

Simplified representation of atmospheric aerosol size distributions using absolute principal component analysis

T. W. Chan^{1*} and M. Mozurkewich¹

¹Department of Chemistry and Centre for Atmospheric Chemistry, York University, Toronto, Ontario, Canada

*now at: Environment Canada, Toronto, Ontario, Canada

Abstract. Principal component analysis provides a fast and robust method to reduce the data dimensionality of an aerosol size distribution data set. Here we describe a methodology for applying principal component analysis to aerosol size distribution measurements. We illustrate the method by applying it to data obtained during five field studies. Most variations in the sub-micrometer aerosol size distribution over periods of weeks can be described using 5 components. Using 6 to 8 components preserves virtually all the information in the original data. A key aspect of our approach is the introduction of a new method to weight the data; this preserves the orthogonality of the components while taking the measurement uncertainties into account. We also describe a new method for identifying the approximate number of aerosol components needed to represent the measurement quantitatively. Applying Varimax rotation to the resultant components decomposes a distribution into independent monomodal distributions. Normalizing the components provides physical meaning to the component scores. The method is relatively simple, computationally fast, and numerically robust. The resulting data simplification provides an efficient method of representing complex data sets and should greatly assist in the analysis of size distribution data.

1 Introduction

Atmospheric aerosol particles affect the global climate both directly, by scattering and absorbing solar radiation, and indirectly, by increasing cloud lifetime and the number of cloud droplets (Schwartz, 1996; Twomey, 1991). Aerosol particles can also significantly degrade visibility (Cheng and Tsai, 2000; Barthelmie and Pryor, 1998). In addition, acute exposure to the atmospheric particulate matter leads to increased

respiratory diseases and mortality rates (Maynard and Maynard, 2002; Peters et al., 2000; Spurny, 2000; Schwela, 1996). Owing to the importance of the atmospheric particles in affecting atmospheric processes and human health, it is important to understand the processes that convert gas phase species to particulate matter and that modify particle size distributions.

Measurements of time series of aerosol size distributions can help us to understand how atmospheric aerosol particles evolve under the influence of processes such as nucleation, coagulation, and condensational growth. However, aerosol size distribution data sets can be difficult to handle and interpret due to the large amounts of data involved. To analyse the size distributions efficiently, data simplification is usually required prior to data analysis.

One classic method used to simplify aerosol size distributions is that of Whitby (Whitby, 1978; Knutson and Whitby, 1975) in which distributions are fit to three log-normal functions for the nucleation, accumulation, and coarse particle modes. More recent measurements have shown that there is often an additional Aitken particle mode; this is located between the nucleation and the accumulation particle modes. Thus four log-normal functions, three of which are in the submicrometer range, are often required to fit the entire size distribution.

One difficulty in fitting size distributions with log-normal functions is that it is usually necessary to specify pre-defined size ranges for the different particle modes; this is done to reduce the computational time and increase the numerical stability of the fitting program. In many cases, these ranges are determined based on the particle formation mechanisms. For example, Mäkelä et al. (2000) fit one year of 3–500 nm size distribution data from a forested site in southern Finland with three log-normal functions, to represent the nucleation, Aitken, and accumulation modes. They found that it was generally not possible to define fixed size ranges for these modes; this was due to particle growth from one mode to

Correspondence to: M. Mozurkewich
(mozurkew@yorku.ca)

another. They also found that the decision to include the nucleation mode in the fit was often difficult due to the high level of noise. In contrast, Mönkkönen et al. (2005) applied a similar fitting procedure to two weeks of 3–800 nm size distribution measurements obtained from New Delhi, India. In this case, since the three distinct maxima could easily be identified throughout the entire study, the three mode log-normal fits worked very well. However, atmospheric aerosol size distributions may possess more than three modes. Birmili et al. (2001) fit 17 months of 3–800 nm size distribution data, and found that number of modes required varied from two to five, depending on the origin of the air mass.

These examples illustrate the basic problem encountered in fitting atmospheric aerosol data to log-normal functions: the number of observed maxima in the distributions is a variable. Varying the number of modes used in the fit greatly complicates both the fitting process and the interpretation of the results. Trying to fit more modes than there are maxima leads to numerical instability, unless the parameters are tightly constrained. Forcing the modes into pre-defined size ranges becomes problematic when particle growth covers a wide range of sizes. Fitting atmospheric size distributions often requires more than three log-normal functions (Birmili et al., 2001); since each function requires three fitting parameters, this can lead to an excessive number of parameters. Thus, a better method of simplifying size distribution data is desirable.

Principal component analysis is an effective alternative for reducing the dimensionality of large data sets; that is, the number of components needed to describe most of the variance in the original data is generally much smaller than the original number of variables. This method uses correlations between variables to discover a smaller number of new variables, called components, that contain maximum information about the data. In analysing particle size distribution data, the measured input variables are the number concentrations measured in different size bins. The components obtained from the analysis will have the form of distribution functions. This occurs because the first component accounts for the maximum amount of data variance that can be represented by a single variable. Each successive component accounts for the maximum amount of the remaining unexplained variance in the data. An analysis that retains the first N components gives the best possible fit with N orthogonal variables. Since all components are orthogonal to each other, the regression fitting of the data to the components is simple and always numerically stable, no matter how many components are included in the analysis.

In atmospheric chemistry, factor analysis methods such as principal component analysis (Thurston and Spengler, 1985) and positive matrix factorization (Paatero and Tapper, 1994) have been mainly used for source identification and apportionment. However, the objective here is only to produce a simplified representation of size distribution data sets in order to assist in data interpretation and analysis. In the follow-

ing paper (Chan and Mozurkewich, 2007) we will show how these simplified representations can assist in source identification by using them in a conventional factor analysis.

Since the objective here is to provide a convenient means of simplifying data sets, we use principal component analysis rather than more complex techniques, such as positive matrix factorization (Paatero and Tapper, 1994). The former method is numerically much simpler to implement. The chief advantage of the latter method is that it guarantees non-negative results. Since all of the quantities to be obtained in the present analysis should be non-negative, this might be seen as a major advantage. However, in practice this does not seem to be the case, the results of our analysis gives components in which all loadings that are not near zero have the same sign; these can be chosen to be positive. Changing the sign of the loadings also changes the sign of the corresponding scores; when the large magnitude loadings are given a positive sign, the scores also tend to be positive. When this is done, the negative values of loadings and scores that occur are in the nature of noise; that is, they are part of fluctuations about zero.

Standard principal component analysis was developed largely for handling social science data. It begins by subtracting variable means from the data and then dividing by the variable standard deviations. This scaling is appropriate for social science data, where all variables are assumed to be equally significant and the absolute values have no meaning, but, it is inappropriate for physical data. Modifications can be introduced to remove the scaling effects from the rotated components and obtain absolute results (e.g., Thurston and Spengler, 1985); however, it is simpler to skip the scaling step. This is sometimes referred to as applying the analysis to covariances (e.g., Jackson, 1991).

Aerosol size distributions require significant additional modifications to the procedure. One difficulty is that aerosol size distributions, whether number or volume, possess very large variations in concentration; if the data are not weighted, this tends to force all the components into a limited portion of the size range as the procedure tries to account for small relative changes at those sizes for which the concentrations are the highest. If the data are scaled so that each variable has equal variance, then size bins in which the signal variance is largely due to instrumental noise are given equal importance with ones that have much less noise. Examples of how these effects can degrade the interpretation of data are given by Keenan and Kotula (2004), who have introduced a weighting method suitable for mass spectral data with pure Poisson noise. Our approach is similar in intent, but we introduce a more general method of weighting the data.

The absolute principal component analysis (APCA) used in this study produces a weighted least square fit to the data. The procedure described here follows the standard principal component analysis in selecting a subset of components to retain and then rotating the axes to obtain components that are more physically meaningful. We suggest a modification

of the standard scree plot for identifying the appropriate number of aerosol components to retain. We use the standard Varimax orthogonal rotation; the resulting components are a set of monomodal distributions with distinct peaks and noise about zero away from the peak. Fitting the measurements to the rotated components produces a time series of component scores; these represent the number concentration of each component present. In this paper, we use size distribution measurements obtained from five different field studies to illustrate the methodology.

2 Size distribution data sources

The size distribution measurements used here were obtained from five field studies: Egbert 2003, Pacific 2001, Hamilton 2000, Simcoe 2000, and Hamilton 1999. All size distributions were measured with a DMA-CPC system over 5-min intervals with 16 size bins per decade resolution. Ambient particles were size selected with a TSI 3071 differential mobility analyser (DMA) operating in a fast scan mode (Wang and Flagan, 1990). Particles exiting the DMA were counted by either a TSI 3010 or a TSI 3025 condensation particle counter (CPC). Details are given by Mozurkewich et al. (2004).

The Egbert 2003 data set was taken at the Meteorological Service of Canada's Centre for Atmospheric Research Experiments at Egbert; a rural site located about 80 km north of Toronto. Data were available for 22 days. Air flow from the south is often heavily influenced by urban emissions, whereas air from the north is relatively clean.

The Pacific 2001 data set was taken at Sumas Mountain (Eagle Ridge) in Abbotsford, Vancouver, B.C. Data were available for 17 days. This site is elevated by about 251 m above the nearby urban area and farmland, and sits above the inversion layer at night. A full description of this study is given by Li (2004). At this site, combination of biogenic and anthropogenic emissions are expected from various locations both at the site and away from the site. A detailed discussion of both the sampling site and the size distributions observed in this study is given by Mozurkewich et al. (2004).

The other three data sets were taken as part of the SON-TAS study. The Hamilton 2000 and Hamilton 1999 data sets were taken at Kelly station, an urban air quality monitoring site of the Ontario Ministry of the Environment, located in downtown Hamilton, Ontario. Air at this site is expected to be strongly impacted by local traffic and industrial emissions. Data were available for 11 days in 1999 and for 25 days in 2000. The Simcoe 2000 data were taken at a rural site about 70 km SW of Hamilton. This site is strongly impacted by trans-boundary transport from the United States. Data were available for 15 days.

3 Methodology

3.1 Applying weights to size distribution data

All the size distribution data used in this study were weighted using estimated uncertainties in order to ensure the production of more reliable results (Cochran and Horne, 1977). When no weights were applied to the data, all the rotated component loadings tended to be located below 200 nm; this is due to the highly varying number concentrations of particles below 200 nm. Weighting the data produced more reasonable results, with components distributed over the full particle size range. One consequence of weighting the data is that conversion to a surface or volume distribution should not alter the results; the scaling factors applied to make these conversions would also have to be applied to the weights and so would cancel out in the weighted data.

3.1.1 Estimation of measurement uncertainties

The size distribution data sets used in this study are in the form of a $(i \times j)$ matrix, with i scans and j size bins. Each entry contains the number concentration in the form of $dN/d\ln D_p$. A reasonable estimated weight, W_{ij} , (or the inverse of the square root of the variance) for each data point can be expressed as

$$W_{ij} = \left[(k_1 C_j)^2 + (C_j A_{ij}) + (k_2 A_{ij})^2 \right]^{-1/2} \quad (1)$$

where C_j is the concentration, $dN/d\ln D_p$, corresponding to one CPC count for size j ; A_{ij} is the measured concentration for size j in scan i ; k_1 and k_2 are constants.

The first term in Eq. (1) is the minimum counting increment, and serves to prevent W_{ij} from approaching infinity as the measured concentration approaches zero. In our DMA-CPC system, the TSI 3010 CPC appeared to count particles in multiples of 5 when the concentration is low, therefore k_1 was set to 5. For size distribution data measured by the TSI 3025 CPC, k_1 was set to unity. The second term represents the uncertainty due to counting statistics. When the measured concentration is high, the counting statistics term produces unreasonably low estimates of the uncertainty. To improve the uncertainty estimates, we also include the fractional error term (the third term). The fractional error term accounts for the combined flow fluctuation errors in both the DMA and CPC, which we expect to be proportional to the measured concentration. This fractional error also includes uncertainties associated with the fact that the DMA is a scanning sampler; the particle concentration recorded at a particular size may be different from the average concentration during the scan due to variations in the aerosol during the time required for the scan. Unfortunately, we do not have an independent estimate of this proportionality constant. However, setting k_2 to too small a value gave large values of chi square (i.e., numerically poor fits) for distributions with visually excellent fits. We found that $k_2=0.05$

produced reasonable results in that large values of chi square were associated with visually poor fits.

3.1.2 Determination of the row and column weights

To ensure that the orthogonal aerosol components obtained from principal component analysis are meaningful, the weights must be applied to the data prior to the analysis and must be removed from the aerosol components after the analysis. It appears that the most general weighting scheme that allows a preservation of the component orthogonality is the row and column weighting scheme of Cochran and Horne (1977). This assumes that the applied weights, V_{ij} , can be expressed as a product of a row weight (X_i) and a column weight (Y_j):

$$V_{ij} = X_i Y_j. \quad (2)$$

Typically, the actual weights given in Eq. (1) can not be factored in the form of Eq. (2). Therefore, we find the row and column weights that give the best possible estimate to the actual weights, W_{ij} . The optimum row and column weights are determined by minimizing the sum of the squares of the percentage deviations between V_{ij} and W_{ij} . The derivation given in the Appendix shows that the best estimated row and column weights are given by

$$X_i = \langle W_{ij} \rangle_i / \langle W_{ij} \rangle^{1/2} \quad (3)$$

and

$$Y_j = \langle W_{ij} \rangle_j / \langle W_{ij} \rangle^{1/2} \quad (4)$$

where $\langle W_{ij} \rangle$ is the geometric mean of all values in the weight matrix, $\langle W_{ij} \rangle_i$ is the geometric mean of all values in row i , and $\langle W_{ij} \rangle_j$ is the geometric mean of all values in column j . This is similar to the ad hoc procedure, using arithmetic means, suggested by Keenan and Kotula (2004) for Poisson noise.

Once the optimum row and column weights have been obtained, the unweighted data matrix, \mathbf{A} , is converted into a weighted data matrix, \mathbf{Z} , according to

$$\mathbf{Z} = \mathbf{XAY} \quad (5)$$

where \mathbf{X} and \mathbf{Y} are diagonal matrices that contain the row weights, X_i , on the main diagonal of \mathbf{X} , and the column weights, Y_j , on the main diagonal of \mathbf{Y} .

3.2 Absolute principal component analysis

Principal component analysis begins by determining the covariances between all pairs of variables in the data set. The covariance matrix, \mathbf{R} , can be expressed by

$$\mathbf{R} = \mathbf{Z}^T \mathbf{Z} / nb \quad (6)$$

where n and b are, respectively, the total number of scans (rows), and the total number of variables (columns) in \mathbf{Z} .

To obtain the principal components, the covariance matrix \mathbf{R} is diagonalized:

$$\mathbf{RQ} = \mathbf{QA} \quad (7)$$

where \mathbf{Q} is an orthonormal square matrix that contains the eigenvectors as the columns, and \mathbf{A} is a diagonal matrix with the corresponding eigenvalues along the main diagonal for the corresponding columns in \mathbf{Q} . For a data set with b variables, matrices \mathbf{R} , \mathbf{Q} , and \mathbf{A} all have b rows and b columns.

The eigenvectors in \mathbf{Q} are linear transformations of the original variables; we refer these as ‘‘components’’. The amount of variance explained by each component is given by the corresponding eigenvalue in \mathbf{A} . Finally, we sort the eigenvalues in \mathbf{A} in descending order, and arrange the columns in \mathbf{Q} according to the corresponding order of their eigenvalues.

Components with small eigenvalues are referred to as ‘‘noise,’’ and are meaningless in explaining the general trends in the original data. Hotelling (1933) proposed that all the noise components should have equal eigenvalues based on the assumption that these components all have equal random variations. For a finite data set, we expect some variations in the eigenvalues for the noise components. Therefore, sorting all eigenvalues yields a gradual decrease in the noise eigenvalues. We use the term ‘‘signal components’’ to refer to aerosol components that contribute significantly to the total variance; therefore, these are worth retaining. Data dimensionality can be reduced by retaining only the most important signal components, and discarding all noise components. In Sect. 4.2, we describe how can we separate the signal components from the noise components; this allows one to retain the suitable numbers of aerosol components.

3.3 Removing weights from scores and loadings

Removing the weights is essential to making the components and scores physically meaningful (Keiding et al., 1988). To do this, consider an unweighted ($n \times b$) data matrix, \mathbf{A} , that can be represented by an unweighted ($b \times b$) component loadings matrix, \mathbf{L} , and an unweighted ($n \times b$) scores matrix, \mathbf{S} , via

$$\mathbf{A} = \mathbf{SL}^T. \quad (8)$$

Similarly, a weighted data matrix, \mathbf{Z} , can be expressed by a weighted component loadings matrix, \mathbf{Q} , and a weighted scores matrix, \mathbf{P} , via

$$\mathbf{Z} = \mathbf{PQ}^T. \quad (9)$$

Note that \mathbf{S} , \mathbf{L} , \mathbf{P} , and \mathbf{Q} are all orthonormal matrices. Substituting Eq. (5) into (9), rearranging, and comparing with Eq. (8) shows that the weighted and unweighted scores and component loadings are related to the row and column weights by

$$\mathbf{L} = \mathbf{Y}^{-1} \mathbf{Q} \quad (10)$$

and

$$\mathbf{S} = \mathbf{X}^{-1}\mathbf{P} = \mathbf{A}\mathbf{Y}\mathbf{Q}. \quad (11)$$

Since both \mathbf{X} and \mathbf{Y} are diagonal matrices, \mathbf{X}^{-1} and \mathbf{Y}^{-1} are also diagonal. In \mathbf{X}^{-1} and \mathbf{Y}^{-1} , the diagonal entries are the reciprocals of the corresponding entries in \mathbf{X} and \mathbf{Y} . Equations (10) and (11) show that the weights can be removed by dividing each row in the scores and each column in the loadings by the corresponding row and column weights. The second equality in Eq. (11) is simpler to use since it avoids the intermediate step of computing \mathbf{P} .

Owing to the orthogonality of the component loadings, Eqs. (8) through (11) are valid for any subset of components. To obtain a subset of unweighted components, \mathbf{L} , and scores, \mathbf{S} , we first obtain the complete set of eigenvectors, \mathbf{Q} , sort the columns according to descending eigenvalue, and then eliminate the columns in \mathbf{Q} and \mathbf{L} that correspond to the undesired components. The issue of how many components to retain is addressed in Sect. 3.4. Once the subset of \mathbf{Q} is obtained, Eqs. (10) and (11) give the subset of unweighted components and scores. Using these subsets in Eq. (8) yields an approximation to the original data matrix, \mathbf{A} . Since all the aerosol components are orthogonal, this approximation is identical to the weighted least squares fit to the data by the retained components. After eliminating undesired columns in \mathbf{Q} and before removing the weights, a rotation is applied to the coordinate system; this is described in Sect. 3.5. Since the rotation is an orthogonal transformation, it does not alter the fit to the original data. As a result, \mathbf{Q}_R , the representation of \mathbf{Q} in the new coordinate system, may be used instead of \mathbf{Q} in Eqs. (10) and (11); this may be verified by substitution. Physically, the weights apply to the measured variables, independent of the coordinate system used.

3.4 Choosing the number of components to retain

The dimensionality of a data set is reduced by retaining fewer components than the number of original variables. However, there is no fixed method of deciding how many components should be retained. One classic method is to plot the eigenvalues in descending order against the component number; this is called a scree plot. The general rule is to look for a point at which there is a sharp change in the slope of the plot (Cattell, 1966), as suggested by the reasoning of Hotelling (1933) described in Sect. 3.2. When we applied this method to the aerosol size distribution measurements, we found that it always indicated fewer aerosol components than are needed to capture the visible features in the original measurements.

As pointed out by Cochran and Horne (1977), if we have an accurate estimate of the true measurement uncertainties, the eigenvalues for loadings that represent noise should be approximately $1/b$, where b is the total number of size bins in the measurements. The eigenvalues obtained for our size distribution measurements are much larger than $1/b$; thus,

Eq. (1) does not represent all possible errors in the size distributions. For example, the reported distribution data is implicitly treated as a true average over one measurement scan time. For DMA-CPC data, this is not true in a dynamic environment in which significant changes in the size distribution may occur within the time required for one scan. Additional errors are introduced by the fact that the applied weights, V_{ij} , are approximations to the true weights, W_{ij} . These difficulties prevent us from using the criteria of Cochran and Horne to select the number of components to retain.

Nine different methods that are available in the literature for determining the number of components to retain have been tested by Ferre (1995). He concluded that there is no universal method which works for every application. The most suitable method for determining the appropriate number of components to retain depends on both the nature of the data set and the objective of the user: whether the aim is to obtain a “good explanation” (good fit to the data), or to obtain a “good prediction” (a good estimation of the parameters of a model).

Another approach to selecting the proper number of components to retain is to compare the original measurements with the fitted data generated using various numbers of retained components. The decision is somewhat subjective since it depends on what is deemed an adequate fit. For example, is it sufficient for the fit to capture the general trends in the measurements, or should it reproduce all significant features of the data? This procedure is also cumbersome to apply. In the following we use this method to judge the success of our procedure for selecting the number of components to retain.

From analysing different aerosol size distribution measurements, we found that the most successful and effective method to determine the number of components to retain is based on χ_j , which we define as the square root of the sum of the unused eigenvalues:

$$\chi_j \equiv \sqrt{\sum_{i=j+1}^b \Lambda_i} \quad (12)$$

where Λ_i is the eigenvalue for eigenvector i , b is the total number of variables in the measurement, and j is the number of retained components. Since the covariance matrix in Eq. (6) is standardized by the total number of data points in the data set (nb), all eigenvectors will have unit length, and the eigenvalue of each component represents the scaled variance contributed by the corresponding component (Jolliffe, 1986). In Eq. (12), $\sum \Lambda_i$, represents the total variance associated with the unused components. The square root of this quantity represents the deviation between the original data and the fitted data based on retaining j aerosol components.

The procedure for using this is to make a plot of χ_j against the number of retained aerosol components; we call this the “modified scree plot,” because of its similarity to the traditional scree plot. Ideally, a sharp break in the plot would

distinguish the signal components from the noise components. In practice, the break is gradual, so that there is a range of possible values for the number of components to be retained. The actual number retained depends on the extent to which an accurate fit to the data is desired. This will be discussed in detail in Sect. 4.2.

3.5 Rotation and normalization of loadings and scores

The component loadings obtained directly from the absolute principal component analysis are mathematical functions that have no physical meaning. In order to provide a physical meaning to each component, rotation of the retained components is required (Buharma et al., 1998). We adapt the widely used Varimax procedure (Comrey and Lee, 1992) to obtain the optimal rotation matrix, \mathbf{T} . Some workers incorrectly use the term rotation to refer to other types of linear transformations, we don't. This matrix relates the rotated components, \mathbf{Q}_R , to the original non-rotated components, \mathbf{Q} , via

$$\mathbf{Q}_R = \mathbf{Q}\mathbf{T}. \quad (13)$$

Since both \mathbf{Q} and \mathbf{T} are orthonormal, \mathbf{Q}_R is also orthonormal. Once \mathbf{Q}_R is obtained, it can be used in place of \mathbf{Q} in Eqs. (10) and (11) to obtain the unweighted components and scores.

For aerosol data, it is desirable for the component loadings to be in the form of probability distributions, so that the corresponding score represents the absolute concentration of particles associated with the component. To do this, the loadings for each component are normalized in the probability sense. The normalization factor for each component is obtained by integrating its loadings over the entire size range, taking into account the logarithmic spacing of the size bins. Then the loadings are divided by the normalization factor, and the corresponding scores are multiplied by the same factor. This normalization procedure causes the aerosol components to be no longer normalized in the vector algebra sense; therefore, this procedure is done after using Eq. (11) to obtain the component scores.

4 Results and discussion

4.1 Nature of the rotated components

Figure 1 shows the rotated component loadings obtained for each field study. In each case, the results shown are those obtained when retaining the maximum number of components indicated by the method described in Sect. 3.4. Once the components have been rotated and the effects of weights have been removed, the dominant feature in each rotated component has a shape similar to a single mode size distribution. In addition, away from the peak there are oscillations about zero; this is a consequence of the orthogonality condition. We believe that these oscillations should be regarded as a

kind of systematic noise. Note that the loadings are distributed over the entire measured size range. These basic features of the rotated components are preserved when fewer mixed components are retained. However, when too few or too many components are retained, the single mode structures are not obtained.

In several of the data sets, the smallest diameter component is truncated and has a greater amplitude than the others. This is a consequence of requiring each component to have unit area in order to transform the component loadings to a probability function (see Sect. 3.5). The 9nm component is missing in the Hamilton 2000 data set, apparently as a consequence of the cloudy and rainy weather encountered during that study. As mentioned earlier, we can not assign the identified components directly to specific sources. However, when combined with other data, these components are useful in identifying sources; this will be addressed in the accompanying paper (Chan and Mozurkewich, 2007).

4.2 Number of components retained

The modified scree plots for the Egbert 2003, Pacific 2001, Hamilton 2000, Simcoe 2000, and Hamilton 1999 data sets are shown in Fig. 2. For a data set with b variables, there are $(b-1)$ points in these plots. The point for retaining all b components is not included in the plot, since this always gives a perfect fit with zero deviation. The point for retaining zero components is also omitted because it has no practical use.

The points on the modified scree plot are fit to a four parameter function, which is defined as the greater of two straight lines. This divides the points into three categories: signal, noise, and mixed components. Components that fall on the first straight line segment are classified as signal components, while those that fall on the second straight line segment are classified as noise components. The rationale for this is the same as for the standard scree plot. The mixed components are the ones that contain significant amount of both signal and noise.

This interpretation is supported by tests with synthetic data. Those tests suggest that the signal components represent critical features in the original data set and should always be retained, while the noise components represent unimportant features and should always be discarded. They also showed that the mixed components tend to represent features that appear only in a portion of the data set. Therefore, the choice of how many mixed components should be retained depends on how important these small features are to the user. Specifically, retaining only the signal components seems to be sufficient to fit the general trends in the data set, while some or all the mixed components are needed to be able to fit all significant visual features in the data set.

From Fig. 2, we see that from 5 to 7 components should be retained for the Egbert 2003 data set, from 4 to 8 should be retained for Pacific 2001, from 4 to 5 for Hamilton 2000,

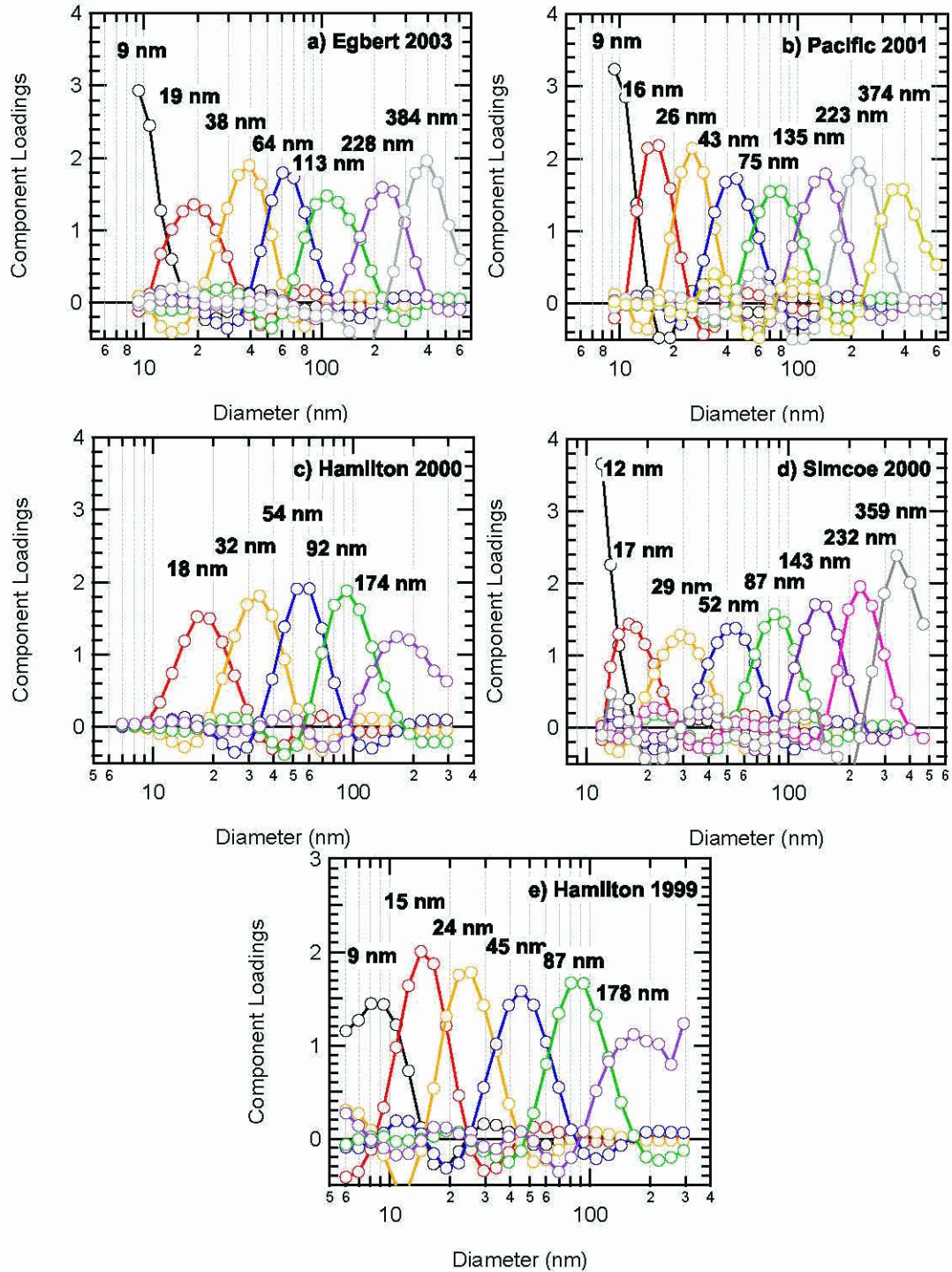


Fig. 1. The relative positions and shapes of the components after Varimax rotation and probability normalization for the five field studies. Each component is labelled with its modal diameter, as determined by fitting the component loadings to log-normal distributions.

from 6 to 10 for Simcoe 2000, and from 5 to 6 for Hamilton 1999. For the Pacific 2001 data set, we conclude that the minimum number of components to retain should be increased to five due to the shape of the rotated components;

with four components, one of the rotated components shows a bimodal structure, while with five components, all rotated components are monomodal.

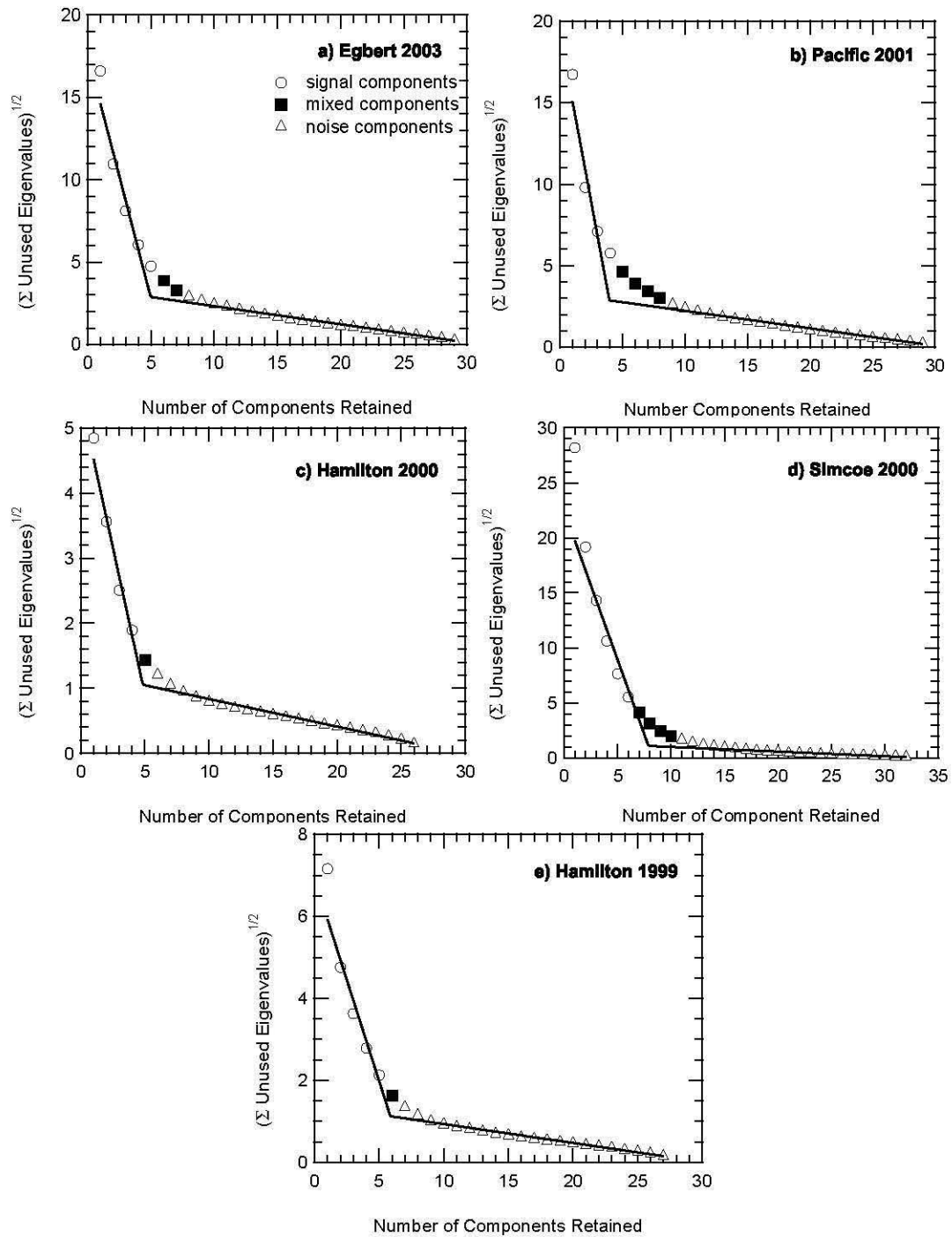


Fig. 2. Modified scree plots for various field studies. Open circles represent signal components, open triangles represent noise components, and solid squares represent mixed components. The solid line indicates the best fit to a four parameter function defined as the greater of two straight lines.

4.3 Quality of fits

As noted above, the modified scree plot does not provide an unambiguous result for the number of components to retain;

deciding whether to include the mixed components is somewhat subjective. In all five data sets considered here, we found that when all the mixed components are retained we obtain excellent fits to the original data throughout each data

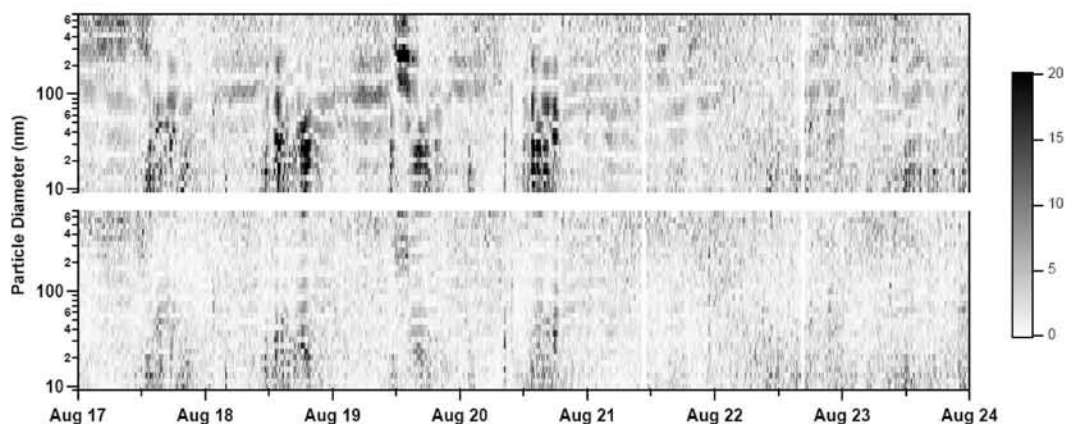


Fig. 3. Relative deviations between the measured and fitted data for one week of data from Pacific 2001. The components used were determined from the entire 17 day data set. The top panel is for the retention of 5 components and the bottom panel is for the retention of 8 components. The darkness scale corresponds with the product of the absolute deviations with their approximate weights.

set. In the case of the Simcoe 2000 data set, we found that only 8 components (instead of 10, indicated by the modified scree plot) are sufficient to capture virtually all features in the original measurements. When some or all of the mixed components are omitted, the fits are degraded slightly during most time periods and substantially during others.

An example of these comparisons is shown in Fig. 3, for the retention of either 5 or 8 components in the Pacific 2001 data. Although the figure shows just six days of data, the components used were derived from the entire study and the results in Fig. 3 are representative of the entire study. For the comparison, we multiply the absolute deviations (that is, the absolute values of the differences between the measured and fitted data) by the estimated weights (V_{ij} , Eq. 2); these relative deviations provide an indication of how large the deviations are in comparison with what would be expected from the measurement uncertainty. In Fig. 3, we see that during the second half of the time period (21 August to 24 August), both the 5 and 8 component fits reproduce the original data well, with no large systematic deviations. In contrast, during the first half of the period (17 August to 21 August), the 5 component fit shows some large systematic deviations. This shows that although 5 components are adequate to fit most of the data set, more mixed components are needed in order to fit the entire data set quantitatively. When we apply principal component analysis separately to the periods from 17 August to 21 and 21 August to 24, the corresponding modified scree plots show that the former period requires 6 to 9 components, while the latter period requires 3 to 6 components.

In Fig. 4, we show the comparison between the Pacific 2001 measurements and fits obtained by using either 5 or 8 components. For clarity, only the period from 15 August to 21 August is shown, the results are representative of the entire 17 day study. At most times, both fits reproduce the measurements very well. However, the 5 component fit has some significant deviations during the two circled peri-

ods. These results are typical of those obtained from all five data sets.

As a result of the normalization procedure; the component scores represent the absolute concentrations of particles associated with each component. Thus, the sum of all rotated component scores should be equal to the total number concentration of the measured size distribution. We tested this for all field data sets; the corresponding r.m.s. deviations between the sum of all scores and the integrated DMA total number concentration are summarized in Table 1. The comparisons were carried out using both the minimum and maximum numbers of retained components. In the former case, when including only the signal components, the r.m.s. deviation varies from 1.0% to 2.5%. As expected, when the mixed components were included, the r.m.s. deviations are slightly smaller, ranging from 0.75% to 1.9%. Note that since the components are orthogonal, the scores for individual non-rotated components do not depend on how many components are retained. Thus, the difference between the two sets of r.m.s. deviations in Table 1 is caused by the additional components that are retained in each data set. The small differences of the two sets of r.m.s. deviations in Table 1 are indicative of the relatively small impact of the mixed components on fitting the data set as a whole.

4.4 Interpretation

We find that the number of mixed components in a modified scree plot appears to relate to the amount of atmospheric processing of the sampled aerosol. Among the five field studies considered, both Hamilton 1999 and 2000 data sets are the simplest, with the fewest number of mixed components (see Fig. 2). This may be because the measurement site is located within a source region, where the air is strongly affected by local vehicle and industrial emissions. As a result, we might expect that the individual components would be most nearly

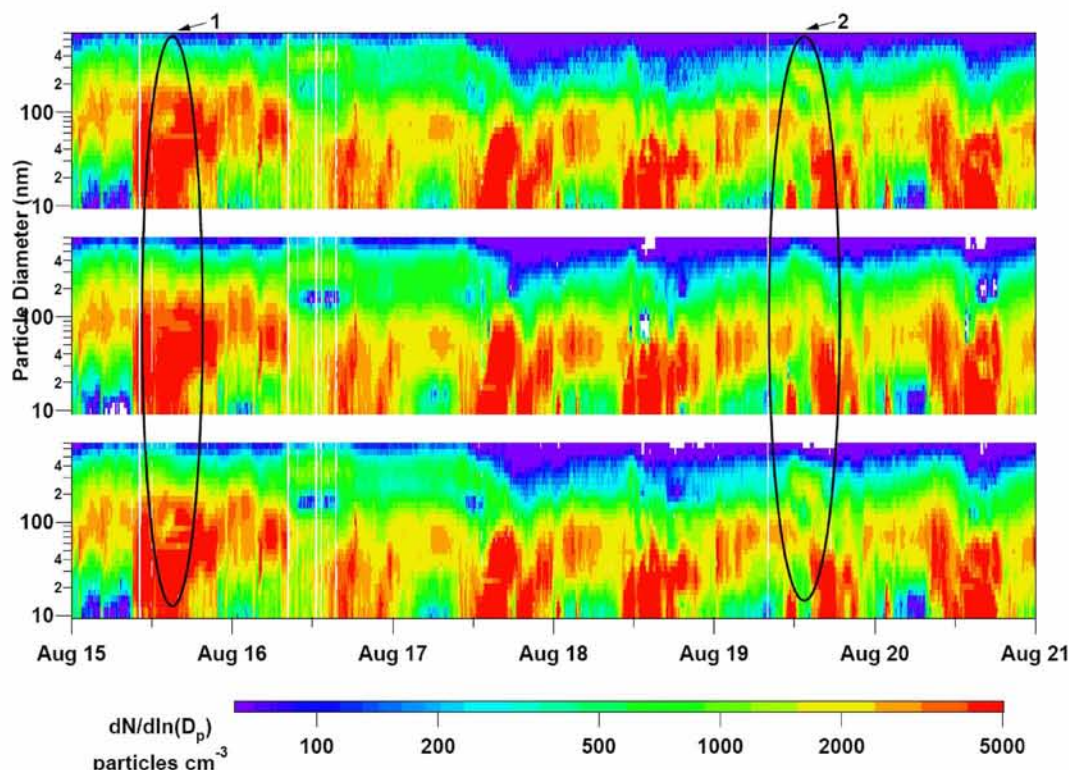


Fig. 4. Comparison of measured and fitted data for a six day segment of the Pacific 2001 data set. The components used were determined from the entire 17 day data set. The top panel shows the measurements, the middle panel shows fitted data using 5 components, and the bottom panel shows fitted data using 8 components. Circled are two intervals during which the 5 component fits show significant deviations from the measurements.

Table 1. Comparisons of the r.m.s. deviation between the sum of all component scores and the integrated DMA total number concentrations for maximum and minimum numbers of retained components.

Field study	Components	Deviation	Components	Deviation
Egbert 2003	5	2.5%	7	1.9%
Pacific 2001	5	1.0%	8	0.85%
Hamilton 2000	4	2.5%	5	1.9%
Simcoe 2000	6	2.0%	8	1.6%
Hamilton 1999	5	2.0%	6	0.75%

associated with specific sources at this site. In contrast, the Pacific 2001 sampling site was located at a considerable distance from a number of sources; this leads to greater atmospheric processing which may be the reason for the larger number of mixed components. In terms of the size distributions, this is seen as a greater variability in the locations and shapes of the various fine particle modes. Compared to the Pacific 2001 sampling site, the Simcoe 2000 data were obtained at a rural site that occasionally receives local pollution from Nanticoke but mostly experiences regional scale pollution, largely transported from the United States. Thus, air in Simcoe is also highly processed but not as variable as

that sampled during the Pacific 2001 study. The Egbert 2003 data were measured at a rural site that has a major nearby pollution source from Toronto as well as being impacted by regional scale pollution; it has an intermediate number of mixed components.

It does not appear that there is a direct connection between the individual components obtained by this method and any specific sources. However, we believe that this procedure will be extremely useful in simplifying the analysis of size distribution data since it enables a large number of size bins to be replaced with a much smaller number of components. At a minimum these components can be thought of as a way of “binning” the data that preserves maximum information. We find it remarkable that only 4 or 5 such “bins” are needed to reproduce most features of the size distributions and that just 6 to 8 components can preserve virtually all details of the distributions. This is made possible by the fact that the data themselves are used to determine the optimal “binning”. In analysing data, the scores may be treated as being analogous to the numbers of particles in various size ranges (such as nucleation, Aitken, and accumulation modes). However, using the principal components should be much preferred to using predefined size ranges since the components retain much more of the information present in the size distribution data.

For this reason, we believe that this procedure has the potential to greatly aid data analysis. As example of this is given in the following paper (Chan and Mozurkewich, 2007).

5 Conclusions

We have described how to apply absolute principal component analysis to atmospheric aerosol number size distribution measurements. This method provides a useful means to reduce the data dimensionality prior to analysis; DMA-CPC size distribution data with initially about 30 size bins can be accurately summarized using just a few components. One use of these components is as a way of “binning” the data that preserves maximum information. Only 4 or 5 components are needed to reproduce most features of the size distributions and just 6 to 8 components can preserve virtually all details of the distributions. As a result, this has the potential to greatly simplify data analysis.

In particular, we believe that this produces a simplified representation of size distribution data that is very advantageous in comparison with fitting multiple log-normal modes. Numerically, principal component analysis is extremely stable, so its application can be readily automated; this is not usually the case with fitting multiple modes. The number of time varying parameters needed to fit the distributions is typically fewer than for fitting multiple modes. Finally, the principal component results are fully continuous whereas the number of modes used in fits may vary with time.

We find that there are a number of steps that must be taken in order to successfully apply absolute principal component analysis to aerosol size distribution data. First, the weighting scheme used for social science data is often not appropriate for size distribution data. Therefore, the data mean should not be subtracted from the data prior to the analysis and the individual size bins should not be scaled according to their standard deviations. An appropriate data weighting is essential to produce realistic results. This can be accomplished by adopting the row and column weighting scheme of Cochran and Horne (1977). To make it possible to do this, we introduce a method of finding the row and column weights that give the best estimate to the actual individual data point weights derived from instrumental uncertainties. The weight for each row or column is the geometric mean of all weights in that row or column divided by the square root of the geometric mean of all the weights.

We have found that a modification of the widely used scree plot provides an effective method for determining the minimum and maximum number of components to retain; the exact number of components to retain depends on the user objectives. Application of the Varimax rotation to the retained component loadings and scores generates meaningful results. Each rotated component has a distinct maximum with low amplitude oscillations away from the peak. After removing the effect of weights, normalizing the rotated com-

ponents gives the corresponding component scores physical units of absolute concentrations. We believe that principal component analysis will be a useful method to simplify the representation of aerosol size distribution data and aid in the analysis of these data sets. However, some experience will be required to determine the best applications of the results. A first application of this is described in the accompanying paper (Chan and Mozurkewich, 2007).

Appendix A

Estimation of row and column weights

We begin with a set of weights, W_{ij} , for each individual data point, such as those calculated from Eq. (A1). To apply weights in principal component analysis, we need to factor the measured uncertainties into a set of row weights, X_i , and column weights, Y_j . The products of these generate a set of approximate weights, V_{ij} , given by

$$V_{ij} = X_i Y_j. \quad (\text{A1})$$

Since the row and column weights can not be determined directly, we find the row and column weights that provide the best estimate of the actual measured weights, W_{ij} . The optimum row and column weights are obtained by minimizing the sum of squares of the deviations, S_s^2 , between the logarithms of V_{ij} and W_{ij} ; which is given by Eq. (A2)

$$S_s^2 = \sum_i^n \sum_j^b [\ln W_{ij} - \ln (X_i Y_j)]^2. \quad (\text{A2})$$

Our objective is to minimize Eq. (A2) so that the ratios of V_{ij} to W_{ij} are as near as possible to unity. We choose to use percentage deviations over absolute deviations because W_{ij} vary over a wide range and we see no reason why the larger weights should be more accurately estimated than the smaller ones. Minimizing Eq. (A2) makes the percentage deviations independent of the magnitudes of the W_{ij} . To get the optimum row weights, we set the derivative of S_s^2 with respect to any one X_i equal to zero, this yields

$$0 = \sum_j^b [\ln W_{ij} - \ln (X_i Y_j)]. \quad (\text{A3})$$

Solving Eq. (A3) for the optimum row weight, X_i , yields

$$\ln X_i = \left(\frac{1}{b} \sum_{j=1}^b \ln W_{ij} \right) - \left(\frac{1}{b} \sum_{j=1}^b \ln Y_j \right). \quad (\text{A4})$$

The first term on the right hand side of Eq. (A4) is the logarithm of $\langle W_{ij} \rangle_i$, which we define as the geometric mean of the individual weights in row i . The last term in Eq. (A4) is the logarithm of the geometric mean of the column weights.

Equation (A4) indicates that the row weights should be proportional to $\langle W_{ij} \rangle_i$ due to the fact that all rows in any particular column have the same column weight. Also, since the V_{ij} should have the same geometric mean as the W_{ij} , we adjust the proportionality constant to $\langle W_{ij} \rangle^{1/2}$ where $\langle W_{ij} \rangle$ is the geometric mean of all values in W_{ij} . Then rearranging Eq. (A4) yields the optimum row weights:

$$X_i = \langle W_{ij} \rangle_i / \langle W_{ij} \rangle^{1/2}. \quad (\text{A5})$$

To obtain the optimum column weights, we differentiate Eq. (A2) with respect to the column weights, and using the same logic, we find that the optimum column weights to be

$$Y_j = \langle W_{ij} \rangle_j / \langle W_{ij} \rangle^{1/2} \quad (\text{A6})$$

where $\langle W_{ij} \rangle_j$ is the geometric mean of the individual weights in column j . Equations (A5) and (A6) are the same as Eqs. (3) and (4).

Acknowledgements. Funding for this research was provided by the Natural Science and Engineering Research Council of Canada and by the Canadian Foundation for Climate and Atmospheric Sciences.

Edited by: C. George

References

- Barthelmie, R. J. and Pryor, S. C.: Implications of Ammonia emissions for fine aerosol formation and visibility impairment – a case study from the Lower Fraser Valley, British Columbia, *Atmos. Environ.*, 32, 345–352, 1998.
- Birmili, W., Wiedensohler, A., Heintzenberg, J., and Lehmann, K.: Atmospheric particle number size distribution in central Europe: statistical relations to air masses and meteorology, *J. Geophys. Res.*, 106, 32 005–32 018, 2001.
- Buhamra, S. S., Bouhamra, W. S., and Elkilani, A. S.: Assessment of air quality in ninety-nine residences of Kuwait, *Environ. Technol.*, 19, 357–367, 1998.
- Cattell, R. B.: The Scree test for the number of factors, *J. Multiv. Behav. Res.*, 1, 245–276, 1966.
- Chan, T. W. and Mozurkewich, M.: Application of absolute principal component analysis to size distribution data: Identification of particle origins, *Atmos. Chem. Phys.*, 7, 887–897, 2007, <http://www.atmos-chem-phys.net/7/887/2007/>.
- Cheng, M. T. and Tsai, Y. I.: Characterization of visibility and atmospheric aerosols in urban, suburban, and remote areas, *The Science of the Total Environment*, 263, 101–114, 2000.
- Cochran, R. N. and Horne, F. H.: Statistically weighted principal component analysis of rapid scanning wavelength kinetics experiments, *Anal. Chem.*, 49, 846–853, 1977.
- Comrey, A. L. and Lee, H. B.: A first course in factor analysis (2nd edition), Lawrence Erlbaum Associates, Hillsdale, N. J., 1992.
- Ferre, L.: Selection of components in principal component analysis: A comparison of methods, *Comput. Stat. Data Anal.*, 19, 669–682, 1995.
- Hotelling, H.: Analysis of a complex of statistical variables into principal components, *J. Educ. Psychol.*, 24, 417–441, 1933.
- Jackson, J. E.: A user's guide to principal components, Wiley-Interscience, New York, 1991.
- Jolliffe, I. T.: Principal component analysis, Chapter 1. Springer-Verlag, New York, 1986.
- Keenan, M. R. and Kotula, P. G.: Accounting for Poisson noise in the multivariate analysis of ToF-SIMS spectrum images, *Surf. Interface Anal.*, 36, 203–212, 2004.
- Keiding, K., Sørensen, M. S., and Pind, N.: A receptor model for urban aerosols, based on oblique factor analysis, *Anal. Chim. Acta*, 193, 295–307, 1987.
- Knutson, E. O. and Whitby, K. T.: Aerosol classification by electric mobility: apparatus, theory, and applications, *J. Aerosol Sci.*, 6, 443–451, 1975.
- Li, S. M.: A Concerted effort to understand the ambient particulate matter in the Lower Fraser Valley: The Pacific 2001 air quality study, *Atmos. Environ.*, 38, 5717–5894, 2004.
- Mäkelä, J. M., Koponen, I. K., Aalto, P., and Kulmala, M.: One-year data of submicron size modes of tropospheric background aerosol in southern Finland, *J. Aerosol Sci.*, 31, 595–611, 2000.
- Maynard, A. D. and Maynard, R. L.: A derived association between ambient aerosol surface area and excess mortality using historic time series data, *Atmos. Environ.*, 36, 5561–5567, 2002.
- Mönkkönen, P., Koponen, I. K., Lehtinen, K. E. J., Hämeri, K., Uma, R., and Kulmala, M.: Measurements in a highly polluted Asian mega city: observations of aerosol number size distribution, modal parameters and nucleation events, *Atmos. Chem. Phys.*, 5, 57–66, 2005, <http://www.atmos-chem-phys.net/5/57/2005/>.
- Mozurkewich, M., Chan, T. W., Aklilu, Y. A., and Verheggen, B.: Aerosol particle size distributions in the Lower Fraser Valley: evidence for particle nucleation and growth, *Atmos. Chem. Phys.*, 4, 1047–1062, 2004, <http://www.atmos-chem-phys.net/4/1047/2004/>.
- Paatero, P. and Tapper, U.: Positive matrix factorization: a non-negative factor model with optimal utilization of error estimates of data values, *Environmetrics*, 5, 111–126, 1994.
- Peters, A., Skorkovsky, J., Kotesovec, F., Brynda, J., Spix, C., Wichmann, H. E., and Heinrich, J.: Associations between mortality and air pollution in Central Europe, *Environmental Health Perspectives*, 108, 283–287, 2000.
- Schwartz, S. E.: The whitehouse effect – shortwave radiative forcing of climate by anthropogenic aerosols: an overview, *J. Aerosol Sci.*, 27, 359–382, 1996.
- Schwela, D.: Exposure to environmental chemicals relevant for respiratory hypersensitivity: global aspects, *Toxicol. Lett.*, 86, 131–142, 1996.
- Spurny, K. R.: Aerosol chemistry and its environmental effects in Aerosol chemical processes in the environment, edited by: Spurny, K. R., pp. 3–21, Lewis Publishers, New York, 2000.
- Thurston, G. D. and Spengler, J. D.: A quantitative assessment of source contributions to inhalable particulate matter pollution in metropolitan Boston, *Atmos. Environ.*, 19, 9–25, 1985.
- Twomey, S.: Aerosols, clouds and radiation, *Atmos. Environ.*, 25A, 2435–2442, 1991.
- Wang, S. C. and Flagan, R. C.: Scanning Electrical Mobility Spectrometer, *Aerosol Sci. Technol.*, 13, 230–240, 1990.
- Whitby, K. T.: The physical characteristics of sulfur aerosol, *Atmos. Environ.*, 12, 135–159, 1978.