ENVIRONMETRICS
Environmetrics 2008; **19**: 618–629
Published online 18 December 2007 in Wiley InterScience
(www.interscience.wiley.com) DOI: 10.1002/env.902

Dirichlet based Bayesian multivariate receptor modeling

Jeff W. Lingwall[1], William F. Christensen[2*,†] and C. Shane Reese[2]

[1]*Department of Statistics, Carnegie Mellon University, Pittsburgh, PA 15213, U.S.A.*
[2]*Department of Statistics, Brigham Young University, Provo, UT 84602, U.S.A.*

SUMMARY

We propose a simple, fully Bayesian approach for multivariate receptor modeling that allows for flexible and consistent incorporation of *a priori* information. The model uses a generalization of the Dirichlet distribution as the prior distribution on source profiles that allows great flexibility in the specification of prior information. Heavy-tailed lognormal distributions are used as priors on source contributions to match the nature of particulate concentrations. A simulation study based on the Washington, DC airshed shows that the model compares favorably to Positive Matrix Factorization, a standard analysis approach used for pollution source apportionment. A significant advantage of the proposed approach compared to most popularly used methods is that the Bayesian framework yields complete distributional results for each parameter of interest (including distributions for each element of the source profile and source contribution matrices). These distributions offer a great deal of power and versatility when addressing complex questions of interest to the researcher. Copyright © 2007 John Wiley & Sons, Ltd.

KEY WORDS: pollution source apportionment; chemical mass balance; air pollution; source attribution; Bayesian methods

1. INTRODUCTION

Receptor modeling approaches attempt to apportion ambient air pollutant measurements (such as particulate matter species measurements) into identifiable sources. In multivariate receptor modeling, we derive from the data both the source profiles (the particulate 'fingerprint' of each source) and their contributions (the impact of each source at the receptor). Because studies are showing links between ambient particulates and public health (see Dockery *et al.*, 1993; Dominici *et al.*, 2000), a reliable source apportionment approach is necessary to give decision makers valuable information about the impact of various particulate sources in the airshed.

The basic multivariate receptor model for K sources can be written as

$$\underset{P\times T}{\mathbf{Y}} = \underset{P\times K}{\boldsymbol{\Lambda}}\;\underset{K\times T}{\mathbf{F}} + \underset{P\times T}{\boldsymbol{\epsilon}} \tag{1}$$

where $\mathbf{Y}$ is the observed particulate concentrations of P chemical species over T time periods, $\boldsymbol{\Lambda}$ is the matrix of profiles for the K sources, and $\mathbf{F}$ is the matrix of source contributions. The source contributions will be expressed in the same metric as the observed $\mathbf{Y}$ (e.g. $\mu g/m^3$). For example, Christensen *et al.*

[*]Correspondence to: W. F. Christensen, Department of Statistics, Brigham Young University, Provo, UT 84602, U.S.A.
[†]E-mail: william@stat.byu.edu

Received 27 February 2007
Accepted 6 November 2007

(2006) identified eight sources effecting the Washington DC airshed including three secondary sources, a vehicular source, soil dust, wood burning, sea salt, and a lead source.

This model differs from a traditional factor analysis model in that each element of $\mathbf{Y}$, $\boldsymbol{\Lambda}$, and $\mathbf{F}$ is constrained to be non-negative to match the physical reality of the problem. In the approach described herein, we incorporate this and other *a priori* information about source profiles and contributions into the analysis.

There are many approaches to solving the non-negative factor analytic equation in the literature (see Christensen *et al.*, 2006 for an overview). Perhaps the most common approach is Positive Matrix Factorization (PMF), an algorithm developed by Paatero and Tapper (1994) that quickly estimates $\boldsymbol{\Lambda}$ and $\mathbf{F}$ by iterative methods. *A priori* information about source profiles may be introduced into PMF through a process we will refer to as 'source profile targeting' (also known as 'Gkeying') in which target profiles are given along with their uncertainties. Another process pulls specified elements of the profile matrix toward zero (also known as 'Fkeying').

Though there is much literature on Bayesian source separation (Knuth, 1998, 1999; Knuth and Vaughan, 1998; Rowe, 2003; Miskin and MacKay, 2001) and Bayesian latent variable analysis (Evans *et al.*, 1989; Dunson, 2000), there is little that deals specifically with non-negative matrix factorization. Ichir and Mohammad-Djafari (2004) consider a Bayesian approach to source separation that constrains the sources (contributions) to be positive. Again from the perspective of source separation, Moussaoui *et al.* (2004) use a Bayesian approach to constrain sources and mixing elements to be non-negative, taking Gamma prior distributions for both sources and mixing elements. They use maximum *a posteriori* estimation to fit the model in a generalization of the PMF algorithm.

Bayesian methods have also begun to be applied directly to PSA. Park *et al.* (2001), published an analysis of data from Atlanta, utilizing MCMC and accounting for the temporal dependence in the data. Using the judgment of an environmental engineer and source measurements, they established zeros in the source profiles and used MCMC to fit a hierarchical model. They assume normally distributed errors in the bulk of their analysis and leave the study of heavy-tailed errors open.

In this manuscript we propose a simple and flexible Bayesian approach for pollution source apportionment that can be implemented in a wide variety of situations. Using a generalization of the Dirichlet distribution as the prior on source profiles allows a natural interpretation of the profile as a proportional relationship among species emitted. At the same time, our approach allows great flexibility in specifying the error structure within the profile. This Generalized Dirichlet is much more scientifically satisfying than using truncated normal or gamma distributions for prior elements, because it inherently captures the multivariate nature of source profiles. For the source contributions and in the likelihood for the observed data, the use of heavy-tailed lognormal distributions is appealing because it matches the often skewed nature of particulate observations.

The methods proposed offer a framework capable of incorporating a wide range of *a priori* information in an internally consistent manner. This is a substantial advantage over currently used techniques that either adapt the natural estimates to conform to suspected values or require the researcher to continually adjust and re-run analysis software until results conform to pre-conceived notions. In this paper we offer a simple example that illustrates the use of this broader framework which is potentially much more flexible than existing techniques.

Section 2 introduces the models used in the research, including discussion of the Generalized Dirichlet distribution. Section 3 presents the model for the air quality data and discusses computational issues for the Markov Chain Monte Carlo sampling. As a benchmark for evaluating the model, Section 4 shows the Bayesian model's performance compared to PMF on a simulated airshed using both informative and flat prior distributions. Although the Bayesian approach yields full distributions for unknown parameters

DOI: 10.1002/env

(such as the elements of the profile and contribution matrices), we use the medians of the Bayesian posterior distributions in order to compare with PMF's point estimates for $\mathbf{F}$ and $\mathbf{\Lambda}$. The availability of full posterior distributions for the model parameters for analysis is a significant advantage over methods that solely return point estimates. From the posterior distribution we may obtain *distributional* results about any quantities of interest in the analysis.

2. THE GENERALIZED DIRICHLET DISTRIBUTION

Beginning with the source profiles, we have

$$\underset{P\times K}{\mathbf{\Lambda}} = [\boldsymbol{\lambda}_1, \ldots, \boldsymbol{\lambda}_K]$$

where P is the number of chemical species measured in $\mathbf{Y}$ and K is the number of pollution sources. (We assume K to be known for the purposes of this paper. Park *et al.* (2002) consider model selection within a Bayesian framework and this is an important topic of current research.) The Dirichlet distribution naturally incorporates our assumptions about source profile elements into the model, keeping $0 < \lambda_{pk} < 1$ and $\sum_{p=1}^{P} \lambda_{pk} = 1$. This leads to the simple interpretation of each profile as the fractionization of the chemical species emitted by the source.

The challenge with using the Dirichlet distribution for source profiles is the strict relationship between its expectation and variance, where $Var(\lambda_{pk}) = \frac{a_{pk}(c_k - a_{pk})}{c_k^2(c_k+1)}$, a_{pk} is the parameter of the Dirichlet corresponding to λ_{pk}, and $\sum_{p=1}^{P} a_{pk} = c_k$ is the sum of the parameters of the Dirichlet for $\boldsymbol{\lambda}_k$. This relationship might not be desirable for the pollution modeler, where differing amounts of information might exist about the representation of species in a profile. That is, the uncertainty associated with each element of the profile is often denoted specifically. Yet the standard Dirichlet requires the uncertainty for a profile to be denoted for the profile as a whole.

As an alternative, we consider a generalization of the Dirichlet distribution (see Rogers and Young, 1973) where the relationship between the element-wise variance and expectation is more relaxed. The elements of the standard Dirichlet distribution can be constructed from independent Gamma random variables with common scale parameter 1, divided by their sum. In contrast, the elements of our Generalized Dirichlet can be constructed from independent Gammas with *differing* scale parameters divided by their sum. This allows us to generalize the relationship between the expectation and the variance of the distribution to introduce more flexibility into the model, while still keeping the desirable properties of the Dirichlet ($0 < \lambda_{pk} < 1$ and $\sum_{p=1}^{P} \lambda_{pk} = 1$). Note that after scaling the independent Gamma variables to sum to 1, the uncertainty of each element of the resulting Generalized Dirichlet vector will not (in general) be equal to the desired uncertainty. However, in practice the parameters of the individual independent Gammas can be tuned so that the uncertainty associated with each element of the Generalized Dirichlet can be matched to an encouragingly high degree.

For example, consider a smelter where it is known with high precision that 20% of the $PM_{2.5}$ emissions are lead, with other chemical species constituting the remaining 80%. If we have less information about the non-lead emissions, we would seek a model that lets us specify relatively loose priors around the non-lead species and a tight prior around lead. The standard Dirichlet cannot directly incorporate such information, because the element-wise variance is tied directly to the expectation of the distribution, but the Generalized Dirichlet easily allows for such flexibility.

DOI: 10.1002/env

3. BAYESIAN MODELING

We assume each column $\boldsymbol{\lambda}_k$ of $\boldsymbol{\Lambda}$ is distributed as *Generalized Dirichlet*($\boldsymbol{\kappa}_k, \boldsymbol{\beta}_k$), so that

$$p(\boldsymbol{\lambda}_k) \propto \left(\sum_{p=1}^{P} \lambda_{pk}/\beta_{pk}\right)^{-\sum_{p=1}^{P}\kappa_{pk}} \prod_{p=1}^{P} \lambda_{pk}^{(\kappa_{pk}-1)} \tag{2}$$

and

$$p(\boldsymbol{\Lambda}) \propto \prod_{k=1}^{K} \left(\sum_{p=1}^{P} \lambda_{pk}/\beta_{pk}\right)^{-\sum_{p=1}^{P}\kappa_{pk}} \prod_{p=1}^{P} \lambda_{pk}^{(\kappa_{pk}-1)} \tag{3}$$

If we set $\beta_{pk} = 1$ for all p, k we obtain the usual Dirichlet distribution.

A common practice in pollution source apportionment is to specify *a priori* zeros in source profiles to establish identifiability of the model (Park *et al.*, 2001; Lee *et al.*, 2006). Our flexible framework allows incorporation of this information in two ways. First, if M zeros are assumed to be known in a profile, the dimensionality of the Generalized Dirichlet over the profile is simply reduced and becomes $P - M$. The zeros are then left as constants during the MCMC fitting process. Second, if zeros are hypothesized but not known, prior distributions may be centered near zero.

If no prior information is known (i.e., a purely exploratory analysis is to be performed), we return to the usual Dirichlet distribution, setting each of the parameters of the profile to $1/P$. This gives a 'flat' prior distribution in the sense that the marginal prior distribution of each profile element is identical to the others, while the distribution gives more prior weight to small profile elements. Because profiles usually consist of a small number of dominant species with the rest closer to zero, we find this formulation of a flat prior appealing.

Often, there exists at least partial *a priori* information about a source's composition in the form of a hypothesized profile $\tilde{\boldsymbol{\lambda}}$ accompanied by individual uncertainties $\tilde{\boldsymbol{\delta}}$ for each element of the profile. For such cases, we use a genetic algorithm that tunes the parameters of the Generalized Dirichlet $(\beta_1, \ldots, \beta_P, \kappa_1, \ldots, \kappa_P)$ to best match the information in $\tilde{\boldsymbol{\lambda}}$ and $\tilde{\boldsymbol{\delta}}$. Values are found which are close to minimizing the Euclidean distance between the marginal standard deviations of elements of the Generalized Dirichlet with the elementwise target uncertainties. To compare the *a priori* profile information with the tuned Generalized Dirichlet, Figure 1 plots (in dashed lines) the elementwise distributions implied by hypothesized auto-diesel profile information (assuming a Gamma distribution for $\lambda_p, p = 1, \ldots, P$). Also plotted (in solid lines) are the marginal distributions of the tuned Generalized Dirichlet. Although the Generalized Dirichlet does not perfectly reproduce the assumed uncertainties associated with a hypothesized profile, it closely mimics the distribution implied by the *a priori* information while incorporating the scientifically meaningful structure of the Dirichlet distribution. That is, the Generalized Dirichlet preserves the property that the profile elements sum to one, whereas the set of P independent random variables has no such property.

For the prior distributions on the source contributions, $\underset{K\times T}{\mathbf{F}}$, we take $f_{kt} \sim LN(\gamma_{kt}, \delta_{kt})$, where LN is the lognormal distribution with expectation γ_{kt} and standard deviation δ_{kt}. The use of the lognormal distribution is scientifically satisfying, since we expect long-tailed concentrations of particulate measurements and require non-negativity. This formulation of the lognormal distribution gives rise

DOI: 10.1002/env

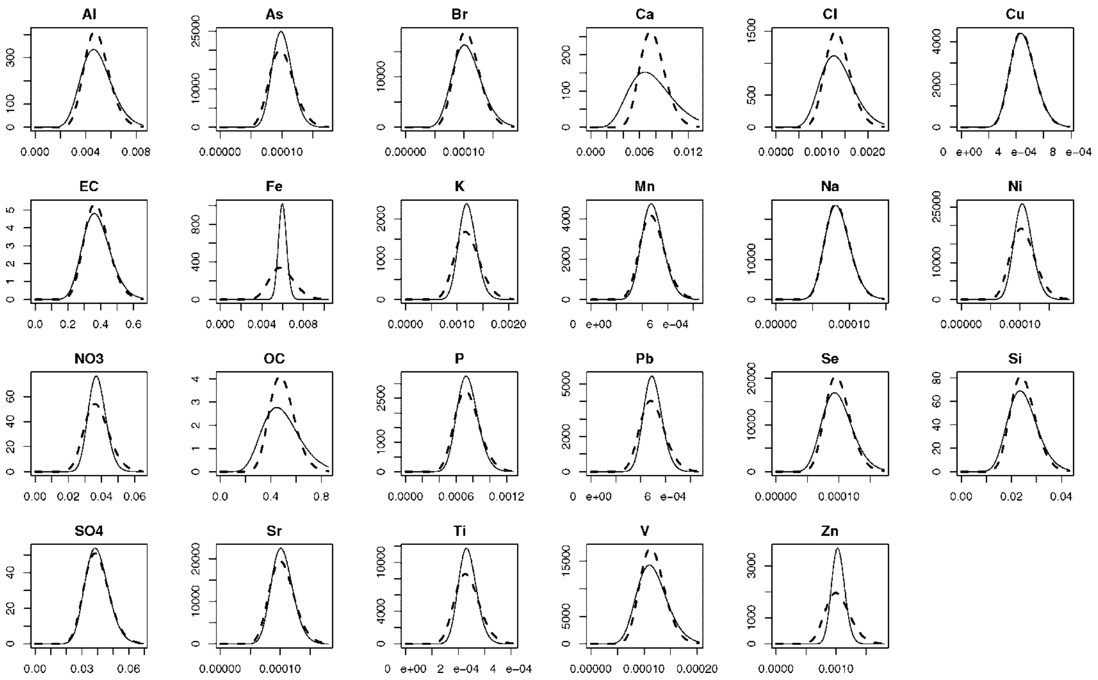

Figure 1. Example of output from the genetic algorithm that tunes parameters of the Generalized Dirichlet to match the *a priori* information associated with an auto-diesel profile. The dashed lines represent distributions implied by the *a priori* information and the solid lines illustrate the marginal distributions of the tuned Generalized Dirichlet

to the likelihood for **F** as

$$p(\mathbf{F}) \propto \prod_{t=1}^{T}\prod_{k=1}^{K} \frac{1}{f_{kt}} \exp\left[-\frac{\left(\log(f_{kt}) - \left[\log(\gamma_{kt}) - 0.5\log\left(\frac{\gamma_{kt}^2+\delta_{kt}^2}{\gamma_{kt}^2}\right)\right]\right)^2}{2\log\left((\gamma_{kt}^2+\delta_{kt}^2)/\gamma_{kt}^2\right)}\right] \tag{4}$$

Finally, we assume that $y_{pt} \sim LN(\boldsymbol{\lambda}_p\mathbf{f}_t, \nu_{pt})$, where $\boldsymbol{\lambda}_p$ is the pth row of $\boldsymbol{\Lambda}$, $\mathbf{F}_t$ is the tth column of **F**, and ν_{pt} is the observational uncertainty information about y_{pt} from the receptor. The expectation of $\mathbf{y}_t$ (the multivariate mass measured at time t) becomes $\boldsymbol{\Lambda}\mathbf{f}_t$, a linear combination of the contributions of the K sources at time t. The lognormal distribution is satisfying scientifically because the error term (incorporating measurement and equation error) tends to be strongly skewed for environmental measurements such as these.

Draws from the normalized posterior are obtained by using Markov Chain Monte Carlo simulation, updating parameters via Metropolis steps. To update the source profiles, we update one element of each profile at a time by using a normal proposal, rescaling the profile to one, and then accepting or rejecting by a Metropolis step. This is performed for each element of each source profile, so in essence P updates are proposed for each profile during each iteration of the algorithm. The standard deviations of the normal candidate distributions for both **F** and $\boldsymbol{\Lambda}$ are automatically tuned during all or part of the burn-in phase to achieve good sampling, based on the number of draws being accepted (see Gelman *et al.*, 2004,

DOI: 10.1002/env

pp. 305–306). For example, for the majority of our sampling in the next section, after each 20 MCMC iterations the candidate sigmas for **F** were decreased by a tenth if acceptance ratios fell beneath 0.25 and increased by a tenth if acceptance ratios rose above 0.55.

Once we obtain draws from the posterior distribution, we obtain a substantial advantage over standard pollution source apportionment methods with regard to inference. We can obtain well-defined distributions for complex quantities related to air quality standards, such as: (i) the number of times a source contribution threshold is exceeded, (ii) the daily probability that a source contribution exceeds a threshold, (iii) the daily probability that the combined contributions of two related sources exceeded a threshold, or (iv) the probability that a source's contribution exceeded a specified daily threshold for seven consecutive days at some point during a study period. Note that while such issues would be difficult or impossible to address using traditional source apportionment methods, the desired analyses are easily obtained after creating posterior distributions for each of the parameters of interest. Additionally, the use of the Bayesian prior distribution allows for internally consistent incorporation of *a priori* information. This is preferable to the lengthy series of methodological adjustments and tuning that are often used *a posteriori* to ensure that source apportionment results coincide with *a priori* knowledge about existing pollution sources. We also prefer this use of *a priori* information to the introduction of target transformations, which 'squeeze' the data to best match the profiles (see Hopke, 1988).

4. SIMULATION STUDY AND RESULTS

To evaluate our approach, we perform a simulation study and compare the results of the Bayesian source apportionment to PMF (Paatero and Tapper, 1994), a commonly used approach for pollution source apportionment. We simulated artificial airsheds based on an analysis of $PM_{2.5}$ data from Washington DC (Christensen *et al.*, 2006). Our approach allows us to build natural temporal correlation into our simulated data while still allowing us to know the true sources. We use five of the sources derived from the Washington DC analysis, including sea salt, secondary sulfate, secondary nitrate, soil dust, and auto/diesel (vehicular). An error-free airshed matrix was first created by multiplying the source profile matrix with the source contribution matrix. Each element of the simulated airsheds was then created by taking a random draw from a lognormal distribution centered at the corresponding element of the error-free airshed matrix and with a coefficient of variation of $CV_{\mathbf{Y}} = 0.3$ or 0.6. Our resulting data sets ($\mathbf{Y}$) have $T = 100$ measurements on $P = 23$ different chemical species.

For purposes of the simulation, our *a priori* (assumed) source profile matrix ($\tilde{\mathbf{\Lambda}}$) used to inform the prior distributions on $\mathbf{\Lambda}$ was obtained from the 'true' source profile matrix by taking a random draw from a lognormal distribution centered at the true value for each profile element and with a coefficient of variation of $CV_{\mathbf{\Lambda}} = 0.2$, 0.4, or 0.6. The profiles were then scaled to sum to one. Our estimated uncertainties, both for $\mathbf{Y}$ and $\tilde{\mathbf{\Lambda}}$, were taken as the simulated values times their respective CV. For the Bayesian model, the genetic algorithm discussed in the previous section was used to find Generalized Dirichlet parameters that would match the elementwise information.

For the prior distributions on elements of **F** in this simulation study, we used the right-skewed lognormal distribution shown in Figure 2. Rather than attempting the use of element-specific informative prior distributions for **F**, we choose a prior distribution that reflects the general shape of the elements of F. By using the same generic prior distribution for every element of **F**, we are better able to compare with PMF, where there is no studied method for introducing target shapes for source contributions. The use of informative priors on **F** by incorporating seasonal and atmospheric data is a focus of current research.

DOI: 10.1002/env

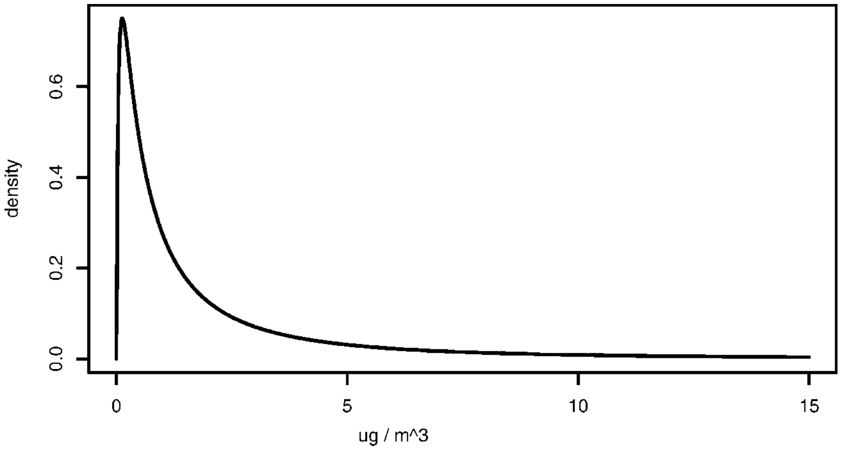

Figure 2. Lognormal distribution used on each element of **F** for the simulation study

Two different scenarios for source apportionment are considered in our simulations. In the first scenario, we consider the use of *a priori* information $\tilde{\mathbf{\Lambda}}$. For the Bayesian approach, this information is incorporated via the Generalized Dirichlet prior. For the PMF approach, the *a priori* information in $\tilde{\mathbf{\Lambda}}$ and the associated uncertainty matrix ($\mathrm{CV}_{\mathbf{\Lambda}} \times \mathbf{\Lambda}$) are incorporated using the source-profile-targeting approach outlined by Paatero (2003). As recommended by Paatero (2003), profile targeting is only used after an initial run of PMF is executed to obtain starting values. In addition to the *a priori* information on λ, we consider a second source apportionment scenario which uses no *a priori* information about $\mathbf{\Lambda}$. That is, we consider the performance of the Bayesian and PMF approaches in the exploratory setting.

Using the prior specifications for $\mathbf{\Lambda}$, the Bayesian model was fit to each simulated data set. The processing time took about 45 min per data set on a Pentium 4 processor for 5000 draws from the posterior (including a burn-in of 3000 draws). Each draw consisted of values for each of the 615 parameters in the model (500 elements of **F** and 115 elements of $\mathbf{\Lambda}$). For the exploratory setting (no *a priori* information about $\mathbf{\Lambda}$), 50 000 draws from the posterior (including a burn-in of 30 000 draws) were used to obtain Bayesian posterior distributions, taking about 8 h per data set. The estimation required a large number of draws because without *a priori* information about profiles, the MCMC required more draws to converge to a stable distribution. A total of 30 simulated datasets (each with 100 days of estimates) were considered for each combination of $\mathrm{CV}_{\mathbf{Y}}$ and $\mathrm{CV}_{\mathbf{\Lambda}}$.

To quantify our results and compare the Bayesian and PMF approaches, we calculate the Total Median Absolute Error (TMAE) of the estimates, obtained in the following manner. For the Bayesian approach, let each element of $\hat{\mathbf{\Lambda}}$ be the posterior median of the marginal density for each parameter in $\mathbf{\Lambda}$. When using PMF, $\hat{\mathbf{\Lambda}}$ is simply the estimated profile matrix. Then, let the TMAE for $\hat{\mathbf{\Lambda}}$ be the sum of the median absolute error for each of the K sources in the model, where the median absolute error for the kth source is calculated as the median of $|\hat{\lambda}_{pk} - \lambda_{pk}|$, $p = 1, \ldots, P$. The TMAE for $\hat{\mathbf{F}}$ is calculated in a similar manner, again using the posterior medians for $\hat{\mathbf{F}}$ when the Bayesian approach is employed.

Table 1 gives TMAE values for the Bayesian and PMF approaches under the scenario where *a priori* knowledge about $\mathbf{\Lambda}$ is available and the exploratory analysis scenario. We can see that in the scenario where $\tilde{\mathbf{\Lambda}}$ is available, using the Bayesian approach instead of PMF reduces the TMAE for the source contribution estimates by an average of 4% in the low profile error case, by 15% in the medium profile error case, and by 16% in the high profile error case. For the profile estimates, the Bayesian

DOI: 10.1002/env

Table 1. Total Median Absolute Error for estimating source contributions and source profiles when using four different estimation approaches: PMF (with and without *a priori* information on $\mathbf{\Lambda}$) and the Bayesian approach (with and without *a priori* information on $\mathbf{\Lambda}$). The four approaches are labeled 'PMF$_{\tilde{\Lambda}}$,' 'PMF,' 'Bayesian$_{\tilde{\Lambda}}$,' and 'Bayesian'

Parameters	$\mathrm{CV}_{\mathbf{Y}}$	$\mathrm{CV}_{\mathbf{\Lambda}}$	PMF$_{\tilde{\Lambda}}$	Bayesian$_{\tilde{\Lambda}}$	PMF	Bayesian
F	0.3	0.2	4.22	4.02	6.84	4.77
$\mathbf{\Lambda}$	0.3	0.2	0.0048	0.0016	0.0136	0.0031
F	0.3	0.4	5.17	4.36	6.84	4.77
$\mathbf{\Lambda}$	0.3	0.4	0.0065	0.0019	0.0136	0.0031
F	0.3	0.6	5.01	4.42	6.84	4.77
$\mathbf{\Lambda}$	0.3	0.6	0.0069	0.0023	0.0136	0.0031
F	0.6	0.2	7.24	7.05	10.21	7.87
$\mathbf{\Lambda}$	0.6	0.2	0.0019	0.0022	0.0283	0.0056
F	0.6	0.4	9.13	7.67	10.21	7.87
$\mathbf{\Lambda}$	0.6	0.4	0.0265	0.0035	0.0283	0.0056
F	0.6	0.6	9.80	8.05	10.21	7.87
$\mathbf{\Lambda}$	0.6	0.6	0.0296	0.0051	0.0283	0.0056

approach reduces the TMAE in the low, medium, and high profile error cases by an average of 25%, 79%, and 75%, respectively. When we consider the exploratory analysis scenario, the superiority of the Bayesian approach is slightly more pronounced. On average, the Bayesian TMAE is 27% smaller for the source contribution estimates and 79% smaller for the profile estimates. While this simulation does not attempt to exhaustively compare PMF's performance with our Bayesian approach, it seems to indicate that the Bayesian approach compares favorably to PMF when using data affected by a wide range of measurement error variance values.

To illustrate the power of Bayesian inference, we consider estimates obtained from a data set generated with $\mathrm{CV}_{\mathbf{Y}} = 0.3$ and $\mathrm{CV}_{\mathbf{\Lambda}} = 0.2$. Figure 3 illustrates the posterior distribution for the secondary sulfate source by denoting the logged marginal density estimates for the contributions (with lowest density areas trimmed to a constant). The posterior medians are denoted with a black line. Instead of returning only point estimates as in most of the widely-used pollution source apportionment approaches, we are given complete probability distributions for each quantity of interest. While other methods might allow some insight into the distributional quality of the results (see Eberly, 2005), the distributions of interest follow naturally when using a Bayesian framework for the problem.

Using the posterior distribution for each estimated parameter in the model, we can obtain well-defined distributions for complex quantities related to air quality standards. For example, because secondary sources comprise a large fraction of the total $\mathrm{PM}_{2.5}$ in many areas, there may be interest in evaluating the daily probability of high total secondary formation (e.g., total secondary source contributions exceed $20\,\mu\mathrm{g/m}^3$ on a given day). More technically, we are interested in the 615-dimensional integral

$$\int \cdots \int I_{\{f_{\mathrm{SS},t}+f_{\mathrm{SN},t}>20\}}\, \pi(\mathbf{\Theta}|\mathbf{Y})\, \mathrm{d}\mathbf{\Theta}, \quad t = 1, \ldots, T \tag{5}$$

where $f_{\mathrm{SS},t}$ and $f_{\mathrm{SN},t}$ are the contributions of the secondary sulfate and secondary nitrate sources on day t, $\mathbf{\Theta} = (\lambda_{1,1}, \ldots, \lambda_{23,5}, f_{1,1}, \ldots, f_{5,100})$ is the collection of all 615 parameters in the model, $\pi(\mathbf{\Theta}|\mathbf{Y})$ is the joint posterior distribution for $\mathbf{\Theta}$ given $\mathbf{Y}$, and $I_{\{A\}}$ is the indicator function taking the value of 1 when the condition A is true and 0 otherwise. Although standard approaches such as PMF can yield a sum of estimates, the probability that the sum of actual contributions exceeds a threshold is complicated

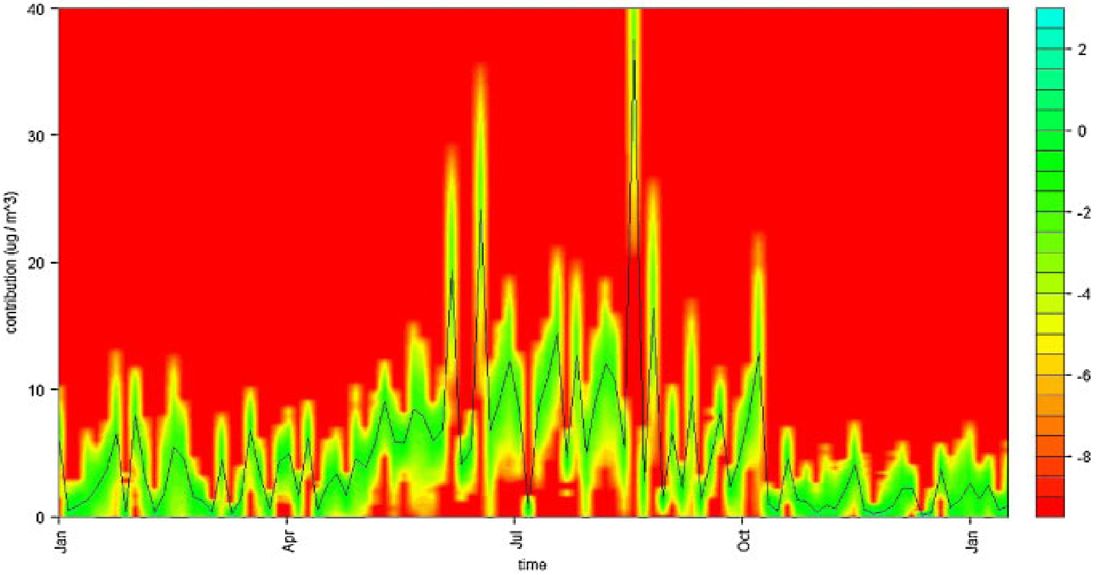

Figure 3. Smoothed (logged) densities for the secondary sulfate source contributions. Posterior medians are shown in black

because threshold exceedance probabilities require substantial information about the joint distribution of the estimates. Further, a sum of source contributions is particularly difficult to characterize because of the negative correlation between the estimation errors. However, the Bayesian approach recommended here easily yields the quantity of interest from a simple analysis of the post-burn-in draws from the Markov chain, and the solution properly accounts for the uncertainty associated with each of the 615 model parameters. Figure 4 gives the daily probability that the summed secondary formation sources

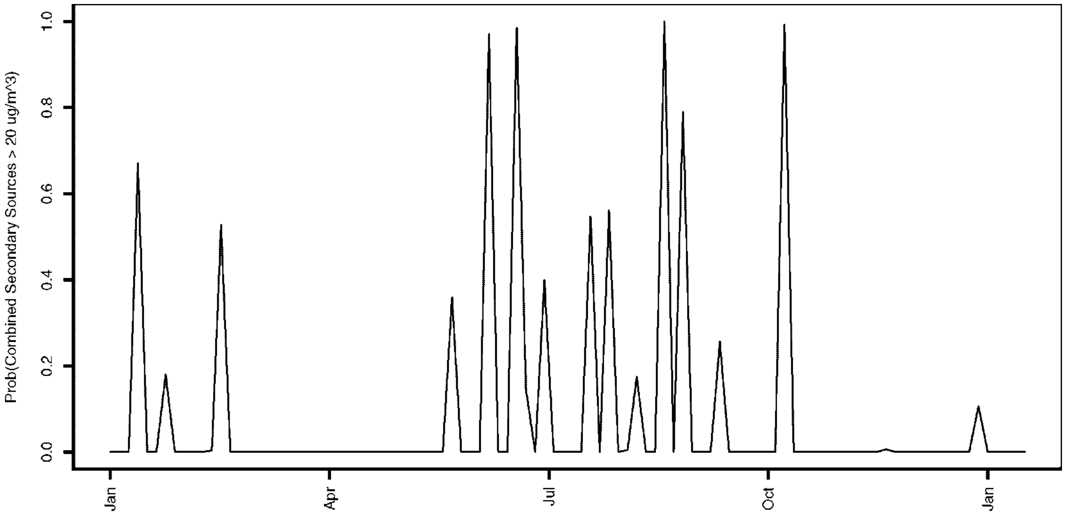

Figure 4. Daily probability that the combined contributions of the secondary sulfate and secondary nitrate sources exceeds 20 μg/m^3

DOI: 10.1002/env

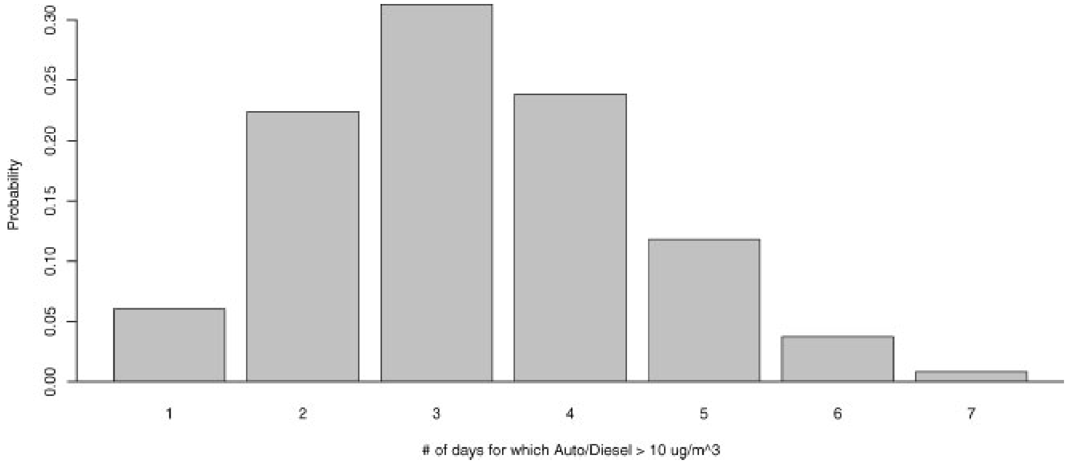

Figure 5. Probability distribution for the number of days (out of the total of 100) in which the auto/diesel source exceeds $10\,\mu\text{g/m}^3$

exceeds $20\,\mu\text{g/m}^3$. Sulfate formation is heaviest during the summer and nitrate formation is heaviest during the winter. The combined contributions from these secondary sources is most likely to exceed the specified threshold during August, July, June, and January. One might also be interested in doing inference on a combination of source estimates that may be correlated and poorly separated. For example, differentiating gasoline vehicle emissions and diesel emissions can be very difficult, so inference about their sum may be an attractive alternative.

As another example of the flexibility of the Bayesian approach, one might be interested in the number of exceedance days for a specific source. Reducing the exceedance days for auto/diesel emissions may be a sub-goal related to a larger aim of reducing the number of $\text{PM}_{2.5}$ threshold exceedance days. Let κ be the number of days (out of the total of 100) in which the auto/diesel source exceeds $10\,\mu\text{g/m}^3$. Figure 5 gives the probability distribution for κ given $\mathbf{Y}$. Using the same notation as in Equation (5), this distribution can be more technically defined as

$$g(\kappa|\mathbf{Y}) = \int \cdots \int I_{\{(\sum_{t=1}^{T} I_{\{f_{\text{AD},t}>10\}})=\kappa\}}\, \pi(\boldsymbol{\Theta}|\mathbf{Y})\,\text{d}\boldsymbol{\Theta}, \quad \kappa = 0, 1, \ldots, T \tag{6}$$

where $f_{\text{AD},t}$ is the contribution of the auto/diesel source on day t. If we consider the posterior median as a point estimate for the auto/diesel source contribution, only 3 of the 100 study days have point estimates in excess of $10\,\mu\text{g/m}^3$. But Figure 5 gives a more complete understanding of this variable. For example, the expected number of auto/diesel exceedance days is roughly 3.3 and the probability that the number of auto/diesel exceedance days surpasses 4 days is roughly 16%.

5. CONCLUSIONS

The proposed Bayesian approach provides a useful alternative to other methods used in multivariate receptor modeling. The fully Bayesian approach is attractive because it easily incorporates a wide range of *a priori* information into analysis and gives full distributional results rather than just point

DOI: 10.1002/env

estimates for source profiles and contributions. The novel use of heavy-tailed lognormal distributions for the source contributions and for the distribution of the particulate measurements is scientifically satisfying. The use of a Generalized Dirichlet distribution for source profiles allows for great flexibility in multivariate specification of the prior information about emission sources while constraining the solution to be physically meaningful.

The Bayesian approach allows us to consistently incorporate the *a priori* information into an analysis rather than adjusting results after a model has been fit or introducing target transformations *a posteriori*. In simulation, the approach has been found to compete favorably with PMF. The full distributional results obtained from the Bayesian approach gives the researcher a great deal of flexibility in addressing questions associated with potentially complex functions of estimated parameters. This Bayesian framework for receptor modeling should be of great interest to researchers seeking a cohesive approach for combining their airshed knowledge with their airshed measurements and wanting to obtain full distributional results to facilitate their specific scientific inquiries.

ACKNOWLEDGEMENTS

This work was supported by the STAR Research Assistance Agreement No. RD-83216001-0 awarded by the U.S. Environmental Protection Agency. The paper has not been formally reviewed by the EPA. The views expressed in this document are solely those of the authors and the EPA does not endorse any products or commercial services mentioned in this publication.

REFERENCES

Christensen WF, Schauer JJ, Lingwall JW. 2006. Iterated confirmatory factor analysis for pollution source apportionment. *Environmetrics* **17**: 663–681.

Dockery DW, Pope CA, Xu X, Spengler JD, Ware JH, Fay ME, Ferris BG, Speizer FE. 1993. An association between air pollution and Mortality in six U.S. cities. *The New England Journal of Medicine* **329**: 1753–1759.

Dominici F, Samet JM, Zeger SL. 2000. Combining evidence on air pollution and daily mortality from the 20 largest US cities: a hierarchical modelling strategy. *Journal of the Royal Statistical Society. Series A (Statistics in Society)* **163**: 263–302.

Dunson DB. 2000. Bayesian latent variable analysis for clustered mixed outcomes. *Journal of the Royal Statistical Society. Series B (Statistical Methodology)* **62**: 355–366.

Eberly S. 2005. *EPA PMF 1.1 User's Guide*. U.S. Environmental Protection Agency: Research Triangle Park, NC.

Evans MJ, Gilula Z, Guttman I. 1989. Latent class analysis of two-way contingency tables by Bayesian methods. *Biometrika* **76**: 557–563.

Gelman A, Carlin JB, Stern HS, Rubin DB. 2004. *Bayesian Data Analysis* (2nd edn). Chapman & Hall/CRC: Washington DC.

Hopke PK. 1988. Target transformation factor analysis as an aerosol mass apportionment method: a review and sensitivity study. *Atmospheric Environment* **22**: 1777–1792.

Ichir MM, Mohammad-Djafari A. 2004. Bayesian blind source separation of positive non-stationary sources. In *American Institute of Physics Conference Proceedings: 24th International Workshop on Bayesian Inference and Maximum Entropy Methods in Science and Engineering*, Vol. 735; 493–500.

Knuth KH. 1998. Bayesian source separation and localization. In *SPIE'98 Proceedings: Bayesian Inference for Inverse Problems*, Mohammad-Djafari A (ed.). San Diego, CA; 147–158.

Knuth KH. 1999. A Bayesian approach to source separation. In *Proceedings of the First International Workshop on Independent Component Analysis and Signal Separation: ICA'99*, Cardoso JF, Jutten C, Loubaton P (eds). Aussois, France; 283–288.

Knuth KH, Vaughan HG, Jr. 1999. Convergent Bayesian formulations of blind source separation and electromagnetic source estimation. In *Maximum Entropy and Bayesian Methods, Munich 1998*, von der Linden W, Dose V, Fischer R, Preuss R. (eds). Dordrecht: Kluwer; 217–226.

Lee JH, Hopke PK, Turner JR. 2006. Source identification of airborne $PM_{2.5}$ at the St. Louis-Midwest Supersite. *Journal of Geophysical Research* **111**: D10S10. DOI:10.1029/2005JD006329

Miskin J, MacKay D. 2001. Ensemble learning for blind source separation. In *Independent Component Analysis: Principles and Practice*, Roberts S, Everson R (eds). Cambridge University Press: Cambridge, UK; 209–233.

Moussaoui S, Brie D, Caspary O, Mohammad-Djafari A. 2004. A Bayesian method for positive source separation. In *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP'04)*, Vol. 5; 485–488.

Paatero P. 2003. *User's Guide for Positive Matrix Factorization Programs PMF2 and PMF3, Part 1: Tutorial*. University of Helsinki: Helsinki, Finland.

Paatero P, 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 ES, Guttorp P, Henry RC. 2001. Multivariate receptor modeling for temporally correlated data by using MCMC. *Journal of the American Statistical Association* **96**: 1171–1183.

Park ES, Oh MS, Guttorp P. 2002. Multivariate receptor modeling and model uncertainty. *Chemometrics and Intelligent Laboratory Systems* **60**: 49–67.

Rogers GS, Young DL. 1973. On the products of powers of generalized Dirichlet components with an application. *The Canadian Journal of Statistics* **1**: 159–169.

Rowe DB. 2003. *Multivariate Bayesian Statistics: Models for Source Separation and Signal Unmixing*. Chapman & Hall/CRC: Boca Raton, Florida.

DOI: 10.1002/env