The Annals of Applied Statistics
2012, Vol. 6, No. 4, 1552–1587
DOI: 10.1214/12-AOAS587

GAP BOOTSTRAP METHODS FOR MASSIVE DATA SETS WITH AN APPLICATION TO TRANSPORTATION ENGINEERING[1]

By S. N. Lahiri, C. Spiegelman, J. Appiah and L. Rilett

Texas A&M University, Texas A&M University, University of Nebraska-Lincoln and University of Nebraska-Lincoln

In this paper we describe two bootstrap methods for massive data sets. Naive applications of common resampling methodology are often impractical for massive data sets due to computational burden and due to complex patterns of inhomogeneity. In contrast, the proposed methods exploit certain structural properties of a large class of massive data sets to break up the original problem into a set of simpler subproblems, solve each subproblem separately where the data exhibit approximate uniformity and where computational complexity can be reduced to a manageable level, and then combine the results through certain analytical considerations. The validity of the proposed methods is proved and their finite sample properties are studied through a moderately large simulation study. The methodology is illustrated with a real data example from Transportation Engineering, which motivated the development of the proposed methods.

1. Introduction. Statistical analysis and inference for massive data sets present unique challenges. Naive applications of standard statistical methodology often become impractical, especially due to increase in computational complexity. While large data size is desirable from a statistical inference perspective, suitable modification of existing statistical methodology is needed to handle such challenges associated with massive data sets. In this paper, we propose a novel resampling methodology, called the *Gap Bootstrap*, for a large class of massive data sets that possess certain structural properties. The proposed methodology cleverly exploits the data structure to break up the original inference problem into smaller parts, use standard resampling methodology to each part to reduce the computational complexity,

Received April 2012.

[1]Supported in part by grant from the U.S. Department of Transportation, University Transportation Centers Program to the University Transportation Center for Mobility (DTRT06-G-0044) and by NSF Grants no. DMS-07-07139 and DMS-10-07703.

Key words and phrases. Exchangeability, multivariate time series, nonstationarity, OD matrix estimation, OD split proportion, resampling methods.

and then use some analytical considerations to put the individual pieces together, thereby alleviating the computational issues associated with large data sets to a great extent.

The class of problems we consider here is the estimation of standard errors of estimators of population parameters based on massive multivariate data sets that may have heterogeneous distributions. A primary example is the origin-destination (OD) model in transportation engineering. In an OD model, which motivates this work and which is described in detail in Section 2 below, the data represent traffic volumes at a number of origins and destinations collected over short intervals of time (e.g., 5 minute intervals) daily, over a long period (several months), thereby leading to a massive data set. Here, the main goals of statistical analysis are (i) uncertainty quantification associated with the estimation of the parameters in the OD model and (ii) to improve prediction of traffic volumes at the origins and the destinations over a given stretch of the highway. Other examples of massive data sets having the required structural property include (i) receptor modeling in environmental monitoring, where spatio-temporal data are collected for many pollution receptors over a long time, and (ii) toxicological models for dietary intakes and drugs, where blood levels of a large number of toxins and organic compounds are monitored in repeated samples for a large number of patients. The key feature of these data sets is the presence of "gaps" which allow one to partition the original data set into smaller subsets with nice properties.

The "largeness" and potential inhomogeneity of such data sets present challenges for estimated model uncertainty evaluation. The standard propagation of error formula or the delta method relies on assumptions of independence and identical distributions, stationarity (for space–time data) or other kinds of uniformity which, in most instances, are not appropriate for such data sets. Alternatively, one may try to apply the bootstrap and other resampling methods to assess the uncertainty. It is known that the ordinary bootstrap method typically underestimates the standard error for parameters when the data are dependent (positively correlated). The block bootstrap has become a popular tool for dealing with dependent data. By using blocks, the local dependence structure in the data is maintained and, hence, the resulting estimates from the block bootstrap tend to be less biased than those from the traditional (i.i.d.) bootstrap. For more details, see Lahiri (1999, 2003). However, computational complexity of naive block bootstrap methods increases significantly with the size of the data sets, as the given estimator has to be computed *repeatedly* based on resamples that have the same size as the original data set. In this paper, we propose two resampling methods, generally both referred to as *Gap Bootstraps*, that exploit the "gap" in the dependence structure of such large-scale data sets to reduce the computational burden. Specifically, the gap bootstrap estimator of the standard error is appropriate for data that can be partitioned into approximately exchangeable or homogeneous subsets. While the distribution of the entire data set is not exchangeable or homogeneous, it is entirely reasonable that many multivariate subsets will be exchangeable or homogeneous. If the estimation method that is being used is accurate, then we show that the gap bootstrap gives a consistent and asymptotically unbiased estimate of standard errors. The *key idea is to employ the bootstrap method to each of the homogeneous subsets of the data separately and then combine the estimators from different subsets in a suitable way to produce a valid estimator of the standard error of a given estimator based on the entire data set*. The proposed method is computationally much simpler than the existing resampling methods that require repeated computation of the original estimator, which may not be feasible simply due to computational complexity of the original estimator, at the scale of the whole data set.

The rest of the paper is organized as follows. In Section 2 we describe the OD model and the data structure that motivate the proposed methodology. In Section 3 we give the descriptions of two variants of the Gap Bootstrap. Section 4 asserts consistency of the proposed Gap Bootstrap variance estimators. In Section 5 we report results from a moderately large simulation study, which shows that the proposed methods attain high levels of accuracy for moderately large data sets under various types of gap-dependence structures. In Section 6 we revisit the OD models and apply the methodology to a real data set from a study of traffic patterns, conducted by an intelligent traffic management system on a test bed in San Antonio, TX. Some concluding remarks are made in Section 7. Conditions for the validity of the theoretical results and outlines of the proofs are given in the Appendix.

2. The OD models and the estimation problem.

2.1. *Background.* The key component of an origin-destination (OD) model is an OD trip matrix that reflects the volume of traffic (number of trips, amount of freight, etc.) between all possible origins and destinations in a transportation network over a given time interval. The OD matrix can be measured directly, albeit with much effort and at great costs, by conducting individual interviews, license plate surveys, or by taking aerial photographs [cf. Cramer and Keller (1987)]. Because of the cost involved in collecting direct measurements to populate a traffic matrix, there has been considerable effort in recent years to develop synthetic techniques which provide "reasonable" values for the unknown OD matrix entries in a more indirect way, such as using observed data from link volume counts from inductive loop detectors. Over the past two decades, numerous approaches to synthetic OD matrix estimation have been proposed [Cascetta (1984), Bell (1991),

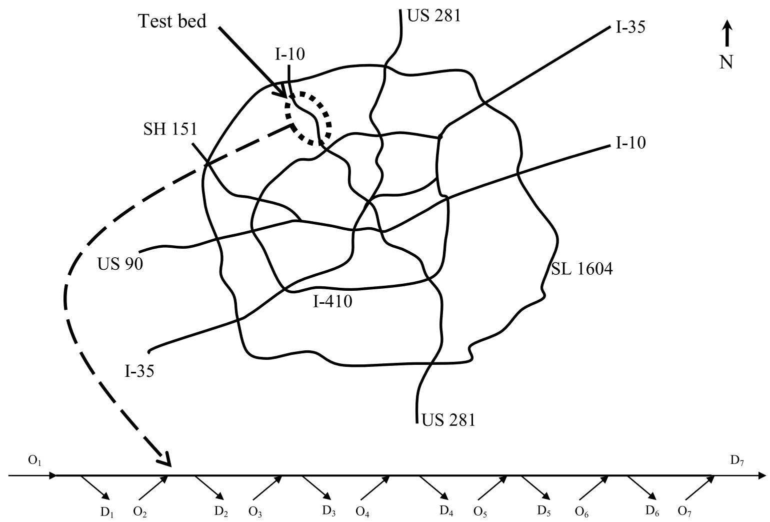

FIG. 1. *The transportation network in San Antonio, TX under study.*

Okutani (1987), Dixon and Rilett (2000)]. One common approach for estimating the OD matrix from link volume counts is based on the least squares regression where the unknown OD matrix is estimated by minimizing the squared Euclidean distance between the observed link and the estimated link volumes.

2.2. *Data structure.* The data are in the form of a time series of link volume counts measured at several on/off ramp locations on a freeway using an inductive loop detector, such as in Figure 1.

Here O_k and D_k, respectively, represent the traffic volumes at the kth origin and the kth destination over a given stretch of a highway. The analysis period is divided into T time periods of equal duration Δt. The time series of link volume counts is generally *periodic* and *weakly dependent*, that is, the dependence dies off as the separation of the time intervals becomes large. For example, daily data over each given time slot of duration Δt are similar, but data over well separated time slots (e.g., time slots in Monday morning and Monday afternoon) can be different. This implies that the traffic data have a periodic structure. Further, Monday at 8:00–8:05 am data have nontrivial correlation with Monday at 8:05–8:10 am data, but neither data set says anything about Tuesday data at 8:00–8:05 am (showing *approximate independence*). Accordingly, let $\mathbf{Y}_t, t = 1, 2 \ldots,$ be a d-dimensional time series, representing the link volume counts at a given set of on/off ramp locations over the tth time interval. Suppose that we are interested in reconstructing the OD matrix for p-many short intervals during the morning rush hours, such as 36 link volume counts over $\Delta t = 5$-minute intervals, extending from 8:00 am through 11:00 am, at several on/off ramp locations. Thus, the observed data for the OD modeling is a part of the $\mathbf{Y}_t$ series,

$$\{\mathbf{X}_1, \ldots, \mathbf{X}_p; \ldots; \mathbf{X}_{(m-1)p+1}, \ldots, \mathbf{X}_{mp}\},$$

where the link volume counts are observed over the p-intervals on each day, for m days, giving a d-dimensional multivariate sample of size $n = mp$. There are $q = T - p$ time slots between the last observation on any given day and the first observation on the next day, which introduces the "gap" structure in the $\mathbf{X}_t$-series. Specifically, in terms of the $\mathbf{Y}_t$-series, the $\mathbf{X}_t$-variables are given by

$$\mathbf{X}_{ip+j} = \mathbf{Y}_{i(p+q)+j}, \qquad j = 1, \ldots, p, i = 0, \ldots, m-1.$$

For data collected over a large transportation network and over a long period of time, d and m are large, leading to a massive data set. Observe that the $\mathbf{X}_t$-variables can be arranged in a $p \times m$ matrix, where each element of the matrix-array gives a d-dimensional data value:

$$\mathbb{X} = \begin{pmatrix} \mathbf{X}_1 & \mathbf{X}_{p+1} & \ldots & \mathbf{X}_{(m-1)p+1} \\ \mathbf{X}_2 & \mathbf{X}_{p+2} & \ldots & \mathbf{X}_{(m-1)p+2} \\ \cdot & \cdot & \ldots & \cdot \\ \cdot & \cdot & \ldots & \cdot \\ \mathbf{X}_p & \mathbf{X}_{2p} & \ldots & \mathbf{X}_{mp} \end{pmatrix}. \tag{2.1}$$

Due to the arrangement of the p time slots in the jth day along the jth column in (2.1), the rows in the array (2.1) correspond to a fixed time slot over days and are expected to exhibit a similar distribution of the link volume counts; although a day-of-week variation might be present, the standard practice in the Transportation engineering is to treat the weekdays as similar [cf. Roess, Prassas and McShane (2004), Mannering, Washburn and Kilareski (2009)]. On the other hand, due to the "gap" between the last time slot on the jth day and the first time slot of the $(j+1)$st day, the variables in the jth and the $(j+1)$st columns are essentially independent. Hence, this yields a data structure where

$$\left.\begin{array}{l} \text{(a) the variables within each column have serial correlations} \\ \quad\ \text{and possibly nonstationary distributions,} \\ \text{(b) the variables in each row are identically distributed, and} \\ \text{(c) the columns are approximately independent arrays} \\ \quad\ \text{of random vectors.} \end{array}\right\} \tag{2.2}$$

In the transportation engineering application, each random vector $\mathbf{X}_t$ represents the link volume counts in a transportation network corresponding to r origin (entrance) ramps and s destination (exit) ramps as shown in Figure 1. Let $o_{\ell t}$ and d_{kt}, respectively, denote the link volumes at origin ℓ and at destination k at time t. Then the components of $\mathbf{X}_t$ for each t are given by the $d \equiv (r+s)$-variables $\{o_{\ell t} : \ell = 1, \ldots, r\} \cup \{d_{kt} : k = 1, \ldots, s\}$. Given the link volume counts on all origin and destination ramps, the fraction $p_{k\ell}$ (known as the *OD split proportion*) of vehicles that exit the system at destination ramp k given that they entered at origin ramp ℓ can be calculated. This is because the link volume at destination k at time t, d_{kt}, is a linear combination of the OD split proportions and the origin volumes at time t, $o_{\ell t}$'s. In the synthetic OD model, $p_{k\ell}$'s are the *unknown* system parameters and have to be estimated. Once the split proportions are available, the OD matrix for each time period can be identified as a linear combination of the split proportion matrix and the vector of origin volumes. The key statistical inference issue here is to quantify the size of the standard errors of the estimated split proportions in the synthetic OD model.

3. Resampling methodology.

3.1. *Basic framework.* To describe the resampling methodology, we adopt a framework that mimics the "gap structure" of the OD model in Section 2. Let $\{\mathbf{X}_1,\dots,\mathbf{X}_p;\dots;\mathbf{X}_{(m-1)p+1},\dots,\mathbf{X}_{mp}\}$ be a d-dimensional time series with stationary components $\{\mathbf{X}_{ip+j}:i=0,\dots,m-1\}$ for $j=1,\dots,p$ such that the corresponding array (2.1) satisfies (2.2). For example, such a time series results from a periodic, multivariate parent time series $\mathbf{Y}_t$ that is m_0-dependent for some $m_0 \geq 0$ and that is observed with "gaps" of length $q > m_0$. In general, the dependence structure of the original time series $\mathbf{Y}_t$ is retained within each complete period $\{\mathbf{X}_{ip+j}:j=1,\dots,p\}$, $i=0,\dots,m$, but the random variables belonging to two different periods are essentially independent. Let θ be a vector-valued parameter of interest and let $\hat{\theta}_n$ be an estimator of θ based on $\mathbf{X}_1,\dots\mathbf{X}_n$, where $n=mp$ denotes the sample size. We now formulate two resampling methods for estimating the standard error of $\hat{\theta}_n$ that are suitable for massive data sets with such "gap" structures. The first method is applicable when the p rows of the array (2.1) are *exchangeable* and the second one is applicable where the rows are possibly nonidentically distributed and where the variables within each column have serial dependence.

3.2. *Gap Bootstrap I.* Let $\mathbf{X}_{(j)}=(\mathbf{X}_{ip+j}:i=0,\dots,m-1)$ denote the jth row of the array $\mathbb{X}$ in (2.1). For the time being, assume that the rows of $\mathbb{X}$ are exchangeable, that is, for any permutation $(j_1,\dots,j_p)$ of the integers $(1,\dots,p)$, $\{\mathbf{X}_{(j_1)},\dots,\mathbf{X}_{(j_p)}\}$ have the same joint distribution as $\{\mathbf{X}_{(1)},\dots,\mathbf{X}_{(p)}\}$, although we do not need the full force of exchangeability for the validity of the method (cf. Section 4). For notational compactness, set $\mathbf{X}_{(0)}=\mathbb{X}$. Next suppose that the parameter θ can be estimated by using the row variables $\mathbf{X}_{(j)}$ as well as using the complete data set, through estimating equations of the form

$$\Psi_j(\mathbf{X}_{(j)};\theta)=0, \qquad j=0,1,\dots,p,$$

resulting in the estimators $\hat{\theta}_{jn}$, based on the jth row, for $j = 1, \dots, p$, and the estimator $\hat{\theta}_n = \hat{\theta}_{0n}$ for $j = 0$ based on the entire data set, respectively. It is obvious that for large values of p, the computation of $\hat{\theta}_{jn}$'s can be much simpler than that of $\hat{\theta}_n$, as the estimators $\hat{\theta}_{jn}$'s are based on a fraction (namely, $\frac{1}{p}$) of the total observations. On the other hand, the individual $\hat{\theta}_{jn}$'s lose efficiency, as they are based on a subset of the data. However, under some mild conditions on the score functions, the M-estimators can be asymptotically linearized by using the averages of the influence functions over the respective data sets $\mathbf{X}_{(j)}$ [cf. Chapter 7, Serfling (1980)]. As a result, under such regularity conditions,

$$\bar{\theta}_n \equiv p^{-1} \sum_{j=1}^{p} \hat{\theta}_{jn} \tag{3.1}$$

gives an asymptotically equivalent approximation to $\hat{\theta}_n$. Now an estimator of the variance of the original estimator $\hat{\theta}_n$ can be obtained by combining the variance estimators of the $\hat{\theta}_{jn}$'s through the equation

$$\operatorname{Var}(\bar{\theta}_n) = p^{-2} \left[\sum_{j=1}^{p} \operatorname{Var}(\hat{\theta}_{jn}) + \sum_{1 \le j \ne k \le p} \operatorname{Cov}(\hat{\theta}_{jn}, \hat{\theta}_{kn}) \right]. \tag{3.2}$$

Note that using the i.i.d. assumption on the row variables, an estimator of $\operatorname{Var}(\hat{\theta}_{jn})$ can be found by the ordinary bootstrap method (also referred to as the *i.i.d. bootstrap* in here) of Efron (1979) that selects a with replacement sample of size m from the jth row of data values. We denote this by $\widehat{\operatorname{Var}}(\hat{\theta}_{jn})$ (and also by $\hat{\Sigma}_{jn}$), $j = 1, \dots, p$. Further, under the exchangeability assumption, all the covariance terms are equal and, hence, we may estimate the cross-covariance terms by estimating the variance of the pairwise differences as follows:

$$\widetilde{\operatorname{Var}}(\hat{\theta}_{j_0 n} - \hat{\theta}_{k_0 n}) = \frac{\sum_{1 \le j \ne k \le p} (\hat{\theta}_{jn} - \hat{\theta}_{kn})(\hat{\theta}_{jn} - \hat{\theta}_{kn})'}{p(p-1)}, \qquad 1 \le j_0 \ne k_0 \le p.$$

Then, the cross covariance estimator is given by

$$\widetilde{\operatorname{Cov}}(\hat{\theta}_{j_0 n}, \hat{\theta}_{k_0 n}) = [\hat{\Sigma}_{j_0 n} + \hat{\Sigma}_{k_0 n} - \widetilde{\operatorname{Var}}(\hat{\theta}_{j_0 n} - \hat{\theta}_{k_0 n})]/2.$$

Plugging in these estimators of the variance and the covariance terms in (3.2) yields the *Gap Bootstrap Method I estimator of the variance* of $\hat{\theta}_n$ as

$$\widehat{\operatorname{Var}}_{\text{GB-I}}(\hat{\theta}_n) = p^{-2} \left[\sum_{j=1}^{p} \widehat{\operatorname{Var}}(\hat{\theta}_{jn}) + \sum_{1 \le j \ne k \le p} \widetilde{\operatorname{Cov}}(\hat{\theta}_{jn}, \hat{\theta}_{kn}) \right]. \tag{3.3}$$

Note that the estimator proposed here only requires computation of the parameter estimators based on the p subsets, which can cut down on the computational complexity significantly when p is large.

3.3. *Gap Bootstrap II.* In this section we describe a Gap Bootstrap method for the more general case where the rows $\mathbf{X}_{(j)}$'s in (2.1) are not necessarily exchangeable and, hence, do not have the same distribution. Further, we allow the columns of $\mathbb{X}$ to have certain serial dependence. This, for example, is the situation when the $\mathbf{X}_t$-series is obtained from a weakly dependent parent series $\{\mathbf{Y}_t\}$ by systematic deletion of q-components, creating the "gap" structure in the observed $\mathbf{X}_t$-series as described in Section 2. If the $\mathbf{Y}_t$-series is m_0-dependent with an $m_0 < q$, then $\{\mathbf{X}_t\}$ satisfies the conditions in (2.2). For a mixing sequence $\mathbf{Y}_t$, the gapped segments are never exactly independent, but the effect of the dependence on the gapped segments are practically negligible for large enough "gaps," so that approximate independence of the columns holds when q is large. We restrict attention to the simplified structure (2.2) to motivate the main ideas and to keep the exposition simple. Validity of the theoretical results continue to hold under weak dependence among the columns of the array (2.1); see Section 4 for further details.

As in the case of Gap Bootstrap I, we suppose that the parameter θ can be estimated by using the row variables $\mathbf{X}_{(j)}$ as well as using the complete data set, resulting in the estimator $\hat{\theta}_{jn}$, based on the jth row for $j = 1, \dots, p$ and the estimator $\hat{\theta}_n = \hat{\theta}_{0n}$ (for $j = 0$) based on the entire data set, respectively. The estimation method can be any standard method, including those based on score functions and quasi-maximum likelihood methods, such that the following *asymptotic linearity condition* holds:

There exist known weights $w_{1n}, \dots, w_{pn} \in [0, 1]$ with $\sum_{j=1}^{p} w_{jn} = 1$ such that

$$\hat{\theta}_n - \sum_{j=1}^{p} w_{jn}\hat{\theta}_{jn} = o_P(n^{-1/2}) \qquad \text{as } n \to \infty. \tag{3.4}$$

Classes of such estimators are given by (i) L-, M- and R-estimators of location parameters [cf. Koul and Mukherjee (1993)], (ii) differentiable functionals of the (weighted) empirical process [cf. Serfling (1980), Koul (2002)], and (iii) estimators satisfying the smooth function model [cf. Hall (1992), Lahiri (2003)]. An explicit example of an estimator satisfying (3.4) is given in Remark 3.5 below [cf. (3.9)] and the details of verification of (3.4) are given in the Appendix.

Note that under (3.4), the asymptotic variance of $n^{1/2}(\hat{\theta}_n - \theta)$ is given by the asymptotic variance of $\sum_{j=1}^{p} w_{jn} n^{1/2}(\hat{\theta}_{jn} - \theta)$. The latter involves both variances and covariances of the row-wise estimators $\hat{\theta}_{jn}$'s. The Gap Bootstrap method II estimator of the variance of $\hat{\theta}_n$ is obtained by combining individual variance estimators of the marginal estimators $\hat{\theta}_{jn}$'s with estimators of their cross covariances. Note that as the row-wise estimators

$\hat{\theta}_{jn}$ are based on (approximately) i.i.d. data, as in the case of Gap Bootstrap method I, one can use the i.i.d. bootstrap method of Efron (1979) *within* each row $\mathbf{X}_{(j)}$ and obtain an estimator of the standard error of each $\hat{\theta}_{jn}$. We continue to denote these by $\widehat{\mathrm{Var}}(\hat{\theta}_{jn})$, $1 \le j \le p$, as in Section 3.2. However, since we now allow the presence of temporal dependence *among* the rows, resampling individual observations is not enough [cf. Singh (1981)] for cross-covariance estimation and some version of block resampling is needed [cf. Künsch (1989), Lahiri (2003)]. As explained earlier, repeated computation of the estimator $\hat{\theta}_n$ based on replicates of the full sample may not be feasible merely due to the associated computational costs. Instead, computation of the replicates on smaller portions of the data may be much faster (as it avoids *repeated* resampling) and stable. This motivates us to consider the sampling window method of Politis and Romano (1994) and Hall and Jing (1996) for cross-covariance estimation. Compared to the block bootstrap methods, the sampling window method is computationally much faster but at the same time, it typically achieves the same level of accuracy as the block bootstrap covariance estimators, asymptotically [cf. Lahiri (2003)]. The main steps of the Gap Bootstrap Method II are as follows.

3.3.1. *The univariate parameter case.* For simplicity, we first describe the steps of the Gap Bootstrap Method II for the case where the parameter θ is *one-dimensional*:

Steps:

(I) Use i.i.d. resampling of individual observations within each row to construct a bootstrap estimator $\widehat{\mathrm{Var}}(\hat{\theta}_{jn})$ of $\mathrm{Var}(\hat{\theta}_{jn})$, $j = 1, \ldots, p$, as in the case of Gap Bootstrap method I. In the one-dimensional case, we will denote these by $\hat{\sigma}^2_{jn}$, $j = 1, \ldots, p$.

(II) The Gap Bootstrap II estimator of the asymptotic variance of $\hat{\theta}_n$ is given by

$$\bar{\tau}^2_n = \sum_{j=1}^{p} \sum_{k=1}^{p} w_{jn} w_{kn} \hat{\sigma}_{jn} \hat{\sigma}_{kn} \tilde{\rho}_n(j,k), \tag{3.5}$$

where $\hat{\sigma}^2_{jn}$ is as in Step I and where $\tilde{\rho}_n(j,k)$ is the sampling window estimator of the asymptotic correlation between $\hat{\theta}_{jn}$ and $\hat{\theta}_{kn}$, described below.

(III) To estimate the correlation $\rho_n(j,k)$ between $\hat{\theta}_{jn}$ and $\hat{\theta}_{kn}$ by the sampling window method [cf. Politis and Romano (1994) and Hall and Jing (1996)], first fix an integer $\ell \in (1, m)$. Also, let

$$\mathbf{X}^{(1)} = (\mathbf{X}_1, \ldots, \mathbf{X}_p), \qquad \mathbf{X}^{(2)} = (\mathbf{X}_{p+1}, \ldots, \mathbf{X}_{2p}), \ldots,$$
$$\mathbf{X}^{(m)} = (\mathbf{X}_{(m-1)p+1}, \ldots, X_{mp})$$

denote the columns of the matrix array (2.1). The version of the sampling window method that we will employ here will be based on (overlapping) subseries of ℓ columns. The following are the main steps of the sampling window method:

(IIIa) Define the overlapping subseries of the column-variables $\mathbf{X}^{(\cdot)}$ of length ℓ as

$$\mathcal{X}_i = (\mathbf{X}^{(i)}, \ldots, \mathbf{X}^{(i+\ell-1)}), \qquad i = 1, \ldots, I,$$

where $I = m - \ell + 1$. Note that each subseries $\mathcal{X}_i$ contains ℓ complete columns or periods and consists of ℓp-many $\mathbf{X}_t$-variables.

(IIIb) Next, for each $i = 1, \ldots, I$, we employ the given estimation algorithm to the $\mathbf{X}_t$-variables in $\mathcal{X}_i$ to construct the subseries version $\tilde{\theta}_{jn}^{(i)}$ of $\hat{\theta}_{jn}$, $j = 1, \ldots, p$. (There is a slight abuse of notation here, as the sample size for the ith subseries of $\mathbf{X}_t$-variables is ℓp, not $n = mp$ and, hence, we should be using $\tilde{\theta}_{j(\ell p)}^{(i)}$ instead of $\tilde{\theta}_{jn}^{(i)}$, but we drop the more elaborate notation for simplicity).

(IIIc) For $1 \le j < k \le p$, the sampling window estimator of the correlation between $\hat{\theta}_{jn}$ and $\hat{\theta}_{kn}$ is given by

$$\tilde{\rho}_n(j,k) = \frac{I^{-1}\sum_{i=1}^{I}(\tilde{\theta}_{jn}^{(i)} - \hat{\theta}_n)(\tilde{\theta}_{kn}^{(i)} - \hat{\theta}_n)}{[I^{-1}\sum_{i=1}^{I}(\tilde{\theta}_{jn}^{(i)} - \hat{\theta}_n)^2]^{1/2}[I^{-1}\sum_{i=1}^{I}(\tilde{\theta}_{kn}^{(i)} - \hat{\theta}_n)^2]^{1/2}}. \tag{3.6}$$

3.3.2. *The multivariate parameter case.* The multivariate version of the Gap bootstrap estimator of the variance matrix of a vector parameter estimator $\hat{\theta}_n$ can be derived using the same arguments, with routine changes in the notation. Let $\hat{\Sigma}_{jn}$ denote the bootstrap estimator of $\mathrm{Var}(\hat{\theta}_{jn})$, based on the i.i.d. bootstrap method of Efron (1979). Next, with the subsampling replicates $\tilde{\theta}_{jn}^{(i)}$, $j = 1, \ldots, p$, based on the overlapping blocks $\{\mathcal{X}_i : i = 1, \ldots, I\}$ of ℓ columns each (cf. Step [III] of Section 3.3.1), define the sampling window estimator $\tilde{\mathcal{R}}_n(j,k)$ of the correlation matrix of $\hat{\theta}_{jn}$ and $\hat{\theta}_{kn}$ as

$$\begin{aligned}\tilde{\mathcal{R}}_n(j,k) = {} & \left[I^{-1}\sum_{i=1}^{I}(\tilde{\theta}_{jn}^{(i)} - \hat{\theta}_n)(\tilde{\theta}_{jn}^{(i)} - \hat{\theta}_n)'\right]^{-1/2} \\ & \times \left\{I^{-1}\sum_{i=1}^{I}(\tilde{\theta}_{jn}^{(i)} - \hat{\theta}_n)(\tilde{\theta}_{km}^{(i)} - \hat{\theta}_n)'\right\} \\ & \times \left[I^{-1}\sum_{i=1}^{I}(\tilde{\theta}_{km}^{(i)} - \hat{\theta}_n)(\tilde{\theta}_{km}^{(i)} - \hat{\theta}_n)'\right]^{-1/2}.\end{aligned}$$

Then the variance estimator based on Gap bootstrap II is given by

$$\widehat{\mathrm{Var}}_{\text{GB-II}}(\hat{\theta}_n) = \sum_{j=1}^{p}\sum_{k=1}^{p} w_{jn} w_{kn} \hat{\Sigma}_{jn}^{1/2} \tilde{\mathcal{R}}_n(j,k) \hat{\Sigma}_{kn}^{1/2}. \tag{3.7}$$

3.3.3. *Some comments on Method II.*

REMARK 3.1. Note that for estimators $\{\tilde{\theta}_{jn} : j = 1, \ldots, p\}$ with large asymptotic variances, estimation of the correlation coefficients by the sampling window method is more stable, as these are *bounded* (and have a compact support). On the other hand, the asymptotic variances of $\hat{\theta}_{jn}$'s have an unbounded range of values and therefore are more difficult to estimate accurately. Since variance estimation by Efron (1979)'s bootstrap has a higher level of accuracy [e.g., $O_P(n^{-1/2})$] compared to the sampling window method variance estimation [with the slower rate $O_P([\ell/n]^{1/2} + \ell^{-1})$; see Lahiri (2003)], the proposed approach is expected to lead to a better overall performance than a direct application of the sampling window method to estimate the variance of $\hat{\theta}_n$.

REMARK 3.2. Note that all estimators computed here (apart from a one-time computation of $\hat{\theta}_n$ in the sampling window method) are based on subsamples and hence are computationally simpler than repeated computation of $\hat{\theta}_n$ required by naive applications of the block resampling methods.

REMARK 3.3. For applying Gap Bootstrap II, the user needs to specify the block length l. Several standard block length selection rules are available in the block resampling literature [cf. Chapter 7, Lahiri (2003)] for estimating the variance–covariance parameters. Any of these are applicable in our problem. Specifically, we mention the plug-in method of Patton, Politis and White (2009) that is computationally simple and, hence, is specially suited for large data sets.

REMARK 3.4. The proposed estimator remains valid (i.e., consistent) under more general conditions than (2.2), where the columns of the array (2.1) are not necessarily independent. In particular, the proposed estimator in (3.7) remains consistent even when the $\mathbf{X}_t$ variables in the array (2.1) are obtained by creating "gaps" in a weakly dependent (e.g., strongly mixing) parent time series $\mathbf{Y}_t$. This is because the subsampling window method employed in the construction of the cross-correlation can effectively capture the residual dependence structure among the columns of the array (2.1). The use of i.i.d. bootstrap to construct the variance estimators $\hat{\Sigma}_{jn}$ is adequate when the gap is large, as the separation of two consecutive random variables within a row makes the correlation negligible. See Theorem 4.2 below and its proof in the Appendix.

REMARK 3.5. An alternative, intuitive approach to estimating the variance of $\hat{\theta}_n$ is to consider the data array (2.1) by columns rather than by rows. Let $\hat{\theta}^{(1)}, \ldots, \hat{\theta}^{(m)}$ denote the estimates of θ based on the m *columns* of the data matrix $\mathbb{X}$. Then, assuming that the columns of $\mathbb{X}$ are (approximately) independent and assuming that $\hat{\theta}^{(1)}, \ldots, \hat{\theta}^{(m)}$ are identically distributed, one may be tempted to estimate $\mathrm{Var}(\hat{\theta}_n)$ by using the sample variance of the $\hat{\theta}^{(1)}, \ldots, \hat{\theta}^{(m)}$, based on the following analog of (3.1):

$$\hat{\theta}_n \approx m^{-1} \sum_{k=1}^{m} \hat{\theta}^{(k)}. \tag{3.8}$$

However, when p is small compared to m, such an approximation is suboptimal, and this approach may drastically fail if p is fixed. As an illustrating example, consider the case where the $\mathbf{X}_i$'s are 1-dimensional random variables, $p \geq 1$ is fixed (i.e., it does not depend on the sample size), $n = mp$, and the columns $\mathbf{X}^{(k)}, k = 1, \ldots, m$, have an "identical distribution" with mean vector $(\mu, \ldots, \mu)' \in \mathbb{R}^p$ and $p \times p$ covariance matrix Σ. For simplicity, also suppose that the diagonal elements of Σ are all equal to $\sigma^2 \in (0, \infty)$. Let

$$\hat{\theta}_n = n^{-1} \sum_{i=1}^{n} (X_i - \bar{X}_n)^2,$$

an estimator of $\theta = p^{-1}\,\mathrm{trace}(\Sigma) = \sigma^2$. Let $\hat{\theta}^{(k)}$ and $\hat{\theta}_{jn}$, respectively, denote the sample variance of the X_t's in the kth column and the jth row, $k = 1, \ldots, m$ and $j = 1, \ldots, p$. Then, in Appendix A.1, we show that

$$\hat{\theta}_n = p^{-1} \sum_{j=1}^{p} \hat{\theta}_{jn} + o_p(n^{-1/2}), \tag{3.9}$$

while

$$\hat{\theta}_n = m^{-1} \sum_{k=1}^{m} \hat{\theta}^{(k)} + p^{-2}\mathbf{1}'\Sigma\mathbf{1} + O_p(n^{-1/2}), \tag{3.10}$$

where $\mathbf{1}$ is the $p \times 1$ vector of 1's. Thus, in this example, (3.4) holds with $w_{jn} = p^{-1}$ for $1 \leq j \leq p$. However, (3.10) shows that the column-wise approach based on (3.8) results in a very crude approximation which *fails* to satisfy an analog of (3.4). For estimating the variance of $\hat{\theta}_n$, the deterministic term $p^{-2}\mathbf{1}'\Sigma\mathbf{1}$ has no effect, but the $O_p(n^{-1/2})$-term in (3.10) has a nontrivial contribution to the bias of the resulting column-based variance estimator, which can *not* be made negligible. As a result, this alternative approach fails to produce a consistent estimator for fixed p. In general, caution must be exercised while applying the column-wise method for small p.

4. Theoretical results.

4.1. *Consistency of Gap Bootstrap I estimator.* The Gap Bootstrap I estimator $\widehat{\mathrm{Var}}_{\mathrm{GP\text{-}I}}(\hat{\theta}_n)$ of the (asymptotic) variance matrix of $\hat{\theta}_n$ is consistent under fairly mild conditions, as stated in Appendix A.2. Briefly, these conditions require (i) homogeneity of pairwise distributions of the centered and scaled estimators $\{m^{1/2}(\hat{\theta}_{jn} - \theta) : 1 \le j \le p\}$, (ii) some moment and weak dependence conditions on the $m^{1/2}(\hat{\theta}_{jn} - \theta)$'s, and (iii) $p \to \infty$ as $n \to \infty$. In particular, the rows of $\mathbb{X}$ need *not* be exchangeable. Condition (iii) is needed to ensure consistency of the estimator of the covariance term(s) in (3.3), which is defined in terms of the average of the $p(p-1)$ pair-wise differences $\{\hat{\theta}_{jn} - \hat{\theta}_{kn} : 1 \le j \ne k \le p\}$. Thus, for employing the Gap Bootstrap I method in an application, $p(p-1)$ should not be too small,

The following result asserts consistency of the Gap Bootstrap I variance (matrix) estimator.

THEOREM 4.1. *Under conditions* (A.1) *and* (A.2) *given in the Appendix, as* $n \to \infty$,

$$n[\widehat{\mathrm{Var}}_{\mathrm{GB\text{-}I}}(\hat{\theta}_n) - \mathrm{Var}(\bar{\theta}_n)] \to 0 \qquad \textit{in probability.}$$

4.2. *Consistency of Gap Bootstrap II estimator.* Next consider the Gap Bootstrap II estimator of the (asymptotic) variance matrix of $\hat{\theta}_n$. Consistency of $\widehat{\mathrm{Var}}_{\mathrm{GB\text{-}II}}(\hat{\theta}_n)$ holds here under suitable regularity conditions on the estimators $\{\hat{\theta}_{jn} : 1 \le j \le p\}$ and the length of the "gap" q for a large class of time series that allows the rows of the array (2.1) to have nonidentical distributions. See the Appendix for details of the conditions and their implications. It is worth noting that unlike Gap Bootstrap I, here the column dimension p need not go to infinity for consistency.

THEOREM 4.2. *Under conditions* (C.1)–(C.4), *given in the Appendix, as* $n \to \infty$,

$$n[\widehat{\mathrm{Var}}_{\mathrm{GB\text{-}II}}(\hat{\theta}_n) - \mathrm{Var}(\hat{\theta}_n)] \to 0 \qquad \textit{in probability.}$$

5. Simulation results. To investigate finite sample properties of the proposed methods, we conducted a moderately large simulation study involving different univariate and multivariate time series models. For the univariate case, we considered three models:

(I) Autoregressive (AR) models of order two ($X_t = \mu + Y_t$ where $Y_t = \alpha_1 Y_{t-1} + \alpha_2 Y_{t-2} + W_t$).

(II) Moving average (MA) models of order two ($X_t = \mu + Y_t$ where $Y_t = \beta_1 W_{t-1} + \beta_2 W_{t-2} + W_t$).

(III) A periodic time series model ($X_t = \mu_t + W_t$, $W_t = \sigma \varepsilon_t$), where $W_t = \sigma\varepsilon_t$ and $\{\varepsilon_t\}$ are i.i.d. random variables with zero mean and unit variance. The parameter values of the AR models are $\alpha_1 = 0.8, \alpha_2 = 0.1$ with constant mean $\mu = 0.1$ and with $\sigma = 0.2$. Similarly, for the MA models, we took the MA-parameters as $\beta_1 = 0.3$, $\beta_2 = 0.5$, and set $\sigma = 0.2$ and $\mu = 0.1$. For the third model, the mean of the X_t-variables were taken as a periodic function of time t:

$$\mu_t = \mu + \cos 2\pi t/p + \sin 2\pi t/p$$

with $\mu = 1.0$ and $p \in \{5, 10, 20\}$ and with $\sigma = 0.2$. In all three cases, the ε_t are generated from two distributions, namely, (i) $N(0,1)$-distribution and (ii) a centered Exponential (1) distribution, to compare the effects of nonnormality on the performance of the two methods. Note that the rows of the generated $\mathbb{X}$ are identically distributed for models I and II but not for model III. We considered six combinations of (n,p) where n denotes the sample size and p the number of time slots (or the periodicity). The parameter of interest θ was the population mean and the estimator $\hat{\theta}_n$ was taken to be the sample mean. Thus, the row-wise estimators $\hat{\theta}_{jn}$ were the sample means of the row-variables and the weights in (3.4) were $w_{jn} = 1/p$ for all $j = 1, \ldots, p$. In all, there are $(3 \times 2 \times 6 =)$ 36 possible combinations of (n,p)-pairs, the error distributions, and the three models. To keep the size of the paper to a reasonable length, we shall only present 3 combinations of (n,p) in the tables, while we present side-by-side box-plots for all 6 combinations of (n,p), arranged by the error distributions. All results are based on 500 simulation runs.

Figures 2 and 3 give the box-plots of the differences between the Gap Bootstrap I standard error estimates and the true standard errors in the one-dimensional case under centered exponential and under normal error distributions, respectively. Here box-plots in the top panels are based on the AR(2) model, the middle panels are based on the MA(2) model, while the bottom panels are based on the periodic model. For each model, the combinations of (n,p) are given by $(n,p) = (200,5), (500,10), (1800,30), (3500,50), (6000,75), (10{,}000,100)$.

Similarly, Figures 4 and 5 give the corresponding box-plots for the Gap Bootstrap II method under centered exponential and under normal error distributions, respectively.

From the Figures 4 and 5, it is evident that the variability of the standard error estimates from the Gap Bootstrap I Method is higher under Models I and II than under Model III for both error distributions. However, the bias under Model III is persistently higher even for larger values of the sample size. This can be explained by noting that for Method I, the assumption of approximate exchangeability of the rows is violated under the periodic mean structure of Model III, leading to a bigger bias. In comparison, Gap

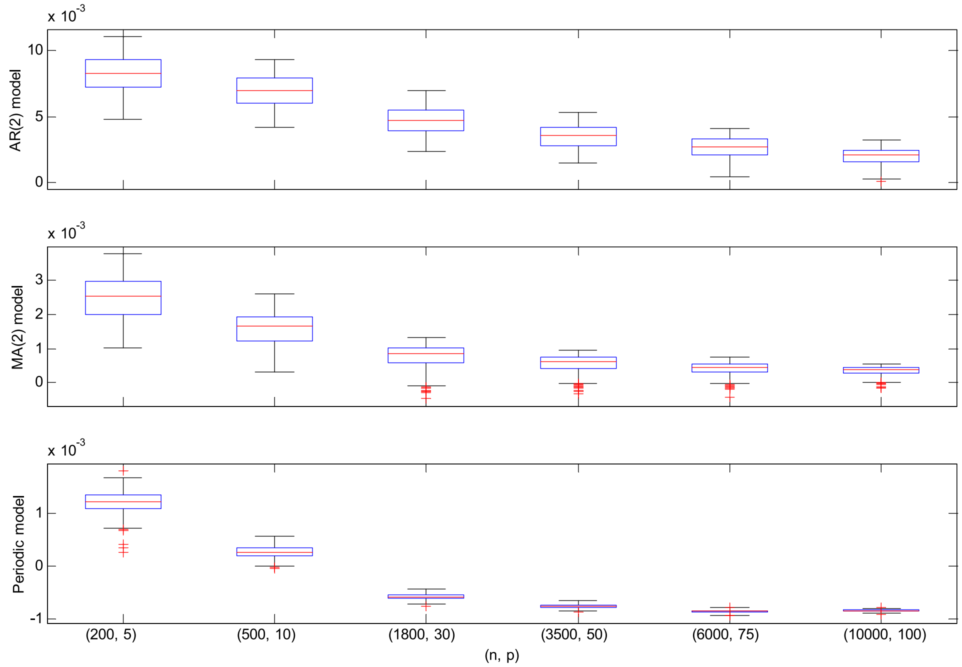

FIG. 2. *Box-plots of the differences between the standard error estimates based on Gap Bootstrap I and the true standard errors in the one-dimensional case using* 500 *simulation runs. Here, plots in the first panel are based on Model I, those in the second and third panels are based on Models II and III, respectively. The values of* (n, p) *for each box-plot are given at the bottom of the third panel. The innovation distribution is centered exponential.*

Bootstrap II estimates tend to center around the target value (i.e., with differences around zero) even for the periodic model. Table 1 gives the true values of the standard errors of $\hat{\theta}_n$ based on Monte-Carlo simulation and the corresponding summary measures for Gap Bootstrap methods I and II in 18 out of the 36 cases [we report only the first 3 combinations of (n, p) to save space. A similar pattern was observed in the other 18 cases].

From the table, we make the following observations:

(i) The biases of the Gap Bootstrap I estimators are consistently higher than those based on Method II under Models I and II for both normal and nonnormal errors, resulting in higher overall MSEs for Gap Bootstrap I estimators.

(ii) Unlike under Models I and II, here the biases of the two methods can have opposite signs.

(iii) From the last column of Table 1 (which gives the ratios of the MSEs of estimators based on Methods I and II), it follows that the Gap Bootstrap II works significantly better than Gap Bootstrap I for Models I and II. For

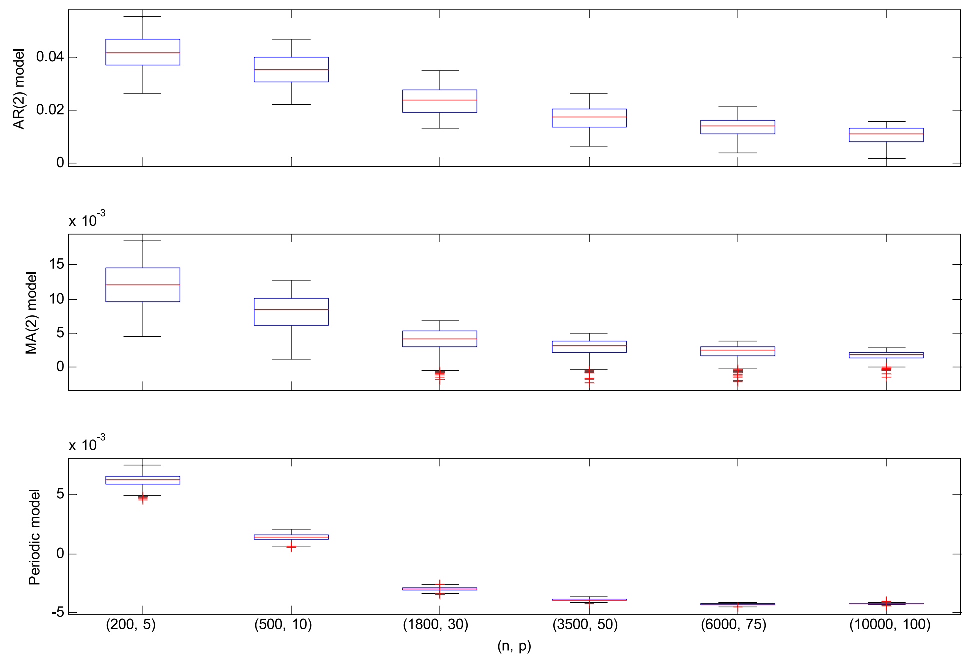

FIG. 3. *Box-plots for the differences of Gap Bootstrap I estimates and the true standard errors as in Figure 2, but under normal innovation distribution.*

Model III, neither method dominates the other in terms of bias and/or MSE. MSE comparison shows a curious behavior of Method I at $(n, p) = (500, 10)$ for the periodic model.

(iv) The nonnormality of the $\mathbf{X}_t$'s does not seem to have significant effects on the relative accuracy of the two methods.

Next we consider performance of the two gap Bootstrap methods for multivariate data. The models we consider are analogs of (I)–(III) above, with the general structure

$$\mathbf{Y}_t = (0.2, 0.3, 0.4, 0.5)' + \mathbf{Z}_t, \qquad t \geq 1,$$

where $\mathbf{Z}_t$ is taken to be the following: (IV) a multivariate autoregressive (MAR) process, (V) a multivariate moving average (MMA) process, and (VI) a multivariate periodic process. For the MAR process,

$$\mathbf{Z}_t = \Psi \mathbf{Z}_{t-1} + \mathbf{e}_t,$$

where

$$\Psi = \begin{bmatrix} 0.5 & 0 & 0 & 0 \\ 0.1 & 0.6 & 0 & 0 \\ 0 & 0 & -0.2 & 0 \\ 0 & 0.1 & 0 & 0.4 \end{bmatrix}$$

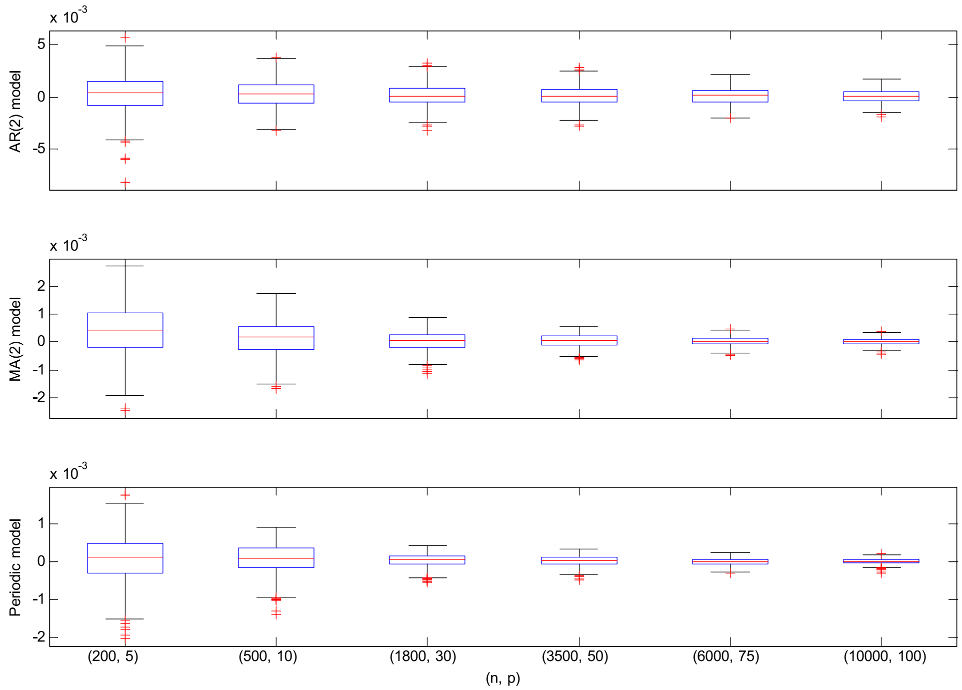

FIG. 4. *Box-plots of the differences of standard error estimates based on Gap Bootstrap II and the true standard errors in the one-dimensional case, as in Figure 2, under the centered exponential innovation distribution.*

and the $\mathbf{e}_t$ are i.i.d. $d = 4$ dimensional normal random vectors with mean 0 and covariance matrix Σ_0, where we consider two choices of Σ_0:

(i) Σ_0 is the identity matrix of order 4;
(ii) Σ_0 has (i, j)th element given by $(-\rho)^{|i-j|}$, $1 \le i, j \le 4$, with $\rho = 0.55$.

For the MMA model, we take

$$\mathbf{Z}_t = \Phi_1 \mathbf{e}_{t-1} + \Phi_2 \mathbf{e}_{t-2} + \mathbf{e}_t,$$

where $\mathbf{e}_t$ are as above. The matrix of MA coefficients are given by

$$\Phi_1 = \begin{bmatrix} 1 & 0 & 0 & 0 \\ * & 2 & 0 & 0 \\ * & * & 2 & 0 \\ * & * & * & 2 \end{bmatrix} \quad \text{and} \quad \Phi_2 = \frac{1}{8} \begin{bmatrix} 1 & 0 & 0 & 0 \\ * & 1 & 0 & 0 \\ * & * & 1 & 0 \\ * & * & * & 1 \end{bmatrix},$$

where, in both Φ_1 and Φ_2, the $*$'s are generated by using a random sample from the UNIFORM $(0, 1)$ distribution [i.e., random numbers in $(0, 1)$] and are held fixed throughout the simulation. We take Φ_1 and Φ_2 as lower triangular matrices to mimic the structure of the OD model for the real data

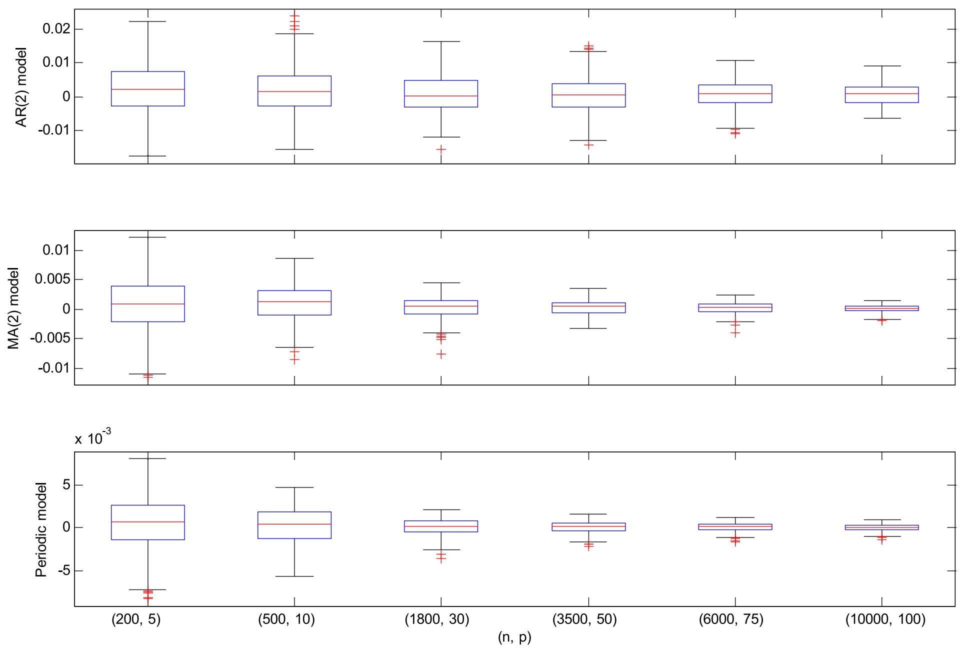

FIG. 5. *Box-plots of the differences of standard error estimates based on Gap Bootstrap II and the true standard errors in the one-dimensional case, as in Figure 2, under the normal innovation distribution.*

example that will be considered in Section 6 below. Finally, the observations $\mathbf{X}_t$ under the periodic model (VI) are generated by stacking the univariate case with the same p, but with μ changed to the the vector $(0.2, 0.3, 0.4, 0.5)$. The component-wise values of α_1 and α_2 are kept the same and the ε_t's for the 4 components are now given by the $\mathbf{e}_t$'s, with the two choices of the covariance matrix.

The parameter of interest is the mean of component-wise means, that is,

$$\theta = \bar{\mu} = [0.2 + 0.3 + 0.4 + 0.5]/4.$$

The estimator $\hat{\theta}_n$ is the mean of the component-wise means of the *entire* data set and $\tilde{\theta}^{(i)}$ is given by the mean of the component-wise means coming from the ith row of n/p-many data vectors, for $j = 1, \ldots, p$. Box-plots of the differences between the true standard errors of $\hat{\theta}_n$ and their estimates obtained by the two Gap Bootstrap methods are reported in Figures 6 and 7, respectively. We only report the results for the models with covariance structure (ii) above (to save space).

The number of simulation runs is 500 as in the univariate case. From the figures it follows that the relative patterns of the box-plots mimic those in the case of the univariate case, with Gap Bootstrap I leading to systematic biases

TABLE 1

Bias and MSEs of Standard Error estimates from Gap Bootstraps I and II for univariate data for Models I–III. For each model, the two sets of 3 rows correspond to $(n,p) = (200,5),(500,10),(1800,30)$ *under the normal (denoted by N in the first column) and the centered Exponential (denoted by E) error distributions, respectively. Here B-I = Bias of Gap Bootstrap I* $\times 10^2$*, M-I = MSE of Gap Bootstrap I* $\times 10^4$*, B-II = Bias of Gap Bootstrap II* $\times 10^3$*, and M-II = MSE of Gap Bootstrap II* $\times 10^4$*. Column 2 gives the target parameter evaluated by Monte-Carlo simulations and the last column is the ratio of columns 4 and 6*

Model	True-se	B-I	M-I	B-II	M-II	Ratio (fix)
I.N.1	0.013	−0.831	0.708	−0.376	0.029	24.4
I.N.2	0.011	−0.700	0.503	−0.118	0.0202	25.2
I.N.3	0.008	−0.481	0.241	−0.256	0.0142	17.2
I.E.1	0.065	−4.18	17.8	−1.97	0.623	28.6
I.E.2	0.053	−3.54	12.8	−1.52	0.451	28.4
I.E.3	0.038	−2.41	6.04	−0.844	0.348	17.4
II.N.1	0.005	−0.240	0.061	−0.178	0.008	7.6
II.N.2	0.003	−0.154	0.026	−0.122	0.004	6.5
II.N.3	0.002	−0.081	0.007	−0.087	0.001	7.0
II.E.1	0.023	−1.22E	1.59	−1.18	0.183	8.9
II.E.2	0.015	−0.767	0.657	−0.288	0.101	6.5
II.E.3	0.008	−0.398	0.184	−0.092	0.025	7.4
III.N.1	0.003	−0.125	0.016	−0.183	0.005	3.2
III.N.2	0.002	−0.0263	0.0008	−0.065	0.002	0.4
III.N.3	0.001	0.059	0.004	−0.028	0.0004	10.0
III.E.1	0.014	−0.619	0.386	−0.549	0.094	4.1
III.E.2	0.009	−0.158	0.026	−0.506	0.042	0.6
III.E.3	0.005	0.292	0.086	−0.216	0.010	8.6

under the periodic mean structure. For comparison, we have also considered the performance of more standard methods, namely, the overlapping versions of the Subsampling (SS) and the Block Bootstrap (BB).

Figures 8 and 9 give box-plots of the differences between the true standard errors of $\hat{\theta}_n$ and their estimates obtained by SS and BB methods, under Models (IV)–(VI) with covariance structure (ii). The choice of the block size was based on the block length selection rule of Patton, Politis and White (2009). From the figures, it follows that the relative performances of the SS and the BB methods are qualitatively similar and both methods handily outperform Gap Bootstrap I.

These qualitative observations are more precisely quantified in Table 2 which gives the MSEs of all 4 methods for models (IV)–(VI) for all six combinations of (n,p) under covariance structure (ii). It follows from the table that Gap Bootstrap Method II has the best overall performance in terms of

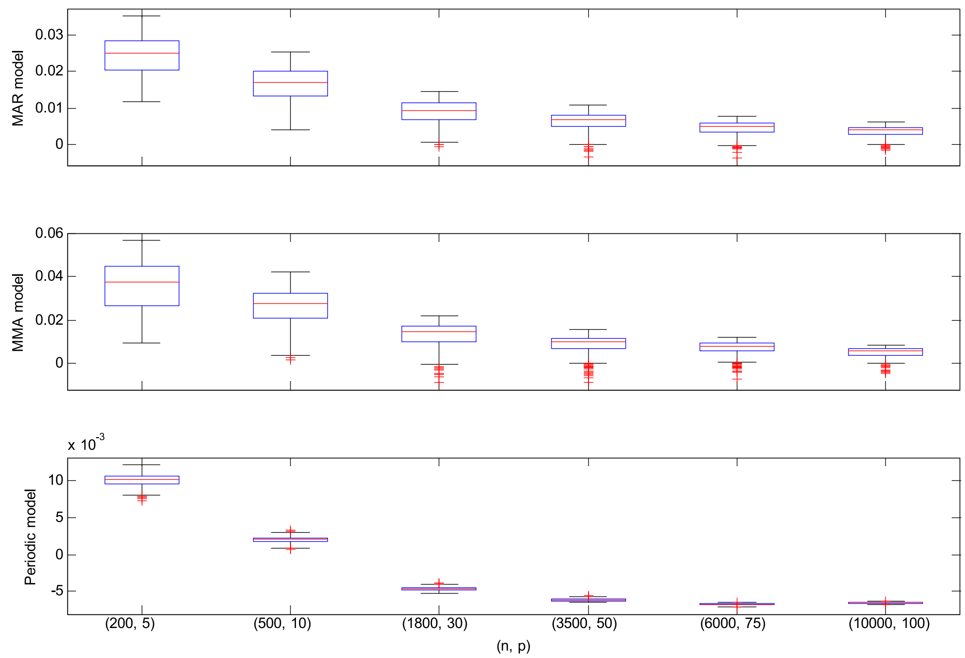

FIG. 6. *Box-plots of the differences of standard error estimates based on Gap Bootstrap I and the true standard errors in the multivariate case, under the Type II error distribution. The number of simulation runs is 500. Also, the models and the values of* (n,p) *are depicted on the panels as in Figure 2.*

the MSE. This may appear somewhat counter-intuitive at first glance, but the gain in efficiency of Gap Bootstrap II can be explained by noting that it results from judicious choices of resampling methods for different parts of the target parameter, as explained in Section 3.3.3 (cf. Remark 3.1). On the other hand, in terms of computational time, Gap Bootstrap I had the best possible performance, followed by the SS, Gap Bootstrap II and the BB methods, respectively. Since the basic estimator $\hat{\theta}_n$ is computationally very simple (being the sample mean), the computational time may exhibit a very different relative pattern (e.g., for $\hat{\theta}_n$ requiring high-dimensional matrix inversion, the BB method based on the entire data set may be totally infeasible).

6. A real data example: The OD estimation problem.

6.1. *Data description.* A 4.9 mile section of Interstate 10 (I-10) in San Antonio, Texas was chosen as the test bed for this study. This section of freeway is monitored as part of San Antonio's TransGuide Traffic Management Center, an intelligent transportation systems application that provides motorists with advanced information regarding travel times, congestion, acci-

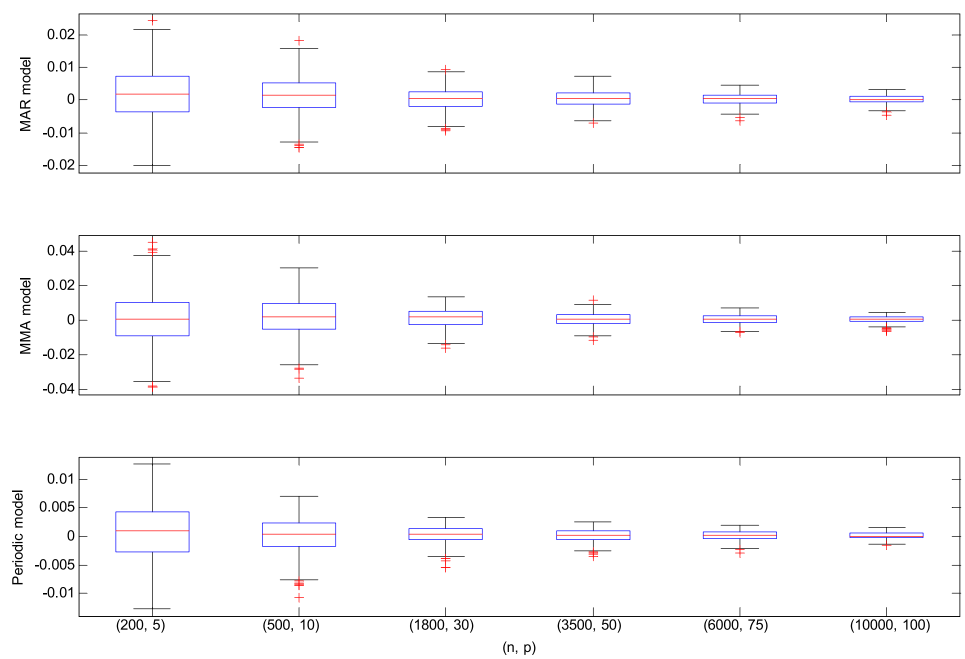

FIG. 7. *Box-plots of the differences of standard error estimates based on Gap Bootstrap II and the true standard errors in the multivariate case, under the setup of Figure 6.*

dents and other traffic conditions. Archived link volume counts from a series of 14 inductive loop detector locations (2 main lane locations, 6 on-ramps and 6 off-ramps) were used in this study (see Figure 1). The analysis is based on 575 days of peak AM (6:30 to 9:30) traffic count data (All weekdays—January 1, 2007 to March 13, 2009). Each day's data were summarized into 36 volume counts of 5-minute duration. Thus, there were a total of 20,700 time points, and each time point giving 14 origin-destination traffic data, resulting in *more than a quarter-million data-values*. Figures 10 and 11 are plots showing the periodic behavior of the link volume count data at the 7 origin (O1 to O7) and 7 destination (D1 to D7) locations, respectively.

6.2. *A synthetic OD model.* As described in Section 2, the OD trip matrix is required in many traffic applications such as traffic simulation models, traffic management, transportation planning and economic development. However, due to the high cost of direct measurements, the OD entries are constructed using synthetic OD models [Cascetta (1984), Bell (1991), Okutani (1987), Dixon and Rilett (2000)]. One common approach for estimating the OD matrix from link volume counts is based on the least squares regression where the unknown OD matrix is estimated by minimizing the squared Euclidean distance between the observed link volumes and the estimated link volumes.

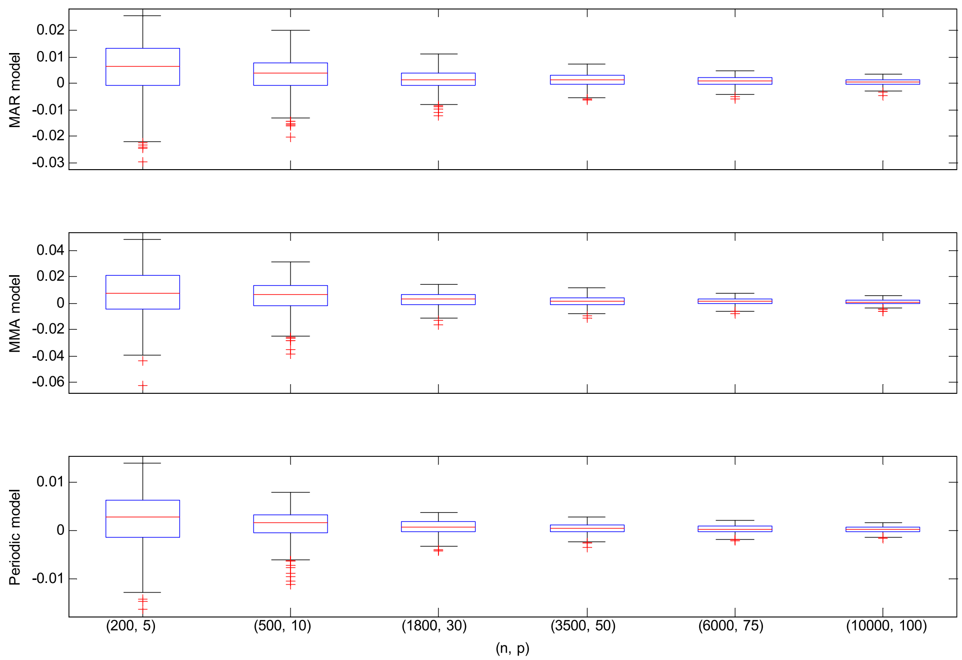

FIG. 8. *Box-plots of the difference of standard error estimates based on Subsampling and the true standard errors in the multivariate case, under the setup of Figure 6.*

Given the link volume counts on all origin and destination ramps, the OD split proportion, p_{ij} (assumed homogeneous over the morning rush-hours), is the fraction of vehicles that exit the system at destination ramp d_{jt} given that they enter at origin ramp o_{it} at time point t (cf. Section 2). Once the split proportions are known, the OD matrix for each time period can be identified as a linear combination of the split proportion matrix and the vector of origin volumes. It should be noted that because the origin volumes are dynamic, the estimated OD matrix is also dynamic. However, the split proportions are typically assumed constant so that the OD matrices by time slice are linear functions of each other [Gajewski et al. (2002)]. While this is a reasonable assumption for short freeway segments over a time span with homogeneous traffic patterns like the ones used in this study, it elicits the question as to when trips began and ended when used on larger networks over a longer tie span. It is also assumed that all vehicles that enter the system from each origin ramp during a given time period exit the system during the same time period. That is, it is assumed that conservation of vehicles holds, so that the sum of the trip proportions from each origin ramp equals 1. Caution should be exercised in situations where a large proportion of trips begin and end during different time periods [Gajewski et al. (2002)]. Note also that some split proportions such as p_{21} are not feasible because of the

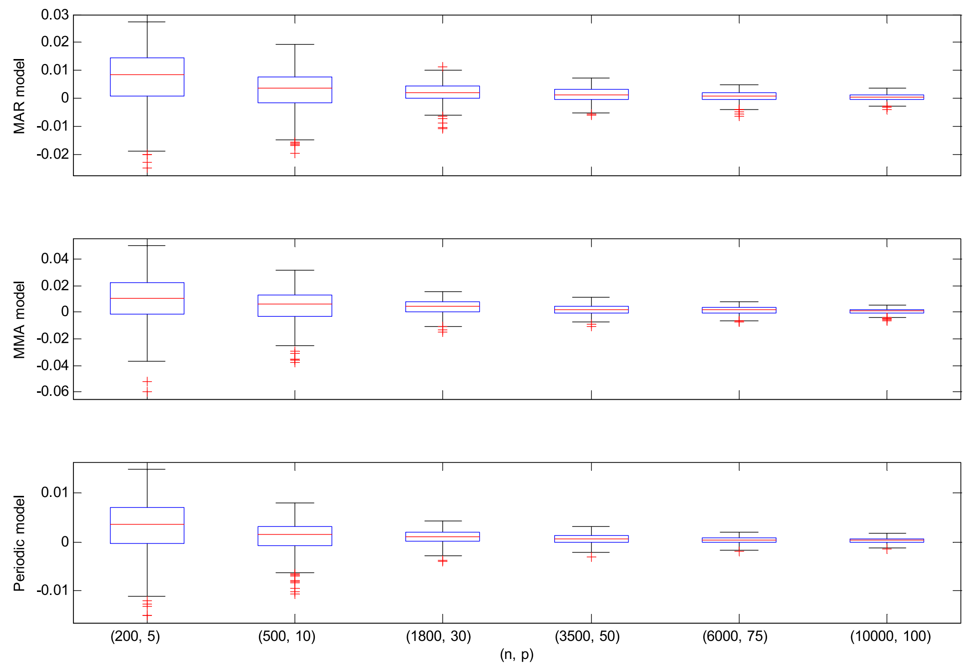

FIG. 9. *Box-plots of the differences of standard error estimates based on the block bootstrap and the true standard errors in the multivariate case, under the setup of Figure 6.*

structure of the network. Moreover, all vehicles that enter the freeway from origin ramp 7 go through destination ramp 7 so that $p_{77} = 1$. All of these constraints need to be incorporated into the estimation process.

Let d_{jt} denote the volume at destination j over the tth time interval (of duration 5 minutes) and o_{jt} denote the jth origin volume over the same period. Let p_{ij} be the proportion of origin i volume contributing to the destination j volume (assumed not to change over time). Then, the synthetic OD model for the link volume counts can be described as follows:

For each t,

$$\begin{aligned}
d_{1t} &= o_{1t}p_{11} + \varepsilon_{1t},\\
d_{2t} &= o_{1t}p_{12} + o_{2t}p_{22} + \varepsilon_{2t},\\
d_{3t} &= o_{1t}p_{13} + o_{2t}p_{23} + o_{3t}p_{33} + \varepsilon_{3t},\\
d_{4t} &= o_{1t}p_{14} + o_{2t}p_{24} + o_{3t}p_{34} + o_{4t}p_{44} + \varepsilon_{4t}, \qquad (6.1)\\
d_{5t} &= o_{1t}p_{15} + o_{2t}p_{25} + o_{3t}p_{35} + o_{4t}p_{45} + o_{5t}p_{55} + \varepsilon_{5t},\\
d_{6t} &= o_{1t}p_{16} + o_{2t}p_{26} + o_{3t}p_{36} + o_{4t}p_{46} + o_{5t}p_{56} + o_{6t}p_{66} + \varepsilon_{6t},\\
d_{7t} &= o_{1t}p_{17} + o_{2t}p_{27} + o_{3t}p_{37} + o_{4t}p_{47} + o_{5t}p_{57} + o_{6t}p_{67}\\
&\quad + o_{7t}p_{77} + \varepsilon_{7t},
\end{aligned}$$

Table 2

MSEs of Standard Error estimates from Gap Bootstraps I *and* II *and the Subsampling (SS) and Block Bootstrap (BB) methods for the multivariate data for Models IV–VI under covariance matrix of type* (ii). *The six rows under each model correspond to* $(n,p)=(200,5),(500,10),(1800,30),(3500,50),(6000,75),(10{,}000,100)$. *Further, the entries in the table gives the values of the MSEs multiplied* 10^4, 10^4 *and* 10^5 *for Models IV–VI, respectively*

Model	**True-se**	**GB-I**	**GB-II**	**SS**	**BB**
IV.1	0.044	6.190	0.634	1.390	1.510
IV.2	0.030	2.970	0.353	0.568	0.567
IV.3	0.017	0.873	0.116	0.151	0.162
IV.4	0.012	0.451	0.064	0.078	0.082
IV.5	0.009	0.247	0.034	0.040	0.042
IV.6	0.007	0.155	0.017	0.020	0.020
V.1	0.076	14.300	2.350	3.690	4.040
V.2	0.053	7.560	1.190	1.650	1.690
V.3	0.028	2.060	0.300	0.374	0.427
V.4	0.019	0.930	0.144	0.165	0.176
V.5	0.015	0.590	0.080	0.094	0.099
V.6	0.011	0.297	0.037	0.043	0.045
VI.1	0.022	10.300	2.400	3.150	3.440
VI.2	0.014	4.250	0.918	1.110	1.100
VI.3	0.007	2.230	0.215	0.257	0.291
VI.4	0.005	3.860	0.111	0.134	0.140
VI.5	0.004	4.620	0.069	0.073	0.074
VI.6	0.003	4.350	0.032	0.036	0.038

where ε_{jt} are (correlated) error variables. Note that the parameters p_{ij} satisfy the conditions

$$\sum_{j=i}^{7} p_{ij}=1 \qquad \text{for } i=1,\ldots,7. \tag{6.2}$$

In particular, $p_{77}=1$. Because of the above linear restrictions on the p_{ij}'s, it is enough to estimate the parameter vector $\mathbf{p}=(p_{11},p_{12},\ldots,p_{16};p_{22},\ldots,p_{26};\ldots;p_{66})'$. We relabel the components and write $\mathbf{p}=(\theta^{[1]},\ldots,\theta^{[21]})'\equiv\theta$. We will estimate these parameters by the least squares method using the entire data, resulting in the estimator $\hat{\theta}_n$ and using the daily data over each of the 36 time intervals of length 5 minutes, yielding $\hat{\theta}_{jn}$, $j=1,\ldots,24$. For notational simplicity, we set $\hat{\theta}_{0n}=\hat{\theta}_n$.

For $t=1,\ldots,20{,}700$, let $D_t=(d_{1t},\ldots,d_{6t},d_{7t}-\sum_{i=1}^{7}o_{1i})'$ and let O_t be the 7×21 matrix given by

$$O_t=[O_t^{[1]}:\ldots:O_t^{[6]}],$$

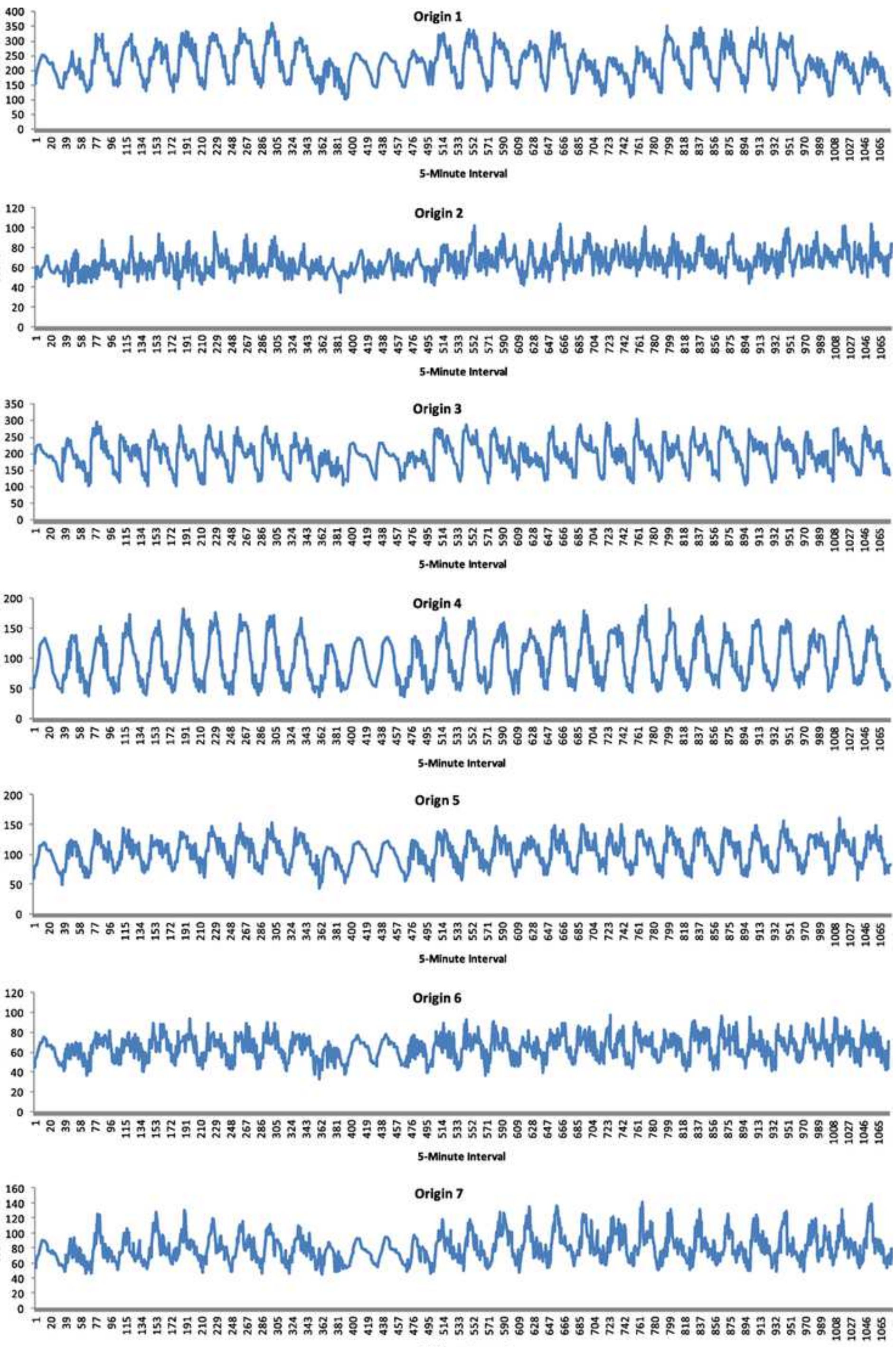

FIG. 10. *Plots of the origin volume counts for the San Antonio, TX data (including weekend days).*

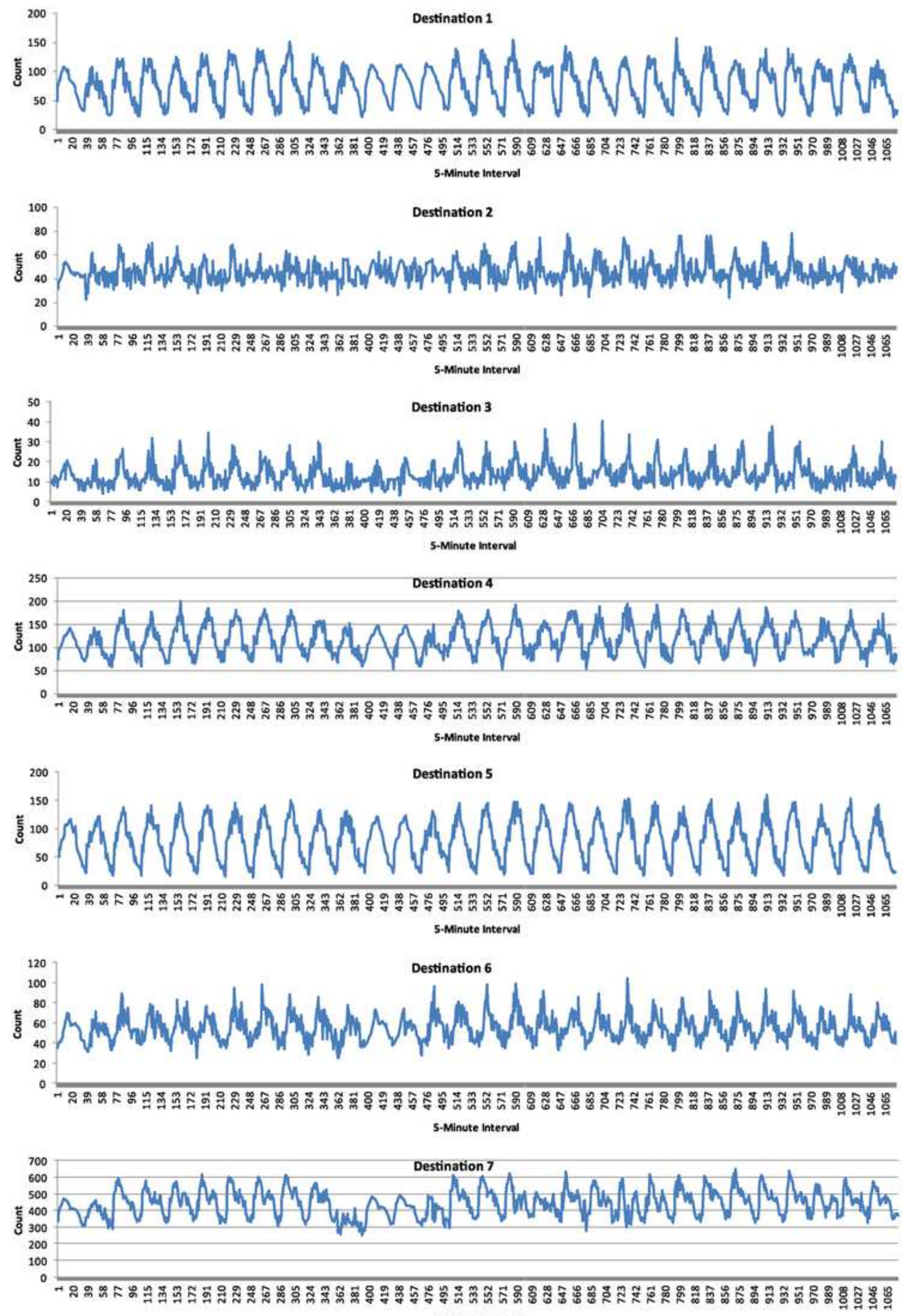

FIG. 11. *Plots of the destination volume counts for the San Antonio, TX data (including weekend days).*

where, for $k = 1, \ldots, 6$, $O_t^{[k]}$ is a $7 \times (7-k)$ matrix with its last row given by $(-o_{kt}, \ldots, -o_{kt})$ and the rest of the elements by

$$(O_t^{[k]})_{ij} = o_{kt}\mathbb{1}(i \geq k)\mathbb{1}(j = i-k+1), \qquad i = 1, \ldots, 6, j = 1, \ldots, 7-k.$$

For $j = 0, 1, \ldots, 36$, let

$$\hat{\theta}_{jn} = \left[\sum_{t \in T_j} O_t' O_t\right]^{-1} \sum_{t \in T_j} O_t' D_t, \tag{6.3}$$

where $T_j = \{j, j+36, \ldots, j + (574 \times 36)\}$ for $j = 1, \ldots, 36$ and where $T_0 = \{1, \ldots, 720\}$. Note that each of $T_1, \ldots, T_{36}$ has size 575 (the total number of days) and corresponds to the counts data over the respective 5 minute period, while T_0 has size 20,700 and it corresponds to the entire data set. For applying Gap Bootstrap II, we need a minor extension of the formulas given in Section 3.3, as the weights in (3.4) now vary component-wise. For $j = 0, 1, \ldots, 36$, define $\Gamma_{jn} = \sum_{t \in T_j} O_t' O_t$. Then, the following version of (3.4) holds [without the $o_p(1)$ term]:

$$\hat{\theta}_n = \sum_{j=1}^{36} W_{jn}\hat{\theta}_{jn},$$

where $W_{jn} = \Gamma_{0n}^{-1}\Gamma_{jn}$. This can be proved by noting that

$$\hat{\theta}_n = \Gamma_{0n}^{-1} \sum_{t \in T_0} O_t' D_t = \Gamma_{0n}^{-1} \sum_{j=1}^{36} \sum_{t \in T_j} O_t' D_t \equiv \sum_{j=1}^{36} W_{jn}\hat{\theta}_{jn}.$$

The Gap Bootstrap II estimator of the variance of the individual components $\hat{\theta}_n^{[1]}, \ldots, \hat{\theta}_n^{[21]}$ of the estimator $\hat{\theta}_n$ is now given by

$$\widehat{\mathrm{Var}}(\hat{\theta}_n^{[a]}) = \sum_{k=1}^{36}\sum_{l=1}^{36} \hat{\sigma}_{ak}\hat{\sigma}_{al}\tilde{\rho}_a(k,l), \qquad a = 1, \ldots, 21,$$

where $\hat{\sigma}_{ak}^2 = \mathbf{w}_{ak}'\hat{\Sigma}^{(k)}\mathbf{w}_{ak}$, $\hat{\Sigma}^{(k)}$ is the i.i.d. bootstrap based estimator of the variance matrix of $\hat{\theta}_{kn}$, $\tilde{\rho}_a(k,j)$ is the sampling window estimator of the correlation between the ath component of the kth and jth row-wise estimators of θ and $\mathbf{w}_{ak}$'s are weights based on W_{jn}'s. Indeed, with $\mathbf{e}_1 = (1, 0, \ldots, 0)', \ldots, \mathbf{e}_{21} = (0, \ldots, 1)'$, we have $\mathbf{w}_{aj} = \mathbf{e}_a'\Gamma_{0n}^{-1}\Gamma_{jn}$, $1 \leq j \leq 36$. To find $\tilde{\rho}_a(k,j)$'s, we applied the sampling window method estimator with $\ell = 17$ and the following formula for $\tilde{\rho}_a(k,j)$:

$$\tilde{\rho}_a(k,j) = \frac{I^{-1}\sum_{i=1}^{I}(\mathbf{w}_{ak}'[\tilde{\theta}_{kn}^{(i)} - \hat{\theta}_n])(\mathbf{w}_{aj}'[\tilde{\theta}_{jn}^{(i)} - \hat{\theta}_n])}{[I^{-1}\sum_{i=1}^{I}(\mathbf{w}_{ak}'[\tilde{\theta}_{kn}^{(i)} - \hat{\theta}_n])^2]^{1/2}[I^{-1}\sum_{i=1}^{I}(\mathbf{w}_{aj}'[\tilde{\theta}_{jn}^{(i)} - \hat{\theta}_n])^2]^{1/2}},$$

TABLE 3

Standard Error estimates from Gap Bootstraps I and II (denoted by STD-I and STD-II, resp.) for the San Antonio, TX data

p_{ij}	Estimates	STD-I	STD-II	p_{ij}	Estimates	STD-I	STD-II
p_{11}	0.355	0.0009	0.0019	p_{33}	0.046	0.0026	0.0041
p_{12}	0.104	0.0018	0.0042	p_{34}	0.232	0.0032	0.0132
p_{13}	0.011	0.0006	0.0015	p_{35}	0.106	0.0061	0.0082
p_{14}	0.064	0.0043	0.0131	p_{36}	0.039	0.0025	0.0080
p_{15}	0.047	0.0024	0.0073	p_{44}	0.436	0.0100	0.0155
p_{16}	0.022	0.0017	0.0042	p_{45}	0.240	0.0123	0.0094
p_{22}	0.385	0.0079	0.0118	p_{46}	0.105	0.0057	0.0141
p_{23}	0.083	0.0044	0.0066	p_{55}	0.233	0.0080	0.0130
p_{24}	0.242	0.0053	0.0237	p_{56}	0.109	0.0045	0.0168
p_{25}	0.112	0.0107	0.0144	p_{66}	0.537	0.0093	0.0263
p_{26}	0.064	0.0037	0.0058				

$j,k=1,\ldots,36$, $a=1,\ldots,21$, where $\tilde{\theta}_{kn}^{(i)}$'s is the ith subsample version of $\hat{\theta}_{kn}$ and $I=575-\ell+1=559$. Following the result on the optimal order of the block size for estimation of (co)-variances in the block resampling literature [cf. Lahiri (2003)], here we have set $\ell=cN^{1/3}$ with $N=575$ and $c=2$.

Table 3 gives the estimated standard errors of the least squares estimators of the 21 parameters $\theta_1,\ldots,\theta_{21}$.

From the table, it is evident that the estimates generated by Gap Bootstrap I are consistently smaller than those produced by Gap Bootstrap II. To verify the presence of serial correlation within columns, we also computed the component-wise sample autocorrelation functions (ACFs) for each of origin and destination time series (not shown here). From these, we found that there is nontrivial correlation in all other series up to lag 14 and that the ACFs are of different shapes. In view of the nonstationarity of the components and the presence of nontrivial serial correlation, it seems reasonable to infer that Gap Bootstrap I underestimates the standard error of the split proportion estimates in the synthetic OD model and, hence, Gap Bootstrap II estimates may be used for further analysis and decision making.

7. Concluding remarks. In this paper we have presented two resampling methods that are suitable for carrying out inference on a class of massive data sets that have a special structural property. While naive applications of the existing resampling methodology are severely constrained by the computational issues associated with massive data sets, the proposed methods exploit the so-called "gap" structure of massive data sets to split them into well-behaved smaller subsets where judicious combinations of known resampling techniques can be employed to obtain subset-wise accurate solutions. Some simple analytical considerations are then used to combine the piecewise results to solve the original problem that is otherwise intractable. As is evident from the discussions earlier, the versions of the proposed Gap Bootstrap methods require different sets of regularity conditions for their validity. Method I requires that the different subsets (in our notation, rows) have approximately the same distribution and that the number of such subsets be large. In comparison, Method II allows for nonstationarity among the different subsets and does not require the number of subsets itself to go to infinity. However, the price paid for a wider range of validity for Method II is that it requires some analytical considerations [cf. (3.4)] and that it uses more complex resampling methodology. We show that the analytical considerations are often simple, specifically for asymptotically linear estimators, which cover a number of commonly used classes of estimators. Even in the nonstationary setup, such as in the regression models associated with the real data example, finding the weights in (3.4) is not very difficult. In the moderate scale simulation of Section 5, Method II typically outperformed all the resampling methods considered here, including, perhaps surprisingly, the block bootstrap on the entire data set; This can be explained by noting that unlike the block bootstrap method, Method II crucially exploits the gap structure to estimate different parts by using a suitable resampling method for each part separately. On the other hand, Method I gives a "quick and simple" alternative for massive data sets that has a reasonably good performance whenever the data subsets are relatively homogeneous and the number of subsets is large.

APPENDIX: PROOFS

For clarity of exposition, we first give a relatively detailed proof of Theorem 4.2 in Section A.1 and then outline a proof of Theorem 4.1 in Section A.2.

A.1. Proof of consistency of Method II.

A.1.1. *Conditions.* Let $\{\mathbf{Y}_t\}_{t\in\mathbb{Z}}$ be a d-dimensional time series on a probability space $(\Omega, \mathcal{F}, P)$ with strong mixing coefficient

$$\alpha(n) \equiv \sup\{|P(A\cap B) - P(A)P(B)| : A \in \mathcal{F}^a_\infty, B \in \mathcal{F}^\infty_{a+n}, a \in \mathbb{Z}\}, \qquad n \ge 1,$$

where $\mathbb{Z} = \{0, \pm 1, \pm 2, \dots\}$ and where $\mathcal{F}^b_a = \sigma\langle \mathbf{Y}_t : t \in [a,b]\cap\mathbb{Z}\rangle$ for $-\infty \le a \le b \le \infty$. We suppose that the observations $\{\mathbf{X}_t : t = 1,\dots,n\}$ are obtained from the $\mathbf{Y}_t$-series with systematic deletion of $\mathbf{Y}_t$-subseries of length q, as described in Section 2.2, leaving a gap of q in between two columns of $\mathbb{X}$, that is, $(\mathbf{X}_1,\dots,\mathbf{X}_p) = (\mathbf{Y}_1,\dots,\mathbf{Y}_p)$, $(\mathbf{X}_{p+1},\dots,\mathbf{X}_{2p} = (\mathbf{Y}_{p+q+1},\dots,\mathbf{Y}_{2p+q})$, etc. Thus, for $i = 0,\dots,m-1$ and $j = 1,\dots,p$,

$$\mathbf{X}_{ip+j} = \mathbf{Y}_{i(p+q)+j}.$$

Further, suppose that the vectorized process $\{(\mathbf{X}_{ip+1},\dots,\mathbf{X}_{(i+1)p}) : i \ge 0\}$ is stationary. Thus, the original process $\{\mathbf{Y}_t\}$ is nonstationary, but it has a periodic structure over a suitable subset of the index set, as is the case in the transportation data example. Note that these assumptions are somewhat weaker than the requirements in (2.2). Also, for each $j = 1, \dots, p$, denote the i.i.d. bootstrap observations generated by Efron (1979)'s bootstrap by $\{\mathbf{X}^*_{ip+j} : i = 0, \dots, m-1\}$ and the bootstrap version of $\hat{\theta}_{jn}$ by θ^*_{jn}. Write E_* and Var_* to denote the conditional expectation and variance of the bootstrap variables.

To prove the consistency of the Gap bootstrap II variance estimator, we will make use of the following conditions:

(C.1) There exist $C \in (0, \infty)$ and $\delta \in (0, \infty)$ such that for $j = 1, \dots, p$,

$$E\psi_j(\mathbf{X}_j) = 0, \qquad E|\psi_j(\mathbf{X}_j)|^{2+\delta} < C$$

and $\sum_{n=1}^{\infty} \alpha(n)^{\delta/(2+\delta)} < \infty$.

(C.2) $[\hat{\theta}_n - \sum_{j=1}^{p} w_{jn}\hat{\theta}_{jn}] = o(n^{-1/2})$ in $L^2(P)$.

(C.3) (i) For $j = 1, \dots, p$,

$$\hat{\theta}_{jn} = m^{-1} \sum_{i=0}^{m-1} \psi_j(\mathbf{X}_{ip+j}) + o(m^{-1/2}) \qquad \text{in } L^2(P).$$

(ii) For $j = 1, \dots, p$,

$$\theta^*_{jn} = m^{-1} \sum_{i=0}^{m-1} \psi_j(\mathbf{X}^*_{ip+j}) + R^*_{jn} \quad \text{and} \quad E[E_*\{R^*_{jn}\}^2] = o(m^{-1/2}),$$

$$\tilde{\theta}^{(i)}_{jn} = \sum_{a=i}^{i+\ell-1} \psi_j(\mathbf{X}_{(a-1)p+j}) + o(\ell^{-1/2}) \qquad \text{in } L^2(P), i = 1, \dots, I.$$

(C.4) $q \to \infty$ and $p \sum_{j=1}^{p} w^2_{jn} = O(1)$ as $n \to \infty$.

We now briefly comment on the conditions. Condition (C.1) is a standard moment and mixing condition used in the literature for convergence of the series $\sum_{k=1}^{\infty} \mathrm{Cov}(\psi_j(\mathbf{X}_j), \psi_j(\mathbf{X}_{kp+j}))$ [cf. Ibragimov and Linnik (1971)]. Condition (C.2) is a stronger form of (3.4). It guarantees asymptotic equivalence of the variances of $\hat{\theta}_n$ and its subsample (row)-based approximation $\sum_{j=1}^{p} w_{jn}\hat{\theta}_{jn}$. Condition (C.3) in turn allows us to obtain an explicit expression for the asymptotic variance of $\hat{\theta}_{jn}$ and, hence, of $\hat{\theta}_n$. Note that the linear representation of $\hat{\theta}_{jn}$ in (C.3) holds for many common estimators, including M, L and R estimators, where the $L^2(P)$ convergence is replaced by convergence in probability. The $L(P)$ convergence holds for M-estimators under suitable monotonicity conditions on the score function; for L and R-estimators, it also holds under suitable moment condition on $\mathbf{X}_j$'s and under suitable growth conditions on the weight functions. Condition (C.3)(ii) requires that a linear representation similar to that of the row-wise estimator $\hat{\theta}_{jn}$ holds for its i.i.d. bootstrap version θ^*_{jn}. If the bootstrap variables $\mathbf{X}^*_{ip+j}$ are defined on $(\Omega, \mathcal{F}, P)$ (which can always be done on a possibly enlarged probability space), then the iterated expectation $E[E_*\{R^*_{jn}\}^2]$ is the same as the unconditional expectation $E\{R^*_{jn}\}^2$, and the first part of (C.2)(ii) can be simply stated as

$$\theta^*_{jn} = m^{-1} \sum_{i=0}^{m-1} \psi_j(\mathbf{X}^*_{ip+j}) + o(m^{-1/2}) \qquad \text{in } L^2(P).$$

The second part of (C.2)(ii) is an analog of (C.2)(i) for the subsample versions of the estimators $\hat{\theta}_{jn}$'s. The remainder term here is $o(\ell^{-1/2})$, as the subsampling estimators are now based on ℓ columns of $\mathbf{X}_t$-variables as opposed to m columns for $\hat{\theta}_{jn}$'s. All the representations in condition (C.3) hold for suitable classes of M, L and R estimators, as described above.

Next consider condition (C.4). It requires that the gap between the $\mathbf{Y}_t$ variables in two consecutive columns of $\mathbb{X}$ go to infinity, *at an arbitrary rate*. This condition guarantees that the i.i.d. bootstrap of Efron (1979) yields consistent variance estimators for the row-wise estimators $\hat{\theta}_{jn}$'s, even in presence of (weak) serial correlation. The second part of condition (C.4) is equivalent to requiring $w_{jn} = O(1)$ for each $j = 1, \dots, p$, when p is fixed. For simplicity, in the following we only prove Theorem 4.2 for the case p is fixed. However, in some applications, "$p \to \infty$" may be a more realistic assumption and, in this case, Theorem 4.2 remains valid provided the order symbols in (C.3) have the rate $o(m^{-1/2})$ uniformly over $j \in \{1, \dots, p\}$, in addition to the other conditions.

A.1.2. *Proofs.* Let $\theta^\dagger_n = \sum_{j=1}^p w_{jn} \hat{\theta}_{jn}$ and $\theta^\dagger_{jn} = m^{-1} \sum_{i=0}^{m-1} \psi_j(\mathbf{X}_{ip+j})$, $j = 1, \dots, p$. Let K denote a generic constant in $(0, \infty)$ that does not depend on n. Also, unless otherwise specified, limits in order symbols are taken by letting $n \to \infty$.

PROOF OF THEOREM 4.2. First we show that

$$n\left|\operatorname{Var}(\hat{\theta}_n) - \sum_{j=1}^p \sum_{k=1}^p w_{jn} w_{kn} \operatorname{Cov}(\hat{\theta}_{jn}, \hat{\theta}_{kn})\right| = o(1). \tag{A.1}$$

Let $\Delta_n = \hat{\theta}_n - \theta^\dagger_n$. Note that by condition (C.2), $E\Delta_n^2 = o(1)$. Hence, by the Cauchy–Schwarz inequality, the left side of (A.1) equals

$$\begin{aligned} &n|E(\hat{\theta}_n - E\hat{\theta}_n)^2 - E(\theta^\dagger_n - E\theta^\dagger_n)^2| \\ &\quad \le 2n|E(\theta^\dagger_n - E\theta^\dagger_n)(\Delta_n - E\Delta_n)| + n \operatorname{Var}(\Delta_n) \end{aligned}$$

$$\begin{aligned}&\le 2n[\mathrm{Var}(\theta_n^\dagger)]^{1/2}(E\Delta_n^2)^{1/2}+E\Delta_n^2\\&=o(1),\end{aligned}$$

provided $\mathrm{Var}(\theta_n^\dagger)=O(1)$.

To see that $\mathrm{Var}(\theta_n^\dagger)=O(1)$, note that

$$\begin{aligned}m\,\mathrm{Var}(\theta_{jn}^\dagger)&=m^{-1}\,\mathrm{Var}\left(\sum_{i=0}^{m-1}\psi_j(\mathbf{X}_{ip+j})\right)\\&=E\psi_j(X_j)^2+2m^{-1}\sum_{k=1}^{m-1}(m-k)E\psi_j(\mathbf{X}_j)\psi_j(\mathbf{X}_{kp+j})\qquad\text{(A.2)}\\&=E\psi_j(X_j)^2+o(1)\end{aligned}$$

as, by conditions (C.1) and (C.4),

$$\begin{aligned}&2m^{-1}\sum_{k=1}^{m-1}(m-k)|E\psi_j(\mathbf{Y}_j)\psi_j(\mathbf{Y}_{k(p+q)+j})|\\&\quad\le K\sum_{k=1}^{m-1}\alpha(k[p+q])^{\delta/(2+\delta)}(E|\psi_j(\mathbf{X}_j)|^{2+\delta})^{2/(2+\delta)}\\&\quad\le C^{2/(2+\delta)}K\sum_{k=p+q}^{\infty}\alpha(k)^{\delta/(2+\delta)}=o(1).\end{aligned}$$

By similar arguments, for any $1\le j,k\le p$,

$$m\,\mathrm{Cov}(\theta_{jn}^\dagger,\theta_{kn}^\dagger)=E\psi_j(\mathbf{X}_j)\psi_k(\mathbf{X}_k)+o(1).\qquad\text{(A.3)}$$

Also, by (A.2) and conditions (C.3) and (C.4),

$$\begin{aligned}n\,\mathrm{Var}(\theta_n^\dagger)&=n\sum_{j=1}^p w_{jn}^2\,\mathrm{Var}(\hat\theta_{jn})+2n\sum_{1\le j<k\le p}|w_{jn}w_{kn}||\mathrm{Cov}(\hat\theta_{jn},\hat\theta_{kn})|\\&=O\left(\left[\sum_{j=1}^p|w_{jn}|\right]^2nm^{-1}\right)=O(1).\end{aligned}$$

Hence, (A.1) follows.

To complete the proof of the theorem, by (A.1), it now remains to show that

$$m[\hat\sigma_{jn}^2-\mathrm{Var}(\hat\theta_{jn})]=o_p(1),\qquad\text{(A.4)}$$

$$\hat\rho_n(j,k)-\rho_n(j,k)=o_p(1)\qquad\text{(A.5)}$$

for all $1\le j,k\le p$, where $\rho_n(j,k)$ is the correlation between $\hat\theta_{jn}$ and $\hat\theta_{kn}$. First consider (A.4). Note that by (A.2), $m\,\mathrm{Var}(\hat\theta_{jn})=E\psi_j(X_j)^2+o(1)$ and by condition (C.3)(ii),

$$m\hat{\sigma}_{jn}^2 = m\,\mathrm{Var}_*\left(m^{-1}\sum_{i=0}^{m-1}\psi_j(\mathbf{X}_{ip+j}^*)\right) + o_p(1).$$

By using a truncation argument and the mixing condition (C.4), it is easy to show that

$$m^{-1}\sum_{i=0}^{m-1}[\psi_j(\mathbf{X}_{ip+j})]^r = E[\psi_j(\mathbf{X}_{ip+j})]^r + o_p(1), \qquad r = 1, 2.$$

Hence, (A.4) follows. Next, to prove (A.5), note that by condition (C.3), (A.2) and (A.3),

$$\rho_n(j,k) = \frac{E\psi_j(\mathbf{X}_j)\psi_k(\mathbf{X}_k)}{[E\psi_j(\mathbf{X}_j)^2]^{1/2}[E\psi_k(\mathbf{X}_k)^2]^{1/2}} + o(1)$$

for all j,k. Also, by conditions (C.3)–(C.4) and standard variance bound under the moment and mixing conditions of (C.4), for all j,k,

$$I^{-1}\sum_{i=1}^{I}\tilde{\theta}_{jn}^{(i)}\tilde{\theta}_{kn}^{(i)} = I^{-1}\sum_{i=1}^{I}\theta_{jn}^{\dagger(i)}\theta_{kn}^{\dagger(i)} + o_p(\ell^{-1/2}),$$

where $\theta_{jn}^{\dagger(i)} = \sum_{a=i}^{i+\ell-1}\psi_j(\mathbf{X}_{(a-1)p+j})$, $i = 1,\dots,I$. The consistency of the sampling window estimator of $\rho_n(j,k)$ can now be proved by using conditions (C.2), (C.3) and standard results [cf. Theorem 3.1, Lahiri (2003)]. This completes the proof of (A.5) and hence of Theorem 4.2.

PROOFS OF (3.9) AND (3.10). For notational simplicity, w.l.g., we set $\mu = 0$. (Otherwise, replace X_t by $X_t - \mu$ for all t in the following steps.) Write $\bar{X}_{jn}$ and $\bar{X}^{(k)}$, respectively, for the sample averages of the jth row and kth column, $1 \le j \le p$ and $1 \le k \le m$. First consider (3.9). Since $\mu = 0$, it follows that for each $j \in \{1,\dots,p\}$,

$$\hat{\theta}_{jn} = m^{-1}\sum_{i=1}^{m}X_{(i-1)p+j}^2 - \bar{X}_{jn}^2 = m^{-1}\sum_{i=1}^{m}X_{(i-1)p+j}^2 + O_p(n^{-1}).$$

Since $n = mp$, using a similar argument, it follows that $\hat{\theta}_n = n^{-1}\sum_{i=1}^{n}X_i^2 + O_p(n^{-1}) = p^{-1}\sum_{j=1}^{p}\hat{\theta}_{jn} + O_p(n^{-1})$. Hence, (3.9) holds.

Next consider (3.10). It is easy to check that for all $k = 1,\dots,m$, $\hat{\theta}^{(k)} = p^{-1}\sum_{i=1}^{p}X_{(k-1)p+i}^2 - [\bar{X}^{(k)}]^2$ and $E[\bar{X}^{(k)}]^2 = p^{-2}\mathbf{1}'\Sigma\mathbf{1}$. Hence, with $W_k = [\bar{X}^{(k)}]^2 - E[\bar{X}^{(k)}]^2$,

$$\hat{\theta}_n = n^{-1}\sum_{i=1}^{n}X_i^2 + O_p(n^{-1})$$

$$
\begin{aligned}
&= m^{-1}\sum_{k=1}^{m}[\hat{\theta}^{(k)} + \{\bar{X}^{(k)}\}^2] + O_p(n^{-1}) \\
&= m^{-1}\sum_{k=1}^{m}\hat{\theta}^{(k)} + p^{-2}\mathbf{1}'\Sigma\mathbf{1} + m^{-1}\sum_{k=1}^{m} W_k + O_p(n^{-1}) \\
&= m^{-1}\sum_{k=1}^{m}\hat{\theta}^{(k)} + p^{-2}\mathbf{1}'\Sigma\mathbf{1} + O_p(n^{-1/2}),
\end{aligned}
$$

provided condition (C.1) holds with $\psi_j(x) = x^2$ for all j. Further, note that the leading part of the $O_p(n^{-1/2})$-term is $n^{-1/2} \times m^{-1/2}\sum_{k=1}^{m} W_k$ and $m^{-1/2} \times \sum_{k=1}^{m} W_k$ is asymptotically normal with mean zero and variance $\sigma_W^2 \equiv \mathrm{Var}(W_1) + 2\sum_{i=1}^{\infty}\mathrm{Cov}(W_1, W_{i+1})$. As a result, the $O_p(n^{-1/2})$-term cannot be of a smaller order (except in the special case of $\sigma_W^2 = 0$). $\square$

A.2. Proof of consistency of Method I.

A.2.1. *Conditions.* We shall continue to use the notation and conventions of Section A.1.2. In addition to assuming that $\mathbb{X}$ satisfies (2.2), we shall make use of the following conditions:

(A.1) (i) Pairwise distributions of $\{m^{1/2}(\hat{\theta}_{jn} - \theta) : 1 \le j \le p\}$ are identical.
(ii) $\{m^{1/2}(\hat{\theta}_{jn} - \theta) : 1 \le j \le p\}$ are m_0-dependent with $m_0 = o(p)$.

(A.2) (i) $m\,\mathrm{Var}(\hat{\theta}_{1n}) \to \Sigma$ and $m\,\mathrm{Cov}(\hat{\theta}_{1n}, \hat{\theta}_{2n}) \to \Lambda$ as $n \to \infty$.
(ii) $\{[m^{1/2}(\hat{\theta}_{1n} - \theta)]^2 : n \ge 1\}$ is uniformly integrable.
(iii) $mp^{-1}\sum_{j=1}^{p}\widehat{\mathrm{Var}}(\hat{\theta}_{jn}) \to_p \Sigma$ as $n \to \infty$.

Now we briefly comment on the conditions. As indicated earlier, for the validity of the Gap Bootstrap I method, we do not need the exchangeability of the rows of $\mathbb{X}$; the amount of homogeneity of the centered and scaled row-wise estimators $\{m^{1/2}(\hat{\theta}_{jn} - \theta) : 1 \le j \le p\}$, as specified by condition (A.1)(i), is all that is needed. (A.1)(i) also provides the motivation behind the definition of the variance estimator of the pair-wise differences right above (3.3). Condition (A.1)(ii) has two implications. First, it quantifies the approximate independence condition in (2.2). A suitable strong mixing condition can be used instead, as in the proof of Theorem 4.2, but we do not attempt such generalizations to keep the proof short. A second implication of (A.1)(ii) is that $p \to \infty$ as $n \to \infty$, that is, the number of subsample estimators $\hat{\theta}_{jn}$'s must be large. In comparison, m_0 may or may not go to infinity with $n \to \infty$. Next consider condition (A.2). Condition (A.2)(i) says that the row-wise estimators are root-m consistent and that for any pair $j \ne k$, the covariance between $m^{1/2}(\hat{\theta}_{jn} - \theta)$ and $m^{1/2}(\hat{\theta}_{kn} - \theta)$ has a common limit, which is what we are indirectly trying to estimate using $m\widetilde{\mathrm{Var}}(\hat{\theta}_{jn} - \hat{\theta}_{kn})$. Condition (A.2)(ii) is a uniform integrability condition that is implied by $E|m^{1/2}(\hat{\theta}_{1n} - \hat{\theta}_{2n})|^{2+\delta} = O(1)$ [cf. condition (C.1)] for some $\delta > 0$. Part (iii) of condition (A.2) says that the i.i.d. bootstrap variance estimator applied to the (average of the) row-wise estimators be consistent. A proof of this can be easily constructed using the arguments given in the proof of Theorem 4.2, by requiring some standard regularity conditions on the score functions that define the $\hat{\theta}_{jn}$'s in Section 3.2. We decided to state it as a high level condition to avoid repetition of similar arguments and to save space.

A.2.2. *Proof of Theorem 4.1.* In view of condition (A.2)(iii) and (3.3), it is enough to show that

$$m[\widetilde{\mathrm{Var}}(\hat{\theta}_{1n} - \hat{\theta}_{2n}) - E(\hat{\theta}_{1n} - \hat{\theta}_{2n})(\hat{\theta}_{1n} - \hat{\theta}_{2n})'] \to_p 0.$$

Since this is equivalent to showing component-wise consistency, without loss of generality, we may suppose that the $\hat{\theta}_{jn}$'s are one-dimensional. Define $V_{jk} = m(\hat{\theta}_{jn} - \hat{\theta}_{kn})^2 \mathbb{1}(|m^{1/2}(\hat{\theta}_{jn} - \hat{\theta}_{kn})| > a_n)$, $W_{jk} = m(\hat{\theta}_{jn} - \hat{\theta}_{kn})^2 \mathbb{1}(|m^{1/2}(\hat{\theta}_{jn} - \hat{\theta}_{kn})| \le a_n)$, for some $a_n \in (0, \infty)$ to be specified later. It is now enough to show that

$$Q_{1n} \equiv p^{-2} \sum_{1 \le j \ne k \le p} |V_{jk} - EV_{jk}| \to_p 0,$$

$$Q_{2n} \equiv p^{-2} \left| \sum_{1 \le j \ne k \le p} [W_{jk} - EW_{jk}] \right| \to_p 0.$$

By condition (A.2)(ii), $\{[m^{1/2}(\hat{\theta}_{1n} - \hat{\theta}_{2n})]^2 : n \ge 1\}$ is also uniformly integrable and, hence,

$$EQ_{1n} \le 2E|m^{1/2}(\hat{\theta}_{1n} - \hat{\theta}_{2n})^2 \mathbb{1}(|m^{1/2}(\hat{\theta}_{1n} - \hat{\theta}_{2n})| > a_n) = o(1)$$

whenever $a_n \to \infty$ as $n \to \infty$. Next consider Q_{2n}. Define the sets $J_1 = \{(j,k) : 1 \le j \ne k \le p\}$, $A_{j,k} = \{(j_1, k_1) \in J_1 : \min\{|j - j_1|, |k - k_1|\} \le m_0\}$ and $B_{j,k} = J_1 \setminus A_{j,k}$, $(j,k) \in J_1$. Then, for any $(j,k) \in J_1$, by the m_0-dependence condition,

$$\mathrm{Cov}(W_{jk}, W_{a,b}) = 0 \qquad \text{for all } (a,b) \in B_{j,k}.$$

Further, note that $|A_{j,k}| \equiv$ the size of $A_{j,k}$ is at most $2m_0 p$ for all $(j,k) \in J_1$. Hence, it follows that

$$\begin{aligned} EQ_{2n}^2 &\le p^{-4}\left[\sum_{(j,k)\in J_1} \mathrm{Var}(W_{jk}) + \sum_{(j,k)\in J_1} \sum_{(a,b)\ne(j,k)} \mathrm{Cov}(W_{jk}, W_{ab})\right] \\ &\le p^{-4}\left[p^2 EW_{12}^2 + \sum_{(j,k)\in J_1} \sum_{(a,b)\in A_{j,k}} |\mathrm{Cov}(W_{jk}, W_{ab})|\right] \end{aligned}$$

$$\begin{aligned}&\le p^{-4}[p^2 a_n^2 E|W_{12}| + p^2 \cdot 2m_0 p \cdot a_n^2 E|W_{12}|]\\&= O(p^{-1} m_0 a_n^2)\end{aligned}$$

as $E|W_{12}| \le mE(\hat{\theta}_{1n} - \hat{\theta}_{2n})^2 = O(1)$. Now choosing $a_n = [p/m_0]^{1/3}$ (say), we get $Q_{kn} \to_p 0$ for $k = 1, 2$, proving (A.6). This completes the proof of Theorem 4.1. $\square$

Acknowledgments. The authors thank the referees, the Associate Editor and the Editor for a number of constructive suggestions that significantly improved an earlier draft of the paper.

REFERENCES

Bell, M. G. H. (1991). The estimation of origin-destination matrices by constrained generalised least squares. *Transportation Res.* **25B** 13–22. MR1093618

Cascetta, E. (1984). Estimation of trip matrices from traffic counts and survey data: A generalized least squares estimator. *Transportation Res.* **18B** 289–299.

Cremer, M. and Keller, H. (1987). A new class of dynamic methods for identification of origin-destination flows. *Transportation Res.* **21B** 117–132.

Dixon, M. P. and Rilett, L. R. (2000). Real-time origin-destination estimation using automatic vehicle identification data. In *Proceedings of the 79th Annual Meeting of the Transportation Research Board CD-ROM*. Washington, DC.

Efron, B. (1979). Bootstrap methods: Another look at the jackknife. *Ann. Statist.* **7** 1–26. MR0515681

Gajewski, B. J., Rilett, L. R., Dixon, P. M. and Spiegelman, C. H. (2002). Robust estimation of origin-destination matrices. *Journal of Transportation and Statistics* **5** 37–56.

Hall, P. (1992). *The Bootstrap and Edgeworth Expansion*. Springer, New York.

Hall, P. and Jing, B. (1996). On sample reuse methods for dependent data. *J. Roy. Statist. Soc. Ser. B* **58** 727–737. MR1410187

Ibragimov, I. A. and Linnik, Y. V. (1971). *Independent and Stationary Sequences of Random Variables*. Wolters-Noordhoff Publishing, Groningen. MR0322926

Koul, H. L. (2002). *Weighted Empirical Processes in Dynamic Nonlinear Models. Lecture Notes in Statistics* **166**. Springer, New York. MR1911855

Koul, H. L. and Mukherjee, K. (1993). Asymptotics of R-, MD- and LAD-estimators in linear regression models with long range dependent errors. *Probab. Theory Related Fields* **95** 535–553. MR1217450

Künsch, H. R. (1989). The jackknife and the bootstrap for general stationary observations. *Ann. Statist.* **17** 1217–1241. MR1015147

Lahiri, S. N. (1999). Theoretical comparisons of block bootstrap methods. *Ann. Statist.* **27** 386–404. MR1701117

Lahiri, S. N. (2003). *Resampling Methods for Dependent Data*. Springer, New York. MR2001447

Mannering, F. L., Washburn, S. S. and Kilareski, W. P. (2009). *Principles of Highway Engineering and Traffic Analysis*, 4th ed. Wiley, Hoboken, NJ.

Okutani, I. (1987). The Kalman filtering approaches in some transportation and traffic problems. In *Transportation and Traffic Theory (Cambridge, MA, 1987)* 397–416. Elsevier, New York. MR1095119

Patton, A., Politis, D. N. and White, H. (2009). Correction to "Automatic block-length selection for the dependent bootstrap" by D. Politis and H. White. *Econometric Rev.* **28** 372–375. MR2532278

Politis, D. N. and Romano, J. P. (1994). Large sample confidence regions based on subsamples under minimal assumptions. *Ann. Statist.* **22** 2031–2050. MR1329181

Roess, R. P., Prassas, E. S. and McShane, W. R. (2004). *Traffic Engineering*, 3rd ed. Prentice Hall, Englewood Cliffs, NJ.

Serfling, R. J. (1980). *Approximation Theorems of Mathematical Statistics*. Wiley, New York. MR0595165

Singh, K. (1981). On the asymptotic accuracy of Efron's bootstrap. *Ann. Statist.* **9** 1187–1195. MR0630102

S. N. Lahiri
C. Spiegelman
Department of Statistics
Texas A&M University
College Station, Texas 77843
USA
E-mail: snlahiri@stat.tamu.edu

J. Appiah
L. Rilett
Nebraska Transportation Center
University of Nebraska-Lincoln
262D Whittier Research Center
P.O. Box 880531
Lincoln, Nebraska 68588-0531
USA