BNL-81483-2008-JA

Joint Cluster and Non-Negative Least Squares Analysis for Aerosol Mass Spectrum Data

Tianyi Zhang[1], Wei Zhu[1] and Robert McGraw[2]

[1]Department of Applied Mathematics and Statistics, Stony Brook University, Stony Brook, NY 11794-3600
[2]Environmental Sciences Department, Brookhaven National Laboratory, Upton, NY 11973-5000

E-mail: zhu@ams.sunysb.edu

Abstract. Aerosol mass spectrum (AMS) data contain hundreds of mass to charge ratios and their corresponding intensities from air collected through the mass spectrometer. The observations are usually taken sequentially in time to monitor the air composition, quality and temporal change in an area of interest. An important goal of AMS data analysis is to reduce the dimensionality of the original data yielding a small set of representing tracers for various atmospheric and climatic models. In this work, we present an approach to jointly apply the cluster analysis and the non-negative least squares method towards this goal. Application to a relevant study demonstrates the effectiveness of this new approach. Comparisons are made to other relevant multivariate statistical techniques including the principal component analysis and the positive matrix factorization method, and guidelines are provided.

1. Introduction

Atmospheric aerosols are significantly related to environmental issues such as rainfall, air pollution, and climate change. In recent years, quantitative aerosol studies based on mass spectrometry data have evolved rapidly in such field as particle formation research and air pollution analysis (see, e.g., [1]-[6]). Another area of great interest is dimension reduction. The goal is to obtain, from the original high dimensional aerosol mass spectrum (AMS) data, a much smaller set of representative tracers that will elucidate the source and interaction of different aerosol components, and to be plugged in as part of the model variables, to the next generation of climate models.

Several multivariate analysis techniques have been applied individually towards the dimension reduction of AMS data. These include principal component analysis [7] and cluster analysis [8]. The most widely used technique in the atmospheric research community however, is Positive Matrix Factorization (PMF), also known as the non-negative matrix factorization [9]. It has been developed [10] to yield factors with non-negative coefficients which would be more intuitive and interpretable than factors with a mixture of positive and negative coefficients. Plenty of work using PMF have been done [10-14] with two review papers published recently [15, 16]. However, a number of undesirable features inherent in PMF including subjectivity and non-uniqueness [17] have made it clear that better method must be developed to better serve the climate research communities.

In this paper, we propose a joint cluster analysis on variables (VARCLUS) [18-20] and non-negative least squares (NNLS) [21] approach to better achieve the goal of dimension reduction. Each method alone is not new; however, their combination is novel and as shown later, is rather effective in solving the problem at hand. The rest of the paper is arranged as follows. Section 2 provides a brief introduction to the AMS data from Milagro. Section 3 describes the proposed method in details while Section 4 presents the analysis of the Milagro data using the new method. Finally, Section 5 concludes with a comparison between the proposed method and the existing methods, especially the PMF.

2. Aerosol Mass Spectrum Data

2.1 Aerodyne Mass Spectrometer

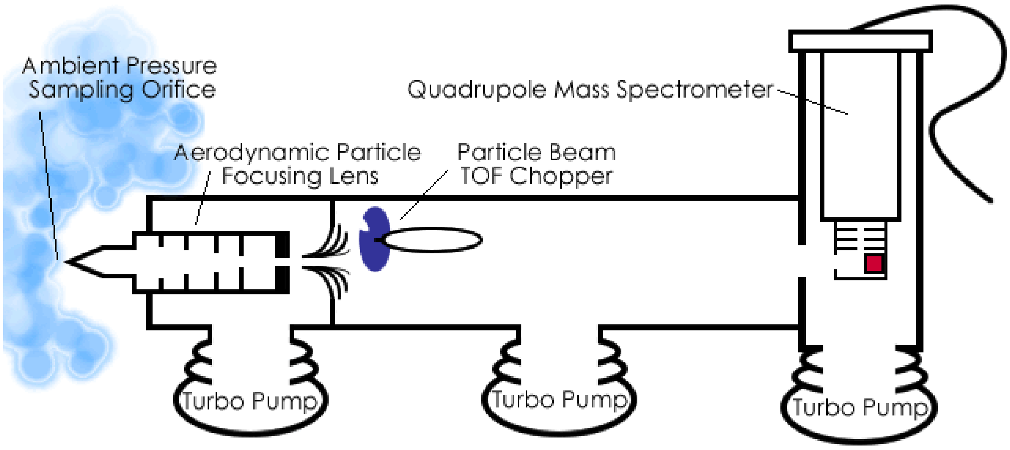

Figure 1. Basic schematic of the Aerodyne aerosol mass spectrometer.
(http://cires.colorado.edu/~jjose/ams.html)

The aerosol mass spectrum data analyzed in this work are recorded by the Aerodyne mass spectrometer. Its basic mechanism is illustrated in Figure 1. In summary, the air flows into the orifice. Particles are then focused in the focusing lens and passed through the chopper which is set to either an open or a closed mode. Vaporized and ionized, particles are introduced into the spectrometer sequentially with their weights and frequencies recorded. The instrument alternates in the MS mode and the TOF mode. During the MS mode, the chopper alternates between the open and blocked position every 5 seconds and the particle rates are record, while during the TOF mode mass spectrometer is set to a singe m/z and sampled at a user defined rate. In this work, only data from the MS mode are involved. A full description of the Aerodyne mass spectrometer can be found in [1].

2.2 Data Preprocessing

The data we analyzed were collected on March 19, 2006 from Milagro containing signals with mass to charge ratios (m/z) ranging from 1 to 452 measured every 12 seconds for a total of 943 measurements in time. Two modes of data were obtained -- data recorded when the aerosol beam was unobstructed (MSSopen) and data collected when the beam was blocked by the chopper (MSSClosed). Each mode also has its own baselines, designated as MSSOpenBaseL and MSSClosedBaseL. We used the difference between the open mode and the closed mode to remove the contribution from background gas in the detector. Thus the data file we analyzed in this paper is named MSSDiff and defined as:

$$\text{MSSDiff} = (\text{MSSOpen - MSSOpenBaseL}) - (\text{MSSClosed - MSSClosedBaseL})$$

Data in MSSDiff contain negative values, which are expected for signals that are close to zero. Since MSSDiff contains negative signals, it does not satisfy the requirement to apply PMF directly. Although replacing all negative values with zeros would be a viable approach, it will nonetheless destroy the linearity of the AMS data. Since the proposed joint cluster and NNLS approach would allow negative values, we are able to retain all negative values for the analysis.

Our analysis is focused on the subset organic aerosols ranging in m/z values from 1 to 100. Twenty-eight inorganic aerosols in the range (m/z's 1-11, 14, 16-18, 20-23, 28, 32-36, 39-40, and 47) were removed by setting the corresponding intensities to 0. The dataset was further normalized by dividing by the standard deviation at each m/z value. In addition, two time steps (time = 353, 643 seconds) appeared as outliers and were removed. The processed data are shown in Figure 2.

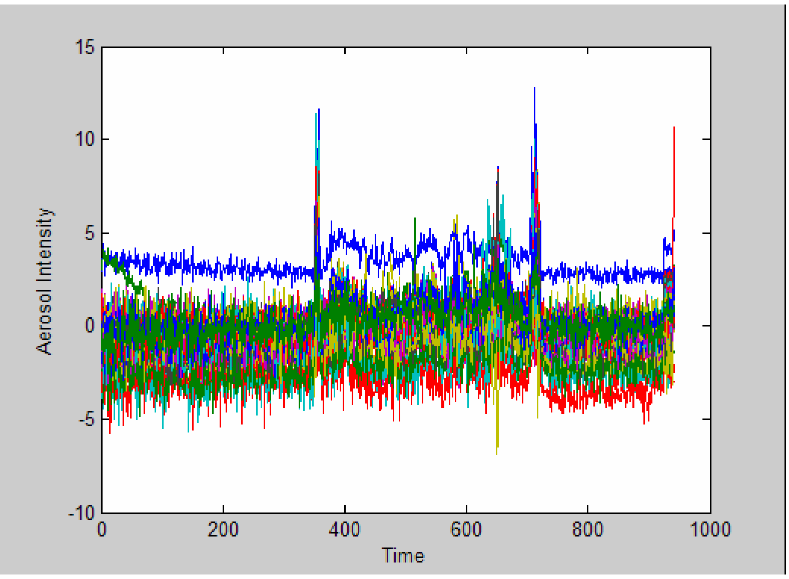

Figure 2. Pre-processed organic aerosol data where each colour represents the time series of a unique m/z value.

3. Method

3.1 Notations and Overview

Let **ORG** represent the pre-processed AMS data with Ni columns for m/z values ranging from 1 to Ni, and Nt rows for Nt temporal measurements. Thus the ith column of the matrix **ORG** is the time series corresponding to m/z value i, and the tth row of **ORG** is the mass spectrum at time point t. Our goal is to find a mass spectrum basis, **MS**, with lower dimension of Nc, such that

$$\mathbf{ORG}=\mathbf{C}*\mathbf{MS}+\mathbf{E} \qquad (1)$$

Here, **ORG** is an Nt*Ni matrix (in our work Nt=941, Ni=100). **MS** is an Nc*Ni matrix, Nc<<Nt. Our goal includes two parts: first, we need to decide Nc, the dimension of the basis; next, we need to estimate **MS**.

The above decomposition can be solved by the principal component analysis (PCA), factor analysis (FA) or positive matrix factorization (PMF). PCA and FA could yield a suggestive Nc based on the variation explained by the basis. However, there are two apparent disadvantages to these methods. First, it is hard to interpret the output basis; second, the output basis usually contains negative values that are often unacceptable for practical purposes. The PMF method could avoid the second problem but not the first one. Furthermore, PMF returns different output using different starting matrix. That is, it is not unique [9, 17, 22].

To resolve these issues, we propose a combined cluster analysis on variables (VARCLUS), PCA and non-negative least square (NNLS) approach to better achieve the goal of dimension reduction. In summary, first we use VARCLUS to determine the dimension Nc. Next we perform the PCA to obtain the matrix **C**. Finally we apply the NNLS to estimate the non-negative matrix **MS**. More details are given below.

3.2 VARCLUS and PCA

The goal of VARCLUS is to divide the m/z values into disjoint clusters based on their correlations. The inherent linearity of AMS data usually returns explainable clusters with m/z values belonging to the same aerosol chemical class grouped together. Therefore the number of major aerosol classes can be estimated by the number of major clusters. Since the first principal component for each cluster is a weighted linear combination of all m/z values in the given cluster, with non-negative coefficients, and would explain the most (and often the majority) variation in the given cluster, it is the natural choice of the representing tracer for each cluster. Thus the dimension of basis Nc equals to the number of major clusters. And the basis consists of the first principal components from each major cluster. The matrix C is thus determined.

The VARCLUS procedure follows the following steps. [23, 24]

1. We treat the m/z values as variables, and different time points as observations. The data set is normalized by dividing by the standard deviation at each m/z value.
2. A cluster is chosen for splitting, for which the percentage of variation explained by its cluster component is the smallest. (The whole dataset is in the same cluster at the very beginning.)
3. The chosen cluster is split into two clusters by finding the first two principal components, performing an orthoblique rotation, and assigning each variable to the rotated component with which it has the higher squared correlation.
4. Variables are iteratively reassigned to clusters to maximize the variance accounted for by the cluster components.

Steps 2-4 are repeated until the stop criterion (total percentage of variation accounted for) is satisfied. A hierarchical tree is produced as the final output.

In each cluster, we use the first principal component (PC) to represent the whole cluster. Each PC forms a column of the matrix **C**. The PC's from all major clusters thus form the basis (the tracer set). We denote columns of **ORG** as $\mathbf{MZ_1,\ldots,MZ_{100}}$ (each of length 941), while rows of **ORG** as $\mathbf{T_1,\ldots T_{941}}$.(each of length 100). Thus, the i^th^ column of **C** is $C_i = \sum_{k \in Cluster\#i} a_k MZ_k$ **.** VARCLUS produces disjoint clusters so that each $\mathbf{C_i}$ is a weighted summation of disjoint subset of the m/z's (**$\mathbf{MZ_k}$'s)**. For practical reasons, clusters that would explain a small amount of variation in the data are either merged to its nearest major clusters, or simply discarded.

3.3 Non-Negative Least Squares

Now that we have obtained the basis dimension Nc and the basis matrix C, our next goal is to find the matrix MS in

$$\underset{941*100}{ORG} = \underset{941*Nc}{C} * \underset{Nc*100}{MS} + E$$

such that we can express each original m/z value as a linear combination of the basis with non-negative coefficients. This will further elucidate the relations between the original AMS data and the newly obtained tracer set. In other words, we need to calculate the non-negative coefficients β in the following equation system:

$$\begin{aligned} MZ_1 &= \beta_{1,1}C_1 + \beta_{1,2}C_2 + \ldots + \beta_{1,Nc}C_{Nc} + e_1 \\ MZ_2 &= \beta_{2,1}C_1 + \beta_{2,2}C_2 + \ldots + \beta_{2,Nc}C_{Nc} + e_2 \\ &\ldots\ldots \\ MZ_{100} &= \beta_{100,1}C_1 + \beta_{100,2}C_2 + \ldots + \beta_{100,Nc}C_{Nc} + e_{100} \end{aligned}$$

This is achieved through the NNLS algorithm by minimizing $\| ORG - \widehat{ORG} \|^2$, with the constraints that each β is non-negative. The details of the NNLS algorithm can be found in [21]. For better fit, we add constant terms in the above linear equation system as follows.

$$\begin{aligned}
&MZ_1 = \beta_{1,0} + \beta_{1,1}C_1 + \beta_{1,2}C_2 + \ldots + \beta_{1,Nc}C_{Nc} + e_1 \\
&MZ_2 = \beta_{2,0} + \beta_{2,1}C_1 + \beta_{2,2}C_2 + \ldots + \beta_{2,Nc}C_{Nc} + e_2 \qquad (2) \\
&\ldots\ldots \\
&MZ_{100} = \beta_{100,0} + \beta_{100,1}C_1 + \beta_{100,2}C_2 + \ldots + \beta_{100,Nc}C_{Nc} + e_{100}
\end{aligned}$$

In matrix form we have: $\underset{941*100}{ORG} = \underset{941*(Nc+1)}{\widetilde{C}} * \underset{(Nc+1)*100}{\widetilde{MS}}$, where $\widetilde{C}$ is **C** with an extra column of 1, and $\widetilde{MS}$ is **MS** with the added row $(\beta_{1,0}, \beta_{2,0} \ldots, \beta_{100,0})$. Notice that $(\beta_{1,0}, \beta_{2,0} \ldots, \beta_{100,0})$ is allowed to be negative. We used the NNLS command in MATLAB to estimate β.

4. Results

4.1 VARCLUS and NNLS output

The VARCLUS output, a hierarchical tree, is shown in Figure 3. Based on the output tree and the related aerosol information, we have, clearly, 4 major clusters.

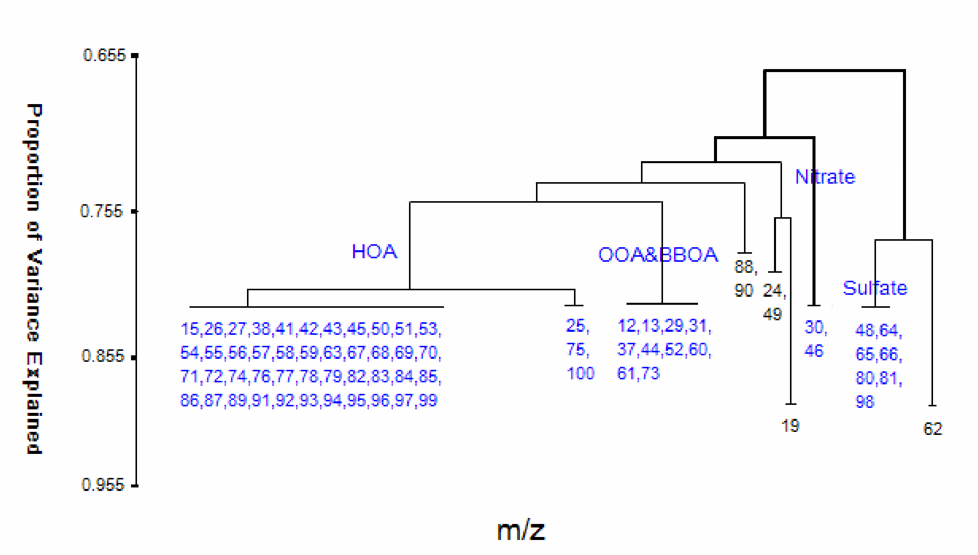

Figure 3. VARCLUS output: the hierarchical tree. From the tree structural, we decided four major clusters. Numbers in blue indicate the m/z values of cluster members while numbers in black are m/z values not belonging to any major cluster. The proportion of variance explained at each step of the binary tree split is displayed along the vertical axis.

Table 1 contains detailed information on these four major clusters. As discussed in Section 3.2, the advantage of using VARCLUS is that the clustering results are usually explainable if there is inherent linearity in the data. The AMS data are inherently linear with organic aerosols usually coming from several main sources. For example, hydrocarbon-like organic aerosols (HOA) mostly come from fossil fuel, while oxygenated organic aerosols (OOA) mostly come from secondary organic aerosols (SOA) [7]. This property supports the VARCLUS data assumptions and in our output we can clearly identify these 4 clusters. All 1st PCs in these clusters represent a high percentage of variation explained with non-negative coefficients. Thus, it is reasonable to use these four 1st PCs as the basis (tracer set) for the given AMS data.

Table 1. VARCLUS and PCA output.
Total variation explained by the basis consisting of four PC's is **71.14%**.

Cluster	**Members (m/z values)**	**Variation Explained**	**1st PC Variation in Each Cluster**	**Label**
Cluster#1	**15,25:27,38,41:43,45,50 51,53:59,63,67:72,74:79, 82:87,89,91:97,99,100**	**49.72%**	**76.16%**	**HOA**
Cluster#2	**48,64,65,66,80,81,98**	**9.15%**	**94.16%**	**Sulfate**
Cluster#3	**12,13,29,31,37,44, 52,60,61,73**	**9.50%**	**68.36%**	**OOA and BBOA**
Cluster#4	**30, 46**	**2.77%**	**99.68%**	**Nitrate**

Based on the VARCLUS output, we obtained the matrix **C** mentioned above. The matrix **MS** was estimated using NNLS. The output is shown in Figure 4.

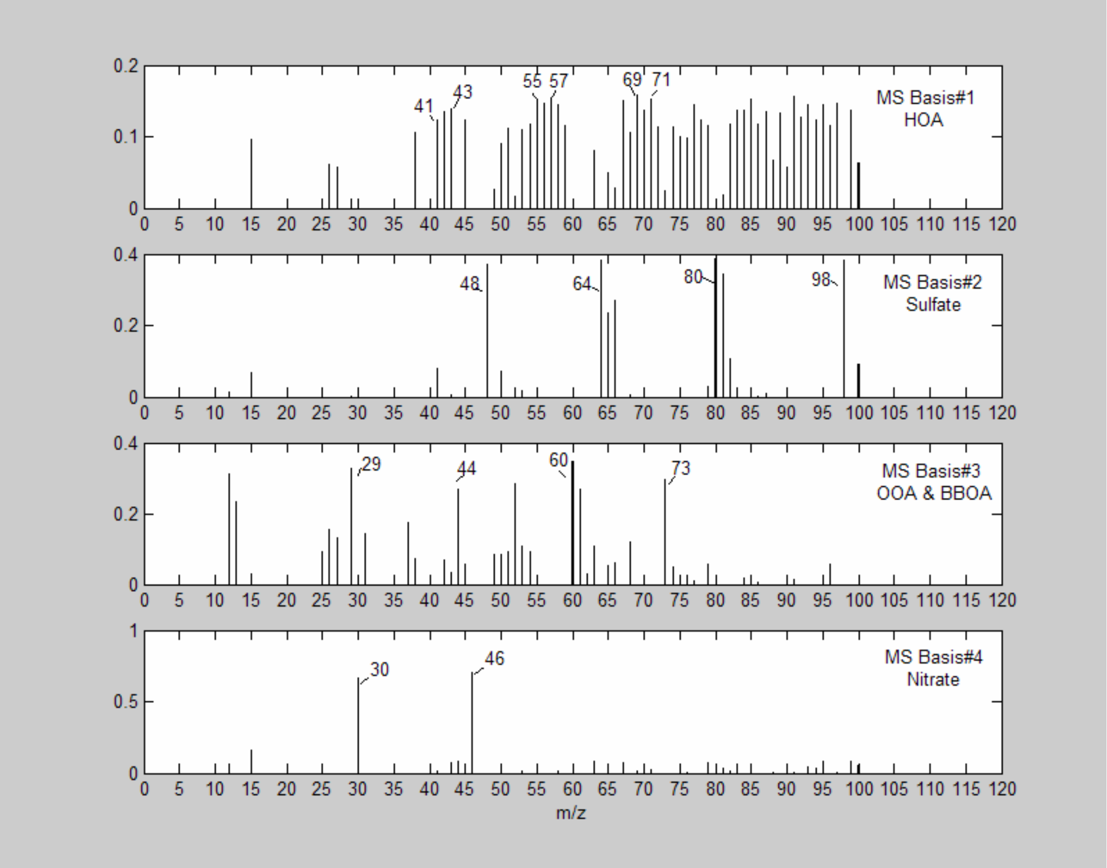

Figure 4. The basis of four 1st PC's are linear combinations of individual m/z values within each cluster. Significant peaks on each 1st PC are labelled.

4.2 Validity of Tracers

We checked the validity of the four tracers from three aspects: (a) whether they are interpretable for the given aerosol study; (b) how much of the total variation they could explain; (c) whether they could represent most of the aerosols in the original AMS data.

The organic aerosol factors from theoretic analysis are listed in Table 2 to check for (a). Compared with Table 1, the VARCLUS and PCA outputs are consistent with these factors for each major cluster containing all the corresponding critical m/z's correctly.

For (b), we found that 71.14% of the total variation is explained by the tracers as shown in Table 1. Figure 4 shows the time series of the four tracers. Compared to the single m/z tracer used by Zhang et al (2005), we found that the new tracers have larger variability as illustrated in Figure 5.

Table 2. Organic aerosol factors. The emphasized m/z's are important signals for the given factor.

Organic aerosol factors	Ionized at 600C	m/z	O: C ratio
HOA hydrocarbon-like organic aerosols	$C_nH_m \rightarrow C_{n-x}H_{m-y}^{+}$	27,29,**41,43,55,57**,69,71,…	Less Oxidized ↓ More Oxidized
OOA2 Oxygenated organic aerosols type II	$C_nH_mO \rightarrow C_2H_3O^+, C_3H_3O^+, R'^+$	**43,55**,…	
BBOA biomass burning organic aerosols	$R \rightarrow R'^+, C_2H_4O_2^+, C_3H_3O_2^+$	**44**,45,…	
OOA1 Oxygenated organic aerosols type I	$C_nH_mO_2 \rightarrow CO_2^+, HCO_2^+, R'^+$	**60,73**,…	

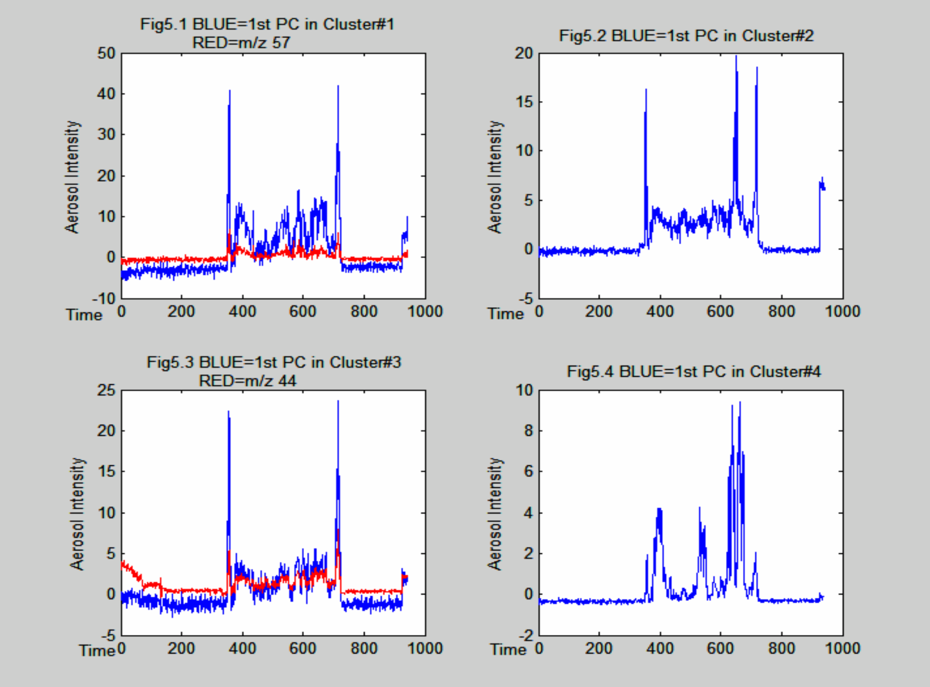

Figure 5. Time series of 1st PCs in clusters. Comparison between 1st PC in cluster#1 (HOA) and m/z 57 is shown in Fig5.1; comparison between 1st PC in Cluster#3 and m/z 44 is shown in Fig5.3. The single m/z values of m/z 44 and m/z 57 were used as tracers for OOA and HOA in Zhang et al (2005).

To check for (c), we need to verify that most of the peaks in the aerosol mass spectrum could be represented by our tracers. We first used the ordinary least squares regression (OLS) for this purpose. For each m/z=i, we fit the equation $MZ_i = \alpha_{i,0} + \alpha_{i,1} * C_1 + \alpha_{i,2} * C_2 + \alpha_{i,3} * C_3 + \alpha_{i,4} * C_4 + e_i$, where C_j represents the 1st PC in the jth major cluster and i=1, 2,…, 100. The goodness of fit for each regression, as evaluated by the coefficient of determinant R^2 for each regression equation, is reported in Figure 6. We found that most R^2 are bigger than 0.8 (the mean of R^2 is 0.81, the standard deviation is 0.23). For the few m/z's with small R^2 such as m/z19, it is because they are not in any major cluster and thus their signals can not be expected to be explained by the tracers. Lastly, stepwise regression was performed. The results support that these four tracers are not reducible.

We therefore conclude that the given tracers are highly satisfactory gauged by the three criteria (a), (b) and (c).

4.3 Error Analysis

To further validate our analysis, especially the NNLS portion of the analysis, we performed an error analysis by examining and comparing the coefficient of determination (R^2) from both OLS and NNLS.

Let $SSE_i = \sum_{j=1}^{941} (ORG_{ji} - \widehat{ORG_{ji}})^2$ be the sum of squared errors (SSE) for m/z=i (the i-th column in **ORG**), and $SST_i = \sum_{j=1}^{941} (ORG_{ji} - \overline{ORG_{ji}})^2$ be the corresponding total sum of squares (SST). The coefficient of determination (at m/z=i) is defined as [25]:

$$R_i^2 = 1 - \frac{SSE_i}{SST_i}$$

R^2 for each m/z using OLS or NNLS were calculated to examine the regression goodness-of-fit. We can see OLS and NNLS give similar R^2 for most m/z's in the given example. R^2=0.36, which corresponds to a multiple correlation of 60%, was used as a threshold to determine outliers. Six outliers were found: they're m/z 19, 24, 49, 62, 88, 90. The reason causing these outliers is quite clear. These are the six m/z's that are not assigned to any of the four major clusters as shown in Figure 7. Hence, all the other m/z's could be well represented by the four tracers (first PC's of the 4 major clusters) with non-negative constraints except for these outsiders.

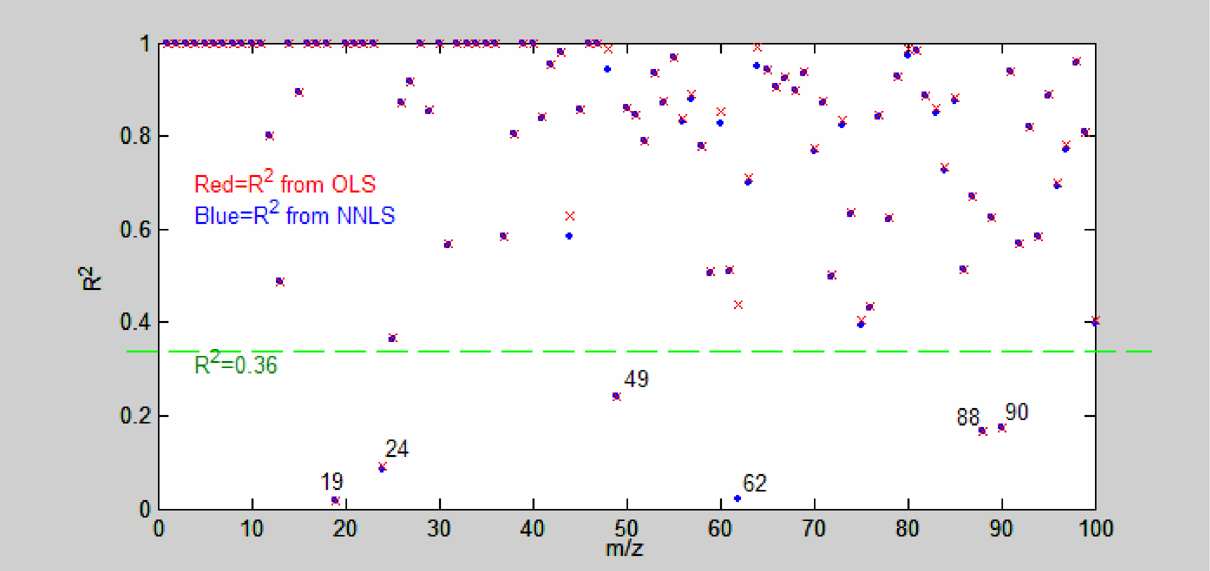

Figure 6. The coefficient of determination (R^2) for each m/z using OLS or NNLS. Outliers are specified.

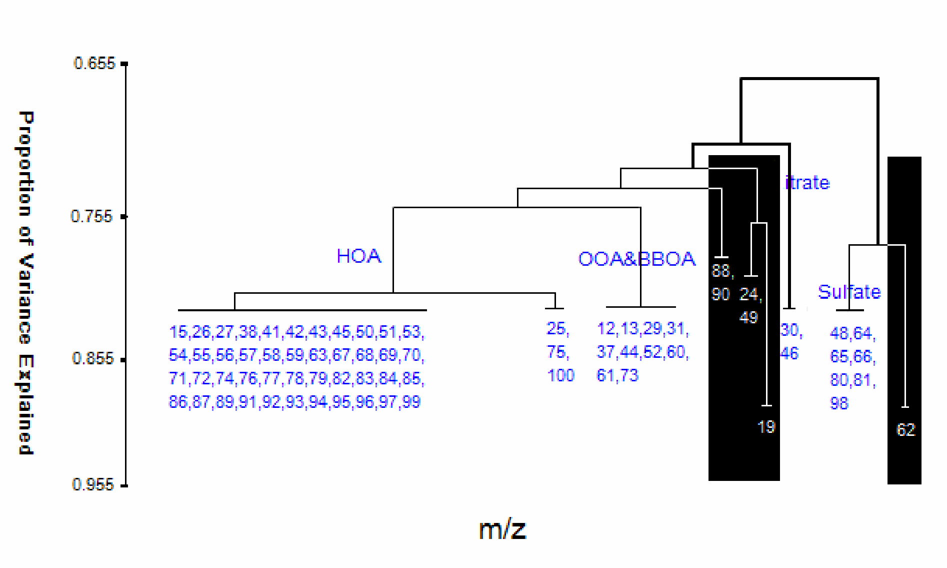

Figure 7. Outliers are displayed in the hierarchical clustering tree.

5. Discussion

Many commonly used multivariate statistical methods are not suitable in the intended AMS data analysis of dimension reduction and tracer extraction due to negative coefficients of the resulting basis. Both the PMF and the proposed method of joint cluster and NNLS analysis can achieve non-negative tracer coefficients. The major draw-back of the PMF lies in its non-uniqueness and subjectivity.

Generally speaking, PMF is not unique. For example, if X=ABC, where A, B and C are all positive matrix, then we have at least two different PMF's in X=(AB)*C=A*(BC). Different factorizations are called "rotations" in factor analysis. For the proposed joint analysis method, there is little flexibility in deciding the major clusters and thus the basis. Subsequently, given the fixed **C** matrix, the **MS** matrix determined by the NNLS procedure would be unique.

As P. Paatero (See [17]) pointed out "It is unfortunate that introducing a priori information also introduces some subjectivity in the analysis." To apply the PMF analysis, one needs to first decide upon at least two things: the number of factors and the "rotation" parameter. Otherwise one will find the solution non-unique. Both of these two choices have determinant effect on the final output. For our method, we still have to refer to some prior knowledge from the field. However, the "subjectivity" exists in only one step – the determination of major clusters based on the VARCLUS output. Furthermore, even for this step, our decision is based mainly on objective criterion such as the percent variation explained by each cluster as shown in the given study.

There is space yet for us to improve upon in the proposed method. First, since the sources of organic aerosols are complicated, the disjoint clustering method may not perform well for overlapping aerosol groups. Secondly, if the first principal component for a given cluster could only explain a modest amount of variation, one would need to find additional tracer(s) for the given cluster. Perhaps a combined VARCLUS and PMF approach with the PMF done within each cluster would better serve our purposes. Further research is warranted in this area.

References

[1] Jimenez J L, *et al.* 2003 Ambient aerosol sampling using the Aerodyne aerosol mass spectrometer *J Geophysical Research* **108**(D7):8425

[2] Allan J D, *et al.* 2003 Quantitative sampling using an Aerodyne aerosol mass spectrometer, 1: Techniques of data interpretation and error analysis *J. Geophysical Research* **108**(D3) 4090

[3] Allan J D, *et al.* 2003 Quantitative sampling using an Aerodyne aerosol mass spectrometer, 2: Measurements of fine particulate chemical composition in two U.K. cities *J. Geophysical Research* **108**(D3) 4091

[4] Zhang Q, *et al.* 2005 Deconvolution and quantification of hydrocarbon-like and oxygenated organic aerosols based on aerosol mass spectrometry *Environmental Science & Tech***.** **39:**4938–4952

[5] Zhang Q, *et al.* 2005 Time- and size-resolved chemical composition of submicron particles in Pittsburgh: Implications for aerosol sources and processes *J. Geophysical Research* **110:**D07S09

[6] DeCarlo P F, *et al.* 2006 A field-deployable high-resolution time-of-flight aerosol mass spectrometer *Analytical Chemistry* **78:**8281–8289

[7] Zhang Q, *et al.* 2007 Ubiquity and dominance of oxygenated species in organic aerosols in anthropogenically-influenced Northern Hemisphere midlatitudes *Geophysical Research Letters* **34:**L13801

[8] Marcolli C, *et al.* 2006 Cluster analysis of the organic peaks in bulk mass spectra obtained during the 2002 New England air quality study with an Aerodyne aerosol mass spectrometer *Atmos. Chem. Phys.* **6:**5649–5666

[9] Lee D D and Seung H S 1999 Learning the parts of objects by non-negative matrix factorization *Nature* **401**(6755): 788–791

[10] Paatero P and Tapper U 1994 Positive matrix factorization: A non-negative factor model with optimal utilization of error estimates of data values *Environmetrics* **5:**111–126

[11] Lee E, *et al.* 1999 Application of positive matrix factorization in source apportionment of particulate pollutants in Hong Kong *Atmos. Environ.* **33:**3201–3212

[12] Ramadan Z, *et al.* 2000 Identification of sources of Phoenix aerosol by positive matrix factorization *J. Air Waste Manag. Assoc.* **50:**1308–1320

[13] Larsen R K and Baker J E 2003 Source apportionment of polycyclic aromatic hydrocarbons in the urban atmosphere: A comparison of three methods *Environ. Sci. Technol.,* **37:**1873–1881

[14] Maykut N N, *et al.* 2003 Source apportionment of PM2.5 at an urban improve site in Seattle, Washington *Environ. Sci. Technol.* **37:**5135–5142

[15] Engel-Cox J and Weber S A 2007 Compilation and assessment of recent positive matrix factorization and UNMIX receptor model studies on fine particulate matter source apportionment for the eastern United States *J. Air Waste Manag. Assoc.* **57:**1307–1316

[16] Reff A, *et al.* 2007 Receptor modeling of ambient particulate matter data using positive matrix factorization: Review of existing methods *J. Air Waste Manag. Assoc.* **57:**146–154

[17] Ulbrich I M, *et al.* 2008 Interpretation of organic components from positive matrix factorization of aerosol mass spectrometric data *Atmos. Chem. Phys. Discuss.* **8:**6729-6791

[18] Harman H H 1976 *Modern Factor Analysis* 3rd ed (University of Chicago Press)

[19] Cattell R B 1965 Factor analysis: An introduction to essentials, II: The role of factor analysis in research *Biometrics* **21**(2):405–435

[20] Rummel R J 1970 *Applied Factor Analysis* (Northewestern Univ. Press: Evanston, Illinois)

[21] Lawson C L and Hanson R J 1974 *Solving Least Squares Problems* (Prentice-Hall) chapter 23

[22] Lee D D and Seung H S 2001 Algorithms for non-negative matrix factorization *Advances in Neural Information Processing Systems 13: Proc. 2000 Conference,* 556–562

[23] SAS Institute Inc 1994 *SAS/STAT User's Guide* Version 6, 4th ed

[24] Sun S 2000 *Multivariate Statistical Analysis and Statistical Software* (Beijing Medical University Press)
[25] http://en.wikipedia.org/wiki/Coefficient_of_determination.