Biostatistics (2013), **0**, 0, *pp.* 1–23
doi:10.1093/biostatistics/HDDLM

Extending Distributed Lag Models to Higher Degrees

MATTHEW J. HEATON*

Institute for Mathematics Applied to Geosciences, National Center for Atmospheric Research, Box 3000,
Boulder CO, 80307-3000, USA

heaton@ucar.edu

and ROGER D. PENG

Department of Biostatisics, Johns Hopkins Bloomberg School of Public Health, Baltimore, MD

Summary

Distributed lag models relate lagged covariates to a response and are a popular statistical model used in a wide variety of disciplines to analyze exposure-response data. However, classical distributed lag models do not account for possible interactions between lagged predictors. In the presence of interactions between lagged covariates, the total effect of a change on the response is not merely a sum of lagged effects as is typically assumed. This article proposes a new class of models, called high degree distributed lag models, that extend basic distributed lag models to incorporate hypothesized interactions between lagged predictors. The modeling strategy utilizes Gaussian processes to counterbalance predictor collinearity and as a dimension reduction tool. To choose the degree and maximum lags used within the models, a computationally manageable model comparison method is proposed based on maximum *a posteriori* estimators. The models and methods are illustrated via simulation and application to investigating the effect of heat exposure on mortality in Los Angeles and New York.

*To whom correspondence should be addressed.

Key words: Dimension reduction; Gaussian process; Heat Exposure; Lagged interaction; NMMAPS dataset.

1. Introduction

Given projected increases in global temperatures as a result of climate change (IPCC, 2007), investigating the potential impact of such increases on public health is receiving escalated attention. Numerous studies have linked high temperatures to excess mortality (O'Neill *and others*, 2003; Kovats and Hajat, 2008; Anderson and Bell, 2009) and more recently such impacts have been projected into the future (Li *and others*, 2011; Peng *and others*, 2011). When investigating the effect of heat on mortality (or, likewise, morbidity), it has been hypothesized that the effects of heat may extend over multiple days and that consecutive days of heat may be more harmful than individual hot days. The distributed lag (DL) model is a common tool to investigate this hypothesis and various studies have used it to relate high heat exposure over long time periods to spikes in mortality or hospitalizations (Hajat *and others*, 2005; Anderson and Bell, 2009).

Basic distributed lag models relate a response at time t, say Y_t, to temporally lagged covariates $X_t, X_{t-1}, \dots$ via,

$$g(\mathbb{E}(Y_t)) = \alpha + \boldsymbol{z}_t'\boldsymbol{\gamma} + \sum_{\ell=0}^{M} \theta_\ell X_{t-\ell} \tag{1.1}$$

where $g(\cdot)$ is an appropriate link function and $\boldsymbol{z}_t'$ is a vector of confounding covariates with associated coefficients $\boldsymbol{\gamma}$. In nearly all cases, Y_t is follows a distribution in the exponential family but this is not necessary. The basic DL model in (1.1), however, can not directly examine whether the effect of multiple days of heat go beyond an additive effect. For example, one social hypothesis is that individuals adjust their behavior for avoiding excessive heat on day t based on the heat on day $t-1$ (e.g. if it was raining on day $t-1$, resulting in a cool day, then more people may head outdoors on day t than had day $t-1$ been sunny (hot), suggesting an interaction between lagged effects). Under this hypothesis, the effect of two consecutive days of high heat on public health may not simply be the sum of the effect of high heat on day t and the effect of high heat on day $t-1$ due to the adjustment in behavior. Motivated by the epidemiological need to investigate hypotheses about lagged interactions in heat-mortality studies, this article proposes a class of models, called high degree DL models (HDDLMs), that extend the basic DL model to incorporate such hypothesized lagged interactions up to a given order δ.

Extending DL models to account for high degree interactions presents a number of interesting statistical modeling challenges. For example, if $\{X_{t-\ell}\}_{\ell=0}^{M}$ exhibit strong autocorrelation (high collinearity), then the variance of the estimates of $\{\theta_\ell\}_{\ell=0}^{M}$ in (1.1) are inflated, rendering it difficult to detect significance. This collinearity issue is well documented with various solutions having been put forward such as averaging over temporal windows (Caffo *and others*, 2011), constraining the coefficients to follow a function (Schwartz, 2000) or building strong prior contraints (Welty *and others*, 2009). For HDDLMs, not only are first order predictors $\{X_{t-\ell}\}$ included but also higher order interactions which can magnify collinearity problems. To counter-balance collinearity, this article relies on and extends the Gaussian process (GP) prior of Heaton and Peng (2012) to construct a predictive process prior to enforce *a priori* correlation among higher order interaction coefficients. This predictive process prior also restricts the dimension of the parameter space for problems that consider a high degree of interactions.

As with all DL models, choosing how many temporal lags to include is an important aspect. The majority of previous approaches simply fix the lag length based on *a priori* knowledge and fit the associated model. An exception is Heaton and Peng (2012) who treat the maximum lag as an additional parameter and estimate it by sampling from its posterior distribution but note the computational complexity in doing so. As a computationally efficient alternative, this article proposes a method for choosing maximum lags using maximum *a posteriori* (MAP) estimators within information criterions. Because prior constraints are essential for handling collinearity, using MAP estimators (rather than MLEs) within information criterion properly incorporates necessary prior constraints into classical model comparison criterion.

To summarize, the primary contributions of this article are to (i) extend basic DL models to include interactions between lagged predictors, (ii) propose a prior structure that, not only deals with strong collinearity in the predictors, but also offers dimension reduction, (iii) propose a computationally efficient approach to estimating maximum lags and (iv) illustrate the usefulness of the proposed methods for investigating heat effects on mortality. Section 2 describes the modeling strategy for higher degree DL models. Section 3 describes techniques for parameter estimation. Section 4 evaluates the proposed modeling strategy using simulation. Section 5 illustrates the methods by quantifying the risk of high temperatures on mortality and Section 6 concludes and outlines further extensions.

2. Methodology

2.1 *Model Definition*

Let Y_t denote a response variable of interest observed at time $t \in \mathbb{Z}$ and let X_t denote a covariate which is (potentially) associated with Y_t. Throughout this article, Y_t is related to $\{X_{t-\ell} : \ell = 0, 1, \dots\}$ via a high degree distributed lag model of degree $\delta \in \{1, 2, \dots\}$ (denoted DL$^\delta$) which is defined as,

$$
\begin{aligned}
g(\mathbb{E}(Y_t)) = \alpha + \boldsymbol{z}_t'\boldsymbol{\gamma} + \sum_{\ell=0}^{M_1} \theta_\ell X_{t-\ell} + \sum_{\ell_1=0}^{M_2}\sum_{\ell_2=\ell_1}^{M_2} \theta_{\ell_1\ell_2} X_{t-\ell_1}X_{t-\ell_2} + \cdots \\
+ \sum_{\ell_1=0}^{M_\delta} \cdots \sum_{\ell_\delta=\ell_{(\delta-1)}}^{M_\delta} \theta_{\ell_1\cdots\ell_\delta} X_{t-\ell_1}\cdots X_{t-\ell_\delta} \qquad (2.2)
\end{aligned}
$$

where $g(\cdot)$ is a link function (e.g. identity, log, logit, etc.), α is an intercept, $\boldsymbol{z}_t$ is a vector of confounding covariates with associated coefficients $\boldsymbol{\gamma}$, $\boldsymbol{\theta}_{(1)} = (\theta_0, \dots, \theta_{M_1})'$ is the vector of first degree (linear) lagged effects, $\boldsymbol{\theta}_{(2)} = (\theta_{00}, \theta_{01}, \dots, \theta_{M_2M_2})'$ is the vector of second degree (quadratic) lagged effects, and (in an obvious extension of notation) $\boldsymbol{\theta}_{(i)}$ is the vector of i^{th}-degree lagged effects for $i = 1, \dots, \delta$. As with typical regression models, DL$^\delta$ models will rarely be considered for $\delta > 3$ because of the large increase in the number of parameters for such high degree DL models. However, because the HDDLM framework presented below extends to any degree $\delta \in \{1, 2, \dots\}$, general DL$^\delta$ models are considered in this section.

In traditional distributed lag models $\boldsymbol{\theta}_{(1)}$ is termed the "distributed lag function" and quantifies the *linear* relationship between Y_t and the lagged covariates $X_{t-\ell}$. Due to higher degree terms in (2.2), $\boldsymbol{\theta}^{(1)}$ will be referred to here as the "first degree distributed lag surface." Likewise, $\boldsymbol{\theta}_{(i)}$ will be referred to as the i^{th} degree distributed lag surface and quantifies the effect of i^{th} degree polynomials and i^{th} degree DL interactions on Y_t. That is, the first degree DL surface $\boldsymbol{\theta}_{(1)}$ represents main effects and $\boldsymbol{\theta}_{(i)}$ for $i > 1$ represent interaction effects. For example, θ_{01} represents the added effect on Y_t due to an interaction between X_t and X_{t-1}. We note that when $\delta > 1$, the interpretation of the coefficients $\{\boldsymbol{\theta}_{(i)}\}$ becomes muddled due to the presence of interactions. Because of this, rather than attempting to interpret specific coefficients, interpretation for DL$^\delta$ models should focus on the change in Y_t due to a change in the X's. For example, by what percent does the expected value of Y_t change if X_t changes by 1? This method of interpretation captures the cumulative effect (main effect and interactions) of a change in the X's on Y_t. See the example in Section 5 for further explanation and illustration.

The model in Equation (2.2) is able to capture non-linear relationships between Y and X. For example, assuming $\delta = 3$, the most simplistic form for (2.2) is to let $M_i = 0$ for $i = 1, \ldots, \delta$ leading to the (non-linear in X_t yet linear in the coefficients) model $g(\mathbb{E}(Y_t)) = \alpha + \boldsymbol{z}_t'\boldsymbol{\gamma} + \theta_0 X_t + \theta_{00} X_t^2 + \theta_{000} X_t^3$. The class of non-linear functions captured by DL$^\delta$ models are constrained to be polynomials. The distributed lag non-linear models of Gasparrini *and others* (2010) are able to capture more general non-linear lagged effects but they do not directly incorporate interactions between lagged effects as is done here.

When considering a modeling strategy for the coefficients in (2.2) above, several issues immediately come to the forefront. First, the covariates in (2.2) may exhibit collinearity which result in inflated variances (standard errors) of the parameter estimates. Second, the dimensionality of the parameter space grows quickly with M_i. For example, if $\delta = 3$ and $M_i = 5$ for $i = 1, \ldots, 3$ (a moderately small lag structure) then $P = 117$ coefficients would need to be estimated. For a more realistic lag structure of $M_i = 14$ for $i = 1, \ldots, 3$, $P = 966$. And, third, any prior knowledge regarding the distributed lag surfaces $\boldsymbol{\theta}_{(1)}, \ldots, \boldsymbol{\theta}_{(\delta)}$ is meager.

To develop a modeling strategy for the DL surfaces $\boldsymbol{\theta}_{(i)}$ that deals with these issues, note that $\boldsymbol{\theta}_{(i)}$ can be viewed as a surface over the set of points $\mathcal{L}_i = \{(\ell_1, \ldots, \ell_i)' : \ell_1 \leqslant \ell_2 \leqslant \cdots \leqslant \ell_i \leqslant M_i\}$. The indices of each element of $\boldsymbol{\theta}_{(i)}$ indicate the "location" of a point on a surface over $\mathcal{L}_i$. In a slight change of notation from above, let $\boldsymbol{\theta}_{(i)} = \{\theta_{\boldsymbol{\ell}_{ij}} : j = 1, \ldots, \dim(\boldsymbol{\theta}_{(i)})\}$ and $\boldsymbol{\ell}_{ij} \in \mathcal{L}_i$ for all j such that $\boldsymbol{\ell}_{ij}$ indicates the location of θ on $\mathcal{L}_i$. For example, if $\theta_{\boldsymbol{\ell}_{3j}} = \theta_{012}$ then $\boldsymbol{\ell}_{ij} = (0, 1, 2) \in \mathcal{L}_3$. From this viewpoint, modeling $\boldsymbol{\theta}_{(i)}$ is equivalent to modeling a nonlinear surface on $\mathcal{L}_i$. Gaussian processes (GPs) are a well suited tool for modeling non-linear surfaces (see Heaton and Peng, 2012). The appeal of a GP prior specification for $\boldsymbol{\theta}_{(i)}$ is the ability to flexibly fit a wide variety of surfaces while accounting for collinearity through the use of *a priori* correlation between the coefficients thereby enforcing smoothness in the DL surfaces and borrowing of information across lags to reduce standard errors. However, for the DL$^\delta$ models considered here, GP priors do not directly relieve the issue related to dimensionality of the parameter space in (2.2). In this regard, we propose the Gaussian predictive process of Banerjee *and others* (2008) as an elegant solution.

Consider a knot vector $\boldsymbol{\theta}^\star_{(i)} = \{\theta_{\boldsymbol{\ell}^\star_{ij}}\}$ such that $\dim(\boldsymbol{\theta}^\star_{(i)}) \leqslant \dim(\boldsymbol{\theta}_{(i)})$ where $\boldsymbol{\ell}^\star_{ij}$ denotes the lag "location" of the j^{th} knot on $\mathcal{L}_i$. Let $\boldsymbol{\theta}^\star_{(i)}$ follow a zero-mean Gaussian process with covariance function $\sigma_i^2\mathcal{M}_{\nu_i}(\cdot;\psi_i)$ where σ_i^2 is the common variance and $\mathcal{M}_\nu(\cdot;\psi)$ is the isotropic Matern correlation function with smoothness ν and decay parameter ψ. In other words, let $\boldsymbol{\theta}^\star_{(i)} \sim \mathcal{N}(\mathbf{0}, \mathbb{V}\mathrm{ar}(\boldsymbol{\theta}^\star_{(i)}))$ where $\mathbf{0}$ is the zero vector and $\mathbb{V}\mathrm{ar}(\boldsymbol{\theta}^\star_{(i)}) = \{\sigma_i^2\mathcal{M}_{\nu_i}(\|\boldsymbol{\ell}^\star_{ij} - \boldsymbol{\ell}^\star_{ik}\|;\psi_i)\}_{j,k}$ is the variance matrix. The model for $\boldsymbol{\theta}_{(i)}$ is given by the predictive process (Banerjee *and others*, 2008) interpolator,

$$\boldsymbol{\theta}_{(i)} = \mathbb{C}\mathrm{ov}(\boldsymbol{\theta}_{(i)}, \boldsymbol{\theta}^\star_{(i)})\mathbb{V}\mathrm{ar}^{-1}\left(\boldsymbol{\theta}^\star_{(i)}\right)\boldsymbol{\theta}^\star_{(i)} \tag{2.3}$$

where $\mathbb{C}\mathrm{ov}(\boldsymbol{\theta}_{(i)}, \boldsymbol{\theta}^\star_{(i)}) = \{\sigma_i^2\mathcal{M}_{\nu_i}(\|\boldsymbol{\ell}_{ij} - \boldsymbol{\ell}^\star_{ik}\|;\psi_i) : j = 1, \ldots, \dim(\boldsymbol{\theta}_{(i)}); k = 1, \ldots, \dim(\boldsymbol{\theta}^\star_{(i)})\}$ is the covariance of $\boldsymbol{\theta}_{(i)}$ and $\boldsymbol{\theta}^\star_{(i)}$ under the GP prior. As a brief aside, we note that other correlation functions could be used here but the Matérn class is the most common due to its flexibility.

Importantly, notice that the dimension of $\boldsymbol{\theta}_{(i)}$ in (2.3) is now $\dim(\boldsymbol{\theta}^\star_{(i)})$ because the knot vector $\boldsymbol{\theta}^\star_{(i)}$ completely determines the DL surface. Intuitively, the predictive process in (2.3) models $\boldsymbol{\theta}_{(i)}$ by a linear basis function expansion where the basis functions are given in the matrix $\mathbb{C}\mathrm{ov}(\boldsymbol{\theta}_{(i)}, \boldsymbol{\theta}^\star_{(i)})\mathbb{V}\mathrm{ar}^{-1}(\boldsymbol{\theta}^\star_{(i)})$ and the associated coefficients are represented by the knot vector $\boldsymbol{\theta}^\star_{(i)}$. This dimension reduction achieved by a basis function expansion is useful in DL modeling for a few reasons. First, and most obviously, the number of parameters required to estimate the i^{th} degree DL surface is reduced from $\dim(\boldsymbol{\theta}_{(i)})$ to $\dim(\boldsymbol{\theta}^\star_{(i)})$. And, second, in the presence of high collinearity among the X_t, the parameter estimates are correlated so as to borrow information across lags to stabilize estimation and reduce standard errors.

2.2 *Constraining the DL Surfaces*

Distributed lag surfaces are often subject to constraints. For example, when considering the first degree distributed lag surface $\boldsymbol{\theta}_{(1)} = (\theta_0, \ldots, \theta_{M_1})'$, a common assumption is for $\theta_j \to 0$ smoothly as $j \to M_1$. This is due to the physical intuition that $X_{t-\ell}$ for $\ell \gg 0$ should have a smaller effect on Y_t than $X_{t-\ell}$ for $\ell \approx 0$. By the same reasoning, the higher order distributed lag surfaces should decrease to zero as the lag time increases. That is, $\theta_{\boldsymbol{\ell}_{ij}} \to 0$ as $\max(\boldsymbol{\ell}_{ij}) \to M_i$ where $\max(\boldsymbol{\ell}_{ij})$ is the maximum element of $\boldsymbol{\ell}_{ij}$ (i.e. $\max(\boldsymbol{\ell}_{ij}) = \max\{\ell_{ij1}, \ldots, \ell_{iji}\}$).

To build the aforementioned constraints on DL surfaces into the model specification, consider introducing a set of lag times $L_i \ll M_i$ for $i = 1, \ldots, \delta$ such that $\theta_{\boldsymbol{\ell}_{ij}} = 0$ if $\max(\boldsymbol{\ell}_{ij}) > L_i$. Write $\boldsymbol{\theta}_{(i)} = (\boldsymbol{\theta}'_{(i;1)}, \boldsymbol{\theta}'_{(i;2)})'$ where $\boldsymbol{\theta}_{(i;1)} = \{\theta_{\boldsymbol{\ell}_{ij}} : \max(\boldsymbol{\ell}_{ij}) \leqslant L_i\}$ and $\boldsymbol{\theta}_{(i;2)} = \{\theta_{\boldsymbol{\ell}_{ij}} : \max(\boldsymbol{\ell}_{i;j}) > L_i\}$. Conditioning the model for $\boldsymbol{\theta}_{(i)}$ on L_i reduces to finding the conditional distribution $[\boldsymbol{\theta}_{(i;1)} \mid \boldsymbol{\theta}_{(i;2)} = \mathbf{0}]$. Using properties of the multivariate Gaussian distribution, the distribution $[\boldsymbol{\theta}_{(i;1)} \mid \boldsymbol{\theta}_{(i;2)} = \mathbf{0}]$ is Gaussian with mean $\mathbf{0}$ and $\mathbb{V}\text{ar}(\boldsymbol{\theta}_{(i;1)} \mid \boldsymbol{\theta}_{(i;2)} = \mathbf{0}) = \sigma_i^2(\boldsymbol{R}_{11} - \boldsymbol{R}_{12}\boldsymbol{R}_{22}^{-1}\boldsymbol{R}'_{12})$ where $\boldsymbol{R}_{11} = \mathbb{V}\text{ar}(\boldsymbol{\theta}_{(i;1)})$, $\boldsymbol{R}_{22} = \mathbb{V}\text{ar}(\boldsymbol{\theta}_{(i;2)})$ and $\boldsymbol{R}_{12} = \mathbb{C}\text{ov}(\boldsymbol{\theta}_{(i;1)}, \boldsymbol{\theta}_{(i;2)})$. Via this conditioning, the coefficients at higher lags are constrained (via a small prior variance) to be closer to zero than those at smaller lags. Further detail on this constraint is provided in Heaton and Peng (2012) and in the online supplementary materials.

Even though conditioning on L_i reduces the dimension of the i^{th} degree DL surface from $\dim(\boldsymbol{\theta}_{(i)})$ to $\dim(\boldsymbol{\theta}_{(i;1)})$, dimension reduction for $\boldsymbol{\theta}_{(i;1)}$ is still desired and may still be necessary to fit the DL$^\delta$ model. As such, let $\boldsymbol{\theta}_{(i;1)}$ be given by the predictive process interpolator *after* having conditioned on L_i; that is, let

$$\boldsymbol{\theta}_{(i;1)} = \mathbb{C}\text{ov}\left(\boldsymbol{\theta}_{(i;1)}, \boldsymbol{\theta}^\star_{(i)} \mid \boldsymbol{\theta}_{(i;2)} = \mathbf{0}\right) \mathbb{V}\text{ar}^{-1}\left(\boldsymbol{\theta}^\star_{(i)} \mid \boldsymbol{\theta}_{(i;2)} = \mathbf{0}\right) \boldsymbol{\theta}^\star_{(i)}. \tag{2.4}$$

Calculating the predictive process basis functions used in (2.4) is done by simple a three step process: (i) find the joint covariance matrix $\mathbb{V}\text{ar}((\boldsymbol{\theta}'_{(i;1)}, \boldsymbol{\theta}'^{\star}_{(i)}, \boldsymbol{\theta}'_{(i;2)})')$ using the Gaussian process prior, (ii) find $\mathbb{V}\text{ar}(\boldsymbol{\theta}_{(i;1)}, \boldsymbol{\theta}^\star_{(i)} \mid \boldsymbol{\theta}_{(i;2)} = \mathbf{0})$ by properties of the multivariate Gaussian distribution and (iii) calculate the basis function matrix $\mathbb{C}\text{ov}(\boldsymbol{\theta}_{(i;1)}, \boldsymbol{\theta}^\star_{(i)} \mid \boldsymbol{\theta}_{(i;2)} = \mathbf{0})\mathbb{V}\text{ar}^{-1}(\boldsymbol{\theta}^\star_{(i)} \mid \boldsymbol{\theta}_{(i;2)} = \mathbf{0})$. Notice also that by conditioning on $\boldsymbol{\theta}_{(i;2)} = \mathbf{0}$, the predictive process knot locations $\{\boldsymbol{\ell}^{\star}_{ik}\}$ need not be distributed over $\mathcal{L}_i = \{(\ell_1, \ldots, \ell_i)' : \ell_1 \leqslant \ell_2 \leqslant \cdots \leqslant \ell_i \leqslant M_i\}$. Rather, the knot locations need only be distributed over the subset of $\mathcal{L}_i$ given by $\{(\ell_1, \ldots, \ell_i)' : \ell_1 \leqslant \ell_2 \leqslant \cdots \leqslant \ell_i \leqslant L_i\}$.

To complete the model specification, prior distributions are required for the intercept α, the variance parameters $\{\sigma_i^2\}$ and the parameters of the Matérn correlation functions $\{\nu_i, \psi_i\}$. For α, assume a vague prior distribution where $\alpha \sim \mathcal{N}(0, s_\alpha^2)$ where s_α^2 is fixed at a "large" value. We let $\sigma_i^2 \sim \mathcal{IG}(a_\sigma, b_\sigma)$ where $\mathcal{IG}(a, b)$ denotes the inverse gamma distribution with shape parameter a and scale b (e.g. if $X \sim \mathcal{IG}(a, b)$ then $\mathbb{E}(X) = b/(a-1)$ for $a > 1$). For the studies in Section 4 and 5 below, $a_\sigma = 2$, $b_\sigma = 1$ and $s_\alpha^2 = 100^2$. The parameter ν_i controls the smoothness of the parameter surface $\boldsymbol{\theta}_{(i)}$. Estimating this smoothness parameter is a notoriously difficult problem even for observed spatial surfaces (Gneiting *and others*, 2012); hence, we fix $\nu_i = 3$ for all i which allows for the resulting DL surfaces to each be twice differentiable. In early stages of this work, we tried to estimate $\{\psi_i\}$ but found that $\{\psi_i\}$ was not identifiable (the prior and posterior were, for practical purposes, equal). This finding shouldn't be too surprising given work by Zhang (2004) who showed that for the isotropic Matérn class of covariance functions, weakly consistent estimators for σ_i^2 and ψ_i do not exist. The implication of this is that $\{\psi_i\}$ can be fixed *a priori* without sacrificing flexibility so long as σ_i^2 is assigned a vague prior (see also Du *and others*, 2009; Zhang and Wang, 2010). Hence, $\{\psi_i\}$ is treated as fixed *a priori* and a discussion of the choice of $\{\psi_i\}$ is deferred to Section 3.3.

3. Estimation

3.1 *Estimating the DL Surfaces*

Let $\boldsymbol{Y} = (Y_1, \ldots, Y_T)'$ denote a vector of response variables measured at T time periods. Assume, for the time being, that the set of lag times $\{L_i < M_i\}_{i=1}^{\delta}$ is known. Let $\boldsymbol{X}$ denote the $T \times \sum_i \dim(\boldsymbol{\theta}_{(i;1)})$ matrix of lagged explanatory variables and their interactions according to model (2.2). For example, the t^{th} row of $\boldsymbol{X}$

would be $(X_t, \ldots, X_{t-L_1}, X_t^2, X_t X_{t-1}, \ldots\ , X_{t-L_\delta}^\delta)$. From (2.2), the DL$^\delta$ model for $\boldsymbol{Y}$ is given by,

$$g\left(\mathbb{E}(\boldsymbol{Y})\right) = \alpha \mathbf{1}_T + \boldsymbol{Z}\boldsymbol{\gamma} + \boldsymbol{X}\boldsymbol{\theta} \tag{3.5}$$

where $\mathbf{1}_T$ is a length T vector of ones, $\boldsymbol{Z}$ is the design matrix of confounding variables and $\boldsymbol{\theta} = (\boldsymbol{\theta}'_{(1;1)}, \ldots, \boldsymbol{\theta}'_{(\delta;1)})'$ is the concatenated vector of distributed lag surfaces. By the predictive process specification for $\boldsymbol{\theta}$ in (2.4), $\boldsymbol{\theta} = \boldsymbol{B}\boldsymbol{\theta}^\star$ where $\boldsymbol{B}$ is block diagonal with i^{th} block $\mathbb{C}\text{ov}(\boldsymbol{\theta}_{(i;1)}, \boldsymbol{\theta}^\star_{(i)} \mid \boldsymbol{\theta}_{(i;2)} = \mathbf{0})\mathbb{V}\text{ar}^{-1}(\boldsymbol{\theta}^\star_{(i)} \mid \boldsymbol{\theta}_{(i;2)} = \mathbf{0})$ and $\boldsymbol{\theta}^\star$ is the concatenated vector of predictive process coefficients. Inserting $\boldsymbol{\theta} = \boldsymbol{B}\boldsymbol{\theta}^\star$ into (3.5) leads to,

$$g(\mathbb{E}(\boldsymbol{Y})) = \alpha \mathbf{1}_T + \boldsymbol{Z}\boldsymbol{\gamma} + \boldsymbol{D}\boldsymbol{\theta}^\star \tag{3.6}$$

where $\boldsymbol{D}$ is the $T \times \sum_i \dim(\boldsymbol{\theta}^\star_{(i)})$ design matrix for the DL$^\delta$ model. For brevity, let $P = \sum_i \dim(\boldsymbol{\theta}^\star_{(i)})$ be the total number of predictive process knots used to define the distributed lag surfaces such that $\boldsymbol{D}$ has dimension $T \times P$.

Conditional on the set of maximum lags $\{L_i\}$, a DL$^\delta$ model contains $P + \delta + 1$ parameters given by $\boldsymbol{\theta}^\star$, $\{\sigma_i^2\}_{i=1}^\delta$, and the common intercept α. Given that the DL$^\delta$ model in (3.6) is a generalized linear model, inference for $(\alpha, \boldsymbol{\theta}^\star, \{\sigma_i^2\})$ can be done in a straight forward fashion from a frequentist or Bayesian viewpoint. From the frequentist view, the maximum likelihood estimate $\hat{\boldsymbol{\theta}}^\star$ for $\boldsymbol{\theta}^\star$ is subject to the regularization criteria imposed by the priors $\boldsymbol{\theta}^\star_{(i)} \sim \mathcal{N}(\mathbf{0}, \mathbb{V}\text{ar}(\boldsymbol{\theta}^\star_{(i)} \mid \boldsymbol{\theta}_{(i;2)} = \mathbf{0}))$. From the Bayesian view, draws of $(\alpha, \boldsymbol{\theta}^\star, \{\sigma_i^2\})$ are obtained from the posterior using well-established Markov chain Monte Carlo (MCMC) techniques (see, e.g., Gamerman and Lopes, 2006).

3.2 *Estimating the Maximum Lags*

Traditionally, L_i is fixed *a priori* (see Welty and Zeger, 2005; Welty *and others*, 2009; Gasparrini *and others*, 2010). This approach may be effective for selecting L_1; however, little information, if any, is available for the maximum lag of higher order DL surfaces. Hence, selecting reasonable values for L_i where $i > 1$ is more difficult. In order to obtain a balance between estimating $\{L_i\}$ (which is computationally burdensome) and fixing $\{L_i\}$ (which may be inaccurate based on a lack of *a priori* information), this article proposes the following method for estimating $\{L_i\}$ based on maximum *a posteriori* (MAP) estimators. Let,

$$(\hat{\alpha}_{\{L_i\}}, \hat{\boldsymbol{\theta}}^{\star}_{\{L_i\}}, \{\hat{\sigma}^2_{i;\{L_i\}}\}) = \max_{(\alpha,\boldsymbol{\theta}^{\star},\{\sigma^2_i\})} \mathcal{LH}\left(\alpha, \boldsymbol{\theta}^{\star}, \{\sigma^2_i\} \mid \{L_i\}, \boldsymbol{Y}, \boldsymbol{X}\right) [\alpha] \left[\boldsymbol{\theta}^{\star} \mid \{\sigma^2_i\}\right] \prod_{i=1}^{\delta} \left[\sigma^2_i\right] \tag{3.7}$$

be the MAP estimator for $(\alpha, \boldsymbol{\theta}^{\star}, \{\sigma^2_i\})$ conditional on $\{L_i\}$ (hence, each is a function $\{L_i\}$), where $[\cdot]$ denotes a prior density function and $\mathcal{LH}(\cdot)$ denotes the likelihood function. For the studies in Sections 4 and 5 below, let $(L_1, \ldots, L_\delta) = (\hat{L}_1, \ldots, \hat{L}_\delta)$ where,

$$(\hat{L}_1, \ldots, \hat{L}_\delta) = \min_{(L_1,\ldots,L_\delta)} \text{AIC}(\hat{\alpha}_{\{L_i\}}, \hat{\boldsymbol{\theta}}^{\star}_{\{L_i\}}, \{\hat{\sigma}^2_{i;\{L_i\}}\}) \tag{3.8}$$

such that $(\hat{L}_1, \ldots, \hat{L}_\delta)$ are the values for $\{L_i\}$ that minimize the Akaike information criterion (AIC) evaluated at the MAP estimators.

By using MAP estimators, the prior constraints on each parameter (particularly the prior constraints for $\boldsymbol{\theta}^{\star}$) are accounted for in the minimization done in (3.8). Alternatively, other criterions such as BIC could be substituted into (3.8) but the important point is to evaluate the likelihood at the MAP estimates $(\hat{\alpha}_{\{L_i\}}, \hat{\boldsymbol{\theta}}^{\star}_{\{L_i\}}, \{\hat{\sigma}^2_{i;\{L_i\}}\})$ so that the prior constraints are appropriately accounted for.

3.3 *Decay Parameter and Knot Selection*

Because the primary function of inducing *a priori* correlation into DL$^\delta$ models is to control for collinearity by enforcing smoothness in the DL surface, the choice of ψ_i should be based on the amount of collinearity present in the $X's$. Let $\boldsymbol{X}_{(i)}$ be the columns of $\boldsymbol{X}$ associated with the i^{th} degree DL surface $\boldsymbol{\theta}_{(i;1)}$. From the model setup in Section 3.1, each column of $\boldsymbol{X}_{(i)}$ is associated with a lag location given by $\boldsymbol{\ell}_{ij}$ where $j = 1, \ldots, \dim(\boldsymbol{\theta}_{(i;1)})$. Hence, the empirical correlations between each column of $\boldsymbol{X}_{(i)}$ coupled with the distances $\|\boldsymbol{\ell}_{ij_1} - \boldsymbol{\ell}_{ij_2}\|$ for all $j_1 \neq j_2$ define an empirical variogram. The approach here is to choose ψ_i based on a fit to this empirical variogram. In this way, the *a priori* correlation for $\boldsymbol{\theta}_{(i;1)}$ is tied to the correlation in the $X's$. That is, if the $X's$ display a high degree of correlation then $\boldsymbol{\theta}_{(i;1)}$ will have high *a priori* correlation to counter-balance the collinearity. While this is an intuitive default specification for choosing $\{\psi_i\}$, theoretical and empirical studies by Zhang (2004), Du *and others* (2009) and Zhang and Wang (2010) indicate that fixing $\{\psi_i\}$ in this manner will not sacrifice the flexibility of the GP prior to fit a given surface even if fixed at the "wrong" values. We do note, however, choosing ψ_i based on the correlation in the X's is not always appropriate. That is, there may be situations where the autocorrelation in the X's is small yet smoothing the coefficients is still desired.

An important issue related to DL$^\delta$ models is choosing the knots $\{\boldsymbol{\ell}^\star_{ij}\}_{i,j}$ to specify the predictive process. Poor location of knots will lead to more error in the predictive process approximation of the parent process. For this reason, the knot locations $\{\boldsymbol{\ell}^\star_{ij}\}_j$ should be well dispersed over $\{(\ell_1,\dots,\ell_i)' : \ell_1 \leqslant \ell_2 \leqslant \cdots \leqslant \ell_i \leqslant L_i\}$ in order to learn about the i^{th} DL surface across this set. The strategy used here is to first select a surface-specific reduction factor $r_i \in [0,1)$ that reduces the dimension of $\boldsymbol{\theta}_{(i)}$ by $100 \times r_i\%$. Values of r_i near 0 will result in more knot locations (less dimension reduction) and, potentially, less error in recovering the DL surface but at the cost of computation and degrees of freedom. In contrast, values of r_i near 1 will have fewer knots (more dimension reduction) yet more error in estimating the DL surface. Given a choice of r_i, $\lceil \dim(\boldsymbol{\theta}_{(i)}) \times (1-r_i) \rceil$ knots are chosen (where $\lceil\cdot\rceil$ denotes the ceiling) using a space filling design over $\{(\ell_1,\dots,\ell_i)' : \ell_1 \leqslant \ell_2 \leqslant \cdots \leqslant \ell_i \leqslant L_i\}$. By using a space-filling design, we ensure the knot locations are well dispersed over the original domain. The choice for r_i is investigated further via simulation study in Section 4 to arrive at some guidance for choosing r_i.

4. Simulation Study

4.1 *Simulation Outline*

Twenty-five sets of 3^{rd} degree DL coefficients $\{\boldsymbol{\theta}_{(i)}\}_{i=1}^3$ were simulated independently according to (2.4) with no dimension reduction where $(L_1, L_2, L_3) = (6, 4, 2)$, $\sigma_i^2 = 0.10^2$ for all i and $\{\psi_i\}$ were fixed according to the methods outlined in Section 3 above. For each of the 25 DL models, 50 (1250 total) data sets of $n = 915$ observations were simulated from a Poisson distribution with mean given by (3.5) using a log link function. Values for σ_i^2, α and $\boldsymbol{X}$ were structured after the National Morbidity, Mortality and Air Pollution (NMMAPS) study for Dallas, TX. Specifically, $\sigma_i^2 = 0.10^2$ was chosen to align the simulated means within the range of observed number of deaths for the "older than 75" age group, α was fixed at the log of the mean number of deaths in the "older than 75" age group and $\boldsymbol{X}$ was constructed from summer (April-September) average daily temperatures in Dallas, TX between the years 2001-2005 (183 "summer" days over 5 years equates to $n = 915$ total days). Because each of the DL surfaces were simulated on the same scale ($\sigma_i^2 = 0.1^2$ for all i), the columns of $\boldsymbol{X}$ were centered and scaled.

Each simulated data set was fit using four different models for comparison: (A) a DL$^\delta$ model where the degree δ and maximum lags $\{L_i\}$ were treated as unknown and estimated from the data, (B) a DL$^\delta$ model where δ and (L_1, L_2, L_3) are assumed known, (C) a DL^1 model where the maximum lag L_1 was treated as unknown and estimated from the data and (D) an unconstrained maximum likelihood approach where δ and (L_1, L_2, L_3) are, again, treated as known. The primary reason for including model (D) is to highlight the importance of incorporating model constraints when estimating DL$^\delta$ models. For (A), δ and (L_1, L_2, L_3) were estimated by searching over all models of degree less than or equal to 3 and with maximum possible maximum lags of $(7, 5, 3)$, respectively (i.e. models with $L_1 > 7$ or $L_2 > 5$ or $L_3 > 3$ were not considered). This equated to searching over 387 total DL$^\delta$ models (9, 63, and 315 models of degree 1, 2, and 3, respectively) and choosing the one which minimized AIC evaluated at the MAP estimate. Likewise, for (C), L_1 was estimated by searching over all models where $L_1 < 8$. To investigate the effect of the dimension reduction factor r_i on model fit, (A), (B) and (C) were fit using a common reduction factor of $r_i = r \in \{0, 0.1, 0.2, 0.3, 0.6, 0.9\}$ where $r = 0$ is the ground "truth." For (A) and (C), the model selection was performed for each value of r. No dimension reduction was used for (D) as additional assumptions would be required.

4.2 *Simulation Results*

The four model fits are compared in terms of root mean square error (RMSE) of the posterior mean (or MLE), coverage (CVG) and width of a 95% credible (confidence) interval. For example, the RMSE for the

i^{th} DL surface using a dimension reduction factor of r (RMSE$_{ir}$) is calculated as,

$$\text{RMSE}_{ir} = \frac{1}{\dim(\boldsymbol{\theta}_{(i)})} \sum_{k=1}^{\dim(\boldsymbol{\theta}_{(i)})} \sqrt{\frac{\sum_{j=1}^{25}\sum_{s=1}^{50}(\theta_{(i,k)}^{j} - \hat{\theta}_{(i,k)}^{jsr})^2}{25 \times 50}}$$

where $\theta_{(i,k)}^{j}$ is the k^{th} parameter of the i^{th} DL surface in the j^{th} simulated model and $\hat{\theta}_{(i,k)}^{jsr}$ is the corresponding posterior mean (or MLE) from the s^{th} data set using a dimension reduction factor of r. Coverage of a 95% credible (or confidence) interval is calculated by simply looking at the proportion of 95% intervals which capture the true parameter. Interval width is calculated as the average distance between the upper and lower 95% interval limits.

Figure 1 displays RMSE, CVG, and width as a function of the dimension reduction factor separated by the DL surface (i.e. columns 1, 2, and 3 of Figure 1 correspond to results from the first, second, and third degree DL surface, respectively). The value in the upper right corner corresponds to the results from the unconstrained maximum likelihood fit (this is constant across r because no dimension reduction tool was used). Comparing models (A) and (B) to (D), clearly, fitting a constrained DL model improves performance. Specifically, the RMSE of the three DL surfaces was reduced by an average of 75% , 85% , and 93% under models (A) and (B) when using dimension reduction factors less than 0.5. Likewise, 95% credible interval widths for the three DL surfaces in models (A) and (B) were, on average, 71%, 90% and 95% shorter than the corresponding 95% confidence interval in model (D).

As evidenced by comparing the results from models (A) and (B) to (C), including higher order terms in the model is beneficial when the corresponding coefficients are non-zero. That is, the "best" DL^1 model had large RMSE and low CVG for the coefficients of the first degree DL surface when the true degree of the underlying model was greater than one.

Comparing the simulation results in Figure 1 between models (A) and (B), notice that estimating the degree (δ) and maximum lags ($\{L_i\}$), results in more error in recovering the true DL surfaces. This fact is apparent in that model (A) always had greater RMSE for dimension reduction factors less than 0.5. Greater error in model (A), however, is expected as more opportunity for error exists when treating δ and $\{L_i\}$ as unknowns.

Under model (A), a DL^3 model was correctly chosen 95% of the time suggesting the model selection method in Section 3.2 is useful in finding an appropriate degree for the HDDLM in this simulation setting. Yet, as displayed in Table 1, there was large variation in the maximum lag chosen for each DL surface. For example, according to Table 1, model (A) correctly estimated $L_2 = 4$ only 30% of the time. Perhaps alarmingly, model (A) chose the correct DL^3 model (i.e. the model with $(L_1, L_2, L_3) = (6, 4, 2)$) less than 5% of the time. This is due to the amount of noise present in the simulated datasets. To validate the model selection procedure described above, we performed a separate simulation study using Gaussian errors with very little noise (details of this simulation study are provided in the online supplementary materials). In the low noise setting, the model selection procedure found $(L_1, L_2, L_3) = (6, 4, 2)$ 100% of the time for reduction factors as high as $r = 0.3$. Higher values of r however, led to more error in model selection. Because this noisy simulation was built to mimic real data, we feel this noisy simulation is more realistic than the low noise setting in displaying how the HDDLM's perform in practice.

From Figure 1, model (A) had lower 95% credible interval CVG and width compared to when δ and $\{L_i\}$ were treated as known. We hypothesize that this lower coverage is due to the fact that we fit only the "best" model according to AIC. A full Bayesian analysis should treat δ and $\{L_i\}$ as parameters and average across model fits. We hypothesize that by model averaging, the model uncertainty would be reflected in the credible intervals by increasing credible interval widths such that coverage would be nearer to the nominal rate. However, model averaging in this setting is computationally demanding. Hence, there is a need to develop computationally feasible methods to appropriately account for model uncertainty in DL^δ models so as to have near nominal coverage rates. The development of such methods is beyond the scope of this article and left for future work.

For this simulation study, dimension reduction factors as high as 0.6 maintained a high performance in terms or RMSE and CVG while reducing the width of 95% credible intervals. For example, for the first degree DL surface, by using $r = 0.6$ compared to $r = 0$, the average 95% credible interval width was reduced by approximately 20% in model (B) while maintaining lower RMSE and a respectable 95% coverage rate of

92%. In Figure 1, the behavior of RMSE, CVG and WDTH are relatively stable for $r \in \{0, 0.1, 0.2, 0.3\}$ but less so for $r_i > 0.3$. This behavior suggest that, in practice, the value of r_i can be chosen by fitting the model for various values of r_i and plotting the estimates of the coefficients against r_i to look for large changes in the estimates. The value of r_i can then be chosen as the maximum reduction factor such that estimates are still stable for all values less than r_i.

5. NMMAPS Application

Mortality and temperature data for Los Angeles and New York were obtained from the National Morbidity, Mortality and Air Pollution (NMMAPS) study (Samet *and others*, 2000). Let Y_{ct} denote the mortality count for people over the age of 65 on day t in city c. The model used for this analysis is given by,

$$Y_{ct} \sim \mathcal{P}\left(\exp\left\{\alpha_c + \boldsymbol{x}'_{ct}\boldsymbol{\theta}_c + \boldsymbol{z}'_{ct}\boldsymbol{\gamma}_c\right\}\right) \tag{5.9}$$

where $\boldsymbol{x}_{ct}$ represents the vector of lagged average daily temperatures and their interactions as in model (2.2), α is an intercept, and $\boldsymbol{z}_{ct}$ is a vector of confounding covariates and includes a smooth function of time (specifically, a natural cubic spline with 5 degrees of freedom per year) and a day of the week effect. We use the HDDLM in Section 2 as a model for $\boldsymbol{\theta}_c$ and vague independent priors distributions were used for for α_c and $\boldsymbol{\gamma}_c$.

To choose the maximum lags, we considered all DL$^\delta$ models up to $\delta = 3$ with maximum lags up to and including $(L_1, L_2, L_3) = (10, 7, 7)$. Using the methods outlined in Section 3, Table 2 displays the top 5 DL$^\delta$ models for each city according to the AIC criterion evaluated at the MAP estimates and the "best" DL^1 model for comparison. From Table 2, notice that Los Angeles seems to have longer maximum lags than New York suggesting that "heat" effects are spread over a longer period of time in Los Angeles than in New York. In both New York and Los Angeles, however, interactions between lagged heat covariates are preferred. That is DL$^\delta$ models with $\delta > 1$ are preferred to DL^1 models. The preference for higher order interactions between lagged covariates is stronger in New York than it is in Los Angles. For New York, the "best" DL^1 model ranked 240th among the 804 considered models. For Los Angeles, the best DL^1 model ranked 4th suggesting that an additive model may be sufficient for Los Angeles but not for New York. As a final point, the models displayed in Table 2 are similar in terms of AIC suggesting that distinguishing a "best" model for this data set is difficult. However, when considering all 804 models, the spread in AIC is much larger than seen in Table 2 (from 4445 to 4536 for NY and from 4387 to 4461 for LA) suggesting the data is able to distinguish between models that fit well and those that fit poorly.

We estimated the preferred model from Table 2 using different values for a common dimension reduction factor r but found little difference in the posterior means of $\boldsymbol{\theta}_c$ for any $r \leqslant 0.5$ (although there were differences in the posterior standard deviation). The estimates of $\boldsymbol{\theta}_c$ were less stable using $r > 0.5$ so we chose to set $r = 0.5$ for this analysis to reduce the dimension and posterior standard deviations as much as possible.

Figure 2 displays the the first degree DL surfaces for New York and Los Angeles according to the "best" model in Table 2. Both New York and Los Angeles were found to have quite similar first degree DL surfaces with large effects occurring at more recent lags before dipping below zero at moderate lags and eventually tapering off to zero. These curves are consistent with previous studies which find a "displacement" effect of heat on mortality (Braga *and others*, 2002; Heaton and Peng, 2012). Figure 2 also displays the second degree DL surfaces for New York and Los Angeles. For New York, 95% credible intervals showed that the $(0, 1)$ and $(1, 1)$ coefficients were different from zero while for Los Angeles, the $(1, 1)$ effect was the only interaction effect different from zero. In both cities, the coefficient for X_{t-1}^2 was significantly positive showing an non-linear increase in mortality due to heat on the previous day. New York also had a significant positive coefficient for the interaction term X_tX_{t-1} suggesting that high heat on successive days increased mortality counts beyond what traditional DL models would suggest.

Due to the difficulty of interpreting higher degree DL surfaces, Figure 3 presents a more interpretable way of viewing the effect of lagged heat exposure on expected mortality. Figure 3 displays the posterior mean of the percentage change in expected mortality counts as a function of the deviation from a temperature of 75° F on days t and $t-1$ holding the temperature on days $t-3, t-4, \ldots$ constant and is a summary of the effect of all lagged effects on expected mortality. From Figure 3, the effect of including interactions between lagged covariates is apparent in that the contours of equal height are not straight lines. For example, in both cities the effect of temperature changes on mortality is non-linear and changes across the temperature domain.

6. Conclusions and Extensions

This article proposed higher degree DL models that extend the basic DL model to consider interactions between lagged covariates. The basic DL model is easily seen to be a special case of a higher degree DL model. Appropriate modeling constraints were imposed on the coefficients via a Gaussian process prior. Due to the potentially high dimension of these models, predictive processes were derived from the Gaussian process prior and used as a natural dimension reduction tool in this context. The usefulness of high degree DL models, along with the effectiveness of the dimension reduction strategy, were illustrated via simulation and in investigating the effect of high heat exposure on mortality in Los Angeles and New York using data from the NMMAPS study. We also proposed a method for selecting the degree and maximum lags using MAP estimators.

Beyond applied research, several new statistical research avenues for DL$^{\delta}$ models remain. Foremost is the need to explore model selection strategies for selecting the degree and maximum lags in DL$^{\delta}$ models. As was seen in the simulation studies in Section 4, choosing the maximum lag using the procedure outlined in this article was often difficult to do due to the noise in the data. Much of this difficulty is due to the inherent complex nature of DL$^{\delta}$ models. However, continued statistical work is needed to improve on model selection techniques.

In conjunction with improved model selection techniques, the DL$^{\delta}$ models used here assume a single maximum lag for each degree. For example, L_2 determines the maximum lag for second degree interactions. Under this assumption, if $L_2 = 6$ then the $(1, 6)$ lagged interaction is treated equally as the $(5, 6)$ interaction despite the fact that the $(5, 6)$ interaction is *a priori* more likely to be zero. A more realistic assumption is allow the maximum lag to change depending on the interaction. Additionally, assuming stationarity across lags, as is done implicitly here, is not realistic, That is, we expect larger differences between effects at smaller lags than larger lags which suggests non-stationarity in modeling the DL coefficients. These extensions are left for future work.

Finally, note that the DL$^{\delta}$ models presented here are, in reality, only univariate because they rely on a single lagged covariate. Multiple distributed lag models have also been used but typically do not include interactions between lag and variables. Extending the DL$^{\delta}$ models here to multiple-lagged covariates is currently under investigation.

Funding

The project described was supported by the National Institute of Environmental Health Sciences [contract awards R21ES020152 and R01ES019560]. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institute of Environmental Health Sciences or the National Institutes of Health. The first author was partially supported by National Aeronautics and Space Administration [contract award NNX10AK79G].

Acknowledgements

The National Center for Atmospheric Research is managed by the University Corporation for Atmospheric Research under the sponsorship of the National Science Foundation.

References

Anderson, B. G. and Bell, M. L. (2009). Weather-related mortality: How heat, cold, and heat waves affect mortality in the united states. *Epidemiology* **20**, 205–213.

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

Braga, A. L., Zanobetti, A. and Schwartz, J. (2002). The effect of weather on respiratory and cardiovascular deaths in 12 u.s. cities. *Environmental Health Perspectives* **110**(9), 859–863.

Caffo, B. S., Peng, R. D., Dominici, F., Louis, T. A. and Zeger, S. L. (2011). Parallel MCMC Imputation for Multiple Distributed Lag Models: A Case Study in Environmental Epidemiology. In: *The Handbook of Markov chain Monte Carlo*. Chapman and Hall/CRC Press, pp. 493–510.

Du, Juan, Zhang, Hao and Mandrekar, V. S. (2009). Fixed-Domain Asymptotic Properties of Tapered Maximum Likelihood Estimators. *The Annals of Statistics* **37**(6A), 3330–3361.

Gamerman, D. and Lopes, H. F. (2006). *Markov Chain Monte Carlo: Stochastic Simulation for Bayesian Inference*, 2 edition. Boca Raton, FL: Chapman and Hall/CRC.

Gasparrini, A., Armstrong, B. and Kenward, M. G. (2010). Distributed Lag Non-Linear Models. *Statistics in Medicine* **29**, 2224–2234.

Gneiting, T., Ševčíková, H. and Percival, D. B. (2012). Estimators of Fractal Dimension: Assessing the Roughness of Time Series and Spatial Data. *Statistical Science* **27**(2), 247–277.

Hajat, S., Armstrong, B. G., Gouveia, N. and Wilkinson, P. (2005). Mortality displacement of heat-related deaths. *Epidemiology* **16**, 613–620.

Heaton, Matthew J. and Peng, Roger D. (2012). Flexible Distributed Lag Models using Random Functions with Application to Estimating Mortality Displacement from Heat-Related Deaths. *Journal of Agricultural, Biological, and Environmental Statistics* **17**(3), 313–331.

IPCC. (2007). *Climate Change 2007: Impacts, Adaptation and Vulnerability. Contribution of Working Group II to the fourth assessment report of the Intergovernmental Panel on Climate Change*. Cambridge University Press.

Kovats, R. S. and Hajat, S. (2008). Heat Stress and Public Health: A Critical Review. *Annual Review of Public Health* **29**, 41–55.

Li, B., Sain, S., Mearns, L. O., Anderson, H. A., Kovats, S., Ebi, K. L., Bekkedal, M., Kanarek, M. S. and Patz, J. A. (2011). The Impact of Extreme Heat on Morbidity in Milwaukee, Wisconsin. *Climatic Change* **110**, 959–976. DOI: 10.1007/s10584-011-0120-y.

O'Neill, M. S., Zanobetti, A. and Schwartz, J. (2003). Modifiers of the temperature and mortality association in seven us cities. *American Journal of Epidemiology* **157**, 1074–1082.

Peng, R. D., Bobb, J. F., Tebaldi, C., McDaniel, L., Bell, M. L. and Dominici, F. (2011). Toward a quantitative estimate of future heat wave mortality under global climate change. *Environmental Health Perspectives* **119**, 701–706.

Samet, J. M., Zeger, S. L., Dominici, F., Curriero, F., Coursac, I., Dockery, D. W., Schwartz, J. and Zanobetti, A. (2000). The National morbidity, Mortality, and Air Pollution Study Part II: morbidity and mortality from air pollution in the united states. *Research Report Health Effects Institute* **94**, 5–79.

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

Welty, L. J., Peng, R. D., Zeger, S. L. and Dominici, F. (2009). Bayesian distributed lag models: Estimating the effects of particulate matter air pollution on daily mortality. *Biometrics* **65**, 282–291.

Welty, L. J. and Zeger, S. L. (2005). 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. *Americal Journal of Epidemiology* **162**, 80–88.

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

Zhang, H. and Wang, Y. (2010). Kriging and Cross-Validation for Massive Spatial Data. *Environmetrics* **21**, 290–304.

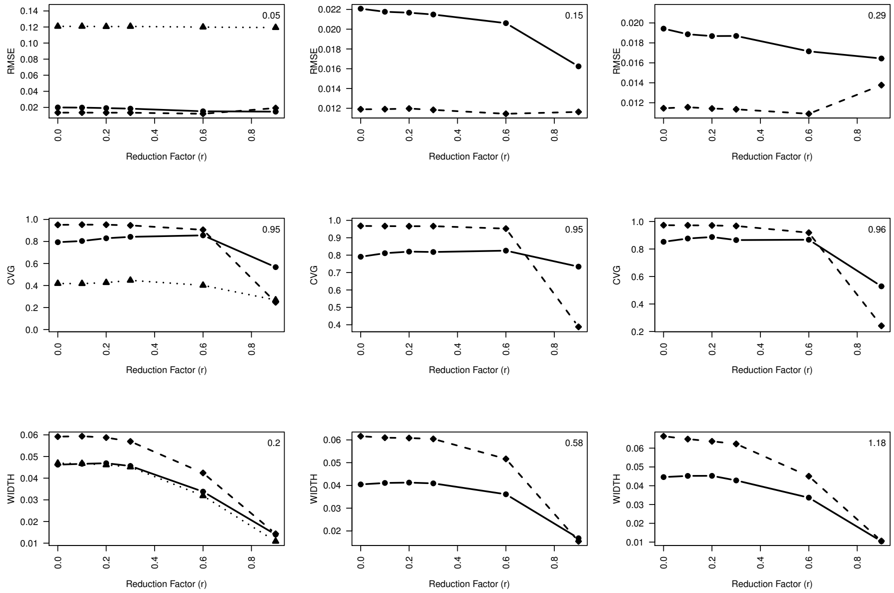

Fig. 1. Simulations results as a function of the dimension reduction factor (r). The solid, dashed and dotted lines correspond to the simulations results when estimating the degree (model A), when treating it as fixed (model B) and when fitting the best DL[1] model (model C), respectively. The number in the top right corner corresponds to the unconstrained maximum likelihood fit (model D). The first, second, and third column of plots correspond to results from the first, second, and third degree DL surface, respectively. Results from model C are excluded from the second and third columns because only the first degree surface was estimated.

Table 1. *Proportion of time $L_i = \ell$ for a given maximum lag ℓ under model (A) in the simulation study. "NA" stands for "Not Applicable" and indicates that models with $L_i = \ell$ were not considered in the explored model space. As an example, only 30% of the time model (A) correctly chose $L_2 = 4$.*

Maximum Lag	L_1	L_2	L_3
0	0.12	0.17	0.14
1	0.08	0.08	0.30
2	0.09	0.12	0.41
3	0.12	0.25	0.15
4	0.17	0.30	NA
5	0.15	0.08	NA
6	0.19	NA	NA
7	0.07	NA	NA

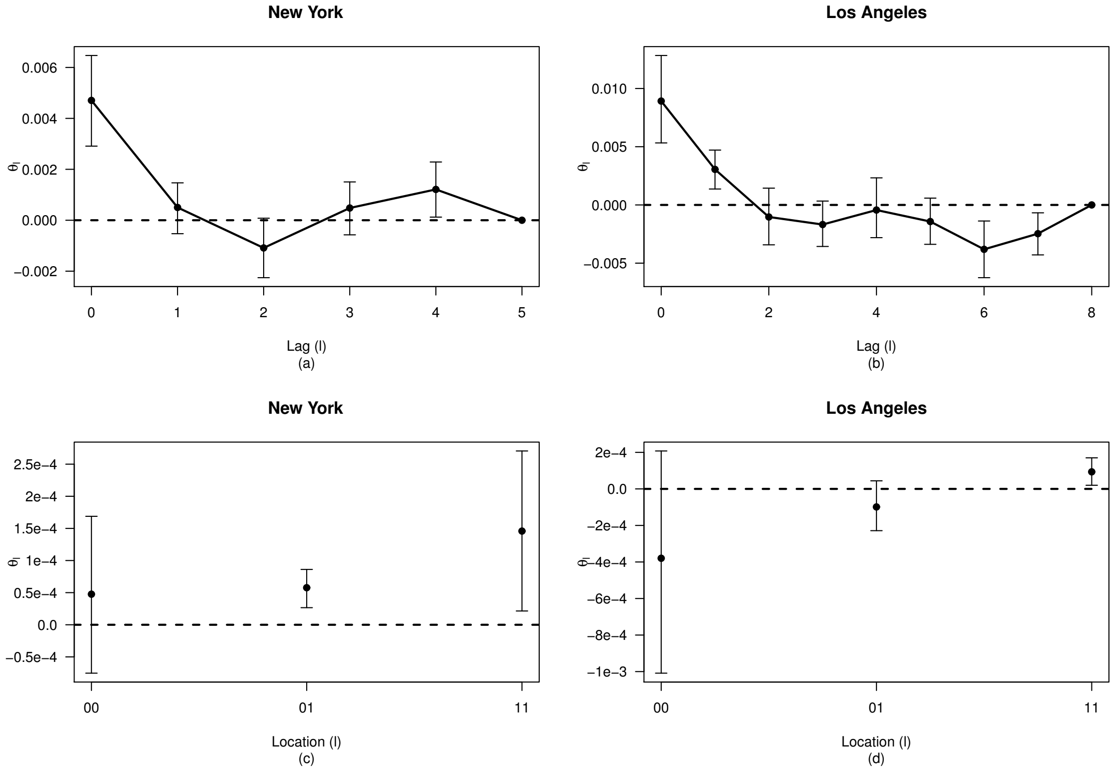

Fig. 2. Estimate first degree DL surfaces for (a) New York and (b) Los Angeles. Estimated second degree DL coefficients for (a) New York and (b) Los Angeles. The solid points correspond to the posterior mean while the error bars represent a 95% central credible interval.

Table 2. *Top 5 DL$^\delta$ models in the NMMAPS analysis of New York and Los Angeles mortality counts.*

	New York		Los Angeles	
Rank	$\{L_i\}$	AIC(MAP)	$\{L_i\}$	AIC(MAP)
1	$\{4, 1\}$	4445.48	$\{7, 1\}$	4387.12
2	$\{4, 3\}$	4445.81	$\{7, 0, 0\}$	4388.46
3	$\{5, 1\}$	4445.85	$\{6, 0\}$	4388.65
4	$\{3, 1\}$	4446.03	$\{8\}$	4388.84
5	$\{4, 1, 1\}$	4446.15	$\{7, 0\}$	4388.87
Best DL^1 Model	$\{4\}$	4458.24	$\{8\}$	4388.84

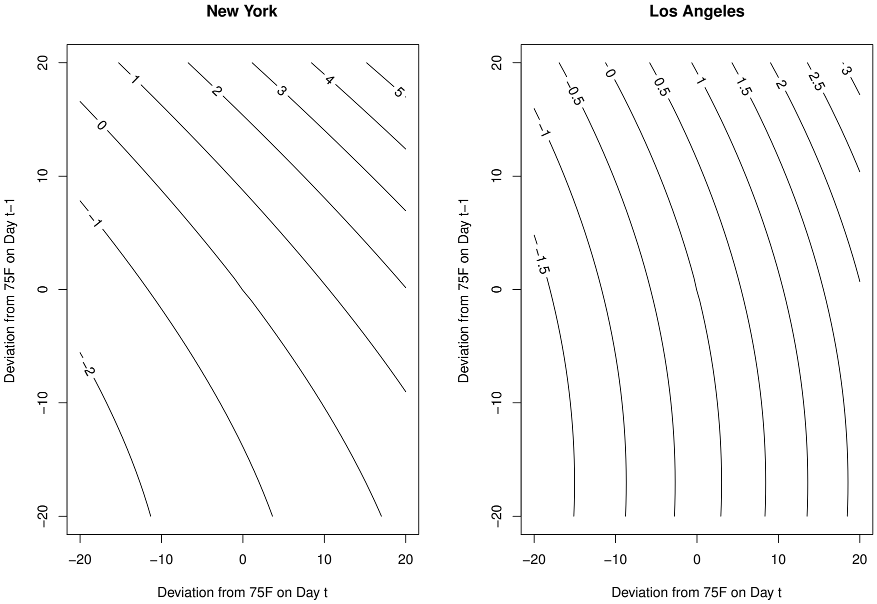

Fig. 3. Percent change in mortality as a function of the deviation from a temperature of 75° F on days t and $t-1$. Allowing for interactions between lagged covariates models the curvature of contour lines.