UNIVERSITY OF CALIFORNIA, SAN DIEGO

Fixed Smoothing Asymptotic Theory in Over-identified Econometric Models in the Presence of Time-series and Clustered Dependence

A dissertation submitted in partial satisfaction of the requirements for the degree
Doctor of Philosophy

in

Economics

by

Jungbin Hwang

Committee in charge:

Professor Yixiao Sun, Chair
Professor Ian Abramson
Professor Graham Elliott
Professor Dimitris N. Politis
Professor Andres Santos

2016

The dissertation of Jungbin Hwang is approved, and it is acceptable in quality and form for publication on microfilm:

____

____

____

____

____

Chair

University of California, San Diego

2016

DEDICATION

To my parents and older brother.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGEMENTS

I do not have enough words to express my deepest gratitude to my parents and older brother whom I deeply love. They patiently bear with me even when I cannot stand myself. With your care, your support, today I have come to end of this road. I also feel special love to Harim Kim, who has been with me and supported my journey, despite her own challenges in seeking for a Ph.D.

I have another deepest gratitude to my advisor Professor Yixiao Sun, for his thoughtful guidance, care, and patience. I could not have completed my Ph.D. without his help, comments, supervision, and patience. I am also in debt to other committee members, Professor Andres Santos, Professor Graham Elliott, Professor Dimitris Politis, and Professor Ian Abramson for guiding my research for the past years.

Lastly, I am sincerely grateful to my friends in San Diego during these years, especially, Gonzalo Valdes, Henrique Romero, and people in Seoul National University Alumni Association (SNUAA), San Diego. They have helped me to develop not only my academic abilities but also social skills. I never felt homesick thanks to them and gratefully acknowledge the scholarship from SNUAA, San Diego.

Chapter 2 and 3, in full, are co-authored with Yixiao Sun and have been submitted for publication of material.

VITA

2008	B.S. in Economics , Seoul National University, Korea
2010	M.A in Economics, Seoul National University, Korea
2016	Ph.D. in Economics, University of California, San Diego

ABSTRACT OF THE DISSERTATION

Fixed Smoothing Asymptotic Theory in Over-identified Econometric Models in the Presence of Time-series and Clustered Dependence

by

Jungbin Hwang

Doctor of Philosophy in Economics

University of California, San Diego, 2016

Professor Yixiao Sun, Chair

In the widely used over-identified econometric model, the two-step Generalized Methods of Moments (GMM) estimator and inference, first suggested by (Hansen, 1982), require the estimation of optimal weighting matrix at the initial stages. For time series data and clustered dependent data, which is our focus here, the optimal weighting matrix is usually referred to as the long run variance (LRV) of the (scaled) sample moment conditions. To maintain generality and avoid misspecification, nowadays we do not model serial dependence and within-cluster dependence parametrically but use the heteroscedasticity and autocorrelation robust (HAR) variance estimator in standard practice. These estimators are nonparametric in nature with high variation in finite samples, but the conventional increasing smoothing asymptotics, so called small-bandwidth asymptotics, completely ignores the finite sample variation of the estimated GMM weighting matrix. As a consequence, empirical researchers are often in danger of making unreliable inferences and false assessments of the (efficient) two-step GMM methods. Motivated by

this issue, my dissertation consists of three papers which explore the efficiency and approximation issues in the two-step GMM methods by developing new, more accurate, and easy-to-use approximations to the GMM weighting matrix.

The first chapter, "Simple and Trustworthy Cluster-Robust GMM Inference″ explores new asymptotic theory for two-step GMM estimation and inference in the presence of clustered dependence. Clustering is a common phenomenon for many cross-sectional and panel data sets in applied economics, where individuals in the same cluster will be interdependent while those from different clusters are more likely to be independent. The core of new approximation scheme here is that we treat the number of clusters G fixed as the sample size increases. Under the new fixed-G asymptotics, the centered two-step GMM estimator and two continuously-updating estimators have the same asymptotic mixed normal distribution. Also, the t statistic, J statistic, as well as the trinity of two-step GMM statistics (QLR, LM and Wald) are all asymptotically pivotal, and each can be modified to have an asymptotic standard F distribution or t distribution. We also suggest a finite sample variance correction further to improve the accuracy of the F or t approximation. Our proposed asymptotic F and t tests are very appealing to practitioners, as test statistics are simple modifications of the usual test statistics, and the F or t critical values are readily available from standard statistical tables. We also apply our methods to an empirical study on the causal effect of access to domestic and international markets on household consumption in rural China.

The second paper "Should we go one step further? An Accurate Comparison of One-step and Two-step procedures in a Generalized Method of Moments Framework″ (coauthored with Yixiao Sun) focuses on GMM procedure in time-series setting and provides an accurate comparison of one-step and two-step GMM procedures in a fixed-smoothing asymptotics framework. The theory developed in this paper shows that the two-step procedure outperforms the one-step method only when the benefit of using the optimal weighting matrix outweighs the cost of estimating it. We also provide clear guidance on how to choose a more efficient (or powerful) GMM estimator (or test) in practice.

While our fixed smoothing asymptotic theory accurately describes sampling distribution of two-step GMM test statistic, the limiting distribution of conventional GMM statistics is non-standard, and its critical values need to be simulated or approximated by standard distributions in practice. In the last chapter, "Asymptotic F and t Tests in an Efficient GMM Setting" (coauthored with Yixiao Sun), we propose a simple and easy-to-implement modification to the trinity (QLM, LM, and Wald) of two-step GMM statistics and show that the modified test statistics are all asymptotically F distributed under the fixed-smoothing asymptotics. The modification is multiplicative and only involves the J statistic for testing over-identifying restrictions. In fact, what we propose can be regarded as the multiplicative variance correction for two-step GMM statistics that takes into account the additional asymptotic variance term under the fixed-smoothing asymptotics. The results in this paper can be immediately generalized to the GMM setting in the presence of clustered dependence.

Chapter 1

Simple and Trustworthy Cluster-Robust GMM Inference

Abstract. This paper develops a new asymptotic theory for two-step GMM estimation and inference in the presence of clustered dependence. While conventional asymptotic theory completely ignores the variability in the cluster-robust GMM weighting matrix, the new asymptotic theory takes it into account, leading to more accurate approximations. The key difference between these two types of asymptotics is whether the number of clusters G is regarded as fixed or growing when the sample size increases. Under the new fixed-G asymptotics, the centered two-step GMM estimator and the two continuously-updating estimators have the same asymptotic mixed normal distribution. In addition, the J-statistic, the trinity of two-step GMM statistics (QLR, LM and Wald), and the t-statistic are all asymptotically pivotal, and each can be modified to have an asymptotic standard F distribution or t distribution. We suggest a finite sample variance correction to further improve the accuracy of the F and t approximations. Our proposed asymptotic F and t tests are very appealing to practitioners because our test statistics are simple modifications of the usual test statistics, and the F and t critical values are readily available from standard statistical tables. A Monte Carlo study shows that our proposed tests are much more accurate than existing tests. We also ap-

ply our methods to an empirical study on the causal effect of access to domestic and international markets on household consumption in rural China. The results suggest that the effect of access to markets may be lower than the previous finding.

1.1 Introduction

Clustering is a common feature for many cross-sectional and panel data sets in applied economics. The data often come from a number of independent clusters with a general dependence structure within each cluster. For example, in development economics, data are often clustered by geographical regions, such as village, county and province, e.g., (De Brauw and Giles, 2012; Pepper, 2002; Dube et al., 2010). In empirical finance and industrial organization, firm level data are often clustered at the industry level ,e.g., Samila and Sorenson, 2011; Bharath et al., 2014, and in many educational studies, students' test scores are clustered at the classroom or school level (Andrabi et al., 2011). Because of learning from daily interactions, the presence of common shocks, and for many other reasons, individuals in the same cluster will be interdependent while those from different clusters tend to be independent. Failure to control for within group or cluster correlation often leads to downwardly biased standard errors and spurious statistical significance.

Seeking to robustify inference, many practical methods employ clustered covariance estimators (CCE). See White (1980), Liang and Zeger (1986), and Arellano and Bond (1991) for overviews of the CCE and its applications. It is now well known that standard test statistics based on the CCE are either asymptotically chi-squared or normal. The chi-squared and normal approximations are obtained under the so-called large-G asymptotic specification, which requires the number of clusters G to grow with the sample size. The key ingredient behind these approximations is that the CCE becomes concentrated at the true asymptotic variance as G diverges to infinity. In effect, this type of asymptotics ignores the estimation uncertainty in the CCE despite its high variation in finite samples, especially when the number of clusters is small. In practice, it is not unusual to have a data set that has a small number of clusters. For example, if clustering is based on large geographical regions such as U.S. states and regional blocks of neighboring countries, (e.g., Duflo et al., 2004; Obstfeld et al., 2008; Bester et al., 2011; Ibragimov and Müller; 2011), we cannot convincingly claim that the number of cluster is large so that the large-G asymptotic approximations are applicable. In fact, there is ample simulation evidence that the large-G approximation can be very poor when the number of clusters is not large (e.g., Donald and Lang, 2007; Cameron et al., 2008 ; Bester et al., 2011; MacKinnon and Webb, 2014).

In this paper, we introduce a new approach that yields more accurate approximations, and that works well whether or not the number of clusters is large. In fact, our approximations work especially well when the chi-squared and normal approximations are poor. They are obtained from a limiting thought experiment where the number of clusters G is held fixed. Under this fixed-G asymptotics, the CCE no longer asymptotically degenerates; instead, it converges in distribution to a random matrix that is proportional to the true asymptotic variance. The random limit of the CCE has profound implications for the analyses of the asymptotic properties of GMM estimators and the corresponding test statistics.

We start with the first-step GMM estimator where the underlying model is possibly over-identified and show that suitably modified Wald and t-statistics converge weakly to standard F and t distributions, respectively. The modification is easy to implement because it involves only a known multiplicative factor. Similar results have been obtained by Hansen (2007) and Bester et al. (2011) ,which employ a CCE type HAC estimator but consider only linear regressions and M-estimators for an exactly identified model.

We then consider the two-step GMM estimator that uses the CCE as a weighting matrix. Because the weighting matrix is random even in the limit, the two-step estimator is not asymptotically normal. The form of the limiting distribution depends on how the CCE is constructed. If the CCE is based on the uncentered moment process, we obtain the so-called uncentered two-step GMM estimator. We show that the asymptotic distribution of this two-step GMM estimator is highly nonstandard. As a result, the associated Wald statistic is not asymptotically pivotal. However, it is surprising that the J-statistic is still asymptotically pivotal. Furthermore, we show that the limiting distribution of the J-statistic can be represented as an increasing function of a standard F random variable. So critical values are readily available from standard statistical tables and software packages.

Next, we establish the asymptotic properties of the "centered" two-step GMM estimator[1] whose weighting matrix is constructed using recentered moment conditions. Invoking centering is not innocuous for an over-identified GMM model because the empirical moment conditions, in this case, are not equal to zero in general. Under the traditional large-G asymptotics, recentering does not matter in large samples because the empirical moment conditions are asymptotically zero and here are ignorable, even though they are not identically zero in finite sample. In contrast, under the fixed-G asymptotics, recentering plays two important roles: it removes the first order effect of the estimation error in the first-step estimator, and it ensures that the weighting matrix is asymptotically independent of the empirical moment conditions. With the recentered CCE as the weighting matrix, the two-step GMM estimator is asymptotically mixed normal. The mixed normality reflects the high variation of the feasible two-step GMM estimator as compared to the infeasible two-step GMM estimator, which is obtained under the assumption that the 'efficient' weighing matrix is known. The mixed-normality allows us to construct the Wald and t-statistics that are asymptotically nuisance parameter free.

We also consider two types of continuous updating (CU) estimators. The first type continuously updates the first order conditions (FOC) underlying the two-step GMM estimator. Given that FOC's can be regarded as the empirical version of generalized estimating equations (GEE), we call this type of CU estimator the CU-GEE estimator. The second type continuously updates the GMM criterion function, leading to the CU-GMM estimator, which was first suggested

[1] Our definition of the centered two-step GMM estimator is originated from the recentered (or demeaned) GMM weighting matrix, and it should not be confused with "centering" the estimator itself.

by Hansen et al. (1996). Both CU estimators are designed to improve the finite sample performance of two-step GMM estimators. Interestingly, we show that the continuous updating scheme has a built-in recentering feature. So in terms of the first order asymptotics, it does not matter whether the empirical moment conditions are recentered or not. We find that the centered two-step GMM estimator and the two CU estimators are all first-order asymptotically equivalent under the fixed-G asymptotics. This result provides a theoretical justification for using the recentered CCE in a two-step GMM framework.

To relate the fixed-G asymptotic pivotal distributions to standard distributions, we introduce simple modifications to the Wald and t statistics associated with the centered two-step GMM and CU estimators. We show that the modified Wald and t statistics are asymptotically F and t distributed, respectively. This result resembles the corresponding result that is based on the first-step GMM estimator. It is important to point out that the proposed modifications are indispensable for our asymptotic F and t theory. In the absence of the modifications, the Wald and t statistics converge in distribution to nonstandard distributions, and as a result, critical values have to be simulated. The modifications involve only the standard J-statistic, and it is very easy to implement because the modified test statistics are scaled versions of the original Wald test statistics with the scaling factor depending on the J-statistic. Significantly, the combination of the Wald statistic and the J-statistic enables us to develop the F approximation theory.

Finally, although recentering removes the first order effect of the first-step estimation error, the centered two-step GMM estimator still faces some extra estimation uncertainty in the first-step estimator. The main source of the problem is that we have to estimate the unobserved moment process based on the first-step estimator. To capture the higher order effect, we propose to retain one more term in our stochastic approximation that is asymptotically negligible. The expansion helps us develop a finite sample correction to the asymptotic variance estimator. Our correction resembles that of Windmeijer, (2005) , which considers variance correction for a two-step GMM estimator but only in the i.i.d. setting. We show that the finite sample variance correction does not change the limiting distributions of the test statistics, but they can help improve the finite sample performance of our tests.

Monte Carlo simulations clearly show that our new tests have a much more accurate size than existing tests via standard normal and chi-square critical values, especially when the number of clusters G is not large. An advantage of our procedure is that the test statistics do not entail much extra computational cost because the main ingredient for the modification is the usual J-statistic. There is also no need to simulate critical values because the F and t critical values can be readily obtained from standard statistical tables.

Our fixed-G asymptotics is related to fixed-smoothing asymptotics for a long run variance (LRV) estimation in a time series setting. The latter was initiated and developed in econometric literature by Kiefer et al. (2000), Kiefer and Vogelsang (2002b), Müller (2007), Sun et al. (2008), Sun (2014a, 2014b), and Politis, (2011) among others. Our new asymptotics is in the same spirit in that both lines of research attempt to capture the estimation uncertainty in covariance estimation. With regards to orthonormal series LRV estimation, a recent paper by Hwang and Sun (2015b) modifies the two-step GMM statistics using the J-statistic, and shows that the modified statistics are asymptotically F and t distributed. The F and t limit theory presented in this paper is similar, but our cluster-robust limiting distributions differ from those of our predecessors in terms of the multiplicative adjustment and the degrees of freedom. Moreover, we propose a finite sample variance correction to capture the uncertainty embodied in the estimated moment process adequately. To our knowledge, the finite sample variance correction provided in this paper has not been considered in the literature on the fixed-smoothing asymptotics.

There is also a growing literature that uses the fixed-G asymptotics to design more accurate cluster-robust inference. For instance, Ibragimov and Müller (2010,

2011) proposes a t-test for a scalar parameter that is robust to potentially heterogeneous clusters. Hansen (2007), Stock and Watson (2008), and Bester et al. (2011) propose a cluster-robust F or t tests under cluster-size homogeneity. Bell and McCaffrey (2002) and Imbens and Kolesar (2012) suggest an adjusted t-critical value employing data-determined degrees of freedom. Recently, Canay et al., (2014) establishes a theory of randomization tests and suggests an alternative cluster-robust test. For other approaches, see Carter et al. (2013) which proposes a measure of the effective number of clusters under the large-G asymptotics.; Cameron et al. (2008), MacKinnon and Webb (2014) which provide cluster bootstrap approaches with asymptotic refinement. All these studies, however, mainly focus on a simple location model or linear regressions that are special cases of exactly identified models.

The remainder of the paper is organized as follows. Section 1.2 presents the basic setting and establishes the approximation results for the first-step GMM estimator under the fixed-G asymptotics. Sections 1.3 and 1.4 establish the fixed-G asymptotics for two-step GMM estimators and the CU estimators, respectively. Section 1.5 is devoted to developing asymptotic F and t tests based on the centered two-step GMM estimator and the CU estimators. Section 1.6 proposes a finite sample variance correction. The next two sections apply our methods to the popular linear dynamic panel model and report a simulation evidence in the context of this model. Section 1.10 applies our methods to an empirical study on the causal effect of access to markets on household consumption in some rural Chinese areas. The last section concludes. Proofs are given in the appendix

1.2 Basic Setting and the First-step GMM Estimator

We want to estimate the $d \times 1$ vector of parameters $\theta \in \Theta$. The true parameter vector θ_0 is assumed to be an interior point of $\Theta \subseteq \mathbb{R}^d$, which is a compact parameter space. The moment condition

$$Ef(Y_i, \theta) = 0 \text{ holds if and only if } \theta = \theta_0, \tag{1.1}$$

where $f_i(\theta) = f(Y_i, \theta)$ is an $m \times 1$ vector of twice continuously differentiable functions. We assume that $q = m - d \geq 0$ and the rank of $\Gamma = E\left[\partial f(Y_i, \theta_0)/\partial\theta'\right]$ is d. So the model is possibly over-identified with the degree of over-identification q. The number of observations is N.

Define $g_N(\theta) = N^{-1}\sum_{i=1}^{N} f_i(\theta)$. Given the moment condition in (1.1), the initial "first-step" GMM estimator of θ_0 is given by

$$\hat{\theta}_1 = \arg\min_{\theta\in\Theta} g_N(\theta)' W_N^{-1} g_N(\theta),$$

where W_N is an $m \times m$ positive definite and a symmetric weighting matrix that does not depend on the unknown parameter θ_0 and $\text{plim}_{N\to\infty} W_N = W > 0$. In the context of instrumental variable (IV) regression, one popular choice for W_N is $Z'Z/N$ where Z is the data matrix of instruments.

Let

$$\hat{\Gamma}(\theta) = N^{-1}\sum_{i=1}^{N} \frac{\partial f_i(\theta)}{\partial\theta'}.$$

To establish the asymptotic properties of $\hat{\theta}_1$, we assume that for any $\sqrt{N}$ consistent estimator $\tilde{\theta}$, $\text{plim}_{N\to\infty}\hat{\Gamma}(\tilde{\theta}) = \Gamma$ and that Γ is of full column rank. Also, under some regularity conditions, we have the following Central Limit Theorem (CLT):

$$\begin{aligned} &\sqrt{N} g_N(\theta_0) \xrightarrow{d} N(0, \Omega) \text{ where} \\ &\Omega = \lim_{N\to\infty} \frac{1}{N} E\left(\sum_{i=1}^{N} f_i(\theta_0)\right)\left(\sum_{j=1}^{N} f_j(\theta_0)\right)'. \end{aligned} \tag{1.2}$$

Here Ω is analogous to the long run variance in a time series setting but the components of Ω are contributed by cross-sectional dependences over all locations. For easy reference, we follow Sun and Kim (2015) and call Ω the global variance. Primitive conditions for the above CLT in the presence of cross-sectional dependence are provided in Jenish and Prucha (2009, 2012). Under these conditions, we have

$$\sqrt{N}(\hat{\theta}_1 - \theta_0) \xrightarrow{d} N\left[0, (\Gamma' W^{-1}\Gamma)^{-1}\Gamma' W^{-1}\Omega W^{-1}\Gamma(\Gamma' W^{-1}\Gamma)^{-1}\right].$$

Since Γ and W can be accurately estimated by $\hat{\Gamma}(\hat{\theta}_1)$ and W_N, we need only estimate Ω to make reliable inference about θ_0. The main issue is how to properly account for cross-sectional dependence in the moment process $\{f_j(\theta_0)\}_{j=1}^N$. In this paper, we assume that the cross-sectional dependence has a cluster structure, which is not uncommon in many microeconomic applications. More specifically, our data consists of a number of independent clusters, each of which has an unknown dependence structure. Let G be the total number of clusters and L_g be the size of cluster g. For simplicity, we assume that every cluster has the common size L_g, i.e., $L = L_1 = L_2 = ... = L_G$. The identical cluster size assumption can be relaxed to the assumption that each cluster has the same size asymptotically, i.e., $\lim_{N\to\infty} L_g/(G^{-1}\sum_{i=1}^G L_i) = 1$ for every $g = 1, ..., G$. The following assumption formally characterizes the cluster dependence.

Assumption 1 *(i) The data $\{Y_j\}_{j=1}^N$ consists of G clusters. (ii) Observations are independent across clusters. (iii) The number of clusters G is fixed, and the size of each cluster L grows with the total sample size N.*

Assumption 1-i) implies that the set $\{f_i(\theta_0), i = 1, 2, ..., N\}$ can be partitioned into G nonoverlapping clusters $\cup_{g=1}^G \mathcal{G}_g$ where $\mathcal{G}_g = \{f_k^g(\theta_0) : k = 1, ..., L\}$. In the context of this clustered structure, Assumption 1-ii) implies that the within-cluster dependence for each cluster can be arbitrary but $Ef_k^g(\theta_0)f_l^h(\theta_0) = 0$ if $g \neq h$. That is, $f_k^g(\theta_0)$ and $f_l^h(\theta_0)$ are independent if they belong to different clusters. Independence across clusters in Assumption 1-ii) can be generalized to allow weak dependence among clusters by restricting the number of observations located on the boundaries between clusters. See Bester et al. (2011) for the detailed primitive conditions. Under Assumption 1-ii), we have

$$\Omega = \lim_{N\to\infty} \frac{1}{N} E\left(\sum_{i=1}^{N} f_i(\theta_0)\right)\left(\sum_{j=1}^{N} f_j(\theta_0)\right)'$$
$$= \lim_{N\to\infty} \frac{1}{N} \sum_{i=1}^{N}\sum_{j=1}^{N} 1(i,\ j\ \in \text{same cluster}) E f_i(\theta_0) f_j(\theta_0)'. \tag{1.3}$$

Assumption 1-iii) specifies the direction of asymptotics we consider. Under this fixed-G asymptotic specification, we have

$$\Omega = \frac{1}{G}\sum_{g=1}^{G} \lim_{L\to\infty} Var\left(\frac{1}{\sqrt{L}}\sum_{i=1}^{L} f_i^g(\theta_0)\right) := \frac{1}{G}\sum_{g=1}^{G}\Omega_g.$$

Thus, the global covariance matrix Ω can be represented as the simple average of Ω_g, $g = 1, ..., G$, where Ω_g's are the limiting variances within individual clusters. Motivated by this, we construct the clustered covariance estimator (CCE) as follows:

$$\hat{\Omega}(\hat{\theta}_1) = \frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{N} 1(i, j \in \text{the same group}) f_i(\hat{\theta}_1) f_j(\hat{\theta}_1)'$$
$$= \frac{1}{G}\sum_{g=1}^{G}\left\{\left(\frac{1}{\sqrt{L}}\sum_{i=1}^{L} f_i^g(\hat{\theta}_1)\right)\left(\frac{1}{\sqrt{L}}\sum_{j=1}^{L} f_j^g(\hat{\theta}_1)\right)'\right\}.$$

To ensure that $\hat{\Omega}(\hat{\theta}_1)$ is positive definite, we assume that $G \geq m$, and we maintain this condition throughout the rest of the paper.

Suppose we want to test the null hypothesis $H_0 : R\theta_0 = r$ against the alternative $H_1 : R\theta_0 \neq r$, where R is a $p \times d$ matrix. We focus on linear restrictions without loss of generality because the Delta method can be used to convert nonlinear restrictions into linear ones in an asymptotic sense. The F-test version of the Wald test statistic is given by

$$F(\hat{\theta}_1) := (R\hat{\theta}_1 - r)'\left\{R\widehat{var}(\hat{\theta}_1)R'\right\}^{-1}(R\hat{\theta}_1 - r)/p,$$

where

$$\widehat{var}(\hat{\theta}_1)$$
$$= \frac{1}{N}\left[\hat{\Gamma}(\hat{\theta}_1)'W_N^{-1}\hat{\Gamma}(\hat{\theta}_1)\right]^{-1}\left[\hat{\Gamma}(\hat{\theta}_1)'W_N^{-1}\hat{\Omega}(\hat{\theta}_1)W_N^{-1}\hat{\Gamma}(\hat{\theta}_1)\right]\left[\hat{\Gamma}(\hat{\theta}_1)'W_N^{-1}\hat{\Gamma}(\hat{\theta}_1)\right]^{-1}.$$

When $p = 1$ and the alternative is one sided, we can construct the t-statistic:

$$t(\hat{\theta}_1) := \frac{R\hat{\theta}_1 - r}{\sqrt{R\widehat{var}(\hat{\theta}_1)R'}}.$$

To formally characterize the asymptotic distributions of $F\backslash(\hat{\theta}_1)$ and $t(\hat{\theta}_1)$ under the fixed-G asymptotics, we further maintain the following high level conditions.

Assumption 2 $\hat{\theta}_1 \xrightarrow{p} \theta_0$.

Assumption 3 *(i) For each $g = 1, ..., G$, let*

$$\Gamma_g(\theta) := \lim_{L_N \to \infty} E\left[\frac{1}{L}\sum_{k=1}^{L}\frac{\partial f_k^g(\theta)}{\partial \theta'}\right].$$

Then,

$$\sup_{\theta \in \mathcal{N}(\theta_0)}\left\|\frac{1}{L}\sum_{k=1}^{L}\frac{\partial f_k^g(\theta)}{\partial \theta'} - \Gamma_g(\theta)\right\| \xrightarrow{p} 0,$$

holds, where $\mathcal{N}(\theta_0)$ is an open neighborhood of θ_0 and $\|\cdot\|$ is the Euclidean norm. (ii) $\Gamma_g(\theta)$ is continuous at $\theta = \theta_0$, and for $\Gamma_g = \Gamma_g(\theta_0)$, $\Gamma = G^{-1}\sum_{g=1}^{G}\Gamma_g$ has full rank.

Assumption 4 *Let $B_{m,g} \sim i.i.d. N(0, I_m)$ for $g = 1, ..., G$, then*

$$P\left(\frac{1}{\sqrt{L}}\sum_{k=1}^{L} f_k^g(\theta_0) \leq x\right) = P\left(\Lambda_g B_{m,g} \leq x\ \right) + o(1) \text{ as } L \to \infty.$$

for each $g = 1, ..., G$ where $x \in \mathbb{R}^m$ and Λ_g is the matrix square root of Ω_g.

Assumption 5 *(Homogeneity of Γ_g) For all $g = 1, ..., G$, $\Gamma_g = \Gamma$.*

Assumption 6 *(Homogeneity of Ω_g) For all $g = 1, ..., G$, $\Omega_g = \Omega$.*

Assumption 2 is made for convenience, and primitive sufficient conditions are available from the standard GMM asymptotic theory. Assumption 3 is a uniform law of large numbers (ULLN), from which we obtain $\hat{\Gamma}(\hat{\theta}_1) = G^{-1}\sum_{g=1}^{G}\Gamma_g + o_p(1) = \Gamma + o_p(1)$. Together with Assumption 1-(ii), Assumption 4 implies that $L^{-1/2}\sum_{j=1}^{L} f_j^g(\theta_0)$ follows a central limit theorem jointly over $g = 1, ..., G$ with zero asymptotic covariance between any two clusters. The homogeneity conditions in Assumptions 5 and 6 guarantee the asymptotic pivotality of the cluster-robust GMM statistics we consider. Similar assumptions are made in Bester et al. (2011) and Sun and Kim (2015), which develop asymptotically valid F tests that are robust to spatial autocorrelation in the same spirit as our fixed-G asymptotics. Let

$$\bar{B}_m := G^{-1}\sum_{g=1}^{G} B_{m,g} \text{ and } \bar{\mathbb{S}} := G^{-1}\sum_{g=1}^{G}\left(B_{m,g} - \bar{B}_m\right)\left(B_{m,g} - \bar{B}_m\right)'$$

where $B'_{m,g}$ as in Assumption 4. Also, let $\mathbb{W}_p(K, \Pi)$ denote a Wishart distribution with K degrees of freedom and $p \times p$ positive definite scale matrix Π. By construction, $\sqrt{G}\bar{B}_m \sim N(0, I_m)$, $\bar{\mathbb{S}} \sim G^{-1}\mathbb{W}_p(G-1, I_m)$ and $\bar{B}_m \perp \bar{\mathbb{S}}$. To present our asymptotic results, we partition $\bar{B}_m$ and $\bar{\mathbb{S}}$ as follows:

$$\bar{B}_m = \begin{pmatrix} \underset{d\times 1}{\bar{B}_d} \\ \underset{q\times 1}{\bar{B}_q} \end{pmatrix}, \bar{B}_d = \begin{pmatrix} \underset{p\times 1}{\bar{B}_p} \\ \underset{(d-p)\times 1}{\bar{B}_{d-p}} \end{pmatrix}, \bar{\mathbb{S}} = \begin{pmatrix} \underset{d\times d}{\bar{\mathbb{S}}_{dd}} & \underset{d\times q}{\bar{\mathbb{S}}_{dq}} \\ \underset{q\times d}{\bar{\mathbb{S}}_{qd}} & \underset{q\times q}{\bar{\mathbb{S}}_{qq}} \end{pmatrix},$$

$$\bar{\mathbb{S}}_{dd} = \begin{pmatrix} \underset{p\times p}{\bar{\mathbb{S}}_{pp}} & \underset{p\times(d-p)}{\bar{\mathbb{S}}_{p,d-p}} \\ \underset{(d-p)\times p}{\bar{\mathbb{S}}_{d-p,p}} & \underset{(d-p)\times(d-p)}{\bar{\mathbb{S}}_{d-p,d-p}} \end{pmatrix}, \text{ and } \bar{\mathbb{S}}_{dq} = \begin{pmatrix} \underset{p\times q}{\bar{\mathbb{S}}_{pq}} \\ \underset{(d-p)\times q}{\bar{\mathbb{S}}_{d-p,q}} \end{pmatrix}.$$

Proposition 1 *Let Assumptions 1∼6 hold. Then*

(a) $F(\hat{\theta}_1) \xrightarrow{d} \mathbb{F}_{1\infty} := G\bar{B}_p'\bar{\mathbb{S}}_{pp}^{-1}\bar{B}_p/p$;

(b) $t(\hat{\theta}_1) \xrightarrow{d} \mathbb{T}_{1\infty} := \frac{N(0,1)}{\sqrt{\chi^2_{G-1}/G}}$ *where* $N(0,1) \perp \sqrt{\chi^2_{G-1}}$.

Remark 2 *The limiting distribution $\mathbb{F}_{1\infty}$ follows Hotelling's T^2distribution. Us-*

ing the well-known relationship between the T^2 and standard F distributions, we obtain $\mathbb{F}_{1\infty} \overset{d}{=} (G/G-p)\mathcal{F}_{p,G-p}$ where $\mathcal{F}_{p,G-p}$ is a random variable that follows the F distribution with degree of freedom $(p, G-p)$. Similarly, $\mathbb{T}_{1\infty} \overset{d}{=} (G/G-1)t_{G-1}$ where t_{G-1} is a random variable that follows the t distribution with degree of freedom $G-1$.

Remark 3 *As an example of the general GMM setting, consider the linear regression model $y_j = x_j'\theta + \epsilon_j$. Under the assumption that $cov(x_j, \epsilon_j) = 0$, the moment function is $f_j(\theta_0) = x_j(y_j - x_j'\theta)$. With the moment condition $Ef_j(\theta_0) = 0$, the model is exactly identified. This setting was employed in Hansen, 2007; Stock and Watson, 2008; Bester et al., 2011, indeed, our F and t approximations in Proposition 1 are identical to what is obtained in these papers.*

Remark 4 *Under the large-G asymptotics where $G \to \infty$ but L is fixed, one can show that the CCE $\hat{\Omega}(\hat{\theta}_1)$ converges in probability to Ω for*

$$\Omega = \lim_{G\to\infty} \frac{1}{G} \sum_{g=1}^{G} Var\left(\frac{1}{\sqrt{L}} \sum_{k=1}^{L} f_k^g(\theta_0) \right).$$

The convergence of $\hat{\Omega}(\hat{\theta}_1)$ to Ω does not require the homogeneity of Ω_g in Assumption 6 (Hansen, 2007; Carter et al., 2013). Under this type of asymptotics, the test statistics $F(\hat{\theta}_1)$ and $t(\hat{\theta}_1)$ are asymptotically χ_p^2/p and $N(0,1)$. Let $\mathcal{F}_{p,G-p}^{1-\alpha}$ and $\chi_p^{1-\alpha}$ be the $1-\alpha$ quantiles of $F_{p,G-p}$ and the χ_p^2 distributions, respectively. As $G/(G-p) > 1$ and $\mathcal{F}_{p,G-p}^{1-\alpha} > \chi_p^{1-\alpha}/p$, it is easy to see that

$$\frac{G}{G-p}\mathcal{F}_{p,G-p}^{1-\alpha} > \chi_p^{1-\alpha}/p.$$

However, the difference between the two critical values $G(G-p)^{-1}\mathcal{F}_{p,G-p}^{1-\alpha}$ and $\chi_p^{1-\alpha}/p$ shrinks to zero as G increases. Therefore, the fixed-G critical value $G(G-p)^{-1}\mathcal{F}_{p,G-p}^{1-\alpha}$ is asymptotically valid under the large-G asymptotics. The asymptotic validity holds even if the homogeneity conditions of Assumptions 5 and 6 are not *satisfied. The fixed-G critical value is robust in the sense that it works whether G is small or large.*

Remark 5 *Let* Λ *the matrix square root of* Ω*,i.e.* $\Lambda\Lambda' = \Omega$*. Then, it follows from the proof of Proposition 1 that* $\hat{\Omega}(\hat{\theta}_1)$ *converges in distribution to a random matrix* $\Omega_{1\infty}$ *given by*

$$\Omega_{1\infty} = \Lambda\tilde{\mathbb{D}}\Lambda' \text{ where } \mathbb{D} = \frac{1}{G}\sum_{g=1}^{G}\tilde{D}_g\tilde{D}_g'$$
$$\tilde{D}_g = B_{m,g} - \Gamma_\Lambda(\Gamma_\Lambda' W_\Lambda^{-1}\Gamma_\Lambda)^{-1}\Gamma_\Lambda' W_\Lambda^{-1}\bar{B}_m \qquad (1.4)$$

for $\Gamma_\Lambda = \Lambda^{-1}\Gamma$ *and* $W_\Lambda = \Lambda^{-1}W(\Lambda')^{-1}$*.* $\tilde{D}_g$ *is a quasi-demeaned version of* $B_{m,g}$ *with quasi-demeaning attributable to the estimation error in* $\hat{\theta}_1$*. Note that the quasi-demeaning factor* $\Gamma_\Lambda(\Gamma_\Lambda' W_\Lambda^{-1}\Gamma_\Lambda)^{-1}\Gamma_\Lambda' W_\Lambda^{-1}$ *depends on all of* Γ, Ω *and* W*, and cannot be further simplified in general. The estimation error in* $\hat{\theta}_1$ *affects* $\Omega_{1\infty}$ *in a complicated way. However, for the first-step Wald and t statistics, we do not care about* $\hat{\Omega}(\hat{\theta}_1)$ per se. *Instead, we care about the scaled covariance matrix* $\hat{\Gamma}(\hat{\theta}_1)'W_N^{-1}\hat{\Omega}(\hat{\theta}_1)W_N^{-1}\hat{\Gamma}(\hat{\theta}_1)$*, which converges in distribution to* $\Gamma'W^{-1}\Omega_{1\infty}W^{-1}\Gamma$*. But*

$$\Gamma_\Lambda' W_\Lambda^{-1}\tilde{D}_g = \Gamma_\Lambda' W_\Lambda^{-1}\left(B_{m,g} - \bar{B}_m\right),$$

and thus

$$\Gamma'W^{-1}\Omega_{1\infty}W^{-1}\Gamma = \Gamma_\Lambda' W_\Lambda^{-1}\tilde{\mathbb{D}}W_\Lambda^{-1}\Gamma_\Lambda = \frac{1}{G}\sum_{g=1}^{G}\Gamma_\Lambda' W_\Lambda^{-1}\tilde{D}_g\left(\Gamma_\Lambda' W_\Lambda^{-1}\tilde{D}_g\right)'$$
$$\stackrel{d}{=} \Gamma_\Lambda' W_\Lambda^{-1}\frac{1}{G}\sum_{g=1}^{G}\left(B_{m,g} - \bar{B}_m\right)\left(B_{m,g} - \bar{B}_m\right)'\left(\Gamma_\Lambda' W_\Lambda^{-1}\right)'.$$

So, to the first order fixed-G asymptotics, the estimation error in $\hat{\theta}_1$ *affects* $\Gamma'W^{-1}$ $\Omega_{1\infty}W^{-1}\Gamma$ *via simple demeaning only. This is a key result that drives the asymptotic pivotality of* $F(\hat{\theta}_1)$ *and* $t(\hat{\theta}_1)$*.*

1.3 Two-step GMM Estimation and Inference

In an overidentified GMM framework, we often employ a two-step procedure to improve the efficiency of the initial GMM estimator and the power of the associated tests. It is now well-known that the optimal weighting matrix is the (inverted) asymptotic variance of the sample moment conditions. There are two different ways to estimate the asymptotic variance, and these lead to two different estimators $\hat{\Omega}(\hat{\theta}_1)$ and $\hat{\Omega}^c(\hat{\theta}_1)$ where

$$\hat{\Omega}(\theta) = \frac{1}{G}\sum_{g=1}^{G}\left(\frac{1}{\sqrt{L}}\sum_{k=1}^{L} f_k^g(\theta)\right)\left(\frac{1}{\sqrt{L}}\sum_{l=1}^{L} f_l^g(\theta)\right)',$$
$$\hat{\Omega}^c(\theta) = \frac{1}{G}\sum_{g=1}^{G}\left\{\frac{1}{\sqrt{L}}\sum_{k=1}^{L}\left[f_k^g(\theta) - g_N(\theta)\right]\right\}\left\{\frac{1}{\sqrt{L}}\sum_{l=1}^{L}\left[f_l^g(\theta) - g_N(\theta)\right]\right\}'.$$

While $\hat{\Omega}(\hat{\theta}_1)$ employs the uncentered moment process $\{f_i^g(\hat{\theta}_1)\}_{i=1}^N$, $\hat{\Omega}^c(\hat{\theta}_1)$ employs the recentered moment process $\{f_i^g(\hat{\theta}_1) - g_N(\hat{\theta}_1)\}_{i=1}^G$. For inference based on the first-step estimator $\hat{\theta}_1$, it does not matter which asymptotic variance estimator is used. This is so because for any asymptotic variance estimator $\hat{\Omega}(\hat{\theta}_1)$, the Wald statistic depends on $\hat{\Omega}(\hat{\theta}_1)$ only via $\hat{\Gamma}(\hat{\theta}_1)'W_N^{-1}\hat{\Omega}(\hat{\theta}_1)W_N^{-1}\hat{\Gamma}(\hat{\theta}_1)$. It is easy to show that the following asymptotic equivalence:

$$\begin{aligned}
&\hat{\Gamma}(\hat{\theta}_1)'W_N^{-1}\hat{\Omega}(\hat{\theta}_1)W_N^{-1}\hat{\Gamma}(\hat{\theta}_1) \\
&= \hat{\Gamma}(\hat{\theta}_1)'W_N^{-1}\hat{\Omega}^c(\hat{\theta}_1)W_N^{-1}\hat{\Gamma}(\hat{\theta}_1) + o_p(1) \\
&= \Gamma'W^{-1}\hat{\Omega}^c(\theta_0)W^{-1}\Gamma + o_p(1).
\end{aligned}$$

Thus, the limiting distribution of the Wald statistic is the same whether the estimated moment process is recentered or not. It is important to point out that the asymptotic equivalence holds because two asymptotic variance estimators are premultiplied by $\hat{\Gamma}(\hat{\theta}_1)'W_N^{-1}$ and post-multiplied by $W_N^{-1}\hat{\Gamma}(\hat{\theta}_1)$. The two asymptotic variance estimators are not asymptotically equivalent by themselves under fixed-G

asymptotics.

Depending on whether we use $\hat{\Omega}(\hat{\theta}_1)$ or $\hat{\Omega}^c(\hat{\theta}_1)$, we have different two-step GMM estimators:

$$\hat{\theta}_2 = \arg\min_{\theta\in\Theta} g_N(\theta)' \left[\hat{\Omega}(\hat{\theta}_1)\right]^{-1} g_N(\theta),$$
$$\hat{\theta}_2^c = \arg\min_{\theta\in\Theta} g_N(\theta)' \left[\hat{\Omega}^c(\hat{\theta}_1)\right]^{-1} g_N(\theta).$$

Given that $\hat{\Omega}(\hat{\theta}_1)$ and $\hat{\Omega}^c(\hat{\theta}_1)$ are not asymptotically equivalent and that they enter the definitions of $\hat{\theta}_2$ and $\hat{\theta}_2^c$ by themselves, the two estimators have different asymptotic behaviors, as shown in the next two subsections.

1.3.1 Uncentered Two-step GMM Estimator

In this subsection, we consider the two-step GMM estimator $\hat{\theta}_2$ based on the uncentered moment process. We establish the asymptotic properties of $\hat{\theta}_2$ and the associated Wald statistic and J-statistic. We show that the J-statistic is asymptotically pivotal, even though the Wald statistic is not.

It follows from standard asymptotic arguments that

$$\sqrt{N}(\hat{\theta}_2 - \theta_0) = -\left[\Gamma'\hat{\Omega}^{-1}(\hat{\theta}_1)\Gamma\right]^{-1} \Gamma'\hat{\Omega}^{-1}(\hat{\theta}_1)\frac{1}{\sqrt{G}}\sum_{g=1}^{G}\left(\frac{1}{\sqrt{L}}\sum_{k=1}^{L} f_k^g(\theta_0)\right) + o_p(1). \tag{1.5}$$

Using the joint convergence of the following

$$\hat{\Omega}(\hat{\theta}_1) \xrightarrow{d} \Omega_{1\infty} = \Lambda\tilde{\mathbb{D}}\Lambda' \text{ and } \frac{1}{\sqrt{G}}\sum_{g=1}^{G}\left(\frac{1}{\sqrt{L}}\sum_{k=1}^{L} f_k^g(\theta_0)\right) \xrightarrow{d} \sqrt{G}\Lambda\bar{B}_m, \tag{1.6}$$

we obtain:

$$\sqrt{N}(\hat{\theta}_2 - \theta_0) \xrightarrow{d} -\left[\Gamma_\Lambda'\left(\tilde{\mathbb{D}}\right)^{-1}\Gamma_\Lambda\right]^{-1}\Gamma_\Lambda'\left(\tilde{\mathbb{D}}\right)^{-1}\sqrt{G}\bar{B}_m$$

where as before

$$\tilde{\mathbb{D}} = \frac{1}{G}\sum_{g=1}^{G}\tilde{D}_g\tilde{D}_g' \text{ for } \tilde{D}_g = B_{m,g} - \Gamma_\Lambda(\Gamma_\Lambda' W_\Lambda^{-1}\Gamma_\Lambda)^{-1}\Gamma_\Lambda' W_\Lambda^{-1}\bar{B}_m.$$

Since $\tilde{\mathbb{D}}$ is random, the limiting distribution is not normal. Even though both $\tilde{D}_g$ and $\bar{B}_m$ are normal, there is a nonzero correlation between them. As a result, $\tilde{\mathbb{D}}$ and $\bar{B}_m$ are correlated, too. This makes the limiting distribution of $\sqrt{N}(\hat{\theta}_2 - \theta_0)$ highly nonstandard.

To understand the limiting distribution, we define the infeasible estimator $\tilde{\theta}_2$ by assuming that $\hat{\Omega}(\theta_0)$ is known, which leads to

$$\tilde{\theta}_2 = \arg\min_{\theta\in\Theta} g_N(\theta)'\hat{\Omega}^{-1}(\theta_0)g_N(\theta).$$

Now

$$\sqrt{N}(\tilde{\theta}_2 - \theta_0) \xrightarrow{d} -\left[\Gamma_\Lambda'\mathbb{S}^{-1}\Gamma_\Lambda\right]^{-1}\Gamma_\Lambda'\mathbb{S}^{-1}\sqrt{G}\bar{B}_m$$

where $\mathbb{S} = G^{-1}\sum_{g=1}^{G} B_{m,g}B_{m,g}'$. The only difference between the asymptotic distributions of $\sqrt{N}(\hat{\theta}_2 - \theta_0)$ and $\sqrt{N}(\tilde{\theta}_2 - \theta_0)$ is the quasi-demeaning embedded in the definition of $\tilde{D}_g$. This difference captures the first order effect of having to estimate the optimal weighting matrix, which is needed to construct the feasible two-step estimator $\hat{\theta}_2$.

To make further links between the limiting distributions, let's partition $\mathbb{S}$ in the same way that $\bar{\mathbb{S}}$ is partitioned. Also, define U to be the $m \times m$ matrix of the eigen vectors of $\Gamma_\Lambda'\Gamma_\Lambda = \Gamma'\Omega^{-1}\Gamma$ and $U\Sigma V'$ be a singular value decomposition (SVD) of Γ_Λ. By construction, $U'U = UU' = I_m$, $V'V = V'V = I_d$, and $\Sigma' = \left[\begin{array}{cc} A_{d\times d} & O_{d\times q}\end{array}\right]$. We then define $\tilde{W} = U'W_\Lambda U$ and partition $\tilde{W}$ as before. We also introduce

$$\beta_{\mathbb{S}} = \mathbb{S}_{dq}\mathbb{S}_{qq}^{-1},\ \beta_{\tilde{W}} = \tilde{W}_{dq}\tilde{W}_{qq}^{-1} \text{ and } \kappa_G = G\cdot\bar{B}_q'\mathbb{S}_{qq}^{-1}\bar{B}_q.$$

By construction, $\beta_{\mathbb{S}}$ is the "random" regression coefficient induced by $\mathbb{S}$ while $\beta_{\tilde{W}}$

is the regression coefficient induced by the constant matrix $\tilde{W}$. Also, κ_G is the quadratic form of normal random vector $\sqrt{G}\,\bar{B}_q$ with random matrix $\mathbb{S}_{qq}$. Finally, on the basis of $\hat{\theta}_2$, the J-statistic for testing over-identification restrictions is

$$J(\hat{\theta}_2) := N g_N(\hat{\theta}_2)' \left(\hat{\Omega}(\hat{\theta}_1)\right)^{-1} g_N(\hat{\theta}_2)/q. \tag{1.7}$$

The following proposition characterizes and connects the limiting distributions of the three estimators: the first-step estimator $\hat{\theta}_1$, the feasible two-step estimator $\hat{\theta}_2$, and the infeasible two-step estimator $\tilde{\theta}_2$.

Proposition 6 *Let Assumptions 1∼6 hold. Then*

(a) $\sqrt{N}(\hat{\theta}_1 - \theta_0) \xrightarrow{d} -VA^{-1}\sqrt{G}(\bar{B}_d - \beta_{\tilde{W}}\bar{B}_q)$;

(b) $\sqrt{N}(\tilde{\theta}_2 - \theta_0) \xrightarrow{d} -VA^{-1}\sqrt{G}(\bar{B}_d - \beta_{\mathbb{S}}\bar{B}_q)$;

(c) $\sqrt{N}(\hat{\theta}_2 - \theta_0) \xrightarrow{d} -VA^{-1}\sqrt{G}(\bar{B}_d - \beta_{\mathbb{S}}\bar{B}_q) - VA^{-1}\sqrt{G}(\bar{B}_d - \beta_{\tilde{W}}\bar{B}_q)\cdot(\kappa_G/G)$;

(d) $\sqrt{N}(\hat{\theta}_2 - \theta_0) = \sqrt{N}(\tilde{\theta}_2 - \theta_0) + \sqrt{N}(\hat{\theta}_1 - \theta_0)\cdot(\kappa_G/G) + o_p(1)$;

(e) $J(\hat{\theta}_2) \xrightarrow{d} \kappa_G$ *where* (a), (b), (c), *and* (e) *hold jointly.*

Part (d) of the proposition shows that $\sqrt{N}(\hat{\theta}_2 - \theta_0)$ is asymptotically equivalent to a linear combination of the infeasible two-step estimator $\sqrt{N}(\tilde{\theta}_2 - \theta_0)$ and the first-step estimator $\sqrt{N}(\hat{\theta}_1 - \theta_0)$. This contrasts with the conventional GMM asymptotics, wherein feasible and infeasible estimators are asymptotically equivalent.

It is interesting to see that the linear coefficient in Parts (c) and (d) is proportional to the limit of the J-statistic. Given $\kappa_G = O_p(1)$ as G increases, the limiting distribution of $\sqrt{N}(\hat{\theta}_2 - \theta_0)$ becomes closer to that of $\sqrt{N}(\tilde{\theta}_2 - \theta_0)$. In the special case where $q = 0$, i.e., when the model is exactly identified, $\kappa_G = 0$ and $\sqrt{N}(\hat{\theta}_2 - \theta_0)$ and $\sqrt{N}(\tilde{\theta}_2 - \theta_0)$ have the same limiting distribution. This is expected given that the weighting matrix is irrelevant in the exactly identified GMM model.

Using the Sherman–Morrison formula[2], it is straightforward to show

$$\kappa_G \overset{d}{=} \left(\frac{G}{q}\right)\frac{\frac{q}{G-q}\mathcal{F}_{q,G-q}}{1+\frac{q}{G-q}\mathcal{F}_{q,G-q}}.$$

It is perhaps surprising that while the asymptotic distributions of $\hat{\theta}_2$ is complicated and nonstandard, the limiting distribution of the J-statistic is not only pivotal but is also an increasing function of the standard F distribution. For the J test at the significance level α, say 5%, the critical value from κ_G can be obtained from

$$\left(\frac{G}{q}\right)\frac{\frac{q}{G-q}\mathcal{F}^{1-\alpha}_{q,G-q}}{1+\frac{q}{G-q}\mathcal{F}^{1-\alpha}_{q,G-q}}.$$

Equivalently, we have

$$\frac{G-q}{q}\frac{q\kappa_G}{G-q\kappa_G} \overset{d}{=} \mathcal{F}_{q,G-q},$$

and so

$$\tilde{J}(\hat{\theta}_2) := \frac{G-q}{q}\frac{qJ(\hat{\theta}_2)}{G-qJ(\hat{\theta}_2)} \overset{d}{\to} \mathcal{F}_{q,G-q}.$$

That is, the transformed J-statistic $\tilde{J}(\hat{\theta}_2)$ is asymptotically F distributed. This is very convenient in empirical applications.

It is important to point out that the convenient F limit of $\tilde{J}(\hat{\theta}_2)$ holds only if the J-statistic is equal to the GMM criterion function evaluated at the two-step GMM estimator $\hat{\theta}_2$. This effectively imposes a constraint on the weighting matrix. If we use a weighting matrix that is different from $\hat{\Omega}(\hat{\theta}_1)$, then the resulting J-statistic may not be asymptotically pivotal any longer.

Define the F-statistic and variance estimate for the two-step estimator $\hat{\theta}_2$

[2] $(C+ab')^{-1} = C^{-1} - \frac{C^{-1}ab'C^{-1}}{1+b'C^{-1}a}$ for any invertable square matrix C and conforming column vectors such that $1+b'C^{-1}a \neq 0$.

as

$$F_{\hat{\Omega}(\hat{\theta}_1)}(\hat{\theta}_2) = (R\hat{\theta}_2 - r)' \left(R\widehat{var}_{\hat{\Omega}(\hat{\theta}_1)}(\hat{\theta}_2)R' \right)^{-1} (R\hat{\theta}_2 - r)/p \text{ for}$$
$$\widehat{var}_{\hat{\Omega}(\hat{\theta}_1)}(\hat{\theta}_2) = \frac{1}{N} \left(\hat{\Gamma}(\hat{\theta}_2)'\hat{\Omega}^{-1}(\hat{\theta}_1)\hat{\Gamma}(\hat{\theta}_2) \right)^{-1}.$$

In the above definitions, we use a subscript notation $\hat{\Omega}(\hat{\theta}_1)$ to clarify the choice of CCE in $F_{\cdot}(\hat{\theta}_2)$ and $\widehat{var}_{\cdot}(\hat{\theta}_2)$. Is the above F-statistic asymptotically pivotal as the J-statistic $J(\hat{\theta}_2)$? Unfortunately, the answer is no, as implied by the following proposition which uses the additional notation:

$$\mathbb{E}_{p+q,p+q} := \begin{pmatrix} \mathbb{E}_{pp} & \mathbb{E}_{pq} \\ \mathbb{E}'_{pq} & \mathbb{E}_{qq} \end{pmatrix} = \begin{pmatrix} \mathbb{S}_{pp} & \mathbb{S}_{pq} \\ \mathbb{S}'_{pq} & \mathbb{S}_{qq} \end{pmatrix} + \begin{pmatrix} \tilde{\beta}^p_{\tilde{W}} \bar{B}_q \bar{B}'_q (\tilde{\beta}^p_{\tilde{W}})' & \tilde{\beta}^p_{\tilde{W}} \bar{B}_q \bar{B}'_q \\ \bar{B}_q \bar{B}'_q (\tilde{\beta}^p_{\tilde{W}})' & \bar{B}_q \bar{B}'_q \end{pmatrix}$$

where $\tilde{\beta}^p_{\tilde{W}}$ is the $p \times q$ matrix and consists of the first p rows of $\tilde{V}'\beta_{\tilde{W}}$ where $\tilde{V}$ is the $d \times d$ matrix of the eigen vector of $(RVA^{-1})' RVA^{-1}$.

Proposition 7 *Let Assumptions 1∼6 hold. Then*

$$F_{\hat{\Omega}(\hat{\theta}_1)}(\hat{\theta}_2) \xrightarrow{d} \frac{G}{p} (\bar{B}_p - \mathbb{E}_{pq}\mathbb{E}^{-1}_{qq}\bar{B}_q)' (\mathbb{E}_{pp\cdot q})^{-1} (\bar{B}_p - \mathbb{E}_{pq}\mathbb{E}^{-1}_{qq}\bar{B}_q)$$
$$= \frac{1}{p} \left[G \begin{pmatrix} \bar{B}_p \\ \bar{B}_q \end{pmatrix}' \begin{pmatrix} \mathbb{E}_{pp} & \mathbb{E}_{pq} \\ \mathbb{E}'_{pq} & \mathbb{E}_{qq} \end{pmatrix}^{-1} \begin{pmatrix} \bar{B}_p \\ \bar{B}_q \end{pmatrix} - G\bar{B}'_q \mathbb{E}^{-1}_{qq} \bar{B}_q \right], \tag{1.8}$$

where

$$\mathbb{E}_{pp\cdot q} = \mathbb{E}_{pp} - \mathbb{E}_{pq}\mathbb{E}^{-1}_{qq}\mathbb{E}'_{pq}.$$

Due to the presence of the second term in $\mathbb{E}_{p+q,p+q}$, which depends on $\tilde{\beta}_{\tilde{W}}$, the F-statistic is not asymptotically pivotal. It depends on several nuisance parameters including Ω. To see this, we note that the second term in (1.8) is the same as $(G/q) \cdot \bar{B}'_q \mathbb{S}^{-1}_{qq} \bar{B}_q = \kappa_G$. So the second term is the limit of the J-statistic, which is nuisance parameter free. However, the first term in (1.8) is not pivotal because we have

$$G\begin{pmatrix}\bar{B}_p\\ \bar{B}_q\end{pmatrix}'\begin{pmatrix}\mathbb{E}_{pp} & \mathbb{E}_{pq}\\ \mathbb{E}'_{pq} & \mathbb{E}_{qq}\end{pmatrix}^{-1}\begin{pmatrix}\bar{B}_p\\ \bar{B}_q\end{pmatrix}$$
$$= G\left[\begin{pmatrix}\bar{B}_p\\ \bar{B}_q\end{pmatrix}'\begin{pmatrix}\bar{\mathbb{S}}_{pp} & \bar{\mathbb{S}}_{pq}\\ \bar{\mathbb{S}}'_{pq} & \bar{\mathbb{S}}_{qq}\end{pmatrix}^{-1}\begin{pmatrix}\bar{B}_p\\ \bar{B}_q\end{pmatrix} - \frac{\left(\bar{B}'_{p+q}\bar{\mathbb{S}}^{-1}_{p+q,p+q}\tilde{w}\bar{B}_q\right)^2}{1+\bar{B}'_q\tilde{w}'_{p+q}\bar{\mathbb{S}}^{-1}_{p+q}\tilde{w}\bar{B}_q}\right]$$

where $\tilde{w} = \left((\tilde{\beta}^p_W)', I_q\right)'$. Here, as in the case of the J-statistic, the first term in the above equation is nuisance parameter free. But the second term is clearly a nonconstant function of $\tilde{\beta}^p_W$, which, in turn, depends on R, Γ, W and Ω.

1.3.2 Centered Two-step GMM estimator

Given that the estimation error in $\hat{\theta}_1$ affects the limiting distribution of $\hat{\Omega}(\hat{\theta}_1)$, the Wald statistic based on the uncentered two-step GMM estimator $\hat{\theta}_2$ is not asymptotically pivotal. In view of (1.4), the effect of the estimator error is manifested via a location shift in $\tilde{D}_g$; the shifting amount depends on $\hat{\theta}_1$. A key observation is that the location shift is the same for all groups under the homogeneity Assumptions 5 and 6. So if we demean the empirical moment process, we can remove the location shift that is caused by the estimator error in $\hat{\theta}_1$. This leads to the recentered asymptotic variance estimator and a pivotal inference for both the Wald test and J test.

It is important to note that recentering is not innocuous for an over-identified GMM model because $N^{-1}\sum_{i=1}^{N} f_i(\hat{\theta}_1)$ is not zero in general. In the time series HAR variance estimation, recentering is known to have several advantages. For example, as Hall (2000) observes, in conventional increasing smoothing asymptotic theory, recentering can potentially improve the power of the J-test using a HAR variance estimator when the model is misspecified. Building on this intuition, Lee (2014) recently proposes a nonparametric misspecification robust GMM bootstrap employing the recentered GMM weight matrix. Also, shows that, under the fixed smoothing asymptotics, recentering is necessary to yield an asymptotically pivotal inference from the two-step Wald test statistic.

In our fixed-G asymptotic framework, recentering plays an important role in the CCE estimation. It ensures that the limiting distribution of $\hat{\Omega}^c(\hat{\theta}_1)$ is invariant to the initial estimator $\hat{\theta}_1$. The following lemma proves a more general result and characterizes the fixed-G limiting distribution of the centered CCE matrix for any $\sqrt{N}$ consistent estimator $\tilde{\theta}$.

Lemma 8 *Let Assumptions 1∼6 hold. Let $\tilde{\theta}$ be any $\sqrt{N}$ consistent estimator of θ_0. Then*

(a) $\hat{\Omega}^c(\tilde{\theta}) = \hat{\Omega}^c(\theta_0) + o_p(1)$;

(b) $\hat{\Omega}^c(\theta_0) \xrightarrow{d} \Omega^c_\infty$ *where* $\Omega^c_\infty = \Lambda \bar{\mathbb{S}} \Lambda'$.

Lemma 8 indicates that the centered CCE $\Omega^c(\hat{\theta}_1)$ converges in distribution to the random matrix limit $\Omega^c_\infty = \Lambda \bar{\mathbb{S}} \Lambda'$, which follows a (scaled) Wishart distribution $G^{-1}\mathbb{W}_m(G-1, \Omega)$. Using Lemma 8, it is possible to show

$$\sqrt{N}(\hat{\theta}^c_2 - \theta_0) \xrightarrow{d} -\left[\Gamma' \left(\Omega^c_\infty\right)^{-1} \Gamma\right]^{-1} \Gamma' \left(\Omega^c_\infty\right)^{-1} \Lambda \sqrt{G} \bar{B}_m. \tag{1.9}$$

Since $(\Omega^c_\infty)^{-1}$ is independent with $\sqrt{G}\Lambda \bar{B}_m \sim N(0, \Omega)$, the limiting distribution of $\hat{\theta}^c_2$ is mixed normal.

On the basis of $\hat{\theta}^c_2$, we can construct the "trinity" of GMM test statistics. The first one is the normalized Wald statistic defined by

$$F_{\hat{\Omega}^c(\hat{\theta}_1)}(\hat{\theta}^c_2) := (R\hat{\theta}^c_2 - r)' \{R\widehat{var}_{\hat{\Omega}^c(\hat{\theta}^c_2)}(\hat{\theta}^c_2) R'\}^{-1} (R\hat{\theta}^c_2 - r)/p \text{ where} \tag{1.10}$$

$$\widehat{var}_{\hat{\Omega}^c(\hat{\theta}_1)}(\hat{\theta}^c_2) = \frac{1}{N}\left(\hat{\Gamma}(\hat{\theta}^c_2)' \left(\hat{\Omega}^c(\hat{\theta}^c_2)\right)^{-1} \hat{\Gamma}(\hat{\theta}^c_2)\right)^{-1}.$$

When $p = 1$ and the alternative is one sided, we can construct the t-statistic below:

$$t_{\hat{\Omega}^c(\hat{\theta}_1)}(\hat{\theta}^c_2) := \frac{\left(R\hat{\theta}^c_2 - r\right)}{\{R\widehat{var}_{\hat{\Omega}^c(\hat{\theta}_1)}(\hat{\theta}^c_2) R'\}^{1/2}}.$$

The second test statistic is the Quasi-Likelihood Ratio (QLR) type of statistic. Define the restricted and centered two-step estimator $\hat{\theta}_2^{c,r}$:

$$\hat{\theta}_2^{c,r} = \arg\min_{\theta\in\Theta} g_N(\theta)' \left[\hat{\Omega}^c(\hat{\theta}_1)\right]^{-1} g_N(\theta) \text{ s.t. } R\theta = r.$$

The QLR statistic is given by

$$LR_{\hat{\Omega}^c(\hat{\theta}_1)}(\hat{\theta}_2^c, \hat{\theta}_2^{c,r}) := N\left\{g_N(\hat{\theta}_2^c)' \left[\hat{\Omega}^c(\hat{\theta}_1)\right]^{-1} g_N(\hat{\theta}_2^c) - g_N(\hat{\theta}_2^{c,r})' \left[\hat{\Omega}^c(\hat{\theta}_1)\right]^{-1} g_N(\hat{\theta}_2^{c,r})\right\}/p.$$

The third test statistic is the Lagrange Multiplier (LM) or score statistic in the GMM setting. Let $\Delta_{\hat{\Omega}^c(\cdot)}(\theta)$ be the gradient of the GMM criterion function $\hat{\Gamma}(\theta)'$ $\left[\hat{\Omega}^c(\cdot)\right]^{-1} g_N(\theta)$, then the GMM score test statistic is given by

$$LM_{\hat{\Omega}^c(\hat{\theta}_1)}(\hat{\theta}_2^{c,r}) := N\left[\Delta_{\hat{\Omega}^c(\hat{\theta}_1)}(\hat{\theta}_2^{c,r})\right]' \left\{\hat{\Gamma}(\hat{\theta}_2^{c,r})' \left[\hat{\Omega}^c(\hat{\theta}_1)\right]^{-1} \hat{\Gamma}(\hat{\theta}_2^{c,r})\right\}^{-1} \left[\Delta_{\hat{\Omega}^c(\hat{\theta}_1)}(\hat{\theta}_2^{c,r})\right]/p.$$

In the definition of all three types of the GMM test statistics, we plug the first-step estimator $\hat{\theta}_1$ into $\hat{\Omega}^c(\cdot)$, but Lemma 8 indicates that replacing $\hat{\theta}_1$ with any $\sqrt{N}$ consistent estimator (e.g., $\hat{\theta}_2$ and $\hat{\theta}_2^c$) does not affect the fixed-G asymptotic results. This contrasts with the fixed-G asymptotics for the uncentered two-step estimator $\hat{\theta}_2$. Lastly, we also construct the standard J- statistic based on $\hat{\theta}_2^c$:

$$J(\hat{\theta}_2^c) := Ng_N(\hat{\theta}_2^c)' \left(\hat{\Omega}^c(\hat{\theta}_1)\right)^{-1} g_N(\hat{\theta}_2^c)/q,$$

where $\hat{\Omega}^c(\hat{\theta}_1)$ can be replaced by $\hat{\Omega}^c(\hat{\theta}_2^c)$ without affecting the limiting distribution of the J statistic.

Using (1.9) and Lemma 8, we have $F_{\hat{\Omega}^c(\hat{\theta}_1)}(\hat{\theta}_2^c) \xrightarrow{d} \mathbb{F}_{2\infty}$ where

$$\begin{aligned}\mathbb{F}_{2\infty} = G\left[R\left(\Gamma_\Lambda'\bar{\mathbb{S}}^{-1}\Gamma_\Lambda\right)^{-1}\Gamma_\Lambda'\bar{\mathbb{S}}^{-1}\bar{B}_m\right]' \left[R\left(\Gamma_\Lambda'\bar{\mathbb{S}}^{-1}\Gamma_\Lambda\right)^{-1}R'\right]^{-1} \\ \times\left[R\left(\Gamma_\Lambda'\bar{\mathbb{S}}^{-1}\Gamma_\Lambda\right)^{-1}\Gamma_\Lambda'\bar{\mathbb{S}}^{-1}\bar{B}_m\right]/p.\end{aligned} \quad (1.11)$$

When $p = 1$, we get $t_{\hat{\Omega}^c(\hat{\theta}_1)}(\hat{\theta}_2^c) \xrightarrow{d} \mathbb{T}_{2\infty}$ with

$$\mathbb{T}_{2\infty} = \frac{R\left(\Gamma_\Lambda'\bar{\mathbb{S}}^{-1}\Gamma_\Lambda\right)^{-1}\Gamma_\Lambda'\bar{\mathbb{S}}^{-1}\sqrt{G}\bar{B}_m}{\sqrt{R\left(\Gamma_\Lambda'\bar{\mathbb{S}}^{-1}\Gamma_\Lambda\right)^{-1}R'}}.$$

Also, it follows in a similar way that

$$\begin{aligned} J(\hat{\theta}_2^c) \xrightarrow{d} \mathbb{J}_\infty := G\left\{\bar{B}_m - \Gamma_\Lambda\left(\Gamma_\Lambda'\bar{\mathbb{S}}^{-1}\Gamma_\Lambda\right)^{-1}\Gamma_\Lambda'\bar{\mathbb{S}}^{-1}\bar{B}_m\right\}'\bar{\mathbb{S}}^{-1} \\ \times\left\{\bar{B}_m - \Gamma_\Lambda\left(\Gamma_\Lambda'\bar{\mathbb{S}}^{-1}\Gamma_\Lambda\right)^{-1}\Gamma_\Lambda'\bar{\mathbb{S}}^{-1}\bar{B}_m\right\}/q. \end{aligned} \tag{1.12}$$

The remaining question is whether the above representations for $\mathbb{F}_{2\infty}$ and $\mathbb{J}_\infty$ are free of nuisance parameters. The following proposition provides a positive answer.

Proposition 9 *Let Assumptions 1∼6 hold and define* $\bar{\mathbb{S}}_{pp\cdot q} = \bar{\mathbb{S}}_{pp} - \bar{\mathbb{S}}_{pq}\bar{S}_{qq}^{-1}\bar{\mathbb{S}}_{qp}$.

(*a*) $F_{\hat{\Omega}^c(\hat{\theta}_1)}(\hat{\theta}_2^c) \xrightarrow{d} G\left(\bar{B}_p - \bar{\mathbb{S}}_{pq}\bar{\mathbb{S}}_{qq}^{-1}\bar{B}_q\right)'\bar{\mathbb{S}}_{pp\cdot q}^{-1}\left(\bar{B}_p - \bar{\mathbb{S}}_{pq}\bar{\mathbb{S}}_{qq}^{-1}\bar{B}_q\right)'/p \overset{d}{=} \mathbb{F}_{2\infty}$;

(*b*) $t_{\hat{\Omega}^c(\hat{\theta}_1)}(\hat{\theta}_2^c) \xrightarrow{d} \sqrt{G}\left(\bar{B}_p - \bar{\mathbb{S}}_{pq}\bar{\mathbb{S}}_{qq}^{-1}\bar{B}_q\right)/\sqrt{\bar{\mathbb{S}}_{pp\cdot q}} \overset{d}{=} \mathbb{T}_{2\infty}$ *for* $p = 1$;

(*c*) $LR_{\hat{\Omega}^c(\hat{\theta}_1)}(\hat{\theta}_2^c, \hat{\theta}_2^{c,r}) = F_{\hat{\Omega}^c(\hat{\theta}_1)}(\hat{\theta}_2^c) + o_p(1)$;

(*d*) $LM_{\hat{\Omega}^c(\hat{\theta}_1)}(\hat{\theta}_2^{c,r}) = F_{\hat{\Omega}^c(\hat{\theta}_1)}(\hat{\theta}_2^c) + o_p(1)$;

(*e*) $J(\hat{\theta}_2^c) \xrightarrow{d} (G/q)\bar{B}_q'\bar{\mathbb{S}}_{qq}^{-1}\bar{B}_q \overset{d}{=} \mathbb{J}_\infty$.

To simplify the representations of $\mathbb{F}_{2\infty}$ and $\mathbb{T}_{2\infty}$ in the above proposition, we note that

$$G\begin{bmatrix} \bar{\mathbb{S}}_{pp} & \bar{\mathbb{S}}_{pq} \\ \bar{\mathbb{S}}_{qp} & \bar{\mathbb{S}}_{qq} \end{bmatrix} \overset{d}{=} \sum_{g=1}^{G}\left(B_{p+q,g} - \bar{B}_{p+q}\right)\left(B_{p+q,g} - \bar{B}_{p+q}\right)',$$

where $B_{p+q,g} := (B_{p,g}', B_{p,g}')'$. The above random matrix has a standard Wishart distribution $\mathbb{W}_{p+q}(G-1, I_{p+q})$. It follows from the well-known properties of a Wishart distribution that $\bar{\mathbb{S}}_{pp\cdot q} \sim \mathbb{W}_p(G-1-q, I_p)/G$ and $\bar{\mathbb{S}}_{pp\cdot q}$ is independent of $\bar{\mathbb{S}}_{pq}$ and $\bar{\mathbb{S}}_{qq}$.[3] Therefore, if we condition on $\Delta := \bar{\mathbb{S}}_{pq}\bar{\mathbb{S}}_{qq}^{-1}\sqrt{G}\bar{B}_q$, the limiting distribution $\mathbb{F}_{2\infty}$ satisfies

$$\frac{G-p-q}{G}\mathbb{F}_{2\infty} \overset{d}{=} \frac{G-p-q}{G}\frac{(\sqrt{G}\bar{B}_p+\Delta)'\bar{\mathbb{S}}_{pp\cdot q}(\sqrt{G}\bar{B}_p+\Delta)}{p} \overset{d}{=} F_{p,G-p-q}\left(\|\Delta\|^2\right), \tag{1.13}$$

where $F_{p,G-p-q}(\|\Delta\|^2)$ is a noncentral F distribution with random noncentrality parameter $\|\Delta\|^2$. Similarly, the limiting distribution $\mathbb{T}_{2\infty}$ can be represented as

$$\sqrt{\frac{G-1-q}{G}}\mathbb{T}_{2\infty} \overset{d}{=} \sqrt{\frac{G-1-q}{G}}\frac{\sqrt{G}\bar{B}_p+\Delta}{\sqrt{\bar{\mathbb{S}}_{pp\cdot q}}} \overset{d}{=} t_{G-1-q}(\Delta), \tag{1.14}$$

which is a noncentral t distribution with a noncentrality parameter Δ. The nonstandard limiting distributions are similar to those in Sun (2014) which provides the fixed-smoothing asymptotic result in the case of the series LRV estimation. However, in our setting of clustered dependence, the scale adjustment and degrees of freedom parameter in (1.13) and (1.14) are different from those in Sun (2014).

The critical values from the nonstandard limiting distribution $\mathbb{F}_{2\infty}$ can be obtained through simulation, but Sun (2014b) shows that $\mathbb{F}_{2\infty}$ can be approximated by a noncentral F distribution. With regard to the QLR and LM types of test statistics, Proposition 9-(c) and (d) shows that they are asymptotically equivalent to $F_{\hat{\Omega}^c(\hat{\theta}_1)}(\hat{\theta}_2^c)$. This also implies that all three types of test statistics share the same fixed-G limit as given in (1.13) and (1.14). Similar results are obtained by Sun (2014b) and Hwang and Sun (2015a; 2015b), which focus on two-step GMM estimation and HAR inference in a time series setting.

For the J-statistic $J(\hat{\theta}_2^c)$, it follows from Proposition 9-(e) that

$$\frac{G-q}{G}J(\hat{\theta}_2^c) \overset{d}{\to} \mathbb{J}_\infty \overset{d}{=} \frac{G-q}{G}\bar{B}_q'\bar{\mathbb{S}}_{qq}^{-1}\bar{B}_q \overset{d}{=} F_{q,G-q}.$$

This is consistent with Kim and Sun's (2012) results except that our adjustment and degrees of freedom parameter are different.

[3] See Proposition 7.9 in Bilodeau and Brenner.

1.4 Iterative Two-step and Continuous Updating Schemes

Another class of popular GMM estimators is the continuous updating (CU) estimators, which are designed to improve the poor finite sample performance of two-step GMM estimators. For more discussion on the CU estimators, see Hansen et al. (1996).

Here we consider two types of continuous updating schemes. The first is the iterative scheme that iterates the second steps in the two-step GMM estimation until convergence. The j-th iterated GMM estimator $\hat{\theta}^j_{IE}$ is defined as the solution of the following minimization problem:

$$\hat{\theta}^j_{IE} = \underset{\theta\in\Theta}{\arg\min} g_N(\theta)'\hat{\Omega}^{-1}(\hat{\theta}^{j-1}_{IE})g_N(\theta) \text{ for } j \geq 1,$$

where $\hat{\theta}^0_{IE} = \hat{\theta}_2$ is the two-step estimator $\hat{\theta}_2$. The FOC for $\hat{\theta}^j_{IE}$ is

$$\hat{\Gamma}(\hat{\theta}^j_{IE})'\hat{\Omega}^{-1}(\hat{\theta}^{j-1}_{IE})g_N(\hat{\theta}^j_{IE}) = 0 \text{ for } j \geq 1.$$

In view of the above FOC, $\hat{\theta}^j_{IE}$ can be regarded as a generalized-estimating-equations (GEE) estimator, which is a class of estimators first studied by Liang and Zeger (1986).

When the number of iterations j goes to infinity until $\hat{\theta}^j_{IE}$ converges, we obtain the continuously updated generalized estimating equations (CU-GEE) estimator $\hat{\theta}^{\text{cu}}_{\text{GEE}}$. The FOC for $\hat{\theta}^{\text{cu}}_{\text{GEE}}$ is given by

$$\hat{\Gamma}(\hat{\theta}^{\text{cu}}_{\text{GEE}})'\hat{\Omega}^{-1}(\hat{\theta}_{\text{CU-GEE}})g_N(\hat{\theta}^{\text{cu}}_{\text{GEE}}) = 0. \tag{1.15}$$

In the above definition of $\hat{\theta}^{\text{cu}}_{\text{GEE}}$, we employ the uncentered CCE, $\hat{\Omega}(\cdot)$. However, it is not difficult to show that

$$\begin{aligned}&\hat{\Gamma}(\hat{\theta}^{\text{cu}}_{\text{GEE}})'\hat{\Omega}^{-1}(\hat{\theta}_{\text{CU-GEE}})g_N(\hat{\theta}^{\text{cu}}_{\text{GEE}})\\&=\hat{\Gamma}(\hat{\theta}^{\text{cu}}_{\text{GEE}})'\left(\hat{\Omega}^c(\hat{\theta}^{\text{cu}}_{\text{GEE}})\right)^{-1}g_N(\hat{\theta}_{\text{CU-GEE}})\cdot\frac{1}{1+\nu_N(\hat{\theta}^{\text{cu}}_{\text{GEE}})}\end{aligned}$$

where

$$\nu_N(\hat{\theta}^{\text{cu}}_{\text{GEE}})=L\cdot g_N(\hat{\theta}_{\text{CU-GEE}})'\left(\hat{\Omega}^c(\hat{\theta}_{\text{CU-GEE}})\right)^{-1}g_N(\hat{\theta}^{\text{cu}}_{\text{GEE}}).$$

Since $1/[1+\nu_N(\hat{\theta}^{\text{cu}}_{\text{GEE}})]$ is always positive, the first-order condition in (1.15) holds if and only if

$$\hat{\Gamma}(\hat{\theta}^{\text{cu}}_{\text{GEE}})'\left[\hat{\Omega}^c(\hat{\theta}^{\text{cu}}_{\text{GEE}})\right]^{-1}g_N(\hat{\theta}_{\text{CU-GEE}})=0. \tag{1.16}$$

So recentering has no effect on the CU-GEE estimator.

The second CU scheme continuously updates the GMM criterion function, which leads to the familiar continuous updating GMM (CU-GMM) estimator:

$$\hat{\theta}^{cu}_{\text{GMM}}=\arg\min_{\theta\in\Theta} g_N(\theta)'\hat{\Omega}^{-1}(\theta)g_N(\theta).$$

Although we use the uncentered CEE $\hat{\Omega}(\theta)$ in the above definition, the original definition of $\hat{\theta}^{cu}_{\text{GMM}}$ in Hansen et al. (1996) is based on the centered CCE weighting matrix $\hat{\Omega}^c(\theta)$. It is easy to show that

$$\begin{aligned}Lg_N(\theta)'\hat{\Omega}^{-1}(\theta)g_N(\theta)&=Lg_N(\theta)'\hat{\Omega}^{-1}(\theta)\left[\hat{\Omega}(\theta)-Lg_N(\theta)g_N(\theta)'\right]\left[\hat{\Omega}^c(\theta)\right]^{-1}g_N(\theta)\\&=Lg_N(\theta)'\left(\hat{\Omega}^c(\theta)\right)^{-1}g_N(\theta)\left\{1-Lg_N(\theta)'\hat{\Omega}^{-1}(\theta)g_N(\theta)\right\}.\end{aligned}$$

So we have

$$Lg_N(\theta)'\left(\hat{\Omega}^c(\theta)\right)^{-1}g_N(\theta)=\frac{Lg_N(\theta)'\hat{\Omega}^{-1}(\theta)g_N(\theta)}{1-Lg_N(\theta)'\hat{\Omega}^{-1}(\theta)g_N(\theta)}.$$

The above equation reveals the fact that the CU-GMM estimator will not change if the uncentered weighting matrix $\hat{\Omega}(\theta)$ is replaced by the centered one $\hat{\Omega}^c(\theta)$, i.e.,

$$\hat{\theta}^{cu}_{\text{GMM}} = \arg\min_{\theta\in\Theta} g_N(\theta)' \left(\hat{\Omega}^c(\theta)\right)^{-1}(\theta) g_N(\theta). \quad (1.17)$$

Similar to the centered two-step GMM estimator, the two CU estimators can be regarded as having a built-in recentering mechanism. For this reason, the limiting distributions of the two CU estimators are the same as that of the centered two-step GMM estimator, as is shown below.

Proposition 10 *Let Assumptions 1, 3∼6 hold. Assume that $\hat{\theta}^{cu}_{GEE}$ and $\hat{\theta}_{CU\text{-}GMM}$ are $\sqrt{N}$ consistent. Then*

$$\sqrt{N}(\hat{\theta}^{cu}_{GEE} - \theta_0) \xrightarrow{d} -\left[\Gamma' \left(\Omega^c_\infty\right)^{-1} \Gamma\right]^{-1} \Gamma' \left(\Omega^c_\infty\right)^{-1} \Lambda\sqrt{G}\bar{B}_m,$$
$$\sqrt{N}(\hat{\theta}^{cu}_{GMM} - \theta_0) \xrightarrow{d} -\left[\Gamma' \left(\Omega^c_\infty\right)^{-1} \Gamma\right]^{-1} \Gamma' \left(\Omega^c_\infty\right)^{-1} \Lambda\sqrt{G}\bar{B}_m.$$

The proposition shows that the CU estimators and the centered two-step GMM estimator are asymptotically equivalent under the fixed-G asymptotics.

We can construct the Wald statistics based on the two CU estimators as follows:

$$F_{\hat{\Omega}^c(\hat{\theta}^{\text{cu}}_{\text{GEE}})}(\hat{\theta}^{\text{cu}}_{\text{GEE}}) = (R\hat{\theta}^{\text{cu}}_{\text{GEE}} - r)'\{R\widehat{var}_{\hat{\Omega}^c(\hat{\theta}^{\text{cu}}_{\text{GEE}})}(\hat{\theta}_{\text{CU-GEE}})R'\}^{-1}(R\hat{\theta}^{\text{cu}}_{\text{GEE}} - r)/p$$
$$F_{\hat{\Omega}^c(\hat{\theta}^{cu}_{\text{GMM}})}(\hat{\theta}^{cu}_{\text{GMM}}) = (R\hat{\theta}^{cu}_{\text{GMM}} - r)'\{R\widehat{var}_{\hat{\Omega}^c(\hat{\theta}^{cu}_{\text{GMM}})}(\hat{\theta}^{cu}_{\text{GMM}})R'\}^{-1}(R\hat{\theta}^{cu}_{\text{GMM}} - r)/p$$

We construct $t_{\hat{\Omega}^c(\hat{\theta}^{\text{cu}}_{\text{GEE}})}(\hat{\theta}_{\text{CU-GEE}})$ and $t_{\hat{\Omega}^c(\hat{\theta}^{\text{cu}}_{\text{GMM}})}(\hat{\theta}^{cu}_{\text{GMM}})$ in a similar way when $p = 1$. It follows from Proposition 10 that the Wald statistics based on $\hat{\theta}_{\text{CU-GEE}}$ and $\hat{\theta}^{cu}_{\text{GMM}}$ are asymptotically equivalent to $F_{\hat{\Omega}^c(\hat{\theta}_\cdot)}(\hat{\theta}^c_2)$. As a result,

$$F_{\hat{\Omega}^c(\hat{\theta}^{\text{cu}}_{\text{GEE}})}(\hat{\theta}^{\text{cu}}_{\text{GEE}}) \xrightarrow{d} \mathbb{F}_{2\infty} \text{ and } F_{\hat{\Omega}^c(\hat{\theta}^{cu}_{\text{GMM}})}(\hat{\theta}^{cu}_{\text{GMM}}) \xrightarrow{d} \mathbb{F}_{2\infty}.$$

Similarly,

$$t_{\hat{\Omega}^c(\hat{\theta}^{\text{cu}}_{\text{GEE}})}(\hat{\theta}^{\text{cu}}_{\text{GEE}}) \xrightarrow{d} \mathbb{T}_{2\infty} \text{ and } t_{\hat{\Omega}^c(\hat{\theta}^{cu}_{\text{GMM}}}(\hat{\theta}^{cu}_{\text{GMM}}) \xrightarrow{d} \mathbb{T}_{2\infty}.$$

In summary, we have shown that all three estimators $\hat{\theta}_2^c$, $\hat{\theta}_{\text{GEE}}^{\text{cu}}$ and $\hat{\theta}_{\text{GMM}}^{cu}$, and the corresponding Wald test statistics converge in distribution to the same nonstandard distributions. Proposition 9-(c) and (d) continues to hold for the CU estimators, leading to the asymptotic equivalence of the three test statistics based on the CU estimators.

The findings in this subsection are quite interesting. Under the first order large-G asymptotics, the CU estimators and the default (uncentered) two-step GMM are all asymptotically equivalent. In other words, the first-order large-G asymptotics is not informative about the merits of the CU estimators. One may develop a high order expansion under the large-G asymptotics to reveal the advantages of CU estimators. In fact, Newey and Smith (2004) develops the stochastic expansion of CU estimators in the i.i.d setting and shows that the CU schemes automatically remove the high order estimation error of two-step estimator which is caused by the non-optimal weighting matrix in the first-step estimator. See also Anatolyev (2005) which extends the work of Newey and Smith (2004) to a time series setting. We could adopt these approaches, instead of the fixed-G asymptotics, to capture the estimation uncertainty of the first-step estimator in the default (uncentered) two-step GMM procedures. But the high order asymptotic analysis is technically very challenging and often requires strong assumptions on the smoothness of moment process. Although the fixed-G asymptotics we develop here is just a first order theory, it is powerful enough to reveal the asymptotic difference between the CU and the plain uncentered two-step GMM estimators. Moreover, the built-in recentering function behind the CU estimators provides some justification for the use of the centered CCE in a two-step GMM framework.

1.5 Asymptotic F and t tests

Under the fixed-G asymptotics, the limiting distributions of two-step test statistics, including Wald, QLR and LM, and the t statistics, are nonstandard and hence critical values have to be simulated in practice. This contrasts with the conventional large-G asymptotics, where the limiting distributions are the standard chi-square and normal distributions. In this section, we show that a simple modification of the two-step Wald and t statistics enables us to develop the standard F and t asymptotic theory under the fixed-G asymptotics. The asymptotic F and t tests are more appealing in empirical applications because the standard F and t distributions are more accessible than the nonstandard $\mathbb{F}_{2\infty}$ and $\mathbb{T}_{2\infty}$ distributions.

The modified two-step Wald, QLR and LM statistics are

$$\tilde{F}_{\hat{\Omega}^c(\hat{\theta}_2^c)}(\hat{\theta}_2^c) := \frac{G-p-q}{G} \cdot \frac{F_{\hat{\Omega}^c(\hat{\theta}_2^c)}(\hat{\theta}_2^c)}{1+\frac{q}{G}J(\hat{\theta}_2^c)}, \tag{1.18}$$

$$\widetilde{LR}_{\hat{\Omega}^c(\hat{\theta}_1)}(\hat{\theta}_2^c, \hat{\theta}_2^{c,r}) := \frac{G-p-q}{G} \cdot \frac{LR_{\hat{\Omega}^c(\hat{\theta}_1)}(\hat{\theta}_2^c, \hat{\theta}_2^{c,r})}{1+\frac{q}{G}J(\hat{\theta}_2^c)},$$

$$\widetilde{LM}_{\hat{\Omega}^c(\hat{\theta}_1)}(\hat{\theta}_2^{c,r}) := \frac{G-p-q}{G} \cdot \frac{LM_{\hat{\Omega}^c(\hat{\theta}_1)}(\hat{\theta}_2^{c,r})}{1+\frac{q}{G}J(\hat{\theta}_2^c)},$$

and the corresponding version of the t-statistic is

$$\tilde{t}_{\hat{\Omega}^c(\hat{\theta}_2^c)}(\hat{\theta}_2^c) := \sqrt{\frac{G-1-q}{G}} \cdot \frac{t_{\hat{\Omega}^c(\hat{\theta}_2^c)}(\hat{\theta}_2^c)}{\sqrt{1+\frac{q}{G}J(\hat{\theta}_2^c)}}.$$

The modified test statistics involve a scale multiplication factor that uses the usual J-statistic and a constant factor that adjusts the degrees of freedom.

It follows from Proposition 9 and Theorem 12 that

$$\left(F_{\hat{\Omega}^c(\hat{\theta}_2^c)}(\hat{\theta}_2^c), J(\hat{\theta}_2^c)\right) \xrightarrow{d} (\mathbb{F}_{2\infty}, \mathbb{J}_{\infty}) \tag{1.19}$$

$$\stackrel{d}{=} \left(G\left(\bar{B}_p - \mathbb{S}_{pq}\mathbb{S}_{qq}^{-1}\bar{B}_q\right)' \mathbb{S}_{pp\cdot q}^{-1}\left(\bar{B}_p - \mathbb{S}_{pq}\mathbb{S}_{qq}^{-1}\bar{B}_q\right)'/p, (G/q)\bar{B}_q'\mathbb{S}_{qq}^{-1}\bar{B}_q\right) \tag{1.20}$$

So

$$F_{\hat{\Omega}^c(\hat{\theta}_2^c)}(\hat{\theta}_2^c) \xrightarrow{d} \frac{G-p-q}{G}\frac{\mathbb{F}_{2\infty}}{1+\frac{q}{G}\mathbb{J}_{\infty}} \stackrel{d}{=} \frac{G-p-q}{pG}\xi_p'\tilde{\mathbb{S}}_{pp\cdot q}^{-1}\xi_p,$$

where

$$\xi_p := \frac{\sqrt{G}(\bar{B}_p - \bar{\mathbb{S}}_{pq}\bar{\mathbb{S}}_{qq}^{-1}\bar{B}_q)}{\sqrt{1+\bar{B}_q'\bar{\mathbb{S}}_{qq}^{-1}\bar{B}_q}}.$$

Similarly,

$$\tilde{t}_{\hat{\Omega}^c(\hat{\theta}_2^c)}(\hat{\theta}_2^c) \xrightarrow{d} \sqrt{\frac{G-1-q}{G}} \cdot \frac{\mathbb{T}_{2\infty}}{\sqrt{1+\frac{q}{G}\mathbb{J}_\infty}} \overset{d}{=} \frac{\xi_p}{\sqrt{\tilde{\mathbb{S}}_{pp\cdot q}}}.$$

In the proof of Theorem 11 we show that ξ_p follows a standard normal distribution $N(0, I_p)$ and that ξ_p is independent of $\tilde{S}_{pp\cdot q}^{-1}$. So the limiting distribution of $\tilde{F}_{\hat{\Omega}^c(\hat{\theta}_2^c)}(\hat{\theta}_2^c)$ is proportional to a quadratic form in the standard normal vector ξ_p with an independent inverse-Wishart distributed weighting matrix $\tilde{\mathbb{S}}_{pp\cdot q}^{-1}$. It follows from a theory of multivariate statistics that the limiting distribution of $\tilde{F}_{\Omega^c(\hat{\theta}_2^c)}(\hat{\theta}_2^c)$ is $F_{p,G-p-q}$. Similarly, the limiting distribution of $\tilde{t}_{\hat{\Omega}^c(\hat{\theta}_2^c)}(\hat{\theta}_2^c)$ is t_{G-1-q}. This is formalized in the following theorem.

Theorem 11 *Let Assumptions 1∼6 hold. Then*

(a) $\tilde{F}_{\hat{\Omega}^c(\hat{\theta}_2^c)}(\hat{\theta}_2^c) \xrightarrow{d} F_{p,G-p-q}$;

(b) $\widetilde{LR}_{\hat{\Omega}^c(\hat{\theta}_2^c)}(\hat{\theta}_2^c, \hat{\theta}_2^{c,r}) \xrightarrow{d} F_{p,G-p-q}$;

(c) $\widetilde{LM}_{\hat{\Omega}^c(\hat{\theta}_1)}(\hat{\theta}_2^{c,r}) \xrightarrow{d} F_{p,G-p-q}$;

(d) $\tilde{t}_{\hat{\Omega}^c}(\hat{\theta}_2^c) \xrightarrow{d} t_{G-1-q}$.

Together with the asymptotic equivalence between $\hat{\theta}_2^c, \hat{\theta}_{\text{CU-GEE}}$ and $\hat{\theta}_{\text{GMM}}^{cu}$ established in Proposition 10, the proof of Theorem 11 implies that the modified Wald, LR,LM, and t statistics based on $\hat{\theta}_{\text{CU-GEE}}$ and $\hat{\theta}_{\text{GMM}}^{cu}$ are all asymptotically F and t distributed under the fixed-G asymptotics. This equivalence relationship is consistent with the recent paper by Hwang and Sun (2015b) which establishes the asymptotic F and t limit theory of two-step GMM in time series setting. But our cluster-robust limiting distributions in Theorem 11 are different from Hwang and Sun (2015b) in terms of the multiplicative adjustment and the degrees of freedom correction.

It follows from the proofs of Theorem 11 and Proposition 9 that

$$\begin{aligned} \sqrt{N}\left(\hat{\theta}_2^c-\theta_0\right) &\xrightarrow{d} MN\left(0,\left(\Gamma'\Omega^{-1}\Gamma\right)^{-1}\cdot\left(1+\bar{B}_q'\bar{\mathbb{S}}_{qq}^{-1}\bar{B}_q\right)\right) \\ \text{and} \qquad J(\hat{\theta}_2^c) &\xrightarrow{d} (G/q)\bar{B}_q'\bar{\mathbb{S}}_{qq}^{-1}\bar{B}_q \end{aligned} \tag{1.21}$$

holds jointly under fixed-G asymptotics. Here, $MN(0,\mathbb{V})$ denotes a random variable that follows a mixed normal distribution with conditional variance $\mathbb{V}$. The random multiplication term $(1+\bar{B}_q'\tilde{S}_{qq}^{-1}\bar{B}_q)$ in (1.21) reflects the estimation uncertainty of CCE weighting matrix on the limiting distribution of $\sqrt{N}(\hat{\theta}_2^c-\theta_0)$. The fixed-$G$ limiting distribution in (1.21) is in sharp contrast to that of under the conventional large-G asymptotics as the latter completely ignores the variability in the cluster-robust GMM weighting matrix. By continuous mapping theorem,

$$\frac{\sqrt{N}\left(\hat{\theta}_2^c-\theta_0\right)}{\sqrt{1+(G/q)J(\hat{\theta}_2^c)}}\xrightarrow{d} N\left(0,\left(\Gamma'\Omega^{-1}\Gamma\right)^{-1}\right). \tag{1.22}$$

and this shows that the J-statistic modification factor in the denominator effectively cancels out the uncertainty of CCE to recover the limiting distribution of $\sqrt{N}(\hat{\theta}_2^c-\theta_0)$ under the conventional large-G asymptotics. In view of (1.22), the finite sample distribution of $\sqrt{N}(\hat{\theta}_2^c-\theta_0)$ can be well-approximated by $N(0,\widetilde{var}_{\hat{\Omega}^c(\hat{\theta}_1)}(\hat{\theta}_2^c))$ where

$$\widetilde{var}_{\hat{\Omega}^c(\hat{\theta}_1)}(\hat{\theta}_2^c):=\widehat{var}_{\hat{\Omega}^c(\hat{\theta}_1)}(\hat{\theta}_2^c)\cdot\left(1+\frac{q}{G}J(\hat{\theta}_2^c)\right). \tag{1.23}$$

The modification term $(1+(q/G)J(\hat{\theta}_2^c))^{-1}$ degenerates to one as G increases so that the two variance estimates in (1.23) become close to each other. Thus, the multiplicative term $(1+(q/G)J(\hat{\theta}_2^c))^{-1}$ in (1.18) can be regarded as a finite sample modification to the standard variance estimate $\widehat{var}_{\hat{\Omega}^c(\hat{\theta}_1)}(\hat{\theta}_2^c)$ under the large-G asymptotics. For more discussions about the role of J-statistic modification, see Hwang and Sun (2015b) which casts the two-step GMM problems into OLS estimation and inference in classical normal linear regression.

1.6 Finite Sample Variance Correction

1.6.1 Centered Two-step GMM Estimation

Define the infeasible two-step GMM estimator with the centered CCE weighting matrix $\hat{\Omega}^c(\theta_0)$:

$$\tilde{\theta}_2^c = \arg\min_{\theta\in\Theta} g_N(\theta)' \left(\hat{\Omega}^c(\theta_0)\right)^{-1} g_N(\theta).$$

Then

$$\sqrt{N}\left(\tilde{\theta}_2^c - \theta_0\right) = -\left[\Gamma' \left(\hat{\Omega}^c(\theta_0)\right)^{-1} \Gamma\right]^{-1} \Gamma' \left(\hat{\Omega}^c(\theta_0)\right)^{-1} \sqrt{N} g_N(\theta_0) + o_p(1)$$

. But we also have

$$\sqrt{N}\left(\hat{\theta}_2^c - \theta_0\right) = -\left[\Gamma' \left(\hat{\Omega}^c(\hat{\theta}_1)\right)^{-1} \Gamma\right]^{-1} \Gamma' \left(\hat{\Omega}^c(\hat{\theta}_1)\right)^{-1} \sqrt{N} g_N(\theta_0) + o_p(1) \quad (1.24)$$

Together with Lemma 8, this implies that

$$\sqrt{N}(\hat{\theta}_2^c - \theta_0) = \sqrt{N}(\tilde{\theta}_2^c - \theta_0) + o_p(1).$$

That is, the estimation error in $\hat{\theta}_1$ has no effect on the asymptotic distribution of $\sqrt{N}(\hat{\theta}_2^c - \theta_0)$ in the first-order asymptotic analysis. However, in finite samples $\hat{\theta}_2^c$ does have higher variation than $\tilde{\theta}_2^c$, and this can be attributed to the high variation in $\hat{\Omega}^c(\hat{\theta}_1)$ than $\hat{\Omega}^c(\theta_0)$. To account for this extra variation, we could develop a higher order asymptotic theory under the fixed-G asymptotics. But this is a formidable task that requires new technical machinery and lengthy calculations. Instead, we keep one additional term in the stochastic expansion of $\sqrt{N}(\hat{\theta}_2^c - \theta_0)$ in hopes of developing a finite sample correction to our asymptotic variance estimator.

To this end, we first introduce the notion of asymptotic equivalence in distribution $\xi_N \overset{a}{\sim} \eta_N$ for two stochastically bounded sequences of random vectors $\xi_N \in \mathbb{R}^\ell$ and $\eta_N \in \mathbb{R}^\ell$ when ξ_N and η_N coverge in distribution to each other. Now under the fixed-G asymptotics we have:

$$\begin{aligned}\sqrt{N}(\hat{\theta}_2^c - \theta_0) \overset{a}{\sim} -\left\{\Gamma' \left[\hat{\Omega}^c(\theta_0)\right]^{-1} \Gamma\right\}^{-1} \Gamma' \left[\hat{\Omega}^c(\theta_0)\right]^{-1} \sqrt{N} g_N(\theta_0) \\ + (\mathcal{E}_1 + \mathcal{E}_2)\sqrt{N}(\hat{\theta}_1 - \theta_0)\end{aligned}$$

where

$$\mathcal{E}_1 = -\left.\frac{\partial\left\{\Gamma' \left[\hat{\Omega}^c(\theta)\right]^{-1} \Gamma\right\}^{-1}}{\partial \theta'}\right|_{\theta=\theta_0} \Gamma' \left[\hat{\Omega}^c(\theta)\right]^{-1} g_N(\theta_0)$$

$$\mathcal{E}_2 = -\left\{\Gamma' \left[\hat{\Omega}^c(\theta)\right]^{-1} \Gamma\right\}^{-1} \left.\frac{\partial \Gamma' \left[\hat{\Omega}^c(\theta)\right]^{-1} g_N(\theta_0)}{\partial \theta'}\right|_{\theta=\theta_0}$$

are $d \times d$ matrices. In finite samples, if we estimate the term $\Gamma' \left[\hat{\Omega}^c(\theta_0)\right]^{-1} g_N(\theta_0)$ in $\mathcal{E}_1$ by $\hat{\Gamma}(\hat{\theta}_2^c)[\hat{\Omega}^c(\hat{\theta}_1)]^{-1} g_N(\hat{\theta}_2^c)$, then the estimate will be identically zero because of the FOC's. For this reason, we drop $\mathcal{E}_1$ and keep only $\mathcal{E}_2$, which leads to the distributional approximation:

$$\sqrt{N}(\hat{\theta}_2^c - \theta_0) \overset{a}{\sim} -\left\{\Gamma' \left[\hat{\Omega}^c(\theta_0)\right]^{-1} \Gamma\right\}^{-1} \Gamma' \left[\hat{\Omega}^c(\theta_0)\right]^{-1} \sqrt{N} g_N(\theta_0) + \mathcal{E}_2 \sqrt{N}(\hat{\theta}_1 - \theta_0). \tag{1.25}$$

Using element by element differentiation with respect to θ_j for $1 \le j \le d$, we can write the j-th column of $\mathcal{E}_2$ as

$$\mathcal{E}_2[.,j] = -\left\{\Gamma' \left[\hat{\Omega}^c(\theta_0)\right]^{-1} \Gamma\right\}^{-1} \Gamma' \left[\hat{\Omega}^c(\theta_0)\right]^{-1} \left.\frac{\partial \hat{\Omega}^c(\theta)}{\partial \theta_j}\right|_{\theta=\theta_0} \left[\hat{\Omega}^c(\theta_0)\right]^{-1} g_N(\theta_0), \tag{1.26}$$

where

$$\frac{\partial \hat{\Omega}^c(\theta_0)}{\partial \theta_j} = \Upsilon_j(\theta_0) + \Upsilon_j'(\theta_0) \text{ and}$$
$$\Upsilon_j(\theta_0) = \frac{1}{G}\sum_{g=1}^{G}\left[\frac{1}{\sqrt{L}}\sum_{r=1}^{L}\left(f_r^g(\theta_0) - \frac{1}{N}\sum_{s=1}^{N} f_s(\theta_0)\right)\right.$$
$$\left. \cdot \frac{1}{\sqrt{L}}\sum_{s=1}^{L}\left(\frac{\partial f_s^g(\theta_0)}{\partial \theta_j} - \frac{1}{N}\sum_{s=1}^{N}\frac{\partial f_s(\theta_0)}{\partial \theta_j}\right)'\right]. \quad (1.27)$$

Note that the term $\mathcal{E}_2\sqrt{N}(\hat{\theta}_1 - \theta_0)$ has no first order effect on the asymptotic distribution of $\sqrt{N}(\hat{\theta}_2^c - \theta_0)$. This is true because $\mathcal{E}_2$ converges to zero in probability. In fact, it follows from (1.26) and (1.27) that $\mathcal{E}_2 = O_p(N^{-1/2})$.

It follows from (1.25) that

$$\sqrt{N}(\hat{\theta}_2^c - \theta_0) \overset{a}{\sim} -\left(\left[\Gamma' \left(\Omega_\infty^c\right)^{-1}\Gamma\right]^{-1} \quad \mathcal{E}_N(\Gamma' W^{-1}\Gamma)^{-1} \right)\begin{pmatrix} \Gamma' \left(\Omega_\infty^c\right)^{-1}\Lambda Z \\ \Gamma' W^{-1}\Lambda Z \end{pmatrix} \quad (1.28)$$

where $Z \sim N(0, I_d)$, Z is independent of Ω_∞^c, $\mathcal{E}_N$ has the same marginal distribution as $\mathcal{E}_2$ but it is independent of Z and Ω_∞^c. It then follows that $\sqrt{N}(\hat{\theta}_2^c - \theta_0)$ is asymptotically equivalent in distribution to the mixed normal distribution with the conditional variance given by

$$\Xi_N = \begin{pmatrix} \left[\Gamma' \left(\Omega_\infty^c\right)^{-1}\Gamma\right]^{-1} \\ (\Gamma' W^{-1}\Gamma)^{-1}\mathcal{E}_N' \end{pmatrix}' \begin{pmatrix} \Gamma' \left(\Omega_\infty^c\right)^{-1}\Omega \left(\Omega_\infty^c\right)^{-1}\Gamma & \Gamma' \left(\Omega_\infty^c\right)^{-1}\Omega W^{-1}\Gamma \\ \Gamma' W^{-1}\Omega' \left(\Omega_\infty^c\right)^{-1}\Gamma & \Gamma' W^{-1}\Omega W^{-1}\Gamma \end{pmatrix}$$
$$\cdot \begin{pmatrix} \left[\Gamma' \left(\Omega_\infty^c\right)^{-1}\Gamma\right]^{-1} \\ (\Gamma' W^{-1}\Gamma)^{-1}\mathcal{E}_N' \end{pmatrix}.$$

Motivated by the above approximation, we propose to use the following corrected variance estimator:

$$\begin{aligned}
&\widehat{var}^{\text{adj}}_{\hat{\Omega}^c(\hat{\theta}_1)}(\hat{\theta}_2^c) = \frac{1}{N}\hat{\Xi}_N \\
&= \frac{1}{N}\left(\left[\hat{\Gamma}' \left[\hat{\Omega}^c(\hat{\theta}_1)\right]^{-1} \hat{\Gamma}\right]^{-1} \quad \widehat{\mathcal{E}}_N(\hat{\Gamma}' W_N^{-1}\hat{\Gamma})^{-1} \right) \\
&\times \begin{pmatrix} \hat{\Gamma}' \left[\hat{\Omega}^c\left(\hat{\theta}_1\right)\right]^{-1} \hat{\Gamma} & \hat{\Gamma}' W_N^{-1}\hat{\Gamma} \\ \hat{\Gamma}' W_N^{-1}\hat{\Gamma} & \hat{\Gamma}' W_N^{-1}\hat{\Omega}^c\left(\hat{\theta}_1\right) W_N^{-1}\hat{\Gamma} \end{pmatrix} \\
&\times \begin{pmatrix} \left[\hat{\Gamma}' \left[\hat{\Omega}^c\left(\hat{\theta}_1\right)\right]^{-1} \hat{\Gamma}'\right]^{-1} \\ (\hat{\Gamma}' W_N^{-1}\hat{\Gamma})^{-1}\widehat{\mathcal{E}}_N' \end{pmatrix} \\
&= \widehat{var}_{\hat{\Omega}^c(\hat{\theta}_1)}(\hat{\theta}_2^c) + \widehat{\mathcal{E}}_N \widehat{var}_{\hat{\Omega}^c(\hat{\theta}_1)}(\hat{\theta}_2^c) + \widehat{var}_{\hat{\Omega}^c(\hat{\theta}_1)}(\hat{\theta}_2^c)\widehat{\mathcal{E}}_N' + \widehat{\mathcal{E}}_N \widehat{var}(\hat{\theta}_1)\widehat{\mathcal{E}}_N' \qquad (1.29)
\end{aligned}$$

where

$$\begin{aligned}
\widehat{\mathcal{E}}_N[.,j] &= \left\{\hat{\Gamma}' \left[\hat{\Omega}^c(\hat{\theta}_1)\right]^{-1} \hat{\Gamma}'\right\}^{-1} \Gamma_N' \left\{ \left[\hat{\Omega}^c(\hat{\theta}_1)\right]^{-1} \frac{\partial \hat{\Omega}^c(\theta)}{\partial \theta_j}\bigg|_{\theta=\hat{\theta}_1} \left[\hat{\Omega}^c(\hat{\theta}_1)\right]^{-1} \right\} g_N(\hat{\theta}_2^c), \\
\hat{\Gamma} &= \hat{\Gamma}(\hat{\theta}_2^c).
\end{aligned}$$

The last three terms in (1.29), which are of smaller order, serve as a finite sample correction to the original variance estimator.

Windmeijer, (2005), too, has used the idea of variance correction, and his proposed correction has been widely implemented in applied work for simple models such as linear IV models and linear dynamic panel data models. However, Windmeijer, (2005) considers only an i.i.d. setting. Two principal differences distinguish Windmeijer's approach and ours. First, our asymptotic variance estimator involves a centered CCE; in contrast, Windmeijer's involves only a plain variance estimator. Second, we consider the fixed-G asymptotics; Windmeijer, (2005) considers the traditional asymptotics. More broadly, we often have to keep higher-order terms to develop a high order Edgeworth expansion. Here we choose to focus on variance correction instead of distribution correction, which is often the real target behind the Edgeworth expansion. In addition to technical reasons, a principal reason for our choice is that we have already developed more accurate fixed-G asymptotic approximations.

With the finite sample corrected variance estimator, we can construct the variance-corrected Wald statistic:

$$F^{\text{adj}}_{\hat{\Omega}^c(\hat{\theta}_1)}(\hat{\theta}^c_2) = (R\hat{\theta}^c_2 - r)' \left[R\widehat{var}^{\text{adj}}_{\hat{\Omega}^c(\hat{\theta}_1)}(\hat{\theta}^c_2) R' \right]^{-1} (R\hat{\theta}^c_2 - r)/p.$$

When $p = 1$ and for one-sided alternative hypotheses, we can construct the variance-corrected t-statistic:

$$t^{\text{adj}}_{\hat{\Omega}^c(\hat{\theta}_1)}(\hat{\theta}^c_2) = \frac{(R\hat{\theta}^c_2 - r)}{\sqrt{R\widehat{var}^{\mathbf{c}}_{\hat{\Omega}^c(\hat{\theta}_1)}(\hat{\theta}^c_2))R'}}.$$

Given that the variance correction terms are of smaller order, the variance-corrected statistic will have the same limiting distribution as the original statistic.

Assumption 7 *For each $g = 1, ..., G$ and $s = 1, ..., d$, define $Q^g_s(\theta)$ as*

$$Q^g_s(\theta) = \lim_{L \to \infty} E \left[\frac{1}{L} \sum_{k=1}^{L} \frac{\partial}{\partial \theta'} \left(\frac{\partial f^g_k(\theta)}{\partial \theta_s} \right) \right]$$

Then,

$$\sup_{\theta \in \mathcal{N}(\theta_0)} \left\| \frac{1}{L} \sum_{k=1}^{L} \frac{\partial}{\partial \theta'} \left(\frac{\partial f^g_k(\theta)}{\partial \theta_s} \right) - Q^g_s(\theta) \right\| \xrightarrow{p} 0.$$

holds for each $g = 1, ..., G$ and $s = 1, ..., d$ where $\mathcal{N}(\theta_0)$ is an open neighborhood of θ_0 and $\|\cdot\|$ is the Euclidean norm. Also, $Q^g_s(\theta_0) = Q_s(\theta_0)$ for $g = 1, ...G$.

This assumption trivially holds if the moment conditions are linear in parameters.

Theorem 12 *Let Assumptions 1∼7 hold. Then*

$$F^{adj}_{\hat{\Omega}^c(\hat{\theta}_1)}(\hat{\theta}^c_2) = F_{\hat{\Omega}^c(\hat{\theta}_1)}(\hat{\theta}^c_2) + o_p(1) \text{ and}$$
$$t^{adj}_{\hat{\Omega}^c(\hat{\theta}_1)}(\hat{\theta}^c_2) = t_{\hat{\Omega}^c(\hat{\theta}_1)}(\hat{\theta}^c_2) + o_p(1).$$

In the proof of Theorem 12, we show that $\widehat{\mathcal{E}}_N = (1 + o_p(1))\mathcal{E}_2$. That is, the high order correction term has been consistently estimated in a relative sense. This guarantees that $\widehat{\mathcal{E}}_N$ is a reasonable estimator for $\mathcal{E}_2$, which is of order $o_p(1)$.

As a direct implication of Theorem 12, the fixed-G asymptotic distributions of $F^{\mathbf{c}}_{\hat{\Omega}^c(\hat{\theta}_1)}(\hat{\theta}^c_2)$ and $t^{\mathbf{c}}_{\hat{\Omega}^c(\hat{\theta}_1)}(\hat{\theta}^c_2)$ are

$$F^{\text{adj}}_{\hat{\Omega}^c(\hat{\theta}_1)}(\hat{\theta}^c_2) \xrightarrow{d} \mathbb{F}_{2\infty} \text{ and } t^{\text{adj}}_{\hat{\Omega}^c(\hat{\theta}_1)}(\hat{\theta}^c_2) \xrightarrow{d} \mathbb{T}_{2\infty}.$$

Note that the corrected variance estimator is not necessarily larger than the original estimator in finite samples. In the simulation work we consider later, we observe that the smaller value of corrected variance estimate rather deteriorates the finite sample performance of variance-corrected statistics. To avoid this undesirable situation, we make an adjustment to $\widehat{var}^{\text{adj}}_{\hat{\Omega}^c(\hat{\theta}_1)}(\hat{\theta}^c_2)$ so that $\widehat{var}^{\text{adj}}_{\hat{\Omega}^c(\hat{\theta}_1)}(\hat{\theta}^c_2) - \widehat{var}_{\hat{\Omega}^c(\hat{\theta}_1)}(\hat{\theta}^c_2)$ is guaranteed to be positive semidefinite. This is an easy task. Let

$$M_N = \widehat{var}^{\text{adj}}_{\hat{\Omega}^c(\hat{\theta}_1)}(\hat{\theta}^c_2) - \widehat{var}_{\hat{\Omega}^c(\hat{\theta}_1)}(\hat{\theta}^c_2),$$

$\{\lambda_i\}^d_{i=1}$ be the eigenvalues of M_N and $V_N L_N V_N'$ be the eigen-decomposition of M_N where $P_N = diag\,(\lambda_i) \in \mathbb{R}^{d\times d}$. Define

$$\tilde{P}_N = diag\,(\max(\lambda_i, 0)) \text{ and } \tilde{M}_N = V_N \tilde{P}_N V_N'.$$

The corresponding regularized version of $\widehat{var}^{\text{adj}}_{\hat{\Omega}^c(\hat{\theta}_1)}(\hat{\theta}^c_2)$ is given by

$$\widehat{var}^{\text{adj+}}_{\hat{\Omega}^c(\hat{\theta}_1)}(\hat{\theta}^c_2) = \widehat{var}^{\text{adj}}_{\hat{\Omega}^c(\hat{\theta}_1)}(\hat{\theta}^c_2) + \tilde{M}_N \tag{1.30}$$

The corresponding modified Wald statistic is

$$F^{\text{adj+}}_{\hat{\Omega}^c(\hat{\theta}_1)}(\hat{\theta}^c_2) = (R\hat{\theta}^c_2 - r)' \left[R\widehat{var}^{\text{adj+}}_{\hat{\Omega}^c(\hat{\theta}_1)}(\hat{\theta}^c_2)R' \right]^{-1} (R\hat{\theta}^c_2 - r)/p. \tag{1.31}$$

Similarly, the modified t-statistic is

$$t^{\text{adj+}}_{\hat{\Omega}^c(\hat{\theta}_1)}(\hat{\theta}_2) = \frac{(R\hat{\theta}^c_2 - r)}{\sqrt{R\widehat{var}^{\text{adj+}}_{\hat{\Omega}^c(\hat{\theta}_1)}(\hat{\theta}^c_2)R'}}.$$

The limiting distributions of the modified Wald and t statistics are again $\mathbb{F}_{2\infty}$ and $\mathbb{T}_{2\infty}$.

1.6.2 CU Estimation

For the CU-GEE estimator, we have the following expansion

$$\sqrt{N}(\hat{\theta}^{\text{cu}}_{\text{GEE}} - \theta_0)$$
$$= -\left(\Gamma'\left(\hat{\Omega}^c(\theta_0)\right)^{-1}\Gamma\right)^{-1}\Gamma'\left(\hat{\Omega}^c(\theta_0)\right)^{-1}\sqrt{N}g_N(\theta_0) + \mathcal{E}_2\sqrt{N}(\hat{\theta}_{\text{CU-GEE}} - \theta_0) \tag{1.32}$$
$$+ o_p(1). \tag{1.33}$$

This can be regarded as a special case of (1.25) wherein the first-step estimator $\hat{\theta}_1$ is replaced by the CU-GEE estimator. So

$$\sqrt{N}(\hat{\theta}^{\text{cu}}_{\text{GEE}} - \theta_0) \overset{a}{\sim} -(I_d - \mathcal{E}_2)^{-1}\left(\Gamma'\left(\hat{\Omega}^c(\theta_0)\right)^{-1}\Gamma\right)^{-1}\Gamma'\left(\hat{\Omega}^c(\theta_0)\right)^{-1}\sqrt{N}g_N(\theta_0). \tag{1.34}$$

We can obtain the same expression for the CU-GMM estimator $\sqrt{N}(\hat{\theta}_{\text{CU-GMM}} - \theta_0)$.

In view of the representation in (1.34), the corrected variance estimator for the CU type estimators can be constructed as follows:

$$\widehat{var}^{\text{adj}}_{\hat{\Omega}^c(\hat{\theta}^{\text{cu}}_{\text{GEE}})}(\hat{\theta}^{\text{cu}}_{\text{GEE}}) = \left(I_d - \widehat{\mathcal{E}}_{\text{CU-GEE}}\right)^{-1} \widehat{var}\left(\hat{\theta}^{\text{cu}}_{\text{GEE}}\right)\left(I_d - \widehat{\mathcal{E}}'_{\text{CU-GEE}}\right)^{-1}$$
$$\widehat{var}^{\text{adj}}_{\hat{\Omega}^c(\hat{\theta}^{\text{cu}}_{\text{GEE}})}(\hat{\theta}^{cu}_{\text{GMM}}) = \left(I_d - \widehat{\mathcal{E}}_{\text{CU-GMM}}\right)^{-1} \widehat{var}\left(\hat{\theta}^{cu}_{\text{GMM}}\right)\left(I_d - \widehat{\mathcal{E}}'_{\text{CU-GMM}}\right)^{-1}$$

where

$$\widehat{\mathcal{E}}_{\text{CU-GEE}}[.,j] = \left\{\hat{\Gamma}'\left[\hat{\Omega}^c(\hat{\theta}^{\text{cu}}_{\text{GEE}})\right]^{-1}\hat{\Gamma}'\right\}^{-1}$$
$$\times\,\hat{\Gamma}'\left\{\left[\hat{\Omega}^c(\hat{\theta}_{\text{CU-GEE}})\right]^{-1}\frac{\partial\hat{\Omega}^c(\hat{\theta}_{\text{CU-GEE}})}{\partial\theta_j}\left[\hat{\Omega}^c(\hat{\theta}_{\text{CU-GEE}})\right]^{-1}\right\}g_N(\hat{\theta}^{\text{cu}}_{\text{GEE}})$$

and $\widehat{\mathcal{E}}_{\text{CU-GMM}}$ is defined in the same way but with $\hat{\theta}^{\text{cu}}_{\text{GEE}}$ replaced by $\hat{\theta}_{\text{CU-GMM}}$. The adjusted variance estimators can be defined using the same formula provided in (1.30).

The adjusted (and regularized) test statistics associated with the CU type estimators are

$$F^{\text{adj}+}_{\hat{\Omega}^c(\hat{\theta}^{\text{cu}}_{\text{GEE}})}(\hat{\theta}_{\text{CU-GEE}})$$
$$= \left(R\hat{\theta}^{\text{cu}}_{\text{GEE}} - r\right)'\left(R\widehat{var}^{\text{adj}+}_{\hat{\Omega}^c(\hat{\theta}^{\text{cu}}_{\text{GEE}})}(\hat{\theta}_{\text{CU-GEE}})R'\right)^{-1}\left(R\hat{\theta}_{\text{CU-GEE}} - r\right)'/p,$$

and

$$t^{\text{adj}+}_{2,\hat{\Omega}^c(\hat{\theta}^{\text{cu}}_{\text{GEE}})}(\hat{\theta}_{\text{CU-GEE}}) = \frac{(R\hat{\theta}^{\text{cu}}_{\text{GEE}} - r)}{\sqrt{R\widehat{var}^{\text{adj}+}_{\hat{\Omega}^c(\hat{\theta}^{\text{cu}}_{\text{GEE}})}(\hat{\theta}_{\text{CU-GEE}})R'}},$$

when $p = 1$. We can easily show that the Wald statistic converge weakly to $\mathbb{F}_{2\infty}$ and the t-statistic converge weakly to $\mathbb{T}_{2\infty}$.

With the finite sample corrected and adjusted variance estimators in place, the test statistics based on all three estimators $\hat{\theta}^c_2, \hat{\theta}_{\text{CU-GEE}}$ and $\hat{\theta}^{cu}_{\text{GMM}}$ converge in distribution to the same nonstandard distributions. A multiplicative modification provided in Section 1.5 can then turn the nonstandard distributions $\mathbb{F}_{2\infty}$ and $\mathbb{T}_{2\infty}$ into standard F and t distributions.

1.7 Application to Linear Dynamic Panel Model

This section discusses how to implement our inference procedures in the context of a linear dynamic panel model:

$$y_{it} = \gamma y_{it-1} + x_{it}'\beta + \eta_i + u_{it}, \tag{1.35}$$

for $i = 1, ..., N$, $t = 1, ..., T$, where $x_{it} = (x_{it}^1, ..., x_{it}^{d-1})' \in \mathbb{R}^{d-1}$. The unknown parameter vector is $\theta = (\gamma, \beta')' \in \mathbb{R}^d$. We assume that the vector of regressors $w_{it} = (y_{it-1}, x_{it}')'$ is correlated with η_i and is predetermined with respect to u_{it}, i.e., $E(w_{it}u_{it+s}) = 0$ for $s = 0, ..., T-t$.

When T is small, popular panel estimators such as the fixed-effects estimator or first-differenced estimator suffer from the Nickel bias (Nickell, 1981). Arellano and Bond (1991) consider the first-differenced equation

$$\Delta y_{it} = \Delta w_{it}'\theta + \Delta u_{it}, \ t = 2, ..., T$$

and propose a consistent IV estimator that employs the lagged w_{it} as the instrument. Building upon Anderson and Hsiao (1981), Arellano and Bond (1991, AB hereafter) examine the problem in a GMM framework and find $dT(T-1)/2$ sequential instruments:

$$\underset{(T-1)\times d(T-1)T/2}{Z_i} = diag(z_{i2}', ..., z_{iT}')$$
$$z_{it} = (y_{i0}, ..., y_{it-2}, x_{i1}', ..., x_{it-1}')', 2 \le t \le T.$$

The moment conditions are then given by

$$E\left(Z_i'\Delta u_i\right) = 0,$$

where Δu_i is the $(T-1)$ vector $(\Delta u_{i2}, ..., \Delta u_{iT})'$. The original AB method assumes away cross-sectional dependence, but clustered dependence can be easily accommodated. Here we assume that the moment vector $\{Z_i'\Delta u_i\}_{i=1}^N$ can be partitioned into independent clusters. That is, $\{Z_i'\Delta u_i\}_{i=1}^N = \cup_{g=1}^G \{Z_i^{g\prime}\Delta u_i^g\}_{i=1}^{L_N}$ with $Z_i^{g\prime}\Delta u_i^g$ and $Z_j^{h\prime}\Delta u_j^h$ being independent for all $g \neq h$.

The first-step GMM estimator with weighting matrix W_N^{-1} is given by

$$\hat{\theta}_1 = \left(\Delta w' Z W_N^{-1} Z' \Delta w\right)^{-1} \Delta w' Z W_N^{-1} Z' \Delta y,$$

where Z' is the $dT(T-1)/2 \times N(T-1)$ matrix $(Z_1', Z_2', ..., Z_N')$, Δw_i is the $(T-1) \times d$ matrix $(\Delta w_{i2}, ..., \Delta w_{iT})'$, Δy_i is the $(\Delta y_{i2}, ..., \Delta y_{iT})'$, Δw and Δy are $(\Delta w_1', ..., \Delta w_N')'$ and $(\Delta y_1', ..., \Delta y_N')'$, respectively. The corresponding Wald statistic[4] for testing $H_0 : R\theta_0 = r$ vs $H_1 : R\theta_0 \neq r$ is given by

$$F(\hat{\theta}_1) := (R\hat{\theta}_1 - r)' \left\{ R\widehat{var}(\hat{\theta}_1)R' \right\}^{-1} (R\hat{\theta}_1 - r)/p$$

where

$$\widehat{var}(\hat{\theta}_1) = N\left(\Delta w' Z W_N^{-1} Z' \Delta w\right)^{-1} \left(\Delta w' Z W_N^{-1} \hat{\Omega}(\hat{\theta}_1) W_N^{-1} Z' \Delta w\right) \left(\Delta w' Z W_N^{-1} Z' \Delta w\right)^{-1}.$$

Let $Z_{(g)}$ be the $L_N(T-1) \times dT(T-1)/2$ matrix obtained by stacking all Z_i's belonging to cluster g. Similarly, let $\Delta\hat{u}_{(g)}$ be the $L_N(T-1)$ stacked vector of the estimated first-differenced errors $\Delta\hat{u}_i = \Delta y_i - \Delta w_i'\hat{\theta}_1$. Then, in the presence of clustered dependence, the CCE and centered CCE are constructed as follows:

$$\hat{\Omega}(\hat{\theta}_1) = \frac{1}{G}\sum_{g=1}^G \left(\frac{Z_{(g)}'\Delta\hat{u}_{(g)}}{\sqrt{L_N}}\right)\left(\frac{Z_{(g)}'\Delta\hat{u}_{(g)}}{\sqrt{L_N}}\right)'$$

$$\hat{\Omega}^c(\hat{\theta}_1) = \frac{1}{G}\sum_{g=1}^G \left(\frac{Z_{(g)}'\Delta\hat{u}_{(g)}}{\sqrt{L_N}} - \frac{1}{G}\sum_{h=1}^G \frac{Z_{(h)}'\Delta\hat{u}_{(h)}}{\sqrt{L_N}}\right)\left(\frac{Z_{(g)}'\Delta\hat{u}_{(g)}}{\sqrt{L_N}} - \frac{1}{G}\sum_{h=1}^G \frac{Z_{(h)}'\Delta\hat{u}_{(h)}}{\sqrt{L_N}}\right)'.$$

Using the centered CCE $\hat{\Omega}^c(\hat{\theta}_1)$ as the weighting matrix, the two-step GMM

[4] The formula for the t-statistic, which is omitted here, is straightforward.

estimator $\hat{\theta}_2^c$ is

$$\hat{\theta}_2^c = \left\{ \Delta w' Z \left[\hat{\Omega}^c(\hat{\theta}_1) \right]^{-1} Z' \Delta w \right\}^{-1} \Delta w' Z \left[\hat{\Omega}^c(\hat{\theta}_1) \right]^{-1} Z' \Delta y,$$

and the Wald statistic for $\hat{\theta}_2^c$ is

$$F_{\hat{\Omega}^c(\hat{\theta}_1)}(\hat{\theta}_2^c) := (R\hat{\theta}_2^c - r)' \{ R\widehat{var}_{\hat{\Omega}^c(\hat{\theta}_1)}(\hat{\theta}_2) R' \}^{-1} (R\hat{\theta}_2^c - r)/p,$$
$$\widehat{var}_{\hat{\Omega}^c(\hat{\theta}_1)}(\hat{\theta}_2^c) = N \left\{ \Delta w' Z \left[\hat{\Omega}^c(\hat{\theta}_1) \right]^{-1} Z' \Delta w \right\}^{-1}.$$

Under the conventional large-G asymptotics, both $F(\hat{\theta}_1)$ and $F_{\hat{\Omega}^c(\hat{\theta}_1)}(\hat{\theta}_2^c)$ are asymptotically χ_p^2/p. Under our fixed-G asymptotics, we have

$$F(\hat{\theta}_1) \xrightarrow{d} \frac{G}{G-p} F_{p,G-p} \text{ and}$$
$$F_{\hat{\Omega}^c(\hat{\theta}_1)}(\hat{\theta}_2^c) \xrightarrow{d} \frac{G}{G-p-q} F_{p,G-p-q}\left(\|\Delta\|^2 \right). \tag{1.36}$$

In addition to utilizing these new approximations, we suggest a variance correction in order to capture the higher order effect of $\hat{\theta}_1$ on $\hat{\Omega}^c(\hat{\theta}_1)$. The finite sample corrected variance is

$$\widehat{var}^{\mathbf{c}}_{\hat{\Omega}^c(\hat{\theta}_1)}(\hat{\theta}_2^c) = \widehat{var}_{\hat{\Omega}^c(\hat{\theta}_1)}(\hat{\theta}_2^c) + \widehat{\mathcal{E}}_N \widehat{var}_{\hat{\Omega}^c(\hat{\theta}_1)}(\hat{\theta}_2^c) + \widehat{var}_{\hat{\Omega}^c(\hat{\theta}_1)}(\hat{\theta}_2^c) \widehat{\mathcal{E}}_N' + \widehat{\mathcal{E}}_N \widehat{var}(\hat{\theta}_1) \widehat{\mathcal{E}}_N' \tag{1.37}$$

where the j-th column is given by

$$\widehat{\mathcal{E}}_N[.,j] = -\left\{ \Delta w' Z \left[\hat{\Omega}^c(\hat{\theta}_1) \right]^{-1} Z' \Delta w \right\}^{-1} \Delta w' Z \left[\hat{\Omega}^c(\hat{\theta}_1) \right]^{-1}$$
$$\cdot \left. \frac{\partial \hat{\Omega}^c(\theta)}{\partial \theta_j} \right|_{\theta=\theta} \left[\hat{\Omega}^c(\theta) \right]^{-1} Z' \Delta \hat{u}_2,$$
$$\Delta \hat{u}_2 = \Delta y - \Delta w \hat{\theta}_2^c$$

and

$$\left.\frac{\partial \hat{\Omega}^c(\theta)}{\partial \theta_j}\right|_{\theta=\hat{\theta}_1} = \Upsilon_j(\hat{\theta}_1) + \Upsilon_j'(\hat{\theta}_1),$$

$$\Upsilon_j(\hat{\theta}_1) = -\frac{1}{G}\sum_{g=1}^{G}\left(\frac{Z_{(g)}'\Delta w_{j,(g)}}{\sqrt{L}} - \frac{1}{G}\sum_{h=1}^{G}\frac{Z_{(h)}'\Delta w_{j,(h)}}{\sqrt{L}}\right)$$

$$\cdot\left(\frac{Z_{(g)}'\Delta \hat{u}_{(g)}}{\sqrt{L}} - \frac{1}{G}\sum_{h=1}^{G}\frac{Z_{(h)}'\Delta \hat{u}_{(h)}}{\sqrt{L}}\right)'$$

$$\underset{(T-1)\times d}{\Delta w_i} = (\Delta w_{1,i}, ..., \Delta w_{d,i}) \text{ and } \underset{L_N(T-1)\times d}{\Delta w_{(g)}} = (\Delta w_{1,(g)}, ..., \Delta w_{d,(g)})$$

for each $j = 1, ..., d$. Here, $\Delta w_{(g)}$ is a $L_N(T-1)\times d$ matrix that stacks $\{\Delta w_i\}_{i=1}^N$ belonging to the group g. The extra adjustment toward a 'larger' corrected estimator $\widehat{var}^{\text{adj+}}_{\hat{\Omega}^c(\hat{\theta}_1)}(\hat{\theta}_2^c)$ directly follows from (1.30).

Based on the finite sample corrected variance estimator in (1.37) and the usual J-statistic, we construct the modified Wald statistic:

$$\tilde{F}^{\text{adj+}}_{\hat{\Omega}^c(\hat{\theta}_1)}(\hat{\theta}_2^c) = \frac{G-p-q}{G}\frac{F^{\text{adj+}}_{\hat{\Omega}^c(\hat{\theta}_1)}(\hat{\theta}_2^c)}{1+\frac{q}{G}J(\hat{\theta}_2^c)} \tag{1.38}$$

where

$$J(\hat{\theta}_2^c) := N g_N(\hat{\theta}_2^c)'\left(\hat{\Omega}^c(\hat{\theta}_2^c)\right)^{-1}(\theta) g_N(\hat{\theta}_2^c).$$

From the F limit theory in Section 1.5, we have

$$\tilde{F}^{\text{adj+}}_{\hat{\Omega}^c(\hat{\theta}_1)}(\hat{\theta}_2^c) \xrightarrow{d} F_{p,G-p-q} \tag{1.39}$$

and

$$\frac{G-q}{G}J(\hat{\theta}_2^c) \xrightarrow{d} F_{q,G-q}.$$

Critical values for from the F distribution are readily available from statistical tables.

1.8 Simulation Evidence

1.8.1 Design

We continue focusing on the dynamic panel data model in the previous section with $d = 4$,

$$y_{it} = \gamma y_{it-1} + x_{1,it}\beta_1 + x_{2,it}\beta_2 + x_{3,it}\beta_3 + \eta_i + u_{it}.$$

The unknown parameter vector is $\theta = (\gamma, \beta_1, \beta_2, \beta_3)$ and the corresponding covariates are $w_{it} = (y_{it-1}, x_{it})'$ with $x_{it} = (x_{1,it}, x_{2,it}, x_{3,it})'$. The true value of θ is chosen as $\theta_0 = (0.5, 1, 1, 1)$. We denote $s^g_{it} = (s^g_{1,it}, ..., s^g_{k,it})'$ as any vector valued observations in cluster g, and stack all observations at same period by cluster to define $s_{(g),t} = (s^g_{1t}, ..., s^g_{L_N t})'$. The j-th predetermined regressor $x^g_{j,it}$ are generated according to the following process:

$$x^g_{j,it} = \rho x^g_{j,it-1} + \eta^g_i + \rho u^g_{it-1} + e^g_{j,it}$$

for $j = 1, 2, 3$, $i = 1, ..., L_N$, and $t = 1, ..., T$. We characterize the within-cluster dependence in $\eta_{(g)}$, $e_{(g),t}$ and $u_{(g),t}$ by spatial locations that are indexed by a one-dimensional lattice. Define Σ_η and Σ_u to be $L_N \times L_N$ matrices whose (i, j)-th elements are $\sigma^\eta_{ij} = \lambda^{|i-j|}$ and $\sigma^u_{ij} = \lambda^{|i-j|}$, respectively, and Σ_e to be a $3L_N \times 3L_N$ block diagonal matrix with diagonal matrix $\Sigma_{d,e}$ of size $L_N \times L_N$ for $d = 1, ..., 3$. The (i, j)-th element of $\Sigma_{d,e}$ is $\sigma^e_{d,ij} = \lambda^{|i-j|}$ for $d = 1, ..., 3$. The parameter λ governs the degree of spatial dependence in each cluster. When $\lambda = 0$, there is no clustered dependence and our model reduces to that of Windmeijer, (2005) which considers a static panel data model with only one regressor.

The individual fixed effects and shocks in group g are generated by:

$$\begin{aligned} \eta_{(g)} &\sim \text{i.i.d.}N(0, \Sigma_\eta), \text{ vec}(e_{(g),t}) \sim \text{i.i.d.}N(0, \Sigma_e), \\ u_{(g),t} &= \tau_t \Sigma_u^{1/2} (\delta_1^g \omega_{1t}^g, , .., \delta_{L_N}^g \omega_{L_N t}^g)', \\ \delta_i^g &\sim \text{i.i.d.}U[0.5, 1.5], \text{ and } \omega_{it}^g \sim i.i.d.\chi_1^2 - 1 \end{aligned} \tag{1.40}$$

for $i = 1, ..., L_N$ and $t = 1, ..., T$ where $\tau_t = 0.5 + 0.1(t-1)$. The DGP of individual shock $u_{(g),t}$ in (1.41) features a non-Gaussian process which is heteroskedastic over both time t and individual i. Also, the clustered dependence structure implies

$$\{\eta_{(g)}, \text{vec}(e_{(g),t}), \delta_{(g)}, \omega_{(g),t}\} z \perp \{\eta_{(h)}, \text{vec}(e_{(h),t}), \delta_{(h)}, \omega_{(h),t}\}$$

for $g \neq h$ at any t and s.

Before we draw an estimation sample for $t = 1, ..., T$, 50 initial values are generated with $\tau_t = 0.5$ for $t = -49, ..., 0$, $x_{d,i,-49}^g \sim$i.i.d.$N(\eta_i^g/(1-\rho), (1-\rho)^{-1}\Sigma_{d,e})$ for $d = 1, ..., 3$, and $y_{i,-49}^g = (\sum_{d=1}^3 x_{d,i,-49}\beta_d + \eta_i^g + u_{i,-49}^g)/(1-\gamma)$. We fix the values of λ and ρ at 0.75 and 0.70, respectively; thus each observation is reasonably persistent with respect to both time and spatial dimensions. We set the number of time periods to be $T = 4$. The parameters are estimated by the first differenced GMM (AB estimator) as described in the previous section. The initial first-step estimator is the two stage least square (2SLS) with $W_N = (1/N)\sum_{i=1}^N Z_i' H Z_i$ where H is a matrix that consists with 2's on the main diagonal, with -1's on the main diagonal, and zeros elsewhere. With all possible lagged instruments, the number of moment conditions for the AB estimator is $dT(T-1)/2 = 24$ and the degrees of over-identification is $q = 20$. It could be better to use only a subset of full moment conditions because using this full set of instruments may lead to poor finite sample properties, especially when the number of clusters G is small. Thus, we also employ a reduced set of instruments; that is, we use the most recent lag $z_{it} = (y_{it-2}, x_{it-1}')'$, leading to $d(T-1) = 12$ moment conditions.

1.8.2 Choice of tests

We focus on the Wald type of tests as the Monte Carlo results for other types of tests are qualitatively similar. We examine the empirical size of a variety of testing procedures, all of which are based on first-step or two-step GMM estimators. For the first-step procedures, we consider the unmodified F-statistic $F_1 := F_1(\hat{\theta}_1)$ and the degrees-of-freedom modified F-statistic $[(G-P)/G]\,F_1$ where the associated critical values are $\chi_p^{1-\alpha}/p$ and $\mathcal{F}_{p,G-p}^{1-\alpha}$, respectively. These two tests have the same size-adjusted power, because the modification only involves a constant multiplier factor.

For the two-step GMM estimation and related tests, we examine five different procedures. The first three tests use different test statistics but the same critical value $\chi_p^{1-\alpha}/p$. The first test uses the "plain" F-statistic $F_2 := F_{\hat{\Omega}^c(\hat{\theta}_1)}(\hat{\theta}_2^c)$ as defined in (1.10). The second test uses the statistic $[(G-p-q)/G] \cdot F_2$ where $(G-p-q)/G$ is the degree-of-freedom correction factor. The third test uses $\tilde{F}_2 := F_{\hat{\Omega}^c(\hat{\theta}_1)}(\hat{\theta}_2^c)$ as defined in (1.18). Note that

$$\tilde{F}_2 = \frac{(G-p-q)}{G} \cdot \frac{F_2}{1+(q/G)J(\hat{\theta}_2^c)}.$$

Compared to the second test statistic, $\tilde{F}_2$ has the additional J-statistic correction factor $(1+(q/G)J(\hat{\theta}_2^c))^{-1}$. The three tests use increasingly more sophisticated test statistics. Because $[(G-p-q)/G] \to 1$ and $(1+(q/G)J(\hat{\theta}_2^c))^{-1} \to 1$ as $G \to \infty$, both corrections can be regarded as devices for finite sample improvement under the large-G asymptotics.

1.9 Simulation Evidence

1.9.1 Design

We continue focusing on the dynamic panel data model in the previous section with $d = 4$,

$$y_{it} = \gamma y_{it-1} + x_{1,it}\beta_1 + x_{2,it}\beta_2 + x_{3,it}\beta_3 + \eta_i + u_{it}.$$

The unknown parameter vector is $\theta = (\gamma, \beta_1, \beta_2, \beta_3)$ and the corresponding covariates are $w_{it} = (y_{it-1}, x_{it})'$ with $x_{it} = (x_{1,it}, x_{2,it}, x_{3,it})'$. The true value of θ is chosen as $\theta_0 = (0.5, 1, 1, 1)$. We denote $s_{it}^g = (s_{1,it}^g, ..., s_{k,it}^g)'$ as any vector valued observations in cluster g, and stack all observations at same period by cluster to define $s_{(g),t} = (s_{1t}^g, ..., s_{L_N t}^g)'$. The j-th predetermined regressor $x_{j,it}^g$ are generated according to the following process:

$$x_{j,it}^g = \rho x_{j,it-1}^g + \eta_i^g + \rho u_{it-1}^g + e_{j,it}^g$$

for $j = 1, 2, 3$, $i = 1, ..., L_N$, and $t = 1, ..., T$. We characterize the within-cluster dependence in $\eta_{(g)}$, $e_{(g),t}$ and $u_{(g),t}$ by spatial locations that are indexed by a one-dimensional lattice. Define Σ_η and Σ_u to be $L_N \times L_N$ matrices whose (i, j)-th elements are $\sigma_{ij}^\eta = \lambda^{|i-j|}$ and $\sigma_{ij}^u = \lambda^{|i-j|}$, respectively, and Σ_e to be a $3L_N \times 3L_N$ block diagonal matrix with diagonal matrix $\Sigma_{d,e}$ of size $L_N \times L_N$ for $d = 1, ..., 3$. The (i, j)-th element of $\Sigma_{d,e}$ is $\sigma_{d,ij}^e = \lambda^{|i-j|}$ for $d = 1, ..., 3$. The parameter λ governs the degree of spatial dependence in each cluster. When $\lambda = 0$, there is no clustered dependence and our model reduces to that of Windmeijer, (2005) which considers a static panel data model with only one regressor.

The individual fixed effects and shocks in group g are generated by:

$$\begin{aligned} \eta_{(g)} &\sim \text{i.i.d.} N(0, \Sigma_\eta),\ \text{vec}(e_{(g),t}) \sim \text{i.i.d.} N(0, \Sigma_e), \\ u_{(g),t} &= \tau_t \Sigma_u^{1/2} (\delta_1^g \omega_{1t}^g, , .., \delta_{L_N}^g \omega_{L_N t}^g)', \\ \delta_i^g &\sim \text{i.i.d.} U[0.5, 1.5],\ \text{and}\ \omega_{it}^g \sim i.i.d. \chi_1^2 - 1 \end{aligned} \tag{1.41}$$

for $i = 1, ..., L_N$ and $t = 1, ..., T$ where $\tau_t = 0.5 + 0.1(t-1)$. The DGP of individual shock $u_{(g),t}$ in (1.41) features a non-Gaussian process which is heteroskedastic over both time t and individual i. Also, the clustered dependence structure implies

$$\{\eta_{(g)}, \text{vec}(e_{(g),t}), \delta_{(g)}, \omega_{(g),t}\} z \perp \{\eta_{(h)}, \text{vec}(e_{(h),t}), \delta_{(h)}, \omega_{(h),t}\}$$

for $g \neq h$ at any t and s.

Before we draw an estimation sample for $t = 1, ..., T$, 50 initial values are generated with $\tau_t = 0.5$ for $t = -49, ..., 0$, $x_{d,i,-49}^g \sim$i.i.d.$N(\eta_i^g/(1-\rho), (1-\rho)^{-1}\Sigma_{d,e})$ for $d = 1, ..., 3$, and $y_{i,-49}^g = (\sum_{d=1}^3 x_{d,i,-49}\beta_d + \eta_i^g + u_{i,-49}^g)/(1-\gamma)$. We fix the values of λ and ρ at 0.75 and 0.70, respectively; thus each observation is reasonably persistent with respect to both time and spatial dimensions. We set the number of time periods to be $T = 4$. The parameters are estimated by the first differenced GMM (AB estimator) as described in the previous section. The initial first-step estimator is the two stage least square (2SLS) with $W_N = (1/N)\sum_{i=1}^N Z_i' H Z_i$ where H is a matrix that consists with 2's on the main diagonal, with -1's on the main diagonal, and zeros elsewhere. With all possible lagged instruments, the number of moment conditions for the AB estimator is $dT(T-1)/2 = 24$ and the degrees of over-identification is $q = 20$. It could be better to use only a subset of full moment conditions because using this full set of instruments may lead to poor finite sample properties, especially when the number of clusters G is small. Thus, we also employ a reduced set of instruments; that is, we use the most recent lag $z_{it} = (y_{it-2}, x_{it-1}')'$, leading to $d(T-1) = 12$ moment conditions.

1.9.2 Choice of tests

We focus on the Wald type of tests as the Monte Carlo results for other types of tests are qualitatively similar. We examine the empirical size of a variety of testing procedures, all of which are based on first-step or two-step GMM estimators. For the first-step procedures, we consider the unmodified F-statistic $F_1 := F_1(\hat{\theta}_1)$ and the degrees-of-freedom modified F-statistic $[(G-P)/G]\, F_1$ where the associated critical values are $\chi_p^{1-\alpha}/p$ and $\mathcal{F}_{p,G-p}^{1-\alpha}$, respectively. These two tests have the same size-adjusted power, because the modification only involves a constant multiplier factor.

For the two-step GMM estimation and related tests, we examine five different procedures. The first three tests use different test statistics but the same critical value $\chi_p^{1-\alpha}/p$. The first test uses the "plain" F-statistic $F_2 := F_{\hat{\Omega}^c(\hat{\theta}_1)}(\hat{\theta}_2^c)$ as defined in (1.10). The second test uses the statistic $[(G-p-q)/G] \cdot F_2$ where $(G-p-q)/G$ is the degree-of-freedom correction factor. The third test uses $\tilde{F}_2 := F_{\hat{\Omega}^c(\hat{\theta}_1)}(\hat{\theta}_2^c)$ as defined in (1.18). Note that

$$\tilde{F}_2 = \frac{(G-p-q)}{G} \cdot \frac{F_2}{1+(q/G)J(\hat{\theta}_2^c)}.$$

Compared to the second test statistic, $\tilde{F}_2$ has the additional J-statistic correction factor $(1+(q/G)J(\hat{\theta}_2^c))^{-1}$. The three tests use increasingly more sophisticated test statistics. Because $[(G-p-q)/G] \to 1$ and $(1+(q/G)J(\hat{\theta}_2^c))^{-1} \to 1$ as $G \to \infty$, both corrections can be regarded as devices for finite sample improvement under the large-G asymptotics.

1.9.3 Results

Balanced Cluster Size

We first consider the case when all clusters have an equal number of individuals and take different values of $G \in \{30, 35, 50, 100\}$ and $L_N \in \{50, 100\}$. The null hypotheses of interests are

$$H_{01} : \beta_{10} = 1$$
$$H_{02} : \beta_{10} = \beta_{20} = 1$$
$$H_{03} : \beta_{10} = \beta_{20} = \beta_{30} = 1$$

with the corresponding number of joint hypotheses $p = 1, 2$ and 3, respectively, and the significance level is 5%. The number of simulation replications is 5000.

Tables 1.2∼1.5 report the empirical size of the first-step and two-step tests for different values of $G \in \{30, 35, 50, 100\}$ and $L_N = \{50, 100\}$. The results indicate that both the first-step and two-step tests based on unmodified statistics F_1 and F_2 suffer from severe size distortions, when the conventional chi-square critical values are used. For example, with $G = 50$, $L_N = 50$, and $p = 3$, the empirical size of the first-step chi-square test (using the full set of IVs, and $m = 24$) is around 43%. This size distortion becomes more severe, as the number of clusters becomes smaller, say, for example when G is between 30 and 35. The empirical size of the first-step F test with $G = 50$ reduces to 36.3% when the F critical value is employed. This finding is consistent with the findings in Bester et al., (2011) and which highlights the improved finite sample performance of the fixed-G approximation in some exactly identified models. Tables 1.2∼1.5 also indicate that the finite sample size distortion of all tests become less severe as the number of moment conditions decreases or the cluster size increases.

For the two-step test that employs the plain two-step statistic F_2 and chi-squared critical value, the empirical size is 63.4% for the above mentioned values of L_N, G, m, and p. In view of the large size distortion, we can conclude that the two-step chi-square test suffers more size distortion than the first-step chi-square test. This relatively large size distortion reflects the additional cost in estimating the weighting matrix, which is not captured by the chi-square approximation. The degrees-of-freedom adjusted F_2 reduces the size distortion by almost one third, but the empirical size of 40.1% is still far away from the nominal size of 5%. This motivates us to implement an additional correction via the J-statistic multiplier coupled with the new critical value $\mathcal{F}^{1-\alpha}_{p,G-p-q}$. Tables 1.2∼1.5 show that using the additional modification and the F critical value significantly alleviates the remaining size distortion. The size distortion in the previous example becomes 13.5% which is much closer to the targeted level 5%. Lastly, we find evidence that the most refined statistic $\tilde{F}_2^{\text{adj}+}$, equipped with the finite sample variance correction, successfully captures the higher order estimation uncertainty and yields more accurate finite sample size. For instance, while the empirical size of the most basic two-step chi-square test is 63.4%, the empirical size of the most refined two-step F test is 5.7%, which is very close to the nominal size of 5%. Figure 1.1 summarizes the outstanding performance of our modified two-step tests with F critical values.

Next we investigate the finite sample power performances of the first-step procedure and the two-step procedures F_2, $\tilde{F}_2$, and $\tilde{F}_2^{\text{adj}+}$. 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 that the parameters are generated from the local null alternatives $\beta_1 = \beta_{10} + c/\sqrt{N}$ for $c \in [0, 15]$, and $d = 2$ and $p = 1$. Figures 1.2∼1.5 report the power curves for the first-step and two-step tests for $G \in \{30, 35, 50, 100\}$. The degree of over-identification q considered here is 10 for the full instrument set, and is 4 for the reduced instrument set. The results first indicate that there is no real difference between power curves of the modified ($\tilde{F}_2$) and unmodified (F_2) two-step tests. In fact, some simulation results not reported here indicate the modified F test can be slightly more powerful as the number of parameters gets larger. Also, the finite sample corrected test $\tilde{F}_2^{\text{adj}+}$ does not lead to a loss of power compared with the uncorrected one $\tilde{F}_2$.

Figures 1.2∼1.5 also indicate that two-step tests are more powerful than first-step tests. The power gain of the two-step procedures becomes more significant as the number of G increases. This is because the two-step estimator becomes more efficient. However, there is a cost in estimating the CCE weighting matrix, the power of first-step procedures might dominate the power of the two-step ones in other scenarios, i.e., when the cost of employing CCE weighting matrix outweighs the benefit of estimating it. Some simulation results not reported here show that the power of the first-step test can be higher than that of a two-step test when the number of parameters d and the number of joint hypotheses p are large.

Lastly, Tables 1.2∼1.5 show that the finite sample size distortion of the (centered) J test and the transformed (uncentered) J test is substantially reduced when we employ F critical values instead of conventional chi-squared critical values.

In sum, our simulation evidence clearly demonstrates the size accuracy of our most refined F test regardless of whether the number of clusters G is small or moderate.

Unbalanced Cluster Size

Although our fixed-G asymptotics is valid as long as the cluster sizes are approximately equal, we remain wary of the effect of the cluster size heterogeneity on the quality of the fixed-G approximation. In this subsection, we turn to simulation designs with heterogeneous cluster sizes.

Each simulated data set consists of $5,000$ observations that are divided into 50 clusters. The sequence of alternative cluster-size designs starts by assigning 120 individuals to each of first 10 clusters and 95 individuals to each of next 40 clusters. In each succeeding cluster-size design, we subtract 10 individuals from the second group of clusters and add them to the first group of clusters. In this manner, we construct a series of four cluster-size designs, in which the proportion of the samples in the first group of clusters grows monotonically from 24% to 48%. The design is similar to Carter, Schnepel and Steigerwald (2013) which investigates the behavior of cluster-robust t-statistic under cluster heterogeneity. Table 1.6 describes the heterogeneous cluster-size designs we consider. All other parameter values are the same as before.

Tables 1.7∼1.8 report the empirical size of the first-step, two-step, and J tests for $q = 20$ and $p = 3$. The results immediately indicate that the two-step tests suffer from severe size distortion when the conventional chi-square critical value is employed. For example, under design II, the empirical size of the "plain" two-step chi-square test is around 60.4% for $G = 50$, $q = 20$, and $p = 3$. The size distortion become more severe when the degree of heterogeneity across cluster-size increases. However, our fixed-G asymptotics still performs very well as they reduce the empirical sizes to 4.3% $\sim$ 7.9%, which are much closer to the nominal size of 5%. Figures 1.6∼1.9 summarize the outstanding performance of our modified two-step F tests, even with unbalanced cluster sizes. The results of J tests are omitted here as they are qualitatively similar to those of the F tests.

1.10 Empirical Application

In this section we employ the proposed procedures to revisit the study of Emran and Hou (2013), which investigates the casual effects of access to domestic and international markets on household consumption in rural China. They use a survey data of 7998 rural households across 19 provinces in China. The survey data comes from Chinese Household Income Project (ICPSR 3012) in 1995.[5]

The regression equation for per capita consumption for household i, C_i, in

[5]The data set is downloadable from the *Review of Economics and Statistics* website.

1995 (yuan) is specified as

$$C_i = \alpha_0 + \alpha_p + \beta_d A_i^d + \beta_s A_i^s + \beta_{ds}(A_i^d \times A_i^s) + X_i' \beta_h + X_i^{v\prime} \beta_v + X_i^{c'} \beta_c + \epsilon_i, \quad (1.42)$$

where A_i^d and A_i^s are the log distances of access to domestic (km) and international markets (km), respectively. X_i is the vector of household characteristics that may affect consumption choice, and X_i^v, X_i^c are village, county level controls, respectively, which capture the heterogeneity in economic environments across different regions, and α_p is the province level fixed effect.

Among the unknown parameters in vector $\theta = (\alpha_0, \alpha_p, \beta_m', \beta_h', \beta_v', \beta_c')'$, our focus of interest is $\beta_m = (\beta_d, \beta_s, \beta_{ds})'$ which measures the casual effect of access to domestic and international markets on household consumption in the rural areas. To identify these parameters, Emran and Hou (2013) employs geographic instrumental variables that capture exogeneous variations in access to markets, e.g., straight-line distances to the nearest navigable river and coastline, along with the topographic and agroclimatic features of the counties.[6] There are 21 instrument variables and 31 control variables, including province dummy variables so that the number of moment conditions m is 52. The number of estimated parameters d is 34, and the degree of over-identification q is 18. Because of the close economic and cultural ties within the same county in rural Chinese areas, the study clusters the data by the county level and estimates the model using 2SLS and two-step GMM with uncentered cluster-robust weighting matrix. The data set consists of 7462 observations divided into 86 clusters where the number of households vary across from a low of 49 to a high of 270. Statistical inferences in Emran and Hou (2013) are conducted using the large-G asymptotics only. We apply our more accurate asymptotics to Emran and Hou's study. The inference methods we use here are described in Tables 1.9 and 1.10 which present the test statistics, the reference dis-

[6]For the detailed description of the control variables and instrument variables, see the appendix in Emran and Hou (2013).

tribution, and the standard error formula (finite sample corrected or not) for each method. Here we view all corrections, including the degree-of-freedom correction, the J correction, and the finite sample variance corrections as corrections to the variance estimator.

Table 1.11 shows the point estimation results, standard error estimates, and associated confidence intervals (CIs) for each of 2SLS and the uncentered and centered two-step GMM estimators. Similar to Emran and Hou (2013), our results show that the better access to domestic and international markets has a substantial positive effect on household consumption, and that the domestic market effect is significantly higher. For the 2SLS method, there are no much differences in confidence interval and standard error between the large-G and fixed-G results. This is well expected because the number of clusters $G = 86$ is large enough so that the large-G and fixed-G approximations are close to each other.

The uncentered two-step GMM estimate of the effect of access to domestic market is $\beta_d = -2722.22$. The reported standard error 400.5 is about 40% smaller than that of 2SLS. However, the plain two-step standard error estimate might be downward biased because the variation of the cluster-robust weighting matrix is not considered. The centered two-step GMM estimator gives a smaller effect of market access $\beta_d = -2670.0$ with the modified standard error of 519.2, which is 25% larger than the plain two-step standard error. However, the modified standard error is still smaller than that based on the 2SLS method. So the two-step estimator still enjoys the benefit of using the cluster-robust weighting matrix. The results for other parameters deliver similar qualitative messages. Table 1.11 also provides the finite sample corrected standard error estimates of two-step estimators that capture the extra variation of feasible CCE, leading to slightly larger standard errors and wider CIs than the uncorrected ones. Overall, our results suggest that the effect of access to markets may be lower than the previous finding after we take into account the randomness of the estimated optimal GMM weighting matrix.

1.11 Conclusion

This paper studies GMM estimation and inference under clustered dependence. To obtain more accurate asymptotic approximations, we utilize an alternative asymptotics under which the sample size of each cluster is growing, but the number of cluster size G is fixed. The paper is comprehensive in that it covers the first-step GMM, the second-step GMM, and continuously-updating GMM estimators. For the two-step GMM estimator, we show that only if centered moment processes are used in constructing the weighting matrix can we obtain asymptotically pivotal Wald statistic and t-statistic. We also find that the centered two-step GMM estimator and CU estimators are all first-order equivalent under the fixed-G asymptotics. With the help of the standard J-statistic, the Wald statistic and t-statistic based on these estimators can be modified to have to standard F and t limiting distributions. A finite sample variance correction is suggested to further improve the performance of the asymptotic F and t tests. The advantages of our procedures are clearly reflected in finite samples as demonstrated by our simulation study and empirical application.

In an overidentified GMM model, the set of moment conditions can be divided into two blocks: the moment conditions that are for identifying unknown parameters, and the rest of ones for improving the efficiency of the GMM estimator. We expect that the spatial dependence between these two blocks of moment conditions is the key information to assess the relative power performance of first-step and two-step tests. Recently, Hwang and Sun (2015a) compares these two types of tests by employing more accurate asymptotic approximations in a time series GMM framework. We leave the extension of this analysis to the spatial setting to future research.

1.12 Acknowledgements

I am sincerely grateful to my advisor Yixiao Sun for his invaluable guidance and support to write this paper. I am also indebt to Graham Elliott, Andres Santos, and Minseong Kim that helped greatly improve this paper. I would also like to thank Jorge Agüero, Eli Berman, Kenneth Couch, Gordon Dahl, James D. Hamilton, Harim Kim, Stepehn L. Ross, Eun-young (Grace) Shim, Gonzalo Valdes, and other seminar participants at UCSD and University of Connecticut for their comments and discussions.

1.13 Figures and Tables

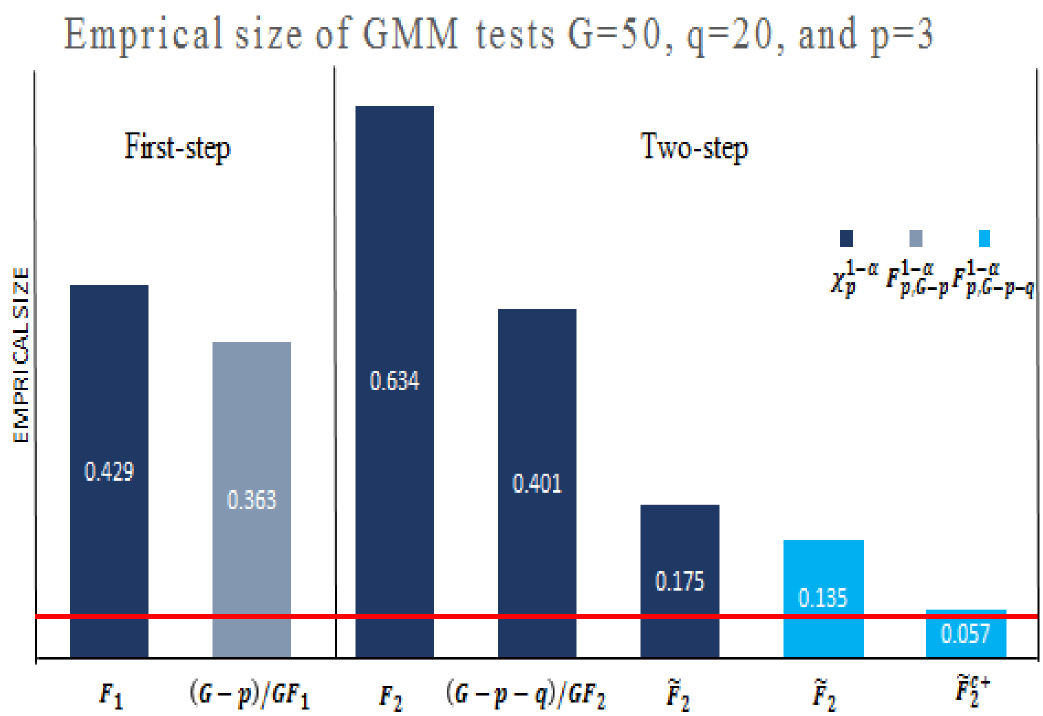

Figure 1.1: Emprical size of the first-step and two-step tests when $G = 50, L_N = 50, m = 24, d = 4$, and $p = 3$ with the nominal size 5% (red line).

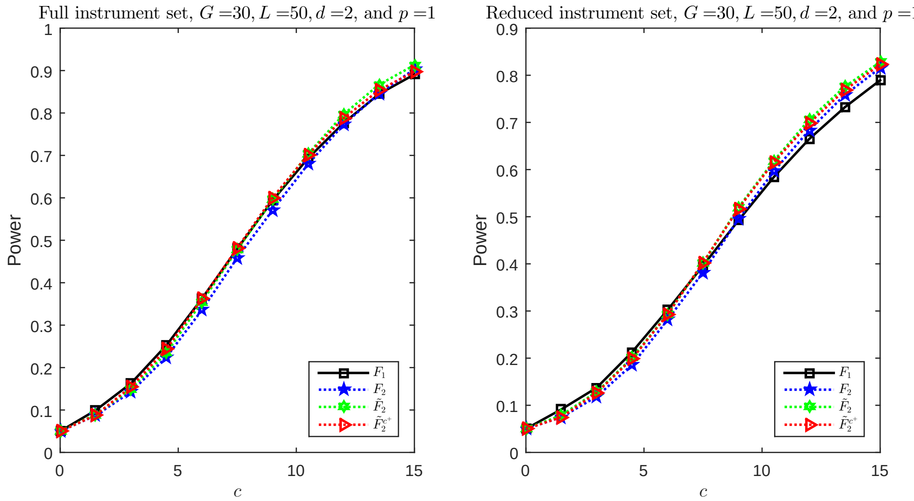

Figure 1.2: Size-adjusted power of the first-step (2SLS) and two-step tests with $G = 30$ and $L_N = 50$.

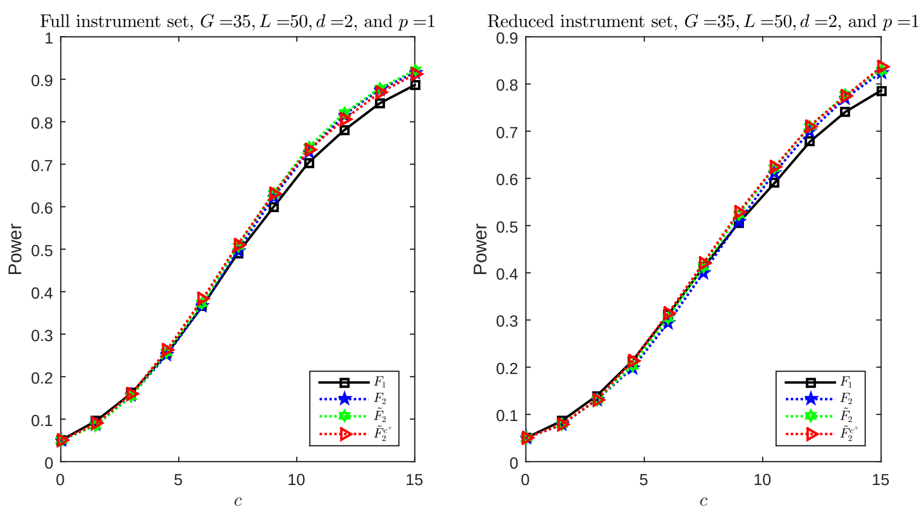

Figure 1.3: Size-adjusted power of the first-step (2SLS) and two-step tests with $G = 35$ and $L_N = 50$.

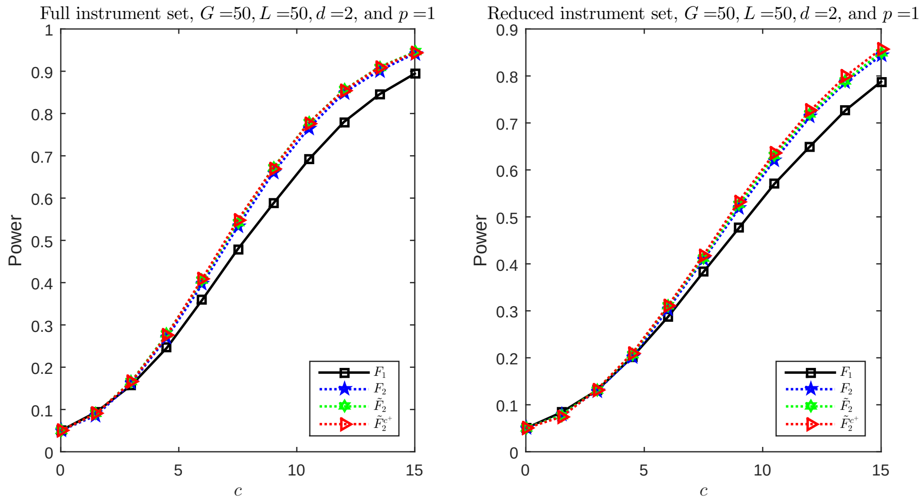

Figure 1.4: Size-adjusted power of the first-step (2SLS) and two-step tests with $G = 50$ and $L_N = 50$.

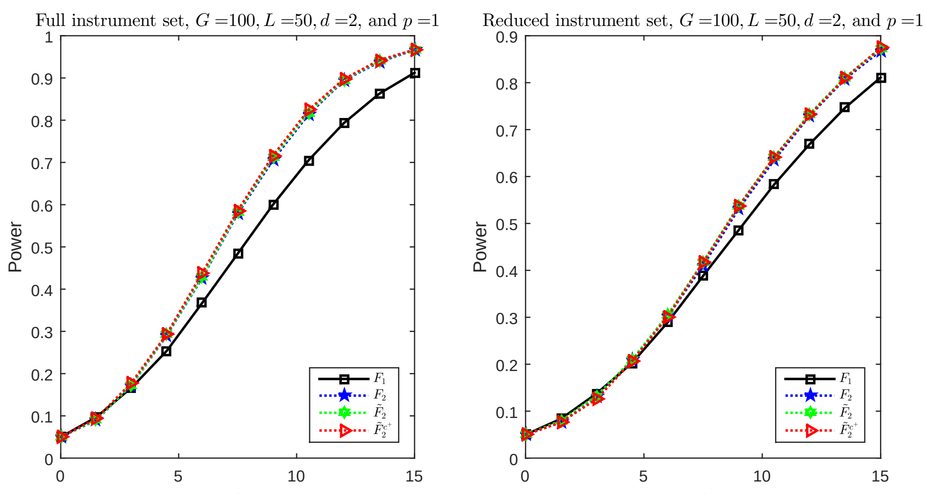

Figure 1.5: Size-adjusted power of the first-step (2SLS) and two-step tests with $G = 100$ and $L_N = 50$.

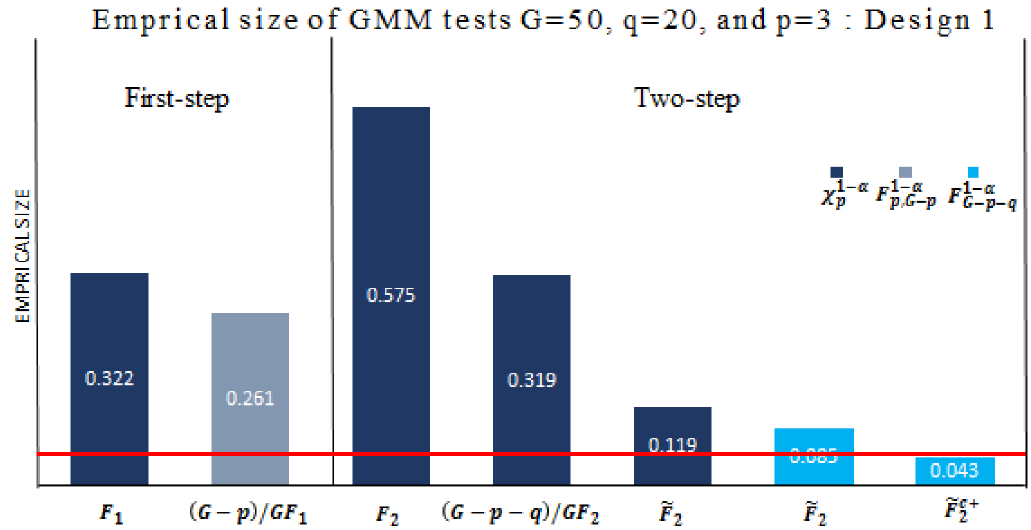

Figure 1.6: Empirical size of first-step and two-step tests based on the centered CCE when there is a heterogeneity in cluster size with the nominal size 5% (red line): Design I with $G = 50, q = 20$, and $p = 3$.

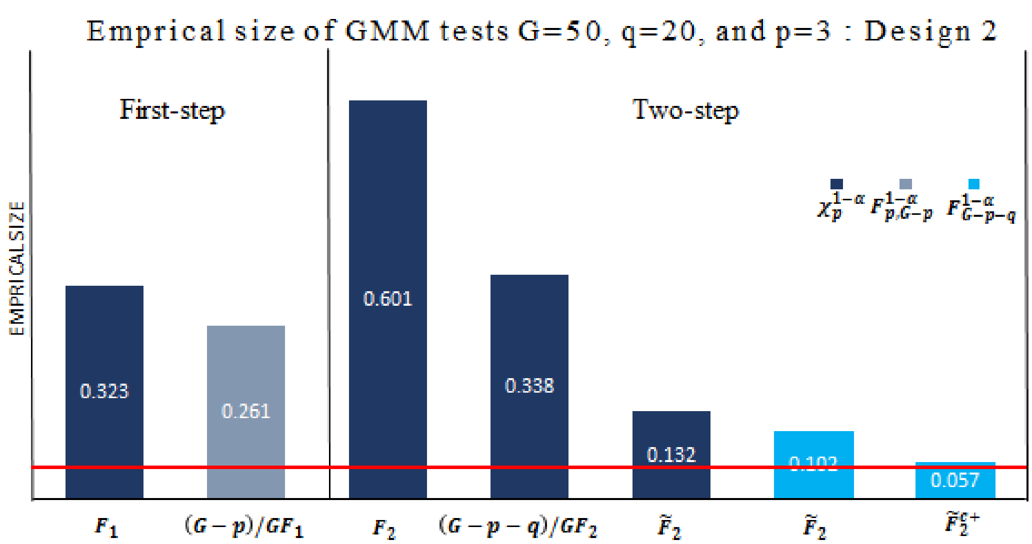

Figure 1.7: Empirical size of first-step and two-step tests based on the centered CCE when there is a heterogeneity in cluster size with the nominal size 5% (red line): Design II with $G = 50, q = 20$, and $p = 3$.

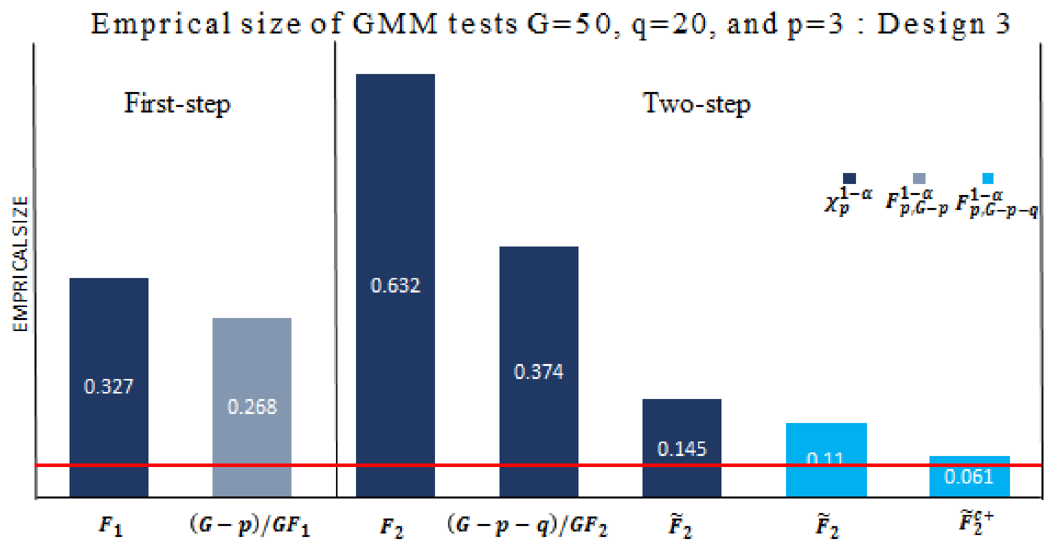

Figure 1.8: Empirical size of first-step and two-step tests based on the centered CCE when there is a heterogeneity in cluster size with the nominal size 5% (red line): Design III with $G = 50, q = 20$, and $p = 3$.

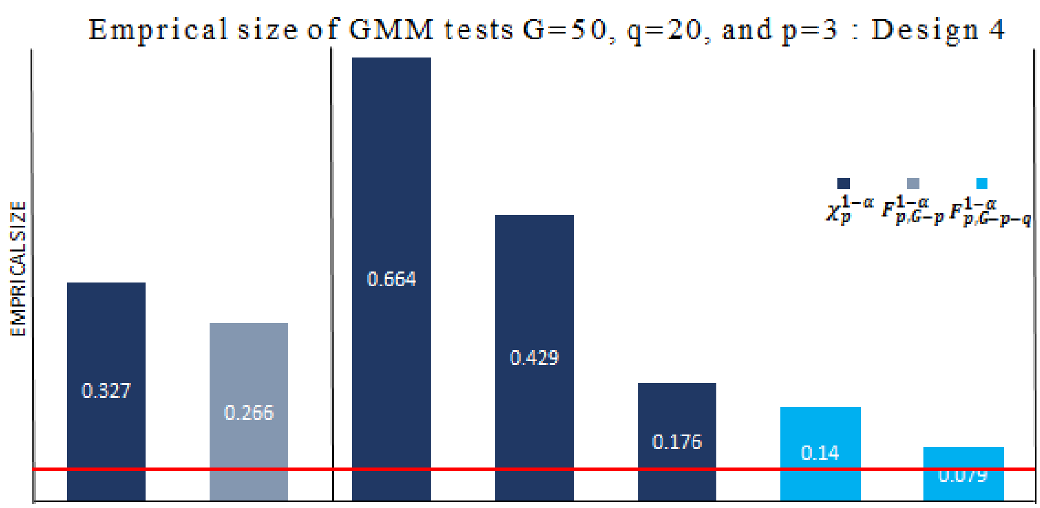

Figure 1.9: Empirical size of first-step and two-step tests based on the centered CCE when there is a heterogeneity in cluster size with the nominal size 5% (red line): Design IV with $G = 50, q = 20$, and $p = 3$.

Table 1.1: Summary of the first-step, two-step tests, and J test

		First-step GMM tests		
statistic	d.f. adj.	$\mathcal{F}^{1-\alpha}_{p,G-p}$	·	·
F_1	–	–	–	–
$\frac{G-p}{G}F_1$	Y	Y	–	–
		Two-step GMM tests		
statistic	d.f. adj.	J-modification	$\mathcal{F}^{1-\alpha}_{p,G-p-q}$	variance correction
F_2	–	–	–	–
$\frac{G-p-q}{G}F_2$	Y	–	–	–
$\tilde{F}_2$	Y	Y	–	–
$\tilde{F}_2$	Y	Y	Y	–
$\tilde{F}^{\text{adj}+}$	Y	Y	Y	Y
		J-tests		
statistic	d.f. adj.	$\mathcal{F}^{1-\alpha}_{q,G-q}$	·	·
J	–	–	–	–
$\frac{G-q}{G}J^c$	Y	Y	–	–

Notes: The first-step tests are based on the first-step GMM estimator $\hat{\theta}_{2SLS}$. They use the associated F-statistic $F_1 = F_1(\hat{\theta}_1)$ with critical value $\chi^{1-\alpha}_p/p$ or $F^{1-\alpha}_{p,G-p}$. The first J test employs the statistic $J(\hat{\theta}_2)$ and critical value $\chi^{1-\alpha}_q$, and the second J test employs the statistic $\frac{G-q}{G}J^c = \frac{G-q}{G}J(\hat{\theta}^c_2)$ and critical value $\mathcal{F}^{1-\alpha}_{q,G-q}$. All two-step tests are based on the centered two-step GMM estimator $\hat{\theta}^c_2$ but use different test statistics and critical values: the unmodified F-statistic $F_2 = F_{\hat{\Omega}^c(\hat{\theta}_1)}(\hat{\theta}^c_2)$, J-statistic and degrees-of-freedom corrected statistic $\tilde{F}_2 = \tilde{F}_{\hat{\Omega}^c(\hat{\theta}_1)}(\hat{\theta}^c_2)$, and J-statistic, degrees-of-freedom, and finite-sample-variance corrected F-statistic $\tilde{F}^{\text{adj}+}_2 = \tilde{F}^{\text{adj}+}_{\hat{\Omega}^c(\hat{\theta}_1)}(\hat{\theta}^c_2)$, coupled with critical value $\chi^{1-\alpha}_p/p$ or $\mathcal{F}^{1-\alpha}_{p,G-p-q}$.

Table 1.2: Empirical size of first-step and two-step tests based on the centered CCE when $L_N = 50$, number of clusters $G = 30$ and 35, number of joint hypothesis $p = 1 \sim 3$ and number of moment conditions $m = 12, 24$.

$G = 30$	Test statistic	critical values	$m = 24$			$m = 12$		
			$p = 1$	$p = 2$	$p = 3$	$p = 1$	$p = 2$	$p = 3$
first-step	F_1	$\chi_p^{1-\alpha}/p$	0.374	0.475	0.528	0.274	0.302	0.332
	$\frac{G-p}{G}F_1$	$\mathcal{F}_{p,G-p}^{1-\alpha}$	0.333	0.389	0.417	0.236	0.232	0.220
two-step	F_2	$\chi_p^{1-\alpha}/p$	0.659	0.857	0.939	0.308	0.415	0.492
	$\frac{G-p-q}{G}F_2$	$\chi_p^{1-\alpha}/p$	0.414	0.562	0.658	0.229	0.276	0.308
	$\tilde{F}_2$	$\chi_p^{1-\alpha}/p$	0.128	0.159	0.192	0.147	0.164	0.182
	$\tilde{F}_2$	$\mathcal{F}_{p,G-p-q}^{1-\alpha}$	0.087	0.082	0.079	0.124	0.128	0.129
	$\tilde{F}_2^{\text{adj}+}$	$\mathcal{F}_{p,G-p-q}^{1-\alpha}$	0.015	0.015	0.012	0.063	0.063	0.058
J-test	J	$\chi_q^{1-\alpha}$	–	0.935	–	–	0.329	–
	$\frac{G-q}{q}\frac{qJ}{G-qJ}$	$\mathcal{F}_{q,G-q}^{1-\alpha}$	–	0.069	–	–	0.086	–
	$\frac{G-q}{G}J^c$	$\mathcal{F}_{q,G-q}^{1-\alpha}$	–	0.129	–	–	0.071	–
$G = 35$	Test Statistic	Critical values	$m = 24$			$m = 12$		
			$p = 1$	$p = 2$	$p = 3$	$p = 1$	$p = 2$	$p = 3$
first-step	F_1	$\chi_p^{1-\alpha}/p$	0.381	0.482	0.538	0.268	0.292	0.316
	$\frac{G-p}{G}F_1$	$\mathcal{F}_{p,G-p}^{1-\alpha}$	0.348	0.407	0.429	0.236	0.236	0.218
two-step	F_2	$\chi_p^{1-\alpha}/p$	0.572	0.756	0.864	0.292	0.357	0.416
	$\frac{G-p-q}{G}F_2$	$\chi_p^{1-\alpha}/p$	0.366	0.483	0.561	0.199	0.248	0.275
	$\tilde{F}_2$	$\chi_p^{1-\alpha}/p$	0.145	0.174	0.195	0.143	0.156	0.159
	$\tilde{F}_2$	$\mathcal{F}_{p,G-p-q}^{1-\alpha}$	0.111	0.119	0.115	0.128	0.126	0.117
	$\tilde{F}_2^{\text{adj}+}$	$\mathcal{F}_{p,G-p-q}^{1-\alpha}$	0.038	0.034	0.033	0.066	0.060	0.057
J-test	J	$\chi_q^{1-\alpha}$	–	0.869	–	–	0.307	–
	$\frac{G-q}{q}\frac{qJ}{G-qJ}$	$\mathcal{F}_{q,G-q}^{1-\alpha}$	–	0.080	–	–	0.091	–
	$\frac{G-q}{Gq}J^c$	$\mathcal{F}_{q,G-q}^{1-\alpha}$	–	0.091	–	–	0.073	–

Notes: See footnote to Table 1.1.

Table 1.3: Empirical size of first-step and two-step tests based on the centered CCE when $L_N = 50$, number of clusters $G = 50$ and 100, number of joint hypothesis $p = 1 \sim 3$ and number of moment conditions $m = 12, 24$.

$G = 50$	Test statistic	critical values	$m = 24$			$m = 12$		
			$p = 1$	$p = 2$	$p = 3$	$p = 1$	$p = 2$	$p = 3$
first-step	F_1	$\chi_p^{1-\alpha}/p$	0.347	0.399	0.429	0.247	0.258	0.273
	$\frac{G-p}{G}F_1$	$\mathcal{F}_{p,G-p}^{1-\alpha}$	0.325	0.356	0.363	0.227	0.222	0.217
two-step	F_2	$\chi_p^{1-\alpha}/p$	0.399	0.538	0.634	0.208	0.255	0.299
	$\frac{G-p-q}{G}F_2$	$\chi_p^{1-\alpha}/p$	0.269	0.344	0.401	0.168	0.192	0.206
	$\tilde{F}_2$	$\chi_p^{1-\alpha}/p$	0.142	0.158	0.175	0.129	0.140	0.139
	$\tilde{F}_2$	$\mathcal{F}_{p,G-p-q}^{1-\alpha}$	0.123	0.129	0.135	0.119	0.116	0.115
	$\tilde{F}_2^{\text{adj}+}$	$\mathcal{F}_{p,G-p-q}^{1-\alpha}$	0.067	0.064	0.057	0.064	0.060	0.058
J-test	J	$\chi_q^{1-\alpha}$	–	0.666	–	–	0.235	–
	$\frac{G-q}{q}\frac{qJ}{G-qJ}$	$\mathcal{F}_{q,G-q}^{1-\alpha}$	–	0.093	–	–	0.098	–
	J^c	$\mathcal{F}_{q,G-q}^{1-\alpha}$	–	0.072	–	–	0.072	–

$G = 100$	Test Statistic	Critical values	$m = 24$			$m = 12$		
			$p = 1$	$p = 2$	$p = 3$	$p = 1$	$p = 2$	$p = 3$
first-step	F_1	$\chi_p^{1-\alpha}/p$	0.266	0.292	0.303	0.195	0.200	0.200
	$\frac{G-p}{G}F_1$	$\mathcal{F}_{p,G-p}^{1-\alpha}$	0.254	0.274	0.274	0.188	0.183	0.172
two-step	F_2	$\chi_p^{1-\alpha}/p$	0.213	0.271	0.305	0.120	0.142	0.156
	$\frac{G-p-q}{G}F_2$	$\chi_p^{1-\alpha}/p$	0.163	0.185	0.202	0.102	0.105	0.110
	$\tilde{F}_2$	$\chi_p^{1-\alpha}/p$	0.115	0.119	0.113	0.086	0.085	0.084
	$\tilde{F}_2$	$\mathcal{F}_{p,G-p-q}^{1-\alpha}$	0.109	0.108	0.099	0.082	0.079	0.076
	$\tilde{F}_2^{\text{adj}+}$	$\mathcal{F}_{p,G-p-q}^{1-\alpha}$	0.078	0.069	0.063	0.053	0.049	0.049
J-test	J	$\chi_q^{1-\alpha}$	–	0.342	–	–	0.156	–
	$\frac{G-q}{q}\frac{J}{G-J}$	$\mathcal{F}_{q,G-q}^{1-\alpha}$	–	0.106	–	–	0.095	–
	J^c	$\mathcal{F}_{q,G-q}^{1-\alpha}$	–	0.069	–	–	0.065	–

Notes: See footnote to Table 1.1.

Table 1.4: Empirical size of first-step and two-step tests based on the centered CCE when $L_N = 50$, number of clusters $G = 30$ and 35, number of joint hypothesis $p = 1 \sim 3$ and number of moment conditions $m = 12, 24$.

$G = 30$	Test Statistic	Critical values	$m = 24$			$m = 12$		
			$p = 1$	$p = 2$	$p = 3$	$p = 1$	$p = 2$	$p = 3$
first-step	F_1	$\chi_p^{1-\alpha}/p$	0.344	0.402	0.440	0.249	0.117	0.289
	$\frac{G-p}{G} F_1$	$\mathcal{F}_{p,G-p}^{1-\alpha}$	0.300	0.328	0.333	0.217	0.208	0.195
two-step	F_2	$\chi_p^{1-\alpha}/p$	0.610	0.798	0.883	0.275	0.347	0.436
	$\frac{G-p-q}{G} F_2$	$\chi_p^{1-\alpha}/p$	0.371	0.517	0.612	0.191	0.220	0.246
	$\tilde{F}_2$	$\chi_p^{1-\alpha}/p$	0.101	0.130	0.164	0.116	0.117	0.124
	$\tilde{F}_2$	$\mathcal{F}_{p,G-p-q}^{1-\alpha}$	0.061	0.064	0.059	0.095	0.088	0.124
	$\tilde{F}_2^{\text{adj}+}$	$\mathcal{F}_{p,G-p-q}^{1-\alpha}$	0.017	0.015	0.013	0.045	0.043	0.040
J-test	J	$\chi_q^{1-\alpha}$	–	0.930	–	–	0.336	–
	$\frac{G-q}{q} \frac{qJ}{G-qJ}$	$\mathcal{F}_{q,G-q}^{1-\alpha}$	–	0.073	–	–	0.098	–
	$\frac{G-q}{Gq} J^c$	$\mathcal{F}_{q,G-q}^{1-\alpha}$	–	.114	–	–	0.077	–
	Test		$m = 24$			$m = 12$		
$G = 35$	Statistic	Ref.dist.	$p = 1$	$p = 2$	$p = 3$	$p = 1$	$p = 2$	$p = 3$
first-step	F_1	$\chi_p^{1-\alpha}/p$	0.321	0.373	0.393	0.237	0.257	0.269
	$\frac{G-p}{G} F_1$	$\mathcal{F}_{p,G-p}^{1-\alpha}$	0.292	0.303	0.300	0.213	0.207	0.187
two-step	F_2	$\chi_p^{1-\alpha}$	0.522	0.697	0.826	0.224	0.301	0.362
	$\frac{G-p-q}{G} F_2$	$\chi_p^{1-\alpha}/p$	0.306	0.411	0.484	0.162	0.196	0.224
	$\tilde{F}_2$	$\chi_p^{1-\alpha}$	0.105	0.125	0.136	0.107	0.115	0.118
	$\tilde{F}_2$	$\mathcal{F}_{p,G-p-q}^{1-\alpha}$	0.079	0.079	0.076	0.090	0.090	0.087
	$\tilde{F}_2^{\text{adj}+}$	$\mathcal{F}_{p,G-p-q}^{1-\alpha}$	0.038	0.032	0.026	0.049	0.049	0.044
J-test	J	$\chi_q^{1-\alpha}$	–	0.853	–	–	0.288	–
	$\frac{G-q}{q} \frac{J}{G-J}$	$\mathcal{F}_{q,G-q}^{1-\alpha}$	–	0.080	–	–	0.098	–
	$\frac{G-q}{Gq} J^c$	$\mathcal{F}_{q,G-q}^{1-\alpha}$	–	0.078	–	–	0.075	–

Notes: See footnote to Table 1.1.

Table 1.5: Empirical size of first-step and two-step tests based on the centered CCE when $L_N = 50$, number of clusters $G = 50$ and 100, number of joint hypothesis $p = 1 \sim 3$ and number of moment conditions $m = 12, 24$.

$G = 50$	Test Statistic	Critical values	$m = 24$			$m = 12$		
			$p = 1$	$p = 2$	$p = 3$	$p = 1$	$p = 2$	$p = 3$
first-step	F_1	$\chi_p^{1-\alpha}/p$	0.269	0.305	0.324	0.206	0.214	0.222
	$\frac{G-p}{G} F_1$	$\mathcal{F}_{p,G-p}^{1-\alpha}$	0.250	0.260	0.257	0.189	0.180	0.175
two-step	F_2	$\chi_p^{1-\alpha}/p$	0.341	0.469	0.575	0.176	0.211	0.250
	$\frac{G-p-q}{G} F_2$	$\chi_p^{1-\alpha}/p$	0.209	0.268	0.324	0.129	0.146	0.164
	$\tilde{F}_2$	$\chi_p^{1-\alpha}/p$	0.099	0.115	0.125	0.094	0.098	0.105
	$\tilde{F}_2$	$\mathcal{F}_{p,G-p-q}^{1-\alpha}$	0.088	0.091	0.093	0.086	0.083	0.084
	$\tilde{F}_2^{\text{adj}+}$	$\mathcal{F}_{p,G-p-q}^{1-\alpha}$	0.059	0.056	0.052	0.057	0.052	0.055
J-test	J	$\chi_q^{1-\alpha}$	–	0.629	–	–	0.221	–
	$\frac{G-q}{q} \frac{qJ}{G-qJ}$	$\mathcal{F}_{q,G-q}^{1-\alpha}$	–	0.087	–	–	0.094	–
	J^c	$\mathcal{F}_{q,G-q}^{1-\alpha}$	–	0.064	–	–	0.063	–
$G = 100$	Test Statistic	Critical values	$m = 24$			$m = 12$		
			$p = 1$	$p = 2$	$p = 3$	$p = 1$	$p = 2$	$p = 3$
first-step	F_1	$\chi_p^{1-\alpha}/p$	0.185	0.203	0.209	0.149	0.147	0.152
	$\frac{G-p}{G} F_1$	$\mathcal{F}_{p,G-p}^{1-\alpha}$	0.176	0.185	0.182	0.143	0.136	0.128
two-step	F_2	$\chi_p^{1-\alpha}/p$	0.168	0.225	0.258	0.100	0.120	0.129
	$\frac{G-p-q}{G} F_2$	$\chi_p^{1-\alpha}/p$	0.125	0.149	0.161	0.082	0.089	0.093
	$\tilde{F}_2$	$\chi_p^{1-\alpha}/p$	0.085	0.091	0.094	0.069	0.073	0.071
	$\tilde{F}_2$	$\mathcal{F}_{p,G-p-q}^{1-\alpha}$	0.081	0.082	0.083	0.065	0.067	0.064
	$\tilde{F}_2^{\text{adj}+}$	$\mathcal{F}_{p,G-p-q}^{1-\alpha}$	0.067	0.065	0.064	0.056	0.055	0.051
J-test	J	$\chi_q^{1-\alpha}$	–	0.293	–	–	0.135	–
	$\frac{G-q}{q} \frac{J}{G-J}$	$\mathcal{F}_{q,G-q}^{1-\alpha}$	–	0.081	–	–	0.081	–
	J^c	$\mathcal{F}_{q,G-q}^{1-\alpha}$	–	0.055	–	–	0.053	–

Notes: See footnote to Table 1.1.

Table 1.6: Design of heterogeneity in cluster size

$G = 50$	$L_{N_1} = ... = L_{N_{10}}$	$L_{N_{11}} = ... = L_{N_{50}}$	N
Design I	120	95	5000
Design II	160	85	5000
Design III	200	75	5000
Design IV	240	65	5000

Table 1.7: Empirical size of first-step and two-step tests based on the centered CCE when there is a heterogeneity in cluster size: Design I∼II

Design I								
	test statistic	critical values	$m = 24$			$m = 12$		
			$p = 1$	$p = 2$	$p = 3$	$p = 1$	$p = 2$	$p = 3$
first-step	F_1	$\chi_p^{1-\alpha}/p$	0.265	0.303	0.322	0.205	0.201	0.217
	$\frac{G-p}{G} F_1$	$\mathcal{F}_{p,G-p}^{1-\alpha}$	0.246	0.264	0.261	0.186	0.168	0.165
two-step	F_2	$\chi_p^{1-\alpha}/p$	0.335	0.472	0.575	0.153	0.194	0.242
	$\frac{G-p-q}{G} F_2$	$\chi_p^{1-\alpha}/p$	0.209	0.265	0.319	0.118	0.132	0.153
	$\tilde{F}_2$	$\chi_p^{1-\alpha}/p$	0.095	0.104	0.119	0.086	0.087	0.091
	$\tilde{F}_2$	$\mathcal{F}_{p,G-p-q}^{1-\alpha}$	0.082	0.083	0.085	0.074	0.071	0.072
	$\tilde{F}_2^{\text{adj+}}$	$\mathcal{F}_{p,G-p-q}^{1-\alpha}$	0.054	0.048	0.043	0.046	0.046	0.047
J-test	J	$\chi_q^{1-\alpha}$	–	0.632	–	–	0.228	–
	$\frac{G-q}{q} \frac{qJ}{G-qJ}$	$\mathcal{F}_{q,G-q}^{1-\alpha}$	–	0.089	–	–	0.100	–
	$\frac{G-q}{Gq} J^c$	$\mathcal{F}_{q,G-q}^{1-\alpha}$	–	0.067	–	–	0.068	–
Design II								
			$m = 24$			$m = 12$		
$G = 35$	test statistic	critical values	$p = 1$	$p = 2$	$p = 3$	$p = 1$	$p = 2$	$p = 3$
first-step	F_1	$\chi_p^{1-\alpha}/p$	0.268	0.304	0.323	0.206	0.213	0.220
	$\frac{G-p}{G} F_1$	$\mathcal{F}_{p,G-p}^{1-\alpha}$	0.250	0.265	0.261	0.193	0.182	0.171
two-step	F_2	$\chi_p^{1-\alpha}/p$	0.361	0.501	0.601	0.160	0.209	0.254
	$\frac{G-p-q}{G} F_2$	$\chi_p^{1-\alpha}/p$	0.231	0.291	0.338	0.124	0.141	0.170
	$\tilde{F}_2$	$\chi_p^{1-\alpha}/p$	0.112	0.120	0.132	0.088	0.092	0.100
	$\tilde{F}_2$	$\mathcal{F}_{p,G-p-q}^{1-\alpha}$	0.097	0.098	0.102	0.079	0.076	0.082
	$\tilde{F}_2^{\text{adj+}}$	$\mathcal{F}_{p,G-p-q}^{1-\alpha}$	0.063	0.060	0.057	0.053	0.052	0.056
J-test	J	$\chi_q^{1-\alpha}$	–	0.638	–	–	0.214	–
	$\frac{G-q}{q} \frac{qJ}{G-qJ}$	$\mathcal{F}_{q,G-q}^{1-\alpha}$	–	0.083	–	–	0.094	–
	$\frac{G-q}{Gq} J^c$	$\mathcal{F}_{q,G-q}^{1-\alpha}$	–	0.081	–	–	0.072	–

Notes: See footnote to Table 1.1.

Table 1.8: Empirical size of first-step and two-step tests based on the centered CCE when there is a heterogeneity in cluster size: Design III∼IV

Design III								
	test statistic	critical values	$m = 24$			$m = 12$		
			$p = 1$	$p = 2$	$p = 3$	$p = 1$	$p = 2$	$p = 3$
first-step	F_1	$\chi_p^{1-\alpha}/p$	0.276	0.315	0.327	0.203	0.215	0.228
	$\frac{G-p}{G} F_1$	$\mathcal{F}_{p,G-p}^{1-\alpha}$	0.254	0.276	0.268	0.186	0.186	0.178
two-step	F_2	$\chi_p^{1-\alpha}/p$	0.378	0.532	0.632	0.168	0.222	0.274
	$\frac{G-p-q}{G} F_2$	$\chi_p^{1-\alpha}/p$	0.244	0.323	0.374	0.134	0.164	0.184
	$\tilde{F}_2$	$\chi_p^{1-\alpha}/p$	0.117	0.132	0.145	0.097	0.109	0.117
	$\tilde{F}_2$	$\mathcal{F}_{p,G-p-q}^{1-\alpha}$	0.101	0.107	0.110	0.088	0.091	0.092
	$\tilde{F}_2^{\text{adj+}}$	$\mathcal{F}_{p,G-p-q}^{1-\alpha}$	0.04	0.057	0.061	0.061	0.062	0.060
J-test	J	$\chi_q^{1-\alpha}$	–	0.631	–	–	0.213	–
	$\frac{G-q}{q} \frac{qJ}{G-qJ}$	$\mathcal{F}_{q,G-q}^{1-\alpha}$	–	0.076	–	–	0.089	–
	$\frac{G-q}{Gq} J^c$	$\mathcal{F}_{q,G-q}^{1-\alpha}$	–	0.102	–	–	0.079	–
Design IV								
			$m = 24$			$m = 12$		
$G = 35$	test statistic	critical values	$p = 1$	$p = 2$	$p = 3$	$p = 1$	$p = 2$	$p = 3$
first-step	F_1	$\chi_p^{1-\alpha}/p$	0.255	0.306	0.327	0.200	0.214	0.226
	$\frac{G-p}{G} F_1$	$\mathcal{F}_{p,G-p}^{1-\alpha}$	0.232	0.265	0.266	0.180	0.179	0.171
two-step	F_2	$\chi_p^{1-\alpha}/p$	0.397	0.555	0.664	0.185	0.243	0.304
	$\frac{G-p-q}{G} F_2$	$\chi_p^{1-\alpha}/p$	0.264	0.363	0.429	0.139	0.174	0.205
	$\tilde{F}_2$	$\chi_p^{1-\alpha}/p$	0.131	0.155	0.176	0.104	0.120	0.136
	$\tilde{F}_2$	$\mathcal{F}_{p,G-p-q}^{1-\alpha}$	0.115	0.125	0.140	0.094	0.102	0.110
	$\tilde{F}_2^{\text{adj+}}$	$\mathcal{F}_{p,G-p-q}^{1-\alpha}$	0.075	0.077	0.079	0.062	0.070	0.077
J-test	J	$\chi_q^{1-\alpha}$	–	0.621	–	–	0.206	–
	$\frac{G-q}{q} \frac{qJ}{G-qJ}$	$\mathcal{F}_{q,G-q}^{1-\alpha}$	–	0.070	–	–	0.080	–
	$\frac{G-q}{Gq} J^c$	$\mathcal{F}_{q,G-q}^{1-\alpha}$	–	0.151	–	–	0.092	–

Notes: See footnote to Table 1.1.

Table 1.9: Summary of the difference between the conventional large-G asympototics and alternative fixed-G asympotics for the first-step (2SLS) and two-step GMM methods.

2SLS							
Asymptotics	Weight	Test statistic			Reference distribution		
·		Wald	t	·	Wald	t	·
large-G	$Z'HZ/N$	$F(\hat{\theta}_1)$	$t(\hat{\theta}_1)$	·	χ^2_p/p	$N(0,1)$	·
fixed-G	$Z'HZ/N$	$\frac{G-p}{G}F(\hat{\theta}_1)$	$\sqrt{\frac{G-1}{G}}t(\hat{\theta}_1)$	·	$F_{p,G-p}$	t_{G-1}	·
Two-step GMM							
Asymptotics	Weight	Test statistic			Reference distribution		
·		Wald	t	J	Wald	t	J
large-G	$\hat{\Omega}$	$F(\hat{\theta}_2)$	$t(\hat{\theta}_2)$	$J(\hat{\theta}_2)$	χ^2_p/p	$N(0,1)$	χ^2_q
fixed-G	$\hat{\Omega}^c$	$\tilde{F}(\hat{\theta}^c_2)$	$\tilde{t}(\hat{\theta}^c_2)$	$\frac{G-q}{G}J(\hat{\theta}^c_2)$	$F_{p,G-p-q}$	t_{G-1-q}	$F_{q,G-q}$
	$\hat{\Omega}^c$	$\tilde{F}^{\text{adj+}}(\hat{\theta}^c_2)$	$\tilde{t}^{\text{adj+}}(\hat{\theta}^c_2)$	·	$F_{p,G-p-q}$	t_{G-1-q}	·

Table 1.10: Summary of standard error formula when $p = 1$

	2SLS	
	large-G	fixed-G
standard error	$\left[R\widehat{var}(\hat{\theta}_1)R'\right]^{1/2}$	$\left[\frac{G-1}{G}R\widehat{var}_{\Omega(\hat{\theta}_1)}(\hat{\theta}_1)R'\right]^{1/2}$
	Two-step GMM	
	large-G	fixed-G
standard error	$\left[R\widehat{var}(\hat{\theta}_1)R'\right]^{1/2}$	$\left[\frac{G}{G-1-q}R\widehat{var}_{\hat{\Omega}(\hat{\theta}_1)}(\hat{\theta}_2^c)R'\left(1+(q/G)J(\hat{\theta}_2^c)\right)\right]^{1/2}$
corrected standard error	$\cdot$	$\left[\frac{G}{G-1-q}R\widehat{var}^{\text{adj+}}_{\hat{\Omega}(\hat{\theta}_1)}(\hat{\theta})R'\left(1+(q/G)J(\hat{\theta}_2^c)\right)\right]^{1/2}$

Table 1.11: Results for Emran and Hou (2013) data

	2SLS	
Variables	Large-G Asymptotics	Fixed-G Asymptotics
Domestic market (A_i^d)	-2713.2 (712.1)	-2713.2 (716.8)
	$[-4109.9, -1316.4]$	$[-4138.0, -1288.0]$
International market (A_i^s)	-1993.5 (514.8)	-1993.5 (517.9)
	$[-3002.5, -984.4]$	$[-3023.10, -963.8]$
Interaction ($A_i^d \times A_i^s$)	345.8 (105.0)	345.8 (105.6)
	$[140.0, 551.7]$	$[135.8, 555.9]$
$H_0 : \beta_d = \beta_s$	-2.3218 (2.02%)	-2.771 (2.26%)
	Two-step GMM	
Variables	Large-G Asymptotics	Fixed-G Asymptotics
Domestic market (A_i^d)	-2722.8 (400.5)	-2670.0 (519.2)
		(520.7)
	$[-3507.7, -1937.9]$	$[-3706.2, -1633.8]$
		$[-3709.2, -1630.7]$
International market (A_i^s)	-2000.2 (344.3)	-1981.3 (446.4)
		(447.7)
	$[-2675.0, -1325.5]$	$[-2872.3, -1090.3]$
		$[-2874.9, -1087.7]$
Interaction ($A_i^d \times A_i^s$)	362.7 (68.7)	364.1 (89.1)
		(89.4)
	$[228.0, 497.3]$	$[186.2, 541.9]$
		$[187.5, 542.4]$
$H_0 : \beta_d = \beta_s$	-5.239 (0%)	-3.3318 (0%)
		-3.3217 (0%)
J-statistic ($q = 18$)	1.1708 (99.8%)	0.3096 (45.83%)

Notes: Standard errors for 2SLS and the weighting matrix for (centered) two-step GMM estimators are clustered at the county level. Numbers in parentheses are standard errors and intervals are 95% confidence intervals. For hypothesis testing, the numbers in parentheses are P-values.

1.14 Appendix of Proofs

Proof of Proposition 1. **Part (a).** For each $g = 1, ..., G$,

$$\frac{1}{\sqrt{L}}\sum_{i=1}^{L} f_i^g(\hat{\theta}_1) = \frac{1}{\sqrt{L}}\sum_{i=1}^{L}\left\{ f_i^g(\theta_0) + \frac{\partial f_i^g(\tilde{\theta}^*)}{\partial\theta'}\left[\hat{\theta}_1 - \theta_0\right]\right\},$$

where $\tilde{\theta}^*$ is between $\hat{\theta}_1$ and θ_0. Here, $\tilde{\theta}^*$ may be different for different rows of $\partial f_k^g(\tilde{\theta}^*)/\partial\theta'$. For notational simplicity, we do not make this explicit. By Assumptions 2 and 5, we have

$$\begin{aligned}
&\frac{1}{\sqrt{L}}\sum_{i=1}^{L} f_i^g(\hat{\theta}_1) = \frac{1}{\sqrt{L}}\sum_{i=1}^{L} f_i^g(\theta_0) \\
&-\frac{1}{L}\sum_{i=1}^{L}\frac{\partial f_i^g(\tilde{\theta}^*)}{\partial\theta'}(\Gamma'W^{-1}\Gamma)^{-1}\Gamma'W^{-1}\frac{1}{G}\sum_{h=1}^{G}\left(\frac{1}{\sqrt{L}}\sum_{j=1}^{L} f_j^h(\theta_0)\right) + o_p(1) \\
&= \frac{1}{\sqrt{L}}\sum_{i=1}^{L} f_i^g(\theta_0) - \Gamma_g(\Gamma'W^{-1}\Gamma)^{-1}\Gamma'W^{-1}\frac{1}{G}\sum_{h=1}^{G}\left(\frac{1}{\sqrt{L}}\sum_{i=1}^{L} f_i^h(\theta_0)\right) + o_p(1)\,.
\end{aligned} \tag{1.43}$$

Using Assumptions 4–6, we then have

$$\begin{aligned}
&\frac{1}{\sqrt{L}}\sum_{i=1}^{L} f_i^g(\hat{\theta}_1) \\
&\overset{d}{\to} \Lambda_g B_{m,g} - \Gamma_g(\Gamma'W^{-1}\Gamma)^{-1}\Gamma'W^{-1}\Lambda\bar{B}_m \\
&= \Lambda B_{m,g} - \Gamma(\Gamma'W^{-1}\Gamma)^{-1}\Gamma'W^{-1}\Lambda\bar{B}_m
\end{aligned}$$

where $\bar{B}_m := G^{-1}\sum_{g=1}^{G} B_{m,g}$. It follows that

$$\begin{aligned}
&\hat{\Gamma}(\hat{\theta}_1)'W_N^{-1}\frac{1}{\sqrt{L}}\sum_{i=1}^{L} f_i^g(\hat{\theta}_1) \\
&\overset{d}{\to} \Gamma'W^{-1}\left[\Lambda B_{m,g} - \Gamma(\Gamma'W^{-1}\Gamma)^{-1}\Gamma'W^{-1}\Lambda\bar{B}_m\right] \\
&= \Gamma'W^{-1}\Lambda B_{m,g} - \Gamma'W^{-1}\Lambda\bar{B}_m = \Gamma'W^{-1}\Lambda\left(B_{m,g} - \bar{B}_m\right).
\end{aligned}$$

So, the scaled CCE matrix converges in distribution to a random matrix:

$$
\begin{aligned}
&\hat{\Gamma}(\hat{\theta}_1)'W_N^{-1}\hat{\Omega}(\hat{\theta}_1)W_N^{-1}\hat{\Gamma}(\hat{\theta}_1) \\
&= \frac{1}{G}\sum_{g=1}^{G}\left\{\left(\hat{\Gamma}(\hat{\theta}_1)'W_N^{-1}\frac{1}{\sqrt{L}}\sum_{i=1}^{L} f_i^g(\hat{\theta}_1)\right)\left(\frac{1}{\sqrt{L}}\sum_{j=1}^{L} f_j^g(\hat{\theta}_1)W_N^{-1}\hat{\Gamma}(\hat{\theta}_1)\right)'\right\} \\
&\xrightarrow{d} \Gamma'W^{-1}\Lambda\left\{\frac{1}{G}\sum_{g=1}^{G}\left(B_{m,g}-\bar{B}_m\right)\left(B_{m,g}-\bar{B}_m\right)'\right\}\left(\Gamma'W^{-1}\Lambda\right)'.
\end{aligned}
$$

Therefore,

$$
\begin{aligned}
&NR\widehat{var}(\hat{\theta}_1)R' \\
&= \left[\hat{\Gamma}(\hat{\theta}_1)'W_N^{-1}\hat{\Gamma}(\hat{\theta}_1)\right]^{-1}\left[\hat{\Gamma}(\hat{\theta}_1)'W_N^{-1}\hat{\Omega}(\hat{\theta}_1)W_N^{-1}\hat{\Gamma}(\hat{\theta}_1)\right]\left[\hat{\Gamma}(\hat{\theta}_1)'W_N^{-1}\hat{\Gamma}(\hat{\theta}_1)\right]^{-1} \\
&= R\left[\Gamma'W^{-1}\Gamma\right]^{-1}\Gamma'W^{-1}\Lambda\left\{\frac{1}{G}\sum_{g=1}^{G}\left(B_{m,g}-\bar{B}_m\right)\left(B_{m,g}-\bar{B}_m\right)'\right\} \\
&\cdot \Lambda W^{-1}\Gamma\left[\Gamma'W^{-1}\Gamma\right]^{-1}R' \\
&= \tilde{R}\left\{\frac{1}{G}\sum_{g=1}^{G}\left(B_{m,g}-\bar{B}_m\right)\left(B_{m,g}-\bar{B}_m\right)'\right\}\tilde{R}'
\end{aligned}
$$

where $\tilde{R} := R\left[\Gamma'W^{-1}\Gamma\right]^{-1}\Gamma'W^{-1}\Lambda$. Also, it follows by Assumption 4 that

$$
\begin{aligned}
\sqrt{N}(R\hat{\theta}_1 - r) &= -R(\Gamma'W^{-1}\Gamma)^{-1}\Gamma'W^{-1}\sqrt{N}g_N(\theta_0) + o_p(1) \\
&= -R(\Gamma'W^{-1}\Gamma)^{-1}\Gamma'W^{-1}\frac{1}{\sqrt{G}}\sum_{g=1}^{G}\left(\frac{1}{\sqrt{L}}\sum_{i=1}^{L} f_i^g(\theta_0)\right) + o_p(1) \\
&\xrightarrow{d} -\tilde{R}\frac{1}{\sqrt{G}}\sum_{g=1}^{G} B_{m,g} = -\tilde{R}\sqrt{G}\bar{B}_m.
\end{aligned}
$$

Combining the results so far yields:

$$
\begin{aligned}
F(\hat{\theta}_1) &\xrightarrow{d} \left(\tilde{R}\sqrt{G}\bar{B}_m\right)'\left\{\tilde{R}\frac{1}{G}\sum_{g=1}^{G}\left(B_{m,g}-\bar{B}_m\right)\left(B_{m,g}-\bar{B}_m\right)'\tilde{R}'\right\}^{-1}\tilde{R}\sqrt{G}\bar{B}_m/p \\
&= \mathbb{F}_{1\infty}.
\end{aligned}
$$

Define the $p\times p$ matrix $\tilde{\Lambda}$ such that $\tilde{\Lambda}\tilde{\Lambda}'=\tilde{R}\tilde{R}'$. Then we have the following distributional equivalence

$$\left[\ \tilde{R}\sqrt{G}\bar{B}_m,\quad \tilde{R}\frac{1}{G}\sum_{g=1}^{G}\left(B_{m,g}-\bar{B}_m\right)\left(B_{m,g}-\bar{B}_m\right)'\tilde{R}'\ \right]\overset{d}{=}\left[\ \sqrt{G}\tilde{\Lambda}\bar{B}_p,\quad \tilde{\Lambda}\mathbb{S}_{pp}\tilde{\Lambda}'\ \right].$$

Using this, we get

$$\mathbb{F}_{1\infty}\overset{d}{=}G\bar{B}_p'\mathbb{S}^{-1}\bar{B}_p/p$$

as desired for Part (a). Part (b) can be similarly proved. ■

Proof of Proposition 6. Parts (a), (b) and (c). All three estimators can be represented in the following form

$$-(\Gamma'M^{-1}\Gamma)^{-1}\Gamma'M^{-1}\Lambda\sqrt{G}\bar{B}_m+o_p(1)$$

for some weighing matrix M which may be random. Let

$$M_\Lambda=\Lambda^{-1}M\left(\Lambda'\right)^{-1}\ \text{ and }\Gamma_\Lambda=\Lambda^{-1}\Gamma.$$

Then

$$-(\Gamma'M^{-1}\Gamma)^{-1}\Gamma'M^{-1}\Lambda\sqrt{G}\bar{B}_m=-(\Gamma_\Lambda M_\Lambda^{-1}\Gamma_\Lambda)^{-1}\Gamma_\Lambda M_\Lambda^{-1}\sqrt{G}\bar{B}_m,$$

Let $U\Sigma V'$ be a singular value decomposition (SVD) of Γ_Λ. By construction, $U'U=UU'=I_m$, $V'V=V'V=I_d$, and

$$\Sigma=\left[\begin{array}{c}A_{d\times d}\\ O_{q\times d}\end{array}\right],$$

where A is a diagonal matrix. Denoting

$$\tilde{M}=U'M_\Lambda U=\begin{pmatrix}\tilde{M}_{11} & \tilde{M}_{12}\\ \tilde{M}_{21} & \tilde{M}_{22}\end{pmatrix},$$

we have

$$
\begin{aligned}
&(\Gamma_\Lambda M_\Lambda^{-1}\Gamma_\Lambda)^{-1}\Gamma_\Lambda M_\Lambda^{-1} \\
&= \left[V\Sigma' U' M_\Lambda^{-1} U \Sigma V'\right]^{-1} V\Sigma' U' M_\Lambda^{-1} \\
&= \left[V\Sigma' \left(U' M_\Lambda U\right)^{-1} \Sigma V'\right]^{-1} V\Sigma' \left(U' M_\Lambda U\right)^{-1} U' \\
&= V\left(A'\tilde{M}^{11}A\right)^{-1} \begin{pmatrix} A'_{d\times d}, & O'_{q\times d} \end{pmatrix} \begin{pmatrix} \tilde{M}^{11} & \tilde{M}^{12} \\ \tilde{M}^{21} & \tilde{M}^{22} \end{pmatrix} U' \\
&= VA^{-1}(\tilde{M}^{11})^{-1} \begin{pmatrix} I_d, & O_{d\times q} \end{pmatrix} \begin{pmatrix} \tilde{M}^{11} & \tilde{M}^{12} \\ \tilde{M}^{21} & \tilde{M}^{22} \end{pmatrix} U' \\
&= VA^{-1}(\tilde{M}^{11})^{-1} \begin{pmatrix} \tilde{M}^{11}, & \tilde{M}^{12} \end{pmatrix} U' = VA^{-1} \begin{pmatrix} I_d, & (\tilde{M}^{11})^{-1}\tilde{M}^{12} \end{pmatrix} U' \\
&= VA^{-1} \begin{pmatrix} I_d, & -\tilde{M}_{12}\tilde{M}_{22}^{-1} \end{pmatrix} U'.
\end{aligned}
$$

So

$$
-(\Gamma_\Lambda M_\Lambda^{-1}\Gamma_\Lambda)^{-1}\Gamma_\Lambda M_\Lambda^{-1}\sqrt{G}\bar{B}_m = -VA^{-1} \begin{pmatrix} I_d, & -\tilde{M}_{12}\tilde{M}_{22}^{-1} \end{pmatrix} U'\sqrt{G}\bar{B}_m.
$$

For $\hat{\theta}_1$, the matrix M is W, and so

$$
\tilde{M} = \tilde{W} = (\Lambda U)^{-1} W \left[(\Lambda U)^{-1}\right]' = \begin{pmatrix} \tilde{W}_{11} & \tilde{W}_{12} \\ \tilde{W}_{21} & \tilde{W}_{22} \end{pmatrix}.
$$

Therefore

$$
\sqrt{N}(\hat{\theta}_1 - \theta_0) \xrightarrow{d} -\sqrt{G}VA^{-1}\left(\bar{B}_d - \beta_{\tilde{W}}\bar{B}_q\right),
$$

where we have used $U'\bar{B}_m \stackrel{d}{=} \bar{B}_m = (\bar{B}_d', \bar{B}_q')'$ for any orthonormal matrix U.

For $\tilde{\theta}$, the matrix M_Λ is $\mathbb{S}$, and so

$$\sqrt{N}(\tilde{\theta}_2 - \theta_0) \overset{d}{\to} -\left[V\Sigma'U'M_\Lambda^{-1}U\Sigma V'\right]^{-1} V\Sigma'U'M_\Lambda^{-1}\sqrt{G}U'\bar{B}_m$$
$$= -V(\Sigma U'\mathbb{S}^{-1}U\Sigma')^{-1}\Sigma U'\mathbb{S}^{-1}U\sqrt{G}U'\bar{B}_m$$
$$\overset{d}{=} -V(\Sigma\mathbb{S}^{-1}\Sigma')^{-1}\Sigma\mathbb{S}^{-1}\sqrt{G}\bar{B}_m,$$

using the asymptotic equivalence $(\mathbb{S}, \bar{B}_m) \overset{d}{=} (U'\mathbb{S}U, U'\bar{B}_m)$ for any orthonormal matrix U. Therefore,

$$\sqrt{N}(\tilde{\theta}_2 - \theta_0) \overset{d}{\to} -VA^{-1}\sqrt{G}(\bar{B}_d - \beta_{\mathbb{S}}\bar{B}_q).$$

For the estimator $\hat{\theta}_2$, the matrix M_Λ is D_∞. We have

$$\sqrt{N}(\hat{\theta}_2 - \theta_0) \overset{d}{\to} -\left[\Gamma_\Lambda' D_\infty^{-1}\Gamma_\Lambda\right]^{-1}\Gamma_\Lambda' D_\infty^{-1}\sqrt{G}\bar{B}_m$$
$$= -\left[V\Sigma'\left(U'D_\infty U\right)^{-1}\Sigma V'\right]^{-1} V\Sigma\left(U'D_\infty U\right)^{-1}U'\sqrt{G}\bar{B}_m$$
$$= -V\left[\Sigma'\mathbb{D}_\infty^{-1}\Sigma\right]^{-1}\Sigma'\mathbb{D}_\infty^{-1}U'\sqrt{G}\bar{B}_m$$
$$= -VA^{-1}\begin{pmatrix} I_d, & -\mathbb{D}_{12}\mathbb{D}_{22}^{-1}\end{pmatrix}U'\sqrt{G}\bar{B}_m \tag{1.44}$$

where

$$\mathbb{D}_\infty = U'D_\infty U = \begin{pmatrix} \underset{d\times d}{\mathbb{D}_{11}} & \underset{d\times q}{\mathbb{D}_{12}} \\ \underset{q\times d}{\mathbb{D}_{21}} & \underset{q\times q}{\mathbb{D}_{22}} \end{pmatrix}.$$

To investigate each component of $\mathbb{D}_\infty = G^{-1}\sum_{g=1}^{G} U'\tilde{D}_g\tilde{D}_g'U$, we first look at the term $U'\tilde{D}_g$ for each $g = 1, ..., G$:

$$U'\tilde{D}_g = U'B_{m,g} - U'\Gamma_\Lambda(\Gamma_\Lambda' W_\Lambda^{-1}\Gamma_\Lambda)^{-1}\Gamma_\Lambda' W_\Lambda^{-1}\bar{B}_m$$
$$= U'B_{m,g} - U'U\Sigma V'(\Gamma_\Lambda' W_\Lambda^{-1}\Gamma_\Lambda)^{-1}V\Sigma'U'W_\Lambda^{-1}UU'\bar{B}_m$$
$$= B_{m,g}^U - \Sigma(\Sigma'U'W_\Lambda^{-1}U\Sigma)^{-1}\Sigma'U'W_\Lambda^{-1}U\bar{B}_m^U \tag{1.45}$$

where $B^U_{m,g} = U' B_{m,g}$ and $\bar{B}^U_m = U'\bar{B}_m$. But

$$
\begin{aligned}
& B^U_{m,g} - \Sigma(\Sigma'\tilde{W}^{-1}\Sigma)^{-1}\Sigma'\tilde{W}^{-1}\bar{B}^U_m \\
&= B^U_{m,g} - \begin{bmatrix} A \\ O \end{bmatrix} (A\tilde{W}^{11}A)^{-1} \begin{bmatrix} A & O' \end{bmatrix} \begin{pmatrix} \tilde{W}^{11} & \tilde{W}^{12} \\ \tilde{W}^{21} & \tilde{W}^{22} \end{pmatrix} \bar{B}^U_m \\
&= B^U_{m,g} - \begin{pmatrix} \left(\tilde{W}^{11}\right)^{-1} & O' \\ O & O \end{pmatrix} \begin{pmatrix} \tilde{W}^{11} & \tilde{W}^{12} \\ \tilde{W}^{21} & \tilde{W}^{22} \end{pmatrix} \bar{B}^U_m \\
&= B^U_{m,g} - \begin{pmatrix} I & \left(\tilde{W}^{11}\right)^{-1}\tilde{W}^{12} \\ O & O \end{pmatrix} \bar{B}^U_m \\
&= B^U_{m,g} - \begin{bmatrix} \bar{B}^U_d - \beta_{\tilde{W}}\bar{B}^U_q \\ O \end{bmatrix} = (B^U_{m,g} - \bar{B}^U_m) + w\bar{B}^U_q
\end{aligned}
$$

for

$$
w = \begin{pmatrix} \beta_{\tilde{W}} \\ I_q \end{pmatrix} \in \mathbb{R}^{m\times q}.
$$

So, the matrix $\mathbb{D}_\infty$ can be represented by

$$
\begin{aligned}
\mathbb{D}_\infty &= \frac{1}{G}\sum_{g=1}^{G} \left(B^U_{m,g} - \bar{B}^U_m + w\bar{B}^U_q\right)\left(B^U_{m,g} - \bar{B}^U_m + w\bar{B}^U_q\right)' \\
&= \frac{1}{G}\sum_{g=1}^{G} (B^U_{m,g} - \bar{B}^U_m)(B^U_{m,g} - \bar{B}^U_m)' + w\bar{B}^U_q(\bar{B}^U_q)'w' \\
&:= \tilde{S}^U_\infty + w\bar{B}^U_q(\bar{B}^U_q)'w'.
\end{aligned}
$$

From this, the block matrix components of $\mathbb{D}_\infty$ are

$$
\begin{aligned}
\mathbb{D}_{11} &= \tilde{S}^U_{\infty,11} + \beta_{\tilde{W}}\bar{B}^U_q(\bar{B}^U_q)'\beta'_{\tilde{W}}, \\
\mathbb{D}_{12} &= \tilde{S}^U_{\infty,12} + \beta_{\tilde{W}}\bar{B}^U_q(\bar{B}^U_q)', \\
\mathbb{D}_{21} &= \tilde{S}^U_{\infty,21} + \bar{B}^U_q(\bar{B}^U_q)'\beta'_{\tilde{W}}, \\
\mathbb{D}_{22} &= \tilde{S}^U_{\infty,22} + \bar{B}^U_q(\bar{B}^U_q)' = S^U_{\infty,22}. \qquad (1.46)
\end{aligned}
$$

Using these representations, we can rewrite (1.44) as

$$\begin{aligned}
&\sqrt{N}(\hat{\theta}_2-\theta_0) \xrightarrow{d} -VA^{-1}\begin{pmatrix} I_d, & -\mathbb{D}_{12}\mathbb{D}_{22}^{-1}\end{pmatrix}\sqrt{G}\bar{B}_m^U \\
&= VA^{-1}\sqrt{G}\left[\bar{B}_d^U - \mathbb{D}_{12}\mathbb{D}_{22}^{-1}\bar{B}_q^U\right] \\
&= -VA^{-1}\sqrt{G}\left[\bar{B}_d^U - \left(\tilde{S}_{\infty,12}^U + \beta_{\tilde{W}}\bar{B}_q^U(\bar{B}_q^U)'\right)\left(S_{\infty,22}^U\right)^{-1}\bar{B}_q^U\right] \\
&= -VA^{-1}\sqrt{G}\left\{\bar{B}_d^U - \left[S_{\infty,12}^U - (\bar{B}_d^U - \beta_{\tilde{W}}\bar{B}_q^U)(\bar{B}_q^U)'\right]\left(S_{\infty,22}^U\right)^{-1}\bar{B}_q^U\right\} \\
&\overset{d}{=} -VA^{-1}\sqrt{G}\left(\bar{B}_d - \beta_{S_\infty}\bar{B}_q\right) - VA^{-1}\sqrt{G}\left\{\bar{B}_d - \beta_{\tilde{W}}\bar{B}_q\right\}\cdot(\kappa_G/G).
\end{aligned}$$

(d) It is easy to check that the weak convergences in (a)∼(c) hold jointly. By continuous mapping theorem we have

$$\sqrt{N}(\hat{\theta}_2-\theta_0) - \sqrt{N}(\tilde{\theta}_2-\theta_0) - \sqrt{N}(\hat{\theta}_1-\theta_0)\cdot(\kappa_G/G) \xrightarrow{d} 0,$$

which implies that

$$\sqrt{N}(\hat{\theta}_2-\theta_0) - \sqrt{N}(\tilde{\theta}_2-\theta_0) - \sqrt{N}(\hat{\theta}_1-\theta_0)\cdot(\kappa_G/G) = o_p(1).$$

That is

$$\sqrt{N}(\hat{\theta}_2-\theta_0) = \sqrt{N}(\tilde{\theta}_2-\theta_0) + \sqrt{N}(\hat{\theta}_1-\theta_0)\cdot(\kappa_G/G) + o_p(1).$$

(e) Using the same argument in the proof of Proposition 1, we have

$$\begin{aligned}
\sqrt{N}g_N(\hat{\theta}_2) &= \frac{1}{\sqrt{G}}\sum_{g=1}^{G}\left(\frac{1}{\sqrt{L_N}}\sum_{i=1}^{L_N} f_i^g(\hat{\theta}_2)\right) \\
&\xrightarrow{d} \Lambda\sqrt{G}\left(UU'\bar{B}_m - \Gamma_\Lambda\left[\Gamma_\Lambda' D_\infty^{-1}\Gamma_\Lambda\right]^{-1}\Gamma_\Lambda' D_\infty^{-1}\bar{B}_m\right) \\
&\overset{d}{=} \Lambda\sqrt{G}\left[U\bar{B}_m^U - \Gamma_\Lambda VA^{-1}\left(\bar{B}_d^U - \mathbb{D}_{12}\mathbb{D}_{22}^{-1}\bar{B}_q^U\right)\right]
\end{aligned}$$

with $\mathbb{D}_{12}$ and $\mathbb{D}_{22}$ given in (1.46). Therefore,

$$
\begin{aligned}
&J(\hat{\theta}_2) = N g_N(\hat{\theta}_2)' \left(\hat{\Omega}(\hat{\theta}_1)\right)^{-1} g_N(\hat{\theta}_2) \\
&\xrightarrow{d} G\left\{U\bar{B}_m^U - \Gamma_\Lambda V A^{-1}\left(\bar{B}_d^U - \mathbb{D}_{12}\mathbb{D}_{22}^{-1}\bar{B}_q^U\right)\right\}' \times \Lambda\left(\Lambda D_\infty \Lambda'\right)^{-1}\Lambda \\
&\times \left\{U\bar{B}_m^U - \Gamma_\Lambda V A^{-1}\left(\bar{B}_d^U - \mathbb{D}_{12}\mathbb{D}_{22}^{-1}\bar{B}_q^U\right)\right\} \\
&= G\left\{\bar{B}_m^U - U'\Gamma_\Lambda V A^{-1}\left(\bar{B}_d^U - \mathbb{D}_{12}\mathbb{D}_{22}^{-1}\bar{B}_q^U\right)\right\}' U' D_\infty^{-1} U \\
&\times \left\{\bar{B}_m^U - U'\Gamma_\Lambda V A^{-1}\left(\bar{B}_d^U - \mathbb{D}_{12}\mathbb{D}_{22}^{-1}\bar{B}_q^U\right)\right\} \\
&= G\left\{\bar{B}_m^U - \begin{bmatrix} I_{d\times d} \\ O_{q\times d}\end{bmatrix}\left(\bar{B}_d^U - \mathbb{D}_{12}\mathbb{D}_{22}^{-1}\bar{B}_q^U\right)\right\}' \mathbb{D}_\infty^{-1} \\
&\times \left\{\bar{B}_m^U - \begin{bmatrix} I_{d\times d} \\ O_{q\times d}\end{bmatrix}\left(\bar{B}_d^U - \mathbb{D}_{12}\mathbb{D}_{22}^{-1}\bar{B}_q^U\right)\right\} \\
&= G\begin{pmatrix} \mathbb{D}_{12}\mathbb{D}_{22}^{-1}\bar{B}_q^U \\ \bar{B}_q^U \end{pmatrix}' \mathbb{D}_\infty^{-1} \begin{pmatrix} \mathbb{D}_{12}\mathbb{D}_{22}^{-1}\bar{B}_q^U \\ \bar{B}_q^U \end{pmatrix} \\
&= G(\bar{B}_q^U)'\mathbb{D}_{22}^{-1}\bar{B}_q^U =^d G\bar{B}_q' S_{\infty,22}^{-1}\bar{B}_q = \kappa_G,
\end{aligned}
$$

where the second last equality follows from straightforward calculations. The joint convergence can be proved easily. ■

Proof of Proposition 7. It follows from

$$\sqrt{N}(\hat{\theta}_2 - \theta_0) \xrightarrow{d} V A^{-1}\sqrt{G}\left[\bar{B}_d^U - \mathbb{D}_{12}\mathbb{D}_{22}^{-1}\bar{B}_q^U\right] \text{ and } \hat{\Omega}(\hat{\theta}_1) \xrightarrow{d} \Lambda D_\infty \Lambda'$$

jointly that

$$
\begin{aligned}
&F_{2,\hat{\Omega}(\hat{\theta}_1)}(\hat{\theta}_2)\\
&= \left[R(\hat{\theta}_2-\theta_0)\right]'\left(R\widehat{var}_{\hat{\Omega}(\hat{\theta}_1)}(\hat{\theta}_2)R'\right)^{-1}R(\hat{\theta}_2-\theta_0)/p\\
&\overset{d}{\to} G(\bar{B}_d^U-\mathbb{D}_{12}\mathbb{D}_{22}^{-1}\bar{B}_q^U)'A^{-1\prime}V'R'\left[R\left(\Gamma'\left(\Lambda D_\infty\Lambda'\right)^{-1}\Gamma\right)^{-1}R'\right]^{-1}\\
&\times RVA^{-1}(\bar{B}_d^U-\mathbb{D}_{12}\mathbb{D}_{22}^{-1}\bar{B}_q^U)/p\\
&= G(\bar{B}_d^U-\mathbb{D}_{12}\mathbb{D}_{22}^{-1}\bar{B}_q^U)'A^{-1\prime}V'R'\cdot\left\{R\left[\Gamma'\left(\Lambda'\right)^{-1}U\left(U'D_\infty U\right)^{-1}U'\Lambda^{-1}\Gamma\right]^{-1}R'\right\}^{-1}\\
&\times RVA^{-1}(\bar{B}_d^U-\mathbb{D}_{12}\mathbb{D}_{22}^{-1}\bar{B}_d^U)/p\\
&= G(\bar{B}_d^U-\mathbb{D}_{12}\mathbb{D}_{22}^{-1}\bar{B}_q^U)'A^{-1\prime}V'R'\left\{RVA^{-1}\mathbb{D}_{11\cdot 2}A^{-1\prime}V'R'\right\}^{-1}\\
&\times RVA^{-1}(\bar{B}_d^U-\mathbb{D}_{12}\mathbb{D}_{22}^{-1}\bar{B}_d^U)/p.
\end{aligned}
$$

Let $\tilde{U}_{p\times p}\tilde{\Sigma}\tilde{V}'_{d\times d}$ be a SVD of RVA^{-1}, where $\tilde{\Sigma}=\left(\tilde{A}_{p\times p},O_{p\times(d-p)}\right)$. By definition, $\tilde{V}$ is the matrix of eigenvectors of $(RVA^{-1})'(RVA^{-1})$. Let

$$
\mathbb{V}=\begin{pmatrix}\tilde{V}_{d\times d} & O\\ O & I_{q\times q}\end{pmatrix}
$$

and define

$$
\tilde{\mathbb{D}}=\begin{pmatrix}\tilde{\mathbb{D}}_{11} & \tilde{\mathbb{D}}_{12}\\ \tilde{\mathbb{D}}_{21} & \tilde{\mathbb{D}}_{22}\end{pmatrix}=\begin{pmatrix}\tilde{V}_{d\times d} & O\\ O & I_q\end{pmatrix}'\begin{pmatrix}\mathbb{D}_{11} & \mathbb{D}_{12}\\ \mathbb{D}_{21} & \mathbb{D}_{22}\end{pmatrix}\begin{pmatrix}\tilde{V}_{d\times d} & O\\ O & I_q\end{pmatrix}=\mathbb{V}'\mathbb{D}_\infty\mathbb{V}.
$$

Then

$$
\begin{aligned}
\tilde{\mathbb{D}} &= \frac{1}{G}\sum_{g=1}^{G}\mathbb{V}'U'(B_{m,g}-\bar{B}_m)(B_{m,g}-\bar{B}_m)'\mathbb{V}U+\begin{pmatrix}\tilde{V}'\beta_{\tilde{W}}\\ I_q\end{pmatrix}\bar{B}_q^U(\bar{B}_q^U)'\begin{pmatrix}\tilde{V}'\beta_{\tilde{W}}\\ I_q\end{pmatrix}'\\
&=^d \frac{1}{G}\sum_{g=1}^{G}(B_{m,g}-\bar{B}_m)(B_{m,g}-\bar{B}_m)'+\begin{pmatrix}\tilde{V}'\beta_{\tilde{W}}\\ I_q\end{pmatrix}\bar{B}_q\bar{B}_q'\begin{pmatrix}\tilde{V}'\beta_{\tilde{W}}\\ I_q\end{pmatrix}',
\end{aligned}
$$

which implies that

$$\tilde{\mathbb{D}}_{11} := \begin{pmatrix} \tilde{\mathbb{D}}_{pp} & \tilde{\mathbb{D}}_{p,d-p} \\ \tilde{\mathbb{D}}_{d-p,p} & \tilde{\mathbb{D}}_{d-p,d-p} \end{pmatrix} \tag{1.47}$$

$$\stackrel{d}{=} \frac{1}{G}\sum_{g=1}^{G}(B_{d,g} - \bar{B}_d)(B_{d,g} - \bar{B}_d)' \tag{1.48}$$

$$+ \left(\tilde{V}'\beta_{\tilde{W}}\right)\bar{B}_q\bar{B}_q'\left(\tilde{V}'\beta_{\tilde{W}}\right)', \tag{1.49}$$

and

$$\tilde{\mathbb{D}}_{12} := \begin{pmatrix} \tilde{\mathbb{D}}_{pq} \\ \tilde{\mathbb{D}}_{d-p,q} \end{pmatrix} \stackrel{d}{=} \frac{1}{G}\sum_{g=1}^{G}(B_{d,g} - \bar{B}_d)(B_{q,g} - \bar{B}_q)' + \left(\tilde{V}'\beta_{\tilde{W}}\right)\bar{B}_q\bar{B}_q'. \tag{1.50}$$

Now

$$\begin{aligned}
&F_{2,\hat{\Omega}(\hat{\theta}_1)}(\hat{\theta}_2) \\
&\stackrel{d}{\to} G(\bar{B}_d^U - \mathbb{D}_{12}\mathbb{D}_{22}^{-1}\bar{B}_q^U)'\tilde{V}\tilde{\Sigma}'\tilde{U}'\left\{\tilde{U}\tilde{\Sigma}\tilde{V}'\mathbb{D}_{11\cdot 2}\tilde{V}\tilde{\Sigma}'\tilde{U}'\right\}^{-1}\tilde{U}\tilde{\Sigma}\tilde{V}'(\bar{B}_d^U - \mathbb{D}_{12}\mathbb{D}_{22}^{-1}\bar{B}_q^U)/p \\
&= G(\bar{B}_d^U - \mathbb{D}_{12}\mathbb{D}_{22}^{-1}\bar{B}_q^U)'\tilde{V}\tilde{\Sigma}'\cdot\left\{\tilde{\Sigma}\tilde{V}'\mathbb{D}_{11\cdot 2}\tilde{V}\tilde{\Sigma}'\right\}^{-1}\cdot\tilde{\Sigma}\tilde{V}'(\bar{B}_d^U - \mathbb{D}_{12}\mathbb{D}_{22}^{-1}\bar{B}_q^U)/p \\
&= G(\tilde{V}'\bar{B}_d^U - \tilde{\mathbb{D}}_{12}\tilde{\mathbb{D}}_{22}^{-1}\bar{B}_q^U)'\cdot\tilde{\Sigma}'\left\{\tilde{\Sigma}\tilde{\mathbb{D}}_{11\cdot 2}\tilde{\Sigma}'\right\}^{-1}\tilde{\Sigma} \\
&\times(\tilde{V}'\bar{B}_d^U - \tilde{\mathbb{D}}_{12}\tilde{\mathbb{D}}_{22}^{-1}\bar{B}_q^U)/p \\
&\stackrel{d}{=} G\left[\bar{B}_p - \tilde{\mathbb{D}}_{pq}\tilde{\mathbb{D}}_{qq}^{-1}\bar{B}_q\right]'\tilde{A}'\left\{\tilde{A}\left(\tilde{\mathbb{D}}_{pp} - \tilde{\mathbb{D}}_{pq}\tilde{\mathbb{D}}_{qq}^{-1}\tilde{\mathbb{D}}_{qp}\right)\tilde{A}'\right\}^{-1}\tilde{A}\left[\bar{B}_p - \tilde{\mathbb{D}}_{pq}\tilde{\mathbb{D}}_{qq}^{-1}\bar{B}_q\right]/p \\
&\stackrel{d}{=} G\left[\bar{B}_p - \tilde{\mathbb{D}}_{pq}\tilde{\mathbb{D}}_{qq}^{-1}\bar{B}_q\right]'\left(\tilde{\mathbb{D}}_{pp} - \tilde{\mathbb{D}}_{pq}\tilde{\mathbb{D}}_{qq}^{-1}\tilde{\mathbb{D}}_{qp}\right)^{-1}\left[\bar{B}_p - \tilde{\mathbb{D}}_{pq}\tilde{\mathbb{D}}_{qq}^{-1}\bar{B}_q\right]/p,
\end{aligned}$$

where $\tilde{\mathbb{D}}_{pq}$, $\tilde{\mathbb{D}}_{qq}$, and $\tilde{\mathbb{D}}_{qp}$ in the last two equalities are understood to equals the corresponding components on the right hand sides of (1.49) and (1.50). Here we have abused the notation a little bit. We have

$$\begin{pmatrix} \tilde{\mathbb{D}}_{pp} & \tilde{\mathbb{D}}_{pq} \\ \tilde{\mathbb{D}}_{pq}' & \tilde{\mathbb{D}}_{qq} \end{pmatrix} = \begin{pmatrix} \tilde{S}_{pp} & \tilde{S}_{pq} \\ \tilde{S}_{pq}' & \tilde{S}_{qq} \end{pmatrix} + \tilde{w}\bar{B}_q\bar{B}_q'\tilde{w}' \tag{1.51}$$

for

$$\tilde{w} = \begin{pmatrix} \tilde{\beta}^{p}_{\tilde{W}} \\ I_q \end{pmatrix} \in \mathbb{R}^{(p+q)\times q}.$$

We have therefore shown that the first representation of the limit of $F_{2,\hat{\Omega}(\hat{\theta}_1)}(\hat{\theta}_2)$ holds. Direct calculations show that the second representation is numerically identical to the first representation. This completes the proof of Proposition 7. ■

Proof of Lemma 8. The centered CCE $\Omega^c(\breve{\theta}_N)$ can be represented as:

$$\hat{\Omega}^c(\breve{\theta}_N) = \frac{1}{G}\sum_{h=1}^{G}\left\{\frac{1}{\sqrt{L_N}}\sum_{i=1}^{L_N}\left(f_i^h(\breve{\theta}_N) - \frac{1}{N}\sum_{g=1}^{G}\sum_{s=1}^{L_N} f_s^g(\breve{\theta}_N)\right)\right.$$
$$\left.\times \frac{1}{\sqrt{L_N}}\sum_{j=1}^{L_N}\left(f_j^h(\breve{\theta}_N) - \frac{1}{N}\sum_{g=1}^{G}\sum_{s=1}^{L_N} f_s^g(\breve{\theta}_N)\right)'\right\}.$$

To prove Part (a), it suffices to show that

$$\frac{1}{\sqrt{L_N}}\sum_{i=1}^{L_N}\left(f_i^h(\breve{\theta}_N) - \frac{1}{N}\sum_{g=1}^{G}\sum_{s=1}^{L_N} f_s^g(\breve{\theta}_N)\right) \tag{1.52}$$
$$= \frac{1}{\sqrt{L_N}}\sum_{i=1}^{L_N}\left(f_i^h(\theta_0) - \frac{1}{N}\sum_{g=1}^{G}\sum_{s=1}^{L_N} f_s^g(\theta_0)\right)(1+o_p(1))$$

holds for each $h = 1, ..., G$. By Assumption 3 and using a Taylor expansion, we have

$$\frac{1}{\sqrt{L_N}}\sum_{i=1}^{L_N} f_i^h(\breve{\theta}_N) = (1+o_p(1))\left(\frac{1}{\sqrt{L_N}}\sum_{i=1}^{L_N} f_i^h(\theta_0) + \frac{1}{L_N}\sum_{i=1}^{L_N}\frac{\partial f_i^h(\breve{\theta}_N)}{\partial\theta'}\sqrt{L_N}(\breve{\theta}_N - \theta_0)\right)$$

Using $\sqrt{N}(\breve{\theta}_N - \theta_0) = O_p(1)$ and Assumption 5, we have

$$\frac{1}{\sqrt{L_N}}\sum_{i=1}^{L_N} f_i^h(\breve{\theta}_N) = (1+o_p(1))\left(\frac{1}{\sqrt{L_N}}\sum_{i=1}^{L_N} f_i^h(\theta_0) + \Gamma\sqrt{L_N}(\breve{\theta}_N - \theta_0)\right)$$

for each $h = 1, ..., G$. That is, the effect of the estimation uncertainty in $\breve{\theta}_N$ does not change with the cluster. It then follows that

$$\frac{1}{\sqrt{L_N}} \sum_{i=1}^{L_N} \left(f_i^h(\breve{\theta}_N) - \frac{1}{N} \sum_{g=1}^{G} \sum_{s=1}^{L_N} f_s^g(\breve{\theta}_N) \right)$$
$$= (1 + o_p(1)) \left(\frac{1}{\sqrt{L_N}} \sum_{i=1}^{L_N} f_i^h(\theta_0) - \frac{1}{G} \sum_{g=1}^{G} \frac{1}{\sqrt{L_N}} \sum_{i=1}^{L_N} f_i^g(\theta_0) \right),$$

which completes the proof of part (a).

To prove Part (b), we apply CLT in Assumption 4 together with 6 to obtain:

$$\frac{1}{\sqrt{L_N}} \sum_{i=1}^{L_N} f_i^h(\theta_0) - \frac{1}{G} \sum_{g=1}^{G} \frac{1}{\sqrt{L_N}} \sum_{i=1}^{L_N} f_i^g(\theta_0) \xrightarrow{d} \Lambda \left(B_{m,h} - \bar{B}_m \right),$$

where the convergence holds jointly for $h = 1, ..., G$. As a result,

$$\hat{\Omega}^c(\hat{\theta}_1) \xrightarrow{d} \frac{1}{G} \Lambda \sum_{g=1}^{G} \left(B_{m,g} - \bar{B}_m \right) \left(B_{m,g} - \bar{B}_m \right)' \Lambda'.$$

■

Proof of Proposition 9. The proof of part (a) is essentially the same as the proof of Proposition 7. The only difference is that the second term in (1.51) will not be present for the centered two-step GMM estimator $\hat{\theta}_2^c$. The proof of part (b) is similar. The proof of part (e) is similar to that of Proposition 6(e).

To prove part (c), recall that the restricted two-step GMM estimator $\hat{\theta}_2^{c,r}$ minimizes

$$g_N(\theta)' \left[\hat{\Omega}^c(\hat{\theta}_1) \right]^{-1} g_N(\theta)/2 + \lambda_N'(R\theta - r). \tag{1.53}$$

The first order conditions are

$$\Gamma_N(\hat{\theta}_2^{c,r}) \left[\hat{\Omega}^c(\hat{\theta}_1) \right]^{-1} g_N(\hat{\theta}_2^{c,r}) + R'\lambda_N = 0, \tag{1.54}$$
$$R\hat{\theta}_2^{c,r} - r = 0.$$

Using a Taylor expansion and Assumption 3, we can combine two FOC's to get

$$\begin{aligned}\sqrt{N}\left(\hat{\theta}_2^{c,r}-\theta_0\right) &= -\tilde{\Phi}^{-1}\Gamma'\left[\hat{\Omega}^c(\hat{\theta}_1)\right]^{-1}\sqrt{N}g_N(\theta_0) \\ &- \tilde{\Phi}^{-1}R'\left(R\tilde{\Phi}^{-1}R'\right)^{-1}R\tilde{\Phi}^{-1}\Gamma'\left[\hat{\Omega}^c(\hat{\theta}_1)\right]^{-1}\sqrt{N}g_N(\theta_0)+o_p(1),\end{aligned} \tag{1.55}$$

where $\tilde{\Phi} := \Gamma'\left[\hat{\Omega}^c(\hat{\theta}_1)\right]^{-1}\Gamma$. Subtracting (1.55) from (1.9), we have

$$\sqrt{N}\left(\hat{\theta}_2^{c}-\hat{\theta}_2^{c,r}\right) = -\tilde{\Phi}^{-1}R'\left(R\tilde{\Phi}^{-1}R'\right)^{-1}R\tilde{\Phi}^{-1}\Gamma'\left[\hat{\Omega}^c(\theta_0)\right]^{-1}\sqrt{N}g_N(\theta_0)+o_p(1). \tag{1.56}$$

By Taylor expansion and Assumption 3 we have

$$g_N'(\hat{\theta}_{2N}^{c,r}) = g_N'(\hat{\theta}_2^{c}) + (\hat{\theta}_2^{c,r}-\hat{\theta}_2^{c})'\Gamma' + o_p(N^{-1/2})$$

and

$$\begin{aligned}&Ng_N'(\hat{\theta}_2^{c})\left[\hat{\Omega}^c(\hat{\theta}_1)\right]^{-1}g_N(\hat{\theta}_{2N}^{c}) - Ng_N'(\hat{\theta}_2^{c,r})\left[\hat{\Omega}^c(\hat{\theta}_1)\right]^{-1}g_N(\hat{\theta}_2^{c}) \\ &= N(\hat{\theta}_2^{c}-\hat{\theta}_2^{c,r})'\Gamma_N'(\hat{\theta}_2^{c})\left[\hat{\Omega}^c(\hat{\theta}_1)\right]^{-1}g_N(\hat{\theta}_2^{c}) + O_p\left(\frac{1}{\sqrt{N}}\right) \\ &= O_p\left(\frac{1}{\sqrt{N}}\right).\end{aligned} \tag{1.57}$$

Here the last equality follows from the FOC's for $\hat{\theta}_2^c$. In a similar way, we can write the second term in (1.57) as

$$\begin{aligned}&Ng'(\hat{\theta}_2^{c,r})\left[\hat{\Omega}^c(\hat{\theta}_1)\right]^{-1}g_N(\hat{\theta}_2^{c}) \\ &= Ng_N'(\hat{\theta}_2^{c,r})\left[\hat{\Omega}^c(\hat{\theta}_1)\right]g_N(\hat{\theta}_2^{c,r}) + N(\hat{\theta}_2^{c}-\hat{\theta}_2^{c,r})'\tilde{\Phi}(\hat{\theta}_2^{c}-\hat{\theta}_2^{c,r}) + o_p(1).\end{aligned}$$

Combining this and (1.56), we get

$$
\begin{aligned}
& LR_{\hat{\Omega}^c(\hat{\theta}_1)}(\hat{\theta}_2^c, \hat{\theta}_2^{c,r}) \\
&= \left\{ N g_N(\hat{\theta}_2^c)' \left[\hat{\Omega}^c(\hat{\theta}_1)\right]^{-1} g_N(\hat{\theta}_2^c) - N g_N(\hat{\theta}_2^{c,r})' \left[\hat{\Omega}^c(\hat{\theta}_1)\right]^{-1} g_N(\hat{\theta}_2^{c,r}) \right\} /p \\
&= N(\hat{\theta}_2^{c,r} - \hat{\theta}_2^c)' \tilde{\Phi} (\hat{\theta}_2^{c,r} - \hat{\theta}_2^c)/p + o_p(1) \\
&= \sqrt{N} g_N'(\theta_0) \left[\hat{\Omega}^c(\theta_0)\right]^{-1} \Gamma \tilde{\Phi}^{-1} R' \left(R \tilde{\Phi}^{-1} R'\right)^{-1} R \tilde{\Phi}^{-1} \Gamma' \sqrt{N} g_N(\theta_0)/p + o_p(1) \\
&= \sqrt{N} \left(R\hat{\theta}_2^c - r\right)' \left(R\tilde{\Phi}^{-1}R'\right)^{-1} \sqrt{N} \left(R\hat{\theta}_2^c - r\right)/p + o_p(1) \\
&= F_{2,\hat{\Omega}^c(\hat{\theta}_1)}(\hat{\theta}_2^c) + o_p(1).
\end{aligned}
$$

as desired.

To prove part (d), we rewrite the FOC in (1.54) as

$$
\begin{aligned}
\sqrt{N} \Delta_{\hat{\Omega}^c(\hat{\theta}_1)}(\hat{\theta}_2^{c,r}) &= -R' \sqrt{N} \lambda_N \\
&= -R' \left(R\tilde{\Phi}^{-1}R'\right)^{-1} R \tilde{\Phi}^{-1} \Gamma' \left[\hat{\Omega}^c(\hat{\theta}_1)\right]^{-1} \sqrt{N} g_N(\theta_0) + o_p(1) \\
&= \tilde{\Phi} \sqrt{N} \left(\hat{\theta}_{2N}^c - \hat{\theta}_2^{c,r}\right) + o_p(1).
\end{aligned}
$$

So,

$$
\begin{aligned}
LM_{\hat{\Omega}^c(\hat{\theta}_1)}(\hat{\theta}_2^{c,r}) &= N \left[\Delta_{\hat{\Omega}^c(\hat{\theta}_1)}(\hat{\theta}_2^{c,r})\right]' \tilde{\Phi}^{-1} \left[\Delta_{\hat{\Omega}^c(\hat{\theta}_1)}(\hat{\theta}_2^{c,r})\right]/p \\
&= N \left(\hat{\theta}_2^c - \hat{\theta}_2^{c,r}\right)' \tilde{\Phi} \sqrt{N} \left(\hat{\theta}_{2N}^c - \hat{\theta}_2^{c,r}\right)/p + o_p(1) \\
&= LR_{\hat{\Omega}^c(\hat{\theta}_1)}(\hat{\theta}_{2N}^c, \hat{\theta}_2^{c,r}) + o_p(1) \\
&= F_{2,\hat{\Omega}^c(\hat{\theta}_1)}(\hat{\theta}_2^c) + o_p(1).
\end{aligned}
$$

■

Proof of Proposition 10. For the result with CU-GEE estimator $\hat{\theta}_{\text{GEE}}^{\text{cu}}$, we have

$$\sqrt{N}(\hat{\theta}^{\text{cu}}_{\text{GEE}} - \theta_0) = -\left(\Gamma' \left(\hat{\Omega}^c(\hat{\theta}^{\text{cu}}_{\text{GEE}})\right)^{-1} \Gamma\right)^{-1} \Gamma' \left(\hat{\Omega}^c(\hat{\theta}^{cu}_{\text{GEE}})\right)^{-1} \sqrt{N} g_N(\theta_0) + o_p(1).$$

Since $\hat{\theta}^{\text{cu}}_{\text{GEE}}$ is $\sqrt{N}$ consistent, we can apply Lemma 8 to obtain $\hat{\Omega}^c(\hat{\theta}^{\text{cu}}_{\text{GEE}}) = \hat{\Omega}^c(\theta_0) + o_p(1)$. Invoking the continuous mapping theorem yields

$$\sqrt{N}(\hat{\theta}^{\text{cu}}_{\text{GEE}} - \theta_0) \xrightarrow{d} -\left\{\Gamma' \left(\Omega^c_\infty\right)^{-1} \Gamma\right\}^{-1} \left\{\Gamma' \left(\Omega^c_\infty\right)^{-1} \Lambda \sqrt{G} \bar{B}_m\right\}$$

as desired.

For the CU-GMM estimator, we let $\Gamma^j_N(\hat{\theta}^{cu}_{\text{GMM}})$ be the j-th column of $\Gamma_N(\hat{\theta}^{cu}_{\text{GMM}})$. Then, the FOC with respect to the j-th element of $\hat{\theta}^{cu}_{\text{CUE}}$ is

$$\begin{aligned} 0 &= \Gamma^j_N(\hat{\theta}^{cu}_{\text{GMM}})' \left[\Omega^c_N(\hat{\theta}^{cu}_{\text{GMM}})\right]^{-1} g_N(\hat{\theta}^{cu}_{\text{GMM}}) \\ &\quad - g_N(\hat{\theta}^{cu}_{\text{GMM}})' \left[\hat{\Omega}^c(\hat{\theta}^{cu}_{\text{GMM}})\right]^{-1} \Upsilon_j(\hat{\theta}_{\text{CU-GMM}}) \left[\hat{\Omega}^c(\hat{\theta}^{\text{cu}}_{\text{GMM}})\right]^{-1} g_N(\hat{\theta}^{cu}_{\text{GMM}}), \end{aligned} \tag{1.58}$$

where

$$\Upsilon_j(\theta) = \frac{1}{N} \sum_{g=1}^{G} \left(\sum_{r=1}^{L_N} f^g_r(\theta)\right) \left(\sum_{s=1}^{L_N} \frac{\partial f_s(\theta)}{\partial \theta_j}\right)' - L_N g_N(\theta) \left(\frac{\partial g_N(\theta)}{\partial \theta_j}\right)'.$$

The second term in (1.58) can be written as

$$\begin{aligned} & g_N(\hat{\theta}^{cu}_{\text{GMM}})' \left[\hat{\Omega}^c(\hat{\theta}^{cu}_{\text{GMM}})\right]^{-1} \Upsilon_j(\hat{\theta}^{cu}_{\text{GMM}}) \left[\hat{\Omega}^c(\hat{\theta}^{\text{cu}}_{\text{GMM}})\right]^{-1} g_N(\hat{\theta}^{cu}_{\text{GMM}}) \\ &= \sqrt{L_N} g_N(\hat{\theta}^{cu}_{\text{GMM}})' \left[\Omega^c_N(\hat{\theta}^{cu}_{\text{GMM}})\right]^{-1} \left[\frac{1}{G} \sum_{g=1}^{G} \left(\frac{1}{L_N} \sum_{r=1}^{L_N} f^g_r(\hat{\theta}^{cu}_{\text{GMM}})\right)\right. \\ &\cdot \left.\left\{\left(\frac{1}{L_N} \sum_{s=1}^{L_N} \frac{\partial f^g_s(\hat{\theta}^{cu}_{\text{GMM}})}{\partial \theta}\right) - \frac{1}{G} \sum_{g=1}^{G} \left(\frac{1}{L_N} \sum_{s=1}^{L_N} \frac{\partial f^g_s(\hat{\theta}^{cu}_{\text{GMM}})}{\partial \theta}\right)\right\}'\right] \\ &\cdot \left(\hat{\Omega}^c(\hat{\theta}^{cu}_{\text{GMM}})\right)^{-1} \sqrt{L_N} g_N(\hat{\theta}^{cu}_{\text{GMM}}). \end{aligned}$$

Given that $\hat{\theta}^{cu}_{\text{GMM}} = \theta_0 + O_p(L_N^{-1/2})$, we have

$$\begin{aligned}
\hat{\Omega}^c(\hat{\theta}^{cu}_{\text{GMM}}) &= O_p(1) \\
\sqrt{L_N} g_N(\hat{\theta}^{cu}_{\text{GMM}}) &= \frac{1}{G}\sum_{g=1}^{G}\left(\frac{1}{\sqrt{L_N}}\sum_{s=1}^{L_N} f_s^g(\theta_0)\right) + \Gamma\sqrt{L_N}(\hat{\theta}^{cu}_{\text{GMM}} - \theta_0) + o_p(1) \\
&= O_p(1) \\
\frac{1}{L_N}\sum_{s=1}^{L_N} f_s^g(\hat{\theta}^{cu}_{\text{GMM}}) &= \frac{1}{L_N}\sum_{r=1}^{L_N} f_s^g(\theta_0) + \frac{1}{L_N}\sum_{s=1}^{L_N}\frac{\partial f_s^g(\tilde{\theta})}{\partial\theta}(\hat{\theta}^{cu}_{\text{GMM}} - \theta_0) \\
&= O_p\left(\frac{1}{\sqrt{L_N}}\right)
\end{aligned}$$

and for each $g = 1, ..., G$,

$$\begin{aligned}
&\left(\frac{1}{L_N}\sum_{r=1}^{L_N} f_r^g(\hat{\theta}^{cu}_{\text{GMM}})\right)\left\{\left(\frac{1}{L_N}\sum_{s=1}^{L_N}\frac{\partial f_s^g(\hat{\theta}^{cu}_{\text{GMM}})}{\partial\theta}\right.\right. \\
&\left.\left. -\frac{1}{G}\sum_{g=1}^{G}\left(\frac{1}{L_N}\sum_{s=1}^{L_N}\frac{\partial f_s^g(\hat{\theta}^{cu}_{\text{GMM}})}{\partial\theta}\right)\right)'\right\} \\
&= O_p\left(\frac{1}{\sqrt{L_N}}\right)\cdot o_p(1) = o_p\left(\frac{1}{\sqrt{L_N}}\right).
\end{aligned}$$

Combining these together, the second term in FOC in (1.58) is $o_p(L_N^{-1/2})$. As a result,

$$\Gamma_N(\hat{\theta}^{cu}_{\text{GMM}})'\left[\hat{\Omega}^c(\hat{\theta}^{cu}_{\text{GMM}})\right]^{-1} g_N(\hat{\theta}^{cu}_{\text{GMM}}) = o_p\left(\frac{1}{\sqrt{L_N}}\right),$$

and so

$$\begin{aligned}
\sqrt{N}(\hat{\theta}^{cu}_{\text{GMM}} - \theta_0) &= -\left\{\Gamma'\left[\hat{\Omega}^c(\hat{\theta}^{cu}_{\text{GMM}})\right]^{-1}\Gamma\right\}^{-1}\Gamma'\left[\hat{\Omega}^c(\hat{\theta}^{cu}_{\text{GMM}})\right]^{-1}\sqrt{N}g_N(\theta_0) \\
&+ o_p(1) \xrightarrow{d} -\left\{\Gamma'\left(\Omega^c_\infty\right)^{-1}\Gamma\right\}^{-1}\Gamma'\left(\Omega^c_\infty\right)^{-1}\Lambda\sqrt{G}\bar{B}_m.
\end{aligned}$$

■

Proof of Theorem 11. Define $\mathbf{B}_q' = (B_{q,1}', ..., B_{q,G}')'$ and denote

$$v_g = \left(B_{q,g} - \bar{B}_q\right)' \left[\sum_{g=1}^{G} \left(B_{q,g} - \bar{B}_q\right)\left(B_{q,g} - \bar{B}_q\right)'\right]^{-1} \bar{B}_q.$$

Then, the distribution of $\sqrt{G}\tilde{S}_{pq}\tilde{S}_{qq}^{-1}\bar{B}_q$ conditional on $\mathbf{B}_q$ can be represented as

$$\begin{aligned}
&\sqrt{G}\left(\sum_{g=1}^{G}\left(B_{p,g} - \bar{B}_p\right)\left(B_{q,g} - \bar{B}_q\right)'\right)\left(\sum_{g=1}^{G}\left(B_{q,g} - \bar{B}_q\right)\left(B_{q,g} - \bar{B}_q\right)'\right)^{-1} \bar{B}_q \\
&= \sqrt{G}\sum_{g=1}^{G}\left(B_{p,g} - \bar{B}_p\right) v_g = \sqrt{G}\sum_{g=1}^{G} B_{p,g} v_g - \sqrt{G}\bar{B}_p \sum_{g=1}^{G} v_g \\
&\overset{d}{=} N\left(0, G\sum_{g=1}^{G} v_g^2 \cdot I_p\right)
\end{aligned}$$

where the last line holds because $\sum_{g=1}^{G} v_g = 0$. Note that

$$\begin{aligned}
G\sum_{g=1}^{G} v_g^2 = G\sum_{g=1}^{G}&\left\{\left(B_{q,g} - \bar{B}_q\right)' \left[\sum_{g=1}^{G}\left(B_{q,g} - \bar{B}_q\right)\left(B_{q,g} - \bar{B}_q\right)'\right]^{-1} \bar{B}_q\right. \\
&\left.\cdot \bar{B}_q' \left[\sum_{g=1}^{G}\left(B_{q,g} - \bar{B}_q\right)\left(B_{q,g} - \bar{B}_q\right)'\right]^{-1} \left(B_{q,g} - \bar{B}_q\right)\right\} \\
&= G\bar{B}_q' \left[\sum_{g=1}^{G}\left(B_{q,g} - \bar{B}_q\right)\left(B_{q,g} - \bar{B}_q\right)'\right]^{-1} \left[\sum_{g=1}^{G}\left(B_{q,g} - \bar{B}_q\right)\right. \\
&\left.\times \left(B_{q,g} - \bar{B}_q\right)'\right] \left[\sum_{g=1}^{G}\left(B_{q,g} - \bar{B}_q\right)\left(B_{q,g} - \bar{B}_q\right)'\right] \bar{B}_q \\
&= \bar{B}_q' \left[\sum_{g=1}^{G}\left(B_{q,g} - \bar{B}_q\right)\left(B_{q,g} - \bar{B}_q\right)'/G\right]^{-1} \bar{B}_q \\
&= \bar{B}_q'\tilde{S}_{qq}^{-1}\bar{B}_q.
\end{aligned}$$

So conditional on $\mathbf{B}_q$, $\sqrt{G}\tilde{S}_{pq}\tilde{S}_{qq}^{-1}\bar{B}_q$ is distributed as $N(0, \bar{B}_q'\tilde{S}_{qq}^{-1}\bar{B}_q \cdot I_p)$. It then follows that the distribution of $\sqrt{G}\left(\bar{B}_p - \tilde{S}_{pq}\tilde{S}_{qq}^{-1}\bar{B}_q\right)$ conditional on $\mathbf{B}_q$ is

$$\sqrt{G}\left(\bar{B}_p - \tilde{S}_{pq}\tilde{S}_{qq}^{-1}\bar{B}_q\right) \sim N\left(0, (1 + \bar{B}_q'\tilde{S}_{qq}^{-1}\bar{B}_q) \cdot I_p\right)$$

using the independence of $\bar{B}_p$ from $\tilde{S}_{pq}\tilde{S}_{qq}^{-1}\bar{B}_q$ conditional on $\mathbf{B}_q$. Therefore the conditional distribution of ξ_p is

$$\xi_p := \frac{\sqrt{G}(\bar{B}_p - \tilde{S}_{pq}\tilde{S}_{qq}^{-1}\bar{B}_q)}{\sqrt{1 + \bar{B}_q'\tilde{S}_{qq}^{-1}\bar{B}_q}} \sim N(0, I_p).$$

Given that the conditional distribution of ξ_p does not depend on $\mathbf{B}_q$, the unconditional distribution of ξ_p is also $N(0, I_p)$.

Using $\xi_p \sim N(0, I_p)$, $\tilde{S}_{pp\cdot q} \sim G^{-1}\mathbb{W}_m(G - q - 1, I_p)$ and ξ_p is independent of $\tilde{S}_{pp\cdot q}$, we have

$$\xi_p'\left(\frac{G\tilde{S}_{pp\cdot q}}{G - q - 1}\right)^{-1}\xi_p \sim \text{Hotelling's } T^2 \text{ distribution } T^2_{p,G-q-1}.$$

It then follows that

$$\frac{G - p - q}{p\,(G - q - 1)}\xi_p'\left(\frac{G\tilde{S}_{pp\cdot q}}{G - q - 1}\right)^{-1}\xi_p \sim F_{p,G-p-q}.$$

That is

$$\frac{G - p - q}{pG}\xi_p'\left(\tilde{S}_{pp\cdot q}\right)^{-1}\xi_p \sim F_{p,G-p-q}.$$

Together with Proposition 9(c)(d), this completes the proof of the F limit theory in parts (a), (b) and (c). The proof of the t limit theory is similar and is omitted here. ■

Proof of Theorem 12.

We first show that $\widehat{\mathcal{E}}_N = \mathcal{E}_N\left(1+o_p\left(1\right)\right).$ For each $j = 1, ..., d,$ we have

$$\begin{aligned}
\widehat{\mathcal{E}}_N[.,j] &= \left\{\hat{\Gamma}_N'\left[\hat{\Omega}^c(\hat{\theta}_1)\right]^{-1}\hat{\Gamma}_N\right\}^{-1}\hat{\Gamma}_N'\left[\hat{\Omega}^c(\hat{\theta}_1)\right]^{-1}\left.\frac{\partial\hat{\Omega}^c(\theta)}{\partial\theta_j}\right|_{\theta=\hat{\theta}_1} \\
&\times\left[\hat{\Omega}^c(\hat{\theta}_1)\right]^{-1}g_N(\hat{\theta}_2^c) \\
&= \left\{\Gamma'\left[\hat{\Omega}^c(\theta_0)\right]^{-1}\Gamma\right\}^{-1}\Gamma'\left[\hat{\Omega}^c(\theta_0)\right]^{-1}\left.\frac{\partial\hat{\Omega}^c(\theta)}{\partial\theta_j}\right|_{\theta=\hat{\theta}_1}\left[\Omega_N^c(\theta_0)\right]^{-1}g_N(\hat{\theta}_2^c) \\
&\cdot(1+o_p(1))
\end{aligned}$$

where the second equality holds by Assumption 3, 5 and Lemma 8. Using a Taylor expansion, we have

$$g_N(\hat{\theta}_2^c) = g_N(\theta_0) - \Gamma\left\{\Gamma'\left[\hat{\Omega}^c(\theta_0)\right]^{-1}\Gamma\right\}^{-1}\Gamma'\left[\hat{\Omega}^c(\theta_0)\right]^{-1}g_N(\theta_0)(1+o_p(1)).$$

So

$$\begin{aligned}
\widehat{\mathcal{E}}_N[.,j] &= \left\{\Gamma'\left[\Omega_N^c(\theta_0)\right]^{-1}\Gamma\right\}^{-1}\Gamma'\left[\hat{\Omega}^c(\theta_0)\right]^{-1}\left.\frac{\partial\Omega_N^c(\theta)}{\partial\theta_j}\right|_{\theta=\hat{\theta}_1}\left[\hat{\Omega}^c(\theta_0)\right]^{-1}g_N(\hat{\theta}_2^c) \\
&\cdot(1+o_p(1)) \\
&-\left\{\Gamma'\left[\hat{\Omega}^c(\theta_0)\right]^{-1}\Gamma\right\}^{-1}\Gamma'\left[\hat{\Omega}^c(\theta_0)\right]^{-1}\left.\frac{\partial\hat{\Omega}^c(\theta)}{\partial\theta_j}\right|_{\theta=\hat{\theta}_1}\left[\Omega_N^c(\theta_0)\right]^{-1}\Gamma \\
&\times\left\{\Gamma'\left[\hat{\Omega}^c(\theta_0)\right]^{-1}\Gamma\right\}^{-1}\Gamma'\left[\Omega_N^c(\theta_0)\right]^{-1}g_N(\theta_0)\Bigg\}(1+o_p(1))
\end{aligned}$$

for each $j = 1, ..., d$. For the term, $\left.\frac{\partial\Omega_N^c(\theta)}{\partial\theta_j}\right|_{\theta=\hat{\theta}_1}$, recall that

$$\left.\frac{\partial\hat{\Omega}^c(\theta)}{\partial\theta_j}\right|_{\theta=\hat{\theta}_{1N}} = \Upsilon_j(\hat{\theta}_1) + \Upsilon_j'(\hat{\theta}_1),$$

$$\Upsilon_j(\theta) = \frac{1}{N}\sum_{g=1}^{G}\left[\sum_{r=1}^{L_N}\left(f_r^g(\theta) - \frac{1}{N}\sum_{s=1}^{N}f_s(\theta)\right)\left(\sum_{s=1}^{L_N}\left(\frac{\partial f_s^g(\theta)}{\partial\theta_j} - \frac{1}{N}\sum_{s=1}^{N}\frac{\partial f_s(\theta)}{\partial\theta_j}\right)\right)'\right].$$

It remains to show that $\Upsilon_j(\hat{\theta}_1) = \Upsilon_j(\theta_0)(1 + o_p(1))$. From the proof of Lemma 8, we have

$$\frac{1}{\sqrt{L_N}} \sum_{r=1}^{L_N} \left(f_r^g(\hat{\theta}_1) - \frac{1}{N} \sum_{s=1}^{N} f_s(\hat{\theta}_1) \right)$$
$$= \frac{1}{\sqrt{L_N}} \sum_{r=1}^{L_N} \left(f_r^g(\theta_0) - \frac{1}{N} \sum_{s=1}^{N} f_s(\theta_0) \right) (1 + o_p(1)) \tag{1.59}$$

for each $g = 1, ..., G$. By Assumption 3, 7 and a Taylor expansion, we have:

$$\frac{1}{\sqrt{L_N}} \sum_{s=1}^{L_N} \frac{\partial f_s^g(\hat{\theta}_1)}{\partial \theta_j}$$
$$= \left(\frac{1}{\sqrt{L_N}} \sum_{i=1}^{L_N} \frac{\partial f_s^g(\theta_0)}{\partial \theta_j} + \frac{1}{L_N} \sum_{s=1}^{L_N} \frac{\partial}{\partial \theta'} \left(\frac{\partial f_s^g(\theta_0)}{\partial \theta_j} \right) \sqrt{L_N}(\hat{\theta}_1 - \theta_0) \right)$$
$$\cdot (1 + o_p(1))$$
$$:= \left(\frac{1}{\sqrt{L_N}} \sum_{i=1}^{L_N} \frac{\partial f_s^g(\theta_0)}{\partial \theta_j} + Q(\theta_0)\sqrt{L_N}(\hat{\theta}_1 - \theta_0) \right) (1 + o_p(1))$$

for $j = 1, ..., d$ and $g = 1, ..., G$. This implies that

$$\sum_{s=1}^{L_N} \left(\frac{\partial f_s^g(\hat{\theta}_1)}{\partial \theta_j} - \frac{1}{N} \sum_{s=1}^{N} \frac{\partial f_s(\hat{\theta}_1)}{\partial \theta_j} \right)$$
$$= \sum_{s=1}^{L_N} \left(\frac{\partial f_s^g(\theta_0)}{\partial \theta_j} - \frac{1}{N} \sum_{s=1}^{N} \frac{\partial f_s(\theta_0)}{\partial \theta_j} \right) (1 + o_p(1)) \tag{1.60}$$

Combining these together, we have $\Upsilon(\hat{\theta}_1) = \Upsilon(\theta_0)(1 + o_p(1))$ from which we obtain the desired result

$$\widehat{\mathcal{E}}_N = \mathcal{E}_N \left(1 + o_p\left(1\right)\right). \tag{1.61}$$

Now, define the infeasible corrected variance

$$
\begin{aligned}
&\widehat{var}^{\mathbf{c},\inf}_{\hat{\Omega}^c(\hat{\theta}_1)}(\hat{\theta}^c_2) \\
&= \widehat{var}_{\hat{\Omega}^c(\hat{\theta}_1)}(\hat{\theta}^c_2) \\
&+ \mathcal{E}_N \widehat{var}_{\hat{\Omega}^c(\hat{\theta}_1)}(\hat{\theta}^c_{2N}) + \widehat{var}_{\hat{\Omega}^c(\hat{\theta}_1)}(\hat{\theta}^c_2)\mathcal{E}'_N + \mathcal{E}'_N \widehat{var}(\hat{\theta}_1)\widehat{\mathcal{E}}'_N
\end{aligned}
$$

and the corresponding infeasible Wald statistic

$$
\mathbb{F}^{\mathbf{c}}_{2,\hat{\Omega}^c(\hat{\theta}_1)}(\hat{\theta}^c_2) = (R\hat{\theta}^c_2 - r)' \left[R\widehat{var}^{\mathbf{c},\inf}_{\Omega^c_N(\hat{\theta}_1)}(\hat{\theta}^c_2)R' \right]^{-1} (R\hat{\theta}^c_2 - r)/p.
$$

The result in (1.61) implies

$$
F^{\mathbf{c}}_{2,\hat{\Omega}^c(\hat{\theta}_1)}(\hat{\theta}^c_2) = \mathbb{F}^{\mathbf{c}}_{2,\hat{\Omega}^c(\hat{\theta}_1)}(\hat{\theta}^c_2)(1 + o_p(1)).
$$

Also, $\mathcal{E}_N = o_p(1)$ and we have

$$
\widehat{var}^{\mathbf{c},\inf}_{\hat{\Omega}^c(\hat{\theta}_1)}(\hat{\theta}^c_2) = \widehat{var}^{\mathbf{c}}_{\hat{\Omega}^c(\hat{\theta}_1)}(\hat{\theta}^c_2)(1 + o_p(1)),
$$

and so

$$
\begin{aligned}
F^{c}_{2,\hat{\Omega}^c(\hat{\theta}_1)}(\hat{\theta}^c_2) &= \mathbb{F}^{\mathbf{c}}_{2,\hat{\Omega}^c(\hat{\theta}_1)}(\hat{\theta}^c_2) + o_p(1) \\
&= F_{2,\hat{\Omega}^c(\hat{\theta}_1)}(\hat{\theta}^c_2) + o_p(1).
\end{aligned}
$$

■

Chapter 2

Should We Go One Step Further? An Accurate Comparison of One-step and Two-step Procedures in a Generalized Method of Moments Framework

Abstract. According to the conventional asymptotic theory, the two-step Generalized Method of Moments (GMM) estimator and test perform at least as well as the one-step estimator and test in large samples. The conventional asymptotic theory, 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.

2.1 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 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

[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.

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 (1986) 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 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 et al. (2000) and Kiefer and Vogelsang (2002b, 2005) in the econometrics literature. For other studies, see, for example, Jansson)2004), Sun, Phillips and Jin (2008), Sun and Phillips (2009), Goncalves and Vogelsang (2011), and Zhang et al. (2013) in the time series setting; Bester et al. (2016) in the spatial setting; and Gonçalves (2011), Kim and Sun (2013), and Vogelsang (2012) in the panel data setting. 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 (2011a, 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 2.3 compares the two procedures from the perspective of point estimation. Section 2.4 compares them from the testing perspective. Section 2.5 extends the ideas to a general GMM framework. Section 2.6 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.

2.2 A Simple Overidentified GMM Framework

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

$$\begin{aligned} y_{1t} &= \theta_0 + u_{1t},\ y_{1t} \in \mathbb{R}^d, \\ y_{2t} &= u_{2t},\ y_{2t} \in \mathbb{R}^q \end{aligned} \tag{2.1}$$

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 :

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

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

[2] If

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

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

---

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.

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:

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

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

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

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

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

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 $\hat{\Omega}$, $Q_h(r,s)$ is a symmetric weighting function that depends on the smoothing parameter h. For conventional kernel LRV estimators, $Q_h(r,s) = k((r-s)/b)$ and we take $h = 1/b$. For the orthonormal series (OS) LRV estimators, $Q_h(r,s) = K^{-1}\sum_{j=1}^{K}\phi_j(r)\phi_j(s)$ and we take $h = K$, where $\{\phi_j(r)\}$ are orthonormal basis functions on $L^2[0,1]$ satisfying $\int_0^1 \phi_j(r)\,dr = 0$. 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 $\hat{\Omega}(\hat{u})$. For the location model, $\hat{\Omega}(\hat{u})$ is numerically identical to $\hat{\Omega}(u)$ 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.

2.3 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., $\sqrt{T}(\tilde{\theta}_{2T} - \hat{\theta}_{2T}) = o_p(1)$. As a direct consequence, we have

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

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 8 on the kernel function and basis functions.

Assumption 8 *(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 8 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 8 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 9 *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 $j = 0,1,...,J$ for every fixed J;

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

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

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 10 $\sum_{j=-\infty}^{\infty} \| Eu_t u_{t-j}' \| < \infty$.

Proposition 13 *Let Assumptions 8–10 hold. As $T \to \infty$ for a fixed $h > 0$, we have:*

(a) $\hat{\Omega} \overset{d}{\Longrightarrow} \Omega_\infty$ 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) $\sqrt{T}\left(\hat{\theta}_{2T} - \theta_0\right) \overset{d}{\Longrightarrow} \begin{pmatrix} I_d, & -\beta_\infty \end{pmatrix}\Omega_{1/2}B_m(1)$ 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 $\sqrt{T}(\hat{\theta}_{1T} - \theta_0) \overset{d}{\Longrightarrow} N(0, \Omega_{11})$ 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

$$\tilde{\beta}_\infty := \tilde{\beta}_\infty(h, d, q) := \tilde{\Omega}_{\infty,12}\tilde{\Omega}_{\infty,22}^{-1}, \tag{2.2}$$

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

$$\beta_\infty = \Omega_{1\cdot 2}^{1/2}\tilde{\beta}_\infty\Omega_{22}^{-1/2} + \Omega_{12}\Omega_{22}^{-1}. \tag{2.3}$$

Using this, we have:

$$\begin{aligned} \text{avar}(\hat{\theta}_{2T}) - \text{avar}(\hat{\theta}_{1T}) &= \underbrace{\text{avar}(\hat{\theta}_{2T}) - \text{avar}(\tilde{\theta}_{2T})}_{\text{cost}} - \underbrace{[\text{avar}(\hat{\theta}_{1T}) - \text{avar}(\tilde{\theta}_{2T})]}_{\text{benefit}} \\ &= \Omega_{1\cdot 2}^{1/2} E\tilde{\beta}_{\infty}\tilde{\beta}_{\infty}'(\Omega_{1\cdot 2}^{1/2})' - \Omega_{12}\Omega_{22}^{-1}\Omega_{21}. \end{aligned} \tag{2.4}$$

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 14 *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 15 *Let Assumptions 8–10 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 $\nu\left(\rho\rho'\right) = \nu\left(\Omega_{12}\Omega_{22}^{-1}\Omega_{21}\Omega_{11}^{-1}\right)$. The eigenvalues of $\rho\rho'$ are the squared long run correlation coefficients between $c_1'u_{1t}$ and $c_2'u_{2t}$ for some c_1 and c_2, 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 $\nu_{\max}\left(\rho\rho'\right) < g(h,q)$ 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 $\nu_{\min}\left(\rho\rho'\right) = 1$ and we have $\nu_{\min}\left(\rho\rho'\right) > g(h,q)$. In this case, it is worthwhile estimating the long run variance and using it to improve the efficiency $\hat{\theta}_{2T}$.

The two conditions $\nu_{\min}\left(\rho\rho'\right) > g(h,q)$ and $\nu_{\max}\left(\rho\rho'\right) < g(h,q)$ 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 $\nu_{\min}\left(\rho\rho'\right) > g(h,q)$ and $\nu_{\max}\left(\rho\rho'\right) \leq g(h,q)$ 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 $\nu_{\max}\left(\rho\rho'\right) >$

$g(h,q)$. 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 two conditions $\nu_{\min}(\rho\rho') > g(h,q)$ and $\nu_{\max}(\rho\rho') \leq g(h,q)$ becomes 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 2.1–2.3 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 16 *Let Assumptions 8–10 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 2.6. Figure 2.1 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 2.1, 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 (2.4). More specifically, the conditional asymptotic variance of $\hat{\theta}_{2T}$ is

$$\text{avar}(\hat{\theta}_{2T}|\tilde{\beta}_\infty) = V_2 = \Omega_{1\cdot 2}^{1/2}\tilde{\beta}_\infty\tilde{\beta}_\infty'(\Omega_{1\cdot 2}^{1/2})' + \Omega_{1\cdot 2} = 0.9(1+\tilde{\beta}_\infty^2). \tag{2.5}$$

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 $\hat{\theta}_{1T}$.

2.4 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:

$$\mathbb{T}_{1T} \;:=\; \frac{\sqrt{T}\left(R\hat{\theta}_{1T} - r\right)}{\sqrt{R\hat{\Omega}_{11}R'}} \tag{2.6}$$

$$\mathbb{T}_{2T} \;:=\; \frac{\sqrt{T}\left(R\hat{\theta}_{2T} - r\right)}{\sqrt{R\hat{\Omega}_{1\cdot 2}R'}}. \tag{2.7}$$

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

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

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 $\chi_p^2(\lambda^2)$ 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 17 *Let Assumptions 8–10 hold. As $T \to \infty$ for a fixed h, we have, under $H_1 : R\theta_0 = r + \delta_0/\sqrt{T}$:*
(a) $\mathbb{W}_{1T} \overset{d}{\Longrightarrow} \mathbb{W}_{1\infty}(\left\|\Lambda_1^{-1}\delta_0\right\|^2)$ *where*

$$\mathbb{W}_{1\infty}(\|\xi\|^2) = [B_p(1) + \xi]' C_{pp}^{-1} [B_p(1) + \xi] \text{ for } \xi \in \mathbb{R}^p. \tag{2.8}$$

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

$$\mathbb{W}_{2\infty}(\|\xi\|^2) = \left[B_p(1) - C_{pq}C_{qq}^{-1}B_q(1) + \xi\right]' D_{pp}^{-1} \left[B_p(1) - C_{pq}C_{qq}^{-1}B_q(1) + \xi\right]. \tag{2.9}$$

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

In Proposition 17, we use the notation $\mathbb{W}_{1\infty}(\|\xi\|^2)$, which implies that the right hand side of (2.8) depends on ξ only through $\|\xi\|^2$. This is true, because for any orthogonal matrix H :

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

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

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

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 (2.9) 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:

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

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 18 *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 18(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

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

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 18(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 parameter move in opposite directions. It is not straightforward to compare the two power functions. However, Proposition 18 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

$$\rho_R = \left(R\Omega_{11}R'\right)^{-1/2}\left(R\Omega_{12}\right)\Omega_{22}^{-1/2}, \tag{2.10}$$

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 18(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

$$\begin{aligned}
&\left\|\Lambda_2^{-1}\delta_0\right\|^2 - \tau(\left\|\Lambda_2^{-1}\delta_0\right\|^2)\left\|\Lambda_1^{-1}\delta_0\right\|^2 \\
= &\sum_{i=1}^{p}\left(\frac{1}{1-\nu_{i,R}} - \tau(\left\|\Lambda_2^{-1}\delta_0\right\|^2)\right)\left[a_{i,R}'\Lambda_1^{-1}\delta_0\right]^2 \\
= &\sum_{i=1}^{p}\frac{1}{1-\nu_{i,R}}\left(\nu_{i,R} - \frac{\tau(\left\|\Lambda_2^{-1}\delta_0\right\|^2)-1}{\tau(\left\|\Lambda_2^{-1}\delta_0\right\|^2)}\right)\left[a_{i,R}'\Lambda_1^{-1}\delta_0\right]^2\tau(\left\|\Lambda_2^{-1}\delta_0\right\|^2) \\
= &\sum_{i=1}^{p}\frac{1}{1-\nu_{i,R}}\left(\nu_{i,R} - f(\left\|\Lambda_2^{-1}\delta_0\right\|^2)\right)\left[a_{i,R}'\Lambda_1^{-1}\delta_0\right]^2\tau(\left\|\Lambda_2^{-1}\delta_0\right\|^2) \qquad (2.11)
\end{aligned}$$

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 19 *Let Assumptions 8–10 hold. Define*

$$\mathfrak{A}(\lambda_0) = \{\delta : \delta' (R\Omega_{1\cdot 2}R')^{-1} \delta = \lambda_0\}.$$

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 $\nu_{\max}(\rho_R \rho_R') < f(\lambda_0; h, p, q, \alpha)$, 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

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

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 19(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 of the value of δ_0 that characterizes the direction 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.

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 19 is in parallel with Proposition 15. 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 20 *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 20 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 20 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 2.4, 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.

2.5 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 $q = m - d$.

2.5.1 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:

$$\tilde{\theta}_{1T} \quad : \quad = \arg\min_{\theta\in\Theta} \breve{g}_T(\theta)'\breve{\Sigma}^{-1}\breve{g}_T(\theta), \tag{2.12}$$

$$\tilde{\theta}_{2T} \quad : \quad = \arg\min_{\theta\in\Theta} \breve{g}_T(\theta)'\breve{\Omega}^{-1}\breve{g}_T(\theta). \tag{2.13}$$

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

$$\breve{\Sigma}_{est}(\theta) \quad : \quad = \frac{1}{T}\sum_{t=1}^{T} \breve{f}(v_t,\theta)\breve{f}(v_t,\theta)', \tag{2.14}$$

$$\breve{\Omega}_{est}(\theta) \quad : \quad = \frac{1}{T}\sum_{s=1}^{T}\sum_{t=1}^{T} Q_h^*(\frac{s}{T},\frac{t}{T})\breve{f}(v_t,\theta)\breve{f}(v_s,\theta)'. \tag{2.15}$$

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:

$$\mathbb{W}_{1T} \quad : \quad = T(R\hat{\theta}_{1T} - r)'\left\{R\hat{\mathcal{V}}_{1T}R'\right\}^{-1}(R\hat{\theta}_{1T} - r), \tag{2.16}$$

$$\mathbb{W}_{2T} \quad : \quad = T(R\hat{\theta}_{2T} - r)'\left\{R\hat{\mathcal{V}}_{2T}R'\right\}^{-1}(R\hat{\theta}_{2T} - r),$$

where

$$\hat{\mathcal{V}}_{1T} = \left[\breve{G}_{1T}'\breve{\Sigma}_{est}^{-1}(\hat{\theta}_{1T})\breve{G}_{1T}\right]^{-1}\left[\breve{G}_{1T}'\breve{\Sigma}_{est}^{-1}(\hat{\theta}_{1T})\breve{\Omega}_{est}(\hat{\theta}_{1T})\breve{\Sigma}_{est}^{-1}(\hat{\theta}_{1T})\breve{G}_{1T}\right] \quad (2.17)$$

$$\times\left[\breve{G}_{1T}'\breve{\Sigma}_{est}^{-1}(\hat{\theta}_{1T})\breve{G}_{1T}\right]^{-1} \quad (2.18)$$

$$\hat{\mathcal{V}}_{2T} = \left[\breve{G}_{2T}\breve{\Omega}_{est}^{-1}(\hat{\theta}_{2T})\breve{G}_{2T}\right]^{-1}$$

and

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

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 11 *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 12 *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 13: $\breve{\Omega}_{\infty} = \breve{\Omega}_{1/2} \tilde{\Omega}_{\infty} \breve{\Omega}'_{1/2}$.

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 2.3 and 2.4. To confront this challenge, we let

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

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

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

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:

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

where

$$\Sigma_{1/2}^* = \begin{pmatrix} (\Sigma_{1\cdot 2}^*)^{1/2} & \Sigma_{12}^* (\Sigma_{22}^*)^{-1/2} \\ 0 & (\Sigma_{22}^*)^{1/2} \end{pmatrix}. \quad (2.19)$$

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 21 *Let Assumptions 8–12 hold with u_t replaced by $f(v_t,\theta_0)$ in Assumptions 9 and 10. Then as $T \to \infty$ for a fixed $h > 0$,*

$$(\Sigma_{1\cdot 2}^*)^{-1/2} AV'\sqrt{T}(\hat{\theta}_{1T} - \theta_0) = \frac{1}{\sqrt{T}} \sum_{t=1}^{T} f_1(v_t,\theta_0) + o_p(1) \overset{d}{\Longrightarrow} N(0,\Omega_{11}) \quad (2.20)$$

$$\begin{aligned} (\Sigma_{1\cdot 2}^*)^{-1/2} AV'\sqrt{T}(\hat{\theta}_{2T} - \theta_0) &= \frac{1}{\sqrt{T}} \sum_{t=1}^{T} [f_1(v_t,\theta_0) - \beta_\infty f_2(v_t,\theta_0)] + o_p(1) \quad (2.21) \\ &\overset{d}{\Longrightarrow} MN\left(0, \Omega_{11} - \Omega_{12}\beta_\infty' - \beta_\infty\Omega_{21} + \beta_\infty\Omega_{22}\beta_\infty'\right) \end{aligned}$$

where $\beta_\infty := \Omega_{\infty,12}\Omega_{\infty,22}^{-1}$ is the same as in Proposition 13.

Lemma 21 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 21 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 (2.20) and (2.21) 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

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

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

$$\tilde{R} = RVA^{-1}\left(\Sigma_{1\cdot 2}^{*}\right)^{1/2}, \tag{2.22}$$

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 (2.2). 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

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

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 (2.10).

Theorem 22 *Let the assumptions in Lemma 21 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 22(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 22 can also be applied to a general location model with a nonscalar error variance, in which case $\tilde{R} = R\left(\Sigma_{1\cdot 2}^{*}\right)^{1/2}$.

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 (2.16), 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

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

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 21, we can show that

$$(\Sigma_{1\cdot 2}^*)^{-1/2} AV'\sqrt{T}(\hat{\theta}_{aT} - \theta_0) = \frac{1}{\sqrt{T}}\sum_{t=1}^{T} \left[f_1(v_t, \theta_0) - \beta_a f_2(v_t, \theta_0)\right] + o_p(1). \quad (2.23)$$

The above representation is the same as that in (2.21) 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

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

Theorem 23 *Let the assumptions in Lemma 21 hold. Assume further that* $\breve{W}_T(\hat{\theta}_{0T}) \overset{p}{\to} \breve{W}$*, 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 23 is entirely analogous to Theorem 22. 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 23(a) and (b) gives the conditions under which $\hat{\theta}_{2T}$ is asymptotically more 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 23(d), the test based on $R\hat{\theta}_{2T}$ could be more powerful than that based on $R\hat{\theta}_{aT}$ in large samples.

2.6 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 (2.1) 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:

$$u_{1t}^{*i} = \psi u_{1t-1}^{*i} + \frac{\gamma}{\sqrt{q}} \sum_{j=1}^{q} u_{2t-1}^{*j} + e_{1t}^{*i} \text{ for } i = 1, ..., d \tag{2.24}$$

$$u_{2t}^{*i} = \psi u_{2t-1}^{*i} + e_{2t}^{*i} \text{ for } i = 1, ..., q$$

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

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

and

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

Let $u_{1t} = (\Sigma_{1\cdot 2}^*)^{-1/2} [u_{1t}^* - \Sigma_{12}^* (\Sigma_{22}^*)^{-1} u_{2t}^*]$ 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

$$\rho\rho' = \left(d + \frac{(1-\psi^2)^2}{\gamma^2}\right)^{-1} J_{d,d}. \tag{2.25}$$

So the maximum eigenvalue of $\rho\rho'$ is given by $\nu_{\max}(\rho\rho') = [1 + (1-\psi^2)^2/(d\gamma^2)]^{-1}$, 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^* :

$$u_{1t}^{*i} = \psi u_{1t-1}^{*i} + e_{1t}^{*i} + \frac{\gamma}{\sqrt{q}} \sum_{j=1}^{q} e_{2,t-1}^{*j} \text{ for } i = 1, ..., d \tag{2.26}$$

$$u_{2t}^{*i} = \psi u_{2t-1}^{*i} + e_{2t}^{*i} \text{ for } i = 1, ..., q$$

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

$$\rho\rho' = \left(d + \frac{1}{(1-\psi)^2 \gamma^2}\right)^{-1} J_{d,d}, \tag{2.27}$$

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

Under the VARMA(1,1) design, a working weighting matrix $\breve{W}(\hat{\theta}_{aT})$ using VAR(1) is misspecified and it is not hard to obtain the probability limit of $\breve{W}(\hat{\theta}_{aT})$ 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):

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

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 $10,000$.

2.6.1 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 (2.25) or (2.27).

According to Proposition 15, 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 2.5~2.6 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 $\gamma^2 = (\rho\rho'(1-\psi^2)^2) \,/\, (d(1-\rho\rho')$ 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 13: $\hat{\theta}_{2T}$ 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 2.5 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 16.

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 2.1.

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 2.7∼2.8 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 23(a) and (b).

2.6.2 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 (2.25) and (2.27) 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:

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

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 2.9~2.16 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 $\mathbb{W}_{1\infty}^{\alpha}$ and $\mathbb{W}_{2\infty}^{\alpha}$ 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 c) 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 2.11~2.16.

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 2.4 and 2.5.

Figures 2.2∼2.3 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 19. For example, with $K = 14$ and $p = 1$, the power curves in Figure 2.2 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 19 and Table 2.4 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 **??**∼2.3 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 2.4∼2.7 for the cases of kernel LRV estimators deliver the same qualitative messages.

2.6.3 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 (2.15).

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

$$\hat{\Sigma}^*_{1/2} = \begin{pmatrix} \left(\hat{\Sigma}^*_{1\cdot 2}\right)^{1/2} & \hat{\Sigma}^*_{12}\left(\hat{\Sigma}^*_{22}\right)^{-1/2} \\ 0 & \left(\hat{\Sigma}^*_{22}\right)^{1/2} \end{pmatrix}. \tag{2.28}$$

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.

2.7 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.

2.8 Acknowledgements

Chapter 2 is coauthored with Yixiao Sun, and in full, has been submitted for publication of material. 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.

2.9 Figures and Tables

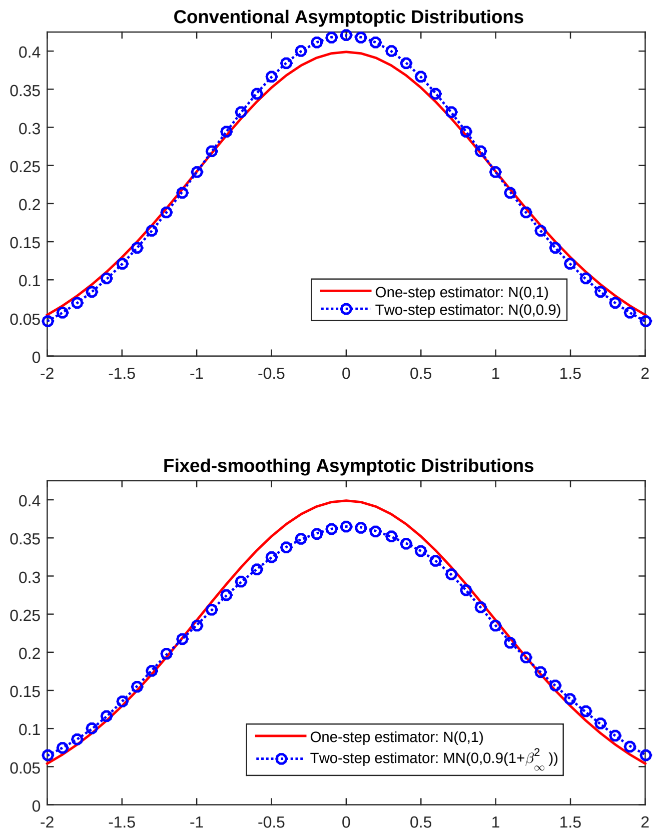

Figure 2.1: Limiting distributions of $\hat{\theta}_{1T}$ and $\hat{\theta}_{2T}$ based on the OS LRV estimator with $K = 4$.

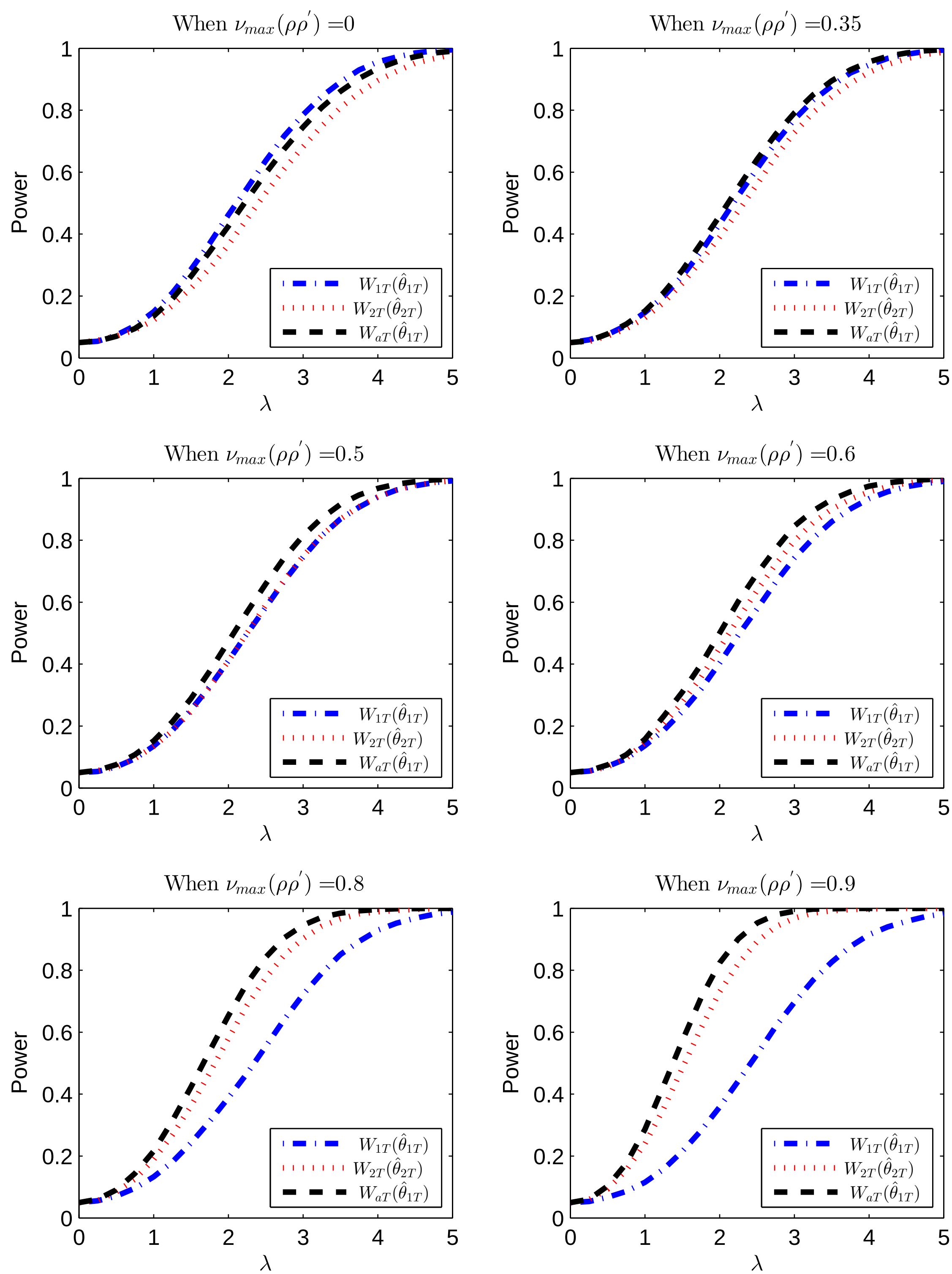

Figure 2.2: Size-adjusted power of the three tests based on the OS LRV estimator under VAR(1) error with $p = 1$, $q = 3$, $\psi = 0.75$, $T = 200$, and $K = 14$.

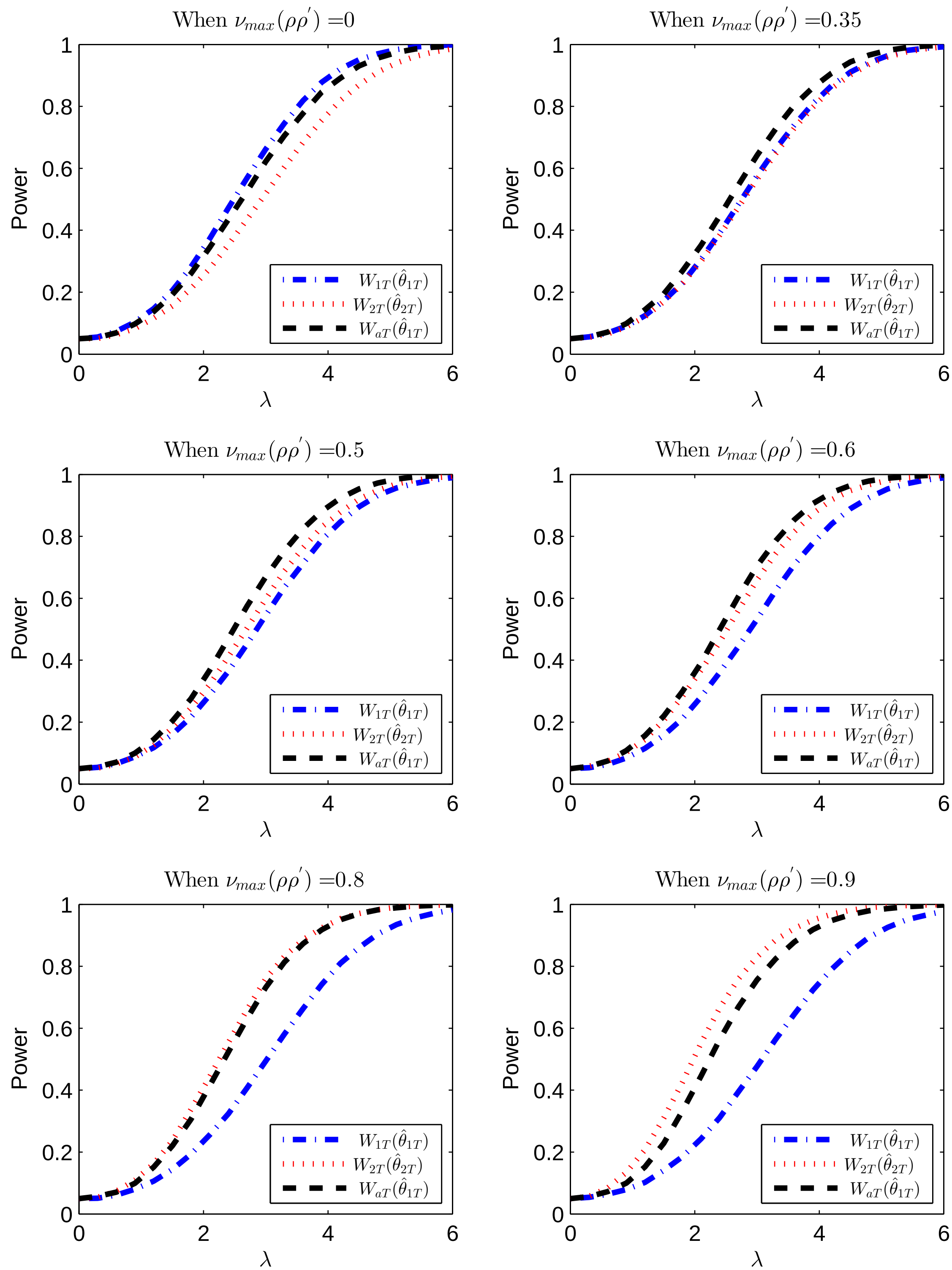

Figure 2.3: Size-adjusted power of the three tests based on the OS LRV estimator under VARMA(1,1) error with $p = 2$, $q = 3$, $\psi = 0.75$, $T = 200$, and $K = 14$.

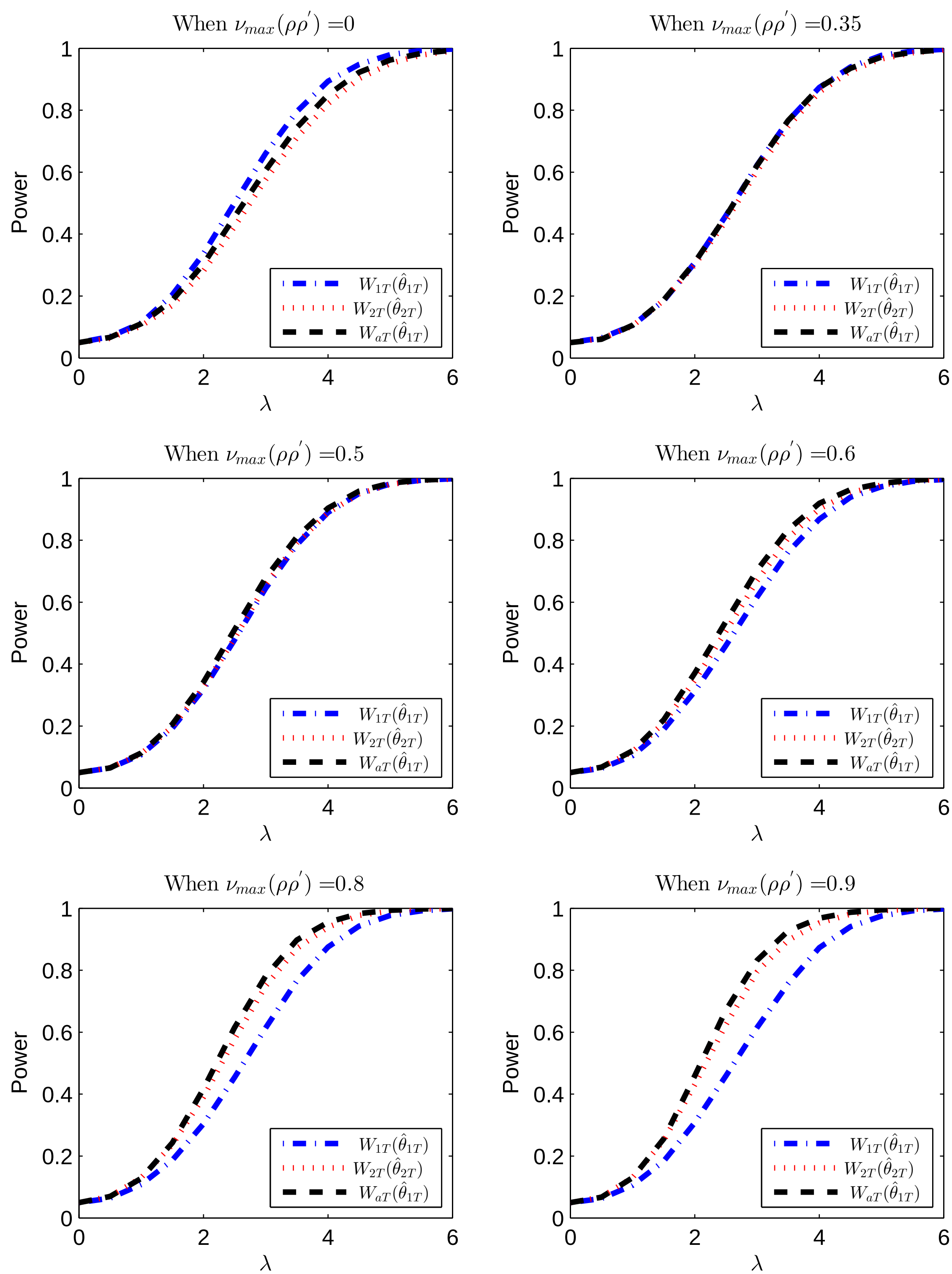

Figure 2.4: Size-adjusted power of the three tests based on the Bartlett LRV estimator under VAR(1) error with $p = 2$, $q = 3$, $\psi = 0.75$, $T = 200$, and $b = 0.078$.

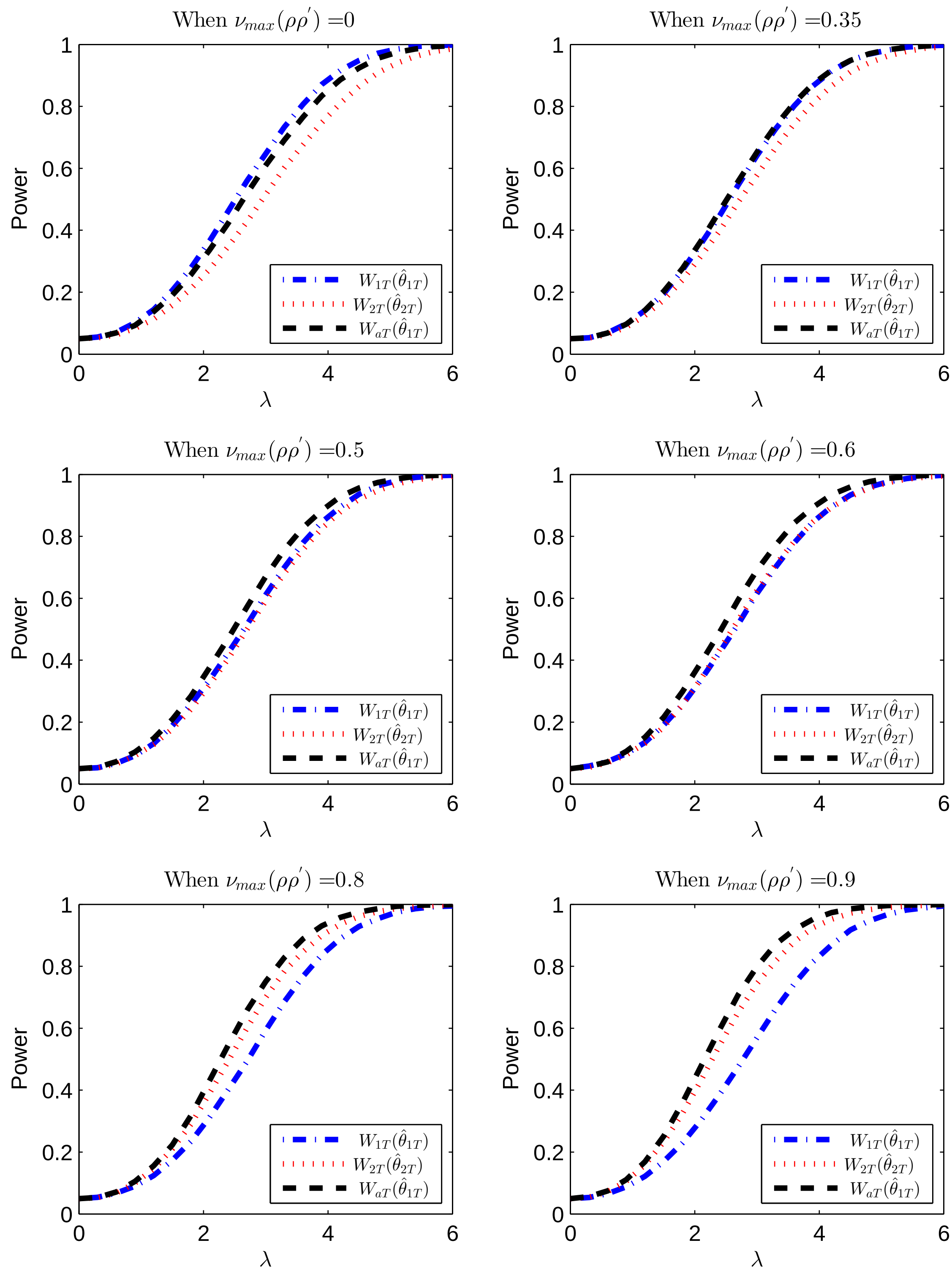

Figure 2.5: Size-adjusted power of the three tests based on the OS LRV estimator under VAR(1) error with $p = 2$, $q = 3$, $\psi = 0.75$, $T = 200$, and $K = 14$.

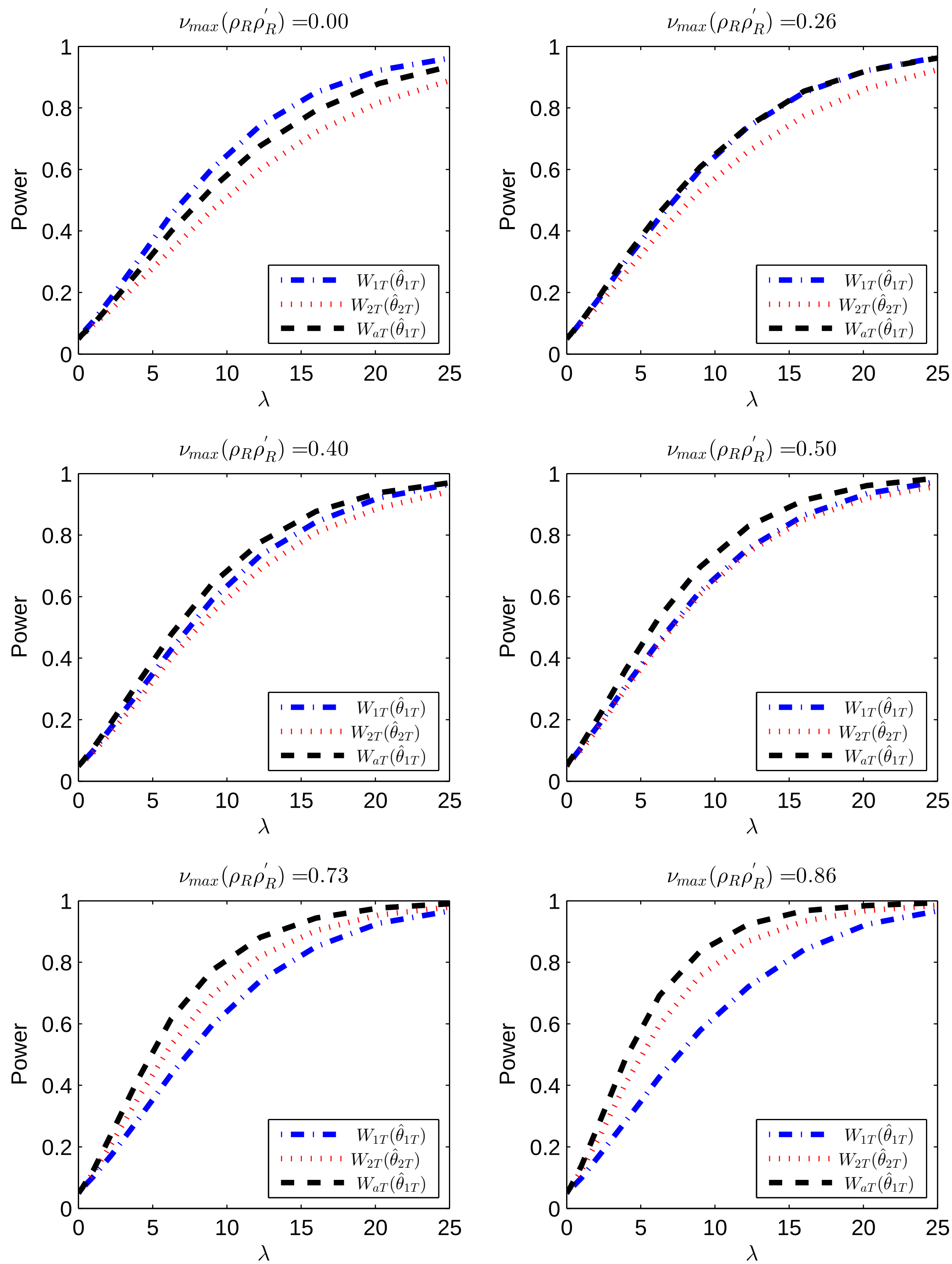

Figure 2.6: Size-adjusted power of the three tests based on the Parzen LRV estimator under VAR(1) error with $p = 2$, $q = 3$, $\psi = 0.75$, $T = 200$, and $b = 0.16$.

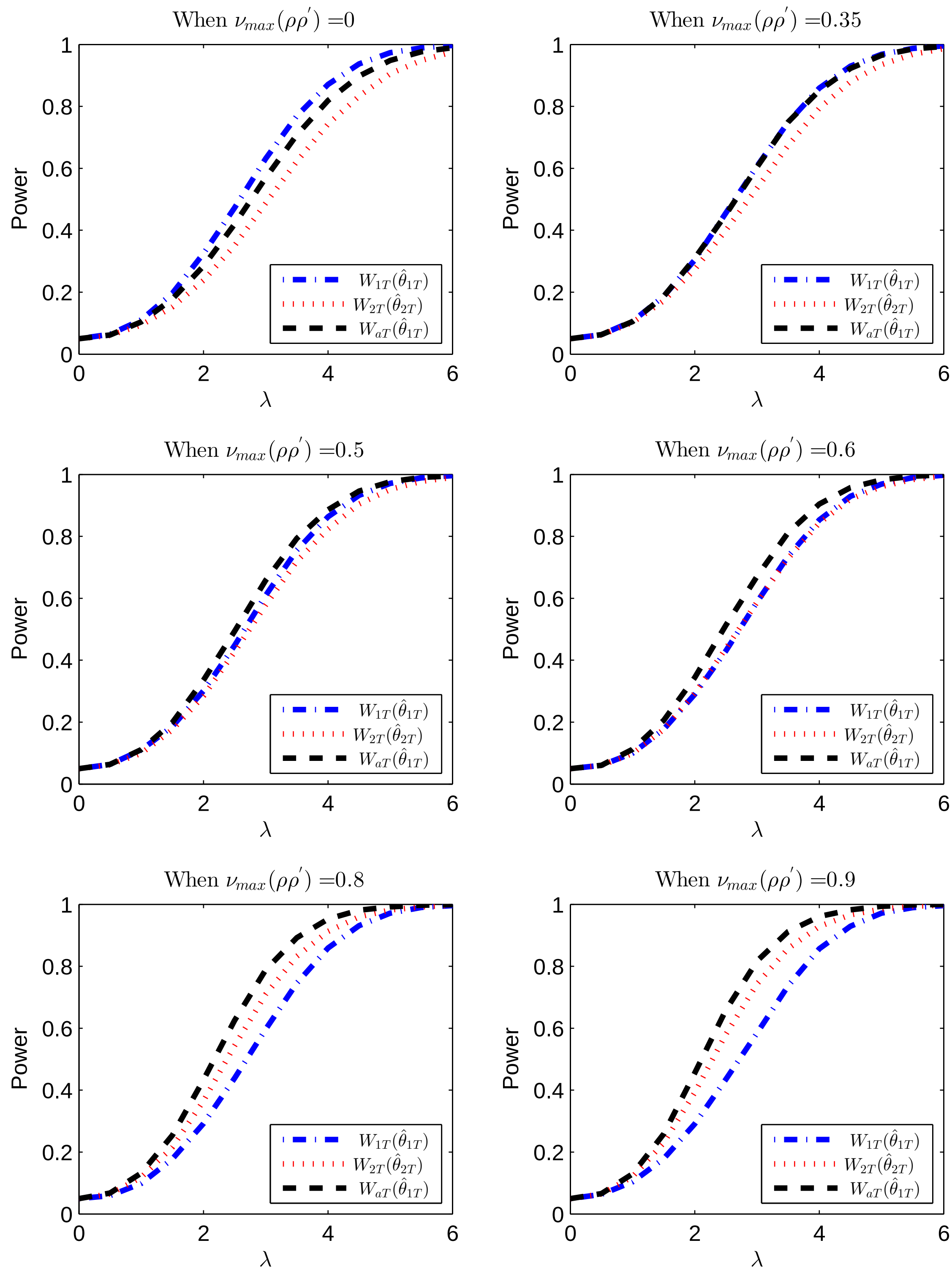

Figure 2.7: Size-adjusted power of the three tests based on the QS LRV estimator under VAR(1) error with $p = 2$, $q = 3$, $\psi = 0.75$, $T = 200$, and $b = 0.079$.

Table 2.1: 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

Note: $h=1/b$ 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 2.2: 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

Note: See notes to Table 2.1

Table 2.3: 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

Note: See notes to Table 2.1.

Table 2.4: 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	$q=1$	$q=2$	$q=3$	$q=1$	$q=2$	$q=3$	$q=1$	$q=2$	$q=3$
8	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
	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
10	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
	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
12	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
	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
14	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
	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

Note: 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 2.5: 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 $q = 3$.

$\nu_{\max}(\rho\rho')$	Var($\hat{\theta}_{1T}$)	Var($\hat{\theta}_{2T}$)				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 2.6: 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 $q = 4$.

$\nu_{\max}(\rho\rho')$	Var($\hat{\theta}_{1T}$)	Var($\hat{\theta}_{2T}$)				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 2.7: 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 $q = 3$

$\nu_{\max}(\rho\rho')$	$\text{Var}(\hat{\theta}_{1T})$	$\text{Var}(\hat{\theta}_{2T})$				$\nu_{\max}(\rho_a\rho_a')$	$\text{Var}(\hat{\theta}_{aT})$
·	·	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	0.000	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 2.8: 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 $q = 4$.

$\nu_{\max}(\rho\rho')$	$\text{Var}(\hat{\theta}_{1T})$	$\text{Var}(\hat{\theta}_{2T})$				$\nu_{\max}(\rho_a\rho_a')$	$\text{Var}(\hat{\theta}_{aT})$
·	·	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 2.9: 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 $q = 3$						
·	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 $q = 3$						
·	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

Note: "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 2.10: 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 $q = 3$						
.	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

Note: See notes to Table 2.9.

Table 2.11: 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 $q = 3$						
·	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 $q = 3$						
·	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

Note: See notes to Table 2.9.

Table 2.12: 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 $q = 3$						
.	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 $q = 3$						
.	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

Note: See notes to Table 2.9.

Table 2.13: 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 $q = 3$						
.	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 $q = 3$						
.	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

Note: See notes to Table 2.9.

Table 2.14: 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 $q = 3$						
·	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 $q = 3$						
·	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

Note: See notes to Table 2.9.

Table 2.15: 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 $q = 3$						
.	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 $q = 3$						
.	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

Note: See notes to Table 2.9.

Table 2.16: 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 $q = 3$						
.	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 $q = 3$						
.	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

Note: See notes to Table 2.9.

2.10 Appendix of Proofs

Proof of Proposition 13. 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 14. For any $a \in \mathbb{R}^d$, we have

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

where

$$\kappa(h,q) = Etr\left(\int_0^1\int_0^1 Q_h^*\left(r,s\right)dB_q\left(r\right)dB_q'\left(s\right)\right)^{-2} \cdot \left[\int_0^1\int_0^1\left(\int_0^1 Q_h^*\left(r,\tau\right)Q_h^*\left(\tau,s\right)d\tau\right)dB_q\left(r\right)dB_q'\left(s\right)\right].$$

So

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

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

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

■

Proof of Proposition 15. Using (2.4) and Lemma 14, we have

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

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 $Q_\rho' g(h,q)Q_\rho >$

Λ_ρ, 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 $\text{avar}(\hat{\theta}_{2T}) > \text{avar}(\hat{\theta}_{1T})$ if and only if $\nu_{\max}(\rho\rho') < g(h,q)$. Similarly, we can prove that $\text{avar}(\hat{\theta}_{2T}) < \text{avar}(\hat{\theta}_{1T})$ if and only if $\nu_{\min}(\rho\rho') > g(h,q)$. ■

Proof of Corollary 16. 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 (**??**), 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 15. ■

Proof of Proposition 17. 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 13(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 $C_{dd} \perp B_d(1)$. 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

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

as desired. ■

Proof of Proposition 18.

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 $P\left(\chi_p^2(\delta^2) > x\right)$ 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_{p+2j}^2 is a (central) chi-squared variate with degrees of freedom $p+2j$, we have

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

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 $P\left(\chi_p^2(\delta^2) > x\right)$ in δ^2, we have

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

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 $D_{pp}^{-1} = \sum_{i=1}^{p} \lambda_{Di} d_i d_i'$ 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

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

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 $x > 0$,

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

Using the above argument repeatedly, we have

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

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

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

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

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

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

$$
\begin{aligned}
& P\left[\mathbb{W}_{1\infty}\left(\|\xi_1\|^2\right) > x\right] \\
& < P\left\{\lambda_{C1}\left[c_1' B_p(1) + \|\xi_2\| c_1' e_p\right]^2 + \lambda_{C2}\left[c_2' B_p(1) + \|\xi_2\| c_2' e_p\right]^2 \right. \\
& \left. + \sum_{i=3}^{p} \lambda_{Ci}\left[c_i' B_p(1) + \|\xi_1\| c_i' e_p\right]^2 > x\right\} \\
& < P\left\{\sum_{i=1}^{p} \lambda_{Ci}\left[c_i' B_p(1) + \|\xi_2\| c_i' e_p\right]^2 > x\right\} \\
& = P\left\{\left[B_p(1) + \xi_2\right]' C_{pp}^{-1}\left[B_p(1) + \xi_2\right] > x\right\} = P\left[\mathbb{W}_{1\infty}\left(\|\xi_2\|^2\right) > x\right]
\end{aligned}
$$

as desired.

Part (c). We note that

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

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 19. We prove part (b) only as part (a) can be proved using the same argument. Using (2.11), we have, for $\lambda_0=\left\|\Lambda_2^{-1}\delta_0\right\|^2$:

$$\begin{aligned}
&\left\|\Lambda_2^{-1}\delta_0\right\|^2-\tau(\lambda_0)\left\|\Lambda_1^{-1}\delta_0\right\|^2\\
=&\ \tau(\lambda_0)\sum_{i=1}^{p}\frac{1}{1-\nu_{i,R}}\left[\nu_{i,R}-f(\lambda_0)\right]\left(a_{i,R}'\Lambda_1^{-1}\delta_0\right)^2\\
=&\ \tau(\lambda_0)\left\|\Lambda_1^{-1}\delta_0\right\|^2\sum_{i=1}^{p}\frac{1}{1-\nu_{i,R}}\left[\nu_{i,R}-f(\lambda_0)\right]\left\langle a_{i,R},\frac{\Lambda_1^{-1}\delta_0}{\left\|\Lambda_1^{-1}\delta_0\right\|}\right\rangle^2, \quad (2.29)
\end{aligned}$$

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

$$\left\|\Lambda_2^{-1}\delta_0\right\|^2 - \tau(\lambda_0)\left\|\Lambda_1^{-1}\delta_0\right\|^2 = \frac{\left\|\Lambda_1^{-1}\delta_0\right\|^2}{1-\nu_{i^*,R}}\left[\nu_{i^*,R} - f(\lambda_0)\right]\tau(\lambda_0) \leq 0, \qquad (2.30)$$

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 20. 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:

$$\underset{(p+q)\times 1}{Y} = \begin{pmatrix} \underset{(p\times 1)}{Y_1} \\ \underset{(q\times 1)}{Y_2} \end{pmatrix} \sim N_{p+q}(\mu, \Omega) \text{ with } \mu = \begin{pmatrix} \underset{(p\times 1)}{\delta_0} \\ \underset{(q\times 1)}{0} \end{pmatrix}, \Omega = \begin{pmatrix} \underset{(p\times p)}{\Omega_{11}} & \underset{(p\times q)}{0} \\ \underset{(q\times p)}{0} & \underset{(q\times q)}{\Omega_{22}} \end{pmatrix}$$

$$\underset{(p+q)\times(p+q)}{S} = \begin{pmatrix} \underset{(p\times p)}{S_{11}} & \underset{(p\times q)}{S_{12}} \\ \underset{(q\times p)}{S_{21}} & \underset{(q\times q)}{S_{22}} \end{pmatrix} \sim \frac{\mathcal{W}_{p+q}(K, \Omega)}{K}$$

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) andMarden and Perlman (1980).

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

$$\begin{aligned} & f(Y, S|\delta_0, \Omega_{11}, \Omega_{22}) \\ = \ & \alpha(\delta_0, \Omega_{11}, \Omega_{22}) h(S) \\ \cdot \ & \exp\left\{-\frac{1}{2} tr\left[\Omega_{11}^{-1}(Y_1 Y_1' + KS_{11}) + \Omega_{22}^{-1}(Y_2 Y_2' + KS_{22})\right] + Y_1'\Omega_{11}^{-1}\delta_0\right\} \end{aligned}$$

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

$$\Pi := (Y_1, S_{11}, Y_2 Y_2' + KS_{22})$$

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 $Y_2 Y_2' + KS_{22} \sim \mathcal{W}_q(K + 1, \Omega_{22})$ 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

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

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

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

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

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

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 et al. (1986)). 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 21. 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

$$\begin{aligned} \breve{G}'\breve{M}^{-1}\breve{G} &= V\Xi'(U'\breve{M}U)^{-1}\Xi V' \\ &= VA\begin{pmatrix} I_d, & O \end{pmatrix}(M^*)^{-1}\begin{pmatrix} I_d, & O \end{pmatrix}' AV' \\ &= VA\begin{pmatrix} I_d, & O \end{pmatrix}(\Sigma_{1/2}^{*-1})'\left[\Sigma_{1/2}^{*-1}M^*(\Sigma_{1/2}^{*-1})'\right]^{-1}\Sigma_{1/2}^{*-1}\begin{pmatrix} I_d, & O \end{pmatrix}' AV' \\ &= VA\begin{pmatrix} I_d, & O \end{pmatrix}(\Sigma_{1/2}^{*-1})'M^{-1}\Sigma_{1/2}^{*-1}\begin{pmatrix} I_d, & O \end{pmatrix}' AV' \\ &= VA\,(\Sigma_{1\cdot 2}^{*})^{-1/2}\begin{pmatrix} I_d, & O \end{pmatrix}M^{-1}\left[VA\,(\Sigma_{1\cdot 2}^{*})^{-1/2}\begin{pmatrix} I_d, & O \end{pmatrix}\right]' \\ &= VA\,(\Sigma_{1\cdot 2}^{*})^{-1/2}\,M_{1\cdot 2}^{-1}\,(\Sigma_{1\cdot 2}^{*})^{-1/2}\,AV', \end{aligned} \tag{2.31}$$

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

$$\begin{aligned}
&\frac{1}{\sqrt{T}}\sum_{t=1}^{T}\left[\breve{G}'\breve{M}^{-1}\breve{G}\right]^{-1}\breve{G}'\breve{M}^{-1}\breve{f}(v_t,\theta_0)\\
=\;&\frac{1}{\sqrt{T}}\sum_{t=1}^{T}\left[VA\left(\Sigma_{1\cdot 2}^{*}\right)^{-1/2}M_{1\cdot 2}^{-1}\left(\Sigma_{1\cdot 2}^{*}\right)^{-1/2}AV'\right]^{-1}\left[VA\left(\Sigma_{1\cdot 2}^{*}\right)^{-1/2}M_{1\cdot 2}^{-1}\right]\\
&\times\left[f_1\left(v_t,\theta_0\right)-M_{12}M_{22}^{-1}f_2\left(v_t,\theta_0\right)\right]\\
=\;&\frac{1}{\sqrt{T}}\sum_{t=1}^{T}VA^{-1}\left(\Sigma_{1\cdot 2}^{*}\right)^{1/2}\left[f_1\left(v_t,\theta_0\right)-M_{12}M_{22}^{-1}f_2\left(v_t,\theta_0\right)\right]. \qquad (2.32)
\end{aligned}$$

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

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

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

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

It then follows that

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

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

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

Using this, we have

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

It then follows that

$$\begin{aligned}\left(\Sigma_{1\cdot 2}^{*}\right)^{-1/2}AV'\sqrt{T}(\hat{\theta}_{2T}-\theta_0) &= \frac{1}{\sqrt{T}}\sum_{t=1}^{T}\left[f_1(v_t,\theta_0)-\beta_{\infty}f_2(v_t,\theta_0)\right]+o_p(1) \qquad (2.33)\\ &\overset{d}{\Longrightarrow} MN\left(0,\Omega_{11}-\Omega_{12}\beta_{\infty}'-\beta_{\infty}\Omega_{21}+\beta_{\infty}\Omega_{22}\beta_{\infty}'\right).\end{aligned}$$

■

Proof of Theorem 22. 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 $\hat{\theta}_{1T}^{R}$ and two-step estimator $\hat{\theta}_{2T}^{R}$ of θ_0^{R} in the following simple location model:

$$\begin{cases} y_{1t}^{R}=\theta_0^{R}+u_{1t}^{R}\in\mathbb{R}^{p}\\ y_{2t}=u_{2t}\in\mathbb{R}^{q}\end{cases} \qquad (2.34)$$

where

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

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

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

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

So the above location model has exactly the same form as the model in Section 2.3. We can invoke Proposition 15 to complete the proof.

The long run canonical correlation coefficients between u_{1t}^R and u_{2t} are the same as those between $\tilde{R}u_{1t}$ and u_{2t}. This follows because u_{1t}^R 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_{1t}^R 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 15.

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 (2.34). We use Proposition 19 to complete the proof. For the above location model, the asymptotic variance of the infeasible two-step GMM estimator is

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

In addition, the local alternative 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 19 is given by

$$\begin{aligned} \mathfrak{A}_{loc}(\lambda_0) &= \left\{\delta : \left[(\tilde{R}\tilde{R}')^{-1/2}\delta\right]' \left(\Omega_{1\cdot 2}^R\right)^{-1}\left[(\tilde{R}\tilde{R}')^{-1/2}\delta\right] = \lambda_0\right\} \\ &= \left\{\delta : \delta'(\tilde{R}\Omega_{1\cdot 2}\tilde{R}')^{-1}\delta = \lambda_0\right\}. \end{aligned} \quad (2.35)$$

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

Using (2.31) with $\breve{M}^{-1} = \breve{\Omega}^{-1}$, we have $M = \Omega$ and so

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

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

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

So the set of δ_0's considered in the theorem is exactly the same as that given in (2.35). ■

Proof of Theorem 23. The theorem is similar to Theorem 22. 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 $\mathbb{W}_{aT} \overset{d}{\Longrightarrow} \mathbb{W}_{1\infty}(||\mathcal{V}^{-1/2}_{a,R}\delta_0||^2)$ where

$$\begin{aligned} \mathcal{V}_{a,R} &= R(\breve{G}'\breve{W}^{-1}\breve{G})^{-1}\breve{G}'\breve{W}^{-1}\breve{\Omega}\breve{W}^{-1}\breve{G}(\breve{G}'\breve{W}^{-1}\breve{G})^{-1}R' \\ &= RVA^{-1}\left(\Sigma^*_{1\cdot 2}\right)^{1/2}\left(I_d, -\beta_a\right)\Omega\begin{pmatrix} I_d \\ -\beta_a' \end{pmatrix}\left[RVA^{-1}\left(\Sigma^*_{1\cdot 2}\right)^{1/2}\right]'. \quad (2.36) \end{aligned}$$

Similarly, we have

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

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 $R\sqrt{T}\left(\tilde{\theta}_{2T}-\theta_0\right)$ 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

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

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

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

which has the same form as the representation given in (2.29). The rest of the proof is then identical to the proof of Proposition 19 and is omitted here. ■

Chapter 3

Asymptotic F and t Tests in an Efficient GMM Setting

Abstract. This paper considers two-step efficient GMM estimation and inference where the weighting matrix and asymptotic variance matrix are based on the series long run variance estimator. We propose a simple and easy-to-implement modification to the trinity of test statistics in the two-step efficient GMM setting and show that the modified test statistics are all asymptotically F distributed under the so-called fixed-smoothing asymptotics. The modification is multiplicative and involves the J statistic for testing over-identifying restrictions. This leads to convenient asymptotic F tests that use standard F critical values. Simulation shows that, in terms of both size and power, the asymptotic F tests perform as well as the nonstandard tests proposed recently by Sun (2014b) in finite samples. But the F tests are more appealing as the critical values are readily available from standard statistical tables. Compared to the conventional chi-square tests, the F tests are as powerful, but are much more accurate in size.

3.1 Introduction

This paper considers the optimal two-step GMM estimator and the associated tests in a time series setting. In the presence of nonparametric temporal dependence, the optimal weighting matrix is the inverted long run variance (LRV) of the moment process. To implement the two-step GMM method, we often estimate the LRV using the nonparametric kernel or series method. Given the nonparametric nature of the LRV estimator, there is a high variation in the weighting matrix with consequent effects on the two-step point estimator and the associated tests. Recently Sun (2014b) employs the fixed-smoothing asymptotics and establishes a new asymptotic approximation that captures the estimation uncertainty in the LRV estimator. Under the fixed-smoothing asymptotics, the point estimator is asymptotically mixed normal and the test statistics converge to a nonstandard distribution. In the case of series LRV estimation, Sun (2014b) shows that the nonstandard limiting distribution can be approximated by a noncentral F distribution.

In this paper, we follow Sun (2014b) but focus on the series LRV estimator. We modify the usual test statistics, including the Wald statistic, the quasi LR statistic, and the LM statistic and show that the modified test statistics are all asymptotically standard F distributed. The standard F distribution is the exact limiting distribution. No additional approximation is needed. This is in contrast to Sun (2014b) where the noncentral F distribution is an approximation to the fixed-smoothing limiting distribution. The standard F distribution is more accessible than the noncentral F distribution, as standard F critical values are readily available from standard statistical tables.

The modification involves the usual J statistic for testing overidentifying restrictions. The modified test statistics are scaled versions of the original test statistics with the scaling factor depending on the J statistic. So the modification is very easy to implement. To understand the modification, we cast the two-step

GMM estimation and inference into OLS estimation and inference in a classical normal linear regression (CNLR). We show that the modified Wald statistic in the GMM framework is exactly the usual Wald statistic constructed in the standard way in the CNLR framework. Our proposed asymptotic F tests, which are based on the modified test statistics and use the standard F approximation, can be regarded as conditional tests conditioning on the J statistic. The conditioning argument is entirely analogous to that used in the linear regression model with stochastic regressors that are independent of the regression error.

Monte Carlo simulations show that our proposed asymptotic F tests are as accurate in size as the corresponding nonstandard tests of Sun (2014b). They are also as powerful as the latter tests. So there is no power loss in using the asymptotic F tests. Like the nonstandard tests of Sun (2014b), the asymptotic F tests are much more accurate in size than the usual chi-square tests without any power sacrifice. Given the convenience of the standard F approximation, we recommend the asymptotic F tests for practical use.

The paper contributes to a growing body of literature on the fixed-smoothing asymptotics. For kernel LRV estimators such as the Newey-West estimator (Newey and West (1987)), the fixed-smoothing asymptotics is the so-called the fixed-b asymptotics first studied by Vogelsang (2002a, 2002b, 2005) in the econometrics literature. Subsequent research includes Jansson (2004), Sun, Phillips, Jin (2008), Sun and Phillips (2009), Gonçlaves and Vogelsang (2011) and among others. Papers that are most closely related to this paper are those that use the series LRV estimators. In this case, the fixed-smoothing asymptotics is the so-called fixed-K asymptotics. Some examples of these papers are Phillips (2005), Müller (2007), Sun (2011, 2013, 2014a&b), and Sun and Kim (2012).

In the case of series LRV estimation, the F limit theory has been established in Sun (2011) for trend regression, Sun (2013) for stationary moment processes, and Sun (2014c) for highly persistent moment processes. See also Sun and Kim (2012, 2015) for the J test and the Wald test in the spatial setting. All these papers focus on the first-step GMM estimator or OLS estimator. This paper is the first to establish the F limit theory for the trinity of test statistics in a two-step efficient GMM framework. This is not trivial, as the asymptotic pivotality of these statistics under the fixed-smoothing asymptotics was not established until very recently in Sun (2014b).

The rest of the paper is organized as follows. Section 3.2 presents the basic setting and introduces the modified test statistics. Section 3.3 establishes the fixed-smoothing asymptotics of the modified test statistics and develops the asymptotic F and t tests. Section 3.4 casts the GMM estimator as an OLS estimator in a regression setting and shows that the modified Wald statistic is the usual Wald statistic in a CNLR model. The next section reports simulation evidence. The last section concludes. Proofs are given in the appendix.

3.2 Two-step GMM Estimation and Testing

We consider the standard GMM setting with moment conditions

$$Ef\left(v_t, \theta_0\right) = 0,\ \ t = 1, 2, \ldots, T, \tag{3.1}$$

where v_t is the vector of observations at time t, $\theta_0 \in \Theta \subseteq \mathbb{R}^d$ is the parameter of interest, and $f\left(v_t, \theta\right)$ is the $m \times 1$ vector of moment conditions that are twice continuously differentiable. We assume that $Ef\left(v_t, \theta\right) = 0$ if and only if $\theta = \theta_0$ so that θ_0 is point identified. The model may be overidentified with the degree of overidentification $q = m - d \geq 0$. We allow $\{f\left(v_t, \theta_0\right)\}$ to have autocorrelation of unknown forms.

Define

$$g_t\left(\theta\right) = \frac{1}{T}\sum_{j=1}^{t} f(v_j, \theta),$$

then the GMM estimator of θ_0 is given by

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

where W_T is a positive definite weighting matrix. The initial first-step GMM estimator can be obtained by choosing W_T to be a matrix $W_{o,T}$ that does not depend on any unknown parameter. This gives rise to

$$\tilde{\theta}_T = \arg\min_{\theta\in\Theta} g_T(\theta)' W_{o,T}^{-1} g_T(\theta).$$

Here $W_{o,T}$ may depend on the sample size T but we assume that $W_{o,T} \overset{p}{\to} W_{o,\infty}$, a matrix that is positive definite almost surely.

With the first step estimator $\tilde{\theta}_T$, we can construct the optimal weighting matrix W_T, which is the asymptotic variance matrix of $\sqrt{T} g_T(\theta_0)$. See Hansen (1982). Most, if not all, estimators of the asymptotic variance take the following form

$$W_T\left(\tilde{\theta}_T\right) = \frac{1}{T}\sum_{t=1}^{T}\sum_{s=1}^{T} Q_h\left(\frac{t}{T},\frac{s}{T}\right)\left(f(v_t,\tilde{\theta}_T) - \frac{1}{T}\sum_{\tau=1}^{T} f(v_\tau,\tilde{\theta}_T)\right) \cdot \left(f(v_s,\tilde{\theta}_T) - \frac{1}{T}\sum_{\tau=1}^{T} f(v_\tau,\tilde{\theta}_T)\right)',$$

where $Q_h(r,s)$ is a symmetric weighting function that depends on the smoothing parameter h. In this paper, we focus on the series LRV estimator with

$$Q_K(r,s) = \frac{1}{K}\sum_{j=1}^{K} \Phi_j(r)\Phi_j(s),$$

where $\{\Phi_j(r)\}$ are orthonormal basis functions on $L^2[0,1]$ satisfying $\int_0^1 \Phi_j(r)\,dr = 0$. In the econometric literature, the series LRV estimator has been recently used, for example, in Phillips (2005), Müller (2007), and Sun (2011, 2013, 2014a&b).

Define the projection coefficient

$$\Lambda_j(\theta_0) = \frac{1}{\sqrt{T}} \sum_{t=1}^{T} \Phi_j(\frac{t}{T}) \left[f(v_t, \theta_0) - \frac{1}{T} \sum_{\tau=1}^{T} f(v_\tau, \theta_0) \right] \text{ for } j = 1, 2, \ldots, K.$$

Then

$$W_T(\theta_0) = \frac{1}{K} \sum_{j=1}^{K} \Lambda_j(\theta_0) \Lambda_j'(\theta_0). \tag{3.2}$$

In essence, each outer product $\Lambda_j(\theta_0)\Lambda_j'(\theta_0)$ is an approximately unbiased estimator of the LRV, and the series LRV estimator is a simple average of these estimators. Here K is the smoothing parameter underlying the series LRV estimator W_T. If $\Phi_j(r) = \sqrt{2}\sin(2\pi jr)$ or $\sqrt{2}\cos(2\pi jr)$, then the series LRV estimator is proportional to the spectral density estimator at the origin that takes a simple average of the first K periodograms. The averaged periodogram estimator is a common spectral density estimator. In the traditional asymptotic framework, it can be shown that the averaged periodogram estimator is asymptotically equivalent to the kernel LRV estimator based on the Daniell kernel; See for example Phillips (2005). Sun (2013) provides more discussion on the relationship between the kernel LRV and series LRV estimators. To ensure that W_T is positive semidefinite, we assume that $K \geq m$ throughout the rest of the paper.

With the optimal weighting matrix estimator $W_T(\tilde{\theta}_T)$, the two-step GMM estimator is:

$$\hat{\theta}_T = \arg\min_{\theta \in \Theta} g_T(\theta)' W_T^{-1}(\tilde{\theta}_T) g_T(\theta).$$

Suppose that we want to perform hypothesis testing based on $\hat{\theta}_T$. Without loss of generality, we consider the linear null hypothesis $H_0 : R\theta_0 = r$ against the alternative $H_1 : R\theta_0 \neq r$ where R is a $p \times d$ matrix with full row rank. As in Sun (2014b), we consider the "trinity" of test statistics in the GMM setting. The first test statistic is the (normalized) Wald statistic given by

$$\mathbb{W}_T := \mathbb{W}_T(\hat{\theta}_T) = T(R\hat{\theta}_T - r)' \left\{ R \left[G_T(\hat{\theta}_T)' W_T^{-1}(\hat{\theta}_T) G_T(\hat{\theta}_T) \right]^{-1} R' \right\}^{-1} (R\hat{\theta}_T - r)/p, \tag{3.3}$$

where $G_T(\theta) = \frac{\partial g_T(\theta)}{\partial \theta'}$. When $p = 1$ and for one-sided alternative hypotheses, we can construct the t statistic:

$$t_T(\hat{\theta}_T) = \frac{\sqrt{T}(R\hat{\theta}_T - r)}{\left\{ R \left[G_T(\hat{\theta}_T)' W_T^{-1}(\hat{\theta}_T) G_T(\hat{\theta}_T) \right]^{-1} R' \right\}^{1/2}}.$$

The second test statistic is the GMM criterion function statistic, which can be regarded as the LR analogue in the GMM setting. Let $\hat{\theta}_{T,R}$ be the restricted second-step GMM estimator:

$$\hat{\theta}_{T,R} = \arg\min_{\theta \in \Theta} g_T(\theta)' W_T^{-1}(\tilde{\theta}_T) g_T(\theta) \;\; s.t. \; R\theta = r.$$

The GMM criterion function statistic is given by

$$\mathbb{D}_T := \left[T g_T(\hat{\theta}_{T,R})' W_T^{-1}(\tilde{\theta}_T) g_T(\hat{\theta}_{T,R}) - T g_T(\hat{\theta}_T)' W_T^{-1}(\tilde{\theta}_T) g_T(\hat{\theta}_T) \right] /p,$$

which is often referred to as the quasi LR statistic.

The third test statistic is the GMM counterpart of the score or LM statistic. Let $\Delta_T(\theta) = G_T'(\theta) W_T^{-1}(\tilde{\theta}_T) g_T(\theta)$ be the gradient of the GMM criterion function. The score type test statistic is given by

$$\mathbb{S}_T = T \left[\Delta_T(\hat{\theta}_{T,R}) \right]' \left[G_T'(\hat{\theta}_{T,R}) W_T^{-1}(\tilde{\theta}_T) G_T(\hat{\theta}_{T,R}) \right]^{-1} \Delta_T(\hat{\theta}_{T,R})/p.$$

In the definitions of $\mathbb{D}_T$ and $\mathbb{S}_T$, $\tilde{\theta}_T$ can be replaced by $\hat{\theta}_T$ or any other $\sqrt{T}$ consistent estimator without affecting our asymptotic results.

To introduce the modified or corrected versions of the above three test statistics, we construct the standard J statistic for testing the over-identifying restrictions:

$$J_T := J_T(\hat{\theta}_T) = T g_T(\hat{\theta}_T)' W_T^{-1}(\hat{\theta}_T) g_T(\hat{\theta}_T).$$

The modified or corrected versions of $\mathbb{W}_T, \mathbb{D}_T$ and $\mathbb{S}_T$ are

$$\mathbb{W}_T^c := \mathbb{W}_T^c(\hat{\theta}_T) = \frac{K-p-q+1}{K} \frac{\mathbb{W}_T(\hat{\theta}_T)}{1+\frac{1}{K} J_T(\hat{\theta}_T)},$$
$$\mathbb{D}_T^c := \mathbb{D}_T^c(\hat{\theta}_T) = \frac{K-p-q+1}{K} \frac{\mathbb{D}_T(\hat{\theta}_T)}{1+\frac{1}{K} J_T(\hat{\theta}_T)},$$
$$\mathbb{S}_T^c := \mathbb{S}_T^c(\hat{\theta}_T) = \frac{K-p-q+1}{K} \frac{\mathbb{S}_T(\hat{\theta}_T)}{1+\frac{1}{K} J_T(\hat{\theta}_T)}.$$

The multiplicative corrections are the same for all three statistics. The corresponding version of the t statistic is

$$t_T^c(\hat{\theta}_T) = \sqrt{\frac{K-q}{K}} \frac{t_T(\hat{\theta}_T)}{\sqrt{1+\frac{1}{K} J_T(\hat{\theta}_T)}}.$$

Under the conventional asymptotic theory where K diverges to ∞ with the sample size T but $K/T \to 0$, both correction factors $K-p-q+1/K$ and $(1+J_T(\hat{\theta}_T)/K)^{-1}$ approach unity in probability. So they do not matter in large samples and can thus be regarded as finite sample corrections. Under this type of asymptotics, $\mathbb{W}_T, \mathbb{D}_T$ and $\mathbb{S}_T$ and hence $\mathbb{W}_T^c, \mathbb{D}_T^c$ and $\mathbb{S}_T^c$ are all asymptotically χ_p^2/p distributed. It is now well known that the chi-square approximation is not accurate in finite samples. This motivates the more accurate fixed-smoothing asymptotics under which K is held fixed as $T \to \infty$. We point out in passing that the fixed-K specification is an asymptotic device to help establish a more accurate approximation. We do not have to use a fixed K value in finite samples.

3.3 The Asymptotic F and t Tests

Define

$$G_t(\theta) = \frac{\partial g_t(\theta)}{\partial \theta'} = \frac{1}{T}\sum_{j=1}^{t} \frac{\partial f(v_j, \theta)}{\partial \theta'} \text{ for } t \geq 1.$$

Let $u_t = f(v_t, \theta_0)$ and $\Phi_0(t) \equiv 1$, $e_t \sim iidN(0, I_m)$. We make the following assumptions on the basis functions, the GMM estimators, and the data generating process. These assumptions are the same as those in Sun (2014b) and are commonly used in the literature on the fixed-smoothing asymptotics.

Assumption 13 *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 14 *As $T \to \infty$, $\hat{\theta}_T = \theta_0 + o_p(1)$, $\tilde{\theta}_T = \theta_0 + o_p(1)$ for an interior point $\theta_0 \in \Theta$, a compact parameter space.*

Assumption 15 $\sum_{j=-\infty}^{\infty} \|\Gamma_j\| < \infty$ *where* $\Gamma_j = Eu_t u_{t-j}'$.

Assumption 16 *(a) $f(v_t, \theta)$ is twice continuously differentiable in θ for almost all v_t. (b) For any $\theta_T = \theta_0 + o_p(1)$, $plim_{T\to\infty} G_{[rT]}(\theta_T) = rG$ uniformly in r where $G = G(\theta_0)$ has rank d and $G(\theta) = E\partial f(v_t, \theta)/\partial\theta'$.*

Assumption 17 *(a) $T^{-1/2}\sum_{t=1}^{T} \Phi_j(t/T)\, u_t$ converges weakly to a continuous distribution, jointly over $j = 0, 1, \ldots, J$ for every finite J.*

(b) The following holds:

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

for every finite J where $x \in \mathbb{R}^m$ and Λ is the matrix square root of Ω, i.e., $\Lambda\Lambda' = \Omega := \sum_{j=-\infty}^{\infty} \Gamma_j$. (c) Ω is of full rank.

Let

$$B_{p+q}(r) := \left(B_p'(r),\ B_q'(r)\right)',$$

where $B_p(r)$ and $B_q(r)$ are independent standard Brownian motion processes of dimensions p and q, respectively. Denote

$$C_{pp} = \int_0^1\int_0^1 Q_K(r,s)dB_p(r)dB_p(s)',\ C_{pq} = \int_0^1\int_0^1 Q_K(r,s)dB_p(r)dB_q(s)' \quad (3.4)$$
$$C_{qq} = \int_0^1\int_0^1 Q_K(r,s)dB_q(r)dB_q(s)',\ D_{pp} = C_{pp} - C_{pq}C_{qq}^{-1}C_{pq}'.$$

Theorem 24 *Let Assumptions 13-17 hold. Then, for a fixed* K, *the following weak convergence results hold jointly as* $T \to \infty$:

(*a*) $\mathbb{W}_T(\hat{\theta}_T) \overset{d}{\to} \left[B_p(1) - C_{pq}C_{qq}^{-1}B_q(1)\right]' D_{pp}^{-1}\left[B_p(1) - C_{pq}C_{qq}^{-1}B_q(1)\right]/p :\overset{d}{=} F_\infty$,

(*b*) $t_T(\hat{\theta}_T) \overset{d}{\to} \left[B_p(1) - C_{pq}C_{qq}^{-1}B_q(1)\right]/\sqrt{D_{pp}} :\overset{d}{=} t_\infty$,

(*c*) $J_T(\hat{\theta}_T) \overset{d}{\to} B_q'(1)C_{qq}^{-1}\left[B_q(1)\right] :\overset{d}{=} J_\infty$,

where $\left(B_p'(1), B_q'(1)\right)'$ *is independent of* (C_{pq}, C_{qq}, D_{pp}) *and* D_{pp} *is independent of* (C_{pq}, C_{qq}).

The weak convergence of the marginal distributions in Theorem 24(a,b) and 24(c) has been established in Sun (2014b) and Sun and Kim (2012), respectively. It suffices to show that the weak convergence holds jointly. A proof is given in the appendix.

Remark 25 *If* $Q_K(\cdot,\cdot)$ *is replaced by a kernel function, then under some condition on the kernel function, Theorem 24 also holds. A key advantage of using the series LRV estimator is that*

$$C_K := K\begin{bmatrix} C_{pp} & C_{pq} \\ C_{pq}' & C_{qq}\end{bmatrix} = \sum_{j=1}^K\left[\int_0^1 \Phi_j(r)\,dB_{p+q}(r)\right]\left[\int_0^1 \Phi_j(r)\,dB_{p+q}(r)\right]'$$

follows a standard Wishart distribution $\mathcal{W}_{p+q}(K, I_{p+q})$. *A well-known property of a Wishart random matrix is that* $D_{pp} = C_{pp} - C_{pq}C_{qq}^{-1}C_{pq}' \sim \mathcal{W}_p(K-q, I_p)/K$.

The fact that D_{pp} *follows a Wishart distribution and its independence of* (C_{pq}, C_{qq}) *are the two key properties of* D_{pp} *that drive our F limit theory. For kernel LRV estimation,* D_{pp} *will not be Wishart and will not be independent of* (C_{pq}, C_{qq}). *So an exact F limit theory is not possible.*

Remark 26 *Note that* $\Delta = C_{pq}C_{qq}^{-1}B_q(1)$ *is independent of* $B_p(1)$ *and* D_{pp}, *the limiting distribution* F_∞ *in Theorem 24(a) conditional on* Δ *satisfies*

$$\frac{K-p-q+1}{K}F_\infty \overset{d}{=} \frac{K-p-q+1}{K}\frac{[B_p(1)-\Delta]' D_{pp}^{-1}[B_p(1)-\Delta]}{p}$$
$$\overset{d}{=} F_{p,K-p-q+1}\left(\|\Delta\|^2\right),$$

which is a noncentral F distribution with noncentrality parameter $\|\Delta\|^2$. *Unconditionally,* $\frac{K-p-q+1}{K}F_\infty$ *follows a mixed noncentral F distribution, i.e., a noncentral F distribution with a random noncentrality parameter. The noncentral F test proposed in Sun (2014b) is based on the noncentral F approximation to the mixed F distribution.*

Remark 27 *It follows from Theorem 24(c) that*

$$\frac{K-q+1}{Kq}J_T(\hat{\theta}_T) \overset{d}{\to} F_{q,K-q+1}, \tag{3.5}$$

where $F_{q,K-q+1}$ *is the standard F distribution with degrees of freedom* q *and* $K-q+1$. *This is a result first established in Sun and Kim (2012).*

Using Theorem 24, we have

$$\mathbb{W}_T^c(\hat{\theta}_T) = \frac{K-p-q+1}{K}\frac{\mathbb{W}_T(\hat{\theta}_T)}{1+\frac{1}{K}J_T(\hat{\theta}_T)}$$
$$\overset{d}{\to} \frac{K-p-q+1}{K}\frac{F_\infty}{1+\frac{1}{K}J_\infty} = \frac{K-p-q+1}{K}\xi_p' D_{pp}^{-1}\xi_p$$

where

$$\xi_p := \frac{B_p(1) - C_{pq}C_{qq}^{-1}B_q(1)}{\sqrt{1+\frac{1}{K}J_\infty}} = \frac{B_p(1) - C_{pq}C_{qq}^{-1}B_q(1)}{\sqrt{1+\frac{1}{K}B_q'(1)C_{qq}^{-1}B_q(1)}}.$$

Another key result that drives the F limit theory is that $\xi_p \sim N(0, I_p)$. This holds for the case of series LRV estimation but not for the kernel LRV estimation. The result is proved in the proof of Theorem 28 using the conditioning argument with J_∞ as the conditioning variable. This is in contrast with Sun (2014b) which uses Δ or $\|\Delta\|^2$ as the conditioning variable. Given that $\xi_p \sim N(0, I_p)$ and that ξ_p is independent of D_{pp}, $F_\infty \left(1 + K^{-1}J_\infty\right)^{-1} = \xi_p' D_{pp}^{-1} \xi_p$ follows Hotelling's T^2 distribution. Using the relationship between the T^2 distribution and the standard F distribution, we obtain Part (a) of Theorem 28. Other parts can be similarly obtained. In particular, Parts (b) and (c) follow because, as shown by Sun (2014b), the asymptotic equivalence of $\mathbb{W}_T$, $\mathbb{D}_T$, and $\mathbb{S}_T$ continues to hold under the fixed-smoothing asymptotics.

Theorem 28 *Let Assumptions 13-17 hold. Then, for a fixed K as $T \to \infty$, we have:*

(a) $\mathbb{W}_T^c(\hat{\theta}_T) \overset{d}{\to} F_{p,K-p-q+1}$;

(b) $\mathbb{D}_T^c(\hat{\theta}_T) \overset{d}{\to} F_{p,K-p-q+1}$;

(c) $\mathbb{S}_T^c(\hat{\theta}_T) \overset{d}{\to} F_{p,K-p-q+1}$;

(d) $t_T^c(\hat{\theta}_T) \overset{d}{\to} t_{K-q}$.

Remark 29 *When $q = 0$, we have $J_T(\hat{\theta}_T) = 0$ and the multiplicative correction degenerates. In this case, we have*

$$\frac{K-p+1}{K}\mathbb{W}_T(\hat{\theta}_T) \overset{d}{\to} F_{p,K-p+1}.$$

This is identical to a result obtained in Sun (2013) for the Wald test based on the first-step estimator. This is expected, as when $q = 0$, the optimal weighting matrix becomes irrelevant and the first-step estimator and two-step estimator become numerically identical.

Remark 30 *It follows from (3.5) that*

$$\frac{1}{K}J_T(\hat{\theta}_T) \overset{d}{\to} \frac{q}{K-q+1}F_{q,K-q+1} \overset{d}{=} \frac{\chi_q^2}{\chi_{K-q+1}^2}$$

for two independent chi-square random variables χ_q^2 *and* χ_{K-q+1}^2*. So, as* K *increases for a fixed* q*,* $J_T(\hat{\theta}_T)/K$ *approaches zero and the modified Wald statistic becomes close to the original Wald statistic. The multiplicative correction* $1 + J_T(\hat{\theta}_T)/K$ *can be regarded as a finite sample correction under the conventional increasing-smoothing asymptotics. For the same reason, the other multiplicative correction* $(K-p-q+1)/K$ *can be regarded as a finite sample correction under the conventional increasing-smoothing asymptotics, as* $(K-p-q+1)/K \to 1$ *as* $K \to \infty$*. This correction factor can be motivated from the Bartlett correction. See Sun (2013) for more discussion.*

Remark 31 *Let* $F_{p,K-p-q+1}^{\alpha}$ *be the* $(1-\alpha)$ *quantile of the F distribution* $F_{p,K-p-q+1}$*. According to Theorem 28, the critical value for the original test statistic* $\mathbb{W}_T(\hat{\theta}_T)$ *can be taken to be*

$$\left[1 + \frac{1}{K} J_T(\hat{\theta}_T)\right] \left[\frac{K}{K-p-q+1}\right] F_{p,K-p-q+1}^{\alpha}. \tag{3.6}$$

Compare with the chi-square critical value χ_p^{α}/p *where* χ_p^{α} *is the* $(1-\alpha)$ *quantile of the chi-squared distribution* χ_p^2*, the above critical value is larger for three reasons. First,* $F_{p,K-p-q+1}^{\alpha} > \chi_p^{\alpha}/p$ *due to the random denominator in the F distribution. Second,* $K/(K-p-q+1) > 1$ *for* $q > 1$ *or* $p > 1$*. Third,* $1 + J_T(\hat{\theta}_T)/K > 1$ *almost surely. A direct implication is that the chi-square critical values are too small, especially when* q *is large and* K *is relatively small. The small value of* K *can be empirically very relevant, as the moment process in economic applications often has high autocorrelation (e.g., Müller, 2014), which calls for a small value of* K*. Using the chi-square critical value can therefore lead to the finding of statistical significance that does not actually exist.*

Remark 32 *If we use the kernel LRV estimator, then we can choose an equivalent* K *value and use the critical value in (3.6). According to Sun and Kim (2012), the* *equivalent K value is given by the integer that is closest to*

$$\frac{\left[\int_0^1 k_b\left(r,r\right)dr\right]^2}{\int_0^1\int_0^1\left[k_b(r,s)\right]^2 drds}, \tag{3.7}$$

where

$$k_b(t,\tau)=k(\frac{t-\tau}{b})-\int_0^1 k(\frac{s-\tau}{b})ds-\int_0^1 k(\frac{t-s}{b})ds+\int_0^1\int_0^1 k(\frac{r-s}{b})drds,$$

$b = M/T$ for the truncation lag parameter M, and $k(\cdot)$ is the kernel function used in the LRV estimation. This procedure can be justified under the conventional asymptotics under which $b \to 0$, $bT \to \infty$ as $T \to \infty$, as in this case, the equivalent K value approaches ∞ and the critical value in (3.6) approaches the chi-squared critical value χ_p^α/p. In fact, as $b \to 0$, we can take

$$K=\frac{1}{b\left[\int_{-\infty}^{\infty}k^2(x)dx\right]},$$

which provides a good approximation to (3.7). Here $\int_{-\infty}^{\infty}k^2(x)dx = 2/3, 0.54,$ and 1 for the Bartlett, Parzen, and the quadratic spectral kernels, respectively. However, under the fixed-b asymptotics, the standard F distribution is not the exact limiting distribution. So, strictly speaking, we cannot justify this procedure under the fixed-b asymptotics. For this reason, one may argue that we should just simulate the nonstandard distribution and use the exact nonstandard critical value. However, the approximate critical value in (3.6) with an equivalent K is convenient to use and may be more appealing in applied research.

Remark 33 *In the proof of the theorem, we show that conditional on $B_q(\cdot)$, $\xi_p \sim N(0, I_p)$. Since the conditional distribution does not depend on $B_q(\cdot)$, we can conclude that ξ_p is independent of $B_q(\cdot)$. As a result, ξ_p is independent of $B_q(1)$ and C_{qq}. Note that D_{pp} is also independent of $B_q(1)$ and C_{qq}. So $\xi_p' D_{pp}^{-1}\xi_p$ is* *independent of* $B_q(1)' C_{qq}^{-1} B_q(1)$. *Now*

$$\begin{aligned}
&\frac{K-p-q+1}{K} F_\infty \\
&\overset{d}{=} \frac{K-p-q+1}{Kp}\left(\xi_p' D_{pp}^{-1}\xi_p\right)\left[1+\frac{1}{K}B_q'(1)C_{qq}^{-1}B_q(1)\right] \\
&\overset{d}{=} \mathcal{F}_{p,K-p-q+1}\cdot\left(1+\frac{1}{K}\mathcal{J}_\infty\right) \overset{d}{=} \mathcal{F}_{p,K-p-q+1}\cdot\left(1+\frac{q}{K-q+1}\mathcal{F}_{p,K-q+1}\right)
\end{aligned}$$

where $\mathcal{F}_{p,K-p-q+1} \sim F_{p,K-p-q+1}$, $\mathcal{J}_\infty \sim J_\infty$, $\mathcal{F}_{p,K-q+1} \sim F_{p,K-q+1}$ *and* $\mathcal{F}_{p,K-p-q+1}$ *is independent of* $\mathcal{J}_\infty$ *and* $\mathcal{F}_{p,K-q+1}$. *This gives another characterization of the nonstandard limiting distribution developed by Sun (2014b). It can be used to simplify the simulation of the nonstandard distribution* F_∞.

Remark 34 *Let* cv^α *be the nonstandard critical value for* $[(K-p-q+1)/K] \cdot \mathbb{W}_T(\hat{\theta}_T)$ *as proposed in Sun (2014b). Using the characterization in the previous remark, we have*

$$\begin{aligned}
&\lim_{T\to\infty} P\left(\frac{K-p-q+1}{K}\mathbb{W}_T(\hat{\theta}_T) > cv^\alpha\right) \\
&= P\left[\mathcal{F}_{p,K-p-q+1}\cdot\left(1+\frac{1}{K}\mathcal{J}_\infty\right) > cv^\alpha\right] \\
&= 1 - EG\left(\left.\frac{cv^\alpha}{1+\mathcal{J}_\infty/K}\right| p, K-p-q+1\right) = \alpha.
\end{aligned}$$

where $G(x|\, d_1, d_2)$ *denotes a CDF of F-distribution with parameters* d_1 *and* d_2. *That is, the asymptotic level of the nonstandard test is* α *when averaging over all realizations of* $\mathcal{J}_\infty$. *Conditional on* $\mathcal{J}_\infty$, *the asymptotic level is*

$$1 - G\left(\left.\frac{cv^\alpha}{1+\mathcal{J}_\infty/K}\right| p, K-p-q+1\right)$$

which is strictly increasing in $\mathcal{J}_\infty$. *So when the* J *statistic is large, which implies a large* $\mathcal{J}_\infty$ *in large samples, the nonstandard Wald test is expected to reject the null more often. In contrast, the critical value in (3.6) is based on the conditional distribution of* $[(K-p-q+1)/K]\,\mathbb{W}_T(\hat{\theta}_T)$ *conditional on* $J_T(\hat{\theta}_T)$. *With the conditional* *critical value, the asymptotic conditional level of the test is fixed at α regardless of the value of $J_T(\hat{\theta}_T)$.*

3.4 Understanding the Asymptotic F and t Tests

The asymptotic F and t tests may appear mysterious at first sight. To shed some light on the two tests, we consider the location model, which is perhaps the simplest model in an overidentified GMM setting:

$$\begin{aligned} y_{1t} &= \theta_0 + u_{1t},\ y_{1t} \in \mathbb{R}^p, \\ y_{2t} &= u_{2t},\ y_{2t} \in \mathbb{R}^q, \end{aligned} \tag{3.8}$$

where θ_0 is the parameter of interest, and $u_t = (u_{1t}', u_{2t}')' \in \mathbb{R}^{p+q}$ is a mean zero stationary process that can exhibit autocorrelation of unknown forms. The long run variance of u_t is

$$\Omega = \begin{pmatrix} \Omega_{11} & \Omega_{12} \\ \Omega_{21} & \Omega_{22} \end{pmatrix},$$

which has been partitioned conformably with the two blocks of equations. As simple as it is, the location model captures all the essentials in a GMM setting. In fact, a general GMM model can be reduced to the above location model in an asymptotic sense. The location model is an ideal framework to present the basic ideas and intuition, as it abstracts away the unnecessary details and complications. For more discussions, see Hwang and Sun (2015).

At the mechanical level, the parameter θ_0 can be estimated using the GMM. The moment conditions are

$$E \begin{pmatrix} y_{1t} - \theta_0 \\ y_{2t} \end{pmatrix} = 0,$$

and the GMM estimator of θ is $\hat{\theta}_{GMM} = \arg\min_{\theta\in\Theta} g_T'(\theta) W_T^{-1} g_T(\theta)$ with

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

If we take $W_{o,T} = I_{p+q}$, we obtain the initial GMM estimator $\tilde{\theta}_T = \bar{y}_1 := \frac{1}{T}\sum_{t=1}^T y_{1t}$, which is the OLS estimator based on the first block of equations. If we take W_T to be the long run variance estimator:

$$\hat{\Omega} = \begin{pmatrix} \hat{\Omega}_{11} & \hat{\Omega}_{12} \\ \hat{\Omega}_{21} & \hat{\Omega}_{22} \end{pmatrix} = \frac{1}{T}\sum_{t=1}^T\sum_{s=1}^T Q_K\left(\frac{t}{T}, \frac{s}{T}\right)(y_t - \bar{y})(y_s - \bar{y}), \tag{3.9}$$

where $y_t = (y_{1t}', y_{2t}')'$, we obtain the efficient two-step GMM estimator: $\hat{\theta}_T = \bar{y}_1 - \hat{\beta}\bar{y}_2$ with

$$\hat{\beta} = \hat{\Omega}_{12}\hat{\Omega}_{22}^{-1},$$

which is an estimator of the long run regression coefficient $\beta_0 = \Omega_{12}\Omega_{22}^{-1}$. Compared with the initial estimator $\tilde{\theta}_T$, which ignores the second block of equations, the two-step estimator $\hat{\theta}_T$ aims to explore the additional information embodied in the second block. As a special case of the GMM setting, the location model permits the asymptotic F tests and t test as described in the previous section.

To demystify the asymptotic F and t tests, we cast the GMM estimator as an OLS estimator in a linear regression model. Let

$$\omega_i(y_1) = \frac{1}{\sqrt{T}}\sum_{t=1}^T \Phi_i\left(\frac{t}{T}\right) y_{1t}, \ \omega_i(y_2) = \frac{1}{\sqrt{T}}\sum_{t=1}^T \Phi_i\left(\frac{t}{T}\right) y_{2t}$$
$$\omega_i(u_1) = \frac{1}{\sqrt{T}}\sum_{t=1}^T \Phi_i\left(\frac{t}{T}\right) u_{1t}, \ \omega_i(u_2) = \omega_i(y_2),$$
$$x_i = \frac{1}{\sqrt{T}}\sum_{t=1}^T \Phi_i\left(\frac{t}{T}\right) \text{ for } i = 0, 1, \ldots, K.$$

These transforms are analogous to the Fourier transforms and are designed to capture the long run behavior of the underlying processes. Then

$$\omega_i(y_1) = \theta_0 x_i + \omega_i(u_1)$$
$$\omega_i(y_2) = \omega_i(u_2)$$

for $i = 0, 1, \ldots, K$. This can be regarded as a system of cross-sectional regressions with dependent variables $\omega_i(y_1)$ and $\omega_i(y_2)$ and sample size $K+1$.

To obtain an efficient estimator of θ_0, we use $\omega_i(u_2)$ to predict and hence reduce the error term in the first block of equations. This is equivalent to adding $\omega_i(y_2)$ to the first block of equations, leading to the regression model of the form:

$$\omega_i(y_1) = \theta_0 x_i + \beta_0 \omega_i(y_2) + \omega_i(\varepsilon),$$

where as before $\beta_0 = \Omega_{12}\Omega_{22}^{-1} \in \mathbb{R}^{p\times q}$, $\varepsilon = u_1 - \beta_0 u_2$, and $\omega_i(\varepsilon) = \omega_i(u_1) - \beta_0\omega_i(u_2)$ is the new error term. Under Assumptions 13–17 for the location model, of which Assumptions 14 and 16 hold trivially, we have

$$\begin{pmatrix} \omega_i(u_1) \\ \omega_i(u_2) \end{pmatrix} \xrightarrow{d} iidN(0, \Omega).$$

Hence the error term $\omega_i(\varepsilon)$ is asymptotically normal. More specifically, $\omega_i(\varepsilon)$ is asymptotically iid $N(0, \Omega_{11\cdot 2})$ where

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

In addition, $\omega_i(\varepsilon)$ is asymptotically independent of $\omega_i(y_2)$.

The above model is close to a CNLR model with fixed regressors. However, there are three differences. First, the normality of the error term and its independence from the regressors hold only asymptotically. To remove this difference and for simplicity, we assume that normality holds exactly from now on, i.e., $\omega_i(\varepsilon) \sim iid\ N(0, \Omega_{11\cdot 2})$ and that $\omega_i(\varepsilon)$ is independent of $\omega_i(y_2)$. The finite sample results obtained under these assumptions then hold asymptotically without these assumptions. Second, when $p > 1$, we have a system of regressions while there is typically only one regression in a CNLR model. Of course, we can focus on the case of $p = 1$ to gain some insights but we will consider a general p. Third, $\omega_i(y_2)$ is random rather than fixed. This is innocuous, as we can follow the standard practice and use the conditioning argument.

Let

$$\omega_1 = \begin{pmatrix} \omega_0'(y_1) \\ \omega_1'(y_1) \\ \cdots \\ \omega_K'(y_1) \end{pmatrix}_{(K+1)\times p}, \omega_2 = \begin{pmatrix} \omega_0'(y_2) \\ \omega_1'(y_2) \\ \cdots \\ \omega_K'(y_2) \end{pmatrix}_{(K+1)\times q},$$

$$\omega_\varepsilon = \begin{pmatrix} \omega_0'(\varepsilon) \\ \omega_1'(\varepsilon) \\ \cdots \\ \omega_K'(\varepsilon) \end{pmatrix}_{(K+1)\times p}, \text{ and } X = \begin{pmatrix} x_0 \\ x_1 \\ \cdots \\ x_K \end{pmatrix}_{(K+1)\times 1}.$$

Then

$$\omega_1 = X\theta_0' + \omega_2\beta_0' + \omega_\varepsilon.$$

Based on this, we obtain the OLS estimator of θ_0' below:

$$\hat{\theta}_{T,OLS}' = (X'M_2X)^{-1}(X'M_2\omega_1),$$

where $M_2 = I_{K+1} - \omega_2(\omega_2'\omega_2)^{-1}\omega_2'$. Conditional on ω_2, we have

$$(\hat{\theta}_{T,OLS}' - \theta_0') \sim N\left[0, \Omega_{11\cdot 2}(X'M_2X)^{-1}\right].$$

Hence it is mixed normal unconditionally. This result is analogous to the asymptotic mixed normality of the two-step GMM estimator. In fact we can show that $\hat{\theta}_{T,OLS}$ and the two-step GMM estimator $\hat{\theta}_{T,GMM} \equiv \hat{\theta}_T$ are numerically identical under a slightly stronger condition on the basis functions. Here we add the subscript 'GMM' to $\hat{\theta}_T$ to signify its origin.

Proposition 35 *Let Assumption 13 hold with* $\int_0^1 \Phi_k(r)\,dr = 0$ *replaced by* $T^{-1}\sum_{t=1}^{T}\Phi_k(t/T) = 0$ *for* $k = 1, 2, \ldots, K$, *then* $\hat{\theta}_{T,OLS} = \hat{\theta}_{T,GMM}$. *If* $\int_0^1 \Phi_k(r)\,dr = 0$ *but not* $T^{-1}\sum_{t=1}^{T}\Phi_k(t/T) = 0$ *for* $k = 1, 2, \ldots, K$, *then under Assumptions 13–17, we have* $\sqrt{T}(\hat{\theta}_{T,OLS} - \hat{\theta}_{T,GMM}) = o_p(1)$ *for a fixed* K *as* $T \to \infty$.

While the asymptotic equivalence between $\hat{\theta}_{T,OLS}$ and $\hat{\theta}_{T,GMM}$ is well expected, it is nontrivial to show that they are numerically identical under the assumption that $T^{-1}\sum_{t=1}^{T}\Phi_k(t/T) = 0$. This assumption holds for $\Phi_k(t/T) = \sqrt{2}\sin(2\pi kt/T)$, $\sqrt{2}\cos(2\pi kt/T)$, which are the basis functions in common use for the series LRV estimation.

The conditional distribution of $(\hat{\theta}'_{T,OLS} - \theta'_0)$ conditional on ω_2 depends on ω_2 only through $(X'M_2X)^{-1}$. It then follows that the conditional distribution of $(\hat{\theta}'_{T,OLS} - \theta'_0)$ conditional on $(X'M_2X)^{-1}$ is also $N[0, \Omega_{11\cdot 2}(X'M_2X)^{-1}]$. In the proof of the proposition, it is shown that $(X'M_2X)^{-1} = (1 + T\bar{y}_2'\hat{\Omega}_{22}^{-1}\bar{y}_2/K)/T$. Therefore, we can take $T\bar{y}_2'\hat{\Omega}_{22}^{-1}\bar{y}_2$ as the conditioning variable. But $T\bar{y}_2'\hat{\Omega}_{22}^{-1}\bar{y}_2$ is exactly the J statistic in the overidentified location model. So the minimal conditioning variable in the CNLR coincides with the conditioning variable we use in the GMM framework.

Now suppose that we follow the mechanics in the CNLR framework to conduct inference. Conditional on $(X'M_2X)^{-1}$, the variance of $\hat{\theta}_{T,OLS}$ is $\Omega_{11\cdot 2}(X'M_2X)^{-1}$. Following a routine in the CNLR framework, we can estimate the conditional variance by $\tilde{\Omega}_{11\cdot 2}(X'M_2X)^{-1}$ where

$$\tilde{\Omega}_{11\cdot 2} = \frac{1}{K-q}\left(\omega_1 - X\hat{\theta}'_{T,OLS} - \omega_2\hat{\beta}'_{T,OLS}\right)'\left(\omega_1 - X\hat{\theta}'_{T,OLS} - \omega_2\hat{\beta}'_{T,OLS}\right)$$

and $\hat{\beta}'_{T,OLS}$ is the OLS estimator of β_0. Here we have used $1/(K-q) = 1/(K+1-q-1)$ instead of $1/(K+1)$ as the scaling function. This is the usual degree-of-freedom correction in a standard linear regression model. Constructing the Wald statistic for testing $H_0 : \theta_0 = r$ in the same way as what we would do in a CNLR framework, we obtain the (normalized) Wald statistic

$$\mathbb{W}_{CNLR} = \sqrt{T}\left(\hat{\theta}_{T,OLS} - r\right)' \left[\tilde{\Omega}_{11\cdot 2}\left(\frac{X'M_2X}{T}\right)^{-1}\right]^{-1} \sqrt{T}\left(\hat{\theta}_{T,OLS} - r\right)/p.$$

We can also construct other type statistics such as the LR, LM and t statistics but we focus on the Wald statistic here.

To formally compare $\mathbb{W}_{CNLR}$ with the unmodified GMM Wald statistic as given in (3.3), we note that for the location model $G_T(\hat{\theta}_T) = (I_p, O_{p\times q})'$. Using this and plugging $W_T(\hat{\theta}_{T,GMM}) = \hat{\Omega}$ and $R = I_p$ into (3.3), we obtain

$$\mathbb{W}_T = \sqrt{T}(\hat{\theta}_{T,GMM} - r)' \left[\hat{\Omega}_{11\cdot 2}\right]^{-1} \sqrt{T}(\hat{\theta}_{T,GMM} - r)/p, \tag{3.10}$$

where $\hat{\Omega}_{11\cdot 2} = \hat{\Omega}_{11} - \hat{\Omega}_{12}\hat{\Omega}_{22}^{-1}\hat{\Omega}_{21}$ and $\hat{\Omega}_{ij}$ are given in (3.9). A formal comparison of $\mathbb{W}_{CNLR}$ with $\mathbb{W}_T$ reveals that $\mathbb{W}_{CNLR}$ has the additional factor $(X'M_2X/T)^{-1}$ in the variance estimator that the GMM Wald statistic $\mathbb{W}_T$ ignores. The reason that $\mathbb{W}_T$ ignores this factor is that the underlying variance estimator is based on the conventional "sandwich" formula, which is derived under the conventional increasing-smoothing asymptotics where $K \to \infty$ as $T \to \infty$. Under this type of asymptotics, $(X'M_2X/T)^{-1} \to^p 1$ and so the factor is negligible in large samples. Under the fixed-smoothing asymptotics, it follows from Hwang and Sun (2015b) that

$$\begin{aligned}\sqrt{T}\left(\hat{\theta}_{T,GMM} - \theta_0\right) &= \begin{pmatrix} I_p, & -\hat{\beta} \end{pmatrix} \begin{pmatrix} \frac{1}{\sqrt{T}}\sum_{t=1}^{T}(y_{1t} - Ey_{1t}) \\ \frac{1}{\sqrt{T}}\sum_{t=1}^{T} y_{2t} \end{pmatrix} \\ &\to^d \begin{pmatrix} I_p, & -\beta_\infty \end{pmatrix} \Lambda B_{p+q}(1),\end{aligned}$$

where

$$\beta_\infty = \Omega_{11\cdot 2}^{1/2}\tilde{\beta}_\infty \Omega_{22}^{-1/2} + \Omega_{12}\Omega_{22}^{-1} \text{ and } \tilde{\beta}_\infty = C_{pq}C_{qq}^{-1}. \tag{3.11}$$

Some simple calculations show that the asymptotic variance of $\hat{\theta}_{T,GMM}$ conditional on $\tilde{\beta}_\infty$ satisfies:

$$\text{avar}(\hat{\theta}_{T,GMM}) = \Omega_{11\cdot 2}^{1/2}\left(I_p + \tilde{\beta}_\infty \tilde{\beta}_\infty'\right)(\Omega_{11\cdot 2}^{1/2})' = \Omega_{11\cdot 2} + \Omega_{11\cdot 2}^{1/2}\tilde{\beta}_\infty \tilde{\beta}_\infty'(\Omega_{11\cdot 2}^{1/2})'.$$

When we use the conventional "sandwich" formula for variance estimation, which attempts to estimate $\Omega_{11\cdot 2}$ only, we effectively ignore the term that involves $\tilde{\beta}_\infty \tilde{\beta}_\infty'$. This will not cause any problem for asymptotic pivotal inference but will prevent us from developing an F limit theory. The modification we propose can be regarded as the multiplicative variance correction that takes into account the extra asymptotic variance term under the fixed-smoothing asymptotics. More specifically, instead of using $\hat{\Omega}_{11\cdot 2}$, we use $\hat{\Omega}_{11\cdot 2}(1 + \hat{J}_T/K)$ as the asymptotic variance estimator.

The following proposition establishes the connection between $\mathbb{W}_{CNLR}$ and $\mathbb{W}_T^c$ rigorously.

Proposition 36 *Let Assumption 13 hold with* $\int_0^1 \Phi_k(r)\,dr = 0$ *replaced by* $T^{-1}\sum_{t=1}^T \Phi_k(t/T) = 0$ *for* $i = 1, 2, \ldots, K$. *Then*

$$\mathbb{W}_{CNLR} = \frac{K-q}{K-p-q+1}\mathbb{W}_T^c.$$

In particular, $\mathbb{W}_{CNLR} = \mathbb{W}_T^c$ *when* $p = 1$. *If* $\int_0^1 \Phi_k(r)\,dr = 0$ *but not* $T^{-1}\sum_{t=1}^T \Phi_k(t/T) = 0$, *then under Assumptions 13–17, we have* $\mathbb{W}_{CNLR} = \frac{K-q}{K-p-q+1}\mathbb{W}_T^c + o_p(1)$ *for a fixed* K *as* $T \to \infty$.

Remark 37 *When* $p = 1$, *the proposition shows that the Wald statistic constructed in the standard way is numerically identical to the modified Wald statistic we propose in the GMM setting. While the modification can be motivated on the ground of obtaining a convenient standard F limiting distribution, it is a built-in feature of the standard Wald statistic in a linear regression. The modification may appear to be mysterious at first sight but it becomes natural from the regression perspective.*

Remark 38 *When* $p > 1$, $\mathbb{W}_{CNLR}$ *does not follow an* F *distribution but a rescaled version does:*

$$\frac{K-p-q+1}{K-q}\mathbb{W}_{CNLR} \sim F_{p,K-p-q+1}.$$

This follows from Theorem 28 and Proposition 36. Of course this can be proved directly in the CNLR setting but there is no need to do so, as the limit theory established in the GMM setting is directly applicable to the CNLR model.

Remark 39 *Looking at the GMM problem from the regression perspective motivates us to use the modified Wald statistic even if there is no serial dependence. In this case, we can take* $K = T$ *and the modified Wald statistic becomes*

$$\mathbb{W}_T^c := \frac{T-p-q+1}{T}\frac{\mathbb{W}_T}{1+\frac{1}{T}J_T},$$

where $\mathbb{W}_T$ *and* J_T *are the standard Wald and J statistics in the GMM framework with iid data. In addition, we use* $F_{p,T-p-q+1}$ *instead of* χ_p^2/p *as the reference distribution. From an asymptotic point of view,* $\mathbb{W}_T^c = \mathbb{W}_T + o_p(1)$ *and* $F_{p,T-p-q+1}^{\alpha} = \chi_p^{\alpha}/p + o(1)$ *as* $T \to \infty$*. So the modified Wald test based on the* F *approximation can be justified in the same manner as the conventional chi-square test. However, in finite samples, the new test can be more accurate in size.*

3.5 Simulation Evidence

3.5.1 Asymptotic Size and Power

We follow Sun (2014b) and consider a linear model of the form:

$$y_t = x_t'\theta + \varepsilon_{y,t}$$

where $x_t := (1, x_{1,t}, x_{2,t}, x_{3,t})'$ is a vector of endogeneous regressors. The unknown parameter vector is $\theta = (\gamma_0, \gamma_1, ..., \gamma_d)' \in \mathbb{R}^d$. We have m instruments

$z_{0,t}, z_{1,t}, \ldots, z_{m-1,t}$ with $z_{0,t} \equiv 1$. The reduced-form equations for $x_{1,t}$, $x_{2,t}$ and $x_{3,t}$ are given by

$$x_{j,t} = z_{j,t} + \sum_{i=d}^{m-1} z_{i,t} + \varepsilon_{x_j,t} \text{ for } j = 1, ..., d-1.$$

We consider two different experiment designs: the autoregressive (AR) design and the centered moving average (CMA) design. In the AR design, each $z_{i,t}$ follows an AR(1) process of the form $z_{i,t} = \rho z_{i,t-1} + \sqrt{1-\rho^2} e_{z_i,t}$ where $e_{z_i,t} = \left(e_{zt}^i + e_{zt}^0\right)/\sqrt{2}$ and $e_t = [e_{zt}^0, e_{zt}^1, \ldots, e_{zt}^{m-1}]' \sim iidN(0, I_m)$. By construction, z_{it} has unit variance for all for $i \geq 1$, and the correlation coefficient between the non-constant $z_{i,t}$ and $z_{j,t}$ for $i \neq j$ is 0.5. The DGP for $\varepsilon_t = (\varepsilon_{yt}, \varepsilon_{x_1t}, \varepsilon_{x_2t}, \varepsilon_{x_3t})'$ is the same as that for $(z_{1,t}, \ldots, z_{m-1,t})$ except that there is a difference in the dimension. The two vector processes ε_t and $(z_{1,t}, \ldots, z_{m-1,t})$ are independent from each other. We take $\rho = -0.5, 0.0, 0.5$, 0.8 and 0.9.

In the CMA design, $\varepsilon_{y,t}$ is a scaled and centered moving average of an iid sequence $\varepsilon_{y,t} = \sum_{j=-L}^{L} e_{t+j}/\sqrt{2L+1}$ where $e_t \sim iidN(0,1)$ and L is the number of leads and lags in the average. The instruments are generated according to $z_{it} = [e_{t-L+i-1} - (2L+1)^{-1} \sum_{j=-L}^{L} e_{t+j}] \sqrt{(2L+1)/2L}$ for $i = 1, \ldots, m-1$. The error term in the reduced-form equation is given by $\varepsilon_{x_j,t} = \left(\varepsilon_{y,t} + e_{x_j,t}\right)/\sqrt{2}$ where $e_{x_j,t} \sim iidN(0,1)$ and is independent of the sequence $\{e_t\}$. We take $L = 3$, 6, and 9.

We consider $q = 0, 1, 2$ and $d = 4$ with corresponding numbers of moment conditions $m = 4, 5, 6$. The null hypotheses of interest are

$$H_{01} : \gamma_1 = 0,$$
$$H_{02} : \gamma_1 = \gamma_2 = 0,$$
$$H_{03} : \gamma_1 = \gamma_2 = \gamma_3 = 0.$$

The numbers of joint hypotheses are $p = 1$, 2 and 3, respectively. We consider three different sample sizes $T = 100, 200, 500$ and two significance levels $\alpha = 5\%$ and $\alpha = 10\%$. We focus on the Wald type of test but the simulation results are qualitatively similar for other type of tests.

We examine the empirical size of four different two-step tests. The first three tests are based on the same unmodified Wald test statistic, so they have the same size-adjusted power. The difference lies in the critical values used. We employ the following critical values: $\chi_p^{1-\alpha}/p$, $\frac{K}{K-p-q+1}\mathcal{F}_{p,K-p-q+1}^{1-\alpha}(\delta^2)$ with $\delta^2 = pq/(K-q-1)$, and $\mathcal{F}_\infty^{1-\alpha}$, leading to the χ^2 test, the NCF (noncentral F) test and the nonstandard F_∞ test. The χ^2 test uses the conventional chi-square approximation. The NCF test uses the noncentral F approximation. The F_∞ test uses the nonstandard F_∞ approximation with simulated critical values. The NCF test and the F_∞ test are developed in Sun (2014b). The fourth test is the test proposed in this paper, which is based on the modified Wald statistic $\mathbb{W}_T^c$ and uses the standard F critical value $\mathcal{F}_{p,K-p-q+1}^{1-\alpha}$. Equivalently, our proposed test is based the unmodified Wald test statistic as the first three tests but uses the critical values given in (3.6). For easy reference, we now refer to our test as the standard F test, which should not be confused with the standard F test in a CNLR model. For each test, the initial first-step estimator is the IV estimator with weight matrix $W_o = Z'Z/T$ where Z is the matrix of instruments.

We use the following basis functions $\Phi_{2j-1}(x) = \sqrt{2}\cos 2j\pi x$, $\Phi_{2j}(x) = \sqrt{2}\sin 2j\pi x$, $j = 1, \ldots, K/2$ and assume that K is even. In this case, the series LRV estimator can be computed using discrete Fourier transforms. We select K based on the AMSE criterion implemented using the VAR(1) plug-in procedure in Phillips (2005), which is similar to the plug-in procedure of Andrews (1991). We compute the data-driven K on the basis of the initial first step estimator $\tilde{\theta}_T$ and use it in computing both $W_T(\tilde{\theta}_T)$ and $W_T(\hat{\theta}_T)$.

We also compare the size-adjusted power of the proposed standard F test with that of the nonstandard F_∞ test. The data is generated under the local alternative $H_1 : R\theta = c_0\ell_p/\sqrt{T}$ where c_0 is a scalar and ℓ_p is the p-vector of ones. The two tests use the same data driven smoothing parameter K. To make the power comparison meaningful, we compute the power using the empirical finite sample critical values obtained from the null distribution. That is, we compare the size-adjusted power. It should be pointed out that size-adjustment is not feasible in practice.

Tables 3.1 and 3.2 report the finite sample size of the four tests for $T = 100$ and $\alpha = 5\%$. The number of simulation replications is 10000. It is clear that the standard F test has as accurate size as the nonstandard F_∞ test and noncentral F test. Like the latter two tests, the standard F test is much more accurate in size than the conventional chi-square test, which can be highly size-distorted. These qualitative observations remain valid for other sample sizes and significance levels.

Figures 3.1 and **??** report the size-adjusted power of the nonstandard F_∞ test and the standard F test for $\alpha = 5\%$ and $T = 100$. There is no real difference between the two power curves. In fact, the standard F test can be slightly more powerful in some scenarios. Note that the size-adjusted power of the nonstandard F_∞ test is the same as that of the conventional chi-square test, the standard F test is therefore as powerful as the conventional chi-square test.

Our simulation evidence lends a strong support to the standard F test: it enjoys the same good size and power properties as the nonstandard F_∞ test but it is easier to use, as the critical values are readily available from statistical tables and no simulation or approximation is needed.

3.6 Conclusion

This paper has proposed a modification to the trinity of test statistics in an efficient two-step GMM framework. Each modified test statistic is a function of the original test statistic and the usual J statistic for testing overidentification. We show that the modified test statistics are all asymptotically F distributed. This leads to standard F tests that are based on the modified test statistics and use the standard F critical values. Simulation shows that the standard F tests have the same finite sample performance as the nonstandard tests recently proposed by Sun (2014b) but the standard F tests are much easier to use.

The paper complements Sun (2011a, 2013, 2014a) and Sun and Kim (2012) which establish the F limit theory for the tests based on the first-step GMM estimation and the J test. When the series LRV estimator is used, the F limit theory appears to be applicable to all common tests in the first-step and two-step GMM settings. The results of the paper can be easily extended to the continuous updating GMM (CU-GMM) framework. Recently, Zhang (2015) has shown that the Wald statistic based on the CU-GMM estimator has the same fixed-smoothing limit as what Sun (2014b) obtains in the two-step GMM framework. Given this, it is easy to see that our result holds without change if the CU-GMM estimator is used instead. Following the work of Bester et al. (2016) and Sun and Kim (2015), we also do not imagine much difficulty in extending our results to the spatial setting.

3.7 Acknowledgements

Chapter 3, in full, is co-authored with Yixiao Sun and has been submitted for publication of material.

3.8 Figures and Tables

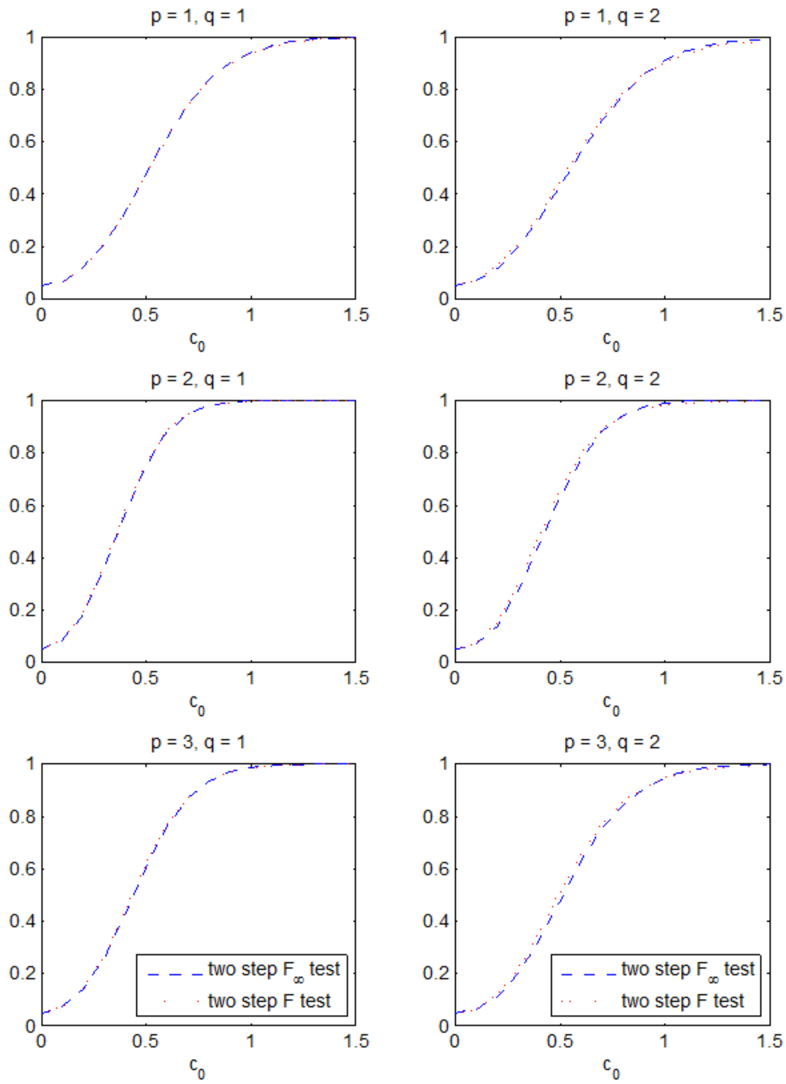

Figure 3.1: Size-adjusted power of two-step 5% F_∞ and F tests based on the series LRV estimator under the AR design with $\rho = 0.5$ and $T = 100$

Table 3.1: Empirical size of the nominal 5% χ^2 test, noncentral F test, nonstandard F_∞ test and standard F test based on the series LRV estimator under the AR design with $T = 100$, number of joint hypotheses p, and number of overidentifying restrictions q

	χ^2	NCF	F_∞	F	χ^2	NCF	F_∞	F	χ^2	NCF	F_∞	F
ρ	$p=1, q=0$				$p=2, q=0$				$p=3, q=0$			
-0.95	0.187	0.136	0.137	0.136	0.332	0.202	0.197	0.202	0.492	0.263	0.267	0.263
-0.8	0.114	0.072	0.073	0.072	0.197	0.087	0.084	0.087	0.310	0.109	0.111	0.109
-0.5	0.081	0.060	0.059	0.060	0.117	0.066	0.066	0.066	0.174	0.077	0.078	0.077
0.0	0.063	0.051	0.050	0.051	0.083	0.052	0.053	0.052	0.112	0.060	0.062	0.060
0.5	0.094	0.063	0.063	0.063	0.142	0.065	0.065	0.065	0.222	0.077	0.078	0.077
0.8	0.134	0.086	0.088	0.086	0.229	0.100	0.097	0.100	0.355	0.119	0.122	0.119
0.9	0.166	0.117	0.120	0.117	0.290	0.150	0.146	0.150	0.437	0.181	0.184	0.181
0.95	0.203	0.145	0.148	0.145	0.344	0.202	0.197	0.202	0.500	0.257	0.261	0.257
	$p=1, q=1$				$p=2, q=1$				$p=3, q=1$			
-0.95	0.316	0.181	0.176	0.168	0.521	0.249	0.246	0.234	0.705	0.305	0.315	0.281
-0.8	0.186	0.081	0.077	0.079	0.307	0.088	0.087	0.086	0.457	0.107	0.113	0.105
-0.5	0.113	0.065	0.065	0.064	0.175	0.069	0.068	0.067	0.247	0.079	0.080	0.078
0.0	0.081	0.053	0.052	0.052	0.113	0.057	0.056	0.056	0.155	0.060	0.060	0.060
0.5	0.128	0.064	0.063	0.063	0.204	0.073	0.072	0.071	0.308	0.079	0.080	0.078
0.8	0.196	0.089	0.086	0.087	0.331	0.101	0.099	0.100	0.489	0.112	0.118	0.112
0.9	0.252	0.126	0.123	0.122	0.420	0.155	0.153	0.147	0.589	0.183	0.192	0.172
0.95	0.306	0.172	0.169	0.162	0.504	0.234	0.232	0.216	0.681	0.277	0.286	0.255
	$p=1, q=2$				$p=2, q=2$				$p=3, q=2$			
-0.95	0.402	0.185	0.182	0.160	0.634	0.241	0.227	0.205	0.812	0.271	0.267	0.231
-0.8	0.260	0.080	0.079	0.077	0.425	0.090	0.083	0.084	0.602	0.100	0.097	0.090
-0.5	0.154	0.061	0.062	0.061	0.244	0.065	0.061	0.065	0.351	0.074	0.073	0.072
0.0	0.104	0.055	0.055	0.055	0.148	0.062	0.058	0.060	0.211	0.065	0.063	0.065
0.5	0.171	0.065	0.066	0.064	0.279	0.065	0.061	0.062	0.415	0.073	0.072	0.070
0.8	0.268	0.085	0.082	0.080	0.449	0.090	0.082	0.085	0.623	0.100	0.097	0.088
0.9	0.332	0.124	0.121	0.109	0.529	0.148	0.137	0.132	0.712	0.160	0.157	0.139
0.95	0.391	0.170	0.166	0.144	0.613	0.220	0.208	0.181	0.782	0.244	0.240	0.199

Notes: The first three tests χ^2, NCF and F_∞ in Table 3.1 are based on the same unmodified Wald statistic but use different critical values. The χ^2 test uses the chi-squared critical value; the NCF test uses the noncentral F critical value; and the F_∞ test uses simulated nonstandard critical value. The standard F test is based on the modified Wald statistic and uses the standard F critical value.

Table 3.2: Empirical size of the nominal 5% χ^2 test, noncentral F test, nonstandard F_∞ test and standard F test based on the series LRV estimator under the centered MA design with $T = 100$, number of joint hypotheses p, and number of overidentifying restrictions q

	χ^2	NCF	F_∞	F	χ^2	NCF	F_∞	F	χ^2	NCF	F_∞	F
L		$p=1, q=0$				$p=2, q=0$				$p=3, q=0$		
3	0.017	0.007	0.007	0.007	0.089	0.030	0.029	0.030	0.201	0.047	0.048	0.047
6	0.030	0.017	0.017	0.017	0.068	0.023	0.022	0.023	0.134	0.028	0.029	0.028
9	0.048	0.029	0.030	0.029	0.079	0.027	0.026	0.027	0.142	0.032	0.033	0.032
		$p=1, q=1$				$p=2, q=1$				$p=3, q=1$		
3	0.102	0.033	0.031	0.036	0.229	0.057	0.056	0.059	0.299	0.047	0.050	0.049
6	0.106	0.039	0.037	0.046	0.169	0.031	0.031	0.036	0.275	0.034	0.037	0.039
9	0.108	0.035	0.034	0.039	0.159	0.032	0.031	0.034	0.259	0.033	0.036	0.036
		$p=1, q=2$				$p=2, q=2$				$p=3, q=2$		
3	0.180	0.046	0.046	0.042	0.286	0.050	0.046	0.042	0.387	0.043	0.043	0.035
6	0.164	0.039	0.037	0.045	0.265	0.036	0.032	0.040	0.425	0.036	0.036	0.039
9	0.165	0.040	0.039	0.043	0.265	0.032	0.029	0.034	0.402	0.032	0.031	0.034

See footnotes to Table 3.1

3.9 Appendix of Proofs

Proof of Theorem 24. The marginal weak convergence results in (a) and (b) have been proved in Sun (2014b, Theorem 1), and the result in (c) has been proved in Sun and Kim (2012, Theorem 1 and equation (7)). It remains to show that the convergence results hold jointly. As a representative example, we prove that (a) and (c) hold jointly.

Let

$$\tilde{W}_{\infty}=\int_{0}^{1}\int_{0}^{1}Q_{K}\left(r,s\right)dB_{m}\left(r\right)dB_{m}\left(s\right)$$

and $G_{\Lambda}=\Lambda^{-1}G$, which is an $m\times d$ matrix, then it follows from Sun (2014b) and Sun and Kim (2012) that

$$\begin{aligned}\mathbb{W}_{T}(\hat{\theta}_{T}) &\overset{d}{\to}\left\{R\left[G_{\Lambda}'\tilde{W}_{\infty}^{-1}G_{\Lambda}\right]^{-1}G_{\Lambda}'\tilde{W}_{\infty}^{-1}B_{m}(1)\right\}'\left\{R\left[G_{\Lambda}'\tilde{W}_{\infty}^{-1}G_{\Lambda}\right]^{-1}R'\right\}^{-1}\\ &\times\left\{R\left[G_{\Lambda}'\tilde{W}_{\infty}^{-1}G_{\Lambda}\right]^{-1}G_{\Lambda}'\tilde{W}_{\infty}^{-1}B_{m}(1)\right\}/p:=F_{\infty},\end{aligned}$$

$$\begin{aligned}J_{T}(\hat{\theta}_{T}) &\overset{d}{\to}\left\{B_{m}(1)-G_{\Lambda}\left[G_{\Lambda}'\tilde{W}_{\infty}^{-1}G_{\Lambda}\right]^{-1}G_{\Lambda}'\tilde{W}_{\infty}^{-1}B_{m}(1)\right\}'\tilde{W}_{\infty}^{-1}\\ &\times\left\{B_{m}(1)-G_{\Lambda}\left[G_{\Lambda}'\tilde{W}_{\infty}^{-1}G_{\Lambda}\right]^{-1}G_{\Lambda}'\tilde{W}_{\infty}^{-1}B_{m}(1)\right\}:=J_{\infty}.\end{aligned} \tag{3.12}$$

In addition, a careful inspection shows that the above convergence results hold jointly. It remain to show that (F_{∞},J_{∞}) is equivalent in distribution to

$$\left(\left[B_{p}\left(1\right)-C_{pq}C_{qq}^{-1}B_{q}\left(1\right)\right]'D_{pp}^{-1}\left[B_{p}\left(1\right)-C_{pq}C_{qq}^{-1}B_{q}\left(1\right)\right]/p,\ B_{q}'(1)C_{qq}^{-1}B_{q}(1)\right).$$

Let $U_{m\times m}\Sigma_{m\times d}V_{d\times d}'$ be a singular value decomposition (SVD) of G_{Λ}. By definition, $U'U=UU'=I_{m}$, $VV'=V'V=I_{d}$ and

$$\Sigma=\begin{bmatrix}A_{d\times d}\\ O_{q\times d}\end{bmatrix},$$

where A is a diagonal matrix with singular values on the main diagonal and O is a matrix of zeros. Then we have:

$$\begin{aligned}
&\left[G_\Lambda'\tilde{W}_\infty^{-1}G_\Lambda\right]^{-1}G_\Lambda'\tilde{W}_\infty^{-1}B_m(1)=\left[V\Sigma'U'\tilde{W}_\infty^{-1}U\Sigma V'\right]^{-1}V\Sigma'U'\tilde{W}_\infty^{-1}B_m(1)\\
&=V\left[\Sigma'U'\tilde{W}_\infty^{-1}U\Sigma\right]^{-1}\Sigma'U'\tilde{W}_\infty^{-1}B_m(1)\\
&=V\left[\Sigma'U'\tilde{W}_\infty^{-1}U\Sigma\right]^{-1}\Sigma'\left[U'\tilde{W}_\infty^{-1}U\right]\left[U'B_m(1)\right]
\end{aligned}$$

and

$$\begin{aligned}
&B_m(1)-G_\Lambda\left[G_\Lambda'\tilde{W}_\infty^{-1}G_\Lambda\right]^{-1}G_\Lambda'\tilde{W}_\infty^{-1}B_m(1)\\
&=B_m(1)-U\Sigma V'\left[V\Sigma'U'\tilde{W}_\infty^{-1}U\Sigma V'\right]^{-1}V\Sigma'U'\tilde{W}_\infty^{-1}B_m(1)\\
&=U\left\{U'B_m(1)-\Sigma\left[\Sigma'U'\tilde{W}_\infty^{-1}U\Sigma\right]^{-1}\Sigma'\left(U'\tilde{W}_\infty^{-1}U\right)U'B_m(1)\right\}.
\end{aligned}$$

Since $[U'\tilde{W}_\infty^{-1}U,\, U'B_m(1)]$ has the same joint distribution as $[\tilde{W}_\infty^{-1}, B_m(1)]$, we can write

$$\begin{pmatrix} F_\infty \\ J_\infty \end{pmatrix} \overset{d}{=} \begin{pmatrix} \tilde{F}_\infty \\ \tilde{J}_\infty \end{pmatrix}$$

where

$$\begin{aligned}
\tilde{F}_\infty=\;&B_m(1)'\left[RV\left[\Sigma'\tilde{W}_\infty^{-1}\Sigma\right]^{-1}\Sigma'\tilde{W}_\infty^{-1}\right]'\left\{RV\left[\Sigma'\tilde{W}_\infty^{-1}\Sigma\right]^{-1}V'R'\right\}^{-1}\\
&\times\left[RV\left[\Sigma'\tilde{W}_\infty^{-1}\Sigma\right]^{-1}\Sigma'\tilde{W}_\infty^{-1}\right]B_m(1),
\end{aligned}$$

and

$$\begin{aligned}
\tilde{J}_\infty=\;&B_m'(1)\left\{I_m-\Sigma\left[\Sigma'\tilde{W}_\infty^{-1}\Sigma\right]^{-1}\Sigma'\tilde{W}_\infty^{-1}\right\}'\tilde{W}_\infty^{-1}\\
&\times\left\{I_m-\Sigma\left[\Sigma'\tilde{W}_\infty^{-1}\Sigma\right]^{-1}\Sigma'\tilde{W}_\infty^{-1}\right\}B_m(1).
\end{aligned}$$

We proceed to simplify $\tilde{F}_\infty$ and $\tilde{J}_\infty$ starting with $\tilde{F}_\infty$. We let

$$\tilde{W}_\infty = \begin{pmatrix} C_{11} & C_{12} \\ C_{21} & C_{22} \end{pmatrix} \text{ and } \tilde{W}_\infty^{-1} = \begin{pmatrix} C^{11} & C^{12} \\ C^{21} & C^{22} \end{pmatrix}$$

where C_{11} and C^{11} are $d \times d$ matrices, C_{22} and C^{22} are $q \times q$ matrices, and $C_{12} = C_{21}'$, $C^{12} = (C^{21})'$. By definition,

$$C_{11} = \int_0^1 \int_0^1 Q_K(r,s)dB_d(r)dB_d(s)' = \begin{pmatrix} C_{pp} & C_{p,d-p} \\ C_{p,d-p}' & C_{d-p,d-p} \end{pmatrix} \quad (3.13)$$

$$C_{12} = \int_0^1 \int_0^1 Q_K(r,s)dB_d(r)dB_q(s)' = \begin{pmatrix} C_{pq} \\ C_{d-p,q} \end{pmatrix} \quad (3.14)$$

$$C_{22} = \int_0^1 \int_0^1 Q_K(r,s)dB_q(r)dB_q(s)' = C_{qq} \quad (3.15)$$

where C_{pp}, C_{pq}, and C_{qq} are defined in (3.4), and $C_{d-p,d-p}$, $C_{p,d-p}$ and $C_{d-p,q}$ are similarly defined. It follows from the partitioned inverse formula that

$$C^{11} = \left[C_{11} - C_{12}C_{22}^{-1}C_{21}\right]^{-1}, \; C^{12} = -C^{11}C_{12}C_{22}^{-1}.$$

With the above partition of $\tilde{W}_\infty^{-1}$, we have

$$\begin{aligned}\left[\Sigma'\tilde{W}_\infty^{-1}\Sigma\right]^{-1} &= \left\{ \begin{pmatrix} A' & O' \end{pmatrix} \begin{pmatrix} C^{11} & C^{12} \\ C^{21} & C^{22} \end{pmatrix} \begin{pmatrix} A \\ O \end{pmatrix} \right\}^{-1} \\ &= \left[A'C^{11}A\right]^{-1} = A^{-1}\left(C^{11}\right)^{-1}\left(A'\right)^{-1},\end{aligned}$$

and so

$$\begin{aligned}
& RV\left[\Sigma'\tilde{W}_{\infty}^{-1}\Sigma\right]^{-1}\Sigma'\tilde{W}_{\infty}^{-1} \\
&= RVA^{-1}\left(C^{11}\right)^{-1}(A')^{-1}\left(\begin{array}{cc} A' & O' \end{array}\right)\left(\begin{array}{cc} C^{11} & C^{12} \\ C^{21} & C^{22} \end{array}\right) \\
&= RVA^{-1}\left(C^{11}\right)^{-1}(A')^{-1}A'\left(\begin{array}{cc} C^{11} & C^{12} \end{array}\right) \\
&= RVA^{-1}\left(\begin{array}{cc} I_d, & \left(C^{11}\right)^{-1}C^{12} \end{array}\right), \qquad (3.16)
\end{aligned}$$

and

$$RV\left[\Sigma'\tilde{W}_{\infty}^{-1}\Sigma\right]^{-1}V'R' = RVA^{-1}\left(C^{11}\right)^{-1}(A')^{-1}V'R'.$$

As a result,

$$\begin{aligned}
\tilde{F}_{\infty} &= B_m(1)'\left[RVA^{-1}\left(\begin{array}{cc} I_d, & \left(C^{11}\right)^{-1}C^{12} \end{array}\right)\right]'\left[RVA^{-1}\left(C^{11}\right)^{-1}(A')^{-1}V'R'\right]^{-1} \\
&\times\left[RVA^{-1}\left(\begin{array}{cc} I_d, & \left(C^{11}\right)^{-1}C^{12} \end{array}\right)\right]B_m(1)/p \\
&= B_m(1)'\left[RVA^{-1}\left(\begin{array}{cc} I_d, & -C_{12}C_{22}^{-1} \end{array}\right)\right]'\left[RVA^{-1}\left(C^{11}\right)^{-1}(A')^{-1}V'R'\right]^{-1} \\
&\times\left[RVA^{-1}\left(\begin{array}{cc} I_d, & -C_{12}C_{22}^{-1} \end{array}\right)\right]B_m(1)/p.
\end{aligned}$$

Let $B_m(1) = \left[B_d'(1), B_q'(1)\right]'$ and $\tilde{U}_{p\times p}\tilde{\Sigma}_{p\times d}\tilde{V}_{d\times d}'$ be a SVD of RVA^{-1}, where

$$\tilde{\Sigma}_{p\times d} = \left(\begin{array}{cc} \tilde{A}_{p\times p}, & \tilde{O}_{p\times(d-p)} \end{array}\right).$$

Then

$$
\begin{aligned}
\tilde{F}_\infty &= \left\{\tilde{U}\tilde{\Sigma}\tilde{V}'\left[B_d(1) - C_{12}C_{22}^{-1}B_q(1)\right]\right\}'\left[\tilde{U}\tilde{\Sigma}\tilde{V}'\left(C^{11}\right)^{-1}\tilde{V}\tilde{\Sigma}'\tilde{U}'\right]^{-1} \\
&\times \tilde{U}\tilde{\Sigma}\tilde{V}'\left[B_d(1) - C_{12}C_{22}^{-1}B_q(1)\right]/p \\
&= \left\{\tilde{\Sigma}\tilde{V}'\left[B_d(1) - C_{12}C_{22}^{-1}B_q(1)\right]\right\}'\left[\tilde{\Sigma}\tilde{V}'\left(C^{11}\right)^{-1}\tilde{V}\tilde{\Sigma}'\right]^{-1} \\
&\times \tilde{\Sigma}\tilde{V}'\left[B_d(1) - C_{12}C_{22}^{-1}B_q(1)\right]/p \\
&= \left\{\tilde{\Sigma}\left[\tilde{V}'B_d(1) - \tilde{V}'C_{12}C_{22}^{-1}B_q(1)\right]\right\}'\left[\tilde{\Sigma}\tilde{V}'\left(C^{11}\right)^{-1}\tilde{V}\tilde{\Sigma}'\right]^{-1} \\
&\times \tilde{\Sigma}\left[\tilde{V}'B_d(1) - \tilde{V}'C_{12}C_{22}^{-1}B_q(1)\right].
\end{aligned}
$$

Using the same steps, we have

$$
\begin{aligned}
& I_m - \Sigma\left[\Sigma'\tilde{W}_\infty^{-1}\Sigma\right]^{-1}\Sigma'\tilde{W}_\infty^{-1} \\
&= I_m - \begin{pmatrix} A \\ O \end{pmatrix}\left[A'C^{11}A\right]^{-1}\begin{pmatrix} A' & O' \end{pmatrix}\begin{pmatrix} C^{11} & C^{12} \\ C^{21} & C^{22} \end{pmatrix} \\
&= I_m - \begin{pmatrix} \left(C^{11}\right)^{-1} & O_{12} \\ O_{21} & O_{22} \end{pmatrix}\begin{pmatrix} C^{11} & C^{12} \\ C^{21} & C^{22} \end{pmatrix} \\
&= I_m - \begin{pmatrix} I_d & \left(C^{11}\right)^{-1}C^{12} \\ O_{21} & O_{22} \end{pmatrix} = \begin{pmatrix} O_{11} & -\left(C^{11}\right)^{-1}C^{12} \\ O_{21} & I_q \end{pmatrix}
\end{aligned}
$$

where O_{ij} are matrices of zeros with the dimensions as C_{ij}. So

$$\begin{aligned}
\tilde{J}_\infty &= \left[\begin{pmatrix} O_{11} & -\left(C^{11}\right)^{-1}C^{12} \\ O_{21} & I_q \end{pmatrix} B_m(1)\right]' \tilde{W}_\infty^{-1} \\
&\times \left[\begin{pmatrix} O_{11} & -\left(C^{11}\right)^{-1}C^{12} \\ O_{21} & I_q \end{pmatrix} B_m(1)\right] \\
&= \begin{pmatrix} -\left(C^{11}\right)^{-1}C^{12}B_q(1) \\ B_q(1) \end{pmatrix}' \begin{pmatrix} C^{11} & C^{12} \\ C^{21} & C^{22} \end{pmatrix} \begin{pmatrix} -\left(C^{11}\right)^{-1}C^{12}B_q(1) \\ B_q(1) \end{pmatrix} \\
&= \begin{pmatrix} -\left(C^{11}\right)^{-1}C^{12}B_q(1) \\ B_q(1) \end{pmatrix}' \begin{pmatrix} O \\ \left[C^{22} - C^{21}\left(C^{11}\right)^{-1}C^{21}\right] B_q(1) \end{pmatrix} \\
&= B_q(1)' \left[C^{22} - C^{21}\left(C^{11}\right)^{-1}C^{21}\right] B_q(1) \\
&= B_q(1)' C_{qq}^{-1} B_q(1).
\end{aligned}$$

In the last equality, we have used $\left[C^{22} - C^{21}\left(C^{11}\right)^{-1}C^{21}\right]^{-1} = C_{22} = C_{qq}$, which follows from the partitioned inverse formula.

Noting that the joint distribution of $\left[\tilde{V}'B_d\left(1\right), \tilde{V}'C_{12}, C_{22}, \ \tilde{V}'\left(C^{11}\right)^{-1}\tilde{V}\right]$ is invariant to the orthonormal matrix $\tilde{V}$, we have

$$\begin{pmatrix} \tilde{F}_\infty \\ \tilde{J}_\infty \end{pmatrix} \overset{d}{=} \begin{pmatrix} \tilde{F}_\infty^* \\ \tilde{J}_\infty^* \end{pmatrix}$$

where

$$\begin{aligned}
\tilde{F}_\infty^* &= \left\{\tilde{\Sigma}\left[B_d\left(1\right) - C_{12}C_{22}^{-1}B_q\left(1\right)\right]\right\}' \left[\tilde{\Sigma}\left(C^{11}\right)^{-1}\tilde{\Sigma}'\right]^{-1} \\
&\times \tilde{\Sigma}\left[B_d\left(1\right) - C_{12}C_{22}^{-1}B_q\left(1\right)\right]/p \\
&= \left\{\begin{pmatrix} \tilde{A}, & \tilde{O} \end{pmatrix}\left[B_d\left(1\right) - C_{12}C_{22}^{-1}B_q\left(1\right)\right]\right\}' \\
&\times \left[\begin{pmatrix} \tilde{A}, & \tilde{O} \end{pmatrix}\left(C^{11}\right)^{-1}\begin{pmatrix} \tilde{A}, & \tilde{O} \end{pmatrix}'\right]^{-1} \\
&\times \left\{\begin{pmatrix} \tilde{A}, & \tilde{O} \end{pmatrix}\left[B_d\left(1\right) - C_{12}C_{22}^{-1}B_q\left(1\right)\right]\right\}/p,
\end{aligned}$$

and

$$\tilde{J}_{\infty}^{*}=B_q(1)'C_{qq}^{-1}B_q(1).$$

Writing

$$\left(C^{11}\right)^{-1}=C_{11}-C_{12}C_{22}^{-1}C_{21}=\begin{pmatrix} D_{pp} & D^{12} \\ D^{21} & D^{22} \end{pmatrix}$$

where $D_{pp}=C_{pp}-C_{pq}C_{qq}^{-1}C_{pq}'$ and D^{22} is a $(d-p)\times(d-p)$ matrix and using equations (3.13)–(3.15), we have

$$\tilde{F}_{\infty}^{*}=\left[B_p(1)-C_{pq}C_{qq}^{-1}B_q(1)\right]'D_{pp}^{-1}\left[B_p(1)-C_{pq}C_{qq}^{-1}B_q(1)\right]/p.$$

So

$$\begin{pmatrix} \mathbb{W}_T(\hat{\theta}_T) \\ J_T(\hat{\theta}_T) \end{pmatrix}\overset{d}{\to}\begin{pmatrix} F_{\infty} \\ J_{\infty} \end{pmatrix}\overset{d}{=}\begin{pmatrix} \tilde{F}_{\infty}^{*} \\ \tilde{J}_{\infty}^{*} \end{pmatrix}.$$

The theorem then follows if we let $(F_{\infty},J_{\infty})'=(F_{\infty}^{*},J_{\infty}^{*})'$, which is innocuous for the weak convergence result. ■

Proof of Theorem 28. Part (a). Conditional on $B_q(\cdot):=\{B_q(r):r\in[0,1]\}$, both $B_p(1)$ and C_{pq} are normal. Hence conditional on $B_q(\cdot)$, we have

$$B_p(1)-C_{pq}C_{qq}^{-1}B_q(1)\sim N\left(0,I_p+E\left[C_{pq}C_{qq}^{-1}B_q(1)B_q(1)'C_{qq}^{-1}C_{qp}|B_q(\cdot)\right]\right). \tag{3.17}$$

Let $B_p^{(i)}(r)$ be the i-th element of $B_p(r)$. Define

$$C_{p_{(i)},q}=\int Q_K(r,s)\,dB_p^{(i)}(r)\,dB_q'(r)\in\mathbb{R}^{1\times q}$$
$$C_{q,p_{(j)}}=\int Q_K(r,s)\,dB_q(r)\,dB_p^{(j)}(r)\in\mathbb{R}^{q\times 1}$$

which are the i-th row of C_{pq} and j-th column of C_{qp}, respectively. Then the

(i,j)-th element of the conditional variance in (3.17) can be written as

$$
\begin{aligned}
& E\left\{C_{p(i),q}C_{qq}^{-1}B_q(1)B_q(1)'C_{qq}^{-1}C_{q,p(j)}|B_q(\cdot)\right\} \\
& = E\left\{\frac{1}{K}\sum_{\ell_1=1}^{K}\left(\int_0^1 \Phi_{\ell_1}(r)\,dB_p^{(i)}(r)\right)\right. \\
& \times\left(\int_0^1 \Phi_{\ell_1}(s)\,dB_q'(s)\right)C_{qq}^{-1}B_q(1)B_q(1)'C_{qq}^{-1}\frac{1}{K}\sum_{\ell_2=1}^{K}\left(\int_0^1 \Phi_{\ell_2}(\tilde{r})\,dB_q(\tilde{r})\right) \\
& \left.\left(\int_0^1 \Phi_{\ell_2}(\tilde{s})\,dB_p^{(j)}(\tilde{s})\right)\right| B_q(\cdot)\Big\} \\
& = E\left\{\frac{1}{K^2}\sum_{\ell_1,\ell_2}\left(\int_0^1 \Phi_{\ell_1}(r)\,dB_p^{(i)}(r)\right)\right. \\
& \times\underbrace{\left(\int_0^1 \Phi_{\ell_1}(s)\,dB_q'(s)\right)C_{qq}^{-1}B_q(1)B_q(1)'C_{qq}^{-1}\left(\int_0^1 \Phi_{\ell_2}(\tilde{r})\,dB_q(\tilde{r})\right)}_{a\ scalar} \\
& \times\left.\left(\int_0^1 \Phi_{\ell_2}(\tilde{s})\,dB_p^{(j)}(\tilde{s})\right)\right| B_q(\cdot)\Big\} \\
& = \delta_{ij}\frac{1}{K^2}\sum_{\ell_1,\ell_2}\left(\int_0^1 \Phi_{\ell_1}(r)\Phi_{\ell_2}(r)\,dr\right)\left(\int_0^1 \Phi_{\ell_1}(s)\,dB_q'(s)\right) \\
& \times C_{qq}^{-1}B_q(1)B_q(1)'C_{qq}^{-1}\left(\int_0^1 \Phi_{\ell_2}(\tilde{r})\,dB_q(\tilde{r})\right) \\
& = \delta_{ij}\frac{1}{K^2}\sum_{\ell_1=1}^{K}\left(\int_0^1 \Phi_{\ell_1}(s)\,dB_q'(s)\right)C_{qq}^{-1}B_q(1)B_q(1)'C_{qq}^{-1}\left(\int_0^1 \Phi_{\ell_1}(\tilde{r})\,dB_q(\tilde{r})\right) \\
& = \delta_{ij}\frac{1}{K^2}\sum_{\ell_1=1}^{K}B_q(1)'C_{qq}^{-1}\left(\int_0^1 \Phi_{\ell_1}(s)\,dB_q(s)\right)\left(\int_0^1 \Phi_{\ell_1}(\tilde{r})\,dB_q'(\tilde{r})\right)C_{qq}^{-1}B_q(1) \\
& = \frac{\delta_{ij}}{K}B_q(1)'C_{qq}^{-1}B_q(1),
\end{aligned}
$$

where $\delta_{ij} = 1\{i=j\}$. So, conditional on $B_q(\cdot)$,

$$B_p(1) - C_{pq}C_{qq}^{-1}B_q(1) \sim N\left[0, I_p\left(1+\frac{1}{K}B_q(1)'C_{qq}^{-1}B_q(1)\right)\right].$$

That is, conditional on $B_q(\cdot)$,

$$\frac{B_p(1) - C_{pq}C_{qq}^{-1}B_q(1)}{\sqrt{1 + \frac{1}{K}B_q(1)' C_{qq}^{-1}B_q(1)}} \sim N(0, I_p).$$

But $N(0, I_p)$ does not depend on $B_q(\cdot)$, so

$$\xi_p = \frac{B_p(1) - C_{pq}C_{qq}^{-1}B_q(1)}{\sqrt{1 + \frac{1}{K}B_q(1)' C_{qq}^{-1}B_q(1)}} \sim N(0, I_p)$$

unconditionally. In addition, ξ_p is independent of D_{pp}. Using these results, we have

$$\begin{aligned}
&\frac{F_\infty}{1 + \frac{1}{K}B_q'(1)\,C_{qq}^{-1}B_q(1)} \overset{d}{=} \frac{\xi_p' D_{pp}^{-1}\xi_p}{p} \overset{d}{=} \frac{\chi_p^2/p}{\chi_{K-p-q+1}^2/K} \\
&\overset{d}{=} \frac{K}{(K-p-q+1)} \frac{\chi_p^2/p}{\chi_{K-p-q+1}^2/(K-p-q+1)} \\
&\overset{d}{=} \frac{K}{(K-p-q+1)} F_{p,K-p-q+1}.
\end{aligned}$$

In view of Theorem 24, we have

$$\frac{K-p-q+1}{K} \frac{\mathbb{W}_T(\hat{\theta}_T)}{1 + \frac{q}{K}J_T(\hat{\theta}_T)} \overset{d}{\to} F_{p,K-p-q+1},$$

completing the proof of Part (a).

Using the same argument, we can prove Part (d). Parts (b) and (c) hold because the asymptotic equivalence of $\mathbb{W}_T(\hat{\theta}_T)$, $\mathbb{D}_T(\hat{\theta}_T)$ and $\mathbb{S}_T(\hat{\theta}_T)$ still holds under the fixed-smoothing asymptotics. For more details, see Sun (2014b). ■

Proof of Proposition 35. If $T^{-1}\sum_{t=1}^{T}\Phi_k(t/T) = 0$ for $k = 1, 2, \ldots, K$, then

$X = \sqrt{T} e_{K+1}$ where $e_{K+1} = (1, 0, ..., 0)'$ is the first unit vector in $\mathbb{R}^{K+1}$. So

$$\begin{aligned}
&\sqrt{T}\left(\hat{\theta}'_{T,OLS} - \theta'_0\right) \\
&= \left[\frac{X'}{\sqrt{T}} M_2 \frac{X}{\sqrt{T}}\right]^{-1} \frac{X'}{\sqrt{T}} M_2 \omega_\varepsilon = \left[e'_{K+1} M_2 e_{K+1}\right]^{-1} e'_{K+1} M_2 \omega_\varepsilon \\
&= \left[1 - e'_{K+1}\omega_2 \left(\omega'_2\omega_2\right)^{-1} \omega'_2 e_{K+1}\right]^{-1} \left[e'_{K+1}\omega_{u_1} - e'_{K+1}\omega_2 \left(\omega'_2\omega_2\right)^{-1} \omega'_2 \omega_{u_1}\right]
\end{aligned}$$

where ω_{u_1} is defined in the same way as ω_ε is defined. Let

$$S_{22} = \sum_{i=1}^{K} \omega_i(y_2)\,\omega_i(y_2)' \text{ and } S_{21} = \sum_{i=1}^{K} \omega_i(y_2)\,\omega_i(u_1)'.$$

Using the Sherman-Morrison formula, we have

$$\begin{aligned}
&e'_{K+1}\omega_2 \left(\omega'_2\omega_2\right)^{-1} \omega'_2 e_{K+1} \\
&= \omega_0(y_2)' \left(\omega_0(y_2)\,\omega'_0(y_2) + \sum_{i=1}^{K} \omega_i(y_2)\,\omega_i(y_2)'\right)^{-1} \omega_0(y_2) \\
&= T\bar{y}'_2 \left(T\bar{y}_2\bar{y}'_2 + S_{22}\right)^{-1} \bar{y}_2 = T\bar{y}'_2 \left[S_{22}^{-1} - \frac{S_{22}^{-1}\left(T\bar{y}_2\bar{y}'_2\right) S_{22}^{-1}}{1 + T\bar{y}'_2 S_{22}^{-1}\bar{y}_2}\right] \bar{y}_2 \\
&= T\bar{y}'_2 S_{22}^{-1}\bar{y}_2 - \frac{\left(T\bar{y}'_2 S_{22}^{-1}\bar{y}_2\right)^2}{1 + T\bar{y}'_2 S_{22}^{-1}\bar{y}_2} = \frac{T\bar{y}'_2 S_{22}^{-1}\bar{y}_2}{1 + T\bar{y}'_2 S_{22}^{-1}\bar{y}_2},
\end{aligned}$$

and

$$
\begin{aligned}
& e_{K+1}^{\prime}\omega_2\left(\omega_2^{\prime}\omega_2\right)^{-1}\omega_2^{\prime}\omega_{u_1} \\
& =\omega_0\left(y_2\right)^{\prime}\left[\omega_0\left(y_2\right)\omega_0^{\prime}\left(y_2\right)+\sum_{i=1}^{K}\omega_i\left(y_2\right)\omega_i\left(y_2\right)^{\prime}\right]^{-1} \\
& \left[\omega_0\left(y_2\right)\omega_0^{\prime}\left(u_1\right)+\sum_{i=1}^{K}\omega_i\left(y_2\right)\omega_i^{\prime}\left(u_1\right)\right] \\
& =\sqrt{T}\bar{y}_2^{\prime}\left(T\bar{y}_2\bar{y}_2^{\prime}+S_{22}\right)^{-1}\left[T\bar{y}_2\bar{u}_1^{\prime}+S_{21}\right] \\
& =\sqrt{T}\bar{y}_2^{\prime}\left[S_{22}^{-1}-\frac{S_{22}^{-1}\left(T\bar{y}_2\bar{y}_2^{\prime}\right)S_{22}^{-1}}{1+T\bar{y}_2^{\prime}S_{22}^{-1}\bar{y}_2}\right]\left[T\bar{y}_2\bar{u}_1^{\prime}+S_{21}\right] \\
& =\left(T\bar{y}_2^{\prime}S_{22}^{-1}\bar{y}_2\right)\sqrt{T}\bar{u}_1^{\prime}+\sqrt{T}\bar{y}_2^{\prime}S_{22}^{-1}S_{21} \\
& -\frac{\left(T\bar{y}_2^{\prime}S_{22}^{-1}\bar{y}_2\right)\times\left(T\bar{y}_2^{\prime}S_{22}^{-1}\bar{y}_2\right)\sqrt{T}\bar{u}_1^{\prime}}{1+T\bar{y}_2^{\prime}S_{22}^{-1}\bar{y}_2}-\frac{\left(T\bar{y}_2^{\prime}S_{22}^{-1}\bar{y}_2\right)\times\left(\sqrt{T}\bar{y}_2^{\prime}S_{22}^{-1}\right)}{1+T\bar{y}_2^{\prime}S_{22}^{-1}\bar{y}_2}S_{21} \\
& =\frac{\left(T\bar{y}_2^{\prime}S_{22}^{-1}\bar{y}_2\right)\sqrt{T}\bar{u}_1^{\prime}+\sqrt{T}\bar{y}_2^{\prime}S_{22}^{-1}S_{21}}{1+T\bar{y}_2^{\prime}S_{22}^{-1}\bar{y}_2}.
\end{aligned}
$$

Hence

$$
e_{K+1}^{\prime}M_2e_{K+1}=1-e_{K+1}^{\prime}\omega_2\left(\omega_2^{\prime}\omega_2\right)^{-1}\omega_2^{\prime}e_{K+1}=1-\frac{T\bar{y}_2^{\prime}S_{22}^{-1}\bar{y}_2}{1+T\bar{y}_2^{\prime}S_{22}^{-1}\bar{y}_2}=\frac{1}{1+T\bar{y}_2^{\prime}S_{22}^{-1}\bar{y}_2},
$$

$$
\begin{aligned}
e_{K+1}^{\prime}M_2\omega_{\varepsilon} &= e_{K+1}^{\prime}\omega_{u_1}-e_{K+1}^{\prime}\omega_2\left(\omega_2^{\prime}\omega_2\right)^{-1}\omega_2^{\prime}\omega_{u_1} \\
&=\sqrt{T}\bar{u}_1^{\prime}-\frac{\left(T\bar{y}_2^{\prime}S_{22}^{-1}\bar{y}_2\right)\sqrt{T}\bar{u}_1^{\prime}+\sqrt{T}\bar{y}_2^{\prime}S_{22}^{-1}S_{21}}{1+T\bar{y}_2^{\prime}S_{22}^{-1}\bar{y}_2}=\frac{\sqrt{T}\bar{u}_1^{\prime}-\sqrt{T}\bar{y}_2^{\prime}S_{22}^{-1}S_{21}}{1+T\bar{y}_2^{\prime}S_{22}^{-1}\bar{y}_2}.
\end{aligned}
$$

It then follows that

$$
\sqrt{T}\left(\hat{\theta}_{T,OLS}^{\prime}-\theta_0^{\prime}\right)=\sqrt{T}\left(\bar{u}_1^{\prime}-\bar{y}_2^{\prime}S_{22}^{-1}S_{21}\right)=\sqrt{T}\left(\bar{u}_1^{\prime}-\bar{u}_2^{\prime}S_{22}^{-1}S_{21}\right).
$$

It is easy to see that $S_{22}^{-1}S_{21}=\hat{\Omega}_{22}^{-1}\hat{\Omega}_{21}$. So

$$
\hat{\theta}_{T,OLS}-\theta_0=\bar{u}_1-\hat{\beta}\bar{u}_2=\hat{\theta}_{T,GMM}-\theta_0.
$$

This implies that $\hat{\theta}_{T,OLS} = \hat{\theta}_{T,GMM}$, as desired.

If $\int_0^1 \Phi_k(r)\,dr = 0$ but not $T^{-1}\sum_{t=1}^T \Phi_k(t/T) = 0$, then we have $X = \sqrt{T}e_{K+1} + O\left(1/\sqrt{T}\right)$. Using this and the assumptions in the proposition, we have

$$\sqrt{T}\left(\hat{\theta}'_{T,OLS} - \theta'_0\right) = \left[e'_{K+1}M_2 e_{K+1}\right]^{-1} e'_{K+1}M_2\omega_\varepsilon + o_p(1).$$

Following the same argument as above, we have $\sqrt{T}(\hat{\theta}_{T,OLS} - \theta_0) = \sqrt{T}(\hat{\theta}_{T,GMM} - \theta_0) + o_p(1)$, which implies that $\sqrt{T}(\hat{\theta}_{T,OLS} - \hat{\theta}_{T,GMM}) = o_p(1)$. ■

Proof of Proposition 36. We first give a representation of $\mathbb{W}_{CNLR}$. We focus on the case that $T^{-1}\sum_{t=1}^T \Phi_k(t/T) = 0$ for $k = 1, 2, \ldots, K$, as the other case follows from the similar arguments. Using $(\hat{\theta}_{T,OLS} - r)' = [X'M_2X]^{-1} X'M_2\omega_\varepsilon$, we have

$$\begin{aligned}
\mathbb{W}_{CNLR} &= \left(\hat{\theta}_{T,OLS} - r\right)' \left\{\tilde{\Omega}_{11\cdot 2}\left(X'M_2X\right)^{-1}\right\}^{-1}\left(\hat{\theta}_{T,OLS} - r\right)/p \\
&= [X'M_2X]^{-1} X'M_2\omega_\varepsilon \left\{\tilde{\Omega}_{11\cdot 2}\left(X'M_2X\right)^{-1}\right\}^{-1} \omega'_\varepsilon M_2 X\,[X'M_2X]^{-1} \\
&= \frac{X'M_2\omega_\varepsilon \times \tilde{\Omega}_{11\cdot 2}^{-1} \times \omega'_\varepsilon M_2 X}{X'M_2X}\frac{1}{p}
\end{aligned}$$

using the fact that $X'M_2X$ is a scalar.

In the proof of Proposition 35, we have shown that

$$\begin{aligned}
X'M_2\omega_\varepsilon &= \sqrt{T}e'_{K+1}M_2\omega_\varepsilon = \frac{T\left(\bar{u}'_1 - \bar{y}'_2 S_{22}^{-1}S_{21}\right)}{1 + T\bar{y}'_2 S_{22}^{-1}\bar{y}_2} \text{ and} \\
X'M_2X &= Te'_{K+1}M_2e_{K+1} = \frac{T}{1 + T\bar{y}'_2 S_{22}^{-1}\bar{y}_2}.
\end{aligned}$$

Hence

$$\mathbb{W}_{CNLR} = \frac{\sqrt{T}\left(\bar{u}'_1 - \bar{y}'_2 S_{22}^{-1}S_{21}\right)}{\sqrt{1 + T\bar{y}'_2 S_{22}^{-1}\bar{y}_2}} \times \tilde{\Omega}_{11\cdot 2}^{-1} \times \frac{\sqrt{T}\left(\bar{u}_1 - S_{12}S_{22}^{-1}\bar{y}_2\right)}{\sqrt{1 + T\bar{y}'_2 S_{22}^{-1}\bar{y}_2}}\frac{1}{p}.$$

To simplify $\tilde{\Omega}_{11\cdot 2}^{-1}$, we note that

$$\hat{\beta}'_{T,OLS} = \left[\omega'_2 M_X \omega_2\right]^{-1} \omega'_2 M_X \omega_1$$

where

$$M_X = I_{K+1} - X(X'X)^{-1}X' = I_{K+1} - e_{K+1}e'_{K+1} = \begin{pmatrix} 0 & 0 \\ 0 & I_K \end{pmatrix}.$$

So $\hat{\beta}'_{OLS} = S_{22}^{-1}S_{21}$. Plugging this and $\hat{\theta}_{OLS}$ into the estimated residuals yields

$$\begin{aligned}
&\omega_1 - X\hat{\theta}'_{T,OLS} - \omega_2\hat{\beta}'_{T,OLS} \\
&= \omega_\varepsilon - X\left(\hat{\theta}'_{T,OLS} - \theta'_0\right) - \omega_2\left(\hat{\beta}'_{T,OLS} - \beta'_0\right) \\
&= \omega_\varepsilon - X\left(\bar{u}'_1 - \bar{u}'_2 S_{22}^{-1}S_{21}\right) - \omega_2 S_{22}^{-1}S_{21} + \omega_2\beta'_0 \\
&= \begin{pmatrix} \sqrt{T}\left(\bar{\varepsilon}' - \bar{u}'_1 + \bar{u}'_2 S_{22}^{-1}S_{21} - \bar{u}'_2 S_{22}^{-1}S_{21} + \bar{u}_2\beta'_0\right) \\ \omega'_1(\varepsilon) - \omega'_2(u_2)\, S_{22}^{-1}S_{21} + \omega'_2(u_2)\,\beta'_0 \\ \dots \\ \omega'_K(\varepsilon) - \omega'_K(u_2)\, S_{22}^{-1}S_{21} + \omega'_K(u_2)\,\beta'_0 \end{pmatrix} \\
&= \begin{pmatrix} 0 \\ \omega'_1(u_1) - \omega'_2(u_2)\, S_{22}^{-1}S_{21} \\ \dots \\ \omega'_K(u_1) - \omega'_2(u_2)\, S_{22}^{-1}S_{21} \end{pmatrix}.
\end{aligned}$$

Therefore

$$\begin{aligned}
\tilde{\Omega}_{11\cdot 2} &= \frac{1}{K-q}\sum_{i=1}^{K}\left[\omega_i(u_1) - S_{12}S_{22}^{-1}\omega_i(u_2)\right]\left[\omega_i(u_1) - S_{12}S_{22}^{-1}\omega_i(u_2)\right]' \\
&= \frac{1}{K-q}\left(S_{11} - S_{12}S_{22}^{-1}S_{21}\right).
\end{aligned}$$

Using this and noting that $S_{ij} = K\hat{\Omega}_{ij}$, we have

$$\tilde{\Omega}_{11\cdot 2} = \frac{K}{K-q}\left(\hat{\Omega}_{11} - \hat{\Omega}_{12}\hat{\Omega}_{22}^{-1}\hat{\Omega}_{21}\right),$$

and so

$$\begin{aligned}\mathbb{W}_{CNLR} &= \frac{K-q}{K}\frac{\sqrt{T}\left(\bar{u}_1 - S_{12}S_{22}^{-1}\bar{y}_2\right)'}{\sqrt{1+T\bar{y}_2'S_{22}^{-1}\bar{y}_2}} \times \left(\hat{\Omega}_{11} - \hat{\Omega}_{12}\hat{\Omega}_{22}^{-1}\hat{\Omega}_{21}\right)^{-1} \\ &\quad\times \frac{\sqrt{T}\left(\bar{u}_1 - S_{12}S_{22}^{-1}\bar{y}_2\right)}{\sqrt{1+T\bar{y}_2'S_{22}^{-1}\bar{y}_{22}}}\frac{1}{p} \\ &= \frac{K-q}{K}\frac{\sqrt{T}\left(\bar{u}_1 - \hat{\beta}\bar{u}_2\right)'\left(\hat{\Omega}_{11} - \hat{\Omega}_{12}\hat{\Omega}_{22}^{-1}\hat{\Omega}_{21}\right)^{-1}\sqrt{T}\left(\bar{u}_1 - \hat{\beta}\bar{u}_2\right)}{1+\frac{1}{K}(\sqrt{T}\bar{u}_2)'\hat{\Omega}_{22}^{-1}(\sqrt{T}\bar{u}_2)}\frac{1}{p},\end{aligned}$$

where we have used $S_{12}S_{22}^{-1} = \hat{\beta}_{T,OLS} = \hat{\Omega}_{12}\hat{\Omega}_{22}^{-1} = \hat{\beta}$.

Next, we give a representation of $\mathbb{W}_T^c(\hat{\theta}_{T,GMM})$ when $R = I_p$. For the location model, $G_T(\hat{\theta}_{T,GMM})' = (I_p, O_{p\times q})$. We have

$$\mathbb{W}_T := \sqrt{T}(\hat{\theta}_{T,GMM} - r)'\left(\hat{\Omega}_{11} - \hat{\Omega}_{12}\hat{\Omega}_{22}^{-1}\hat{\Omega}_{21}\right)^{-1}\sqrt{T}(\hat{\theta}_{T,GMM} - r)/p.$$

Combining this with

$$J_T = (\sqrt{T}\bar{u}_2)'\hat{\Omega}_{22}^{-1}(\sqrt{T}\bar{u}_2),$$

we have

$$\begin{aligned}\mathbb{W}_T^c(\hat{\theta}_{T,GMM}) =& \frac{K-p-q+1}{K} \\ &\times \frac{\sqrt{T}\left(\bar{u}_1 - \hat{\beta}\bar{u}_2\right)'\left(\hat{\Omega}_{11} - \hat{\Omega}_{12}\hat{\Omega}_{22}^{-1}\hat{\Omega}_{21}\right)^{-1}\sqrt{T}\left(\bar{u}_1 - \hat{\beta}\bar{u}_2\right)}{1+\frac{1}{K}(\sqrt{T}\bar{u}_2)'\hat{\Omega}_{22}^{-1}(\sqrt{T}\bar{u}_2)}\frac{1}{p}.\end{aligned}$$

So

$$\mathbb{W}_{CNLR} = \frac{K-q}{K-p-q+1}\mathbb{W}_T^c(\hat{\theta}_{T,GMM}).$$

In particular, $\mathbb{W}_{CNLR} = \mathbb{W}_T^c(\hat{\theta}_{T,GMM})$ when $p = 1$. ■

References

[1] Anatolyev, S., 2005: Gmm, gel, serial correlation, and asymptotic bias. *Econometrica*, **73**(3), 983–1002.

[2] Anderson, T. W., and Hsiao, C., 1981: Estimation of dynamic models with error components. *Journal of the American statistical Association*, **76**(375), 598–606.

[3] Andrabi, T., Das, J., Khwaja, A. I., and Zajonc, T., 2011: Do value-added estimates add value? accounting for learning dynamics. *American Economic Journal: Applied Economics*, 29–54.

[4] Andrews, D. W., 1991: Heteroskedasticity and autocorrelation consistent covariance matrix estimation. *Econometrica: Journal of the Econometric Society*, 817–858.

[5] Arellano, M., and Bond, S., 1991: Some tests of specification for panel data: Monte carlo evidence and an application to employment equations. *The review of economic studies*, **58**(2), 277–297.

[6] Bell, R. M., and McCaffrey, D. F., 2002: Bias reduction in standard errors for linear regression with multi-stage samples. *Survey Methodology*, **28**(2), 169–182.

[7] 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.

[8] Bester, C. A., Conley, T. G., Hansen, C. B., and Vogelsang, T. J., 2016: Fixed-b asymptotics for spatially dependent robust nonparametric covariance matrix estimators. *Econometric Theory*, **32**(01), 154–186.

[9] Bharath, S., Dittmar, A., and Sivadasan, J., 2014: Do going-private transactions affect plant efficiency and investment? *Review of Financial Studies*, **27**(7), 1929–1976.

[10] Cameron, A. C., Gelbach, J. B., and Miller, D. L., 2008: Bootstrap-based improvements for inference with clustered errors. *The Review of Economics and Statistics*, **90**(3), 414–427.

[11] Canay, I. A., Romano, J. P., and Shaikh, A. M., 2014: Randomization tests under an approximate symmetry assumption. *manuscript, July*, **20**.

[12] Carter, A. V., Schnepel, K. T., and Steigerwald, D. G., 2013: Asymptotic behavior of at test robust to cluster heterogeneity.'. *University of California-Santa Barbara. Unpublished.*

[13] De Brauw, A., and Giles, J., 2012: Migrant labor markets and the welfare of rural households in the developing world: Evidence from china.

[14] Donald, S. G., and Lang, K., 2007: Inference with difference-in-differences and other panel data. *The review of Economics and Statistics*, **89**(2), 221–233.

[15] Dube, A., Lester, T. W., and Reich, M., 2010: Minimum wage effects across state borders: Estimates using contiguous counties. *The review of economics and statistics*, **92**(4), 945–964.

[16] Duflo, E., Mullainathan, S., and Bertrand, M., 2004: Vhow much should we trust difference in differences estimates. *Quarterly Journal of Economics119 (1): 249s275.*

[17] Emran, M. S., and Hou, Z., 2013: Access to markets and rural poverty: evidence from household consumption in china. *Review of Economics and Statistics*, **95**(2), 682–697.

[18] Gonçalves, S., 2011: The moving blocks bootstrap for panel linear regression models with individual fixed effects. *Econometric Theory*, **27**(05), 1048–1082.

[19] Goncalves, S., and Vogelsang, T. J., 2011: Block bootstrap hac robust tests: the sophistication of the naive bootstrap. *Econometric Theory*, **27**(04), 745–791.

[20] Hall, A. R., 2000: Covariance matrix estimation and the power of the overidentifying restrictions test. *Econometrica*, **68**(6), 1517–1527.

[21] Hansen, C. B., 2007: Asymptotic properties of a robust variance matrix estimator for panel data when T is large. *Journal of Econometrics*, **141**(2), 597–620.

[22] Hansen, L. P., 1982: Large sample properties of generalized method of moments estimators. *Econometrica: Journal of the Econometric Society*, 1029–1054.

[23] Hansen, L. P., Heaton, J., and Yaron, A., 1996: Finite-sample properties of some alternative gmm estimators. *Journal of Business and Economic Statistics*, **14**(3), 262–280.

[24] Hwang, J., and Sun, Y., 2015a: Should we go one step further? an accurate comparison of one-step and two-step procedures in a generalized method of moments framework. *An Accurate Comparison of One-Step and Two-Step Procedures in a Generalized Method of Moments Framework (July 24, 2015).*

[25] Hwang, J., and Sun, Y., 2015b: Asymptotic f and t tests in an efficient gmm setting.

[26] Ibragimov, R., and Müller, U. K., 2010: t-statistic based correlation and heterogeneity robust inference. *Journal of Business & Economic Statistics*, **28**(4), 453–468.

[27] Ibragimov, R., and Müller, U. K., 2011: Inference with few heterogeneous clusters. *Review of Economics and Statistics*, (00).

[28] Imbens, G. W., and Kolesar, M., 2012: Robust standard errors in small samples: Some practical advice. *Review of Economics and Statistics*, (0).

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

[30] Jenish, N., and Prucha, I. R., 2009: Central limit theorems and uniform laws of large numbers for arrays of random fields. *Journal of econometrics*, **150**(1), 86–98.

[31] Jenish, N., and Prucha, I. R., 2012: On spatial processes and asymptotic inference under near-epoch dependence. *Journal of econometrics*, **170**(1), 178–190.

[32] Kiefer, N. M., and Vogelsang, T. J., 2002a: Heteroskedasticity-autocorrelation robust testing using bandwidth equal to sample size. *Econometric Theory*, **18**(06), 1350–1366.

[33] Kiefer, N. M., and Vogelsang, T. J., 2002b: Heteroskedasticity-autocorrelation robust standard errors using the bartlett kernel without truncation. *Econometrica*, **70**(5), 2093–2095.

[34] Kiefer, N. M., and Vogelsang, T. J., 2005: A new asymptotic theory for heteroskedasticity-autocorrelation robust tests. *Econometric Theory*, **21**(06), 1130–1164.

[35] Kiefer, N. M., Vogelsang, T. J., and Bunzel, H., 2000: Simple robust testing of regression hypotheses. *Econometrica*, **68**(3), 695–714.

[36] 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.

[37] Lee, S., 2014: Asymptotic refinements of a misspecification-robust bootstrap for generalized method of moments estimators. *Journal of Econometrics*, **178**, 398–413.

[38] Lehmann, E. L., Romano, J. P., and Casella, G., 1986: *Testing statistical hypotheses*, volume 150. Wiley New York et al.

[39] Liang, K.-Y., and Zeger, S. L., 1986: Longitudinal data analysis using generalized linear models. *Biometrika*, **73**(1), 13–22.

[40] MacKinnon, J. G., and Webb, M. D., 2014: Wild bootstrap inference for wildly different cluster sizes. Technical report, Queen's Economics Department Working Paper.

[41] Marden, J., and Perlman, M. D., 1980: Invariant tests for means with covariates. *The Annals of Statistics*, 25–63.

[42] Muirhead, R. J., 2009: *Aspects of multivariate statistical theory*, volume 197. John Wiley & Sons.

[43] Müller, U. K., 2007: A theory of robust long-run variance estimation. *Journal of Econometrics*, **141**(2), 1331–1352.

[44] Newey, W. K., and Smith, R. J., 2004: Higher order properties of gmm and generalized empirical likelihood estimators. *Econometrica*, **72**(1), 219–255.

[45] Newey, W. K., and West, K. D., 1986: A simple, positive semi-definite, heteroskedasticity and autocorrelationconsistent covariance matrix.

[46] Nickell, S., 1981: Biases in dynamic models with fixed effects. *Econometrica: Journal of the Econometric Society*, 1417–1426.

[47] Obstfeld, M., Shambaugh, J. C., and Taylor, A. M., 2008: Financial stability, the trilemma, and international reserves. Technical report, National Bureau of Economic Research.

[48] Pepper, J. V., 2002: Robust inferences from random clustered samples: an application using data from the panel study of income dynamics. *Economics Letters*, **75**(3), 341–345.

[49] Phillips, P. C., 2005: Hac estimation by automated regression. *Econometric Theory*, **21**(01), 116–142.

[50] Politis, D. N., 2011: Higher-order accurate, positive semidefinite estimation of large-sample covariance and spectral density matrices. *Econometric Theory*, **27**(04), 703–744.

[51] Samila, S., and Sorenson, O., 2011: Venture capital, entrepreneurship, and economic growth. *The Review of Economics and Statistics*, **93**(1), 338–349.

[52] Shao, X., 2010: A self-normalized approach to confidence interval construction in time series. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, **72**(3), 343–366.

[53] Stock, J. H., and Watson, M. W., 2008: Heteroskedasticity-robust standard errors for fixed effects panel data regression. *Econometrica*, **76**(1), 155–174.

[54] Sun, Y., 2011a: Robust trend inference with series variance estimator and testing-optimal smoothing parameter. *Journal of Econometrics*, **164**(2), 345–366.

[55] Sun, Y., 2013: A heteroskedasticity and autocorrelation robust f test using an orthonormal series variance estimator. *The Econometrics Journal*, **16**(1), 1–26.

[56] Sun, Y., 2014a: Lets fix it: Fixed-b asymptotics versus small-b asymptotics in heteroskedasticity and autocorrelation robust inference. *Journal of Econometrics*, **178**, 659–677.

[57] Sun, Y., 2014b: Fixed-smoothing asymptotics in a two-step generalized method of moments framework. *Econometrica*, **82**(6), 2327–2370.

[58] Sun, Y., 2014c: Fixed-smoothing asymptotics and asymptotic f and t tests in the presence of strong autocorrelation. In *Essays in Honor of Peter CB Phillips*, 23–63. Emerald Group Publishing Limited.

[59] Sun, Y., and Kim, M. S., 2012: Simple and powerful gmm over-identification tests with accurate size. *Journal of Econometrics*, **166**(2), 267–281.

[60] Sun, Y., and Kim, M. S., 2015: Asymptotic f-test in a gmm framework with cross-sectional dependence. *Review of Economics and Statistics*, **97**(1), 210–223.

[61] Sun, Y., and Phillips, P., 2009: Bandwidth choice for interval estimation in gmm regression. working paper. *Department of Economics, UC San Diego.*

[62] Sun, Y., Phillips, P. C., and Jin, S., 2008: Optimal bandwidth selection in heteroskedasticity–autocorrelation robust testing. *Econometrica*, **76**(1), 175–194.

[63] 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.

[64] White, H., 1980: A heteroskedasticity-consistent covariance matrix estimator and a direct test for heteroskedasticity. *Econometrica: Journal of the Econometric Society*, 817–838.

[65] Windmeijer, F., 2005: A finite sample correction for the variance of linear efficient two-step gmm estimators. *Journal of econometrics*, **126**(1), 25–51.

[66] Zhang, X., 2015: Fixed-smoothing asymptotics in the generalized empirical likelihood estimation framework. *Department of Statistics, University of MissouriColumbia.*

[67] Zhang, X., Shao, X., et al., 2013: Fixed-smoothing asymptotics for time series. *The Annals of Statistics*, **41**(3), 1329–1349.