Atmos. Chem. Phys. Discuss., 8, S2059–S2068, 2008
www.atmos-chem-phys-discuss.net/8/S2059/2008/

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.

P. Paatero (Referee)

Pentti.Paatero@helsinki.fi

Received and published: 28 April 2008

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.

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

Full Screen / Esc

Printer-friendly Version

Interactive Discussion

Discussion Paper

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.

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 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.

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. 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

Full Screen / Esc

Printer-friendly Version

Interactive Discussion

Discussion Paper

in the updated version of the paper.

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

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

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 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. 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 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.

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. – 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.

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 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.

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 the corresponding paragraphs in discussion and/or conclusions.

Behavior of the PMF model when too many factors are used

Full Screen / Esc

Printer-friendly Version

Interactive Discussion

Discussion Paper

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]',

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.

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

Full Screen / Esc

Printer-friendly Version

Interactive Discussion

Discussion Paper

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 be clearly stated in the manuscript.

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 values.

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

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 bold-faced 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.

- 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".

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

- 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

Full Screen / Esc

Printer-friendly Version

Interactive Discussion

Discussion Paper

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.

- p.6740, ll 26-27: "Note that "solid body" geometric rotations of the factors are only a subset of the possible linear transformations." The remark is OK but the wording is unusual. The word "orthogonal" is typically used, e.g. "Note that orthogonal or "solid body" rotations of the factors are ..."

- in Eq(4), what does the term 0.001 mean? Do you mean that in all x_ij values (in Hz), a fixed amount of 0.001 was added. Why? Or do you mean that the std-dev of the Gaussian variate was increased by 0.001? If yes, correct the equation accordingly!

- p.6744, ll 3-4: "any ions of importance have enough counts to reach a Gaussian distribution to good approximation" This is true. However, the value of Q also encompasses a huge number of ions that are -not- important. If the simulations for those are not properly done, the obtained Q may deviate from Qexp simply because of the poor approximation used in the simulation. I am not sure if this risk is or is not important. Please watch out!

- p.6752, l 6: Was the SVD computed of the unscaled residual matrix? If not, what form of scaling was used? Note that the SVD of the unscaled residuals is quite inefficient because the std-dev of different values are so different. Also cf. p.6748, l 10.

- p.6754, lines 17-18: "As the simulated noise is white and thus has a large number of degrees of freedom..." This contains two errors: first, the noise is not white. Second, the form of distribution of the noise does not influence the number of degrees of freedom, hence the word "thus" is wrong. The correct formulation might be "The simulated noise has a large number of degrees of freedom..."

- p. 6755, ll 17-18: "The third factor lies 6 degrees out of the plane of the HOA and OOA factors." How was the "6 degrees" defined and computed? Was some form of

Full Screen / Esc

Printer-friendly Version

Interactive Discussion

Discussion Paper

weighting used in order to compensate for the different magnitudes of low and high m/z profile values?

- after Eq(A1): The shape of distribution of single-ion signals is not relevant. Omit "Gaussian". Write e.g. "account for the random variation of the heights of single-ion signals,"

- Eq(A2): Under the square root, the expression is a sum of products, such as in (matrix by vector). Please write the correct mathematical expression, with summation sign.

Other details

I do not understand figure A2. It says "errors". Does this mean error estimates sigma_ij, computed for real and for synthetic values x_ij? Or does it mean residuals? Use precise math notation, once again! What is the difference between "Time Variants" on x-axis and "Time Varying" on y-axis? For some points, the errors are much larger in horizontal direction, i.e. for the synthetic data set. Why is this so? I thought that the same expression was used in both cases for computing the sigma values.

Figures 9 are very informative when they are plotted in large scale. Instead of discussion of correlations, please pay attention to the details in these figures. It would be good to go through the details of how these figures change because of rotations. Also, connect the changes with the additions and subtractions that occur because of different FPEAK. Such discussions would be useful for your readers.

Better graphics are needed

This section is addressed both to the authors and the publishers! Many of the diagrams of this work contain a wealth of information. Unfortunately, this information is hidden by the graphical technique. It is true that by expanding the graphics on the computer screen, one may see the details. But many of us wish to have paper copies, e.g. in order to make annotations. In their original size, the diagrams are much too small for seeing the details (this is true both for the html version and for the printer-friendly

Full Screen / Esc

Printer-friendly Version

Interactive Discussion

Discussion Paper

version). On the other hand, I did not find any method at all for producing enlarged prints out of Adobe Acrobat Reader. The only method of producing enlarged prints seems to be via screen capture! In this way, it was possible to produce A4-sized (same as letter-sized) pictures of e.g. each half of figure 9. Only with this magnification, the important details in these figures became clearly visible. The authors should do their best in order to help the readers in this respect. Figures should be expanded horizontally so that they fill the full width of the page. In some cases, the figures might need to be split in two horizontal halves. Use enough markers on coordinate axes, etc. The publishers should work hard in order to create a presentation that allows the output of expanded diagrams directly out of Adobe Acrobat or other presentation format. After all, computers should enhance the flexibility of our work. With paper copiers, it was easy to enlarge the figures we copied. Why should this be impossible when using computers?

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

Full Screen / Esc

Printer-friendly Version

Interactive Discussion

Discussion Paper