Atmos. Chem. Phys. Discuss.,
doi:10.5194/acp-2017-124-RC2, 2017

Atmospheric Chemistry and Physics Discussions

Interactive comment on "Long-term chemical analysis and organic aerosol source apportionment at 9 sites in Central Europe: Source identification and uncertainty assessment" *by* Kaspar R. Daellenbach et al.

P. Paatero (Referee)

Pentti.Paatero@helsinki.fi

Received and published: 6 April 2017

This work develops first a mathematical machinery whereby they attempt to formulate a small number of factor spectra (2 assumed "known" mass spectra plus 4 fitted mass spectra) whereby a large number (819) of mass spectra, measured in 9 locations in different times of the year, could be sucessfully fitted.

In the second part, they perform a very large number of repetitions of the modeling, so that different details of the model are varied. From variation of the results, they deduce reliability estimates for the obtained spectral profiles and contribution time series.

The mathematical operations are not adequately explained. It is impossible to know what they have done (and why) in different stages of the work. One of their key concepts is a constraining of F factor elements. Unfortunately, the defining equation (3) is so unclear (and possibly contains a typo) that I cannot even guess what they might mean by this equation.

This review is limited to the first part of the work, modeling the measured mass spectra by a bilinear model. Although the second part is important, and has required a large amount of work, it cannot be analyzed in the time that would be available for such analysis. Thus I do not comment on the second part.

I recommend that this manuscript should be published after the listed problems in mathematical presentation and elsewhere have been corrected.

Problems in the mathematical presentation

The fundamental principle of science is repeatability. Used methods should be defined clearly and in sufficient detail so that colleagues will be able (at least in principle) to repeat what was done by the authors. In the following, I discuss mathematics that have not been described in an understandable manner. In general, equations are the language of mathematics. Mathematical work should be described using equations. Verbal explanations may only help in understanding the equations, they cannot replace equations. In order to use equations, it is necessary to define symbols for various quantities. If necessary, use two-letter symbols. Consider typographic questions: The prime should be avoided. Complicated notation in subscripts or in superscripts is often very difficult to read.

P5 L9-11 (page 5, lines 9-11)

> .. by the estimated organic matter (OM) concentration,

> calculated as the product of the OC concentrations .. and

> the OM/OC ratios from the AMS measurements.

Reading this verbal explanation, it is very hard to understand what was done. In fact, I would need to try to write the equations in order to understand. Please do this equation writing for the benefit of your readers!

P5 EQ(3): I do not understand eq(3) at all.

- what are f_kn and f'_kn (or is it f_kn'), what difference is there between them?

- is there a typo in the equation? As shown, the equation appears impossible.

This is a key detail in the manuscript as it describes constraints that perhaps have never been published in similar work. It must be described so that it is understandable.

What was included in matrix X, to be modeled by PMF.

It remains unclear what information was contained in X. Section 2 lists a number of variables whose concentrations were measured. Were all these included in X, or were some of these included, or none of them?

P5 L8-9 say: Input data and error matrices included 202 organic ions..

Do you mean that input data consisted of 202 organic ions? The formulation you used may also be interpreted so that among other data, input data also included 202 organic ions.

Also regarding matrix X:

Input data ... were rescaled by the estimated organic matter (OM) concentration ...

In PMF, one is allowed to rescale data rows in any way, provided that error rows are also scaled in the same manner. Thus your scaling is OK. However, it would help the reader a lot if you state briefly why to scale, what advantage was achieved by scaling. Your scaling (which does not change profile spectra in any way) is useful for plotting figure 9. On the other hand, your scaling influences (improves or worsens) correlations between factor values and marker concentrations. Please explain and/or correct. Note

that this scaling does not change the computations performed according to Eq(4) in any way.

P5 L5-7 (input errors):

> The input errors ... include ... the uncertainty related to ion

> counting statistics and ion-to-ion signal variability at the detector.

I understand that you counted the ions. Then, ion-to-ion signal variability is *not* a source of uncertainty. If ion current is measured, instead of counting ions, then ion signal variablity *is* a source of error. Please correct or clarify.

P5 L20-24

describe a complicated method for filling knowledge gaps in the known/assumed fixed-factor spectra ("reference profiles") of HOA and COA. I am not fully convinced about the performance of this method. The "natural" alternative method is to leave the unknown elements in reference profiles (profiles of HOA and COA) as ordinary factor elements, to be fitted by ME-2 together with all "normal" non-fixed F factor elements, as explained in detail below. Please include this remark in the corrected ms, so that future colleagues are encouraged to follow the safer and simpler method instead of your complicated method.

Using "constrained factors" based on known profiles of HOA and COA.

This topic was very difficult to understand at first. It was not clear what is "constrained" by what. Now I assume that you mean the following: In all PMF runs, two constant F factors were used, i.e. two rows of factor F were defined as a-priori fixed, so that the values of these constant factors were set equal to previously known mass spectra of HOA and COA. Is this what you mean? – Using constant or constrained factors is not familiar to PMF users, not at all. Such unusual methodology must be carefully explained so that all readers have the possibility to understand what you have done. In particular, you should explain that using fixed factors is not the same as using inequal-

ities in order to constrain factors to lie between upper and lower limits, set very close to each other. Also, you should go into technical details here, because it is possible to implement constant factors in two different ways in ME-2. You should guide your readers to the optimal usage. It is possible to use "constant factors" that reside in a different matrix, which is clumsy. The alternative is to keep all F factors in the same matrix but define that the elements in two first (say) rows of F are "locked", not allowed to change during the fitting process. These elements are set equal to the known profile values before initiating the fit. If there are gaps in the knowledge of HOA and COA, then those unknown elements in "locked" F rows should simply not be locked at all, so that they may obtain their best possible values during the fit. Use of constant factors or constrained factors often causes so-called "normalization conflicts". How did you protect your bilinear model against normalization conflicts? This is another important detail that should be communicated to colleagues who might follow your example.

P7 L16-17

> For each of the four PMF datasets, 2420 PMF runs were performed for

> evaluating the sensitivity of the model to the chosen a-value and the seed.

This statement mentions sensitivity of the model to random seed. The random seed determines the pseudorandom initial values of PMF fit. In plain language, this statement says that there were local solutions so that depending on seed, PMF iteration converged to different local solutions. Presumably, these solutions had comparable Q values because otherwise, Q values would be used for selecting between solutions. Now these different solutions are somehow pooled together and their presence is otherwise ignored.

The presence of multiple solutions should be properly reported (e.g. how many of PMF runs had multiple solutions, how many different individual solutions per PMF run were obtained at most and on the average, are the solutions rotationally equivalent having identical residuals of fit, etc.) There are no fixed rules on what to do with multiple so-

lutions. On one extreme, it has been suggested that scientists may at will pick the one solution they like most and ignore the others. At the other extreme, PMF modeling of such data may be considered failing if there are several local solutions with comparable Q values.

If DISP is used for uncertainty estimation of a case with several local solutions, one often obtains the outcome that the model is "Not Well Defined" or "NWD". I would not suggest what the authors should do with their many-solution cases in addition to discussing them. Whatever they opt to do, they should describe it: what was done and why.

Section 3.3, sensitivity analysis

I cannot comment more on this analysis because I do not understand what the a-values are and how they were used.

P6 L26-28 say:

> Paatero et al. (2014) compare the effectiveness in estimating modelling errors

> using two different approaches: the displacement (DISP) and bootstrap analysis

> (BS), respectively.

Here seems to be a terminological problem: in the quoted paper, Paatero et al. estimated the uncertainties of estimated F factor elements in the situation when no modelling errors are present. These F uncertainties depend mainly on random error in X and on rotational freedom of factor matrices G and F. It was specifically emphasized that the obtained uncertainties do not cover effects of modelling errors in the results. (Examples of modelling errors: non-constant factor profiles, wrong uncertainties assigned to data values.) Thus modelling errors were not estimated in the quoted paper.

P6 L29:

> DISP involves running PMF several times using randomly perturbed

> factor profile elements of a reference solution

In fact, DISP estimation does not involve any randomness at all. F factor elements are pertubed in a systematic fashion by DISP. Perhaps here is confusion with Monte Carlo methods where random perturbations may be applied. DISP is not Monte Carlo.

In table (3), uncertainty estimates of percentage concentrations are not correctly computed.

Notation:

In supplement, subscript "i" is used as a subscript of Q. It is not defined what "i" means here. Does it mean number of factors? If yes, then the symbol used for number of factors should be used. If i does not have a definable meaning, then it might be clearer to omit the subscript in this case. In general, systematic use of subscripts would be a help for the reader. E.g. use i only as the index of sample (time), j as index of column of X, k index of factor. For other quantities, select other symbols and define what they mean.

In different places, factor elements are denoted as G_ik and as g_ik.

This may confuse readers. Eventually they will recognize that this difference does not mean anything, but first they will waste time trying to understand. Either, select one notation (preferred), or, in the section "Notation" (to be written), specify that G_ik and g_ik mean the same, and also F_kj and f_kj mean the same.

Eq (2) is incorrect.

If you wish to use matrix element notation, then summation over k must be indicated. If you wish to use vector-matrix notation, then its use should be defined, especially because many of your readers are not familiar with such notation. In vector-matrix notation, index k would not be visible.

P5 L10: sunset –> Sunset OC/EC analyzer

P5 L21

> Fitted ions in our datasets missing in the reference ...

What do you mean by "fitted ions"? I do not understand this sentence.

P7 L21:

> The identity of HOA and COA were identified first as their mass spectra were

> initially constrained.

Why do you need to identify HOA and COA? I would assume that they are on preselected rows of F, such as rows 1 and 2. No identification is needed for factors that are in known positions. What is wrong here? Am I understanding all this completely wrong?

Supplement

Figure SI.1

shows ratios of obtained Q vs. expected Qexp. How were Qexp computed? Did you take into account that downweighted columns of X contribute very little to Qexp? How many downweighted columns were present?

Obtained ratios Q/Qexp are of the order of 10 for 6 factors. This indicates that there are one or several significant modeling errors. It has been assumed among atmospheric scientists that a ratio of 4, say, would not be significant, that it could be caused by random variations from the expected Q. This assumption is totally wrong. Such ratios (>1.5, say) always have a cause that preferably should be understood in a project where careful mathematical analysis is attempted.

Possible causes of modeling errors are: underestimation of random errors in mass spectra, variation of factor profiles with time and between sites, systematic errors in preprocessing m/z spectra, and spurious sporadic local sources that cannot be mod-

eled by PMF. An attempt should be made at understanding and discussing those errors, even if the effect to obtained results cannot be eliminated any more at this final stage. One useful diagnostic is to examine contributions to Q from different m/z values, from different times of day and days of week, and so on.

Table SI.1

Why is there a table for mass closure criteria (used for rejecting bad solutions) when all entries in this table are identical? The criteria seem to concern distribution of residuals of OC fitting. Is it really so that the fit is rejected if 1st quartile point is negative and 3rd quartile point is positive? In other words, the fit is rejected if residuals are symmetrical around zero. Usually, such residuals would be considered desirable.

Interactive comment on Atmos. Chem. Phys. Discuss., doi:10.5194/acp-2017-124, 2017.