

# Use of Latent Variable Models in Air Quality Monitoring

William F. Christensen and Stephan R. Sain

Biostatistics Center  
Department of Statistical Science  
Southern Methodist University  
Dallas, Texas 75275-0332

## Abstract

Latent variable analysis is a statistical approach for modeling the underlying structure in multivariate data in terms of a smaller number of latent variables or factors. In the environmental sciences, factor analytic models such as the multivariate receptor model have been used to assess the number of pollution sources affecting the air quality at a monitoring site. Because air quality data often exhibit temporal and/or spatial dependence, we consider the importance of accounting for such correlation in estimating model parameters and making valid statistical inferences. For the multivariate receptor model and latent variable models in general, we propose a new approach for incorporating dependence structure directly into estimation and inference via a nested block bootstrap approach. The method requires no assumptions about data distributions, priors, or forms of covariance functions, and it avoids many of the complications associated with the modeling of multivariate correlated data. The application of the approach is facilitated by a new multivariate extension of an existing block size determination algorithm. The proposed approaches are evaluated by simulation. An analysis of hourly measurements of volatile organic compounds in the El Paso, Texas/Ciudad Juarez, Mexico area yields the conclusion that the compounds are products of two underlying pollution sources, one associated with auto exhaust and the other associated with industrial emissions.

## 1 Introduction

In recent years, there has been an increasing interest in air quality monitoring. Specifically, researchers are interested in identifying the sources and compositions of pollutants in order to implement air pollution control programs. Rather than directly observing the quantity of various pollutants emitted from all potential pollu-

tion sources (which is usually impossible), receptor models are used to analyze concentrations of airborne gases or particles measured over time in order to gain insight about the unobserved pollution sources. For example, Henry, Lewis, and Collins (1994) use hourly measurements of 37 volatile organic compounds in order to describe the composition of the three vehicle-related volatile hydrocarbon sources in Atlanta. Koutrakis and Spengler (1987) concluded that the ambient particle data on 18 elements measured in Steubenville, Ohio were driven by six sources. Similar analyses have been carried out by Alpert and Hopke (1980), Thurston and Spengler (1985), Spiegelman and Dattner (1993), Park, Henry, and Spiegelman (1999), Park, Spiegelman, and Henry (1999), Park, Guttorp, and Henry (1999), and others.

As in the closely related factor analysis models, the choice of the number of pollution sources ( $k$ ) used in receptor models is pivotal. It is generally assumed known or is chosen using one of many methods (some ad hoc) advocated in the literature. A discussion of several of these methods can be found in Park, Henry, and Spiegelman (1999). After fitting the  $k$ -source model, interest often lies in describing the primary pollutants emitted from each source. At this point, inference about the estimated pollutant concentrations is natural, but statistical inferential tools for such data has not received much discussion in the literature. Because of the dependence often exhibited by pollution data collected over time and/or space, the standard  $\chi^2$  goodness-of-fit statistic as well as the usual procedures for doing inference for model parameters are often invalid. In this paper, we present an approach for incorporating dependence into a factor analysis through the use of the block bootstrap and then illustrate the usefulness of this procedure in multivariate receptor modeling using data from the El Paso, Texas/Ciudad Juarez, Mexico area.

## 2 Factor Analysis and Receptor Models

Latent variable modeling is a statistical approach for modeling the underlying structure in multivariate data in terms of a smaller number of latent variables or factors. Once used predominantly in the social and behavioral sciences where abstract quantities such as feelings, aptitudes, and intelligence are difficult to measure, latent variable models are now widely used in many disciplines including the physical and environmental sciences. The classical latent variable model is used to describe the behavior of a random sample of  $p$ -variate observations  $\mathbf{X}_i$ ,  $i = 1, \dots, n$ . The latent variable model states that  $\mathbf{X}_i$  depends on an unobserved factor vector  $\mathbf{f}_i$ , where the length of factor vector ( $k$ ) is less than  $p$ . The latent variable model allows one to explore the relationships among the elements of  $\mathbf{X}_i$ , looking for underlying structure. A model in the usual statistical sense, one can formulate a specific hypothesis about the nature of the relationship among the variables and then use inferential tools to assess the model's goodness-of-fit, construct confidence intervals for model parameters, or conduct relevant hypothesis tests. Unlike purely exploratory multivariate techniques, latent variable modeling allows the research to incorporate subject matter knowledge directly into the model in order to give the fitted model interpretable meaning.

The simplest latent variable model is the linear factor analysis model

$$\mathbf{X}_i = \boldsymbol{\gamma} + \mathbf{\Lambda}\mathbf{f}_i + \mathbf{e}_i, \quad i = 1, \dots, n, \quad (1)$$

where  $\boldsymbol{\gamma}$  is an intercept,  $\mathbf{\Lambda}$  is a  $p \times k$  matrix of factor coefficients or factor loadings, and  $\mathbf{e}_i$  is a  $p$ -vector of error terms. Though the model written in this form is not uniquely identified, one can use subject matter knowledge in order to constrain some of the factor loadings in  $\mathbf{\Lambda}$  to be constants (e.g., zeros and ones) so that the factors cannot be transformed without altering the model form. An oft-used parameterization sets  $k$  of the observed variables to be equal to one of the factors plus its error term, essentially putting the rows of a  $k \times k$  identity matrix in  $k$  of the rows of  $\mathbf{\Lambda}$ . The errors are generally considered to be independent of each other and independent of the factors, but we allow the factors to be correlated since this parameterization is most useful in practice. Thus,

$$\begin{aligned} \text{var}(\mathbf{e}_i) &= \boldsymbol{\Psi} = \begin{pmatrix} \psi_{11} & & 0 \\ & \ddots & \\ 0 & & \psi_{pp} \end{pmatrix} \\ \text{var}(\mathbf{f}_i) &= \boldsymbol{\Phi}. \end{aligned}$$

Letting  $\boldsymbol{\lambda}$  be the non-constant elements of  $\mathbf{\Lambda}$ , we can define a parameter vector  $\boldsymbol{\theta} = (\boldsymbol{\lambda}', (\text{vech } \boldsymbol{\Phi})', \text{diag}(\boldsymbol{\Psi})')'$ , where “vech  $\boldsymbol{\Phi}$ ” is a vector containing the  $k(k+1)/2$  unique elements of the symmetric matrix  $\boldsymbol{\Phi}$ . Given  $\boldsymbol{\theta}$ , we note that the covariance matrix under the model is  $\boldsymbol{\Sigma}(\boldsymbol{\theta}) = \mathbf{\Lambda}\boldsymbol{\Phi}\mathbf{\Lambda}' + \boldsymbol{\Psi}$ . Estimates of  $\boldsymbol{\theta}$  will then be values that in some sense minimizes the difference between  $\boldsymbol{\Sigma}(\boldsymbol{\theta})$  and  $\mathbf{S}$ , the sample covariance matrix for  $\mathbf{X}$ .

We now compare the factor analysis model (1) with the receptor model used in the literature:

$$\mathbf{X}_t = \mathbf{\Lambda}\mathbf{f}_t + \mathbf{e}_t, \quad t = 1, \dots, n, \quad (2)$$

where  $\mathbf{X}_t$  is a  $p$ -vector of observed concentrations of ambient particles observed at time  $t$ ,  $\mathbf{\Lambda}$  is a  $p \times k$  matrix of non-negative source compositions, and  $\mathbf{f}_t$  is a  $k$ -vector of non-negative pollution source contributions. Note that the  $k$ th column of  $\mathbf{\Lambda}$  consists of the proportions of the  $k$ th pollution source that correspond to the  $p$  ambient particles. Thus, each column of  $\mathbf{\Lambda}$  sums to no more than one. For example, if the first element in  $\mathbf{X}_t$  is acetylene and there are two pollution sources (auto exhaust and industrial emissions), the receptor model states that  $X_{1t}$ , the concentration of acetylene in the atmosphere at time  $t$ , can be written

$$\begin{aligned} X_{1t} &= \lambda_{11}f_{1t} + \lambda_{12}f_{2t} + e_{1t} \\ &= [\% \text{ acetylene in auto exhaust}] \times \\ &\quad [\text{concentration of auto exhaust in atmosphere}] \\ &\quad + [\% \text{ acetylene in industrial emissions}] \times \\ &\quad [\text{concentration of industrial emissions in atmos.}] \\ &\quad + e_{1t}. \end{aligned}$$

When  $\mathbf{\Lambda}$  is known, the literature usually refers to (2) as the *chemical mass balance model*. The pollution source contributions can then be estimated using regression. However, it is more often the case that many of the elements of the source composition matrix  $\mathbf{\Lambda}$  need to be estimated along with the pollution source contributions  $\mathbf{f}_t$ ,  $t = 1, \dots, n$ . In this situation, (2) is called a *multivariate receptor model* and its similarities with (1) are many. Much of the multivariate receptor modeling studies in the literature use factor analytic techniques to identify the number of pollution sources, the pollution source compositions, and the pollution source contributions.

## 3 Temporal Dependence

The major difference between the use of multivariate receptor models in the literature and the use of classical

linear factor analysis models is that the  $n$  observations in a pollution data set are rarely if ever from a random sample. Rather, multivariate receptor models are used to model data that exhibit temporal and/or spatial dependence. There are several potential hazards of ignoring dependence structure when carrying out a factor analysis. First, after the model has been fitted, one can predict the values of the factor process (the “factor scores”) for each  $t$ . When temporal dependence is ignored, the predicted factor process will often fail to exhibit temporal continuity. Second, parameter estimation may be biased and inefficient. Third, inferential techniques commonly associated with factor analysis (including goodness-of-fit tests, hypothesis tests, and confidence intervals for model parameters) will be invalid. In the present discussion, we focus on the effect of dependence on inference about the model parameters.

Throughout this paper, we consider a simulated 8-variate time series of length 300 which depends upon 2 temporally-correlated factors according to the model

$$\begin{aligned}
 x_{1t} &= 0.5f_{1t} && + e_{1t} \\
 x_{2t} &= 1.5f_{1t} && + e_{2t} \\
 x_{3t} &= && 0.5f_{2t} + e_{3t} \\
 x_{4t} &= && 1.5f_{2t} + e_{4t} \\
 x_{5t} &= 0.25f_{1t} + 0.25f_{2t} && + e_{5t} \\
 x_{6t} &= 0.75f_{1t} + 0.75f_{2t} && + e_{6t} \\
 x_{7t} &= f_{1t} && + e_{7t} \\
 x_{8t} &= && f_{2t} + e_{8t},
 \end{aligned} \tag{3}$$

$t = 1, \dots, 300$ , where each of  $f_{1t}, f_{2t}, e_{1t}, \dots, e_{8t}$  is a lognormal AR(1) process with autoregressive coefficient  $\phi_1 = 0.6$ . Ignoring the dependence in the data, the following model was fit using pseudo-independent, pseudo-normal maximum likelihood:

$$\begin{aligned}
 x_{1t} &= \lambda_{11}f_{1t} + \lambda_{12}f_{2t} + e_{1t} \\
 x_{2t} &= \lambda_{21}f_{1t} + \lambda_{22}f_{2t} + e_{2t} \\
 x_{3t} &= \lambda_{31}f_{1t} + \lambda_{32}f_{2t} + e_{3t} \\
 x_{4t} &= \lambda_{41}f_{1t} + \lambda_{42}f_{2t} + e_{4t} \\
 x_{5t} &= \lambda_{51}f_{1t} + \lambda_{52}f_{2t} + e_{5t} \\
 x_{6t} &= \lambda_{61}f_{1t} + \lambda_{62}f_{2t} + e_{6t} \\
 x_{7t} &= f_{1t} && + e_{7t} \\
 x_{8t} &= && f_{2t} + e_{8t}.
 \end{aligned} \tag{4}$$

For each parameter in the model (factor loadings, factor variances and covariance, and error variances), a 95% confidence interval was created and the percent of replications in which the parameter was contained in the interval was calculated. Average coverage probabilities for each group of parameters is given in Table 1 as is the Type I Error for the  $\chi^2$  goodness-of-fit test. Note that the average coverage probability for the factor loadings (of primary interest in most studies) is only 83%. Additionally, 68% of the time the goodness-of-fit test rejects

Table 1: Ignoring temporal dependence—average coverage probability for each set of parameters for a nominal 95% confidence interval, and Type I Error rate for  $\chi^2$  goodness-of-fit test.

Parameters	Average coverage probability (nominal 95%)	Type I error for $\chi^2$ GOF test (nominal 5%)
$\lambda$	83%	
$\Phi$	45%	68%
$\Psi$	60%	

the model from which the data was generated. Thus, positive dependence in the data will erroneously indicate that more factors are needed in the model.

The inadequacy of the  $\chi^2$  goodness-of-fit test when data exhibit dependence has been noted, but the source of the problem is often not explored further. In a study of social interactions over time, the authors noted that “a common problem with factor analytic studies of intraindividual variability across time is the occurrence of spurious covariances between adjacent measurement periods. This problem of excessive serial covariance leads to the appearance of more factors than are actually the case. Thus, the chi-square...will not be as informative” (Hershberger, Corneal, and Molenaar, 1994, p. 41). The authors then recommend other model fit diagnostics. Park, Henry, and Spiegelman (1999) compare the  $\chi^2$  goodness-of-fit test with 7 other model-fit diagnostics for selecting the number of factors to retain in two different factor analysis models of pollution data (collected in Atlanta and the Grand Canyon National Park). For both factor analysis models, the  $\chi^2$  goodness-of-fit test indicated the need for more factors than all of the other model-fit diagnostics considered. However, the dependence in the pollution data and its effect on the  $\chi^2$  goodness-of-fit test is never addressed.

The effect of temporal or spatial dependence on estimation and inference can be understood by an examination of the generalized least squares (GLS) estimator of  $\theta$ , the parameter vector for the factor analysis model. In Anderson and Amemiya (1988) and Amemiya and Anderson (1990), the normal maximum likelihood and GLS estimators based on a random sample of observations are shown to be consistent, and have asymptotic normal distributions, even when factors and errors are nonnormal. Additionally, associated  $\chi^2$  goodness-of-fit statistics are shown to have asymptotic  $\chi^2$  distributions under the same general conditions. A GLS estimator can

be obtained by minimizing

$$SS(\boldsymbol{\theta}) = n(\text{vech } \mathbf{S} - \text{vech } \boldsymbol{\Sigma}(\boldsymbol{\theta}))' \hat{\boldsymbol{\Gamma}}^{-1} (\text{vech } \mathbf{S} - \text{vech } \boldsymbol{\Sigma}(\boldsymbol{\theta})) \quad (5)$$

where

$$\boldsymbol{\Gamma} = n \text{Var}(\text{vech } \mathbf{S}). \quad (6)$$

Since the goodness of fit statistic  $SS(\hat{\boldsymbol{\theta}})$  in (5) and the covariance matrix for  $\hat{\boldsymbol{\theta}}$  both depend on  $\boldsymbol{\Gamma}$ , proper estimation of this fourth-moment matrix is crucial. When observations are independent and normal, then an unbiased estimator  $\hat{\boldsymbol{\Gamma}}_N$  for  $\boldsymbol{\Gamma}$  can be obtained as a simple function of sample second moments (see Fuller, 1987, pp. 386-387). However, when the data are not independent, both this normality-based fourth moment matrix and the sample fourth moment matrix  $\hat{\boldsymbol{\Gamma}}_S$  are biased for  $\boldsymbol{\Gamma}$ . For example, in the presence of positive spatial or temporal dependence among observations, the diagonal elements of  $\hat{\boldsymbol{\Gamma}}_N$  and  $\hat{\boldsymbol{\Gamma}}_S$  will underestimate the diagonal elements of  $\boldsymbol{\Gamma}$ , leading to deflated standard errors for the elements of  $\hat{\boldsymbol{\theta}}$  and inflated values for the goodness-of-fit statistic  $SS(\hat{\boldsymbol{\theta}})$ .

Notwithstanding the hazards of ignoring temporal and spatial dependence, the independence assumption is implicitly assumed in almost every study involving receptor modeling. One noteworthy exception is Park, Guttorp, and Henry (2000) which incorporates temporal dependence structure directly into a hierarchical model and then estimates the model parameters (including factor process auto-regressive coefficients) using an MCMC method.

We propose an alternative approach to the sophisticated hierarchical modeling of Park, Guttorp, and Henry (2000) which is more exploratory in nature. We use the simple multivariate receptor/factor analysis model (2) which can be fit using existing software packages. An advantage of our approach is that it requires no assumptions about distributional forms, prior distributions for parameters, or forms of covariance functions. We account for dependence structure by assuming only second-order stationary for the temporal/spatial process and then using an estimate of  $\boldsymbol{\Gamma}$  in (6) for estimation and inference which is approximately unbiased. A version of the block bootstrap is used to estimate  $\boldsymbol{\Gamma}$ . We favor a bootstrapped estimate of  $\boldsymbol{\Gamma}$  to an estimate based on models of the  $p^2$  (cross-)covariograms for the multivariate process because of the problems associated with ensuring the positive definiteness of  $\boldsymbol{\Gamma}$ . (For a discussion of the problems associated with modeling and prediction of multivariate spatial processes, see Ver Hoef and Barry, 1998). Additionally, the nonparametric block bootstrap requires no assumptions about covariance forms.

In the following sections, we present a version of the

block bootstrap which is appropriate for the estimation of moment matrices for dependent data and propose a new multivariate extension of the optimal block size selection algorithm originally proposed for univariate processes by Hall, Horowitz, and Jing (1995). Though the approach can be applied to spatially-correlated data as well as temporally-correlated data, we henceforth focus our discussion on the multivariate time series scenario for simplicity. We conclude with an illustration of multivariate receptor modeling of volatile organic compounds from the El Paso/Ciudad Juarez area.

## 4 The Block Bootstrap

The basic principle underlying the block bootstrap for dependent observations is that by resampling contiguous blocks of data of length  $\ell$ , then pasting the resampled blocks together to form a bootstrap replicate, one can preserve much of the dependence properties of the original data in each bootstrap replicate. One can then explore the distribution of a function of the data by observing the function of interest as it is applied to each of the bootstrap replicates. Consider a univariate time series

$$\mathbf{x} = \{x_1 \ x_2 \ x_3 \ x_4 \ x_5 \ x_6 \ x_7 \ x_8 \ x_9 \ x_{10} \ x_{11} \ x_{12}\}$$

and a function  $g(\mathbf{x})$ . We may wish to estimate  $\text{Var}(g(\mathbf{x}))$ . One could use  $B$  standard (block size  $\ell = 1$ ) bootstrap replicates

$$\mathbf{x}_1^* = \{x_5 \ x_3 \ x_6 \ x_{12} \ x_9 \ x_3 \ x_8 \ x_5 \ x_{10} \ x_3 \ x_1 \ x_5\}$$

⋮

$$\mathbf{x}_B^* = \{x_5 \ x_{12} \ x_5 \ x_{12} \ x_2 \ x_3 \ x_4 \ x_7 \ x_9 \ x_7 \ x_8 \ x_2\}$$

to calculate  $\bar{g}^* = \frac{1}{B} \sum_{i=1}^B g(\mathbf{x}_i^*)$  and  $\widehat{\text{Var}}(g(\mathbf{x})) = \frac{1}{B-1} \sum_{i=1}^B (g(\mathbf{x}_i^*) - \bar{g}^*)^2$ . But if positive dependence is exhibited in  $\mathbf{x}$ , then  $\widehat{\text{Var}}(g(\mathbf{x}))$  will underestimate  $\text{Var}(g(\mathbf{x}))$ .

Instead of using the standard bootstrap (resampling blocks of length  $\ell = 1$ ) as above, one could resample blocks  $\mathbf{y}_j$  of length  $\ell = 3$  from

$$\mathbf{x} = \underbrace{\{x_1 \ x_2 \ x_3\}}_{\mathbf{y}_1} \underbrace{\{x_4 \ x_5 \ x_6\}}_{\mathbf{y}_2} \underbrace{\{x_7 \ x_8 \ x_9\}}_{\mathbf{y}_3} \underbrace{\{x_{10} \ x_{11} \ x_{12}\}}_{\mathbf{y}_4}$$

to create bootstrap replicates

$$\mathbf{x}_1^* = \underbrace{\{x_4 \ x_5 \ x_6\}}_{\mathbf{y}_2} \underbrace{\{x_7 \ x_8 \ x_9\}}_{\mathbf{y}_3} \underbrace{\{x_4 \ x_5 \ x_6\}}_{\mathbf{y}_2} \underbrace{\{x_{10} \ x_{11} \ x_{12}\}}_{\mathbf{y}_4}$$

⋮

$$\mathbf{x}_B^* = \underbrace{\{x_{10} \ x_{11} \ x_{12}\}}_{\mathbf{y}_4} \underbrace{\{x_1 \ x_2 \ x_3\}}_{\mathbf{y}_1} \underbrace{\{x_7 \ x_8 \ x_9\}}_{\mathbf{y}_3} \underbrace{\{x_1 \ x_2 \ x_3\}}_{\mathbf{y}_1}.$$

If the block size  $\ell$  is large enough, the dependence in the original data will be sufficiently preserved in each bootstrap replicate. In fact, for the optimal choice of  $\ell$ ,  $\frac{1}{B-1} \sum_{i=1}^B (g(\mathbf{x}_i^*) - \bar{g}^*)^2 \cong \text{Var}(g(\mathbf{x}))$ .

Block resampling for spatial data was proposed by Hall (1985). Carlstein (1986) and Künch (1989) recommended block resampling for time series data using nonoverlapping blocks and overlapping blocks, respectively. Politis and Romano (1994) proposed the stationary bootstrap which creates bootstrap replicates by randomly drawing an observation  $x_k$  from the data  $\mathbf{x} = (x_1, \dots, x_n)$  and defining  $x_1^* = x_k$ . Then,

$$x_2^* = \begin{cases} x_{k+1} & \text{w.p. } 1-p \\ \text{random draw from } \mathbf{x} & \text{w.p. } p, \end{cases}$$

and  $x_3^*, \dots, x_n^*$  are defined in similar fashion. When  $p = 1/\ell$ , the mean block length for the stationary bootstrap ( $\ell$ ) is equal to the block length under the Carlstein (1986) and Künch (1989) resampling rules.

Our purpose in using block resampling is to obtain an estimate of  $\mathbf{\Gamma} = n \text{Var}(\text{vech } \mathbf{S})$  for use in estimation (via minimization of (5)) and inference. An ideal estimator is one that is unbiased and has low variance, but such an estimator is difficult to obtain. For our purposes of estimation and inference, we wish to obtain a “good” estimator in the sense of having low bias and low variance. Let  $\mathbf{X} = (\mathbf{x}_1 \ \mathbf{x}_2 \ \dots \ \mathbf{x}_n)'$  be the  $n \times p$  multivariate time series data matrix, where  $\mathbf{x}_t$  is the  $p \times 1$  observation vector at time  $t$ . To apply the block bootstrap to estimate  $\mathbf{\Gamma}$  from  $\mathbf{X}$ , we first create bootstrap replicates  $\mathbf{X}_1^*, \dots, \mathbf{X}_B^*$  using either non-overlapping blocks (Carlstein, 1986), overlapping blocks (Künch, 1989), or the stationary bootstrap (Politis and Romano, 1994). For example, using Carlstein or Künch’s rules with block size  $\ell$ , one would be pasting randomly chosen  $\ell \times p$  blocks of  $\mathbf{X}$  in order to form a bootstrap replicate  $\mathbf{X}_i^*$ . For simplicity of notation, we define the function

$$S(\mathbf{y}_1, \dots, \mathbf{y}_K) = \frac{1}{K-1} \sum_{k=1}^K (\mathbf{y}_k - \bar{\mathbf{y}})(\mathbf{y}_k - \bar{\mathbf{y}})' \quad (7)$$

which calculates the standard sample covariance matrix of  $\mathbf{y}$  based on observations  $\mathbf{y}_1, \dots, \mathbf{y}_K$ , where  $\bar{\mathbf{y}} = \frac{1}{K} \sum_{k=1}^K \mathbf{y}_k$ . Now, for each  $n \times p$  bootstrap replicate  $\mathbf{X}_i^* = (\mathbf{x}_{i1}^* \ \mathbf{x}_{i2}^* \ \dots \ \mathbf{x}_{in}^*)'$ , we use the function (7) to obtain

$$\mathbf{S}_i^* = S(\mathbf{x}_{i1}^*, \dots, \mathbf{x}_{in}^*),$$

the sample covariance matrix for  $\mathbf{x}_i^*$ . Then, our initial bootstrapped estimate of  $\mathbf{\Gamma}$  is

$$\mathbf{\Gamma}^* = n S(\text{vech } \mathbf{S}_1^*, \dots, \text{vech } \mathbf{S}_B^*), \quad (8)$$

the sample covariance matrix for  $\text{vech } \mathbf{S}^*$ , where  $S(\cdot, \cdot)$  is defined in (7).

Consider the simulation described at the beginning of Section 3 in which we generated time series of length 300 according to model (3). The “true”  $\mathbf{\Gamma}$  was obtained by generating 1000 realizations of the time series and calculating  $\mathbf{\Gamma} = n \text{Var}(\mathbf{S})$ . Block bootstrap estimates  $\mathbf{\Gamma}_{(\ell)}^*$  were obtained from 400 realizations using (8) with block size  $\ell = 1, 5, 10, 15, 20$ , and 30. Because the goodness-of-fit statistic  $\text{SS}(\hat{\boldsymbol{\theta}})$  in (5) depends upon the inverse of  $\mathbf{\Gamma}$ , we not only wish that  $\mathbf{\Gamma}^*$  be a good estimator of  $\mathbf{\Gamma}$ , but also that  $(\mathbf{\Gamma}^*)^{-1}$  be a good estimator of  $\mathbf{\Gamma}^{-1}$ . These are not redundant goals since an alteration in the smallest eigenvalues of an unbiased estimator  $\hat{\mathbf{\Gamma}}$  will often yield another relatively unbiased estimator  $\tilde{\mathbf{\Gamma}}$ , but  $\hat{\mathbf{\Gamma}}^{-1}$  might be dramatically different from  $\tilde{\mathbf{\Gamma}}^{-1}$ . In order to address the issue of estimation of both  $\mathbf{\Gamma}$  and  $\mathbf{\Gamma}^{-1}$ , we consider the bias in the 36 ordered log-eigenvalues of  $\mathbf{\Gamma}_{(\ell)}^*$ ,  $\ell = 1, 5, 10, 15, 20$ , and 30. Figure 1 compares the 36 ordered log-eigenvalues of the true (simulated)  $\mathbf{\Gamma}$  with the 36 ordered log-eigenvalues of  $\mathbf{\Gamma}_{(\ell)}^*$ ,  $\ell = 1, 5, 10, 15, 20$ , and 30 (averaged over the 400 realizations of each estimator). Note that although the trace of  $\mathbf{\Gamma}_{(\ell)}^*$  essentially plateaus as  $\ell$  increases to 5~10 (see Figure 2), the underestimation of the smallest estimated eigenvalues of  $\mathbf{\Gamma}_{(\ell)}^*$  worsens as  $\ell$  increases. This increasing bias in the minimum eigenvalue estimates can be explained by the fact that more linear dependencies exist among bootstrap replicates as the block size increases. Though these linear dependencies have little effect on the estimated moments of any single component of the multivariate time series, they can distort the eigen-structure of estimated moment matrices enough to significantly alter the estimated moment matrix inverses.

To adjust for the bias in the eigenvalues of  $\mathbf{\Gamma}^*$  in estimating  $\mathbf{\Gamma}$ , we use a *nested block bootstrap* similar to the *double bootstrap* or *nested bootstrap* discussed by Chapman and Hinkley (1986) and Davison and Hinkley (1997, pp.103-113). Recall that the bootstrap replicates  $\mathbf{X}_1^*, \dots, \mathbf{X}_B^*$  used to obtain  $\mathbf{\Gamma}^*$  in (8) were created from the data matrix  $\mathbf{X}$ . We note that from the  $i$ th first-level bootstrap replicate  $\mathbf{X}_i^*$  we can obtain  $B$  second-level bootstrap replicates  $\mathbf{X}_{i1}^{**}, \dots, \mathbf{X}_{iB}^{**}$  and corresponding sample covariance matrices  $\mathbf{S}_{i1}^{**}, \dots, \mathbf{S}_{iB}^{**}$ , where  $\mathbf{X}_{ij}^{**} = (\mathbf{x}_{ij1}^{**} \ \mathbf{x}_{ij2}^{**} \ \dots \ \mathbf{x}_{ijn}^{**})'$ ,

$$\mathbf{S}_{ij}^{**} = S(\mathbf{x}_{ij1}^{**}, \dots, \mathbf{x}_{ijn}^{**}),$$

$j = 1, \dots, B$ , and  $S(\cdot, \cdot)$  is defined in (7). Then the second-level bootstrap estimate of  $\mathbf{\Gamma}$  based on  $\mathbf{X}_i^*$  can be calculated using

$$\mathbf{\Gamma}_i^{**} = n S(\text{vech } \mathbf{S}_{i1}^{**}, \dots, \text{vech } \mathbf{S}_{iB}^{**}). \quad (9)$$

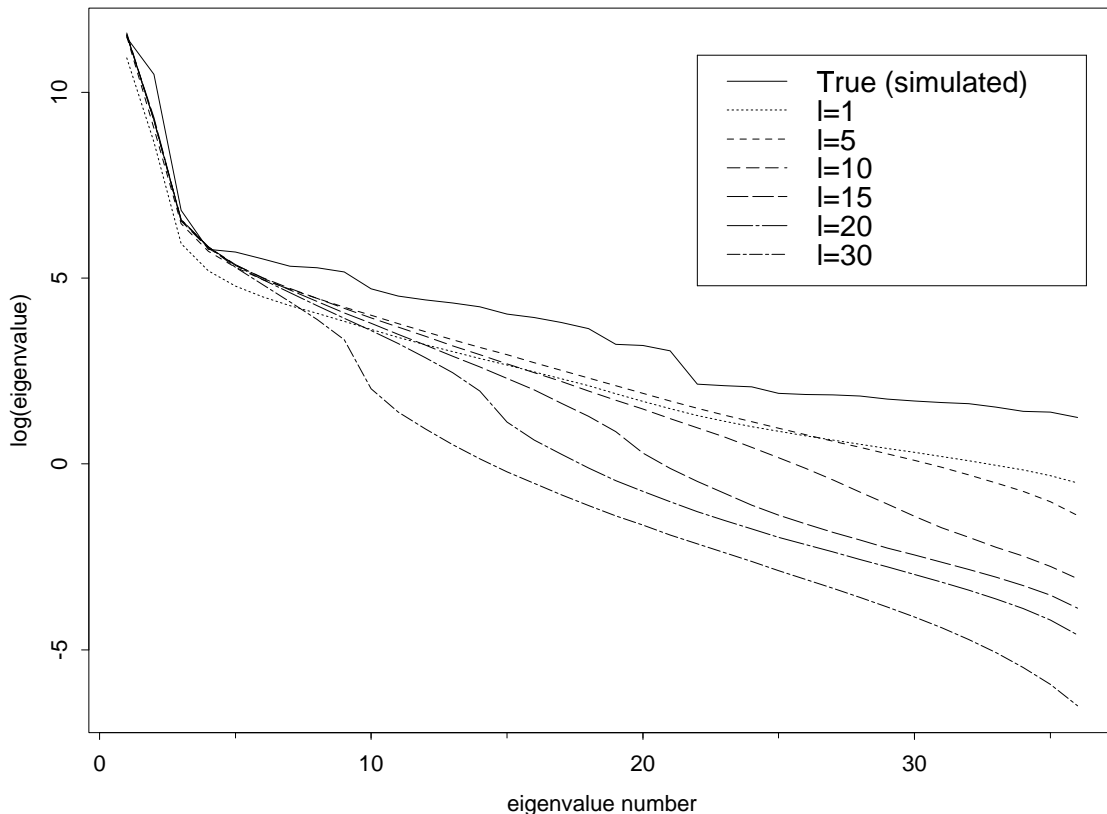


Figure 1: The 36 ordered log-eigenvalues of the true (simulated)  $\mathbf{\Gamma}$ , and the 36 ordered log-eigenvalues of  $\mathbf{\Gamma}^*(\ell)$ ,  $\ell = 1, 5, 10, 15, 20$ , and  $30$  (averaged over the 400 realizations).

The nested block bootstrap is based on the idea that the bias of  $\mathbf{\Gamma}^*$  (8) in estimating  $\mathbf{\Gamma}$  can be approximated by the bias of  $\mathbf{\Gamma}_i^{**}$  (9) in estimating  $\mathbf{\Gamma}^*$ , which can be found empirically. Beran and Srivastava (1985) recommend pivots based on the log-eigenvalues when creating bootstrap confidence regions in order to stabilize the variance of sample eigenvalues. We use similar reasoning to construct our nested block bootstrap estimate by considering the bias of the log-eigenvalues of  $\mathbf{\Gamma}^{**}$ . The bias-adjusted estimate is

$$\tilde{\mathbf{\Gamma}}^* = \mathbf{C}^* \tilde{\mathbf{D}} \mathbf{C}^{*'}, \quad (10)$$

where  $\mathbf{\Gamma}^* = \mathbf{C}^* \mathbf{D}^* \mathbf{C}^{*'}$  and  $\mathbf{\Gamma}_i^{**} = \mathbf{C}_i^{**} \mathbf{D}_i^{**} \mathbf{C}_i^{**'}$  are the spectral decompositions of  $\mathbf{\Gamma}^*$  and  $\mathbf{\Gamma}_i^{**}$ , respectively,

$$\tilde{\mathbf{D}} = \exp \left\{ 2 \log \mathbf{D}^* - \frac{1}{B_2} \sum_{i=1}^{B_2} \log \mathbf{D}_i^{**} \right\},$$

and  $B_2$  ( $\leq B$ ) is the number of first-level bootstrap repli-

cates from which  $\mathbf{\Gamma}_i^{**}$  in (9) will be calculated. In order to reduce the number of computations involved in obtaining  $\tilde{\mathbf{\Gamma}}^*$ , one may obtain second-level bootstrap replicates from each of only the first  $B_2$  first-level bootstrap replicates, where  $B_2$  might be on the order of 10 to 30. Using only  $B_2$  of the  $B$  possible matrices  $\mathbf{\Gamma}_i^{**}$  when calculating  $\tilde{\mathbf{D}}$  reduces the total number of replicates required from  $B(1+B)$  to  $B(1+B_2)$ —that is,  $B$  at the first-level of the bootstrap and  $B \times B_2$  at the second level.

We now consider the use of  $\tilde{\mathbf{\Gamma}}^*$  in (10) as the weight matrix in the GLS minimization of  $SS(\hat{\boldsymbol{\theta}})$  in (5). Specifically, we are interested in the inferential properties associated with the goodness-of-fit statistic  $SS(\hat{\boldsymbol{\theta}})$  and the confidence intervals for the elements of  $\hat{\boldsymbol{\theta}}$  when using  $\tilde{\mathbf{\Gamma}}^*$ . To evaluate these properties, we again generate time series of length 300 according to model (3). For each of the 400 generated realizations of the time series,  $\tilde{\mathbf{\Gamma}}^*$  of (10) was calculated using  $B = 100$ ,  $B_2 = 10$ , and block sizes

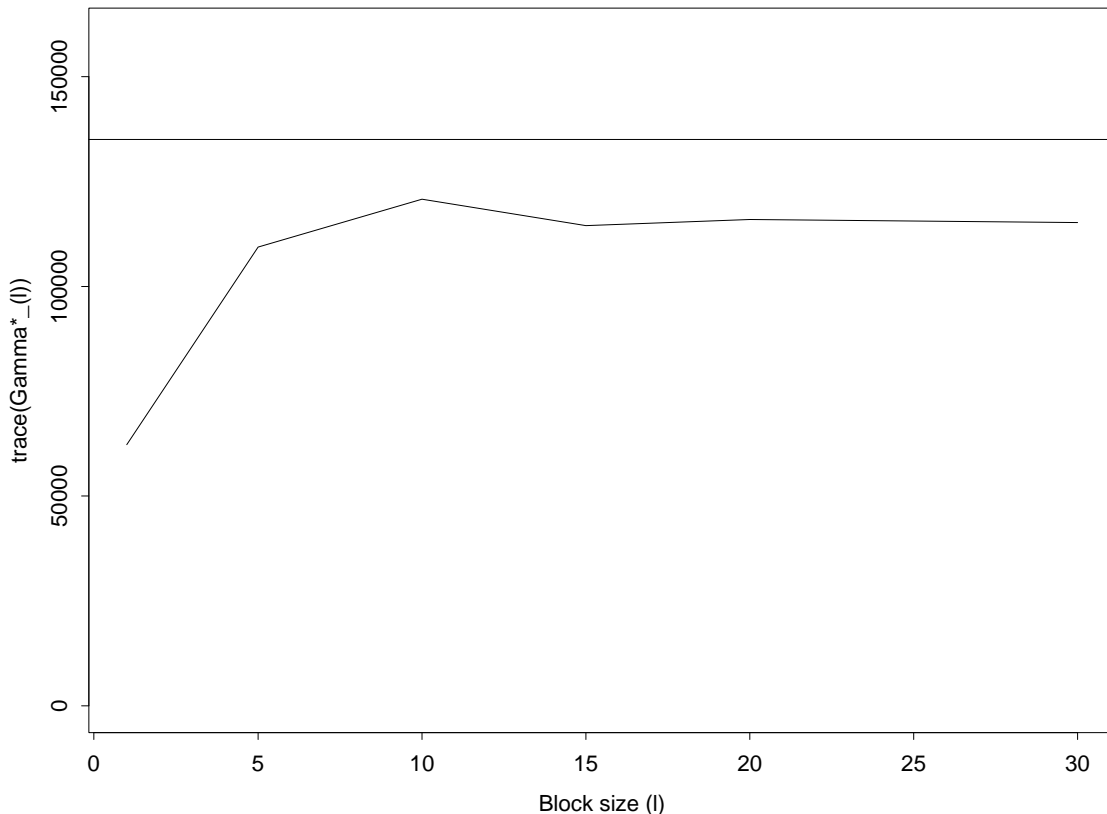


Figure 2: Trace of  $\Gamma^*(\ell)$ ,  $\ell = 1, 5, 10, 15, 20$ , and  $30$ . Trace of true (simulated)  $\Gamma$  ( $\approx 135,000$ ) given as solid line.

$\ell = 1, 5, 10$ , and  $15$ . (We denote these estimates  $\tilde{\Gamma}_{(\ell)}^*$ , where the subscript in parentheses denotes the block size.) Then, each estimate  $\tilde{\Gamma}_{(\ell)}^*$  was used to estimate  $\theta$  via minimization of (5) and to obtain  $\chi^2$  goodness-of-fit statistics and confidence intervals for the factor loadings ( $\lambda$ ). For comparison, we also use pseudo-independent pseudo-normal maximum likelihood (found in the software packages) to fit the model and carry out inference.

Figure 3 plots the average coverage probability for the 12 fitted factor loadings and the proportion of replications for which the null model was not rejected. For this study, we use  $\alpha = 0.05$  so that we wish to observe 95% coverage for the factor loading confidence intervals and a 95% “acceptance” rate for the goodness-of-fit test of the null model. The maximum likelihood approach and the approach using  $\tilde{\Gamma}_{(1)}^*$  both yield factor loading confidence intervals with low coverage, and both have goodness-of-fit test Type I error rates in excess of 60%. Apparently, the optimal value for  $\ell$  is in the neighborhood of

(5,10), since the goodness-of-fit “acceptance” rate when using  $\ell = 5$  and  $\ell = 10$  is 0.961 and 0.947, respectively. Also, the average coverage probability of confidence intervals for the factor loadings is 0.944 when  $\ell = 5$  and 0.968 when  $\ell = 10$ . Confidence intervals based on  $\tilde{\Gamma}_{(15)}^*$  are slightly conservative, with average coverage probability of 0.983. As previously noted, the larger block size leads to more substantial underestimation of the smallest eigenvalues of the fourth moment matrix. Consequently, the diagonal elements of  $(\tilde{\Gamma}_{(15)}^*)^{-1}$  are inflated and the  $\chi^2$  goodness-of-fit statistics are too large, leading to a null model acceptance rate of 0.857.

## 5 Block Size Determination

A plot such as that given in Figure 2 is helpful in determining the approximate value of the optimal block size. That is, we expect the optimal block size ( $\ell$ ) to correspond to the location on a plot of  $\text{tr}(\tilde{\Gamma}_{(\ell)}^*)$  versus  $\ell$  at

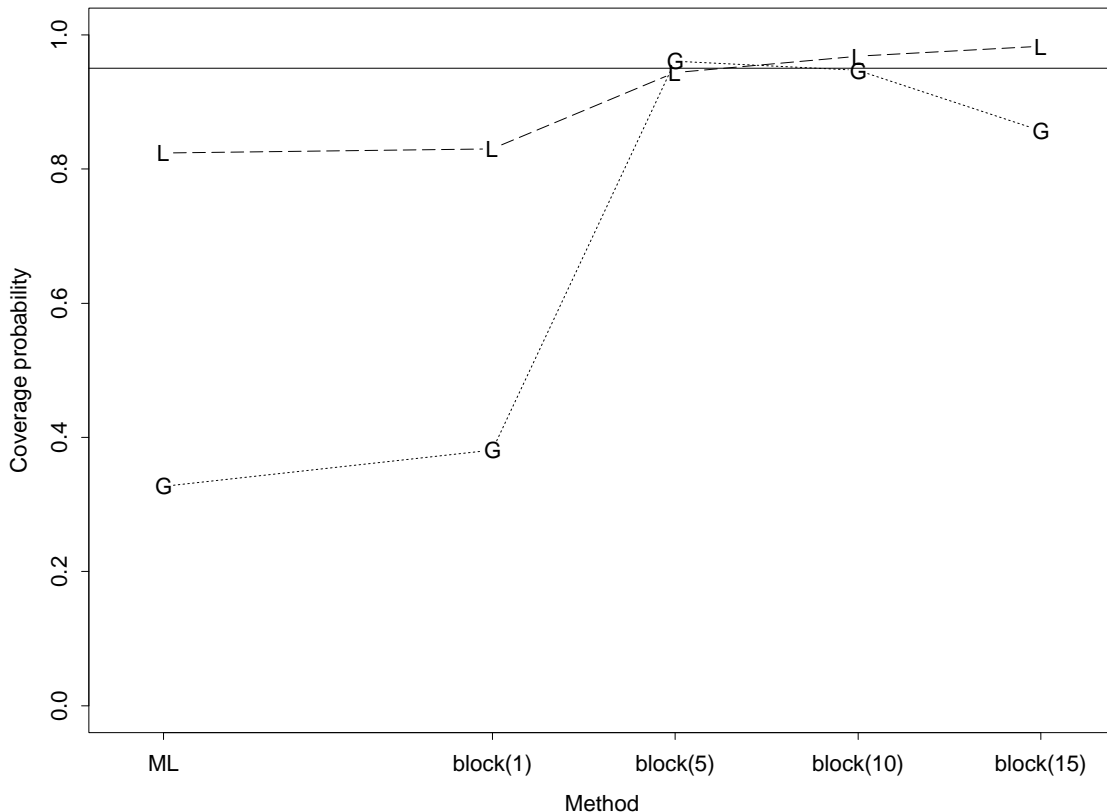


Figure 3: Average coverage probability for factor loadings (dashed/“L”) and “acceptance” rate for  $\chi^2$  goodness-of-fit test (dotted/“G”). The nominal  $1 - \alpha = 0.95$  is noted with a solid line.

which  $\text{tr}(\tilde{\Gamma}_{(\ell)}^*)$  begins to plateau. However, we wish to have an approach which explicitly determines the best choice for  $\ell$ .

The similarities between the block size selection problem and the classical bandwidth selection problem have been noted by Hall, Horowitz, and Jing (1995). When selecting a bandwidth (smoothing parameter) for estimating a density  $f(x)$  based on  $x_1, \dots, x_n$ , increasing the size of the bandwidth simultaneously increases the squared bias and decreases the variance of the estimated density. In the block size selection problem there is also a variance/bias trade-off, although the relationship between the two is more complicated, particularly in the multivariate setting where we are interested in estimating a variance-covariance matrix instead of just a single variance.

Figure 4 illustrates typical relationships between the block size ( $\ell$ ) and  $\tilde{\Gamma}_{(\ell)}^*$  when positive dependence is ex-

hibited by the multivariate time series. Specifically, we are interested in the behavior of  $f_1(\ell) =$  the squared bias of an element of  $\tilde{\Gamma}_{(\ell)}^*$ ,  $f_2(\ell) =$  the variance of an element of  $\tilde{\Gamma}_{(\ell)}^*$ , and  $f_3(\ell) =$  the squared bias of the minimum eigenvalue of  $\tilde{\Gamma}_{(\ell)}^*$ . The elements of  $\tilde{\Gamma}_{(1)}^*$  are biased, but the bias improves dramatically as  $\ell$  approaches the optimal value of  $\ell$  (where “optimal” refers to the value yielding the  $\tilde{\Gamma}_{(\ell)}^*$  with the best inferential properties). The function  $f_1(\ell)$  remains relatively constant (note the plateau in Figure 2) before increasing as  $\ell$  approaches the length of the time series ( $n$ ). The variance of an element of  $\tilde{\Gamma}_{(1)}^*$  ( $f_2(\ell)$ ) increases with  $\ell$  until the block size approaches the length of the time series, at which point the bootstrap replicates are nearly identical and the bootstrap breaks down. Recall that one of the requirements for good properties of the  $\chi^2$  test is low bias in the smallest eigenvalues of  $\tilde{\Gamma}_{(\ell)}^*$  (since the goodness-

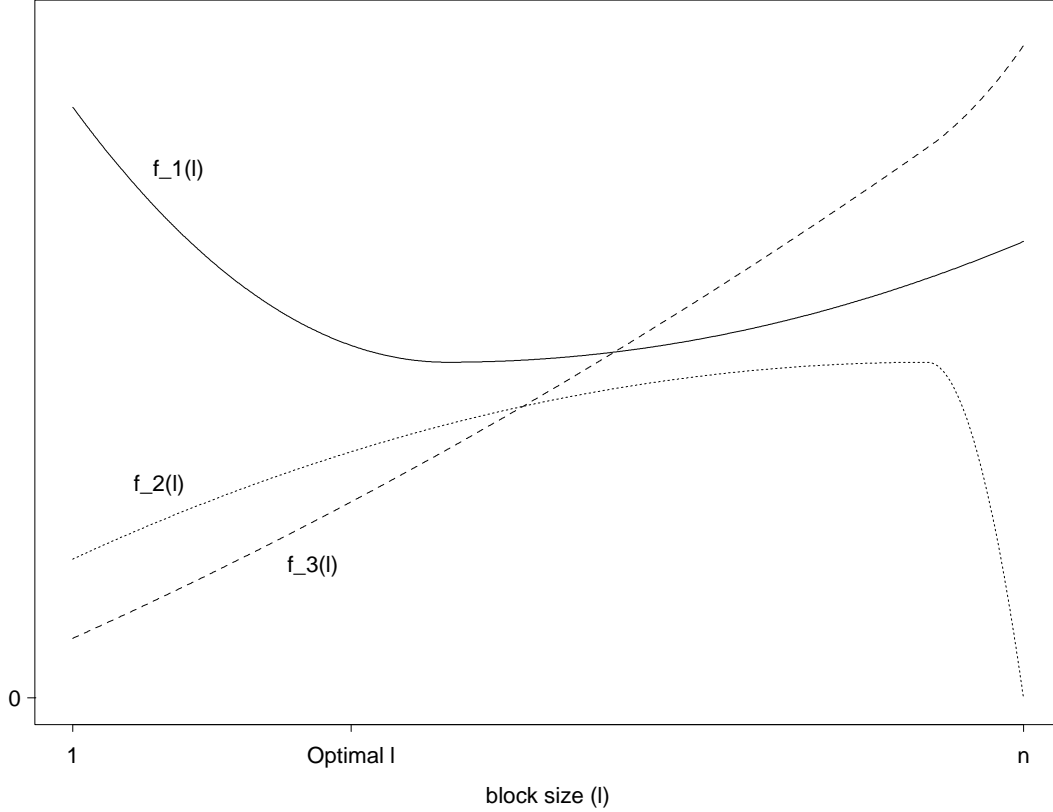


Figure 4: Behavior of  $f_1(\ell)$  = the squared bias of an element of  $\tilde{\Gamma}_{(\ell)}^*$ ,  $f_2(\ell)$  = the variance of an element of  $\tilde{\Gamma}_{(\ell)}^*$ , and  $f_3(\ell)$  = the squared bias of the minimum eigenvalue of  $\tilde{\Gamma}_{(\ell)}^*$ , as functions of the block size  $\ell \in (1, 2, \dots, n)$ .

of-fit statistic  $SS(\hat{\theta})$  of (5) depends upon  $\tilde{\Gamma}_{(\ell)}^*$  through  $(\tilde{\Gamma}_{(\ell)}^*)^{-1}$ . Consequently, our choice of  $\ell$  needs to keep  $f_3(\ell)$  at a low level. Our overall strategy is to select the block size  $\ell$  which minimizes some combination of  $f_1(\ell)$ ,  $f_2(\ell)$ , and  $f_3(\ell)$ .

Hall, Horowitz, and Jing (1995) propose an algorithm for determining the optimal block size in the univariate time series setting. Because interest is in a scalar variance, their approach can ignore  $f_3(\ell)$  and selects the block size which minimizes  $MSE(\ell) = f_1(\ell) + f_2(\ell)$ . The authors show that when estimating the variance  $\gamma$  of a univariate statistic  $g(\mathbf{x})$  based on data  $\mathbf{x} = (x_1, \dots, x_n)$  and using block size  $\ell$ , the MSE of the estimator  $\hat{\gamma}_{(\ell)}$  is minimized when using  $\ell = c \times n^{1/3}$ . The approach given by the authors is to estimate  $c$  by finding the optimal block size for a subset of  $m < n$  contiguous observations. We consider here a new multivariate extension of

the algorithm of Hall, Horowitz, and Jing (1995) which is appropriate for estimating moment matrices from an  $n \times p$  data matrix  $\mathbf{X} = (\mathbf{x}_1, \dots, \mathbf{x}_n)'$ .

1. From initial guess  $\ell_0$ , obtain  $\tilde{\Gamma}_{(\ell_0)}^*$  using the block bootstrap with block size  $\ell_0$ .
2. Randomly choose  $D$  “chunks” of size  $m \times p$  from  $\mathbf{X}$ . E.g.,  $d$ th chunk:  $\mathbf{X}^{(d)} = (\mathbf{x}_7, \dots, \mathbf{x}_{7+m-1})$ .
3. For  $d = 1, \dots, D$ , calculate  $\Gamma_{(k)}^{*(d)}$  using  $\mathbf{X}^{(d)}$  and the block bootstrap with block size  $k$ .
4. Let  $\hat{k}$  be the value of  $k$  minimizing the *weighted MSE*:

$$WMSE(\Gamma_{(k)}^*) = \frac{1}{D} \sum_{d=1}^D (\text{vech } \Gamma_{(k)}^{*(d)} - \text{vech } \tilde{\Gamma}_{(\ell_0)}^*)' \mathbf{\Omega}^{-1} \times (\text{vech } \Gamma_{(k)}^{*(d)} - \text{vech } \tilde{\Gamma}_{(\ell_0)}^*) \quad (11)$$

where  $\Omega \propto \text{Var}(\text{vech } \tilde{\Gamma}_{(\ell_0)}^*)$

$$5. \hat{\ell} = \hat{k} \times (n/m)^{1/3}$$

We now discuss several issues to be considered in implementing this algorithm. First, the algorithm can be iterated by replacing  $\ell_0$  in step 1 with  $\hat{\ell}$  in step 5. Hall, Horowitz, and Jing (1995) found that most often, the algorithm converges after the first or second iteration. Our experience verifies this finding in the multivariate setting.

In Step 3, we calculate  $\Gamma_{(k)}^{*(d)}$  in (8) instead of the less biased but more computationally intensive  $\tilde{\Gamma}_{(k)}^{*(d)}$  in (10). But in the calculation of  $\text{WMSE}(\Gamma_{(k)}^*)$  in Step 4, instead of subtracting the biased first-level bootstrap estimate  $\Gamma_{(\ell_0)}^*$  from  $\text{vech } \Gamma_{(k)}^{*(d)}$ , the bias-adjusted nested bootstrap estimate  $\text{vech } \tilde{\Gamma}_{(\ell_0)}^*$  is subtracted to ensure that  $\text{WMSE}$  is not dominated by the variance. This has the affect of selecting a block size  $\ell$  that is slightly larger because  $\text{WMSE}$  has the form  $f_1(\ell) + f_2(\ell)$  rather than the form  $a f_1(\ell) + f_2(\ell)$ , for  $a < 1$  (see Figure 4).

As noted above, the weight matrix  $\Omega$  should be proportional to  $\text{Var}(\text{vech } \tilde{\Gamma}_{(\ell_0)}^*)$ . When  $\Omega$  is too large or too difficult to estimate, one can simplify Step 4 minimizing the *diagonal weighted MSE*

$$\begin{aligned} \text{DWMSE}(\Gamma_{(k)}^*) &= \frac{1}{D} \sum_{d=1}^D (\text{diag } \Gamma_{(k)}^{*(d)} - \text{diag } \tilde{\Gamma}_{(\ell_0)}^*)' \Omega^{-1} \\ &\quad \times (\text{diag } \Gamma_{(k)}^{*(d)} - \text{diag } \tilde{\Gamma}_{(\ell_0)}^*) \end{aligned} \quad (12)$$

where  $\Omega \propto \text{Var}(\text{diag } \tilde{\Gamma}_{(\ell_0)}^*)$ . In the simulations and data analysis to follow, we use (12) with  $\Omega$  being a diagonal matrix with each diagonal element equal to the square of the corresponding element of  $\text{diag } \tilde{\Gamma}_{(\ell_0)}^*$ . Both (11) and (12) are functions of  $f_1(\ell)$  and  $f_2(\ell)$  only. In Step 4, one might also consider minimizing functions such as

$$a_1 \text{DWMSE}(\Gamma_{(k)}^*) + a_2 f_3(\ell)$$

or

$$a_1 \text{DWMSE}(\Gamma_{(k)}^*) + a_2 \text{DWMSE}((\Gamma_{(k)}^*)^{-1}),$$

where

$$\begin{aligned} \text{DWMSE}((\Gamma_{(k)}^*)^{-1}) &= \\ \frac{1}{D} \sum_{d=1}^D &(\text{diag } (\Gamma_{(k)}^{*(d)})^{-1} - \text{diag } (\tilde{\Gamma}_{(\ell_0)}^*)^{-1})' \Omega^{-1} \times \\ &(\text{diag } (\Gamma_{(k)}^{*(d)})^{-1} - \text{diag } (\tilde{\Gamma}_{(\ell_0)}^*)^{-1}), \end{aligned}$$

and  $\Omega$  is a diagonal matrix with each diagonal element equal to the square of the corresponding element of

Table 2: Distribution of empirically chosen optimal block length (lognormal factor and error processes).

$\hat{\ell}$	2	3	6	10	13	19	29	39
Freq.	.15	.21	.15	.30	.09	.06	.03	.03

Table 3: Distribution of empirically chosen optimal block length (normal factor and error processes).

$\hat{\ell}$	2	3	6	10	13	19	29	39
Freq.	.06	.31	.51	.09	.03	.00	.003	.00

$\text{diag } (\tilde{\Gamma}_{(\ell_0)}^*)^{-1}$ . The performance of this algorithm when minimizing functions other than (12) is currently being investigated.

Lastly, we note that the algorithm generalizes in a straightforward manner for the estimation of any variance-covariance matrix such as the  $p \times p$  matrix  $\text{Var}(\bar{\mathbf{X}})$ . In fact,  $\Gamma$  can be considered to be of the form  $\Gamma \cong n \text{Var}(\bar{\mathbf{A}})$ , where  $\mathbf{A}_i = \text{vech } (\mathbf{X}_i - \bar{\mathbf{X}})(\mathbf{X}_i - \bar{\mathbf{X}})'$ ,  $i = 1, \dots, n$ .

We consider the use of the algorithm for estimating the optimal block size by generating time series of length 300 according to model (3). We consider data generated from (3) where the AR(1) factor and error processes are lognormal and normal. The autoregressive coefficient for each of the two factor and eight error processes is  $\phi_1 = 0.6$ , and when calculating  $\tilde{\Gamma}^*$  in (10), we use  $B = 50$  and  $B_2 = 20$ . As described in Section 4 and Figure 3, the block size yielding optimal inferential properties when fitting model (4) to these data is in the range (5,10). For each of 400 generated realizations of the time series, the algorithm given above was used to estimate the optimal block size, except that we minimized (12) instead of (11) in Step 4. In Step 2, we use  $D = 20$  ‘‘chunks’’ of size  $m \times p$  with  $m = 72$ . For  $k$  in Steps 3-4, the values  $k = 1, 2, 4, 6, 7, 12, 18$ , and 24 were used. The value of  $k$  minimizing (12) was then used in Step 5 to obtain  $\hat{\ell} \approx 2, 3, 6, 10, 13, 19, 29$ , or 39, respectively. Tables 2 and 3 give the distribution of the empirically chosen optimal block length when factor and error processes are lognormal and normal, respectively. We note that the mode of each distribution is the range which is optimal for inference. When the data are normally distributed, the algorithm selects a block length in (3,10) over 90% of the time. When the data are lognormally distributed, only 65% of chosen block lengths are between 3 and 10.

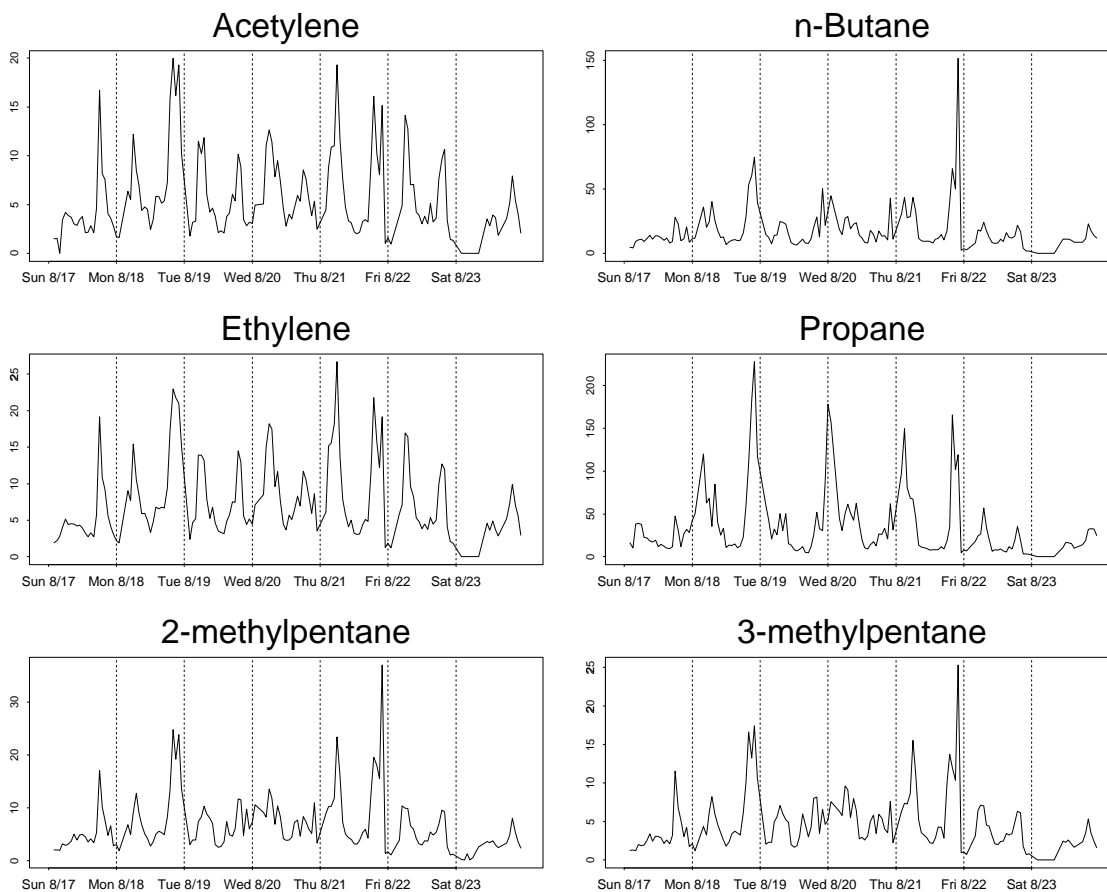


Figure 5: Hourly measurements of six VOCs over a 7-day period.

## 6 Modeling of the El Paso/Juarez Pollution Data

We turn now to a collection of hourly measurements of volatile organic compounds (VOCs) obtained during a 55-day period of the summer of 1997 in the El Paso, Texas/Ciudad Juarez, Mexico area (see Fujita, 1998). We consider 10 of these VOCs including: 2-methylpentane (2Mp), 3-methylpentane (3Mp), Benzene, Cyclohexane, 2-methylhexane (2Mh), 2,2,4-trimethylpentane (224Tmp), Propane, Ethylene, Acetylene, and n-Butane. Our purpose is to determine the number of pollution sources from which these VOCs were emitted. Fujita (1998) reported that between 50 and 67 percent of nonmethane hydrocarbon in this region is due to vehicle exhaust, and that the second largest contributor of nonmethane hydrocarbon is industrial emissions, accounting for 10 to 30 percent. We wish to evaluate the hypothesis that the 10 VOC concentrations are the

result of two latent sources, namely auto exhaust and industrial emissions.

The dependence in the data is illustrated in Figure 5, which gives plots for six of these VOCs over a 7-day segment of the 55-day observation period. Some of these VOCs are known to be closely associated with auto exhaust, including acetylene and ethylene, and to a lesser degree, 2-methylpentane and 3-methylpentane. These pollutants have daily peaks which roughly correspond to the morning and evening rush hours and are less pronounced on the Sunday and Saturday shown. Other VOCs plotted in Figure 5 are known to be components of emissions from local industry, including n-Butane, Propane, and to a lesser degree, 2-methylpentane. Comparatively small components of auto exhaust, n-Butane and Propane exhibit dependence structures which are similar to each other, but decidedly different from acetylene and ethylene. Though we cannot observe the time series of emission volumes from auto exhaust or industrial emissions, we wish to evaluate the hypothesis that

the observed VOCs are linear combinations of these two unobserved time series (factors), along with 10 error processes which are independent of one another. It is clear that statistical methods employed to analyze these data should be robust to nonnormality, as these and other environmental data often exhibit skewness.

We employ the errors-in-variables parameterization of the linear factor analysis model. Using our knowledge about the nature of these pollution sources, we select a “tracer” pollutant for each of auto exhaust and industrial emissions (Acetylene and n-Butane, respectively). Though the choice of which observed variable is defined to be equal to a factor plus error does not affect the goodness-of-fit statistic, using such a parameterization identifies the model and makes possible a physical interpretation of the parameter estimates. Our hypothesized model is

$$\begin{aligned}
 2\text{Mp} &= \lambda_{11}f_{1t} + \lambda_{12}f_{2t} + e_{1t} \\
 3\text{Mp} &= \lambda_{21}f_{1t} + \lambda_{22}f_{2t} + e_{2t} \\
 \text{Benzene} &= \lambda_{31}f_{1t} + \lambda_{32}f_{2t} + e_{3t} \\
 \text{Cyclohex} &= \lambda_{41}f_{1t} + \lambda_{42}f_{2t} + e_{4t} \\
 2\text{Mh} &= \lambda_{51}f_{1t} + \lambda_{52}f_{2t} + e_{5t} \\
 224\text{Tmp} &= \lambda_{61}f_{1t} + \lambda_{62}f_{2t} + e_{6t} \\
 \text{Propane} &= \lambda_{71}f_{1t} + \lambda_{72}f_{2t} + e_{7t} \\
 \text{Ethylene} &= \lambda_{81}f_{1t} + \lambda_{82}f_{2t} + e_{8t} \\
 \text{Acetylene} &= f_{1t} + e_{9t} \\
 \text{n-Butane} &= f_{2t} + e_{10t}.
 \end{aligned} \tag{13}$$

In order to fit the model, we need an estimate of  $\mathbf{\Gamma}$  in (6). Since measurements were taken hourly for 55 days, our data matrix  $\mathbf{X}$  has 1320 rows, but 280 of these measurements are missing. Notwithstanding, one can still obtain the block bootstrap estimate by resampling  $\ell \times p$  blocks of  $\mathbf{X}$ , and performing the necessary calculations using the non-missing observations in each bootstrap replicate. First, the block size was selected by using the algorithm given in Section 5. When implementing the algorithm, we used:  $B = 200$  and  $B_2 = 10$  when calculating  $\tilde{\mathbf{\Gamma}}_{(\ell_0)}^*$  in Step 1,  $D = 20$  chunks of size  $m \times p$  with  $m = 400$  in Step 2, and  $B = 200$  when calculating  $\mathbf{\Gamma}_{(k)}^{*(d)}$  in Step 3. In Step 4, we used (12) instead of (11), and found this function to be minimized when  $k$  was in the range (6,20), yielding  $\hat{\ell}$  in the range (9,30) in Step 5. In order to lessen the bias in the eigenvalues of  $\tilde{\mathbf{\Gamma}}^*$  (which worsens as  $\ell$  increases), we choose  $\hat{\ell} = 10$ , a value at the low end of the estimated optimal range.

For comparison, we first fit model (13) using pseudo-independent pseudo-normal maximum likelihood, the default estimation technique in several of the software packages. Using an  $\alpha = 0.05$  significance level, the 2-factor model was rejected with a  $\chi^2 = 763.27$  and  $p$ -value  $< 10^{-10}$ , indicating the need for more factors in

the model. The 1-factor, 3-factor, and 4-factor models were also fit, but these had similar values for the  $\chi^2$  goodness-of-fit test associated with the maximum likelihood fit. In contrast, when we obtain  $\hat{\theta}$  by minimizing (5), replacing  $\mathbf{\Gamma}$  with  $\tilde{\mathbf{\Gamma}}_{(10)}^*$ , we obtain a goodness-of-fit statistic  $\chi^2 = \text{SS}(\hat{\theta})$  which accounts for the dependence in the data. Using this approach, the 1-factor model is rejected with  $\chi^2 = 86.58$  and  $p$ -value = 0.0000075, but the hypothesized 2-factor model given in (13) does not show lack of fit, with  $\chi^2 = 40.81$  and  $p$ -value = 0.0715. Thus, we conclude that the 10 VOCs can be considered to be products of 2 underlying pollution sources, one closely associated with vehicle exhaust and the other closely associated with industrial emissions.

## 7 Conclusion

Hypotheses about pollution source apportionment can be assessed by a class of latent variable models known as multivariate receptor models. When using subject-matter knowledge, one can specify multivariate receptor models that are identified and the estimates of the model parameters can have a physical interpretation with practical value. Though dependence structure in air quality monitoring data is often ignored at the expense of valid inference, one can incorporate dependence structure directly into estimation and inference via the block bootstrap. This approach is particularly attractive when observations are high-dimensional, since the modeling of multivariate temporal and/or spatial covariance functions is in general very complicated. The bias-adjusted, nested block bootstrap introduced herein provides appropriate parameter estimation and inference when using latent variable models for dependent data. The application of the methodology is facilitated by a new multivariate extension of the block size determination algorithm of Hall, Horowitz, and Jing (1995). The methodology is used to determine that 10 volatile organic compounds measured hourly during the summer of 1997 in the El Paso/Ciudad Juarez can be considered to be products of 2 underlying pollution sources, one closely associated with vehicle exhaust and the other closely associated with industrial emissions.

## References

Alpert, D. J., and Hopke, P. K. (1980), “A Quantitative Determination of Sources in the Boston Urban Aerosol,” *Atmospheric Environment*, Vol. 14, pp. 1137-1146.

- Amemiya, Y., and Anderson, T. W. (1990), "Asymptotic Chi-square Tests for a Large Class of Factor Analysis Models," *The Annals of Statistics*, Vol. 18, pp. 1453-1463.
- Anderson, T. W., and Amemiya, Y. (1988), "The Asymptotic Normal Distribution of Estimators in Factor Analysis Under General Conditions," *The Annals of Statistics*, Vol. 16, pp. 759-771.
- Beran, R., and Srivastava, M. S. (1985), "Bootstrap Tests and Confidence Regions for Functions of a Covariance Matrix," *The Annals of Statistics*, Vol. 13, pp. 95-115.
- Carlstein, E. (1986), "The Use of Subseries Values for Estimating the Variance of a General Statistic from a Stationary Sequence," *The Annals of Statistics*, Vol. 14, pp. 1171-1179.
- Chapman, P., and Hinkley, D. V. (1986), "The Double Bootstrap, Pivots and Confidence Limits," Center for Statistical Sciences, University of Texas at Austin, Technical Report 34.
- Davison, A. C., and Hinkley, D. V. (1997), *Bootstrap Methods and their Application*, Cambridge University Press, Cambridge, UK.
- Fujita, E. M. (1998), "Hydrocarbon Source Apportionment for the 1996 Paso Del Norte Ozone Study," Final Report, EPA Contract 68-D3-0030, Work Assignment III-130.
- Fuller, W. A. (1987), *Measurement Error Models*, John Wiley & Sons, Inc., New York.
- Hall, P. (1985), "Resampling a Coverage Process," *Stochastic Processes Applications*, Vol. 20, pp. 231-246.
- Hall, Horowitz, and Jing (1995), "On Blocking Rules for the Bootstrap with Dependent Data," *Biometrika*, Vol. 82, pp. 561-574.
- Henry, R. C., Lewis, C. W., and Collins, J. F. (1994), "Vehicle-Related Hydrocarbon Source Compositions from Ambient Data: The GRACE/SAFER Method," *Environmental Science Technology*, Vol. 28, pp. 823-832.
- Hershberger, S. L., Corneal, S. E., and Molenaar, P. C. M. (1994), "Dynamic Factor Analysis: An Application to Emotional Response Patterns Underlying Daughter/Father and Stepdaughter/Stepfather Relationships," *Structural Equation Modeling*, Vol. 2, pp. 31-52.
- Koutrakis, P. and Spengler, J. D. (1987), "Source Apportionment of Ambient Particles in Steubenville, OH Using Specific Rotation Factor Analysis," *Atmospheric Environment*, Vol. 21, pp. 1511-1519.
- Künsch, H. (1989), "The Jackknife and the Bootstrap for General Stationary Observations," *The Annals of Statistics*, Vol. 17, pp. 1217-1241.
- Park, E. S., Guttorp, P., and Henry, R. C. (2000), "Multivariate Receptor Modeling for Temporally Correlated Data by Using MCMC," National Research Center for Statistics and the Environment, TRS #034, [www.nrcse.washington.edu/research/reports/reports.asp](http://www.nrcse.washington.edu/research/reports/reports.asp).
- Park, E. S., Henry, R. C., and Spiegelman, C. H. (1999), "Determining the Number of Major Pollution Sources in Multivariate Air Quality Receptor Models," National Research Center for Statistics and the Environment, TRS #034, [www.nrcse.washington.edu/research/reports/reports.asp](http://www.nrcse.washington.edu/research/reports/reports.asp).
- Park, E. S., Spiegelman, C. H., and Henry, R. C. (1999), "Bilinear Estimation of Pollution Source Profiles in Receptor Models," National Research Center for Statistics and the Environment, TRS #019, [www.nrcse.washington.edu/research/reports/reports.asp](http://www.nrcse.washington.edu/research/reports/reports.asp).
- Politis, D. N. and Romano, J. P. (1994), "The Stationary Bootstrap," *Journal of the American Statistical Association*, Vol. 89, pp. 1303-1313.
- Spiegelman, C. H. and Dattner, S. (1993), "Multivariate Chemometrics, a Case Study: Applying and Developing Receptor Models for the 1990 El Paso Winter PM<sub>10</sub> Receptor Modeling Scoping Study," in *Multivariate Environmental Statistics*, Patil, G. P., and Rao, C. R., eds., Elsevier Science Publishers, pp. 509-524.
- Thurston, G. D., and Spengler, J. D. (1985), "A Quantitative Assessment of Source Contributions to Inhalable Particulate Matter Pollution in Metropolitan Boston," *Atmospheric Environment*, Vol. 19, pp. 9-25.
- Ver Hoef, J. M. and Barry, R. D. (1998), "Modeling Crossvariograms for Cokriging and Multivariable Spatial Prediction," *Journal of Statistical Planning and Inference*, Vol. 69, pp. 275-294.