2012-03-12

Bayesian Pollution Source Apportionment Incorporating Multiple Simultaneous Measurements

Jonathan Casey Christensen
Brigham Young University - Provo

Bayesian Pollution Source Apportionment Incorporating Multiple Simultaneous Measurements

Jonathan Christensen

A selected project submitted to the faculty of
Brigham Young University
in partial fulfillment of the requirements for the degree of

Master of Science

William Christensen, Chair
Natalie Blades
Del Scott

Department of Statistics

Brigham Young University

April 2012

ABSTRACT

Bayesian Pollution Source Apportionment Incorporating Multiple Simultaneous Measurements

Jonathan Christensen
Department of Statistics, BYU
Master of Science

We describe a method to estimate pollution profiles and contribution levels for distinct prominent pollution sources in a region based on daily pollutant concentration measurements from multiple measurement stations over a period of time. In an extension of existing work, we will estimate common source profiles but distinct contribution levels based on measurements from each station. In addition, we will explore the possibility of extending existing work to allow adjustments for synoptic regimes—large scale weather patterns which may effect the amount of pollution measured from individual sources as well as for particular pollutants. For both extensions we propose Bayesian methods to estimate pollution source profiles and contributions.

Keywords: Pollution Source Apportionment, Bayesian Methods, Multivariate Receptor Modeling

ACKNOWLEDGMENTS

I would like to thank William Christensen for taking a chance on me and giving me the opportunity to participate in this research, as well as his advice, recommendations, and comments throughout the process. I would also like to thank Wisconsin Focus on Energy for the grant which has supported this endeavor. Natalie Blades and Del Scott both made recommendations which resulted in a better final product, for which I thank them.

CONTENTS

CHAPTER 1

INTRODUCTION TO POLLUTION SOURCE APPORTIONMENT

The last two decades have seen increasing recognition of the health hazards posed by air pollutants, including particulate matter. As local and national governments have tightened environmental standards in response, interest has also increased in statistical methods which identify the primary pollution sources and estimate the characteristic profile of pollutants as well as the amount of pollution emitted by each. These statistical methods can be broadly grouped under the heading of pollution source apportionment (PSA). PSA methods range from regression models, in cases where the major sources of pollution in a region with their respective profiles are known, to purely exploratory models. Our approach in this project falls nearer the exploratory end of the spectrum, in that we do not assume strong prior knowledge of the sources and their profiles.

1.1 The PSA Model

The basic pollution source apportionment model can be expressed by the equation

$$\mathbf{Y} = \mathbf{\Lambda F} + \mathbf{E}, \tag{1.1}$$

where $\mathbf{Y}$ is the concentrations matrix, $\mathbf{\Lambda}$ is the profile matrix, $\mathbf{F}$ is the contributions matrix, and $\mathbf{E}$ is random error. That is, given a $p \times t$ matrix $\mathbf{Y}$ containing concentration measurements for p pollutants at each of t time periods, exploratory PSA methods return a factorization into matrices $\mathbf{\Lambda}$ and $\mathbf{F}$, which represent the source profiles and contribution levels and have dimensions $p \times k$ and $k \times t$, respectively. The number of sources k to be fit is determined in advance and is largely a judgment call which may be informed by expert knowledge. PSA methods which determine a suitable value for k as part of the process are a subject of continuing research interest.

For the output to be physically meaningful, the elements of $\mathbf{\Lambda}$ and $\mathbf{F}$ must all be nonnegative, as are the elements of $\mathbf{Y}$. In addition, the entries in each column of $\mathbf{\Lambda}$ are constrained to sum to 1, so as to represent the proportion which each pollutant makes up of the emissions from a given source.

When approaching PSA from a Bayesian perspective, we put priors on the individual elements of $\mathbf{\Lambda}$ and $\mathbf{F}$ and a likelihood for the elements of $\mathbf{Y}$ given $\mathbf{\Lambda F}$. Estimation is then performed using Markov Chain Monte Carlo methods.

1.2 A Caveat Regarding Estimability

It should be immediately noted that absent some amount of prior information on the sources the model is severely under-determined. In particular, the introduction of an arbitrary $k \times k$ orthogonal matrix $\mathbf{T}$ and its inverse (transpose) does not change the likelihood of $\mathbf{Y}$ given $\mathbf{\Lambda}$ and $\mathbf{F}$, since

$$\mathbf{\Lambda F} = \mathbf{\Lambda T T' F}.$$

Thus, the likelihood is invariant to reflections and rotations in the k-dimensional source space. The matrices which can be introduced here are limited by the requirement that $\mathbf{\Lambda T}$ and $\mathbf{T'F}$ still be element-wise nonnegative, but there are still infinitely many allowable transformations. To consider a special case, $\mathbf{T}$ can be any permutation matrix, indicating that the columns of $\mathbf{\Lambda}$ and the rows of $\mathbf{F}$ can be permuted arbitrarily. From a Bayesian point of view, this means that given identical prior information for each of the sources, as would seem natural to do in a purely exploratory setting, the posterior distributions for each of the sources are identical and multi-modal.

In practice the MCMC methods used to estimate the posterior distributions are very unlikely to estimate this multi-modality accurately. This is because the *conditional* posterior distributions for each source given the current draws for the other sources in the Gibbs sampler are typically unimodal and the algorithm gets “stuck” at this mode. Even given accurate estimation of the identical posterior, summarizing the information it contains in a useful way would require considerable effort—probably including a high-dimensional mixture deconvolution. To the best of our knowledge, this problem has not been explored in the literature.

As a result, we will not consider a purely exploratory model within the scope of this project. Instead, we will use weakly-informative priors based on a separate, non-Bayesian analysis by James Schauer at the University of Wisconsin at Madison (Schauer, 2011).

1.3 Assumptions

Scientifically, the fundamental assumption of the PSA model is that pollutants are measured at the same (relative) concentrations at which they are emitted. This is known as the mass-balance assumption, and implies that different pollutants spread through the air in the same way and that individual pollutants are not created or destroyed between emission and measurement. In fact, both of these conditions are at best questionable: pollutants may diffuse differently and react with other pollutants or oxygen and water vapor in the air. Moreover, the extent to which the assumption is violated may vary over time depending on atmospheric conditions. Unfortunately, this violation is intractable: we can only perform an analysis on the pollution levels which we measure, however those may differ from the pollutants actually emitted. At best attempts can be made to minimize the impact of this violation, as in Park, Guttorp, and Henry (2001), where, in an analysis of volatile organic compounds, species were specifically chosen which had long half-lives relative to the estimated time between emission and measurement.

Two additional statistical assumptions should be noted. The fundamental statistical assumption which allows us to perform PSA is that source profiles do not vary between measurement periods, so that the profile for any given source is the same throughout the data collection period. If the proportions of pollutants emitted from a given source change either gradually or abruptly during the measurement time frame PSA methods may not accurately reflect the contributions and profile of that source. This may also result in inaccurate estimates of the profiles and contributions of other sources, as the method attempts to adjust to account for all the observed pollution. Heaton, Reese, and Christensen (2010) use a Bayesian model which allows some variation in the profile over time, but this lies beyond the scope of this project.

Second, most PSA methods implicitly assume independence of observations over time, an assumption which is often inappropriate. Park, Guttorp, and Henry (2001) address this assumption using a time series approach; for the sake of simplicity, we do not attempt to model temporal autocorrelation in this project.

CHAPTER 2

LITERATURE REVIEW

2.1 Non-Bayesian PSA

Early work on pollution source apportionment included simple multivariate regression methods, requiring knowledge of the relevant sources and their respective profiles to estimate contribution levels.

Miller *et al.* (1972) conduct an analysis of source contributions in the Pasadena, California area using known profiles for four sources (sea salt, soil dust, automobile emissions, and fuel oil), and cite earlier work indicating that as much as a third of anthropogenic pollutants measured in the Los Angeles area come from secondary sources such as atmospheric reactions, casting considerable doubt on the validity of the mass-balance assumption.

Gatz (1975) applies the methods of Miller *et al.* (1972) to calculate pollution contribution levels in the Chicago area, based on data collected at a number of sites in the city and published source profiles for six sources, including automobile exhaust, four industrial sources, and soil dust. He notes that the mass-balance assumption is probably violated, as "a systematic change in elemental relative abundances" is "known or suspected to occur" in air-borne sea spray and soil dust. He shrugs the difficulty off, however, noting that the effects have "not been expressed quantitatively and will be ignored in this paper."

Where the sources are known and estimates of their profiles are available, Watson *et al.* (1984) provide an iteratively reweighted least squares procedure known as effective variance (EV), which makes use of the estimated profiles and the levels of uncertainty about the individual entries.

In cases where the source profiles are not known in advance, factor analysis models were used by Thurston and Spengler (1985) in an analysis of particulate matter in Boston and Koutrakis and Spengler (1987) for a site in Ohio. These approaches suffer from a lack of uniqueness as discussed above. Paatero and Tapper (1994) propose the positive matrix factorization method, which has been extended (Paatero 1998) to allow ad hoc incorporation of expected source profiles in order to provide identifiability.

A series of papers in the 1990s addresses issues with uniqueness by proposing various conditions that provide identifiability, including setting some elements of the profile matrix equal to zero. By making the model conditional on these additional assumptions, however, these methods lessen the appeal. Park *et al.* (2002) attempt to address this issue, as well as the issue of determining the number of sources k, by fitting a series of models with different identifiability conditions within a Bayesian framework and calculating which model has the highest posterior probability.

2.2 Bayesian PSA

Park *et al.* (2001) introduce a Bayesian approach to estimating pollution source profiles and contributions, including an autoregressive component to account for temporal correlation in the data. The flexibility of a Bayesian approach allows it to deal with a range of situations, from cases where the researcher is confident that the primary sources and their profiles are known with considerable accuracy to an exploratory analysis, though we face problems of identifiability in purely exploratory analyses as discussed above.

Lingwall, Christensen, and Reese (2008) propose a Bayesian model which is the foundation of this work, and show that it compares favorably to Positive Matrix Factorization. Heaton, Reese, and Christensen (2010) extend this work to allow variation of the profile associated with a single source over time, relaxing the assumption that source profiles are constant.

CHAPTER 3

METHODOLOGICAL DEVELOPMENTS

In this project we propose two extensions to existing work in Bayesian PSA:

1. Dealing with pollution measurements from multiple sites in the same region, fitting common sources but allowing contribution levels to vary by site.

2. Exploring the relationship between synoptic regimes, or large-scale weather patterns, and pollution contributions, and whether this additional information can be exploited to better estimate the contributions themselves.

A Bayesian model is proposed for the first extension. For the second, we propose a *post hoc* statistical analysis on the estimated profiles, as well as a Bayesian model which may potentially help include the information about synoptic regimes in the estimation process. Bayesian inference will be performed using MCMC code written in MATLAB to generate draws from the posterior distributions. Data have been provided for pollution measurements from three sites in the Milwaukee, Wisconsin area, as well as synoptic regime data covering the same time period, under the scope of a grant from Wisconsin Focus on Energy.

Both of these avenues appear to have been little explored in the literature. Gatz (1975) performed analyses using data from multiple sites in Chicago, but his work assumed perfect knowledge of the source profiles and did not combine information from different sites to improve estimation.

3.1 Multiple Locations

We propose an approach to estimate pollution source profiles and contributions when pollution measurements are available from several locations in the same region (i.e., close enough that the major pollution sources are expected to be the same). In particular, given data matrices $\mathbf{Y}_1, \mathbf{Y}_2, \ldots, \mathbf{Y}_m$ from the same region, we propose the model

$$\begin{aligned}
\mathbf{Y}_1 &= \mathbf{\Lambda F}_1 + \mathbf{E}_1 \\
\mathbf{Y}_2 &= \mathbf{\Lambda F}_2 + \mathbf{E}_2 \\
&\vdots \\
\mathbf{Y}_m &= \mathbf{\Lambda F}_m + \mathbf{E}_m,
\end{aligned}$$

where we estimate common profiles across all m sites but separate contributions matrices. This model fits with the assumption of source profile invariance, while making allowance for the fact that different amount of pollution from each source will end up at each measurement station due to geography and possible station-specific effects, such as local air flow patterns.

As long as the data from the locations are reasonably consistent, we expect that we will obtain not only better estimates of Λ but also better estimates of the contribution matrices for each location than if we were to make use of data only from that location.

Under the Wisconsin Focus on Energy grant, we have been provided data from three EPA measurement sites in the Milwaukee area (Milwaukee, Waukesha, and Mayfield). These data will be used for this portion of the project.

3.2 Synoptic Regimes

Because of the potentially large affect of weather patterns on air pollution, we will explore whether there is a relationship between the synoptic regimes identified by a meteorologist and the estimated source contributions. As an initial investigation, we will perform a simple MANOVA using the synoptic regime classifications for each measurement period as a independent variable and the estimated contribution levels as responses. The synoptic regimes may effect pollution measurements in several ways; the easiest to identify would be source masking, in which particular synoptic regimes cause the pollution emitted by one or more sources not to reach the measurement stations. For example, particulate matter may be washed out of the air by a rain storm, or prevailing winds may blow it away from the measurement station. In these cases, though the source may still be emitting pollution, it will not be measured at typical levels. This type of effect should be easily identified by the MANOVA procedure if the model estimation is able to pick it up.

If this analysis indicates that the synoptic regimes are indeed related to weather patterns and may have such a masking effect, we will consider another extension to the basic model which allows for the estimation of these masking effects. Namely, for each time period a classification is provided into one of the r synoptic regimes. For simplicity we limit ourselves to a single site for this analysis. We construct the $r \times t$ regime matrix $\mathbf{R}$ as a binary matrix where the the (i, j) entry is 1 if time period j is classified into regime i and 0 otherwise. Thus, 1 appears exactly once in each column of $\mathbf{R}$.

We also introduce a $k \times r$ matrix $\mathbf{M}$, which we call the *masking* matrix. This is also a binary matrix, but is estimated during the MCMC process. A 1 in the (i, j) entry of $\mathbf{M}$ indicates that the ith source is measured normally under the jth regime; a 0 indicates that under the jth regime that source is masked.

The basic model is then given by

$$\mathbf{Y} = \mathbf{\Lambda} \left[\mathbf{F} \circ (\mathbf{M}\mathbf{R})\right] + \mathbf{E}, \tag{3.1}$$

where $\circ$ indicates the Hadamard product (element-wise multiplication). In effect, for each element of $\mathbf{F}$ this model looks up whether the source in question was masked at the given time period, and zeros out that location if so. This model may be estimated with standard MCMC techniques.

CHAPTER 4

DATA ANALYSIS

4.1 Inference using Multiple Sites

We were able to successfully fit the multiple sites model described using data from the Milwaukee, Waukesha, and Mayfield measurement stations. Because data was collected at different intervals at the different sites, only 343 observations were available in the collection period between 2002 and 2008 with data for all three sites; this yielded roughly one observation per week.

In order to evaluate the trade-off between using sparser data from all three sites and using more frequent measurements from a single site, we fit a model to the 789 observations from the Milwaukee site, which has the most observations (one every three days). Both models used the same prior information on $\mathbf{\Lambda}$, which is of primary interest, and the same non-distinguishing priors on the elements of $\mathbf{F}$. On average, equal-tail 95% credible intervals for the elements of $\mathbf{\Lambda}$ from the model using data from all three sites were about 85% as wide as equivalent credible intervals from the model using data from just the Milwaukee site. It is clear that drawing observations from multiple sites can increase model estimation precision even when working with a much smaller data set. If multiple sites with the same frequency of measurements were available, this would allow fitting a model and obtaining equivalent precision using data from a much shorter time span than if only data from one site were used, which would help mitigate the effects of violations of the assumption of profile invariance over time.

Other than having, on average, narrower credible intervals, the posteriors of the elements of the $\mathbf{\Lambda}$ matrix estimated using data from multiple sites match those estimated from the data for a single site very closely. The correlations of posterior means, medians, and the endpoints of equal-tail 95% credible intervals are all above 0.99, and very few of the posterior means differ by more than 0.01 between the two estimates. The estimated $\mathbf{\Lambda}$ matrices using all three sites and using just the Milwaukee data are included as an appendix.

The estimated contribution levels for all three sites are plotted in Figure 4.1. Major tick marks are placed at January 1 of each year. Because the details of these results are not of primary interest in this project, we will not discuss them at length; however, we note two interesting features of the estimated contributions. There is a clear seasonal pattern in the contribution levels of the Secondary Nitrate source, with higher contribution levels during the winter months. In addition, we note the relatively low contributions by industry and high contributions by soil dust at the Mayfield site, which is more rural than either of the Milwaukee and Waukesha sites.

4.2 Synoptic Regimes

MANOVA Analysis

To determine whether synoptic regimes appear to have an effect on measured pollution concentration levels we fit a Bayesian PSA model following Lingwall, Christensen, and Reese (2008), using weakly informative priors on $\mathbf{\Lambda}$ based on Schauer's analysis mentioned previously. Based on Schauer's work, we fit a model using nine sources to predict the concentrations of twenty-eight pollutants measured at the Milwaukee measuring station. The mean of the posterior distribution for the contribution level is used as a point estimate for the contributions measured for each source on each day. We then perform a MANOVA analysis using these point estimates for each of the nine sources as response variables and the regime classifications, provided by Benjamin de Foy, a collaborator on the Wisconsin Focus on Energy grant, as the predictor variable[1]. Both the overall MANOVA test and individual ANOVA

[1] Each measurement day is classified into one of eight regimes; unfortunately, without access to either the source data used for classification or descriptions of the regimes, I am unable to provide any insight into the characteristics of the regimes apart from the numbers one through eight.

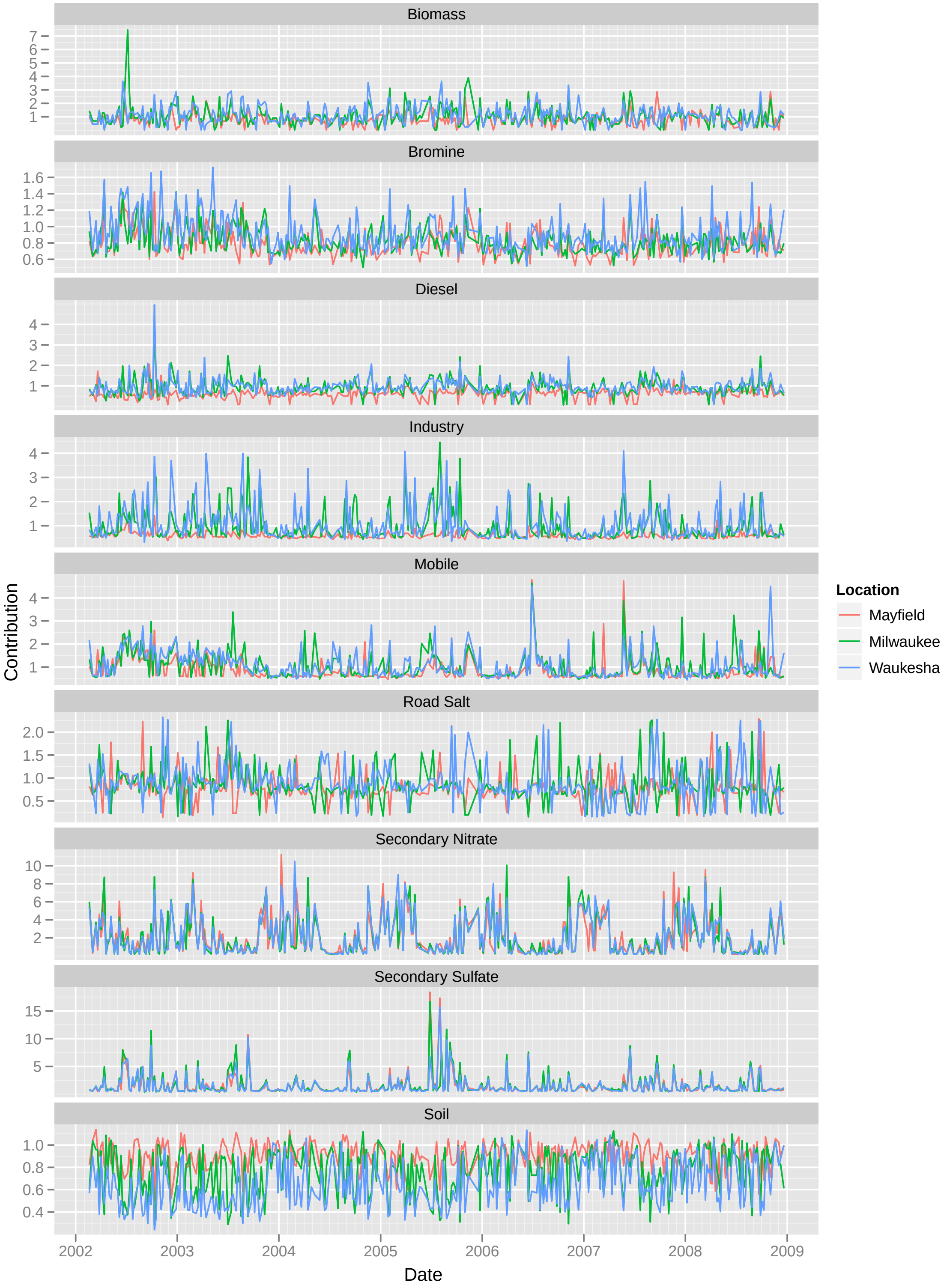

Figure 4.1: Contribution levels for each of the three sites as estimated using the common model.

tests on each source's contributions are highly significant, with all p-values < 0.0001. Plots of the estimated coefficients, showing clear regime effects, are given in Figure 4.3.

The first two discriminant functions, accounting for over 90% of the total variation, are given in Table 4.1. The first discriminant function is driven by the contrast between the generic bromine source and the sum of the soil dust and diesel fuel sources. The second discriminant function is primarily a contrast between soil dust and automobile exhaust. Unfortunately, neither of these discriminant functions offers obvious insight into the behavior of the sources.

Table 4.1: The first two discriminant functions for the relationship between synoptic regimes and source contribution levels as measured in Milwaukee.

	z_1	z_2
Secondary Nitrate	-0.13	0.03
Secondary Sulfate	-0.10	0.06
Biomass	-0.18	-0.02
Mobile	-0.01	0.53
Soil	-0.24	-0.76
Road Salt	-0.05	0.00
Industry	0.01	0.33
Bromine	0.52	0.08
Diesel	-0.25	0.35

A plot of the mean discriminant scores for each of the regimes is given in Figure 4.2. We note the clustering of regimes 1, 3, and 8, and the distance of regime 5, in particular, from the others.

We note that while synoptic regimes do have an effect on measured pollution data, none of the estimates have credible intervals extending below about $0.4\mu g/m^3$. This may be the result of a real lack of a complete masking effect or may be an effect of the model used, as the log normal prior on the elements of F will tend to pull low estimates away from zero. In any case, this analysis clearly indicates that regime data can help estimate contribution levels.

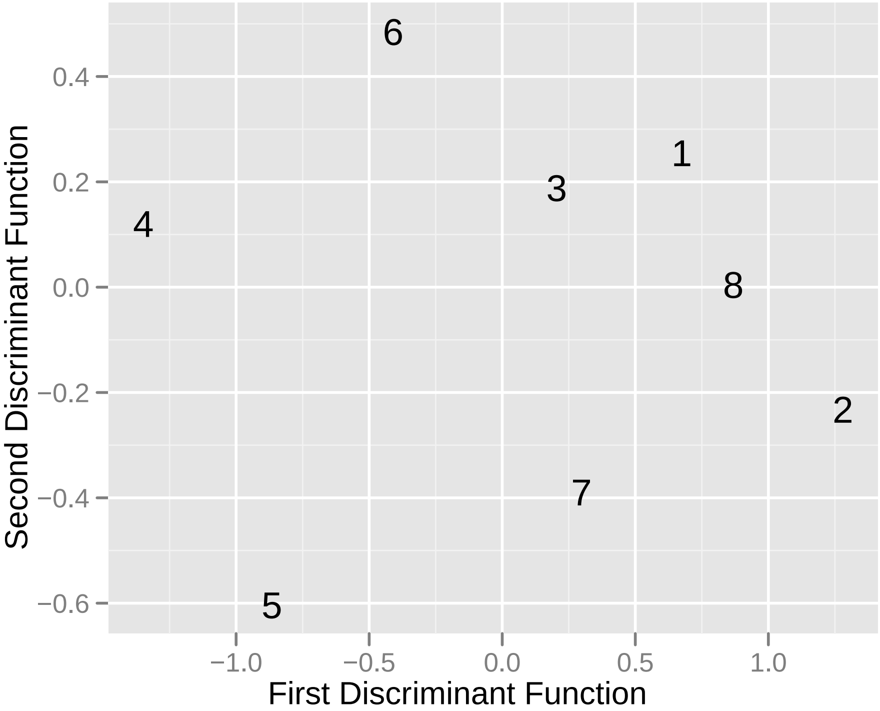

Figure 4.2: Mean discriminant scores of each of the eight regimes for the first two discriminant functions.

A Bayesian Model Incorporating Synoptic Regimes

We attempt to capitalize on this knowledge by fitting the model described in equation (2) above. In particular, we let

$$\begin{aligned}
\mathbf{Y} &\sim \mathrm{LN}(\mathbf{\Lambda F}, Y_{\text{error}}) \\
\mathbf{\Lambda} &\sim \mathrm{Dirichlet}(\alpha) \\
\mathbf{F}_{ij} &\sim \begin{cases} \mathrm{LN}(\mathbf{F}_{\text{prior}}, \sigma^2_{\text{prior}}) & (MR)_{ij} = 1 \\ \mathrm{Exp}(\mathbf{F}^*) & (MR)_{ij} = 0 \end{cases} \\
\mathbf{M} &\sim \mathrm{Bernoulli}(\pi_0).
\end{aligned}$$

This parameterization differs slightly from that originally proposed in that rather than zeroing **F** out entirely when the source is masked we replace the log normal distribution, which

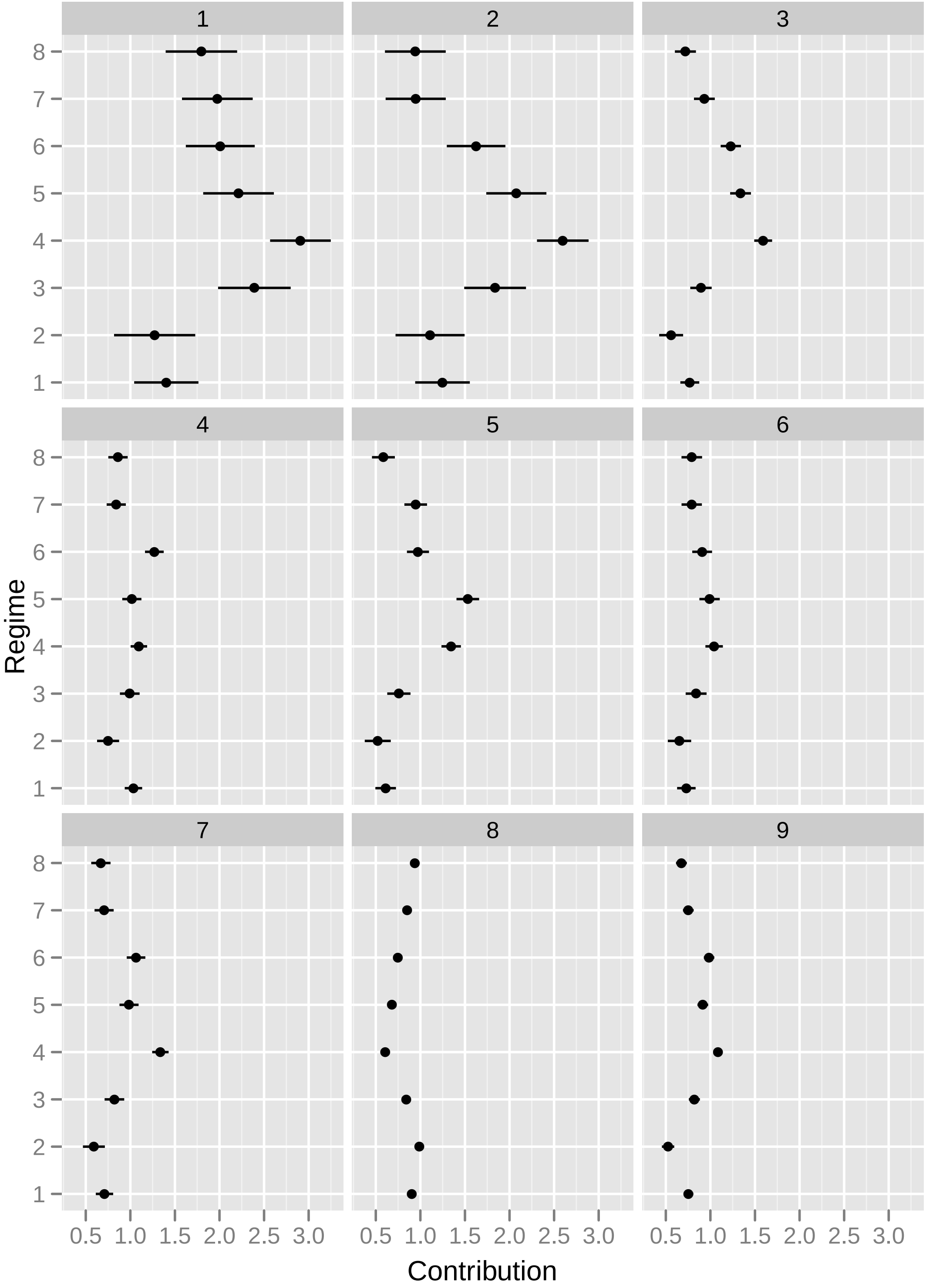

Figure 4.3: Estimated regime effects (point estimates and 95% confidence intervals) for each of the nine sources.

will pull estimates away from zero, with an exponential distribution, which will push estimates toward zero. This approach retains the ability to estimate the masking matrix while also allowing regimes to mask a source only partially. The exponential distribution is chosen to cross with the previously chosen log normal prior at an arbitrary point (in this case, $\frac{1}{3}\mu g/m^3$), and so may vary from element to element if informative priors on the contributions matrix are used. The crossing of the log normal and exponential distributions is illustrated in Figure 4.4. Thus, if the estimated contribution level for a given source on a given day is below $1/3\mu g/m^3$, the estimate will be pushed toward zero and the model will favor masking that source under the regime to which that day belongs; otherwise, the estimate will be pulled away from zero and the model will favor not masking that source. This approach has been influenced in part by Johnson's work on non-local priors (2010). Unfortunately, in practice I have been unable to fit this model (or the slightly simpler one originally proposed) and obtain anything resembling MCMC convergence with either real or toy datasets.

4.3 Summary

In this project we have pursued two extensions to existing work in Bayesian Pollution Source Apportionment. In the first part we have extended existing models to allow for incorporation of data from multiple measurement sites in the same area. We show that by including multiple sites we can use far fewer observations—even with fewer than half as many observations, credible intervals estimated using the data from three sites were on average only 85% as wide as credible intervals estimated using the data from a single site, and posterior means differed little between the two estimation methods. If multiple sites with the same frequency of measurement were used, the time range of the data needed would be cut considerably, allowing for more confidence in the assumption of profile invariance over time. Further work may investigate the possibility of using all the available data from each site, even if measurements do not line up, as when measurements are taken at different frequencies at different sites.

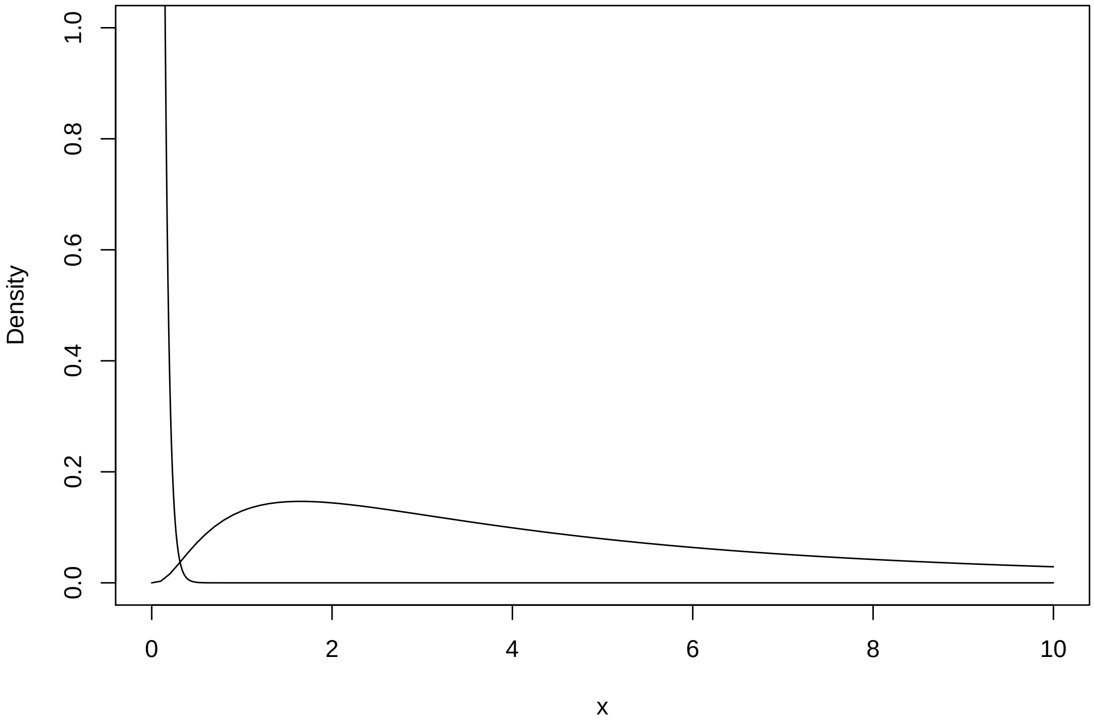

Figure 4.4: The two components of the prior distribution on the elements of **F**. The exponential distribution pushes small estimates toward zero and favors masking; the log normal distribution pulls slightly larger estimates away from zero and favors not masking.

In the second part we investigated the relationship between weather patterns, using a classification of each measurement period into one of eight synoptic regimes, and source contribution levels estimated from the Milwaukee data. We found that the classification into synoptic regimes is predictive of contribution levels. Unfortunately, due to technical difficulties with the MCMC procedure we were unable to fit our proposed model and estimate source masking effects.

It is unclear what caused this failure to fit the source-masking model. It may be that the introduction of the source masking effects leads to a lack of identifiability in the model beyond the difficulties previously noted with the **F** and $\mathbf{\Lambda}$ matrices. Perhaps related to this, the alternative priors on the elements of **F** may result in severe multimodality in the posteriors, causing problems with the MCMC sampling procedure. It is possible that the difficulties encountered are not issues with the model, and may be overcome by implementing a more sophisticated sampling algorithm to find the posterior; however, we suspect that investigation of alternative model formulations (such as a regression model with regime effects for each source) will be more fruitful. Though our current approach has been unsuccessful, there appears to be no fundamental reason why we should not be able to incorporate regime data in our model.

CHAPTER 5

APPENDIX I: ESTIMATED PROFILES

Table 5.1: Posterior mean of $\boldsymbol{\Lambda}$ matrix estimated using only observations from the Milwaukee site. The first two sources are Secondary Nitrate and Secondary Sulfate, respectively.

	Sec. Nit.	Sec. Sulf.	Biomass	Mobile	Soil	Road Salt	Industry	Bromine	Diesel
OC	0.03459	0.07302	0.61164	0.88001	0.54950	0.48498	0.49926	0.52894	0.30406
EC	0.00487	0.00757	0.04149	0.01233	0.04179	0.03960	0.04783	0.02142	0.46379
NO3-	0.93197	0.01385	0.05223	0.01662	0.04249	0.06261	0.05483	0.04642	0.02919
SO42-	0.02345	0.89791	0.21637	0.07557	0.25400	0.36650	0.30689	0.28024	0.17074
NH4+	0.00226	0.00435	0.01291	0.00775	0.01293	0.01700	0.01557	0.02567	0.00891
Sb	0.00011	0.00008	0.00074	0.00031	0.00087	0.00122	0.00121	0.00882	0.00162
As	0.00000	0.00000	0.00006	0.00000	0.00010	0.00013	0.00008	0.00089	0.00020
Al	0.00034	0.00014	0.00087	0.00052	0.00094	0.00094	0.00088	0.01111	0.00307
Br	0.00003	0.00000	0.00001	0.00001	0.00001	0.00002	0.00001	0.00109	0.00001
Ca	0.00014	0.00041	0.00474	0.00083	0.02051	0.00118	0.00868	0.00106	0.00110
Cr	0.00008	0.00006	0.00064	0.00023	0.00081	0.00132	0.00123	0.00865	0.00174
Cu	0.00000	0.00000	0.00004	0.00001	0.00002	0.00012	0.00001	0.00103	0.00005
Cl	0.00000	0.00001	0.00003	0.00001	0.00004	0.00704	0.00003	0.00005	0.00002
Fe	0.00087	0.00094	0.00771	0.00189	0.01243	0.00510	0.05442	0.00455	0.00544
Pb	0.00000	0.00000	0.00001	0.00000	0.00002	0.00011	0.00001	0.00102	0.00001
Mn	0.00000	0.00000	0.00001	0.00001	0.00001	0.00005	0.00001	0.00089	0.00001
Mg	0.00018	0.00017	0.00126	0.00051	0.00155	0.00150	0.00138	0.01066	0.00175
Se	0.00000	0.00000	0.00007	0.00000	0.00009	0.00014	0.00009	0.00089	0.00020
Ti	0.00001	0.00000	0.00004	0.00001	0.00002	0.00015	0.00001	0.00097	0.00015
V	0.00000	0.00000	0.00006	0.00000	0.00009	0.00013	0.00008	0.00088	0.00023
Si	0.00019	0.00027	0.00130	0.00044	0.05397	0.00107	0.00108	0.00088	0.00054
Zn	0.00000	0.00000	0.00001	0.00001	0.00001	0.00003	0.00001	0.00040	0.00001
Sr	0.00000	0.00000	0.00010	0.00000	0.00006	0.00014	0.00005	0.00088	0.00024
Tb	0.00000	0.00000	0.00004	0.00000	0.00010	0.00015	0.00009	0.00088	0.00018
Rb	0.00000	0.00000	0.00007	0.00000	0.00008	0.00014	0.00009	0.00087	0.00022
K	0.00003	0.00037	0.04200	0.00003	0.00227	0.00010	0.00011	0.00007	0.00005
Na	0.00087	0.00082	0.00548	0.00290	0.00522	0.00841	0.00597	0.03985	0.00628
Zr	0.00000	0.00000	0.00007	0.00000	0.00007	0.00015	0.00008	0.00089	0.00021

Table 5.2: Posterior mean of $\mathbf{\Lambda}$ matrix estimated using observations from all three sites. The first two sources are Secondary Nitrate and Secondary Sulfate, respectively.

	Sec. Nit.	Sec. Sulf.	Biomass	Mobile	Soil	Road Salt	Industry	Bromine	Diesel
OC	0.02446	0.06258	0.53628	0.89619	0.47868	0.50227	0.57831	0.54184	0.33642
EC	0.00442	0.00568	0.03715	0.00964	0.01490	0.02426	0.03283	0.02614	0.46634
NO3-	0.94837	0.01400	0.05484	0.01597	0.04602	0.05324	0.04525	0.05446	0.02625
SO42-	0.01872	0.91105	0.29372	0.06496	0.33254	0.37350	0.23289	0.32815	0.14593
NH4+	0.00189	0.00398	0.01287	0.00687	0.02327	0.01617	0.01343	0.01752	0.00838
Sb	0.00007	0.00004	0.00081	0.00027	0.00772	0.00181	0.00137	0.00246	0.00090
As	0.00000	0.00000	0.00007	0.00000	0.00077	0.00017	0.00012	0.00027	0.00011
Al	0.00024	0.00012	0.00125	0.00037	0.00971	0.00249	0.00098	0.00291	0.00134
Br	0.00001	0.00000	0.00003	0.00001	0.00082	0.00016	0.00004	0.00032	0.00001
Ca	0.00009	0.00085	0.00544	0.00086	0.00091	0.00119	0.02345	0.00129	0.00101
Cr	0.00004	0.00004	0.00068	0.00022	0.00744	0.00179	0.00145	0.00236	0.00121
Cu	0.00000	0.00000	0.00001	0.00000	0.00101	0.00001	0.00000	0.00001	0.00000
Cl	0.00000	0.00000	0.00002	0.00001	0.00003	0.00436	0.00002	0.00002	0.00001
Fe	0.00075	0.00080	0.00554	0.00152	0.00410	0.00485	0.05957	0.00486	0.00532
Pb	0.00000	0.00000	0.00001	0.00000	0.00100	0.00010	0.00001	0.00002	0.00001
Mn	0.00000	0.00001	0.00001	0.00001	0.00098	0.00004	0.00001	0.00003	0.00001
Mg	0.00011	0.00011	0.00096	0.00042	0.00959	0.00231	0.00256	0.00265	0.00109
Se	0.00000	0.00000	0.00005	0.00000	0.00082	0.00018	0.00013	0.00029	0.00003
Ti	0.00000	0.00000	0.00003	0.00000	0.00090	0.00015	0.00001	0.00023	0.00005
V	0.00000	0.00000	0.00007	0.00000	0.00077	0.00017	0.00013	0.00027	0.00010
Si	0.00013	0.00011	0.00056	0.00031	0.01761	0.00077	0.00046	0.00080	0.00033
Zn	0.00000	0.00000	0.00001	0.00001	0.00043	0.00003	0.00001	0.00004	0.00001
Sr	0.00000	0.00000	0.00005	0.00000	0.00077	0.00017	0.00013	0.00028	0.00012
Tb	0.00000	0.00000	0.00002	0.00000	0.00078	0.00015	0.00014	0.00028	0.00013
Rb	0.00000	0.00000	0.00007	0.00000	0.00077	0.00017	0.00012	0.00026	0.00012
K	0.00000	0.00001	0.04375	0.00001	0.00002	0.00002	0.00002	0.00002	0.00001
Na	0.00069	0.00062	0.00568	0.00235	0.03684	0.00927	0.00644	0.01194	0.00469
Zr	0.00000	0.00000	0.00006	0.00000	0.00078	0.00018	0.00013	0.00027	0.00008

CHAPTER 6

APPENDIX II: CODE

6.1 MCMC.M

```
% Bayesian PSA code by Jonathan Christensen

%% Load data

cd C:/Research/Final;
Y = load(’mil.csv’)’;
Y_unc = load(’mil_unc.csv’)’;
Y(Y==0) = 0.001; % Since we take logs of Y, Y==0 causes problems.
Y_unc(Y_unc==0) = 0.001;

%% Model parameters

[p, t] = size(Y);
k = 9;

%% F matrix
% F is the contribution matrix. Rows correspond to the postulated sources.
% Columns correspond to time points (measurements). The priors on the
% elements of F are lognormal.

F_prior_mean = ones(k,t);
```

```
F_prior_var = ones(k,t);
F = F_prior_mean;

%% Lambda matrix;
% The Lambda matrix represents the source profiles. Rows correspond to
% measured species; columns correspond to postulated sources, and sum to 1.

L_prior_mean = load(’Lambda_prior_mean.csv’);
L_prec = 1000*ones(1,k);
L = L_prior_mean;

%% MCMC parameters

L_cand = 0.2*L;
F_cand = 0.2*F;

thin = 10;
adjust = 100;

%% Acceptance rates

L_accept = zeros(p,k);
F_accept = zeros(size(F));

%% MCMC

burnin = 0;
```

```
length = 2000;

for i = 2:(burnin+length)
   % update Lambda
   for j = 1:k
       for h = 1:p
           L_new = L;
           L_new(h,j) = normrnd(L(h,j),L_cand(h,j));
           if L_new(h,j) > 0;
               scaling = sum(L_new(:,j));
               L_new(:,j) = L_new(:,j)/scaling;
               L_rat = L_cc(L_new,L_prior_mean,L_prec,F,Y,Y_unc,j) - ...
                       L_cc(L,L_prior_mean,L_prec,F,Y,Y_unc,j);
               if log(rand(1)) < L_rat;
                   L = L_new;
                   L_accept(h,j) = L_accept(h,j) + 1;
               end
           end
       end
   end

   % update F

   for j = 1:k
       F_new = F;
       F_new(j,:) = normrnd(F(j,:),F_cand(j,:));
       neg = (F_new(j,:) <= 0);
```

```
    F_new(j,neg) = F(j,neg);
    F_rat = F_cc(F_new,F_prior_mean,F_prior_var,L,Y,Y_unc,j) - ...
            F_cc(F,F_prior_mean,F_prior_var,L,Y,Y_unc,j);
    prob = log(rand(1,t));
    F(j,(prob < F_rat)) = F_new(j,(prob < F_rat));
    F_accept(j, (prob<F_rat & ~neg)) = ...
      F_accept(j, (prob<F_rat & ~neg)) + 1;
end

% write to file

if i > burnin && mod(i,thin)==0
    dlmwrite(’output/Lambda.txt’,L,’-append’,’roffset’,1);

    for j = 1:k
         dlmwrite(strcat(’output/f’,int2str(j),’.txt’),...
           F(j,:),’-append’,’roffset’,1);
    end
end

% update tuning parameters

if mod(i,adjust) == 0
     i
     if i<=burnin
         for j = 1:k
             for h = 1:p
```

```
                    if L_accept(h,j)/adjust < 0.2
                        L_cand(h,j) = L_cand(h,j).*0.9;
                    elseif L_accept(h,j)/adjust > 0.45
                        L_cand(h,j) = L_cand(h,j).*1.1;
                    end
                end
                for h = 1:t
                    if F_accept(j,h)/adjust < 0.2
                        F_cand(j,h) = F_cand(j,h).*0.9;
                    elseif F_accept(j,h)/adjust > 0.45
                        F_cand(j,h) = F_cand(j,h).*1.1;
                    end
                end
            end
            L_accept = zeros(p,k);
            F_accept = zeros(k,t);
        end
    end
end
```

6.2 F_cc.m

```
function cc  = F_cc( F, F_prior_mean, F_prior_var, L, Y, Y_unc, j )
%lambda_cc Calculates complete conditional for Lambda

prior = -log(F(j,:)) - (log(F(j,:)) - F_prior_mean(j,:)) .^ 2 ./ ...
  F_prior_var(j,:);
```

```
m = L * F;
logfrac = log((m .^ 2 + Y_unc .^ 2)./(m .^ 2));
likelihood = sum(-log(Y) - (log(Y) - log(m) + 0.5*logfrac).^2 ./ ...
  (2 * logfrac));

cc = prior + likelihood;

end
```

6.3 L_cc.m

```
function  cc  = L_cc( L, L_prior_mean, L_prec, F, Y, Y_unc, j )
%lambda_cc Calculates complete conditional for Lambda

prior = log_dirichlet_pdf(L(:,j),L_prior_mean(:,j)*L_prec(j));

m = L * F;
logfrac = log((m .^ 2 + Y_unc .^ 2)./(m .^ 2));
likelihood = sum(sum(-log(Y) - (log(Y) - log(m) + 0.5*logfrac).^2 ./ ...
  (2 * logfrac)));

cc = prior + likelihood;

end
```

6.4 log_dirichlet_pdf.m

```
function dirich = dirichlet_pdf(x,alpha)
```

```
dirich = sum((alpha-1).*log(x));
```

6.5 mcmc3.m

```
% Bayesian PSA code by Jonathan Christensen


%% Load data


cd C:/Research/Final;

Y1 = load('mil3.csv')';

Y2 = load('wau3.csv')';

Y3 = load('may3.csv')';

Y1_unc = load('mil3_unc.csv')';

Y2_unc = load('wau3_unc.csv')';

Y3_unc = load('may3_unc.csv')';

Y1(Y1==0) = 0.001; % Since we take logs of Y, Y==0 causes problems.

Y2(Y2==0) = 0.001;

Y3(Y3==0) = 0.001;

Y1_unc(Y1_unc==0) = 0.001;

Y2_unc(Y2_unc==0) = 0.001;

Y3_unc(Y3_unc==0) = 0.001;


%% Model parameters


[p, t] = size(Y1);

k = 9;


%% F matrix
```

```
% F is the contribution matrix. Rows correspond to the postulated sources.
% Columns correspond to time points (measurements). The priors on the
% elements of F are lognormal.

F_prior_mean = ones(k,t);
F_prior_var = ones(k,t);
F1 = F_prior_mean;
F2 = F_prior_mean;
F3 = F_prior_mean;

%% Lambda matrix;
% The Lambda matrix represents the source profiles. Rows correspond to
% measured species; columns correspond to postulated sources, and sum to 1.

L_prior_mean = load('Lambda_prior_mean.csv');
L_prec = 1000*ones(1,k);
L = L_prior_mean;

%% MCMC parameters

L_cand = 0.2*L;
F1_cand = 0.2*F1;
F2_cand = 0.2*F2;
F3_cand = 0.2*F3;

thin = 10;
adjust = 100;
```

```
%% Acceptance rates

L_accept = zeros(p,k);
F1_accept = zeros(size(F1));
F2_accept = zeros(size(F2));
F3_accept = zeros(size(F3));

%% MCMC

burnin = 0;
length = 2000;

for i = 2:(burnin+length)
   % update Lambda
   for j = 1:k
       for h = 1:p
           L_new = L;
           L_new(h,j) = normrnd(L(h,j),L_cand(h,j));
           if L_new(h,j) > 0;
               scaling = sum(L_new(:,j));
               L_new(:,j) = L_new(:,j)/scaling;
               L_rat = L_cc3(L_new,L_prior_mean,L_prec,F1,F2,F3,...
                         Y1,Y2,Y3,Y1_unc,Y2_unc,Y3_unc,j) - ...
                       L_cc3(L,L_prior_mean,L_prec,F1,F2,F3,Y1,Y2,Y3,...
                         Y1_unc,Y2_unc,Y3_unc,j);
               if log(rand(1)) < L_rat;
```

```
                L = L_new;
                L_accept(h,j) = L_accept(h,j) + 1;
            end
        end
    end
end

% update F1

for j = 1:k
    F_new = F1;
    F_new(j,:) = normrnd(F1(j,:),F1_cand(j,:));
    neg = (F_new(j,:) <= 0);
    F_new(j,neg) = F1(j,neg);
    F_rat = F_cc(F_new,F_prior_mean,F_prior_var,L,Y1,Y1_unc,j) - ...
            F_cc(F1,F_prior_mean,F_prior_var,L,Y1,Y1_unc,j);
    prob = log(rand(1,t));
    F1(j,(prob < F_rat)) = F_new(j,(prob < F_rat));
    F1_accept(j, (prob<F_rat & ~neg)) = ...
      F1_accept(j, (prob<F_rat & ~neg)) + 1;
end

% update F2

for j = 1:k
    F_new = F2;
    F_new(j,:) = normrnd(F2(j,:),F2_cand(j,:));
    neg = (F_new(j,:) <= 0);
    F_new(j,neg) = F2(j,neg);
    F_rat = F_cc(F_new,F_prior_mean,F_prior_var,L,Y2,Y2_unc,j) - ...
            F_cc(F2,F_prior_mean,F_prior_var,L,Y2,Y2_unc,j);
    prob = log(rand(1,t));
    F2(j,(prob < F_rat)) = F_new(j,(prob < F_rat));
    F2_accept(j, (prob<F_rat & ~neg)) = ...
      F2_accept(j, (prob<F_rat & ~neg)) + 1;
end

% update F3

for j = 1:k
    F_new = F3;
    F_new(j,:) = normrnd(F3(j,:),F3_cand(j,:));
    neg = (F_new(j,:) <= 0);
    F_new(j,neg) = F3(j,neg);
    F_rat = F_cc(F_new,F_prior_mean,F_prior_var,L,Y3,Y3_unc,j) - ...
            F_cc(F3,F_prior_mean,F_prior_var,L,Y3,Y3_unc,j);
    prob = log(rand(1,t));
    F3(j,(prob < F_rat)) = F_new(j,(prob < F_rat));
    F3_accept(j, (prob<F_rat & ~neg)) = ...
      F3_accept(j, (prob<F_rat & ~neg)) + 1;
end


% write to file
```

```
if i > burnin && mod(i,thin)==0
    dlmwrite('output3/Lambda.txt',L,'-append','roffset',1);

    for j = 1:k
         dlmwrite(strcat('output3/f1.',int2str(j),'.txt'),...
           F1(j,:),'-append','roffset',1);
         dlmwrite(strcat('output3/f2.',int2str(j),'.txt'),...
           F2(j,:),'-append','roffset',1);
         dlmwrite(strcat('output3/f3.',int2str(j),'.txt'),...
           F3(j,:),'-append','roffset',1);
    end
end

% update tuning parameters

if mod(i,adjust) == 0
     i
     if i<=burnin
         for j = 1:k
             for h = 1:p
                 if L_accept(h,j)/adjust < 0.2
                     L_cand(h,j) = L_cand(h,j).*0.9;
                 elseif L_accept(h,j)/adjust > 0.45
                     L_cand(h,j) = L_cand(h,j).*1.1;
                 end
             end
```

```
        for h = 1:t
            if F1_accept(j,h)/adjust < 0.2
                F1_cand(j,h) = F1_cand(j,h).*0.9;
            elseif F1_accept(j,h)/adjust > 0.45
                F1_cand(j,h) = F1_cand(j,h).*1.1;
            end
        end
        for h = 1:t
            if F2_accept(j,h)/adjust < 0.2
                F2_cand(j,h) = F2_cand(j,h).*0.9;
            elseif F2_accept(j,h)/adjust > 0.45
                F2_cand(j,h) = F2_cand(j,h).*1.1;
            end
        end
        for h = 1:t
            if F3_accept(j,h)/adjust < 0.2
                F3_cand(j,h) = F3_cand(j,h).*0.9;
            elseif F3_accept(j,h)/adjust > 0.45
                F3_cand(j,h) = F3_cand(j,h).*1.1;
            end
        end
    end
    L_accept = zeros(p,k);
    F1_accept = zeros(k,t);
    F2_accept = zeros(k,t);
    F3_accept = zeros(k,t);
end
```

```
    end
end
```

6.6 L_cc3.m

```
function  cc  = L_cc3( L, L_prior_mean, L_prec, F1, F2, F3, Y1, Y2, Y3,...
  Y1_unc, Y2_unc, Y3_unc, j )
%lambda_cc Calculates complete conditional for Lambda

prior = log_dirichlet_pdf(L(:,j),L_prior_mean(:,j)*L_prec(j));

m1 = L * F1;
logfrac1 = log((m1 .^ 2 + Y1_unc .^ 2)./(m1 .^ 2));
likelihood1 = sum(sum(-log(Y1) - (log(Y1) - log(m1) + ...
  0.5*logfrac1).^2 ./ (2 * logfrac1)));

m2 = L * F2;
logfrac2 = log((m2 .^ 2 + Y2_unc .^ 2)./(m2 .^ 2));
likelihood2 = sum(sum(-log(Y2) - (log(Y2) - log(m2) + ...
  0.5*logfrac2).^2 ./ (2 * logfrac2)));

m3 = L * F3;
logfrac3 = log((m3 .^ 2 + Y3_unc .^ 2)./(m3 .^ 2));
likelihood3 = sum(sum(-log(Y3) - (log(Y3) - log(m3) + ...
  0.5*logfrac3).^2 ./ (2 * logfrac3)));

cc = prior + likelihood1 + likelihood2 + likelihood3;
```

```
end
```

BIBLIOGRAPHY

Gatz, D. F. (1975), "Relative Contributions of Different Sources of Urban Aerosols: applications of a new estimation method to multiple sites in Chicago," *Atmospheric Environment*, 9, 1–18.

Heaton, M. J., Reese, C. S., and Christensen, W. F. (2010), "Incorporating Time-Dependent Source Profiles Using the Dirichlet Distribution in Multivariate Receptor Models," *Technometrics*, 52, 67–79.

Johnson, V. E., and Rossell, D. (2010), "On the Use of Non-Local Prior Densities in Bayesian Hypothesis Tests," *Journal of the Royal Statistical Society, Series B*, 72, 143–170.

Koutrakis, P., and Spengler, J. D. (1987), "Source Apportionment of Ambient Particles in Steubenville, OH Using Specific Rotation Factor Analysis," *Atmospheric Environment*, 21, 1511–1519.

Lingwall, J. W., Christensen, W. F., and Reese, C. S. (2008), "Dirichlet Based Bayesian Multivariate Receptor Modeling," *Environmetrics*, 19, 618–629.

Miller, M. S., Friedlander, S. K., and Hidy, G. M. (1972), "A Chemical Element Balance for the Pasadena Aerosol," *Journal of Colloid and Interface Science*, 39, 165–176.

Paatero, P. (1998), *Users Guide for Positive Matrix Factorization Programs PMF2 and PMF3*, University of Helsinki, Helsinki, Finland.

Paatero, P., and Tapper, U. (1994), "Positive Matrix Factorization: a non-negative factor model with optimal utilization of error estimate of data values," *Environmentrics*, 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., Oh, M.-S., and Guttorp, P. (2002), "Multivariate receptor models and model uncertainty," *Chemometrics and Intelligent Laboratory Systems*, 60, 49 – 67.

Schauer, J. (2011), "Data and synoptic regime info," Private communication.

Thurston, G. D., and Spengler, J. D. (1985), "A Quantitative Assessment of Source Contributions to Inhalable Particulate Matter Pollution in Metropolitan Boston," *Atmospheric Environment*, 19, 9–25.

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