Title

Robust distributed lag models using data adaptive shrinkage.

Permalink
https://escholarship.org/uc/item/8g0602dj

Journal
Biostatistics (Oxford, England), 19(4)

ISSN
1465-4644

Authors
Chen, Yin-Hsiu
Mukherjee, Bhramar
Adar, Sara D
et al.

Publication Date
2018-10-01

DOI
10.1093/biostatistics/kxx041

Peer reviewed

Biostatistics (2018) **19**, 4, *pp.* 461–478
doi:10.1093/biostatistics/kxx041
Advance Access publication on October 12, 2017

Robust distributed lag models using data adaptive shrinkage

YIN-HSIU CHEN, BHRAMAR MUKHERJEE*

*Department of Biostatistics, University of Michigan, 1415 Washington Heights,
Ann Arbor, MI 48109, USA*

bhramar@umich.edu

SARA D. ADAR

*Department of Epidemiology, University of Michigan, 1415 Washington Heights,
Ann Arbor, MI 48109, USA*

VERONICA J. BERROCAL

*Department of Biostatistics, University of Michigan, 1415 Washington Heights,
Ann Arbor, MI 48109, USA*

BRENT A. COULL

Department of Biostatistics, Harvard University, 655 Huntington Avenue, Boston, MA 02115, USA

Summary

Distributed lag models (DLMs) have been widely used in environmental epidemiology to quantify the lagged effects of air pollution on an outcome of interest such as mortality or cardiovascular events. Generally speaking, DLMs can be applied to time-series data where the current measure of an independent variable and its lagged measures collectively affect the current measure of a dependent variable. The corresponding distributed lag (DL) function represents the relationship between the lags and the coefficients of the lagged exposure variables. Common choices include polynomials and splines. On one hand, such a constrained DLM specifies the coefficients as a function of lags and reduces the number of parameters to be estimated; hence, higher efficiency can be achieved. On the other hand, under violation of the assumption about the DL function, effect estimates can be severely biased. In this article, we propose a general framework for shrinking coefficient estimates from an unconstrained DLM, that are unbiased but potentially inefficient, toward the coefficient estimates from a constrained DLM to achieve a bias-variance trade-off. The amount of shrinkage can be determined in various ways, and we explore several such methods: empirical Bayes-type shrinkage, a hierarchical Bayes approach, and generalized ridge regression. We also consider a two-stage shrinkage approach that enforces the effect estimates to approach zero as lags increase. We contrast the various methods via an extensive simulation study and show that the shrinkage methods have better average performance across different scenarios in terms of mean squared error (MSE).

Keywords: Bayesian; Distributed lag model; Penalized regression; Shrinkage; Smoothing splines; Time series.

*To whom correspondence should be addressed.

We illustrate the methods by using data from the National Morbidity, Mortality, and Air Pollution Study (NMMAPS) to explore the association between PM_{10}, O_3, and SO_2 on three types of disease event counts in Chicago, IL, from 1987 to 2000.

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 distributed lag models (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 distributed lag (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 *and others*, 2000). 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 *and others*, 2000) use splines to represent the DL function. Muggeo (2008) proposed a flexible segmented break point model with doubly penalized B-splines. Distributed lag non-linear models (DLNMs) (Gasparrini *and others*, 2010) were developed to simultaneously model the non-linear exposure-response dependencies and non-linear DL function. Bayesian DLM (BDLM) (Welty *and others*, 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 *and others*, 2009). Obermeier *and others* (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.

In this article, we consider several alternative approaches for shrinkage and smoothing of the distributed lag function. We propose a class of shrinkage methods that shrinks the unconstrained DLM estimator toward a model-dependent constrained DLM estimator. The notion is to retain the flexibility of unconstrained DLM and gain estimation efficiency from a parsimonious constrained DLM. The first approach is to perform component-wise shrinkage by combining the two estimators using an empirical-Bayes (EB) type of weighting (Mukherjee and Chatterjee, 2008; Chen *and others*, 2009). The second approach is a new hierarchical Bayes (HB) approach. The third approach is generalized ridge regression (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 *and others*, 1998) and generalized cross-validation (GCV) (Golub *and others*, 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. In Section 2, we introduce our shrinkage approaches in detail. In Section 3, we conduct an extensive simulation study to compare the proposed approaches to existing alternatives. In Section 4, we illustrate our methods by analyzing data from the National Morbidity, Mortality, and Air Pollution Study (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, IL, from 1987 to 2000. Section 5 contains concluding remarks.

2. Methods

We use the following notation throughout the article. 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 $g[E(y_t|x_t, x_{t-1}, ..., x_{t-L}, \boldsymbol{z}_t)] = \alpha_0 + \boldsymbol{z}_t^T\boldsymbol{\alpha}_1 + \sum_{\ell=0}^{L}\beta_\ell x_{t-\ell}$ where α_0 is the intercept, $\boldsymbol{\alpha}_1$ represents the effect of covariates, L is the pre-determined maximum number of lags, and $\boldsymbol{\beta} = (\beta_0, \beta_1, ..., \beta_L)^T$ is the vector of lagged effects. We first consider the log-linear Poisson model:

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

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

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 maximum likelihood estimation (MLE). Let $\boldsymbol{X_t} = (x_t, x_{t-1}, ..., x_{t-L})^T$. The unconstrained DLM estimator $\hat{\boldsymbol{\beta}}_{\text{UDLM}}$ can be written as

$$\hat{\boldsymbol{\beta}}_{\text{UDLM}} = \arg\max_{\boldsymbol{\beta}} \ell_u(\boldsymbol{\beta}) = \arg\max_{\boldsymbol{\beta}} \sum_{t=1}^{T}[y_t\boldsymbol{\beta}^T\boldsymbol{X_t} - e^{\boldsymbol{\beta}^T\boldsymbol{X_t}} - \log(y_t!)]. \tag{2.1}$$

Constrained DLM imposes structure on $\boldsymbol{\beta}$ by assuming β_ℓ is a known function of ℓ for $\ell = 0, \ldots, L$. We assume that $B_1(\cdot), \ldots, B_p(\cdot)$ are the p basis functions that generate the class of functions in which $\boldsymbol{\beta}$ lies. A transformation matrix $\boldsymbol{C}$ (Gasparrini *and others*, 2010) is defined as a $(L+1) \times p$ matrix where the element $(\ell+1, j)$ is the jth 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 $\boldsymbol{W}_t = \boldsymbol{C}^T \boldsymbol{X}_t$ 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 $\hat{\boldsymbol{\beta}}_{\text{CDLM}} = \boldsymbol{C}\hat{\boldsymbol{\theta}}$ where

$$\hat{\boldsymbol{\theta}} = \arg\max_{\boldsymbol{\theta}} \ell_c(\boldsymbol{\theta}) = \arg\max_{\boldsymbol{\theta}} \sum_{t=1}^{T} [y_t \boldsymbol{\theta}^T \boldsymbol{W_t} - e^{\boldsymbol{\theta}^T \boldsymbol{W_t}} - \log(y_t!)]. \tag{2.2}$$

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

Note that the choice of basis functions for constructing $\boldsymbol{C}$ is unique only up to a full-rank linear transformation. In Sections 2.2–2.4, we will introduce different approaches to shrink $\hat{\boldsymbol{\beta}}_{\text{UDLM}}$ toward $\hat{\boldsymbol{\beta}}_{\text{CDLM}}$ in a data-adaptive manner.

2.1. *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 (2.2). 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 (2.1) subject to the constraints. The constraints can be represented by $\boldsymbol{R}\boldsymbol{\beta} = \boldsymbol{0}$ where $\boldsymbol{R}$ is the $(L + 1 - p) \times (L + 1)$ constraint matrix. The basis functions in $\boldsymbol{C}$ span the solution space of $\boldsymbol{R}\boldsymbol{\beta} = \boldsymbol{0}$, 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) $\boldsymbol{C}_e^T = \boldsymbol{U}_C\boldsymbol{D}_C\boldsymbol{V}_C^T$ 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 $\boldsymbol{C}_e^T$ 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 $\boldsymbol{V}_C^T$. More detailed description of the connection between $\boldsymbol{R}$ and $\boldsymbol{C}$ is provided in the supplementary material available at *Biostatistics* online. We summarize two important results that are going to be used in the subsequent development.

Result 1: $\hat{\boldsymbol{\beta}}_{\text{CDLM}} = \boldsymbol{C}\hat{\boldsymbol{\theta}}$, where $\hat{\boldsymbol{\theta}}$ is as given in (2.2), is equivalent to the maximizer of the likelihood function in (2.1) subject to the constraint $\boldsymbol{R}\boldsymbol{\beta} = \boldsymbol{0}$, 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 article but Results 1 and 2 suggest that the methods are generalizable to other more flexible DLMs.

2.2. *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 *and others* (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

$$\hat{\boldsymbol{\beta}}_{\mathrm{EB}} = \hat{\boldsymbol{\beta}}_{\mathrm{UDLM}} + \boldsymbol{K}(\hat{\boldsymbol{\beta}}_{\mathrm{CDLM}} - \hat{\boldsymbol{\beta}}_{\mathrm{UDLM}}) \tag{2.3}$$

with $\boldsymbol{K} = (\hat{\boldsymbol{V}} \circ \boldsymbol{I}_{L+1})[(\hat{\boldsymbol{V}} + \hat{\boldsymbol{\psi}}\hat{\boldsymbol{\psi}}^T) \circ \boldsymbol{I}_{L+1}]^{-1}$. $\hat{\boldsymbol{V}}$ is the estimated variance-covariance matrix of $\hat{\boldsymbol{\beta}}_{\mathrm{UDLM}}$, $\hat{\boldsymbol{\psi}} = \hat{\boldsymbol{\beta}}_{\mathrm{CDLM}} - \hat{\boldsymbol{\beta}}_{\mathrm{UDLM}}$, $\boldsymbol{I}_{L+1}$ is a $(L+1) \times (L+1)$ identity matrix, and $\circ$ is the Hadamard product. The shrinkage factor can be represented by $\boldsymbol{K} = \mathrm{diag}[k_1, ..., k_{L+1}]$ with $k_i = v_i/(v_i + \hat{\psi}_i^2)$ where $\hat{\psi}_i^2$ is the ith diagonal component of $\hat{\boldsymbol{\psi}}\hat{\boldsymbol{\psi}}^T$, and v_i is the ith 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 $\hat{\boldsymbol{\psi}}\hat{\boldsymbol{\psi}}^T$ in (2.3). The expression and derivation of the variance–covariance estimate of $\hat{\boldsymbol{\psi}}$ are given in the supplementary material available at *Biostatistics* online. From now on, we will denote the EB estimator in (2.3) as EB1 and the EB estimator that replaces $\hat{\boldsymbol{\psi}}\hat{\boldsymbol{\psi}}^T$ with $\hat{\mathrm{Cov}}(\hat{\boldsymbol{\psi}})$ in (2.3) 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 $\hat{\boldsymbol{\beta}}_{\mathrm{EB}} = \hat{\boldsymbol{\beta}}_{\mathrm{CDLM}}$. At the other extreme, $\boldsymbol{K} = \boldsymbol{0}$ yields $\hat{\boldsymbol{\beta}}_{\mathrm{EB}} = \hat{\boldsymbol{\beta}}_{\mathrm{UDLM}}$. When the working DL function in CDLM is not correctly specified, $\hat{\boldsymbol{\beta}}_{\mathrm{EB}}$ is asymptotically equivalent to $\hat{\boldsymbol{\beta}}_{\mathrm{UDLM}}$ and therefore $\hat{\boldsymbol{\beta}}_{\mathrm{EB}}$ is consistent. The expression of the asymptotic variance–covariance of $\hat{\boldsymbol{\beta}}_{\mathrm{EB}}$ and its derivation are provided in the supplementary material available at *Biostatistics* online. The limiting distribution of $\sqrt{N}(\hat{\boldsymbol{\beta}}_{\mathrm{EB}} - \boldsymbol{\beta})$ is not a normal distribution as expected for most model averaged estimators (Claeskens and Carroll, 2007). However, Chen *and others* (2009) showed that the normal approximation works well and is acceptable in practice.

2.3. *Hierarchical Bayes model*

We propose a HB approach that sets up a non-null 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 (2.1) 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 $\hat{\boldsymbol{\beta}}_{\mathrm{UDLM}}$ toward $\hat{\boldsymbol{\beta}}_{\mathrm{CDLM}}$. The brief proof is provided in the supplementary material available at *Biostatistics* online. 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 $\boldsymbol{X} = (\boldsymbol{X}_1, \cdots, \boldsymbol{X}_T)^T$ and an outcome vector $\boldsymbol{Y} = (y_1, \cdots, y_T)^T$ of length T. In order to allow uncertainty on the variance component σ_π^2, we specify the full HB model as:

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

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 given by

$$f(\sigma_\pi^2|\boldsymbol{\beta},\boldsymbol{Y}) \propto IG(a_\pi + M/2, b_\pi + \boldsymbol{\beta}^T\boldsymbol{R}^T\boldsymbol{R}\boldsymbol{\beta}/2)$$

$$f(\boldsymbol{\beta}|\sigma_\pi^2,\boldsymbol{Y}) \propto \prod_{t=1}^{T}[\exp(y_t\boldsymbol{X}_t^T\boldsymbol{\beta} - e^{\boldsymbol{X}_t^T\boldsymbol{\beta}})]\cdot\exp\left(-\frac{\boldsymbol{\beta}^T\boldsymbol{R}^T\boldsymbol{R}\boldsymbol{\beta}}{2\sigma_\pi^2}\right).$$

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 $\hat{\beta}_{\text{HB}}$ 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

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

where $\boldsymbol{R}$ is defined previously and λ is the tuning parameter. We can use the Newton–Raphson algorithm (Gill *and others*, 1981) to obtain GRR estimator $\hat{\boldsymbol{\beta}}_{\text{GRR}}$ by minimizing $\ell_p(\boldsymbol{\beta})$ given λ. GCV (Golub *and others*, 1979) and AICC (Hurvich *and others*, 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 $\hat{\boldsymbol{\beta}}_{\text{GRR}} \to \hat{\boldsymbol{\beta}}_{\text{CDLM}}$ as $\lambda \to \infty$ and $\hat{\boldsymbol{\beta}}_{\text{GRR}} \to \hat{\boldsymbol{\beta}}_{\text{UDLM}}$ 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 $\hat{\boldsymbol{\beta}}_{\text{GRR}}$ is a monotonic decreasing function of λ, the asymptotic bias of $\hat{\boldsymbol{\beta}}_{\text{GRR}}$ is a monotonic increasing function of λ, and the asymptotic mean square errors (MSE) of $\hat{\boldsymbol{\beta}}_{\text{GRR}}$ is lower than the asymptotic MSE of $\hat{\boldsymbol{\beta}}_{\text{UDLM}}$. The proofs are provided in the supplementary material available at *Biostatistics* online. The described asymptotic properties assume that the tuning parameter λ is fixed. Choosing λ from data would induce another layer of uncertainty in $\hat{\boldsymbol{\beta}}_{\text{GRR}}$ 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 3.

2.4. *Two-stage shrinkage*

The Bayesian distributed lag model (BDLM) proposed by Welty *and others* (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 $\boldsymbol{V}(\omega) = \text{diag}[1,\exp(\omega),\exp(2\omega),...,\exp(L\omega)]$, $\boldsymbol{W}(\omega_2) = \boldsymbol{V}(\omega_2)\boldsymbol{V}(\omega_2)^T +$ $\{\boldsymbol{I}_{L+1} - \boldsymbol{V}(\omega_2)\}\mathbf{1}_{L+1}\mathbf{1}_{L+1}^T\{\boldsymbol{I}_{L+1} - \boldsymbol{V}(\omega_2)\}^T$, $\boldsymbol{I}_{L+1}$ 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 $\boldsymbol{\omega} = (\omega_1,\omega_2)^T$, Welty *and others* (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 $\hat{\boldsymbol{\beta}}_{\text{UDLM}}$ toward $\hat{\boldsymbol{\beta}}_{\text{CDLM}}$ through one of the shrinkage approaches introduced in Sections 2.2–2.3. 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 $\hat{\boldsymbol{\beta}}_{\text{EB}}$ 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 2.2, is given by:

$$\hat{\boldsymbol{\beta}}_{\text{EB}}|\boldsymbol{\beta} \sim \mathcal{N}(\boldsymbol{\beta}, \boldsymbol{G\Sigma G}^T)$$

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

where $\boldsymbol{V}(\omega)$ and $\boldsymbol{W}(\omega)$ are as defined in Section 2.3. If we let $\boldsymbol{\omega} = (\omega_1, \omega_2)^T$ have a discrete uniform prior distribution, the full conditional distributions of $\boldsymbol{\beta}, \sigma^2$, and $\boldsymbol{\omega}$ are given by:

$$f(\boldsymbol{\beta}|\hat{\boldsymbol{\beta}}_{\text{EB}}, \boldsymbol{\omega}, \sigma^2) \sim N([1/\sigma^2\boldsymbol{\Omega}(\boldsymbol{\omega})^{-1} + (\boldsymbol{G\Sigma G}^T)^{-1}]^{-1}(\boldsymbol{G\Sigma G}^T)^{-1}\hat{\boldsymbol{\beta}}_{EB}, [1/\sigma^2\boldsymbol{\Omega}(\boldsymbol{\omega})^{-1} + (\boldsymbol{G\Sigma G})^{-1}]^{-1})$$

$$p(\boldsymbol{\omega}|\hat{\boldsymbol{\beta}}_{\text{EB}}, \boldsymbol{\beta}, \sigma^2) = \frac{|\boldsymbol{\Omega}(\boldsymbol{\omega})|^{-1/2}\exp[-\frac{1}{2\sigma^2}\boldsymbol{\beta}^T\boldsymbol{\Omega}(\boldsymbol{\omega})^{-1}\boldsymbol{\beta}]}{\sum_{\boldsymbol{\omega}^*}|\boldsymbol{\Omega}(\boldsymbol{\omega}^*)|^{-1/2}\exp[-\frac{1}{2\sigma^2}\boldsymbol{\beta}^T\boldsymbol{\Omega}(\boldsymbol{\omega}^*)^{-1}\boldsymbol{\beta}]}$$

$$f(\sigma^2|\hat{\boldsymbol{\beta}}_{\text{EB}}, \boldsymbol{\beta}, \boldsymbol{\omega}) \sim IG(a_0 + (L+1)/2, b_0 + \boldsymbol{\beta}^T\boldsymbol{\Omega}(\boldsymbol{\omega})^{-1}\boldsymbol{\beta}/2).$$

The joint posterior distribution can be obtained via a Gibbs sampling technique and the two-stage shrinkage estimate $\hat{\boldsymbol{\beta}}_{\text{HB2}}$ can be obtained accordingly.

The analogue of the previous two-stage Hierarchical Bayesian 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 $\hat{\boldsymbol{\beta}}_{\text{EB}}$ 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

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

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 $\hat{\boldsymbol{\beta}}_{\text{GRR}}$ is chosen as the shrinkage estimator from the first stage, a similar framework can be followed.

3. Simulation study

3.1. *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 2.2 and 2.3 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 $\boldsymbol{\omega} = (\omega_1, \omega_2)^T$ 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, with both inverse gamma prior parameters set to 0.001.

3.1.1. *Simulation settings.* We first generated an exposure series of length 200 with mean 0 and first order autocorrelation 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 *and others* (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 $\boldsymbol{\beta} = (\beta_0, ..., \beta_{10})^T$ 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. The three combinations of true coefficients and specified working DL function reflect the first three scenarios of interest for comparing various methods: (i) the working DL function completely matches true DL function, (ii) the working DL function moderately departs from the true DL function, and (iii) 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 (i) constant, (ii) linear, (iii) cubic, (iv) cubic-like smooth function with slight departure, and (v) 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 the supplementary material available at *Biostatistics* online. We generated 1000 data sets for each scenario to evaluate the estimation performance.

3.1.2. *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 mean squared errors (MSE) of UDLM estimates to the MSE of the estimate under each method. The expressions of the metrics used for comparison are summarized in the supplementary material available at *Biostatistics* online.

3.1.3. *Results.* The simulation results for the estimated lagged coefficient vector ($\hat{\boldsymbol{\beta}}$) are summarized in the upper part of Table 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

Table 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*

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

Bold values corresponding to the best performer in each row. †The working distributed lag (DL) function in CDLM for CDLM, EB1, EB2, GRR, 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 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

Table 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*

	EB1	HB2-EB1	HP-EB1	GRR	HB2-GRR	HP-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

Bold values corresponding to the best performer in each row.

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.

3.2. *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, HB2 with EB1 from the first stage (HB2–EB1), HP with EB1 from the first stage (HP–EB1), GRR, HB2 with GRR from the first stage (HB2–GRR), HP with GRR from the first stage (HP–GRR), and BDLM. For BDLM and HB2, the prior on $\boldsymbol{\omega} = (\omega_1, \omega_2)^T$ 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 HP, $\boldsymbol{\omega}$ was chosen as the minimizer of the hyper-penalized criterion. The tuning parameters in HP–EB1 and HP–GRR were selected based on 5-fold cross-validation. The working DL function in CDLM was specified as a cubic polynomial throughout.

3.2.1. *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 $\boldsymbol{\beta} = (0.07, 0.135, 0.2, 0.210, 0.18, 0.125, 0.06, 0.02, 0, 0, 0, 0, 0, 0, 0, 0)^T$ 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.

3.2.2. *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.

3.2.3. *Results.* The results are presented in Table 2. Overall, the two-stage approaches are effective in increasing efficiency when L is misspecified. Both HB2 and HP further reduce MSE and reduce the mean distance compared to the shrinkage estimator obtained in the first stage. Specifically, compared to EB1 (1.83), HB2–EB1 (1.95) and HP–EB1 (1.98) have higher efficiencies while all three are less efficient than

BDLM (3.42); in contrast, HB2–GRR (10.47) and HP–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 (HB2–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 the supplementary material available at *Biostatistics* online. 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 generalized additive models 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 HP–GRR in the analysis of NMMAPS data.

4. Application to NMMAPS data

We first explore the association of (i) daily particular matter with aerodynamic diameter less than 10 microns (PM_{10}), (ii) daily ozone concentration (O_3), and (iii) daily sulfur dioxide concentration (SO_2) with (1) daily non-accidental mortality counts, (2) daily cardiovascular mortality counts, and (3) daily respiratory mortality counts in Chicago, IL 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, HB2–GRR, HP–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 *and others* (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 2 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 $\mathbf{z}_{tk}$ 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 $\mathbf{z}_{tk}$ is specified in the same way as in previous analysis by Dominici *and others* (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

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

where $\boldsymbol{X}_t = (x_t, ..., x_{t-14})^T$, $\boldsymbol{\beta} = (\beta_0, ..., \beta_{14})^T$, I$(\cdot)$ is the indicator function,ns$(\cdot)$ denotes the natural spline with specified 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, respectively. 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 morality 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 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 *and others*, 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 supplementary material available at *Biostatistics* online. One can notice the slower decay and stronger autocorrelation in O$_3$ time series than in PM$_{10}$ time series. Figure 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.

4.1. *Estimation of lag coefficients*

With respect to PM$_{10}$, the strongest association occurs at lag 3 for UDLM, EB1, GRR, BDLM, and HP–GRR and at lag 2 for CDLM, HB, and HB2–GRR. The interquartile range of PM$_{10}$ is 21.49μg/m^3. The

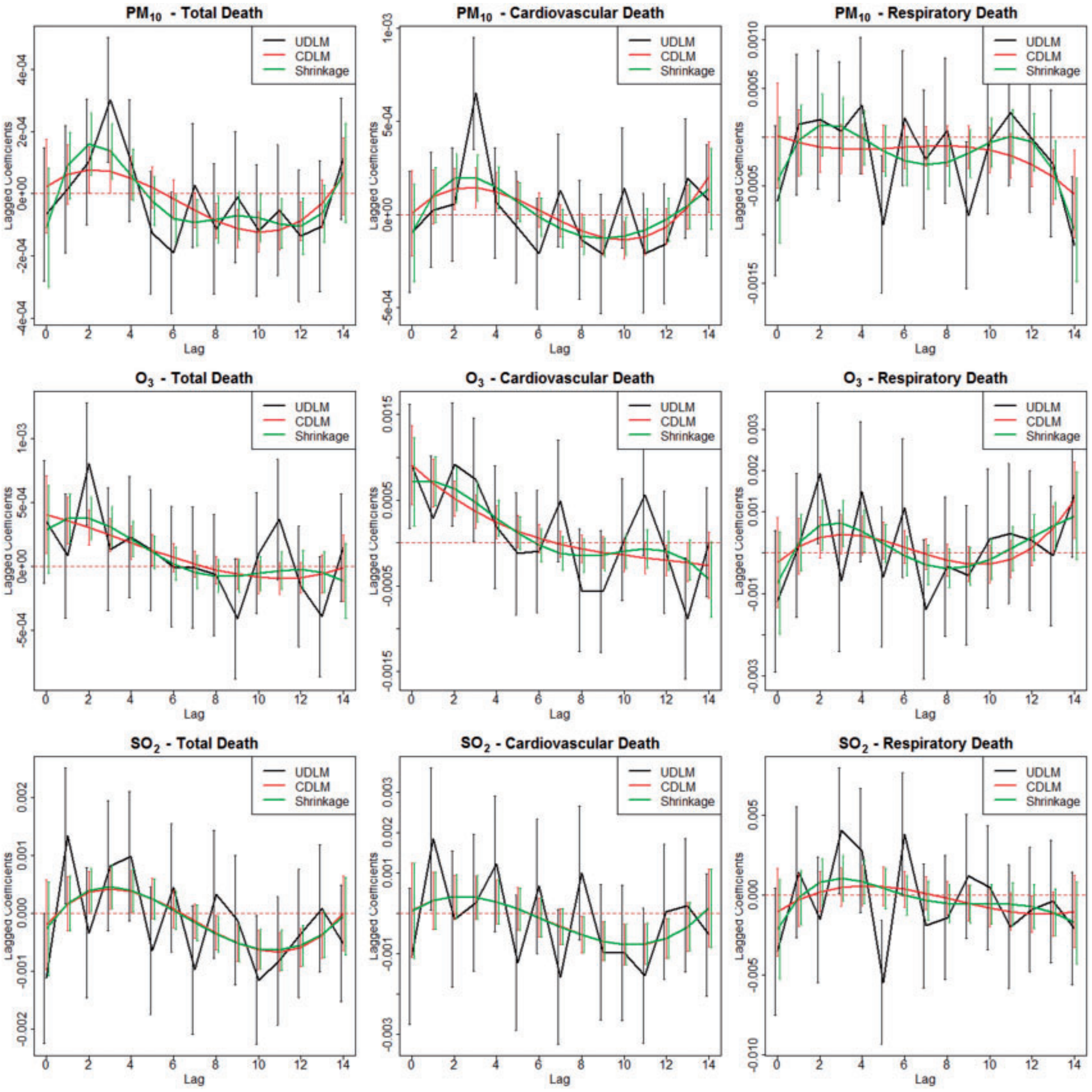

Fig. 1. Estimated distributed lag functions up to 14 days for PM_{10}, O_3, and SO_2 on total morality, cardiovascular mortality, and respiratory mortality with 95% confidence/credible interval at each lag in Chicago, IL from 1987 to 2000 based on the NMMAPS data.

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 HP-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, 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

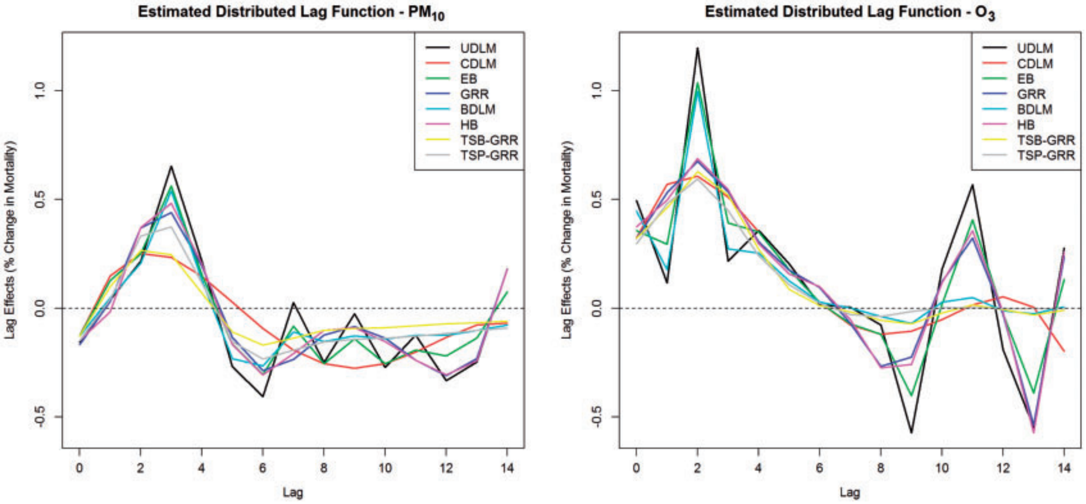

Fig. 2. Estimated distributed lag functions up to 14 days for PM_{10} (left) and O_3 (right) on mortality in Chicago, IL from 1987 to 2000 based on the data from the NMMAPS under eight estimation methods. The lag effects are presented as the percentage change in mortality with an interquartile range increase in the exposure level (PM_{10}: $21.49\mu g/m^3$, O_3: 14.65 ppb).

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 8 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 measure 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 supplementary material available at *Biostatistics* online.

4.2. *Estimation of cumulative lag coefficients*

Supplementary material available at *Biostatistics* online 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 3. 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

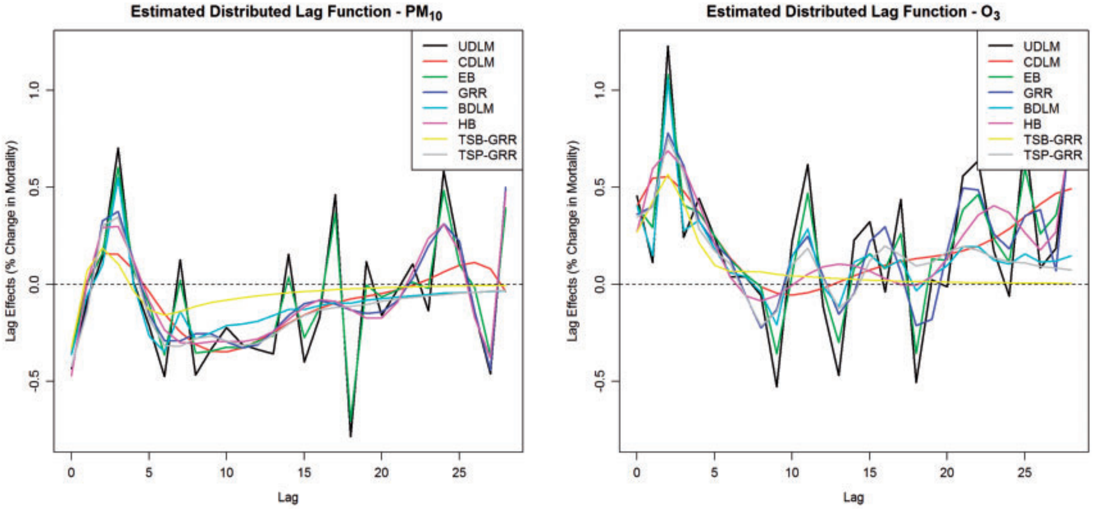

Fig. 3. Estimated mean and 95% confidence/credible interval of the cumulative lagged effect (% change in mortality count) up to 3, 7, and 14 days of PM_{10} (left) and O_3 (right) on mortality with an interquartile range increase in exposure level (PM_{10}: 21.49 $\mu g/m^3$, O_3: 14.65 ppb) in Chicago,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

(Zanobetti *and others*, 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, we can also observe that the two-stage shrinkage methods HB2–GRR and HP–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).

5. Discussion

In this article, 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, supplementary material available at *Biostatistics* online 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 combinations. 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 non-null 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 article. 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.

Supplementary material

Supplementary material is available at http://biostatistics.oxfordjournals.org.

Acknowledgments

Conflict of Interest: None declared.

FUNDING

National Science Foundation (DMS 1406712); National Institutes of Health (ES 20811).

REFERENCES

ALMON, S. (1965). The distributed lag between capital appropriations and expenditures. *Econometrica: Journal of the Econometric Society* **33**, 178–196.

CHEN, Y., CHATTERJEE, N. AND CARROLL, R. J. (2009). Shrinkage estimators for robust and efficient inference in haplotype-based case-control studies. *Journal of the American Statistical Association* **104**, 220–233.

CLAESKENS, G. AND CARROLL, R. J. (2007). An asymptotic theory for model selection inference in general semiparametric problems. *Biometrika* **94**, 249–265.

CORRADI, C. (1977). Smooth distributed lag estimators and smoothing spline functions in Hilbert spaces. *Journal of Econometrics* **5**, 211–219.

DOMINICI, F., MCDERMOTT, A., DANIELS, M., ZEGER, S. L. AND SAMET, J. M. (2005). 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**, 1071–1092.

FARRAR, D. E. AND GLAUBER, R. R. (1967). Multicollinearity in regression analysis: the problem revisited. *The Review of Economic and Statistics*, 92–107.

GASPARRINI, A., ARMSTRONG, B. AND KENWARD, M. G. (2010). Distributed lag non-linear models. *Statistics in Medicine* **29**, 2224–2234.

GILL, P. E., MURRAY, W. AND WRIGHT, M. H. (1981). *Practical optimization*. Bingley, United Kingdom: Emerald Group Publishing Limited.

GOLUB, G. H., HEATH, M. AND WAHBA, G. (1979). Generalized cross-validation as a method for choosing a good ridge parameter. *Technometrics* **21**, 215–223.

HASTIE, T. AND TIBSHIRANI, R. (1993). Varying-coefficient models. *Journal of the Royal Statistical Society. Series B (Methodological)*, 757–796.

HURVICH, C. M., SIMONOFF, J. S. AND TSAI, C. -L. (1998). Smoothing parameter selection in nonparametric regression using an improved Akaike information criterion. *Journal of the Royal Statistical Society: Series B* (*Statistical Methodology*) **60**, 271–293.

MARRA, G. AND WOOD, S. N. (2012). Coverage properties of confidence intervals for generalized additive model components. *Scandinavian Journal of Statistics* **39**, 53–74.

MUGGEO, V. M. R. (2008). Modeling temperature effects on mortality: multiple segmented relationships with common break points. *Biostatistics* **9**, 613–620.

MUKHERJEE, B. AND CHATTERJEE, N. (2008). Exploiting gene-environment independence for analysis of case–control studies: an empirical Bayes-type shrinkage estimator to trade-off between bias and efficiency. *Biometrics* **64**, 685–694.

OBERMEIER, V., SCHEIPL, F., HEUMANN, C., WASSERMANN, J. AND KÜCHENHOFF, H. (2015). Flexible distributed lags for modelling earthquake data. *Journal of the Royal Statistical Society: Series C* (*Applied Statistics*) **64**, 395–412.

PENG, R. D., DOMINICI, F. AND WELTY, L. J. (2009). 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**, 3–24.

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

SCHWARTZ, J. (1994). Air pollution and daily mortality: a review and meta analysis. *Environmental Research* **64**, 36–52.

SCHWARTZ, J. (2000). The distributed lag between air pollution and daily deaths. *Epidemiology* **11**, 320–326.

SCHWARTZ, J. AND DOCKERY, D. W. (1992). Increased mortality in philadelphia associated with daily air pollution concentrations. *American Review of Respiratory Disease* **145**, 600–604.

SHILLER, R. J. (1973). A distributed lag estimator derived from smoothness priors. *Econometrica: Journal of the Econometric Society*, 775–788.

WELTY, L. J., PENG, R. D., ZEGER, S. L. AND DOMINICI, F. (2009). Bayesian distributed lag models: estimating effects of particulate matter air pollution on daily mortality. *Biometrics* **65**, 282–291.

ZANOBETTI, A. AND SCHWARTZ, J. (2008). 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**, 184–189.

ZANOBETTI, A., SCHWARTZ, J., SAMOLI, E., GRYPARIS, A., TOULOUMI, G., ATKINSON, R., LE TERTRE, A., BOBROS, J., CELKO, M., GOREN, A. *and others*. (2002). The temporal pattern of mortality responses to air pollution: a multicity assessment of mortality displacement. *Epidemiology* **13**, 87–93.

ZANOBETTI, A., WAND, M. P., SCHWARTZ, J. AND RYAN, L. M. (2000). Generalized additive distributed lag models: quantifying mortality displacement. *Biostatistics* **1**, 279–292.

[*Received April 18, 2016; revised March 10, 2017; accepted for publication May 11, 2017*]