Atmos. Chem. Phys. Discuss., 12, C6511–C6517, 2012
www.atmos-chem-phys-discuss.net/12/C6511/2012/

Atmospheric Chemistry and Physics Discussions

Interactive comment on "Analysis of secondary organic aerosol formation and aging using positive matrix factorization of high-resolution aerosol mass spectra: application to the dodecane low-NO_x system" *by* J. S. Craven et al.

P. Paatero (Referee)

Pentti.Paatero@helsinki.fi

Received and published: 3 September 2012

This paper describes results from an important and sucessful experiment that produced a massive amount of results. This reviewer evaluation is mostly concerned with the factor analytic modeling of the matrix of AMS spectra.

Because of the amount of results to report, some details are understandably omitted from the paper. Nevertheless, some hard facts should be reported, either in the main paper or in supporting documentation. Some of these details might be deduced by

careful reading of all text. However, the most basic info should be more easily available.

- What was (typically) the size of the matrix analyzed by PMF?

- Did the columns of the matrix correspond to individual m/z

values (or individual fractional m/z values) or to individual identified ions? Did you consider using a mixed matrix, containing both raw (perhaps fractional) m/z values and also a selected set of individual ions?

In PMF modeling, the following details are of special importance and are hence discussed here:

1. Formulation of std-dev values for matrix elements

2. Selecting number of factors to fit

3. Inspection of residuals

4. Reporting of Q contributions

5. Examination of rotational ambiguity of the model

6. If rotational ambiguity exists, controlling the rotational status of the solution so that a useful solution is obtained. Reporting of questions relating to rotations.

1. Std-dev values are carefully documented in the text, on pages 16665 to 16667.

2. The number of factors is well considered in the manuscript. The conclusion that at least 3 factors are needed appears solid, see also Fig. A4.

3. & 4. Figure A4 shows the behavior of residuals on columns and on rows for different numbers of factors. This is a well-though-of figure, much recommended. However, see also suggestions at the very end of this note.

5. Figure A5 is very useful for examining the effect of free rotations. It demonstrates how different factor shapes may be obtained with different rotational states of the solu-

tion. However, here we see a problem, see item 6.

6. The authors attempted to control the rotational status of the solution. Unfortunately, their attempt only went half-way, see detailed discussion and recommendations, below.

—— The question of rotational ambiguity in this work

When modeling speciated aerosol measurements, we may call one certain rotational status "correct". In such a correct rotation, the F factors (rows of F matrix) are identical (within experimental error) with profiles of existing physical sources. Achieving the correct rotation usually requires that in addition to matrix X, some additional information is available.

When modeling the formation of SOA, the situation is different. There are no external profiles to match. It does not make sense to call a rotation "correct" or "incorrect". Each set of G and F factors is equally "right" if they fit matrix X equally well. There is no direct physical interpretation of columns of G or rows of F.

However, there may be criteria for regarding one pair (G,F) more useful or "better" than another pair (G,F). Non-negativity is an obvious criterion: all G and all F should be >=0. Usually, non-negativity is not sufficient for reaching unambiguous factors.

The present results are close to violating the non-negativity requirement. If the experiment would be continued for a few hours, the present factor G1 (first time factor) would go negative, as is seen in Figure 9. Of course, if a longer experiment would be performed, a different rotation would be obtained so that G1 factor would again stay positive. The point is, however, that in such longer experiment, the obtained F factor profiles (mass spectra) would be different because of a rotation with respect to present results.

It is seen that the rotation obtained in present analysis is not "good" because it depends on the length of the experiment. In this work, rotation was controlled by varying the Fpeak parameter, as shown in Figure A5. The alternative rotations shown in Figure A5

appear worse than the 3-factor result in Figure 9.

The authors failed to realize that Fpeak is not a precision tool for examining possible competing rotations in any given situation. In so doing, they repeated the very common misunderstanding among PMF users: it is believed that if Fpeak does not provide a competing rotation, then no such competing rotation exists. Fpeak examines a one-dimensional path in the many-dimensional space of all rotations. (The rotational space is 6-dimensional for 3 factors and 12-dimensional for 4 factors). Only in exceptional cases does one find the desired alternative rotation by using Fpeak.

The proper way of examining rotations is to explicitly pull the factors in desired directions, see PMF User's Guide and

Pentti Paatero, Philip K. Hopke, Xin-Hua Song, and Ziad Ramadan, Understanding and controlling rotations in factor analytic models. Chemometrics and Intelligent Laboratory Systems 60 (2002) 253-264.

In the present study, a physically meaningful solution might be such where the G1 factor decreases with (approximately) exponential slope, without ever reaching the value of zero. Thus it would be necessary to try pulling up the last few elements of G1 factor. If an exponential-like decrease is obtained for G1 in this way without too much increase in Q, then the rotational problem of factor G1 has been solved. If, however, such rotation proves impossible because of too much rise in Q, then one should conclude that three factors do not allow a satisfactory fit. Then it would be necessary to repeat the modeling using four factors - a result that could not be reached simply by examining the Q values and residuals.

—— Recommendations for publication of this work

I definitely recommend this manuscript for publication, essentially in its present shape. I do not recommend that the results should be recomputed by applying a different rotation. Such exercise would require an unduly amount of work and its significance

on interpretation of results would be minor. However, I feel that is essential to prevent current computations from becoming a standard procedure, to be followed in future publications. In particular, future colleagues should not be relying solely on Fpeak in the search of alternative rotations. Thus I recommend that the authors should include more detailed discussion of the rotational question in the final version of this work, perhaps adapting some sentences from this note in their presentation.

—— Minor details

lines 16-17 on page 16665 are now: greater than 0.2. Signals with a S/N between 0.2 and 2 were down-weighted by a factor or 3, as recommended by Paatero (2003).

There are two errors. Corrected text is: greater than 0.2. Signals with a S/N between 0.2 and 2 were down-weighted by a factor of 3, as recommended by Paatero and Hopke (2003).

lines 26-27 on page 16671 (in References) are now: Paatero, P.: Discarding or down-weighting high-noise variables in factor analytic models, Anal. Chim. Acta, 490, 277-289, 2003. 16665

Corrected text is: Paatero, P. and Hopke, P.K.: Discarding or downweighting high-noise variables in factor analytic models, Anal. Chim. Acta, 490, 277-289, 2003. 16665

Equation A1 on page 16665: In both summations, the initial index value is given as 0. This seems to be a typo, the value of 1 should appear instead of 0 in both places.

Figures: please make the figures as wide as possible, using full width of the printable area. This request is especially important for figures 3, 8, 9, 10, and A4. These figures contain a wealth of detail that is not visible in the current size of the figures.

—— Suggestions for future experiments

As this experiment is first of its kind, future studies will be modeled after this first publication. In order to guide those follow-up studies, I wish to offer two suggestions that

might enhance the quality of follow-up results. I do -not- suggest that parts of this work be redone along these lines.

—— Examination of residuals

Average sizes of residuals on columns and on rows of X were successfully reported in this work. It would be useful to also examine the distribution of negative and positive residuals. Such examination might throw light e.g. on the problem discussed on lines 5-15 on p.16667.

I recommend a rectangular "map plot", displaying a dot for each residual element r_ij of X (or of part of X) so that positive residuals r_ij > sigma_ij are plotted with a red dot, and negative residuals r_ij < -sigma_ij are plotted with a blue dot. Small residuals, e.g. -sigma_ij < r_ij < sigma_ij may be omitted from the plot altogether, in order to make the plot less crowded. Alternatively, the command "spy" from matlab might be used.

Any regular patterns of red and/or of blue dots may indicate that some assumptions of the PMF model do not hold at that time and/or for those ions. A random pattern indicates a successful model.

—— Time resolution of the model

The used measurement techniques allow high time resolution. On the other hand, the system under study is not expected to display rapid variations. Thus the available time resolution is much higher than what is needed for adequate description of the physical-chemical processes under study. In this work, the normal attitude of scientists was followed: collect and use all information that you can get.

The bilinear PMF model has the peculiar counter-intuitive property that sometimes, additional information may be harmful for bilinear analysis of large matrices. This applies both to PMF and also to SVD (singular Value Decomposition). This property was discovered in 2003 and published as Pentti Paatero and Philip K. Hopke, Discarding or downweighting high-noise variables in factor analytic models. Analytica Chimica Acta

490 (2003) 277-289. The original publication was concerned with information carried in low-signal high-noise columns of the matrix. However, the problem is more general: any information that is irrelevant for the problem at hand may be harmful and should be downweighted or discarded from the matrix before performing the PMF analysis.

I propose that in future chamber studies of SOA formation, high-resolution data should be preprocessed before PMF modeling as follows: Combine sets of consecutive points together, effectively forming longer integration periods, so that e.g. points 1-5, 6-10, 11-15 and so on would be averaged together in sets of five. Possibly, the averaging periods might be longer, such as 10 or 20 points. This approach has several advantages:

- computations go faster and graphics become clearer

- removal of high-frequency noise will enable better rotations and enable detection of weaker components in PMF modeling

- enhanced signal-to-noise ratio may enable that such information becomes useful that had to be excluded from PMF in the present study, see p.16666.

- uncertainty of individual averaged points may perhaps be obtained from the average variation within each averaging group.

What about disadvantages of this approach? I am aware of only one problem: your colleagues may at first complain that you have thrown away some useful information, must not do so!

Interactive comment on Atmos. Chem. Phys. Discuss., 12, 16647, 2012.