

Using the chemical mass balance model to estimate pollution source contributions from correlated air quality observations

William F. Christensen
Department of Statistical Science
Southern Methodist University
william@mail.smu.edu

Abstract

In the environmental sciences, receptor models are used to evaluate the contribution of various pollution sources to the air composition at a location. Using pollution source profiles, profile uncertainties, and measurement error variances, the chemical mass balance (CMB) model can be fit in order to partition the ambient pollutants measured at the receptor into a collection of source contributions. We discuss the use of the CMB model for the analysis of a multivariate time series of air quality measurements, and we consider estimation and inference procedures which account for the multiple sources of correlation in the data. Using computer simulation, approaches are compared under various scenarios in which standard model assumptions are violated.

1 Introduction

Receptor modeling involves the analysis of ambient pollutant concentrations in order to partition pollutants observed at a receptor to their sources. The desired intent of receptor modeling is to assist in the implementation and evaluation of air pollution control programs. For example, a government agency may be interested in making conclusions such as “ $X\%$ of ambient lead particulates observed at the location of interest is due to auto emissions.”

The principle of conservation of mass underlies most of the analysis approaches used in the receptor modeling. For example, the amount of lead particulates observed at a receptor is assumed to be the simple sum of the contributions from sources such as auto exhaust, industrial emissions, lead smelters, etc. Thus, the amount of the i th species (c_i) is modeled as a linear combination of contributions from k pollution sources

$$c_i = \sum_{j=1}^k a_{ij}s_j, \quad i = 1, \dots, p, \quad (1)$$

where s_j is the mass contribution of source j to the atmosphere at the receptor, and (a_{1j}, \dots, a_{pj}) , $j = 1, \dots, k$, represents the composition or “profile” of the j th source, with a_{ij} taking values in $[0,1]$ and $\sum_{i=1}^p a_{ij} \leq 1$. Mass balance models based on (1) were introduced by Miller, Friedlander, and Hidy (1972) and Winchester and Nifong (1971). For the case where the number of observed pollutants (p) exceeds the number of sources (k), the model may be written

$$\mathbf{c} = \mathbf{A} \mathbf{s} + \mathbf{e} \quad (2)$$

where \mathbf{c} is a p -vector of ambient species, \mathbf{A} is a $p \times k$ source profile matrix with the j th column describing the composition of the j th source, and \mathbf{e} is a p -vector of errors.

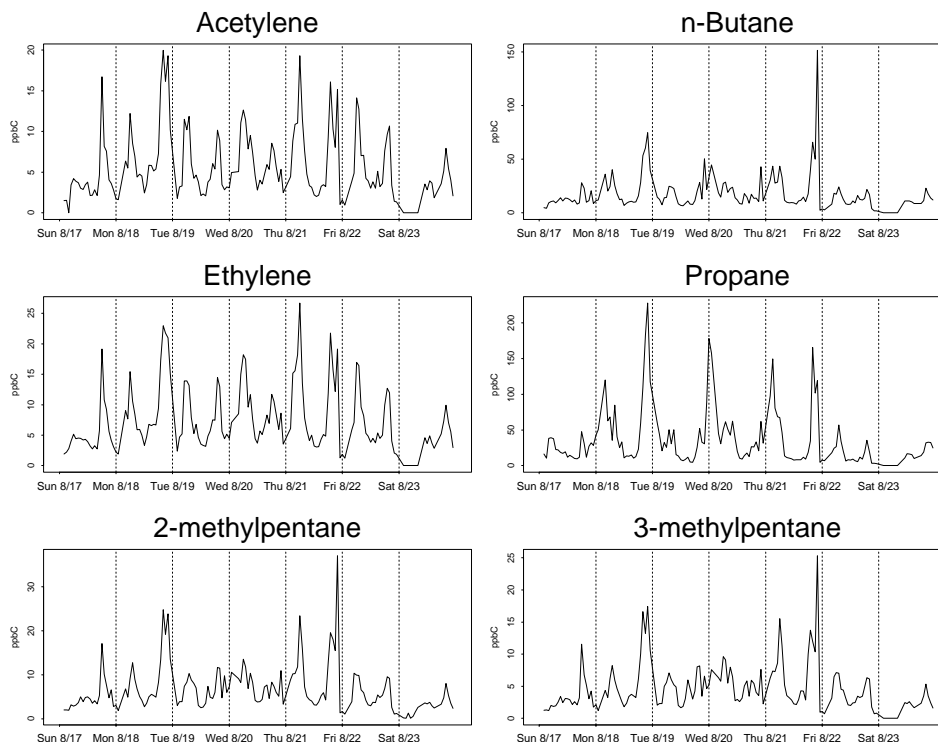


Figure 1: Hourly ambient measurements for 6 volatile organic compounds.

An example of ambient data used for receptor modeling is the El Paso volatile organic compound (VOC) concentrations shown in Figure 1. These data include hourly measurements of over 50 VOCs. From the plots for acetylene and ethylene, it is apparent from the temporal dependence structure that these VOCs have one or more sources in common. In fact, it is known that acetylene and ethylene are originating in great part from auto exhaust—note the peaks corresponding to weekday morning and weekday evening rush hour. In estimating the hourly pollution source contributions from these data, we would like to use both the source profile information and the observed dependence structure in the ambient observations.

Although (2) is widely used in the receptor modeling literature, the name of this model varies, depending on what is known about the pollution source profiles. When the profile matrix \mathbf{A} is assumed known, (2) is often referred to as the chemical mass balance (CMB) model. Friedlander (1973) used a least squares fit to obtain estimates of pollution source contributions from the CMB model, and numerous analyses of pollution data have been conducted using regression methods (e.g., Gatz, 1975; Mayrsohn and Crabtree, 1976; Kowalczyk, Choquette, and Gordon, 1978). However, when little or nothing is known about the number and nature of the pollution sources, (2) has been viewed as a factor analysis model. For such cases, the researcher is interested in simultaneously estimating the pollution source profiles (which take the role of “factor loadings”) and the pollution source contributions (which take the role of “factor scores”). See, e.g., Thurston and Spengler (1985), Koutrakis and Spengler (1987), and Henry, Lewis, and Collins (1994). Note that while only a single ambient observation is needed to estimate \mathbf{s} in the CMB setting,

approach when the ambient data consist of multiple temporally-correlated observations. In Section 3 we introduce multivariate estimators which account for temporal dependence. Section 4 discusses a simulation study used to assess the properties of the proposed estimators.

2 The effective variance solution

In the influential work of Watson, Cooper, and Huntzicker (1984), the authors employ an “effective variance” weighted least squares approach for fitting the measurement error model (3). The approach is implemented in the United States Environmental Protection Agency’s EPA-CMB8.2 program (EPA, 2000). The effective variance solution of Watson, Cooper, and Huntzicker (1984) is a simplification of the generalized least squares solution of Britt and Luecke (1973) and estimates the k -vector \mathbf{s} using: the ambient measure \mathbf{c} , an initial source profile matrix estimate $\tilde{\mathbf{A}}$, and estimates of the measurement error variances and source profile error variances (\mathbf{V}_c and \mathbf{V}_A , respectively). Both \mathbf{V}_c and \mathbf{V}_A are diagonal matrices. The EPA-CMB8.2 algorithm is given in Watson, Cooper, and Huntzicker (1984), which gives formulas for most of the quantities of interest in either summation form or one-column-at-time form. For simplicity of discussion and ease of computational implementation, we give the algorithm written in matrix form. Throughout the paper we use the standard matrix notations “ \otimes ” (denoting the kronecker product) and “vec,” where $\text{vec } \mathbf{X} = (\mathbf{x}'_1, \dots, \mathbf{x}'_n)'$ for any matrix \mathbf{X} with columns $\mathbf{x}_1, \dots, \mathbf{x}_n$.

EPA-CMB8.2 Algorithm

Step EPA.0—Initialize:

$$\tilde{\mathbf{A}}^{(0)} = \tilde{\mathbf{A}} \text{ and } \mathbf{s}^{(0)} = \mathbf{0}$$

Step EPA.1—Update “effective variance”:

$$\mathbf{V}_e^{(i)} = \mathbf{V}_c + \text{diag} \left\{ \left(\mathbf{s}^{(i)'} \otimes \mathbf{I}_p \right) \mathbf{V}_A \left(\mathbf{s}^{(i)} \otimes \mathbf{I}_p \right) \right\}$$

Step EPA.2—Update $\mathbf{s}^{(i)}$:

$$\mathbf{s}^{(i+1)} = \mathbf{s}^{(i)} + \left(\tilde{\mathbf{A}}^{(i)'} \left(\mathbf{V}_e^{(i)} \right)^{-1} \tilde{\mathbf{A}}^{(i)} \right)^{-1} \tilde{\mathbf{A}}^{(i)'} \left(\mathbf{V}_e^{(i)} \right)^{-1} \left[\mathbf{c} - \tilde{\mathbf{A}} \mathbf{s}^{(i)} \right]$$

Step EPA.3—Update $\tilde{\mathbf{A}}^{(i)}$:

$$\begin{aligned} \text{vec } \tilde{\mathbf{A}}^{(i+1)} = & \text{vec } \tilde{\mathbf{A}}^{(i)} + \mathbf{V}_A \left(\mathbf{s}^{(i)} \otimes \mathbf{I}_p \right) \left(\mathbf{V}_e^{(i)} \right)^{-1} \times \\ & \left[\mathbf{I}_{pn} - \left(\mathbf{I}_n \otimes \tilde{\mathbf{A}}^{(i)} \right) \left\{ \left(\mathbf{I}_n \otimes \tilde{\mathbf{A}}^{(i)'} \right) \left(\mathbf{V}_e^{(i)} \right)^{-1} \left(\mathbf{I}_n \otimes \tilde{\mathbf{A}}^{(i)} \right) \right\}^{-1} \times \right. \\ & \left. \left(\mathbf{I}_n \otimes \tilde{\mathbf{A}}^{(i)'} \right) \left(\mathbf{V}_e^{(i)} \right)^{-1} \right] \left[\mathbf{c} - \tilde{\mathbf{A}} \mathbf{s}^{(i)} \right] \end{aligned}$$

Note that when \mathbf{A} is measured without error, the optimal weight matrix for the generalized least squares estimation of \mathbf{s} is simply $\mathbf{V}_c = \text{Var}\{\mathbf{e}\}$. However, when \mathbf{A}

is measured with error, the optimal weight matrix for the generalized least squares estimation of \mathbf{s} is the “effective variance” of \mathbf{c} , denoted $\mathbf{V}_e = \text{Var}\{\tilde{\mathbf{A}}\mathbf{s} + \mathbf{e}\}$. The effective variance incorporates uncertainty due to both measurement error (\mathbf{V}_c) and profile error (\mathbf{V}_A).

If we omit the update of $\tilde{\mathbf{A}}^{(i)}$ in Step EPA.3 and iterate Steps EPA.1-2 until

$$\max_{j=1,\dots,k} \left(\frac{s_j^{(i+1)} - s_j^{(i)}}{s_j^{(i)}} \right) < .01, \quad (4)$$

we denote the estimate \mathbf{s}_{EPA} , where the “EPA” subscript refers to the EPA-CMB8.2 program. Note that when the update of $\tilde{\mathbf{A}}^{(i)}$ in Step EPA.3 is omitted, the estimate of $\mathbf{s}^{(i+1)}$ in Step EPA.2 simplifies to

$$\mathbf{s}^{(i+1)} = \left(\tilde{\mathbf{A}}' \left(\mathbf{V}_e^{(i)} \right)^{-1} \tilde{\mathbf{A}} \right)^{-1} \tilde{\mathbf{A}}' \left(\mathbf{V}_e^{(i)} \right)^{-1} \mathbf{c}.$$

When Steps EPA.1-3 are iterated until convergence, we denote the estimate $\tilde{\mathbf{s}}_{\text{EPA}}$, where the tilde indicates that the estimator updates $\tilde{\mathbf{A}}$ in each iteration. This is the estimate obtained when the “Britt and Luecke” option is selected in EPA-CMB8.2. Statistical inference for \mathbf{s}_{EPA} or $\tilde{\mathbf{s}}_{\text{EPA}}$ can be carried out using using

$$\text{Var}\{\mathbf{s}^{(i+1)}\} = \left(\tilde{\mathbf{A}}^{(i)'} \left(\mathbf{V}_e^{(i)} \right)^{-1} \tilde{\mathbf{A}}^{(i)} \right)^{-1}.$$

3 The multiple ambient measures scenario

We consider now the scenario in which we have n temporally-correlated ambient observations ($\mathbf{c}_1, \dots, \mathbf{c}_n$) and are interested in estimating the associated pollution source contributions ($\mathbf{s}_1, \dots, \mathbf{s}_n$). We write our model

$$\begin{aligned} \mathbf{C} &= \mathbf{A} \mathbf{S} + \mathbf{E} \\ \tilde{\mathbf{A}} &= \mathbf{A} + \mathbf{U} \end{aligned} \quad (5)$$

where \mathbf{C} is a $p \times n$ matrix with columns $\mathbf{c}_1, \dots, \mathbf{c}_n$, \mathbf{S} is a $k \times n$ matrix with columns $\mathbf{s}_1, \dots, \mathbf{s}_n$, and \mathbf{E} is a $p \times n$ matrix with columns $\mathbf{e}_1, \dots, \mathbf{e}_n$. Also, $\text{Var}\{\text{vec } \mathbf{E}\} = \mathbf{V}_C$, and $\text{Var}\{\text{vec } \mathbf{U}\} = \mathbf{V}_A$. Note that even when \mathbf{A} is measured without error ($\mathbf{U} = \mathbf{0}$), model (5) has a different form than the classical multivariate multiple regression model. For example, our dependent variable matrix \mathbf{C} has each observation in a separate column, so obtaining an additional observation has no effect on the predictor variable matrix \mathbf{A} but results in an additional column of regression coefficients in \mathbf{S} . Thus, when the n observations are independent and identically distributed, $\text{Var}\{\text{vec } \mathbf{E}\} = \mathbf{I}_n \otimes \boldsymbol{\Sigma}$, where $\boldsymbol{\Sigma} = \text{Var}\{\mathbf{e}_t\}$, $t = 1, \dots, n$. Then, unlike the standard multivariate multiple regression scenario, the ordinary least squares estimator of \mathbf{S} is not the best linear unbiased estimator for the IID sample setting. Rather, for the case where $\mathbf{U} = \mathbf{0}$ and observations are independent, the BLUE for \mathbf{S} is

$$\hat{\mathbf{S}} = (\mathbf{A}' \boldsymbol{\Sigma}^{-1} \mathbf{A})^{-1} \mathbf{A}' \boldsymbol{\Sigma}^{-1} \mathbf{C}. \quad (6)$$

Two of the pivotal CMB model assumptions (Watson, Chow, and Pace, 1991) are: (i) measurement errors are random, uncorrelated, and normally distributed,

and (ii) chemical species do not react with each other (i.e., they add linearly). Several potential drawbacks exist when using the EPA-CMB8.2 algorithm to estimate pollution source contributions from multiple ambient measurements such as the El Paso data illustrated in Figure 1. First, although the EPA-CMB8.2 algorithm assumes no correlation among measurement errors, in practice the p ambient species are collected (and often measured) simultaneously. Thus one might be concerned by the effect of measurement error correlations among the species. Second, because of meteorological effects and because the species often *do* chemically react in the atmosphere, there may exist a temporally-correlated component in the error term. Third, because the columns of \mathbf{S} constitute a k -variate time series process, when there is error in estimating \mathbf{A} , the fitted model will include a temporally-correlated error component represented by \mathbf{T} in

$$\begin{aligned}\mathbf{C} &= \tilde{\mathbf{A}} \mathbf{S} + (\mathbf{A} - \tilde{\mathbf{A}}) \mathbf{S} + \mathbf{E} \\ &= \tilde{\mathbf{A}} \mathbf{S} + \mathbf{T} + \mathbf{E}.\end{aligned}$$

Thus, we wish to develop an estimator which accounts for both multivariate and temporal correlations when estimating \mathbf{S} and carrying out statistical inference.

We noted in the previous section that estimating the vector \mathbf{s} from a single ambient observation \mathbf{c} via the EPA-CMB8.2 algorithm is similar to a generalized least squares estimation of \mathbf{s} using the effective variance of \mathbf{c} as the weight matrix. We now consider generalizations of the effective variance solution which account for multiple temporally-correlated ambient observations. To do this, we first make an assumption about $\text{Var}\{\text{vec } \mathbf{E}\}$ which facilitates an iterative estimation procedure that separately accounts for the multivariate and temporal dependencies in the data. Specifically, we assume $\text{cov}\{\mathbf{e}_t, \mathbf{e}_{t+h}\} = \rho^*(|h|) \boldsymbol{\Sigma}$, where $\rho^*(\cdot)$ is a univariate correlation function and $\boldsymbol{\Sigma} = \text{Var}\{\mathbf{e}_t\}$, $t = 1, \dots, n$. Making this assumption about the error process implies

$$\text{Var}\{\text{vec } \mathbf{E}\} = \mathbf{P} \otimes \boldsymbol{\Sigma} \tag{7}$$

where $\mathbf{P} = (\rho_{ij})_{i,j=1,\dots,n}$, and $\rho_{ij} = \rho^*(|i-j|)$. Similar assumptions have been used in the analysis of multivariate dependent data (see, e.g., Mardia, 1984) and is similar to the separability assumption used in the modeling of spatio-temporally correlated data (see, e.g., Rodriguez-Iturbe and Mejia, 1974). While constraining the error process to satisfy (7) may often be unjustified, we consider its use as a tool for approximating the complex multivariate and temporal dependencies among observations. In the simulation study of Section 4, we show that even when model (7) is not correct, using it to evaluate the correlation structure still yields good statistical properties in terms of efficiency of estimation and validity of inference.

3.1 Estimation of \mathbf{S} assuming \mathbf{A} is known

We first consider an estimator for \mathbf{S} in (5) for the case in which $\mathbf{U} = \mathbf{0}$. That is, we assume $\tilde{\mathbf{A}} = \mathbf{A}$ is known. For this case, we note that the model (5) can be written as

$$\text{vec } \mathbf{C} = (\mathbf{I}_n \otimes \mathbf{A}) \text{vec } \mathbf{S} + \text{vec } \mathbf{E},$$

and the effective variance of $\text{vec } \mathbf{C}$ is simply

$$\mathbf{V}_e = \text{Var}\{\text{vec } \mathbf{E}\} = \mathbf{P} \otimes \boldsymbol{\Sigma}.$$

So, the generalized least squares estimator for \mathbf{S} is

$$\begin{aligned}
\text{vec } \ddot{\mathbf{S}} &= \left[\left(\mathbf{I}_n \otimes \tilde{\mathbf{A}}' \right) \left(\mathbf{P} \otimes \boldsymbol{\Sigma} \right)^{-1} \left(\mathbf{I}_n \otimes \tilde{\mathbf{A}} \right) \right]^{-1} \left(\mathbf{I}_n \otimes \tilde{\mathbf{A}}' \right) \left(\mathbf{P} \otimes \boldsymbol{\Sigma} \right)^{-1} \text{vec } \mathbf{C} \\
&= \left[\mathbf{I}_n \otimes \left(\tilde{\mathbf{A}}' \boldsymbol{\Sigma}^{-1} \tilde{\mathbf{A}} \right)^{-1} \tilde{\mathbf{A}}' \boldsymbol{\Sigma}^{-1} \right] \text{vec } \mathbf{C} \\
&\text{or} \\
\ddot{\mathbf{S}} &= \left(\tilde{\mathbf{A}}' \boldsymbol{\Sigma}^{-1} \tilde{\mathbf{A}} \right)^{-1} \tilde{\mathbf{A}}' \boldsymbol{\Sigma}^{-1} \mathbf{C}. \tag{8}
\end{aligned}$$

Note that when \mathbf{A} is known and (7) holds, the generalized least squares estimator for the time series \mathbf{S} (8) is identical to the estimator (6) for the independent observation case. The fact that (8) is free of \mathbf{P} facilitates a simple estimation procedure. While we realize that the $\mathbf{U} = \mathbf{0}$ assumption will rarely hold in practice, we might expect an estimator based on (8) to perform reasonably as a potential estimator because of its simple form. However, we would not expect that inference based on an estimator of

$$\text{Var} \left\{ \text{vec } \ddot{\mathbf{S}} \right\} = \mathbf{P} \otimes \left(\tilde{\mathbf{A}}' \boldsymbol{\Sigma}^{-1} \tilde{\mathbf{A}} \right)^{-1}$$

would be valid because of the source profile error \mathbf{U} . So, to carry out inference for (8) we use instead the ‘‘sandwich formula’’

$$\text{Var} \left\{ \text{vec } \ddot{\mathbf{S}} \right\} = \left(\mathbf{I}_n \otimes \mathbf{M} \right) \mathbf{V}_e^* \left(\mathbf{I}_n \otimes \mathbf{M}' \right) \tag{9}$$

where $\mathbf{M} = \left(\tilde{\mathbf{A}}' \boldsymbol{\Sigma}^{-1} \tilde{\mathbf{A}} \right)^{-1} \tilde{\mathbf{A}}' \boldsymbol{\Sigma}^{-1}$ and

$$\mathbf{V}_e^* = \mathbf{P} \otimes \boldsymbol{\Sigma} + \left(\ddot{\mathbf{S}}' \otimes \mathbf{I}_p \right) \mathbf{V}_A \left(\ddot{\mathbf{S}} \otimes \mathbf{I}_p \right).$$

An iterative algorithm for obtaining (8) and (9) is given below.

Algorithm I
(for obtaining (8) and (9))

Step I.0—Initialize:

$$\underbrace{\boldsymbol{\Sigma}^{(0)}}_{p \times p} = \left(\sigma_{\ell m}^{(0)} \right) = \frac{1}{n} \sum_{t=1}^n \mathbf{V}_{\mathbf{c}_t}$$

where

$$\mathbf{V}_{\mathbf{C}} = \begin{bmatrix} \mathbf{V}_{\mathbf{c}_1} & & \mathbf{0} \\ & \ddots & \\ \mathbf{0} & & \mathbf{V}_{\mathbf{c}_n} \end{bmatrix}$$

Step I.1—Update $\ddot{\mathbf{S}}^{(i)}$:

$$\ddot{\mathbf{S}}^{(i+1)} = \left(\tilde{\mathbf{A}}' \left(\boldsymbol{\Sigma}^{(i)} \right)^{-1} \tilde{\mathbf{A}} \right)^{-1} \tilde{\mathbf{A}}' \left(\boldsymbol{\Sigma}^{(i)} \right)^{-1} \mathbf{C}$$

Step I.2—Update $\boldsymbol{\Sigma}^{(i)}$:

$$\underbrace{\mathbf{R}}_{p \times n} = \mathbf{C} - \tilde{\mathbf{A}} \mathbf{S}^{(i+1)}$$

$$\bar{\mathbf{R}} = \frac{1}{n} \sum_{t=1}^n \mathbf{R}_t$$

$$\underbrace{\mathbf{H}}_{p \times p} = \frac{1}{n-1} \sum_{t=1}^n (\mathbf{R}_t - \bar{\mathbf{R}})(\mathbf{R}_t - \bar{\mathbf{R}})'$$

$$\ddot{\mathbf{H}} = (\ddot{h}_{\ell m}) = \mathbf{Q} \begin{bmatrix} \lambda_1 & & & & & & & \mathbf{0} \\ & \ddots & & & & & & \\ & & \lambda_{p-k} & & & & & \\ & & & \lambda_{p-k} & & & & \\ & & & & \ddots & & & \\ \mathbf{0} & & & & & & & \lambda_{p-k} \end{bmatrix} \mathbf{Q}'$$

where $\mathbf{Q} = \begin{bmatrix} \underbrace{\mathbf{Q}_1}_{p \times (p-k)} & \underbrace{\mathbf{Q}_2}_{p \times k} \end{bmatrix}$ and $(\lambda_1, \dots, \lambda_{p-k}, 0, \dots, 0)$ are the eigenvectors and ordered eigenvalues of \mathbf{H} .

$$\Sigma^{(i+1)} = \ddot{\mathbf{H}} + \begin{bmatrix} \max\{(\sigma_{11}^{(0)} - \ddot{h}_{11}), 0\} & & & \mathbf{0} \\ & \ddots & & \\ & & & \\ \mathbf{0} & & & \max\{(\sigma_{pp}^{(0)} - \ddot{h}_{pp}), 0\} \end{bmatrix}$$

Step I.3—Obtain estimate of \mathbf{P} :

$$\underbrace{\ddot{\mathbf{e}}}_{1 \times n} = \frac{1}{p-k} \mathbf{1}'_{p-k} \begin{bmatrix} \lambda_1^{-1/2} & & & \mathbf{0} \\ & \ddots & & \\ & & & \\ \mathbf{0} & & & \lambda_{p-k}^{-1/2} \end{bmatrix} \mathbf{Q}'_1 \mathbf{R}$$

Model $\rho(\cdot)$ using $\ddot{\mathbf{e}}$ and construct $\hat{\mathbf{P}} = (\hat{\rho}_{ij})$, $i, j = 1, \dots, n$, where $\hat{\rho}_{ij} = \hat{\rho}(|i-j|)$. (E.g., let $\rho(\cdot)$ be an AR(1) correlation function.)

Note that the estimated error matrix \mathbf{R} in Step I.2 has rank $p-k$. So, the covariance matrix \mathbf{H} has only $p-k$ nonzero eigenvalues. The adjusted estimate $\ddot{\mathbf{H}}$ is a full rank matrix with its smallest $k+1$ eigenvalues equal to the smallest nonzero eigenvalue of \mathbf{H} . To lend stability to the algorithm (particularly in the first few iterations), the p diagonal elements of $\Sigma^{(i+1)}$ are constrained to be no less than the corresponding average measurement error variances $\sigma_{\ell\ell}^{(0)}$, $\ell = 1, \dots, p$, calculated in Step I.0.

Because the updated estimates of \mathbf{S} and Σ in Steps I.1 and I.2 are free of \mathbf{P} , we only have need to iterate these two steps until convergence to obtain the final estimate which we denote $\ddot{\mathbf{S}}_V$, where the “V” subscript denotes the use of a full variance-covariance matrix weight in the GLS estimation of Step I.1 and the double dots indicate the estimator ignores the source profile error. After obtaining $\ddot{\mathbf{S}}_V$ and

a final estimate of Σ , we then proceed to estimate \mathbf{P} in Step I.3. Hypothesis tests and confidence intervals based on \mathbf{S}_V may be constructed using

$$\ddot{\mathbf{\Gamma}}_V = \left(\mathbf{I}_n \otimes \mathbf{M}^{(i)} \right) \mathbf{V}_e^{*(i)} \left(\mathbf{I}_n \otimes \mathbf{M}^{(i)'} \right) \quad (10)$$

as an estimator for (9), where

$$\mathbf{M}^{(i)} = \left(\tilde{\mathbf{A}}'(\Sigma^{(i)})^{-1} \tilde{\mathbf{A}} \right)^{-1} \tilde{\mathbf{A}}'(\Sigma^{(i)})^{-1}$$

and

$$\mathbf{V}_e^{*(i)} = \mathbf{P} \otimes \Sigma^{(i)} + \left(\ddot{\mathbf{S}}^{(i+1)'} \otimes \mathbf{I}_p \right) \mathbf{V}_A \left(\ddot{\mathbf{S}}^{(i+1)} \otimes \mathbf{I}_p \right). \quad (11)$$

As an alternative to $\ddot{\mathbf{S}}_V$, we also consider substituting $\mathbf{D}^{(i)} = \text{diag}\{\Sigma^{(i)}\}$ in place of $\Sigma^{(i)}$ in the GLS estimation of Step I.1. We refer to this simpler estimator as $\hat{\mathbf{S}}_D$, where the ‘‘D’’ subscript refers to the use of a diagonal weight matrix. Statistical inference for \mathbf{S}_D can be carried out using

$$\ddot{\mathbf{\Gamma}}_D = \left(\mathbf{I}_n \otimes \mathbf{M}^{(i)} \right) \mathbf{V}_e^{*(i)} \left(\mathbf{I}_n \otimes \mathbf{M}^{(i)'} \right) \quad (12)$$

as an estimator for (9), where $\mathbf{V}_e^{*(i)}$ is defined as in (11) but with

$$\mathbf{M}^{(i)} = \left(\tilde{\mathbf{A}}'(\mathbf{D}^{(i)})^{-1} \tilde{\mathbf{A}} \right)^{-1} \tilde{\mathbf{A}}'(\mathbf{D}^{(i)})^{-1}.$$

3.2 Estimation of \mathbf{S} when \mathbf{A} is unknown

When the source profile measurement $\tilde{\mathbf{A}}$ is known to be contaminated by source profile errors with known variances as in (5), one can formulate a multiple-observation effective variance solution by first defining the ‘‘effective variance’’ of $\text{vec } \mathbf{C}$ as

$$\mathbf{V}_e = \mathbf{P} \otimes \Sigma + (\mathbf{S}' \otimes \mathbf{I}_p) \mathbf{V}_A (\mathbf{S} \otimes \mathbf{I}_p).$$

Then, the GLS estimator for \mathbf{S} is

$$\text{vec } \hat{\mathbf{S}} = \left[(\mathbf{I}_n \otimes \mathbf{A}') \mathbf{V}_e^{-1} (\mathbf{I}_n \otimes \mathbf{A}) \right]^{-1} (\mathbf{I}_n \otimes \mathbf{A}') \mathbf{V}_e^{-1} (\text{vec } \mathbf{C}) \quad (13)$$

and

$$\text{Var} \left\{ \text{vec } \hat{\mathbf{S}} \right\} = \left[(\mathbf{I}_n \otimes \mathbf{A}') \mathbf{V}_e^{-1} (\mathbf{I}_n \otimes \mathbf{A}) \right]^{-1}. \quad (14)$$

An iterative algorithm for obtaining (13) and (14) is given below.

Algorithm II (for obtaining (13) and (14))

Step II.0—Initialize:

$$\mathbf{S}^{(0)} = \mathbf{0}, \quad \tilde{\mathbf{A}}^{(0)} = \mathbf{A}, \quad \text{and} \quad \mathbf{V}_e^{(0)} = (v_{lm}^{(0)}) = \mathbf{V}_C$$

Step II.1—Update $\mathbf{S}^{(i)}$:

$$\begin{aligned} \text{vec } \mathbf{S}^{(i+1)} &= \text{vec } \mathbf{S}^{(i)} + \left[\left(\mathbf{I}_n \otimes \tilde{\mathbf{A}}^{(i)'} \right) \left(\mathbf{V}_e^{(i)} \right)^{-1} \left(\mathbf{I}_n \otimes \tilde{\mathbf{A}}^{(i)} \right) \right]^{-1} \times \\ &\quad \left(\mathbf{I}_n \otimes \tilde{\mathbf{A}}^{(i)'} \right) \left(\mathbf{V}_e^{(i)} \right)^{-1} \text{vec} \left[\mathbf{C} - \mathbf{A} \mathbf{S}^{(i)} \right] \end{aligned}$$

Step II.2—Update $\Sigma^{(i)}$ and $\mathbf{P}^{(i)}$ (as in Steps I.2-3 of Algorithm I)

Step II.3—Update $\mathbf{V}_e^{(i)}$:

$$\mathbf{G} = (g_{lm}) = \mathbf{P}^{(i+1)} \otimes \ddot{\mathbf{H}}$$

$$\begin{aligned} \mathbf{V}_e^{(i+1)} &= \mathbf{G} + \begin{bmatrix} \max\{(v_{11}^{(0)} - g_{11}), 0\} & & \mathbf{0} \\ & \ddots & \\ \mathbf{0} & & \max\{(v_{pp}^{(0)} - g_{pp}), 0\} \end{bmatrix} \\ &+ \left(\mathbf{S}^{(i)'} \otimes \mathbf{I}_p \right) \mathbf{V}_A \left(\mathbf{S}^{(i)} \otimes \mathbf{I}_p \right) \end{aligned}$$

Step II.4—Update $\tilde{\mathbf{A}}^{(i)}$:

$$\begin{aligned} \text{vec } \tilde{\mathbf{A}}^{(i+1)} &= \text{vec } \tilde{\mathbf{A}}^{(i)} + \mathbf{V}_A \left(\mathbf{S}^{(i+1)} \otimes \mathbf{I}_p \right) \left(\mathbf{V}_e^{(i+1)} \right)^{-1} \times \\ &\quad \left[\mathbf{I}_{pn} - \left(\mathbf{I}_n \otimes \tilde{\mathbf{A}}^{(i)} \right) \left\{ \left(\mathbf{I}_n \otimes \tilde{\mathbf{A}}^{(i)'} \right) \left(\mathbf{V}_e^{(i+1)} \right)^{-1} \left(\mathbf{I}_n \otimes \tilde{\mathbf{A}}^{(i)} \right) \right\}^{-1} \times \right. \\ &\quad \left. \left(\mathbf{I}_n \otimes \tilde{\mathbf{A}}^{(i)'} \right) \left(\mathbf{V}_e^{(i+1)} \right)^{-1} \right] \text{vec} \left[\mathbf{C} - \mathbf{A} \mathbf{S}^{(i+1)} \right] \end{aligned}$$

In Algorithm I, it was not necessary to estimate \mathbf{P} until after the estimate of \mathbf{S} was obtained. However, in Algorithm II (Step II.2), we are required to update $\mathbf{P}^{(i)}$ in each iteration since it is involved in the calculation of $\mathbf{V}_e^{(i+1)}$ in Step II.3.

The update of $\tilde{\mathbf{A}}^{(i)}$ in Step II.4 can be omitted in the interest of algorithmic stability. That is, we can let $\tilde{\mathbf{A}}^{(i)} = \tilde{\mathbf{A}}^{(0)}$, for $i = 1, 2, \dots$. We denote \mathbf{S}_V to be the estimator obtained by iterating Steps II.1-3 until convergence, where the “V” subscript denotes the use of a full effective variance matrix weight in the GLS estimation of Step II.1. Iterating Steps II.1-4 until convergence of $\mathbf{S}^{(i)}$ yields the estimator $\tilde{\mathbf{S}}_V$, where the tilde denotes the update of $\tilde{\mathbf{A}}^{(i)}$ in each iteration.

As alternatives to \mathbf{S}_V and $\tilde{\mathbf{S}}_V$, we also consider substituting $\mathbf{D}^{(i)} = \text{diag}\{\mathbf{V}_e^{(i)}\}$ in place of $\mathbf{V}_e^{(i)}$ in the GLS estimation of Step II.1. We refer to the resulting estimators as \mathbf{S}_D and $\tilde{\mathbf{S}}_D$, where the “D” subscript refers to the use of a diagonal weight matrix. Statistical inference procedures for \mathbf{S}_V or $\tilde{\mathbf{S}}_V$ can be constructed using

$$\Gamma_V = \left[\left(\mathbf{I}_n \otimes \tilde{\mathbf{A}}^{(i)'} \right) \left(\mathbf{V}_e^{(i)} \right)^{-1} \left(\mathbf{I}_n \otimes \tilde{\mathbf{A}}^{(i)} \right) \right]^{-1} \quad (15)$$

as an estimate of (14). Inference for \mathbf{S}_D or $\tilde{\mathbf{S}}_D$ can be carried out using

$$\Gamma_D = \mathbf{M}^{(i)} \mathbf{V}_e^{(i)} \mathbf{M}^{(i)'} \quad (16)$$

as an estimate of (14), where

$$\mathbf{M}^{(i)} = \left(\left(\mathbf{I}_n \otimes \tilde{\mathbf{A}}^{(i)'} \right) \left(\mathbf{D}^{(i)} \right)^{-1} \left(\mathbf{I}_n \otimes \tilde{\mathbf{A}}^{(i)} \right) \right)^{-1} \left(\mathbf{I}_n \otimes \tilde{\mathbf{A}}^{(i)'} \right) \left(\mathbf{D}^{(i)} \right)^{-1}$$

4 Simulation Study

In order to compare the new estimators ($\check{\mathbf{S}}_V$, $\check{\mathbf{S}}_D$, \mathbf{S}_V , \mathbf{S}_D , $\tilde{\mathbf{S}}_V$, and $\tilde{\mathbf{S}}_D$) with the existing estimators (\mathbf{S}_{CMB} and $\check{\mathbf{S}}_{\text{CMB}}$), we consider the simulation of a simple airshed using source profiles obtained from Javitz, Watson, and Robinson (1988). Table 1 gives the source profile matrix (\mathbf{A}) with columns corresponding to pollutants from soil, a coal-fired power plant, vehicle exhaust, and wood burning. Measurements of the 17 ambient elements were generated for each of 25 “days,” assuming that the average daily source contribution to the air at the receptor is $5 \mu\text{g}/\text{m}^{-3}$ for each of the pollution sources. Thus the model from which the data were generated is

$$\underbrace{\mathbf{C}}_{17 \times 25} = \underbrace{\mathbf{A}}_{17 \times 4} \underbrace{\mathbf{S}}_{4 \times 25} + \overbrace{\mathbf{W} + \mathbf{E}_{\text{me}}}^{\mathbf{E}} \quad (17)$$

$$\tilde{\mathbf{A}} = \mathbf{A} + \mathbf{U}$$

The columns of \mathbf{S} constitute a 4-variate time series with each element of the mean vector equal to $5 \mu\text{g}/\text{m}^{-3}$. The columns of \mathbf{W} constitute a zero-mean 17-variate time series representing a temporally-correlated error component in addition to the error \mathbf{E}_{me} associated with the ambient measurements. Such a \mathbf{W} process may exist due to meteorological effects or interactions among the chemical species.

The (i, j) element of the source profile error matrix ($\tilde{\mathbf{U}}$) was generated from a normal distribution with zero mean and standard deviation equal to $\ell \times a_{ij}$, where a_{ij} is the corresponding element of \mathbf{A} . In our simulations, ℓ took on values in $(0.1, 0.5)$. The \mathbf{S} process was generated as a lognormal multivariate AR(1) process with mean vector of $(5, 5, 5, 5)'$ and AR(1) coefficient matrix of

$$\Phi_{\mathbf{S}} = \mathbf{0}, \quad \begin{bmatrix} .3 & 0 & 0 & 0 \\ 0 & .6 & 0 & 0 \\ 0 & 0 & .7 & 0 \\ 0 & 0 & 0 & .4 \end{bmatrix}, \quad \text{or} \quad \begin{bmatrix} .1 & .1 & .1 & .1 \\ .1 & .3 & .1 & .1 \\ .1 & .1 & .4 & .1 \\ .1 & .1 & .1 & .1 \end{bmatrix}.$$

Note that each of the latter two possibilities for $\Phi_{\mathbf{S}}$ results in observed data for which the factorization assumption (7) does not hold. Notwithstanding, it will be illustrated that estimators and associated inference procedures based on (7) still performed well under all choices for $\Phi_{\mathbf{S}}$. The coefficient of variation for each element of \mathbf{S}_t is equal to 0.5. (That is, the standard deviation for each element of \mathbf{S}_t is 2.5.) Because different pollution source contributions tend to be correlated, we generate disturbances $(\mathbf{S}_t - \Phi_{\mathbf{S}} \mathbf{S}_{t-1})$ for the pollution source contributions that have pairwise correlations equal to 0.5.

The 17-variate \mathbf{W} process was generated as a zero-mean, lognormal multivariate AR(1) process with structure similar to the 4-variate \mathbf{S} process. The standard

Element	Soil	Coal-fired power plant	Vehicle exhaust	Wood burning
OC	0.04	0.01	0.407	0.475
EC	0.005	0.005	0.222	0.128
Al	0.0783	0.146	0.0004	0.00021
Si	0.305	0.219	0.001	0
Cl	0.00032	0.00052	0.022	0.00509
K	0.0282	0.0122	0.0001	0.0086
Ca	0.0287	0.0121	0.0005	0.00067
Ti	0.00474	0.0087	0	0
V	0.000095	0.00068	0	0
Cr	0.000070	0.00062	0	0
Mn	0.00069	0.00043	0.0011	0
Fe	0.0354	0.0809	0.0014	0
Ni	0.000044	0.00039	0	0
Cu	0.000030	0.00029	0	0
Zn	0.000060	0.00067	0.001	0.00037
Br	0.000003	0	0.0243	0
Pb	0.000015	0.00042	0.0815	0

Table 1: Pollution source profile matrix (\mathbf{A}). From Javitz, Watson, and Robinson (1988).

deviation of each element of \mathbf{W}_t was equal to m times the corresponding element of $\mathbf{A} \mathbf{S}$, where m took on values between 0 (equivalent to no \mathbf{W} process) and 0.2. The standard deviation of each element of \mathbf{E}_{me} was set equal to q times the corresponding element of $\mathbf{A} \mathbf{S}$, where q took on values in (0.05, 0.3). Because the p ambient species collected at the receptor are often measured simultaneously (e.g., using gas chromatography or x-ray fluorescence methods), we consider correlated measurement errors by setting each pairwise intercorrelation to a value in (0, 0.3).

After a series of observations \mathbf{C} and an estimated source profile matrix $\hat{\mathbf{A}}$ are generated, each of the estimators of \mathbf{S} was obtained. In keeping with the standard metrics in the receptor modeling literature, we evaluate the goodness of each estimator using the Average Absolute Error (AAE) of $\hat{\mathbf{S}}$:

$$\text{AAE}(\hat{\mathbf{S}}) = \frac{1}{kn} \sum_{j=1}^k \sum_{t=1}^n \frac{|\hat{S}_{jt} - S_{jt}|}{S_{jt}}.$$

Because we are also interested in statistical inference properties, we consider also the average coverage probability for the elements of $\hat{\mathbf{S}}$ when using nominal 95% confidence intervals. In the discussion below, we report the AAE and coverage, averaged over all replications of the simulation study.

Several trends were apparent from the simulation results. First, the new estimators that update $\mathbf{A}^{(i)}$ in each iteration ($\tilde{\mathbf{S}}_V$ and $\tilde{\mathbf{S}}_D$) perform much worse than the non-updating estimators \mathbf{S}_V and \mathbf{S}_D . In fact, we exclude $\tilde{\mathbf{S}}_V$ and $\tilde{\mathbf{S}}_D$ from our discussion at this point because of their inability to compete with the other estimators. Second, the estimators using diagonal weight matrices in the GLS estimation ($\check{\mathbf{S}}_D$ and \mathbf{S}_D) generally outperform the estimators using a full covariance matrix weight. Finally, the new estimators (particularly $\check{\mathbf{S}}_D$ and \mathbf{S}_D) significantly outperform the current standards (\mathbf{S}_{EPA} and $\tilde{\mathbf{S}}_{EPA}$) when: (i) the source profile errors \mathbf{U} were large, (ii) a temporally-correlated error component (\mathbf{W}) exists, and

	m.e. corr's = 0		m.e. corr's = 0.3	
	AAE*	Coverage†	AAE*	Coverage†
	for \hat{S}_{jt}	for \hat{S}_{jt}	for \hat{S}_{jt}	for \hat{S}_{jt}
\mathbf{S}_{CMB}	0.55	0.87	0.57	0.86
$\tilde{\mathbf{S}}_{\text{CMB}}$	1.89	0.77	2.44	0.77
$\tilde{\mathbf{S}}_{\text{D}}$	0.34	0.95	0.36	0.94
$\tilde{\mathbf{S}}_{\text{V}}$	0.33	0.98	0.35	0.98
\mathbf{S}_{D}	0.42	0.94	0.44	0.94
\mathbf{S}_{V}	0.53	0.75	0.54	0.74

* Average Absolute Error (expressed as proportion of true value)

† Nominal 95% confidence interval

Table 2: Performance of estimators when source profile error variances are large (standard deviation of each source profile error u_{ij} equals $0.5 \times a_{ij}$).

(iii) correlations among pollution source contributions are large, e.g.,

$$\Phi_{\mathbf{S}} = \begin{bmatrix} .1 & .1 & .1 & .1 \\ .1 & .3 & .1 & .1 \\ .1 & .1 & .4 & .1 \\ .1 & .1 & .1 & .1 \end{bmatrix}.$$

In all scenarios considered, the existing estimators \mathbf{S}_{EPA} and $\tilde{\mathbf{S}}_{\text{EPA}}$ never significantly outperform the new estimators.

Tables 2 and 3 illustrate typical cases in the simulation. For the results in Table 2, the standard deviation of each source profile error u_{ij} equals $0.5 \times a_{ij}$,

$$\Phi_{\mathbf{S}} = \begin{bmatrix} .3 & 0 & 0 & 0 \\ 0 & .6 & 0 & 0 \\ 0 & 0 & .7 & 0 \\ 0 & 0 & 0 & .4 \end{bmatrix},$$

the standard deviation of each element of \mathbf{E}_{me} is equal to 0.1 times the corresponding element of $\mathbf{A S}$, and no temporally-correlated error component \mathbf{W} is included. We note that the multivariate estimators are more efficient than the existing estimators when source profile error variances are large. Additionally, the coverage probability is near the nominal for $\tilde{\mathbf{S}}_{\text{D}}$, $\tilde{\mathbf{S}}_{\text{V}}$, and \mathbf{S}_{D} , while the coverage for \mathbf{S}_{EPA} and $\tilde{\mathbf{S}}_{\text{EPA}}$ is well below the nominal.

Table 3 illustrates the performance under a similar scenario, but with small source profile error standard deviations (u_{ij} equals $0.1 \times a_{ij}$), and slightly larger ambient measurement errors (the standard deviation of each element of \mathbf{E}_{me} is equal to 0.3 times the corresponding element of $\mathbf{A S}$). In this scenario, $\tilde{\mathbf{S}}_{\text{D}}$, $\tilde{\mathbf{S}}_{\text{V}}$, and \mathbf{S}_{D} are only slightly more efficient than \mathbf{S}_{EPA} and $\tilde{\mathbf{S}}_{\text{EPA}}$, but the coverage probabilities are better for the multivariate estimators.

	m.e. corr's = 0		m.e. corr's = 0.3	
	AAE* for \hat{S}_{jt}	Coverage† for \hat{S}_{jt}	AAE* for \hat{S}_{jt}	Coverage† for \hat{S}_{jt}
\mathbf{S}_{CMB}	0.35	0.77	0.31	0.80
$\tilde{\mathbf{S}}_{\text{CMB}}$	0.35	0.78	0.31	0.80
$\tilde{\mathbf{S}}_{\text{D}}$	0.28	0.94	0.27	0.94
$\tilde{\mathbf{S}}_{\text{V}}$	0.31	0.85	0.28	0.92
\mathbf{S}_{D}	0.29	0.88	0.27	0.85
\mathbf{S}_{V}	0.35	0.68	0.30	0.72

* Average Absolute Error (expressed as proportion of true value)

† Nominal 95% confidence interval

Table 3: Performance of estimators when source profile error variances are small (standard deviation of each source profile error u_{ij} equals $0.1 \times a_{ij}$).

5 Conclusion

Many pollution source apportionment studies use multiple measurements which are correlated in time (and space). These correlations should be exploited when estimating \mathbf{S} , and must be properly accounted for when carrying out inference for \mathbf{S} . In this paper, we consider the use of an error covariance matrix factorization which allows the multivariate and temporal correlations to be estimated separately in an iterative algorithm. The multivariate estimators proposed appear to be more efficient than existing estimators, and associated confidence intervals have coverage probabilities that are closer to the nominal confidence level.

References

- Britt, H. I., and Luecke, R. H. (1973), "The Estimation of Parameters in Nonlinear, Implicit Models," *Technometrics*, 15:233-247.
- Christensen, W. F., and Sain, S. R. (2000), "Use of latent variable models in air quality monitoring," in *Computing Science and Statistics: Proceedings of the 32nd Symposium on the Interface*.
- Environmental Protection Agency (2000), "EPA-CMB8.2 User's Manual," EPA Publication No. EPA-454/R-00-XXX, Office of Air Quality Planning & Standards, Research Triangle Park, NC.
- Friedlander, S. K. (1973), "Chemical Element Balances and Identification of Air Pollution Sources," *Environmental Science Technology*, 7, 235-240.
- Gatz, D. F. (1975), "Relative Contributions of Different Sources of Urban Aerosols: Application of a New Estimation Method to Multiple Sites in Chicago," *Atmospheric Environment*, 9, 1-18.
- Gleser, L. J. (1997), "Some Thoughts on Chemical Mass Balance Models," *Chemo-metrics and Intelligent Laboratory Systems*, 37, 15-22.
- Henry, R. C., Lewis, C. W., and Collins, J. F. (1994), "Vehicle-Related Hydrocarbon Source Compositions from Ambient Data: The GRACE/SAFER Method," *Environmental Science Technology*, 28, 823-832.

- Javitz, H. S., Watson, J. G., and Robinson, N. F. (1988), "Performance of the Chemical Mass Balance Model with Simulated Local-Scale Aerosols," *Atmospheric Environment*, 22, 2309-2322.
- Koutrakis, P. and Spengler, J. D. (1987), "Source Apportionment of Ambient Particles in Steubenville, OH Using Specific Rotation Factor Analysis," *Atmospheric Environment*, 21, 1511-1519.
- Kowalczyk, G. S., Choquette, C. E., and Gordon, G. E. (1978), "Chemical Element Balances and Identification of Air Pollution Sources in Washington, D.C.," *Atmospheric Environment*, 12, 1143-1153.
- Mardia, K. V. (1984), "Spatial Discrimination and Classification Maps," *Communications in Statistics*, 16, 2181-2197.
- Mayrsohn, H., and Crabtree, J. H. (1976), "Source Reconciliation of Atmospheric Hydrocarbons," *Atmospheric Environment*, 10, 137-143.
- Miller, M. S., Friedlander, S. K., and Hidy, G. M. (1972), "A Chemical Element Balance for the Pasadena Aerosol," *Journal of Colloid Interface Science*, 39, 65-176.
- Park, E. S., Guttorp, P., and Henry, R. C. (2000), "Multivariate Receptor Modeling for Temporally Correlated Data by Using MCMC," National Research Center for Statistics and the Environment, TRS #034, www.nrcse.washington.edu/research/reports/reports.asp.
- Park, E. S., Oh, M.-S., and Guttorp, P. (2000), "Multivariate Receptor Models and Model Uncertainty," National Research Center for Statistics and the Environment, TRS #060, www.nrcse.washington.edu/research/reports/reports.asp.
- Rodriguez-Iturbe, I., and Mejia, J. M. (1974), "The Design of Rainfall Networks in Time and Space," *Water Resources Research*, 10, 713-728.
- Thurston, G. D., and Spengler, J. D. (1985), "A Quantitative Assessment of Source Contributions to Inhalable Particulate Matter Pollution in Metropolitan Boston," *Atmospheric Environment*, 19, 9-25.
- Watson, J. G., Chow, J. C., and Pace, T. G. (1991), "Chemical Mass Balance," in *Receptor Modeling for Air Quality Management*, ed. P. K. Hopke, New York: Elsevier Science Publishers, pp. 83-116.
- Watson, J. G., Cooper, J. A., and Huntzicker, J. J. (1984), "The Effective Variance Weighting for Least Squares Calculations Applied to the Mass Balance Receptor Model," *Atmospheric Environment*, 18, 1347-1355.
- Winchester, J. W., and Nifong, G. D. (1971), "Water Pollution in Lake Michigan by Trace Elements from Pollution Aerosol Fallout," *Water, Air, and Soil Pollution*, 1, 50-64.
- Yang, H. (1994), "Confirmatory Factor Analysis and Its Application to Receptor Modeling," unpublished Ph.D. dissertation, University of Pittsburgh, Department of Mathematics and Statistics.