ARTICLE IN PRESS

ELSEVIER

Atmospheric Environment ▮ (▮▮▮▮) ▮▮▮–▮▮▮

ATMOSPHERIC ENVIRONMENT

www.elsevier.com/locate/atmosenv

Measurement error models in chemical mass balance analysis of air quality data

William F. Christensen[a,*], Richard F. Gunst[b]

[a] Department of Statistics, Brigham Young University, Provo, UT 84602-6575, USA
[b] Department of Statistical Science, Southern Methodist University, P.O. Box 750332, Dallas, TX 75275-0332, USA

Received 20 June 2003; received in revised form 18 September 2003; accepted 6 October 2003

UNCORRECTED PROOF

Abstract

The chemical mass balance (CMB) equations have been used to apportion observed pollutant concentrations to their various pollution sources. Typical analyses incorporate estimated pollution source profiles, estimated source profile error variances, and error variances associated with the ambient measurement process. Often the CMB model is fit to the data using an iteratively re-weighted least-squares algorithm to obtain the effective variance solution. We consider the chemical mass balance model within the framework of the statistical measurement error model (e.g., Measurement Error Models, Wiley, NewYork, 1987), and we illustrate that the models assumed by each of the approaches to the CMB equations are in fact special cases of a general measurement error model. We compare alternative source contribution estimators with the commonly used effective variance estimator when standard assumptions are valid and when such assumptions are violated. Four approaches for source contribution estimation and inference are compared using computer simulation: weighted least squares (with standard errors adjusted for source profile error), the effective variance approach of Watson et al. (Atmos, Environ., 18, 1984, 1347), the Britt and Luecke (Technometrics, 15, 1973, 233) approach, and a method of moments approach given in Fuller (1987, p. 193). For the scenarios we consider, the simplistic weighted least-squares approach performs as well as the more widely used effective variance solution in most cases, and is slightly superior to the effective variance solution when source profile variability is large. The four estimation approaches are illustrated using real $PM_{2.5}$ data from Fresno and the conclusions drawn from the computer simulation are validated.

© 2003 Published by Elsevier Ltd.

Keywords: Source apportionment; Receptor model; Chemical mass balance model; Effective-variance weighting

1. Introduction

The chemical mass balance (CMB) model is used to apportion ambient pollutants to the sources from whence they came. Introduced by Miller et al. (1972) and Winchester and Nifong (1971), the model is based on the principle of conservation of mass. That is, the amount of a chemical species observed ambiently is a simple sum of the pollutant contributions emanating from a finite number of pollution sources in the region. Hence, the mass concentration of the ith species (y_i) is a linear combination of contributions from k pollution sources:

$$y_i = \sum_{j=1}^{k} x_{ij}\beta_j + e_i, \quad i = 1, ..., p, \tag{1}$$

where β_j is the mass contribution of source j to the atmosphere at the receptor, e_i is the measurement error at the receptor for the ith species, and $\mathbf{x}_{(j)} = (x_{1j}, \ldots, x_{pj})'$, $j = 1, \ldots, k$, represents the composition or "profile" of the jth source. The components of each profile are non-negative proportions that sum to no more than 1. Thus, y_i, β_j, and e_i are expressed in units of mass concentration such as $\mu g\,m^{-3}$, while x_{1j} is unitless. For notational

*Corresponding author. Tel.: 1-801-422-7057; fax: 1-801-422-0635.
E-mail address: william@stat.byu.edu (W.F. Christensen).

1352-2310/$ - see front matter © 2003 Published by Elsevier Ltd.
doi:10.1016/j.atmosenv.2003.10.018

simplicity, we often refer to the model for $\mathbf{y}=(y_1,\ldots,y_p)'$ as

$$\mathbf{y} = \mathbf{X\beta} + \mathbf{e}, \tag{2}$$

where $\mathbf{X}=[\mathbf{x}_{(1)},\mathbf{x}_{(2)},\ldots,\mathbf{x}_{(k)}]$, $\boldsymbol{\beta}=(\beta_1,\beta_2,\ldots,\beta_k)$, and $\mathbf{e}=(e_1,\ldots,e_p)$. We can also describe the composition of $\mathbf{X}$ in terms of its rows $\mathbf{x}_i'$, $i=1,\ldots,p$, so that

$$\mathbf{X} = \begin{bmatrix} \mathbf{x}_1' \\ \mathbf{x}_2' \\ \vdots \\ \mathbf{x}_p' \end{bmatrix}.$$

The CMB model assumptions are discussed by Watson et al. (1991). We briefly state them as follows:

1. Source emission compositions are constant over time.
2. Chemical species do not react with each other (they add linearly).
3. All influential sources are speciated.
4. Source compositions are linearly independent.
5. The number of sources does not exceed the number of chemical species.
6. Measurement uncertainties are random, uncorrelated, and normally distributed.

Because we never expect all of these assumptions to be met in practice (Watson et al., 1991), several studies have evaluated the sensitivity of the CMB model to violations of the assumptions, including Henry (1982), Currie et al. (1984), Javitz et al. (1988a, b), and others.

In this study we pay particular attention to the following violations: non-constant source compositions (Assumption 1), correlated source compositions (Assumption 4), non-zero measurement error correlations (Assumption 6), non-zero correlations within a profile across species and within a species across profiles (Assumption 6), and lognormally distributed ambient measurement errors and source profile errors (Assumption 6).

In Section 2, we consider the CMB equations from the perspective of the traditional measurement error model in the statistical science literature (see e.g., Fuller, 1987) and we discuss various estimators of the pollution source contributions. Section 3 contains simulation experiments used to compare the competing estimators and associated inferential procedures. In Section 4, we compare the estimators when applied to the Fresno $PM_{2.5}$ data of Chow et al. (1992). The final section contains some conclusions and remarks.

2. CMB equations as a measurement error model

Because both the ambient measures and the source profiles are subject to measurement error, we view the CMB equations as a measurement error model. To avoid confusion, we refer to the measurement error associated with the ambiently measured concentrations as "measurement error" and we refer to the error associated with the source profile estimates as "source profile error" or just "profile error." In most situations, it is believed that profile error is much larger and more influential than the measurement error at the receptor (see e.g., Javitz et al., 1988b).

We define "diag($a_1,\ldots,a_n$)" to be a matrix with $a_1,\ldots,a_n$ as its diagonal elements and 0 s as the off-diagonal elements. Then, for the model considered below, we use the following notation:

ψ_i unobservable actual mass of species i at the receptor

V_i measurement error in the measurement of species i at the receptor

Σ_{vv} variance–covariance matrix for $\mathbf{v}=(v_1,\ldots,v_p)'$. If the elements of $\mathbf{v}$ are uncorrelated, then $\Sigma_{vv}=\mathrm{diag}(\mathrm{var}(v_1),\ldots,\mathrm{var}(v_p))=\mathrm{diag}(\sigma_{vv11},\ldots,\sigma_{vvpp})$

y_I measured mass of species i at the receptor (so that $y_i=\psi_i+v_i$)

q_I mass of species i that is not accounted for by all sources in the model (model error)

σ_{qq} var(q_i)

π_{Ij} actual source profile (unobservable proportional contribution to species i of source j)

u_{ij} source profile error (measurement error in the measured profile)

Σ_{uuii} Variance–covariance matrix for $\mathbf{u}_i=(u_{i1},\ldots,u_{ik})'$. If the elements of $\mathbf{u}_i$ exhibit no "across profile correlation" (that is, the elements are uncorrelated), then we have $\Sigma_{uuii} = \mathrm{diag}(\mathrm{var}(u_{i1}),\ldots,\mathrm{var}(u_{ik})) = \mathrm{diag}(\sigma_{uuii11},\ldots,\sigma_{uuiikk})$

Σ_{uu} Variance–covariance matrix for $\mathbf{u}=(\mathbf{u}_1',\ldots,\mathbf{u}_p')'$. If the elements of $\mathbf{u}_i$ are uncorrelated with the elements of $\mathbf{u}_{i'}$ for all $i\neq i'$ (i.e., no "within profile correlation" is exhibited), then $\Sigma_{uu} = \mathrm{diag}(\Sigma_{uu11},\ldots,\Sigma_{uupp})$ or

$$\Sigma_{uu} = \begin{bmatrix} \Sigma_{uu11} & & \mathbf{0} \\ & \ddots & \\ \mathbf{0} & & \Sigma_{uupp} \end{bmatrix}$$

Σ_{uvii} $(\mathrm{cov}(u_{i1},v_i),\ldots,\mathrm{cov}(u_{ik},v_i))'$

x_{Ij} measured source profile (so that $x_{ij}=\pi_{ij}+u_{ij}$)

β_j magnitude (mass) of source contribution j

The "measurement-error-free model" is written

$$\psi_i = \pi_{i1}\beta_1 + \cdots + \pi_{ik}\beta_k + q_i, \quad i=1,\ldots,p, \tag{3}$$

where the model errors q_i are normally and identically distributed (NID) with zero mean and variance σ_{qq} (i.e., $q_i \sim \mathrm{NID}(0,\sigma_{qq})$). The "observable model" is

UNCORRECTED PROOF

$$y_i = x_{i1}\beta_1 + \cdots + x_{ik}\beta_k + e_i, \quad i = 1, \ldots, p, \tag{4}$$

but since $y_i = \psi_i + v_i$ and $x_{ij} = \pi_{ij} + u_{ij}$, the error e_i in Eq. (4) is

$$e_i = q_i + v_i + \sum_{j=1}^{k} u_{ij}\beta_j, \quad i = 1, ..., p. \tag{5}$$

Note that because of source profile error, x_{ij} is stochastic and e_i and x_{ij} are correlated.

2.1. Ordinary least-squares solution

The ordinary least-squares (OLS) solution to the measurement error model requires the special case of Eq. (4) that is the classical linear statistical model. Also called "unweighted least squares," OLS was first used to solve the CMB equations by Miller et al. (1972) and Winchester and Nifong (1971). In addition to the assumptions of model (4), OLS also requires:

Assumption 1. $u_{ij} = 0$ for all *i,j* (i.e., there are no source profile errors so that $x_{ij} = \pi_{ij}$ is known and fixed for all *i,j*).

Assumption 2. $e_i = q_i + v_i$,

Assumption 3. e_i is independent of $e_{i'}$ for all $i \neq i'$.

Assumption 4. $\mathrm{var}(e_i) = \sigma_{ee}$ for all i , where σ_{ee} is known.

Under these restrictive assumptions, the best linear unbiased estimate of $\boldsymbol{\beta} = (\beta_1, \beta_2, \ldots, \beta_k)$ is

$$\hat{\boldsymbol{\beta}}_{\mathrm{OLS}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}. \tag{6}$$

Because Assumptions 1 and 4 are violated in virtually all receptor modeling studies, the OLS estimator (6) is biased and inefficient. Further, the standard errors associated with the estimates are incorrect when these assumptions are violated.

2.2. Weighted least-squares solution

We can relax Assumption 4 in Section 2.1 to obtain

Assumption 4′. $\mathrm{var}(e_i) = \sigma_{eeii}$ where σ_{eeii} is known for all i

Then, under Assumptions 1, 2, 3 and 4′, the best linear unbiased estimate of $\boldsymbol{\beta} = (\beta_1, \beta_2, \ldots, \beta_k)$ is the weighted least-squares (WLS) solution:

$$\hat{\boldsymbol{\beta}}_{\mathrm{WLS}} = (\mathbf{X}'\Sigma_{ee}^{-1}\mathbf{X})^{-1}\mathbf{X}'\Sigma_{ee}^{-1}\mathbf{y},$$

where $\boldsymbol{\Sigma}_{ee} = \mathrm{diag}(\sigma_{ee11}, \ldots, \sigma_{eepp})$ (a diagonal matrix with $\sigma_{ee11}, \ldots, \sigma_{eepp}$ as its diagonal elements). In practice, we use

$$\hat{\boldsymbol{\beta}}_{\mathrm{WLS}} = (\mathbf{X}'\boldsymbol{\Sigma}_{vv}^{-1}\mathbf{X})^{-1}\mathbf{X}'\boldsymbol{\Sigma}_{vv}^{-1}\mathbf{y}, \tag{7}$$

where $\boldsymbol{\Sigma}_{vv} = \mathrm{diag}(\sigma_{vv11}, \ldots, \sigma_{vvpp})$. The use of WLS was first applied to the CMB problem by Friedlander (1973). Because Assumption 1 is usually violated, the WLS estimator (7) still suffers from bias, but is more efficient than the OLS estimator. Additionally, estimated standard errors based on

$$\mathrm{var}(\hat{\boldsymbol{\beta}}_{\mathrm{WLS}}) = (\mathbf{X}'\boldsymbol{\Sigma}_{vv}^{-1}\mathbf{X})^{-1}$$

are in general too small in receptor modeling studies because of the existence of source profile error.

Because the WLS estimator has some attractive properties in terms of computational stability, one might consider a procedure which employs Eq. (7) to estimate $\boldsymbol{\beta}$, but then incorporates the variability due to source profile error into the standard error for $\hat{\boldsymbol{\beta}}$. Assuming that the model errors $q_i = 0$ for all species measurements, the variance of e_i in Eq. (5) can be estimated as

$$\tilde{\sigma}_{eeii} = \sigma_{vvii} + \hat{\boldsymbol{\beta}}'_{\mathrm{WLS}}\boldsymbol{\Sigma}_{uuii}\hat{\boldsymbol{\beta}}_{\mathrm{WLS}}, \quad i = 1, ..., p. \tag{8}$$

Then assuming that all measurement errors and source profile errors are uncorrelated across species, the variance–covariance matrix for **e** is estimated as

$$\tilde{\boldsymbol{\Sigma}}_{ee} = \mathrm{diag}(\tilde{\sigma}_{ee11}, ..., \tilde{\sigma}_{eepp}). \tag{9}$$

and adjusted standard errors can be found using the "sandwich" formula

$$\mathrm{var}(\hat{\boldsymbol{\beta}}_{\mathrm{WLS}}) = (\mathbf{X}'\boldsymbol{\Sigma}_{vv}^{-1}\mathbf{X})^{-1}\mathbf{X}'\boldsymbol{\Sigma}_{vv}^{-1}\tilde{\boldsymbol{\Sigma}}_{ee}\boldsymbol{\Sigma}_{vv}^{-1}\mathbf{X}(\mathbf{X}'\boldsymbol{\Sigma}_{vv}^{-1}\mathbf{X})^{-1}. \tag{10}$$

2.3. Britt and Luecke solution

Britt and Luecke (1973) consider the problem of estimating $\boldsymbol{\beta}$ from the perspective of a nonlinear measurement error model

$$f(\boldsymbol{\xi}, \boldsymbol{\beta}) = 0, \tag{11}$$

where

$$\boldsymbol{\xi} = \begin{pmatrix} \boldsymbol{\psi} \\ \boldsymbol{\pi} \end{pmatrix}, \boldsymbol{\psi} = \begin{pmatrix} \psi_1 \\ \vdots \\ \psi_p \end{pmatrix}, \boldsymbol{\pi} = \begin{pmatrix} \boldsymbol{\pi}_1 \\ \vdots \\ \boldsymbol{\pi}_p \end{pmatrix},$$

$$\text{and } \boldsymbol{\xi}_i = \begin{pmatrix} \psi_i \\ \boldsymbol{\pi}_i \end{pmatrix}.$$

Note that because $\boldsymbol{\pi}_i$ is of length k, $\boldsymbol{\pi}$ is a column vector of length pk, $\boldsymbol{\xi}_i$ is a column vector of length $(k + 1)$ and ξ is a column vector of length $pk{+}p$. If we assume no model error ($q_i = 0$ for all i), then in the context of our problem, we can write the model as

$$f(\boldsymbol{\xi}_i, \boldsymbol{\beta}) = \psi_i - \sum_{j=1}^{k} \pi_{ij}\beta_j = \boldsymbol{\xi}_i'\boldsymbol{\alpha} = 0,$$

$$f(\boldsymbol{\xi}, \boldsymbol{\beta}) = \boldsymbol{\psi} - \boldsymbol{\Pi}\boldsymbol{\beta} = 0, \tag{12}$$

where

UNCORRECTED PROOF

UNCORRECTED PROOF

$$\boldsymbol{\alpha}=\begin{pmatrix}1\\-\boldsymbol{\beta}\end{pmatrix},\boldsymbol{\psi}=\begin{pmatrix}\psi_1\\\vdots\\\psi_p\end{pmatrix},\boldsymbol{\pi}_i'=(\pi_{i1},\ldots,\pi_{ik}),$$

$$\text{and }\boldsymbol{\Pi}=\begin{bmatrix}\boldsymbol{\pi}_1'\\\vdots\\\boldsymbol{\pi}_p'\end{bmatrix}=\left[\boldsymbol{\pi}_{(1)},\boldsymbol{\pi}_{(2)},\ldots,\boldsymbol{\pi}_{(k)}\right].$$

Note that Eq. (12) can be viewed as a nonlinear model in the sense that both π_{ij} and β_j must be estimated.

The observed model can be written as

$$\mathbf{z}=\begin{pmatrix}\mathbf{y}\\\mathbf{x}\end{pmatrix}=\begin{pmatrix}\boldsymbol{\psi}+\mathbf{v}\\\boldsymbol{\pi}+\mathbf{u}\end{pmatrix}\sim N\{\boldsymbol{\xi},\boldsymbol{\Sigma}_{zz}\}$$
$$=N\left\{\begin{pmatrix}\boldsymbol{\Pi\beta}\\\boldsymbol{\pi}\end{pmatrix},\begin{bmatrix}\boldsymbol{\Sigma}_{vv}&\boldsymbol{\Sigma}_{vu}\\\boldsymbol{\Sigma}_{uv}&\boldsymbol{\Sigma}_{uu}\end{bmatrix}\right\},\tag{13}$$

where $\mathbf{y}$, and ψ, and $\boldsymbol{\pi}$ are defined as above,

$$\mathbf{v}=\begin{pmatrix}v_1\\\vdots\\v_p\end{pmatrix},\mathbf{x}=\begin{pmatrix}\mathbf{x}_1\\\vdots\\\mathbf{x}_p\end{pmatrix},\mathbf{u}=\begin{pmatrix}\mathbf{u}_1\\\vdots\\\mathbf{u}_p\end{pmatrix},$$

and

$$\mathbf{U}=\begin{bmatrix}\mathbf{u}_1'\\\vdots\\\mathbf{u}_p'\end{bmatrix}=\left[\mathbf{u}_{(1)},\mathbf{u}_{(2)},\ldots,\mathbf{u}_{(k)}\right]$$

is the $p\times k$ source profile error matrix. The $\mathbf{u}$ and $\mathbf{x}$ vectors have the same length (pk) as the $\boldsymbol{\pi}$ vector described after Eq. (11).

For the model presented in Eqs. (12) and (13) we can omit Assumptions 1, 3 and 4 given in Section 2.1 and alter Assumption 2 to obtain the following assumption in addition to those associated with model (4):

Assumption 2′. There is no model error ($q_i=0$).

Because we account for source profile error and allow for a more general form for measurement errors, the model used by Britt and Luecke is substantially less restrictive than the ordinary and generalized least-squares models.

The Britt and Luecke algorithm attempts to choose vectors $\boldsymbol{\beta}$ and $\boldsymbol{\xi}$ which minimize

$$(\mathbf{z}-\boldsymbol{\xi})'\boldsymbol{\Sigma}_{zz}^{-1}(\mathbf{z}-\boldsymbol{\xi})\tag{14}$$

subject to constraint (12). We linearize the problem by writing

$$f(\boldsymbol{\xi},\boldsymbol{\beta})\approx f(\tilde{\boldsymbol{\xi}},\tilde{\boldsymbol{\beta}})+F_{\boldsymbol{\beta}}(\tilde{\boldsymbol{\xi}},\tilde{\boldsymbol{\beta}})(\tilde{\boldsymbol{\beta}}-\boldsymbol{\beta})+F_{\boldsymbol{\xi}}(\tilde{\boldsymbol{\xi}},\tilde{\boldsymbol{\beta}})(\tilde{\boldsymbol{\xi}}-\boldsymbol{\xi})$$

where $F_{\boldsymbol{\beta}}(\boldsymbol{\xi},\boldsymbol{\beta})$ and $F_{\boldsymbol{\xi}}(\boldsymbol{\xi},\boldsymbol{\beta})$ denote the first derivatives of $f(\boldsymbol{\xi},\boldsymbol{\beta})=\psi-\boldsymbol{\Pi\beta}$ with respect to $\boldsymbol{\beta}$ and $\boldsymbol{\xi}$. Note that because $f(\boldsymbol{\xi},\boldsymbol{\beta})$ is a vector of length p, the derivatives $F_{\boldsymbol{\beta}}(\boldsymbol{\xi},\boldsymbol{\beta})$ and $F_{\boldsymbol{\xi}}(\boldsymbol{\xi},\boldsymbol{\beta})$ are matrices of dimension $p\times k$ and $p\times(pk+p)$, respectively. The derivatives are calculated

$$F_{\boldsymbol{\beta}}=\left[\frac{\partial f(\boldsymbol{\xi},\boldsymbol{\beta})}{\partial\boldsymbol{\beta}'}\right]=-\boldsymbol{\Pi}$$
$$\frac{\partial f(\boldsymbol{\xi}_i,\boldsymbol{\beta})}{\partial\boldsymbol{\xi}}=\frac{\partial(\psi_i-\boldsymbol{\pi}_i'\boldsymbol{\beta})}{\partial\boldsymbol{\xi}}=\boldsymbol{\alpha}=\begin{pmatrix}1\\-\boldsymbol{\beta}\end{pmatrix}$$

and

$$F_{\boldsymbol{\xi}}=\left[\frac{\partial f(\boldsymbol{\xi},\boldsymbol{\beta})}{\partial\boldsymbol{\xi}'}\right]$$
$$=\begin{bmatrix}1&0&\cdots&0&-\boldsymbol{\beta}'&\mathbf{0}'&\cdots&\mathbf{0}'\\0&1&\ddots&0&\mathbf{0}'&-\boldsymbol{\beta}'&\ddots&\mathbf{0}'\\\vdots&\ddots&\ddots&\vdots&\vdots&\ddots&\ddots&\vdots\\0&0&\cdots&1&\mathbf{0}'&\mathbf{0}'&\cdots&-\boldsymbol{\beta}'\end{bmatrix}$$
$$=\left[\mathbf{I}_p-(\mathbf{I}_p\otimes\boldsymbol{\beta}')\right].$$

The constrained estimation carried out by choosing $\boldsymbol{\beta}$, $\boldsymbol{\xi}$, and $\boldsymbol{\lambda}$ (a vector of p Lagrangian multipliers) to minimize

$$(\mathbf{z}-\boldsymbol{\xi})'\boldsymbol{\Sigma}_{zz}^{-1}(\mathbf{z}-\boldsymbol{\xi})$$
$$+\boldsymbol{\lambda}'\{f(\tilde{\boldsymbol{\xi}},\tilde{\boldsymbol{\beta}})+F_{\boldsymbol{\beta}}(\tilde{\boldsymbol{\xi}},\tilde{\boldsymbol{\beta}})(\tilde{\boldsymbol{\beta}}-\boldsymbol{\beta})+F_{\boldsymbol{\xi}}(\tilde{\boldsymbol{\xi}},\tilde{\boldsymbol{\beta}})(\tilde{\boldsymbol{\xi}}-\boldsymbol{\xi})\},$$

where $\tilde{\boldsymbol{\beta}}$ and $\tilde{\boldsymbol{\xi}}$ are estimated values. The solution can be obtained using an iteratively reweighted generalized least-squares algorithm where the estimates at the $(m+1)$th step are

$$\tilde{\boldsymbol{\beta}}^{(m+1)}=\tilde{\boldsymbol{\beta}}-\left[\tilde{F}'_{\boldsymbol{\beta}}\left\{\tilde{\boldsymbol{\Sigma}}_{ee}^{(m)}\right\}^{-1}\tilde{F}_{\boldsymbol{\beta}}\right]^{-1}\tilde{F}_{\boldsymbol{\beta}}'\left\{\tilde{\boldsymbol{\Sigma}}_{ee}^{(m)}\right\}^{-1}$$
$$\times\left\{\mathrm{f}(\tilde{\boldsymbol{\xi}},\tilde{\boldsymbol{\beta}})+\tilde{F}_{\boldsymbol{\xi}}(\mathbf{z}-\tilde{\boldsymbol{\xi}})\right\}$$
$$\tilde{\boldsymbol{\xi}}^{(m+1)}=\mathbf{z}-\boldsymbol{\Sigma}_{zz}\tilde{F}_{\boldsymbol{\xi}}'\left\{\tilde{\boldsymbol{\Sigma}}_{ee}^{(m)}\right\}^{-1}$$
$$\times\left\{f(\tilde{\boldsymbol{\xi}},\tilde{\boldsymbol{\beta}})+\tilde{F}_{\boldsymbol{\beta}}\left(\tilde{\boldsymbol{\beta}}^{(m+1)}-\tilde{\boldsymbol{\beta}}\right)+\tilde{F}_{\boldsymbol{\xi}}(\mathbf{z}-\tilde{\boldsymbol{\xi}})\right\}$$
$$\tilde{\boldsymbol{\Sigma}}_{ee}^{(m+1)}=\tilde{F}_{\boldsymbol{\xi}}\boldsymbol{\Sigma}_{zz}\tilde{F}_{\boldsymbol{\xi}}'\tag{15}$$

where $\tilde{\boldsymbol{\beta}}=\tilde{\boldsymbol{\beta}}^{(m)},\tilde{F}_{\boldsymbol{\beta}}=\tilde{F}_{\boldsymbol{\beta}}\left(\tilde{\boldsymbol{\xi}}^{(m)},\tilde{\boldsymbol{\beta}}^{(m)}\right)$, and so forth. In essence, the algorithm alternates between updating the source contribution estimates and the source profiles. We refer to the final estimates $\tilde{\boldsymbol{\beta}}^{(m+1)}$ and $\tilde{\boldsymbol{\Sigma}}_{ee}^{(m+1)}$ as $\hat{\boldsymbol{\beta}}_{\mathrm{BL}}$ and $\hat{\boldsymbol{\Sigma}}_{ee\mathrm{BL}}$. In the US Environmental Protection Agency's EPA-CMB8.2 program (EPA, 2000), the "Britt and Luecke" option is a simplified version of the algorithm which uses only the diagonal elements of $\tilde{\boldsymbol{\Sigma}}_{ee}^{(m+1)}$ as described by Watson et al. (1984). In an effort to mimic the output of EPA-CMB8.2, we use only the diagonal elements of $\tilde{\boldsymbol{\Sigma}}_{ee}^{(m+1)}$. Statistical inference is carried out using

$$\mathrm{var}(\hat{\boldsymbol{\beta}}_{\mathrm{BL}})=(\mathbf{X}'\hat{\boldsymbol{\Sigma}}_{ee\mathrm{BL}}^{-1}\mathbf{X})^{-1}.\tag{16}$$

2.4. Effective variance solution

We can further simplify the Britt and Luecke solutions if we begin with model (4) and are willing to make the following additional assumptions:

UNCORRECTED PROOF

Assumption 2′. There is no model error ($q_i = 0$).

Assumption 5. Species errors are uncorrelated ($\boldsymbol{\Sigma}_{vv} = \text{diag}(\sigma_{vv11}, ..., \sigma_{vvpp})$).

Assumption 6. Measurement errors and profile errors are uncorrelated ($\boldsymbol{\Sigma}_{uv} = \mathbf{0}$).

Assumption 7. Profile errors are uncorrelated across columns of $\mathbf{U}$, that is, $\boldsymbol{\Sigma}_{uu} = \text{diag}(\boldsymbol{\Sigma}_{uu11}, ..., \boldsymbol{\Sigma}_{uupp})$. We say that the data exhibit no "across profile error correlation (APC)."

Assumption 8. Profile errors are uncorrelated within columns of $\mathbf{U}$, that is, $\boldsymbol{\Sigma}_{uuii} = \text{diag}(\sigma_{uuii11}, ..., \sigma_{uuiikk})$. We say that the data exhibit no "within profile error correlation (WPC)."

The "effective variance (EV)" approach was employed by Watson et al. (1984) and implemented in software such as the US Environmental Protection Agency's EPA-CMB8.2 program (EPA, 2000). This approach uses an algorithm that further simplifies the Britt and Luecke estimation in order to achieve more stable parameter estimates. Specifically, instead of updating $\tilde{\boldsymbol{\xi}}^{(m)}$ at each iteration of the algorithm, we fix $\tilde{\boldsymbol{\xi}}^{(m)} = \mathbf{z}$ for all iterations. The effective variance solution is obtained from an iteratively reweighted least-squares algorithm that employs weights that are functions of both measurement and source profile errors. At the $(m+1)$st iteration, we update

$$\tilde{\boldsymbol{\beta}}^{(m+1)} = \left[\mathbf{X}'\left\{\tilde{\boldsymbol{\Sigma}}_{ee}^{(m)}\right\}^{-1}\mathbf{X}\right]^{-1}\mathbf{X}'\left\{\tilde{\boldsymbol{\Sigma}}_{ee}^{(m)}\right\}^{-1}\mathbf{y} \qquad (17)$$

and

$$\tilde{\boldsymbol{\Sigma}}_{ee}^{(m+1)} = \text{diag}\left(\sigma_{vvii} + \sum_{j=1}^{k}\sigma_{uuiijj}\left(\tilde{\boldsymbol{\beta}}_j^{(m+1)}\right)^2\right). \qquad (18)$$

We refer to the final estimates $\tilde{\boldsymbol{\beta}}^{(m+1)}$ and $\tilde{\boldsymbol{\Sigma}}_{ee}^{(m+1)}$ as $\hat{\boldsymbol{\beta}}_{\text{EV}}$ and $\hat{\boldsymbol{\Sigma}}_{ee\text{EV}}$. Statistical inference is carried out using

$$\text{var}(\hat{\boldsymbol{\beta}}_{\text{EV}}) = (\mathbf{X}'\hat{\boldsymbol{\Sigma}}_{ee\text{EV}}^{-1}\mathbf{X})^{-1}. \qquad (19)$$

2.5. Method of moments solution

The final solution we consider in depth is a method of moments solution to the measurement error model discussed by Fuller (1987, p. 193–194). The method of moments solution allows us to relax Assumptions 2′, 6 and 8 given in Section 2.4. Thus, in addition to the assumptions required for model (4), we require only

Assumption 5. Species errors are uncorrelated ($\boldsymbol{\Sigma}_{vv} = \text{diag}(\sigma_{vv11}, ..., \sigma_{vvpp})$).

Assumption 7. Profile errors are uncorrelated across columns of $\mathbf{U}$, that is, $\boldsymbol{\Sigma}_{uu} = \text{diag}(\boldsymbol{\Sigma}_{uu11}, ..., \boldsymbol{\Sigma}_{uupp})$. We say that the data exhibit no "across profile error correlation (APC)."

The solution is obtained by updating the following at each iteration of the algorithm:

$$\tilde{\boldsymbol{\beta}}^{(m+1)} = \left[\sum_{i=1}^{p}\tilde{\sigma}_{eeii}^{(m)}\left(\mathbf{x}_i\mathbf{x}_i' - \boldsymbol{\Sigma}_{uuii}\right)\right]^{-1} \left\{\sum_{i=1}^{p}\tilde{\sigma}_{eeii}^{(m)}\left(\mathbf{x}_i\mathbf{y}_i - \boldsymbol{\Sigma}_{uvii}\right)\right\}, \qquad (20)$$

$$\tilde{\sigma}_{eeii}^{(m+1)} = \sigma_{vvii} + \sigma_{qq} + \sum_{j=1}^{\text{k}}\sigma_{uuiijj}\left(\tilde{\beta}_j^{(m+1)}\right)^2, \quad i = 1, ..., p \qquad (21)$$

We refer to the final estimates $\tilde{\boldsymbol{\beta}}^{(m+1)}$ and $\tilde{\sigma}_{eeii}^{(m+1)}$ as $\hat{\boldsymbol{\beta}}_{\text{MM}}$ and $\hat{\sigma}_{ee\text{MM}}$. Statistical inference is carried out using

$$\begin{aligned}\text{var}(\hat{\boldsymbol{\beta}}_{\text{MM}}) = &\left[\sum_{i=1}^{p}\hat{\sigma}_{ee\text{MM}}(\mathbf{x}_i\mathbf{x}_i' - \boldsymbol{\Sigma}_{uuii})\right]^{-1} \\ &\times\left[\sum_{i=1}^{p}\frac{1}{\hat{\sigma}_{ee\text{MM}}}\left(\mathbf{x}_i\mathbf{x}_i' + \frac{1}{\hat{\sigma}_{ee\text{MM}}}\boldsymbol{\Sigma}_{uvii}\boldsymbol{\Sigma}_{uvii}'\right)\right] \\ &\times\left[\sum_{i=1}^{p}\hat{\sigma}_{ee\text{MM}}(\mathbf{x}_i\mathbf{x}_i' - \boldsymbol{\Sigma}_{uuii})\right]^{-1}. \qquad (22)\end{aligned}$$

If we assume that profile errors are uncorrelated with measurement errors, the above simplifies to

$$\begin{aligned}\text{var}(\hat{\boldsymbol{\beta}}_{\text{MM}}) = &\left[\sum_{i=1}^{p}\hat{\sigma}_{ee\text{MM}}(\mathbf{x}_i\mathbf{x}_i' - \boldsymbol{\Sigma}_{uuii})\right]^{-1} \\ &\times\left[\sum_{i=1}^{p}\frac{1}{\hat{\sigma}_{ee\text{MM}}}(\mathbf{x}_i\mathbf{x}_i')\right] \\ &\times\left[\sum_{i=1}^{p}\hat{\sigma}_{ee\text{MM}}(\mathbf{x}_i\mathbf{x}_i' - \boldsymbol{\Sigma}_{uuii})\right]^{-1}. \qquad (23)\end{aligned}$$

2.6. Other solutions

The solutions to the CMB equations discussed above constitute only a subset of the approaches that have been (or could be) applied to the source apportionment problem. Other approaches have included exploratory factor analysis (e.g., Thurston and Spengler, 1985; Henry et al., 1994), confirmatory factor analysis (e.g., Gleser, 1997; Christensen and Sain, 2002), positive matrix factorization (Paatero and Tapper, 1994), Unmix analysis (Henry, 1997), and Bayesian analysis (e.g., Park et al., 2001, 2002). These approaches are not directly comparable to the measurement error model solutions discussed in Sections 2.1–2.5 in that they do not utilize (or only partially utilize) the a priori information about the source profiles and profile uncertainties. An approach that is related to the measurement error model solutions is the orthogonal regression approach, called "total least squares (TLS)" by Golub and Van Loan (1980). The TLS approach can be attractive because, like the solutions discussed in Sections 2.3–2.5, it accounts for errors in both the measurements of the responses and the profiles. However, while the statistical model associated with TLS recognizes the existence of the profile errors, the TLS solution does not utilize any a priori information about the magnitude of profile or receptor measurement uncertainties.

3. Comparison of CMB solutions

In this section, we evaluate the statistical properties of the class of source contribution estimators that utilize a priori information about both source profiles and receptor measurement uncertainties. This most frequently employed class of estimators includes the WLS estimator ($\hat{\boldsymbol{\beta}}_{\mathrm{WLS}}$) from Section 2.2, the Britt and Luecke estimator ($\hat{\boldsymbol{\beta}}_{\mathrm{BL}}$) from Section 2.3, the effective variance estimator ($\hat{\boldsymbol{\beta}}_{\mathrm{EV}}$) from Section 2.4, and the method of moments estimator ($\hat{\boldsymbol{\beta}}_{\mathrm{MM}}$) from Section 2.5. Cheng et al. (1988) compared the estimators $\hat{\boldsymbol{\beta}}_{\mathrm{WLS}}$ and $\hat{\boldsymbol{\beta}}_{\mathrm{EV}}$ for scenarios with relatively low levels of receptor measurement error and profile error (i.e., noise-to-signal levels in the 5–10% range). They concluded that for low levels of error, $\hat{\boldsymbol{\beta}}_{\mathrm{EV}}$ did not yield more accurate source contribution estimates than $\hat{\boldsymbol{\beta}}_{\mathrm{WLS}}$, even when the underlying normality assumption is valid.

In this section, we compare $\hat{\boldsymbol{\beta}}_{\mathrm{EV}}$ and $\hat{\boldsymbol{\beta}}_{\mathrm{WLS}}$ with the more sophisticated (but potentially more computationally instable) $\hat{\boldsymbol{\beta}}_{\mathrm{BL}}$ and $\hat{\boldsymbol{\beta}}_{\mathrm{MM}}$ via computer simulation. In our study, we consider simulated data that violate standard assumptions in a variety of ways including: non-constant source compositions, correlated source compositions, non-zero measurement error correlations, non-zero correlations within a profile across species and within a species across profiles, and lognormally distributed ambient measurement errors and source profile errors. In contrast to the work of Cheng et al. (1988), our simulation study includes scenarios in which the source profile uncertainties are allowed to be very large, with profile noise-to-signal ratios as high as 100%. To compare the estimators, we use simulated data based on source profiles given in Javitz et al. (1988b) and reproduced in Table 1. For our first set of simulations, we consider an airshed influenced by four pollution sources: soil, a coal-fired power plant, vehicle exhaust, and wood burning. We use the first four columns of Table 1 as our assumed or estimated source profile matrix **X**.

Simulating data. Because pollution profiles tend to vary from day to day in the real world, we allow the true source profile matrix $\boldsymbol{\Pi}$ to vary between replications. The quantity π_{ij} (an element of $\boldsymbol{\Pi}$) follows a lognormal distribution with mean π_{ij} and coefficient of variation ("source profile CV") of $\mathrm{CV}_u \in (0, 200\%)$. Elements of $\boldsymbol{\Pi}$

Table 1
Pollution source profiles obtained from Javitz, Watson, and Robinson (1988)

Source contribution ($\mu g\,m^{-3}$)	Source types			
	Soil (geological) 5	Coal-fired power plant 5	Motor vehicle exhaust 5	Vegitative burning 5
Species	Source profiles			
OC	0.040000	0.010000	0.407000	0.475000
EC	0.005000	0.005000	0.222000	0.128000
Al	0.078300	0.146000	0.000400	0.000210
Si	0.305000	0.219000	0.001000	0.000000
Cl	0.000320	0.000520	0.022000	0.005090
K	0.028200	0.012200	0.000100	0.008600
Ca	0.028700	0.012100	0.000500	0.000670
Ti	0.004740	0.008700	0.000000	0.000000
V	0.000095	0.000680	0.000000	0.000000
Cr	0.000070	0.000620	0.000000	0.000000
Mn	0.000690	0.000430	0.001100	0.000000
Fe	0.035400	0.080900	0.001400	0.000000
Ni	0.000044	0.000390	0.000000	0.000000
Cu	0.000030	0.000290	0.000000	0.000000
Zn	0.000060	0.000670	0.001000	0.000370
Br	0.000003	0.000000	0.024300	0.000000
Pb	0.000015	0.000420	0.081500	0.000000
Mass of the 17 species ($\mu g\,m^{-3}$)	2.6333	2.4896	3.8115	3.0897

UNCORRECTED PROOF

UNCORRECTED PROOF

were generated such that $r_{\mathrm{APC}} = \mathrm{corr}(\log\{\pi_{ij}\}, \log\{\pi_{ij'}\})$ was equal to a specified "across profile correlation" value in the range (0, 0.6) for all $j \neq j'$. Similarly the "within profile correlation" defined by $r_{\mathrm{WPC}} = \mathrm{corr}(\log\{\pi_{ij}\}, log\{\pi_{i'j}\}) \in (0, 0.3)$ was specified for all $i \neq i'$.

Each source contribution in $\boldsymbol{\beta}$ was allowed to vary between replications following a lognormal distribution with mean of ($5\,\mu\mathrm{g\,m}^{-3}$, $5\,\mu\mathrm{g\,m}^{-3}$, $5\,\mu\mathrm{g\,m}^{-3}$, $5\,\mu\mathrm{g\,m}^{-3}$) and a coefficient of variation equal to 50%. Because source contributions are often intercorrelated, we generated contributions such that $r_\beta = \mathrm{corr}(\log\{\beta_j\}, \log\{\beta_{j'}\}) = 0.5$ for all $j \neq j'$.

Finally, the observed pollutant concentrations $\mathbf{y}$ was obtained such that y_i followed a lognormal distribution with mean of $\pi_i'\beta$ and a coefficient of variation ("species measurement error CV") of $\mathrm{CV}_v \in (0, 20\%)$. For some simulations we allow observed species concentrations to be correlated by setting $r_v = \mathrm{corr}(\log\{y_j\}, \log\{y_{i'}\})$ to a specified value in the range (0, 0.3) for all $i \neq i'$.

Fitting the model. When fitting the model, we use error covariance matrix estimates $\hat{\boldsymbol{\Sigma}}_{vv} = (\hat{\sigma}_{vvii'})$ and $\hat{\boldsymbol{\Sigma}}_{uuii} = (\hat{\sigma}_{uuiijj'})$, where the typical elements of these matrices are

$$\hat{\sigma}_{vvii'} = \begin{cases} (\mathrm{CV}_v)^2 y_i y_{i'} & \text{if } i = i' \\ (\mathrm{CV}_v)^2 y_i y_{i'} r_v & \text{if } i \neq i' \end{cases} \quad \text{for } i, i' = 1, \ldots, p,$$

$$\hat{\sigma}_{uuiijj'} = \begin{cases} (\mathrm{CV}_u)^2 x_{ij} x_{ij'} & \text{if } j = j' \\ (\mathrm{CV}_u)^2 x_{ij} x_{ij'} r_{\mathrm{APC}} & \text{if } j \neq j' \end{cases} \quad \text{for } j, j' = 1, \ldots, k.$$

Note that despite the fact that we have estimates $\hat{\boldsymbol{\Sigma}}_{vv}$ and $\hat{\boldsymbol{\Sigma}}_{uuii}$ which are not diagonal, only the method of moments approach of Section 2.5 actually makes full use of the off-diagonal elements of the matrices. The within profile correlation that we allow in some scenarios violates the assumptions for all the approaches considered. When using the method of moments approach, we assume that $\boldsymbol{\Sigma}_{uv} = \mathbf{0}$.

Comparing estimation and inference approaches. Using the data generation approach described above, 1000 experiments were simulated. For each day's data, the weighted least squares, effective variance, Britt and Luecke, and method of moments estimates of each pollution source were calculated and a 95% confidence interval was obtained using

$$\hat{\beta}_j \pm 1.96\sqrt{\hat{\mathrm{var}}(\hat{\beta}_j)}. \tag{24}$$

To evaluate the quality of the estimates obtained using each approach, the average absolute error (AAE) for each estimator was calculated and expressed as a percentage of the true source contribution. That is, for the jth source $\hat{\beta}_j$, we use the n replicates of the simulation study to calculate

$$\mathrm{AAE}_j = \frac{1}{n} \sum_{i=1}^{n} \left| \frac{\hat{\beta}_j - \beta_j}{\beta_j} \right| \times 100\%.$$

To evaluate the validity of the statistical inference associated with each approach, we calculated the empirical coverage probability for each approach by finding the percentage of the confidence intervals (24) that actually contains β_j. Ideally, the empirical coverage probability should be equal to the nominal value of 95%. Table 2 gives AAE and coverage probability averaged over all four sources. For example, the AAE reported for a given scenario is actually $(\mathrm{AAE}_1 + \mathrm{AAE}_2 + \mathrm{AAE}_3 + \mathrm{AAE}_4)/4$.

As discussed in Javitz et al. (1988b), the AAE for all estimators becomes larger as the source profile coefficient of variation (CV_u) increases. This increase in AAE is more dramatic for the more complex estimators. The four estimators are comparable in terms of both AAE and coverage when $\mathrm{CV}_u \leqslant 25\%$. The equivalence of the estimators for relatively small values of CV_u is in line with the conclusions of Cheng, Hopke, and Jennings (1988), but is extended here to the Britt and Luecke and

Table 2
AAE and empirical coverage probability for four solutions to the chemical mass balance equations. Absolute errors are expressed as a percentage of the true values. Empirical coverage probabilities for nominal 95% confidence intervals are expressed as percentages

Source profile CV (%)	Species ME CV (%)	Across profile Corr's	Within profile Corr's	Species ME Corr's	WLS		Effective variance		Britt and Luecke		Method of moments	
					AAE (%)	Coverage (%)	AAE	Coverage	AAE	Coverage	AAE	Coverage
10	10	0	0	0	10	93	10	94	10	94	10	95
10	10	0.3	0	0	11	94	10	93	10	93	11	94
10	10	0	0	0.3	10	92	10	92	10	92	11	93
10	10	0.3	0.3	0.3	10	92	11	91	11	91	11	92
25	20	0	0	0	24	85	23	87	23	89	25	94
25	20	0.3	0.3	0.3	23	88	23	87	24	88	27	91
50	20	0	0	0	37	73	37	75	42	79	2,914	96
50	20	0.3	0.3	0.3	34	80	35	79	40	82	1,277	95
100	20	0	0	0	55	54	61	54	524	65	484,068	83
100	20	0.3	0.3	0.3	53	65	56	62	423	70	22,878	76

UNCORRECTED PROOF

method of moments estimators. However, when $\mathrm{CV}_u \geqslant 50\%$ the WLS and effective variance (EV) solutions are superior to the other two, and when $\mathrm{CV}_u = 100\%$ the WLS solution performs slightly better than the EV solution. Although the $\hat{\boldsymbol{\beta}}_{\mathrm{BL}}$ and $\hat{\boldsymbol{\beta}}_{\mathrm{MM}}$ estimators require fewer assumptions than the more simplistic $\hat{\boldsymbol{\beta}}_{\mathrm{WLS}}$ and $\hat{\boldsymbol{\beta}}_{\mathrm{EV}}$, we see that the additional quantities that must be estimated in the Britt and Luecke and method of moments algorithms lends instability to the estimation process. For example, Eq. (15) in the Britt and Luecke algorithm contains three quantities that must be estimated at each iteration. With only 17 species, estimation of these quantities will be more volatile, particularly when the noise-to-signal ratios associated with the profile estimates are large. In Eq. (20) of the method of moments algorithm, we invert the sum

$$\sum_{i=1}^{p} \tilde{\sigma}^{(m)}_{\mathrm{ee}ii}\left(\mathbf{x}_i\mathbf{x}_i' - \boldsymbol{\Sigma}_{\mathrm{uu}ii}\right).$$

While this sum of $k \times k$ matrices is likely to have a stable inverse when the elements of $\boldsymbol{\Sigma}_{uuii}$ are small, the inverted sum will be increasingly volatile as the noise-to-signal ratio for the profile estimates increases. Thus the increased sophistication of the Britt and Luecke and method of moments solutions acts as a liability when profile estimates are associated with high degrees of uncertainty. While these estimators are generally valuable tools for scenarios in which the errors associated with the regressors (e.g., profiles) are relatively small, they are simply unable to perform well when such errors are as large as those we see in many CMB settings.

In addition to its estimation properties, which are comparable to those of the effective variance estimator, the adjusted standard error of $\hat{\boldsymbol{\beta}}_{\mathrm{WLS}}$ obtained from Eq. (10) yields coverage probabilities that are roughly as good or better than the coverage probabilities associated with any of the other estimators. Note that for most of the scenarios considered, the presence of profile error and measurement error correlations had little or no discernable effect on the AAE or coverage. The only exception to this pattern was the performance of the method of moments estimator when profile errors were large ($\mathrm{CV}_u \geqslant 50\%$). In such cases, the AAE for the method of moments solution was lower with correlated errors, but so were the empirical coverage probabilities.

4. Analysis of Fresno $\mathrm{PM}_{2.5}$ data

We consider here the Fresno, California $\mathrm{PM}_{2.5}$ data from the San Joaquin Valley Air Quality Study (see Chow et al., 1992), available as a test case with the EPA-CMB8.2 program (EPA, 2000). These data were collected every 6 days between June 1988 and June 1989. After removing missing values, we have a total of 35 observations for the Fresno site. Mimicking the analysis of Chow et al. (1992), we use six source profiles provided by the authors, including: road dust, wood burning, crude oil combustion, motor vehicles, ammonium sulfate, and ammonium nitrate. We begin by using each of the four approaches (WLS, effective variance, Britt and Luecke, and method of moments) to estimate the source contribution estimate for each source and day. Fig. 1 gives a plot of the source contribution estimates associated with the WLS, effective variance, and method of moments approaches (the Britt and Luecke estimates were dramatically different and are not shown). Although the solutions from the three methods are generally in agreement, we cannot say which apportionment is ‘‘correct.’’ We can however evaluate the expected absolute error associated with each solution which arises from the profile uncertainties.

We begin by acting as if the source profile matrix used for Fresno from the EPA-CMB8.2 program (EPA, 2000) and reproduced here in Table 3 is the ‘‘true’’ but unknown profile matrix ($\boldsymbol{\Pi}$) and the accompanying source contributions are the ‘‘true’’ contributions. We then simulate the real world scenario in which an ‘‘observed’’ profile matrix ($\mathbf{X}$) is only approximately equal to the true profile matrix. Each element of the observed profile matrix (x_{ij}) is a random draw from a lognormal distribution with mean equal to the corresponding element of the true profile matrix (π_{ij}) and a standard deviation of ($m \times \sigma_{uuiijj}$), where σ_{uuiijj} is the uncertainty associated with the (i,j) element of the Fresno source profile matrix. Thus m=1 would represent a scenario in which the estimated uncertainties for the source profiles available in EPA-CMB8.2 (EPA, 2000) are correct, and $m > 1$ would represent a scenario in which the estimated uncertainties are in fact larger than those reported—perhaps due to biased or non-representative measurements of the desired source.

After generating an ‘‘observed source profile matrix’’ ($\mathbf{X}$) from the ‘‘true’’ profiles ($\boldsymbol{\Pi}$) and profile uncertainties, we then estimated the source contribution estimates using each of the four approaches. Since the contributions from the original (‘‘true’’) profiles yielded four separate sets of contribution estimates and none of these can be declared the correct apportionment, we compare the WLS estimates obtained from the randomly generated profiles with the WLS estimates obtained from the ‘‘true’’ profiles, we compare the effective variance estimates obtained from the randomly generated profiles with the effective variance estimates obtained from the ‘‘true’’ profiles, and so forth. In this way, we are able to evaluate the stability of the various estimators in the presence of varying levels of source profile uncertainty.

For each of m=1, 2, and 3, we generated 200 different estimates of the source profile matrix, and calculated the

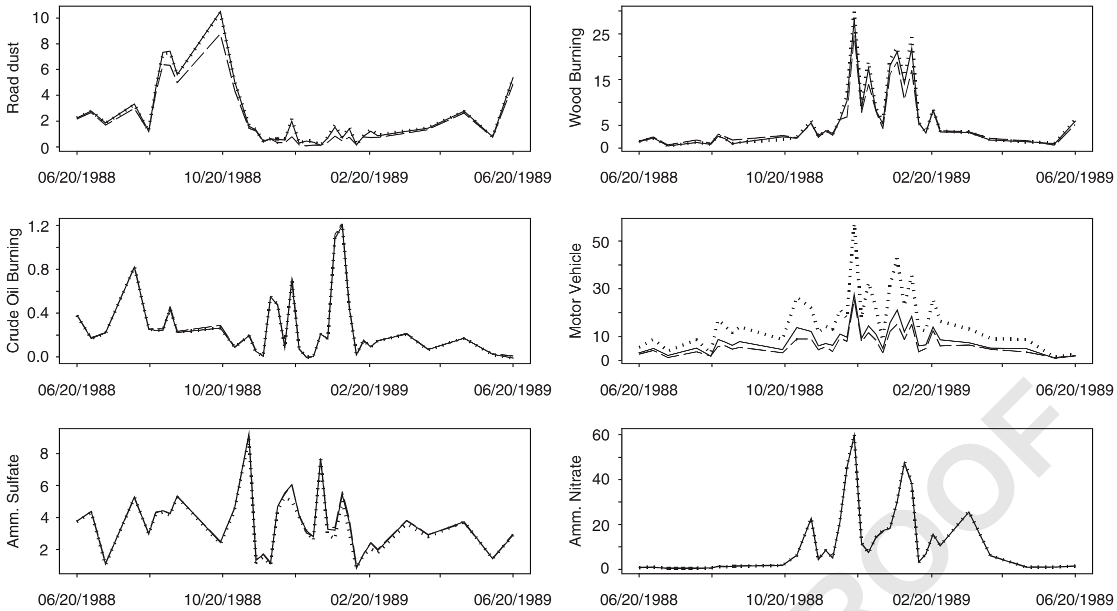

Fig. 1. Source contribution estimates for six sources in Fresno. Solid—effective variance solution, dashed—WLS solution, and dotted—method of moments solution.

UNCORRECTED PROOF

Table 3
Source profiles (± profile uncertainty) used for source apportionment of Fresno data from EPA (2000)

Species	Road dust	Wood burning	Crude oil combustion	Motor vehicles	Ammonium sulfate	Ammonium nitrate
NO_3	0.0011±0.0058	0.0046±0.0012	0.0000±0.0000	0.0200±0.0200	0.0000±0.0000	0.7750±0.0775
SO_4	0.0110±0.0183	0.0142±0.0042	0.2032±0.0424	0.0311±0.0355	0.7270±0.0727	0.0000±0.0000
NH_4	0.0001±0.0002	0.0009±0.0006	0.0001±0.0001	0.0000±0.0100	0.2730±0.0273	0.2255±0.0226
Soluble K	0.0052±0.0013	0.0399±0.0124	0.0006±0.0001	0.0000±0.0100	0.0000±0.0000	0.0000±0.0000
Soluble Na	0.0085±0.0088	0.0014±0.0005	0.0076±0.0040	0.0000±0.0100	0.0000±0.0000	0.0000±0.0000
EC	0.0157±0.0107	0.1589±0.0580	0.0000±0.0007	0.5415±0.1978	0.0000±0.0000	0.0000±0.0000
OC	0.1868±0.0224	0.4460±0.0794	0.0009±0.0012	0.4981±0.2415	0.0000±0.0000	0.0000±0.0000
Al	0.0731±0.0083	0.0000±0.0003	0.0000±0.0001	0.0008±0.0005	0.0000±0.0000	0.0000±0.0000
Si	0.1939±0.0220	0.0000±0.0001	0.0001±0.0002	0.0096±0.0139	0.0000±0.0000	0.0000±0.0000
S	0.0050±0.0006	0.0052±0.0018	0.0545±0.0039	0.0104±0.0118	0.2427±0.0243	0.0000±0.0000
Cl	0.0036±0.0004	0.0191±0.0064	0.0002±0.0000	0.0003±0.0002	0.0000±0.0000	0.0000±0.0000
K	0.0195±0.0022	0.0399±0.0124	0.0004±0.0001	0.0001±0.0001	0.0000±0.0000	0.0000±0.0000
Ca	0.0465±0.0053	0.0007±0.0006	0.0006±0.0000	0.0007±0.0008	0.0000±0.0000	0.0000±0.0000
Ti	0.0047±0.0005	0.0000±0.0002	0.0001±0.0000	0.0000±0.0000	0.0000±0.0000	0.0000±0.0000
V	0.0003±0.0001	0.0000±0.0001	0.0082±0.0006	0.0000±0.0000	0.0000±0.0000	0.0000±0.0000
Cr	0.0003±0.0000	0.0000±0.0000	0.0001±0.0002	0.0000±0.0000	0.0000±0.0000	0.0000±0.0000
Mn	0.0010±0.0001	0.0000±0.0000	0.0001±0.0000	0.0003±0.0002	0.0000±0.0000	0.0000±0.0000
Fe	0.0572±0.0065	0.0000±0.0000	0.0021±0.0002	0.0000±0.0001	0.0000±0.0000	0.0000±0.0000
Ni	0.0001±0.0000	0.0000±0.0000	0.0079±0.0009	0.0000±0.0000	0.0000±0.0000	0.0000±0.0000
Cu	0.0002±0.0000	0.0000±0.0000	0.0000±0.0000	0.0001±0.0000	0.0000±0.0000	0.0000±0.0000
Zn	0.0023±0.0002	0.0009±0.0004	0.0026±0.0003	0.0005±0.0003	0.0000±0.0000	0.0000±0.0000
Br	0.0001±0.0000	0.0001±0.0000	0.0000±0.0000	0.0026±0.0015	0.0000±0.0000	0.0000±0.0000
Pb	0.0020±0.0002	0.0000±0.0000	0.0000±0.0000	0.0037±0.0021	0.0000±0.0000	0.0000±0.0000

average absolute error (in $\mu g\,m^{-3}$) for each of the six sources and refer to this as "expected absolute error" (see Table 4). As expected from the simulation results in the previous section, the WLS and effective variance solutions perform substantially better than the other two solutions, and they have roughly comparable expected absolute errors when source profile errors are moderate (m=1). Further, as the source profile estimates become

Table 4
Expected absolute error (in $\mu g\,m^{-3}$) for each of the six sources in Fresno

m	Source	WLS	Effective variance	Britt and Luecke	Method of moments
1	Road dust	0.164	**0.149**	0.256	0.330
	Wood	**1.082**	1.204	**2.413**	2.710
	Crude oil	0.032	**0.025**	0.227	0.198
	Motor vehicles	**2.163**	2.167	274.653	213.447
	Amm. sulfate	0.292	**0.274**	5.952	2.852
	Amm. nitrate	0.846	**0.764**	5.461	5.356
2	Road dust	**0.279**	0.325	0.700	0.743
	Wood	**2.074**	2.910	27.879	78.868
	Crude oil	0.070	0.049	0.478	0.071
	Motor vehicles	**3.445**	5.620	839.082	59.906
	Amm. sulfate	**0.521**	0.536	14.216	2.879
	Amm. nitrate	1.658	**1.566**	14.796	2.441
3	Road dust	**0.393**	0.496	1.079	1.919
	Wood	**2.393**	4.648	428.512	113.500
	Crude oil	0.121	**0.075**	0.832	0.217
	Motor vehicles	**4.534**	11.056	1619.850	26.078
	Amm. sulfate	**0.703**	0.843	30.690	2.464
	Amm. nitrate	**2.231**	2.280	22.764	3.780

The constant *m* denotes the magnitude of the source profile errors. For each set of simulations, the source profile uncertainties are equal to *m* times the uncertainties reported in Chow et al. (1992). The smallest expected absolute error in each row is boldfaced.

more uncertain (have higher variability), the WLS estimator becomes more and more superior to the effective variance solution. When *m*=3, the effective variance solution yields better estimates for only crude oil combustion (the smallest of the six sources) and is dramatically inferior for the motor vehicle and wood burning sources.

In this simulation, we again see that the Britt and Luecke and method of moments solutions do poorly when the noise-to-signal ratio (or CV) is high. For example, even in the *m*=1 case, the median of the CV's (uncertainty/profile element value) associated with non-zero elements of the profile matrix is 0.6. For the *m*=2 and 3 cases, the median CVs associated with non-zero profile elements are 1.2 and 1.8, respectively. Note that while the simulation results in Section 3 indicate that the Britt and Luecke and method of moments solutions become computationally instable for noise-to-signal ratios near 100%, we see in this simulation that for even larger degrees of profile uncertainty, even the effective variance solution can begin breaking down. In the effective variance algorithm, the estimate of the effective variance in Eq. (18) depends heavily on the profile uncertainties and on the previous iteration's source contribution estimate in Eq. (17). We see in our simulations that for the case when the majority of the actual noise-to-signal ratios associated with the profile estimates are greater than 100%, the advantage of the flexibility of the effective variance weighting in Eq. (17) is overcome by the disadvantageous propensity for the iterative procedure to progressively move the source contribution estimate away from the true value. Because the WLS estimator is not a function of the source profile uncertainties, it is biased but unlikely to yield wildly implausible source contribution estimates. Thus, it can be the most practical option for scenarios in which profile uncertainties are very large.

5. Conclusions and remarks

The traditional chemical mass balance model is a special case of the measurement error model. The parameters of the model can be estimated by a variety of approaches, the choice of which depends on the assumptions made about the form of the ambient chemical species observations. As the number of restrictive assumptions placed on the model increases, the forms of the estimation and inference approaches become simpler. The method of moments solution found in Fuller (1987) and the Britt and Luecke (1973) solution allow for the most flexible measurement error models, but can perform poorly when the source profile errors are large. The gold standard for pollution source contribution estimation in the CMB setting has been the effective variance (EV) solution (Watson et al., 1984), in part because it incorporates source profile error and receptor measurement error into the estimation of

UNCORRECTED PROOF

both the pollution source contributions (β_i) and the standard errors of $\hat{\beta}_i$.

The most restrictive assumptions are required when using the weighted least-squares (WLS) estimator. Most importantly, the WLS estimator assumes that the source profile matrix is observed without error. This assumption of no source profile error is, of course, never true in practice. In fact, we suspect that many estimated source profile uncertainties for classes of polluters (such as automobiles or smelters) are often underestimated when they are based on samples of class members that do not accurately represent the wide variety of polluters in the class. Notwithstanding, we have proposed using the simple WLS estimator (7) for estimation because it is computationally stable and thus yields better average absolute error for scenarios in which the magnitude of source profile errors and measurement errors are large. Because the Britt and Luecke and method of moments estimators depend most heavily on the profile uncertainties, they perform poorly when profile noise-to-signal ratios are larger than 25%. In the simulation based on the $PM_{2.5}$ data from Fresno, California, the effective variance solution begins breaking down when profile noise-to-signal ratios exceed 100%. Because the WLS estimator is not a function of the source profile uncertainties, it is biased but is computationally stable. Thus, it has statistical properties which are comparable to the other estimators for small profile noise-to-signal ratios ($CV \leqslant 100\%$), and superior to the other estimators for large noise-to-signal ratios ($CV > 100\%$). For valid statistical inference associated with the WLS estimates, we have recommended standard errors based on the "sandwich formula" (10), which accounts for both source profile and measurement errors and yields coverage probabilities that are comparable to those obtained using the EV approach. Although the popular effective variance solution is an equally reasonable approach for most scenarios with small to moderate levels of uncertainty associated with the source profiles, the WLS approach with corrected standard errors is shown to be a more attractive all-around alternative, particularly when source profile uncertainties are large.

References

Britt, H.I., Luecke, R.H., 1973. The estimation of parameters in nonlinear, implicit models. Technometrics 15, 233–247.

Cheng, M.D., Hopke, P.K., Jennings, D.E., 1988. The effects of measurement errors, collinearity, and their interactions on aerosol source apportionment computations. Chemometrics and Intelligent Laboratory Systems 4, 239–250.

Chow, J.C., Watson, J.G., Lowenthal, D.H., Solomon, P.A., Magliano, K.L., Ziman, S.D., Richards, L.W., 1992. PM10 source apportionment in California's San Joaquin Valley. Atmospheric Environment 26A, 3335–3354.

Christensen, W.F., Sain, S.R., 2002. Accounting for dependence in a flexible multivariate receptor model. Technometrics 44, 328–337.

Currie, L.A., Gerlach, R.W., Lewis, C.W., Balfour, W.D., Cooper, J.A., Dattner, S.L., DeCesar, R.T., Gordon, G.E., Heisler, S.L., Hopke, P.K., Shah, J.J., Thurston, G.D., Williamson, H.J., 1984. Interlaboratory comparison of source apportionment procedures: results for simulated data sets. Atmospheric Environment 18, 1555–1566.

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 and Technology 7, 235–240.

Fuller, W.A., 1987. Measurement Error Models. Wiley, New York.

Gleser, L.J., 1997. Some thoughts on chemical mass balance models. Chemometrics and Intelligent Laboratory Systems 37, 15–22.

Golub, G.H., Van Loan, C.F., 1980. An analysis of the total least squares problem. SIAM Journal of Numerical Analysis 17, 883–893.

Henry, R.C., 1982. Stability analysis of receptor models that use least squares fitting. In: Dattner, S.L., Hopke, P.K. (Eds.), Receptor Models Applied to Contemporary Air Pollution Problems. Proceedings No. SP-48. Air and Waste Management Association, Pittsburgh, PA, pp. 141–157.

Henry, R.C., 1997. History and fundamentals of multivariate air quality receptor models. Chemometrics and Intelligent Laboratory Systems 37, 525–530.

Henry, R.C., Lewis, C.W., Collins, J.F., 1994. Vehicle-related hydrocarbon source compositions from ambient data: the GRACE/SAFER method. Environmental Science and Technology 28, 823–832.

Javitz, H.S., Watson, J.G., Guertin, J.P., Mueller, P.K., 1988a. Results of a receptor modeling feasibility study. Journal of the Air Pollution Control Association 38, 661–667.

Javitz, H.S., Watson, J.G., Robinson, N.F., 1988b. Performance of the chemical mass balance model with simulated local-scale aerosols. Atmospheric Environment 22, 2309–2322.

Miller, M.S., Friedlander, S.K., Hidy, G.M., 1972. A chemical element balance for the Pasadena aerosol. Journal of Colloid Interface Science 39, 65–176.

Paatero, P., Tapper, U., 1994. Positive matrix factorization: a non-negative factor model with optimal utilization of error estimate of data values. Environmetrics 5, 111–126.

Park, E.S., Guttorp, P., Henry, R.C., 2001. Multivariate receptor modeling for temporally correlated data by using MCMC. Journal of the American Statistical Association 96, 1171–1183.

Park, E.S., Oh, M.-S., Guttorp, P., 2002. Multivariate receptor models and model uncertainty. Chemometrics and Intelligent Laboratory Systems 60, 49–67.

Thurston, G.D., Spengler, J.D., 1985. A quantitative assessment of source contributions to inhalable particulate matter pollution in metropolitan Boston. Atmospheric Environment 19, 9–25.

UNCORRECTED PROOF

Watson, J.G., Cooper, J.A., Huntzicker, J.J., 1984. The effective variance weighting for least squares calculations applied to the mass balance receptor model. Atmospheric Environment 18, 1347–1355.

Watson, J.G., Chow, J.C., Pace, T.G., 1991. Chemical mass balance. In: Hopke, P.K. (Ed.), Receptor Modeling for Air Quality Management. Elsevier Science Publishers, New York, pp. 83–116.

Winchester, J.W., Nifong, G.D., 1971. Water pollution in Lake Michigan by trace elements from pollution aerosol fallout. Water, Air, and Soil Pollution 1, 50–64.

UNCORRECTED PROOF