Chemometrics and Intelligent Laboratory Systems 149 (2015) 215–226

Contents lists available at ScienceDirect

Chemometrics and Intelligent Laboratory Systems

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

Robust Bayesian multivariate receptor modeling

Eun Sug Park [a,*], Man-Suk Oh [b,1]

[a] Texas A&M Transportation Institute, 3135 TAMU, College Station, TX 77843-3135, USA
[b] Department of Statistics, Ewha Womans University, Seoul 120-750, Republic of Korea

ARTICLE INFO

Article history:
Received 14 May 2015
Received in revised form 24 August 2015
Accepted 25 August 2015
Available online 2 September 2015

Keywords:
Source apportionment
Uncertainty estimation
Model uncertainty
Outliers

ABSTRACT

Multivariate receptor modeling aims to unfold the multivariate air pollution data into components associated with different sources of air pollution based on ambient measurements of air pollutants. It is now a widely accepted approach in source identification and apportionment. An evolving area of research in multivariate receptor modeling is to quantify uncertainty in estimated source contributions as well as model uncertainty caused by the unknown identifiability conditions, sometimes referred to as *rotational ambiguity*. Unlike the uncertainty estimates for the source composition profiles that have been available in commonly used receptor modeling tools such as positive matrix factorization, little research has been conducted on the uncertainty estimation for the source contributions or the identifiability conditions. Bayesian multivariate receptor modeling based on Markov chain Monte Carol methods is an attractive approach as it offers a great deal of flexibility in both modeling and estimation of parameter uncertainty and model uncertainty. In this paper, we propose a robust Bayesian multivariate receptor modeling approach that can simultaneously estimate uncertainty in source contributions as well as in compositions and uncertainty due to the unknown identifiability conditions by extending the previous Bayesian multivariate receptor modeling in two ways. First, we explicitly account for nonnegativity constraints on the source contributions, in addition to the nonnegativity constraints on the source compositions, in both parameter estimation and model uncertainty estimation. Second, we account for outliers that may often exist in the air pollution data in estimation by considering a heavy-tailed error distribution. The approach is illustrated with both simulated data and real $PM_{2.5}$ speciation data from Phoenix, Arizona, USA.

1. Introduction

Multivariate receptor modeling is a collection of methods for identifying major pollution sources/source categories and quantifying their impacts based on ambient measurements of air pollutants. Estimation of the source composition profiles (that can serve as chemical fingerprints of pollution source categories) and contributions (amounts of pollution) from different source categories have been the primary concerns in multivariate receptor modeling. Hopke [6,8] and Tauler et al. [23] provide comprehensive reviews of multivariate receptor models. Despite their increasingly widespread use, however, there has been little research on how to quantify uncertainty associated with estimated source contributions as well as coping with uncertainty associated with the unknown identifiability conditions.

Statistically, multivariate receptor models may be viewed as latent variables models because they assume that the correlations among the observed multivariate data are induced by a set of latent variables (source contributions). More specifically, they can be regarded as factor analysis models, other than the nonnegativity constraints on source compositions and contributions, in the sense that both the source composition profiles (factor loadings) and the source contributions (factor scores or latent variables) are unknown parameters to be estimated based on the observed multivariate data. Estimation of the source compositions and contributions in multivariate receptor models is a statistically challenging problem because of the well-known factor indeterminacy (sometimes referred to as rotational ambiguity) problem (parameterization is not unique without placing additional constraints) even under the assumption of the known number of sources. As a matter of fact, the uncertainty (model uncertainty) associated with factor indeterminacy or the unknown number of sources in multivariate receptor modeling has been largely ignored, and there have been only a handful of studies that addressed model uncertainty issues in multivariate receptor models (see, e.g., [17,20,21]). Park et al. [20,21] discussed model uncertainty associated with the unknown number of sources and factor indeterminacy in multivariate receptor models. Paatero et al. [17] discussed capturing the uncertainty of positive matrix factorization (PMF) analyses due to random errors and rotational ambiguity in the estimation of the source composition matrix under the assumption of the known number of sources.

Also, the uncertainty estimation of the source contributions had not received much attention unlike the uncertainty estimation of the source composition profiles. Recently, the level of interest in uncertainty

[*] Corresponding author. Tel.: +1 979 845 9942; fax: +1 979 845 6008.
E-mail addresses: e-park@tamu.edu (E.S. Park), msoh@ewha.ac.kr (M.-S. Oh).
[1] Tel.: +82 2 3277 2374; fax: +82 2 3277 3607.

http://dx.doi.org/10.1016/j.chemolab.2015.08.021
0169-7439/© 2015 Elsevier B.V. All rights reserved.

Table 1
Candidate models in simulation.

Model number	Source	Pre-specified position of zeros in *P*
1	1	2, 3
	2	1, 4
	3	5, 6
2	1	2, 4
	2	1, 5
	3	3, 6
3	1	4, 6
	2	2, 5
	3	3, 7

estimation of source contributions has increased dramatically. This increased interest and appeal lies in, at least partially, the need for quantification of uncertainty in source-specific exposures used in the assessment of health effects associated with the major sources.

Bayesian multivariate receptor modeling naturally accounts for uncertainty in estimated source contributions and source compositions and has the capability to deal with model uncertainty in a more coherent manner. Park et al. [20,21] proposed Bayesian approaches in multivariate receptor modeling, based on Markov chain Monte Carlo (MCMC) methods, that can simultaneously provide the uncertainty estimates for the number of sources/identifiability conditions as well as for the source contributions and compositions. In Park et al. [20,21], the nonnegativity constraints were imposed only on the source compositions because the centered source contributions were utilized in modeling and a multivariate Gaussian distribution was assumed for the error term. As a matter of fact, not incorporating nonnegativity constraints on the source contributions explicitly in modeling may lead some source contribution estimates to be negative especially when the true source contributions are small (see, e.g., [9]).

Also, outliers often exist in the air pollution data in practice, and the use of Gaussian error distribution may not account for those outliers. Paatero [16] discussed three possible reasons for outliers in the air pollution data (a weak local source that may be visible only occasionally, a laboratory error or a contamination in the field, and extreme values in source contributions), and presented a robust factorization method based on the Huber influence function that can handle the first two kinds of outliers, omitting the third type. As mentioned in Paatero [16], it is a question of definition whether the third type is considered an outlier or not because the patterns of composition for the observation of the third type would still be the same as those for lower concentration samples. We also omit the third type of outliers from this discussion. Gajewski and Spiegelman [3] developed estimators of source composition profiles that are robust to outliers (outlying errors) by minimizing the L_2E objective function. Neither uncertainty estimation of source contributions nor model uncertainty estimation was discussed in Gajewski and Spiegelman [3], however.

In this paper, we propose a Bayesian approach to robust multivariate receptor modeling that can account for nonnegativity constraints on the source contributions and possible outliers (corresponding to the first two types of outliers discussed in [16]) in both parameter estimation and model uncertainty (caused by factor indeterminacy) estimation under the assumption of the known number of sources. Section 2 introduces robust Bayesian multivariate receptor models that account for nonnegativity constraints on the source contributions and outlying errors in modeling and estimation. Section 3 presents estimation of parameter uncertainty and model uncertainty by MCMC. Section 4 contains the simulation study. In Section 5, the proposed method is applied to the real $PM_{2.5}$ speciation data for Phoenix, AZ, USA. Finally, concluding remarks are made in Section 6.

2. Model

The basic physical model can be written as follows:

$$Y_t = A_t P + E_t\,, \quad t = 1, \cdots, T, \tag{1}$$

where $Y_t = (Y_{t1}, Y_{t2}, \cdots, Y_{tJ})$: tth observation consisting of concentrations of J pollutants (chemical species) measured in time t, T = # of observations, q = # of major pollution sources, P: $q \times J$ source composition matrix of which rows are the source composition profiles ($P_{k,}$, $k = 1, \cdots, q$), $P_k = (P_{k1}, P_{k2}, \cdots, P_{kJ})$: kth source composition profile consisting of the fractional amount of each chemical species in the emissions from the kth source, $A_t = (A_{t1}, A_{t2}, \cdots, A_{tq})$: source contribution vector in time t where A_{tk} is the contribution from the kth source, and $E_t = (E_{t1}, \cdots, E_{tJ})$: measurement error in pollutant concentrations in time t. Our main goal is to estimate, A and P along with their uncertainties.

Remark 1. Note that in reality source composition profiles in P of Eq. (1) are not truly constant and may change over time. Therefore, the errors in Eq. (1) actually represent both measurement error and variability in the source compositions. If the source composition profiles are relatively stable over time, however, the measurement error component of the model assuming constant source composition profiles can adequately handle random fluctuations of source emissions about stable central values. In such cases, the assumption of constant source compositions may not be unreasonable and the source composition profiles that we estimate would correspond to the average source composition profiles (where the average is taken over time). As a matter of fact, one of the key assumptions of the PMF model is also that the composition of the emission sources is constant over the period of sampling at the receptors (see, e.g., [10,17]). If there are systematic trends in the source composition profiles (i.e., if the number of sources changes and/or the source composition profiles systematically change over time), then it will make more sense to analyze the subsets of data (grouped according to different time periods) separately so that the source composition profiles are approximately the same within each time period (as in [10]).

It is well-known in multivariate receptor modeling and factor analysis that parameters A and P cannot be uniquely estimated (even under the assumption that q is known) without enforcing additional

Table 2
Log of marginal likelihoods (LogMD) for three candidate models estimated by Method T assuming a heavy-tailed distribution (T_4) and Method G assuming a Gaussian distribution for errors when the data contain outliers.

Dataset	Model number							
	1		2		3		Selected model by Method T	Selected model by Method G
	T	G	T	G	T	G		
1	**−2781**	**−3719**	−3386	−4178	−2919	−4190	1	1
2	**−2647**	**−3315**	−3312	−3549	−3186	−3803	1	1
3	**−2588**	**−3239**	−3178	−3684	−2724	−3288	1	1
4	**−2749**	−3265	−3310	−3763	−2889	**−3241**	1	3
5	**−2760**	**−3485**	−3335	−3725	−2864	−3557	1	1

Notes: 1. Only 5 cases are shown for illustration. 2. ‘T’ stands for Method T and ‘G’ stand for Method G. 3. The largest LogMD among the three models obtained by each method for each dataset is shown in bold.