l$ representing departure from underlying true $\boldsymbol{\Theta}$.

$$
\begin{aligned}
& Q(\boldsymbol{\Theta} + n^{-1/2}\boldsymbol{\mu}) - Q(\boldsymbol{\Theta}) \\
=& - l(\boldsymbol{\Theta} + n^{-1/2}\boldsymbol{\mu}) + l(\boldsymbol{\Theta}) + \lambda_1 \sum_{j=1}^{S} w_j(\boldsymbol{\beta}_j + n^{-1/2}\boldsymbol{\mu}_j)||\boldsymbol{\beta}_j + n^{-1/2}\boldsymbol{\mu}_j||_2 - \lambda_1 \sum_{j=1}^{S} w_j(\boldsymbol{\beta}_j)||\boldsymbol{\beta}_j||_2 \\
& + \lambda_2 \sum_{j<j'} w_{jj'}(\boldsymbol{\eta}_{jj'} + n^{-1/2}\boldsymbol{\mu}_{jj'})||(\boldsymbol{\eta}_{jj'} + n^{-1/2}\boldsymbol{\mu}_{jj'})||_2 - \lambda_2 \sum_{j<j'} w_{jj'}(\boldsymbol{\eta}_{jj'})||\boldsymbol{\eta}_{jj'}||_2 \\
=& - l(\boldsymbol{\Theta} + n^{-1/2}\boldsymbol{\mu}) + l(\boldsymbol{\Theta}) + \lambda_1 \sum_{j=1}^{S} w_j(\boldsymbol{\beta}_j + n^{-1/2}\boldsymbol{\mu}_j)||\boldsymbol{\beta}_j + n^{-1/2}\boldsymbol{\mu}_j||_2 - \lambda_1 \sum_{j=1}^{s_0} w_j(\boldsymbol{\beta}_j)||\boldsymbol{\beta}_j||_2 \\
& + \lambda_2 \sum_{j<j'} w_{jj'}((\boldsymbol{\eta}_{jj'} + n^{-1/2}\boldsymbol{\mu}_{jj'}))||(\boldsymbol{\eta}_{jj'} + n^{-1/2}\boldsymbol{\mu}_{jj'})||_2 - \sum_{j,j' \in \mathcal{I}} w_{jj'}(\boldsymbol{\eta}_{jj'})||\boldsymbol{\eta}_{jj'}||_2 \\
\geq& - l(\boldsymbol{\Theta} + n^{-1/2}\boldsymbol{\mu}) + l(\boldsymbol{\Theta}) + \lambda_1 \sum_{j=1}^{s_0} w_j(\boldsymbol{\beta}_j)(||\boldsymbol{\beta}_j + n^{-1/2}\boldsymbol{\mu}_j||_2 - ||\boldsymbol{\beta}_j||_2) \\
& + \lambda_2 \sum_{j,j' \in \mathcal{I}} w_{jj'}(\boldsymbol{\eta}_{jj'})(||\boldsymbol{\eta}_j + n^{-1/2}\boldsymbol{\mu}_j||_2 - ||\boldsymbol{\eta}_j||_2) + o_p(1) \\
\geq& - l(\boldsymbol{\Theta} + n^{-1/2}\boldsymbol{\mu}) + l(\boldsymbol{\Theta}) - s_0\lambda_1\sqrt{n}||\boldsymbol{\mu}||_2 - i_0\lambda_2\sqrt{n}||\boldsymbol{\mu}||_2 + o_p(1) \\
=& - n^{-1/2}[\nabla l(\boldsymbol{\Theta})]^\top \boldsymbol{\mu} - \frac{1}{2n}\boldsymbol{\mu}^\top[\nabla^2 l(\boldsymbol{\Theta})]\boldsymbol{\mu}[1 + o_p(1)] - s_0\lambda_1\sqrt{n}||\boldsymbol{\mu}||_2 - i_0\lambda_2\sqrt{n}||\boldsymbol{\mu}||_2 + o_p(1) \\
=& - O_p(1)\boldsymbol{\mu} + \frac{1}{2}\boldsymbol{\mu}^\top I(\boldsymbol{\Theta})\boldsymbol{\mu}[1 + o_p(1)] - s_0\lambda_1\sqrt{n}||\boldsymbol{\mu}||_2 - i_0\lambda_2\sqrt{n}||\boldsymbol{\mu}||_2 + o_p(1)
\end{aligned}
\tag{5.13}
$$

Under the condition that $a_n n^{-1/2} \to_p 0$, we have the last three terms of (5.13) converge in probability to a constant, that is

$$-s_0\lambda_1\sqrt{n}||\boldsymbol{\mu}||_2 - i_0\lambda_2\sqrt{n}||\boldsymbol{\mu}||_2 + o_p(1) = o_p(1).$$

The first term converges in probability to a linear function of $\boldsymbol{\mu}$ and the second term is quadratic in $\boldsymbol{\mu}$ and dominates the rest terms. Subsequently, for any $\epsilon > 0$, we always can find a sufficiently large constant C such that

$$\liminf_n P[\inf_{||\boldsymbol{\mu}||_2 = C} Q(\boldsymbol{\Theta} + n^{-1/2}\boldsymbol{\mu}) > Q(\boldsymbol{\Theta})] > 1 - \epsilon.$$

Hence, there exists a local minimizer such that $||\hat{\boldsymbol{\Theta}} - \boldsymbol{\Theta}||_2 = O_p(n^{-1/2})$.

Theorem 2 (Sparsity): Assume $a_n n^{-1/2} \to_p 0$ and $\sigma_n \to 0$ as $n \to \infty$. Under the regular conditions (A)-(D) of Andersen and Gill [1982], the local minimizer in Theorem 1 satisfies $P(||\hat{\boldsymbol{\beta}}^{(0)}||_2 = 0) \to 1$ and $P(||\hat{\boldsymbol{\gamma}}^{(0)}||_2 = 0) \to 1$.

Proof: Let $\boldsymbol{X}_M^{(0)}$ and $\boldsymbol{X}_M^{(1)}$ denote the main-effect design matrices corresponding to $\boldsymbol{\beta}^{(0)}$ and $\boldsymbol{\beta}^{(1)}$, respectively. Let $\boldsymbol{X}_I^{(0)}$ and $\boldsymbol{X}_I^{(1)}$ denote the main-effect design matrices corresponding to $\boldsymbol{\gamma}^{(0)}$ and $\boldsymbol{\gamma}^{(1)}$, respectively. Denote $\boldsymbol{d}_{M1}$ and $\boldsymbol{d}_{M2}$ denote the first-order derivative of the first and second penalty terms in (5.6) with respect to $\boldsymbol{\beta}^{(0)}$. Likewise, denote $\boldsymbol{d}_{I1}$ and $\boldsymbol{d}_{I2}$ denote the first-order derivative of the first and second penalty terms with respect to $\boldsymbol{\gamma}^{(0)}$. The score equation to solve for $\hat{\boldsymbol{\beta}}^{(0)}$ is given by

$$
\begin{aligned}
&\boldsymbol{X}_M^{(0)\top}(\boldsymbol{y} - \mathbb{X}\hat{\boldsymbol{\Theta}}) + \lambda_1 \hat{\boldsymbol{d}}_{M1} + \lambda_2 \hat{\boldsymbol{d}}_{M2} = \mathbf{0} \\
\to &\frac{1}{\sqrt{n}} \boldsymbol{X}_M^{(0)\top}(\boldsymbol{y} - \mathbb{X}\boldsymbol{\Theta}) - (\frac{1}{n}\boldsymbol{X}_M^{(0)\top}\boldsymbol{X}_M^{(0)})\sqrt{n}(\hat{\boldsymbol{\beta}}^{(0)} - \boldsymbol{\beta}^{(0)}) - (\frac{1}{n}\boldsymbol{X}_M^{(0)\top}\boldsymbol{X}_M^{(1)})\sqrt{n}(\hat{\boldsymbol{\beta}}^{(1)} - \boldsymbol{\beta}^{(1)}) \\
&- (\frac{1}{n}\boldsymbol{X}_M^{(0)\top}\boldsymbol{X}_I^{(0)})\sqrt{n}(\hat{\boldsymbol{\gamma}}^{(0)} - \boldsymbol{\gamma}^{(0)}) - (\frac{1}{n}\boldsymbol{X}_M^{(0)\top}\boldsymbol{X}_I^{(1)})\sqrt{n}(\hat{\boldsymbol{\gamma}}^{(1)} - \boldsymbol{\beta}^{(1)}) \\
&+ \frac{1}{\sqrt{n}}\lambda_1 \hat{\boldsymbol{d}}_{M1} + \frac{1}{\sqrt{n}}\lambda_2 \hat{\boldsymbol{d}}_{M2} = \mathbf{0} \qquad (5.14)
\end{aligned}
$$

The first five terms in (5.14) are $O_p(n^{-1/2})$ according to $\sqrt{n}$ estimation consistency shown in Theorem 1. If $\hat{\boldsymbol{\beta}}^{(0)} \neq \mathbf{0}$, $\lambda_1 w_j/\sqrt{n}$ and $\lambda_2 w_{jj'}/\sqrt{n}$ go to infinity. Therefore, the last two terms go to infinity with probability 1 and they dominate the score equation. In other words, the score equation cannot be satisfied with sufficiently large sample size. Therefore, we conclude that $P(||\hat{\boldsymbol{\beta}}^{(0)}||_2 = 0) \to 1$. $P(||\hat{\boldsymbol{\gamma}}^{(0)}||_2 = 0) \to 1$.

5.4 Simulation Study

We compare the performance of HiGLASSO to other alternative approaches for selecting main effects and interaction effects. Here we emphasize that the group structure we consider is defined through a set of basis functions representing nonlinearity for all group-based approaches. The competing methods accounting for linear main-effect terms and

linear pairwise interaction terms include multiple regression (MR-lin), LASSO (LASSO-lin), adaptive LASSO (aLASSO-lin), and hierNet (hierNet). The competing alternatives accounting for nonlinear main-effect terms and all pairwise interaction terms include multiple regression (MR-nonlin), LASSO (LASSO-nonlin), adaptive LASSO (aLASSO-nonlin), group LASSO (gLASSO), adaptive group LASSO (agLASSO), and VANISH (VANISH). Each individual variable in methods accounting for linear effects only is expanded to three variables in methods accounting for nonlinear effects using cubic splines. Specifically, each scalar variable x_j is expanded to $(x_j, x_j^2, x_j^3)^\top$, In other words, $p_s = 3\ \forall s = 1, \cdots, S$ for all nonlinear methods with group structure including gLASSO, agLASSO, VANISH, and HiGLASSO. We use R package glmnet to implement LASSO-lin, aLASSO-lin, LASSO-nonlin, and aLASSO-nonlin, R package hierNet to implement hierNet, and R package gglasso to implement gLASSO and agLASSO. For VANISH implementation, we used the R script provided by the author Radchenko and James [2010]. The iterative algorithm for HiGLASSO is implemented in R.

5.4.1 Simulation Setting

We consider 10 simulation scenarios in total to compare different variable selection methods. In each scenario, we generate 100 simulated data sets with $n = 200$ and number of predictors (p) equal to 4 or 8 using regression model

$$\boldsymbol{y} = f(\boldsymbol{x}_1, \cdots, \boldsymbol{x}_p) + \boldsymbol{\epsilon}$$

where $\boldsymbol{\epsilon} \sim N(\mathbf{0}, \sigma^2 \boldsymbol{I}_n)$ with $\sigma^2 = 9$ or 36. All predictor values are independently generated from a standard normal distribution. Different specifications of mean function $f(\cdot)$ across different simulation scenarios are considered and the exact specifications are presented in Table 5.1. Considering the cubic spline expansion with two-way interaction, $p = 4$ and $p = 8$ correspond to 66 and 276 effective predictors, respectively. In the latter case, the number of effective predictors is greater than sample size $n = 200$ so MR-nonlin fit is not available and consequently aLASSO-nonlin and agLASSO fits are not available either. We use out-of-sample mean squared prediction error (MSPE) to choose tuning parameter

for all methods except MR-lin and MR-nonlin. Two separate data sets are generated for each repetition in each simulation scenario. Given a value of tuning parameter, MSPE is computed as the MSE of the second data set based on the coefficient estimate obtained from the first data set. The model with the tuning parameter corresponding to the lowest MSPE is the final fit.

5.4.2 Evaluation Metrics

We present the simulation results based on the following five metrics.

1. **False negative main effects rate (FNM)**: average number of times of non-null main-effect terms not selected by a model.

2. **False positive main effects rate (FPM)**: average number of times of null main-effect terms selected by a model.

3. **False negative interaction effects rate (FNI)**: average number of times of non-null interaction-effect terms not selected by a model.

4. **False positive interaction effects rate (FPI)**: average number of times of null interaction-effect terms selected by a model.

5. Number of occurrences of violating strong heredity constraints.

MR-lin and MR-nonlin do not induce sparsity. Hence, we rely on p-value with a cutoff at 0.05 to determine whether particular term(s) are selected or not. Specifically, a full model with the term being tested and a reduced model without the term are fitted. A F-test is subsequently conducted to obtain the p-value. The first four metrics are scaled to a range between 0 and 1, reflecting the average error rate per simulated data set and per important/unimportant term. Note that smaller values of all five metrics indicate superior performance.

5.4.3 Simulation Results

We present the results in Figure 5.1 - 5.5. The left and right plots in Figure 5.1 refer to the scenario with linear main effects only and the scenario with linear main and interaction effects, respectively. As we can see that all linear methods are able to detect the important main effects and interaction effects. Among the nonlinear methods, MR-nonlin and VANISH frequently miss the signal. HiGLASSO has lower FNM in selecting main effects but higher FNI in selecting interaction effects than agLASSO. In terms of false discovery, all methods except MR-nonlin, VANISH, and HiGLASSO select unimportant terms at an unignorable frequency. VANISH is extremely conservative and HiGLASSO appears to be slightly conservative in this case. Figure 5.2 refers to the same scenarios as Figure 5.1 except that the signal now becomes weaker. We can observe that all nonlinear methods suffer from power loss in capturing the linear main and interaction effects. Linear methods except MR-lin have better performance of not missing the important signals, whereas they select unimportant variables occasionally as well.

Figure 5.3 refers to the scenario with nonlinear main effects only (left) and the scenario with nonlinear main and interaction effects (right). All linear methods and nonlinear methods except MR-lin successfully detect the relevant main effects and interaction effects. However, except MR-nonlin and HiGLASSO, all methods have large false discover rate. MR-nonlin and HiGLASSO are top performers in this case. Figure 5.4 presents the results when true model is under weak heredity constraints (left) and is violating heredity constraints (right). As expected, HiGLASSO rarely selects the relevant interaction terms due to the strong heredity constraints. On the other hand, hierNet never misses the important signals but it raises FNM and FNI to ensure that strong heredity constraints are maintained. Figure 5.5 refers to similar scenarios as Figure 5.3 except that the number of predictors is larger and now the number of effective parameters is greater than the sample size. As we can observe, all methods except VANISH successfully identify important main and interaction effects. However, HiGLASSO is the only method capable of controlling false discovery rate at a minimum level.

We summarized the counts of violation of strong heredity constraints out of 100 data

repetitions in Table 5.2. As expected, hierNet, VANISH, and HiGLASSO never violate the constraints by construction. All other methods frequently select an interaction term without having both of its corresponding main effects terms in the model. Most of the selected non-hierarchical interaction terms are false discovery. In summary, performance of HiGLASSO is very competitive across all the simulation settings in terms of various metrics. The major strength of HiGLASSO is that its false identification of unimportant effects (based on FNM and FNI) is much smaller compared to the other methods. It also has comparable identification rate of important effects (based on FPM and FPI). In addition, HiGLASSO admits a simple characterization of imposing heredity. If the interaction effects in true model does not violate strong heredity constraints, HiGLASSO is superior to other alternatives.

5.5 Application to NMMAPS

5.5.1 Data Overview and Modeling

We first illustrate HiGLASSO and other competing variable selection approaches using NMMAPS data. Daily time series of five pollutants ($S = 5$) including (1) PM_{10}, (2) O_3, (3) SO_2, (4) NO_2, and (5) CO with pairwise interactions are jointly modelled in associated with non-accidental mortality counts in Chicago, Illinois between 1987 and 2000. The goal is to identify important pollutant(s) and pollutant-pollutant interaction(s) that contribute the most in explaining the variability in mortality counts. To avoid repetition, we refer to http://www.ihapss.jhsph.edu/data/NMMAPS/ for data configuration in details.

Following Zanobetti et al. [2000], we set $L_s = 14$ for $s = 1, \cdots, 5$. Since the most commonly seen patterns of the estimated DL function are either monotonic decreasing over time or increasing at early lags and decreasing at later lags, cubic polynomial is able to capture the feature, including the possible mortality displacement, and is the most common parametric choice of the DL function in quantifying the short-term lagged effects of air pollution [Zanobetti and Schwartz, 2008, Bhaskaran et al., 2013]. We therefore choose

the DL functions of the five pollutants to follow a cubic polynomial (with intercept). In other words, a 15×4 transformation matrix $\boldsymbol{C}_s$ is applied to vector of lagged coefficients for $s = 1, \cdots, 5$. In addition, transformation matrix $\boldsymbol{C}_{ss'} = \boldsymbol{C}_s \otimes \boldsymbol{C}_{s'}$ is applied to the vector of interaction effects between pollutant s and pollutant s', for $1 \leq s < s' \leq 5$. The covariates are adjusted in the same way as Dominici et al. [2005] and we refer to Sections 2.5 and 4.4 for the details.

We emphasize that the group structure of a pollutant in this application corresponds to its serial measurements. We still only consider the linear association between pollutants and mortality. LASSO (R package `glmnet`), group LASSO (R package `grplasso`), and HiGLASSO are applied to the data set. Adaptive LASSO, adaptive group LASSO, and VANISH are excluded because the R package/function does not support Poisson loglinear model. We use 5-fold cross-validation to select tuning parameter.

5.5.2 Variable Selection Results

We present the selection results in Table 5.3. Five main-effect terms and 10 interaction terms are ranked from top to bottom based on their importance in each of the three methods. The terms selected by individual methods are highlighted in bold. In summary, 3, 4, 4 main-effect terms and 6, 6, 0 interaction terms are selected by LASSO, group LASSO, and HiGLASSO, respectively.

HiGLASSO selects all pollutants except NO_2, without selecting any crossproduct terms. We can observe that LASSO and group LASSO select more variables than HiGLASSO, akin to the observations from simulation study. Apparently, LASSO and group LASSO are not subject to heredity constraints and some of the interaction terms are present in the model prior to their corresponding main effect terms. Despite disparate orderings across the three methods, the results suggest stronger effects from PM_{10}, SO_2, O_3, and the interaction between PM_{10} and O_3 across the board as they appear earlier in the models. The findings reiterate the results from single-pollutant models in Chapter II and the results from two-pollutant models in Chapter IV and can be useful for future research on more comprehensive understanding of the short-term joint effects of multiple pollutants on public

health.

5.6 Application to Brigham and Women's Hospital (BWH) Prospective Cohort Study

5.6.1 Data Overview

We utilize the dataset from an ongoing Brigham and Women's Hospital (BWH) prospective pregnancy/birth cohort study that collects biological samples and detailed clinical data. The target population is a subset of women who initiated their care at the BWH Maternal-Fetal Medicine (MFM) clinic and intend to deliver at BWH. Exclusion criteria are women who had their initial prenatal visit at $>$15 weeks gestation. Our total sample size is 161 women.

Exposure to phthalates, phenols and polycyclic aromatic hydrocarbons (PAHs) has been documented in nearly 100% of the U.S. general population [Crinnion, 2010]. Phthalates, diesters of 1 2-benzenedicarboxylic acid, are a group of chemicals that are widely used as plasticizers or solvents in diverse products in food packaging, cosmetics, and other industrial products. They can enter the human body through daily ingestion and inhalation [Schettler, 2006]. Continuous daily exposure leads to effects similar to those caused by bioaccumulative compounds [Wang et al., 2014]. Phenols is a class of chemical compounds used in the manufacture of polycarbonate plastics and epoxy resins. Applications include use in some food and drink packaging, compact discs, and medical devices [Rezg et al., 2014]. The most commonly seen phenolic compound is bisphenol a (BPA). BPA is well-known to possess estrogenic activity influencing reproductive and its induction of oxidative stress [Rezg et al., 2014] has been demonstrated. PAHs are a class of chemicals that occur naturally in coal, crude oil, and gasoline. Human exposure can result from inhalation but also through ingestion of certain foods such as grilled and smoked meats. Exposure to PAHs has also been linked with cancer, cardiovascular disease and poor fetal development.

Oxidative stress is a condition of imbalance between reactive oxygen species and neutralizing antioxidant capacity within a system. Much of the damage caused by oxidative stress arises from its modification of the DNA inside a cell's nucleus which gives rise to

mutations. Conditions arisen from the damage caused by oxidative stress include neurodegenerative disorders, lung diseases, and heart and blood vessel disorders [Betteridge, 2000]. Oxidative stress may in turn play an important role in the etiology of adverse health outcomes such as preterm birth [Ferguson et al., 2015] and neurodegenerative disorders [Uttara et al., 2009].

Blood and urine were collected at the time of the participants' enrollment visits and analyzed using liquid chromatography-mass spectrometry (LC-MS) method [Li et al., 2008, Onyemauwa et al., 2009]. In total, we consider 9 phthatlate metabolites, 11 phenols, and 8 PAHs. They are tabulated in Table 5.4. Their concentrations in urine have been shown to be variable over time within individuals [Meeker et al., 2012] so we average the biomarker measurements from multiple visits to reduce measurement error. We consider 8-isoprostane as the outcome variable for oxidative stress marker in our analysis. Participant's weight, blood pressure, health status, new diagnoses, BMI, self-identified race, occupation, family history, income, educational attainment, and specific gravity are collected. All these variables are the covariates to be adjusted for in the analysis.

5.6.2 Exploratory Analysis

There are 28 exposure variables that are considered in the analysis with sample size $n = 161$. We first fit a multiple regression model with 8-isoprostane as the dependent variable and only the linear terms of the 28 exposure variables as independent variables. Seven out of 28 variables have a p-value less than 0.05. They are mono-benzyl (MBzP), monoethyl (MEP), Bisphenol A (BPA), benzophenone-3 (BP3), butyl paraben (BuPB), 4-hydroxyphenanthrene (4-PHE), and 1-hydroxypyrene (1-PYR). We present the marginal scatter plots between 8-isoprostane and each of the seven exposure superimposed with a Locally Weighted Scatterplot Smoothing (LOWESS) curve in Figure 5.6. As we can see that, the LOWESS fits suggest that some exposures might display nonlinear relationship in association with 8-isoprostane. The finding reaffirms that a model accounting for nonlinearity is desired.

The mode of action for simultaneous burden from multiple exposures can be synergistic

or antagonistic. We explore 378 possible pairwise linear interactions among 28 exposure variables by regressing 8-isoprostane on a two-way interaction term one at a time, retaining 28 main effect terms in the model throughout. Out of 378 pairs, 35 of them have a p-value less than 0.05. For example, we observe possible interaction between mono(2-ethyl-5-hydroxyhexyl) (MEHHP) and mono(3-carboxypropyl) (MCPP) and possible interaction between 2-hydroxyfluorene (2-FLU) and 2- and 3-hydroxyphenanthrene (2,3-PHE). We summarize the p-values in a heatmap in Figure 5.7. We can observe that the pairs with smaller p-values are somewhat clustered within the same category (i.e. phthatlates, phenols, or PAHs). The phenomenon can partly attribute to the correlation between the exposure measurements within the same category. Since we only account for linear main effects in this exploratory analysis, the plausible interaction can relate to either true interaction effect or higher-order main effects. In the next section, we will resort to methods that account for both nonlinear main effects and nonlinear interaction effects and attempt to clear the ambiguity.

5.6.3 Variable Selection Results

We expand each of the 28 exposure variables into a group of two variables using quadratic splines for nonlinear methods. The effective number of predictors becomes 1568. None of the typical adaptive penalization-based approach can be applied. We therefore consider LASSO-lin, hierNet, LASSO-nonlin, gLASSO, and HiGLASSO in the analysis. We present the effects selected by the five methods in Table 5.5, with the main and interaction effects selected more than once highlighted in bold. As we can see that, hierNet only selects two main effects without any interactions and it is the most conservative method. On the other hand, LASSO-nonlin and gLASSO tend to select more terms, especially interaction effects, than other methods in this case. Based on the findings in our simulation study and from previous studies, some of the selected effects might be false discovery and cannot be validated by competing methods. HiGLASSO is conservative among three nonlinear methods and the strong heredity is maintained as expected. Across the five methods, PAH 1-PYR is the most frequently selected main effect and BP3 by 1-hydroxynapthalene

(1-NAP) is the most frequently selected interaction effect. We note that MBzP, BP3, and 1-NAP are selected by group-based nonlinear methods but not by linear methods. The result suggests potential nonlinearity in the main effects of the three compounds and further investigation is needed. These results can be useful for future research on understanding the effects of mixtures of chemical compounds on oxidative stress.

5.7 Discussion

In this chapter, we developed a HiGLASSO approach to perform variable selection at a group level while maintaining the strong heredity constraint. The approach enforces sparsity in the solution for variable selection at a group level with strong heredity on pairwise interaction, and estimates weights and model parameters in an integrative manner. The proposed approach can handle the situations where number of effective predictors is greater than number of sample size.

In our simulation studies and BWH application, groups of variables are defined through nonlinear expansion to explicitly account for nonlinear main and interaction effects. One major disadvantage of using a general model like HiGLASSO is potential loss of power. One possible remedy is to use a hybrid model where nonlinear main-effect terms are retained but only linear interaction terms are included. The other possible solution is to induce sparsity within groups, as outlined at the end of Section 5.3.2. Inducing sparsity within groups can facilitate with identifying specific nature of association, whether it is linear, quadratic, or in a more complicated functional form. Also, we did not explicitly address the issue of correlation between groups. HiGLASSO with sparsity within groups might partly alleviate the problem. Future theoretical and empirical work is needed to address the issue.

We derived the consistency properties of HiGLASSO when sample size grows at a rate relative to number of predictors p. Future line of theoretical works includes deriving the properties when number of predictor goes to infinity. Computationally, two aspects can be improved. The running time of HiGLASSO greatly depends on how close the initial guess of parameter estimates are to the optimum. OLS estimates are not available when

number of effective predictors is greater than sample size. Stability of using (adaptive) group LASSO estimates as initial guess should be further inspected. Furthermore, the scalability of our algorithm is suboptimal at this point. The major bottleneck is that matrix inversion is inevitable to get exact updates at each iterative step and it is computationally expensive. One potential solution is to make use of an algorithm such as gradient descent to avoid matrix inversion. Sacrificing slight precision at early stage of iterations is acceptable and matrix inversion is only needed a couple of times when the estimates are approaching optimum to resume precision.

Table 5.1: Model specifications in 10 simulation scenarios. "L" indicates linear main effects only, "N" indicates nonlinear main effects only, "LL" indicates linear main and interaction effects, "NN" indicates nonlinear main and interaction effects, "WH" indicates interaction with weak heredity, and "NH" indicates interaction violating heredity. p represents the number of predictors, σ^2 represents the error variance, and true Effects column provides the indices for nonnull main and interaction effects.

Case	p	σ^2	Mean Function	True Effects
L	4	9	$E(\boldsymbol{y}) = \boldsymbol{x}_2$	$\boldsymbol{x}_2$
LL	4	9	$E(\boldsymbol{y}) = \boldsymbol{x}_2 + \boldsymbol{x}_3 + \boldsymbol{x}_2\boldsymbol{x}_3$	$\boldsymbol{x}_2, \boldsymbol{x}_3, \boldsymbol{x}_2\boldsymbol{x}_3$
L	4	36	$E(\boldsymbol{y}) = \boldsymbol{x}_2$	$\boldsymbol{x}_2$
LL	4	36	$E(\boldsymbol{y}) = \boldsymbol{x}_2 + \boldsymbol{x}_3 + \boldsymbol{x}_2\boldsymbol{x}_3$	$\boldsymbol{x}_2, \boldsymbol{x}_3, \boldsymbol{x}_2\boldsymbol{x}_3$
N	4	9	$E(\boldsymbol{y}) = -2.5\boldsymbol{x}_1 + 1.4\boldsymbol{x}_1^2 + 1.3\boldsymbol{x}_1^3 - 1.3\boldsymbol{x}_3 - 1.4\boldsymbol{x}_3^2 + 0.2\boldsymbol{x}_3^3$	$\boldsymbol{x}_1, \boldsymbol{x}_3$
NN	4	9	$E(\boldsymbol{y}) = 2.1\boldsymbol{x}_1 + 2.5\boldsymbol{x}_1^2 - 0.2\boldsymbol{x}_1^3 - 8.2\boldsymbol{x}_2 - 0.7\boldsymbol{x}_2^2 - 0.9\boldsymbol{x}_2^3$ $+8.6\boldsymbol{x}_1\boldsymbol{x}_2 + 1.8\boldsymbol{x}_1\boldsymbol{x}_2^2 - 0.8\boldsymbol{x}_1\boldsymbol{x}_2^3 - 6.5\boldsymbol{x}_1^2\boldsymbol{x}_2 - 1.3\boldsymbol{x}_1^2\boldsymbol{x}_2^2$ $+0.6\boldsymbol{x}_1^2\boldsymbol{x}_2^3 - 1.1\boldsymbol{x}_1^3\boldsymbol{x}_2 - 0.2\boldsymbol{x}_1^3\boldsymbol{x}_2^2 + 0.1\boldsymbol{x}_1^3\boldsymbol{x}_2^3$	$\boldsymbol{x}_1, \boldsymbol{x}_2, \boldsymbol{x}_1\boldsymbol{x}_2$
WH	4	9	$E(\boldsymbol{y}) = \boldsymbol{x}_2 + \boldsymbol{x}_2\boldsymbol{x}_3$	$\boldsymbol{x}_2, \boldsymbol{x}_2\boldsymbol{x}_3$
NH	4	9	$E(\boldsymbol{y}) = \boldsymbol{x}_2 + \boldsymbol{x}_1\boldsymbol{x}_4$	$\boldsymbol{x}_2, \boldsymbol{x}_1\boldsymbol{x}_4$
N	8	9	$E(\boldsymbol{y}) = 2.1\boldsymbol{x}_4 + 2.5\boldsymbol{x}_4^2 - 0.2\boldsymbol{x}_4^3 - 8.2\boldsymbol{x}_6 - 0.7\boldsymbol{x}_6^2 - 0.9\boldsymbol{x}_6^3$	$\boldsymbol{x}_4, \boldsymbol{x}_6$
NN	8	9	$E(\boldsymbol{y}) = 2.1\boldsymbol{x}_4 + 2.5\boldsymbol{x}_4^2 - 0.2\boldsymbol{x}_4^3 - 8.2\boldsymbol{x}_6 - 0.7\boldsymbol{x}_6^2 - 0.9\boldsymbol{x}_6^3$ $+8.6\boldsymbol{x}_4\boldsymbol{x}_6 + 1.8\boldsymbol{x}_4\boldsymbol{x}_6^2 - 0.8\boldsymbol{x}_4\boldsymbol{x}_6^3 - 6.5\boldsymbol{x}_4^2\boldsymbol{x}_6 - 1.3\boldsymbol{x}_4^2\boldsymbol{x}_6^2$ $+0.6\boldsymbol{x}_4^2\boldsymbol{x}_6^3 - 1.1\boldsymbol{x}_4^3\boldsymbol{x}_6 - 0.2\boldsymbol{x}_4^3\boldsymbol{x}_6^2 + 0.1\boldsymbol{x}_4^3\boldsymbol{x}_6^3$	$\boldsymbol{x}_4, \boldsymbol{x}_6, \boldsymbol{x}_4\boldsymbol{x}_6$

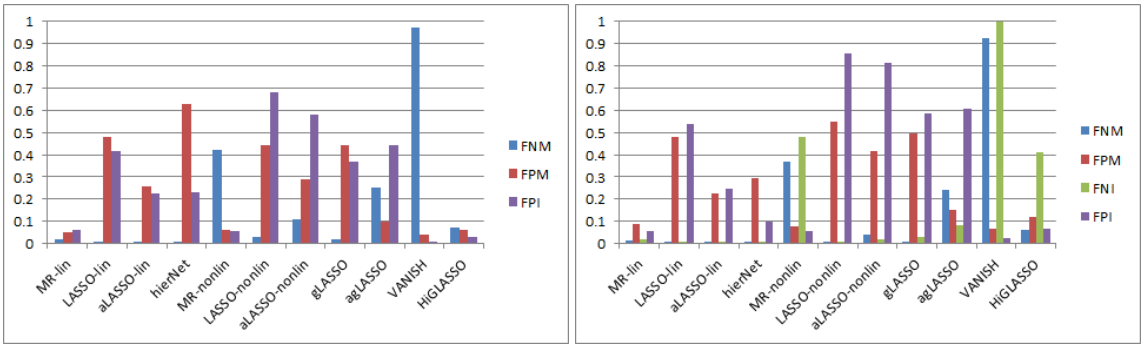

Figure 5.1: False negative main effects rate (FNM), false positive main effects rate (FPM), false negative interaction effects rate (FNI), and false positive interaction effects rate (FPI) across 11 different models in the low-noise scenarios with true linear main effects only (left) and with true linear main and interaction effects (right) based on 100 simulated data sets.

Table 5.2: Number of occurrences of violating strong heredity constraints across 9 methods based on 100 simulated data sets.

Scenario	**LASSO-lin**	**aLASSO-lin**	**hierNet**	**LASSO-nonlin**	**aLASSO-nonlin**	**gLASSO**	**agLASSO**	**VANISH**	**HiGLASSO**
L	17	13	0	32	31	14	43	0	0
LL	24	13	0	39	48	28	51	0	0
N	18	15	0	31	30	21	36	0	0
NN	3	5	0	0	0	0	0	0	0
WH	22	15	0	25	42	21	46	0	0
NH	16	14	0	31	31	25	58	0	0

Table 5.3: Ordered pollutants and pollutant-pollutant interactions (the most important from top) in association to mortality in Chicago, Illinois from 1987 to 2000 based on the National Morbidity, Mortality, and Air Pollution Sutdy (NMMAPS) data. The selected terms by LASSO, group LASSO, and HiGLASSO are in bold.

LASSO	*Group LASSO*	*HiGLASSO*
$PM_{10} \times O_3$	**PM_{10}**	**PM_{10}**
$NO_2 \times O_3$	**SO_2**	**SO_2**
SO_2	**$NO_2 \times O_3$**	**O_3**
PM_{10}	**O_3**	**CO**
NO2$\times$ SO_2	**CO**	$PM_{10} \times SO_2$
O_3	**CO $\times$ SO_2**	$PM_{10} \times O_3$
$PM_{10} \times SO_2$	**$PM_{10} \times O_3$**	CO $\times$ SO_2
CO $\times$ SO_2	**$SO_2 \times O_3$**	$SO_2 \times O_3$
CO$\times$ O_3	**NO2$\times$ SO_2**	CO$\times$ O_3
CO	**CO$\times$ O_3**	$PM_{10} \times$ CO
$PM_{10} \times$ CO	$PM_{10} \times$ CO	NO_2
$SO_2 \times O_3$	$PM_{10} \times SO_2$	$PM_{10} \times NO_2$
CO$\times$ NO_2	$PM_{10} \times NO_2$	$NO_2 \times O_3$
$PM_{10} \times NO_2$	CO$\times$ NO_2	NO2$\times$ SO_2
NO_2	NO_2	CO$\times$ NO_2

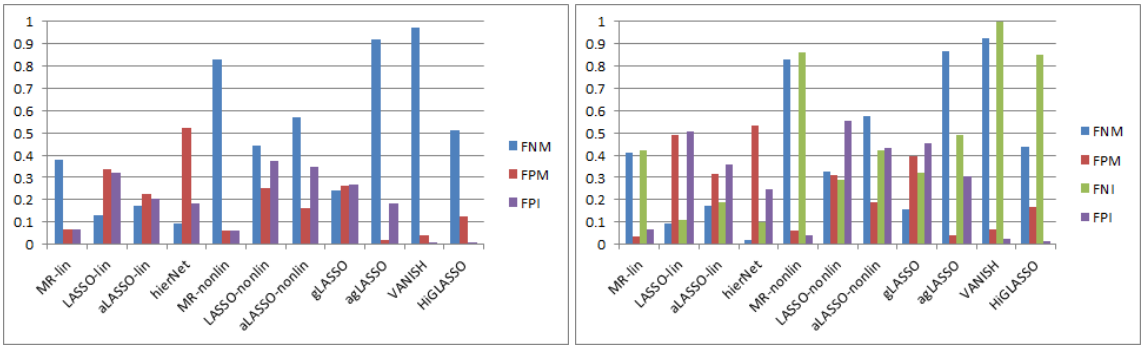

Figure 5.2: False negative main effects rate (FNM), false positive main effects rate (FPM), false negative interaction effects rate (FNI), and false positive interaction effects rate (FPI) across 11 different models in the high-noise scenarios with true linear main effects only (left) and with true linear main and interaction effects (right) based on 100 simulated data sets.

Table 5.4: List of 28 exposure measurements including 9 phthatlates, 11 phenols, and 8 PAHs used in Brigham and Women's Hospital (BWH) analysis.

Exposure	Full Name (Acronym)
Phthatlates	mono(2-ethylhexyl) (MEHP) mono(2-ethyl-5-hydroxyhexyl) (MEHHP) mono(2-ethyl-5-oxohexyl) (MEOHP) mono(2-ethyl-5-carboxypentyl) (MECPP) monobenzyl (MBzP) mono-n-butyl (MBP) monoisobutyl (MiBP) monoethyl (MEP) mono(3-carboxypropyl) (MCPP)
Phenols	Bisphenol A (BPA) Bisphenol S (BPS) 2,4-Dichlorophenol (2,4-DCP) 2,5-Dichlorophenol (2,5-DCP) benzophenone-3 (BP3) butyl paraben (BuPB) ethyl paraben (EtPB) methyl paraben (MePB) propyl paraben (PrPB) triclocarban (TCB) triclosan (TCS)
PAHs	2-hydroxynapthalene (2-NAP) 1-hydroxynapthalene (1-NAP) 2-hydroxyfluorene (2-FLU) 2- and 3-hydroxyphenanthrene (2,3-PHE) 9-hydroxyphenanthrene (9-PHE) 1-hydroxyphenanthrene (1-PHE) 4-hydroxyphenanthrene (4-PHE) 1-hydroxypyrene (1-PYR)

Table 5.5: Brigham and Women's Hospital (BWH) prospective cohort study data set: selected main effects and interaction effects.

Method	Selected Variables
LASSO-lin	**2-FLU**, **1-PYR**, MCPP×BuPB, MCPP×TCB, BP3×BuPB, BP3×TCB, **BP3**×**1-NAP**, BuPB×TCB
hierNet	**2-FLU**, **1-PYR**
LASSO-nonlin	**2-FLU**, **1-PYR**, MBP×2-FLU, **MiBP**×**MEP**, **MiBP**×**2-FLU**, **MCPP**×**MePB**, MCPP×TCB, BP3×2-NAP, **BP3**×**1-NAP**, 2-NAP×1-PYR, 1-NAP×9-PHE, 1-NAP×1-PHE, 2-FLU×2- and 3-PHE, 2-FLU×1-PYR
gLASSO	**MBzP**, **BP3**, BuPB, 2-NAP, **1-NAP**, **2-FLU**, **1-PYR**, MECPP×MEP, MECPP×1-NAP, **MBZP**×**1-PYR**, MBP×2,5-DCP, MBP×1-PYR, **MiBP**×**MEP**, MiBP×2,4-DCP MiBP×2,5-DCP, **MiBP**×**2-FLU**, MEP×BP3, **MCPP**×**MePB** MCPP×PrPB, MCPP×2-FLU, MCPP×2- and 3-PHE, BPA×BPS BPA×MePB, BPS×BP3, BPS×TCB, 2,4-DCP×1-PHE **BP3**×**1-NAP**, BP3×2-FLU, TCS×2-NAP, TCS×2-FLU 2-FLU×1-PYR, 9-PHE×1-PHE
HiGLASSO	**MBzP**, **BP3**, **1-NAP**, **1-PYR**, **MBzP**×**1-PYR**, **BP3**×**1-NAP**

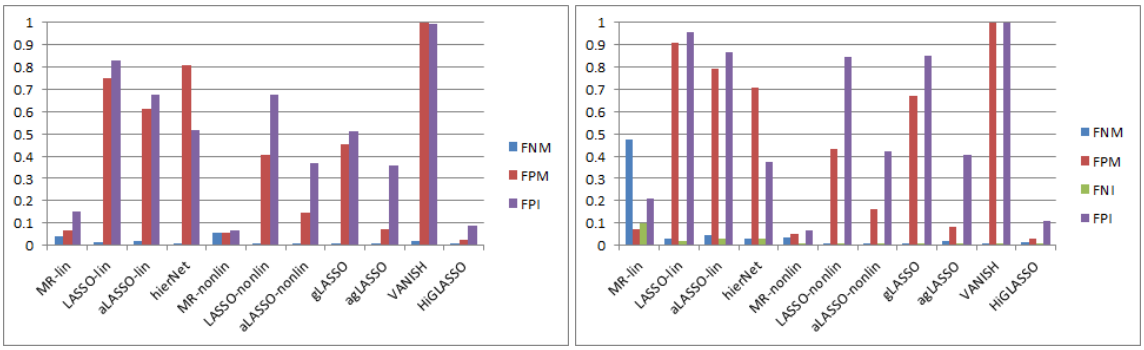

Figure 5.3: False negative main effects rate (FNM), false positive main effects rate (FPM), false negative interaction effects rate (FNI), and false positive interaction effects rate (FPI) across 11 different models in the scenarios with true nonlinear main effects (left) and with true nonlinear main and interaction effects (right) based on 100 simulated data sets.

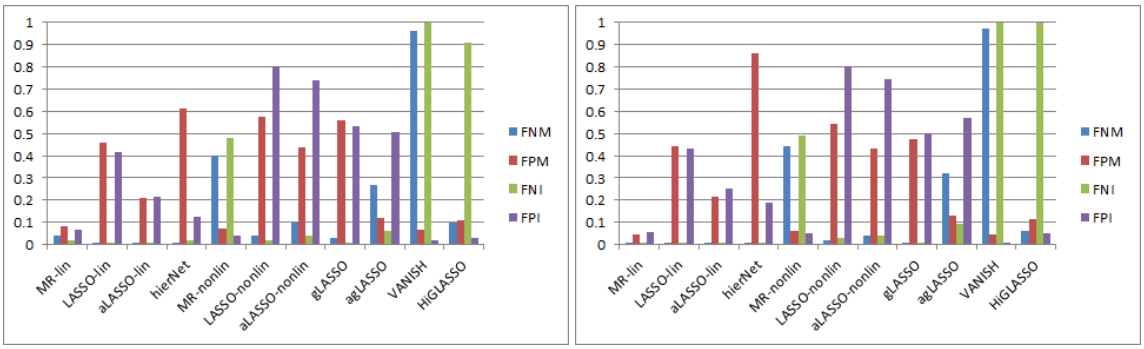

Figure 5.4: False negative main effects rate (FNM), false positive main effects rate (FPM), false negative interaction effects rate (FNI), and false positive interaction effects rate (FPI) across 11 different models in the scenarios with interaction under weak heredity (left) and interaction violating heredity constraint (right) based on 100 simulated data sets.

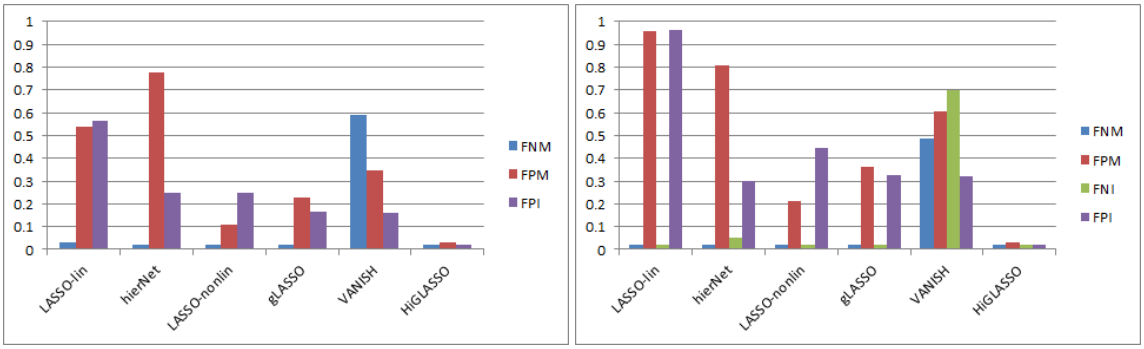

Figure 5.5: False negative main effects rate (FNM), false positive main effects rate (FPM), false negative interaction effects rate (FNI), and false positive interaction effects rate (FPI) across 6 different models in the scenarios with true nonlinear main effects (left) and with true nonlinear main and interaction effects (right) based on 100 simulated data sets. Number of effective predictors is greater than number of sample size.

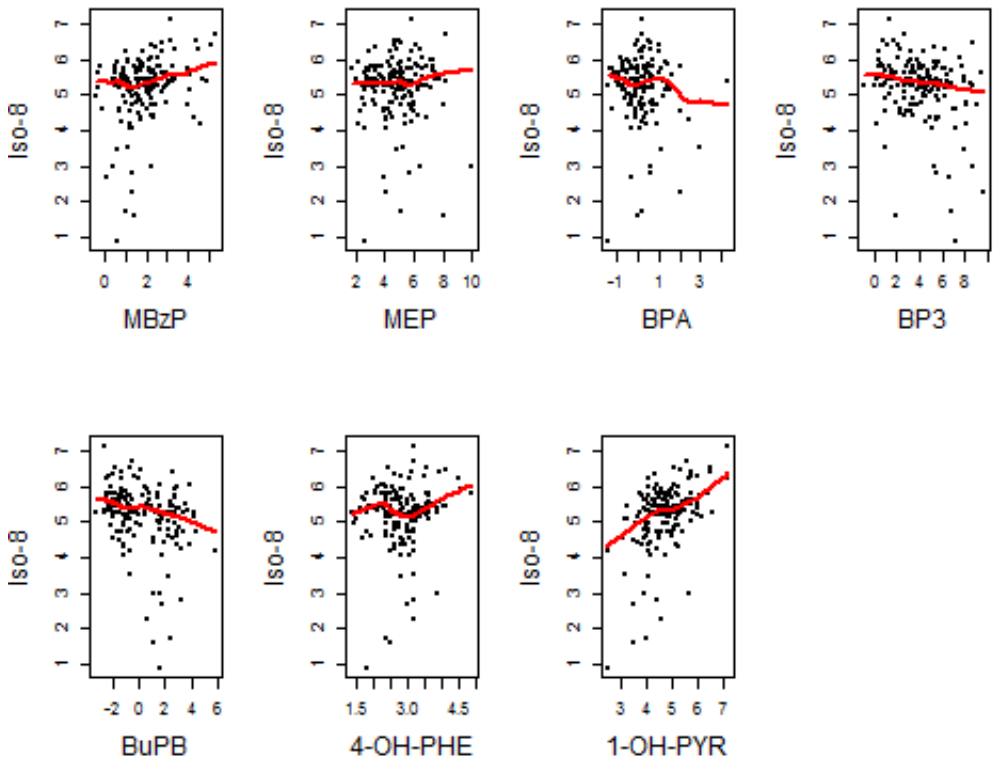

Figure 5.6: Scatter plots between seven exposures and 8-isoprostane superimposed with a Locally Weighted Scatterplot Smoothing (LOWESS curve). The seven exposures are mono-benzyl (MBzP), monoethyl (MEP), Bisphenol A (BPA), benzophenone-3 (BP3), butyl paraben (BuPB), 4-hydroxyphenanthrene (4-PHE), and 1-hydroxypyrene (1-PYR).

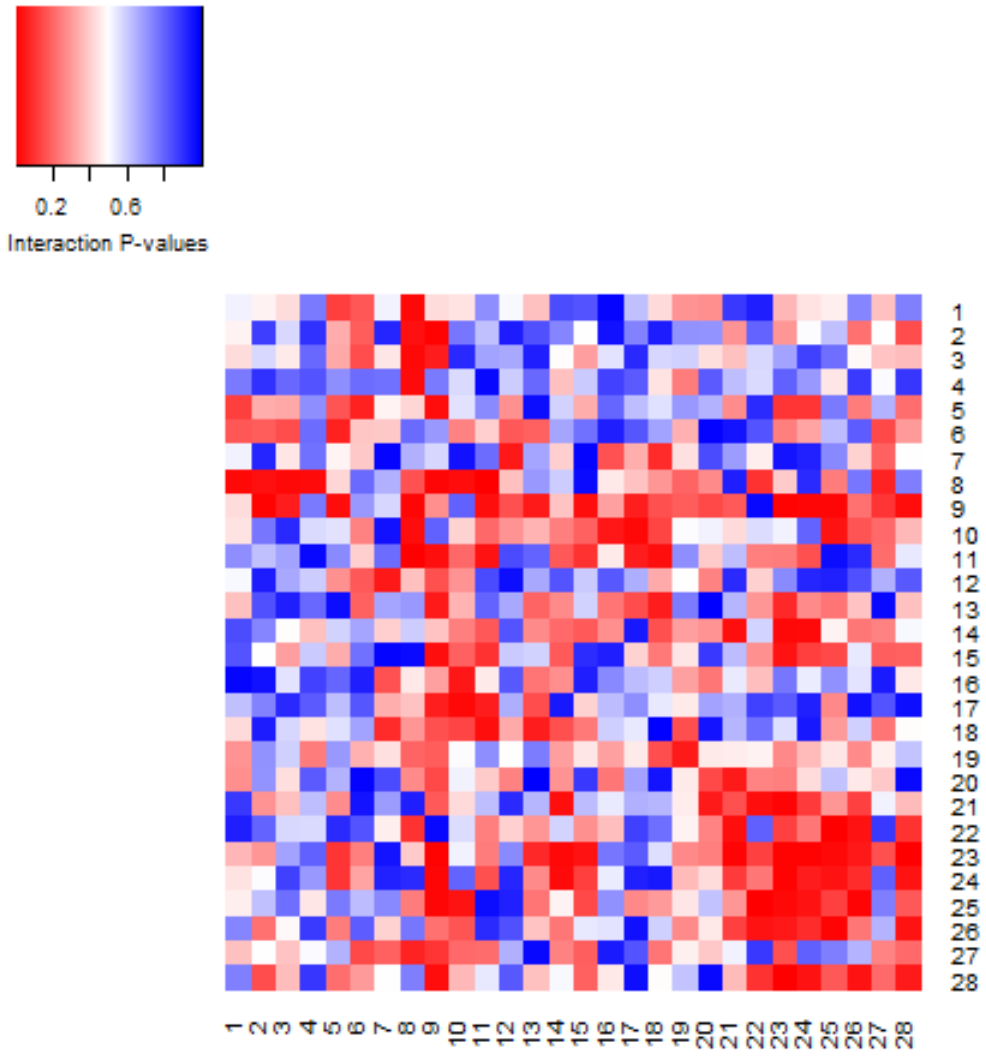

Figure 5.7: Heatmap for pairwise interaction p-values between 28 exposure variables. Each p-value is obtained from a multiple regression model with 28 exposure main-effect terms and a single interaction term. 1-9 are phthalates, 10-20 are phenols, and 21-28 are PAHs.

CHAPTER 6

Conclusion

DLM has been extensively employed to study lagged effects of environmental exposures on health outcomes in environmental epidemiology. Most of the existing DLMs are limited to modeling one pollutant at a time. Existing multi-pollutants methods do not consider the temporal dynamics of lags. In this dissertation, our main goal is to extend one-dimensional DLMs to two or more pollutants. To achieve this goal, we first proposed methods to make single pollutant DLMs robust and then consider two or more pollutants in a related framework. The methods proposed in this dissertation all broadly relate to the concept of shrinkage and selection.

Most existing shrinkage methods shrink an estimator toward the null (i.e. zero) to avoid overfitting. In contrast, the approaches we introduced in Chapter II and Chapter IV shrink an unconstrained estimator toward a meaningful nonnull shrinkage target. The target can be tailored according to the prior knowledge in subject-matter domain. The flexibility to data adaptively shrink between multiple nonnull targets protects against model misspecifications without losing all of the efficiency advantage of a parametric model. The simulation results indicate that these methods are robust to choice of nonnull target and optimal bias-variance tradeoff can be achieved across different simulation scenarios. The powerful VCST for testing a particular DLM structure against a general alternative developed in Chapter III can serve as a screening procedure for choosing a proper parametric distributed lag structure as a nonnull target. Instead of relying on agnostic robust DLMs, a VCST can be conducted as a first-step screening method before a specific parametric DLM is fitted. One can then use this parametric DLM as a target for shrinkage.

In Chapter IV, we considered different strategies to model pollutant-by-pollutant inter-

action in two-dimensional DLMs. Assuming the basis functions underlying the interaction DL surface as the tensor products of the basis functions underlying the main-effect DL functions is a natural step. The marginal DL functions can still be expressed as the linear combination of the same sets of basis functions, making interpretation more convenient. Tukey's single parameter form of interaction is a well-known way of modeling the interaction surface and it is known to be powerful for hypothesis testing. We extend Tukey's form of interaction to DLMs: a major innovation in this dissertation. We also provided shrinkage versions of the two-dimensional DLMs to protect against misclassification. A novel way of interpreting changes in marginal DL function of one pollutant when the other pollutant varies across different values was introduced. In Chapter V, we proposed a variable selection framework HiGLASSO that is a promising approach capable of incorporating weights into variable selection at a group level with consideration of two-way interaction, while maintaining the strong heredity constraints. Empirically, it outperforms other alternatives in selection consistency and other metrics for variable selection.

The work presented in this dissertation indicates many areas of potential future research. The robust DLMs aim at minimizing MSE by introducing small bias in exchange for reduction in variance. Although they are capable of reducing MSE asymptotically, the asymptotic bias is never zero. One possible extension is to perform correction based on the expression of asymptotic bias to de-bias robust DLMs. While VCST provides a global test for DLM, Bayesian variable selection methods can help with identification of nonnull specific lag coefficients. Current version of HiGLASSO is not computationally efficient in ultra-high dimensional variable selection. One possible way to streamline the algorithm is to construct the solution path by leveraging information from solutions with adjacent tuning parameters. The solutions with adjacent tuning parameters typically differ by no more than one group of variables and a single sweep through all unselected variables may be sufficient to ensure whether an unselected predictor is to enter the model or a selected predictor is to exit the model. HiGLASSO only induces inter-group sparsity. In other words, all the variables corresponding to a main or interaction effect are set to be zero or nonzero. Inducing sparsity within groups can potentially assist in identifying which variable(s) within each group contributes to the association. Extending HiGLASSO to a situation allowing

within-group sparsity can be another direction of future research.

Although we illustrated the proposed methods using examples in environmental epidemiology, these methods can potentially be adapted to a wide spectrum of problems. Shrinkage methods that are adaptive and data-driven and certainly play an important role in machine learning and data mining. The dissertation enables us to characterize lag effects of one pollutant when the other is set at a given value and thus can help with guiding multi-pollutant policy for air quality standard. We hope that this dissertation contributes to statistical methods for chemical mixtures and has broader relevance in the statistical literature on selection and shrinkage for interaction models.

BIBLIOGRAPHY

Michelle L Bell, Jonathan M Samet, and Francesca Dominici. Time-series studies of particulate matter. *Annu. Rev. Public Health*, 25:247–280, 2004a.

Majid Ezzati, Alan D Lopez, Anthony Rodgers, Stephen Vander Hoorn, Christopher JL Murray, et al. Selected major risk factors and global and regional burden of disease. *The Lancet*, 360(9343):1347–1360, 2002.

A Le Tertre, S Medina, E Samoli, B Forsberg, P Michelozzi, A Boumghar, JM Vonk, A Bellini, R Atkinson, JG Ayres, et al. Short-term effects of particulate air pollution on cardiovascular diseases in eight european cities. *Journal of epidemiology and community health*, 56(10):773–779, 2002.

Claudia Spix, H Ross Anderson, Joel Schwartz, Maria Angela Vigotti, Alain Letertre, Judith M Vonk, Giota Touloumi, Franck Balducci, Tomasz Piekarski, Ljuba Bacharova, et al. Short-term effects of air pollution on hospital admissions of respiratory diseases in europe: a quantitative summary of aphea study results. *Archives of Environmental Health: An International Journal*, 53(1):54–64, 1998.

Klea Katsouyanni, D Zmirou, C Spix, J Sunyer, JP Schouten, A Ponka, HR Anderson, Y Le Moullec, B Wojtyniak, MA Vigotti, et al. Short-term effects of air pollution on health: a european approach using epidemiological time-series data. the aphea project: background, objectives, design. *European Respiratory Journal*, 8(6):1030–1038, 1995.

G Touloumi, SJ Pocock, K Katsouyanni, and D Trichopoulos. Short-term effects of air pollution on daily mortality in athens: a time-series analysis. *International journal of epidemiology*, 23(5):957–967, 1994.

DAVID Berglund and D Abbey. Long-term effects of air pollution. *Western journal of medicine*, 165(3):140, 1996.

Bert Brunekreef and Stephen T Holgate. Air pollution and health. *The lancet*, 360(9341): 1233–1242, 2002.

Kristin A Miller, David S Siscovick, Lianne Sheppard, Kristen Shepherd, Jeffrey H Sullivan, Garnet L Anderson, and Joel D Kaufman. Long-term exposure to air pollution and incidence of cardiovascular events in women. *New England Journal of Medicine*, 356 (5):447–458, 2007.

Jonathan M Samet, Scott L Zeger, Francesca Dominici, Frank Curriero, Ivan Coursac, Douglas W Dockery, Joel Schwartz, and Antonella Zanobetti. The national morbidity, mortality, and air pollution study. *Part II: morbidity and mortality from air pollution in the United States Res Rep Health Eff Inst*, 94(pt 2):5–79, 2000.

Joel Schwartz. The distributed lag between air pollution and daily deaths. *Epidemiology*, 11(3):320–326, 2000.

Alfésio Luís Ferreira Braga, Antonella Zanobetti, and Joel Schwartz. The lag structure between particulate air pollution and respiratory and cardiovascular deaths in 10 us cities. *Journal of Occupational and Environmental Medicine*, 43(11):927–933, 2001.

Antonella Zanobetti, Joel Schwartz, Evi Samoli, Alexandros Gryparis, Giota Touloumi, Janet Peacock, Ross H Anderson, Alain Le Tertre, Janos Bobros, Martin Celko, et al. The temporal pattern of respiratory and heart disease mortality in response to air pollution. *Environmental health perspectives*, 111(9):1188, 2003.

Shirley Almon. The distributed lag between capital appropriations and expenditures. *Econometrica: Journal of the Econometric Society*, pages 178–196, 1965.

Trevor Hastie and Robert Tibshirani. Varying-coefficient models. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 757–796, 1993.

A Zanobetti, MP Wand, J Schwartz, and LM Ryan. Generalized additive distributed lag models: quantifying mortality displacement. *Biostatistics*, 1(3):279–292, 2000.

Trevor J Hastie and Robert J Tibshirani. *Generalized additive models*, volume 43. CRC Press, 1990.

A Gasparrini, B Armstrong, and MG Kenward. Distributed lag non-linear models. *Statistics in medicine*, 29(21):2224, 2010.

Leah J Welty, RD Peng, SL Zeger, and F Dominici. Bayesian distributed lag models: estimating effects of particulate matter air pollution on daily mortality. *Biometrics*, 65 (1):282–291, 2009.

Francesca Dominici, Roger D Peng, Christopher D Barr, and Michelle L Bell. Protecting human health from air pollution: shifting from a single-pollutant to a multi-pollutant approach. *Epidemiology (Cambridge, Mass.)*, 21(2):187, 2010.

Michelle L Bell, Jee Young Kim, and Francesca Dominici. Potential confounding of particulate matter on the short-term association between ozone and mortality in multisite time-series studies. *Environmental health perspectives*, pages 1591–1595, 2007.

Rosalba Rojas-Martinez, Rogelio Perez-Padilla, Gustavo Olaiz-Fernandez, Laura Mendoza-Alvarado, Hortensia Moreno-Macias, Teresa Fortoul, William McDonnell, Dana Loomis, and Isabelle Romieu. Lung function growth in children with long-term exposure to air pollutants in mexico city. *American journal of respiratory and critical care medicine*, 176(4):377–384, 2007.

Joe L Mauderly, Richard T Burnett, Margarita Castillejos, Halûk Özkaynak, Jonathan M Samet, David M Stieb, Sverre Vedal, and Ronald E Wyzga. Is the air pollution health research community prepared to support a multipollutant air quality management framework? *Inhalation toxicology*, 22(sup1):1–19, 2010.

Donald E Farrar and Robert R Glauber. Multicollinearity in regression analysis: the problem revisited. *The Review of Economic and Statistics*, pages 92–107, 1967.

Wenbiao Hu, Kerrie Mengersen, Anthony McMichael, and Shilu Tong. Temperature, air pollution and total mortality during summers in sydney, 1994–2004. *International journal of biometeorology*, 52(7):689–696, 2008.

Sandra E Sinisi and Mark J van der Laan. Deletion/substitution/addition algorithm in learning with applications in genomics. *Statistical applications in genetics and molecular biology*, 3(1):1–38, 2004.

Francesca Dominici, Chi Wang, Ciprian Crainiceanu, and Giovanni Parmigiani. Model selection and health effect estimation in environmental epidemiology. *Epidemiology*, 19 (4):558–560, 2008.

Robert Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 267–288, 1996.

RT Burnett, J Brook, T Dann, C Delocla, O Philips, S Cakmak, R Vincent, MS Goldberg, and D Krewski. Association between particulate-and gas-phase components of urban air pollution and daily mortality in eight canadian cities. 2001.

Zhengmin Qian, Junfeng Jim Zhang, Leo R Korn, Fusheng Wei, and Robert S Chapman. Factor analysis of household factors: are they associated with respiratory conditions in chinese children? *International journal of epidemiology*, 33(3):582–588, 2004.

Ahmed A Arif and Syed M Shah. Association between personal exposure to volatile organic compounds and asthma among us adult population. *International archives of occupational and environmental health*, 80(8):711–719, 2007.

Steven Roberts. Interactions between particulate air pollution and temperature in air pollution mortality time series studies. *Environmental research*, 96(3):328–337, 2004.

Matthew J Heaton and Roger D Peng. Extending distributed lag models to higher degrees. *Biostatistics*, 15(2):398–412, 2014.

Vito MR Muggeo. Bivariate distributed lag models for the analysis of temperature-by-pollutant interaction effect on mortality. *Environmetrics*, 18(3):231–243, 2007.

John W Tukey. One degree of freedom for non-additivity. *Biometrics*, 5(3):232–242, 1949.

Yi-An Ko, Bhramar Mukherjee, Jennifer A Smith, Sung Kyun Park, Sharon LR Kardia, Matthew A Allison, Pantel S Vokonas, Jinbo Chen, and Ana V Diez-Roux. Testing departure from additivity in tukey's model using shrinkage: application to a longitudinal setting. *Statistics in medicine*, 33(29):5177–5191, 2014.

Joel Schwartz and Douglas W Dockery. Increased mortality in philadelphia associated with daily air pollution concentrations. *American review of respiratory disease*, 145(3): 600–604, 1992.

Joel Schwartz. Air pollution and daily mortality: a review and meta analysis. *Environmental research*, 64(1):36–52, 1994.

Steven Roberts. An investigation of distributed lag models in the context of air pollution and mortality time series analysis. *Journal of the Air & Waste Management Association*, 55(3):273–282, 2005.

Corrado Corradi. Smooth distributed lag estimators and smoothing spline functions in hilbert spaces. *Journal of Econometrics*, 5(2):211–219, 1977.

Antonella Zanobetti, Joel Schwartz, Evi Samoli, Alexandros Gryparis, Giota Touloumi, Richard Atkinson, Alain Le Tertre, Janos Bobros, Martin Celko, Ayana Goren, et al. The temporal pattern of mortality responses to air pollution: a multicity assessment of mortality displacement. *Epidemiology*, 13(1):87–93, 2002.

Vito MR Muggeo. Modeling temperature effects on mortality: multiple segmented relationships with common break points. *Biostatistics*, 9(4):613–620, 2008.

Roger D Peng, Francesca Dominici, and Leah J Welty. A bayesian hierarchical distributed lag model for estimating the time course of risk of hospitalization associated with particulate matter air pollution. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 58(1):3–24, 2009.

Viola Obermeier, Fabian Scheipl, Christian Heumann, Joachim Wassermann, and Helmut Küchenhoff. Flexible distributed lags for modelling earthquake data. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 64(2):395–412, 2015.

Bhramar Mukherjee and Nilanjan Chatterjee. Exploiting gene-environment independence for analysis of case–control studies: An empirical bayes-type shrinkage estimator to trade-off between bias and efficiency. *Biometrics*, 64(3):685–694, 2008.

Yi-Hau Chen, Nilanjan Chatterjee, and Raymond J Carroll. Shrinkage estimators for robust and efficient inference in haplotype-based case-control studies. *Journal of the American Statistical Association*, 104(485):220–233, 2009.

Clifford M Hurvich, Jeffrey S Simonoff, and Chih-Ling Tsai. Smoothing parameter selection in nonparametric regression using an improved akaike information criterion. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 60(2):271–293, 1998.

Gene H Golub, Michael Heath, and Grace Wahba. Generalized cross-validation as a method for choosing a good ridge parameter. *Technometrics*, 21(2):215–223, 1979.

Herbert Schimmel and Thaddeus J Murawski. The relation of air pollution to mortality. *Journal of Occupational and Environmental Medicine*, 18(5):316–333, 1976.

Bradley P Carlin and Thomas A Louis. Bayes and empirical bayes methods for data analysis. *Statistics and Computing*, 7(2):153–154, 1997.

Robert J Shiller. A distributed lag estimator derived from smoothness priors. *Econometrica: journal of the Econometric Society*, pages 775–788, 1973.

Gerda Claeskens and Raymond J Carroll. An asymptotic theory for model selection inference in general semiparametric problems. *Biometrika*, 94(2):249–265, 2007.

Philip E Gill, Walter Murray, and Margaret H Wright. Practical optimization. 1981.

Andrew Gelman, Aleks Jakulin, Maria Grazia Pittau, and Yu-Sung Su. A weakly informative default prior distribution for logistic and other regression models. *The Annals of Applied Statistics*, pages 1360–1383, 2008.

Giampiero Marra and Simon N Wood. Coverage properties of confidence intervals for generalized additive model components. *Scandinavian Journal of Statistics*, 39(1):53–74, 2012.

Francesca Dominici, Aidan McDermott, Michael Daniels, Scott L Zeger, and Jonathan M Samet. Revised analyses of the national morbidity, mortality, and air pollution study: mortality among residents of 90 cities. *Journal of Toxicology and Environmental Health, Part A*, 68(13-14):1071–1092, 2005.

Antonella Zanobetti and Joel Schwartz. Mortality displacement in the association of ozone with mortality: an analysis of 48 cities in the united states. *American Journal of Respiratory and Critical Care Medicine*, 177(2):184–189, 2008.

M Majid and Rosylin Mohd Yusof. Long-run relationship between islamic stock returns and macroeconomic variables: An application of the autoregressive distributed lag model. *Humanomics*, 25(2):127–141, 2009.

Tzu-Kuang Hsu. Exploring relationship between the stock price of taiwan and the exchange rate: An autoregressive distributed lag model with a quantile regression. *International Journal of Economics and Finance*, 8(1):72, 2015.

Yin-Hsiu Chen, Bhramar Mukherjee, Sara D Adar, Veronica J Berrocal, and Brent A Coull. Robust distributed lag models using data adaptive shrinkage. *Biostatistics (Accepted)*, 2017.

Xihong Lin. Variance component testing in generalised linear models with random effects. *Biometrika*, 84(2):309–326, 1997.

Daowen Zhang and Xihong Lin. Hypothesis testing in semiparametric additive mixed models. *Biostatistics*, 4(1):57–74, 2003.

Robert B Davies. Algorithm as 155: The distribution of a linear combination of χ 2 random variables. *Applied Statistics*, pages 323–333, 1980a.

Harald Bohman. A method to calculate the distribution function when the characteristic function is known. *BIT Numerical Mathematics*, 10(3):237–242, 1970.

Robert B Davies. Algorithm as 155: The distribution of a linear combination of chi-square random variables. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, 29(3):323–333, 1980b.

C Arden Pope and Douglas W Dockery. Health effects of fine particulate air pollution: lines that connect. *Journal of the air & waste management association*, 56(6):709–742, 2006.

Robert D Brook, Sanjay Rajagopalan, C Arden Pope, Jeffrey R Brook, Aruni Bhatnagar, Ana V Diez-Roux, Fernando Holguin, Yuling Hong, Russell V Luepker, Murray A Mittleman, et al. Particulate matter air pollution and cardiovascular disease an update to the scientific statement from the american heart association. *Circulation*, 121(21):2331–2378, 2010.

C Arden Pope. Mortality effects of longer term exposures to fine particulate air pollution: review of recent epidemiological evidence. *Inhalation Toxicology*, 19(sup1):33–38, 2007.

C Arden Pope, Douglas W Dockery, and Joel Schwartz. Review of epidemiological evidence of health effects of particulate air pollution. *Inhalation toxicology*, 7(1):1–18, 1995.

Klea Katsouyanni, Giotta Touloumi, Claudia Spix, Joel Schwartz, Franck Balducci, Silvya Medina, G Rossi, Bogdan Wojtyniak, Jordi Sunyer, Ljuba Bacharova, et al. Short term effects of ambient sulphur dioxide and particulate matter on mortality in 12 european cities: results from time series data from the aphea project. *Bmj*, 314(7095):1658, 1997.

Michelle L Bell, Aidan McDermott, Scott L Zeger, Jonathan M Samet, and Francesca Dominici. Ozone and short-term mortality in 95 us urban communities, 1987-2000. *Jama*, 292(19):2372–2378, 2004b.

Francesca Dominici, Roger D Peng, Michelle L Bell, Luu Pham, Aidan McDermott, Scott L Zeger, and Jonathan M Samet. Fine particulate air pollution and hospital admission for cardiovascular and respiratory diseases. *Jama*, 295(10):1127–1134, 2006.

Ariana Zeka and Joel Schwartz. Estimating the independent effects of multiple pollutants in the presence of measurement error: an application of a measurement-error-resistant technique. *Environmental health perspectives*, pages 1686–1690, 2004.

Cécile Billionnet, Duane Sherrill, Isabella Annesi-Maesano, et al. Estimating the health effects of exposure to multi-pollutant mixture. *Annals of epidemiology*, 22(2):126–141, 2012.

BA Coull, JF Bobb, GA Wellenius, MA Kioumourtzoglou, MA Mittleman, P Koutrakis, and JJ Godleski. Part 1. statistical learning methods for the effects of multiple air pollution constituents. *Research report (Health Effects Institute)*, (183 Pt 1-2):5–50, 2015.

Joe L Mauderly. Toxicological approaches to complex mixtures. *Environmental health perspectives*, 101(Suppl 4):155, 1993.

Sander Greenland. Basic problems in interaction assessment. *Environmental health perspectives*, 101(Suppl 4):59, 1993.

Zhichao Sun, Yebin Tao, Shi Li, Kelly K Ferguson, John D Meeker, Sung Kyun Park, Stuart A Batterman, and Bhramar Mukherjee. Statistical strategies for constructing health risk models with multiple pollutants and their interactions: possible choices and comparisons. *Environ Health*, 12(1):85, 2013.

Hui Zou and Trevor Hastie. Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 67(2):301–320, 2005.

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

Jennifer F Bobb, Linda Valeri, Birgit Claus Henn, David C Christiani, Robert O Wright, Maitreyi Mazumdar, John J Godleski, and Brent A Coull. Bayesian kernel machine regression for estimating the health effects of multi-pollutant mixtures. *Biostatistics*, page kxu058, 2014.

Brian D Marx and Paul HC Eilers. Direct generalized additive modeling with penalized likelihood. *Computational Statistics & Data Analysis*, 28(2):193–209, 1998.

Nilanjan Chatterjee, Zeynep Kalaylioglu, Roxana Moslehi, Ulrike Peters, and Sholom Wacholder. Powerful multilocus tests of genetic association in the presence of gene-gene and gene-environment interactions. *The American Journal of Human Genetics*, 79(6): 1002–1016, 2006.

Arnab Maity, Raymond J Carroll, Enno Mammen, and Nilanjan Chatterjee. Testing in semiparametric models with interaction, with applications to gene–environment interactions. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 71(1): 75–96, 2009.

Yaping Wang, Donghui Li, and Peng Wei. Powerful tukeys one degree-of-freedom test for detecting gene–gene and gene–environment interactions. *Cancer informatics*, 14(Suppl 2):209, 2015.

John C Beatty and Brian A Barsky. *An introduction to splines for use in computer graphics and geometric modeling*. Morgan Kaufmann, 1987.

Paul Dierckx. *Curve and surface fitting with splines*. Oxford University Press, 1995.

Carl De Boor, Carl De Boor, Carl De Boor, and Carl De Boor. *A practical guide to splines*, volume 27. Springer-Verlag New York, 1978.

Sudipto Banerjee, Alan E Gelfand, Andrew O Finley, and Huiyan Sang. Gaussian predictive process models for large spatial data sets. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 70(4):825–848, 2008.

Tilmann Gneiting, Hana Ševčíková, and Donald B Percival. Estimators of fractal dimension: Assessing the roughness of time series and spatial data. *Statistical Science*, pages 247–277, 2012.

Hao Zhang. Inconsistent estimation and asymptotically equal interpolations in model-based geostatistics. *Journal of the American Statistical Association*, 99(465):250–261, 2004.

Juan Du, Hao Zhang, VS Mandrekar, et al. Fixed-domain asymptotic properties of tapered maximum likelihood estimators. *the Annals of Statistics*, 37(6A):3330–3361, 2009.

Hao Zhang and Yong Wang. Kriging and cross-validation for massive spatial data. *Environmetrics*, 21(3-4):290–304, 2010.

Abhirup Datta, Sudipto Banerjee, Andrew O Finley, and Alan E Gelfand. Hierarchical nearest-neighbor gaussian process models for large geostatistical datasets. *Journal of the American Statistical Association*, (just-accepted):00–00, 2015.

Dani Gamerman and Hedibert F Lopes. *Markov chain Monte Carlo: stochastic simulation for Bayesian inference*. CRC Press, 2006.

Douglas M Bates and Donald G Watts. Nonlinear regression: iterative estimation and linear approximations. *Nonlinear Regression Analysis and Its Applications*, pages 32–66, 1988.

Bradley Efron. Nonparametric estimates of standard error: the jackknife, the bootstrap and other methods. *Biometrika*, 68(3):589–599, 1981.

Andrew Gelman et al. Prior distributions for variance parameters in hierarchical models (comment on article by browne and draper). *Bayesian analysis*, 1(3):515–534, 2006.

W Keith Hastings. Monte carlo sampling methods using markov chains and their applications. *Biometrika*, 57(1):97–109, 1970.

Stuart Geman and Donald Geman. Stochastic relaxation, gibbs distributions, and the bayesian restoration of images. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, (6):721–741, 1984.

George EP Box and George C Tiao. *Bayesian inference in statistical analysis*, volume 40. John Wiley & Sons, 2011.

David Lunn, David Spiegelhalter, Andrew Thomas, and Nicky Best. The bugs project: Evolution, critique and future directions. *Statistics in medicine*, 28(25):3049–3067, 2009.

Patrick G Goodman, Douglas W Dockery, and Luke Clancy. Cause-specific mortality and the extended effects of particulate pollution and temperature exposure. *Environmental Health Perspectives*, 112(2):179, 2004.

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

Leah J Welty and Scott L Zeger. Are the acute effects of particulate matter on mortality in the national morbidity, mortality, and air pollution study the result of inadequate control for weather and season? a sensitivity analysis using flexible distributed lag models. *American Journal of Epidemiology*, 162(1):80–88, 2005.

David J Spiegelhalter, Nicola G Best, Bradley P Carlin, and Angelika Van Der Linde. Bayesian measures of model complexity and fit. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 64(4):583–639, 2002.

Martyn Plummer. Penalized loss functions for bayesian model comparison. *Biostatistics*, 9(3):523–539, 2008.

Joe L Mauderly and Jonathan M Samet. Is there evidence for synergy among air pollutants in causing health effects? *Environmental health perspectives*, 117(1):1, 2009.

Jacob Bien, Jonathan Taylor, and Robert Tibshirani. A lasso for hierarchical interactions. *Annals of statistics*, 41(3):1111, 2013.

Hansheng Wang and Chenlei Leng. A note on adaptive group lasso. *Computational statistics & data analysis*, 52(12):5277–5286, 2008.

Dennis D Boos, Leonard A Stefanski, and Yujun Wu. Fast fsr variable selection with applications to clinical trials. *Biometrics*, 65(3):692–700, 2009.

Larry Wasserman and Kathryn Roeder. High dimensional variable selection. *Annals of statistics*, 37(5A):2178, 2009.

Shikai Luo and Subhashis Ghosal. Prediction consistency of forward iterated regression and selection technique. *Statistics & Probability Letters*, 107:79–83, 2015.

Arthur E Hoerl and Robert W Kennard. Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*, 12(1):55–67, 1970.

Bradley Efron, Trevor Hastie, Iain Johnstone, Robert Tibshirani, et al. Least angle regression. *The Annals of statistics*, 32(2):407–499, 2004.

Wenjiang J Fu. Penalized regressions: the bridge versus the lasso. *Journal of computational and graphical statistics*, 7(3):397–416, 1998.

Scott Shaobing Chen, David L Donoho, and Michael A Saunders. Atomic decomposition by basis pursuit. *SIAM review*, 43(1):129–159, 2001.

Adel Javanmard and Andrea Montanari. De-biasing the lasso: Optimal sample size for gaussian designs. *arXiv preprint arXiv:1508.02757*, 2015.

Xiaoying Tian, Joshua R Loftus, and Jonathan E Taylor. Selective inference with unknown variance via the square-root lasso. *arXiv preprint arXiv:1504.08031*, 2015.

Hui Zou. The adaptive lasso and its oracle properties. *Journal of the American statistical association*, 101(476):1418–1429, 2006.

Leo Breiman. Better subset regression using the nonnegative garrote. *Technometrics*, 37 (4):373–384, 1995.

Ming Yuan and Yi Lin. Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 68 (1):49–67, 2006.

Yi Lin, Hao Helen Zhang, et al. Component selection and smoothing in multivariate nonparametric regression. *The Annals of Statistics*, 34(5):2272–2297, 2006.

Jerome Friedman, Trevor Hastie, and Robert Tibshirani. A note on the group lasso and a sparse group lasso. *arXiv preprint arXiv:1001.0736*, 2010.

Michael Hamada and CF Jeff Wu. Analysis of designed experiments with complex aliasing. *Journal of Quality Technology;(United States)*, 24(3), 1992.

Hugh Chipman. Bayesian variable selection with related predictors. *Canadian Journal of Statistics*, 24(1):17–36, 1996.

JA Nelder. A reformulation of linear models. *Journal of the Royal Statistical Society. Series A (General)*, pages 48–77, 1977.

Julio L Peixoto. Hierarchical variable selection in polynomial regression models. *The American Statistician*, 41(4):311–313, 1987.

Peter McCullagh. Generalized linear models. *European Journal of Operational Research*, 16(3):285–292, 1984.

David R Cox. Interaction. *International Statistical Review/Revue Internationale de Statistique*, pages 1–24, 1984.

Jing Wu, Bernie Devlin, Steven Ringquist, Massimo Trucco, and Kathryn Roeder. Screen and clean: a tool for identifying interactions in genome-wide association studies. *Genetic epidemiology*, 34(3):275–285, 2010.

Peter J Bickel, Yaacov Ritov, Alexandre B Tsybakov, et al. Hierarchical selection of variables in sparse high-dimensional regression. In *Borrowing strength: theory powering applications–a Festschrift for Lawrence D. Brown*, pages 56–69. Institute of Mathematical Statistics, 2010.

Hugh B Crews, Dennis D Boos, and Leonard A Stefanski. Fsr methods for second-order regression models. *Computational statistics & data analysis*, 55(6):2026–2037, 2011.

Ning Hao and Hao Helen Zhang. Interaction screening for ultrahigh-dimensional data. *Journal of the American Statistical Association*, 109(507):1285–1301, 2014.

Naveen N Narisetty, Bhramar Mukherjee, Yin-Hsiu Chen, and Richard Gonzalez. Selection of non-linear interactions by a forward stepwise algorithm: Application to characterizing the health effects of environmental chemical mixtures. 2017.

Ming Yuan, V Roshan Joseph, and Hui Zou. Structured variable selection and estimation. *The Annals of Applied Statistics*, pages 1738–1757, 2009.

Nam Hee Choi, William Li, and Ji Zhu. Variable selection with the strong heredity constraint and its oracle property. *Journal of the American Statistical Association*, 105(489): 354–364, 2010.

Stephen Boyd, Neal Parikh, Eric Chu, Borja Peleato, and Jonathan Eckstein. Distributed optimization and statistical learning via the alternating direction method of multipliers. *Foundations and Trends® in Machine Learning*, 3(1):1–122, 2011.

Asad Haris, Daniela Witten, and Noah Simon. Convex modeling of interactions with strong heredity. *Journal of Computational and Graphical Statistics*, 25(4):981–1004, 2016.

Michael Lim and Trevor Hastie. Learning interactions through hierarchical group-lasso regularization. *arXiv preprint arXiv:1308.2719*, 2013.

Amir Beck and Marc Teboulle. A fast iterative shrinkage-thresholding algorithm for linear inverse problems. *SIAM journal on imaging sciences*, 2(1):183–202, 2009.

Peng Zhao, Guilherme Rocha, and Bin Yu. The composite absolute penalties family for grouped and hierarchical variable selection. *The Annals of Statistics*, pages 3468–3497, 2009.

Peter Radchenko and Gareth M James. Variable selection using adaptive nonlinear interaction structures in high dimensions. *Journal of the American Statistical Association*, 105 (492):1541–1553, 2010.

Daniel J Bauer and Li Cai. Consequences of unmodeled nonlinear effects in multilevel models. *Journal of Educational and Behavioral Statistics*, 34(1):97–114, 2009.

Marilyn C Cornelis, Eric J Tchetgen Tchetgen, Liming Liang, Lu Qi, Nilanjan Chatterjee, Frank B Hu, and Peter Kraft. Gene-environment interactions in genome-wide association studies: a comparative study of tests applied to empirical studies of type 2 diabetes. *American journal of epidemiology*, 175(3):191–202, 2012.

Bhramar Mukherjee, Jaeil Ahn, Stephen B Gruber, and Nilanjan Chatterjee. Testing gene-environment interaction in large-scale case-control association studies: possible choices and comparisons. *American journal of epidemiology*, 175(3):177–190, 2012.

Zihuai He, Min Zhang, Seunggeun Lee, Jennifer A Smith, Sharon LR Kardia, Ana V Diez Roux, and Bhramar Mukherjee. Set-based tests for gene-environment interaction in longitudinal studies. *Journal of the American Statistical Association*, (just-accepted), 2016.

Chenlei Leng, Yi Lin, and Grace Wahba. A note on the lasso and related procedures in model selection. *Statistica Sinica*, pages 1273–1284, 2006.

Ming Yuan and Yi Lin. On the non-negative garrotte estimator. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 69(2):143–161, 2007.

Hui Zou and Hao Helen Zhang. On the adaptive elastic-net with a diverging number of parameters. *Annals of statistics*, 37(4):1733, 2009.

Hansheng Wang, Guodong Li, and Guohua Jiang. Robust regression shrinkage and consistent variable selection through the lad-lasso. *Journal of Business & Economic Statistics*, 25(3):347–355, 2007.

Hao Helen Zhang and Wenbin Lu. Adaptive lasso for cox's proportional hazards model. *Biometrika*, 94(3):691–703, 2007.

Qing Pan and Yunpeng Zhao. Integrative weighted group lasso and generalized local quadratic approximation. *Computational Statistics & Data Analysis*, 104:66–78, 2016.

Jianqing Fan and Runze Li. Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American statistical Association*, 96(456):1348–1360, 2001.

John E Dennis Jr and Robert B Schnabel. *Numerical methods for unconstrained optimization and nonlinear equations*. SIAM, 1996.

Larry Armijo. Minimization of functions having lipschitz continuous first partial derivatives. *Pacific Journal of mathematics*, 16(1):1–3, 1966.

Per Kragh Andersen and Richard David Gill. Cox's regression model for counting processes: a large sample study. *The annals of statistics*, pages 1100–1120, 1982.

Krishnan Bhaskaran, Antonio Gasparrini, Shakoor Hajat, Liam Smeeth, and Ben Armstrong. Time series regression studies in environmental epidemiology. *International journal of epidemiology*, page dyt092, 2013.

Walter J Crinnion. The cdc fourth national report on human exposure to environmental chemicals: what it tells us about our toxic burden and how it assist environmental medicine physicians. *Altern Med Rev*, 15(2):101–109, 2010.

TED Schettler. Human exposure to phthalates via consumer products. *International journal of andrology*, 29(1):134–139, 2006.

I-Jen Wang, Ching-Chun Lin, Yen-Ju Lin, Wu-Shiun Hsieh, and Pau-Chung Chen. Early life phthalate exposure and atopic disorders in children: a prospective birth cohort study. *Environment international*, 62:48–54, 2014.

Raja Rezg, Saloua El-Fazaa, Najoua Gharbi, and Bessem Mornagui. Bisphenol a and human chronic diseases: current evidences, possible mechanisms, and future perspectives. *Environment international*, 64:83–90, 2014.

D John Betteridge. What is oxidative stress? *Metabolism*, 49(2):3–8, 2000.

Kelly K Ferguson, Thomas F McElrath, Yin-Hsiu Chen, Rita Loch-Caruso, Bhramar Mukherjee, and John D Meeker. Repeated measures of urinary oxidative stress biomarkers during pregnancy and preterm birth. *American journal of obstetrics and gynecology*, 212(2):208–e1, 2015.

Bayani Uttara, Ajay V Singh, Paolo Zamboni, and RT Mahajan. Oxidative stress and neurodegenerative diseases: a review of upstream and downstream antioxidant therapeutic options. *Current neuropharmacology*, 7(1):65–74, 2009.

Zheng Li, Courtney D Sandau, Lovisa C Romanoff, Samuel P Caudill, Andreas Sjodin, Larry L Needham, and Donald G Patterson. Concentration and profile of 22 urinary polycyclic aromatic hydrocarbon metabolites in the us population. *Environmental research*, 107(3):320–331, 2008.

Frank Onyemauwa, Stephen M Rappaport, Jon R Sobus, Dagmar Gajdošová, Renan Wu, and Suramya Waidyanatha. Using liquid chromatography–tandem mass spectrometry to quantify monohydroxylated metabolites of polycyclic aromatic hydrocarbons in urine. *Journal of Chromatography B*, 877(11):1117–1125, 2009.

John D Meeker, Antonia M Calafat, and Russ Hauser. Urinary phthalate metabolites and their biotransformation products: predictors and temporal variability among men and women. *Journal of Exposure Science and Environmental Epidemiology*, 22(4):376–385, 2012.