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 and Yixiao Sun*

Department of Economics,
University of California, San Diego†

July 24, 2015

Abstract

According to the conventional asymptotic theory, the two-step Generalized Method of Moments (GMM) estimator and test perform as least as well as the one-step estimator and test in large samples. The conventional asymptotic theory, as elegant and convenient as it is, 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 fixed-smoothing asymptotic theory that accounts for the estimation uncertainty, and compare the performance of the one-step and two-step procedures in this more accurate asymptotic framework. We show 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. This qualitative message applies to both the asymptotic variance comparison and power comparison of the associated tests. A Monte Carlo study lends support to our asymptotic results.

JEL Classification: C12, C32

Keywords: Asymptotic Efficiency, Asymptotic Mixed Normality, Fixed-smoothing Asymptotics, Heteroskedasticity and Autocorrelation Robust, Increasing-smoothing Asymptotics, Nonstandard Asymptotics, Two-step GMM Estimation

1 Introduction

Efficiency is one of the most important problems in statistics and econometrics. In the widely-used GMM framework, it is standard practice to employ a two-step procedure to improve the

*For helpful comments and suggestions, we would like to thank Brendan Beare, Graham Elliott, Bruce Hansen, Jonathan Hill, Min Seong Kim, Oliver Linton, 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, 2014 Shanghai Jiao Tong University and Singapore Management University Bi-party Conference, and 2014 Shandong Econometrics Conference, China.

†Email: j6hwang@ucsd.edu, yisun@ucsd.edu. Correspondence to: Department of Economics, University of California, San Diego, 9500 Gilman Drive, La Jolla, CA 92093-0508.

efficiency of the GMM estimator and the power of the associated tests. The two-step procedure requires the estimation of a weighting matrix. According to the Hansen (1982), the optimal weighting matrix is the asymptotic variance of the (scaled) sample moment conditions. For time series data, which is our focus here, 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 or series 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 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 large estimation uncertainty. A good approximation 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 weighing matrix is accounted for. There are two ways to capture the estimation uncertainty. One is to use the high order conventional asymptotic theory under which the amount of nonparametric smoothing in the LRV estimator increases with the sample size but at a slower rate. While the estimation uncertainty vanishes in the first order asymptotics, we expect it to remain in high order asymptotics. The second way is to use an alternative asymptotic approximation that can capture the estimation uncertainty even with just a first-order asymptotics. 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 in the recent literature.

In this paper, we employ 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 this new asymptotic theory. So 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 that 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. This qualitative message applies to both the asymptotic variance comparison and the local asymptotic power comparison of the associated

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

tests. This is in sharp contrast with the conventional asymptotics 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 an honest assessment of their relative merits. This is confirmed by a Monte Carlo study.

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 first studied by Kiefer, Vogelsang and Bunzel (2002) and Kiefer and Vogelsang (2002a, 2002b, 2005) in the econometrics literature. For other studies, see, for example, Jansson (2004), Sun, Phillips and Jin (2008), Sun and Phillips (2009), Gonçlaves and Vogelsang (2011), and Zhang and Shao (2013) in the time series setting; Bester, Conley, Hansen and Vogelsang (2014) in the spatial setting; and Gonçalves (2011), Kim and Sun (2013), and Vogelsang (2012) in the panel data setting. For orthonormal series LRV estimators, the fixed-smoothing asymptotics is the so-called fixed-K asymptotics. For its theoretical development and related simulation evidence, see, for example, Phillips (2005), Müller (2007), Sun (2011, 2013) and Sun and Kim (2015). 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, Conley, and Hansen (2011). The fixed-smoothing asymptotics can be regarded as a convenient device to obtain some high order terms under the conventional increasing-smoothing asymptotics.

The rest of the paper is organized as follows. The next section presents a simple overidentified GMM framework. Section 3 compares the two procedures from the perspective of point estimation. Section 4 compares them from the testing perspective. Section 5 extends the ideas to a general GMM framework. Section 6 reports simulation evidence and provides some practical guidance. The last section concludes. Proofs are provided in the Appendix.

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 and 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 be different at different occurrences. For two random variables X and 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.

2 A 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 $t = 1, ..., T$ where $\theta_0 \in \mathbb{R}^d$ is the parameter of interest and the vector process $u_t := (u_{1t}', u_{2t}')'$ is stationary with mean zero. We allow u_t to have autocorrelation of unknown forms so that the

long run variance Ω of u_t :

$$\Omega = lrvar(u_t) = \sum_{j=-\infty}^{\infty} Eu_t u_{t-j}'$$

takes a general form. However, for simplicity, we assume that $var(u_t) = \sigma^2 I_{d+q}$ for the moment[2]. Our model is just a location model. We initially consider a general GMM framework but later find out that our points can be made more clearly in 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.

Embedding the location model in a GMM framework, the moment conditions are

$$E(y_t) - \begin{pmatrix} \theta_0 \\ \mathbf{0}_{q\times 1} \end{pmatrix} = 0,$$

where $y_t = (y_{1t}', y_{2t}')'$. Let

$$g_T(\theta) = \begin{pmatrix} \frac{1}{\sqrt{T}} \sum_{t=1}^{T} (y_{1t} - \theta) \\ \frac{1}{\sqrt{T}} \sum_{t=1}^{T} y_{2t} \end{pmatrix}.$$

Then a GMM estimator of θ_0 can be defined as

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

for some positive definite weighting matrix W_T. Writing

$$W_T = \begin{pmatrix} W_{11} & W_{12} \\ W_{21} & W_{22} \end{pmatrix},$$

where W_{11} is a $d \times d$ matrix and W_{22} is a $q \times q$ matrix, then it is easy to show that

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

There are at least two different choices of W_T. First, we can take W_T to be the identity matrix $W_T = I_m$ for $m = d + q$. In this case, $\beta_W = 0$ and the GMM estimator $\hat{\theta}_{1T}$ is simply

$$\hat{\theta}_{1T} = \frac{1}{T} \sum_{t=1}^{T} y_{1t}.$$

Second, we can take W_T to be the 'optimal' weighting matrix $W_T = \Omega$. With this choice, we obtain the GMM estimator:

$$\tilde{\theta}_{2T} = \frac{1}{T} \sum_{t=1}^{T} (y_{1t} - \beta y_{2t}),$$

---

[2] If

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

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 $\mathbb{V}_{1/2}^{-1} (y_{1t}', y_{2t}')'$ 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.

where $\beta = \Omega_{12}\Omega_{22}^{-1}$ is the long run regression coefficient matrix. While $\hat{\theta}_{1T}$ completely ignores the information in $\{y_{2t}\}$, $\tilde{\theta}_{2T}$ takes advantage of this source of information.

Under some moment and mixing conditions, we have

$$\sqrt{T}\left(\hat{\theta}_{1T} - \theta_0\right) \Longrightarrow N(0, \Omega_{11}) \text{ and } \sqrt{T}\left(\tilde{\theta}_{2T} - \theta_0\right) \Longrightarrow N\left(0, \Omega_{1\cdot 2}\right),$$

where

$$\Omega_{1\cdot 2} = \Omega_{11} - \Omega_{12}\Omega_{22}^{-1}\Omega_{21}.$$

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

i) Estimate the long run covariance matrix by

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

where $\hat{u}_t = (y_{1t}' - \hat{\theta}_{1T}', y_{2t}')'$.

ii) Obtain the feasible two-step estimator $\hat{\theta}_{2T} = T^{-1}\sum_{t=1}^{T}(y_{1t} - \hat{\beta}y_{2t})$ where $\hat{\beta} = \hat{\Omega}_{12}\hat{\Omega}_{22}^{-1}$.

In the above definition of $\hat{\Omega}$, $Q_h\left(r, s\right)$ is a symmetric weighting function that depends on the smoothing parameter h. For conventional kernel LRV estimators, $Q_h\left(r, s\right) = k\left(\left(r - s\right)/b\right)$ and we take $h = 1/b$. For the orthonormal series (OS) LRV estimators, $Q_h\left(r, s\right) = K^{-1}\sum_{j=1}^{K}\phi_j\left(r\right)\phi_j\left(s\right)$ and we take $h = K$, where $\left\{\phi_j\left(r\right)\right\}$ are orthonormal basis functions on $L^2[0,1]$ satisfying $\int_0^1 \phi_j\left(r\right)dr = 0$. We parametrize h in such a way so that h indicates the level or amount of smoothing for both types of LRV estimators.

Note that we use the demeaned process $\{\hat{u}_t - T^{-1}\sum_{\tau=1}^{T}\hat{u}_\tau\}$ in constructing $\hat{\Omega}\left(\hat{u}\right)$. For the location model, $\hat{\Omega}\left(\hat{u}\right)$ is numerically identical to $\hat{\Omega}\left(u\right)$ where the unknown error process $\{u_t\}$ is used. The moment estimation uncertainty is reflected in the demeaning operation. Had we known the true value of θ_0 and hence the true moment process $\{u_t\}$, we would not need to demean $\{u_t\}$.

While $\tilde{\theta}_{2T}$ is asymptotically more efficient than $\hat{\theta}_{1T}$, is $\hat{\theta}_{2T}$ necessarily more efficient than $\hat{\theta}_{1T}$ and in what sense? Is the Wald test based on $\hat{\theta}_{2T}$ necessary more powerful than that based on $\hat{\theta}_{1T}$? One of the objectives of this paper is to address these questions.

3 A Tale of Two Asymptotics: Point Estimation

We first consider the conventional asymptotics where $h \to \infty$ as $T \to \infty$ but at a slower rate, i.e., $h/T \to 0$. Sun (2014a, 2014b) calls this type of asymptotics the "Increasing-smoothing Asymptotics," as h increases with the sample size. Under this type of asymptotics and some regularity conditions, we have $\hat{\Omega} \xrightarrow{p} \Omega$. It can then be shown that $\hat{\theta}_{2T}$ is asymptotically equivalent to $\tilde{\theta}_{2T}$, i.e., $\sqrt{T}(\tilde{\theta}_{2T} - \hat{\theta}_{2T}) = o_p\left(1\right)$. As a direct consequence, we have

$$\sqrt{T}\left(\hat{\theta}_{1T} - \theta_0\right) \Longrightarrow N(0, \Omega_{11}), \sqrt{T}\left(\hat{\theta}_{2T} - \theta_0\right) \Longrightarrow N\left[0, \Omega_{11} - \Omega_{12}\Omega_{22}^{-1}\Omega_{21}\right].$$

So $\hat{\theta}_{2T}$ is still asymptotically more efficient than $\hat{\theta}_{1T}$.

The conventional asymptotics, as elegant and convenient as it is, 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 $\sqrt{T}(\hat{\theta}_{2T} - \theta_0)$, we could develop a high order increasing-smoothing asymptotics that reflects the estimation uncertainty in $\hat{\Omega}$. This is possible but requires strong assumptions that cannot be easily verified. In addition, it is also technically challenging and tedious to rigorously justify the high order asymptotic theory. Instead of high order asymptotic theory under the conventional asymptotics, we adopt the type of asymptotics that holds h fixed (at a positive value) as $T \to \infty$. Given that h is fixed, we follow Sun (2014a, 2014b) and call this type of asymptotics the "Fixed-smoothing Asymptotics." This type of asymptotics takes the sampling variability of $\hat{\Omega}$ into consideration.

Sun (2013, 2014a) has shown that critical values from the fixed-smoothing asymptotic distribution are higher order correct under the conventional increasing-smoothing asymptotics. So the fixed-smoothing asymptotics can be regarded as a convenient device to obtain some higher order terms under the conventional increasing-smoothing asymptotics.

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

Assumption 1 (*i*) *For kernel LRV estimators, the kernel function* $k(\cdot)$ *satisfies the following conditions: for any* $b \in (0,1]$, $k_b(x) = k(x/b)$ *is symmetric, continuous, piecewise monotonic, and piecewise continuously differentiable on* $[-1,1]$. (*ii*) *For the OS LRV variance estimator, the basis functions* $\phi_j(\cdot)$ *are piecewise monotonic, continuously differentiable and orthonormal in* $L^2[0,1]$ *and* $\int_0^1 \phi_j(x)\,dx = 0$.

Assumption 1 on the kernel function is very mild. It includes many commonly used kernel functions such as the Bartlett kernel, Parzen kernel, and Quadratic Spectral (QS) kernel.

Define

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

which is a centered version of $Q_h(r,s)$, and

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

Assumption 1 ensures that $\tilde{\Omega}$ and $\hat{\Omega}$ are asymptotically equivalent. Furthermore, under this assumption, Sun (2014a) shows that, for both kernel LRV and OS LRV estimation, the centered weighting function $Q_h^*(r,s)$ satisfies :

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

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 converges to $Q_h^*(r,s)$ absolutely and uniformly over $(r,s) \in$

$[0,1] \times [0,1]$. The representation can be regarded as a spectral decomposition of the compact Fredholm operator with kernel $Q_h^*(r,s)$. See Sun (2014a) for more discussion.

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 *The vector process $\{u_t\}_{t=1}^T$ satisfies:*

(i) $T^{-1/2}\sum_{t=1}^T \Phi_j(t/T)u_t$ converges weakly to a continuous distribution, jointly over $j = 0, 1, \ldots, J$ for every fixed J;

(ii) For every fixed J and $x \in \mathbb{R}^m$,

$$P\left(\frac{1}{\sqrt{T}}\sum_{t=1}^{T}\Phi_j(\frac{t}{T})u_t \leq x \text{ for } j = 0,1,\ldots,J\right)$$
$$= P\left(\Omega_{1/2}\frac{1}{\sqrt{T}}\sum_{t=1}^{T}\Phi_j(\frac{t}{T})e_t \leq x \text{ for } j = 0,1,\ldots,J\right) + o(1) \text{ as } T \to \infty$$

where

$$\Omega_{1/2} = \begin{pmatrix} \Omega_{1\cdot 2}^{1/2} & \Omega_{12}\Omega_{22}^{-1/2} \\ 0 & \Omega_{22}^{1/2} \end{pmatrix} > 0$$

is a matrix square root of the nonsingular LRV matrix $\Omega = \sum_{j=-\infty}^{\infty} Eu_t u_{t-j}'$ and $e_t \sim iid\ N(0, I_m)$.

Assumption 3 $\sum_{j=-\infty}^{\infty} \| Eu_t u_{t-j}' \| < \infty$.

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

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

$$\Omega_\infty = \Omega_{1/2}\tilde{\Omega}_\infty\Omega_{1/2}' := \begin{pmatrix} \Omega_{\infty,11} & \Omega_{\infty,12} \\ \Omega_{\infty,21} & \Omega_{\infty,22} \end{pmatrix}$$
$$\tilde{\Omega}_\infty = \int_0^1\int_0^1 Q_h^*(r,s)dB_m(r)dB_m(s)' := \begin{pmatrix} \tilde{\Omega}_{\infty,11} & \tilde{\Omega}_{\infty,12} \\ \tilde{\Omega}_{\infty,21} & \tilde{\Omega}_{\infty,22} \end{pmatrix}$$

and $B_m(\cdot)$ is a standard Brownian motion of dimension $m = d + q$;

(b) $\sqrt{T}\left(\hat{\theta}_{2T} - \theta_0\right) \xrightarrow{d} \begin{pmatrix} I_d, & -\beta_\infty \end{pmatrix}\Omega_{1/2}B_m(1)$ where $\beta_\infty = \beta_\infty(h,d,q) := \Omega_{\infty,12}\Omega_{\infty,22}^{-1}$ is independent of $B_m(1)$.

Conditional on β_∞, the asymptotic distribution of $\sqrt{T}(\hat{\theta}_{2T} - \theta_0)$ is a normal distribution with variance

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

Given that V_2 is random, $\sqrt{T}(\hat{\theta}_{2T} - \theta_0)$ is asymptotically mixed-normal rather than normal. Since

$$\begin{aligned} \text{avar}(\hat{\theta}_{2T}) - \text{avar}(\tilde{\theta}_{2T}) &= EV_2 - \left(\Omega_{11} - \Omega_{12}\Omega_{22}^{-1}\Omega_{21}\right) \\ &= E\left(\Omega_{12}\Omega_{22}^{-1}\Omega_{21} - \Omega_{12}\beta_\infty' - \beta_\infty\Omega_{21} + \beta_\infty\Omega_{22}\beta_\infty'\right) \\ &= E\left(\Omega_{12}\Omega_{22}^{-1} - \beta_\infty\right)\Omega_{22}\left(\Omega_{12}\Omega_{22}^{-1} - \beta_\infty\right)' \geq 0, \end{aligned}$$

the feasible estimator $\hat{\theta}_{2T}$ has a large variation than the infeasible estimator $\tilde{\theta}_{2T}$. This is consistent with our intuition. The difference $\text{avar}(\hat{\theta}_{2T}) - \text{avar}(\tilde{\theta}_{2T})$ can be regarded as the cost of implementing the two-step estimator, i.e., the cost of having to estimate the weighting matrix.

Under the fixed-smoothing asymptotics, we still have $\sqrt{T}(\hat{\theta}_{1T} - \theta_0) \xrightarrow{d} N(0, \Omega_{11})$, as $\hat{\theta}_{1T}$ does not depend on the smoothing parameter h. So

$$\text{avar}(\hat{\theta}_{1T}) - \text{avar}(\tilde{\theta}_{2T}) := \Omega_{11} - \left(\Omega_{11} - \Omega_{12}\Omega_{22}^{-1}\Omega_{21}\right) = \Omega_{12}\Omega_{22}^{-1}\Omega_{21} \geq 0,$$

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

To compare the asymptotic variances of $\sqrt{T}(\hat{\theta}_{1T} - \theta_0)$ and $\sqrt{T}(\hat{\theta}_{2T} - \theta_0)$, we need to evaluate the relative magnitudes of the cost and the benefit. Define

$$\tilde{\beta}_\infty := \tilde{\beta}_\infty(h, d, q) := \tilde{\Omega}_{\infty,12}\tilde{\Omega}_{\infty,22}^{-1}, \tag{2}$$

which does not depend on any nuisance parameter but depends on h, d, q. For notational economy, we sometimes suppress this dependence. Direct calculations show that

$$\beta_\infty = \Omega_{1\cdot 2}^{1/2}\tilde{\beta}_\infty\Omega_{22}^{-1/2} + \Omega_{12}\Omega_{22}^{-1}. \tag{3}$$

Using this, we have:

$$\begin{aligned} \text{avar}(\hat{\theta}_{2T}) - \text{avar}(\hat{\theta}_{1T}) &= \underbrace{\text{avar}(\hat{\theta}_{2T}) - \text{avar}(\tilde{\theta}_{2T})}_{\text{cost}} - \underbrace{[\text{avar}(\hat{\theta}_{1T}) - \text{avar}(\tilde{\theta}_{2T})]}_{\text{benefit}} \\ &= \Omega_{1\cdot 2}^{1/2}E\tilde{\beta}_\infty\tilde{\beta}_\infty'(\Omega_{1\cdot 2}^{1/2})' - \Omega_{12}\Omega_{22}^{-1}\Omega_{21}. \end{aligned} \tag{4}$$

If the cost is larger than the benefit, i.e., $\Omega_{1\cdot 2}^{1/2}E\tilde{\beta}_\infty\tilde{\beta}_\infty'(\Omega_{1\cdot 2}^{1/2})' > \Omega_{12}\Omega_{22}^{-1}\Omega_{21}$, then the asymptotic variance of $\hat{\theta}_{2T}$ is larger than that of $\hat{\theta}_{1T}$.

The following lemma gives a characterization of $E\tilde{\beta}_\infty(h, d, q)\,\tilde{\beta}_\infty(h, d, q)'$.

Lemma 2 *For any $d \geq 1$, we have $E\tilde{\beta}_\infty(h, d, q)\,\tilde{\beta}_\infty(h, d, q)' = \left(E||\tilde{\beta}_\infty(h, 1, q)||^2\right) \times I_d$.*

Using the lemma, we can prove that

$$\text{avar}(\hat{\theta}_{2T}) - \text{avar}(\hat{\theta}_{1T}) = (1 + E||\tilde{\beta}_\infty(h, 1, q)||^2)\Omega_{11}^{1/2}\left[g(h, q)I_d - \rho\rho'\right](\Omega_{11}^{1/2})',$$

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

$$\rho = \Omega_{11}^{-1/2}\Omega_{12}\Omega_{22}^{-1/2} \in \mathbb{R}^{d\times q},$$

which is the long run correlation matrix between u_{1t} and u_{2t}. The proposition below then follows immediately.

Proposition 3 *Let Assumptions 1–3 hold. Consider the fixed-smoothing asymptotics.*
(a) If $\nu_{\max}(\rho\rho') < g(h, q)$, then $\hat{\theta}_{2T}$ has a larger asymptotic variance than $\hat{\theta}_{1T}$.
(b) If $\nu_{\min}(\rho\rho') > g(h, q)$, then $\hat{\theta}_{2T}$ has a smaller asymptotic variance than $\hat{\theta}_{1T}$.

To compute the eigenvalues of $\rho\rho'$, we can use the fact that $\nu(\rho\rho') = \nu\left(\Omega_{12}\Omega_{22}^{-1}\Omega_{21}\Omega_{11}^{-1}\right)$. 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, i.e., the squared long run canonical correlation coefficients between u_{1t} and u_{2t}. So the conditions in the proposition can be presented in terms of the smallest and largest square long run canonical correlation coefficients.

If $\rho = 0$, then $\nu_{\max}(\rho\rho') < g(h,q)$ holds trivially. In this case, the asymptotic variance of $\hat{\theta}_{2T}$ is larger than the asymptotic variance of $\hat{\theta}_{1T}$. Intuitively, when the long run correlation is zero, there is no information that can be explored to improve efficiency. If we insist on using the long run correlation matrix in attempt to improve the efficiency, we may end up with a less efficient estimator, due to 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 $\nu_{\min}(\rho\rho') = 1$ and we have $\nu_{\min}(\rho\rho') > g(h,q)$. In this case, it is worthwhile estimating the long run variance and using it to improve the efficiency of the two-step GMM estimator.

The two conditions $\nu_{\min}(\rho\rho') > g(h,q)$ and $\nu_{\max}(\rho\rho') < g(h,q)$ in the proposition may appear to be strong. However, the conclusions are also very strong. For example, $\hat{\theta}_{2T}$ has a smaller asymptotic variance than $\hat{\theta}_{1T}$ means that $avar(R\hat{\theta}_{2T}) \leq \text{avar}(R\hat{\theta}_{1T})$ for *any* matrix $R \in \mathbb{R}^{p\times d}$ and for all $1 \leq p \leq d$. In fact, in the proof of the proposition, we show that the conditions are both necessary and sufficient.

The two conditions $\nu_{\min}(\rho\rho') > g(h,q)$ and $\nu_{\max}(\rho\rho') \leq g(h,q)$ are not mutually exclusive unless $d = 1$. When $d > 1$, it is possible that neither of two conditions is satisfied, in which case $\text{avar}(\hat{\theta}_{2T}) - \text{avar}(\hat{\theta}_{1T})$ is indefinite. So, as a whole vector, the relative asymptotic efficiency of $\hat{\theta}_{2T}$ to $\hat{\theta}_{1T}$ cannot be compared. However, there exist two 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 $\text{avar}(R^+\hat{\theta}_{2T}) \leq \text{avar}(R^+\hat{\theta}_{1T})$ and $\text{avar}(R^-\hat{\theta}_{2T}) \geq \text{avar}(R^-\hat{\theta}_{1T})$. An example of the indefinite case is when $q < d$ and $\nu_{\max}(\rho\rho') > g(h,q)$. In this case, $\nu_{\min}(\rho\rho') = 0$ and $\nu_{\min}(\rho\rho') > g(h,q)$ does not hold. A direct implication is that $\text{avar}(R^-\hat{\theta}_{2T}) > \text{avar}(R^-\hat{\theta}_{1T})$ for some R^-. So when the degree of overidentification is not large enough, there are some directions characterized by R^- along which the two-step estimator is less efficient than the one-step estimator.

When $d = 1$, $\rho\rho'$ is a scalar, and the two conditions $\nu_{\min}(\rho\rho') > g(h,q)$ and $\nu_{\max}(\rho\rho') \leq g(h,q)$ become mutually exclusive. So if $\rho\rho' > g(h,q)$, then $\hat{\theta}_{2T}$ is asymptotically more efficient than $\hat{\theta}_{1T}$. Otherwise, it is asymptotically less efficient.

In the case of kernel LRV estimation, it is hard to obtain an analytical expression for $E||\tilde{\beta}_\infty(h,1,q)||^2$ and hence $g(h,q)$, although we can always simulate $g(h,q)$ numerically. The threshold $g(h,q)$ depends on the smoothing parameter $h = 1/b$ and the degree of overidentification q. Tables 1–3 report the simulated values of $g(h,q)$ for $b = 0.00 : 0.01 : 0.20$ and $q = 1 \sim 5$. These values are nontrivial in that they are close to neither zero nor one. It is clear that $g(h,q)$ increases with q and decreases with the smoothing parameter $h = 1/b$.

When the OS LRV estimation is used, we do not need to simulate $g(h,q)$, as we can obtain a closed-form expression.

Corollary 4 *Let Assumptions 1–3 hold. In the case of OS LRV estimation, we have*

$$g(h,q) = \frac{q}{K-1}.$$

So if $\nu_{\max}(\rho\rho') < \frac{q}{K-1}$ (or $\nu_{\min}(\rho\rho') > \frac{q}{K-1}$), then $\hat{\theta}_{2T}$ has a larger (or smaller) asymptotic variance than $\hat{\theta}_{1T}$ under the fixed-smoothing asymptotics.

Since $\hat{\theta}_{2T}$ is not asymptotically normal, asymptotic variance comparison does not paint the whole picture. To compare the asymptotic distributions of $\hat{\theta}_{1T}$ and $\hat{\theta}_{2T}$, we consider the case of OS LRV estimation with $d = q = 1$ and $K = 4$ as an example. We use the sine and cosine basis functions as given in (**??**) later in Section 6. Figure 1 reports the shapes of probability density functions when $(\Omega_{11}, \Omega_{12}^2, \Omega_{22}) = (1, 0.10, 1)$. In this case, $\Omega_{1\cdot 2} = \Omega_{11} - \Omega_{12}\Omega_{22}^{-1}\Omega_{21} = 0.9$. The first graph shows $\sqrt{T}(\hat{\theta}_{1T} - \theta_0) \overset{a}{\sim} N(0, 1)$ and $\sqrt{T}(\hat{\theta}_{2T} - \theta_0) \overset{a}{\sim} N(0, 0.9)$ under the conventional asymptotics. The conventional limiting distributions for $\sqrt{T}(\hat{\theta}_{1T} - \theta_0)$ and $\sqrt{T}(\hat{\theta}_{2T} - \theta_0)$ are both normal but the latter has a smaller variance, so the asymptotic efficiency of $\hat{\theta}_{2T}$ is always guaranteed. However, this is not true in the second graph of Figure 1, which represents the limiting distributions under the fixed-smoothing asymptotics. While we still have $\sqrt{T}(\hat{\theta}_{1T} - \theta_0) \overset{a}{\sim} N(0, 1)$, $\sqrt{T}(\hat{\theta}_{2T} - \theta_0) \overset{a}{\sim} MN[0, 0.9(1 + \tilde{\beta}_\infty^2)]$. The mixed normality can be obtained by using a conditional version of (4). More specifically, the conditional asymptotic variance of $\hat{\theta}_{2T}$ is

$$\text{avar}(\hat{\theta}_{2T}|\tilde{\beta}_\infty) = V_2 = \Omega_{1\cdot 2}^{1/2}\tilde{\beta}_\infty\tilde{\beta}_\infty'(\Omega_{1\cdot 2}^{1/2})' + \Omega_{1\cdot 2} = 0.9(1 + \tilde{\beta}_\infty^2). \tag{5}$$

Comparing these two different families of distributions, we find that the asymptotic distribution of $\hat{\theta}_{2T}$ has fatter tail than that of $\hat{\theta}_{1T}$. The asymptotic variance of $\hat{\theta}_{2T}$ is

$$\text{avar}(\hat{\theta}_{2T}) = EV_2 = \Omega_{1\cdot 2}\{1 + E[||\tilde{\beta}_\infty(h, 1, q)||^2]\} = \Omega_{1\cdot 2}\frac{K-1}{K-q-1} = 0.9 \times \frac{3}{2} = 1.35,$$

which is larger than the asymptotic variance of $\hat{\theta}_{1T}$.

4 A Tale of Two Asymptotics: Hypothesis Testing

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

$$\mathbb{W}_{1T} := T(R\hat{\theta}_{1T} - r)'\left(R\hat{\Omega}_{11}R'\right)^{-1}(R\hat{\theta}_{1T} - r)$$
$$\mathbb{W}_{2T} := T(R\hat{\theta}_{2T} - r)'\left(R\hat{\Omega}_{1\cdot 2}R'\right)^{-1}(R\hat{\theta}_{2T} - r)$$

where $\hat{\Omega}_{1\cdot 2} = \hat{\Omega}_{11} - \hat{\Omega}_{12}\hat{\Omega}_{22}^{-1}\hat{\Omega}_{21}$. When $p = 1$ and the alternative is one sided, we can construct the following two t statistics:

$$\mathbb{T}_{1T} \quad := \quad \frac{\sqrt{T}\left(R\hat{\theta}_{1T} - r\right)}{\sqrt{R\hat{\Omega}_{11}R'}} \tag{6}$$

$$\mathbb{T}_{2T} \quad := \quad \frac{\sqrt{T}\left(R\hat{\theta}_{2T} - r\right)}{\sqrt{R\hat{\Omega}_{1\cdot 2}R'}}. \tag{7}$$

No matter whether the test is based on $\hat{\theta}_{1T}$ or $\hat{\theta}_{2T}$, we have to employ the long run covariance estimator $\hat{\Omega}$. Define the $p \times p$ matrices Λ_1 and Λ_2 according to

$$\Lambda_1\Lambda_1' = R\Omega_{11}R' \text{ and } \Lambda_2\Lambda_2' = R\Omega_{1\cdot 2}R'.$$

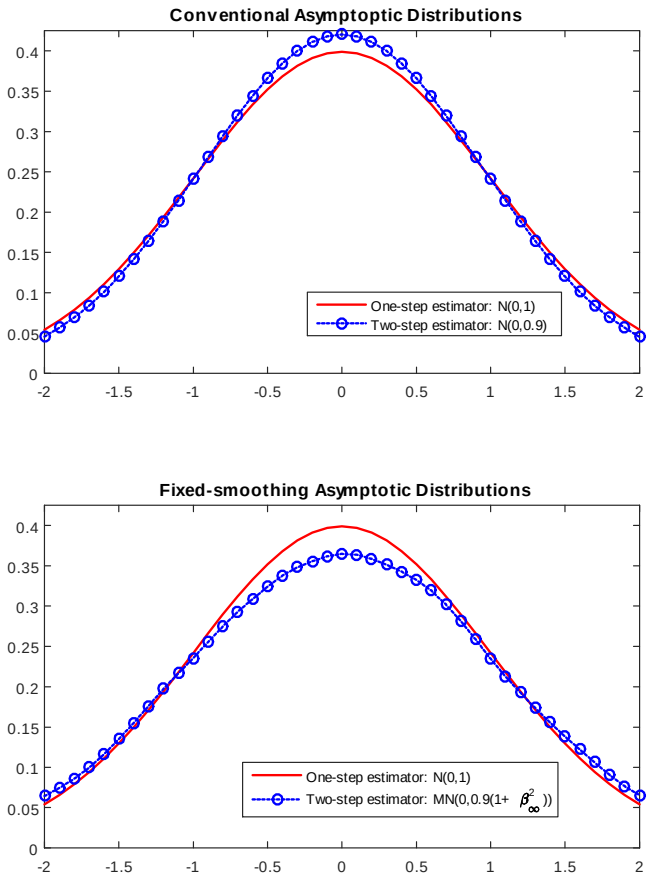

Figure 1: Limiting distributions of $\hat{\theta}_{1T}$ and $\hat{\theta}_{2T}$ based on the OS LRV estimator with $K = 4$.

In other words, Λ_1 and Λ_2 are matrix square roots of $R\Omega_{11}R'$ and $R\Omega_{1\cdot 2}R'$ respectively.

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

$$\mathbb{W}_{1T} \overset{d}{\Longrightarrow} \chi_p^2(\left\|\Lambda_1^{-1}\delta_0\right\|^2),\ \mathbb{W}_{2T} \overset{d}{\Longrightarrow} \chi_p^2(\left\|\Lambda_2^{-1}\delta_0\right\|^2),$$
$$\mathbb{T}_{1T} \overset{d}{\Longrightarrow} N(\Lambda_1^{-1}\delta_0, 1),\ \mathbb{T}_{2T} \overset{d}{\Longrightarrow} N(\Lambda_2^{-1}\delta_0, 1),$$

where $\chi_p^2\left(\lambda^2\right)$ is the noncentral chi-square distribution with noncentrality parameter λ^2. When $\delta_0 = 0$, we obtain the null distributions:

$$\mathbb{W}_{1T}, \mathbb{W}_{2T} \overset{d}{\Longrightarrow} \chi_p^2 \text{ and } \mathbb{T}_{1T}, \mathbb{T}_{2T} \overset{d}{\Longrightarrow} N(0,1).$$

So under the conventional increasing-smoothing asymptotics, the null limiting distributions of $\mathbb{W}_{1T}$ and $\mathbb{W}_{2T}$ are identical. Since $\left\|\Lambda_1^{-1}\delta_0\right\|^2 \leq \left\|\Lambda_2^{-1}\delta_0\right\|^2$, under the conventional asymptotics, the local asymptotic power function of the test based on $\mathbb{W}_{2T}$ is higher than that based on $\mathbb{W}_{1T}$.

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 section, we employ the fixed-smoothing asymptotics to derive more accurate distributional approximations. Let

$$C_{pp} = \int_0^1 \int_0^1 Q_h^*(r,s) dB_p(r) dB_p(s)', C_{pq} = \int_0^1 \int_0^1 Q_h^*(r,s) dB_p(r) dB_q(s)'$$
$$C_{qq} = \int_0^1 \int_0^1 Q_h^*(r,s) dB_q(r) dB_q(s)', C_{qp} = C_{pq}'$$

and

$$D_{pp} = C_{pp} - C_{pq} C_{qq}^{-1} C_{pq}'$$

where $B_p(\cdot) \in \mathbb{R}^p$ and $B_q(\cdot) \in \mathbb{R}^q$ are independent standard Brownian motion processes.

Proposition 5 *Let Assumptions 1–3 hold. As $T \to \infty$ for a fixed h, we have, under $H_1 : R\theta_0 = r + \delta_0/\sqrt{T}$:*

(a) $\mathbb{W}_{1T} \xrightarrow{d} \mathbb{W}_{1\infty}(\left\|\Lambda_1^{-1}\delta_0\right\|^2)$ *where*

$$\mathbb{W}_{1\infty}(\|\xi\|^2) = [B_p(1) + \xi]' C_{pp}^{-1} [B_p(1) + \xi] \text{ for } \xi \in \mathbb{R}^p. \tag{8}$$

(b) $\mathbb{W}_{2T} \xrightarrow{d} \mathbb{W}_{2\infty}(\left\|\Lambda_2^{-1}\delta_0\right\|^2)$ *where*

$$\mathbb{W}_{2\infty}(\|\xi\|^2) = \left[B_p(1) - C_{pq}C_{qq}^{-1}B_q(1) + \xi\right]' D_{pp}^{-1} \left[B_p(1) - C_{pq}C_{qq}^{-1}B_q(1) + \xi\right]. \tag{9}$$

(c) $\mathbb{T}_{1T} \xrightarrow{d} \mathbb{T}_{1\infty}\left(\Lambda_1^{-1}\delta_0\right) := \left[B_p(1) + \Lambda_1^{-1}\delta_0\right]/\sqrt{C_{pp}}$ *for* $p = 1$.
(d) $\mathbb{T}_{2T} \xrightarrow{d} \mathbb{T}_{2\infty}\left(\Lambda_2^{-1}\delta_0\right) := \left[B_p(1) - C_{pq}C_{qq}^{-1}B_q(1) + \Lambda_2^{-1}\delta_0\right]/\sqrt{D_{pp}}$ *for* $p = 1$.

In Proposition 5, we use the notation $\mathbb{W}_{1\infty}(\|\xi\|^2)$, which implies that the right hand side of (8) depends on ξ only through $\|\xi\|^2$. This is true, because for any orthogonal matrix H :

$$[B_p(1) + \xi]' C_{pp}^{-1} [B_p(1) + \xi] = [HB_p(1) + H\xi]' HC_{pp}^{-1}H' [HB_p(1) + H\xi]$$
$$\stackrel{d}{=} [B_p(1) + H\xi]' C_{pp}^{-1} [B_p(1) + H\xi].$$

If we choose $H = (\xi/\|\xi\|, \tilde{H})'$ for some $\tilde{H}$ such that H is orthogonal, then

$$[B_p(1) + \xi]' C_{pp}^{-1} [B_p(1) + \xi] \stackrel{d}{=} [B_p(1) + \|\xi\| e_p]' C_{pp}^{-1} [B_p(1) + \|\xi\| e_p],$$

where $e_p = (1, 0, ..., 0)' \in \mathbb{R}^p$. So the distribution of $[B_p(1) + \xi]' C_{pp}^{-1} [B_p(1) + \xi]$ depends on ξ only through $\|\xi\|$. Similarly, the distribution of the right hand side of (9) depends only on $\|\xi\|^2$.

When $\delta_0 = 0$, we obtain the limiting distributions of $\mathbb{W}_{1T}, \mathbb{W}_{2T}, \mathbb{T}_{1T}$ and $\mathbb{T}_{2T}$ under the null hypothesis:

$$\mathbb{W}_{1T} \xrightarrow{d} \mathbb{W}_{1\infty} := \mathbb{W}_{1\infty}(0) = B_p(1)' C_{pp}^{-1} B_p(1),$$
$$\mathbb{W}_{2T} \xrightarrow{d} \mathbb{W}_{2\infty} := \mathbb{W}_{2\infty}(0) = \left[B_p(1) - C_{pq}C_{qq}^{-1}B_q(1)\right]' D_{pp}^{-1} \left[B_p(1) - C_{pq}C_{qq}^{-1}B_q(1)\right],$$
$$\mathbb{T}_{1T} \xrightarrow{d} \mathbb{T}_{1\infty} := \mathbb{T}_{1\infty}(0) = B_p(1)/\sqrt{C_{pp}},$$
$$\mathbb{T}_{2T} \xrightarrow{d} \mathbb{T}_{2\infty} := \mathbb{T}_{2\infty}(0) = \left[B_p(1) - C_{pq}C_{qq}^{-1}B_q(1)\right]/\sqrt{D_{pp}}.$$

These distributions are different from those under the conventional asymptotics. For $\mathbb{W}_{1T}$ and $\mathbb{T}_{1T}$, the difference lies in the random scaling factor C_{pp} or $\sqrt{C_{pp}}$. The random scaling factor captures the estimation uncertainty of the LRV estimator. For $\mathbb{W}_{2T}$ and $\mathbb{T}_{2T}$, there is an additional difference embodied by the random location shift $C_{pq}C_{qq}^{-1}B_q(1)$ with a consequent change in the random scaling factor.

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

Proposition 6 *For any $x > 0$, the following hold:*

(*a*) $\mathbb{W}_{2\infty}(0)$ *first-order stochastically dominates* $\mathbb{W}_{1\infty}(0)$ *in that*

$$P\left[\mathbb{W}_{2\infty}(0) \geq x\right] > P\left[\mathbb{W}_{1\infty}(0) \geq x\right].$$

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

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

Proposition 6(a) is intuitive. $\mathbb{W}_{2\infty}$ first-order stochastically dominates $\mathbb{W}_{1\infty}$ because $\mathbb{W}_{2\infty}$ first-order stochastically dominates $B_p(1)' D_{pp}^{-1} B_p(1)$, which in turn first-order stochastically $B_p(1)' C_{pp}^{-1} B_p(1)$, which is just $\mathbb{W}_{1\infty}$. According to 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 is that the asymptotic critical values for $\mathbb{W}_{2T}$ are larger than the corresponding ones for $\mathbb{W}_{1T}$. The difference in critical values has implications on the power properties of the two tests.

For $x > 0$, we have

$$P\left(\mathbb{T}_{1\infty} > x\right) = \frac{1}{2} P\left(\mathbb{W}_{1\infty} \geq x^2\right) \text{ and } P\left(\mathbb{T}_{2\infty} > x\right) = \frac{1}{2} P\left(\mathbb{W}_{2\infty} \geq x^2\right).$$

It then follows from Proposition 6(a) that $P\left(\mathbb{T}_{2\infty} > x\right) \geq P\left(\mathbb{T}_{1\infty} > x\right)$ for $x > 0$. So for a one-sided test with the alternative $H_1 : R\theta_0 > r$, critical values from $\mathbb{T}_{2\infty}$ are larger than those from $\mathbb{T}_{1\infty}$. Similarly, we have $P\left(\mathbb{T}_{2\infty} < x\right) \geq P\left(\mathbb{T}_{1\infty} < x\right)$ for $x < 0$. This implies that for a one-sided test with the alternative $H_1 : R\theta_0 < r$, critical values from $\mathbb{T}_{2\infty}$ are smaller than those from $\mathbb{T}_{1\infty}$.

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

$$\pi_1\left(\left\|\Lambda_1^{-1}\delta_0\right\|^2\right) := \pi_1\left(\left\|\Lambda_1^{-1}\delta_0\right\|^2; h, p, q, \alpha\right) = P\left[\mathbb{W}_{1\infty}(\left\|\Lambda_1^{-1}\delta_0\right\|^2) > \mathbb{W}_{1\infty}^{\alpha}\right],$$
$$\pi_2\left(\left\|\Lambda_2^{-1}\delta_0\right\|^2\right) := \pi_2\left(\left\|\Lambda_1^{-1}\delta_0\right\|^2; h, p, q, \alpha\right) = P\left[\mathbb{W}_{2\infty}(\left\|\Lambda_2^{-1}\delta_0\right\|^2) > \mathbb{W}_{2\infty}^{\alpha}\right].$$

While $\left\|\Lambda_2^{-1}\delta_0\right\|^2 \geq \left\|\Lambda_1^{-1}\delta_0\right\|^2$, we also have $\mathbb{W}_{2\infty}^{\alpha} > \mathbb{W}_{1\infty}^{\alpha}$. 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 6 suggests that if the difference in the noncentrality parameters $\left\|\Lambda_2^{-1}\delta_0\right\|^2 - \left\|\Lambda_1^{-1}\delta_0\right\|^2$ is large enough to offset the increase in critical values, then the two-step test based on $\mathbb{W}_{2T}$ will be more powerful.

To evaluate $\left\|\Lambda_2^{-1}\delta_0\right\|^2 - \left\|\Lambda_1^{-1}\delta_0\right\|^2$, we define

$$\rho_R = \left(R\Omega_{11}R'\right)^{-1/2}\left(R\Omega_{12}\right)\Omega_{22}^{-1/2}, \tag{10}$$

which is the long run correlation matrix ρ_R between Ru_{1t} and u_{2t}. In terms of $\rho_R \in \mathbb{R}^{p\times q}$ we have

$$\begin{aligned}
&\left\|\Lambda_2^{-1}\delta_0\right\|^2 - \left\|\Lambda_1^{-1}\delta_0\right\|^2 \\
&= \delta_0'\left(R\Omega_{11}R' - R\Omega_{12}\Omega_{22}^{-1}\Omega_{21}R'\right)^{-1}\delta_0 - \delta_0'\left(R\Omega_{11}R'\right)^{-1}\delta_0 \\
&= \delta_0'\left(\Lambda_1'\right)^{-1}\left[I_p - \Lambda_1^{-1}R\Omega_{12}\Omega_{22}^{-1}\Omega_{21}R'\left(\Lambda_1'\right)^{-1}\right]^{-1}\left(\Lambda_1^{-1}\delta_0\right) - \delta_0'\left(\Lambda_1'\right)^{-1}\left(\Lambda_1^{-1}\delta_0\right) \\
&= \delta_0'\left(\Lambda_1'\right)^{-1}\left\{\left[I_p - \rho_R\rho_R'\right]^{-1} - I_p\right\}\left(\Lambda_1^{-1}\delta_0\right).
\end{aligned}$$

So the difference in the noncentrality parameters depends on the matrix $\rho_R\rho_R'$.

Let $\rho_R\rho_R' = \sum_{i=1}^p \nu_{i,R}a_{i,R}a_{i,R}'$ be the eigen decomposition of $\rho_R\rho_R'$, where $\{\nu_{i,R}\}$ are the eigenvalues of $\rho_R\rho_R'$ and $\{a_{i,R}\}$ are the corresponding eigenvectors. Sorted in the descending order, $\{\nu_{i,R}\}$ are the (squared) long run canonical correlation coefficients between Ru_{1t} and u_{2t}. Then

$$\left\|\Lambda_2^{-1}\delta_0\right\|^2 - \left\|\Lambda_1^{-1}\delta_0\right\|^2 = \sum_{i=1}^p \frac{\nu_{i,R}}{1-\nu_{i,R}}\left[a_{i,R}'\Lambda_1^{-1}\delta_0\right]^2.$$

Consider a special case that $\nu_{p,R} := \min_{i=1}^p\{\nu_{i,R}\}$ approaches 1. If $a_{p,R}'\Lambda_1^{-1}\delta_0 \neq 0$, then $||\Lambda_2^{-1}\delta_0||^2 - \left\|\Lambda_1^{-1}\delta_0\right\|^2$ and hence $||\Lambda_2^{-1}\delta_0||^2$ approaches ∞ as $\nu_{p,R}$ approaches 1 from below. This case happens when the second block of moment conditions has very high long run prediction power for the first block. In this case, we expect the $\mathbb{W}_{2T}$ test to be more powerful, as $\lim_{\nu_{p,R}\to 1}\pi_2(\left\|\Lambda_2^{-1}\delta_0\right\|^2) = 1$. Consider another special case that $\max_{i=1}^p\{\nu_{i,R}\} = 0$, i.e., ρ_R is a matrix of zeros. In this case, the second block of moment conditions contains no additional information, and we have $\left\|\Lambda_2^{-1}\delta_0\right\|^2 = \left\|\Lambda_1^{-1}\delta_0\right\|^2$. In this case, we expect the $\mathbb{W}_{2T}$ test to be less powerful.

It follows from Proposition 6(b) and (c) that for any λ, there exists a unique $\tau(\lambda) := \tau(\lambda; h, p, q, \alpha)$ such that

$$\pi_2(\lambda) = \pi_1\left(\frac{\lambda}{\tau}\right).$$

As a function of λ, $\tau(\lambda)$ is defined implicitly via the above equation. Then $\pi_2(\left\|\Lambda_2^{-1}\delta_0\right\|^2) < \pi_1(\left\|\Lambda_1^{-1}\delta_0\right\|^2)$ if and only if $\left\|\Lambda_2^{-1}\delta_0\right\|^2 < \tau(\left\|\Lambda_2^{-1}\delta_0\right\|^2)\cdot\left\|\Lambda_1^{-1}\delta_0\right\|^2$. Using

$$\begin{aligned}
&\left\|\Lambda_2^{-1}\delta_0\right\|^2 - \tau(\left\|\Lambda_2^{-1}\delta_0\right\|^2)\left\|\Lambda_1^{-1}\delta_0\right\|^2 \\
= &\sum_{i=1}^p\left(\frac{1}{1-\nu_{i,R}} - \tau(\left\|\Lambda_2^{-1}\delta_0\right\|^2)\right)\left[a_{i,R}'\Lambda_1^{-1}\delta_0\right]^2 \\
= &\sum_{i=1}^p\frac{1}{1-\nu_{i,R}}\left(\nu_{i,R} - \frac{\tau(\left\|\Lambda_2^{-1}\delta_0\right\|^2) - 1}{\tau(\left\|\Lambda_2^{-1}\delta_0\right\|^2)}\right)\left[a_{i,R}'\Lambda_1^{-1}\delta_0\right]^2\tau(\left\|\Lambda_2^{-1}\delta_0\right\|^2) \\
= &\sum_{i=1}^p\frac{1}{1-\nu_{i,R}}\left(\nu_{i,R} - f(\left\|\Lambda_2^{-1}\delta_0\right\|^2)\right)\left[a_{i,R}'\Lambda_1^{-1}\delta_0\right]^2\tau(\left\|\Lambda_2^{-1}\delta_0\right\|^2)
\end{aligned} \tag{11}$$

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

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

we can prove the proposition below.

Proposition 7 *Let Assumptions 1–3 hold. Define*

$$\mathfrak{A}(\lambda_0) = \{\delta : \delta' \left(R\Omega_{1.2}R'\right)^{-1} \delta = \lambda_0\}.$$

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 $\nu_{\max}(\rho_R\rho_R') < f(\lambda_0; h, p, q, \alpha)$*, then the two-step test based on* $\mathbb{W}_{2T}$ *has a lower local asymptotic power than the one-step test based on* $\mathbb{W}_{1T}$ *for any* $\delta_0 \in \mathfrak{A}(\lambda_0)$*.*

(b) If $\nu_{\min}(\rho_R\rho_R') > f(\lambda_0; h, p, q, \alpha)$*, then the two-step test based on* $\mathbb{W}_{2T}$ *has a higher local asymptotic power than the one-step test based on* $\mathbb{W}_{1T}$ *for any* $\delta_0 \in \mathfrak{A}(\lambda_0)$*.*

To compute $\nu_{\max}(\rho_R\rho_R')$ and $\nu_{\min}(\rho_R\rho_R')$, we can use the relationship that

$$\nu\left(\rho_R\rho_R'\right) = \nu\left\{\left(R\Omega_{12}\Omega_{22}^{-1}\Omega_{21}R'\right)\left(R\Omega_{11}R'\right)^{-1}\right\}.$$

There is no need to compute the matrix square roots $(R\Omega_{11}R')^{-1/2}$ and $\Omega_{22}^{-1/2}$.

As in the case of variance comparison, the conditions on the canonical correlation coefficients in Proposition 7(a) and (b) are both sufficient and necessary. See the proof of the proposition for details. The conditions may appear to be strong but the conclusions are equally strong — the power comparison results hold regardless the directions of the local departure. If we have a particular direction in mind so that δ_0 is fixed and given, then we can evaluate $\left\|\Lambda_2^{-1}\delta_0\right\|^2 - \tau\left(\Lambda_2^{-1}\delta_0\right)\left\|\Lambda_1^{-1}\delta_0\right\|^2$ directly for the given δ_0. If $\left\|\Lambda_2^{-1}\delta_0\right\|^2 - \tau\left(\Lambda_2^{-1}\delta_0\right)\left\|\Lambda_1^{-1}\delta_0\right\|^2$ is positive (negative), then the two-step test has a higher (lower) local asymptotic power along the given direction.

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. 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. More specifically, 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 less powerful.

Proposition 7 is in parallel with Proposition 3. The qualitative messages of these two propositions are the same — when the long run correlation is high enough, we should estimate and exploit it to reduce the variation of our point estimator and improve the power of the associated tests. However, the thresholds are different quantitatively. The two propositions fully characterize the threshold for each criterion under consideration.

Proposition 8 *Consider the case of OS LRV estimation. For any* $\lambda \in \mathbb{R}^+$*, we have* $\pi_1(\lambda) > \pi_2(\lambda)$ *and hence* $\tau(\lambda; h, p, q, \alpha) > 1$ *and* $f(\lambda; h, p, q, \alpha) > 0$*.*

Proposition 8 is intuitive. When there is no long run correlation between Ru_{1t} and u_{2t}, we have $\left\|\Lambda_2^{-1}\delta_0\right\|^2 = \left\|\Lambda_1^{-1}\delta_0\right\|^2$. In this case, the two-step $\mathbb{W}_{2T}$ test is necessarily less powerful. The proof uses the theory of uniformly most powerful invariant tests and the theory of complete and

sufficient statistics. It is an open question whether the same strategy can be adopted to prove Proposition 8 in the case of kernel LRV estimation. Our extensive numerical work supports that $\tau(\lambda; h, p, q, \alpha) > 1$ and $f(\lambda; h, p, q, \alpha) > 0$ continue to hold in the kernel case.

It is not easy to give an analytical expression for $f(\lambda; h, p, q, \alpha)$ but we can compute it numerically without any difficulty. In Table 4, we consider the case of OS LRV estimation and compute the values of $f(\lambda; K, p, q, \alpha)$ for $\lambda = 1 \sim 25, K = 8, 10, 12, 14$, $p = 1 \sim 3$ and $q = 1 \sim 3$. The values are nontrivial in that they are not close to the bounary value of zero or one. Similar to the asymptotic variance comparison, we find that these threshold values increase as the degree of overidentification increases and decrease as the smoothing parameter K increases.

For the case of kernel LRV estimation, results not reported here show that $f(\lambda; h, p, q, \alpha)$ increases with q and decreases with h. This is entirely analogous to the case of OS LRV estimation.

5 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 θ_0 is the true value, an interior point of the parameter space Θ. The moment conditions

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

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 equal to d. That is, we consider a model that is possibly overidentified with the degree of overidentification $q = m - d$.

5.1 One-step and Two-step Estimation and Inference

Define the $m \times m$ contemporaneous covariance matrix $\breve{\Sigma}$ and the LRV matrix $\breve{\Omega}$ as:

$$\breve{\Sigma} = E\breve{f}(v_t, \theta_0)\breve{f}(v_t, \theta_0)' \text{ and } \breve{\Omega} = \sum_{j=-\infty}^{\infty} \breve{\Omega}_j \text{ where } \breve{\Omega}_j = E\breve{f}(v_t, \theta_0)\breve{f}(v_{t-j}, \theta_0)'.$$

Let

$$\breve{g}_t(\theta) = \frac{1}{\sqrt{T}} \sum_{j=1}^{t} \breve{f}(v_j, \theta).$$

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

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

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 $\breve{\Sigma}$ or $\breve{\Omega}$ as the weighting matrix, we obtain the following two (infeasible) GMM estimators:

$$\tilde{\theta}_{1T} \quad : \quad = \arg\min_{\theta \in \Theta} \breve{g}_T(\theta)' \breve{\Sigma}^{-1} \breve{g}_T(\theta), \tag{12}$$

$$\tilde{\theta}_{2T} \quad : \quad = \arg\min_{\theta \in \Theta} \breve{g}_T(\theta)' \breve{\Omega}^{-1} \breve{g}_T(\theta). \tag{13}$$

For the estimator $\tilde{\theta}_{1T}$, we use the contemporaneous covariance matrix $\breve{\Sigma}$ as the weighting matrix and ignore all the serial dependency in the moment vector process $\{\breve{f}(v_t,\theta_0)\}_{t=1}^T$. In contrast to this procedure, the second estimator $\tilde{\theta}_{2T}$ accounts for the long run dependency. The feasible versions of these two estimators $\hat{\theta}_{1T}$ and $\hat{\theta}_{2T}$ can be naturally defined by replacing $\breve{\Sigma}$ and $\breve{\Omega}$ with their estimates $\breve{\Sigma}_{est}(\hat{\theta}_{0T})$ and $\breve{\Omega}_{est}(\hat{\theta}_{0T})$ where

$$\breve{\Sigma}_{est}(\theta) \;:=\; \frac{1}{T}\sum_{t=1}^{T}\breve{f}(v_t,\theta)\breve{f}(v_t,\theta)', \tag{14}$$

$$\breve{\Omega}_{est}(\theta) \;:=\; \frac{1}{T}\sum_{s=1}^{T}\sum_{t=1}^{T}Q_h^*(\frac{s}{T},\frac{t}{T})\breve{f}(v_t,\theta)\breve{f}(v_s,\theta)'. \tag{15}$$

To test the null hypothesis $H_0 : R\theta_0 = r$ against $H_1 : R\theta_0 = r + \delta_0/\sqrt{T}$, we construct two different Wald statistics as follows:

$$\begin{aligned}\mathbb{W}_{1T} &:= T(R\hat{\theta}_{1T}-r)'\left\{R\hat{\mathcal{V}}_{1T}R'\right\}^{-1}(R\hat{\theta}_{1T}-r), \\ \mathbb{W}_{2T} &:= T(R\hat{\theta}_{2T}-r)'\left\{R\hat{\mathcal{V}}_{2T}R'\right\}^{-1}(R\hat{\theta}_{2T}-r),\end{aligned} \tag{16}$$

where

$$\begin{aligned}\hat{\mathcal{V}}_{1T} &= \left[\breve{G}_{1T}'\breve{\Sigma}_{est}^{-1}(\hat{\theta}_{1T})\breve{G}_{1T}\right]^{-1}\left[\breve{G}_{1T}'\breve{\Sigma}_{est}^{-1}(\hat{\theta}_{1T})\breve{\Omega}_{est}(\hat{\theta}_{1T})\breve{\Sigma}_{est}^{-1}(\hat{\theta}_{1T})\breve{G}_{1T}\right]\left[\breve{G}_{1T}'\breve{\Sigma}_{est}^{-1}(\hat{\theta}_{1T})\breve{G}_{1T}\right]^{-1} \\ \hat{\mathcal{V}}_{2T} &= \left[\breve{G}_{2T}\breve{\Omega}_{est}^{-1}(\hat{\theta}_{2T})\breve{G}_{2T}\right]^{-1}\end{aligned} \tag{17}$$

and

$$\breve{G}_{1T} = \frac{1}{T}\sum_{t=1}^{T}\left.\frac{\partial\breve{f}(v_t,\theta)}{\partial\theta'}\right|_{\theta=\hat{\theta}_{1T}},\ \breve{G}_{2T} = \frac{1}{T}\sum_{t=1}^{T}\left.\frac{\partial\breve{f}(v_t,\theta)}{\partial\theta'}\right|_{\theta=\hat{\theta}_{2T}}.$$

These are the standard Wald test statistics in the GMM framework.

To compare the two estimators $\hat{\theta}_{1T}$ and $\hat{\theta}_{2T}$ and associated tests, we maintain the standard assumptions below.

Assumption 4 *As $T\to\infty$ for a fixed h, $\hat{\theta}_{0T}=\theta_0+o_p(1)$, $\hat{\theta}_{1T}=\theta_0+o_p(1)$, $\hat{\theta}_{2T}=\theta_0+o_p(1)$ for an interior point $\theta_0\in\Theta$.*

Assumption 5 *Define*

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

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

With these assumptions and some mild conditions, the standard GMM theory gives us

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

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

$$\sqrt{T}(\hat{\theta}_{2T} - \theta_0) = \frac{1}{\sqrt{T}} \sum_{t=1}^{T} \left[\breve{G}' \breve{\Omega}_\infty^{-1} \breve{G} \right]^{-1} \breve{G}' \breve{\Omega}_\infty^{-1} \breve{f}(v_t, \theta_0) + o_p(1),$$

where $\breve{\Omega}_\infty$ is defined in the similar way as Ω_∞ in Proposition 1: $\breve{\Omega}_\infty = \breve{\Omega}_{1/2} \tilde{\Omega}_\infty \breve{\Omega}_{1/2}'$.

Due to the complicated structure of two transformed moment vector processes, it is not straightforward to compare the asymptotic distributions of $\hat{\theta}_{1T}$ and $\hat{\theta}_{2T}$ as in Sections 3 and 4. To confront this challenge, we let

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

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 $\{f^*(v_t, \theta_0)\}$ are

$$\Sigma^* := U' \breve{\Sigma} U = \begin{pmatrix} \Sigma_{11}^* & \Sigma_{12}^* \\ \Sigma_{21}^* & \Sigma_{22}^* \end{pmatrix},$$

and $\Omega^* := U' \breve{\Omega} U$, respectively. To convert the variance matrix into an identity matrix, we define the normalized moment conditions below:

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

where

$$\Sigma_{1/2}^* = \begin{pmatrix} (\Sigma_{1\cdot 2}^*)^{1/2} & \Sigma_{12}^* (\Sigma_{22}^*)^{-1/2} \\ 0 & (\Sigma_{22}^*)^{1/2} \end{pmatrix}. \tag{18}$$

More specifically,

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

Then the contemporaneous variance of the time series $\{f(v_t, \theta_0)\}$ is I_m and the long run variance is $\Omega := (\Sigma_{1/2}^*)^{-1} \Omega^* (\Sigma_{1/2}^{*\prime})^{-1}$.

Lemma 9 *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 $h > 0$,*

$$(\Sigma_{1\cdot 2}^*)^{-1/2} AV' \sqrt{T}(\hat{\theta}_{1T} - \theta_0) = \frac{1}{\sqrt{T}} \sum_{t=1}^{T} f_1(v_t, \theta_0) + o_p(1) \Longrightarrow N(0, \Omega_{11}) \tag{19}$$

$$\begin{aligned} (\Sigma_{1\cdot 2}^*)^{-1/2} AV' \sqrt{T}(\hat{\theta}_{2T} - \theta_0) &= \frac{1}{\sqrt{T}} \sum_{t=1}^{T} \left[f_1(v_t, \theta_0) - \beta_\infty f_2(v_t, \theta_0) \right] + o_p(1) \\ &\Longrightarrow MN\left(0, \Omega_{11} - \Omega_{12}\beta_\infty' - \beta_\infty \Omega_{21} + \beta_\infty \Omega_{22} \beta_\infty'\right) \end{aligned} \tag{20}$$

where $\beta_\infty := \Omega_{\infty,12} \Omega_{\infty,22}^{-1}$ is the same as in Proposition 1.

Lemma 9 casts the stochastic expansions of two estimators in the same form. To the best of our knowledge, these representations are new in the econometric literature and may be of independent interest. Lemma 9 enables us to directly compare the asymptotic properties of one-step and two-step estimators and the associated tests.

It follows from the proof of the lemma that

$$(\Sigma^*_{1\cdot 2})^{-1/2} AV'\sqrt{T}(\tilde{\theta}_{2T} - \theta_0) = \frac{1}{\sqrt{T}} \sum_{t=1}^{T} [f_1(v_t, \theta_0) - \beta_0 f_2(v_t, \theta_0)] + o_p(1),$$

where $\beta_0 = \Omega_{12}\Omega_{22}^{-1}$ as defined before. So the difference between the feasible and infeasible two-step GMM estimators lies in the uncertainty in estimating β_0. While the true value of β appears in the asymptotic distribution of the infeasible estimator $\tilde{\theta}_{2T}$, the fixed-smoothing limit of the implied estimator $\hat{\beta} := \hat{\Omega}_{12}\hat{\Omega}_{22}^{-1}$ appears in that of the feasible estimator $\hat{\theta}_{2T}$. It is important to point out that the estimation uncertainty in the whole weighting matrix $\breve{\Omega}_{est}$ matters only through that in $\hat{\beta}$.

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 (19) and (20) are exactly the same as what we would obtain in the location model. The location model, as simple as it is, has implications for general settings from an asymptotic point of view. More specifically, define

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

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

To present our next theorem, we transform R into $\tilde{R}$ using

$$\tilde{R} = RVA^{-1} (\Sigma^*_{1\cdot 2})^{1/2}, \tag{21}$$

which has the same dimension as R. We let

$$\tilde{\beta}_\infty (h, p, q) = \left[\int_0^1 \int_0^1 Q_h^*(r, s) dB_p(r) dB_q(s)'\right] \left[\int_0^1 \int_0^1 Q_h^*(r, s) dB_q(r) dB_q(s)'\right]^{-1},$$

which is compatible with the definition in (2). We define

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

While ρ is the long run correlation matrix between $f_1(v_t, \theta_0)$ and $f_2(v_t, \theta_0)$, ρ_R is the long run correlation matrix between $\tilde{R}f_1(v_t, \theta_0)$ and $f_2(v_t, \theta_0)$. The corresponding long run canonical correlation coefficients are

$$\nu\left(\rho\rho'\right) = \nu\left\{\left(\Omega_{12}\Omega_{22}^{-1}\Omega_{21}\right)\Omega_{11}^{-1}\right\} \text{ and } \nu\left(\rho_R\rho_R'\right) = \nu\left\{(\tilde{R}\Omega_{12}\Omega_{22}^{-1}\Omega_{21}\tilde{R}')(\tilde{R}\Omega_{11}\tilde{R}')^{-1}\right\}.$$

For the location model considered before, $\breve{G} = (I_d, O_{d\times q})'$ and so $U = I_m$, $A = I_d$ and $V = I_d$. Given the assumption that $\breve{\Sigma} = \Sigma^* = I_m$, which implies that $\Sigma^*_{1\cdot 2} = I_d$, we have $\tilde{R} = R$. So the above definition of ρ_R is identical to that in (10).

Theorem 10 *Let the assumptions in Lemma 9 hold. Define*

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

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 $\nu_{\max}(\rho_R\rho_R') < g(h,q)$*, then* $R\hat{\theta}_{2T}$ *has a larger asymptotic variance than* $R\hat{\theta}_{1T}$*.*

(b) If $\nu_{\min}(\rho_R\rho_R') > g(h,q)$*, then* $R\hat{\theta}_{2T}$ *has a smaller asymptotic variance than* $R\hat{\theta}_{1T}$*.*

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

If $R = I_d$, then $\tilde{R}$ is a square matrix with a full rank. Since the long canonical correlation coefficient is invariant to a full-rank linear transformation, we have $\nu(\rho_R\rho_R') = \nu(\rho\rho')$. It then follows from Theorem 10(a) (b) that

(i) if $\nu_{\max}(\rho\rho') < g(h,q)$, then $\text{avar}(\hat{\theta}_{2T}) > \text{avar}(\hat{\theta}_{1T})$.

(ii) if $\nu_{\min}(\rho\rho') > g(h,q)$, then $\text{avar}(\hat{\theta}_{2T}) < \text{avar}(\hat{\theta}_{1T})$.

These results are identical to what we obtain for the location model. The only difference is that in the general GMM case we need to rotate and standardize the original moment conditions before computing the long run correlation matrix. Theorem 10 can also be applied to a general location model with a nonscalar error variance, in which case $\tilde{R} = R(\Sigma_{1\cdot 2}^*)^{1/2}$.

5.2 GMM Estimation and Inference with a Working Weighting Matrix

In the previous subsection, we employ two specific weighting matrices — the variance and long run variance estimators. In this subsection, we consider a general weighting matrix $\breve{W}_T(\hat{\theta}_{0T})$, which may depend on the initial estimator $\hat{\theta}_{0T}$ and the sample size T, leading to yet another GMM estimator:

$$\hat{\theta}_{aT} = \arg\min_{\theta\in\Theta} \breve{g}_T(\theta)' \left[\breve{W}_T(\hat{\theta}_{0T})\right]^{-1} \breve{g}_T(\theta)$$

where the subscript 'a' signifies 'another' or 'alternative'.

An example of $\breve{W}_T(\hat{\theta}_{0T})$ is the implied LRV matrix when we employ a simple approximating parametric model to capture the dynamics in the moment process. We could also use the general LRV estimator but we choose a large h so that the variation in $\breve{W}_T(\hat{\theta}_{0T})$ is small. In the kernel LRV estimation, this amounts to including only autocovariances of low orders in constructing $\breve{W}_T(\hat{\theta}_{0T})$. We assume that $\breve{W}_T(\hat{\theta}_{0T}) \xrightarrow{p} \breve{W}$, a positive definite nonrandom matrix under the fixed-smoothing asymptotics. $\breve{W}$ may not be equal to the variance or long run variance of the moment process. We call $\breve{W}_T(\hat{\theta}_{0T})$ a working weighting matrix. This is in the same spirit of using a working correlation matrix rather than a true correlation matrix in the generalized estimating equations (GEE) setting. See, for example, Liang and Zeger (1986).

In parallel to (16), we construct the test statistic

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

where, for $\breve{G}_{aT} = \frac{1}{T}\sum_{t=1}^{T} \partial \breve{f}(v_t,\theta)/\partial\theta'\Big|_{\theta=\hat{\theta}_{aT}}$, $\hat{\mathcal{V}}_{aT}$ is defined according to

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

which is a standard variance estimator for $\hat{\theta}_{aT}$.

Define

$$W^* = U'\breve{W}U \text{ and } W = \Sigma_{1/2}^{*-1}W^*(\Sigma_{1/2}^{*\prime})^{-1} = \begin{pmatrix} W_{11} & W_{12} \\ W_{21} & W_{22} \end{pmatrix}$$

and $\beta_a = W_{12}W_{22}^{-1}$.

Using the same argument for proving Lemma 9, we can show that

$$(\Sigma_{1\cdot 2}^*)^{-1/2}AV'\sqrt{T}(\hat{\theta}_{aT} - \theta_0) = \frac{1}{\sqrt{T}}\sum_{t=1}^{T}\left[f_1(v_t,\theta_0) - \beta_a f_2(v_t,\theta_0)\right] + o_p(1). \tag{22}$$

The above representation is the same as that in (20) except that β_∞ is now replaced by β_a.

Let $\mathcal{V}_a$ and $\mathcal{V}_{a,R}$ be the long run variances of

$$\left[f_1(v_t,\theta_0) - \beta_a f_2(v_t,\theta_0)\right] \text{ and } \tilde{R}\left[f_1(v_t,\theta_0) - \beta_a f_2(v_t,\theta_0)\right],$$

respectively. The long run correlation matrices are

$$\rho_a = \mathcal{V}_a^{-1/2}(\Omega_{12} - \beta_a\Omega_{22})\,\Omega_{22}^{-1/2} \text{ and } \rho_{a,R} = \mathcal{V}_{a,R}^{-1/2}\left[\tilde{R}(\Omega_{12} - \beta_a\Omega_{22})\right]\Omega_{22}^{-1/2}.$$

The corresponding long run canonical correlation coefficients are

$$\begin{aligned} \nu\left(\rho_a\rho_a'\right) &= \nu\left\{(\Omega_{12} - \beta_a\Omega_{22})\,\Omega_{22}^{-1}(\Omega_{12} - \beta_a\Omega_{22})'\,\mathcal{V}_a^{-1}\right\} \text{ and} \\ \nu\left(\rho_{a,R}\rho_{a,R}'\right) &= \nu\left\{\tilde{R}(\Omega_{12} - \beta_a\Omega_{22})\,\Omega_{22}^{-1}(\Omega_{12} - \beta_a\Omega_{22})'\,\tilde{R}'\mathcal{V}_{a,R}^{-1}\right\}. \end{aligned}$$

Theorem 11 *Let the assumptions in Lemma 9 hold. Assume further that $\breve{W}_T(\hat{\theta}_{0T}) \xrightarrow{p} \breve{W}$, a positive definite nonrandom matrix. Consider the local alternative $H_1(\lambda_0)$ and the fixed-smoothing asymptotics.*

(a) If $\nu_{\max}(\rho_{a,R}\rho_{a,R}') < g(h,q)$, then $R\hat{\theta}_{2T}$ has a larger asymptotic variance than $R\hat{\theta}_{aT}$.

(b) If $\nu_{\min}(\rho_{a,R}\rho_{a,R}') > g(h,q)$, then $R\hat{\theta}_{2T}$ has a smaller asymptotic variance than $R\hat{\theta}_{aT}$.

(c) If $\nu_{\max}(\rho_{a,R}\rho_{a,R}') < f(\lambda_0;h,p,q,\alpha)$, then the two-step test based on $\mathbb{W}_2$ is asymptotically less powerful than the test based on $\mathbb{W}_a$ for any $\delta_0 \in \mathfrak{A}(\lambda_0)$.

(d) If $\nu_{\min}(\rho_{a,R}\rho_{a,R}') > f(\lambda_0;h,p,q,\alpha)$, then the two-step test based on $\mathbb{W}_2$ is asymptotically more powerful than the test based on $\mathbb{W}_a$ for any $\delta_0 \in \mathfrak{A}(\lambda_0)$.

Theorem 11 is entirely analogous to Theorem 10. The only difference is that the second block of moment conditions is removed from the first block using the implied matrix coefficient β_a before computing the long run correlation coefficient.

When $R = I_d$, $\tilde{R}$ becomes a square matrix, and we have $\nu(\rho_{a,R}\rho_{a,R}') = \nu(\rho_a\rho_a')$. Theorem 11(a) and (b) gives the conditions under which $\hat{\theta}_{2T}$ is asymptotically more (or less) efficient than $\hat{\theta}_{aT}$.

To understand the theorem, we can see that the effective moment conditions behind $R\hat{\theta}_{aT}$ are:

$$Ef_{1a}(v_t,\theta_0)=0 \text{ for } f_{1a}(v_t,\theta_0)=\tilde{R}\left[f_1(v_t,\theta_0)-\beta_a f_2(v_t,\theta_0)\right].$$

$R\hat{\theta}_{aT}$ uses the information in $Ef_2(v_t,\theta_0)=0$ to some extent, but it ignores the residual information that is still potentially available from $Ef_2(v_t,\theta_0)=0$. In contrast, $R\hat{\theta}_{2T}$ attempts to explore the residual information. If there is no long run correlation between $f_{1a}(v_t,\theta_0)$ and $f_2(v_t,\theta_0)$, i.e., $\rho_{a,R}=0$, then all the information in $Ef_2(v_t,\theta_0)=0$ has been fully captured by the effective moment conditions underlying $R\hat{\theta}_{aT}$. As a result, the test based on $R\hat{\theta}_{aT}$ necessarily outperforms that based on $R\hat{\theta}_{2T}$. If the long run correlation $\rho_{a,R}$ is large enough in the sense given in Theorem 11(d), the test based on $R\hat{\theta}_{2T}$ could be more powerful than that based on $R\hat{\theta}_{aT}$ in large samples.

6 Simulation Evidence and Practical Guidance

This section compares the finite sample performances of one-step and two-step estimators and tests using the fixed-smoothing approximations. We consider the location model given in (1) with the true parameter value $\theta_0=(0,...,0)\in\mathbb{R}^d$ but we allow for a nonscalar error variance. The error $\{u_t^*\}$ follows a VAR(1) process:

$$\begin{aligned} u_{1t}^{*i} &= \psi u_{1t-1}^{*i}+\frac{\gamma}{\sqrt{q}}\sum_{j=1}^{q}u_{2t-1}^{*j}+e_{1t}^{*i} \text{ for } i=1,...,d \\ u_{2t}^{*i} &= \psi u_{2t-1}^{*i}+e_{2t}^{*i} \text{ for } i=1,...,q \end{aligned} \tag{23}$$

where $e_{1t}^{*i}\sim iid\ N(0,1)$ across i and t, $e_{2t}^{*i}\sim iid\ N(0,1)$ across i and t, and $\{e_{1t}^*,t=1,2,...,T\}$ are independent of $\{e_{2t}^*,t=1,2,...,T\}$. Let $u_t^*:=((u_{1t}^*)',(u_{2t}^*)')'\in\mathbb{R}^m\in\mathbb{R}^m$, then $u_t^*=\tilde{\Gamma}u_{t-1}^*+e_t^*$ where

$$\underset{m\times m}{\Gamma}=\begin{pmatrix}\psi I_d & \frac{\gamma}{\sqrt{q}}J_{d,q}\\ 0 & \psi I_q\end{pmatrix},\ e_t^*=\begin{pmatrix}e_{1t}^*\\ e_{2t}^*\end{pmatrix}\sim iid\ N(0,I_m),$$

and $J_{d,q}$ is the $d\times q$ matrix of ones. Direct calculations give us the expressions for the long run and contemporaneous variances of $\{u_t^*\}$ as

$$\begin{aligned}\Omega^* &= \sum_{j=-\infty}^{\infty}Eu_t^*(u_{t-j}^*)'=(I_m-\Gamma)^{-1}(I_m-\Gamma')^{-1}\\ &= \begin{pmatrix}\frac{1}{(1-\psi)^2}I_d+\frac{\gamma^2}{(1-\psi)^4}J_{d,d} & \frac{\gamma}{(1-\psi)^3\sqrt{q}}J_{d,q}\\ \frac{\gamma}{(1-\psi)^3\sqrt{q}}J_{q,d} & \frac{1}{(1-\psi)^2}I_q\end{pmatrix}\end{aligned}$$

and

$$\Sigma^*=var(u_t^*)=\begin{pmatrix}\frac{1}{1-\psi^2}I_d+\frac{\gamma^2(1+\psi^2)}{(1-\psi^2)^3}J_{d,d} & \frac{\gamma}{\sqrt{q}}\frac{\psi}{(1-\psi^2)^2}J_{d,q}\\ \frac{\gamma}{\sqrt{q}}\frac{\psi}{(1-\psi^2)^2}J_{q,d} & \frac{1}{1-\psi^2}I_q\end{pmatrix}.$$

Let $u_{1t}=(\Sigma_{1\cdot2}^*)^{-1/2}\left[u_{1t}^*-\Sigma_{12}^*(\Sigma_{22}^*)^{-1}u_{2t}^*\right]$ and $u_{2t}=(\Sigma_{22}^*)^{-1/2}u_{2t}^*$ and ρ be the long run correlation matrix between u_{1t} and u_{2t}. With some algebraic manipulations, we have

$$\rho\rho'=\left(d+\frac{(1-\psi^2)^2}{\gamma^2}\right)^{-1}J_{d,d}. \tag{24}$$

So the maximum eigenvalue of $\rho\rho'$ 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.

In addition to the VAR(1) error process, we also consider the following VARMA(1,1) process for u_t^* :

$$\begin{aligned} u_{1t}^{*i} &= \psi u_{1t-1}^{*i} + e_{1t}^{*i} + \frac{\gamma}{\sqrt{q}} \sum_{j=1}^{q} e_{2,t-1}^{*j} \text{ for } i = 1, ..., d \\ u_{2t}^{*i} &= \psi u_{2t-1}^{*i} + e_{2t}^{*i} \text{ for } i = 1, ..., q \end{aligned} \tag{25}$$

where $e_t^* \overset{i.i.d}{\sim} N(0, I_m)$. The corresponding long run covariance matrix Ω^* and contemporaneous covariance matrix Σ^* are

$$\Omega^* = \begin{pmatrix} \frac{1}{(1-\psi)^2} I_d + \frac{\gamma^2}{(1-\psi)^2} \cdot J_{d,d} & \frac{\gamma}{(1-\psi)^2\sqrt{q}} \cdot J_{d,q} \\ \frac{\gamma}{(1-\psi)^2\sqrt{q}} \cdot J_{q,d} & \frac{1}{(1-\psi)^2} \cdot I_q \end{pmatrix}$$

and

$$\Sigma^* = \begin{pmatrix} \frac{1}{1-\psi^2} I_d + \frac{\gamma^2}{1-\psi^2} J_{d,d} & \frac{1}{\sqrt{q}} \frac{\psi\gamma}{1-\psi^2} \cdot J_{d,q} \\ \frac{1}{\sqrt{q}} \frac{\psi\gamma}{1-\psi^2} \cdot J_{q,d} & \frac{1}{1-\psi^2} \cdot I_q \end{pmatrix}.$$

With some additional algebras, we have

$$\rho\rho' = \left(d + \frac{1}{(1-\psi)^2 \gamma^2}\right)^{-1} J_{d,d}, \tag{26}$$

and $\nu_{\max}(\rho\rho') = (1 + 1/[d(1-\psi)^2\gamma^2])^{-1}$.

Under the VARMA(1,1) design, the approximating AR(1) model is misspecified. It is not hard to obtain the probability limit of $\breve{W}(\tilde{\theta}_{aT})$ as

$$\breve{W} = \left(I_m - \tilde{\Gamma} - \tilde{\Lambda}(\Sigma^*)^{-1}\right)^{-1} \left(I - \tilde{\Lambda}(\Sigma^*)^{-1}\tilde{\Lambda}' + \tilde{\Lambda}\tilde{\Lambda}'\right) \left(I_m - \tilde{\Gamma}' - (\Sigma^*)^{-1}\tilde{\Lambda}'\right)^{-1},$$

which is different from the true long run variance matrix Ω^*. Based on $\breve{W}$, Ω^*, and Σ^*, we can compute $\rho_a\rho_a'$ and $\rho_{a,R}\rho_{a,R}'$.

For the basis functions in OS LRV estimation, we choose the following orthonormal basis functions $\{\Phi_j\}_{j=1}^{\infty}$ in the $L^2[0,1]$ space:

$$\Phi_{2j-1}(x) = \sqrt{2}\cos(2j\pi x) \text{ and } \Phi_{2j}(x) = \sqrt{2}\sin(2j\pi x) \text{ for } j = 1, ..., K/2,$$

where K is an even integer. We also consider kernel based LRV estimators with the three commonly-used kernels: Bartlett, Parzen, QS kernels. For the choice of K in OS LRV estimation, we employ the following AMSE-optimal formula in Phillips (2005):

$$K_{MSE} = 2 \times \left\lceil 0.5 \left(\frac{tr\left[(I_{m^2} + \mathbb{K}_{mm})(\Omega^* \otimes \Omega^*)\right]}{4vec(B^*)'vec(B^*)} \right)^{1/5} T^{4/5} \right\rceil$$

where $\lceil \cdot \rceil$ is the ceiling function, $\mathbb{K}_{mm}$ is $m^2 \times m^2$ commutation matrix and

$$B^* = -\frac{\pi^2}{6} \sum_{j=-\infty}^{\infty} j^2 E u_t^* u_{t-j}^{*\prime}.$$

Similarly, in the case of kernel LRV estimation, we select the smoothing parameter b according to the AMSE-optimal formula in Andrews (1991). The unknown parameters in the AMSE are either calibrated or data-driven using the VAR(1) plug-in approach. The qualitative messages 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 s $10,000$.

6.1 Point Estimation

We focus on the case with $d = 1$, under which $\rho\rho'$ is a scalar and $\nu_{\max}(\rho\rho') = \rho\rho'$. For both simulation designs, $\nu_{\max}(\rho\rho')$ is increasing in γ^2 for a given ψ. We fix the value of ψ at 0.75 so that each time series is reasonably persistent. For this value of ψ, we consider $\nu_{\max}(\rho\rho') = 0, 0.09, 0.18, ..., 0.90, 0.99$, which are obtained by setting γ to appropriate values using (24) or (26).

According to Proposition 3, if $\rho\rho'$ is greater than a threshold value, then $Var(\hat{\theta}_{2T})$ is expected to be smaller than $Var(\hat{\theta}_{1T})$. Otherwise, $Var(\hat{\theta}_{2T})$ is expected to be larger. We simulate $Var(\hat{\theta}_{1T})$, $Var(\hat{\theta}_{2T})$ and $Var(\hat{\theta}_{aT})$. Here, $\hat{\theta}_{aT}$ is based on a working weighting matrix $\breve{W}(\hat{\theta}_{0T})$ using VAR(1) as the approximating model for the estimated error process $\{\hat{u}_t^*(\hat{\theta}_{0T})\}$.

Tables 5∼6 report the simulated variances under the VAR(1) design with $q = 3$ and 4 for some given values of K and b. These values 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 $d = 1$ and $\rho\rho' = 0.40$. We first discuss the case when the OS LRV estimator is used. It is clear that $Var(\hat{\theta}_{2T})$ becomes smaller than $Var(\hat{\theta}_{1T})$ only when $\rho\rho'$ is large enough. For example, when $q = 4$ and there is no long run correlation, i.e., $\rho\rho' = 0$, we have $Var(\hat{\theta}_{1T}) = 0.081 < Var(\hat{\theta}_{2T}) = 0.112$, and so $\hat{\theta}_{1T}$ is more efficient than $\hat{\theta}_{2T}$ with 28% efficiency gain. These numerical observations are consistent with our theoretical result in Proposition 1: $\hat{\theta}_{2T}$ becomes more efficient relative to $\hat{\theta}_{1T}$ only when the benefit of using the LRV matrix outweighs the cost of estimating it. With the choice of $K = 14$ and $q = 4$, Table 5 indicates that $Var(\hat{\theta}_{2T})$ starts to become smaller than $Var(\hat{\theta}_{1T})$ when $\rho\rho'$ crosses a value in the interval $[0.270, 0.360]$ from below. This agrees with the theoretical threshold value $\rho\rho' = q/(K-1) \approx 0.307$ given in Corollary 4.

In the case of kernel LRV estimation, we get exactly the same qualitative messages. For example, consider the case with the Bartlett kernel, $b = 0.08$, and $q = 3$. We observe that $Var(\hat{\theta}_{2T})$ starts to become smaller than $Var(\hat{\theta}_{1T})$ when $\rho\rho'$ crosses a value in the interval $[0.09, 0.18]$ from below. This is compatible with the threshold value 0.152 given in Table 1.

Finally, we note that $Var(\hat{\theta}_{aT})$ is smaller than $Var(\hat{\theta}_{2T})$ for all values of $\rho\rho'$ considered. This is well expected. In constructing $\hat{\theta}_{aT}$, we employ a correctly specified parametric model to estimate the weighting matrix and so $\breve{W}(\hat{\theta}_{0T})$ converges in probability to the true long run variance matrix Ω^*. However, when the true DGP is VARMA(1,1), the results in Tables 7∼8 indicate that the efficiency of $\hat{\theta}_{aT}$ is reduced due to the misspecification bias in the working weighting matrix $\breve{W}(\hat{\theta}_{aT})$. The tables also report the values of $\rho_a\rho_a'$. We find that $\hat{\theta}_{aT}$ is more efficient than $\hat{\theta}_{2T}$ only when $\rho_a\rho_a'$ is below a certain threshold value. This confirms the qualitative messages in Theorem 11(a) and (b).

6.2 Hypothesis Testing

We implement three testing procedures on the basis of $\mathbb{W}_{1T}$, $\mathbb{W}_{2T}$ and $\mathbb{W}_{aT}$. Here, $\mathbb{W}_{aT}$ is based on the same working weighting matrix $\breve{W}(\hat{\theta}_{0T})$ as in the point estimation case. The nominal

significance level is $\alpha = 0.05$. As before, $\psi = 0.75$. We use (24) and (26) to set γ and obtain $\nu_{\max}(\rho\rho') \in \{0.00, 0.35, 0.50, 0.60, 0.80, 0.90\}$. We focus on the case with $d = 3$ and $q = 3$. The null hypotheses of interest are:

$$\begin{aligned} H_{01} &: \theta_1 = 0, \\ H_{02} &: \theta_1 = \theta_2 = 0 \end{aligned}$$

where $p = 1, 2$ respectively. For the smoothing parameters, we employ the data driven AMSE optimal bandwidth through VAR(1) plug-in implementation developed by Andrews (1991) and Phillips (2005).

Tables 9∼16 report the empirical size of three nominal 5% testing procedures based on the two types of asymptotic approximations. It is clear that all of the three tests based on $\mathbb{W}_{1T}, \mathbb{W}_{aT}$ and $\mathbb{W}_{2T}$ suffer from severe size distortion if the conventional normal (or chi-square) critical values are used. For example, when the DGP is VAR(1) and OS LRV estimation is implemented, the empirical sizes of the three tests using the OS LRV estimator are reported to be around $14\% \sim 29\%$ when $p = 2$. The relatively large size distortion of the $\mathbb{W}_{2T}$ test comes from the additional cost in estimating the weighting matrix. However, if the nonstandard critical values $\mathbb{W}^{\alpha}_{1\infty}$ and $\mathbb{W}^{\alpha}_{2\infty}$ are used, we observe that the size distortion of all three procedures is substantially reduced. The result agrees with the previous literature such as Sun (2013, 2014a&b) and Kiefer and Vogelsang (2005) which highlight the higher accuracy of the fixed-smoothing approximations. Also, we observe that when the fixed-smoothing approximations are used, the $\mathbb{W}_{1T}$ test is more size-distorted than the $\mathbb{W}_{2T}$ test in most cases. Similar results for the kernel cases are reported in Tables 11∼16.

Next, we investigate the finite sample power performances of the three procedures. We use the finite sample critical values under the null, so the power is size-adjusted and the power comparison is meaningful. The DGPs are the same as before except the parameters are from the local null alternatives $R\theta_0 = r + \delta_0/\sqrt{T}$. The degree of overidentification considered here is $q = 3$. Also, the domain of each power curve is rescaled to be $\lambda := \delta_0'(\tilde{R}\Omega_{1\cdot 2}\tilde{R}')^{-1}\delta_0$ with $\tilde{R} = R(\Sigma^*_{1\cdot 2})^{1/2}$ as in Section 4 and 5.

Figures 2∼4 show the size-adjusted finite sample power of the three procedures in the case of OS LRV estimation. We can see that in all figures, 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 from certain point $\nu_{\max}(\rho_R\rho_R') \in (0, 1)$. This is consistent with Proposition 7. For example, with $K = 14$ and $p = 1$, the power curves in Figure 2 show that the power curve of the two-step test $\mathbb{W}_{2T}$ starts to dominate that of the one-step test $\mathbb{W}_{1T}$ when $\nu_{\max}(\rho_R\rho_R')$ reaches 0.25. This matches our theoretical results in Proposition 7 and Table 4 which indicate that the threshold value $\max_{\lambda\in[1,25]} f(\lambda; K, p, q, \alpha)$ is about 0.275 when $K = 14, p = 1$ and $q = 3$. Also, if $\nu_{\max}(\rho_R\rho_R')$ is as high as 0.75, we can see that the two-step test is more powerful than the one-step test in most of cases.

Lastly, in the presence of VAR(1) error, the performance of $\mathbb{W}_{aT}$ dominates that of $\mathbb{W}_{1T}$ and $\mathbb{W}_{2T}$ for all $\nu_{\max}(\rho_R\rho_R') \in (0, 1)$. This is analogous to the point estimation results. The working weighting matrix $\breve{W}(\hat{\theta}_{0T})$ based on VAR(1) plug-in model is close to the true long run variance matrix Ω^*. This leads to power improvement whenever there is some long run correlation between u^*_{1t} and u^*_{2t}. However, under the VARMA(1,1) error, Figures 3∼4 show that the advantages of $\mathbb{W}_{aT}$ are reduced and $\mathbb{W}_{aT}$ is more powerful than the two-step test $\mathbb{W}_{2T}$ only when $\nu_{\max}(\rho_{a,R}\rho_{a,R}')$ is below the threshold value $f(\lambda_0; K, p, q, \alpha)$. This is due to the misspecification bias in $\breve{W}(\hat{\theta}_{0T})$ which is attributed to the use of a wrong plug-in model. Nevertheless, we still observe comparable

performances of $\mathbb{W}_{aT}$ for most of non-zero $\nu_{\max}(\rho_{a,R}\rho'_{a,R})$ values. Figures 5∼8 for the cases of kernel LRV estimation deliver the same qualitative messages.

6.3 Practical Recommendation

Both our theoretical result and 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 large enough. In empirical applications, we often care about only a 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 problem 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 to use the two-step estimator and test only when the estimated long run canonical correlation coefficient is larger than our theoretical threshold by a margin, say 10%. On the other hand, when the estimated long run canonical correlation coefficient is smaller than our theoretical threshold by 10%, we stick with the first-step estimator and test. When the estimated long run canonical correlation coefficient is within 10% of the theoretical threshold, we propose to use the GMM estimator and test based on a working weighting matrix using VAR(1) as the approximating parametric model. Our recommendation in the not so clear-cut case is based on the simulation evidence that the working weighting matrix can deliver a robust performance in finite samples.

We now formalize our recommendation using hypothesis testing as an example. Given the set of moment conditions $E\breve{f}(v_t,\theta_0)=0$ and the data $\{v_t\}$, suppose that we want to test $H_0: R\theta_0 = r$ against $R\theta_0 \neq r$ for some $R \in \mathbb{R}^{p\times d}$. We follow the steps below to decide on which test to use.

1. Compute the initial estimator $\hat{\theta}_{0T} = \arg\min_{\theta\in\Theta} \left\| \sum_{t=1}^{T} \breve{f}(v_t,\theta) \right\|^2$.

2. On the basis of $\hat{\theta}_{0T}$, use a data-driven method such as Andrews (1991) or Phillips (2005) to select the smoothing parameter. Denote the data-driven value by $\hat{h}$.

3. Based on the smoothing parameter $\hat{h}$, compute $\breve{\Sigma}_{est}(\hat{\theta}_{0T})$ and $\breve{\Omega}_{est}(\hat{\theta}_{0T})$ using the formulae in (15).

4. Compute $\breve{G}_T(\hat{\theta}_{0T}) = \frac{1}{T}\sum_{t=1}^{T} \frac{\partial \breve{f}(v_t,\theta)}{\partial\theta'}|_{\theta=\hat{\theta}_{0T}}$ 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

$$\hat{\Sigma}^* := \hat{U}'\breve{\Sigma}_{est}(\hat{\theta}_{0T})\hat{U} \text{ and } \hat{\Omega}^* := \hat{U}'\breve{\Omega}_{est}(\hat{\theta}_{0T})\hat{U}.$$

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

$$\hat{\Sigma}^*_{1/2} = \begin{pmatrix} \left(\hat{\Sigma}^*_{1\cdot 2}\right)^{1/2} & \hat{\Sigma}^*_{12}\left(\hat{\Sigma}^*_{22}\right)^{-1/2} \\ 0 & \left(\hat{\Sigma}^*_{22}\right)^{1/2} \end{pmatrix}. \tag{27}$$

7. Let $\tilde{R}_{est} = R\hat{V}\hat{A}^{-1}(\hat{\Sigma}^*_{1\cdot 2})^{1/2}$. Compute the eigenvalues:

$$\nu\left(\hat{\rho}_R\hat{\rho}_R'\right) = \nu\left[(\tilde{R}_{est}\hat{\Omega}_{12}\hat{\Omega}_{22}^{-1}\hat{\Omega}_{12}\tilde{R}_{est}')(\tilde{R}_{est}\hat{\Omega}_{11}\tilde{R}_{est}')^{-1}\right].$$

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 $P\left(\chi_p^2(\lambda^o) > \chi_p^{1-\alpha}\right) = 75\%$. This choice of λ^o is consistent with the optimal testing literature. We may also choose a value of λ^o to reflect scientific interest or economic significance.

9. (a) If $\nu_{\min}\left(\hat{\rho}_R\hat{\rho}_R'\right) > 1.1f(\lambda^o;\hat{h},p,q,\alpha)$, then we use the second-step test based on $\mathbb{W}_{2T}$.

(b) If $\nu_{\max}\left(\hat{\rho}_R\hat{\rho}_R'\right) < 0.9f(\lambda^o;\hat{h},p,q,\alpha)$, then we use the first-step test based on $\mathbb{W}_{1T}$.

(c) If neither condition (a) nor condition (b) is satisfied, then we use the testing procedure based on $\mathbb{W}_{aT}$ using the VAR(1) as the approximating parametric model to estimate the weighting matrix.

7 Conclusion

In this paper we have provided more accurate and honest comparisons between the popular one-step and two-step GMM estimators and the associated inference procedures. We have given some clear guidance on when we should go one step further and use a two-step procedure. Qualitatively, we want to go one step further only if the benefit of doing so clearly outweighs the cost. When the benefit and cost comparison is not clear-cut, we recommend using the GMM procedure with a working weighting matrix.

The qualitative message of the paper is applicable more broadly. As long as there is additional nonparametric estimation uncertainty in a two-step procedure relative to 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, it may do more harm than good sometimes if we blindly use the two-step procedure.

There are many extensions of the paper. We give some examples here. First, we can use the more accurate approximations to compare the continuous updating GMM and other generalized empirical likelihood estimators with the one-step and two-step GMM estimators. While the fixed-smoothing asymptotics captures the nonparametric estimation uncertainty of the weighting matrix estimator, it does not fully capture the estimation uncertainty embodied in the first-step estimator. The source of the problem is that we do not observe the moment process and have to use the estimated moment process based on the first-step estimator to construct the nonparametric variance estimator. It is interesting to develop a further refinement to the fixed-smoothing approximation to capture the first-step estimation uncertainty more adequately. Finally, it will be also very interesting to give an honest assessment of the relative merits of the OLS and GLS estimators which are popular in empirical applications.

Table 1: Threshold values $g(h,q)$ for asymptotic variance comparison with 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
0.020	0.014	0.027	0.040	0.053	0.065
0.030	0.020	0.040	0.059	0.078	0.097
0.040	0.027	0.053	0.079	0.104	0.128
0.050	0.034	0.066	0.098	0.128	0.157
0.060	0.040	0.079	0.116	0.152	0.185
0.070	0.047	0.092	0.135	0.175	0.211
0.080	0.054	0.104	0.152	0.197	0.237
0.090	0.061	0.117	0.170	0.218	0.260
0.100	0.068	0.129	0.186	0.238	0.282
0.110	0.074	0.141	0.203	0.257	0.303
0.120	0.081	0.153	0.218	0.274	0.322
0.130	0.088	0.164	0.233	0.291	0.340
0.140	0.094	0.175	0.247	0.306	0.356
0.150	0.101	0.186	0.260	0.321	0.371
0.160	0.107	0.196	0.273	0.334	0.384
0.170	0.113	0.206	0.284	0.347	0.397
0.180	0.119	0.216	0.295	0.358	0.407
0.190	0.124	0.226	0.306	0.369	0.417
0.200	0.130	0.235	0.316	0.380	0.425

Notes: $h = 1/b$ indicates the level of smoothing and q is the degrees of overidentification. If the largest squared long run canonical correlation between the two blocks of (rotated and transformed) moment conditions is less than $g(h,q)$, then the two-step estimator $\hat{\theta}_{2T}$ is asymptotically less efficient than the one-step estimator $\hat{\theta}_{1T}$. If the smallest squared long run canonical correlation is greater than $g(h,q)$, then the two-step estimator $\hat{\theta}_{2T}$ is asymptotically more efficient than the one-step estimator $\hat{\theta}_{1T}$.

Table 2: Threshold values $g(h, q)$ for asymptotic variance comparison with Parzen kernel

b	$q = 1$	$q = 2$	$q = 3$	$q = 4$	$q = 5$
0.010	0.006	0.011	0.016	0.022	0.027
0.020	0.011	0.022	0.033	0.043	0.054
0.030	0.017	0.033	0.049	0.065	0.081
0.040	0.022	0.044	0.065	0.087	0.107
0.050	0.028	0.055	0.082	0.108	0.134
0.060	0.033	0.066	0.099	0.130	0.161
0.070	0.039	0.077	0.115	0.152	0.187
0.080	0.045	0.088	0.132	0.173	0.213
0.090	0.051	0.100	0.148	0.194	0.238
0.100	0.057	0.111	0.164	0.215	0.263
0.110	0.063	0.122	0.181	0.236	0.288
0.120	0.069	0.133	0.197	0.257	0.312
0.130	0.075	0.145	0.213	0.277	0.336
0.140	0.081	0.156	0.229	0.297	0.359
0.150	0.087	0.168	0.245	0.317	0.382
0.160	0.093	0.179	0.261	0.337	0.404
0.170	0.100	0.191	0.277	0.356	0.426
0.180	0.106	0.202	0.293	0.375	0.448
0.190	0.112	0.214	0.308	0.393	0.469
0.200	0.118	0.225	0.323	0.411	0.489

See notes to Table 1

Table 3: Threshold values $g(h, q)$ for asymptotic variance comparison with QS kernel

b	$q = 1$	$q = 2$	$q = 3$	$q = 4$	$q = 5$
0.010	0.010	0.020	0.030	0.040	0.050
0.020	0.021	0.041	0.061	0.082	0.102
0.030	0.031	0.062	0.093	0.124	0.154
0.040	0.042	0.084	0.126	0.166	0.206
0.050	0.053	0.106	0.158	0.209	0.258
0.060	0.065	0.128	0.191	0.252	0.311
0.070	0.077	0.151	0.225	0.296	0.362
0.080	0.089	0.175	0.259	0.340	0.414
0.090	0.102	0.198	0.293	0.382	0.464
0.100	0.115	0.222	0.326	0.423	0.516
0.110	0.127	0.247	0.359	0.463	0.565
0.120	0.140	0.271	0.392	0.502	0.612
0.130	0.153	0.296	0.426	0.542	0.655
0.140	0.166	0.321	0.458	0.581	0.697
0.150	0.179	0.346	0.489	0.619	0.736
0.160	0.193	0.371	0.520	0.655	0.773
0.170	0.206	0.395	0.549	0.690	0.806
0.180	0.220	0.418	0.578	0.722	0.834
0.190	0.233	0.441	0.605	0.752	0.859
0.200	0.246	0.463	0.630	0.779	0.879

See notes to Table 1.

Table 4: Threshold Values $f(\lambda; K, p, q, \alpha)$ for power comparison with OS LRV estimation when $\alpha = 0.05$ and K = 8,10,12,14.

		$p=1$			$p=2$			$p=3$		
K	λ	$q=1$	$q=2$	$q=3$	$q=1$	$q=2$	$q=3$	$q=1$	$q=2$	$q=3$
	1.000	0.162	0.378	0.514	0.223	0.367	0.581	0.242	0.433	0.576
	5.000	0.151	0.364	0.503	0.214	0.370	0.582	0.225	0.469	0.623
	9.000	0.154	0.352	0.493	0.213	0.377	0.597	0.226	0.488	0.639
8	13.000	0.153	0.345	0.496	0.213	0.397	0.600	0.226	0.495	0.645
	17.000	0.160	0.352	0.489	0.217	0.399	0.608	0.230	0.498	0.652
	21.000	0.165	0.356	0.493	0.211	0.405	0.604	0.234	0.503	0.657
	25.000	0.171	0.355	0.492	0.208	0.399	0.611	0.231	0.510	0.665
	1.000	0.082	0.283	0.474	0.162	0.277	0.461	0.171	0.369	0.507
	5.000	0.130	0.281	0.426	0.133	0.310	0.439	0.192	0.348	0.507
	9.000	0.138	0.269	0.423	0.136	0.305	0.431	0.196	0.328	0.506
10	13.000	0.135	0.261	0.416	0.132	0.308	0.432	0.200	0.339	0.507
	17.000	0.128	0.267	0.406	0.137	0.308	0.431	0.209	0.341	0.509
	21.000	0.136	0.276	0.406	0.137	0.308	0.436	0.210	0.346	0.508
	25.000	0.134	0.270	0.418	0.135	0.308	0.439	0.203	0.344	0.509
	1.000	0.085	0.198	0.322	0.128	0.203	0.345	0.151	0.325	0.314
	5.000	0.106	0.218	0.298	0.127	0.244	0.336	0.129	0.301	0.345
	9.000	0.103	0.210	0.301	0.122	0.233	0.353	0.119	0.284	0.352
12	13.000	0.098	0.205	0.308	0.125	0.232	0.353	0.124	0.274	0.359
	17.000	0.105	0.193	0.318	0.128	0.230	0.359	0.124	0.277	0.366
	21.000	0.100	0.197	0.325	0.119	0.243	0.363	0.123	0.274	0.369
	25.000	0.118	0.197	0.325	0.110	0.236	0.360	0.121	0.284	0.378
	1.000	0.062	0.316	0.260	0.089	0.184	0.367	0.155	0.287	0.394
	5.000	0.091	0.232	0.275	0.133	0.181	0.287	0.112	0.220	0.341
	9.000	0.093	0.214	0.274	0.117	0.188	0.273	0.124	0.209	0.341
14	13.000	0.087	0.211	0.265	0.109	0.192	0.281	0.126	0.213	0.338
	17.000	0.097	0.200	0.263	0.109	0.201	0.285	0.125	0.214	0.338
	21.000	0.093	0.213	0.257	0.105	0.197	0.285	0.130	0.208	0.332
	25.000	0.110	0.226	0.268	0.101	0.191	0.289	0.122	0.209	0.334

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

Table 5: Finite sample variance comparison for the three estimators $\hat{\theta}_{1T}$, $\hat{\theta}_{2T}$ and $\hat{\theta}_{aT}$ under VAR(1) error with $T = 200$, and $q = 3$.

$\nu_{\max}(\rho\rho')$	$\text{Var}(\hat{\theta}_{1T})$	$\text{Var}(\hat{\theta}_{2T})$				$\text{Var}(\hat{\theta}_{aT})$
·	·	OS	Bartlett	Parzen	QS	·
·	·	K=14	b=0.08	b=0.15	b=0.08	·
0.000	0.081	0.103	0.100	0.108	0.109	0.089
0.090	0.093	0.105	0.103	0.110	0.111	0.093
0.180	0.107	0.108	0.105	0.112	0.113	0.096
0.270	0.124	0.111	0.108	0.114	0.115	0.099
0.360	0.146	0.115	0.111	0.117	0.118	0.102
0.450	0.174	0.120	0.116	0.120	0.122	0.106
0.540	0.214	0.127	0.122	0.125	0.127	0.110
0.630	0.272	0.137	0.131	0.132	0.134	0.116
0.720	0.368	0.154	0.145	0.144	0.146	0.123
0.810	0.554	0.185	0.174	0.166	0.170	0.135
0.900	1.073	0.274	0.253	0.227	0.235	0.166
0.990	10.892	1.937	1.731	1.372	1.451	0.714

Table 6: Finite sample variance comparison for the three estimators $\hat{\theta}_{1T}$, $\hat{\theta}_{2T}$ and $\hat{\theta}_{aT}$ under VAR(1) error with $T = 200$, and $q = 4$.

$\nu_{\max}(\rho\rho')$	$\text{Var}(\hat{\theta}_{1T})$	$\text{Var}(\hat{\theta}_{2T})$				$\text{Var}(\hat{\theta}_{aT})$
·	·	OS	Bartlett	Parzen	QS	·
·	·	K=14	b=0.07	b=0.150	b=0.07	·
0.000	0.081	0.112	0.104	0.120	0.114	0.089
0.090	0.092	0.114	0.106	0.121	0.115	0.093
0.180	0.106	0.117	0.108	0.123	0.118	0.096
0.270	0.122	0.124	0.111	0.126	0.120	0.100
0.360	0.146	0.125	0.115	0.129	0.124	0.105
0.450	0.175	0.130	0.121	0.133	0.129	0.110
0.540	0.217	0.139	0.129	0.139	0.135	0.116
0.630	0.278	0.151	0.141	0.148	0.146	0.123
0.720	0.379	0.172	0.160	0.163	0.162	0.134
0.810	0.576	0.213	0.198	0.193	0.196	0.152
0.900	1.128	0.328	0.305	0.276	0.289	0.197
0.990	11.627	2.538	2.364	1.884	2.089	1.013

Table 7: Finite sample variance comparison for the three estimators $\hat{\theta}_{1T}$, $\hat{\theta}_{2T}$ and $\hat{\theta}_{aT}$ under VARMA(1,1) error with $T = 200$, and $q = 3$

$\nu_{\max}(\rho\rho')$	$\mathrm{Var}(\hat{\theta}_{1T})$	$\mathrm{Var}(\hat{\theta}_{2T})$				$\nu_{\max}(\rho_a\rho_a')$	$\mathrm{Var}(\hat{\theta}_{aT})$
·	·	OS	Bartlett	Parzen	QS	·	·
·	·	$K = 14$	$b = 0.08$	$b = 0.15$	$b = 0.08$	·	·
0.000	0.081	0.103	0.100	0.108	0.109	0.000	0.089
0.090	0.104	0.105	0.102	0.110	0.110	0.152	0.087
0.180	0.129	0.107	0.103	0.111	0.112	0.199	0.090
0.270	0.161	0.109	0.105	0.113	0.114	0.250	0.096
0.360	0.202	0.112	0.108	0.116	0.117	0.306	0.104
0.450	0.255	0.116	0.111	0.119	0.121	0.368	0.115
0.540	0.329	0.121	0.116	0.124	0.126	0.439	0.130
0.630	0.439	0.129	0.123	0.131	0.133	0.519	0.153
0.720	0.620	0.143	0.134	0.143	0.145	0.611	0.191
0.810	0.970	0.168	0.155	0.165	0.168	0.716	0.265
0.900	1.950	0.240	0.215	0.228	0.233	0.838	0.471
0.990	20.496	1.589	1.357	1.411	1.462	0.982	4.356

Table 8: Finite sample variance comparison for the three estimators $\hat{\theta}_{1T}$, $\hat{\theta}_{2T}$ and $\hat{\theta}_{aT}$ under VARMA(1,1) error with $T = 200$, and $q = 4$.

$\nu_{\max}(\rho\rho')$	$\mathrm{Var}(\hat{\theta}_{1T})$	$\mathrm{Var}(\hat{\theta}_{2T})$				$\nu_{\max}(\rho_a\rho_a')$	$\mathrm{Var}(\hat{\theta}_{aT})$
·	·	OS	Bartlett	Parzen	QS	·	·
·	·	$K = 14$	$b = 0.07$	$b = 0.15$	$b = 0.07$	·	·
0.000	0.081	0.112	0.104	0.120	0.114	0.000	0.089
0.090	0.103	0.113	0.105	0.121	0.115	0.152	0.086
0.180	0.132	0.115	0.106	0.123	0.117	0.199	0.091
0.270	0.167	0.118	0.108	0.125	0.119	0.250	0.098
0.360	0.212	0.121	0.111	0.128	0.122	0.306	0.109
0.450	0.272	0.126	0.114	0.132	0.126	0.368	0.123
0.540	0.356	0.132	0.119	0.137	0.131	0.439	0.144
0.630	0.481	0.142	0.127	0.145	0.139	0.519	0.174
0.720	0.686	0.158	0.140	0.159	0.153	0.611	0.225
0.810	1.086	0.190	0.164	0.186	0.180	0.716	0.325
0.900	2.206	0.279	0.235	0.262	0.257	0.838	0.605
0.990	23.519	2.013	1.598	1.735	1.742	0.982	5.954

Table 9: Empirical size of one-step and two-step tests based on the series LRV estimator under VAR(1) error when $\psi = 0.75, p = 1 \sim 2$, and $T = 200$

$p = 1$ and $q = 3$						
$\cdot$	One Step($\hat{\Sigma}^*$)		One Step($\breve{W}$)		Two Step	
$\nu_{\max}(\rho_R \rho_R')$	χ^2	$\mathbb{W}_{1\infty}$	χ^2	$\mathbb{W}_{1\infty}$	χ^2	$\mathbb{W}_{2\infty}$
0.00	0.128	0.098	0.151	0.119	0.187	0.076
0.15	0.126	0.096	0.135	0.103	0.177	0.061
0.25	0.135	0.102	0.138	0.105	0.187	0.063
0.33	0.135	0.105	0.127	0.094	0.174	0.059
0.57	0.139	0.107	0.086	0.061	0.154	0.044
0.75	0.143	0.116	0.046	0.031	0.118	0.032
$p = 2$ and $q = 3$						
$\cdot$	One Step($\hat{\Sigma}^*$)		One Step($\breve{W}$)		Two Step	
$\nu_{\max}(\rho_R \rho_R')$	χ^2	$\mathbb{W}_{1\infty}$	χ^2	$\mathbb{W}_{1\infty}$	χ^2	$\mathbb{W}_{2\infty}$
0.00	0.181	0.111	0.222	0.138	0.290	0.077
0.26	0.191	0.118	0.219	0.136	0.296	0.069
0.40	0.192	0.115	0.201	0.120	0.290	0.065
0.50	0.195	0.119	0.194	0.112	0.290	0.057
0.73	0.206	0.120	0.168	0.095	0.272	0.057
0.86	0.206	0.124	0.143	0.082	0.245	0.051

Notes: "One Step($\hat{\Sigma}^*$) test" is based on the first-step GMM estimator using the contemporaneous variance estimator as the weighing matrix; "One Step($\breve{W}$) test" is based on the GMM estimator using the VAR(1) parametric plug-in LRV estimator as the weighing matrix; "Two Step test" is based on the two-step GMM estimator using the data driven nonparametric LRV estimator as the weighing matrix.

Table 10: Empirical size of one-step and two-step tests based on the series LRV estimator under VARMA(1,1) error when $\psi = 0.75, p = 1 \sim 2$, and $T = 200$

$p = 1$ and $q = 3$						
$\cdot$	One Step($\hat{\Sigma}^*$)		One Step($\breve{W}$)		Two Step	
$\nu_{\max}(\rho_R \rho_R')$	χ^2	$\mathbb{W}_{1\infty}$	χ^2	$\mathbb{W}_{1\infty}$	χ^2	$\mathbb{W}_{2\infty}$
0.00	0.117	0.091	0.138	0.108	0.181	0.068
0.15	0.140	0.113	0.142	0.113	0.173	0.071
0.25	0.144	0.117	0.140	0.113	0.165	0.065
0.33	0.155	0.127	0.141	0.111	0.160	0.060
0.57	0.167	0.138	0.128	0.106	0.121	0.043
0.75	0.168	0.141	0.118	0.096	0.087	0.025
$p = 2$ and $q = 3$						
	One Step($\hat{\Sigma}^*$)		One Step($\breve{W}$)		Two Step	
$\nu_{\max}(\rho_R \rho_R')$	χ^2	$\mathbb{W}_{1\infty}$	χ^2	$\mathbb{W}_{1\infty}$	χ^2	$\mathbb{W}_{2\infty}$
0.00	0.188	0.119	0.227	0.146	0.290	0.080
0.26	0.202	0.129	0.209	0.136	0.270	0.073
0.40	0.206	0.135	0.204	0.134	0.254	0.069
0.50	0.223	0.148	0.215	0.144	0.251	0.065
0.73	0.221	0.148	0.205	0.138	0.214	0.053
0.86	0.222	0.156	0.194	0.132	0.178	0.044

See notes to Table 9.

Table 11: Empirical size of one-step and two-step tests based on the Bartlett kernel variance estimator under VAR(1) error when $\psi = 0.75$, $p = 1 \sim 2$ and $T = 200$

$p = 1$ and $q = 3$						
$\cdot$	One Step($\hat{\Sigma}^*$)		One Step($\breve{W}$)		Two Step	
$\nu_{\max}(\rho_R \rho_R')$	χ^2	$\mathbb{W}_{1\infty}$	χ^2	$\mathbb{W}_{1\infty}$	χ^2	$\mathbb{W}_{2\infty}$
0.00	0.156	0.138	0.192	0.172	0.201	0.133
0.15	0.163	0.138	0.175	0.154	0.201	0.120
0.25	0.161	0.138	0.164	0.141	0.196	0.112
0.33	0.154	0.127	0.140	0.115	0.181	0.100
0.57	0.147	0.119	0.085	0.066	0.144	0.069
0.75	0.152	0.128	0.035	0.023	0.115	0.053
$p = 2$ and $q = 3$						
$\cdot$	One Step($\hat{\Sigma}^*$)		One Step($\breve{W}$)		Two Step	
$\nu_{\max}(\rho_R \rho_R')$	χ^2	$\mathbb{W}_{1\infty}$	χ^2	$\mathbb{W}_{1\infty}$	χ^2	$\mathbb{W}_{2\infty}$
0.00	0.239	0.183	0.287	0.228	0.305	0.177
0.26	0.230	0.166	0.263	0.196	0.298	0.150
0.40	0.231	0.169	0.243	0.170	0.296	0.138
0.50	0.228	0.161	0.234	0.159	0.286	0.130
0.73	0.228	0.157	0.179	0.118	0.263	0.108
0.86	0.230	0.159	0.161	0.108	0.240	0.098

See notes to Table 9.

Table 12: Empirical size of one-step and two-step tests based on the Bartlett kernel variance estimator under VARMA(1,1) error when $\psi = 0.75$, $p = 1 \sim 2$ and $T = 200$

$p = 1$ and $q = 3$						
$\cdot$	One Step($\hat{\Sigma}^*$)		One Step($\breve{W}$)		Two Step	
$\nu_{\max}(\rho_R \rho_R')$	χ^2	$\mathbb{W}_{1\infty}$	χ^2	$\mathbb{W}_{1\infty}$	χ^2	$\mathbb{W}_{2\infty}$
0.00	0.161	0.142	0.196	0.177	0.203	0.134
0.15	0.147	0.127	0.165	0.144	0.188	0.116
0.25	0.140	0.117	0.149	0.129	0.174	0.105
0.33	0.131	0.115	0.134	0.113	0.158	0.090
0.57	0.117	0.099	0.083	0.068	0.109	0.051
0.75	0.109	0.092	0.035	0.026	0.058	0.024
$p = 2$ and $q = 3$						
$\cdot$	One Step($\hat{\Sigma}^*$)		One Step($\breve{W}$)		Two Step	
$\nu_{\max}(\rho_R \rho_R')$	χ^2	$\mathbb{W}_{1\infty}$	χ^2	$\mathbb{W}_{1\infty}$	χ^2	$\mathbb{W}_{2\infty}$
0.00	0.235	0.180	0.292	0.230	0.307	0.174
0.26	0.213	0.157	0.239	0.181	0.278	0.146
0.40	0.203	0.147	0.224	0.165	0.262	0.124
0.50	0.205	0.146	0.209	0.151	0.246	0.115
0.73	0.191	0.136	0.167	0.114	0.195	0.085
0.86	0.190	0.133	0.147	0.105	0.174	0.078

See notes to Table 9.

Table 13: Empirical size of one-step and two-step tests based on the Parzen kernel variance estimator under VAR(1) error when $\psi = 0.75$, $p = 1 \sim 2$ and $T = 200$

$p = 1$ and $q = 3$						
$\cdot$	One Step($\hat{\Sigma}^*$)		One Step($\breve{W}$)		Two Step	
$\nu_{\max}(\rho_R \rho_R')$	χ^2	$\mathbb{W}_{1\infty}$	χ^2	$\mathbb{W}_{1\infty}$	χ^2	$\mathbb{W}_{2\infty}$
0.00	0.145	0.108	0.182	0.139	0.214	0.090
0.15	0.148	0.105	0.173	0.125	0.223	0.076
0.25	0.142	0.102	0.161	0.115	0.220	0.070
0.33	0.142	0.101	0.142	0.099	0.211	0.063
0.57	0.150	0.105	0.107	0.068	0.186	0.050
0.75	0.141	0.101	0.054	0.030	0.147	0.034
$p = 2$ and $q = 3$						
$\cdot$	One Step($\hat{\Sigma}^*$)		One Step($\breve{W}$)		Two Step	
$\nu_{\max}(\rho_R \rho_R')$	χ^2	$\mathbb{W}_{1\infty}$	χ^2	$\mathbb{W}_{1\infty}$	χ^2	$\mathbb{W}_{2\infty}$
0.00	0.216	0.123	0.278	0.169	0.340	0.102
0.26	0.225	0.117	0.267	0.149	0.348	0.085
0.40	0.221	0.117	0.260	0.140	0.346	0.081
0.50	0.219	0.112	0.241	0.123	0.331	0.072
0.73	0.217	0.102	0.199	0.097	0.310	0.059
0.86	0.226	0.116	0.175	0.080	0.292	0.054

See notes to Table 9.

Table 14: Empirical size of one-step and two-step tests based on the Parzen kernel variance estimator under VARMA(1,1) error when $\psi = 0.75$, $p = 1 \sim 2$ and $T = 200$

$p = 1$ and $q = 3$						
$\cdot$	One Step($\hat{\Sigma}^*$)		One Step($\breve{W}$)		Two Step	
$\nu_{\max}(\rho_R \rho_R')$	χ^2	$\mathbb{W}_{1\infty}$	χ^2	$\mathbb{W}_{1\infty}$	χ^2	$\mathbb{W}_{2\infty}$
0.00	0.142	0.104	0.186	0.141	0.218	0.088
0.15	0.134	0.099	0.164	0.125	0.210	0.082
0.25	0.136	0.099	0.155	0.117	0.200	0.076
0.33	0.127	0.096	0.150	0.113	0.191	0.074
0.57	0.122	0.087	0.110	0.079	0.156	0.052
0.75	0.111	0.082	0.070	0.046	0.114	0.033
$p = 2$ and $q = 3$						
$\cdot$	One Step($\hat{\Sigma}^*$)		One Step($\breve{W}$)		Two Step	
$\nu_{\max}(\rho_R \rho_R')$	χ^2	$\mathbb{W}_{1\infty}$	χ^2	$\mathbb{W}_{1\infty}$	χ^2	$\mathbb{W}_{2\infty}$
0.00	0.220	0.124	0.279	0.171	0.338	0.100
0.26	0.204	0.112	0.248	0.142	0.320	0.094
0.40	0.198	0.108	0.226	0.135	0.303	0.083
0.50	0.196	0.112	0.225	0.131	0.291	0.085
0.73	0.186	0.106	0.188	0.102	0.255	0.063
0.86	0.182	0.105	0.156	0.083	0.219	0.055

See notes to Table 9.

Table 15: Empirical size of one-step and two-step tests based on the QS kernel variance estimator under VAR(1) error when $\psi = 0.75$, $p = 1 \sim 2$ and $T = 200$

$p = 1$ and $q = 3$						
$\cdot$	One Step($\hat{\Sigma}^*$)		One Step($\breve{W}$)		Two Step	
$\nu_{\max}(\rho_R \rho_R')$	χ^2	$\mathbb{W}_{1\infty}$	χ^2	$\mathbb{W}_{1\infty}$	χ^2	$\mathbb{W}_{2\infty}$
0.00	0.138	0.107	0.174	0.144	0.204	0.089
0.15	0.138	0.103	0.164	0.126	0.209	0.077
0.25	0.141	0.106	0.151	0.115	0.214	0.076
0.33	0.135	0.099	0.145	0.106	0.208	0.069
0.57	0.149	0.110	0.101	0.068	0.187	0.056
0.75	0.132	0.099	0.049	0.029	0.136	0.036
$p = 2$ and $q = 3$						
$\cdot$	One Step($\hat{\Sigma}^*$)		One Step($\breve{W}$)		Two Step	
$\nu_{\max}(\rho_R \rho_R')$	χ^2	$\mathbb{W}_{1\infty}$	χ^2	$\mathbb{W}_{1\infty}$	χ^2	$\mathbb{W}_{2\infty}$
0.00	0.210	0.124	0.265	0.168	0.312	0.101
0.26	0.217	0.122	0.261	0.151	0.335	0.089
0.40	0.216	0.119	0.244	0.141	0.327	0.084
0.50	0.214	0.114	0.234	0.130	0.332	0.077
0.73	0.204	0.113	0.188	0.099	0.295	0.063
0.86	0.214	0.121	0.158	0.082	0.277	0.063

See notes to Table 9.

Table 16: Empirical size of one-step and two-step tests based on the QS kernel variance estimator under VARMA(1,1) error when $\psi = 0.75$, $p = 1 \sim 2$ and $T = 200$

$p = 1$ and $q = 3$						
$\cdot$	One Step($\hat{\Sigma}^*$)		One Step($\breve{W}$)		Two Step	
$\nu_{\max}(\rho_R \rho_R')$	χ^2	$\mathbb{W}_{1\infty}$	χ^2	$\mathbb{W}_{1\infty}$	χ^2	$\mathbb{W}_{2\infty}$
0.00	0.141	0.112	0.175	0.141	0.204	0.090
0.15	0.137	0.110	0.164	0.132	0.201	0.089
0.25	0.130	0.104	0.149	0.117	0.188	0.076
0.33	0.123	0.096	0.140	0.111	0.178	0.074
0.57	0.117	0.094	0.113	0.088	0.152	0.058
0.75	0.110	0.085	0.060	0.042	0.110	0.034
$p = 2$ and $q = 3$						
$\cdot$	One Step($\hat{\Sigma}^*$)		One Step($\breve{W}$)		Two Step	
$\nu_{\max}(\rho_R \rho_R')$	χ^2	$\mathbb{W}_{1\infty}$	χ^2	$\mathbb{W}_{1\infty}$	χ^2	$\mathbb{W}_{2\infty}$
0.00	0.213	0.128	0.271	0.176	0.323	0.106
0.26	0.199	0.123	0.249	0.160	0.310	0.104
0.40	0.194	0.122	0.231	0.147	0.297	0.096
0.50	0.183	0.108	0.212	0.130	0.278	0.083
0.73	0.188	0.114	0.187	0.113	0.250	0.072
0.86	0.182	0.113	0.156	0.091	0.217	0.061

See notes to Table 9.

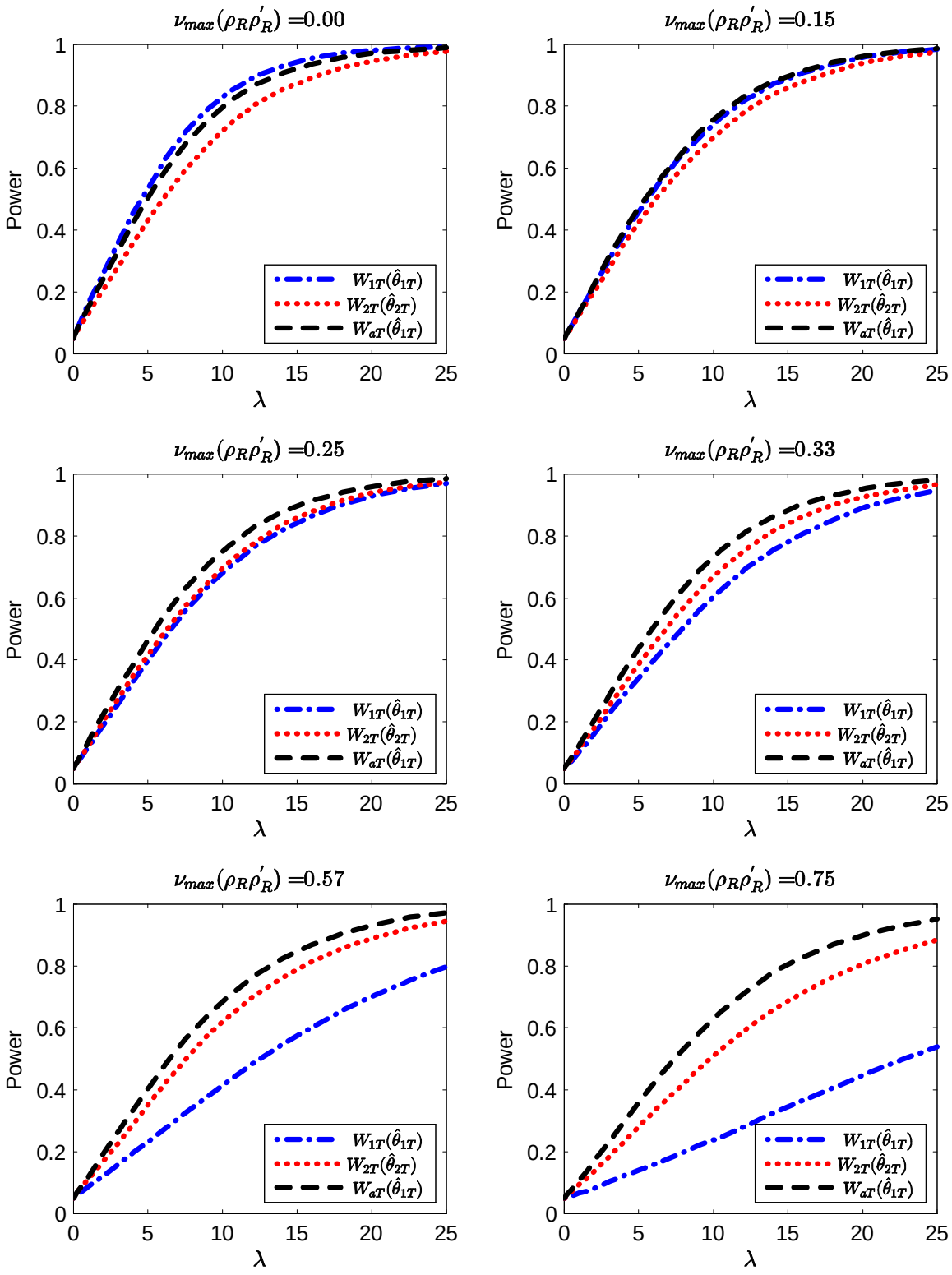

Figure 2: Size-adjusted power of the three tests based on the OS LRV estimator under VAR(1) error with $p = 1$, $q = 3$, $\psi = 0.75$, $T = 200$, and $K = 14$.

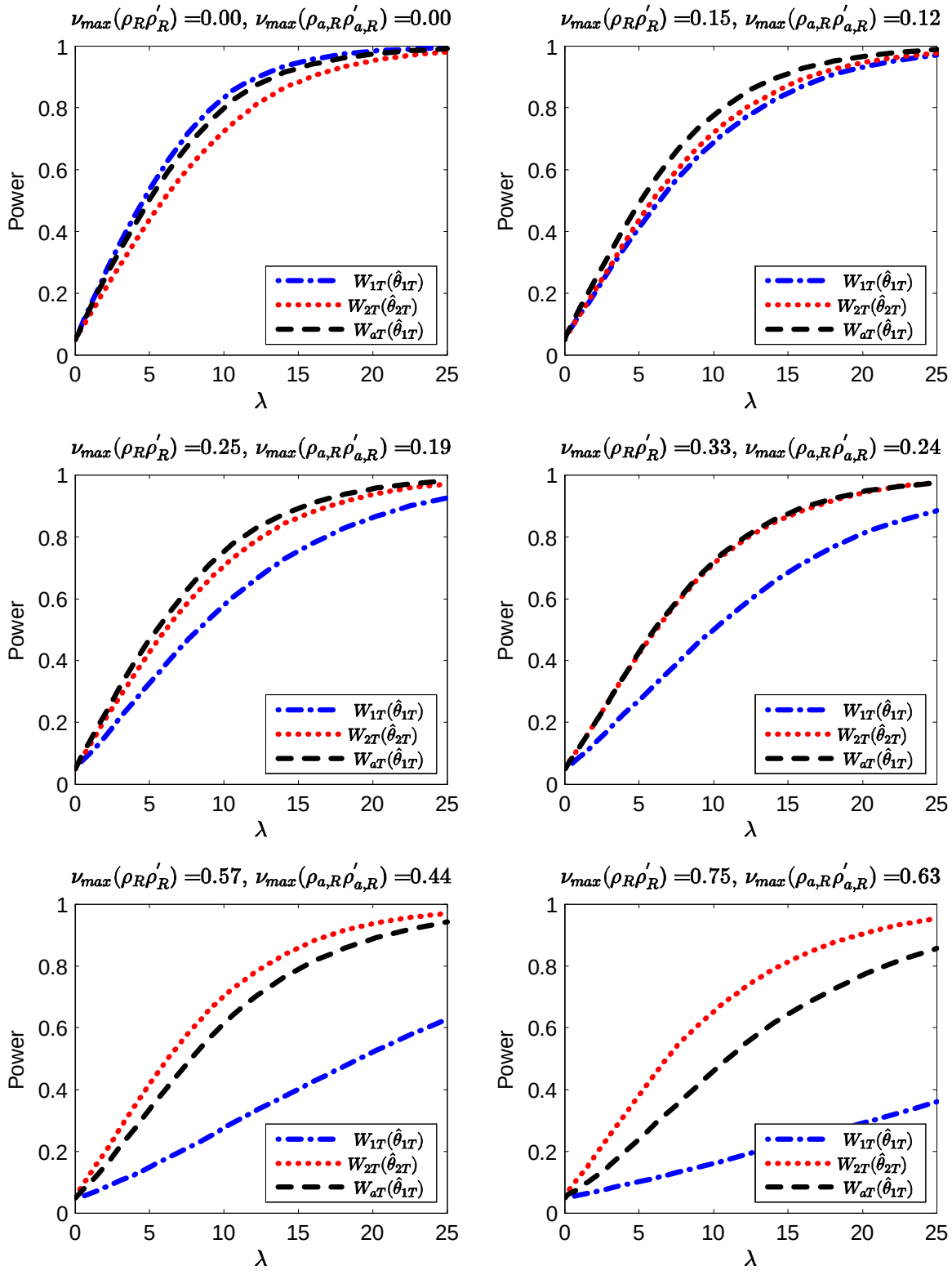

Figure 3: Size-adjusted power of the three tests based on the OS LRV estimator under VARMA(1,1) error with $p = 1$, $q = 3$, $\psi = 0.75$, $T = 200$, and $K = 14$.

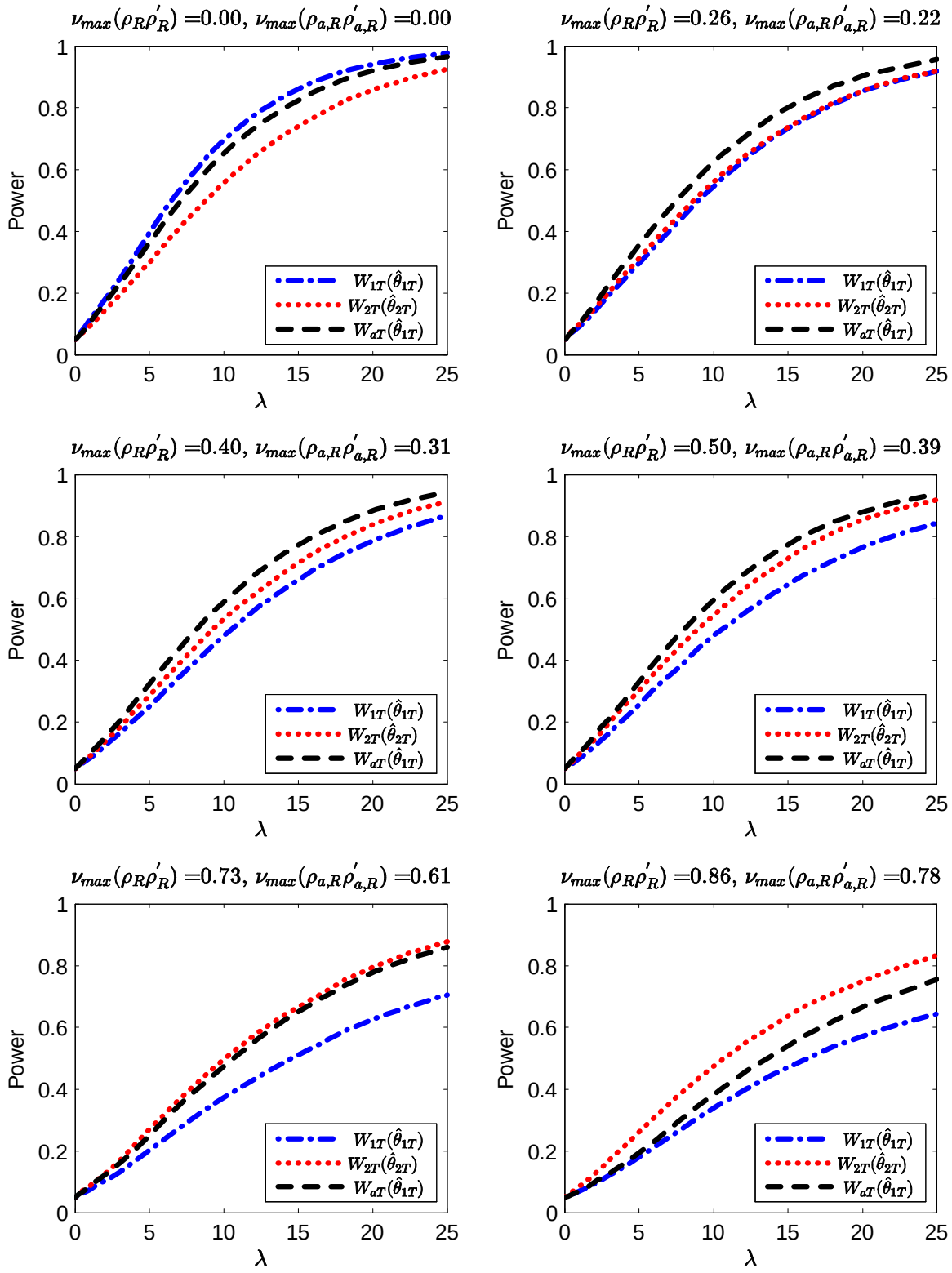

Figure 4: Size-adjusted power of the three tests based on the OS LRV estimator under VARMA(1,1) error with $p = 2$, $q = 3$, $\psi = 0.75$, $T = 200$, and $K = 14$.

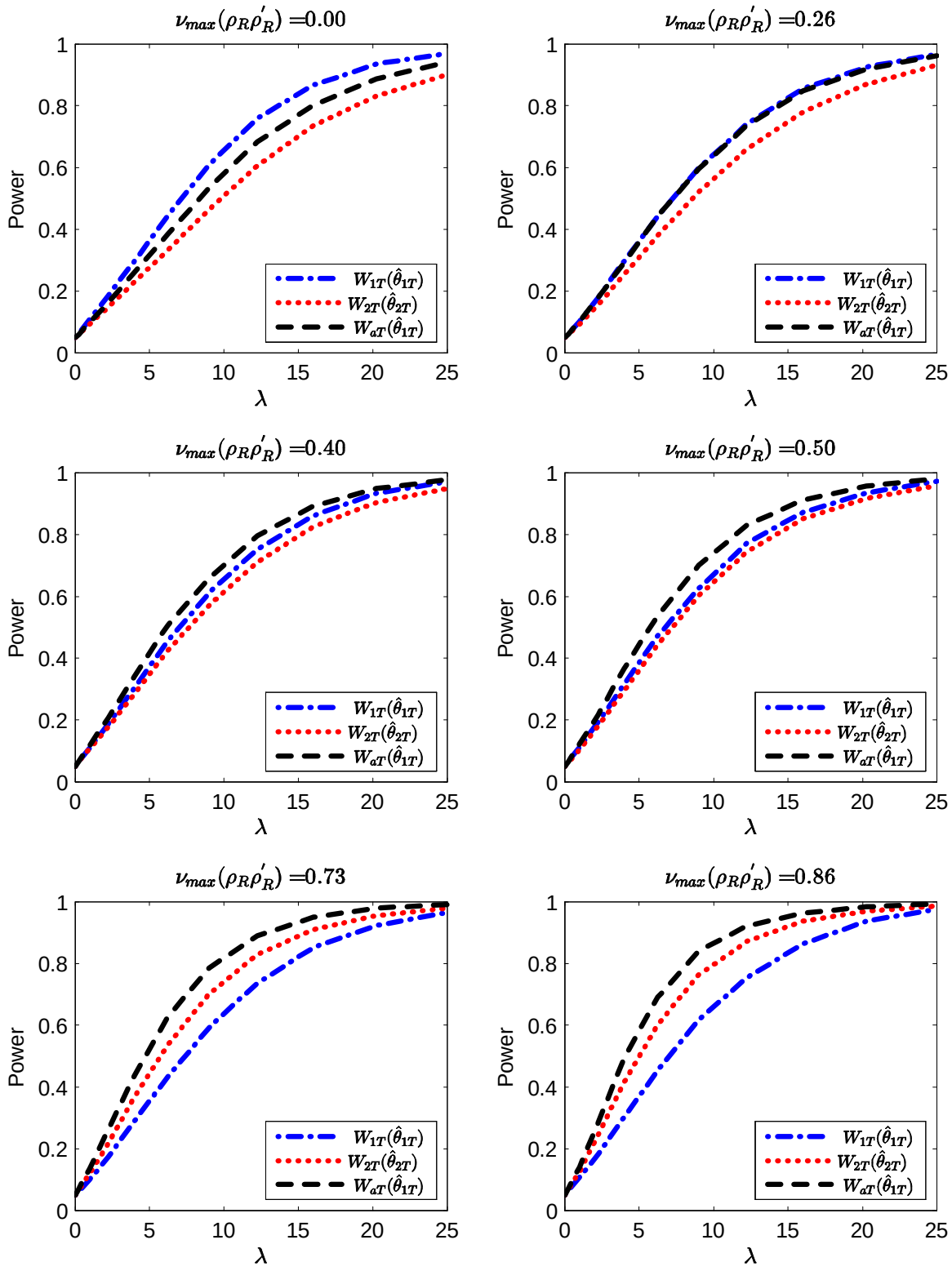

Figure 5: Size-adjusted power of the three tests based on the Bartlett LRV estimator under VAR(1) error with $p = 2$, $q = 3$, $\psi = 0.75$, $T = 200$, and $b = 0.078$.

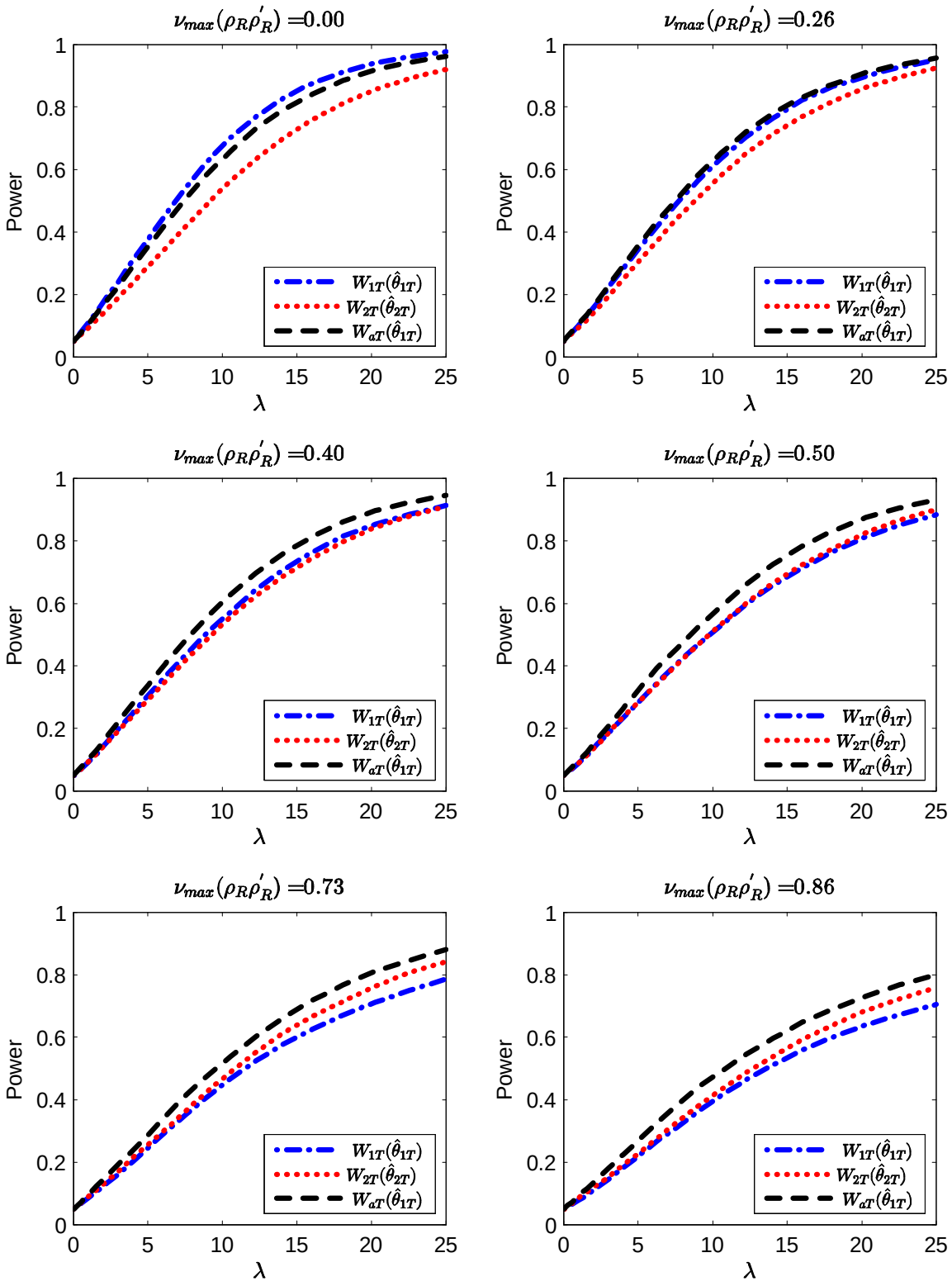

Figure 6: Size-adjusted power of the three tests based on the OS LRV estimator under VAR(1) error with $p = 2$, $q = 3$, $\psi = 0.75$, $T = 200$, and $K = 14$.

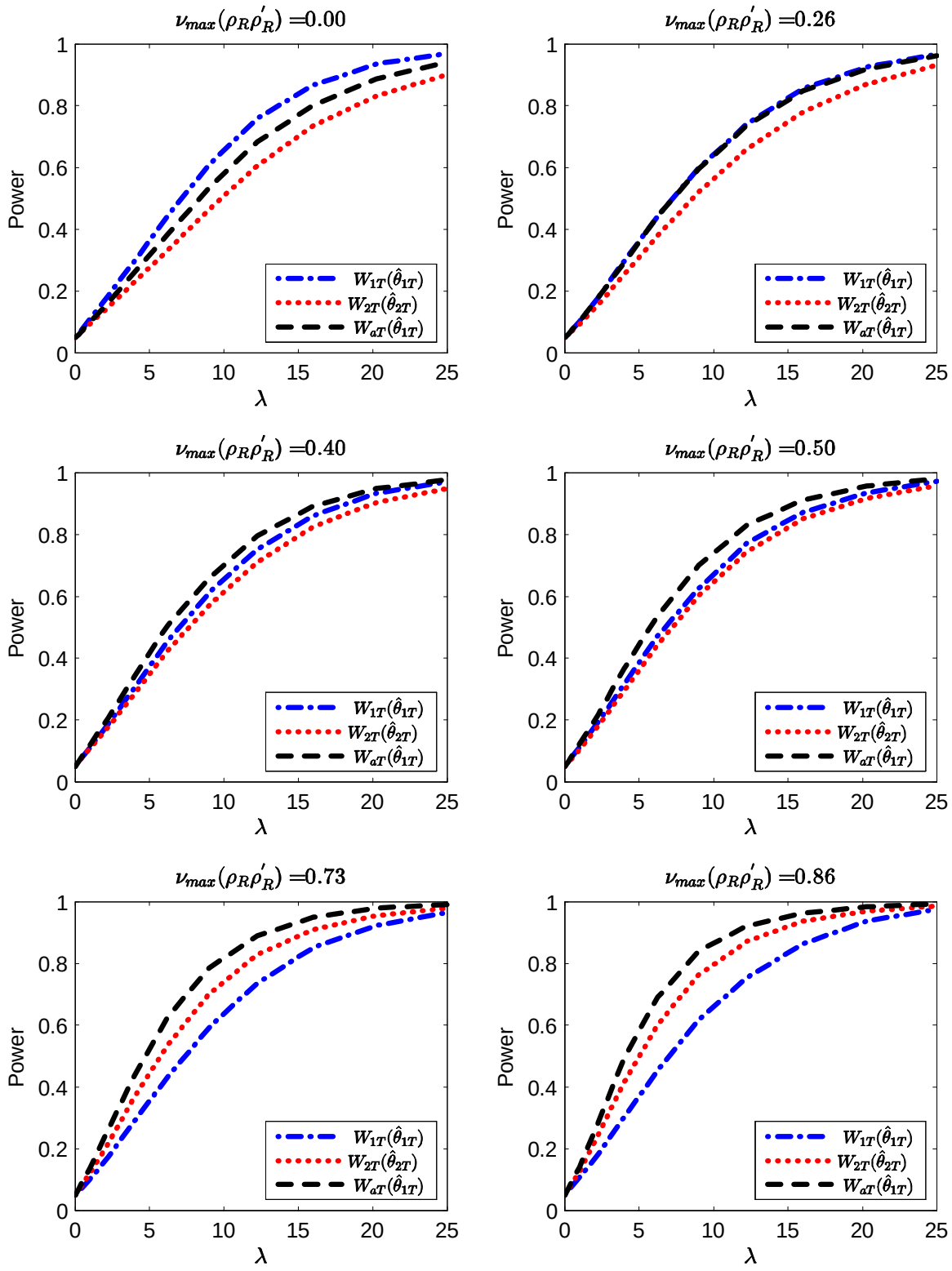

Figure 7: Size-adjusted power of the three tests based on the Parzen LRV estimator under VAR(1) error with $p = 2$, $q = 3$, $\psi = 0.75$, $T = 200$, and $b = 0.16$.

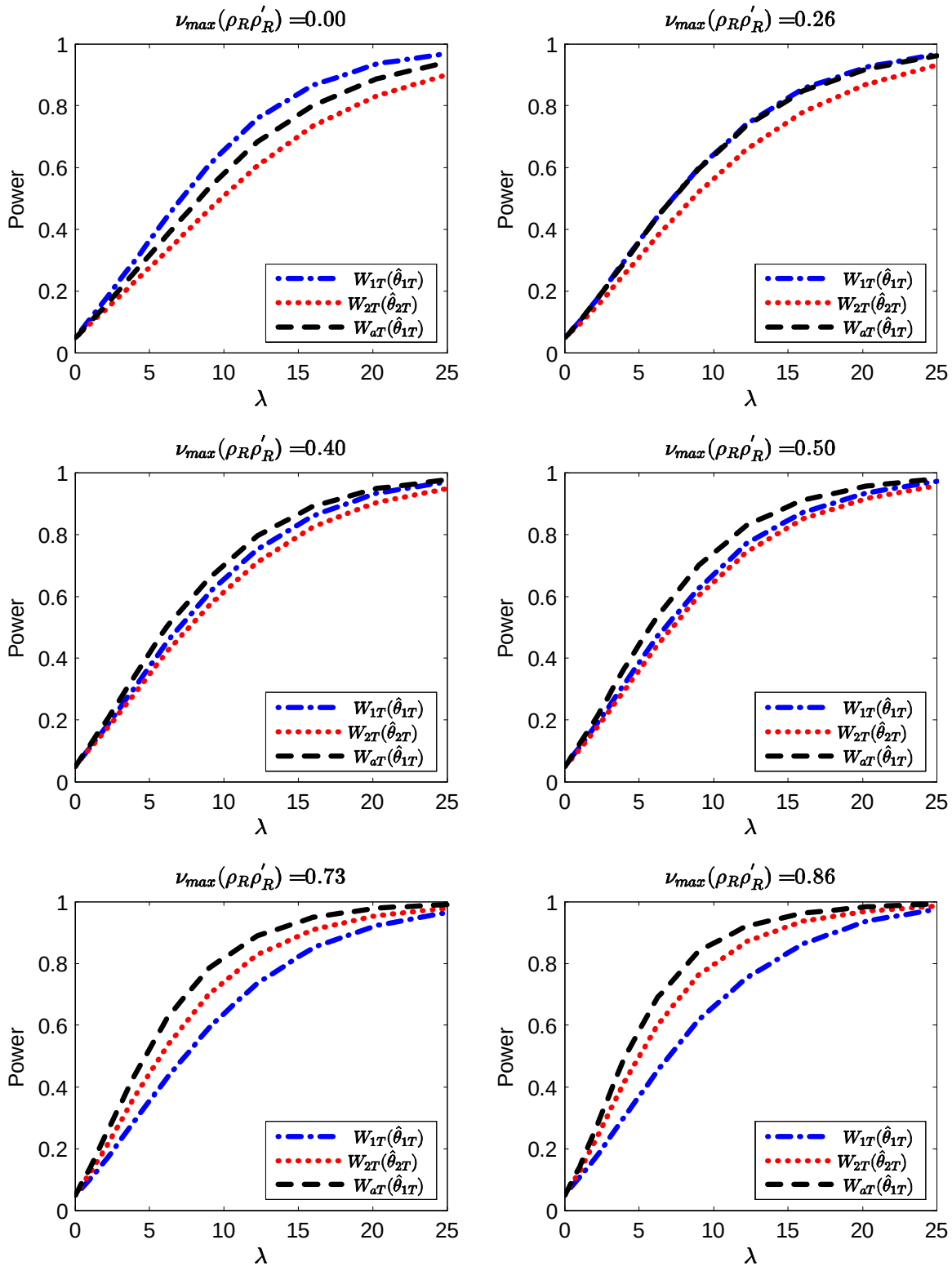

Figure 8: Size-adjusted power of the three tests based on the QS LRV estimator under VAR(1) error with $p = 2$, $q = 3$, $\psi = 0.75$, $T = 200$, and $b = 0.079$.

8 Appendix of Proofs

Proof of Proposition 1. Part (a) follows from Lemma 1 of Sun (2014b). For part (b), we note that $\hat{\beta} \overset{d}{\Longrightarrow} \beta_\infty$ and so

$$\begin{aligned}\sqrt{T}\left(\hat{\theta}_{2T}-\theta_0\right) &= \frac{1}{\sqrt{T}}\sum_{t=1}^{T}\left[\left(y_{1t}-Ey_{1t}\right)-\hat{\beta}y_{2t}\right] \\ &= \left(\begin{array}{cc} I_d, & -\hat{\beta} \end{array}\right)\left(\begin{array}{c} \frac{1}{\sqrt{T}}\sum_{t=1}^{T}\left(y_{1t}-Ey_{1t}\right) \\ \frac{1}{\sqrt{T}}\sum_{t=1}^{T}y_{2t} \end{array}\right) \overset{d}{\Longrightarrow} \left(\begin{array}{cc} I_d, & -\beta_\infty \end{array}\right)\Omega_{1/2}B_m(1).\end{aligned}$$

■

Proof of Lemma 2. For any $a \in \mathbb{R}^d$, we have

$$\begin{aligned} & Ea'\tilde{\beta}_\infty(h,d,q)\,\tilde{\beta}_\infty(h,d,q)'\,a \\ = \;& E\left[tra'\left(\int_0^1\int_0^1 Q_h^*(r,s)\,dB_d(r)\,dB_q'(s)\right)\right. \\ & \left.\times\left(\int_0^1\int_0^1 Q_h^*(r,s)\,dB_q(r)\,dB_q'(s)\right)^{-2}\left(\int_0^1\int_0^1 Q_h^*(r,s)\,dB_q(r)\,dB_d'(s)\right)a\right] \\ = \;& E\left[tr\left(\int_0^1\int_0^1 Q_h^*(r,s)\,dB_q(r)\,dB_q'(s)\right)^{-2}\right. \\ & \left.\times\left(\int_0^1\int_0^1 Q_h^*(r,s)\,dB_q(r)\,dB_d'(s)\right)aa'\left(\int_0^1\int_0^1 Q_h^*(r,s)\,dB_d(r)\,dB_q'(s)\right)\right] \\ = \;& E\left[tr\left(\int_0^1\int_0^1 Q_h^*(r,s)\,dB_q(r)\,dB_q'(s)\right)^{-2}\right. \\ & \left.\times\left(\int_0^1\int_0^1 Q_h^*(r,s)\,dB_q(r)\left[a'dB_d(s)\right]\right)\left(\int_0^1\int_0^1 Q_h^*(r,s)\left[dB_d'(r)\,a\right]dB_q'(s)\right)\right] \\ :\;& = \kappa(h,q)a'a, \end{aligned}$$

where

$$\begin{aligned} & \kappa(h,q) \\ = \;& Etr\left(\int_0^1\int_0^1 Q_h^*(r,s)\,dB_q(r)\,dB_q'(s)\right)^{-2}\left[\int_0^1\int_0^1\left(\int_0^1 Q_h^*(r,\tau)\,Q_h^*(\tau,s)\,d\tau\right)dB_q(r)\,dB_q'(s)\right]. \end{aligned} \tag{28}$$

So

$$E\tilde{\beta}_\infty(h,d,q)\,\tilde{\beta}_\infty(h,d,q)' = \kappa(h,q)\cdot I_d.$$

Since this holds for any d, we have $E\tilde{\beta}_\infty(h,1,q)\,\tilde{\beta}_\infty(h,1,q)' = \kappa(h,q)$. It then follows that

$$E\tilde{\beta}_\infty(h,d,q)\,\tilde{\beta}_\infty(h,d,q)' = \left(E\left\|\tilde{\beta}_\infty(h,1,q)\right\|^2\right)\cdot I_d.$$

■

Proof of Proposition 3. Using (4) and Lemma 2, we have

$$\begin{aligned}
& \operatorname{avar}(\hat{\theta}_{2T}) - \operatorname{avar}(\hat{\theta}_{1T}) \\
= \; & (E||\tilde{\beta}_{\infty}(h,1,q)||^2)\Omega_{1\cdot 2} - \Omega_{12}\Omega_{22}^{-1}\Omega_{21} \\
= \; & (E||\tilde{\beta}_{\infty}(h,1,q)||^2)\Omega_{11} - (1+E||\tilde{\beta}_{\infty}(h,1,q)||^2)\Omega_{12}\Omega_{22}^{-1}\Omega_{21} \\
= \; & (1+E||\tilde{\beta}_{\infty}(h,1,q)||^2)\left[g(h,q)\Omega_{11} - \Omega_{12}\Omega_{22}^{-1}\Omega_{21}\right] \\
= \; & (1+E||\tilde{\beta}_{\infty}(h,1,q)||^2)\Omega_{11}^{1/2}\left[g(h,q)I_d - \Omega_{11}^{-1/2}\Omega_{12}\Omega_{22}^{-1}\Omega_{21}(\Omega_{11}^{-1/2})'\right](\Omega_{11}^{1/2})' \\
= \; & (1+E||\tilde{\beta}_{\infty}(h,1,q)||^2)\Omega_{11}^{1/2}\left[g(h,q)I_d - \rho\rho'\right](\Omega_{11}^{1/2})'.
\end{aligned}$$

So $\operatorname{avar}(\hat{\theta}_{2T}) > \operatorname{avar}(\hat{\theta}_{1T})$ if and only if $g(h,q)I_d > \rho\rho'$. Let $\rho\rho' = Q_\rho \Lambda_\rho Q_\rho'$ be the eigen-decomposition of $\rho\rho'$ where Λ_ρ is a diagonal matrix with the eigenvalues of $\rho\rho'$ as the diagonal elements and Q_ρ is an orthogonal matrix that consists of the corresponding eigenvectors. Then $g(h,q)I_d > \rho\rho'$ if and only if $Q_\rho' g(h,q) Q_\rho > \Lambda_\rho$, which is equivalent to $g(h,q)I_d - \Lambda_\rho > 0$. The latter holds if and only if $\nu_{\max}(\rho\rho') < g(h,q)$. We have therefore proved that $\operatorname{avar}(\hat{\theta}_{2T}) > \operatorname{avar}(\hat{\theta}_{1T})$ if and only if $\nu_{\max}(\rho\rho') < g(h,q)$. Similarly, we can prove that $\operatorname{avar}(\hat{\theta}_{2T}) < \operatorname{avar}(\hat{\theta}_{1T})$ if and only if $\nu_{\min}(\rho\rho') > g(h,q)$. ■

Proof of Corollary 4. For the OS LRV estimator, we have

$$Q_h^*(r,s) = \frac{1}{K}\sum_{i=1}^{K} \Phi_i(r)\,\Phi_i(s),$$

and so

$$\begin{aligned}
\int_0^1 Q_h^*(r,\tau)\,Q_h^*(\tau,s)\,d\tau &= \int_0^1 \frac{1}{K}\sum_{i=1}^{K}\Phi_i(r)\,\Phi_i(\tau)\,\frac{1}{K}\sum_{j=1}^{K}\Phi_j(\tau)\,\Phi_j(s)\,d\tau \\
&= \frac{1}{K^2}\sum_{i=1}^{K}\Phi_i(r)\,\Phi_i(s) = \frac{1}{K}Q_h^*(r,s).
\end{aligned}$$

As a result, for $\kappa(h,q)$ defined in (28), we have:

$$\kappa(h,q) = \frac{1}{K} Etr\left(\int_0^1\int_0^1 Q_h^*(r,s)\,dB_q(r)\,dB_q'(s)\right)^{-1}.$$

Let

$$\xi_j = \int_0^1 \Phi_j(r)\,dB_q(r) \sim iidN(0, I_q),$$

then

$$\kappa(h,q) = trE\left[\left(\sum_{j=1}^{K}\xi_j\xi_j'\right)^{-1}\right] = \frac{q}{K-q-1},$$

where the last equality follows from the mean of an inverse Wishart distribution. Using this, we have

$$g(h,q) = \frac{\kappa(h,q)}{1+\kappa(h,q)} = \frac{q/(K-q-1)}{1+q/(K-q-1)} = \frac{q}{K-1}.$$

The corollary then follows from Proposition 3. ■

Proof of Proposition 5. It suffices to prove parts (a) and (b) as parts (c) and (d) follow from similar arguments. 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 $R\theta_0 = r + \delta_0/\sqrt{T}$, we have:

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

Using Proposition 1(a), we have

$$(R\hat{\Omega}_{11}R') \overset{d}{\Longrightarrow} R\Omega_{11}^{1/2}C_{dd}(R\Omega_{11}^{1/2})'$$

where $C_{dd} = \int_0^1\int_0^1 Q_h^*(r,s)dB_d(r)dB_d(s)'$ and $C_{dd} \perp B_d(1)$. Invoking the continuous mapping theorem yields

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

Now, $\left[R\Omega_{11}^{1/2}B_d(1), R\Omega_{11}^{1/2}C_{dd}(R\Omega_{11}^{1/2})'\right]$ is distributionally equivalent to $[\Lambda_1 B_p(1), \Lambda_1 C_{pp}\Lambda_1']$, and so

$$\begin{aligned}&\mathbb{W}_{1T} \overset{d}{\Longrightarrow} [\Lambda_1 B_p(1) + \delta_0]' \left[\Lambda_1 C_{pp}\Lambda_1'\right]^{-1} [\Lambda_1 B_p(1) + \delta_0] \\ &\overset{d}{=} \left[B_p(1) + \Lambda_1^{-1}\delta_0\right]' C_{pp}^{-1} \left[B_p(1) + \Lambda_1^{-1}\delta_0\right] \overset{d}{=} \mathbb{W}_{1\infty}(\left\|\Lambda_1^{-1}\delta_0\right\|^2),\end{aligned}$$

as desired. ■

Proof of Proposition 6.

Part (a) Let $\chi_p^2(\delta^2)$ be a random variable following the noncentral chi-squared distribution with degrees of freedom p and noncentrality parameter δ^2. We first prove that $P(\chi_p^2(\delta^2) > x)$ increases with δ^2 for any integer p and $x > 0$. Note that

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

where χ^2_{p+2j} is a (central) chi-squared variate with degrees of freedom $p+2j$, we have

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

as needed.

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

$$\begin{aligned}P(\|\phi+\psi\|^2 &> x) = E\left[P(\chi^2_1\left(\psi^2\right)>x)|\psi^2\right]\\ &> P(\chi^2_1>x)=P(\|\phi\|^2>x) \text{ for any } x.\end{aligned}$$

Now we proceed to prove the theorem. Note that $B_p(1)$ and $B_q(1)$ are independent of C_{pq}, C_{pp}, and C_{qq}. Let $D_{pp}^{-1}=\sum_{i=1}^p\lambda_{Di}d_id_i'$ be the spectral decomposition of D_{pp}^{-1} where $\lambda_{Di}\geq 0$ almost surely and $\{d_i\}$ are orthonormal in $\mathbb{R}^p$. Then

$$\begin{aligned}&\left[B_p(1)-C_{pq}C_{qq}^{-1}B_q(1)\right]'D_{pp}^{-1}\left[B_p(1)-C_{pq}C_{qq}^{-1}B_q(1)\right]\\ &=\sum_{i=1}^p\lambda_{Di}\left[d_i'B_p(1)-d_i'C_{pq}C_{qq}^{-1}B_q(1)\right]^2=\sum_{i=1}^p\lambda_{Di}\left(\phi_i+\psi_i\right)^2\end{aligned}$$

where $\phi_i=d_i'B_p(1)$, $\psi_i=-d_i'C_{pq}C_{qq}^{-1}B_q(1)$, $\{\phi_i\}$ is independent of $\{\psi_i\}$ conditional on C_{pq}, C_{pp}, and C_{qq}. In addition, $\phi_i\sim iidN(0,1)$ conditionally on C_{pq}, C_{pp}, and C_{qq} and unconditionally. So for any $x>0$,

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

Using the above argument repeatedly, we have

$$\begin{aligned}
&P\left(\mathbb{W}_{2\infty}\left(0\right)>x\right)\geq EP\left(\sum_{i=1}^{p}\lambda_{Di}\phi_{i}^{2}>x|C_{pq},C_{pp},C_{qq}\right)\\
&=P\left(\sum_{i=1}^{p}\lambda_{Di}\phi_{i}^{2}>x\right)=P\left[B_{p}\left(1\right)'D_{pp}^{-1}B_{p}\left(1\right)>x\right]\\
&>P\left[B_{p}\left(1\right)'C_{pp}^{-1}B_{p}\left(1\right)>x\right]=P\left(\mathbb{W}_{1\infty}\left(0\right)>x\right),
\end{aligned}$$

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

Part (b). Let $C_{pp}^{-1}=\sum_{i=1}^{p}\lambda_{Ci}c_{i}c_{i}'$ be the spectral decomposition of C_{pp}^{-1}. Since $C_{pp}>0$ with probability one, $\lambda_{ci}>0$ with probability one. We have

$$\begin{aligned}
&\mathbb{W}_{1\infty}\left(\left\|\xi\right\|^{2}\right)\overset{d}{=}\left[B_{p}\left(1\right)+\left\|\xi\right\|e_{p}\right]'C_{pp}^{-1}\left[B_{p}\left(1\right)+\left\|\xi\right\|e_{p}\right]\\
&=\sum_{i=1}^{p}\lambda_{Ci}\left[c_{i}'B_{p}\left(1\right)+\left\|\xi\right\|c_{i}'e_{p}\right]^{2}
\end{aligned}$$

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

$$\begin{aligned}
&P\left[\mathbb{W}_{1\infty}\left(\left\|\xi_{1}\right\|^{2}\right)>x\right]\\
&=P\left\{\sum_{i=1}^{p}\lambda_{Ci}\left[c_{i}'B_{p}\left(1\right)+\left\|\xi_{1}\right\|c_{i}'e_{p}\right]^{2}>x\right\}\\
&=EP\left\{\lambda_{C1}\left[c_{1}'B_{p}\left(1\right)+\left\|\xi_{1}\right\|c_{1}'e_{p}\right]^{2}>x-\sum_{i=2}^{p}\lambda_{Ci}\left[c_{i}'B_{p}\left(1\right)+\left\|\xi_{1}\right\|c_{i}'e_{p}\right]^{2}\middle|\left\{\lambda_{Ci}\right\}_{i=1}^{p},\left\{c_{i}\right\}_{i=1}^{p}\right\}\\
&<EP\left\{\lambda_{C1}\left[c_{1}'B_{p}\left(1\right)+\left\|\xi_{2}\right\|c_{1}'e_{p}\right]^{2}>x-\sum_{i=2}^{p}\lambda_{Ci}\left[c_{i}'B_{p}\left(1\right)+\left\|\xi_{1}\right\|c_{i}'e_{p}\right]^{2}\middle|\left\{\lambda_{Ci}\right\}_{i=1}^{p},\left\{c_{i}\right\}_{i=1}^{p}\right\}\\
&=P\left\{\lambda_{C1}\left[c_{1}'B_{p}\left(1\right)+\left\|\xi_{2}\right\|c_{1}'e_{p}\right]^{2}+\sum_{i=2}^{p}\lambda_{Ci}\left[c_{i}'B_{p}\left(1\right)+\left\|\xi_{1}\right\|c_{i}'e_{p}\right]^{2}>x\right\}
\end{aligned}$$

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

$$\begin{aligned}
&P\left[\mathbb{W}_{1\infty}\left(\left\|\xi_{1}\right\|^{2}\right)>x\right]\\
&<P\left\{\lambda_{C1}\left[c_{1}'B_{p}\left(1\right)+\left\|\xi_{2}\right\|c_{1}'e_{p}\right]^{2}+\lambda_{C2}\left[c_{2}'B_{p}\left(1\right)+\left\|\xi_{2}\right\|c_{2}'e_{p}\right]^{2}+\sum_{i=3}^{p}\lambda_{Ci}\left[c_{i}'B_{p}\left(1\right)+\left\|\xi_{1}\right\|c_{i}'e_{p}\right]^{2}>x\right\}\\
&<P\left\{\sum_{i=1}^{p}\lambda_{Ci}\left[c_{i}'B_{p}\left(1\right)+\left\|\xi_{2}\right\|c_{i}'e_{p}\right]^{2}>x\right\}\\
&=P\left\{\left[B_{p}\left(1\right)+\xi_{2}\right]'C_{pp}^{-1}\left[B_{p}\left(1\right)+\xi_{2}\right]>x\right\}=P\left[\mathbb{W}_{1\infty}\left(\left\|\xi_{2}\right\|^{2}\right)>x\right]
\end{aligned}$$

as desired.

Part (c). We note that

$$\begin{aligned}
&\mathbb{W}_{2\infty}\left(\|\xi\|^2\right) \\
&= \left[B_p(1) - C_{pq}C_{qq}^{-1}B_q(1) + \|\xi\| e_p\right]' D_{pp}^{-1} \left[B_p(1) - C_{pq}C_{qq}^{-1}B_q(1) + \|\xi\| e_p\right] \\
&= \left\{\left[I_p + C_{pq}C_{qq}^{-1}C_{qq}^{-1}C_{qp}\right]^{-1/2}\left[B_p(1) - C_{pq}C_{qq}^{-1}B_q(1)\right] + \|\xi\|\tilde{e}_p\right\}' \\
&\times \left[I_p + C_{pq}C_{qq}^{-1}C_{qq}^{-1}C_{qp}\right]^{1/2} D_{pp}^{-1}\left[I_p + C_{pq}C_{qq}^{-1}C_{qq}^{-1}C_{qp}\right]^{1/2} \\
&\times \left\{\left[I_p + C_{pq}C_{qq}^{-1}C_{qq}^{-1}C_{qp}\right]^{-1/2}\left[B_p(1) - C_{pq}C_{qq}^{-1}B_q(1)\right] + \|\xi\|\tilde{e}_p\right\}
\end{aligned}$$

where

$$\tilde{e}_p = \left[I_p + C_{pq}C_{qq}^{-1}C_{qq}^{-1}C_{qp}\right]^{-1/2} e_p.$$

Let $\sum_{i=1}^p \tilde{\lambda}_{D_i}\tilde{d}_i\tilde{d}_i'$ be the spectral decomposition of $\left[I_p + C_{pq}C_{qq}^{-1}C_{qq}^{-1}C_{qp}\right]^{1/2} D_{pp}^{-1}\left[I_p + C_{pq}C_{qq}^{-1}C_{qq}^{-1}C_{qp}\right]^{1/2}$. Define

$$\tilde{\phi}_{di} = \tilde{d}_i'\left[I_p + C_{pq}C_{qq}^{-1}C_{qq}^{-1}C_{qp}\right]^{-1/2}\left[B_p(1) - C_{pq}C_{qq}^{-1}B_q(1)\right].$$

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

$$\begin{aligned}
&\mathbb{W}_{2\infty}(\|\xi_1\|^2) \\
&= \sum_{i=1}^p \tilde{\lambda}_{D_i}\left\{\tilde{d}_i'\left[I_p + C_{pq}C_{qq}^{-1}C_{qq}^{-1}C_{qp}\right]^{-1/2}\left[B_p(1) - C_{pq}C_{qq}^{-1}B_q(1)\right] + \|\xi_1\| d_i'\tilde{e}_p\right\}^2 \\
&= \sum_{i=1}^p \tilde{\lambda}_{D_i}\left(\tilde{\phi}_{di} + \|\xi_1\|\tilde{d}_i'\tilde{e}_p\right)^2,
\end{aligned}$$

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

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

■

Proof of Proposition 7. We prove part (b) only as part (a) can be proved using the same argument. Using (11), we have, for $\lambda_0 = \left\|\Lambda_2^{-1}\delta_0\right\|^2$:

$$\begin{aligned} & \left\|\Lambda_2^{-1}\delta_0\right\|^2 - \tau(\lambda_0)\left\|\Lambda_1^{-1}\delta_0\right\|^2 \\ = & \tau(\lambda_0)\sum_{i=1}^{p}\frac{1}{1-\nu_{i,R}}\left[\nu_{i,R} - f(\lambda_0)\right]\left(a_{i,R}'\Lambda_1^{-1}\delta_0\right)^2 \\ = & \tau(\lambda_0)\left\|\Lambda_1^{-1}\delta_0\right\|^2\sum_{i=1}^{p}\frac{1}{1-\nu_{i,R}}\left[\nu_{i,R} - f(\lambda_0)\right]\left\langle a_{i,R}, \frac{\Lambda_1^{-1}\delta_0}{\left\|\Lambda_1^{-1}\delta_0\right\|}\right\rangle^2, \end{aligned} \tag{29}$$

where $\nu_{i,R} \in [0,1)$ and $\langle\cdot,\cdot\rangle$ is the usual inner product.

We proceed to show that $\left\|\Lambda_2^{-1}\delta_0\right\|^2 - \tau(\lambda_0)\left\|\Lambda_1^{-1}\delta_0\right\|^2 > 0$ for all $\delta_0 \in \mathfrak{A}(\lambda_0)$ if and only if $\nu_{i,R} - f(\lambda_0) > 0$ for all $i = 1, ..., p$. The "if" part is obvious. To show the "only if" part, we prove by contradiction. Suppose that $\left\|\Lambda_2^{-1}\delta_0\right\|^2 - f(\lambda_0)\left\|\Lambda_1^{-1}\delta_0\right\|^2 > 0$ for all $\delta_0 \in \mathfrak{A}(\lambda_0)$ but there exists an i^* such that $\nu_{i^*,R} - f(\lambda_0) \leq 0$. Choosing $\delta_0 \in \mathfrak{A}(\lambda_0)$ such that $\left(\Lambda_1^{-1}\delta_0\right)/\left\|\Lambda_1^{-1}\delta_0\right\| = a_{i^*,R}$, we have

$$\left\|\Lambda_2^{-1}\delta_0\right\|^2 - \tau(\lambda_0)\left\|\Lambda_1^{-1}\delta_0\right\|^2 = \frac{\left\|\Lambda_1^{-1}\delta_0\right\|^2}{1-\nu_{i^*,R}}\left[\nu_{i^*,R} - f(\lambda_0)\right]\tau(\lambda_0) \leq 0, \tag{30}$$

leading to a contradiction.

Note that the condition $\nu_{i,R} - f(\lambda_0) > 0$ for all $i = 1, ..., p$ is equivalent to $\min\{\nu_{i,R}\} > f(\lambda_0)$, which is the same as $\nu_{\min}(\rho_R\rho_R') > f(\lambda_0; h, p, q, \alpha)$. This completes the proof of part (b). ■

Proof of Proposition 8. Instead of directly proving $\pi_1(\lambda) > \pi_2(\lambda)$ for any $\lambda > 0$, we consider the following testing problem: we observe $(Y, S) \in \mathbb{R}^{p+q} \times \mathbb{R}^{(p+q)\times(p+q)}$ with $Y \perp S$ from the following distributions:

$$\begin{aligned} \underset{(p+q)\times 1}{Y} &= \begin{pmatrix} \underset{(p\times 1)}{Y_1} \\ \underset{(q\times 1)}{Y_2} \end{pmatrix} \sim N_{p+q}(\mu, \Omega) \text{ with } \mu = \begin{pmatrix} \underset{(p\times 1)}{\delta_0} \\ \underset{(q\times 1)}{0} \end{pmatrix}, \Omega = \begin{pmatrix} \underset{(p\times p)}{\Omega_{11}} & \underset{(p\times q)}{0} \\ \underset{(q\times p)}{0} & \underset{(q\times q)}{\Omega_{22}} \end{pmatrix} \\ \underset{(p+q)\times(p+q)}{S} &= \begin{pmatrix} \underset{(p\times p)}{S_{11}} & \underset{(p\times q)}{S_{12}} \\ \underset{(q\times p)}{S_{21}} & \underset{(q\times q)}{S_{22}} \end{pmatrix} \sim \frac{\mathcal{W}_{p+q}(K, \Omega)}{K} \end{aligned}$$

where Ω_{11} and Ω_{22} are non-singular matrices and $\mathcal{W}_{p+q}(K, \Omega)$ is the Wishart distribution with K degrees of freedom. We want to test $H_0 : \delta_0 = 0$ against $H_1 : \delta_0 \neq 0$. The testing problem is partially motivated by Das Gupta and Perlman (1974) and Marden and Perlman (1980).

The joint pdf of (Y, S) can be written as

$$\begin{aligned} & f(Y, S|\delta_0, \Omega_{11}, \Omega_{22}) \\ = & \alpha(\delta_0, \Omega_{11}, \Omega_{22})h(S)\exp\left\{-\frac{1}{2}tr\left[\Omega_{11}^{-1}(Y_1Y_1' + KS_{11}) + \Omega_{22}^{-1}(Y_2Y_2' + KS_{22})\right] + Y_1'\Omega_{11}^{-1}\delta_0\right\} \end{aligned}$$

for some functions $\alpha(\cdot)$ and $h(\cdot)$. It follows from the exponential structure that

$$\Pi := (Y_1, S_{11}, Y_2Y_2' + KS_{22})$$

is a complete sufficient statistic for

$$\Gamma := (\delta_0, \Omega_{11}, \Omega_{22}).$$

We note that $Y_1 \sim N(\delta_0, \Omega_{11})$, $KS_{11} \sim \mathcal{W}_p(K, \Omega_{11})$ and $Y_2Y_2' + KS_{22} \sim \mathcal{W}_q(K+1, \Omega_{22})$ and these three random variables are mutually independent.

Now, we define the following two test functions for testing $H_0 : \delta_0 = 0$ against $H_1 : \delta_0 \neq 0$:

$$\begin{aligned} \phi_1(\Pi) &: = 1(\mathbb{V}_1(\Pi) > \mathbb{W}_{1\infty}^{\alpha}) \\ \phi_2(\Pi) &: = E[1(\mathbb{W}_2(Y,S) > \mathbb{W}_{2\infty}^{\alpha})|\Pi] \end{aligned}$$

where

$$\mathbb{V}_1(\Pi) := Y_1'S_{11}^{-1}Y_1 \text{ and } \mathbb{W}_2(Y,S) := (Y_1 - S_{12}S_{22}^{-1}Y_2)'(S_{11} - S_{12}S_{22}^{-1}S_{21})^{-1}(Y_1 - S_{12}S_{22}^{-1}Y_2).$$

We can show that the distributions of $\mathbb{V}_1(\Pi)$ and $\mathbb{W}_2(Y,S)$ depend on the parameter Γ only via $\delta_0'\Omega_{11}^{-1}\delta_0$. First, it is easy to show that

$$\mathbb{W}_2(Y,S) = \begin{pmatrix} Y_1 \\ Y_2 \end{pmatrix}' \begin{pmatrix} S_{11} & S_{12} \\ S_{21} & S_{22} \end{pmatrix}^{-1} \begin{pmatrix} Y_1 \\ Y_2 \end{pmatrix} - Y_2'S_{22}^{-1}Y_2.$$

Let

$$\begin{aligned} \tilde{Y} &: = \begin{pmatrix} \tilde{Y}_1 \\ \tilde{Y}_2 \end{pmatrix} = \Omega^{-1/2} \begin{pmatrix} Y_1 \\ Y_2 \end{pmatrix} \sim N(\tilde{\delta}, I_{p+q}),\ \tilde{\delta} = \begin{pmatrix} \Omega_{11}^{-1/2}\delta_0 \\ 0 \end{pmatrix} \text{ and} \\ \tilde{S} &: = \begin{pmatrix} \tilde{S}_{11} & \tilde{S}_{12} \\ \tilde{S}_{21} & \tilde{S}_{22} \end{pmatrix} = \Omega^{-1/2} \begin{pmatrix} S_{11} & S_{12} \\ S_{21} & S_{22} \end{pmatrix} (\Omega^{-1/2})' \sim \frac{\mathcal{W}_{p+q}(K, I_{p+q})}{K}. \end{aligned}$$

Then $\tilde{Y} \perp \tilde{S}$ and

$$\mathbb{W}_2(Y,S) = \left(\tilde{Y} + \tilde{\delta}\right)' \tilde{S}^{-1} \left(\tilde{Y} + \tilde{\delta}\right) - \tilde{Y}_2'\tilde{S}_{22}^{-1}\tilde{Y}_2.$$

It is now obvious that the distribution of $\mathbb{W}_2(Y,S)$ depends on Γ only via $||\tilde{\delta}||^2$, which is equal to $\delta_0'\Omega_{11}^{-1}\delta_0$. Second, we have

$$\mathbb{V}_1(\Pi) = \left(\tilde{Y}_1 + \Omega_{11}^{-1/2}\delta_0\right)' \tilde{S}_{11}^{-1} \left(\tilde{Y}_1 + \Omega_{11}^{-1/2}\delta_0\right)$$

and so the distribution of $\mathbb{V}_1(\Pi)$ depends on Γ only via $\left\|\Omega_{11}^{-1/2}\delta_0\right\|^2$ which is also equal to $\delta_0'\Omega_{11}^{-1}\delta_0$.

It is easy to show that the null distributions of $\mathbb{V}_1(\Pi)$ and $\mathbb{W}_2(Y,S)$ are the same as $\mathbb{W}_{1\infty}$ and $\mathbb{W}_{2\infty}$, respectively. In view of the critical values used, both the tests $\phi_1(\Pi)$ and $\phi_2(\Pi)$ have the correct level α. Since

$$E\phi_1(\Pi) = P(\mathbb{V}_1(\Pi) > \mathbb{W}_{1\infty}^{\alpha}) \text{ and } E\phi_2(\Pi) = E\left\{E[1(\mathbb{W}_2(Y,S) > \mathbb{W}_{2\infty}^{\alpha})|\Pi]\right\} = P(\mathbb{W}_2(Y,S) > \mathbb{W}_{1\infty}^{\alpha}),$$

the power functions of the two tests $\phi_1(\Pi)$ and $\phi_2(\Pi)$ are $\pi_1(\delta_0'\Omega_{11}^{-1}\delta_0)$ and $\pi_2(\delta_0'\Omega_{11}^{-1}\delta_0)$, respectively.

We consider a group of transformations G, which consists of the elements in $\mathbb{A}^{p\times p} := \{A \in \mathbb{R}^p \times \mathbb{R}^p : A \text{ is a } (p\times p) \text{ non-singular matrix}\}$ and acts on the sample space $\mathbf{\Pi} := \mathbb{R}^p \times \mathbb{R}^{p\times p} \times \mathbb{R}^{q\times q}$ for the sufficient statistic Π through the mapping

$$G : (Y_1, S_{11}, Y_2Y_2' + KS_{22}) \Rightarrow (AY_1, AS_{11}A', Y_2Y_2' + KS_{22}).$$

The induced group of transformations $\bar{G}$ acting on the parameter space $\mathbf{\Gamma} := \mathbb{R}^p \times \mathbb{S}^{p\times p} \times \mathbb{S}^{q\times q}$ is given by

$$\bar{G} : \Gamma = (\delta_0, \Omega_{11}, \Omega_{22}) \Rightarrow (A\delta_0, A\Omega_{11}A', \Omega_{22}).$$

Our testing problem is obviously invariant to this group of transformations.

Define

$$\mathbb{V}(\Pi) := (Y_1'S_{11}^{-1}Y_1, Y_2Y_2' + KS_{22}) := (\mathbb{V}_1(\Pi), \mathbb{V}_2(\Pi)).$$

It is clear that $\mathbb{V}(\Pi)$ is invariant under G. We can also show that $\mathbb{V}(\Pi)$ is maximal invariant under G. To do so, we consider two different samples $\Pi := (Y_1, S_{11}, Y_2Y_2' + KS_{22})$ and $\breve{\Pi} := (\breve{Y}_1, \breve{S}_{11}, \breve{Y}_2\breve{Y}_2' + K\breve{S}_{22})$ such that $\mathbb{V}(\Pi) = \mathbb{V}(\breve{\Pi})$. We want to show that there exists a $p \times p$ non-singular matrix A such that $Y_1 = A\breve{Y}_1$ and $S_{11} = A\breve{S}_{11}A'$ whenever $Y_1'S_{11}^{-1}Y_1 = \breve{Y}_1'\breve{S}_{11}^{-1}\breve{Y}_1$. By Theorem A9.5 (Vinograd's Theorem) in Muirhead (2009), there exists an orthogonal $p\times p$ matrix H such that $S_{11}^{-1/2}Y_1 = H\breve{S}_{11}^{-1/2}\breve{Y}_1$ and this gives us the non-singular matrix $A := S_{11}^{1/2}H\breve{S}_{11}^{-1/2}$ satisfying $Y_1 = A\breve{Y}_1$ and $S_{11} = A\breve{S}_{11}A'$. Similarly, we can show that

$$v(\Gamma) := (\delta_0'\Omega_{11}^{-1}\delta_0, \Omega_{22})$$

is maximal invariant under the induced group $\bar{G}$. Therefore, restricting attention to G-invariant tests, testing $H_0 : \delta_0 = 0$ against $H_1 : \delta_0 \neq 0$ reduces to testing

$$H_0' : \delta_0'\Omega_{11}^{-1}\delta_0 = 0 \text{ against } H_1' : \delta_0'\Omega_{11}^{-1}\delta_0 > 0$$

based on the maximal invariant statistic $\mathbb{V}(\Pi)$.

Let $f(\mathbb{V}_1; \delta_0'\Omega_{11}^{-1}\delta_0)$ and $f(\mathbb{V}_2; \Omega_{22})$ be the marginal pdf's of $\mathbb{V}_1 := \mathbb{V}_1(\Pi)$ and $\mathbb{V}_2 := \mathbb{V}_2(\Pi)$. By construction, $\mathbb{V}_1(\Pi)K/(K-p+1)$ follows the noncentral F distribution $F_{p,K-p+1}(\delta_0'\Omega_{11}^{-1}\delta_0)$. So $f(\mathbb{V}_1; \delta_0'\Omega_{11}^{-1}\delta_0)$ is the (scaled) pdf of the noncentral F distribution. It is well known that the noncentral F distribution has the Monotone Likelihood Ratio (MLR) property in $\mathbb{V}_1$ with respect to the parameter $\delta_0'\Omega_{11}^{-1}\delta_0$ (e.g. Chapter 7.9 in Lehmann and Romano (2008)). Also, in view of the independence between $\mathbb{V}_1$ and $\mathbb{V}_2$, the joint distribution of $\mathbb{V}(\Pi)$ also has the MLR property in $\mathbb{V}_1$. By the virtue of the Neyman-Pearson lemma, the test $\phi_1(\Pi) := 1(\mathbb{V}_1(\Pi) > \mathbb{W}_{1\infty}^{\alpha})$ is the unique Uniformly Most Powerful Invariant (UMPI) test among all G-invariant tests based on the complete sufficient statistic Π. So if $\phi_2(\Pi)$ is equivalent to a G-invariant test, then $\pi_1(\delta_0'\Omega_{11}^{-1}\delta_0) > \pi_2(\delta_0'\Omega_{11}^{-1}\delta_0)$ for any $\delta_0'\Omega_{11}^{-1}\delta_0 > 0$. To show that $\phi_2(\Pi)$ has this property, we let $g \in G$ be any element of G with the corresponding matrix A_g and induced transformation $\bar{g} \in \bar{G}$. Then,

$$\begin{aligned} E_\Gamma[\phi_2(g\Pi)] &= E_{\bar{g}\Gamma}[\phi_2(\Pi)] = \pi_2\left((A_g\delta_0)'(A_g\Omega_{11}A_g')^{-1}(A_g\delta_0)\right) \\ &= \pi_2(\delta_0'\Omega_{11}^{-1}\delta_0) = E_\Gamma[\phi_2(\Pi)] \end{aligned}$$

for all Γ. It follows from the completeness of Π that $\phi_2(g\Pi) = \phi_2(\Pi)$ almost surely and this drives the desired result. ■

Proof of Lemma 9. 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 $M_{1\cdot 2}=M_{11}-M_{12}M_{22}^{-1}M_{21}$ 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

$$\begin{aligned}
\breve{G}'\breve{M}^{-1}\breve{G} &= V\Xi'(U'\breve{M}U)^{-1}\Xi V'\\
&= VA\begin{pmatrix} I_d, & O\end{pmatrix}(M^*)^{-1}\begin{pmatrix} I_d, & O\end{pmatrix}'AV'\\
&= VA\begin{pmatrix} I_d, & O\end{pmatrix}(\Sigma_{1/2}^{*-1})'\left[\Sigma_{1/2}^{*-1}M^*(\Sigma_{1/2}^{*-1})'\right]^{-1}\Sigma_{1/2}^{*-1}\begin{pmatrix} I_d, & O\end{pmatrix}'AV'\\
&= VA\begin{pmatrix} I_d, & O\end{pmatrix}(\Sigma_{1/2}^{*-1})'M^{-1}\Sigma_{1/2}^{*-1}\begin{pmatrix} I_d, & O\end{pmatrix}'AV'\\
&= VA(\Sigma_{1\cdot 2}^*)^{-1/2}\begin{pmatrix} I_d, & O\end{pmatrix}M^{-1}\left[VA(\Sigma_{1\cdot 2}^*)^{-1/2}\begin{pmatrix} I_d, & O\end{pmatrix}\right]'\\
&= VA(\Sigma_{1\cdot 2}^*)^{-1/2}M_{1\cdot 2}^{-1}(\Sigma_{1\cdot 2}^*)^{-1/2}AV', \qquad (31)
\end{aligned}$$

where we have used

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

In addition,

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

Hence

$$\begin{aligned}
&\frac{1}{\sqrt{T}}\sum_{t=1}^{T}\left[\breve{G}'\breve{M}^{-1}\breve{G}\right]^{-1}\breve{G}'\breve{M}^{-1}\breve{f}(v_t,\theta_0)\\
=\;&\frac{1}{\sqrt{T}}\sum_{t=1}^{T}\left[VA\left(\Sigma^*_{1\cdot2}\right)^{-1/2}M_{1\cdot2}^{-1}\left(\Sigma^*_{1\cdot2}\right)^{-1/2}AV'\right]^{-1}\left[VA\left(\Sigma^*_{1\cdot2}\right)^{-1/2}M_{1\cdot2}^{-1}\right]\\
&\times\left[f_1\left(v_t,\theta_0\right)-M_{12}M_{22}^{-1}f_2\left(v_t,\theta_0\right)\right]\\
=\;&\frac{1}{\sqrt{T}}\sum_{t=1}^{T}VA^{-1}\left(\Sigma^*_{1\cdot2}\right)^{1/2}\left[f_1\left(v_t,\theta_0\right)-M_{12}M_{22}^{-1}f_2\left(v_t,\theta_0\right)\right]. \qquad (32)
\end{aligned}$$

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

$$\frac{1}{\sqrt{T}}\sum_{t=1}^{T}\left[\breve{G}'\breve{M}^{-1}\breve{G}\right]^{-1}\breve{G}'\breve{M}^{-1}\breve{f}(v_t,\theta_0)=\frac{1}{\sqrt{T}}\sum_{t=1}^{T}VA^{-1}\left(\Sigma^*_{1\cdot2}\right)^{1/2}f_1\left(v_t,\theta_0\right).$$

Using this and the stochastic expansion of $\sqrt{T}(\hat{\theta}_{1T}-\theta_0)$, we have

$$\sqrt{T}(\hat{\theta}_{1T}-\theta_0)=\frac{1}{\sqrt{T}}\sum_{t=1}^{T}VA^{-1}\left(\Sigma^*_{1\cdot2}\right)^{1/2}f_1\left(v_t,\theta_0\right)+o_p(1).$$

It then follows that

$$\left(\Sigma^*_{1\cdot2}\right)^{-1/2}AV'\sqrt{T}(\hat{\theta}_{1T}-\theta_0)=\frac{1}{\sqrt{T}}\sum_{t=1}^{T}f_1(v_t,\theta_0)+o_p(1)\overset{d}{\Longrightarrow}N(0,\Omega_{11}).$$

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

$$\frac{1}{\sqrt{T}}\sum_{t=1}^{T}\left[\breve{G}'\breve{\Omega}_\infty^{-1}\breve{G}\right]^{-1}\breve{G}'\breve{\Omega}_\infty^{-1}\breve{f}(v_t,\theta_0)=\frac{1}{\sqrt{T}}\sum_{t=1}^{T}VA^{-1}\left(\Sigma^*_{1\cdot2}\right)^{1/2}\left[f_1\left(v_t,\theta_0\right)-\beta_\infty f_2\left(v_t,\theta_0\right)\right].$$

Using this, we have

$$\begin{aligned}
\sqrt{T}(\hat{\theta}_{2T}-\theta_0)&=\frac{1}{\sqrt{T}}\sum_{t=1}^{T}\left[\breve{G}'\breve{\Omega}_\infty^{-1}\breve{G}\right]^{-1}\breve{G}'\breve{\Omega}_\infty^{-1}\breve{f}(v_t,\theta_0)+o_p\left(1\right)\\
&=\frac{1}{\sqrt{T}}VA^{-1}\left(\Sigma^*_{1\cdot2}\right)^{1/2}\sum_{t=1}^{T}\left(f_1(v_t,\theta_0)-\beta_\infty f_2(v_t,\theta_0)\right)+o_p(1).
\end{aligned}$$

It then follows that

$$\begin{aligned}
\left(\Sigma^*_{1\cdot2}\right)^{-1/2}AV'\sqrt{T}(\hat{\theta}_{2T}-\theta_0)&=\frac{1}{\sqrt{T}}\sum_{t=1}^{T}\left[f_1(v_t,\theta_0)-\beta_\infty f_2(v_t,\theta_0)\right]+o_p(1) \qquad (33)\\
&\overset{d}{\Longrightarrow}MN\left(0,\Omega_{11}-\Omega_{12}\beta'_\infty-\beta_\infty\Omega_{21}+\beta_\infty\Omega_{22}\beta'_\infty\right).
\end{aligned}$$

■

Proof of Theorem 10. Parts (a) and (b). Instead of comparing the asymptotic variances of $R\sqrt{T}(\hat{\theta}_{1T}-\theta_0)$ and $R\sqrt{T}(\hat{\theta}_{2T}-\theta_0)$ directly, we equivalently compare the asymptotic variances of $(\tilde{R}\tilde{R}')^{-1/2}R\sqrt{T}(\hat{\theta}_{1T}-\theta_0)$ and $(\tilde{R}\tilde{R}')^{-1/2}R\sqrt{T}(\hat{\theta}_{2T}-\theta_0)$. We can do so because $(\tilde{R}\tilde{R}')^{-1/2}$ is nonsingular. Note that the latter two asymptotic variances are the same as those of the respective one-step estimator $\hat{\theta}^R_{1T}$ and two-step estimator $\hat{\theta}^R_{2T}$ of θ^R_0 in the following simple location model:

$$\left\{\begin{array}{l} y^R_{1t}=\theta^R_0+u^R_{1t}\in\mathbb{R}^p \\ y_{2t}=u_{2t}\in\mathbb{R}^q \end{array}\right. \tag{34}$$

where

$$\theta^R_0=(\tilde{R}\tilde{R}')^{-1/2}R\theta_0,\ u^R_{1t}=(\tilde{R}\tilde{R}')^{-1/2}\tilde{R}u_{1t},$$

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

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

$$var(u^R_t)=\left(\begin{array}{cc} I_p & O \\ O & I_q \end{array}\right)=I_{p+q}.$$

So the above location model has exactly the same form as the model in Section 3. We can invoke Proposition 3 to complete the proof.

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

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

So the long run canonical correlation coefficients between u^R_{1t} and u_{2t} are the eigenvalues of $\rho_R\rho'_R$, i.e., $\nu\left(\rho_R\rho'_R\right)$. Parts (a) and (b) then follow from Proposition 3.

Parts (c) and (d). The local asymptotic power of the one-step test and two-step test are the same as the local asymptotic power of respective one-step and two-step tests in the location model given in (34). We use Proposition 7 to complete the proof. For the above location model, the asymptotic variance of the infeasible two-step GMM estimator is

$$\Omega^R_{1\cdot2}=\left[(\tilde{R}\tilde{R}')^{-1/2}\tilde{R}\right]\Omega_{1\cdot2}\left[(\tilde{R}\tilde{R}')^{-1/2}\tilde{R}\right]'.$$

In addition, the local alternative parameter corresponding to $H_1: R\theta_0=r+\delta_0/\sqrt{T}$ for the location model is $(\tilde{R}\tilde{R}')^{-1/2}\delta_0/\sqrt{T}$. So the set of δ_0's considered in Proposition 7 is given by

$$\begin{aligned}\mathfrak{A}_{loc}\left(\lambda_0\right) &= \left\{\delta:\left[(\tilde{R}\tilde{R}')^{-1/2}\delta\right]'\left(\Omega^R_{1\cdot2}\right)^{-1}\left[(\tilde{R}\tilde{R}')^{-1/2}\delta\right]=\lambda_0\right\} \\ &= \left\{\delta:\delta'(\tilde{R}\Omega_{1\cdot2}\tilde{R}')^{-1}\delta=\lambda_0\right\}. \end{aligned}\tag{35}$$

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

Using (31) with $\breve{M}^{-1} = \breve{\Omega}^{-1}$, we have $M = \Omega$ and so

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

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

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

So the set of δ_0's considered in the theorem is exactly the same as that given in (35). ■

Proof of Theorem 11. The theorem is similar to Theorem 10. We only give the proof for part (d) in some details. It is easy to show that under the local alterative $H_1 : R\theta_0 = r + \delta_0/\sqrt{T}$, we have $\mathbb{W}_{aT} \overset{d}{\Longrightarrow} \mathbb{W}_{1\infty}(||\mathcal{V}_{a,R}^{-1/2}\delta_0||^2)$ where

$$\begin{aligned}
\mathcal{V}_{a,R} &= R(\breve{G}'\breve{W}^{-1}\breve{G})^{-1}\breve{G}'\breve{W}^{-1}\breve{\Omega}\breve{W}^{-1}\breve{G}(\breve{G}'\breve{W}^{-1}\breve{G})^{-1}R' \\
&= RVA^{-1}\left(\Sigma_{1\cdot 2}^{*}\right)^{1/2}\left(I_d, -\beta_a\right)\Omega\begin{pmatrix} I_d \\ -\beta_a' \end{pmatrix}\left[RVA^{-1}\left(\Sigma_{1\cdot 2}^{*}\right)^{1/2}\right]'. \qquad (36)
\end{aligned}$$

Similarly, we have

$$\mathbb{W}_{2T} \overset{d}{\Longrightarrow} \mathbb{W}_{2\infty}(||\mathcal{V}_{2,R}^{-1/2}\delta_0||^2),$$

where

$$\begin{aligned}
\mathcal{V}_{2,R} &= R(\breve{G}'\breve{\Omega}^{-1}\breve{G})^{-1}R' \\
&= RVA^{-1}\left(\Sigma_{1\cdot 2}^{*}\right)^{1/2}\left(I_d, -\beta_0\right)\Omega\begin{pmatrix} I_d \\ -\beta_0' \end{pmatrix}\left[RVA^{-1}\left(\Sigma_{1\cdot 2}^{*}\right)^{1/2}\right]',
\end{aligned}$$

which is the asymptotic variance of $R\sqrt{T}(\tilde{\theta}_{2T} - \theta_0)$ with $\tilde{\theta}_{2T}$ being the infeasible optimal two-step GMM estimator.

The difference in the two matrices $\mathcal{V}_{a,R}$ and $\mathcal{V}_{2,R}$ is

$$\mathcal{V}_{a,R} - \mathcal{V}_{2,R} = RVA^{-1}\left(\Sigma_{1\cdot 2}^{*}\right)^{1/2}\left(\beta_a - \beta_0\right)\Omega_{22}\left(\beta_a - \beta_0\right)'\left[RVA^{-1}\left(\Sigma_{1\cdot 2}^{*}\right)^{1/2}\right]'.$$

Now for any $\tau > 0$,

$$\begin{aligned}
& ||\mathcal{V}_{2,R}^{-1/2}\delta_0||^2 - \tau||\mathcal{V}_{a,R}^{-1/2}\delta_0||^2 = \delta_0'[\mathcal{V}_{2,R}^{-1} - \tau\mathcal{V}_{a,R}^{-1}]\delta_0 \\
= \; & \delta_0'\left[\mathcal{V}_{a,R}^{-1/2}\right]'\left\{[\mathcal{V}_{a,R}^{1/2}]'\mathcal{V}_{2,R}^{-1}\mathcal{V}_{a,R}^{1/2} - \tau I_p\right\}\mathcal{V}_{a,R}^{-1/2}\delta_0 \\
= \; & \delta_0'\left[\mathcal{V}_{a,R}^{-1/2}\right]'\left\{\left[\mathcal{V}_{a,R}^{-1/2}\mathcal{V}_{2,R}(\mathcal{V}_{a,R}^{-1/2})'\right]^{-1} - \tau I_p\right\}\mathcal{V}_{a,R}^{-1/2}\delta_0.
\end{aligned}$$

But

$$\mathcal{V}_{a,R}^{-1/2}\mathcal{V}_{2,R}[\mathcal{V}_{a,R}^{-1/2}]' = I_p - \mathcal{V}_{a,R}^{-1/2}\left(\mathcal{V}_{a,R} - \mathcal{V}_{2,R}\right)[\mathcal{V}_{a,R}^{-1/2}]',$$

and

$$\begin{aligned}
& \mathcal{V}_{a,R}^{-1/2}\left(\mathcal{V}_{a,R} - \mathcal{V}_{2,R}\right)[\mathcal{V}_{a,R}^{-1/2}]' \\
= \; & \mathcal{V}_{a,R}^{-1/2} RVA^{-1}\left(\Sigma_{1\cdot 2}^{*}\right)^{1/2}\left(\beta_a - \Omega_{12}\Omega_{22}^{-1}\right)\Omega_{22}\left(\beta_a - \Omega_{12}\Omega_{22}^{-1}\right)'\left[RVA^{-1}\left(\Sigma_{1\cdot 2}^{*}\right)^{1/2}\right]'[\mathcal{V}_{a,R}^{-1/2}]' \\
= \; & \rho_{a,R}\rho_{a,R}'.
\end{aligned}$$

So

$$||\mathcal{V}_{2,R}^{-1/2}\delta_0||^2 - \tau||\mathcal{V}_{a,R}^{-1/2}\delta_0||^2 = \delta_0'\left[\mathcal{V}_{a,R}^{-1/2}\right]'\left[\left(I_p - \rho_{a,R}\rho_{a,R}'\right)^{-1} - \tau I_p\right]\mathcal{V}_{a,R}^{-1/2}\delta_0 \cdot \tau$$

for any τ.

Let $\rho_{a,R}\rho_{a,R}' = \sum_{i=1}^{p}\nu_{i,a,R}b_{i,a,R}b_{i,a,R}'$ be the eigen decomposition of $\rho_{a,R}\rho_{a,R}'$, then

$$\begin{aligned}
& ||\mathcal{V}_{2,R}^{-1/2}\delta_0||^2 - \tau\left(\lambda_0\right)||\mathcal{V}_{a,R}^{-1/2}\delta_0||^2 \\
= \; & \sum_{i=1}^{p}\left[\frac{1}{1-\nu_{i,a,R}} - \tau\left(\lambda_0\right)\right]\left(b_{i,a,R}'\mathcal{V}_{a,R}^{-1/2}\delta_0\right)^2 \\
= \; & \tau\left(\lambda_0\right)\left\|\mathcal{V}_{a,R}^{-1/2}\delta_0\right\|^2\sum_{i=1}^{p}\frac{\nu_{i,a,R} - f\left(\lambda_0\right)}{1-\nu_{i,a,R}}\left\langle b_{i,a,R}, \frac{\mathcal{V}_{a,R}^{-1/2}\delta_0}{\left\|\mathcal{V}_{a,R}^{-1/2}\delta_0\right\|}\right\rangle^2,
\end{aligned}$$

which has the same form as the representation given in (29) . The rest of the proof is then identical to the proof of Proposition 7 and is omitted here. ■

References

[1] Andrews, D. W. (1991): "Heteroskedasticity and Autocorrelation Consistent Covariance Matrix Estimation." *Econometrica* 59(3), 817–858.

[2] Bester, C. A., Conley, T. G., and Hansen, C. B. (2011): "Inference with Dependent Data Using Cluster Covariance Estimators." *Journal of Econometrics* 165(2), 137–151.

[3] Bester, C. A., Conley, T. G., Hansen, C. B., and Vogelsang, T. J. (2014): "Fixed-b Asymptotics for Spatially Dependent Robust Nonparametric Covariance Matrix Estimators." Forthcoming, *Econometric Theory.*

[4] Goncalves, S. and Vogelsang, T. (2011): "Block Bootstrap HAC Robust Tests: The Sophistication of the Naive Bootstrap." *Econometric Theory* 27(4), 745–791.

[5] Goncalves, S. (2011). "The Moving Blocks Bootstrap for Panel Linear Regression Models with Individual Fixed Effects." *Econometric Theory* 27(5), 1048–1082.

[6] Gupta, S.D. and Perlman, M. D. (1974). On the Power of the Noncentral F-test: Effect of Additional Variates on Hotelling's T^2 test. *Journal of the American Statistical Association* 69, 174-180

[7] Hansen, L. P. (1982): "Large Sample Properties of Generalized Method of Moments Estimators." *Econometrica* 50, 1029–1054.

[8] Ibragimov, R. and Müller, U. K. (2010): "t-Statistic Based Correlation and Heterogeneity Robust Inference." *Journal of Business and Economic Statistics* 28, 453–468.

[9] Jansson, M. (2004): "The Error in Rejection Probability of Simple Autocorrelation Robust Tests." *Econometrica* 72(3), 937–946.

[10] Kiefer, N. M., Vogelsang, T. J. and Bunzel, H. (2002): "Simple Robust Testing of Regression Hypotheses" *Econometrica* 68 (3), 695–714.

[11] Kiefer, N. M. and Vogelsang, T. J. (2002a): "Heteroskedasticity-autocorrelation Robust Testing Using Bandwidth Equal to Sample Size." *Econometric Theory* 18, 1350–1366.

[12] ———— (2002b): "Heteroskedasticity-autocorrelation Robust Standard Errors Using the Bartlett Kernel without Truncation." *Econometrica* 70(5), 2093–2095.

[13] ———— (2005): "A New Asymptotic Theory for Heteroskedasticity-Autocorrelation Robust Tests." *Econometric Theory* 21, 1130–1164.

[14] Kim, M. S., and Sun, Y. (2013): "Heteroskedasticity and Spatiotemporal Dependence Robust Inference for Linear Panel Models with Fixed Effects." *Journal of Econometrics* 177(1), 85–108.

[15] Lehmann, E.L. and Romano, J. (2008): *Testing Statistical Hypotheses*, Springer: New York.

[16] Liang, K.-Y. and Zeger, S. (1986). Longitudinal Data Analysis Using Generalized Linear Models. *Biometrika* 73(1):13–22.

[17] Marden, J. and Perlman, M. D. (1980): "Invariant Tests for Means with Covariates." *Annals of Statistics* 8(1), 25–63.

[18] Müller, U. K. (2007): "A Theory of Robust Long-run Variance Estimation." *Journal of Econometrics* 141(2), 1331–1352.

[19] Muirhead, R. J. (2009): *Aspects of Multivariate Statistical Theory.* Vol. 197. Wiley.

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

[21] Phillips, P. C. B. (2005). "HAC Estimation by Automated Regression." *Econometric Theory* 21(1), 116–142.

[22] Shao, X. (2010): "A Self-normalized Approach to Confidence Interval Construction in Time Series." *Journal of the Royal Statistical Society* B, 72, 343–66.

[23] Sun, Y. (2011): "Robust Trend Inference with Series Variance Estimator and Testing-optimal Smoothing Parameter." *Journal of Econometrics* 164(2), 345–66.

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

[25] Sun, Y. (2014a): "Let's Fix It: Fixed-b Asymptotics versus Small-b Asymptotics in Heteroscedasticity and Autocorrelation Robust Inference." *Journal of Econometrics* 178(3), 659–677.

[26] Sun, Y. (2014b): "Fixed-smoothing Asymptotics in a Two-step GMM Framework." *Econometrica* 82(6), 2327–2370.

[27] Sun, Y. and Kim, M.S. (2015): "Asymptotic F Test in GMM Framework with Cross Sectional Dependence," *Review of Economics and Statistics* 97(1), 210–223.

[28] Sun, Y., Phillips, P. C. B. and Jin, S. (2008). "Optimal Bandwidth Selection in Heteroskedasticity-Autocorrelation Robust Testing." *Econometrica* 76(1), 175–94.

[29] Sun, Y. and Phillips, P. C. B. (2009): "Bandwidth Choice for Interval Estimation in GMM Regression." Working Paper, Department of Economics, UC San Diego.

[30] Vogelsang, T. J. (2012). "Heteroskedasticity, Autocorrelation, and Spatial Correlation Robust Inference in Linear Panel Models with Fixed-effects." *Journal of Econometrics* 166(2), 303–319.

[31] Zhang, X., and Shao, X. (2013). "Fixed-smoothing Asymptotics for Time Series." *Annals of Statistics* 41(3), 1329–1349.