Atmospheric Environment 43 (2009) 510–519

Contents lists available at ScienceDirect

Atmospheric Environment

journal homepage: www.elsevier.com/locate/atmosenv

Bayesian treatment of a chemical mass balance receptor model with multiplicative error structure

Andrew Keats [a,*], Man-Ting Cheng [b], Eugene Yee [c], Fue-Sang Lien [a]

[a] Department of Mechanical Engineering, University of Waterloo, Waterloo, ON N2L 3G1, Canada
[b] Department of Environmental Engineering, National Chung Hsing University, 250 Kuokuang Road, Taichung 402, Taiwan, ROC
[c] Defence R&D Canada – Suffield, P.O. Box 4000, Medicine Hat, AB T1A 8K6, Canada

ARTICLE INFO

Article history:
Received 25 January 2008
Received in revised form
8 October 2008
Accepted 9 October 2008

Keywords:
Receptor model
Chemical mass balance
Bayesian inference
Multiplicative error
Source apportionment
Hamiltonian Markov chain Monte Carlo

ABSTRACT

The chemical mass balance (CMB) receptor model is commonly used in source apportionment studies as a means for attributing measured airborne particulate matter (PM) to its constituent emission sources. Traditionally, error terms (e.g., measurement and source profile uncertainty) associated with the model have been treated in an additive sense. In this work, however, arguments are made for the assumption of multiplicative errors, and the effects of this assumption are realized in a Bayesian probabilistic formulation which incorporates a 'modified' receptor model. One practical, beneficial effect of the multiplicative error assumption is that it automatically precludes the possibility of negative source contributions, without requiring additional constraints on the problem. The present Bayesian treatment further differs from traditional approaches in that the source profiles are inferred alongside the source contributions. Existing knowledge regarding the source profiles is incorporated as prior information to be updated through the Bayesian inferential scheme. Hundreds of parameters are therefore present in the expression for the joint probability of the source contributions and profiles (the posterior probability density function, or PDF), whose domain is explored efficiently using the Hamiltonian Markov chain Monte Carlo method. The overall methodology is evaluated and results compared to the US Environmental Protection Agency's standard CMB model using a test case based on PM data from Fresno, California.

1. Introduction

Source apportionment studies performed using receptor models employ information about the chemical makeup of sampled particulate matter (PM) in order to infer the relative contributions made by emission sources. Given a set of relevant source profiles (which indicate the relative proportions of constituent species present in the PM released by each individual emission source) as well as knowledge of the constituent elemental and molecular species which comprise the sampled PM, the receptor model provides a source–receptor relationship which is used to estimate the most important individual contributors to the sample.

The chemical mass balance (CMB) receptor model (Watson et al., 1990), which is commonly used to perform source apportionment (Watson and Chow, 2001; Kim and Henry, 1999; Chio et al., 2004; Marmur et al., 2007), essentially reduces the problem to one of constrained multiple linear regression. In the standard CMB model, mass concentrations of individual species are assumed to be linear combinations of emission source contributions:

$$y_i = \sum_{j=1}^{N} X_{ij}\beta_j + \varepsilon_i, \tag{1}$$

where N is the number of source profiles; y_i is the measured concentration of the ith species, and X_{ij} is the fractional amount of the ith species originating from the jth source. The target quantities, β_j, are the calculated contributions of the jth source to the receptor. Given that there are M species of interest (both elemental and molecular), the X_{ij} constitute a matrix, $\mathbb{X}^{M\times N}$, which is usually not square. The vector β can be obtained using, e.g., a least-squares method (Watson et al., 1990). The CMB model accounts for measurement uncertainty (and to a certain extent, model uncertainty) through the term ε_i, which is commonly assumed to be additive, uncorrelated and Gaussian in nature.

In the literature, the CMB model has been reinterpreted and implemented in several different but related ways, with the primary differences lying in the treatment of its associated measurement and model errors. Aside from the ordinary least-squares approach, Christensen and Gunst (2004) outline four

* Corresponding author. Tel.: +1 416 2608934.
E-mail address: wakeats@alumni.uwaterloo.ca (A. Keats).

1352-2310/$ – see front matter © 2008 Elsevier Ltd. All rights reserved.
doi:10.1016/j.atmosenv.2008.10.031

alternative approaches to source apportionment based on the CMB model, each of which yields a specific estimator for β along with its variance. One popular method, known as the effective variance solution, explicitly accounts for uncertainty in the source composition matrix $\mathbb{X}$ (we hereafter denote this uncertainty by $\sigma_{X_{ij}}$) in addition to measurement uncertainties (Watson et al., 1990). For the methods described by Christensen and Gunst which do account for model error explicitly, this error is assumed to be the mass of the ith species that is not accounted for by all sources in the model. By contrast, the present work presumes a more holistic (although less precise) interpretation of the model error – we assume it to be representative of any prospective failure of the CMB model to calculate the true species concentration y_i, given the source contributions β_j. As with any other predictive model, this failure might arise as a consequence of physical reality's deviation from any or all of the basic assumptions behind the CMB model (lists of these assumptions can be found in the works of Christensen and Gunst, 2004 or Seinfeld and Pandis, 2006).

Bayesian approaches to source apportionment are becoming increasingly common. Bayesian inference has already proven to be a useful tool for multivariate statistical analysis, and its application to receptor models is a part of this pattern. Chan et al. (1996) applied Bayesian inference to a receptor modelling problem in Taipei and used Markov chain Monte Carlo (MCMC) to sample from a joint posterior distribution for the source contributions β and the combined model and measurement uncertainties. In a similar vein to the present work, they log-transformed the data and used a modified CMB equation of the form

$$\log y_i = \log\left(\sum_{j=1}^{N} X_{ij}\beta_j\right) + \varepsilon_i, \tag{2}$$

with ε_i being normally distributed and uncorrelated. We expand on their analysis, however, to consider model, measurement and source profile errors separately, and adopt informative prior distributions for these errors.

A more general problem based on the chemical mass balance approach is source identification (and subsequent apportionment), frequently conducted using techniques of multivariate receptor modelling. Bayesian inference provides a valuable framework for addressing this problem, as seen in the work of Park et al. (2002). They consider both measurement and model parameter uncertainties and formulate a posterior distribution based on truncated normal and conjugate priors for these quantities. Such an approach presumes additive behaviour for the errors, whereas the present work adopts a multiplicative assumption (although Park et al. did conduct investigations using log-normally distributed data). Park et al. (2001) extended the Bayesian approach to account for temporal correlations in the data, improving estimates for source compositions.

Billheimer (2001) and Kashiwagi (2004) also address source apportionment in a Bayesian framework. They both treat the data compositionally (viz., as vectors which convey relative proportions). Kashiwagi examines the case where one source is unknown, and compares a few different distributions for the model error (including log-normal and truncated normal), evaluating their suitability using estimates of the posterior mean and variance. The present work is more similar to that of Billheimer in that positivity is maintained and prior information about the source profiles and measurement errors is incorporated in a cogent way. However, we do not explicitly treat the data as being compositional, and whereas Billheimer uses the logistic normal distribution for model parameters, we adopt a log-normal distribution. It is interesting to note that Billheimer obtains source contribution estimates based on posterior medians, a reflection of the fact that the median is a better indicator of central tendency in such positively skewed distributions (Slob, 1994). Recent work of Lingwall et al. (2008) adopts a Bayesian approach which maintains positivity while treating data compositionally through the use of a generalized Dirichlet distribution. Lingwall et al. address both exploratory aspects (source profiles considered unknown) and confirmatory (CMB apportionment) aspects of the problem.

The question of how to treat errors in the chemical mass balance model (additively vs. multiplicatively) has not received a great deal of attention in the literature, perhaps because the log-normal distribution is not as firmly embedded in the scientific consciousness as the normal distribution. Some would argue that the log-normal occurs just as frequently as the normal (Limpert et al., 2001), as a consequence of multiplicative error behaviour. Such errors require alternate techniques of analysis, such as using the coefficient of variation (CV) and median as measures of spread and central tendency, respectively. In the work by Watson and Chow (2001), source profile entries for certain elements can be seen to display uncertainties which are almost two orders of magnitude larger than the average abundances. Although these uncertainties (and averages) are commonly expressed in additive terms, we advocate reconsidering them in a multiplicative sense.

In the next section, we phrase the source apportionment problem in a Bayesian probabilistic sense by giving consideration to the types of uncertainty present. Quantifying these uncertainties leads to a comprehensive probabilistic expression for the source and profile parameters (this expression is known as the posterior distribution), which must be sampled in order to obtain statistics (e.g., estimates of expected source contributions). Section 3 describes the specific MCMC technique used for sampling from the posterior distribution. Since there are potentially hundreds of parameters present, conventional Markov chain Monte Carlo sampling techniques remain challenging to execute (due to the high dimensionality of the parameter space). Indeed, we employ a Hamiltonian MCMC algorithm, utilizing the gradients of the log-posterior in order to accomplish the sampling in a reasonable amount of time. The overall methodology is evaluated in Section 4 using $PM_{2.5}$ (particulate matter with an aerodynamic diameter less than 2.5 μm) data from Fresno, California.

2. Bayesian formulation

Phrasing the source apportionment problem in a Bayesian framework allows one to obtain an expression for the probability of the source contributions which is consistent with any assumptions made about the nature of measurement and model errors. The Bayesian methodology culminates in an expression for the posterior distribution,

$$P(\boldsymbol{\beta}, \mathbb{X}|\boldsymbol{y}, I), \tag{3}$$

where $\boldsymbol{\beta}$ is the vector of N unknown source contributions β_j, $\mathbb{X}^{M\times N}$ is the matrix of source profile information X_{ij}, $\boldsymbol{y}$ is the vector of M elemental and chemical species measurements y_i, and I represents background information pertaining to our problem. In traditional CMB-based source apportionment, the source profiles $\mathbb{X}$ are often considered to be fixed, and uncertainty surrounding them enters the calculation in the form of the standard deviations $\sigma_{X_{ij}}$ (which are reported along with the profiles). However, for the present approach we consider $\mathbb{X}$ to be unknown (although our prior knowledge about the entries X_{ij} and their standard deviations acts to 'constrain' the range of plausible values for $\mathbb{X}$) and requiring inference. According to Bayes' theorem, the posterior probability is proportional to the likelihood multiplied by the prior:

$$P(\boldsymbol{\beta}, \mathbb{X}|\boldsymbol{y}, I) \propto P(\boldsymbol{y}|\boldsymbol{\beta}, \mathbb{X}, I)P(\boldsymbol{\beta}, \mathbb{X}|I). \tag{4}$$

Here we assume a priori that β and $\mathbb{X}$ are statistically independent.[1] In this case, the prior factors as follows:

$$P(\beta, \mathbb{X}|I) = P(\beta|I)P(\mathbb{X}|I). \tag{5}$$

When the species measurements y_i are statistically independent (in this work we do not consider the dependent case), the likelihood factors as follows:

$$P(\mathbf{y}|\beta, \mathbb{X}, I) = \prod_{i=1}^{M} P(y_i|\beta, \mathbf{X}_i, I), \tag{6}$$

where $\mathbf{X}_i$ is a vector of length N consisting of the ith row of the matrix $\mathbb{X}$. Under similar assumptions of independence, the prior distributions for $\mathbb{X}$ and β factor as follows:

$$P(\mathbb{X}|I) = \prod_{i=1}^{M}\prod_{j=1}^{N} P(X_{ij}|I), \tag{7}$$

$$P(\beta|I) = \prod_{j=1}^{N} P(\beta_j|I). \tag{8}$$

Note that the Bayesian formulation presented above is independent of the specific choice of model used for source apportionment. Before proceeding with the model definition, however, we reiterate that in this work, both measurement and model errors are considered separately, so at this point it is important to distinguish y_i, the measured concentration of the ith species, from $\check{y}_i$ ($= \sum_{j=1}^{N} X_{ij}\beta_j$) the modelled concentration of the ith species.

2.1. Sources and types of uncertainty

The probability $P(\mathbf{y}|\beta, \mathbb{X}, I)$ is effectively determined by quantifying the model and measurement errors, while the prior probability $P(\mathbb{X}|I)$ is specified based on the degree of uncertainty surrounding the source profiles. Before proceeding with the expressions for the likelihood, prior and posterior PDFs (Sections 2.3–2.5, respectively), we first explain the roles of errors in the CMB approach by relating the modelled and measured quantities to their true (but unknown) counterparts.

2.1.1. Measurement and model errors

Under a multiplicative noise assumption, the logarithm of the noise is additive and Gaussian, rendering the untransformed noise log-normally distributed. The measured species concentrations y_i are therefore related to the true quantities in the following way:

$$\begin{aligned} \log(y_i) &= \log(y_i^{\text{true}}) + \log(\varepsilon_i), \quad \log(\varepsilon_i) \sim N(0, \sigma_i); \\ y_i &= \varepsilon_i y_i^{\text{true}}, \qquad\qquad\quad \varepsilon_i \sim LN(0, \sigma_i), \end{aligned} \tag{9}$$

where $LN(0, \sigma_i)$ is a log-normal distribution based on a Gaussian distribution parameterized by a zero mean and a standard deviation of σ_i. We also assume that model uncertainty, like the measurement uncertainty, can be characterized as log-normal noise:

$$\begin{aligned} \log(\check{y}_i) &= \log(y_i^{\text{true}}) + \log(\check{\varepsilon}_i), \quad \log(\check{\varepsilon}_i) \sim N(0, \check{\sigma}_i); \\ \check{y}_i &= \check{\varepsilon}_i y_i^{\text{true}}, \qquad\qquad\quad \check{\varepsilon}_i \sim LN(0, \check{\sigma}_i). \end{aligned} \tag{10}$$

[1] Note that a prior assumption of independence does not rule out posterior correlation.

The probability density function (PDF) for the errors ε_i is given by

$$P(\varepsilon_i) = \frac{1}{\varepsilon_i \sigma_i \sqrt{2\pi}} \exp\left[-\frac{\log^2(\varepsilon_i)}{2\sigma_i^2}\right]. \tag{11}$$

One consequence of assuming a multiplicative error structure is that existing data on measurements, source profiles and their variances must be converted to a form compatible with the log-normal distribution. For example, in the PDF of Eq. (11), the quantity σ_i expresses the standard deviation of the log-transformed ε_i. In order for Eq. (11) to properly describe the untransformed ε_i, tabulated measurement errors must be converted to express coefficients of variation (CVs) instead of standard deviations. The σ_i should then be chosen based on the CVs, instead of being extracted directly from the data. This conversion will be addressed in Section 2.2.

2.1.2. Source profile uncertainty

Source profile entries X_{ij} are assumed to be distributed log-normally about median values $X_{o,ij}$:

$$P(X_{ij}|I) = \frac{1}{X_{ij}\sigma_{X_{ij}}\sqrt{2\pi}} \exp\left[\frac{-\log^2(X_{ij}/X_{o,ij})}{2\sigma_{X_{ij}}^2}\right]. \tag{12}$$

In Watson and Chow (2001), source profile estimates are stated in terms of average fractional abundances and their variances. We wish to choose the $X_{o,ij}$ to be equivalent to the median fractional abundances, so a conversion from mean to median fractional abundances is necessary. This conversion depends on the stated source profile variances, and is given by Eq. (15) in Section 2.2.1.

For certain source profiles, the list of species present is exhaustive (profile entries sum to 1). For these species, we specify the constraint equation:

$$\sum_{i=1}^{M} X_{ij} = 1, \text{ for fully specified source profile } j. \tag{13}$$

2.1.3. Multiplicative vs. additive error

To justify using log-normal distributions to characterize both our measurement and model noise, we invoke the principle of maximum entropy (MaxEnt), as put forward by Jaynes (2003). MaxEnt provides a way of selecting distributions which are maximally noncommittal with respect to any information about them which remains unknown. If all that is known about our noise are its mean and variance, then according to the MaxEnt principle, the least informative distribution that can be chosen is the Gaussian.

In the CMB problem, we manipulate concentration and source profile data which may only take positive values. Uncertainties related to such information are commonly dealt with in a 'scaling fashion' in which uncertainties are specified as relative percentages of the measurement, rather than absolute plus-or-minus values (Jaynes, 2003; Sivia and Skilling, 2006). Indeed, the source profiles themselves are vectors of positive, fractional amounts. It therefore makes sense to deal with concentration data and source profiles in a logarithmic setting (Tarantola, 2006). Adopting a Gaussian distribution for the logarithmic noise implies that the untransformed noise should be considered to belong to a log-normal distribution.

In some cases, individual measurements and source profile entries are specified as being zero. While the information contained in such zero values is undoubtedly important (such data could have a large impact on the source apportionment by permitting the outright rejection of source profiles identified by certain elements which act as markers), directly adopting a log-normal error distribution (noise prior) in these cases leads to an overemphasis on these zero values within the Bayesian framework. This overemphasis can be rectified by recognizing that measurements and source profiles are subject to lower limits of detection, which are

almost certainly non-zero. We therefore replace zero entries by an estimate for the lower limit of detection and supply an appropriate value for the uncertainty.

2.2. Assigning distribution parameters

Although we wish to characterize uncertainties as log-normal, the measured concentration data and source profile information are actually supplied as sample mean and standard deviation information. Furthermore, the model uncertainty is not explicitly known and must be either assumed or inferred.

2.2.1. Specifying measurement error parameter σ_i and profile parameters $\sigma_{X_{ij}}$, $X_{o,ij}$

Out of the available measures of central tendency (e.g., the mean, median, and mode), the mean is typically the more popular. However, when the data are known to be log-normally distributed, the median is a more appropriate measure of central tendency than the mean, for the following reasons (Slob, 1994):

(1) The median of the log-normal distribution is the 'multiplicative analogue' of the mean: if the mean of $\log(x)$ is μ, then the median of x is $x_0 = \exp(\mu)$. In contrast, the mean of x is $\exp(\mu + \sigma^2/2)$, which depends on the variance of the untransformed data [$\log(x)$].
(2) The median is less sensitive than the mean to extreme values in the data. Log-normally distributed data are positive, and may be spread over several orders of magnitude. Interpreting such data from an absolute rather than a logarithmic standpoint can result in a mean value which is significantly higher than the location of the distribution's central region.

Slob goes on to argue that "the coefficient of variation is a natural measure of uncertainty in the log-normal distribution, since $\exp(\sigma^2)$ [...] can be recognized as the multiplicative analogue of σ^2, which has an additive nature." The coefficient of variation is the ratio of the standard deviation to the mean, which for the log-normal distribution is:

$$\mathrm{CV} = \frac{\text{standard deviation}}{\text{mean}} = \sqrt{\exp(\sigma^2) - 1}. \tag{14}$$

Adopting the CV as our preferred measure of uncertainty is especially useful in the context of the CMB model, where uncertainties are best represented as fractions of measured values (with the exception of zero-measurements, in which case the uncertainty would be representative of the lower limit of detection). We therefore assign measurement uncertainties σ_i using Eq. (14), with the CV taking the percentage value of the measurement standard deviation. For example, if a measurement takes a value of 100 units, and the measurement standard deviation is specified as 10 units, then σ_i is chosen such that CV = 10%.

The source profile distribution parameters $X_{o,ij}$ are specified based on quoted estimates for the average abundances, which we denote $\mu_{X_{ij}}$; in other words, $\mu_{X_{ij}}$ represents the mean of the (i,j)th log-normal distribution for X_{ij}. Manipulating relationships for the mean and CV of a log-normal distribution, an expression is obtained which allows one to obtain the median value given the mean and CV:

$$X_{o,ij} = \mu_{X_{ij}}\left(1 + \mathrm{CV}_{X_{ij}}^2\right)^{-1/2}. \tag{15}$$

The parameters $\sigma_{X_{ij}}$ are determined using Eq. (14), which implies that:

$$\sigma_{X_{ij}}^2 = \log\left(1 + \mathrm{CV}_{X_{ij}}^2\right). \tag{16}$$

2.2.2. Model error $\breve{\sigma}_i$

The CMB model rests on a number of assumptions which may be violated in unpredictable ways. It is difficult to derive a single useful noise estimate for the model because it is implemented across many different scenarios. Traditional least-squares approaches to estimating source contributions through the CMB model carry a built-in assumption of a Gaussian, independent error structure. This assumption often leads to negative mass estimates which must be truncated or otherwise remedied in an ad hoc manner. The present approach avoids this problem by characterizing errors as multiplicative (log-normal), and guarantees that the estimated masses are positive. However, the scale of these errors remains unknown, and we require a methodology to deal with this 'uncertainty of the uncertainty'. The following is a partial list of possible approaches (in order of decreasing subjectivity):

(1) Arbitrarily specify quantities for the model errors $\breve{\sigma}_i$. Equivalently, specify the CVs for the model errors.
(2) Select model errors to be proportional to a percentage of the concentration measurement errors (with both errors expressed as CVs). Use a single constant of proportionality to relate all of the model errors (we have just added a parameter to the inference). Specify a prior probability for this constant and incorporate it into the Bayesian inference scheme.
(3) Consider each model error to be an unknown parameter which must be inferred. The parameter values can either be sampled using MCMC, or else analytically marginalized (if possible). Whether these parameters are sampled or marginalized will not affect the marginal posterior distributions for any of the other parameters.

In this work we select option # 2 to account for model error. The posterior distribution must therefore incorporate a prior which assumes a state of ignorance with respect to the 'scale' parameter ρ. In a manner similar to Sivia and Skilling (2006), we adopt a form for the prior which does not require the specification of a finite upper bound on ρ:

$$P(\rho|I) = \frac{\rho_{\min}}{\rho^2}, \quad \rho \in (\rho_{\min}, \infty), \tag{17}$$

where $\rho_{\min}$ is a conservative lower bound for ρ (selected by the user). This limit ensures that the PDF of Eq. (17) is normalizable. The model CVs are rewritten in terms of the measurement CVs: $\breve{\mathrm{CV}}_i \to \rho \mathrm{CV}_i$, $i = 1, 2, \ldots, M$.

2.3. Assignment of the likelihood $P(\boldsymbol{y}|\boldsymbol{\beta}, \mathbb{X}, I)$

The likelihood is obtained by marginalizing the joint PDF of the measured (y_i) and true (y_i^{true}) data given the modelled $(\breve{y}_i = \sum X_{ij}\beta_j)$ data:

$$\begin{aligned} P(y_i|\boldsymbol{\beta}, \boldsymbol{X}_i, I) &= \int_{\text{all } y_i^{\text{true}}} \mathrm{d}y_i^{\text{true}} P(y_i, y_i^{\text{true}}|\boldsymbol{\beta}, \boldsymbol{X}_i, I) \\ &= \int_{\text{all } y_i^{\text{true}}} \mathrm{d}y_i^{\text{true}} P(y_i|y_i^{\text{true}}, \boldsymbol{\beta}, \boldsymbol{X}_i, I) P(y_i^{\text{true}}|\boldsymbol{\beta}, \boldsymbol{X}_i, I). \end{aligned} \tag{18}$$

Expanding and combining both expressions in the integrand (each is a log-normal PDF[2]), the likelihood can be written as:

[2] A detailed derivation would show that each expression is obtained through a convolution integral, $\int \mathrm{d}\varepsilon_i P(\varepsilon_i)\delta(y_i - \varepsilon_i y_i^{\text{true}})$, with $P(\varepsilon_i)$ specified by Eq. (11).

$$P(y_i|\beta, \boldsymbol{X}_i, I) = \int_{\text{all } y_i^{\text{true}}} \mathrm{d}y_i^{\text{true}} \frac{1}{2\pi\sigma_i\check{\sigma}_i y_i \check{y}_i} \times \exp\left[-\frac{\log^2(y_i/y_i^{\text{true}})}{2\sigma_i^2} - \frac{\log^2(\check{y}_i/y_i^{\text{true}})}{2\check{\sigma}_i^2}\right], \quad (19)$$

where $\check{\sigma}_i^2$ is the variance pertaining to the log of the ith modelled datum ($\log \check{y}_i$) and σ_i^2 is the variance pertaining to the log of the ith measured datum ($\log y_i$). Integration over $y_i^{\text{true}} \in (0, \infty)$ yields the likelihood:

$$P(y_i|\beta, \boldsymbol{X}_i, I) = a_i\sqrt{\frac{\pi}{r_i}} \exp\left[\frac{(q_i+1)^2}{4r_i} - p_i\right],$$
$$\text{where } a_i = \frac{1}{2\pi\sigma_i\check{\sigma}_i y_i \check{y}_i}, \quad p_i = \frac{\log^2 y_i}{2\sigma_i^2} + \frac{\log^2 \check{y}_i}{2\check{\sigma}_i^2},$$
$$q_i = \frac{\log y_i}{\sigma_i^2} + \frac{\log \check{y}_i}{\check{\sigma}_i^2}, \quad r_i = \frac{1}{2\sigma_i^2} + \frac{1}{2\check{\sigma}_i^2}. \quad (20)$$

2.4. Assignment of the prior probabilities

We assume a state of ignorance with respect to each parameter β_j and therefore adopt a Jeffreys' prior (Sivia and Skilling, 2006), constrained by our estimates for the lower and upper bounds for β_j:

$$P(\beta_j|I) = \frac{1}{\beta_j \log\left(\beta_j^{\max}/\beta_j^{\min}\right)}, \quad \beta_j \in \left[\beta_j^{\min}, \beta_j^{\max}\right]. \quad (21)$$

Note that under this prior, $P(\log \beta_j|I) \sim$ constant for $\beta_j \in [\beta_j^{\min}, \beta_j^{\max}]$.

The prior distribution $P(X_{ij}|I)$ was given in Eq. (12), and as mentioned above, the prior distribution $P(\rho|I)$, Eq. (17), must also form part of the posterior distribution.

2.5. The full posterior distribution

Calculations are most easily performed using the logarithm of the posterior distribution, which takes the following form:

$$\log P(\beta, \mathbb{X}, \rho|\boldsymbol{y}, I) = \sum_{i=1}^{M}\left[\log\left(a_i\sqrt{\frac{\pi}{r_i}}\right) + \frac{(q_i+1)^2}{4r_i} - p_i\right] - \sum_{i=1}^{M}\sum_{j=1}^{N}\left[\log\left(X_{ij}\sigma_{X_{ij}}\sqrt{2\pi}\right) + \frac{\log^2(X_{ij}/X_{\mathrm{o},ij})}{2\sigma_{X_{ij}}^2}\right] - \sum_{j=1}^{N}\log\left[\beta_j \log\left(\beta_j^{\max}/\beta_j^{\min}\right)\right] - 2\log\rho + \log\rho_{\min} + C, \quad (22)$$

where C is a constant derived from taking the logarithm of the (unknown) normalization constant for the posterior distribution.

2.5.1. Gradients of the negative log-posterior

As explained below in Section 3, we require knowledge of the partial derivatives of the negative log-posterior PDF in order to effectively sample from it. Fortunately, these expressions are available analytically, and are reproduced below for completeness.

$$\frac{\partial}{\partial\beta_j}(-\log P) = \sum_{i=1}^{M}\frac{X_{ij}}{\check{y}_i}\left(\frac{\check{\sigma}_i^2 + \log\check{y}_i - \log y_i}{\check{\sigma}_i^2 + \sigma_i^2}\right) + \frac{1}{\beta_j} \quad (23a)$$

$$\frac{\partial}{\partial X_{ij}}(-\log P) = \frac{\beta_j}{\check{y}_i}\left(\frac{\check{\sigma}_i^2 + \log\check{y}_i - \log y_i}{\check{\sigma}_i^2 + \sigma_i^2}\right) + \frac{\log(X_{ij}/X_{\mathrm{o},ij})}{\sigma_{X_{ij}}^2 X_{ij}} + \frac{1}{X_{ij}} \quad (23b)$$

$$\frac{\partial}{\partial\rho}(-\log P) = \frac{2}{\rho} + \sum_{i=1}^{M}\frac{\rho\mathrm{CV}_i^2}{\left(\rho^2\mathrm{CV}_i^2+1\right)\left[\sigma_i^2 + \log\left(\rho^2\mathrm{CV}_i^2+1\right)\right]^2} \times\left[-\sigma_i^4 + \sigma_i^2(2\log(\check{y}_i/y_i)+1) - \log^2(y_i/\check{y}_i) + \log\left(\rho^2\mathrm{CV}_i^2+1\right)\right] \quad (23c)$$

3. Exploring the posterior distribution with Markov chain Monte Carlo

For a source apportionment involving N sources and M species, the posterior distribution involves up to $N \times M + N + 1$ parameters. For typical problems, this number could lie between 10^2 and 10^3. Clearly, tesselating and evaluating the posterior PDF over such a high-dimensional parameter space is computationally impractical, so here we use Markov chain Monte Carlo (MCMC) as a technique for sampling from the posterior distribution.

MCMC methods are well-documented in the literature (Neal, 1993; Gilks et al., 1996; MacKay, 2003; Gregory, 2005) and provide a way to explore high-dimensional parameter spaces more efficiently than conventional Monte Carlo integration. MCMC algorithms work by generating a Markov chain of samples whose distribution tends asymptotically to a target distribution (in our case, the posterior PDF). This property ensures that time is not wasted generating samples from areas of the parameter space which contribute negligibly to the overall probability mass. We denote the sequence by $\boldsymbol{m}^{(k)} \in \mathbb{R}^{N_d}$, where N_d is the dimensionality of $\boldsymbol{m}^{(k)}$; $\boldsymbol{m}^{(k)}$ is the kth sample; and m_i is the ith component of the model parameter vector, $\boldsymbol{m}$.

One popular MCMC method, the Metropolis–Hastings algorithm (Metropolis et al., 1953; Hastings, 1970), is particularly simple to implement and requires only the ability to calculate values of the target distribution to within a constant of proportionality. Starting from an initial state $\boldsymbol{m}^{(1)}$, the Metropolis–Hastings method operates by generating trial steps $\tilde{\boldsymbol{m}}$ from a proposal distribution $q(\tilde{\boldsymbol{m}}|\boldsymbol{m})$. These trials are then either accepted with probability

$$\alpha\left(\tilde{\boldsymbol{m}}, \boldsymbol{m}^{(k)}\right) = \min\left(1, \frac{p\left(\tilde{\boldsymbol{m}}\right)q\left(\boldsymbol{m}^{(k)}\middle|\tilde{\boldsymbol{m}}\right)}{p\left(\boldsymbol{m}^{(k)}\right)q\left(\tilde{\boldsymbol{m}}\middle|\boldsymbol{m}^{(k)}\right)}\right), \quad (24)$$

or else rejected. Here, $p(\boldsymbol{m})$ is shorthand for the full posterior PDF of the parameters $\boldsymbol{m}$. Frequently, the proposal distribution q is chosen to be symmetric (e.g., uniform or normal, centred on the chain's location at the kth iteration), which simplifies the acceptance probability.

The Metropolis–Hastings MCMC algorithm is difficult to implement for the present source apportionment problem because, as Hajian (2007) points out, the efficiency[3] of this algorithm is inversely proportional to the number of parameters involved. Hajian advocates the use of Hamiltonian MCMC (originally due to

[3] The statistical efficiency of a chain is basically a measure of the MCMC method's effectiveness at generating samples which accurately describe the target PDF. An inefficient chain will suffer from large sample mean variance. Details can be found in Gilks et al. (1996) and Hajian (2007).

Duane et al. 1987, the method is also known as 'hybrid Monte Carlo'), for which the efficiency remains constant with dimensionality. This method is slightly more difficult to implement than Metropolis–Hastings, as it requires the use of auxiliary momentum variables and knowledge of the partial derivatives of the logarithm of the posterior PDF. Furthermore, all of our parameters of interest, $\boldsymbol{m} = (\beta, \mathbb{X}, \rho)$, are necessarily positive and must be transformed so that exploration of the posterior distribution (which must also be transformed) can be carried out in log-parameter space. However, due to the hundreds of parameters present in our source apportionment problem, and due to the fact that partial derivatives of the log-posterior are available analytically, Hamiltonian MCMC remains a valuable technique for exploring our parameter space.

3.1. Hamiltonian MCMC: implementation

In the Hamiltonian MCMC method, we augment the model parameters $\boldsymbol{m}$ with a vector of auxiliary momentum variables $\boldsymbol{p}$ and construct the Hamiltonian,

$$H(\boldsymbol{m}, \boldsymbol{p}) = U(\boldsymbol{m}) + K(\boldsymbol{p}), \tag{25}$$

where U, the 'potential energy', is the negative logarithm of the target (posterior) PDF, and K, the 'kinetic energy', is a quadratic function of the momenta: $K(\boldsymbol{p}) = (1/2)\boldsymbol{p}^{\mathrm{T}}\boldsymbol{p}$. The Hamiltonian (25) is explored in the way described by Algorithm 1.

Algorithm 1 Hamiltonian MCMC

select initial parameter values: $\boldsymbol{m}^{(0)}$
select leapfrog step sizes: ς
FOR $k = 1, 2, \ldots$ DO
  $\boldsymbol{p}^{(0)} \leftarrow$ sample from $N(0, 1)$
  $\tau \leftarrow 0$
  $\tilde{\boldsymbol{m}}^{(\tau)} \leftarrow \boldsymbol{m}^{(k-1)}$
  FOR $n = 1, 2, \ldots, N_{\text{leap}}$ DO {execute leapfrogging}
    $\boldsymbol{p}^{(\tau+\varsigma/2)} \leftarrow \boldsymbol{p}^{(\tau)} - (1/2)\varsigma^{\mathrm{T}}\nabla_{\boldsymbol{m}}U(\tilde{\boldsymbol{m}}^{(\tau)})$
    $\tilde{\boldsymbol{m}}^{(\tau+\varsigma)} \leftarrow \tilde{\boldsymbol{m}}^{(\tau)} + \varsigma^{\mathrm{T}}\boldsymbol{p}^{(\tau+\varsigma/2)}$
    $\boldsymbol{p}^{(\tau+\varsigma)} \leftarrow \boldsymbol{p}^{(\tau+\varsigma/2)} - (1/2)\varsigma^{\mathrm{T}}\nabla_{\boldsymbol{m}}U(\tilde{\boldsymbol{m}}^{(\tau+\varsigma)})$
    $\tau \leftarrow \tau + \varsigma$
  END FOR {end leapfrogging}
  calculate acceptance probability:
  $\alpha \leftarrow \min(1, \exp[H(\boldsymbol{m}^{(k-1)}, \boldsymbol{p}^{(0)}) - H(\tilde{\boldsymbol{m}}^{(\tau)}, \boldsymbol{p}^{(\tau)})])$
  $u \leftarrow$ sample from uniform(0, 1)
  IF $u < \alpha$ THEN
    $\boldsymbol{m}^{(k)} \leftarrow \tilde{\boldsymbol{m}}^{(\tau)}$ {accept the sample}
  ELSE
    $\boldsymbol{m}^{(k)} \leftarrow \boldsymbol{m}^{(k-1)}$ {reject the sample}
  END IF
END FOR

The leapfrog moves serve to transport the proposal $\tilde{\boldsymbol{m}}$ through the Hamiltonian along trajectories of constant energy. In this way, mixing (the tendency of the chain to explore different regions of parameter space) happens with more intensity than in a chain generated by the Metropolis–Hastings algorithm. In theory, the acceptance rate for the method should be 100%; however, because the Hamiltonian dynamics are not simulated exactly (the leapfrog algorithm discretely approximates the true trajectory; step sizes ς are finite), the acceptance rate does not necessarily reach 100%. Good acceptance rates[4] are obtained when each step size is chosen to be proportional to the standard deviation of the chain: $\varsigma_i \propto \mathrm{sd}(m_i)$. The number of leapfrog steps taken, N_{leap}, is also determined based on the desired rates of acceptance and chain exploration, and is subjected to small random perturbations at each step k in order to guarantee ergodicity. These issues, along with an analysis of Hamiltonian MCMC, are discussed in detail by Neal (1993).

3.2. Assessing chain convergence

Because the Markov chain of samples generated by an MCMC algorithm tends asymptotically to the target distribution, it is necessary to assess how well a chain of finite length approximates the target distribution. In the rest of this work, we adopt the method of Dunkley et al. (2005) for assessing chain 'convergence'. Briefly, Dunkley et al. analyzed the spectral properties of a Markov chain which had progressed beyond the initial 'burn-in' stage,[5] testing for convergence by examining parameter values obtained through the fitting of a model template power spectrum. They imposed the following two requirements:

(1) The distribution is not biased by correlated points, and the chain is drawing points throughout the full region of high probability. This is equivalent to saying that for the spectral representation of the chain (obtained using, e.g., a fast Fourier transform), frequencies lying below some pre-determined cut-off value have entered the white noise regime (viz., spectrally flat).
(2) The 'convergence ratio' is specified by an estimate for the ratio of the sample mean variance ($\sigma_{\bar{x}}^2$) to the variance of the underlying distribution (σ_0^2). This ratio is required to lie below some cut-off value, such as 0.01.

Full details regarding these convergence criteria can be found in the work of Dunkley et al. (2005). Hajian (2007) compares power spectra obtained using both Hamiltonian and Metropolis–Hastings MCMC, demonstrating the ability of the Hamiltonian method to achieve convergence more rapidly than Metropolis–Hastings for high-dimensional parameter spaces.

4. Test case: San Joaquin Valley Fine (SJVF) data

We evaluate the proposed Bayesian source apportionment methodology using a subset of the data collected at the Fresno site in California's San Joaquin Valley (SJV) as part of the 1988–1989 Valley Air Quality Study (VAQS). Details pertaining to this study (including the area's geography, dominant sources and measurement procedures) are presented in the work of Chow et al. (1992), where a source apportionment is performed using the US Environmental Protection Agency's (EPA) CMB software version 7.0 (Watson et al., 1990). In a similar vein to the study of Christensen and Gunst (2004), we perform a source apportionment using the $PM_{2.5}$ data and compare our results to those supplied in the EPA's CMB version 8.2 software users manual (Coulter, 2004). It should be noted that while Christensen and Gunst performed the source apportionment using data from each time period, and Seinfeld and Pandis (2006) performed the apportionment using annual average concentrations, our results are based on data from only a single observation (Feb 27, 1989, as listed in the manual). The full data set is available at the US EPA's internet site.[6]

4.1. Source contribution estimates

Estimates for the contributions of each individual source are obtained directly from the chains of samples generated by the

[4] In practice, a compromise is reached between acceptance rate and the speed of each chain's exploration. For this method applied to the source apportionment problem, we generally aim for acceptance rates higher than 50%.

[5] The burn-in period occurs at the beginning of the chain and consists of the initial set of samples which are generated as the chain moves from the starting conditions toward the target distribution. These samples are typically discarded.

[6] http://www.epa.gov/scram001/receptor_cmb.htm.

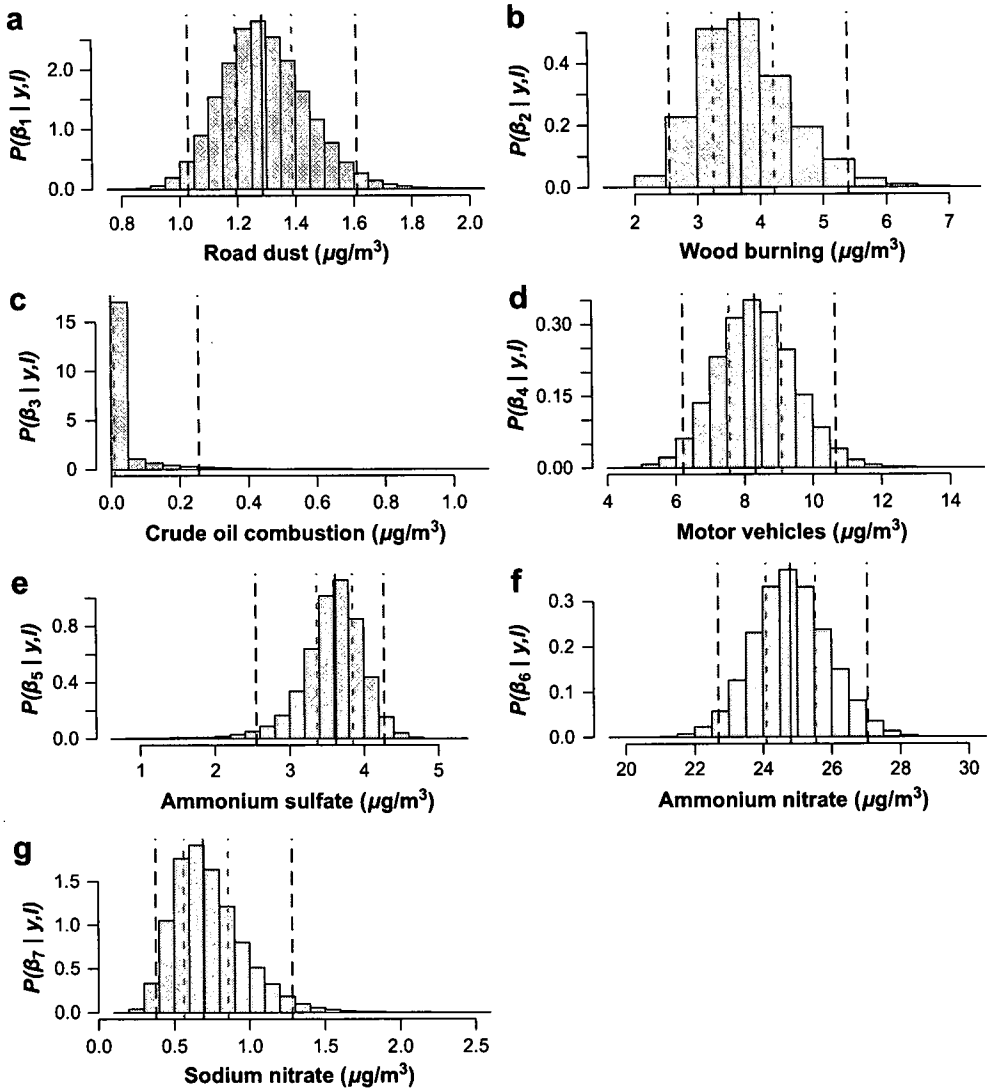


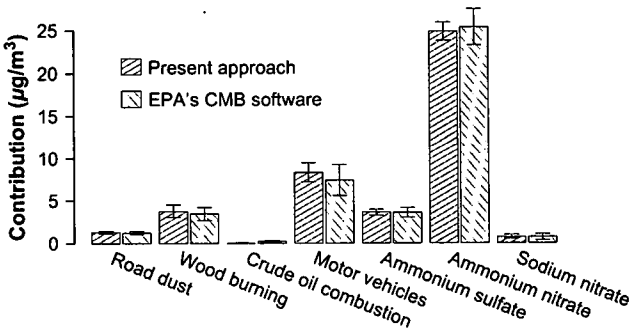
Fig. 1. Marginal parameter distributions generated from MCMC samples. The solid lines are at median values, and the long and short dashed lines delineate 95% and 50% credible intervals (CIs), respectively.

MCMC algorithm. Fig. 1 illustrates the sample histograms corresponding to each source contribution parameter β_j. For each histogram, the median value along with 50% and 95% credible intervals (each credible interval contains a given percentage of the probability mass) are shown.

A comparison between the apportionment results illustrated in Fig. 1 and those obtained using the EPA's CMB software (Coulter, 2004) is graphed and tabulated in Fig. 2. The results are similar, although the uncertainty bounds are generally decreased. Furthermore, the estimate for crude oil burning lies much closer to zero.

4.2. Source profile estimates

One of the strengths of the proposed method lies in the way it accounts for uncertainty in the source profiles. The Bayesian methodology allows us to obtain new estimates for the source profile concentrations ($\tilde{X}_{o,ij}$) and their uncertainties, based on prior information which consists of the profiles reported in the literature. Many source profile species concentrations are listed in the literature as being zero; however, as mentioned in Section 2.1.3, we replace these entries by an estimate for the lower limit of detection. For the present test case, this lower limit is assumed to be equal to the lowest non-zero quoted measurement uncertainty. Replacing these zero entries permits one to compare prior distributions for source profile species to their posterior distributions. Fig. 3 compares prior and posterior distributions of species for the motor vehicles profile. Shifts in median estimates for certain species are



Source	Present approach		EPA's CMB software	
	Median value	68% CI width	Estimate	2×uncertainty
Road dust	1.29	0.29	1.23	0.32
Wood burning	3.74	1.43	3.45	1.54
Crude oil combustion	1.3×10^{-5}	4.24×10^{-2}	0.20	0.26
Motor vehicles	8.30	2.24	7.39	3.65
Ammonium sulfate	3.62	0.73	3.57	1.11
Ammonium nitrate	24.8	2.16	25.4	4.22
Sodium nitrate	0.693	0.438	0.680	0.702

Fig. 2. Comparison between apportionments obtained using the present approach vs. those obtained using the EPA's CMB software. Error bars delineate 68% credible intervals (for the present approach) and estimated standard deviations (quoted by the EPA's CMB software).

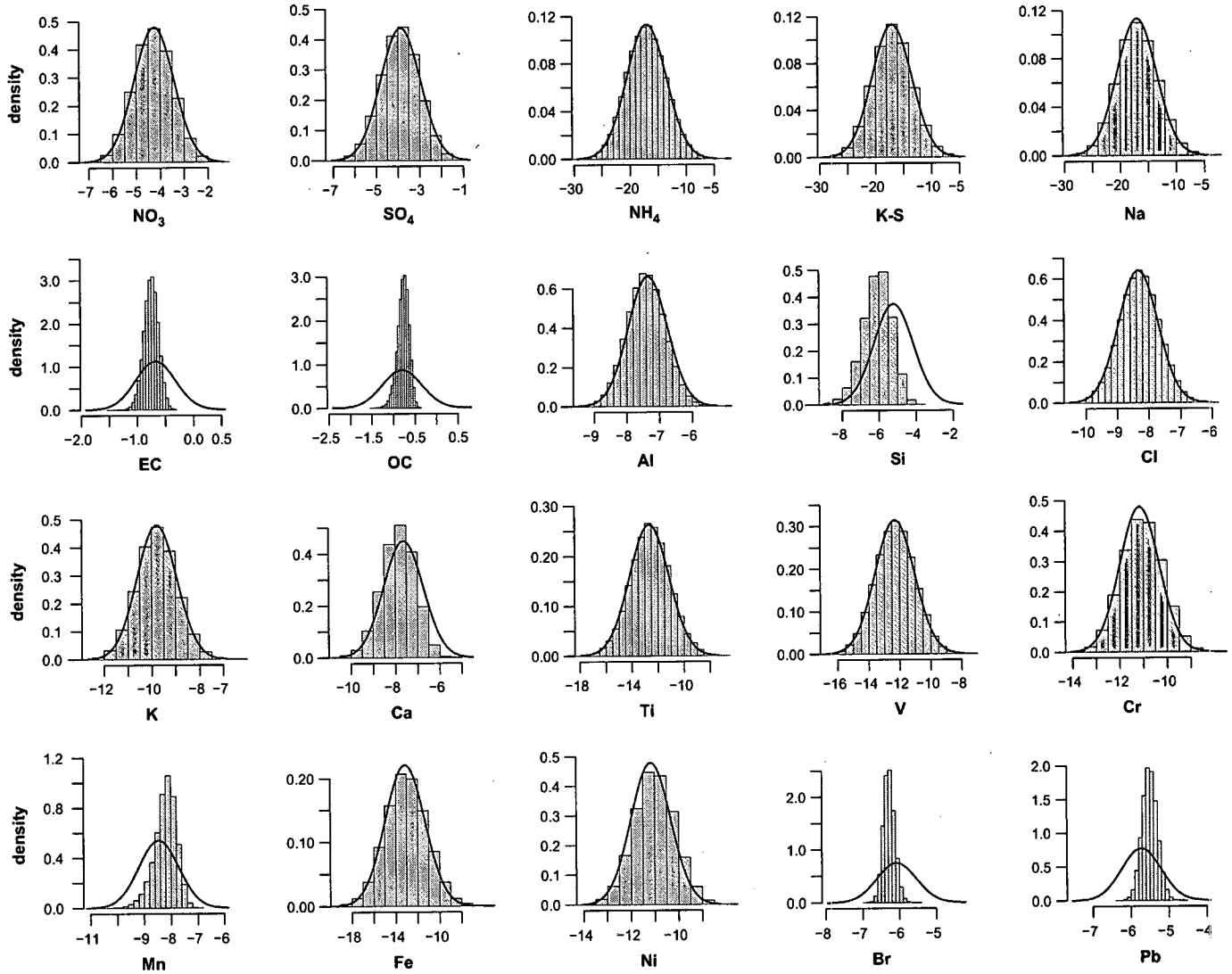


Fig. 3. Motor vehicles source profile: comparison of prior (solid line) to posterior density (histogram) for each of the species in the profile. Horizontal axes are logarithmic.

evident (e.g., silicon) while for other species, the posterior distribution appears much narrower than the prior.

Fig. 4 illustrates the ratio $(\check{X}_0/X_0)_{ij}$ for two different profiles; namely, wood burning and motor vehicles. Relatively large differences exist between the estimated and reported profiles for motor vehicles, whereas in contrast, the posterior estimate for the wood burning profile does not deviate greatly from its prior specification.

4.3. *Markov chain convergence*

To complete this section on the validation of our Bayesian methodology, we present some details regarding the application of the MCMC algorithm (described in Section 3.1) to the SJVF test case.

The Hamiltonian MCMC method used to obtain the samples (binned and summarized in Fig. 1) was run for 525×10^3 iterations, with each leapfrogging loop performing approximately 15 steps (this number was randomized to lie uniformly between 11 and 19). The algorithm resulted in an overall acceptance ratio of 69%. Initial values and step sizes ς were arrived at by running the algorithm iteratively using shorter chain lengths (on the order of 10^3–10^4 samples) in order to drive the chain behaviour to achieve a desirable acceptance ratio. The total number of iterations was chosen to be slightly larger than a power of two ($2^{19} = 524{,}288$) due to the fact that the chain convergence criteria are assessed using a spectral technique which requires a fast Fourier transform (FFT). Commonly available FFT routines operate most efficiently on data whose length is a power of two, and convergence was not found to happen after 2^{17} iterations.

Convergence was assessed using the criteria of Dunkley et al. (2005), and convergence ratios were found to lie well below 0.01 for all parameters β and $\mathbb{X}$. A diagram of the thinned power spectrum[7] for the chain corresponding to β_1 (road dust), along with a parametric fit to the spectrum, can be seen in Fig. 5. The parametric fit models the power spectrum according to a template proposed by Dunkley et al:

$$P(k) = P_0 \frac{\left(k^*/k\right)^{\alpha}}{\left(k^*/k\right)^{\alpha}+1}, \tag{26}$$

which effectively divides the spectrum into 'white noise' (the flat region to the left) and 'correlated noise' (the downward-sloping region at the right). Dunkley et al. showed that estimating P_0 (or,

[7] The full spectrum was obtained using 2^{19} samples.

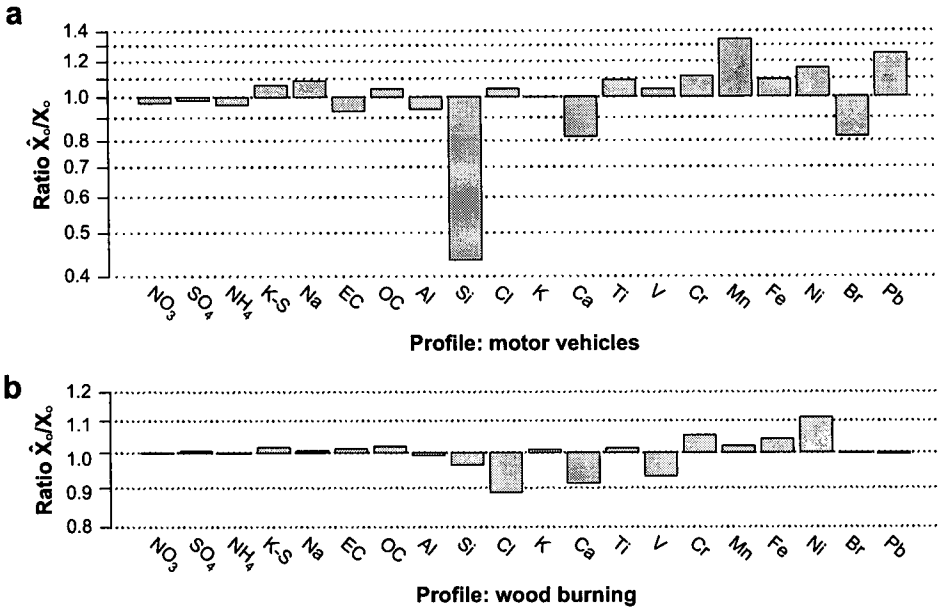


Fig. 4. Ratios of inferred to reported source profiles for motor vehicles and wood burning sources. Vertical scale is logarithmic.

$P(k)$ as $k \to 0$) is equivalent to estimating the sample mean variance of a long chain. The parametric fit allows one to estimate this quantity easily, as well as the quantity k^*, which indicates the extent of the white noise regime. Chains which have entered the white noise regime no longer experience correlations at the largest scales, and are likely to be exploring the full region of high posterior probability.

5. Conclusions

Applying the multiplicative error assumption to the CMB model within a Bayesian framework has proven to yield estimates for the source contributions (and their associated uncertainties) which are consistent with results obtained using more traditional methods. Despite the apparent similarity between these sets of estimates, the following attributes of the proposed methodology are noteworthy:

(1) The source profiles form part of the inference. We obtain posterior estimates for the source profiles which effectively improve model concentration estimates $\hat{\boldsymbol{y}}$ with respect to the measured PM data, $\boldsymbol{y}$. It should be mentioned that source profile estimates cannot be obtained if the posterior distribution is analytically marginalized over the parameters $\mathbb{X}$.

(2) The ability to account for an unknown model uncertainty. It is often difficult to objectively assess the impact that intrinsic features and assumptions of the model will have on the agreement between model results and the true (but unknown) data. Accounting for model uncertainty through an extra parameter ρ provides insight into how appropriate the CMB model is for a given scenario. For example, if the posterior estimate for the 'model error factor' (ρ) was large, it might indicate that the tabulated source profiles ($\mathbb{X}_o$) are inappropriate for the geographical area or data set under study (examination of the posterior estimates for the source profiles themselves would help to support or refute this hypothesis). Furthermore, many other sources of model uncertainty are possible, such as the assumption that species do not react as they are transported from source to receptor. The Bayesian methodology permits one to account for model uncertainty in any number of ways, allowing competing hypotheses (regarding the applicability of the model) to be tested.

(3) Hamiltonian MCMC permits the resulting high-dimensional parameter space to be sampled efficiently. A fortunate consequence of using the CMB receptor model is that gradients of the posterior distribution are available analytically, allowing one to use 'directed' MCMC methods such as the Hamiltonian method described in Section 3. For this method, efficiency remains constant with dimensionality, which is especially important since the SJVF test case involved 148 parameters.

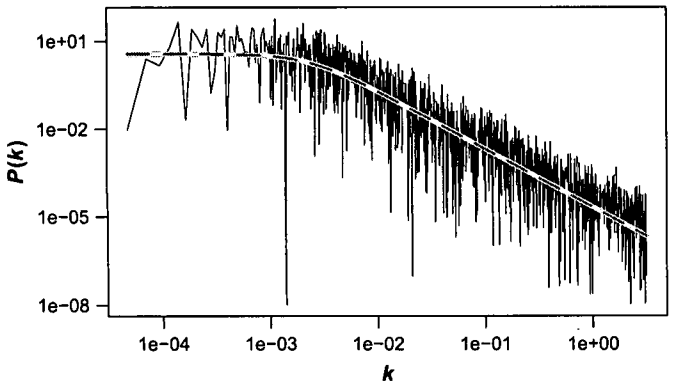


Fig. 5. Power spectrum ($P(k)$, cosmetically thinned) for the chain corresponding to the parameter β_1, generated by the Hamiltonian MCMC algorithm. The horizontally flat, leftmost portion of the parametric fit (dashed line) indicates that the chain has entered the white noise regime.

To conclude, the proposed approach is of greater generality and flexibility than traditional methods, yet manages to yield estimates with competing (or better) degrees of uncertainty. While it is beyond the scope of this paper to address issues surrounding source identification and selection, Bayesian approaches to model selection and hypothesis testing could be applied in a way compatible with the present framework.

Acknowledgements

The authors wish to acknowledge support from the Natural Sciences and Engineering Research Council (NSERC) of Canada and the National Science Council (NSC) of the Republic of China. Figures and numerical analysis were done using R (2005).

References

Billheimer, D., 2001. Compositional receptor modeling. Environmetrics 12, 451–467.
Chan, C.-C., Nien, C.-K., Hwang, J.-S., 1996. Receptor modeling of VOCs, CO, NO_x, and THC in Taipei. Atmospheric Environment 30, 25–33.
Chio, C.-P., Cheng, M.-T., Wang, C.-F., 2004. Source apportionment to PM10 in different air quality conditions for Taichung urban and coastal areas, Taiwan. Atmospheric Environment 38, 6893–6905.
Chow, J.C., Watson, J.G., Lowenthal, D.H., Solomon, P.A., Magliano, K.L., Ziman, S.D., Richards, L.W., 1992. PM_{10} source apportionment in California's San Joaquin Valley. Atmospheric Environment 26A, 3335–3354.
Christensen, W.F., Gunst, R.F., 2004. Measurement error models in chemical mass balance analysis of air quality data. Atmospheric Environment 38, 733–744.
Coulter, C.T., Dec 2004. EPA-CMB8.2. Users Manual EPA-452/R-04-011. Office of Air Quality Planning & Standards, Research Triangle Park, NC.
Duane, S., Kennedy, A.D., Pendleton, B.J., Roweth, D., 1987. Hybrid Monte Carlo. Physics Letters B 195, 216–222.
Dunkley, J., Bucher, M., Ferreira, P.G., Moodley, K., Skordis, C., 2005. Fast and reliable MCMC for cosmological parameter estimation. Monthly Notices of the Royal Astronomical Society 356, 925–936.
Gilks, W.R., Richardson, S., Spiegelhalter, D.J. (Eds.), 1996. Markov Chain Monte Carlo in Practice. Chapman & Hall/CRC, London.
Gregory, P.C., 2005. Bayesian Logical Data Analysis for the Physical Sciences: A Comparative Approach with *Mathematica*® Support. Cambridge University Press, Cambridge.
Hajian, A., 2007. Efficient cosmological parameter estimation with Hamiltonian Monte Carlo technique. Physical Review D 75, 083525. 11 pages.
Hastings, W.K., 1970. Monte Carlo sampling methods using Markov chains and their applications. Biometrika 57, 97–109.
Jaynes, E.T., 2003. Probability Theory: The Logic of Science. Cambridge University Press, Cambridge.
Kashiwagi, N., 2004. Chemical mass balance when an unknown source exists. Environmetrics 15, 777–796.
Kim, B.M., Henry, R.C., 1999. Diagnostics for determining influential species in the chemical mass balance receptor model. Journal of the Air & Waste Management Association 49, 1449–1455.
Limpert, E., Stahl, W.A., Abbt, M., 2001. Log-normal distributions across the sciences: keys and clues. Bioscience 51, 341–352.
Lingwall, J.W., Christensen, W.F., Reese, C.S., 2008. Dirichlet based Bayesian multivariate receptor modeling. Environmetrics 18, 618–629.
MacKay, D.J.C., 2003. Information Theory, Inference, and Learning Algorithms. Cambridge University Press, Cambridge.
Marmur, A., Mulholland, J.A., Russell, A.G., 2007. Optimized variable source-profile approach for source apportionment. Atmospheric Environment 41, 493–505.
Metropolis, N., Rosenbluth, A.W., Rosenbluth, M.N., Teller, A.H., Teller, E., 1953. Equations of state calculations by fast computing machine. Journal of Chemical Physics 21, 1087–1091.
Neal, R.M., 1993. Probabilistic Inference Using Markov Chain Monte Carlo Methods. Technical Report CRG-TR-93-1, Department of Computer Science, University of Toronto.
Park, E.S., Guttorp, P., 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., Guttorp, P., 2002. Multivariate receptor models and model uncertainty. Chemometrics and Intelligent Laboratory Systems 60, 49–67.
R Development Core Team, 2005. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria, ISBN: 3-900051-07-0. Available from: <http://www.R-project.org>.
Seinfeld, J.H., Pandis, S.N., 2006. Atmospheric Chemistry and Physics: From Air Pollution to Climate Change. John Wiley and Sons, Hoboken.
Sivia, D.S., Skilling, J., 2006. Data Analysis: A Bayesian Tutorial. Oxford University Press.
Slob, W., 1994. Uncertainty analysis in multiplicative models. Risk Analysis 14, 571–576.
Tarantola, A., 2006. Elements for Physics: Qualities, Quantities, and Intrinsic Theories. Springer, Berlin.
Watson, J.G., Chow, J.C., 2001. Source characterization of major emission sources in the Imperial and Mexicali Valleys along the US/Mexico border. The Science of the Total Environment 276, 33–47.
Watson, J.G., Robinson, N.F., Chow, J.C., Henry, R.C., Kim, B.M., Pace, T.G., Meyer, E.L., Nguyen, Q., 1990. The USEPA/DRI Chemical Mass Balance receptor model, CMB 7.0. Environmental Software 5, 38–49.