The Annals of Statistics
2022, Vol. 50, No. 3, 1376–1400
https://doi.org/10.1214/21-AOS2154

OPTIMAL DIFFERENCE-BASED VARIANCE ESTIMATORS IN TIME SERIES: A GENERAL FRAMEWORK

BY KIN WAI CHAN[a]

Department of Statistics, The Chinese University of Hong Kong, [a]*kinwaichan@cuhk.edu.hk*

Variance estimation is important for statistical inference. It becomes nontrivial when observations are masked by serial dependence structures and time-varying mean structures. Existing methods either ignore or suboptimally handle these nuisance structures. This paper develops a general framework for the estimation of the long-run variance for time series with nonconstant means. The building blocks are difference statistics. The proposed class of estimators is general enough to cover many existing estimators. Necessary and sufficient conditions for consistency are investigated. The first asymptotically optimal estimator is derived. Our proposed estimator is theoretically proven to be invariant to arbitrary mean structures, which may include trends and a possibly divergent number of discontinuities.

1. Introduction.

1.1. *Motivation and background.* Let the observed time series $X_{1:n} = \{X_1, \ldots, X_n\}$ be generated from the signal-plus-noise model:

$$X_i = \mu_i + Z_i, \quad i = 1, \ldots, n, \tag{1.1}$$

where the deterministic signals μ_i and the zero-mean stationary noises Z_i are not directly observable. Many statistics designed for inferring $\mu_{1:n} = \{\mu_1, \ldots, \mu_n\}$ admit the form $T_n = T_n(\widehat{v})$, where $\widehat{v}$ is an estimator of the long-run variance (LRV) $v = \lim_{n\to\infty} n \operatorname{Var}(\bar{Z}_n)$ of $\bar{Z}_n = \sum_{i=1}^n Z_i/n$. Deriving a good estimator $\widehat{v}$ is, therefore, important, and is the major goal of this article.

Examples of such $T_n(\widehat{v})$ include, but are not restricted to, the Kolmogorov–Smirnoff (KS) change point test and its variants (Crainiceanu and Vogelsang (2007), Górecki, Horváth and Kokoszka (2018), Horváth, Kokoszka and Steinebach (1999), Juhl and Xiao (2009)), mean constancy tests (Dalla, Giraitis and Phillips (2015), Wu (2004)), mass excess tests of relevant mean changes (Dette and Wu (2019)), tests for monotone trends (Wu, Woodroofe and Mentz (2001)), simultaneous confidence bands (SCBs) for trends (Wu and Zhao (2007)), etc. Serving as a normalizer in $T_n(\widehat{v})$, the estimator $\widehat{v}$ measures the significance of the signals μ_i relative to the noises Z_i. Constructing a good $\widehat{v}$ is nevertheless difficult due to two nuisance structures.

1. *Nuisance structure 1: variability of* $\mu_{1:n}$. The stochastic variability of $Z_{1:n} = \{Z_1, \ldots, Z_n\}$ is masked by the deterministic variability of $\mu_{1:n}$; see Figure 1. Disentangling the variabilities of $\mu_{1:n}$ and $Z_{1:n}$ can be challenging. Without the nuisance structure 2 below, this task was studied by, for example, Hall, Kay and Titterington (1990). Similar and extended results include Anderson (1971), Rice (1984) and Levine and Tecuapetla-Gómez (2019).

Received January 2021; revised November 2021.
MSC2020 subject classifications. Primary 62G05; secondary 62G20.
Key words and phrases. Change point detection, nonlinear time series, optimal bandwidth selection, trend inference, variate difference method.

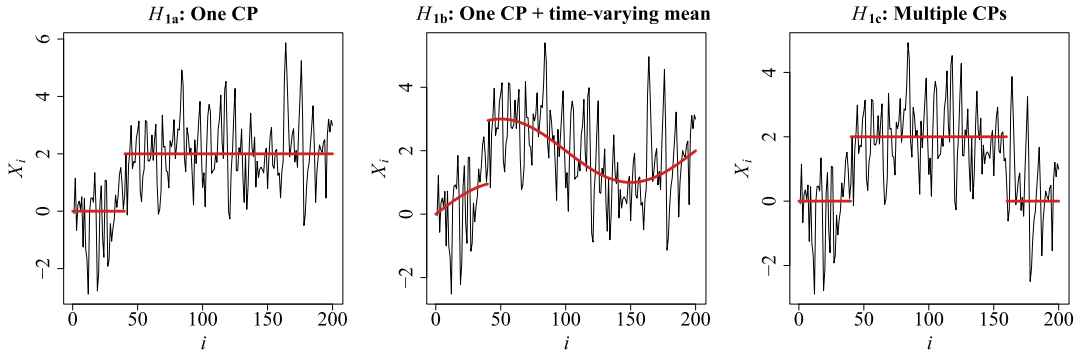

FIG. 1. *The black and red lines denote $X_{1:n}$ and $\mu_{1:n}$, respectively. Here, H_{1a}: $\mu_i = \Xi\mathbb{1}(i > 2n/10)$; H_{1b}: $\mu_i = \Xi\{\mathbb{1}(i > 2n/10) + \sin(2\pi i/n)/2\}$; H_{1c}: $\mu_i = \Xi\mathbb{1}(2n/10 < i < 8n/10)$, where $\Xi \in \mathbb{R}$ measures the magnitude of jump(s) and/or the amplitude of trend. The noises $Z_{1:n}$ are generated from an autoregressive AR(2) model: $Z_i = Z_{i-1}/2 + Z_{i-2}/5 + \varepsilon_i$ for each i, where ε_i follow* $\mathrm{N}(0, 1)$ *independently. In particular, $n = 200$ and $\Xi = 2$ are used in the above plots.*

2. *Nuisance structure 2: serial dependence of $Z_{1:n}$.* Under regularity conditions, $v = \sum_{k\in\mathbb{Z}} \gamma_k$ is a sum of infinitely many unknowns, where $\gamma_k = \mathsf{E}(Z_0 Z_k)$ is the autocovariance function (ACVF). So, estimation of v is hard. Without the nuisance structure 1 above, this task was studied by, for example, Chen and Schmeiser (2013), Carlstein (1986), Newey and West (1987), Künsch (1989), Andrews (1991), Politis (2011), Yau and Chan (2016) and Chan and Yau (2017a).

In this article, we propose a *general* framework of estimators of v; see Definition 1 and equation (2.3). Necessary and sufficient conditions for consistency are derived; see Theorems 4.1 and 4.2. They are proven to achieve the *optimal* $\mathcal{L}^2$ rate of convergence under various strengths of serial dependence (see Theorems 5.1 and 5.2) and are *robust* against a wide class of mean structures (see Theorem 3.1). The optimal mth order difference-based variance estimator $\widehat{v}_{(m)}$ is given in Corollary 5.3, where the optimality refers to the best possible difference statistics used in the estimator. In particular, one special case (with $m = 3$) is given by

$$\widehat{v}_{(3)} = \sum_{|k|<\ell} \left\{1 - \left(\frac{|k|}{\ell}\right)^2\right\} \frac{1}{n} \sum_{i=mh+|k|+1}^{n} D_i D_{i-|k|},$$

where $D_i = 0.1942X_i + 0.2809X_{i-h} + 0.3832X_{i-2h} - 0.8582X_{i-3h}$ for each i, $\ell = O(n^{1/5})$, and $h = 2\ell$. The proposed estimator with the optimally selected ℓ is presented in (6.3). This estimator outperforms all existing estimators in terms of the mean-squared error (MSE) asymptotically; see (5.4). We conclude this subsection with an example to illustrate the importance of this project.

EXAMPLE 1.1 (Change point detection). Suppose we want to test $H_0 : \mu_1 = \cdots = \mu_n$. The celebrated KS change point (CP) test statistic (see, e.g., Csörgő and Horváth (1997)) is defined as

$$T_n(v) = \frac{1}{\sqrt{nv}} \max_{k\in\{1,\ldots,n\}} \left|\sum_{i=1}^{k} (X_i - \bar{X}_n)\right|, \quad \text{where } \bar{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i. \tag{1.2}$$

We reject H_0 at size 5% if $T_n(v) > 1.358$. Although this test is designed for a one-CP alternative (H_{1a}), it is still applicable to more complicated situations, for example, a one-CP alternative in the presence of a smooth trend (H_{1b}), and a multiple-CP alternative (H_{1c}); see Figure 1. No matter which situation we consider, having a good estimator of v is still necessary. We compare two estimators: the classical Bartlett kernel estimator $\widehat{v}_{(\mathrm{A})}$ with a bandwidth

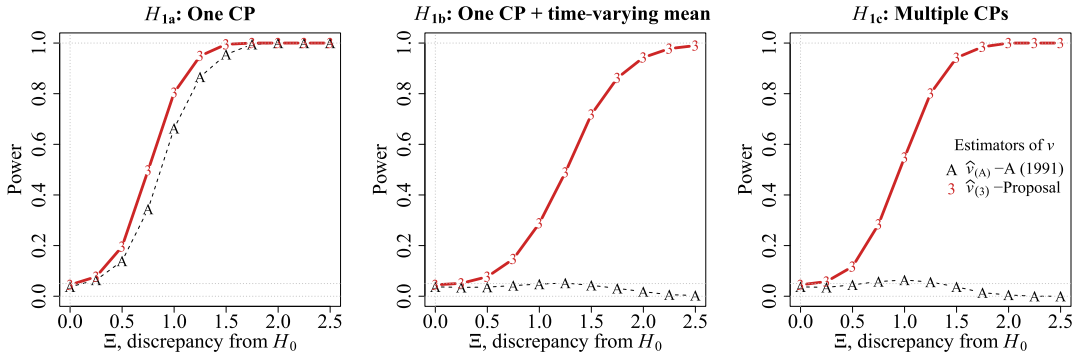

FIG. 2. *Power curves of the KS tests with the classical estimator* $\widehat{v}_{(A)}$ *and the proposed estimator* $\widehat{v}_{(3)}$ *in three different types of alternative hypotheses stated in Figure* 1.

selected by fitting an AR(1) model proposed in Andrews (1991), and our proposed estimator $\widehat{v}_{(3)}$ with optimally selected parameters to be discussed in (6.3).

Consider the time series defined in Figure 1 with different magnitudes of jump Ξ. We compute the power of the classical test $T_n(\widehat{v}_{(A)})$ and the proposed test $T_n(\widehat{v}_{(3)})$ against Ξ; see Figure 2. Under H_{1a}, the test $T_n(\widehat{v}_{(A)})$ is valid but less powerful than $T_n(\widehat{v}_{(3)})$ because $\widehat{v}_{(A)}$ is inaccurate for v when $\Xi \neq 0$. Under H_{1b} or H_{1c}, the test $T_n(\widehat{v}_{(A)})$ even fails to demonstrate monotone power when Ξ increases because $\widehat{v}_{(A)} \to \infty$ in probability as $\Xi \to \infty$ in these cases. So a robust and efficient estimator of v is crucial.

1.2. *Notation and mathematical background.* Let $\mu_i = \mu(i/n)$ for $i = 1, \ldots, n$, where $\mu : [0, 1] \to \mathbb{R}$ is a mean function. Suppose that $\mu(\cdot)$ consists of a continuous part $c(\cdot)$ and a step-discontinuous part $s(\cdot)$ such that

$$\mu(t) = c(t) + s(t), \quad s(t) = \sum_{j=0}^{\mathcal{J}} \xi_j \mathbb{1}(T_j/n \le t < T_{j+1}/n), \tag{1.3}$$

where $\mathcal{J}$ is the number of discontinuities, $1 \equiv T_0 < T_1 < \cdots < T_{\mathcal{J}} < T_{\mathcal{J}+1} \equiv n + 1$ are the times of discontinuities, and $\xi_0, \ldots, \xi_{\mathcal{J}}$ are the step sizes such that $\xi_j \neq \xi_{j-1}$ for each j. Note that $c(\cdot)$, $\mathcal{J}$, $\xi_0, \ldots, \xi_{\mathcal{J}}$, and $T_1, \ldots, T_{\mathcal{J}}$ are possibly dependent on n. For example, $\mathcal{J}$ and $\xi_0, \ldots, \xi_{\mathcal{J}}$ can be divergent with n. Denote the minimal gap between two consecutive CP times by

$$\mathcal{G} = \min_{0 \le j \le \mathcal{J}} (T_{j+1} - T_j).$$

We measure the smoothness of $c(\cdot)$ by $\mathcal{C}$, the maximum step magnitude of $s(\cdot)$ by $\mathcal{S}$, and the overall variability of $\mu(\cdot)$ by $\mathcal{V}$, where

$$\mathcal{C} = \sup_{0 \le t' < t \le 1} \left| \frac{c(t) - c(t')}{t - t'} \right|, \qquad \mathcal{S} = \sup_{1 \le j \le \mathcal{J}} |\xi_j - \xi_{j-1}|, \qquad \mathcal{V} = \int_0^1 \{\mu(t) - \bar{\mu}\}^2 \, \mathrm{d}t,$$

and $\bar{\mu} = \int_0^1 \mu(t) \, \mathrm{d}t$. Clearly, $\mathcal{C} = 0$ iff there is no trend effect; $\mathcal{S} = 0$ or $\mathcal{J} = 0$ iff there is no discontinuity; and $\mathcal{V} = 0$ iff the mean function is a constant.

Let $Z_i = g(\mathcal{F}_i)$ for some measurable function g, where $\mathcal{F}_i = (\ldots, \varepsilon_{i-1}, \varepsilon_i)$ and $\{\varepsilon_i\}_{i \in \mathbb{Z}}$ are independent and identically distributed (i.i.d.) innovations. Let ε'_j be an i.i.d. copy of ε_j, $\mathcal{F}_{i,\{j\}} = (\mathcal{F}_{j-1}, \varepsilon'_j, \varepsilon_{j+1}, \ldots, \varepsilon_i)$, and $Z_{i,\{j\}} = g(\mathcal{F}_{i,\{j\}})$. Define $\mathcal{P}_i \cdot = \mathsf{E}(\cdot \mid \mathcal{F}_i) - \mathsf{E}(\cdot \mid \mathcal{F}_{i-1})$. For $p \ge 1$, define the *physical dependence measure* and its aggregated value by

$$\theta_{p,i} = \|Z_i - Z_{i,\{0\}}\|_p \quad \text{and} \quad \Theta_p = \sum_{i=0}^{\infty} \theta_{p,i}, \tag{1.4}$$

respectively, where $\|\cdot\|_p = (\mathsf{E}|\cdot|^p)^{1/p}$. The finiteness of Θ_p provides a mild and easily verifiable condition for asymptotic theory; see Wu (2005, 2007, 2011).

ASSUMPTION 1 (Weak dependence). The noise sequence $\{Z_i\}_{i\in\mathbb{Z}}$ is a zero-mean strictly stationary time series that satisfies $\mathsf{E}(Z_1^r) < \infty$ for some $r > 4$, and $\Theta_4 < \infty$.

Indeed, Assumption 1 implies that the ACVFs are absolutely summable, that is, $u_0 := \sum_{k\in\mathbb{Z}} |\gamma_k| < \infty$, which ensures the existence of $v = \sum_{k\in\mathbb{Z}} \gamma_k$. We remark that there exist other ways of quantifying dependence, including various types of mixing coefficients (Rosenblatt (1956), Volkonskiĭ and Rozanov (1959)) and near epoch approach (Ibragimov (1962)). They have been widely adopted and studied; see Taqqu and Eberlein (1986) and Bradley (2005) for some surveys of results. It is certainly interesting to develop our theoretical results under these settings, however, it is beyond the scope of this paper. We leave it for further study.

The following notation is used. Let $\mathbb{N} = \{1, 2, 3, \ldots\}$, $\mathbb{N}_0 = \{0, 1, 2, \ldots\}$, and $\mathbb{R}^+ = (0, \infty)$. For any statement E, $\mathbb{1}_{(E)} = 1$ if E is true, otherwise $\mathbb{1}_{(E)} = 0$. For any $a, b \in \mathbb{R}$, $a^+ = \max(a, 0)$ and $a \wedge b = \min(a, b)$. For any $\{a_n\}_{n\in\mathbb{N}}$ and $\{b_n\}_{n\in\mathbb{N}}$ with $a_n, b_n \in \mathbb{R}^+$, the relation $a_n \sim b_n$ means $a_n/b_n \to 1$; $a_n \asymp b_n$ means there is $C \in \mathbb{R}^+$ such that $1/C \le a_n/b_n \le C$ for all large n; $a_n \ll b_n$ or $a_n = o(b_n)$ means $a_n/b_n \to 0$; $a_n \lesssim b_n$ or $a_n = O(b_n)$ means there is $C > 0$ such that $a_n/b_n \le C$ for all large n. Convergence in probability and convergence in distribution are denoted by $\xrightarrow{\text{pr}}$ and $\Rightarrow$, respectively. Write $\|\cdot\| = \|\cdot\|_2$. For any sequence of random variables $\{Z_n\}_{n\in\mathbb{N}}$, $Z_n = O_p(a_n)$ means for any $\epsilon > 0$ there exist $C \in \mathbb{R}^+$ and $N \in \mathbb{N}$ such that $\mathsf{P}(|Z_n/a_n| > C) < \epsilon$ for all $n > N$; $Z_n = o_p(a_n)$ means $Z_n/a_n \xrightarrow{\text{pr}} 0$. For any estimator $\widehat{\theta}$ of θ, denote $\mathsf{Bias}(\widehat{\theta}; \theta) = \mathsf{Bias}(\widehat{\theta}) = \mathsf{E}(\widehat{\theta}) - \theta$ and $\mathsf{MSE}(\widehat{\theta}; \theta) = \mathsf{MSE}(\widehat{\theta}) = \mathsf{E}(\widehat{\theta} - \theta)^2$.

In this article, we propose and study a general framework for estimating

$$v = \lim_{n\to\infty} n\,\mathsf{Var}(\bar{Z}_n) = \sum_{k\in\mathbb{Z}} \gamma_k, \quad \text{where } \gamma_k = \mathsf{Cov}(Z_0, Z_k), \tag{1.5}$$

by using difference statistics. This article is structured as follows. Section 2 defines the proposed class of estimators. We show that it covers many existing estimators as special cases. Section 3 demonstrates its invariance to mean structures. Section 4 derives the necessary and sufficient conditions for consistency. Section 5 shows that the proposed estimator is asymptotically optimal. Section 6 addresses implementation issues and generalization. Section 7 presents simulation experiments, applications and real-data examples. We conclude the paper with a summary of major contributions and possible future work in Section 8. All proofs are deferred to a separate supplementary note (Chan (2022b)). An R-package `"dlrv"` is available on the author's website.

2. A general framework for variance estimation.

2.1. *Difference-based statistics.* Variance estimators usually require *centering* to achieve mean invariance. For example, if $\mu_1 = \cdots = \mu_n$, one may *globally* center the data as $D_i' = X_i - \bar{X}_n$; if $\mu_{1:n}$ are not constant, one may *locally* center each X_i by the kernel method and the lag-1 difference:

$$D_i'' = X_i - \frac{\sum_j H(\frac{i-j}{m/2}) X_j}{\sum_{j'} H(\frac{i-j'}{m/2})} \quad \text{and} \quad D_i''' = X_i - X_{i-1}, \tag{2.1}$$

where $H(\cdot)$ is a kernel, and $m/2$ is a bandwidth. The statistics D_i', D_i'' and D_i''' are special cases of the following class of general difference statistics.

DEFINITION 1 (Difference statistics). A real-valued sequence $\{d_j\}_{j=0}^m$ is said to be an mth order *difference sequence* if $d_0 + \cdots + d_m = 0$. If, in addition, $\delta_0 = d_0^2 + \cdots + d_m^2 = 1$, then $\{d_j\}$ is said to be *normalized*. For $h \in \mathbb{N}$, the mth order lag-h *difference statistics* are defined as

$$D_i = \sum_{j=0}^{m} d_j X_{i-jh}, \quad i = mh+1, \ldots, n. \tag{2.2}$$

The zeroth-order difference statistics are $D_i = X_i - \bar{X}_n$ for $i = 1, \ldots, n$. Also denote $\delta_s = \sum_{j=|s|}^m d_j d_{j-|s|}$ for $|s| \le m$ and $\delta_s = 0$ for $|s| > m$.

The condition $\sum_{j=0}^m d_j = 0$ is used to ensure that $\mathsf{E}(D_i) \approx 0$ when $\mu_i \approx \mu_{i-h} \approx \cdots \approx \mu_{i-mh}$. This property is important for deriving asymptotic mean invariance of statistics based on D_i; see Section 3 for a precise and rigorous definition of mean invariance. The requirement $\delta_0 = 1$ is used to regularize D_i such that $\mathsf{Var}(D_i) = \mathsf{Var}(X_i)$ when $X_{1:n}$ are serially uncorrelated. One can easily normalize d_j by $d_j/\sqrt{\delta_0}$ provided that $\delta_0 \neq 0$. From now on, we assume the difference sequence $\{d_j\}$ is normalized. The lag parameter h is used to control how frequent the observations are used for constructing one difference statistic. When the data are independent, $h = 1$ works well. When the data are serially dependent, a larger h can be used to reduce the serial dependence among the observations that are used in the same difference statistic. Some difference sequences are shown in Example 2.1.

EXAMPLE 2.1. Some commonly used difference sequences $\{d_j\}_{j=0}^m$ are listed below:

- Binomial differencing: $d_j = \binom{m}{j}(-1)^j / \binom{2m}{m}^{1/2}$ for $j = 0, \ldots, m$. It gives $\delta_k = (-1)^k \times (m!)^2/\{(m+k)!(m-k)!\}$ for $k = 0, 1, \ldots, m$.
- Local differencing: $d_0 = \sqrt{m/(m+1)}$ and $d_j = -1/\sqrt{m^2+m}$ for $j = 1, \ldots, m$. It gives $\delta_0 = 1$ and $\delta_k = -k/(m^2+m)$ for $k = 1, \ldots, m$.
- Hall, Kay and Titterington (1990): Define $\{d_j\}_{j=0}^m$ by minimizing $\sum_{k=1}^m \delta_k^2$; see Table 1 for the solution. It gives $\delta_0 = 1$ and $\delta_k = -1/(2m)$ for $k = 1, \ldots, m$.

Note that $m = m_n$ is allowed to diverge with n. In this case, we need the following assumption to regularize the difference sequence.

ASSUMPTION 2. The difference sequence $\{d_j\}$ satisfies (i) $\sup_{n\in\mathbb{N}} \sum_{j=0}^m |d_j| < \infty$, and (ii) $\sup_{n\in\mathbb{N}} \sum_{|s|\le m} |\delta_s| < \infty$.

2.2. *Proposed difference-based variance estimator.* Since $D_{mh+1}, \ldots, D_n$ are approximately centered at zero, it motivates us to utilize them as building blocks for estimating v.

TABLE 1
Hall, Kay and Titterington's (1990) difference sequence $\{d_j\}_{j=0}^m$ *for* $m = 1, \ldots, 4$

m	d_0	d_1	d_2	d_3	d_4
1	0.7071	−0.7071	–	–	-
2	0.8090	−0.5000	−0.3090	–	–
3	0.1942	0.2809	0.3832	−0.8582	–
4	0.2708	−0.0142	0.6909	−0.4858	−0.4617

We define the *mth order difference-based estimator* of v by

$$\widehat{v} = \sum_{|k|<\ell} K\left(\frac{k}{\ell}\right)\widehat{\gamma}_k^D, \quad \text{where } \widehat{\gamma}_k^D = \frac{1}{n}\sum_{i=mh+|k|+1}^{n} D_i D_{i-|k|}. \tag{2.3}$$

We may write $\widehat{v}$ as $\widehat{v}_{(m)}$ to emphasize the order m. In (2.3),

$$m = m_n \in \mathbb{N}_0, \qquad \ell = \ell_n \in \{1, \ldots, n\}, \qquad h = h_n \in \{1, \ldots, n\} \tag{2.4}$$

are the order of differencing, bandwidth parameter and lag parameter of the estimator $\widehat{v}$, respectively. The function $K : \mathbb{R} \to \mathbb{R}$ is called a kernel, which satisfies that $K(0) = 1$, $K(t) = K(-t)$ for all t, $K(t) = 0$ for $|t| \geq 1$ and K is continuous on $(-1, 1)$. Popular kernels include the Bartlett kernel $K_{\text{Bart}}(t) = (1 - |t|)^+$ and the flat-top truncated kernel $K_{\text{Flat}}(t) = \mathbb{1}_{(|t|<1)}$. In (2.3), one may alternatively use $\sum_{i=mh+|k|+1}^{n}(D_i - \bar{D}_n)(D_{i-|k|} - \bar{D}_n)/n$ instead of $\widehat{\gamma}_k^D$, where $\bar{D}_n = \sum_{i=mh+1}^{n} D_i/n$. It does not affect the asymptotic results in this article.

The kernel estimator $\widehat{v}$ can be written in a subsampling form. For each $i = \ell, \ldots, n$, define the ith subsample of size ℓ as $\{D_t : t \in \Lambda_i\}$, where $\Lambda_i = \{i - \ell + 1, \ldots, i\}$. If $\mathcal{I} \subseteq \{mh + \ell + 1, \ldots, n\}$ is the set of subsample indices to be used, then the *subsampling estimator* of v is defined as

$$\widehat{v}' = \frac{1}{|\mathcal{I}|}\sum_{i\in\mathcal{I}} \widehat{v}'(i), \quad \text{where } \widehat{v}'(i) = \sum_{t,t'\in\Lambda_i} \frac{K(|t - t'|/\ell)}{\ell - |t - t'|} D_t D_{t'}, \tag{2.5}$$

and $|\mathcal{I}|$ is the total number of subsamples. The overlapping subsamples and nonoverlapping subsamples utilize $\mathcal{I}_1 = \{mh + 1 + \ell, \ldots, n\}$ and

$$\mathcal{I}_0 = \big\{mh + 1 + \ell, 2(mh + 1 + \ell), \ldots, \big\lfloor n/(mh + 1 + \ell)\big\rfloor (mh + 1 + \ell)\big\},$$

respectively. Similar ideas can be found in, for example, Carlstein (1986) and Welch (1987). The estimator $\widehat{v}'$ can be regarded as a "bagged" estimator of v by averaging the rough estimators (or weak "learners") $\{\widehat{v}'(i)\}_{i\in\mathcal{I}}$. If computational time is a concern, we may use the nonoverlapping subsamples. However, its statistical efficiency is reduced; see, for example, Alexopoulos, Goldsman and Wilson (2011). On the other hand, if the overlapping subsamples are used, the estimators $\widehat{v}'$ and $\widehat{v}$ are asymptotically equivalent in the following sense.

PROPOSITION 2.1 (Asymptotic equivalence of $\widehat{v}$ and $\widehat{v}'$). *Consider $\widehat{v}'$ with the overlapping subsamples $\mathcal{I} = \mathcal{I}_1$, and the order of differencing $m = m_n$. If Assumptions 1–2 hold, $1/\ell + (\ell + mh)/n \to 0$, and $\mathcal{G} \gtrsim \ell + mh$, then for any $\mu(\cdot)$ and $K(\cdot)$, we have*

$$\|\widehat{v} - \widehat{v}'\| = O\left(\frac{\ell + mh}{n}\right)(1 + \|\widehat{v}\|) + r_{\text{sub}}, \tag{2.6}$$

where $r_{\text{sub}} = O\{\mathcal{C}^2(\ell + mh)^4/n^3 + \mathcal{S}^2/n\}$.

The proof of Proposition 2.1 can be found in Section A.1 of the supplement. By Minkowski's inequality, (2.6) implies $\|\widehat{v} - \widehat{v}'\| \leq O\{(\ell + mh)/n\}(1 + v + \|\widehat{v} - v\|) + r_{\text{sub}}$, where the root-MSE $\|\widehat{v} - v\| \to 0$ if $\widehat{v}$ is $\mathcal{L}^2$ consistent for v. So, (2.6) reduces to $\|\widehat{v} - \widehat{v}'\| = O\{(\ell + mh)/n\} + r_{\text{sub}}$ if $\widehat{v}$ is $\mathcal{L}^2$ consistent. The remainder term r_{sub} is negligible if $\ell + mh$ is not too large. For example, if $\ell + mh = O(n^\theta)$ for some $\theta \in (0, 1/2]$, then $r_{\text{sub}} = O\{(\mathcal{C}^2 + \mathcal{S}^2)/n\}$. We emphasize that Proposition 2.1 is true even for a possibly divergent $m = m_n$ and a possibly nonconstant $\mu(\cdot)$ under the regularity conditions in Proposition 2.1.

2.3. *Existing variance estimators*. Many popular variance estimators admit the forms of (2.3) or (2.5). They are presented and categorized in Examples 2.2–2.5 according to the values of h and m.

EXAMPLE 2.2 (Serially uncorrelated case: $h = 1$). Assume $\gamma_1 = \gamma_2 = \cdots = 0$. Then $v = \sum_{k\in\mathbb{Z}} \gamma_k$ reduces to $v = \gamma_0$. Hall, Kay and Titterington (1990) proposed to estimate v by the $\widehat{v}$ in (2.3) with $h = 1$. Such $\widehat{v}$ is just an estimator of the marginal variance γ_0 but not $v = \sum_{k\in\mathbb{Z}} \gamma_k$. Recently, Tecuapetla-Gómez and Munk (2017) and Levine and Tecuapetla-Gómez (2019) extended it to estimation of γ_k for M-dependent time series, that is, $\gamma_k = 0$ for $|k| > M$. Their proposal is a special case of $\widehat{\gamma}_k^D$ in (2.3) when $n \to \infty$. The assumption of M-dependence can be restrictive in real applications. Most importantly, they did not consider the estimation of $v = \sum_{k\in\mathbb{Z}} \gamma_k$.

EXAMPLE 2.3 (Constant-mean case: $m = 0$). Assume $\mu_1 = \cdots = \mu_n$. Let $D_i' = X_i - \bar{X}_n$ for each i. Then v is estimated by $\widehat{v}$ and $\widehat{v}'$ with $m = 0$:

$$\widehat{v} = \sum_{|k|\le\ell} \frac{K(k/\ell)}{n} \sum_{i=1+|k|}^{n} D_i' D_{i-|k|}', \qquad \widehat{v}' = \frac{1}{|\mathcal{I}|} \sum_{i\in\mathcal{I}} \sum_{t,t'\in\Lambda_i} \frac{K(|t-t'|/\ell)}{\ell - |t-t'|} D_t' D_{t'}'. \tag{2.7}$$

The kernel estimator $\widehat{v}$ in (2.7) has a long history in statistics, econometrics and operational research; see, for example, Newey and West (1987), Andrews (1991) and Politis, Romano and Wolf (1999). The subsampling estimator $\widehat{v}'$ in (2.7) is studied in, for example, Song and Schmeiser (1995) and Chan and Yau (2017b). If $K = K_{\text{Bart}}$ and $\mathcal{I} = \mathcal{I}_1$, then $\widehat{v}' = \sum_{i=\ell}^{n} (\sum_{j=i-\ell+1}^{i} D_j')^2 / \{\ell(n-\ell+1)\}$ is the well-known overlapping batch means (OBM) estimator (Chen and Schmeiser (2013)). Besides, Carlstein (1986) and Alexopoulos, Goldsman and Wilson (2011) studied the nonoverlapping and partially overlapping subsamples, however, these schemes are suboptimal in terms of MSE.

EXAMPLE 2.4 (General case: $m = 1$, $h \to \infty$). In the presence of both serial dependence and time-varying means, the estimation of v is less well studied. Let $\widetilde{v}' = \ell \sum_{i\in\mathcal{I}} (S_i - S_{i-\ell})^2 / \{2|\mathcal{I}|\}$, where $S_i = \sum_{j=i-\ell+1}^{i} X_j/\ell$ is the ith subsample mean. This class of estimators is independently proposed by various authors. For example, Dette and Wu (2019) used $\mathcal{I} = \mathcal{I}_1$, whereas Wu, Woodroofe and Mentz (2001), Wu (2004), Wu and Zhao (2007), Dette, Eckle and Vetter (2020) and Chen, Wang and Wu (2021) used $\mathcal{I} = \mathcal{I}_0$. In either case, $\widetilde{v}'$ is just a special case of $\widehat{v}'$ with $m = 1$, $h = \ell = O(n^{1/3})$ and $K = K_{\text{Bart}}$. Moreover, none of them provides the optimal value of ℓ.

EXAMPLE 2.5 (General case: $m \to \infty$, $h = 1$). Altissimo and Corradi (2003) proposed to locally center X_i by using the D_i'' defined in (2.1) with $H = K_{\text{Flat}}$, $K = K_{\text{Bart}}$ and $m \to \infty$. Their estimator is asymptotically equivalent to $\widehat{v}$ with $d_j = \{\mathbb{1}_{(j=\lfloor m/2 \rfloor)} - w_j\}/c$ for $j = 0, \ldots, m$, where $c, w_0, \ldots, w_m$ are some constants such that $w_0 + \cdots + w_m = 1$ and $\delta_0 = 1$. A similar proposal can be found in Juhl and Xiao (2009).

Some estimators cannot be expressed as (2.3) or (2.5); see Examples 2.6–2.8. All of them are suboptimal or require restrictive assumptions.

EXAMPLE 2.6 (Removal of one CP). Crainiceanu and Vogelsang (2007) proposed to estimate one single potential CP T_1 by the standard CUSUM-type estimator $\widehat{T}_1$. After centering $\{X_i\}_{i=1}^{\widehat{T}_1 - 1}$ and $\{X_i\}_{i=\widehat{T}_1}^{n}$ by their respective sample means, one may apply (2.7) to the centered series to estimate v. This method is vulnerable to the one-CP assumption. Although it can be extended to handle multiple CPs, the accumulated errors may ruin the final estimator.

Recently, Dehling, Fried and Wendler (2020) proposed to split $X_{1:n}$ into three disjoint subsamples of (approximately) equal length so that (2.7) can be applied to each of the three subsamples. The final estimator is the sample median of the three estimators. This method incurs a huge loss of efficiency. It is remarked that their method is applied to ranks of $X_{1:n}$ instead of $X_{1:n}$, but their idea is still applicable generally.

EXAMPLE 2.7 (Mean and median of absolute deviations). Apart from the estimator $\widetilde{v}'$ in Example 2.4, Wu and Zhao (2007) also proposed two other estimators that utilize the sample mean and sample median of $\{|S_i - S_{i-\ell}|\}_{i\in\mathcal{I}_1}$, that is,

$$\widetilde{v}'' = \frac{\pi}{4(\lfloor n/\ell \rfloor - 1)^2} \sum_{i\in\mathcal{I}_1} |S_i - S_{i-\ell}| \quad \text{and} \quad \widetilde{v}''' = \frac{1}{2z_{3/4}^2} \operatorname*{median}_{i\in\mathcal{I}_1} |S_i - S_{i-\ell}|,$$

where $\operatorname{median}_{k\in\mathcal{K}} x_k$ is the sample median of $\{x_k\}_{k\in\mathcal{K}}$, and z_p is the $100p\%$ quantile of $\mathrm{N}(0,1)$. They proved that the convergence rates of $\widetilde{v}''$ and $\widetilde{v}'''$ are much slower than that of the $\widetilde{v}'$ in Example 2.4.

EXAMPLE 2.8 (Insufficient differencing). Chan (2022a) proposed an estimator that is asymptotically equivalent to $\overline{v} = \sum_{|k|\le\ell} K(k/\ell)\overline{\gamma}_k$, where $\overline{\gamma}_k = \sum_{i=k+\ell+1}^{n} X_i(X_{i-k} - X_{i-k+\ell})/n$. It is an incomplete special case of $\widehat{v}$ with $m = 1$ and $h = \ell$. It is an incomplete version because $\overline{\gamma}_k$ is constructed by the product of the raw observation X_i and the difference statistic $X_{i-k} - X_{i-k+\ell}$, whereas our proposed statistic $\widehat{\gamma}_k^D$ in (2.3) is constructed by the product of two difference statistics D_i and D_{i-k}. We prove that $\widehat{v}$ is uniformly better than this "insufficient" difference-based estimator $\overline{v}$.

We also remark that some statistical procedures do not require estimation of the LRV by utilizing self-normalization; see, for example, Lobato (2001), Shao (2010), and Cheng and Chan (2022). However, different specifically designed self-normalizers may be needed for handling different types of mean structure; see, for example, Zhao (2011), Zhang and Lavitas (2018) and Pešta and Wendler (2020). This alternative approach may also lead to a decrease in power or statistical efficiency. Nevertheless, they enjoy some added appealing properties. We refer interested readers to an excellent review by Shao (2015).

2.4. *Interpretation and representation.* Recall the definitions of h and ℓ in (2.4). We parametrize $h = \ell\lambda \in \mathbb{N}$ for some $\lambda := \lambda_n \to \lambda_\infty \in [0,\infty]$. The goal of this section is to provide statistical interpretations of $\widehat{v}$ under different values of λ.

PROPOSITION 2.2. *Suppose Assumptions 1–2 hold, and* $1/\ell + (\ell + mh)/n = o(1)$. *Let the* differencing kernel *be*

$$K_{\mathsf{diff}}(t) = \sum_{s=\lceil -(1+t)/\lambda\rceil}^{\lfloor (1-t)/\lambda\rfloor} \delta_{|s|} K(t+\lambda s), \quad t\in\mathbb{R}. \tag{2.8}$$

Define the differencing kernel estimator *as* $\widehat{v}_{\mathsf{diff}} = \sum_{|k|\le\ell+mh} K_{\mathsf{diff}}(k/\ell)\widehat{\gamma}_k^X$, *where* $\widehat{\gamma}_k^X = \sum_{i=|k|+1}^{n}(X_i - \bar{X}_n)(X_{i-|k|} - \bar{X}_n)/n$.

1. (*Representation*) *If* $\mu(t) \equiv \mu_0$ *for all* $t\in[0,1]$, *then as* $n\to\infty$,

$$\|\widehat{v} - \widehat{v}_{\mathsf{diff}}\| = O\{(\ell + mh)/n\}. \tag{2.9}$$

It remains true if $\widehat{\gamma}_k^X$ *is replaced by* $\widehat{\gamma}_k^Z = \sum_{i=|k|+1}^{n} Z_i Z_{i-|k|}/n$ *in* $\widehat{v}_{\mathsf{diff}}$.

2. (*Differencing property*) *If* $m = 0$, *then* $K_{\mathsf{diff}} = K$. *If* $m > 0$, *then* K_{diff} *satisfies that* $\sum_{|k|\le\ell+mh} K_{\mathsf{diff}}(k/\ell) = 0$.

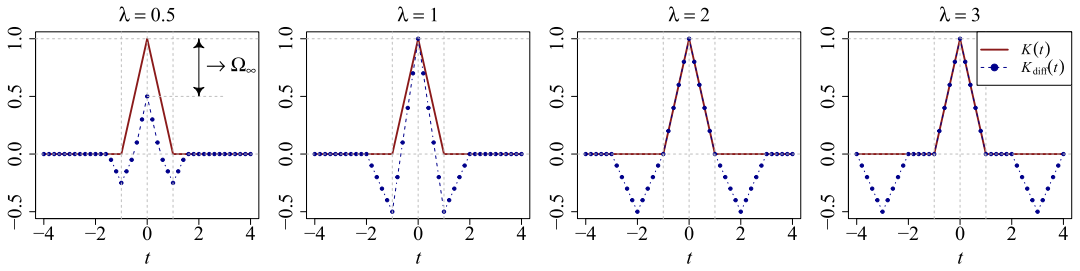

FIG. 3. *Comparisons between K and K_{diff} when $m = 1$, $K = K_{\text{Bart}}$ and $\lambda \in \{1/2, 1, 2, 3\}$. The quantity Ω_∞ in the first plot is defined in Assumption 3.*

3. (*Kernel necessity*) *If* $\lambda \geq 1$, *then* $K_{\text{diff}}(0) = 1$.
4. (*Matching property*) *If* $\lambda \geq 2$, *then* $K_{\text{diff}}(t) = K(t)$ *for all* $t \in [-1, 1]$.

The proof can be found in Section A.2 of the supplement. Proposition 2.2(1) states that $\widehat{v}$ smoothes $\{\widehat{\gamma}_k^X\}$ by the distorted kernel K_{diff} instead of the intended kernel K. So, $\widehat{v}$ may not inherit the properties of K, for example, the higher-order property of achieving a faster convergence rate of $\widehat{v}$ as in Andrews (1991). Proposition 2.2(2) states that $\widehat{v}_{\text{diff}}$ is invariant to $\mu(\cdot)$ because the kernel weights must sum to zero. This property is not achieved by most commonly used kernels. Proposition 2.2(3) states that K_{diff} satisfies the minimal requirement as a kernel if $\lambda \geq 1$. A striking fact conveyed by Proposition 2.2(4) is that if $\lambda = h/\ell$ is large enough (≥ 2), then

$$\widehat{v} \equiv \sum_{|k| \leq \ell} K\left(\frac{k}{\ell}\right) \widehat{\gamma}_k^D = \sum_{|k| \leq \ell} K\left(\frac{k}{\ell}\right) \widehat{\gamma}_k^X + r_{\text{diff}},$$

where the remainder term r_{diff} involves only $\{\widehat{\gamma}_k\}_{|k|>\ell}$, which are expected to have a negligible contribution to $\widehat{v}$ owing to weak dependence (Assumption 1). In this case, $\widehat{v}$ correctly uses the intended kernel $K(\cdot)$ to smooth $\{\widehat{\gamma}_k\}_{|k|\leq\ell}$. However, the existing estimators in Example 2.4, which employ $\lambda = h/\ell = 1$, incorrectly utilize another kernel. Example 2.9 below visualizes this fact.

EXAMPLE 2.9. Consider $m = 1$ and $K = K_{\text{Bart}}$. Figure 3 visualizes how K_{diff} changes with λ. When $\lambda < 1$, we have $K_{\text{diff}}(0) \neq 1$. So, K_{diff} is not even qualified as a kernel for estimating v. When $1 \leq \lambda < 2$, we have $K_{\text{diff}}(0) = 1$ but $K_{\text{diff}}(t) \not\equiv K(t)$ even for $|t| \leq 1$. It implies that K_{diff} does not share the same properties as K. When $\lambda \geq 2$, we have $K_{\text{diff}}(t) \equiv K(t)$ for all $|t| \leq 1$. In this case, K_{diff} copies most properties of the kernel K.

REMARK 2.1. The differencing kernel K_{diff} may depend on n when $\lambda = \lambda_n$ or $m = m_n$ depends on n. We do not recommend to use $\widehat{v}_{\text{diff}}$ in practice as it has a larger influence by a nonconstant $\mu(\cdot)$ than our proposed $\widehat{v}$ and $\widehat{v}'$ in (2.3) and (2.5). The representation (2.9) is true only under the constant $\mu(\cdot)$ assumption. If $\mu(\cdot)$ is not a constant, the upper bound for $\|\widehat{v} - \widehat{v}_{\text{diff}}\|$ is larger than that in (2.9). It may not be enough to establish asymptotic equivalence. Nevertheless, it is informative to use K_{diff} to understand the proposed $\widehat{v}$.

2.5. *Dual representations*. There are two types of ambiguous dual representations of $\widehat{v}$. First, a high order but sparse difference sequence can be represented by a lower-order sequence with a larger lag h, for example, the fourth-order sequence $\{-1/\sqrt{2}, 0, 0, 0, 1/\sqrt{2}\}$ with $h = 1$ is equivalent to the first-order sequence $\{-1/\sqrt{2}, 1/\sqrt{2}\}$ with $h = 4$. Sparse difference sequences lead to $\delta_s = 0$ for all small s. Second, consider $\widehat{v}$ with kernel $K^\circ(\cdot)$, bandwidth ℓ°, and lag h°. The estimator does not change if we stretch the kernel to $K(\cdot) = K^\circ(C\cdot)$

with $\ell = C\ell^\circ$ and $h = h^\circ$, where $C > 1$. This type of stretched kernel truncates earlier than ± 1, that is, $K(t) = 0$ for $|t| \geq 1/C$. Assumption 3 below rules out these ambiguous dual representations.

ASSUMPTION 3 (Unambiguity). For $\lambda < 1$, $2\sum_{s=1}^{\lfloor 1/\lambda \rfloor} \delta_s K(\lambda s) \to \Omega_\infty \neq 0$.

Most commonly used kernels and difference sequences satisfy Assumption 3. For example, if $K = K_{\text{Bart}}$, Assumption 3 is satisfied for all $0 \leq m \leq 10$, all h, and all $\{d_j\}$ in Example 2.1. Indeed, Ω_∞ measures the limiting gap between 1 and $K_{\text{diff}}(0)$, that is, $K_{\text{diff}}(0) - 1 \to \Omega_\infty$ as $n \to \infty$; see Figure 3.

However, it is possible that Assumption 3 is not satisfied when $m \to \infty$. It may happen when the difference sequence is approximately "uncorrelated," that is, $\delta_s \approx 0$ for $s \neq 0$. We formalize this situation by the following assumption.

ASSUMPTION 4 (Approximately uncorrelated differencing). There are $c', c'' > 0$ such that $-c''/m \leq \delta_s \leq -c'/m$ for all $s = \pm 1, \ldots, \pm m$.

Note that Assumption 4 is satisfied by the local difference sequence and Hall, Kay and Titterington's (1990) difference sequence (see Example 2.1). We will show in Section 5.2 that the first type of sequence is nearly optimal whereas the second type is asymptotically optimal.

3. Invariance to time-varying means.

3.1. *Strength of robustness.* When the true mean function is $\mu(\cdot)$, we denote the bias, variance and MSE of $\widehat{v}$ by $\mathsf{Bias}_\mu(\widehat{v}; v)$, $\mathsf{Var}_\mu(\widehat{v})$ and $\mathsf{MSE}_\mu(\widehat{v}; v)$, respectively. If $\mu(t) \equiv 0$, we write them as $\mathsf{Bias}_0(\widehat{v}; v)$, $\mathsf{Var}_0(\widehat{v})$ and $\mathsf{MSE}_0(\widehat{v}; v)$ to emphasize that it is the ideal situation as $X_i = Z_i$ for all i. If the estimand is v, we may omit the argument "v" in the bias and MSE.

DEFINITION 2 (Robustness against mean functions). Let $\mathcal{M}$ be a family of mean functions. An estimator $\widehat{\theta}$ of θ is said to be *strictly robust in* $\mathcal{M}$ if $\mathsf{MSE}_\mu(\widehat{\theta}; \theta) \sim \mathsf{MSE}_0(\widehat{\theta}; \theta)$ for all $\mu \in \mathcal{M}$; and *loosely robust in* $\mathcal{M}$ if $\mathsf{MSE}_0(\widehat{\theta}; \theta) \to 0$ implies $\mathsf{MSE}_\mu(\widehat{\theta}; \theta) \to 0$ for all $\mu \in \mathcal{M}$.

If $\widehat{\theta}$ is strictly robust, its first-order $\mathcal{L}^2$ asymptotic properties are the same for any $\mu \in \mathcal{M}$. It is the most desirable. If $\widehat{\theta}$ is loosely robust, its $\mathcal{L}^2$ consistency is maintained within $\mathcal{M}$ but the convergence rate and MSE may be different.

THEOREM 3.1 (Robustness). *Let* $\kappa := \int_{-1}^{1} K(t)\,\mathrm{d}t \neq 0$. *Suppose Assumptions 1–2 hold,* $\ell/n \to 0$, *and* $\mathcal{G} \gtrsim \ell + mh$. *Also let* $\ell, h \in \mathbb{N}$ *and* $m \in \mathbb{N}_0$, *which are possibly divergent. Then*

$$\mathsf{Bias}_\mu(\widehat{v}; v) = \mathsf{Bias}_0(\widehat{v}; v) + \left\{\kappa \ell \mathcal{V} + O\left(\frac{\ell}{n}\right)\right\} \mathbb{1}_{(m=0)} + R_{\text{bias}},$$

$$\sqrt{\mathsf{Var}_\mu(\widehat{v})} = \sqrt{\mathsf{Var}_0(\widehat{v})} + O\left\{\frac{\ell(\mathcal{C} + \mathcal{S}\mathcal{J})}{\sqrt{n}}\right\} \mathbb{1}_{(m=0)} + R_{\text{se}},$$

where

$$R_{\text{bias}} = O\left[\frac{\ell}{n}\left\{(\ell \mathbb{1}_{m=0} + mh)\mathcal{S}^2\mathcal{J} + \frac{(\ell \mathbb{1}_{m=0} + mh)^2 \mathcal{C}^2}{n}\right\}\right],$$

$$R_{\text{se}} = O\left[\frac{\ell}{\sqrt{n}}\left\{\frac{mh\mathcal{C}}{n} + \mathcal{S}\left(\frac{mh\mathcal{J}}{n}\right)^{1/2}\right\}\right].$$

The proof can be found in Section A.3 of the supplement. Theorem 3.1 states that the bias and variance of $\widehat{v}$ are governed by (i) the performance of $\widehat{v}$ when $\mu \equiv 0$, (ii) the order m and (iii) the mean function $\mu(\cdot)$.

Factor (i) is the idealistic performance. The estimator $\widehat{v}$ is good if $\mathsf{Bias}_\mu(\widehat{v}) \sim \mathsf{Bias}_0(\widehat{v})$ and $\mathsf{Var}_\mu(\widehat{v}) \sim \mathsf{Var}_0(\widehat{v})$. Factor (ii) has the greatest impact on $\widehat{v}$. When $m = 0$, the estimator reduces to the classical estimators in Example 2.3. It has a divergent bias if the mean is a fixed nonconstant function. If $\mathcal{V} = o(1/\ell)$, $\widehat{v}$ is still consistent. However, in this case, the variability of $\mu(\cdot)$ is diminishing as $n \to \infty$. This robustness is insufficient for most real applications. Similar results have been documented in, for example, Gonçalves and White (2002). Factor (iii) depends on $\mu(\cdot)$ only through $\mathcal{C}$, $\mathcal{S}$ and $\mathcal{J}$. When $m > 0$, $\mu(\cdot)$ only affects R_{bias} and R_{se}, which are typically negligible; see Section 3.2.

Besides, Theorem 3.1 is also applicable to the estimators in Examples 2.2–2.5. We compare them in Example 3.1 below.

EXAMPLE 3.1 (Comparison). In the presence of time-varying mean and autocorrelation, our proposed estimator $\widehat{v}$ and the existing estimators in Examples 2.4–2.8 can be used for estimating v. We compare their robustness as follows:

- Our framework in (2.3) covers the proposed $\widehat{v}$ and the estimators in Examples 2.4–2.5. Although they use different values of $m > 0$ and h, all of them satisfy
$$\mathsf{MSE}_\mu(\widehat{v}) = \{\mathsf{Bias}_0(\widehat{v}) + R_{\mathsf{bias}}\}^2 + \{\sqrt{\mathsf{Var}_0(\widehat{v})} + R_{\mathsf{se}}\}^2$$
according to Theorem 3.1 with
$$R_{\mathsf{bias}} = O\left[\frac{\ell}{n}\left\{H\mathcal{S}^2\mathcal{J} + \frac{H^2\mathcal{C}^2}{n}\right\}\right] \quad \text{and} \quad R_{\mathsf{se}} = O\left\{\frac{\ell}{\sqrt{n}}\left(\frac{H\mathcal{C}}{n} + \mathcal{S}\left(\frac{H\mathcal{J}}{n}\right)^{1/2}\right)\right\},$$
where $H := mh$. Clearly, as long as ℓ and H remain unchanged, the orders of R_{bias} and R_{se} do not change with h and m. In other words, all these estimators are equally robust against the mean functions asymptotically. It is worth mentioning that we further enhance the finite-sample robustness of our proposed estimator in Section 6.1.
- Since the estimators in Examples 2.6–2.8 do not fall into our framework, we compare their robustness via simulation in Section 7.1. It indicates that our proposed $\widehat{v}$ is more robust against the mean functions than all competitors.

3.2. *Class of well-behaved mean functions.* We will prove in Sections 4–5 that, under the assumption $u_q := \sum_{k\in\mathbb{Z}} |k|^q |\gamma_k| < \infty$, the estimator $\widehat{v}$ is consistent and rate optimal with $\mathsf{MSE}_0(\widehat{v}) \asymp n^{-2q/(1+2q)}$ when $m \in \mathbb{N}$, $h/\ell = \lambda \in (0, \infty)$, and $\ell \asymp n^{1/(1+2q)}$; see (5.3). The same optimal MSE is also achieved by the standard estimators; see, for example, Andrews (1991). Using the baseline $\mathsf{MSE}_0(\widehat{v}) \asymp n^{-2q/(1+2q)}$, Theorem 3.1 shows that $\widehat{v}$ is strictly robust in

$$\begin{aligned}\mathcal{M}_q &:= \{\mu(\cdot) : R_{\mathsf{bias}}^2 + R_{\mathsf{se}}^2 = o(\mathsf{MSE}_0(\widehat{v}))\} \\ &= \{\mu(\cdot) : \mathcal{C}^2 = o(n^{\frac{3q-1}{1+2q}}), \mathcal{S}^2\mathcal{J} = o(n^{\frac{q-1}{1+2q}}), \mathcal{G} \gtrsim n^{\frac{1}{1+2q}}\},\end{aligned} \tag{3.1}$$

which covers a large class of mean functions. A larger class can similarly be derived if we only require $\widehat{v}$ to be loosely robust. Example 3.2 discuss a special case when $q = 2$.

EXAMPLE 3.2 (Strict robustness of $\widehat{v}$ with $q = 2$). Suppose $u_2 < \infty$. The optimal MSE satisfies $\mathsf{MSE}_0(\widehat{v}) = O(n^{-4/5})$. Then $\mathcal{M}_2 = \{\mu(\cdot) : \mathcal{C}^2 = o(n), \mathcal{S}^2\mathcal{J} = o(n^{1/5}), \mathcal{G} \gtrsim n^{1/5}\}$, which includes (i) all Lipschitz continuous functions with $o(n^{1/2})$ Lipschitz constants, (ii) all step functions with $o(n^{1/5})$ number of finite-jump discontinuities that are separated by at least $O(n^{1/5})$ and (iii) a sum of (i) and (ii). For example, all mean functions in Figure 1 are members of $\mathcal{M}_2$.

4. Classes of consistent and rate-optimal estimators. It is unclear whether $\widehat{v}$ is consistent for $v = \sum_{k\in\mathbb{Z}} \gamma_k$ even under the constant mean assumption because the ACVF of $\{D_i\}$ is not equal to γ_k:

$$\gamma_k^D := \mathsf{Cov}(D_i, D_{i+k}) = \sum_{j,j'=0}^{m} d_j d_{j'} \gamma_{h(j-j')+k} = \sum_{|s|\le m} \delta_s \gamma_{hs+k} \neq \gamma_k. \tag{4.1}$$

In this section, we study the conditions for consistency and rate optimality for $\widehat{v}$ when the mean function $\mu(\cdot)$ is mildly nonconstant in the sense that $\mathcal{J}, \mathcal{S}, \mathcal{C} \asymp 1$. Our asymptotic theory requires the following regularity conditions on K.

ASSUMPTION 5 (Near-origin property). The kernel K satisfies that there exist $q \in \mathbb{N}$ and $B \in \mathbb{R} \setminus \{0\}$ such that $\{K(t) - K(0)\}/|t|^q \to B$ as $t \downarrow 0$.

ASSUMPTION 6 (Near-boundary property). The kernel K satisfies that there exist $q' \in \mathbb{N}$ and $B' \in \mathbb{R} \setminus \{0\}$ such that $\{K(1) - K(1-t)\}/|t|^{q'} \to B'$ as $t \downarrow 0$.

Assumption 5 is standard. The index q is called the characteristic exponent (CE) of $K(\cdot)$; see Parzen (1957). We say that a kernel K is of order q if Assumption 5 is satisfied. The larger the value of q, the flatter the kernel is around 0. It governs the order of bias of $\widehat{v}$ in the stationary case. In particular, for the kernel estimator with a qth order kernel in Example 2.3, if $u_q = \sum_{k\in\mathbb{Z}} |k|^q |\gamma_k| < \infty$, the best possible bias is $O(1/\ell^q)$, and the resulting optimal MSE is $O(n^{-2q/(1+2q)})$; see, for example, Andrews (1991). Therefore, we say that $\widehat{v}$ is rate optimal if its MSE attains $O(n^{-2q/(1+2q)})$ for all time series that satisfy $u_q < \infty$. Some commonly used kernels are shown in Table 2. We suggest to use Parzen's (1957) kernel $K_q(t) = (1 - |t|^q)^+$ as a convenient choice as it satisfies Assumption 5 with any specified $q \in \mathbb{N}$.

Assumption 6 is nonstandard. It states the flatness of $K(t)$ when $t \uparrow 1$; see Table 2 for a summary of kernels that satisfy Assumption 6. Although Assumption 6 does not affect the convergence rate of the classical estimators in Example 2.3, it plays an important role for difference-based estimators when $h/\ell \to 1$.

4.1. *Fixed-m difference-based estimators.* Theorem 4.1 below studies the consistency and rate-optimality of $\widehat{v}$ when $0 < m < \infty$ in different regimes according to the limiting value of h/ℓ; see (2.4) for the definitions of m, ℓ and h.

TABLE 2

Some commonly used kernels. The last two columns indicate the values of q and q' so that Assumptions 5 and 6 are satisfied, respectively. In lugsail kernel, $r \ge 1$, $c \in [0, 1)$ and K_0 is any initial kernel. In trapezoidal kernel, $c' \in (0, 1]$. Trapezoidal and truncated kernels do not satisfy Assumption 5 because $B = 0$ for any $q \in \mathbb{N}$

Kernel	Definition	Assumption 5	Assumption 6
Bartlett (Newey and West (1987))	$K(t) = (1 - \lvert t\rvert)^+$	$q = 1$	$q' = 1$
Tukey–Hanning (Andrews (1991))	$K(t) = \{1 + \cos(\pi t)\}\mathbb{1}(\lvert t\rvert \le 1)/2$	$q = 2$	$q' = 2$
Parzen (Gallant (1987))	$K(t) = \begin{cases} 1 - 6t^2 + 6\lvert t\rvert^3, & \lvert t\rvert \le 1/2; \\ 2\{(1 - \lvert t\rvert)^+\}^3, & \lvert t\rvert > 1/2. \end{cases}$	$q = 2$	$q' = 3$
qth order polynomial (Parzen (1957))	$K(t) = (1 - \lvert t\rvert^q)^+$	$q \in \mathbb{N}$	$q' = 1$
Lugsail (Vats and Flegal (2021))	$K(t) = \{K_0(t) - cK_0(rt)\}/(1 - c)$	Same as K_0	Depends on K_0
Trapezoidal (Politis and Romano (1995))	$K(t) = \begin{cases} 1, & \lvert t\rvert \le c'; \\ (1 - \lvert t\rvert)^+/(1 - c'), & \lvert t\rvert > c'. \end{cases}$	Not satisfied	$q' = 1$
Truncated (White (1984))	$K(t) = \mathbb{1}(\lvert t\rvert < 1)$	Not satisfied	Not satisfied
Modified qth order polynomial	Equation (4.2)	$q \in \mathbb{N}$	$q' = q$

THEOREM 4.1 (Finite-m regime). *Suppose that $\mu(\cdot)$ satisfies (1.3) with $\mathcal{J}, \mathcal{S}, \mathcal{C} \asymp 1$; and $\{Z_i\}_{i\in\mathbb{Z}}$ satisfies Assumption 1 and $u_q = \sum_{k\in\mathbb{Z}} |k|^q |\gamma_k| < \infty$ for some $q \in \mathbb{N}$. Let ℓ be an unknown free bandwidth satisfying $1/\ell + (\ell + mh)/n \to 0$. Suppose $0 < m < \infty$ is fixed, and Assumption 5 holds. Under the least favorable data generating mechanism, we have the following results*:

1. *If $h/\ell \to 0$, then $\widehat{v}$ is inconsistent in $\mathcal{L}^2$.*
2. *If $h/\ell \to \lambda_\infty \in (0, 1)$, then under Assumption 3, $\widehat{v}$ is inconsistent in $\mathcal{L}^2$.*
3. *If $h/\ell \to 1$, then under Assumption 6 and $|h - \ell| = O(1)$, the best possible MSE is $\mathsf{MSE}_\mu(\widehat{v}) \asymp n^{-2(q\wedge q')/\{1+2(q\wedge q')\}}$, which is achieved by $\ell \asymp n^{1/\{1+2(q\wedge q')\}}$.*
4. *If $h/\ell \to \lambda_\infty \in (1, \infty)$, then the best possible MSE is $\mathsf{MSE}_\mu(\widehat{v}) \asymp n^{-2q/(1+2q)}$, which is achieved by $\ell \asymp n^{1/(1+2q)}$.*
5. *Suppose $h/\ell \to \infty$.*

   (a) *If $q = 1$, then $\widehat{v}$ is rate suboptimal in $\mathcal{L}^2$.*

   (b) *If $q > 1$, then the best possible MSE is $\mathsf{MSE}_\mu(\widehat{v}) \asymp n^{-2q/(1+2q)}$, which is achieved by $\ell \asymp n^{1/(1+2q)}$ and $n^{1/(1+2q)} \ll h \lesssim n^{q/(1+2q)}$.*

The proof can be found in Section A.4 of the supplement. From Theorem 4.1(1)–(2), $\widehat{v}$ with $h/\ell \to \lambda_\infty \in [0, 1)$ should never be used as it is guaranteed to be inconsistent for v. Theorem 4.1(3)–(5) state the fastest possible convergence rate of $\widehat{v}$. Under Assumption 5, the optimal MSE in the stationary case is $O\{n^{-2q/(1+2q)}\}$; see Andrews (1991). In case (5), the rate optimality cannot be achieved for handling time series that satisfies $u_q < \infty$ with $q = 1$ only. In case (3), the rate optimality cannot be achieved by all qth order kernels unless they satisfy Assumption 6 with $q' \geq q$, which means that $K(t)$ is flatter or equally flat near the boundary $t \uparrow 1$ than near the origin $t \downarrow 0$. The condition $|h - \ell| = O(1)$ means that $h/\ell \to 1$ sufficiently quickly. The requirement $q' \geq q$ is not satisfied by all kernels; see Table 2. For example, Parzen's (1957) kernel $K_q(t) = (1 - |t|^q)^+$ satisfies Assumption 5 for any $q \in \mathbb{N}$, but it only satisfies Assumption 6 with $q' = 1$. Hence, the rate optimality cannot be achieved when $q > 1$. One may design a smooth and differentiable kernel with $q' = q$ as follows:

$$\widetilde{K}(t) = \begin{cases} 1 - |t|^q + a|t|^{q+1} + b|t|^{q+2} & \text{if } |t| \leq 1/2; \\ (1 - |t|)^q - a(1 - |t|)^{q+1} - b(1 - |t|)^{q+2} & \text{if } 1/2 < |t| \leq 1; \\ 0 & \text{if } |t| > 1, \end{cases} \tag{4.2}$$

where $a = 4 - (q + 1)2^q$ and $b = q2^{q+1} - 4$. In this case, $\mathsf{MSE}(\widehat{v}) = O(n^{-2q/(1+2q)})$ if $\ell \asymp n^{1/(1+2q)}$. In case (4), it is more well behaved as $\widehat{v}$ is rate optimal for all $K(\cdot)$ without any additional assumption. We remark that users always know whether $\widehat{v}$ is consistent or not as the values of m, ℓ and h are specified by users. In practice, we suggest to select $h = \lambda_\infty \ell$ whenever it is possible so that h/ℓ equals to λ_∞ not only in the limit but also in finite samples.

4.2. *Divergent-m difference-based estimator.* This section investigates the convergence properties of $\widehat{v}$ with $m = m_n \to \infty$ as $n \to \infty$.

THEOREM 4.2 (Divergent-m regime). *Assume all conditions in Theorem 4.1 except that $0 < m < \infty$ is replaced by $m = m_n \to \infty$. In addition, suppose Assumption 2 holds. Under the least favorable data generating mechanism, we have the following results*:

1. *Suppose $h/\ell \to 0$.*

   (a) *If Assumption 3 holds, then $\widehat{v}$ is inconsistent in $\mathcal{L}^2$.*

(b) *If Assumption 4 holds and K is decreasing on* $[0,1]$, *then* (i) $\widehat{v}$ *is inconsistent in* $\mathcal{L}^2$ *for* $\ell/h \gtrsim m$, *and* (ii) $\widehat{v}$ *is rate suboptimal in* $\mathcal{L}^2$ *for* $\ell/h \ll m$.

2. *Suppose* $h/\ell \to \lambda_\infty \in (0,1)$.

(a) *If Assumption 3 holds, then* $\widehat{v}$ *is inconsistent in* $\mathcal{L}^2$.
(b) *If Assumption 4 holds and K is decreasing on* $[0,1]$, *then* $\widehat{v}$ *is rate suboptimal in* $\mathcal{L}^2$.

3. *Suppose* $h/\ell \to 1$. *Suppose further that* $|h-\ell| = O(1)$, *and Assumptions 4 and 6 hold.*

(a) *If* $q=1$ *or* $\ell^q \gg m\ell^{q'}$, *then* $\widehat{v}$ *is rate suboptimal in* $\mathcal{L}^2$.
(b) *If* $q>1$ *and* $\ell^q \lesssim m\ell^{q'}$, *then the best possible MSE is* $\mathsf{MSE}_\mu(\widehat{v}) \asymp n^{-2q/(1+2q)}$, *which is achieved by* $\ell \asymp n^{1/(1+2q)}$ *and any* $m \to \infty$ *such that* $n^{(q-q')/(1+2q)} \lesssim m \lesssim n^{(q-1)/(1+2q)}$.

4. *Suppose* $h/\ell \to \lambda_\infty \in (1,\infty)$ *and Assumption 4 holds.*

(a) *If* $q=1$, *then* $\widehat{v}$ *is rate suboptimal in* $\mathcal{L}^2$.
(b) *If* $q>1$, *then the best possible MSE is* $\mathsf{MSE}_\mu(\widehat{v}) \asymp n^{-2q/(1+2q)}$, *which is achieved by* $\ell \asymp n^{1/(1+2q)}$ *and* $1 \ll m \lesssim n^{(q-1)/(1+2q)}$.

5. *Suppose* $h/\ell \to \infty$.

(a) *If* $q=1$, *then* $\widehat{v}$ *is rate suboptimal in* $\mathcal{L}^2$.
(b) *If* $q>1$, *then the best possible MSE is* $\mathsf{MSE}_\mu(\widehat{v}) \asymp n^{-2q/(1+2q)}$, *which is achieved by* $h \gg \ell \asymp n^{1/(1+2q)}$ *and* $n^{1/(1+2q)} \ll mh \lesssim n^{q/(1+2q)}$.

The proof can be found in Section A.5 of the supplement. Theorem 4.2 implies that $\widehat{v}$ with a divergent m is inconsistent or suboptimal in Cases 1–2. Although $\widehat{v}$ is consistent in Cases 3–5, the rate optimality cannot be achieved for handling time series that satisfies $u_q < \infty$ with $q=1$ only. It is worth emphasizing that the variance estimators in Example 2.5 utilize a local centering technique with $h=1$ and $m, \ell \to \infty$. Since they fall in the regime $m \to \infty$ and $h/\ell \to 0$, the resulting estimators are inadmissible. Theorems 4.1 and 4.2 are summarized in Table 3.

5. Asymptotic optimality of variance estimators.

5.1. *Mean squared error.* From Section 4, $\widehat{v}$ is always rate optimal iff $h/\ell \to \lambda_\infty \in [1,\infty)$ and $m < \infty$. So, we study $\widehat{v}$ in this regime. From now on, we use a fixed $\lambda \equiv h/\ell = \lambda_\infty \in [1,\infty)$ and a fixed m for simplicity.

TABLE 3

Convergence properties of $\widehat{v}$ *when* μ *satisfies (1.3) with* $\mathcal{J}, \mathcal{S}, \mathcal{C} \asymp 1$ *and* $u_q = \sum_{k\in\mathbb{Z}} |k|^q |\gamma_k| < \infty$. *"Inconsistent" means that* $\mathsf{MSE}_\mu(\widehat{v}) \nrightarrow 0$ *for some time series. "Suboptimal" means that* $\mathsf{MSE}_\mu(\widehat{v}) \to 0$ *at a suboptimal rate for some* $q \in \mathbb{N}$. *"Optimal" means that* $\mathsf{MSE}_\mu(\widehat{v}) \to 0$ *at the optimal rate for all* $q \in \mathbb{N}$. *"May be optimal" means that* $\mathsf{MSE}_\mu(\widehat{v}) \to 0$ *at the optimal rate for all* $q \in \mathbb{N}$ *iff the kernel satisfies* $q' \geq q$; *see Assumption 6. An asterisk "*" means that additional regularity conditions on* K, $\{d_j\}$, h *or* ℓ *are needed*

Regimes	$0 < m < \infty$	$m \to \infty$
$h/\ell \to 0$	Inconsistent	Inconsistent or suboptimal*
$h/\ell \to \lambda_\infty \in (0,1)$	Inconsistent*	Inconsistent or suboptimal*
$h/\ell \to 1$	May be optimal*	Suboptimal*
$h/\ell \to \lambda_\infty \in (1,\infty)$	Optimal	Suboptimal*
$h/\ell \to \infty$	Suboptimal	Suboptimal

THEOREM 5.1 (Bias). *Let $\{X_i\}_{i\in\mathbb{Z}}$ be a stationary time series with $\{Z_i\}_{i\in\mathbb{Z}}$ satisfying Assumption 1 and $u_q < \infty$ for some $q \in \mathbb{N}$. Suppose $1/\ell + \ell/n \to 0$, $h/\ell = \lambda \in [1, \infty)$, and $m < \infty$.*

1. *Let $r_{\mathsf{bias}} = o(1/\ell^q) + O(\ell/n)$. Then*

$$\mathsf{Bias}_0(\widehat{v}) = \sum_{|k|\le\ell} \left\{ K_{\mathsf{diff}}\left(\frac{k}{\ell}\right) - 1 \right\} \gamma_k + r_{\mathsf{bias}}. \tag{5.1}$$

2. *If, in addition, $\lambda \in [2, \infty)$, and $K(\cdot)$ satisfies Assumption 5, then $\mathsf{Bias}(\widehat{v}) = B v_q/\ell^q + r_{\mathsf{bias}}$, where B is defined in Assumption 5, and $v_q = \sum_{k\in\mathbb{Z}} |k|^q \gamma_k$.*

THEOREM 5.2 (Variance). *Let $\{X_i\}_{i\in\mathbb{Z}}$ be a stationary time series with $\{Z_i\}_{i\in\mathbb{Z}}$ satisfying Assumption 1. Suppose $1/\ell + \ell/n \to 0$, $h/\ell = \lambda \in [2, \infty)$, and $m < \infty$. Let $A = \int_0^1 K^2(t)\,\mathrm{d}t$, $\Delta_m = \sum_{|s|\le m} \delta_s^2$ and $r_{\mathsf{se}}^2 = o(\ell/n)$. Then*

$$\mathsf{Var}_0(\widehat{v}) = \frac{4A\ell v^2 \Delta_m}{n} + r_{\mathsf{se}}^2.$$

The proofs of Theorems 5.1 and 5.2 can be found in Sections A.6 and A.7 of the supplement, respectively. Note that if $\ell = o\{n^{1/(1+q)}\}$, then $r_{\mathsf{bias}}^2 + r_{\mathsf{se}}^2 = o\{\mathsf{MSE}_0(\widehat{v})\}$. In this case, Theorem 5.1(2) and Theorem 5.2 imply that

$$\mathsf{MSE}_0(\widehat{v}) \sim \frac{B^2 v_q^2}{\ell^{2q}} + \frac{4A\ell v^2 \Delta_m}{n},$$

whose magnitude is controlled by the following factors:

- (Kernel K) The constants B^2 and A are determined by the user-specified kernel $K(\cdot)$. In practice, different users may use different kernels. Our theory is flexible enough to support general kernels, but the existing estimators in Example 2.4 only support the Bartlett kernel.
- (Serial dependence $\{\gamma_k\}$) The process-dependent constants v and v_q are functions of $\{\gamma_k\}_{k\in\mathbb{Z}}$. They govern the magnitudes of variance and squared bias, respectively. Although users cannot control them, it is possible to select the best ℓ to adapt to the observed dependence structure of $X_{1:n}$; see Section 5.2.
- (Lag h) The value of $h = \lambda\ell$ has a great impact on the bias. When λ is small ($\lambda \in [1, 2)$), the bias (5.1) admits no simple form. When λ is large enough ($\lambda \in [2, \infty)$), the bias is asymptotically unaffected by the differencing operation because of the matching property in Proposition 2.2 (4). We recommend $\lambda = 2$ due to finite-sample consideration.
- (Difference sequence $\{d_j\}$) In the regime $\lambda \in [2, \infty)$, the variance neatly depends on m and $\{d_j\}_{j=0}^m$ only through Δ_m. In practice, we should pick the $\{d_j\}$ to minimize the MSE; see Section 5.2.
- (Bandwidth ℓ and sample size n) The bandwidth ℓ affects the squared bias and variance in an opposite direction. A large ℓ leads to a large variance and a small squared bias. So, the well-known bias-variance tradeoff occurs. Section 5.2 discusses the selection of ℓ.
- (Reminder term r_{bias}) It has two parts: $o(1/\ell^q)$ and $O(\ell/n)$. The term $o(1/\ell^q)$ comes from approximating $\sum_{|k|\le\ell} |k/\ell|^q \gamma_k$ by v_q/ℓ^q. The other one comes from approximating $\widehat{v}$ by $\widehat{v}_{\mathsf{diff}}$; see Proposition 2.2.
- (Reminder term r_{se}) It appears when we apply the invariance principle (Theorem 3 of Wu (2011)) to approximate the moments of $\widehat{v}$ by the moments of a Brownian motion.

5.2. *Optimal parameters selection.* From now on, we use $\lambda = 2$ as it has the best properties. The optimal ℓ and $\{d_j\}$ are derived for each m below.

COROLLARY 5.3. *Suppose the conditions stated in Theorem 5.1(2) and Theorem 5.2 hold. In addition, assume* $v_q \neq 0$, $\ell = o\{n^{1/(1+q)}\}$ *and* $\lambda = 2$.

1. *The MSE-optimal* ℓ *is given by*

$$\ell^\star = \left\{ \frac{q(v_q/v)^2 B^2 n}{2A\Delta_m} \right\}^{1/(1+2q)}. \tag{5.2}$$

2. *For each* m, *the optimal* $\{d_j\}_{j=0}^m$, *denoted by* $\{d_j^\star\}_{j=0}^m$, *satisfies* $\delta_1 = \cdots = \delta_m = -1/(2m)$, *which implies* $\Delta_m = 1 + 1/(2m)$. *The solution* $\{d_j^\star\}_{j=0}^m$ *is a sequence of universal constants that depends on* m *only.*

The proof can be found in Section A.8 of the supplement. From Corollary 5.3, if the optimal $\ell^\star$ is used, the optimal MSE satisfies

$$n^{2q/(1+2q)}\,\mathsf{MSE}_0(\widehat{v})/v^2 \to (1+2q)\left\{ B^2 \left(\frac{2A\Delta_m}{q} \right)^{2q} (v_q/v)^2 \right\}^{1/(1+2q)} =: M_{(m)}, \tag{5.3}$$

which can be generalized to $n^{2q/(1+2q)}\,\mathsf{MSE}_\mu(\widehat{v})/v^2 \to M_{(m)}$ if $\mu \in \mathcal{M}_q$; see (3.1) for the definition of $\mathcal{M}_q$. So, the $\mathcal{L}^2$ convergence rate of $\widehat{v}$ is $n^{q/(1+2q)}$, which is the best possible.

If, in addition, the optimal $\{d_j^\star\}_{j=0}^m$ is used, we have $M_{(1)} > M_{(2)} > \cdots$. Indeed, for any $\epsilon > 0$, there is $m \in \mathbb{N}$ such that $|M_{(m)} - M_{(0)}| < \epsilon$, where $M_{(0)}$ is the best possible MSE achieved by the classical (nonrobust) estimator $\widehat{v}_{(0)}$. Hence, the proposed framework covers the optimal estimator asymptotically. The optimal $\{d_j^\star\}_{j=0}^m$ can be obtained numerically by the innovation algorithm (Definition 8.3.1 of Brockwell and Davis (2006)); see Table 1 for the solution. It is worth mentioning that the mth order local differencing in Example 2.1 is nearly optimal in the sense that $\Delta_m = 1 + (2m+1)/(3m^2+3m) \approx 1$ when m is large. It can be a good and convenient choice in practice.

For reference, we compare our proposed estimator $\widehat{v}_{(3)}$ (i.e., using $m = 3$) with the best existing robust estimator $\widehat{v}_{(\mathrm{C})}$ in Chan (2022a) (see Example 2.8) and the best proposal $\widehat{v}_{(\mathrm{W})}$ in Wu and Zhao (2007) (see Example 2.4). If $K = K_{\mathrm{Bart}}$, then $\widehat{v}_{(3)}$ uniformly dominates $\widehat{v}_{(\mathrm{C})}$ and $\widehat{v}_{(\mathrm{W})}$ in the sense that

$$\frac{\mathsf{MSE}_0\{\widehat{v}_{(\mathrm{C})}\}}{\mathsf{MSE}_0\{\widehat{v}_{(3)}\}} \to \left(\frac{12}{7}\right)^{2/3} \approx 1.43 \quad \text{and} \quad \frac{\mathsf{MSE}_0\{\widehat{v}_{(\mathrm{W})}\}}{\mathsf{MSE}_0\{\widehat{v}_{(3)}\}} \to \left(\frac{3}{2}\right)^{4/3} \approx 1.71; \tag{5.4}$$

see Section B.8 of the supplement for a detailed derivation.

Since $\ell^\star$ depends on the unknowns v_q and v_0, we need also to estimate v_q. Similar to (2.3), we propose to estimate v_p ($p \in \mathbb{N}_0$) by

$$\widehat{v}_p = \sum_{|k|<\ell} |k|^p K\left(\frac{k}{\ell}\right) \widehat{\gamma}_k^D.$$

We may write $\widehat{v}_p$ as $\widehat{v}_{p,(m)}$ to emphasize the order m. The following corollaries show that $\widehat{v}_p$ is a consistent and robust estimator of v_p.

COROLLARY 5.4 (Consistency). *Let* $p \in \mathbb{N}_0$ *and* $\{X_i\}_{i\in\mathbb{Z}}$ *be a stationary time series with* $\{Z_i\}_{i\in\mathbb{Z}}$ *satisfying Assumption 1 and* $u_{p+q} < \infty$ *for some* $q \in \mathbb{N}$. *Suppose* $K(\cdot)$ *satisfies Assumption 5,* $1/\ell + \ell^{1+2p}/n \to 0$, $h/\ell = \lambda \in [2, \infty)$ *and* $m < \infty$. *Then*

$$\mathsf{MSE}_0(\widehat{v}_p; v_p) = \left(\frac{B v_{p+q}}{\ell^q} + r_{p,\mathsf{bias}} \right)^2 + \left(\frac{4A_p \ell^{1+2p} v^2 \Delta_m}{n} + r_{p,\mathsf{se}}^2 \right),$$

where $r_{p,\mathsf{bias}} = o(1/\ell^q) + O(\ell^{1+p}/n)$; $r_{p,\mathsf{se}}^2 = o(\ell^{1+2p}/n)$; *the constant* B *is defined in Assumption 5; and* $A_p = \int_0^1 |t|^{2p} K^2(t)\,\mathrm{d}t$.

COROLLARY 5.5 (Robustness). *Assume the conditions in Theorem 3.1. Let* $p \in \mathbb{N}_0$, $R_{p,\mathsf{bias}} = O(\ell^p R_{\mathsf{bias}})$ *and* $R_{p,\mathsf{se}} = O(\ell^p R_{\mathsf{se}})$. *Then*

$$\mathsf{Bias}_\mu(\widehat{v}_p; v_p) = \mathsf{Bias}_0(\widehat{v}_p; v_p) + \left\{\kappa \ell^{1+p}\mathcal{V} + O\left(\frac{\ell^{1+p}}{n}\right)\right\}\mathbb{1}_{(m=0)} + R_{p,\mathsf{bias}},$$

$$\sqrt{\mathsf{Var}_\mu(\widehat{v}_p)} = \sqrt{\mathsf{Var}_0(\widehat{v}_p)} + O\left\{\frac{\ell^{1+p}(\mathcal{C}+\mathcal{SJ})}{\sqrt{n}}\right\}\mathbb{1}_{(m=0)} + R_{p,\mathsf{se}},$$

The proofs of Corollaries 5.4 and 5.5 can be found in Sections A.9 and A.10 of the supplement, respectively. By Corollary 5.4, we know that, in the constant mean case, $\widehat{v}_p$ converges in $\mathcal{L}^2$ optimally, that is, $\mathsf{MSE}_0(\widehat{v}_p; v_p) = O(n^{-2q/(1+2p+2q)})$, if $\ell \asymp n^{1/(1+2p+2q)}$. By Corollary 5.5, if $m \in \mathbb{N}$, $h/\ell = \lambda \in (0,\infty)$ and $\ell \asymp n^{1/(1+2p+2q)}$ are used for $\widehat{v}_p$, then $\widehat{v}_p$ is strictly robust in

$$\begin{aligned}\mathcal{M}_{p,q} &:= \{\mu(\cdot) : R_{p,\mathsf{bias}}^2 + R_{p,\mathsf{se}}^2 = o(\mathsf{MSE}_0(\widehat{v}_p; v_p))\}\\ &= \{\mu(\cdot) : \mathcal{C}^2 = o(n^{\frac{3p+3q-1}{1+2p+2q}}), \mathcal{S}^2\mathcal{J} = o(n^{\frac{p+q-1}{1+2p+2q}}), \mathcal{G} \gtrsim n^{\frac{1}{1+2p+2q}}\}.\end{aligned}$$

So, under $u_{p+q} < \infty$, we have $\mathsf{MSE}_\mu(\widehat{v}_p; v_p) \sim \mathsf{MSE}_0(\widehat{v}_p; v_p)$ for $\mu \in \mathcal{M}_{p,q}$.

6. Implementation issue and generalization.

6.1. *Rough centering procedure.* If there are obvious jumps and trends, one may *roughly* remove them. It improves finite-sample performance. We only *roughly* center $X_{1:n}$ because consistent centering either distorts the autocovariance structure or deteriorates the convergence rate of $\widehat{v}$; see Theorem 4.2. We propose a two-step *rough centering procedure* (RCP). The first and second steps remove obvious jumps and trends, respectively.

In step 1, we locate the N most obvious CP times $t_1, \ldots, t_N$, where $N \le N'$ for some $N' < \infty$, for example, $N' = 10$. We initialize $X_i^{(1)} = X_i$ for each i, and iterate the following steps for $k = 1, 2, \ldots$ Let $\xi_i^{(k)} = \sum_{i'=i}^{i+b-1} X_{i'}^{(k)}/b - \sum_{i'=i-b+1}^{i} X_{i'}^{(k)}/b$ be the local batch-mean difference at time i for $i = b, \ldots, n-b+1$, where $b = \lfloor n^{1/3} \rfloor$. An unusually large $\xi_i^{(k)}$ indicates that i is a potential CP. Denote the distance of $\xi_i^{(k)}$ from Tukey's fences by $O_i^{(k)} = \max\{0, \xi_i^{(k)} - 4Q_1^{(k)} + 3Q_3^{(k)}, 4Q_3^{(k)} - 3Q_1^{(k)} - \xi_i^{(k)}\}$, where $Q_1^{(k)}$ and $Q_3^{(k)}$ are the lower and upper quartiles of $\{\xi_i^{(k)}\}_{i=b}^{n-b+1}$, respectively. By Tukey's rule, if

$$I^{(k)} := \{i \in \{b, \ldots, n-b+1\} \setminus \{t_1, \ldots, t_{k-1}\} : O_i^{(k)} > 0\} \neq \varnothing,$$

we define the kth most obvious CP as $t_k = \arg\max_{i \in I^{(k)}} O_i^{(k)}$, and let $X_i^{(k+1)} = X_i^{(k)} - [X_{t_k}^{(k)} - X_{t_k-1}^{(k)}]_{-M}^{M}\mathbb{1}(i \ge t_k)$, where $[\cdot]_{-M}^{M} = \max\{-M, \min(\cdot, M)\}$ and $M = M'\{\sum_{i=2}^{n}(X_i - X_{i-1})^2/(2n)\}^{1/2}$ for some $M' \in \mathbb{R}^+$, for example, $M' = 100$. The iteration stops if $I^{(k)} = \varnothing$ or $k = N'$. So, the jump-removed series is

$$X_i^\dagger = X_i - \sum_{j \in \{1,\ldots,N\}: t_j \le i} [X_{t_j} - X_{t_j-1}]_{-M}^{M}. \tag{6.1}$$

In step 2, we run a segmented linear regression on $X^\dagger_{t_j},\dots,X^\dagger_{t_{j+1}-1}$ against $t_j,\dots,t_j-1$ for each $j=0,\dots,N$, where $t_0=1$ and $t_{N+1}=n+1$. After shifting the segmented regression lines to ensure continuity, we obtain

$$X^\ddagger_i = X^\dagger_i - \sum_{j=0}^{N}\{\widehat{\beta}_{j,0}+\widehat{\beta}_{j,1}(i-t_j)\}\mathbb{1}(t_j\le i\le t_{j+1}-1), \tag{6.2}$$

where, for each $j=0,\dots,N$, $\widehat{\beta}_{j,1}=\widehat{\alpha}_{j,1}$, $\widehat{\beta}_{j,0}=\sum_{j'=0}^{j-1}\widehat{\alpha}_{j',1}(t_{j'+1}-1-t_{j'})$, $(\widehat{\alpha}_{j,0},\widehat{\alpha}_{j,1})^\intercal=(\mathbb{Z}_j^\intercal\mathbb{Z}_j)^{-1}\mathbb{Z}_j^\intercal\mathbb{X}_j$; and $\mathbb{X}_j=(X^\dagger_{t_j},\dots,X^\dagger_{t_{j+1}-1})^\intercal$, $\mathbb{Z}_j$ is a $(t_{j+1}-t_j)\times 2$ matrix whose first column is a vector of 1 and the second column is $[0,\dots,t_{j+1}-t_j-1]^\intercal$. Applying $\widehat{v}_p$ on the roughly centered time series $\{X^\ddagger_i\}$, we obtain a finite-sample-adjusted estimator $\widehat{v}^\ddagger_p$. The following corollary ensures that the RCP does not inflate the asymptotic MSE.

COROLLARY 6.1. *Assume the conditions in Corollary 5.4. If $p\in\mathbb{N}_0$, $q\in\mathbb{N}$, $m\in\mathbb{N}$ and $\ell=O(n^{1/(1+2p+2q)})$, then $\mathsf{MSE}_\mu(\widehat{v}^\ddagger_p;v_p)\sim\mathsf{MSE}_\mu(\widehat{v}_p;v_p)$ for any $\mu\in\mathcal{M}_{p,q}$.*

The proof can be found in Section A.11 of the supplement. For clarity, denote $\widehat{v}_p=\widehat{v}_p(X_{1:n};\ell,K,m,d)$ if data $X_{1:n}$, bandwidth $\lceil\ell\rceil$, kernel $K(\cdot)$, order m, difference sequence $\{d_j\}_{j=0}^m$ and $\lambda=h/\ell=2$ are used. In practice, we always use the (universally) optimal difference sequence $\{d^\star_j\}$ in Table 1. The suggested estimator of the LRV v is

$$\widehat{v}^\star=\widehat{v}_0(X^\ddagger_{1:n};\widehat{\ell}^\star,K,m,d^\star),\quad \text{where } \widehat{\ell}^\star=\left\{\frac{q(\widehat{v}^\sharp_q/\widehat{v}^\sharp)^2B^2n}{2A\Delta_m}\right\}^{1/(1+2q)}, \tag{6.3}$$

where $\widehat{v}^\sharp_q=\widehat{v}_q(X^\ddagger_{1:n};2n^{1/(5+2q)},K_2,m,d^\star)$ and $\widehat{v}^\sharp=\widehat{v}_0(X^\ddagger_{1:n};2n^{1/5},K_2,m,d^\star)$ are pilot estimators of v_q and v, respectively. We recommend $m=3$ and Parzen's (1957) kernel $K_q(t)=(1-|t|^q)^+$ with $q=2$; see Section 7 for some simulation evidence.

6.2. *Multivariate time series.* We generalize (1.1) to the multivariate setting such that $\{Z_i\}_{i\in\mathbb{Z}}$ is a sequence of S-dimensional zero-mean stationary noises with $\gamma_k=\mathsf{E}(Z_0Z_k^\intercal)$; and $\mu_i=\mu(i/n)\in\mathbb{R}^S$ are S-dimensional signals. We also use the decomposition in (1.3) but $\xi_0,\dots,\xi_{\mathcal{J}}\in\mathbb{R}^S$. Denote the sth element of a vector e by $e^{[s]}$, and the (r,s)th element of a matrix E by $E^{[r,s]}$.

We assume $Z_i=g(\dots,\varepsilon_{i-1},\varepsilon_i)$, where $\{\varepsilon_i\}_{i\in\mathbb{Z}}$ are i.i.d. $\mathbb{R}^{S'}$-dimensional innovations. Generalize (1.4) to $\theta^{[s]}_{p,i}:=\|Z^{[s]}_i-Z^{[s]}_{i,\{0\}}\|_p$ and $\Theta^{[s]}_p:=\sum_{i=0}^\infty\theta^{[s]}_{p,i}$, for $s=1,\dots,S$. Assumption 1 is generalized as follows.

ASSUMPTION 1* (Weak dependence). The S-dimensional zero-mean strictly stationary $\{Z_i\}$ satisfies $\mathsf{E}(Z^{[s]}_1)^4<\infty$ and $\Theta^{[s]}_4<\infty$ for all $s=1,\dots,S$.

The quantity of interest is the long-run variance (-covariance matrix) $v=\lim_{n\to\infty}n\mathsf{E}(\bar{Z}_n\bar{Z}_n^\intercal)=\sum_{k\in\mathbb{Z}}\gamma_k\in\mathbb{R}^{S\times S}$. Our proposed estimator of v admits the same form as (2.3) with $\widehat{\gamma}^D_k=\sum_{i=mh+|k|+1}^n D_iD_{i-|k|}^\intercal/n$. We define $v_p=\sum_{k\in\mathbb{Z}}|k|^p\gamma_k$ and $u_p=\sum_{k\in\mathbb{Z}}|k|^p|\gamma_k|$, where $|\gamma_k|$ is the entrywise absolute value of γ_k. Let w be a $S\times S$ matrix whose (r,s)th element is

$$w^{[r,s]}=\big(v^{[r,r]}v^{[s,s]}+v^{[r,s]}v^{[r,s]}\big)/2,\quad r,s\in\{1,\dots,S\}.$$

COROLLARY 6.2. *The results in Corollaries 5.4 and 5.5 remain valid if $(\widehat{v}_p,v_p,v_{p+q},v^2)$ is replaced by $(\widehat{v}^{[r,s]}_p,v^{[r,s]}_p,v^{[r,s]}_{p+q},w^{[r,s]})$ for each $r,s\in\{1,\dots,S\}$, provided that Assumption 1 is replaced by Assumption 1*, and $u_{p+q}<\infty$ is replaced by $u^{[r,s]}_{p+q}<\infty$ for all r,s.*

The proof can be found in Section A.12 of the supplement.

7. Experiments, applications and real-data examples. We consider a nonlinear time series model for all simulation studies in this section. Let $\{Z_i'\}_{i=1}^n$ be generated from a threshold autoregressive (TAR) model:

$$Z_i' = \begin{cases} \theta_1 Z_{i-1}' + \varepsilon_i & \text{if } Z_{i-1}' \geq 0; \\ \theta_2 Z_{i-1}' + \varepsilon_i & \text{if } Z_{i-1}' < 0, \end{cases} \tag{7.1}$$

where θ_1, θ_2 are the AR parameters in regimes 1 and 2, respectively, and ε_i follow $\mathrm{N}(0, 1)$ independently. We use $\theta_1 \in \{0.1, \ldots, 0.9\}$ and $\theta_2 = 0.5$. Let $Z_i = Z_i'/\sqrt{v}$, where $v = v_{\theta_1,\theta_2}$ is the LRV for the time series (7.1). So, $\{Z_i\}$ is stationary and satisfies Assumption 1; see Wu (2011).

7.1. *Efficiency and robustness.* Let $\mu(t) = \Xi\{e^t + \mathbb{1}(t > 0.3) + 2\mathbb{1}(t > 0.6) + 4\mathbb{1}(t > 0.8)\}$, where $\Xi \in \{0, 1, \ldots, 4\}$. When $\Xi \neq 0$, it contains three jumps and an exponentially increasing trend. We study the following estimators of v: (i) $\widehat{v}_{(\mathrm{W})}$ the best proposal in Wu and Zhao (2007) (see Example 2.4), (ii) $\widehat{v}_{(\mathrm{D})}$ the OBM estimator with Dehling, Fried and Wendler's (2020) adjustment (see Example 2.6), (iii) $\widehat{v}_{(\mathrm{C})}$ the suggested estimator in Chan (2022a) (see Example 2.8), (iv) $\widehat{v}_{(0)}^{\star}$ a classical estimator (see Example 2.3) and (v) $\widehat{v}_{(m)}^{\star}$ the proposed estimator with $m = 1, 2, 3$. We use $K(t) = (1 - t^2)^+$ in $\widehat{v}_{(0)}^{\star}, \ldots, \widehat{v}_{(3)}^{\star}$. We compare their (a) efficiency and (b) robustness via 10^4 replications.

For (a), the values of $\mathsf{MSE}_0(\cdot)$ are reported for different values of θ_1. For (b), the values of $\mathsf{MSE}_\mu(\cdot)$ are reported for different values of Ξ when $\theta_1 = 0.4$. The results when $n = 200$ are plotted in Figure 4. Our worst proposal $\widehat{v}_{(1)}^{\star}$ is already considerably more efficient than the existing estimators. The improvement of $\widehat{v}_{(2)}^{\star}$ over $\widehat{v}_{(1)}^{\star}$ is substantial. Although the improvement of $\widehat{v}_{(3)}^{\star}$ over $\widehat{v}_{(2)}^{\star}$ looks incremental, the advantage of $\widehat{v}_{(3)}^{\star}$ becomes obvious when n increases; see Figure C.2 of the supplement for the results when $n = 400$. When $\Xi \neq 0$, the existing estimators still show a certain degree of robustness relative to the non-robust $\widehat{v}_{(0)}^{\star}$. But their MSEs are substantially affected. All of our proposed estimators are of nearly constant risk for all Ξ. It is remarked that the asymptotic relative efficiency does not improve significantly when $m \geq 4$ because

$$\frac{\mathsf{MSE}(\widehat{v}_{(2)})}{\mathsf{MSE}(\widehat{v}_{(1)})} - 1 \approx -13.6\%, \qquad \frac{\mathsf{MSE}(\widehat{v}_{(3)})}{\mathsf{MSE}(\widehat{v}_{(2)})} - 1 \approx -5.4\%, \qquad \frac{\mathsf{MSE}(\widehat{v}_{(4)})}{\mathsf{MSE}(\widehat{v}_{(3)})} - 1 \approx -2.9\%.$$

A similar finding is also documented in Hall, Kay and Titterington (1990). Moreover, an excessively large m may affect the robustness of $\widehat{v}_{(m)}$ in finite samples. Hence, in practice, we suggest using $m = 3$.

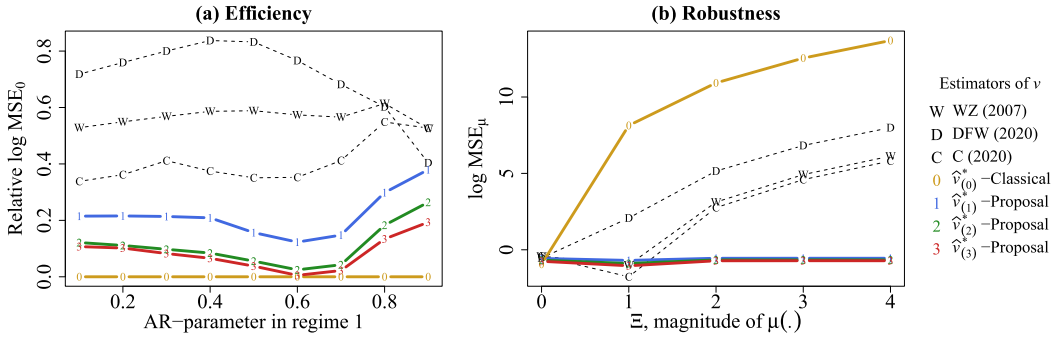

FIG. 4. *(a) The values of* $\log\{\mathsf{MSE}_0(\cdot)/\mathsf{MSE}_0(\widehat{v}_{(0)}^{\star})\}$ *when* $\mu_1 = \cdots = \mu_n = 0$. *(b) The values of* $\log \mathsf{MSE}_\mu(\cdot)$ *against* Ξ (*i.e., the magnitude of the mean function*).

For reference, we also compute the following estimators: (vi) the estimators in Wu and Zhao (2007) that use the sum of absolute (SA) differences and the median of absolute (MA) differences (see Example 2.7), (vii) Altissimo and Corradi's (2003) estimator (see Example 2.5), (viii) Crainiceanu and Vogelsang's (2007) estimator (see Example 2.6) and (ix) Juhl and Xiao's (2009) estimator (see Example 2.5). Their performances are obviously worse than other estimators either in terms of efficiency or robustness; see Figure C.1 of the supplement. Additional simulation experiments that further investigate the robustness (in terms of $\mathcal{S}$ and $\mathcal{C}$) can be found in Section D of the supplement.

7.2. *Hypothesis tests for structural breaks.* An estimator of v is needed in many hypothesis testing problems. We present two examples here: (a) the KS CP test, and (b) a structural break test in the presence of trends (Wu and Zhao (2007)). The test statistic (a) is defined in (1.2). The test statistic (b) is defined as

$$T_n(v) = \frac{1}{k_n\sqrt{v}} \max_{k_n \le i \le n-k_n} \left| \sum_{j=i+1}^{k_n+i} X_j - \sum_{j=i-k_n+1}^{i} X_j \right|, \tag{7.2}$$

where $k_n = n^\beta$ and $1/2 < \beta < 2/3$. The estimators (i)–(v) with $m = 3$ described in Section 7.1 are used to estimate the v in (1.2) and (7.2).

In reality, when a structural break occurs, the mean may suddenly jump to a high level but return to a lower level after that. So, we consider $\mu(t) = \Xi\{10\mathbb{1}(t > 0.3) - 9\mathbb{1}(t > 0.35)\}$. The mean jumps from 0 to 10Ξ at $t = 0.3$ and drops to Ξ at $t = 0.35$. Figure 5 shows the powers and size-adjusted powers of the tests (a) and (b) when $n = 200$ and the nominal size is 5%.

First, the sizes (type-I error rates) of the tests with $\widehat{v}_{(\mathrm{W})}$, $\widehat{v}_{(\mathrm{D})}$ or $\widehat{v}_{(\mathrm{C})}$ are not well controlled at the nominal value because these estimators of v are not accurate under the null hypothesis $H_0 : \mu(t) \equiv 0$. Second, the powers of the tests with $\widehat{v}_{(\mathrm{W})}$, $\widehat{v}_{(\mathrm{D})}$ and $\widehat{v}_{(\mathrm{C})}$ are low because these

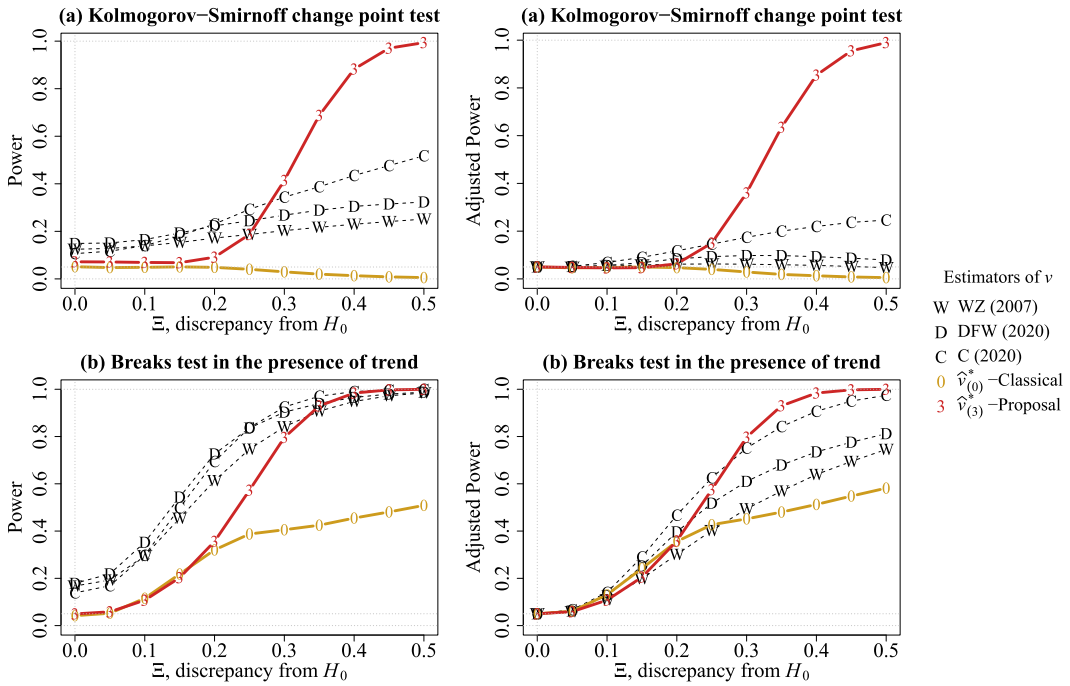

FIG. 5. *The powers and adjusted powers of different CP tests with various estimators of v.*

estimators are not robust against μ under the alternative hypothesis H_1. For test (a) with $\widehat{v}_{(\mathrm{W})}$ and $\widehat{v}_{(\mathrm{D})}$, it even fails to be monotonically powerful with respect to Ξ. It means that it is harder to reject a more obviously wrong H_1. Clearly, the tests (a) and (b) with our proposed estimator $\widehat{v}^{\star}_{(3)}$ control the size well and are monotonically powerful.

7.3. *Simultaneous confidence bands for trends*. An estimator of v is an essential component in many automatic procedures, for example, bandwidth selection for nonparametric regression, and construction of SCBs, etc. In this section, we present the local linear regression estimator (Wu and Zhao (2007)):

$$\widehat{\mu}_b(t) = 2\bar{\mu}_b(t) - \bar{\mu}_{b\sqrt{2}}(t), \quad \text{where } \bar{\mu}_b(t) = \sum_{i=1}^{n} \frac{H\{(t-i/n)/b\}}{nb} X_i, \tag{7.3}$$

and $H(\cdot)$ is a kernel, for example, the Gaussian kernel, and b is a bandwidth. They suggest that b could be selected as $b_n = 2(\widehat{v}/\widehat{\gamma}_0)^{1/5} b^*$, where $\widehat{\gamma}_0 = \sum_{i=1}^{n}\{X_i - \widehat{\mu}_{b^*}(i/n)\}^2/n$, b^* is the optimal bandwidth under the i.i.d. assumption (Ruppert, Sheather and Wand (1995)), and $\widehat{v}$ is an estimator of v. Since b_n is crucial to the performance of $\widehat{\mu}_{b_n}(\cdot)$, an efficient estimator of v is important.

SCBs for $\mu(\cdot)$ directly depend on $\widehat{v}$. In particular, 95% SCBs are given by $\widehat{\mu}_{b_n}(t) \pm \sqrt{\widehat{v}} q_{0.95}$, where $q_{0.95}$ is the 95% quantile of $\sup_{0\le t\le 1} |\widehat{\mu}^{\circ}_{b^{\circ}_n}(t)|$ with $\widehat{\mu}^{\circ}_{b^{\circ}_n}(t)$ and b°_n computed on i.i.d. data $X^{\circ}_1, \ldots, X^{\circ}_n \sim \mathrm{N}(0,1)$. The quantile $q_{0.95}$ can be easily obtained from simulation; see Table 2 of Wu and Zhao (2007). A simulation experiment is performed to compare the coverage probability and the expected half-width of the 95% SCBs when $n = 200$ and the true mean function is $\mu(t) = \cos(2\pi t)$. We try different bandwidth $0.05 \le b \le 0.1$. The results are shown in Table 4. The SCBs with $\widehat{v}^{\star}_{(0)}$ or $\widehat{v}_{(\mathrm{D})}$ are overcovered, and their expected half-widths are large. The SCBs with $\widehat{v}_{(\mathrm{W})}$ or $\widehat{v}_{(\mathrm{C})}$ are undercovered. The SCBs with our proposed $\widehat{v}^{\star}_{(3)}$ have a quite accurate coverage rate and a reasonable expected width.

7.4. *Southern hemispheric land and ocean temperature*. The earth's surface temperature has been an actively discussed topic in various fields. This section studies the southern hemispheric land and ocean monthly temperature from 1880 to 2018 ($n = 139 \times 12$). The data set is freely accessible from the website of NOAA's National Centers for Environmental Information.

Since land temperature changes more rapidly than ocean temperature, the land temperature is expected to be more volatile. The LRV v is a measure of the stochastic variability of the average. Using $\widehat{v}^{\star}_{(3)}$, the long-run standard deviations ($\sqrt{v}$) for the land and ocean series are about 0.525 and 0.313, respectively. We can also compute the long-run correlation between

TABLE 4

The coverage probabilities of 95% SCBs for $\mu(\cdot)$ under different bandwidth b and different estimators of v. The numbers inside parentheses are the expected half-widths

b	WZ (2007)	DFW (2020)	C (2020)	$\widehat{v}^{\star}_{(0)}$ (Classical)	$\widehat{v}^{\star}_{(3)}$ (Proposal)
0.05	92.0%(1.4)	99.4%(2.1)	91.8%(1.4)	100%(3.1)	94.9%(1.5)
0.06	91.7%(1.2)	99.2%(1.9)	91.7%(1.2)	100%(2.8)	94.7%(1.3)
0.07	91.4%(1.2)	99.1%(1.8)	91.3%(1.1)	100%(2.6)	94.4%(1.2)
0.08	91.4%(1.1)	99.1%(1.6)	91.3%(1.1)	100%(2.4)	94.4%(1.2)
0.09	91.3%(1.0)	99.1%(1.5)	91.2%(1.0)	100%(2.3)	94.3%(1.1)
0.1	90.9%(1.0)	99.0%(1.4)	91.1%(1.0)	100%(2.2)	94.2%(1.0)

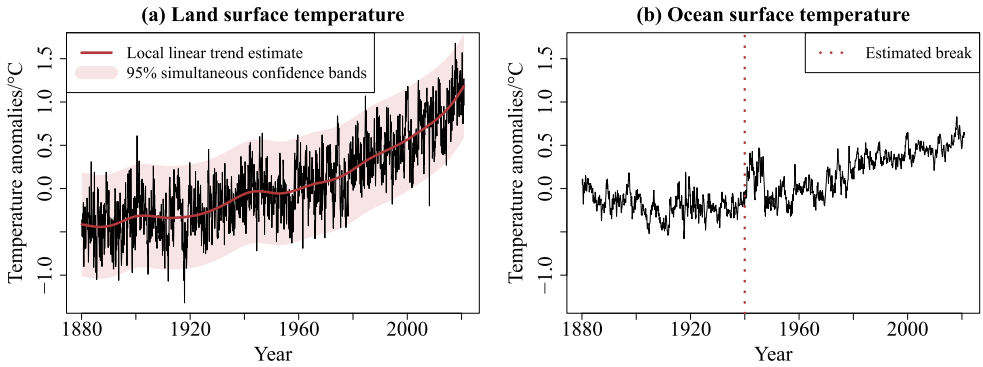

FIG. 6. (*a*) *Trend estimate and* 95% *SCBs for the mean of the land surface temperature.* (*b*) *Estimated break location for the mean of the ocean surface temperature.*

the land and ocean average temperature. It is about 0.651, which indicates a moderately strong correlation. If the nonrobust $\widehat{v}^{\star}_{(0)}$ is used, it is inflated to 0.966 as $\widehat{v}^{\star}_{(0)}$ mistakenly regards comovement of trends as correlation.

Next, we test whether the global temperature has a structural break even it has a possibly increasing trend. The test in Wu and Zhao (2007) is used. The p-values for the land series and ocean series are 15.4% and $< 10^{-5}$, respectively. Since the land mean temperature is smooth, we produce a trend estimate together with 95% SCBs in Figure 6(a). The bandwidth selected for $\widehat{\mu}_{b_n}(\cdot)$ in (7.3) is $b_n = 2(0.276/0.062)^{1/5}(0.017) \approx 0.046$. For the ocean temperature, the structural break is detected at around 1940 using the break location estimator proposed in Wu and Zhao (2007); see Figure 6(b). We suspect that it is easier to detect a structural break in the ocean series because the ocean temperature has a smaller LRV relative to the variation of $\mu(\cdot)$.

8. Summary, discussion and future work. This article presents a general class of difference-based estimators $\widehat{v}$ (2.3) for the long-run variance (1.5). Many existing estimators are special cases. We derive the regimes in which $\widehat{v}$ is consistent and rate optimal; see Table 3. In particular, the intuitive estimator with locally centered $X_{1:n}$ is inadmissible. We also derive detailed $\mathcal{L}^2$ properties of $\widehat{v}$. It is proven to be asymptotically optimal even in the presence of trends and a possibly divergent number of change points. The suggested estimator is stated in (6.3). We list some possible future work below:

- From Theorem 3.1, a possible estimator of $\mathcal{V} = \int_0^1 \{\mu(t) - \bar{\mu}\}^2 \,\mathrm{d}t$ is $\widehat{\mathcal{V}}_{(m)} = \{\widehat{v}_{(0)} - \widehat{v}_{(m)}\}/\{\ell \int_{-1}^{1} K(t)\,\mathrm{d}t\}$ for some $m > 0$; see To and Chan (2022). It can be used to test constancy of $\mu(\cdot)$, that is, $H_0 : \mathcal{V} = 0$. The resulting test is expected to perform well because the estimator $\widehat{v}_{(m)}$ has a small MSE.
- To test for a variance change point in the presence of a nonconstant mean, a possible test statistic can be constructed by using $\widehat{Q}_{(m)} = \sum_i D_i^2/n$ with parameters chosen to minimize the long-run variance of $\log \widehat{Q}_{(m)}$, where D_i is our proposed mth order difference statistics; see Leung and Chan (2022). This test could be used as an alternative to the recent work by Gao et al. (2019).

All of these directions rely on the optimal framework proposed in this article.

Acknowledgments. The authors would like to thank the anonymous referees, an associate editor and the editor for their constructive comments that improved the scope and presentation of the paper.

Funding. This research was partially supported by grants GRF-2130730 and GRF-2130788 provided by Research Grants Council of HKSAR.

SUPPLEMENTARY MATERIAL

Supplement to "Optimal difference-based variance estimators in time series: A general framework" (DOI: 10.1214/21-AOS2154SUPP; .pdf). Appendix A: Proofs of main results. The proofs of Propositions 2.1, 2.2, Theorems 3.1, 4.1, 4.2, 5.1, 5.2, Corollaries 5.3, 5.4 and Corollaries 6.1, 6.2 are placed in Sections A.1–A.12, respectively. Appendix B: Auxiliary results. Technical results of independent interest are stated in Sections B.1–B.7. The derivation of (5.4) is stated in Section B.8. Appendix C: Additional plots. It contains additional simulation results for Section 7.1. Appendix D: Additional simulation experiments. It contains additional simulation experiments about the robustness against mean functions.

REFERENCES

ALEXOPOULOS, C., GOLDSMAN, D. and WILSON, J. R. (2011). Overlapping batch means: Something more for nothing? *Winter Simul. Conf.* 401–411.

ALTISSIMO, F. and CORRADI, V. (2003). Strong rules for detecting the number of breaks in a time series. *J. Econometrics* **117** 207–244. MR2008769 https://doi.org/10.1016/S0304-4076(03)00147-7

ANDERSON, T. W. (1971). *The Statistical Analysis of Time Series*. Wiley, New York. MR0283939

ANDREWS, D. W. K. (1991). Heteroskedasticity and autocorrelation consistent covariance matrix estimation. *Econometrica* **59** 817–858. MR1106513 https://doi.org/10.2307/2938229

BRADLEY, R. C. (2005). Basic properties of strong mixing conditions. A survey and some open questions. *Probab. Surv.* **2** 107–144. MR2178042 https://doi.org/10.1214/154957805100000104

BROCKWELL, P. J. and DAVIS, R. A. (2006). *Time Series*: *Theory and Methods*. *Springer Series in Statistics*. Springer, New York. MR2839251

CARLSTEIN, E. (1986). The use of subseries values for estimating the variance of a general statistic from a stationary sequence. *Ann. Statist.* **14** 1171–1179. MR0856813 https://doi.org/10.1214/aos/1176350057

CHAN, K. W. (2022a). Mean-Structure and Autocorrelation Consistent Covariance Matrix Estimation. *J. Bus. Econom. Statist.* **40** 201–215. MR4356567 https://doi.org/10.1080/07350015.2020.1796397

CHAN, K. W. (2022b). Supplement to "Optimal difference-based variance estimators in time series: A general framework." https://doi.org/10.1214/21-AOS2154SUPP

CHAN, K. W. and YAU, C. Y. (2017a). Automatic optimal batch size selection for recursive estimators of time-average covariance matrix. *J. Amer. Statist. Assoc.* **112** 1076–1089. MR3735361 https://doi.org/10.1080/01621459.2016.1189337

CHAN, K. W. and YAU, C. Y. (2017b). High-order corrected estimator of asymptotic variance with optimal bandwidth. *Scand. J. Stat.* **44** 866–898. MR3730019 https://doi.org/10.1111/sjos.12279

CHEN, H. and SCHMEISER, B. (2013). I-SMOOTH: Iteratively smoothing mean-constrained and nonnegative piecewise-constant functions. *INFORMS J. Comput.* **25** 432–445. MR3085324 https://doi.org/10.1287/ijoc.1120.0512

CHEN, L., WANG, W. and WU, W. B. (2021). Inference of breakpoints in high-dimensional time series. *J. Amer. Statist. Assoc.* To appear.

CHENG, C. H. and CHAN, K. W. (2022). *A general framework for constructing locally self-normalized multiple-change-point tests*. Manuscript.

CRAINICEANU, C. M. and VOGELSANG, T. J. (2007). Nonmonotonic power for tests of a mean shift in a time series. *J. Stat. Comput. Simul.* **77** 457–476. MR2405424 https://doi.org/10.1080/10629360600569394

CSÖRGŐ, M. and HORVÁTH, L. (1997). *Limit Theorems in Change-Point Analysis*. *Wiley Series in Probability and Statistics*. Wiley, Chichester. MR2743035

DALLA, V., GIRAITIS, L. and PHILLIPS, P. C. B. (2015). *Testing Mean Stability of Heteroskedastic Time Series*. Manuscript.

DEHLING, H., FRIED, R. and WENDLER, M. (2020). A robust method for shift detection in time series. *Biometrika* **107** 647–660. MR4138981 https://doi.org/10.1093/biomet/asaa004

DETTE, H., ECKLE, T. and VETTER, M. (2020). Multiscale change point detection for dependent data. *Scand. J. Stat.* **47** 1243–1274. MR4178193 https://doi.org/10.1111/sjos.12465

DETTE, H. and WU, W. (2019). Detecting relevant changes in the mean of nonstationary processes—a mass excess approach. *Ann. Statist.* **47** 3578–3608. MR4025752 https://doi.org/10.1214/19-AOS1811

GALLANT, A. R. (1987). *Nonlinear Statistical Models*. *Wiley Series in Probability and Mathematical Statistics*: *Applied Probability and Statistics*. Wiley, New York. MR0921029 https://doi.org/10.1002/9780470316719

GAO, Z., SHANG, Z., DU, P. and ROBERTSON, J. L. (2019). Variance change point detection under a smoothly-changing mean trend with application to liver procurement. *J. Amer. Statist. Assoc.* **114** 773–781. MR3963179 https://doi.org/10.1080/01621459.2018.1442341

GONÇALVES, S. and WHITE, H. (2002). The bootstrap of the mean for dependent heterogeneous arrays. *Econometric Theory* **18** 1367–1384. MR1945417 https://doi.org/10.1017/S0266466602186051

GÓRECKI, T., HORVÁTH, L. and KOKOSZKA, P. (2018). Change point detection in heteroscedastic time series. *Econom. Stat.* **7** 63–88. MR3824127 https://doi.org/10.1016/j.ecosta.2017.07.005

HALL, P., KAY, J. W. and TITTERINGTON, D. M. (1990). Asymptotically optimal difference-based estimation of variance in nonparametric regression. *Biometrika* **77** 521–528. MR1087842 https://doi.org/10.1093/biomet/77.3.521

HORVÁTH, L., KOKOSZKA, P. and STEINEBACH, J. (1999). Testing for changes in multivariate dependent observations with an application to temperature changes. *J. Multivariate Anal.* **68** 96–119. MR1668911 https://doi.org/10.1006/jmva.1998.1780

IBRAGIMOV, I. A. (1962). Some limit theorems for stationary processes. *Teor. Veroyatn. Primen.* **7** 361–392. MR0148125

JUHL, T. and XIAO, Z. (2009). Tests for changing mean with monotonic power. *J. Econometrics* **148** 14–24. MR2494814 https://doi.org/10.1016/j.jeconom.2008.08.020

KÜNSCH, H. R. (1989). The jackknife and the bootstrap for general stationary observations. *Ann. Statist.* **17** 1217–1241. MR1015147 https://doi.org/10.1214/aos/1176347265

LEUNG, C. W. D. and CHAN, K. W. (2022). *Testing for variance changes under varying mean and serial correlation*. Manuscript.

LEVINE, M. and TECUAPETLA-GÓMEZ, I. (2019). ACF estimation via difference schemes for a semiparametric model with m-dependent errors. ArXiv Preprint. Available at arXiv:1905.04578.

LOBATO, I. N. (2001). Testing that a dependent process is uncorrelated. *J. Amer. Statist. Assoc.* **96** 1066–1076. MR1947254 https://doi.org/10.1198/016214501753208726

NEWEY, W. K. and WEST, K. D. (1987). A simple, positive semidefinite, heteroskedasticity and autocorrelation consistent covariance matrix. *Econometrica* **55** 703–708. MR0890864 https://doi.org/10.2307/1913610

PARZEN, E. (1957). On consistent estimates of the spectrum of a stationary time series. *Ann. Math. Stat.* **28** 329–348. MR0088833 https://doi.org/10.1214/aoms/1177706962

PEŠTA, M. and WENDLER, M. (2020). Nuisance-parameter-free changepoint detection in non-stationary series. *TEST* **29** 379–408. MR4095034 https://doi.org/10.1007/s11749-019-00659-1

POLITIS, D. N. (2011). Higher-order accurate, positive semidefinite estimation of large-sample covariance and spectral density matrices. *Econometric Theory* **27** 703–744. MR2822363 https://doi.org/10.1017/S0266466610000484

POLITIS, D. N. and ROMANO, J. P. (1995). Bias-corrected nonparametric spectral estimation. *J. Time Series Anal.* **16** 67–103. MR1323618 https://doi.org/10.1111/j.1467-9892.1995.tb00223.x

POLITIS, D. N., ROMANO, J. P. and WOLF, M. (1999). *Subsampling*. *Springer Series in Statistics*. Springer, New York. MR1707286 https://doi.org/10.1007/978-1-4612-1554-7

RICE, J. (1984). Bandwidth choice for nonparametric regression. *Ann. Statist.* **12** 1215–1230. MR0760684 https://doi.org/10.1214/aos/1176346788

ROSENBLATT, M. (1956). A central limit theorem and a strong mixing condition. *Proc. Natl. Acad. Sci. USA* **42** 43–47. MR0074711 https://doi.org/10.1073/pnas.42.1.43

RUPPERT, D., SHEATHER, S. J. and WAND, M. P. (1995). An effective bandwidth selector for local least squares regression. *J. Amer. Statist. Assoc.* **90** 1257–1270. MR1379468

SHAO, X. (2010). A self-normalized approach to confidence interval construction in time series. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **72** 343–366. MR2758116 https://doi.org/10.1111/j.1467-9868.2009.00737.x

SHAO, X. (2015). Self-normalization for time series: A review of recent developments. *J. Amer. Statist. Assoc.* **110** 1797–1817. MR3449074 https://doi.org/10.1080/01621459.2015.1050493

SONG, W. T. and SCHMEISER, B. W. (1995). Optimal mean-squared-error batch sizes. *Manage. Sci.* **41** 110–123.

TAQQU, M. and EBERLEIN (1986). *Dependence in Probability and Statistics*. Birkhäuser, Basel.

TECUAPETLA-GÓMEZ, I. and MUNK, A. (2017). Autocovariance estimation in regression with a discontinuous signal and m-dependent errors: A difference-based approach. *Scand. J. Stat.* **44** 346–368. MR3658518 https://doi.org/10.1111/sjos.12256

TO, H. K. and CHAN, K. W. (2022). *Mean stationarity test in time series*: *A signal variance-based approach*. Manuscript.

VATS, D. and FLEGAL, J. M. (2021). Lugsail lag windows for estimating time-average covariance matrices. *Biometrika*. To appear.

VOLKONSKIĬ, V. A. and ROZANOV, Y. A. (1959). Some limit theorems for random functions. I. *Theory Probab. Appl.* **4** 178–197. MR0121856 https://doi.org/10.1137/1104015

WELCH, P. D. (1987). On the relationship between batch means, overlapping batch means and spectral estimation. *Winter Simul. Conf.* 320–323.

WHITE, H. (1984). *Asymptotic Theory for Econometricians*. Academic Press, New York.

WU, W. B. (2004). A test for detecting changes in mean. In *Time Series Analysis and Applications to Geophysical Systems* (D. R. Brillinger, E. A. Robinson and F. Schoenberg, eds.) **139** 105–122. Springer, New York.

WU, W. B. (2005). Nonlinear system theory: Another look at dependence. *Proc. Natl. Acad. Sci. USA* **102** 14150–14154. MR2172215 https://doi.org/10.1073/pnas.0506715102

WU, W. B. (2007). Strong invariance principles for dependent random variables. *Ann. Probab.* **35** 2294–2320. MR2353389 https://doi.org/10.1214/009117907000000060

WU, W. B. (2011). Asymptotic theory for stationary processes. *Stat. Interface* **4** 207–226. MR2812816 https://doi.org/10.4310/SII.2011.v4.n2.a15

WU, W. B., WOODROOFE, M. and MENTZ, G. (2001). Isotonic regression: Another look at the changepoint problem. *Biometrika* **88** 793–804. MR1859410 https://doi.org/10.1093/biomet/88.3.793

WU, W. B. and ZHAO, Z. (2007). Inference of trends in time series. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **69** 391–410. MR2323759 https://doi.org/10.1111/j.1467-9868.2007.00594.x

YAU, C. Y. and CHAN, K. W. (2016). New recursive estimators of the time-average variance constant. *Stat. Comput.* **26** 609–627. MR3489860 https://doi.org/10.1007/s11222-015-9548-7

ZHANG, T. and LAVITAS, L. (2018). Unsupervised self-normalized change-point testing for time series. *J. Amer. Statist. Assoc.* **113** 637–648. MR3832215 https://doi.org/10.1080/01621459.2016.1270214

ZHAO, Z. (2011). A self-normalized confidence interval for the mean of a class of nonstationary processes. *Biometrika* **98** 81–90. MR2804211 https://doi.org/10.1093/biomet/asq076