Available online at www.sciencedirect.com
ScienceDirect
Chemometrics and Intelligent Laboratory Systems xx (2007) xxx–xxx
Chemometrics and intelligent laboratory systems
www.elsevier.com/locate/chemolab

Pollution source apportionment using *a priori* information and positive matrix factorization

Jeff W. Lingwall [a], William F. Christensen [b,*]

[a] *Department of Statistics, Carnegie Mellon University, Pittsburgh, PA 15213, U.S.A.*
[b] *Department of Statistics, Brigham Young University, Provo, UT 84602, U.S.A.*

Received 2 June 2006; received in revised form 20 March 2007

Abstract

The use of *a priori* information in positive matrix factorization (PMF) is examined in the context of pollution source apportionment. The impact of PMF's general run control settings is evaluated and simulation experiments are employed to illustrate the relative advantages and hazards associated with different uses of *a priori* information. Pulling source profile elements to zero appears to be uniformly beneficial when using data with low measurement error and no contamination from unknown sources. However, the benefit of F element pulling is less pronounced when data are subject to higher degrees of measurement error and when some elements are erroneously pulled to zero. The use of source profile targeting shows much promise, both for incorporating well-established knowledge about pollution sources and as a tool for incremental exploratory analysis of the data. A data analysis of the latter type is illustrated using $PM_{2.5}$ data from the St. Louis Supersite.

Keywords: PMF; Multivariate receptor modeling; Chemical mass balance; Air pollution; Source attribution

1. Introduction

1.1. Pollution source apportionment

Pollution source apportionment is the practice of using ambient air pollution data to derive information about pollution sources and the amount of pollution each emits. Various methods are employed, based on differing amounts of information that can be assumed about the number of polluting sources and their compositions. Factor analysis techniques can be used when the pollution sources are unknown, using the equation

$$\mathbf{X} = \mathbf{GF} + \mathbf{E} \tag{1}$$

where $\mathbf{G}$ is an $n \times p$ matrix containing pollution source contributions for the p sources, $\mathbf{F}$ is a $p \times m$ matrix of pollution source profiles, $\mathbf{E}$ is a matrix of errors, and $\mathbf{X}$ is an $n \times m$ matrix of measurements of m different chemical species observed at n times. Thus, for example, the concentration of species j observed at time i, x_{ij}, measured at a receptor can be explained as

$$\underset{1\times 1}{x_{ij}} = \underset{1\times pg_i\ p\times 1f_j}{\mathbf{g}_i\mathbf{f}_j} + e_{ij} \tag{2}$$

where $\mathbf{g}_i$ is the ith row of $\mathbf{G}$ and $\mathbf{f}_j$ is the jth column of $\mathbf{F}$.

1.2. Positive matrix factorization

Traditional factor analysis fails in an important aspect of the pollution source apportionment problem. In traditional factor analysis there are no non-negativity constraints on the results. Since air pollution studies require that contributions and profiles be non-negative, traditional factor analysis fails to provide the best solution. Alternatives exist, one of which is positive matrix factorization (PMF).

PMF [1] is a method related to factor analysis. In contrast to traditional factor analysis methods which decompose $\mathbf{X}$ based on the correlation or covariance matrix, PMF solves the factor

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

0169-7439/$ - see front matter © 2007 Elsevier B.V. All rights reserved.
doi:10.1016/j.chemolab.2007.03.007

Please cite this article as: J.W. Lingwall, W.F. Christensen, Pollution source apportionment using *a priori* information and positive matrix factorization, Chemometrics and Intelligent Laboratory Systems (2007), doi:10.1016/j.chemolab.2007.03.007

analysis equations by iteratively computing **F** and **G** via the minimization of

$$Q = \sum_{i=1}^{n} \sum_{j=1}^{m} e_{ij}^2 / s_{ij}^2 \quad (3)$$

where e_{ij} is calculated as $x_{ij} - \mathbf{g}_i \mathbf{f}_j$ and s_{ij} is the standard deviation or uncertainty associated with each data point [1]. PMF constrains **G** and **F** to be non-negative, thus satisfying the constraints necessary for realistic pollution source apportionment models. Additionally, when a "significant" number of zeros appear in all the columns of **F** and **G** the results of PMF may be unique [1]. In 1997, Paatero introduced a revised PMF algorithm called PMF2, which is examined in this paper and will be referred to as PMF [2].

1.3. A priori information in PMF

PMF allows the user to introduce prior information in its solution, giving researchers the opportunity to incorporate knowledge from previous studies or facts known about a source believed to be present in the airshed. In theory, the introduction of correct information should improve the pollution source estimates. PMF offers the user two different ways to introduce this information, by pulling source profile elements to zero and by using target source profiles. Both of these techniques in PMF use a "key" matrix, which give an integer value to each element of **F** or **G**, and pulls that element to zero based on the integer value.

This paper examines the use of a *priori* information in PMF, illustrating the use of these methods in the context of pollution source apportionment. A simulation study gives guidelines to the researcher about the effectiveness of incorporating *a priori* information about an airshed into a PMF analysis, suggesting that the use of target profiles can dramatically improve estimation of both source profiles and source contributions. Finally, the use of target source profiles as an exploratory analysis tool is illustrated using $PM_{2.5}$ data from the St. Louis airshed.

The United States Environmental Protection Agency recently introduced a free version of PMF available through its website [3]. This simplified version of PMF does not allow rotational adjustment through the use of Fpeak but does allow incorporation of *a priori* information through modification of the ".ini" file.

2. Simulation study methods

Six hundred simulated data sets were created, based on five sources: sea salt, secondary sulfate, winter secondary, soil, and

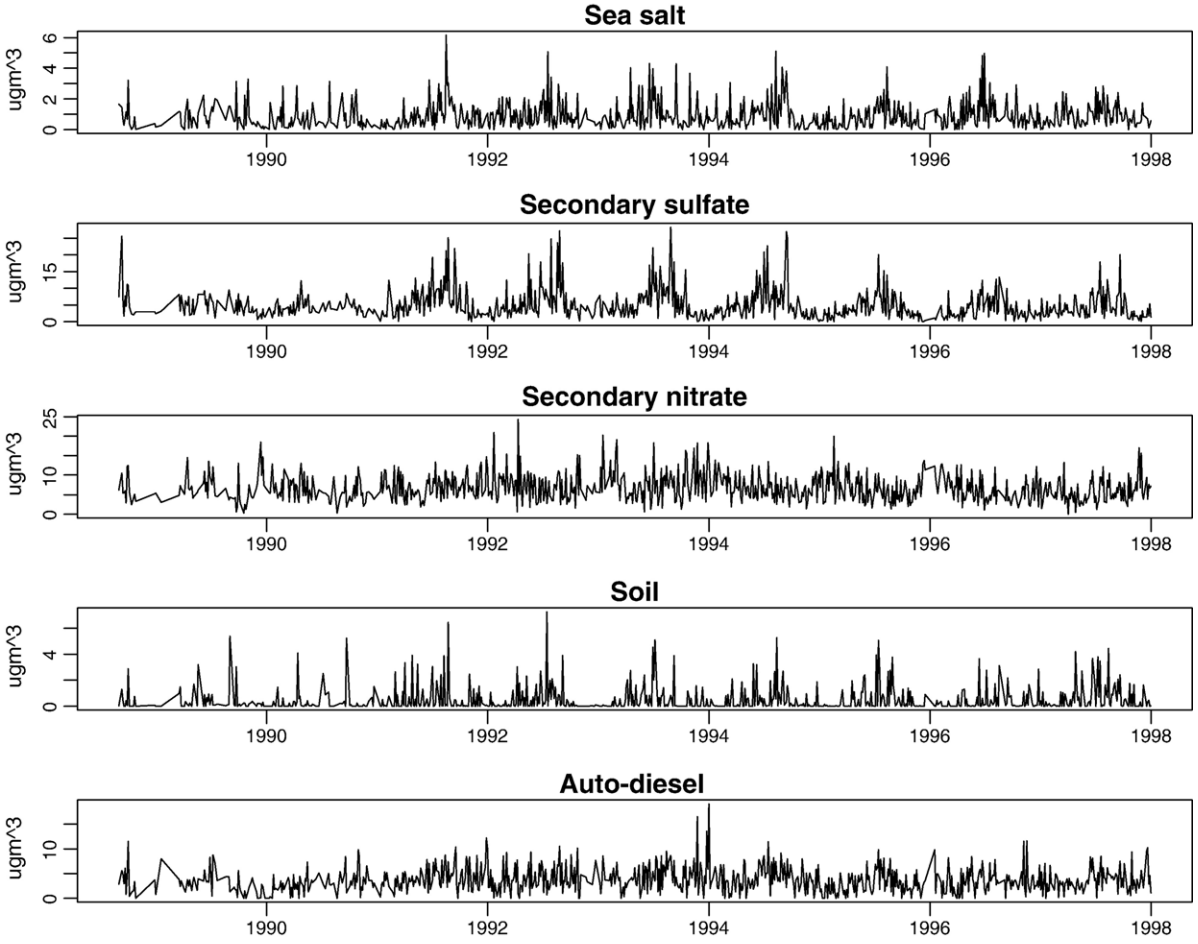

Fig. 1. True source contributions for simulated data (**G**).

Please cite this article as: J.W. Lingwall, W.F. Christensen, Pollution source apportionment using *a priori* information and positive matrix factorization, Chemometrics and Intelligent Laboratory Systems (2007), doi:10.1016/j.chemolab.2007.03.007

auto-diesel. Each simulated observation constituted measurements on 23 different chemical species. The five source profiles were combined to create an array of source profiles, **F**. Instead of randomly creating the source contributions, **G**, data that was reflective of the real world was used based on an analysis of an actual data set from Washington, DC. The results from the analysis were treated as if they were actually the source contributions. Thus, a natural degree of temporal correlation, seasonality, and daily and weekly fluctuations is built into the data. Although these five source contributions represent only one possibility of the wildly varying airsheds that exist in actual analysis, they allow a starting point that reflects real sources for study. These source contributions are shown in Fig. 1, while the source profiles are included as part of Table 3 (see Section 4).

Two different types of simulated data were considered. The first type of simulated data was based only on the five sources listed above, and the second was additionally influenced by three minor sources. These three minor sources, (a wood source, a lead smelter source, and an additional secondary source) would serve as unidentified minor sources in the model. In other words, we say that incorporating these three minor sources (described in Table 1) yields a 5-source airshed plus "contamination" by minor sources.

For the "uncontaminated data" simulation, data with three different degrees of noise were created, where an observation x_{ij} was obtained as a draw from a lognormal distribution with mean $\mathbf{g}_i\mathbf{f}_j$ and a coefficient of variation (CV) of either 0.2, 0.5, or 1.0. Note that draws from a lognormal distribution with mean of η and CV of ξ can be obtained by generating from $Z \sim N(\theta, \sigma^2)$ (where θ and σ are the mean and the standard deviation of Z) and then exponentiating the observed z. Using the expressions

$$\theta = \log(\eta) - \frac{1}{2}\log\left(\xi^2 + 1\right)$$
$$\sigma = \sqrt{\log\left(\xi^2 + 1\right)},$$

the values of θ and σ can be chosen so that $\exp(z)$ is lognormally distributed with a mean of η and a CV of ξ. For the "contaminated data" simulation, each observation x_{ij} was obtained as a draw from a lognormal distribution with mean $\mathbf{g}_i^*\mathbf{f}_j^*$, where $\mathbf{f}_j^*$ is the jth column of $\mathbf{F}^*$ which contains the three minor source profiles in addition to the five major sources, and $\mathbf{g}_i^*$ is the ith row of $\mathbf{G}^*$, which contains the three minor source contributions as well as the five major sources. The combined contributions for the three minor sources increase the particulate mass by an average of less than 5% (see [4]). As in the uncontaminated data scenario, three different CV's were used: 0.2, 0.5, and 1.0. One hundred data sets of each combination of CV (0.2, 0.5, and 1.0) and contamination status (uncontaminated and contaminated) were created, resulting in 600 simulated data sets that represent a wide range of potential airsheds.

After the creation of the simulated data, PMF was used to estimate **G** and **F** for each data set (with estimates denoted $\hat{\mathbf{G}}$ and $\hat{\mathbf{F}}$). Where PMF requires the receptor measurement uncertainties, we use CV×**X**, where CV is the same coefficient used when randomly generating the data matrix **X** (as described above). Because the sources derived by PMF are outputted in an arbitrary order, they needed to be rearranged to match the true sources before assessing the similarity between $\hat{\mathbf{G}}$ and **G** and between $\hat{\mathbf{F}}$ and **F**. Consequently, the labels associated with the five true sources were assigned to the PMF sources in a way that minimized the MSE (mean of the squared deviations) between columns of $\hat{\mathbf{G}}$ and columns of **G**.

Table 1
Profiles of "unidentified" minor sources which contaminate the airshed but are not included in the model fit

Species	Wood	Secondary	Lead smelter
Al	0.00113	0.00870	0.00174
As	0	0.00273	0.02905
Br	0	0.00063	0.00157
Ca	0.00113	0.02161	0.01337
Cl	0.01330	0.03419	0.01640
Cu	0	0.00467	0.05425
EC	0.39645	0.18306	0.55505
Fe	0	0.00907	0.02222
K	0.03389	0.00269	0.00343
Mn	0	0.00033	0.00026
Na	0.00452	0.07802	0
Ni	0	0.01011	0.00156
NO3	0.02259	0.18306	0
OC	0.50837	0.18306	0
P	0	0.00059	0.00013
Pb	0	0.02338	0.24938
Se	0	0.00011	0.00026
Si	0.00339	0.05063	0.01908
SO_4	0.01356	0.18306	0
Sr	0	0	0.00053
Ti	0	0.00156	0.00021
V	0	0.01216	0
Zn	0.00167	0.00660	0.03150

After the sorting of source labels, the similarity between the PMF estimates ($\hat{\mathbf{G}}$ and $\hat{\mathbf{F}}$) and the actual **G** and **F** matrices was determined using average absolute error (AAE). AAE is a frequently-used metric in source apportionment that uses the absolute value of the error instead of the squared error [5,6]. The AAE for $\hat{\mathbf{F}}$ was calculated as

$$\mathrm{AAE} = \frac{1}{m}\sum_{j=1}^{m}\sum_{h=1}^{p}\left|\hat{f}_{hj} - f_{hj}\right| \quad (4)$$

where $m=23$ is the number of chemical species in the source profile, and $p=5$ is the number of profiles in **F**. AAE is thus simply the average of the absolute error, $|\hat{f}_{hj} - f_{hj}|$, for each source summed over the number of sources. The AAE of $\hat{\mathbf{G}}$ was calculated in a similar manner.

3. Optimal settings for PMF

PMF offers the user a variety of settings to influence the output of program. Uncertainties associated with the data may be incorporated with a variety of models, rotations may be introduced, limit values on iteration steps may be adjusted, and so forth. A study was performed on the simulated data to attempt to optimize these settings in the context of pollution source apportionment.

Please cite this article as: J.W. Lingwall, W.F. Christensen, Pollution source apportionment using *a priori* information and positive matrix factorization, Chemometrics and Intelligent Laboratory Systems (2007), doi:10.1016/j.chemolab.2007.03.007

Table 2
Hypothetical elements of an Fkey for secondary sulfate

Species	Profile	Fkey
Al	0.00495	0
As	0.00032	0
Br	0	8
Ca	0.03108	0
Cl	0.00735	0
Cu	0	8
EC	0.17706	0
Fe	0.01616	0
K	0.0076	0
Mn	0.00022	0
Na	0	13
Ni	0.01422	0
NO3	0.00102	0
OC	0.06379	0
P	0.00047	0
Pb	0.00163	0
Se	0	9
Si	0.01342	0
SO_4	0.63788	0
Sr	0	9
Ti	0.00057	0
V	0.01714	0
Zn	0.0051	0

Under the default settings, PMF performs almost as expected. The AAE for $\hat{\mathbf{G}}$ and $\hat{\mathbf{F}}$ increase as we increase the CV associated with measurement errors for the data. Interestingly, when we increase the CV to 1.0, there is little difference between the results for the contaminated data and the results for the uncontaminated data. That is, when the data are quite noisy it matters less to identify all the sources, since the high noise blurs out precise identifiability. We summarize our findings for some settings below, with details available in [7].

3.1. Outlier threshold distance

This setting controls what PMF processes as an outlier. For these data, the AAE was lower on average when the outlier threshold distance was lowered to 2.0 (down from the default value of 4.0).

3.2. Fpeak

This setting controls rotations in the data and is used by trial and error. Large positive values of Fpeak improved estimation in noisy data (CV = 1.0), while slight negative values (e.g. −0.5) improved estimation in very clean data (CV = 0.2, uncontaminated).

3.3. Errormodel

This setting controls how PMF handles the uncertainties associated with the observations, whether to update them as the program iterates, which formula to use for the error calculations, and so forth. For these simulations, setting the standard deviations for the data array equal to the uncertainties associated with the ambient data (i.e., Errormodel = −12 with no adaptation of the inputted uncertainties) yielded the lowest AAE values on average. However, this finding is likely due to

Table 3
Third Fkey, which represents some correct and some erroneous information about source profile zeros

	Sea salt		Secondary sulfate		Winter secondary		Soil		Auto-diesel	
	Profile	Fkey	Profile	Fkey	Profile	Fkey	Profile	Fkey	Profile	Fkey
Al	0	9	0.00495	0	0.03773	0	0.11493	0	0.00448	0
As	0	0	0.00032	0	0.00336	0	0	0	0	0
Br	0	0	0	0	0.00089	0	0	0	0.00006	0
Ca	0.009	0	0.03108	0	0.01546	0	0.00599	0	0.00750	0
Cl	0.265	0	0.00735	0	0.01921	0	0	0	0.00142	0
Cu	0	9	0	0	0.00658	0	0	0	0.00038	0
EC	0	0	0.17706	0	0	0	0.04891	0	0.36060	0
Fe	0	0	0.01616	0	0.04907	0	0.08181	0	0.00517	0
K	0.037	0	0.00760	0	0.00979	0	0.06708	0	0.00146	0
Mn	0	0	0.00022	0	0.00070	0	0.00125	9	0.00045	0
Na	0.689	0	0	0	0.05882	0	0.01565	0	0	0
Ni	0	0	0.01422	0	0.00605	0	0.00005	0	0	0
NO3	0	0	0.00102	0	0.28525	0	0	0	0.03768	0
OC	0	0	0.06379	0	0.02224	0	0.19562	0	0.52138	9
P	0	0	0.00047	0	0.00024	0	0	0	0.00069	0
Pb	0	0	0.00163	0	0.02884	0	0	0	0.00057	0
Se	0	0	0	9	0.00028	0	0	0	0	0
Si	0	0	0.01342	9	0.15755	0	0.45046	0	0.02284	0
SO_4	0	0	0.63788	0	0.27670	0	0	0	0.03368	0
Sr	0	0	0	0	0.00083	0	0.00186	0	0.00006	0
Ti	0	0	0.00057	0	0.00608	9	0.01622	0	0.00026	0
V	0	9	0.01714	0	0.00758	0	0.00017	0	0	9
Zn	0	0	0.00510	9	0.00676	0	0	0	0.00131	0

For each source an *a priori* value for the profile is given along with the corresponding Fkey value. Increasing Fkey values indicate increasing levels of certainty that the corresponding profile element is equal to zero.

Please cite this article as: J.W. Lingwall, W.F. Christensen, Pollution source apportionment using *a priori* information and positive matrix factorization, Chemometrics and Intelligent Laboratory Systems (2007), doi:10.1016/j.chemolab.2007.03.007

our use of simulated data with known uncertainties. We expect that in cases where uncertainty estimates are of lower quality, other errormodel settings may perform better.

Because the evaluation of the simulation experiment yielded no settings which exhibited dramatically improved estimation properties, we confirm the appropriateness of the PMF default settings. Our recommendation is to simply use the default settings, adjusting them as necessary for convergence or as adjustment yields clearer results. In the study of *a priori* information in PMF, we use PMF settings that are very similar to the default values: outlier threshold distance = 4.0, Fpeak = 0, and Errormodel = −12 (with C1 = 0, C2 = 0, and C3 = 0.01).

4. Pulling source profile elements to zero ("F element pulling")

The first method we examine for incorporating *a priori* information into a PMF analysis is that of "pulling" elements of the source profiles to zero. This method uses a matrix that indicates the location of suspected zeros in source profiles or contributions. Since here we are concerned with the profiles, this information is given in the form of integer values in an Fkey. The more sure the researcher is that a certain element of a source profile is zero, the larger the integer value that is specified. Because we are dealing with simulated data, the locations of zero elements in the profile matrix are known. However, we consider here the behavior of F element pulling when the researcher does not have certain information about the location of zero elements. For example, suppose that for the secondary sulfate profile, the researcher believes that the elements corresponding to Na, Br, Cu, Se, and Br are zero, but is more certain about the zero associated with Na and less certain about the zero associated with Br, Cu, Se, and Sr. For such a scenario, Table 2 illustrates an Fkey which appropriately represents the hypothetical researcher's *a priori* certainty about profile elements believed to be zero.

Four different types of Fkeys were used to examine the performance of this technique. First, 5 elements of the sea salt factor that are actually zero were pulled to zero with a "medium strong" pull (integer value of 9). This represents a case where a researcher has knowledge about one of the factors, but not the others. Second, 10 elements chosen from the actual zeros scattered throughout **F** were pulled to zero with a value of 9, representing a case where the researcher has some knowledge about most of the sources. Third, 10 elements were again pulled to zero, this time with five pulled correctly to zero and five mistakenly pulled to zero, including one major element in the auto-diesel factor, as shown in Table 3. The fourth type of Fkey pulled 10 elements correctly to zero, setting the integer values to 5, 6,..., 14, so that 10 Fkeys were created. By varying the integer value from 5 to 14 we can observe the effect of F element pulling as we change the degree of certainty.

The AAE calculated using the first three Fkeys is shown in Figs. 2 ($\hat{\mathbf{F}}$) and 3 ($\hat{\mathbf{G}}$). For each Fkey scenario and each data generation scenario, a density plot is given for the AAE values obtained from the 100 simulated data sets, where AAE is defined in Eq. (4). AAE values are plotted on the horizontal axis with the observed density of the AAE values plotted on the vertical axis. The density plot is analogous to a smoothed histogram with higher density indicating AAE values that are more likely. The vertical axes for the plots within each figure differ from one another in order to accentuate differences within each plot, but the area under each density curve is equal to 1. Within each figure, the horizontal axis is the same for each plot so that differences in performance between data generation scenarios can be more easily detected.

For each of the Fkey scenarios, as the degree of error in the data (denoted by the CV) increases, the AAE increases for $\hat{\mathbf{F}}$ and $\hat{\mathbf{G}}$. If the data are contaminated by unidentified minor sources, the AAE values also increase. When we correctly identify 5 zeros in the sea salt factor we improve the overall estimation slightly in the uncontaminated data (the thin black line in the figures). When we correctly specify 10 profile zeros, we again improve estimation in the uncontaminated data (the dashed line in the figures). When we correctly specify five profile zeros but incorrectly specify five others, as seen in the dotted line in Figs. 2 and 3, we can dramatically worsen estimation in the low CV/uncontaminated data scenarios. However, when the data are contaminated by unidentified sources or are subject to a high degree of error (CV = 1.0), none of the Fkeys substantially increase or decrease the AAE of the PMF estimates. The features of the PMF estimates for the remaining scenarios (uncontaminated data with CV ≤ 0.5) can be noted in Figs. 2 and 3. The density plots associated with correct zero element specification (thin black and dashed lines) are centered at lower AAE values than the density plot associated with no *F* element pulling (thick black line), and the density plots associated with partially incorrect pulling (dotted line) are centered at higher AAE values. Thus, correctly pulling zero elements to zero can improve estimation, but misspecified zeros can substantially increase the AAE when data have CV values that are less than or equal to 0.5.

Finally, the results from the fourth Fkey, where we correctly pull ten elements of the source profiles to zero with varying degrees of strength, are shown in Table 4. For the cleanest data (CV ≤ 0.5, uncontaminated), the use of F element pulling improves estimation almost as we would expect, with stronger degrees of pulling strength (corresponding to higher levels of certainty) generally leading to better estimation when CV ≤ 0.5. When the data are contaminated by unidentified pollution sources or have CV = 1.0, there is little evidence of improvement in AAE when increasing the pulling strength on the actual zero locations.

In general, the use of F element pulling on "clean data" (low CV and no contamination from unidentified sources) reduces the AAE. For noisier data, even correct specification of zeros yields no positive effect on the AAE. In the simulation, misspecification of zero locations often resulted in much higher AAE values. On the basis of this simulation, the researcher should carefully avoid misspecifying zeros and use appropriate levels of certainty or "pulling strength." Although the process of F element pulling is somewhat complicated (requiring a

Please cite this article as: J.W. Lingwall, W.F. Christensen, Pollution source apportionment using *a priori* information and positive matrix factorization, Chemometrics and Intelligent Laboratory Systems (2007), doi:10.1016/j.chemolab.2007.03.007

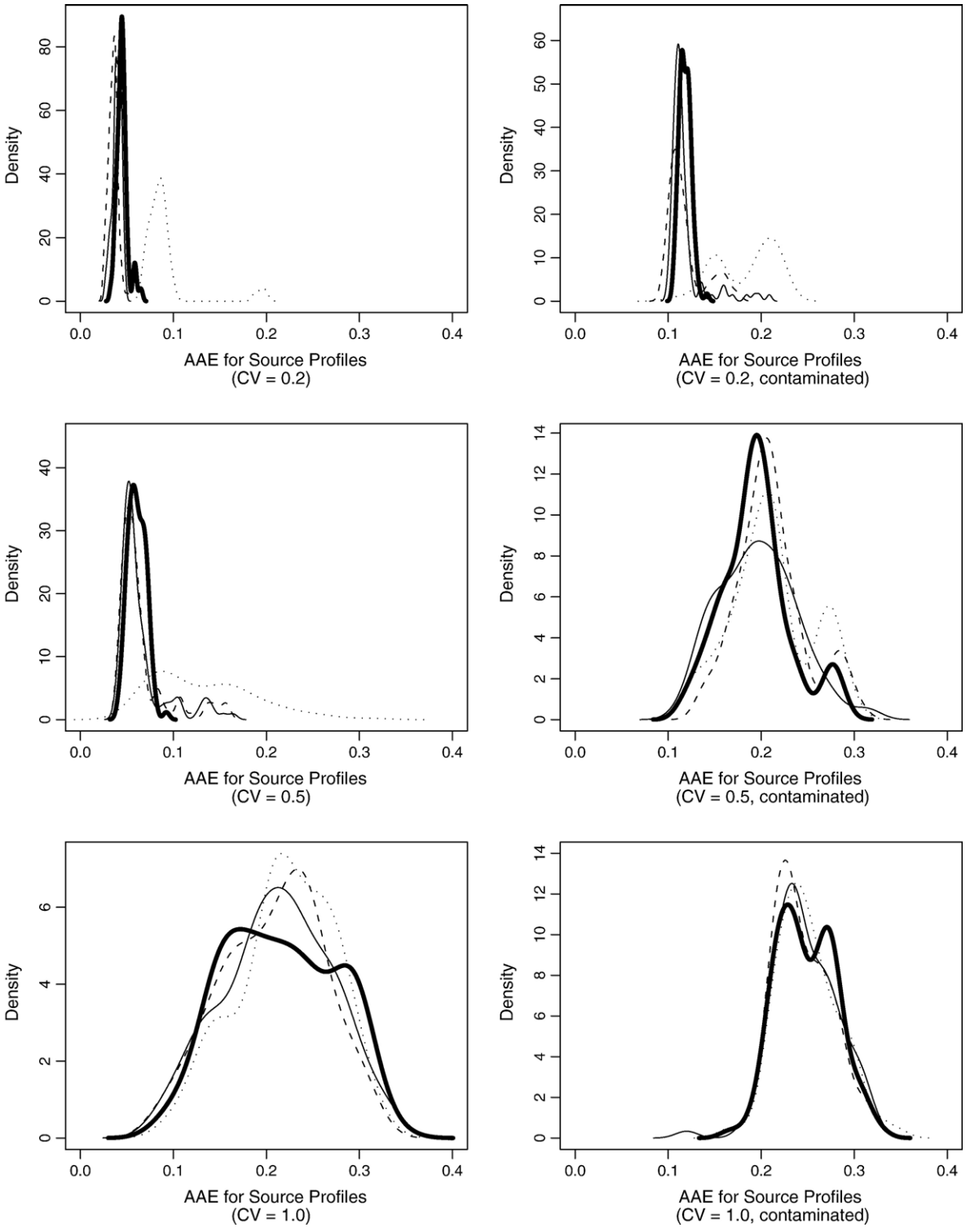

Fig. 2. Density estimates of the AAE for $\hat{\mathbf{F}}$, using three different Fkeys. The thick black line represents results for PMF's default settings, the thin black line represents correctly pulling 5 elements of one factor to zero, the dashed line represents pulling 10 elements of multiple factors correctly to zero, and the dotted line represents a mixture of correct and incorrect pulling.

modification of the PMF initialization script), we note that plans for future releases of EPA PMF [3] include an enhancement of the graphical user interface that will allow F element pulling to be specified.

5. Source profile targeting

The second method available for introducing prior information into a PMF analysis is source profile targeting. For a local

Please cite this article as: J.W. Lingwall, W.F. Christensen, Pollution source apportionment using *a priori* information and positive matrix factorization, Chemometrics and Intelligent Laboratory Systems (2007), doi:10.1016/j.chemolab.2007.03.007

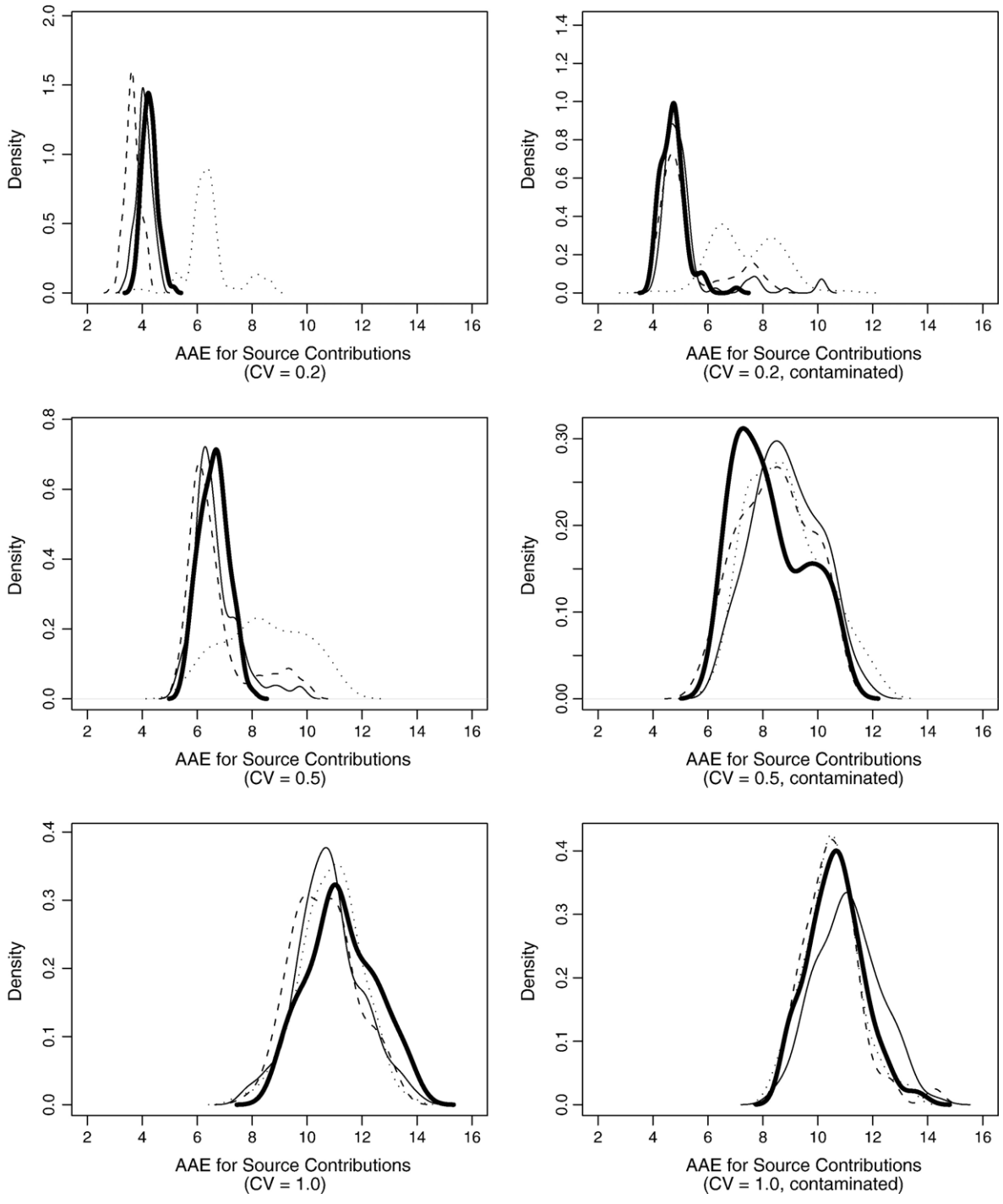

Fig. 3. Density estimates of the AAE for $\hat{\mathbf{G}}$, using three different Fkeys. The thick black line represents results for PMF's default settings, the thin black line represents correctly pulling 5 elements of one factor to zero, the dashed line represents pulling 10 elements of multiple factors correctly to zero, and the dotted line represents a mixture of correct and incorrect pulling.

pollution source of interest, target source profiles are often available from past studies, laboratory measurements, or source profile inventories such as the U.S. EPA's Speciate database. In many cases, available source profiles are accompanied by an estimate of the uncertainty associated with each profile element. Thus, incorporation of profile uncertainty in source profile targeting is often very natural and potentially free of subjectivity. PMF's method for source profile targeting introduces *a priori* information about one or more of the profiles in the form of a matrix of target profiles, $\tilde{\mathbf{F}}$. Each element of $\tilde{\mathbf{F}}$ is given a specified amount of uncertainty. See [8] for implementation details.

As in the study of F element pulling, once PMF had computed $\hat{\mathbf{F}}$ and $\hat{\mathbf{G}}$, the columns of $\hat{\mathbf{G}}$ were sorted to minimize the MSE between $\hat{\mathbf{G}}$ and $\mathbf{G}$. In the simulations, if PMF failed to converge using the target profiles, the results from runs using the default values of PMF (without source profile targets) were

Please cite this article as: J.W. Lingwall, W.F. Christensen, Pollution source apportionment using *a priori* information and positive matrix factorization, Chemometrics and Intelligent Laboratory Systems (2007), doi:10.1016/j.chemolab.2007.03.007

Table 4
Means of the AAE results for the fourth Fkey, pulling 10 elements correctly to zero with differing amounts of strength

	CV=0.2		CV=0.5		CV=1.0	
	Default	Contaminated	Default	Contaminated	Default	Contaminated
Source contribution AAE						
No Fkey	4.266	4.730	6.600	8.263	11.261	10.648
5	3.835	4.475	6.685	7.797	9.868	10.089
6	3.730	4.480	6.624	7.595	9.810	10.505
7	3.633	4.812	6.536	7.588	9.945	10.841
8	3.675	4.734	6.354	7.617	9.671	10.852
9	3.631	4.795	6.265	7.687	9.582	10.840
10	3.586	4.848	6.312	7.649	9.631	10.890
11	3.588	4.864	6.218	7.779	9.508	10.717
12	3.595	4.866	6.256	7.835	9.650	10.562
13	3.652	4.885	6.291	7.727	9.672	10.516
14	3.789	4.848	6.340	8.052	9.742	10.544
Source profile AAE						
No Fkey	0.045	0.118	0.061	0.194	0.215	0.248
5	0.037	0.118	0.063	0.190	0.213	0.255
6	0.035	0.108	0.067	0.188	0.216	0.253
7	0.034	0.115	0.069	0.181	0.215	0.255
8	0.033	0.116	0.063	0.182	0.209	0.252
9	0.032	0.116	0.064	0.177	0.206	0.257
10	0.031	0.111	0.056	0.178	0.211	0.258
11	0.032	0.109	0.050	0.188	0.208	0.255
12	0.030	0.108	0.051	0.192	0.211	0.253
13	0.031	0.107	0.051	0.186	0.209	0.251
14	0.033	0.107	0.052	0.198	0.209	0.251

Results are shown for the uncontaminated (default) data setting and the contaminated data setting.

substituted. This corrects for the possibility that PMF is not converging on high-noise data sets, although in practice the settings might be changed for a particular data set so that PMF converges while using the target profiles.

To use the target profiles in PMF, the following steps are used (as outlined in [8]). First, the target profiles, $\tilde{\mathbf{F}}$, are appended to the beginning of the data matrix $\mathbf{X}$, so that now $\mathbf{X}$ is $(p+n)\times m$ instead of $n\times m$. The $p\times m$ matrix containing the uncertainties associated with $\tilde{\mathbf{F}}$ is appended to the beginning of the matrix of receptor measurement uncertainties. As a result of these steps, our resulting $\hat{\mathbf{G}}$ will be $(p+n)\times p$ instead of $n\times p$. A Gkey is then constructed to pull these extra $p\times p$ elements of $\hat{\mathbf{G}}$ to zero. Starting values obtained from running PMF without the use of target profiles are used for the algorithm, after appending a $p\times p$ matrix of zeros to the starting values of $\hat{\mathbf{G}}$ to correspond to the extra $p\times p$ elements.

Since the use of starting values complicates the process, it is desirable to see if their use actually improves the results. To test this, PMF was run on over the 600 simulated data sets, using target profile information, with and without starting values. As noted in [8], the use of starting values generally improved PMF's estimates or at least had no negative impact. For this study, the true source profiles were used as $\tilde{\mathbf{F}}$, with the uncertainties associated with them obtained by multiplying the true source profiles by 0.2. The multiplier of 0.2 was chosen because assumed profile element uncertainties in source profile databases (such as the U.S. EPA's Speciate database) usually take on values of 10% to 20% of the profile element. As noted below, our simulations also consider some cases in which the assumed 20% error value is understating the actual uncertainty in the profile.

5.1. Differing amounts of information in $\tilde{F}$

We next consider how the use of target source profiles improves estimation when we have varying amounts of *a priori* information. This section examines the issue by using approximate profiles with varying degrees of uncertainty. For purposes of the simulation study, matrices of estimated factor profiles were used as $\tilde{\mathbf{F}}$, obtained in the following manner. Let f_{hj} be the proportion contribution of the jth species to the hth profile, and let $\tilde{f}_{hj}$ be its assumed value. We obtain $\tilde{f}_{hj}$ as a draw from a lognormal distribution with a mean equal to f_{hj} and a coefficient of variation equal to either 1.0, 0.5, or 0.25. The array of uncertainties associated with $\tilde{\mathbf{F}}$ was obtained by multiplying $\tilde{\mathbf{F}}$ by 0.2. Choosing the multiplier of 0.2 allows us to evaluate the performance of source profile targeting when the assumed profile uncertainties are dramatically understated, somewhat understated, or approximately correct. The degree of our knowledge about the profile matrix ($\mathbf{F}$) is reflected by the level of the CV used to obtain $\tilde{\mathbf{F}}$. Thus, a CV of 0.25 reflects greater information than the CV of 1.0. Three hundred different $\tilde{\mathbf{F}}$'s were thus created, one hundred at each CV. Each value for $\tilde{\mathbf{F}}$ was used to estimate the profiles in both the contaminated and uncontaminated data setting.

Analogous to Figs. 2 and 3 discussed in Section 4, the AAE for $\hat{\mathbf{F}}$ and $\hat{\mathbf{G}}$ are shown in Figs. 4 and 5, respectively. Density estimates for the AAE associated with $\tilde{\mathbf{F}}_{CV=0.25}$, $\tilde{\mathbf{F}}_{CV=0.5}$, and

Please cite this article as: J.W. Lingwall, W.F. Christensen, Pollution source apportionment using *a priori* information and positive matrix factorization, Chemometrics and Intelligent Laboratory Systems (2007), doi:10.1016/j.chemolab.2007.03.007

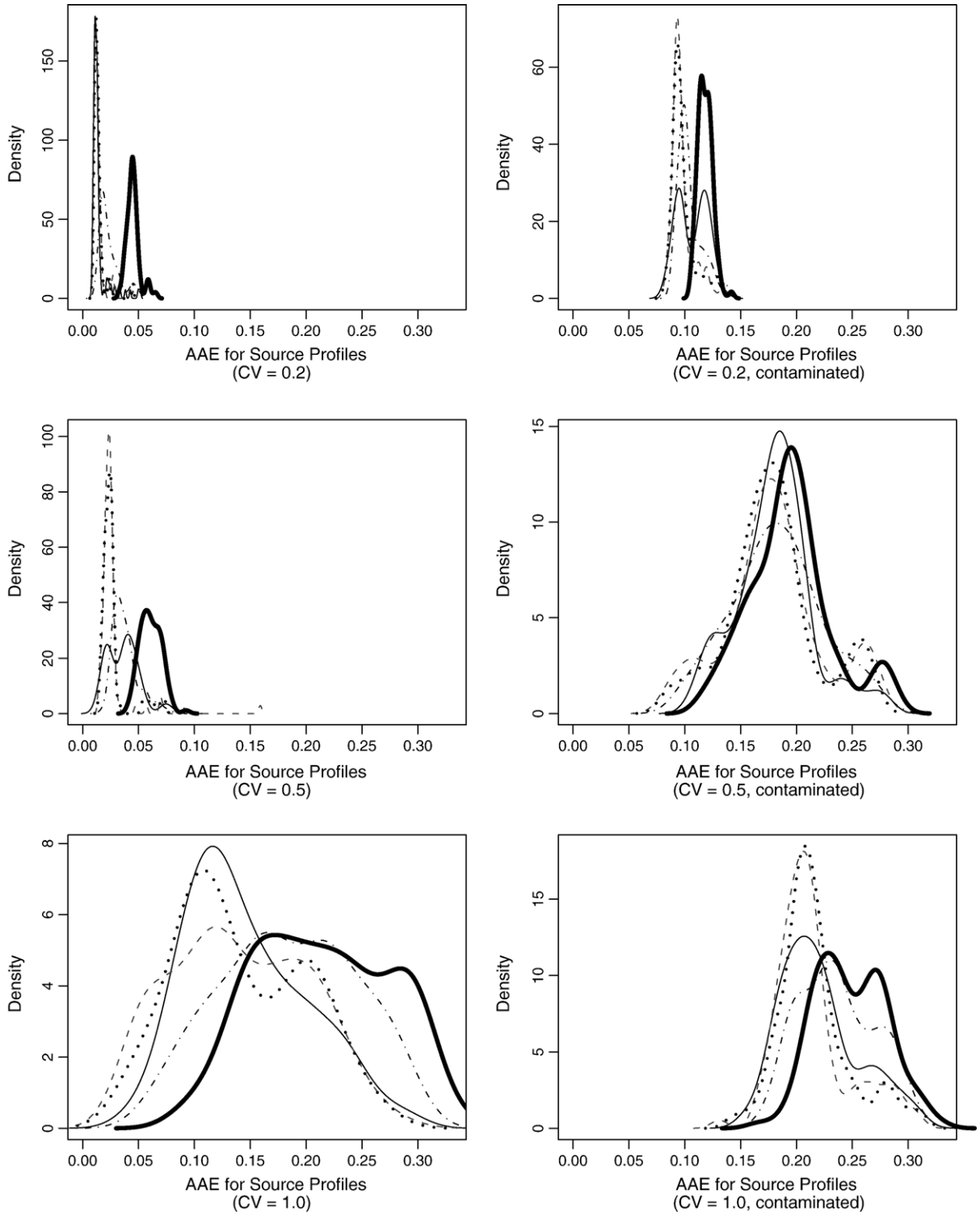

Fig. 4. Density estimates of the AAE for $\hat{\mathbf{F}}$, using target profiles in PMF. Thick black line represents results for PMF without using *a priori* information. Thin black line represents results for PMF when using the correct source profiles as targets. Gray lines represent results obtained when using target profiles that are affected by increasing degrees of error: CV = 0.25 (dash), CV = 0.5 (dot), and CV = 1.0 (dot–dash).

$\tilde{\mathbf{F}}_{CV=1.0}$ are shown using the dashed gray line, dotted gray line, and dot–dash gray line, respectively. The AAE values for the default PMF estimates (obtained without using *a priori* information) are shown by the thick black line, and the thin black line shows the density estimates obtained by using the correct profiles (**F**) as targets. As can be noted in Figs. 2 and 3, the use of *a priori* information consistently improved estimation for both **F** and **G**. That is, in every plot in Figs. 2 and 3, the AAE values for the source-profile-targeting estimates are smaller than the default PMF estimates. This improvement due to source

Please cite this article as: J.W. Lingwall, W.F. Christensen, Pollution source apportionment using *a priori* information and positive matrix factorization, Chemometrics and Intelligent Laboratory Systems (2007), doi:10.1016/j.chemolab.2007.03.007

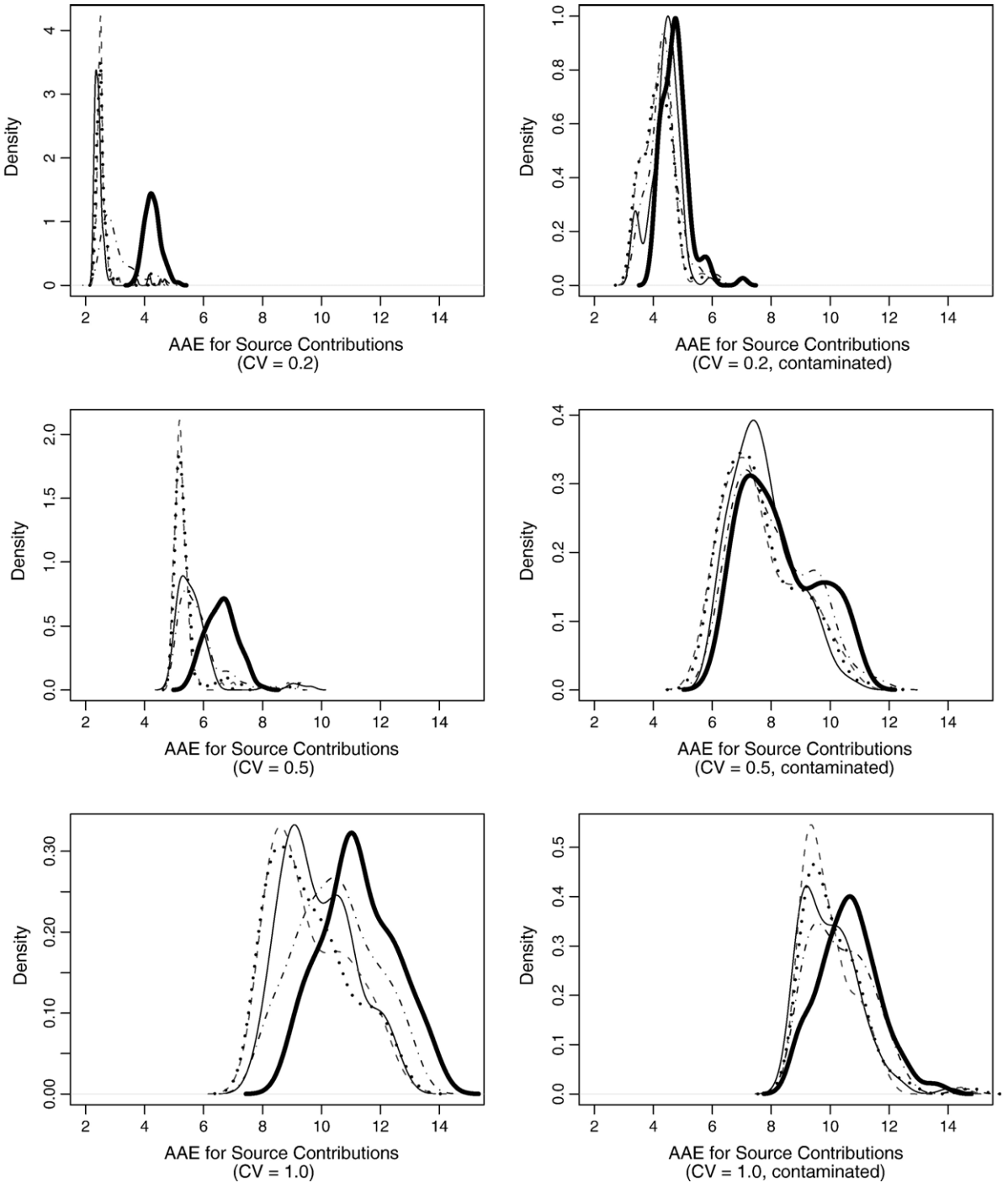

Fig. 5. Density estimates of the AAE for $\hat{\mathbf{G}}$, using target profiles in PMF. Thick black line represents results for PMF without using *a priori* information. Thin black line represents results for PMF when using the correct source profiles as targets. Gray lines represent results obtained when using target profiles that are affected by increasing degrees of error: CV = 0.25 (dash), CV = 0.5 (dot), and CV = 1.0 (dot–dash).

profile targeting even holds true for the contaminated and low data quality (CV = 1.0) scenarios, although the improvement is less pronounced.

It appears that source profiles need not be exact for source profile targeting to improve estimation. In fact, the AAE values obtained using the approximate profiles (gray lines) are roughly the same as the AAE values obtained using the correct profiles (thin black line). This indicates that these data have sufficient structure to extract accurate parameter estimates even with *a priori* information that is less than exact. In the simulation, we are able to use the true profiles and contributions to assess the effect of error in the *a priori* source profiles when estimating model parameters. However, in practice, one can utilize model fit diagnostics such as Q from Eq. (3) to avoid using erroneous target profiles which distort the profiles and contributions of the actual sources. In cases where a *priori* information is detrimental, Q can

Please cite this article as: J.W. Lingwall, W.F. Christensen, Pollution source apportionment using *a priori* information and positive matrix factorization, Chemometrics and Intelligent Laboratory Systems (2007), doi:10.1016/j.chemolab.2007.03.007

be expected to be dramatically larger than when PMF is used with no *a priori* profile information. (See [8], p. 17.)

6. Illustration of source profile targeting with St. Louis Supersite data

To illustrate the use of source profile targeting in pollution source apportionment, we consider a positive matrix factorization of daily $PM_{2.5}$ measurements from the St. Louis Supersite. The species used here include metals, organic and elemental carbon (OC and EC), sulfate (SO_4), and nitrate (NO_3). A total of 661 complete measurements were utilized, obtained during the period of May 2001 to May 2003. Details about measurement methods as well as a more thorough PMF analysis of a similar data set can be found in [9]. Although source profile targeting is primarily a tool for utilizing well-established knowledge about source profiles, we demonstrate an alternative use of this approach—model building in an exploratory setting. Specifically, we begin with multiple exploratory runs of PMF and identify factors that are important for purposes of model interpretability or simplicity. We then use source profile targeting to incrementally improve upon PMF solutions by retaining key factors from previous runs.

We begin by considering several k-factor solutions ($k=3,\ldots, 13$) without the use of any *a priori* information. For each solution, the profiles and contribution plots were inspected in order to identify likely interpretations for each outputted source. As can be seen in Table 5, the nature of the PMF solution depends heavily upon the number of sources used in the model. The collection of identified sources changes substantially as sources are added to the model, with some sources disappearing and then reappearing as the number of sources in the model is reduced from 13 to 3. No source remains the same as the value for k is reduced, with most source contribution estimates increasing as the mass from previously removed sources is partitioned among the remaining sources.

Unlike the analysis of [9], this analysis does not include fractionization of OC and EC which is reported to assist in separating gasoline-based and diesel-based mobile emissions. Another important difference relates to the use of fireworks event days (July 4 and 5) which were excluded in the analysis of [9] but were included in the current analysis to illustrate the utility of source profile targeting in the identification of small but distinct sources.

After examining Table 5, we concluded that these data were able to sufficiently resolve 9 sources in an interpretable manner: lead smelter, fireworks, copper smelter, zinc smelter, summer secondary, soil dust, steel manufacturing, mobile (combined gasoline/diesel), and winter secondary. The estimated profile associated with a given source changes as k is varied, but from among all the profiles associated with a given source, we selected a profile that best captured the known nature of each source. For example, several of the sources that are difficult to identify in the 13-source analysis are heavy in sulfate. As the number of sources (k) is reduced and some of these sulfate-heavy sources disappear from the collection of resolved sources, the amount of SO_4 in many of the sources increases. Thus, for the lead, copper, and zinc smelters, we chose profiles obtained from high k solutions ($k=11$, 12, and 13, respectively) in order to minimize this artificial leaking of the sulfate into likely unrelated sources. For similar reasons, a profile from a high k solution was desired for the fireworks source. The profile for the fireworks source was taken from the 10-source solution, because the 12- and 13-source solutions split fireworks into two sources and the profile from the 11-source solution was contaminated by NO_3, yielding small but unrealistic fireworks emissions during the winter months. But for other sources such as summer secondary and mobile, we selected profiles from the $k=6$ solutions because we surmised that some of the sources that disappeared when moving from high k to low k analyses (such as "S/SO_4" and "Mobile 2") are in fact subsets of the broadly defined summer secondary and mobile sources. The profiles from the soil, steel manufacturing, and winter secondary sources were taken from the 10-source solution because of the instability of these profiles for $k<10$.

In combination with the assumed profile matrix, a matrix of profile uncertainty values must be specified which is appended to the top of the matrix of receptor measurement uncertainties when running PMF. Because the fireworks source is the most likely to be lost from the 9-source solution when using these 9 profiles, we set the uncertainties for the fireworks column of the

Table 5
Source contributions (in $\mu g/m^3$) for k-source PMF solutions

# of sources	Explained mass	Lead smelter	Copper smelter	Fire-works	Zinc smelter	S/SO_4 source	Summer secondary	Soil dust	Steel mill	Mobile	Mobile 2	OC/SO_4 source	Winter secondary
3	11.63				0.78		7.44			3.41			
4	11.68				0.58		6.61		1.42	3.07			
5	13.20				0.37		6.20		1.13	2.56			2.94
6	13.21		0.25		0.26		6.40		0.96	2.46			2.88
7	13.36		0.26		0.30		6.11		0.81	1.98	1.04		2.86
8	13.52		0.20		0.32		6.50	0.47	0.49	1.83	1.23		2.48
9	13.45	0.28	0.22	0.27	0.23		6.21		0.78	1.96		0.49	3.01
10	13.50	0.20	0.24	0.25	0.26		6.31	0.40	0.75	1.83		0.51	2.75
11	13.49	0.09	0.27	0.20	0.31		6.16	0.41	1.00	1.15	1.32	0.05	2.53
12	13.83	0.09	0.13	0.35[a]	0.27		6.21	0.38	0.49	2.80	0.79	0.12	2.20
13	13.83	0.08	0.15	0.31[a]	0.23	0.36	6.11	0.38	0.51	2.66	0.61	0.20	2.23

[a] For the 12- and 13-source solutions, two fireworks sources were resolved and their mean contributions are summed here.

Please cite this article as: J.W. Lingwall, W.F. Christensen, Pollution source apportionment using *a priori* information and positive matrix factorization, Chemometrics and Intelligent Laboratory Systems (2007), doi:10.1016/j.chemolab.2007.03.007

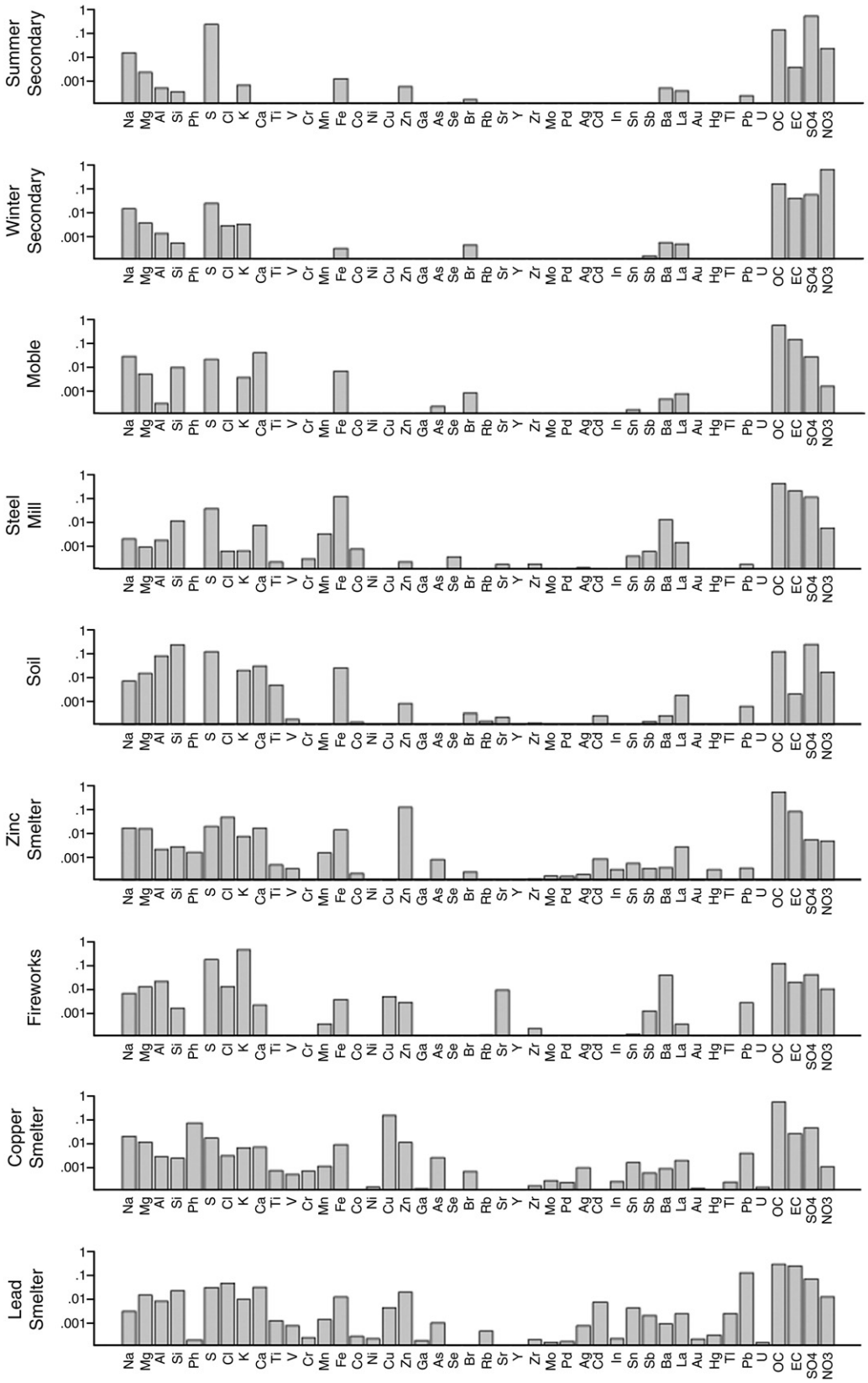

Fig. 6. Source profile estimates for 9-source model using source profile targeting. The various chemical species are plotted on the horizontal axis with the profile concentration on the vertical axis.

Please cite this article as: J.W. Lingwall, W.F. Christensen, Pollution source apportionment using *a priori* information and positive matrix factorization, Chemometrics and Intelligent Laboratory Systems (2007), doi:10.1016/j.chemolab.2007.03.007

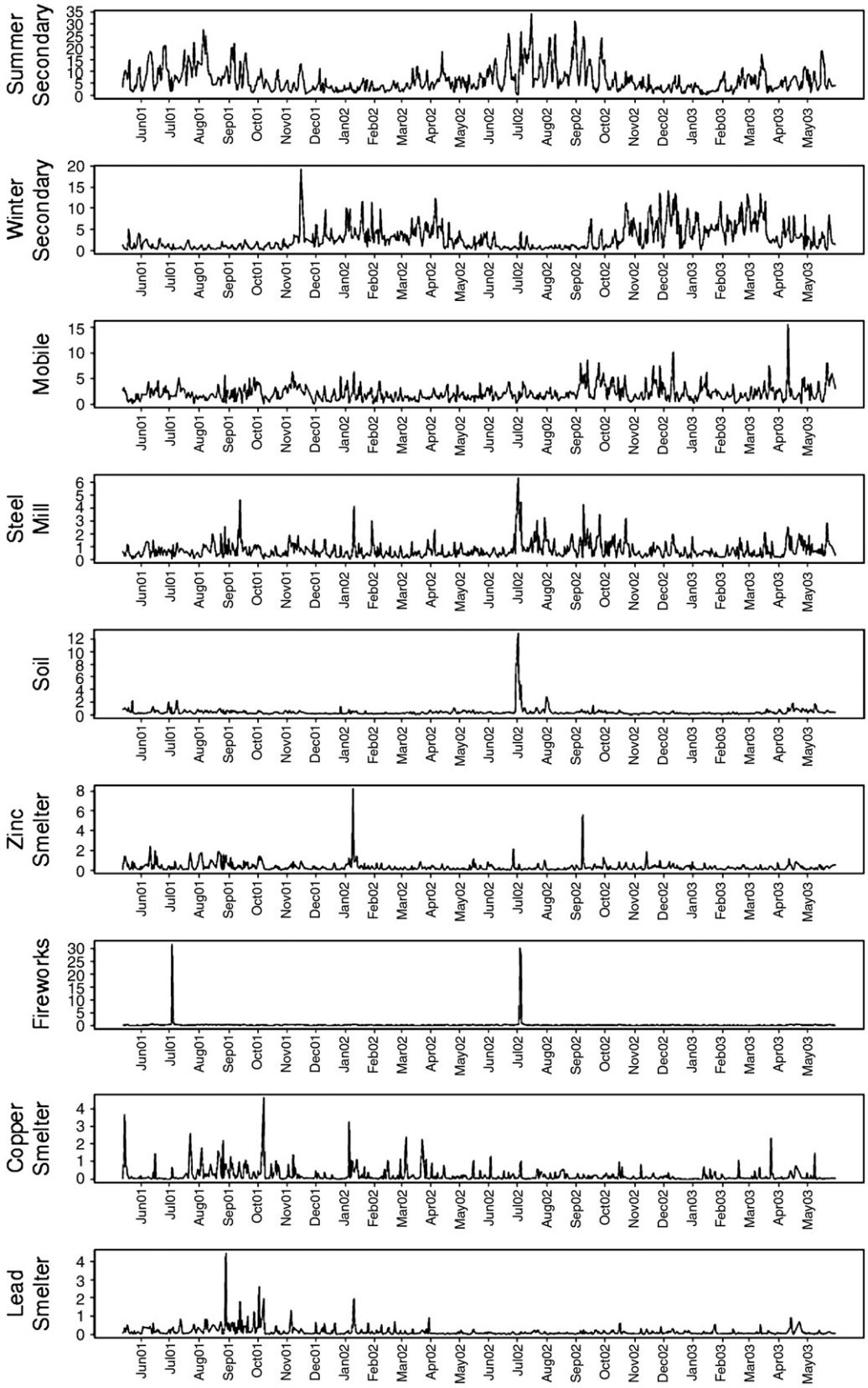

Fig. 7. Source contribution estimates in μg/m^3 for the 9-source model using source profile targeting. Estimated source contributions are plotted on the vertical axis are plotted over time (horizontal axis).

Please cite this article as: J.W. Lingwall, W.F. Christensen, Pollution source apportionment using *a priori* information and positive matrix factorization, Chemometrics and Intelligent Laboratory Systems (2007), doi:10.1016/j.chemolab.2007.03.007

profile matrix equal to 0.001. The three smelter sources and the steel manufacturing source (also relatively minor contributors to the total $PM_{2.5}$ mass) have uncertainties set to 0.01. As with fireworks, these are set to a low level so that they are not "lost" when fitting the 9-source model. Because we suspect that at least two of the columns in Table 5 relate to mobile source emissions, we set the uncertainties associated with the mobile source profile to the relatively high value of 10. This allows for additional flexibility in estimating the mobile source profile. The uncertainties associated with all other source profile elements were set to the value of 1. Although the specific profile uncertainty values (10, 1, 0.01, 0.001) were chosen in somewhat of a trial-and-error fashion to ensure retention of smaller factors such as fireworks, the relative values of the uncertainties were chosen to reflect our relative willingness to let PMF alter the nature of the previously derived source.

PMF was re-run with this *a priori* information used in the source profile targeting framework. The estimates of the profiles and contributions for this analysis are illustrated in Figs. 6 and 7. Fig. 7 exhibits a few evidences of potential factor-separation problems. For example, the coincident peaks of the steel mill and soil sources (in early July 2002) and the coincident peaks of the zinc and copper smelters (in early January 2002). Notwithstanding, the estimated contributions generally meet expectations related to rough magnitudes and adherence to well-understood seasonal patterns. The mean daily contribution and percentage of explained $PM_{2.5}$ mass for each source are as follows (in order of importance): summer secondary (6.47 $\mu g/m^3$ or 47.8%), winter secondary (2.98 $\mu g/m^3$ or 22.0%), mobile (2.1 $\mu g/m^3$ or 15.5%), steel mill (0.72 $\mu g/m^3$ or 5.3%), soil (0.38 $\mu g/m^3$ or 2.8%), zinc smelter (0.31 $\mu g/m^3$ or 2.3%), fireworks (0.23 $\mu g/m^3$ or 1.7%), copper smelter (0.20 $\mu g/m^3$ or 1.5%), and lead smelter (0.14 $\mu g/m^3$ or 1.0%). Thus, not only is source profile targeting a useful tool when profiles are known *a priori*, but it is also beneficial because it assists the researcher in formulating an interpretable solution in an exploratory framework.

7. Conclusions

In this manuscript, we have considered the empirical properties of positive matrix factorization (PMF) in order to gain insight about its optimal use in practice. Although no simulated data set is able to fully capture the complexities associated with real-world data analysis, studies based on simulation (such as this one) can be useful in providing insights about the relative advantages and hazards of different analysis methods and approaches. Few of the basic PMF run control settings had a dramatic effect on estimation performance, but there were several lessons learned from the investigation of F element pulling and source profile targeting. For "clean" ambient data (i.e., data that have a low degree of measurement error and are not affected by unidentified/unmodeled sources), correctly specifying source profile zeros using F element pulling can improve estimation of both source profiles and source contributions. However, the incorrect specification of zero elements can worsen estimation. Specifying a high degree of certainty about the correct location of zeros in the profile matrix seems to improve estimation for clean data, but has little effect when data exhibit a high degree of noise. The use of target source profiles can dramatically improve estimation, even with less than exact *a priori* profile information. The use of target source profiles may also be used in an incremental model fitting process to obtain parsimonious and interpretable models of the airshed. Such an approach was used to obtain an interpretable source apportionment of $PM_{2.5}$ constituents at the St. Louis—Midwest Supersite.

Acknowledgments

This work was supported by the STAR Research Assistance Agreement No. RD-83216001-0 awarded by the U.S. Environmental Protection Agency. 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.

The authors thank Dr. Jay Turner and Dr. James Schauer for assistance in accessing data from the St. Louis Supersite. We are also grateful for the helpful comments of the editor and anonymous reviewers.

References

[1] P. Paatero, U. Tapper, Environmetrics 5 (1994) 111–126.
[2] P. Paatero, Chemometrics and Intelligent Laboratory Systems 37 (1997) 23–35.
[3] S. Eberly, EPA PMF 1.1 Users Guide, U.S. Environmental Protection Agency, Research Triangle Park, NC, 2005.
[4] W.F. Christensen, J.J. Schauer, J.W. Lingwall, Environmetrics 17 (2006) 663–681.
[5] H.S. Javitz, N.F. Robinson, J.G. Watson, Atmospheric Environment 22 (1988) 2309–2322.
[6] W.F. Christensen, R.F. Gunst, Atmospheric Environment 38 (2004) 733–744.
[7] J.W. Lingwall, Bayesian and Positive Matrix Factorization approaches to pollution source apportionment. M.S. Thesis, Department of Statistics, Brigham Young University, Provo, UT, 2006.
[8] P. Paatero, Users Guide for Positive Matrix Factorization programs PMF2 and PMF3, Part 1: Tutorial, University of Helsinki, Helsinki, Finland, 2004.
[9] J.H. Lee, P.K. Hopke, J.R. Turner, Journal of Geophysical Research 111 (2006) D10S10, doi:10.1029/2005JD006329.

Please cite this article as: J.W. Lingwall, W.F. Christensen, Pollution source apportionment using *a priori* information and positive matrix factorization, Chemometrics and Intelligent Laboratory Systems (2007), doi:10.1016/j.chemolab.2007.03.007