Atmos. Chem. Phys. Discuss., 8, S11919–S11933, 2009
www.atmos-chem-phys-discuss.net/8/S11919/2009/

Atmospheric Chemistry and Physics Discussions

Interactive comment on "Interpretation of organic components from positive matrix factorization of aerosol mass spectrometric data" *by* I. M. Ulbrich et al.

I. M. Ulbrich et al.

Received and published: 1 April 2009

General Response

We thank referee P. Paatero for his detailed review and thoughtful comments on this work, which we think have helped us improve the work significantly. We've repeated his comments here in italics, numbering the items for easy cross-reference. Our replies follow each excerpt. Our response to these comments has been divided into two parts to fit within the allowed page length for Author Comments prescribed by ACPD (15 pgs.). This part includes responses through item 22 of P. Paatero's comments. Changes to the manuscript text are presented in bold. This response is intended to be a stand-alone, complete response to Prof. Paatero's comments, and for that reason some text

from our preliminary response to these comments is repeated here.

Detailed Response to Individual Comments

1. This manuscript contains an extensive analysis of what happens when Q-mode Aerosol Mass Spectrometer (Q-AMS) data are analyzed by the PMF model, using the program PMF2. The authors have attempted to find limits of what may be reliably deduced about the sources of organic aerosol in such analysis. As a result of their work, they express caveats against interpreting the results too strongly. This part of their work is mostly good.

[Response]: We thank the reviewer for this support of our work.

2. However, the work suffers from some basic errors in data analytic practices and understanding. In order to correct these errors, the work must be partly redone. Also, the presentation should be improved at the same time. More efficient use of graphics is needed, and mathematical notation should be made more clear and less ambiguous.

[Response]: Additional analyses have been performed in response to points from the reviewer (as indicated in points 4, 5, 8, and 12-13 below) and updated figures have been made available (see responses to each specific point below). Note that we can run and analyze many additional numerical experiments quickly because of our investment in the PMF Evaluation Tool (Fig. 2 in the paper).

3. Also, I recommend that the analysis of synthetic measurements (including solvability studies of variants of the basic case) should be separated from this work and published as a separate paper after correcting the essential problems in the present formulation. In this way, the results become accessible for a wider readership, including non-AMS and even non-aerosol scientists. Such separation should also be in the best interests of the authors: the main results should be published as soon as possible. Postponing the work that is needed for redoing the synthetic analysis and solvability studies helps in getting the main results out quickly. It is possible to cite the main results from the

ACPD

8, S11919–S11933, 2009

postponed part (in particular, the occurrence of similar factors when too many factors are used) in the main paper even although the final version of the postponed part is not yet submitted.

[Response]: As we have worked to address all of the comments, we found that we could address most of them except the one related to the number of near-zero values in the two-factor synthetic cases exploring retrievability of correlated factors (see point 16 below), due to the extreme complexity of this topic. Consistent with the suggestion from the reviewer, we have decided to remove Figs. 11 and 12 and the associated text in Sect. 2.3.2, 3.2.2, Abstract, Discussion, and Conclusions. We will further explore this subject in a future publication if time permits.

*4. The main method in the present work is the use of correlations of factor vectors (both of time series and of m/z profiles). One problem is that Pearson correlations do ignore the constant parts of the vectors. As viewed through correlations, the following two vectors are exactly identical: 10 11 12 12 11 10 and 0 1 2 2 1 0 Yet, physically, they represent two quite different behaviors of the physical system. Instead of correlation between two vectors x and y, the simpler "uncentered correlation", defined by expression (x!y)/sqrt(x!x * y!y) might be considered (here, the character ! denotes the dot product of the vectors, this is a non-standard notation). A very large number of correlations have been computed in this work. Yet, very little information has been gained from the correlations. Correlations or "uncentered correlations" may be quite useful in qualitative comparisons, such as comparisons between aerosol time series and time series of gases, RH, or other environmental variables.*

[Response]: Prof. Paatero explains this point further in his second set of comments (item 3) and we make a full response here. We agree that the Pearson correlation coefficient is not a perfect metric of correlation. We have compared the two correlation statistics for MS and TS in this study: the uncentered correlation metric that Prof. Paatero suggests is very similar to the Pearson's R for mass spectra, and quite correlated with Pearson's R for time series (when computed with a large number of different

ACPD

8, S11919–S11933, 2009

MS and TS; see figures available at http://tinyurl.com/4cohxf). We do use an alternative metric of correlation (Pearson's R for *m/z* $>$ 44 only) in Fig.6, and as stated in the paper (P6747 L2-9) we evaluated several other metrics of correlation (Spearman's R, and custom variations of Spearman's R with e.g. threshold to eliminate very small values) and found that they did not provide significantly more information than R and $R_{m/z>44}$.

Our own main concern with Pearson R is for comparing MS, which have values that vary over several orders of magnitude and tend to be dominated by a small subset of *m/z*'s. The uncentered correlation for MS has very similar values to the Pearson R and does not provide additional insight about the similarity or difference of the MS.

For the reasons suggested by the reviewer and to encourage its use by future researchers, we have used the uncentered correlation coefficient throughout the manuscript. In order to also provide values that can be compared to several previous works that only use Pearson R (e.g. Zhang et al., 2005ab; Lanz et al., 2007, 2008) we have versions of Figures 6 and 11-13 with Pearson's R in the supplementary information (Figs. S4, S11, S12). We encourage future authors to explore the use of both metrics.

5. In contrast, studies about rotations should mostly concentrate on appearance and disappearance of zero values and other distinct details in the factors. Tabulating correlations as functions of rotations do not appear too useful. I recommend that the use of correlations be strongly diminished in the updated version of the paper.

[Response]: We agree with the referee about the importance of zero values in the rotations. As described in the response to point 16 below, we have removed from the paper the 2-factor synthetic cases constructed with varying correlation between the factors, and the rotations of the solutions of these cases. However, we have retained Figure 10 from the ACPD paper (also Fig. 10 in the final version) because we feel that the qualitative characteristics of these plots do give some insight to the practitioner about the

ACPD

8, S11919–S11933, 2009

changes in the solutions. In general in our cases, positive FPEAKs tend to increase the correlation between the TS and decrease the correlation between the MS, while negative FPEAKs tend to increase the correlation between the MS and decrease the correlation between the TS. This type of analysis helps to narrow the set of reasonable solutions for consideration. We explain in the Discussion of the revised version that the use of factor correlations should be qualitative.

6. The language of the ms is good. The abstract appears rather long. No key words are suggested in the ms.

[Response]: We feel that we need length of the current abstract to address the many concepts in the manuscript. We suggest the following key words in the revised manuscript: "**Aerosol Mass Spectrometer, Synthetic Data, Mass Spectrometry, Factor Analysis, Primary Organic Aerosol (POA), Secondary Organic Aerosol (SOA)**"

7. It is not possible to cover all problems of the work in this review in sufficient detail. Thus, a new review appears necessary when the updated version is ready.

[Response]: We feel that this issue has been clarified in P. Paatero's second set of comments. Our full response appears there (see point 1 in that response).

- In the following, the problems are discussed one by one.

8. Estimation of data uncertainties and expected Q values.

Equation A1 in Appendix A defines how uncertainties (sigma values) were computed for data values x_ij. This equation is basically correct for the uncertainties of original x_ij caused by ion counting statistics (the sqrt expression) and by the variation of pulse heights from individual ions (the coefficient alpha=1.2). Unfortunately, the data values x_ij used in this work are not the original measured values but values obtained by a box-car smoother of width three, applied on the time series of each m/z value. Such smoothing decreases the sigma values of individual smoothed x_ij by a factor

ACPD

8, S11919–S11933, 2009

of sqrt(3)as is easily computed. Thus the sigma values used in this work have been too large by a factor of sqrt(3). Thus the correct expected Q values, expected for the sigma values used in the manuscript, are only one third of the values that appear in the manuscript. Thus the obtained Q values are approximately three times the expected Q, instead of being approximately equal as claimed in the ms. - When the corrected Q_exp is used, figure 7b becomes more plausible and resembles figure 7c: the smallest Q/Qexp values are approximately equal to unity, as they should be.

[Response]: As pointed out by Prof. Paatero, there was a mistake in the specification of the error matrix in the "Pittsburgh Real" case (but not on the synthetic cases) due to the application of 3-point boxcar smoothing to the data to reduce high-frequency noise. We apologize for this unfortunate mistake. Our evaluation of this effect indicates that all changes for the "Pittsburgh Real" case are very minor, and our conclusions could be evaluated with the figures presented in the paper. Updated versions of all affected figures were made available at http://tinyurl.com/4klulg. There are many accompanying plots with scatter plots between the versions from the submitted paper and the updated figures available from http://tinyurl.com/4emw7y. The factors obtained when the mistake is corrected are virtually identical to the factors reported in the discussion paper, but the Q-contribution values do increase. The factors obtained from the final analysis (with additional error treatments suggested by P. Paatero in his second set of comments, item 5) change the factors slightly. The Q-contributions do make more sense now; Q/Q_{exp} is near 1 during a few periods dominated by the aged OOA-1 whose spectrum appears to be quite stable and Q/Q_{exp} is larger than 1 during other periods, especially when HOA and OOA-2 make a larger contribution.

9. The larger values deviate up, so that the largest Q/Qexp contributions are between 10 and 15. Such values apparently occur for samples where "something happens". It would be of prime importance to find out what is this "something". The present ms does not solve this problem. It is good if the problem can be solved in the updated version of the ms. However, most likely a solution cannot be found very soon. Then it

ACPD

8, S11919–S11933, 2009

Interactive Comment

is more important to publish this work without such solution, outlining where the large Q contributions occur and maybe discussing possible reasons, even if no reason may be identified with certainty. The publication should not be delayed long in the hope of finding the reason for large Q contributions.

[Response]: We agree with the reviewer that it would be of prime importance to understand what happens during the periods of high residual which are not fit by factors with constant MS. We did begin to explain this in the ACPD manuscript (pg. 6752, L8-12) by saying, "The residual at specific *m/z*'s during periods of high OOA-II and high Q/Q_{exp} changes for many significant OOA-II *m/z*'s in modest amounts, fairly continuously, over periods of 10-20 min. This is likely caused by variations in the true OOA-II spectrum (which could occur, e.g., during condensation or evaporation of SVOCs) that cannot be represented by the constant-MS factor, nor are constant enough to become their own factor." A plot of the diurnal cycles of the Q/Q_{exp} and the residual is added to the revised version of the paper (Fig. S5), which highlights the fact that the Q-contribution is lowest when OOA-2 is near zero and it increases as soon as OOA-2 starts to appear in the evening. The variability in HOA may also play a role as suggested by its similar diurnal profile; however, a more detailed examination of the TS points to OOA-2 as a greater contributor to the Q-values.

Although most of the periods of sustained high Q/Q_{exp} appear to be due to this reason, the highest, short-lived spike is likely due to a specific HOA plume (e.g., a specific combustion source) whose spectrum is similar to but has some differences from the main HOA factor during the study and shows variation in *m/z* peaks with higher contribution to HOA than OOA-2. We have mentioned the effect of this HOA plume in the revised paper.

We are the first to attempt to give a specific explanation for the structure in the residual of a factor analysis of an AMS organic matrix. While the residual TS has been presented in other works (Zhang 2005a; Lanz et al 2007, 2008), no detailed explanation of the cause of this residual was given. Zhang et al. (2005a) notes several reasons why

ACPD

8, S11919–S11933, 2009

high residual could occur (chiefly the existence of more than 2 factors in the dataset) though some of those would be removed by the less-constrained PMF2 solution algorithm. Lanz et al. (2007) discussed the residual briefly and stated that during their study the high residual occurred during periods of high photochemical activity. This is not observed in our case, where the diurnal profile of the residual does not have a clear diurnal cycle (see time series and diurnal cycles at http://tinyurl.com/cr8b7t). Thus, we believe that the phenomenology in our case is different and we have described it as completely as we can with the information available.

In summary, we believe that the phenomena described above provide a plausible explanation for a large fraction of the Q above the theoretical limit.

10. A detailed analysis of the features of the obtained results does not appear motivated as long as there are essential but unknown weaknesses in the setup of the analysis.

[Response]: Our understanding after our preliminary response and P. Paatero's second set of comments is that this comment has been addressed.

11. - While considering this topic, the authors might wish to also check some earlier AMS publications: it is possible that sigma values of x_ij have been computed incorrectly in some earlier publications while this error has gone unnoticed so far.

[Response]: We are certain that this mistake has not been made in other factor analyses of AMS organics matrices published by our group. We have generally not smoothed the data in past analyses with PMF; however, we decided to smooth this dataset to maximize the comparability with the already-published results of Zhang et al. (2005ab), who did perform smoothing to remove high-frequency noise.

12. Construction and use of synthetic measurements

One "true case" (with variations) was used for generating the synthetic measurements. This true case was obtained as the PMF2 solution of the real data set. In this PMF2

ACPD

8, S11919–S11933, 2009

modeling, no rotational forcing was exercised ("FPEAK=0"). By definition, PMF2 then produces a "most central" solution, such that the solution avoids the border regions (small factor element values) as much as possible and also makes the strengths of individual sources (factors) as similar as possible.

[Response]: Here we clarify the construction of the 2- and 3-factor base cases. We will continue to use the language of the manuscript here to describe the synthetic cases. What P. Paatero calls "true case" we have called a "base case."

The 2-factor synthetic base case is not constructed from PMF as stated by the reviewer; it is the output of Zhang et al.'s (2005a) Custom Principal Component Analysis (CPCA) method (as already described in the ACPD paper P6744/L10-12), which solves the same bilinear model as PMF2, but uses a different algorithm (initial guess based on a priori knowledge about *m/z* 44 and 57 in AMS spectra), and doesn't impose the positivity constraint. This algorithm, similarly to PMF2, seeks to minimize the total chi-square but no rotational forcing is applied. Thus it is not clear that the solution of the CPCA should be reproduced by PMF2 with FPEAK=0.

The 3-factor synthetic base case is constructed as described by P. Paatero from factors obtained without rotational forcing.

13. The "true case" was then analyzed by PMF2. Different values of the FPEAK rotational forcing parameter were tried. It is no wonder that the true results are best recovered with a zero or near-zero FPEAK. This property is built-in in the true values themselves!

Consider a different scenario: the true values are taken from a PMF2 modeling with FPEAK=1, say. Then we expect that analysis of the synthetic data set gives best results with FPEAK=1! It is seen that the ms applies circular reasoning in this question. No real information has been gained about the best values of FPEAK for any situation where real data are analyzed. The reasoning about "best FPEAK" must be either removed or rewritten so that this circular reasoning is clearly explained. Also rewrite

ACPD

8, S11919–S11933, 2009

Interactive Comment

the corresponding paragraphs in discussion and/or conclusions.

[Response]: We carried out the analysis suggested by the reviewer and it appears that even solutions generated with FPEAK <> 0 still reproduce the input solution very well at FPEAK close to 0 for these particular cases. Some figures describing these results are available from http://tinyurl.com/4uzbsl. However, we agree that the empirical results for the synthetic cases constructed in this study do not prove that FPEAKs near 0 necessarily give the best solutions in real cases. We will note that these observations relate to this specific case, and will therefore remove the statements about FPEAKs from the abstract and conclusions. In order to explore the range of rotational ambiguity in other AMS datasets, in the revised version of the paper we will encourage other researchers to include plots of Q/Q_{exp} vs. FPEAK when presenting PMF solutions of AMS datasets.

P. Paatero further discusses this point in his second set of comments (see point 2 in those comments). In response to these comments and point 16 below, we have removed the part of Sect. 3.2.2 that dealt with correlations and FPEAKs. We have kept the descriptions of the effect of FPEAK on the real case and the synthetic base cases in Sect. 3.1.2 and 3.2.1.

14. Behavior of the PMF model when too many factors are used

This section is interesting in itself. To my knowledge, a similar analysis has never been published. However, the analysis should be more mathematical, less empirical. The discussion could run along the following lines. Consider an error-free 2-factor bilinear model

*X = G * F*

where G = [a b] and F' = [s t], and a, b, s, and t denote column vectors. It is easily seen that the following three-factor solution fits X exactly, i.e.

*X = [e f b] * [s s t]',*

ACPD

8, S11919–S11933, 2009

Interactive Comment

if e+f = a. This formulation gives us two families of basic three-factor solutions that are equivalent to the original two-factor solution. Different four-factor solutions are also possible. The following is one example of a basic four-factor solution:

*X = [e f b b] * [s s u v]',*

where e+f = a and u+v = t.

More solutions can be obtained as rotations of the preceding formulations, e.g.

*X = [e f b] * T * inv(T) * [s s t]',*

where e+f = a and T is a non-singular 3x3 matrix such that the rotated factor matrices

*[e f b] * T and inv(T) * [s s t]'*

do not contain negative values.

In contrast to the basic solutions, the rotated solutions need not contain repetitions of identical columns. Thus the repetition of identical factors is not necessarily present even if too many factors are used.

[Response]: We thank the reviewer for this useful suggestion. In this paper we do take the vantage point of doing an empirical evaluation of what PMF does when confronted with typical AMS data since we think this approach will inform future analyses of AMS with PMF. However, the suggested mathematical discussion will be useful, so we have incorporated it in the revised manuscript in Sect 2.2.1 in the discussion of choosing the number of factors. We have cited this comment and added an acknowledgement for this suggestion.

15. When performing experiments with "too many" factors, it is essential to compute from many random starts. Even if the "good" analyses might have globally unique solutions, the "too-many-factors solutions" may have several local solutions. - The ms does not say if random starts were used in the too-many-factors study. If no local solutions were obtained while multiple random starts were made, then this fact must

ACPD

8, S11919–S11933, 2009

be clearly stated in the manuscript.

[Response]: Trials with 64 multiple starts have been calculated for the real Pittsburgh case with solutions up to 5 factors and for the 2- and 3-factor synthetic base cases with solutions up to 6 factors. No other local solutions were observed in the solutions for the real case or the 3-factor synthetic base case. Solution of the 2-factor base case with 3 or more factors exhibited local solutions within a small range of Q/Q_{exp} values. The solutions for each number of factors fall into groups in which two factors represent HOA and OOA with high similarity to the input factors and the other factor represents various degrees of mixing of the HOA and OOA factors.

We have added more information about the solutions obtained with random starts to the paper, including the Q range for the cases described above in Sect. 3.1.1 and 3.2.1. We have also added a figure to the supplemental info of the solutions of 3-5 factors in the real case plots with the TS for all seeds overlaid and plots with the MS from the seed 0 case and the max and min values for each *m/z*. Running with multiple starts is a feature in the PMF Evaluation Panel and this is mentioned in the text. We will also recommend that other researchers run their analyses with multiple random starts.

16. Solvability studies

The ms explores the solvability of bilinear models as a function of the amount of correlations among the left and among the right factor vectors, by using variations of the synthetic case. In principle, this is an important topic. The following problem complicates the analysis, however: both the correlations of the factors and the numbers of zero values (and numbers of near-zero values, too) in the factors have a strong influence in solvability. When the variants were formulated in the ms, only the correlations were considered. However, the number and significance of (near) zero values do also change when variations are generated. This variation was not considered. Thus satisfactory interpretation of the obtained results is not possible because it is not clear what effects were caused by variations in correlations, what by variations in (near) zero

ACPD

8, S11919–S11933, 2009

Interactive Comment

values.

[Response]: We agree with the referee about the importance of zeros constraining the solvability of the solutions. We did explore this in earlier versions of our work, but it was not reported in the ACPD paper, partially due to concerns about excessive length.

We have performed preliminary analysis to determine definitions of "zero" and "near-zero" levels in factor MS and TS in order to address the criticism of this part of the work. We find that the complexity of this topic requires additional work and it is not possible to do it justice in the present paper. We therefore have removed all parts of the paper that deal with the variations on the two-factor base case constructed from correlations (Figs. 11 and 12 and the associated text in Sect. 2.3.2, 3.2.2, Abstract, Discussion, and Conclusions). If time permits, we will explore these topics in a future publication.

17. Correctness and clarity of mathematical presentation

This work is about a mathematical method. Yet, there are only four equations in the main part, and only one of them was formulated in this work. This is not how mathematical work should be published.

[Response]: We have added more mathematical discussion to the paper (see point 14 above about discussing "split" factors and point 5 above about rotation). From our point of view, this work is mainly about the application of the method to a specific type of data. We believe that it is useful to take a mainly empirical vantage point (see response to point 14 above), much the same way as a PMF user views the PMF output. Many people who use PMF do so in a "black-box" way by looking mainly at the solutions, so the evaluation procedures and guidance from the empirical approach should be useful to future practitioners. We agree that a very large amount of mathematical work along these lines and on further topics could be done, but this is beyond the scope of the present paper.

18. You should define concepts and notation, and show equations based on these,

ACPD

8, S11919–S11933, 2009

Interactive Comment

instead of writing hard-to-understand verbal descriptions. E.g. there are verbal expressions in the style of "correlation of HOA with OOA". What is this? Correlation of what with what? The reader has to research back and forth the text in order to find out. Please adopt a clear notation, and define it before first use. One possibility might be

corr(TS(HOA),TS(OOA))

which is almost self-explanatory. The same notation should be used everywhere, in the text, in figure captions, in figures, in tables, etc. As the ms is now, correlation is indicated in three (or more) different ways. You may choose your notation. But once chosen, please stick to it. Once a concise notation is adopted, many sentences become shorter or are replaced by tabular presentations. This will make the work easier to read.

[Response]: We appreciate Prof. Paatero's comment on this difficult point about notation. We have tried to clarify the notation, and have adapted the suggestion made here to adopt a new convention of the form to denote the uncentered correlation between the time series of HOA and OOA. We have applied this notation everywhere in the manuscript, including in the text, figure legends, figure captions, and tables.

19. Math details: - how to indicate matrix elements? It is permissible to use capital letters, e.g. X_ij (underscore denotes here subscripting) for matrix elements. The modern method is, however, to use lower-case for matrix elements (e.g. x_ij) and boldfaced upper case for the entire matrix. This has the advantage that different versions of a matrix can be indicated by subscripts (or superscripts) appended to the capital letter, without causing confusion between the entire matrix and the matrix element notation. Please select one method and use it consistently.

[Response]: We thank the referee for this clarifying suggestion. We have changed the references to matrices and matrix elements to be consistent with the suggested notation in Eqs. 1, 2, 4 and in the accompanying descriptive text.

ACPD

8, S11919–S11933, 2009

Interactive Comment

20. - Q is not "residual" p.6739 l 22: "A first criterion is the total scaled residual, Q." Caption of Fig.3: "Values of the normalized residual Q/Qexp and .." Both of these, and similar uses of Q, are wrong. The word residual denotes the (signed) difference (measured-fitted) for any data value x_ij. The symbol Q denotes the sum of squares of scaled residuals, summed over all data values. The sum may be simply called "Q value". Sums of squares of scaled residuals over parts of data matrix may be called "Q contributions". If in doubt, you may include these definitions in your text. But do not call Q a "residual".

[Response]: In the ACPD version, we were indeed using the term "residual" loosely to also include Q values. We have made the revised manuscript consistent with the reviewer's comment.

21. - in eq(1), the summation sign is needed. This is not a formal math publication where perhaps a "summation convention" might hold.

[Response]: This omission has been corrected in the revised manuscript.

22. - p.6739, l 28: The assumption of normally distributed errors does not belong to the assumptions needed for applying the bilinear model. The false statement occurs rather frequently in discussions of least-squares-based models. All least-squares (LS) models, including PCA and PMF, may well be applied to data whose errors are not normally distributed. Please avoid repeating false statements even if they occur in prior work. It is true that under the normality assumption, the LS results possess a certain optimality. But because this optimality is not needed nor mentioned in the present work, normality is not relevant as an assumption of the model.

[Response]: We appreciate the clarification from the reviewer, as indeed some previous literature is misleading on this topic. We will remove the mention of normally-distributed errors in the revised version of the manuscript.

Interactive comment on Atmos. Chem. Phys. Discuss., 8, 6729, 2008.