GMM with Many Weak Moment Conditions

GMM with Many Weak Moment Conditions*

Whitney K. Newey
Department of Economics
M.I.T.

Frank Windmeijer
Department of Economics
University of Bristol

September 2004
Revised, February 2008

Abstract

Using many moment conditions can improve efficiency but makes the usual GMM inferences inaccurate. Two step GMM is biased. Generalized empirical likelihood (GEL) has smaller bias but the usual standard errors are too small in instrumental variable settings. In this paper we give a new variance estimator for GEL that addresses this problem. It is consistent under the usual asymptotics and under many weak moment asymptotics is larger than the usual one, and is consistent. We also show that the Kleibergen (2005) Lagrange multiplier and conditional likelihood ratio statistics are valid under many weak moments. In addition we introduce a jackknife GMM estimator, but find that GEL is asymptotically more efficient under many weak moments. In Monte Carlo examples we find that t-statistics based on the new variance estimator have nearly correct size in a wide range of cases.

JEL Classification: C12, C13, C23
Keywords: GMM, Continuous Updating, Many Moments, Variance Adjustment

*The NSF provided financial support for this paper under Grant No. 0136869. Helpful comments were provided by J. Hausman, J. Powell, J. Stock, T. Rothenberg and four referees. This paper has been presented at the North American Winter Meeting and World Meeting of the Econometric Society, Cal Tech, Chicago, Copenhagen, Leuven, Manchester, NYU, Ohio State, Stanford, UC Berkeley, UCL, UCLA, and USC.

1 Introduction

Many applications of generalized method of moments (GMM, Hansen, 1982) have low precision. Examples include some natural experiments (Angrist and Krueger, 1991), consumption asset pricing models (Hansen and Singleton, 1982), and dynamic panel models (Holtz-Eakin, Newey and Rosen, 1988). In these settings the use of many moments can improve estimator accuracy. For example, Hansen, Hausman and Newey (2008) have recently found that in an application from Angrist and Krueger (1991), using 180 instruments, rather than 3, shrinks correct confidence intervals substantially.

A problem with using many moments is that the usual Gaussian asymptotic approximation can be poor. The two-step GMM estimator can be very biased. Generalized empirical likelihood (GEL, Smith 1997) and other estimators have smaller bias but the usual standard errors are found to be too small in examples in Han and Phillips (2006) and here. In this paper we use alternative asymptotics that addresses this problem in overidentified instrumental variable models that are weakly identified. Such environments seem quite common in econometric applications. Under the alternative asymptotics we find that GEL has a Gaussian limit distribution with asymptotic variance larger than the usual one. We give a new, "sandwich" variance estimator that is consistent under standard and many weak moment asymptotics. We find in Monte Carlo examples that, in a range of cases where identification is not very weak, t-ratios based on the new variance estimator have a better Gaussian approximation than the usual ones. We also show that the Kleibergen (2005) Lagrange multiplier (LM) and conditional likelihood ratio statistics, the Stock and Wright (2000) statistic, and the overidentifying statistic have asymptotically correct level under these asymptotics, but that the likelihood ratio statistic does not.

For comparison purposes we also consider a jackknife GMM estimator that generalizes jackknife instrumental variable (IV) estimators of Phillips and Hale (1977), Angrist, Imbens and Krueger (1999), and Blomquist and Dahlberg (1999). This estimator should also be less biased than the two-step GMM estimator. In the linear IV case Chao and

Swanson (2004) derived its limiting distribution under the alternative asymptotics. Here we show that jackknife GMM is asymptotically less efficient than GEL.

The alternative asymptotics is based on many weak moment sequences like those of Chao and Swanson (2004, 2005), Stock and Yogo (2005a), and Han and Phillips (2006). This paper picks up where Han and Phillips (2006) leave off, by showing asymptotic normality with an explicit formula for the asymptotic variance that is larger than the usual one and by giving a consistent variance estimator. This paper also extends Han and Phillips (2006) by giving primitive conditions for consistency and a limiting distribution when a heteroskedasticity consistent weight matrix is used for the continuous updating estimator (CUE), by analyzing GEL estimators other than the CUE, and by consideration of jackknife GMM.

The standard errors we give can be thought of as an extension of the Bekker (1994) standard errors from homoskedasticity and the limited information maximum likelihood (LIML) estimator to heteroskedasticity and GEL. Under homoskedasticity these standard errors and Bekker's (1994) have the same limit but the ones here are consistent under heteroskedasticity.

The asymptotics here is well suited for IV estimators but will not be particularly helpful for the type of minimum distance estimator considered in Altonji and Segal (1996). Estimation of the weighting matrix can strongly affect the properties of minimum distance estimators but the asymptotics here treats it as fixed.

The limiting distribution for GEL can be derived by increasing the number of moments in the Stock and Wright (2000) limiting distribution of the continuous updating estimator (CUE). This derivation corresponds to sequential asymptotics, where one lets the number of observations go to infinity and then lets the number of moments grow. We give here simultaneous asymptotics, where the number of moments grows along with, but slower than, the sample size.

One might also consider asymptotics where the number of moments increases at the same rate as the sample size, as Bekker (1994) did for LIML. It is harder to do this for GEL than for LIML, because GEL uses a heteroskedasticity consistent weighting matrix.

Consequently, estimation of all the elements of this weighting matrix has to be allowed for rather than just estimation of a scalar variance term. If the number of instruments grows as fast as the sample size the number of elements of the weight matrix grows as fast as the square of the sample size. It seems difficult to simultaneously control the estimation error for all these elements. Many weak moment asymptotics sidesteps this problem by allowing the number of moments to grow more slowly than the sample size, while accounting for the presence of many instruments by letting identification shrink.

In the linear heteroskedastic model we give primitive conditions for consistency and asymptotic normality of GEL estimators under many weak moments. For consistency of the CUE these conditions include a requirement that the number of moments m and the sample size n satisfy $m^2/n \longrightarrow 0$. This condition seems minimal given the need to control estimation of the weighting matrix. For asymptotic normality we require $m^3/n \longrightarrow 0$ for the CUE. We impose somewhat stronger rate conditions for other GEL estimators. In comparison, under homoskedasticity Stock and Yogo (2005a) require $m^2/n \longrightarrow 0$, Hansen, Hausman and Newey (2008) can allow m to grow at the same rate as n but restrict m to grow slower than the square of the concentration parameter, and Andrews and Stock (2006) require $m^3/n \longrightarrow 0$ when normality is not imposed. Of course one might expect somewhat stronger conditions with a heteroskedasticity consistent weighting matrix.

The new variance estimator from the many weak instrument asymptotics is different than Windmeijer (2005). That paper adjusts for the variability of the weight matrix while the many instrument asymptotics adjusts for the variability of the moment derivative.

In Section 2 we describe the model, the estimators, and the new asymptotic variance estimator. Test statistics that are robust to weak instruments and many weak instruments are described in Section 3. The alternative asymptotics is set up in Section 4. Section 5 calculates the asymptotic variance. Section 6 gives precise large sample results for GEL. Section 7 reports some Monte Carlo results. Section 8 offers some conclusions and some possible directions for future work. The Appendix gives proofs.

2 The Model and Estimators

The model we consider is for i.i.d. data where there is a countable number of moment restrictions. In the asymptotics we allow the data generating process to depend on the sample size. To describe the model, let w_i, $(i = 1, ..., n)$, be i.i.d. observations on a data vector w. Also, let β be a $p \times 1$ parameter vector and $g(w, \beta) = (g_1^m(w, \beta), ..., g_m^m(w, \beta))'$ be an $m \times 1$ vector of functions of the data observation w and the parameter, where $m \geq p$. For notational convenience we suppress an m superscript on $g(w, \beta)$. The model has a true parameter β_0 satisfying the moment condition

$$E[g(w_i, \beta_0)] = 0,$$

where $E[.]$ denotes expectation taken with respect to the distribution of w_i for sample size n, and we suppress the dependence on n for notational convenience. To describe the estimators and the asymptotic approximation we will use some notation. Let e_j denote the j^{th} unit vector and

$$\begin{aligned}
g_i(\beta) &= g(w_i, \beta), \hat{g}(\beta) = \sum_{i=1}^{n} g_i(\beta)/n, \hat{\Omega}(\beta) = \sum_{i=1}^{n} g_i(\beta) g_i(\beta)'/n, \\
\bar{g}(\beta) &= E[g_i(\beta)], g_i = g_i(\beta_0), \Omega(\beta) = E[g_i(\beta) g_i(\beta)'], \Omega = \Omega(\beta_0), \\
\hat{G}(\beta) &= \partial \hat{g}(\beta)/\partial \beta, G(\beta) = E[\partial g_i(\beta)/\partial \beta], G_i(\beta) = \partial g_i(\beta)/\partial \beta, G_i = G_i(\beta_0), \\
G &= G(\beta_0), B^j = \Omega^{-1} E[g_i e_j' G_i'], U_i^j = G_i e_j - G e_j - B^{j\prime} g_i, U_i = [U_i^1, ..., U_i^p].
\end{aligned}$$

An important example of this model is a single linear equation with instruments orthogonal to disturbances and heteroskedasticity of unknown form. This model is given by

$$y_i = x_i' \beta_0 + \varepsilon_i, \; x_i = \Upsilon_i + \eta_i, \tag{2.1}$$

$$E[\varepsilon_i | Z_i, \Upsilon_i] = 0, E[\eta_i | Z_i, \Upsilon_i] = \; 0,$$

where y_i is a scalar, x_i is a $p \times 1$ vector of right-hand side variables, Z_i is an $m \times 1$ vector of instrumental variables, and Υ_i is a $p \times 1$ vector of reduced form values. In this setting the moment functions are

$$g(w_i, \beta) = Z_i (y_i - x_i' \beta).$$

The notation for the linear model is then

$$\begin{aligned} g_i(\beta) &= Z_i(y_i - x_i'\beta), \hat{g}(\beta) = \sum_{i=1}^n Z_i(y_i - x_i'\beta)/n, \hat{\Omega}(\beta) = \sum_{i=1}^n Z_i Z_i'(y_i - x_i'\beta)^2/n, \\ \bar{g}(\beta) &= -E[Z_i \Upsilon_i'](\beta - \beta_0), \Omega(\beta) = E[Z_i Z_i'(y_i - x_i'\beta)^2], \Omega = E[Z_i Z_i' \varepsilon_i^2], g_i = Z_i \varepsilon_i, \\ \hat{G}(\beta) &= -\sum_{i=1}^n Z_i x_i'/n, G(\beta) = G = -E[Z_i \Upsilon_i'], G_i(\beta) = G_i = -Z_i x_i', \\ B^j &= -\Omega^{-1} E[Z_i Z_i' \varepsilon_i x_{ij}], U_i^j = -Z_i x_{ij} + E[Z_i x_{ij}] - B^{j\prime} Z_i \varepsilon_i. \end{aligned}$$

To describe the Hansen (1982) two step GMM estimator let $\dot{\beta}$ be a preliminary estimator and B be a compact set of parameter values. This estimator is given by

$$\ddot{\beta} = \arg\min_{\beta \in B} \ddot{Q}(\beta), \ddot{Q}(\beta) = \hat{g}(\beta)' \hat{W} \hat{g}(\beta)/2, \hat{W} = \hat{\Omega}(\dot{\beta})^{-1}.$$

The weighting matrix $\hat{W} = \hat{\Omega}(\dot{\beta})^{-1}$ is optimal in minimizing the asymptotic variance of $\ddot{\beta}$ under standard asymptotics.

The CUE has an analogous form where the objective function is simultaneously minimized over β in $\hat{\Omega}(\beta)$, i.e.

$$\hat{\beta} = \arg\min_{\beta \in B} \hat{Q}(\beta), \hat{Q}(\beta) = \hat{g}(\beta)' \hat{\Omega}(\beta)^{-1} \hat{g}(\beta)/2.$$

To describe a GEL estimator let $\rho(v)$ be a function of a scalar v that is concave on an open interval $\mathcal{V}$ containing zero and let $\rho_j(0) = \partial^j \rho(0)/\partial v^j$. We normalize $\rho(v)$ so that $\rho(0) = 0$, $\rho_1(0) = -1$ and $\rho_2(0) = -1$. Let $\hat{L}(\beta) = \{\lambda : \lambda' g_i(\beta) \in \mathcal{V}, i = 1, ..., n\}$. A GEL estimator is given by

$$\hat{\beta} = \arg\min_{\beta \in B} \hat{Q}(\beta), \ \hat{Q}(\beta) = \sup_{\lambda \in \hat{L}(\beta)} \sum_{i=1}^n \rho(\lambda' g_i(\beta))/n,$$

as in Smith (1997). The empirical likelihood (EL; Qin and Lawless, 1994, Imbens, 1997) estimator is obtained when $\rho(v) = \ln(1 - v)$ (and $\mathcal{V} = (-\infty, 1)$), and exponential tilting (ET, Imbens, 1997, Kitamura and Stutzer, 1997) when $\rho(v) = -e^v + 1$. When $\rho(v) = -v - v^2/2$ the objective function has an explicit form $\hat{Q}(\beta) = \hat{g}(\beta)' \hat{\Omega}(\beta)^{-1} \hat{g}(\beta)/2$ (Newey and Smith, 2004) and GEL is CUE.

To describe the new variance estimator for GEL, assume that

$$\hat{\lambda}(\beta) = \arg\max_{\lambda \in \hat{L}(\beta)} \sum_{i=1}^n \rho(\lambda' g_i(\beta))/n$$

exists (which will be true with probability approaching one in large samples) and let

$$\hat{D}(\beta) = \sum_{i=1}^{n} \hat{\pi}_i(\beta) \frac{\partial g_i(\beta)}{\partial \beta}, \hat{\pi}_i(\beta) = \frac{\rho_1(\hat{\lambda}(\beta)' g_i(\beta))}{\sum_{j=1}^{n} \rho_1(\hat{\lambda}(\beta)' g_j(\beta))}, (i = 1, ..., n).$$

For the CUE, the j^{th} column $\hat{D}_j(\beta)$ of $\hat{D}(\beta)$ will be taken to be

$$\hat{D}_j(\beta) = \frac{1}{n} \sum_{i=1}^{n} \frac{\partial g_i(\beta)}{\partial \beta_j} - \frac{1}{n} \sum_{i=1}^{n} \frac{\partial g_i(\beta)}{\partial \beta_j} g_i(\beta)' \hat{\Omega}(\beta)^{-1} \hat{g}(\beta).$$

In general, $\hat{D} = \hat{D}(\hat{\beta})$ is an efficient estimator of $G = E[\partial g_i(\beta_0)/\partial \beta]$, like that considered by Brown and Newey (1998). Also let

$$\hat{\Omega} = \hat{\Omega}(\hat{\beta}), \hat{H} = \frac{\partial^2 \hat{Q}(\hat{\beta})}{\partial \beta \partial \beta'}.$$

The estimator of the asymptotic variance of $\hat{\beta}$ is $\hat{V}/n$ where

$$\hat{V} = \hat{H}^{-1} \hat{D}' \hat{\Omega}^{-1} \hat{D} \hat{H}^{-1}.$$

When m is fixed and identification is strong, i.e. under "textbook" asymptotics, $\hat{V}$ will be consistent. In that case $\hat{g}(\hat{\beta}) \xrightarrow{p} 0$ so that $\hat{D} \xrightarrow{p} G$, and hence $\hat{V} \xrightarrow{p} (G' \Omega^{-1} G)^{-1}$, the textbook GMM asymptotic variance. The virtue of $\hat{V}$ is that it also consistent under the alternative, many weak moment asymptotics (when normalized appropriately).

Under the alternative asymptotics the asymptotic variance of $\hat{\beta}$ has a "sandwich" form that is estimated by $\hat{V}/n$. The matrix $\hat{G}' \hat{\Omega}^{-1} \hat{G}$, where $\hat{G} = \partial \hat{g}(\hat{\beta})/\partial \beta$, cannot be used in place of $\hat{H}$ in $\hat{V}$ because $\hat{G}' \hat{\Omega}^{-1} \hat{G}$ has a bias. This bias can be removed by using $\breve{H} = \sum_{i \neq j} \hat{G}_i' \hat{\Omega}^{-1} \hat{G}_j / n^2$ for $\hat{G}_i = \partial g_i(\hat{\beta})/\partial \beta$, but we do not consider this further because it did not work well in trial simulations. The middle term $\hat{D}' \hat{\Omega}^{-1} \hat{D}$ in $\hat{V}$ estimates a different, larger object than $\hat{H}$. It is an estimator of the asymptotic variance of $\partial \hat{Q}(\beta_0)/\partial \beta$ under weak identification due to Kleibergen (2005) for CUE and Guggenberger and Smith (2005) for other GEL objective functions. They show that this estimator can be used to construct a test statistic under weak identification with fixed m. Here we give conditions for consistency of a properly normalized version of $\hat{V}$ when m is allowed to grow with the sample size.The jackknife GMM estimator is obtained by deleting "own observation"

terms from the double sum that makes up the two-step GMM estimator, as

$$\breve{\beta} = \arg\min_{\beta \in B} \breve{Q}(\beta), \breve{Q}(\beta) = \sum_{i \neq j} g_i(\beta)' \breve{W} g_j(\beta)/2n^2, \breve{W} = \hat{\Omega}(\dot{\beta})^{-1},$$

where $\dot{\beta}$ is a preliminary jackknife GMM estimator based on a known choice of $\breve{W}$ (analogously to two step optimal GMM). For example, consider the linear model and let $\breve{P}_{ij} = Z_i' \breve{W} Z_j$. Here the jackknife GMM estimator is

$$\breve{\beta} = \left(\sum_{i \neq j} \breve{P}_{ij} x_i x_j' \right)^{-1} \sum_{i \neq j} \breve{P}_{ij} x_i y_j.$$

This estimator is a generalization of JIVE2 of Angrist, Imbens, and Krueger (1999) to allow a general weighting matrix $\breve{W}$.

To describe the variance estimator for jackknife GMM, let $\breve{\Omega} = \hat{\Omega}(\breve{\beta})$, $\breve{G}_i = G_i(\breve{\beta})$, $\breve{g}_i = g_i(\breve{\beta})$, and $\breve{G} = \sum_{i=1}^n \breve{G}_i/n$. Also let

$$\breve{H} = \sum_{i \neq j} \breve{G}_i' \breve{\Omega}^{-1} \breve{G}_j/n^2, \breve{\Lambda}_J = \sum_{i \neq j} \breve{G}_j' \breve{\Omega}^{-1} \breve{g}_i \breve{g}_j' \breve{\Omega}^{-1} \breve{G}_i/[n^2(n-1)].$$

The estimator of the asymptotic variance of $\breve{\beta}$ is $\breve{V}/n$ where

$$\breve{V} = \breve{H}^{-1} (\breve{G}' \breve{\Omega}^{-1} \breve{G} + \breve{\Lambda}_J) \breve{H}^{-1}.$$

This has a sandwich form like $\hat{V}$, with a jackknife estimator $\breve{H}$ of H rather than the Hessian $\hat{H}$ and an explicit adjustment term $\breve{\Lambda}_J$ for many moments. $\breve{V}$ will be consistent under both standard and many weak moment asymptotics, though we do not show this result here.

The many moment bias of two-step GMM with nonrandom $\hat{W}$ has a quite simple explanation that motivates CUE, GEL, and jackknife GMM. This explanation is also valid under many weak moments with a random $\hat{W}$, because estimation of $\hat{W}$ does not affect the limiting distribution. The absence of weighting matrix effects from many weak moment asymptotics indicates these asymptotics may not be a good approximation for minimum distance settings like those of Altonji and Segal (1996), where estimation of the weighting matrix is important.

Following Han and Phillips (2006), the bias is explained by the fact that the expectation of the objective function is not minimized at the truth. Since the objective function will be close to its expectation in large samples, the estimator will tend to be close to the minimum of the expectation, leading to bias. When $\hat{W}$ equals a nonrandom matrix W, the expectation of the GMM objective function is

$$\begin{aligned} E[\hat{g}(\beta)'W\hat{g}(\beta)/2] &= E[\sum_{i\neq j} g_i(\beta)'Wg_j(\beta) + \sum_{i=1}^{n} g_i(\beta)'Wg_i(\beta)]/2n^2 \\ &= (1-n^{-1})\bar{g}(\beta)'W\bar{g}(\beta)/2 + E[g_i(\beta)'Wg_i(\beta)]/2n \\ &= (1-n^{-1})\bar{g}(\beta)'W\bar{g}(\beta)/2 + tr(W\Omega(\beta))/2n. \end{aligned} \tag{2.2}$$

The term $(1-n^{-1})\bar{g}(\beta)'W\bar{g}(\beta)$ is a "signal" term that is minimized at β_0. The second term is a bias (or "noise") term that generally does not have zero derivative at β_0 (and hence is not minimized at β_0), when G_i is correlated with g_i, e.g. when endogeneity is present in the linear model. Also, when G_i and g_i are correlated the second term generally increases in size with the number of moments m. This increasing bias term leads to inconsistency of the two-step GMM estimator under many weak moments, as shown by Han and Phillips (2006). This bias also corresponds to the higher order bias term B_G in Newey and Smith (2004) that is important with endogeneity.

One way to remove this bias is to choose W so the bias does not depend on β. Note that if $W = \Omega(\beta)^{-1}$, then the bias term becomes $tr(W\Omega(\beta))/2n = tr(\Omega(\beta)^{-1}\Omega(\beta))/2n = m/2n$, which does not depend on β. A feasible version of this bias correction is to choose $\hat{W} = \hat{\Omega}(\beta)^{-1}$, leading to the objective function

$$\begin{aligned} \hat{Q}(\beta) &= \hat{g}(\beta)'\hat{\Omega}(\beta)^{-1}\hat{g}(\beta)/2 \\ &= \sum_{i\neq j} g_i(\beta)'\hat{\Omega}(\beta)^{-1}g_j(\beta)/2n^2 + \sum_{i=1}^{n} g_i(\beta)'\hat{\Omega}(\beta)^{-1}g_i(\beta)/2n^2 \\ &= \sum_{i\neq j} g_i(\beta)'\hat{\Omega}(\beta)^{-1}g_j(\beta)/2n^2 + m/2n. \end{aligned} \tag{2.3}$$

The estimator $\hat{\beta} = \arg\min_{\beta\in B}\ \hat{Q}(\beta)$ that minimizes this objective function is the CUE. It is interesting to note that it also has a jackknife GMM form.

Another way to remove the bias is to simply subtract an estimator $tr(\hat{W}\hat{\Omega}(\beta))/2n$ of

the bias term from the GMM objective function, giving

$$\breve{Q}(\beta) = \ddot{Q}(\beta) - tr(\hat{W}\hat{\Omega}(\beta))/2n = \sum_{i \neq j}[g_i(\beta)'\hat{W}g_j(\beta)]/2n^2,$$

giving the jackknife GMM objective function. The corresponding estimator will be consistent under many weak moment asymptotics because the own observation terms are the source of the bias in equation (2.2).

In what follows we will focus most of our attention on the GEL estimators. As shown below, when $\hat{W}$ is optimal in the usual GMM sense, the GEL estimators will be asymptotically more efficient than the jackknife GMM estimators under many weak moments. They are also inefficient relative to GEL in our Monte Carlo study, giving us further reason for our GEL focus.

3 Large Sample Inference

As shown by Dufour (1997) in linear models, if the parameter set is allowed to include values where the model is not identified then a correct confidence interval for a structural parameter must be unbounded with positive probability. Hence, bounded confidence intervals, such as Wald intervals formed in the usual way from $\hat{V}$, cannot be correct. Also, under the weak identification sequence of Stock and Wright (2000) the Wald confidence intervals will not be correct, i.e. the new variance estimator is not robust to weak identification. These observations motivate consideration of statistics that are asymptotically correct with weak or many weak moment conditions.

One identification robust statistic proposed by Stock and Wright (2000) is a GMM version of the Anderson Rubin statistic. For the null hypothesis $H_0 : \beta_0 = \beta$, where β is known, the GEL version of this statistic, as given by Guggenberger and Smith (2005), is

$$AR(\beta) = 2n\hat{Q}(\beta). \tag{3.4}$$

Under the null hypothesis and weak identification, or many weak moments, treating this as if it were distributed as $\chi^2(m)$ will be asymptotically correct. As a result we can form a joint confidence interval for the vector β by inverting $AR(\beta)$. Specifically, for the $1-\alpha$

quantile q_α^m of a $\chi^2(m)$ distribution an asymptotic $1-\alpha$ confidence interval for β will be $\{\beta : AR(\beta) \leq q_\alpha^m\}$. This confidence interval will be valid under weak identification and under many weak moments. However, there are other confidence intervals that have this property but are smaller in large samples, thus producing more accurate inference.

One of these is the Kleibergen (2005) and Guggenberger and Smith (2005) Lagrange multiplier (LM) statistic for GEL. For the null hypothesis $H_0 : \beta_0 = \beta$, where β is known, the LM statistic is

$$LM(\beta) = n\frac{\partial \hat{Q}(\beta)}{\partial \beta}' [\hat{D}(\beta)'\hat{\Omega}(\beta)^{-1}\hat{D}(\beta)]^{-1}\frac{\partial \hat{Q}(\beta)}{\partial \beta}. \tag{3.5}$$

Under the null hypothesis and weak identification or many weak moments this statistic will have a $\chi^2(p)$ limiting distribution. As a result we can form joint confidence intervals for the vector β_0 by inverting $LM(\beta)$. Specifically, for the $1-\alpha$ quantile q_α^p of a $\chi^2(p)$ distribution, an asymptotic $1-\alpha$ confidence interval is $\{\beta : LM(\beta) \leq q_\alpha^p\}$. These confidence intervals are also correct in the weak identification setting of Stock and Wright (2000).

Kleibergen (2005) also proposed a GMM analog of the conditional likelihood ratio (CLR) test of Moreira (2003), motivated by the superior performance of the analogous CLR statistic, relative to LM, in the linear homoskedastic model. Smith (2006) extended this statistic to GEL. Here we consider one version.

Let $\hat{R}(\beta)$ be some statistic which should be large if the parameters are identified and small if not, and with fixed m depends only on $\hat{D}(\beta)$ asymptotically. Kleibergen (2005) suggests to use a statistic of a null hypothesis about the rank of $\hat{D}(\beta)$. We consider a simple choice of $\hat{R}(\beta)$ given by

$$\hat{R}(\beta) = n\xi_{\min}(\hat{D}(\beta)'\hat{\Omega}(\beta)^{-1}\hat{D}(\beta)),$$

where $\xi_{\min}(A)$ denotes the smallest eigenvalue of A. A version of the GEL-CLR statistic is

$$CLR(\beta) = \frac{1}{2}\left\{AR(\beta) - \hat{R}(\beta) + \left[\left(AR(\beta) - \hat{R}(\beta)\right)^2 + 4LM(\beta)\hat{R}(\beta)\right]^{1/2}\right\}. \tag{3.6}$$

Under the null hypothesis $H_0 : \beta_0 = \beta$ a level α critical value $\hat{q}_\alpha(\beta)$ for this test statistic can be simulated. Let (q_s^{m-p}, q_s^p), $s = 1, ..., S$, be i.i.d. draws (independent from each other and over s) of $\chi^2(m-p)$ and $\chi^2(p)$ random variables. Let $\hat{q}_\alpha(\beta)$ be the $1-\alpha$ quantile of

$$\left\{ \frac{1}{2} \left\{ q_s^{m-p} + q_s^p - \hat{R}(\beta) + \left[\left(q_s^{m-p} + q_s^p - \hat{R}(\beta) \right)^2 + 4 q_s^p \hat{R}(\beta) \right]^{1/2} \right\} ; s = 1, ..., S \right\}.$$

An asymptotic $1-\alpha$ confidence interval can then be formed as $\{\beta : CLR(\beta) \leq \hat{q}_\alpha(\beta)\}$. These confidence intervals will be correct under weak identification and also under many weak moment conditions.

Another test statistic of interest is the overidentification statistic $AR(\hat{\beta})$. This statistic is often used to test all the overidentifying restrictions associated with the moment conditions. Under a fixed number of moment conditions this statistic converges in distribution to $\chi^2(m-p)$ and the critical value for this distribution remains valid under many weak moments. Thus, it will be the case that $\Pr(AR(\hat{\beta}) > q_\alpha^{m-p}) \longrightarrow \alpha$.

In addition to these statistics Hansen, Heaton, and Yaron (1996) considered the likelihood ratio statistic corresponding to the CUE. For GEL this statistic takes the form

$$LR(\beta) = 2n \left[\hat{Q}(\beta) - \hat{Q}(\hat{\beta}) \right].$$

As discussed in Stock and Wright (2000), this statistic does not have a chi-squared limiting distribution under weak identification. We show that it also does not under many weak moments. We find that the critical value for a chi-squared distribution leads to overrejection, so that the confidence interval based on this statistic is too small.

Under local alternatives and many weak moments, one could compare the power of some of these test statistics as a test of $H_0 : \beta = \beta_0$. The Wald statistic is $\hat{T} = n(\hat{\beta} - \beta_0)' \hat{V}^{-1} (\hat{\beta} - \beta_0)$. We will show that there is a bounded sequence $\{c_n\}$ with c_n bounded positive such that

$$LM(\beta_0) = \hat{T} + o_p(1); \; CLR(\beta_0) = c_n \hat{T} + o_p(1).$$

Thus, the Wald test based on $\hat{T}$ will be asymptotically equivalent under the null hypothesis and contiguous alternatives to the LM and CLR tests. The implied asymptotic

equivalence of LM and CLR is a GMM version of a result of Andrews and Stock (2006). In contrast, a test based on $AR(\beta_0)$ will have asymptotic local power equal to size, because its degrees of freedom goes to infinity. However these comparisons do not hold up under weak identification. No power ranking of these statistics is known in that case.

The new variance estimator seems useful despite the lack of robustness to weak instruments. Standard errors are commonly used in practice as a measure of uncertainty associated with an estimate. Also, for multidimensional parameters the confidence intervals based on the LM or CLR are more difficult to compute. Confidence ellipses can be formed in the usual way from $\hat{\beta}$ and $\hat{V}$ while LM or CLR confidence sets need to be calculated by an exhaustive grid search. Furthermore, the conditions for an accurate many weak moment approximation seem to occur often in applications, as further discussed below. For all these reasons, the standard errors given here seem useful for econometric practice.

It does seem wise to check for weak moments in practice. One could develop GMM versions of the Hahn and Hausman (2004) and/or Stock and Yogo (2005b) tests. One could also compare a Wald test based on the corrected standard errors with a test based on an identification robust statistic.

4 Many Weak Moment Approximation

As always, asymptotic theory is meant to provide an approximation to the distribution of objects of interest in applications. The theory and Monte Carlo results below indicate that many weak moment asymptotics, applied to $\hat{\beta}$ and $\hat{V}$, should provide an improvement in 1) overidentified models where 2) the variance of the Jacobian of the moment functions is large relative to its average and 3) the parameters are quite well identified. Condition 2) is often true in IV settings, tending to hold when reduced form R^2s are low. Condition 3) is also often true in IV settings (e.g. see the brief applications survey in Hansen, Hausman and Newey, 2008).

The many weak moment asymptotics will not provide an improved approximation in

minimum distance settings where $g(w,\beta) = g_1(w) - g_2(\beta)$. In that setting $\partial g_i(\beta_0)/\partial\beta$ is constant, so that condition 2) will not hold. In fact, the asymptotic variance under many weak moments will be the same as the usual variance.

Conditions 1), 2), and 3) are simultaneously imposed in the asymptotics, where 1) m grows, 2) some components of $G'\Omega^{-1}G$ go to zero, so that the variance of $\partial g_i(\beta_0)/\partial\beta$ is large relative to G, and 3) $nG'\Omega^{-1}G$ grows, so that the parameters are identified. The following specific condition incorporates each of 1), 2), and 3).

ASSUMPTION 1: *i) There is a $p \times p$ matrix $S_n = \tilde{S}_n \operatorname{diag}(\mu_{1n}, ..., \mu_{pn})$ such that $\tilde{S}_n$ is bounded, the smallest eigenvalue of $\tilde{S}_n\tilde{S}_n'$ is bounded away from zero, for each j either $\mu_{jn} = \sqrt{n}$ or $\mu_{jn}/\sqrt{n} \longrightarrow 0$, $\mu_n = \min_{1\leq j\leq p} \mu_{jn} \longrightarrow \infty$, and m/μ_n^2 is bounded; ii) $nS_n^{-1}G'\Omega^{-1}GS_n^{-1\prime} \longrightarrow H$ and H is nonsingular.*

This assumption allows for linear combinations of β to have different degrees of identification, similarly to Hansen, Hausman and Newey (2008). For example, when a constant is included one might consider the corresponding reduced form coefficient to be strongly identified. This will correspond to $\mu_{jn} = \sqrt{n}$. For less strong identification μ_{jn} will be allowed to grow slower than $\sqrt{n}$. This condition is a GMM version of one of Chao and Swanson (2005) for IV. It generalizes Han and Phillips (2006) to allow μ_{jn} to differ across j.

The linear model of equation (2.1) is an example. Suppose that it has reduced form and instruments given by

$$x_i = (z_{1i}', x_{2i}')', x_{2i} = \pi_{21}z_{1i} + \frac{\mu_n}{\sqrt{n}}z_{2i} + \eta_{2i}, Z_i = (z_{1i}', Z_{2i}')',$$

where z_{1i} is a $p_1 \times 1$ vector of included exogenous variables, z_{2i} is a $(p - p_1) \times 1$ vector of excluded exogenous variables, and Z_{2i} is an $(m - p_1) \times 1$ vector of instruments. This specification allows for constants in the structural equation and reduced form by allowing an element of z_{1i} to be 1. The variables z_{2i} may not be observed by the econometrician. For example, we could have $z_{2i} = f_0(w_i)$ for a vector of underlying exogenous variables w_i and an unknown vector of functions $f_0(w)$. In this case the instrument vector could

be $Z_i = (z_{1i}', p_{1,m-p_1}(w_i), ..., p_{m-p_1,m-p_1}(w_i))'$, where $p_{j,m-p_1}(w_i)$, $(j = 1, ..., m - p_1)$ are approximating functions, such as power series or splines. In this case the model is like Newey (1990), except that the coefficient the unknown function $f_0(w_i)$ goes to zero to model weaker identification.

To see how Assumption 1 is satisfied in this example, let

$$\tilde{S}_n = \begin{pmatrix} I_{p_1} & 0 \\ \pi_{21} & I_{p-p_1} \end{pmatrix}, \mu_{jn} = \begin{cases} \sqrt{n} : j = 1, ..., p_1 \\ \mu_n : j = p_1 + 1, ..., p \end{cases}.$$

Then for $z_i = (z_{1i}', z_{2i}')'$ the reduced form is

$$\Upsilon_i = \begin{pmatrix} z_{1i} \\ \pi_{21} z_{1i} + \frac{\mu_n}{\sqrt{n}} z_{2i} \end{pmatrix} = S_n z_i / \sqrt{n}, G = -E[Z_i \Upsilon_i'] = -E[Z_i z_i'] S_n' / \sqrt{n}.$$

Assume that z_i and Z_i are functions of some variables $\tilde{z}_i$ and let $\sigma_i^2 = E[\varepsilon_i^2 | \tilde{z}_i] > 0$ and $z_i^* = z_i / \sigma_i^2$. Then

$$\begin{aligned} n S_n^{-1} G' \Omega^{-1} G S_n^{-1\prime} &= E[z_i Z_i'] \Omega^{-1} E[Z_i z_i'] \\ &= E[\sigma_i^2 z_i^* Z_i'] (E[\sigma_i^2 Z_i Z_i'])^{-1} E[\sigma_i^2 Z_i z_i^{*\prime}]. \end{aligned}$$

The expression following the second equality is the mean square error of a linear projection of z_i^* on Z_i, weighted by σ_i^2. Therefore, if linear combinations of Z_i can approximate z_i^*, i.e. if there is π_m such that $\lim_{m \longrightarrow \infty} E[\sigma_i^2 \left\| z_i^* - \pi_m Z_i \right\|^2] = 0$, then

$$n S_n^{-1} G' \Omega^{-1} G S_n^{-1\prime} \longrightarrow E[\sigma_i^2 z_i^* z_i^{*\prime}] = E[\sigma_i^{-2} z_i z_i'].$$

Then it suffices for Assumption 1 to assume that $E[\sigma_i^{-2} z_i z_i']$ is nonsingular.

Asymptotic normality will lead to different convergence rates for linear combinations of the coefficients. In the linear model example just considered, where $\beta = (\beta_1', \beta_2')'$, it will be the case that

$$S_n'(\hat{\beta} - \beta) = \begin{pmatrix} \sqrt{n}[(\hat{\beta}_1 - \beta_1) + \pi_{21}'(\hat{\beta}_2 - \beta_2)] \\ \mu_n(\hat{\beta}_2 - \beta_2) \end{pmatrix}$$

is jointly asymptotically normal. Thus, the coefficients $\hat{\beta}_2$ of the endogenous variables converge at rate $1/\mu_n$ but the coefficients of included exogenous variables $\hat{\beta}_1$ need not converge at rate $1/\sqrt{n}$. Instead, it is the linear combination $\hat{\beta}_1 + \pi_{21}' \hat{\beta}_2$ that converges

at rate $1/\sqrt{n}$. Note that $\beta_1 + \pi_{21}'\beta_2$ is the coefficient of z_{1i} in the reduced form equation for y_i. Thus, it is the reduced form coefficient that converges to the truth at rate $1/\sqrt{n}$. In general, all the structural coefficients may converge at the rate $1/\mu_n$. In that case the asymptotic variance matrix of $\mu_n(\hat{\beta} - \beta_0)$ will be singular with rank equal to p_2. Wald tests of up to p_2 linear combinations can still have the usual asymptotic distribution, but tests of more than p_2 linear combinations would need to account for singularity of the asymptotic variance of $\mu_n(\hat{\beta} - \beta_0)$.

The many weak moment asymptotic variance is larger than the usual one when m grows at the same rate as μ_n^2, e.g. when $\mu_n^2 = m$. In the linear model this corresponds to a reduced form

$$x_{2i} = \pi_{21} z_{1i} + \frac{\sqrt{m}}{\sqrt{n}} z_{2i} + \eta_{i2}.$$

This sequence of models is a knife-edge case where the additional variance due to many instruments is the same size as the usual one. If μ_n^2 grew faster than m the usual variance would dominate while if μ_n^2 grew slower than m the additional term would dominate in the asymptotic variance. The case with μ_n^2 growing slower than m is ruled out by Assumption 1 but is allowed in some work on the linear model, e.g. see Chao and Swanson (2004) and Hansen, Hausman and Newey (2008).

One specification where μ_n^2 and m grow at the same rate has

$$z_{2i} = C \sum_{j=1}^{m-p_1} Z_{2ij}/\sqrt{m}, E[Z_{2i} Z_{2i}'] = I_{m-p_1},$$

where C is an unknown constant. In that case the reduced form is

$$x_{2i} = \pi_{21} z_{1i} + \sum_{j=1}^{m-p_1} \frac{C}{\sqrt{n}} Z_{2ij} + \eta_{i2}.$$

This is a many weak instrument specification like that considered by Chao and Swanson (2004, 2005).

Despite the knife-edge feature of these asymptotics, we find in simulations below that using the asymptotic variance estimate provides greatly improved approximation in a wide range of cases. Given these favorable results one might expect that the new variance estimator provides an improved approximation more generally than just when

m grows at the same rate as μ_n^2. Hansen, Hausman and Newey (2008) did find such a result for the Bekker (1994) variance in a homoskedastic linear model, and the new variance here extends that to GEL and heteroskedasticity, so we might expect a similar result here. Showing such a result is beyond the scope of this paper though we provide some theoretical support for the linear model example in the next Section.

5 Asymptotic Variances

To explain and interpret the results we first give a formal derivation of the asymptotic variance for GEL and jackknife GMM. We begin with jackknife GMM because it is somewhat easier to work with. The usual Taylor expansion of the first-order condition $\partial\breve{Q}(\breve{\beta})/\partial\beta = 0$ gives

$$S_n'(\breve{\beta} - \beta_0) = -\bar{H}^{-1} n S_n^{-1} \partial\breve{Q}(\beta_0)/\partial\beta, \bar{H} = n S_n^{-1} \partial^2 \breve{Q}(\bar{\beta})/\partial\beta\partial\beta' S_n^{-1\prime},$$

where $\bar{\beta}$ is an intermediate value for β, being on the line joining $\breve{\beta}$ and β_0 (that actually differs from row to row of $\bar{H}$). Under regularity conditions it will be the case that

$$\bar{H} \xrightarrow{p} H_W = \lim_{n\longrightarrow\infty} n S_n^{-1} G' W G S_n^{-1\prime},$$

where we assume that $\hat{W}$ estimates a matrix W in such a way that the remainders are small and that the limit of $n S_n^{-1} G' W G S_n^{-1\prime}$ exists. The asymptotic distribution of $S_n'(\breve{\beta} - \beta)$ will then equal the asymptotic distribution of $-H_W^{-1} n S_n^{-1} \partial\breve{Q}(\beta_0)/\partial\beta$.

The estimation of the weighting matrix will not affect the asymptotic distribution, so that differentiating the jackknife GMM objective function and replacing $\hat{W}$ with its limit W, gives

$$\begin{aligned} n S_n^{-1} \partial\breve{Q}(\beta_0)/\partial\beta &= \sum_{i\neq j} S_n^{-1} G_i' W g_j/n + o_p(1) \\ &= (1 - n^{-1})\sqrt{n} S_n^{-1} G' W \sqrt{n}\hat{g}(\beta_0) + \sum_{j<i} \psi_{ij}^J/n + o_p(1), \\ \psi_{ij}^J &= S_n^{-1}(G_j - G)' W g_i + S_n^{-1}(G_i - G)' W g_j, \end{aligned}$$

where the second equality holds by adding and subtracting G to G_i. The $\sqrt{n} S_n^{-1} G' W \sqrt{n}\hat{g}(\beta_0)$ term is the usual GMM one, having asymptotic variance $H_\Omega = \lim_{n\longrightarrow\infty} n S_n^{-1} G' W \Omega W G S_n^{-1\prime}$,

assumed to exist. The other term $\sum_{j<i} \psi_{ij}^J/n$ is a degenerate U-statistic, a martingale sum that turns out to be asymptotically normal under regularity conditions, as in Lemma A10 of the Appendix. Its asymptotic variance will be the limit of

$$\begin{aligned} E[\psi_{ij}^J \psi_{ij}^{J\prime}]/2 &= S_n^{-1}\{E[(G_j - G)'W g_i g_i' W(G_j - G)] + E[(G_j - G)'W g_i g_j' W(G_i - G)]\}S_n^{-1\prime} \\ &= S_n^{-1}\{E[(G_j - G)'W\Omega W(G_j - G)] + E[G_j' W g_i g_j' W G_i]\}S_n^{-1\prime} \\ &= S_n^{-1}(E[G_j' W\Omega W G_j] - G'W\Omega WG + E[G_j' W g_i g_j' W G_i])S_n^{-1\prime}. \end{aligned}$$

This limit is equal to

$$\Lambda_J = \lim_{n\longrightarrow\infty} E[\psi_{ij}^J \psi_{ij}^{J\prime}]/2 = \lim_{n\longrightarrow\infty} S_n^{-1}(E[G_j' W\Omega W G_j] + E[G_j' W g_i g_j' W G_i])S_n^{-1\prime}.$$

The U-statistic term is uncorrelated with the usual GMM term, so by the central limit theorem, $nS_n^{-1}\partial\breve{Q}(\beta_0)/\partial\beta \xrightarrow{d} N(0, H_\Omega + \Lambda_J)$. It then will follow that

$$S_n'(\breve{\beta} - \beta_0) \xrightarrow{d} N(0, V_J), V_J = H_W^{-1} H_\Omega H_W^{-1} + H_W^{-1}\Lambda_J H_W^{-1},$$

a result that was previously derived for the JIVE2 estimator by Chao and Swanson (2004).

For GEL we will focus on the asymptotic variance of the CUE because the explicit form of the CUE simplifies the discussion. The other GEL estimators will have the same asymptotic variance, essentially because $\hat{Q}(\beta)$ will be quadratic in $\hat{g}(\beta)$ near β_0.

To derive the CUE asymptotic variance we expand the first-order conditions similarly to jackknife GMM. That gives an analogous expression for $S_n'(\hat{\beta} - \beta_0)$ with the CUE objective function $\hat{Q}(\beta)$ replacing the jackknife GMM objective $\breve{Q}(\beta)$. It will turn out that $nS_n^{-1}\partial^2\hat{Q}(\bar{\beta})/\partial\beta\partial\beta' S_n^{-1\prime} \xrightarrow{p} H$ from Assumption 1, so that the Hessian term is the same for the CUE as for jackknife GMM. However, the other term in the variance will be different. To derive it, recall the definitions of B^j and U_i from Section 2, and note that the columns of U_i are the population residuals from least squares regression of columns of $G_i - G$ on g_i. Assuming we can differentiate under the integral we have

$$\frac{\partial\Omega(\beta_0)^{-1}}{\partial\beta_j} = -\Omega^{-1}\left[\frac{\partial\Omega(\beta_0)}{\partial\beta_j}\right]\Omega^{-1} = -B^j\Omega^{-1} - \Omega^{-1}B^{j\prime}.$$

Then differentiating the CUE objective function with $\Omega(\beta)^{-1}$ replacing $\hat{\Omega}(\beta)^{-1}$ we have

$$\begin{aligned} nS_n^{-1}\frac{\partial \hat{Q}(\beta_0)}{\partial \beta} &= nS_n^{-1}\frac{\partial}{\partial \beta}\left\{\hat{g}(\beta)'\Omega^{-1}\hat{g}(\beta)+\hat{g}(\beta_0)'\Omega(\beta)^{-1}\hat{g}(\beta_0)\right\}|_{\beta=\beta_0}/2 \\ &= S_n^{-1}\frac{1}{n}\sum_{i,j=1}^{n}(G+U_i)'\Omega^{-1}g_j \\ &= \sqrt{n}S_n^{-1}G'\Omega^{-1}\sqrt{n}\hat{g}(\beta_0)+\sum_{j<i}\psi_{ij}/n+S_n^{-1}\sum_{i=1}^{n}U_i'\Omega^{-1}g_i/n, \\ \psi_{ij} &= S_n^{-1}(U_j'\Omega^{-1}g_i+U_i'\Omega^{-1}g_j). \end{aligned}$$

By the projection residual form of U_i, each component of U_i is uncorrelated with every component of g_i. Then by the law of large numbers, $S_n^{-1}\sum_{i=1}^{n}U_i'\Omega^{-1}g_i/n \xrightarrow{p} 0$. Also note that $E[\psi_{ij}\psi_{ij}']/2 = S_n^{-1}E[U_i'\Omega^{-1}U_i]S_n^{-1\prime}$. It then follows similarly to the jackknife GMM that $nS_n^{-1}\partial\hat{Q}(\beta_0)/\partial\beta \xrightarrow{d} N(0, H+\Lambda)$, $\Lambda = \lim_{n\longrightarrow\infty} S_n^{-1}E[U_i'\Omega^{-1}U_i]S_n^{-1\prime}$. Then it follows that

$$S_n'(\hat{\beta}-\beta_0) \xrightarrow{d} N(0,V), V = H^{-1}+H^{-1}\Lambda H^{-1}.$$

We now show that GEL is asymptotically efficient relative to the jackknife GMM, i.e. that $V \leq V_J$ in the positive semidefinite sense, when the jackknife GMM has $W = \Omega^{-1}$. Let $\Delta_{ij} = \psi_{ij}^J - \psi_{ij}$. Under $W = \Omega^{-1}$ each element of Δ_{ij} depends on the data only through $(1, g_i')'(1, g_j')$. Therefore, by each element of U_i uncorrelated with every component of g_i, it follows that $E[\psi_{ij}\Delta_{ij}'] = 0$. Therefore we have

$$E[\psi_{ij}^J\psi_{ij}^{J\prime}] = E[(\psi_{ij}+\Delta_{ij})(\psi_{ij}+\Delta_{ij})'] = E[\psi_{ij}\psi_{ij}'] + E[\Delta_{ij}\Delta_{ij}'] \geq E[\psi_{ij}\psi_{ij}'],$$

so that

$$\Lambda = \lim_{n\longrightarrow\infty}\frac{1}{2}E[\psi_{ij}\psi_{ij}'] \leq \frac{1}{2}\lim_{n\longrightarrow\infty}E[\psi_{ij}^J\psi_{ij}^{J\prime}] = \Lambda_J.$$

Thus we have

$$V = H^{-1}+H^{-1}\Lambda H^{-1} \leq H^{-1}+H^{-1}\Lambda_J H^{-1} = V_J,$$

showing the asymptotic efficiency of GEL relative to a jackknife GMM estimator with $W = \Omega^{-1}$.

The linear model provides an example of the asymptotic variance, where from the earlier notation,

$$B^j = -\Omega^{-1}E[Z_iZ_i'\eta_{ij}\varepsilon_i], U_i^j = -Z_i\Upsilon_{ij}+E[Z_i\Upsilon_{ij}]+u_{ij},\ u_{ij} = -Z_i\eta_{ij}+B^{j\prime}Z_i\varepsilon_i.$$

Then for $u_i = [u_i^1, ..., u_i^p]$ we have, by $\Upsilon_i = S_n z_i/\sqrt{n}$

$$S_n^{-1} E[U_i'\Omega^{-1}U_i]S_n^{-1\prime} = S_n^{-1} E[u_i'\Omega^{-1}u_i]S_n^{-1\prime} + E[\{Z_i z_i' - E[Z_i z_i']\}'\Omega^{-1}\{Z_i z_i' - E[Z_i z_i']\}]/n.$$

The second term will be small as long as m grows slowly enough relative to n (when Z_{ij} is uniformly bounded $m/n \longrightarrow 0$ will suffice), so that

$$\Lambda = \lim_{n\longrightarrow\infty} S_n^{-1} E[u_i'\Omega^{-1}u_i]S_n^{-1\prime}.$$

For instance, in the homoskedastic case where $E[\varepsilon_i^2|Z_i] = \sigma_\varepsilon^2$, $E[\eta_i\eta_i'|Z_i] = \Sigma_\eta$, $E[\varepsilon_i\eta_i|Z_i] = \sigma_{\eta\varepsilon}$, we have $u_i = -Z_i(\eta_i' - \sigma_{\eta\varepsilon}'\varepsilon_i/\sigma_\varepsilon^2)$, so that

$$\begin{aligned} S_n^{-1} E[u_i'\Omega^{-1}u_i]S_n^{-1\prime} &= S_n^{-1} E[(\eta_i - \sigma_{\eta\varepsilon}\varepsilon_i/\sigma_\varepsilon^2)(\eta_i - \sigma_{\eta\varepsilon}\varepsilon_i/\sigma_\varepsilon^2)' Z_i'\Omega^{-1}Z_i]S_n^{-1\prime} \\ &= S_n^{-1}(\Sigma_\eta - \sigma_{\eta\varepsilon}\sigma_{\eta\varepsilon}'/\sigma_\varepsilon^2) E[Z_i'(\sigma_\varepsilon^2 I)^{-1}Z_i]S_n^{-1\prime} \\ &= mS_n^{-1}(\sigma_\varepsilon^2\Sigma_\eta - \sigma_{\eta\varepsilon}\sigma_{\eta\varepsilon}')S_n^{-1\prime}/\sigma_\varepsilon^4. \end{aligned}$$

Then, assuming $E[z_i Z_i']E[Z_i z_i'] \longrightarrow E[z_i z_i'] = \sigma_\varepsilon^2 H$ and $\sqrt{m}S_n^{-1} \longrightarrow S_0$, the asymptotic variance matrix for $S_n'(\hat{\beta} - \beta_0)$ will be

$$V = H^{-1} + H^{-1}S_0(\sigma_\varepsilon^2\Sigma_\eta - \sigma_{\eta\varepsilon}\sigma_{\eta\varepsilon}')S_0'H^{-1}/\sigma_\varepsilon^4.$$

This variance for GEL is identical to the asymptotic variance of LIML under many weak instrument asymptotics derived by Stock and Yogo (2005a). Thus we find that in the linear homoskedastic model GEL and LIML have the same asymptotic variance under many weak moment asymptotics. As shown by Hansen, Hausman and Newey (2008), the Bekker (1994) standard errors are consistent under many weak instruments, so that $S_n'\hat{V}S_n/n$ will have the same limit as the Bekker standard errors in a homoskedastic linear model. Since $S_n'\hat{V}S_n/n$ will also be consistent with heteroskedasticity, one can think of $\hat{V}$ as an extension of the Bekker (1994) variance estimator to GEL with heteroskedasticity.

It is interesting to compare the asymptotic variance V of the CUE with the usual asymptotic variance formula H^{-1} for GMM. When m grows slower than μ_n^2 or $\partial g_i(\beta_0)/\partial\beta$ is constant $V = H^{-1}$, but otherwise the variance here is larger than the standard formula. For further comparison we consider a corresponding variance approximation V_n for $\hat{\beta}$

for a sample size of size n. Replacing H by $H_n = nS_n^{-1}G'\Omega^{-1}GS_n^{-1\prime}$ and Λ by $\Lambda_n = S_n^{-1}E[U_i'\Omega^{-1}U_i]S_n^{-1\prime}$, and premultiplying by $(S_n')^{-1}$ and postmultiplying by S_n^{-1}gives the variance approximation for sample size n of

$$\begin{aligned} V_n &= S_n^{-1\prime}(H_n^{-1} + H_n^{-1}\Lambda_n H_n^{-1})S_n^{-1} \\ &= (nG'\Omega^{-1}G)^{-1} + (nG'\Omega^{-1}G)^{-1}E[U_i'\Omega^{-1}U_i](nG'\Omega^{-1}G)^{-1} \\ &= n^{-1}\{(G'\Omega^{-1}G)^{-1} + (G'\Omega^{-1}G)^{-1}(E[U_i'\Omega^{-1}U_i]/n)(G'\Omega^{-1}G)^{-1}\} \end{aligned}$$

The usual variance approximation for $\hat{\beta}$ is $(G'\Omega^{-1}G)^{-1}/n$. The approximate variance V_n includes an additional term which can be important in practice. When $Var(\Omega^{-1/2}\partial g_i(\beta_0)/\partial\beta)$ is large relative to $G'\Omega^{-1}G$ (condition 2 of Section 4), $E[U_i'\Omega^{-1}U_i]$ may be very large relative to $G'\Omega^{-1}G$, leading to the additional term being important, even when n is large.

It is interesting to note that the usual term is divided by n and the additional term by n^2. In asymptotic theory with fixed m this makes the additional term a "higher-order" variance term. Indeed, by inspection of Donald and Newey (2003), one can see that the additional term corresponds to a higher order variance term involving estimation of G. There are also additional higher order terms that come from the estimation of the weight matrix, but the Jacobian term dominates when identification is not strong. For example, in the linear homoskedastic example suppose that $E[\varepsilon_i^3|Z_i] = 0$ and $E[\varepsilon_i^4|Z_i] = E[\varepsilon_i^4]$, and let $\kappa = E[\varepsilon_i^4]/(E[\varepsilon_i^2])$. For $A_n = E[z_iZ_i']E[Z_iz_i']$ the higher-order variance approximation for GEL from Donald and Newey (2003) is

$$\begin{aligned} V_n &= \sigma_\varepsilon^2 A_n^{-1}/n + (m/n)\sigma_\varepsilon^2 A_n^{-1}(\Sigma_\eta - \sigma_{\eta\varepsilon}\sigma_{\eta\varepsilon}'/\sigma_\varepsilon^2)A_n^{-1}/n \\ &\quad + [(5-\kappa) + \rho_3(0)(3-\kappa)]\sigma_\varepsilon^2 A_n^{-1}E[Z_i'Z_i\Upsilon_i^2]A_n^{-1}/n^2. \end{aligned}$$

The last term corresponds to weight matrix estimation and will tend to be small when Υ_i is small, as it is under the asymptotics we consider. In this sense the many weak moment asymptotics accounts well for variability of the derivative of the moment conditions but takes no account of variability of the weight matrix. Also, it is interesting to note that this last term will be asymptotically small relative to the second even when m does not grow at the same rate as μ_n^2. For example, if z_i is bounded and $\mu_{jn} = \mu_n$ for each j, then

$\Upsilon_i^2 \leq C\mu_n/\sqrt{n}$, so as long as μ_n grows slower than $\sqrt{n}$ the third (weight matrix) term will be small relative to the second (Jacobian) term. Here the new variance estimator corresponds to the higher-order variance, showing that it provides an improved variance approximation more generally than in the knife-edge case where m and μ_n^2 grow at the same rate.

6 Large Sample Theory

We give results for GEL, leaving a precise treatment of jackknife GMM to another paper. It is helpful to strengthen Assumption 1. For a matrix A let $\|A\| = trace(A'A)^{1/2}$ denote its Euclidean norm and for symmetric A let $\xi_{\min}(A)$ and $\xi_{\max}(A)$ denote its smallest and largest eigenvalues, respectively. Also, let $\delta(\beta) = S_n'(\beta - \beta_0)/\mu_n$, where we suppress an n subscript on $\delta(\beta)$ for notational convenience.

ASSUMPTION 2: *i) Assumption 1 i) is satisfied; ii) There is $C > 0$ with $\|\delta(\beta)\| \leq C\sqrt{n}\,\|\bar{g}(\beta)\|/\mu_n$ for all $\beta \in B$ iii) there is $C > 0$ and $\hat{M} = O_p(1)$ such that with probability approaching one, $\|\delta(\beta)\| \leq C\sqrt{n}\,\|\hat{g}(\beta)\|/\mu_n + \hat{M}$ for all $\beta \in B$;*

This condition implies global identification of β_0. We also need conditions on convergence of $\hat{g}(\beta)$, as imposed in the following condition.

ASSUMPTION 3: *$g_i(\beta)$ is continuous in β and there is $C > 0$ such that i) $\sup_{\beta\in B} E[\{g_i(\beta)'g_i(\beta)\}^2]/n \longrightarrow 0$; ii) $1/C \leq \xi_{\min}(\Omega(\beta))$ and $\xi_{\max}(\Omega(\beta)) \leq C$ for all $\beta \in B$; iii) $\sup_{\beta\in B}\|\hat{\Omega}(\beta) - \Omega(\beta)\| \xrightarrow{p} 0$; iv) $|a'[\Omega(\tilde{\beta}) - \Omega(\beta)]b| \leq C\|a\|\,\|b\|\left\|\tilde{\beta} - \beta\right\|$ for all $a, b \in \Re^m, \tilde{\beta}, \beta \in B$; v) for every $\tilde{C} > 0$ there is C and $\hat{M} = O_p(1)$ such that for all $\tilde{\beta}, \beta \in B, \left\|\delta(\tilde{\beta})\right\| \leq \tilde{C}, \|\delta(\beta)\| \leq \tilde{C}$, $\sqrt{n}\left\|\bar{g}(\tilde{\beta}) - \bar{g}(\beta)\right\|/\mu_n \leq C\left\|\delta(\tilde{\beta} - \beta)\right\|$ and $\sqrt{n}\left\|\hat{g}(\tilde{\beta}) - \hat{g}(\beta)\right\|/\mu_n \leq \hat{M}\left\|\delta(\tilde{\beta} - \beta)\right\|$.*

These conditions restrict the rate at which m can grow with the sample size. If $E[g_{ij}(\beta)^4]$ is bounded uniformly in j, m, and β then a sufficient condition for $\|\hat{\Omega}(\beta) - \Omega(\beta)\| \xrightarrow{p} 0$ at each β is that $m^2/n \longrightarrow 0$. Uniform convergence may require further conditions.

For GEL estimators other than the CUE we need an additional condition.

ASSUMPTION 4: *$\hat{\beta}$ is the CUE or i) $\rho(v)$ is three times continuously differentiable and ii) there is $\gamma > 2$ such that $n^{1/\gamma}(E[\sup_{\beta \in B} \|g_i(\beta)\|^\gamma])^{1/\gamma}\sqrt{m/n} \longrightarrow 0$.*

When $\hat{\beta}$ is not the CUE this condition puts further restrictions on the growth rate of m. If the elements of $g_i(\beta)$ were bounded uniformly in n then this condition is $m^2/n^{1-2/\gamma} \longrightarrow 0$, that is only slightly stronger than $m^2/n \longrightarrow 0$. The following is a consistency result for CUE.

THEOREM 1: *If Assumptions 2 - 4 are satisfied then $S_n'(\hat{\beta} - \beta_0)/\mu_n \xrightarrow{p} 0$.*

We also give more primitive regularity conditions for consistency for the linear model example. Let $\tilde{\eta}_i$ be a vector of the nonzero elements of η_i and $\Sigma_i = E[(\varepsilon_i, \tilde{\eta}_i')'(\varepsilon_i, \tilde{\eta}_i')|Z_i, \Upsilon_i]$.

ASSUMPTION 5: *The linear model holds, $\Upsilon_i = S_n z_i/\sqrt{n}$, and there is a constant C with $E[\varepsilon_i^4|Z_i, \Upsilon_i] \leq C$, $E[\|\eta_i\|^4|Z_i, \Upsilon_i] \leq C$, $\|\Upsilon_i\| \leq C$, $\xi_{\min}(\Sigma_i) \geq 1/C$, $E[Z_i Z_i'] = I_m$, $E[(Z_i'Z_i)^2]/n \longrightarrow 0$, $E[\|z_i\|^4] < C$, and either $\hat{\beta}$ is the CUE or for $\gamma > 2$ we have $E[|\varepsilon_i|^\gamma|Z_i] \leq C$, $E[\|\eta_i\|^\gamma|Z_i] \leq C$, $n^{1/\gamma}(E[\|Z_i\|^\gamma])^{1/\gamma}\sqrt{m/n} \longrightarrow 0$.*

The conditions put restrictions on the rate at which m can grow with the sample size. If Z_{ij} is bounded uniformly in j and m, then these conditions will hold for the CUE if $m^2/n \longrightarrow 0$ (for in that case, $E[(Z_i'Z_i)^2]/n = O(m^2/n) \longrightarrow 0$) and if $m^2/n^{1-2/\gamma} \longrightarrow 0$ for other GEL estimators.

THEOREM 2: *If Assumptions 1 and 5 are satisfied then $S_n'(\hat{\beta} - \beta_0)/\mu_n \xrightarrow{p} 0$.*

For asymptotic normality some additional conditions are needed.

ASSUMPTION 6: *$g(z, \beta)$ is twice continuously differentiable in a neighborhood N of β_0, $(E[\|g_i\|^4] + E[\|G_i\|^4])\, m/n \longrightarrow 0$, and for a constant C and $j = 1, ..., p$,*

$$\xi_{\max}(E[G_i G_i']) \leq C, \xi_{\max}(E[\frac{\partial G_i(\beta_0)}{\partial \beta_j}\frac{\partial G_i(\beta_0)}{\partial \beta_j}']) \leq C, \sqrt{n}\left\|E[\frac{\partial G_i(\beta_0)}{\partial \beta_j}]S_n^{-1\prime}\right\| \leq C.$$

This condition imposes a stronger restriction on the growth rate of the number of moment conditions than was imposed for consistency. If $g_{ij}(\beta_0)$ were uniformly bounded a sufficient condition would be that $m^3/n \longrightarrow 0$.

ASSUMPTION 7: *If* $\bar{\beta} \xrightarrow{p} \beta_0$ *then* $\|\sqrt{n}[\hat{G}(\bar{\beta}) - \hat{G}(\beta_0)]S_n^{-1\prime}\| \xrightarrow{p} 0$, $\|\sqrt{n}[\partial\hat{G}(\bar{\beta})/\partial\beta_k - \partial\hat{G}(\beta_0)/\partial\beta_k]S_n^{-1\prime}\| \xrightarrow{p} 0, k = 1, ..., p.$

This condition restricts how the derivatives of the moments vary with the parameters. It is automatically satisfied in the linear model. For the next Assumption let

$$\begin{aligned}
\hat{\Omega}^k(\beta) &= \frac{1}{n}\sum_{i=1}^n g_i(\beta)\frac{\partial g_i(\beta)'}{\partial\beta_k}, \Omega^k(\beta) = E[\hat{\Omega}^k(\beta)], \\
\hat{\Omega}^{k\ell}(\beta) &= \frac{1}{n}\sum_{i=1}^n g_i(\beta)\frac{\partial^2 g_i(\beta)'}{\partial\beta_k\partial\beta_\ell}, \Omega^{k\ell}(\beta) = E[\hat{\Omega}^{k\ell}(\beta)], \\
\hat{\Omega}^{k,\ell}(\beta) &= \frac{1}{n}\sum_{i=1}^n \frac{\partial g_i(\beta)}{\partial\beta_k}\frac{\partial g_i(\beta)'}{\partial\beta_\ell}, \Omega^{k,\ell}(\beta) = E[\hat{\Omega}^{k,\ell}(\beta)].
\end{aligned}$$

ASSUMPTION 8: *For all* β *on a neighborhood* N *of* β_0 and A equal to $\Omega^k, \Omega^{k\ell}$,or $\Omega^{k,\ell}$; *i)* $\sup_{\beta\in N}\left\|\hat{A}(\beta) - A(\beta)\right\| \xrightarrow{p} 0$, *ii)* $|a'[A(\tilde{\beta}) - A(\beta)]b| \leq C\|a\|\|b\|\left\|\tilde{\beta} - \beta\right\|$.

This condition imposes uniform convergence and smoothness conditions similar to those already required for $\hat{\Omega}(\beta)$ and $\Omega(\beta)$ above.

ASSUMPTION 9: $\hat{\beta}$ *is the CUE or i) there is* $\gamma > 2$ *such that* $n^{1/\gamma}(E[\sup_{\beta\in B}\|g_i(\beta)\|^\gamma])^{1/\gamma}(m + \mu_n)/\sqrt{n} \longrightarrow 0$; *and ii)* $\mu_n^2 E[d_i^4]/n \longrightarrow 0$ *for*

$$d_i = \max_{\beta\in B}\max_j\{\|g_i(\beta)\|, \|\partial g_i(\beta)/\partial\beta\|, \left\|\partial^2 g_i(\beta)/\partial\beta\partial\beta_j\right\|\}.$$

This condition imposes some additional restrictions on the growth of m and μ_n. In the primary case of interest where μ_n^2 and m grow at the same rate then μ_n^2 can be replaced by m in this condition. If $\hat{\beta}$ is not the CUE, m^3 must grow slower than n. The next condition imposes corresponding requirements for the linear model case.

ASSUMPTION 10: *The linear model holds,* $mE[\|Z_i\|^4]/n \longrightarrow 0$, *and* $\hat{\beta}$ *is the CUE or* $n^{1/\gamma}(E[\|Z_i\|^\gamma])^{1/\gamma}(m+\mu_n)/\sqrt{n} \longrightarrow 0$ *and* $\mu_n^2 E[\|Z_i\|^4]/n \longrightarrow 0$.

Under these and other regularity conditions we can show that $\hat{\beta}$ is asymptotically normal and that the variance estimator is consistent. Recall the definition of U_i from Section 2.

THEOREM 3: *If Assumption 1 is satisfied,* $S_n^{-1}E[U_i'\Omega^{-1}U_i]S_n^{-1\prime} \longrightarrow \Lambda$, *and Assumptions 2 - 4 and 6-9 are satisfied or the linear model Assumptions 1, 5, and 10 are satisfied, then for* $V = H^{-1} + H^{-1}\Lambda H^{-1}$

$$S_n'(\hat{\beta} - \beta_0) \xrightarrow{d} N(0, V), S_n'\hat{V}S_n/n \xrightarrow{p} V.$$

Furthermore, if there is r_n *and* $c^* \neq 0$ *such that* $r_n S_n^{-1} c \longrightarrow c^*$ *then*

$$\frac{c'(\hat{\beta} - \beta_0)}{\sqrt{c'\hat{V}c/n}} \xrightarrow{d} N(0,1).$$

This result includes the linear model case. The next result shows that $\chi^2(m)$ asymptotic approximation for the Anderson-Rubin statistic is correct. Let q_α^m be the $1-\alpha^{th}$ quantile of a $\chi^2(m)$ distribution.

THEOREM 4: *If i)* $mE[\|g_i\|^4]/n \longrightarrow 0$ *and* $\xi_{\min}(\Omega) \geq C$ *or the linear model holds with* $E[\varepsilon_i^4|Z_i] \leq C$, $E[\varepsilon_i^2|Z_i] \geq C$, $E[Z_iZ_i'] = I$, *and* $mE[\|Z_i\|^4]/n \longrightarrow 0$; *ii)* $\hat{\beta}$ *is the CUE or there is* $\gamma > 2$ *such that* $n^{1/\gamma}E[\|g_i\|^\gamma]m/\sqrt{n} \longrightarrow 0$ *then*

$$\Pr(AR(\beta_0) \geq q_\alpha^m) \longrightarrow \alpha.$$

The last result shows that the Wald, LM, CLR, and overidentification statistics described in Section 3 are asymptotically equivalent and have asymptotically correct level under many weak moments, but that the likelihood ratio does not. Let $\hat{T} = n(\hat{\beta} - \beta_0)'\hat{V}^{-1}(\hat{\beta} - \beta_0)$.

THEOREM 5: *If* $S_n^{-1}E[U_i'\Omega^{-1}U_i]S_n^{-1\prime} \longrightarrow \Lambda$ *and either Assumptions 1 - 4 and 6-9 are satisfied or the linear model Assumptions 1, 5, and 10 are satisfied, then* $\hat{T} \xrightarrow{d} \chi^2(p)$,

$$LM(\beta_0) = \hat{T} + o_p(1),$$

$$\Pr(2n\hat{Q}(\hat{\beta}) \geq q_\alpha^{m-p}) \longrightarrow \alpha, \Pr(2n[\hat{Q}(\beta_0) - \hat{Q}(\hat{\beta})] \geq q_\alpha^p) \geq \alpha + o(1).$$

In addition, if there is $C > 0$ *such that* $\xi_{\min}(\mu_n^{-2}S_nHVHS_n') - m/\mu_n^2 > C$ *for all* n *large enough then there is a bounded sequence* $\{c_n\}$, $c_n \geq 0$, *that is bounded away from zero such that*

$$CLR(\beta_0) = c_n\hat{T} + o_p(1), \hat{q}_\alpha(\beta_0) = c_n q_\alpha^p + o_p(1).$$

and $\Pr(CLR(\beta_0) \geq \hat{q}_\alpha(\beta_0)) \longrightarrow \alpha$.

7 Monte Carlo Results

We first carry out a Monte Carlo for the simple linear IV model where the disturbances and instruments have a Gaussian distribution. The parameters of this experiment are the correlation coefficient ρ between the structural and reduced form errors, the concentration parameter and the number of instruments m.

The data generating process is given by

$$\begin{aligned} y_i &= x_i\beta_0 + \varepsilon_i \\ x_i &= z_i'\pi + \eta_i \\ \varepsilon_i &= \rho\eta_i + \sqrt{1-\rho^2}v_i \\ \eta_i &\sim N(0,1);\ v_i \sim N(0,1);\ z_i \sim N(0, I_m) \\ \pi &= \sqrt{\frac{CP}{mn}}\iota_m, \end{aligned}$$

where ι_m is an m-vector of ones. The concentration parameter in this design is equal to CP. We generate samples of size $n = 200$, with values of CP equal to 10, 20 or 35; number of instruments m equal to 3, 10 or 15; values of ρ equal to 0.3, 0.5 or 0.9; and $\beta_0 = 0$. This design covers cases with very weak instruments. For example, when $CP = 10$ and $m = 15$, the first stage F-statistic equals $CP/m = 0.67$.

Table 1 presents the estimation results for $10,000$ Monte Carlo replications. We report median bias and interquartile range (IQR) of 2SLS, GMM, LIML, CUE and the Jackknife GMM estimator, denoted JGMM. The CUE is obtained by a standard iterative minimization routine, taking the minimum obtained from five different starting values of β, $\{-2,-1,...,2\}$. The results for 2SLS and GMM are as expected. They are upward biased, with the bias increasing with the number of instruments, the degree of endogeneity and a decreasing concentration parameter. LIML and CUE are close to being median unbiased, although they display some small biases, accompanied by large interquartile ranges, when $CP = 10$ and the number of instruments is larger than 3. JGMM displays larger median biases than LIML and CUE in general, and especially in the very weak instrument case when $CP = 10$ and $m = 15$, with this bias increasing with ρ. There is a clear reduction in IQR for LIML, CUE and JGMM when both the number of instruments and the concentration parameter increase, whereas the biases for 2SLS and GMM remain. As expected, the IQR for JGMM is larger than the IQR for CUE, which in turn is larger than that of LIML. The superior performance of LIML might be expected here and in the Wald tests below, because it is a homoskedastic design and LIML imposes homoskedasticity on the weighting matrix. Doing so is often thought to improve small sample performance in homoskedastic cases.

Table 2 presents rejection frequencies of Wald tests at 5% nominal level. The purpose here is to analyze our proposed general methods in this well-understood setting. The estimators and standard errors utilized in the Wald tests are the two-step GMM estimator with the usual standard errors (GMM), with the Windmeijer (2005) standard errors (GMMC), the continuous updating estimator with the usual standard errors (CUE) and with the many weak instruments standard errors (CUEC), and equivalent for JGMM. For purposes of comparison we also give results for 2SLS and LIML with their usual standard errors and LIML with Bekker (1994) standard errors (LIMLC), and the GEL-LM statistic (LM) as defined in (3.5). We have also investigated the size properties of the GEL-AR and GEL-CLR statistics as defined in (3.4) and (3.6) respectively, and found them in these settings to be very similar to those of the LM statistic. They are therefore

not reported separately.

Table 1a. Simulation results for linear IV model $\rho = 0.3$

	$CP = 10$		$CP = 20$		$CP = 35$	
	Med Bias	IQR	Med Bias	IQR	Med Bias	IQR
$m = 3$						
2SLS	0.0509	0.3893	0.0312	0.2831	0.0184	0.2226
GMM	0.0516	0.3942	0.0307	0.2885	0.0186	0.2233
LIML	-0.0016	0.4893	0.0027	0.3184	0.0020	0.2398
CUE	-0.0031	0.4963	0.0034	0.3257	0.0013	0.2410
JGMM	-0.0127	0.5665	-0.0123	0.3474	-0.0064	0.2495
$m = 10$						
2SLS	0.1496	0.3059	0.0967	0.2486	0.0630	0.1996
GMM	0.1479	0.3153	0.0956	0.2562	0.0644	0.2056
LIML	0.0152	0.6060	0.0006	0.3846	-0.0001	0.2568
CUE	0.0230	0.6501	0.0002	0.4067	0.0007	0.2762
JGMM	0.0438	0.7242	-0.0117	0.4290	-0.0088	0.2785
$m = 15$						
2SLS	0.1814	0.2645	0.1237	0.2281	0.0839	0.1863
GMM	0.1809	0.2772	0.1248	0.2397	0.0846	0.1981
LIML	0.0262	0.6605	-0.0024	0.4102	-0.0047	0.2729
CUE	0.0375	0.7178	-0.0008	0.4629	-0.0034	0.3126
JGMM	0.0781	0.7855	-0.0065	0.4769	-0.0104	0.3128

Notes: $n = 200$; $\beta_0 = 0$; $10,000$ replications

Table 1b. Simulation results for linear IV model, $\rho = 0.5$

	$CP = 10$		$CP = 20$		$CP = 35$	
	Med Bias	IQR	Med Bias	IQR	Med Bias	IQR
$m = 3$						
2SLS	0.0957	0.3720	0.0498	0.2830	0.0300	0.2191
GMM	0.0961	0.3773	0.0501	0.2850	0.0296	0.2210
LIML	0.0053	0.4761	0.0028	0.3219	0.0018	0.2364
CUE	0.0082	0.4773	0.0031	0.3233	0.0009	0.2376
JGMM	-0.0189	0.5886	-0.0227	0.3576	-0.0130	0.2514
$m = 10$						
2SLS	0.2422	0.2768	0.1603	0.2302	0.1044	0.1910
GMM	0.2434	0.2900	0.1606	0.2360	0.1052	0.1969
LIML	0.0169	0.5640	0.0025	0.3641	-0.0016	0.2529
CUE	0.0212	0.6044	0.0045	0.3851	0.0035	0.2676
JGMM	0.0451	0.7086	-0.0137	0.4330	-0.0118	0.2875
$m = 15$						
2SLS	0.3000	0.2492	0.2108	0.2114	0.1432	0.1831
GMM	0.3021	0.2615	0.2115	0.2233	0.1437	0.1911
LIML	0.0320	0.6377	0.0026	0.3920	-0.0022	0.2718
CUE	0.0484	0.7039	0.0081	0.4408	0.0003	0.3027
JGMM	0.1051	0.7808	-0.0027	0.4890	-0.0122	0.3207

Notes: $n = 200$; $\beta_0 = 0$; $10,000$ replications

Table 1c. Simulation results for linear IV model, $\rho = 0.9$

	$CP = 10$		$CP = 20$		$CP = 35$	
	Med Bias	IQR	Med Bias	IQR	Med Bias	IQR
$m = 3$						
2SLS	0.1621	0.3254	0.0855	0.2601	0.0495	0.2077
GMM	0.1614	0.3313	0.0848	0.2650	0.0503	0.2106
LIML	-0.0053	0.4490	-0.0061	0.3054	-0.0046	0.2283
CUE	-0.0036	0.4559	-0.0038	0.3094	-0.0034	0.2291
JGMM	-0.0536	0.6532	-0.0441	0.3863	-0.0268	0.2613
$m = 10$						
2SLS	0.4348	0.1984	0.2842	0.1836	0.1870	0.1630
GMM	0.4363	0.2083	0.2853	0.1896	0.1856	0.1699
LIML	-0.0036	0.4823	-0.0057	0.3264	-0.0049	0.2391
CUE	-0.0034	0.5184	-0.0070	0.3477	-0.0059	0.2555
JGMM	0.0385	0.7737	-0.0347	0.4890	-0.0259	0.3155
$m = 15$						
2SLS	0.5333	0.1682	0.3747	0.1588	0.2608	0.1435
GMM	0.5333	0.1800	0.3748	0.1686	0.2609	0.1517
LIML	0.0018	0.5117	-0.0035	0.3331	0.0041	0.2391
CUE	0.0066	0.5778	-0.0013	0.3705	0.0042	0.2655
JGMM	0.1186	0.7972	-0.0232	0.5377	-0.0182	0.3378

Notes: $n = 200$; $\beta_0 = 0$; $10,000$ replications

Table 2. Rejection frequencies of Wald tests for linear IV model

	$\rho = 0.3$			$\rho = 0.5$		
	$CP = 10$	$CP = 20$	$CP = 35$	$CP = 10$	$CP = 20$	$CP = 35$
$m = 3$						
2SLS	0.0451	0.0440	0.0477	0.0780	0.0653	0.0593
GMM	0.0489	0.0492	0.0535	0.0835	0.0674	0.0621
GMMC	0.0468	0.0463	0.0510	0.0806	0.0644	0.0579
LIML	0.0384	0.0392	0.0428	0.0535	0.0470	0.0446
LIMLC	0.0317	0.0329	0.0374	0.0439	0.0415	0.0413
CUE	0.0744	0.0638	0.0621	0.0902	0.0638	0.0600
CUEC	0.0348	0.0382	0.0418	0.0500	0.0433	0.0429
JGMM	0.1080	0.0734	0.0676	0.1085	0.0724	0.0672
JGMMC	0.0217	0.0282	0.0370	0.0366	0.0378	0.0401
LM	0.0477	0.0444	0.0440	0.0428	0.0455	0.0446
$m = 10$						
2SLS	0.1148	0.0924	0.0793	0.2500	0.1833	0.1384
GMM	0.1423	0.1157	0.1001	0.2763	0.2089	0.1635
GMMC	0.1147	0.0910	0.0789	0.2291	0.1683	0.1305
LIML	0.0812	0.0663	0.0627	0.1015	0.0724	0.0587
LIMLC	0.0414	0.0367	0.0392	0.0585	0.0462	0.0423
CUE	0.3450	0.2277	0.1628	0.3080	0.2026	0.1470
CUEC	0.0587	0.0488	0.0450	0.0770	0.0532	0.0433
JGMM	0.3676	0.2513	0.1686	0.3657	0.2415	0.1629
JGMMC	0.0224	0.0327	0.0411	0.0472	0.0473	0.0458
LM	0.0398	0.0374	0.0363	0.0345	0.0356	0.0329
$m = 15$						
2SLS	0.1641	0.1339	0.1080	0.4081	0.3037	0.2283
GMM	0.2056	0.1749	0.1425	0.4547	0.3494	0.2704
GMMC	0.1534	0.1269	0.1008	0.3701	0.2720	0.2034
LIML	0.0995	0.0894	0.0786	0.1285	0.0935	0.0749
LIMLC	0.0393	0.0397	0.0413	0.0594	0.0510	0.0473
CUE	0.4721	0.3450	0.2535	0.4628	0.3234	0.2376
CUEC	0.0709	0.0637	0.0536	0.1001	0.0701	0.0509
JGMM	0.4668	0.3544	0.2397	0.4810	0.3487	0.2475
JGMMC	0.0244	0.0341	0.0420	0.0581	0.0571	0.0531
LM	0.0318	0.0317	0.0323	0.0342	0.0337	0.0299

Notes: $n = 200$; $H_0 : \beta_0 = 0$; $10,000$ replications, 5% nominal size

Table 2 continued. Rejection frequencies of Wald tests for linear IV model

	$\rho = 0.9$			
	$CP = 10$	$CP = 20$	$CP = 35$	
$m = 3$				
2SLS	0.1898	0.1274	0.0969	
GMM	0.1940	0.1312	0.1007	
GMMC	0.1818	0.1217	0.0933	
LIML	0.0799	0.0637	0.0556	
LIMLC	0.0767	0.0625	0.0551	
CUE	0.0967	0.0769	0.0675	
CUEC	0.0779	0.0648	0.0564	
JGMM	0.1265	0.0769	0.0676	
JGMMC	0.0708	0.0543	0.0482	
LM	0.0448	0.0451	0.0459	
$m = 10$				
2SLS	0.7315	0.5252	0.3572	
GMM	0.7446	0.5423	0.3847	
GMMC	0.7034	0.4850	0.3251	
LIML	0.0937	0.0739	0.0612	
LIMLC	0.0789	0.0663	0.0571	
CUE	0.2159	0.1462	0.1138	
CUEC	0.0833	0.0645	0.0527	
JGMM	0.3848	0.2520	0.1698	
JGMMC	0.1107	0.0747	0.0614	
LM	0.0334	0.0336	0.0345	
$m = 15$				
2SLS	0.9329	0.7935	0.6130	
GMM	0.9388	0.8092	0.6483	
GMMC	0.9165	0.7535	0.5663	
LIML	0.1062	0.0788	0.0662	
LIMLC	0.0827	0.0661	0.0596	
CUE	0.3350	0.2209	0.1712	
CUEC	0.0887	0.0665	0.0559	
JGMM	0.5054	0.3625	0.2545	
JGMMC	0.1497	0.0936	0.0744	
LM	0.0314	0.0271	0.0281	

Notes: $n = 200$; $H_0 : \beta_0 = 0$; $10,000$ replications, 5% nominal size

The LIML Wald test using the Bekker standard errors (LIMLC) has rejection frequencies very close to the nominal size, correcting the usual asymptotic Wald test which tends to be oversized with an increasing number of instruments. The LM-statistic shows a tendency to be undersized with an increasing number of instruments. The results for the rejection frequencies of the Wald test show that even with low numbers of instruments the corrected standard errors for the continuous updating estimator produce large improvements in the accuracy of the approximation. When the instruments are not too weak, i.e. when $CP = 20$ and larger, the observed rejection frequencies are very close to the nominal size for all values of m, whereas those based on the usual asymptotic standard errors are much larger than the nominal size. When we consider the "diagonal" elements, i.e. increasing the number of instruments and the concentration parameter at the same time, we see that the CUEC Wald test performs very well in terms of size. Similar improvements are found for the JGMMC Wald test, although this test overrejects more when $\rho = 0.9$ and $m = 15$.

We next analyze the properties of the CUE using the many weak instrument asymptotics for the estimation of the parameters in a panel data process, generated as in Windmeijer (2005):

$$\begin{aligned} y_{it} &= \beta_0 x_{it} + u_{it};\ u_{it} = \eta_i + v_{it};\ i = 1, ..., n;\ t = 1, ..., T; \\ x_{it} &= \gamma x_{it-1} + \eta_i + 0.5 v_{it-1} + \varepsilon_{it};\ \eta_i \sim N(0,1);\ \varepsilon_{it} \sim N(0,1); \\ v_{it} &= \delta_i \tau_t \omega_{it};\ \omega_{it} \sim \left(\chi^2_1 - 1\right);\ \delta_i \sim U[0.5, 1.5];\ \tau_t = 0.5 + 0.1(t-1). \end{aligned}$$

Fifty time periods are generated, with $\tau_t = 0.5$ for $t = -49, ..., 0$ and $x_{i,-49} \sim N\left(\frac{\eta_i}{1-\gamma}, \frac{1}{1-\gamma^2}\right)$, before the estimation sample is drawn. $n = 250$, $T = 6$, $\beta_0 = 1$ and $10,000$ replications are drawn. For this data generating process the regressor x_{it} is correlated with the unobserved constant heterogeneity term η_i and is predetermined due to its correlation with v_{it-1}. The idiosyncratic shocks v_{it} are heteroskedastic over time and at the individual level, and have a skewed chi-squared distribution. The model parameter β_0 is estimated by first-differenced GMM (see Arellano and Bond (1991)). As x_{it} is predetermined the

sequential moment conditions used are

$$g_i(\beta) = Z_i' \Delta u_i(\beta),$$

where

$$Z_i = \begin{bmatrix} x_{i1} & 0 & 0 & \cdots & 0 & \cdots & 0 \\ 0 & x_{i1} & x_{i2} & \cdots & 0 & \cdots & 0 \\ . & . & . & \cdots & . & \cdots & . \\ 0 & 0 & 0 & \cdots & x_{i1} & \cdots & x_{iT-1} \end{bmatrix},$$

$$\Delta u_i(\beta) = \begin{bmatrix} \Delta u_{i2}(\beta) \\ \Delta u_{i3}(\beta) \\ \vdots \\ \Delta u_{iT}(\beta) \end{bmatrix} = \begin{bmatrix} \Delta y_{i2} - \beta \Delta x_{i2} \\ \Delta y_{i3} - \beta \Delta x_{i3} \\ \vdots \\ \Delta y_{iT} - \beta \Delta x_{iT} \end{bmatrix}.$$

This results in a total of 15 moment conditions in this case, but only a maximum of 5 instruments for the cross section in the last time period.

The first two sets of results in Table 3 are the estimation results for values of $\gamma = 0.40$ and $\gamma = 0.85$ respectively. When $\gamma = 0.40$ the instruments are relatively strong, but they are weaker for $\gamma = 0.85$. The reported empirical concentration parameter is an object corresponding to the reduced form of this panel data model and is equal to 261 when $\gamma = 0.4$ and 35 when $\gamma = 0.85$. This is estimated simply from the linear reduced form estimated by OLS and ignores serial correlation and heteroskedasticity over time. This CP is therefore only indicative and does not play the same role as in the linear homoskedastic IV model. Median bias and interquartile range (IQR) are reported for the standard linear one-step and two-step GMM estimators, the CUE and JGMM. When $\gamma = 0.40$, median biases are negligible for GMM, CUE and JGMM, with comparable interquartile ranges. When $\gamma = 0.85$ and the instruments are weaker, the linear GMM estimators are downward biased, whereas the CUE and JGMM are median unbiased but exhibit a larger interquartile range than the linear GMM estimators.

Table 3. Simulation results for panel data model, $N = 250$, $T = 6$

	$\gamma = 0.40$ $(CP = 261)$		$\gamma = 0.85$ $(CP = 35)$		$\gamma = 0.85$ $(CP = 54)$	
	Med Bias	IQR	Med Bias	IQR	Med Bias	IQR
GMM1	-0.0082	0.0797	-0.0644	0.2077	-0.0836	0.1743
GMM2	-0.0047	0.0712	-0.0492	0.1952	-0.0608	0.1627
CUE	0.0002	0.0734	0.0010	0.2615	-0.0068	0.2218
JGMM	0.0003	0.0737	0.0018	0.2707	-0.0038	0.2280
Instr:	$x_{it-1}, ..., x_{i1}$		$x_{it-1}, ..., x_{i1}$		$x_{it-1}, ..., x_{i1}; y_{it-2}, ..y_{i1}$	

Figure 1 presents p-value plots for the Wald tests for the hypothesis $H_0 : \beta_0 = 1$ when $\gamma = 0.85$, based on one-step GMM estimates (W_{GMM1}), on two-step GMM estimates (W_{GMM2}), on the Windmeijer (2005) corrected two-step Wald (W_{GMM2C}), on the CUE using the conventional asymptotic variance (W_{CUE}), on the CUE using the variance estimate $\hat{V}$ described in Section 2 (W_{CUEC}), and equivalently on the JGMM (W_J and W_{JC}). Further displayed is the p-value plot for the LM statistic (LM). It is clear that the usual asymptotic variance estimates for the CUE and JGMM are too small. This problem is similar to that of the linear two-step GMM estimator, leading to rejection frequencies that are much larger than the nominal size. In contrast, use of the variance estimators under many weak instrument asymptotics leads to rejection frequencies that are very close to the nominal size.

The third set of results presented in Table 3 is for the design with $\gamma = 0.85$, but with lags of the dependent variable y_{it} included as sequential instruments $(y_{i,t-2}, ..., y_{i1})$, additional to the sequential lags of x_{it}. As there is feedback from y_{it-1} to x_{it} and x_{it} is correlated with η_i the lagged values of y_{it} could improve the strength of the instrument set. The total number of instruments increases to 25, with a maximum of 11 for the cross section in the final period. The empirical concentration parameter increases from 35 to 54. The GMM estimators are more downward biased when the extra instruments are included. The CUE and JGMM are still median unbiased and their IQRs have decreased by 15%. As the p-value plot in Figure 2 shows, use of the proposed variance estimators result in rejection frequencies that are virtually equal to the nominal size. Although W_{GMM2C} had good size properties when using the smaller instrument set, use of the additional instruments leads to rejection frequencies that are larger than the nominal

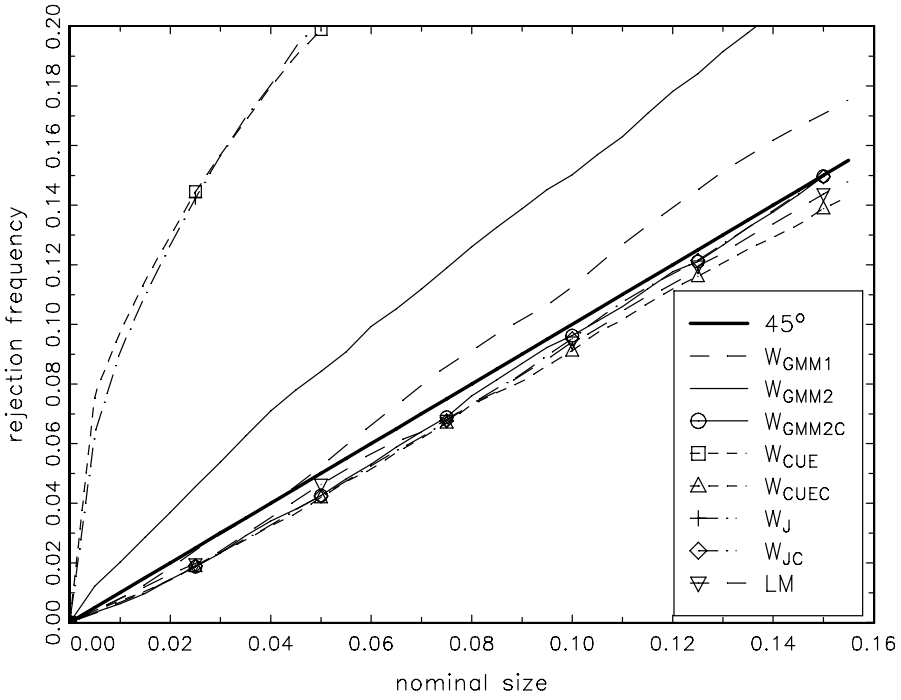

Fig. 1. P-value plot, $\gamma = 0.85$, $H_0 : \beta_0 = 1$, Panel data model

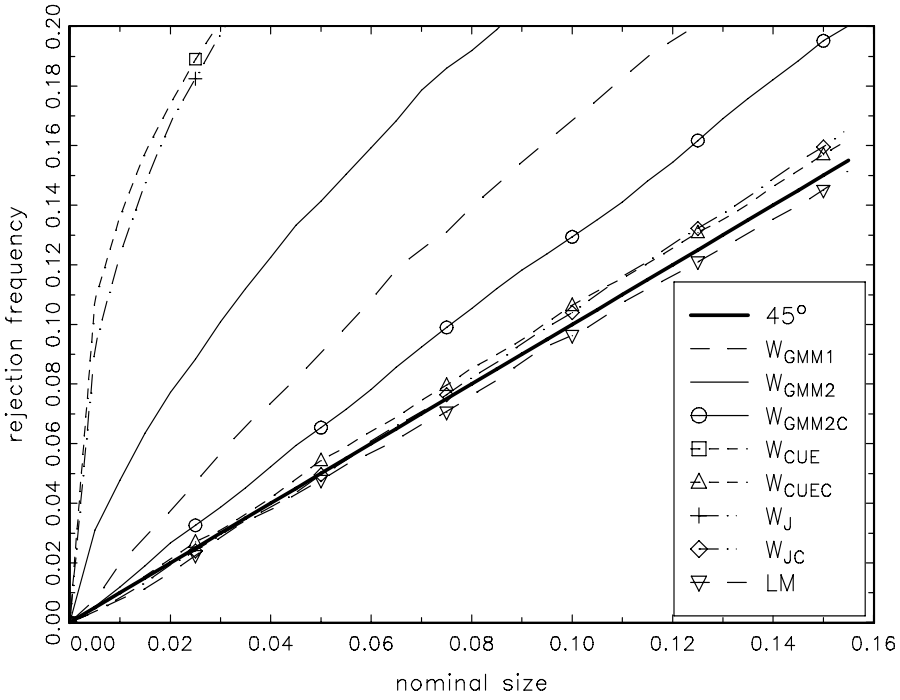

Fig. 2. P-value plot, $\gamma = 0.85$, $H_0 : \beta_0 = 1$, Panel data model, additional instruments.

size.

When further investigating the size properties of the AR and CLR tests, we find that the behaviour of the CLR test is virtually indistinguishable from that of the LM test, whereas the AR test tends to have rejection frequencies that are slightly smaller than the nominal size, especially with the larger instrument set. For the power of these tests in the latter example, we find that the CLR and LM tests have identical power, which is slightly less than that of W_{CUEC}, with the AR test having much lower power.

8 Conclusion

We have given a new variance estimator for GEL that is consistent under standard asymptotics and also accounts for many weak moment conditions. This approximation is shown to perform well in a simple linear IV and panel data Monte Carlo.

One possible topic for future research is higher order asymptotics when m grows so slowly that the standard asymptotic variance formula is correct. As discussed in the paper, we conjecture that the new variance estimator would provide an improved approximation in a range of such cases. Hansen, Hausman and Newey (2008) have shown such a result for the Bekker (1994) variance in the homoskedastic linear model.

Another interesting topic is the choice of moment conditions under many weak moment conditions. Donald, Imbens, and Newey (2003) give a criteria for moment choice for GMM and GEL that is quite complicated. Under many weak moment conditions this criteria should simplify. It would be useful in practice to have a simple criteria for choosing the moment conditions.

A third topic for future research is the extension of these results to dependent observations. It appears that the variance estimator for the CUE would be the same except that $\hat{\Omega}$ would include autocorrelation terms. It should also be possible to obtain similar results for GEL estimators based on time smoothed moment conditions, like those considered in Kitamura and Stutzer (1997).

References

Andrews, D.W.K. and J.H. Stock (2006): "Inference with Weak Instruments," in Blundell, R., W. Newey, T. Persson eds., *Advances in Economics and Econometrics, Vol. 3.*

Altonji, J. and L.M. Segal (1996): "Small Sample Bias in GMM Estimation of Covariance Structures," *Journal of Economic and Business Statistics* 14, 353-366.

Angrist, J. and A. Krueger (1991): "Does Compulsory School Attendance Affect Schooling and Earnings", *Quarterly Journal of Economics 106*, 979–1014.

Angrist, J. and G. Imbens, and A. Krueger (1999): "Jackknife Instrumental Variables Estimation," *Journal of Applied Econometrics* 14, 57-67.

Arellano, M. and S.R. Bond (1991): "Some Tests of Specification for Panel Data: Monte Carlo Evidence and an Application to Employment Equations," *Review of Economic Studies* 58, 277-297.

Bekker, P.A. (1994): "Alternative Approximations to the Distributions of Instrumental Variables Estimators," *Econometrica* 63, 657-681.

Blomquist, S. and M. Dahlberg (1999): "Small Sample Properties of LIML and Jackknife IV Estimators: Experiments with Weak Instruments," *Journal of Applied Econometrics* 14, 69-88.

Brown, B.W. and W.K. Newey (1998): "Efficient Semiparametric Estimation of Expectations," *Econometrica* 66, 453-464.

Chao, J.C. and N.R. Swanson (2004): "Estimation and Testing Using Jackknife IV in Heteroskedastic Regression with Many Weak Instruments," working paper, Rutgers University.

Chao, J.C. and N.R. Swanson (2005): "Consistent Estimation With a Large Number of Weak Instruments," *Econometrica* 73, 1673-692.

Donald, S.G., G.W. Imbens, and W.K. Newey (2003): "Empirical Likelihood Estimation and Consistent Tests With Conditional Moment Restrictions," *Journal of Econometrics* 117, 55-93.

Donald, S. G. and W. K. Newey (2003) "Choosing the Number of Moments in GMM and GEL Estimation," working paper, MIT.

Dufour, J-M. (1997) "Some Impossibility Theorems in Econometrics With Applications to Structural and Dynamic Models," *Econometrica* 65, 1365-1387.

Guggenberger, P. and R.J. Smith (2005): "Generalized Empirical Likelihood Estimators and Tests Under Partial, Weak, and Strong Identification," *Econometric Theory* 21, 667-709.

Hahn, J. and J.A. Hausman (2004) "A New Specification Test for the Validity of Instrumental Variables," *Econometrica* 70, 163-189.

Han, C. and P.C.B. Phillips (2006): "GMM with Many Moment Conditions," *Econometrica* 74, 147-192.

Hansen, C., J.A. Hausman and W.K. Newey (2008): "Estimation with Many Instrumental Variables," *Journal of Business and Economic Statistics,* forthcoming.

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

Hansen, L.P. and K.J. Singleton. (1982): "Generalized Instrumental Variables Estimation of Nonlinear Rational Expectations Models," *Econometrica* 50, 1269-1286.

Hansen, L.P., J. Heaton and A. Yaron (1996): "Finite-Sample Properties of Some Alternative GMM Estimators," *Journal of Business and Economic Statistics* 14, 262-280.

Holtz-Eakin, D., W.K. Newey and H. Rosen (1988), "Estimating Vector Autoregressions With Panel Data," *Econometrica* 56, 1371-1396.

Imbens, G. (1997): "One-step Estimators for Over-identified Generalized Method of Moments Models," *Review of Economic Studies* 64, 359-383.

Kitamura, Y., and M. Stutzer (1997): "An Information-Theoretic Alternative to Generalized Method of Moments Estimation", *Econometrica* 65, 861-874.

Kleibergen, F. (2005): "Testing Parameters in GMM Without Assuming They are Identified," *Econometrica* 73, 1103-1123.

Moreira, M. (2003) "A Conditional Likelihood Ratio Test for Structural Models," *Econometrica* 71, 1027-1048.

Newey, W.K. (1990): "Efficient Instrumental Variables Estimation of Nonlinear Models," *Econometrica* 58, 809-837.

Newey, W.K. (1991): "Uniform Convergence in Probability and Stochastic Equicontinuity," *Econometrica* 59, 1161-1167.

Newey, W.K., and R.J. Smith (2004): "Higher-order Properties of GMM and Generalized Empirical Likelihood Estimators," *Econometrica* 72, 219-255.

Phillips, G.D.A. and C. Hale (1977): "The Bias of Instrumental Variable Estimators of Simultaneous Equation Systems," *International Economic Review* 18, 219-228.

Qin, J. and Lawless, J. (1994): "Empirical Likelihood and General Estimating Equations," *Annals of Statistics* 22, 300-325.

Smith, R. J. (1997): "Alternative Semi-Parametric Likelihood Approaches to Generalized Method of Moments Estimation," *Economic Journal* 107, 503-519.

Smith, R. J. (2006): "Weak Instruments and Empirical Likelihood: A Discussion of the Papers by D.W.K. Andrews and J.H. Stock and Y. Kitamura," in Blundell, R., W. Newey, T. Persson eds., , *Advances in Economics and Econometrics, Vol. 3.*

Stock, J. and J. Wright (2000): "GMM With Weak Identification," *Econometrica* 68, 1055-1096.

Stock, J. and M. Yogo (2005a): "Asymptotic Distributions of Instrumental Variables Statistics with Many Instruments," in D.W.K. Andrews and J.H. Stock, eds., *Identification and Inference for Econometric Models: Essays in Honor of Thomas Rothenberg*, 109-120.

Stock, J. and M. Yogo (2005b): "Testing for Weak Instruments in Linear IV Regression," in D.W.K. Andrews and J.H. Stock, eds., *Identification and Inference for Econometric Models: Essays in Honor of Thomas Rothenberg*, 80-108.

Windmeijer, F. (2005): "A Finite Sample Correction for the Variance of Linear Efficient Two-Step GMM Estimators," *Journal of Econometrics* 126, 25-51.

Supplement to "GMM with Many Weak Moment Conditions": Appendix

Whitney K. Newey Frank Windmeijer

September 2004
Revised, February 2008

1 Appendix: Proofs.

Throughout the Appendices, let C denote a generic positive constant that may be different in different uses. Let CS, M, and T denote the Cauchy-Schwartz, Markov, and triangle inequalities respectively. Let S denote the Slutzky Lemma and CMT the Continuous Mapping Theorem. Also, let CM denote the conditional Markov inequality that if $E[|A_n||B_n] = O_p(\varepsilon_n)$ then $A_n = O_p(\varepsilon_n)$ and let w.p.a.1 stand for "with probability approaching one." The following standard matrix result is used repeatedly.

LEMMA A0: *If A and B are symmetric, positive semidefinite matrices then*

$$|\xi_{\min}(A) - \xi_{\min}(B)| \leq \|A - B\|, |\xi_{\max}(A) - \xi_{\max}(B)| \leq \|A - B\|.$$

Also, if $\left\|\hat{A} - A\right\| \xrightarrow{p} 0$, $\xi_{\min}(A) \geq 1/C$, and $\xi_{\max}(A) \leq C$, then w.p.a.1 $\xi_{\min}(\hat{A}) \geq 1/2C, \xi_{\max}(\hat{A}) \leq 2C$.

1.1 Consistency Proofs for General CUE

For Lemmas A1 and A10, let $Y_i, Z_i, (i = 1, ..., n)$ be i.i.d. $m \times 1$ random vectors with 4th moments, that can depend on n, but where we suppress an n subscript for notational convenience. Also, let

$$\bar{Y} = \sum_{i=1}^{n} Y_i/n, \mu_Y = E[Y_i], \Sigma_{YY} = E[Y_i Y_i'], \Sigma_{YZ} = E[Y_i Z_i']$$

and let objects with Z in place of Y be defined in the corresponding way.

LEMMA A1: *If* $(Y_i, Z_i), (i = 1, ..., n)$, are i.i.d., $\xi_{\max}(AA') \leq C$, $\xi_{\max}(A'A) \leq C$, $\xi_{\max}(\Sigma_{YY}) \leq C$, $\xi_{\max}(\Sigma_{ZZ}) \leq C, m/a_n^2 \longrightarrow 0$, $a_n/n \leq C$, $E[(Y_i'Y_i)^2]/na_n^2 \longrightarrow 0$, $E[(Z_i'Z_i)^2]/na_n^2 \longrightarrow 0$, $n\mu_Y'\mu_Y/a_n^2 \longrightarrow 0$, $n\mu_Z'\mu_Z/a_n^2 \longrightarrow 0$, *then*

$$n\bar{Y}'A\bar{Z}/a_n = \operatorname{tr}\left(A\Sigma_{YZ}'\right)/a_n + n\mu_Y'A\mu_Z/a_n + o_p(1).$$

Proof: Let $W_i = AZ_i$. Then $A\Sigma_{YZ}' = \Sigma_{YW}'$, $A\mu_Z = \mu_W$,

$$\begin{aligned} \xi_{\max}(E[W_iW_i']) &= \xi_{\max}(A\Sigma_{ZZ}A') \leq C\xi_{\max}(AA') \leq C, \\ E[(W_i'W_i)^2]/na_n^2 &= E[(Z_i'A'AZ_i)^2]/na_n^2 \leq CE[(Z_i'Z_i)^2]/na_n^2 \longrightarrow 0. \end{aligned}$$

Thus the hypotheses and conclusion are satisfied with W in place of Z and $A = I$. Therefore, it suffices to show the result with $A = I$.

Note that

$$\begin{aligned} E[(Y_i'Z_i)^2] &\leq E[(Y_i'Y_i)^2] + E[(Z_i'Z_i)^2], \\ E[Y_i'Z_jZ_j'Y_i] &= E[Y_i'\Sigma_{ZZ}Y_i] \leq CE[Y_i'Y_i] = Ctr(\Sigma_{YY}) \leq Cm, \\ |E[Y_i'Z_jY_j'Z_i]| &\leq C(E[Y_i'Z_jZ_j'Y_i] + E[Y_j'Z_iZ_i'Y_j]) \leq Cm. \end{aligned}$$

For the moment suppose $\mu_Y = \mu_Z = 0$. Let $W_n = n\bar{Y}'\bar{Z}/a_n$. Then $E[W_n] = E[Y_i'Z_i]/a_n = \operatorname{tr}\left(\Sigma_{YZ}\right)/a_n$ and

$$E[W_n]^2/n \leq E\left[(Y_i'Z_i)^2\right]/na_n^2 \leq \{E[(Y_i'Y_i)^2] + E[(Z_i'Z_i)^2]\}/na_n^2 \longrightarrow 0.$$

We also have

$$\begin{aligned} E[W_n^2] &= E\left[\sum_{i,j,k,\ell} Y_i'Z_jY_k'Z_\ell/n^2a_n^2\right] = E\left[(Y_i'Z_i)^2\right]/na_n^2 + (1 - 1/n)\{E[W_n]^2 \\ &\quad + E[Y_i'Z_jY_j'Z_i]/a_n^2 + E[Y_i'Z_jZ_j'Y_i]/a_n^2\} = E[W_n]^2 + o(1), \end{aligned}$$

so that by M,

$$W_n = \operatorname{tr}\left(\Sigma_{YZ}'\right)/a_n + o_p(1).$$

In general, when μ_Y or μ_Z are nonzero, note that $E[\{(Y_i-\mu_Y)'(Y_i-\mu_Y)\}^2] \leq CE[(Y_i'Y_i)^2]$ and $\xi_{\max}(Var(Y_i)) \leq \xi_{\max}(\Sigma_{YY})$, so the hypotheses are satisfied with $Y_i - \mu_Y$ replacing Y_i and $Z_i - \mu_Z$ replacing Y_i and Z_i respectively. Also,

$$\begin{aligned} W_n &= n\bar{Y}'\bar{Z}/a_n = n\left(\bar{Y}-\mu_Y\right)'(\bar{Z}-\mu_Z)/a_n + n\mu_Y'(\bar{Z}-\mu_Z)/a_n \\ &\quad + n(\bar{Y}-\mu_Y)'\mu_Z/a_n + n\mu_Y'\mu_Z/a_n. \end{aligned} \tag{1.1}$$

Note that

$$E\left[\left\{n\mu_Y'(\bar{Z}-\mu_Z)/a_n\right\}^2\right] = n\mu_Y'(\Sigma_{ZZ}-\mu_Z\mu_Z')\mu_Y/a_n^2 \leq n\mu_Y'\Sigma_{ZZ}\mu_Y/a_n^2 \leq Cn\mu_Y'\mu_Y/a_n^2 \longrightarrow 0.$$

so by M, the second and third terms in eq. (1.1) (with Y and Z interchanged) are $o_p(1)$. Also, $tr(\mu_Z\mu_Y')/a_n = a_n n^{-1}(n\mu_Y'\mu_Z/a_n^2) \longrightarrow 0$. Applying the result for the zero mean case then gives

$$W_n = tr(\Sigma_{YZ}'-\mu_Z\mu_Y')/a_n + n\mu_Y'\mu_Z/a_n + o_p(1) = tr(\Sigma_{YZ}')/m + n\mu_Y'\mu_Z/m + o_p(1).\ Q.E.D.$$

It is useful to work with a reparameterization

$$\delta = S_n'(\beta-\beta_0)/\mu_n.$$

For notational simplicity we simply change the argument to denote the reparameterized functions, e.g. $\hat{Q}(\delta)$ will denote $\hat{Q}(\beta_0 + \mu_n S_n^{-1\prime}\delta)$. Let $\hat{Q}^*(\delta) = \hat{g}(\delta)'\hat{\Omega}(\delta)^{-1}\hat{g}(\delta)/2$ be the objective function for quadratic $\rho(v)$, $\tilde{Q}(\delta) = \hat{g}(\delta)'\Omega(\delta)^{-1}\hat{g}(\delta)/2$, and $Q(\delta) = \bar{g}(\delta)'\Omega(\delta)^{-1}\bar{g}(\delta)/2 + m/2n$.

LEMMA A2: *If Assumption 3 is satisfied then for any $C > 0$, $\sup_{\beta\in B,\|\delta\|\leq C} \mu_n^{-2} n|\hat{Q}^*(\delta) - Q(\delta)| \xrightarrow{p} 0$.*

Proof: Note that by Assumption 3 ii), $\mu_n^{-2}nE[\|\hat{g}(0)\|^2] = \mu_n^{-2}tr(\Omega(\beta_0)) \leq C$, so by Assumption 3 v) and T,

$$\sup_{\|\delta\|\leq C}\|\hat{g}(\delta)\| \leq \|\hat{g}(0)\| + \sup_{\|\delta\|\leq C}\|\hat{g}(\delta)-\hat{g}(0)\| = O_p(\mu_n/\sqrt{n}).$$

Let $\hat{a}(\delta) = \mu_n^{-1}\sqrt{n}\Omega(\delta)^{-1}\hat{g}(\delta)$. By Assumption 3 ii)

$$\|\hat{a}(\delta)\|^2 = \mu_n^{-2} n \hat{g}(\delta)'\Omega(\delta)^{-\frac{1}{2}}\Omega(\delta)^{-1}\Omega(\delta)^{-\frac{1}{2}}\hat{g}(\delta) \leq C\mu_n^{-2} n \|\hat{g}(\delta)\|^2,$$

so that $\sup_{\|\delta\|\leq C}\|\hat{a}(\delta)\| = O_p(1)$. Also, by Assumption 3 iii) we have

$$\left|\xi_{\min}(\hat{\Omega}(\delta)) - \xi_{\min}(\Omega(\delta))\right| \leq \sup_{\|\delta\|\leq C}\left\|\hat{\Omega}(\delta) - \Omega(\delta)\right\| \xrightarrow{p} 0,$$

so that $\xi_{\min}(\hat{\Omega}(\delta)) \geq C$, and hence $\xi_{\max}(\hat{\Omega}(\delta)^{-1}) \leq C$ for all $\|\delta\| \leq C$, w.p.a.1. Therefore,

$$\begin{aligned}
\mu_n^{-2} n \left|\hat{Q}^*(\delta) - \tilde{Q}(\delta)\right| &\leq \left|\hat{a}(\delta)'\left[\hat{\Omega}(\delta) - \Omega(\delta)\right]\hat{a}(\delta)\right| \\
&\quad + \left|\hat{a}(\delta)'\left[\hat{\Omega}(\delta) - \Omega(\delta)\right]\hat{\Omega}(\delta)^{-1}\left[\hat{\Omega}(\delta) - \Omega(\delta)\right]\hat{a}(\delta)\right| \\
&\leq \|\hat{a}(\delta)\|^2\left(\left\|\hat{\Omega}(\delta) - \Omega(\delta)\right\| + C\left\|\hat{\Omega}(\delta) - \Omega(\delta)\right\|^2\right) \xrightarrow{p} 0.
\end{aligned}$$

Next, let $a(\tilde{\delta}, \delta) = \mu_n^{-1}\sqrt{n}\Omega(\delta)^{-1}\bar{g}(\tilde{\delta})$ and $Q(\tilde{\delta}, \delta) = \bar{g}(\tilde{\delta})'\Omega(\delta)^{-1}\bar{g}(\tilde{\delta})/2 + m/2n$. By Assumption 3, $\sup_{\|\delta\|\leq C, \|\tilde{\delta}\|\leq C}\left\|a(\delta, \tilde{\delta})\right\| \leq C$. Then by Assumption 3 iv), for $\|\delta\| \leq C, \left\|\tilde{\delta}\right\| \leq C$, it follows by $\mu_n S_n^{-1}$ bounded,

$$\begin{aligned}
\mu_n^{-2} n \left|Q(\tilde{\delta}, \tilde{\delta}) - Q(\tilde{\delta}, \delta)\right| &= \left|a(\tilde{\delta}, \tilde{\delta})'\left[\Omega(\tilde{\delta}) - \Omega(\delta)\right]a(\tilde{\delta}, \delta)\right| \\
&\leq C\left\|\mu_n S_n^{-1\prime}(\tilde{\delta} - \delta)\right\| \leq C\left\|\tilde{\delta} - \delta\right\|.
\end{aligned}$$

Also, by T and Assumption 3, for $\|\delta\| \leq C, \left\|\tilde{\delta}\right\| \leq C$,

$$\mu_n^{-2} n \left|Q(\tilde{\delta}, \delta) - Q(\delta, \delta)\right| \leq C\mu_n^{-2} n(\left\|\bar{g}(\tilde{\delta}) - \bar{g}(\delta)\right\|^2 + \|\bar{g}(\delta)\|\left\|\bar{g}(\tilde{\delta}) - \bar{g}(\delta)\right\|) \leq C\left\|\tilde{\delta} - \delta\right\|.$$

Then by T it follows that $\mu_n^{-2} n \left|Q(\tilde{\delta}) - Q(\delta)\right| = \mu_n^{-2} n \left|Q(\tilde{\delta}, \tilde{\delta}) - Q(\delta, \delta)\right| \leq C\left\|\tilde{\delta} - \delta\right\|$. Therefore, $\mu_n^{-2} n Q(\delta)$ is equicontinuous on $\left\|\tilde{\delta}\right\| \leq C, \|\delta\| \leq C$. An analogous argument with $\hat{a}(\tilde{\delta}, \delta) = \mu_n^{-1}\sqrt{n}\Omega(\delta)^{-1}\hat{g}(\tilde{\delta})$ and $\tilde{Q}(\tilde{\delta}, \delta) = \hat{g}(\tilde{\delta})'\Omega(\delta)^{-1}\hat{g}(\tilde{\delta})$ replacing $a(\tilde{\delta}, \delta)$ and $Q(\tilde{\delta}, \delta)$ respectively implies that $\mu_n^{-2} n \left|\tilde{Q}(\tilde{\delta}) - \tilde{Q}(\delta)\right| = \mu_n^{-2} n \left|\tilde{Q}(\tilde{\delta}, \tilde{\delta}) - \tilde{Q}(\delta, \delta)\right| \leq \hat{M}\left\|\tilde{\delta} - \delta\right\|$ on $\left\|\tilde{\delta}\right\| \leq C, \|\delta\| \leq C$, with $\hat{M} = O_p(1)$, giving stochastic equicontinuity of $\mu_n^{-2} n \tilde{Q}(\delta)$.

Since $\mu_n^{-2} n \tilde{Q}(\delta)$ and $\mu_n^{-2} n Q(\delta)$ are stochastically equicontinuous, it suffices by Newey (1991, Theorem 2.1) to show that $\mu_n^{-2} n \tilde{Q}(\delta) = \mu_n^{-2} n Q(\delta) + o_p(1)$ for each δ. Apply Lemma A1 with $Y_i = Z_i = g_i(\delta)$, $A = \Omega(\delta)^{-1}$, and $a_n = \mu_n^2$. By Assumption 3, $\xi_{\max}(A'A) =$

$\xi_{\max}(AA') = \xi_{\max}(\Omega(\delta)^{-2}) \leq C$, $\xi_{\max}(\Sigma_{YY}) = \xi_{\max}(\Omega(\delta)) \leq C$, $E[(Y_i'Y_i)^2]/na_n^2 = E[\{g_i(\delta)'g_i(\delta)\}^2]/n\mu_n^4 \longrightarrow 0$, and $n\mu_Y'\mu_Y/a_n^2 \leq Cn\bar{g}(\delta)'\Omega(\delta)^{-1}\bar{g}(\delta)/\mu_n^4 = C\left(nQ(\delta)/\mu_n^2 - m/\mu_n^2\right)/\mu_n^2 \longrightarrow 0$ where the last follows by equicontinuity of $\mu_n^{-2}nQ(\delta)$. Thus, the hypotheses of Lemma A1 are satisfied. Note that $A\Sigma_{YZ}' = A\Sigma_{ZZ} = A\Sigma_{YY} = mI_m/\mu_n^2$, so by the conclusion of Lemma A1

$$\mu_n^{-2}n\tilde{Q}(\delta) = tr(I_m)/\mu_n^2 + \mu_n^{-2}n\bar{g}(\delta)'\Omega(\delta)^{-1}\bar{g}(\delta) + o_p(1) = \mu_n^{-2}nQ(\delta) + o_p(1).$$

Q.E.D.

Let $\hat{P}(\beta,\lambda) = \sum_{i=1}^n \rho(\lambda' g_i(\beta))/n$.

LEMMA A3: *If Assumptions 3 and 4 are satisfied then w.p.a.1* $\hat{\beta} = \arg\min_{\beta\in B}\hat{Q}(\beta)$, $\hat{\lambda} = \operatorname{argmax}_{\lambda\in\hat{L}_n(\hat{\beta})}\hat{P}(\hat{\beta},\lambda)$, *and* $\tilde{\lambda} = \operatorname{argmax}_{\lambda\in\hat{L}(\beta_0)}\hat{P}(\beta_0,\lambda)$ *exist,* $\left\|\tilde{\lambda}\right\| = O_p(\sqrt{m/n})$, $\left\|\hat{\lambda}\right\| = O_p(\sqrt{m/n})$, $\left\|\hat{g}(\hat{\beta})\right\| = O_p(\sqrt{m/n})$, *and* $\hat{Q}^*(\hat{\beta}) \leq \hat{Q}^*(\beta_0) + o_p(m/n)$.

Proof: Let $b_i = \sup_{\beta\in B}\|g_i(\beta)\|$. A standard result gives $\max_{i\leq n} b_i = O_p(n^{1/\gamma}(E[b_i^\gamma])^{1/\gamma})$. Also, by Assumption 4 there exists τ_n such that $\sqrt{m/n} = o(\tau_n)$ and $\tau_n = o(n^{-1/\gamma}(E[b_i^\gamma])^{-1/\gamma})$. Let $L_n = \{\lambda : \|\lambda\| \leq \tau_n\}$. Note that

$$\sup_{\lambda\in L_n,\beta\in B,i\leq n}|\lambda' g_i(\beta)| \leq \tau_n \max_{i\leq n} b_i = O_p(\tau_n n^{1/\gamma}(E[b_i^\gamma])^{1/\gamma}) \longrightarrow 0.$$

Then there is C such that w.p.a.1, for all $\beta\in B$, $\lambda\in L_n$, and $i\leq n$, we have

$$L_n \subset \hat{L}(\beta), -C \leq \rho_2(\lambda' g_i(\beta)) \leq -C^{-1},\ |\rho_3(\lambda' g_i(\beta))| \leq C.$$

By a Taylor expansion around $\lambda = 0$ with Lagrange remainder, for all $\lambda\in L_n$

$$\hat{P}(\beta,\lambda) = -\lambda'\hat{g}(\beta) + \lambda'\left[\sum_{i=1}^n \rho_2(\bar{\lambda}' g_i(\beta))g_i(\beta)g_i(\beta)'/n\right]\lambda,$$

where $\bar{\lambda}$ lies on the line joining λ and 0. Then by Lemma A0, w.p.a.1 for all $\beta\in B$ and $\lambda\in L_n$,

$$-\lambda'\hat{g}(\beta) - C\|\lambda\|^2 \leq \hat{P}(\beta,\lambda) \leq -\lambda'\hat{g}(\beta) - C^{-1}\|\lambda\|^2 \leq \|\lambda\|\,\|\hat{g}(\beta)\| - C^{-1}\|\lambda\|^2. \tag{1.2}$$

Let $\tilde{g} = \hat{g}(\beta_0)$ and $\tilde{\lambda} = \text{argmax}_{\lambda \in L_n} \hat{P}(\beta_0, \lambda)$. By $\xi_{\max}(\Omega(\beta_0)) \leq C$ it follows that $E\left[\|\tilde{g}\|^2\right] = \text{tr}(\Omega)/n \leq C\, m/n$, so by M, $\|\tilde{g}\| = O_p\left(\sqrt{m/n}\right)$. By the right hand inequality in eq. (1.2),

$$0 = \hat{P}(\beta_0, 0) \leq \hat{P}(\beta_0, \tilde{\lambda}) \leq \left\|\tilde{\lambda}\right\| \|\tilde{g}\| - C^{-1} \left\|\tilde{\lambda}\right\|^2$$

Subtracting $C^{-1}\left\|\tilde{\lambda}\right\|^2$ from both sides and dividing through by $C^{-1}\left\|\tilde{\lambda}\right\|$ gives

$$\left\|\tilde{\lambda}\right\| \leq C \|\tilde{g}\| = O_p\left(\sqrt{m/n}\right).$$

Since $\sqrt{m/n} = o(\tau_n)$ it follows that w.p.a.1, $\tilde{\lambda} \in \text{int}(L_n)$, and is therefore a local maximum of $\hat{P}(\beta_0, \lambda)$ in $\hat{L}(\beta)$. By concavity of $P(\beta_0, \lambda)$ in λ, a local maximum is a global maximum, i.e.

$$\hat{P}(\beta_0, \tilde{\lambda}) = \max_{\lambda \in \hat{L}(\beta_0)} \hat{P}(\beta_0, \lambda) = \hat{Q}(\beta_0).$$

Summarizing, w.p.a.1 $\tilde{\lambda} = \arg\max_{\lambda \in \hat{L}(\beta_0)} \hat{P}(\beta_0, \lambda)$ exists and $\left\|\tilde{\lambda}\right\| = O_p\left(\sqrt{m/n}\right)$. Also, plugging $\tilde{\lambda}$ back in the previous inequality gives

$$\hat{Q}(\beta_0) = O_p(m/n).$$

Next, let $\hat{Q}_{\tau_n}(\beta) = \max_{\lambda \in L_n} \hat{P}(\beta, \lambda)$. By continuity of $g_i(\beta)$ and $\rho(v)$ and by the theorem of the maximum $\hat{Q}_{\tau_n}(\beta)$ is continuous on B, so $\hat{\beta}_{\tau_n} = \arg\min_{\beta \in B} \hat{Q}_{\tau_n}(\beta)$ exists by compactness of B. Let $\hat{g}_{\tau_n} = \hat{g}(\hat{\beta}_{\tau_n})$. By the left-hand inequality in eq. (1.2), for all $\lambda \in L_n$

$$-\lambda' \hat{g}_{\tau_n} - C\|\lambda\|^2 \leq \hat{P}(\hat{\beta}_{\tau_n}, \lambda) \leq \hat{Q}_{\tau_n}(\hat{\beta}_{\tau_n}) \leq \hat{Q}_{\tau_n}(\beta_0) \leq \hat{Q}(\beta_0) = O_p(m/n). \tag{1.3}$$

Consider $\lambda = -(\hat{g}_{\tau_n}/\|\hat{g}_{\tau_n}\|)\tau_n$. Plugging this in eq. (1.3) gives

$$\tau_n \|\hat{g}_{\tau_n}\| - c\tau_n^2 = O_p(m/n).$$

Note that for n large enough, $m/n \leq C\tau_n^2$, so that dividing by τ_n^2

$$\|\hat{g}_{\tau_n}\| \leq O_p\left(\tau_n^{-1} m/n\right) + C\tau_n = O_p(\tau_n)$$

Consider any $\alpha_n \longrightarrow 0$ and let $\breve{\lambda} = -\alpha_n \hat{g}_{\tau_n}$. Then $\left\|\breve{\lambda}\right\| = o_p(\tau_n)$ so that $\breve{\lambda} \in L_n$ w.p.a.1 Substituting this $\breve{\lambda}$ in the above inequality gives

$$\alpha_n \left\|\hat{g}_{\tau_n}\right\|^2 - C\alpha_n^2 \left\|\hat{g}_{\tau_n}\right\|^2 = \alpha_n(1 - C\alpha_n) \left\|\hat{g}_{\tau_n}\right\|^2 = O_p\left(\frac{m}{n}\right).$$

Note that $1 - C\alpha_n \to 1$, so that this inequality implies that $\alpha_n \left\|\hat{g}_{\tau_n}\right\|^2 = O_p\left(m/n\right)$. Since α_n goes to zero as slowly as desired, it follows that

$$\left\|\hat{g}(\hat{\beta}_{\tau_n})\right\| = \left\|\hat{g}_{\tau_n}\right\| = O_p\left(\sqrt{m/n}\right).$$

Let $\hat{\lambda} = \arg\max_{\lambda \in L_n} \hat{P}(\hat{\beta}_{\tau_n}, \lambda)$. It follows exactly as for $\tilde{\lambda}$, with $\hat{\beta}_{\tau_n}$ replacing β_0, that $\left\|\hat{\lambda}\right\| = O_p(\sqrt{m/n})$ and w.p.a.1, $\hat{\lambda} = \arg\max_{\lambda \in \hat{L}(\beta)} \hat{P}(\hat{\beta}_{\tau_n}, \lambda)$, so that

$$\hat{Q}_{\tau_n}(\hat{\beta}_{\tau_n}) = \hat{P}(\hat{\beta}_{\tau_n}, \hat{\lambda}) = \max_{\lambda \in \hat{L}(\beta)} \hat{P}(\hat{\beta}_{\tau_n}, \lambda) = \hat{Q}(\hat{\beta}_{\tau_n}).$$

Then w.p.a.1, by the definition of $\hat{Q}_{\tau_n}(\beta)$ and $\hat{\beta}_{\tau_n}$, for all $\beta \in B$,

$$\hat{Q}(\hat{\beta}_{\tau_n}) = \hat{Q}_{\tau_n}(\hat{\beta}_{\tau_n}) \leq \hat{Q}_{\tau_n}(\beta) = \max_{\lambda \in L_n} \hat{P}(\beta, \lambda) \leq \hat{Q}(\beta).$$

Thus, w.p.a.1 we can take $\hat{\beta} = \hat{\beta}_{\tau_n}$.

Now expand around $\lambda = 0$ to obtain, for $\hat{g}_i = g_i(\hat{\beta})$ and $\hat{\Omega} = \hat{\Omega}(\hat{\beta})$, w.p.a.1,

$$\hat{Q}(\hat{\beta}) = \hat{P}(\hat{\beta}, \hat{\lambda}) = -\hat{g}'\hat{\lambda} - \frac{1}{2}\hat{\lambda}'\hat{\Omega}\hat{\lambda} + \hat{r}, \hat{r} = \frac{1}{6}\sum \rho_3(\bar{\lambda}'\hat{g}_i)(\hat{\lambda}'\hat{g}_i)^3/n,$$

where $\left\|\bar{\lambda}\right\| \leq \left\|\hat{\lambda}\right\|$ and $\hat{r} = 0$ for the CUE (where $\rho(v)$ is quadratic). When $\hat{\beta}$ is not the CUE, w.p.a.1

$$|\hat{r}| \leq \left\|\hat{\lambda}\right\| \max_i b_i C \hat{\lambda}'\hat{\Omega}(\hat{\beta})\hat{\lambda} \leq O_p(\sqrt{m/n} n^{1/\gamma} \left(E[b_i^\gamma]\right)^{1/\gamma}) C \left\|\bar{\lambda}\right\|^2 = o_p(m/n).$$

Also, $\hat{\lambda}$ satisfies the first order conditions $\sum_{i=1}^n \rho_1(\hat{\lambda}'\hat{g}_i)\hat{g}_i/n = 0$. By an expansion $\rho_1(\hat{\lambda}'\hat{g}_i) = -1 - \hat{\lambda}'\hat{g}_i + \rho_3(\bar{v}_i)(\lambda'\hat{g}_i)^2/2$ where $\bar{v}_i$ lies in between 0 and $\lambda'\hat{g}_i$ and either $\rho_3(\bar{v}_i) = 0$ for the CUE or $\max_{i \leq n} |\bar{v}_i| \leq \max_{i \leq n} |\hat{\lambda}'\hat{g}_i| \leq \tau_n \longrightarrow 0$. Expanding around $\lambda = 0$ gives

$$0 = -\hat{g} - \hat{\Omega}\hat{\lambda} + \hat{R}, \hat{R} = \frac{1}{2}\sum_{i=1}^n \rho_3(\bar{v}_i)(\hat{\lambda}'\hat{g}_i)^2 \hat{g}_i/n = 0.$$

Then either $\hat{R} = 0$ for the CUE or we have

$$\left\|\hat{R}\right\| \leq C \max_i b_i |\rho_3(\bar{v}_i)| \hat{\lambda}' \hat{\Omega} \hat{\lambda} = O_p(n^{1/\gamma}(E[b_i^\gamma])^{1/\gamma} m/n) = o_p(\sqrt{m/n}).$$

solving for $\hat{\lambda} = \hat{\Omega}^{-1}(-\hat{g} + \hat{R})$ and plugging into the expansion for $\hat{Q}(\hat{\beta})$ gives

$$\begin{aligned} \hat{Q}(\hat{\beta}) &= -\hat{g}'\hat{\Omega}^{-1}(-\hat{g} + \hat{R}) - \frac{1}{2}(-\hat{g} + \hat{R})'\hat{\Omega}^{-1}(-\hat{g} + \hat{R}) + o_p(m/n) \\ &= \hat{Q}^*(\hat{\beta}) - \hat{R}'\hat{\Omega}^{-1}\hat{R}/2 + o_p(m/n) = \hat{Q}^*(\hat{\beta}) + o_p(m/n). \end{aligned}$$

An exactly analogous expansion, replacing $\hat{\beta}$ with β_0, gives

$$\hat{Q}(\hat{\beta}) = \hat{Q}^*(\beta_0) + o_p(m/n).$$

Then by the definition of $\hat{\beta}$,

$$\hat{Q}^*(\hat{\beta}) = \hat{Q}(\hat{\beta}) + o_p(m/n) \leq \hat{Q}(\beta_0) + o_p(m/n) = \hat{Q}^*(\beta_0) + o_p(m/n). Q.E.D.$$

LEMMA A4: *If Assumptions 2 - 4 are satisfied then* $\left\|\hat{\delta}\right\| = O_p(1)$.

Proof: By Lemma A3, w.p.a.1 $\left\|\hat{g}(\hat{\beta})\right\| = O_p(\sqrt{m/n})$, so that Assumption 2 iii) and $m/\mu_n^2 \leq C$,

$$\left\|\hat{\delta}\right\| \leq C\mu_n^{-1}\sqrt{n}\left\|\hat{g}(\hat{\beta})\right\| + O_p(1) = O_p(\sqrt{m}/\mu_n) + O_p(1) = O_p(1). Q.E.D.$$

Proof of Theorem 1: By Lemma A3 and $m/\mu_n^2 \leq C$ it follows that, parameterizing in terms of $\delta = S_n'(\beta - \beta_0)/\mu_n$ (where $\delta_0 = 0$),

$$\mu_n^{-2} n \hat{Q}^*(\hat{\delta}) \leq \mu_n^{-2} n \hat{Q}^*(0) + o_p(1).$$

Consider any $\varepsilon, \gamma > 0$. By Lemma A4 there is C such that $\Pr(\mathcal{A}_1) \geq 1 - \varepsilon/3$ for $\mathcal{A}_1 = \{\left\|\hat{\delta}\right\| \leq C\}$. In the notation of Lemma A2 let $\mathcal{A}_2 = \{\sup_{\|\delta\| \leq C} \mu_n^{-2} n \left|\hat{Q}^*(\delta) - Q(\delta)\right| < \gamma/3\}$ and $\mathcal{A}_3 = \{\mu_n^{-2} n \hat{Q}^*(\hat{\delta}) \leq \mu_n^{-2} n \hat{Q}^*(0) + \gamma/3\}$ By Lemma A2, for all n large enough $\Pr(\mathcal{A}_2) \geq 1 - \varepsilon/3$ and by Lemma A3 $\Pr(\mathcal{A}_3) \geq 1 - \varepsilon/3$. Then $\Pr(\mathcal{A}_1 \cap \mathcal{A}_2 \cap \mathcal{A}_3) \geq 1 - \varepsilon$, and on $\mathcal{A}_1 \cap \mathcal{A}_2 \cap \mathcal{A}_3$,

$$\mu_n^{-2} n Q(\hat{\delta}) \leq \mu_n^{-2} n \hat{Q}^*(\hat{\delta}) + \gamma/3 \leq \mu_n^{-2} n \hat{Q}^*(0) + 2\gamma/3 \leq \mu_n^{-2} n Q(0) + \gamma = m/\mu_n^2 + \gamma,$$

where the second inequality follows by $\hat{\delta} \in \mathcal{A}_3$. Subtracting m/μ_n^2 from both sides it follows that $\mathcal{A}$ implies $\mu_n^{-2} n \bar{g}(\hat{\delta})' \Omega(\hat{\delta})^{-1} \bar{g}(\hat{\delta}) \leq \gamma$. Since ε, γ can be any positive constants, we have $\mu_n^{-2} n \bar{g}(\hat{\delta})' \Omega(\hat{\delta})^{-1} \bar{g}(\hat{\delta}) \xrightarrow{p} 0$. Then, by Assumption 2 ii) and 3 ii),

$$\mu_n^{-2} n \bar{g}(\hat{\delta})' \Omega(\hat{\delta})^{-1} \bar{g}(\hat{\delta}) \geq C \mu_n^{-2} n \bar{g}(\bar{\beta})' \bar{g}(\bar{\beta}) \geq C \left\| \hat{\delta} \right\|^2,$$

so that $\left\| \hat{\delta} \right\| \xrightarrow{p} 0$. Q.E.D.

1.2 Conditions for the Linear Model

LEMMA A5: *If Assumption 5 is satisfied then* $\xi_{\min}(E[(y_i - x_i'\beta)^2 | Z_i, \Upsilon_i]) \geq C$. *Also, for* $X_i = (y_i, x_i')'$, $E[\|X_i\|^4 | Z_i, \Upsilon_i] \leq C$.

Proof: Let $\Delta = \beta_0 - \beta$ and $\tilde{\Delta}$ the elements of Δ corresponding to the vector $\tilde{\eta}_i$ of nonzero elements of η_i from Assumption 5. Then $y_i - x_i'\beta = \varepsilon_i + \tilde{\eta}_i' \tilde{\Delta} + \Upsilon_i' \Delta$, so that

$$E[(y_i - x_i'\beta)^2 | Z_i, \Upsilon_i] \geq E[(\varepsilon_i + \tilde{\eta}_i' \tilde{\Delta})^2 | Z_i, \Upsilon_i] = (1, \tilde{\Delta}') \Sigma_i (1, \tilde{\Delta}')' \geq \xi_{\min}(\Sigma_i)(1 + \tilde{\Delta}'\tilde{\Delta}) \geq C,$$

giving the first conclusion. Also, $E[\|x_i\|^4 | Z_i, \Upsilon_i] \leq C E[\|\eta_i\|^4 | Z_i, \Upsilon_i] + C E[\|\Upsilon_i\|^4 | Z_i, \Upsilon_i] \leq C$ and $E[y_i^4 | Z_i, \Upsilon_i] \leq C E[\|x_i\|^4 \|\beta_0\|^4 | Z_i, \Upsilon_i] + E[\varepsilon_i^4 | Z_i, \Upsilon_i] \leq C$, giving the second conclusion. Q.E.D.

LEMMA A6: *If Assumption 5 is satisfied then there is a constant* C *such that for every* $\beta \in B$ *and* m, $C^{-1} I_m \leq \Omega(\beta) \leq C I_m$.

Proof: By Lemma A4 $C^{-1} \leq E[(y_i - x_i'\beta)^2 | Z_i] \leq C$, so that the conclusion follows by $I_m = E[Z_i Z_i']$ and $\Omega(\beta) = E[Z_i Z_i' E[(y_i - x_i'\beta)^2 | Z_i]]$. Q.E.D.

LEMMA A7: *If Assumption 5 is satisfied then Assumption 3 v) is satisfied,* $\|n^{-1} \sum_i Z_i z_i' - E[Z_i z_i']\| \xrightarrow{p} 0$, *and* $\|n^{-1} \sum_i Z_i \eta_i'\| = O_p(\sqrt{m/n})$.

Proof: For the last conclusion, by $E[\eta_i'\eta_i|Z_i] \leq C$ we have

$$E[\left\| n^{-1}\sum_i Z_i\eta_i' \right\|^2] = n^{-1}E[Z_i'Z_i\eta_i'\eta_i] \leq Cn^{-1}E[Z_i'Z_i] = Cm/n,$$

so the last conclusion follows by M. For the second to last conclusion, we have

$$E[\left\| n^{-1}\sum_i Z_i z_i' - E[Z_i z_i'] \right\|^2] \leq E[Z_i'Z_i z_i'z_i]/n \leq \sqrt{E[\|Z_i\|^4]/n}\sqrt{E[\|z_i\|^4]/n} \longrightarrow 0,$$

so it also follows by M.

Next, by Assumption 5 Lemma A6 we have

$$\|E[Z_i z_i']\|^2 = tr\left\{E[z_i Z_i'](E[Z_i Z_i'])^{-1}E[Z_i z_i']\right\} \leq tr(E[z_i z_i']) \leq C.$$

Then we have by CS, $\Upsilon_i = S_n z_i/\sqrt{n}$, $G = -E[Z_i z_i']S_n'/\sqrt{n}$

$$\begin{aligned} \mu_n^{-1}\sqrt{n}\left\|\bar{g}(\tilde{\beta}) - \bar{g}(\beta)\right\| &= \mu_n^{-1}\sqrt{n}\left\|G(\tilde{\beta}-\beta)\right\| = \left\|E[Z_i z_i']\left(\tilde{\delta}-\delta\right)\right\| \\ &\leq \|E[Z_i z_i']\|\left\|\tilde{\delta}-\delta\right\| \leq C\left\|\tilde{\delta}-\delta\right\|. \end{aligned}$$

Also, by $\hat{G} = \hat{G}(\beta)$ not depending on β, by $\|S_n^{-1\prime}\| \leq C/\mu_n$, and by T,

$$\begin{aligned} \left\|\hat{G}\sqrt{n}S_n^{-1\prime}\right\| &\leq \left\|\frac{1}{\sqrt{n}}\sum_i Z_i\eta_i'S_n^{-1\prime}\right\| + \left\|\frac{1}{n}\sum_i Z_i z_i' - E[Z_i z_i']\right\| + \|E[Z_i z_i']\| \\ &= O_p(\frac{\sqrt{n}}{\mu_n}\sqrt{\frac{m}{n}}) + o_p(1) + O(1) = O_p(1), \end{aligned}$$

so that for $\hat{M} = \left\|\hat{G}\sqrt{n}S_n^{-1\prime}\right\| = O_p(1)$, by CS,

$$\mu_n^{-1}\sqrt{n}\left\|\hat{g}(\tilde{\beta}) - \hat{g}(\beta)\right\| = \mu_n^{-1}\sqrt{n}\left\|\hat{G}(\tilde{\beta}-\beta)\right\| = \left\|\hat{G}\sqrt{n}S_n^{-1\prime}(\tilde{\delta}-\delta)\right\| \leq \hat{M}\left\|\tilde{\delta}-\delta\right\|.Q.E.D.$$

LEMMA A8: *If Assumption 5 is satisfied then Assumption 3 iii) and Assumption 8 i) are satisfied.*

Proof: Let $X_i = (y_i, x_i')'$ and $\alpha = (1, -\beta')'$, so that $y_i - x_i'\beta = X_i'\alpha$. Note that

$$\hat{\Omega}(\beta) - \Omega(\beta) = \sum_{k,\ell=1}^{p+1}\hat{F}_{k\ell}\alpha_k\alpha_\ell, \hat{F}_{k\ell} = \sum_{i=1}^{n} Z_i Z_i' X_{ik}X_{i\ell}/n - E[Z_i Z_i' X_{ik}X_{i\ell}].$$

Then $E[X_{ik}^2 X_{i\ell}^2|Z_i] \leq C$ by Lemma A4 so that

$$E[\left\|\hat{F}_{k\ell}\right\|^2] \leq CE[(Z_i'Z_i)^2 E[X_{ik}^2 X_{i\ell}^2|Z_i]]/n \leq CE[(Z_i'Z_i)^2]/n \longrightarrow 0.$$

Then $\sup_{\beta\in B}\left\|\hat{\Omega}(\beta) - \Omega(\beta)\right\| \xrightarrow{p} 0$ follows by B bounded. The other parts of Assumption 8 i) follow similarly upon noting that

$$\hat{\Omega}^k(\beta) - \Omega^k(\beta) = \sum_{\ell=1}^{p+1} \hat{F}_{k\ell}\alpha_\ell, \hat{\Omega}^{k,\ell}(\beta) - \Omega^{k,\ell}(\beta) = \hat{F}_{k\ell}, \hat{\Omega}^{k\ell}(\beta) = \Omega^{k\ell}(\beta) = 0. Q.E.D.$$

LEMMA A9: *If Assumption 5 is satisfied, then Assumption 3 iv) Assumption 8 ii) are satisfied.*

Proof: Let $\tilde{\Sigma}_i = E[X_i X_i'|Z_i]$, which is bounded by Lemma A5. Then by $\alpha = (1, -\beta)$ bounded on B we have $|\tilde{\alpha}'\tilde{\Sigma}_i\tilde{\alpha} - \alpha'\tilde{\Sigma}_i\alpha| \leq C\left\|\tilde{\beta} - \beta\right\|$. Also, $E[(a'Z_i)^2] = a'E[Z_iZ_i']a = \|a\|^2$. Therefore,

$$\begin{aligned} & |a'\Omega(\tilde{\beta})b - a'\Omega(\beta)b| = |E[(a'Z_i)(b'Z_i)E[(X_i'\tilde{\alpha})^2 - (X_i'\alpha)^2|Z_i]]| \\ \leq\ & E[|a'Z_i|\,|b'Z_i|\,|\tilde{\alpha}'\tilde{\Sigma}_i\tilde{\alpha} - \alpha'\tilde{\Sigma}_i\alpha|] \leq CE[(a'Z_i)^2]^{1/2}E[(b'Z_i)^2]^{1/2}\|\tilde{\beta} - \beta\| \leq C\|a\|\|b\|\|\tilde{\beta} - \beta\|. \end{aligned}$$

We also have

$$\begin{aligned} |a'\Omega^k(\tilde{\beta})b - a'\Omega^k(\beta)b| &= |2E[(a'Z_i)(b'Z_i)E[x_{ik}X_i'(\tilde{\alpha} - \alpha)|Z_i]]| \\ &\leq CE[|a'Z_i|\,|b'Z_i|\,E[|x_{ij}|\|X_i\||Z_i]]\|\tilde{\beta} - \beta\| \\ &\leq C\|a\|\|b\|\|\tilde{\beta} - \beta\|. \end{aligned}$$

The other parts of Assumption 8 ii) follow by $\Omega^{k,\ell}(\beta)$ and $\Omega^{k\ell}(\beta)$ not depending on β. Q.E.D.

Proof of Theorem 2: The result will follow by Theorem 1 upon showing that Assumptions 2 and 3 are true. We now verify Assumption 2. Assumption 2 i) holds by hypothesis. For Assumption 2 ii), note that by $G = -E[Z_iz_i']S_n'/\sqrt{n}$,

$$\mu_n^{-1}\sqrt{n}\bar{g}(\beta) = \sqrt{n}G(\beta - \beta_0)/\mu_n = -\sqrt{n}GS_n^{-1\prime}\delta.$$

Then by $nS_n^{-1}G'GS_n^{-1\prime} \geq CnS_n^{-1}G'\Omega^{-1}GS_n^{-1\prime}$ and Assumption 1 we have

$$\mu_n^{-1}\sqrt{n}\left\|\bar{g}(\beta)\right\| = \left(\delta'\left[nS_n^{-1}G'GS_n^{-1\prime}\right]\delta\right)^{1/2} \geq C\left\|\delta\right\|.$$

Next, let $\hat{R} = \sum_i \left(Z_i z_i' - E[Z_i z_i']\right)/n$, and note that

$$\hat{g}(\beta) = \hat{g}(\beta_0) - \frac{1}{n}\sum_i Z_i x_i'(\beta - \beta_0) = \hat{g}(\beta_0) - \frac{1}{n}\sum_i Z_i\eta_i'(\beta - \beta_0) + \mu_n n^{-1/2}(-\hat{R} + E[Z_i z_i'])\delta.$$

By Lemma A7, $\left\|\hat{R}\right\| \xrightarrow{p} 0$, so that by T and CS, w.p.a.1,

$$\left\|(-\hat{R} + E[Z_i z_i'])\delta\right\| \geq \left\|E[Z_i z_i']\delta\right\| - \left\|\hat{R}\delta\right\| \geq (C - \left\|\hat{R}\right\|)\left\|\delta\right\| \geq C\left\|\delta\right\|.$$

Also, as previously discussed, $\mu_n^{-1}\sqrt{n}\left\|\hat{g}(\beta_0)\right\| = O_p(1)$ and by Lemma A7 $\mu_n^{-1}\sqrt{n}\left\|\sum_i Z_i\eta_i'/n\right\| = O_p(1)$, so that by B compact

$$\hat{M} = \mu_n^{-1}\sqrt{n}\sup_{\beta\in B}\left\|\hat{g}(\beta_0) - \frac{1}{n}\sum_i Z_i\eta_i'(\beta - \beta_0)\right\| = O_p(1).$$

Then by T it follows that w.p.a.1 for all $\beta \in B$,

$$\left\|\delta\right\| \leq C\left\|(-\hat{R} + E[Z_i z_i'])\delta\right\| \leq \mu_n^{-1}\sqrt{n}\left\|\hat{g}(\beta)\right\| + \hat{M},$$

giving Assumption 2 iii).

Next, Assumption 3 i) holds by Lemma A5 and $E[(Z_i'Z_i)^2]/n \longrightarrow 0$, ii) by Lemma A6, iii) by Lemma A9, iv) by Lemma A8, and v) by Lemma A7. *Q.E.D.*

1.3 Asymptotic Normality

The next result is a general result on asymptotic normality of the sum of a linear and a quadratic form. Let X_i denote a scalar random variable where we also suppress dependence on n, let Z_i and Y_i be $m \times 1$ random vectors as in Lemma A1, $\Psi = \Sigma_{ZZ}\Sigma_{YY} + \Sigma_{ZY}^2$, $\bar{\xi}_Z = \xi_{\max}(\Sigma_{ZZ})$, and $\bar{\xi}_Y = \xi_{\max}(\Sigma_{YY})$.

LEMMA A10: *If* $(X_i, Y_i, Z_i), (i = 1, ..., n)$ *are i.i.d.,* $E[X_i] = 0$, $E[Z_i] = E[Y_i] = 0$, Σ_{ZZ} *and* Σ_{YY} *exist,* $nE[X_i^2] \to A$, $n^2\operatorname{tr}(\Psi) \to \Lambda$, $nE[X_i^4] \longrightarrow 0$, $mn^4\bar{\xi}_Z^2\bar{\xi}_Y^2 \longrightarrow 0$, $n^3(\bar{\xi}_Z^2 E[\left\|Y_i\right\|^4] + \bar{\xi}_Y^2 E[\left\|Z_i\right\|^4]) \longrightarrow 0$, *and* $n^2 E[\left\|Y_i\right\|^4]E[\left\|Z_i\right\|^4] \longrightarrow 0$ *then*

$$\sum_{i=1}^n X_i + \sum_{i\neq j} Z_i'Y_j \xrightarrow{d} N(0, A + \Lambda).$$

Proof: Let $w_i = (X_i, Y_i, Z_i)$ and for any $j < i$, $\psi_{ij} = Z_i'Y_j + Z_j'Y_i$. Note that

$$E[\psi_{ij}|w_{i-1}, ..., w_1] = 0, E[\psi_{ij}^2] = E[(Z_i'Y_j)^2 + (Z_j'Y_i)^2 + 2Z_i'Y_jZ_j'Y_i] = 2tr(\Psi).$$

We have

$$\sum_{i=1}^{n} X_i + \sum_{i \neq j} Z_i'Y_j = \sum_{i=2}^{n}(X_i + B_{in}) + X_1, B_{in} = \sum_{j<i} \psi_{ij} = (\sum_{j<i} Z_j)'Y_i + (\sum_{j<i} Y_j)'Z_i.$$

Note that $E[X_1^2] = (nE[X_i^2])/n \longrightarrow 0$, so $X_1 \xrightarrow{p} 0$ by M. Also, $E[X_iB_{in}] = 0$ and

$$E\left[B_{in}^2\right] = E\left[\sum_{j,k<i} \psi_{ij}\psi_{ik}\right] = (i-1)E[\psi_{ij}^2] = 2(i-1)\,\mathrm{tr}(\Psi).$$

Therefore

$$\begin{aligned} s_n &= \sum_{i=2}^{n} E[(X_i + B_{in})^2] = (n-1)E[X_i^2] + 2\sum_{i=2}^{n}(i-1)\,\mathrm{tr}(\Psi) \\ &= \frac{n-1}{n} nE[X_i^2] + \left(\frac{n^2-n}{n^2}\right) n^2\,\mathrm{tr}(\Psi) \longrightarrow A + \Lambda. \end{aligned} \tag{1.4}$$

Next, note that

$$\begin{aligned} E[B_{in}^2|w_{i-1}, ..., w_1] &= T_{1i} + T_{2i} + 2T_{3i},\ T_{1i} = (\sum_{j<i} Z_j')\Sigma_{YY}(\sum_{j<i} Z_j), \\ T_{2i} &= (\sum_{j<i} Y_j')\Sigma_{ZZ}(\sum_{j<i} Y_j), T_{3i} = (\sum_{j<i} Y_j')\Sigma_{ZY}(\sum_{j<i} Z_j). \end{aligned}$$

We also have

$$\begin{aligned} T_{3i} - E[T_{3i}] &= T_{31i} + T_{32i} + T_{33i},\ T_{31i} = \sum_{j<i} R_j,\ R_j = [Y_j'\Sigma_{ZY}Z_j - tr(\Sigma_{ZY}^2)], \\ T_{32i} &= \sum_{k<i} S_k,\ S_k = (\sum_{j<k} Y_j)'\Sigma_{ZY}Z_k,\ T_{33i} = \sum_{j<k<i} Y_k'\Sigma_{ZY}Z_j. \end{aligned}$$

By $E[(Y_i', Z_i')'(Y_i', Z_i')']$ being p.s.d. it follows that $|Y_j'\Sigma_{ZY}Z_j| \leq (Y_j'\Sigma_{ZZ}Y_j + Z_j'\Sigma_{YY}Z_j)/2$. Note that

$$E[(Y_j'\Sigma_{ZY}Z_j)^2] \leq CE[(Y_j'\Sigma_{ZZ}Y_j)^2] + CE[(Z_j'\Sigma_{YY}Z_j)^2] \leq C\bar{\xi}_Z^2 E[\|Y_j\|^4] + C\bar{\xi}_Y^2 E[\|Z_j\|^4].$$

Note that $\sum_{i=2}^{n} T_{31i} = \sum_{i=2}^{n}(n-i+1)R_i$ so that

$$E[\left(\sum_{i=2}^{n} T_{31i}\right)^2] \leq E[(Y_j'\Sigma_{ZY}Z_j)^2]\sum_{i=2}^{n}(n-i+1)^2 \leq Cn^3\{\bar{\xi}_Z^2 E[\|Y_j\|^4] + \bar{\xi}_Y^2 E[\|Z_j\|^4]\} \longrightarrow 0,$$

so that $\sum_{i=2}^{n} T_{31i} \xrightarrow{p} 0$ by M. We also have,

$$\begin{aligned} E[Y_i'\Sigma_{ZY}\Sigma_{ZZ}\Sigma_{YZ}Y_i] &\leq \bar{\xi}_Z E[Y_i'\Sigma_{ZY}\Sigma_{YZ}Y_i] = \bar{\xi}_Z tr(\Sigma_{YZ}\Sigma_{YY}\Sigma_{ZY}) \leq \bar{\xi}_Z\bar{\xi}_Y tr(\Sigma_{YZ}\Sigma_{ZY}) \\ &\leq \bar{\xi}_Z^2\bar{\xi}_Y tr(\Sigma_{YZ}\Sigma_{ZZ}^{-1}\Sigma_{ZY}) \leq \bar{\xi}_Z^2\bar{\xi}_Y tr(\Sigma_{YY}) \leq m\bar{\xi}_Z^2\bar{\xi}_Y^2, \end{aligned}$$

so that $E[S_i^2] \leq (i-1)m\bar{\xi}_Z^2\bar{\xi}_Y^2$. In addition $E[S_i|w_{i-1},...,w_1] = 0$, so that

$$\begin{aligned} E[\left(\sum_{i=3}^{n} T_{32i}\right)^2] &= E[\{\sum_{i=3}^{n}(n-i+1)S_i\}^2] = \sum_{i=3}^{n}(n-i+1)^2 E[S_i^2] \\ &\leq \sum_{i=3}^{n}(n-i+1)^2(i-1)m\bar{\xi}_Z^2\bar{\xi}_Y^2 \leq mn^4\bar{\xi}_Z^2\bar{\xi}_Y^2 \longrightarrow 0, \end{aligned}$$

and hence $\sum_{i=3}^{n} T_{32i} \xrightarrow{p} 0$. It follows analogously that $\sum_{i=3}^{n} T_{33i} \xrightarrow{p} 0$, so by T, $\sum_{i=3}^{n}\{T_{3i} - E[T_{3i}]\} \xrightarrow{p} 0$. By similar arguments we have $\sum_{i=2}^{n}\{T_{ri} - E[T_{ri}]\} \xrightarrow{p} 0$, $(r = 1, 2)$, so by T,

$$\sum_{i=2}^{n}(E[B_{in}^2|w_{i-1},...,w_1] - E[B_{in}^2]) \xrightarrow{p} 0.$$

Note also that $E[X_i^2] = E[X_i^2 \mid w_{i-1},...,w_1]$ and that

$$\begin{aligned} \sum_{i=2}^{n} E[X_iB_{in}|w_{i-1},...,w_1] &= \sum_{i=2}^{n}\sum_{j<i} E\left[X_i\left(Z_i'Y_j + Z_j'Y_i\right) \mid w_{i-1},...,w_1\right] \\ &= \sum_{i=2}^{n}\{E[X_iZ_i']\left(\sum_{j<i} Y_j\right) + E[X_iY_i']\left(\sum_{j<i} Z_j\right)\} \\ &= E[X_iZ_i']\sum_{i=1}^{n-1}(n-i)Y_i + E[X_iY_i']\sum_{i=1}^{n-1}(n-i)Z_i. \end{aligned}$$

Therefore

$$\begin{aligned} & E\left[\left(\sum_{i=2}^{n} E[X_iB_{in}|w_{i-1},...,w_1]\right)^2\right] \\ \leq\ & C(E[X_iY_i']\Sigma_{ZZ}E[Y_iX_i] + E[X_iZ_i']\Sigma_{YY}E[Z_iX_i])\sum_{i=1}^{n-1}(n-i)^2 \\ \leq\ & Cn^3\bar{\xi}_Y\bar{\xi}_Z E[X_i^2] \leq C\bar{\xi}_Y\bar{\xi}_Z n^2 = C(mn^4\bar{\xi}_Y^2\bar{\xi}_Z^2)^{1/2}/m^{1/2} \longrightarrow 0. \end{aligned}$$

Then by M, we have

$$\sum_{i=2}^{n} E[X_iB_{in} \mid w_{i-1},...,w_1] \xrightarrow{p} 0.$$

By T it then follows that

$$\sum_{i=2}^{n}\{E\left[(X_i+B_{in})^2 \mid w_{i-1},...,w_1\right]-E\left[(X_i+B_{in})^2\right]\} \tag{1.5}$$
$$= \sum_{i=2}^{n}\left(E\left[B_{in}^2 \mid w_{i-1},...,w_1\right]-E[B_{in}^2]\right)+2\sum_{i=2}^{n}E[X_iB_{in} \mid w_{i-1},...,w_1] \xrightarrow{p} 0$$

Next, note that

$$\begin{aligned}\sum_{i=2}^{n}E[(\sum_{j<i}Y_j'Z_i)^4] &= \sum_{i=2}^{n}\sum_{j,k,\ell,m<i}E[Y_j'Z_iY_k'Z_iY_\ell'Z_iY_m'Z_i]\\ &= \sum_{i=2}^{n}\{3\sum_{j\neq k<i}E[Z_i'Y_jY_j'Z_iZ_i'Y_kY_k'Z_i]+\sum_{j<i}E[(Z_i'Y_j)^4]\}\\ &= E[(Z_1'\Sigma_{YY}Z_1)^2]\sum_{i=2}^{n}3(i-1)(i-2)+E[(Z_1'Y_2)^4]\sum_{i=2}^{n}(i-1)\\ &\leq n^3\bar{\xi}_Y^2E[\|Z_i\|^4]+n^2E[\|Z_i\|^4]E[\|Y_i\|^4]\longrightarrow 0.\end{aligned}$$

It follows similarly that $\sum_{i=2}^{n}E[(\sum_{j<i}Z_j'Y_i)^4]\longrightarrow 0$. Then by T,

$$\sum_{i=2}^{n}E[B_{in}^4]\leq\sum_{i=2}^{n}\{CE[(\sum_{j<i}Y_j'Z_i)^4]+CE[(\sum_{j<i}Z_j'Y_i)^4]\}\longrightarrow 0.$$

Therefore,

$$\sum_{i=2}^{n}E\left[(X_i+B_{in})^4\right]\leq CnE[X_i^4]+C\sum_{i=1}^{n}E[B_{in}^4]\to 0. \tag{1.6}$$

The conclusion then follows from eqs. (1.4), (1.5), and (1.6) and the martingale central limit theorem applied to $\sum_{i=2}^{n}(X_i+B_{in})$. Q.E.D.

We again consider the parameterization where $\delta = S_n'(\beta-\beta_0)/\mu_n$ and $\beta = \beta_0 + \mu_nS_n^{-1\prime}\delta$. We will let a δ subscript denote derivatives with respect to δ, e.g. so that $g_{i\delta_k} = \partial g_i(0)/\partial\delta_k = G_iS_n^{-1\prime}e_k\mu_n$, where e_k is the k^{th} unit vector. Also let $\tilde{\Omega} = \hat{\Omega}(\beta_0)$, $\tilde{\Omega}^k = \sum_{i=1}^{n}g_ig_{i\delta_k}'/n$, $\Omega^k = E[\tilde{\Omega}^k]$, $\tilde{B}^k = \tilde{\Omega}^{-1}\tilde{\Omega}^k$, and $B^k = \Omega^{-1}\Omega^k$.

LEMMA A11: *If Assumptions 1-4 and 6-9 are satisfied then*

$$\sqrt{m}\|\tilde{\Omega}-\Omega\|\xrightarrow{p}0,\ \mu_n\sqrt{m}\|\tilde{\Omega}^k-\Omega^k\|\xrightarrow{p}0,\ \sqrt{m}\|\tilde{B}^k-B^k\|\xrightarrow{p}0.$$

Proof: Note that $\mu_n S_n^{-1}$ is bounded, so that $\|g_{i\delta_k}\| \leq C\|G_i\|$. Then by standard arguments and Assumption 6,

$$E[m\|\tilde{\Omega}-\Omega\|^2] \leq CmE[\|g_i\|^4]/n \longrightarrow 0,\ E[m\|\tilde{\Omega}^k-\Omega^k\|^2] \leq CmE[\|g_{i\delta_k}\|^2\|g_i\|^2]/n \longrightarrow 0,$$

so the first two conclusions hold by M. Also, note that $\Omega^{k\prime}\Omega^k \leq C\Omega^{k\prime}\Omega^{-1}\Omega^k \leq CE[g_{i\delta_k}g'_{i\delta_k}]$, so that by Assumption 6, $\xi_{\max}\left(\Omega^{k\prime}\Omega^k\right) \leq C$. Also, $B^{k\prime}B^k \leq C\Omega^{k\prime}\Omega^k \leq CE[g_{i\delta_k}g'_{i\delta_k}]$. Then w.p.a.1,

$$\begin{aligned}\sqrt{m}\|\tilde{B}^k-B^k\| &\leq \sqrt{m}\|(\tilde{\Omega}^{k\prime}-\Omega^{k\prime})\tilde{\Omega}^{-1}\| + \sqrt{m}\|B^{k\prime}(\Omega-\tilde{\Omega})\tilde{\Omega}^{-1}\| \\ &\leq C\sqrt{m}\|\tilde{\Omega}^k-\Omega^k\| + C\sqrt{m}\|\tilde{\Omega}-\Omega\| \xrightarrow{p} 0.\ Q.E.D.\end{aligned}$$

LEMMA A12: *If Assumption 1-4 and 6-9 are satisfied then,*

$$nS_n^{-1}\frac{\partial\hat{Q}(\beta_0)}{\partial\beta} = \mu_n^{-1}n\frac{\partial\hat{Q}(0)}{\partial\delta} \xrightarrow{d} N(0,H+\Lambda) = N(0,HVH).$$

Proof: Let $\tilde{g} = \hat{g}(\beta_0)$, $\tilde{g}_{\delta_k} = \partial\hat{g}(0)/\partial\delta_k = \sum_i G_i S_n^{-1\prime}e_k\mu_n/n$, $\bar{g}_{\delta_k} = E[\partial g_i(0)/\partial\delta_k] = GS_n^{-1\prime}e_k\mu_n$, $\hat{U}^k = \tilde{g}_{\delta_k} - \bar{g}_{\delta_k} - \tilde{B}^{k\prime}\tilde{g}$, and let $\tilde{\lambda}$ be as defined in Lemma A3. Consider an expansion $\rho_1(\tilde{\lambda}'g_i) = -1 - \tilde{\lambda}'g_i + \rho_3(\bar{v}_i)(\tilde{\lambda}'g_i)^2/2$, where $|\bar{v}_i| \leq \left|\tilde{\lambda}'g_i\right|$. By the envelope theorem and by $\hat{Q}(\delta) = \hat{Q}(\beta_0 + \mu_n S_n^{-1\prime}\delta)$

$$\begin{aligned}ne'_kS_n^{-1}\partial\hat{Q}(\beta_0)/\partial\beta &= n[\partial\hat{Q}(\beta_0)/\partial\beta]'S_n^{-1\prime}e_k = \mu_n^{-1}n\frac{\partial\hat{Q}}{\partial\delta_k}(0) \\ &= \mu_n^{-1}\sum_i \tilde{\lambda}'g_{i\delta_k}\rho_1(\tilde{\lambda}'g_i) = -\mu_n^{-1}n\tilde{g}'_{\delta_k}\tilde{\lambda} - \mu_n^{-1}n\tilde{\lambda}'\tilde{\Omega}^k\tilde{\lambda} + \hat{r}, \\ \hat{r} &= \mu_n^{-1}\sum_i \tilde{\lambda}'g_{i\delta_k}\rho_3\left(\bar{v}_i\right)\left(\tilde{\lambda}'g_i\right)^2/2.\end{aligned}$$

By Lemma A3, $\left\|\tilde{\lambda}\right\| = O_p\left(\sqrt{m/n}\right)$. Note that either β is the CUE or $\max_{i\leq n}|\bar{v}_i| \leq \left\|\tilde{\lambda}\right\|\hat{b}$ for $\hat{b} = \max_{i\leq n}\|g_i\|$, and that $\hat{b} = O_p(n^{1/\gamma}(E[b_i^\gamma])^{1/\gamma})$ by a standard result. Therefore, by Assumption 9, either $\hat{\beta}$ is the CUE or $\max_{i\leq n}|\bar{v}_i| \leq O_p(\sqrt{m/n})\hat{b} = O_p(n^{1/\gamma}(E[b_i^\gamma])^{1/\gamma}\sqrt{m/n})$ $\xrightarrow{p} 0$. It follows that $\max_{i\leq n}\rho_3\left(\bar{\xi}'_i g_i\right) \leq C$ w.p.a.1 and, by $\xi_{\max}\left(\tilde{\Omega}\right) = O_p(1)$, $\sqrt{m}/\mu_n \leq C$, and Assumption 9 that either $\hat{r} = 0$ for the CUE or

$$|\hat{r}| \leq \mu_n^{-1}C\left\|\tilde{\lambda}\right\|\hat{b}n\bar{\xi}'\tilde{\Omega}\bar{\xi} = O_p(\mu_n^{-1}m^{3/2}n^{1/\gamma}(E[b_i^\gamma])^{1/\gamma}/\sqrt{n}) = O_p(n^{1/\gamma}(E[b_i^\gamma])^{1/\gamma}m/\sqrt{n}) \xrightarrow{p} 0.$$

As in Lemma A3, w.p.a.1 $\tilde{\lambda}$ satisfies the first-order conditions

$$\sum_i \rho_1\left(\tilde{\lambda}' g_i\right) g_i/n = 0.$$

Plugging in the expansion for $\rho_1(\tilde{\lambda}' g_i)$ and solving give

$$\tilde{\lambda} = -\tilde{\Omega}^{-1}\tilde{g} + \hat{R}, \hat{R} = \tilde{\Omega}^{-1}\textstyle\sum_i \rho_3\left(\bar{v}_i\right) g_i \left(\tilde{\lambda}' g_i\right)^2 /n.$$

Either $\hat{R} = 0$ for the CUE or by $\xi_{\max}(\tilde{\Omega}^{-1}) \leq C$ and $\xi_{\max}(\tilde{\Omega}) \leq C$ w.p.a.1,

$$\left\|\hat{R}\right\| \leq C \max_{i \leq n} \|g_i\| \, \tilde{\lambda}'\tilde{\Omega}\tilde{\lambda} \leq C\hat{b}\left\|\tilde{\lambda}\right\|^2 = O_p(n^{1/\gamma}(E[b_i^\gamma])^{1/\gamma} m/n).$$

Now, plug $\hat{\lambda}$ back in the expression for $\partial\hat{Q}(0)/\partial\delta_k$ to obtain

$$\begin{aligned}\mu_n^{-1} n \frac{\partial \hat{Q}}{\partial \delta_k}(0) &= \mu_n^{-1} n \tilde{g}_{\delta_k}' \tilde{\Omega}^{-1}\tilde{g} - \mu_n^{-1} n \tilde{g}' \tilde{B}^k \tilde{\Omega}^{-1}\tilde{g} + \hat{r} \\ &\quad + \mu_n^{-1} n \tilde{g}_{\delta_k}' \hat{R} - \mu_n^{-1} n \hat{R}' \tilde{\Omega}^k \hat{R} + \mu_n^{-1} n \hat{R}\left(\tilde{\Omega}^k + \tilde{\Omega}^{k\prime}\right)\tilde{\Omega}^{-1}\tilde{g}.\end{aligned}$$

Note that by Assumption 6 and $\mu_n S_n^{-1}$ bounded, $E[\|g_{i\delta_k}\|^2] = tr(E[g_{i\delta_k} g_{i\delta_k}']) \leq Cm\xi_{\max}(E[G_i G_i']) \leq Cm$. Therefore, $\|\tilde{g}_{\delta_k} - \bar{g}_{\delta_k}\| = O_p(\sqrt{m/n})$. We also have $\|\mu_n^{-1}\sqrt{n}\bar{g}_{\delta_k}\| \leq \|\sqrt{n} G S_n^{-1\prime}\| \leq C$, so that $\|\bar{g}_{\delta_k}\| = O(\mu_n/\sqrt{n})$. Therefore, by $\sqrt{m}/\mu_n \leq C$ and T, $\|\tilde{g}_{\delta_k}\| = O_p(\mu_n/\sqrt{n})$, so by CS

$$\left|\mu_n^{-1} n \tilde{g}_{\delta_k}' \hat{R}\right| \leq \mu_n^{-1} n \left\|\tilde{g}_{\delta_k}\right\| \left\|\hat{R}\right\| = O_p\left(\sqrt{n} n^{1/\gamma}(E[b_i^\gamma])^{1/\gamma} m/n\right) \xrightarrow{p} 0.$$

Let $\tilde{\Omega}^{k,k} = \sum_i g_{i\delta_k} g_{i\delta_k}'/n$ and $\Omega^{k,k} = E\left[g_{i\delta_k} g_{i\delta_k}'\right]$. By Assumption 6 and M we have $\left\|\tilde{\Omega}^{k,k} - \Omega^{k,k}\right\| \xrightarrow{p} 0$, so by Lemma A0, Assumption 6 and $\mu_n S_n^{-1}$ bounded, w.p.a1

$$\xi_{\max}(\tilde{\Omega}^{k,k}) \leq \xi_{\max}(\Omega^{k,k}) + 1 \leq C\xi_{\max}(E[G_i G_i']) + 1.$$

Therefore, $\hat{M} = \sqrt{\xi_{\max}(\tilde{\Omega})\xi_{\max}(\tilde{\Omega}^{k,k})} = O_p(1)$, so that for any a, b, by CS,

$$\left|a'\tilde{\Omega}^k b\right| \leq [a'\tilde{\Omega} a b'\tilde{\Omega}^{k,k} b]^{1/2} \leq \hat{M}\|a\|\|b\|.$$

Then

$$\left|\mu_n^{-1} n \hat{R}'\tilde{\Omega}^k \hat{R}\right| \leq \hat{M}\mu_n^{-1} n \left\|\hat{R}\right\|^2 = O_p\left(\mu_n^{-1}\{n^{1/\gamma}(E[b_i^\gamma])^{1/\gamma} m/\sqrt{n}\}^2\right) \xrightarrow{p} 0.$$

We also have $\left\|\tilde{\Omega}^{-1}\tilde{g}\right\| = O_p\left(\sqrt{m/n}\right)$, so that by $\sqrt{m}/\mu_n \leq C$,

$$\left|\mu_n^{-1} n\hat{R}\left(\tilde{\Omega}^k + \tilde{\Omega}^{k\prime}\right)\tilde{\Omega}^{-1}\tilde{g}\right| \leq C\hat{M}\mu_n^{-1} n\left\|\hat{R}\right\|\left\|\tilde{\Omega}^{-1}\tilde{g}\right\| = O_p\left(n^{1/\gamma}(E[b_i^\gamma])^{1/\gamma} m/\sqrt{n}\right) \xrightarrow{p} 0.$$

By T it now follows that

$$\begin{aligned}\mu_n^{-1} n\frac{\partial \hat{Q}}{\partial \delta_k}(0) &= \mu_n^{-1} n\tilde{g}_{\delta_k}'\tilde{\Omega}^{-1}\tilde{g} - \mu_n^{-1} n\tilde{g}'\tilde{B}^k\tilde{\Omega}^{-1}\tilde{g} + o_p(1) \\ &= \bar{g}_{\delta_k}'\tilde{\Omega}^{-1}\tilde{g} + \hat{U}^{k\prime}\tilde{\Omega}^{-1}\tilde{g} + o_p(1),\end{aligned}$$

where $\hat{U}^k = \tilde{g}_{\delta_k} - \bar{g}_{\delta_k} - \tilde{B}^{k\prime}\tilde{g}$. For B^k defined preceding Lemma A11 let $\tilde{U}^k = \tilde{g}_{\delta_k} - \bar{g}_{\delta_k} - B^{k\prime}\tilde{g}$. Note that $n\|\tilde{g}\|^2 = O_p(m)$. By Lemma A11 and $m/\mu_n^2 \leq C$ we have

$$n\mu_n^{-1}|(\hat{U}^{k\prime}\tilde{\Omega}^{-1} - \tilde{U}^{k\prime}\tilde{\Omega}^{-1})\tilde{g}| \leq Cn\mu_n^{-1}|\tilde{g}'(\tilde{B}^k - B^k)\tilde{\Omega}^{-1}\tilde{g}| \leq Cn\mu_n^{-1}\|\tilde{g}\|^2\|\tilde{B}^k - B^k\| \xrightarrow{p} 0.$$

Note also that by the usual properties of projections and Assumption 6, $nE[\left\|\tilde{U}^k\right\|^2] \leq CE[\|g_{i\delta_k}\|^2] \leq Cm$, so that $n\mu_n^{-1}|\tilde{U}^{k\prime}(\Omega^{-1} - \tilde{\Omega}^{-1})\tilde{g}| \xrightarrow{p} 0$. Similarly we have $\mu_n^{-1}\bar{g}_{\delta_k}'(\tilde{\Omega}^{-1} - \Omega^{-1})\tilde{g} \xrightarrow{p} 0$, so that by T

$$n\mu_n^{-1}\frac{\partial \hat{Q}}{\partial \delta_k}(0) = n\mu_n^{-1}(\bar{g}_{\delta_k} + \tilde{U}^k)'\Omega^{-1}\tilde{g} + o_p(1).$$

It is straightforward to check that for U_i defined in Section 2 we have

$$\tilde{U}^k = n^{-1}\sum_{i=1}^n U_i S_n^{-1\prime} e_k \mu_n,\ \bar{g}_{\delta_k} = GS_n^{-1\prime} e_k \mu_n.$$

Then stacking over k gives

$$n\mu_n^{-1}\frac{\partial \hat{Q}}{\partial \delta}(0) = nS_n^{-1}[G'\Omega^{-1}\tilde{g} + n^{-1}\sum_{i=1}^n U_i'\Omega^{-1}\tilde{g}] + o_p(1). \tag{1.7}$$

For any vector λ with $\|\lambda\| = 1$ let $X_i = \lambda' S_n^{-1} G'\Omega^{-1} g_i$, $Y_i = \Omega^{-1} g_i$, $Z_i = U_i S_n^{-1\prime}\lambda/n$, and $A = \lambda' H\lambda$. Then from the previous equation we have

$$n\mu_n^{-1}\lambda'\frac{\partial \hat{Q}}{\partial \delta}(0) = \sum_{i=1}^n X_i + \sum_{i,j=1}^n Y_i' Z_i + o_p(1).$$

Note that $E[Z_i' Y_i] = 0$ by each component of U_i being uncorrelated with every component of g_i. Also, by $\|S_n^{-1}\| \leq C/\mu_n$,

$$nE[|Y_i' Z_i|^2] \leq CE[\left\|g_i'\Omega^{-1}U_i\right\|^2]/n\mu_n^2 \leq C(E[\|g_i\|^4] + E[\|G_i\|^4])/n\mu_n^2 \longrightarrow 0.$$

Then $\sum_{i=1}^n Z_i'Y_i \xrightarrow{p} 0$ by M. Then by eq. (1.7),

$$n\mu_n^{-1}\lambda'\frac{\partial \hat{Q}(0)}{\partial \delta} = \sum_{i=1}^n X_i + \sum_{i\neq j} Z_i'Y_j + o_p(1).$$

Now apply Lemma A10. Note that $\Sigma_{YY} = \Omega^{-1}$ and $\Sigma_{ZY} = 0$, so that $\Psi = \Sigma_{ZZ}\Sigma_{YY} = n^{-2}E[U_iS_n^{-1\prime}\lambda\lambda'S_n^{-1}U_i']\Omega^{-1}$. By Assumption 1 and the hypothesis of Theorem 3, we have

$$\begin{aligned} nE[X_i^2] &= n\lambda'S_n^{-1}G'\Omega^{-1}GS_n^{-1\prime}\lambda \longrightarrow \lambda'H\lambda = A, \\ n^2tr(\Psi) &= \lambda'S_n^{-1}E[U_i'\Omega^{-1}U_i]S_n^{-1\prime}\lambda \longrightarrow \lambda'\Lambda\lambda. \end{aligned}$$

Also, note that $\xi_{\max}(S_n^{-1\prime}\lambda\lambda'S_n^{-1}) \leq C/\mu_n^2$, so that $\bar{\xi}_Z \leq C/\mu_n^2n^2$. We also have $\|\sqrt{n}S_n^{-1}G'\Omega^{-1}\| \leq C$ by Assumption 1 and $\xi_{\max}(\Omega^{-1}) \leq C$. Then

$$\begin{aligned} nE[|X_i|^4] &\leq nE[\left\|\lambda'\sqrt{n}S_n^{-1}G'\Omega^{-1}g_i\right\|^4]/n^2 \leq CE[\|g_i\|^4]/n \longrightarrow 0, \\ mn^4\bar{\xi}_Y^2\bar{\xi}_Z^2 &\leq Cmn^4/(\mu_n^2n^2)^2 \leq Cm/\mu_n^4 \longrightarrow 0, \\ n^3(\bar{\xi}_Z^2E[\|Y_i\|^4] + \bar{\xi}_Y^2E[\|Z_i\|^4]) &\leq n^3C(E[\|g_i\|^4] + E[\|G_i\|^4])/\mu_n^4n^4 \longrightarrow 0, \\ n^2E[\|Y_i\|^4]E[\|Z_i\|^4] &\leq n^2CE[\|g_i\|^4](E[\|g_i\|^4] + E[\|G_i\|^4])/\mu_n^4n^4 \longrightarrow 0. \end{aligned}$$

The conclusion then follows by the conclusion of Lemma A10 and the Cramer-Wold device. Q.E.D.

LEMMA A13: *If Assumptions 1-4 and 6-9 are satisfied then there is an open convex set N_n such that $0 \in N_n$ and w.p.a.1 $\hat{\delta} \in N_n$, $\hat{Q}(\delta)$ is twice continuous differentiable on N_n, and for any $\bar{\delta}$ that is an element of N_n w.p.a.1,*

$$nS_n^{-1}[\partial^2\hat{Q}(\bar{\delta})/\partial\beta\partial\beta']S_n^{-1\prime} = \mu_n^{-2}n\partial^2\hat{Q}(\bar{\delta})/\partial\delta\partial\delta' \xrightarrow{p} H$$

Proof: By Theorem 1 $\hat{\delta} \xrightarrow{p} 0$. Then there is $\zeta_n \longrightarrow 0$ such that w.p.a.1 $\bar{\delta} \in N_n = \{\delta : \|\delta\| < \zeta_n\}$. By Assumption 3, for all $\delta \in N_n$

$$\mu_n^{-1}\sqrt{n}\,\|\hat{g}(\delta) - \hat{g}(0)\| \leq \hat{M}\,\|\delta\| \leq \hat{M}\zeta_n \xrightarrow{p} 0.$$

As previously shown, $\mu_n^{-1}\sqrt{n}\,\|\hat{g}(0)\| = O_p\left(\mu_n^{-1}\sqrt{n}\sqrt{m/n}\right) = O_p(1)$, so $\sup_{\delta\in N_n}\mu_n^{-1}\sqrt{n}\,\|\hat{g}(\delta)\| = O_p(1)$ by T. Now let τ_n to go to zero slower than $\mu_n/\sqrt{n}$ but faster than

$n^{-1/\gamma}E[\sup_{\beta\in B}\|g_i(\beta)\|^\gamma]^{-1/\gamma}$, which is possible by Assumption 9, and let $L_n = \{\lambda : \|\lambda\| \leq \tau_n\}$. Then $\max_{i\leq n}\sup_{\beta\in B,\lambda\in L_n}|\lambda' g_i(\beta)| \xrightarrow{p} 0$ similarly to the proof of Lemma A3. For all $\delta \in N_n$ let $\hat{\lambda}(\delta) = \underset{\lambda\in L_n}{\operatorname{argmax}} \hat{P}(\delta,\lambda)$. By an argument similar to the proof of Lemma A3, an expansion of $S\left(\delta,\hat{\lambda}(\delta)\right)$ around $\lambda = 0$ gives

$$\begin{aligned} 0 &= \hat{P}(\delta,0) \leq \hat{P}\left(\delta,\hat{\lambda}(\delta)\right) = \hat{g}(\delta)'\hat{\lambda}(\delta) + \frac{1}{2}\hat{\lambda}(\delta)'\left[\sum_{i=1}^n \rho_2(\lambda' g_i(\delta))g_i(\delta)g_i(\delta)'/n\right]\hat{\lambda}(\delta) \\ &\leq \|\hat{g}(\delta)\|\left\|\hat{\lambda}(\delta)\right\| - C\left\|\hat{\lambda}(\delta)\right\|^2. \end{aligned}$$

Adding $C\left\|\hat{\lambda}(\delta)\right\|^2$ and dividing through by $C\left\|\hat{\lambda}(\delta)\right\|$ gives

$$\left\|\hat{\lambda}(\delta)\right\| \leq C\|\hat{g}(\delta)\| \leq C\sup_{\delta\in N_n}\|\hat{g}(\delta)\| = O_p\left(\mu_n/\sqrt{n}\right). \tag{1.8}$$

It follows that w.p.a.1 $\hat{\lambda}(\delta) \in \operatorname{int} L_n$ for all $\delta \in N_n$. Since a local maximum of a concave function is a global maximum, w.p.a.1 for all $\delta \in N_n$,

$$\hat{Q}(\delta) = \hat{P}(\delta,\hat{\lambda}(\delta)).$$

Furthermore w.p.a.1 the first-order conditions

$$\sum_{i=1}^n \rho_1\left(\hat{\lambda}(\delta)' g_i(\delta)\right) g_i(\delta)/n = 0$$

will be satisfied for all δ, so that by the implicit function theorem $\hat{\lambda}(\delta)$ is twice continuously differentiable in $\delta \in N_n$ and hence so is $\hat{Q}(\delta)$.

Here let $\hat{g}_i = g_i(\bar{\delta})$, $\hat{g} = \hat{g}(\bar{\delta})$, $\hat{\lambda} = \hat{\lambda}(\bar{\delta})$, $\hat{\Omega} = -\sum_{i=1}^n \rho_2(\hat{\lambda}'\hat{g}_i)\hat{g}_i\hat{g}_i'/n$, $\hat{g}_{i\delta_k} = \partial g_i(\bar{\delta})/\partial\delta_k$, $\hat{g}_{\delta_k} = \partial\hat{g}(\bar{\delta})/\partial\delta_k$, $\hat{\Omega}^k = -\sum_i \rho_2\left(\hat{\lambda}'\hat{g}_i\right)\hat{g}_i\hat{g}_{i\delta_k}'/n$. Then expanding $\rho_1(\hat{\lambda}'\hat{g}_i) = -1+\rho_2(\bar{v}_i)\hat{\lambda}'\hat{g}_i$, for $|\bar{v}_i| \leq \left|\hat{\lambda}'\hat{g}_i\right|$, and letting $\bar{\Omega}^k = -\sum_i \rho_2(\bar{v}_i)\hat{g}_i\hat{g}_{i\delta_k}'/n$, the implicit function theorem gives

$$\begin{aligned} \hat{\lambda}_{\delta_k} &= \frac{\partial\hat{\lambda}}{\partial\delta_k}(\bar{\delta}) = \hat{\Omega}^{-1}\left[\sum_i \rho_1\left(\hat{\lambda}'\hat{g}_i\right)\hat{g}_{i\delta_k}/n - \hat{\Omega}^k\hat{\lambda}/n.\right] \\ &= -\hat{\Omega}^{-1}\left[\hat{g}_{\delta_k} + \left(\bar{\Omega}^{k\prime} + \hat{\Omega}^k\right)\hat{\lambda}\right], \end{aligned}$$

Also, for $\bar{\Omega} = -\sum_i \rho_2(\bar{v}_i)\hat{g}_i\hat{g}_i'/n$, the first order conditions $0 = \sum_i \rho_1(\hat{\lambda}'\hat{g}_i)\hat{g}_i/n = -\hat{g}-\bar{\Omega}\hat{\lambda}$ imply that

$$\hat{\lambda} = -\bar{\Omega}^{-1}\hat{g}.$$

Next, by the envelope theorem it follows that

$$\hat{Q}_{\delta_k}\left(\bar{\delta}\right) = \sum_i \rho_1\left(\hat{\lambda}'\hat{g}_i\right)\hat{\lambda}'\hat{g}_{i\delta_k}/n.$$

Let $\hat{g}_{i\delta_k\delta_\ell} = \partial^2 g_i(\hat{\delta})/\partial\delta_k\partial\delta_\ell$, $\hat{g}_{\delta_k\delta_\ell} = \partial^2\hat{g}(\hat{\delta})/\partial\delta_k\partial\delta_\ell$, $\hat{\Omega}^{k,\ell} = -\sum_i \rho_2(\hat{\lambda}'\hat{g}_i)\hat{g}_{i\delta_k}\hat{g}'_{i\delta_\ell}/n$, $\bar{\Omega}^{k\ell} = -\sum_i \rho_2\left(\bar{v}_i\right)\hat{g}_i\hat{g}'_{i\delta_k\delta_\ell}/n$. Differentiating again

$$\begin{aligned}
\hat{Q}_{\delta_k\delta_\ell}\left(\bar{\delta}\right) &= \sum_i[\rho_1\left(\hat{\lambda}'g_i\right)\left(\hat{\lambda}'_{\delta_k}\hat{g}_{i\delta_\ell} + \hat{\lambda}'\hat{g}_{i\delta_k\delta_\ell}\right) + \rho_2\left(\hat{\lambda}'\hat{g}_i\right)\left(\hat{\lambda}'_{\delta_k}\hat{g}_i + \hat{\lambda}'\hat{g}_{i\delta_k}\right)\hat{\lambda}'\hat{g}_{i\delta_\ell}]/n \\
&= n^{-1}\sum_i\left[\left(-1+\rho_2\left(\bar{v}_i\right)\hat{\lambda}'\hat{g}_i\right)\left(\hat{\lambda}'_{\delta_k}\hat{g}_{i\delta_\ell} + \hat{\lambda}'\hat{g}_{i\delta_k\delta_\ell}\right)\right] - \hat{\lambda}'_{\delta_k}\hat{\Omega}^{\ell}\hat{\lambda} - \hat{\lambda}'\hat{\Omega}^{k,\ell}\hat{\lambda} \\
&= -\hat{\lambda}'_{\delta_k}\hat{g}_{\delta_\ell} - \hat{\lambda}'\hat{g}_{\delta_k\delta_\ell} - \hat{\lambda}'(\bar{\Omega}^{\ell} + \hat{\Omega}^{\ell\prime})\hat{\lambda}_{\delta_k} - \hat{\lambda}'\left(\bar{\Omega}^{k\ell} + \hat{\Omega}^{k,\ell}\right)\hat{\lambda}.
\end{aligned}$$

Substituting in the formula for $\hat{\lambda}_{\delta_k}$ and then $\hat{\lambda}$ we obtain

$$\begin{aligned}
\hat{Q}_{\delta_k\delta_\ell}\left(\bar{\delta}\right) &= \hat{g}'_{\delta_k}\hat{\Omega}^{-1}\hat{g}_{\delta_\ell} + \hat{\lambda}'\left(\bar{\Omega}^{k} + \hat{\Omega}^{k\prime}\right)\hat{\Omega}^{-1}\hat{g}_{\delta_\ell} - \hat{\lambda}'\hat{g}_{\delta_k\delta_\ell} + \hat{\lambda}'(\bar{\Omega}^{\ell} + \hat{\Omega}^{\ell\prime})\hat{\Omega}^{-1}\hat{g}_{\delta_k} \qquad (1.9) \\
&\quad +\hat{\lambda}'(\bar{\Omega}^{\ell} + \hat{\Omega}^{\ell\prime})\hat{\Omega}^{-1}\left(\bar{\Omega}^{k\prime} + \hat{\Omega}^{k}\right)\hat{\lambda} - \hat{\lambda}'\left(\bar{\Omega}^{k\ell} + \hat{\Omega}^{k,\ell}\right)\hat{\lambda} \\
&= \hat{g}'_{\delta_k}\hat{\Omega}^{-1}\hat{g}_{\delta_\ell} + \hat{g}'\bar{\Omega}^{-1}\hat{g}_{\delta_k\delta_\ell} - \hat{g}'\bar{\Omega}^{-1}\left(\bar{\Omega}^{k} + \hat{\Omega}^{k\prime}\right)\hat{\Omega}^{-1}\hat{g}_{\delta_\ell} - \hat{g}'\bar{\Omega}^{-1}(\bar{\Omega}^{\ell} + \hat{\Omega}^{\ell\prime})\hat{\Omega}^{-1}\hat{g}_{\delta_k} \\
&\quad +\hat{g}'\bar{\Omega}^{-1}(\bar{\Omega}^{\ell} + \hat{\Omega}^{\ell\prime})\hat{\Omega}^{-1}\left(\bar{\Omega}^{k\prime} + \hat{\Omega}^{k}\right)\bar{\Omega}^{-1}\hat{g} - \hat{g}'\bar{\Omega}^{-1}\left(\bar{\Omega}^{k\ell} + \hat{\Omega}^{k,\ell}\right)\bar{\Omega}^{-1}\hat{g}.
\end{aligned}$$

Next, let $\breve{\Omega}^k = \sum_i \hat{g}_i\hat{g}'_{i\delta_k}/n$. Note that $|1+\rho_2(\bar{v}_i)| \leq C|\bar{v}_i| \leq C|\hat{\lambda}'\hat{g}_i|$, so that by CS and M,

$$\begin{aligned}
\left\|\bar{\Omega}^k - \breve{\Omega}^k\right\| &\leq C\sum_i |\bar{v}_i|\left\|\hat{g}_i\right\|\left\|\hat{g}_{i\delta_k}\right\|/n \leq \left(C\sum_i \bar{v}_i^2/n\right)^{1/2}\left(\sum_i \left\|\hat{g}_i\right\|^2\left\|\hat{g}_{i\delta_k}\right\|^2/n\right)^{1/2} \\
&\leq C\left(\hat{\lambda}'\hat{\Omega}\hat{\lambda}\right)^{1/2}\left[\sum_i\left(\left\|\hat{g}_i\right\|^4 + \left\|\hat{g}_{i\delta_k}\right\|^4\right)/n\right]^{1/2} = O_p(\{\mu_n^2 E[d_i^4]/n\}^{1/2}) \xrightarrow{p} 0.
\end{aligned}$$

Also, for $\Omega^k(\delta) = E[g_i(\delta)g_{i\delta_k}(\delta)']$, by Assumption 8 i) and $S_n^{-1}\mu_n$ bounded we have $\left\|\breve{\Omega}^k - \Omega^k(\bar{\delta})\right\| \xrightarrow{p} 0$. Then by T,

$$\left\|\bar{\Omega}^k - \Omega^k(\bar{\delta})\right\| \xrightarrow{p} 0.$$

Let $\Omega^{k,\ell}(\delta) = E[g_{i\delta_k}(\delta)g_{i\delta_\ell}(\delta)']$ and $\Omega^{k\ell}(\delta) = E[g_i(\delta)g_{i\delta_k\delta_\ell}(\delta)']$. Then it follows by arguments exactly analogous to those just given that

$$\begin{aligned}
&\left\|\hat{\Omega} - \Omega(\bar{\delta})\right\| \xrightarrow{p} 0, \left\|\bar{\Omega} - \Omega(\bar{\delta})\right\| \xrightarrow{p} 0, \left\|\hat{\Omega}^k - \Omega^k(\bar{\delta})\right\| \xrightarrow{p} 0, \\
&\left\|\hat{\Omega}^{k,\ell} - \Omega^{k,\ell}(\bar{\delta})\right\| \xrightarrow{p} 0, \left\|\bar{\Omega}^{k\ell} - \Omega^{k\ell}(\bar{\delta})\right\| \xrightarrow{p} 0.
\end{aligned}$$

Next, as previously shown, $\mu_n^{-1}\sqrt{n}\left\|\hat{g}(\bar{\delta})\right\| = O_p(1)$. It follows similarly from Assumption 7 that

$$\mu_n^{-1}\sqrt{n}\left\|\partial\hat{g}(\bar{\delta})/\partial\delta\right\| = \sqrt{n}\left\|\hat{G}(\bar{\beta})S_n^{-1\prime}\right\| = \sqrt{n}\left\|\hat{G}(\beta_0)S_n^{-1\prime}\right\| + o_p(1).$$

Then by Assumption 6 $E[\|G_i\|^2] \leq Cm$, so by M,

$$\left(\sqrt{n}\left\|[\hat{G}(\beta_0) - G]S_n^{-1\prime}\right\|\right)^2 = O_p\left(E[\|G_i\|^2]\right)/\mu_n^2 = O_p(1).$$

Also by Assumptions 1 and 3 we have $\sqrt{n}\,\|GS_n^{-1\prime}\| \leq C$. Then by T and Assumption 1,

$$\sqrt{n}\left\|\hat{G}(\beta_0)S_n^{-1}\right\| \leq \sqrt{n}\left\|[\hat{G}(\beta_0) - G]S_n^{-1\prime}\right\| + \sqrt{n}\left\|GS_n^{-1\prime}\right\| = O_p(1).$$

Then by T it follows that

$$\mu_n^{-1}\sqrt{n}\left\|\partial\hat{g}(\bar{\delta})/\partial\delta\right\| = O_p(1).$$

By similar arguments it follows by Assumption 6 that

$$\mu_n^{-1}\sqrt{n}\left\|\partial^2\hat{g}(\bar{\delta})/\partial\delta\partial\delta_k\right\| = O_p(1).$$

Next, for notational convenience let $\tilde{\Omega} = \Omega(\bar{\delta})$ and $\tilde{\Omega}^k = \Omega^k(\bar{\delta})$. By Assumption 2 $\xi_{\max}(\tilde{\Omega}^{-1}) \leq C$ so that $\xi_{\max}(\tilde{\Omega}^{-2}) \leq C$. It follows as previously that $\xi_{\max}(\bar{\Omega}^{-2}) \leq C$, and $\xi_{\max}(\hat{\Omega}^{k\prime}\bar{\Omega}^{-2}\hat{\Omega}^k) \leq C$ w.p.a.1, so that

$$\begin{aligned}\left\|\bar{\Omega}^{-1}\hat{\Omega}^k\bar{\Omega}^{-1} - \tilde{\Omega}^{-1}\tilde{\Omega}^k\tilde{\Omega}^{-1}\right\| \;\leq\; & \left\|\bar{\Omega}^{-1}\hat{\Omega}^k(\bar{\Omega}^{-1} - \tilde{\Omega}^{-1})\right\| + \left\|\hat{\Omega}^{-1}(\hat{\Omega}^k - \tilde{\Omega}^k)\tilde{\Omega}^{-1}\right\| \\ & + \left\|(\hat{\Omega}^{-1} - \tilde{\Omega}^{-1})\tilde{\Omega}^k\tilde{\Omega}^{-1}\right\| \xrightarrow{p} 0.\end{aligned}$$

Then by Assumption 8 it follows that

$$\mu_n^{-2}n\left|\hat{g}'\bar{\Omega}^{-1}\bar{\Omega}^k\hat{\Omega}^{-1}\hat{g}_{\delta_\ell} - \hat{g}'\tilde{\Omega}^{-1}\tilde{\Omega}^k\tilde{\Omega}^{-1}\hat{g}_{\delta_\ell}\right| \leq O_p(1)\left\|\bar{\Omega}^{-1}\bar{\Omega}^k\hat{\Omega}^{-1} - \tilde{\Omega}^{-1}\tilde{\Omega}^k\tilde{\Omega}^{-1}\right\| \xrightarrow{p} 0.$$

Therefore, we can replace $\bar{\Omega}$ and $\hat{\Omega}$ by $\tilde{\Omega}$ in the third term in eq. (1.9) without affecting its probability limit. Let $\tilde{Q}_{k,\ell}(\delta)$ denote the expression following the second equality in eq. (1.9), with $\tilde{\Omega}$ replacing $\bar{\Omega}$ and $\hat{\Omega}$ throughout. Then applying a similar argument to

the one just given to each of the six terms following the second equality in eq. (1.9), it follows by T that

$$\mu_n^{-2} n \left| \hat{Q}_{\delta_k \delta_\ell} \left(\bar{\delta} \right) - \tilde{Q}_{k,\ell}(\bar{\delta}) \right| \xrightarrow{p} 0.$$

Next, we will show that

$$\mu_n^{-2} n \left| \tilde{Q}_{k,\ell}(\bar{\delta}) - \tilde{Q}_{k,\ell}(0) \right| \xrightarrow{p} 0.$$

Working again with the third term, let $F(\delta) = \Omega(\delta)^{-1} \Omega^k(\delta) \Omega(\delta)^{-1}$. It follows from Assumptions 3 and 8 similarly to previous that for any a and b, $\left| a'[F(\bar{\delta}) - F(0)]b \right| \leq C \|a\| \|b\| \left\| \bar{\delta} \right\|$. Also, by Assumptions 3 and 7 we have $\mu_n^{-1} \sqrt{n} \left\| \hat{g}(\bar{\delta}) - \hat{g}(0) \right\| \xrightarrow{p} 0$ and $\mu_n^{-1} \sqrt{n} \left\| \hat{g}_{\delta_k}(\bar{\delta}) - \hat{g}_{\delta_k}(0) \right\| \xrightarrow{p} 0$. It then follows by CS and T that

$$\begin{aligned} & \mu_n^{-2} n \left| \hat{g}(\bar{\delta})' F(\bar{\delta}) \hat{g}_{\delta_k}(\bar{\delta}) - \hat{g}(0)' F(0) \hat{g}_{\delta_k}(0) \right| \\ \leq \; & \mu_n^{-2} n C (\left\| \hat{g}(\bar{\delta}) \right\| \left\| \hat{g}_{\delta_k}(\bar{\delta}) \right\| \left\| \bar{\delta} \right\| + \left\| \hat{g}(\bar{\delta}) - \hat{g}(0) \right\| \left\| \hat{g}_{\delta_k}(\bar{\delta}) \right\| \\ & + \|\hat{g}(0)\| \left\| \hat{g}_{\delta_k}(\bar{\delta}) - \hat{g}_{\delta_k}(0) \right\|) \xrightarrow{p} 0. \end{aligned}$$

Applying a similar argument for each of the other six term and using T gives $\mu_n^{-2} n \left| \tilde{Q}_{k,\ell}(\bar{\delta}) - \tilde{Q}_{k,\ell}(0) \right| \xrightarrow{p} 0$. It therefore suffices to show that $\mu_n^{-2} n \tilde{Q}_{k,\ell}(0) \xrightarrow{p} H_{k\ell}$.

Next, let $\Omega^k = \Omega^k(\beta_0)$, $\Omega^{k\ell} = \Omega^{k\ell}(\beta_0)$, $\Omega^{k,\ell} = \Omega^{k,\ell}(\beta_0)$, $\tilde{g} = \hat{g}(\beta_0)$, $\tilde{g}_{\delta_k} = \partial \hat{g}(0) / \partial \delta_k$, and $\tilde{g}_{\delta_k \delta_\ell} = \partial^2 \hat{g}(0) / \partial \delta_\ell \partial \delta_k$. Note that

$$\begin{aligned} \tilde{Q}_{k,\ell}(0) \;=\; & \tilde{g}'_{\delta_k} \Omega^{-1} \tilde{g}_{\delta_\ell} + \tilde{g}' \Omega^{-1} \tilde{g}_{\delta_k \delta_\ell} - \tilde{g}' \Omega^{-1} \left(\Omega^k + \Omega^{k\prime} \right) \Omega^{-1} \tilde{g}_{\delta_\ell} - \tilde{g}' \Omega^{-1} (\Omega^\ell + \Omega^{\ell\prime}) \Omega^{-1} \tilde{g}_{\delta_k} \\ & + \tilde{g}' \Omega^{-1} (\Omega^\ell + \Omega^{\ell\prime}) \Omega^{-1} \left(\Omega^{k\prime} + \Omega^k \right) \Omega^{-1} \tilde{g} - \tilde{g}' \Omega^{-1} \left(\Omega^{k\ell} + \Omega^{k,\ell} \right) \Omega^{-1} \tilde{g}. \end{aligned}$$

Consider once again the third term in $\tilde{Q}_{k,\ell}(0)$, that is $\tilde{g}' A \tilde{g}_{\delta_\ell}$ where $A = -\Omega^{-1} \left(\Omega^k + \Omega^{k\prime} \right) \Omega^{-1}$. Now apply Lemma A1 with $Y_i = g_i$, $Z_i = G_i S_n^{-1\prime} \mu_n e_k$, and $a_n = \mu_n^2$ to obtain

$$\mu_n^{-2} n \tilde{g}' A \tilde{g}_{\delta_\ell} = -tr(\Omega^{-1} \left(\Omega^k + \Omega^{k\prime} \right) \Omega^{-1} \Omega^{\ell\prime}) / \mu_n^2 + o_p(1).$$

Let $H_n = n S_n^{-1} G' \Omega^{-1} G S_n^{-1\prime}$. Then applying a similar argument to each term in $\tilde{Q}_{k,\ell}(0)$

gives

$$\begin{aligned}\mu_n^{-2} n\tilde{Q}_{k,\ell}(0) &= H_{nk,\ell} + \mu_n^{-2} tr[\Omega^{-1}\Omega^{k,\ell\prime} + \Omega^{-1}\Omega^{k\ell\prime} - \Omega^{-1}\left(\Omega^{k} + \Omega^{k\prime}\right)\Omega^{-1}\Omega^{\ell\prime} \\ &\quad -\Omega^{-1}(\Omega^{\ell} + \Omega^{\ell\prime})\Omega^{-1}\Omega^{k\prime} + \Omega^{-1}(\Omega^{\ell} + \Omega^{\ell\prime})\Omega^{-1}\left(\Omega^{k\prime} + \Omega^{k}\right) \\ &\quad -\Omega^{-1}\left(\Omega^{k\ell} + \Omega^{k,\ell}\right)] + o_p(1) \\ &= H_{nk,\ell} + \mu_n^{-2} tr[\Omega^{-1}(\Omega^{k,\ell\prime} - \Omega^{k,\ell}) + \Omega^{-1}(\Omega^{k\ell\prime} - \Omega^{k\ell}) \\ &\quad -\Omega^{-1}\left(\Omega^{k} + \Omega^{k\prime}\right)\Omega^{-1}\Omega^{\ell\prime} + \Omega^{-1}(\Omega^{\ell} + \Omega^{\ell\prime})\Omega^{-1}\Omega^{k}] + o_p(1).\end{aligned}$$

By $tr(AB) = tr(BA)$ for any conformable matrices A and B, we have

$$tr[(\Omega^{-1}\Omega^{\ell\prime})(\Omega^{-1}\Omega^{k})] = tr(\Omega^{-1}\Omega^{k}\Omega^{-1}\Omega^{\ell\prime})$$

Also, for a symmetric matrix A, $tr(AB) = tr(B'A) = tr(AB')$, so that

$$\begin{aligned} tr(\Omega^{-1}\Omega^{k,\ell\prime}) &= tr(\Omega^{-1}\Omega^{k,\ell}),\ tr(\Omega^{-1}\Omega^{k\ell\prime}) = tr(\Omega^{-1}\Omega^{k\ell}), \\ tr[\Omega^{-1}(\Omega^{\ell}\Omega^{-1}\Omega^{k})] &= tr[\Omega^{-1}(\Omega^{k\prime}\Omega^{-1}\Omega^{\ell\prime})]. \end{aligned}$$

Then we have $\mu_n^{-2} n\tilde{Q}_{k,\ell}(0) = H_{nk,\ell} + o_p(1)$, so that the conclusion follows by T. Q.E.D.

LEMMA A14: *If Assumptions 1-4 and 6-9 are satisfied then* $nS_n^{-1}\hat{D}(\hat{\beta})'\hat{\Omega}^{-1}\hat{D}(\hat{\beta})S_n^{-1\prime} \xrightarrow{p} H + \Lambda = HVH$.

Proof: For $\hat{g}_i = g_i(\hat{\beta})$, an expansion like those above gives $\rho_1(\hat{\lambda}'\hat{g}_i) = -1 - \hat{\lambda}'\hat{g}_i + \rho_3(\bar{v}_i)(\hat{\lambda}'\hat{g}_i)^2$, so that w.p.a.1

$$\frac{1}{n}\sum_i \rho_1(\hat{\lambda}'\hat{g}_i) = -1 - \hat{\lambda}'\hat{g} + r, |r| \leq C \max_i |\rho_3(\bar{v}_i)| \hat{\lambda}'\hat{\Omega}(\hat{\beta})\hat{\lambda} \leq C \left\|\hat{\lambda}\right\|^2.$$

By $\left\|\hat{\lambda}\right\| = O_p(\sqrt{m/n})$ and $\|\hat{g}\| = O_p(\sqrt{m/n})$ we have $\left|\hat{\lambda}'\hat{g}\right| = O_p(m/n) \xrightarrow{p} 0$. Also, $|r| = O_p(m/n) \xrightarrow{p} 0$, so that by T,

$$\frac{1}{n}\sum_i \rho_1(\hat{\lambda}'\hat{g}_i) \xrightarrow{p} -1. \tag{1.10}$$

Next, consider the expansion $\rho_1(\hat{\lambda}'\hat{g}_i) = -1 + \rho_2(\bar{v}_i)\hat{\lambda}'\hat{g}_i$ as in the proof of Lemma A13. As discussed there $\hat{\lambda}$ satisfies the first order condition $0 = \sum_i \rho_1(\hat{\lambda}'\hat{g}_i)\hat{g}_i/n = -\hat{g} - \bar{\Omega}\hat{\lambda}$

for $\bar{\Omega} = -\sum_i \rho_2(\bar{v}_i)\hat{g}_i\hat{g}_i'/n$, so that for $\hat{g}_{i\delta_k} = \partial\hat{g}_i(\hat{\delta})/\partial\delta_k$, $\hat{g}_{\delta_k} = \partial\hat{g}(\hat{\delta})/\partial\delta_k$, and $\bar{\Omega}^k = -\sum_i \rho_2(\bar{v}_i)\hat{g}_i'\hat{g}_{i\delta_k}/n$ we have

$$\hat{\lambda} = -\bar{\Omega}^{-1}\hat{g},\ \sum_i \rho_1(\hat{\lambda}'\hat{g}_i)\hat{g}_{i\delta_k}/n = -\hat{g}_{\delta_k} - \bar{\Omega}^{k\prime}\hat{\lambda} = -\hat{g}_{\delta_k} + \bar{\Omega}^{k\prime}\bar{\Omega}^{-1}\hat{g}.$$

Also, note that for $\bar{U} = \sum_{i=1}^n U_i/n$, we have $\bar{U}S_n^{-1\prime}e_k\mu_n = \tilde{g}_{\delta_k} - \bar{g}_{\delta_k} - \Omega^{k\prime}\Omega^{-1}\tilde{g}$. Then, in terms of the notation of Lemma A13, it follows similarly to the arguments given there that

$$\begin{aligned}
&\left[\frac{1}{n}\sum_i \rho_1(\hat{\lambda}'\hat{g}_i)\right]^2 e_k' n S_n^{-1}\hat{D}(\hat{\beta})'\hat{\Omega}^{-1}\hat{D}(\hat{\beta})S_n^{-1\prime}e_\ell \\
&= \mu_n^{-2} n(\hat{g}_{\delta_k}'\hat{\Omega}^{-1}\hat{g}_{\delta_\ell} - \hat{g}_{\delta_k}\hat{\Omega}^{-1}\bar{\Omega}^{\ell\prime}\bar{\Omega}^{-1}\hat{g} - \hat{g}'\bar{\Omega}^{-1}\bar{\Omega}^k\hat{\Omega}^{-1}\hat{g}_{\delta_\ell} + \hat{g}'\bar{\Omega}^{-1}\bar{\Omega}^k\hat{\Omega}^{-1}\bar{\Omega}^{\ell\prime}\bar{\Omega}^{-1}\hat{g}) \\
&= \mu_n^{-2} n(\tilde{g}_{\delta_k}'\Omega^{-1}\tilde{g}_{\delta_\ell} - \tilde{g}_{\delta_k}\Omega^{-1}\Omega^{\ell\prime}\Omega^{-1}\tilde{g} - \tilde{g}'\Omega^{-1}\Omega^k\Omega^{-1}\tilde{g}_{\delta_\ell} + \tilde{g}'\Omega^{-1}\Omega^k\Omega^{-1}\Omega^{\ell\prime}\Omega^{-1}\tilde{g}) + o_p(1) \\
&= \mu_n^{-2} n(\tilde{g}_{\delta_k} - \Omega^{k\prime}\Omega^{-1}\tilde{g})'\Omega^{-1}(\tilde{g}_{\delta_\ell} - \Omega^{\ell\prime}\Omega^{-1}\tilde{g}) + o_p(1) \\
&= n e_k' S_n^{-1}(G + \bar{U})'\Omega^{-1}(G + \bar{U})S_n^{-1\prime}e_\ell + o_p(1).
\end{aligned}$$

Note that by Assumption 1, $nS_n^{-1}G'\Omega^{-1}GS_n^{-1\prime} \longrightarrow H$. Also, $\xi_{\max}(E[U_iS_n^{-1\prime}e_\ell e_\ell' S_n^{-1}U_i']) \leq C/\mu_n^2$, so that

$$\begin{aligned}
E[(ne_k'S_n^{-1}G'\Omega^{-1}\bar{U}S_n^{-1\prime}e_\ell)^2] &= ne_k'S_n^{-1}G'\Omega^{-1}E[U_iS_n^{-1\prime}e_\ell e_\ell' S_n^{-1}U_i']\Omega^{-1}GS_n^{-1\prime}e_k \\
&\leq Cne_k'S_n^{-1}G'\Omega^{-2}GS_n^{-1\prime}e_k/\mu_n^2 \leq CH_{nkk}/\mu_n^2 \longrightarrow 0.
\end{aligned}$$

Now apply Lemma A1 to $ne_k'S_n^{-1}\bar{U}'\Omega^{-1}\bar{U}S_n^{-1\prime}e_\ell$, for $A = \Omega^{-1}$, $Y_i = U_iS_n^{-1\prime}e_k\mu_n$, $Z_i = U_iS_n^{-1\prime}e_\ell\mu_n$, and $\mu_n^2 = a_n$. Note that $\xi_{\max}(A'A) = \xi_{\max}(AA') = \xi_{\max}(\Omega^{-2}) \leq C$. Also, by $S_n^{-1}\mu_n$ bounded, $\xi_{\max}(\Sigma_{YY}) \leq \xi_{\max}(E[U_iU_i']) \leq C$ and $\xi_{\max}(\Sigma_{ZZ}) \leq C$. Furthermore, $m/a_n^2 = m/\mu_n^4 \longrightarrow 0$, $a_n/n = \mu_n^2/n \leq C$, $\mu_Y = \mu_Z = 0$, and

$$E[(Y_i'Y_i)^2]/na_n^2 \leq CE[\|U_i\|^4]/na_n^2 \leq CE[\|g_i\|^4 + \|G_i\|^4]/na_n^2 \longrightarrow 0.$$

Then by the conclusion of Lemma A1,

$$\begin{aligned}
ne_k'S_n^{-1}\bar{U}'\Omega^{-1}\bar{U}S_n^{-1\prime}e_\ell &= n\bar{Y}'A\bar{Z}/a_n = \mathrm{tr}\left(A\Sigma_{YZ}'\right)/a_n + o_p(1) \\
&= tr(\Omega^{-1}E[U_iS_n^{-1\prime}e_\ell e_k'S_n^{-1}U_i']) + o_p(1) \\
&= e_k'S_n^{-1}E[U_i'\Omega^{-1}U_i]S_n^{-1\prime}e_\ell + o_p(1) \xrightarrow{p} \Lambda_{k\ell}.
\end{aligned}$$

Then by T,

$$e_k' n S_n^{-1} \hat{D}(\hat{\beta})' \hat{\Omega}^{-1} \hat{D}(\hat{\beta}) S_n^{-1\prime} e_\ell \xrightarrow{p} H_{k\ell} + \Lambda_{k\ell}.$$

The conclusion then follows by applying this result for each k and ℓ. Q.E.D.

Proof of Theorem 3: Let $Y_n = n\mu_n^{-1} \partial \hat{Q}(0)/\partial \delta$. Then expanding the first-order conditions as outlined in Section 5 gives

$$0 = n\mu_n^{-1} \frac{\partial \hat{Q}(\hat{\delta})}{\partial \delta} = n\mu_n^{-1} \frac{\partial \hat{Q}(0)}{\partial \delta} + n\mu_n^{-2} \frac{\partial^2 \hat{Q}(\bar{\delta})}{\partial \delta \partial \delta'} \mu_n \hat{\delta}.$$

By Lemma 13 $n\mu_n^{-2} \partial^2 \hat{Q}(\bar{\delta})/\partial \delta \partial \delta'$ is nonsingular w.p.a.1. Then by CMT, Lemmas A12, A13, and S,

$$\mu_n \hat{\delta} = S_n'(\hat{\beta} - \beta_0) = \left[n\mu_n^{-2} \frac{\partial^2 \hat{Q}(\bar{\delta})}{\partial \delta \partial \delta'} \right]^{-1} n\mu_n^{-1} \frac{\partial \hat{Q}(0)}{\partial \delta} = H^{-1} Y_n + o_p(1).$$

Then by Lemma A12 and S,

$$S_n'(\hat{\beta} - \beta_0) \xrightarrow{d} H^{-1} N(0, H + \Lambda) = N(0, V).$$

Also, by Lemmas A13 and A14,

$$n S_n^{-1} \hat{H} S_n^{-1\prime} = \mu_n^{-2} n \frac{\partial^2 \hat{Q}(\hat{\delta})}{\partial \delta \partial \delta'} \xrightarrow{p} H, n S_n^{-1} \hat{D}' \hat{\Omega}^{-1} \hat{D} S_n^{-1\prime} \xrightarrow{p} HVH.$$

Also, $\hat{H}$ is nonsingular w.p.a.1, so that

$$S_n' V S_n / n = (n S_n^{-1} \hat{H} S_n^{-1\prime})^{-1} n S_n^{-1} \hat{D}' \hat{\Omega}^{-1} \hat{D} S_n^{-1\prime} (n S_n^{-1} \hat{H} S_n^{-1\prime})^{-1} \xrightarrow{p} H^{-1} HVHH^{-1} = V.$$

To prove the last conclusion, note that $r_n S_n^{-1} c \longrightarrow c^*$ and S imply that

$$\begin{aligned} r_n c'(\hat{\beta} - \beta_0) &= r_n c' S_n^{-1\prime} S_n'(\hat{\beta} - \beta_0) \xrightarrow{d} N(0, c^{*\prime} V c^*), \\ r_n^2 c' \hat{V} c / n &= r_n c' S_n^{-1\prime} (S_n' \hat{V} S_n / n) S_n^{-1} c r_n \xrightarrow{p} c^{*\prime} V c^*. \end{aligned}$$

Therefore by CMT and S,

$$\frac{c'(\hat{\beta} - \beta_0)}{\sqrt{c' \hat{V} c / n}} = \frac{r_n c' S_n^{-1\prime} S_n'(\hat{\beta} - \beta_0)}{\sqrt{r_n^2 c' S_n^{-1\prime} \left(S_n' \hat{V} S_n / n \right) S_n^{-1} c}} \xrightarrow{d} \frac{N(0, c^{*\prime} V c^*)}{\sqrt{c^{*\prime} V c^*}} = N(0, 1).$$

For the linear model we proceed by verifying all of the hypotheses of the general case. Note that $g_i(\beta) = Z_i(y_i - x_i'\beta)$ is twice continuously differentiable and that its first derivative does not depend on β, so Assumption 7 is satisfied. Also, by Lemma A5,

$$\begin{aligned}(E[\|g_i\|^4] + E[\|G_i\|^4])m/n &\leq CE[\|Z_i\|^4]m/n \longrightarrow 0, \\ \xi_{\max}(E[G_iG_i']) &\leq \sum_{j=1}^{p} \xi_{\max}(E[Z_iZ_i'x_{ij}^2]) \leq C\xi_{\max}(CI_m) \leq C,\end{aligned}$$

so that Assumption 6 is satisfied. Assumption 8 is satisfied by Lemmas A8 and A9. Assumptions 2 - 4 were shown to hold in the proof of Theorem 2. Assumption 9 can be shown to be satisfied similarly to the proof of Theorem 2. *Q.E.D.*

1.4 Large Sample Inference Proofs

The following result improves upon Theorem 6.2 of Donald, Imbens, and Newey (2003). Let $\tilde{g} = \hat{g}(\beta_0)$ by only requiring that $m/n \longrightarrow 0$ in the case where the elements of g_i are uniformly bounded.

LEMMA A15: *If* $E[(g_i'\Omega^{-1}g_i)^2]/mn \longrightarrow 0$ *then*

$$\frac{n\tilde{g}'\Omega^{-1}\tilde{g} - m}{\sqrt{2m}} \xrightarrow{d} N(0,1).$$

Proof: Note that $E[g_i'\Omega^{-1}g_i] = m$ so that by M,

$$\frac{\sum_{i=1}^n g_i'\Omega^{-1}g_i/n - m}{\sqrt{2m}} = O_p(\{E[\{g_i'\Omega^{-1}g_i\}^2]/nm\}^{1/2}) \xrightarrow{p} 0.$$

Now apply Lemma A9 with $Y_i = Z_i = \Omega^{-1/2}g_i/\sqrt{n}(2m)^{1/4}$, so that $\bar{\xi}_Z = \bar{\xi}_Y = n^{-1}(2m)^{-1/2}$. Note that $\Psi = \Sigma_{YY}\Sigma_{ZZ} + \Sigma_{YZ}^2 = 2I_m/n^2 2m = I_m/n^2 m$, so that $n^2 tr(\Psi) = n^2 tr(I_m/n^2 m) = 1$. Also note that

$$\begin{aligned}mn^4\bar{\xi}_Z^2\bar{\xi}_Y^2 &= m/4m^2 \longrightarrow 0, n^3(\bar{\xi}_Z^2 E[\|Y_i\|^4] + \bar{\xi}_Y^2 E[\|Z_i\|^4]) \\ &\leq n^3 2\{n^{-2}(2m)^{-1}E[\{g_i'\Omega^{-1}g_i\}^2/n^2 2m]\} \longrightarrow 0, \\ n^2 E[\|Y_i\|^4]E[\|Z_i\|^4] &= n^2\{E[\{g_i'\Omega^{-1}g_i\}^2]n^{-2}(2m)^{-1}\}^2 \longrightarrow 0.\end{aligned}$$

It then follows by Lemma A10 that $\sum_{i\neq j} g_i'\Omega^{-1}g_j/\sqrt{2m} \xrightarrow{d} N(0,1)$, so the conclusion follows by T. *Q.E.D.*

Proof of Theorem 4: By an expansion in λ around $\lambda = 0$ we have

$$\hat{Q}(\beta_0) = -\tilde{\lambda}'\tilde{g} - \tilde{\lambda}'\bar{\Omega}\tilde{\lambda}/2,$$

where $\bar{\Omega} = -\sum_i \rho_2(\bar{v}_i) g_i g_i'/n$, $\bar{v}_i = \bar{\xi}' g_i$, and $\left\|\bar{\xi}\right\| \leq \left\|\tilde{\lambda}\right\|$. Also, by an expansion around 0 we have $\rho_1(\tilde{\lambda}' g_i) = -1 + \rho_2(\breve{v}_i)\tilde{\lambda}' g_i$ with $|\breve{v}_i| \leq \left|\tilde{\lambda}' g_i\right|$, so that for $\breve{\Omega} = -\sum_i \rho_2(\breve{v}_i) g_i g_i'/n$ the first order conditions for $\tilde{\lambda}$ give $0 = -\tilde{g} - \breve{\Omega}\tilde{\lambda}$. Note that for $\Delta_n = n^{1/\gamma} \left(E\left[b_i^\gamma\right]\right)^{1/\gamma} \sqrt{m/n}$ we have

$$\max_{i \leq n} |1 + \rho_2(\breve{v}_i)| \leq C \left\|\tilde{\lambda}\right\| \max_{i \leq n} g_i = O_p(\Delta_n).$$

Let $\tilde{\Omega} = \sum_i g_i g_i'/n$. By Lemma A0 $\xi_{\max}(\tilde{\Omega}) \leq C$ w.p.a.1, so that for any a, b,

$$\begin{aligned} \left|a'\left(\breve{\Omega} - \tilde{\Omega}\right) b\right| &\leq \sum_i |1 + \rho(\breve{v}_i)| \left|a' g_i\right| \left|b' g_i\right| / n \\ &\leq O_p(\Delta_n) \sqrt{a'\tilde{\Omega} a b'\tilde{\Omega} b} = O_p(\Delta_n) \left\|a\right\| \left\|b\right\|. \end{aligned}$$

It follows similarly that

$$\left|a'\left(\breve{\Omega} - \tilde{\Omega}\right) b\right| \leq O_p(\Delta_n) \left\|a\right\| \left\|b\right\|.$$

It then follows from $\Delta_n \longrightarrow 0$, similarly to Lemma A0, that $\xi_{\min}(\breve{\Omega}) \geq C$ w.p.a.1., so $\tilde{\lambda} = -\breve{\Omega}^{-1}\tilde{g}$. Plugging into the above expansion gives

$$\hat{Q}(\beta_0) = \tilde{g}'\breve{\Omega}^{-1}\tilde{g} - \tilde{g}'\breve{\Omega}^{-1}\bar{\Omega}\breve{\Omega}^{-1}\tilde{g}/2.$$

As above $\xi_{\min}(\tilde{\Omega}) \geq C$ w.p.a.1, so that $\left\|\tilde{\Omega}^{-1}\tilde{g}\right\| \leq C \left\|\tilde{g}\right\| = O_p(\sqrt{m/n})$ and $\left\|\tilde{\Omega}^{-1}\tilde{g}\right\| = O_p(\sqrt{m/n})$. Therefore, by $\Delta_n\sqrt{m} \longrightarrow 0$,

$$\left|\tilde{g}'(\breve{\Omega}^{-1} - \tilde{\Omega}^{-1})\tilde{g}\right| = \left|\tilde{g}'\breve{\Omega}^{-1}(\tilde{\Omega} - \breve{\Omega})\tilde{\Omega}^{-1})\tilde{g}\right| \leq O_p(\Delta_n) O_p(m/n) = o_p(\sqrt{m}/n).$$

It follows similarly that $\left|\tilde{g}'(\breve{\Omega}^{-1}\bar{\Omega}\breve{\Omega}^{-1} - \tilde{\Omega}^{-1})\tilde{g}\right| = o_p(\sqrt{m}/n)$, so that by T,

$$\hat{Q}(\beta_0) = \tilde{g}'\tilde{\Omega}^{-1}\tilde{g}/2 + o_p(\sqrt{m}/n).$$

It follows by $mE[\|g_i\|^4]/n \longrightarrow 0$ that $\left\|\tilde{\Omega} - \Omega\right\| = o_p(1/\sqrt{m})$, so that $\tilde{g}'\tilde{\Omega}^{-1}\tilde{g} = \tilde{g}'\Omega^{-1}\tilde{g} + o_p(\sqrt{m}/n)$, and, by T,

$$\hat{Q}(\beta_0) = \tilde{g}'\Omega^{-1}\tilde{g}/2 + o_p(\sqrt{m}/n).$$

It then follows that

$$\frac{2n\hat{Q}(\beta_0)-m}{\sqrt{m}}-\frac{n\tilde{g}'\Omega^{-1}\tilde{g}-m}{\sqrt{m}}=\frac{2n}{\sqrt{m}}\left[\hat{Q}(\beta_0)-\tilde{g}'\Omega^{-1}\tilde{g}/2\right]=o_p(1).$$

Then by Lemma A15 and S we have

$$\frac{2n\hat{Q}(\beta_0)-m}{\sqrt{m}}\xrightarrow{d}N(0,1).$$

Also, by standard results for the chi-squared distribution, as $m\longrightarrow\infty$ the $1-\alpha^{th}$ quantile q_α^m of a $\chi^2(m)$ distribution has the property that $[q_\alpha^m-m]/\sqrt{2m}$ converges to the $1-\alpha^{th}$ quantile q_α of $N(0,1)$. Hence we have

$$\Pr(2n\hat{Q}(\beta_0)\geq q_\alpha^m)=\Pr\left(\frac{2n\hat{Q}(\beta_0)-m}{\sqrt{2m}}\geq\frac{q_\alpha^m-m}{\sqrt{2m}}\right)\longrightarrow\alpha.\ Q.E.D.$$

Proof of Theorem 5: Let $\hat{B}=nS_n^{-1}\hat{D}(\beta_0)'\hat{\Omega}(\beta_0)^{-1}\hat{D}(\beta_0)S_n^{-1\prime}$ and $B=HVH$. It follows from Lemma A14, replacing $\hat{\beta}$ with β_0, that $\hat{B}\xrightarrow{p}B$. By the proof of Theorem 3, S, and CM we have

$$\hat{T}=(\hat{\beta}-\beta_0)'S_n(S_n'\hat{V}S_n/n)^{-1}S_n'(\hat{\beta}-\beta_0)=Y_n'B^{-1}Y_n+o_p(1)\xrightarrow{d}\chi^2(p).$$

Then by Lemma A12

$$LM(\beta_0)=n\frac{\partial\hat{Q}(\beta_0)}{\partial\beta}'S_n^{-1\prime}\hat{B}^{-1}nS_n^{-1}\frac{\partial\hat{Q}(\beta_0)}{\partial\beta}=Y_n'(B+o_p(1))^{-1}Y_n=Y_n'B^{-1}Y_n+o_p(1).$$

Therefore we have $LM(\beta_0)=\hat{T}+o_p(1)$.

Next, by an expansion, for $\bar{H}=nS_n^{-1}\partial^2\hat{Q}(\bar{\beta})/\partial\beta\partial\beta'S_n^{-1\prime}$,

$$\begin{aligned}2n[\hat{Q}(\beta_0)-\hat{Q}(\hat{\beta})]&=n(\hat{\beta}-\beta_0)'[\partial^2\hat{Q}(\bar{\beta})/\partial\beta\partial\beta'](\hat{\beta}-\beta_0)\\&=(\hat{\beta}-\beta_0)'S_n\bar{H}S_n'(\hat{\beta}-\beta_0),\end{aligned}$$

where $\bar{\beta}$ lies on the line joining $\hat{\beta}$ and β_0 and $\bar{H}\xrightarrow{p}H$ by Lemma A13. Then by the proof of Theorem 3 and the CMT,

$$\begin{aligned}2n[\hat{Q}(\beta_0)-\hat{Q}(\hat{\beta})]&=\{Y_n'H^{-1}+o_p(1)\}\{H+o_p(1)\}\{H^{-1}Y_n+o_p(1)\}\\&=Y_n'H^{-1}Y_n+o_p(1).\end{aligned}$$

It follows that $2n[\hat{Q}(\beta_0) - \hat{Q}(\hat{\beta})] = O_p(1)$, so that

$$2n[\hat{Q}(\beta_0) - \hat{Q}(\hat{\beta})]/\sqrt{m-p} \xrightarrow{p} 0.$$

Therefore, it follows as in the proof of Theorem 4 that

$$\begin{aligned}\frac{2n\hat{Q}(\hat{\beta}) - (m-p)}{\sqrt{m-p}} &= \frac{2n\hat{Q}(\beta_0) - (m-p)}{\sqrt{m-p}} + o_p(1) \\ &= \sqrt{\frac{m}{m-p}}\frac{2n\hat{Q}(\beta_0) - m}{\sqrt{m}} + \frac{p}{\sqrt{m-p}} + o_p(1) \xrightarrow{d} N(0,1).\end{aligned}$$

Next, note that $H^{-1} \leq V$ in the p.s.d. sense so that $V^{-1} \leq H$. It follows that

$$Y_n' H^{-1} Y_n \geq Y_n' B^{-1} Y_n \xrightarrow{d} \chi^2(p).$$

Then $\Pr(\ 2n[\hat{Q}(\beta_0) - \hat{Q}(\hat{\beta})] > q_\alpha^p) = \Pr(Y_n' H^{-1} Y_n > q_\alpha^p) + o(1) \geq \alpha$.

Next, in considering the CLR test, for notational convenience evaluate at β_0 and drop the β argument, e.g. so that $\hat{R} = \hat{R}(\beta_0)$. By have $\hat{B} \xrightarrow{p} B$ it follows that $\hat{B} \geq (1-\varepsilon)B$ w.p.a.1 for all for $\varepsilon > 0$. Also by m/μ_n^2 bounded, for any C there is ε small enough so that $(1-\varepsilon)C - \varepsilon m/\mu_n^2$ is positive and bounded away from zero, i.e. so that $(1-\varepsilon)C - \varepsilon m/\mu_n^2 \geq C$ (the C's are different). Then by hypothesis and multiplying through by $1-\varepsilon$ and subtracting $\varepsilon m/\mu_n^2$ from both sides it will be the case that

$$\xi_{\min}(\mu_n^{-2} S_n (1-\varepsilon) B S_n') - \left(m/\mu_n^2\right) \geq (1-\varepsilon)C - \varepsilon m/\mu_n^2 \geq C.$$

Then, w.p.a.1

$$\hat{F} = \frac{\hat{R} - m}{\mu_n^2} = \xi_{\min}(\mu_n^{-2} S_n \hat{B} S_n') - m/\mu_n^2 \geq \xi_{\min}(\mu_n^{-2} S_n (1-\varepsilon) B S_n') - \left(m/\mu_n^2\right) \geq C.$$

Also, by the proof of theorem 4,

$$\frac{AR - m}{\mu_n^2} = \frac{\sqrt{m}}{\mu_n^2}\frac{AR - m}{\sqrt{m}} \xrightarrow{p} 0.$$

Therefore we have, w.p.a.1,

$$\frac{AR - \hat{R}}{\mu_n^2} = \frac{AR - m}{\mu_n^2} - \hat{F} \leq -C.$$

It follows that w.p.a.1,

$$\frac{AR}{\hat{R}} = \frac{(AR-m)/\mu_n^2 + m/\mu_n^2}{\hat{F} + m/\mu_n^2} \leq \frac{C/2 + m/\mu_n^2}{C + m/\mu_n^2} \leq 1 - C.$$

Therefore by $\hat{R} \geq C\mu_n^2 + m \longrightarrow \infty$, w.p.a.1,

$$\frac{\hat{R}}{\left(AR - \hat{R}\right)^2} = \frac{1}{\hat{R}} \frac{1}{\left(1 - AR/\hat{R}\right)^2} \xrightarrow{p} 0.$$

Note that $AR - \hat{R} < 0$ w.p.a.1, so that $\left|AR - \hat{R}\right| = \hat{R} - AR$. Also, similarly to Andrews and Stock (2006), by a mean value expansion $\sqrt{1+x} = 1 + (1/2)\left(x + o(1)\right)$, so that

$$\begin{aligned}
C\hat{L}R &= \frac{1}{2}\left\{AR - \hat{R} + \left[\left(AR - \hat{R}\right)^2 + 4LM \cdot \hat{R}\right]^{1/2}\right\} \\
&= \frac{1}{2}\left\{AR - \hat{R} + \left|AR - \hat{R}\right|\left[1 + \frac{4LM \cdot \hat{R}}{\left(AR - \hat{R}\right)^2}\right]^{1/2}\right\} \\
&= \frac{1}{2}\left\{AR - \hat{R} + \left|AR - \hat{R}\right|\left[1 + 2LM\frac{\hat{R}}{\left(AR - \hat{R}\right)^2}\left(1 + o_p(1)\right)\right]\right\} \\
&= L\hat{M}\frac{\hat{R}}{\hat{R} - AR}(1 + o_p(1)).
\end{aligned}$$

Let $r_n = \xi_{\min}(S_n B S_n'/\mu_n^2)$. Then $r_n - m/\mu_n^2 \geq C$ by hypothesis. Then $\hat{R}/\mu_n^2 = r_n + o_p(1)$, as shown below. It then follows that

$$\frac{\hat{R}}{\hat{R} - AR} = \frac{\hat{R}/\mu_n^2}{(\hat{R} - m)/\mu_n^2 - (AR - m)/\mu_n^2} = \frac{r_n + o_p(1)}{r_n - m/\mu_n^2 + o_p(1)} = \frac{r_n}{r_n - m/\mu_n^2} + o_p(1).$$

It then follows that

$$CLR = \left(\frac{r_n}{r_n - m/\mu_n^2}\right) LM + o_p(1).$$

Carrying out these same arguments with $q_s^{m-p} + q_s^p$ replacing AR it follows that

$$\begin{aligned}
\hat{q}_s &= \frac{1}{2}\left\{q_s^{m-p} + q_s^p - \hat{R} + \left[\left(q_s^{m-p} + q_s^p - \hat{R}\right)^2 + 4q_s^p \cdot \hat{R}\right]^{1/2}\right\} \\
&= \left(\frac{r_n}{r_n - m/\mu_n^2}\right) q_s^p + o_p(1),
\end{aligned}$$

giving the conclusion with $c_n = r_n/(r_n - m/\mu_n^2)$.

It now remains to show that $\hat{R}/\mu_n^2 = r_n + o_p(1)$. Note that for $\bar{S}_n = S_n/\mu_n$,

$$\hat{R}/\mu_n^2 = \min_{\|x\|=1} x'\bar{S}_n\hat{B}\bar{S}_n'x, r_n = \min_{\|x\|=1} x'\bar{S}_nB\bar{S}_n'x.$$

By Assumption 1 we can assume without loss of generality that $\mu_n = \mu_{1n}$ and

$$\bar{S}_n = \tilde{S}_n diag(1, \mu_{2n}/\mu_n, ..., \mu_{pn}/\mu_n).$$

Let e_j denote the j^{th} unit vector and consider x_n such that $x_n'S_ne_j = 0, (j = 2, ..., p)$, and $\|x_n\| = 1$. Then by $\tilde{S}_n$ bounded and CS,

$$\left\|x_n'\bar{S}_n\right\| = \left\|x_n'\tilde{S}_n\left[e_1 + \sum_{j=2}^{p}(\mu_{jn}/\mu_n)e_j\right]\right\| = \left\|x_n'\tilde{S}_ne_1\right\| \leq \left\|\tilde{S}_n\right\| \leq C.$$

Also, by $\hat{B} \xrightarrow{p} B$ there is C such $\left\|\hat{B}\right\| \leq C$ and $\xi_{\min}(\hat{B}) \geq 1/C$. w.p.a.1. Let $\hat{x} = \arg\min_{\|x\|=1} x'\bar{S}_n\hat{B}\bar{S}_n'x$ and $x_n^* = \arg\min_{\|x\|=1} x'\bar{S}_nB\bar{S}_n'x$. Then w.p.a.1,

$$C^{-1}\left\|\hat{x}'\bar{S}_n\right\|^2 \leq \hat{R}/\mu_n^2 \leq x_n'\bar{S}_n\hat{B}\bar{S}_n'x_n \leq C, C^{-1}\left\|x_n^{*\prime}\bar{S}_n\right\|^2 \leq r_n \leq x_n'\bar{S}_nB\bar{S}_n'x_n \leq C,$$

so that there is $\bar{C}$ such that w.p.a.1,

$$\left\|\hat{x}'\bar{S}_n\right\| \leq \bar{C}, \left\|x_n^{*\prime}\bar{S}_n\right\| \leq \bar{C}.$$

Consider any $\varepsilon > 0$. By $\hat{B} \xrightarrow{p} B$, w.p.a.1 $\left\|\hat{B} - B\right\| \leq \varepsilon/\bar{C}^2$. Then, w.p.a.1,

$$\begin{aligned}
\hat{R}/\mu_n^2 &\leq x_n^{*\prime}\bar{S}_n\hat{B}\bar{S}_nx_n^* = r_n + x_n^{*\prime}\bar{S}_n(\hat{B} - B)\bar{S}_nx_n^* \leq r_n + \left|x_n^{*\prime}\bar{S}_n(\hat{B} - B)\bar{S}_nx_n^*\right| \\
&\leq r_n + \left\|x_n^{*\prime}\bar{S}_n\right\|^2\left\|\hat{B} - B\right\| \leq r_n + \bar{C}^2(\varepsilon/\bar{C}^2) = r_n + \varepsilon. \\
r_n &\leq \hat{x}'\bar{S}_nB\bar{S}_n\hat{x} = \hat{R}/\mu_n^2 + \hat{x}'\bar{S}_n(B - \hat{B})\bar{S}_n\hat{x} \leq \hat{R}/\mu_n^2 + \varepsilon.
\end{aligned}$$

Thus, w.p.a.1, $r_n - \hat{R}/\mu_n^2 \leq \varepsilon$ and $\hat{R}/\mu_n^2 - r_n \leq \varepsilon$, implying $\left|\hat{R}/\mu_n^2 - r_n\right| \leq \varepsilon$, showing $\left|\hat{R}/\mu_n^2 - r_n\right| \xrightarrow{p} 0$. Q.E.D.

References

Andrews, D.W.K. and J.H. Stock (2006): "Inference with Weak Instruments," in Blundell, R., W. Newey, T. Persson eds., *Advances in Economics and Econometrics, Vol. 3.*

Donald, S.G., G.W. Imbens, and W.K. Newey (2003): "Empirical Likelihood Estimation and Consistent Tests With Conditional Moment Restrictions," *Journal of Econometrics* 117, 55-93.

Newey, W.K. (1991): "Uniform Convergence in Probability and Stochastic Equicontinuity," *Econometrica* 59, 1161-1167.