

Multivariate receptor models and model uncertainty

Eun Sug Park^{a,*}, Man-Suk Oh^b, Peter Guttorp^a

^aNational Research Center for Statistics and the Environment, University of Washington, Seattle, WA 98195, USA

^bDepartment of Statistics, Ewha Women's University, Seoul 120-750, South Korea

Abstract

Estimation of the number of major pollution sources, the source composition profiles, and the source contributions are the main interests in multivariate receptor modeling. Due to lack of identifiability of the receptor model, however, the estimation cannot be done without some additional assumptions.

A common approach to this problem is to estimate the number of sources, q , at the first stage, and then estimate source profiles and contributions at the second stage, given additional constraints (identifiability conditions) to prevent source rotation/transformation and the assumption that the q -source model is correct. These assumptions on the parameters (the number of sources and identifiability conditions) are the main source of model uncertainty in multivariate receptor modeling.

In this paper, we suggest a Bayesian approach to deal with model uncertainties in multivariate receptor models by using Markov chain Monte Carlo (MCMC) schemes. Specifically, we suggest a method which can simultaneously estimate parameters (compositions and contributions), parameter uncertainties, and model uncertainties (number of sources and identifiability conditions). Simulation results and an application to air pollution data are presented. © 2002 Elsevier Science B.V. All rights reserved.

Keywords: Latent variable models; Factor analysis models; Model uncertainty; Model identifiability; Number of sources; Posterior model probability; Marginal likelihood

1. Introduction

Multivariate receptor modeling aims to identify the pollution sources and assess the amounts of pollution by resolving the measured mixture of chemical species into the contributions from the individual source types. Its basic physical model comes from the laws of chemistry (the principles of mass conservation and

chemical mass balance [1]). Let p be the number of chemical species (measured variables) and q be the number of sources. Based on the chemical mass balance equation and the assumption that the relative amounts of the chemical species remain approximately the same as particles/gases travel from sources to the receptor, a multivariate receptor model takes the form of:

$$y_t = \sum_{k=1}^q \alpha_{tk} P_k + \varepsilon_t, \quad t = 1, \dots, n. \quad (1.1)$$

Here, $y_t = (y_{t1}, y_{t2}, \dots, y_{tp})$ is the t th observation at the receptor, $P_k = (p_{k1}, p_{k2}, \dots, p_{kp})$ is the k th

* Corresponding author. Texas Transportation Institute, The Texas A&M University System, 3135 TAMU, College Station, TX 77843-3135, USA.

E-mail address: e-park@ttimail.tamu.edu (E.S. Park).

source composition profile consisting of the fractional amount of each chemical species in the emissions from the k th source, α_{tk} is the contribution from the k th source at time t , and $\varepsilon_t = (\varepsilon_{t1}, \varepsilon_{t2}, \dots, \varepsilon_{tp})$ is the measurement error in the t th observation. In a vector form, model (1.1) can be equivalently written as:

$$y_t = \alpha_t P + \varepsilon_t \quad (1.2)$$

where $\alpha_t = (\alpha_{t1}, \alpha_{t2}, \dots, \alpha_{tq})$ and P is a $q \times p$ source composition matrix of which rows are the source composition profiles.

Statistically, model (1.2) can be considered as a latent variable model (see, e.g., Ref. [2]), specifically, as a factor analysis model in which y is a set of p variables that can be directly observed (manifest variables), α is a set of q variables that are unobservable (latent variables or factors), P is the unknown $q \times p$ factor loading matrix, and q is the unknown number of factors. The goal of latent variable models (multivariate receptor models) is to make inferences on q (the number of major sources), P (source composition profiles), and α (source contributions) based on y (data). As a matter of fact, this goal cannot be achieved without additional assumptions on the model.

The unknown number of factors (sources), q , is the first obstacle that we encounter, since P and α explicitly depend on q in model (1.2). Traditionally, q has been first estimated based on the sample correlation matrix or sample covariance matrix without necessarily using model (1.2) (see Refs. [3,4] and references therein for practical methods of estimating the number of factors in a non-Bayesian context). Except for Bartlett's modification to the likelihood ratio test, most of the commonly used methods such as Percent trace (choosing enough eigenvalues to account for a suitable proportion, say 90%, of the trace of the sample correlation matrix or the sample covariance matrix), Rule-of-one (choosing only eigenvalues of the sample correlation matrix which are greater than one), or Scree plot (choosing q at the 'knee' in the curve of the plot of sample eigenvalues) are ad-hoc in nature. Even Bartlett's test is not strictly valid as a sequential test procedure because it does not control the overall significance level. Other methods mostly used in chemometrics, such as Mali-

nowski's indicator function (see, e.g., Ref. [5]), cross-validation [6], and the NUMFACT [3] approach also lack a full theoretical justification in the sense that they do not provide standard errors for the estimates. Once q is estimated, inferences on P and α are usually made conditionally on the q -source model. Note that this approach ignores the uncertainty involved in q , which can be a big part of overall uncertainty.

Secondly, the parameters in model (1.2) are not uniquely defined even under the assumption that q is known, i.e., there are other parameterizations that produce the same data (rotational indeterminacy of factors plays a major role). This is nonidentifiability in latent variable models/multivariate receptor models, and additional restrictions on the parameters are required to remove it. These assumptions on the parameters are called "identifiability conditions." Park et al. [7] discussed a range of identifiability conditions for multivariate receptor models from a statistical point of view when the number of sources q is assumed to be known. Just like the number of sources q , these identifiability conditions are chosen in advance, and the estimation of P and α is carried out conditionally on that. This again ignores the uncertainty involved in the selection of identifiability conditions.

Each possible combination of q and identifiability conditions defines a different model. The previous approaches in multivariate receptor modeling select a single model and make inferences, conditionally on that model, *without taking account of model uncertainty*. In this paper, we adopt a Bayesian approach to provide the estimates of the model uncertainties, as well as the estimates for the parameters and their uncertainties within each model.

Practical model selection procedures often consist of two stages: choose a class of reasonable models and then select the best model within the class. For the number of sources q , there is an upper bound such that $q < p$ (as a matter of fact, it is often much less than p , see Ref. [8]). For the set of identifiability conditions, however, there is no such bound, and there could be, in principle, infinitely many different identifiability conditions. For this reason, we restrict the type of identifiability conditions to be compared to those that are often used in a receptor modeling context. One such type of identifiability conditions is prespecifica-

tion of zero elements in the source composition matrix:

C1. There are at least $q - 1$ zero elements in each row of P .

C2. The rank of $P^{(k)}$ is $q - 1$, where $P^{(k)}$ is the matrix composed of the columns containing the assigned 0's in the k th row with those assigned 0's deleted.

These conditions imply that some pollutants are not contributed by a particular source type. If an investigator does not have a priori information on the position of zeros, then one may start with several candidate positions for zeros and select the one giving the highest posterior probability. Alternatively, we may consider preassigning zeros in the source contribution matrix (the matrix of α 's), which implies that each source is missing on some days [7].

Although it has not been introduced in the receptor modeling literature, the Schwarz criterion (also known as the Bayesian Information Criterion or BIC) has been a popular choice for model selection in other contexts (including latent variable models). By penalizing the likelihood by a function of the number of parameters and the sample size, it obtains a trade-off between the bias introduced by fitting the wrong number of parameters and the precision with which the parameters are estimated. The BIC is, however, ad-hoc in nature because it is a rough approximation to twice the logarithm of the Bayes factor [9], and the choice of the sample size and the number of parameters in BIC is often nontrivial.

We calculate the posterior probabilities of the competing models, which follow easily from the marginal likelihoods. The marginal likelihood is a key quantity for a Bayesian model comparison and accounting for model uncertainty [9,10]. Regardless of its theoretical justification and ease of interpretation, the marginal likelihood has not been widely used due to its computational difficulties. Recently, Markov chain Monte Carlo (MCMC) has proven to be useful in many statistical applications. In particular, Chen [11] and Oh [12] proposed simple methods for estimating marginal likelihoods, and hence, the posterior probabilities by using the MCMC output.

The remaining part of the paper is organized as follows. In Section 2, we restate the model from a

statistical point of view. Section 3 contains estimation of parameters using MCMC within a model. Model comparison is discussed in Section 4. Section 5 presents a simulation study. In Section 6, our method is applied to air pollution data consisting of ambient measurements on PM_{10} (particulate matter with median aerodynamic diameter less than 10 μm) in the Seattle area. Finally, concluding remarks are made in Section 7.

2. The model

Suppose, as in Section 1, that y is a p -dimensional vector of observed variables, and α is a q -dimensional vector of latent variables. Though there could be two different types of models depending on whether α is treated as random or fixed (structural model and functional model, respectively) from a frequentist perspective, it is not essential to differentiate these two models from the Bayesian standpoint, since all the parameters are viewed as random variables. A latent variable model consists of two parts, the prior distribution (the terminology is due to Bartholomew and Knott [2]) of the latent variables and the conditional distribution of the observed variables given the latent variables (which depends on the distribution of the errors). The purpose of the latent variable models is to explain the correlations among the observed variables by a set of q ($< p$) latent variables α . This implies that ε_{ij} and $\varepsilon_{i'j'}$ are independent for $j \neq j'$ in model (1.2) if all the major sources are accounted for. We assumed that in model (1.2), the errors ε_t follow a multivariate normal distribution with a mean vector $\mathbf{0}$ and the diagonal covariance matrix $\Sigma = \text{diag}(\sigma_1^2, \sigma_2^2, \dots, \sigma_p^2)$, i.e.,

$$\varepsilon_t \sim N_p(\mathbf{0}, \Sigma), \quad t = 1, \dots, n. \quad (2.1)$$

This specifies the conditional distribution of the observed variables given the latent variables as $y_t | \alpha_t \sim N_p(\alpha_t P, \Sigma)$.

We now need to specify the prior distribution of the α 's. As noted in Ref. [2], the form of the prior distribution of the latent variables is essentially arbitrary and largely a matter of convention. A Gaussian distribution is commonly chosen for the distribution

of α 's. We assume that the α 's have mean vector ξ and covariance matrix Φ ,

$$\alpha_t \sim N_q(\xi, \Phi). \quad (2.2)$$

Let $\gamma = \alpha - \xi$ and $\mu = \xi P$. Then, Eq. (1.2) can be re-parameterized using the centered latent variable γ as:

$$y_t = \mu + \gamma_t P + \varepsilon_t, \quad t = 1, \dots, n. \quad (2.3)$$

Since any change in the scale of γ can also be absorbed into P , without loss of generality, the γ 's may be assumed to have unit standard deviations.¹ If the factors (source contributions in receptor models) are primarily assumed to be uncorrelated, Φ can be taken as an identity matrix, i.e.,

$$\gamma \sim N_q(\mathbf{0}, \mathbf{I}). \quad (2.4)$$

Equivalently, the model may be written in terms of probability distributions as:

$$y | \gamma \sim N_p(\mu + \gamma P, \Sigma) \quad (2.5)$$

and

$$\gamma \sim N_q(\mathbf{0}, \mathbf{I}).$$

This defines a standard factor analysis model. Note that the model depends on the number of factors (sources) q , which is unknown. Moreover, the model is not identified even for known q , i.e., there are other models (parameterizations) which lead exactly to the same joint distribution for the observed y variables. Translation invariance [13] and rotation (see Ref. [2]) are the major sources of nonidentifiability. To remove translation invariance, P is assumed to be of full-row rank. In this paper, we do not consider the case where P is rank-deficient, which corresponds to a collinearity problem in receptor modeling. We leave that problem as one of the model limitations (both in latent variable models and multivariate receptor models) rather than as model uncertainties.

Rotational indeterminacy of the model can be removed by imposing one of the many different types of identifiability conditions (see Ref. [8]). The question is whether those conditions are realistic in the given context. We consider the type of identifiability conditions given in Section 1, C1–C2, which are

often reasonable assumptions in the receptor modeling context. Even within this scheme, there could be several different choices (when there is no certain prior information on zeros) for positions of zeros in P . Note that each possible combination of q and positions of zeros in P defines a different model.

3. Estimation within a model

In this section, our inferences are made conditionally on the model that resulted from a particular choice of q and the set of zeros in P . It follows from Eqs. (2.4) and (2.5) that:

$$y \sim N_p(\mu, P' P + \Sigma). \quad (3.1)$$

This is an integrated likelihood, which is used when fitting the model by maximum likelihood (with the restrictions on the parameters). Although the maximum likelihood estimate (MLE) of μ can be easily shown to be \bar{x} , there is no explicit formula for the MLEs of P and Σ . A numerical maximization needs to be used.

Bayesian inference is based on the posterior distribution, which is proportional to the product of the likelihood and the priors for the parameters. The term 'likelihood' is ambiguous (it could mean either an integrated likelihood or a conditional likelihood) in the present context. We use the conditional likelihood of $Y = \{y_t, t = 1, \dots, n\}$ given the latent variables $\Gamma = \{\gamma_t, t = 1, \dots, n\}$,

$$f(Y | \dots) = |2\pi\Sigma|^{-\frac{n}{2}} \exp \left\{ -\frac{1}{2} \text{tr} \Sigma^{-1} \times \sum_{t=1}^n (y_t - \mu - \gamma_t P)' (y_t - \mu - \gamma_t P) \right\}, \quad (3.2)$$

for the 'likelihood' where ' $|\dots$ ' denotes conditioning on all other variables. At any rate, it does not make any difference in the posterior distribution whether to include the distribution of the latent variables as a part of the likelihood or as a part of the priors.

We assume independent priors $p(\mu, P, \Sigma, \Gamma) = p(\mu)p(P)p(\Sigma)p(\Gamma)$. The prior distribution of $\Gamma = \{\gamma_t, t = 1, \dots, n\}$ was specified in Section 2 as:

$$p(\Gamma) = |2\pi|^{-\frac{n}{2}} \exp \left\{ -\frac{1}{2} \text{tr} \sum_{t=1}^n \gamma_t' \gamma_t \right\}. \quad (3.3)$$

¹ This is just a way of eliminating scale invariance of factors by a constant multiplication.

For μ , we take a p -variate normal prior

$$p(\mu) \sim N_p(m_0, M_0) \tag{3.4}$$

with a $p \times p$ diagonal covariance matrix M_0 .

For the prior distribution for P , we assume a point mass at zero for the $q \times (q - 1)$ pre-selected elements (for identifiability). Let $\text{vec}P^0$ denote the $(q^2 - q) \times 1$ vector of these elements, and let $\text{vec}P^*$ denote the $(pq - q^2 - q) \times 1$ vector of the remaining elements of P stacked columnwise. We use a truncated normal distribution for $\text{vec}P^*$ to incorporate nonnegativity of the source compositions,

$$\text{vec}P^* \sim N_{pq-q^2-q}(c_0, C_0)\mathbf{I}(\text{vec}P^* \geq 0), \tag{3.5}$$

where c_0 is a $(pq - q^2 - q)$ -dimensional vector and C_0 is a $(pq - q^2 - q) \times (pq - q^2 - q)$ -dimensional diagonal matrix.

For the diagonal elements of Σ , we assume a common inverse gamma prior,

$$\sigma_j^{-2} \sim \Gamma(\alpha_0, \beta_0), \quad j = 1, \dots, p, \tag{3.6}$$

with the parameterization in which the mean and variance are α_0/β_0 and α_0/β_0^2 , respectively.

From Eqs. (3.2)–(3.6), the joint posterior distribution for (μ, P, Σ, Γ) , is given by

$$\begin{aligned} &\pi(\mu, P, \Sigma, \Gamma | Y \alpha) f(Y | \dots) p(\mu, P, \Sigma, \Gamma) \\ &= |2\pi\Sigma|^{-\frac{n}{2}} \exp\left\{-\frac{1}{2} \text{tr}\Sigma^{-1} \sum_{t=1}^n (y_t - \mu - \gamma_t P)'\right. \\ &\quad \times \left. (y_t - \mu - \gamma_t P)\right\} |2\pi|^{-\frac{n}{2}} \exp\left\{-\frac{1}{2} \text{tr} \sum_{t=1}^n \gamma_t' \gamma_t\right\} \\ &\quad \times |2\pi M_0|^{-\frac{1}{2}} \exp\left\{-\frac{1}{2}(\mu - m_0)\right. \\ &\quad \times \left. M_0^{-1}(\mu - m_0)'\right\} \\ &\quad \times |2\pi C_0|^{-\frac{1}{2}} \exp\left\{-\frac{1}{2}(\text{vec}P^* - c_0)'\right. \\ &\quad \times \left. C_0^{-1}(\text{vec}P^* - c_0)\right\} \mathbf{I}(\text{vec}P^* \geq 0) \mathbf{I}(\text{vec}P^0 = 0) \\ &\quad \times \prod_{j=1}^p \frac{\beta_0^{\alpha_0}}{\Gamma(\alpha_0)} \left(\frac{1}{\sigma_j^2}\right)^{\alpha_0+1} \exp\left(-\frac{\beta_0}{\sigma_j^2}\right). \tag{3.7} \end{aligned}$$

Posterior inferences on the parameters require high-dimensional integration of the joint posterior density. Obviously, the integrals are analytically intractable in this case, and a direct simulation from this density is not possible either due to complexity of Eq. (3.7). We therefore employ a Markov chain Monte Carlo (MCMC) approach (see, e.g., Refs. [14,15]). In particular, we use the Gibbs sampling algorithm by Gelfand and Smith [16] since all of the full conditional distributions can be easily obtained. In our Gibbs sampling algorithm, one sweep consists of four updating procedures: updating μ , updating P , updating Σ , updating γ . Now we give details of each updating procedure.

3.1. Updating μ

The full conditional posterior distribution of μ is given by:

$$\mu | \dots \sim N_p(m, M),$$

where $M = (n\Sigma^{-1} + M_0^{-1})^{-1}$ and $m = \{n(\bar{y} - \bar{\gamma}P)\Sigma^{-1} + m_0 M_0^{-1}\}M$. Sample generation from the distribution is straightforward.

3.2. Updating P

Under the truncated normal prior Eq. (3.5), the full conditional posterior distribution $\pi(\text{vec}P^* | \dots)$ is again a truncated normal distribution. Due to high-dimensionality of $\text{vec}P^*$, it is much more efficient to sample a sub-vector of $\text{vec}P^*$ that corresponds to each column of P (after deleting zero elements) rather than sampling the entire vector $\text{vec}P^*$. Let P_j^* be the j th sub-vector of $\text{vec}P^*$ that corresponds to the j th column of P (after deleting zero elements if there is any). For the columns of P with no zero elements, we have:

$$P_j^* | \dots \sim N_q(c_j, C_j) \cdot \mathbf{I}(P_{kj} \geq 0, \quad k = 1, \dots, q),$$

where $c_j = C_j\{\sigma_j^{-2}\Gamma'(Y_j - \mu_j 1_n) + C_{0j}^{-1}c_{0j}\}$, $C_j = (\sigma_j^{-2}\Gamma'\Gamma + C_{0j}^{-1})^{-1}$, c_{0j} is a q -dimensional prior mean vector of P_j^* , C_{0j} is a corresponding submatrix of C_0 , and Y_j is the j th column of Y . For the columns of P containing zero elements, let q^* be the number of nonzero elements for that column. Then,

$$P_j^* | \dots \sim N_{q^*}(c_j^*, C_j^*) \cdot \mathbf{I}(P_{kj} \geq 0, \quad k = 1, \dots, q),$$

where $C_j^* = (\sigma_j^{-2} \Gamma_j^{*'} \Gamma_j^* + C_{0j}^{*-1})^{-1}$, $c_j^* = C_j^* \Gamma_j^{*'} \{ \sigma_j^{-2} \Gamma_j^{*'} (Y_j - \mu_j 1_n) + C_{0j}^{*-1} c_{0j}^* \}$, c_{0j}^* is a q^* -dimensional prior mean vector of P_j^* , C_{0j}^* is a corresponding submatrix of C_0 , and Γ_j^* consists of the columns of Γ corresponding to nonzero elements of the j th column of P . Sample generation from the truncated multivariate normal distribution can be done by rejection sampling or Metropolis–Hastings algorithm (see Ref. [14]) or by applying the Gibbs sampler for each element of P_j^* .

3.3. Updating Σ

The full conditional posterior distribution of the j th diagonal element σ_j^2 of Σ is given by:

$$\sigma_j^{-2} | \dots \sim \text{Gamma} \left(\alpha_0 + \frac{1}{2}n, \beta_0 + \frac{1}{2}d_j \right),$$

where d_j is the j th diagonal element of $d = (Y - 1_n \otimes \mu - \Gamma P)' (Y - 1_n \otimes \mu - \Gamma P)$, and sample generation is easy.

3.4. Updating Γ

It can easily be shown that the full conditional posterior distribution of the t th row γ_t of Γ is given by:

$$\gamma_t | \dots \sim N_q(b_t, B)$$

where $B = (P \Sigma^{-1} P' + I_q)^{-1}$, $b_t = (y_t - \mu) \Sigma^{-1} P' B$. Again, sample generation is straightforward.

4. Model comparison

Assume that there are G candidate models. Under the g th model,

$$M_g : y = \mu_g + \gamma_g P_g + \varepsilon, \quad \varepsilon \sim N_p(\mathbf{0}, \Sigma_g),$$

$$g = 1, \dots, G.$$

Here, each model comes from different combinations of the number of sources q and identifiability conditions. When there are G competing models, a typical Bayesian model selection procedure computes the posterior model probability, $P(M_g|Y)$, of model M_g given the data Y , for each $g = 1; \dots, G$, and then selects the model with the highest posterior model probability.

From a basic probability law, the posterior model probability $P(M_g|Y)$ is given by:

$$P(M_g | Y) \propto l(Y | M_g) p(M_g),$$

where $p(M_g)$ is the prior probability of model M_g . The prior $p(M_g)$ is often chosen to be uniform so as not to favor one model over another a priori. Under the indifference model prior probabilities, the posterior model probability is proportional to $l(Y|M_g)$. The quantity $l(Y|M_g)$ is called *the marginal likelihood* or *integrated likelihood* of model M_g which is given by, again from the basic probability law,

$$l(Y | M_g) = \int l(Y | \theta_g, M_g) p(\theta_g | M_g) d\theta_g, \quad (4.1)$$

where θ_g is the vector of unknown parameters in model M_g , $l(Y|\theta_g, M_g)$ is the likelihood of θ_g under model M_g , and $p(\theta_g|M_g)$ is the prior of θ_g under model M_g .

In latent variable models (multivariate receptor models), however, Eq. (4.1) is not given in a closed form and a numerical approximation is necessary. Among many methods for approximating the marginal likelihood, the method proposed by Oh [12] can be easily implemented here. From the relation:

$$\pi(\theta_g | Y, M_g) = \frac{l(Y | \theta_g, M_g) p(\theta_g | M_g)}{l(Y | M_g)},$$

one can estimate the marginal likelihood of model M_g by

$$\hat{l}(Y | M_g) = \frac{l(Y | \theta_g^*, M_g) p(\theta_g^* | M_g)}{\hat{\pi}(\theta_g^* | Y, M_g)}, \quad (4.2)$$

where θ_g^* is a point of θ_g and $\hat{\pi}(\theta_g^* | Y, M_g)$ is the estimated posterior density function of θ_g given Y under model M_g . Thus, one only needs to obtain $\hat{\pi}(\theta_g^* | Y, M_g)$, for each $g = 1; \dots, G$.

Now we give a brief description of Oh's method for $\hat{\pi}(\theta_g^* | Y, M_g)$. For simplicity, we suppress the index g for the rest of the section. Let $\theta = (\theta_1; \dots, \theta_m)$ where θ_i is the i th block of θ which can be an element or a vector of elements. Oh [12] showed that

$$\begin{aligned} \pi(\theta^* | Y, M) &= E[\pi(\theta_1^* | \theta_2^*, \dots, \theta_m^*) \\ &\quad \times \pi(\theta_2^* | \theta_1, \theta_3^*, \dots, \theta_m^*) \cdots \\ &\quad \times \pi(\theta_3^* | \theta_1, \dots, \theta_{m-1})], \end{aligned} \quad (4.3)$$

where the expectation is with respect to the joint distribution of θ under model M , and hence, it can be estimated by the sample average of the product of the full conditional posterior density functions, using the posterior sample of θ under model M . Great advantages of the method are that estimation of $\pi(\theta^*|Y, M)$ can be done during the routine MCMC simulation without generating additional samples, and that it can be very easily implemented when all the full conditional posterior density functions are known. In theory, the point θ^* can be arbitrary. For efficiency, however, θ^* should be chosen from the region with high posterior density. An approximate mode of θ , which can be obtained from a preliminary MCMC run, would be a reasonable choice for θ^* .

For the standard factor analysis models with restrictions on P , we can apply the method with μ , Γ , Σ , and each nonzero element of P as blocks of θ . Note that the full conditional posterior distribution of μ , Γ , and the diagonal elements of Σ are multivariate normal, multivariate normal, and Inverse Gamma, respectively, and that of any nonzero element of P is a univariate truncated normal distribution. Thus, all the necessary full conditional posterior density functions for Eq. (4.3) are given, and estimation of $\pi(\mu^*, \Gamma^*, P^*, \Sigma^* | Y, M)$ is straightforward.

5. Simulation

5.1. Application to simulated data

The first data sets are generated as follows: the sample size n is taken to be 100, the number of variables p is 9, and the true number of sources q_0 is 3. The true model has the factor loading matrix (source composition matrix),

$$P = \begin{bmatrix} 0.10 & 0 & 0 & 0.99 & 0.25 & 0.05 & 0.05 & 0.05 & 0.50 \\ 0 & 0.35 & 0 & 0.05 & 0.05 & 0.95 & 0.60 & 0.05 & 0.50 \\ 0.70 & 0 & 0.50 & 0 & 0.50 & 0.05 & 0.90 & 0.90 & 0.30 \end{bmatrix},$$

and the overall mean $\mu = 5 \cdot \mathbf{1}_p$ where $\mathbf{1}_p$ is a p -dimensional row vector of $\mathbf{1}$'s. The factors are generated randomly and independently from $\gamma_t \sim N(\mathbf{0}, \mathbf{I}_3)$, $t = 1, \dots, n$, and the errors are generated randomly

and independently from $\varepsilon_t \sim N(\mathbf{0}, \Sigma)$, $t = 1, \dots, n$ where

$$\Sigma = \text{diag}(0.03, 0.02, 0.03, 0.02, 0.01, 0.04, 0.02, 0.03, 0.03),$$

which results in approximately 13–34% of the error standard deviations to the model standard deviations. Then the y 's are obtained using Eq. (2.3),

$$y_t = \mu + \gamma_t P + \varepsilon_t \quad t = 1, \dots, n.$$

In our simulation, the candidate models may be defined by varying the number of factors (q) and the identifiability conditions (the position of zeros). Recall that under the indifference prior model probabilities, the posterior probability, $P(M_g|Y)$, of model M_g is proportional to the marginal likelihood, $l(Y|M_g)$, of model M_g . Thus, we only need to calculate the marginal likelihood of each model for model comparison. For simplicity of presentation, we first change the number of factors ($q = 1, 2, 3, 4, 5$) with the most plausible identifiability conditions for each q -factor model. Note that there may possibly be confounding effects between the number of factors and the identifiability conditions on the marginal likelihoods. For the purpose of model selection, it does not matter as our interest is to see whether the estimated marginal likelihood is the highest for the true model.

The simulation is repeated 50 times. Throughout the simulation, the values for P , μ , Γ , and Σ remain the same as given above, and only the errors are regenerated to obtain the observations at each simulation. The following hyperparameter values are used for generating MCMC samples: $\alpha_0 = 2$, $\beta_0 = 1$ for Σ , $m_0 = 5 \cdot \mathbf{1}_p$, $M_0 = 100 \cdot \mathbf{I}_p$ for μ , and $c_0 = 0.5$, $C_0 = 10$ for nonzero elements of P , which yield vague priors. The estimated marginal likelihoods for each q -factor model are reported in Table 1 on a log scale (only 10 cases are shown for space). Recall that, with indifference prior for competing models, the posterior model probability is proportional to the marginal likelihood.

We also calculate BIC for each model. The BIC is defined as:

$$\text{BIC} = -2 \log(\text{max likelihood}) + (\log N)(\text{number of parameters}). \quad (5.1)$$

Table 1

Log of marginal likelihood of q (within an additive constant) and BIC for $q=1, 2, 3, 4, 5$ ($q_0=3$), $n=100, p=9$

Methods	Data set	Number of factors (q)					Selected number of factors
		1	2	3	4	5	
LogMD	1	-773.26	-598.95	-541.01	-615.21	-655.05	3
	2	-757.36	-606.81	-538.85	-616.58	-653.20	3
	3	-763.48	-604.69	-550.82	-626.49	-672.71	3
	4	-788.20	-773.84	-564.68	-641.31	-671.84	3
	5	-764.11	-598.38	-540.38	-612.81	-663.31	3
	6	-777.91	-769.70	-553.47	-621.07	-664.61	3
	7	-770.69	-607.16	-548.04	-638.96	-674.52	3
	8	-778.43	-604.55	-539.08	-605.47	-662.29	3
	9	-757.52	-589.23	-536.48	-612.77	-642.69	3
	10	-774.98	-605.71	-553.10	-641.27	-656.85	3
BIC	1	1427.85	911.93	711.35	891.45	1028.76	3
	2	1395.07	935.40	683.13	901.21	1013.56	3
	3	1401.83	932.98	709.26	932.55	1051.33	3
	4	1464.57	1339.75	752.60	962.42	1064.17	3
	5	1408.79	907.77	685.75	882.71	1005.87	3
	6	1437.50	1323.34	725.21	919.26	1054.64	3
	7	1418.25	913.23	731.88	948.16	1045.54	3
	8	1433.06	928.37	696.14	885.26	1012.51	3
	9	1390.03	895.20	669.03	890.62	993.47	3
	10	1434.51	931.58	725.75	939.91	1053.72	3

For max likelihood in Eq. (5.1), we use the integrated likelihood in Eq. (3.1), with MLE for (μ, P, Σ) plugged in. It is well known that the MLE for (μ, P, γ, Σ) , based on the conditional likelihood in Eq. (3.2), do not exist (see, e.g., Ref. [8]), and as mentioned in Section 3.1, even the MLE based on the integrated likelihood requires the use of some sort of an iterative procedure or an EM algorithm. Here, we can easily obtain the approximate MLE for (μ, P, Σ) by directly evaluating the integrated likelihood function using the posterior samples generated from MCMC. Note, however, that the number of observations, N , and the number of parameters in Eq. (5.1) are often not clearly defined. We use $N=n$, and

$$\begin{aligned}
 (\text{number of parameters}) &= pq + p + p - q(q - 1) \\
 &= p(q + 2) - q(q - 1),
 \end{aligned}$$

which is the number of free parameters in the integrated likelihood. The calculated BIC for each model is also reported in Table 1.

Table 2 summarizes the performance of each method. The method based on the marginal likelihood chooses q having the maximum logMD (log of marginal likelihood) and BIC chooses q having the minimum BIC. Both methods select the true model ($q=3$) for all of 50 simulations. We also monitor the R^2 values between the true factor loadings and the estimated factor loadings for $q=3$. Throughout the simulation, R^2 values are all close to 0.99, which indicates that the estimated loadings agree well with the true loadings once the true model is selected.

Table 2
Comparison of model uncertainty assessment methods based on 50 simulated data sets, $n=100, p=9, q_0=3$

Method	q				
	$q=1$	$q=2$	$q=3$	$q=4$	$q=5$
LogMD	0	0	50	0	0
BIC	0	0	50	0	0

Table 3
Log of marginal likelihood of q (within an additive constant) and BIC for $q=3, 4, 5, 6, (q_0=5), n=100, p=9$

Methods	Data set	Number of factors (q)				Selected number of factors
		3	4	5	6	
LogMD	1	-832.48	-751.09	-730.51	-788.95	5
	2	-889.30	-745.87	-738.74	-806.37	5
	3	-826.35	-743.81	-732.03	-785.78	5
	4	-832.21	-737.14	-729.93	-784.59	5
	5	-831.22	-733.67	-721.69	-772.05	5
	6	-816.78	-725.58	-723.00	-776.74	5
	7	-826.81	-736.83	-733.27	-792.06	5
	8	-836.27	-747.40	-729.22	-781.81	5
	9	-830.81	-731.77	-718.95	-783.63	5
	10	-839.08	-742.46	-739.87	-786.85	5
BIC	1	1511.62	1279.24	1235.38	1393.11	5
	2	1609.07	1294.33	1233.03	1416.47	5
	3	1504.86	1287.04	1216.44	1404.25	5
	4	1509.25	1251.24	1203.21	1386.09	5
	5	1505.54	1251.32	1174.42	1370.92	5
	6	1479.15	1244.52	1197.39	1359.30	5
	7	1487.11	1273.86	1205.45	1380.78	5
	8	1523.58	1282.80	1211.89	1395.95	5
	9	1489.98	1241.19	1195.51	1364.05	5
	10	1525.09	1274.42	1231.22	1404.69	5

Secondly, we consider the case where the true number of factors (q_0) is 5. The factor loading matrix is given as follows:

$$P = \begin{bmatrix} 0.10 & 0 & 0 & 0.99 & 0.25 & 0.0 & 0.0 & 0.05 & 0.50 \\ 0 & 0.35 & 0 & 0 & 0 & 0.95 & 0.60 & 0.05 & 0.50 \\ 0.70 & 0 & 0.50 & 0 & 0 & 0 & 0.90 & 0.90 & 0.30 \\ 0.10 & 0 & 0.80 & 0.10 & 0.20 & 0 & 0 & 0 & 0.90 \\ 0.10 & 0.05 & 0.05 & 0 & 0 & 0.70 & 0 & 0 & 0.40 \end{bmatrix}$$

For μ and Σ , the same values as in the 3-factor model case are used, i.e., $\mu = 5 \cdot \mathbf{1}_p$ and $\Sigma = \text{diag}(0.01, 0.05,$

$0.03, 0.01, 0.01, 0.08, 0.06, 0.08, 0.08)$. We use the following hyperparameter values in generating MCMC samples: $\alpha_0 = 2, \beta_0 = 0.5$ for $\Sigma, m_0 = 5 \cdot \mathbf{1}_p, M_0 = 100 \cdot \mathbf{1}_p$ for μ , and $c_0 = 0.5, C_0 = 2$ for nonzero elements of P , which yield vague priors. The marginal likelihood and BIC of each model with the number of factors ($q = 3, 4, 5, 6$) are reported in Table 3. Both methods perform well in choosing the true model ($q = 5$) as can be seen in Table 4. The

Table 4
Comparison of model uncertainty assessment methods based on 50 simulated data, $n = 100, p = 9, q_0 = 5$

Method	q			
	$q=3$	$q=4$	$q=5$	$q=6$
LogMD	0	0	50	0
BIC	0	0	50	0

Table 5
Hyperparameter specifications, $m_0 = a \cdot \mathbf{1}_p, M_0 = b \cdot \mathbf{1}_p, c_0 = c \cdot \mathbf{1}_{pq-q^2-q}, C_0 = d \cdot \mathbf{1}_{pq-q^2-q}$

	α_0	β_0	a	b	c	d
I	2	0.1	5	100	0.5	100
II	2	0.01	5	100	0.5	100
III	2	1	5	100	0.5	100
IV	2	1	5	100	1	100
V	2	1	5	100	0	100
VI	2	1	5	10	0	100
VII	2	1	5	10	0	10
VIII	2	1	5	1	0	10
VI	2	0.1	5	1	1	10
X	2	0.5	5	5	0.5	3

R^2 values between the true P and the estimated P (with $q=5$) are also close to 0.99 throughout the simulation.

5.2. Sensitivity analysis and robustness

It is known that the marginal likelihoods might be sensitive to the priors. To make sure that our analysis results do not change with prior specification, a sensitivity analysis is carried out with a range of different priors. The data are generated from the 3-factor model given in Section 5.1. We try 10 different sets of hyperparameter values given in Table 5. The log of marginal likelihoods for each set of hyperparameters are shown in Table 6. Although the values change with hyperparameter specification, the overall pattern of them is consistent (showing the maximum at $q=3$), and so our decision is not affected. Also, estimates for μ , P , and Σ under the chosen model (3-factor model) show only negligible changes.

It was mentioned in Section 2 that the form of prior distribution of γ is largely a matter of convention. To ascertain that the method is robust to misspecification of the prior distribution of γ , we first simulate $n=100$ observations from the $q=3$ factor model with the same values for μ , P , and Σ as in Section 5.1 but different distributions for γ : truncated normal distribution and lognormal distribution. The simulation is repeated 10 times for each case. Our method (with a standard normal prior on γ) chooses the correct model in all simulations (and parameter estimates under the selected model are all close to the true values).

The method is also applied to the generated data using correlated factors. We simulate $n=200$ obser-

vations from a $q=3$ factor model defined by parameters:

$$P = \begin{bmatrix} 0.90 & 0 & 0 & 0.99 & 0.25 & 0.05 & 0.05 & 0.05 & 0.50 \\ 0 & 0.90 & 0 & 0.05 & 0.05 & 0.95 & 0.60 & 0.05 & 0.50 \\ 0.70 & 0 & 0.90 & 0 & 0.50 & 0.05 & 0.90 & 0.90 & 0.30 \end{bmatrix},$$

$\mu = 5 \cdot \mathbf{1}_p$, $\Sigma = \text{diag}(0.015, 0.01, 0.015, 0.01, 0.005, 0.02, 0.01, 0.015, 0.015)$, and $\gamma \sim N(\mathbf{0}, \Phi)$ where,

$$\Phi = \begin{bmatrix} 1 & 0.7 & 0.7 \\ 0.7 & 1 & 0.7 \\ 0.7 & 0.7 & 1 \end{bmatrix}.$$

Simulation is repeated 10 times, and again, our method (with a standard normal prior on γ) chooses the correct model 100% of time, and the estimated parameters under the chosen model are all close to the true values. The sample correlation matrix of the estimated Γ is given as:

$$R_\Gamma = \begin{bmatrix} 1 & 0.69 & 0.67 \\ 0.69 & 1 & 0.72 \\ 0.67 & 0.72 & 1 \end{bmatrix},$$

which resembles the true Φ . Note that maximum likelihood estimation, based on the integrated likelihood Eq. (3.1), would not be able to find this correlated factor structure because Γ is integrated out using the assumption that $\gamma \sim N_q(\mathbf{0}, \mathbf{I})$. From a Bayesian standpoint, Φ can be viewed as a hyperparameter of the prior distribution for factors rather than the underlying assumption in the model. Although $\Phi = \mathbf{I}$ is misspecified for these data, the correlation structure in γ may be uncovered by estimated γ 's. Finally, we look at the case when γ is generated from the lognormal distribution with correlated factors, i.e., $\log \gamma \sim N_q(\mathbf{0}, \Phi)$. Again, the method shows robustness to violations of both assumptions (the distributional form and the correlation structure in γ).

6. Analysis of Seattle PM₁₀ data

We apply our methods to PM₁₀ data obtained from 10 monitoring sites in the Seattle area during 1992–1996. The monitoring sites are from north to

Table 6
LogMD for each set of hyperparameters

	$q=1$	$q=2$	$q=3$	$q=4$	$q=5$
I	-765.07	-509.52	-384.66	-459.73	-541.98
II	-792.71	-536.67	-384.74	-505.69	-556.59
III	-783.67	-620.87	-555.55	-625.16	-697.40
IV	-784.85	-615.78	-557.25	-611.85	-662.50
V	-783.37	-621.28	-559.97	-617.08	-681.92
VI	-772.00	-604.99	-553.19	-623.27	-663.01
VII	-762.51	-588.93	-518.87	-599.48	-635.68
VIII	-752.39	-584.65	-515.89	-591.27	-625.70
VI	-735.87	-472.31	-338.70	-453.85	-478.37
X	-736.35	-698.58	-431.93	-486.52	-530.44

south (see the map in Fig. 1): Marysville, Everett, Lake Forest Park, Harbor Island, Duwamish, South Park, Kent, two sites in Tacoma (one from a resi-

dential area, Tacoma-R, and the other from an industrial area, Tacoma-I), and Puyallup. At most of the monitoring sites, PM_{10} was measured only

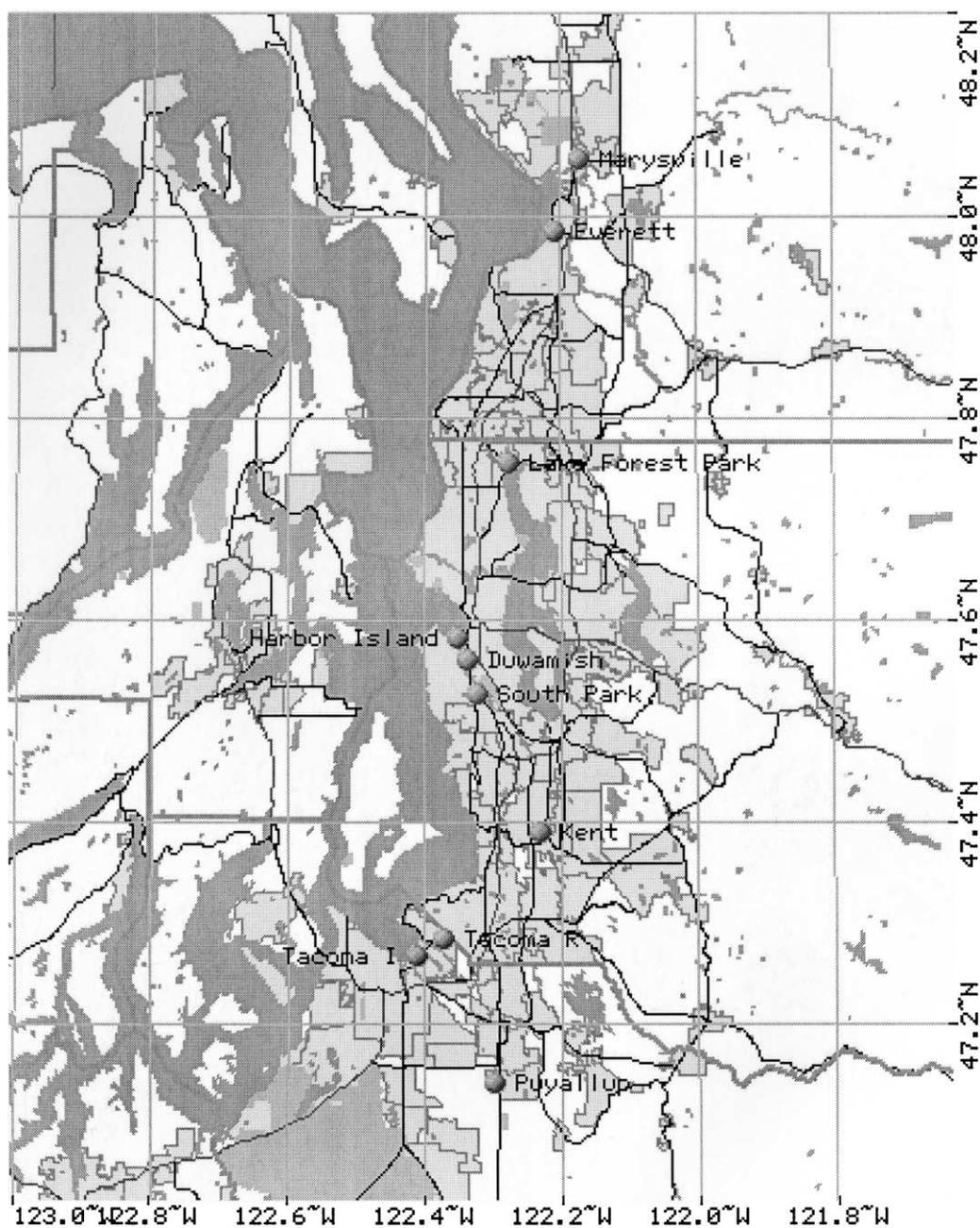


Fig. 1. PM_{10} monitoring sites in Greater Seattle.

every 6 days (as 24-h average concentrations), so we used those 6-day measurements for the analysis. The goal is to identify major sources (source regions) of PM_{10} . Here, the 10 monitoring sites play the role of different variables in our basic multivariate receptor model. The source profile, consisting of the relative amounts of PM_{10} that are conveyed to the 10 monitoring sites in this case, represents the spatial pattern of underlying PM_{10} concentration from each source. These source profiles (spatial profiles) were used in Ref. [7] to locate the major source regions in the Grand Canyon. The underlying assumptions for this approach are:

A1. There are a few underlying spatial patterns (P), and they do not vary with time.

A2. The environmental factors such as wind do not interact with P , i.e., the overall spatial wind flow patterns (on which the spatial source patterns depend) are approximately constant.

It is suspected that in Seattle, there might be some changes in the source regions between the dry season (July to September, referred to as ‘Summer’ hereafter) and the wet season (October to March, referred to as ‘Winter’ hereafter). The PM_{10} level is higher on average during Winter than during Summer (the difference is as big as $10 \mu\text{g}/\text{m}^3$ for some sites such as Marysville, Lake Forest Park, Duwamish, and Tacoma-I, see Table 7). Table 7 also shows that there is a big variation in the PM_{10} level during Winter. It is of interest to determine if there is an additional (major) source during Winter in Seattle.

We analyze the data separately for Summer and Winter to deal with seasonal variation. For each season, the assumption A1 seems to be justified. Also, it is unlikely that there is a significant change in the major source regions within each season during the 5 years of observation (no big point source was added to or removed from this area during that period, Ref. [17]). The methods can be extended to account for violation of A1 by using dynamically varying mean μ_t and source profiles P_t , but this is beyond the scope of this paper.

Analyzing the data separately for each season also has an advantage of coping with the dependence of the regional pattern of PM_{10} concentration on shifting wind patterns. It is known that the main variability of regional wind pattern is between seasons, and it is fairly constant in any given season of the year. For Seattle, the prevailing wind direction is southerly in Winter and northerly in Summer, which is almost aligned with the monitoring sites. This also justifies A2.

After deletion of missing values, 66 observations are retained for Summer, and 129 for Winter. For q , we try the values $q=1, 2, 3, 4$, for each season. As noted in Section 2, each possible combination of q and positions of zeros in the source composition matrix P yields a different model. For candidate positions of zeros in P , we use the results from UNMIX [18] rather than going through an infinite number of possible combinations. For each q ($q=1, 2, 3, 4$), we first obtain UNMIX source composition matrix P_{UNMIX} , and try the elements giving the low

Table 7
Sample mean and standard deviation of PM_{10} data for each season

	Summer ($n=66$)		Winter ($n=129$)	
	Sample mean (in $\mu\text{g}/\text{m}^3$)	Sample standard deviation	Sample mean (in $\mu\text{g}/\text{m}^3$)	Sample standard deviation
Marysville	19.61	7.23	28.16	19.18
Everett	19.82	6.78	21.76	12.20
Lake Forest Park	17.30	5.11	30.09	16.85
Harbor Island	26.03	9.32	33.41	15.72
Duwamish	27.58	10.22	36.43	17.89
South Park	22.50	9.85	28.10	15.92
Kent	26.67	9.86	28.56	16.57
Tacoma-R	25.41	12.36	25.47	16.84
Tacoma-I	25.85	10.21	36.48	20.31
Puyallup	22.82	12.42	27.03	18.02

proportions in P_{UNMIX} as the candidate zeros ($q - 1$ zeros for each row). Note that UNMIX profiles are used only to find out the plausible sets of identifiability conditions under each q -source model. Other than that, the candidate models do not depend on the UNMIX analysis. It is possible to try different sets of zeros within each q . For Summer, we come up with a total of 11 candidate models, and 15 candidate models for Winter (see Tables 8 and 9).

Our MCMC analysis, conducted separately for Summer and Winter, uses the following hyperparameters for the prior distributions. For μ , $m_0 = 20 \cdot \mathbf{1}_p$ for Summer and $m_0 = 30 \cdot \mathbf{1}_p$ for Winter are used since it is a priori expected that the mean concentration of PM_{10} would be much higher in Winter than in Summer. For Σ , $\beta_0 = 10$ for Summer and $\beta_0 = 20$ for Winter due to the similar reason as before. For all other hyperparameters, we use the same values for Summer and

for Winter: $c_0 = 5$, $C_0 = 100$ for nonzero elements of P , $\alpha_0 = 2$, and $M_0 = 100 \cdot \mathbf{I}_p$ for all models compared. For each model, an approximate posterior mode is obtained from a preliminary MCMC run, and this is used for $\theta^* = (\mu^*, P^*, \Sigma^*, \Gamma^*)$ at which the marginal likelihood is calculated. For the preliminary MCMC run, the iterations are started from \bar{Y} for μ , a uniform random matrix with zeros preassigned for P , and $\text{diag}(s_1^2; \dots; s_p^2)/20$, where s_j^2 is the sample variance of the j th column of Y for Σ . An approximate posterior mode is obtained by evaluating the joint posterior density for 20,000 iterations after the first 10,000 draws are discarded. A main MCMC is then started from $\theta^* = (\mu^*, P^*, \Sigma^*, \Gamma^*)$, and the samples are collected for 30,000 iterations without additional burn-in. The marginal likelihood for each model is calculated in sample generation without storing the samples.

Table 8
Candidate models, logMD, and BIC for Seattle PM_{10} data, Summer (7–9)

Model number	q	Position of zeros in P	LogMD	BIC
1	1	none	-2056.15 (0.99996)	4090.11
2	2	profile 1: Marysville profile 2: Lake Forest Park	-2070.16 (8.19×10^{-7})	4145.50
3	2	profile 1: Everett profile 2: Puyallup	-2068.54 (4.13×10^{-6})	4107.30
4	2	profile 1: Marysville profile 2: Puyallup	-2066.50 (3.19×10^{-5})	4104.61
5	2	profile 1: Lake Forest Park profile 2: Puyallup	-2071.12 (3.14×10^{-7})	4108.60
6	2	profile 1: Kent profile 2: Puyallup	-2073.85 (2.05×10^{-8})	4111.91
7	3	profile 1: Lake Forest Park, Puyallup profile 2: Lake Forest Park, Tacoma-R profile 3: Tacoma-R, Puyallup	-2095.09 (1.22×10^{-17})	4165.39
8	3	profile 1: Lake Forest Park, Puyallup profile 2: Marysville, Everett profile 3: Tacoma-R, Puyallup	-2072.65 (6.82×10^{-8})	4139.74
9	3	profile 1: Marysville, Puyallup profile 2: Everett, Lake Forest Park profile 3: Marysville, Lake Forest Park	-2100.92 (3.58×10^{-20})	4194.48
10	4	profile 1: South Park, Tacoma-R, Puyallup profile 2: Marysville, Everett, Lake Forest Park profile 3: Marysville, Everett, Tacoma-R profile 4: Harbor Island, South Park, Tacoma I	-2075.48 (4.01×10^{-9})	4150.55
11	4	profile 1: Everett, Lake Forest Park, Puyallup profile 2: Marysville, Everett, Tacoma-R profile 3: Harbor Island, Duwamish, Tacoma-I profile 4: South Park, Tacoma-R, Puyallup	-2087.20 (3.28×10^{-14})	4138.40

The posterior probability of each model under the indifference prior is given in parenthesis after logMD.

Table 9
Candidate models, logMD and BIC for Seattle PM₁₀ data, Winter (10–3)

Model number	q	Position of zeros in P	LogMD	BIC
1	1	none	–4729.95 (1.06×10^{-17})	9436.88
2	2	profile 1: Lake Forest Park profile 2: Tacoma-R	–4711.14 (1.56×10^{-9})	9440.90
3	2	profile 1: Everett profile 2: Tacoma-R	–4728.38 (5.07×10^{-17})	9447.52
4	2	profile 1: Marysville profile 2: Puyallup	–4747.12 (3.69×10^{-25})	9504.62
5	2	profile 1: Marysville profile 2: Tacoma-R	–4720.32 (1.61×10^{-13})	9424.38
6	3	profile 1: Harbor Island, Tacoma-R profile 2: Marysville, Everett profile 3: Lake Forest Park, Puyallup	–4690.86 (0.99986)	9388.25
7	3	profile 1: South Park, Tacoma-R profile 2: Marysville, Everett profile 3: Lake Forest Park, Puyallup	–4711.07 (1.67×10^{-9})	9401.38
8	3	profile 1: Tacoma-R, Tacoma-I profile 2: Marysville, Everett profile 3: Marysville, Puyallup	–4707.11 (8.76×10^{-8})	9411.95
9	3	profile 1: Harbor Island, Tacoma-R profile 2: Marysville, Everett profile 3: Marysville, Lake Forest Park	–4699.90 (0.00012)	9405.67
10	3	profile 1: Harbor Island, Tacoma-R profile 2: Marysville, Everett profile 3: Marysville, Puyallup	–4701.51 (2.37×10^{-5})	9405.93
11	3	profile 1: South Park, Tacoma-R profile 2: Marysville, Everett profile 3: Marysville, Puyallup	–4709.93 (5.22×10^{-9})	9415.81
12	3	profile 1: South Park, Tacoma-R profile 2: Marysville, Everett profile 3: Marysville, Lake Forest Park	–4707.97 (3.71×10^{-8})	9419.13
13	4	profile 1: Harbor Island, South Park, Tacoma-R profile 2: Marysville, Everett, Tacoma-R profile 3: Lake Forest Park, Harbor Island, Duwamish profile 4: Marysville, Tacoma-R, Puyallup	–4704.14 (1.71×10^{-6})	9443.66
14	4	profile 1: Harbor Island, Kent, Tacoma-I profile 2: Marysville, Tacoma-R, Puyallup profile 3: Lake Forest Park, Harbor Island, Duwamish profile 4: Marysville, Everett, Tacoma-R	–4716.94 (4.72×10^{-12})	9434.19
15	4	profile 1: Kent, Tacoma-R, Tacoma-I profile 2: Marysville, Tacoma-R, Puyallup profile 3: Lake Forest Park, Harbor Island, Duwamish profile 4: Marysville, Everett, Tacoma-R	–4710.43 (3.17×10^{-9})	9440.40

The posterior probability of each model under the indifference prior is given in parenthesis after logMD.

Tables 8 and 9 contain the estimated marginal likelihood (in log) and BIC for each model for Summer and Winter, respectively. The posterior probability of each model under the indifference prior is also given in the parenthesis at the bottom of estimated marginal likelihood. For Summer, both the

marginal likelihood criterion and BIC select Model 1, which corresponds to 1-source model. For Winter, Model 6, corresponding to 3-source model, is selected as the best model based on both criteria. This is consistent with our expectation that there would be additional pollution source/sources during Winter.

Table 10
Posterior summary for the parameters P , μ , and Σ of Model 1, Summer

	P		μ		Σ	
	Mean	Standard deviation	Mean	Standard deviation	Mean	Standard deviation
Marysville	6.96	0.75	19.24	0.83	9.87	1.99
Everett	6.48	0.71	19.47	0.79	9.39	1.85
Lake Forest Park	4.76	0.54	17.05	0.60	6.49	1.22
Harbor Island	9.20	0.94	25.53	1.08	12.66	2.64
Duwamish	10.08	1.03	27.03	1.18	15.28	3.18
South Park	9.81	0.99	21.97	1.13	12.61	2.66
Kent	9.60	1.01	26.14	1.14	16.35	3.33
Tacoma-R	11.48	1.29	24.77	1.43	36.20	6.85
Tacoma-I	9.89	1.05	25.31	1.17	18.22	3.69
Puyallup	8.46	1.46	22.35	1.47	89.49	16.05

We report some of the posterior summaries for parameters μ , P , and Σ , in Tables 10 and 11 for the best model selected for each season. Posterior intervals and simultaneous posterior regions for the parameters can also be easily constructed based on the posterior samples, though we do not report those results here due to limited space (see, e.g., Ref. [19]). Note that the estimates for P are with reference to scaling (normalization) that makes the corresponding source contributions (γ) have unit variances. Because we are using a sampling-based method, we can easily obtain the estimates for P in terms of any other normalization as well from the same MCMC samples. For instance, if one prefers the source profiles expressed in terms of proportions (so that the rows of P sum to 100%), then one can obtain the

samples of new P by applying the corresponding transformation (normalization) on the samples of P , and can carry out estimation of P by using the new samples. From Tables 10 and 11, it can be seen that the source profiles (spatial profiles) for each season show a different pattern. During Winter, the major source regions seem to be near (Marysville, Lake Forest Park, Puyallup), (Tacoma-I, Puyallup), and (Harbor Island, Duwamish, South park), and during Summer, the source profile is fairly spread out over the regions Harbor Island, Duwamish, South Park, Kent, and Tacoma. One of the source regions in Winter, (Marysville, Lake Forest Park, Puyallup) coincides with a high wood smoke area, which does not appear in Summer. This supports that wood smoke is an additional source of PM_{10} in Winter. Fig. 2

Table 11
Posterior summary for the parameters P , μ , and Σ of Model 6, Winter

	Source 1		Source 2		Source 3		μ		Σ	
	Mean	Standard deviation	Mean	Standard deviation	Mean	Standard deviation	Mean	Standard deviation	Mean	Standard deviation
Marysville	15.72	1.35	0	0	6.14	1.39	28.16	1.48	8.79	4.58
Everett	6.76	1.00	0	0	6.42	0.98	21.77	0.92	27.02	4.03
Lake Forest Park	12.13	1.28	5.43	1.08	0	0	30.08	1.33	60.17	8.73
Harbor Island	0	0	7.95	1.26	11.16	1.15	33.33	1.22	20.78	4.20
Duwamish	2.27	0.81	8.31	1.26	11.70	1.17	36.34	1.30	26.64	4.79
South Park	2.52	0.71	8.29	1.06	9.31	0.98	28.04	1.14	19.06	3.41
Kent	3.67	0.93	9.06	1.09	7.55	1.03	28.50	1.18	39.31	5.62
Tacoma-R	0	0	9.80	1.43	6.86	1.45	25.44	1.37	109.44	14.58
Tacoma-I	4.95	1.14	12.56	1.29	7.71	1.16	36.38	1.44	42.90	6.46
Puyallup	9.85	1.36	11.58	1.03	0	0	26.99	1.32	10.52	5.21

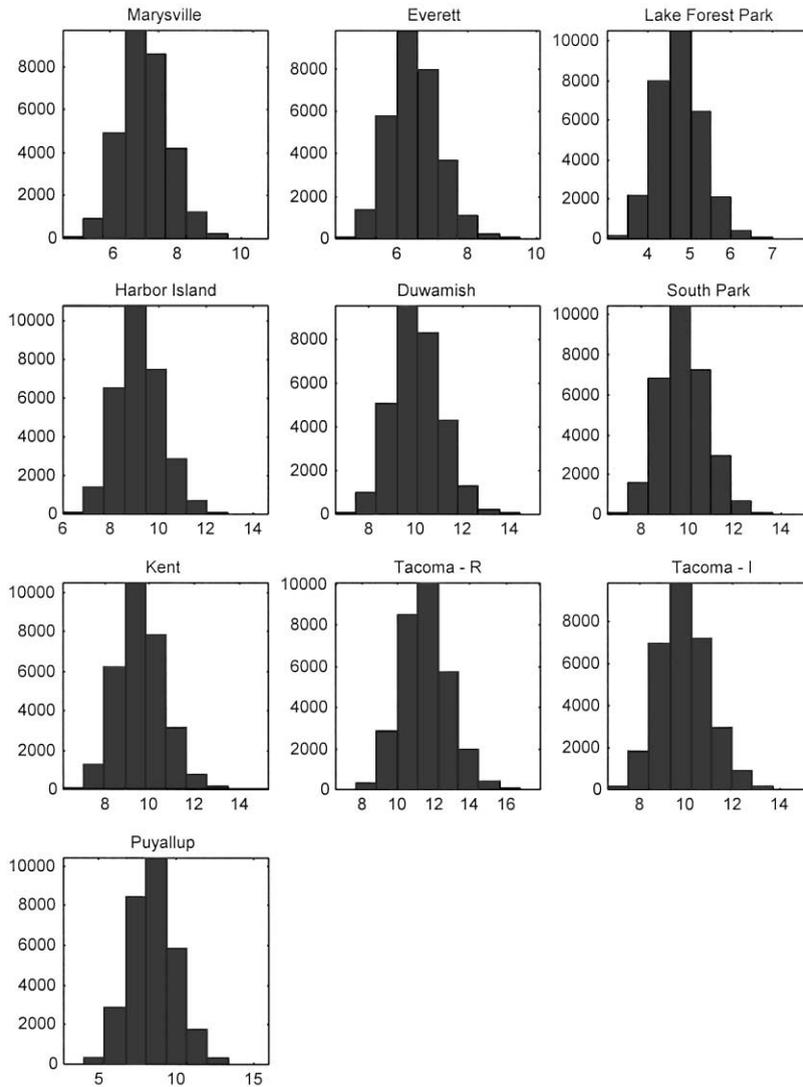


Fig. 2. Posterior sample histogram for source profile, Summer (July–September).

contains the histogram of 30,000 posterior samples for each element of P for Summer, and Fig. 3 contains the histogram for each element of Source 1 profile (that seems to correspond to wood smoke spatial profile) for Winter.

For illustrative purposes, we also apply some other commonly used methods for determining the number of factors (see Section 1) to these data. For Summer data, the 90 Percent trace method gives 1 (based on the covariance matrix) or 3 (based on the correlation

matrix), Rule-of-one gives 1, Barlett's test gives 9 (based on the covariance matrix) or 5 (based on the correlation matrix), Malinowski's indicator function gives 2, World's cross-validation approach gives 1, NUMFACT gives 3, and modified NUMFACT gives 2.

For Winter data, the 90 Percent trace method gives 1 (based on the covariance matrix) or 3 (based on the correlation matrix), Rule-of-one gives 1, Bartlett's test gives 7 (based on the covariance matrix) or 6 (based

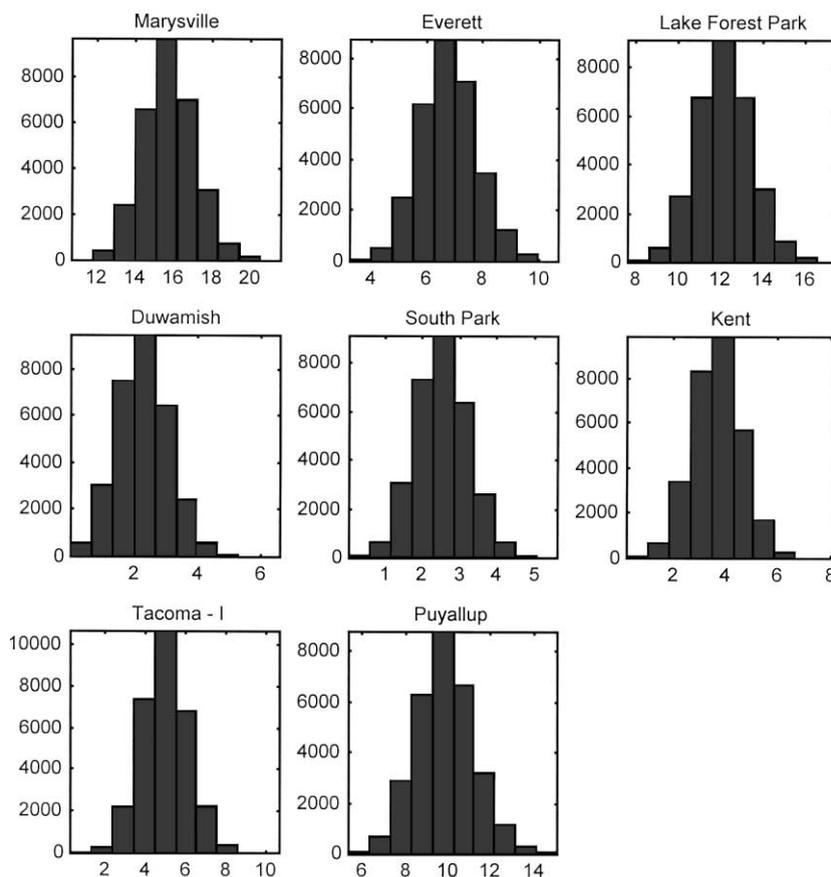


Fig. 3. Posterior sample histogram for Source 1 profile, Winter (October–March).

on the correlation matrix), Malinowki's indicator function gives 1, World's cross-validation approach gives 1, NUMFACT gives 3 (based on the covariance matrix) or 4 (based on the correlation matrix), and modified NUMFACT gives 2 (based on the covariance matrix) or 3 (based on the correlation matrix). There seems to be a fairly large uncertainty in the number of factors for these data, which is why it is important to report model uncertainty estimates (as in Tables 9 and 10).

7. Discussions

In this article, we have developed a general approach for assessing model uncertainty in multivariate receptor models (and standard factor analysis

models). In earlier multivariate receptor modeling, the number of sources and identifiability conditions were determined first, and the inferences of the remaining model parameters were made conditionally on that. This approach ignores the uncertainty involved in the number of sources and selection of identifiability conditions. We approached the problem using the marginal likelihood. The marginal likelihood of each model can easily be converted to the posterior probability of the model, which may well serve as an uncertainty estimate of the model. Although marginal likelihoods used to be computationally intractable, recent developments in MCMC methodology make accurate estimation of them possible. The methods using MCMC (for calculating the marginal likelihoods) have not yet been applied in many statistical problems. The main advantage of

the MCMC approach introduced here is that the marginal likelihood of each model can be calculated based on the same posterior sample that is used to make inferences on the parameters (without requiring any additional sampling). Thus, using a single posterior sample for each model, we can simultaneously obtain the model uncertainty estimate, the estimates for the parameters and their uncertainties. Although we confined ourselves, for brevity of presentation, to one type of identifiability conditions (zeros in the source composition matrix P), the method can be applied to other types of identifiability conditions (e.g., zeros in the source contribution matrix A) with a slight modification in MCMC algorithm.

Throughout this article, we have assumed that the errors are normally distributed, which makes all the full conditionals available in closed forms. When a nonnormal distribution for errors is assumed, some of the full conditionals might be difficult to determine. In that case, the general methodology in Section 4 can be extended using the importance-weighted method of Chen [11] and Oh [12] for the unknown conditionals, replacing each of them by an arbitrary (weighting) conditional density times the ratio of the posterior kernels. Choosing a good weighting conditional density can be a challenging problem due to high dimensionality of each block of parameters in multivariate receptor models.

Another assumption we have made (in the priors) is that the factors (the source contributions) are uncorrelated. Although our method is shown to be robust to violation of this assumption through a simulation study, a further extension of the model (and the method) would be to treat Φ as an additional parameter in the model, with its own prior. This brings the model into the form of a Bayesian hierarchical model (the prior distribution of factors depends on the unknown hyperparameter Φ). In addition to the identifiability conditions (C1–C2) to remove rotational indeterminacy, q additional linearly independent restrictions on the parameters such that Φ is a correlation matrix or one of the nonzero elements in each row of P is known, are needed to cope with indeterminacy of factors by a constant multiplication. Inter-comparisons of models, when Φ is assumed known and when it is unknown, using marginal likelihoods are still under investigation.

Acknowledgements

Although the research described in this article has been funded by the United States Environmental Protection Agency through agreement CR825173-01-0 to the University of Washington, it has not been subjected to the agency's required peer and policy review and therefore does not necessarily reflect the views of the Agency and no official endorsement should be inferred.

References

- [1] L. Hopke, An introduction to receptor modeling, *Chemometrics and Intelligent Laboratory Systems* 10 (1991) 21–43.
- [2] D.J. Bartholomew, M. Knot, *Latent Variable Models and Factor Analysis*, 2nd edn., Oxford Univ. Press, New York, 1999.
- [3] R.C. Henry, E.S. Park, C.H. Spiegelman, Comparing a new algorithm with the classic methods for estimating the number of factors, *Chemometrics and Intelligent Laboratory Systems* 48 (1999) 91–97.
- [4] E.S. Park, R.C. Henry, C.H. Spiegelman, Estimating the number of factors to include in a high-dimensional multivariate bilinear model, *Communications In Statistics, B* 29 (2000) 723–746.
- [5] E.R. Malinowski, D.G. Howery, *Factor Analysis in Chemistry*, Wiley, New York, 1980, p. 82.
- [6] S. Wold, Cross-validators estimation of the number of components in factor and principal components models, *Technometrics* 20 (1978) 397–405.
- [7] E.S. Park, C.H. Spiegelman, R.C. Henry, Bilinear estimation of pollution source profiles and amounts by using multivariate receptor models, to appear in *Environmetrics* (2001).
- [8] T.W. Anderson, *An Introduction to Multivariate Statistical Analysis*, 2nd edn., Wiley, New York, 1984.
- [9] R.A. Kass, A.E. Raftery, Bayes' factor, *Journal of the American Statistical Association* 90 (1995) 395–773.
- [10] A.E. Raftery, Hypothesis testing and model selection, in: W.R. Gilks, S. Richardson, D.J. Spiegelhalter (Eds.), *Markov Chain Monte Carlo in Practice*, Chapman & Hall, New York, 1996.
- [11] M.-H. Chen, Importance-weighted marginal Bayesian posterior density estimation, *Journal of the American Statistical Association* 89 (1994) 818–824.
- [12] M.S. Oh, Estimation of posterior density functions from a posterior sample, *Computational Statistics and Data Analysis* 29 (1999) 411–427.
- [13] J.F. Geweke, K.J. Singleton, Interpreting the likelihood ratio statistic in factor models when sample size is small, *Journal of the American Statistical Association* 75 (1980) 133–137.
- [14] L. Tierney, Markov chains for exploring posterior distributions, *Annals of Statistics* 22 (1994) 1701–1762.

- [15] W.R. Gilks, S. Richardson, D.J. Spiegelhalter, *Markov Chain Monte Carlo in Practice*, Chapman & Hall, New York, 1996.
- [16] A.E. Gelfand, A.F.M. Smith, Sampling-based approaches to calculating marginal densities, *Journal of the American Statistical Association* 85 (1990) 398–409.
- [17] T. Larson, Personal communication (2000).
- [18] R.C. Henry, Personal communication (2000).
- [19] J. Besag, P. Green, D. Higdon, K. Mengersen, Bayesian computation and stochastic systems, *Statistical Science* 10 (1995) 3–41.