ENVIRONMETRICS
Environmetrics 2000; **12**:25–40

Robust estimation of chemical profiles of air-borne particulate matter

D. Wiens[1], L. Z. Florence[2]* and M. Hiltz[2]

[1] *Statistics Centre, Department of Mathematical Sciences, University of Alberta, Edmonton, Alberta, Canada T6G 2G1*
[2] *Forest Resources Unit, Alberta Research Council, Vegreville, Alberta, Canada T9C 1T4*

SUMMARY

We present a modification of the Chemical Mass Balance model commonly used for apportioning pollutants measured at a receptor site to particular sources, given profile data from these sources. The standard Effective Variance model is included as a special case. We present a package of estimation methods for these models; a 'robustness option' is highlighted. A simulation study is carried out to compare and contrast the various approaches. Copyright © 2001 John Wiley & Sons, Ltd.

KEY WORDS: ambient profiles; chemical mass balance; effective variance; Generalized M-estimate; iteratively reweighted least squares; least absolute deviations; least squares; maximum likelihood; Newton–Raphson; regression; source profiles

1. INTRODUCTION

Inhalable particulate matter (PM) in the atmosphere is a major environmental and pubic health concern in North America and elsewhere (Burnett *et al.*, 1995; Dockery *et al.*, 1993). PM collected at a receptor site is generally grouped according to size fractions: fine (<2.5 μg) and coarse (2.5–10 μg). The chemicals associated with these PM fractions offer unique challenges beyond their physico-chemical attributes because they represent complex mixtures of often multiple point sources of pollutants (see Hopke, 1991 and references therein). The objectives for monitoring air quality at a receptor site then become those of sampling (defining frequency and duration), estimation (determining ambient chemical concentrations) and apportionment (allocating the total ambient particulate mass among all regional sources detected at the receptor, both natural and anthropogenic, given the sources' chemical profiles). These activities are often further

* Correspondence to: Zack Florence, FVA Inc., 1316 Slater Street, Victoria, BC V8X 2P9, Canada.

Received 2 December 1999
Revised 22 March 2000

extended to include both temporal and spatial variation (Brook *et al*., 1997; CEPA/FPAC, 1998; Chow *et al*., 1992).

Statistical and chemometric methods that have been used for partitioning ambient pollutants measured at a receptor site include principal component analysis (including factor analysis), multiple linear regression and chemical mass balance (CMB) models. These techniques have been used singly and in combination. The CMB model has been used in several studies in Canada and the United States because the theory has been well developed over the past 30 years and MS-Windows-based software has been made publicly available by the U.S. Environment Protection Agency (EPA) (see Lowenthal *et al*., 1997 and references therein).

Watson *et al*. (1984; henceforth referred to as WC&H) developed an Effective Variance (EV) CMB model. This and subsequent CMB applications, developed for the EPA largely by J. G. Watson and colleagues at the Desert Research Institute (DRI), Reno, Nevada (U.S.A.), make a number of assumptions which are to be met before fitting ambient and source chemical profiles. (The current version of DRI's CMB software can be downloaded by anonymous FTP from `eafs.sage.dri.edu/model/cmb8MMDD.exe`, where MMDD stand for month and day.) These assumptions include (from Watson *et al*., 1991): '(1) compositions of source emissions are constant over the period of ambient and source sampling; (2) chemical species do not react with each other, i.e. they add linearly; (3) all sources with potential for significantly contributing to the receptor have been identified and have had their emissions characterized; (4) the source compositions are linearly independent of each other; (5) the number of sources of source categories is less than or equal to the number of chemical species; (6) measurement uncertainties are random, uncorrelated, and normally distributed.'

Of course practitioners often apply methods developed under possibly untenable assumptions, in the hopes that assumptions which are 'close' to being satisfied will result in applications which are 'close' to being appropriate. There is now a wealth of robustness studies including that such an attitude can be seriously misguided, and that seemingly minor violations of assumptions such as normality or independence can result in a very significant deterioration in the performance of an otherwise appropriate or even optimal statistical procedure. Mathematical descriptions of the difficulty can be phrased in terms of discontinuities in the quality of the procedures, at those points at which the assumptions are violated.

While we do not argue the utility and contributions made by the CMB method developed at DRI, we suggest that adding robustness to the estimation methods can reduce the risk of spurious conclusions regarding apportionment of emission sources based upon results where assumptions are not or cannot be met. Most often, these sorts of violations would arise because: (1) the user of the CMB software would be using source profiles obtained from data libraries containing chemical profiles compiled in many different locations, not from actual data locally obtained (see e.g. Lowenthal *et al*., 1997) and not always having a knowledge about the data's quality during gathering and handling, and (2) it would often be impossible, or economically infeasible, to test whether or not all assumptions were met.

In Section 2 of this article we present a modification to the CMB model. We discuss the similarities with, and differences from, other approaches in the literature. In Section 3 we develop estimation methods for this model based on least squares, and then a set of robust alternatives. In a simulation study carried out in Section 4 we compare our methods with analyses carried out using the DRI effective variance CMB model and previously published data. We argue that the new methods afford additional and necessary security against erroneous allocations of PM chemistry among emission sources.

2. THE MODIFIED CMB MODEL

In this section we use the notation

$\mathbf{y} = n \times 1$ vector of ambient measurements; thus the ith element y_i is the ambient amount of species i. Typically all measurement units are $\mu g/m^3$.

$\mathbf{A} = n \times p$ matrix whose jth column $\mathbf{a}_j$ consists of the n 'true' profile values at source j; thus a_{ij} refers to species i, source j and is the amount of species i in the emissions from source j as perceived at the receptor.

$\mathbf{X} = n \times p$ matrix whose jth column $\mathbf{x}_j = (X_{1j}, \ldots, X_{nj})^T$ consists of the measured profile values at source j.

$\boldsymbol{\theta} = p \times 1$ vector of total mass contributions of the sources to the receptor; θ_j refers to source j.

Assume that, apart from random error, one has

$$\mathbf{y} = \sum_{j=1}^{p} \mathbf{a}_j \theta_j \tag{1}$$

for unknown source contributions θ_j, to be estimated. The ambient amounts y_i are measured with error ε_i, the variation of an error depending on the species. Assume that these n errors in the measurement of $\mathbf{y}$ are independent of each other. Thus, with $\boldsymbol{\varepsilon} = (\varepsilon_1, \ldots, \varepsilon_n)^{\mathrm{T}}$,

$$\mathbf{y} = \mathbf{A}\boldsymbol{\theta} + \boldsymbol{\varepsilon}; \tag{2}$$

$$E[\boldsymbol{\varepsilon}] = \mathbf{0}, \quad \mathrm{COV}[\boldsymbol{\varepsilon}] = \boldsymbol{\Sigma}_{\varepsilon}. \tag{3}$$

Here $\boldsymbol{\Sigma}_{\varepsilon} = \mathrm{diag}(\sigma_1^2, \ldots, \sigma_n^2)$ is a diagonal matrix with diagonal elements of $\sigma_i^2 = \mathrm{VAR}[\varepsilon_i]$.

The $\mathbf{a}_j$ are not known and are observed with error; i.e. one observes a random vector $\mathbf{x}_j$ rather than $\mathbf{a}_j$. The errors $\mathbf{x}_j - \mathbf{a}_j = \boldsymbol{\delta}_j$ may be correlated. It is assumed that the variances may vary both with the species and with the source, but that the *correlation* structure within each profile is the same across sources. Thus

$$\mathbf{x}_j = \mathbf{a}_j + \boldsymbol{\delta}_j; \tag{4}$$

$$E[\boldsymbol{\delta}_j] = \mathbf{0}, \quad \mathrm{COV}[\boldsymbol{\delta}_j] = \boldsymbol{\Sigma}_j, \tag{5}$$

where the structure of the $n \times n$ covariance matrix $\boldsymbol{\Sigma}_j$ is

$$\boldsymbol{\Sigma}_j = \boldsymbol{\Lambda}_j^{1/2} \boldsymbol{\Omega} \boldsymbol{\Lambda}_j^{1/2}.$$

We assume that the errors $\boldsymbol{\delta}_j$ are independent of $\boldsymbol{\varepsilon}$. In the expression above,

$$\boldsymbol{\Lambda}_j = \mathrm{diag}(\sigma_j^2, \ldots, \sigma_{nj}^2)$$

is a diagonal matrix of variances for the species within source j, and $\boldsymbol{\Omega}$ is a correlation matrix. Since the $n \times n$ matrix $\boldsymbol{\Omega}$ must be estimated from only p observation vectors, it appears that

further structure must be imposed. Assume that all off-diagonal entries of $\mathbf{\Omega}$ are equal to a common value ρ, necessarily $\geqslant -1/(n-1)$ in order that $\mathbf{\Omega}$ be positive semi-definite.

In the development (WC&H) of the Effective Variance (EV) CMB model, it is assumed that the measurements X_{ij} are independently and normally distributed with means a_{ij} and variances σ_{ij}^2. WC&H thus assume, in the notation above, that $\mathbf{\Omega} = \mathbf{I}_n$, the $n \times n$ identity matrix. The σ_i^2 and σ_{ij}^2 are assumed known in the theoretical developments of the EV model and the models proposed here. In the implementations these variances are estimated (often before the data are submitted to an analyst) and the estimates are substituted for the true values.

Data given to CMB analysts tend to be 'noisy' and 'dirty'. It is typically difficult to have much faith in the accuracy of much of the data and in particular in the variance estimates. Thus, as well as using classical least squares based methods, we shall propose robust procedures which are not overly sensitive to gross errors in the variance estimates and in other features of the data.

Practitioners might also question the assumption, in the formulation of the EV model, that the X_{ij} are *independently* distributed. Our assumption that the correlation structure is constant across sources, and of a constant value, is also somewhat questionable. It is however less so than the assumption of WC&H that it is constant with correlation matrix $\mathbf{\Omega} = \mathbf{I}_n$.

The normality assumption is not used explicitly by WC&H, although it is used implicitly to justify the use of Least Squares as an estimation procedure. Least Squares is well-known not to be robust against long-tailed (i.e. longer than normal) error distributions.

Our assumptions that the errors ε_i are uncorrelated is necessary when the data include only one ambient value y_i per species or a mean over time or locations, so that estimation of correlations between the ambient measurements is not always possible. Ohtaki *et al*. (1997) adopt a model somewhat similar to the one described here. However, they assume the availability of data from multiple receptors; this allows for estimation of COV[**y**] by the sample covariance matrix, summing across receptors.

Ohtaki *et al*. (1997) consider $\boldsymbol{\theta}$ as a realization of a random vector. The mean contributions are assumed to satisfy

$$0 \leqslant \theta_j \leqslant 1, \sum_{j=1}^{p} \theta_j = 1,$$

but the non-negativity is then addressed in an *ad hoc* (but sensible) manner which does not guarantee that the solution will satisfy this constraint. In fact Hopke (1985, p. 134) comments '... in a mass balance, source contributions should only be positive. It is possible to use a constrained least-squares fit, but this approach has not yet been seriously explored.' In particular, these constraints are not assumed in the EV model or its CMB implementation. Since a primary purpose of the present article is to extend the EV and CMB techniques by adding considerations of robustness, the constraints are also not imposed here. We do however make a *post hoc* modification to the parameter estimates to ensure non-negativity.

3. ESTIMATION METHODS

We first outline the estimation methods used, assuming that the regressions will be carried out by least squares. A robust alternative will then be described and evaluated.

By rearranging Equations (2)–(5) the observed vector $\mathbf{y}$ may be represented as

$$\mathbf{y} = \mathbf{X}\boldsymbol{\theta} + \mathbf{f} \tag{6}$$

where

$$\mathbf{f} = \boldsymbol{\varepsilon} - \sum_{j=1}^{p} \boldsymbol{\delta}_j \theta_j$$

has mean $\mathbf{0}$ and covariance matrix $\mathrm{COV}[\mathbf{f}] = \mathbf{V}$, given by

$$\mathbf{V} = \boldsymbol{\Sigma}_\varepsilon + \sum_{j=1}^{p} \theta_j^2 \boldsymbol{\Sigma}_j.$$

Two estimation approaches, Option 1 and Option 2, are investigated and discussed in this study. Option 1 relies on the observation that if $\mathbf{V}$ were known then one could apply Generalized Least Squares (GLS) to (6) to estimate $\boldsymbol{\theta}$:

$$\begin{aligned}\hat{\boldsymbol{\theta}} &= \arg\min(\mathbf{y} - \mathbf{X}\boldsymbol{\theta})^{\mathrm{T}} \mathbf{V}^{-1} (\mathbf{y} - \mathbf{X})\boldsymbol{\theta}) \\ &= (\mathbf{X}^{\mathrm{T}} \mathbf{V}^{-1} \mathbf{X})^{-1} \mathbf{X}^{\mathrm{T}} \mathbf{V}^{-1} \mathbf{y}\end{aligned}$$

If $\boldsymbol{\theta}$ were known then one could estimate $\mathbf{V}$, in a manner described below.

Option 2 relies on the observation that if $\mathbf{A}$ were known one could apply GLS to (2) to estimate $\boldsymbol{\theta}$:

$$\begin{aligned}\hat{\boldsymbol{\theta}} &= \arg\min(\mathbf{y} - \mathbf{A}\boldsymbol{\theta})^{\mathrm{T}} \boldsymbol{\Sigma}_\varepsilon^{-1} (\mathbf{y} - \mathbf{A}\boldsymbol{\theta}) \\ &= (\mathbf{A}^{\mathrm{T}} \boldsymbol{\Sigma}_\varepsilon^{-1} \mathbf{A})^{-1} \mathbf{A}^{\mathrm{T}} \boldsymbol{\Sigma}_\varepsilon^{-1} \mathbf{y}.\end{aligned}$$

Again, if $\boldsymbol{\theta}$ were known then one could estimate $\mathbf{A}$.

Each option suggests an iterative procedure: estimate $\boldsymbol{\theta}$; use this estimate to estimate $\mathbf{V}$ or $\mathbf{A}$; re-estimate $\boldsymbol{\theta}$ as above, then re-estimate $\mathbf{V}$ or $\mathbf{A}$; iterate to convergence. We shall first describe each estimation in detail. Section 3.5, in which the various steps are put together to outline the entire procedure, also serves as a summary and comparison of the two options.

3.1. Estimation of A

If all parameters except $\mathbf{A}$ are known, and if the errors are normally distributed, then from Equations (2)–(5) the log-likelihood for $\mathbf{A}$, apart from some inessential constants, is given by

$$-2\log l = (\mathbf{y} - \mathbf{A}\boldsymbol{\theta})^{\mathrm{T}} \boldsymbol{\Sigma}_\varepsilon^{-1} (\mathbf{y} - \mathbf{A}\boldsymbol{\theta}) + \sum_{j=1}^{p} (\mathbf{x}_j - \mathbf{a}_j)^{\mathrm{T}} \boldsymbol{\Sigma}_j^{-1} (\mathbf{x}_j - \mathbf{a}_j). \tag{7}$$

To obtain the maximum likelihood estimate (MLE) one maximizes $\log l$. The matrix of partial derivatives, with (i, j)th element $\partial \log l / \partial a_{ij}$, has jth column

$$\mathbf{g}_j = \boldsymbol{\Sigma}_\varepsilon^{-1}(\mathbf{y}-\mathbf{A}\boldsymbol{\theta})\theta_j + \boldsymbol{\Sigma}_j^{-1}(\mathbf{x}_j-\mathbf{a}_j).$$

Solving the equations $\mathbf{g}_1 = \mathbf{g}_2 = \cdots = \mathbf{g}_p = \mathbf{0}$ gives the MLE $\hat{\mathbf{A}}$, with jth column

$$\hat{\mathbf{a}}_j = \mathbf{x}_j + \theta_j \boldsymbol{\Sigma}_j \mathbf{V}^{-1}(\mathbf{y}-\mathbf{X}\boldsymbol{\theta}). \tag{8}$$

Although $\hat{\mathbf{a}}_j$ is the MLE only under an assumption of normality, it is in any event a reasonable estimator. It adjusts $\mathbf{x}_j$, which is the (only) estimate of $\mathbf{a}_j$ if no other data are available, by taking into account the regression on $\mathbf{y}$. This adjustment vanishes if $\mathbf{y}-\mathbf{X}\boldsymbol{\theta} = \mathbf{0}$, as it should since then $\mathbf{y}$ is perfectly predicted by $\mathbf{X}$ and gives us no information which is not already contained in $\mathbf{X}$.

3.2. *Estimation of* $\boldsymbol{\Omega}$

Since $\boldsymbol{\Lambda}_j^{-1/2}\boldsymbol{\delta}_j = \boldsymbol{\Lambda}_j^{-1/2}(\mathbf{x}_j-\mathbf{a}_j)$ has correlation matrix $\boldsymbol{\Omega}$, if all other parameters are known then an estimate of $\boldsymbol{\Omega}$ (ignoring the structural assumption discussed in Section 2) is given by the correlation matrix obtained from the p columns

$$\boldsymbol{\Lambda}_j^{1/2}(\mathbf{x}_j-\hat{\mathbf{a}}_j) = -\theta_j \boldsymbol{\Lambda}_j^{-1/2}\boldsymbol{\Sigma}_j\mathbf{V}^{-1}\mathbf{y}-\mathbf{X}\boldsymbol{\theta}) = -\theta_j\boldsymbol{\Omega}\boldsymbol{\Lambda}_j^{1/2}\mathbf{V}^{-1}(\mathbf{y}-\mathbf{X}\boldsymbol{\theta}). \tag{9}$$

If Option 1 is chosen and the robustness option described in Section 3.7 is not, then we use this last expression, with $\boldsymbol{\Omega}$ evaluated at its value in the previous iteration of the numerical procedure and with $\boldsymbol{\Lambda}_j$ and $\mathbf{V}$ estimated as shown below. Otherwise we use the first expression in (9). We compute an α-trimmed correlation matrix $\mathbf{R}$ from these columns, and then ρ is estimated by the α-trimmed mean of the off-diagonal elements of $\mathbf{R}$. We take $\alpha = 0$ for the least squares option being described here; the robust option employs $\alpha = 0.1$. In either case the required structure is imposed on the estimate of $\boldsymbol{\Omega}$ by defining $\boldsymbol{\Omega}_{ij}$ to be $(x-1/(n-1), \hat{\rho})$ for $i \neq j$.

3.3. *Estimation of* $\boldsymbol{\Lambda}_j$ and $\boldsymbol{\Sigma}_j$

Estimates of s_{ij}^2 of σ_{ij}^2 form a part of the data; one can estimate $\boldsymbol{\Lambda}_j$ by

$$\mathbf{S}_j = \mathrm{diag}(s_{1j}^2, \ldots, s_{nj}^2)$$

and then $\boldsymbol{\Sigma}_j$ by

$$\hat{\boldsymbol{\Sigma}}_j = \mathbf{S}_j^{1/2}\hat{\boldsymbol{\Omega}}\mathbf{S}_j^{1/2},$$

with typical element

$$[\hat{\boldsymbol{\Sigma}}_j]_{i,k} = \hat{\boldsymbol{\Omega}}_{ik}s_{ij}s_{kj}, \quad 1 \leqslant i,k \leqslant n.$$

3.4. Estimation of $\mathbf{V}$

Given estimates s_i^2 of σ_i^2, $\hat{\boldsymbol{\theta}}$ of $\boldsymbol{\theta}$ and $\hat{\boldsymbol{\Sigma}}_j$ of $\boldsymbol{\Sigma}_j$ can one estimate $\mathbf{V} = \boldsymbol{\Sigma}_\varepsilon + \Sigma_{j=1}^{p} \theta_j^2 \boldsymbol{\Sigma}_j$ by

$$\hat{\mathbf{V}} = \mathbf{S}_\varepsilon + \sum_{j=1}^{p} \hat{\theta}_j^2 \hat{\boldsymbol{\Sigma}}_j,$$

where $\mathbf{S}_\varepsilon = \mathrm{diag}(s_1^2, \ldots, s_n^2)$.

3.5. Iterative procedure for Options 1 and 2; least squares

We describe here the numerical procedure by which the estimates are obtained. The parameters $\boldsymbol{\theta}$, $\boldsymbol{\Omega}$, $\boldsymbol{\Sigma}_j$, $\mathbf{V}$ and possibly $\mathbf{A}$ are first set equal to simple initial values, then successively updated until the values of $\hat{\boldsymbol{\theta}}$ stabilize.

Step 0. Initialization step:

$$\boldsymbol{\theta} \leftarrow \boldsymbol{\theta}^{(0)} = \mathbf{0},$$
$$\boldsymbol{\Omega} \leftarrow \boldsymbol{\Omega}^{(0)} = \mathbf{I}_n,$$
$$\boldsymbol{\Sigma}_j \leftarrow \boldsymbol{\Sigma}_j^{(0)} = \mathbf{S}_j,$$
$$\mathbf{V} \leftarrow \mathbf{V}^{(0)} = \mathbf{S}_\varepsilon,$$

For Option 2 only;

$$\mathbf{A} \leftarrow \mathbf{A}^{(0)} = \mathbf{X}.$$

Step $k \geqslant 1$. Updating. Compute, in the indicated order:

$$\boldsymbol{\theta} \leftarrow \boldsymbol{\theta}^{(k)} = \begin{cases} \arg\min(\mathbf{y} - \mathbf{X}\boldsymbol{\theta})^{\mathrm{T}} \mathbf{V}^{-1} (\mathbf{y} - \mathbf{X}\boldsymbol{\theta}) = (\mathbf{X}^{\mathrm{T}} \mathbf{V}^{-1} \mathbf{X})^{-1} \mathbf{X}^{\mathrm{T}} \mathbf{V}^{-1} \mathbf{y}, & \text{Option 1,} \\ \arg\min(\mathbf{y} - \mathbf{A}\boldsymbol{\theta})^{\mathrm{T}} (\mathbf{S}_\varepsilon^{-1} (\mathbf{y} - \mathbf{A}\boldsymbol{\theta}) = (\mathbf{A}^{\mathrm{T}} S_\varepsilon^{-1} \mathbf{A})^{-1} \mathbf{A}^{\mathrm{T}} \mathbf{S}_\varepsilon^{-1} \mathbf{y}, & \text{Option 2,} \end{cases}$$

then truncate at 0: $\theta_j^{(k)} \leftarrow \mathrm{ma}\ (\theta_j^{(k)}, 0)$ for $j = 1, \ldots, p$,

$$\boldsymbol{\Omega} \leftarrow \boldsymbol{\Omega}^{(k)} \text{ as described in Section 3.2,}$$
$$\boldsymbol{\Sigma}_j \leftarrow \boldsymbol{\Sigma}_j^{(k)} = \mathbf{S}_j^{1/2} \boldsymbol{\Omega} \mathbf{S}_j^{1/2},$$
$$\mathbf{V} \leftarrow \mathbf{V}^{(k)} = \mathbf{S}_\varepsilon + \sum_{j=1}^{p} \theta_j^2 \boldsymbol{\Sigma}_j,$$

For Option 2 only:

$$\mathbf{A} \leftarrow \mathbf{A}^{(k)}, \text{ with } j\text{th column } \mathbf{a}_j = \mathbf{x}_j + \theta_j \boldsymbol{\Sigma}_j \mathbf{V}^{-1} (\mathbf{y} - \mathbf{X}\boldsymbol{\theta}),$$

then truncate at 0: $a_{ij} \leftarrow \mathrm{ma}\ (a_{ij}, 0)$.

Iterate until convergence is attained. Convergence is defined in terms of relative convergence $\boldsymbol{\theta}^{(k)}$, i.e. convergence is declared when $\|\boldsymbol{\theta}^{(k)}-\boldsymbol{\theta}^{(k-1)}\|/\|\boldsymbol{\theta}^{(k-1)}\| <$ tolerance for, say, tolerance $= 0.01$. Here $\|\cdot\|$ is the Euclidean norm.

The algorithm employed by WC&H is that of Option 1, with the difference that $\boldsymbol{\Omega}$ is never updated; it remains $= \mathbf{I}_n$.

3.6. *Inferences*

For Option 1, approximately valid inference procedures are obtained by applying standard regression theory to the model $\mathbf{y} = \mathbf{X}\boldsymbol{\tau}+\mathbf{f}$, COV[$\mathbf{f}$] $= \mathbf{V}$, X fixed, V known. Then $\mathbf{V}$ is replaced by $\hat{\mathbf{V}}$ (=the value of $\mathbf{V}$ at the termination of the iterative estimation procedure). Option 2 can be handled in an analogous manner. This gives

$$E[\hat{\boldsymbol{\theta}}] \approx \boldsymbol{\theta}; \text{est.cov.}(\hat{\boldsymbol{\theta}}) = \begin{cases} (\mathbf{X}^{\mathrm{T}}\hat{\mathbf{V}}^{-1}\mathbf{X})^{-1}, & \text{Option 1;} \\ (\hat{\mathbf{A}}^{\mathrm{T}}\mathbf{S}_{\varepsilon}^{-1}\hat{\mathbf{A}})^{-1}, & \text{Option 2} \end{cases}.$$

The p-values are computed using a t_{n-p} approximation to the distribution of the standardized ratio

$$t = \frac{\hat{\theta}_j - \theta_j}{s(\hat{\theta}_j)},$$

where $s^2(\hat{\theta}_j) = [\text{est. cov.}(\hat{\boldsymbol{\theta}})]_{jj}$ is the estimated variance of $\hat{\theta}_j$. The use of the t_{n-p}, rather than the normal, reference distribution is the usual penalty paid for estimation of the standard error of the regression estimate.

An ANOVA (Analysis of Variance) breakdown starts by transforming to weighted data:

$$\begin{aligned} (\tilde{\mathbf{y}}, \tilde{\mathbf{X}}) &= (\hat{\mathbf{V}}^{-1/2}\mathbf{y}, \hat{\mathbf{V}}^{-1/2}\mathbf{X}) \quad \text{(Option 1)}, \\ (\tilde{\mathbf{y}}, \tilde{\mathbf{A}}) &= (\mathbf{S}_{\varepsilon}^{-1/2}\mathbf{y}, \mathbf{S}_{\varepsilon}^{-1/2}\mathbf{A}) \quad \text{(Option 2)}. \end{aligned} \tag{10}$$

Then the Total Sum of Squares $SST = \|\tilde{\mathbf{y}}\|^2$ is broken down into the Sum of Squares due to the Regression SSR (=the sum of squares $\|\tilde{\mathbf{X}}(\tilde{\mathbf{X}}^{\mathrm{T}}\tilde{\mathbf{X}})^{-1}\tilde{\mathbf{X}}^{\mathrm{T}}\tilde{\mathbf{y}}\|^2$ or $\|\tilde{\mathbf{A}}(\tilde{\mathbf{A}}^{\mathrm{T}}\tilde{\mathbf{A}})^{-1}\tilde{\mathbf{A}}^{\mathrm{T}}\tilde{\mathbf{y}}\|^2$, as appropriate, of the fitted values in terms of the "˜" data) and the Sum of Squares due to Error $SSE = SST - SSR$.

For Option 1 the fitted values $\tilde{\mathbf{y}} = \mathbf{X}\tilde{\boldsymbol{\theta}} = \mathbf{K}\mathbf{y}$ where $\mathbf{K} = \mathbf{X}(\mathbf{X}^{\mathrm{T}}\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{V}^{-1}$, and residuals $\mathbf{e} = (\mathbf{I}-\mathbf{K})\mathbf{y}$ have approximate covariance matrices $\mathbf{KVK}$ and $(\mathbf{I}-\mathbf{K})\mathbf{V}(\mathbf{I}-\mathbf{K})$ respectively. These are estimated by replacing $\mathbf{V}$ by $\hat{\mathbf{V}}$. For Option 2 $\mathbf{X}$ is replaced by $\hat{\mathbf{A}}$, $\hat{\mathbf{V}}$ by $\mathbf{S}_{\varepsilon}$. This sets the stage for the usual range of diagnostic procedures based on residual analyses.

3.7. *Adding robustness to the CMB analysis*

Robustness is achieved for the two options partly by substituting the following into the least squares regressions described in Section 3.5:

$$\hat{\boldsymbol{\theta}} = \begin{cases} \arg\min_{\boldsymbol{\theta}} \sum_{i=1}^{n} w_i \xi\left(\dfrac{\tilde{y}_i - \tilde{\mathbf{X}}_i^{\mathrm{T}}\boldsymbol{\theta}}{S}\right), & \text{Option 1;} \\ \arg\min_{\boldsymbol{\theta}} \sum_{i=1}^{n} w_i \xi\left(\dfrac{\tilde{y}_i - \tilde{\mathbf{A}}_i^{\mathrm{T}}\boldsymbol{\theta}}{S}\right), & \text{Option 2.} \end{cases} \tag{11}$$

Here $\tilde{\mathbf{y}}$, $\tilde{\mathbf{X}}$ and $\tilde{\mathbf{A}}$ are as at (10), S is a robust measure of scale and the w_i are weights designed to bound the influence of outlying regressors. Thus $\hat{\boldsymbol{\theta}}$ is a Mallows-type Generalized M-estimate. If $\xi(t) = t^2$ and $w_i \equiv 1$ then (11) gives the least squares estimates of Section 3.5. Alternative forms of ξ, for robustness against outliers, are obtained by replacing $t^2/2 = \int_0^t x\,\mathrm{d}x$ by $\xi(t) = \int_0^t \psi(x)\,\mathrm{d}x$, where $\psi(x)$ is a bounded score function. Common choices are 'Huber's ψ function'

$$\psi(x) = \begin{cases} x, & |x| \leqslant k, \\ k \cdot \mathrm{sign}(x), & |x| \geqslant k; \end{cases}$$

for a user-chosen value of k, and 'Hampel's 3-part redescending ψ function'

$$\psi(x) = \begin{cases} x, & |x| \leqslant k_1, \\ k_1 \cdot \mathrm{sign}(x), & k_1 \leqslant |x| \leqslant k_2; \\ k_1 \dfrac{k_3 - |x|}{k_3 - k_2}\mathrm{sign}(x), & k_2 |x| \leqslant k_3; \\ 0 & k_3 \leqslant |x|. \end{cases}$$

for user-chosen values $k_1 < k_2 < k_3$. In both cases, letting $k \to \infty$ or $k \to \infty$ results in the least squares estimate, with $\psi(x) = x$. Finite values of these tuning constants result in estimates which bound the influence of large residuals on the fit. The Huber estimate gives all sufficiently large residuals the same influence, while the Hampel estimate cuts the influence of very large residuals to zero. See Hampel *et al.* (1986) for discussion. The default values used here are $k = 0.5$, $(k_1, k_2, k_3) = (0.5, 1.5, 5)$; for these choices plots are given in Figure 1.

In our simulations we have taken S to be the median absolute deviation (around the median) of the residuals, normalized for consistency at the Gaussian distribution. We use weights

Figure 1. Huber (a) and Hampel (b) ψ functions, using default values of the tuning constants. Horizontal axis represents regression residual, vertical axis the relative influence of this residual on the regression fit.

$w_i = (1-h_{ii})/\sqrt{h_{ii}}$ (a suggestion of Welsch, 1980), where the leverages h_{ii} are the diagonal elements of the 'hat' matrix formed from the regressors. Regressors far from their centroid yield leverages and thus small weights. These weights are known not to be completely robust, since clusters of outliers can draw the centroid towards themselves, thus diminishing their apparent leverages. However, more robust weighting schemes are typically much more computationally demanding and require a relatively large number of observations, hence are not feasible for the large numbers of simulations, with $n = 8$ only, carried out in this study. See Du and Wiens (2000) for a discussion.

The solutions to (11) are initiated by first computing the least absolute deviations estimator, which minimizes the sum of the absolute values of the residuals rather than of their squares. This is followed by three iterations of a Newton–Raphson algorithm for (11). Finally, one step of the iteratively reweighted least squares regression algorithm is performed. See Simpson and Chang (1997) for details.

Robust up-dating of the matrix $\mathbf{A}$ is performed as follows. Rather than minimizing (7), we minimize its analogue

$$\xi(\|\boldsymbol{\Sigma}_\varepsilon^{-1/2}(\mathbf{y}-\mathbf{A}\boldsymbol{\theta})\|)+\sum_{j=1}^{p}\xi(\|(\boldsymbol{\Sigma}_j^{-1/2}(\mathbf{x}_j-\mathbf{a}_j)\|).$$

Define a function $u(t) = \psi(\sqrt{t})/\sqrt{t}$ for $t > 0$, $= 1$ for $t = 0$. By differentiation we obtain the equation

$$\mathbf{A} = \mathbf{F}(\mathbf{A}) \tag{12}$$

where

$$\mathbf{F}(\mathbf{A})_{n\times p} = \mathbf{X}+\tau_0(\mathbf{A})\left(\frac{\boldsymbol{\Sigma}_1\mathbf{w}(\mathbf{A})\theta_1}{\tau_1(\mathbf{A})}\vdots\ldots\vdots\frac{\boldsymbol{\Sigma}_p\mathbf{w}(\mathbf{A})\theta_p}{\tau_p(\mathbf{A})}\right),$$

$$\mathbf{w}(\mathbf{A}) = \boldsymbol{\Sigma}_\varepsilon^{-1}(\mathbf{y}-\mathbf{A}\boldsymbol{\theta}),$$

$$\tau_0(\mathbf{A}) = u((\mathbf{y}-\mathbf{A}\boldsymbol{\theta})^{\mathrm{T}}\boldsymbol{\Sigma}_\varepsilon^{-1}(\mathbf{y}-\mathbf{A}\boldsymbol{\theta})),$$

$$\tau_j(\mathbf{A}) = u((_j-\mathbf{a}_j)^{\mathrm{T}}\,\boldsymbol{\Sigma}_j^{-1}(\mathbf{x}_j-\mathbf{a}_j)).$$

Equation (12) suggests a natural iterative scheme. We first replace $\boldsymbol{\Sigma}_j$, $\boldsymbol{\Sigma}_\varepsilon$, $\boldsymbol{\theta}$ by their current estimates $\boldsymbol{\Sigma}_j^{(k)}$, $\mathbf{S}_\varepsilon$, $\boldsymbol{\theta}^{(k)}$. Then with $\mathbf{A}_0 := \mathbf{A}^{(k)}$, for $m = 0, 1, \ldots$ we compute

$$\alpha_m = \min\left(\frac{\text{tolerance}\cdot\|\mathbf{A}_m\|}{\|\mathbf{F}(\mathbf{A}_m)-\mathbf{A}_m\|}, 1\right),$$

$$\mathbf{A}_{m+1} = (1-\alpha_m)\mathbf{A}_m+\alpha_m\mathbf{F}(\mathbf{A}_m).$$

If both sequences converge, with α_m having a non-zero limit, then the limit of $\mathbf{A}_m$ satisfies (12). Note that $\alpha_m < \Leftrightarrow \|\mathbf{F}(\mathbf{A}_m)-\mathbf{A}_m|/\|\mathbf{A}_m| >$ tolerance, so that an approximate solution is obtained by iterating until $\alpha_m = 1$, which also implies that $\mathbf{A}_{m+1} = \mathbf{F}(\mathbf{A}_m)$. Our program iterates until $m = 20$ or $\alpha_m = 1$; if $\alpha_{20} < \alpha_0$ we take $\mathbf{A}^{(k+1)} = \mathbf{A}^{(k)}$, i.e. no updating is performed at this kth stage.

The equalities in (9) no longer hold, since they are derived from (8). To update $\boldsymbol{\Omega}$ under the robustness option we compute $\mathbf{R}$ from the columns given by the first term in (9).

The inferential procedures of Section 3.6 remain asymptotically valid, once the estimate of the covariance matrix of $\hat{\boldsymbol{\theta}}$ is appropriately modified. Following Hinkley (1977) and Wu (1986) we use the one-step weighted jackknife estimate as proposed in Du and Wiens (2000), together with a finite-sample correction factor of Huber (1981). The estimate is described as follows. Let the matrix $\mathbf{Z}$ be either $\tilde{\mathbf{X}}$ (for Option 1) or $\tilde{\mathbf{A}}$ (for Option 2), with rows $\mathbf{z}_i$. Define

$$\mathbf{P} = \sum_{i=1}^{n} w_i \mathbf{z}_i \mathbf{z}_i^{\mathrm{T}}, p_i = w_i \mathbf{z}_i^{\mathrm{T}} \mathbf{P}^{-1} \mathbf{z}_i,$$

$$\mathbf{Q}_J = \sum_{i=1}^{n} \frac{w_1^2}{1-p_i} \mathbf{z}_i \mathbf{z}_i^{\mathrm{T}}, \kappa = 1 + \frac{p}{n} \cdot \frac{\mathrm{var}(\psi')}{\mathrm{aver}(\psi')},$$

where aver(ψ') and var(ψ') are the sample mean and variance of the $\psi'((\tilde{y}_i - \mathbf{z}_i^{\mathrm{T}} \hat{\boldsymbol{\theta}})/S)$. Then the estimated covariance matrix is

$$\mathbf{C}_J = \kappa^2 \cdot S^2 \cdot \frac{\frac{1}{n-p} \sum_{i=1}^{n} \left(\frac{\tilde{y}_i - \mathbf{z}_i^{\mathrm{T}} \tilde{\boldsymbol{\theta}}}{S} \right)}{\left(\frac{1}{n} \sum_{i=1}^{n} \psi' \left(\frac{\tilde{y}_i - \mathbf{Y}_i^{\mathrm{T}} \hat{\boldsymbol{\theta}}}{S} \right) \right)^2} \cdot \mathbf{P}^{-1} \mathbf{Q}_J \mathbf{P}^{-1}.$$

4. SIMULATIONS

WC&H describe some calculations of their model, and include some typical 'true' profile values a_{ij}. As in the smaller of their simulation studies, we chose their $n = 8$ species: Na, Al, Si, Cl, V, Ni, Br and Pb. Therefore, the values of the $\mathbf{A}$ matrix are as represented in their Table 1; the $p = 4$ possible emission sources are Marine (M), Urban dust (UD), Auto exhaust (AE) and Residual oil (RO). The relevant values from Table 1 of WC&H are reproduced in our Table I, together

Table I. Partial replications of 'values of variables for generating simulated data and solving mass balance equations' from WC&H.*

Aerosol properties	Marine a_{i1}	Urban dust a_{i2}	Auto exhaust a_{i3}	Residual oil a_{i4}
Na	0.40	0.0125	0	0.035
Al	0	0.0884	0.011	0.0053
Si	0	0.223	0.0082	0.0096
Cl	0.40	0	0.03	0
V	0	0.00023	0	0.0344
Ni	0	0.000093	0.00018	0.0536
Br	0	0.0002	0.05	0.00013
Pb	0	0.0037	0.20	0.0011

* The '10-set averages' of the estimates, from Table 2 of WC&H ($\pm$ one 'known' standard deviation of $\hat{\theta}_j$), were 17.7 ± 2.4, 32.3 ± 2.9, 32.0 ± 5.0, 14.7 ± 1.0.

with a summary of the estimates obtained by WC&H. The values s_{ij}^2 are as described in the footnotes to their Table 1 – $s_{ij}^2 = (0.1 \cdot x_{ij})^2$, with a few exceptions. Similarly $s_i^2 = (0.1 \cdot (\mathbf{X}\boldsymbol{\theta})_i)^2$. The true values of source contributions are $\boldsymbol{\theta} = (20, 35, 30, 15)^{\mathrm{T}}$. In our computations any s_i^2 or s_{ij}^2 equal to 0 was replaced by the α-trimmed mean (with α as in Section 3.2) of all *positive* variance estimates for that species. This admittedly *ad hoc* measure ensured the invertibility of $\mathbf{S}_\varepsilon$ and $\hat{\mathbf{V}}$. All the results from our simulations and discussion which follow are based upon code developed in the S-Plus software package (MathSoft Inc., Seattle, Washington, U.S.A.) and available from us (contact `doug.wiens@ualberta.ca`).

We first simulated independent vectors $\boldsymbol{\delta}_1, \ldots, \boldsymbol{\delta}_p$, where $\boldsymbol{\delta}_j$ was normally distributed with mean vector $\mathbf{0}$ and covariance matrix $\boldsymbol{\Sigma}_j = \boldsymbol{\Lambda}_j^{1/2} \boldsymbol{\Omega} \boldsymbol{\Lambda}_j^{1/2}$. Here, as in WC&H, $\boldsymbol{\Lambda}_j = \mathbf{S}_j = \mathrm{diag}(s_{1j}^2, \ldots, s_{nj}^2)$; i.e. the 'estimated' variances are in fact exactly correct. Then $\mathbf{X} = \mathbf{A} + \|\boldsymbol{\delta}_1, \ldots, \boldsymbol{\delta}_p\|$ was computed and truncated at 0 so that all elements would be non-negative. A response vector was then simulated: $\mathbf{y} = \mathbf{A}\boldsymbol{\theta} + \boldsymbol{\varepsilon}$, where $\boldsymbol{\varepsilon}$ was normally distributed with mean vector $\mathbf{0}$ and covariance matrix $\boldsymbol{\Sigma}_\varepsilon = \mathbf{S}_\varepsilon = \mathrm{diag}(s_1^2, \ldots, s_n^2)$. The first set of simulations used $\boldsymbol{\Omega} = \mathbf{I}_n$ (independent measurements within species across sources) for a comparison with the WC&H simulations; this series of analyses served as our internal control to verify both our understanding of and concordance with results by WC&H and to test our algorithms.

The procedure outlined above yielded one simulated data set $(\mathbf{y}, \mathbf{X})$, from which estimates $\hat{\boldsymbol{\theta}}$ were computed. This procedure was repeated N times. The results are summarized in our Table II for $N = 1000$, $\boldsymbol{\Omega} = \mathbf{I}_n$. Table III reports the results in the case $\boldsymbol{\Omega} = \boldsymbol{\Omega}_0$, an equi-correlation matrix with all off-diagonal elements equal to 0.2.

In these tables the 'self-estimated standard deviations' are the sample averages of the 1000 standard errors computed along with the estimates themselves, using the covariance matrix estimates from Sections 3.6 and 3.7. These should be compared with the simulated standard deviations (accompanying the averages of the simulated estimates of the parameters), presented in the tables in the form 'average $\pm$ one simulated standard deviation'. The latter standard deviations are obtained by taking all 1000 of the simulated estimates, and then calculating their sample standard deviations. They should be viewed as the 'true' standard deviations of the regression parameter estimates.

From Table II we conclude that our Options 1 and 2 do not suffer from unnecessarily estimating $\boldsymbol{\Omega}$, compared to the WC&H's EV method which, correctly, in this case, assumes that $\boldsymbol{\Omega} = \mathbf{I}$. The EV method – which is identical to Option 1 using least squares and not estimating $\boldsymbol{\Omega}$, apart from the protocol detailed above for handling variance estimates which equal 0 – performed somewhat better for us than for WC&H. See the footnote to Table I for some summary values from WC&H.

On the 'clean' data used for Tables II and III most methods performed well, with only moderate biases. Note, however, that the least squares Option 2 method resulted in large biases and huge standard deviations in the estimates for UD when the correlation parameter was not estimated; this was ameliorated when ρ was estimated. The least squares based methods with Option 2 tended to underestimate the standard errors, whereas the other methods tended to overestimate them. The latter is generally preferable, since it leads to confidence intervals which, although wider, have coverages at least as great as the nominal levels. As seen from Table III, the estimation of ρ when in fact ρ was non-zero did not result in any significant improvement. This can be expected to change in larger datasets.

The values shown in Tables II and III appear to be *too good*. This may be due to the fact, pointed out above, that the 'estimated' variances were in fact *exactly* correct. To investigate the robustness of the methods against incorrectly estimated variances, we multiplied each variance

Table II. Simulation results: $N = 1000$, $\mathbf{\Omega} = \mathbf{I}_n$, $n = 8$, $p = 4$. 'Clean' data: 'estimated' variances are exactly correct.

	True values (θ_j)	M 20	UD 35	AE 30	RO 15
Least squares; ρ not estimated ($\hat{\rho} \equiv 0$)					
Option 1	Averages of simulated values*	20.0 ± 2.8	35.3 ± 3.7	29.9 ± 3.9	15.0 ± 1.6
	Self-estimated standard deviations†	3.3	3.8	4.1	2.0
Option 2	Averages of simulated values*	19.4 ± 2.9	42.8 ± 228.9	30.5 ± 5.0	15.3 ± 18.2
	Self-estimated standard deviations†	1.6	2.7	2.1	1.1
Least squares; ρ estimated					
Option 1	Averages of simulated values*	20.0 ± 2.8	35.3 ± 3.8	29.9 ± 3.9	15.0 ± 1.6
	Self-estimated standard deviations†	3.3	3.8	4.0	2.0
Option 2	Averages of simulated values*	19.4 ± 2.9	44.9 ± 295.3	30.5 ± 5.0	21.6 ± 216.9
	Self-estimated standard deviations†	1.6	2.7	2.1	1.1
Robust fit; ρ not estimated ($\hat{\rho} \equiv 0$)					
Option 1	Averages of simulated values*	21.7 ± 7.1	34.2 ± 4.5	31.6 ± 7.2	15.1 ± 1.9
	Self-estimated standard deviations†	18.0	9.5	19.5	5.7
Option 2	Averages of simulated values*	18.8 ± 3.7	35.5 ± 5.1	32.3 ± 10.2	15.1 ± 2.1
	Self-estimated standard deviations†	4.3	7.3	5.9	2.9
Robust fit; ρ estimated					
Option 1	Averages of simulated values*	21.8 ± 9.1	34.3 ± 4.9	30.5 ± 5.0	15.2 ± 2.1
	Self-estimated standard deviations†	15.3	8.8	11.2	5.2
Option 2	Averages of stimulated values*	18.7 ± 3.6	35.4 ± 5.1	32.4 ± 10.3	15.1 ± 2.2
	Self-estimated standard deviations†	4.8	8.1	6.8	3.2

* ± one standard deviation of $\hat{\theta}_j$, as estimated from the simulations.
† Obtained by averaging the estimated standard deviations over the simulations.

estimate by an independent realization of $|C|$, where C followed a Cauchy distribution. To investigate robustness against outliers, 10 per cent of the receptor measurements were randomly chosen and multiplied by 1/3 and 10 per cent were randomly chosen and multiplied by 3. The simulations were then re-run on these severely corrupted data, with $\mathbf{\Omega} = \mathbf{I}$. We suggest (see Table IV) that our robust estimation methods fared somewhat better than the least squares based estimates, primarily with respect to the accuracy of the standard errors.

In these simulations the robustness was implemented using a 'Hampel' score function with the default tuning constants. Results obtained with the 'Huber' were quite similar to those reported here.

5. SUMMARY AND CONCLUSIONS

We have presented a modified CMB model, together with a package of estimation procedures including a robustness option. A special case of our methods is the EV method of WC&H. A

Table III. Simulation results; $N = 1000$, $\mathbf{\Omega} = \mathbf{\Omega}_0$, $n = 8$, $p = 4$, 'clean' data.

	True values (θ_j)	M 20	UD 35	AE 30	RO 15
Least squares; ρ not estimated ($\hat{\rho} \equiv 0$)					
Option 1	Averages of simulated values*	20.1 ± 2.9	35.4 ± 3.9	30.0 ± 4.1	15.0 ± 1.7
	Self-estimated standard deviations†	3.3	3.8	4.1	2.0
Option 2	Averages of simulated values*	19.4 ± 3.1	42.9 ± 232.2	30.5 ± 5.1	15.3 ± 19.7
	Self-estimated standard deviations†	1.6	2.7	2.1	1.1
Least squares; ρ estimated					
Option 1	Averages of simulated values*	20.0 ± 2.9	35.4 ± 3.9	29.9 ± 4.1	15.0 ± 1.7
	Self-estimated standard deviations†	3.3	3.8	4.1	2.0
Option 2	Averages of simulated values*	19.3 ± 3.1	44.4 ± 278.0	30.5 ± 5.0	21.8 ± 222.0
	Self-estimated standard deviations†	1.6	2.7	2.1	1.1
Robust fit; ρ not estimated ($\hat{\rho} \equiv 0$)					
Option 1	Averages of simulated values*	21.7 ± 7.7	34.4 ± 4.5	31.6 ± 7.3	15.1 ± 2.0
	Self-estimated standard deviations†	17.2	9.0	18.8	5.5
Option 2	Averages of simulated values*	19.0 ± 3.8	35.5 ± 5.2	32.4 ± 10.5	15.1 ± 2.1
	Self-estimated standard deviations†	4.2	7.1	6.0	2.8
Robust fit; ρ estimated					
Option 1	Averages of simulated values*	21.8 ± 10.0	34.8 ± 5.0	30.5 ± 5.2	15.2 ± 2.2
	Self-estimated standard deviations†	15.1	8.0	15.9	5.8
Option 2	Averages of simulated values*	18.9 ± 3.8	35.5 ± 5.2	32.3 ± 10.6	15.1 ± 2.1
	Self-estimated standard deviations†	4.8	8.1	6.6	3.2

* ± one standard deviation of $\hat{\theta}_j$, as estimated from the simulations.
† Obtained by averaging the estimated standard deviations over the simulations.

simulation study in a particularly arduous situation ($n = 8, p = 4$) has shown that here most of the various methods have similar performances with respect to the accuracy of the estimates, but can vary widely in the estimation of their own standard errors. The protection against outliers afforded by the robust methods can be expected to become more relevant with larger values of n. Of course, the analyst is not restricted to the use of just one method. We recommend that a thorough analysis includes a comparison of methods based on least squares with those of our robust approach. Significant differences in the results should be interpreted as warnings of particularly anomalous features in the data.

ACKNOWLEDGEMENTS

We acknowledge the support of the Investment Fund Committee, Alberta Research Council and Dr D. McNabb, Business Unit Leader, Forest Resources. We are also grateful to R. Burnett, L. Cheng, R. Henry,

Table IV. Simulation results; $N = 1000$, $\mathbf{\Omega} = \mathbf{I}_n$, $n = 8$, $p = 4$, 'corrupted' data.

	True values (θ_j)	M 20	UD 35	AE 30	RO 15
Least squares; ρ not estimated ($\hat{\rho} \equiv 0$)					
Option 1	Averages of simulated values*	19.2 ± 8.0	35.2 ± 24.2	27.8 ± 16.1	18.9 ± 27.0
	Self-estimated standard deviations†	3.3	3.9	4.0	2.7
Option 2	Averages of simulated values*	19.6 ± 9.1	34.7 ± 16.3	38.8 ± 66.6	19.4 ± 31.5
	Self-estimated standard deviations†	1.6	2.7	2.3	1.2
Least squares; ρ estimated					
Option 1	Averages of simulated values*	19.9 ± 9.8	35.9 ± 26.1	29.1 ± 22.7	18.2 ± 27.0
	Self-estimated standard deviations†	3.3	3.8	4.0	2.6
Option 2	Averages of simulated values*	19.1 ± 8.6	35.1 ± 21.2	39.4 ± 67.2	19.7 ± 33.6
	Self-estimated standard deviations†	1.6	2.7	2.1	1.2
Robust fit; ρ not estimated ($\hat{\rho} \equiv 0$)					
Option 1	Averages of simulated values*	20.9 ± 11.6	34.1 ± 16.0	27.4 ± 18.2	16.5 ± 24.9
	Self-estimated standard deviations†	28.2	15.8	25.7	17.1
Option 2	Averages of simulated values*	20.0 ± 7.9	34.2 ± 16.9	28.1 ± 25.4	17.7 ± 29.6
	Self-estimated standard deviations†	14.8	25.9	25.3	18.2
Robust fit; ρ estimated					
Option 1	Averages of simulated values*	21.0 ± 11.7	33.6 ± 15.6	26.9 ± 17.8	16.6 ± 26.1
	Self-estimated standard deviations†	27.1	17.1	24.3	15.5
Option 2	Averages of simulated values*	20.0 ± 7.9	34.2 ± 16.9	28.1 ± 25.4	17.7 ± 29.6
	Self-estimated standard deviations†	14.8	25.9	25.3	18.2

* ± one standard deviation of $\hat{\theta}_j$, as estimated from the simulations.
† Obtained by averaging the estimated standard deviations over the simulations.

C. Lewis, D. Lowenthal, H. Nguyen, G. Morris, E. Peake, B. Pulsipher and J. Watson for reviewing an earlier version (with associated software) of this work, and to three anonymous referees.

REFERENCES

Brook JR, Dann TF, Burnett RT. 1997. The relationship among TSP, PM_{10}, $PM_{2.5}$ and inorganic constituents of atmospheric particulate matter at multiple Canadian locations. *Journal of Air and Waste Management Association* **47**:2–19.

Burnett RT, Dales RE, Krewski D, Vincent R, Dann TF, Brook JR. 1995. Associations between ambient particulate sulfate and admission to Ontario hospitals for cardiac and respiratory diseases. *American Journal of Epidemiology* **142**:15–22.

CEPA/FPAC (Canadian Environmental Protection Act/Federal-Provincial Advisory Committee) 1998. *National Ambient Air Quality Objectives for Particulate Matter: Executive Summary*, *Part 1*. Science Assessment Document, 19p.

Chow JC, Watson JG, Lowenthal DH, Solomon PA, Magliano KL, Ziman SD, Richards LW. 1992. PM_{10} source apportionment in California's San Joaquin Valley. *Atmospheric Environment* **26A**:3335–3354.

Dockery DW, Pope III CA, Xu X, Spengler JD, Ware ME, Fay BG, Ferris J, Speizer FE. 1993. An association between air pollution and mortality in six U.S. cities. *New England Journal of Medicine* **329**:1753–1759.
Du Z, Wiens DP. 2000. Jackknifing, weighting, diagnostics and variance estimation in generalized M-estimation. *Statistics and Probability Letters* **46**:287–299.
Hampel FR, Ronchetti EM, Rousseeuw PJ, Stahel WA. 1986. *Robust Statistics: The Approach Based on Influence Functions*. Wiley: New York.
Hinkley DV. 1977. Jackknifing in unbalanced situations. *Technometrics* **19**:285–292.
Hopke PK. 1985. *Receptor Modeling in Environmental Chemistry*. Wiley: New York.
Hopke PK (ed.). 1991. *Receptor modelling for air quality management*. *Data Handing in Science and Technology*, Vol. 7. Elsevier: Amsterdam.
Huber PJ. 1981. *Robust Statistics*. Wiley: New York.
Lowenthal DH, Wittorff D, Gertler AW, Sakiyama S. 1997. CMB source apportionment during REVEAL. *Journal of Environmental Engineering* **123**:80–87.
Ohtaki M, Sato M, Nitta H. 1997. Estimating source apportionment of particulate matters based on source profiles with fluctuations. *Environmetrics* **8**:341–350.
Simpson DG, Chang Y-CI. 1997. Reweighting approximate GM estimators: asymptotics and residual-based graphs. *Journal of Statistical Planning and Inference* **57**:273–293.
Watson JG, Cooper JA, Huntzicker JJ. 1984. The effective variance weighting for Least Squares calculations applied to the Mass Balance Receptor Model. *Atmospheric Environment* **18**:1347–1355.
Watson JG, Chow JC, Pace TG. 1991. Chemical mass balance. In *Receptor Modeling for Air Quality Management*, Hopke PK (ed.). Elsevier: Amsterdam; 83–116.
Welsch RE. 1980. Regression sensitivity analysis and bounded influence estimation. In *Evaluation of Econometric Models*, Kmenta J, Ramsey JB (eds). Academic Press: New York; 153–167.
Wu CFJ. 1986. Jackknife, bootstrap and other resampling methods in regression analysis (with discussion). *Annals of Statistics* **14**:1261–1295.

