ENVIRONMETRICS
Environmetrics 2006; **17**: 663–681
Published online in Wiley InterScience (www.interscience.wiley.com). DOI: 10.1002/env.782

Iterated confirmatory factor analysis for pollution source apportionment

William F. Christensen[1,*,†], James J. Schauer[2] and Jeff W. Lingwall[1]

[1]*Department of Statistics, Brigham Young University, Provo, UT 84602-6575, USA*
[2]*Water Chemistry Program, University of Wisconsin, Madison, WI, USA*

SUMMARY

Many approaches for pollution source apportionment have been considered in the literature, most of which are based on the chemical mass balance equations. The simplest approaches for identifying the pollution source contributions require that the pollution source profiles are known. When little or nothing is known about the nature of the pollution sources, exploratory factor analysis, confirmatory factor analysis, and other multivariate approaches have been employed. In recent years, there has been increased interest in more flexible approaches, which assume little knowledge about the nature of the pollution source profiles, but are still able to produce nonnegative and physically realistic estimates of pollution source contributions. Confirmatory factor analysis can yield a physically interpretable and uniquely estimable solution, but requires that at least some of the rows of the source profile matrix be known. In the present discussion, we discuss the iterated confirmatory factor analysis (ICFA) approach. ICFA can take on aspects of chemical mass balance analysis, exploratory factor analysis, and confirmatory factor analysis by assigning varying degrees of constraint to the elements of the source profile matrix when iteratively adapting the hypothesized profiles to conform to the data. ICFA is illustrated using $PM_{2.5}$ data from Washington D.C., and a simulation study illustrates the relative strengths of ICFA, chemical mass balance approaches, and positive matrix factorization (PMF). Copyright © 2006 John Wiley & Sons, Ltd.

KEY WORDS: air quality modeling; latent variable models; chemical mass balance

1. INTRODUCTION

Pollution source apportionment is the attempt to partition ambient pollutants to the sources from which they were emitted. In recent years, it has been hypothesized that similar pollutants originating from different sources may have different levels of toxicity. For example, elemental carbon emanating from diesel exhaust and incinerators may have different impacts on human health. Consequently, quantifying the impact of various pollution sources is an important precursor to understanding the relationship between pollution and health, and to the creation and evaluation of sound environmental policy.

*Correspondence to: William F. Christensen, Department of Statistics, Brigham Young University, 219 TMCB, Provo, UT 84602-6575, USA.
†E-mail: william@stat.bru.edu

Contract/grant sponsor: Health Effects Institute.
Contract/grant sponsor: STAR Research Assistance Agreement; contract/grant number: RD-83216001-0.
The article has not been formally reviewed by the EPA. The views expressed in this document are solely those of the authors and the EPA does not endorse any products or commercial services mentioned in this publication.

Received 29 October 2004
Revised 14 July 2005

Several approaches for pollution source apportionment have been considered in the literature, each of which is based on the simple model

$$\boldsymbol{x}_t = \boldsymbol{\Lambda}\boldsymbol{f}_t + \boldsymbol{e}_t \tag{1}$$

where $\boldsymbol{x}_t$ is the vector of p chemical species observed at time t, $\boldsymbol{f}_t$ is the vector of k pollution source contributions at time t, $\boldsymbol{e}_t$ is the vector of measurement errors associated with the p species observed at time t, and $\boldsymbol{\Lambda}$ is a $p \times k$ matrix with the k pollution source profiles as its columns. Each profile contains the fractions of a pollution source that correspond to various chemical species. The simplest approach for identifying the pollution source contributions $\boldsymbol{f}_t$ involves standard regression methods (e.g., Gatz, 1975; Mayrsohn and Crabtree, 1976; Kowalczyk *et al*., 1978), but this approach requires that the pollution source profiles $\boldsymbol{\Lambda}$ are known to the researcher. When little or nothing is known about the nature of the pollution sources, exploratory factor analysis models have been employed (e.g., Thurston and Spengler, 1985; Koutrakis and Spengler, 1987; Henry *et al.*, 1994). Figure 1 illustrates the continuum of *a priori* knowledge about the nature of the pollution sources and the variety of approaches that exist. Other techniques include confirmatory factor analysis models (e.g., Yang, 1994;

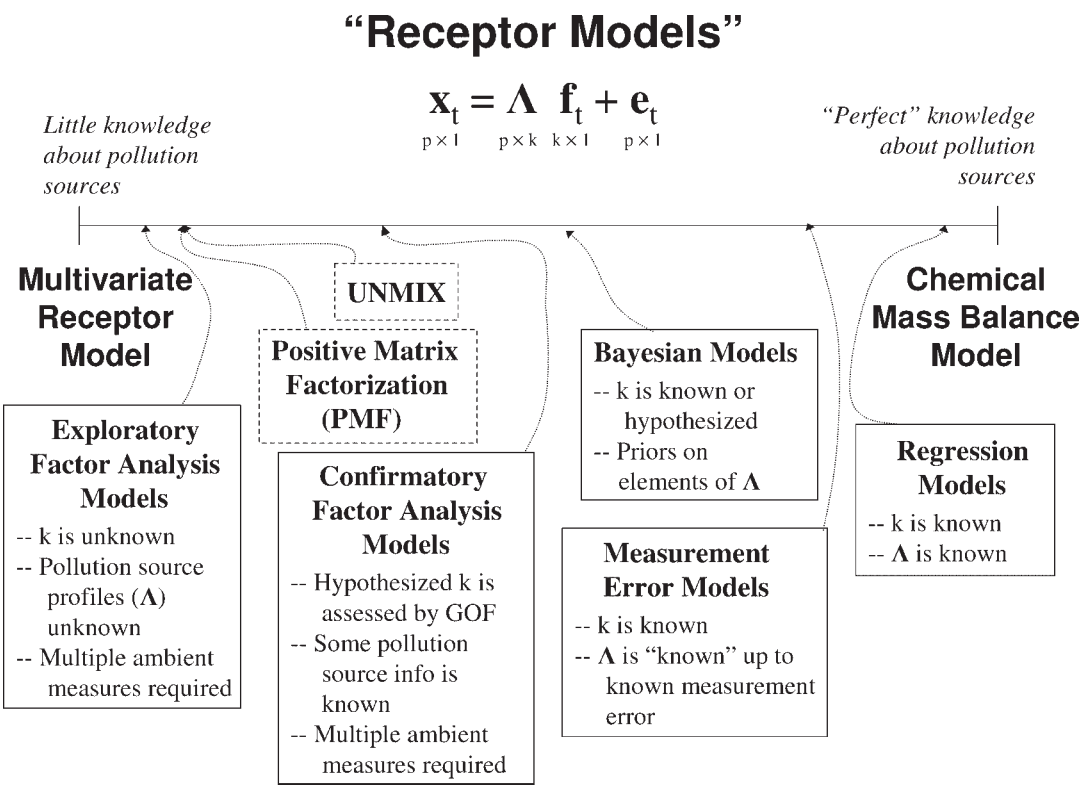

Figure 1. A classification of approaches for estimating pollution source contributions using receptor models. On the far left, little is known about the number and nature and pollution sources. On the far right, 'perfect' knowledge about the pollution sources is available. Approaches for estimating pollution source contributions depend upon the location on the continuum of profile knowledge

Gleser, 1997; Christensen and Sain, 2002), measurement error modeling (Watson *et al*., 1984; Christensen and Gunst, 2004), and Bayesian analysis (e.g., Park *et al*., 2001, 2002). Although the somewhat simplistic Figure 1 places Bayesian models at the middle of the continuum of *a priori* knowledge, we note that the flexibility of the Bayesian approach allows one to use these models at virtually any place on the continuum. That is, one can place prior distributions of varying degrees of vagueness/specificity on source profiles, source contribution amounts, or (as noted by one of the referees) even the number of sources in the model.

In a common pollution source apportionment scenario, the researcher has estimates of pollution source profiles (and their uncertainties) along with ambient species measurements (and their uncertainties). For this common setting, the effective variance (EV) solution of Watson *et al*. (1984) has become a gold standard. Implemented in software available from the U.S. Environmental Protection Agency (see EPA, 2004), the EV approach utilizes uncertainties associated with the profiles and the ambient measure in order to estimate an optimal weight in an iteratively re-weighted least squares algorithm. The weighted least squares (WLS) solution to Equation (1) is similar to the EV solution except that it does not utilize profile uncertainty information. Although the WLS solution does not make full use of available information, it has been argued that the WLS approach may be preferable to the EV approach, particularly when profile estimates are subject to substantial error or the appropriate set of pollution sources has not been correctly identified (Christensen and Gunst, 2004; Christensen, 2004). The EV and WLS solutions and their associated assumptions are discussed in detail by Christensen and Gunst (2004).

In recent years, there has been increased interest in more flexible approaches, which assume little knowledge about the nature of the pollution source profiles. Both the Unmix approach (outlined in Henry, 1997) and the positive matrix factorization (PMF) approach (Paatero and Tapper, 1994) are examples of this type of analysis. Unmix and PMF attempt to extract nonnegative estimates of source profiles and source contributions from the measured chemical species using a factor analytic model. While each approach attempts to reduce factor indeterminacy through the use of nonnegativity constraints and other computational tools, neither these nor any other purely exploratory approach can guarantee a uniquely identified solution without additional constraints on the source profiles (which act like 'factor loadings' when the receptor model is viewed as a factor analysis model). Confirmatory factor analysis can guarantee a uniquely estimable solution, but requires that at least k rows of $\mathbf{\Lambda}$ be fixed to equal a full rank matrix of constants.

A notable member of the class of factor analysis approaches used in source apportionment studies is the target transformation factor analysis (TTFA) which was developed by Weiner *et al*. (1970) and implemented in receptor modeling by Hopke *et al*. (1980) and others. TTFA uses exploratory factor analysis to extract profiles (or factor loadings) from the data and then rotates the factor loadings to align as closely as possible with the hypothesized profiles. This approach is similar to the iterated confirmatory factor analysis (ICFA) approach to be discussed here in that each approach utilizes both the factor analysis structure and *a priori* profile information when estimating source profiles. As we will discuss later, the ICFA approach has some advantages in that it is more robust in the presence of poorly specified a priori source profiles.

In the present discussion, we present the ICFA approach, which can take on aspects of both confirmatory factor analysis (CFA) and exploratory factor analysis (EFA) by assigning varying degrees of constraint to each element of $\mathbf{\Lambda}$ during the estimation process. The degree of constraint placed upon an assumed source profile is based on prior knowledge about the nature of the source. In fact, the ICFA approach has been developed to mimic the flexibility of the Bayesian approach but in a simple and computationally inexpensive fashion. Section 2 describes the ICFA approach and Section 3

gives an illustration of the method applied to $PM_{2.5}$ data from Washington D.C. In Section 4, we provide simulations comparing ICFA with the EV solution, WLS, and PMF. Conclusions and discussion are given in Section 5.

2. ITERATED CONFIRMATORY FACTOR ANALYSIS (ICFA)

Before discussing our approach for estimating the source profile matrix, we review a few key aspects of confirmatory factor analysis. Recall that when a k-factor model holds, the factor loadings for the k factors are uniquely estimable if a $k \times k$ submatrix of the $p \times k$ factor loading matrix is fixed to equal a $k \times k$ full-rank matrix of constants. For multivariate receptor modeling, an implication of this property is that when a $k \times k$ submatrix of the source profile matrix is fixed, the factors represented by the estimated factor loadings will retain the physical interpretation implied by the choice of the 'constraint matrix.' We consider here the case in which some or all of the elements in a hypothesized source profile (factor loading) matrix are estimated, but with error. We allow for some other elements to be unknown. When the k-factor model holds, but errors are present in the estimated profiles, we can use the factor analysis model to update $p - q$ rows of the factor loading matrix based on a full-rank constraint matrix consisting of q rows of the source profile matrix, where $q \geq k$. Thus, our approach has similarities with the EM algorithm, which uses known quantities to update estimates of unknown quantities.

Beginning with all elements of the profile matrix, which are available to the researcher, one can obtain estimates of the unknown elements via a factor analysis solution. We employ a maximum likelihood estimate of the factor model parameters, and constrain factor loading estimates to lie in the interval [0,1]. Such constraints along with restrictions to ensure nonnegativity of EVs and factor variances are easily implemented in standard software such as the CALIS procedure in the SAS Software System (SAS Institute Inc., 1999). After obtaining a complete $p \times k$ factor loading matrix in the first iteration, we then allow the factor analysis model to update the factor loading estimates to be consistent with the observed ambient data by successively re-estimating $p - q$ randomly chosen rows of the factor loading matrix. Each update uses the data and the qk constraint matrix consisting of a full-column-rank subset of $q > k$ rows of the factor loading matrix estimate from the previous iteration. If the k-factor model holds and approximate source profiles have been identified, each iteration will yield source profile matrix estimates that improve upon the starting values. If one or more of the real (and important) source profiles are erroneously omitted from the initial estimate of $\mathbf{\Lambda}$ or if an incorrect profile is used, the source profile matrix estimates will diverge from the starting values towards a more plausible solution.

After repeated iterations of the algorithm, during which time each element of the factor loading matrix has been re-estimated many times, the source profile matrix will converge toward an estimate which is associated with a smaller chi-squared goodness-of-fit statistic. In each iteration of our implementation, we randomly select a row of the source profile matrix to be re-estimated with probability in (0.1, 0.5) so that each iteration will estimate a random number of rows. One could also randomly select a fixed number of rows to be re-estimated in each iteration. However, one chooses the rows to be re-estimated in each iteration, in order for the model in each update step to be 'conditionally identified' (i.e., identified given the $q \times k$ 'constraint matrix' obtained from the previous update) it is vital that the constraint matrix not only have at least k rows, but also be of rank k. If the model in each update is not conditionally identified, then the algorithm can yield final profile estimates that have rotated away from a starting value matrix that was correct and completely specified. In practice, the conditional identification condition prevents the profile matrix estimates from vacillating wildly

among equally well-fitting but dramatically different solutions, many of which will be physically uninterpretable.

In some situations, we may be reasonably certain about the nature of one or more of the source profiles, but have only an approximate estimate for other source profiles. For example, we might have an accurate estimate of a gasoline vehicle exhaust profile, but have only a suspicion that a chlorine-rich profile may be either an incinerator or a smelter. In such cases, we can constrain the profile elements associated with the well-understood source $\lambda_{i1}, i = 1, \ldots, p$, to remain in the interval $(4/5 \times \lambda_{i1}, 5/4 \times \lambda_{i1})$ while allowing the profile elements associated with the poorly understood source $\lambda_{i2}, i = 1, \ldots, p$, to take on updated values in the interval $(1/5 \times \lambda_{i2}, 5 \times \lambda_{i2})$. Profile error variances can also be used directly. Other times we may wish to leave some source profiles completely unrestricted in order to account for potentially unknown sources. In these scenarios, we would like a model fitting procedure that will give us uniqueness (or near-uniqueness) for the source profile estimates associated with the well-understood profiles and still have the flexibility to estimate unknown source profiles. Specifying bounds on elements of $\mathbf{\Lambda}$ allows the flexible approach to take on aspects of chemical mass balance analysis (for well defined profiles), confirmatory factor analysis (for partially or weakly defined profiles), and exploratory factor analysis (for unidentified profiles). From a Bayesian perspective, one might view these bounds as a uniform prior on each profile matrix element. While we implicitly assume a stationary structure for the repeated observations, the statistical properties of maximum likelihood estimation of the factor analysis model parameters are unaffected by correlation structure and trend in the factor vectors under very weak assumptions about the factor process (Anderson and Amemiya, 1988; Christensen and Amemiya, 2003).

The ICFA approach differs from TTFA in that each update of the ICFA algorithm will move the estimated profiles away from the erroneous values (which are not substantiated by the data structure) towards more plausible profiles. In contrast, the TTFA approach first finds an EFA solution, and then rotates the profiles until they are most closely in line with the target (or hypothesized) profiles. If in fact, the correct source profile is excluded from the columns of the target profile matrix and an incorrect profile is included in its place, the TTFA approach will rotate the solution in order to find transformed factor loadings that resemble the incorrect profiles. Consequently, all other profiles will be inaccurately represented by the rotated solution. In this sense, TTFA 'squeezes' the data into the hypothesized profiles while ICFA adapts the hypothesized profiles to conform to the data. However, it should be noted that if one or more major sources are missing from the profile matrix (i.e., too few sources are used in the model), one might 'rotate away' from the correct profile in order for the factors to capture the major components of variability in the ambient data. For this reason, continuing work in model goodness-of-it assessment is important to source apportionment research (see e.g., Park *et al.*, 2002; Christensen and Sain, 2002).

Once the ICFA approach has converged to reasonably stable updated estimates of the source profiles, there is usually interest in calculating source contribution estimates for each ambient observation. We consider two approaches for constraining the contribution estimates to be non-negative. In a nonlinear estimation approach, we define the contribution from the jth source at time t as $f_{jt} = h_{jt}^2$ and we employ a nonlinear model of the form

$$x_{it} = \hat{\lambda}_{i1} h_{1t}^2 + \cdots + \hat{\lambda}_{ik} h_{kt}^2, \quad i = 1, \ldots, p \tag{2}$$

where x_{it} is the abundance of the ith species at time t and $\hat{\lambda}_{ij}$ is the proportional representation of the ith species in the jth source. Estimating h_{jt} in the nonlinear model instead of estimating f_{jt} in a linear

model guarantees that the source contributions will be positive. A Gauss–Newton algorithm is used to fit the model to the data. In order to ensure an estimate that is not just a local minimum, the model was fit multiple times using different randomly chosen starting values each time. Among the candidate solutions, the one with smallest SSE was chosen as the final source contribution estimate.

A second approach is to obtain the WLS estimate of $(f_{1t}, \ldots, f_{kt})$ using the linear model

$$x_{it} = \hat{\lambda}_{i1} f_{1t} + \cdots + \hat{\lambda}_{ik} f_{kt}, \quad i = 1, \ldots, p \tag{3}$$

Let $\hat{f}_{i't}$ be the minimum contribution estimate among $\hat{f}_{1t}, \ldots, \hat{f}_{kt}$. If $\hat{f}_{i't}$ is negative, we set it to 0 and re-estimate the other elements of $\boldsymbol{f}_t$ using only $(\hat{\lambda}_{1t}, \ldots, \hat{\lambda}_{(i'-1)t}, \hat{\lambda}_{(i'+1)t}, \ldots, \hat{\lambda}_{kt})$. This process is repeated until all elements of $\boldsymbol{f}_t$ are greater than or equal to 0. In simulation studies such as those illustrated in Section 4, we found that the constrained linear estimation of source contributions often yields more stable estimates than the nonlinear estimation approach. However, when working with some real-world data sets the constrained linear estimation produces a high proportion of contribution estimates equal to zero. Because an abundance of zero-valued contributions is physically unrealistic for many pollution sources associated with real-world data sets, the nonlinear estimates are sometimes preferred to the constrained linear estimates because they yield more realistic physical interpretations.

3. ANALYSIS OF WASHINGTON DC $PM_{2.5}$ DATA

We illustrate the ICFA approach using 23 species of $PM_{2.5}$ measured approximately twice weekly from 1988 until 1997. The data include trace elements (by X-ray fluorescence), anions (sulfate and nitrate), and organic and elemental carbon (from thermal optical analysis). A more detailed description of the data is found on the IMPROVE web site (http://vista.cira.colostate.edu/improve) and in Kim and Hopke (2004). The data and metadata from Washington DC and other IMPROVE sites are freely accessible from the Visibility Information Exchange Web System at http://vista.cira.colostate.edu/views/Web/Data/DataWizard.aspx. For this analysis, we have available profiles for wood burning, an oil-fired power plant, soil dust, auto/diesel emissions, sea salt, and a lead smelter. In addition to the six available profiles listed above, we included two unspecified profiles to account for secondary or unknown sources affecting the airshed. The initial profile matrix for this analysis is given in Table 1.

Recall that if we have greater certainty about the validity of some profiles relative to others, we can reflect this in the analysis by imposing tighter bounds on the updated elements of the profile. That is, when an element is re-estimated in an iteration of the algorithm, we can constrain that estimate to fall within a small (or large) distance from the starting value. For the analysis of the Washington DC data, we simply required the elements of each of the six profiles to fall within the range $(0.5 \times \lambda_{ij}, 2.0 \times \lambda_{ij})$, where λ_{ij} is the ith element of the jth initial profile. If the ith species was not speciated in the jth profile (i.e., if λ_{ij} is unknown), then the estimated element of the profile is only constrained to fall in [0,1]. All elements of the unspecified profiles are constrained only to fall within [0,1]. All computations were carried out using SAS Proc Calis with the SAS Macro language employed to carry out the iterative calculations (SAS Institute Inc., 1999).

In the first iteration of the algorithm, we estimate all of the missing components of the profile matrix while fixing all other elements to their initially specified values. The values in the profile matrix after the first iteration are given in Table 2. Note that three of the profiles now sum to a value greater than one. The scaling is arbitrary for these profiles and will be re-scaled to sum to one after the final iteration.

Table 1. Initial profile matrix for Washington D.C. data

	Wood (Spec.)	Oil power (Spec.)	Un-known	Soil (Spec.)	Un-known	Auto/dies. (Watson *et al*., 1994)	Sea salt (Javitz *et al*., 1988)	Lead smelter (Spec.)
Al	0.0010	0.0010	—	0.0599	—	0.0041	0.0000	0.0053
As	0.0000	0.0003	—	0.0000	—	0.0000	0.0000	0.1047
Br	0.0000	0.0000	—	0.0000	—	0.0003	0.0000	0.0014
Ca	0.0010	0.0061	—	0.0049	—	0.0071	0.0140	0.0120
Cl	0.0100	0.0014	—	0.0000	—	0.0034	0.4000	0.0148
Cu	0.0000	0.0000	—	0.0000	—	0.0007	0.0000	0.0489
EC	0.1000	0.1280	—	0.0100	—	0.1350	0.0000	—
Fe	0.0000	0.0127	—	0.0378	—	0.0068	0.0000	0.0200
K	0.0300	0.0015	—	0.0209	—	0.0025	0.0140	0.0124
Mn	0.0000	0.0002	—	0.0010	—	0.0010	0.0000	0.0009
Na	0.0010	0.0000	—	0.0032	—	0.0000	0.2600	—
Ni	0.0000	0.0112	—	0.0000	—	0.0000	0.0000	0.0014
NO_3	0.0050	0.0002	—	0.0000	—	0.0389	0.0000	—
OC	0.4500	0.0500	—	0.0400	—	0.3008	0.0000	—
P	0.0000	0.0003	—	0.0000	—	0.0011	0.0000	0.0005
Pb	0.0000	0.0003	—	0.0000	—	0.0016	0.0000	0.2247
Se	0.0000	0.0000	—	0.0000	—	0.0000	0.0000	0.0009
Si	0.0030	0.0026	—	0.3500	—	0.0164	0.0000	0.0172
SO_4	0.0030	0.5000	—	0.0000	—	0.0229	0.0000	—
Sr	0.0000	0.0000	—	0.0004	—	0.0000	0.0000	0.0005
Ti	0.0000	0.0005	—	0.0057	—	0.0007	0.0000	0.0007
V	0.0000	0.0052	—	0.0001	—	0.0000	0.0000	0.0000
Zn	0.0004	0.0010	—	0.0000	—	0.0027	0.0000	0.1135

Sources for initial profiles are given. 'Spec.' indicates that the profile is based on source(s) found in Speciate, version 3.2, which is a source profile database available from U.S. E.P.A. Elements marked with '—' are not available.

In the second and all subsequent iterations, a small number of rows of the profile matrix are re-estimated while all other rows are held constant. In each iteration, we re-estimate a row of $\mathbf{\Lambda}$ with probability of 0.25 but we check to ensure that at least 8 rows of the most recent estimate of $\mathbf{\Lambda}$ are used as a constant matrix in the current iteration. To illustrate, the values in the profile matrix after the second iteration are given in Table 3. The rows randomly selected for re-estimation in the second iteration corresponded to the species EC, Na, Ni, Pb, and V.

In each successive iteration, the χ^2 goodness-of-fit statistic decreases, indicating better agreement between the data and the current profiles. We continue the process until the improvement of the goodness-of-fit is small. After 100 iterations, the changes in the goodness-of-fit statistic are barely changing and we consider the algorithm to have converged. At this point, many of the profiles (columns of $\mathbf{\Lambda}$) now sum to greater than 1, so they are re-scaled to sum to 1. We recognize that the 23 elements are not likely to constitute a complete speciation of the source being re-scaled. However, absent any other information about the combined contribution of the 23 species to the total $PM_{2.5}$ produced by the source, we are required to choose some value in [0,1] to which we must scale the summed profile. By scaling to 1, we provide a lower bound on the corresponding source contribution estimates. The final profile matrix after re-scaling problematic profiles is given in Table 4. Figure 2 illustrates the differences between the initial and final profiles.

Table 2. Profile matrix after the first iteration

	Wood	Oil power	Un-known	Soil	Un-known	Auto/dies.	Sea salt	Lead smelter
Al	0.0010	0.0010	**0.2418**	0.0599	**0.0626**	0.0041	0.0000	0.0053
As	0.0000	0.0003	**0.0178**	0.0000	**0.0276**	0.0000	0.0000	0.1047
Br	0.0000	0.0000	**0.0079**	0.0000	**0.0078**	0.0003	0.0000	0.0014
Ca	0.0010	0.0061	**0.0855**	0.0049	**0.0622**	0.0071	0.0140	0.0120
Cl	0.0100	0.0014	**0.1314**	0.0000	**0.4115**	0.0034	0.4000	0.0148
Cu	0.0000	0.0000	**0.0174**	0.0000	**0.0175**	0.0007	0.0000	0.0489
EC	0.1000	0.1280	**0.5006**	0.0100	**0.0124**	0.1350	0.0000	**0.4996**
Fe	0.0000	0.0127	**0.3535**	0.0378	**0.2186**	0.0068	0.0000	0.0200
K	0.0300	0.0015	**0.3134**	0.0209	**0.0977**	0.0025	0.0140	0.0124
Mn	0.0000	0.0002	**0.0044**	0.0010	**0.0016**	0.0010	0.0000	0.0009
Na	0.0010	0.0000	**0.0478**	0.0032	**0.0792**	0.0000	0.2600	**0.5581**
Ni	0.0000	0.0112	**0.0438**	0.0000	**0.1017**	0.0000	0.0000	0.0014
NO_3	0.0050	0.0002	**0.5170**	0.0000	**0.5102**	0.0389	0.0000	**0.7112**
OC	0.4500	0.0500	**0.5052**	0.0400	**0.4998**	0.3008	0.0000	**0.4995**
P	0.0000	0.0003	**0.0001**	0.0000	**0.0049**	0.0011	0.0000	0.0005
Pb	0.0000	0.0003	**0.0459**	0.0000	**0.0627**	0.0016	0.0000	0.2247
Se	0.0000	0.0000	**0.0027**	0.0000	**0.0025**	0.0000	0.0000	0.0009
Si	0.0030	0.0026	**0.5825**	0.3500	**0.0177**	0.0164	0.0000	0.0172
SO_4	0.0030	0.5000	**0.0661**	0.0000	**0.4871**	0.0229	0.0000	**0.4945**
Sr	0.0000	0.0000	**0.0047**	0.0004	**0.0000**	0.0000	0.0000	0.0005
Ti	0.0000	0.0005	**0.0209**	0.0057	**0.0000**	0.0007	0.0000	0.0007
V	0.0000	0.0052	**0.0313**	0.0001	**0.0627**	0.0000	0.0000	0.0000
Zn	0.0004	0.0010	**0.0517**	0.0000	**0.0628**	0.0027	0.0000	0.1135

Elements of the matrix estimated in this iteration appear in bold face.

The names and interpretations for some of the profiles had to be changed to reflect the structure of the final profile. Specifically, it was recognized the heavy sulfates in the oil-fired power plant profile could also be due to a secondary sulfate. Because these sources are indistinguishable using these species, we changed the name of this profile to 'Secondary Sulfate/Oil-Fired Power Plant.' The first of the two originally unspecified sources is heavy in NO_3, and is suspected to be a winter-heavy secondary source. The second of the two originally unspecified sources is more evenly split among EC, NO_3, OC, and SO_4. It is tentatively referred to as 'Secondary 2,' although the large proportion of EC and the spiky nature of the contributions indicate that this source may be due in part (or in whole) to a point source such as ships. Alternatively, this source could be an incremental diesel truck effect, which is prevalent on weekdays. In fact, the estimated source is slightly more prevalent on weekdays when diesels constitute a higher fraction of the traffic vehicle mix. On weekdays we might expect higher concentrations of EC from the roadway as well as metals from truck-related resuspended road dust (Pb and Zn). Interestingly, this source has a substantial amount of each of these elements. Notwithstanding the problems with assuming this source as secondary formation and the presence of alternate interpretations, we temporarily retain the label 'Secondary 2' because there is no other conclusion that is clearly satisfactory. Perhaps most likely is an interpretation that this source is a combination of several sources. These types of source estimates are reminders that the pollution source problem is inherently complex, and interpretations should be made cautiously. Lastly, the final version of the profile originally called 'Lead Smelter' changed enough that we now refer to it as simply 'Lead Source.' We surmise that this may be an industrial source.

Table 3. Profile matrix after the second iteration

	Wood	Oil power	Un-known	Soil	Un-known	Auto/dies.	Sea salt	Lead smelter
Al	0.0010	0.0010	0.2418	0.0599	0.0626	0.0041	0.0000	0.0053
As	0.0000	0.0003	0.0178	0.0000	0.0276	0.0000	0.0000	0.1047
Br	0.0000	0.0000	0.0079	0.0000	0.0078	0.0003	0.0000	0.0014
Ca	0.0010	0.0061	0.0855	0.0049	0.0622	0.0071	0.0140	0.0120
Cl	0.0100	0.0014	0.1314	0.0000	0.4115	0.0034	0.4000	0.0148
Cu	0.0000	0.0000	0.0174	0.0000	0.0175	0.0007	0.0000	0.0489
EC	**0.0500**	**0.0640**	**0.5168**	**0.0050**	**0.0828**	**0.4217**	**0.0000**	**0.4802**
Fe	0.0000	0.0127	0.3535	0.0378	0.2186	0.0068	0.0000	0.0200
K	0.0300	0.0015	0.3134	0.0209	0.0977	0.0025	0.0140	0.0124
Mn	0.0000	0.0002	0.0044	0.0010	0.0016	0.0010	0.0000	0.0009
Na	**0.0020**	**0.0000**	**0.3300**	**0.0064**	**0.3653**	**0.0000**	**0.2158**	**0.4360**
Ni	**0.0000**	**0.0056**	**0.0155**	**0.0000**	**0.0550**	**0.0000**	**0.0000**	**0.0007**
NO_3	0.0050	0.0002	0.5170	0.0000	0.5102	0.0389	0.0000	0.7112
OC	0.4500	0.0500	0.5052	0.0400	0.4998	0.3008	0.0000	0.4995
P	0.0000	0.0003	0.0001	0.0000	0.0049	0.0011	0.0000	0.0005
Pb	**0.0000**	**0.0002**	**0.0597**	**0.0000**	**0.0760**	**0.0016**	**0.0000**	**0.3268**
Se	0.0000	0.0000	0.0027	0.0000	0.0025	0.0000	0.0000	0.0009
Si	0.0030	0.0026	0.5825	0.3500	0.0177	0.0164	0.0000	0.0172
SO_4	0.0030	0.5000	0.0661	0.0000	0.4871	0.0229	0.0000	0.4945
Sr	0.0000	0.0000	0.0047	0.0004	0.0000	0.0000	0.0000	0.0005
Ti	0.0000	0.0005	0.0209	0.0057	0.0000	0.0007	0.0000	0.0007
V	**0.0000**	**0.0106**	**0.0336**	**0.0001**	**0.1086**	**0.0000**	**0.0000**	**0.0000**
Zn	0.0004	0.0010	0.0517	0.0000	0.0628	0.0027	0.0000	0.1135

Elements of the matrix estimated in this iteration appear in bold face.

The interpretation of the updated profiles is facilitated by inspecting time series plots of the source contributions. For these data, we estimated the source contributions using model (2) and the nonlinear estimation approach discussed in Section 2. The nonlinear approach was preferred to the constrained linear estimation in this example because the resulting estimates exhibited more physically interpretable output. Plots of the source contributions are given in Figure 3. As we should expect, wood burning and the nitrate-heavy (winter) secondary source have wintertime peaks. Also, the secondary sulfate/oil-fired power plant source, the soil source, and the sea salt source each have summertime peaks as expected.

4. SIMULATION STUDIES

Our purpose for these computer simulations is to compare the ICFA estimator with the EV estimator (Watson *et al*., 1984), the WLS approach advocated in Christensen and Gunst (2004) and Christensen (2004), and the PMF method (Paatero and Tapper, 1994). Our focus in the present discussion is on estimation as opposed to statistical inference.

In practice, we never know the precise values for the profiles in $\mathbf{\Lambda}$. Additionally, it is often the case that we either fail to include all important sources in $\mathbf{\Lambda}$ or we include profiles for nonpresent or unimportant sources. In these simulations, we consider the performance of estimators when source profiles are subject to measurement error and when source profiles are subject to measurement error and profile misspecification.

Table 4. Final profile matrix (after the hundredth iteration)

	Wood	Second. sulfate/ Oil power	Winter second.	Soil	Sec. 2?	Auto/diesel	Sea salt	Lead source (industrial?)
Al	0.0011	0.0049	0.0377	0.1149	0.0087	0.0045	0.0000	0.0017
As	0.0000	0.0003	0.0034	0.0000	0.0027	0.0000	0.0000	0.0291
Br	0.0000	0.0000	0.0009	0.0000	0.0006	0.0001	0.0000	0.0016
Ca	0.0011	0.0311	0.0155	0.0060	0.0216	0.0075	0.0093	0.0134
Cl	0.0133	0.0073	0.0192	0.0000	0.0342	0.0014	0.2649	0.0164
Cu	0.0000	0.0000	0.0066	0.0000	0.0047	0.0004	0.0000	0.0543
EC	0.3965	0.1771	0.0000	0.0489	0.1831	0.3606	0.0000	0.5551
Fe	0.0000	0.0162	0.0491	0.0818	0.0091	0.0052	0.0000	0.0222
K	0.0339	0.0076	0.0098	0.0671	0.0027	0.0015	0.0371	0.0034
Mn	0.0000	0.0002	0.0007	0.0012	0.0003	0.0005	0.0000	0.0003
Na	0.0045	0.0000	0.0588	0.0156	0.0780	0.0000	0.6887	0.0000
Ni	0.0000	0.0142	0.0060	0.0000	0.0101	0.0000	0.0000	0.0016
NO_3	0.0226	0.0010	0.2852	0.0000	0.1831	0.0377	0.0000	0.0000
OC	0.5084	0.0638	0.0222	0.1956	0.1831	0.5214	0.0000	0.0000
P	0.0000	0.0005	0.0002	0.0000	0.0006	0.0007	0.0000	0.0001
Pb	0.0000	0.0016	0.0288	0.0000	0.0234	0.0006	0.0000	0.2494
Se	0.0000	0.0000	0.0003	0.0000	0.0001	0.0000	0.0000	0.0003
Si	0.0034	0.0134	0.1575	0.4505	0.0506	0.0228	0.0000	0.0191
SO_4	0.0136	0.6379	0.2767	0.0000	0.1831	0.0337	0.0000	0.0000
Sr	0.0000	0.0000	0.0008	0.0019	0.0000	0.0001	0.0000	0.0005
Ti	0.0000	0.0006	0.0061	0.0162	0.0016	0.0003	0.0000	0.0002
V	0.0000	0.0171	0.0076	0.0002	0.0122	0.0000	0.0000	0.0000
Zn	0.0017	0.0051	0.0068	0.0000	0.0066	0.0013	0.0000	0.0315

Note that the name and interpretation of some profiles have changed due to the data-driven adjustment of initial profiles.

Note that the performance of the EV, WLS, and ICFA approaches for estimating source contributions is directly related to the *a priori* information about the source profile matrix. In its most basic formulation, the PMF approach does not utilize any *a priori* information. However, when using so-called 'G-keying' approaches in PMF (Paatero, 1998), one can introduce 'target shapes' for the profiles. In each replication of our simulations, we use the PMF2 program in two ways: the first estimate of profiles and contributions is based on the standard approach with no additional constraints for limiting rotational ambiguity; the second estimate uses G-keying for incorporating the *a priori* information about the source profile matrix as target profiles in the analysis. After conducting a preliminary simulation study in order to optimize the somewhat complex PMF settings for these data, PMF settings were selected which accomplished the purposes of near optimization with relative simplicity. For the simulations discussed herein, the standard deviations of the ambient data matrix ($\mathbf{X}$) were calculated using EM $= -12$, reading ambient data uncertainties into PMF as a $\mathbf{T}$ matrix, and setting C1, C2, and C3 equal to 0, 0, and 0.01, respectively. Thus the standard deviation matrix for the $\mathbf{X}$ matrix is

$$\mathbf{S} = \mathbf{T} + 0.01\,|\mathbf{X}|$$

Fpeak and outlier threshold distance were left at the default values of 0 and 4, respectively.

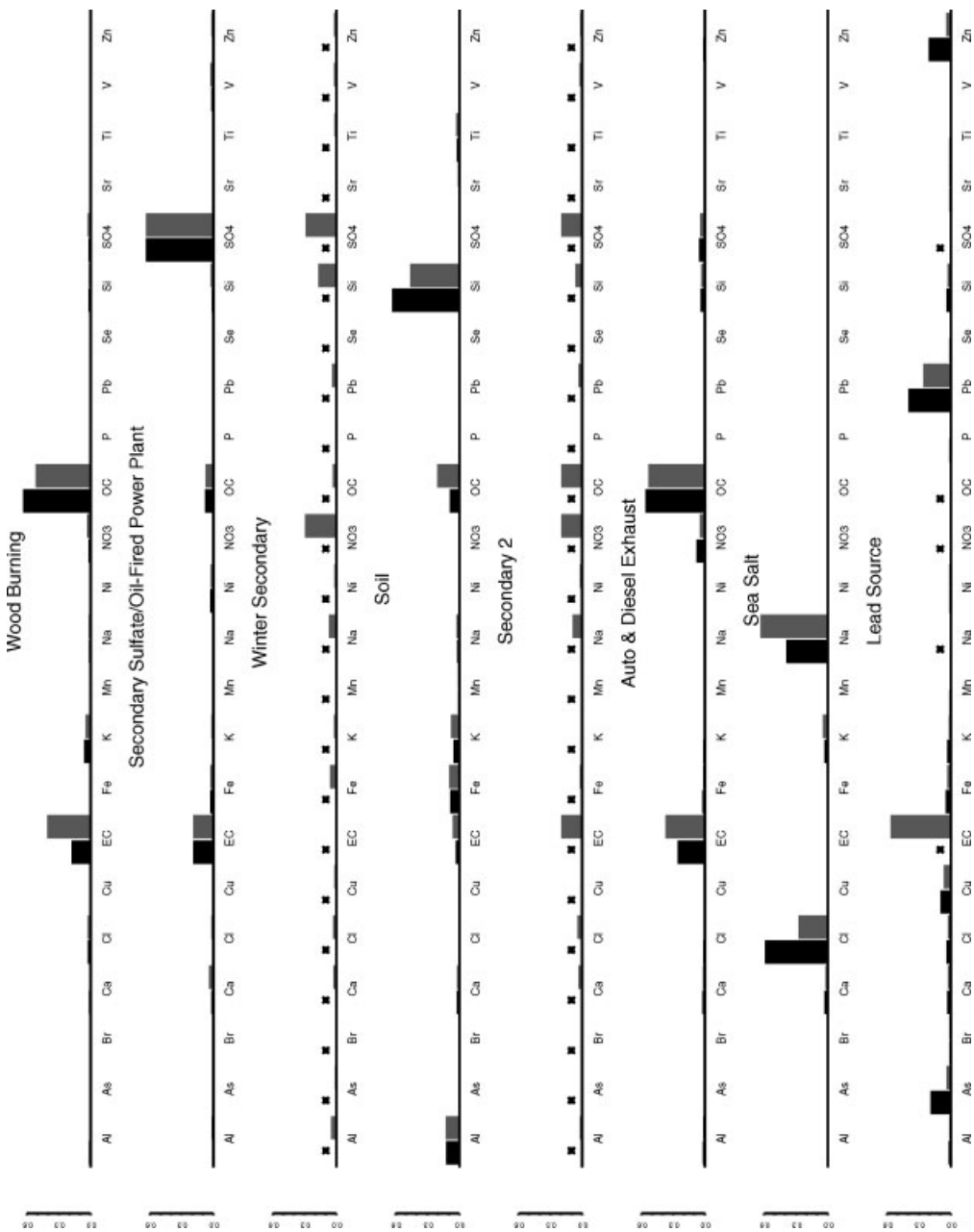

Figure 2. Profiles for the eight fitted sources. Black profiles are initial values and grey profiles are final values. An asterisk denotes that the element's value in the initial profile was unspecified

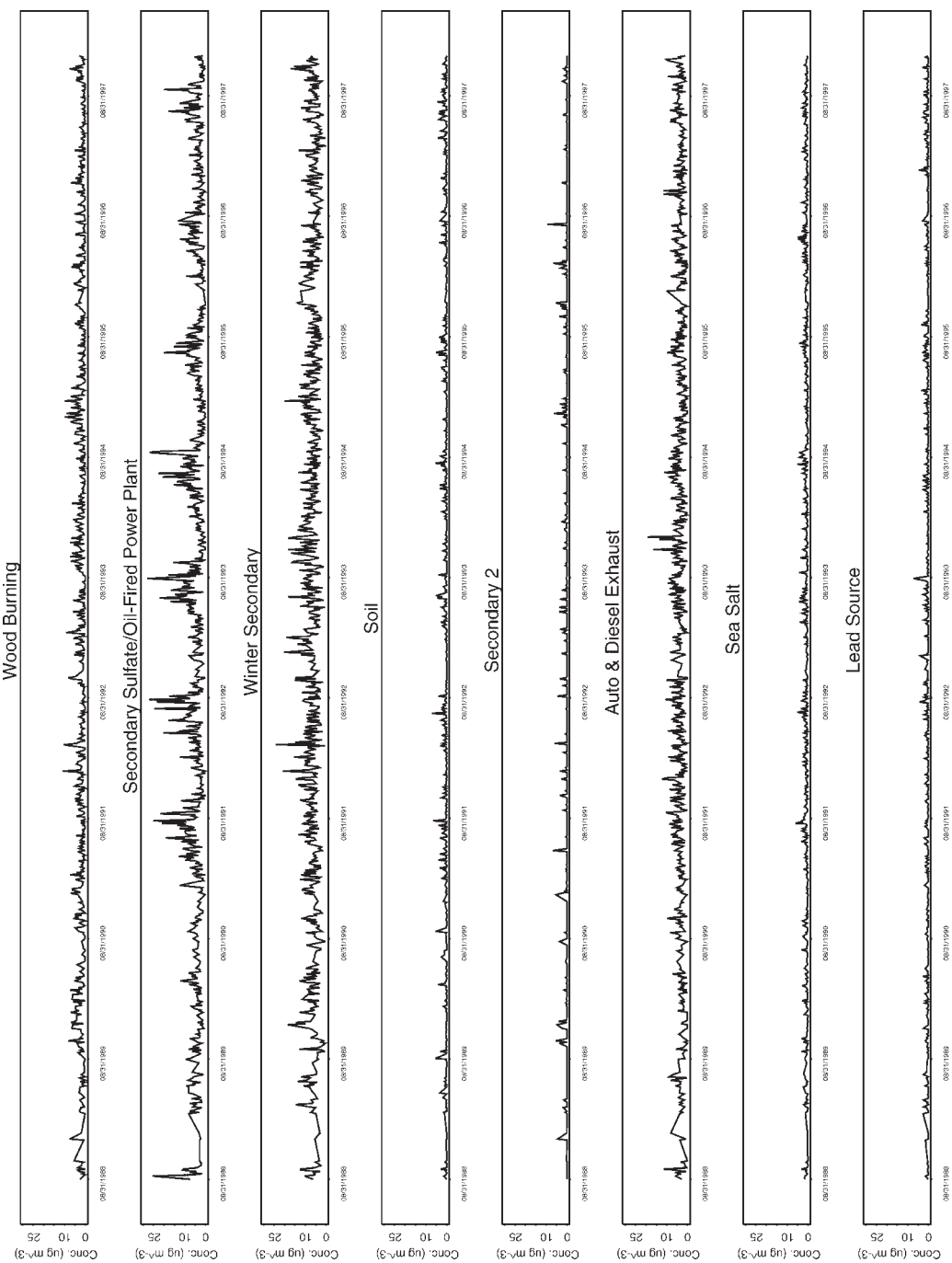

Figure 3. Contributions for the eight fitted sources

We note here that in every setting considered in the simulations, estimates of profiles and contributions were only slightly affected by the use of G-keying, rarely reducing the nonG-keying error rate by more than 5 per cent. Thus, in the interest of simplicity, all reported summaries for PMF are based on the first PMF estimate (with no G-keying). That is, the PMF estimates in these simulations are unaffected by the correctness of the *a priori* source profile matrix. Optimizing the G-keying and other settings for PMF is a topic of current research.

4.1. Simulation #1: No contamination of airshed by unidentified sources

To evaluate the performance of the ICFA and other source apportionment approaches, we simulate data based on the structure of the Washington DC data. Specifically, we simulate an airshed affected by five sources: secondary sulfate, winter secondary, soil, auto/diesel, and sea salt. The estimated profiles for these sources (found in Table 4) are used as the true profiles ($\mathbf{\Lambda}$) in the simulations, and the corresponding contribution estimates (found in Figure 3) are used as the true contributions ($\mathbf{F}$). To mimic the typically right-skewed nature of ambient pollutant distributions, we create an $n \times p$ ambient pollutant matrix $\mathbf{X} = (x_{ti})$ equal to $\mathbf{F}\mathbf{\Lambda}'$ plus a lognormal error. Specifically

$$x_{ti} \sim \text{Lognormal}(\lambda_i' \boldsymbol{f}_t, \text{CV} = 0.20) \tag{4}$$

where λ_i is the ith row of $\mathbf{\Lambda}$ and CV is the coefficient of variation, indicating that the standard deviation of the simulated species x_{ti} is 20 per cent the size of its true value ($\lambda_i' \boldsymbol{f}_t$). A total of 100 replications of the simulated data were obtained, each with a total of 788 observations.

Recall that $\mathbf{\Lambda} = (\lambda_{ij})$ is the true profile matrix for the airshed. In this simulation, we consider *a priori* profile matrices of the form $\mathbf{\Lambda}_c = (\lambda_{cij})$ or $\mathbf{\Lambda}_c^* = (\lambda_{cij}^*)$, $i = 1, \ldots, p$, $j = 1, \ldots, k$, where

$$\lambda_{cij} \sim \text{Lognormal}(\lambda_{ij}, \text{CV} = c) \tag{5}$$

and c is the coefficient of variation, taking on values 0.25, 0.5, or 1. Thus the standard deviation of λ_{ij} is $c \times \lambda_{cij}$. The *a priori* profile matrix $\mathbf{\Lambda}_c^*$ is the same as $\mathbf{\Lambda}_c$ except that $\mathbf{\Lambda}_c^*$ has a misspecified profile for secondary sulfate and the winter secondary source is unspecified. A comparison of the true and misspecified secondary sulfate profiles is given in Table 5. When $\mathbf{\Lambda}_c^*$ is used as the *a priori* profile matrix, the EV and WLS approaches are unable to estimate the contribution from winter secondary (since no profile is available). Hence, the EV and WLS approaches estimate only the other four sources, ignoring the potential presence of the unidentified winter secondary source. Despite the absence of complete profile information, both the ICFA and PMF approaches are able to fit the five source model.

The ICFA approach allows us to constrain the estimates of the elements of $\mathbf{\Lambda}$ according to the degree of credibility associated with each hypothesized profile. In this simulation, we create bounds for estimates of λ_{ij} using the uncertainty associated with this element of $\mathbf{\Lambda}$. As stated previously, we believe that the distribution of these profile errors are right-skewed (lognormal) so we do not want bounds that are symmetric around λ_{ij}. Instead we construct a $C \times 100$ per cent normality-based confidence region for each element on the log scale and then exponentiate the end points to obtain bounds for our estimates of λ_{ij}. For the elements of the profile for secondary sulfate, we use $C = 0.999$ to reflect a low degree of credibility for the *a priori* profile. For the elements of the other profiles we use $C = 0.90$ to reflect a moderate degree of credibility about the *a priori* profile. Thus, because the normal table value for a 99.9 per cent confidence interval is roughly twice the size as the table value for

Table 5. True and misspecified profiles for secondary sulfate source

	True	Misspecified
Al	0.0050	0.0187
As	0.0003	0.0001
Br	0.0000	0.0001
Ca	0.0311	0.0401
Cl	0.0074	0.0003
Cu	0.0000	0.0021
EC	0.1771	0.0038
Fe	0.0162	0.0251
K	0.0076	0.0032
Mn	0.0002	0.0004
Na	0.0000	0.0476
Ni	0.0142	0.0286
NO_3	0.0010	0.0003
OC	0.0638	0.0302
P	0.0005	0.0155
Pb	0.0016	0.0149
Se	0.0000	0.0000
Si	0.0134	0.0509
SO_4	0.6379	0.6760
Sr	0.0000	0.0005
Ti	0.0006	0.0023
V	0.0171	0.0281
Zn	0.0051	0.0111

a 90 per cent confidence interval, the bounds on the profile elements associated with a low degree of credibility will have twice the relative spread as the bounds for profiles associated with a moderate degree of credibility. When a source profile is completely unspecified (as with the winter secondary profile in $\mathbf{\Lambda}_c^*$) we let the estimated elements fall anywhere in [0,1]. After convergence is attained, all columns of the estimate of $\mathbf{\Lambda}$ are re-scaled to sum to no greater than 1.

To compare the performance of the various estimators, we use the true source contributions ($\mathbf{F}$) and the estimates ($\hat{\mathbf{F}}$) from a given estimator to calculate the average absolute error (AAE_j) associated with the jth source:

$$\mathrm{AAE}_j = \frac{1}{n}\sum_{t=1}^{n} |\hat{f}_{tj} - f_{tj}| \tag{6}$$

where f_{tj} is the contribution at time t for the jth source. When using $\mathbf{\Lambda}_c$ as the *a priori* profile matrix, each of the approaches (EV, WLS, ICFA, and PMF) can be used to obtain estimated contributions from all five sources. We then calculate the total contribution AAE using all five sources

$$\mathrm{AAE}_5 = \mathrm{AAE}_{\mathrm{sec.sulf}} + \mathrm{AAE}_{\mathrm{winter.sec}} + \mathrm{AAE}_{\mathrm{soil}} + \mathrm{AAE}_{\mathrm{auto/diesel}} + \mathrm{AAE}_{\mathrm{salt}} \tag{7}$$

When using $\mathbf{\Lambda}_c^*$ as the *a priori* profile matrix, the winter secondary source contributions cannot be estimated with the EV and WLS approaches. Consequently, when using $\mathbf{\Lambda}_c^*$, in order to compare all

four approaches, we consider both AAE_5 in Equation (7) and the total contribution AAE of only the four 'known' sources (secondary sulfate, soil, auto/diesel, and salt) using

$$AAE_4 = AAE_{sec.sulf} + AAE_{soil} + AAE_{auto/diesel} + AAE_{salt} \tag{8}$$

We note here that because PMF (absent G-keying or F-keying) gives an arbitrary ordering of the five sources, we make sure to conduct all of our analyses of the PMF output using the ordering of the PMF sources that minimizes the mean squared error between the PMF estimates and the true source contributions.

Total contribution AAE under six different settings for the *a priori* profile information is given in Table 6. As discussed in Christensen and Gunst (2004) and Christensen (2004), the EV and WLS methods are comparable in terms of AAE in low profile error scenarios, but the WLS approach is superior when profile errors are large or when unidentified sources affect the airshed. For the scenarios using $\hat{\mathbf{\Lambda}}_c$, when error variances at their lowest level, the EV approach performed best among the methods. When error variances were at the middle and highest levels, ICFA and PMF performed best, respectively. When profiles are not fully specified (i.e., when using $\hat{\mathbf{\Lambda}}_c^*$), PMF performs best among the estimators. An interesting facet of ICFA is that when profile misspecification exists, the approach has lower AAE values when the profile error variances are larger. As profile errors increase, the constraints placed on the elements of the estimated profile matrix are loosened, allowing the profile matrix to be more dramatically altered by the data structure. Thus, when misspecification exists and the profile error variances are larger, the increased flexibility in accounting for the misspecification more than compensates for the increased variability introduced into the estimation problem.

We are also interested in how well the methods estimate the true source profile matrix $\mathbf{\Lambda}$. We follow a similar approach for quantifying the performance of the methods. When using $\mathbf{\Lambda}_c$, we calculate

$$AAE_{profile,j} = \frac{1}{p}\sum_{i=1}^{p} |\hat{\lambda}_{ij} - \lambda_{ij}| \tag{9}$$

where λ_{ij} is the true proportional representation of the *i*th species in the *j*th source. Next we calculate the total profile AAE for the five sources

$$\begin{aligned} AAE_{5profile} = {} & AAE_{profile,sec.sulf} + AAE_{profile,winter.sec} + AAE_{profile,soil} \\ & + AAE_{profile,auto/diesel} + AAE_{profile,salt} \end{aligned} \tag{10}$$

Table 6. Simulation when the airshed is free of unidentified sources

A priori profile matrix	EV	WLS	ICFA	PMF
$\hat{\mathbf{\Lambda}}_{0.25}$ (AAE_5)	**2.96**	3.55	3.38	4.32
$\hat{\mathbf{\Lambda}}_{0.50}$ (AAE_5)	4.60	5.30	**3.84**	4.32
$\hat{\mathbf{\Lambda}}_{1.00}$ (AAE_5)	12.47	8.47	5.18	**4.32**
$\hat{\mathbf{\Lambda}}^*_{0.25}$ (AAE_5/AAE_4)	na/*8.94*	na/*9.23*	8.31/*6.28*	**4.32**/*3.55*
$\hat{\mathbf{\Lambda}}^*_{0.50}$ (AAE_5/AAE_4)	na/*18.47*	na/*9.63*	7.16/*5.71*	**4.32**/*3.55*
$\hat{\mathbf{\Lambda}}^*_{1.00}$ (AAE_5/AAE_4)	na/*40.93*	na/*10.40*	5.23/*3.67*	**4.32**/*3.55*

AAE_5 from Equation (7) is reported for six different settings for the *a priori* profile information. PMF estimates did not utilize *a priori* info about $\mathbf{\Lambda}$. For settings using $\hat{\mathbf{\Lambda}}_c^*$ as the *a priori* profile matrix, AAE_4 from Equation (8) is also reported (in *italics*). The best AAE_5 in each row is boldfaced.

Table 7. Simulation when the airshed is free of unidentified sources

A priori profile matrix	EV	WLS	ICFA	PMF
$\hat{\mathbf{\Lambda}}_{0.25}$ ($AAE_{5profile}$)	0.036	0.036	**0.019**	0.045
$\hat{\mathbf{\Lambda}}_{0.50}$ ($AAE_{5profile}$)	0.065	0.065	**0.026**	0.045
$\hat{\mathbf{\Lambda}}_{1.00}$ ($AAE_{5profile}$)	0.107	0.107	**0.041**	0.045
$\hat{\mathbf{\Lambda}}^*_{0.25}$ ($AAE_{5profile}/AAE_{4profile}$)	na/*0.043*	na/*0.043*	0.061/*0.040*	**0.045**/*0.038*
$\hat{\mathbf{\Lambda}}^*_{0.50}$ ($AAE_{5profile}/AAE_{4profile}$)	na/*0.061*	na/*0.061*	0.065/*0.044*	**0.045**/*0.038*
$\hat{\mathbf{\Lambda}}^*_{1.00}$ ($AAE_{5profile}/AAE_{4profile}$)	na/*0.085*	na/*0.085*	0.063/*0.043*	**0.045**/*0.038*

$AAE_{5profile}$ from Equation (10) is reported for six different settings for the *a priori* profile information. PMF estimates did not utilize *a priori* info about $\mathbf{\Lambda}$. For settings using $\hat{\mathbf{\Lambda}}^*_c$ as the *a priori* profile matrix, $AAE_{4profile}$ from Equation (11) is also reported (in *italics*). The best $AAE_{5profile}$ in each row is boldfaced.

When using $\hat{\mathbf{\Lambda}}^*_c$, estimates for winter secondary cannot be obtained because the EV and WLS approaches cannot incorporate sources with unknown profiles in the estimation process. Consequently, when using $\hat{\mathbf{\Lambda}}^*_c$, we base our comparisons on both $AAE_{5profile}$ in Equation (10) and the total profile AAE of only the four 'known' sources using

$$AAE_{4profile} = AAE_{profile,sec.sulf} + AAE_{profile,soil} + AAE_{profile,auto/diesel} + AAE_{profile,salt} \quad (11)$$

Because the EV and WLS approaches do not update the profile matrix beyond the *a priori* values, $AAE_{5profile}$ and $AAE_{4profile}$ for these methods reflect only the accuracy of the *a priori* profiles as an estimate of the true profiles in $\mathbf{\Lambda}$. That is, $AAE_{5profile}$ and $AAE_{4profile}$ serve as baselines for comparing other approaches. Total profile AAE under six different settings for the *a priori* profile information is given in Table 7. For all scenarios using $\hat{\mathbf{\Lambda}}_c$, ICFA yielded the best estimates of $\mathbf{\Lambda}$. For all scenarios using $\hat{\mathbf{\Lambda}}^*_c$, PMF yielded the best estimates of $\mathbf{\Lambda}$. Thus, for this scenario in which the data are not affected by excessive error or unidentified minor sources and the proper number of major sources is selected for the model, ICFA only outperforms PMF for profile estimation when the *a priori* profile information is reasonably correct. As with the estimation of source contributions, when using $\hat{\mathbf{\Lambda}}^*_c$ as the *a priori* profile matrix, an increase in the variability associated with the estimation is offset by the increased flexibility in accounting for misspecification so that the AAE for ICFA is not affected as c increases from 0.25 to 1.00.

4.2. Simulation #2: Airshed contaminated by unidentified sources

The second simulation has the same structure as the simulation in subsection 4.1, but now the ambient data is affected by small amounts of unidentified minor sources—the wood burning, lead, and secondary #2 sources illustrated in Figure 3. Specifically, the minor sources amount to 20 per cent of the wood burning contributions, 50 per cent of the lead source contributions, and 50 per cent of the secondary #2 source contributions. In total, the unidentified sources increase the total daily $PM_{2.5}$ by an average of less than 5 per cent. Hence, we consider these to be minor sources and not of direct interest in our problem. The total source AAE and total profile AAE for these simulations are given in Tables 8 and 9, respectively. Many of the conclusions are similar to those for the simulations in subsection 4.1, but there are some new insights. First, we observe the hazard of using the EV approach when even minor unidentified sources affect the airshed as noted by Christensen (2004). Second, although PMF still performs best in terms of estimating source contributions, we see that ICFA is

Table 8. Simulation when the airshed is contaminated by minor unidentified sources

A priori profile matrix	EV	WLS	ICFA	PMF
$\hat{\mathbf{\Lambda}}_{0.25}$ (AAE_5)	**3.31**	3.63	4.23	4.56
$\hat{\mathbf{\Lambda}}_{0.50}$ (AAE_5)	5.38	5.46	5.02	**4.56**
$\hat{\mathbf{\Lambda}}_{1.00}$ (AAE_5)	13.92	8.76	6.26	**4.56**
$\hat{\mathbf{\Lambda}}^*_{0.25}$ (AAE_5/AAE_4)	na/*10.20*	na/*9.95*	10.19/*6.85*	**4.56**/*3.70*
$\hat{\mathbf{\Lambda}}^*_{0.50}$ (AAE_5/AAE_4)	na/*20.21*	na/*10.32*	9.25/*6.31*	**4.56**/*3.70*
$\hat{\mathbf{\Lambda}}^*_{1.00}$ (AAE_5/AAE_4)	na/*45.26*	na/*11.06*	8.01/*5.02*	**4.56**/*3.70*

AAE_5 from Equation (7) is reported for six different settings for the *a priori* profile information. PMF estimates did not utilize *a priori* info about $\mathbf{\Lambda}$. For settings using $\hat{\mathbf{\Lambda}}^*_c$ as the *a priori* profile matrix, AAE_4 from Equation (8) is also reported (in *italics*). The best AAE_5 in each row is boldfaced.

Table 9. Simulation when the airshed is contaminated by minor unidentified sources

A priori profile matrix	EV	WLS	ICFA	PMF
$\hat{\mathbf{\Lambda}}_{0.25}$ ($AAE_{5\text{profile}}$)	0.036	0.036	**0.023**	0.117
$\hat{\mathbf{\Lambda}}_{0.50}$ ($AAE_{5\text{profile}}$)	0.065	0.065	**0.029**	0.117
$\hat{\mathbf{\Lambda}}_{1.00}$ ($AAE_{5\text{profile}}$)	0.107	0.107	**0.042**	0.117
$\hat{\mathbf{\Lambda}}^*_{0.25}$ ($AAE_{5\text{profile}}/AAE_{4profile}$)	na/*0.043*	na/*0.043*	**0.070**/*0.040*	0.117/*0.112*
$\hat{\mathbf{\Lambda}}^*_{0.50}$ ($AAE_{5\text{profile}}/AAE_{4profile}$)	na/*0.061*	na/*0.061*	**0.074**/*0.047*	0.117/*0.112*
$\hat{\mathbf{\Lambda}}^*_{1.00}$ ($AAE_{5\text{profile}}/AAE_{4profile}$)	na/*0.085*	na/*0.085*	**0.077**/*0.051*	0.117/*0.112*

$AAE_{5\text{profile}}$ from Equation (10) is reported for six different settings for the *a priori* profile information. PMF estimates did not utilize *a priori* info about $\mathbf{\Lambda}$. For settings using $\hat{\mathbf{\Lambda}}^*_c$ as the *a priori* profile matrix, $AAE_{4\text{profile}}$ from Equation (11) is also reported (in *italics*). The best $AAE_{5\text{profile}}$ in each row is boldfaced.

clearly better than the other approaches with respect to total profile AAE even when the degree of profile error is high and the profiles are misspecified.

5. DISCUSSION AND CONCLUSIONS

The motivation behind the development of ICFA is to address the indeterminacy or identifiability issues that affect factor analysis approaches to source apportionment. Approaches like PMF and Unmix reduce the degree of indeterminacy or 'rotational ambiguity' of solutions by incorporating nonnegativity constraints, but are still subject to debatable interpretations when the practically unrealistic assumptions for a unique solution are not met. In fact, this identifiability problem applies to virtually all multivariate receptor modeling tools including ICFA. In developing ICFA, we propose an approach that attempts to maximally utilize *a priori* source profile information to reduce indeterminacy and enhance interpretability in a multivariate receptor modeling scenario. In this sense, one could think of ICFA as an inexpensive Bayesian approach to the problem. Current research involves a more fully Bayesian approach to the source apportionment problem.

The ICFA approach provides unique insights and promises value as it is developed as a new source apportionment approach. We note that although the estimates of source contributions are guaranteed to be nonnegative when using Equation (2) or (3), PMF and Unmix have the advantage of simultaneously

estimating the profiles and contributions, whereas ICFA sequentially estimates the profiles and the contributions.

In terms of source contribution estimation, ICFA in its current implementation does not perform as well as PMF when profile errors are large or when misspecified sources exist. However, ICFA estimates source profiles better than the other methods considered, especially when there is contamination by unidentified sources. The primary interest in pollution source apportionment studies lies in the quantification of source contributions. However, because of the good performance of ICFA in estimating source profiles (particularly when important profiles are initially missing or misspecified), improvements in the second phase of the ICFA procedure (estimating source contributions based on the previously estimated profile matrix) is a topic of current research that is well-motivated.

In these analyses and discussions, we have made many of the restrictive assumptions that have been associated with the source apportionment problem for over 30 years. Watson *et al*. (1991) provide a review of such assumptions and the hazards of violating them. Some challenges to current source apportionment methodology include properly accounting for temporal variability in ambient concentrations, incorporating meteorological and spatial effects, addressing the temporally varying nature of pollution source profiles, and handling the problems associated with multicollinear source profiles or 'profile clusters.' The issue of multicollinearity is an important issue both technically and philosophically because what one might call a source profile might in fact be a group of sources that are related in constitution or are temporally correlated. Thus, great caution must be used in interpreting an estimated source, particularly when the nature of the profile evolves dramatically during the ICFA algorithm. Whether profile evolution is due to original misspecification of the profile, contamination from unidentified sources, multicollinearity, or temporal indistinguishability with other sources is a question that still must be carefully considered. Addressing these issues in both frequentist and fully Bayesian frameworks is matter of current research.

ACKNOWLEDGMENTS

This work was supported by the Health Effects Institute and by the STAR Research Assistance Agreement No. RD-83216001-0 awarded by the U.S. Environmental Protection Agency. The helpful comments of the editor and anonymous reviewers led to improvements in this article and are acknowledged with gratitude.

REFERENCES

Anderson TW, Amemiya Y. 1988. The asymptotic normal distribution of estimators in factor analysis under general conditions. *The Annals of Statistics* **16**: 759–771.

Christensen WF. 2004. Chemical mass balance analysis of air quality data when unknown pollution sources are present. *Atmospheric Environment* **38**: 4305–4317.

Christensen WF, Amemiya Y. 2003. Modeling and prediction for multivariate spatial factor analysis. *Journal of Statistical Planning and Inference* **115**: 543–564.

Christensen WF, Gunst RF. 2004. Measurement error models in chemical mass balance analysis of air quality data. *Atmospheric Environment* **38**: 733–744.

Christensen WF, Sain SR. 2002. Accounting for dependence in a flexible multivariate receptor model. *Technometrics* **44**: 328–337.

Environmental Protection Agency. 2004. EPA-CMB8.2 User's Manual, EPA Publication No. EPA-452/R-04-011. Office of Air Quality Planning & Standards, Research Triangle Park, NC.

Gatz DF. 1975. Relative contributions of different sources of urban aerosols: application of a new estimation method to multiple sites in Chicago. *Atmospheric Environment* **9**: 1–18.

Gleser LJ. 1997. Some thoughts on chemical mass balance models. *Chemometrics and Intelligent Laboratory Systems* **37**: 15–22.

Henry RC. 1997. History and fundamentals of multivariate air quality receptor models. *Chemometrics and Intelligent Laboratory Systems* **37**: 525–530.
Henry RC, Lewis CW, Collins JF. 1994. Vehicle-related hydrocarbon source compositions from ambient data: The GRACE/SAFER method. *Environmental Science Technology* **28**: 823–832.
Hopke PK, Lamb RF, Natusch DF. 1980. Multi-elemental characterization of urban roadway dust. *Environmental Science Technology* **14**: 164–172.
Javitz HS, Watson JG, Robinson N. 1988. Performance of the chemical mass balance model with simulated local-scale aerosols. *Atmospheric Environment* **22**: 2309–2322.
Kim E, Hopke PK. 2004. Source apportionment of fine particles at Washington DC utilizing temperature resolved carbon fractions. *Journal of Air and Waste Management Association* **54**: 773–785.
Koutrakis P, Spengler JD. 1987. Source apportionment of ambient particles in Steubenville, OH using specific rotation factor analysis. *Atmospheric Environment* **21**: 1511–1519.
Kowalczyk GS, Choquette CE, Gordon GE. 1978. Chemical element balances and identification of air pollution sources in Washington DC. *Atmospheric Environment* **12**: 1143–1153.
Mayrsohn H, Crabtree, JH. 1976. Source reconciliation of atmospheric hydrocarbons. *Atmospheric Environment* **10**: 137–143.
Paatero P. 1998. *User's guide for positive matrix factorization programs PMF2 and PMF3*.
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 ES, Guttorp P, Henry RC. 2001. Multivariate receptor modeling for temporally correlated data by using MCMC. *Journal of the American Statistical Association* **96**: 1171–1183.
Park ES, Oh MS, Guttorp P. 2002. Multivariate receptor models and model uncertainty. *Chemometrics and Intelligent Laboratory Systems* **60**: 49–67.
SAS Institute Inc. 1999. *SAS/STAT User's Guide, Version 8*. SAS Institute Inc., Cary, NC.
Thurston GD, Spengler JD. 1985. A quantitative assessment of source contributions to inhalable particulate matter pollution in metropolitan Boston. *Atmospheric Environment* **19**: 9–25.
Watson JG, Chow JC, Lowenthal DH, Pritchett LC, Frazier CA, Neuroth GR, Robbins R. 1994. Differences in the carbon composition of source profiles for diesel- and gasoline-powered vehicles. *Atmospheric Environment* **28**: 2493–2505.
Watson JG, Chow JC, Pace TG. 1991. Chemical mass balance. In *Receptor Modeling for Air Quality Management*, Hopke PK (ed). Elsevier Science Publishers: New York; 83–116.
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.
Weiner PH, Malinowski ER, Levinston AR. 1970. Factor analysis of solvent shifts in proton magnetic resonance. *Journal of Physical Chemistry* **74**: 4537–4542.
Yang H. 1994. Confirmatory factor analysis and its application to receptor modeling. Unpublished Ph.D. dissertation, University of Pittsburgh, Department of Mathematics and Statistics.

