Atmos. Meas. Tech. Discuss., 6, C3658–C3663, 2013
www.atmos-meas-tech-discuss.net/6/C3658/2013/

Atmospheric Measurement Techniques Discussions

Interactive comment on "Methods for estimating uncertainty in factor analytic solutions" *by* P. Paatero et al.

P. Paatero et al.

norris.gary@epa.gov

Received and published: 13 December 2013

Anonymous Referee #1 Received and published: 23 September 2013 >This is a potentially useful paper that provides a description of a new and interesting approach to estimation of the uncertainties in part of factor analysis solutions. However, it seems to be a strange paper to put in a measurement method journal like AMT. That is an editor's decision, but it would seem more appropriate in a statistical or chemometric journal.

Response: Factor analytic methods are commonly used to evaluate measurement data and analysis methods should be provided along with measurement methods. AMT states that "The main subject areas comprise the development, intercomparison and

validation of measurement instruments and techniques of data processing and information retrieval for gases, aerosols, and clouds." This paper focuses on data processing techniques used for analyzing gaseous, aerosol, and other environmental data, so it is within the scope of the journal.

First a general response to Reviewers 1 and 2: Reviewers 1 and 2 suggest enhancements (e.g. error estimates for G factor elements and more simulations, e.g. varying the number of zeros in assumed true factors) that would strengthen the paper and help users in applying the presented EE methods. In principle, we agree with the reviewers: the suggested enhancements would be useful. However, there are several reasons why these suggestions cannot be followed. (1) Purpose of simulations. The simulations are not intended, and cannot act as a proof of the usefulness of the methods because real data cause different kinds of unspecified modeling errors. Also, simulations are not a reliable basis for determining the most useful parameter values of the method, such as the maximum allowed change in Q (dQmax) in DISP and BS-DISP. Experience with modeling various types of real data is necessary. Simulations are presented in the hope that they illustrate the EE process and give examples of possible outcomes that may be obtained. (2) Length of paper. This paper is already fairly long, possibly causing that some readers will skip parts of it. (3) Time. Reviewers hope that the EE methods should be available by the end of year 2013. If more simulation studies should be carried out, that would need several months for the simulations. After that, much of the report of simulation results would need to be rewritten, and the whole review cycle, beginning with internal EPA review, would be repeated. The delay might be almost one year. (4)Resources. The entire project of EE development proved to be much more laborious than originally estimated. Much of writing this paper has been achieved based on unpaid (volunteer) work done by authors. More of such unpaid resources are not currently available, and further funding is unclear. If new resources should become available in 2014, it would be more useful (from the viewpoint of EE users) to devote those resources to enhancing the convergence rate of the algorithms applied in EE computations. In theory, an order-of-magnitude acceleration appears

possible. If it could be achieved, it would be a real help in practical work.

Regarding analyses of real data: in a companion paper, to be submitted soon, several real data sets are analyzed using the three methods BS, DISP, and BS-DISP.

>On page 5, there should be some discussion of the potential for unique solutions given a sufficient number of true zero (edge points) in the data set based on the Anderson (An Introduction to Multivariate Statistical Analysis, (2nd edition). Wiley: New York, 1984). People need to be reminded that the ability to resolve sources depends on the availability of edge points. This point can be reinforced as part of the discussion of self-modeling curve resolution in analytical chemistry.

Response: This discussion has been rewritten with more emphasis on the significance of zero values in factors. Added a reference to Anderson

>In the discussion of rotational ambiguity, it would be useful to refer the interested reader to the Paatero and Hopke (2009) paper discussing rotations in more detail. Another question about rotation that should be at least raised is how much of the rotational ambiguity is removed when you force the use of some fixed profiles as has been done by Amato and collaborators.

Response: References to Paatero and Hopke and to Amato et al were inserted.

Enhancing the model by constraints applied on source profiles is important whenever the customary PMF model contains too much of rotational ambiguity. We will generate a future paper that focuses on source profile constraints applied to initial results (ME or EPA PMF 5.0) to show that such constraints usually reduce the size of the error estimates by reducing rotational ambiguity. The impact of source profile constraints is of course data dependent and data sets with sufficient numbers of zero points are not strongly impacted by constraints. Detailed analysis of profile constraints is outside of the scope of the present paper, see also our general response, above. Because of the importance of applying constraints, this approach is also mentioned in Conclusions.

>Another key issue that is not adequately discussed is the variability of profiles in environmental data. There is a true, absolute absorbance spectrum for any given compound. The degree to which the measurements match that spectrum depend on things like slit width, dispersion in the monochromator, signal to noise in the detector, etc. Thus, it is not only the better precision with which AC measurements can be made, but the absolutely fixed shape of the profile. Such fixed profiles do not exist in the environmental receptor modeling problem.

Response: We agree with the reviewer about the importance of the variability of source profiles. We already discuss it as one of the sources of modeling error, in last paragraph on P.5. In order to further underline this importance, we insert the following 2 sentences to end of this paragraph: “As emphasized by an anonymous reviewer, it is to be expected that in environmental data, modeling errors are much more significant than in analytic chemistry measurements. In a follow-up paper, to be submitted soon, we apply these error estimation methods to three real data sets where modeling errors may be present.”

>All of the discussion is focused on the estimation of the errors in the F matrix, but the output of the model with policy implications is the G matrix because that points to those sources that contribute significant mass to the samples, particularly those samples that drive the violation of standards for which you are likely to be collecting and analyzing samples. Thus, there has to be some discussion as to why errors in G could not be estimated either in an analogous manner or if one could take the asymmetric intervals in the F matrix and estimate a range for each g value so we have some idea as to the likely accuracy of the contribution estimates. This is where the rubber meets the road with respect to the application of this technology to practical solutions of PM pollution issues. With EPA as an active participant in this paper, it is hard to understand why that perspective is not reflected anywhere in it.

Response: We understand the need for G error estimates and they are on our priority list. A new subsection 3.9 has been inserted in order to discuss the problems

associated with evaluating G uncertainties.

>From my experience with beta testing of V5.0, the use of DISP or DISP-BP is EXTREMELY time intensive. Thus, some discussion of the extent of computational resources needed to make the calculations should be provided. People have come to expect relatively instant results and here we are looking at many hours of computer time to produce a DISP solution for even a few input species. To really use it, you want to set it running for a weekend on a computer with a UPS to be sure there are no interruptions.

Response: New subsection 3.8 about computational workloads of different alternatives has been inserted. It is shown that an almost complete DISP run (with most species active) is comparable to a very limited BS-DISP run where only one element per factor is active. We assume that a combination of DISP and BS-DISP may be the optimal choice.

>The conclusions section has essentially no conclusions. It is a summary. It would be good to have some clear conclusions as to the value of the DISP and DISP-BP approaches relative to just the BP and some recommendations/guidelines as to when to apply what method. Right now it leaves the reader uniformed as to what this work means. Although it is not possible to provide guidance that covers all situations, there should be some ideas as to how to proceed to use these error estimation tools and how to interpret the results as to what of them are meaningful and what are not. Right now, there is not much clarity in how to apply them. It only says here they are. Minor Issues Below equation (1), it says “capital bold-face letters denote entire matrices, gk denotes columns of the factor contribution matrix G,” gk should be in bold in parallel with fk

Response: We have updated the conclusion to include the importance of the DISP swap diagnostics as well as a general comment on how to interpret the three error estimates.

>One hopes that the program will actually be released in 2013 since it has been a long time in beta testing.

Response: The uncertainty in EPA funding as well as an external peer review have delayed the release of EPA PMF 5.0 and the software should be completed in early 2014.

Interactive comment on Atmos. Meas. Tech. Discuss., 6, 7593, 2013.