2007-05-31

Temporally Correlated Dirichlet Processes in Pollution Receptor Modeling

Matthew J. Heaton
Brigham Young University - Provo

TEMPORALLY CORRELATED DIRICHLET PROCESSES IN POLLUTION RECEPTOR MODELING

by

Matthew J. Heaton

A thesis submitted to the faculty of

Brigham Young University

in partial fulfillment of the requirements for the degree of

Master of Science

Department of Statistics

Brigham Young University

August 2007

BRIGHAM YOUNG UNIVERSITY

GRADUATE COMMITTEE APPROVAL

of a thesis submitted by

Matthew J. Heaton

This thesis has been read by each member of the following graduate committee and by majority vote has been found to be satisfactory.

Date	C. Shane Reese, Chair
Date	William F. Christensen
Date	Scott D. Grimshaw

BRIGHAM YOUNG UNIVERSITY

As chair of the candidate's graduate committee, I have read the thesis of Matthew J. Heaton in its final form and have found that (1) its format, citations, and bibliographical style are consistent and acceptable and fulfill university and department style requirements; (2) its illustrative materials including figures, tables, and charts are in place; and (3) the final manuscript is satisfactory to the graduate committee and is ready for submission to the university library.

Date

C. Shane Reese
Chair, Graduate Committee

Accepted for the Department

Scott D. Grimshaw
Graduate Coordinator

Accepted for the College

Thomas W. Sederberg
Associate Dean, College of Physical and Mathematical Sciences

ABSTRACT

TEMPORALLY CORRELATED DIRICHLET PROCESSES IN POLLUTION RECEPTOR MODELING

Matthew J. Heaton
Department of Statistics
Master of Science

Understanding the effect of human-induced pollution on the environment is an important precursor to promoting public health and environmental stability. One aspect of understanding pollution is understanding pollution sources. Various methods have been used and developed to understand pollution sources and the amount of pollution those sources emit. Multivariate receptor modeling seeks to estimate pollution source profiles and pollution emissions from concentrations of pollutants such as particulate matter (PM) in the air. Previous approaches to multivariate receptor modeling make the following two key assumptions: (1) PM measurements are independent and (2) source profiles are constant through time. Notwithstanding these assumptions, the existence of temporal correlation among PM measurements and time-varying source profiles is commonly accepted. In this thesis an approach to multivariate receptor modeling is developed in which the temporal structure of PM measurements is accounted for by modeling source profiles as a time-dependent Dirichlet process. The Dirichlet process (DP) pollution model developed herein is evaluated using several simulated data sets. In the presence of time-varying source

profiles, the DP model more accurately estimates source profiles and source contributions than other multivariate receptor model approaches. Additionally, when source profiles are constant through time, the DP model outperforms other pollution receptor models by more accurately estimating source profiles and source contributions.

ACKNOWLEDGEMENTS

Education is never the work of just one person. I would like to personally thank the following people who have helped and supported me throughout my education at Brigham Young University. My wife Cami, for her patience, love, and confidence in me. Shane Reese, for teaching me how to do statistics the right way. William Christensen, for always making research enjoyable. Scott Grimshaw, for convincing me to pursue a degree in statistics rather than economics. My parents for their financial support of my academic goals. Finally, all the other faculty and staff of the Statistics department at Brigham Young University who had the patience to put up with me all these years.

CONTENTS

TABLES

Table

FIGURES

Figure

1. THE PROBLEM OF POLLUTION SOURCE APPORTIONMENT

1.1 Pollution Source Apportionment

In response to an increase in health concerns arising from ambient air particles, the Unites States Environmental Protection Agency (USEPA) established the "supersites" program. As part of this program, sites were established throughout the United States (Atlanta, St. Louis, etc.) to measure concentrations of various chemical species in the air. Pollution Source Apportionment (PSA) derives information regarding pollution sources from ambient air pollution data. Pollution source apportionment has two main goals: (1) acquire information regarding pollution sources through estimating pollution source profiles, and (2) estimate the contribution of each source to measured ambient air pollution.

Identifying pollution sources is the first step and goal of PSA. The major contributors to pollution can be identified by a unique pollution source profile. Let the p-vector, $\boldsymbol{\lambda}_k = (\lambda_{1k}, \lambda_{2k}, \ldots, \lambda_{Pk})'$, be the pollution source profile for the k^{th} pollution source. Each $\lambda_{pk} \in \boldsymbol{\lambda}_k$ represents the proportion of chemical p in pollution emitted from source k. Hence, $\sum_{p=1}^{P} \lambda_{pk} = 1$ if all chemical species emitted from source k are included in the pollution model, otherwise $\sum_{p=1}^{P} \lambda_{pk} < 1$. By obtaining estimates of the vector $\boldsymbol{\lambda}_k$, pollution sources can be identified by matching the chemical makeup of pollution emitted from the given source to the estimated source profile.

After major pollution sources have been identified, PSA seeks to estimate the contribution of that source to ambient air pollution. Let f_{kt} represent the contribution of the k^{th} source to ambient air pollution at time t as measured from the pollution receptor site. If f_{kt} exceeds the USEPA's legal limit, then the k^{th} pollution source can be assessed fines or other penalties for emitting too much pollution. Additionally, tracking changes in f_{kt} leads to an understanding of the amount of pollution being

emitted by the regulatory bodies through time. Thus, obtaining accurate estimates of f_{kt} is important in regulating pollution emissions. If pollution emissions can be regulated, then health risks to the environment and the human population surrounding pollution sources can be minimized. In these aspects, PSA is an important problem to environmental stability and human health.

1.2 Pollution Sampling Methods

Data used in PSA are concentrations (in micrograms per cubic meter) of various chemical species in ambient air. The supersites established as part of the USEPA's supersites program are rich sources of measured chemical concentrations in ambient air. Chemicals are measured over different time periods, D; integration times, H; and frequencies, R. While D can remain constant for all chemical species, H and R vary from chemical to chemical. Typically, $D \in \{30 \text{ Days}, \ldots, 2 \text{ Years}\}$, $H \in \{5 \text{ minutes}, \ldots, 24 \text{ hours}\}$, and $R \in \{\text{Continuous, Semi-continuous, 1 in 6 days}\}$.

For example, consider measuring a single chemical over an integration time of $H = 1$ hour and frequency $R =$ continuous during some time period D. On day $t \in D$, the chemical is allowed to accumulate on a filter continuously ($R =$ Continuous) over $H = 1$ hour intervals. At the end of each interval of length H, the weight of the chemical on the filter is measured and reported. This process is continued for all $t \in D$. For chemicals measured at a frequency of $R =$ semi-continuous, measurements are taken over a fraction of the integration time. For the St. Louis data set used throughout this thesis $D = 640$ Days, $H = 1$ Hour, and $R =$ Continuous.

Because chemical concentrations are measured over consecutive hours and days, measurements are temporally correlated. Therefore, proper statistical modeling of chemical concentrations needs to account for this correlation.

1.3 The Basic PSA Model

The basic PSA model, first introduced by Winchester and Nifong (1971) and Miller et al. (1972), is

$$y_{pt} = \sum_{k=1}^{K} \lambda_{pk} f_{kt} + e_{pt}; \quad p = 1, \ldots, P \quad t = 1, \ldots, N, \tag{1.1}$$

where y_{pt} is the concentration of the p^{th} chemical species in ambient air measured at time t, λ_{pk} is the proportion of the p^{th} chemical from the k^{th} source, f_{kt} is the contribution of the k^{th} source to the atmospheric pollution at time t, e_{pt} is the model error of the p^{th} chemical at time t, K is the number of pollution sources, N is the total number of time periods, and P is the total number of chemical species. Writing Equation 1.1 in matrix notation yields

$$\underset{P\times 1}{\mathbf{y}_t} = \underset{P\times K}{\mathbf{\Lambda}}\ \underset{K\times 1}{\mathbf{f}_t} + \underset{P\times 1}{\mathbf{e}_t}, \quad t = 1, \ldots, N, \tag{1.2}$$

where $\mathbf{y_t}$ is the vector of P chemical species measured at time t, $\mathbf{\Lambda}$ is the matrix of K pollution source profiles, and $\mathbf{f}_t = (f_{1t}, \ldots, f_{Kt})'$. Expanding Equation 1.1 to take into account all N time periods, Equation 1.2 becomes

$$\underset{P\times N}{\mathbf{Y}} = \underset{P\times K}{\mathbf{\Lambda}}\ \underset{K\times N}{\mathbf{F}} + \underset{P\times N}{\mathbf{E}}, \tag{1.3}$$

where $\mathbf{Y}$ is the matrix of P chemical species measured over N time periods, $\mathbf{\Lambda}$ is the matrix of pollution source profiles for the K different sources, $\mathbf{F}$ is the matrix of source contributions at each of the N time periods, and $\mathbf{E}$ is the matrix of model errors. The goal of PSA is to estimate $\mathbf{\Lambda}$ and $\mathbf{F}$ and hence acquire information about pollution sources as discussed in Section 1.1.

Assumptions implied by the basic model proposed in Equation 1.3 are summarized by Christensen and Gunst (2004) as the following:

(1) Each source profile is constant through time ($\mathbf{\Lambda}_t = \mathbf{\Lambda}$ for $t = 1, \ldots, N$).

(2) Chemical species are additive.

(3) All influential sources are accounted for in the model.

(4) K is fixed through all time periods.

While methods which account for violations of these assumptions have been developed, most PSA research has been focused on the sensitivity of various methods to violations of these assumptions (see Christensen and Gunst 2004). This thesis relaxes the assumption that source profiles are constant over time (Assumption 1). The goal of this thesis is to develop a model that allows source profiles to vary through time. Mathematically speaking, $\mathbf{\Lambda}_t \neq \mathbf{\Lambda}$ for all t. Relaxing this assumption makes intuitive sense in that pollution sources will vary the relative amount of chemicals emitted across time. For example, auto exhaust emissions vary in chemical composition from day to day depending on the number of diesel trucks versus regular gasoline cars being driven on a given day. Diesel truck emissions have a different chemical makeup than gasoline vehicles. For these reasons, the source profile for auto emissions will vary through time.

1.4 Difficulties in Pollution Source Apportionment

While the model proposed in Equation 1.3 is basic, the problem of pollution source apportionment is quite complex. One particular problem is determining the number of sources (K) to include in the model. If too many sources are included in the model then the model over-fits the data. Additionally, including too many sources in the model can lead to model identifiability problems. If too few sources are included then some potentially important sources might have been overlooked. To add to the difficulty in selecting K, the number of sources may not even be constant over time due to factory closures and openings. Park et al. (1999) discusses various methods for choosing the number of sources to include in the model. Lopes (2000)

outlines a Bayesian approach to selecting the number of pollution sources through the use of a reversible jump Markov chain Monte Carlo (MCMC) algorithm. The problem of choosing the number of sources will not be addressed in this thesis.

Another potential problem in PSA is the temporal correlation present in the chemical concentration vectors $\mathbf{y}_1, \ldots, \mathbf{y}_N$. This correlation could arise from association among source contributions $\mathbf{f}_1, \ldots, \mathbf{f}_N$ or $\mathbf{\Lambda}$ being non-constant and correlated over time, or both. Ignoring temporal correlation can lead to invalid standard errors and goodness-of-fit tests (see Christensen and Sain 2002). In spite of this, most PSA research has been done assuming chemical concentrations are uncorrelated. Some research has been done, however, that attempts to model the correlation structure present in the chemical concentrations (see Park et al. 2001).

One final difficulty in estimating source contributions is that estimates of $\mathbf{\Lambda}$ and $\mathbf{F}$ are subject to non-negativity constraints. Negative source contributions and negative source profiles have no meaning in PSA models. Positivity constraints are difficult to accomodate using traditional estimation methods such as weighted least squares (WLS) where negative contribution estimates are possible. Additionally, source profiles (columns of $\mathbf{\Lambda}$) need to sum to no more than 1. A few methods that incorporate positivity constraints on $\mathbf{\Lambda}$ and $\mathbf{F}$ are discussed in Chapter 2.

1.5 A Bayesian Approach to PSA through Dynamic Linear Models

Temporal correlation and positivity constraints will be addressed directly through an explicit Bayesian formulation of the PSA model in Equation 1.3. While estimation could be done using frequentist methods, Bayesian methods provide a framework which makes estimation of parameters subject to model constraints straightforward. In addition, Bayesian methods allow for the distributions of each of the unknown parameters to be estimated, thus providing more knowledge than a simple point estimate. Uncertainty in estimation can be accounted for through the use of prior

distributions on parameter values. For these reasons, a Bayesian approach to pollution modeling is favorable to frequentist approaches.

Specifically, this thesis proposes the use of a dynamic linear model (DLM) to account for time-varying source profiles. As described in West and Harrison (1997), DLMs are a standard approach to modeling time-varying parameters from a Bayesian perspective. A DLM is expressed through three equations: (1) observation equation, (2) system equation, and (3) initial knowledge equation. The observation equation states the model that is believed to be observed. The observation equation can also be thought of as the likelihood function of the observed random variable. The system equation states how the researcher believes model parameters are evolving over time. Finally, the initial knowledge equation represents the *a priori* knowledge of the researcher about the observed process at state zero.

As a simple example, West and Harrison (1997) write the first-order polynomial DLM as having the observation equation,

$$Y_t \sim \mathrm{N}[\mu_t, V_t], \tag{1.4}$$

where μ_t follows the system equation,

$$\mu_t \sim \mathrm{N}[\mu_{t-1}, W_t], \tag{1.5}$$

with initial information at time 0,

$$(\mu_0 | D_0) \sim \mathrm{N}[m_0, C_0], \tag{1.6}$$

where Y_t represents a Gaussian process defined over time, μ_t is the mean of Y_t and evolves over time as a random walk, V_t is the observational error of Y_t, and W_t is the evolution error. In the initial information equation, D_0 represents all available information known at time 0. In this thesis, the above model is adapted to be more complex by allowing $\mu_t = \mathbf{\Lambda}_t \mathbf{f}_t$.

Chapter 2 discusses some of the previous work in pollution source apportionment. Chapter 3 discusses the research methods to be employed in this thesis, including model selection and description. Prior distributions, model estimation, and results are discussed in Chapter 4. Finally, conclusions and future research opportunities are discussed in Chapter 5.

2. APPROACHES TO POLLUTION SOURCE APPORTIONMENT

2.1 Introduction

Recall that the basic PSA model introduced in Chapter 1 is

$$\underset{P\times N}{\mathbf{Y}} = \underset{P\times K}{\mathbf{\Lambda}}\underset{K\times N}{\mathbf{F}} + \underset{P\times N}{\mathbf{E}}, \tag{2.1}$$

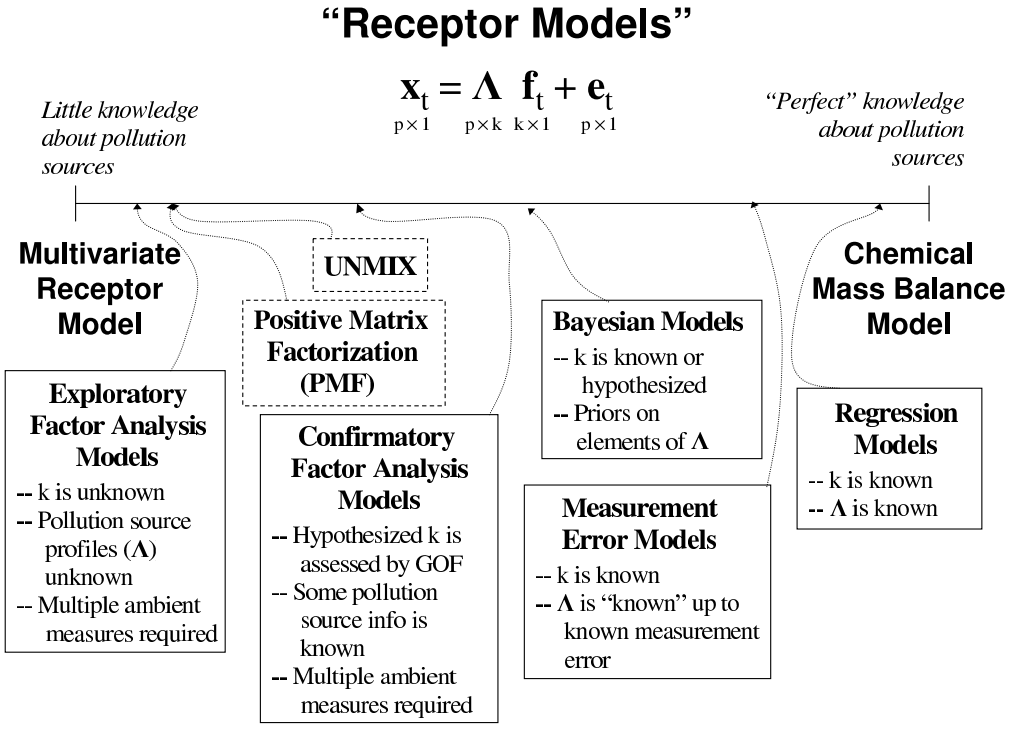
where $\mathbf{Y}$ is a matrix of P chemical species measured at N time periods, $\mathbf{\Lambda}$ is the matrix of pollution source profiles for K different sources, $\mathbf{F}$ is the matrix of source contributions at each of N time periods, and $\mathbf{E}$ is the matrix of model errors. Past approaches to PSA can be classified according to how much prior information a researcher has about $\mathbf{\Lambda}$. Chemical mass balance (CMB) modeling assumes that each $\lambda_{pk} \in \mathbf{\Lambda}$ is fixed and known to within some measurement error u_{pk}. Multivariate receptor modeling assumes that $\mathbf{\Lambda}$ is unknown and needs to be estimated in addition to $\mathbf{F}$. Figure 2.1 shows the continuum of approaches used in PSA based upon available prior information.

In this chapter, CMB modeling is reviewed in Section 2.2, and multivariate receptor modeling is reviewed in Section 2.3. The Bayesian framework of PSA is introduced in Section 2.4 and past approaches to PSA using Bayesian methods are discussed in Section 2.5. Other approaches to PSA are discussed in Section 2.6.

2.2 Chemical Mass Balance Modeling

Chemical mass balance modeling assumes that $\mathbf{\Lambda}$ is fixed and known to within some measurement error u_{pk}. Therefore, in CMB, only $\mathbf{F}$ needs to be estimated. Perhaps the simplest of all CMB models is weighted least squares (WLS). The application of WLS to CMB models was first introduced by Friedlander (1973). Weighted least squares requires assumptions about Equation 2.1 in addition to the previously

Figure 2.1: Pollution Source Apportionment Continuum (Christensen et al. 2006)



stated assumptions in Section 1.3. The necessary assumptions are as follows:

(5) Model errors, e_{pt}, are normal, independent, and identically distributed with variance-covariance matrix $\mathbf{\Sigma}$ for all p, t.

(6) The variance of the model errors, $var(e_{pt}) = \sigma_{pt}^2$, is known for all p, t.

(7) The source profile matrix $\mathbf{\Lambda}$ is known and fixed with no measurement error ($u_{pk} = 0$ for all p, k).

(8) The number of sources does not exceed the number of chemical species.

Under the above assumptions, the best linear unbiased estimate of the source contribution matrix $\mathbf{F}$ is

$$\hat{\mathbf{F}}_{WLS} = (\mathbf{\Lambda'\Sigma^{-1}\Lambda})^{-1}\mathbf{\Lambda'\Sigma^{-1}Y}, \tag{2.2}$$

where $\mathbf{\Lambda}$ is the known source profile matrix, $\mathbf{\Sigma}$ is the variance-covariance matrix of model errors, and $\mathbf{Y}$ is the chemical concentrations matrix. While the use of WLS is practical, the assumptions required for its use rarely hold in practice. For example, violation of Assumption 7 results in the WLS model being susceptible to bias because measurement error is introduced into the source profile matrix (Christensen and Gunst 2004). Additionally, Christensen and Gunst (2004) note that the use of standard errors based on the formula $var(\hat{\mathbf{F}}_{WLS}) = (\mathbf{\Lambda'\Sigma^{-1}\Lambda})^{-1}$ "are in general too small in receptor modeling studies because of the existence of source profile error."

An alternative approach to WLS in estimating $\mathbf{F}$ is to use the effective variance (EV) solution. Effective variance was originally proposed by Watson et al. (1984) and has been implemented in various software programs, including the software used by the USEPA in CMB analysis (Coulter 2000). Effective variance differs from WLS in that it takes into account the measurement uncertainties associated with $\mathbf{\Lambda}$ and y_{pt} in order to estimate the optimal weight in an iteratively reweighted least squares algorithm.

Let u_{pk} be the measurement error for λ_{pk}, let v_{pt} be the measurement error for y_{pt}, and let q_{pt} be the model error for y_{pt}. Under these assumptions $e_{pt} = v_{pt} + q_{pt}$. Let $\tilde{\mathbf{F}}^m$ and $\tilde{\mathbf{\Sigma}}^m$ denote the estimate of $\mathbf{F}$ and $\mathbf{\Sigma}$ at the m^{th} iteration of the EV algorithm. At the m^{th} iteration, $\tilde{\mathbf{F}}^m$ is updated via

$$\tilde{\mathbf{F}}^m = \left[\mathbf{\Lambda}' \left\{\tilde{\mathbf{\Sigma}}^{m-1}\right\}^{-1} \mathbf{\Lambda}\right]^{-1} \mathbf{\Lambda}' \left\{\tilde{\mathbf{\Sigma}}^{m-1}\right\}^{-1} \mathbf{Y},$$

and $\tilde{\mathbf{\Sigma}}^m$ is updated via

$$\tilde{\mathbf{\Sigma}}^m = diag\left(\sigma^2_{vpt} + \sum_{k=1}^{K} \sigma^2_{upk} \left(\tilde{f}^m_{kt}\right)^2\right),$$

where σ^2_{vpt} is the variance of the measurement error of y_{pt} and σ^2_{upk} is the variance of the measurement error of λ_{pk}. After convergence, the final estimates of $\mathbf{F}$ and $\mathbf{\Sigma}$ are denoted by $\hat{\mathbf{F}}_{EV}$ and $\hat{\mathbf{\Sigma}}_{EV}$, respectively.

The method of moments (MoM) solution also employs an iterative algorithm which converges to estimates of source contributions. Let $\mathbf{f}_t$ denote the t^{th} column of $\mathbf{F}$, $\tilde{\mathbf{f}}^m_t$ denote the estimate of $\mathbf{f}_t$ at the m^{th} iteration, and $\tilde{\sigma}^m_{eepp}$ denote the estimate of the p^{th} diagonal element of $\mathbf{\Sigma}$ at the m^{th} iteration of the MoM algorithm. At each iteration, the MoM algorithm updates $\mathbf{f}_t$ through

$$\tilde{\mathbf{f}}^m_t = \left[\sum_{p=1}^{P} \tilde{\sigma}^{m-1}_{eepp} \left(\mathbf{\Lambda}_p \mathbf{\Lambda}'_p - \mathbf{\Sigma}_{up}\right)\right]^{-1} \left[\sum_{p=1}^{P} \tilde{\sigma}^{m-1}_{eepp} \left(\mathbf{\Lambda}_p y_{pt} - \mathbf{\Sigma}_{uvp}\right)\right],$$

and

$$\tilde{\sigma}^m_{eepp} = \sigma^2_{vp} + \sigma^2_q + \sum_{k=1}^{K} \sigma^2_{upk} \left(\tilde{f}^m_{kt}\right)^2, \quad \forall\ p,$$

where σ^2_{vp} and σ^2_{upk} are defined as above, $\mathbf{\Lambda}_p$ is the p^{th} row of $\mathbf{\Lambda}$, y_{pt} is the p^{th} element of $\mathbf{y}_t$, $\mathbf{\Sigma}_{up}$ is the variance-covariance matrix of the measurement error for the p^{th} chemical species across the K sources, $\mathbf{\Sigma}_{uvp}$ is the covariance matrix of u_{pk} with v_{pt}, and σ^2_q is the variance of model errors.

Christensen and Gunst (2004) compare and contrast the four above methods using a variety of data sets generated under different circumstances. They propose

"using the simple WLS estimator for estimation because it is computationally stable and thus yields better average absolute error for scenarios in which the magnitude of source profile errors and measurement errors are large." The EV solution is shown to be equally reasonable to WLS in most cases. They also show that the MoM solution performs poorly when large coefficients of variation are present.

2.3 Multivariate Receptor Modeling

Unlike CMB models, multivariate receptor models assume $\mathbf{\Lambda}$ is unknown and needs to be estimated in addition to $\mathbf{F}$. Exploratory factor analysis (EFA) is one of the earliest multivariate receptor models used in PSA. EFA exploits the correlation between each of the chemical species by factoring the sample variance-covariance matrix ($\mathbf{S}$) into a matrix of source profiles (factor loadings) and source contributions (factors). In EFA, if $\mathbf{S}$ is the covariance matrix of $\mathbf{y}_t$, then $\mathbf{\Lambda}$ is estimated by $\mathbf{CD}^{\frac{1}{2}}$ the matrices $\mathbf{C}$ and $\mathbf{D}$ are the spectral decomposition matrices of $\mathbf{S} = \mathbf{CDC}'$. An estimate of $\mathbf{F}$ can be obtained using WLS, as discussed in Section 2.2.

The main problem with EFA methods in multivariate receptor modeling is the non-uniqueness of source profile estimates. In EFA, the estimated source profile matrix can be multiplied by any orthogonal matrix to obtain different estimates of the source profile and source contribution matrices. For this reason, many statisticians argue that EFA methods are too subjective to the discretion of the researcher. Henry (1987) goes so far as to argue that "factor analysis attempts to get more information out of the data than is really there." In addition, EFA does not account for positivity constraints on λ_{pk} and f_{kt} as well as the constraint that source profiles sum to no more than one.

Positive matrix factorization (see Paatero and Tapper 1994) and Unmix (see Henry 1997) provide good alternatives to the rather subjective procedure of EFA. Both positive matrix factorization (PMF) and Unmix seek to obtain nonnegative

source profile estimates using minimization. In PMF, various settings and algorithms can be used to solve for source profiles and source contributions. The properties of PMF under these settings and algorithms are discussed in Lingwall (2006). Because PMF is currently the most commonly used multivariate receptor model, the DLM proposed in this thesis is compared to PMF.

As noted by Christensen et al. (2006), neither PMF nor Unmix can guarantee a uniquely identified solution without additional constraints on the source profiles. Confirmatory factor analysis (CFA) applies these additional constraints to the traditional factor model by fixing $J > K$ rows of $\mathbf{\Lambda}$ to produce unique estimates of the source profile matrix.

Recent research on multivariate receptor models has been focused on developing an iterated confirmatory factor analysis (ICFA) solution. ICFA reflects both CFA and EFA through application of varying degrees of constraints to each $\lambda_{pkt} \in \mathbf{\Lambda}$ (Christensen et al. 2006). ICFA obtains an initial estimate of the source profile matrix $\mathbf{\Lambda}$ by incorporating *a priori* information in the source profiles. Source profiles with no prior information are estimated using traditional factor analysis methods. At each iteration, $q > K$ randomly chosen rows of the source profile matrix are constrained to be constant, thus guaranteeing a unique solution. The remaining $p - q$ rows are "updated" and the chi-square goodness-of-fit statistic is recalculated. After several iterations, the goodness-of-fit statistic is minimized and the source profiles are scaled to sum to one. Source contributions are then estimated using a linear model where the contribution estimates are constrained to be greater than zero. ICFA methods are discussed in detail in Christensen et al. (2006).

2.4 The Framework of the Bayesian Approach to PSA

Bayesian methods are unique and powerful tools that are useful in solving pollution source apportionment problems. The basis of all Bayesian methods is specifying

prior distributions for each unknown model parameter. The joint posterior distribution of all model parameters can then be solved for through the application of Bayes' Theorem.

In terms of Equation 2.1, let $\mathbf{\Theta} = \{\mathbf{\Lambda}, \mathbf{F}\}$. Bayes' Theorem states that the probability density function of $\mathbf{\Theta}$, given a series of observations $\mathbf{Y} = [\mathbf{y_1}, \ldots, \mathbf{y_N}]$, denoted as $\pi(\mathbf{\Theta}|\mathbf{Y})$, is given by

$$\pi(\mathbf{\Theta}|\mathbf{Y}) = \frac{p(\mathbf{\Theta}, \mathbf{Y})}{\int_\Theta p(\mathbf{\Theta}, \mathbf{Y}) d\mathbf{\Theta}} \tag{2.3}$$

$$= \frac{f(\mathbf{Y}|\mathbf{\Theta})\pi(\mathbf{\Theta})}{\int_\Theta f(\mathbf{Y}|\mathbf{\Theta})\pi(\mathbf{\Theta}) d\mathbf{\Theta}}, \tag{2.4}$$

where $\pi(\mathbf{\Theta})$ denotes the joint prior probability density function and $f(\mathbf{Y}|\mathbf{\Theta})$ denotes the likelihood function. If each $\lambda_{pk} \in \mathbf{\Lambda}$ and $f_{kt} \in \mathbf{F}$ are assumed *a priori* independent, then Equation 2.4 can be rewritten as

$$\pi(\mathbf{\Theta}|\mathbf{Y}) = \frac{f(\mathbf{Y}|\mathbf{\Theta}) \prod_{t=1}^{N} \prod_{p=1}^{P} \prod_{k=1}^{K} \pi(\lambda_{pk})\pi(f_{kt})}{\int_\Theta f(\mathbf{Y}|\mathbf{\Theta})\pi(\mathbf{\Theta}) d\mathbf{\Theta}}.$$

Using Bayesian terminology, $\pi(\mathbf{\Theta}|\mathbf{Y})$ is called the posterior distribution and each $\pi(\lambda_{pk})$ and $\pi(f_{kt})$ are called prior distributions because they represent a researcher's *a priori* knowledge of the unknown parameter.

The difficult aspect of Bayesian statistics is calculating the posterior distribution when the form of the distribution is unknown. In particular, calculating the normalizing constant, $c = \int_\Theta f(\mathbf{Y}|\mathbf{\Theta})\pi(\mathbf{\Theta}) d\mathbf{\Theta}$, is cumbersome because of the involvement of an $R = (P \times K) + (K \times N)$ dimensional integral. One solution to this problem is to obtain draws from the normalized posterior distribution instead of attempting to calculate the normalizing constant c and to use these draws to do posterior analysis. The most common simulation method is called Markov chain Monte Carlo (MCMC) simulation.

2.4.1 Markov Chain Monte Carlo Methods

The general idea of Markov chain Monte Carlo (MCMC) methods is to establish a Markov chain that has a stationary distribution equal to the distribution of interest. A Markov chain has a stationary distribution if the chain is ergodic. The "trick" of MCMC methods is to construct an ergodic Markov chain such that the limiting distribution is the distribution of interest.

The Gibbs sampler is a Markov chain algorithm that, in the limit, produces draws from the joint posterior distribution of all model parameters. Once again, let $\mathbf{\Theta} = \{\mathbf{\Lambda}, \mathbf{F}\}$ and let θ_i denote the i^{th} parameter in $\mathbf{\Theta}$. The Gibbs algorithm to obtain draws from $\pi(\mathbf{\Theta}|\mathbf{Y})$ requires the following steps:

(1) Set starting values for all $\theta_i \in \mathbf{\Theta}$. Collectively, call these starting values $\mathbf{\Theta}^0$.

(2) Set $b = 1$.

(3) For $i = 1, \ldots, R$, where R is the dimensionality of $\mathbf{\Theta}$, draw from $[\theta_i] = f(\theta_i|\theta_1^b, \ldots, \theta_{i-1}^b, \theta_{i+1}^{b-1}, \ldots, \theta_R^{b-1}, \mathbf{Y})$. The distribution $[\theta_i]$ is referred to as the complete conditional distribution of θ_i.

(4) Let $\mathbf{\Theta}^b = \{\theta_1^b, \ldots, \theta_R^b\}$, the draws obtained in step 3.

(5) Repeat steps 3-4 for $b = 2, \ldots, I$, where I is the number of iterations for the Gibbs algorithm.

One important note is that in the successive sampling of $[\theta_i]$, the most recently drawn value of $\theta_{i'}$ is used to draw from $[\theta_i]$. The Gibbs algorithm ensures that as $I \to \infty$, each $\mathbf{\Theta}^b$ is a draw from $\pi(\mathbf{\Theta}|\mathbf{Y})$.

The form of $[\theta_i]$ in the above Gibbs algorithm is often unknown and hence cannot be drawn from directly. The Metropolis and Metropolis-Hastings algorithms are Markov chain algorithms that, in the limit, produce draws from $[\theta_i]$. A full mathematical explanation of the Metropolis-Hastings algorithm is provided in Chib

and Greenberg (1995) but only the steps of the algorithm itself are discussed here. Consider the single parameter $f_{kt} \in \mathbf{F}$ from the PSA model in Equation 2.1. The Metropolis-Hastings steps to generate draws from $[f_{kt}]$ are the following:

(1) Set a starting value for $f_{kt}^{(0)}$.

(2) Set $i = 1$.

(3) Generate a proposal value, X_i, from a candidate distribution $q(f_{kt}^{(i-1)}, X_i)$.

(4) Calculate $\alpha(X_i, f_{kt}^{(i-1)}) = \frac{g(X_i)q(X_i, f_{kt}^{(i-1)})}{g(f_{kt}^{(i-1)})q(f_{kt}^{(i-1)}, X_i)}$ where $g(\cdot)$ is the non-normalized posterior density function for f_{kt}; $q(X_i, f_{kt}^{(i-1)})$ is the candidate distribution evaluated at $f_{kt}^{(i-1)}$; and $q(f_{kt}^{(i-1)}, X_i)$ is the candidate distribution evaluated at X_i.

(5) If $u < \alpha(X_i, f_{kt}^{(i-1)})$ then set $f_{kt}^{(i)} = X_i$; otherwise, set $f_{kt}^{(i)} = f_{kt}^{(i-1)}$ where u is a random variable distributed uniformly on the unit interval.

(6) Repeat steps 2 - 5 for $i = 2, \ldots, I$ where I is the desired number of iterations.

The Metropolis algorithm is a special case of the Metropolis-Hastings algorithm where the candidate distribution is symmetric. Mathematically, in the Metropolis algorithm $q(X_i, f_{kt}^{(i-1)}) = q(f_{kt}^{(i-1)}, X_i)$ so the fraction $\alpha(X_i, f_{kt}^{(i-1)})$ reduces to $\frac{g(X_i)}{g\left(f_{kt}^{(i-1)}\right)}$ and simplifies the acceptance probability. Typically, a normal distribution is used in implementing the Metropolis algorithm. For more information on the Metropolis and Metropolis-Hastings algorithm see Gelman et al. (2004).

The first $B < I$ iterations of the algorithm are called "burn." During the burn-in time, the algorithm converges to the correct stationary distribution. These burn-in iterations are important to keep track of to check convergence and mixing of the algorithm but are typically not included in the final posterior analysis.

One challenge of both the Metropolis and Metropolis-Hastings algorithms is determining the proper standard deviation, σ_c, for the candidate distribution described

in step 3. If σ_c is too small then the algorithm is not able to adequately explore the space of the posterior distribution. On the other hand, if σ_c is too large the algorithm does not effectively explore the space of the posterior distribution. Both of these problems are manifest through the number of iterations a value is kept. If the value obtained at the previous iteration is kept too often, then σ_c is too large. If the algorithm repeatedly keeps the proposal value, σ_c is too small. Both of these problems can often be solved by adjusting σ_c.

2.5 Bayesian Approaches to PSA

As previously mentioned, Bayesian methods are powerful tools in PSA. One example is found in Lingwall (2006) where each element of $\mathbf{\Lambda}$ and each element of $\mathbf{F}$ were assumed to follow a lognormal distribution. The above prior specifications for $\mathbf{\Lambda}$ and $\mathbf{F}$ are logical because the lognormal priors restrict all estimates to be positive. The joint posterior distribution for each source profile and each element of the source contribution matrix were calculated using MCMC techniques.

A very notable example that has particular application to the purpose of this thesis is found in Park et al. (2001). Park et al. proposes the use of an AR(1) model to represent the correlation in $\mathbf{f}_1, \ldots, \mathbf{f}_N$. Specifically, Park et al. propose a DLM with observation equation

$$\mathbf{y}_t = \mathbf{\Lambda}\mathbf{f}_t + \boldsymbol{\eta}_t + \mathbf{e}_t, \quad \mathbf{e}_t \sim N_P(\mathbf{0}, \mathbf{\Sigma}),$$

where $\mathbf{f}_t$ and $\boldsymbol{\eta}_t$ evolve through time according to the system equations

$$\mathbf{f}_t = \boldsymbol{\xi} + (\mathbf{f}_{t-1} - \boldsymbol{\xi})\boldsymbol{\gamma} + \boldsymbol{u}_t \quad , \quad \boldsymbol{u}_t \sim N_K(\mathbf{0}, \mathbf{U}),$$
$$\boldsymbol{\eta}_t = \boldsymbol{\eta}_{t-1}\boldsymbol{\phi} + \boldsymbol{\nu}_t \quad , \quad \boldsymbol{\nu}_t \sim N_P(\mathbf{0}, \mathbf{V}),$$

where $\boldsymbol{\xi}$ is the mean of the multivariate AR(1) process for $\mathbf{f}_t$ and $\boldsymbol{\eta}_t$ is the variability of $\mathbf{y}_t$ correlated in time which is also assumed to follow a multivariate AR(1) process

with mean $\mathbf{0}$. Park et al. uses a "block at a time" Metropolis-Hastings algorithm to sequentially update the parameter vector $\boldsymbol{\Theta} = \{\boldsymbol{\Lambda}, \mathbf{F}, \boldsymbol{\xi}, \boldsymbol{\gamma}, \boldsymbol{\phi}, \boldsymbol{\Sigma}, \boldsymbol{\eta}\}$.

Current research in Bayesian approaches to PSA include the use of the Dirichlet distribution as a prior distribution for each pollution source profile, $\boldsymbol{\lambda}_k$. The Dirichlet distribution[1] is a multivariate distribution that restricts all source profiles to be non-negative and sum to 1. For this reason, the Dirichlet distribution is logical to use as a prior distribution for source profiles. This thesis expands the use of the Dirichlet distribution as a prior distribution by specifying Dirichlet process priors for each source profile.

2.6 Other Approaches to PSA

Many other approaches to PSA exist and are currently being researched. One area that shows promise is the use of a process convolution model to portray the spatial and temporal structure in the day-to-day air samples measured from a location. As described in Calder (2003), a factor analytic model can be convolved with a spatial temporal process according to the observation equation

$$\underset{Q\times P}{\mathbf{Y}_t} = \underset{Q\times M}{\mathbf{K}} \; \underset{M\times K}{\mathbf{X}_t} \; \underset{K\times P}{\boldsymbol{\Lambda}'} + \underset{Q\times P}{\boldsymbol{\nu}},$$

where $\mathbf{X}_t$ evolves according to the system equation

$$\mathbf{X}_t = B(\mathbf{X}_{t-1}), \tag{2.5}$$

where $\mathbf{Y}_t$ represents the concentration of the P species at time t measured at Q different locations, $B(\cdot)$ is some function of past values of $\mathbf{X}_t$, and $\boldsymbol{\nu}$ is the error matrix. In context of the problem being addressed here, $Q = 1$ and $t = 1, \ldots, N$. The process convolution approach has the ability to express temporal correlation through the system equation in Equation 2.5. This approach is still in an early research phase.

[1] For information about the Dirichlet distribution, see Appendix A.

3. FORMULATING THE DIRICHLET PROCESS MODEL

3.1 Introduction

The majority of previous research in PSA has assumed independence among each observation vector $\mathbf{y}_1, \ldots, \mathbf{y}_N$, and many methods developed for PSA are based upon this assumption. However, the existence of temporal correlation among observations $\mathbf{y}_1, \ldots, \mathbf{y}_N$ is commonly ignored in data analytic treatments of PSA. As discussed in Chapter 2, failing to include the correlation structure in the PSA model can lead to incorrect statistical inference. In this chapter, a multivariate receptor model that seeks to incorporate the correlation among observations is proposed. Such a model should more accurately reflect true source profile behavior as well as incorporate the true temporal structure of species concentrations.

Section 3.2 discusses the temporal correlation present in pollution source apportionment problems. The Dirchlet process (DP) model for incorporating this temporal correlation is proposed and discussed in Section 3.3. Model estimation is discussed in Section 3.4 and model evaluation methods are discussed in Section 3.5.

3.2 Temporal Correlation

Because data are collected over consecutive time periods, a certain amount of temporal correlation exists in air pollution data. The main question is where this correlation is to be modeled. Figure 3.1 shows the autocorrelation functions (ACF) for a few chemical species using data collected from the St. Louis supersite.

As illustrated in Figure 3.1, different chemical species exhibit different degrees of autocorrelation. Also, some chemical species, such as tin (Sn), do not seem to exhibit any autocorrelation. In general, a day-to-day correlation seems to exist with

most chemical species. Observations closer in time seem to be more highly correlated than observations farther apart in time.

Part of the temporal correlation that exists in $\mathbf{Y}$ may be attributed to the fact that $\mathbf{\Lambda}$ is not constant and correlated through time. To investigate the behavior of $\mathbf{\Lambda}$ through time, a subset of the St. Louis data set containing measurements on eight chemical species over 640 days was divided into 32 separate data sets of 20 days each. Positive matrix factorization was used to estimate the four-source PSA model of Equation 1.3 for each of the 32 data sets. Figure 3.2 shows how each of the eight chemicals observed change through time for the winter secondary source profile. The dashed line in each graph is the mean composition for the respective element. Figure 3.3 shows the ACF plots of the same winter secondary source profile estimates ($\hat{\mathbf{\Lambda}}$) obtained using PMF.

As Figure 3.3 displays, $\hat{\mathbf{\Lambda}}$ shows varying degrees of autocorrelation across species. In general, the chemicals that make up the majority of the source profile exhibit more autocorrelation than those chemicals that do not constitute the majority of the source profile. For example, Figure 3.2 shows that the chemicals OC, EC, and NO constitute the majority of emissions from the winter secondary source. These chemicals also show the greatest degree of autocorrelation, as shown in Figure 3.3. Based on the PMF estimates from the 32 separate data sets, representing the autocorrelation in $\hat{\mathbf{\Lambda}}$ as more than an AR(1) process is not justified.

Estimates of source contributions across the 32 data sets showed that $\mathbf{F}$ also exhibits decreasing autocorrelation in time. As previously mentioned, Park et al. (2001) models this temporal structure as an AR(1) process. For the purposes of this thesis, the temporal structure of $\mathbf{F}$ is ignored and emphasis is placed on modeling the temporal structure of $\mathbf{\Lambda}$.

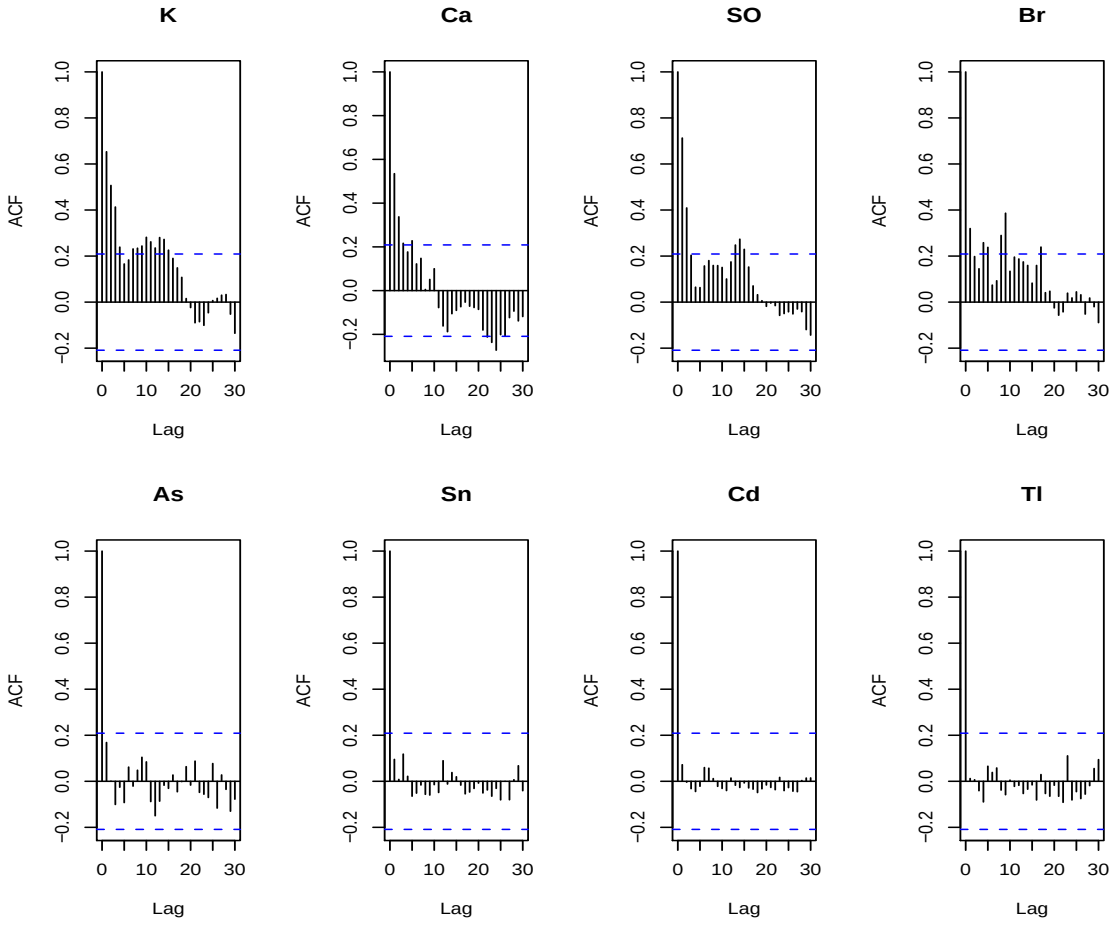


Figure 3.1: ACF plots of Y for the St. Louis Data. Most chemical species exhibit decreasing correlation with time. Some species, however, have little autocorrelation. In general, observations closer in time are more highly correlated than observations farther apart in time.

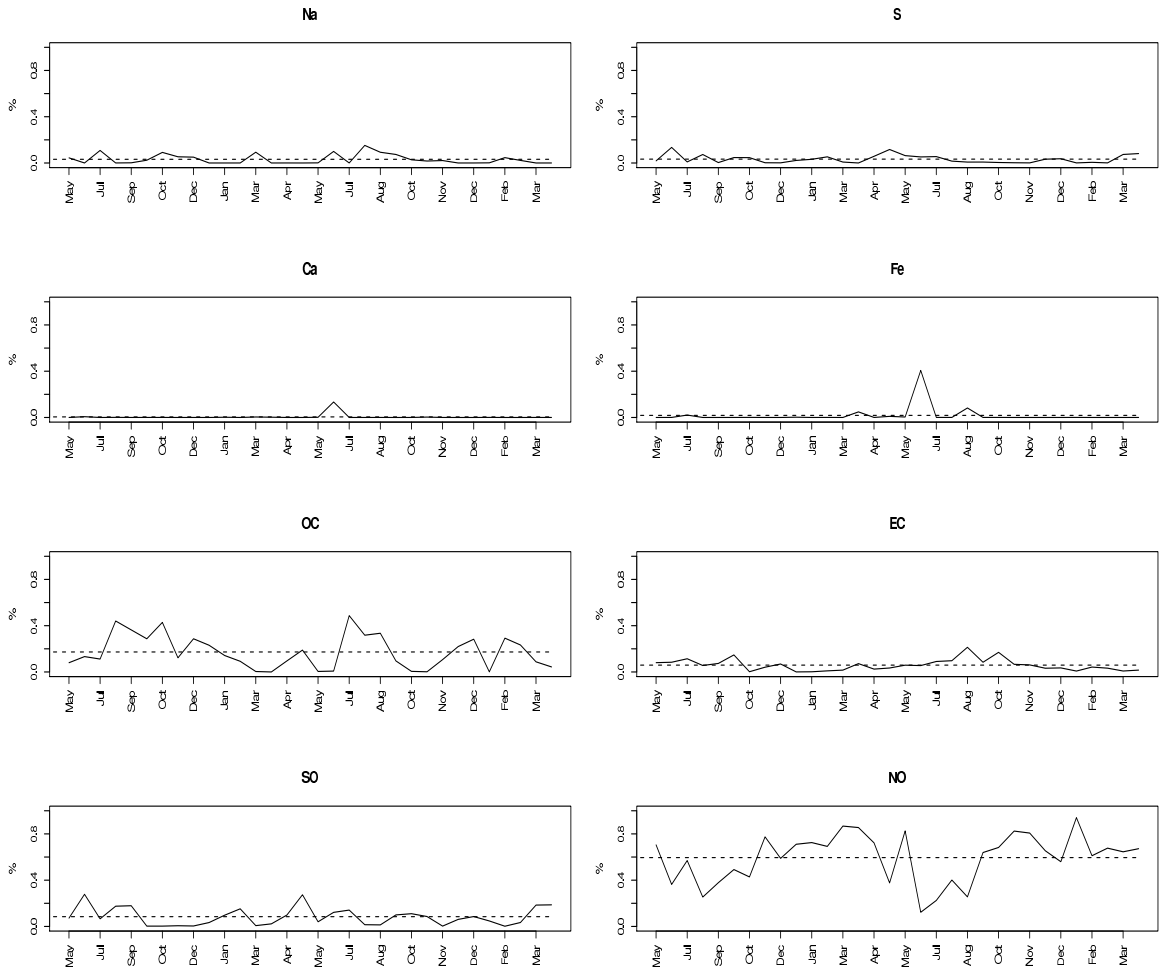


Figure 3.2: PMF estimate of the winter secondary source profile through time. The dashed line is the mean composition averaged over the 32 data sets. Chemicals prevalent in the winter secondary source show larger fluctuations than those chemicals that are not as prevalent.

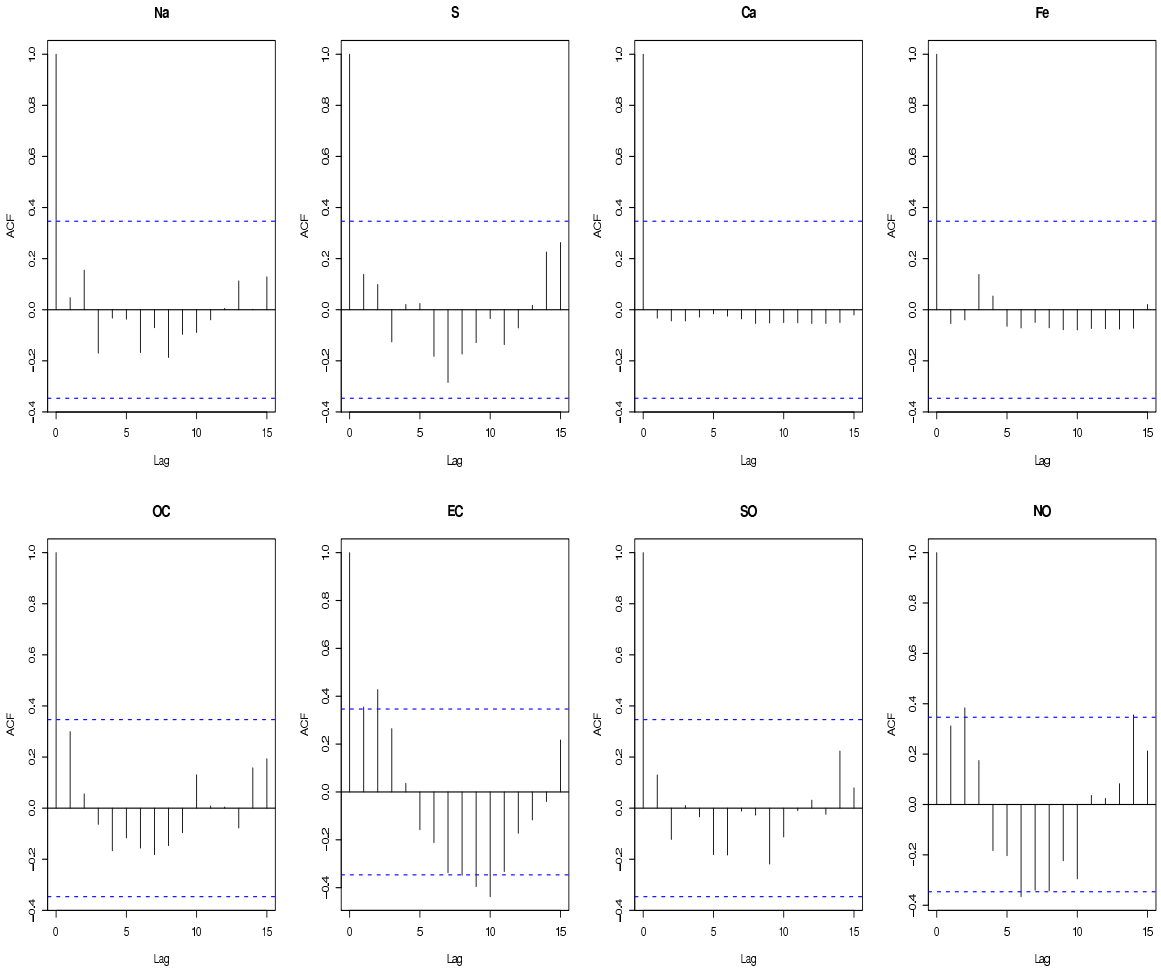


Figure 3.3: ACF plots of the winter secondary source profile. The majority of autocorrelation occurs in chemicals that are prevalent in the winter secondary source.

3.3 The Dirichlet Process Model

In order to incorporate temporal correlation in a PSA model, Equation 1.2 is rewritten in a Dynamic Linear Model (DLM) context. The observation equation is taken to be

$$\mathbf{y}_t \sim \text{LN}[\mathbf{\Lambda}_t \mathbf{f}_t, w_{pt}] \quad t = 1, \ldots, N, \tag{3.1}$$

where each column of $\mathbf{\Lambda}_t$, denoted by $\boldsymbol{\lambda}_{kt}$, follow the system equation

$$\boldsymbol{\lambda}_{kt} \sim \text{DIR}[g_k \boldsymbol{\lambda}_{k(t-1)}] \quad k = 1, \ldots, K \quad t = 1, \ldots, N, \tag{3.2}$$

and the initial information at time 0 is given by

$$\boldsymbol{\lambda}_{k0} \sim \text{DIR}[\mathbf{m}_k] \quad k = 1, \ldots, K, \text{ and} \tag{3.3}$$

$$f_{kt} \sim \text{LN}[a_{kt}, b_{kt}] \quad k = 1, \ldots, K \quad t = 1, \ldots N. \tag{3.4}$$

First, the observation equation (Equation 3.1) states that the concentration of each chemical measured at time t, y_{pt}, is assumed to follow a lognormal distribution with mean $\mathbf{\Lambda}_{pt}\mathbf{f}_t$ and coefficient of variation (CV) w_{pt}, where $\mathbf{\Lambda}_{pt}$ is the p^{th} row of $\mathbf{\Lambda}_t$. The lognormal distribution[1] is both logical and mathematically satisfying as a likelihood for concentrations because each y_{pt} is constrained to be greater than zero. The use of the lognormal distribution for y_{pt} is also consistent with air pollution data. Figure 3.4 displays box plots of 10 randomly selected chemicals from the St. Louis data set. As shown in from Figure 3.4, large positive chemical concentrations are possible. The lognormal distribution accounts for this heavily right-skewed distributional behavior.

Second, the system equation (Equation 3.2) states that each source profile, $\boldsymbol{\lambda}_{kt}$, is assumed to follow a Dirichlet[2] distribution. The Dirichlet distribution is appropriate because it naturally applies the proper constraints to each element of $\boldsymbol{\lambda}_{kt}$ by

[1] For notation of the log-normal distribution, see Appendix A.
[2] For notation of the Dirichlet distribution, see Appendix A.

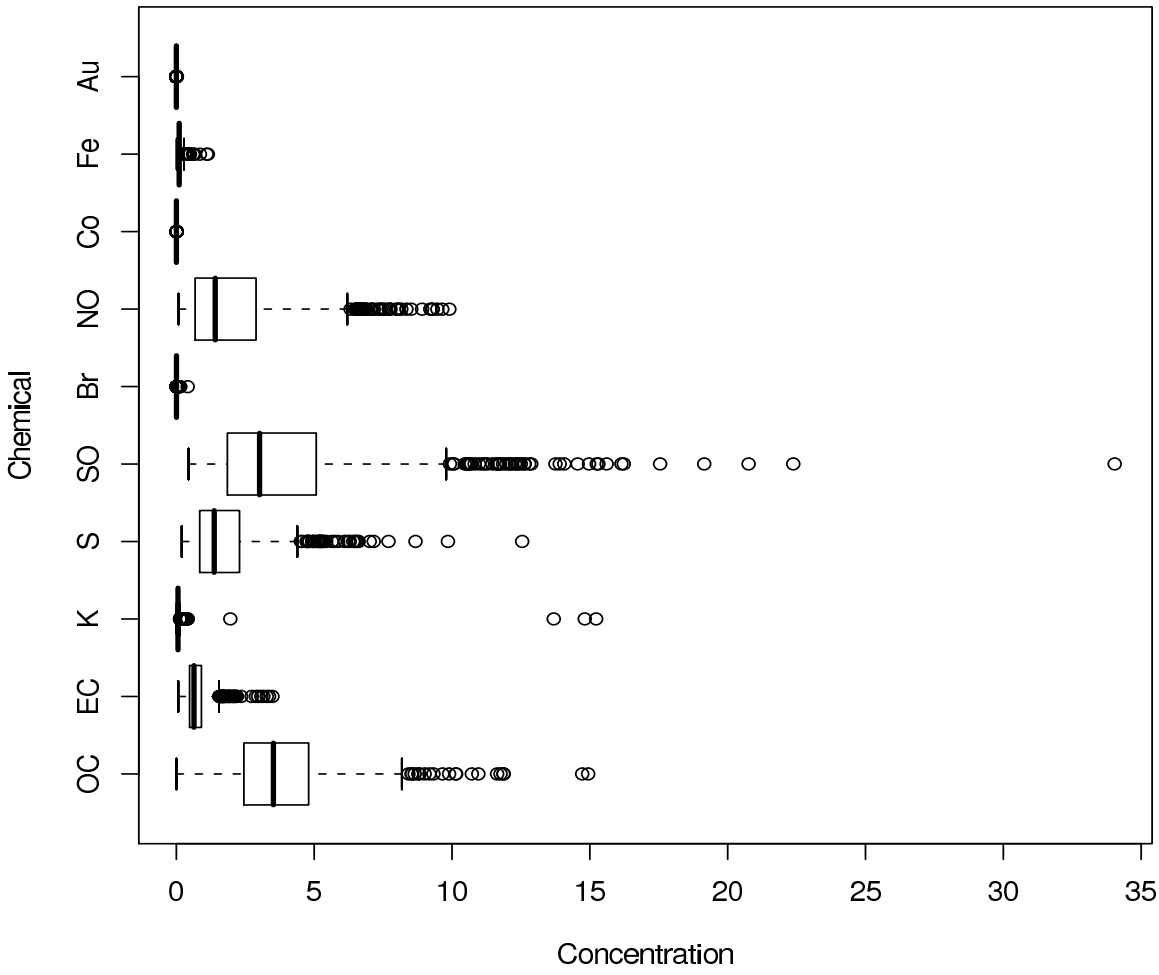


Figure 3.4: Box plots of concentrations for 10 randomly selected chemicals from the St. Louis data set. Chemical species concentrations can exhibit heavily right-skewed distributional behavior.

allowing $0 < \lambda_{pkt} < 1$ and $\sum_{p=1}^{P} \lambda_{pkt} = 1$. The use of the Dirichlet distribution as a process prior for $\boldsymbol{\lambda}_{kt}$ is superior to previous modeling approaches which make the simplifying and dubious assumption that each $\lambda_{pkt} \in \boldsymbol{\lambda}_{kt}$ are mutually independent. The Dirichlet distribution maintains the correct dependence structure as well as the correct constraints for λ_{pkt}.

Furthermore, according to Equation 3.2 $\boldsymbol{\lambda}_{kt}$ evolves as a Dirichlet process where the value of $\boldsymbol{\lambda}_{kt}$ depends only on $\boldsymbol{\lambda}_{k(t-1)}$. For this reason the model proposed in Equations 3.1-3.4 is referred to collectively as the Dirichlet process (DP) model. This dynamic model specification is consistent with the time series DLM (TSDLM) described in West and Harrison (1997) where $\mathbf{G} = g_k\mathbf{I}$. In this process, g_k affects the variance of the Dirichlet process. To illustrate the effect of g_k, consider a single chemical λ_{pkt} from the source profile $\boldsymbol{\lambda}_{kt}$. Under the specified model in Equation 3.2, the expected value of λ_{pkt} is $\lambda_{pk(t-1)}$, with variance $\lambda_{pk(t-1)}(1 - \lambda_{pk(t-1)})/(g_k + 1)$. Therefore, as the value of g_k increases, the variance of the Dirichlet process decreases. Figure 3.5 displays how various levels of g_k effect the OC level of the winter secondary source profile shown in Figure 3.2.

The above proposed DP model makes the assumption that each element of $\boldsymbol{\lambda}_{kt}$ exhibits the same degree of smoothness in the Dirichlet process. As noted in Section 3.2, each $\lambda_{pkt} \in \boldsymbol{\lambda}_{kt}$ exhibits different degrees of autocorrelation. A model that allows for the differences in autocorrelation among each λ_{pkt} is discussed in Section 5.2.

Finally, the initial information equation states that at time 0, $\boldsymbol{\lambda}_{k0}$ follows a Dirichlet distribution with parameter vector $\mathbf{m}_k$. The initial information equation acts as a prior specification for $\boldsymbol{\lambda}_{k0}$ where $\mathbf{m}_k$ is determined by the researcher based upon *a priori* information.

Additionally, the initial information equation states that each $f_{kt} \in \mathbf{F}$ is assumed to be distributed lognormally with expectation a_{kt} and CV b_{kt}. Source contributions could be allowed to follow some sort of time-dependent structure; however,

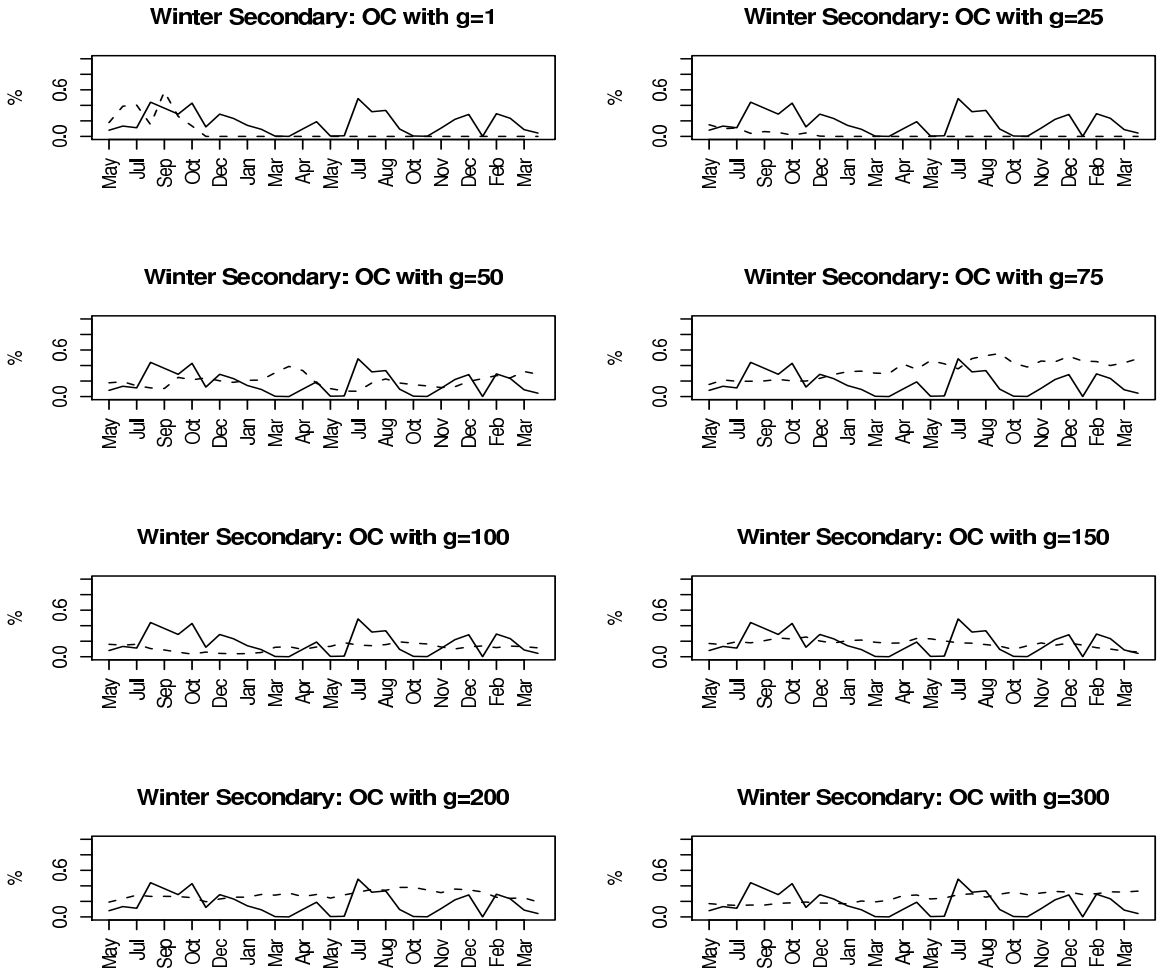


Figure 3.5: Affect of g_k on the OC level of the winter secondary source profile of Figure 3.2. The solid line is the value of OC estimated using PMF on each of the 32 data sets. The dashed line is the value of OC simulated according to Equation 3.2. As g_k increases the variance of the process decreases.

for the purposes of this thesis, no correlation structure is assumed and the focus is placed on modeling the time structure of the source profiles. Despite this fact, the lognormal distribution is once again mathematically satisfying as a distribution for source contributions because each element is constrained to be non-negative.

The entire set of parameters in the above model is $\mathbf{\Theta} = \{g_1, \ldots, g_K, \mathbf{\Lambda_1}, \ldots, \mathbf{\Lambda_N}, f_{11}, \ldots, f_{KN}\}$. The dimensionality of $\mathbf{\Theta} = K + (P \times K \times N) + (K \times N)$. For example, Lingwall (2006) simulates data sets with $K = 5$, $P = 23$, $N = 788$. Applying the above model to these data sets, the dimensionality of $\mathbf{\Theta} = 5 + (23 \times 5 \times 788) + (5 \times 788) = 94,565$. Because the dimensionality of $\mathbf{\Theta}$ increases quickly as N increases, data sets applied to the above model will have a relatively small N.

3.4 Estimation Method

Due to the high-dimensionality of $\mathbf{\Theta}$ and the number of constraints imposed on the model, MCMC simulation is an attractive approach to parameter estimation. Given the series of observations $\mathbf{y_1}, \ldots, \mathbf{y_N}$, the distribution of interest in MCMC simulation is the joint posterior distribution $\pi(\mathbf{\Theta}|\mathbf{y_1}, \ldots, \mathbf{y_N})$ of all model parameters. The full posterior distribution is calculated via Bayes' Theorem as

$$\pi(\mathbf{\Theta}|\mathbf{y_1}, \ldots, \mathbf{y_N}) \propto f(\mathbf{y_1}, \ldots, \mathbf{y_N}|\mathbf{\Theta})\pi(\mathbf{\Theta}), \tag{3.5}$$

where $f(\mathbf{y_1}, \ldots, \mathbf{y_N}|\mathbf{\Theta})$ is the likelihood function and $\pi(\mathbf{\Theta})$ is the joint prior distribution of all model parameters.

As displayed in Equation 3.1, each chemical species is assumed to be distributed lognormal with mean $\mathbf{\Lambda}_{pt}\mathbf{f}_t$ and CV w_{pt}. Therefore, the likelihood for the P chemical species at all N time periods is

$$f(\mathbf{y_1}, \ldots, \mathbf{y_N}|\mathbf{\Theta}) \propto \prod_{t=1}^{N}\prod_{p=1}^{P} \frac{1}{y_{pt}} \exp\left\{-\frac{(\ln(y_{pt}) - \ln(\mathbf{\Lambda_{pt}}\mathbf{f_t}) + \frac{1}{2}\ln(w_{pt}^2+1))^2}{2\ln(w_{pt}^2+1)}\right\}. \tag{3.6}$$

For this thesis, g_k and $\mathbf{F}$ are assumed *a priori* independent for all k and t. However, $\mathbf{\Lambda}_t$ is independent of $\mathbf{F}$ but dependent on $\mathbf{\Lambda}_{t-1}$ and g_k. Under these assumptions, the

full joint prior distribution is

$$\pi(\boldsymbol{\Theta}) = \prod_{t=1}^{N}\prod_{k=1}^{K} \pi(\boldsymbol{\lambda}_{kt}|\boldsymbol{\lambda}_{k(t-1)}, g_k)\pi(f_{kt})\pi(g_k), \tag{3.7}$$

where each g_k are assumed independent and follow a normal distribution.

The MCMC simulation will contain the following steps to sample from the full posterior distribution:

(1) Update $g_1, \ldots, g_K$.

(2) Update $\boldsymbol{\Lambda}_1, \ldots, \boldsymbol{\Lambda}_N$ by sequentially updating $\boldsymbol{\lambda}_{1,1}, \ldots, \boldsymbol{\lambda}_{K,N}$.

(3) Update $f_{1,1}, \ldots, f_{K,N}$.

To update each f_{kt} and g_k, a Metropolis algorithm will be used with proposals generated from a normal candidate distribution. For $\boldsymbol{\Lambda}_t$, a Metropolis-Hastings algorithm will be used. For the MCMC algorithm, a total of $I = 50,000$ iterations will be used, with the first $B = 10,000$ being used as the burn-in phase.

3.5 Model Evaluation Methods

To evaluate the performace of the DP model, the parameter estimates will be compared to parameter estimates using PMF. Because PMF is currently the most commonly used approach to multivariate receptor modeling, the performance of PMF provides a good benchmark from which to evaluate the DP model. Comparison methods between the DP model and PMF are outlined below.

3.5.1 Simulating Data Sets

The above model will be evaluated based on various simulated data sets with $P = 44$, $K = 9$ and $N = 50$. Thus, the total number of model parameters estimated is $9 + (9 \times 44 \times 50) + (9 \times 50) = 20,259$ for each simulated data set.

To generate the chemical weights, y_{pt} will be drawn from a lognormal distribution with mean $\mathbf{\Lambda}_{pt}\mathbf{f}_t$ and CV w_{pt} where $w_{pt} \in \{.2, .8\}$. To do so, values of $\mathbf{\Lambda}_t$ will be generated as draws from a Dirichlet distribution with $g_k \in \{100, 250\}$.

In order to specify values of $\boldsymbol{\lambda}_{k0}$ in generating $\mathbf{\Lambda}_t$, PMF was performed on the data obtained from the St. Louis supersite. Table 3.1 displays the PMF estimate of λ_{pkt} in each of the nine source profiles. These source profile estimates will be used as the values of $\boldsymbol{\lambda}_{k0}$ and the subsequent values of $\boldsymbol{\lambda}_{kt}$ will be generated according to Equation 3.2. Using the output of PMF to indicate values for $\boldsymbol{\lambda}_{k0}$ is preferable to personally specifying $\boldsymbol{\lambda}_{k0}$ because PMF estimates will be more realistic. Values for $f_{kt} \in \mathbf{F}$ will also be obtained from the same PMF output used to obtain values of $\boldsymbol{\lambda}_{k0}$.

Each of the generated values of $\mathbf{\Lambda}_t$, as well as the specified values for g_k and $\mathbf{F}$, will be treated as the known parameter values. Estimates of these parameters obtained from MCMC simulation will be compared to the "true" values to evaluate model fit as discussed in Section 3.5.2. A 2^2 factorial design will be used with 15 data sets being generated at each specification of g_k and w_{pt}.

In addition to evaluating the performance of the DP model under time varying-profiles, evaluating the performance of the DP model when $\boldsymbol{\lambda}_{kt} = \boldsymbol{\lambda}_k$ for all t is also of interest. For this reason, 15 data sets will be simulated using the values of $\boldsymbol{\lambda}_{k0}$ given in Table 3.1 as the constant source profile matrix for each of $w_{pt} \in \{.2, .8\}$. Both the DP model and PMF will be applied to these 30 data sets and the performance of each method will be compared using the same criterion as outlined in Section 3.5.2. Table 3.2 summarizes how each data set will be simulated.

Table 3.1: Values of $\mathbf{\Lambda}_0$ used in simulating data sets.

Chemical Name	Source 1	Source 2	Source 3	Source 4	Source 5	Source 6	Source 7	Source 8	Source 9
Na	0.015	0.014	0.032	0.002	0.007	0.018	0.007	0.020	0.003
Mg	0.002	0.004	0.005	0.001	0.015	0.016	0.012	0.011	0.014
Al	$< .001$	0.001	$< .001$	0.002	0.085	0.002	0.021	0.003	0.008
Si	$< .001$	$< .001$	0.009	0.011	0.253	0.003	0.002	0.002	0.022
Ph	$< .001$	$< .001$	$< .001$	$< .001$	$< .001$	0.002	$< .001$	0.075	$< .001$
S	0.249	0.025	0.020	0.038	0.131	0.020	0.183	0.017	0.031
Cl	$< .001$	0.003	$< .001$	0.001	$< .001$	0.049	0.013	0.003	0.047
K	0.001	0.003	0.003	0.001	0.020	0.008	0.494	0.006	0.009
Ca	$< .001$	$< .001$	0.046	0.007	0.032	0.017	0.002	0.007	0.032
Ti	$< .001$	$< .001$	$< .001$	$< .001$	0.005	$< .001$	$< .001$	0.001	0.001
V	$< .001$	$< .001$	$< .001$	$< .001$	$< .001$	$< .001$	$< .001$	$< .001$	0.001
Cr	$< .001$	$< .001$	$< .001$	$< .001$	$< .001$	$< .001$	$< .001$	0.001	$< .001$
Mn	$< .001$	$< .001$	$< .001$	0.003	$< .001$	0.001	$< .001$	0.001	0.001
Fe	0.001	$< .001$	0.006	0.125	0.026	0.014	0.004	0.009	0.012
Co	$< .001$	$< .001$	$< .001$	0.001	$< .001$	$< .001$	$< .001$	$< .001$	$< .001$
Ni	$< .001$	$< .001$	$< .001$	$< .001$	$< .001$	$< .001$	$< .001$	$< .001$	$< .001$
Cu	$< .001$	$< .001$	$< .001$	$< .001$	$< .001$	$< .001$	0.005	0.156	0.004
Zn	0.001	$< .001$	$< .001$	$< .001$	0.001	0.133	0.003	0.011	0.019
Ga	$< .001$	$< .001$	$< .001$	$< .001$	$< .001$	$< .001$	$< .001$	$< .001$	$< .001$
As	$< .001$	$< .001$	$< .001$	$< .001$	$< .001$	0.001	$< .001$	0.002	0.001
Se	$< .001$	$< .001$	$< .001$	$< .001$	$< .001$	$< .001$	$< .001$	$< .001$	$< .001$
Br	$< .001$	$< .001$	0.001	$< .001$	$< .001$	$< .001$	$< .001$	0.001	$< .001$
Rb	$< .001$	$< .001$	$< .001$	$< .001$	$< .001$	$< .001$	$< .001$	$< .001$	$< .001$
Sr	$< .001$	$< .001$	$< .001$	$< .001$	$< .001$	$< .001$	0.009	$< .001$	$< .001$
Y	$< .001$	$< .001$	$< .001$	$< .001$	$< .001$	$< .001$	$< .001$	$< .001$	$< .001$
Zr	$< .001$	$< .001$	$< .001$	$< .001$	$< .001$	$< .001$	$< .001$	$< .001$	$< .001$
Mo	$< .001$	$< .001$	$< .001$	$< .001$	$< .001$	$< .001$	$< .001$	$< .001$	$< .001$
Pd	$< .001$	$< .001$	$< .001$	$< .001$	$< .001$	$< .001$	$< .001$	$< .001$	$< .001$
Ag	$< .001$	$< .001$	$< .001$	$< .001$	$< .001$	$< .001$	$< .001$	0.001	0.001
Cd	$< .001$	$< .001$	$< .001$	$< .001$	$< .001$	0.001	$< .001$	$< .001$	0.007
In	$< .001$	$< .001$	$< .001$	$< .001$	$< .001$	$< .001$	$< .001$	$< .001$	$< .001$
Sn	$< .001$	$< .001$	$< .001$	$< .001$	$< .001$	0.001	$< .001$	0.002	0.004
Sb	$< .001$	$< .001$	$< .001$	0.001	$< .001$	$< .001$	0.001	0.001	0.002
Ba	$< .001$	0.001	$< .001$	0.013	$< .001$	$< .001$	0.040	0.001	0.001
La	$< .001$	$< .001$	0.001	0.001	0.002	0.003	$< .001$	0.002	0.002
Au	$< .001$	$< .001$	$< .001$	$< .001$	$< .001$	$< .001$	$< .001$	$< .001$	$< .001$
Hg	$< .001$	$< .001$	$< .001$	$< .001$	$< .001$	$< .001$	$< .001$	$< .001$	$< .001$
Tl	$< .001$	$< .001$	$< .001$	$< .001$	$< .001$	$< .001$	$< .001$	$< .001$	0.002
Pb	$< .001$	$< .001$	$< .001$	$< .001$	0.001	$< .001$	0.003	0.004	0.133
U	$< .001$	$< .001$	$< .001$	$< .001$	$< .001$	$< .001$	$< .001$	$< .001$	$< .001$
OC	0.142	0.167	0.675	0.452	0.132	0.609	0.127	0.591	0.302
EC	0.004	0.040	0.168	0.219	0.002	0.088	0.020	0.026	0.255
SO	0.561	0.057	0.030	0.116	0.269	0.006	0.042	0.046	0.070
NO	0.023	0.682	0.001	0.006	0.017	0.005	0.010	0.001	0.012

Table 3.2: Specifications for simulating each data set.

Data Set	$\mathbf{g}_k$	w_{pt}
1-15	100	.2
16-30	100	.8
31-45	250	.2
46-60	250	.8
61-75	∞	.2
76-90	∞	.8

3.5.2 Model Comparison Criterion

Median absolute error (MAE) will be used to compare the estimates of $\mathbf{\Lambda}_t$ and $\mathbf{F}$ to estimates obtained using PMF. Median absolute error for $\mathbf{\Lambda}_t$ is calculated as

$$\text{MAE}_\Lambda = \sum_{p=1}^{P} |\lambda_{pkt} - \hat{\lambda}_{pkt}|, \tag{3.8}$$

where λ_{pkt} is the "true" value of λ_{pkt} and $\hat{\lambda}_{pkt}$ is the median of the 40,000 post-burn-in draws from the posterior distribution obtained using MCMC. Median absolute error will be similarly calculated for $\mathbf{F}$ as

$$\text{MAE}_F = \sum_{t=1}^{N} |f_{kt} - \hat{f}_{kt}|, \tag{3.9}$$

where f_{kt} is the "true" value of f_{kt} and $\hat{f}_{kt}$ is the median of the 40,000 post-burn-in draws from the posterior distribution. MAE is obviously strictly positive where smaller values indicate good performance and large values indicate poor performance.

4. EVALUATION OF THE DIRCHLET PROCESS MODEL

4.1 Introduction

As discussed in Chapter 3, a Dirichlet Process (DP) DLM is proposed to account for time-varying source profiles in the basic pollution source apportionment model. The observation equation for the DP model is

$$\mathbf{y}_t \sim \text{LN}[\mathbf{\Lambda}_t \mathbf{f}_t, \mathbf{w}_t] \quad t = 1, \ldots, N, \tag{4.1}$$

where the columns of $\mathbf{\Lambda}_t$, denoted by $\boldsymbol{\lambda}_{kt}$, evolve through time according to the system equation

$$\boldsymbol{\lambda}_{kt} \sim \text{DIR}[g_k \boldsymbol{\lambda}_{k(t-1)}] \quad k = 1, \ldots, K, \quad t = 1, \ldots, N, \tag{4.2}$$

and the initial information at time 0 is given by

$$\boldsymbol{\lambda}_{k0} \sim \text{DIR}[\mathbf{m}_k] \quad k = 1, \ldots, K, \text{ and} \tag{4.3}$$

$$f_{kt} \sim \text{LN}[a_{kt}, b_{kt}] \quad k = 1, \ldots, K \quad t = 1, \ldots N, \tag{4.4}$$

where $\mathbf{y}_t$ is the vector of P chemical species measured at time t, $\boldsymbol{\lambda}_{kt}$ is the source profile vector for the k^{th} source at time t, and f_{kt} is the k^{th} source contribution at time t.

Parameter estimation for the above Dirichlet process (DP) model presented in Equations 4.1-4.4 was done using successive sampling MCMC methods. Section 4.2 discusses the prior and complete conditional distributions for each model parameter. Section 4.3 discusses the details of the MCMC algorithm used for parameter estimation and Section 4.4 discusses the results of the MCMC algorithm for the simulated data sets using the criterion described in Section 3.5.2.

4.2 Prior and Complete Conditional Distributions

Let $\boldsymbol{\Theta}$ denote the parameter space for the DP model. Use of successive sampling MCMC methods requires the specification of prior distributions and calculation of the complete conditional distributions, denoted by $[\theta_i]$, for all $\theta_i \in \boldsymbol{\Theta}$. In order to sample from the complete joint posterior density function $\pi(\boldsymbol{\Theta}|\mathbf{Y})$, successive samples need to be taken from each complete conditional distribution.

The complete conditional distribution for a given parameter (or parameter vector), $\boldsymbol{\theta}_i \in \boldsymbol{\Theta}$, is defined to be

$$[\boldsymbol{\theta}_i] = p(\boldsymbol{\theta}_i|\boldsymbol{\Theta}_{-i}, \mathbf{Y}) \propto f(\mathbf{Y}|\boldsymbol{\Theta})\pi(\boldsymbol{\theta}_i), \tag{4.5}$$

where $\boldsymbol{\Theta}_{-i}$ denotes all parameters in $\boldsymbol{\Theta}$ not including $\boldsymbol{\theta}_i$, $\mathbf{Y}$ represents the matrix of PM measurements, $f(\mathbf{Y}|\boldsymbol{\Theta})$ is the likelihood function, and $\pi(\boldsymbol{\theta}_i)$ is the joint prior distribution for the parameter vector $\boldsymbol{\theta}_i$. Prior and complete conditional distributions need to be specified for all $\boldsymbol{\theta}_i \in \boldsymbol{\Theta}$.

Equation 4.4 states that the prior distribution for all $f_{kt} \in \mathbf{F}$ is lognormal with mean a_{kt} and CV b_{kt}. Each $f_{kt} \in \mathbf{F}$ is assumed *a priori* independent and therefore the complete conditional for each f_{kt} can be written as

$$[f_{kt}] \propto f(\mathbf{y}_1, \ldots, \mathbf{y}_N|\boldsymbol{\Theta})\pi(f_{kt}),$$

where $f(\mathbf{y_1}, \ldots, \mathbf{y_N}|\boldsymbol{\Theta})$ is given in Equation 3.6 and

$$\pi(f_{kt}) \propto \frac{1}{f_{kt}} \exp\left\{-\frac{(\ln(f_{kt}) - \ln(a_{kt}) + \frac{1}{2}\ln(b_{kt}^2+1))^2}{2\ln(b_{kt}^2+1)}\right\}.$$

Source contribution behavior is such that large contributions are possible. Prior specifications of a_{kt} and b_{kt} for all f_{kt} need to allow for these potentially large contributions. According to Phalen (2002), source contributions cannot exceed 65 on any given day and may not exceed 15 as a yearly average. Allowing $a_{kt} = 20$ and $b_{kt} = 1$ for all $f_{kt} \in \mathbf{F}$ allows for large contributions, while the majority of the probability mass is

less than 10. This prior specification for all $f_{kt} \in \mathbf{F}$ is consistent with the information provided in Phalen (2002).

The prior distribution for $\boldsymbol{\lambda}_{kt}$, as stated in Equation 4.2, is Dirichlet and dependent upon $\boldsymbol{\lambda}_{k(t-1)}$ and g_k. Additionally, the prior distribution for $\boldsymbol{\lambda}_{k(t+1)}$ is dependent upon $\boldsymbol{\lambda}_{kt}$. Therefore, the complete conditional distribution for $\boldsymbol{\lambda}_{kt}$ must also include part of the prior distribution for $\boldsymbol{\lambda}_{k,t+1}$. Under these conditions

$$[\boldsymbol{\lambda}_{kt}] \propto f(\mathbf{y}_1 \ldots, \mathbf{y}_N|\boldsymbol{\Theta}) \frac{\Gamma(g_k)}{\prod_{p=1}^{P} \Gamma(g_k \lambda_{pkt})} \prod_{p=1}^{P} \lambda_{pkt}^{g_k \lambda_{pk(t-1)}} \lambda_{pk(t+1)}^{g_k \lambda_{pkt}}.$$

Allowing $\boldsymbol{\lambda}_{kt}$ to follow a Dirichlet process through time poses a small problem in implementing the MCMC algorithm to update $\boldsymbol{\lambda}_{k1}$ and $\boldsymbol{\lambda}_{k50}$. Under this prior specification, the complete conditional distributions for $\boldsymbol{\lambda}_{k1}$ and $\boldsymbol{\lambda}_{k50}$ are not fully specified because $\boldsymbol{\lambda}_{k0}$ and $\boldsymbol{\lambda}_{k51}$ are unknown and will not be estimated in the MCMC algorithm. The solution to this problem is discussed in detail in Section 4.3.

The variable g_k is known as a hyperparameter and is not a parameter resulting from the likelihood function $f(\mathbf{y_1}, \ldots, \mathbf{y_N}|\boldsymbol{\Theta})$. Rather, g_k arises from the prior specification of $\boldsymbol{\lambda}_{kt}$. Therefore, the complete conditional distribution for g_k is

$$[g_k] \propto \prod_{t=1}^{N} \pi(\boldsymbol{\lambda}_{kt}|g_k)\pi(g_k), \tag{4.6}$$

where $\pi(\boldsymbol{\lambda}_{kt}|g_k)$ represents the Dirichlet density function and $\pi(g_k)$ represents the prior distribution for g_k. For simplicity, g_k is assumed to follow a normal distribution with mean 150 and variance 2500 for all k. The use of the normal distribution as a prior distribution for g_k is logical in that the normal distribution has density on the interval (-∞, ∞) and thus allows the data, not the prior distribution, to dictate the estimate of g_k.

4.3 Details of the MCMC Algorithm

For computational efficiency and speed in updating all 20,259 parameters, the MCMC algorithm was coded using MATLAB. At each iteration of the MCMC al-

gorithm, all g_k and f_{kt} are updated individually using a Metropolis step and each profile vector $\boldsymbol{\lambda}_{kt}$ is updated using Metropolis-Hastings. Approximately 8 hours of computation time was required to run 50,000 MCMC iterations for each data set, requiring $8 \times 60 = 720$ hours $= 30$ days of computation time.

4.3.1 Dealing with Tuning Parameters

As noted in Section 2.4.1, one difficulty in using Metropolis or Metropolis-Hastings steps is the choice of standard deviation, σ_c, for the candidate distributions. Due to the large dimensionality of the proposed model, manually altering each proposal distribution would be very time consuming. The MCMC algorithm keeps track of the number of draws accepted or rejected and updates the scale of the candidate distribution accordingly. For example, if less than 20% of the proposals from Metropolis-Hastings were being accepted, the algorithm would decrease the spread of the candidate distribution by a fixed amount. The spread of the candidate distribution is updated during the burn-in period but held constant otherwise.

4.3.2 Updating $\boldsymbol{\lambda}_{k1}$ and $\boldsymbol{\lambda}_{k50}$

As previously mentioned, one difficulty of the MCMC algorithm is updating $\boldsymbol{\lambda}_{k1}$ and $\boldsymbol{\lambda}_{k50}$. The complete conditional distribution for $\boldsymbol{\lambda}_{k1}$ requires that $\boldsymbol{\lambda}_{k0}$ be known or estimated. Similarly, the complete conditional for $\boldsymbol{\lambda}_{k50}$ requires that $\boldsymbol{\lambda}_{k51}$ be known or estimated.

To solve this problem for $\boldsymbol{\lambda}_{k1}$, several approaches can be taken. For the simulated data sets described in Section 3.5.1, the value of the source profile $\boldsymbol{\lambda}_{k0}$ is known. A rather naive solution to this problem would be to fix $\boldsymbol{\lambda}_{k0}$ at the known parameter value. As can be seen from the first row in Figure 4.1, this approach allows $\boldsymbol{\lambda}_{k1}$ to be estimated accurately. However, using the known value of $\boldsymbol{\lambda}_{k0}$ is not the best approach because $\boldsymbol{\lambda}_{k0}$ is typically not known and using the known value would underestimate

the uncertainty about the value of $\boldsymbol{\lambda}_{k1}$. An alternative approach would be to fix every element of $\boldsymbol{\lambda}_{k0}$ at $1/P$ to give equal weight to each element, thus accurately portraying prior knowledge about $\boldsymbol{\lambda}_{k0}$. However, when every element of $\boldsymbol{\lambda}_{k0}$ is fixed at $1/P$, the MCMC is restricted in its estimate of $\boldsymbol{\lambda}_{kt}$ for t near 0 to compensate for the fixed value of $\boldsymbol{\lambda}_{k0}$. This restriction in estimation of $\boldsymbol{\lambda}_{kt}$ for t near 0 is demonstrated by the left-tail behavior in the second row of Figure 4.1. The approach used in this thesis is to draw $\boldsymbol{\lambda}_{k0}$ from a Dirichlet distribution with parameter vector $\alpha = c/P$, where c is a known fixed constant. This approach accurately reflects the uncertainty about $\boldsymbol{\lambda}_{k0}$ while avoiding the left-tail behavior present when each element of $\boldsymbol{\lambda}_{k0}$ is fixed at $1/P$. Figure 4.1 shows a time plot of a single source profile estimate under each of the three above approaches.

The solution to the problem in updating $\boldsymbol{\lambda}_{k50}$ is similar to that of the solution in updating $\boldsymbol{\lambda}_{k1}$. Because the prior distribution for $\boldsymbol{\lambda}_{k51}$ is defined by g_k and $\boldsymbol{\lambda}_{k50}$, a draw from the prior distribution for $\boldsymbol{\lambda}_{k51}$ using the current value of g_k and $\boldsymbol{\lambda}_{k50}$ will be used in calculating the acceptance probability for $\boldsymbol{\lambda}_{k50}$ in the MCMC algorithm.

4.3.3 Assessing Convergence

Figures 4.2 and 4.3 show successive draws of f_{kt} and g_k as obtained during the MCMC algorithm. Both of these plots support the hypothesis that the MCMC algorithm obtained convergence due to the random scatter of succesive draws. Additionally, Figures 4.2 and 4.3 show that the algorithm achieved proper acceptance ratios.

Figure 4.4 displays successive draws for the primary elements of $\boldsymbol{\lambda}_{kt}$ as obtained from MCMC sampling methods. As is shown in Figure 4.4, for these large elements of $\boldsymbol{\lambda}_{kt}$, the MCMC algorithm seems to have converged with acceptable mixing properties. The slight lack of mixing shown by Figure 4.4 in contrast to the proper mixing properties shown in Figures 4.2 and 4.3 arises due to the correlation structure among

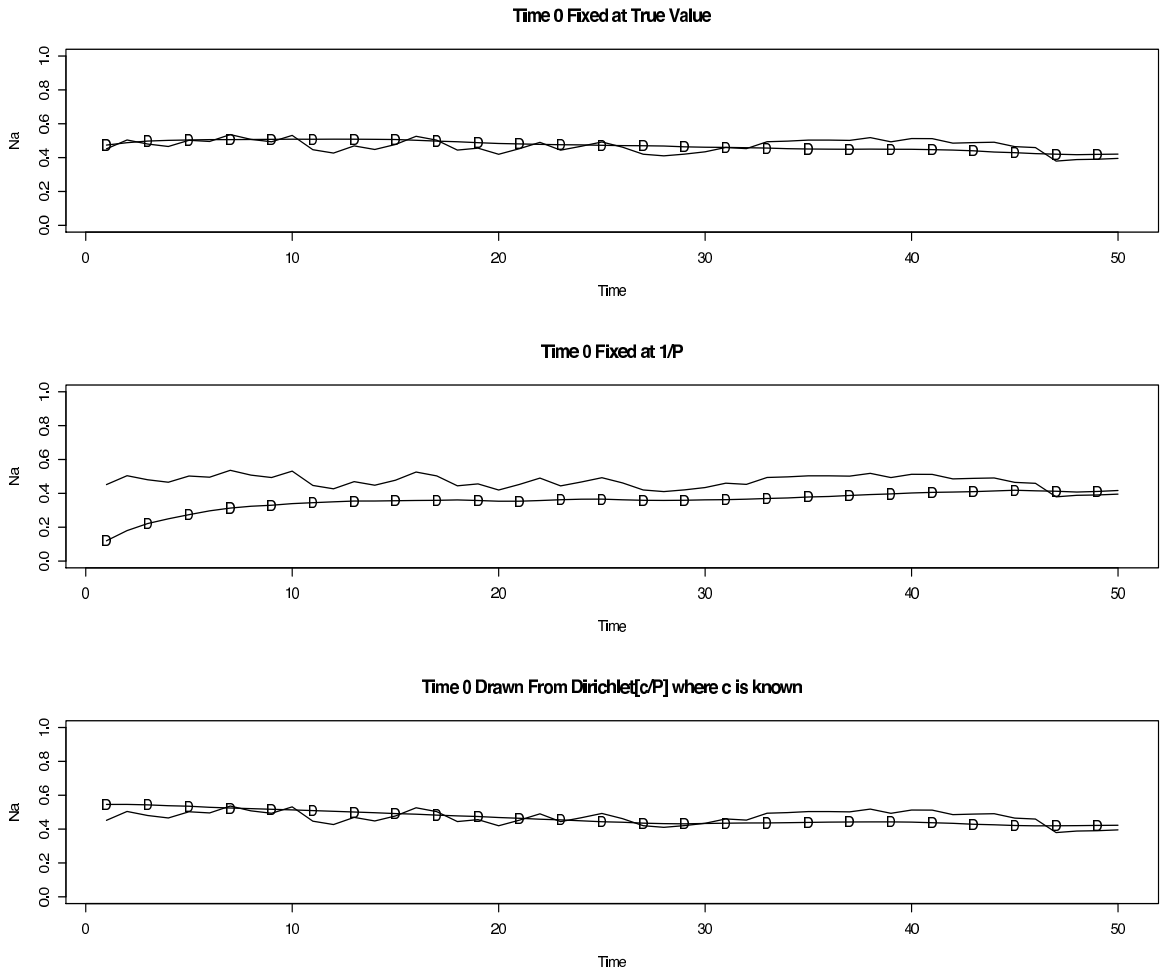


Figure 4.1: Comparison of approaches to the problem of updating $\boldsymbol{\lambda}_{k1}$ when $\boldsymbol{\lambda}_{k0}$ is unknown. The solid line is the true value of $\boldsymbol{\lambda}_{kt}$ and the solid line with inserted "D" is the DP model estimate. Using the true value of $\boldsymbol{\lambda}_{k0}$ (first row) underestimates the uncertainty associated with $\boldsymbol{\lambda}_{k1}$. Fixing each element of $\boldsymbol{\lambda}_{k0}$ at $1/P$ (second row) results in undesirable left-tail behavior. Drawing $\boldsymbol{\lambda}_{k0}$ from Dir$[c/P]$ accurately quantifies the uncertainty about $\boldsymbol{\lambda}_{k0}$ while maintaining good tail behavior.

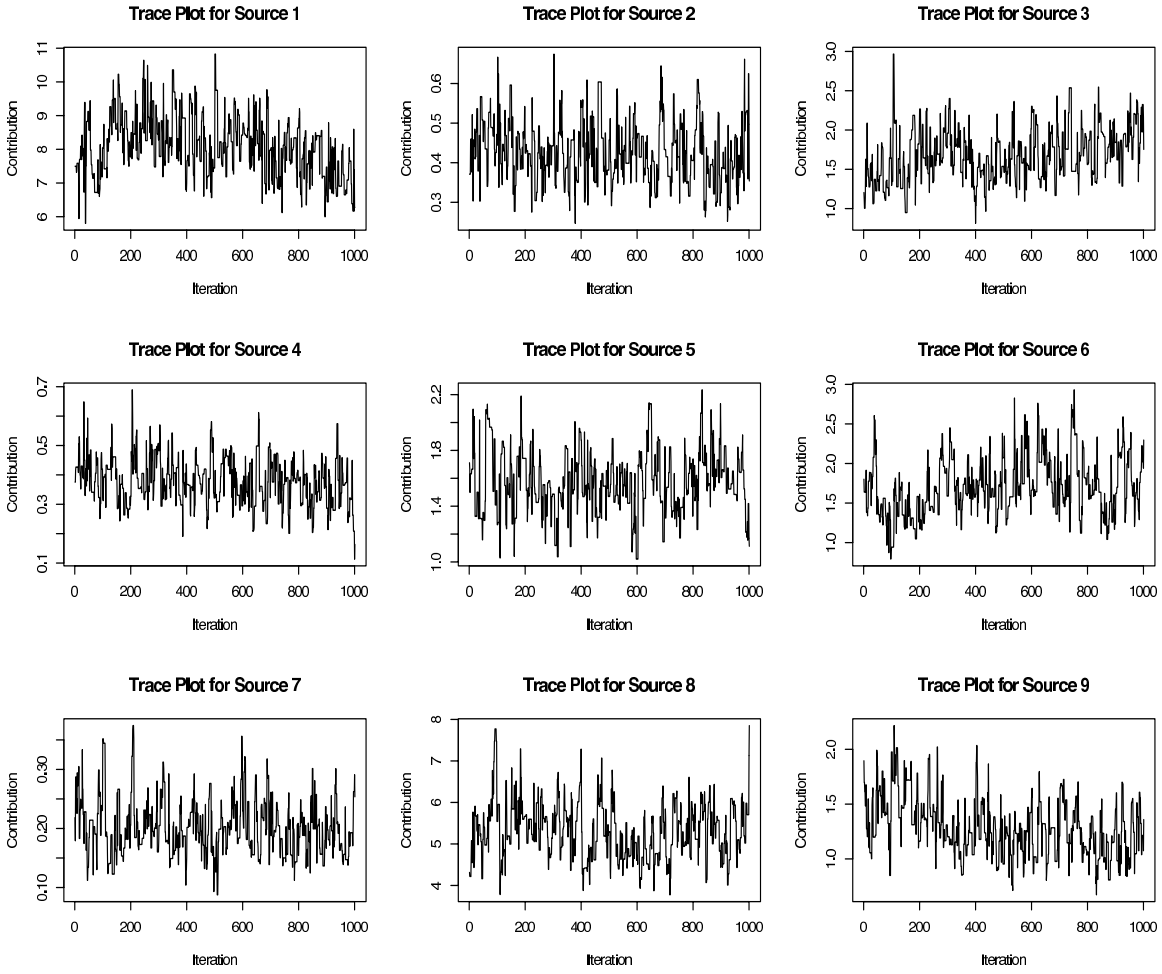


Figure 4.2: Successive draws of f_{kt} as obtained by MCMC sampling methods. The random scatter of successive draws supports the hypothesis that the MCMC algorithm achieved convergence.

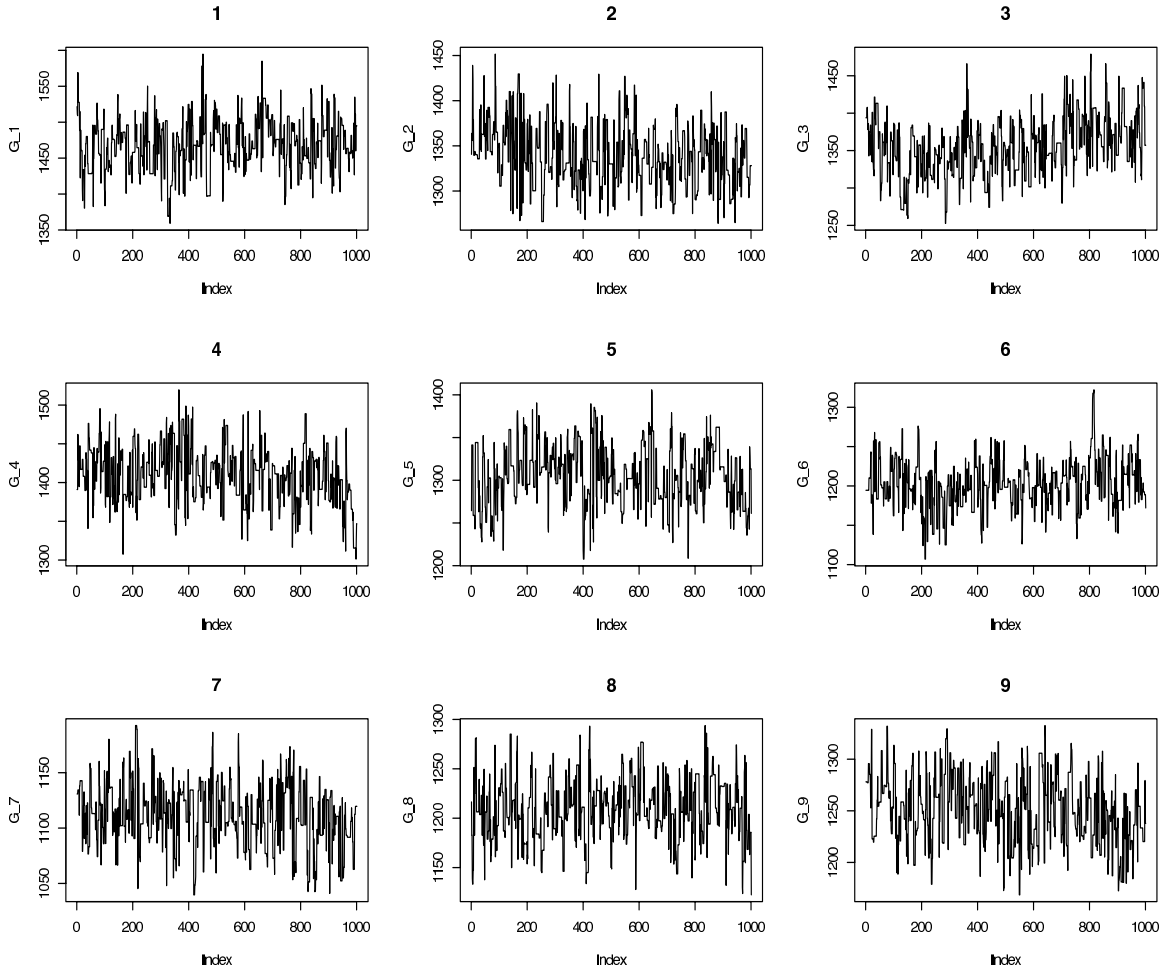


Figure 4.3: Successive draws of g_k as obtained by MCMC sampling methods. The random scatter of successive draws supports the hypothesis that the MCMC algorithm achieved convergence.

the elements of $\boldsymbol{\lambda}_{kt}$. Because each $\lambda_{pkt} \in \boldsymbol{\lambda}_{kt}$ is correlated, the mixing achieved in λ_{pkt} is less than that achieved in f_{kt}, where each $f_{kt} \in \mathbf{F}$ is independent.

Figure 4.5 displays successive draws of those λ_{pkt} that do not make up the majority of $\boldsymbol{\lambda}_{kt}$. The mixing achieved by the MCMC algorithm for those elements of $\boldsymbol{\lambda}_{kt}$ which are relatively small is even less than the mixing achieved for the primary elements of $\boldsymbol{\lambda}_{kt}$. This is due to a slight deficiency in the DP model. Recall that in the DP model, each λ_{pkt} is constrained to be greater than zero. If λ_{pkt} were allowed to be zero through a multivariate point mass mixture prior (to be discussed in Section 5.2), then the mixing would improve because the algorithm could let $\lambda_{pkt} = 0$ rather than settling for values of λ_{pkt} near zero. However, due to the small practical importance of the secondary elements of $\boldsymbol{\lambda}_{kt}$, the lack of mixing presents no practical problems.

4.4 Discussion

In this section, the performance of the DP model is compared to PMF based on the previously discussed simulated data sets.

4.4.1 Model Performance in the Presence of Time Varying Profiles

Figure 4.6 displays a plot of a single $\lambda_{pkt} \in \boldsymbol{\lambda}_{kt}$ over time for $g_k \in \{100, 250\}$ and $w_{pt} \in \{.2, .8\}$. The first column of plots in Figure 4.8 contrasts MAE_Λ using the DP model to using PMF. As shown in Figure 4.6, the Dirichlet process model consistently outperforms PMF in estimating $\boldsymbol{\lambda}_{kt}$ despite the value of g_k and w_{pt}. This performance is due to the fact that the Dirichlet process model provides a smoothed estimate of the $\boldsymbol{\lambda}_{kt}$ process while PMF provides merely a single point estimate across all time periods. Thus, the DP model is not only more flexible than PMF in estimating $\boldsymbol{\lambda}_{kt}$, but more accurate than PMF in estimating source profiles which vary over time.

Table 4.1 compares the five-number summaries of MAE_Λ for the DP model and for PMF. When $g_k = 250$ the DP model provides accurate estimates of the underlying

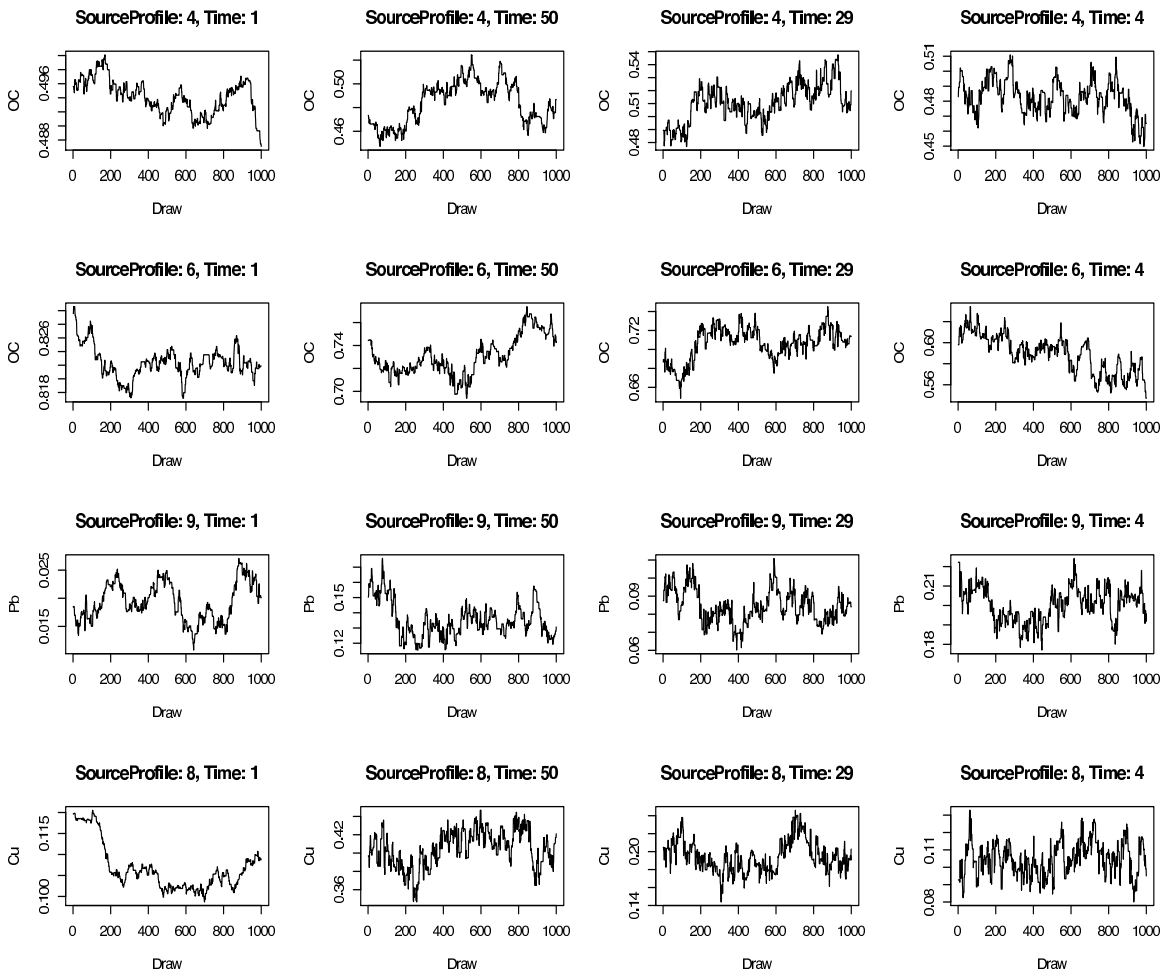


Figure 4.4: Successive draws of the primary elements of $\boldsymbol{\lambda}_{kt}$. The slight lack of mixing shown here could be due to the high correlation between each $\lambda_{pkt} \in \boldsymbol{\lambda}_{kt}$.

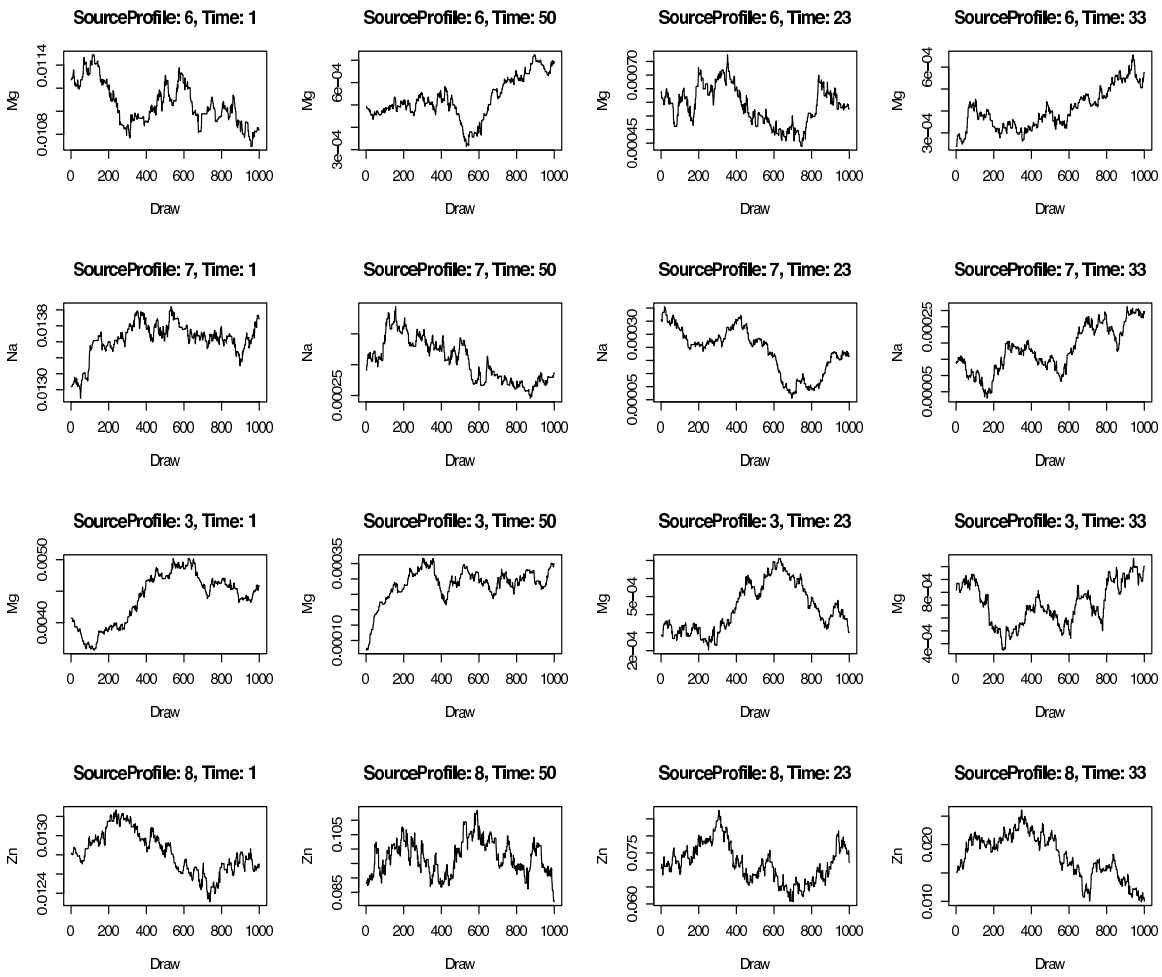


Figure 4.5: Successive draws of smaller elements of $\boldsymbol{\lambda}_{kt}$. The lack of mixing is due to the constraint that $\lambda_{pkt} > 0$.

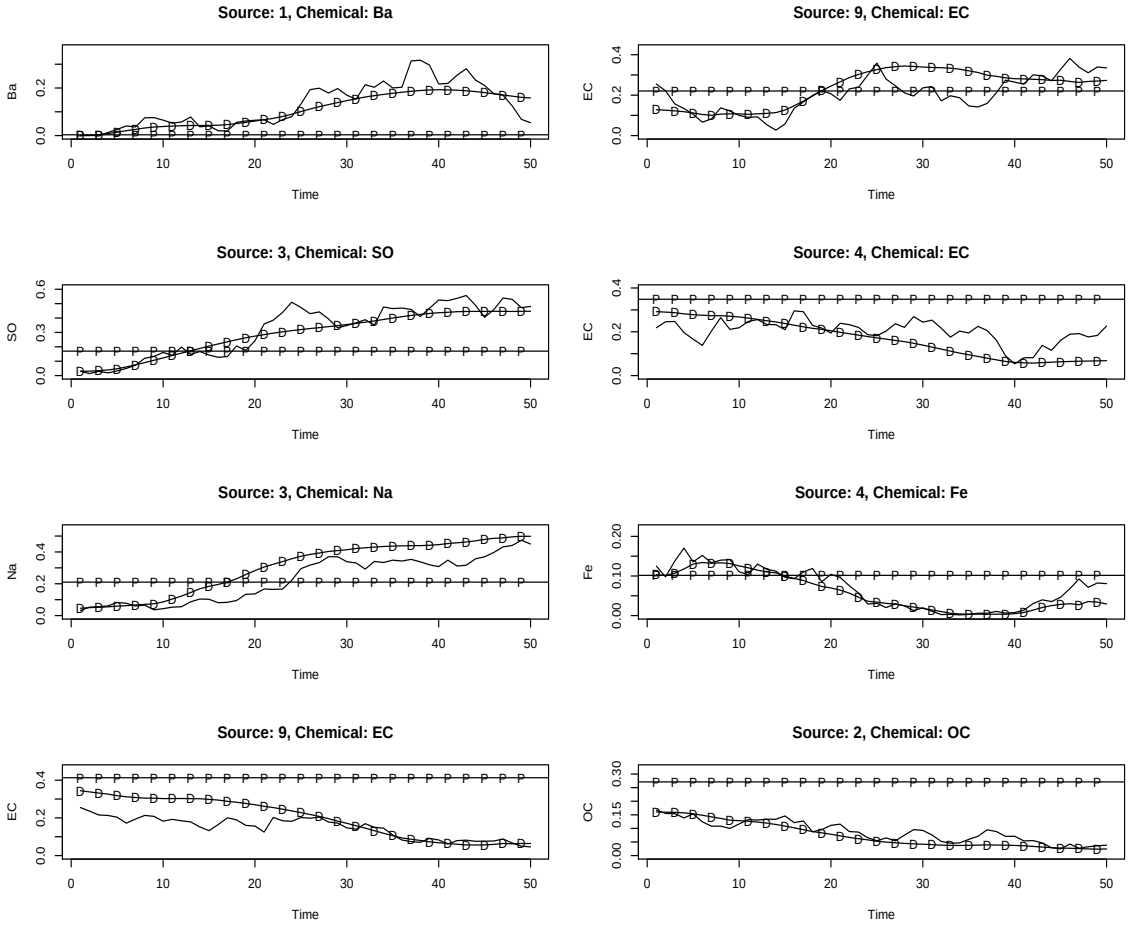


Figure 4.6: Time plot of one element of a source profile, $\boldsymbol{\lambda}_{kt}$, across values of g_k and w_{pt}. The rows correspond to (g_k, w_{pt}) = (100, .2), (100, .8), (250,. 2), (250, .8), respectively. The solid line is the true value of λ_{pkt}, the line marked by "D" is the DP estimate, and the line marked with "P" is the PMF estimate. In all cases, the DP model more accurately estimates the underlying Dirichlet process.

Table 4.1: Comparison of five-number summary of MAE_Λ when source profiles are time-variant.

g_k	CV	Model	Min	Q1	M	Q3	Max
100	0.2	DP Model	.01	.21	.41	.62	1.48
		PMF	.10	.57	1.08	1.66	1.93
100	0.8	DP Model	.03	.32	.66	.95	1.69
		PMF	.15	.77	.97	1.81	2.00
250	0.2	DP Model	.02	.23	.38	.61	1.30
		PMF	.11	.60	1.01	1.42	1.92
250	0.8	DP Model	.04	.26	.45	.77	1.34
		PMF	.22	.66	1.11	1.52	1.95

process, regardless of the level of w_{pt} with a median MAE_Λ of 0.39 when $w_{pt} = .2$ and 0.45 when $w_{pt} = .8$. However, when $g_k = 100$ and $w_{pt} = .8$, the DP model, while providing better estimates of $\boldsymbol{\lambda}_{kt}$ than PMF, does not approximate the underlying process to the same accuracy as in the other values of g_k and w_{pt}. The median MAE_Λ for $g_k = 100$ and $w_{pt} = .8$ is 0.66, compared to 0.41 when w_{pt} is reduced to 0.2.

MAE_Λ increases as the amount of variability in $\boldsymbol{\lambda}_{kt}$ and y_{pt} increases. Additionally, MAE_Λ is consistently higher for PMF than the DP model as shown in Figure 4.8. This is most likely due to the flexibility of the DP model in estimating time-varying profiles. PMF settles with a time-constant estimate of $\boldsymbol{\lambda}_{kt}$, which results in a higher MAE_Λ.

The performance of the DP model as compared to PMF in estimating source contributions is dependent on the variation present in the PM measurements. Figure 4.7 displays a time plot of source contributions with the median of the posterior draws of the DP model and the PMF contribution estimates. The second column of Figure 4.8 compares MAE_F from the DP model to MAE_F from PMF. Five-number summaries of MAE_F for both models is shown in Table 4.2.

When the variation in the PM measurements is small ($w_{pt} = .2$), the median MAE_F under the Dirichlet process model is 3.66 with an inter-quartile range of (2.65,

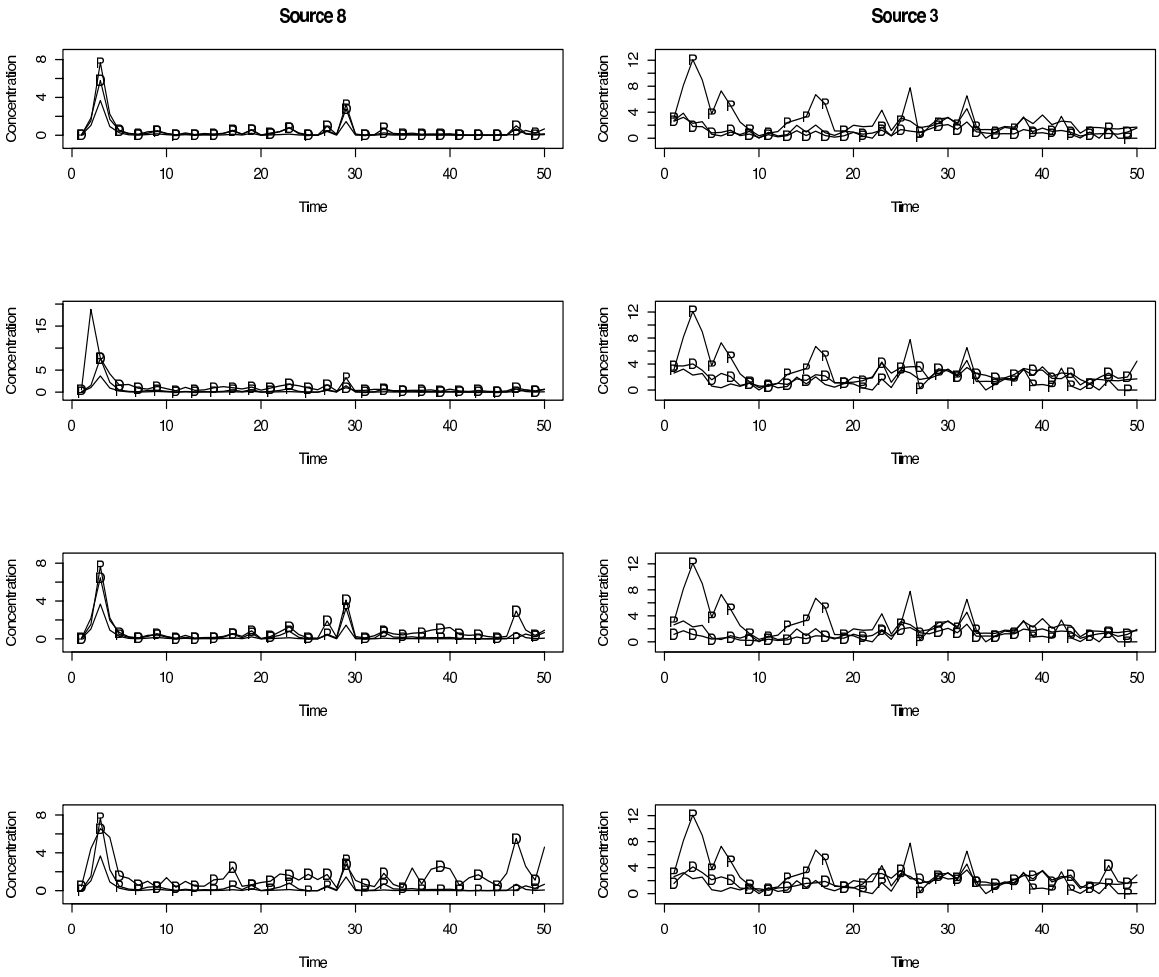


Figure 4.7: Time plot of source contributions f_{kt}. The rows correspond to $(g_k, w_{pt}) = (100, .2),\ (100, .8),\ (250, .2),\ (250, .8)$, respectively. The solid line is the true value of f_{kt}, the line marked by "D" is the DP model estimate, and the line marked by "P" is the PMF estimate. When $w_{pt} = 0.2$ (first and third rows), the DP model outperforms PMF. The DP model and PMF perform similarly when $w_{pt} = 0.8$ (second and fourth rows).

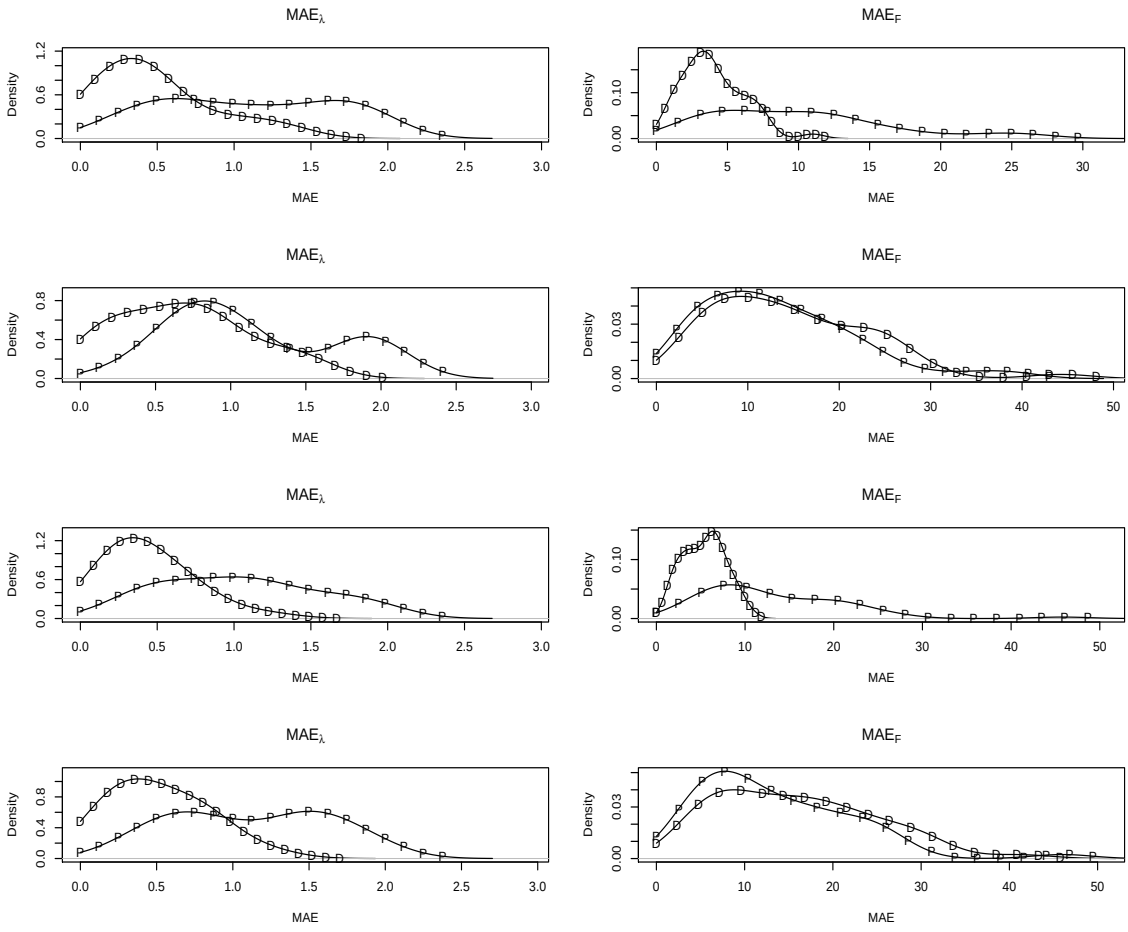


Figure 4.8: MAE density plots for the DP model and PMF for various levels of g_k and w_{pt}. The rows correspond to $(g_k, w_{pt}) = (100, .2)$, $(100, .8)$, $(250, .2)$, and $(250, .8)$, respectively. The DP model has lower MAE_F when $w_{pt} = .2$; however, when $w_{pt} = 0.8$, MAE_F under the DP model and PMF is comparable.

Table 4.2: Comparison of five-number summary of MAE_F when source profiles are time-variant.

g_k	CV	Model	Min	Q1	M	Q3	Max
100	.2	DP Model	.69	2.65	3.66	5.34	11.00
		PMF	1.87	5.07	9.32	12.75	26.67
100	.8	DP Model	3.74	7.42	13.49	20.90	44.24
		PMF	2.27	7.51	11.68	18.00	39.24
250	.2	DP Model	1.51	3.52	5.71	6.87	10.52
		PMF	2.60	7.65	10.06	10.58	45.76
250	.8	DP Model	3.44	8.16	15.28	21.88	41.14
		PMF	3.04	7.25	11.31	18.39	46.01

5.34), compared to the larger median MAE_F of 9.32 with an inter-quartile range of (5.07, 12.75) under PMF. Thus, the Dirichlet process model outperforms the PMF estimate in the presence of low variability among PM measurements. However, when the variance of PM measurements is large ($w_{pt} = .8$), the median MAE_F is 13.49 with an interquartile range of (7.42, 20.90) under the DP model, compared to a median MAE_F of 11.68 and interquartile range (7.51, 18.00) under PMF. Thus, in the presence of larger variation among PM measurements the Dirichlet process model performs comparably to PMF.

4.4.2 Model Performance in the Presence of Time-Constant Profiles

As shown in the previous section, the DP model provided better estimates of $\boldsymbol{\lambda}_{kt}$ when $\boldsymbol{\lambda}_{kt}$ was allowed to vary through time. Additionally, the DP model had lower MAE_F than PMF in the presence of small variation among PM measurements. In this section, the performance of the DP model is compared to PMF in the presence of constant source profiles.

Figure 4.9 displays a time plot of two different constant source profile elements when $w_{pt} = .2$ (first row) and $w_{pt} = .8$ (second row). Figure 4.9 also displays the DP model and PMF estimates of the constant source profile element. Five-number summary comparisons of MAE_Λ in the presence of constant source profiles are shown in Table 4.3. In the case of constant source profiles, the DP model has lower MAE_Λ than PMF with a median MAE_Λ of 0.10, compared to the median MAE_Λ under PMF of 0.28 when $w_{pt} = .2$. When $w_{pt} = .8$, the DP model had a median MAE_Λ of 0.11 compared to a median MAE_Λ of 1.41 under PMF. Similar to the case where source profiles vary through time, the DP model outperforms PMF in estimating source profiles. Given that a key assumption of PMF is that source profiles are constant through time, one would assume *a priori* that PMF would perform at least as well as the DP model; this is, however, not the case. The DP model is flexible enough to

Table 4.3: Comparison of five-number summary of MAE_Λ when source profiles are time-invariant.

g_k	CV	Model	Min	Q1	M	Q3	Max
Constant	.2	DP Model	.04	.06	.10	.15	.21
		PMF	.08	.14	.28	.37	1.22
Constant	.8	DP Model	.05	.09	.11	.15	.16
		PMF	.35	.57	.83	.97	1.14

incorporate the case where source profiles are constant through time.

A comparison of the DP model and PMF estimates of f_{kt} is displayed in Figure 4.10 and Table 4.4. Figure 4.10 displays a time plot of two source contributions with the DP model and PMF estimates overlaid. Table 4.4 compares the five-number summaries of MAE_F in the presence of constant source profiles. Despite the value of w_{pt}, the DP model estimates source contributions better than PMF when source profiles are constant through time. The median MAE_F for the DP model is 2.15 when $w_{pt} = .2$, compared to a median MAE_F of 3.00 when using PMF. Additionally, when $w_{pt} = .8$ the median MAE_F for the DP model is 9.53, compared to 16.83 for PMF. Thus, in both cases, the DP model more accurately estimates f_{kt}.

The density plots of MAE_Λ and MAE_F when $\boldsymbol{\lambda}_{kt}$ is held constant are shown in Figure 4.11. As mentioned above, the DP model surpasses PMF in accuracy of parameter estimates when the assumption of constant source profiles holds.

4.4.3 Estimation of the Precision Parameter g_k

Now consider examining how the DP model estimates the precision parameter g_k when $\boldsymbol{\lambda}_{kt} \neq \boldsymbol{\lambda}_k$ for all t and when $\boldsymbol{\lambda}_{kt} = \boldsymbol{\lambda}_k$ for all t. The estimate of the parameter g_k represents the model's understanding of the smoothness of the underlying Dirichlet process for $\boldsymbol{\lambda}_{kt}$. The higher the estimate, the smoother the process.

Figure 4.12 displays marginal density estimates of g_k for two different sources

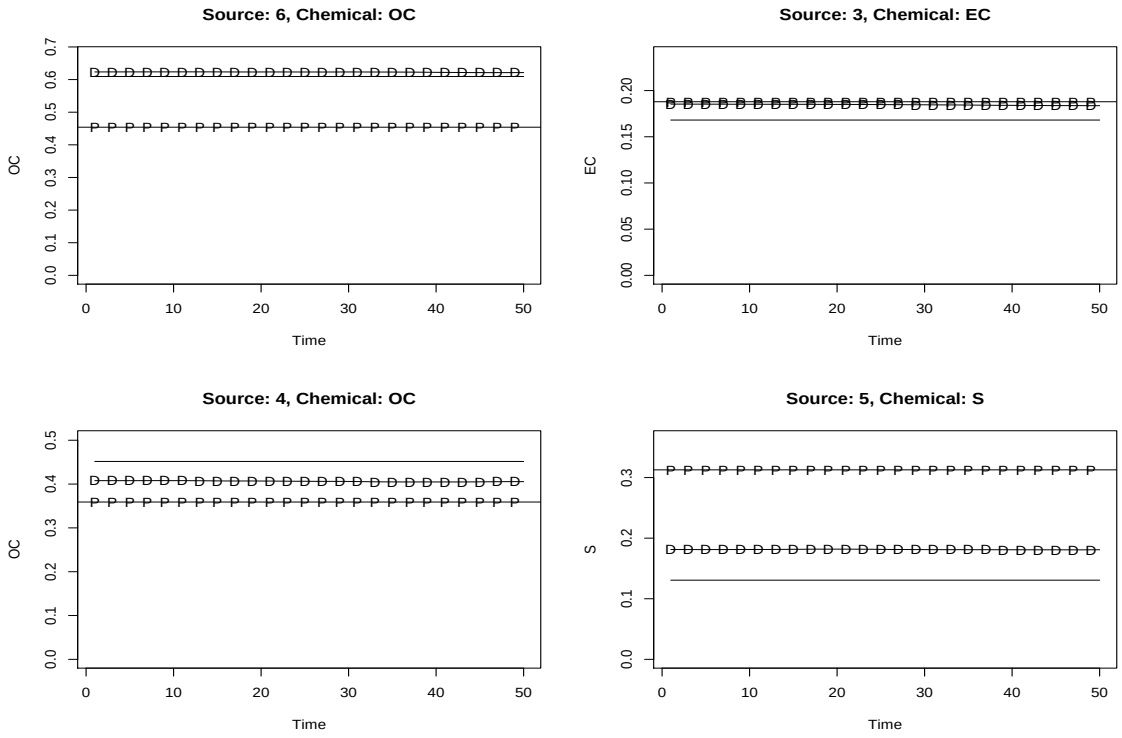


Figure 4.9: Time plot of two source profile elements. The solid line is the true value of the source profile, the "D" line is the DP model estimate and the "P" line is the PMF estimate. The DP model correctly, and more accurately than PMF, estimates $\boldsymbol{\lambda}_{p,t}$ when $w_{pt} = .2$ (first row) and $w_{pt} = .8$.

Table 4.4: Comparison of five-number summary of MAE_F when source profiles are time-invariant.

g_k	CV	Model	Min	Q1	M	Q3	Max
Constant	.2	DP Model	.41	1.14	2.15	3.29	7.55
		PMF	.80	1.70	2.44	3.41	7.01
Constant	.8	DP Model	2.91	6.81	9.53	14.18	46.36
		PMF	1.97	5.79	10.83	16.09	44.21

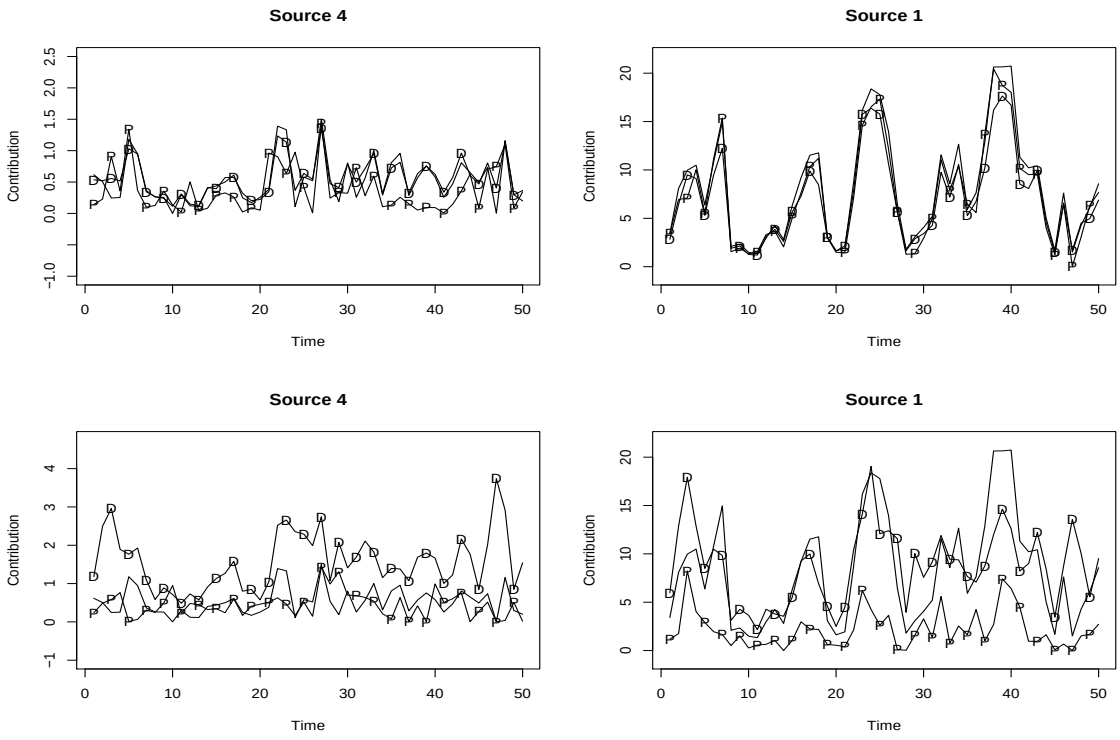


Figure 4.10: Time plot of f_{kt} under constant source profiles. The unmarked solid line is the true value of f_{kt}, the solid line marked "D" is the DP model estimate and the solid line marked "P" is the PMF estimate. For $w_{pt} = .2$ (first row) and $w_{pt} = .8$ (second row), the DP model outperforms PMF.

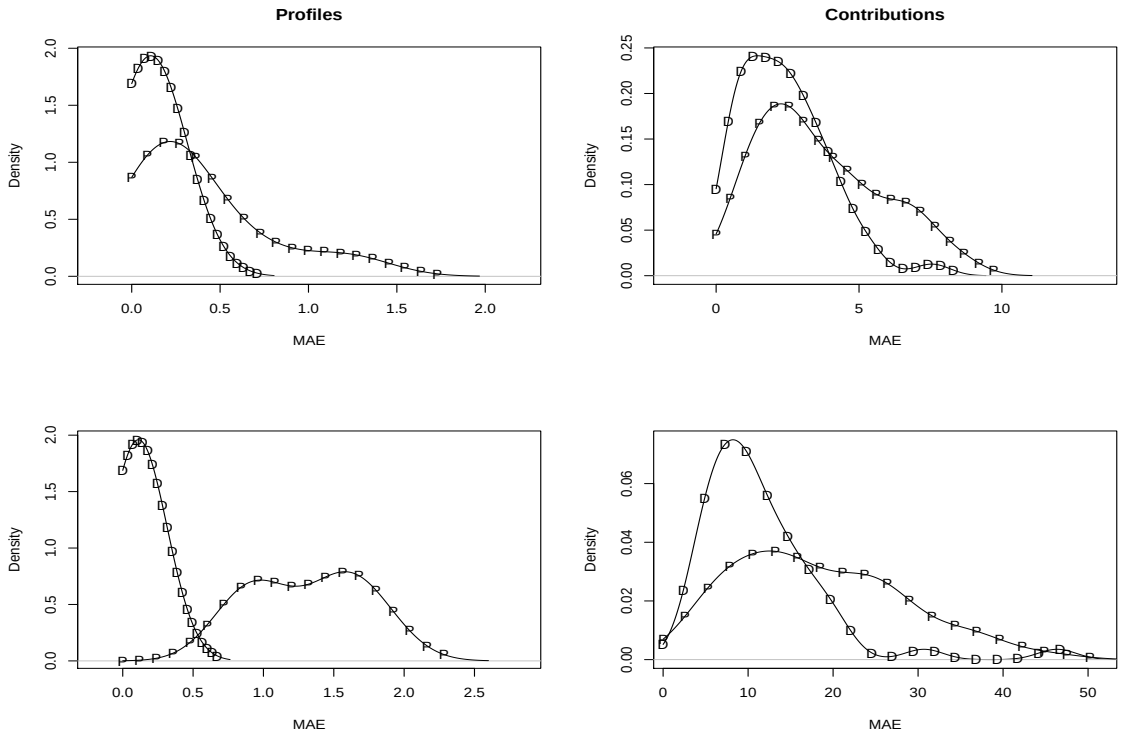


Figure 4.11: Comparison of MAE under constant source profiles when $w_{pt} = .2$ (first row) and $w_{pt} = .8$ (second row). Under both levels of w_{pt}, the DP model has lower MAE than PMF.

when the true value of $g_k = 100$ (first row), when the true value of $g_k = 250$ (second row), and when $\boldsymbol{\lambda}_{kt}$ is constant for all t (third row). When $g_k \in \{100, 250\}$, the DP model largely overestimates the value of g_k. Averaged across all sources, the posterior mean of g_k is approximately 2000, regardless of the value of g_k, with an average MAE_g of 1890. These large estimates of g_k suggest that MCMC methods estimate the underlying process to be much smoother than the true underlying process. This large overestimation of g_k is of little concern, however, due to the fact that the DP model still estimates $\boldsymbol{\lambda}_{kt}$ and f_{kt} well.

When the source profiles are constant over time, the DP model estimates g_k to be approximately 1,000,000 (see the third row of Figure 4.12). This large estimate of g_k arises from the fact that the underlying process is flat and hence very smooth. Thus, when $\boldsymbol{\lambda}_{kt}$ is constant for all t, the DP model correctly estimates a large value of g_k.

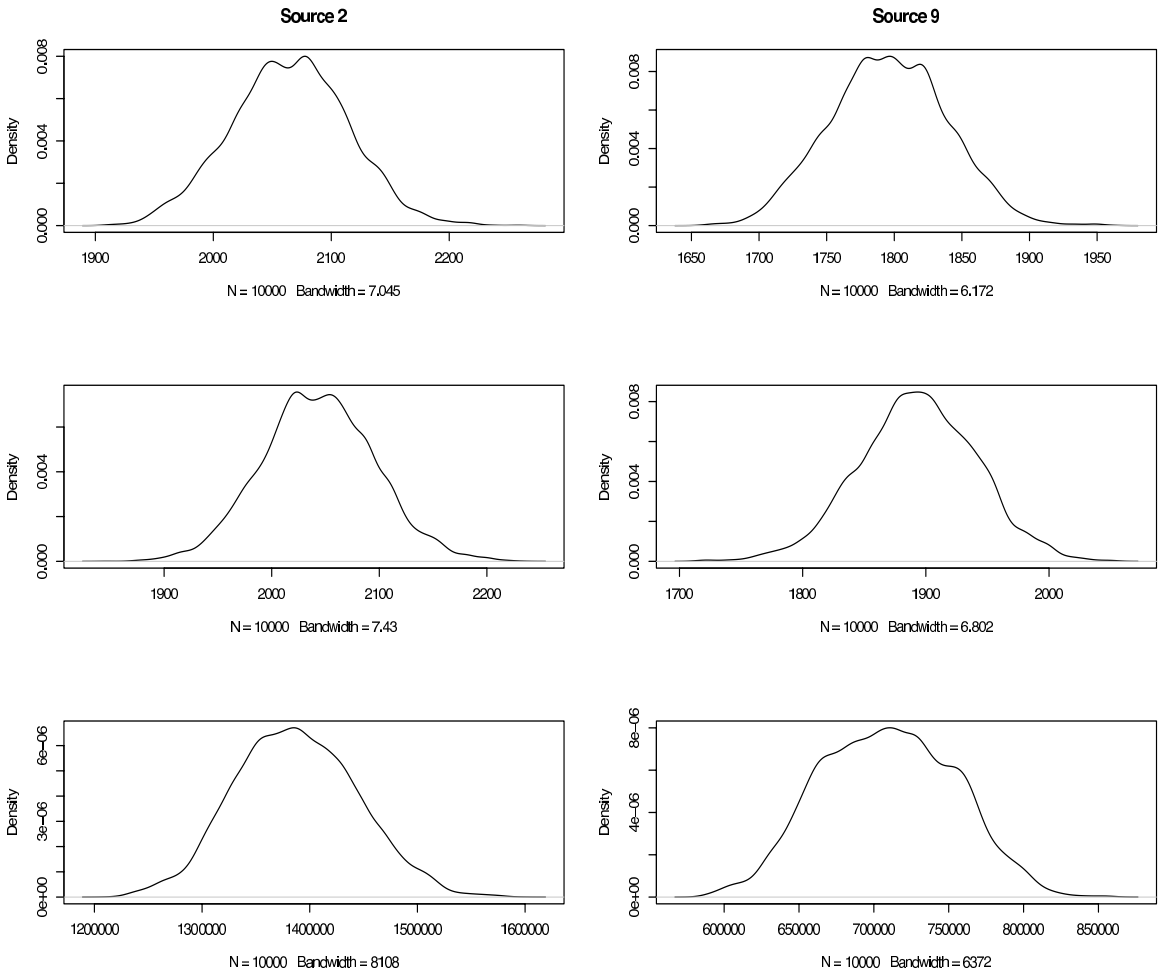


Figure 4.12: Density estimates of g_k across two sources. The rows of plots correspond to $g_k = 100$, $g_k = 250$, and time constant profiles respectively. When $g_k \in \{100, 250\}$ the DP model largely overestimates the smoothness of the underlying process. However, when profiles are held constant, the DP model correctly estimates a large value of g_k.

5. FUTURE RESEARCH

In this chapter, a few possible areas for future work regarding PSA and the DP model are presented here for the reader's information.

5.1 Incorporating the temporal structure in $\mathbf{F}$

As discussed in Chapter 3, not only does $\mathbf{\Lambda}_t$ exhibit temporal correlation, but elements of $\mathbf{F}$ also exhibit temporal correlation. For the purposes of this thesis, the temporal correlation in $\mathbf{F}$ was ignored and emphasis was placed on estimating the Dirichlet process involving $\boldsymbol{\lambda}_{kt}$. However, a complete air pollution model would also incorporate the temporal correlation in $\mathbf{F}$. Park et al. (2001) proposed using a multivariate auto-regressive (AR) model to account for the correlation among f_{kt}. A possible extension of the DP model would include f_{kt} in the system equation of the DLM and unify the methods outlined in this thesis as well as those methods developed by Park et al. (2001) in proposing a model that accounts for both sources of temporal correlation.

5.2 Point mass mixture priors for $\boldsymbol{\lambda}_{kt}$

Not all pollution sources emit all P chemical species in the pollution model. Identifying which elements of the source profile are 0 is desirable in that the source profile would then better represent the pollution being emitted from the given pollution source. From a modeling perspective, this can be represented by allowing some $\lambda_{pkt} = 0$. While the DP model does not directly allow for $\lambda_{pkt} = 0$, the DP model does allow elements of $\boldsymbol{\lambda}_{kt}$ to be very small. Extending the DP model to allow $\lambda_{pkt} = 0$ can be done by using a point mass mixture prior on each element $\boldsymbol{\lambda}_{kt}$. If each element of $\boldsymbol{\lambda}_{kt}$ were allowed to be independent, then a point mass mixture prior on λ_{pkt} could

be of the form

$$\pi(\lambda_{pkt}) = \begin{cases} c \text{ if } \lambda_{pkt} = 0 \\ (1-c) \times \text{LN}[a,b] \text{ if } \lambda_{pkt} > 0, \end{cases} \tag{5.1}$$

where c is a known and specified constant. The use of the lognormal distribution in this example is somewhat mathematically satisfying in that each $\lambda_{pkt} > 0$ but does not restrict $\lambda_{pkt} < 1$. Additionally, Model 5.1 does not accurately apply the constraint $\sum_{p=1}^{P} \lambda_{pkt} = 1$.

Allowing $\lambda_{pkt} = 0$ becomes much more complex when independence of λ_{pkt} is not assumed. If λ_{pkt} is not independent, a multivariate point mass mixture prior for $\boldsymbol{\lambda}_{kt}$ would have to be specified to allow some elements of $\boldsymbol{\lambda}_{kt} = 0$ and the remaining elements of $\boldsymbol{\lambda}_{kt}$ to follow the Dirichlet distribution. Specifying a point mass mixture prior distribution for the vector $\boldsymbol{\lambda}_{kt}$ is difficult because the density of the Dirichlet distribution depends on the number of non-zero elements of $\boldsymbol{\lambda}_{kt}$.

5.3 Controlling for different levels of correlation between elements of $\boldsymbol{\lambda}_{kt}$

For this thesis, each $\lambda_{pkt} \in \boldsymbol{\lambda}_{kt}$ was assumed to exhibit the same amount of temporal correlation. This temporal correlation was represented by the smoothness parameter g_k in Equation 6.3. However, as shown in Section 3.2, the amount of temporal correlation in λ_{pkt} related to how prevalent the chemical species was in the corresponding pollution source profile (see Figure 3.3). These varying degrees of temporal correlation can be accounted for in the DP model by altering the system equation to be

$$\boldsymbol{\lambda}_{kt} \sim \text{DIR}[\mathbf{G}_k \boldsymbol{\lambda}_{k(t-1)}] \quad k = 1, \ldots, K \quad t = 1, \ldots, N, \tag{5.2}$$

where $\mathbf{G}_k = \text{diag}(g_{1k}, g_{2k}, \ldots, g_{Pk})$, a diagonal matrix of smoothness parameters. Each $g_{pk} \in \mathbf{G}_k$ represents how closely tied λ_{pkt} is to $\lambda_{pk(t-1)}$.

5.4 Reverse Jump MCMC to estimate the number of sources

For this thesis, the number of pollution sources was assumed to be known. Most often, this is not a legitimate assumption. In fact, the number of pollution sources may not be constant for all t. Reverse jump MCMC (RJMCMC) is a technique that allows a Markov chain to jump between model spaces while ensuring that the reversibility condition is met. This reversibility is constructed by specifying map functions that map model parameters between the two model spaces. The use of RJMCMC could be applied to estimating the number of sources in the pollution model. Ideally, RJMCMC would be used to allow for different time periods to have a different number of pollution sources.

6. CONCLUSIONS

Pollution source apportionment (PSA) is the practice of deriving information about pollution sources from measurements of particulate matter (PM) to help regulate pollution emissions from these pollution sources. The basic PSA model is written as

$$\underset{P\times N}{\mathbf{Y}} = \underset{P\times K}{\mathbf{\Lambda}}\underset{K\times N}{\mathbf{F}} + \underset{P\times N}{\mathbf{E}}, \tag{6.1}$$

where $\mathbf{Y}$ is a matrix of measurements on P chemical species measured over N time periods, $\mathbf{\Lambda}$ is the matrix of pollution source profiles, and each $\lambda_{pk} \in \mathbf{\Lambda}$ represents the proportion of chemical p in the chemical makeup of pollution emitted from source k, $\mathbf{F}$ is the matrix of pollution source contributions, and $\mathbf{E}$ is the error term matrix. The k^{th} column of $\mathbf{\Lambda}$ is referred to as the k^{th} pollution source profile and is denoted as $\boldsymbol{\lambda}_k$.

Several of the assumptions made by Model 6.1 are nonrepresentative of air pollution data. Two such assumptions are (1) PM measurements are independent and (2) $\mathbf{\Lambda}$ is time-invariant. Therefore, this thesis proposes a dynamic linear model (DLM) with observation equation

$$\mathbf{y}_t \sim \text{LN}[\mathbf{\Lambda}_t\mathbf{f}_t, w_{pt}], \;\; p = 1,\ldots,P \;\; t = 1,\ldots,N, \tag{6.2}$$

where each $\boldsymbol{\lambda}_{kt}$ follows the system equation given by

$$\boldsymbol{\lambda}_{kt} \sim \text{DIR}[g_k\boldsymbol{\lambda}_{k(t-1)}] \;\; k = 1,\ldots,K \;\; t = 1,\ldots,N, \tag{6.3}$$

and the time 0 initial information is given by

$$\boldsymbol{\lambda}_{k,0} \sim \text{DIR}[\mathbf{m}_k] \;\; k = 1,\ldots,K, \text{ and} \tag{6.4}$$

$$f_{kt} \sim \text{LN}[a_{kt}, b_{kt}] \;\; k = 1,\ldots,K \;\; t = 1,\ldots N, \tag{6.5}$$

to account for these inadequacies. Models 6.2-6.5 relax the assumptions of independence and constant source profiles by allowing profile vectors ($\boldsymbol{\lambda}_{kt}$) to vary through time according to a Dirchlet process. Hence, Models 6.2-6.5 are collectively referred to as the Dirichlet process (DP) model. Allowing $\boldsymbol{\lambda}_{kt}$ to follow a Dirichlet process is satisfying in that the Dirichlet distribution accurately applies the constraint that $\sum_{p=1}^{P} \lambda_{pkt} = 1$. Additionally, specifying each $f_{kt} \in \mathbf{F}$ to have a lognormal distribution is also logical in that each f_{kt} is constrained to be greater than 0.

In Chapter 4, the effectiveness of the DP model was evaluated based on several simulated data sets where the data sets were simulated under different degrees of variability. In the presence of time-varying source profiles the DP model was superior to PMF in estimating $\boldsymbol{\lambda}_{kt}$. Additionally, when the variation among the PM measurements was small ($w_{pt} = .2$), the DP model more accurately estimated elements of $\mathbf{F}$ than PMF. PMF and the DP model performed similarly in estimating source contributions when the variation among PM measurements was large (see Figure 4.8).

The DP model was also compared to PMF when the assumption of constant source profiles was true. In this case, the DP model was found to be superior to PMF in estimating both $\boldsymbol{\Lambda}_t$ and $\mathbf{F}$, regardless of the amount of variability present among PM measurements. This discovery is somewhat alarming in that a key assumption of PMF is that source profiles are constant through time. The DP model makes no such assumption and yet outperforms PMF in estimating $\boldsymbol{\Lambda}$ and $\mathbf{F}$ when this assumption holds true.

For the reasons stated above, the DP model is found to be a more flexible model than the basic PSA model in that the DP model has less assumptions required for its use and yet still produces desirable results. The DP model also has the added benefit of making probability statements about model parameters because the DP model is estimated using Bayesian MCMC methods.

Bibliography

Calder, C. A. (2003), "Exploring Latent Structure in Spatial Temporal Processes Using Process Convolutions," Ph.D. thesis, Duke University.

Chib, S. and Greenberg, E. (1995), "Understanding the Metropolis-Hasting Algorithm," *The American Statistician*, 49, 327–335.

Christensen, W. F. and Gunst, R. F. (2004), "Measurement error models in chemical mass balance analysis of air quality data," *Atmospheric Environment*, 38, 733–744.

Christensen, W. F. and Sain, S. R. (2002), "Accounting for Dependence in a Flexible Multivariate Receptor Model," *Technometrics*, 44, 328–337.

Christensen, W. F., Schauer, J. J., and Lingwall, J. W. (2006), "Iterated Confirmatory Factor Analysis for Pollution Source Apportionment," *Environmetrics*.

Coulter, C. T. (2000), *EPA-CMB8.2 User's Manual*, Environmental Protection Agency.

Friedlander, S. (1973), "Chemical Element Balances and Identification of Air Pollution Sources." *Environmental Science and Technology*, 7, 235–240.

Gelman, A., Carlin, J. B., Stern, H. S., and Rubin, D. B. (2004), *Bayesian Data Analysis*, Chapman and Hall, 2nd ed.

Henry, R. (1997), "History and Fundamentals of Multivariate Air Quality Receptor Models," *Chemometrics and Intelligent Laboratory Systems*, 37, 525–530.

Henry, R. C. (1987), "Current Factor Analysis Models are Ill-Posed," *Atmospheric Environment*, 21, 1815–1820.

Lingwall, J. W. (2006), "Bayesian and Positive Matrix Factorization Approaches to Pollution Source Apportionment," Master's thesis, Brigham Young University.

Lopes, H. F. (2000), "Bayesian Analysis in Latent Factor and Longitudinal Models," Ph.D. thesis, Duke University.

Miller, M., Friedlander, S., and Hidy, G. (1972), "A Chemical Element Balance for the Pasadena Aerosol," *Journal of Colloid Interface Science*, 39, 65–176.

Paatero, P. and Tapper, U. (1994), "Positive Matrix Factorization: a nonnegative factor model with optimal utilization of error estimates of data values," *Environmetrics*, 5, 111–126.

Park, E. S., Guttorp, P., and Henry, R. C. (2001), "Multivariate Receptor Modeling for Temporally Correlated Data by Using MCMC," *Journal of the American Statistical Association*, 96, 1171–1183.

Park, E. S., Henry, R. C., and Spiegelman, C. H. (1999), "Determining the Number of Major Pollution Sources in Multivariate Air Quality Receptor Models," Tech. Rep. 34, National Research Center for Statistics and the Environment.

Phalen, R. F. (2002), *The Particulate Air Pollution Controversy*, Kluwer Adademic Publishers.

Watson, J. G., Cooper, J. A., and Huntzicker, J. J. (1984), "The Effective Variance Weighting for Least Squares Calculations Applied to the Mass Balance Receptor Model," *Atmospheric Environment*, 18, 1347–1355.

West, M. and Harrison, J. (1997), *Bayesian Forecasting and Dynamic Linear Models*, Springer.

Winchester, J. and Nifong, G. (1971), "Water Pollution in Lake Michigan by Trace Elements from Pollution Aerosol Fallout," *Water, Air, and Soil Pollution*, 1, 50–64.

A. DISTRIBUTION NOTATION

A.1 The Lognormal Distribution

A random variable X follows a lognormal distribution if the logarithm of the random variable is distributed normally with mean μ and variance σ^2. The probability density function of X is

$$p(X|\mu,\sigma) = \frac{1}{x\sigma\sqrt{2\pi}} e^{-\frac{(\ln(x)-\mu)^2}{2\sigma^2}}, \tag{A.1}$$

where μ and σ are the mean and standard deviation of the variable's logarithm, respectively.

The expected value and variance of a lognormally distributed random variable are

$$E[X] = e^{\mu+\frac{\sigma^2}{2}}, \tag{A.2}$$

and

$$VAR[X] = (e^{\sigma^2}-1)e^{2\mu+\sigma^2}. \tag{A.3}$$

Throughout this thesis, the notation $LN[\mu^*,\nu]$ is used to represent the lognormal distribution with mean μ^* and coefficient of variation (CV) ν. Setting $\mu^* = E[X]$ and $\nu = \frac{\sqrt{VAR[X]}}{E[X]}$, μ and σ^2 can be solved for in terms of μ^* and ν. Doing so yields

$$\mu = \ln(\mu^*) - \frac{1}{2}\ln(\nu^2+1), \tag{A.4}$$

and

$$\sigma^2 = \ln(\nu^2+1). \tag{A.5}$$

A.2 The Dirichlet Distribution

The Dirichlet distribution is the multivariate generalization of the univariate beta distribution. The Dirichlet distribution is parameterized by a K-dimensional

vector α such that $\alpha_i > 0$ for all $\alpha_i \in \alpha$. If a K-vector $\mathbf{X}$ is Dirichlet-distributed, then $\mathbf{X}$ has the probability density function

$$p(\mathbf{X}|\alpha) = \frac{\Gamma\left(\sum_{i=1}^{K} \alpha_i\right)}{\prod_{i=1}^{K} \Gamma(\alpha_i)} \prod_{i=1}^{K} X_i^{\alpha_i - 1}, \tag{A.6}$$

where $\sum_{i=1}^{K} X_i = 1$ and $\Gamma(\cdot)$ is the gamma function. Because the Dirichlet distribution is the multivariate generalization of the beta distribution, the marginal distribution of each $X_i \in \mathbf{X}$ is distributed as a beta distribution with parameters α_i and $\beta = (\alpha_0 - \alpha_i)$, where $\alpha_0 = \sum_{i=1}^{K} \alpha_i$.

The univariate expected value and variance of a Dirichlet-distributed variable is,

$$E[X_i] = \frac{\alpha_i}{\alpha_0} \tag{A.7}$$

and

$$VAR[X_i] = \frac{\alpha_i(\alpha_0 - \alpha_i)}{\alpha_0^2(\alpha_0 + 1)}. \tag{A.8}$$

The notation DIR$[\alpha]$ is used throughout this thesis to denote a Dirichlet distributed random variable with parameter vector α.

A K-dimensional Dirichlet-distributed random variable can be generated with the following steps:

(1) For $i \in \{1, \ldots, K\}$, generate $Y_i \sim \text{Gamma}[\text{shape} = \alpha_i, \text{scale} = 1]$.

(2) Calculate $V = \sum_{i=1}^{K} Y_i$.

(3) Set $x_i = \frac{Y_i}{V}$.

The random variable, $\mathbf{x} = \{x_1, \ldots, x_K\}$ is then Dirichlet-distributed with parameter vector $\alpha = \{\alpha_1, \ldots, \alpha_K\}$.

```
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%

%%%Main MATLAB Code for Running MCMC to Estimate Time Varying Profiles%%%

%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%

%%Specify How Many Species, Days, Sources, Length, Burn are desired
NumSpecies = 44;
NumSources = 9;
NumDays = 50;
NumIter = 50000;
Burn = 40000;
allgg = [100];
allcv = [.2];
begsets = [16];
endsets = [16];
totloops = size(begsets);
totloops = totloops(2);

for hhh = 1:totloops

startset = begsets(hhh);
endset = endsets(hhh);
gg = allgg(hhh);
CVY = allcv(hhh);
cvout = 100*CVY;

for dd = startset:endset
    tic
Current_Project = ['Currently Running MCMC for G=' int2str(gg) '
    CV=' int2str(cvout) ' Set=' int2str(dd)]

%%Specify The File Names that contain the data, starting values,
%and Lambda_0
data = ['../SimulatedDataSets/SimData' int2str(gg) int2str(cvout)
    int2str(dd) '.txt'];
SV_Lambda = ['../SimulationParameters/LambdaT' int2str(gg) '.txt'];
SV_F = '../SimulationParameters/Contributions.txt';
```

```
SV_G = ['../SimulationParameters/g' int2str(gg) '.txt'];
Time0 = '../SimulationParameters/Profiles.txt';
File_Cand_Sig_G = '../SimulationParameters/CandSigG.txt';
File_Cand_Sig_F = '../SimulationParameters/CandSigF.txt';
File_Cand_Sig_L = ['../SimulationParameters/Cand_Sig_L_One_' int2str(gg) '.txt'];

%%Set Prior Values for F and G
F_mean = 20;
F_cv = 1;
G_mean = 175;
G_var = 75^2;

%%Get Original Candidate Sigmas for G, F, and Lambda
Cand_Sig_G = dlmread(File_Cand_Sig_G);
Cand_Sig_G = Cand_Sig_G(1:NumSources,:);
Cand_Sig_F = dlmread(File_Cand_Sig_F);
Cand_Sig_F = Cand_Sig_F(1:NumDays,1:NumSources)';
tempL = dlmread(File_Cand_Sig_L);
Cand_Sig_L = zeros(NumSpecies,NumSources,NumDays);
splits = 1:NumSpecies:((NumSpecies*NumSources)+1);
for t = 1:NumDays
    for k = 1:(size(splits,2)-1)
        Cand_Sig_L(:,k,t) = tempL(splits(:,k):(splits(:,k+1)-1),t);
    end
end clear File_Cand_Sig_G File_Cand_Sig_F File_Cand_Sig_L;

%%Create Matrices to keep track of Acceptance Rates
Accept_Rate_G = zeros(NumSources,1);
Accept_Rate_F = zeros(NumSources,NumDays);
Accept_Rate_L = zeros(NumSources,NumDays);
Rate_Counter = 0;
Total_L = Accept_Rate_L;
Total_G = Accept_Rate_G;
Total_F = Accept_Rate_F;

%%Read in the NumSpecies by NumDays Matrix of Chemical Concentrations
Y = dlmread(data);
Y = Y(1:NumSpecies,1:NumDays);
clear data;

%%Read in the NumSpecies by NumSources Time 0 Profile Matrix
```

```
Lambda_0 = dlmread(Time0);
clear Time0;

%%Initialize Parameter Matrices
Lambda_1 = zeros(NumSpecies,NumSources,NumIter-Burn);

%%Repeat for Lambda_2 through Lambda_50%%

Ghold = zeros(NumSources,NumIter-Burn);
Fhold = zeros(NumSources,NumDays,NumIter-Burn);
wpthold = zeros(NumIter-Burn,1);

%%Read in the Starting Values for F and Lambda_t
temp = dlmread(SV_Lambda);
splits =
1:NumSpecies:((NumSpecies*NumSources)+1);
Lambda_t =
zeros(NumSpecies,NumSources,NumDays);
for t = 1:NumDays
    for k = 1:(size(splits,2)-1)
        Lambda_t(:,k,t) = temp(splits(:,k):(splits(:,k+1)-1),t);
    end
end
trueL = Lambda_t;
tempF = dlmread(SV_F);
F = tempF(1:NumDays,1:NumSources)';
trueF = F;
tempG = dlmread(SV_G);
G = tempG(:,1);
wpt = CVY;
clear tempF tempG SV_F SV_G SV_Lambda temp t k;

for ii = 2:NumIter

  %%Increment Rate_Counter to adjust Candidate Sigmas
  Rate_Counter = Rate_Counter+1;

  for kk=1:NumSources
    new = normrnd(G(kk,:),Cand_Sig_G(kk,:));
    if new > 0
        u = unifrnd(0,1,1,1);
```

```
          alpha = likelambda(Lambda_t(:,kk,:),new,Lambda_0(:,kk)) + prior_g(new,G_mean,G_var) ...
              - likelambda(Lambda_t(:,kk,:),G(kk,:),Lambda_0(:,kk)) - prior_g(G(kk,:),G_mean,G_var);
          if log(u) < alpha
          G(kk,:) = new;
          Accept_Rate_G(kk,1) = Accept_Rate_G(kk,1)+1;
          end
  end
  clear new u alpha accept_g;
end

update_Lambda_t = Lambda_t;
%Update Lambda_2 through Lambda_N-1
for jj = 2:(NumDays-1)
  for kk=1:NumSources
  old = update_Lambda_t(:,kk,jj);
  new = normrnd(old,Cand_Sig_L(:,kk,jj));
  new = new/sum(new);
  if min(new>0) == 1
      newmat = update_Lambda_t(:,:,jj);
      newmat(:,kk) = new;
      u = unifrnd(0,1,1,1);
          alpha = likey(Y(:,jj),newmat,F(:,jj),wpt) ...
              + prior_lambda(update_Lambda_t(:,kk,jj-1),new,update_Lambda_t(:,kk,jj+1),G(kk,:))...
              - likey(Y(:,jj),update_Lambda_t(:,:,jj),F(:,jj),wpt) ...
              - prior_lambda(update_Lambda_t(:,kk,jj-1),update_Lambda_t(:,kk,jj),...
                  update_Lambda_t(:,kk,jj+1),G(kk,:));
          if log(u) < alpha
              update_Lambda_t(:,kk,jj) = new;
              Accept_Rate_L(kk,jj) = Accept_Rate_L(kk,jj) + 1;
          end
  end
  end
end

%Update Lambda_1
for kk=1:NumSources
  old = update_Lambda_t(:,kk,1);
  new = normrnd(old,Cand_Sig_L(:,kk,1));
  new = new/sum(new);
  if min(new>0) == 1
      newmat = update_Lambda_t(:,:,1);
```

```
      newmat(:,kk) = new;
      u = unifrnd(0,1,1,1);
      t0 = rdirich(G(kk,:)*update_Lambda_t(:,kk,1));
          alpha = likey(Y(:,1),newmat,F(:,1),wpt) ...
              + prior_lambda(Lambda_0(:,kk),new,update_Lambda_t(:,kk,2),G(kk,:))...
              - likey(Y(:,1),update_Lambda_t(:,:,1),F(:,1),wpt) ...
              - prior_lambda(Lambda_0(:,kk),update_Lambda_t(:,kk,1),...
                  update_Lambda_t(:,kk,2),G(kk,:));
          if log(u) < alpha
              update_Lambda_t(:,kk,1) = new;
              Accept_Rate_L(kk,1) = Accept_Rate_L(kk,1) + 1;
          end
  end
end

%Update Lambda_NumDays
for kk=1:NumSources
      old = update_Lambda_t(:,kk,NumDays);
  new = normrnd(old,Cand_Sig_L(:,kk,NumDays));
  new = new/sum(new);
  if min(new>0) == 1
      newmat = update_Lambda_t(:,:,NumDays);
      newmat(:,kk) = new;
      u = unifrnd(0,1,1,1);
      t51 = rdirich(G(kk,:)*update_Lambda_t(:,kk,NumDays));
          alpha = likey(Y(:,NumDays),newmat,F(:,NumDays),wpt) ...
              + prior_lambda(update_Lambda_t(:,kk,NumDays-1),new,t51,G(kk,:))...
              - likey(Y(:,NumDays),update_Lambda_t(:,:,NumDays),F(:,NumDays),wpt) ...
              - prior_lambda(update_Lambda_t(:,kk,NumDays-1),...
                  update_Lambda_t(:,kk,NumDays),t51,G(kk,:));
          if log(u) < alpha
              update_Lambda_t(:,kk,NumDays) = new;
              Accept_Rate_L(kk,NumDays) = Accept_Rate_L(kk,NumDays) + 1;
          end
  end
end
Lambda_t = update_Lambda_t;

update = F;
for jj=1:NumDays
    for kk=1:NumSources
```

```
        old = update(kk,jj);
        new = normrnd(old,Cand_Sig_F(kk,jj));
        if new>0
            update(kk,jj) = new;
            u = unifrnd(0,1,1,1);
            alpha = like_f(Y(:,jj),Lambda_t(:,:,jj),update(:,jj),wpt) + prior_f(new,F_mean,F_cv) ...
                - like_f(Y(:,jj),Lambda_t(:,:,jj),F(:,jj),wpt) - prior_f(old,F_mean,F_cv);
            update(kk,jj) = old;
            if log(u) < alpha
                update(kk,jj) = new;
                Accept_Rate_F(kk,jj) = Accept_Rate_F(kk,jj) + 1;
            end
        end
    end
end
F(:,:) = update;

%%Update wpt
wptnew = normrnd(wpt,.05);
if wptnew > 0
    likewptold = 0;
    likewptnew = 0;
    for tt=1:NumDays
        likewptold = like_f(Y(:,tt),Lambda_t(:,:,tt),F(:,tt),wpt)+likewptold;
        likewptnew = like_f(Y(:,tt),Lambda_t(:,:,tt),F(:,tt),wptnew)+likewptnew;
    end
    alpha = likewptnew + log(wptnew) - wptnew/CVY - likewptold -...
    log(wpt) + wpt/CVY;
    u = unifrnd(0,1,1,1);
    if log(u) < alpha
        wpt = wptnew;
    end
end

%%Tweak the Candidate Sigmas if Acceptance Ratios are Bad
if Rate_Counter == 10
    Total_L = Total_L + Accept_Rate_L;
    Total_F = Total_F + Accept_Rate_F;
    Total_G = Total_G + Accept_Rate_G;
    Cand_Sig_G(Accept_Rate_G >= 6) = Cand_Sig_G(Accept_Rate_G >= 6)*1.1;
    Cand_Sig_G(Accept_Rate_G <= 2) = Cand_Sig_G(Accept_Rate_G <= 2)*.9;
```

```
        Cand_Sig_F(Accept_Rate_F >= 6) = Cand_Sig_F(Accept_Rate_F >= 6)*1.1;
        Cand_Sig_F(Accept_Rate_F <= 2) = Cand_Sig_F(Accept_Rate_F <= 2)*.9;
        for i=1:NumDays
        Cand_Sig_L(:,Accept_Rate_L(:,i) >= 6,i) = Cand_Sig_L(:,Accept_Rate_L(:,i) >= 6,i)*1.1;
        Cand_Sig_L(:,Accept_Rate_L(:,i) <= 2,i) = Cand_Sig_L(:,Accept_Rate_L(:,i) <= 2,i)*.9;
        end
        Accept_Rate_G = zeros(NumSources,1);
        Accept_Rate_F = zeros(NumSources,NumDays);
        Accept_Rate_L = zeros(NumSources,NumDays);
        Rate_Counter = 0;
    end

    if ii > Burn
    %Set Lambda_t to new values of each Lambda Matrix
    Lambda_1(:,:,ii-Burn) = Lambda_t(:,:,1);

    %%Repeat for Lambda_2 through Lambda_50%%

    Fhold(:,:,ii-Burn) = F;
    Ghold(:,ii-Burn) = G;
    wpthold(ii-Burn,:) = wpt;
    end

end %End MCMC

    %%Initialize Matrices to Hold Percentiles
    Lambda_1_pct = zeros(NumSpecies,NumSources,101);
    %%Repeat for Lambda_2_pct through Lambda_50_pct%%

    %%For the Lambda Parameters, Keep Only the Quantiles
    for jj=1:NumSpecies
    for kk=1:NumSources
    Lambda_1_pct(jj,kk,:) = quantile(Lambda_1(jj,kk,:),0:.01:1,3);

    %%Repeat for Lambda_2_pct through Lambda_50_pct%%

    end
    end
    clear Lambda_1 Lambda_2 Lambda_3 Lambda_4 Lambda_5 Lambda_6 Lambda_7 Lambda_8;
    clear Lambda_17 Lambda_16 Lambda_15 Lambda_14 Lambda_13 Lambda_12 Lambda_11 Lambda_10 Lambda_9;
    clear Lambda_18 Lambda_19 Lambda_20 Lambda_21 Lambda_22 Lambda_23 Lambda_24 Lambda_25 Lambda_26;
```

```
clear Lambda_35 Lambda_34 Lambda_33 Lambda_32 Lambda_31 Lambda_30 Lambda_29 Lambda_28 Lambda_27;
clear Lambda_36 Lambda_37 Lambda_38 Lambda_39 Lambda_40 Lambda_41 Lambda_42 Lambda_43 Lambda_44;
clear Lambda_50 Lambda_49 Lambda_48 Lambda_47 Lambda_46 Lambda_45;

%%Output the Draws to A file
lamfile = ['./Output/G' int2str(gg) 'CV' int2str(cvout) '/Set' int2str(dd) '/Profiles/'];
dlmwrite([lamfile 'Lambda1.dat'],Lambda_1_pct,'precision','%.7f');

%%Repeat Write out for Lambda2.dat through Lambda_50.dat%%

MAEF = zeros(1,NumDays);
for kk=1:NumDays
    ffile = ['./Output/G' int2str(gg) 'CV' int2str(cvout) '/Set' int2str(dd)...
       '/Contributions/f' int2str(kk) '.dat'];
    temppct = zeros(NumSources,101);
    temp = Fhold(:,kk,:);
    for pp=1:NumSources
        temppct(pp,:) = quantile(temp(pp,:,:),0:.01:1);
    end
    dlmwrite(ffile,temppct,'precision','%.4f');
    MAEF(:,kk) = sum(abs(temppct(:,51)-trueF(:,kk)));
    clear temp;
end
temp = Ghold';
dlmwrite(['./Output/G' int2str(gg) 'CV' int2str(cvout) '/Set' int2str(dd)...
   '/G/g.dat'],temp,'precision','%.1f');
dlmwrite(['./Output/G' ...
          int2str(gg) 'CV' int2str(cvout) '/Set' int2str(dd) '/wpt/wpt.dat'],wpthold,'precision','%.3f');

%%Calculate MAE_Lambda by day
MAE1 = sum(abs(Lambda_1_pct(:,:,51)-trueL(:,:,1)));
%%Repeat for MAE2 through MAE50%%
TotMAE = [MAE1;MAE2;MAE3;MAE4;MAE5;MAE6;MAE7;MAE8;MAE9;MAE10;
          MAE11;MAE12;MAE13;MAE14;MAE15;MAE16;MAE17;MAE18;MAE19;MAE20;
          MAE21;MAE22;MAE23;MAE24;MAE25;MAE26;MAE27;MAE28;MAE29;MAE30;
          MAE31;MAE32;MAE33;MAE34;MAE35;MAE36;MAE37;MAE38;MAE39;MAE40;
          MAE41;MAE42;MAE43;MAE44;MAE45;MAE46;MAE47;MAE48;MAE49; ...
          MAE50];
dlmwrite(['./Output/G' int2str(gg) 'CV' int2str(cvout) ...
          '/MAE/ProfileMAE.dat'],TotMAE,'-append','precision','%.2f');
dlmwrite(['./Output/G' int2str(gg) 'CV' int2str(cvout) ...
```

```
            '/MAE/FMAE.dat'],MAEF,'-append','precision','%.2f');
  ['Completed']
  toc
end %End Do Loop Over Data Sets
end %End Do Loop Across Levels
```