Statistical Approaches to Aerosol Dynamics for Climate Simulation

Final Technical Report

DOE Award Number: DE-FC02-07ER25817

Wei Zhu (PI), Ph.D. Professor & Deputy Chair

The Research Foundation of State University of New York

Department of Applied Mathematics and Statistics

Stony Brook University

Stony Brook, NY 11794-3600

Wei.Zhu@StonyBrook.edu

Part I.

Compound Regression and Constrained Regression: Nonparametric Regression Frameworks for EIV Models

Ling LENG, Wei ZHU

Ling Leng is Statistical Engineer, Amazon.com Inc., Seattle, WA 98144-2734 (Email: lingleng@amazon.com). Wei Zhu is Professor of Statistics, Department of Applied Mathematics and Statistics, State University of New York at Stony Brook, Stony Brook, NY 11794-3600 (Email: weizhu@notes.cc.sunysb.edu). This work was sponsored by the U.S. Department of Energy Grant ER25817 on Climate Modeling.

Abstract

In this work, we introduce two general non-parametric regression analysis methods for errors-in-variable (EIV) models: the compound regression, and the constrained regression. It is shown that these approaches are equivalent to each other and, to the general parametric structural modeling approach. The advantages of these methods lie in their intuitive geometric representations, their distribution free nature, and their ability to offer a practical solution when the ratio of the error variances is unknown. Each includes the classic non-parametric regression methods of ordinary least squares, geometric mean regression, and orthogonal regression as special cases. Both methods can be readily generalized to multiple linear regression with two or more random regressors.

Key Words: Compound regression; Constrained regression; Geometric mean regression; Maximum likelihood method; Ordinary least squares regression; Orthogonal regression.

1. INTRODUCTION

As the renowned physicist E.T. Jaynes pointed out in his celebrated monogram “the most common problem of inference faced by experimental scientists: linear regression with both variables subject to unknown error” (Jaynes 2004, pg 497). We readily agree to this observation as the errors-in-variable (EIV) modeling problem arises in gauging the relationships between two random variables which is indispensible in any research or business practice. For example, with the rapid development of gene measurement platforms, an urgent task is to calibrate between the fading gene microarray platform and the incoming RNA sequencing (RNAseq) technology to ascertain the validity of current disease biomarkers. Another example that motivated our research here came from a climate modeling project in collaboration with scientists from the Brookhaven National Laboratory. In gauging the relationship between the concentrations of organic aerosols and anthropogenic carbon monoxide (CO) (Kleinman et al. 2007), both variables, measured by the mass spectrometer and the UV fluorescence analyzer respectively, contain measurement errors and other volatilities due to air dynamics. Two commonly used non-parametric regression methods for EIV models, the orthogonal regression and the geometric mean regression, yielded different regression equations as expected. Indeed, these are not the only possible solutions. If one can assume the often unattainable bivariate normal distribution for the variables, one can apply the general parametric structural modeling approach for EIV models that will yield an infinite class of regression lines with the optimal choice depends on the ratio of the error variances, which is usually unknown (Lindley 1947; Wong 1989).

To overcome these dilemmas, we have developed two general, and equivalent, nonparametric regression approaches entitled the compound regression and the constrained regression that will provide intuitive and practical solutions for all EIV modeling problems for simple linear regression analysis including our own. These new methods are introduced in Section 3 and illustrated through two examples in Section 4, following a brief review of the current EIV modeling methods.

(discuss unknown lambda)

2. EXISTING METHODS for EIV Models

The general parametric EIV structural model for a simple linear regression model is as follows (Sprent 1969; Wong 1989):

$$\begin{aligned} X &= \xi + \delta & \qquad \delta &\sim N\left(0 \sigma_\delta^2\right) \\ Y &= \eta + \varepsilon & \qquad \varepsilon &\sim N\left(0 \sigma_\varepsilon^2\right) \\ \eta &= \beta_0 + \beta_1 \xi & & \end{aligned}$$

Here δ and ε are independent random errors. Furthermore, ξ is a random variable following normal distribution with mean μ and varianceτ^2, and independent to both random errors. This implies that X and Y follow a bivariate normal distribution:

$$\begin{pmatrix} X \\ Y \end{pmatrix} \sim N\left(\begin{pmatrix} \mu \\ \beta_0 + \beta_1 \mu \end{pmatrix}, \begin{pmatrix} \tau^2 + \sigma_\delta^2 & \beta_1 \tau^2 \\ \beta_1 \tau^2 & \beta_1^2 \tau^2 + \sigma_\varepsilon^2 \end{pmatrix} \right)$$

Given a random sample of observed X's and Y's, the maximum likelihood estimator (MLE) of the regression slope is given by

$$\hat{\beta}_1 = \frac{S_{YY} - \lambda S_{XX} + \sqrt{(S_{YY} - \lambda S_{XX})^2 + 4\lambda S_{XY}^2}}{2S_{XY}}$$

Its value depends on the ratio of the two error variances $\lambda = \sigma_{\varepsilon}^2 / \sigma_{\delta}^2$, which is generally unknown and unable to be estimated from the data alone (Lindley 1947).

It has been shown that the ordinary least squares regressions (OLS) and the two most commonly used nonparametric regression methods when both X and Y are random, the orthogonal regression (OR) and the geometric mean regression (GMR), can be considered special cases in this structural model approach, with the distinction that these specific methods do not reply on the bivariate normal assumption.

The OLS slope estimator with Y or X as the dependent variable will minimize the squared vertical or horizontal distances from the points to the regression line, and corresponds to the MLE of the slope in the structural model approach when $\lambda = \infty$ or $\lambda = 0$. The OLS is suitable when only one of the two variables is random.

The orthogonal regression takes the middle ground by minimizing the sum of squared orthogonal distances from the observed data points to the regression line. The resulting estimator of the slope is (Jackson and Dunlevy 1988):

$$\hat{\beta}_1 = \frac{\frac{S_{YY} - S_{XX}}{S_{XY}} + \sqrt{\Delta}}{2} = \frac{S_{YY} - S_{XX} + \sqrt{(S_{YY} - S_{XX})^2 + 4S_{XY}{}^2}}{2S_{XY}}$$

It is the same as the MLE in the structural model approach when $\lambda = 1$, implying the OR is suitable when the error variances are equal.

The geometric mean regression takes the geometric mean of the slope of y on x, and the reciprocal of the slope of x on y OLS regression lines resulting in the estimated slope

$$\hat{\beta}_1 = sign(S_{XY})\sqrt{\hat{\beta}_{OLS,\text{ Y on X}} * (\hat{\beta}_{OLS,\text{ X on Y}})^{-1}} = sign(S_{XY})\sqrt{\frac{S_{YY}}{S_{XX}}}$$

The GMR can also be obtained by minimizing the sum of the triangular areas bounded by the vertical and the horizontal projections from the data points to the regression line and the regression line itself (Barker et al. 1988). Comparing to the parametric structural model approach, the GMR estimator corresponds to the MLE when $\lambda = S_{YY}/S_{XX}$ (Sprent and Dolby 1980). This means that the GMR approach is suitable when the randomness comes from the random errors only.

3. COMPOUND AND CONSTRAINED REGRESSION ANALYSES

The parametric structural model approach has two fundamental difficulties for real life applications. First, it requires the variables to follow a joint bivariate normal distribution. Second, it requires the knowledge of λ - the ratio of the error variances, which is usually unknown and cannot be estimated from the data statistically (Lindley 1947; Wong 1989). In addition to these predicaments, the structural model approach has also lost the intuitive geometric interpretations enjoyed by the other, albeit more specialized non-parametric regression methods such as OLS, OR or GMR.

In this section, we present the compound regression and the constrained regression methods – two general nonparametric regression frameworks for EIV modeling. Both methods enjoy clear geometric interpretations. They are equivalent to each other, and to the structural model approach when the joint distribution is bivariate normal. We owe our inspiration for these new regression approaches to pioneer statisticians in the optimal design field where they first coined the compound optimal design (Läuter 1974, 1976) and constrained optimal design (Lee 1987, 1988) concepts for designs with multiple objectives, and subsequently proved their equivalency under certain conditions (Cook and Wong 1994; Clyde and Chaloner 1996).

3.1 Compound Regression Analysis

For the OLS on Y and X separately, variation exists in the Y or X direction only and thus one would minimize the sum of squared distances along the vertical or horizontal axis only to obtain the best regression line for each scenario. When both Y and X are random, one would naturally wish to find a regression line $Y = \beta_0 + \beta_1 X$ that will minimize variations in both directions. This can be accomplished by minimizing a weighted average of the squared vertical and horizontal distances, as illustrated in Figure 1 below, as follows:

$$\begin{aligned} SS_\gamma &= \gamma \sum_{i=1}^{\boldsymbol{n}} (Y_i - Y_i)^2 + (1-\gamma) \sum_{i=1}^{\boldsymbol{n}} (X_i - X_i)^2 \\ &= \gamma \sum_{i=1}^{n} (Y_i - \beta_0 - \beta_1 X_i)^2 + (1-\gamma) \sum_{i=1}^{n} (X_i - \frac{Y_i - \beta_0}{\beta_1})^2, \qquad 0 \le \gamma \le 1. \end{aligned}$$

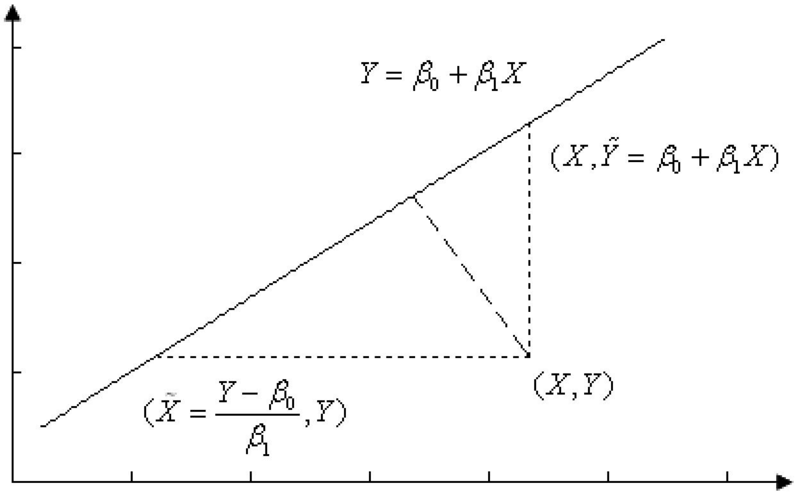
$Y = \beta_0 + \beta_1 X$
$(X, \tilde{Y} = \beta_0 + \beta_1 X)$
(X, Y)
$(\tilde{X} = \frac{Y - \beta_0}{\beta_1}, Y)$

Figure 1. Illustration of the compound regression analysis method.

At the two extreme values of $\gamma = 1$ and $\gamma = 0$, we obtain the OLS on Y or X respectively. For each γ, we can obtain the least squares estimators of the regression parameters by solving $\frac{\partial SS_r}{\partial \beta_0} = 0$ and $\frac{\partial SS_r}{\partial \beta_1} = 0$ simultaneously. Straight-forward derivation shows that the resulting compound regression model estimators $\hat{\beta}_0$ and $\hat{\beta}_1$ would satisfy

$$\beta_0 = \bar{Y} - \beta_1 \bar{X} \quad \text{and} \quad \frac{\gamma}{1-\gamma}\beta_1^4 S_{XX} - \frac{\gamma}{1-\gamma}\beta_1^3 S_{XY} + \beta_1 S_{XY} - S_{YY} = 0 \qquad (1)$$

Solutions can be obtained using any standard numerical software such as MATLAB.

3.2 Equivalence between Compound Regression and Structural Model

In this section, we prove that there is a one-to-one correspondence between the MLE in the structural model approach (under different λ) and the least squares estimator in compound regression analysis (under different γ) for the slope parameter β_1, and thus for the corresponding regression line because each line passes through the point $\left(\bar{X}, \bar{Y}\right)$.

Theorem 1. (a) The compound regression and the structural model approach are equivalent to each other under the bivariate normality assumption. (b) Furthermore, there is a monotone relationship between γ and $\hat{\beta}_1$, and between λ and $\hat{\beta}_1$. (Proof is provided in the Appendix.)

Now that we have shown the equivalence between the structural model and the compound regression approaches, our problem transfers from finding the desirable regression line from a class of unknown λ's to a class of unknown γ's. The constrained regression analysis method, shown to be equivalent to the compound regression analysis approach, will further elucidate our path to a practical non-parametric solution to the EIV modeling problem.

3.3 Constrained Regression Analysis and Regression Efficiencies

We define the constrained regression as follows. As illustrated in Figure 1, given the constraint of $\sum^{n}(Y_i - Y_i)^2 \leq c$ where c is a user selected non-negative constant, the

$i=1$

compound regression line will minimize $\sum_{i=1}^{n}(X_i - X_i)^2$.

Theorem 2. The constrained regression is equivalent to the compound regression in that there is a 1-1 correspondence between $c\left(c \geq 0\right)$ and γ $\left(0 \leq \gamma \leq 1\right)$. For a given c we have $\gamma = \dfrac{S_{YY} - \hat{\beta}_1 S_{XY}}{S_{YY} - \hat{\beta}_1 S_{XY} + \hat{\beta}_1^4 S_{XX} - \hat{\beta}_1^3 S_{XY}}$ and $\hat{\beta}_1 = \dfrac{S_{XY} + sign(S_{XY})\sqrt{S_{XY}^2 - S_{XX}(S_{YY} - c)}}{S_{XX}}$.

(Proof is provided in the Appendix.)

The constrained regression can be stated equivalently in terms of the novel concepts of regression efficiencies for Y and X defined as

$$e_1 = \frac{\min \sum_{i=1}^{n} (Y_i - Y_i)^2}{\sum_{i=1}^{n} (Y_i - Y_i)^2} = \frac{\sum_{i=1}^{n} (Y_i - \hat{Y}_i^{OLS(Y)})^2}{\sum_{i=1}^{n} (Y_i - Y_i)^2} \quad \text{and} \quad e_2 = \frac{\min \sum_{i=1}^{n} (X_i - X_i)^2}{\sum_{i=1}^{n} (X_i - X_i)^2} = \frac{\sum_{i=1}^{n} (X_i - \hat{X}_i^{OLS(X)})^2}{\sum_{i=1}^{n} (X_i - X)^2}$$

respectively. For a given $c^* \in [0,1]$, the constrained regression line will maximize e_2 subject to $e_1 \geq c^*$.

With the equivalence of the constrained and the compound regression approaches, we can first calculate all the compound regression lines given that they are computationally more efficient than their constrained regression counterparts. Then we plot the efficiency curves for all possible γ $(0 \leq \gamma \leq 1)$, and select, from which, the value of γ^* corresponding to a desired c^* (and thus a desirable constrained regression line with intuitive interpretations). The intersection of the line $\gamma = \gamma^*$ and the curve of e_2 in the efficiency plot would yield the best efficiency we can achieve for the estimation of X given the constraint on the required efficiency for Y. By symmetry, one can reverse the order of the importance for X and Y and obtain the best regression line for Y subject to $e_2 \geq c^{**}$. Now that we have circumvented the dilemma of the unknown error variance ratio λ, we will demonstrate our approaches with two examples next.

4. EXAMPLES

In this section, we illustrate the newly proposed nonparametric regression approaches for simple linear regression EIV models through two examples. The first example is from the now classic mathematical statistics textbook by Casella and Berger (2001) where they provided a wonderful introduction to the EIV model. The second example is to model the relationship between the concentrations of organic aerosols and anthropogenic carbon monoxide – the original atmospheric science problem that has motivated this work.

4.1 Example 1

Casella and Berger (2001, pages 542, 579) introduced this classic example for EIV model. Figure 2 shows the scatter plot of the data along with the entire class of compound regression lines ranging from OLS(X) to OLS(Y). The question is which regression line is the 'best' among the entire class of infinitely many regression lines.

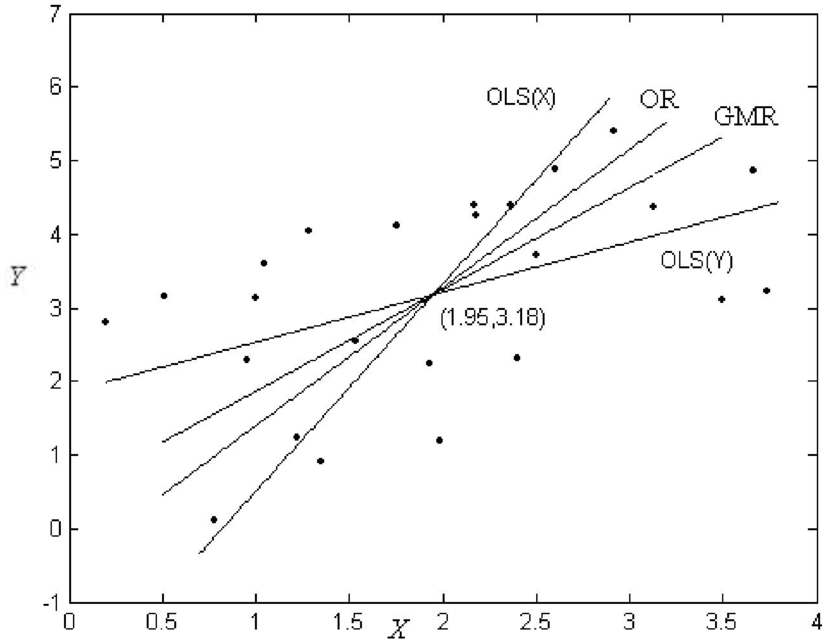

Figure 2. Span of Compound Regression Lines for Example 1.

Table 1. Selected Compound Regression Lines for Example 1.

γ	$\hat{\beta}_1$	$\hat{\beta}_0$	$\sum_{i=1}^{n}(Y_i-\hat{Y}_i)^2$	$\sum_{i=1}^{n}(X_i-\hat{X}_i)^2$	e_1	e_2	e_1+e_2	λ
0 (OLS_X)	2.82	–2.31	137.53	17.33	0.24	1.00	1.24	0.00
0.07(OR)	1.88	–0.48	65.87	18.71	0.50	0.93	1.43	1.00
0.10	1.79	–0.30	61.09	19.16	0.54	0.90	1.44	1.13
0.20	1.57	0.13	51.06	20.84	0.65	0.83	1.48	1.50
0.30	1.43	0.39	46.00	22.50	0.72	0.77	1.49	1.79
0.34(GMR)	1.39	0.48	44.43	23.24	0.75	0.75	1.50	1.91
0.40	1.33	0.59	42.72	24.25	0.78	0.71	1.49	2.07
0.50	1.24	0.76	40.31	26.22	0.82	0.66	1.48	2.36
0.60	1.16	0.92	38.40	28.56	0.86	0.61	1.47	2.71
0.70	1.08	1.08	36.79	31.56	0.90	0.55	1.45	3.17
0.80	0.99	1.24	35.39	35.84	0.94	0.48	1.43	3.90
0.90	0.89	1.45	34.11	43.35	0.97	0.40	1.37	5.57
1 (OLS_Y)	0.68	1.86	33.12	71.94	1.00	0.24	1.24	∞

Table 1 above tabulates selected compound regression lines including OLS(X), OLS(Y), OR and GMR. The efficiencies for estimating X and Y ranging from 0.24 to 1 in opposite directions as the compound regression coefficient γ goes from 0 to 1. The OR line is more efficient in reducing variations in the X direction than the Y direction with efficiencies for X and Y being 0.93 and 0.50 respectively. The GMR provides a nice balance between the two estimations yielding equal efficiencies (0.75) for both X and Y, and moreover, a maximum total efficiency of 1.50. Such is not a mere coincidence; in fact, it is universally true as stated in the following theorem with proof in the Appendix.

Theorem 3. (a) The Geometric Mean Regression would always yield equal efficiencies for the estimations of X and Y respectively. That is, $e_1 = e_2$ for GMR. Furthermore, it also maximizes the total regression efficiency ($e_1 + e_2$) among all compound regression lines. (b) The Ordinary Least Squares Regressions for X and Y have the same efficiencies, albeit in reverse order, for X and Y. That is, $e_1\_OLS(X) = e_2\_OLS(Y)$ and $e_2\_OLS(X) = e_1\_OLS(Y)$.

For each given data set, users can select the desired regression line from the entire class of compound regression lines using the regression efficiency plot as shown in Figure 3. Suppose that the user want the desired line to be at least 95% efficient for the estimation of Y. We will find from the efficiency plot that $e_1 = 0.95$ corresponds to $\gamma = 0.8401$ and $e_2 = 0.453$. Alternatively, if the user desires for at least 85% efficiency for the estimation of Y, he/she will find from Figure 3 that $e_1 = 0.85$ corresponds to $\gamma = 0.5686$ and $e_2 = 0.624$.

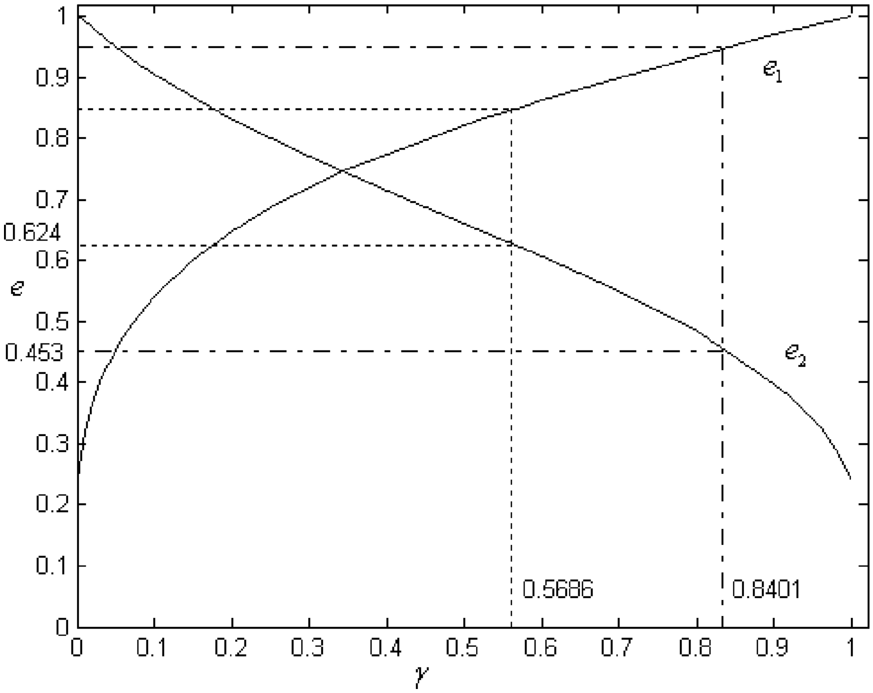

Figure 3. Regression efficiency plot for Example 1.

4.2 Example 2

Our second example from aerosol analysis also motivated this work. The data consist of 113 pairs of CO and organic aerosol concentrations observed above the Mexico City (Kleinman et al. 2007). Our goal is to quantify the linear relationship between the natural log transformed CO concentration (X), and the concentration of organic aerosol (Y). The data and the span of the compound regression lines ranging from OLS(X) to OLS(Y) are shown in Figure 4.

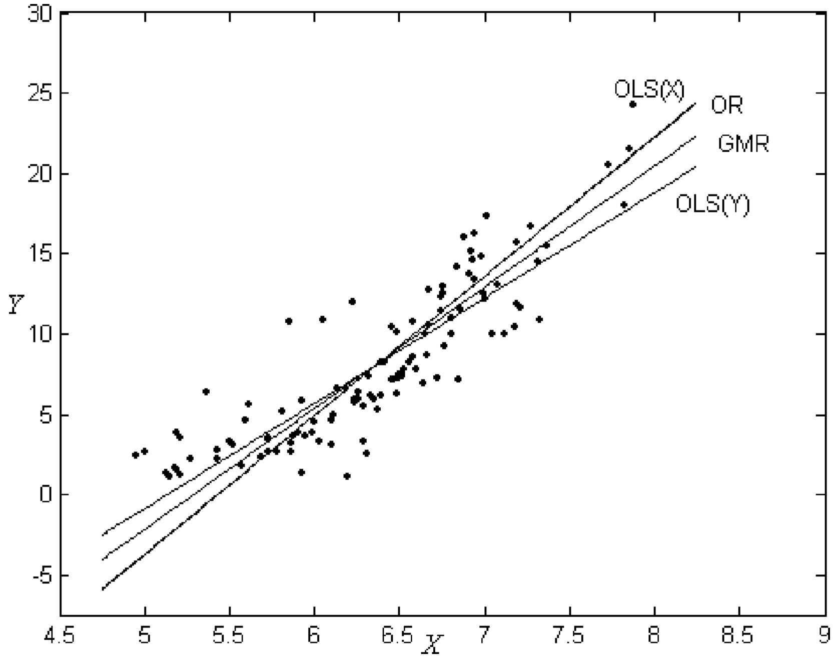

Figure 4. Span of Compound Regression Lines for Example 2.

Unlike Example 1 where the scales of the two random variables are comparable, here the scale of Y can be three times as large as that of X. Subsequently the sum of squares for Y would be much larger than that for X which means the former would dominate the minimization of the compound regression sum of squares SS_γ for most γ. We will still obtain the entire class of compound regression lines however the efficiency plot would be flat in the middle and then change abruptly at the end of the interval for γ -- hampering the visual inspection and selection of desired compound regression lines. This scale inequality, however, can be easily corrected by standardizing the compound regression sum of squares as follows:

$$SS_{\gamma}^{sd} = \gamma \frac{\sum_{i=1}^{n}(Y_i - Y_i)^2}{\min \sum_{i=1}^{n}(Y_i - Y_i)^2} + (1-\gamma)\frac{\sum_{i=1}^{n}(X_i - X_i)^2}{\min \sum_{i=1}^{n}(X_i - X_i)^2}$$

$$= \gamma \frac{\sum_{i=1}^{n}(Y_i - Y_i)^2}{\sum_{i=1}^{n}(Y_i - \hat{Y}_i^{OLS(Y)})^2} + (1-\gamma)\frac{\sum_{i=1}^{n}(X_i - X_i)^2}{\sum_{i=1}^{n}(X_i - \hat{X}_i^{OLS(X)})^2} \qquad 0 \le \gamma \le 1$$

The resulting standardized compound regression is easily shown to be equivalent to the constrained regression as well as the structural model. It also has several added benefits as the GMR now corresponds to γ = 0.5, and the regression efficiencies for X and Y yield perfect symmetrical patterns as shown in Table 2 below.

Table 2. Selected Compound Regression Lines for Example 2.

γ	β_1	β_0	$\sum_{i=1}^{n}(Y_i - Y_i)^2$	$\sum_{i=1}^{n}(X_i - X_i)^2$	e_1	e_2	$e_1 + e_2$	λ
0 (OLS_X)	8.68	-47.13	895.63	11.891	0.755	1.000	1.755	0.00
0.01(OR)	8.64	-46.90	888.08	11.892	0.761	1.000	1.761	1.00
0.10	8.36	-45.13	835.29	11.943	0.810	0.996	1.806	9.56
0.20	8.12	-43.56	795.04	12.066	0.851	0.985	1.836	19.08
0.30	7.91	-42.23	765.36	12.239	0.884	0.972	1.856	29.47
0.40	7.72	-41.03	742.24	12.458	0.911	0.955	1.866	41.62
0.50(GMR)	7.54	-39.90	723.63	12.725	0.935	0.935	1.870	56.87
0.60	7.37	-38.80	708.42	13.052	0.955	0.911	1.866	77.70
0.70	7.19	-37.68	696.01	13.459	0.972	0.884	1.856	109.73
0.80	7.01	-36.50	686.17	13.981	0.985	0.851	1.836	169.44
0.90	6.80	-35.19	679.17	14.689	0.996	0.810	1.806	338.29
1 (OLS_Y)	6.55	-33.62	676.20	15.750	1.000	0.755	1.755	∞

In addition, Table 2 shows that the efficiency of predicting Y increases from 0.755 to 1 while the efficiency of predicting X decreases from 1 to 0.755 as γ goes from 0 to 1.

We also observe that the GMR yields equal efficiencies (0.935) for the estimations of X and Y and a maximum total efficiency of 1.870 as proven in Theorem 3. We would highly recommend GMR as a balanced choice for the aerosol study. If a slightly higher efficiency for the estimation of Y, say 95%, is desired, we can then adopt the corresponding compound regression line with$\gamma = 0.575$ that has a 92% efficiency for estimating X as illustrated in Figure 5.

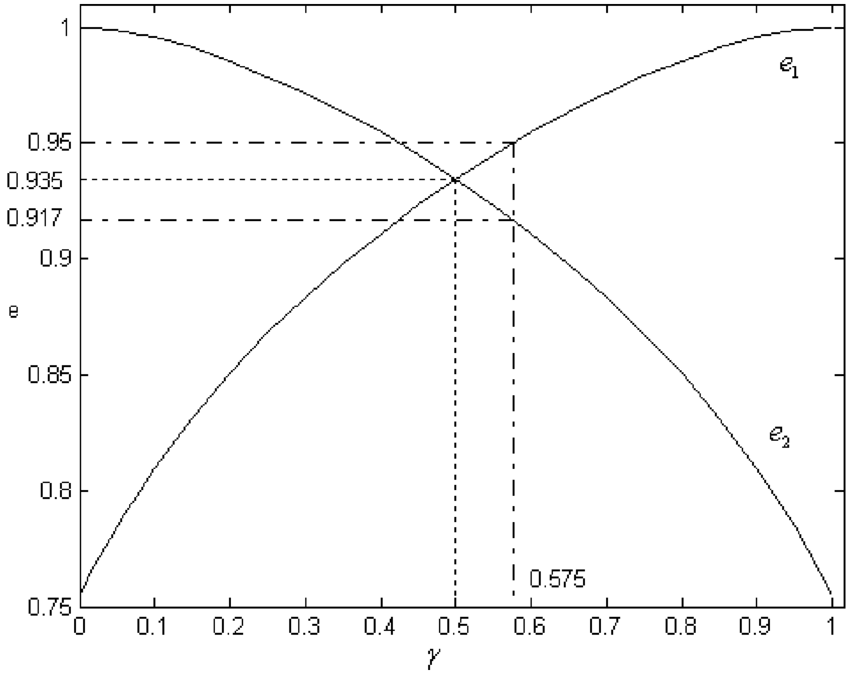

Figure 5. Regression efficiency plot for Example 2.

5. Remarks

In this work, we present two practical solutions to one of the most common problems in applied sciences, namely, regression with random regressors, through the novel non-parametric regression frameworks of compound regression and constrained regression. For EIV models in a simple linear regression setting, we have shown that these two nonparametric regression frameworks are equivalent to each other, and to the traditional parametric structural model approach albeit the latter requires the bivariate normality assumption. Furthermore, each nonparametric regression framework contains the three classic nonparametric regression methods: the ordinary least squares regression, the geometric mean regression and the orthogonal regression as special cases. This lends a systematic approach to examine the properties and relative merit of these classic methods – the geometric mean regression emerges victoriously with two wonderful properties of (1) equal regression efficiencies, and (2) maximum total efficiencies for the estimation of the response variable *Y* and the random regressor *X*.

Compound regression analysis can be readily extended to multiple linear regression with *k* random regressors by minimizing the compound sum of squares of:

$$SS_\gamma = \gamma_1 \sum_{i=1}^{n} (Y_i - Y_i)^2 + \sum_{j=1}^{k} \left[\gamma_j \sum_{i=1}^{n} (X_{ji} - X_{ji})^2 \right], \text{ where } 0 \le \gamma_j \le 1, \ \forall j, \text{ and } \sum\nolimits_{j=1}^{k} \gamma_j = 1.$$

The constrained regression analysis can be extended, by, say, minimizing $\sum_{i=1}^{n} (X_{ki} - X_{ki})^2$ subject to the constraints that $\sum_{i=1}^{n} (Y_i - Y_i)^2 \le c_1$ and $\sum_{i=1}^{n} (X_{ji} - X_{ji})^2 \le c_j$, for $c_1, c_j \ge 0, \ j = 1, \ \ , k-1$. Their relations, properties and relations to the traditional parametric structural model (Kerridge 1967; Patefield 1981) and to the geometric mean regression and orthogonal regression, however, await further elucidation for such higher dimensional errors-in-variable models.

APPENDIX: PROOFS OF RESULTS

Theorem 1. Equivalence of the compound regression and the parametric structural model approach (when the bivariate normality assumption holds for the latter).

Proof. We first prove there is a monotonic relationship between γ and $\hat{\beta}_1$, and between λ and $\hat{\beta}_1$. Given the limits of γ and λ correspond to the same regression lines, OLS(X) and OLS(Y) respectively, it follows immediately that there is a one-to-one correspondence between the compound regression and the structural model approach when the bivariate normality assumption holds for the latter.

From equation (1), we have

$$k=\frac{\gamma}{1-\gamma}=\frac{S_{YY}-\hat{\beta}_1 S_{XY}}{\hat{\beta}_1{}^4 S_{XX}-\hat{\beta}_1{}^3 S_{XY}}=\frac{S_{YY}-\hat{\beta}_1 S_{XY}}{\hat{\beta}_1 S_{XX}-S_{XY}}\frac{1}{\hat{\beta}_1^3}$$

If $S_{XY}\geq 0$, then we have $\frac{S_{XY}}{S_{XX}}\leq\hat{\beta}_1\leq\frac{S_{YY}}{S_{XY}}$, and thus $\frac{S_{YY}-\hat{\beta}_1 S_{XY}}{\hat{\beta}_1 S_{XX}-S_{XY}}>0$. Hence γ is a decreasing function of $\hat{\beta}_1$ and vice versa.

If $S_{XY}<0$, $\frac{S_{YY}}{S_{XY}}\leq\hat{\beta}_1\leq\frac{S_{XY}}{S_{XX}}$ and $\frac{1}{\hat{\beta}_1^3}\frac{S_{YY}-\hat{\beta}_1 S_{XY}}{\hat{\beta}_1 S_{XX}-S_{XY}}=\frac{1}{\hat{\beta}_1^3}\frac{S_{XY}\ \hat{\beta}_1-S_{YY}\blacksquare S_{XY}}{S_{XX}\ S_{XY}\blacksquare S_{XX}-\hat{\beta}_1}$. It follows immediately that γ is an increasing function of $\hat{\beta}_1$ and vice versa.

Next we show that $\hat{\beta}_1$ is a monotonic function of λ. From *equation* (1) we have

$$\frac{\lambda}{1-\lambda}\beta_1{}^4 S_{XX}-\frac{\lambda}{1-\lambda}\beta_1{}^3 S_{XY}+\beta_1 S_{XY}-S_{YY}=0$$

Using the derivative of implicit function, we have

$$\lambda\beta_1^4 S_{XX} - \lambda\beta_1^3 S_{XY} + (1-\lambda)\beta_1 S_{XY} - (1-\lambda)S_{YY} = 0$$

$$\Rightarrow \beta_1^4 S_{XX} + 4\lambda\beta_1^3 S_{XX}\frac{\partial\beta_1}{\partial\lambda} - \beta_1^3 S_{XY} - 3\lambda\beta_1^2 S_{XY}\frac{\partial\beta_1}{\partial\lambda} - \beta_1 S_{XY} + (1-\lambda)S_{XY}\frac{\partial\beta_1}{\partial\lambda} + S_{YY} = 0$$

$$\Rightarrow \frac{\partial\beta_1}{\partial\lambda} = \frac{\beta_1^3 S_{XY} - \beta_1^4 S_{XX} + \beta_1 S_{XY} - S_{YY}}{4\lambda\beta_1^3 S_{XX} - 3\lambda\beta_1^2 S_{XY} + (1-\lambda)S_{XY}}$$

Further simplification of the numerator and denominator lead immediately to the conclusion that λ is a decreasing function of $\hat{\beta}_1$ if $S_{XY} \geq 0$ and vice versa, and an increasing function of $\hat{\beta}_1$ if $S_{XY} < 0$ and vice versa.

Theorem 2. Equivalence of the constrained regression and the compound regression.

Proof. (1) $\forall\gamma$, we have $\hat{\beta}_\gamma$ minimizing $(1-\gamma)\sum_{i=1}^{n}(X_i - X_i)^2 + \gamma\sum_{i=1}^{n}(Y_i - Y_i)^2$.

Let $c = S_{YY}(\hat{\beta}_\gamma) = \sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2 |_{\beta=\hat{\beta}_\gamma}$, $\hat{\beta}_\gamma$ will minimize $\sum_{i=1}^{n}(X_i - X_i)^2$ under the above constraint; otherwise, we will have $\hat{\beta}\,'$satisfy: (i) $S_{YY}(\hat{\beta}\,') \leq c$; (ii) $S_{XX}(\hat{\beta}\,') \leq S_{XX}(\hat{\beta}_\gamma)$ which yields

$$\gamma S_{YY}(\hat{\beta}\,') + (1-\gamma)S_{XX}(\hat{\beta}\,') \leq \gamma c + (1-\gamma)S_{XX}(\hat{\beta}_\gamma) \leq \gamma S_{YY}(\hat{\beta}_\gamma) + (1-\gamma)S_{XX}(\hat{\beta}_\gamma)$$

This contradicts the fact that "$\hat{\beta}_\gamma$ would minimize $(1-\gamma)\sum_{i=1}^{n}(X_i - X_i)^2 + \gamma\sum_{i=1}^{n}(Y_i - Y_i)^2$ ".

Therefore there exists a value c for which $\hat{\beta}_\gamma$ will minimize the corresponding constrained regression.

(2) Similarly one can easily show that $\forall c$, under the constraint of $\sum_{i=1}^{n}(Y_i - Y_i)^2 \leq c$, suppose the estimator minimizing $\sum_{i=1}^{n}(X_i - X_i)^2$ is $\hat{\beta}_c$, then there exists an γ , such that $\hat{\beta}_c$,

where

$$\hat{\beta}_c = \frac{S_{XY} + sign(S_{XY})\sqrt{S_{XY}^2 - S_{XX}(S_{YY} - c)}}{S_{XX}}$$

will also minimize $(1-\gamma)\sum_{i=1}^{n}(X_i - X_i)^2 + \gamma\sum_{i=1}^{n}(Y_i - Y_i)^2$. Furthermore, we have:

$$\gamma = \frac{S_{YY} - \hat{\beta}_c S_{XY}}{S_{YY} - \beta_c \text{姿} S_{XY} + \beta_c^4 S_{XX} - \beta_c^{?3} S_{XY}},$$

Theorem 3. (a) The Geometric Mean Regression would always yield equal efficiencies for the estimations of X and Y respectively. (b) The Ordinary Least Squares Regressions for X and Y have the same efficiencies, albeit in reverse order, for X and Y.

Proof. (a) As described above $\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2 = S_{YY} + \hat{\beta}_1^2 S_{XX} - 2\hat{\beta}_1 S_{XY}$

$$\sum_{i=1}^{n}(X_i - \hat{X}_i)^2 = \frac{1}{\hat{\beta}_1^2}\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2 = \frac{1}{\hat{\beta}_1^2}S_{YY} + S_{XX} - 2\frac{1}{\hat{\beta}_1}S_{XY}$$

For geometric mean regression, we have $\hat{\beta}_1 = sign(S_{XY})\sqrt{S_{YY} / S_{XX}}$; hence,

$$e_1 = \frac{\min\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2}{\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2} = \frac{\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2 |_{\hat{\beta}_1 = \frac{S_{XY}}{S_{XX}}}}{\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2 |_{\hat{\beta}_1 = \sqrt{\frac{S_{YY}}{S_{XX}}}}} = \frac{S_{YY} + \frac{S_{XY}^2}{S_{XX}} - 2\frac{S_{XY}^2}{S_{XX}}}{S_{YY} + S_{YY} - 2S_{XY}\sqrt{\frac{S_{YY}}{S_{XX}}}} = \frac{S_{XX}S_{YY} - S_{XY}^2}{2S_{XX}S_{YY} - 2S_{XY}\sqrt{S_{XX}S_{YY}}}$$

$$e_2 = \frac{\min\sum_{i=1}^{n}(X_i - \hat{X}_i)^2}{\sum_{i=1}^{n}(X_i - \hat{X}_i)^2} = \frac{\frac{1}{\hat{\beta}_1^2}\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2 |_{\hat{\beta}_1 = \frac{S_{YY}}{S_{XY}}}}{\frac{1}{\hat{\beta}_2^2}\sum^{n}(Y - \hat{Y})_{22}^2 | \sqrt{}} = S_{XX}S_{YY} - S_{XY}^2$$

$$\hat{\beta} = \quad s_{YY} \quad \begin{matrix} 2S_{XX}\ S_{XX}S_{YY} \\ S_{YY} - \quad i \quad\ i \\ \beta_1 2S_{XY} \end{matrix} \quad 1$$

$$i=1 \qquad\qquad i=1 \qquad\qquad s_{XX}$$

Thus we have proven that $e_1 = e_2$, for the geometric mean regression.

Furthermore, the total regression efficiency simplifies to:

$$e_1 + e_2 = \left(S_{XX}S_{YY} - S_{XY}^2\right)\frac{\dfrac{1}{S_{XX}} + \dfrac{\hat{\beta}_1^2}{S_{YY}}}{S_{YY} + \hat{\beta}_1^2 S_{XX} - 2\hat{\beta}_1 S_{XY}}$$

Its derivative with respect to $\hat{\beta}_1$ simplifies to:

$$\frac{\partial(e_1 + e_2)}{\partial\hat{\beta}_1} = \left(\frac{1}{S_{XX}} - \frac{\hat{\beta}_1^2}{S_{YY}}\right)\frac{2S_{XY}\left(S_{XX}S_{YY} - S_{XY}^2\right)}{\left(S_{YY} + \hat{\beta}_1^2 S_{XX} - 2\hat{\beta}_1 S_{XY}\right)^2}$$

Setting the above to zero, we find immediately that the geometric mean regression maximizes the total regression efficiency with $\hat{\beta}_1 = sign(S_{XY})\sqrt{S_{YY}/S_{XX}}$.

(b) The equality of e1 (for $\gamma = 0$) and e2 (for $\gamma = 1$) are easily proven as follows:

$$\gamma = 0,\ \beta_1 = \frac{S_{XY}}{S_{XX}},\quad e_1 = \frac{S_{YY} + \left(S_{XY}^2/S_{XX}^2\right)S_{XX} - 2\left(S_{XY}^2/S_{XX}\right)}{S_{YY}} = \frac{S_{YY}S_{XX} - S_{XY}^2}{S_{YY}S_{XX}} = \frac{S_{XX}S_{YY} - S_{XY}^2}{S_{YY}S_{XX}}$$

$$\gamma = 1,\ \beta_1 = \frac{S_{YY}}{S_{XY}},\quad e_2 = \frac{\left(S_{YY}^2/S_{XY}^2\right)S_{XX} + S_{YY} - 2\left(S_{YY}^2/S_{XY}\right)}{\left(S_{YY}^2/S_{XY}^2\right)S_{XX}} = \frac{S_{YY}S_{XX} - S_{XY}^2}{S_{YY}S_{XX}} = \frac{S_{XX}S_{YY} - S_{XY}^2}{S_{YY}S_{XX}}$$

REFERENCES

Barker, F., Soh,Y. C., and Evans, R. J. (1988), “Properties of the Geometric Mean Functional Relationship,” *Biometrics*, 44, 279-281.

Casella, G., and Berger, R. L. (2001), “Statistical Inference (Second Edition)”. Duxbury.

Cook, R. D., and Wong, W. K. (1994), “On the Equivalence of Constrained and Compound Optimal Designs,” *Journal of the American Statistical Association,* 89, 687-692

Clyde, M., and Chaloner, K. (1996), “The Equivalence of Constrained and Weighted Designs in Multiple Objective Design Problems,” *Journal of the American Statistical Association*, 91, 1236 -1244.

Jackson, J. D., and Dunlevy, J. A. (1988), “Orthogonal Least Squares and the Interchangeability of Alternative Proxy Variables in the Social Sciences,” *The Statistician*, 37, 7-14.

Jaynes, E. T. (2004), “The Logic of Science”. Cambridge University Press.

Jobson, B. T., Mckeen, S. A, Parrish, D.D., Fehsenfeld, F. C., Blake, D. R., Goldstein, A. H., Schauffler, S. M., and Elkins, J. W. (1999), “Trace Gas Mixing Ratio Variablility Versus Lifetime in the Troposphere and Stratosphere: Observations,” *Journal of Geophysical Research*, 104, 16091-16113

Kerridge, D. (1967), “Errors of Prediction in Multiple Regression with Stochastic Regressior Variables,” *Technometrics*, 9, 309-311.

Kleinman, L. I., Springston, S. R., Daum, P. H., Lee, Y.-N., Nunnermacker, L. J., Senum, G.I., Wang, J., Weinstein-Lloyd, J., Alexander, M. L., Hubbe, J., Ortega, J., Canagaratna, M. R., and Jayne, J. (2007), “The Time Evolution of Aerosol Composition over the Mexico City Plateau,” *Atmospheric Chemistry and Physics*, 7, 1-49.

Lindley, D. V. (1947) *Supplement to the Journal of the Royal Statistical Society*, 9, 218-244.

Läuter, E. (1974), "Experimental Planning in A Class of Models," *Mathematische Operationsforschung Und Statistik*, 5, 673-708.

Läuter, E. (1976), "Optimal Multipurpose Designs for Regression Models," M*athematische Operationsforschung Und Statistik*, 7, 1-68.

Lee, C. M. S. (1987), "Constrained Optimal Designs for Regression Models," *Communications in Statistics*, Part A-Theory and Methods, 16, 765-783.

Lee, C. M. S. (1988), "Constrained Optimal Design," *Journal of Statistical Planning and Inference*, 18, 377-389.

Patefield, W. M. (1981), “Multivariate Linear Relationships: Maximum Likelihood Estimation and Regression Bounds,” *Journal of Royal Statistics Society B*, 43, 342-352.

Sprent, P., and Dolby, G. R. (1980), “Query: the Geometric Mean Functional Relationship,” *Biometrics*, 36, 547-550.

Sprent, P. (1969), “Models in Regression and Related Topics,” *Methuen's Statistical Monographs.*

Wong, M.Y. (1989), “Likelihood Estimation of A Simple Linear Regression Model

When Both Variables Have Error," *Biometrica,* 76, 141-148.

Part II.
Joint Multivariate Analysis of Aerosol Mass Spectra

Tianyi ZHANG[1], Wei ZHU[1] and Robert MCGRAW[2]

[1]Department of Applied Mathematics and Statistics, State University of New York, Stony Brook, NY 11794-3600

[2]Atmospheric Sciences Division, Brookhaven National Laboratory, Upton, NY 11973-5000.

Abstract: Aerodyne aerosol mass spectra (AMS) datasets typically contain hundreds of mass to charge ratios and their corresponding intensities from air sampled through the mass spectrometer. The observations are usually taken time sequentially to monitor the air composition, quality and change in an area of interest. An important goal of the present AMS data analysis is to reduce the dimensionality of the original data yielding a small set of representative tracers for next generation atmospheric models. In this work, we present an approach to jointly apply three multivariate analysis techniques -- cluster analysis, principle component analysis and non-negative least squares towards this goal. Application to a recent field study demonstrates the effectiveness of this new approach. Comparisons are made to singly applied multivariate statistical techniques including principal component analysis and positive matrix factorization, and guidelines are provided.

Key Words *Aerodyne Aerosol Mass Spectra, Cluster Analysis, Non-Negative Least Squares, Positive Matrix Factorization, Principal Component Analysis*

INTRODUCTION

Atmospheric aerosols are known to play important roles in climate and climate change. These include the aerosol direct effect, whereby particles directly scatter solar radiation back into space and aerosol indirect effects brought about by the fact that aerosol particles, which serve as sites for cloud droplet condensation, influence cloud properties. Thus, greater numbers of aerosol particles tend to result in a greater concentration of smaller cloud droplets and brighter clouds (this, the so-called first indirect effect, also results in the scattering of more solar radiation back into space). Greater numbers of smaller droplets also tend to make clouds more stable - reducing precipitation rates and leading to changing patterns of rainfall and increased cloud lifetime. All of these effects are dependent on aerosol properties, especially particle number, size, and chemical composition. In recent years the study of aerosol composition has been greatly advanced through the development of aerosol mass spectrometers and though measurements taken by many investigators deploying these instruments at a variety of sites around the globe (see, e.g., Jimenez et al. 2003 ; Allan et al. 2003[ab] ; Allan et al. 2004 ; Zhang et al. 2005[b] ; DeCarlo et al. 2006 ; Volkamer et al. 2006). Due to the complexity of Aerodyne aerosol mass spectra (AMS) data, indicative of the complexity of the aerosol itself, an area of great interest in quantitative study is dimension reduction. The goal here is to obtain, from the original high dimensional datasets, a few representative tracers that will elucidate the sources and interactions of different aerosol components, and to identify the most important tracking variables for the next generation of climate models.

In previous studies, several multivariate analysis techniques have been singly applied individually towards the dimensional reduction of Aerodyne AMS data. These include principal component analysis (Zhang et al. 2007) and cluster analysis (Marcolli et al. 2006). The most widely used technique to extract information on organic aerosol types from Aerodyne AMS data in the atmospheric research community is Positive Matrix Factorization (PMF), also known as the non-negative matrix factorization (Lee and Seung 1999). PMF has been developed (Paatero and Tapper 1994) to yield factors, which combined with non-negative coefficients are reflective of the positive mixing of a basis set of chemical components (the factors) that contribute to aerosol mixing state. Recent references include (Lee et al. 1999; Ramadan et al. 2000; Larsen and Baker 2003; Maykut et al. 2003) and two review papers (Engel-Cox and Weber 2007; Reff et al. 2007). However, a number of weaknesses are inherent in PMF, especially its subjectivity and non-uniqueness (Ulbrich et al. 2009), make it clear improvements are needed and that better methods must be developed to better serve the atmospheric research community.

In this paper, we propose a joint cluster analysis on variables (VARCLUS) (Harman

1976; Cattell 1965; Rummel 1970), principle component analysis (PCA) (Jolliffe 2002) and non-negative least squares (NNLS) (Lawson and Hanson 1974; Donoho and Stodden 2003) approach to better achieve the goal of dimension reduction. Each method alone is not new; however, their combination is novel and, as shown later, rather effective in solving the problem at hand. The paper is arranged as follows: a brief introduction is provided to a set of AMS data collected from an aircraft flying in and downwind of Mexico City; the proposed method is described in detail in the method section while the corresponding output is displayed with application to the given AMS data; Finally, it concludes with a comparison between the proposed method and the existing methods, especially PCA and PMF.

THE AEROSOL MASS SPECTRA DATASET

Instrumentation

The aerosol mass spectra analyzed in this work were recorded using an Aerodyne aerosol mass spectrometer during a flight of the US Department of Energy (DOE) G-1 aircraft in and downwind of Mexico City during a morning/pre-noon flight on March 19, 2006 (Kleinman et al. 2008). The basic mechanism of operation of the spectrometer can be found in (Jayne et al. 2000). In summary, the air sample flows through an orifice and particles are focused by an aerodynamic lens and passed through a chopper, which is set to either an open or a blocked mode. Vaporized and ionized, particle species are introduced into the quadrupole mass spectrometer sequentially with their weights and frequencies recorded. A more detailed description of the Aerodyne mass spectrometer can be found in (Jimenez et al. 2003).

Data Pre-processing

Each mass spectrum contains signals with m/z values ranging from 1 to 452, and spectra were recorded every 12 seconds with a total of 943 spectra in the initial dataset. During initial processing, two types 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 has its own baselines, designated as MSSOpenBaseL and MSSClosedBaseL. The difference between the open mode and the closed mode was used to remove the contribution from background gas in the detector, resulting in the data file analyzed in this paper named MSSDiff and defined as:

MSSDiff = (MSSOpen - MSSOpenBaseL) - (MSSClosed - MSSClosedBaseL)

Data in MSSDiff contain negative values, which are expected for signals that are close to zero in the presence of noise. We retain the negative values for the analysis because the proposed joint cluster and NNLS approach allows negative values.

In this paper, we focus on the subset of m/z values from 1 to 100, because mainly low-m/z fragments are generated from larger m/z organic species with electron impact ionization. Therefore the majority of signal in an instrument like the Aerodyne AMS is found in the m/z range below 100. Twenty-eight peaks related to large air/water or no useful aerosol signals (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. Other inorganic peaks like those due to sulphate or nitrate particles were retained for analysis. Two spectra (time = 353, 643 seconds) showed a signal several times higher than the third highest spectra. They were clear outliers and thus removed.

METHOD

Notations and Overview

Let **DATA** represent the pre-processed AMS data with N_i columns for m/z values ranging from 1 to N_i (N_i =100), and N_t (N_t =941) rows for the N_t spectra. Thus the i^{th} column of the matrix **DATA** is the time series corresponding to m/z value i, while the t^{th} row of **DATA** is the mass spectrum sampled at time point t. Our goal is to find a mass spectrum basis, **MS**, with a small number of entries N_c, such that

$$\mathbf{DATA} = \mathbf{C}{*}\mathbf{MS} + \mathbf{E} \tag{1}$$

Here, **DATA** and the error terms **E** are N_t*N_i matrix (in our work N_t = 941, N_i = 100). **MS** is an N_c*N_i matrix, $N_c << N_t$. Our goal includes two parts: first, we need to decide upon N_c, the number of basis spectra (equal to the number of candidate tracers); second, we need to estimate **MS**. The above model was first presented in Zhang et al. (2005^a) in a regression form. In her work, she selected two "tracers" (m/z = 44 and m/z = 57), and a two-step regression was applied. Both the values of "tracers" (which is **C** in equation (1)) and the coefficients (which is **MS** in equation (1)) have physical meanings. Ulbrich et al. (2008) presented the decomposition process with the PMF approach. In general, the decomposition scheme is shown in Figure 1.

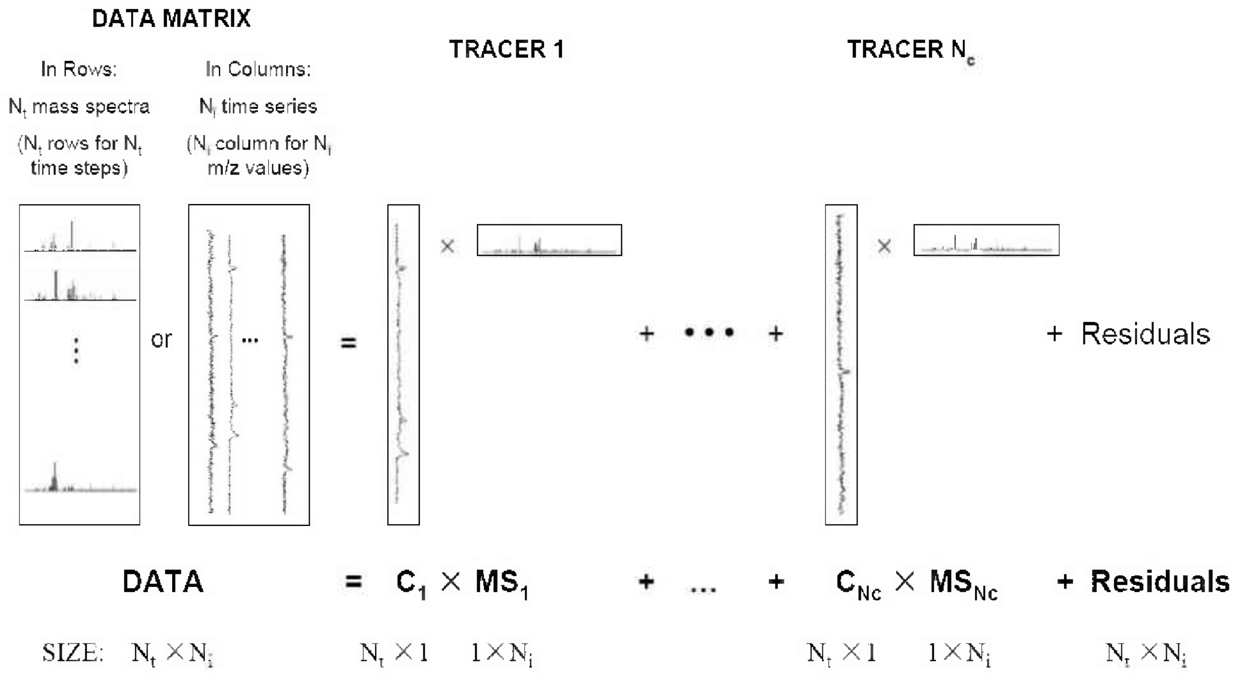

Fig. 1 Schema of the decomposition of the AMS dataset. The time series of the tracers make up the matrix **C** and the mass spectra of the tracers make up the matrix **MS** in Equation (1).

The above decomposition can be solved by the principal component analysis (PCA), factor analysis (FA) or positive matrix factorization (PMF). Both PCA and FA yield a suggestive N_c based on the variation explained by the basis. However, their output factors, with both positive and negative coefficients, are often unacceptable. The PMF method yields non-negative factors and coefficients. However, these factors are not always interpretable and furthermore, the PMF output is not unique (Lee and Seung 1999, 2001; Ulbrich et al. 2008).

Hence 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 N_c. 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.

VARCLUS and PCA

The AMS data is usually sufficiently linear so as to return 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 (PC1) for each cluster is a weighted linear combination of all m/z values in the given cluster, usually with non-negative coefficients, and would explain the most (and often the majority) variation in the given cluster, it is the natural choice as a representative tracer for each cluster. VARCLUS is a procedure complemented in SAS. VACLUS iteratively splits variables into a binary tree 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 (SAS Institute, 2008). Hence, VARCLUS is the most suitable hierarchical clustering technique consistent with within-cluster PCA.

Thus the dimension of basis N_c 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 with its columns consisting of PC1's from all major clusters, which we refer to as the basis/tracer set. We denote the columns of **DATA** as $\mathbf{MZ_1}$**, …,** $\mathbf{MZ_{100}}$ (each of length 941), while the rows of **DATA** as $\mathbf{T_1}$**, …,** $\mathbf{T_{941}}$ (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 C_i is a weighted summation of disjoint subset of the m/z's (i.e., $\mathbf{MZ_k}$**'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.

Non-Negative Least Squares (NNLS)

Now that we have obtained the basis dimension N_c and the basis matrix C, our next goal is to find the matrix **MS** in $\underset{941*100}{DATA} = \underset{941*Nc}{C} * \underset{Nc*100}{MS} + E$ 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 coefficients β, required to be non-negative, 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 $\| DATA - DATA \|^2$, with the constraints that each β is non-negative. The details of the NNLS algorithm can be found in Lawson and Hanson (1974). For better fit, we add the intercepts in the above linear equation system as follows.

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

In matrix form we have: $\underset{941*100}{DATA} = \underset{941*(Nc+1)}{C} * \underset{(Nc+1)*100}{MS}$, where C is **C** with an extra column of 1's, and MS is **MS** with the added row $(\beta_{1,0}, \beta_{2,0} \ldots, \beta_{100,0})$, a vector of intercepts allowed to be negative as it is unrelated to any mass spectrum.

RESULTS

The VARCLUS output, a hierarchical tree, is shown in Figure 2. Based on the output tree and the related aerosol information, we have, clearly, 4 major clusters. More importantly, each of these clusters corresponds to a unique aerosol class of interest with their members (m/z's) highly consistent to the corresponding factors from theoretic analysis as shown in Tables 1 and 2 for the organic and the inorganic aerosols respectively.

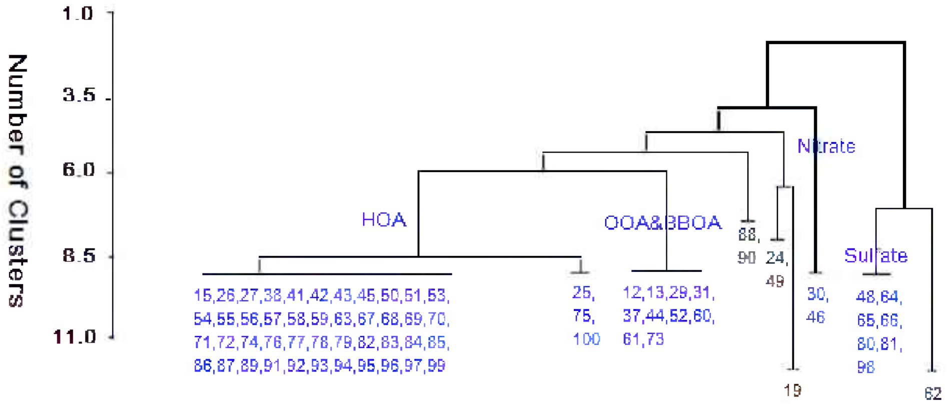

Fig. 2 VARCLUS output: the hierarchical tree showing 4 major clusters with member m/z values written in blue. Only six m/z's (in black) are excluded from these major clusters. Cluster 1 (HOA), 2 (Sulfate), 3 (OOA and BBOA) and 4 (Nitrate) contain 47, 7, 10 and 2 m/z values with its PC1 explaining 91, 99, 90 and 100% of the cluster variations respectively.

Table 1 Organic aerosol factors. The underlined m/z's are critical signals for each 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
OOA2 Oxygenated organic aerosols type II	$C_nH_mO \rightarrow C_2H_3O^+$, $C_3H_3O^+$,	43, 55, ...	⬇
BBOA biomass burning organic aerosols	$R \rightarrow R'^+$, $C_2H_4O_2{}^+$, $C_3H_3O_2$	44, 45, ...	Mor e
OOA1 Oxygenated organic aerosols type I	$C_nH_mO_2 \rightarrow CO_2{}^+$, $HCO_2{}^+$, R	60, 73, ...	Oxid ized

Table 2 Inorganic aerosol factors, with major signals (m/z's) listed.

Inorganic aerosol factors	Major components	m/z
Sulfate	SO, SO_2, SO_3^{2-}, HSO_3^-, H_2SO_4	48, 64, 80, 81, 98
Nitrate	NO, NO_2	30, 46

For the VARCLUS algorithm, the similarity measure is the Pearson correlation. For the particular atmospheric application we reported here, the correlation measure is most meaningful because of the inherent linearity of the aerosol mass spectra (AMS) data. In AMS data, organic aerosols usually come from several main sources. For example, hydrocarbon-like organic aerosols (HOA) mostly come from fossil fuel, while oxygenated organic aerosols (OOA) mostly from secondary organic aerosols (SOA) (Zhang et al. 2007). This property lent theoretical foundation and explanation to our VARCLUS approach and results – where our clusters clearly correspond to the four major aerosol classes. In addition, all PC1's in these clusters represent a high percentage of variation explained. Thus, it is reasonable to use these PC1's as the basis (tracer set) for the given AMS data. The proportion of total variation explained, calculated as the ratio of the summation of all PC1 variances divided by the total variances, is 95%.

Based on the VARCLUS clusters and the subsequent PCA output, we obtained the matrix **C** and as illustrated in Figure 1, we proceed to estimate the MS basis matrix **MS** using NNLS. The resulting four non-negative MS basis are as shown in Figure 3.

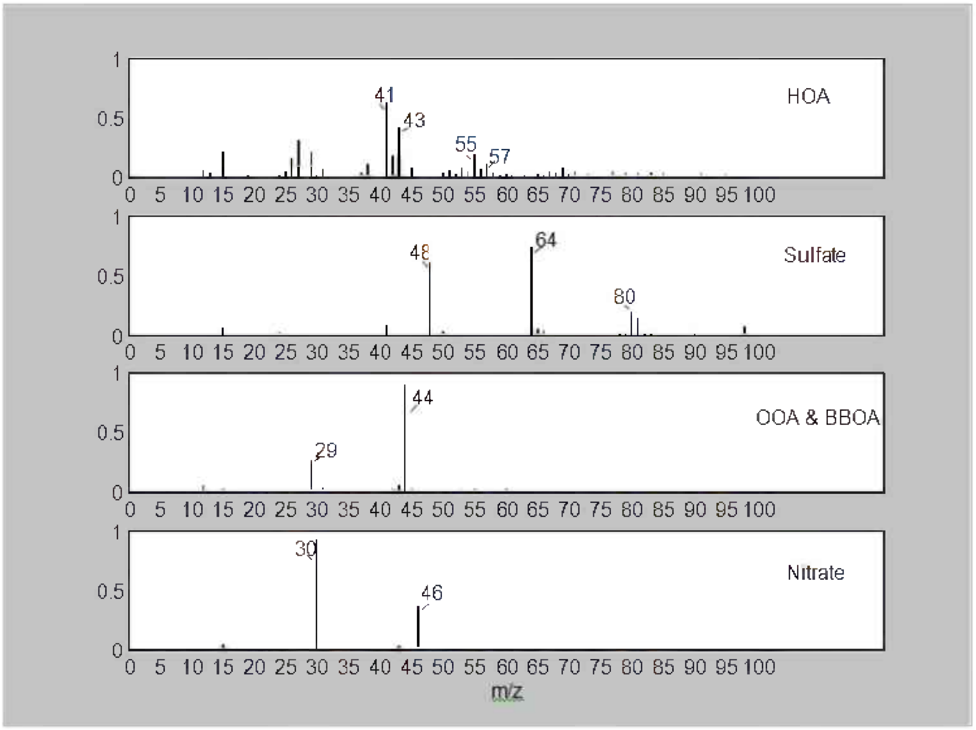

Fig. 3 MS basis obtained using NNLS for the four major clusters of HOA, Sulfate, OOA&BBOA and Nitrate.

SUMMARY AND 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. For example, Figure 4 below shows the mass spectrum basis obtained by using the first four PC's of a single principal component analysis based on the entire mass spectra. It contains large negative values, which is hard to interpret in the mass spectrum language.

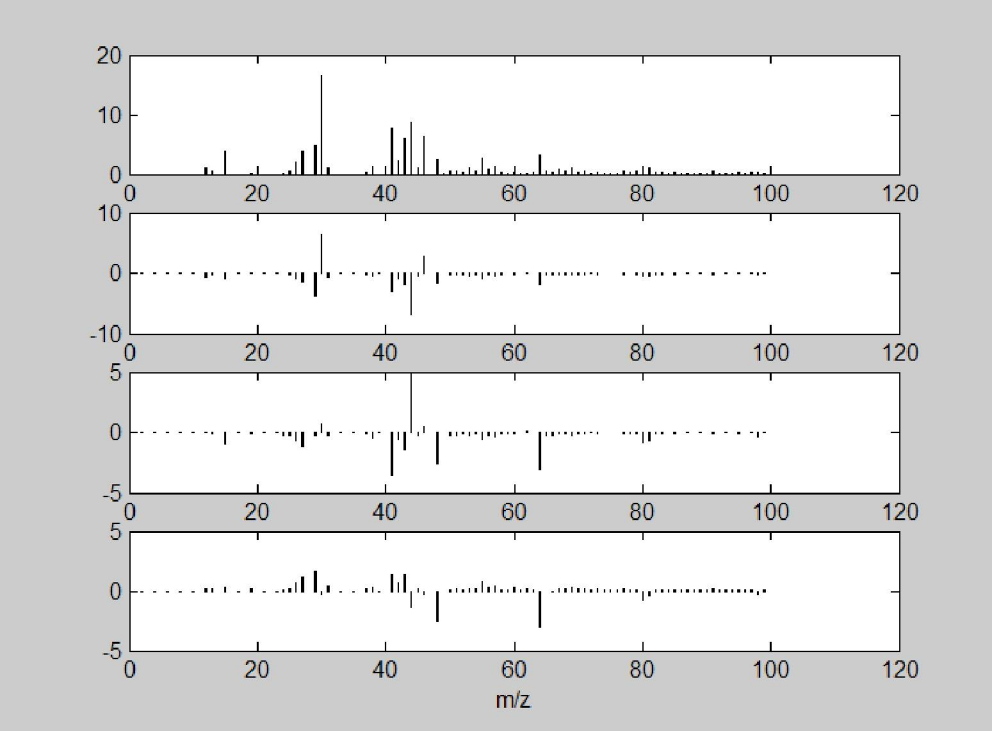

Fig. 4 The mass spectrum basis obtained using the first four principal components of a single principal component analysis based on the entire mass spectrum data set.

Both the PMF and the proposed method of joint VARCLUS and NNLS analysis can achieve non-negative tracer coefficients. The major criticism of the PMF is 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 different PMF's in **X** = (**AB**)***C** = **A***(**BC**). The different factorizations are called "rotation" in factor analysis. For the proposed joint analysis method, there is little flexibility in deciding the major clusters and thus the basis. For example, once the **C** matrix has been computed, the **MS** matrix determined by the NNLS procedure is unique.

As Paatero (See Ulbrich et al. 2008) pointed out "It is unfortunate that introducing a priori information also introduces some subjectivit y in the analysis." To apply PMF analysis, one needs to decide at least two things: the number of factors and the "rotation" parameter. Otherwise one will find the solution non-unique. Both of these choices have determinant effect on the final output. For our method, we still have to refer to some prior knowledge. However, the "subjectivity" exists in only one step – the determination of major clusters based on the VARCLUS output. 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.

A summary of comparison among current methods for aerosol mass spectra data study is shown in Table 3.

There is still space for us to improve 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. Next, 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.

Table 3 Comparison of methods for aerosol mass spectra data study

Methods	How to determine the dimension (N_c).	How to find the tracer matrix **C**	How to find the basis spectra **MS**	Comments
PCA in rows (mass spectrum)	Based on PCA output to get an appreciable proportion of variation explained.	No output for **C**.	PCA applied to find **MS**.	Negative values in **MS (**see Fig. 4**)**.
PCA in columns (time series)	The same as above.	PCA applied to find **C**.	No output for **MS**.	Meaningles s time series.
PMF	Comparing output with different N_c.	Find **C** and **MS** together with certain PMF algorithm.		Non-unique ; subjectivity in determining N_c.
Hierarchi cal	Based on both clustering structure	No output for **C**.	Clustering applied to	**MS** are usually not

Cluster Analysis	chemistry knowledge.		find **MS**.	clearly explainable.
Iterative Regression (Zhang et al. 2005[a])	N_c=2.	Use m/z44 and m/z 57 as two tracers to form matrix **C**$^{(0)}$ at the very beginning; Use Ordinary Least Square (OLS) regression to find **C**$^{(i)}$ based on **MS**$^{(i-1)}$.	Use OLS regression to find **MS**$^{(i)}$ based on given **C**$^{(i)}$.	Information lost; possible to get negative results.
VARCLUS+NNLS	Based on both clustering structure and chemistry knowledge.	Use linear combinations of m/z's in clusters as tracers to form matrix **C**.	Use NNLS to find **MS** based on **C**.	Valid for data structure exploration; non-negative **MS**.

ACKOWLEDGEMENTS Funding for this research was provided by the U.S. Department of Energy's Scientific Discovery through Advanced Computing (SciDAC) program. The MAX-Mex G1 AMS data were collected by John Jayne of Aerodyne Research, Inc., and M. Lizabeth Alexander of PNNL. We are grateful to Dr. Yin-Nan Lee of BNL for his help and useful comments.

REFERENCES

Allan JD, et al. (2003[a]) Quantitative sampling using an aerodyne aerosol mass spectrometer 1, techniques of data interpretation and error analysis. J Geophys Res108: 4090

Allan JD, et al. (2003[b]) Quantitative sampling using an aerodyne aerosol mass spectrometer2, measurements of fine particulate chemical composition in two U.K. cities. J Geophys Res 108: 4091

Allan JD, et al. (2004) A generalised method for the extraction of chemically resolved mass spectra from aerodyne aerosol mass spectrometer data. J Aerosol Sci 35: 909-922

Cattell RB (1965) Factor analysis: An introduction to essentials II. The role of factor analysis in research. Biometrics 21: 405-435

DeCarlo PF, et al. (2006) A field-deployable high-resolution time-of-flight aerosol mass spectrometer. Anal Chem 78: 8281-8289

Donoho D and Stodden V (2003) When does non-negative matrix factorization give a correct decomposition into parts? In: Thrun S, Saul LK, and Schölkopf B (eds) Advances in neural information processing systems 16. MIT Press, Cambridge, pp 1141-1148

Engel-Cox J and Weber SA (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 Manage Assoc 57: 1307–1316

Harman HH (1976) Modern factor analysis, 3rd edn. University of Chicago Press, Chicago

Jayne JT, et al. (2000) Development of an Aerosol Mass Spectrometer for Size and Composition Analysis of Submicron Particles. Aerosol Sci Technol 32: 49-70

Jimenez JL, et al. (2003) Ambient aerosol sampling using the aerodyne aerosol mass spectrometer. J Geophys Res 108: 8425

Jolliffe IT (2002) Principal Component Analysis, Series: Springer Series in Statistics, 2nd ed., Springer, NY.

Kleinman LI, et al. (2008) The time evolution of aerosol composition over the Mexico City plateau. Atmos Chem Phys 8: 1559-1575

Larsen RK and Baker JE (2003) Source apportionment of polycyclic aromatic hydrocarbons in the urban atmosphere: a comparison of three methods. Environ Sci Technol 37: 1873–1881

Lawson CL and Hanson RJ (1974) Solving least squares problems. Prentice-Hall, New Jersey

Lee DD and Seung HS (1999) Learning the parts of objects by non-negative matrix factorization. Nature 401: 788–791

Lee DD and Seung HS (2001) Algorithms for non-negative matrix factorization. In: Leen TK, Diatterich TG and Tresp V (eds) Advances in neural information processing systems 13. MIT Press, Cambridge, pp 556-562

Lee E, Chan CK and Paatero P (1999) Application of positive matrix factorization in source apportionment of particulate pollutants in Hong Kong. Atmos Environ 33: 3201–3212

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

Maykut NN, et al. (2003) Source Apportionment of PM2.5 at an urban improve site in Seattle, Washington. Environ Sci Technol 37: 5135–5142

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

Ramadan Z, Song X and Hopke PK (2000) Identification of sources of Phoenix aerosol by positive matrix factorization. J Air Waste Manage Assoc 50: 1308–1320

Reff A, Eberly SI and Bhave PV (2007) Receptor modeling of ambient particulate matter data using positive matrix factorization: Review of existing methods. J Air Waste Manage Assoc 57: 146–154

Rummel RJ (1970) Applied factor analysis. Northewestern Univ. Press, Evanston

SAS (2008). SAS/STAT 9.2 User's Guide. Cary, NC: SAS Institute Inc.

Ulbrich IM et al. (2009) Interpretation of organic components from positive matrix factorization of aerosol mass spectrometric data. Atmos Chem Phys 9: 2891-2918

Volkamer R et al. (2006) Secondary organic aerosol formation from anthropogenic air pollution: Rapid and higher than expected. Geophys Res Lett 33: L17811

Zhang Q, et al. (2005^{a}) Deconvolution and quantification of hydrocarbon-like and oxygenated organic aerosols based on aerosol mass spectrometry. Environ Sci Technol 39: 4938-4952

Zhang Q, et al. (2005 ^{b}) Time- and size-resolved chemical composition of submicron particles in Pittsburgh: Implications for aerosol sources and processes. J Geophys Res 110: D07S09

Zhang Q, et al. (2007) Ubiquity and dominance of oxygenated species in organic aerosols in anthropogenically-influenced northern Hemisphere midlatitudes. Geophys Res Lett 34: L13801