Journal of Econometrics 207 (2018) 381–405

Contents lists available at ScienceDirect

Journal of Econometrics

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

Should we go one step further? An accurate comparison of one-step and two-step procedures in a generalized method of moments framework☆

Jungbin Hwang [a], Yixiao Sun [b,*]

[a] Department of Economics, University of Connecticut, United States
[b] Department of Economics, University of California, San Diego, United States

ARTICLE INFO

Article history:
Received 16 September 2015
Received in revised form 27 February 2018
Accepted 24 July 2018
Available online 19 September 2018

JEL classification:
C12
C32

Keywords:
Asymptotic mixed normality
Fixed-smoothing asymptotics
Heteroskedasticity and autocorrelation
Nonstandard asymptotics
Two-step GMM estimation

ABSTRACT

According to conventional asymptotic theory, the two-step generalized method of moments (GMM) estimator and test perform at least as well as the one-step estimator and test in large samples. The conventional asymptotic theory completely ignores the estimation uncertainty in the weighting matrix, and as a result it may not reflect finite-sample situations well. In this paper, we employ the more accurate fixed-smoothing asymptotic framework to compare the performances of the one-step and two-step procedures. We show that the two-step procedures outperform the one-step procedures only when the squared long-run canonical correlation coefficients between two blocks of transformed moment conditions are larger than the thresholds established in this paper. The thresholds depend on the criteria of interest. A Monte Carlo study lends support to our asymptotic results.

© 2018 Elsevier B.V. All rights reserved.

1. Introduction

Efficiency is one of the most important issues in statistics and econometrics. In the widely used GMM framework, it is standard practice to employ a two-step procedure to improve the efficiency of the GMM estimator and the power of the associated tests. The two-step procedure entails estimation of a weighting matrix. According to Hansen (1982), the inverse of the optimal weighting matrix is the asymptotic variance of the (scaled) sample moment conditions. For time-series data, which is our focus here, the inverse of the optimal weighting matrix is usually referred to as the long-run variance (LRV) of the moment conditions. To be completely general, we often estimate the LRV using the nonparametric kernel method.

Under the conventional asymptotics, both the one-step and two-step GMM estimators are asymptotically normal.[1] In general, the two-step GMM estimator has a smaller asymptotic variance. Statistical tests based on the two-step estimator

☆ We thank the coeditor Oliver Linton, an associate editor, and anonymous referees for very helpful comments. We gratefully acknowledge constructive suggestions from Brendan Beare, Graham Elliott, Bruce Hansen, Jonathan Hill, Min Seong Kim, Seunghwa Rho, Peter Robinson, Peter Schmidt, Andres Santos, Xiaoxia Shi, Valentin Verdier, Tim Vogelsang, Jeffrey Wooldridge, and seminar participants at LSU, Madison, Michigan State, UNC/Duke/NCSU, the 2014 Shanghai Jiao Tong University and Singapore Management University Bi-party Conference, the 2014 Shandong Econometrics Conference, China, and the 2015 World Congress of the Econometric Society, Montréal, Canada. This research is partially supported by the National Science Foundation under Grant No. SES-1530592.

* Correspondence to: Department of Economics, University of California, San Diego, 9500 Gilman Drive, La Jolla, CA 92093-0508, United States.
E-mail addresses: jungbin.hwang@uconn.edu (J. Hwang), yisun@ucsd.edu (Y. Sun).

[1] In this paper, the term one-step estimator refers to the first-step estimator in a typical two-step GMM framework. This is not to be confused with the continuously updating GMM estimator that involves only one step. We use the terms "one-step" and "first-step" interchangeably. Our use of "one-step" and "two-step" is the same as the use of these terms in the Stata "gmm" command.

https://doi.org/10.1016/j.jeconom.2018.07.006
0304-4076/© 2018 Elsevier B.V. All rights reserved.

are also asymptotically more powerful than those based on the one-step estimator. A driving force behind these results is that the two-step estimator and the associated tests have the same asymptotic properties as the corresponding ones when the optimal weighting matrix is known. However, given that the optimal weighting matrix is estimated nonparametrically in the time-series setting, there is a large estimation uncertainty. Good approximations to the distributions of the two-step estimator and the associated tests should reflect this relatively high estimation uncertainty.

One of the goals of this paper is to compare the asymptotic properties of the one-step and two-step procedures when the estimation uncertainty in the weighting matrix is accounted for. To this end, we consider a limiting thought experiment in which the amount of nonparametric smoothing is held fixed as the sample size increases. This leads to the so-called fixed-smoothing asymptotics, which retains the estimation uncertainty in the LRV estimator or the weighting matrix estimator.

We use the fixed-smoothing asymptotics to compare the one-step and two-step procedures. For the one-step procedure, the LRV estimator is used in computing the standard errors, leading to the popular heteroskedasticity and autocorrelation robust (HAR) standard errors; see, for example, Newey and West (1987) and Andrews (1991). For the two-step procedure, the LRV estimator not only appears in the standard error estimation but also plays the role of the optimal weighting matrix in the second-step GMM criterion function. Under the fixed-smoothing asymptotics, the weighting matrix converges to a random matrix. As a result, the second-step GMM estimator is not asymptotically normal but rather asymptotically mixed normal. The asymptotic mixed normality reflects the estimation uncertainty of the GMM weighting matrix and is expected to be closer to the finite-sample distribution of the second-step GMM estimator. In a recent paper, Sun (2014b) shows that both the one-step and two-step test statistics are asymptotically pivotal under the new asymptotics. Thus a nuisance-parameter-free comparison of the one-step and two-step tests is possible.

Comparing the one-step and two-step procedures under the new asymptotics is fundamentally different from comparing them under the conventional asymptotics. Under the new asymptotics, the two-step procedure outperforms the one-step procedure only when the benefit of using the optimal weighting matrix outweighs the cost of estimating it. The theoretical results in this paper suggest that we should go the extra step –and employ the two-step estimator and test –only when the long-run canonical correlations between two blocks of transformed moment conditions are larger than the thresholds established in this paper. The transformed moment conditions are obtained by rotating and standardizing the original moment conditions. The necessity of rotation and standardization makes it delicate and challenging to compare the performances of the one-step and two-step procedures in a general GMM framework.

To confront this challenge, we transform a general GMM model into a simple overidentified location model. For the latter model, the first block of data identifies the unknown location parameter, and the second block consists of auxiliary information that can be utilized to reduce the long-run variability of the first block of data and hence improve the efficiency of the location estimator. The simple location model enables us to derive a clear qualitative result: We should employ the two-step GMM procedure only when there is a sufficiently large long-run correlation between two blocks of moment conditions. The threshold for the asymptotic variance comparison is different from that for the local asymptotic power comparison. Our theory fully characterizes the threshold for each criterion under consideration.

The theory developed in this paper is in sharp contrast to the conventional asymptotic theory where the cost of estimating the optimal weighting matrix is completely ignored. Since the new asymptotic approximation is more accurate than the conventional asymptotic approximation, comparing the two procedures under this new asymptotics will give a more honest assessment of their relative merits. This is confirmed by a Monte Carlo study.

It is important to point out that the fixed-smoothing asymptotics does not capture the finite-sample bias of the LRV estimator. This is a drawback of this type of asymptotics. For this reason, our asymptotic results may not reflect finite-sample situations very well when there is over-smoothing and hence the bias of the LRV estimator may not be small. While the conventional asymptotics ignores both the bias effect and the variance effect of the LRV estimator, the fixed-smoothing asymptotics accounts for one of the two effects, that is, the variance effect, and hence is a step in the right direction. A higher-order conventional asymptotic theory, which is beyond the scope of this paper, may help shed light on both effects.

There is a large and growing literature on the fixed-smoothing asymptotics. For kernel LRV estimators, the fixed-smoothing asymptotics is the so-called the fixed-b asymptotics, which was first studied by Kiefer et al. (2002) and Kiefer and Vogelsang (2002a, b, 2005) in the econometrics literature. For other studies, see, for example, Jansson (2004), Sun et al. (2008), Sun and Phillips (2009), Gonçalves and Vogelsang (2011), Zhang and Shao (2013), Zhang (2016) and Lazarus et al. (2016) in the time-series setting; Bester et al. (2016) in the spatial setting; and Gonçalves (2011), Kim and Sun (2013), and Vogelsang (2012) in the panel-data setting. In the nonparametric settings, Kim et al. (2017) and Chen et al. (2014) have considered the fixed-smoothing asymptotics for kernel methods and sieve methods, respectively. The approximation approaches in some other papers can also be regarded as special cases of the fixed-smoothing asymptotics. This includes, among others, Ibragimov and Müller (2010), Shao (2010), and Bester et al. (2011). The fixed-smoothing asymptotics can be regarded as a convenient device for obtaining some higher-order terms under the conventional increasing-smoothing asymptotics.

While all these papers except Sun (2014b) consider one-step statistical inference procedures in exactly identified models, this paper and Sun (2014b) are the first to consider the fixed-smoothing asymptotics in a possibly overidentified two-step GMM framework that fully accounts for the estimation uncertainty in the LRV estimator. The difference between Sun (2014b) and this paper is that Sun (2014b) establishes the fixed-smoothing asymptotic theory, while this paper uses the theory to compare the one-step and two-step procedures. Both point estimation and hypothesis testing are considered. While all the papers in the literature use the fixed-smoothing asymptotics to improve the approximation quality of a given procedure, this paper is the first to use this type of asymptotics to compare the performances of different procedures.

The rest of the paper is organized as follows. The next section presents a simple overidentified GMM framework. Section 3 compares the one-step and two-step procedures from the perspective of point estimation and hypothesis testing. Section 4 extends the ideas and results to a general GMM framework. Section 5 presents simulation evidence. On the basis of the theoretical results and the simulation evidence, Section 6 provides some practical guidance. The last section concludes. Proofs are provided in Appendix A. An online supplementary appendix (Appendix B) contains additional theoretical and simulation results and proofs.

A word on notation: For a symmetric matrix A, $A^{1/2}$ (or $A_{1/2}$) is a matrix square root of A such that $A^{1/2}\left(A^{1/2}\right)' = A$. Note that $A^{1/2}$ does not have to be symmetric. We will specify $A^{1/2}$ explicitly when it is not symmetric. If not specified, $A^{1/2}$ is a symmetric matrix square root of A based on its eigen-decomposition. For matrices A, B, we use "$A \geq B$" to signify that $A-B$ is positive (semi)definite. We use "0" and "O" interchangeably to denote a matrix of zeros whose dimension may differ from one occurrence in the paper to another. For two random variables X, Y, we use $X \perp Y$ to indicate that X and Y are independent. For a matrix A, we use $\nu(A)$, $\nu_{\min}(A)$, and $\nu_{\max}(A)$ to denote the set of all singular values, the smallest singular value, and the largest singular value of A, respectively. For an estimator $\hat{\theta}$, we use $\text{avar}(\hat{\theta})$ to denote the asymptotic variance of the limiting distribution of $\sqrt{T}(\hat{\theta} - \text{plim}_{T\to\infty}\hat{\theta})$, where T is the sample size. Lastly, $\chi^2_p\left(\lambda^2\right)$ is the noncentral chi-square distribution with noncentrality parameter λ^2.

2. Simple overidentified GMM framework

To illustrate the basic ideas of this paper, we consider a simple overidentified time-series model of the form

$$\begin{aligned} &y_{1t} = \theta_0 + u_{1t},\ y_{1t} \in \mathbb{R}^d, \\ &y_{2t} = u_{2t},\ y_{2t} \in \mathbb{R}^q \end{aligned} \tag{1}$$

for $t = 1,\ldots, T$, where $\theta_0 \in \mathbb{R}^d$ is the parameter of interest and the vector process $u_t := (u'_{1t}, u'_{2t})'$ is stationary and has mean 0. The long-run variance Ω of u_t is

$$\Omega = \text{lrvar}(u_t) = \sum_{j=-\infty}^{\infty} Eu_t u'_{t-j} = \begin{pmatrix} \Omega_{11} & \Omega_{12} \\ \Omega_{21} & \Omega_{22} \end{pmatrix}, \tag{2}$$

where Ω_{11} is a $d \times d$ matrix and Ω_{22} is a $q \times q$ matrix. We allow u_t to have autocorrelation of unknown forms so that the long-run variance Ω can take a general form. However, for simplicity, we assume for now that $\text{var}(u_t) = \sigma^2 I_{d+q}$.[2] The model considered in this section is just a location model. We initially consider a general GMM framework but later find that our points can be made more clearly in reference to the simple location model. From the asymptotic point of view, we show later that a general GMM framework can be reduced to the above simple location model. The proofs for the general GMM case are also built upon those for the location model. Therefore, we use the location model not only to illustrate the basic ideas but also to establish a technical foundation for the general GMM framework.

If the location model is embedded in a GMM framework, the moment conditions are

$$E(y_t) - \left(\theta'_0, 0'_q\right)' = 0,$$

where 0_q is the $q \times 1$ vector of zeros and $y_t = \left(y'_{1t}, y'_{2t}\right)'$. Let

$$g_T(\theta) = \left(\frac{1}{\sqrt{T}}\sum_{t=1}^{T}(y_{1t} - \theta)', \frac{1}{\sqrt{T}}\sum_{t=1}^{T} y'_{2t}\right)'.$$

Then a GMM estimator of θ_0 can be defined as

$$\hat{\theta}_{\text{GMM}} = \arg\min_{\theta} g_T(\theta)' W_T^{-1} g_T(\theta)$$

for some positive-definite weighting matrix W_T. Writing W_{12} and W_{22} as submatrices of W_T in a way similar to (2), it is easy to show that

$$\hat{\theta}_{\text{GMM}} = \frac{1}{T}\sum_{t=1}^{T}(y_{1t} - \beta_W y_{2t}) \text{ for } \beta_W = W_{12}W_{22}^{-1}.$$

[2] If

$$\text{var}(u_t) = \begin{pmatrix} \mathbb{V}_{11} & \mathbb{V}_{12} \\ \mathbb{V}_{21} & \mathbb{V}_{22} \end{pmatrix} \neq \sigma^2 I_{d+q},$$

for any $\sigma^2 > 0$, we can let

$$\mathbb{V}_{1/2} = \begin{pmatrix} (\mathbb{V}_{1\cdot 2})^{1/2} & \mathbb{V}_{12}(\mathbb{V}_{22})^{-1/2} \\ 0 & (\mathbb{V}_{22})^{1/2} \end{pmatrix},$$

where $\mathbb{V}_{1\cdot 2} = \mathbb{V}_{11} - \mathbb{V}_{12}\mathbb{V}_{22}^{-1}\mathbb{V}_{21}$. Then $\mathbb{V}_{1/2}^{-1}\left(y'_{1t}, y'_{2t}\right)'$ can be written as a location model whose error variance is the identity matrix I_{d+q}. The estimation uncertainty in estimating $\mathbb{V}$ will not affect our asymptotic results.

There are at least two different choices of W_T. First, we can take W_T to be the identity matrix I_m for $m = d+q$. In this case, $\beta_W = 0$, and the GMM estimator $\hat{\theta}_{\text{GMM}}$, which is now written as $\hat{\theta}_1$, is simply a sample mean of the first block of data, that is, $\hat{\theta}_1 = T^{-1}\sum_{t=1}^{T} y_{1t}$. Second, we can take W_T to be the "optimal" weighting matrix Ω. With this choice, we obtain the second GMM estimator, which is now written as $\tilde{\theta}_2 = T^{-1}\sum_{t=1}^{T}(y_{1t} - \beta_0 y_{2t})$, where $\beta_0 = \Omega_{12}\Omega_{22}^{-1}$ is the long-run regression coefficient matrix. While $\hat{\theta}_1$ completely ignores the information in $\{y_{2t}\}$, $\tilde{\theta}_2$ takes advantage of this source of information.

Under some moment and mixing conditions, we have

$$\sqrt{T}(\hat{\theta}_1 - \theta_0) \xrightarrow{d} N(0, \Omega_{11}) \text{ and } \sqrt{T}(\tilde{\theta}_2 - \theta_0) \xrightarrow{d} N(0, \Omega_{1\cdot 2}), \tag{3}$$

where $\Omega_{1\cdot 2} = \Omega_{11} - \Omega_{12}\Omega_{22}^{-1}\Omega_{21}$. Thus $\text{avar}(\tilde{\theta}_2) < \text{avar}(\hat{\theta}_1)$ unless $\Omega_{12} = 0$. This is a well-known result in the literature. Since we do not know Ω in practice, $\tilde{\theta}_2$ is infeasible. However, given the feasible estimator $\hat{\theta}_1$, we can estimate Ω and construct a feasible version of $\tilde{\theta}_2$. The common two-step estimation strategy proceeds as follows.

(i) Estimate the long-run covariance matrix Ω by

$$\hat{\Omega} := \hat{\Omega}(\hat{u}) = \sum_{j=-(T-1)}^{T-1} k\left(\frac{j}{S_T}\right)\hat{\Gamma}_j, \tag{4}$$

where

$$\hat{\Gamma}_j = \begin{cases} T^{-1}\sum_{t=j+1}^{T}[\hat{u}_t - T^{-1}\sum_{\tau=1}^{T}\hat{u}_\tau][\hat{u}_{t-j} - T^{-1}\sum_{\tau=1}^{T}\hat{u}_\tau]' \text{ for } j \geq 0, \\ T^{-1}\sum_{t=j+1}^{T}[\hat{u}_{t+j} - T^{-1}\sum_{\tau=1}^{T}\hat{u}_\tau][\hat{u}_t - T^{-1}\sum_{\tau=1}^{T}\hat{u}_\tau]' \text{ for } j < 0 \end{cases}$$

and $\hat{u}_t = (y_{1t}' - \hat{\theta}_1', y_{2t}')'$. Here S_T is the so-called truncation lag parameter, and $k(\cdot)$ is a kernel function.

(ii) Obtain the feasible two-step estimator $\hat{\theta}_2 = T^{-1}\sum_{t=1}^{T}(y_{1t} - \hat{\beta}y_{2t})$, where $\hat{\beta} = \hat{\Omega}_{12}\hat{\Omega}_{22}^{-1}$, and $\hat{\Omega}_{12}$ and $\hat{\Omega}_{22}$ are the submatrices of $\hat{\Omega}$ that correspond to the partition in (2).

The LRV estimator $\hat{\Omega}$ in (4) is the classical kernel-type heteroskedasticity–autocorrelation-consistent estimator; see, for example, Newey and West (1987) and Andrews (1991). Letting

$$Q_h(s,t) = k\left(\frac{s-t}{b}\right) \text{ with } b = \frac{S_T}{T} \text{ and } h = \frac{1}{b},$$

we can rewrite the kernel LRV estimator in a quadratic form:

$$\hat{\Omega} = \frac{1}{T}\sum_{s=1}^{T}\sum_{t=1}^{T} Q_h\left(\frac{s}{T}, \frac{t}{T}\right)\left(\hat{u}_t - \frac{1}{T}\sum_{\tau=1}^{T}\hat{u}_\tau\right)\left(\hat{u}_s - \frac{1}{T}\sum_{\tau=1}^{T}\hat{u}_\tau\right)'. \tag{5}$$

With a different parametrization of the smoothing parameter, such a representation has the advantage of including other commonly used LRV estimators, for example, the orthogonal series estimators of Phillips (2005), Müller (2007), and Sun (2013, 2014b). In this paper, we focus on kernel LRV estimators and leave the extension to other types of LRV estimators to the working paper Hwang and Sun (2015).

While $\tilde{\theta}_2$ is asymptotically more efficient than $\hat{\theta}_1$, is $\hat{\theta}_2$ necessarily more efficient than $\hat{\theta}_1$ and in what sense? Is the Wald test based on $\hat{\theta}_2$ necessarily more powerful than that based on $\hat{\theta}_1$? One of the objectives of this paper is to address these questions.

3. A tale of two asymptotics

3.1. Point estimation

We first consider the conventional increasing-smoothing asymptotics where $h \to \infty$ as $T \to \infty$ but at a slower rate, that is, $h/T \to 0$. Under this type of asymptotics and some regularity conditions, we have $\hat{\Omega} \xrightarrow{p} \Omega$. It can then be shown that $\hat{\theta}_2$ is asymptotically equivalent to $\tilde{\theta}_2$, that is, $\sqrt{T}(\tilde{\theta}_2 - \hat{\theta}_2) = o_p(1)$. As a direct consequence, the limiting distribution of the feasible estimator $\hat{\theta}_2$ is identical to that of the infeasible estimator $\tilde{\theta}_2$ in (Eq. (3)), and $\hat{\theta}_2$ is still asymptotically more efficient than $\hat{\theta}_1$.

The conventional asymptotics, elegant and convenient though it may be, does not reflect the finite-sample situations well. Under this type of asymptotics, we essentially approximate the distribution of $\hat{\Omega}$ by the degenerate distribution concentrating on Ω. That is, we completely ignore the estimation uncertainty in $\hat{\Omega}$. The degenerate approximation is too optimistic, as $\hat{\Omega}$ is a nonparametric estimator, which by definition can have high variation in finite samples.

To obtain a more accurate distributional approximation of $\sqrt{T}(\hat{\theta}_2 - \theta_0)$, we could develop a higher-order increasing-smoothing asymptotics that reflects the estimation uncertainty in $\hat{\Omega}$. This is possible but requires strong assumptions that cannot easily be verified. In addition, it is technically challenging and tedious to rigorously justify the use of a higher-order asymptotic theory. Instead of a higher-order asymptotic theory under the conventional asymptotics, we adopt the fixed-smoothing asymptotics where h is held fixed (at a positive value) as $T \to \infty$. This type of asymptotics takes the sampling variability of $\hat{\Omega}$ into consideration.

To establish the fixed-smoothing asymptotics, we maintain Assumption 1 on the kernel function.

Assumption 1. For any $b \in (0, 1]$, the scaled kernel function $k_b(x) = k(x/b)$ is symmetric, continuous, piecewise monotonic, and piecewise continuously differentiable on $[-1, 1]$.

Assumption 1 is the same as Assumption 1(a) in Sun (2014b). This assumption is very mild. It allows for many commonly used kernel functions such as the Bartlett kernel, the Parzen kernel, and the quadratic spectral (QS) kernel.

Following Sun (2014b, p. 2333), we define

$$Q_h^*(r,s) = Q_h(r,s) - \int_0^1 Q_h(\tau,s)d\tau - \int_0^1 Q_h(r,\tau)d\tau + \int_0^1\int_0^1 Q_h(\tau_1,\tau_2)d\tau_1 d\tau_2,$$

and the corresponding LRV estimator

$$\tilde{\Omega} = \frac{1}{T}\sum_{s=1}^{T}\sum_{t=1}^{T} Q_h^*\left(\frac{s}{T},\frac{t}{T}\right)\hat{u}_t\hat{u}_s'.$$

Assumption 1 ensures that $\tilde{\Omega}$ and $\hat{\Omega}$ are asymptotically equivalent; see Lemma 1 in Sun (2014b). Furthermore, under this assumption, Sun (2014a) shows that $Q_h^*(r,s)$ can be represented as

$$Q_h^*(r,s) = \sum_{j=1}^{\infty}\lambda_j\Phi_j(r)\Phi_j(s), \tag{6}$$

where $\{\Phi_j(r)\}$ is a sequence of continuously differentiable functions satisfying $\int_0^1\Phi_j(r)dr = 0$, and the series on the right-hand side of (6) converges to $Q_h^*(r,s)$ absolutely and uniformly over $(r,s) \in [0,1]\times[0,1]$.

Now letting $\Phi_0(\cdot) := 1$ and using the basis functions $\{\Phi_j(\cdot)\}_{j=1}^{\infty}$ in the series representation of the weighting function, we make the following assumptions.

Assumption 2. $\sum_{j=-\infty}^{\infty}\|E(u_t u_{t-j}')\| < \infty$.

Assumption 3. As $T \to \infty$, $T^{-1/2}\sum_{t=1}^{T}\Phi_j(t/T)u_t$ converges weakly to a continuous distribution, jointly over $j \in 0, 1, \ldots, J$, for every fixed J, such that for every $x \in \mathbb{R}^m$,

$$P\left(\frac{1}{\sqrt{T}}\sum_{t=1}^{T}\Phi_j(\frac{t}{T})u_t \le x\right) = P\left(\Omega_{1/2}\frac{1}{\sqrt{T}}\sum_{t=1}^{T}\Phi_j(\frac{t}{T})e_t \le x\right) + o(1),$$

where $\Omega_{1/2}$ is a matrix square root of the nonsingular LRV matrix Ω and $e_t \sim iid\ N(0, I_m)$.

Assumption 2 is the same as Assumption 3 in Sun (2014b). Assumption 3 is a multivariate central limit theorem (CLT); it is the same as Assumption 5 in Sun (2014b). Under Assumption 1, Assumption 3 follows from the functional central Limit theorem (FCLT) for the partial-sum process of u_t:

$$\frac{1}{\sqrt{T}}\sum_{t=1}^{[T\cdot]} u_t \xrightarrow{d} \Omega_{1/2}B_m(\cdot), \tag{7}$$

where $B_m(\cdot) = (B_d'(\cdot), B_q'(\cdot))'$ is an m-dimensional standard Brownian motion. We note that the above FCLT is often assumed in the literature on the fixed-smoothing asymptotics, and it implicitly assumes that the unconditional variance of $\{u_t\}$ does not change over time. See, for example, Kiefer and Vogelsang (2002a, b, 2005) and Sun et al. (2008). If the variance of $\{u_t\}$ evolves over time, then the fixed-smoothing asymptotic distributions will not be pivotal in general, whereas the conventional increasing-smoothing asymptotic distributions can still be pivotal under some conditions; see Xu (2012) for further discussion of this. Our paper does not cover the case with unconditional heteroskedasticity.

Proposition 1. *Let Assumptions 1–3 hold. Then as $T \to \infty$ for a fixed $h > 0$, we have:*

(a) $\hat{\Omega} \xrightarrow{d} \Omega_\infty$, *where*

$$\Omega_\infty = \Omega_{1/2}\tilde{\Omega}_\infty\Omega_{1/2}' := \begin{pmatrix} \Omega_{\infty,11} & \Omega_{\infty,12} \\ \Omega_{\infty,21} & \Omega_{\infty,22} \end{pmatrix},$$

$$\tilde{\Omega}_\infty = \int_0^1 \int_0^1 Q_h^*(r,s)dB_m(r)dB_m(s)' := \begin{pmatrix} C_{dd} & C_{dq} \\ C_{qd} & C_{qq} \end{pmatrix}.$$

(b) $\sqrt{T}(\hat{\theta}_2 - \theta_0) \xrightarrow{d} \left(I_d, \quad -\beta_\infty\right)\Omega_{1/2}B_m$, *where* $\beta_\infty = \beta_\infty(h,d,q) := \Omega_{\infty,12}\Omega_{\infty,22}^{-1}$ *is independent of the standard normal vector* $B_m := B_m(1)$.

The proof of Proposition 1 follows directly from Lemma 1 of Sun (2014b) and the continuous mapping theorem. Conditional on β_∞, the asymptotic distribution of $\sqrt{T}(\hat{\theta}_2 - \theta_0)$ is a normal distribution with variance

$$V_2 = \left(I_d, \quad -\beta_\infty\right)\begin{pmatrix} \Omega_{11} & \Omega_{12} \\ \Omega_{21} & \Omega_{22} \end{pmatrix}\begin{pmatrix} I_d \\ -\beta_\infty' \end{pmatrix} = \Omega_{11} - \Omega_{12}\beta_\infty' - \beta_\infty\Omega_{21} + \beta_\infty\Omega_{22}\beta_\infty'.$$

Given that V_2 is random, $\sqrt{T}(\hat{\theta}_2 - \theta_0)$ is asymptotically mixed normal rather than normal. Asymptotic mixed normality in a general GMM framework was first established in Sun (2014b). Since

$$\begin{aligned} &\text{avar}(\hat{\theta}_2) - \text{avar}(\tilde{\theta}_2) = EV_2 - \left(\Omega_{11} - \Omega_{12}\Omega_{22}^{-1}\Omega_{21}\right) \\ &= E\left(\Omega_{12}\Omega_{22}^{-1}\Omega_{21} - \Omega_{12}\beta_\infty' - \beta_\infty\Omega_{21} + \beta_\infty\Omega_{22}\beta_\infty'\right) \\ &= E(\beta_0 - \beta_\infty)\Omega_{22}(\beta_0 - \beta_\infty)' \geq 0, \end{aligned}$$

the feasible estimator $\hat{\theta}_2$ has higher variation than the infeasible estimator $\tilde{\theta}_2$. This is consistent with our intuition. The difference $\text{avar}(\hat{\theta}_2) - \text{avar}(\tilde{\theta}_2)$ can be regarded as the cost of implementing the two-step estimator, that is, the cost of having to estimate the weighting matrix.

Under the fixed-smoothing asymptotics, we still have $\sqrt{T}(\hat{\theta}_1 - \theta_0) \xrightarrow{d} N(0, \Omega_{11})$, as $\hat{\theta}_1$ does not depend on the smoothing parameter h. Therefore,

$$\text{avar}(\hat{\theta}_1) - \text{avar}(\tilde{\theta}_2) := \Omega_{11} - \left(\Omega_{11} - \Omega_{12}\Omega_{22}^{-1}\Omega_{21}\right) = \Omega_{12}\Omega_{22}^{-1}\Omega_{21} \geq 0,$$

which can be regarded as the benefit of going to the second step.

To compare the asymptotic variances of $\sqrt{T}(\hat{\theta}_1 - \theta_0)$ and $\sqrt{T}(\hat{\theta}_2 - \theta_0)$, we need to evaluate the relative magnitudes of the cost and the benefit. Define

$$\tilde{\beta}_\infty := \tilde{\beta}_\infty(h,d,q) := C_{dq}C_{qq}^{-1}, \tag{8}$$

which does not depend on any nuisance parameter but rather on h, d, and q. For notational economy, we sometimes suppress this dependence. Direct calculations show that β_∞ and $\tilde{\beta}_\infty$ are related via

$$\beta_\infty = \Omega_{1\cdot2}^{1/2}\tilde{\beta}_\infty\Omega_{22}^{-1/2} + \Omega_{12}\Omega_{22}^{-1}. \tag{9}$$

Using this, we have

$$\begin{aligned} \text{avar}(\hat{\theta}_2) - \text{avar}(\hat{\theta}_1) &= \underbrace{\text{avar}(\hat{\theta}_2) - \text{avar}(\tilde{\theta}_2)}_{\text{cost}} - \underbrace{[\text{avar}(\hat{\theta}_1) - \text{avar}(\tilde{\theta}_2)]}_{\text{benefit}} \\ &= \Omega_{1\cdot2}^{1/2}E\tilde{\beta}_\infty\tilde{\beta}_\infty'(\Omega_{1\cdot2}^{1/2})' - \Omega_{12}\Omega_{22}^{-1}\Omega_{21}. \end{aligned} \tag{10}$$

If the cost is larger than the benefit, that is, $\Omega_{1\cdot2}^{1/2}E\tilde{\beta}_\infty\tilde{\beta}_\infty'(\Omega_{1\cdot2}^{1/2})' > \Omega_{12}\Omega_{22}^{-1}\Omega_{21}$, then the asymptotic variance of $\hat{\theta}_2$ is larger than that of $\hat{\theta}_1$. In the online appendix, we show that

$$\text{avar}(\hat{\theta}_2) - \text{avar}(\hat{\theta}_1) = (1 + E\|\tilde{\beta}_\infty(h,1,q)\|^2)\Omega_{11}^{1/2}\left[g(h,q)I_d - \rho\rho'\right](\Omega_{11}^{1/2})', \tag{11}$$

where

$$g(h,q) := \frac{E\|\tilde{\beta}_\infty(h,1,q)\|^2}{1 + E\|\tilde{\beta}_\infty(h,1,q)\|^2} \in (0,1)$$

and $\rho = \Omega_{11}^{-1/2}\Omega_{12}\Omega_{22}^{-1/2}$ is the long-run correlation matrix between u_{1t} and u_{2t}. The proposition below then follows immediately.

Proposition 2. *Let Assumptions 1–3 hold, and consider the fixed-smoothing asymptotics.*

(a) *If* $\nu_{\max}\left(\rho\rho'\right) < g(h,q)$, *then* $\hat{\theta}_2$ *has a larger asymptotic variance than* $\hat{\theta}_1$.

(b) *If* $\nu_{\min}\left(\rho\rho'\right) > g(h,q)$, *then* $\hat{\theta}_2$ *has a smaller asymptotic variance than* $\hat{\theta}_1$.

To compute the eigenvalues of $\rho\rho'$, we can use the fact that $\nu\left(\rho\rho'\right) = \nu\left(\Omega_{12}\Omega_{22}^{-1}\Omega_{21}\Omega_{11}^{-1}\right)$. The eigenvalues of $\rho\rho'$ are the squared long-run correlation coefficients between $c_1'u_{1t}$ and $c_2'u_{2t}$ for some c_1 and c_2, that is, the squared long-run canonical correlation coefficients between u_{1t} and u_{2t}. Therefore, the conditions in Proposition 2 can be presented in terms of the smallest and largest squared long-run canonical correlation coefficients.

If $\rho = 0$, then $\nu_{\max}\left(\rho\rho'\right) < g(h, q)$ holds trivially. In this case, the asymptotic variance of $\hat{\theta}_2$ is larger than the asymptotic variance of $\hat{\theta}_1$. Intuitively, when the long-run correlation is zero, there is no information that can be exploited to improve the efficiency. If we insist on using the long-run correlation matrix in an attempt to improve the efficiency, we may end up with a less efficient estimator because of the noise in estimating the zero long-run correlation matrix. On the other hand, if $\rho\rho' = I_d$ after some possible rotation, which holds when the long-run variation of u_{1t} is perfectly predicted by u_{2t}, then $\nu_{\min}\left(\rho\rho'\right) = 1$ and we have $\nu_{\min}\left(\rho\rho'\right) > g(h, q)$. In this case, it is worthwhile to estimate the long-run variance and use it to improve the efficiency $\hat{\theta}_2$.

The conditions $\nu_{\min}\left(\rho\rho'\right) > g(h, q)$ and $\nu_{\max}\left(\rho\rho'\right) < g(h, q)$ in Proposition 2 may appear to be strong. However, the conclusions are also very strong. For example, to say that $\hat{\theta}_2$ has a smaller asymptotic variance than $\hat{\theta}_1$ means that $\text{avar}(R\hat{\theta}_2) \leq \text{avar}(R\hat{\theta}_1)$ for *any* matrix $R \in \mathbb{R}^{p\times d}$ and all p with $1 \leq p \leq d$. In fact, the conditions are both necessary and sufficient. If we care about only a specific linear combination of θ_0, we can use Proposition 2 (and Theorem 7) to evaluate the efficiency of the one-step estimator of the linear combination relative to that of the two-step estimator.

3.2. Hypothesis testing

We are interested in testing the null hypothesis $H_0 : \theta_0 = r$ against the local alternative $H_1 : \theta_0 = r + \delta_0/\sqrt{T}$ for some $d \times 1$ vectors r and δ_0. Nonlinear restrictions can be converted to linear restrictions using the Delta method. We construct the following two Wald statistics:

$$\mathbb{W}_1 := T(\hat{\theta}_1 - r)'\hat{\Omega}_{11}^{-1}(\hat{\theta}_1 - r),$$
$$\mathbb{W}_2 := T(\hat{\theta}_2 - r)'\hat{\Omega}_{1\cdot2}^{-1}(\hat{\theta}_2 - r),$$

where $\hat{\Omega}_{1\cdot2} = \hat{\Omega}_{11} - \hat{\Omega}_{12}\hat{\Omega}_{22}^{-1}\hat{\Omega}_{21}$. Regardless of whether the test is based on $\hat{\theta}_1$ or $\hat{\theta}_2$, we have to employ the long-run covariance estimator $\hat{\Omega}$.

Under the conventional increasing-smoothing asymptotics, it is straightforward to show that under $H_1 : \theta_0 = r + \delta_0/\sqrt{T}$ we have

$$\mathbb{W}_1 \xrightarrow{d} \chi_d^2(\delta_0'\Omega_{11}^{-1}\delta_0) \text{ and } \mathbb{W}_2 \xrightarrow{d} \chi_d^2(\delta_0'\Omega_{1\cdot2}^{-1}\delta_0).$$

When $\delta_0 = 0$, we obtain the null distributions $\mathbb{W}_1, \mathbb{W}_2 \xrightarrow{d} \chi_d^2$. Thus under the conventional increasing-smoothing asymptotics, the null limiting distributions of $\mathbb{W}_1$ and $\mathbb{W}_2$ are identical. Since $\delta_0'\Omega_{11}^{-1}\delta_0 \leq \delta_0'\Omega_{1\cdot2}^{-1}\delta_0$, under the conventional asymptotics the local asymptotic power function of the test based on $\mathbb{W}_2$ is higher than that based on $\mathbb{W}_1$.

The key driving force behind the conventional asymptotics is that we approximate the distribution of $\hat{\Omega}$ by the degenerate distribution concentrating on Ω. The degenerate approximation does not reflect the finite-sample distribution well. As in the previous subsection, we employ the fixed-smoothing asymptotics to derive more accurate distributional approximations. To simplify the notation, we introduce $D_{dd} := C_{dd} - C_{dq}C_{qq}^{-1}C_{qd}'$. Also, we define $B_d := B_d(1)$ and $B_q := B_q(1)$.

Proposition 3. *Let Assumptions 1–3 hold. Then as $T \to \infty$ for a fixed h, we have the following under $H_1 : \theta_0 = r + \delta_0/\sqrt{T}$:*

(a) $\mathbb{W}_1 \xrightarrow{d} \mathbb{W}_{1\infty}(\|\Omega_{11}^{-1/2}\delta_0\|^2)$, *where*

$$\mathbb{W}_{1\infty}(\|\xi\|^2) = [B_d + \xi]'C_{dd}^{-1}[B_d + \xi] \text{ for } \xi \in \mathbb{R}^d. \tag{12}$$

(b) $\mathbb{W}_2 \xrightarrow{d} \mathbb{W}_{2\infty}(\|\Omega_{1\cdot2}^{-1/2}\delta_0\|^2)$, *where*

$$\mathbb{W}_{2\infty}(\|\xi\|^2) = \left[B_d - \tilde{\beta}_\infty B_q + \xi\right]'D_{dd}^{-1}\left[B_d - \tilde{\beta}_\infty B_q + \xi\right] \text{ for } \xi \in \mathbb{R}^d. \tag{13}$$

When $\delta_0 = 0$, we obtain the limiting distributions of $\mathbb{W}_1$ and $\mathbb{W}_2$ under the null hypothesis:

$$\mathbb{W}_1 \xrightarrow{d} \mathbb{W}_{1\infty} := \mathbb{W}_{1\infty}(0) = B_d'C_{dd}^{-1}B_d,$$
$$\mathbb{W}_2 \xrightarrow{d} \mathbb{W}_{2\infty} := \mathbb{W}_{2\infty}(0) = \left[B_d - \tilde{\beta}_\infty B_q\right]'D_{dd}^{-1}\left[B_d - \tilde{\beta}_\infty B_q\right].$$

These distributions are different from those under the conventional asymptotics. For $\mathbb{W}_1$, the difference lies in the random scaling factor C_{dd}. The random scaling factor captures the estimation uncertainty of the LRV estimator. For $\mathbb{W}_2$, there is an additional difference embodied by the random location shift $\tilde{\beta}_\infty B_q = C_{dq}C_{qq}^{-1}B_q$ with a consequent change in the random scaling factor.

The proposition below provides a characterization of the limiting distributions $\mathbb{W}_{1\infty}$ and $\mathbb{W}_{2\infty}$.

Proposition 4. *For any $x > 0$, the following hold:*

(a) $\mathbb{W}_{2\infty}$ first-order stochastically dominates $\mathbb{W}_{1\infty}$ in that

$$P\left[\mathbb{W}_{2\infty} \geq x\right] > P\left[\mathbb{W}_{1\infty} \geq x\right].$$

(b) $P\left[\mathbb{W}_{1\infty}(\|\xi\|^2) \geq x\right]$ is strictly increases in $\|\xi\|^2$ and $\lim_{\|\xi\|\to\infty} P\left[\mathbb{W}_{1\infty}(\|\xi\|^2) \geq x\right] = 1$.

(c) $P\left[\mathbb{W}_{2\infty}(\|\xi\|^2) \geq x\right]$ is strictly increases in $\|\xi\|^2$ and $\lim_{\|\xi\|\to\infty} P\left[\mathbb{W}_{2\infty}(\|\xi\|^2) \geq x\right] = 1$.

Proposition 4(a) is intuitive. Since $\tilde{\beta}_\infty B_q$ is an independent additive component, $\mathbb{W}_{2\infty}$ first-order stochastically dominates $B_d' D_{dd}^{-1} B_d$. Observing that $D_{dd}^{-1} \geq C_{dd}^{-1}$ almost surely, we know that $B_d' D_{dd}^{-1} B_d$ first-order stochastically dominates $B_d' C_{dd}^{-1} B_d$, which is just $\mathbb{W}_{1\infty}$. By the transitivity of first-order stochastic dominance, $\mathbb{W}_{2\infty}$ first-order stochastically dominates $\mathbb{W}_{1\infty}$.

By a property of the first-order stochastic dominance, we have

$$\mathbb{W}_{2\infty} \stackrel{d}{=} \mathbb{W}_{1\infty} + \mathbb{W}_e$$

for some $\mathbb{W}_e > 0$. Intuitively, $\mathbb{W}_{2\infty}$ shifts some of the probability mass of $\mathbb{W}_{1\infty}$ to the right. A direct implication of this is that the asymptotic critical values for $\mathbb{W}_2$ are larger than the corresponding ones for $\mathbb{W}_1$. The difference in critical values has implications for the power properties of the two tests.

Let $\mathbb{W}_{1\infty}^\alpha$ and $\mathbb{W}_{2\infty}^\alpha$ be the $(1-\alpha)$ quantiles from the distributions $\mathbb{W}_{1\infty}$ and $\mathbb{W}_{2\infty}$, respectively. The local asymptotic power functions of the two tests are

$$\pi_1(\delta_0' \Omega_{11}^{-1} \delta_0) := \pi_1(\delta_0' \Omega_{11}^{-1} \delta_0; h, d, q, \alpha) = P\left[\mathbb{W}_{1\infty}(\|\Omega_{11}^{-1/2}\delta_0\|^2) > \mathbb{W}_{1\infty}^\alpha\right],$$
$$\pi_2(\delta_0' \Omega_{1\cdot 2}^{-1} \delta_0) := \pi_2(\delta_0' \Omega_{1\cdot 2}^{-1} \delta_0; h, d, q, \alpha) = P\left[\mathbb{W}_{2\infty}(\|\Omega_{1\cdot 2}^{-1/2}\delta_0\|^2) > \mathbb{W}_{2\infty}^\alpha\right].$$

While $\delta_0' \Omega_{1\cdot 2}^{-1} \delta_0 \geq \delta_0' \Omega_{11}^{-1} \delta_0$, we also have $\mathbb{W}_{2\infty}^\alpha > \mathbb{W}_{1\infty}^\alpha$. The difference in the noncentrality parameters $\delta_0' \Omega_{1\cdot 2}^{-1} \delta_0 - \delta_0' \Omega_{11}^{-1} \delta_0$ indicates the benefit of using the two-step test, while the difference in the critical values $\mathbb{W}_{2\infty}^\alpha - \mathbb{W}_{1\infty}^\alpha$ indicates the associated cost. The effects of the critical values and the noncentrality parameters move in opposite directions. It is not straightforward to compare the two power functions. However, Proposition 4 suggests that if the difference in the noncentrality parameters $\delta_0' \Omega_{1\cdot 2}^{-1} \delta_0 - \delta_0' \Omega_{11}^{-1} \delta_0$ is large enough to offset the increase in the critical values, then the two-step test based on $\mathbb{W}_2$ will be more powerful.

Note that

$$\delta_0' \Omega_{1\cdot 2}^{-1} \delta_0 = (\Omega_{11}^{-1/2}\delta_0)' \left(I_d - \rho\rho'\right)^{-1} (\Omega_{11}^{-1/2}\delta_0).$$

On the one hand, when the second block of moment conditions has very high long-run prediction power for the first block, we have $\nu_{\min}\left(\rho\rho'\right) \to 1$, and as a result $\delta_0' \Omega_{1\cdot 2}^{-1} \delta_0 \to \infty$. In this case, we expect the test based on $\mathbb{W}_2$ to be more powerful, as $\lim_{\nu_{\min}(\rho\rho')\to 1} \pi_2(\delta_0' \Omega_{1\cdot 2}^{-1} \delta_0) = 1$. On the other hand, when $\nu_{\max}\left(\rho\rho'\right) = 0$, that is, ρ is a matrix of zeros, we have $\delta_0' \Omega_{1\cdot 2}^{-1} \delta_0 = \delta_0' \Omega_{11}^{-1} \delta_0$. In this case, we expect the test based on $\mathbb{W}_2$ to be less powerful.

It follows from parts (b) and (c) of Proposition 4 that for every $\lambda > 0$, there exists a unique $\tau(\lambda) := \tau(\lambda; h, d, q, \alpha)$ such that

$$\pi_2(\lambda) = \pi_1(\lambda/\tau).$$

As a function of λ, $\tau(\lambda)$ is defined implicitly via the above equation. Then $\pi_2(\delta_0' \Omega_{1\cdot 2}^{-1} \delta_0) < \pi_1(\delta_0' \Omega_{11}^{-1} \delta_0)$ if and only if $\delta_0' \Omega_{1\cdot 2}^{-1} \delta_0 < \tau \cdot \delta_0' \Omega_{11}^{-1} \delta_0$. Using

$$\begin{aligned}
&\delta_0' \Omega_{1\cdot 2}^{-1} \delta_0 - \tau \cdot \delta_0' \Omega_{11}^{-1} \delta_0 \\
&= (\Omega_{11}^{-1/2}\delta_0)' [\left(I_d - \rho\rho'\right)^{-1} - \tau I_d] (\Omega_{11}^{-1/2}\delta_0) \\
&= \tau [\left(I_d - \rho\rho'\right)^{-1/2} \Omega_{11}^{-1/2}\delta_0]' \left[\rho\rho' - f(\lambda) \cdot I_d\right] [\left(I_d - \rho\rho'\right)^{-1/2} \Omega_{11}^{-1/2}\delta_0],
\end{aligned}$$

where $f(\cdot)$ is defined according to

$$f(\lambda) := f(\lambda; h, d, q, \alpha) = \frac{\tau(\lambda; h, d, q, \alpha) - 1}{\tau(\lambda; h, d, q, \alpha)},$$

we can prove the proposition below.

Proposition 5. *Let Assumptions 1–3 hold. Define*

$$\mathfrak{A}(\lambda_0) = \{\delta : \delta' \Omega_{1\cdot 2}^{-1} \delta = \lambda_0\} \text{ for some } \lambda_0 > 0.$$

Consider the local alternative $H_1(\lambda_0) : \theta_0 = r + \delta_0/\sqrt{T}$ for $\delta_0 \in \mathfrak{A}(\lambda_0)$ and the fixed-smoothing asymptotics.

(a) If $\nu_{\max}\left(\rho\rho'\right) < f(\lambda_0; h, p, q, \alpha)$, then the two-step test based on $\mathbb{W}_2$ has a lower local asymptotic power than the one-step test based on $\mathbb{W}_1$ for any $\delta_0 \in \mathfrak{A}(\lambda_0)$.

(b) If $\nu_{\min}\left(\rho\rho'\right) > f(\lambda_0; h, p, q, \alpha)$, then the two-step test based on $\mathbb{W}_2$ has a higher local asymptotic power than the one-step test based on $\mathbb{W}_1$ for any $\delta_0 \in \mathfrak{A}(\lambda_0)$.

Table 1
Threshold values $g(b, q)$ for asymptotic variance comparison with the Bartlett kernel.

b	$q = 1$	$q = 2$	$q = 3$	$q = 4$	$q = 5$
0.010	0.007	0.014	0.020	0.027	0.033
0.050	0.034	0.066	0.098	0.128	0.157
0.070	0.047	0.092	0.135	0.175	0.211
0.100	0.068	0.129	0.186	0.238	0.282
0.150	0.101	0.186	0.260	0.321	0.371
0.200	0.130	0.235	0.316	0.380	0.425

$1/b$ indicates the level of smoothing, and q is the degree of overidentification. If the largest (smallest) squared long-run canonical correlation between the two blocks of (rotated and transformed) moment conditions is less (greater) than $g(b, q)$, then the two-step estimator $\hat{\theta}_2$ is asymptotically less (more) efficient than the one-step estimator $\hat{\theta}_1$.

Table 2
Threshold values $f(\lambda; b, d, q, \alpha)$ for power comparison with the Bartlett kernel when $\alpha = 0.05$ and $b = 0.08, 0.16, 0.24, 0.32$.

		$d = 1$			$d = 2$			$d = 3$		
b	λ	$q = 1$	$q = 2$	$q = 3$	$q = 1$	$q = 2$	$q = 3$	$q = 1$	$q = 2$	$q = 3$
0.08	1.000	0.038	0.100	0.193	0.047	0.140	0.132	0.113	0.112	0.083
	9.000	0.020	0.102	0.164	0.082	0.163	0.180	0.094	0.126	0.134
	17.000	0.046	0.102	0.160	0.081	0.146	0.161	0.089	0.113	0.142
	25.000	0.044	0.104	0.155	0.065	0.125	0.188	0.091	0.134	0.163
0.16	1.000	0.070	0.178	0.238	0.213	0.269	0.306	0.165	0.143	0.076
	9.000	0.070	0.169	0.248	0.131	0.208	0.261	0.125	0.153	0.170
	17.000	0.081	0.181	0.246	0.127	0.215	0.274	0.135	0.171	0.197
	25.000	0.083	0.196	0.272	0.124	0.220	0.283	0.135	0.184	0.212
0.24	1.000	0.152	0.272	0.321	0.148	0.282	0.342	0.205	0.320	0.127
	9.000	0.115	0.236	0.302	0.131	0.222	0.259	0.104	0.196	0.145
	17.000	0.122	0.245	0.307	0.128	0.243	0.290	0.106	0.188	0.162
	25.000	0.151	0.268	0.328	0.137	0.245	0.296	0.118	0.194	0.165
0.32	1.000	0.149	0.226	0.349	0.121	0.217	0.285	0.132	0.171	0.094
	9.000	0.124	0.236	0.291	0.131	0.197	0.234	0.111	0.153	0.123
	17.000	0.133	0.253	0.306	0.147	0.216	0.255	0.095	0.145	0.113
	25.000	0.160	0.265	0.318	0.141	0.228	0.268	0.102	0.144	0.130

Notes: If the largest (smallest) squared long-run canonical correlation between the two blocks of (rotated and transformed) moment conditions is smaller (greater) than $f(\lambda; b, d, q, \alpha)$, then the two-step test is asymptotically less (more) powerful.

As in the case of the variance comparison, the conditions on the canonical correlation coefficients in Proposition 5 are both necessary and sufficient. They may appear to be strong, but the conclusions are equally strong—the power comparison results hold regardless of the direction of the local departure. If we have a particular direction in mind so that δ_0 is fixed, then we can evaluate $\delta_0'\Omega_{1\cdot2}^{-1}\delta_0 - \tau \cdot \delta_0'\Omega_{11}^{-1}\delta_0$ directly for the given δ_0. If $\delta_0'\Omega_{1\cdot2}^{-1}\delta_0 - \tau \cdot \delta_0'\Omega_{11}^{-1}\delta_0$ is positive (negative), then the two-step test has a higher (lower) local asymptotic power along that direction.

Proposition 5 goes hand in hand with Proposition 2. The conclusions of these two propositions are qualitatively the same: When the long-run correlation is sufficiently high, we should estimate and exploit it to reduce the variation of the point estimator and improve the power of the associated tests. However, the thresholds are quantitatively different. The two propositions fully characterize the threshold for each criterion under consideration.

3.3. Threshold values

In general, it is not easy to obtain analytical expressions for the threshold functions $g(h, q)$ and $f(\lambda; h, d, q, \alpha)$, but we can easily simulate them.

Table 1 reports the values of $g(b, q)$ for some empirically relevant values of $b = 0.010$–0.20 and $q = 1$–5 when the Bartlett kernel is used.[3] The table shows that the values of $g(b, q)$ are nontrivial in that they are close to neither 0 nor 1. It is clear that $g(b, q)$ is increasing in q and decreasing in the smoothing parameter $h = 1/b$ with $b = S_T/T$. Results presented in the online supplementary appendix show that the same observations apply to the Parzen and QS kernels.

Table 2 reports the values of $f(\lambda; b, d, q, \alpha)$ for $\lambda = 1$–25; $b = 0.08, 0.16, 0.24, 0.32$; and $d, q = 1$–3 when the Bartlett kernel is used. These values are also nontrivial in that they are not close to either of the boundary values 0, 1. As with $g(b, q)$, we find that the threshold value increases as the degree of overidentification increases, and that it decreases as the smoothing parameter $h = 1/b$ increases.

[3] With some abuse of notation, we use $g(h, q)$ and $g(b, q)$ interchangeably with the understanding that $g(b, q)$ is equal to $g(h, q)$ for $h = 1/b$. Similarly, $f(\lambda; b, d, q, \alpha)$ is equal to $f(\lambda; h, d, q, \alpha)$ for $h = 1/b$.

4. General overidentified GMM framework

In this section, we consider the general GMM framework. The parameter of interest is a $d \times 1$ vector $\theta \in \Theta \subseteq \mathbb{R}^d$. Let $v_t \in \mathbb{R}^{d_v}$ denote the vector of observations at time t. We assume that the true value θ_0 is an interior point of the parameter space Θ. The moment conditions

$$E\breve{f}(v_t, \theta) = 0, t = 1, 2, \ldots, T$$

hold if and only if $\theta = \theta_0$, where $\breve{f}(v_t, \cdot)$ is an $m \times 1$ vector of continuously differentiable functions. The process $\breve{f}(v_t, \theta_0)$ may exhibit autocorrelation of unknown forms. We assume that $m \geq d$ and that the rank of $E[\partial \breve{f}(v_t, \theta_0)/\partial\theta']$ is d. That is, we consider a model that is possibly overidentified with degree of overidentification $q = m - d$.

Given a simple positive-definite weighting matrix $\breve{W}_{0T}$ that does not depend on any unknown parameter, we can obtain an initial GMM estimator of θ_0 as

$$\hat{\theta}_o = \arg\min_{\theta\in\Theta} \breve{g}_T(\theta)'\breve{W}_{0T}^{-1}\breve{g}_T(\theta),$$

where $\breve{g}_t(\theta) = T^{-1/2}\sum_{j=1}^{T}\breve{f}(v_j, \theta)$. For example, we may set $\breve{W}_{0T}$ equal to I_m. In the case of IV regression, we may set $\breve{W}_{0T}$ equal to $Z_T'Z_T/T$, where Z_T is the matrix of the instruments. Using the contemporaneous covariance matrix $\breve{\Sigma}$ or the LRV matrix $\breve{\Omega}$ of the moment process $\{\breve{f}(v_t, \theta_0)\}$ to construct the weighting matrix, we obtain the following two (infeasible) GMM estimators:

$$\tilde{\theta}_1 := \arg\min_{\theta\in\Theta} \breve{g}_T(\theta)'\breve{\Sigma}^{-1}\breve{g}_T(\theta) \text{ and } \tilde{\theta}_2 := \arg\min_{\theta\in\Theta} \breve{g}_T(\theta)'\breve{\Omega}^{-1}\breve{g}_T(\theta).$$

For the estimator $\tilde{\theta}_1$, we use the inverse of the contemporaneous covariance matrix as the weighting matrix and ignore all the serial dependency in the moment process $\{\breve{f}(v_t, \theta_0)\}$. In contrast to this procedure, the second estimator $\tilde{\theta}_2$ accounts for the long-run dependency. The feasible versions of these estimators $\hat{\theta}_1$ and $\hat{\theta}_2$ can be naturally defined by replacing $\breve{\Sigma}$ and $\breve{\Omega}$ with their estimates $\breve{\Sigma}_{\text{est}}(\hat{\theta}_o)$ and $\breve{\Omega}_{\text{est}}(\hat{\theta}_o)$ where

$$\begin{aligned}
\breve{\Sigma}_{\text{est}}(\theta) &:= \frac{1}{T}\sum_{t=1}^{T}\left[\breve{f}(v_t, \theta) - \frac{1}{T}\sum_{t=1}^{T}\breve{f}(v_\tau, \theta)\right]\left[\breve{f}(v_t, \theta) - \frac{1}{T}\sum_{t=1}^{T}\breve{f}(v_\tau, \theta)\right]', \\
\breve{\Omega}_{\text{est}}(\theta) &:= \frac{1}{T}\sum_{s=1}^{T}\sum_{t=1}^{T} Q_h^*\left(\frac{s}{T}, \frac{t}{T}\right)\breve{f}(v_t, \theta)\breve{f}(v_s, \theta)'. \qquad (14)
\end{aligned}$$

Suppose that we are interested in certain linear combinations $R\theta_0$ of θ_0 for some $p \times d$ full-rank matrix R. We consider both point estimation of $R\theta_0$ and hypothesis testing about $R\theta_0$. We can make a suitable choice of R according to the problem of interest. For example, if we are interested in the whole vector θ_0, we can choose R to be the $d \times d$ identity matrix. For hypothesis testing, we consider testing the null hypothesis $H_0 : R\theta_0 = r$ against the local alternative $H_1 : R\theta_0 = r + \delta_0/\sqrt{T}$. The usual Wald statistic based on $\hat{\theta}_i$ is

$$\mathbb{W}_i := T(R\hat{\theta}_i - r)'\left\{R\hat{\mathcal{V}}_i R'\right\}^{-1}(R\hat{\theta}_i - r) \qquad (15)$$

for $i = 1, 2$, where

$$\begin{aligned}
\hat{\mathcal{V}}_1 &= \left[\breve{G}_{1T}'\breve{\Sigma}_{\text{est}}^{-1}(\hat{\theta}_1)\breve{G}_{1T}\right]^{-1}\left[\breve{G}_{1T}'\breve{\Sigma}_{\text{est}}^{-1}(\hat{\theta}_1)\breve{\Omega}_{\text{est}}(\hat{\theta}_1)\breve{\Sigma}_{\text{est}}^{-1}(\hat{\theta}_1)\breve{G}_{1T}\right]\left[\breve{G}_{1T}'\breve{\Sigma}_{\text{est}}^{-1}(\hat{\theta}_1)\breve{G}_{1T}\right]^{-1}, \qquad (16) \\
\hat{\mathcal{V}}_2 &= \left[\breve{G}_{2T}\breve{\Omega}_{\text{est}}^{-1}(\hat{\theta}_2)\breve{G}_{2T}\right]^{-1},
\end{aligned}$$

and $\breve{G}_{iT} := \breve{G}_T(\hat{\theta}_i)$ with $\breve{G}_T(\theta) = T^{-1}\sum_{t=1}^{T}\frac{\partial\breve{f}(v_t,\theta)}{\partial\theta'}$. When we are interested in a single restriction, we can construct the usual t statistics. For the asymptotic results on the t-type tests, see Hwang and Sun (2015).

To compare the two-step point estimation and testing to their one-step counterparts, we maintain the assumptions below.

Assumption 4. As $T \to \infty$ for a fixed h, $\hat{\theta}_o = \theta_0 + o_p(1)$, $\hat{\theta}_1 = \theta_0 + o_p(1)$, and $\hat{\theta}_2 = \theta_0 + o_p(1)$ for an interior point $\theta_0 \in \Theta$.

Assumption 5. Define

$$\breve{G}_t(\theta) = \frac{1}{T}\sum_{j=1}^{t}\frac{\partial\breve{f}(v_t, \theta)}{\partial\theta'} \text{ for } t \geq 1, \text{ and } \breve{G}_0(\theta) = 0.$$

For any $\theta_T = \theta_0 + o_p(1)$, the following hold: (*i*) $\text{plim}_{T\to\infty}\breve{G}_{[rT]}(\theta_T) = r\breve{G}$ uniformly in r, where $\breve{G} = \breve{G}(\theta_0)$ and $\breve{G}(\theta) = E\partial\breve{f}(v_t, \theta)/\partial\theta'$; (*ii*) $\breve{\Sigma}_{\text{est}}(\theta_T) \xrightarrow{p} \breve{\Sigma} > 0$; (*iii*) $\breve{\Sigma}$, $\breve{\Omega}$, $\breve{G}'\breve{\Sigma}^{-1}\breve{G}$, and $\breve{G}'\breve{\Omega}^{-1}\breve{G}$ are all nonsingular.

Assumptions 4 and 5 are standard assumptions in the literature on the fixed-smoothing asymptotics. They are the same as those in Kiefer and Vogelsang (2005). With these two assumptions and some mild conditions, the standard GMM theory gives

$$\sqrt{T}(\hat{\theta}_1 - \theta_0) = \frac{1}{\sqrt{T}} \sum_{t=1}^{T} \left[\breve{G}' \breve{\Sigma}^{-1} \breve{G} \right]^{-1} \breve{G}' \breve{\Sigma}^{-1} \breve{f}(v_t, \theta_0) + o_p(1).$$

Under the fixed-smoothing asymptotics, Sun (2014b) establishes the representation

$$\sqrt{T}(\hat{\theta}_2 - \theta_0) = \frac{1}{\sqrt{T}} \sum_{t=1}^{T} \left[\breve{G}' \breve{\Omega}_{\infty}^{-1} \breve{G} \right]^{-1} \breve{G}' \breve{\Omega}_{\infty}^{-1} \breve{f}(v_t, \theta_0) + o_p(1),$$

where $\breve{\Omega}_{\infty}$ is defined similarly to Ω_{∞} in Proposition 1: $\breve{\Omega}_{\infty} = \breve{\Omega}_{1/2} \tilde{\Omega}_{\infty} \breve{\Omega}_{1/2}'$. Each of the above stochastic approximations involves a multiplicative transformation of the original moment process $\{\breve{f}(v_t, \theta_0)\}$.

Because of the complicated structure of the transformed moment processes, it is not straightforward to compare the asymptotic distributions of $\hat{\theta}_1$ and $\hat{\theta}_2$ as in Sections 3.1 and 3.2. To deal with this challenge, we let

$$\breve{G} = \underset{m\times m}{U} \cdot \underset{m\times d}{\Xi} \cdot \underset{d\times d}{V'}$$

be a singular-value decomposition (SVD) of $\breve{G}$, where

$$\Xi' = \left(\underset{d\times d}{A}, \quad \underset{d\times q}{O} \right),$$

A is a $d \times d$ diagonal matrix, and O is a matrix of zeros. Also, we define

$$f^*(v_t, \theta_0) = (f_1^{*\prime}(v_t, \theta_0), f_2^{*\prime}(v_t, \theta_0))' := U' \breve{f}(v_t, \theta_0) \in \mathbb{R}^m,$$

where $f_1^*(v_t, \theta_0) \in \mathbb{R}^d$ and $f_2^*(v_t, \theta_0) \in \mathbb{R}^q$ are the rotated moment conditions. The variance and long-run variance matrices of the rotated moment process $\{f^*(v_t, \theta_0)\}$ are $\Sigma^* := U' \breve{\Sigma} U$ and $\Omega^* := U' \breve{\Omega} U$, respectively. To convert the contemporaneous variance of the rotated moment process to an identity matrix, we define the normalized moment process

$$f(v_t, \theta_0) := (\Sigma_{1/2}^*)^{-1} f^*(v_t, \theta_0) = \left[f_1(v_t, \theta_0)', f_2(v_t, \theta_0)' \right]',$$

where

$$f_1(v_t, \theta_0) := \left(\Sigma_{1\cdot 2}^* \right)^{-1/2} \left[f_1^*(v_t, \theta_0) - \Sigma_{12}^* \left(\Sigma_{22}^* \right)^{-1} f_2^*(v_t, \theta_0) \right] \in \mathbb{R}^d,$$

$$f_2(v_t, \theta_0) := \left(\Sigma_{22}^* \right)^{-1/2} f_2^*(v_t, \theta_0) \in \mathbb{R}^q,$$

and the submatrices Σ_{12}^*, Σ_{22}^*, and $\Sigma_{1\cdot 2}^*$ are defined in the same way as in the previous sections. The long-run variance is $\Omega := (\Sigma_{1/2}^*)^{-1} \Omega^* (\Sigma_{1/2}^{*\prime})^{-1}$. The following display, which also serves as a partial list of notations, summarizes the three versions of the moment processes and their variances and long-run variances.

Original	
moment	$\breve{f}(v_t, \theta_0)$
variance	$\breve{\Sigma}$
lrvar	$\breve{\Omega}$

Rotation $\longrightarrow$

Rotated	
moment	$f^*(v_t, \theta_0) = U' \breve{f}(v_t, \theta_0)$
variance	$\Sigma^* = U' \breve{\Sigma} U$
lrvar	$\Omega^* = U' \breve{\Omega} U$

$\downarrow$Normalization

Rotated and normalized	
moment	$f(v_t, \theta_0) = (\Sigma_{1/2}^*)^{-1} f^*(v_t, \theta_0)$
variance	$\Sigma = I_m$
lrvar	$\Omega = (\Sigma_{1/2}^*)^{-1} \Omega^* (\Sigma_{1/2}^{*\prime})^{-1}$

Lemma 6. *Let Assumptions 1–5 hold with u_t replaced by $f(v_t, \theta_0)$ in Assumptions 2 and 3. Then as $T \to \infty$ for a fixed $h > 0$,*

$$\left(\Sigma_{1\cdot 2}^* \right)^{-1/2} A V' \sqrt{T}(\hat{\theta}_1 - \theta_0) = \frac{1}{\sqrt{T}} \sum_{t=1}^{T} f_1(v_t, \theta_0) + o_p(1) \xrightarrow{d} N(0, \Omega_{11}), \tag{17}$$

$$\begin{aligned} \left(\Sigma_{1\cdot 2}^* \right)^{-1/2} A V' \sqrt{T}(\hat{\theta}_2 - \theta_0) &= \frac{1}{\sqrt{T}} \sum_{t=1}^{T} [f_1(v_t, \theta_0) - \beta_{\infty} f_2(v_t, \theta_0)] + o_p(1) \\ &\xrightarrow{d} MN\left(0, \Omega_{11} - \Omega_{12}\beta_{\infty}' - \beta_{\infty}\Omega_{21} + \beta_{\infty}\Omega_{22}\beta_{\infty}'\right), \end{aligned} \tag{18}$$

where $\beta_{\infty} := \Omega_{\infty,12} \Omega_{\infty,22}^{-1}$ and Ω_{∞} are defined in the same way as in Proposition 1.

Lemma 6 casts the stochastic expansions of the estimators $\hat{\theta}_1$ and $\hat{\theta}_2$ in the same form. To the best of our knowledge, these representations are new in the econometrics literature and may be of independent interest. Lemma 6 enables us to directly compare the asymptotic properties of the one-step and two-step estimators and the associated tests.

If we let $(u_{1t}, u_{2t}) = (f_1(v_t, \theta_0), f_2(v_t, \theta_0))$, then the right-hand sides of (17) and (18) are exactly the same as what we would obtain in the location model. The location model, simple though it may be, has implications for general settings from an asymptotic point of view. In particular, if we define

$$y_{1t} = \left(\Sigma^*_{1\cdot 2}\right)^{-1/2} AV'\theta_0 + u_{1t},$$
$$y_{2t} = u_{2t},$$

where $u_{1t} = f_1(v_t, \theta_0)$ and $u_{2t} = f_2(v_t, \theta_0)$, then the estimation and inference problems in the GMM setting are asymptotically equivalent to those in the above simple location model with $\{y_{1t}, y_{2t}\}_{t=1}^T$ as the observations.

To present our next theorem, we transform R into $\tilde{R} = RVA^{-1}\left(\Sigma^*_{1\cdot 2}\right)^{1/2}$, and we define $\tilde{\beta}_\infty(h, p, q)$ as $\tilde{\beta}_\infty(h, d, q)$ is defined in (8) but with d replaced by p. Let

$$\rho_R = (\tilde{R}\Omega_{11}\tilde{R}')^{-1/2}(\tilde{R}\Omega_{12})\Omega_{22}^{-1/2} \in \mathbb{R}^{p\times q}$$

be the long-run correlation matrix between $\tilde{R}f_1(v_t, \theta_0)$ and $f_2(v_t, \theta_0)$.

Theorem 7. *Let the assumptions in Lemma 6 hold, and define*

$$\mathfrak{A}(\lambda_0) = \{\delta : \delta'[R(\breve{G}'\breve{\Omega}^{-1}\breve{G})^{-1}R']^{-1}\delta = \lambda_0\} \text{ for some } \lambda_0 > 0.$$

Consider the local alternative $H_1(\lambda_0): R\theta_0 = r + \delta_0/\sqrt{T}$ for $\delta_0 \in \mathfrak{A}(\lambda_0)$ and the fixed-smoothing asymptotics.

(a) If $\nu_{\max}\left(\rho_R\rho_R'\right) < g(h, q)$, then $R\hat{\theta}_2$ has a larger asymptotic variance than $R\hat{\theta}_1$.

(b) If $\nu_{\min}\left(\rho_R\rho_R'\right) > g(h, q)$, then $R\hat{\theta}_2$ has a smaller asymptotic variance than $R\hat{\theta}_1$.

(c) If $\nu_{\max}\left(\rho_R\rho_R'\right) < f(\lambda_0; h, p, q, \alpha)$, then the two-step test is asymptotically less powerful than the first-step test for any $\delta_0 \in \mathfrak{A}(\lambda_0)$.

(d) If $\nu_{\min}\left(\rho_R\rho_R'\right) > f(\lambda_0; h, p, q, \alpha)$, then the two-step test is asymptotically more powerful than the first-step test for any $\delta_0 \in \mathfrak{A}(\lambda_0)$.

The threshold functions $g(h, q)$ and $f(\lambda_0; h, p, q, \alpha)$ are the same as those that we obtain for the location model. In fact, Propositions 2 and 5 are special cases of Theorem 7.

When R is the $d \times d$ identity matrix, we have $\nu(\rho_R\rho_R') = \nu(\rho\rho')$ for $\rho = \Omega_{11}^{-1/2}\Omega_{12}\Omega_{22}^{-1/2}$. To evaluate the asymptotic efficiency of $\hat{\theta}_2$ relative to that of $\hat{\theta}_1$, we check whether $\nu_{\max}\left(\rho\rho'\right) \le g(h, q)$ or $\nu_{\min}\left(\rho\rho'\right) > g(h, q)$ holds. These two conditions are not mutually exclusive unless $d = 1$, that is, θ_0 is a scalar. When $d > 1$, it is possible that neither of these two conditions is satisfied, in which case $\text{avar}(\hat{\theta}_2) - \text{avar}(\hat{\theta}_1)$ is indefinite. Thus, as a whole vector, the asymptotic efficiency of $\hat{\theta}_2$ relative to that of $\hat{\theta}_1$ cannot be determined. However, we can show that when $d > 1$, there exist matrices $R^+ \in \mathbb{R}^{d_+\times d}$ and $R^- \in \mathbb{R}^{d_-\times d}$ with $d_+ + d_- = d$, $d_+ < d$, and $d_- < d$ such that $\text{avar}(R^+\hat{\theta}_2) \le \text{avar}(R^+\hat{\theta}_1)$ and $\text{avar}(R^-\hat{\theta}_2) \ge \text{avar}(R^-\hat{\theta}_1)$. An example of the indefinite case is when $q < d$ and $\nu_{\max}\left(\rho\rho'\right) > g(h, q)$. In this case, $\nu_{\min}\left(\rho\rho'\right) = 0$ and so $\nu_{\min}\left(\rho\rho'\right) < g(h, q)$. Therefore, neither $\nu_{\max}\left(\rho\rho'\right) < g(h, q)$ nor $\nu_{\min}\left(\rho\rho'\right) > g(h, q)$ holds. The fact that $\nu_{\min}\left(\rho\rho'\right) = 0$ implies that $\text{avar}(R^-\hat{\theta}_2) > \text{avar}(R^-\hat{\theta}_1)$ for some R^-. Therefore, when the degree of overidentification is not sufficiently large, there are some directions characterized by R^- along which the two-step estimator is asymptotically less efficient than the one-step estimator.

When $d = 1$, $\rho\rho'$ is a scalar, and the conditions $\nu_{\max}\left(\rho\rho'\right) \le g(h, q)$ and $\nu_{\min}\left(\rho\rho'\right) > g(h, q)$ become mutually exclusive. Thus, if $\rho\rho' > g(h, q)$, then $\hat{\theta}_2$ is asymptotically more efficient than $\hat{\theta}_1$. If $\rho\rho' < g(h, q)$, then $\hat{\theta}_2$ is asymptotically less efficient.

Similarly, the conditions $\nu_{\max}\left(\rho_R\rho_R'\right) \le f(\lambda_0; h, p, q, \alpha)$ and $\nu_{\min}\left(\rho_R\rho_R'\right) > f(\lambda_0; h, p, q, \alpha)$ are not mutually exclusive unless $p = 1$. When $p = 1$, which is of ultimate importance in empirical studies, $\rho_R\rho_R'$ is equal to the sum of the squared long-run canonical correlation coefficients between $\tilde{R}f_1(v_t, \theta_0)$ and $f_2(v_t, \theta_0)$. In this case, $f(\lambda_0; h, p, q, \alpha)$ is the threshold value of $\rho_R\rho_R'$ for assessing the relative efficiency of the two tests. In particular, when $\rho_R\rho_R' > f(\lambda_0; h, p, q, \alpha)$, the two-step test is more powerful than the one-step test. Otherwise, the two-step test is not more powerful.

Remark 8. A recent paper by Zhang (2016) develops the fixed-smoothing asymptotic theory for the continuously updating GMM estimator (CU-GMM) defined by

$$\hat{\theta}_{\text{CU}} = \arg\min_{\theta\in\Theta} \breve{g}_T(\theta)'\left[\breve{\Omega}_{\text{est}}(\theta)\right]^{-1}\breve{g}_T(\theta),$$

where $\breve{\Omega}_{\text{est}}(\theta) = T^{-1}\sum_{s=1}^T\sum_{t=1}^T Q_h^*(s/T, t/T)\breve{f}(v_t, \theta)\breve{f}(v_s, \theta)'$. After imposing some additional conditions, Zhang (2016, Theorem 2.1) shows that

$$\sqrt{T}(\hat{\theta}_{\text{CU}} - \theta_0) = \frac{1}{\sqrt{T}}\sum_{t=1}^T\left[\breve{G}'\breve{\Omega}_\infty^{-1}\breve{G}\right]^{-1}\breve{G}'\breve{\Omega}_\infty^{-1}\breve{f}(v_t, \theta_0) + o_p(1).$$

Table 3
Comparison of the finite-sample variances for the estimators $\hat{\theta}_1$ and $\hat{\theta}_2$ with $T = 200$ and $q = 4$.

$q = 4$				
		var($\hat{\theta}_2$)		
$\nu_{\max}(\rho\rho')$	var($\hat{\theta}_1$)	Bartlett	Parzen	QS
		$b = .07$	$b = .15$	$b = .07$
0.000	0.082	0.103	0.117	0.112
0.090	0.092	0.105	0.118	0.114
0.180	0.104	0.107	0.120	0.116
0.270	0.120	0.111	0.123	0.119
0.360	0.141	0.115	0.126	0.122
0.450	0.168	0.120	0.130	0.127
0.540	0.207	0.128	0.136	0.134
0.630	0.264	0.140	0.146	0.144
0.720	0.359	0.159	0.161	0.161
0.810	0.543	0.196	0.190	0.194
0.900	1.061	0.301	0.273	0.286

That is, the CU-GMM estimator shares the same first-order fixed-smoothing limit with the two-step GMM estimator in this paper. By the first-order asymptotic equivalence, Theorem 7 still holds if $\hat{\theta}_2$ is replaced by $\hat{\theta}_{\text{CU}}$.

5. Simulation evidence

This section evaluates whether the theoretical comparison of the one-step and two-step procedures using the fixed-smoothing asymptotics is trustworthy in finite samples. We consider the simple location model given in (1), but we allow a general form of var $(u_t^*) = \Sigma^*$, that is, we do not assume that Σ^* is a scalar matrix $\sigma^2 I_m$. Without loss of generality, we set the true parameter to be $\theta_0 = (0, \ldots, 0)' \in \mathbb{R}^d$. The error $\{u_t^*\}$ follows a VAR(1) process:

$$u_{1t}^{*i} = \psi u_{1t-1}^{*i} + \frac{\gamma}{\sqrt{q}} \sum_{j=1}^{q} u_{2t-1}^{*j} + e_{1t}^{*i}, \ i = 1, \ldots, d, \tag{19}$$
$$u_{2t}^{*i} = \psi u_{2t-1}^{*i} + e_{2t}^{*i}, \ i = 1, \ldots, q,$$

where $e_t^* = (e_{1t}^{*\prime}, e_{2t}^{*\prime})' \sim iid\ N\ (0, I_m)$. The notations used above are consistent with those in the GMM setting. For simulation evidence from other data-generating processes, see Hwang and Sun (2015).

Note that no rotation is needed in the location model, as the implied rotation matrix, i.e., U, is an identity matrix. We need only to normalize the error process by letting $u_{1t} = \left(\Sigma_{\cdot 2}^*\right)^{-1/2}[u_{1t}^* - \Sigma_{12}^*\left(\Sigma_{22}^*\right)^{-1} u_{2t}^*]$ and $u_{2t} = \left(\Sigma_{22}^*\right)^{-1/2} u_{2t}^*$. Let ρ be the long-run correlation matrix between u_{1t} and u_{2t}. With some algebraic manipulation, we have

$$\rho\rho' = \left(d + \frac{(1-\psi^2)^2}{\gamma^2}\right)^{-1} J_{d,d}, \tag{20}$$

where $J_{d,d}$ is the $d \times d$ matrix of ones. Thus the maximum eigenvalue of the squared long-run correlation $\rho\rho'$ is given by $\nu_{\max}(\rho\rho') = \left[1 + (1-\psi^2)^2/(d\gamma^2)\right]^{-1}$, which is also the only nonzero eigenvalue.

We consider three commonly used kernels for the LRV estimation: Bartlett, Parzen, and QS kernels. We select the smoothing parameter b according to the AMSE-optimal formula in Andrews (1991). The unknown parameters in the AMSE are either calibrated to a VAR(1) process with a given AR parameter ψ or data-driven using the VAR(1) plug-in approach. The qualitative results remain the same regardless of how the unknown parameters are obtained. In all our simulations, the sample size T is 200, and the number of simulation replications is 10,000.

We first consider point estimation with $d = 1$, under which $\rho\rho'$ is a scalar and $\nu_{\max}(\rho\rho') = \rho\rho'$. Note that $\nu_{\max}(\rho\rho')$ is increasing in γ^2 for a given ψ. We fix the value of ψ at 0.75, so each moment process is reasonably persistent. For this value of ψ, we consider $\nu_{\max}(\rho\rho') = 0, 0.09, 0.18, \ldots, 0.90$, which we obtain by setting γ to appropriate values using (20).

According to Proposition 2, if $\rho\rho'$ is greater than a threshold value, then var($\hat{\theta}_2$) is expected to be smaller than var($\hat{\theta}_1$); otherwise, var($\hat{\theta}_2$) is expected to be larger. The simulated values of var($\hat{\theta}_1$) and var($\hat{\theta}_2$) when $q = 4$ are given in Table 3 for certain values of b. These values of b are calibrated by using the AMSE optimal formulae under the VAR(1) design with $\psi = 0.75$ and $\gamma^2 = \left(\rho\rho'(1-\psi^2)^2\right)/\left(d(1-\rho\rho')\right)$ for $\rho\rho' = 0.40$ and $d = 1$. We discuss only the case with the Bartlett kernel, as we obtain the same qualitative results with the other kernels. The results indicate that var $(\hat{\theta}_2)$ becomes smaller than var($\hat{\theta}_1$) only when $\rho\rho'$ is sufficiently large. For example, when there is no long-run correlation, that is, $\rho\rho' = 0$, we have var($\hat{\theta}_1$) $= 0.082 <$ var($\hat{\theta}_2$) $= 0.103$, and so $\hat{\theta}_1$ is more efficient than $\hat{\theta}_2$ with 22% efficiency gain. The simulation results are consistent with our theoretical results in Proposition 2: $\hat{\theta}_2$ becomes more efficient relative to $\hat{\theta}_1$ only when the benefit

Table 4
Empirical sizes of the one-step and two-step tests based on the Bartlett kernel under VAR(1) errors when $\psi = 0.75$, $p = 1$–2, and $T = 200$..

	One-step test		Two-step test	
$\nu_{\max}(\rho_R\rho_R')$	χ^2	$\mathbb{W}_{1\infty}$	χ^2	$\mathbb{W}_{2\infty}$
$(p, q) = (1, 3)$				
0.000	0.161	0.142	0.205	0.138
0.350	0.152	0.129	0.181	0.099
0.500	0.151	0.125	0.162	0.080
0.600	0.150	0.124	0.144	0.068
0.750	0.149	0.123	0.111	0.049
$(p, q) = (2, 3)$				
0.000	0.237	0.183	0.307	0.173
0.350	0.232	0.167	0.294	0.142
0.500	0.231	0.164	0.283	0.130
0.600	0.231	0.162	0.272	0.121
0.750	0.231	0.159	0.252	0.106

Notes: The "one-step test" is based on the first-step GMM estimator using the inverse of the contemporaneous variance estimator $\hat{\Sigma}^*$ as the weighting matrix; the "two-step test" is based on the two-step GMM estimator using the inverse of the data-driven kernel LRV estimator $\hat{\Omega}^*$ as the weighting matrix.

of using the LRV matrix outweighs the cost of estimating it. When $b = 0.07$ and $q = 4$, Table 3 shows that var $(\hat{\theta}_2)$ starts to become smaller than var$(\hat{\theta}_1)$ when $\rho\rho'$ crosses a value around 0.180 from below. This agrees with the theoretical threshold value 0.175 given in Table 1.

We also investigate the simulated finite-sample biases for the point estimators $\hat{\theta}_1$ and $\hat{\theta}_2$. To save space, we do not present the results here, but they can be summarized as follows. First, we find that the finite-sample biases are not as significant as the finite-sample variances in Table 3. This is because the AMSE optimal rule in Andrews (1991) selects the truncation lag parameter $S_T = bT$ to be reasonably large in all the DGPs we considered, so we are able to avoid the over-smoothing problem. We also find that an alternative bandwidth selection rule in Wilhelm (2015), which directly minimizes the (higher-order) AMSE of the two-step estimator, can lead to improvements in the finite-sample biases compared to the bandwidth selection rule given in Andrews (1991), especially when the AR(1) parameter ψ gets close to one and the degree of overidentification q increases.

Next, we implement the two sets of testing procedures: the one-step tests based on $\mathbb{W}_1$ and the two-step tests based on $\mathbb{W}_2$. The nominal significance level is $\alpha = 0.05$. As before, $\psi = 0.75$. We use (20) and numerical methods to set γ and obtain $\nu_{\max}(\rho_R\rho_R') \in \{0.00, 0.35, 0.50, 0.60, 0.75\}$. We focus on the case with $d = 3$ and $q = 3$. The null hypotheses of interest are as follows:

$$H_{01} : \theta_1 = 0;\ H_{02} : \theta_1 = \theta_2 = 0.$$

The corresponding numbers of joint restrictions are $p = 1$ and $p = 2$, respectively. We report only the results for $p = 1$, as the main results for $p = 2$ are similar. For the smoothing parameters, we employ the data-driven AMSE optimal bandwidth through the VAR(1) plug-in implementation developed by Andrews (1991).[4] The empirical sizes of the nominal 5% testing procedures based on the two types of asymptotic approximations are given in Table 4. It is clear that both the one-step test and the two-step test suffer from severe size distortions if the conventional chi-square critical values are used. The relatively large size distortion of the two-step test comes from the additional cost of estimating the weighting matrix. However, if the nonstandard critical values $\mathbb{W}^{\alpha}_{1\infty}$ and $\mathbb{W}^{\alpha}_{2\infty}$ are used, the size distortions of both tests are substantially reduced. This result agrees with the existing literature, such as Sun (2013, 2014a, b) and Kiefer and Vogelsang (2005), which highlights the higher accuracy of the fixed-smoothing approximations. Also, we observe that when the fixed-smoothing approximations are used, in most cases the nonstandard test based on $\mathbb{W}_1$ is more size-distorted than that based on $\mathbb{W}_2$. Results for the other types of kernels are given in the online supplemental appendix.

Next, we investigate the finite-sample power performances of the two procedures. We use the finite-sample critical values under the null hypothesis, so the power is size-adjusted and the power comparison is meaningful. The DGPs are the same as before except that the parameters are from the local null alternative $R\theta_0 = r + \delta_0/\sqrt{T}$. The degree of overidentification considered here is $q = 3$. Also, the domain of each power curve is rescaled to be $\lambda := \delta_0'(\tilde{R}\Omega_{1\cdot2}\tilde{R}')^{-1}\delta_0$ with $\tilde{R} = RVA^{-1}(\Sigma^*_{1\cdot2})^{1/2}$ as in Section 4. Note that for the location model, $V = A = I_d$, and so $\tilde{R} = R(\Sigma^*_{1\cdot2})^{1/2}$.

Fig. 1 presents the size-adjusted finite sample powers of the two testing procedures when $p = 1$ and the Bartlett kernel is used. The power curve of the two-step test shifts upward as the degree of the long-run correlation $\nu_{\max}(\rho_R\rho_R')$ increases, and it starts to dominate that of the one-step test at a certain value of $\nu_{\max}(\rho_R\rho_R') \in (0, 1)$. This is consistent with Proposition 5.

[4] We could use the rule proposed by Wilhelm (2015). Neither the rule proposed by Wilhelm (2015) nor the one proposed by Andrews (1991) is oriented toward testing problems.

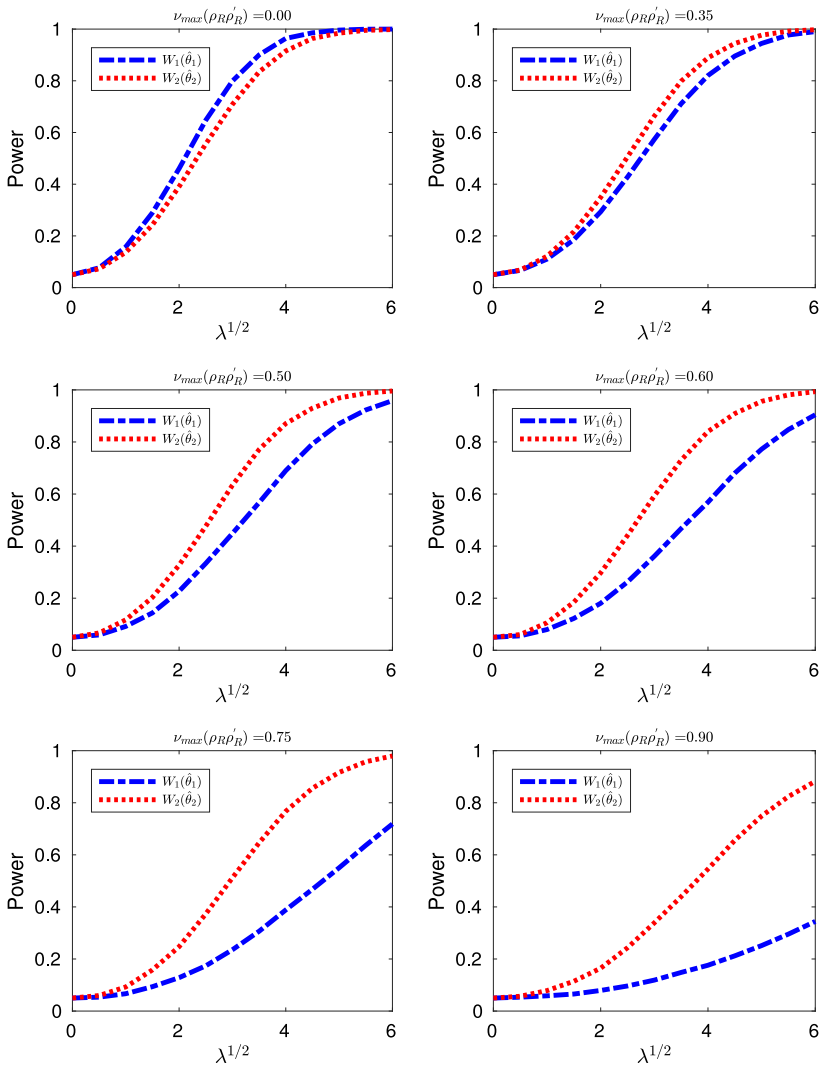


Fig. 1. Size-adjusted powers of the two tests based on the Bartlett kernel under VAR(1) errors with $p = 1$, $q = 3$, $\psi = 0.75$, $T = 200$, and $b = 0.078$.

Fig. 1 shows when $b = 0.078$ and $p = 1$, the power curve of the two-step test starts to dominate that of the one-step test when $\nu_{\max}(\rho_R\rho_R')$ reaches 0.35. This matches our theoretical results in Proposition 5 and Table 2. The table indicates that the threshold value $\max_{\lambda\in[1,25]} f(\lambda; b, p, q, \alpha)$ is about 0.193 when $b = 0.08$, $p = 1$, and $q = 3$. Also, if $\nu_{\max}(\rho_R\rho_R')$ is as high as 0.75, the two-step test is clearly more powerful than the one-step test.

6. Practical recommendation

Both the theoretical results and the simulation evidence suggest that we should go one more step and employ the two-step estimator and test when the long-run canonical correlation coefficients are sufficiently large. In empirical applications, we often care about only a single linear combination of model parameters or a single model parameter. In this case, there is only one long-run canonical correlation coefficient, and it provides the necessary and sufficient condition for going the extra step. However, it is hard to estimate the long-run canonical correlation coefficient with good precision. This is exactly the source of the question as to why the two-step estimator and test may not outperform. In the absence of any prior knowledge of the long-run canonical correlation, we propose using the two-step estimator and test only when the estimated long-run canonical correlation coefficient is larger than our theoretical thresholds. When the estimated long-run canonical correlation coefficient is smaller than our theoretical thresholds, we suggest sticking with the first-step estimator and test.

When we are interested in multiple parameters or restrictions, it is possible that the one-step and two-step procedures cannot be unambiguously ranked. In this case, we propose using the GMM estimator and test based on a working weighting

matrix. Specifically, we fit a parametric VAR(1) model to the estimated moment process $\breve{f}(v_t, \hat{\theta}_o)$ by OLS and treat the estimated model as if it is the true model. We compute the implied long-run variance matrix $\breve{W}_T(\hat{\theta}_o)$ and use its inverse as the weighting matrix in the GMM criterion. This gives rise to the alternative GMM estimator

$$\hat{\theta}_a = \arg\min_{\theta\in\Theta} \breve{g}_T(\theta)' \left[\breve{W}_T(\hat{\theta}_o)\right]^{-1} \breve{g}_T(\theta) \tag{21}$$

and the test statistic

$$\mathbb{W}_a := T(R\hat{\theta}_a - r)' \left\{R\hat{\mathcal{V}}_a R'\right\}^{-1} (R\hat{\theta}_a - r), \tag{22}$$

where, for $\breve{G}_{aT} = \frac{1}{T}\sum_{t=1}^{T} \partial \breve{f}(v_t, \theta)/\partial\theta'\big|_{\theta=\hat{\theta}_a}$, $\hat{\mathcal{V}}_a$ is given by

$$\hat{\mathcal{V}}_a = \left[\breve{G}_{aT}'\breve{W}_T^{-1}(\hat{\theta}_a)\breve{G}_{aT}\right]^{-1} \left[\breve{G}_{aT}'\breve{W}_T^{-1}(\hat{\theta}_a)\breve{\Omega}_{\text{est}}(\hat{\theta}_a)\breve{W}_T^{-1}(\hat{\theta}_a)\breve{G}_{aT}\right] \left[\breve{G}_{aT}'\breve{W}_T^{-1}(\hat{\theta}_a)\breve{G}_{aT}\right]^{-1}.$$

The reader is referred to the online appendix for further details on the working GMM estimator and test statistic based on the VAR(1) plug-in weighting matrix.

With the introduction of the working GMM estimator, we can formalize our recommendations starting with the point estimation. Given the set of moment conditions $E\breve{f}(v_t, \theta_0) = 0$ and the data $\{v_t\}$, we follow the steps below to design our recommended estimator $\hat{\theta}_r$.

1. Compute the initial estimator $\hat{\theta}_o = \arg\min_{\theta\in\Theta} \left\|\sum_{t=1}^{T} \breve{f}(v_t, \theta)\right\|^2$.
2. On the basis of $\hat{\theta}_o$, use a data-driven method such as that given in Andrews (1991) or Wilhelm (2015) to select the smoothing parameter. Denote the data-driven value by $\hat{b}$, and let $\hat{h} = 1/\hat{b}$.
3. Based on the smoothing parameter $\hat{b}$, compute $\breve{\Sigma}_{\text{est}}(\hat{\theta}_o)$ and $\breve{\Omega}_{\text{est}}(\hat{\theta}_o)$ using the formulae in (14).
4. Compute $\breve{G}_T(\hat{\theta}_o) = \frac{1}{T}\sum_{t=1}^{T} \frac{\partial \breve{f}(v_t,\theta)}{\partial\theta'}|_{\theta=\hat{\theta}_o}$ and its singular-value decomposition $\hat{U}\hat{\Xi}\hat{V}'$, where $\hat{\Xi}' = (\hat{A}_{d\times d}, O_{d\times q})$ and $\hat{A}_{d\times d}$ is diagonal.
5. Estimate the variance and the long-run variance of the rotated moment processes by
$$\hat{\Sigma}^* := \hat{U}'\breve{\Sigma}_{\text{est}}(\hat{\theta}_o)\hat{U} \text{ and } \hat{\Omega}^* := \hat{U}'\breve{\Omega}_{\text{est}}(\hat{\theta}_o)\hat{U}.$$
6. Compute the normalized LRV estimator
$$\hat{\Omega} = (\hat{\Sigma}^*_{1/2})^{-1}\hat{\Omega}^*(\hat{\Sigma}^{*\prime}_{1/2})^{-1} := \begin{pmatrix} \hat{\Omega}_{11} & \hat{\Omega}_{12} \\ \hat{\Omega}_{21} & \hat{\Omega}_{22} \end{pmatrix},$$
where
$$\hat{\Sigma}^*_{1/2} = \begin{pmatrix} \left(\hat{\Sigma}^*_{1\cdot 2}\right)^{1/2} & \hat{\Sigma}^*_{12}\left(\hat{\Sigma}^*_{22}\right)^{-1/2} \\ 0 & \left(\hat{\Sigma}^*_{22}\right)^{1/2} \end{pmatrix}.$$
7. Compute the eigenvalues
$$\nu\left(\hat{\rho}\hat{\rho}'\right) = \nu(\hat{\Omega}_{12}\hat{\Omega}_{22}^{-1}\hat{\Omega}_{21}\hat{\Omega}_{11}^{-1}).$$
Let $\nu_{\max}\left(\hat{\rho}\hat{\rho}'\right)$ and $\nu_{\min}\left(\hat{\rho}\hat{\rho}'\right)$ be the largest and smallest eigenvalues, respectively.
8. (a) If $\nu_{\min}\left(\hat{\rho}\hat{\rho}'\right) > g(\hat{h}, q)$, then set $\hat{\theta}_r = \hat{\theta}_2$.
(b) If $\nu_{\max}\left(\hat{\rho}\hat{\rho}'\right) < g(\hat{h}, q)$, then set $\hat{\theta}_r = \hat{\theta}_1$.
(c) If neither condition (a) nor condition (b) is satisfied, then set $\hat{\theta}_r = \hat{\theta}_a$.

For hypothesis testing with the null hypothesis $H_0 : R\theta_0 = r$ and the alternative $H_1 : R\theta_0 \neq r$ for some $R \in \mathbb{R}^{p\times d}$, we first follow steps 1 through 6 for point estimation, and then continue with the steps below:

7. Let $\tilde{R}_{\text{est}} = R\hat{V}\hat{A}^{-1}(\hat{\Sigma}^*_{1\cdot 2})^{1/2}$. Compute the eigenvalues
$$\nu\left(\hat{\rho}_R\hat{\rho}_R'\right) = \nu\left[(\tilde{R}_{\text{est}}\hat{\Omega}_{12}\hat{\Omega}_{22}^{-1}\hat{\Omega}_{21}\tilde{R}_{\text{est}}')(\tilde{R}_{\text{est}}\hat{\Omega}_{11}\tilde{R}_{\text{est}}')^{-1}\right].$$
Let $\nu_{\max}\left(\hat{\rho}_R\hat{\rho}_R'\right)$ and $\nu_{\min}\left(\hat{\rho}_R\hat{\rho}_R'\right)$ be the largest and smallest eigenvalues, respectively.
8. Choose the value of λ^o such that $P\left(\chi_p^2(\lambda^o) > \chi_p^\alpha\right) = 75\%$, where χ_p^α is the $(1-\alpha)$ quantile of the chi-square distribution of p degrees of freedom.
9. (a) If $\nu_{\min}\left(\hat{\rho}_R\hat{\rho}_R'\right) > f(\lambda^o; \hat{h}, p, q, \alpha)$, employ the statistic $\mathbb{W}_2$ and the critical value $\mathbb{W}^\alpha_{2\infty}$.
(b) If $\nu_{\max}\left(\hat{\rho}_R\hat{\rho}_R'\right) < f(\lambda^o; \hat{h}, p, q, \alpha)$, employ the statistic $\mathbb{W}_1$ and the critical value $\mathbb{W}^\alpha_{1\infty}$.
(c) If neither condition (a) nor condition (b) is satisfied, employ the statistic $\mathbb{W}_a$ and the critical value $\mathbb{W}^\alpha_{1\infty}$.

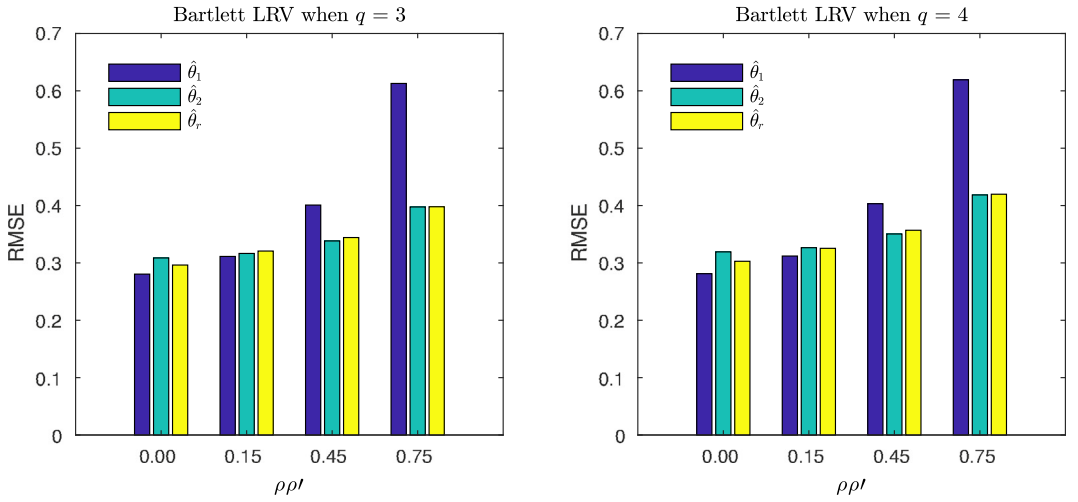


Fig. 2. Comparison of the finite-sample RMSE performances of the one-step estimator ($\hat{\theta}_1$), the two-step estimator ($\hat{\theta}_2$), and the rule-of-thumb estimator ($\hat{\theta}_{\mathrm{r}}$).

We use $\mathbb{W}_{\mathrm{r}}$ to refer to the above rule-of-thumb testing procedure with nonstandard critical values.

In the step 8 above, the choice of the local alternative parameter λ^o is consistent with the optimal testing literature. See, for example, Sun et al. (2008). We may also choose a value of λ^o that reflects a particular scientific interest or economic significance.

Under the increasing-smoothing asymptotics, the null rejection probability of our rule-of-thumb test is equal to the nominal rejection probability as $T \to \infty$. To see this, note that the data-driven h is given by $\hat{h} = \hat{c}T^{\alpha}$, where $\alpha > 0$ is a known parameter that characterizes the MSE-optimal rate for h and $\hat{c}$ is the data-driven estimator based on a parametric plug-in approach. As $T \to \infty$, we have $\hat{c} \xrightarrow{p} c^*$ for some $c^* \neq 0$, and so $\hat{h}$ diverges to infinity with probability approaching one. On the other hand, we can show that as $h \to \infty$, the critical values $\mathbb{W}^{\alpha}_{1\infty}$ and $\mathbb{W}^{\alpha}_{2\infty}$ approach the same chi-squared critical value χ^{α}_p. There is no need to adjust the critical value for each of the three tests in order to account for the selection uncertainty in selecting the test we eventually use. Thus, from the perspective of the increasing-smoothing asymptotics, our rule-of-thumb test has the correct pointwise control of the null rejection probability. Recently, Preinerstorfer and Pötscher (2016) have shown that the existing HAR tests do not have a uniform size control. Our rule-of-thumb test is no exception. See Preinerstorfer (2015) for further discussion of this.

We evaluate the performances of the above estimation and testing procedures relative to those of the one-step and two-step procedures via Monte Carlo simulations. The root mean squared errors (RMSE) of the estimators $\hat{\theta}_1$, $\hat{\theta}_2$, and $\hat{\theta}_{\mathrm{r}}$ are given in Fig. 2 for $q = 3$ and 4. The DGP is the same as that given in Section 5. To save space, we present only the results for the case with the Bartlett kernel. Fig. 2 shows that the RMSE of $\hat{\theta}_2$ becomes smaller than that of $\hat{\theta}_1$ only when $\rho\rho'$ is sufficiently large. This is similar to the simulation evidence in Section 5. However, Table 3 in Section 5 employs fixed values of b for each simulation replication, while Fig. 2 employs data-driven values of b, which could be different for different simulation replications. More importantly, Fig. 2 shows that the rule-of-thumb estimator $\hat{\theta}_{\mathrm{r}}$ has a very competitive RMSE performance relative to $\hat{\theta}_1$ and $\hat{\theta}_2$.

Next, we investigate the testing procedures, $\mathbb{W}_1$, $\mathbb{W}_2$, and $\mathbb{W}_{\mathrm{r}}$ for $p = 1$ and $q = 3$ and 4. Fig. 3 shows the size-adjusted finite sample power against the local alternative $H_1 : R\theta_0 = r + \delta_0/\sqrt{T}$ for $\delta_0'(\tilde{R}\Omega_{1\cdot 2}\tilde{R}')^{-1}\delta_0 = \lambda_0$ with $\tilde{R} = R(\Sigma^*_{1\cdot 2})^{1/2}$ and $R = (1, 0, 0)$. The parameter configuration is the same as in the point estimation case. It is clear that the power of the two-step test $\mathbb{W}_2$ increases with the long-run correlation $\nu_{\max}(\rho_R\rho_R')$ and eventually dominates that of the one-step test $\mathbb{W}_1$. This result is also similar to the simulation results in Section 5 with fixed values of smoothing parameter b. Fig. 3 shows that $\mathbb{W}_{\mathrm{r}}$ succeeds in selecting the more powerful test.

In Figs. 2 and 3, we present results for only a single parameter or restriction. By construction, the plug-in working GMM estimator and test are never used. To improve the finite-sample performance, we can modify our practical recommendations and use $\hat{\theta}_{\mathrm{r}} = \hat{\theta}_a$ (or $\mathbb{W}_{\mathrm{r}} = \mathbb{W}_a$) as long as $\hat{\rho}$ (or $\hat{\rho}_R$) is within 10% of the theoretical threshold. This alternative recommendation, however, has no rigorous theoretical justification, as the percentage margin has to be chosen by the individual practitioner. The online supplementary appendix contains some simulation evidence supporting the modified rule-of-thumb procedure.

7. Conclusion

This paper makes more accurate and more reliable comparisons of the popular one-step and two-step GMM estimators and the associated inference procedures. We transform a general GMM into an asymptotically equivalent overidentified

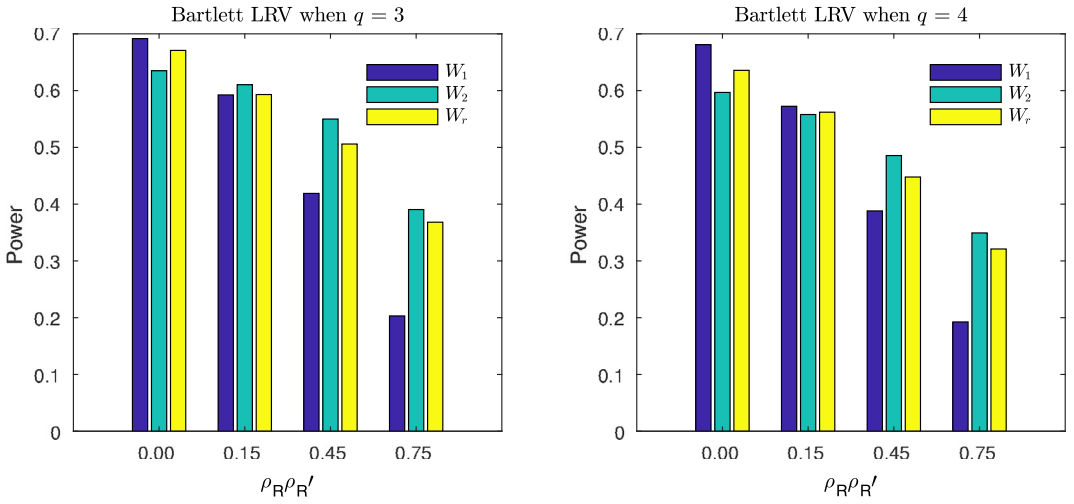


Fig. 3. Comparison of the finite-sample size-adjusted powers of the one-step test ($\mathbb{W}_1$), the two-step test ($\mathbb{W}_2$), and the rule-of-thumb test ($\mathbb{W}_r$) with $p = 1$.

location model and deliver a clear message on when we should go one step further and use a two-step procedure: We should employ a two-step GMM procedure only when the benefit of doing so clearly outweighs the cost. The magnitude of the benefit is quantified by the long-run canonical correlation coefficients between two blocks of moment conditions. In the simple location model, the first block of data represents the signal that identifies the unknown location parameter, and the second block consists of auxiliary information that can be utilized to improve the precision of the location estimator. When the cost-benefit comparison is not clear cut, we recommend using the GMM procedure with a working weighting matrix.

The qualitative results of the paper are applicable more broadly. As long as there is additional nonparametric estimation uncertainty in a two-step procedure relative to that in the one-step procedure, we have to be very cautious about using the two-step procedure. While some asymptotic theory may indicate that the two-step procedure is always more efficient, the efficiency gain may not materialize in finite samples. In fact, blind use of the two-step procedure could do more harm than good.

Appendix A. Appendix of proofs

Proof of Proposition 2. By (11), we have

$$\text{avar}(\hat{\theta}_2) - \text{avar}(\hat{\theta}_1) = (1 + E\|\tilde{\beta}_\infty(h, 1, q)\|^2)\Omega_{11}^{1/2}\left[g(h, q)I_d - \rho\rho'\right](\Omega_{11}^{1/2})'.$$

Thus $\text{avar}(\hat{\theta}_2) > \text{avar}(\hat{\theta}_1)$ if and only if $g(h, q)I_d > \rho\rho'$. But $g(h, q)I_d > \rho\rho'$ if and only if $g(h, q) > \nu_{\max}\left(\rho\rho'\right)$. That is, $\text{avar}(\hat{\theta}_2) > \text{avar}(\hat{\theta}_1)$ if and only if $\nu_{\max}\left(\rho\rho'\right) < g(h, q)$. Similarly, we can prove that $\text{avar}(\hat{\theta}_2) < \text{avar}(\hat{\theta}_1)$ if and only if $\nu_{\min}\left(\rho\rho'\right) > g(h, q)$. ■

Proof of Proposition 3. Part (b) is a special case of Theorem 6(a) of Sun (2014b) with $G = [I_d, O_{d\times q}]'$. It remains to prove part (a). Under $\theta_0 = r + \delta_0/\sqrt{T}$, we have

$$\sqrt{T}(\hat{\theta}_1 - r) = \sqrt{T}(\hat{\theta}_1 - \theta_0) + \delta_0 \xrightarrow{d} \Omega_{11}^{1/2}B_d + \delta_0.$$

By Proposition 1(a), we have $\hat{\Omega}_{11} \xrightarrow{d} \Omega_{11}^{1/2}C_{dd}(\Omega_{11}^{1/2})'$, where $C_{dd} \perp B_d$. Invoking the continuous mapping theorem yields

$$\begin{aligned}\mathbb{W}_1 &:= \sqrt{T}(\hat{\theta}_1 - r)'\hat{\Omega}_{11}^{-1}\sqrt{T}(\hat{\theta}_1 - r)\\ &\xrightarrow{d} \left[\Omega_{11}^{1/2}B_d + \delta_0\right]'\left[\Omega_{11}^{1/2}C_{dd}(\Omega_{11}^{1/2})'\right]^{-1}\left[\Omega_{11}^{1/2}B_d + \delta_0\right].\end{aligned}$$

Thus

$$\mathbb{W}_1 \xrightarrow{d} \left[B_d + \Omega_{11}^{-1/2}\delta_0\right]'C_{dd}^{-1}\left[B_d + \Omega_{11}^{-1/2}\delta_0\right] \overset{d}{=} \mathbb{W}_{1\infty}(\|\Omega_{11}^{-1/2}\delta_0\|^2).$$

The use of the notation $\mathbb{W}_{1\infty}(\|\Omega_{11}^{-1/2}\delta_0\|^2) := \mathbb{W}_{1\infty}(\delta_0'\Omega_{11}^{-1}\delta_0)$ implies that the distribution of $\left[B_d + \Omega_{11}^{-1/2}\delta_0\right]'C_{dd}^{-1}\left[B_d + \Omega_{11}^{-1/2}\delta_0\right]$ depends on $\Omega_{11}^{-1/2}\delta_0$ only through $\delta_0'\Omega_{11}^{-1}\delta_0$. To see this, we let $H = (\Omega_{11}^{-1/2}\delta_0/\sqrt{\delta_0'\Omega_{11}^{-1}\delta_0}, \tilde{H})'$ for some

$\tilde{H}$ such that H is a $d \times d$ orthogonal matrix. Then

$$\begin{aligned}
&\left[B_d + \Omega_{11}^{-1/2}\delta_0\right]' C_{dd}^{-1}\left[B_d + \Omega_{11}^{-1/2}\delta_0\right] \\
&= \left[HB_d + H\Omega_{11}^{-1/2}\delta_0\right]' HC_{dd}^{-1}H'\left[HB_d + H\Omega_{11}^{-1/2}\delta_0\right] \\
&\overset{d}{=} \left[B_d + H\Omega_{11}^{-1/2}\delta_0\right]' C_{dd}^{-1}\left[B_d + H\Omega_{11}^{-1/2}\delta_0\right] \\
&= \left[B_d + \sqrt{\delta_0'\Omega_{11}^{-1}\delta_0}e_d\right]' C_{dd}^{-1}\left[B_d + \sqrt{\delta_0'\Omega_{11}^{-1}\delta_0}e_d\right],
\end{aligned}$$

where $e_d = (1, 0, \ldots, 0)' \in \mathbb{R}^d$. Therefore, the distribution of $\left[B_d + \Omega_{11}^{-1/2}\delta_0\right]' C_{dd}^{-1}\left[B_d + \Omega_{11}^{-1/2}\delta_0\right]$ depends on $\Omega_{11}^{-1/2}\delta_0$ only through $\delta_0'\Omega_{11}^{-1}\delta_0$. ■

Proof of Proposition 4.

Part (a) We first prove that $P\left(\chi_d^2\left(\delta^2\right) > x\right)$ is increasing in δ^2 for any integer d and $x > 0$. Note that

$$P\left(\chi_d^2\left(\delta^2\right) > x\right) = \sum_{j=0}^{\infty} \frac{e^{-\delta^2/2}(\delta^2/2)^j}{j!} P\left(\chi_{d+2j}^2 > x\right),$$

where χ_{d+2j}^2 is a (central) chi-square variate with $d + 2j$ degrees of freedom, we have

$$\begin{aligned}
\frac{\partial P\left(\chi_d^2\left(\delta^2\right) > x\right)}{\partial \delta^2} &= -\frac{1}{2}\sum_{j=0}^{\infty} \frac{(\delta^2/2)^j}{j!} e^{-\delta^2/2} P\left(\chi_{d+2j}^2 > x\right) + \frac{1}{2}\sum_{j=1}^{\infty} \frac{(\delta^2/2)^{j-1}}{(j-1)!} e^{-\delta^2/2} P\left(\chi_{d+2j}^2 > x\right) \\
&= -\frac{1}{2}\sum_{j=0}^{\infty} \frac{(\delta^2/2)^j}{j!} e^{-\delta^2/2} P\left(\chi_{d+2j}^2 > x\right) + \frac{1}{2}\sum_{j=0}^{\infty} \frac{(\delta^2/2)^j}{j!} e^{-\delta^2/2} P\left(\chi_{d+2+2j}^2 > x\right) \\
&= \frac{1}{2}\sum_{j=0}^{\infty} \frac{(\delta^2/2)^j}{j!} e^{-\delta^2/2}\left[P\left(\chi_{d+2+2j}^2 > x\right) - P\left(\chi_{d+2j}^2 > x\right)\right] > 0,
\end{aligned}$$

as claimed.

Let $\phi \sim N(0, 1)$, and let ψ be a 0-mean random variable that satisfies $\psi^2 > 0$ almost surely and $\psi \perp \phi$. Using the monotonicity of $P\left(\chi_d^2\left(\delta^2\right) > x\right)$ in δ^2, we have

$$\begin{aligned}
P(\|\phi + \psi\|^2 > x) &= E\left[P(\chi_1^2\left(\psi^2\right) > x)|\psi^2\right] \\
&> P(\chi_1^2 > x) = P(\|\phi\|^2 > x) \text{ for any } x.
\end{aligned}$$

Now we proceed to the proof of part (a). Note that B_d and B_q are independent of C_{dq}, C_{dd}, and C_{qq}. Let $D_{dd}^{-1} = \sum_{i=1}^{d}\lambda_{Di}d_i d_i'$ be the spectral decomposition of D_{dd}^{-1}, where $\lambda_{Di} \geq 0$ almost surely and d_i's are orthonormal in $\mathbb{R}^d$. Then

$$\begin{aligned}
&\left[B_d - C_{dq}C_{qq}^{-1}B_q\right]' D_{dd}^{-1}\left[B_d - C_{dq}C_{qq}^{-1}B_q\right] \\
&= \sum_{i=1}^{d}\lambda_{Di}\left[d_i'B_d - d_i'C_{dq}C_{qq}^{-1}B_q\right]^2 = \sum_{i=1}^{d}\lambda_{Di}(\phi_i + \psi_i)^2,
\end{aligned}$$

where $\phi_i = d_i'B_d$, $\psi_i = -d_i'C_{dq}C_{qq}^{-1}B_q$, and ϕ_i is independent of $\{\psi_i\}$ conditional on C_{dq}, C_{dd}, and C_{qq}. In addition, $\phi_i \sim iidN(0, 1)$ conditionally on C_{dq}, C_{dd}, and C_{qq} as well as unconditionally. Thus for any $x > 0$,

$$\begin{aligned}
P\left(\mathbb{W}_{2\infty}(0) > x\right) &= EP(\mathbb{W}_{2\infty}(0) > x|C_{dq}, C_{dd}, C_{qq}) \\
&= EP\left(\sum_{i=1}^{d}\lambda_{Di}(\phi_i + \psi_i)^2 > x|C_{dq}, C_{dd}, C_{qq}\right) \\
&= EP\left(\lambda_{D1}(\phi_1 + \psi_1)^2 > x - \sum_{i=2}^{d}\lambda_{Di}(\phi_i + \psi_i)^2|C_{dq}, C_{dd}, C_{qq}, \{\phi_i\}_{i=2}^d, \{\psi_i\}_{i=1}^d\right) \\
&\geq EP\left(\lambda_{D1}\phi_1^2 > x - \sum_{i=2}^{d}\lambda_{Di}(\phi_i + \psi_i)^2|C_{dq}, C_{dd}, C_{qq}, \{\phi_i, \psi_i\}_{i=2}^d\right) \\
&= EP\left(\lambda_{D1}\phi_1^2 > x - \sum_{i=2}^{d}\lambda_{Di}(\phi_i + \psi_i)^2|C_{dq}, C_{dd}, C_{qq}, \{\psi_i\}_{i=2}^d\right).
\end{aligned}$$

Using the above argument repeatedly, we have

$$\begin{aligned}
&P\left(\mathbb{W}_{2\infty}\left(0\right)>x\right)\geq EP\left(\sum_{i=1}^{d}\lambda_{Di}\phi_i^2>x|C_{dq},C_{dd},C_{qq}\right)\\
&=P\left(\sum_{i=1}^{d}\lambda_{Di}\phi_i^2>x\right)=P\left[B_d'D_{dd}^{-1}B_d>x\right]\\
&>P\left[B_d'C_{dd}^{-1}B_d>x\right]=P\left(\mathbb{W}_{1\infty}\left(0\right)>x\right),
\end{aligned}$$

where the last inequality follows from the fact that $D_{dd}^{-1}>C_{dd}^{-1}$ almost surely.

Part (b). Let $C_{dd}^{-1}=\sum_{i=1}^{d}\lambda_{Ci}c_ic_i'$ be the spectral decomposition of C_{dd}^{-1}. Since $C_{dd}>0$ almost surely and $\lambda_{Ci}>0$ almost surely, we have

$$\begin{aligned}
&\mathbb{W}_{1\infty}(\|\xi\|^2)\overset{d}{=}\left[B_d+\|\xi\|e_d\right]'C_{dd}^{-1}\left[B_d+\|\xi\|e_d\right]\\
&=\sum_{i=1}^{d}\lambda_{Ci}\left[c_i'B_d+\|\xi\|c_i'e_d\right]^2,
\end{aligned}$$

where $\left[c_i'B_d+\|\xi\|c_i'e_d\right]^2$ follows the independent noncentral chi-square distribution with noncentrality parameter $\|\xi\|^2\left(c_i'e_d\right)^2$, conditional on $\{\lambda_{Ci}\}_{i=1}^d$ and $\{c_i\}_{i=1}^d$. Now consider two vectors ξ_1 and ξ_2 such that $\|\xi_1\|<\|\xi_2\|$. We have

$$\begin{aligned}
&P\left[\mathbb{W}_{1\infty}\left(\|\xi_1\|^2\right)>x\right]\\
&=P\left\{\sum_{i=1}^{d}\lambda_{Ci}\left[c_i'B_d+\|\xi_1\|c_i'e_d\right]^2>x\right\}\\
&=EP\left\{\lambda_{C1}\left[c_1'B_d+\|\xi_1\|c_1'e_d\right]^2>x-\sum_{i=2}^{d}\lambda_{Ci}\left[c_i'B_d+\|\xi_1\|c_i'e_d\right]^2\Big|\{\lambda_{Ci}\}_{i=1}^d,\{c_i\}_{i=1}^d\right\}\\
&<EP\left\{\lambda_{C1}\left[c_1'B_d+\|\xi_2\|c_1'e_d\right]^2>x-\sum_{i=2}^{d}\lambda_{Ci}\left[c_i'B_d+\|\xi_1\|c_i'e_d\right]^2\Big|\{\lambda_{Ci}\}_{i=1}^d,\{c_i\}_{i=1}^d\right\}\\
&=P\left\{\lambda_{C1}\left[c_1'B_d+\|\xi_2\|c_1'e_d\right]^2+\sum_{i=2}^{d}\lambda_{Ci}\left[c_i'B_d+\|\xi_1\|c_i'e_d\right]^2>x\right\},
\end{aligned}$$

where we have used the strict monotonicity of $P\left(\chi_1^2\left(\delta^2\right)>x\right)$ in δ^2. Repeating the above argument, we have

$$\begin{aligned}
&P\left[\mathbb{W}_{1\infty}(\|\xi_1\|^2)>x\right]\\
&<P\left\{\sum_{i=1}^{d}\lambda_{Ci}\left[c_i'B_d+\|\xi_2\|c_i'e_d\right]^2>x\right\}\\
&=P\left\{\left[B_d+\xi_2\right]'C_{dd}^{-1}\left[B_d+\xi_2\right]>x\right\}=P\left[\mathbb{W}_{1\infty}\left(\|\xi_2\|^2\right)>x\right],
\end{aligned}$$

as desired.

Part (c). We note that

$$\begin{aligned}
&\mathbb{W}_{2\infty}(\|\xi\|^2)\\
&=\left[B_d-C_{dq}C_{qq}^{-1}B_q+\|\xi\|e_d\right]'D_{dd}^{-1}\left[B_d-C_{dq}C_{qq}^{-1}B_q+\|\xi\|e_d\right]\\
&=\left\{\left[I_d+C_{dq}C_{qq}^{-1}C_{qq}^{-1}C_{qd}\right]^{-1/2}\left[B_d-C_{dq}C_{qq}^{-1}B_q\right]+\|\xi\|\tilde{e}_d\right\}'\times\tilde{D}_{dd}^{-1}\\
&\times\left\{\left[I_d+C_{dq}C_{qq}^{-1}C_{qq}^{-1}C_{qd}\right]^{-1/2}\left[B_d-C_{dq}C_{qq}^{-1}B_q\right]+\|\xi\|\tilde{e}_d\right\},
\end{aligned}$$

where

$$\tilde{e}_d=\left[I_d+C_{dq}C_{qq}^{-1}C_{qq}^{-1}C_{qd}\right]^{-1/2}e_d$$

and

$$\tilde{D}_{dd}^{-1}=\left[I_d+C_{dq}C_{qq}^{-1}C_{qq}^{-1}C_{qd}\right]^{1/2}D_{dd}^{-1}\left[I_d+C_{dq}C_{qq}^{-1}C_{qq}^{-1}C_{qd}\right]^{1/2}.$$

Let $\sum_{i=1}^{d}\tilde{\lambda}_{D_i}\tilde{d}_i\tilde{d}_i'$ be the spectral decomposition of $\tilde{D}_{dd}^{-1}$. Define

$$\tilde{\phi}_{di}=\tilde{d}_i'\left[I_d+C_{dq}C_{qq}^{-1}C_{qq}^{-1}C_{qd}\right]^{-1/2}\left[B_d-C_{dq}C_{qq}^{-1}B_q\right].$$

Then conditional on C_{dq}, C_{dd}, and C_{qq}, $\tilde{\phi}_{di} \sim iidN(0, 1)$. Since the conditional distribution does not depend on C_{dq}, C_{dd} or C_{qq}, $\tilde{\phi}_{di} \sim iidN(0, 1)$ unconditionally. Now

$$\begin{aligned}
&\mathbb{W}_{2\infty}(\|\xi_1\|^2) \\
&= \sum_{i=1}^{d} \tilde{\lambda}_{D_i} \left\{ \tilde{d}_i' \left[I_d + C_{dq} C_{qq}^{-1} C_{qq}^{-1} C_{qd} \right]^{-1/2} \left[B_d - C_{dq} C_{qq}^{-1} B_q \right] + \|\xi_1\| \, d_i' \tilde{e}_d \right\}^2 \\
&= \sum_{i=1}^{d} \tilde{\lambda}_{D_i} \left(\tilde{\phi}_{di} + \|\xi_1\| \, \tilde{d}_i' \tilde{e}_d \right)^2,
\end{aligned}$$

and so for two vectors ξ_1 and ξ_2 such that $\|\xi_1\| < \|\xi_2\|$, we have

$$\begin{aligned}
&P\left\{ \mathbb{W}_{2\infty}(\|\xi_1\|^2) > x \right\} \\
&= EP\left\{ \sum_{i=1}^{d} \tilde{\lambda}_{D_i} \left(\tilde{\phi}_{di} + \|\xi_1\| \, \tilde{d}_i' \tilde{e}_d \right)^2 > x \middle| C_{dq}, C_{dd}, C_{qq} \right\} \\
&< EP\left\{ \sum_{i=1}^{d} \tilde{\lambda}_{D_i} \left(\tilde{\phi}_{di} + \|\xi_2\| \, \tilde{d}_i' \tilde{e}_d \right)^2 > x \middle| C_{dq}, C_{dd}, C_{qq} \right\} \\
&= P\left\{ \sum_{i=1}^{d} \tilde{\lambda}_{D_i} \left(\tilde{\phi}_{di} + \|\xi_2\| \, \tilde{d}_i' \tilde{e}_d \right)^2 > x \right\} = P\left\{ \mathbb{W}_{2\infty}(\|\xi_2\|^2) > x \right\}. \quad \blacksquare
\end{aligned}$$

Proof of Proposition 5. We prove only part (b), as part (a) can be proved using the same argument. Let

$$\rho\rho' = \sum_{i=1}^{d} \nu_i a_i a_i'$$

be the eigen-decomposition of $\rho\rho'$ with eigenvalues $\{\nu_i\}$ satisfying $1 > \nu_d \geq \nu_{d-1}, \ldots \geq \nu_1 \geq 0$. Then for $\lambda_0 = \delta_0' \Omega_{1\cdot 2}^{-1} \delta_0$, we have

$$\begin{aligned}
&\delta_0' \Omega_{1\cdot 2}^{-1} \delta_0 - \tau(\lambda_0) \cdot \delta_0' \Omega_{11}^{-1} \delta_0 \\
&= \tau(\lambda_0) \sum_{i=1}^{d} \frac{1}{1 - \nu_i} \left[\nu_i - f(\lambda_0) \right] \left(a_i' \Omega_{11}^{-1/2} \delta_0 \right)^2 \\
&= \tau(\lambda_0) \|\Omega_{11}^{-1/2} \delta_0\|^2 \sum_{i=1}^{d} \frac{1}{1 - \nu_i} \left[\nu_i - f(\lambda_0) \right] \left\langle a_i, \frac{\Omega_{11}^{-1/2} \delta_0}{\|\Omega_{11}^{-1/2} \delta_0\|} \right\rangle^2,
\end{aligned} \tag{23}$$

where $\langle \cdot, \cdot \rangle$ is the usual inner product.

We now show that $\delta_0' \Omega_{1\cdot 2}^{-1} \delta_0 - \tau(\lambda_0) \cdot \delta_0' \Omega_{11}^{-1} \delta_0 > 0$ for all $\delta_0 \in \mathfrak{A}(\lambda_0)$ with $\lambda_0 > 0$ if and only if $\nu_i - f(\lambda_0) > 0$ for all $i \in \{1, \ldots, d\}$. The "if" part is obvious. We prove the "only if" part by contradiction. Suppose that $\delta_0' \Omega_{1\cdot 2}^{-1} \delta_0 - \tau(\lambda_0) \cdot \delta_0' \Omega_{11}^{-1} \delta_0 > 0$ for all $\delta_0 \in \mathfrak{A}(\lambda_0)$ with $\lambda_0 > 0$ but there exists an i^* such that $\nu_{i^*} - f(\lambda_0) \leq 0$. Choosing $\delta_0 \in \mathfrak{A}(\lambda_0)$ such that $\Omega_{11}^{-1/2} \delta_0 / \|\Omega_{11}^{-1/2} \delta_0\| = a_{i^*}$, we have

$$\delta_0' \Omega_{1\cdot 2}^{-1} \delta_0 - \tau(\lambda_0) \cdot \delta_0' \Omega_{11}^{-1} \delta_0 = \frac{\delta_0' \Omega_{11}^{-1} \delta_0}{1 - \nu_{i^*}} \left[\nu_{i^*} - f(\lambda_0) \right] \tau(\lambda_0) \leq 0,$$

leading to a contradiction.

Note that the condition $\nu_i - f(\lambda_0) > 0$ for all $i \in \{1, \ldots, d\}$ is equivalent to $\nu_1 > f(\lambda_0)$, which is the same as $\nu_{\min}\left(\rho\rho'\right) > f(\lambda_0; h, p, q, \alpha)$. This completes the proof of part (b). $\blacksquare$

Proof of Lemma 6. We prove a more general result by establishing a representation for

$$\frac{1}{\sqrt{T}} \sum_{t=1}^{T} \left[\breve{G}' \breve{M}^{-1} \breve{G} \right]^{-1} \breve{G}' \breve{M}^{-1} \breve{f}(v_t, \theta_0)$$

in terms of the rotated and normalized moment conditions for any $m \times m$ (almost surely) positive definite matrix $\breve{M}$, which can be random. Let

$$M^* = U' \breve{M} U, M = \Sigma_{1/2}^{*-1} M^* (\Sigma_{1/2}^{*-1})' = \begin{pmatrix} M_{11} & M_{12} \\ M_{21} & M_{22} \end{pmatrix},$$

and let $M_{1\cdot 2} = M_{11} - M_{12}M_{22}^{-1}M_{21}$, where $M_{11} \in \mathbb{R}^{d\times d}$ and $M_{22} \in \mathbb{R}^{q\times q}$. Using the SVD $U\Xi V'$ of $\breve{G}$, we have

$$\begin{aligned}
\breve{G}'\breve{M}^{-1}\breve{G} &= V\Xi'(U'\breve{M}U)^{-1}\Xi V' \\
&= VA\begin{pmatrix} I_d, & O\end{pmatrix}\left(M^*\right)^{-1}\begin{pmatrix} I_d, & O\end{pmatrix}'AV' \\
&= VA\begin{pmatrix} I_d, & O\end{pmatrix}(\Sigma_{1/2}^{*-1})'\left[\Sigma_{1/2}^{*-1}M^*(\Sigma_{1/2}^{*-1})'\right]^{-1}\Sigma_{1/2}^{*-1}\begin{pmatrix} I_d, & O\end{pmatrix}'AV' \\
&= VA\begin{pmatrix} I_d, & O\end{pmatrix}(\Sigma_{1/2}^{*-1})'M^{-1}\Sigma_{1/2}^{*-1}\begin{pmatrix} I_d, & O\end{pmatrix}'AV' \\
&= VA\left(\Sigma_{1\cdot 2}^*\right)^{-1/2}\begin{pmatrix} I_d, & O\end{pmatrix}M^{-1}\left[VA\left(\Sigma_{1\cdot 2}^*\right)^{-1/2}\begin{pmatrix} I_d, & O\end{pmatrix}\right]' \\
&= VA\left(\Sigma_{1\cdot 2}^*\right)^{-1/2}M_{1\cdot 2}^{-1}\left(\Sigma_{1\cdot 2}^*\right)^{-1/2}AV', \qquad (24)
\end{aligned}$$

where we have used

$$\begin{aligned}
\begin{pmatrix} I_d, & O\end{pmatrix}(\Sigma_{1/2}^{*-1})' &= \begin{pmatrix} I_d, & O\end{pmatrix}\begin{pmatrix} \left(\Sigma_{1\cdot 2}^*\right)^{-1/2} & O \\ -\left[\left(\Sigma_{1\cdot 2}^*\right)^{-1/2}\Sigma_{12}^*\left(\Sigma_{22}^*\right)^{-1}\right]' & \left(\Sigma_{22}^*\right)^{-1/2}\end{pmatrix} \\
&= \left(\left(\Sigma_{1\cdot 2}^*\right)^{-1/2}, \quad O\right) = \left(\Sigma_{1\cdot 2}^*\right)^{-1/2}\begin{pmatrix} I_d, & O\end{pmatrix}.
\end{aligned}$$

In addition,

$$\begin{aligned}
&\breve{G}'\breve{M}^{-1}\breve{f}(v_t,\theta_0) \\
&= V\Xi'(U'\breve{M}U)^{-1}U'\breve{f}(v_t,\theta_0) = VA\begin{pmatrix} I_d, & O\end{pmatrix}\left(M^*\right)^{-1}f^*(v_t,\theta_0) \\
&= VA\begin{pmatrix} I_d, & O\end{pmatrix}(\Sigma_{1/2}^{*-1})'\left[\Sigma_{1/2}^{*-1}M^*(\Sigma_{1/2}^{*-1})'\right]^{-1}\Sigma_{1/2}^{*-1}f^*(v_t,\theta_0) \\
&= VA\begin{pmatrix} I_d, & O\end{pmatrix}(\Sigma_{1/2}^{*-1})'M^{-1}f(v_t,\theta_0) \\
&= VA\left(\Sigma_{1\cdot 2}^*\right)^{-1/2}\begin{pmatrix} I_d, & O\end{pmatrix}M^{-1}f(v_t,\theta_0) \\
&= VA\left(\Sigma_{1\cdot 2}^*\right)^{-1/2}\begin{pmatrix} I_d, & O\end{pmatrix}\begin{pmatrix} M_{1\cdot 2}^{-1} & -M_{1\cdot 2}^{-1}M_{12}M_{22}^{-1} \\ -\left(M_{1\cdot 2}^{-1}M_{12}M_{22}^{-1}\right)' & M_{2\cdot 1}^{-1}\end{pmatrix}f(v_t,\theta_0) \\
&= VA\left(\Sigma_{1\cdot 2}^*\right)^{-1/2}\left(M_{1\cdot 2}^{-1}, \quad -M_{1\cdot 2}^{-1}M_{12}M_{22}^{-1}\right)f(v_t,\theta_0) \\
&= VA\left(\Sigma_{1\cdot 2}^*\right)^{-1/2}M_{1\cdot 2}^{-1}\left[f_1(v_t,\theta_0) - M_{12}M_{22}^{-1}f_2(v_t,\theta_0)\right].
\end{aligned}$$

Thus

$$\begin{aligned}
&\frac{1}{\sqrt{T}}\sum_{t=1}^{T}\left[\breve{G}'\breve{M}^{-1}\breve{G}\right]^{-1}\breve{G}'\breve{M}^{-1}\breve{f}(v_t,\theta_0) \\
&= \frac{1}{\sqrt{T}}\sum_{t=1}^{T}\left[VA\left(\Sigma_{1\cdot 2}^*\right)^{-1/2}M_{1\cdot 2}^{-1}\left(\Sigma_{1\cdot 2}^*\right)^{-1/2}AV'\right]^{-1}\left[VA\left(\Sigma_{1\cdot 2}^*\right)^{-1/2}M_{1\cdot 2}^{-1}\right] \\
&\times \left[f_1(v_t,\theta_0) - M_{12}M_{22}^{-1}f_2(v_t,\theta_0)\right] \\
&= \frac{1}{\sqrt{T}}\sum_{t=1}^{T}VA^{-1}\left(\Sigma_{1\cdot 2}^*\right)^{1/2}\left[f_1(v_t,\theta_0) - M_{12}M_{22}^{-1}f_2(v_t,\theta_0)\right]. \qquad (25)
\end{aligned}$$

Letting $\breve{M} = \breve{\Sigma}$, we have $M^* = U'\breve{\Sigma}U = \Sigma^*$ and $M = \Sigma_{1/2}^{*-1}M^*(\Sigma_{1/2}^{*-1})' = I_m$, and thus $M_{12}M_{22}^{-1} = 0$. As a result,

$$\frac{1}{\sqrt{T}}\sum_{t=1}^{T}\left[\breve{G}'\breve{\Sigma}^{-1}\breve{G}\right]^{-1}\breve{G}'\breve{\Sigma}^{-1}\breve{f}(v_t,\theta_0) = \frac{1}{\sqrt{T}}\sum_{t=1}^{T}VA^{-1}\left(\Sigma_{1\cdot 2}^*\right)^{1/2}f_1(v_t,\theta_0).$$

Using this and the stochastic expansion of $\sqrt{T}(\hat{\theta}_1 - \theta_0)$, we have

$$\sqrt{T}(\hat{\theta}_1 - \theta_0) = \frac{1}{\sqrt{T}}\sum_{t=1}^{T}VA^{-1}\left(\Sigma_{1\cdot 2}^*\right)^{1/2}f_1(v_t,\theta_0) + o_p(1).$$

It then follows that

$$\left(\Sigma_{1\cdot 2}^*\right)^{-1/2}AV'\sqrt{T}(\hat{\theta}_1 - \theta_0) = \frac{1}{\sqrt{T}}\sum_{t=1}^{T}f_1(v_t,\theta_0) + o_p(1) \xrightarrow{d} N(0,\Omega_{11}).$$

Letting $\breve{M} = \breve{\Omega}_\infty$, we have $M = \Sigma_{1/2}^{*-1} U' \breve{\Omega}_\infty U \Sigma_{1/2}^{*-1\prime} = \Omega_\infty$, and thus $M_{12} M_{22}^{-1} = \Omega_{\infty,12} \Omega_{\infty,22}^{-1} = \beta_\infty$. As a result,

$$\frac{1}{\sqrt{T}} \sum_{t=1}^{T} \left[\breve{G}' \breve{\Omega}_\infty^{-1} \breve{G} \right]^{-1} \breve{G}' \breve{\Omega}_\infty^{-1} \breve{f}(v_t, \theta_0) = \frac{1}{\sqrt{T}} \sum_{t=1}^{T} VA^{-1} \left(\Sigma_{1\cdot 2}^* \right)^{1/2} \left[f_1(v_t, \theta_0) - \beta_\infty f_2(v_t, \theta_0) \right].$$

Using this, we have

$$\begin{aligned} \sqrt{T}(\hat{\theta}_2 - \theta_0) &= \frac{1}{\sqrt{T}} \sum_{t=1}^{T} \left[\breve{G}' \breve{\Omega}_\infty^{-1} \breve{G} \right]^{-1} \breve{G}' \breve{\Omega}_\infty^{-1} \breve{f}(v_t, \theta_0) + o_p(1) \\ &= \frac{1}{\sqrt{T}} VA^{-1} \left(\Sigma_{1\cdot 2}^* \right)^{1/2} \sum_{t=1}^{T} (f_1(v_t, \theta_0) - \beta_\infty f_2(v_t, \theta_0)) + o_p(1). \end{aligned}$$

It then follows that

$$\begin{aligned} \left(\Sigma_{1\cdot 2}^* \right)^{-1/2} AV' \sqrt{T}(\hat{\theta}_2 - \theta_0) &= \frac{1}{\sqrt{T}} \sum_{t=1}^{T} [f_1(v_t, \theta_0) - \beta_\infty f_2(v_t, \theta_0)] + o_p(1) \\ &\xrightarrow{d} MN \left(0, \Omega_{11} - \Omega_{12}\beta_\infty' - \beta_\infty \Omega_{21} + \beta_\infty \Omega_{22} \beta_\infty' \right). \quad \blacksquare \end{aligned}$$

Proof of Theorem 7. Parts (a) and (b). Instead of comparing the asymptotic variances of $R\sqrt{T}(\hat{\theta}_1 - \theta_0)$ and $R\sqrt{T}(\hat{\theta}_2 - \theta_0)$ directly, we (equivalently) compare the asymptotic variances of $(\tilde{R}\tilde{R}')^{-1/2} R\sqrt{T}(\hat{\theta}_1 - \theta_0)$ and $(\tilde{R}\tilde{R}')^{-1/2} R\sqrt{T}(\hat{\theta}_2 - \theta_0)$. We can do this, because $(\tilde{R}\tilde{R}')^{-1/2}$ is nonsingular. Note that the latter two asymptotic variances are the same as those of the one-step and two-step estimators $\hat{\theta}_1^R$ and $\hat{\theta}_2^R$, respectively, of θ_0^R in the simple location model

$$\begin{cases} y_{1t}^R = \theta_0^R + u_{1t}^R \in \mathbb{R}^p, \\ y_{2t} = u_{2t} \in \mathbb{R}^q, \end{cases} \tag{26}$$

where

$$\theta_0^R = (\tilde{R}\tilde{R}')^{-1/2} R\theta_0, \;\; u_{1t}^R = (\tilde{R}\tilde{R}')^{-1/2} \tilde{R} u_{1t},$$

and the contemporaneous variance and long-run variances of $u_t = (u_{1t}', u_{2t}')'$ are I_m and Ω respectively.

It suffices to compare the asymptotic variances of $\hat{\theta}_1^R$ and $\hat{\theta}_2^R$ in the above location model. By construction, the variance of $u_t^R := \left(\left(u_{1t}^R \right)', (u_{2t})' \right)'$ is

$$\operatorname{var}(u_t^R) = \begin{pmatrix} I_p & O \\ O & I_q \end{pmatrix} = I_{p+q}.$$

Therefore, the above location model has exactly the same form as the model in Section 3.1, the only difference being that y_{1t}^R is of dimension p instead of d. This difference is not substantive and requires only some notational changes. We can invoke Proposition 2 to complete the proof.

The long-run canonical correlation coefficients between u_{1t}^R and u_{2t} are the same as those between $\tilde{R}u_{1t}$ and u_{2t}. This follows from the facts that u_{1t}^R is equal to $\tilde{R}u_{1t}$ pre-multiplied by a full-rank square matrix and the long-run correlation matrix between $\tilde{R}u_{1t}$ and u_{2t} is

$$(\tilde{R}\Omega_{11}\tilde{R}')^{-1/2} \{\tilde{R}\Omega_{12}\} \times \Omega_{22}^{-1/2} = \rho_R.$$

The long-run canonical correlation coefficients between u_{1t}^R and u_{2t} are then the eigenvalues of $\rho_R \rho_R'$, that is, $\nu\left(\rho_R \rho_R'\right)$. Parts (a) and (b) then follow from Proposition 2.

Parts (c) and (d). The local asymptotic powers of the one-step and two-step tests are the same as the local asymptotic powers of the respective one-step and two-step tests for testing about θ_0^R in the location model given in (26). We use Proposition 5 to complete the proof. For the above location model, the asymptotic variance of the infeasible two-step GMM estimator is

$$\Omega_{1\cdot 2}^R = \left[(\tilde{R}\tilde{R}')^{-1/2} \tilde{R} \right] \Omega_{1\cdot 2} \left[(\tilde{R}\tilde{R}')^{-1/2} \tilde{R} \right]'.$$

In addition, the local alternative hypothesis $H_1 : R\theta_0 = r + \delta_0/\sqrt{T}$ corresponds to

$$H_1 : \theta_0^R = (\tilde{R}\tilde{R}')^{-1/2} r + (\tilde{R}\tilde{R}')^{-1/2} \delta_0/\sqrt{T}$$

in the location model. This implies that the local alternative parameter is $(\tilde{R}\tilde{R}')^{-1/2}\delta_0/\sqrt{T}$. Thus the set of δ_0's considered in Proposition 5 is given by

$$\begin{aligned}\mathfrak{A}_{loc}(\lambda_0) &= \left\{\delta : \left[(\tilde{R}\tilde{R}')^{-1/2}\delta\right]' \left(\Omega_{1\cdot2}^{R}\right)^{-1}\left[(\tilde{R}\tilde{R}')^{-1/2}\delta\right] = \lambda_0\right\} \\ &= \left\{\delta : \delta'(\tilde{R}\Omega_{1\cdot2}\tilde{R}')^{-1}\delta = \lambda_0\right\}. \end{aligned} \tag{27}$$

It remains to show that the above set is the same as the set given in the theorem.

Using (24) with $\breve{M}^{-1} = \breve{\Omega}^{-1}$, we have $M = \Omega$, and so

$$\breve{G}'\breve{\Omega}^{-1}\breve{G} = VA\left(\Sigma_{1\cdot2}^{*}\right)^{-1/2}\Omega_{1\cdot2}^{-1}\left(\Sigma_{1\cdot2}^{*}\right)^{-1/2}AV'.$$

Plugging this into $\delta'\left[R(\breve{G}'\breve{\Omega}^{-1}\breve{G})^{-1}R'\right]^{-1}\delta$ yields

$$\begin{aligned}&\delta'\left[R(\breve{G}'\breve{\Omega}^{-1}\breve{G})^{-1}R'\right]^{-1}\delta \\ &= \delta'\left\{R\left[VA\left(\Sigma_{1\cdot2}^{*}\right)^{-1/2}\Omega_{1\cdot2}^{-1}\left(\Sigma_{1\cdot2}^{*}\right)^{-1/2}AV'\right]^{-1}R'\right\}^{-1}\delta \\ &= \delta'\left\{RVA^{-1}\left(\Sigma_{1\cdot2}^{*}\right)^{1/2}\Omega_{1\cdot2}\left(\Sigma_{1\cdot2}^{*}\right)^{1/2}A^{-1}V'R'\right\}^{-1}\delta = \delta'\left(\tilde{R}\Omega_{1\cdot2}\tilde{R}'\right)^{-1}\delta.\end{aligned}$$

Therefore, the set of δ_0's considered in the theorem is exactly the same as the set given in (27). ■

Appendix B. Supplementary data

Supplementary material related to this article can be found online at https://doi.org/10.1016/j.jeconom.2018.07.006.

References

Andrews, D.W., 1991. Heteroskedasticity and autocorrelation consistent covariance matrix estimation. Econometrica 59 (3), 817–858.

Bester, C.A., Conley, T.G., Hansen, C.B., 2011. Inference with dependent data using cluster covariance estimators. J. Econometrics 165 (2), 137–151.

Bester, C.A., Conley, T.G., Hansen, C.B., Vogelsang, T.J., 2016. Fixed-b asymptotics for spatially dependent robust nonparametric covariance matrix estimators. Econometric Theory 32 (1), 154–186.

Chen, X., Liao, Z., Sun, Y., 2014. Sieve inference on possibly misspecified seminonparametric time series models. J. Econometrics 178, 639–658.

Gonçalves, S., 2011. The moving blocks bootstrap for panel linear regression models with individual fixed effects. Econometric Theory 27 (5), 1048–1082.

Gonçalves, S., Vogelsang, T., 2011. Block bootstrap hac robust tests: the sophistication of the naive bootstrap. Econometric Theory 27 (4), 745–791.

Hansen, L.P., 1982. Large sample properties of generalized method of moments estimators. Econometrica 50, 1029–1054.

Hwang, J., Sun, Y., Should we go one step further? An accurate comparison of one-step and two-step procedures in a generalized method of moments framework. Working paper version of this paper, Department of Economics, UC San Diego.

Ibragimov, R., Müller, U.K., 2010. t-Statistic Based Correlation and Heterogeneity Robust Inference. J. Bus. Econom. Statist. 28, 453–468.

Jansson, M., 2004. The error in rejection probability of simple autocorrelation robust tests. Econometrica 72 (3), 937–946.

Kiefer, N.M., Vogelsang, T.J., 2002a. Heteroskedasticity-autocorrelation robust testing using bandwidth equal to sample size. Econometric Theory 18, 1350–1366.

Kiefer, N.M., Vogelsang, T.J., 2002b. Heteroskedasticity-autocorrelation robust standard errors using the bartlett kernel without truncation. Econometrica 70 (5), 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., 2002. Simple robust testing of regression hypotheses. Econometrica 68 (3), 695–714.

Kim, M.S., Sun, Y., 2013. Heteroskedasticity and spatiotemporal dependence robust inference for linear panel models with fixed effects. J. Econometrics 177 (1), 85–108.

Kim, M.S., Sun, Y., Yang, J., 2017. A fixed-bandwidth view of the pre-asymptotic inference for kernel smoothing with time series data. J. Econometrics 197 (2), 298–322.

Lazarus, E., Lewis, D.J., Stock, J.H., Watson, M.W., 2016. HAR Inference: Kernel Choice, Size Distortions, and Power Losses. Working Paper. Department of Economics, Harvard University.

Müller, U.K., 2007. A theory of robust long-run variance estimation. J. Econometrics 141 (2), 1331–1352.

Newey, W.K., West, K.D., 1987. A simple, positive semi-definite, heteroskedasticity and autocorrelation consistent covariance matrix. Econometrica 55, 703–708.

Phillips, P.C.B., 2005. HAC estimation by automated regression. Econometric Theory 21 (1), 116–142.

Preinerstorfer, D., Finite sample properties of tests based on prewhitened nonparametric covariance estimators. Working paper, Department of Statistics and OR, University of Vienna.

Preinerstorfer, D., Pötscher, B.M., 2016. On size and power of heteroskedasticity and autocorrelation robust tests. Econometric Theory 32 (2), 261–358.

Shao, X., 2010. A self-normalized approach to confidence interval construction in time series. Journal of the Royal Statistical Society B 72, 343–366.

Sun, Y., 2013. A heteroskedasticity and autocorrelation robust f test using orthonormal series variance estimator. Econom. J. 16 (1), 1–26.

Sun, Y., 2014a. Let's fix it: fixed-*b* asymptotics versus small-*b* asymptotics in heteroscedasticity and autocorrelation robust inference. J. Econometrics 178 (3), 659–677.

Sun, Y., 2014b. Fixed-smoothing asymptotics in a two-step gmm framework. Econometrica 82 (6), 2327–2370.

Sun, Y., Phillips, P.C.B., Bandwidth choice for interval estimation in GMM regression. Working Paper, Department of Economics, UC San Diego.

Sun, Y., Phillips, P.C.B., Jin, S., 2008. Optimal bandwidth selection in heteroskedasticity-autocorrelation robust testing. Econometrica 76 (1), 175–194.

Vogelsang, T.J., 2012. Heteroskedasticity, autocorrelation, and spatial correlation robust inference in linear panel models with fixed-effects. J. Econometrics 166 (2), 303–319.
Wilhelm, D., 2015. Optimal bandwidth selection for robust generalized method of moments estimation. Econometric Theory 31 (5), 1054–1077.
Xu, K.-L., 2012. Robustifying multivariate trend tests to nonstationary volatility. J. Econometrics 169 (2), 147–154.
Zhang, X., 2016. Fixed-smoothing asymptotics in the generalized empirical likelihood estimation framework. J. Econometrics 193 (1), 123–146.
Zhang, X., Shao, X., 2013. Fixed-Smoothing asymptotics for time series. Ann. Statist. 41 (3), 1329–1349.