Atmos. Chem. Phys. Discuss., 8, S2132–S2137, 2008
www.atmos-chem-phys-discuss.net/8/S2132/2008/

Atmospheric Chemistry and Physics Discussions

Interactive comment on "Assessing positive matrix factorization model fit: a new method toestimate uncertainty and bias in factor contributions at the daily timescale" *by* J. G. Hemann et al.

J. G. Hemann et al.

Received and published: 29 April 2008

The authors would like to thank the Referee for his/her thoughtful and detailed comments. Regarding the issues raised,

1)
The authors agree that the factor contribution estimates, at least for some factors, are not encouraging. Regarding correlations between input factor profile and correlations between input factor contributions, the following tables are given.

Full Screen / Esc
Printer-friendly Version
Interactive Discussion
Discussion Paper

Factor Profile Correlations (profiles under normalization in Eq. 6)

	A.Sulfate	A.Nitrate	Gasoline	Diesel	Road Dust	Wood	Meat	Nat Gas	Vegetative
A.Sulfate	1	0.20	-0.17	-0.15	-0.14	-0.11	-0.09	-0.05	-0.12
A.Nitrate	-	1	-0.17	-0.14	-0.15	-0.10	-0.08	-0.09	-0.11
Gasoline	-	-	1	0.22	-0.09	-0.30	-0.14	-0.15	-0.30
Diesel	-	-	-	1	0.02	-0.26	-0.15	0.00	-0.13
Road Dust	-	-	-	-	1	-0.30	0.08	-0.24	0.05
Wood	-	-	-	-	-	1	-0.08	-0.20	-0.24
Meat	-	-	-	-	-	-	1	-0.09	-0.12
Nat Gas	-	-	-	-	-	-	-	1	-0.18
Vegetative	-	-	-	-	-	-	-	-	1

Factor Contribution Robust Correlations, Lag = 0

	A.Sulfate	A.Nitrate	Gasoline	Diesel	Road Dust	Wood	Meat	Nat Gas	Vegetative
A.Sulfate	1	0.55	0.89	0.57	0.73	0.34	0.81	0.68	0.49
A.Nitrate	-	1	0.39	0.26	0.32	0.82	0.35	0.80	-.015
Gasoline	-	-	1	0.62	0.74	0.13	0.81	0.64	0.54
Diesel	-	-	-	1	0.61	0.08	0.28	0.33	0.31
Road Dust	-	-	-	-	1	0.13	0.57	0.50	0.37
Wood	-	-	-	-	-	1	0.14	0.77	-0.34
Meat	-	-	-	-	-	-	1	0.64	0.53
Nat Gas	-	-	-	-	-	-	-	1	0.08
Vegetative	-	-	-	-	-	-	-	-	1

Given these correlation structures, the authors feel that the synthetic data are generally realistic. Specifically, the authors would expect small inter-factor profile correlations because the compounds chosen for "measurement" were chosen to yield distinguishable profiles. At the same time, the synthetic data are realistic in showing significant cross-correlation between at least some of the factor contributions time series. The authors

ACPD
8, S2132–S2137, 2008

would expect to see that behavior in real data associated with a location like Denver, in which the common influence of meteorology has a significant effect. With respect to PMF modeling, the relatively high cross-correlations between some of the input factor contribution time series has the implication that some of these factors may be harder to cleanly separate from others.

The authors have also investigated using the new method with various PMF2 settings. For example, the synthetic data set used in this work had an associated, "optimal" FPEAK of 0.24. The method of choosing this value followed the advice given in section 3.4 of Paatero, et al. (2002) regarding the behavior of Q values (the sum of the normalized, squared residuals). Note though that the following advice was also given (pg 257):

> *"...if there are not enough of zero values in either the G and F matrices, then rotating with [FPEAK] has no theoretical justification."*

Unfortunately, there is no subsequent discussion of what "enough" might mean, but it seems clear from the rest of that text that the primary use of FPEAK lays in changing its values to assess the range of possible PMF solutions, rather than its use as a parameter for which some optimum exists. The results of using FPEAK=0.24 were in fact qualitatively different from the results presented (where the default FPEAK value of 0 was used). For example, the plot for Diesel was noticeably more variable and biased than the associated plot in Fig. 3d of the manuscript. Unlike the Diesel factor though, when FPEAK was set to 0.24 the modeled Vegetative Detritus had a better contribution fit to the actual data than when FPEAK was set to 0. In general, however, the authors found that factor profile and contribution estimates were more variable across bootstrap runs when FPEAK was changed from the "assumption free" default of 0 to the "optimal" 0.24. The analysis method presented in the manuscript thus offers an additional way to optimize PMF2 results. Rather than choosing FPEAK based largely on the behavior of

ACPD
8, S2132–S2137, 2008

the Q values, one could also choose FPEAK based upon how reproducible or precise the results are across numerous bootstrap replicate data sets.

In a broader sense, the authors feel that the methods of optimizing PMF2 parameters (of which there are many more besides FPEAK) have been discussed elsewhere, and that that discussion is not directly relevant to the presentation of a new method for *assessing* PMF results.

Finally, the authors have taken note that in many settings in which the bootstrap is applied 500 replicate data sets is not very large, with 1000 being more common. Accordingly, the authors will include updated graphics and simulation statistics based on 1000 replicate data sets, in the hopes of obviating any concern for results based on small sample size.

2)

The authors agree that this is curious behavior. The authors suspect that the factors in Fig. 3c and 3h are ones that are possibly sensitive to extremes in the data, where values that may exist in the base case data set are not resampled uniformly in the bootstrapped data sets, and accordingly, the bootstrapped solutions are skewed relative to the base case solution. Although it is beyond the scope of this paper, more investigation is needed.

The authors' speculation opens the door to touch upon a related issue, one that presents a major challenge: how to devise a bootstrap method that better accounts for the correlation structure in multivariate time series data that are also often heavy-tailed in distribution? Rajagapolan (1999) offers a potential method based on a k-nearest-neighbor resampling from a multivariate state space. However, it has been used primarily with data sets comprised of observations on only a handful of variables, whereas the number of variables in today's speciated $PM_{2.5}$ data sets are often far more numerous. An approach for resampling heavy-tailed data is offered by LePage

ACPD

8, S2132–S2137, 2008

et al. (1998) and involves bootstrapping only the signs of model fit residuals. Recall, however, that for the synthetic data base case PMF fit, eight of the 39 species had residuals that were not independent. The authors are unaware of any published work for which PMF residuals were formally tested for the independence and distributional assumptions that go along with their subsequent use in analysis. For a given data set and PMF fit, if the residuals are not independent and identically distributed then a practitioner finds him/herself in the same situation as with the raw data: how to best resample multivariate data that has rich and varying temporal correlation?

3)
The authors appreciate the questioning of this assumption. The 0.3 value was derived from mass balance of Denver's 2003 daily, speciated $PM_{2.5}$ levels. In creating the synthetic data set (based on 2003's data), SOA was not part of the process, and the authors did not feel it was justified to vary the scaling by season given only one year of data. When working with real data this issue will certainly need to be addressed, and accordingly, the authors will include a qualifying statement to that effect.

4)
The authors appreciate the Referee's concern that the last paragraph of section 4.1 could be misinterpreted by readers. The authors will qualify that the results presented are based on synthetic data and that, especially with respect to certain factors, they should not be generalized. Moreover, focus should not be given to the results themselves but instead on the method of getting those results.

5)
The authors list these descriptive statistics to help give the reader an idea of how Normal the distribution of Q values is. The authors will add an explanation of kurtosis.

ACPD
8, S2132–S2137, 2008

6)
Thank you for catching this typographic error.

7)
Again, thank you for your attention to detail; the plot titles for Fig. 2a and 2b will be corrected. Also, in the interest of enhancing clarity, the authors will reorder the sources as listed in Table 1, such that the ordering in the table matches the ordering of the factors in Fig 2.

References

LePage, R., Podgorski, K., Ryznar, M., White, A.: A Practical Guide to Heavy Tails: Statistical Techniques for Analyzing Heavy Tailed Distributions. R. Adler, M. Taqqu, and R. Feldman eds. Birkhauser, pp 339-358, 1998.

Rajagopalan, B., Lall, U.: A k-nearest-neighbor simulator for daily precipitation and other weather variables, Water Resources Research, 35, 3089-3101, 1999.

Paatero, P., Hopke, P. K., Song, X. H., and Ramadan, Z.: Understanding and controlling rotations in factor analytic models, Chemometrics and Intelligent Laboratory Systems, 60, 253-264, 2002.

Interactive comment on Atmos. Chem. Phys. Discuss., 8, 2977, 2008.