Atmos. Chem. Phys. Discuss., 6, S5541–S5546, 2006
www.atmos-chem-phys-discuss.net/6/S5541/2006/

Atmospheric Chemistry and Physics Discussions

Interactive comment on "Simplified representation of atmospheric aerosol size distributions using absolute principal component analysis" *by* T. W. Chan and M. Mozurkewich

T. W. Chan and M. Mozurkewich

Received and published: 19 December 2006

Main Points

Comment on page S5112 regarding the choice of PCA over PMF. A brief discussion of PCA and PMF is already provided in the paper; we do not think that this is the place for an in-depth discusion. We do not agree with the referee's statement that "almost everyone in atmospheric chemistry uses" PMF; both methods are often used in this field. We also disagree with the referee's statement that "PMF-type methods are in principle more appropriate for atmospheric data". Although emissions should be non-negative, processing can produce anti-correlations which lead to physically real negative values in loadings. A classic example is seen in the sea salt aerosol where non-sea-salt sul-

fate is associated with a chloride deficit. We do point out in the introduction that a non-negativity constraint would be appropriate for the results of this paper. But since the lack of this constraint does not seem to have produced any serious problem, such a constraint would appear to be superfluous. In any event, the objective of this paper is merely data simplification. In the companion paper we do seek to extract physically meaningful components; but there we obtain negative loadings that we believe represent physically real anti-correlations.

Comment on page S5113 regarding Varimax rotation. The referee states that Varimax rotation "rotates the solution so that the first component has the maximum variance ... then does the same with the second component ... This is again not physical." In fact, the situation that the referee describes is produced by the diagonalization of the correlation matrix; it is indeed unphysical. Varimax rotation attempts to rectify this by rotating the coordinates to obtian the closest possible approximation to "simple structure"; that is, loadings that consist entirely of zeros and ones. There is indeed no proof that this should work, although there are hand-waving arguments to support it. But Varimax rotation has proven to be remarkably effective in a wide variety of problems, including ones in atmospheric chemistry. In the present paper it produces loadings that are extremely suitable for the purpose of data simplification. In the companion paper it produces results that appear to be physically real. So far as we are aware, that is the best that can be done with factor analysis techniques.

Comment on page S5113 comment on P10469/L25 regarding the criteria for determining “reasonable results”. This is, of course, a qualitative assessment. The criteria for "reasonable" vary with context. On P10469/L25 the criterion was, as stated, that components were distributed over the full particle size range. On P10470/L23 the criterion was that large values of chi square (i.e., numerically poor fits) were associated with visually poor fits but not with visually excellent fits. In the second paper, P10501/L8, we state that "The results were examined to determine if they provided a reasonable physical interpretation." This is described in general in the surrounding paragraph and

ACPD
6, S5541–S5546, 2006

EGU

in more detail in the discussion of the specific components.

P10470/L23 regarding the value for k2 and the lack of definition of chi square. Chi square is a standard statistical term and does not need to be defined. If the error estimates are good estimates of the true error, then chi square should be on the order of the degrees of freedom (a discussion of this may be found in Press et al., "Numerical Recipes"). This permits one to judge whether one's error estimates are roughly correct, especially when one has thousands of fits, as is the case here. As indicated in the text, that is what was done. The text does not pretend that the value for k2 is other than a rough estimate.

P10477/L18-20 regarding the characterization of the small fluctuations about zero as noise. We agree with almost all of what the referee says with respect to the origin of these fluctuations; this is already included in the text. It seems that the issue is one of semantics. We would maintain that just as errors can be random or systematic, noise can be random or systematic. An example of the later would be a 50 Hz fluctuation in a DC signal caused by coupling to line voltage. The fluctuations in the loadings are similar, except that their origin is in the numerics rather than in the measurement. We disagree with the referee's statement that "the representation of the growth of one mode into another ... necessitates these negative values to represent transitional distributions". The main modes of the loadings overlap, so linear combinations of them are prefectly capable of representing intermediate distributions.

P10479/L5 regarding the monomodality of the components and how the analysis would handle bimodal distributions from vehicle emissions. We remind the referee that the objective of this paper is data simplification and representation, not source identification. Requiring monomodal distributions is perfectly reasonable for this purpose. We think that reliable source identification probably requires data in addition to the size distributions; this is what we do in the companion paper. In that analysis there is no requirement of monodality. In any event, we suspect that in the real atmosphere the chance of seeing two distinct parts of the size distribution that belong to the same com-

ACPD
6, S5541–S5546, 2006

EGU

ponent is rare unless the samples are dominated by a single source (e.g., on a highway following a plume such in Kittelson et al., 2006).

P10480/L27 regarding the number of mixed components and processing. We only note that there "appears" to be a connection; we have no more support for this than is given. It would be interesting to see if future studies show the same tendency. We know of no way to associate specific components with processing in the context of the analysis presented here (in the companion paper, one component does appear to be the result of processing). We will try to clarify the text in this paragraph.

Regarding the detailed points, we will make corrections as suggested, with the exception of the following.

P10466/L25 regarding the difference between retaining additional components in our PCA compared to PMF. We are describing the PCA procedure. A description of the procedure for PMF would be inappropriate.

P10467/L13-16 regarding the sign of the loadings. We are puzzled as to just what is causing the confusion. If the loadings of two variables have the same sign, they are positively correlated. If they have opposite signs, they are negatively correlated. In either case, changing all the signs has no effect on the interpretation. That seems obvious to us. We will try to improve our text here.

P10470 regarding the use of surface and volume distributions for the analysis. With respect to this paper, this would make no difference since the column weights would have to be multiplied by the same scaling factors as the number concentrations. So the weighted data would be unchanged. With respect to the companion paper, the data scaling would also remove the effect of transforming the distribution. We will point this out in the revised text.

P10474/L27-28 regarding the small eigenvalues. In any measurement, some physically real variation will be lost in the noise and may contribute to the noise. As the

ACPD

6, S5541–S5546, 2006

referee points out, that variation will end up in the "noise" components in PCA. However, the uniform decrease in the eigenvalues of these components strongly suggests that they are dominated by noise. As shown in the paper, even omitting the "mixed" components permits almost all significant features to be captured.

P10475/L15 regarding removal of a paragraph to later section. This paragraph is part of a broader discussion of methods of choosing the number of components to retain. It belongs where it is.

P10475/L25 regarding why we used a method not included in Ferre's paper. The point of Ferre's paper is essentially that there is no single "right" method. We introduce a new method because, as stated in the paper, we found that it works best for this situation.

P10477/L20 regarding the rotated components being distributed over the entire size range. This properly belongs where it is: under results and discussion.

P10480/L4 The referee wrote "It would be interesting to see how these cases compare in both the mode diameters of the components and the associated scores." Why? We found nothing of interest here.

P10480/L12-25 regarding the request to include total concentration vs time on Fig 4. Changing Figure 4 would not add anything as far as this section of text would be concerned. The differences referred to here are so small that it would be difficult to devise a figure in which they would be visible. In any case, this is a relatively minor point that simply does not warrant more space in the paper.

P10481/L11 regarding whether any component shows a high correlation with transported pollutants, such as from Nanticoke. This is addressed in the companion paper.

P10482/L18 regarding whether 'principal component results are fully continuous' refer to loading. Yes, this is so.

P10482/L22 regarding the weighting used for social science data. We do say that it is not always appropriate for physical data. That is a key aspect of what we have done

ACPD
6, S5541–S5546, 2006

EGU

here.

P10484/L5-7 regarding logarithmic deviation and percentage deviation. Provided that the deviations are not very large (which is the case here), they are proportional, which is all that matters here. dln(x) = dx/x.

P10488/Table 1 regarding whether the deviation is over- or under-estimate. We don't see how this data can be put on the figure. The deviations are root-mean-square; they have no sign.

P10491/Fig 3 regarding use of absolute deviations. The point of the figure is to see when deviations are large. We have tried preserving the signs, but it actually makes makes the figure harder to understand. The color scale is dimensionless since relative deviations are plotted.

P10491/Fig 4 combining Fig 3 and 4. We think that interspersing the two types of figures, with different color scales and different vertical axes, would be very confusing. We don't understand what the referee thinks this would clarify.

We have corrected the typos and grammatical errors.

Interactive comment on Atmos. Chem. Phys. Discuss., 6, 10463, 2006.

ACPD

6, S5541–S5546, 2006

Interactive Comment