K-sample subsampling in general spaces: The case of independent time series

Dimitris N. Politis [a,*], Joseph P. Romano [b]

[a] *Department of Mathematics, University of California at San Diego, La Jolla, CA 92093-0112, United States*
[b] *Departments of Statistics and Economics, Stanford University, Stanford, CA 94305-4065, United States*

ABSTRACT

The problem of subsampling in two-sample and K-sample settings is addressed where both the data and the statistics of interest take values in general spaces. We focus on the case where each sample is a stationary time series, and construct subsampling confidence intervals and hypothesis tests with asymptotic validity. Some examples are also given, and the problem of optimal block size choice is discussed.

© 2008 Elsevier Inc. All rights reserved.

1. Introduction

Subsampling is a statistical method that is most generally valid for nonparametric inference such as the construction of confidence intervals and hypothesis tests in a large-sample setting. The applications of subsampling are numerous starting from i.i.d. data and regression, and continuing to time series, random fields, marked point processes, etc.; see [1] for a review and extensive list of references.

Interestingly, the two-sample and K-sample i.i.d. set-ups have not been explored yet in the subsampling literature; we attempt to fill this gap here. So consider K independent datasets: $\underline{X}^{(1)}, \ldots, \underline{X}^{(K)}$ where $\underline{X}^{(k)} = (X_1^{(k)}, \ldots, X_{n_k}^{(k)})$ for $k = 1, \ldots, K$. The random variables $X_j^{(k)}$ take values in an arbitrary space $\mathbf{S}$; typically, $\mathbf{S}$ would be $\mathbf{R}^d$ for some d, but $\mathbf{S}$ can very well be a function space.

Although dataset $\underline{X}^{(k)}$ is independent of $\underline{X}^{(k')}$ for $k \neq k'$ there may well be dependence *within* a dataset. Thus, we distinguish two cases:

- *i.i.d. samples.* For each $k = 1, \ldots, K$, the data $X_1^{(k)}, \ldots, X_{n_k}^{(k)}$ are i.i.d.
- *Time series samples.* For each $k = 1, \ldots, K$, the data $X_1^{(k)}, \ldots, X_{n_k}^{(k)}$ represent a stretch from a time series $\{X_t^{(k)}, \ t \in \mathbf{Z}\}$ that is governed by probability law P_k.

---

* Corresponding author.
 E-mail addresses: dpolitis@ucsd.edu (D.N. Politis), romano@stanford.edu (J.P. Romano).

An example in the i.i.d. case above is the usual two-sample set-up in biostatistics where d 'features' (body characteristics, gene expressions, etc.) are measured on a group of patients, and then again measured on a control group. The i.i.d. case was concisely treated in the short announcement of [2].

Since the i.i.d. case is a special case of the time series case, the paper at hand focuses on the latter. An immediate formulation of the set-up of multiple time series is the framework of a multivariate time series—see e.g. [3], [4], or [5]); this multivariate set-up is covered by the general theory of subsampling as discussed in [6]. In particular, [7] were recently able to use subsampling in the context of a bivariate time series with the purpose of comparing the two coordinate time series with each other.

Nevertheless, literature on comparing time series of possibly different length, sampling frequency, and/or synchronicity seems scarce. As an example, consider the problem of comparing the average temperature of San Diego to that of San Francisco where the San Diego measurements are quarterly (say) spanning years 1997 to 2007, and the San Francisco measurements are monthly (say) spanning 2000 to 2005. Because of different lengths, sampling frequencies, and lack of synchronicity, this two-sample temperature dataset could not easily be treated as a multivariate time series.

Three concrete examples are given in the next section together with some key definitions and assumptions for our asymptotic results. The large-sample validity of subsampling-based confidence intervals is shown in Section 3 using both studentized and unstudentized roots. Section 4 shows how similar results can be obtained with subsamples that have only partial overlap that is associated with a reduction of the computational expense. Section 5 focuses on hypothesis tests based on K-sample subsampling, while finally Section 6 addresses the problem of optimal choice of the block sizes, and the need for dealing with estimated rates of convergence.

2. Definitions, examples, and problem set-up

Throughout this paper it is assumed that, for each $k = 1, \ldots, K$, the data $X_1^{(k)}, \ldots, X_{n_k}^{(k)}$ represents a stretch from a time series $\{X_t^{(k)}, \ t \in \mathbf{Z}\}$ which is governed by probability law P_k. Throughout this paper, each time series $\{X_t^{(k)}, \ t \in \mathbf{Z}\}$ will be assumed *strictly stationary* and *strong mixing* with mixing coefficients $\alpha^{(k)}(t) \to 0$ when $t \to \infty$; both the stationarity and the mixing assumption can be somewhat relaxed—see [1], Ch. 4 and Ch. 12 and also [20].

The probability law associated with such a K-sample experiment is $P = (P_1, P_2, \ldots, P_K)$. The goal is inference (confidence regions, hypothesis tests, etc.) regarding some parameter $\theta = \theta(P)$ that takes values in a general normed linear space $\mathbf{B}$ with norm denoted by $\| \cdot \|$. Denote $\mathbf{n} = (n_1, \ldots, n_K)$, and let $\hat{\theta}_\mathbf{n} = \hat{\theta}_\mathbf{n}(\underline{X}^{(1)}, \ldots, \underline{X}^{(K)})$ be a consistent estimator of θ as $\min_k n_k \to \infty$.

2.1. Some motivating examples

We now give three illustrations; in all three examples, $K = 2$ and $\mathbf{S} = \mathbf{R}$, i.e., two real-valued time series samples.

Example 2.1 (*Comparing Population Means*). For simplicity, let $\mathbf{B} = \mathbf{R}$, and denote by μ_k, $\gamma_k(s)$ and $f_k(w)$, the mean, lag-s autocovariance, and spectral density of time series $\{X_t^{(k)}\}$ respectively for $k = 1, 2$. All such parameters are assumed to exist.

The natural statistic for testing $H_0 : \mu_1 = \mu_2$ is the difference of sample means, i.e., $\hat{\theta}_\mathbf{n} = \bar{X}^{(2)} - \bar{X}^{(1)}$ where $\bar{X}^{(2)} = n_2^{-1} \sum_{i=1}^{n_2} X_i^{(2)}$, and $\bar{X}^{(1)} = n_1^{-1} \sum_{i=1}^{n_1} X_i^{(1)}$. Note that $\hat{\theta}_\mathbf{n}$ satisfies $E\hat{\theta}_\mathbf{n} = \mu_2 - \mu_1$, and

$$\mathrm{Var}(\hat{\theta}_\mathbf{n}) = \frac{\sum_{s=-n_2}^{n_2} (1 - |s|n_2^{-1})\gamma_2(s)}{n_2} + \frac{\sum_{s=-n_1}^{n_1} (1 - |s|n_1^{-1})\gamma_1(s)}{n_1},$$

and hence

$$\mathrm{Var}(\hat{\theta}_\mathbf{n}) \sim \frac{2\pi f_2(0)}{n_2} + \frac{2\pi f_1(0)}{n_1} \quad \text{as } \min(n_1, n_2) \to \infty.$$

The statistic $\hat{\theta}_\mathbf{n}$ is asymptotically normal under standard conditions; see e.g. Brockwell and Davis (1991). Thus, an asymptotically valid 95% confidence interval for $\mu_2 - \mu_1$ is simply $\hat{\theta}_\mathbf{n} \pm 1.96\sqrt{\mathrm{Var}(\hat{\theta}_\mathbf{n})}$. Since even the asymptotic expression for $\mathrm{Var}(\hat{\theta}_\mathbf{n})$ depends on the unknown parameters $f_2(0), f_1(0)$, it must be replaced by a consistent estimator in the construction of the confidence interval. Such an estimator is given by $\widehat{\mathrm{Var}}(\hat{\theta}_\mathbf{n}) = 2\pi\hat{f}_{2,n_2}(0)/n_2 + 2\pi\hat{f}_{1,n_1}(0)/n_1$ where $\hat{f}_{2,n_2}(0), \hat{f}_{1,n_1}(0)$ are consistent nonparametric estimates of $f_2(0), f_1(0)$ based on the $X_t^{(k)}$ data of size n_k, for $k = 1, 2$; see e.g. [3]. Alternatively, the large-sample distribution of $\hat{\theta}_\mathbf{n}$ can be directly approximated by subsampling, which would automatically capture the correct asymptotic variance without explicit estimation. $\square$

Example 2.2 (*Comparing Probability Distribution Functions*). Let $G^{(k)}(\cdot)$ denote the probability distribution function of $X_1^{(k)}$, i.e., the first marginal distribution of time series $\{X_t^{(k)}\}$. The goal is to compare $G^{(1)}(\cdot)$ to $G^{(2)}(\cdot)$. Let

$$\theta(P) = \theta(\cdot, P) = G^{(1)}(\cdot) - G^{(2)}(\cdot),$$

regarded as a random element of $D(-\infty, \infty)$ endowed with the sup norm $\| \cdot \|$. Let $\hat{G}_{n_k}^{(k)}(\cdot)$ denote the empirical distribution function of the kth sample. Then, an empirical estimate of $\theta(x, P)$ is given by

$$\hat{\theta}_{\mathbf{n}}(x) = \hat{G}_{n_1}^{(1)}(x) - \hat{G}_{n_2}^{(2)}(x).$$

Note that under regularity conditions, as a random process on $D(-\infty, \infty)$,

$$n_k[\hat{G}_{n_k}^{(k)}(\cdot) - G^{(k)}(\cdot)]$$

converges weakly to a mean zero, Gaussian process; see Deo [8] and Yoshihara [9] who provide sufficient strong mixing conditions in the univariate and multivariate cases, respectively. Let $\tau_{\mathbf{n}}^2 = \min(n_1, n_2)$ and assume the ratio n_1/n_2 stays bounded away from 0 and ∞. It follows that, under sufficient mixing conditions, $\tau_{\mathbf{n}}[\hat{\theta}_{\mathbf{n}}(\cdot) - \theta(\cdot)]$ converges in distribution to a mean zero Gaussian process as well.

The two-sample Kolmogorov–Smirnov test statistic is then given by

$$t_{\mathbf{n}} = \sup_{-\infty < x < \infty} |\hat{\theta}_{\mathbf{n}}(x)|.$$

Under the null hypothesis $H_0 : G^{(1)}(\cdot) = G^{(2)}(\cdot)$, the statistic $\tau_{\mathbf{n}} t_{\mathbf{n}}$ has a well-defined asymptotic distribution under an appropriate mixing condition. However, in contrast to the i.i.d. case that is described in detail by DasGupta [10], this distribution depends on particular characteristics of the two time series, namely their dependence structure. To appreciate why, note that although $E\hat{G}_{n_k}^{(k)}(x) = G^{(k)}(x)$, we have

$$\text{Var}[\hat{G}_{n_k}^{(k)}(x)] = \frac{1}{n_k} \sum_{j=-n_k}^{n_k} \left(1 - \frac{|j|}{n_k}\right) c^{(k)}(x, j)$$

where

$$c^{(k)}(x, j) = \text{Cov}(1\{X_1^{(k)} \le x\}, 1\{X_{1+j}^{(k)} \le x\})$$

and $1\{\cdot\}$ is the indicator function. Therefore,

$$n_k \text{Var}[\hat{G}_{n_k}^{(k)}(x)] \to \sum_{j=-\infty}^{\infty} c^{(k)}(x, j),$$

assuming the series on the right side is convergent. Note that, in the i.i.d. case, exact permutation tests could be constructed, but the construction breaks down if there is dependence present. Nevertheless, subsampling can be used to directly approximate the quantiles of the null limit distribution of $t_{\mathbf{n}}$ so that the test can be performed. □

Example 2.3 (*Comparing Spectral Distribution Functions*). Now let $F^{(k)}(\cdot)$ denote the spectral distribution function of time series $\{X_t^{(k)}\}$. The goal is to compare $F^{(1)}(\cdot)$ to $F^{(2)}(\cdot)$. Let

$$\theta(P) = \theta(\cdot, P) = F^{(1)}(\cdot) - F^{(2)}(\cdot),$$

regarded as a random element of $D[0, \pi]$ endowed with the sup norm $\| \cdot \|$. Let $\hat{F}_{n_k}^{(k)}(\cdot)$ denote the corresponding integrated periodogram estimate, so that a natural estimate of $\theta(\lambda, P)$ is given by

$$\hat{\theta}_{\mathbf{n}}(\lambda) = \hat{F}_{n_1}^{(1)}(\lambda) - \hat{F}_{n_2}^{(2)}(\lambda).$$

For testing $H_0 : F^{(1)}(\cdot) = F^{(2)}(\cdot)$, consider the test statistic given by

$$t_{\mathbf{n}} = \sup_{\lambda \in [0, \pi]} |\hat{\theta}_{\mathbf{n}}(\lambda)|.$$

Under H_0, $\tau_{\mathbf{n}} t_{\mathbf{n}}$ has a well-defined asymptotic distribution for some judicial choice of $\tau_{\mathbf{n}}$; see Ch. 7.5 of [1]. If we can assume that $n_1/n_2 \to \beta \in (0, \infty)$, then we could take $\tau_{\mathbf{n}} = \min(n_1, n_2)$ as the convergence rate. Otherwise, we can take $\tau_{\mathbf{n}} = (\sigma_1^2/n_1 + \sigma_2^2/n_2)^{-1/2}$ where σ_1^2, σ_2^2 are some positive parameters; since σ_1^2, σ_2^2 are unknown, Corollary 6.1 could then be invoked in order to use an estimated rate of convergence $\hat{\tau}_{\mathbf{n}}$.

Alternatively, in order to construct a confidence band for the difference in spectral distribution functions, we can consider the root

$$\tau_{\mathbf{n}} \sup_{\lambda \in [0, \pi]} \|[\hat{F}_{n_1}^{(1)}(\lambda) - \hat{F}_{n_2}^{(2)}(\lambda)] - [F_1^{(1)}(\lambda) - F_2^{(2)}(\lambda)]\|,$$

whose true c.d.f. can be denoted by $J_\mathbf{n}(x, P)$. Evidently, knowledge of $J_\mathbf{n}(x, P)$ would allow for construction of a confidence region for the function-valued parameter $\theta(P)$; this confidence region is tantamount to a (simultaneous) confidence band for the difference of spectral distribution functions.

The subsampling method will offer an asymptotically valid approach to either approximate $J_\mathbf{n}(x, P)$ (resulting into confidence bands), and/or approximate the threshold of the critical region for the test of H_0.   □

2.2. Main assumptions

In order to handle the above and other examples in a unifying way we introduce the following notation. Let $g : \mathbf{B} \to \mathbf{R}$ be a continuous function, and let $J_\mathbf{n}(P)$ denote the probability law of the "root" $g[\tau_\mathbf{n}(\hat{\theta}_\mathbf{n} - \theta(P))]$ under P, with corresponding cumulative distribution function

$$J_\mathbf{n}(x, P) = \mathrm{Prob}_P\{g[\tau_\mathbf{n}(\hat{\theta}_\mathbf{n} - \theta(P))] \le x\} \tag{1}$$

where $\tau_\mathbf{n}$ is a normalizing sequence; in particular, $\tau_\mathbf{n}$ is to be thought of as a fixed function of $\mathbf{n}$ such that $\tau_\mathbf{n} \to \infty$ when $\min_k n_k \to \infty$.

Typically, $g(\theta)$ will either be (a continuous and invertible function of) the norm $\|\theta\|$ or a 'projection', i.e., a linear mapping of $\mathbf{B}$ into $\mathbf{R}$ of particular form. Some motivating examples are given below.

1. *Case* $\mathbf{B} = \mathbf{R}$. Here g may be taken to be the identity function (i.e., projection), or $g(\theta) = |\theta|$ (i.e., the norm).
2. *Case* $\mathbf{B} = \mathbf{R}^p$. Here $\theta = (\theta_1, \ldots, \theta_p)$, and the 'projection' choice corresponds to $g(\theta) = \sum_{i=1}^p c_i \theta_i$ for some vector $c = (c_1, \ldots, c_p)$; in particular, if $c_j = 1$ and all the other coordinates of c are zero, then $g(\theta) = \theta_j$ just picks the jth coordinate of θ. A 'norm' choice is to let $g(\theta) = \|\theta\|$ where $\|\cdot\|$ is some norm on $\mathbf{R}^p$.
3. *Case* $\mathbf{B} = l_p$. Here $\theta = (\theta_1, \theta_2, \cdots)$ with $\|\theta\| = (\sum_{i=1}^\infty |\theta_i|^p)^{1/p}$. If $c = (c_1, c_2, \ldots)$ is a sequence in l_q (with $q^{-1} = 1 - p^{-1}$), then a projection on direction c can be defined as $g(\theta) = \sum_{i=1}^\infty c_i \theta_i$. As above, if $c_j = 1$ and all the other coordinates of c are zero, then $g(\theta) = \theta_j$.
4. *Case* $\mathbf{B} = L_p[a_0, a_1]$. Here $\theta = \{\theta_x \text{ for } x \in [a_0, a_1]\}$ is a real-valued function on $[a_0, a_1]$, and the norm is $\|\theta\| = (\int_{a_0}^{a_1} |\theta_x|^p dx)^{1/p}$. If $c = \{c_x\}$ is a function on $L_q[a_0, a_1]$ (with $q^{-1} = 1 - p^{-1}$), then a projection on direction c can be defined as $g(\theta) = \int_{a_0}^{a_1} c_x \theta_x dx$. Letting $g(\theta) = \theta_{x_0}$ for some particular value x_0 is also a projection.

As in the one-sample case, the basic assumption that is required for subsampling to work is existence of a *bona fide* large-sample distribution, i.e.,

Assumption 2.1. There exists a nondegenerate limiting law $J(P)$ such that $J_\mathbf{n}(P)$ converges weakly to $J(P)$ as $\min_k n_k \to \infty$.

The law $J(P)$ has associated distribution function $J(x, P)$ with its $1 - \alpha$ quantile denoted by $J^{-1}(1 - \alpha, P)$. In general, for any distribution function $F(x)$, we define the quantile-inverse as $F^{-1}(\alpha) = \inf\{x : F(x) \ge \alpha\}$. Similarly, for a distribution of the type $F(x, P)$, the quantile-inverse is defined by $F^{-1}(\alpha, P) = \inf\{x : F(x, P) \ge \alpha\}$.

The basic idea behind subsampling is to be able to recompute a statistic of interest ($\hat{\theta}_\mathbf{n}$ here) not on data with sample sizes of $\mathbf{n} = (n_1, \ldots, n_K)$ but on appropriate subsamples of the original data of sizes $\mathbf{b} = (b_1, \ldots, b_K)$ where each b_k is an integer between 1 and n_k, chosen so that $b_k/n_k \to 0$. These recomputed values will be used to build up the subsampling distribution of a test statistic or a root; explicit constructions will be given in subsequent sections where, in addition to Assumption 2.1, we will use the following mild assumption.

Assumption 2.2. As $\min_k n_k \to \infty$, $\tau_\mathbf{b}\|\hat{\theta}_\mathbf{n} - \theta(P)\| = o_P(1)$.

As a matter of fact, Assumptions 2.1 and 2.2 are implied by the following assumption, as long as $\tau_\mathbf{b}/\tau_\mathbf{n} \to 0$.

Assumption 2.3. As $\min_k n_k \to \infty$, the distribution of $\tau_\mathbf{n}(\hat{\theta}_\mathbf{n} - \theta(P))$ under P converges weakly to some distribution (on the Borel σ-field of the normed linear space $\mathbf{B}$).

Here, weak convergence is understood to be taken in the modern sense of Hoffmann–Jorgensen; see Section 1.3 of [11]. That Assumption 2.3 implies both Assumptions 2.1 and 2.2 follows by the Continuous Mapping Theorem; see Theorem 1.3.6 of [11].

All the above asymptotic limits are under the assumption that $\min_k n_k \to \infty$. However, for technical reasons it will be important that the individual sample sizes n_k are of the same order of magnitude, i.e., that for some positive constants C_*, C^* we have

$$C_* \le n_k/n_{k'} \le C^* \quad \text{for all } k, k'. \tag{2}$$

3. Confidence sets with full-overlap subsampling

3.1. Unstudentized roots

In time series subsampling and/or the block bootstrap, an important issue is the degree of overlap between the extracted blocks to be re/subsampled. The case of fully overlapping blocks is generally thought to be the most efficient; see e.g. [1] or [12], Ch. 5).

Focusing on the kth sample (for some $1 \leq k \leq K$), let

$$T_j^{(k)} = (X_{(j-1)+1}^{(k)}, X_{(j-1)+2}^{(k)}, \ldots, X_{(j-1)+b_k}^{(k)})$$

be the jth block-subsample of size b_k that can be extracted from the series $\{X_1^{(k)}, \ldots, X_{n_k}^{(k)}\}$. The block size b_k is an integer in $[1, n_k]$; note that the overlap between adjacent blocks is the maximum possible, i.e., $T_j^{(k)}$ and $T_{j+1}^{(k)}$ have $b_k - 1$ common elements.

Let $\mathcal{T}_k$ denote the set of all size b_k block-subsamples obtained from the kth sample, i.e., let $\mathcal{T}_k = \{T_j^{(k)}, j = 1, \ldots, q_k\}$ where $q_k = n_k - b_k + 1$. A K-fold subsample is then constructed by choosing one element from each super-set $\mathcal{T}_k$ for $k = 1, \ldots, K$. Thus, a typical K-fold subsample has the form: $T_{i_1}^{(1)}, T_{i_2}^{(2)}, \ldots, T_{i_K}^{(K)}$, where $1 \leq i_k \leq q_k$ for $k = 1, \ldots, K$. It is apparent that the number of possible K-fold subsamples is $q = \prod_{k=1}^{K} q_k$.

So a subsample value of the general statistic $\hat{\theta}_{\mathbf{n}}$ is

$$\hat{\theta}_{\mathbf{i},\mathbf{b}} = \hat{\theta}_{\mathbf{b}}(T_{i_1}^{(1)}, \ldots, T_{i_K}^{(K)}) \tag{3}$$

where $\mathbf{b} = (b_1, \ldots, b_K)$ and $\mathbf{i} = (i_1, \ldots, i_K)$.

The subsampling distribution of statistic $\hat{\theta}_{\mathbf{n}}$ is now defined as

$$L_{\mathbf{n},\mathbf{b}}(x) = \frac{1}{q} \sum_{i_1=1}^{q_1} \sum_{i_2=1}^{q_2} \cdots \sum_{i_K=1}^{q_K} 1\{g[\tau_{\mathbf{b}}(\hat{\theta}_{\mathbf{i},\mathbf{b}} - \hat{\theta}_{\mathbf{n}})] \leq x\}. \tag{4}$$

As in the single-sample case, $L_{\mathbf{n},\mathbf{b}}(x)$ provides a generally consistent approximation to $J_{\mathbf{n}}(x, P)$ of Assumption 2.1. Consequently, the quantiles of $L_{\mathbf{n},\mathbf{b}}(x)$ can be used in place of the unknown quantiles of $J_{\mathbf{n}}(x, P)$ for the construction of large-sample confidence regions for θ.

Theorem 3.1. *Assume Assumptions 2.1 and 2.2, where g is assumed uniformly continuous. Also assume (2), and that, for each $k = 1, \ldots, K$, we have $b_k/n_k \to 0$, $\tau_{\mathbf{b}}/\tau_{\mathbf{n}} \to 0$, and $b_k \to \infty$ as $\min_k n_k \to \infty$.*

(i) Then, $L_{\mathbf{n},\mathbf{b}}(x) \xrightarrow{P} J(x, P)$ for all points of continuity of $J(\cdot, P)$.

(ii) If $J(\cdot, P)$ is continuous at $J^{-1}(1 - \alpha, P)$, then the event

$$g[\tau_{\mathbf{n}}(\hat{\theta}_{\mathbf{n}} - \theta(P))] \leq L_{\mathbf{n},\mathbf{b}}^{-1}(1 - \alpha) \tag{5}$$

has asymptotic probability equal to $1 - \alpha$.

Proof. (i) Let x be a continuity point of $J(\cdot, P)$. We first argue that it suffices to show that

$$U_{\mathbf{n},\mathbf{b}}(x) \xrightarrow{P} J(x, P) \tag{6}$$

where

$$U_{\mathbf{n},\mathbf{b}}(x) = \frac{1}{q} \sum_{i_1=1}^{q_1} \sum_{i_2=1}^{q_2} \cdots \sum_{i_K=1}^{q_K} 1\{g[\tau_{\mathbf{b}}(\hat{\theta}_{\mathbf{i},\mathbf{b}} - \theta(P))] \leq x\}. \tag{7}$$

Assume without loss of generality that $\theta(P) = 0$. We claim that

$$L_{\mathbf{n},\mathbf{b}}(x) - U_{\mathbf{n},\mathbf{b}}(x) \xrightarrow{P} 0. \tag{8}$$

Given $\epsilon > 0$, there exists $\delta > 0$, so that $|g(x) - g(x')| < \epsilon$ if $\|x - x'\| < \delta$. But then, $|g[\tau_{\mathbf{b}}(\hat{\theta}_{\mathbf{i},\mathbf{b}} - \hat{\theta}_{\mathbf{n}})] - g(\tau_{\mathbf{b}}\hat{\theta}_{\mathbf{i},\mathbf{b}})| < \epsilon$ if $\|\tau_{\mathbf{b}}\hat{\theta}_{\mathbf{n}}\| < \delta$; this latter event has probability tending to one. It follows that, for any fixed $\epsilon > 0$,

$$U_{\mathbf{n},\mathbf{b}}(x - \epsilon) \leq L_{\mathbf{n},\mathbf{b}}(x) \leq U_{\mathbf{n},\mathbf{b}}(x + \epsilon)$$

with probability tending to one. So, assuming we can show (6), the result is established by letting $\epsilon \to 0$ through continuity points $x \pm \epsilon$.

To establish (6), note that $E[U_{\mathbf{n},\mathbf{b}}](x) = J_{\mathbf{b}}(x, P) \to J(x, P)$ as $\min_k b_k \to \infty$. So, it suffices to show that $\text{Var}(U_{\mathbf{n},\mathbf{b}}(x)) \to 0$.

To do this, let $\alpha(s) = \max_k \alpha^{(k)}(s)$, and note that $\alpha(s) \to 0$ when $s \to \infty$ since $\alpha^{(k)}(s) \to 0$ for all k. Momentarily treating the K-samples as a multivariate time series, we see that the mixing coefficient at lag s of this K-variate time series is bounded above by

$$C_K \cdot \alpha(s) \quad \text{where } C_K = K; \tag{9}$$

this is a corollary of Theorem 6.2(I) of [13].

For simplicity of presentation we now focus on the two-sample case; the general case is handled similarly. So, in the case $K = 2$, we have

$$\text{Var}(qU_{\mathbf{n},\mathbf{b}}(x)) = \sum_{i_1=1}^{q_1} \sum_{i_2=1}^{q_2} \sum_{j_1=1}^{q_1} \sum_{j_2=1}^{q_2} \text{Cov}(Y_{\mathbf{i}}, Y_{\mathbf{j}}) \tag{10}$$

where $Y_{\mathbf{i}} = 1\{\tau_{\mathbf{b}} g(\hat{\theta}_{\mathbf{i},\mathbf{b}} - \theta(P)) \le x\}$. Now fix an $\mathbf{i}$ and consider the last two sums, namely

$$\sum_{j_1=1}^{q_1} \sum_{j_2=1}^{q_2} \text{Cov}(Y_{\mathbf{i}}, Y_{\mathbf{j}}). \tag{11}$$

Note that $Y_{\mathbf{i}}$ and $Y_{\mathbf{j}}$ are (possibly) strongly dependent when $|i_1 - j_1| \le b_1$ and/or $|i_2 - j_2| \le b_2$; in these cases, we bound $\text{Cov}(Y_{\mathbf{i}}, Y_{\mathbf{j}})$ by $1/4$ which is a crude upper bound to the variance of any Bernoulli random variable such as $Y_{\mathbf{i}}$. Now if $|i_1 - j_1| > b_1$ and $|i_2 - j_2| > b_2$, then $\text{Cov}(Y_{\mathbf{i}}, Y_{\mathbf{j}}) \le 4C_2 \cdot \alpha(\min_k\{|i_k - j_k| - b_k\})$, by a well-known mixing inequality such as Lemma A.0.2 of [1] coupled with (9).

Letting $c_1, c_2, \ldots$ denote some positive constants, we have

$$\left| \sum_{j_1=1}^{q_1} \sum_{j_2=1}^{q_2} \text{Cov}(Y_{\mathbf{i}}, Y_{\mathbf{j}}) \right| \le c_1 b_1 q_2 + c_2 b_2 q_1 + 4C_2 \sum_{s_1=1}^{q_1} \sum_{s_2=1}^{q_2} \alpha(\min(s_1, s_2)). \tag{12}$$

Assume without loss of generality that $q_1 \ge q_2$. Then,

$$\sum_{s_1=1}^{q_1} \sum_{s_2=1}^{q_2} \alpha(\min(s_1, s_2)) \le 2 \sum_{s_1=1}^{q_1} \sum_{s_2=1}^{s_1} \alpha(s_1) = 2 \sum_{s_1=1}^{q_1} s_1 \alpha(s_1) \le 2q_1 \sum_{s_1=1}^{q_1} \alpha(s_1). \tag{13}$$

Plugging in the bounds (12) and (13) to (10), it follows that

$$\begin{aligned} \text{Var}(U_{\mathbf{n},\mathbf{b}}(x)) &\le \frac{1}{q_1^2 q_2^2} \sum_{i_1=1}^{q_1} \sum_{i_2=1}^{q_2} \sum_{j_1=1}^{q_1} \sum_{j_2=1}^{q_2} |\text{Cov}(Y_{\mathbf{i}}, Y_{\mathbf{j}})| \\ &\le \frac{1}{q_1^2 q_2^2} \sum_{i_1=1}^{q_1} \sum_{i_2=1}^{q_2} \left(c_1 b_1 q_2 + c_2 b_2 q_1 + 2q_1 C_2 \sum_{s_1=1}^{q_1} \alpha(s_1) \right) \\ &\le \frac{c_1 b_1}{q_1} + \frac{c_2 b_2}{q_2} + \frac{2q_1 C_2}{q_1 q_2} \sum_{s_1=1}^{q_1} \alpha(s_1). \end{aligned} \tag{14}$$

The first two terms of (14) tends to zero because $b_k/n_k \to 0$ by assumption. Finally, due to (2), the third term on the RHS of (14) is of the order $O(q_1^{-1} \sum_{s_1=1}^{q_1} \alpha(s_1))$ which also tends to zero because of the strong mixing assumption $\alpha(s) \to 0$ as $s \to \infty$.

(ii) The proof of (ii) is very similar to the proof of Theorem 1 of [14] given our result (i). □

Remark 3.1. The uniform continuity assumption for g can be weakened to continuity if Assumptions 2.1 and 2.2 are replaced by Assumption 2.3. However, the proof is much more complicated and relies on a K-sample version of Theorem 7.2.1 of [1].

Remark 3.2. If $g(\cdot) = \| \cdot \|$, then part (ii) of Theorem 3.1 implies that the statement $\tau_{\mathbf{n}} \|\hat{\theta}_{\mathbf{n}} - \theta(P)\| \le L_{\mathbf{n},\mathbf{b}}^{-1}(1 - \alpha)$ defines an asymptotic $(1 - \alpha)100\%$ confidence region for $\theta(P)$ that is centered around $\hat{\theta}_{\mathbf{n}}$. In the special case of a real-valued parameter $\theta(P)$, the above reduces to an approximate $1 - \alpha$ level confidence interval for $\theta(P)$ that is *symmetric* about $\hat{\theta}_{\mathbf{n}}$. If, on the other hand, $g(\cdot)$ is a 'projection', then using the linearity of the projection mapping, (5) can be solved for $g(\theta)$ yielding a one-sided confidence bound for $g(\theta)$ with asymptotic $(1 - \alpha)100\%$ confidence level; putting two such bounds together will result in a confidence interval for $g(\theta)$. To elaborate, if the above is implemented using $\alpha = 0.05$ and $\alpha = 0.95$, the resulting two bounds will form an approximate 90% *equal-tailed* confidence interval for $g(\theta)$.

Remark 3.3. The approach taken here is to estimate the distribution of some real-valued root. In fact, it is generally possible to use subsampling to estimate the distribution of the **B**-valued random object $\tau_{\mathbf{n}}(\hat{\theta}_{\mathbf{n}} - \theta)$, assuming the weak convergence Assumption 2.3. The proof can be based on a generalization of the argument behind Theorem 7.4.1 of [1] from the one-sample to the K-sample setting of the present paper.

3.2. Studentized roots

Consider the t-statistic for comparing the means of two i.i.d. samples, e.g., our Example 2.1 without dependence. This familiar example shows the necessity of considering 'studentized' roots of the type $g[\tau_{\mathbf{n}}(\hat{\theta}_{\mathbf{n}} - \theta(P))]/\hat{\sigma}_{\mathbf{n}}$ where $\hat{\sigma}_{\mathbf{n}} = \hat{\sigma}_{\mathbf{n}}(\underline{X}^{(1)}, \ldots, \underline{X}^{(K)})$ is a nonnegative (real-valued) statistic. If

$$\hat{\sigma}_{\mathbf{n}} \xrightarrow{P} \quad \text{some } \sigma(P) > 0 \text{ as } \min_k n_k \to \infty \tag{15}$$

then the subsampling application is straightforward; see [1], Section 2.5.1). To describe it, let $J_{\mathbf{n}}^*(P)$ denote the probability law of the 'studentized' root $g[\tau_{\mathbf{n}}(\hat{\theta}_{\mathbf{n}} - \theta(P))]/\hat{\sigma}_{\mathbf{n}}$ with associated distribution function

$$J_{\mathbf{n}}^*(x, P) = \text{Prob}_P\{g[\tau_{\mathbf{n}}(\hat{\theta}_{\mathbf{n}} - \theta(P))]/\hat{\sigma}_{\mathbf{n}} \leq x\}. \tag{16}$$

The subsampling distribution of the studentized root is defined as

$$L_{n,b,*}(x) = \frac{1}{q} \sum_{i_1=1}^{q_1} \sum_{i_2=1}^{q_2} \cdots \sum_{i_K=1}^{q_K} 1\{g[\tau_{\mathbf{b}}(\hat{\theta}_{\mathbf{i},\mathbf{b}} - \hat{\theta}_{\mathbf{n}})]/\hat{\sigma}_{\mathbf{i},\mathbf{b}} \leq x\} \tag{17}$$

where $\hat{\sigma}_{\mathbf{i},\mathbf{b}}$ is evaluated on the same K-fold subsample as $\hat{\theta}_{\mathbf{i},\mathbf{b}}$.

Theorem 3.2. *Assume Assumption 2.1 together with (15). Also assume Assumption 2.2 and (2), that g is uniformly continuous, and that for each $k = 1, \ldots, K$, we have $b_k/n_k \to 0$, $\tau_{\mathbf{b}}/\tau_{\mathbf{n}} \to 0$, and $b_k \to \infty$ as $\min_k n_k \to \infty$. Then,*

(i) $L_{n,b,*}(x) \xrightarrow{P} J^*(x, P)$ *for all points of continuity of $J^*(\cdot, P)$.*
(ii) *If $J^*(\cdot, P)$ is continuous at $J_*^{-1}(1 - \alpha, P)$, then the event*

$$g[\tau_{\mathbf{n}}(\hat{\theta}_{\mathbf{n}} - \theta(P))]/\hat{\sigma}_{\mathbf{n}} \leq L_{n,b,*}^{-1}(1 - \alpha) \tag{18}$$

has asymptotic probability equal to $1 - \alpha$.

Proof. Similar to the proof of Theorem 2.5.1 in [1] in view of Theorem 3.1.

The discussion of Remark 3.2 applies in the context of Theorem 3.2 as well. To elaborate, (18) can be solved for either θ ('norm' case) or for $g(\theta)$ ('projection' case) to yield confidence intervals of the 'studentized' type. $\quad\square$

Remark 3.4. In general, it may be necessary to 'studentize' with a quantity that does not necessarily converge in probability, i.e., a case where (15) does not hold. Subsampling can still work in that case; in particular, Theorem 3.2 would remain valid at the expense of a more complicated assumption analogous to Assumption 11.3.1 of [1].

4. Confidence sets with partial-overlap subsampling

As previously mentioned, the full-overlap case of Section 3 is the most efficient in terms of the accuracy of the subsampling approximation. Nevertheless, the number of K-fold subsamples obtained with full overlap is of the order of $\prod_{k=1}^{K} n_k$ which can be a prohibitively large number when $\min_k n_k$ is large. We thus consider the case of partially overlapping subsamples as a practical alternative.

Again, focus on the kth sample where $1 \leq k \leq K$. Let

$$T_j^{(k)} = (X_{(j-1)h_k+1}^{(k)}, X_{(j-1)h_k+2}^{(k)}, \ldots, X_{(j-1)h_k+b_k}^{(k)})$$

be the jth partial-overlap, block-subsample of size b_k that can be extracted from the series $\{X_1^{(k)}, \ldots, X_{n_k}^{(k)}\}$. The block size b_k is an integer in $[1, n_k]$ as before; the parameter h_k is an integer in $[1, b_k]$ and controls the amount of overlap between $T_j^{(k)}$ and $T_{j+1}^{(k)}$. If $h_k = 1$, then the overlap is the maximum possible as in the previous subsection; if $h_k = b_k$, then there is *no* overlap between $T_j^{(k)}$ and $T_{j+1}^{(k)}$.

Let $\mathcal{T}_k$ denote the set of all size b_k block-subsamples extracted from the kth sample corresponding to the overlap parameter h_k, i.e., let $\mathcal{T}_k = \{T_j^{(k)}, j = 1, \ldots, q_k\}$ where $q_k = \lfloor (n_k - b_k)/h_k \rfloor + 1$ and $\lfloor \cdot \rfloor$ denotes the integer part function. No overlap implies $q_k = \lfloor n_k/b_k \rfloor$; maximum overlap implies $q_k = n_k - b_k + 1$ as in the previous subsection. A K-fold subsample is again constructed by choosing one element from each super-set $\mathcal{T}_k$ for $k = 1, \ldots, K$. Thus, a typical K-fold subsample has

Table 1
Comparison of the asymptotic relative efficiency (ARE) of subsampling estimation of the variance of the sample mean according to different overlap schemes with its resulting effect on q_k, the number of subsamples under consideration; ARE is taken with respect to the full-overlap case.

$\lim(h_k/b_k)$	0	1/4	1/2	1
Overlap	full	75%	50%	zero
ARE	1	1.031	1.125	1.5
$q_k \sim$	$n_k - b_k$	$4\, n_k/b_k$	$2\, n_k/b_k$	n_k/b_k

the form: $T_{i_1}^{(1)}, T_{i_2}^{(2)}, \ldots, T_{i_K}^{(K)}$, where $1 \le i_k \le q_k$ for $k = 1, \ldots, K$. The number of possible K-fold subsamples is then given by $q = \prod_{k=1}^{K} q_k$.

A subsample value of the general statistic $\hat{\theta}_\mathbf{n}$ is given by $\hat{\theta}_{\mathbf{i},\mathbf{b}}$ defined in (3). The subsampling distribution of statistic $\hat{\theta}_\mathbf{n}$ is now defined as

$$L_{\mathbf{n},\mathbf{b},\mathbf{h}}(x) = \frac{1}{q} \sum_{i_1=1}^{q_1} \sum_{i_2=1}^{q_2} \cdots \sum_{i_K=1}^{q_K} 1\{g[\tau_\mathbf{b}(\hat{\theta}_{\mathbf{i},\mathbf{b}} - \hat{\theta}_\mathbf{n})] \le x\} \tag{19}$$

where $\mathbf{h} = (h_1, \ldots, h_K)$.

As in the previous subsection, $L_{\mathbf{n},\mathbf{b},\mathbf{h}}(x)$ provides another consistent approximation to $J_\mathbf{n}(x, P)$ of Assumption 2.1.

Theorem 4.1. *Under the assumptions of Theorem 3.1, and with any choice of $\mathbf{h}$ satisfying $1 \le h_k \le b_k$ for all k, we have:*

(i) $L_{\mathbf{n},\mathbf{b},\mathbf{h}}(x) \xrightarrow{P} J(x, P)$ *for all points of continuity of $J(\cdot, P)$.*

(ii) *If $J(\cdot, P)$ is continuous at $J^{-1}(1 - \alpha, P)$, then the event $g[\tau_\mathbf{n}(\hat{\theta}_\mathbf{n} - \theta(P))] \le L_{\mathbf{n},\mathbf{b},\mathbf{h}}^{-1}(1 - \alpha)$ has asymptotic probability equal to $1 - \alpha$.*

Proof. Similar to the proof of Theorem 3.1. □

As before, the subsampling distribution of the studentized root is defined as

$$L_{n,b,h,*}(x) = \frac{1}{q} \sum_{i_1=1}^{q_1} \sum_{i_2=1}^{q_2} \cdots \sum_{i_K=1}^{q_K} 1\{g[\tau_\mathbf{b}(\hat{\theta}_{\mathbf{i},\mathbf{b}} - \hat{\theta}_\mathbf{n})]/\hat{\sigma}_{\mathbf{i},\mathbf{b}} \le x\}. \tag{20}$$

Theorem 4.2. *Under the assumptions of Theorem 3.2, and with any choice of $\mathbf{h}$ satisfying $1 \le h_k \le b_k$ for all k, we have:*

(i) $L_{n,b,h,*}(x) \xrightarrow{P} J^*(x, P)$ *for all points of continuity of $J^*(\cdot, P)$.*

(ii) *If $J^*(\cdot, P)$ is continuous at $J^{*-1}(1-\alpha, P)$, then the event $g[\tau_\mathbf{n}(\hat{\theta}_\mathbf{n} - \theta(P))]/\hat{\sigma}_\mathbf{n} \le L_{n,b,h,*}^{-1}(1-\alpha)$ has asymptotic probability equal to $1 - \alpha$.*

Proof. Similar to the proof of Theorem 3.2. □

Remark 4.1. Insisting on a partial (although not full) overlap is a practical alternative that can be arbitrarily close to being efficient. For example, in the single-sample case when the statistic is the sample mean, a 75% overlap (i.e., $h_k \sim b_k/4$) leads to relative efficiency that is very close to one as compared to the full-overlap case, even though the number of K-fold subsamples is now only of the order of $\prod_{k=1}^{K} 4\lfloor n_k/b_k \rfloor$ instead of $\prod_{k=1}^{K}(n_k - b_k)$. Typically, $b_k \sim c_k n_k^\beta$ for some $c_k > 0$ and $\beta \in (0, 1)$, and so the resulting computational savings are of substantial magnitude. See Table 1 whose compilation was based on eq. (9.3) of [1].

5. Hypothesis testing

Consider the general problem of testing a null hypothesis H_0 that $P = (P_1, \ldots, P_k) \in \mathbf{P}_0$ against H_1 that $P \in \mathbf{P}_1$. The goal is to construct an asymptotically valid null distribution based on some (generally studentized) test statistic of the form $g(\tau_\mathbf{n}\hat{\theta}_\mathbf{n})/\hat{\sigma}_\mathbf{n}$, whose probability law under P is defined to be $G_\mathbf{n}(P)$ with distribution function denoted by $G_\mathbf{n}(\cdot, P)$. The unstudentized case is obtained by letting $\hat{\sigma}_\mathbf{n} = 1$.

A theoretical critical value of the test is a $1 - \alpha$ quantile of $G_\mathbf{n}(\cdot, P)$, i.e., $G_\mathbf{n}^{-1}(1 - \alpha, P)$. However, this critical value is generally unknown, since it depends on P. The subsampling approximation to this critical value is $G_{\mathbf{n},\mathbf{b},\mathbf{h}}^{-1}(1 - \alpha)$ where

$$G_{\mathbf{n},\mathbf{b},\mathbf{h}}(x) = \frac{1}{q} \sum_{i_1=1}^{q_1} \sum_{i_2=1}^{q_2} \cdots \sum_{i_K=1}^{q_K} 1\{g[\tau_\mathbf{b}\hat{\theta}_{\mathbf{i},\mathbf{b}}]/\hat{\sigma}_{\mathbf{i},\mathbf{b}} \le x\}, \tag{21}$$

and the partial-overlap framework of the previous section was used.

We will make use of the following assumption.

Assumption 5.1. If $P \in \mathbf{P}_0$, there exists a nondegenerate limiting law $G(P)$ such that $G_{\mathbf{n}}(P)$ converges weakly to $G(P)$ as $\min_k n_k \to \infty$.

Let $G(\cdot, P)$ denote the c.d.f. corresponding to $G(P)$. Let $G^{-1}(1 - \alpha, P)$ denote a $1 - \alpha$ quantile of $G(P)$. The following result gives the consistency of the procedure under H_0, under a sequence of contiguous alternatives, and under fixed alternatives; see Section 12.3 of [15] for the definition of contiguity.

Theorem 5.1. *Assume* (2) *and that, for each* $k = 1, \ldots, K$, *we have* $b_k/n_k \to 0$, *and* $b_k \to \infty$ *as* $\min_k n_k \to \infty$. *The choice of the overlap parameter can by any vector* $\mathbf{h}$ *satisfying* $1 \le h_k \le b_k$ *for all* k.

(i) *Further suppose Assumption* 5.1 *holds and that* (15) *holds under* $P \in \mathbf{P}_0$. *Assume* $P \in \mathbf{P}_0$. *If* $G(\cdot, P)$ *is continuous at* $G^{-1}(1 - \alpha, P)$, *then*

$$\text{Prob}_P\{g(\tau_{\mathbf{n}}\hat{\theta}_{\mathbf{n}})/\hat{\sigma}_{\mathbf{n}} > G^{-1}_{\mathbf{n,b,h}}(1 - \alpha)\} \to \alpha \quad \text{as } \min_k n_k \to \infty. \tag{22}$$

Furthermore, if $G(\cdot, P)$ *is continuous and strictly increasing at* $G^{-1}(1 - \alpha, P)$, *then*

$$G^{-1}_{\mathbf{n,b,h}}(1 - \alpha) \xrightarrow{P} G^{-1}(1 - \alpha, P). \tag{23}$$

(ii) *Assume the same conditions as (i). Let* $P^{(\mathbf{n})}$ *denote the joint distribution of the data of size* $\mathbf{n}$ *from* $P = (P_1, \ldots, P_K)$, *where as before,* $\mathbf{n} = (n_1, \ldots, n_K)$ *and* n_k *is the number of observations from the time series* P_k. *Let* $Q_{\mathbf{n}}^{(\mathbf{n})}$ *denote the joint distribution of* $\mathbf{n}$ *observations, with* n_k *of those observations from the time series* $Q_{\mathbf{n},k}$; *denote by* $Q_{\mathbf{n},k}^{(n_k)}$ *the joint distribution of these* n_k *observations. Suppose, for each* k, $Q_{\mathbf{n},k}^{(n_k)}$ *is contiguous to* $P_k^{(n_k)}$, *where* $P = (P_1, \ldots, P_K) \in \mathbf{P}_0$. *Then, under such a contiguous sequence,* $g(\tau_{\mathbf{n}}\hat{\theta}_{\mathbf{n}})/\hat{\sigma}_{\mathbf{n}}$ *is tight. Moreover, if it converges in distribution to some random variable* T *and* $G(\cdot, P)$ *is continuous and strictly increasing at* $G^{-1}(1 - \alpha, P)$, *then the limiting power of the test against such a sequence is* $P\{T > G^{-1}(1 - \alpha, P)\}$— *which is the same limiting power as if we used the asymptotic critical value* $G^{-1}(1 - \alpha, P)$.

(iii) *Assume the test statistic is constructed so that* $\hat{\theta}_{\mathbf{n}} \to \theta(P)$ *in probability as* $\min_k n_k \to \infty$, *where* $\theta(P)$ *is a constant which satisfies* $\theta(P) = 0$ *if* $P \in \mathbf{P}_0$ *and* $\theta(P) > 0$ *if* $P \in \mathbf{P}_1$. *Assume that* (15) *holds for* $P \in \mathbf{P}_1$. *Further assume that* $\liminf(\tau_{\mathbf{n}}/\tau_{\mathbf{b}}) > 1$. *Then, for* $P \in \mathbf{P}_1$, *the rejection probability satisfies*

$$\text{Prob}_P\{g(\tau_{\mathbf{n}}\hat{\theta}_{\mathbf{n}}/\hat{\sigma}_{\mathbf{n}}) > G^{-1}_{\mathbf{n,b,h}}(1 - \alpha)\} \to 1.$$

Proof. In the unstudentized case of full overlap, the behavior of $G_{\mathbf{n,b,h}}(x)$ corresponds exactly to that of (7) in the proof of Theorem 3.1 when $\theta(P) = 0$, and the argument is identical. The case of partial overlap is similar, and the studentized extension straightforward. To prove (ii), the behavior of the subsampling critical value to a degenerate limit under P forces the same behavior under a sequence of contiguous alternatives, and so the result follows by Slutsky's Theorem. The proof of (iii) follows the proof of Theorem 2.6.1(ii) of [1], except that the U-statistic argument there is replaced by the argument used to show (6). □

Remark 5.1. For the validity of (23) and of part (ii) of Theorem 5.1, it is important that $G(\cdot, P)$ is assumed *strictly* increasing at $G^{-1}(1 - \alpha, P)$; this condition was inadvertently omitted from Theorem 2.1 of [2], as well as Theorem 2.6.1 of [1], and should be added back to maintain their validity. Note, however, that the added assumption of strict monotonicity is used only to get convergence of quantiles, i.e., (23); it is not needed for the asymptotic attainment of the correct size of the test, i.e., (22).

6. Block size choice and estimated rates of convergence

6.1. The need for estimated rates of convergence and a different view of studentization

For motivation, consider again Example 2.1, and recall that the statistic $\hat{\theta}_{\mathbf{n}} = \bar{X}^{(2)} - \bar{X}^{(1)}$ is asymptotically normal under standard conditions so that Assumption 2.1 is satisfied. Assuming at least one of $f_2(0), f_1(0)$ is nonzero, the convergence rate of $\hat{\theta}_{\mathbf{n}}$ is $\tau_{\mathbf{n}} = (2\pi f_2(0)/n_2 + 2\pi f_1(0)/n_1)^{-1/2}$. Unfortunately, $\tau_{\mathbf{n}}$ is seen to depend on the unknown parameters $f_2(0), f_1(0)$.

Nevertheless, for the purposes of subsampling we can use an *estimated* convergence rate such as $\hat{\tau}_{\mathbf{n}} = (2\pi \hat{f}_{2,n_2}(0)/n_2 + 2\pi \hat{f}_{1,n_1}(0)/n_1)^{-1/2}$ where $\hat{f}_{2,n_2}(0), \hat{f}_{1,n_1}(0)$ are the nonparametric estimates of $f_2(0), f_1(0)$ mentioned in Example 2.1. This definition of $\hat{\tau}_{\mathbf{n}}$ allows us to also construct $\hat{\tau}_{\mathbf{b}} = (2\pi \hat{f}_{2,n_2}(0)/b_2 + 2\pi \hat{f}_{1,n_1}(0)/b_1)^{-1/2}$ which would be the quantity used in the construction of the subsampling distribution.

Note that, in such a case, using an estimated rate is equivalent to looking at the studentized problem from a different perspective since, in effect, we would be working with the studentized root $\hat{\tau}_{\mathbf{n}}(\hat{\theta}_{\mathbf{n}} - \theta)$. This implicit *"studentization via rate estimation"* idea is applicable to the other two examples of Section 2 as well.

We now leave the narrow framework of Example 2.1 to talk about the general case. Let $\hat{\tau}_\mathbf{b}$ be an estimator of $\tau_\mathbf{b}$ based on the *whole* of the available data, i.e., a function of $\{X_1^{(k)}, \ldots, X_{n_k}^{(k)} \text{ for } k = 1, \ldots, K\}$.

Corollary 6.1. *If* $\hat{\tau}_\mathbf{b}/\tau_\mathbf{b} \xrightarrow{P} 1$, *then all the theorems of this paper remain valid under their respective conditions with* $\hat{\tau}_\mathbf{b}$ *taking the place of* $\tau_\mathbf{b}$ *in the construction of the respective subsampling distributions.*

Proof. Similar to the proof of Theorem 8.3.1 of [1]. $\square$

6.2. Optimal choice of block sizes

The problem of optimal choice of the block sizes $b_1, \ldots, b_K$ is as difficult as it is important in practice. For the case $K = 1$, treatments on optimally choosing the block size for subsampling (and for the related method of block bootstrap) have been given by [6,16,1,17,12,18].

In this section, we consider the case $K > 1$, and develop a rough argument on how to choose the block sizes $b_1, \ldots, b_K$ with the specific purpose of optimizing $L_{\mathbf{n},\mathbf{b}}(x)$ as an estimator of $J_\mathbf{n}(x, P)$. To this end, we should investigate the statistical accuracy of $L_{\mathbf{n},\mathbf{b}}(x)$; instead, we will focus on the easier problem of studying the quantity $U_{\mathbf{n},\mathbf{b}}(x)$ that was defined in (7) for two reasons: (a) $U_{\mathbf{n},\mathbf{b}}(x)$ is tantamount to the subsampling distribution $G_{\mathbf{n},\mathbf{b},1}(x)$ that is the main vehicle for subsampling-based hypothesis testing as in Section 5; and (b) as shown in the proof of Theorem 3.1, $U_{\mathbf{n},\mathbf{b}}(x) \approx L_{\mathbf{n},\mathbf{b}}(x)$. As a matter of fact, by an argument similar to one used in [6], p. 2039), in typical situations we expect to have $L_{\mathbf{n},\mathbf{b}}(x) - U_{\mathbf{n},\mathbf{b}}(x) = O(\tau_\mathbf{b}^2/\tau_\mathbf{n}^2)$. But, as will be apparent from what follows, $\tau_\mathbf{b}^2/\tau_\mathbf{n}^2$ is typically of smaller order of magnitude as compared to the error of $L_{\mathbf{n},\mathbf{b}}(x)$ as an estimator of $J_\mathbf{n}(x, P)$. However, the rate $\tau_\mathbf{n}$ is problem-specific and exceptions to the above claim are possible; hence, we focus on $U_{\mathbf{n},\mathbf{b}}(x)$ in what follows.

Recall that, throughout this paper, the rather minimal condition $\alpha(s) = \max_k \alpha^{(k)}(s) \to 0$ as $s \to \infty$ was assumed. However, if it were further assumed that

$$\sum_{s=1}^{\infty} \alpha(s) < \infty, \tag{24}$$

then

$$\frac{\displaystyle\sum_{s=1}^{\max_j n_j} s^{K-1}\alpha(s)}{\displaystyle\prod_{i=1}^{K} n_i} = O\left(\max_i \frac{b_i}{n_i}\right). \tag{25}$$

To see this, note that all the n_i are of same order of magnitude by (2). Hence, $\sum_{s=1}^{\max_j n_j} s^{K-1}\alpha(s) \le \sum_{s=1}^{\max_j n_j} \max_j n_j^{K-1}\alpha(s) = O(n_1^{K-1}\sum_{s=1}^{\max_j n_j}\alpha(s))$ which is $O(n_1^{K-1})$ if $\sum_{s=1}^{\infty}\alpha(s)$ is finite, thereby implying (25). So, assuming (24), the proof of Theorem 3.1 implies that

$$\mathrm{Var}(U_{\mathbf{n},\mathbf{b}}(x)) = O\left(\sum_{i=1}^{K}\frac{b_i}{n_i} + \frac{\displaystyle\sum_{s=1}^{\max_j n_j} s^{K-1}\alpha(s)}{\displaystyle\prod_{i=1}^{K} n_i}\right) = O\left(\max_i \frac{b_i}{n_i}\right). \tag{26}$$

Now assume that the rate of the convergence stated in Assumption 2.1 is known, i.e., assume that an Edgeworth/Berry–Esseen result of the type

$$J_\mathbf{n}(x, P) = J(x, P) + O(a_\mathbf{n}) \tag{27}$$

is available uniformly in x for some known nonnegative function $a_\mathbf{n}$ satisfying $a_\mathbf{n} \to 0$ as $\min_j n_j \to \infty$. In that case, as argued by [19], we would also have $J_\mathbf{b}(x, P) = J(x, P) + O(a_\mathbf{b})$ and therefore

$$J_\mathbf{b}(x, P) = J_\mathbf{n}(x, P) + O(a_\mathbf{b}). \tag{28}$$

Since $EU_{\mathbf{n},\mathbf{b}}(x) = J_\mathbf{b}(x, P)$, putting (26) and (28) together gives

$$U_{\mathbf{n},\mathbf{b}}(x) - J_\mathbf{n}(x, P) = O(a_\mathbf{b}) + O_P\left(\max_i \sqrt{\frac{b_i}{n_i}}\right). \tag{29}$$

Thus, since $a_\mathbf{b} \to 0$ as $\min_j b_j \to \infty$, to optimize the rate of convergence of $U_{\mathbf{n},\mathbf{b}}(x)$ as an estimator of $J_\mathbf{n}(x, P)$ we must select $b_1, \ldots, b_K$ to satisfy

$$a_\mathbf{b} \sim \max_i \sqrt{\frac{b_i}{n_i}}. \tag{30}$$

Of course, the single Eq. (30) is not enough to determine the values of the K free parameters $b_1, \ldots, b_K$. Recall, however, that by (2) all the n_ks have the same order of magnitude. Thus, it is natural to require that all the b_ks have the same order of magnitude as well, i.e., that

$$C_* \le b_k/b_{k'} \le C^* \quad \text{for all } k, k'. \tag{31}$$

In view of (2), a simple way to enforce (31) is to require

$$b_i/b_j \sim n_i/n_j \quad \text{for all } i, j. \tag{32}$$

Relation (32) is intuitive since the relative proportions of the different samples are reflected in the subsamples. For example, if $n_2 = 2\, n_1$, then (32) implies $b_2 = 2\, b_1$. Note that (32) can be equivalently re-written as

$$b_k \sim b_1 n_k/n_1 \quad \text{for } k = 2, \ldots, K. \tag{33}$$

Formula (33) thus provides the additional $K - 1$ constraints that – coupled with (30) – can uniquely determine the optimal *rates* of the K parameters $b_1, \ldots, b_K$. An example of the applicability of this general idea will be given in the next section with the help of a concrete example.

6.3. Block size choice for Example 2.1

We now give an application of the problem of optimal block size choice in the simple set-up of Example 2.1. To fix ideas, consider the case $g(x) = |x|$ leading to two-sided tests and symmetric confidence intervals. Although results such as (27) are not yet available in the literature it is natural to conjecture that – under appropriate conditions – we would have $a_\mathbf{n} = O(1/(n_1 + n_2)) = O(1/n_1)$ since n_1, n_2 have the same order of magnitude. A similar bound for $a_\mathbf{n}$ would be expected to hold in the case $g(x) = x$ as well, provided that the first marginal distribution of each sample is symmetric, or that the two time series have the same distributions (except for mean) and $n_1 = n_2$.

Consequently, (30) would then imply that the optimal choices for b_1, b_2 are given by $b_k \sim c_k n_k^\beta$ for $\beta = 1/3$ and two positive constants c_1, c_2. As mentioned before, the discussion of Section 6.2 only suggests the optimal rates for b_1, b_2; there remains the question of optimally choosing the constant c_1—since c_2 would be determined from (33) given c_1. Nevertheless, the fact that the rate $b \sim n^{1/3}$ is optimal is very useful, and – interestingly – coincides with the optimal block size rate for subsampling estimation of the *individual* standard errors, i.e., looking at the sample mean of each time series separately and focusing on optimizing the subsampling estimator of standard error. The latter is a well-studied problem for which data-dependent block size choice methods are readily available; see e.g. [18].

References

[1] D.N. Politis, J.P. Romano, M. Wolf, Subsampling, Springer, New York, 1999.
[2] D.N. Politis, J.P. Romano, *K*-sample subsampling, in: S. Dabo-Niang, F. Ferraty (Eds.), Functional and Operational Statistics, Springer Verlag, Heidelberg, 2008, pp. 247–254.
[3] E.J. Hannan, Multiple Time Series, John Wiley, New York, 1970.
[4] D.R. Brillinger, Time Series: Data Analysis and Theory, Holden-Day, New York, 1981.
[5] H. Lütkepohl, Introduction to Multiple Time Series Analysis, Springer, New York, 1993.
[6] D.N. Politis, J.P. Romano, Large sample confidence regions based on subsamples under minimal assumptions, Ann. Stat. 22 (1994) 2031–2050.
[7] A.M. Alonso, E.A. Maharaj, Comparison of time series using subsampling, Comput. Stat. Data Anal. 50 (2006) 2589–2599.
[8] C. Deo, A note on empirical process of strong-mixing sequences, Ann. Probab. 1 (1973) 870–875.
[9] K. Yoshihara, Weak convergence of multidimensional empirical processes for strong mixing sequences of stochastic vectors, Z. Wahrscheinlichkeits-theorie und verwandte Gebiete 33 (1975) 133–137.
[10] A. DasGupta, Asymptotic Theory of Statistics and Probability, Springer, New York, 2008.
[11] A. van der Vaart, J. Wellner, Weak Convergence and Empirical Processes, Springer, New York, 1996.
[12] S.N. Lahiri, Resampling Methods for Dependent Data, Springer, New York, 2003.
[13] R.C. Bradley, Introduction to Strong Mixing Conditions, vol. 1, Kendrick Press, Heber City, Utah, 2007.
[14] R. Beran, Bootstrap methods in statistics, Jahresber. Deutsch. Mathematischen Vereins 86 (1984) 14–30.
[15] E.L. Lehmann, J. Romano, Testing Statistical Hypotheses, 3rd edition, Springer, New York, 2005.
[16] P. Hall, J.L. Horowitz, B.-Y. Jing, On blocking rules for the bootstrap with dependent data, Biometrika 82 (1995) 561–574.
[17] F. Götze, A. Rackauskas, Adaptive choice of bootstrap sample sizes, in: State of the Art in Probability and Statistics, in: IMS Lecture Notes Monogr. Ser., vol. 36, Beachwood, OH, 2001, pp. 286–309.
[18] D.N. Politis, H. White, Automatic block-length selection for the dependent bootstrap, Econometric Rev. 23 (1) (2004) 53–70. Correction: vol. 27, 2008.
[19] P. Bertail, Second-order properties of an extrapolated bootstrap without replacement under weak assumptions, Bernoulli 3 (2) (1997) 149–179.
[20] P. Ango Nze, S. Dupoiron, R. Rios, Subsampling under weak dependence conditions, Working paper no. 2003-42, CREST (www.crest.fr), 2003.