# Determining the Number of Major Pollution Sources in Multivariate Air Quality Receptor Models

Eun Sug Park  Ronald C. Henry  Clifford H. Spiegelman

# NRCSE

Technical Report Series

NRCSE-TRS No. 034

The NRCSE was established in 1996 through a cooperative agreement with the United States Environmental Protection Agency which provides the Center's primary funding.

EPA

# Determining the Number of Major Pollution Sources in Multivariate Air Quality Receptor Models

[*]Eun Sug Park[1], Ronald C. Henry[2], and Clifford H. Spiegelman[3]

1. University of Washington

National Research Center for Statistics and the Environment

Seattle, Washington 98195-4323

2. University of Southern California

Department of Civil & Environmental Engineering

3620 South Vermont Ave

Los Angeles, CA 90089-2531

3. Texas A&M University

Department of Statistics

College Station, TX 77843-3143

[*]Corresponding author: Tel.:206-616-9439; Fax:206-616-9443; E-mail:epark@stat.washington.edu

**Abstract and Key Words**

We present two new statistics for estimating the number of factors underlying in a multivariate system. One of the two new methods, the original NUMFACT, has been used in high profile environmental studies. The two new methods are first explained from a geometrical viewpoint. We then present an algebraic development and asymptotic cutoff points. Next we present a simulation study that shows that for skewed data the new methods are typically superior to traditional methods and for normally distributed data the new methods are competitive to the best of the traditional methods. We finally show how the methods compare by using two environmental data sets.

KEY WORDS: Correlation Matrix; Resampling; Eigenvectors; Eigenvalues; NUMFACT.

# 1. INTRODUCTION

Selecting the number of underlying factors is often the most difficult step in building a multivariate model. The methods most commonly used to estimate the number of factors usually disagree. As a result, practitioners typically build several models varying the number of factors over the range given by their favorite methods. The "best" model is the one that makes most sense in the context of the application. The present work was motivated by the second author's experience modeling air quality data (Henry, Lewis, and Collins 1994; Henry 1997; Henry, Spiegelman, Collins, and Park 1997). This issue has also been investigated in many other contexts. For example, in education see Kaiser (1992), and Cattell and Vogelmann (1977), in environmetrics see Juntto and Paatero (1994), in psychology see Everett (1983), and in chemistry see Malinowski (1977).

In this environmental application, the variables are a series of concentrations of airborne gases or particles measured over time, and the number of factors in the model is the number of air pollution sources impacting the sampling site. Of course, if there was no error in the measurements, and there were enough chemical compounds measured and enough observations then the rank of the correlation matrix of the measured chemical compounds would be the number of sources. Air quality data, like most environmental data, has high levels of measurement and sampling uncertainties, so estimating the number of factors by looking for a break in the eigenvalues of the correlation matrix is frequently unproductive. The smaller eigenvalues due to real sources are overwhelmed by the eigenvalues dominated by error. In addition, even if minuscule sources contribute to the data it is not likely that they can be successfully modeled.

The NUMFACT algorithm described in this paper was developed to determine the number of factors that can be seen above the noise level in the data. It does this by using a resampling technique to estimate the stability of the eigenvectors of the correlation matrix. Here we present two statistics, the original NUMFACT statistic $S$ and the modified NUMFACT statistic $MS$, and their associated cutoff values to determine the number of factors in the data that are

distinguishable from random errors. Next, we give a heuristic development of the method, followed by simulation studies comparing our methods to established procedures. Examples using air quality data are also given.

The inspiration for our method, which we call NUMFACT, is geometrical. Assume there are *q* factors in the data set with *p* variables, then the first *q* eigenvectors of the resampled data will span nearly the same *q*-dimensional subspace as the first *q* eigenvectors of the original data. This is true even though the individual eigenvectors of the resampled data may not look like the original eigenvectors or be in the same relative order. Thus, the first *q* eigenvectors of the resampled data will have a large projection on the space spanned by the original eigenvectors. However, this will not be true of the remaining *p*-*q* eigenvectors. Since these are dominated by errors, the directions of these eigenvectors are random and the projection on the space spanned by the same number of the original eigenvectors will often be small. Thus, our method is to resample the data and calculate the signal, which is the length of the *i*th eigenvector of the resampled data, as projected into the space spanned by the first *i* original eigenvectors. This is done a number of times, 40 – 50 are usually sufficient for a moderate sample size (for a very small sample size such as 50 or 60 it needs to be done many more times), and the average squared signal for each eigenvector is calculated. This is identified as the fraction of the eigenvector associated with the common variability in the data. The length of the *i*th eigenvector of the resampled data projected into the space spanned by the remaining *p*-*i* original eigenvectors is identified as the noise. The ratio of the average squared signal and the average squared noise, *W*, is a basis for defining the statistics used to estimate the number of factors in the data. Thus *W* is conceptually related to *F* statistics used in forward variable selection in regression. Our NUMFACT statistics, which we call *S* and *MS*, start with the value of *W*. We assume that for the *i*th eigenvector $W_i = \left(signal_i / noise_i\right)^2$. Furthermore, assume that the eigenvalue $l_i = signal_i + noise_i$. Then, solving these two equations for $signal_i$ and $noise_i$ gives

$$signal_i = \frac{l_i \sqrt{W_i}}{1+\sqrt{W_i}} \text{ and } noise_i = \frac{l_i}{1+\sqrt{W_i}}.$$

Then $\sum_{i=1}^{p}\left(\frac{l_i}{1+\sqrt{W_i}}\right)\Big/(p-1)$ be a reasonable estimate of the average noise level. Thus we let "*noise*" be the average for values of $noise_i$ over $i$ = 1 to *p-1*. We assume that the degrees of freedom of the average noise is *p-q-1* and "$noise_M$" be $noise_M = noise(p-1)/(p-q-1)$. Finally we calculate $S_i = signal_i/noise$ and $MS_i = signal_i/noise_M$. The asymptotic cutoff value of these statistics is 2. The value of 2 is typical for statistics of the form $\hat{\theta}/SE(\hat{\theta})$. We not only derive the asymptotic cutoff value but also show by simulation studies and the real data, that 2 is a good critical (cutoff) point. By this we mean that the number of factors is the number of eigenvectors for which $S_i$ (or $MS_i$) is greater than 2. All the results show that the new estimators work better for lognormal data than do standard tests for rank such as Bartlett's test. In addition they are competitive for normal data.

## 2. NOTATION AND DEFINITIONS

Let $X$ denote the $n \times p$ data matrix ($n$ iid observations, $p$-variate, and $\Re$ denote the population correlation matrix. The eigenvectors of population correlation matrix are denoted by $\beta_1, \ldots, \beta_p$ and the corresponding eigenvalues are $\lambda_1, \ldots, \lambda_p$. We now define sample estimates of the parameters. The sample correlation matrix is denoted as $R$. Let $b_1, \ldots, b_p$ denote the corresponding eigenvectors of $R$ and let $l_1, \ldots, l_p$ denote the corresponding eigenvalues. Next we define the corresponding bootstrap quantities. Bootstrap analogs to the sample estimates have a superscript *. For example $X^*$ (of size $n \times p$ ) denotes a bootstrap sample (drawn independently from $X$ with replacement) and $b_1^*, \ldots, b_p^*$ denote eigenvectors of the bootstrap correlation matrix $R^*$. Finally we let $N$ denote the number of independent bootstrap resamples. In particular $X_j^*$ denotes the $j$th bootstrap sample and $b_{1j}^*, \ldots, b_{pj}^*$ denote the eigenvectors of sample correlation matrix of $X_j^*$.

The NUMFACT statistics depend upon the ratios of the average squared projections of the resampled eigenvectors on the spaces spanned by the original eigenvectors, which can also be viewed as the regression sum of squares and the error sum of squares. The *i*th ratio is defined as

$$W_i = \frac{avg\left\|P\left(b_i^*: span\{b_1, \ldots, b_i\}\right)\right\|^2}{avg\left\|P\left(b_i^*: span\{b_{i+1}, \ldots, b_p\}\right)\right\|^2} = \frac{\sum_{j=1}^{N} b_{ij}^{*t}\left(b_1 b_1^t + b_2 b_2^t + \ldots + b_i b_i^t\right) b_{ij}^*}{\sum_{j=1}^{N} b_{ij}^{*t}\left(b_{i+1} b_{i+1}^t + \ldots + b_p b_p^t\right) b_{ij}^*},$$

$i = 1, \ldots, p-1$, and $W_p \equiv 0$, where *avg* denotes the average over the *N* samples and *P* denotes a projection.

The *i*th original NUMFACT statistic is denoted by $S_i$ where

$$S_i = \frac{signal_i}{noise} = \frac{\frac{l_i \sqrt{W_i}}{1+\sqrt{W_i}}}{\left(\sum_{k=1}^{p-1} \frac{l_k}{1+\sqrt{W_k}}\right) \Big/ (p-1)},$$

$i = 1, \ldots, p-1$, and $S_p = 0$,

and the *i*th modified NUMFACT statistic is denoted by $MS_i$ where

$$MS_i = \frac{signal_i}{noise_M} = \frac{\frac{l_i \sqrt{W_i}}{1+\sqrt{W_i}}}{\left(\sum_{k=1}^{p-1} \frac{l_k}{1+\sqrt{W_k}}\right) (p-q-1)},$$

$i = 1, \ldots, p-1$, and $MS_p = 0$.

Both of the original and the modified NUMFACT statistics increase with *W*. In addition they are bigger for statistics corresponding to relatively big eigenvalues.

*Remark 1.* Note that here the bootstrap samples are used to define the statistics themselves not to approximate the distribution of some statistics.

## 3. HYPOTHESES AND ASYMPTOTIC RESULTS

Let $q$ be the number of major factors. We are interested in testing a series of nested hypotheses:

$H_{0q}$: There are $q$ major factors, for example, major pollution sources. That is,

$$\lambda_1 > \lambda_2 > \ldots > \lambda_q > \lambda_{q+1} = \lambda_{q+2} = \ldots = \lambda_p,$$

where $0 \le q < p-1$.

Under the hypothesis $H_{0q}$ we expect the first $q$ $S_i$'s (or $MS_i$'s) are greater than some cutoff value and the remaining $p$-$q$ are less than it. This sequence of tests starts with $q=1$ (or $0$) and increases $q$ until a hypothesis $H_{0q}$ is accepted. Later we will derive the asymptotic cutoff points, $c_{Sq}$ and $c_{MSq}$, for $q$ source case. If $S_1$ through $S_q$ are greater than $c_{Sq}$ (or $c_{MSq}$), and $S_{q+1}$ through $S_p$ are less than or equal to $c_{Sq}$ (or $c_{MSq}$), we accept the hypothesis $H_{0q}$ that there are $q$ major sources.

Under the assumptions A1 and A2, we obtain the asymptotic values for $S_i$ and $MS_i$, as given in Result 1. The proof is given in the Appendix.

A1. If a matrix has some nondistinct eigenvalues then the corresponding eigenvectors are chosen randomly according to a uniform distribution over the permitted directions.

A2. The statistics are calculated on a computer that computes using a finite number of digits of precision. Thus if the computer has 8 digits of accuracy, 1234567.8 = 1234567.83.

Note that A1 and A2 are a work around that allow us to handle the equal eigenv case cleanly, a case that has not been solved in the theoretical literature. How equal eigenvalues are handled would vary with computer programs. Assumption A1 is a reasonable way to handle the equal eigenvalue case. Our simulation studies and real examples show that our work around allows the calculation of effective critical values for our statistic. This is true even for sample sizes that are considered moderate. We demonstrate this by simulation and scientific examples.

**Result 1:** Under $H_{0q}$: $\lambda_1 > \lambda_2 > ... > \lambda_q > \lambda_{q+1} = \lambda_{q+2} = ... = \lambda_p$ (= $\theta$), suppose $l_1 > l_2 > ... > l_q > l_{q+1} \cong l_{q+2} \cong ... \cong l_p$. Then as $N \to \infty$ and $n \to \infty$,

a1. $$S_i \xrightarrow{p} \frac{2\lambda_i(p-1)}{\theta(p-q-1)}, \qquad i = 1,\ldots,q.$$

a2. $$S_i \xrightarrow{p} \frac{\frac{2\sqrt{\frac{i-q}{p-i}}}{1+\sqrt{\frac{i-q}{p-i}}}}{\frac{p-q-1}{p-1}}, \qquad i = q+1,\ldots,p-1.$$

b1. $$MS_i \xrightarrow{p} \frac{2\lambda_i}{\theta}, \qquad i = 1,\ldots,q.$$

b2. $$MS_i \xrightarrow{p} \frac{2\sqrt{\frac{i-q}{p-i}}}{1+\sqrt{\frac{i-q}{p-i}}}, \qquad i = q+1,\ldots,p-1.$$

*Remark* 2. We know that with probability one, the ordered sample eigenvalues $\{l_i, 1 \le i \le p\}$ are distinct and positive for any finite $n$ (see Okamoto, 1973). As the sample size increases, however, it can be shown that the sample eigenvalues converge to the corresponding population eigenvalues regardless of the multiplicity of the population eigenvalues (see Henry, Park, and Spiegelman, 1997). From simulation experiments we have seen that it needs a very large value for the sample size $n$ for the near equality of the sample eigenvalues to be satisfied. Nonetheless, we shall see from our simulation study that this approximation is useful. From Table 1 we can see that the NUMFACT statistics corresponding to the equal eigenvalue case are not even close to the critical value 2. We expand upon this point later.

*Remark* 3. It seems difficult to derive the limiting distributions of the statistics rather than the asymptotic values. Both of *S* and *MS* depend on *W*, which is based on the inner product of bootstrap eigenvector and the sample eigenvector. For the eigenvector associated with the simple root, this inner product has a degenerate distribution. For the eigenvector associated with the multiple root, the distribution is unknown. Although Anderson (1963) discussed the distribution of

the eigenvectors of the covariance matrix of the normal data when there is the multiplicity among the eigenvalues, his result cannot be directly applied in our case that uses the sample correlation matrix. To the best of our knowledge this limiting distribution remains one of the open and hard problems in multivariate analysis. Note that in Result 1

$$\frac{\sqrt{\frac{j-q}{p-j}}}{1+\sqrt{\frac{j-q}{p-j}}} < 1 < \frac{\lambda_i}{\theta}, \quad i = 1,\mathrm{L}\ ,q, \quad j = q+1,\mathrm{L}\ ,p-1.$$

For the purpose of estimating the number of factors, our main asymptotic results can be restated as follows:

**Result 2:** Under $H_{0q}$: $\lambda_1 > \lambda_2 > ... > \lambda_q > \lambda_{q+1} = \lambda_{q+2} = ... = \lambda_p (= \theta)$, for large enough $n$ and $N$,

a1. $S_i > \dfrac{2(p-1)}{p-q-1}, \quad i = 1,\mathrm{L}\ ,q.$

a2. $S_i < \dfrac{2(p-1)}{p-q-1}, \quad i = q+1,\mathrm{L}\ ,p-1.$

b1. $MS_i > 2, \quad i = 1,\mathrm{L}\ ,q.$

b2. $MS_i < 2, \quad i = q+1,\mathrm{L}\ ,p-1.$

That is, all we require is that these asymptotic inequalities hold rather than the convergence of each statistic to its limiting value.

*Remark* 4. Since $\dfrac{2(p-1)}{p-q-1} \approx 2$ if $p$ is large and $q$ is small, 2 is used as an asymptotic cutoff point in $q$ source case for both of the original and the modified NUMFACT statistics.

*Remark* 5. There are some cases that the hypothesis of interest is equality of the eigenvalues of the population covariance matrix (not of the population correlation matrix), e.g., all measurements are made in the same units (Anderson 1963). If this is the case, then the NUMFACT statistics need to be calculated based on the sample covariance matrix not on the sample correlation matrix.

## 4. EVALUATION

Our asymptotic cutoff values in section 3 are examined by numerical comparisons. The data matrix *X* is generated using two different methods. Eastment and Krzanowski (1982) introduced the method of generating *n*×*p* data matrices of known structure. We first employ their method, which can be described as follows: a set of eigenvalues, $l_1, \ldots, l_p$, are selected and the square roots of the products of *n-1* and these eigenvalues are used as the diagonal elements of a diagonal matrix *T*. An *n*×*p* matrix *Y* of independent uniform entries is generated and decomposed into $USV^t$ via the singular value decomposition. The data matrix *X* to be used in the simulation is then obtained by setting $X = UTV^t$. As noted by Eastment and Krzanowski (1982), "*X* can be viewed as an observation from the set of all *n*×*p* data matrices with the required eigenvalue structure". Note that $l_i$'s are actually the sample eigenvalues, and in this case they can be forced to be equal to the population eigenvalues by using $l_1 = l_2 = \ldots = l_p = 1$ for example. Since equal sample eigenvalues may be unrealistic, we also add some perturbation to reflect more realistic sample eigenvalue pattern. Let $\varepsilon$ be the difference between the subsequent eigenvalues, i.e., $l_i - l_{i+1} = \varepsilon$ ($i = 1, \ldots, p-1$). Table 1 shows how the approximations are affected by adding some perturbation to $l_i$'s. The asymptotic values for *S* and *MS* based on the matrix $\frac{1}{n-1} X^t X$, and the sample means of the statistics over *200* replications are presented for *n = 500, p = 10, q = 0,* and $\varepsilon$ *= 0, .001, .01, .05.* Note that *MS* is the same as *S* in this case (*q = 0*). When $l_i$'s are different by only *.001,* the approximations are still very good. As $\varepsilon$ gets bigger, the deviations between asymptotic values and the sample means of the statistics get bigger, but all the sample values are still less than the cutoff value 2 for no source (*q = 0*) case. The last two columns of the table show the results for the eigenvalue pattern (of no source case) obtained from the sample correlation matrix of normal random matrix of sizes *500* by *10* and *100,000* by 10, respectively.

Secondly, we generate the data matrix $X$ by the model

$$X = AP + Error \qquad (1)$$

where $A$ and $P$ are $n \times q$ matrix and $q \times p$ matrix, respectively. We assume that the rows of $A$ are random. Note that the hypothesis $H_0$: $\lambda_1 > \lambda_2 > ... > \lambda_q = \lambda_{q+1} = \lambda_{q+2} = ... = \lambda_p$ ($\lambda_i$'s are the eigenvalues of $\Re$) is equivalent to $H_{0q}$: $\Sigma = \Psi + K$ where $\Sigma$ is the population covariance matrix, $rank(\Psi) = q$, $K = diag(k_{ii})$, and $k_{ii}$ is proportional to the $i$th diagonal element of $\Psi$ (see Anderson 1963). Under our model, $\Psi$ is the covariance matrix of the rows of $AP$. An $n \times p$ matrix $G$ of independent standard normal entries is generated, and the data matrix $X$ to be used in the simulation is obtained by setting

$$X = AP + c \cdot G \cdot diag\left(\sqrt{k_{11}}, \quad \text{L} \ , \quad \sqrt{k_{pp}}\right) \qquad (2)$$

where $c$ is a constant, $k_{ii}$'s are the diagonal elements of $P^t S_A P$, and $S_A$ is the sample covariance matrix of $A$. The asymptotic values for $S$ and $MS$ based on the sample correlation matrix $R$ and the sample means of the statistics over *200* replications are presented in Table 2 for $n = 500$, $p = 10$, and $q = 0, 2, 4$. There are deviations between asymptotic value and the sample mean for individual statistics, but the approximations for hypothesis testing are close enough. All that we require is that the statistic values associated with the simple roots are larger than the cutoff value and the statistic values associated with the multiple roots are less than the cutoff value. Our simulations support the use of our asymptotic cutoff values. When $p = 10$, the decision rule associated with any of $S$ and $MS$ works well for a relatively large range of $q$.

**{Insert Tables 1-2 here}**

## 5. COMPARISONS WITH OTHER METHODS

We compare our results to many of the number of factor estimating methods that are widely used. These are: the method of choosing enough eigenvalues to account for a suitable proportion

(say 90%) of $p = tr(R)$ (90 percent trace), the method of choosing only eigenvalues which are greater than one (Rule-of-One), Bartlett's modification (1951) of the likelihood ratio test, Malinowski's indicator function (see Malinowski 1980), and Wold's cross validation approach (1978). These methods are described in more detail in Henry, Park, and Spiegelman (1999). In this section, the above five methods, and the original and modified NUMFACT, are compared through the simulations and real data examples.

### 5.1 Simulations

Three factors are considered in generating the simulated data. The factors considered here are distribution of error (normal, lognormal), sample size (*n = 30, 60, 90, 120, 150, 200, 500, 2000*), and number of sources ($q$). For our simulation $q$ is chosen within the range $1 \le q \le q^*$ where $q^*$ is the largest integer satisfying $(p-q)^2 - p - q \ge 0$ (see *p. 565* Anderson 1984 for an example of this choice for $q^*$). Each simulation is repeated *200* times, and the Root Mean Squared Error (RMSE) of the estimator over the 200 replications is computed. The results are displayed in Figures 1-3. RMSE is calculated by

$$RMSE = \sqrt{\sum_{\hat{q}} (q - \hat{q})^2 \, p_{\hat{q}}}$$

where $p_{\hat{q}}$ represents the sample proportion that the value $\hat{q}$ is selected over 200 replications.

For each factor level combination the data is generated by model (1) $X = AP + Error$ of Section 4. In environmental application, $A$ is called a source contribution matrix and $P$ is called a source composition matrix. The number of variables (chemical species) $p$ is fixed to be at a commonly used value of 15, $q$ varies within the range $1 \le q \le 10$. The source composition matrix $P$ is obtained from the uniform random number generator in MATLAB. To avoid getting a source composition matrix with high collinearity, the condition number (the ratio of the smallest and the biggest eigenvalues of the correlation matrix of $P'$) is examined first, and $P$ is redrawn if the condition number is bigger than some threshold (15 is used here). The source composition matrix $P$ is fixed over *200* replications and the same $P$ is used for the different sample sizes or different

error distributions. The source contribution matrix *A* is regenerated at each replication using the uniform random number generator in MATLAB. As in (2), error matrices are generated so that the error variances are proportional to the systematic variances (to satisfy the hypothesis $H_{0q}$) according to Anderson (1963). Throughout the simulations the error standard deviation is about 12~20% of the model standard deviation.

Seven methods, 90 Percent Variation (denoted as *TA*), Rule-of-One (denoted as *TB*), Bartlett's method (denoted as *BA*), Malinowski's indicator function (denoted as *MA*), Wold's cross validation method (denoted as *CV*), *S*, and *MS*, are compared.

Figure 1 contains comparisons of the methods (for *n = 200, 500, 2000*) in terms of RMSE under normal errors. For a sample size *n = 2000*, CV is not included due to the computational burden of the method. The traditional methods *TA* and *TB* are perfect (RMSE is *0*) when *q* is very small like 1 or 2. As *q* increases, however, they seriously underestimate *q* and RMSE increases with *q*. The 5% level (not an overall level) Bartlett's test, *BA*, works fine in general with these sample sizes. The Malinowski's indicator function (applied to standardized data), *MA*, performs very well if *q* is moderate (RMSE is *0*). But at some point (here $q = 7$), it starts to underestimates *q*, and RMSE goes up rapidly (for $q \geq 8$, it always returns 1 as the estimate for *q*). The *CV* method works fine when *q* is small (less than 4), but the bias gets bigger as *q* gets bigger. F it seriously underestimate *q*. The *S* and *MS* methods work fine unless *q* is very large for sample sizes *n = 200, 500*. For a large *q*, they tend to underfactor but the bias is much smaller than *TA, TB, MA,* or *CV*. As the sample size gets larger, the performance of both *S* and *MS* improves. For a very small *q* such as 1 or 2, *MS* works slightly better than *S*, and for a large *q*, *S* works generally better than *MS*. When *n = 2000*, *S* performs constantly better than *BA* (RMSE for *S* is 0 in the entire range of *q*). Summarizing the result for normal error case, *S* and *MS* are comparable to the best of the traditional ones, *BA*.

When the distribution of errors is lognormal (Figure 2), *TA* and *TB* generally fail in detecting the right number of factors except when *q* is only 1 or 2 as in the case of normal error. Now *BA* completely fails regardless of the number of factors or the sample size. It always overfactors (It

selects $\hat{q}$ among the values $q+1, \ldots, p-1$ uniformly over 200 simulations.) As a result, RMSE decreases as *q* increases as opposed to all the other methods. Poor performance of *BA* is a natural consequence of violation of the normality assumption under which *BA* was developed. The *MA* method and *CV* method show almost the same performance regardless of the sample size as in the case of normal error. For lognormal error, *S* and *MS* yield significantly better results than the traditional ones. For these new statistics the results improve even more as the sample size gets bigger, which is not true for the traditional ones.

We also calculated the average RMSE (*avgRMSE*) over the range of *q* (*1 ≤ q ≤ 10*) with varying *n* (*n = 30, 60, 90, 120, 150, 200, 500, 2000* where *avgRMSE* is defined to be

$$avgRMSE = \frac{1}{10}\sum_{q=1}^{10}\sqrt{\sum_{\hat{q}}(q-\hat{q})^2 p_{\hat{q}}}\,.$$

Figure 3 contains plots of *avgRMSE* for each of seven methods under two different error distributions. From the plots we can see overall performance of each method and the effect of sample size. *TA, TB, MA,* and *CV* show high avgRMSE regardless of type of error distribution and show basically no improvement as the sample size increases. Surprisingly, *BA* does not work at all even under normal error when the sample size is as small as 30 or 60 (see Figure 3a). Note that the exact cutoff value for *BA* is unknown for correlation matrix case and as an approximation the cutoff value for covariance matrix case is used. Figure 3a indicates that this approximation could be very poor with a very small sample size. As the sample size increases, *BA* shows expected performance. Both of *S* and *MS* show much better performance as *n* increases. Although it is not shown in the plot, for *n = 2000*, avgRMSE of *S* and *MS* are 0 and 0.49, respectively (avgRMSE of *BA* is 0.39 in this case). For small sample sizes, they still do better than the traditional methods (other than *BA*) and they do not show sudden breakdown (BA does) even when *n* is as small as 30, i.e., when $n/p \le 2$.

When error distribution is lognormal (Figure 3b), none of traditional methods shows improvement as *n* increases. Also note that *BA* shows the highest avgRMSE in this case. Small

sample behavior of *S* and *MS* is better than any traditional statistics, and avgRMSE of these two methods decreases as *n* gets larger. Again for a sample size *n = 2000*, avgRMSE of *S* is 0.08, and avgRMSE of *MS* is 0.54 though they are not plotted.

**{Insert Figure 1-3 here}**

*Remark* 6. When the error distribution is lognormal, one might consider log-transforming the data before applying Bartlett's test. We found, however, the log-transformation did not help in this case. Though it is not reported in detail here, the simulation result showed no improvement over the results given in Figure 2 and Figure 3b.

*Remark* 7. In most examples scientists want to find the number of major factors and not the number of factors. For example, in pollution studies people, plants, and animals are pollution sources but typically they are minor pollution sources. A statistic that indicates additional sources for grass, and dogs in addition to major pollution sources would typically lead scientists and regulators to an unnecessarily complex model. In our experience the additional complexity leads to multicollinearity and poor model performance. For this reason we have not included the following variation of NUMFACT statistic

$$VS_i = \frac{\dfrac{l_i\sqrt{W_i}}{1+\sqrt{W_i}}}{\left(\displaystyle\sum_{k=q}^{p-1}\frac{l_k}{1+\sqrt{W_k}}\right)\ (p-q-1)}.$$

In simulation experiments it outperforms all the methods presented in this study, but in scientific data it finds too many factors and leads to too complex models. In private communication S. Wold indicated that he modified his CV procedure so that it worked better in practice but worse in simulations due to the similar reasons that we indicated above.

## 5.2 Examples

### 5.2.1 Air pollution composition data

The original data consists of 538 hourly averaged concentrations of 37 volatile organic compounds (after screening out the missing values) from the 1990 Atlanta Ozone Precursor Study

(see Henry, Lewis, and Collins 1994). It is known that there are three types of vehicle-related sources specific to Atlanta during the summertime of 1990: emissions from vehicles in motion, evaporation of whole gasoline, and gasoline headspace vapor, i.e., $q = 3$. Eight vehicle-related species our of 37 species are selected. Natural breaks in the sample eigenvalues indicate 1 or 3 factors as shown in Table 3. The 90% trace method, *TA*, gives 1 factor. The rule-of-one method, *TB*, gives 1 factor. Bartlett's chi square test, *BA*, gives 7 factors at the 5% level. Malinowski's method, *MA*, gives 4 (when applied to raw data) or 3 (when applied to standardized data), and Wold's cross validation method, *CV*, gives 1. Table 3 also shows the output of *S* and *MS*. The cutoff value for *S* and *MS* is 2. In this case, both *S* and *MS* choose 3 factors.

5.2.2 Air pollution spatial data

As the second example we consider measurements on PM2.5 (the airborne particulate matter less than 2.5 micrometers in aerodynamic diameter) collected from 11 monitoring sites in the nearby Grand Canyon National Park during the summer of 1992. The resulting data set consists of 53 observations on 11 variables (here monitoring sites). A major constituent of PM2.5 is often sulfate formed in the air by oxidation of sulfur dioxide gas. Physically, there are three known source regions of sulfur dioxide gases in the region, i.e., $q = 3$. These sources are believed to correspond to pollution sources in southern California, copper smelters in southern Arizona and northern Mexico, and electric power plants in the desert southwest. For this data, *TA* gives 4, *TB* gives 3, and *BA* gives 8, *MA* gives 2 (when applied to raw data) or 3 (when applied to standardized data), and *CV* gives 2, respectively, as the number of factors. Table 4 shows the output of *S* and *MS*. Both statistics give 3 as the number of sources.

**{Insert Tables 3-4 here}**

## 5. CONCLUSIONS AND OPEN PROBLEMS

In this paper we presented a resampling method for determining the number of major pollution sources used with success by the second author in high profile environmental applications. We

presented a base level statistical theory for it as well as for new related statistic. We showed by simulation study that the new methods are frequently better than traditional number of factors estimators for skewed data and highly competitive for normal data.

We did not address the issue of optimality of the new estimates. The important issue of how to accurately link the number of factors to the degree each factor affects the data must be addressed. The receptor modeling references by the second author and his co-authors are only examples of how this important information can be used. At this time, the successful handling at a deep level, of the variation in statistics such as *S* and *MS* in model building remains an open problem. Another crucial question is how to select variables for NUMFACT or other number of factors estimator. In many applied problems some variables have a few common factors and some have many more. If the variables used by number of factors estimator come from different sets of factors each with different number of factors the estimated number of factors is not likely to be interpretable. For environmental applications, this issue is partially addressed in the first author's dissertation, where several variable selection algorithms are developed.

## ACKNOWLEDGMENTS

Although the research described in this article has been funded in part (E.S. Park) by the United States Environmental Protection Agency through agreement CR825173-01-0 to the University of Washington, it has as not been subjected to the Agency's required peer and policy review and therefore does not necessarily reflect the views of the Agency and no official endorsement should be inferred. This Research was partially supported by Grant DMS-9523878 from the Chemistry and Statistics and Probability programs at the National Science Foundation (C.H. Spiegelman and E.S. Park).

## APPENDIX: PROOF OF RESULT 1

The result depends on the following lemmas.

*Lemma 1.* Let $\lambda_1 \geq \lambda_2 \geq ... > \lambda_i > \lambda_{i+1} \geq \lambda_{i+2} \geq ... \geq \lambda_p$ be the eigenvalues of the population correlation matrix $P$ and $\beta_1, ..., \beta_p$ be the corresponding eigenvectors. And, let $l_1 \geq l_2 \geq ... \geq l_p$ be the eigenvalues of the sample correlation matrix $R$ and $b_1, ..., b_p$ be the corresponding eigenvectors. Then for large enough $n$,

$$\sqrt{n}(b_i^* - b_i) \sim N(0, \Gamma_i), \qquad i = 1, ..., q$$

where $b_i^*$ is the $i^{th}$ eigenvector of the sample correlation matrix of $X^*$ and

$$\Gamma_i = \{ b_i^t \otimes b(l_i I - L)^+ b^t \} \Psi_n \{ b_i \otimes b(l_i I - L)^+ b^t \}$$

where $b=[b_1 \ ... \ b_p]$, $L = diag(l_1, ..., l_p)$ and

$$\Psi_n = \{I - 1/2(I+K)(I \otimes R)K_d\}(S_d^{-1/2} \otimes S_d^{-1/2})V_n(S_d^{-1/2} \otimes S_d^{-1/2})\{1/2K_d(I \otimes R)(I+K)\}$$

where $V_n = M_{4n}(x) - (vec\ S)(vec\ S)^t$, and $S$ is the sample covariance matrix.

*Proof.* It follows easily from theorem 8 (actually, from a slightly generalized version of theorem 8) of Kollo and Neudecker (1993) and the asymptotic normality of the bootstrap sample correlation matrix $R^*$.

*Lemma 2.* Under the definitions of Lemma 1,

$$\Gamma_i b_i = \mathbf{0}, \qquad i = 1, \ldots, q$$

where $\mathbf{0}$ represents $p$-dimensional zero vector.

*Lemma 3.* Under the hypothesis $H_{0q}$: $\lambda_1 > \lambda_2 > ... > \lambda_q > \lambda_{q+1} = \lambda_{q+2} = ... = \lambda_p$ $(= \theta)$,

$$E\left(b_i^t \beta_k \beta_k^t b_i\right) \xrightarrow{n \to \infty} 0, \qquad k = 1, \ldots, q, \quad i = q+1, \ldots, p$$

and

$$E\left(b_i^t \beta_k \beta_k^t b_i\right) \xrightarrow{n\to\infty} \frac{1}{p-q}, \quad k = q+1, \ldots, p, \quad i = q+1, \ldots, p.$$

*Proof.* By the consistency and orthonormality of the sample eigenvectors, we get,

$$\beta_k^t (b_1 b_1^t + \ldots + b_q b_q^t) \beta_k \xrightarrow{n\to\infty} 1, \quad k = 1, \ldots, q, \text{ in probability.}$$

This and the fact that $\beta_k^t (b_1 b_1^t + \ldots + b_p b_p^t) \beta_k \xrightarrow{n\to\infty} 1$, implies

$$\beta_k^t b_i b_i^t \beta_k \xrightarrow{n\to\infty} 0, \tag{A1}$$

$i = q+1, \ldots, p, \quad k = 1, \ldots, q$ in probability.

Thus

$$(\beta_1 \beta_1^t + \ldots + \beta_q \beta_q^t) b_i \xrightarrow{n\to\infty} 0, \tag{A2}$$

$i = q+1, \ldots, p,$ in probability.

Since $b_i^t (\beta_1 \beta_1^t + \ldots + \beta_p \beta_p^t) b_i = 1$, (A2) implies

$$b_i^t (\beta_{q+1} \beta_{q+1}^t + \ldots + \beta_p \beta_p^t) b_i \xrightarrow{n\to\infty} 1, \quad i = q+1, \ldots, p, \text{ in probability.}$$

Further

$$E\left(b_i^t \beta_k \beta_k^t b_i\right) \xrightarrow{n\to\infty} \frac{1}{p-q}, \quad k = q+1, \ldots, p, \quad i = q+1, \ldots, p$$

as there is no preferred orientation among $\beta_{q+1}$, ..., $\beta_p$ under $H_{0q}$. It follows directly from (A1) that

$$E\left(b_i^t \beta_k \beta_k^t b_i\right) \xrightarrow{n\to\infty} 0, \quad k = 1, \ldots, q, \quad i = q+1, \ldots, p.$$

*Remark A.1.* When the sample size, *n*, is extremely large, the sample eigenvalues would be nearly equal, i.e., $l_{q+1} \approx l_{q+2} \approx \ldots \approx l_p \neq 0$, and we expect the same sort of result holds for the bootstrap eigenvectors as above. That is, when $n \to \infty$,

$$E\left(b_i^{*t} b_k b_k^t b_i^* | X\right) \approx 0, \quad k = 1, \ldots, q, \quad i = q+1, \ldots, p,$$

and

$$E\left(b_i^{*t} b_k b_k^t b_i^* | X\right) \approx \frac{1}{p-q} \quad , k = q+1, \ldots, p, \quad i = q+1, \ldots, p.$$

Note that the assumption A2 in Section 3 guarantees that in the limit the equality holds.

*Lemma 4.* Under $H_{0q}$: $\lambda_1 > \lambda_2 > \ldots > \lambda_q > \lambda_{q+1} = \lambda_{q+2} = \ldots = \lambda_p$ (= $\theta$), suppose $l_1 > l_2 > \ldots > l_q > l_{q+1} \cong l_{q+2} \cong \ldots \cong l_p$. Then as $N \to \infty$ and $n \to \infty$,

(a) $W_i$ is stochastically unbounded, i.e., given any large $M > 0$,

$$P\left(|W_i| > M\right) \longrightarrow 1, \qquad i = 1, \ldots, q.$$

(b) $W_i \xrightarrow{p} \frac{i-q}{p-i}, \qquad i = q+1, \ldots, p-1.$

*Proof.*

(a) It immediately follows from Lemma 1 and Lemma 2 and the weak law of large numbers.

(b) By the weak law of large numbers and Remark A.1, we get as $N \to \infty$, $n \to \infty$,

$$W_i \xrightarrow{p} \frac{(i-q)/(p-q)}{1-(i-q)/(p-q)} = \frac{i-q}{p-i}$$

for $i = q+1, \ldots, p-1$.

*Proof of result.*

Note that the statistics $S_i$ and $MS_i$ are the continuous functions of $W_i$'s and the sample eigenvalues $l_i$'s. It is well known that the sample eigenvalues are consistent estimators of the population eigenvalues when the population eigenvalues have multiplicity 1. For the eigenvalues with multiplicity greater than 1, it can also be shown that the sample eigenvalues converge to the common root (see Henry, Park, and Spiegelman, 1997).

The results follow from direct use of the continuous mapping theorem and Lemma 4. The following lemma is useful in calculating the asymptotic cutoff values of $S_i$'s and $MS_i$'s.

*Lemma 5.* $\displaystyle\sum_{i=q+1}^{p-1} \frac{1}{1+\sqrt{\frac{i-q}{p-i}}} = \frac{p-q-1}{2}.$

*Proof.* $\displaystyle\sum_{i=q+1}^{p-1} \frac{1}{1+\sqrt{\frac{i-q}{p-i}}}$ can be rewritten after rearrangement as a sum of terms $\frac{1}{1+\sqrt{a_i}} + \frac{1}{1+\sqrt{1/a_i}} = 1$ (where $a_i = \frac{i-q}{p-i}$ when $i = q+1, ..., p-1$) and 1 (when $i = q$).

When $p-1-(q+1)+1 = p-q-1$ is even,

$$\sum_{i=q+1}^{p-1} \frac{1}{1+\sqrt{\frac{i-q}{p-i}}} = \sum_{i=q+1}^{q+\frac{p-q-1}{2}} \left( \frac{1}{1+\sqrt{a_i}} + \frac{1}{1+\sqrt{1/a_i}} \right) = \frac{p-q-1}{2}.$$

When $p-q-1$ is odd, there are $(p-q-2)/2$ ones and the middle term that occurs when $i = q + \frac{p-q-2}{2} + 1$. Thus

$$\sum_{i=q+1}^{p-1} \frac{1}{1+\sqrt{\frac{i-q}{p-i}}} = \left\{ \sum_{i=q+1}^{q+\frac{p-q-2}{2}} \left( \frac{1}{1+\sqrt{a_i}} + \frac{1}{1+\sqrt{1/a_i}} \right) \right\} + MiddleTerm.$$

The result follows from

$$Middle\ Term = \frac{1}{1+\sqrt{\frac{\left(\frac{p-q-2}{2}+q+1\right)-q}{p-\left(\frac{p-q-2}{2}+q+1\right)}}} = \frac{1}{1+\sqrt{\frac{(p-q)/2}{(p-q)/2}}} = \frac{1}{2}.$$

## REFERENCES


Anderson, T.W. (1963), "Asymptotic Theory for Principal Component Analysis," *Annals of Mathematical Statistics,* 34, 122-148.

——————— (1984), *An Introduction to Multivariate Statistical Analysis* (2nd ed.), New York: Wiley.

Bartlett, M. S. (1951), "The Effect of Standardization on a $\chi^2$ Approximation in Factor Analysis," *Biometrika* , 38, 337-344.

Cattell, R. B., and Vogelmann, S. (1977), "A comprehensive trial of the scree and KG criteria for determining the number of factors," *Multivariate Behavioral Research,* 12, 289-325.

Eastment, H. T., and Krzanowski, W. J. (1982), "Cross-Validatory Choice of the Number of Components From a Principal Component Analysis," *Technometrics,* 24, 73-77.

Everett, J. E. (1983), "Factor Comparability as a Means of Determining the Number of Factors and Their Rotation," *Multivariate Behavioral Research,* 18, 197-218.

Henry, R. C., Lewis, C. W., and Collins, J. F. (1994), "Vehicle-Related Hydrocarbon Source Composition from Ambient Data: The GRACE/SAFER Method," *Environmental Science and Technology,* 28, 823-832.

Henry, R. C. (1997), "History and Fundamentals of Multivariate Air Quality Receptor Models," *Chemometrics and Intelligent Laboratory Systems,* 37, 525-530.

Henry, R.C., Park, E.S., and Spiegelman, C.H. (1997), "Estimating the Number of Factors to Include in a Multivariate Mixture Model," Technical report, 279, Texas A&M University, Dept. of Statistics.

Henry, R.C., Park, E.S., and Spiegelman, C.H. (1999), "Comparing a New Algorithm with the Classic Methods for Estimating the Number of Factors," *Chemometrics and Intelligent Laboratory Systems,* in press.

Henry, R. C., Spiegelman, C. H., Collins, J. F., and Park, E. S. (1997), "Reported Emissions of Volatile Organic Compounds are not Consistent with Observations," *Proceedings of the National Academy of Sciences,* 94, 6596-6599.

Juntto, S., and Paatero, P. (1994), "Analysis of Daily Precipitation Data by Positive Matrix Factorization," *EnvironMetrics,* 5, 127-144.

Kaiser, H. F. (1992), "On Cliff's Formula, the Kaiser-Guttman Rule, and the Number of Factors," *Perceptual and Motor Skills,* 74, 595-598.

Kollo, T., and Neudecker, H. (1993), "Asymptotics of Eigenvalues and Unit-length Eigenvectors of Sample Variance and Correlation Matrices," *Journal of Multivariate Analysis,* 47, 283-300.

Malinowski, E. R. (1977), "Determination of the Number of Factors and the Experimental Error in a Data Matrix," *Analytical Chemistry*, 49, 612-617.

Malinowski, E. R., and Howery, D. G. (1980), *Factor Analysis in Chemistry*, John Wiley & Sons, New York.

Okamoto, M. (1973),"Distinctness of the Eigenvalues of a Quadratic Form in a Multivariate Sample," *Annals of Statistics,* 1, 763-765.

Wold, S. (1978), "Cross-Validatory Estimation of the Number of Components in Factor and Principal Components Models," *Technometrics,* 20, 397-405.

Table 1. Comparison of the asymptotic and sample mean values for $S$ (=$MS$ ) when data are generated according to Eastment and Krzanowski (1982)

$h$ (# of replications) = *200:*, $n = 500$, $p = 10$, $q = 0$

| | Asymptotic value | Sample mean[a] | Sample mean[b] | Sample mean[c] | Sample mean[d] | sample mean[e] | *sample mean*[f] |
|---|---|---|---|---|---|---|---|
| $S_1$ | 0.5000 | 0.4978 | 0.4937 | 0.6176 | 1.2673 | 1.0707 | 0.5201 |
| $S_2$ | 0.6667 | 0.6660 | 0.6750 | 0.7861 | 1.3110 | 1.1889 | 0.6999 |
| $S_3$ | 0.7913 | 0.7913 | 0.8037 | 0.8857 | 1.2764 | 1.1551 | 0.8272 |
| $S_4$ | 0.8990 | 0.8994 | 0.9100 | 0.9647 | 1.2551 | 1.1948 | 0.9280 |
| $S_5$ | 1.0000 | 1.0010 | 1.0112 | 1.0265 | 1.2281 | 1.2021 | 1.0143 |
| $S_6$ | 1.1010 | 1.1042 | 1.1050 | 1.0910 | 1.1947 | 1.2329 | 1.1047 |
| $S_7$ | 1.2087 | 1.2085 | 1.2057 | 1.1627 | 1.1659 | 1.2004 | 1.1960 |
| $S_8$ | 1.3333 | 1.3343 | 1.3258 | 1.2489 | 1.1399 | 1.2723 | 1.3066 |
| $S_9$ | 1.5000 | 1.4992 | 1.4888 | 1.3825 | 1.1531 | 1.2980 | 1.4644 |

Note: 1. $\varepsilon = l_i - l_{i+1}$, $i = 1,\ldots, p\text{-}1$; [a]$\varepsilon = 0$; [b]$\varepsilon = .001$; [c]$\varepsilon = .01$; [d]$\varepsilon = .05$.

[e]$l_1 = 1.1802$, $l_2 = 1.1476$, $l_3 = 1.0945$, $l_4 = 1.0755$, $l_5 = 1.0328$, $l_6 = 0.0095$, $l_7 = 0.9481$, $l_8 = 0.9246$, $l_9 = 0.8458$, $l_{10} = 0.7413$. This eigenvalue pattern is obtained from the sample correlation matrix of normal random matrix of size 500 by 10.

[f]$l_1 = 1.0141$, $l_2 = 1.0111$, $l_3 = 1.0091$, $l_4 = 1.0044$, $l_5 = 1.0013$, $l_6 = 0.9982$, $l_7 = 0.9936$, $l_8 = 0.9919$, $l_9 = 0.9904$, $l_{10} = 0.9858$. This eigenvalue pattern is obtained from the sample correlation matrix of normal random matrix of size 100,000 by 10.

2. $MS$ is the same as $S$ when $q = 0$.

Table 2. Comparison of the asymptotic and the sample mean values for *S*, and *MS* when data are generated according to $X = AP + Error$

$h$ (# of replications) $= 200,\ \ n = 500,\ p = 10$

| | $q = 0$ * | | $q = 2$ | | $q = 4$ | |
|---|---|---|---|---|---|---|
| | Asymptotic value | Sample mean | Asymptotic value | Sample mean | Asymptotic value | Sample mean |
| $S_1$ | 0.5000 | 1.3225 | 147.1440 | 117.8816 | 192.4710 | 104.8975 |
| $S_2$ | 0.6667 | 1.2468 | 29.4275 | 22.7918 | 31.6392 | 16.1102 |
| $S_3$ | 0.7913 | 1.2221 | 0.7053 | 1.2954 | 19.9085 | 10.1393 |
| $S_4$ | 0.8990 | 1.2021 | 0.9412 | 1.2263 | 10.3813 | 5.2264 |
| $S_5$ | 1.0000 | 1.1841 | 1.1224 | 1.1884 | 1.1125 | 1.2585 |
| $S_6$ | 1.1010 | 1.1752 | 1.2857 | 1.1579 | 1.4912 | 1.1829 |
| $S_7$ | 1.2087 | 1.1625 | 1.4490 | 1.1423 | 1.8000 | 1.1538 |
| $S_8$ | 1.3333 | 1.1542 | 1.6302 | 1.1395 | 2.1088 | 1.1537 |
| $S_9$ | 1.5000 | 1.2012 | 1.8661 | 1.1641 | 2.4875 | 1.1719 |
| $MS_1$ | 0.5000 | 1.3225 | 114.4453 | 91.6857 | 106.9283 | 58.2764 |
| $MS_2$ | 0.6667 | 1.2468 | 22.8880 | 17.7270 | 17.5773 | 8.9501 |
| $MS_3$ | 0.7913 | 1.2221 | 0.5486 | 1.0075 | 11.0603 | 5.6329 |
| $MS_4$ | 0.8990 | 1.2021 | 0.7321 | 0.9538 | 5.7674 | 2.9036 |
| $MS_5$ | 1.0000 | 1.1841 | 0.8730 | 0.9243 | 0.6180 | 0.6992 |
| $MS_6$ | 1.1010 | 1.1752 | 1.0000 | 0.9006 | 0.8284 | 0.6572 |
| $MS_7$ | 1.2087 | 1.1625 | 1.1270 | 0.8885 | 1.0000 | 0.6410 |
| $MS_8$ | 1.3333 | 1.1542 | 1.2679 | 0.8863 | 1.1716 | 0.6409 |
| $MS_9$ | 1.5000 | 1.2012 | 1.4514 | 0.9054 | 1.3820 | 0.6511 |

* When $q = 0$, $X = Error =$ *Normal random matrix* of size 500 by 10.

Table 3. Atlanta air pollution composition data

| *Number* | *Eigenvalue* | *S* | *MS* |
|---|---|---|---|
| 1 | 7.5054 | 520.8791 | 297.6452 |
| 2 | 0.2448 | 14.1598 | 8.0913 |
| 3 | 0.1623 | 10.6843 | 6.1053 |
| 4 | 0.0309 | 1.2351 | 0.7058 |
| 5 | 0.0296 | 1.8425 | 1.0529 |
| 6 | 0.0140 | 0.6655 | 0.3803 |
| 7 | 0.0105 | 0.6986 | 0.3992 |
| 8 | 0.0027 | 0 | 0 |

NOTE: The data consists of 538 observations on 8 chemical compounds. The original NUMFACT statistic, $S$, with cutoff value 2 gives 3 sources; The modified NUMFACT statistic, $MS$, with cut-off value 2 gives 3 sources; The Malinowski's indicator function (applied to raw data) gives 4; The Malinowski's indicator function (applied to standardized data) gives 3; The cross validation approach gives 1 (for both standardized data and raw data); Bartlett's test gives 7 sources at the 5% level; The rule-of-one gives 1 sources; The 90% trace method gives 1 source.

4. Air pollution spatial data

| *Number* | *Eigenvalue* | *S* | *MS* |
|---|---|---|---|
| 1 | 6.5176 | 31.7507 | 22.2255 |
| 2 | 1.9843 | 8.0852 | 5.6596 |
| 3 | 1.2352 | 5.3748 | 3.7624 |
| 4 | 0.3300 | 0.9588 | 0.6712 |
| 5 | 0.2769 | 0.7287 | 0.5101 |
| 6 | 0.2211 | 0.6865 | 0.4805 |
| 7 | 0.1983 | 0.7297 | 0.5108 |
| 8 | 0.1126 | 0.4088 | 0.2861 |
| 9 | 0.0604 | 0.2190 | 0.1533 |
| 10 | 0.0354 | 0.1169 | 0.0818 |
| 11 | 0.0281 | 0 | 0 |

NOTE: The data consists of 53 observations on 11 variables. The original NUMFACT statistic, $S$, with cutoff value 2 gives 3 sources; The modified NUMFACT statistic, $MS$, with cut-off value 2 gives 3 sources; The Malinowski's indicator function (applied to raw data) gives 2; The Malinowski's indicator function (applied to standardardized data) gives 3 sources; The cross validation approach (applied to standardized data) gives 2 sources; Bartlett's test gives 8 sources at the 5% level; The rule-of-one gives 3 sources; The 90% of trace method gives 4 sources.

**Figure Titles and Legends**

Figure 1. Root Mean Squared Error (RMSE) of traditional methods (*TA, TB, BA, MA, CV*) and NUMFACT statistics (*S, MS*) based on *200* replications when error distribution is normal for sample sizes $n = 200$ (Figure 1a), *500* (Figure 1b), and *2000* (Figure 1c). The lines are interpolations between symbols that correspond to RMSE. The CV method is not included in Figure 1c due to computational burden to implement it.

Figure 2. Root Mean Squared Error (RMSE) of traditional methods (*TA, TB, BA, MA, CV*) and NUMFACT statistics (*S, MS*) based on *200* replications when error distribution is lognormal for sample sizes $n = 200$ (Figure 2a), *500* (Figure 2b), and *2000* (Figure 2c). The lines are interpolations between symbols that correspond to RMSE. The CV method is not included in Figure 2c due to computational burden to implement it.

Figure 3. Average Root Mean Squared Error (avgRMSE) of traditional methods (*TA, TB, BA, MA, CV*) and NUMFACT statistics (*S, MS*) over the range of $q$ ($1 \le q \le 10$) with varying $n$ ($n = 30, 60, 90, 120, 150, 200, 500$), based on *200* replications. The lines are interpolations between symbols that correspond to avgRMSE.

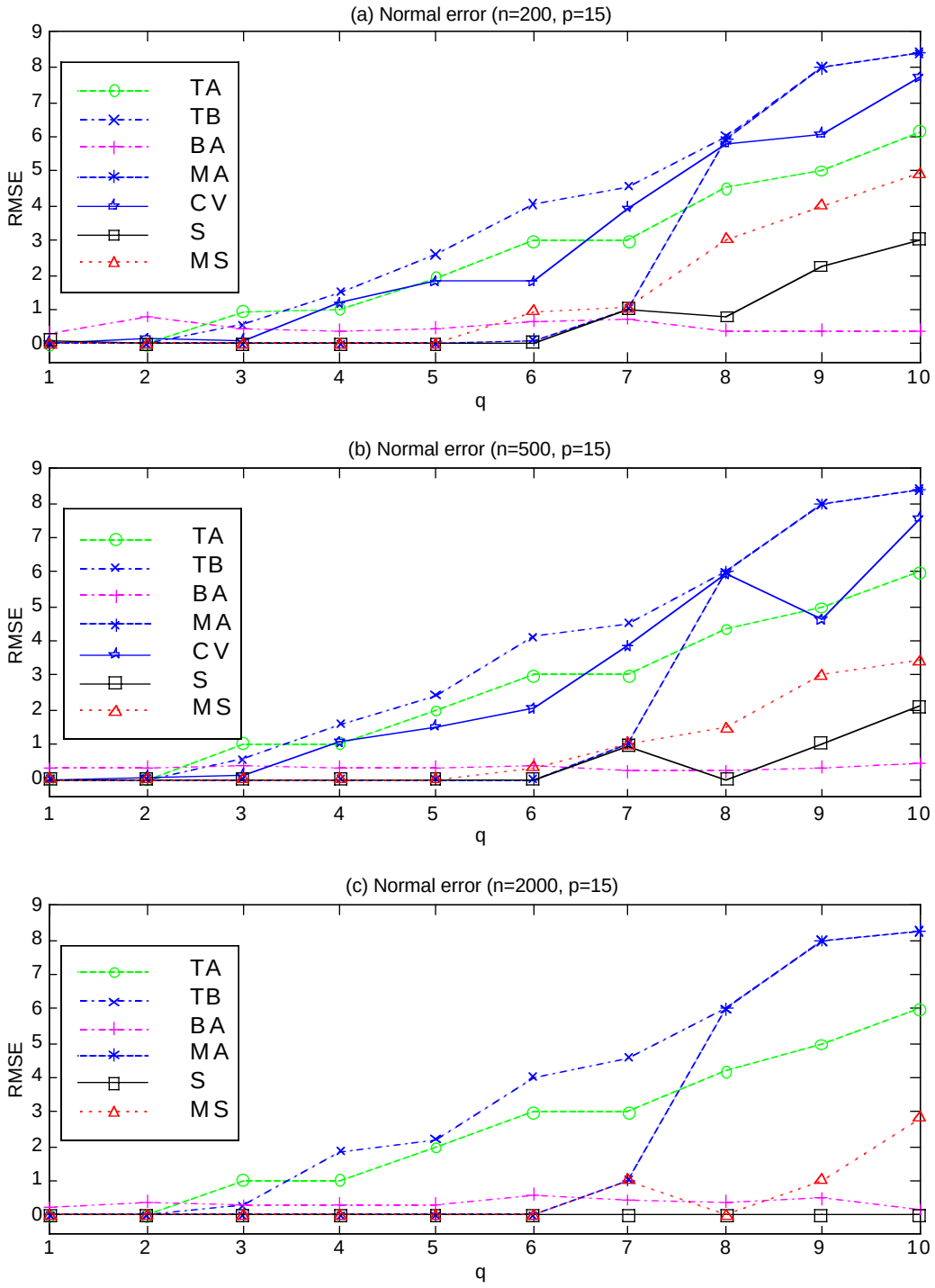

(a) Normal error (n=200, p=15)
RMSE
q
TA
TB
BA
MA
CV
S
MS
(b) Normal error (n=500, p=15)
RMSE
q
TA
TB
BA
MA
CV
S
MS
(c) Normal error (n=2000, p=15)
RMSE
q
TA
TB
BA
MA
S
MS

Figure 1

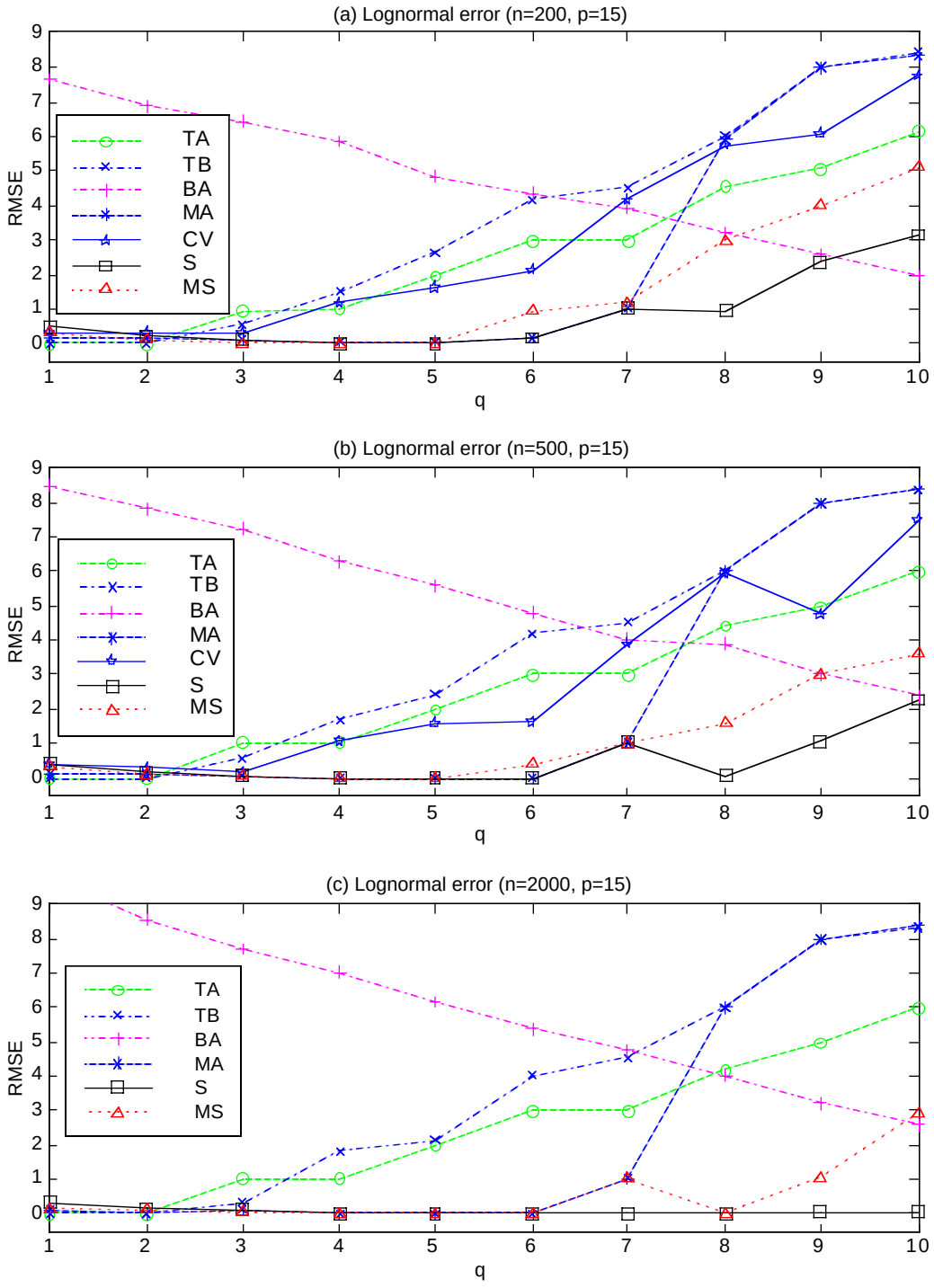
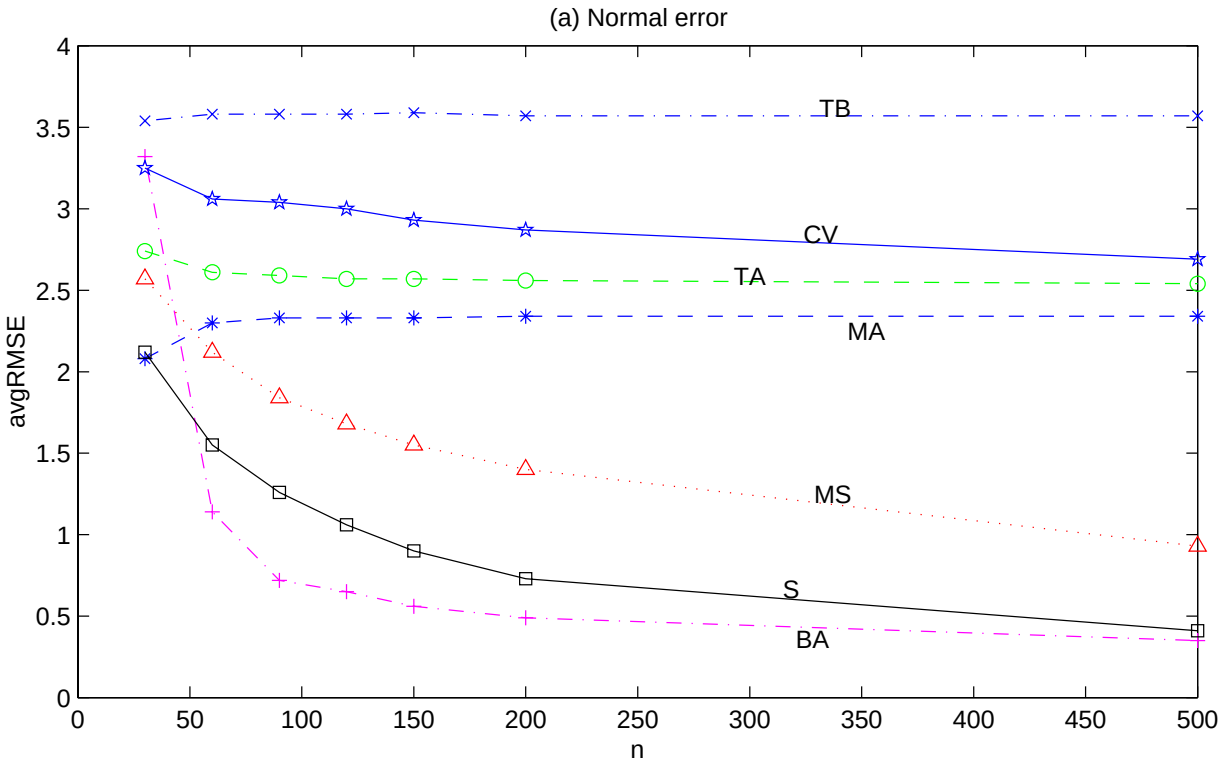

(a) Lognormal error (n=200, p=15)
RMSE
q
TA
TB
BA
MA
CV
S
MS
(b) Lognormal error (n=500, p=15)
RMSE
q
TA
TB
BA
MA
CV
S
MS
(c) Lognormal error (n=2000, p=15)
RMSE
q
TA
TB
BA
MA
S
MS

Figure 2


(a) Normal error
avgRMSE
n
TB
CV
TA
MA
MS
S
BA
0
0.5
1
1.5
2
2.5
3
3.5
4
0
50
100
150
200
250
300
350
400
450
500

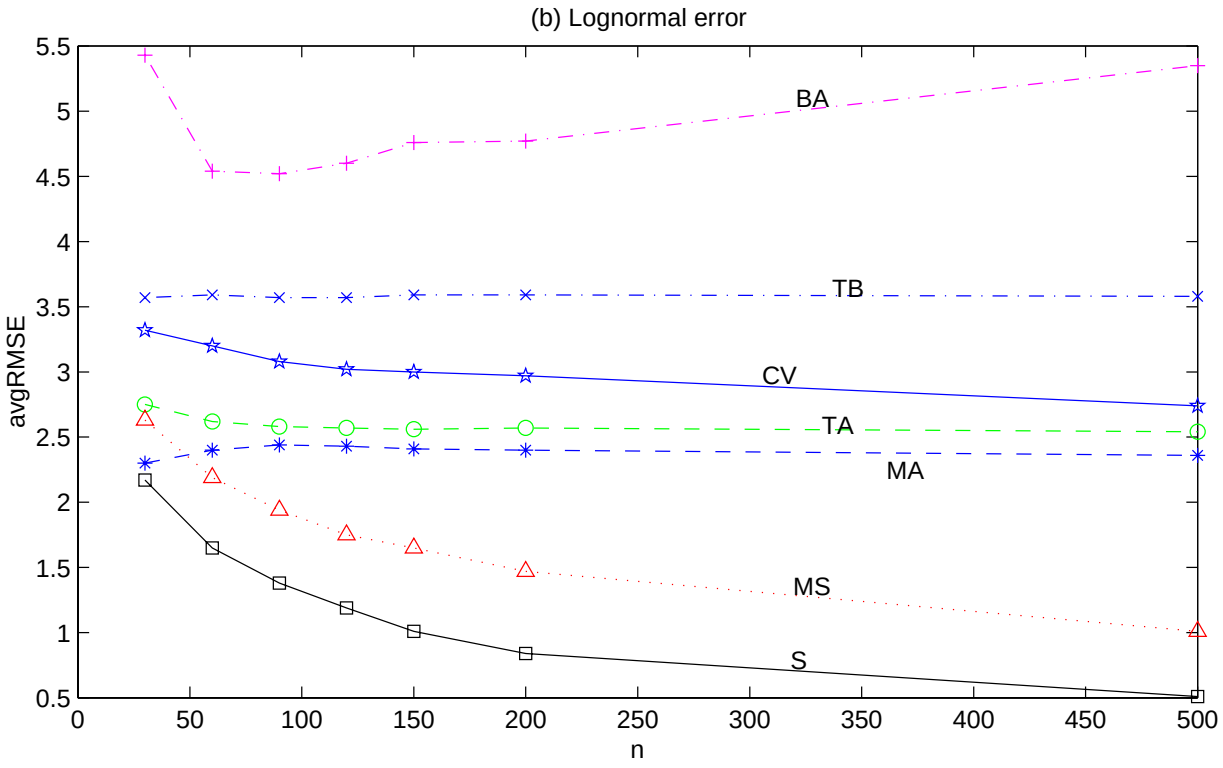

(b) Lognormal error
avgRMSE
n
BA
TB
CV
TA
MA
MS
S
0.5
1
1.5
2
2.5
3
3.5
4
4.5
5
5.5
0
50
100
150
200
250
300
350
400
450
500

Figure 3