Econometrics and Statistics 27 (2023) 36–61

Contents lists available at ScienceDirect

Econometrics and Statistics

journal homepage: www.elsevier.com/locate/ecosta

Robust Covariance Matrix Estimation in Time Series: A Review

Masayuki Hirukawa

Faculty of Economics, Ryukoku University, 67 Tsukamoto-cho, Fukakusa, Fushimi-ku, Kyoto 612-8577, Japan

ARTICLE INFO

Article history:
Received 26 February 2021
Revised 1 December 2021
Accepted 1 December 2021
Available online 16 December 2021

JEL classification:
C12
C13
C22
C32

Keywords:
Bandwidth
Generalized method of moments
Heteroskedasticity and autocorrelation robust inference
Kernel
Long-run variance
Positive semi-definite

ABSTRACT

In the analysis of economic, financial and other time series, long-run variance estimators play an important role in estimating model parameters more efficiently and drawing more accurate statistical inference on the parameters. A non-technical review of long-run variance estimation is provided. Both parametric and nonparametric estimators are discussed. Kernel methods are dominant among all estimation procedures, and therefore recent developments in kernel-smoothed estimators and related inference are presented. The information given can help practitioners decide on a suitable long-run variance estimator.

1. Introduction

Many applications of the analysis of economic, financial and other time series involve estimation of a covariance matrix of a zero-mean stationary vector process. An example of such a process is the cross-product of instruments and the regression error. The process typically exhibits conditional heteroskedasticity and serial correlation. Heteroskedasticity and temporal dependence may be of unknown form.

To estimate model parameters more efficiently and draw more accurate statistical inference on the parameters in this circumstance, we must estimate the long-run variance (LRV) matrix of the process. Heteroskedasticity and autocorrelation robust (HAR) covariance matrix estimators smoothed by kernel functions are frequently employed in LRV estimation. The estimators of this class have long been called heteroskedasticity and autocorrelation consistent (HAC) covariance matrix estimators due to the influential work by Andrews (1991). However, inconsistent kernel-smoothed estimators can yield well-defined test statistics. Therefore, the abbreviation HAR is adopted throughout, unless otherwise noted.

This article provides an overview of LRV estimation. LRV estimation forms an integral part of the generalized method of moments (GMM) estimation by Hansen (1982). With the development of GMM estimation, many excellent surveys on LRV or HAR estimation have been published. Comprehensive surveys on this subject include Den Haan and Levin (1997), Cushing and McGarvey (1999), and West (2008). Hall (2005, Section 3.5) focuses on LRV estimation within the framework of GMM

E-mail address: hirukawa@econ.ryukoku.ac.jp

https://doi.org/10.1016/j.ecosta.2021.12.001
2452-3062/

estimation, emphasizing that the inverse of an LRV estimator serves as the optimal weighting matrix of the efficient GMM criterion function. To the best of our knowledge, the most recent survey is given by Wang and Wu (2012). The primary target of that survey, however, is users of Stata, one of the most popular statistical packages among economists.

There are two objectives in this article. First, nearly a decade has passed since the latest survey was published. So, this article not only presents a variety of LRV estimators, but it also discusses recent developments in HAR estimation and inference that have not been covered before. Second, the article emphasizes the role of LRV estimators as a tool for efficient estimation of model parameters. On the one hand, development of LRV estimation has been paralleled by improvement in HAR inference. Indeed, HAR inference procedures have been refined since the seminal work on test statistics using inconsistent kernel-smoothed HAR estimators by Kiefer et al. (2000) and Kiefer and Vogelsang (2002, 2005). On the other hand, little is known about the implementation of HAR estimators in estimating parameters in GMM and cointegrating regressions efficiently. When employing a kernel HAR estimator for these estimation problems, we typically rely on a so-called automatic bandwidth, which is an estimate of the bandwidth that minimizes the asymptotic mean squared error (AMSE) of the HAR estimator. This bandwidth does not necessarily minimize the AMSE of the parameter estimator of interest. To the best of our knowledge, the only available result is Wilhelm (2015), who derives the bandwidth that optimizes a quadratic loss based on a higher-order approximation to the efficient GMM estimator. It is hoped that the Monte Carlo study in Section 6 can give guidance for improving estimation-optimal bandwidth selection procedures one step further.

Furthermore, preference is given to practical rather than theoretical aspects. In particular, we pay attention to whether each LRV estimator necessarily generates positive semi-definite (PSD) estimates. This property is highly desirable, because the estimator will be used to construct the criterion function for efficient GMM estimation and to compute standard errors and test statistics.

Before proceeding, it is worth remarking on the relation of LRV estimation to spectral density estimation. The LRV matrix of a zero-mean stationary vector process equals 2π times its spectral density matrix evaluated at frequency zero. Therefore, LRV estimation dates back to pioneering works of spectral density estimation in statistical time series analysis such as Jowett (1955), Hannan (1957) and Parzen (1957, 1961). There are also book-length treatments of spectral estimation available. Examples include Hannan (1970), Anderson (1971), Brillinger (1975), and Priestley (1981). Kernel methods are dominant in such earlier works. As GMM estimation became a standard tool in economics, spectral matrix estimation was revived in the name of HAR estimation. Indeed, kernel-smoothed HAR estimators studied by Newey and West (1987), Gallant (1987) and Andrews (1991), for example, inherit their form and their asymptotic properties from the literature on spectral estimation.

In addition, studies on HAR estimation and inference can be found among more than 200 articles in *Econometrics and Statistics*. Wenger and Leschinski (2021) improve HAR inference in the context of structural break testing, whereas Kawka (2020) investigates statistical properties of HAR estimators for locally stationary processes. Details will be described in Sections 5.4 and 5.6.1, respectively.

The remainder of this article is organized as follows. Section 2 provides an overview of LRV estimation. Various applications of LRV estimates in estimation and testing problems are also documented. Procedures for LRV estimation may be classified into two broad categories. Sections 3 focuses on parametric LRV estimators. Section 4 discusses nonparametric LRV estimators with emphasis on sample autocovariance-based ones. While much of the material in Sections 3 and 4 appears in previous surveys, Section 5 presents several recent developments in HAR estimation and inference that have not been covered before. Section 6 reports the results of two Monte Carlo experiments. The simulation study focuses on appraising the influence of different HAR estimation procedures on parameter estimates of GMM and cointegrating regressions. Section 7 summarizes and concludes the article.

The article adopts the following notational conventions: $\mathbf{0}_{p\times q}$ is the $p\times q$ zero matrix; $i=\sqrt{-1}$; $\otimes$ is used to represent the tensor (or Kronecker) product; $\mathbf{I}_p$ is the p-dimensional identity matrix; L is the lag operator, i.e., $L^j\mathbf{Y}_t=\mathbf{Y}_{t-j}$ for $j=1,2,\ldots$; $a\vee b=\max\{a,b\}$ and $a\wedge b=\min\{a,b\}$ for $a,b\in\mathbb{R}$; $\lfloor\cdot\rfloor$ denotes the integer part; $*$ signifies conjugate transpose of a complex-valued matrix $\mathbf{A}$, i.e., $\mathbf{A}^*=\overline{\mathbf{A}}^\top$; $\mathbf{K}_{pp}$ is the $p^2\times p^2$ commutation matrix that transforms $\mathrm{vec}(\mathbf{A})$ into $\mathrm{vec}\left(\mathbf{A}^\top\right)$, i.e., $\mathbf{K}_{pp}=\sum_{j=1}^p\sum_{k=1}^p\mathbf{e}_j\mathbf{e}_k^\top\otimes\mathbf{e}_k\mathbf{e}_j^\top$, where $\mathbf{e}_j$ is the jth elementary p-vector; and $\chi^2(\nu)$ and $t(\upsilon)$ signify the chi-squared distribution with ν degrees of freedom and the Student-t distribution with υ degrees of freedom, respectively. Lastly, limits and orders of magnitude are taken as $T\to\infty$ unless otherwise noted.

2. LRV Estimation: An Overview

2.1. Definition of the LRV

Let $\mathbf{g}_t$ be an $\ell\times 1$ stationary vector process with $E(\mathbf{g}_t)=\mathbf{0}_{\ell\times 1}$. Throughout $\mathbf{g}_t$ is assumed to have an $\ell\times\ell$ spectral matrix. Denote the jth-order autocovariance of $\mathbf{g}_t$ for $j=0,1,\ldots$ as $\mathbf{\Gamma}_j=E(\mathbf{g}_t\mathbf{g}_{t-j}^\top)$, where $\mathbf{\Gamma}_{-j}=\mathbf{\Gamma}_j^\top$. The $\ell\times\ell$ LRV matrix of the process $\mathbf{g}_t$ is in the form of

$$\mathbf{\Omega}=\sum_{j=-\infty}^{\infty}\mathbf{\Gamma}_j=\mathbf{\Gamma}_0+\sum_{j=1}^{\infty}\mathbf{\Gamma}_j+\sum_{j=1}^{\infty}\mathbf{\Gamma}_j^\top=\mathbf{\Gamma}_0+\mathbf{\Gamma}+\mathbf{\Gamma}^\top. \tag{1}$$

Most of the LRV estimators in this article can be obtained in the time domain. However, some are computed in the frequency domain (see, e.g., Sections 5.1.1 and 5.6.2). For this purpose, the spectral density matrix of $\mathbf{g}_t$ at frequency $\omega\in(-\pi,\pi)$ is

also considered. It is given by

$$f_{\mathbf{gg}}(\omega) = \frac{1}{2\pi} \sum_{j=-\infty}^{\infty} \mathbf{\Gamma}_j e^{-ij\omega}. \tag{2}$$

Observe that $\mathbf{\Omega} = 2\pi f_{\mathbf{gg}}(0)$ holds.

2.2. Roles of the LRV in the GMM Framework

There are two major reasons why we estimate the LRV matrix (1). These are (i) efficient estimation and (ii) statistical inference. To clarify the reasons, we further specify the process $\mathbf{g}_t$ as $\mathbf{g}_t = \mathbf{g}_t(\theta_0) = \mathbf{g}(\mathbf{w}_t; \theta_0)$ for a stationary vector process $\mathbf{w}_t$ and a known function $\mathbf{g}(\cdot;\cdot)$ up to the parameter of interest $\theta_0 \in \mathbf{\Theta} \subseteq \mathbb{R}^p$ $(p \leq \ell)$. Also assume that

$$E\{\mathbf{g}_t(\theta)\} = \mathbf{0}_{\ell\times 1} \tag{3}$$

holds if and only if $\theta = \theta_0$.

As a concrete example, consider a system of r equations

$$\mathbf{y}_t = \mathbf{X}_t^\top \theta_0 + \mathbf{u}_t \tag{4}$$

for the $r \times 1$ vector of dependent variables $\mathbf{y}_t$, the $p \times r$ matrix of regressors $\mathbf{X}_t$ and the $r \times 1$ vector of regression errors $\mathbf{u}_t$. If $\mathbf{X}_t$ is endogenous and the $\ell \times r$ matrix of instruments $\mathbf{Z}_t$ is chosen to estimate θ_0, then the orthogonality condition (3) becomes

$$E(\mathbf{Z}_t\mathbf{u}_t) = E\left\{\mathbf{Z}_t\left(\mathbf{y}_t - \mathbf{X}_t^\top\theta_0\right)\right\} = \mathbf{0}_{\ell\times 1}. \tag{5}$$

Alternatively, as in Hansen and Singleton (1982), the $h \times 1$ vector of instruments $\mathbf{R}_t$ may be orthogonal to each element of $\mathbf{u}_t$. In this scenario, we simply put $\mathbf{Z}_t = \mathbf{R}_t \otimes \mathbf{I}_r$ and $\ell = hr$. Observe that $\mathbf{w}_t = \left(\mathbf{y}_t^\top, \mathrm{vec}\,(\mathbf{X}_t)^\top, \mathrm{vec}\,(\mathbf{Z}_t)^\top\right)^\top$ and $\mathbf{w}_t = \left(\mathbf{y}_t^\top, \mathrm{vec}\,(\mathbf{X}_t)^\top, \mathrm{vec}\,(\mathbf{R}_t)^\top\right)^\top$ in the first and second cases, respectively, and that the moment function $\mathbf{g}_t = \mathbf{Z}_t\left(\mathbf{y}_t - \mathbf{X}_t^\top\theta_0\right)$ in both cases.

Given T observations $\{\mathbf{w}_t\}_{t=1}^T$, the orthogonality condition (3) motivates us to estimate θ_0 by GMM so that $\tilde{\theta} = \arg\min_{\theta\in\mathbf{\Theta}} \mathbf{G}(\theta)^\top \mathbf{W} \mathbf{G}(\theta)$, where $\mathbf{G}(\theta) = (1/T)\sum_{t=1}^T \mathbf{g}_t(\theta)$ is the sample counterpart of $E\{\mathbf{g}_t(\theta)\}$, and $\mathbf{W}$ is some $\ell \times \ell$ symmetric PSD weighting matrix. Now the reason (i) becomes clear. In efficient GMM estimation, the inverse of a consistent estimator of (1) is chosen as $\mathbf{W}$. Denote the estimator as $\hat{\mathbf{\Omega}}$. Then, for the efficient GMM estimator

$$\hat{\theta} = \arg\min_{\theta\in\mathbf{\Theta}} \mathbf{G}(\theta)^\top \hat{\mathbf{\Omega}}^{-1} \mathbf{G}(\theta), \tag{6}$$

it holds that $\sqrt{T}\left(\hat{\theta} - \theta_0\right) \xrightarrow{d} N\left(\mathbf{0}_{p\times 1}, \mathbf{V}\right) = N\left(\mathbf{0}_{p\times 1}, \left(\mathbf{D}^\top \mathbf{\Omega}^{-1} \mathbf{D}\right)^{-1}\right)$, where $\mathbf{D} = E\left\{\partial \mathbf{g}_t(\theta)/\partial\theta^\top\big|_{\theta=\theta_0}\right\}$.

The reason (ii) appears at this stage. The asymptotic variance $\mathbf{V}$ can be consistently estimated as $\hat{\mathbf{V}} = \left(\hat{\mathbf{D}}^\top \hat{\mathbf{\Omega}}^{-1} \hat{\mathbf{D}}\right)^{-1}$, where $\hat{\mathbf{D}} = \partial \mathbf{G}(\theta)/\partial\theta^\top\big|_{\theta=\hat{\theta}}$ is the sample counterpart of $\mathbf{D}$. The variance estimator $\hat{\mathbf{V}}$ can be used to compute standard errors for $\hat{\theta}$ and to construct test statistics for θ_0. Another example of inference is the J-test of overidentifying restrictions, where $J = T\mathbf{G}\left(\hat{\theta}\right)^\top \hat{\mathbf{\Omega}}^{-1} \mathbf{G}\left(\hat{\theta}\right) \xrightarrow{d} \chi^2(\ell - p)$ for $\ell > p$, under the null hypothesis that (3) holds true.

2.3. Applications of LRV Estimators

Other than GMM, there are many examples of applying LRV estimators for efficient estimation and inference. In reference to efficient estimation, LRV estimators are used for efficiency gain via generalized least squares-type estimation methods applied to seemingly unrelated regression models (e.g., Creel and Farell, 1996) and two-pass cross-sectional regressions of asset returns (e.g., Ahn et al., 2012). Moreover, in some class of cointegrating regression models (e.g., Phillips and Hansen, 1990; Phillips, 1991; Park, 1992), an estimate of the one-sided LRV matrix

$$\mathbf{\Lambda} = \mathbf{\Gamma}_0 + \mathbf{\Gamma} = \mathbf{\Gamma}_0 + \sum_{j=1}^{\infty} \mathbf{\Gamma}_j, \tag{7}$$

as well as that of the two-sided one (1), is employed to correct the so-called second-order effect that arises when the regressor is endogenous.

LRV estimators are also applied for a variety of testing problems other than inference on parameters in regression and moment-condition models. Examples include unit-root and stationarity tests (e.g., Phillips, 1987; Phillips and Perron, 1988; Kwiatkowski et al., 1992), cointegration tests (e.g., Phillips and Ouliaris, 1990), structural break tests (e.g., Andrews, 1993), and forecast comparison tests (e.g., Diebold and Mariano, 1995; West, 1996; McCracken, 2000; Rossi and Inoue, 2012), to name a few.

2.4. Classifications of LRV Estimation Procedures

Our review of LRV estimation starts from the next section. Particular attention is paid to whether the LRV estimator in discussion necessarily generates PSD estimates. This property is highly desirable, because the estimator will be used to construct the optimal weighting matrix for the efficient GMM criterion function and to compute standard errors and test statistics.

LRV estimation procedures can be categorized into two broad classes. The first class is parametric procedures covered in Section 3. In the parametric procedures, a linear time series model such as a vector autoregressive (VAR) or a vector moving-average (VMA) model is fitted to the vector process $\mathbf{g}_t$. Then, (2π times) the spectral density matrix at frequency zero implied by the model is computed. Parametric LRV estimators become PSD by construction.

The second class is nonparametric procedures. The most popular estimator is in the form of a weighted sum of sample autocovariances, where the weights are determined by a kernel and bandwidth. Section 4 focuses on this class of estimators and refers to a few kernels that necessarily generate PSD estimates.

3. Parametric LRV Estimation

3.1. VMA-Based Estimators

This section begins with the parametric LRV estimator implied by specifying $\mathbf{g}_t$ as a VMA model of known order. Finite-order serial dependence typically arises when $\mathbf{g}_t$ is implied by a multi-period ahead predictive regression or comes from an Euler equation for a rational expectations model.

Suppose that $\mathbf{g}_t$ is known to obey a VMA model of order n a priori, as is the case with an $(n+1)$-period ahead predictive regression. Then, it can be expressed as $\mathbf{g}_t = \varepsilon_t + \mathbf{\Psi}_1\varepsilon_{t-1} + \cdots + \mathbf{\Psi}_n\varepsilon_{t-n} = \mathbf{\Psi}(L)\varepsilon_t$, where ε_t is the $\ell \times 1$ innovation in $\mathbf{g}_t$ with $\mathbf{\Sigma}_\varepsilon = E\left(\varepsilon_t\varepsilon_t^\top\right)$, and $\mathbf{\Psi}(L) = \mathbf{I}_\ell + \mathbf{\Psi}_1 L + \cdots + \mathbf{\Psi}_n L^n$ is the lag polynomial with $\mathbf{\Psi}_1, \ldots, \mathbf{\Psi}_n$ being $\ell \times \ell$ coefficient matrices. It follows that the LRV matrix of $\mathbf{g}_t$ becomes $\mathbf{\Omega} = \mathbf{\Psi}(1)\mathbf{\Sigma}_\varepsilon\mathbf{\Psi}(1)^\top$, and a consistent estimator of $\mathbf{\Omega}$ can be obtained, in principle, by replacing $\mathbf{\Psi}_1, \ldots, \mathbf{\Psi}_n$ with their consistent estimates. This estimator is PSD by construction and $T^{1/2}$-consistent. However, it has not been applied in empirical works, because computational cost of estimating VMA coefficient matrices $\mathbf{\Psi}_1, \ldots, \mathbf{\Psi}_n$ is yet to be inexpensive. The estimation suffers from slow convergence of likelihood functions even for small-dimensional systems.

Instead of directly estimating VMA coefficients, Cumby et al. (1983) and Eichenbaum et al. (1988) first estimate a finite-order (approximate) VAR model for $\mathbf{g}_t$ and then invert the model to recover the VMA coefficients. While the resulting LRV estimate is PSD, its convergence rate may be no longer the parametric one; actually, the rate is determined by the expansion rate of the lag order of the VAR model.

Hodrick (1992) and West (1997) take a different estimation strategy. Their estimators are PSD and $T^{1/2}$-consistent but can be computed with less computational burden. Suppose that the regression error $\mathbf{u}_t$ in (4) is known to obey a VMA(n) model a priori. The focus here is on the estimator proposed by West (1997), because Hodrick's (1992) estimator may be viewed as a special case of West's (1997) estimator in which $\mathbf{u}_t$ is a scalar and all MA coefficients are set equal to unity. Since $r = \dim(\mathbf{u}_t)$ is usually (much) smaller than $\ell = \dim(\mathbf{g}_t)$ (e.g., $1 = r \leq p \leq \ell$ in a single-equation case), it is easier to estimate the VMA(n) model for $\mathbf{u}_t$ than $\mathbf{g}_t$. Then, West (1997) fits a VMA(n) process to the residuals $\hat{\mathbf{u}}_t$ to obtain $\hat{\mathbf{u}}_t = \hat{\varepsilon}_t + \hat{\psi}_1\hat{\varepsilon}_{t-1} + \cdots + \hat{\psi}_n\hat{\varepsilon}_{t-n}$, where the $\hat{\varepsilon}_t$ are fitted innovations with $\hat{\varepsilon}_t \equiv \mathbf{0}_{r\times 1}$ for $t \leq 0$, and $\hat{\psi}_1, \ldots, \hat{\psi}_n$ are consistent estimates of $r \times r$ coefficient matrices. Given the instrument matrix $\mathbf{Z}_t$ defined in (5), let $\hat{\mathbf{d}}_{t+n} = \left(\mathbf{Z}_t + \mathbf{Z}_{t+1}\hat{\psi}_1 + \cdots + \mathbf{Z}_{t+n}\hat{\psi}_n\right)\hat{\varepsilon}_t$ for $t = 1, \ldots, T-n$. Then, the VMA(n)-based LRV estimator by West (1997) is

$$\hat{\mathbf{\Omega}}_{VMA} = \frac{1}{T-n}\sum_{t=1}^{T-n}\hat{\mathbf{d}}_{t+n}\hat{\mathbf{d}}_{t+n}^\top.$$

3.2. VAR-Based Estimator

VAR-based LRV estimators are also available. A benefit of VAR models is low computational cost, because coefficients can be estimated by ordinary least squares (OLS). Den Haan and Levin (1996, 1997, 2000) propose to fit an infinite-order VAR model to $\mathbf{g}_t$ and then compute the LRV estimator as 2π times the spectral density matrix at frequency zero implied by this model. The resulting estimator is called the VARHAC estimator because of this particular procedure.

Before proceeding, a few remarks are in order. First, the idea of approximating univariate time series models by infinite-order AR processes dates back to Berk (1974). In this respect, the VARHAC estimator can be viewed as an application of the multivariate extension of Berk's (1974) method. Second, consistency of the VAR process is essential for consistency of the VARHAC estimator. This is a sharp contrast to VAR-based prewhitening to be discussed in Section 4.2.3. Therefore, the lag order must be chosen in a data-driven manner as documented below.

The VARHAC estimator is implemented as follows. Suppose that $\mathbf{g}_t$ can be specified as a possibly infinite-order VAR model

$$\mathbf{g}_t = \sum_{j=1}^{\infty} \mathbf{\Phi}_j \mathbf{g}_{t-j} + \epsilon_t \tag{8}$$

with $\mathbf{\Sigma}_\epsilon = E\left(\epsilon_t \epsilon_t^\top\right)$, where the $\mathbf{\Phi}_j$ are $\ell \times \ell$ coefficient matrices. The VAR(∞) representation is possible if $\mathbf{g}_t$ obeys an invertible VMA or VARMA model, as well as a finite-order VAR model.

To approximate (8) by a finite-order VAR model, let $M = M_T$ be the maximum lag order. It is assumed that $M_T \to \infty$. Coefficient estimation and order selection can be done in an element-by-element manner. To see this, suppose that an AR(m) model is fitted to $\mathbf{g}_{a,t}$, the ath element of $\mathbf{g}_t$, for $m = 1, \ldots, M$ and $a = 1, \ldots, \ell$. The fitted model can be expressed as

$$\mathbf{g}_{a,t} = \sum_{j=1}^{m} \sum_{b=1}^{\ell} \hat{\phi}_{a,b,j}(m) \mathbf{g}_{b,t-j} + \hat{\epsilon}_{a,t}(m), \; t = M+1, \ldots, T, \tag{9}$$

where the $\hat{\phi}_{a,b,j}(m)$ are OLS estimates of coefficients and $\hat{\epsilon}_{a,t}(m)$ is the residual. In this model, the number of lagged dependent variables $\mathbf{g}_{a,t-j}$ and that of other predetermined elements $\mathbf{g}_{b,t-j}$ for $b \neq a$ are assumed to be the same. Alternatively, it is possible to differ the former from the latter; see Den Haan and Levin (2000) for more details. The optimal lag order for $\mathbf{g}_{a,t}$, denoted as $\hat{m}_a = \hat{m}_{aT} \in \{1, \ldots, M\}$, is determined as the minimizer of either Akaike's (1973) information criterion (AIC)

$$AIC_a(m) = \log\left\{ \frac{1}{T} \sum_{t=M+1}^{T} \hat{\epsilon}_{a,t}^2(m) \right\} + \frac{2}{T} m\ell, \tag{10}$$

or Schwarz' (1978) Bayesian information criterion (BIC)

$$BIC_a(m) = \log\left\{ \frac{1}{T} \sum_{t=M+1}^{T} \hat{\epsilon}_{a,t}^2(m) \right\} + \frac{\log T}{T} m\ell. \tag{11}$$

Denote the maximum optimal lag order as $\hat{m} = \hat{m}_T = \max_{1 \leq a \leq \ell} \hat{m}_a$. Also let $\hat{\mathbf{\Phi}}_j(\hat{m})$ be the $\ell \times \ell$ estimated coefficient matrix for lag $j = 1, \ldots, \hat{m}$, where the (a, b) element of $\hat{\mathbf{\Phi}}_j(\hat{m})$ is $\hat{\phi}_{a,b,j}(\hat{m})$ with $\hat{\phi}_{a,b,j}(\hat{m}) \equiv 0$ for $j > \hat{m}_a$. Then, the estimated lag polynomial $\hat{\mathbf{\Phi}}_{\hat{m}}(L)$ is given by $\hat{\mathbf{\Phi}}_{\hat{m}}(L) = \mathbf{I}_\ell - \sum_{j=1}^{\hat{m}} \hat{\mathbf{\Phi}}_j(\hat{m}) L^j$. For the vector of residuals $\hat{\epsilon}_t(\hat{m}) = \left(\hat{\epsilon}_{1,t}(\hat{m}_1), \ldots, \hat{\epsilon}_{\ell,t}(\hat{m}_\ell)\right)^\top$, $t = M+1, \ldots, T$, the instantaneous variance of the innovation can be estimated by $\hat{\mathbf{\Sigma}}_{\hat{\epsilon}(\hat{m})} = (T-M)^{-1} \sum_{t=M+1}^{T} \hat{\epsilon}_t(\hat{m}) \hat{\epsilon}_t(\hat{m})^\top$. In the end, the VARHAC estimator is defined as

$$\hat{\mathbf{\Omega}}_{VAR} = \hat{\mathbf{\Phi}}_{\hat{m}}(1)^{-1} \hat{\mathbf{\Sigma}}_{\hat{\epsilon}(\hat{m})} \left\{ \hat{\mathbf{\Phi}}_{\hat{m}}(1)^\top \right\}^{-1}. \tag{12}$$

The VARHAC estimator is PSD by construction. In addition, Den Haan and Levin (1996, Theorem 3) demonstrate that if $M_T = O\left(T^{1/3}\right)$ and the optimal lag order $\hat{m}_T$ is determined via BIC (11) and some other regularity conditions hold, then the estimator is nearly $T^{1/2}$-consistent.

3.3. Discussion

Parametric LRV estimators have a few advantages over nonparametric, kernel-smoothed ones. The former necessarily generates PSD estimates and attains (nearly) $T^{1/2}$-consistency. Simulation results by West (1997) and Den Haan and Levin (1997) indicate that test statistics using their parametric LRV estimators often outperform those using nonparametric kernel estimators. On the other hand, whether a kernel-smoothed LRV estimator is PSD depends on the choice of a kernel, and its convergence rate is usually slower than the parametric one.

Nevertheless, parametric LRV estimators are less popular than nonparametric ones in practice. This can be attributed to difficulty in estimating model parameters precisely. First, estimating a VMA model is still computationally challenging. Therefore, a kernel-smoothed estimator with a suitable choice of the bandwidth is preferred even when an economic or other theory suggests that $\mathbf{g}_t$ has finite-order dependence. Second, estimating many parameters in the VAR model is a serious concern for VARHAC. In (9) we use $(T-M)$ observations to estimate $m\ell$ parameters. When ℓ is large relative to T (e.g., for the case of many instruments in (5)), we may resort to regularized VAR estimation (e.g., Basu and Michailidis, 2015). However, VARHAC in this direction is yet to be developed, to the best of our knowledge. Moreover, the process $\mathbf{g}_t$ computed from economic and financial data often exhibits high persistence. Then, OLS estimates of VAR coefficients have considerable finite-sample bias, which is likely to deteriorate finite-sample properties of the VARHAC estimator, as suggested by Sul et al. (2005).

4. Nonparametric LRV Estimation

4.1. Truncated Estimator

Among all nonparametric LRV estimators, sample autocovariance-based ones are dominant. Therefore, below we focus exclusively on this class. Suppose that an economic or other theory suggests that $\mathbf{g}_t$ has zero autocovariances after lag n, as seen in Section 3.1. The LRV matrix reduces to $\mathbf{\Omega} = \sum_{j=-n}^{n} \mathbf{\Gamma}_j$ in this case. To estimate $\mathbf{\Omega}$, define the jth-order sample autocovariance of $\mathbf{g}_t$ for $j = 0, \pm 1, \ldots, \pm(T-1)$ as $\hat{\mathbf{\Gamma}}_j = (1/T)\sum_{t=1\vee(1+j)}^{(T+j)\wedge T} \mathbf{g}_t \mathbf{g}_{t-j}^{\top}$. A natural estimator of $\mathbf{\Omega}$ is its sample analog

$$\hat{\mathbf{\Omega}}_{TR} = \sum_{j=-n}^{n} \hat{\mathbf{\Gamma}}_j.$$

This is called the truncated estimator, and it is studied in the early literature on GMM including Hansen (1982), Hansen and Singleton (1982), and White and Domowitz (1984). Hansen and Hodrick (1980) consider a similar but more restrictive LRV estimator, but its nature is the same.

The truncated estimator is consistent for more general cases. White and Domowitz (1984, Theorem 3.5) establish that even when $\mathbf{\Gamma}_j \neq \mathbf{0}_{\ell\times\ell}$ for all j, $\hat{\mathbf{\Omega}}_{TR}$ becomes consistent if the lag truncation point $n = n_T$ satisfies $1/n_T + n_T/T^{1/3} \to 0$ and some other regularity conditions hold.

A disadvantage of the truncated estimator is that it does not necessarily generate PSD estimates. West (1997) and Hirukawa (2010) report in their simulation studies that $\hat{\mathbf{\Omega}}_{TR}$ quite frequently yields negative variance estimates in the presence of negative serial dependence. Therefore, this estimator is not much used in practice.

4.2. Weighted Autocovariance Estimators

4.2.1. Definition and Kernel Choices

The most popular LRV estimators take the form of a weighted sum of sample autocovariances $\hat{\mathbf{\Omega}} = \sum_{j=-(T-1)}^{T-1} w_j \hat{\mathbf{\Gamma}}_j$. Following the literature on spectral density estimation, Andrews (1991) explores a general framework for the analysis on this class of estimators. In the framework, the weights w_j are determined by a kernel $k(\cdot)$ and a sequence of bandwidth $S_T \in \mathbb{R}_+$ so that $w_j = k(j/S_T)$ and

$$\hat{\mathbf{\Omega}}_{\mathcal{J}} = \sum_{j=-(T-1)}^{T-1} k_{\mathcal{J}}\left(\frac{j}{S_T}\right)\hat{\mathbf{\Gamma}}_j, \tag{13}$$

where the subscript $\mathcal{J}$ signifies that the kernel $k_{\mathcal{J}}(\cdot)$ is chosen. Such LRV estimators have long been called HAC estimators due to the pioneered work by Andrews (1991). However, the estimator (13) using the bandwidth $S_T = T$ is inconsistent but can construct well-defined test statistics, as in Section 5.3. Therefore, we refer to what was called HAC estimators as HAR estimators, whenever there is no confusion.

Some conditions on a kernel $k(\cdot)$ and bandwidth S_T are required for consistency of HAR estimators. A typical set of regularity conditions on $k(\cdot)$ are summarized as Assumption 1 below. The final condition $\int_0^\infty \bar{k}(x)dx < \infty$ ensures that certain Riemann-type sums defined in terms of the kernel converge to their integral representation counterparts; see Jansson (2002, p.1451) for discussion.

Assumption 1. $k : \mathbb{R} \to [-1, 1]$, $k(0) = 1$, $k(x) = k(-x)$, for all $x \in \mathbb{R}$, $k(\cdot)$ is continuous at 0 and almost everywhere, and $\int_0^\infty \bar{k}(x)dx < \infty$ where $\bar{k}(x) = \sup_{y\geq x}|k(y)|$.

Quite a few kernels satisfy Assumption 1, and the bandwidth selection methods in the next section are also designed for these kernels. From the practical perspective, however, preference is given to the following most commonly used kernels.

Truncated: $k_{TR}(x) = \begin{cases} 1 & \text{for } |x| \leq 1 \\ 0 & \text{otherwise} \end{cases}$

Bartlett: $k_{BT}(x) = \begin{cases} 1-|x| & \text{for } |x| \leq 1 \\ 0 & \text{otherwise} \end{cases}$

Parzen: $k_{PZ}(x) = \begin{cases} 1-6x^2+6|x|^3 & \text{for } |x| \leq 1/2 \\ 2(1-|x|)^3 & \text{for } 1/2 < |x| \leq 1 \\ 0 & \text{otherwise} \end{cases}$

Quadratic Spectral: $k_{QS}(x) = \frac{25}{12\pi^2x^2}\left\{\frac{\sin(6\pi x/5)}{6\pi x/5} - \cos\left(\frac{6\pi x}{5}\right)\right\}$

The truncated estimator discussed in the previous section can be viewed as the HAR estimator in which the truncated kernel is used and the bandwidth S_T is set equal to n. The Bartlett, Parzen and quadratic spectral (QS) kernels are investigated by Newey and West (1987), Gallant (1987, p.533) and Andrews (1991), respectively.

Table 1
Characteristic Numbers of Kernels Widely Applied in HAR Estimation

Kernel	q	k_q	$\int_{-\infty}^{\infty} k^2(x)dx$	$\int_{-\infty}^{\infty} \lvert x\rvert^{2q} k^2(x)dx$
Bartlett	1	1	2/3	1/15
Parzen	2	6	151/280	929/295680
QS	2	$18\pi^2/125$	1	∞

Whether a kernel necessarily generates PSD estimates is of particular importance and interest. To answer this question, define the *spectral window generator* of a kernel $k(\cdot)$ (Andrews, 1991) as $K(\lambda) = (2\pi)^{-1}\int_{-\infty}^{\infty} k(x)e^{-ix\lambda}dx$. The kernel $k(\cdot)$ necessarily generates PSD estimates if $K(\lambda) \geq 0$ for all $\lambda \in \mathbb{R}$. The Bartlett, Parzen and QS kernels satisfy this condition, but the truncated kernel does not. Moreover, Andrews (1991) demonstrate that the QS kernel is optimal in the sense that it attains the smallest AMSE among all HAR estimators smoothed by kernels yielding PSD estimates. The spectral window generator of this kernel reduces to the Epanechnikov kernel in the frequency domain, which is known to be optimal among all nonnegative kernels. Optimality of the QS kernel can be understood analogously.

Now we turn to the regularity condition on the bandwidth S_T. For consistency of HAR estimators, let S_T satisfy $1/S_T + S_T/T \to 0$. This assumption, jointly with examples of the kernels satisfying Assumption 1, implies the following intuitions. On the one hand, it follows from $S_T/T \to 0$ that $k(j/S_T) = 0$ (or $k(j/S_T) \to 0$ at least) for $|j|$ near $T-1$. As a consequence, sample autocovariances with long lags, which tend to be estimated imprecisely, are downweighted. On the other hand, $S_T \to \infty$ implies that $k(j/S_T) \to k(0) = 1$ for each j, which leads to consistency of a HAR estimator. The requirement for the bandwidth to diverge also draws the important conclusion that we must set $S_T \to \infty$ even when $\mathbf{g}_t$ is known to have at most n non-zero autocovariances. Putting $S_T = n$ is allowed only if the truncated kernel is employed. However, this comes at the cost of non-PSD estimates.

4.2.2. Data-Driven Bandwidth Selection

How to choose the bandwidth S_T for a given kernel $k(\cdot)$ is an important practical problem. Finite-sample performances of a HAR estimator is largely influenced by the bandwidth, and the choice of a kernel is thought to be of second importance in practice. This section addresses the issue of selecting S_T in a data-driven manner.

To begin with, we refer to measures of smoothness for a kernel and spectral density. A kernel $k(\cdot)$ is said to have the *characteristic exponent* q (Parzen, 1957) if it has the following properties:

$$k_r = \lim_{x\to 0}\left\{\frac{1-k(x)}{|x|^r}\right\}\begin{cases} = 0 & \text{for } r < q \\ \in (0,\infty) & \text{for } r = q, \\ = \infty & \text{for } r > q \end{cases}$$

where k_r and k_q are called the rth *generalized derivative of the kernel at the origin* and the *characteristic coefficient* (Parzen, 1961), respectively. In addition, the rth *generalized derivative of the spectral density* (2) at frequency $\omega \in (-\pi, \pi)$ is defined as $f_{\mathbf{gg}}^{(r)}(\omega) = (2\pi)^{-1}\sum_{j=-\infty}^{\infty}|j|^r\Gamma_j e^{-ij\omega}$. Since the truncated kernel has $k_r \equiv 0$ for any $r > 0$ and thus $q = \infty$, it does not fit with the AMSE-optimal framework below. Furthermore, Parzen (1961, p.179) shows that a kernel that necessarily generates PSD estimates must satisfy $q \leq 2$, and remaining three kernels indeed have $q = 1$ (Bartlett) or 2 (Parzen, QS). Table 1 presents values of q and k_q, as well as some other characteristic numbers, of these kernels.

We adopt the AMSE as the loss function of a HAR estimator. The AMSE approximates the mean squared error (MSE) of the estimator by $AMSE = ABias^2 + AVar$, where $ABias$ and $AVar$ denote dominant terms of the bias and variance of the estimator, respectively. Andrews (1991) demonstrates that if a HAR estimator is built on a kernel having the characteristic exponent q, then under sufficient smoothness of the spectral density at frequency zero, $ABias = O(S_T^{-q})$ and $AVar = O(S_T/T)$. Therefore, $AMSE = ABias^2 + AVar = O(S_T^{-2q} + S_T/T)$, and thus the estimator attains the best possible convergence rate of $T^{q/(2q+1)}$ under the bandwidth $S_T = O(T^{1/(2q+1)})$ that balances $ABias^2$ and $AVar$. This means that $\hat{\mathbf{\Omega}}_{BT} \xrightarrow{p} \mathbf{\Omega}$ at rate $T^{1/3}$ if $S_T = O(T^{1/3})$ and that $\hat{\mathbf{\Omega}}_{PZ}, \hat{\mathbf{\Omega}}_{QS} \xrightarrow{p} \mathbf{\Omega}$ at rate $T^{2/5}$ if $S_T = O(T^{1/5})$.

However, the above result is merely a guidance, and a bandwidth formula of the form $S_T = CT^{1/(2q+1)}$ for some constant $C > 0$ is required for practical purposes. Andrews (1991), Newey and West (1994) and Hirukawa (2010) provide the so-called automatic bandwidth as an estimate of the minimizer of the AMSE of a HAR estimator using a general class of kernels. There are important differences in the definition of the MSE of a HAR estimator and the approach to estimating an unknown quantity in the AMSE-optimal bandwidth. Below we describe the three bandwidth choice rules by illustrating the differences.

Andrews (1991): Andrews (1991) defines the MSE of a HAR estimator $\hat{\mathbf{\Omega}}$ as $E\{\sum_{a=1}^{\ell} w_a(\hat{\mathbf{\Omega}}_{aa} - \mathbf{\Omega}_{aa})^2\}$, where the $w_a \geq 0$ are weights, and $\hat{\mathbf{\Omega}}_{aa}$ and $\mathbf{\Omega}_{aa}$ are the ath diagonal elements of $\hat{\mathbf{\Omega}}$ and $\mathbf{\Omega}$, respectively. The definition yields the AMSE-optimal bandwidth formula as

$$S_T^{\dagger} = \left\{\frac{qk_q^2\alpha(q)}{\int_{-\infty}^{\infty}k^2(x)dx}\right\}^{1/(2q+1)} T^{1/(2q+1)}, \tag{14}$$

where $\alpha(q) = \sum_{a=1}^{\ell} w_a \left(f^{(q)}_{\mathbf{gg},aa} \right)^2 / \sum_{a=1}^{\ell} w_a f^2_{\mathbf{gg},aa}$, and $f^{(q)}_{\mathbf{gg},aa}$ and $f_{\mathbf{gg},aa}$ are the ath diagonal elements of $f^{(q)}_{\mathbf{gg}}(0)$ and $f_{\mathbf{gg}}(0)$, respectively.

To make (14) fully operational, we must prespecify the weights w_a and replace the unknown $f^{(q)}_{\mathbf{gg},aa}$ and $f_{\mathbf{gg},aa}$ with their proxies. For the former, Andrews (1991) set all weights corresponding to slopes equal to unity and the one corresponding to an intercept equal to zero for regression models. For the latter, an AR(1) model as a reference is fitted to each element of $\mathbf{g}_t$. In the end, using the numbers in Table 1 yields Andrews' (1991) data-driven (or automatic) bandwidth as

$$\hat{S}^{\dagger}_T = \begin{cases} 1.1447\{\hat{\alpha}(1)\}^{1/3} T^{1/3} & \text{for Bartlett} \\ 2.6614\{\hat{\alpha}(2)\}^{1/5} T^{1/5} & \text{for Parzen} \\ 1.3221\{\hat{\alpha}(2)\}^{1/5} T^{1/5} & \text{for QS} \end{cases},$$

where

$$\hat{\alpha}(1) = \frac{\sum_{a=1}^{\ell} w_a \left\{ \frac{4\hat{\rho}_a^2 \hat{\sigma}_a^4}{(1-\hat{\rho}_a)^6 (1+\hat{\rho}_a)^2} \right\}}{\sum_{a=1}^{\ell} w_a \left\{ \frac{\hat{\sigma}_a^4}{(1-\hat{\rho}_a)^4} \right\}}, \quad \hat{\alpha}(2) = \frac{\sum_{a=1}^{\ell} w_a \left\{ \frac{4\hat{\rho}_a^2 \hat{\sigma}_a^4}{(1-\hat{\rho}_a)^8} \right\}}{\sum_{a=1}^{\ell} w_a \left\{ \frac{\hat{\sigma}_a^4}{(1-\hat{\rho}_a)^4} \right\}},$$

and $(\hat{\rho}_a, \hat{\sigma}_a)$ are OLS estimates of the coefficient and the instantaneous variance of the innovation in the AR(1) model fitted to the ath element of $\mathbf{g}_t$, respectively.

Andrews' (1991) approach is analogous to the "rule of thumb" for probability density estimation by Silverman (1986, Section 3.4.2). A potential problem is that the data-driven bandwidth is not consistent for the AMSE-optimal bandwidth unless the AR(1) model is elementwise correct specification of the process $\mathbf{g}_t$. Hence, this approach may perform poorly when $\mathbf{g}_t$ is not well approximated by an AR(1) model.

Newey and West (1994). For an $\ell \times 1$ vector of weights $\mathbf{w}$, Newey and West (1994) consider $E\{\mathbf{w}^{\top}(\hat{\mathbf{\Omega}} - \mathbf{\Omega})\mathbf{w}\}^2$ as an alternative definition of the MSE of a HAR estimator $\hat{\mathbf{\Omega}}$. This definition leads to the AMSE-optimal bandwidth formula

$$S^{\dagger}_T = \left[\frac{q k_q^2 \left\{ R^{(q)} \right\}^2}{\int_{-\infty}^{\infty} k^2(x) dx} \right]^{1/(2q+1)} T^{1/(2q+1)}, \tag{15}$$

where

$$R^{(q)} = \frac{\mathbf{w}^{\top} f^{(q)}_{\mathbf{gg}} \mathbf{w}}{\mathbf{w}^{\top} f_{\mathbf{gg}} \mathbf{w}} = \frac{\sum_{j=-\infty}^{\infty} |j|^q \gamma_j}{\sum_{j=-\infty}^{\infty} \gamma_j} = \frac{s^{(q)}}{s^{(0)}},$$

$\gamma_j = E\left(h_t h_{t-j}\right)$ is the jth autocovariance of the process $h_t = \mathbf{w}^{\top}\mathbf{g}_t$, and $s^{(r)} = \sum_{j=-\infty}^{\infty} |j|^r \gamma_j$. Observe that (14) and (15) coincide if $\mathbf{g}_t$ is a scalar process. Because the quantity $R^{(q)}$ is the ratio of a density derivative to the density itself, Hirukawa (2010) refers to it as the *normalized curvature*.

When implementing (15), Newey and West (1994) estimate $R^{(q)}$ nonparametrically using the truncated kernel to avoid possible misspecification of the process h_t. As in Andrews (1991), Newey and West (1994) also set all weights in $\mathbf{w}$ corresponding to slopes equal to unity and the one corresponding to an intercept equal to zero for regression models. The truncated estimator for $R^{(q)}$ using the bandwidth n is

$$\hat{R}^{(q)}_{TR} = \frac{\sum_{j=-n}^{n} |j|^q \hat{\gamma}_j}{\sum_{j=-n}^{n} \hat{\gamma}_j}, \tag{16}$$

where $\hat{\gamma}_j = \sum_{t=1+j}^{T} h_t h_{t-j}/T$ is the jth-order sample autocovariance of h_t for $j = 0, 1, \ldots, T-1$ with $\hat{\gamma}_{-j} = \hat{\gamma}_j$. The problem is that it is hard to find the optimal bandwidth $n^{\dagger} = n^{\dagger}_T$ for this estimator. Rather, Newey and West (1994) determine n in an ad hoc manner. The final form of their data-driven bandwidth is

$$\hat{S}^{\dagger}_T = \begin{cases} 1.1447\left[\left\{ \hat{R}^{(1)}_{TR} \right\}^2 \right]^{1/3} T^{1/3} & \text{for Bartlett} \\ 2.6614\left[\left\{ \hat{R}^{(2)}_{TR} \right\}^2 \right]^{1/5} T^{1/5} & \text{for Parzen} \\ 1.3221\left[\left\{ \hat{R}^{(2)}_{TR} \right\}^2 \right]^{1/5} T^{1/5} & \text{for QS} \end{cases},$$

where

$$n = \begin{cases} \left\lfloor 4\left(\frac{T}{100}\right)^{2/9} \right\rfloor, \left\lfloor 12\left(\frac{T}{100}\right)^{2/9} \right\rfloor & \text{for Bartlett} \\ \left\lfloor 4\left(\frac{T}{100}\right)^{4/25} \right\rfloor, \left\lfloor 12\left(\frac{T}{100}\right)^{4/25} \right\rfloor & \text{for Parzen} \\ \left\lfloor 3\left(\frac{T}{100}\right)^{2/25} \right\rfloor, \left\lfloor 4\left(\frac{T}{100}\right)^{2/25} \right\rfloor & \text{for QS} \end{cases}.$$

Hirukawa (2010): Yet another approach by Hirukawa (2010) is an analog to the bandwidth choice rule for probability density estimation by Sheather and Jones (1991). It is built on the AMSE-optimal bandwidth (15) by Newey and West (1994). Two remarkable differences are: (i) that the normalized curvature $R^{(q)}$ is estimated nonparametrically using the same kernel $k(\cdot)$ but a different bandwidth b_T; and (ii) that the AMSE-optimal bandwidth $b_T^\dagger$ for the normalized curvature estimator

$$\hat{R}^{(q)}(b_T) = \frac{\sum_{j=-(T-1)}^{T-1} k(j/b_T)|j|^q \hat{\gamma}_j}{\sum_{j=-(T-1)}^{T-1} k(j/b_T)\hat{\gamma}_j}$$

is derived. It is

$$b_T^\dagger = \left\{ \frac{qk_q^2\beta^2(q)}{(2q+1)\int_{-\infty}^{\infty}|x|^{2q}k^2(x)dx} \right\}^{1/(4q+1)} T^{1/(4q+1)}, \tag{17}$$

where $\beta(q) = \left(s^{(q)}/s^{(0)}\right)^2 - s^{(2q)}/s^{(0)}$.

Like Newey and West (1994), it is possible to conduct sequential estimation of the normalized curvature (first-stage) and the LRV (second-stage) by using estimates of $b_T^\dagger$ and $S_T^\dagger$. Instead, in the spirit of Sheather and Jones (1991), Hirukawa (2010) adopts the solve-the-equation plug-in (SEPI) approach as the algorithm for estimating $S_T^\dagger$. It starts from constructing a system of two equations with two unknowns $\left(S_T^\dagger, b_T^\dagger\right)$. Solving (15) for T gives

$$T = \left[\frac{\int_{-\infty}^{\infty} k^2(x)dx}{qk_q^2\left\{R^{(q)}\right\}^2} \right] \left(S_T^\dagger\right)^{2q+1}.$$

Substituting this into (17) and invoking that $R^{(q)} = s^{(q)}/s^{(0)}$, we have

$$b_T^\dagger = b_T^\dagger\left(S_T^\dagger\right) = \left\{ \frac{\delta^2(q)\int_{-\infty}^{\infty} k^2(x)dx}{(2q+1)\int_{-\infty}^{\infty}|x|^{2q}k^2(x)dx} \right\}^{1/(4q+1)} \left(S_T^\dagger\right)^{(2q+1)/(4q+1)}, \tag{18}$$

where $\delta(q) = s^{(q)}/s^{(0)} - s^{(2q)}/s^{(q)}$. Finally, replacing $R^{(q)}$ in (15) with $\hat{R}^{(q)}\left(b_T^\dagger\right) = \hat{R}^{(q)}\left(b_T^\dagger\left(S_T^\dagger\right)\right)$ yields

$$S_T^\dagger = \left[\frac{qk_q^2\left\{\hat{R}^{(q)}\left(b_T^\dagger\left(S_T^\dagger\right)\right)\right\}^2}{\int_{-\infty}^{\infty} k^2(x)dx} \right]^{1/(2q+1)} T^{1/(2q+1)}. \tag{19}$$

The SEPI approach numerically solves the system of equations (18) and (19) for $S_T^\dagger$. Notice that this does not work for the HAR estimator using the QS kernel, because Table 1 indicates that $\int_{-\infty}^{\infty}|x|^{2q}k_{QS}^2(x)dx = \infty$.

To operationalize this algorithm, we replace the unknown quantity $\delta(q)$ with its proxy from fitting an AR(1) model to h_t, as in Andrews (1991). Sheather and Jones (1991) justify fitting a parametric model at this stage by arguing that doing so is less crucial than fitting it directly to $R^{(q)}$. The proxies of $\delta(q)$ for $q = 1, 2$ using the OLS estimate of the autoregressive parameter $\hat{\rho}$ are $\hat{\delta}(1) = \left(\hat{\rho}^2+1\right)/\left(\hat{\rho}^2-1\right)$ and $\hat{\delta}(2) = -\left(\hat{\rho}^2+8\hat{\rho}+1\right)/\left(\hat{\rho}-1\right)^2$. Let $\hat{S}_T^\dagger$ be the estimate of $S_T^\dagger$. Also denote the estimate of $b_T^\dagger$ as $\hat{b}_T^\dagger = \hat{b}_T^\dagger\left(\hat{S}_T^\dagger\right)$. Then, the final form of the system is given by

$$\begin{cases} \hat{S}_T^\dagger = 1.1447\left[\left\{\hat{R}_{BT}^{(1)}\left(\hat{b}_T^\dagger\left(\hat{S}_T^\dagger\right)\right)\right\}^2\right]^{1/3} T^{1/3} \\ \hat{b}_T^\dagger\left(\hat{S}_T^\dagger\right) = 1.2723\left\{\hat{\delta}^2(1)\right\}^{1/5}\left(\hat{S}_T^\dagger\right)^{3/5} \end{cases} \quad \text{for Bartlett}$$

$$\begin{cases} \hat{S}_T^\dagger = 2.6614\left[\left\{\hat{R}_{PZ}^{(2)}\left(\hat{b}_T^\dagger\left(\hat{S}_T^\dagger\right)\right)\right\}^2\right]^{1/5} T^{1/5} \\ \hat{b}_T^\dagger\left(\hat{S}_T^\dagger\right) = 1.4812\left\{\hat{\delta}^2(2)\right\}^{1/9}\left(\hat{S}_T^\dagger\right)^{5/9} \end{cases} \quad \text{for Parzen}.$$

By construction, the SEPI approach inherits properties of both Andrews' (1991) parametric and Newey and West's (1994) nonparametric approaches. While fitting a parametric model to the unknown process h_t is required for implementation, the rule itself is built on a nonparametric estimator of the normalized curvature, which limits the influence of the parametric reference. Nice finite-sample properties of HAR estimates combined with the SEPI approach are reported in Hirukawa (2010, 2011). Disadvantages of the approach are excluding the QS kernel and solving the highly nonlinear system numerically. For the latter, there may be multiple roots. In such cases, Hirukawa (2010) recommends taking the largest root as $\hat{S}_T^\dagger$.

4.2.3. Prewhitening

The process $\mathbf{g}_t$ computed from economic and financial data is often highly persistent. Essentially, its spectral density has a sharp peak at frequency zero, and as a result, covariance estimates from HAR estimators tend to be imprecise. It has long been known in the literature on statistical time series analysis (e.g., Press and Tukey, 1956; Blackman and Tukey, 1958) that the spectral density at frequency zero is easier to estimate if it is flat in the vicinity of the origin. This motivates us to filter out the original process $\mathbf{g}_t$ via a simple time series model in order to estimate the spectral density of the filtered process at frequency zero more easily. *Prewhitening* studied by Andrews and Monahan (1992) and Lee and Phillips (1994) is built on the above idea. The procedure of recovering the spectral density of the original process from the one of the filtered process is called *recoloring*. Several simulation results justify prewhitening at the beginning and recoloring at the end of HAR estimation by indicating better finite-sample properties.

Our focus is on VAR-based prewhitening by Andrews and Monahan (1992), which is most popular in practice. Below the HAR estimation with prewhitening and recoloring is described in a step-by-step manner.

1. **Prewhitening:** Fit a low-order VAR model to the original process $\mathbf{g}_t$, and obtain OLS estimates of autoregressive coefficient matrices and the VAR residuals. In the example of a VAR(1) model $\mathbf{g}_t = \hat{\mathbf{A}}\mathbf{g}_{t-1} + \hat{\epsilon}_t,\ t = 2, \ldots, T$, $\hat{\mathbf{A}}$ is the OLS estimate of the coefficient matrix and the $\hat{\epsilon}_t$ are the VAR residuals.
2. **HAR estimation:** Compute the LRV estimate of the residual $\hat{\epsilon}_t$ by HAR estimation in combination of a prespecified kernel $k_{\mathcal{J}}(\cdot)$ with one of its corresponding bandwidth choice methods discussed in Section 4.2.2. The resulting LRV estimate using is denoted as $\hat{\mathbf{\Omega}}^{\hat{\epsilon}}_{\mathcal{J}}$.
3. **Recoloring:** Obtain the LRV estimate of $\mathbf{g}_t$ as

$$\hat{\mathbf{\Omega}}_{\mathcal{J}-PW} = \left(\mathbf{I}_\ell - \hat{\mathbf{A}}\right)^{-1} \hat{\mathbf{\Omega}}^{\hat{\epsilon}}_{\mathcal{J}} \left\{\left(\mathbf{I}_\ell - \hat{\mathbf{A}}\right)^{\top}\right\}^{-1}.$$

There are a few remarks in order. First, the sole purpose of prewhitening is to flatten the spectral density of the filtered process in the neighborhood of frequency zero. Therefore, the time series model used for prewhitening does not need to be correctly specified, unlike the VARHAC estimator in Section 3.2. Second, Lee and Phillips (1994) propose yet another prewhitening method based on an autoregressive moving-average (ARMA) model. Orders and coefficient estimates of the ARMA model are determined by the Hannan-Rissanen recursion (Hannan and Rissanen, 1982). This procedure is rarely applied in practice, because it is designed for a scalar process. Extending it to a vector process and fitting a VARMA model would incur additional computational cost in the MA part, as mentioned in Section 3.1.

4.3. Estimator without Smoothing

Robinson (1998) is concerned with using a single bandwidth for all elements of a kernel-smoothed HAR estimator (13). So far it has been implicitly assumed that all elements of a HAR estimator are smoothed by a single bandwidth. This practice looks unavoidable, because varying bandwidth numbers across elements destroys the property of generating PSD estimates even when the Bartlett, Parzen or QS kernel is employed. Actually, the practice may be suboptimal in that the same, fixed amount of smoothing is imposed on all elements in the spectral matrix that are likely to have a variety of shapes. The concern motivates Robinson (1998) to propose yet another LRV estimator that is based on sample autocovariances but free of kernel smoothing.

Robinson's (1998) LRV estimator is built on the process corresponding to (5) with $\mathbf{Z}_t = \mathbf{R}_t \otimes \mathbf{I}_r$. More specifically, let the $\ell \times 1$ process be $\mathbf{g}_t = \mathbf{u}_t \otimes \mathbf{R}_t$, where $\mathbf{u}_t$ is an $r \times 1$ vector of regression errors with $E(\mathbf{u}_t) = \mathbf{0}_{r\times 1}$, $\mathbf{R}_t$ is an $h \times 1$ vector of instruments, and $\ell = hr$. Let the sample autocovariances of $\mathbf{u}_t$ and $\mathbf{R}_t$ be $\hat{\mathbf{\Gamma}}^{\mathbf{u}}_j = (1/T)\sum_{t=1\vee(1+j)}^{(T+j)\wedge T} \mathbf{u}_t\mathbf{u}_{t-j}^{\top}$ and $\hat{\mathbf{\Gamma}}^{\mathbf{R}}_j = (1/T)\sum_{t=1\vee(1+j)}^{(T+j)\wedge T} \left(\mathbf{R}_t - \bar{\mathbf{R}}\right)\left(\mathbf{R}_{t-j} - \bar{\mathbf{R}}\right)^{\top}$, respectively, where $\bar{\mathbf{R}} = (1/T)\sum_{t=1}^{T} \mathbf{R}_t$. The LRV estimator is then defined as

$$\hat{\mathbf{\Omega}}_R = \sum_{j=-(T-1)}^{T-1} \hat{\mathbf{\Gamma}}^{\mathbf{u}}_j \otimes \hat{\mathbf{\Gamma}}^{\mathbf{R}}_j.$$

Further advantages of $\hat{\mathbf{\Omega}}_R$ are that it is $T^{1/2}$-consistent and that it necessarily generates PSD estimates.

A disadvantage of this estimator is that its consistency relies on the condition

$$E\left\{(\mathbf{u}_t \otimes \mathbf{R}_t)\left(\mathbf{u}_{t-j}^{\top} \otimes \mathbf{R}_{t-j}^{\top}\right)\right\} = E\left(\mathbf{u}_t\mathbf{u}_{t-j}^{\top}\right) \otimes E\left(\mathbf{R}_t\mathbf{R}_{t-j}^{\top}\right)$$

for all j, which does not allow for heteroskedasticity in $\mathbf{u}_t$. Furthermore, all elements of $\mathbf{u}_t$ and $\mathbf{R}_t$ must be stochastic, and thus the natural instrument "1" cannot enter as an element of $\mathbf{R}_t$.

4.4. A Real Data Example

It is of practical importance and interest to see how different HAR estimators produce different LRV estimates. The test for uncovered interest parity (UIP) by Hansen and Hodrick (1980) is chosen as an example. The test is based on the regression

Table 2
Results of Predictive Regressions and UIP Tests

Regressand		$\hat{\beta}_0$	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_3$	$\hat{\beta}_5$	Wald	[p-value]	BW
GBP:		0.0033	0.0300	0.3002	−0.1473	−0.0859			
	TR-HH	(0.0058)	(0.0920)	(0.2278)	(0.1714)	(0.0582)	3.9534	[0.5561]	2.00
	QS-A	(0.0055)	(0.0724)	(0.2385)	(0.1697)	(0.0545)	4.4548	[0.4860]	9.61
	BT-NW	(0.0051)	(0.0840)	(0.2055)	(0.1567)	(0.0517)	4.9356	[0.4238]	3.94
	PZ-H	(0.0052)	(0.0859)	(0.2098)	(0.1599)	(0.0531)	4.7120	[0.4520]	5.47
FFR:		0.0017	−0.0269	0.2760	−0.0344	−0.0283			
	TR-HH	(0.0051)	(0.0971)	(0.1797)	(0.1416)	(0.0596)	5.8645	[0.3196]	2.00
	QS-A	(0.0047)	(0.0916)	(0.1721)	(0.1270)	(0.0536)	6.7428	[0.2405]	8.18
	BT-NW	(0.0041)	(0.0827)	(0.1549)	(0.1278)	(0.0500)	8.5216	[0.1297]	2.28
	PZ-H	(0.0045)	(0.0889)	(0.1665)	(0.1341)	(0.0541)	7.1868	[0.2071]	4.61
DM:		−0.0025	0.0119	0.2990	−0.1524	−0.0059			
	TR-HH	(0.0054)	(0.1046)	(0.2286)	(0.1695)	(0.0774)	3.0711	[0.6890]	2.00
	QS-A	(0.0052)	(0.1080)	(0.2105)	(0.1664)	(0.0543)	4.2387	[0.5156]	9.47
	BT-NW	(0.0036)	(0.0805)	(0.1679)	(0.1320)	(0.0546)	6.1352	[0.2933]	1.30
	PZ-H	(0.0051)	(0.1046)	(0.2228)	(0.1757)	(0.0600)	3.6028	[0.6079]	11.90
JPY:		0.0074	−0.1214	0.2015	−0.3210	0.5616			
	TR-HH	(0.0063)	(0.1154)	(0.3073)	(0.2280)	(0.0950)	48.2608	[0.0000]	2.00
	QS-A	(0.0052)	(0.1069)	(0.3302)	(0.2182)	(0.1222)	27.7054	[0.0000]	11.11
	BT-NW	(0.0051)	(0.1062)	(0.3070)	(0.2086)	(0.1071)	35.7203	[0.0000]	9.55
	PZ-H	(0.0053)	(0.1079)	(0.3132)	(0.2118)	(0.1085)	34.8245	[0.0000]	12.92

Note: Numbers in parentheses, "Wald", "p-value", and "BW/LO" for each LRV estimator are standard errors, values of Wald statistics, their associated p-values from $\chi^2(5)$, and automatic bandwidths $\tilde{S}_T^+$ (for the QS, Bartlett and Parzen estimators) or lag orders n (for the truncated estimator), respectively.

of the forecast error of the exchange rate of a currency c on lagged forecast errors of own and other currencies

$$s_t^c - f_{t-3,3}^c = \beta_0^c + \sum_{j=1}^{4} \beta_j^c \left(s_{t-3}^j - f_{t-6,3}^j \right) + \epsilon_t^c,$$

where s_t^c and $f_{t,3}^c$ are logarithms of spot and 3-month-forward exchange rates of a currency c at time t, respectively. We work with 352 monthly observations from January 1973 to December 1999 in the data set "`fxeff_m.prn`" available on the webpage of Mark (2001). Four exchange rates (expressed in the US dollar), namely, pound sterling (GBP; $c = 1$), French franc (FFR; $c = 2$), Deutsche mark (DM; $c = 3$), and Japanese yen (JPY; $c = 4$), are examined.

The regression can be estimated by OLS, and then the UIP test for the null hypothesis $H_0^c : \beta_j^c = 0$ for $j = 0, \ldots, 4$ is conducted. To compute standard errors of OLS estimates and Wald statistics for the testing of H_0^c, we consider the following four HAR estimators: (i) the truncated estimator with 2 lags employed by Hansen and Hodrick (1980), reflecting that the error term ϵ_t^c behaves like a MA(2) process [TR-HH]; (ii) the QS HAR estimator using Andrews' (1991) automatic bandwidth [QS-A]; (iii) the Bartlett HAR estimator using Newey and West's (1994) automatic bandwidth, where the bandwidth for the truncated normalized curvature estimator (16) is $n = \left\lfloor 4\left(\frac{T}{100}\right)^{2/9} \right\rfloor$ [BT-NW]; and (iv) the Parzen HAR estimator using Hirukawa's (2010) SEPI bandwidth [PZ-H].

Table 2 presents OLS estimates, standard errors, Wald statistics with associated p-values from $\chi^2(5)$, and automatic bandwidths or lag orders. While it is not possible to judge which HAR estimator is most reliable, it is safe to say that the results are largely influenced by the choice of the estimator.

4.5. Discussion

Practitioners' preference has long been given to kernel HAR estimators, and the Bartlett and QS estimators are most popular among all such estimators. This reflects that the convergence rate of a LRV estimator is not well transmitted to its finite-sample properties. Although the truncated and VARHAC estimators have faster convergence rates, simulation results indicate that the former often and the latter sometimes perform more poorly than the slower QS estimator, which is occasionally outperformed by the even slower (and thus theoretically inferior) Bartlett estimator.

However, test statistics using the Bartlett and QS estimators often exhibit unsatisfactory finite-sample performances. In particular, substantial size distortions of test statistics in the presence of strong positive dependence are reported and extensively discussed in recent work (e.g., Kiefer and Vogelsang, 2005; Sul et al., 2005).

While the bandwidth determines finite-sample performance of a HAR estimator, there is uncertainty in the bandwidth choice itself. Even if the bandwidth is chosen in a data-driven manner, practitioners must make some kind of arbitrary decision (e.g., which rule given in Section 4.2.2 to choose and how to estimate the normalized curvature). Moreover, there is no guarantee that the AMSE-optimal bandwidth is equally optimal for parameter estimation or statistical inference. From this viewpoint, two object-oriented (i.e., test- and estimation-optimal) bandwidth choices, as well as several other topics, will be discussed in the next section.

5. Recent Developments in HAR Estimation and Inference

While much material in the previous sections is included in past surveys, research on HAR estimation and inference continues. This section is devoted to developments over the past two decades that are not covered before.

5.1. Kernel HAR Estimators with Faster Convergence Rates

A faster convergence rate in kernel HAR estimation is a theoretically appealing property. A HAR estimator having an accelerated convergence rate is more valuable if it generates PSD estimates. Below we present two examples of such HAR estimators.

5.1.1. Nonparametrically Prewhitened Estimator

The first estimator, proposed by Xiao and Linton (2002) and Hirukawa (2006), attains the rate improvement via a multiplicative bias correction (MBC) of the spectral matrix estimator in the frequency domain. The MBC method, originally proposed by Jones et al. (1995) for probability density estimation, is based on the identity $f(\cdot) \equiv \tilde{f}(\cdot)\left\{f(\cdot)/\tilde{f}(\cdot)\right\} = \tilde{f}(\cdot)\alpha(\cdot)$ for $\tilde{f}(\cdot) > 0$, where $f(\cdot)$ and $\tilde{f}(\cdot)$ are the true density and its initial estimator, respectively, and $\alpha(\cdot)$ can be recognized as the bias correction term. It is anticipated that replacing $\alpha(\cdot)$ with its consistent estimate can yield a less-biased estimator than $\tilde{f}(\cdot)$, and non-negativity of the bias-corrected estimator is guaranteed by construction. The MBC procedure is reminiscent to prewhitening when it is applied to spectral estimation. Then, Xiao and Linton (2002) call the MBC-HAR estimator the nonparametrically prewhitened (NPW) estimator.

Because the MBC is originally designed for scalar quantities, transplanting it to spectral matrix estimation should be done with care. Hirukawa (2006) proposes the following procedure of computing the NPW estimator. Throughout the kernel $k(\cdot)$ is assumed to have the characteristic exponent $q = 2$ and the spectral window generator $K(\cdot) \geq 0$. For such $K(\cdot)$ and the bandwidth S_T, the amplitude window (Parzen, 1963) is given by $K_{S_T}(\lambda) = S_T \sum_{j=-\infty}^{\infty} K\{S_T(\lambda + 2\pi j)\} \approx S_T K(S_T\lambda)$. Then, the spectral matrix of $\mathbf{g}_t$ evaluated at frequency $\omega \in (-\pi, \pi)$ can be estimated as

$$\tilde{f}_{\mathbf{gg}}(\omega) = \frac{2\pi}{T} \sum_{\lambda_j \in B(\omega)} K_{S_T}(\lambda_j - \omega) I_{\mathbf{gg}}(\lambda_j),$$

where $I_{\mathbf{gg}}(\lambda_j) = \zeta_{\mathbf{g}}(\lambda_j)\zeta_{\mathbf{g}}(\lambda_j)^*$ is the periodogram with $\zeta_{\mathbf{g}}(\lambda_j) = (2\pi T)^{-1/2} \sum_{t=1}^{T} \mathbf{g}_t e^{-it\lambda_j}$ being the finite Fourier transform of $\mathbf{g}_t$ evaluated at fundamental frequencies $\lambda_j = 2\pi j/T$ for $j = 0, \pm 1, \ldots, \pm\lfloor T/2 \rfloor$, and $B(\omega) = \left\{\lambda_j \middle| \omega - \pi < \lambda_j < \omega + \pi\right\}$. Observe that $\tilde{f}_{\mathbf{gg}}(\omega)$ is Hermitian and positive definite (PD). Hence, $\tilde{f}_{\mathbf{gg}}(\omega)$ has eigenvalues $\lambda_1, \ldots, \lambda_\ell > 0$, and thus $\tilde{f}_{\mathbf{gg}}^{1/2}(\omega)$ is well-defined by the unitary decomposition $\tilde{f}_{\mathbf{gg}}^{1/2}(\omega) = \mathbf{U}\mathbf{\Lambda}^{1/2}\mathbf{U}^*$, where $\mathbf{\Lambda}^{1/2} = \text{diag}\left\{\lambda_1^{1/2}, \ldots, \lambda_\ell^{1/2}\right\}$ is the diagonal matrix containing the square roots of the eigenvalues, and $\mathbf{U}$ is the unitary matrix, i.e., $\mathbf{U}\mathbf{U}^* = \mathbf{I}_\ell$. In the end, the NPW estimator can be defined as

$$\hat{\mathbf{\Omega}}_{NPW} = \tilde{\mathbf{\Omega}}^{1/2}\tilde{\alpha}(0)\tilde{\mathbf{\Omega}}^{1/2},$$

where $\tilde{\mathbf{\Omega}}^{1/2} = \sqrt{2\pi}\,\tilde{f}_{\mathbf{gg}}^{1/2}(0)$ is the square root of the initial HAR estimator, and

$$\tilde{\alpha}(0) = \sum_{\lambda_j \in B(0)} K_{S_T}(\lambda_j) \frac{2\pi}{T} \tilde{f}_{\mathbf{gg}}^{-1/2}(\lambda_j) I_{\mathbf{gg}}(\lambda_j) \tilde{f}_{\mathbf{gg}}^{-1/2}(\lambda_j)$$

serves as the bias correction term. $\hat{\mathbf{\Omega}}_{NPW}$ is PSD by construction.

For consistency of $\hat{\mathbf{\Omega}}_{NPW}$, let the bandwidth satisfy $1/S_T + S_T^4/T \to 0$. Also define the MSE of $\hat{\mathbf{\Omega}}_{NPW}$ as $E\left\{\text{vec}\left(\hat{\mathbf{\Omega}}_{NPW} - \mathbf{\Omega}\right)^\top \mathbf{W}\,\text{vec}\left(\hat{\mathbf{\Omega}}_{NPW} - \mathbf{\Omega}\right)\right\}$ for some $\ell^2 \times \ell^2$ symmetric PSD weighting matrix $\mathbf{W}$, as in Andrews (1991). Under sufficient smoothness of the spectral density at frequency zero, the AMSE is given by

$$ABias^2 + AVar = (2\pi)^2 k_2^4\,\text{vec}\{\mathbf{\Upsilon}(0)\}^\top \mathbf{W}\,\text{vec}\{\mathbf{\Upsilon}(0)\} S_T^{-8} + \text{tr}\{\mathbf{W}(\mathbf{\Omega} \otimes \mathbf{\Omega})(\mathbf{I}_{\ell^2} + \mathbf{K}_{\ell\ell})\} \int_{-\infty}^{\infty} t_k^2(x)dx \left(\frac{S_T}{T}\right),$$

where

$$\mathbf{\Upsilon}(\omega) = f_{\mathbf{gg}}^{1/2}(\omega)\left\{\frac{d^2}{d\omega^2}\mathbf{\Xi}(\omega)\right\} f_{\mathbf{gg}}^{1/2}(\omega),$$

$$\mathbf{\Xi}(\omega) = f_{\mathbf{gg}}^{-1/2}(\omega)\left\{\frac{d^2}{d\omega^2} f_{\mathbf{gg}}(\omega)\right\} f_{\mathbf{gg}}^{-1/2}(\omega),$$

$t_k(x) = \int_{-\infty}^{\infty} T_K(\lambda) e^{ix\lambda} d\lambda$, and $T_K(\lambda) = 2K(\lambda) - K \circ K(\lambda)$ is the fourth-order spectral window obtained by twicing (Stuetzle and Mittal, 1979) with $K \circ K(\lambda) = \int_{-\infty}^{\infty} K(t)K(\lambda - t)dt$ being the convolution of $K(\lambda)$. Observe that the NPW estimator ac-

celerates the bias convergence from $O\left(S_T^{-2}\right)$ to $O\left(S_T^{-4}\right)$ while maintaining the variance convergence of $O(S_T/T)$. The AMSE-optimal bandwidth is

$$S_T^{\dagger} = \left[\frac{8k_2^4 \operatorname{vec}\{\boldsymbol{\Upsilon}(0)\}^{\top}\mathbf{W}\operatorname{vec}\{\boldsymbol{\Upsilon}(0)\}}{\operatorname{tr}\{\mathbf{W}(f_{\mathbf{gg}}(0)\otimes f_{\mathbf{gg}}(0))(\mathbf{I}_{\ell^2}+\mathbf{K}_{\ell\ell})\}\int_{-\infty}^{\infty}t_k^2(x)dx}\right]^{1/9} T^{1/9}.$$

Therefore, $\hat{\boldsymbol{\Omega}}_{NPW}$ can attain the rate of convergence $T^{4/9}$ when best implemented. It is worth remarking that this result does not contradict Andrews' (1991) optimality argument of the QS kernel, because the NPW estimator falls outside his framework.

Simulation results in Hirukawa (2006) indicate that the NPW estimator often yields more accurate LRV estimates than conventional HAR estimators. A disadvantage of this estimator is that its computational burden increases linearly with T, as $\left\{\tilde{f}_{\mathbf{gg}}^{1/2}\left(\lambda_j\right)\right\}_{j=-\lfloor T/2\rfloor}^{\lfloor T/2\rfloor}$ are required for the bias-correction term $\tilde{\alpha}(0)$.

5.1.2. Flat-Top Kernel Estimator

Politis (2011) investigates yet another HAR estimator with a faster convergence rate. The entire estimation procedure contains three novel features. These are: (i) using the flat-top kernel; (ii) allowing for different bandwidths in different elements of the estimate; and (iii) proposing a new procedure of estimating the optimal bandwidth.

We take a closer look at each feature. For (i), the flat-top kernel is defined as

$$k_{FT}(x) = \begin{cases} 1 & \text{for } |x| \le c \\ g(x) & \text{otherwise} \end{cases},$$

where $c > 0$ is a parameter, the function $g : \mathbb{R} \to [-1, 1]$ is symmetric, continuous at all but a finite number of points and satisfying $g(c) = 1$ and $\int_{-\infty}^{\infty} g^2(x)dx < \infty$. A special case of $k_{FT}(x)$ is the trapezoidal kernel

$$k_{TZ}(x) = \begin{cases} 1 & \text{for } |x| \le c \\ (|x|-1)/(c-1) & \text{for } c < |x| \le 1 \\ 0 & \text{otherwise} \end{cases}.$$

The property of this kernel becomes closer to those of the truncated and Bartlett kernels if $c \approx 1$ and $c \approx 0$, respectively. Politis (2011, pp.715-716) presents other special cases of $k_{FT}(x)$ including analogs to the Parzen and QS kernels.

For (ii) and (iii), the (a, b) element of the flat-top HAR estimator $\tilde{\boldsymbol{\Omega}}_{FT}$ is given by

$$\tilde{\boldsymbol{\Omega}}_{FT,a,b} = \sum_{j=-(T-1)}^{T-1} k_{FT}\left(\frac{j}{S_{T,a,b}}\right)\hat{\boldsymbol{\Gamma}}_{j,a,b},$$

where $\hat{\boldsymbol{\Gamma}}_{j,a,b}$ is the (a, b) element of $\hat{\boldsymbol{\Gamma}}_j$. The bandwidth $S_{T,a,b}$ is chosen elementwise by inspecting correlograms or cross-correlograms, and the convergence rate of the HAR estimator is also derived in an element-by-element manner. This may be thought of as a resolution for Robinson's (1998) concern in Section 4.3. Politis (2011) establishes that $\tilde{\boldsymbol{\Omega}}_{FT}$ can attain the rate of convergence near $T^{1/2}$.

The flat-top kernel has the characteristic exponent $q = \infty$, like the truncated kernel. Hence, $\tilde{\boldsymbol{\Omega}}_{FT}$ does not necessarily generate PSD estimates under a single bandwidth. Assigning different bandwidths for different elements makes the non-PSD problem even more complicated. To generate PSD estimates while maintaining the near-parametric convergence rate, Politis (2011) advocates replacing the diagonal matrix of eigenvalues $\boldsymbol{\Lambda} = \operatorname{diag}\{\lambda_1, \ldots, \lambda_\ell\}$ in the unitary decomposition $\tilde{\boldsymbol{\Omega}}_{FT} = \mathbf{U}\boldsymbol{\Lambda}\mathbf{U}^*$ with $\boldsymbol{\Lambda}^+ = \operatorname{diag}\left\{\lambda_1^+, \ldots, \lambda_\ell^+\right\}$, where $\mathbf{U}$ is the unitary matrix, $\lambda_a^+ = \max\{\lambda_a, \epsilon_T\}$ for $a = 1, \ldots, \ell$, and $\epsilon_T = T^{-\alpha}$ for some $\alpha \in [1, 2]$. Politis (2011) finally demonstrates that the clipped flat-top estimator $\hat{\boldsymbol{\Omega}}_{FT} = \mathbf{U}\boldsymbol{\Lambda}^+\mathbf{U}^*$ still maintains the convergence rate near $T^{1/2}$.

The flat-top estimator has several theoretically attractive properties. In addition, extensive simulations in Politis (2011) indicate that the estimator often exhibits better finite-sample performances than those using conventional kernels. Nonetheless, a difficulty in using it in practice is to choose quite a few tuning parameters. These include: c and $g(x)$ for the shape of the kernel; C_0 and K_T in Politis (2011, p.718) for the bandwidth selection; and the exponent α in ϵ_T for the PSD correction.

5.2. Non-Autocovariance-Based Estimators

Taking a (weighted) sum of autocovariances is not the only way of handling heteroskedasticity and temporal dependence of unknown form. Each of the nonparametric LRV estimators below is free of computing sample autocovariances.

5.2.1. Smoothed-Moments Estimator

The first class of LRV estimators that does not depend on sample autocovariances is investigated by Smith (2005) in the framework of efficient GMM estimation. It still relies on a kernel, but smoothing is made for the error process (or the

moment function) $\mathbf{g}_t = \mathbf{g}_t(\theta)$ itself, not for its autocovariances. For a kernel $k(\cdot)$ satisfying Assumption 1 and a bandwidth S_T, the smoothed moment function is defined as

$$\tilde{\mathbf{g}}_t = \tilde{\mathbf{g}}_t(\theta) = \sum_{j=t-T}^{t-1} k\left(\frac{j}{S_T}\right)\mathbf{g}_{t-j}(\theta),\ t = 1, \ldots, T.$$

The idea of smoothing $\mathbf{g}_t(\theta)$ can be also found in the context of the generalized empirical likelihood estimation using time series data. Because the likelihood function by construction ignores temporal dependence in $\mathbf{g}_t(\theta)$, we run the nonparametric likelihood based on its smoothed counterpart $\tilde{\mathbf{g}}_t(\theta)$ for valid inference; see Kitamura (1997), Kitamura and Stutzer (1997), and Smith (1997, 2011) for more details.

In the context of efficient GMM estimation, Smith (2005) defines the LRV estimator as the normalized outer-product of the smoothed process

$$\hat{\mathbf{\Omega}}_S = \hat{\mathbf{\Omega}}_S\left(\tilde{\theta}\right) = \frac{(1/T)\sum_{t=1}^{T}\tilde{\mathbf{g}}_t\left(\tilde{\theta}\right)\tilde{\mathbf{g}}_t\left(\tilde{\theta}\right)^{\top}}{\sum_{j=1-T}^{T-1}k^2(j/S_T)},$$

where $\tilde{\theta}$ is some initial consistent GMM estimate. By construction this estimator necessarily generates PSD estimates. An interesting feature of this LRV estimator is that the efficient GMM estimator (6) based on $\hat{\mathbf{\Omega}}_S^{-1}$ as the optimal weighting matrix is first-order asymptotically equivalent to the one based on the inverse of the HAR estimator using the same initial estimate $\tilde{\theta}$, the *induced kernel* $k^\star(x) = \int_{-\infty}^{\infty}k(u-x)k(u)du/\int_{-\infty}^{\infty}k^2(u)du$ and the same bandwidth S_T. Smith (2011) shows that induced kernels from the truncated and Bartlett kernels are the Bartlett and Parzen kernels, respectively, via the relation between their spectral window generators $K^\star(\cdot) = 2\pi|K(\cdot)|^2$.

Smith (2005) assumes that a single kernel and bandwidth are used for all elements of $\mathbf{g}_t(\theta)$. In fact, even if we employ different kernels and/or bandwidths across elements like the VARHAC estimator, the resulting estimator is still PSD.

A potential disadvantage of this estimator is that Smith (2005) does not consider its implementation method. The bandwidth S_T for the estimator using a kernel $k(\cdot)$ may be chosen in an analogous manner to its first-order asymptotically equivalent HAR estimator using the induced kernel $k^\star(\cdot)$. However, no simulation results of the LRV estimator are available in Smith (2005), and thus how such a bandwidth choice method is transmitted to finite-sample properties of the estimator is unknown.

5.2.2. Orthonormal Series Estimator

The second class of LRV estimators with no autocovariances is based on an orthonormal series (OS) approximation to the process $\mathbf{g}_t$, and it can be computed as the outer-product of its explained part. This class of estimators is first proposed in the literature by Phillips (2005) and later investigated in the context of HAR inference by Müller (2007) and Sun (2011).

The OS-HAR estimator can be constructed as follows. Let $\{\phi_k(r)\}_{k=1}^{\infty}$ be an orthonormal sequence. For a collection $\{\phi_k(t/T)\}_{k=1}^{K}$ with $K \in \mathbb{N}$ formed by the first K members of the orthonormal sequence, define $\hat{\mathbf{\Lambda}}_k = \left(1/\sqrt{T}\right)\sum_{t=1}^{T}\phi_k(t/T)\mathbf{g}_t$ and $\hat{\mathbf{\Omega}}_k = \hat{\mathbf{\Lambda}}_k\hat{\mathbf{\Lambda}}_k^{\top}$. $\hat{\mathbf{\Lambda}}_k$ is approximately the regression coefficient obtained by regressing $\mathbf{g}_t$ on the regressor $\phi_k(t/T)/\sqrt{T}$, and $\hat{\mathbf{\Omega}}_k$ can be interpreted as the part of the total sum of squares $\sum_{t=1}^{T}\mathbf{g}_t\mathbf{g}_t^{\top}$ explained by this regression. The OS-HAR estimator is finally given by

$$\hat{\mathbf{\Omega}}_{OS} = \frac{1}{K}\sum_{k=1}^{K}\hat{\mathbf{\Omega}}_k. \tag{20}$$

As the sequence $\phi_k(t/T)$, Phillips (2005), Müller (2007) and Sun (2011) take the eigenvectors of the covariance kernel of Brownian motion $\sqrt{2}\sin\{\pi(k-1/2)t/T\}$, the type II discrete cosine transform $\sqrt{2}\cos\{\pi k(t-1/2)/T\}$, and a subset of cosine functions $\sqrt{2}\cos(2\pi kt/T)$, respectively. $\hat{\mathbf{\Omega}}_{OS}$ is PSD by construction.

For consistency of $\hat{\mathbf{\Omega}}_{OS}$, let the number of basis functions $K = K_T$ satisfy $1/K_T + K_T/T \to 0$. Also define the MSE of $\hat{\mathbf{\Omega}}_{OS}$ as $E\{\mathrm{vec}\left(\hat{\mathbf{\Omega}}_{OS} - \mathbf{\Omega}\right)^{\top}\mathbf{W}\,\mathrm{vec}(\hat{\mathbf{\Omega}}_{OS} - \mathbf{\Omega})\}$ for some $\ell^2 \times \ell^2$ symmetric PSD weighting matrix $\mathbf{W}$, as in Andrews (1991). The AMSE implied by this definition is

$$ABias^2 + AVar = \mathrm{vec}\,(\mathbf{B})^{\top}\mathbf{W}\,\mathrm{vec}\,(\mathbf{B})\left(\frac{K_T}{T}\right)^4 + \mathrm{tr}\,\{\mathbf{W}(\mathbf{\Omega}\otimes\mathbf{\Omega})(\mathbf{I}_{\ell^2} + \mathbf{K}_{\ell\ell})\}K_T^{-1},$$

where $\mathbf{B} = \lim_{T\to\infty}[(T/K_T)^2\{E(\hat{\mathbf{\Omega}}_{OS}) - \mathbf{\Omega}\}]$ is the dominant bias term of $\hat{\mathbf{\Omega}}_{OS}$. The AMSE-optimal number of basis functions is

$$K_T^{\dagger} = \left[\frac{\mathrm{tr}\,\{\mathbf{W}(\mathbf{\Omega}\otimes\mathbf{\Omega})(\mathbf{I}_{\ell^2} + \mathbf{K}_{\ell\ell})\}}{4\,\mathrm{vec}\,(\mathbf{B})^{\top}\mathbf{W}\,\mathrm{vec}\,(\mathbf{B})}\right]^{1/5}T^{4/5}.$$

Under this optimal number, $\hat{\mathbf{\Omega}}_{OS}$ attains the best possible convergence rate of $T^{2/5}$, which is the same as the one for the HAR estimator using the Parzen or QS kernel. Implementation of $K_T^{\dagger}$ can be conducted analogously to the methods in Section 4.2.2.

Asymptotic equivalence of kernel and OS-HAR estimators is argued in Phillips (2005) and Sun (2011). In particular, Phillips (2005) shows that $\hat{\mathbf{\Omega}}_{OS}$ using $\phi_k(t/T) = \sqrt{2}\sin\{\pi(k-1/2)t/T\}$ is asymptotically equivalent to the HAR estimator using the Daniell kernel

$$k_{DA}(x) = \frac{\sin(\pi x)}{\pi x}.$$

Therefore, approximations to the bias and variance of this OS-HAR estimator are the same as those of the Daniell HAR estimator.

5.3. HAR Inference

5.3.1. Fixed-b Asymptotics

Unsatisfactory finite-sample performance of test statistics using HAR estimators has long been a concern. Specializing in hypothesis testing of coefficients in linear regression models, Kiefer et al. (2000) and Kiefer and Vogelsang (2002) propose an alternative method of constructing test statistics from the Bartlett HAR estimator by setting the bandwidth $S_T = T$. Kiefer and Vogelsang (2005) extend the analysis to HAR estimators using a general class of kernels and the bandwidth $S_T = bT$ for some constant $b \in (0, 1]$. Such HAR estimators are no longer consistent, and thus have non-degenerate limiting distributions. To be more precise, $\hat{\mathbf{\Omega}} \xrightarrow{d} \mathbf{L}\mathbf{Q}(b)\mathbf{L}^\top$ for an $\ell \times \ell$ matrix $\mathbf{L}$ satisfying $\mathbf{\Omega} = \mathbf{L}\mathbf{L}^\top$, where, for instance,

$$\mathbf{Q}(b) = \begin{cases} \frac{2}{b}\int_0^1 \mathbf{B}(r)\mathbf{B}(r)^\top dr - \frac{1}{b}\int_0^{1-b}\left\{\mathbf{B}(r+b)\mathbf{B}(r)^\top + \mathbf{B}(r)\mathbf{B}(r+b)^\top\right\}dr & \text{for Bartlett} \\ -\frac{1}{b^2}\int_0^1\int_0^1 k''_{QS}\left(\frac{r-s}{b}\right)\mathbf{B}(r)\mathbf{B}(s)^\top dr ds & \text{for QS} \end{cases},$$

and $\mathbf{B}(r) = \mathbf{W}(r) - r\mathbf{W}(1)$ and $\mathbf{W}(r)$ for $r \in [0, 1]$ are $\ell \times 1$ vectors of Brownian bridges and standard Wiener processes, respectively. Nonetheless, the test statistics using inconsistent HAR estimators are well-defined and shown to obey non-standard limiting distributions that explicitly depend on the kernel and b. This alternative approximation is referred to as the *fixed-b asymptotics*. The conventional approximation corresponds to the case with $b \to 0 \Leftrightarrow S_T = o(T)$, and it is called the *small-b asymptotics*. The limiting distribution in this scenario is normal or chi-squared, regardless of the kernel and bandwidth.

Kiefer and Vogelsang (2005) in their simulations show that the fixed-b limiting distributions of test statistics can approximate their finite-sample distributions more accurately than the standard small-b limiting distributions. Jansson (2004) and Sun et al. (2008) confirm this finding theoretically in a Gaussian location model (i.e., $y_t = \beta + u_t$ with a linear Gaussian process u_t), demonstrating that the error in rejection probability (ERP) of the t-test using critical values from the nonstandard fixed-b limiting distribution is $O(\log T/T)$ and $O(1/T)$, respectively. Moreover, the fixed-b theory is extended beyond location and linear regression models. Examples include Sun (2011, 2014a, 2014b) for trend regression, first-step and efficient GMM frameworks, respectively, as well as the applications to be discussed in Section 5.4.

Kiefer and Vogelsang (2005) also find the size-power trade-off between small and large values of b. More specifically, a small b results in higher power at the cost of larger size distortions, whereas a large b leads to greater accuracy in size in exchange for lower power. A rationale is that size distortions (power losses) come from the bias (variance) of the LRV estimate due to a small (large) b. The size-power trade-off leads to the proposal of yet another HAR inference below.

5.3.2. Fixed-ρ Asymptotics for Exponentiated Kernels

To improve power of test statistics while retaining accuracy in size in the framework of no truncation in the bandwidth, Phillips et al. (2006, 2007) propose to exponentiate a kernel. For a kernel $k(\cdot)$ (called the *mother kernel*) and an exponent $\rho \in \mathbb{N}$, the exponentiated kernel is given by $k_\rho(\cdot) = k^\rho(\cdot)$. Widely used kernels including the Bartlett, Parzen and QS kernels can be employed as the mother kernel.

The bandwidth for HAR inference using exponentiated kernels is set equal to T, as in Kiefer et al. (2000) and Kiefer and Vogelsang (2002). The long bandwidth contributes to attractive size properties. The exponent ρ controls the amount of downweighting; as ρ increases, $k_\rho(\cdot)$ becomes more concentrated in the vicinity of the origin than the mother kernel $k(\cdot)$. Therefore, power properties improve with ρ.

Phillips et al. (2006, 2007) investigate two approximations analogous to small- and fixed-b asymptotics. First, letting $\rho \to \infty$ at a suitable rate (called the *large-ρ asymptotics*) leads to consistency of the HAR estimators and size distortions of test statistics, as in small-b cases. Second, a fixed ρ (called the *fixed-ρ asymptotics*) yields inconsistent HAR estimators. Like fixed-b cases, test statistics have greater accuracy in size than the conventional and large-ρ-based ones while preserving power. An intuition is that the asymptotic theory that reflects that ρ is always finite in practice may be able to provide a better approximation to the finite-sample distribution. Sun et al. (2011) establish that the ERP of the t-test for a Gaussian location model using critical values from the nonstandard fixed-ρ limiting distribution is again $O(1/T)$.

5.3.3. Test-Optimal Bandwidth Selection

Practitioners prefer a method that is simple and straightforward to implement, and HAR inference is not an exception. Although the fixed-b (and ρ) asymptotics theoretically refine HAR inference, practitioners may wonder how to choose b and what numbers should be reported as standard errors for parameter estimates. As regards the second question, under the

literally fixed-b setup, the HAR estimator is inconsistent and thus the resulting standard errors cannot be interpreted as estimates for standard deviations of the distributions of parameter estimates.

Sun et al. (2008) propose a method of choosing b for two-sided tests based on HAR estimators using a general class of kernels in a Gaussian location model. Assuming that $b = b_T$ satisfies $b_T + 1/(b_T T) \to 0$, Sun et al. (2008) adopt a weighted sum of Type I and Type II errors as the loss function. For the characteristic exponent of the kernel q, the loss function is of $O\left\{b_T + (b_T T)^{-q}\right\}$, and thus the test-optimal b_T is $b_T^\dagger = O\left(T^{-q/(q+1)}\right)$. The corresponding test-optimal bandwidth is given by $S_T^\dagger = b_T^\dagger T = O\left(T^{1/(q+1)}\right)$. Observe that this is larger by an order of magnitude than the AMSE-optimal bandwidth for the same HAR estimator $S_T^\dagger = O\left(T^{1/(2q+1)}\right)$.

While the test-optimal bandwidth by Sun et al. (2008) may be viewed as a purely theoretical result, Lazarus et al. (2018) take their idea one step further. A weighted sum of the squared size distortions and the squared size-adjusted power is chosen as their loss function, which is again of $O\left\{b_T + (b_T T)^{-q}\right\}$. Then, the test-optimal formula of b_T is derived under some weighting scheme and dependent structure of the error term in the Gaussian location model. From a theoretical point of view and extensive simulation evidence, Lazarus et al. (2018) recommend two practices. First, the combination of the OS-HAR estimator (20) using the type II discrete cosine transform $\phi_k(t/T) = \sqrt{2}\cos\{\pi k(t-1/2)/T\}$, the test-optimal number of basis functions $\hat{K}_T^\dagger = 0.4T^{2/3}$ and fixed-b critical values should be used for inference about a single parameter. Second, the combination of the Bartlett HAR estimator, the test-optimal bandwidth $\hat{S}_T^\dagger = 1.3T^{1/2}$ and fixed-b critical values should be preferred for high-dimensional inference with multiple restrictions. Indeed, simulation results in Lazarus et al. (2018) indicate that the recommended procedures substantially reduce size distortions observed in HAR inference under automatic bandwidths and normal approximations. It is also worth emphasizing that the HAR estimators based on the procedures by Lazarus et al. (2018) are consistent, and thus we may safely report the standard errors computed from the corresponding LRV estimates.

As a by-product, Lazarus et al. (2018, p.547) find that the QS kernel attains the minimum value of their loss function among all nonnegative kernels. Sun and Yang (2020) demonstrate that the test-optimality of the QS kernel continues to hold under the loss function by Sun et al. (2008).

5.3.4. HAR Inference Based on a Fixed Number of Basis Functions

The test-optimal method of choosing the bandwidth or the number of basis functions has been established in the asymptotic framework. Whether asymptotic results can be well transmitted in finite samples is a concern. Two inference procedures below are designed in the context of finite samples; these are built on some OS-HAR estimator using a fixed number of basis functions $K \in \mathbb{N}$.

The first procedure is Ibragimov and Müller's (2010) split-sample t-test for a single parameter. Their inference procedure starts from splitting T observations into K non-overlapping sub-samples of (approximately) equal length. Let β be the parameter of interest and $\hat{\beta}_j$ be the estimator from the sub-sample $j \in \{1, \ldots, K\}$. Then, the usual t-statistic $t_\beta = \sqrt{K}\left(\overline{\hat{\beta}} - \beta_0\right)/s_{\hat{\beta}}$ is constructed to test for the null hypothesis $H_0 : \beta = \beta_0$, where $\overline{\hat{\beta}} = K^{-1}\sum_{j=1}^K \hat{\beta}_j$ and $s_{\hat{\beta}}^2 = (K-1)^{-1}\sum_{j=1}^K \left(\hat{\beta}_j - \overline{\hat{\beta}}\right)^2$. A critical value of $t(K-1)$ can be used for the t-test as it attains the upper bound for rejection probability under the null; to put it another way, this test becomes conservative under different variances across K sub-samples and has exact size under equal variances (as in Student's original approach). Superior finite-sample properties of this procedure in comparison with fixed-b approaches are confirmed in Monte Carlo simulations by Ibragimov and Müller (2010) and Müller (2014). It is worth remarking that Ibragimov and Müller's (2010) test can be interpreted as a special case of HAR inference. At a first glance, the relation of this test with HAR inference is not obvious because of no explicit use of a LRV estimate. However, Lazarus et al. (2021, Proposition S1) demonstrate that $\hat{\mathbf{\Omega}}_{SS} = (T/K)s_{\hat{\beta}}^2$ can be expressed as an equal-weighted OS-HAR estimator (20) using K basis functions of some orthonormal sequence.

Second, Dou (2020) studies optimal HAR inference in finite samples for a Gaussian location model. The test statistic is again the t-statistic, where the standard error is computed by the OS-HAR estimator (20) using K type II discrete cosine transforms. The critical value for a given level of significance can be obtained as an adjustment factor times the critical value from $t(K)$, where the adjustment factor is greater than 1 and thus the usual Student-t critical value is necessarily inflated for implementation. An advantage of these two tests over fixed-b tests is no need to run simulations for critical values, because Ibragimov and Müller's (2010) test is Student-t based and the adjustment factor for Dou's (2020) test can be easily tabulated.

5.4. LRV-Based Inference for Nonstationary Data

Apart from HAR inference discussed above, inference using LRV estimates for (possibly) nonstationary data has evolved uniquely. This section specializes in unit-root and stationarity tests and structural break tests using LRV estimates.

Ng and Perron (2001) and Vogelsang and Wagner (2013) are two improvements of the Phillips-Perron unit-root test (Phillips and Perron, 1988) in different directions. For an AR(1) model $y_t = \alpha y_{t-1} + u_t$ with a serially correlated error u_t, the Phillips-Perron test statistic requires a consistent estimate of the LRV of u_t. Perron and Ng (1996) find that the test statistic using an AR spectral density estimator (i.e., a univariate version of the VARHAC estimator (12)) has less size distortions than

the one using a kernel HAR estimator. To improve size and power properties of the Phillips-Perron test for the regression

$$y_t = \mathbf{z}_t^\top \beta + \alpha y_{t-1} + u_t \tag{21}$$

with deterministic trends $\mathbf{z}_t = (1, t, \ldots, t^m)^\top$ for $m \in \mathbb{N}$ and a serially correlated error u_t, Ng and Perron (2001) propose to combine the detrending procedure by Elliott et al. (1996) and a modified information criterion for the lag order selection of (12) that replaces the existing criteria such as AIC (10) and BIC (11).

In contrast, Vogelsang and Wagner (2013) extend the fixed-b asymptotics to the Phillips-Perron test for (21) using kernel HAR estimators. It is demonstrated that one- and two-step detrending procedures for (21) yield different non-pivotal fixed-b limits. Based on this, Vogelsang and Wagner (2013) propose modified Phillips-Perron tests that allow for asymptotically pivotal fixed-b inference.

Next, to reduce finite-sample bias in the KPSS test statistic for stationarity (Kwiatkowski et al., 1992) with prewhitened HAR estimators employed, Sul et al. (2005) propose to combine two practices. First, Sul et al. (2005) introduce recursive demeaning procedures to correct finite-sample bias in prewhitening coefficient estimates. Second, Andrews and Monahan (1992) recommend the so-called '0.97' rule that replaces roots of the fitted characteristic equation by 0.97 whenever they exceed this number. In place of 0.97, Sul et al. (2005) advocate the '$1 - 1/\sqrt{T}$' boundary condition rule. These practices jointly improve size and power properties of the KPSS test while maintaining its consistency.

Finally, the test statistic for the null of stability in the mean-shift model

$$y_t = \mu + \delta \mathbf{1}(t \geq t^*) + u_t \tag{22}$$

with an unknown break date t^* and a serially correlated error u_t again requires a consistent estimate of the LRV of u_t. It has long been known that empirical rejection rates of the test statistic using a kernel HAR estimate computed under the null do not increase with the magnitude of the break $|\delta|$. Employing a HAR estimate computed under the alternative of a break may be a remedy for the problem of non-monotonic power, whereas this comes at the cost of severe size distortions. Taking these aspects into account, Kejriwal (2009) proposes to estimate the instantaneous variance $E\left(u_t^2\right)$ and bandwidth S_T using residuals under the alternative while estimating autocovariances $E\left(u_t u_{t-j}\right)$ using those under the null. Kejriwal (2009) reports that the test statistic using the hybrid HAR estimate can bypass the non-monotone power problem while maintaining adequate size properties.

Furthermore, Wenger and Leschinski (2021) extend the fixed-b asymptotics to the CUSUM test for a break in the model (22) with a long-memory error u_t. Comparing two alternative kernel HAR estimates using residuals under the null and alternative, Wenger and Leschinski (2021) find that those under the alternative result in good size and power of their test.

5.5. Estimation-Optimal Bandwidth Selection

The problem of bandwidth selection in kernel HAR estimation is not limited to statistical inference. Wilhelm (2015) proposes yet another objective-oriented bandwidth choice method. Assume that $\ell > p$, i.e., the parameter of interest θ_0 is over-identified. Based on a higher-order approximation to the MSE of the two-step GMM estimator, the method selects the optimal bandwidth $S_T^\dagger$ for the parameter estimator of interest, not the one for the kernel HAR estimator. Specifically, Wilhelm (2015, Proposition 1) establishes that the efficient GMM estimator (6) admits the stochastic expansion

$$\sqrt{T}\left(\hat{\theta} - \theta_0\right) = \mathbf{Z}_T + \mathbf{B}_T S_T^{-q} + \mathbf{V}_T \sqrt{\frac{S_T}{T}} + o_p\left(S_T^{-q} + \sqrt{\frac{S_T}{T}}\right),$$

where q is the characteristic exponent of the kernel,

$$\mathbf{Z}_T = -\left(\mathbf{D}^\top \boldsymbol{\Omega}^{-1} \mathbf{D}\right)^{-1} \mathbf{D}^\top \boldsymbol{\Omega}^{-1} \sqrt{T} \mathbf{G}(\theta_0),$$

$$\mathbf{B}_T = \left(\mathbf{D}^\top \boldsymbol{\Omega}^{-1} \mathbf{D}\right)^{-1} \mathbf{D}^\top \boldsymbol{\Omega}^{-1} S_T^q \left\{E\left(\hat{\boldsymbol{\Omega}}\right) - \boldsymbol{\Omega}\right\} \left\{\boldsymbol{\Omega}^{-1} - \boldsymbol{\Omega}^{-1} \mathbf{D}\left(\mathbf{D}^\top \boldsymbol{\Omega}^{-1} \mathbf{D}\right)^{-1} \mathbf{D}^\top \boldsymbol{\Omega}^{-1}\right\} \sqrt{T} \mathbf{G}(\theta_0), \text{ and}$$

$$\mathbf{V}_T = \left(\mathbf{D}^\top \boldsymbol{\Omega}^{-1} \mathbf{D}\right)^{-1} \mathbf{D}^\top \boldsymbol{\Omega}^{-1} \sqrt{\frac{T}{S_T}} \left\{\hat{\boldsymbol{\Omega}} - E\left(\hat{\boldsymbol{\Omega}}\right)\right\} \left\{\boldsymbol{\Omega}^{-1} - \boldsymbol{\Omega}^{-1} \mathbf{D}\left(\mathbf{D}^\top \boldsymbol{\Omega}^{-1} \mathbf{D}\right)^{-1} \mathbf{D}^\top \boldsymbol{\Omega}^{-1}\right\} \sqrt{T} \mathbf{G}(\theta_0).$$

Three $O_p(1)$ quantities $\mathbf{Z}_T$, $\mathbf{B}_T$ and $\mathbf{V}_T$ correspond to an asymptotically normal term, and terms involving the bias and variance of the HAR estimator $\hat{\boldsymbol{\Omega}}$, respectively. It follows from $\mathbf{B}_T S_T^{-q} + \mathbf{V}_T \sqrt{S_T/T} = o_p(1)$ that estimation errors of $\hat{\boldsymbol{\Omega}}$ do not affect the first-order asymptotic result of the efficient GMM estimator $\hat{\theta}$, as is well known.

Following Andrews (1991), Wilhelm (2015) defines the (scaled) MSE of $\hat{\theta}$ as $E\{T(\hat{\theta} - \theta_0)^\top \mathbf{W}(\hat{\theta} - \theta_0)\}$ for some $p \times p$ symmetric PSD weighting matrix $\mathbf{W}$. Then, it can be found that the optimal bandwidth that balances the squared 'bias' and 'variance' terms of order S_T^{-2q} and S_T/T in the MSE is $S_T^\dagger = C^\dagger(q) T^{1/(2q+1)}$ for some constant $C^\dagger(q)$. Therefore, the optimal bandwidth in efficient GMM estimation shares the same expansion rate with the optimal bandwidth in HAR estimation. Nevertheless, the coefficient $C^\dagger(q)$ substantially differs, for instance, from what Andrews (1991) derives; see Section 2.1 of Wilhelm (2015) for a comparison of two coefficients.

Because $C^\dagger(q)$ is highly complicated, Wilhelm (2015) simplifies its estimation procedure by making a few additional assumptions on the dependent structure of $\mathbf{g}_t$. After assuming that $\mathbf{g}_t$ obeys a linear Gaussian process, as in Andrews (1991),

Wilhelm (2015) puts $\mathbf{W} = \mathbf{I}_p$ and fits an AR(1) model to each element of $\mathbf{g}_t$ to obtain proxies of several unknown quantities in $C^\dagger(q)$. Wilhelm (2015) also reports nice finite-sample properties of the efficient GMM estimator $\hat{\theta}$ using the estimation-optimal bandwidth. However, the simplified implementation method could be a disadvantage if $\mathbf{g}_t$ deviates from Gaussianity or the fitted AR(1) model turns out to be misspecified, as is the case with Andrews' (1991) AMSE-optimal bandwidth.

5.6. LRV Estimation beyond Covariance Stationarity

Consistency of LRV estimators is typically built on covariance stationarity with finite fourth-order moments of $\mathbf{g}_t$ in the early literature. Attempts to relax these conditions have been made since then. Examples include consistency proofs by Hansen (1992b) for non-covariance stationary, heterogenous processes and by De Jong and Davidson (2000) and Davidson and De Jong (2002) for near-epoch dependent functions on mixing processes. LRV estimators in the recent literature assume yet other classes of dependent structure. Two examples are presented below.

5.6.1. Locally Stationary Processes

Kawka (2020) studies spectral density estimation when $\mathbf{g}_t$ obeys a locally stationary process. The process can describe slowly-changing dependent structure over time and is characterized by time-varying second-order moments and behavior like a stationary process in a small neighborhood of each time point $t \in \{1, \ldots, T\}$; see Dahlhaus (2012) for a comprehensive review of this class of processes. To examine statistical properties of the HAR estimator smoothed by a general class of kernels $\hat{\mathbf{\Omega}}_{\mathcal{J}}$ given in (13), define $u = t/T$ so that the time domain $t \in \{1, \ldots, T\}$ is rescaled to the unit interval $u \in [0, 1]$. Also let $f_{\mathbf{gg}}(u, \omega)$ be the $\ell \times \ell$ time-varying spectral density matrix of $\mathbf{g}_t$ evaluated at rescaled time u and frequency $\omega \in (-\pi, \pi)$. Kawka (2020) establishes that under some regularity conditions, $\hat{\mathbf{\Omega}}_{\mathcal{J}} \xrightarrow{p} 2\pi \int_0^1 f_{\mathbf{gg}}(u, 0)du$ holds, i.e., the probability limit of $\hat{\mathbf{\Omega}}_{\mathcal{J}}$ is 2π times the time-averaged spectral density matrix evaluated at frequency zero.

5.6.2. Long-Memory Processes

Robinson (2005) proposes a smoothed nonparametric LRV estimator that is consistent against long-memory or antipersistence. Suppose that the spectral density matrix of $\mathbf{g}_t$ satisfies $f_{\mathbf{gg}}(\omega) \sim h(\omega)\mathbf{S}\overline{h(\omega)}$ as $\omega \to 0+$, where $\mathbf{S}$ is an $\ell \times \ell$ finite PD matrix, $h(\omega) = \text{diag}\left\{e^{id_1\pi/2}\omega^{-d_1}, \ldots, e^{id_\ell\pi/2}\omega^{-d_\ell}\right\}$, $d_1, \ldots, d_\ell \in (-1/2, 1/2)$ are memory parameters, and $\sim$ signifies that the ratio of real parts, and of imaginary parts, of corresponding elements of matrices on the left- and right-hand sides converges to unity. The ath element of $\mathbf{g}_t$ is a short-memory, long-memory and antipersistence processes if $d_a = 0$, $d_a \in (0, 1/2)$ and $d_a \in (-1/2, 0)$, respectively. Then,

$$Var\left[\text{diag}\left\{T^{-d_1}, \ldots, T^{-d_\ell}\right\}\frac{1}{\sqrt{T}}\sum_{t=1}^{T}\mathbf{g}_t\right] \to \mathbf{V} = 2\pi\mathbf{S} \circ \mathbf{Q},$$

where the (a, b) element of the $\ell \times \ell$ matrix $\mathbf{Q} = \mathbf{Q}(d_1, \ldots, d_\ell)$ is

$$\frac{\sin(\pi d_a) + \sin(\pi d_b)}{\Gamma(d_a + d_b + 2)\sin\{\pi(d_a + d_b)\}},$$

$\Gamma(\cdot)$ is the gamma function, and $\circ$ signifies the Hadamard product. Robinson's (2005) memory and autocorrelation consistent (MAC) estimator is defined as $\hat{\mathbf{V}} = 2\pi\hat{\mathbf{S}} \circ \hat{\mathbf{Q}}$, where $\hat{\mathbf{Q}} = \mathbf{Q}\left(\hat{d}_1, \ldots, \hat{d}_\ell\right)$, $\hat{d}_1, \ldots, \hat{d}_\ell$ are consistent estimates of memory parameters,

$$\hat{\mathbf{S}} = \frac{1}{2m_T - 1}\sum_{j=-(m_T-1)}^{m_T-1}\hat{h}\left(\lambda_j\right)^{-1}I_{\mathbf{gg}}\left(\lambda_j\right)\overline{\hat{h}\left(\lambda_j\right)}^{-1}$$

for the bandwidth m_T is a natural estimator of $\mathbf{S}$, $I_{\mathbf{gg}}\left(\lambda_j\right)$ is the periodogram evaluated at frequency $\lambda_j = 2\pi j/T$, and $\hat{h}\left(\lambda_j\right) = \text{diag}\left\{e^{i\hat{d}_1\pi/2}\lambda_j^{-\hat{d}_1}, \ldots, e^{i\hat{d}_\ell\pi/2}\lambda_j^{-\hat{d}_\ell}\right\}$. The MAC estimator $\hat{\mathbf{V}}$ is PSD by construction. Moreover, it can be demonstrated that if $m_T \to \infty$ and $m_T = o\left(T/\log^2 T\right)$, then $\hat{\mathbf{V}} \xrightarrow{p} \mathbf{V}$.

Abadir et al. (2009) compare the asymptotic expansion of the HAR estimator using the Bartlett kernel with that of the MAC estimator for a scalar process having the memory parameter $d \in (-1/2, 1/2)$. The AMSE-optimal bandwidth $S_T^\dagger$ for the HAR estimator is shown to depend on the memory d so that

$$S_T^\dagger = \begin{cases} O\left(T^{1/(3+4d)}\right) & \text{for } d \in (-1/2, 1/4) \\ O\left(T^{1/2-d}\right) & \text{for } d \in [1/4, 1/2) \end{cases}.$$

Observe that the optimal bandwidth reduces to the familiar one $S_T^\dagger = O\left(T^{1/3}\right)$ for the short-memory case ($d = 0$). In contrast, the AMSE-optimal bandwidth $m_T^\dagger$ for the MAC estimator is $m_T^\dagger = O\left(T^{4/5}\right)$ regardless of d. Abadir et al. (2009) conclude that the MAC estimator is more robust to the bandwidth choice than the HAR estimator as the expansion rate of the optimal bandwidth for the latter is influenced by d.

5.7. Implementation in Statistical Packages

This section concludes by noting the commands for implementing HAR estimation and inference in standard statistical packages. In R, the package `cointReg` contains two functions `getLongRunVar` and `getBandwidth` for conventional kernel HAR estimation. The former computes kernel HAR estimates, and the latter yields automatic bandwidths by Andrews (1991) and Newey and West (1994). In Stata, the command `lrcov` by Wang and Wu (2012) corresponds to these two functions.

The commands for fixed-b inference are also available in Stata. Ye and Sun (2018) develop several commands based on test-optimal bandwidths and fixed-smoothing critical values. More specifically, F and t tests for linear regression models (for efficient linear GMM estimation) can be implemented with their commands `har` (`gmmhar`) and `hart` (`gmmhart`).

6. Monte Carlo Simulations

While test-optimal bandwidth selection methods have been explored for a few decades, estimation-optimal ones are still under development. Focusing on the latter, we conduct two independent simulation studies. One is based on efficient GMM estimation and the other on cointegrating regressions. In these applications, choices of LRV estimators do not matter in the first-order asymptotic sense but may affect finite-sample performances of estimates of model parameters. Therefore, each of two Monte Carlo studies is concerned with how a variety of LRV estimators and/or bandwidth formulae can influence the quality of parameter estimates in finite samples.

6.1. Design #1: GMM

The focus of the first Monte Carlo study is on efficient GMM estimation. As already seen in the previous section, Wilhelm (2015) proposes an estimation-optimal bandwidth choice method for HAR estimators using a general class of kernels in the GMM framework. However, its implementation is based on several stringent assumptions. In particular, practitioners may wonder whether $\mathbf{g}_t$ is not a linear Gaussian process or whether the fitted AR(1) model turns out to be misspecified.

In light of these concerns, the Monte Carlo design incorporates the cases in which $\mathbf{g}_t$ deviates from Gaussianity or AR(1), whereas it basically follows Wilhelm (2015). Consider a slope-only linear regression model $y_t = \theta x_t + v_t$ with the true value $\theta = 1$. The regression error v_t falls into one of the two cases

$$v_t = \begin{cases} u_{1t} & \text{[HOM]} \\ \frac{1}{4}|x_t|u_{1t} & \text{[HET]} \end{cases},$$

where "HOM" and "HET" stand for "homoskedastic" and "heteroskedastic" errors, respectively. The regressor x_t has the reduced form $x_t = \pi^\top \mathbf{z}_t + u_{2t}$ with a five-dimensional vector of unity π. The bivariate process $\mathbf{u}_t = (u_{1t}, u_{2t})^\top$ obeys either a VAR(1) model

$$\mathbf{u}_t = 0.8\mathbf{I}_2\mathbf{u}_{t-1} + \epsilon_t \text{ [VAR]}$$

or a VMA(1) model

$$\mathbf{u}_t = \epsilon_t + 0.8\mathbf{I}_2\epsilon_{t-1}, \text{ [VMA]}$$

where the innovation $\epsilon_t = (\epsilon_{1t}, \epsilon_{2t})^\top$ is generated by

$$\epsilon_t \overset{iid}{\sim} N_2\left(\begin{bmatrix}0\\0\end{bmatrix}, \begin{bmatrix}1 & 0.8\\0.8 & 1\end{bmatrix}\right).$$

Due to the correlation between ϵ_{1t} and ϵ_{2t}, x_t is correlated with v_t and thus endogenous. Finally, the vector of instruments $\mathbf{z}_t = (z_{1t}, \ldots, z_{5t})^\top$ is also generated by a VAR(1) model $\mathbf{z}_t = \mathbf{\Phi}_z\mathbf{z}_{t-1} + \mathbf{w}_t$, where $\mathbf{\Phi}_z = \text{diag}\{0, 0.8, \ldots, 0.8\}$, $\mathbf{w}_t = (w_{1t}, \ldots, w_{5t})^\top$, and $w_{1t}, \ldots, w_{5t}$ are iid copies of $N(0,1)$, $\chi^2(2) - 2$ or $t(5)$ and independent of ϵ_s for all t, s. For each combination of v_t, $\mathbf{u}_t$ and $\mathbf{w}_t$, 5000 data sets of sample sizes $T = 128$ are simulated.

Observe that $\mathbf{g}_t = \mathbf{g}_t(\theta) = \mathbf{z}_t(y_t - \theta x_t)$ satisfies the orthogonality condition $E(\mathbf{g}_t) = \mathbf{0}_{5\times 1}$. Taking the two-stage least squares (2SLS) estimator $\hat{\theta}_{2SLS}$ as the initial estimator, we have the efficient GMM estimator of θ in a closed form

$$\hat{\theta}_{GMM} = \left\{\left(\sum_{t=1}^T x_t\mathbf{z}_t\right)\hat{\mathbf{\Omega}}^{-1}\left(\sum_{t=1}^T \mathbf{z}_t^\top x_t\right)\right\}^{-1}\left(\sum_{t=1}^T x_t\mathbf{z}_t\right)\hat{\mathbf{\Omega}}^{-1}\left(\sum_{t=1}^T \mathbf{z}_t^\top y_t\right)$$

for some LRV estimator $\hat{\mathbf{\Omega}}$ computed from $\hat{\mathbf{g}}_t = \mathbf{g}_t\left(\hat{\theta}_{2SLS}\right) = \mathbf{z}_t\left(y_t - \hat{\theta}_{2SLS}x_t\right)$.

As $\hat{\mathbf{\Omega}}$, in addition to QS-A, BT-NW and PZ-H used in Section 4.4, we consider three versions of Smith's (2005) smoothed-moments estimator using the truncated kernel [S-BT-A, S-BT-NW, S-BT-H] and the QS, Bartlett and Parzen HAR estimators

using Wilhelm's (2015) estimation-optimal bandwidths [QS-W, BT-W, PZ-W]. Here are some details on Smith's (2005) estimator. As in Example 2.1 of Smith (2011), the smoothed moment function using the truncated kernel can be constructed by

$$\bar{\mathbf{g}}_t = \bar{\mathbf{g}}_t(\theta) = \frac{2}{2m_T+1} \sum_{j=(t-T)\vee(-m_T)}^{(t-1)\wedge m_T} \mathbf{g}_{t-j}(\theta),\ t = 1, \ldots, T.$$

The optimal bandwidth $m_T^\dagger$ is implied by the induced kernel of the truncated kernel, which is the Bartlett kernel. Because no further information on implementing the smoothed-moments estimator is available, we estimate $m_T^\dagger$ indirectly with the aid of automatic bandwidths for the Bartlett HAR estimator by Andrews (1991), Newey and West (1994) and Hirukawa (2010), denoted by BT-A, BT-NW and BT-H, respectively. More specifically, given the relation $S_T^\dagger = \left(2m_T^\dagger + 1\right)/2$ and an automatic bandwidth $\hat{S}_T^\dagger$, $m_T^\dagger$ can be estimated as $\hat{m}_T^\dagger = \left\lfloor \left(2\hat{S}_T^\dagger - 1\right)/2 \right\rfloor$.

Performance of each estimator is measured by the root MSE (RMSE), bias (BIAS) and standard deviation (SD). Monte Carlo averages of bandwidths (BW) are also presented for convenience. The results of $\hat{\theta}_{2SLS}$ are provided as a benchmark.

Table 3 reports simulation results. Except the three cases of VMA-HOM, RMSEs of GMM estimates are smaller than those of the corresponding 2SLS estimates, as the theory predicts. Although there is no dominant estimator, PZ-W, QS-W and BT-W tend to attain the smallest RMSE for VAR-HOM, VAR-HET and VMA-HET, respectively. Therefore, Wilhelm's (2015) bandwidth appears to have some degree of robustness against the deviations from Gaussianity and AR(1).

A few interesting phenomena can be also found. First, Wilhelm's (2015) bandwidths are in general smaller than the corresponding automatic bandwidths. This finding agrees with simulation results in Wilhelm (2015), and it is a sharp contrast to test-optimal bandwidths, which tend to be longer than the automatic bandwidths. Second, finite-sample properties of Smith's (2005) smoothed-moments estimators do not surpass those of the first-order equivalent Bartlett HAR estimator. Third but not least importantly, a careful inspection reveals that regardless of the data generation processes, average bandwidths of S-BT-H and PZ-H are close to those of BT-W and PZ-W, respectively. As a consequence, RMSEs of the former are comparable to those of the latter. Invoke that both S-BT-H and PZ-H are computed by the SEPI algorithm by Hirukawa (2010). This may be a hint for the direction of improving Wilhelm's (2015) estimation-optimal approach.

6.2. Design #2: Cointegrating Regressions

The second Monte Carlo study deals with a linear cointegrating regression model with an endogenous I(1) regressor. Because of the endogeneity and the serially correlated regression error, the OLS estimator of the cointegrating vector is T-consistent but inefficient due to the so-called second-order bias. To gain full efficiency, several authors have proposed bias-correction estimation methods of the cointegrating vector. This study concentrates on two bias-correction procedures using LRV estimates, namely, fully modified least squares (FMLS) by Phillips and Hansen (1990) and canonical cointegrating regression (CCR) estimation by Park (1992). Finite-sample properties of FMLS and CCR are compared with those of integrated modified OLS (IM-OLS) estimation, a LRV estimation-free bias-corrected procedure proposed by Vogelsang and Wagner (2014).

The Monte Carlo design is similar to Hirukawa (2011). The bivariate process $\{(y_t, x_t)\}_{t=1}^T \in \mathbb{R}^2$ admits the triangular representation

$$\begin{aligned} y_t &= \mathbf{x}_t^\top \theta + u_{1t}, \\ \Delta x_t &= u_{2t}, \end{aligned} \tag{23}$$

where $\mathbf{x}_t = (1, x_t)^\top$, $\theta = (\theta_0, \theta_1)^\top$ with true values $\theta_0 = \theta_1 = 1$, and θ_1 is the parameter of interest. The error term $\mathbf{u}_t = (u_{1t}, u_{2t})^\top$ obeys either a VAR(1) model

$$\mathbf{u}_t = \begin{bmatrix} \rho & 0 \\ 0 & 0 \end{bmatrix} \mathbf{u}_{t-1} + \epsilon_t \ \text{[VAR]}$$

or a VMA(1) model

$$\mathbf{u}_t = \epsilon_t + \begin{bmatrix} 0.3 & -0.4 \\ \rho & 0.6 \end{bmatrix} \epsilon_{t-1}, \ \text{[VMA]}$$

where $\rho \in \{0.4, 0.8\}$. The innovation $\epsilon_t = (\epsilon_{1t}, \epsilon_{2t})^\top$ is generated by

$$\epsilon_t \overset{iid}{\sim} N_2\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 1 & \sigma_{21} \\ \sigma_{21} & 1 \end{bmatrix} \right),$$

where $\sigma_{21} \in \{0.4, 0.8\}$ for VAR and $\sigma_{21} \in \{-0.4, -0.8\}$ for VMA. For each combination of (ρ, σ_{21}), 5000 data sets of sample sizes $T = 128$ are simulated.

Table 3
Simulation Results on GMM

				HOM				*HET*			
Model	Distribution of w	Estimator		RMSE	BIAS	SD	BW	RMSE	BIAS	SD	BW
VAR	$N(0,1)$	2SLS		0.0871	0.0147	0.0859	-	0.1728	0.1257	0.1185	-
		GMM:	QS-A	0.0833	0.0145	0.0821	9.69	0.1580	0.1108	0.1126	7.18
			BT-NW	0.0839	0.0150	0.0825	6.27	0.1572	0.1093	0.1130	5.81
			PZ-H	0.0837	0.0148	0.0824	11.43	0.1574	0.1098	0.1128	9.74
			S-BT-A	0.0842	0.0143	0.0829	10.40	0.1609	0.1139	0.1136	7.92
			S-BT-NW	0.0836	0.0145	0.0824	6.27	0.1587	0.1115	0.1130	5.81
			S-BT-H	0.0834	0.0146	0.0821	4.89	0.1576	0.1099	0.1129	4.06
			QS-W	0.0832	0.0148	0.0819	5.53	0.1569	0.1090	0.1129	4.12
			BT-W	0.0848	0.0151	0.0834	4.12	0.1580	0.1089	0.1144	3.08
			PZ-W	0.0830	0.0148	0.0817	11.14	0.1571	0.1095	0.1126	8.29
	$\chi^2(2)-2$	2SLS		0.0439	0.0037	0.0437	-	0.1391	0.0669	0.1219	-
		GMM:	QS-A	0.0426	0.0040	0.0424	9.78	0.1286	0.0596	0.1140	7.16
			BT-NW	0.0430	0.0039	0.0428	6.24	0.1282	0.0587	0.1139	5.87
			PZ-H	0.0428	0.0040	0.0426	11.36	0.1280	0.0592	0.1135	9.72
			S-BT-A	0.0428	0.0041	0.0426	10.45	0.1303	0.0614	0.1149	7.85
			S-BT-NW	0.0427	0.0039	0.0426	6.24	0.1290	0.0599	0.1142	5.87
			S-BT-H	0.0428	0.0039	0.0426	4.86	0.1284	0.0593	0.1139	4.07
			QS-W	0.0427	0.0039	0.0425	5.60	0.1282	0.0585	0.1140	4.15
			BT-W	0.0434	0.0038	0.0433	4.14	0.1290	0.0584	0.1151	3.09
			PZ-W	0.0426	0.0040	0.0424	11.27	0.1281	0.0589	0.1138	8.36
	$t(5)$	2SLS		0.0671	0.0071	0.0667	-	0.1555	0.0985	0.1203	-
		GMM:	QS-A	0.0662	0.0070	0.0659	9.74	0.1435	0.0870	0.1142	7.24
			BT-NW	0.0662	0.0075	0.0658	6.27	0.1429	0.0858	0.1142	5.84
			PZ-H	0.0659	0.0072	0.0655	11.38	0.1428	0.0861	0.1139	9.66
			S-BT-A	0.0663	0.0070	0.0660	10.43	0.1457	0.0894	0.1151	7.94
			S-BT-NW	0.0663	0.0072	0.0659	6.27	0.1438	0.0874	0.1142	5.84
			S-BT-H	0.0662	0.0072	0.0658	4.87	0.1432	0.0864	0.1142	4.02
			QS-W	0.0658	0.0073	0.0654	5.55	0.1427	0.0856	0.1142	4.16
			BT-W	0.0666	0.0073	0.0662	4.12	0.1437	0.0854	0.1155	3.09
			PZ-W	0.0657	0.0073	0.0653	11.18	0.1427	0.0860	0.1139	8.37
VMA	$N(0,1)$	2SLS		0.0431	0.0042	0.0429	-	0.0965	0.0749	0.0608	-
		GMM:	QS-A	0.0449	0.0042	0.0447	5.10	0.0913	0.0679	0.0609	4.33
			BT-NW	0.0444	0.0042	0.0442	5.10	0.0907	0.0678	0.0603	6.00
			PZ-H	0.0443	0.0042	0.0441	6.43	0.0904	0.0674	0.0602	5.97
			S-BT-A	0.0449	0.0040	0.0447	5.80	0.0918	0.0689	0.0606	4.87
			S-BT-NW	0.0447	0.0044	0.0445	5.10	0.0918	0.0690	0.0605	6.00
			S-BT-H	0.0445	0.0044	0.0442	2.37	0.0902	0.0670	0.0603	1.95
			QS-W	0.0444	0.0044	0.0442	2.98	0.0900	0.0667	0.0604	2.58
			BT-W	0.0441	0.0043	0.0439	2.30	0.0894	0.0662	0.0601	2.00
			PZ-W	0.0444	0.0043	0.0442	5.99	0.0903	0.0671	0.0604	5.19
	$\chi^2(2)-2$	2SLS		0.0218	0.0011	0.0218	-	0.0745	0.0389	0.0635	-
		GMM:	QS-A	0.0229	0.0012	0.0229	5.17	0.0707	0.0365	0.0606	4.43
			BT-NW	0.0226	0.0012	0.0226	5.16	0.0701	0.0361	0.0600	6.15
			PZ-H	0.0226	0.0012	0.0226	6.42	0.0700	0.0358	0.0601	6.15
			S-BT-A	0.0229	0.0013	0.0228	5.86	0.0713	0.0370	0.0609	4.94
			S-BT-NW	0.0227	0.0012	0.0227	5.16	0.0708	0.0367	0.0606	6.15
			S-BT-H	0.0227	0.0012	0.0227	2.36	0.0701	0.0358	0.0603	2.00
			QS-W	0.0227	0.0012	0.0227	3.03	0.0699	0.0357	0.0601	2.67
			BT-W	0.0225	0.0012	0.0225	2.34	0.0695	0.0353	0.0599	2.07
			PZ-W	0.0227	0.0012	0.0226	6.10	0.0700	0.0359	0.0601	5.38
	$t(5)$	2SLS		0.0341	0.0017	0.0341	-	0.0858	0.0581	0.0632	-
		GMM:	QS-A	0.0356	0.0018	0.0356	5.15	0.0815	0.0531	0.0618	4.41
			BT-NW	0.0353	0.0018	0.0352	5.24	0.0808	0.0529	0.0612	6.51
			PZ-H	0.0353	0.0018	0.0353	6.41	0.0808	0.0528	0.0612	6.03
			S-BT-A	0.0357	0.0017	0.0356	5.84	0.0820	0.0539	0.0618	4.92
			S-BT-NW	0.0355	0.0018	0.0355	5.24	0.0818	0.0537	0.0617	6.51
			S-BT-H	0.0353	0.0018	0.0353	2.36	0.0808	0.0523	0.0616	1.99
			QS-W	0.0353	0.0018	0.0353	3.01	0.0808	0.0523	0.0616	2.65
			BT-W	0.0351	0.0018	0.0351	2.33	0.0803	0.0519	0.0612	2.05
			PZ-W	0.0353	0.0018	0.0353	6.07	0.0809	0.0526	0.0615	5.33

Note: "RMSE", "Bias", "SD", and "BW" are the RMSE, bias, standard deviation, and Monte Carlo average of bandwidths, respectively. Averages of automatic bandwidths for the Bartlett estimator $\hat{S}_T^{\dagger}$, not those of $\hat{m}_T^{\dagger}$, are reported as those for smoothed-moments estimators.

In this setup $\hat{\theta}_{OLS}$, the OLS estimator of the cointegrating vector θ in (23), is T-consistent but inefficient due to the second-order bias. FMLS and CCR are developed as efficient estimation methods that can eliminate the bias, and these are first-order asymptotically equivalent. Let $\mathbf{\Omega}$ and $\mathbf{\Lambda}$ be the two- and one-sided LRVs of $\mathbf{u}_t$, respectively, where the definition of $\mathbf{\Lambda}$ is given in (7) and these LRVs can be partitioned into

$$\mathbf{\Omega} = \begin{bmatrix} \mathbf{\Omega}_{11} & \mathbf{\Omega}_{12} \\ \mathbf{\Omega}_{21} & \mathbf{\Omega}_{22} \end{bmatrix} \text{ and } \mathbf{\Lambda} = \begin{bmatrix} \mathbf{\Lambda}_{11} & \mathbf{\Lambda}_{12} \\ \mathbf{\Lambda}_{21} & \mathbf{\Lambda}_{22} \end{bmatrix} = \begin{bmatrix} \mathbf{\Lambda}_1 \\ \mathbf{\Lambda}_2 \end{bmatrix}.$$

Also let $\hat{\mathbf{\Omega}}$ and $\hat{\mathbf{\Lambda}}$ be their consistent estimators with $\mathbf{u}_t$ replaced by $\hat{\mathbf{u}}_t = \left(\hat{u}_{1t}, \Delta x_t\right)^\top$, where $\hat{u}_{1t}$ is the OLS residual from (23). The FMLS estimator of θ is given by

$$\hat{\theta}_{FMLS} = \left(\sum_{t=1}^{T} \mathbf{z}_t \mathbf{z}_t^\top \right)^{-1} \left(\sum_{t=1}^{T} \mathbf{z}_t y_t^+ - T\hat{\mathbf{j}}^+ \right),$$

where $y_t^+ = y_t - \hat{\mathbf{\Omega}}_{12}\hat{\mathbf{\Omega}}_{22}^{-1}\Delta x_t$ and $\hat{\mathbf{j}}^+ = \left(0, \left(\hat{\mathbf{\Lambda}}_{21} - \hat{\mathbf{\Lambda}}_{22}\hat{\mathbf{\Omega}}_{22}^{-1}\hat{\mathbf{\Omega}}_{21}\right)^\top\right)^\top$. CCR employs the transformed data

$$x_t^* = x_t - \left(\hat{\mathbf{\Gamma}}_0^{-1}\hat{\mathbf{\Lambda}}_2\right)^\top \hat{\mathbf{u}}_t \text{ and } y_t^* = y_t - \left(\hat{\mathbf{\Gamma}}_0^{-1}\hat{\mathbf{\Lambda}}_2\hat{\theta}_{OLS} + \left(0, \hat{\mathbf{\Omega}}_{12}\hat{\mathbf{\Omega}}_{22}^{-1}\right)^\top\right)^\top \hat{\mathbf{u}}_t,$$

where $\hat{\mathbf{\Gamma}}_0 = \sum_{t=1}^{T} \hat{\mathbf{u}}_t\hat{\mathbf{u}}_t^\top/T$ is a consistent estimator of $\mathbf{\Gamma}_0 = E(\mathbf{u}_t\mathbf{u}_t^\top)$. Denoting $\mathbf{z}_t^* = \left(1, x_t^*\right)^\top$ yields the CCR estimator of θ as

$$\hat{\theta}_{CCR} = \left(\sum_{t=1}^{T} \mathbf{z}_t^* \mathbf{z}_t^{*\top} \right)^{-1} \left(\sum_{t=1}^{T} \mathbf{z}_t^* y_t^* \right).$$

A problem is that so far no estimation-optimal bandwidth choice method for a HAR estimator has been proposed in the context of cointegrating regressions, unlike the GMM case, to the best of our knowledge. As a compromise, we again employ HAR estimators of $\mathbf{\Omega}$ using currently available automatic bandwidths, namely, QS-A, BT-NW and PZ-H, where $\hat{\mathbf{u}}_t$ may be prewhitened by a VAR(1) filter before HAR estimation. The one-sided LRV estimator based on VAR(1)-prewhitening can be found in Hansen (1992a, p.47). In addition, the VARHAC estimator of $\mathbf{\Omega}$ is employed, where the maximum lag order $M = M_T = \lfloor T^{1/3} \rfloor$ and the lag order in each element of $\hat{\mathbf{u}}_t$ is chosen via BIC (11). The one-sided LRV estimator based on VARHAC is implied by Proposition 2 of Park and Ogaki (1991).

Instead of LRV estimation, the second-order bias correction in IM-OLS is made by a partial sum transformation of the regression (23). The IM-OLS estimator of θ is defined as the OLS estimator of the transformed regression

$$S_t^y = \left(\mathbf{S}_t^{\mathbf{x}}\right)^\top \theta + x_t\gamma + S_t^{u_1},$$

where $S_t^y = \sum_{j=1}^{t} y_t$, $\mathbf{S}_t^{\mathbf{x}}$ and $S_t^{u_1}$ are defined analogously, and x_t is added as an extra regressor to establish asymptotic mixed normality when (u_{1t}, u_{2t}) are correlated.

As before, RMSE, BIAS and SD are chosen as performance measures. Monte Carlo averages of bandwidths and lag orders (BW/LO) are also reported. The results of $\hat{\theta}_{OLS}$ are provided as a benchmark.

Finite-sample performances of OLS, IM-OLS, FMLS, and CCR estimators of θ_1 are presented in Table 4. FMLS and CCR decrease RMSEs from OLS, except those using VARHAC. Performances of FMLS and CCR do not have much difference, and prewhitening often contributes bias reduction. Again results are mixed, and no dominant LRV estimator can be found.

VARHAC does not look an appropriate tool for FMLS or CCR. It yields the highest RMSE for all eight cases. In addition, RMSEs of VARHAC often exceed those of OLS for four cases of VMA. Disappointing performance of VARHAC is attributed to both large BIAS (in magnitude) and SD. This reflects difficulty in estimating AR coefficients precisely, and coefficient estimates with poor quality in turn induce additional bias and/or variability in estimates of cointegrating vectors.

IM-OLS is more effective in bias reduction than FMLS and CCR using non-prewhitened (in all cases) and prewhitened HAR estimates (frequently). However, its attractive bias properties come at the cost of higher SD and thus RMSE. These findings coincide with what are reported in Vogelsang and Wagner (2014).

7. Concluding Remarks

The literature on HAR estimation and inference is still growing. This article has discussed and compared a variety of LRV estimators that have been proposed in the past few decades, and it covers both parametric and nonparametric estimators. Particular attention is paid to whether a LRV estimator necessarily generates PSD estimates. Kernel methods have been dominant among all LRV estimation procedures. A section is devoted to recent developments in HAR estimation and inference including fixed-b asymptotics and test-optimal bandwidth selection. This article also covers estimation-optimal bandwidth choice methods for HAR estimation, which are still under development. In this connection, we run Monte Carlo simulations implied by two estimation problems using LRV estimators, namely, GMM and cointegrating regressions.

Table 4
Simulation Results on Cointegrating Regressions

				$\rho = 0.4$				$\rho = 0.8$			
Model	σ_{21}	Estimator		RMSE	BIAS	SD	BW/LO	RMSE	BIAS	SD	BW/LO
VAR	0.4	OLS		0.0503	0.0264	0.0428	-	0.1298	0.0695	0.1096	-
		IM-OLS		0.0570	0.0028	0.0570	-	0.1661	0.0216	0.1646	-
		FMLS:	QS-A	0.0415	0.0067	0.0410	4.48	0.1257	0.0364	0.1203	13.77
			BT-NW	0.0408	0.0082	0.0400	4.49	0.1216	0.0406	0.1146	7.57
			PZ-H	0.0406	0.0078	0.0399	4.96	0.1224	0.0382	0.1163	13.69
			QS-A-PW	0.0408	0.0035	0.0407	0.96	0.1258	0.0184	0.1245	1.20
			BT-NW-PW	0.0417	0.0055	0.0413	5.18	0.1270	0.0213	0.1252	4.30
			PZ-H-PW	0.0408	0.0035	0.0407	0.14	0.1260	0.0187	0.1246	0.43
			VARHAC	0.0460	0.0057	0.0456	1.01,1.01	0.1308	0.0658	0.1131	1.01,1.01
		CCR:	QS-A	0.0414	0.0065	0.0409	4.48	0.1251	0.0377	0.1193	13.77
			BT-NW	0.0407	0.0080	0.0399	4.49	0.1213	0.0409	0.1142	7.57
			PZ-H	0.0405	0.0076	0.0398	4.96	0.1220	0.0389	0.1156	13.69
			QS-A-PW	0.0407	0.0036	0.0405	0.96	0.1234	0.0232	0.1212	1.20
			BT-NW-PW	0.0415	0.0055	0.0412	5.18	0.1249	0.0255	0.1222	4.30
			PZ-H-PW	0.0407	0.0036	0.0405	0.14	0.1236	0.0234	0.1213	0.43
			VARHAC	0.0423	0.0061	0.0419	1.01,1.01	0.1281	0.0669	0.1092	1.01,1.01
	0.8	OLS		0.0727	0.0532	0.0495	-	0.1823	0.1393	0.1176	-
		IM-OLS		0.0477	0.0052	0.0474	-	0.1457	0.0418	0.1396	-
		FMLS:	QS-A	0.0342	0.0133	0.0315	4.67	0.1321	0.0744	0.1092	15.04
			BT-NW	0.0354	0.0163	0.0314	4.41	0.1273	0.0798	0.0993	7.68
			PZ-H	0.0341	0.0154	0.0304	5.05	0.1245	0.0746	0.0997	14.08
			QS-A-PW	0.0290	0.0056	0.0285	0.94	0.0962	0.0253	0.0928	1.10
			BT-NW-PW	0.0366	0.0112	0.0349	5.19	0.1069	0.0349	0.1010	4.40
			PZ-H-PW	0.0291	0.0056	0.0286	0.10	0.0963	0.0247	0.0931	0.26
			VARHAC	0.0982	−0.0085	0.0978	1.02,1.01	0.2180	0.1070	0.1899	1.02,1.01
		CCR:	QS-A	0.0353	0.0151	0.0319	4.67	0.1360	0.0821	0.1084	15.04
			BT-NW	0.0360	0.0173	0.0316	4.41	0.1312	0.0844	0.1005	7.68
			PZ-H	0.0352	0.0167	0.0310	5.05	0.1298	0.0816	0.1010	14.08
			QS-A-PW	0.0321	0.0107	0.0303	0.94	0.1182	0.0638	0.0996	1.10
			BT-NW-PW	0.0375	0.0140	0.0348	5.19	0.1228	0.0658	0.1036	4.40
			PZ-H-PW	0.0321	0.0107	0.0303	0.10	0.1176	0.0631	0.0993	0.26
			VARHAC	0.0396	0.0084	0.0387	1.02,1.01	0.1748	0.1269	0.1203	1.02,1.01
VMA	−0.4	OLS		0.0374	−0.0225	0.0299	-	0.0428	−0.0269	0.0333	-
		IM-OLS		0.0357	−0.0021	0.0356	-	0.0381	−0.0026	0.0380	-
		FMLS:	QS-A	0.0260	−0.0044	0.0256	4.27	0.0289	−0.0056	0.0284	3.81
			BT-NW	0.0269	−0.0072	0.0259	4.40	0.0310	−0.0103	0.0292	4.38
			PZ-H	0.0261	−0.0058	0.0254	5.83	0.0296	−0.0082	0.0284	5.84
			QS-A-PW	0.0275	0.0098	0.0257	1.86	0.0297	0.0093	0.0282	2.00
			BT-NW-PW	0.0268	0.0032	0.0266	4.75	0.0288	0.0035	0.0286	4.33
			PZ-H-PW	0.0275	0.0094	0.0259	2.18	0.0291	0.0071	0.0282	4.37
			VARHAC	0.0519	−0.0115	0.0506	1.55,1.49	0.0495	−0.0168	0.0465	1.39,2.47
		CCR:	QS-A	0.0261	−0.0048	0.0256	4.27	0.0290	−0.0060	0.0284	3.81
			BT-NW	0.0269	−0.0073	0.0259	4.40	0.0311	−0.0105	0.0293	4.38
			PZ-H	0.0261	−0.0060	0.0254	5.83	0.0297	−0.0084	0.0285	5.84
			QS-A-PW	0.0262	0.0078	0.0250	1.86	0.0288	0.0079	0.0277	2.00
			BT-NW-PW	0.0262	0.0022	0.0262	4.75	0.0285	0.0026	0.0283	4.33
			PZ-H-PW	0.0263	0.0076	0.0251	2.18	0.0284	0.0058	0.0278	4.37
			VARHAC	0.0438	−0.0137	0.0416	1.55,1.49	0.0460	−0.0182	0.0423	1.39,2.47
	−0.8	OLS		0.0510	−0.0357	0.0365	-	0.0776	−0.0565	0.0532	-
		IM-OLS		0.0364	−0.0036	0.0362	-	0.0502	−0.0049	0.0500	-
		FMLS:	QS-A	0.0253	−0.0073	0.0242	4.91	0.0399	−0.0129	0.0378	5.12
			BT-NW	0.0252	−0.0089	0.0235	4.56	0.0417	−0.0181	0.0375	4.30
			PZ-H	0.0243	−0.0076	0.0230	5.69	0.0410	−0.0169	0.0374	4.36
			QS-A-PW	0.0228	0.0062	0.0219	1.96	0.0349	0.0073	0.0341	1.66
			BT-NW-PW	0.0230	0.0008	0.0229	4.49	0.0351	−0.0014	0.0351	4.69
			PZ-H-PW	0.0223	0.0036	0.0220	4.62	0.0348	−0.0025	0.0347	7.12
			VARHAC	0.0993	−0.0068	0.0990	1.61,1.12	0.1117	−0.0486	0.1006	1.35,1.89
		CCR:	QS-A	0.0256	−0.0080	0.0243	4.91	0.0405	−0.0143	0.0379	5.12
			BT-NW	0.0254	−0.0093	0.0237	4.56	0.0422	−0.0189	0.0378	4.30
			PZ-H	0.0247	−0.0082	0.0233	5.69	0.0417	−0.0179	0.0377	4.36
			QS-A-PW	0.0221	0.0020	0.0220	1.96	0.0336	0.0020	0.0336	1.66
			BT-NW-PW	0.0230	−0.0019	0.0229	4.49	0.0353	−0.0049	0.0349	4.69
			PZ-H-PW	0.0222	0.0002	0.0222	4.62	0.0353	−0.0057	0.0348	7.12
			VARHAC	0.0496	−0.0208	0.0450	1.61,1.12	0.0885	−0.0529	0.0710	1.35,1.89

Note: HAR estimators with "PW" indicate prewhitened ones. "RMSE", "Bias", "SD", and "BW/LO" are the RMSE, bias, standard deviation, and Monte Carlo average of bandwidths or lag orders, respectively. For VARHAC, average lag orders of the first and second elements of $\hat{\mathbf{u}}_t$ are presented.

This article concludes with possible directions of future research suggested by the Monte Carlo simulations. The first simulation study confirms some degree of robustness in Wilhelm's (2015) estimation-optimal bandwidth choice method for GMM. While it is implemented like Andrews' (1991) automatic bandwidth, it in general works well even when the underlying process is not Gaussian or AR(1). On the other hand, Wilhelm's (2015) and Hirukawa's (2010) bandwidths are very close on average, and the latter appears to have the potential to match the former. In addition, Hirukawa's (2010) SEPI bandwidth can be implemented under less stringent conditions than Wilhelm's (2015). Therefore, natural extensions of Wilhelm's (2015) approach would be: (i) relaxing the conditions for its implementation; (ii) investigating applicability of prewhitening (which is not supported theoretically); and (iii) improving its computation in combination with the SEPI algorithm, if applicable.

The second simulation study indicates that while the parametric VARHAC estimator is easy to implement, estimation errors of AR coefficients induce additional bias and/or variability in the estimates of cointegrating vectors. Hence, preference is still given to nonparametric HAR estimators, whereas no estimation-optimal bandwidth choice method has been proposed in the framework of semiparametric bias-corrected estimators of cointegrating regressions. It is worth tailoring the estimation-optimal bandwidth choice method to FMLS and CCR. As in Wilhelm (2015), this task may start from stochastic expansions of the FMLS and CCR estimators. The contribution of prewhitening to bias reduction should be also justified theoretically.

Acknowledgment

The author would like to thank the Editor, an Associate Editor, two anonymous referees, Liudas Giraitis, Eiji Kurozumi, Artem Prokhorov, Ken West, and the participants of ISF2021 for their constructive comments and suggestions, which helped improve this article substantially. The author also gratefully acknowledges financial support from Japan Society for the Promotion of Science (grant number 19K01595).

References

Abadir, K.M., Distaso, W., Giraitis, L., 2009. Two estimators of the long-run variance: Beyond short memory. Journal of Econometrics 150, 56–70.
Ahn, S.C., Gadarowski, C., Perez, M.F., 2012. Robust two-pass cross-sectional regressions: A minimum distance approach. Journal of Financial Econometrics 10, 669–701.
Akaike, H., 1973. Information theory and an extension of the maximum likelihood principle. In: Petrov, B.N., Csaki, F. (Eds.), Proceedings of the Second International Symposium on Information Theory. Akademiai Kiado, Budapest, pp. 267–281.
Anderson, T.W., 1971. The Statistical Analysis of Time Series. John Wiley & Sons, New York.
Andrews, D.W.K., 1991. Heteroskedasticity and autocorrelation consistent covariance matrix estimation. Econometrica 59, 817–858.
Andrews, D.W.K., 1993. Tests for parameter instability and structural change with unknown change point. Econometrica 61, 821–856.
Andrews, D.W.K., Monahan, J.C., 1992. An improved heteroskedasticity and autocorrelation consistent covariance matrix estimator. Econometrica 60, 953–966.
Basu, S., Michailidis, G., 2015. Regularized estimation in sparse high-dimensional time series models. Annals of Statistics 43, 1535–1567.
Berk, K.N., 1974. Consistent autoregressive spectral estimates. Annals of Statistics 2, 489–502.
Blackman, R.B., Tukey, J.W., 1958. The measurement of power spectra from the point of view of communications engineering - Part I. Bell System Technical Journal 37, 185–282.
Brillinger, D.R., 1975. Time Series: Data Analysis and Theory. Holt, Rinehart and Winston, New York.
Creel, M., Farell, M., 1996. SUR estimation of multiple time-series models with heteroscedasticity and serial correlation of unknown form. Economics Letters 53, 239–245.
Cumby, R.E., Huizinga, J., Obstfeld, M., 1983. Two-step two-stage least squares estimation in models with rational expectations. Journal of Econometrics 21, 333–355.
Cushing, M.J., McGarvey, M.G., 1999. Covariance matrix estimation. In: Mátyás, L. (Ed.), Generalized Method of Moments Estimation. Cambridge University Press, New York, pp. 63–95.
Dahlhaus, R., 2012. Locally stationary processes. In: Rao, T.S., Rao, S.S., Rao, C.R. (Eds.), Handbook of Statistics Volume 30: Time Series Analysis: Methods and Applications. Elsevier, Amsterdam, pp. 351–413.
Davidson, J., De Jong, R.M., 2002. Consistency of kernel variance estimators for sums of semiparametric linear processes. Econometrics Journal 5, 160–175.
De Jong, R.M., Davidson, J., 2000. Consistency of kernel estimators of heteroskedastic and autocorrelated covariance matrices. Econometrica 68, 407–423.
Den Haan, W.J., Levin, A.T., 1996. Inferences from parametric and non-parametric covariance matrix estimation procedures. Technical Working Paper 195. National Bureau of Economic Research.
Den Haan, W.J., Levin, A.T., 1997. A practitioner's guide to robust covariance matrix estimation. In: Maddala, G.S., Rao, C.R. (Eds.), Handbook of Statistics Volume 15: Robust Inference. Elsevier, Amsterdam, pp. 299–342.
Den Haan, W.J., Levin, A.T., 2000. Robust covariance matrix estimation with data-dependent VAR prewhitening order. Technical Working Paper 255. National Bureau of Economic Research.
Diebold, F.X., Mariano, R.S., 1995. Comparing predictive accuracy. Journal of Business & Economic Statistics 13, 253–263.
Dou, L., 2020. Optimal HAR Inference. Unpublished Manuscript, Chinese University of Hong Kong.
Eichenbaum, M.S., Hansen, L.P., Singleton, K.J., 1988. A time series analysis of representative agent models of consumption and leisure choice under uncertainty. Quarterly Journal of Economics 103, 51–78.
Elliott, G., Rothenberg, T.J., Stock, J.H., 1996. Efficient tests for an autoregressive unit root. Econometrica 64, 813–836.
Gallant, A.R., 1987. Nonlinear Statistical Models. John Wiley & Sons, New York.
Hall, A.R., 2005. Generalized Method of Moments. Oxford University Press, New York.
Hannan, E.J., 1957. The variance of the mean of a stationary process. Journal of the Royal Statistical Society, Series B 19, 282–285.
Hannan, E.J., 1970. Multiple Time Series. John Wiley & Sons, New York.
Hannan, E.J., Rissanen, J., 1982. Recursive estimation of mixed autoregressive-moving average order. Biometrika 69, 81–94.
Hansen, B.E., 1992a. Tests for parameter instability in regressions with I(1) processes. Journal of Business & Economic Statistics 10, 45–59.
Hansen, B.E., 1992b. Consistent covariance matrix estimation for dependent heterogenous processes. Econometrica 60, 967–972.
Hansen, L.P., 1982. Large sample properties of generalized method of moments estimators. Econometrica 50, 1029–1054.
Hansen, L.P., Hodrick, R.J., 1980. Forward exchange rates as optimal predictors of future spot rates: An econometric analysis. Journal of Political Economy 88, 829–853.
Hansen, L.P., Singleton, K.J., 1982. Generalized instrumental variables estimation of nonlinear rational expectations models. Econometrica 50, 1269–1286.

Hirukawa, M., 2006. A modified nonparametric prewhitened covariance estimator. Journal of Time Series Analysis 27, 441–476.
Hirukawa, M., 2010. A two-stage plug-in bandwidth selection and its implementation for covariance estimation. Econometric Theory 26, 710–743.
Hirukawa, M., 2011. How useful is yet another data-driven bandwidth in long-run variance estimation?: A simulation study on cointegrating regressions. Economics Letters 111, 170–172.
Hodrick, R.J., 1992. Dividend yields and expected stock returns: Alternative procedures for inference and measurement. Review of Financial Studies 5, 357–386.
Ibragimov, R., Müller, U.K., 2010. t-statistic based correlation and heterogeneity robust inference. Journal of Business & Economic Statistics 28, 453–468.
Jansson, M., 2002. Consistent covariance matrix estimation for linear processes. Econometric Theory 18, 1449–1459.
Jansson, M., 2004. The error in rejection probability of simple autocorrelation robust tests. Econometrica 72, 937–946.
Jones, M.C., Linton, O., Nielsen, J.P., 1995. A simple bias reduction method for density estimation. Biometrika 82, 327–338.
Jowett, G.H., 1955. The comparison of means of sets of observations from sections of independent stochastic series. Journal of the Royal Statistical Society, Series B 17, 208–227.
Kawka, R., 2020. Convergence of spectral density estimators in the locally stationary framework. Econometrics and Statistics. forthcoming
Kejriwal, M., 2009. Tests for a mean shift with good size and monotonic power. Economics Letters 102, 78–82.
Kiefer, N.M., Vogelsang, T.J., 2002. Heteroskedasticity-autocorrelation robust standard errors using the Bartlett kernel without truncation. Econometrica 70, 2093–2095.
Kiefer, N.M., Vogelsang, T.J., 2005. A new asymptotic theory for heteroskedasticity-autocorrelation robust tests. Econometric Theory 21, 1130–1164.
Kiefer, N.M., Vogelsang, T.J., Bunzel, H., 2000. Simple robust testing of regression hypotheses. Econometrica 68, 695–714.
Kitamura, Y., 1997. Empirical likelihood methods with weakly dependent processes. Annals of Statistics 25, 2084–2102.
Kitamura, Y., Stutzer, M., 1997. An information-theoretic alternative to generalized method of moments estimation. Econometrica 65, 861–874.
Kwiatkowski, D., Phillips, P.C.B., Schmidt, P., Shin, Y., 1992. Testing the null hypothesis of stationarity against the alternative of a unit root: How sure are we that economic time series have a unit root? Journal of Econometrics 54, 159–178.
Lazarus, E., Lewis, D.J., Stock, J.H., 2021. Supplement to The Size-Power Tradeoff in HAR Inference (Econometrica, Vol.89, No.5, September 2021, 2497-2516). Econometrica Supplementary Material.
Lazarus, E., Lewis, D.J., Stock, J.H., Watson, M.W., 2018. HAR inference: Recommendations for practice. Journal of Business & Economic Statistics 36, 541–559.
Lee, C.C., Phillips, P.C.B., 1994. An ARMA prewhitened long-run variance estimator. Unpublished Manuscript, Yale University.
Mark, N.C., 2001. International Macroeconomics and Finance: Theory and Econometric Methods. Blackwell, Malden, MA.
McCracken, M.W., 2000. Robust out-of-sample inference. Journal of Econometrics 99, 195–223.
Müller, U.K., 2007. A theory of robust long-run variance estimation. Journal of Econometrics 141, 1331–1352.
Müller, U.K., 2014. HAC corrections for strongly autocorrelated time series. Journal of Business & Economic Statistics 32, 311–322.
Newey, W.K., West, K.D., 1987. A simple, positive semi-definite, heteroskedasticity and autocorrelation consistent covariance matrix. Econometrica 55, 703–708.
Newey, W.K., West, K.D., 1994. Automatic lag selection in covariance matrix estimation. Review of Economic Studies 61, 631–653.
Ng, S., Perron, P., 2001. Lag length selection and the construction of unit root tests with good size and power. Econometrica 69, 1519–1554.
Park, J.Y., 1992. Canonical cointegrating regressions. Econometrica 60, 119–143.
Park, J.Y., Ogaki, M., 1991. Inference in cointegrated models using VAR prewhitening to estimate shortrun dynamics. Rochester Center for Economic Research Working Paper No. 281. University of Rochester.
Parzen, E., 1957. On consistent estimates of the spectrum of a stationary time series. Annals of Mathematical Statistics 28, 329–348.
Parzen, E., 1961. Mathematical considerations in the estimation of spectra. Technometrics 3, 167–190.
Parzen, E., 1963. Notes on Fourier analysis and spectrum windows. Technical Report No. 48. Department of Statistics, Stanford University. (Published in Parzen, E. (1967): *Time Series Analysis Papers*. Holden-Day, San Francisco).
Perron, P., Ng, S., 1996. Useful modifications to some unit root tests with dependent errors and their local asymptotic properties. Review of Economic Studies 63, 435–463.
Phillips, P.C.B., 1987. Time series regression with a unit root. Econometrica 55, 277–301.
Phillips, P.C.B., 1991. Spectral regression for cointegrated time series. In: Barnett, W.A., Powell, J., Tauchen, G.E. (Eds.), Nonparametric and Semiparametric Methods in Econometrics and Statistics: Proceedings of the Fifth International Symposium in Economic Theory and Econometrics. Cambridge University Press, Cambridge, UK, pp. 413–435.
Phillips, P.C.B., 2005. HAC estimation by automated regression. Econometric Theory 21, 116–142.
Phillips, P.C.B., Hansen, B.E., 1990. Statistical inference in instrumental variables regression with I(1) processes. Review of Economic Studies 57, 99–125.
Phillips, P.C.B., Ouliaris, S., 1990. Asymptotic properties of residual based tests for cointegration. Econometrica 58, 165–193.
Phillips, P.C.B., Perron, P., 1988. Testing for a unit root in time series regression. Biometrika 75, 335–346.
Phillips, P.C.B., Sun, Y., Jin, S., 2006. Spectral density estimation and robust hypothesis testing using steep origin kernels without truncation. International Economic Review 47, 837–894.
Phillips, P.C.B., Sun, Y., Jin, S., 2007. Long run variance estimation and robust regression testing using sharp origin kernels with no truncation. Journal of Statistical Planning and Inference 137, 985–1023.
Politis, D.N., 2011. Higher-order accurate, positive semidefinite estimation of large-sample covariance and spectral density matrices. Econometric Theory 27, 703–744.
Press, H., Tukey, J.W., 1956. Power spectral methods of analysis and their application to problems in airplane dynamics. In: Flight Test Manual, NATO Advisory Group for Aeronautical Research and Development, IV-C, pp. 1-41. (reprinted as Bell System Monograph No. 2606.).
Priestley, M.B., 1981. Spectral Analysis and Time Series. Academic Press, San Diego.
Robinson, P.M., 1998. Inference-without-smoothing in the presence of nonparametric autocorrelation. Econometrica 66, 1163–1182.
Robinson, P.M., 2005. Robust covariance matrix estimation: HAC estimates with long memory/antipersistence correction. Econometric Theory 21, 171–180.
Rossi, B., Inoue, A., 2012. Out-of-sample forecast tests robust to the choice of window size. Journal of Business & Economic Statistics 30, 432–453.
Schwarz, G., 1978. Estimating the dimension of a model. Annals of Statistics 6, 461–464.
Sheather, S.J., Jones, M.C., 1991. A reliable data-based bandwidth selection method for kernel density estimation. Journal of the Royal Statistical Society, Series B 53, 683–690.
Silverman, B.W., 1986. Density Estimation for Statistics and Data Analysis. Chapman & Hall, London.
Smith, R.J., 1997. Alternative semi-parametric likelihood approaches to generalised method of moments estimation. Econometrics Journal 107, 503–519.
Smith, R.J., 2005. Automatic positive semidefinite HAC covariance matrix and GMM estimation. Econometric Theory 21, 158–170.
Smith, R.J., 2011. GEL criteria for moment condition models. Econometric Theory 27, 1192–1235.
Stuetzle, W., Mittal, Y., 1979. Some comments on the asymptotic behavior of robust smoothers. In: Gasser, T., Rosenblatt, M. (Eds.), Smoothing Techniques for Curve Estimation: Proceedings of a Workshop Held in Heidelberg, April 2-4, 1979. Springer-Verlag, Berlin, pp. 191–195.
Sul, D., Phillips, P.C.B., Choi, C.Y., 2005. Prewhitening bias in HAC estimation. Oxford Bulletin of Economics and Statistics 67, 517–546.
Sun, Y., 2011. Robust trend inference with series variance estimator and testing-optimal smoothing parameter. Journal of Econometrics 164, 345–366.
Sun, Y., 2014a. Let's fix it: Fixed-b asymptotics versus small-b asymptotics in heteroskedasticity and autocorrelation robust inference. Journal of Econometrics 178, 659–677.
Sun, Y., 2014b. Fixed-smoothing asymptotics in a two-step generalized method of moments framework. Econometrica 82, 2327–2370.
Sun, Y., Phillips, P.C.B., Jin, S., 2008. Optimal bandwidth selection in heteroskedasticity-autocorrelation robust testing. Econometrica 76, 175–194.

Sun, Y., Phillips, P.C.B., Jin, S., 2011. Power maximization and size control in heteroskedasticity and autocorrelation robust tests with exponentiated kernels. Econometric Theory 27, 1320–1368.
Sun, Y., Yang, J., 2020. Testing-optimal kernel choice in HAR inference. Journal of Econometrics 219, 123–136.
Vogelsang, T.J., Wagner, M., 2013. A fixed-*b* perspective on the Phillips-Perron unit root tests. Econometric Theory 29, 609–628.
Vogelsang, T.J., Wagner, M., 2014. Integrated modified OLS estimation and fixed-*b* inference for cointegrating regressions. Journal of Econometrics 178, 741–760.
Wang, Q., Wu, N., 2012. Long-run covariance and its applications in cointegrating regression. Stata Journal 12, 515–542.
Wenger, K., Leschinski, C., 2021. Fixed-bandwidth CUSUM tests under long memory. Econometrics and Statistics 20, 46–61.
West, K.D., 1996. Asymptotic inference about predictive ability. Econometrica 64, 1067–1084.
West, K.D., 1997. Another heteroskedasticity- and autocorrelation-consistent covariance matrix estimator. Journal of Econometrics 76, 171–191.
West, K.D., 2008. Heteroskedasticity and autocorrelation corrections. In: Durlauf, S.N., Brume, L.E. (Eds.), The New Palgrave Dictionary of Economics, Second Edition, Volume 4. Palgrave Macmillan, Hampshire, UK, pp. 6–12.
White, H., Domowitz, I., 1984. Nonlinear regression with dependent observations. Econometrica 52, 143–162.
Wilhelm, D., 2015. Optimal bandwidth selection for robust generalized method of moments estimation. Econometric Theory 31, 1054–1077.
Xiao, Z., Linton, O., 2002. A nonparametric prewhitened covariance estimator. Journal of Time Series Analysis 23, 215–250.
Ye, X., Sun, Y., 2018. Heteroscedasticity and autocorrelation robust *F* and *t* tests in Stata. Stata Journal 18, 951–980.