Journal of Business & Economic Statistics

ISSN: 0735-0015 (Print) 1537-2707 (Online) Journal homepage: https://www.tandfonline.com/loi/ubes20

Mean-Structure and Autocorrelation Consistent Covariance Matrix Estimation

Kin Wai Chan

To cite this article: Kin Wai Chan (2020): Mean-Structure and Autocorrelation Consistent Covariance Matrix Estimation, Journal of Business & Economic Statistics, DOI: 10.1080/07350015.2020.1796397

To link to this article: https://doi.org/10.1080/07350015.2020.1796397

JOURNAL OF BUSINESS & ECONOMIC STATISTICS
2020, VOL. 00, NO. 0, 1–15
https://doi.org/10.1080/07350015.2020.1796397

Mean-Structure and Autocorrelation Consistent Covariance Matrix Estimation

Kin Wai Chan

Department of Statistics, The Chinese University of Hong Kong, Shatin, New Territories, Hong Kong

ABSTRACT

We consider estimation of the asymptotic covariance matrix in nonstationary time series. A nonparametric estimator that is robust against unknown forms of trends and possibly a divergent number of change points (CPs) is proposed. It is algorithmically fast because neither a search for CPs, estimation of trends, nor cross-validation is required. Together with our proposed automatic optimal bandwidth selector, the resulting estimator is both statistically and computationally efficient. It is, therefore, useful in many statistical procedures, for example, CPs detection and construction of simultaneous confidence bands of trends. Empirical studies on four stock market indices are also discussed.

ARTICLE HISTORY
Received May 2019
Accepted June 2020

KEYWORDS
Change point detection; Nonlinear time series; Optimal bandwidth selection; Trend inference; Variate difference method

1. Introduction

In many real applications, the observed time series $\{\boldsymbol{Y}_i\}_{i=1}^n$ is a contaminated version of the ideal stationary time series $\{\boldsymbol{X}_i\}_{i=1}^n$. The contamination may consist of an unknown trend, seasonality, and abrupt change points (CPs). This type of nonstationary time series is commonly encountered in Econometrics, Risk Management, Neurology, Genetics, Ecology, etc. (see, e.g., Horváth, Kokoszka, and Steinebach 1999; Grangera and Hyung 2004; Banerjeea and Urga 2005; Mikkonen et al. 2014; Kirch, Muhsal, and Ombao 2015). As a result, assessing the stationarity of $\boldsymbol{Y}_i$ is usually indispensable before conducting inference and modeling. Many tests for this purpose require estimating the *asymptotic covariance matrix* (ACM) of $\bar{\boldsymbol{X}}_n = n^{-1}\sum_{i=1}^n \boldsymbol{X}_i$, namely, $\boldsymbol{\Sigma} = \lim_{n\to\infty} n\,\mathrm{var}(\bar{\boldsymbol{X}}_n)$. Their performances rely on an *efficient* estimator of $\boldsymbol{\Sigma}$ that is *robust* against the mean and autocorrelation structures. This article addresses the problem of mean-structure and autocorrelation consistent (MAC) estimation of $\boldsymbol{\Sigma}$.

One classical problem in assessing stability is *CP* detection. A large class of CP tests is based on the *cumulative sum* (CUSUM) process (e.g., Brown, Durbin, and Evans 1975; Ploberger and Krämer 1992; Jirak 2015). Among them, the celebrated *Kolmogorov–Smirnoff* (KS) test is arguably the most commonly used in detecting a mean shift. In the univariate case, the KS test statistic usually requires an estimator of the *asymptotic variance constant* (AVC) σ^2, that is, the univariate analog of $\boldsymbol{\Sigma}$, for standardization (see Csörgö and Horváth 1997). However, without a *jump robust* estimator of σ^2, the KS test may not be monotonically powerful with respect to the jump magnitude (see Vogelsang 1999; Crainiceanu and Vogelsang 2007; Juhl and Xiao 2009). Indeed, the power may even completely vanish (see Figure 6 for a visualization of this phenomenon). However, many existing approaches are either restricted to one CP or do not fully eliminate the nonmonotone problem. Furthermore, for multidimensional CP tests (e.g., Horváth, Kokoszka, and Steinebach 1999), there is no robust estimator of $\boldsymbol{\Sigma}$. Although Shao and Zhang (2010) proposed a *self-normalized* KS test, which does not require estimating $\boldsymbol{\Sigma}$, it sacrifices power for that. Its power may even completely vanish under a misspecified alternative (see Section 5.4).

Besides detection of CPs, there are many more dedicated procedures for assessing the stability of mean, for example, testing the existence of structural breaks in trends, constructing simultaneous confidence bands (SCB) of trends, and testing non-constancy of trends (see Wu, Woodroofe, and Mentz 2001; Wu 2004; Wu and Zhao 2007, and references therein). All of the aforementioned procedures require an estimator of σ^2 that is jump robust as well as *trend robust*. It is worth mentioning that even in the absence of CP and trend, estimation of σ^2 is already difficult because it requires specifying a *bandwidth* parameter (see, e.g., Andrews 1991; Newey and West 1994). To the best of our knowledge, there is no jump and trend robust estimator of $\boldsymbol{\Sigma}$ that equips with an optimal bandwidth estimator.

In view of the above problems, this article proposes a *single-pass* (i.e., neither estimation of CP nor trend is required) and *fully nonparametric* estimator of $\boldsymbol{\Sigma}$ for general multidimensional time series. It is consistent and robust even if there are a divergent number of CPs and nonconstant trends of unknown forms. Furthermore, a closed-form formula of the optimal bandwidth is derived so that users do not need to resort to computationally intensive cross-validation. Hence, the resulting estimator is MAC, statistically efficient, and computationally fast.

The remaining part of the article is organized as follows. Section 2 reviews some standard estimation methods of $\boldsymbol{\Sigma}$. Section 3 provides motivation of deriving the proposed jump robust estimator; and presents the key theoretical results.

CONTACT Kin Wai Chan *kinwaichan@cuhk.edu.hk* Department of Statistics, The Chinese University of Hong Kong, Lady Shaw Building, Shatin, New Territories, Hong Kong.

Section 4 gives the extension to trend robustness. Implementation issues and generalization are also discussed. Section 5 illustrates finite sample performance Section 6 presents empirical studies on stock market indices. Section 7 concludes the article.

2. Review of Asymptotic Covariance Estimation

2.1. Mathematical Setup

Suppose the observed time series $\{\boldsymbol{Y}_i \in \mathbb{R}^d\}_{i=1}^n$, $d \in \mathbb{N}$, is generated from $\boldsymbol{Y}_i = \boldsymbol{\mu}(i/n) + \boldsymbol{X}_i$, where $\boldsymbol{\mu} : [0,1] \to \mathbb{R}^d$ is a mean function; and $\{\boldsymbol{X}_i \in \mathbb{R}^d\}_{i\in\mathbb{Z}}$ is strictly stationary and ergodic with mean $\mathrm{E}\boldsymbol{X}_i \equiv \boldsymbol{0}$, $i \in \mathbb{Z}$, and autocovariance function (ACVF) $\boldsymbol{\Gamma}_k := \mathrm{E}\left(\boldsymbol{X}_k \boldsymbol{X}_0^\intercal\right)$, $k \in \mathbb{Z}$. Also denote the symmetrized ACVF by $\boldsymbol{\Pi}_k := (\boldsymbol{\Gamma}_k + \boldsymbol{\Gamma}_{-k})/2$, $k \in \mathbb{Z}$. The ACM of $\bar{\boldsymbol{X}}_n := n^{-1}\sum_{i=1}^n \boldsymbol{X}_i$ is defined by

$$\boldsymbol{\Sigma} := \lim_{n\to\infty} n \operatorname{var}\left(\bar{\boldsymbol{X}}_n\right) = \sum_{k=-\infty}^{\infty} \boldsymbol{\Gamma}_k = \sum_{k=-\infty}^{\infty} \boldsymbol{\Pi}_k, \tag{1}$$

provided that the limit exists. Note that the ACM is also known as the *time-average covariance matrix*, *long-run covariance matrix*, and (scaled) *spectral density at zero frequency*.

The unknown mean $\boldsymbol{\mu}(\cdot)$ combines trend, seasonality and jump discontinuities:

$$\boldsymbol{\mu}(i/n) := \boldsymbol{f}(i/n) + \sum_{j=0}^{J} \boldsymbol{\xi}_j \mathbb{1}\{D_j \le i < D_{j+1}\}, \tag{2}$$

where $\boldsymbol{f} := \boldsymbol{f}_n : [0,1] \to \mathbb{R}^d$ is a sequence of continuous functions; $J := J_n$ is the number of CPs; $\boldsymbol{\xi}_j := \boldsymbol{\xi}_{j,n}$ is the mean-shift from $\boldsymbol{f}$ in the time period $[D_j, D_{j+1})$, for $j = 0, \ldots, J$; and $D_j := D_{j,n}$ is the jth CP, for $j = 1, \ldots, J$, such that $1 \equiv D_0 < D_1 < \cdots < D_J < D_{J+1} \equiv n+1$. Here the indicator $\mathbb{1}\{E\} = 1$ if the event E occurs, otherwise $\mathbb{1}\{E\} = 0$. Without loss of generality, assume $\boldsymbol{\xi}_{j-1} \ne \boldsymbol{\xi}_j$ for all $j = 1, \ldots, J$. For simplicity, we write $\boldsymbol{\mu}_i := \boldsymbol{\mu}(i/n)$.

The unobservable time series $\{\boldsymbol{X}_i\}$ is assumed to admit a causal representation $\boldsymbol{X}_i = \boldsymbol{g}(\mathcal{F}_i)$, where $\boldsymbol{g}(\cdot)$ is a d-dimensional measurable function, $\mathcal{F}_i := (\ldots, \boldsymbol{\varepsilon}_{i-1}, \boldsymbol{\varepsilon}_i)$; and $\{\boldsymbol{\varepsilon}_i\}_{i\in\mathbb{Z}}$ are independent and identically distributed multidimensional vectors of innovations (see Wu 2005). This framework is general enough to cover many commonly-used models, for example, autoregressive moving average (ARMA) model, Volterra series, bilinear (BL) model, threshold AR model, and generalized AR conditional heteroscedastic (GARCH) model (see, e.g., Wu 2011; Degras et al. 2012). More multivariate examples defined under this framework can be found in Sections 1 and 2 of Wu and Zaffaroni (2018).

2.2. Mathematical Notations

The following notations are used in the article. Denote $\mathbb{N} = \{1, 2, \ldots\}$ and $\mathbb{N}_0 = \mathbb{N} \cup \{0\}$. For $a \in \mathbb{R}$, $\lfloor a \rfloor$ and $\lceil a \rceil$ are the floor and ceiling of a, respectively. For $a, b \in \mathbb{R}$, denote $a \vee b = \max(a, b)$ and $a \wedge b = \min(a, b)$. When the sample size n is clear, denote $[\![a]\!] = (2 \vee \lceil a \rceil) \wedge (n-1)$. For real sequences $\{a_n\}$ and $\{b_n\}$, write $a_n \sim b_n$ if $a_n/b_n \to 1$; $a_n = o(b_n)$ if $a_n/b_n \to 0$; and $a_n = O(b_n)$ if there are $M > 0$ and N such that $|a_n/b_n| \le M$ for all $n \ge N$.

Matrices and vectors are written in boldface, while scalars are written in normal face. The (r,s)th element of a matrix $\boldsymbol{A}$ is denoted by $A^{[r,s]}$. The uth component of a vector $\boldsymbol{\mu}$ is denoted by $\mu^{[u]}$. In one-dimensional case (i.e., $d = 1$), the ACM in (1) is written as Σ, $\Sigma^{[1,1]}$ or σ^2, and the mean function in (2) is written as $\mu(\cdot)$ or $\mu^{[1]}(\cdot)$.

For any matrix $\boldsymbol{A}$, denote its entry-wise absolute value by $|\boldsymbol{A}|$, its trace by $\operatorname{tr}(\boldsymbol{A})$, its transpose by $\boldsymbol{A}^\intercal$, its column-by-column vectorization by $\operatorname{vec}(\boldsymbol{A})$, and $\boldsymbol{A}^{\otimes 2} = \boldsymbol{A}\boldsymbol{A}^\intercal$. The diagonalization of a vector $\boldsymbol{v}$ is denoted by $\operatorname{diag}(\boldsymbol{v})$, that is, a diagonal matrix whose diagonal elements are the elements of $\boldsymbol{v}$. Denote the column vector of ones, the column vector zeros, and the identity matrix by $\boldsymbol{1}$, $\boldsymbol{0}$, and $\boldsymbol{I}$, respectively.

For any real random variable Z and any $p \ge 1$, denote $\|Z\|_p = \left(\mathrm{E}\,|Z|^p\right)^{1/p}$. For any vector-valued random variable $\boldsymbol{Z}$, we write $\boldsymbol{Z} \in \mathcal{L}^p$ if $\left\|Z^{[u]}\right\|_p < \infty$ for all u. If $\varepsilon_1, \ldots, \varepsilon_n$ are identically and independently distributed (iid) as the standard normal distribution, we write $\varepsilon_1, \ldots, \varepsilon_n \overset{\text{iid}}{\sim} \mathcal{N}(0,1)$. If $\varepsilon, \varepsilon'$ are iid, then we say that ε' is an iid copy of ε.

2.3. Estimation in Stationary Time Series

Suppose that $\boldsymbol{\mu}_1 = \cdots = \boldsymbol{\mu}_n$. There are three standard classes of methods to estimate $\boldsymbol{\Sigma}$. The first one is the *subsampling method* (see, e.g., Meketon and Schmeiser 1984; Carlstein 1986; Song and Schmeiser 1995; Politis, Romano, and Wolf 1999; Chan and Yau 2017b). For instance, the *overlapping batch means* (OBM) estimator is

$$\widehat{\boldsymbol{\Sigma}}_{\mathrm{OBM},n} := \frac{\ell}{n-\ell+1} \sum_{i=\ell}^{n} \left(\frac{1}{\ell} \sum_{j=i-\ell+1}^{i} \widehat{\boldsymbol{X}}_j \right)^{\otimes 2}, \tag{3}$$

where $\ell \in \mathbb{N} \cap (1, n)$ is the *batch-size*, $\widehat{\boldsymbol{X}}_i := \boldsymbol{Y}_i - \bar{\boldsymbol{Y}}_n$ and $\bar{\boldsymbol{Y}}_n := n^{-1}\sum_{i=1}^n \boldsymbol{Y}_i$. The second one is the *kernel method* (see, e.g., Newey and West 1987; Andrews 1991; Politis 2011). For example, the *Bartlett* kernel and the *quadratic spectral* (QS) kernel estimators are

$$\widehat{\boldsymbol{\Sigma}}_{\mathrm{Bart},n} := \sum_{k=-\ell}^{\ell} \mathrm{Bart}(k/\ell)\, \widehat{\boldsymbol{\Gamma}}_k \quad \text{and}$$

$$\widehat{\boldsymbol{\Sigma}}_{\mathrm{QS},n} := \sum_{k=-(n-1)}^{n-1} \mathrm{QS}(k/\ell)\, \widehat{\boldsymbol{\Gamma}}_k, \tag{4}$$

respectively, where $\widehat{\boldsymbol{\Gamma}}_k := n^{-1}\sum_{i=|k|+1}^n \widehat{\boldsymbol{X}}_i \widehat{\boldsymbol{X}}_{i-|k|}^\intercal$, $\mathrm{Bart}(t) := (1-|t|)\mathbb{1}(|t| \le 1)$, and $\mathrm{QS}(t) := 25\{\sin(6\pi t/5)/(6\pi t/5) - \cos(6\pi t/5)\}/(12\pi^2 t^2)$. The third one is based on the *resampling method* (see, e.g., Künsch 1989; Politis and Romano 1994; Paparoditis and Politis 2001; Lahiri 2003). Recently, a new class of estimators based on orthonormal sequences is proposed (see, e.g., Phillips 2005; Sun 2013). The choice of kernel or orthonormal sequences are discussed in Lazarus et al. (2018). Besides, Müller (2014) studied the problem under strong autocorrelation.

Estimation of $\boldsymbol{\Sigma}$ is important because it is usually required in the inference of $\boldsymbol{\mu}$, for example, construction of SCB for $\boldsymbol{\mu}$, and

Table 1. A summary of the robust estimators introduced in Section 2.4, where AC, CV, J, W, and WZ represent estimators proposed in Altissimoa and Corradic (2003), Crainiceanu and Vogelsang (2007), Jirak (2015), Wu (2004), and Wu and Zhao (2007), respectively.

	Robustness		Generality & optimality	
Estimators	(a) J change points	(b) Continuous trend	(c) Dimension d	(d) Optimal ℓ
AC	Yes, $J < \infty$	No	$d = 1$	Not derived
CV, J	Yes, $J = 1$	No	$d = 1$	Not derived
W, WZ	No	Yes	$d = 1$	Not derived
$\widehat{\mathbf{\Sigma}}_{0,2,n}$ (proposal)	Yes, $J = o(n^{1/5})$	Yes	$d \geq 1$	Derived

NOTE: The last row shows a special case of our proposed estimator $\widehat{\mathbf{\Sigma}}_{p,q,n}$, which will be defined in (8). Note that (a) states the robustness against piecewise-constant means with J CPs. Note also that three estimators (namely WZ1, WZ2, and WZ3) are proposed in WZ, but they share the same facts (a)–(d).

output analysis in Markov chain Monte Carlo (see Flegal and Jones 2010; Chan and Yau 2016, 2017a; Liu and Flegal 2018). All three methods above require specifying an unknown *bandwidth* ℓ (or the *batch size*, *block size*, etc.) In practice, ℓ is crucial to the performance of estimators but its optimal value is notoriously difficult to estimate (see, e.g., Politis 2003; Hirukawa 2010).

2.4. Estimation in Nonstationary Time Series

Suppose that $\boldsymbol{\mu}_i \neq \boldsymbol{\mu}_j$ for some $i \neq j$. In this case, as far as we know, (i) all existing estimators of $\mathbf{\Sigma}$ are either restricted to particular forms of mean-structure; and (ii) the estimators are not equipped with the optimal bandwidth. Some representative estimators are listed below, and are summarized in Table 1. For reference, the precise formulas of the estimators are presented in Section C.1 of the supplementary materials.

Altissimoa and Corradic (2003) proposed estimating σ^2 by applying a standard kernel estimator to the time series after being de-trended by a local mean estimator. The resulting estimator is consistent when the mean is a piecewise constant function with finitely many breaks. However, there are some drawbacks. First, they did not derive the optimal bandwidth. It is possible that the optimal bandwidth of the modified estimator is different from that of the standard kernel estimator. Second, the modified estimator introduces an extra tuning parameter, that is, the bandwidth for the local mean estimator. This bandwidth has to be chosen carefully to have a consistent estimator of σ^2. However, its optimal value is unsolved. A similar method was proposed by Juhl and Xiao (2009) in a hypothesis testing context. However, their estimator is inconsistent under nonstationarity.

Crainiceanu and Vogelsang (2007) found that a CP test has a non-monotonic power if a non-robust estimator of σ^2 is used. They proposed an estimator of σ^2 that is robust to one CP. Their idea is to estimate a potential CP and then de-mean the observed time series before and after the estimated CP separately. So, the standard methods in Section 2.3 can be applied to estimate σ^2. Their remedy mitigates the non-monotone problem, but it still has some drawbacks. First, it allows a single CP only; and the trend must be a piecewise constant. In reality, these assumptions may not be satisfied (see Section 6.1). Second, the optimal bandwidth is estimated by a parametric plug-in method proposed by Andrews (1991). If the parametric model is misspecified, its performance is doubtful. Recently, Jirak (2015) proposed a similar de-trending method for estimating σ^2 robustly, but the optimal bandwidth selection issue was not addressed.

In Wu (2004) and Wu and Zhao (2007), they proposed using the first-order difference of nonoverlapping batch means (NBMs) to construct robust estimators of σ^2. There are some drawbacks. First, NBM-type estimators are less efficient than the overlapping batch means counterpart in terms of mean-squared error (MSE) (see Politis, Romano, and Wolf 1999). Thus, their estimators have a significant loss in $\mathcal{L}^2$ efficiency. It is worth noting that there is no trivial way to extend their NBM-type estimators to the more efficient OBM-type estimators (see Remark C.2 in the supplementary materials). Second, they did not derive the optimal bandwidth.

Remark 2.1. Gonçalves and White (2002) proved that two block bootstrap estimators (Künsch 1989; Politis and Romano 1994) are consistent under a mild nonconstant mean structure, namely $U_n := \sum_{i=1}^n (\mu_i - \bar{\mu}_n)^2/n = o(1/\ell)$, where ℓ is the block size used in the estimators, and $\bar{\mu}_n = \sum_{i=1}^n \mu_i/n$. However, $U_n = o(1/\ell)$ does not hold if there is one nontrivial jump in mean. For example, if $\mu_i = \mathbb{1}(i \geq n/2)$, that is, the mean jumps from 0 to 1 at $n/2$, then $U_n \to 1/4 \neq 0$. Gallant and White (1988) documented similar results in the context of heteroscedasticity and autocorrelation consistent (HAC) variance estimation. In Theorem 6.8, they showed that standard HAC estimators are biased unless the mean is a constant.

3. Jump Robustness

3.1. Motivation

Throughout Section 3, $\boldsymbol{\mu}(\cdot)$ is assumed to be a piecewise constant function with J jumps, that is, $\boldsymbol{f}(x) \equiv \mathbf{0}$. This assumption will be relaxed in Section 4.1. Our proposal is to use a *differencing* technique consecutively (see Remark 3.1). If the number of jumps J and the magnitude of jumps $|\boldsymbol{\xi}_j - \boldsymbol{\xi}_{j-1}|$ are not too large, then each value in the *lag-1 difference sequence* $\{\boldsymbol{Y}_i - \boldsymbol{Y}_{i-1}\}_{i=2}^n$ is of mean zero approximately. In this case, we have

$$\mathrm{E}\left\{\frac{1}{2(n-1)}\sum_{i=2}^{n}(\boldsymbol{Y}_i - \boldsymbol{Y}_{i-1})^{\otimes 2}\right\} \approx \frac{1}{2}\left(2\mathbf{\Gamma}_0 - \mathbf{\Gamma}_1 - \mathbf{\Gamma}_1^{\intercal}\right) = \mathbf{\Pi}_0 - \mathbf{\Pi}_1 .$$

So, the semi-average of the lag-1 difference sequence is a potential estimator of $\mathbf{\Pi}_0 - \mathbf{\Pi}_1$, the spread between the symmetrized ACVF at lag 0 and lag 1. Similarly, a potential estimator of $\mathbf{\Pi}_0 - \mathbf{\Pi}_k$ is the semi-average of the *lag-k difference sequence*

$$\widehat{\mathbf{\Psi}}_k := \frac{1}{2(n-|k|+1)}\sum_{i=|k|+1}^{n}\left(\boldsymbol{Y}_i - \boldsymbol{Y}_{i-|k|}\right)^{\otimes 2}, \quad |k| = 0, 1, \ldots, n-1 . \tag{5}$$

The convention $\widehat{\boldsymbol{\Psi}}_t := \widehat{\boldsymbol{\Psi}}_{\lceil |t| \rceil \wedge (n-1)}$ is used for $t \in \mathbb{R}$. The summability of $\boldsymbol{\Pi}_k$ in (1) implies $\boldsymbol{\Pi}_L \to \mathbf{0}$ as $L \to \infty$. Hence, $\widehat{\boldsymbol{\Psi}}_L$ approximates $\boldsymbol{\Pi}_0$ for large L. The *bi-differencing* estimator

$$\widehat{\boldsymbol{\Pi}}_k(L) := \widehat{\boldsymbol{\Psi}}_L - \widehat{\boldsymbol{\Psi}}_k, \quad |k| = 0, 1, \ldots, n-1, \tag{6}$$

is, thus, a potential estimator of $\boldsymbol{\Pi}_k$ when L is large. Observe that the sample mean $\bar{\boldsymbol{Y}}_n$ is not involved in the definition of $\widehat{\boldsymbol{\Pi}}_k(L)$, therefore, we can estimate the ACVFs without estimating the mean $\boldsymbol{\mu}$. The concept of bi-differencing is new. The first and second differencing operations (5) and (6) remove the first and second-order offsets, that is, $\boldsymbol{\mu}(\cdot)$ and $\boldsymbol{\Pi}_0$, respectively. A graphical illustration of the bi-differencing concept can be found in Section A of the supplementary materials. Using the representation $\boldsymbol{\Sigma} = \sum_{k=-\infty}^{\infty} \boldsymbol{\Pi}_k$ in (1), we may use a "naive" estimator of $\boldsymbol{\Sigma}$ as follows:

$$\widehat{\boldsymbol{\Sigma}}_{\text{naive},n} := \sum_{k=-\ell}^{\ell} K(k/\ell)\widehat{\boldsymbol{\Pi}}_k(L) = \sum_{k=-\ell}^{\ell} K(k/\ell)\left(\widehat{\boldsymbol{\Psi}}_L - \widehat{\boldsymbol{\Psi}}_k\right), \tag{7}$$

where $\ell \in \mathbb{N}$ is a bandwidth, $K(\cdot)$ is a kernel function, and $L = c_0\ell$ for some $c_0 \geq 1$.

However, MSE $\left(\widehat{\Sigma}^{[r,s]}_{\text{naive},n}\right) \to 0$ slowly for all $r, s \in \{1, \ldots, d\}$. We demonstrate it through a simple Monte Carlo experiment in Section B of the supplementary materials. An explanation of the slow convergence of $\widehat{\boldsymbol{\Sigma}}_{\text{naive},n}$ is that the "*same* correction term" $\widehat{\boldsymbol{\Psi}}_L$ is used for all $\widehat{\boldsymbol{\Psi}}_k$, $k = 0, \ldots, \ell$, in (7). So, the variance of the "*aggregated* correction term" $\sum_{k=-\ell}^{\ell} K(k/\ell)\widehat{\boldsymbol{\Psi}}_L = O(\ell)\widehat{\boldsymbol{\Psi}}_L$ increases with ℓ quadratically. This is a huge loss in $\mathcal{L}^2$ efficiency because the variance of a standard ACM estimator only increases with ℓ linearly (see Andrews 1991, Proposition 1(a)). Our strategy is to replace $\widehat{\boldsymbol{\Pi}}_k(L)$ in (7) by $\widehat{\boldsymbol{\Pi}}_k(L_k)$ with an appropriately chosen sequence $\{L_k \in \mathbb{R}^+\}_{k\in\mathbb{Z}}$. This sequence should satisfy the following two conditions.

1. (Bias condition) $L_k \uparrow \infty$ as $\ell \uparrow \infty$ for each k so that $\widehat{\boldsymbol{\Psi}}_{L_k} \approx \boldsymbol{\Pi}_0$ for each k. It ensures that $\widehat{\boldsymbol{\Pi}}_k(L_k) = \widehat{\boldsymbol{\Psi}}_{L_k} - \widehat{\boldsymbol{\Psi}}_k$ is able to accurately approximate $\boldsymbol{\Pi}_k$ with a small bias.
2. (Variance condition) $L_k \uparrow \infty$ as $|k| \uparrow \infty$ for each ℓ so that asymptotically different correction terms $\widehat{\boldsymbol{\Psi}}_{L_0}, \ldots, \widehat{\boldsymbol{\Psi}}_{L_\ell}$ are used to correct $\widehat{\boldsymbol{\Psi}}_0, \ldots, \widehat{\boldsymbol{\Psi}}_\ell$ for each ℓ. Since $\widehat{\boldsymbol{\Psi}}_{L_0}, \ldots, \widehat{\boldsymbol{\Psi}}_{L_\ell}$ are not perfectly correlated, it help reducing the variance of the new "aggregated correction term" $\sum_{k=-\ell}^{\ell} K(k/\ell)\widehat{\boldsymbol{\Psi}}_{L_k}$.

Hence, L_k should be increasing with ℓ and $|k|$. One choice of such L_k is a linear combination of ℓ and $|k|$ with positive weights, that is, $L_k = c_0\ell + c_1|k|$, $c_0, c_1 \in \mathbb{R}^+$. Note that $\widehat{\Sigma}^{[r,s]}_{\text{naive},n}$ sets $c_1 = 0$, which violates the variance condition. In the remaining part of this article, we will demonstrate that $L_k = c_0\ell + c_1|k|$ is sufficient to produce optimal results (see Remark 3.2).

Remark 3.1. Difference-based estimators are not new. It has been used in time series analysis and robust estimation (see, e.g., Anderson 1971; Hall, Kay, and Titterinton 1990; Dette, Munk, and Wagner 1998; Hall and Horowitz 2013). However, they are restricted to the estimation of the marginal variance $\boldsymbol{\Gamma}_0$. Differently, we aim at estimating the ACM $\boldsymbol{\Sigma} = \sum_{k\in\mathbb{Z}} \boldsymbol{\Gamma}_k$. It is a harder problem than estimating $\boldsymbol{\Gamma}_0$, and requires a new technique called bi-differencing. Our bi-differencing technique is partially motivated by the bipower variation (Barndorff-Nielsen and Shephard 2004) in the context of testing for jumps in a continuous time series.

Remark 3.2. As we will show in Theorems 3.1 and 3.2, a linear form of L_k already achieves the optimal convergence rate. Although an incremental improvement maybe possible by using a more general form of L_k, we leave it for future investigation.

3.2. Proposed Robust Estimators and Overview of Main Results

For estimation of $\boldsymbol{\Sigma}$, we can use the polynomial kernel $K_q(x) = (1 - |x|^q)\mathbb{1}\{|x| \leq 1\}$ for some $q \in \mathbb{N}$. Then the jump robust estimator of the ACM $\boldsymbol{\Sigma}$ is defined by

$$\begin{aligned}\widehat{\boldsymbol{\Sigma}}_{0,q,n} :&= \sum_{k=-\ell}^{\ell} K_q(|k|/\ell) \cdot \widehat{\boldsymbol{\Pi}}_k \\ &= \sum_{k=-\ell}^{\ell} \left\{1 - \left|\frac{k}{\ell}\right|^q\right\} \left\{\widehat{\boldsymbol{\Psi}}_{c_0\ell + c_1|k|} - \widehat{\boldsymbol{\Psi}}_k\right\},\end{aligned} \tag{8}$$

where $\ell = \ell_n \in \mathbb{N} \cap (1, n)$. Using other kernels in (8), for example, QS$(\cdot)$, is also possible, however, we only focus on the polynomial kernel $K_q(\cdot)$ in this article to avoid complication. Extension to other kernels is routine. Users may choose their favorite kernel and their favorite sequence L_k. Relative to ℓ, these choices have slightly less impact on $\widehat{\boldsymbol{\Sigma}}_{p,q,n}$ at least in the first-order asymptotic. The effect of kernel choice on higher order asymptotic (see, e.g., Lazarus et al. 2018) is theoretically interesting. However, it is beyond the scope of this article. We leave it for future research.

Suppose that $\boldsymbol{\Sigma}_q := \sum_{k=-\infty}^{\infty} |k|^q \boldsymbol{\Pi}_k$ exists and its entries are finite. Under the conditions in Theorems 3.1, 3.2, and part (1) of Corollary 4.1 (to be presented in Sections 3.3 and 4.1), the value of MSE $\left(\widehat{\Sigma}^{[r,s]}_{0,q,n}\right) = \mathrm{E}\left(\widehat{\Sigma}^{[r,s]}_{0,q,n} - \Sigma^{[r,s]}\right)^2$ is given by

$$\begin{aligned}\text{MSE}\left(\widehat{\Sigma}^{[r,s]}_{0,q,n}\right) \sim{}& \left(\Sigma^{[r,s]}_q\right)^2 \frac{1}{\ell^{2q}} \\ &+ \left[\frac{4q^2\,(1+c_1)\,\{\Sigma^{[r,r]}\Sigma^{[s,s]} + \left(\Sigma^{[r,s]}\right)^2\}}{(q+1)(2q+1)}\right]\frac{\ell}{n}\end{aligned} \tag{9}$$

for each r, s. Hence, if $\ell = O(n^{1/(1+2q)})$, then MSE $\left(\widehat{\Sigma}^{[r,s]}_{0,q,n}\right) = O(n^{-2q/(1+2q)})$, which is the optimal convergence rate achieved by the standard estimators (see, e.g., Andrews 1991). In other words, the proposed robust estimator $\widehat{\boldsymbol{\Sigma}}_{0,q,n}$ is rate-optimal in the $\mathcal{L}^2$ sense.

From (9), the MSE of the proposed estimator $\widehat{\Sigma}^{[r,s]}_{0,q,n}$ depends on $\Sigma^{[r,s]}_q$. Hence, its MSE-optimal bandwidth ℓ also depends on $\boldsymbol{\Sigma}_q$. As a result, a robust estimator of $\boldsymbol{\Sigma}_q$ is also important for estimating the optimal bandwidth. This phenomenon is similar to the classic results in non-robust estimation of $\boldsymbol{\Sigma}$ (see, e.g., Andrews 1991). It motivates us to study robust estimation for all $\boldsymbol{\Sigma}_0, \boldsymbol{\Sigma}_1, \boldsymbol{\Sigma}_2, \ldots$. Similar to (8), our proposed jump robust

Table 2. Summary of the statistical meanings of p, q, P and their associated quantities.

Order	Related quantity	Statistical meaning
p	$\boldsymbol{\Sigma}_p = \sum_{k\in\mathbb{Z}} \lvert k\rvert^p \boldsymbol{\Pi}_k$	The estimand is $\boldsymbol{\Sigma}_p$. If $p = 0$, $\boldsymbol{\Sigma}_0 \equiv \boldsymbol{\Sigma}$ reduces to the ACM (1).
q	$K_q(t) = (1 - \lvert t\rvert^q)\mathbb{1}(\lvert t\rvert \le 1)$	The qth order kernel $K_q(\cdot)$ is used in the estimator $\widehat{\boldsymbol{\Sigma}}_{p,q,n}$.
$P \equiv p + q$	$\boldsymbol{\Upsilon}_P = \sum_{k\in\mathbb{Z}} \lvert k\rvert^p \lvert\boldsymbol{\Pi}_k\rvert$	If $\boldsymbol{\Upsilon}_P$ is finite with a larger P, then the autocorrelation is weaker.

estimator of $\boldsymbol{\Sigma}_p = \sum_{k\in\mathbb{Z}} |k|^p \boldsymbol{\Pi}_k$ ($p \in \mathbb{N}_0$) is defined as

$$\widehat{\boldsymbol{\Sigma}}_{p,q,n} := \widehat{\boldsymbol{\Sigma}}_{p,q,n}(\boldsymbol{Y}_{1:n}, \ell, c_0, c_1) := \sum_{k=-\ell}^{\ell} K_q(|k|/\ell) \cdot |k|^p \cdot \widehat{\boldsymbol{\Pi}}_k. \tag{10}$$

The statistical meanings of p and q are summarized in the first two rows of Table 2.

3.3. Theoretical Results

We develop a general estimation procedure which takes various levels of serial dependence into account. For each $P \in \mathbb{N}_0$, define $\boldsymbol{\Upsilon}_P := \sum_{k=-\infty}^{\infty} |k|^P |\boldsymbol{\Pi}_k|$. The finiteness of $\boldsymbol{\Upsilon}_P$ characterizes the strength of serial dependence, thus it is usually served as an assumption for proving consistency of estimators (see, e.g., Politis 2011, Theorem 1). More precisely, we follow Chan and Yau (2017b) to define the coefficient of serial dependence of $\{\boldsymbol{X}_i\}$ by

$$\mathrm{CSD}(\boldsymbol{X}) := \sup\left\{P \in \mathbb{N}_0 : \Upsilon_P^{[\bullet,\bullet]} < \infty\right\}, \quad \text{where}$$
$$\Upsilon_P^{[\bullet,\bullet]} := \max_{r,s\in\{1,\ldots,d\}} \Upsilon_P^{[r,s]}. \tag{11}$$

Clearly, the larger the value of $\mathrm{CSD}(\boldsymbol{X})$, the weaker the serial dependence. For example, consider a univariate fractional Gaussian noise process (Davies and Harte 1987) defined as a Gaussian process with ACVF $\Gamma_k = a(|k| + c)^{-b}$ for each k, where $a, c > 0$ and $b \ge 1$. In this model, $\Upsilon_P < \infty$ if and only if $P < b - 1$. Hence, $\mathrm{CSD}(X) = \lfloor b - 1 \rfloor$. More examples and their associated values of Υ_P can be found in Appendix B of Chan and Yau (2017b). Some multivariate examples can be found in Section C.3 of Chan and Yau (2017a). As we shall see in Section 3.3.1, the assumption of CSD plays a critical role in controlling the bias of the estimator $\widehat{\boldsymbol{\Sigma}}_{p,q,n}$.

Asymptotic theories are built on the framework of dependence measures (see Wu 2005). Recall that $\boldsymbol{X}_i = \boldsymbol{g}(\mathcal{F}_i)$ and $\mathcal{F}_i := (\ldots, \boldsymbol{\varepsilon}_{i-1}, \boldsymbol{\varepsilon}_i)$ (see Section 2.1). Let $\boldsymbol{\varepsilon}_j'$ be an iid copy of $\boldsymbol{\varepsilon}_j$. Denote $\boldsymbol{X}_{i,\{j\}} := \boldsymbol{g}(\mathcal{F}_{i,\{j\}})$ and $\mathcal{F}_{i,\{j\}} := (\mathcal{F}_{j-1}, \boldsymbol{\varepsilon}_j', \boldsymbol{\varepsilon}_{j+1}, \ldots, \boldsymbol{\varepsilon}_i)$. Define the *physical dependence measure* and its aggregated value by, respectively,

$$\delta_{4,i}^{[u]} := \left\|X_i^{[u]} - X_{i,\{0\}}^{[u]}\right\|_4 \quad \text{and} \quad \Delta_4^{[u]} := \sum_{i=0}^{\infty} \delta_{4,i}^{[u]}.$$

For example, consider a univariate linear process (Brockwell and Davis 1991, Definition 3.2.1) defined as $X_i = \sum_{j=0}^{\infty} c_j \varepsilon_{i-j}$, where $\{c_j\}$ are real coefficients such that $\sum_{j=0}^{\infty} |c_j| < \infty$, and $\{\varepsilon_j\}$ are iid noises such that $\mathrm{E}|\varepsilon_0|^4 < \infty$. Then $\delta_{4,i}^{[1]} = K|c_i|$ for each i, and $\Delta_4^{[1]} = K\sum_{j=0}^{\infty} |c_j| < \infty$, where $K = \|\varepsilon_i - \varepsilon_i'\|_4 < \infty$. More univariate examples and their associated values of physical dependence measures can be found in Examples 1–11 of Wu (2011). Also see Models I–VI in Example 1 of Chan and Yau (2017a) for some multivariate examples. Finiteness of $\Delta_4^{[\bullet]} := \max_{u\in\{1,\ldots,d\}} \Delta_4^{[u]}$ (i.e., Assumption 3.1) is a mild and easily-verifiable condition for studying asymptotic properties (see Wu (2007)).

Assumption 3.1 (Short range dependence). The time series $\{\boldsymbol{X}_i\}$ satisfies $\Delta_4^{[\bullet]} < \infty$.

Assumption 3.1 rules out time series having very strong serial dependence, for example, time series with $\Sigma^{[r,s]} = \infty$. Indeed, Assumption 3.1 implies the existence of $\boldsymbol{\Sigma}$. More importantly, it leads to the invariance principle for the (scaled) partial sum $\sum_{i=1}^{\lfloor tn\rfloor} X_i/\sqrt{n}$ for $0 \le t \le 1$. It is required for deriving the variance of $\widehat{\boldsymbol{\Sigma}}_{p,q,n}$. Assumption 3.1 is satisfied by many important time series models, including the aforementioned linear process, ARMA and BL models (see, e.g., Wu 2005; Liu and Wu 2010). Note also that some parallel formulations of dependence like strong mixing coefficient (Rosenblatt 1985) have been widely adopted by researchers. However, the mixing type assumptions are sometimes difficult to verify. On the contrary, Assumption 3.1 is more easily verifiable (see Wu 2011).

We also need to regularize the size of the bandwidth ℓ. Denote $\mathcal{J} := \{1, \ldots, J\}$ and $\mathcal{J}_u := \left\{j \in \mathcal{J} : \mu_{D_{j-1}}^{[u]} \ne \mu_{D_j}^{[u]}\right\}$ for $u = 1, \ldots, d$.

Assumption 3.2 (Conditions on ℓ). The bandwidth $\ell = \ell_n$ satisfies (i) $\ell \to \infty$ as $n \to \infty$, (ii) $\ell = o(n)$ as $n \to \infty$, and (iii) $\{(c_0 + c_1) \vee 1\}\ell \le \inf_{u\in\{1,\ldots,d\}} \inf_{j\in\mathcal{J}_u} (D_{j+1} - D_j)$.

In Assumption 3.2, conditions (i) and (ii) require that the size of ℓ cannot be too small or too large, respectively. These conditions are commonly required in the small-ℓ subsampling approach (i.e., $\ell/n \to 0$) (see Politis, Romano, and Wolf 1999). Condition (iii) states that two consecutive CPs cannot be too close within the same component of the time series. Indeed, condition (iii) is stronger than needed but it makes derivations easier.

3.3.1. Bias and Variance Expressions

Let $\chi := \mathbb{1}\left\{c_1 \ne 1 - c_0 F_{p,q}\right\}$, where $F_{p,q} := (p+2)(p+q+2)/\{(p+1)(p+q+1)\}$. Also let

$$a_n := \sup_{u\in\{1,\ldots,d\}} \sup_{j\in\mathcal{J}_u} \left|\mu_j^{[u]} - \mu_{j-1}^{[u]}\right|.$$

Also recall that $J = J_n$ denotes the number of CPs (see (2)). The bias of the jump robust estimator, $\mathrm{Bias}\left(\widehat{\Sigma}_{p,q,n}^{[r,s]}\right) := \mathrm{E}\left(\widehat{\Sigma}_{p,q,n}^{[r,s]}\right) - \Sigma_p^{[r,s]}$, is given below.

Theorem 3.1 (Bias of the estimator). Suppose that $\boldsymbol{X}_1 \in \mathcal{L}^2$, $\boldsymbol{f}(x) \equiv \boldsymbol{0}$, $\mathrm{CSD}(\boldsymbol{X}) = P \equiv p + q$, and Assumption 3.2 holds, where $p \in \mathbb{N}_0$ and $q \in \mathbb{N}$. Then, for $c_0, c_1 \in \mathbb{R}^+$ and

$r, s \in \{1, \ldots, d\}$,

$$\operatorname{Bias}\left(\widehat{\Sigma}_{p,q,n}^{[r,s]}\right) = -\frac{1}{\ell^q}\Sigma_{p+q}^{[r,s]} + r_n^{\text{bias}}, \tag{12}$$
$$r_n^{\text{bias}} = O\left\{\frac{\ell^{p+1}}{n}\left(\ell^{\chi} + \frac{\ell^2}{n}\right)a_n^2 J_n\right\} + o\left(\frac{1}{\ell^q}\right).$$

In Theorem 3.1, the assumption $\mathrm{CSD}(\boldsymbol{X}) = p + q$ controls the rate of decay of ACVF. For a fixed p, if the value of $\mathrm{CSD}(\boldsymbol{X})$ is larger, then q is larger, and the autocorrelation is weaker. Consequently, the autocorrelation at large lags only introduce a small bias to $\widehat{\Sigma}_{p,q,n}^{[r,s]}$. Hence, it makes sense that the magnitude of the leading term of the bias in (12), that is, $|\Sigma_{p+q}^{[r,s]}|/\ell^q = O(1/\ell^q)$, is decreasing with q. Besides, J_n and a_n determine the frequency of the CPs and the magnitude of the jumps, respectively. From (12), if $a_n^2 J_n$ is not too large so that $r_n^{\text{bias}} = o(1/\ell^q)$, the dominating term of the asymptotic bias is $-\Sigma_{p+q}^{[r,s]}/\ell^q$. Consequently, $\widehat{\Sigma}_{p,q,n}^{[r,s]}$ is asymptotically unbiased as $\ell \to \infty$. Technical conditions for controlling r_n^{bias} are discussed in Corollary 4.1. Moreover, c_0 and c_1 do not affect the first-order asymptotic bias of $\widehat{\boldsymbol{\Sigma}}_{p,q,n}$.

Define $\Xi^{[r,s]} := \Sigma^{[r,r]}\Sigma^{[s,s]} + \left(\Sigma^{[r,s]}\right)^2$. The variance of $\widehat{\Sigma}_{p,q,n}^{[r,s]}$ is given below.

Theorem 3.2 (Variance of the estimator). Suppose that $\boldsymbol{X}_1 \in \mathcal{L}^v$ for $v > 4$, $\boldsymbol{f}(x) \equiv \boldsymbol{0}$, and Assumptions 3.1 and 3.2 hold. If $p \in \mathbb{N}_0$, $q \in \mathbb{N}$, $c_0, c_1 \in \mathbb{R}^+$ and $r, s \in \{1, \ldots, d\}$, then

$$\operatorname{var}\left(\widehat{\Sigma}_{p,q,n}^{[r,s]}\right) = \frac{4q^2(1+c_1)\,\Xi^{[r,s]}\ell^{1+2p}}{(2p+1)(2p+q+1)(2p+2q+1)n} + r_n^{\text{var}}, \tag{13}$$
$$r_n^{\text{var}} = O\left[\frac{\ell^{1+2p}}{n}\left\{\frac{\ell^2}{n}\left(1 + a_n^2 J_n\right) + o(1)\right\}\right].$$

Theorem 3.2 requires Assumption 3.1 because its proof relies on the invariance principle, which is guaranteed by Assumption 3.1 (see, e.g., Wu 2005). However, the detailed strength of serial dependence (i.e., the CSD) is not important for deriving (13). Besides, unlike the asymptotic bias, the variance of $\widehat{\boldsymbol{\Sigma}}_{p,q,n}$ depends also on L_k. However, only c_1 but not c_0 is relevant. Since c_1 determines the speed of divergence of L_k as $k \to \infty$, the variance in (13) is naturally increasing with c_1. Note that c_0 and c_1 are not tuning parameters for balancing the leading terms of the bias and variance because c_0 and c_1 are not involved in (12). Although c_0 and c_1 can be chosen optimally by balancing the second-order bias and variance, that is, r_n^{bias} and r_n^{var}, the effect on $\widehat{\Sigma}_{p,q,n}^{[r,s]}$ is relatively incremental.

Consider $\ell = O(n^{\theta})$ for some $\theta \in (0, 1)$. The MSE-optimal value of θ can be found by balancing the squared-bias and variance of $\widehat{\Sigma}_{p,q,n}^{[r,s]}$ so that the MSE is minimized. Assume $a_n^2 J_n \to \infty$ sufficiently slow so that $r_n^{\text{bias}} = o(1/\ell^q)$ and $r_n^{\text{var}} = o(\ell^{1+2p}/n)$ (see Corollary 4.1 for explicit conditions to guarantee that). In this case, Theorems 3.1 and 3.2 imply that

$$\operatorname{MSE}\left(\widehat{\Sigma}_{p,q,n}^{[r,s]}\right) = \left\{\operatorname{Bias}\left(\widehat{\Sigma}_{p,q,n}^{[r,s]}\right)\right\}^2 + \operatorname{var}\left(\widehat{\Sigma}_{p,q,n}^{[r,s]}\right)$$
$$= O(1/\ell^{2q}) + O(\ell^{1+2p}/n). \tag{14}$$

If $\ell = O(n^{\theta^{\diamond}})$, then (14) achieves its minimum order, that is, $\operatorname{MSE}\left(\widehat{\Sigma}_{p,q,n}^{[r,s]}\right) = O(n^{-\lambda^{\diamond}})$, where

$$\theta^{\diamond} := 1/(1+2p+2q) \quad \text{and} \quad \lambda^{\diamond} := 2q/(1+2p+2q). \tag{15}$$

Note that the superscript "$\diamond$" indicates optimal values. It is worth mentioning that the robust estimator $\widehat{\boldsymbol{\Sigma}}_{p,q,n}$ achieves the same optimal $\mathcal{L}^2$ convergence rate as the non-robust counterparts (see, e.g., Andrews 1991; Chan and Yau 2017b).

3.3.2. Theoretically Optimal Bandwidth

In this subsection, we derive the optimal $\ell \sim \phi n^{\theta^{\diamond}}$, $\phi \in \mathbb{R}^+$, such that the MSE of $\widehat{\boldsymbol{\Sigma}}_{p,q,n}$ is optimized up to the first order including its proportionality constant.

Suppose $r_n^{\text{bias}} = o(1/\ell^q)$ and $r_n^{\text{var}} = o(\ell^{1+2p}/n)$. Let $\boldsymbol{W}$ be a *weight matrix* specifying the entry-wise importance of $\boldsymbol{\Sigma}_p$. For example, $\boldsymbol{W} = (\mathbb{1}\{r \le s\})_{r,s=1}^d$ puts equal weight on each element of the upper triangular part (including the diagonal) of $\boldsymbol{\Sigma}_p$. Write $\boldsymbol{W} \succ 0$ if $W^{[r,s]} \ge 0$ for all r, s, and $W^{[r,s]} > 0$ for at least one pair of r, s. From now on, assume $\boldsymbol{W} \succ 0$. Denote $\boldsymbol{\mathcal{W}} := \operatorname{diag}\{\operatorname{vec}(\boldsymbol{W})\}$ (see Section 2.2 for the definitions of $\operatorname{diag}(\cdot)$ and $\operatorname{vec}(\cdot)$). Then the optimal value of ϕ is the minimizer of

$$\operatorname{AMSE}_{p,q,\boldsymbol{W}}\left(\widehat{\boldsymbol{\Sigma}}_{p,q,\bullet}\right) := \lim_{n\to\infty} n^{2q/(1+2p+2q)}$$
$$\operatorname{MSE}_{\boldsymbol{W}}\left(\widehat{\boldsymbol{\Sigma}}_{p,q,n}\right), \quad \text{where}$$
$$\operatorname{MSE}_{\boldsymbol{W}}\left(\widehat{\boldsymbol{\Sigma}}_{p,q,n}\right) := \mathrm{E}\left[\left\{\operatorname{vec}\left(\widehat{\boldsymbol{\Sigma}}_{p,q,n} - \boldsymbol{\Sigma}_p\right)\right\}^{\intercal} \boldsymbol{\mathcal{W}}\left\{\operatorname{vec}\left(\widehat{\boldsymbol{\Sigma}}_{p,q,n} - \boldsymbol{\Sigma}_p\right)\right\}\right]. \tag{16}$$

The weighted MSE (16) is a generalization of the $\mathcal{L}^2$ risk under the Frobenius norm $\|\cdot\|_F$ because, for any square matrix $\boldsymbol{A}$, $\{\operatorname{vec}(\boldsymbol{A})\}^{\intercal}\boldsymbol{\mathcal{W}}\{\operatorname{vec}(\boldsymbol{A})\} = \operatorname{tr}(\boldsymbol{A}^{\intercal}\boldsymbol{A}) = \|\boldsymbol{A}\|_F^2$ if $\boldsymbol{W} = \boldsymbol{1}\boldsymbol{1}^{\intercal}$. Similar weighting rule is also adopted by Andrews (1991) and Chan and Yau (2017a). By Theorems 3.1 and 3.2, the optimal value of ϕ is $\phi^{\diamond}$, where

$$\phi^{\diamond} := \phi_{p,q}^{\diamond} := \left\{\frac{(2p+q+1)(2p+2q+1)\kappa_{p+q}}{2q(1+c_1)}\right\}^{\theta^{\diamond}}, \tag{17}$$
$$\kappa_{p+q} := \frac{(\operatorname{vec}\boldsymbol{\Sigma}_{p+q})^{\intercal}\boldsymbol{\mathcal{W}}(\operatorname{vec}\boldsymbol{\Sigma}_{p+q})}{(\operatorname{vec}\boldsymbol{\Xi})^{\intercal}\boldsymbol{\mathcal{W}}(\operatorname{vec}\boldsymbol{\Xi})}$$
$$= \frac{(\operatorname{vec}\boldsymbol{\Sigma}_{p+q})^{\intercal}\boldsymbol{\mathcal{W}}(\operatorname{vec}\boldsymbol{\Sigma}_{p+q})}{(\operatorname{vec}\boldsymbol{\Sigma})^{\intercal}\boldsymbol{\mathcal{W}}(\operatorname{vec}\boldsymbol{\Sigma}) + \operatorname{tr}\{\boldsymbol{\mathcal{W}}(\boldsymbol{\Sigma}\otimes\boldsymbol{\Sigma})\}}. \tag{18}$$

Here $\boldsymbol{A}\otimes\boldsymbol{B}$ is the Kronecker's product of $\boldsymbol{A}$ and $\boldsymbol{B}$. Note that the size of the optimal bandwidth $\ell^{\diamond} \sim \phi^{\diamond} n^{\theta^{\diamond}}$ depends on two parameters $\theta^{\diamond}$ and $\phi^{\diamond}$.

- The parameter $\theta^{\diamond} = 1/(1+2P)$ controls the divergence rate of $\ell^{\diamond} = O(n^{\theta^{\diamond}})$. Recall, from Table 2, that if P is small, then the serial dependence is strong. Hence, it makes sense to have a larger optimal bandwidth $\ell^{\diamond}$ to cover more autocovariances.
- The parameter $\phi^{\diamond}$ controls the leading coefficient of $\ell^{\diamond}$. The value of $\phi^{\diamond}$ depends on the unknown κ_{p+q}. We interpret this quantity for univariate $\{X_i\}$. In this case, $\kappa_{p+q} =$

$(\Sigma_{p+q}/\Sigma)^2/2$, which is not purely increasing with the strength of autocorrelation. Indeed, it also depends on the sign of autocorrelation. For example, if $\Gamma_k = \rho^k$ for some $\rho \in (-1,1)$, then $\kappa_2 = 2\rho^2/(1-\rho)^4$, which is not an increasing function of $|\rho|$. This interesting phenomenon also exists in the standard variance estimation (see Andrews 1991, (5.1) and (5.2)). Estimation of $\phi^\diamond$ is presented in Section 4.3.

Formula (17) handles all entries of $\mathbf{\Sigma}$ simultaneously. If the dependence structures of $\{\boldsymbol{X}_i\}$ vary dramatically across entries, we may construct entry-adaptive optimal bandwidth. Let $\boldsymbol{e}_u := (0,\ldots,0,1,0,\ldots,0)^\intercal$ be the uth elementary d-vector, that is, $e_u^{[v]} = \mathbb{1}\{v = u\}$ for all $v \in \{1,\ldots,d\}$. Setting $\boldsymbol{W} = \boldsymbol{e}_r\boldsymbol{e}_s^\intercal$, we can produce the optimal bandwidth for the (r,s)th entry of $\mathbf{\Sigma}$. The resulting optimal asymptotical MSE (AMSE) of $\widehat{\Sigma}_{p,q,n}^{[r,s]}$ is given by

$$n^{\lambda^\diamond_{p,q}}\mathrm{E}\left(\widehat{\Sigma}_{p,q,n}^{[r,s]} - \Sigma_p^{[r,s]}\right)^2 \to \frac{1}{1-\lambda^\diamond}\left\{\frac{\lambda^\diamond(1+c_1)}{2p+q+1}\right\}^{\lambda^\diamond}\Xi^{[r,s]}\left\{\frac{\left(\Sigma_{p+q}^{[r,s]}\right)^2}{\Xi^{[r,s]}}\right\}^{1-\lambda^\diamond}. \tag{19}$$

4. Extension, Discussion, and Implementation

4.1. Extension to Trend Robustness

In this section, we consider the full generalization of $\{\boldsymbol{Y}_i\}$, that is, the assumption $\boldsymbol{f}(x) \equiv 0$ is removed. We measure the amount of fluctuation of $\boldsymbol{f} = \boldsymbol{f}_n$ by

$$b_n := \sup_{u\in\{1,\ldots,d\}}\sup_{x,x'\in[0,1]}\left|\frac{f_n^{[u]}(x) - f_n^{[u]}(x')}{x-x'}\right|,$$

which is small if the fluctuation of $\boldsymbol{f}_n$ does not grow too fast with n. The following two theorems state the bias and variance of $\widehat{\mathbf{\Sigma}}_{p,q,n}$ when $\boldsymbol{\mu}(\cdot)$ consists of jumps and trends.

Theorem 4.1 (Bias of the estimator). If the assumption $\boldsymbol{f}(x) \equiv 0$ is removed, then, under all other conditions in Theorem 3.1, (12) is satisfied with r_n^{bias} being replaced by $R_n^{\mathrm{bias}} = r_n^{\mathrm{bias}} + O\left\{\ell^{p+3}(b_n^2 + J_n a_n b_n)/n^2\right\}$.

Theorem 4.2 (Variance of the estimator). If the assumption $\boldsymbol{f}(x) \equiv 0$ is removed, then, under all other conditions in Theorem 3.2, (13) is satisfied with r_n^{var} being replaced by $R_n^{\mathrm{var}} = r_n^{\mathrm{var}} + O\left(\ell^{4+2p}b_n^2/n^3\right)$.

The optimal bandwidth in (17) and the optimal MSE in (19) remain valid, provided that $R_n^{\mathrm{bias}} = o(1/\ell^q)$ and $R_n^{\mathrm{var}} = o(\ell^{1+2p}/n)$. Consequently, $\widehat{\mathbf{\Sigma}}_{p,q,n}$ achieves the optimal convergence rate even in the presence of jumps and continuous trends. The remainder terms R_n^{bias} and R_n^{var} are influenced by (i) the jump effect $a_n^2 J_n$, (ii) the trend effect b_n^2, and (iii) their joint effect $a_n b_n J_n$. Using these three factors, we define the following classes of mean functions:

$$\mathcal{M}^\diamond := \left\{\boldsymbol{\mu}(\cdot) \,:\, a_n^2 J_n = o\left(n^{\theta^\diamond(p+q-\chi)}\right),\ a_n b_n J_n + b_n^2 = o\left(n^{\theta^\diamond(3p+3q-1)}\right)\right\},$$
$$\mathcal{M} := \left\{\boldsymbol{\mu}(\cdot) \,:\, a_n^2 J_n = o\left(n^{\theta^\diamond(p+2q-\chi)}\right),\ a_n b_n J_n + b_n^2 = o\left(n^{\theta^\diamond(3p+4q-1)}\right)\right\}.$$

Both $\mathcal{M}^\diamond$ and $\mathcal{M}$ include only reasonably well-behaved mean functions $\boldsymbol{\mu}(\cdot)$ such that the aforementioned effects (i), (ii), and (iii) are small. Clearly, $\mathcal{M}^\diamond \subseteq \mathcal{M}$. Simple conditions to control R_n^{bias} and R_n^{var} are given below.

Corollary 4.1. Assume the conditions in Theorems 4.1 and 4.2. Let $\ell = O(n^{\theta^\diamond})$.

1. If $\boldsymbol{\mu}(\cdot) \in \mathcal{M}^\diamond$, then $R_n^{\mathrm{bias}} = o(1/\ell^q)$ and $R_n^{\mathrm{var}} = o(\ell^{1+2p}/n)$.
2. If $\boldsymbol{\mu}(\cdot) \in \mathcal{M}$, then $R_n^{\mathrm{bias}} = o(1)$ and $R_n^{\mathrm{var}} = o(1)$.

The above results remain valid if R_n^{bias} and R_n^{var} are replaced by r_n^{bias} and r_n^{var}, respectively.

Corollary 4.1 ensures that $\widehat{\mathbf{\Sigma}}_{p,q,n}$ is $\mathcal{L}^2$ consistent if $\boldsymbol{\mu}(\cdot)$ belongs to the well-behaved class $\mathcal{M}$. If $\boldsymbol{\mu}(\cdot)$ belongs to a more well-behaved class $\mathcal{M}^\diamond$, we also have the optimal results (17) and (19), which imply that the convergence rate of the estimator $\widehat{\mathbf{\Sigma}}_{p,q,n}$ is not affected by jumps and trends. However, the standard (non-robust) estimators, for example, $\widehat{\mathbf{\Sigma}}_{\mathrm{OBM},n}$ in (3) and $\widehat{\mathbf{\Sigma}}_{\mathrm{QS},n}$ in (4), are not guaranteed to be consistent if $\boldsymbol{\mu}(\cdot) \in \mathcal{M}$.

For example, consider the estimator $\widehat{\mathbf{\Sigma}}_{0,2,n}$ with $\ell = O(n^{1/5})$, and any $c_0 > 0$ and $c_1 \geq 1$. It is $\mathcal{L}^2$ consistent, and satisfies the optimal results (17) and (19) if $\boldsymbol{\mu}(\cdot)$ belongs to

$$\mathcal{M}^\diamond = \{\boldsymbol{\mu}(\cdot) \,:\, a_n^2 J_n = o(n^{1/5}),\quad a_n b_n J_n + b_n^2 = o(n)\}. \tag{20}$$

The class $\mathcal{M}^\diamond$ in (20) includes (but not restricted to) mean functions having piecewise Lipschitz continuous trends with at most $J_n = o(n^{1/5})$ bounded jumps. Note that such J_n is allowed to be divergent to infinity as $n \to \infty$. See the last row of Table 1 for a summary and a comparison with existing robust estimators. We illustrate Corollary 4.1 through a simple simulation experiment. Let $X_i = 0.5X_{i-1} + 0.5\varepsilon_{i-1} + \varepsilon_i$, where $\varepsilon_i \overset{\mathrm{iid}}{\sim} \mathcal{N}(0,1)$. Consider

$$\mu(t) = 4\mathbb{1}(0.2 \leq t < 0.3) + 2e^{2t} + \sin(8\pi t), \tag{21}$$

which consists of two CPs, an exponentially increasing trend, and a periodic structure. In this case, $a_n = 4$, $b_n = 4(e^2 + 2\pi)$, and $J_n = 2$. Hence, the mean function (21) is a member of the class $\mathcal{M}^\diamond$ defined in (20). Figure 1(a) shows a typical realization of $Y_i = X_i + \mu_i$ $(1 \leq i \leq 400)$. The density functions of $\widehat{\Sigma}_{0,2,n}$ and $\widehat{\Sigma}_{\mathrm{QS},n}$ are shown in Figure 1(b). The proposed estimator $\widehat{\Sigma}_{0,2,n}$ concentrates at around the true value $\Sigma = 9$, however, the standard estimator $\widehat{\Sigma}_{\mathrm{QS},n}$ is obviously off the targeted value.

4.2. Comparison With Standard Estimators

The estimator $\widehat{\mathbf{\Sigma}}_{p,q,n}$ sacrifices statistical efficiency to gain robustness. In this section, we investigate how much efficiency is lost.

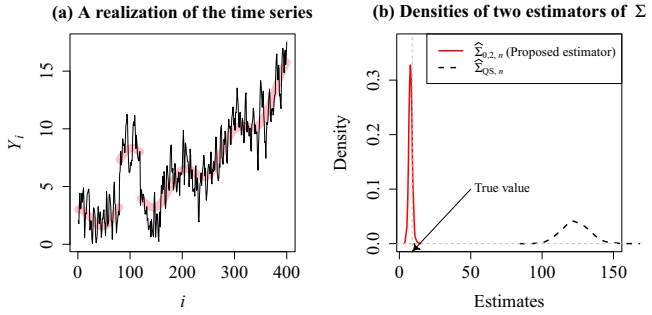

Figure 1. (a) A typical realization of the time series with the mean function defined in (21). (b) The density functions of $\widehat{\Sigma}_{0,2,n}$ and $\widehat{\Sigma}_{\mathrm{QS},n}$ when $n = 400$. The true value is $\Sigma = 9$.

The proposed robust estimator $\widehat{\mathbf{\Sigma}}_{0,1,n}$ uses the Bartlett kernel $K_1(\cdot) \equiv \mathrm{Bart}(\cdot)$. So, we compare it with the standard non-robust Bartlett kernel estimator $\widehat{\mathbf{\Sigma}}_{\mathrm{Bart},n}$ defined in (4). Denote the optimal bandwidths for $\widehat{\mathbf{\Sigma}}_{0,1,n}$ and $\widehat{\mathbf{\Sigma}}_{\mathrm{Bart},n}$ by $\ell^{\diamond}_{0,1}$ and $\ell^{\diamond}_{\mathrm{Bart}}$, respectively. According to (15) and (17), and Equation (5.2) of Andrews (1991), they are given by

$$\ell^{\diamond}_{0,1} \sim \left(\frac{3\kappa_1 n}{2(1+c_1)}\right)^{1/3} \quad \text{and} \quad \ell^{\diamond}_{\mathrm{Bart}} \sim \left(\frac{3\kappa_1 n}{2}\right)^{1/3},$$

respectively, where κ_1 is defined in (18). Denote the resulting optimal estimators by $\widehat{\mathbf{\Sigma}}^{\diamond}_{0,1,n}$ and $\widehat{\mathbf{\Sigma}}^{\diamond}_{\mathrm{Bart},n}$, respectively. The ratio of their weighted MSEs (see (16)) is given below.

Proposition 4.1. Assume the conditions in Theorems 3.1 and 3.2. Let $c_0, c_1 > 0$, $\boldsymbol{W} \succ 0$, and $\boldsymbol{\mu}(t) = \mathbf{0}$ for all $t \in [0,1]$. Then $\mathrm{MSE}_{\boldsymbol{W}}\left(\widehat{\mathbf{\Sigma}}^{\diamond}_{0,1,n}\right)/\mathrm{MSE}_{\boldsymbol{W}}\left(\widehat{\mathbf{\Sigma}}^{\diamond}_{\mathrm{Bart},n}\right) \to (1+c_1)^{2/3} > 1$.

According to Proposition 4.1, the non-robust estimator $\widehat{\mathbf{\Sigma}}^{\diamond}_{\mathrm{Bart},n}$ is more efficient than the robust estimator $\widehat{\mathbf{\Sigma}}^{\diamond}_{0,1,n}$ asymptotically. It makes sense. Note that the efficiency loss is smaller if c_1 is smaller. However, in finite sample, setting $c_1 \approx 0$ may degenerate the estimator to the naive estimator $\widehat{\mathbf{\Sigma}}_{\mathrm{naive},n}$ defined in (7). Hence, using a small $c_1 > 0$ is suggested only if the sample size n is extremely large. Practical suggestion on selecting c_1 is discussed in Section 4.3.

Besides, we also compare our estimator with the most promising (univariate) robust estimator proposed by Wu, Woodroofe, and Mentz (2001), Wu (2004), and Wu and Zhao (2007), namely,

$$\widehat{\sigma}^2_{\mathrm{WZ3},n} := \frac{\ell}{2(m-1)}\sum_{k=2}^{m}(A_k - A_{k-1})^2, \tag{22}$$

where $m = \lfloor n/\ell \rfloor$, and $A_k = \ell^{-1}\sum_{i=1+(k-1)\ell}^{k\ell} Y_i$ is the kth non-overlapping batch mean (NBM) for $k = 1, \ldots, m$. The optimal MSE of $\widehat{\sigma}^2_{\mathrm{WZ3},n}$ was not derived by the authors. For reference, we derive it under the constant mean assumption. Applying similar techniques as in Theorems 3.1 and 3.2, we have $\mathrm{Bias}\left(\widehat{\sigma}^2_{\mathrm{WZ3},n}\right) \sim -\Sigma_1/\ell$ and $\mathrm{var}\left(\widehat{\sigma}^2_{\mathrm{WZ3},n}\right) \sim 7\Sigma^2\ell/(2n)$. The optimal bandwidth is $\ell \sim \{4\Sigma_1^2 n/(7\Sigma^2)\}^{1/3}$. Consequently, $\mathrm{MSE}(\widehat{\Sigma}_{0,1,n})/\mathrm{MSE}(\widehat{\sigma}^2_{\mathrm{WZ3},n}) \to \{8(1+c_1)/21\}^{2/3}$. In particular, when $c_1 = 1$, our estimator $\widehat{\Sigma}_{0,1,n}$ is uniformly better than $\widehat{\sigma}^2_{\mathrm{WZ3},n}$, and satisfies that $\mathrm{MSE}(\widehat{\Sigma}_{\mathrm{Bart},n}) : \mathrm{MSE}(\widehat{\Sigma}_{0,1,n}) : \mathrm{MSE}(\widehat{\sigma}^2_{\mathrm{WZ3},n}) \approx 1.00 : 1.59 : 1.90$ when n is large and their respective optimal bandwidths are used.

4.3. Choices of q, c_0, c_1, and ℓ

The best estimator in Wu and Zhao (2007) has a MSE of size $O(n^{-2/3})$, whereas our proposed estimator $\widehat{\mathbf{\Sigma}}_{0,q,n}$ has a much smaller MSE, that is, $O\{n^{-2q/(1+2q)}\}$, if $q > 1$. In practice, if there is no prior information, we suggest $q = 2$, that is, assuming $\mathrm{CSD}(\boldsymbol{X}) = 2$, which is essentially equivalent to the assumption ($\Upsilon_2 < \infty$) made by Paparoditis and Politis (2001).

Although we develop theories for all $c_1 > 0$, it makes little sense to use $c_1 \in (0,1)$ statistically and intuitively. To see it, observe that $\widehat{\mathbf{\Pi}}_k(L_k) = \widehat{\mathbf{\Psi}}_{L_k} - \widehat{\mathbf{\Psi}}_{|k|}$ is a reasonable estimator of $\mathbf{\Pi}_k$ only if $L_k > |k|$, which is satisfied for all k if and only if $c_1 \geq 1$. Hence, it is sensible (but not necessary) to assume $c_1 \in [1, \infty)$, among which $c_1 = 1$ minimizes the AMSE. So, $c_1 = 1$ is suggested in practice. For $q > 1$, $\widehat{\mathbf{\Sigma}}_{p,q,n}$ has the same AMSE for any $c_0 > 0$, hence, c_0 does not affect the asymptotic behavior. We illustrate in Section C.4 of the supplementary materials that the finite sample performance of $\widehat{\mathbf{\Sigma}}_{p,q,n}$ is essentially the same for any c_0 that is not close to zero. In practice, we suggest using $c_0 = 1$ as a default choice.

If an initial pilot estimate of $\mathbf{\Sigma}_p$ is needed, we can use $\widehat{\mathbf{\Sigma}}_{p,q,n}$ with a rate optimal bandwidth $\ell = O(n^{\theta^{\diamond}})$. In practice, we suggest $\ell = [\![2n^{\theta^{\diamond}}]\!]$, where $[\![t]\!] := (2 \vee \lceil t \rceil) \wedge (n-1)$. According to our simulation experience, this rule-of-thumb bandwidth gives reasonably good performance. Using the notation in (10), we denote the resulting pilot estimator by

$$\widehat{\mathbf{\Sigma}}^{\dagger}_{p,q,n} := \widehat{\mathbf{\Sigma}}_{p,q,n}\left(\boldsymbol{Y}_{1:n}, \ell = [\![2n^{1/(1+2p+2q)}]\!], c_0 = 1, c_1 = 1\right). \tag{23}$$

In particular, for estimating $\mathbf{\Sigma} \equiv \mathbf{\Sigma}_0$, our recommended default estimator is as simple as

$$\widehat{\mathbf{\Sigma}}^{\dagger}_{0,2,n} = \sum_{k=-\ell}^{\ell}\left(1 - \left|\frac{k}{\ell}\right|^2\right)\left(\widehat{\mathbf{\Psi}}_{\ell+|k|} - \widehat{\mathbf{\Psi}}_{|k|}\right), \tag{24}$$

where $\ell = [\![2n^{1/5}]\!]$ and $\widehat{\mathbf{\Psi}}_h = \{2(n-|h|+1)\}^{-1}\sum_{i=|h|+1}^{n}(\boldsymbol{Y}_i - \boldsymbol{Y}_{i-|h|})^{\otimes 2}$. If a more accurate estimate of $\mathbf{\Sigma}_p$ is needed, we can use $\widehat{\mathbf{\Sigma}}_{p,q,n}$ with a fully optimal bandwidth $\ell \sim \phi^{\diamond} n^{\theta^{\diamond}}$. From (17), $\phi^{\diamond}$ is a function of $\mathbf{\Sigma}$ and $\mathbf{\Sigma}_{p+q}$. So, the value of $\phi^{\diamond}$ is unknown. We propose to first estimate $\mathbf{\Sigma}$ and $\mathbf{\Sigma}_{p+q}$ by the pilot estimators $\widehat{\mathbf{\Sigma}}^{\dagger}_{0,2,n}$ and $\widehat{\mathbf{\Sigma}}^{\dagger}_{p+q,2,n}$. Then $\phi^{\diamond}$ is consistently estimated by plugging in these estimated values into (17) and (18), that is,

$$\widehat{\phi}^{\diamond} := \left\{\frac{(2p+q+1)(2p+2q+1)}{\begin{array}{c}(\mathrm{vec}\,\widehat{\mathbf{\Sigma}}^{\dagger}_{p+q,2,n})^{\intercal}\boldsymbol{\mathcal{W}}(\mathrm{vec}\,\widehat{\mathbf{\Sigma}}^{\dagger}_{p+q,2,n}) \\ 2q(1+c_1)(\mathrm{vec}\,\widehat{\mathbf{\Sigma}}^{\dagger}_{0,1,n})^{\intercal}\boldsymbol{\mathcal{W}}(\mathrm{vec}\,\widehat{\mathbf{\Sigma}}^{\dagger}_{0,2,n}) \\ +\mathrm{tr}\{\boldsymbol{\mathcal{W}}(\widehat{\mathbf{\Sigma}}^{\dagger}_{0,2,n} \otimes \widehat{\mathbf{\Sigma}}^{\dagger}_{0,2,n})\}\end{array}}\right\}^{1/(1+2p+2q)}. \tag{25}$$

Using $\widehat{\ell}^{\diamond} := [\![\widehat{\phi}^{\diamond} n^{\theta^{\diamond}}]\!]$, the estimator $\widehat{\mathbf{\Sigma}}_{p,q,n}$ is equipped with the optimal bandwidth asymptotically. The resulting estimator

$$\widehat{\mathbf{\Sigma}}^{\ddagger}_{p,q,n} := \widehat{\mathbf{\Sigma}}_{p,q,n}\left(\boldsymbol{Y}_{1:n}, \ell = [\![\widehat{\phi}^{\diamond} n^{1/(1+2p+2q)}]\!], c_0 = 1, c_1 = 1\right) \tag{26}$$

Algorithm 1: Proposed MAC estimator $\widehat{\mathbf{\Sigma}}^{\ddagger}_{p,q,n}$ for estimating $\mathbf{\Sigma}_p$

Input:
(i) $\boldsymbol{Y}_{1:n}$—d-dimensional time series;
(ii) p—order of the estimand $\mathbf{\Sigma}_p$ (set $p = 0$ for estimation of the ACM $\mathbf{\Sigma}$);
(iii) q—order of the polynomial kernel $K_q(\cdot)$ (set $q = 2$ by default);
(iv) c_0, c_1—parameters (set $c_0 = c_1 = 1$ by default); and
(v) $\boldsymbol{W}$—$d \times d$ weight matrix (set $W^{[r,s]} = \mathbb{1}\{r \le s\}$ for each $1 \le r, s \le d$ by default).
begin
  Compute $\widehat{\mathbf{\Sigma}}^{\dagger}_{0,2,n}$ and $\widehat{\mathbf{\Sigma}}^{\dagger}_{p+q,2,n}$ according to (26);
  Compute $\widehat{\phi}^{\diamond}$ according to (25);
  Compute the estimated optimal bandwidth $\widehat{\ell}^{\diamond} = [\![\widehat{\phi}^{\diamond} n^{1/(1+2p+2q)}]\!]$;
  Compute $\widehat{\mathbf{\Sigma}}^{\ddagger}_{p,q,n} = \widehat{\mathbf{\Sigma}}_{p,q,n}\left(\boldsymbol{Y}_{1:n}, \ell = \widehat{\ell}^{\diamond}, c_0, c_1\right)$ according to (10).
return $\widehat{\mathbf{\Sigma}}^{\ddagger}_{p,q,n}$ – MAC estimator of $\mathbf{\Sigma}_p$.

is called the qth order *MAC* estimator of $\mathbf{\Sigma}_p$. It can be computed by Algorithm 1. The R-package `MAC` is built for implementing it.

4.4. Discussion on Robustness to Heteroscedasticity

Thus far we have assumed that the noise sequence $\{\boldsymbol{X}_i\}$ is stationary (i.e., without heteroscedasticity). Now, suppose that $\{\boldsymbol{X}_i\}$ is not stationary but satisfies $\mathrm{E}(\boldsymbol{X}_i) = 0$. In this case, we define the finite-n version of $\mathbf{\Sigma}_0 = \lim_{n\to\infty} n \operatorname{var}\left(\bar{\boldsymbol{Y}}_n\right)$ by

$$\begin{aligned}\mathbf{\Sigma}_{0,n} &:= n \operatorname{var}\left(\bar{\boldsymbol{Y}}_n\right) = n\mathrm{E}\left(\bar{\boldsymbol{X}}_n \bar{\boldsymbol{X}}_n^{\intercal}\right) \\ &= \frac{1}{n}\sum_{1\le i,j\le n} \mathrm{E}\left(\boldsymbol{X}_i \boldsymbol{X}_j^{\intercal}\right) = \sum_{|k|<n} \mathbf{\Pi}_{k,n}, \end{aligned} \tag{27}$$

where $\mathbf{\Pi}_{k,n} = \sum_{i=1+|k|}^{n} \mathrm{E}(X_i X_{i-|k|}^{\intercal} + X_{i-|k|} X_i^{\intercal})/(2n)$. Following the arguments in Section 3.1, it is not hard to see that $\widehat{\mathbf{\Pi}}_k(L)$ still approximates $\mathbf{\Pi}_{k,n}$. Thus, it is not surprising that the proposed estimator $\widehat{\mathbf{\Sigma}}_{p,q,n}$ continues to be consistent for $\mathbf{\Sigma}_{p,n} := \sum_{|k|<n} |k|^p \mathbf{\Pi}_{k,n}$. Similar to Section 8 of Andrews (1991), we can extend the consistency results to heteroscedastic time series. Suppose the regularity conditions of Theorem 4.1, Theorem 4.2, and part (1) of Corollary 4.1 are satisfied except the following changes.

- The stationarity of the noise sequence $\{\boldsymbol{X}_i\}$ is removed. However, it still satisfies that there is some $\nu > 4$ such that $\boldsymbol{X}_i \in \mathcal{L}^{\nu}$ and $\mathrm{E}(\boldsymbol{X}_i) = \boldsymbol{0}$ for all $i \in \mathbb{Z}$.
- The assumption $\mathrm{CSD}(\boldsymbol{X}) = p + q$ is changed to $\mathrm{CSD}_*(\boldsymbol{X}) = p + q$, where

$$\mathrm{CSD}_*(\boldsymbol{X}) := \sup\left\{P \in \mathbb{N} : \max_{r,s\in\{1,\ldots,d\}} \sum_{k=-\infty}^{\infty} |k|^P \sup_{i\ge 1} \mathrm{E}(X_i^{[r]} X_{i-k}^{[s]}) < \infty\right\}.$$

We also define, for each $P \in \mathbb{N}_0$, that

$$\Sigma^{[r,s]}_{P,*} := \sum_{k=-\infty}^{\infty} |k|^P \sup_{i\ge 1} \mathrm{E}(X_i^{[r]} X_{i-k}^{[s]}) \quad \text{and}$$
$$\Xi^{[r,s]}_{*} := \Sigma^{[r,r]}_{0,*}\Sigma^{[s,s]}_{0,*} + \left(\Sigma^{[r,s]}_{0,*}\right)^2.$$

Note that $\mathrm{CSD}_*(\boldsymbol{X}) = P$ implies that $\Sigma^{[r,s]}_{P,*} < \infty$ and $\Xi^{[r,s]}_{*} < \infty$ for each r, s. Under the modified regularity conditions, the conclusions of Theorems 4.1 and 4.2 are updated to

$$\limsup_{n\to\infty} \ell^{2q}\left\{\mathrm{E}\left(\widehat{\Sigma}^{[r,s]}_{p,q,n}\right) - \Sigma^{[r,s]}_{p,n}\right\}^2 \le \left(\Sigma^{[r,s]}_{p+q,*}\right)^2, \tag{28}$$

$$\limsup_{n\to\infty} \frac{n}{\ell^{1+2p}} \operatorname{var}\left(\widehat{\Sigma}^{[r,s]}_{p,q,n}\right) \le \frac{4q^2(1+c_1)\,\Xi^{[r,s]}_{*}}{(2p+1)(2p+q+1)(2p+2q+1)} \tag{29}$$

for all $r, s \in \{1, \ldots, d\}$. If $\ell = O(n^{1/(1+2p+2q)})$, then (28) and (29) imply that

$$\limsup_{n\to\infty} n^{2q/(1+2p+2q)} \mathrm{E}\left(\widehat{\Sigma}^{[r,s]}_{p,q,n} - \Sigma^{[r,s]}_{p,n}\right)^2 \le C$$

for some $C < \infty$. Hence, $\widehat{\mathbf{\Sigma}}_{p,q,n}$ is a consistent estimator of $\mathbf{\Sigma}_{p,n}$ with the optimal convergence rate. Examples and finite-sample performance of $\widehat{\mathbf{\Sigma}}_{p,q,n}$ in the heteroscedastic case are shown in Section 5.3.

5. Finite Sample Performance

5.1. Efficiency and Robustness Against One Jump

We compare $\widehat{\mathbf{\Sigma}}_{0,q,n}$ with the following estimators in terms of efficiency and robustness.

- (CV) Crainiceanu and Vogelsang (2007) proposed to estimate one potential CP D_1 and then construct a de-trended process, say $\{\widehat{X}_i^{\mathrm{CV}}\}$. The modified OBM estimator $\widehat{\sigma}^2_{\mathrm{CV},n}$ is defined by applying the estimator (3) to $\{\widehat{X}_i^{\mathrm{CV}}\}$ instead of $\{\widehat{X}_i\}$. Andrews (1991)'s AR(1)-plug-in rule is used for selecting the optimal batch size.
- (WZ) Wu and Zhao (2007) used NBMs $\{A_k\}$ to estimate σ^2. They proposed $\widehat{\sigma}^2_{\mathrm{WZ1},n} := \pi\ell\{4(m-1)^2\}^{-1}\sum_{k=2}^{m}|A_k - A_{k-1}|$, $\widehat{\sigma}^2_{\mathrm{WZ2},n} := \ell(2z_{3/4})^{-1}\operatorname{median}_{k\in\{2,\ldots,m\}}|A_k - A_{k-1}|$, and $\widehat{\sigma}^2_{\mathrm{WZ3},n}$ defined in (22), where $\operatorname{median}_{k\in\mathcal{K}} x_k$ denotes the median of $\{x_k\}_{k\in\mathcal{K}}$, and z_p is the $100p\%$ quantile of $\mathcal{N}(0,1)$. They showed, under regularity conditions, that $\widehat{\sigma}^2_{\mathrm{WZ1},n}$ and $\widehat{\sigma}^2_{\mathrm{WZ2},n}$ are weakly consistent if $\ell = [\![n^{5/8}]\!]$, and that $\widehat{\sigma}^2_{\mathrm{WZ2},n}$ is $\mathcal{L}^2$ consistent with $\mathrm{MSE}(\widehat{\sigma}^2_{\mathrm{WZ3},n}) = O(n^{-2/3})$ if $\ell \asymp n^{1/3}$. For $\widehat{\sigma}^2_{\mathrm{WZ3},n}$, we implement it with the an estimated optimal bandwidth by using our proposed estimator (see Section 4.2 for more details). Denote these three estimators by WZ1, WZ2, and WZ3, respectively.
- (AC) Altissimoa and Corradic (2003) proposed using Bartlett kernel estimator after locally detrending the mean. The bandwidth is selected by cross-validation. Denote the resulting estimator by $\widehat{\sigma}^2_{\mathrm{AC},n}$. They proved that $\widehat{\sigma}^2_{\mathrm{AC},n}$ is consistent (see Table 1).

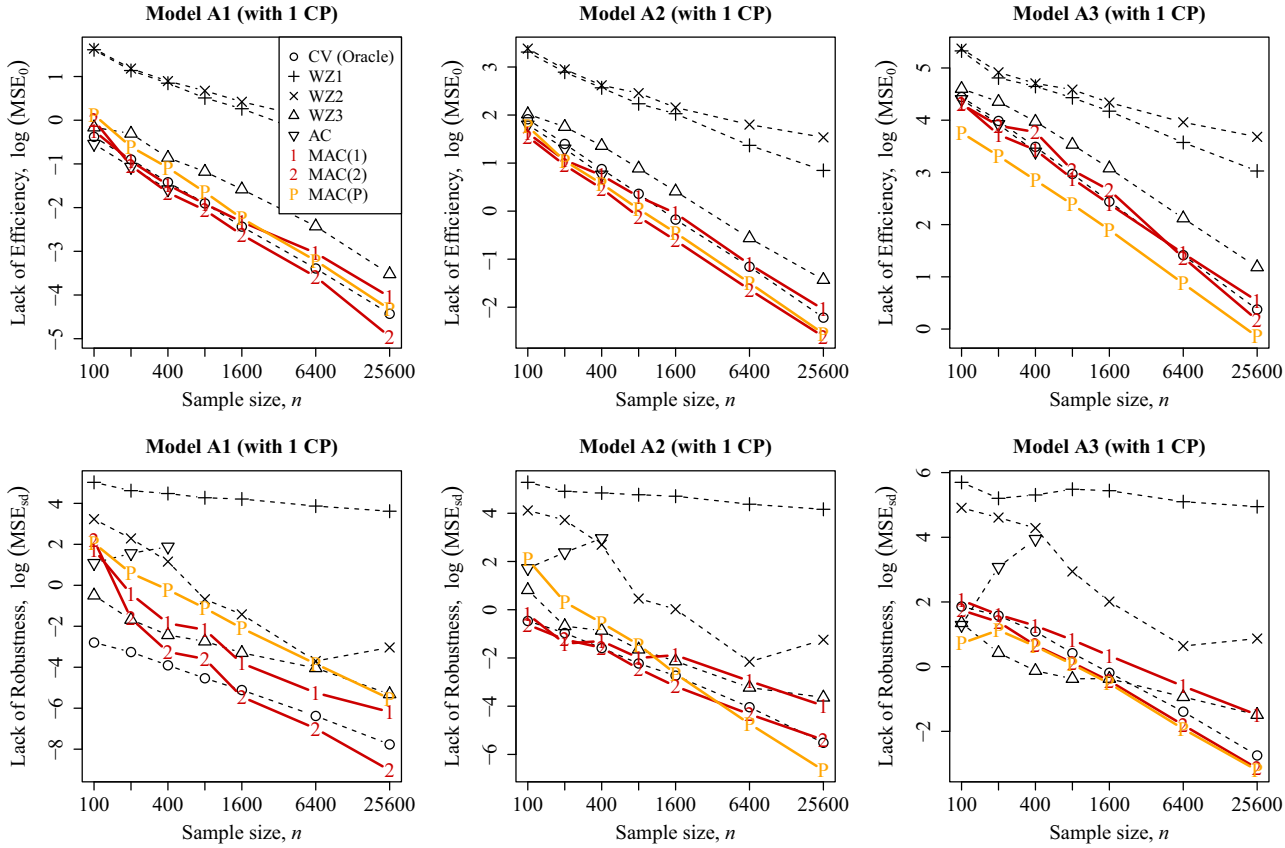

Figure 2. The values of $\log(\text{MSE}_0)$ and $\log(\text{MSE}_{\text{sd}})$ are plotted against n, where MSE_0 denotes the MSE when the jump size $\xi = 0$, and MSE_{sd} denotes the standard deviation of the MSEs across different ξ. Recall that smaller MSE_0 and smaller MSE_{sd} imply higher efficiency and robustness, respectively. Note that $\widehat{\sigma}^2_{\text{AC},n}$ is computed only when $n \le 400$ because it requires a computationally intensive cross-validation step. Note that horizontal axis is plotted in the logarithmic scale for better visualization.

- (MAC) We use the estimators $\widehat{\Sigma}^{\ddagger}_{0,1,n}$ and $\widehat{\Sigma}^{\ddagger}_{0,2,n}$, as well as the pilot estimator $\widehat{\Sigma}^{\dagger}_{0,2,n}$. Denote them by MAC(1), MAC(2), and MAC(P), respectively.

Their detailed formulas are presented in Section C.1 of the supplementary materials for reference. We recall from Table 1 that $\widehat{\sigma}^2_{\text{CV},n}$ is robust to one CP without trend; $\widehat{\sigma}^2_{\text{WZ1},n}$, $\widehat{\sigma}^2_{\text{WZ2},n}$ and $\widehat{\sigma}^2_{\text{WZ3},n}$ are proved to be robust to trends only; $\widehat{\sigma}^2_{\text{AC},n}$ is only proved to be robust to finitely many CPs; and the proposed estimators $\widehat{\Sigma}^{\diamond}_{0,1,n}$ and $\widehat{\Sigma}^{\diamond}_{0,2,n}$ are robust to both trends and a divergent number of CPs. If there is at most one CP, then $\widehat{\sigma}^2_{\text{CV},n}$ is an *oracle* estimator because, in practice, we rarely know that there is at most one CP.

Consider the ARMA(1,1) model: $Y_i = X_i + \mu_i$ where $X_i = aX_{i-1} + b\varepsilon_{i-1} + \varepsilon_i$ and $\varepsilon_i \overset{\text{iid}}{\sim} \mathcal{N}(0,1)$, for $i = 1,\ldots,n$. In particular, consider $a = b = 0.2, 0.4, 0.6$ (Models A1–A3, respectively), $n = 400 \times 4^j$, $j = 0,\ldots,3$, and five different mean sequences $\mu_i = \xi \times \mathbb{1}\{i \le n/2\}$ for $\xi = 0,\ldots,4$. The MSEs are estimated by using 2000 independent replications. The lack of efficiency (MSE_0) is measured by the MSE when $\xi = 0$, whereas the lack of robustness is measured by the standard derivation (MSE_{sd}) of the MSEs across $\xi \in \{0,1,\ldots,4\}$. Smaller MSE_0 and smaller MSE_{sd} imply higher efficiency and robustness, respectively.

The results are shown in Figure 2. Clearly, $\widehat{\sigma}^2_{\text{WZ1},n}$ and $\widehat{\sigma}^2_{\text{WZ2},n}$ perform badly in terms of both efficiency and robustness. The major competitor $\widehat{\sigma}^2_{\text{WZ3},n}$ performs reasonably well in terms of both two measures, however, it is less efficient than all of our proposed estimators ($\widehat{\Sigma}^{\ddagger}_{0,1,n}$, $\widehat{\Sigma}^{\ddagger}_{0,2,n}$, $\widehat{\Sigma}^{\dagger}_{0,2,n}$) in nearly all cases. The estimator $\widehat{\sigma}^2_{\text{AC},n}$ is quite efficient when the mean is a constant, however, it loses all of its efficiency when the jump size is large. For example, when $n = 400$, its MSE inflates 407% when the jump magnitude ξ increases from 0 to 4. Besides, the cross-validation step makes it computationally inefficient.

The proposed estimators $\widehat{\Sigma}^{\ddagger}_{0,1,n}$ and $\widehat{\Sigma}^{\ddagger}_{0,2,n}$ perform the best in nearly all cases. The advantage of $\widehat{\Sigma}^{\ddagger}_{0,2,n}$ is increasingly obvious when n increases. The pilot estimator $\widehat{\Sigma}^{\dagger}_{0,2,n}$ performs quite well, so it is justifiable to use it as an initial guess. It is remarked that $\widehat{\Sigma}^{\dagger}_{0,2,n}$ performs very well in Model A3 because its default tuning parameter accidentally matches the theoretically optimal value. However, this privilege is not general (see, e.g., Figure 5 of another experiment in Section 5.3).

5.2. Robustness Against Trend and Multiple Jumps

In this subsection, we investigate the robustness against both trends and jumps. Consider the same models of $\{X_i\}$ in Section 5.1, but the mean function is replaced by $\mu_i = (i/n)\mathbb{1}\{0.4 \le i/n < 0.7\} + (5i/n - 4)^2\mathbb{1}\{i/n \ge 0.7\}$. Figure 3 shows a typical realization of $\{Y_i\}$ in Model A2. Observe that the trend effect and jump effect are not obvious because they are masked by the intrinsic variability of the noises $\{X_i\}$. This scenario mimics the situation in which the observed time series looks stationary but, indeed, it has been contaminated by a hardly noticeable nonconstant trend and structural breaks.

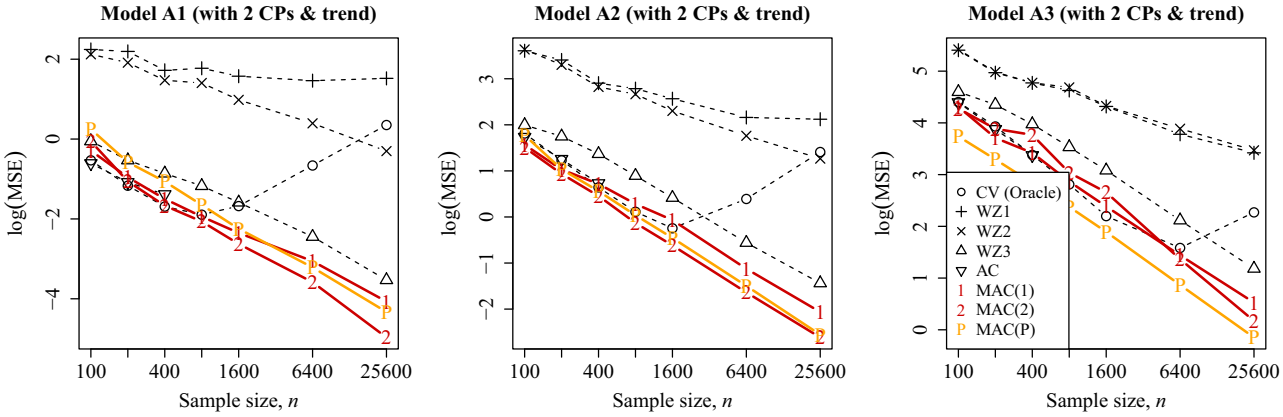

Figure 4. The values of log {MSE(·)} of different estimators are plotted against the sample size n in Models A1–A3. Here the mean function consists of nonconstant trends and multiple jumps (see Section 5.2 and Figure 3). Note that horizontal axis is plotted in the logarithmic scale.

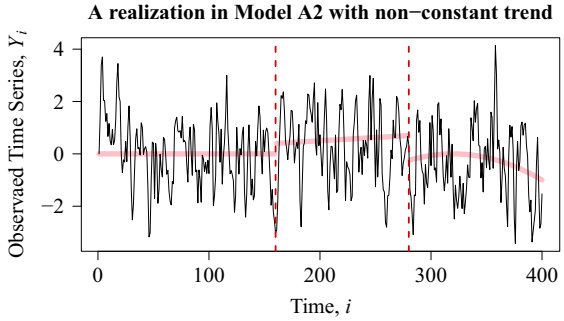

Figure 3. Thin solid line: A realization of $\{Y_i\}$ in Model A2 of length $n = 400$. Thick solid line: The nonconstant mean function $\{\mu_i\}$ in Section 5.2. Dotted vertical lines: The change points.

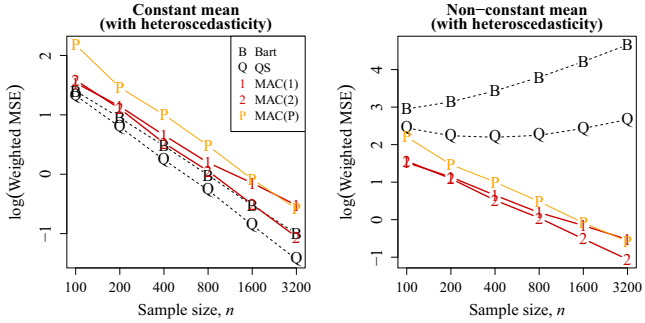

Figure 5. The values of $\log\{\mathrm{MSE}_{\boldsymbol{W}}(\cdot)\}$ for $\widehat{\boldsymbol{\Sigma}}_{\mathrm{Bart},n}$, $\widehat{\boldsymbol{\Sigma}}_{\mathrm{QS},n}$, $\widehat{\boldsymbol{\Sigma}}^{\ddagger}_{0,1,n}$, $\widehat{\boldsymbol{\Sigma}}^{\ddagger}_{0,2,n}$ and $\widehat{\boldsymbol{\Sigma}}^{\dagger}_{0,2,n}$ in the heteroscedastic case are plotted against n, where $\mathrm{vec}(\boldsymbol{W}) = (1, 1/2, 1/2, 1)^{\intercal}$ is used, and $\mathrm{MSE}_{\boldsymbol{W}}(\cdot)$ is defined in (16). The left and right plots show the results in the constant mean and nonconstant mean cases, respectively. Note that the horizontal axes are plotted in the logarithmic scale.

The simulation result is visualized in Figure 4. First, the MSE of the previous oracle estimator $\widehat{\sigma}^2_{\mathrm{CV},n}$ does not decrease with n because it is no longer consistent when the mean is not a piecewise constant function. The estimators $\widehat{\sigma}^2_{\mathrm{WZ1},n}$ and $\widehat{\sigma}^2_{\mathrm{WZ2},n}$ perform poorly again. The estimator $\widehat{\sigma}^2_{\mathrm{WZ3},n}$ and our proposed $\widehat{\Sigma}^{\ddagger}_{0,1,n}$, $\widehat{\Sigma}^{\ddagger}_{0,2,n}$, $\widehat{\Sigma}^{\dagger}_{0,2,n}$ perform well. Among them, $\widehat{\sigma}^2_{\mathrm{WZ3},n}$ performs least well, whereas $\widehat{\Sigma}^{\ddagger}_{0,2,n}$ and $\widehat{\Sigma}^{\dagger}_{0,2,n}$ perform most promisingly. The take-home message is that even if the trend is relatively insignificant, the impact on the estimators of σ^2 can be catastrophic especially when the mean-structure is misspecified.

5.3. Multivariate Time Series With Heteroscedastic Errors

We consider estimation of $\boldsymbol{\Sigma}_{0,n}$ (defined in (27)) for a bivariate time series $\{\boldsymbol{Y}_i = (Y_{i1}, Y_{i2})^{\intercal}\}_{i=1}^n$ with time-varying means and heteroscedastic errors. Let $Y_{ij} = \mu_{ij} + \tau_{ij}X_{ij}$ for $i = 1, \ldots, n$ and $j = 1, 2$, where μ_{ij} is the mean, X_{ij} is a stationary noise, and τ_{ij} creates heteroscedasticity. Two mean sequences are used: (i) $\mu_{ij} = 0$ for all i, j, and (ii) $\mu_{i1} = i/n$, $\mu_{i2} = \mathbb{1}\,(i/n > 1/3)$. We set $\tau_{i1} = 1 + i/(4n)$, $\tau_{i2} = 1 + \sin(4\pi i/n)/(4n)$, and generate $\{X_{ij}\}$ as follows:

$$\begin{bmatrix} X_{i1} \\ X_{i2} \end{bmatrix} = \begin{bmatrix} 0.27 & -0.09 \\ -0.18 & 0.18 \end{bmatrix} \begin{bmatrix} X_{i-1,1} \\ X_{i-1,2} \end{bmatrix} + \begin{bmatrix} 0.01 & -0.14 \\ 0.28 & 0.08 \end{bmatrix} \boldsymbol{\varepsilon}_{i-1} + \boldsymbol{\varepsilon}_i, \quad i = 1, 2, \ldots, n,$$

where $\boldsymbol{\varepsilon}_0, \ldots, \boldsymbol{\varepsilon}_n$ are independent standard bivariate normal random vectors.

The proposed estimators $\widehat{\boldsymbol{\Sigma}}^{\ddagger}_{0,1,n}$, $\widehat{\boldsymbol{\Sigma}}^{\ddagger}_{0,2,n}$, and $\widehat{\boldsymbol{\Sigma}}^{\dagger}_{0,2,n}$ are evaluated. We compare them with the standard Bartlett kernel estimator $\widehat{\boldsymbol{\Sigma}}_{\mathrm{Bart},n}$ and QS kernel estimators $\widehat{\boldsymbol{\Sigma}}_{\mathrm{QS},n}$ (see (4)). The bandwidths of $\widehat{\boldsymbol{\Sigma}}_{\mathrm{Bart},n}$ and $\widehat{\boldsymbol{\Sigma}}_{\mathrm{QS},n}$ are selected by Andrews's (1991) vector AR(1)-plug-in rule. As far as we know, in the multivariate setting, there exists no other estimator that is proved to be consistent and optimal in the presence of nonconstant mean, autocorrelation and heteroscedasticity. The results are shown in Figure 5. We also repeat the experiment with homoscedastic errors, that is, $\tau_{ij} = 1$ for all i, j. Since the results are very similar to the heteroscedastic case, we only present the result in Figure 3 of the supplementary materials.

From Figure 5, all five estimators are consistent in the constant-mean case. However, $\widehat{\boldsymbol{\Sigma}}_{\mathrm{Bart},n}$ and $\widehat{\boldsymbol{\Sigma}}_{\mathrm{QS},n}$ are no longer consistent when the mean is not a constant. On the other hand, the mean-structure does not affect the performance of $\widehat{\boldsymbol{\Sigma}}^{\ddagger}_{0,1,n}$, $\widehat{\boldsymbol{\Sigma}}^{\ddagger}_{0,2,n}$, and $\widehat{\boldsymbol{\Sigma}}^{\dagger}_{0,2,n}$. It verifies the claimed consistency and robustness. In addition, although the pilot estimator $\widehat{\boldsymbol{\Sigma}}^{\dagger}_{0,2,n}$ does not perform as well as the optimal estimator $\widehat{\boldsymbol{\Sigma}}^{\ddagger}_{0,2,n}$, it is still able to give sufficiently good results. It supports the use of $\widehat{\boldsymbol{\Sigma}}^{\dagger}_{0,2,n}$ as an initial estimator in practice.

5.4. Change-Point Detection

In this subsection, we consider the CP detection problem, that is, to test $H_0 : \mathrm{E}Y_1 = \cdots = \mathrm{E}Y_n$ against $H_1 : \exists D_1$ such that $\mathrm{E}Y_1 = \cdots = \mathrm{E}Y_{D_1-1} \neq \mathrm{E}Y_{D_1} = \cdots = \mathrm{E}Y_n$. We analyze (i) whether CP tests are monotonically powerful with respect to the magnitude of jump $|\mathrm{E}Y_{D_1} - \mathrm{E}Y_{D_1-1}|$; and (ii) their power losses under a misspecified alternative hypothesis.

Let $T_n(k) := n^{-1/2}\sum_{i=1}^{k}\widehat{X}_i$ be the CUSUM process of $\widehat{X}_i = Y_i - \bar{Y}_n$. The standard KS test statistic is defined by $T_n := \max_{k\in\{1,\ldots,n\}} |T_n(k)/\widehat{\sigma}|$, where $\widehat{\sigma}$ is a consistent estimator of σ. Then, H_0 is rejected at 5% level if $T_n > 1.358$. Alternatively, a self-normalized KS test (Shao and Zhang 2010) can be used. Following them, we compare

- (SZ) their self-normalized KS test, and
- (KS) the standard KS tests with different estimators of σ^2, namely, $\widehat{\sigma}^2_{\mathrm{A},n}$, $\widehat{\sigma}^2_{\mathrm{CV},n}$ $\widehat{\sigma}^2_{\mathrm{JX},n}$, $\widehat{\sigma}^2_{\mathrm{WZ3},n}$, $\widehat{\sigma}^2_{\mathrm{AC},n}$, and $\widehat{\Sigma}^{\ddagger}_{0,2,n}$, where $\widehat{\sigma}^2_{\mathrm{A},n}$ is Bartlett kernel estimator with Andrew's AR(1) plug-in selector of ℓ; the estimator $\widehat{\sigma}^2_{\mathrm{JX}}$ is proposed by Juhl and Xiao (2009); and all other estimators are defined in Section 5.1.

Detailed formulas of the above CP tests and estimators of σ^2 are presented in Section C.3 of the supplementary materials for reference. Consider the bilinear model: $Y_i = X_i + \mu_i$ where $X_i = (a + b\varepsilon_i)X_{i-1} + \varepsilon_i$ and $\varepsilon_i \overset{\mathrm{iid}}{\sim} \mathcal{N}(0,1)$, for $i = 1,\ldots,n$. The physical dependence measure decays at the rate $\delta^{[1]}_{4,n} = O(\varrho^n)$, where $\varrho = \sqrt{a^2+b^2}$ (see Wu 2005, 2011). If ϱ is larger, the serial dependence is stronger. We use $a = 0.33, 0.36, 0.39$ and $b = 0.5, 0.6, 0.7$ so that $\varrho = 0.6, 0.7, 0.8$, respectively. Denote them by Models B1–B3, respectively.

Both SZ and KS tests assume that the mean function is a piecewise constant with one CP when H_0 is false. If it is actually the case, we call that the alternative hypothesis is correctly specified, otherwise, the alternative hypothesis is said to be misspecified. We consider the following two alternative hypotheses in the experiments:

- (correctly specified alternative) H_1: $\mu_i = \xi \times \mathbb{1}\{i/n > 1/4\}$ for $i = 1,\ldots,n$; and
- (misspecified alternative) H_1': $\mu_i = \xi(1 + e^{-10i/n+5}) \times \mathbb{1}\{i/n > 1/4\}$ for all $i = 1,\ldots,n$,

where the value of $\xi \in \mathbb{R}$ controls the jump magnitude. If $\xi = 0$, both H_1 and H_1' reduce to H_0. In H_1, the mean jumps to ξ at $i = \lfloor n/4 \rfloor + 1$, and then stays constant; whereas, in H_1', the mean shoots up at $i = \lfloor n/4 \rfloor + 1$, and then decays to ξ. In practice, CP may arrive like H_1' instead of H_1, hence, a good CP test should be powerful in both cases. A good size-α CP test should satisfy the following four properties, where $\alpha \in (0,1)$.

- (Size correctness) The probability of rejecting H_0 is close to α when H_0 is correct.
- (Powerfulness) The probability of rejecting H_0 is high when H_0 is incorrect.
- (Monotonicity of power) The power is increasing with the magnitude of jump $|\xi|$.
- (Robustness) The test is still powerful under misspecified alternative hypotheses.

The simulation is conducted for $n = 100, 400, 800$ with nominal size $\alpha = 5\%$. Since the results are similar under different models, we only report the results under Model B2 here (see Figure 6). The full results are deferred to Section C.3 in the supplementary materials. The size-adjusted power curves are also presented in the supplementary materials for reference. Under H_1, all tests except KS(A) and KS(JX) have monotonic powers with respect to $|\xi|$. The test KS(WZ3) commits the Type I error more frequently than the nominal value even when the sample size is large. This over-size phenomenon is due to the use of inefficiency estimator of σ^2. The power curves are largely the same for KS(CV), KS(AC), and KS(MAC(2)) as they are essentially the same test. Observe that SZ is significantly less powerful when $0 < \xi < 1$.

Under H_1', all tests except KS(MAC(2)) and KS(WZ(3)) immediately lose all power when $|\xi| > 0$. In particular, SZ remains powerless even when n and ξ are large. It is not desirable because SZ is very sensitive to whether the alternative hypothesis is well-specified. For KS(A), KS(CV), KS(JX), and KS(AC), they are not powerfully because of using inconsistent or inefficient estimators of σ^2. It gives a sign of warning to use these tests in practice. It is worth emphasizing that KS(WZ3) seems more powerful than KS(MAC(2)). However, it is just because KS(WZ3) rejects too frequently no matter H_0 is true or not. Hence, the apparently more powerful KS(WZ3) test is not reliable. Among all tests above, our proposed test KS(MAC(2)) is the only monotonically powerful test that has accurate size and is insensitive to misspecification of the alternative hypothesis.

6. Empirical Studies

6.1. Change Point Detection in S&P 500 Index

The Standard & Poor's 500 (S&P 500) Index is a stock market index based on 500 representative companies in the USA. The daily adjusted close prices of the index, from 3 January 2006 to 30 December 2011 ($n = 1511$), are investigated. The dataset can be downloaded from *http://finance.yahoo.com/quote/%5EGSPC/history*. The financial crisis in 2008 is believed to have a tremendous impact on the global stock market. We suspect that it led to an abrupt change in the stock market. Testing this claim is important since a noncontinuous impact implies that the economy may have a structural change.

Denote the logarithm of the S&P 500 Index by Y_i. Observe that there is an obvious trend in Y_i (see Figure 7). A standard approach is to study the return series $y_i := Y_i - Y_{i-1}$ to get rid of the trend component. This differencing step is essential for many standard CP tests, for example, SZ and KS tests presented in Section 5.4, because they cannot handle trends. Using the CUMSUM-type CP estimator $\widehat{D}_1$ (see (1) of the supplementary materials for its formula), we estimate the CP to be 10 March 2009. It is remarked that the same CP is detected by the method described in Altissimoa and Corradic (2003). Hence, the CP test fails to capture the 2008 financial crisis. Indeed, testing H_0: "$\mathrm{E}y_1 = \cdots = \mathrm{E}y_n$" by the KS(MAC) test defined in Section 5.4, we fail to reject H_0 at 5% level. We conclude that the 2008 financial crisis has no jump impact on the return y_i. Since taking the difference of Y_i may cancel out the potential jump effect, it seems desirable to analyze Y_i

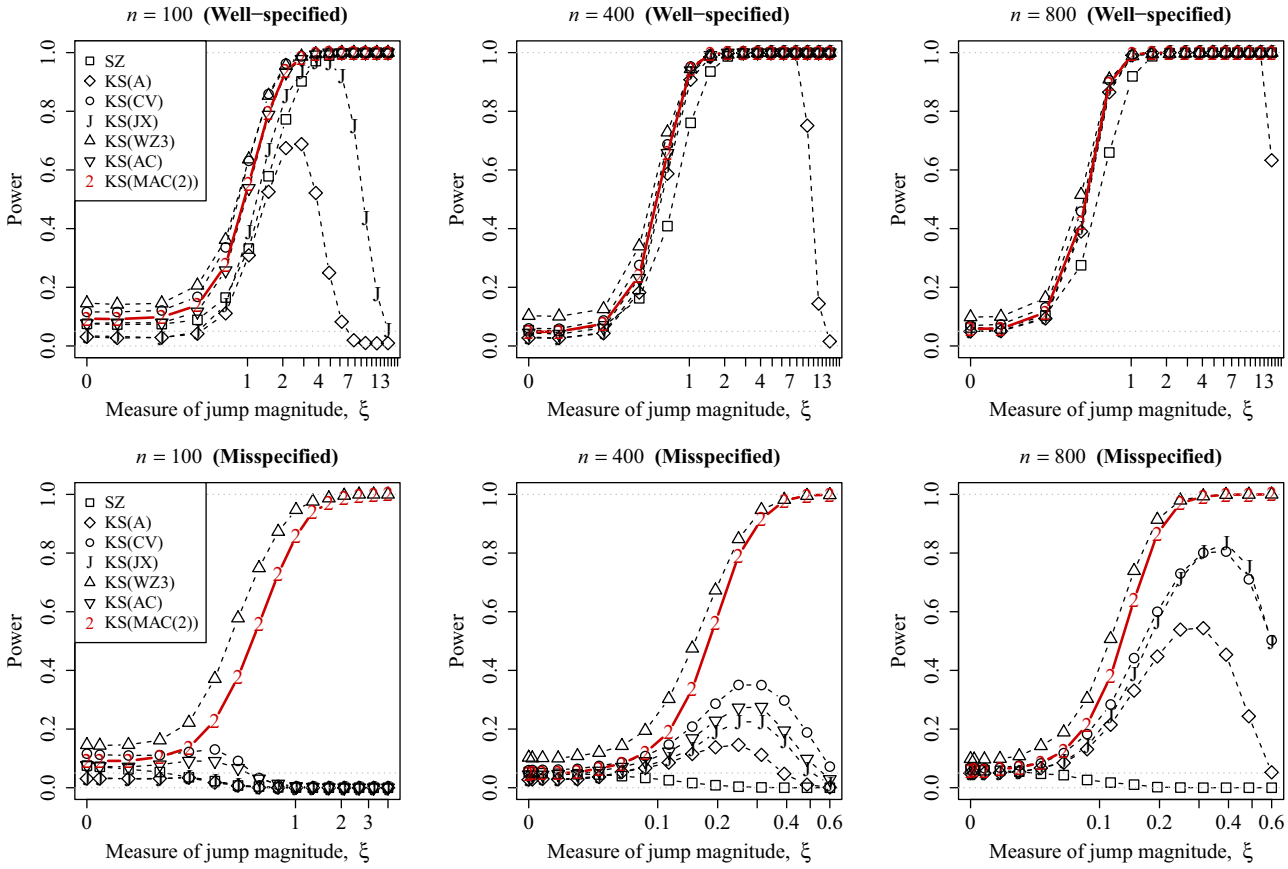

Figure 6. The powers of the CP tests defined in Section 5.4 are plotted against the jump magnitude ξ under Model B2. The scenarios under well-specified alternative H_1 and misspecified alternative H_1' are shown in the upper and lower plots, respectively. Dashed horizontal lines indicate the significance level $\alpha = 5\%$ and zero. Note that horizontal axis is plotted in the logarithmic scale for better visualization.

directly (see Vogelsang 1999 for a similar analysis). Using the CP test proposed by Wu and Zhao (2007), we can test H_0: "The mean function $i \mapsto EY_i$ is continuous" against H_1: "The mean function $i \mapsto EY_i$ has a jump-discontinuity." The test statistic is $Q_n := (k_n\widehat{\sigma})^{-1} \max_{k_n \le i \le n-k_n} \left| \sum_{j=i+1}^{k_n+i} Y_j - \sum_{j=i-k_n+1}^{i} Y_j \right|$, where $\widehat{\sigma}^2$ is a consistent estimator of σ^2, that is, the AVC of $\{Y_i\}$; and $k_n = \lfloor n^{0.6} \rfloor$. Then H_0 is rejected if Q_n is large. Using MAC(2) to estimate σ, we obtain $\widehat{\sigma} = 0.0517$; and found that H_0 is rejected at any reasonable level. It is remarked that σ is estimated to be 0.0434 by using the estimator WZ3. Although this estimate is a bit smaller than our proposed estimate, the same conclusion for testing H_0 is obtained if this estimate is used in the test statistic Q_n. Although Wu and Zhao (2007) did not provide any estimator of the CP, they argue that if $i + 1$ is a discontinuity point, then the difference of the averages inside the statistic Q_n should be large. Following their idea, $\widehat{D}_{\mathrm{WZ}} := 1 + \arg\max_{k_n \le i \le n-k_n} \left| \sum_{j=i+1}^{k_n+i} Y_j - \sum_{j=i-k_n+1}^{i} Y_j \right|$ is a reasonable estimator of the CP. The estimated CP, $\widehat{D}_{\mathrm{WZ}}$, is 7 October 2008 (see Figure 7). It indicates the 2008 financial crisis quite accurately. It coincides with our understanding of the stock market.

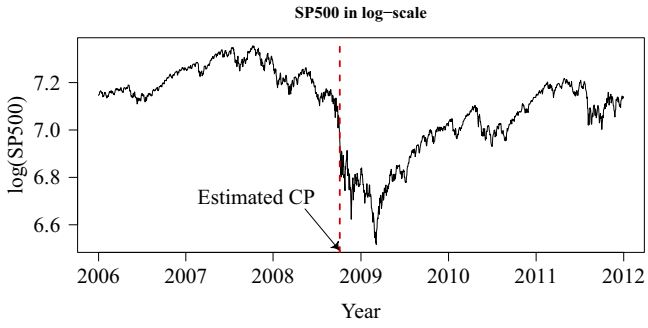

Figure 7. Time series plot of $\{Y_i\}$, that is, the daily S&P 500 Index (3 January 2006–30 December 2011) in the log scale (see Section 6.1). The vertical dotted line indicates the value of $\widehat{D}_{\mathrm{WZ}}$ estimated by the statistic in Wu and Zhao (2007). Here σ^2 is estimated by MAC(2).

6.2. Simultaneous Change Point Detection in Several Indices

Besides S&P 500 Index mentioned in Section 6.1, there are several other stock market indices that are commonly used by traders, for example, Dow Jones Index, Nasdaq Composite, and Russell 2000. In this subsection, we investigate whether we can make use of these four market indices simultaneously to make a more precise detection of the 2008 financial crisis.

Consider the squared daily returns, which can be used as proxies for daily volatilities, of the aforementioned four indices in the period 1 July 2008–30 December 2008 (see Figure 8). Applying the CUSUM-type CP estimator $\widehat{D}_1$ (see (1) of the supplementary materials for its formula) to each index individually, we obtain the same CP 29 September 2018. It is remarked that the no CP null hypothesis is rejected at 5% level by the test KS(MAC(2)) for each individual index.

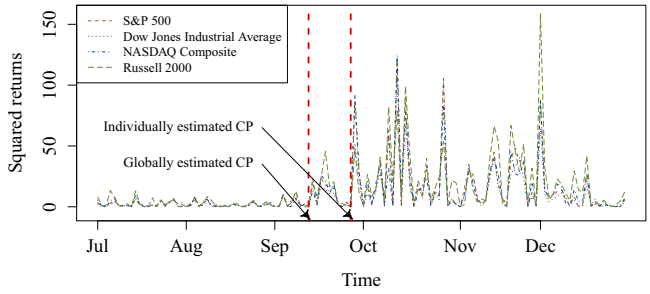

Figure 8. The squared returns of four stock indices (1 July 2008–30 December 2008) (see Section 6.2). The two vertical lines denote the CP locations. The earlier and later CPs are detected by the multivariate and univariate CUSUM CP estimators, respectively.

Since these stock market indices are highly correlated and are believed to follow the market trend very closely, a CP (if any) is likely to appear simultaneously. Hence, using multivariate time series for detecting a CP can be more accurate and precise. Applying the multivariate version of the KS CP test (Horváth, Kokoszka, and Steinebach 1999) to the four indices, we detect the CP to be 15 September 2008. From Figure 8, the squared returns between 15 and 29 September are slightly higher than the first portion of the series. Hence, using multivariate time series helps detecting these small changes. Consequently, multivariate tests are potentially more useful in practice. It is also remarked that the no simultaneous CP alternative is rejected at 5% level by the CP test (Horváth, Kokoszka, and Steinebach 1999) with our proposed MAC(2) estimator.

7. Conclusions

In this article, we propose an estimator of the ACM in nonstationary time series. The estimator has several desirable features: (i) it is *robust* against unknown trends and a divergent number of jumps; (ii) it is *optimal* in the sense that an asymptotically correct optimal bandwidth can be implemented robustly; (iii) it is *statistically efficient* since it has the optimal $\mathcal{L}^2$ convergence rate for different strength of serial dependence; (iv) it is *computationally fast* because neither numerical optimization, trend estimation, nor CPs detection is required; and (v) it is *handy* because its formula can be as simple as (24).

Some applications of the estimator are illustrated. In particular, we found that the CP test equipped with the proposed estimator is the only available test which is monotonically powerful and insensitive to a misspecified alternative hypothesis.

Supplementary Materials

Supplementary materials include graphical illustration, additional simulation results, and proofs. The R-package MAC for computing the proposed estimator is also provided.

Acknowledgments

The author thanks the editor Christian Hansen, the associate editor, and two reviewers for their detailed and insightful comments. The author also gratefully thanks Neil Shephard for his helpful advice on improving the estimator as well as Xiao-Li Meng, Jim Stock and Pierre Jacob for fruitful discussions.

Funding

This research was supported by the Direct Grant (4053356) provided by the Chinese University of Hong Kong, and the Early Career Scheme (24306919) provided by the University Grant Committee of HKSAR.

References

Altissimoa, F., and Corradic, V. (2003), "Strong Rules for Detecting the Number of Breaks in a Time Series," *Journal of Econometrics*, 117, 207–244. [3,9,12]

Anderson, T. W. (1971), *The Statistical Analysis of Time Series*, New York: Wiley. [4]

Andrews, D. W. K. (1991), "Heteroskedasticity and Autocorrelation Consistent Covariance Matrix Estimation," *Econometrica*, 59, 817–858. [1,2,3,4,6,7,8,9,11]

Banerjeea, A., and Urga, G. (2005), "Modelling Structural Breaks, Long Memory and Stock Market Volatility: An Overview," *Journal of Econometrics*, 19, 1–34. [1]

Barndorff-Nielsen, O. E., and Shephard, N. (2004), "Power and Bipower Variation With Stochastic Volatility and Jumps," *Journal of Financial Econometrics*, 2, 1–37. [4]

Brockwell, P. J., and Davis, R. A. (1991), *Time Series: Theory and Methods*, New York: Springer. [5]

Brown, R. L., Durbin, J., and Evans, J. M. (1975), "Techniques for Testing the Constancy of Regression Relationships Over Time," *Journal of the Royal Statistical Society*, Series B, 37, 149–192. [1]

Carlstein, E. (1986), "The Use of Subseries Values for Estimating the Variance of a General Statistic From a Stationary Sequence," *The Annals of Statistics*, 14, 1171–1179. [2]

Chan, K. W., and Yau, C. Y. (2016), "New Recursive Estimators of the Time-Average Variance Constant," *Statistics and Computing*, 26, 609–627. [3]

——— (2017a), "Automatic Optimal Batch Size Selection for Recursive Estimators of Time-Average Covariance Matrix," *Journal of the American Statistical Association*, 112, 1076–1089. [3,5,6]

——— (2017b), "High Order Corrected Estimator of Asymptotic Variance With Optimal Bandwidth," *Scandinavian Journal of Statistics*, 44, 866–898. [2,5,6]

Crainiceanu, C. M., and Vogelsang, T. J. (2007), "Nonmonotonic Power for Tests of a Mean Shift in a Time Series," *Journal of Statistical Computation and Simulation*, 77, 457–476. [1,3,9]

Csörgö, M., and Horváth, L. (1997), *Limit Theorems in Change-Point Analysis*, New York: Wiley. [1]

Davies, R. B., and Harte, D. S. (1987), "Tests for Hurst Effect," *Biometrika*, 74, 95–101. [5]

Degras, D., Xu, Z., Zhang, T., and Wu, W. B. (2012), "Testing for Parallelism Among Trends in Multiple Time Series," *IEEE Transactions on Signal Processing*, 60, 1087–1097. [2]

Dette, H., Munk, A., and Wagner, T. (1998), "Estimating the Variance in Nonparametric Regression—What Is a Reasonable Choice?," *Journal of the Royal Statistical Society*, Series B, 60, 751–764. [4]

Flegal, J. M., and Jones, G. L. (2010), "Batch Means and Spectral Variance Estimation in Markov Chain Monte Carlo," *The Annals of Statistics*, 38, 1034–1070. [3]

Gallant, A. R., and White, H. (1988), *A Unified Theory of Estimation and Inference for Nonlinear Dynamic Models*, New York: Basil Blackwell. [3]

Gonçalves, S., and White, H. (2002), "The Bootstrap of the Mean for Dependent Heterogeneous Arrays," *Econometric Theory*, 18, 1367–1384. [3]

Grangera, C. W. J., and Hyung, N. (2004), "Occasional Structural Breaks and Long Memory With an Application to the S&P 500 Absolute Stock Returns," *Journal of Empirical Finance*, 11, 399–421. [1]

Hall, P., and Horowitz, J. (2013), "A Simple Bootstrap Method for Constructing Nonparametric Confidence Bands for Functions," *The Annals of Statistics*, 41, 1892–1921. [4]

Hall, P., Kay, J. W., and Titterinton, D. M. (1990), "Asymptotically Optimal Difference-Based Estimation of Variance in Nonparametric Regression," *Biometrika*, 77, 521–528. [4]

Hirukawa, M. (2010), "A Two-Stage Plug-In Bandwidth Selection and Its Implementation for Covariance Estimation," *Econometric Theory*, 26, 710–743. [3]

Horváth, L., Kokoszka, P., and Steinebach, J. (1999), "Testing for Changes in Multivariate Dependent Observations With an Application to Temperature Changes," *Journal of Multivariate Analysis*, 68, 96–119. [1,14]

Jirak, M. (2015), "Uniform Change Point Tests in High Dimension," *The Annals of Statistics*, 43, 2451–2483. [1,3]

Juhl, T., and Xiao, Z. (2009), "Tests for Changing Mean With Monotonic Power," *Journal of Econometrics*, 148, 12–24. [1,3,12]

Kirch, C., Muhsal, B., and Ombao, H. (2015), "Detection of Changes in Multivariate Time Series With Application to EEG Data," *Journal of the American Statistical Association*, 110, 1197–1216. [1]

Künsch, H. R. (1989), "The Jackknife and the Bootstrap for General Stationary Observations," *The Annals of Statistics*, 17, 1217–1241. [2,3]

Lahiri, S. N. (2003), *Resampling Methods for Dependent Data*, New York: Springer. [2]

Lazarus, E., Lewis, D. J., Stock, J. H., and Watson, M. W. (2018), "HAR Inference: Recommendations for Practice," *Journal of Business & Economic Statistics*, 36, 541–559. [2,4]

Liu, W., and Wu, W. B. (2010), "Asymptomatic of Spectral Density Estimates," *Econometric Theory*, 26, 1218–1245. [5]

Liu, Y., and Flegal, J. M. (2018), "Weighted Batch Means Estimators in Markov Chain Monte Carlo," *Electronic Journal of Statistics*, 12, 3397–3442. [3]

Meketon, M. S., and Schmeiser, B. (1984), "Overlapping Batch Means: Something for Nothing?," in *Proceedings of the 16th Conference on Winter Simulation*, pp. 226–230. [2]

Mikkonen, S., Laine, M., Mäkelä, H. M., Gregow, H., Tuomenvirta, H., Lahtinen, M., and Laaksonen, A. (2014), "Trends in the Average Temperature in Finland, 1847–2013," *Stochastic Environmental Research and Risk Assessment*, 29, 1521–1529. [1]

Müller, U. K. (2014), "HAC Corrections for Strongly Autocorrelated Time Series," *Journal of Business & Economic Statistics*, 32, 311–322. [2]

Newey, W. K., and West, K. D. (1987), "A Simple, Positive Semi-Definite, Heteroskedasticity and Autocorrelation Consistent Covariance Matrix," *Econometrica*, 55, 703–708. [2]

——— (1994), "Automatic Lag Selection in Covariance Matrix Estimation," *The Review of Economic Studies*, 61, 631–653. [1]

Paparoditis, E., and Politis, D. N. (2001), "Tapered Block Bootstrap," *Biometrika*, 88, 1105–1119. [2,8]

Phillips, P. C. B. (2005), "HAC Estimation by Automated Regression," *Econometric Theory*, 21, 116–142. [2]

Ploberger, W., and Krämer, W. (1992), "The CUSUM Test With OLS Residuals," *Econometrica*, 60, 271–285. [1]

Politis, D. N. (2003), "Adaptive Bandwidth Choice," *Journal of Nonparametric Statistics*, 15, 517–533. [3]

——— (2011), "Higher-Order Accurate, Positive Semidefinite Estimation of Large-Sample Covariance and Spectral Density Matrices," *Econometric Theory*, 27, 703–744. [2,5]

Politis, D. N., and Romano, J. P. (1994), "The Stationary Bootstrap," *Journal of the American Statistical Association*, 89, 1303–1313. [2,3]

Politis, D. N., Romano, J. P., and Wolf, M. (1999), *Subsampling*, New York: Springer. [2,3,5]

Rosenblatt, M. (1985), *Stationary Sequences and Random Fields*, Boston: Birkhäuser. [5]

Shao, X., and Zhang, X. (2010), "Testing for Change Points in Time Series," *Journal of the American Statistical Association*, 105, 1228–1240. [1,12]

Song, W. T., and Schmeiser, B. W. (1995), "Optimal Mean-Squared-Error Batch Sizes," *Management Science*, 41, 110–123. [2]

Sun, Y. (2013), "Heteroscedasticity and Autocorrelation Robust *F* Test Using Orthonormal Series Variance Estimator," *Econometrics Journal*, 16, 1–26. [2]

Vogelsang, T. J. (1999), "Sources of Nonmonotonic Power When Testing for a Shift in Mean of a Dynamic Time Series," *Journal of Econometrics*, 88, 283–299. [1,13]

Wu, W. B. (2004), "A Test for Detecting Changes in Mean," in *Time Series Analysis and Applications to Geophysical Systems*, eds. D. R. Brillinger, E. A. Robinson, and F. Schoenberg (Vol. 139), New York: Springer-Verlag, pp. 105–122. [1,3,8]

——— (2005), "Nonlinear System Theory: Another Look at Dependence," *Proceedings of the National Academy of Sciences of the United States of America*, 102, 14150–14154. [2,5,6,12]

——— (2007), "Strong Invariance Principles for Dependent Random Variables," *The Annals of Probability*, 35, 2294–2320. [5]

——— (2011), "Asymptotic Theory for Stationary Processes," *Statistics and Its Interface*, 4, 207–226. [2,5,12]

Wu, W. B., Woodroofe, M., and Mentz, G. (2001), "Isotonic Regression: Another Look at the Change Point Problem," *Biometrika*, 88, 793–804. [1,8]

Wu, W. B., and Zaffaroni, P. (2018), "Asymptotic Theory for Spectral Density Estimates of General Multivariate Time Series," *Econometric Theory*, 34, 1–22. [2]

Wu, W. B., and Zhao, Z. (2007), "Inference of Trends in Time Series," *Journal of the Royal Statistical Society*, Series B, 69, 391–410. [1,3,8,9,13]