20

Parallel Bayesian MCMC Imputation for Multiple Distributed Lag Models: A Case Study in Environmental Epidemiology

Brian Caffo, Roger Peng, Francesca Dominici, Thomas A. Louis, and Scott Zeger

20.1 Introduction

Patterned missing covariate data is a challenging issue in environmental epidemiology. For example, particulate matter measures of air pollution are often collected only every third day or every sixth day, while morbidity and mortality outcomes are collected daily. In this setting, many desirable models cannot be directly fit. We investigate such a setting in so-called "distributed lag" models when the lagged predictor is collected on a cruder time scale than the response. In multi-site studies with complete predictor data at some sites, multilevel models can be used to inform imputation for the sites with missing data.

We focus on the implementation of such multilevel models, in terms of both model development and computational implementation of the sampler. Specifically, we parallelize single chain runs of sampler. This is of note, since the Markovian structure of Markov chain Monte Carlo (MCMC) samplers typically makes effective parallelization of single chains difficult. However, the conditional independence relationships of our developed model allow us to exploit parallel computing to run the chain. As a first attempt at using parallel MCMC for Bayesian imputation on such data, this chapter largely represents a proof of principle, though we demonstrate some promising potential for the methodology. Specifically, the methodology results in proportional decreases in run-time over the nonparallelized version near one over the number of available nodes.

In addition, we describe a novel software implementation of parallelization that is uniquely suited to disk-based shared memory systems. We use a "blackboard" parallel computing scheme where shared network storage is a used as a blackboard to tally currently completed and queued tasks. This strategy allows for easy addition and subtraction of compute nodes and control of load balancing. Moreover, it builds in automatic checkpointing.

Our investigation is motivated by multi-site time series studies of the short-term effects of air pollution on disease or death rates. A common measure of air pollution used for such studies is the amount in micrograms per cubic meter of particulate matter of a specified maximum aerodynamic diameter. We focus on $PM_{2.5}$ (see Samet et al., 2000). Unfortunately, the definitive source of particulate matter data in the United States, the Environmental Protection Agency's air pollution network of monitoring stations, collects data only a few times per week at some locations. One of the most frequent observed data patterns for

$PM_{2.5}$ is data being recorded every third day. However, the disease rates that we consider are collected daily.

In this setting, directly fitting a model that includes several lags of $PM_{2.5}$ simultaneously is not possible. Such models are useful, for example, to investigate a cumulative weekly effect of air pollution on health. They are also useful to more finely investigate the dynamics of the relationship between the exposure and response. As an example, one might postulate that after an increase in air pollution, high air pollution levels on later days may have a smaller impact, as the risk set has been depleted from the initial increase (Dominici et al., 2002; Schwartz, 2000; Zeger et al., 1999).

We focus on distributed lag models that relate the current-day disease rate to particulate matter levels over the past week. That is, our model includes the current day's $PM_{2.5}$ levels as well as the previous six days. While direct estimation of the effect for any particular lag is possible, joint estimation of the distributed lag model is not possible (see Section 20.3). Moreover, missing-data imputation for counties with patterned missing data is difficult. We consider a situation where several independent time series are observed at different geographical regions, some with complete $PM_{2.5}$ data. We use multilevel models to borrow information across series to fill in the missing data via Bayesian imputation. The hierarchical model is also used to combine county-specific distributed lag effects into national estimates.

The rest of the chapter is organized as follows. In Section 20.2 we outline the data set used for analysis and follow in Section 20.3 with a discussion of Bayesian imputation. In Section 20.4 we describe the distributed lag models of interest, and in Section 20.5 we illustrate a multiple imputation strategy. Section 20.6 uses the imputation algorithm to analyze hospitalization rates of chronic obstructive pulmnonary disease (COPD). Finally, Section 20.7 gives some conclusions, discussion and proposals for future work.

20.2 The Data Set

The Johns Hopkins Environmental Biostatistics and Epidemiology Group has assembled a national database comprising time series data on daily hospital admission rates for respiratory outcomes, fine particles ($PM_{2.5}$), and weather variables for the 206 largest US counties having a population larger than 200,000 and with at least one full year of $PM_{2.5}$ data available. The study population, derived from Medicare claims, includes 21 million adults older than 65 with a place of residence in one of the 206 counties included in the study.

Daily counts of hospital admissions and daily number of people enrolled in the cohort are constructed from the Medicare National Claims History Files. These counts are obtained from billing claims of Medicare enrollees residing in the 206 counties. Each billing claim contains the following information: date of service, treatment, disease (ICD 9 codes), age, gender, race and place of residence (zip and county).

Air pollution data for fine particles are collected and posted by the United States Environmental Protection Agency Aerometric Information Retrieval Service (AIRS, now called the Air Quality System, AQS). To protect against outlying observations, a 10% trimmed mean is used to average across monitors after correction for yearly averages for each monitor. Specifically, after removing a smoothly varying annual trend from each monitor time series, the trimmed mean was computed using the deviations from this smooth trend. Weather data is

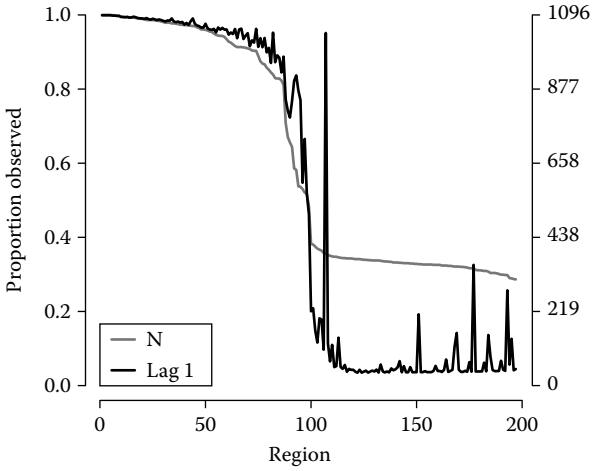

FIGURE 20.1
Summary of the missing-data pattern. The gray line displays the proportion of the total days in the study with observed $PM_{2.5}$ data for each county, with the actual number of days displayed on the right axis. The black line shows the proportion and count of the days with observed air pollution data where the lag-1 day is also observed.

obtained from the National Weather Monitoring Network which comprises daily temperature and daily dew points temperature for approximately 8000 monitoring stations in the USA. We aggregate data across monitors to obtain temperature time series data for each of the 206 counties, of which 196 were used in analysis. Details about aggregation algorithms for the air pollution and weather are posted at http://www.biostat.jhsph.edu/MCAPS and further information about data collection is given in Dominici et al. (2006).

Figure 20.1 illustrates the salient features of the missing-data pattern for $PM_{2.5}$ in this database. This study considered 1096 monitoring days. Figure 20.1 displays the proportion of the 1096 days with observed $PM_{2.5}$ data for each county (dark gray line). The associated number of observed days is displayed on the right scale. This figure also displays the proportion of 1096 days with observed $PM_{2.5}$ data where the lag-1 day was also observed (black line).

The plots show that nearly half of the 196 counties have measurements on roughly one third of the total possible days. Ninety-six of these counties have over 40% of the air pollution data observed and enough instances of seven consecutive observed $PM_{2.5}$ days to estimate the desired distributed lag model (see Section 20.4). For these counties, any missing data is often due to a large contiguous block, for example, several weeks where the monitor malfunctioned. Such uninformative missing data leaves ample daily measurements to estimate distributed lag models, so is ignored in our model. The remaining counties have $PM_{2.5}$ data collected every third day and possibly also have blocks of missing data, hence have data on less than 33% of the days under study. Because of the systematically missing $PM_{2.5}$ data in these counties, there is little hope of fitting a distributed lag model without borrowing information on the exposure-response curve from daily time series data from other counties. The plot further highlights this by showing that direct estimates of the lag-1 autocovariances are not available for roughly half of the counties. However, because of the missing-data pattern, all of the counties have direct estimates of the lag-3 autocovariances.

20.3 Bayesian Imputation

In this section, we discuss the relative merits of Bayesian imputation. We focus on our particular missing-data problem, and refer the reader to Carlin and Louis (2009) and Little and Rubin (2002) for general introductions to Bayesian statistics, missing data, and computation. We argue that imputation for systematic missingness in the predictor time series is relevant for distributed lag models, and particularly for the data set in question, while it is less relevant for single-lag models. In this section, we restrict our discussion to the consideration of a single outcome time series, say Y_t, and single predictor time series, X_t. For context, consider the outcome to be the natural log of the county-specific Medicare emergency admissions rate for COPD, and the predictor to be $PM_{2.5}$ levels for that county. To make this thought experiment more realistic, let Y_t and X_t be the residual time series obtained after having regressed out relevant confounding variables. We assume that the $\{Y_t\}$ are completely observed and the $\{X_t\}$ are observed only every third day, so that $X_0, X_3, X_6, \ldots$ are recorded; and we evaluate whether or not to impute the missing predictors. In our subsequent analysis of the data, we will treat this problem more formally using Poisson regression.

20.3.1 Single-Lag Models

A single-lag model relates the Y_t to X_{t-u} for some $u = 0, 1, 2, \ldots$ via the mean model $\mathrm{E}[Y_t] = \theta_u X_{t-u}$, when an identity link function is used. We argue that, for any such single-lag model, implementing imputation strategies for the missing predictor values is unnecessary. Consider that direct evidence regarding any single-lag model is available in the form of simple lagged cross-correlations. For example, the pairs $(Y_0, X_0), (Y_3, X_3), (Y_6, X_6), \ldots$ provide direct evidence for $u = 0$; the pairs $(Y_1, X_0), (Y_4, X_3), (Y_7, X_6), \ldots$ provide direct evidence for $u = 1$ and so on. Imputing the missing predictors only serves to inject unneeded assumptions. Furthermore, there is a tradeoff where more variation in the predictor series benefits the model's ability to estimate the associated parameter, yet hampers the ability to impute informatively. Hence, in the typical cases where the natural variation in the predictor series is large enough to be of interest, we suggest that imputing systematically missing predictor data for single-lag models is not worth the trouble. In less desirable situations with low variation in the predictor series, imputation for single-lag models may be of use.

20.3.2 Distributed Lag Models

Now consider a distributed lag model, such as $\mathrm{E}[Y_t] = \sum_{u=0}^{d} \theta_u X_{t-u}$. Here, if a county has predictor data recorded every third day, there is no direct information to estimate this relationship. Specifically, let $t = 0, \ldots, T-1$ and $\mathbf{D}$ be the design matrix associated with the distributed lag model and Y be the vector of responses. Then the least squares estimates of the coefficients are $(\frac{1}{T}\mathbf{D}^t\mathbf{D})^{-1}\frac{1}{T}\mathbf{D}^t Y$. The off-diagonal terms of $\frac{1}{T}\mathbf{D}^t\mathbf{D}$ contain the lagged autocovariances in the $\{X_t\}$ series; $\frac{1}{T}\mathbf{D}^t Y$ contains the lagged cross-covariances between the $\{Y_t\}$ and $\{X_t\}$. As was previously noted, these lagged cross-covariances are directly estimable, even with patterned missing data in the predictor series. In contrast, the autocovariances in the predictor series are only directly estimable for lags that are multiples of 3. Thus, without addressing the missing predictor data, the distributed lag model cannot be fit. For our data, this would eliminate information from nearly 50% of the counties studied. Hence, a study of the utility of predicting the missing data is warranted.

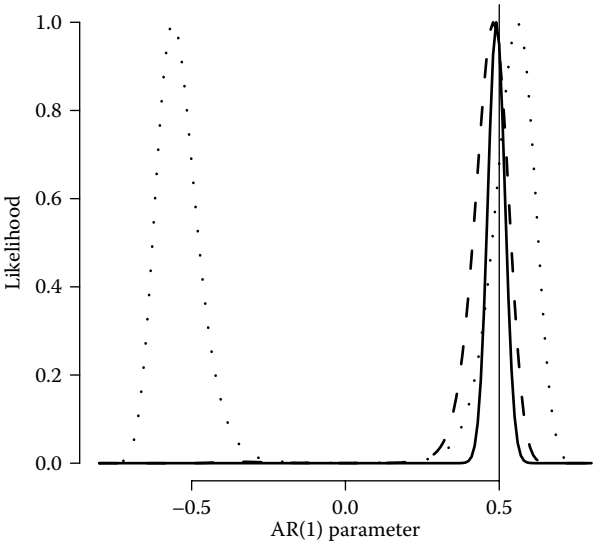

FIGURE 20.2
Likelihood for the AR(1) coefficient for AR(1) simulated data with an assumed correctly known innovation variance of 1 and coefficient of 0.5, for three missing-data patterns: completely observed (solid), observed only every three days (dashed) and observed only every six days (dotted). A solid vertical line denotes the actual coefficient value of 0.2.

One might consider using a model, such as an AR(p), to extrapolate the missing autocorrelations. However, a single time series with this degree of systematic missingness may not have enough information to estimate the parameters. Consider an AR(1) process. To illustrate, Figure 20.2 shows the likelihood for the AR(1) coefficient for data simulated under an AR(1) model with a correctly known innovation variance of 1 and data observed every day (solid), every third day (dashed) and every sixth day (dotted). Any inference for the AR(1) parameter (at 0.5, depicted with a horizontal line) would be imprecise with the systematically missing data. For the data observed every sixth day, notice that the likelihood is multimodal and symmetric about zero. This is because the likelihood only depends on the AR coefficient raised to even powers. This poses a problem even for our every-third-day data, because additional missing observations create patterns of data collected only every sixth day. Such multimodal likelihoods for AR models are described in Wallin and Isaksson (2002) (also see Broersen et al., 2004). In addition, here we assume that the correct model is known exactly, which is unlikely to be true in practice.

In our data set, there is important information in the counties with completely observed data that can be used to help choose models for the predictor time series and estimate parameters. Figure 20.3 demonstrates such model fits with an AR(4) model applied to a detrended version of the log of the $PM_{2.5}$ process. This plot is informative because when the data are available only once every third day, it is not possible to estimate the autocorrelation function using data only from that county. However, it also illustrates that the population distribution of autoregressive parameters appears to be well defined by the counties with mostly observed data. Figure 20.4 shows the estimated residual standard deviations from these model fits, suggesting that these are well estimated even if the autoregressive parameters are not. The data suggest that an AR(1) process is perhaps sufficient, though we continue to focus on AR(4) models to highlight salient points regarding imputation.

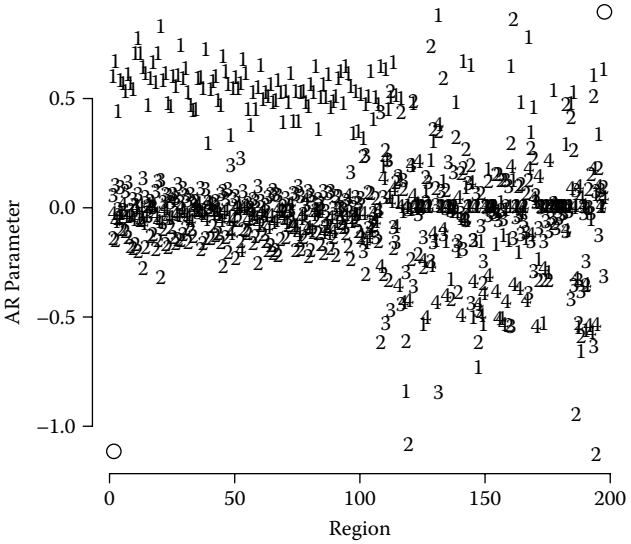

FIGURE 20.3
Estimated AR(4) coefficients, labeled "1" to "4," by counties ordered by decreasing percentage of observed data from left to right. Roughly the first 100 counties have substantial consecutively observed data to estimate the AR parameters while the remaining 100 do not.

20.4 Model and Notation

In this section, we present notation and modeling assumptions. A summary of the most important parameters and hyperparameters is given in Table 20.1. Let Y_{ct}, for county $c = 0, \ldots, C-1$ and day $t = 0, \ldots, T-1$, denote a response time series of counts, such

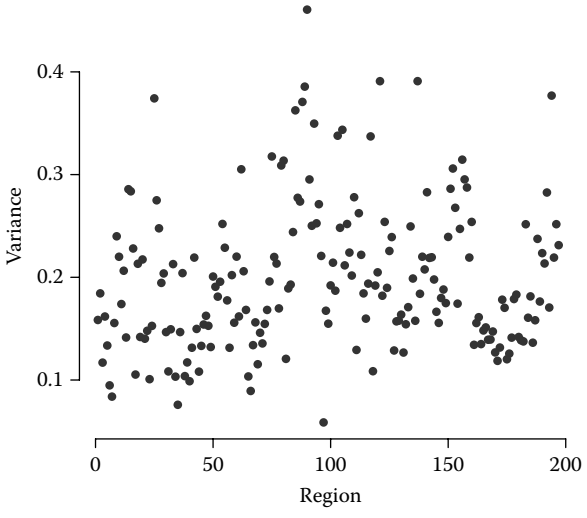

FIGURE 20.4
Variances from individual autoregressive time series by counties ordered by decreasing percentage of observed data from left to right.

TABLE 20.1

Parameters and their Definitions

Y_{ct}	Response count for county c at time t
R_{ct}	Size of the risk set for county c at time t
λ_{ct}	County-specific expected rate for Y_{ct}
X_{ct}	Log $PM_{2.5}$ level in micrograms per cubic meter for county c at time t
θ_{cu}	Distributed lag parameters for county c
θ^*_{cu}	Constrained distributed lag parameters for county c
$\theta^\dagger_{cu}$	Constrained and reparameterized distributed lag parameters for county c
$\tilde{\ell}(\boldsymbol{\theta}_c)$	Profile log likelihood for $\boldsymbol{\theta}_c$
$\mathbf{W}_{ct}$, $\boldsymbol{\psi}_c$	Slowly varying trend model on the log $PM_{2.5}$ series $\log(X_{ct}) = \mathbf{W}_{ct}\boldsymbol{\psi}_c + \epsilon_{ct}$
ϵ_{ct}	Residuals from above model, where we presume $\boldsymbol{\psi}_c$ is known and fixed at the estimated value; these terms are the actually imputed terms
$\boldsymbol{\mu}$	Inter-county mean of the $\boldsymbol{\theta}_c$
$\boldsymbol{\Sigma}_\theta$	Inter-county variance–covariance matrix of the $\boldsymbol{\theta}_c$
$\boldsymbol{\alpha}_c$	County-specific autoregressive parameters on the ϵ_{ct}
$\boldsymbol{\zeta}$	Inter-county mean of the $\boldsymbol{\alpha}_c$
$\boldsymbol{\Sigma}_\alpha$	Inter-county variance–covariance matrix of the $\boldsymbol{\alpha}_c$

as the daily incident counts of COPD. Let R_{ct} denote the number of persons in county c at risk for disease on day t. We assume that both of these processes are completely observed. Let $\lambda_{ct} = \mathrm{E}[Y_{ct} \mid R_{ct}, X_{ct}, \boldsymbol{\theta}_c, \boldsymbol{\beta}_c]$ be the daily mean, X_{ct} denote $PM_{2.5}$, and Z_{ctj}, for $j = 0, \ldots, J-1$, be other covariates of interest, which we assume are completely observed.

We assume that the response process follows a Poisson law with means satisfying:

$$\log(\lambda_{ct}/R_{ct}) = \sum_{u=0}^{d} \theta_{cu} X_{c,t-u} + \sum_{j=0}^{J-1} Z_{ctj}\beta_{cj}.$$

For the hospital admissions data, we consider $d = 6$. That is, the model relates the prior week's air pollution to the current day's mean disease rate.

The sum of the lagged air pollution parameters, labeled the distributed lag "total" or "cumulative" effect, $\sum_{u=0}^{d} \theta_{cu}$, is a parameter of primary interest. The total effect is the change in the log mean response rate given a one-unit across-the-board increase in the $PM_{2.5}$ over the current and prior d days.

To mitigate the variance inflation incurred by including many obviously collinear lagged covariates, sometimes a functional form is placed on the θ_{cu}, especially if d is large (see Zanobetti et al., 2000). A particularly effective approach is to assume that $\theta_{cu} = \mathbf{A}_u\boldsymbol{\theta}^*_c$, where $\mathbf{A}_u$ is column u from a smoothing design matrix, say $\mathbf{A}$, on the time points $0, \ldots, d$. When $d = 6$, choosing $\mathbf{A}$ to be a bin smoother with bins for the current day, days lag 1 to 2, and days lag 3 to 6, has been shown to be a useful approach in the air pollution time series literature (Bell et al., 2004). Such a smoother requires three parameters, $\boldsymbol{\theta}^*_c = (\theta^*_{c1}, \theta^*_{c2}, \theta^*_{c3})$, so that the model becomes

$$\log(\lambda_{ctr}/R_{ctr}) = \theta^*_{c1}x_{ct} + \theta^*_{c2}\sum_{u=1}^{2} X_{c,t-u} + \theta^*_{c3}\sum_{u=3}^{6} X_{c,t-u} + \sum_{j=0}^{k} z_{ctj}\beta_{cjr}.$$

This restriction on the distibuted lag parameters is equivalent to a convenient form for the rate model that considers the seven-day average air pollution and its deviation from

the three-day average and current day:

$$\log(\lambda_{ctr}/R_{ctr}) = \theta^{\dagger}_{c1}\bar{x}^{(6)}_{ct} + \theta^{\dagger}_{c2}(\bar{x}^{(2)}_{ct} - \bar{x}^{(6)}_{ct}) + \theta^{\dagger}_{c3}(x_{ct} - \bar{x}^{(2)}_{ct}) + \sum_{j=0}^{k} z_{ctj}\beta_{cjr}. \tag{20.1}$$

Here $\bar{x}^{(k)}_{ct}$ is the average of the current-day and k previous days' $PM_{2.5}$ values. These parameters are related to the θ^{*}_{i} via the equalities

$$\begin{aligned}
\theta^{*}_{c1} &= \frac{1}{7}\theta^{\dagger}_{c1} + \frac{4}{21}\theta^{\dagger}_{c2} + \frac{6}{7}\theta^{\dagger}_{c3},\\
\theta^{*}_{c2} &= \frac{1}{7}\theta^{\dagger}_{c1} + \frac{4}{21}\theta^{\dagger}_{c2} - \frac{1}{7}\theta^{\dagger}_{c3},\\
\theta^{*}_{c3} &= \frac{1}{7}\theta^{\dagger}_{c1} - \frac{1}{7}\theta^{\dagger}_{c2} - \frac{1}{7}\theta^{\dagger}_{c3}.
\end{aligned}$$

In this constrained model the total effect is $\theta^{\dagger}_{c1} = \theta^{*}_{c1} + 2\theta^{*}_{c2} + 4\theta^{*}_{c3}$. We use the constrained and reparameterized specification from Equation 20.1 for analysis. For convenience, we have dropped the superscript $*$ or $\dagger$ from θ when generically discussing the likelihood or MCMC sampler.

We denote the Poisson log likelihood for county c by $\ell_c(\boldsymbol{\theta}_c, \boldsymbol{\beta}_c)$, where bold face represents a vector of the relevant parameters, such as $\boldsymbol{\theta}_c = (\theta_{c1}, \ldots, \theta_{cd})^t$. Our approach uses Bayesian methodology to explore the joint likelihood by smoothing parameters across counties. However, the number of nuisance parameters makes implementation and prior specification unwieldy. Therefore, we replace the county-specific log likelihoods with the associated profile log likelihoods:

$$\tilde{\ell}_c(\boldsymbol{\theta}_c) = \ell_c\{\boldsymbol{\theta}_c, \hat{\boldsymbol{\beta}}_c(\boldsymbol{\theta}_c)\}, \quad \text{where } \hat{\beta}_c(\boldsymbol{\theta}_c) = \text{argmax}_{\boldsymbol{\beta}_c}\ell_c(\boldsymbol{\theta}_c, \boldsymbol{\beta}_c).$$

This step greatly reduces the complexity of the MCMC fitting algorithm. However, it does so at the cost of theoretical unity, as the profile likelihood used for Bayesian inference is not a proper likelihood (Monahan and Boos, 1992), as well as computing time. We stipulate that this choice may impact the validity of the sampler and inference. Currently, we assess validity by comparing results with maximum likelihood results for counties with complete data.

The model for the air pollution time series contains trend variables and AR(p) distributed errors. We assume that

$$\log(X_{ct}) = \mathbf{W}_{ct}\boldsymbol{\psi}_c + \epsilon_{ct}, \tag{20.2}$$

where the ϵ_{ct} are a stationary autoregressive process of order p with conditional means and variances

$$\text{E}[\epsilon_{ct}|\epsilon_{c,t-1}, \ldots, \epsilon_{c,t-p}] = \sum_{j=1}^{p}\alpha_{cj}\epsilon_{c,t-j}, \quad \text{var}(\epsilon_{ct}|\epsilon_{c,t-1}, \ldots, \epsilon_{c,t-p}) = \sigma^2_c.$$

Here the trend term, $\mathbf{W}_{ct}\boldsymbol{\psi}_c$, represents the slowly varying correlation between air pollution and seasonality. Specifically, we set $\mathbf{W}_{ct}$ to be a natural cubic spline with 24 degrees of freedom per year. Throughout, we set $p = 4$.

20.4.1 Prior and Hierarchical Model Specification

We place a $N(\boldsymbol{\mu}, \boldsymbol{\Sigma}_\theta)$ prior on the distributed lag parameters, and a diffuse normal prior for $\boldsymbol{\mu}$ and an inverted Wishart prior with an identity matrix scale on $\boldsymbol{\Sigma}_\theta$ with 4 degrees of freedom. Here, $\boldsymbol{\mu}$ is a parameter of central interest, estimating the between-county mean distributed lag parameters.

We do not place a prior on the missing-data trend term $\boldsymbol{\psi}_c$, instead fixing it from the onset at the least squares estimated value. For the autoregressive parameters, α_{cj}, we place the prior on the lagged partial autocorrelations (Barnett et al., 1996; Monahan, 1983). We refer the reader to Diggle (1990) for a definition of partial autocorrelations and Huerta and West (1999) for a different perspective for placing priors on autoregressive parameters.

We use a recursive formula of Durbin (1960), to transform the autoregressive parameters to and from the partial autocorrelations. Let $\tilde{\alpha}_{cj}$ represent the p partial autocorrelations for county c; we specify that

$$0.5 \log\{(1 + \tilde{\boldsymbol{\alpha}}_c)/(1 - \tilde{\boldsymbol{\alpha}}_c)\} \sim N(\boldsymbol{\zeta}, \boldsymbol{\Sigma}_\alpha),$$

where the Fisher's Z transformation, $\log\{(1 + a)/(1 - a)\}$, is assumed to operate componentwise on vectors. Here, taking Fisher's Z transformation is useful as the partial correlations are bounded by 1 in absolute value for stationary series.

We use a diffuse normal prior for $\boldsymbol{\zeta}$ and an inverse Wishart distribution centered at an identity matrix with 10 degrees of freedom. The prior on σ_c^{-2} is gamma with a mean set at the county-specific method of moments estimates and a coefficient of variation of 10. Note that we chose not to shrink variance estimates across counties, as they appear to be well estimated from the data.

20.5 Bayesian Imputation

20.5.1 Sampler

Here we give an overview of the Bayesian imputation algorithm. Let brackets generically denote a density, and let $\mathbf{X}_{c,\text{obs}}$ and $\mathbf{X}_{c,\text{miss}}$ be the collection of X_{tc} observed and missing components for county c respectively, $\mathbf{Y}_c$ be the collection of Y_{tc}, $\mathbf{P}_c = \{\psi_c, \alpha_{1c}, \ldots, \alpha_{pc}, \sigma_c\}$, $\mathbf{P}$ be the collection of between-county parameters and $\mathbf{H}$ denote hyperparameters. Then, the full join posterior is

$$\begin{aligned}
&[\mathbf{X}_{0,\text{miss}}, \ldots, \mathbf{X}_{C-1,\text{miss}}, \boldsymbol{\theta}_0, \ldots, \boldsymbol{\theta}_{C-1}, \mathbf{P}_0, \ldots, \mathbf{P}_{C-1}, \mathbf{P} \mid \mathbf{Y}_0, \ldots, \mathbf{Y}_{C-1}, \mathbf{X}_{0,\text{obs}}, \mathbf{X}_{C-1,\text{obs}}, \mathbf{H}] \\
&\quad \propto \left\{ \prod_c [\mathbf{Y}_c \mid \mathbf{X}_{c,\text{miss}}, \mathbf{X}_{c,\text{obs}}, \boldsymbol{\theta}_c][\mathbf{X}_{c,\text{miss}}, \mathbf{X}_{c,\text{obs}} \mid \mathbf{P}_c][\mathbf{P}_c \mid \mathbf{P}, \mathbf{H}] \right\} [\mathbf{P} \mid \mathbf{H}].
\end{aligned}$$

Here, recall that $[\mathbf{Y}_c \mid \mathbf{X}_{c,\text{miss}}, \mathbf{X}_{c,\text{obs}}, \boldsymbol{\theta}_c]$ uses the profile likelihood, rather than the actual likelihood. Our sampler proceeds as follows (where *EE* is "everything else"):

$$[\mathbf{X}_{0,\text{miss}} \mid EE] \propto [\mathbf{Y}_0 \mid \mathbf{X}_{0,\text{miss}}, \mathbf{X}_{0,\text{obs}}, \boldsymbol{\theta}_0][\mathbf{X}_{0,\text{miss}}, \mathbf{X}_{0,\text{obs}} \mid \mathbf{P}_0],$$

$$
\begin{aligned}
&[\mathbf{X}_{1,\text{miss}} \mid EE] \propto [\mathbf{Y}_1 \mid \mathbf{X}_{0,\text{miss}}, \mathbf{X}_{0,\text{obs}}, \boldsymbol{\theta}_1][\mathbf{X}_{0,\text{miss}}, \mathbf{X}_{0,\text{obs}} \mid \mathbf{P}_1],\\
&\vdots\\
&[\mathbf{X}_{C-1,\text{miss}} \mid EE] \propto [\mathbf{Y}_C \mid \mathbf{X}_{C-1,\text{miss}}, \mathbf{X}_{C-1,\text{obs}}, \boldsymbol{\theta}_{C-1}][\mathbf{X}_{C-1,\text{miss}}, \mathbf{X}_{C-1,\text{obs}} \mid \mathbf{P}_{C-1}],\\
&[\mathbf{P}_0 \mid EE] \propto [\mathbf{Y}_0 \mid \mathbf{X}_{c,\text{miss}}, \mathbf{X}_{c,\text{obs}}, \boldsymbol{\theta}_0][\mathbf{X}_{0,\text{miss}}, \mathbf{X}_{0,\text{obs}} \mid \mathbf{P}_0],\\
&[\mathbf{P}_1 \mid EE] \propto [\mathbf{Y}_1 \mid \mathbf{X}_{c,\text{miss}}, \mathbf{X}_{c,\text{obs}}, \boldsymbol{\theta}_1][\mathbf{X}_{1,\text{miss}}, \mathbf{X}_{1,\text{obs}} \mid \mathbf{P}_1],\\
&\vdots\\
&[\mathbf{P}_{C-1} \mid EE] \propto [\mathbf{Y}_{C-1} \mid \mathbf{X}_{C-1,\text{miss}}, \mathbf{X}_{C-1,\text{obs}}, \boldsymbol{\theta}_C][\mathbf{X}_{C-1,\text{miss}}, \mathbf{X}_{C-1,\text{obs}} \mid \mathbf{P}_{C-1}],\\
&[\mathbf{P} \mid EE] \propto [\mathbf{P}_c \mid \mathbf{P}, \mathbf{H}][\mathbf{P} \mid \mathbf{H}].
\end{aligned}
$$

Because of the Gibbs-friendly priors, $\boldsymbol{\mu}$ and $\boldsymbol{\zeta}$ have multivariate normal full conditionals. Moreover, $\boldsymbol{\Sigma}_\theta$ and $\boldsymbol{\Sigma}_\alpha$ have inverse Wishart full conditionals, while the $\{\sigma_c^2\}$ have an inverted gamma. The county-specific distributed lag parameters and AR parameters, $\{\boldsymbol{\theta}_c\}$ and $\boldsymbol{\alpha}_c$, require a Metropolis step. We use a variable-at-a-time, random-walk update. Further details on the full conditionals are given in the Appendix to this chapter.

The update of the missing data deserves special attention. We use a variable-at-a-time Metropolis step to impute ϵ_{tc} for each missing day conditional on the remaining. Consider $p = 4$ and let ϵ_{c5} be a missing day to be imputed. We use the autoregressive prior for the day under consideration given all of the remaining days as the proposal. For example, the distribution of ϵ_{c5} given $\{\epsilon_{c1}, \ldots, \epsilon_{c4}, \epsilon_6, \ldots, \epsilon_9\}$ is used to generate the proposal for ϵ_{c5}, that is, the four neighboring days before and after the day under consideration. Because of the AR(4) assumption, this is equivalent to the distribution of ϵ_5 given all of the days. After imputation, $X_{c5} = \exp(\mathbf{W}_{c5}\psi_c + \epsilon_{c5})$, is calculated. By simulating from the prior distribution of the current missing day given the remainder, only the contribution of X_{c5} to the profile likelihood remains in the Metropolis ratio. To summarize, the distribution of the current day given the remainder, disregarding the profile likelihood, is used to generate proposals; the profile likelihood is then used in a Metropolis correction. Of course, since $PM_{2.5}$ has a relatively weak relationship with the response, the acceptance rate is high.

20.5.2 A Parallel Imputation Algorithm

Given the large number of days that need to be imputed for the counties with missing data, and the difficult calculation of the county-specific profile likelihoods, the time for running such a sampler is quite long. In this section we propose a parallel computing approach that can greatly speed up computations.

Notice that the conditional-independence structure from Section 20.5.1 illustrates *that all of the county-specific full conditionals are conditionally independent*. Thus, the imputation of the missing predictor data, the simulation of the county-specific parameters, and the calculation of the profile likelihoods can be performed simultaneously. Hence, it represents an ideal instance where we can increase the efficiency of the simulation of a single Markov chain with parallel computation.

To elaborate, let f be the time required to update $[\mathbf{P} \mid EE]$, $g_0, g_1, \ldots, g_{C-1}$ be the time required to transfer the relevant information to and from the separate nodes for parallel processing, and $h_0, h_1, \ldots, h_{C-1}$ be the time required to perform the processing, as depicted in Figure 20.5. Suppose that C nodes are available for computation. Then, conceptually, the

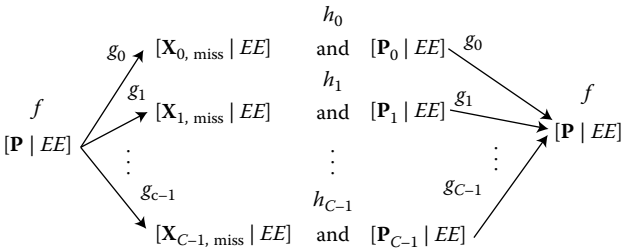

FIGURE 20.5
Parallel computing diagram.

run-time per parallel MCMC iteration is $f + \max_c(g_c + h_c)$. In contrast, the single processor run-time would be $f + \sum_c(h_c)$. Clearly, if the transfer times, $\{g_c\}$ are small relative to the county-specific processing times, $\{h_c\}$, then substantial savings can be made by parallelizing the process, with the gains scaling proportional to the number of conditionally independent full conditionals. This is exactly the setting of the Medicare claims data, where the profile likelihood and imputation county-specific calculations are very time-consuming. Of course, this simple schematic is extremely optimistic in not accounting for several factors, such as variability in the number of available nodes, node-specific run-times and the added time for the software to manage the parallelization. However, it does suggest that substantial gains can be made with parallelization.

While we know of few implementations of parallel MCMC of this scope, this approach to parallelizing Markov chains and its generalizations has been discussed previously (Kontoghiorghes, 2006; Rosenthal, 2000; Winkler, 2003). Moreover, other approaches could be used for parallelizing Markov chains. When applicable, perfect sampling (Fill, 1998; Fill et al., 2000; Propp and Wilson, 1998) could be easily parallelized. Specifically, each perfect sample is an independent and identically distributed draw from the stationary distribution and hence can be generated independently. Also regeneration times (Hobert et al., 2002; Jones et al., 2006; Mykland et al., 1995) create independent tours of the chain from other regeneration times. Therefore, given a starting value at a regeneration time, the tours could be generated in parallel. These two techniques have the drawback that a substantial amount of mathematics needs to be addressed to simply implement the sampler prior to any discussion of parallel implementation. A less theoretically justified, yet computationally simple, approach parallelizes and combines multiple independent chains (Gelman and Rubin, 1992; Geyer, 1992).

Most work on statistical parallel computing algorithms depends on existing network-based parallel computing algorithms, such as Parallel Virtual Machines (Beguelin et al., 1995) or Message Passing Interface (Snir et al., 1995), such as implemented in the R package `SNOW` (Rossini et al., 2007). These programs are not optimized for particular statistical problems or computational infrastructures and, furthermore, require direct computer-to-computer communication. While such parallel computing architectures are used, large computing clusters that employ queuing management software often cannot take advantage of these approaches.

In contrast, our approach uses a disk-based shared memory blackboard system which required building the parallelization software. Specifically in our approach, a collection of tokens, one for each county, are used to represent which counties currently need processing. A collection of identical programs, which we refer to as spiders, randomly select a token from the bin and move it to a bin of tokens representing counties currently being

operated on. We have adopted several strategies to avoid race conditions, where two spiders simultaneously attempt to grab the same token, including: using file system locks and creating small random delays before the spider grabs the token. The spider then performs the county-specific update and moves its token to another bin of counties with finished calculations. The spider then goes back to the original bin and repeats the process. If there are no tokens remaining, the first spider to discover this fact then performs the national update while the remaining sit idle. It then moves the tokens back to the original bin to restart the process. Disk-based shared memory is used for all of the data transfer.

The benefits of this strategy for parallel MCMC are many. Notably, nodes or spiders can be dynamically added or subtracted. Moreover, load balancing can be accomplished easily. In addition, the system allowed us to use a storage area network (SAN) as the shared memory resource (blackboard). While having much slower data transfer than direct computer-to-computer based solutions, this approach allowed us to implement a parallel programming in spite of scheduling software that precludes more direct parallelization. As an added benefit, using the SAN for data transfer builds in automatic checkpointing for the algorithm. We've also found that this approach facilitates good MCMC practice, such as using the ending value from initial runs as the starting value for final runs. Of course, an overwhelming negative property of this approach is the need to create the custom, setting-specific, parallelization software.

20.6 Analysis of the Medicare Data

We analyzed the Medicare data using our parallel MCMC algorithm. We used 30 processors, resulting in a run-time of 5–10 seconds per MCMC iteration. In contrast, the run-time for a single processor was over 2 minutes. That is, there is a 90% decrease in run-time due to parallelization.

The sampler was run for 13,000 iterations. This number was used as simply the largest feasible in the time given. Final values from testing-iterations were used as starting values. Trace plots were investigated to evaluate the behavior of the chains, and were also used to change the step size of the random-walk samplers.

Figure 20.6 displays an example imputation for 1000 monitoring days for a county. The black lines connects observed days while the gray line depicts the estimated trend. The points depict the imputed data set. Figure 20.7 depicts a few days for a county where pollution data is observed every three days; the separate lines are iterations of the MCMC process. These figures illustrate the reasonableness of the imputed data. A possible concern is that the imputed data are slightly less variable than the actual data. Moreover, the data is more regular, without extremely high air pollution days. However, this produces conservatively wider credible intervals for the distributed lag estimates.

Figure 20.8 shows the estimated posterior medians for the exponential of the cumulative effect for a 10-unit increase in air pollution, along with 95% credible intervals. The estimates range from 0.746 to 1.423. The cumulative effect is interpreted as the relative increase or decrease in the rate of COPD corresponding to an across-the-board 10-unit increase in $PM_{2.5}$ over the prior six days. Therefore, for example, 1.007 (the national average) represents a 0.7% increase in the rate of COPD per microgram per 10 cubic meter increase in fine particulate matter over six days.

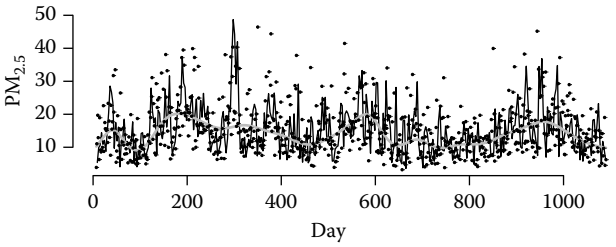

FIGURE 20.6
Example imputation from the MCMC sampler for a specific county. The black line connects observed points while the gray line shows the estimated trend. The points are from a specific iteration of the MCMC sampler.

The national estimate, μ_1, represents the variance-weighted average of the county-specific cumulative effects. The 95% equi-tail credible interval ranges from an estimated 2.6% decrease to a 4.0% increase in the rate of COPD. The posterior median was a 0.7% increase. In contrast, a meta-analysis model using the maximum likelihood fits and variances for only those counties with adequate data for fitting the distributed lag model results in a confidence interval for the national cumulative effect ranging from a 5.1% decrease to a 7.5% increase, while the mean is a 1.0% increase. That is, adding the data from the counties with systematic missing data does not appear to introduce a bias, but does greatly reduce the width of the interval.

The shape of the distributed lag function is of interest to environmental health researchers, as different diseases can have very different profiles, such as rates of hospitalization, recurrence, and complications. Examining the shape of the distributed lag function can shed light on the potential relationship of air pollution and the disease. For example, a decline over time could be evidence of the "harvesting" hypothesis, whereby a large air pollution effect for early lags would deplete the risk set of its frailest members, through hospitalization or mortality. Hence, the latter days would have lower effects. Figure 20.9 shows the exponent of 10 times the distributed lag parameters' posterior medians, θ^*_{c1}, θ^*_{c2}, and θ^*_{c3}, by county. Here θ^*_{c1} is the current-day estimate, while θ^*_{c2} and θ^*_{c3} are cumulative effect for days lag 1 to 2 and 3 to 6, respectively. The current-day effect tends to be much larger, and more variable, by county. The comparatively smaller values for the later lags are supportive of the harvesting hypothesis, though we emphasize that other mechanisms could be in place. Further, this model is not ideal for studying such phenomena, as the bin smoothing of the distributed lag parameters may be too crude to explore the distributed lag function's shape.

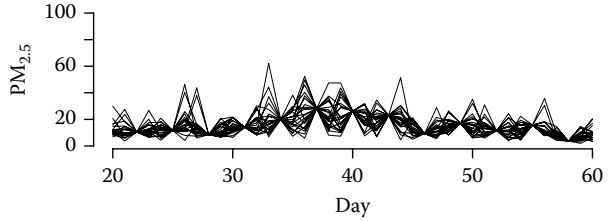

FIGURE 20.7
Several example imputations for a subset of the days for a county. The lines converge on observed days.

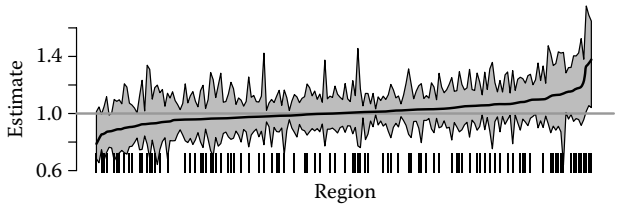

FIGURE 20.8
Estimates and credible intervals for the exponential of the distributed lag cumulative effect by county for a 10-unit increase in air pollution, $10\theta_{c1}$. The solid middle line shows the posterior medians; the gray area shows the estimated 95% equi-tail credible intervals. A horizontal reference line is drawn at 1. Hash marks denote counties with systematic missing data, where the distributed lag model could not be fit without imputation.

Figure 20.10 shows 95% credible intervals for the AR parameters across counties. The counties are organized so that the rightmost 97 counties have the systematic missing data. Notice that, in these counties, their estimate for the AR(1) parameter is attenuated toward zero over the counties with complete data. The estimated posterior median of ζ is $(0.370, -0.010, -0.011, -0.025)^t$. The primary AR(1) parameter is slightly below the 0.5, because of the contribution of those counties with missing data. Regardless, we note that

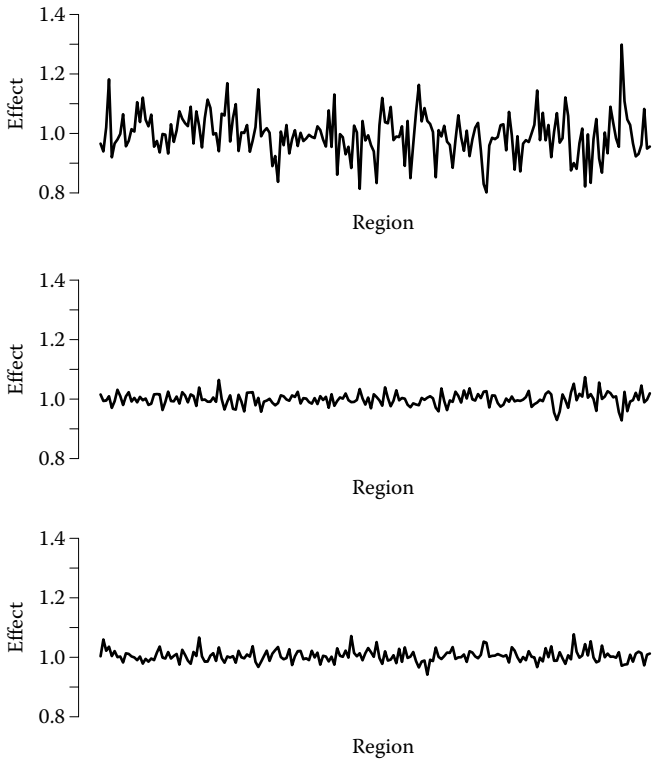

FIGURE 20.9
Exponent of ten times the distributed lag parameters' posterior medians: θ^*_{c1} (top), θ^*_{c2} (middle), and θ^*_{c3} (bottom), by county.

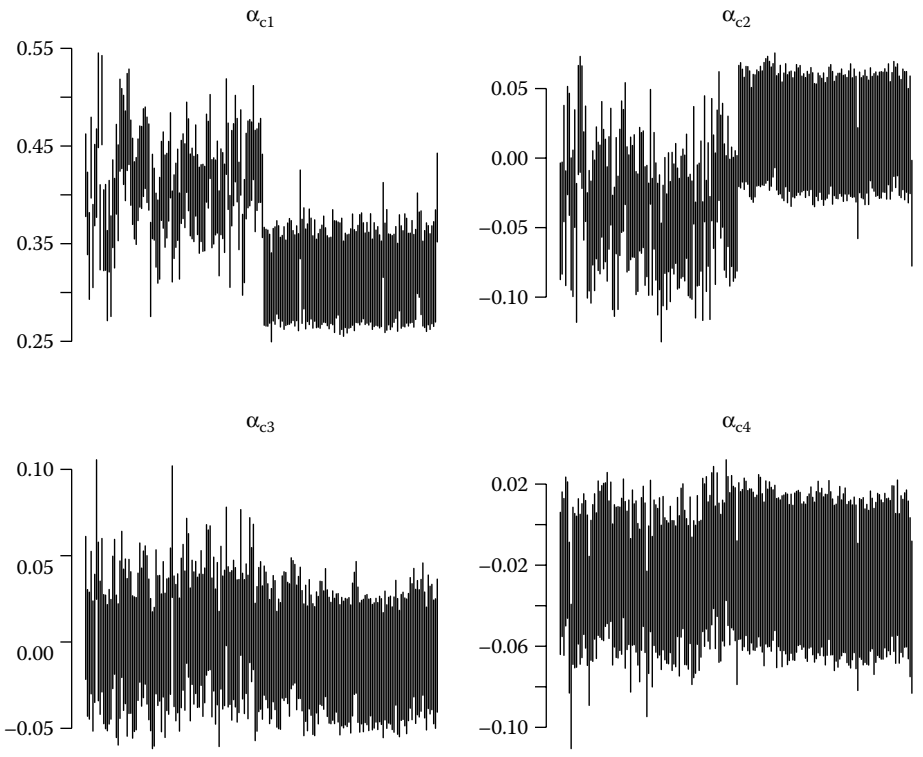

FIGURE 20.10
Posterior credible intervals for the AR parameters by county.

this shrinkage estimation greatly improves on county-specific estimation (see Figure 20.3) for those cites with incomplete data.

20.7 Summary

In this chapter, we propose an MCMC algorithm for fitting distributed lag models for time series with systematically missing predictor data. We emphasize that our analysis only scratches the surface for the analysis of the Medicare claims data. One practical issue is the effect of varying degrees of confounder adjustment, in terms of both the degrees of freedom employed in nonlinear fits and the confounders included. Moreover, a more thorough analysis would consider other health outcomes and different numbers of lags included. Also, commonly air pollution effects are interacted with age or age categories, because of the plausibility of different physical responses to air pollution with aging. In addition, the PM measurements are aggregates of several chemical pollutants. Determining the effects of individual component parts may help explain some of the county variation in the effect of air pollution on health.

The bin smoother on the distributed lag parameters allows for a simpler algorithm and interpretation. However, more reasonable smoothers, such as setting $\mathbf{A}$ to the design matrix for a regression-spline model, should be considered. This would allow much more accurate

exploration of the shape of the distributed lag model, as well as variations in its shape by counties.

The use of the profile likelihood instead of the actual likelihood raises numerous issues and concerns. Foremost is the propriety of the posterior and hence the validity of the sampler and inference. The theoretical consequences of this approach should be evaluated. Moreover, comparisons with other strategies, such as placing independent diffuse priors on the nuisance parameters, are of interest.

Also of interest is to eliminate the attenuation of the estimates of the autoregressive parameters for the counties with missing data. To highlight this problem more clearly, suppose that instead of 97 counties with missing data, we had 9700 with data recorded every other day. Then the accurate information regarding $\boldsymbol{\zeta}$ and $\boldsymbol{\Sigma}_\alpha$ contained in the counties with complete data would be swamped by the noisy bimodal likelihoods from the counties with systematic missingness. More elaborate hierarchies on this component of the model may allow for the counties with observed data to have control over estimation of these parameters.

In addition, the potential informativeness of counties having missing data (see Little and Rubin, 2002) should be investigated. To elaborate, clearly the pattern of missing data is uninformative for any given county; however, whether or not a county collected data every day or every third day may be informative. For example, counties with air pollutions levels well below or above standards may be less likely to collect data every day. Such missingness may impact national estimates.

We also did not use external variables to impute the missing predictor data. Ideally, a completely observed instrumental variable that is causally associated with the predictor, yet not with the response, would be observed. Such variables could be used to impute the predictor, but would not confound its relationship with the response. However, such variables are rarely available. More often variables that are potentially causally associated with the predictor are also potentially causally associated with response. For example, seasonality and temperature are thought to be causally associated with *PM* levels and many health outcomes. Hence, using those variables to impute the missing predictor data would immediately raise the criticism that any association found was due to residual confounding between the response and the variables used for imputation.

These points notwithstanding, this work suggests potential for the ability to impute missing data for distributed lag models. The Bayesian model produces a marked decrease in the width of the inter-county estimate of the cumulative effect. Moreover, the imputed data sets are consistent with the daily observed data, though perhaps being more regular and less variable. However, we note the bias incurred by lower variability in the imputed air pollution is conservative, and would attenuate the distribute lag effects.

A second accomplishment of this chapter is the parallelization algorithm and software development. The computational overhead for the parallelization software was small, and hence the decrease in run-time was nearly linear with the number of nodes.

Race conditions, times when multiple spiders attempted to access the same token, represent a difficult implementation problem. For example, after the county-specific updates finish, all spiders attempt to obtain the token representing the national update simultaneously. A colleague implementing a similar system proposed a potential solution (Fernando Pineda, personal communication). Specifically, he uses atomic operations on a lock file to only allow access to the bin of tokens to one program at a time. As an analogy, this approach has a queue of programs waiting for access to the bins to obtain a token. In contrast, our approach allows simultaneous access to the bins, thus increasing speed, though also increasing the likelihood of race conditions. A fundamental problem we have yet to

solve is the need for truly atomic operations on networked file systems to prevent these race conditions. Our use of file system locks when moving files as the proposed atomic operation made the system very fragile, given the complex nature and inherent lag of an NFS-mounted SAN.

Our current solutions to these problems are inelegant. First, as described earlier, random waiting times were added. Secondly, spiders grabbed tokens in a random order. Finally, a worker program was created that searched for and cleaned up lost tokens and ensured that the appropriate number of spiders were operational. We are currently experimenting with the use of an SQL database, with database queries, rather than file manipulations to manage the tokens.

Appendix: Full Conditionals

We have

$$\boldsymbol{\theta}_c \propto \exp\left(\tilde{\ell}_c(\boldsymbol{\theta}_c) - \frac{1}{2}(\boldsymbol{\theta}_c - \boldsymbol{\mu})^t \boldsymbol{\Sigma}_\theta (\boldsymbol{\theta}_c - \boldsymbol{\mu})\right),$$

$$\boldsymbol{\mu} \sim N\left\{\left(C\boldsymbol{\Sigma}_\theta^{-1} + G_1^{-1}\right)^{-1} \boldsymbol{\Sigma}_\theta^{-1} \sum_{c=0}^{C-1} \boldsymbol{\theta}_c, \left(C\boldsymbol{\Sigma}_\theta^{-1} + G_1^{-1}\right)^{-1}\right\},$$

$$\boldsymbol{\Sigma}_\theta^{-1} \sim \text{Wishart}\left(G_2 + \sum_{c=0}^{C-1} (\boldsymbol{\theta}_c - \boldsymbol{\mu})(\boldsymbol{\theta}_c - \boldsymbol{\mu})^t, df_1 + C\right),$$

$$\boldsymbol{\alpha}_c \sim N\left\{(\mathbf{E}_c^t \mathbf{E}_c/\sigma_c^2 + \boldsymbol{\Sigma}_\alpha^{-1})^{-1} \mathbf{E}_c^t \boldsymbol{\epsilon}_c, (\mathbf{E}_c^t \mathbf{E}_c/\sigma_c^2 + \boldsymbol{\Sigma}_\alpha^{-1})\right\},$$

$$\boldsymbol{\zeta} \sim N\left\{\left(c\boldsymbol{\Sigma}_\alpha^{-1} + G_4^{-1}\right)^{-1} \boldsymbol{\Sigma}_\alpha^{-1} \sum_{c=0}^{C-1} \boldsymbol{\alpha}_c, \left(c\boldsymbol{\Sigma}_\alpha^{-1} + G_4^{-1}\right)^{-1} \boldsymbol{\Sigma}_\alpha^{-1}\right\},$$

$$\boldsymbol{\Sigma}_\alpha \sim \text{Wishart}\left(G_5 + \sum_{c=0}^{C-1} (\boldsymbol{\alpha}_c - \boldsymbol{\zeta})(\boldsymbol{\alpha}_c - \boldsymbol{\zeta})^t), df_2 + C\right),$$

$$\sigma_c^{-2} \sim \Gamma\left\{C/2 + G_6, \sum_{c=0}^{C-1} \left(\epsilon_{ct} - \sum_{u=1}^{p} \epsilon_{c,t-u}\alpha_u\right)^2 + G_7\right\}.$$

Here $\boldsymbol{\epsilon}_c = (\epsilon_{c1}, \ldots, \epsilon_{cT_c})^t$, where ϵ_{ct} are the residuals after fitting model (Equation 20.2). The matrix $\mathbf{E}_c$ denotes the lagged values of $\boldsymbol{\epsilon}_c$. $G_1, G_2, \ldots$ denote generic hyperparameters whose values are described in the chapter, while df_1 and df_2 correspond to prior Wishart degrees of freedom. That is, G_1 is the prior variance on $\boldsymbol{\mu}$; G_2 represents the Wishart scale matrix for $\boldsymbol{\Sigma}_\theta$; (G_3, G_4) represent the prior means and variance on $\boldsymbol{\zeta}$; G_5 represents the Wishart scale matrix for $\boldsymbol{\Sigma}_\alpha$; and G_6 and G_7 are the gamma shape and rate on σ_c^{-2}.

Acknowledgment

The authors would like to thank Dr. Fernando Pineda for helpful discussions on parallel computing.

References

Barnett, G., Kohn, R., and Sheather, S. 1996. Bayesian estimation of an autoregressive model using Markov chain Monte Carlo. *Journal of Econometrics*, 74:237–254.

Beguelin, A., Dongarra, J., Jiang, W., Manchek, R., and Sunderam, V. 1995. *PVM: Parallel Virtual Machine: A Users' Guide and Tutorial for Networked Parallel Computing*. MIT Press, Cambridge, MA.

Bell, M., McDermott, A., Zeger, S., Samet, J., and Dominici, F. 2004. Ozone and short-term mortality in 95 urban communities, 1987–2000. *Journal of the American Medical Association*, 292(19):2372–2378.

Broersen, P., de Waele, S., and Bos, R. 2004. Autoregressive spectral analysis when observations are missing. *Automatica*, 40(9):1495–1504.

Carlin, B. and Louis, T. 2009. *Bayesian Methods for Data Analysis*, 3rd edn. Chapman & Hall/CRC Press, Boca Raton, FL.

Diggle, P. 1990. *Times Series: A Biostatistical Introduction*. Oxford University Press, Oxford.

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

Dominici, F., Peng, R., Bell, M., Pham, L., McDermott, A., Zeger, S., and Samet, J. 2006. Fine particulate air pollution and hospital admission for cardiovascular and respiratory diseases. *Journal of the American Medical Association*, 295(10):1127–1134.

Durbin, J. 1960. The fitting of time series models. *International Statistical Review*, 28:233–244.

Fill, J. 1998. An interruptible algorithm for perfect sampling via Markov chains. *Annals of Applied Probability*, 8(1):131–162.

Fill, J., Machida, M., Murdoch, D., and Rosenthal, J. 2000. Extension of Fill's perfect rejection sampling algorithm to general chains. *Random Structures and Algorithms*, 17(3–4):290–316.

Gelman, A. and Rubin, D. 1992. Inference from iterative simulation using multiple sequences. *Statistical Science*, 7:457–511.

Geyer, C. 1992. Practical Markov chain Monte Carlo. *Statistical Science*, 7:473–511.

Hobert, J., Jones, G., Presnell, B., and Rosenthal, J. 2002. On the applicability of regenerative simulation in Markov chain Monte Carlo. *Biometrika*, 89(4):731–743.

Huerta, G. and West, M. 1999. Priors and component structures in autoregressive time series models. *Journal of the Royal Statistical Society, Series B*, 61(4):881–899.

Jones, G., Haran, M., Caffo, B., and Neath, R. 2006. Fixed-width output analysis for Markov chain Monte Carlo. *Journal of the American Statistical Association*, 101(476):1537–1547.

Kontoghiorghes, E. 2006. *Handbook of Parallel Computing and Statistics*. Chapman & Hall/CRC, Boca Raton, FL.

Little, R. and Rubin, D. 2002. *Statistical Analysis with Missing Data*, 2nd edn. Wiley, Hoboken, NJ.

Monahan, J. 1983. Fully Bayesian analysis of ARMA time series models. *Journal of Econometrics*, 21:307–331.

Monahan, J. and Boos, D. 1992. Proper likelihoods for Bayesian analysis. *Biometrika*, 79(2):271–278.

Mykland, P., Tierney, L., and Yu, B. 1995. Regeneration in Markov chain samplers. *Journal of the American Statistical Association*, 90(429):233–241.

Propp, J. and Wilson, D. 1998. How to get a perfectly random sample from a generic Markov chain and generate a random spanning tree of a directed graph. *Journal of Algorithms*, 27(2):170–217.

Rosenthal, J. 2000. Parallel computing and Monte Carlo algorithms. *Far East Journal of Theoretical Statistics*, 4(2):207–236.

Rossini, A., Tierney, L., and Li, N. 2007. Simple parallel statistical computing in R. *Journal of Computational and Graphical Statistics*, 16(2):399.

Samet, J., Dominici, F., Currieo, F., Coursac, I., and Zeger, S. 2000. Fine particulate air pollution and mortality in 20 U.S. cities 1987-1994. *New England Journal of Medicine*, 343(24):1742–1749.

Schwartz, J. 2000. Harvesting and long term exposure effects in the relation between air pollution and mortality. *American Journal of Epidemiology*, 151(5):440.

Snir, M., Otto, S., Walker, D., Dongarra, J., and Huss-Lederman, S. 1995. *MPI: The Complete Reference*. MIT Press, Cambridge, MA.

Wallin, R. and Isaksson, A. 2002. Multiple optima in identification of ARX models subject to missing data. *EURASIP Journal on Applied Signal Processing*, 1:30–37.

Winkler, G. 2003. *Image Analysis, Random Fields, and Dynamic Monte Carlo Methods*, 2nd edn. Springer, Berlin.

Zanobetti, A., Wand, M. P., Schwartz, J., and Ryan, L. M. 2000. Generalized additive distributed lag models: Quantifying mortality displacement. *Biostatistics*, 1(3):279–292.

Zeger, S., Dominici, F., and Samet, J. 1999. Harvesting-resistant estimates of air pollution effects on mortality. *Epidemiology*, 10(2):171.