Atmos. Chem. Phys. Discuss., 6, S5111–S5118, 2006
www.atmos-chem-phys-discuss.net/6/S5111/2006/

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

Anonymous Referee #1

Received and published: 6 December 2006

Review of Simplified representation of atmospheric aerosol size distributions using absolute principal component analysis. Chan and Mozurkewich, ACPD, 6, 10463-10492 2006.

This paper presents a new method for factor analysis of size distribution data acquired with high size and time resolution (16 bins/decade, 5 min.) using Scanning Mobility Particle Sizing (SMPS) instruments. The application of factor analysis techniques to complex atmospheric data is of interest in order to both reduce the complexity of the data and identify sources, processes, relationships, and sinks of atmospheric constituents. The present paper develops a variant of the 'principal component analysis'

Full Screen / Esc
Printer-friendly Version
Interactive Discussion
Discussion Paper

technique that is based on linear algebraic eigenvalue analysis of the covariance matrix of the observations. The authors appear to have carefully studied the various aspects of the technique and adapted it to their needs. They don't just apply 'canned' routines as many other papers do, but rather they present several technical innovations such as a new method to weigh the input data matrix and a new graphical method to estimate the number of components needed. The paper is well written, interesting, of good technical and scientific quality, and of a subject appropriate for ACP. I recommend it for publication in ACP after the following issues have been addressed.

Main Points

The most important issue with this paper relates to the choice of the PCA technique, as opposed to a PMF-type technique. The main difference between these techniques is most clearly summarized in the companion paper (ACPD 6, 10493-10522) on P10496/L7-10: 'The major difference between the two techniques is that PCA does not have constraints on the values of either the component loadings or scores, but requires that the resulting components be orthogonal, while PMF requires component loadings and scores to be non-negative, but has no orthogonality requirement.' This is indeed an accurate description of the main difference between the techniques. What is not clearly discussed here is which type of technique is more appropriate for atmospheric data. In fact, atmospheric constituent concentrations are non-negative and non-orthogonal. Many studies have shown both characteristics, since different sources influence a site at the same time with some overlap of chemistry and size distributions. Thus PMF-type methods are in principle more appropriate for atmospheric data. This should be discussed in the manuscript. This does not mean that the specific Paatero 'black-box' implementation of PMF (which almost everyone in atmospheric chemistry uses) is perfect or superior to the current method. Clearly there is a benefit in trying different methods and comparing them, and there would be a danger if everyone used a single black-box program for factor analysis of atmospheric chemistry data. We commend the authors for the effort and thought they have put in this development and

ACPD
6, S5111–S5118, 2006

EGU

application, but again these issues should be discussed in more detail in the paper.

A similar issue involves the use of Varimax rotation. This procedure rotates the solution so that the first component has the maximum variance that a single component can capture, then does the same with the second component and the remaining variance, etc. This is again not physical. There is no reason why the main source or component affecting the data should have the maximum amount of variance. This should be discussed in the manuscript.

P10469/L25: What are the criteria for determining 'reasonable results'? This term is used repeatedly in the two manuscripts but is not sufficiently defined.

P10470/L23: use of 'reasonable results' - Value for k2 appears to be chosen because it looks good. Chi square is not defined (or used further in manuscript) and 'visually poor fits' is vague.

P10477/L18-20: The paper states here that 'negative oscillations should be regarded as 'noise'. This is incorrect. These oscillations are clearly systematic and are not measurement or numerical noise. Rather these oscillations appear because of the requirement of orthogonality of the size distribution basis set. In order for the scalar product of two components to be zero (= orthogonality) and since the components do have some overlap in the diameter axis, one of them has to be negative. Thus these oscillations are systematic and inherent to the method, and not just noise. Also the representation of the growth of one mode into another (e.g. nucleation into Aitken) necessitates these negative values to represent transitional distributions between the 'basis' modes. A more detailed and accurate discussion of these points is needed.

P10479/L5: a criterion used by the authors for choosing components is that they are all monomodal. However some particle sources such as motor vehicles typically produce multimodal size distributions, see e.g. Kittelson (1998) and Kittelson et al. (2006). Presumably with this approach the different modes from motor vehicles would be separated into different components. This should be discussed in the manuscript.

ACPD
6, S5111–S5118, 2006

EGU

P10480/L27: Please provide more support for the claim that components with a mix of noise and signal relate to the amount of atmospheric processing of the particles. This claim begs for the inclusion of eigenvalues with the reported component loadings, especially in the accompanying paper of source determination. This would allow the reader to see which of the component size distributions are attributed to processing.

Detailed Points

P10466/L20: Definition of 'reducing dimensionality' should be moved to head of this paragraph, where the term is introduced.

P10466/L25: The idea of 'retaining additional components' in this one PCA solution is very different from choosing the correct number of solutions in many independent PMF solutions. This should be mentioned in the paper.

P10467/L13-16: This would be better served with an example. How can positive/negative be arbitrary here, but have meaning about correlation when examining results? Is this changing of sign different than a rotation or linear transformation?

P10470, section 3.1: Volume and surface area distributions are often calculated from SMPS distributions. An alternative approach would be to look for components on the number, surface area, and volume distributions simultaneously. This would effectively use the same data with three different weights, but this would actually be useful if the data would be used later in surface area or volume form (as it is very often done). Have the authors tried this? I suggest that they briefly discuss this possibility in the revised paper.

P10470/L19: Doesn't 'variation in particle concentrations between the actual and the recorded concentrations' mean 'measurement error'?

P10472/L18: it is unclear what the authors mean here by 'orthogonal fitted surfaces that provide the best characterization of the dataset.' I suggest expanding this important characterization of the method in a couple of additional sentences.

ACPD

6, S5111–S5118, 2006

P10474/L5: The order of steps in the method begins to be confusing here. It appears that the true order is 1) weight data matrix; 2) calculate lambda and Q matrices; 3) rotate QR = QT; 4) remove weights, 5) normalize scores to probability functions. Is the number of components to be retained determined before rotation? This seems challenging if 'the component loadings obtained directly from the absolute principal component analysis are mathematical functions that have no physical meaning' (P10476/L17). Do X and Y need to be correspondingly ordered (when arranging by eigenvalue) and truncated (when retaining components)?

P10474/L27-28: it is concluded here that the error estimated with Eq (1) must be low because the eigenvalues are larger than 1/b. However it is also possible that there are many more components (real size distributions, or variations on size distributions arising from small sources or atmospheric processes) that come and go for short periods of time and that cannot be represented with the 4-9 main components for a given study. In other words, it seems to me that it is more likely that smaller eigenvalues represent a cascade of real atmospheric variability in size distributions, than true measurement error.

P10475/L15: This method does not appear to be used until section 4.3, paragraph 3. It could be moved to that section and condensed.

P10475/L25: This method didn't appear to be in the Ferre paper. Why were those methods mainly rejected and this substituted?

P10476/19: the word 'rotation' is often a misnomer. Factor analytical models typically use full-rank 'linear transformations' in which the final components are linear combinations of the original components. A good description of this difference is e.g. in Paatero et al. (2002), section 1.1.1. I suggest that a brief mention of this issue is added, this will help add some clarity to the field on this point. If the rotations used here are also rotations in the geometrical sense, this should be explained here and contrasted to the more general definition of Paatero et al.

ACPD

6, S5111–S5118, 2006

P10477/L20: Note about loadings being distributed over the entire measured size range would be better emphasized in section 3.3, final paragraph in the discussion of truncation of columns of Q.

P10480/L4: It would be interesting to see how these cases compare in both the mode diameters of the components and the associated scores.

P10480/L5: Figures 3 and 4 should be combined - see notes for P10492/Fig4.

P10480/L19: Add 'as expected' to the statement that the error decreases when more factors are retained (i.e. this is always true in PCA analyses).

P10480/L12-25: I suggest that a time series graph of the total concentration vs. time is added on top of Fig 4, showing the measured concentration and the reconstructed concentration with both options. This will help clarify what's being discussed here.

P10481/L2: The two sentences 'This is probably...' are not supported in this manuscript and would be better placed in the accompanying paper.

P10481/L11: Does the Simcoe APCA data show the difference between local pollution from Nanticoke and regional airflow in the scores of some factors, perhaps with some modes having occasional high scores that might correlate with wind direction?

P10482/L18: Does 'principal component results are fully continuous' refer to the loadings?

P10482/L22: It may add clarity to say that the weighting used in social sciences is not appropriate for physical data.

P10484/L5-7: the logarithmic deviations are being used here. However they are called 'percentage deviations', which mathematically is not the same.

P10488/Table 1: This data may be more useful as an additional line above the modified scree plots. Do these values represent over- or under-estimates of the original data? (Is it possible to overestimate the data when rejecting some components?)

ACPD

6, S5111–S5118, 2006

Interactive Comment

P10489/Fig1: Including a horizontal line at 0 would be helpful to see the oscillations away from the peak. Eigenvectors for each component would be very interesting. Scores for these components would be very interesting, especially if one can see the sizes grow into each other.

P10491/Fig3: Why use the absolute value of the deviation? It may better to show the sign of the deviation as well with a color scale. Also, what are the units in the image plot? These should be labeled.

10492/Fig4: This figure is less clear than Figure 3, but attempts to provide the same information. It would be better if these plots were combined as a series of 5 image plots, with the measured data on top, the 5 component solution next, followed by its deviations, then the 8 component solution followed by its deviations. This may not have been practical for the ACPD online version, but can be done for a full-page figure in the ACP version.

Grammar etc.

P10464/L15: ‘simply’ should be ‘simple’

P10466/L25: Insert ‘remaining’ before ‘unexplained variance’

P10466/L29: The point would be better emphasized by phrasing the first sentence as ‘As opposed to principal component analysis and positive matrix factorization... source identification..., our objective is...’

P10467/L1: Citation for PMF (Paatero and Tapper, 1994) should be made in this line.

P10468/L21: I suggest changing the names of the studies in this line to ‘Egbert-2003; Pacific-2001; etc.’ Otherwise it appears that those are literature references instead of the names of the field studies themselves.

P10469/L14: insert ‘at’ after ‘Air’

P10471/L2: insert ‘that’ after ‘To ensure’

ACPD
6, S5111–S5118, 2006

Interactive Comment

EGU

P10473/L5: Could also refer to section 3.4.

P10475/L17: replace ‘subjecting’ with ‘subjective’

P10476/L6: the greek letter sigma should be capitalized here

P10477/L2: Since scores represent absolute concentrations of particles associated with components, this would be a clearer label for figures with scores (in this paper and the accompanying paper).

P10477/L25: Clarify that the requirement of components having unit area came from the probability distribution scaling step.

References

PARAMETERIZATION OF AEROSOL SIZE DISTRIBUTION USING CONSTRAINED MATRIX FACTORIZATION., HEIKKI JUNNINEN, Markku Kulmala; University of Helsinki, Helsinki, Finland. Paper 9G21, International Aerosol Conference, 2006, Minneapolis, MN, USA.

Understanding and controlling rotations in factor analytic models. Pentti Paatero et al. Chemometrics and Intelligent Laboratory Systems 60 (2002) 253- 264.

David B. Kittelson. ENGINES AND NANOPARTICLES: A REVIEW. J. Aerosol Sci. Vol. 29, No. 5/6, pp. 575-588, 1998.

Kittelson et al. On-road and laboratory evaluation of combustion aerosols - Part 2: Summary of Spark Ignition Engine Results. J. Aerosol Sci. 37:931-949, 2006.

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