Atmos. Chem. Phys. Discuss., 6, S5920–S5926, 2007
www.atmos-chem-phys-discuss.net/6/S5920/2007/

Atmospheric Chemistry and Physics Discussions

Interactive comment on "Source apportionment of submicron organic aerosols at an urban site by linear unmixing of aerosol mass spectra" *by* V. A. Lanz et al.

P. Paatero (Referee)

Pentti.Paatero@helsinki.fi

Received and published: 3 January 2007

This manuscript describes a bilinear (factor analytic, FA) analysis of a large matrix of aerosol mass spectra. To my knowledge, this is the first paper where the non-negatively constrained PMF method has been applied to such data.

Overall, the project has been carried through carefully and many aspects of the work have been well done. There is no reason to question the major results of the work. The manuscript deserves to be published and ACP is a good forum for it.

However, there are technical problems in the data analysis. It is possible that the weakest factor(s) is/are not well determined. Also, this paper will be the model for

other similar studies in this field. Thus it is important that questionable practices are not published in this paper, even if their use might not have significant detrimental effect in the present case. Otherwise, other scientists might consider such practices acceptable and even state-of-the-art, and might apply them in their future projects.

Items 1 to 8, below, discuss these problems in detail. The manuscript should not be published as it is now. The items 1 to 8 should be carefully considered and the manuscript should be updated accordingly. If some of my items are caused by misunderstanding the ms, then the manuscript should be clarified so that future readers will not suffer from the same misunderstandings.

1. Use and report Q values

This ms does not report even a single Q value (Q = sum of squares of scaled residuals). The R-squared diagnostic is not a sufficient replacement of Q because it does not take into account data uncertainties.

- First, the absolute level of Q values should be considered.

If Q values are too large by a significant factor (more than a factor of 2, say) then either, the model is wrong or, more usually, the standard deviations assigned to data values are too small. In the present case, the stochastical error of data values is correctly estimated but the authors have overlooked that there may also be modeling error. Examples of modeling error: emission profiles do not stay constant with time, the instrument has non-linearities or other distortions, aerosol particles undergo chemical reactions before reaching the receptor site. If the Q values appear too large, then one should usually include a proportional component in the standard deviations, in order to accomodate modeling errors. In practice, environmental data sets seem to require a proportional error component of 5% or more.

- Second, consider the decrease of Q values caused by increasing the number of factors by one

ACPD
6, S5920–S5926, 2007

EGU

If the decrease of Q is approximately equal to the sum of the dimensions of the matrix, or smaller, then including the new factor does not represent a better model. Then the new factor should not be included in the model unless the interpretability of the model improved significantly.

Sometimes, a few rows and/or columns of the matrix contribute an excessive amount to the overall Q. This indicates that those rows/columns represent something unusual that cannot be well modeled by the factor analytic model. Note that a FA model is intended for modeling the usual behavior only. Such unusual rows/columns should be either downweighted or omitted from the model. It would be useful to plot the Q contributions from each row, and similarly from each column. This should be especially useful with the present novel data set whose error characteristics are not yet fully understood.

2. Check for high-noise columns

Consider the possibility of having high-noise columns (or rows) in the data set, as discussed in the paper Pentti Paatero and Philip K. Hopke, Discarding or downweighting high-noise variables in factor analytic models. Analytica Chimica Acta 490 (2003) 277-289. In the present case, it appears possible that the masses above 100, say, might be high-noise columns. (Please publish information about average error levels in columns!). If some columns are indeed high-noise, then I recommend that such high-mass columns be added together in suitable groups. Instead of having individual mass columns (e.g. for masses 101, 102, ...300) one might have aggregate columns, e.g. one for masses 101-110, next for 111-120, etc. until 291-300. In such aggregate columns, the ratio of signal to (stochastical) noise becomes better than in individual columns. – In this way, the main part of the information in high-mass columns can be retained while eliminating most of the harmful noise.

3. Terminology

The ms uses the word "algorithm" wrongly. It is important to keep the concepts "model" and "algorithm" well separated. (Top of page 11688): Equation (1) describes the

ACPD
6, S5920–S5926, 2007

EGU

model PMF. This model can be fitted or "solved" by different algorithms: There are three different algorithms in the three programs PMF2, PMF3, and ME-2. All can be used for solving the PMF model of equation (1). Incidentally, none of these is based on alternating regression, as suggested on p. 11684.

In general, computer scientists are dealing with algorithms. Applied science, such as chemistry and physics, formulates different models according to the needs of the real-world situation. These models can be solved by using different algorithms. The choice of the algorithm usually does not matter as long as the algorithm does what it is supposed to do. – It is true that in chemometrics literature, the words algorithm and model are often misused. This fact is no excuse for continuing that practice.

According to standard terminology, the factor analytic model is called bilinear, not linear. The reason is that although the model is linear with respect to scores, and also with respect to loadings (G and F factors in PMF terminology), the model is not linear when considering all unknowns, i.e. scores and loadings together. It is true that the model is linear in the sense that the measured data are approximated by a linear sum of contributions from a number of sources. Despite of this, I recommend that standard mathematical or data analytical terminology be observed. Thus the title of the paper might be

"Source apportionment of submicron organic aerosols at an urban site by bilinear unmixing of aerosol mass spectra".

Another possible title might be

"Source apportionment of submicron organic aerosols at an urban site by factor analytic modeling of aerosol mass spectra".

4. Column scaling

P.11689, lines 14-17 say that columns of data and error matrices (ORG and S) were divided by the median values of data columns. It is claimed that this scaling enables the

ACPD
6, S5920–S5926, 2007

EGU

low-intensity columns to be significant. This claim is erroneous. The authors should consider the role of the uncertainties in equation (3). Equation (3) is formulated so that Q is invariant with respect to joint column scaling of ORG and S. Thus column scaling has no effect on the significance of low-intensity columns. (Of course, column scaling does not do any harm, either, because of the invariance of Q.) If the authors like to scale their data, it is OK. However, the same results are also obtained without the scaling. It is important that erroneous claims are deleted from the ms, so that other authors do not get the message that they MUST perform a similar scaling.

5. Interpretation of the sixth factor

My interpretation of the sixth factor is that its main purpose is to model the average signal in the high-mass columns. Its time behavior is quite different from all other factors, there are no tall sharp peaks. The variation is more stationary than in other factors.

There are two possibilities:

A. There is a special source for the the aerosol that creates the high-mass columns. This source would be of quite different characteristics than the named local sources. I wonder if the source could be the forests? Certainly it does not appear likely that the source could be cooking, i.e. a local source.

B. The sixth factor represents a measurement artefact. Possible reasons for such artefact: (a) varying background in the high-mass columns, (b) variable sensitivity (in relation to low-mass columns) of the high-mass columns

I suggest that cooking should not be named for the sixth factor. Instead, it might be called "unidentified factor, possibly artefact". I am not suggesting that only five factors should be used. It is good that the peculiar sixth factor is shown. The point is, do not over-interpret it.

6. Handling of isolated peaks

ACPD
6, S5920–S5926, 2007

EGU

Lines 20-21 on p. 11686 mention that nearby events (log-fires, charbroiling, and delivery vans) caused isolated organic aerosol peaks. The state of the art of factor analytic modeling is still at a loss about how to best deal with such peaks. Including them in the data set cannot be called an error. However, the result of including them in the data set may not be good, especially if such individual peaks give rise to simultaneous sharp peaks in two (or more) factors. E.g. modeling the peak caused by a van as 70% of hydrocarbons and 30% of wood burning would not be useful. (It is not possible to discern from the figures in the ms if such behavior actually happens in the present analysis).

The purpose of factor analysis is to model the recurrent or usual properties of the data set. Thus it might be good to downweight the unusual samples, e.g. such where a delivery van is known to be the source. One may argue that the mass spectrum of the nearby source is different from the overall spectrum of similar distant sources because the nearby emissions have not had time to oxidize similarly as the distant emissions. Thus inclusion of local sources may distort the factors and make them more difficult to interpret. Downweighting peaks caused by local sources is definitely not an error. Omitting such peaks entirely might sometimes be the best solution.

7. Error estimates in figures 7 and 9.

- Define the symbols in the box-and-whiskers plots in these figures. What are the end-of-box, the T symbols, and the o symbols?

The notches in these figures represent 90% confidence intervals of the hourly grouped median values. It appears that these confidence intervals have been derived under false assumptions: it has been assumed that all items in the population arise from one stationary probability distribution. If this assumption were true, then the intervals would be OK. However, in the present case, the probability distribution is randomly different in each different day. Then the used equation is not applicable, it is wrong. One might try to estimate the uncertainty of the medians by bootstrapping the days in the following

ACPD
6, S5920–S5926, 2007

way:

Consider the weekend/holidays case as an example. There are 7 days in these figures, denoted here by the 7 letters a,b,c,d,e,f,g. The reported median is obtained using the values from all 7 days a to g. In order to obtain an error estimate for the median, compute a number of "bootstrapped median values" by using resampled sets of days such as a,a,c,e,e,e,f and b,b,c,c,d,e,f and so on, 7 randomly chosen days in each set (see textbooks about the bootstrap method). The std-dev of the set of bootstrapped medians is an estimate of the uncertainty of the reported median value.

Instead of computing the uncertainty by way of bootstrapping, one might omit the uncertainty indicators altogether from figures 7 and 9. The wrong uncertainties, as present in the manuscript, must not be published.

8. Figures.

Publish figures 3 and 6 (the main results of the paper) in a more readable format. At least, they should be expanded so that they extend over the full width of the A4 page. Now these figures are absolutely unreadable. – Fortunately, it was possible to examine the figures in the .pdf file by expanding them on the screen. However, the paper should also be readable when printed on the paper. – Be nice towards the reader: align subfigures "a" .. "e" of figure 3 so that same-numbered channels are aligned! Similarly, align parts of figure 6. Indicate weekends and holidays in figure 6.

Interactive comment on Atmos. Chem. Phys. Discuss., 6, 11681, 2006.