Supplemental appendix for Identification strength with a large number of moments

Hyojin Han* Eric Renault†

April 2, 2020

Abstract

This paper studies how identification is affected in GMM estimation as the number of moment conditions increases. We develop a general asymptotic theory extending the set up of of Chao and Swanson (2005) and Antoine and Renault (2009, 2012) to the case where moment conditions have heterogeneous identification strengths and the number of them may diverge to infinity with the sample size. We also allow the models to be locally misspecified and examine how the asymptotic theory is affected by the degree of misspecification. The theory encompasses many cases including GMM models with many moments (Han and Phillips (2006)), partially linear models, and local GMM via kernel smoothing with a large number of conditional moment restrictions. We provide an understanding of the benefits of a large number of moments that compensate the weakness of individual moments by explicitly showing how an increasing number of moments improves the rate of convergence in GMM.

JEL classification: C01, C13, C14

Keywords: Generalized Method of Moments, Weak Identification, Alternative Asymptotic Theory

*Gregory and Paula Chow Center for Economic Research (GCCER) and Department of Statistics, School of Economics, Xiamen University, hyojin_han@xmu.edu.cn

†Department of Economics, University of Warwick, Eric.Renault@warwick.ac.uk

Appendix A. Three motivating examples (Continued)

In this section, we discuss the asymptotic properties of GMM estimators in the three motivating examples continued from section 2.

Example 1 (*GMM with Many Moment Conditions and Diagonal* Λ_T)

Under the setup of a single convergence rate,

$$\tilde{\Lambda}_T = \lambda_T \sqrt{J_T} I_p$$

and R^0 can be thought as a $p \times p$ identity matrix. The role of many moments that improves the rate of convergence of a GMM estimator is also implicitly shown in HP. More specifically, they show that under some conditions, the rate of convergence is $\sqrt{Tc_T}$:

$$\sqrt{Tc_T}(\hat{\theta}_T - \theta^0) = O_p(1).$$

From the comparison we made in subsection 2.1, we can see that

$$\sqrt{Tc_T} = \sqrt{T}\sqrt{\frac{J_T \lambda_T^2}{T}} = \lambda_T \sqrt{J_T}$$

HP imposes a type of a Lipschitz condition that is on par with our assumption 4(ii) for such a rate of convergence (see proposition 10 in HP). Borrowing the notations in HP, define[1]

$$\begin{aligned}
f_T(\theta) &= c_T^{-1} \bar{\psi}_T(\theta)' \bar{\psi}_T(\theta) = \frac{T}{\lambda_T^2 J_T} \bar{\psi}_T(\theta)' \bar{\psi}_T(\theta) \\
\bar{f}_T(\theta) &= c_T^{-1} E\left[\bar{\psi}_T(\theta)\right]' E\left[\bar{\psi}_T(\theta)\right] = c_T^{-1} \frac{\lambda_T^2}{T} \rho_T(\theta)' \rho_T(\theta) = \frac{\rho_T(\theta)' \rho_T(\theta)}{J_T} \\
W_T(\theta) &= [f_T(\theta) - \bar{f}_T(\theta)] - [f_T(\theta^0) - \bar{f}_T(\theta^0)].
\end{aligned}$$

This leads us to see that our assumption 4(i) regards the limit of the normalized objective function $\bar{f}_T$ and our assumption 4(ii) is a type of a Lipschitz condition on the function W_T as in HP.

Another key condition of HP for asymptotic normality with zero mean of a GMM estimator is (see condition (ii) in corollary 14 in HP)

$$\lim_{T\to\infty} \frac{J_T}{\sqrt{Tc_T}} = \lim_{T\to\infty} \frac{\sqrt{J_T}}{\lambda_T} = 0.$$

which is also imposed in our model by assumption 9(i) in order to remove the correlation between

[1] Recall that $c_T = \frac{J_T \lambda_T^2}{T}$.

the moments and the Jacobian.

Newey and Windmeijer (2009) extend HP by allowing the heterogeneity of convergence rates of the GMM estimator although their focus is to provide correct standard errors. They do not explicitly discuss reparameterization in general cases but they also allow for linear combinations of θ to have different strength of identification as in our case. The different rates of convergence are summarized in a matrix that they denote S_T (see their assumption 1) such that

$$S_T = \tilde{S}_T diag(\mu_{1T}, \cdots, \mu_{pT}), \quad \frac{\mu_{i,T}}{\sqrt{T}} \to c \in [0, \infty[, \quad \underset{i}{min}\ \mu_{i,T} \to \infty$$

and $S_T'(\hat{\theta}_T - \theta^0)$ is asymptotically normal. This is conformable with our asymptotic normality result with $diag(\mu_{1T}, \cdots, \mu_{pT}) = \tilde{\Lambda}_T$.

Example 2 (*Local GMM via Kernel Smoothing*)

For the sake of expositional simplicity, we assume as in section 2 that

$$\begin{aligned} J_T &= J_{(1)} + J_{(2)T} \\ \lambda_{j,T} &= \sqrt{T},\ \forall j = 1, 2, ..J_{(1)} \\ \lambda_{j,T} &= \sqrt{Th_T^d},\ \forall j = J_{(1)} + 1, J_{(1)} + 2, .., J_{(1)} + J_{(2)T}, \end{aligned}$$

but also that $J_{(1)}$ is finite and fixed, so that

$$O(J_{(2)T}) = O(J_T).$$

We also impose more conditions on the bandwidth. From the development of asymptotic theory of the GMM estimator, we can deduce that $\sqrt{T}$ and $\sqrt{Th_T^d J_T}$ are the rates of convergence of the local GMM/kernel smoothing estimator. The below bandwidth rates assumption(i) states that $\sqrt{Th_T^d J_T}$ grows slower than $\sqrt{T}$, which is the largest rate of convergence:

$$\sqrt{Th_T^d J_T} = o(\sqrt{T}) \quad \Leftrightarrow \quad h_T^d J_T \to 0.$$

In addition, we need an orthogonality condition (see section 3.2) in order to disentangle the slower rate of convergence from the faster rate. The second part of the below assumption serves this purpose.

$$\sqrt{T} = o(Th_T^d J_T) \quad \Leftrightarrow \quad T^{1/2} h_T^d J_T \to \infty.$$

When $J_T < \infty$, the third condition in the below assumption imposes $Th_T^{d+4} \to c \in [0, \infty)$, which is

a standard condition in the kernel smoothing literature. We need a stronger condition in our setup since it is the rate of asymptotic bias that also increases with J_T. When $c = 0$, the asymptotic bias disappears asymptotically while when $c > 0$, we have the optimal rate of convergence but with non negligible asymptotic bias. The last part of the bandwidth rates assumption is another price to pay for having a large number of moments which corresponds to the rate of bias that arises from adding many moments.

Assumption (Bandwidth Rates): As $T \to \infty$,

$$\begin{aligned} (i)\ h_T^d J_T &\to 0 \\ (ii)\ \sqrt{T} h_T^d J_T &\to \infty \\ (iii)\ T h_T^{d+4} J_T &\to c \in [0, \infty) \\ (iv)\ J_T / T h_T^d &\to 0. \end{aligned}$$

As already discussed in section 3, when J_T is diverging, we need to account for the rate of J_T in order to properly derive the rate of convergence of the GMM estimator. Then, in order to ensure the improved rate of convergence with large J_T, we need to impose the same high-level condition in assumption 4 since this is a nonlinear model. Note that we have two sets of moments in this example so l, which denotes the number of different rates of convergence, is 2 with $\bar{\psi}_{2,T}(\theta) = \sqrt{h_T^d} \hat{E}_T \left[g_2(Y, \theta) | Z = z_0 \right]$ being the sample counterpart of the moments with the weaker rate. Instead of restating the assumption, we will assume that assumption 4 holds for this example with $J_{l,T} = J_T$, $\lambda_{l,T} = \sqrt{T h_T^d}$, and $\rho_{lT}(\theta) = \rho_{2T}(\theta) = E\left[\hat{E}_T \left[g_2(Y, \theta) | Z = z_0 \right] \right]$.

We also assume that

$$rank\left(E\left[\frac{\partial g_1(Z, Y, \theta^0)}{\partial \theta'} \right] \right) = s_1 < p$$

so θ^0 is not identified using $E\left[g_1(Z, Y, \theta) \right]$ only. From assumption(i) above, the rate of convergence ($\sqrt{T}$) of the first set of moments is stronger than that of the second set and we want to separate these so that the s_1 directions of the parameter space can be estimated with the faster rate of convergence. The reparameterization method introduced in section 3.2 serves this purpose with $R^0 = [R_1 \ R_2]$ where R_1 and R_2, are $p \times s_1$ and $p \times (p - s_1)$ matrices, respectively, with $rank[R_1] = s_1$ and $rank[R_2] = p - s_1$ such that

$$E\left\{ \hat{E}_T \left[\frac{\partial g_2(Y, \theta^0)}{\partial \theta'} | Z = z_0 \right] \right\} R_2 = 0.$$

Comparison with our general framework

As stated above, we have two rates of convergence such that

$$\lambda_{1,T}\sqrt{J_{1,T}} = \sqrt{T}, \quad \lambda_{2,T}\sqrt{J_{2,T}} = \sqrt{Th_T^d J_T}.$$

Thus,

$$\tilde{\Lambda}_T = \begin{pmatrix} \sqrt{T}Id_{J_{(1)}} & 0 \\ 0 & \sqrt{Th_T^d J_T}Id_{(J_T - J_{(1)})} \end{pmatrix}.$$

Then assumption 3 holds by the above assumption of bandwidth rates.

Our assumption 5 would hold if

$$\sqrt{Th_T^d}\left\| E\left[\hat{E}_T[g_2(Y,\theta^0)|Z=z_0]\right]\right\| = O(1)$$

and this is what assumption(iii) above implies.

We already maintain the local identification condition in section 2.2 (local identification assumption). We now let $\Gamma_1(\theta^0) = E\left[\frac{\partial g_1(Z,Y,\theta^0)}{\partial\theta'}\right]$ and $\Gamma_{2,T}(\theta^0) = E\left\{\hat{E}_T\left[\frac{\partial g_2(Y,\theta^0)}{\partial\theta'}|Z=z_0\right]\right\}$ and impose:

(i) The eigenvalues of $\Gamma_1'(\theta^0)\Gamma_1(\theta^0)$ are bounded away from zero and infinity, and

(ii) $\underset{T\to\infty}{lim}\ \frac{1}{J_T}\Gamma_{2,T}'(\theta^0)\Gamma_{2,T}(\theta^0)$ exists and its eigenvalues are bounded away from zero and infinity, so that assumption 6(iii) holds. Then we can deduce, using assumption 10, D_T^0 is a block diagonal matrix with the first $J_{(1)} \times s_1$ block is

$$E\left[\frac{\partial g_1(Z,Y,\theta^0)}{\partial\theta'}\right]R_1$$

and the second block ,which is $(J_T - J_{(1)}) \times (p - s_1)$, is

$$\frac{1}{J_T}E\left\{\hat{E}_T\left[\frac{\partial g_2(Y,\theta^0)}{\partial\theta'}|Z=z_0\right]\right\}R_2.$$

Lemma A.1 in Gagliardini et al. (2011) shows that $E\left[\Phi_T(\theta^0)\Phi_T(\theta^0)'\right]$ is a block-diagonal matrix such that

$$\begin{aligned} S_T &= E\left[\Phi_T(\theta^0)\Phi_T(\theta^0)'\right] \\ &= \begin{pmatrix} Var\left[g_1(Z_t,Y_t,\theta^0)\right] & 0 \\ 0 & \omega_2 Var\left[g_2(Y_t,\theta^0)|Z=z_0\right]/f_Z(z_0) \end{pmatrix} \end{aligned}$$

where $w^0 = \int_{\mathbb{R}^d} K(u)^2\ du$ which is assumed to be finite with some choice of the kernel $K(\cdot)$ and $f_Z(z_0)$ is the density function of Z. Then assumption 11 would maintain if the eigenvalues of S_T

are bounded away from zero and infinity under the standard conditions on the choice of kernel. As a result, we have the asymptotic normality result for the kernel smoothing estimator.

Example 3 (*Near Multicollinearity*)

In this section, we continue the analysis of PLM from our subsection 2.4. We can derive the asymptotic distribution of the GMM estimator using the first order closed-form expression

$$\hat{\mu}_T - \mu_T^0 = \left[\frac{\partial \bar{\psi}_T(\hat{\mu}_T)'}{\partial \mu_T} W_T \frac{\partial \bar{\psi}_T(\hat{\mu}_T)}{\partial \mu_T'}\right]^{-1} \frac{\partial \bar{\psi}_T(\hat{\mu}_T)'}{\partial \mu_T} W_T \bar{\psi}_T(\mu_T^0).$$

where $\frac{\partial \bar{\psi}_T(\mu)}{\partial \mu'}$ does not depend on μ since this is a linear model. From $\bar{\psi}_T(\mu)$ derived in section 2.4, we can see that

$$\frac{\partial \bar{\psi}_T(\hat{\mu}_T)}{\partial \mu_T'} = \frac{1}{T}\sum_{t=1}^{T} z_{t,T} z_{t,T}' \Pi_T \frac{\bar{\Lambda}_T}{\sqrt{T}} + \frac{1}{T}\sum_{t=1}^{T} z_{t,T} \varepsilon_{t,T}' \frac{\bar{\Lambda}_T}{\sqrt{T}}$$

$$W_T^{-1} = \frac{1}{T}\sum_{t=1}^{T} z_{t,T} z_{t,T}'.$$

Assumption:

(i) There exists a sequence of positive definite matrices Υ_T such that

$$\left\| \frac{\Pi_T' Z' Z \Pi_T}{T} - \Upsilon_T \right\| \xrightarrow{p} 0.$$

(ii) The errors $\xi_{t,T} = (u_t, \varepsilon_{t,T}')'$ are i.i.d. with a mean zero and positive definite variance matrix $\Sigma_T = \begin{pmatrix} \sigma_{uu} & \Sigma_{\varepsilon u}' \\ \Sigma_{\varepsilon u} & \Sigma_{\varepsilon\varepsilon} \end{pmatrix}$. Also, there exists a positive constant $C < \infty$ such that $E\left[\|\varepsilon_{t,T}\|^4 | z_{t,T}\right] < C$.

(iii) As $T \to \infty$,

$$\frac{J_T^2}{T} \to 0.$$

(iv) $\text{rank}(P_{ZW}) = K_T$ (a.s.) and there is a $C > 0$ such that $M_{tt} \geq C$.

(v) As $T \to \infty$,

$$\sum_{t=1}^{T} \|\Pi_{1T}' z_{t,T}\|^4 / T^2 \to 0.$$

The first part of the above assumption shows that the concentration parameter $\Pi_T' Z' Z \Pi_T$ grows at the rate of the sample size T since we do not have weak instruments. The second part is a usual assumption made in the instrumental variable literature and the third part is imposed to remove the asymptotic bias from having a large number of instruments. With these conditions, we can see

that

$$\left\| T\bar{\Lambda}_T^{-1} \frac{\partial \bar{\psi}_T(\hat{\theta}_T)'}{\partial \mu_T} W_T \frac{\partial \bar{\psi}_T(\hat{\theta}_T)}{\partial \mu_T'} \bar{\Lambda}_T^{-1} - \Upsilon_T \right\| = o_p(1)$$

since

$$\frac{1}{T}\sum_{t=1}^{T} \varepsilon_{t,T} z_{t,T}' \left(\frac{1}{T}\sum_{t=1}^{T} z_{t,T} z_{t,T}' \right)^{-1} \frac{1}{T}\sum_{t=1}^{T} z_{t,T}\varepsilon_{t,T}' = \frac{J_T}{T}\left(\Sigma_{\varepsilon\varepsilon} + O_p(\sqrt{J_T}) \right) \xrightarrow{p} 0.$$

where $\Sigma_{\varepsilon\varepsilon} = O(1)$ using lemma A1 in CS. The fourth part is the assumption PLM (d) in Cattaneo et al. (2018). While the assumption PLM (d) allows K_T/T to converge to some nonzero constant, we already imposed a stricter restriction in the third part of the above assumption that $K_T^2/T \to 0$ ($K_T + p \leq J_T$). The reason is that x and w are endogenous in our setup and we bring an additional bias that comes from having a large number of instruments. By imposing a stricter restriction on the rate of K_T, we do not have to worry about a larger variance arising from a large K_T. The last part of the above assumption allows us to apply central limit theorem to derive our asymptotic normality result as shown below.

Comparing this with assumption 10 in our main framework, we can see that R^0, which is designed to disentangle various rates of convergence, is treated as an identity matrix in this example. As stated in AR12, only the column spaces of the Jacobian $\partial\bar{\psi}_T(\theta)'/\partial\mu_T$ matter. But, in this example, it is the column spaces that have separate rates of convergence from each other rather than being a linear combination of the terms involving all different rates because $\bar{\Lambda}_T$ right-multiplies the moments[2]. Thus, Υ_T is not and does not have to be a block-diagonal matrix.

We now impose a stricter restriction on the rate at which misspecification disappears and it is conformable with assumption 5 (see $\Lambda_T \rho_T(\mu_T^0)$ in section 2.4) by letting

$$\left\| E\left[z_{t,T} \left\{ g^0(w_t) - p_{K_T}'(w_t)\beta_T^0 \right\} \right] \right\| = O\left(\frac{1}{\sqrt{T}} \right). \tag{A.1}$$

Asymptotic normality of $\hat{\theta}_T$

As stated in section 2.4, our interest is θ, the first $p < \infty$ components of μ_T[3]. To present the asymptotic distribution of $\hat{\theta}_T$, we start with introducing some notations.

Let $X = (x_1, \cdots, x_T)'$, $P^{K_T} = (p_{K_T}(w_1), \cdots, p_{K_T}(w_T))'$, $u_T = (u_1, \cdots, u_T)'$, $\varepsilon_T = (\varepsilon_{1,T}, \cdots, \varepsilon_{T,T})'$, and $e_T = (e_1, \cdots, e_T)'$ where $e_t = g^0(w_t) - p_{K_T}'(w_t)\beta_T^0$ for $t = 1, 2, \cdots, T$. Also, let $P_Z = Z(Z'Z)^{-1}Z'$, $P_{ZW} = P_Z P^{K_T}$, and $M_{P_{ZW}} = Id - P_{ZW}$ where Id is an identity matrix. $M_{t,s}$

[2]As pointed out in section 2.4, the weakness of moments does not arise from weak instruments (i.e. local-to-zero first stage coefficients) but from the near multicollinearity.

[3]Recall that μ_T includes a growing number of nuisance parameters.

denotes the (t,s)-th element of $M_{P_{ZW}}$. In addition, $\Pi_T = [\Pi_{1T}\ \Pi_{2T}]$ and $\varepsilon_T = [\varepsilon_{1T}\ \varepsilon_{2T}]$ where Π_{1T} is a $J_T \times p$ matrix and ε_{1T} is a $T \times p$ matrix. $\tilde{\Lambda}_{1T}$ is a $p \times p$ diagonal matrix whose i-th diagonal element is $\sqrt{T}/\delta_{i,T}$. Lastly, $\Upsilon_T = \begin{pmatrix} \Upsilon_1 & \Upsilon_{12,T} \\ \Upsilon'_{12,T} & \Upsilon_{2,T} \end{pmatrix}$ so that $\frac{\Pi'_{1T}Z'Z\Pi_{1T}}{T} \xrightarrow{p} \Upsilon_1$ by the assumption above in this example.

By realizing that $\hat{\mu}_T$ is the OLS estimator of μ_T^0 with x and $p_{K_T}(w)$ projected onto the space of Z, we can derive the following closed-form expression of $\hat{\theta}_T$ from Frisch-Waugh theorem and the reduced form equation given in section 2.4

$$\begin{aligned}\hat{\theta}_T - \theta^0 &= \left[(P_Z X)' M_{P_{ZW}} P_Z X\right]^{-1} (P_Z X)' M_{P_{ZW}} (u_T + e_T) \\ &= \sqrt{T}\bar{\Lambda}_T^{-1} \left[(Z\Pi_{1T} + \varepsilon_{1T})' P_Z M_{P_{ZW}} (Z\Pi_{1T} + \varepsilon_{1T})\right]^{-1} \left[(Z\Pi_{1T} + \varepsilon_{1T})' P_Z M_{P_{ZW}} (u_T + e_T)\right],\end{aligned}$$

which is equivalent to

$$\begin{aligned}\bar{\Lambda}_T(\hat{\theta}_T - \theta^0) = &\left[\frac{(Z\Pi_{1T} + \varepsilon_{1T})' P_Z M_{P_{ZW}} (Z\Pi_{1T} + \varepsilon_{1T})}{T}\right]^{-1} \\ &\times \left[\frac{(Z\Pi_{1T} + \varepsilon_{1T})' P_Z M_{P_{ZW}} (u_T + e_T)}{\sqrt{T}}\right].\end{aligned} \tag{A.2}$$

The last part of the above assumption, together with the i.i.d. assumption given in subsection 2.4 and the fourth part of the above assumption, lets $\Pi'_{1T}Z'M_{P_{ZW}}u_T/\sqrt{T}$ be asymptotically normal with zero mean with variance which is the limit of V_{1T}:

$$V_{1T} = \frac{1}{T}\sum_{t=1}^{T} M_{tt}^2 \Pi'_{1T} E\left[z_{t,T} z'_{t,T} u_t^2 | z_t\right] \Pi_{1T} + \frac{1}{T}\sum_{t=1}^{T}\sum_{s=1, s\neq t}^{T} M_{ts}^2 \Pi'_{1T} E\left[z_{t,T} z'_{t,T} u_s^2 | z_t\right] \Pi_{1T}$$

where the second term converges to zero since, assuming $\Pi'_{1T} z_{t,T} z'_{t,T} \Pi_{1T}$ is uniformly bounded from above by some positive constant C,

$$\frac{1}{T}\sum_{t=1}^{T}\sum_{s=1, s\neq t}^{T} M_{ts}^2 \Pi'_{1T} z_{t,T} z'_{t,T} E\left[u_s^2 | z_t\right] \Pi_{1T} \leq \frac{C}{T}\sum_{t=1}^{T}\sum_{s=1, s\neq t}^{T} M_{ts}^2 \leq C\frac{K_T}{T} \to 0.$$

and

$$\frac{1}{T}\sum_{t=1}^{T} M_{tt}^2 \Pi'_{1T} E\left[z_{t,T} z'_{t,T} u_t^2 | z_t\right] \Pi_{1T} = \Pi'_{1T} E\left[z_{t,T} z'_{t,T} u_t^2\right] \Pi_{1T} + o_p(1)$$

by using lemma A2 in Chao et al. (2012).

Also, using lemma 1 in Cattaneo et al. (2018), we can show that

$$\frac{\Pi_{1T}' Z' M_{P_{ZW}} Z \Pi_{1T}}{T} = \frac{1}{T} \sum_{t=1}^{T} \Pi_{1T}' z_{t,T} z_{t,T}' \Pi_{1T} + o_p(1) = \Upsilon_1 + o_p(1)$$

where the last equality holds by the above assumption(i).

From the above assumption(iii) that $J_T^2/T \to 0$, all the remaining terms in (A.2) converge to 0. Thus, $\bar{\Lambda}_T(\hat{\theta}_T - \theta^0)$ is asymptotically normal with some bias due to the approximation error (misspecification) of $g^0(w)$ and variance $\Upsilon_1^{-1} \Pi_{1T}' E\left[z_{t,T} z_{t,T}' u_t^2\right] \Pi_{1T} \Upsilon_1^{-1}$. This term is indeed the first $p \times p$ sub-matrix of $D^{-1} V D^{-1}$ in our theorem 3.

The rate of convergence of the estimation is possibly slower than $\sqrt{T}$, even though the instruments do not display weakness, due to the near multicollinearity problem. When all elements in $\bar{\Lambda}_T$ diverges at the rate of $\sqrt{T}$ and x and w are both exogenous, we are back to the case considered in Cattaneo et al. (2018).

Appendix B. Comparison with the example of Chao and Swanson (2005) (CS)

We consider a linear model with the following equations

$$y_{1T} = Y_{2T}\theta + u_T$$
$$Y_{2T} = Z_T \Pi_T + V_T,$$

where y_{1T} and Y_{2T} and a $T \times 1$ vector and a $T \times p$ matrix of endogenous observables, respectively, Z_T is a $T \times J_T$ matrix of instrumental variables, and u_T and V_T are a $T \times 1$ vector and $T \times p$ matrix of error terms. Note that $J_T \to \infty$ corresponds to an increasing number of instruments. By assuming that Z_T is a matrix of valid instruments, we have the following orthogonality condition

$$E\left[\psi_{t,T}(\theta^0)\right] = E\left[z_{t,T}(y_{1,t} - y_{2,t}'\theta^0)\right] = 0.$$

where $z_{t,T}$, $y_{1,t}$ and $y_{2,t}$ are elements of Z_T, y_{1T} and Y_{2T} for observation t with $z_{t,T}$ vector of dimension J_T. The identification strength of instruments is expressed through the local to zero coefficient Π_T such that

$$\Pi_T = \frac{A_T}{\sqrt{T}} C_T,$$

where A_T is a J_T-dimensional positive definite matrix such that the largest eigenvalue, $\bar{\lambda}_T$, is $O(\sqrt{T})$. One example of this local to zero coefficient is the case with

$$\Pi_T = \frac{A_T}{\sqrt{T}} C_T = \frac{\Lambda_T}{\sqrt{T}} C_T,$$

so that each instrument is associated with coefficients that may converge to zero at different rates.

We allow the identification strength to be different across instruments although CS does not explicitly discuss such heterogeneity. It is not hard to see that the following moment conditions hold

$$E\left[\bar{\psi}_T(\theta)\right] = \frac{1}{\sqrt{T}} E\left[z_{t,T} z_{t,T}'\right] A_T C_T(\theta^0 - \theta) = \frac{\Lambda_T}{\sqrt{T}} \rho_T(\theta).$$

where $\rho_T(\theta) = C_T(\theta^0 - \theta)$ and $E[z_{t,T} z_{t,T}']$ is assumed to be positive definite for all T. Then the function $\delta(\theta)$ introduced by assumption 1(i) is

$$\delta(\theta) = \lim_{T\to\infty} \frac{1}{J_T} (\theta^0 - \theta)' C_T' C_T (\theta^0 - \theta).$$

Therefore, a sufficient condition for assumption 1(i) and (ii) is:

$$\lim_{T\to\infty} \frac{C_T' C_T}{J_T} = \Delta \tag{B.1}$$

with Δ positive definite matrix. Note that this assumption is natural since

$$\frac{C_T' C_T}{J_T} = \frac{1}{J_T} \sum_{j=1}^{J_T} C_T^j C_T^{j\prime},$$

where $C_T^{j\prime}, j = 1, ..., J_T$ stand for the J_T rows of the matrix C_T. In the case of a finite number J of instruments ($J_T = J$), it is tantamount to the standard rank condition for identification that the matrix C_T (or equivalently the matrix $\Pi_T = \frac{A_T}{\sqrt{T}} C_T$) is of full column rank p. In the general case, it is worth relating this assumption to the asymptotic behavior of the concentration matrix

$$\Psi_T = \Pi_T' Z_T' Z_T \Pi_T = C_T' A_T' \frac{Z_T' Z_T}{T} A_T C_T \sim C_T' \Lambda_T^2 C_T.$$

In the case with homogeneous strength of identification across all instruments, that is $A_T = \Lambda_T = \lambda_T I_{J_T}$ (the only case considered by CS),

$$\Psi_T \sim \lambda_T^2 C_T' C_T.$$

CS measures the identification strength by a sequence r_T such that

$$\lim_{T\to\infty} \frac{\Psi_T}{r_T} = \Psi \tag{B.2}$$

with Ψ positive definite matrix. They acknowledge that consistency of 2SLS estimators (and k-class estimators as well) takes that

$$\lim_{T\to\infty} \frac{J_T}{r_T} = 0. \tag{B.3}$$

It is worth noting that this assumption maintained by CS (as far as consistency of 2SLS is concerned) is generalized by our assumption 1 and the condition (2.10) in theorem 1. More precisely, in the particular setup $\Lambda_T = \lambda_T I_{J_T}$, the conditions (B.1) for assumption 1 and (B.2) for CS are equivalent for

$$r_T = J_T \lambda_T^2 \tag{B.4}$$

and then, the CS condition (B.3) just means that $\lambda_T \to \infty$, that is precisely our assumption (2.10) in theorem 1 in the particular case $\Lambda_T = \lambda_T I_{J_T}$.

Our assumption (2.10) is more general due to the heterogeneity of identification strengths. In order to see that, consider for instance the case with $A_T = \Lambda_T$ and $E[z_{t,T} z'_{t,T}] = Id_{J_T}$ where Λ_T is a diagonal matrix. We note that in this case some occurrences of weak identification (some diagonal coefficients λ_{jT} of Λ_T do not converge to infinity) are not necessarily at odds with our assumption (2.10), as illustrated by the remark right after theorem 1 in subsection 2.3. Note also that (B.4) shows that identification strength should be measured by the rate of convergence of the concentration matrix $J_T \lambda_T^2$. Interestingly enough, our asymptotic distribution theory in section 3 will provide a multivariate generalization of this measure. We show that the possible rates of convergence in different directions of the parameter space are precisely defined by the products $\lambda_{jT}\sqrt{J_T}$. In the much more general nonlinear setup of this paper, it is interesting to note that we still have identification strength measured by something like the rate of convergence of a concentration parameter, namely the rate of convergence of λ_{jT} (smaller than $\sqrt{T}$ if and only if identification is not strong) times the square root of the number J_T of moments. As already announced, the fact that the number of moment conditions does go to infinity with the sample size allows us to reinforce identification.

While our assumption (2.10) in theorem 1 places the same restriction on the strength of instruments as Chao and Swanson (2005) for the 2SLS estimator, we want to stress that consistent estimation is still possible for weaker instruments by other estimators such as k-class estimators including LIML and Fuller that are already proposed in the weak instrument literature. Those estimators are consistent when all instruments are genuinely weak (constant λ_T) or even weaker than weak and the number of instruments grows with the sample size.

Note that this means that the true parameter θ^0 is actually identified even when the instruments

are weak. For the case of the 2SLS estimator, the moments are induced from the conditional moment

$$E\left[y_{1t} - y_{2,t}'\theta | z_{t,T}\right] = 0,$$

with instrument $z_{t,T}$. This conditional moment also weakly identifies θ^0 since, by assuming $\Lambda_T = \lambda_T I_{J_T}$,

$$\begin{aligned} E\left[y_{1t} - y_{2,t}'\theta | z_{t,T}\right] &= E\left[y_{2,t}'(\theta^0 - \theta) + u_t | z_{t,T}\right] \\ &= E\left[(z_{t,T}'\Pi_T + v_t)(\theta^0 - \theta) | z_{t,T}\right] \\ &= z_{t,T}'\frac{\lambda_T}{\sqrt{T}} C_T. \end{aligned}$$

However, the consistency results by the k-class estimators show that the above conditional moment still identifies θ^0. It is rather a problem of choosing $z_{t,T}$ as the instruments to form unconditional moments that fail the identification. As noted by Domínguez and Lobato (2004), an arbitrary choice of moment conditions do not ensure the parameter identifiability even when the parameters are identified by the conditional moments. We see now that the same identification issue arises in this linear IV model and an appropriate choice of instruments can remedy the problem when the conditional moments have a sufficient identification strength with a large number of instruments.

Finally, as we announce in section 3.1, we now verify that the second part of our assumption 4 holds under this framework. Note that $\theta_T^* = \theta^0$ for all T as $\rho_T(\theta^0) = 0$. Let $z_{t,T} = [z_{i,t,T}]_{1 \le i \le l}$ where $z_{i,t,T}$ be the $J_{i,T}-$dimensional vector of instruments. For expositional simplicity, assume that $A_T = \Lambda_T$ where Λ_T is a diagonal matrix and that $W_T = (Z_T'Z_T/T)^{-1} = E[z_{t,T}'z_{t,T}] = Id_{J_T}$ (hence, $W_{iT} = (Z_{iT}'Z_{iT}/T)^{-1} = Id_{J_{i,T}}$ where $Z_T' = [Z_{iT}']_{1 \le i \le l}$). Then, for $i = 1, \cdots, l$, we have

$$\begin{aligned} &X_{iT}(\theta) - X_{iT}(\theta^0) \\ &= \frac{T}{r_{i,T}^2}\left[\frac{1}{T}\sum_{t=1}^T z_{i,t,T}y_{2,t}'(\theta^0 - \theta) + \frac{1}{T}\sum_{t=1}^T z_{i,t,T}u_t\right]'\left[\frac{1}{T}\sum_{t=1}^T z_{i,t,T}y_{2,t}'(\theta^0 - \theta) + \frac{1}{T}\sum_{t=1}^T z_{i,t,T}u_t\right] \\ &- \frac{\lambda_{i,T}^2}{r_{i,T}^2}(\theta^0 - \theta)'C_i'C_i(\theta^0 - \theta) - \frac{T}{r_{i,T}^2}\left[\frac{1}{T}\sum_{t=1}^T z_{i,t,T}u_t\right]'\left[\frac{1}{T}\sum_{t=1}^T z_{i,t,T}u_t\right], \end{aligned} \tag{B.5}$$

where $r_{i,T}^2 = \lambda_{i,T}^2 J_{i,T}, i = 1, \cdots, l$, $V_T = (v_t)_{1 \le t \le T}$, $u_T = (u_t)_{1 \le t \le T}$ and where C_i is a $J_{i,T} \times p$

matrix of coefficients such that

$$C_T = \begin{pmatrix} C_1 \\ \vdots \\ C_l. \end{pmatrix}$$

$$\frac{C_i' C_i}{J_{i,T}} \to \rho_i, \ \rho_i \text{ is p.d., } i = 1, \cdots, l \ .$$

Since

$$y_{2,t} = \Pi_T' z_{t,T} + v_t = C_T' \frac{\Lambda_T}{\sqrt{T}} z_{t,T} + v_t,$$

we have

$$\begin{aligned} \frac{1}{T}\sum_{t=1}^{T} z_{i,t,T} y_{2,T}'(\theta^0 - \theta) &= \frac{1}{T}\sum_{t=1}^{T} z_{i,t,T} z_{t,T}' \frac{\Lambda_T}{\sqrt{T}} C_T(\theta^0 - \theta) + \frac{1}{T}\sum_{t=1}^{T} z_{i,t,T} v_t'(\theta^0 - \theta) \\ &= \frac{\lambda_{i,T}}{\sqrt{T}} C_i(\theta^0 - \theta) + \frac{1}{T}\sum_{t=1}^{T} z_{i,t,T} v_t'(\theta^0 - \theta), \end{aligned}$$

Then we can rewrite (B.5) as follows,

$$X_{iT}(\theta) - X_{iT}(\theta^0) = 2\Psi_{i,1T} + 2\Psi_{i,2T} + 2\Psi_{i,3T} + \Psi_{i,4T}$$

where

$$\begin{aligned} \Psi_{i,1T} &= \frac{\lambda_{i,T}\sqrt{T}}{r_{i,T}^2}(\theta^0 - \theta)' C_i' \left[\frac{1}{T}\sum_{t=1}^{T} z_{i,t,T} v_t'\right](\theta^0 - \theta) = \frac{\lambda_{i,T}}{r_{i,T}^2}(\theta^0 - \theta)' C_i' \frac{Z_{iT}' V_T}{\sqrt{T}}(\theta^0 - \theta) \\ \Psi_{i,2T} &= \frac{\lambda_{i,T}}{r_{i,T}^2}(\theta^0 - \theta)' C_i' \left[\frac{1}{T}\sum_{t=1}^{T} z_{i,t,T} u_t\right] = \frac{\lambda_{i,T}}{r_{i,T}^2}(\theta^0 - \theta)' C_i' \frac{Z_{iT}' u_T}{\sqrt{T}} \\ \Psi_{i,3T} &= \frac{T}{r_{i,T}^2}(\theta^0 - \theta)' \left[\frac{1}{T}\sum_{t=1}^{T} v_t z_{i,t,T}'\right]\left[\frac{1}{T}\sum_{t=1}^{T} z_{i,t,T} u_t\right] = \frac{1}{r_{i,T}^2}(\theta^0 - \theta)' V_T' P_{Z_{iT}} u_T \\ \Psi_{i,4T} &= \frac{T}{r_{i,T}^2}(\theta^0 - \theta)' \left[\frac{1}{T}\sum_{t=1}^{T} v_t z_{i,t,T}'\right]\left[\frac{1}{T}\sum_{t=1}^{T} z_{i,t,T} v_t'\right](\theta^0 - \theta) = \frac{T}{r_{i,T}^2}(\theta^0 - \theta)' V_T' P_{Z_{iT}} V_T(\theta^0 - \theta) \end{aligned}$$

with $P_{Z_{iT}} = Z_{iT}(Z_{iT}' Z_{iT})^{-1} Z_{iT}' = Z_{iT} Z_{iT}'/T$. Hence,

$$\lambda_{l,T}\sqrt{J_{l,T}}\left|X_{iT}(\theta) - X_{iT}(\theta^0)\right| = \lambda_{l,T}\sqrt{J_{l,T}}\left|2\Psi_{i,1T} + 2\Psi_{i,2T} + 2\Psi_{i,3T} + \Psi_{i,4T}\right|.$$

It can now be easily checked that these terms $\Psi_{i,1T}, \cdots, \Psi_{i,4T}$ are controlled for all $i = 1, \cdots, l$. To see that, one can for instance maintain the assumptions of Chao and Swanson (2005) and apply

their lemma A1 in page 1681. Following their notations, let

$$Var(v_t u_t) = \Sigma = \begin{pmatrix} \sigma_{uu} & \sigma'_{Vu} \\ \sigma_{Vu} & \Sigma_{VV} \end{pmatrix}$$

with σ^j_{Vu} denoting the i-th element of σ_{Vu} for $j = 1, \cdots, p$. Using their proof, we have, by letting C be a generic constant that take different values in different equations,

$$\begin{aligned}
E\left[\underset{\|\theta^0-\theta\|<\epsilon}{sup}\left|\lambda_{l,T}\sqrt{J_{l,T}}\Psi_{i,1T}\right|^2\right] &\leq C\epsilon\frac{\lambda^2_{l,T}J_{l,T}\lambda^2_{i,T}}{(\lambda^2_{i,T}J_{i,T})^2}E\left[\frac{\|V'_T Z_{iT}C_{iT}\|^2}{T}\right] \\
&\leq C\epsilon\frac{\lambda^2_{l,T}J_{l,T}}{\lambda^2_{i,T}J_{i,T}}\frac{1}{J_{i,T}}E\left[\frac{Tr\left(C'_{iT}Z'_{iT}V_T V'_T Z_{iT}C_{iT}\right)}{T}\right] \\
&= C\epsilon\frac{\lambda^2_{l,T}J_{l,T}}{\lambda^2_{i,T}J_{i,T}}\frac{1}{J_{i,T}}Tr(\Sigma_{VV})Tr\left(C'_{iT}\left[\frac{Z'_{iT}Z_{iT}}{T}\right]C_{iT}\right) \\
&\leq C\epsilon\frac{1}{J_{i,T}}Tr\left(C'_{iT}C_{iT}\right) \\
&= C\epsilon, \\
E\left[\underset{\|\theta^0-\theta\|<\epsilon}{sup}\left|\lambda_{l,T}\sqrt{J_{l,T}}\Psi_{i,3T}\right|^2\right] &\leq C\epsilon\frac{\lambda^2_{l,T}J_{l,T}}{(\lambda^2_{i,T}J_{i,T})^2}J^2_{i,T}E\left[\left\|\frac{V'_T P_{Z_{iT}}u_T}{J_{i,T}}\right\|^2\right] \\
&= C\epsilon\frac{\lambda^2_{l,T}J_{l,T}}{\lambda^4_{i,T}}\sum_{j=1}^{p}\left((\sigma^j_{Vu})^2 + O(1/J_{i,T})\right) \\
&\leq C\epsilon,
\end{aligned}$$

by assuming that $\lambda^2_{l,T}J_{l,T}/\lambda^4_{i,T} = O(1)$ for all $i = 1, \cdots, l$, and the same results hold for $\Psi_{i,2T}$ and $\Psi_{i,4T}$ as well.

Appendix C. Lemmas and proofs

Lemmas and proofs are provided. For a $K \times 1$ vector x, $\|x\| = \sqrt{\sum_{k=1}^K x_k^2}$ and for a $K \times J$ matrix A, $\|A\| = \sqrt{\lambda_{max}(A'A)}$. $\lambda_{max}(X)$ and $\lambda_{min}(X)$ denote the maximum and the minimum eigenvalues of a square matrix X, respectively. C is a generic positive constant that may be different in different uses.

Proof of Lemma 1

Assumption 2 suggests that we see the sample objective function as

$$\hat{Q}_T(\theta) = \left[\frac{\Phi_T(\theta)}{\sqrt{T}} + \frac{\Lambda_T}{\sqrt{T}}\rho_T(\theta)\right]' W_T \left[\frac{\Phi_T(\theta)}{\sqrt{T}} + \frac{\Lambda_T}{\sqrt{T}}\rho_T(\theta)\right].$$

We want to show that

$$x_T = \frac{\Lambda_T \rho_T(\hat{\theta}_T)}{\sqrt{J_T}} = O_P(1).$$

We note that

$$\frac{T}{J_T}\hat{Q}_T(\hat{\theta}_T) = x_T' W_T x_T + 2\frac{\Phi_T(\hat{\theta}_T)'}{\sqrt{J_T}} W_T x_T + \frac{1}{J_T}\Phi_T(\hat{\theta}_T)' W_T \Phi_T(\hat{\theta}_T).$$

Thus, if $\underline{\mu}_T$ and $\bar{\mu}_T$ stand respectively for the smallest and the largest eigenvalue of W_T, we have

$$\frac{T}{J_T}\hat{Q}_T(\hat{\theta}_T) \geq \underline{\mu}_T \|x_T\|^2 - 2\bar{\mu}_T \left\| \frac{\Phi_T(\hat{\theta}_T)}{\sqrt{J_T}} \right\| \|x_T\| + \underline{\mu}_T \left\| \frac{\Phi_T(\hat{\theta}_T)}{\sqrt{J_T}} \right\|^2.$$

Since by definition of the GMM estimator

$$\frac{T}{J_T}\hat{Q}_T(\hat{\theta}_T) \leq \frac{T}{J_T}\hat{Q}_T(\theta^0),$$

we conclude that

$$\|x_T\|^2 - 2\frac{\bar{\mu}_T}{\underline{\mu}_T} \left\| \frac{\Phi_T(\hat{\theta}_T)}{\sqrt{J_T}} \right\| \|x_T\| + \left\| \frac{\Phi_T(\hat{\theta}_T)}{\sqrt{J_T}} \right\|^2 - \frac{T}{J_T}\frac{\hat{Q}_T(\theta^0)}{\underline{\mu}_T} \leq 0.$$

Thus

$$\|x_T\| \leq \frac{\bar{\mu}_T}{\underline{\mu}_T} \left\| \frac{\Phi_T(\hat{\theta}_T)}{\sqrt{J_T}} \right\| + \sqrt{\Delta_T},$$

with

$$\Delta_T = \left[\frac{\bar{\mu}_T}{\underline{\mu}_T} \left\| \frac{\Phi_T(\hat{\theta}_T)}{\sqrt{J_T}} \right\| \right]^2 - \left\| \frac{\Phi_T(\hat{\theta}_T)}{\sqrt{J_T}} \right\|^2 + \frac{T}{J_T}\frac{\hat{Q}_T(\theta^0)}{\underline{\mu}_T}.$$

By assumption 2

$$\left\| \frac{\Phi_T(\hat{\theta}_T)}{\sqrt{J_T}} \right\| = O_P(1),$$

and thus we also have

$$\frac{\bar{\mu}_T}{\underline{\mu}_T} \left\| \frac{\Phi_T(\hat{\theta}_T)}{\sqrt{J_T}} \right\| = O_P(1),$$

since

$$\begin{aligned} \bar{\mu}_T &= \|W_T\| = O_P(1) \\ \frac{1}{\underline{\mu}_T} &= \left\|W_T^{-1}\right\| = O_P(1). \end{aligned}$$

Therefore, we will be able to conclude that $x_T = O_P(1)$ if we show that

$$\frac{T}{J_T}\frac{\hat{Q}_T(\theta^0)}{\underline{\mu}_T} = O_P(1)$$

or (since $1/\underline{\mu}_T = O_P(1)$) that:

$$\frac{T}{J_T}\hat{Q}_T(\theta^0) = O_P(1). \tag{C.1}$$

By a computation similar to the one used above for $\hat{Q}_T(\hat{\theta}_T)$, we have

$$\begin{aligned}\frac{T}{J_T}\hat{Q}_T(\theta^0) &= y_T' W_T y_T + 2\frac{\Phi_T(\theta^0)'}{\sqrt{J_T}} W_T y_T + \frac{1}{J_T}\Phi_T(\theta^0)' W_T \Phi_T(\theta^0) \\ &\leq \bar{\mu}_T \left\|y_T\right\|^2 + 2\bar{\mu}_T \left\|y_T\right\| \left\|\frac{\Phi_T(\theta^0)}{\sqrt{J_T}}\right\| + \bar{\mu}_T \left\|\frac{\Phi_T(\theta^0)}{\sqrt{J_T}}\right\|^2,\end{aligned}$$

where

$$y_T = \frac{\Lambda_T \rho_T(\theta^0)}{\sqrt{J_T}}.$$

We can then conclude (C.1) since

$$\bar{\mu}_T = \left\|W_T\right\| = O_P(1)$$

and by assumptions 1 and 2 respectively

$$y_T = O(1),$$

and:

$$\left\|\frac{\Phi_T(\theta^0)}{\sqrt{J_T}}\right\| = O_P(1).$$

QED

Proof of Theorem 1

By the assumption that $\bar{\lambda}_T^2(\hat{\theta}) \to +\infty$, we have

$$\left\|\rho_T(\hat{\theta}_T)\right\| = o_P\left(\left\|\Lambda_T \rho_T(\hat{\theta}_T)\right\|\right).$$

Then, we can deduce from lemma 1 that

$$\left\|\rho_T(\hat{\theta}_T)\right\| = o_P\left(\sqrt{J_T}\right). \tag{C.2}$$

Then, an argument similar to the one in AR12 will allow us to conclude that $\hat{\theta}_T$ is a weakly consistent estimator of θ^0.

Let $\varepsilon > 0$. By assumption 1(ii)

$$\inf_{\|\theta-\theta^0\|>\varepsilon} \delta(\theta) = \alpha > 0.$$

Let us define

$$\omega_T = \sup_{\theta\in\Theta} \left\{ \delta(\theta) - \frac{\|\rho_T(\theta)\|^2}{J_T} \right\}.$$

Then

$$\frac{\left\|\rho_T(\hat{\theta}_T)\right\|^2}{J_T} + \omega_T \geq \delta\left(\hat{\theta}_T\right)$$

so that

$$\Pr\left[\left\|\hat{\theta}_T - \theta^0\right\| > \varepsilon\right] \leq \Pr\left[\delta(\hat{\theta}_T) \geq \alpha\right] \leq \Pr\left[\frac{\left\|\rho_T(\hat{\theta}_T)\right\|^2}{J_T} + \omega_T \geq \alpha\right]. \tag{C.3}$$

However, we know by assumption 1(i) and (C.2) that:

$$\frac{\left\|\rho_T(\hat{\theta}_T)\right\|^2}{J_T} + \omega_T = o_P(1)$$

and thus, for all $\alpha > 0$

$$\lim_{T\to\infty} \Pr\left[\frac{\left\|\rho_T(\hat{\theta}_T)\right\|^2}{J_T} + \omega_T \geq \alpha\right] = 0,$$

and we deduce from (C.3) that

$$\lim_{T\to\infty} \Pr\left[\left\|\hat{\theta}_T - \theta^0\right\| > \varepsilon\right] = 0.$$

This can be done for all $\varepsilon > 0$ and thus proves that

$$P \lim_{T\to\infty} \hat{\theta}_T = \theta^0.$$

QED

Proof of Theorem 2

We follow the proof of Proposition 21 of HP with $\alpha = 2$ and $\beta = 1$.

Fix $M > 0$. Let $r_T = \sqrt{\lambda_{l,T}^2 J_{l,T}}$. Let $S_{j,T} = \{\theta : 2^{j-1} < r_T\|\theta - \theta_T^*\| < 2^j\}$. Then, for every

$\eta > 0$,

$$P\left(r_T\|\hat{\theta}_T - \theta_T^*\| > 2^M\right) \leq \sum_{j\geq M, 2^j \leq \eta r_T} P\left(\inf_{\theta\in S_{j,T}} T\left[\bar{\psi}_T(\theta)'W_T\bar{\psi}_T(\theta) - \bar{\psi}_T(\theta_T^*)'W_T\bar{\psi}_T(\theta_T^*)\right] \leq 0\right)$$
$$+ P\left(2\left\|\hat{\theta}_T - \theta_T^*\right\| \geq \eta\right). \tag{C.4}$$

Since $\hat{\theta}_T$ is consistent for θ^0 as theorem 1 shows and θ_T^* converges to θ^0, the second probability on the right converges to 0 for every η as $T \to \infty$. Also since, for a constant C that takes different values in different uses,

$$\begin{aligned}
&\inf T\left[\bar{\psi}_T(\hat{\theta}_T)'W_T\bar{\psi}_T(\hat{\theta}_T) - \bar{\psi}_T(\theta_T^*)'W_T\bar{\psi}_T(\theta_T^*)\right] \\
&\geq \inf C\sum_{i=1}^{l} T\left[\bar{\psi}_{iT}(\hat{\theta}_T)'W_{iT}\bar{\psi}_{iT}(\hat{\theta}_T) - \bar{\psi}_{iT}(\theta_T^*)'W_{iT}\bar{\psi}_{iT}(\theta_T^*)\right] \\
&\geq \inf C\sum_{i=1}^{l} \frac{T}{\lambda_{i,T}^2 J_{i,T}}\left[\bar{\psi}_{iT}(\hat{\theta}_T)'W_{iT}\bar{\psi}_{iT}(\hat{\theta}_T) - \bar{\psi}_{iT}(\theta_T^*)'W_{iT}\bar{\psi}_{iT}(\theta_T^*)\right] \\
&\geq \inf C\sum_{i=1}^{l}\left[X_{iT}(\theta) - X_{iT}(\theta_T^*)\right] + \inf C\sum_{i=1}^{l}\frac{1}{J_{i,T}}\left[\rho_{iT}(\theta)'W_{iT}\rho_{iT}(\theta) - \rho_{iT}(\theta_T^*)'W_{iT}\rho_{iT}(\theta_T^*)\right],
\end{aligned}$$

for any common set over which the inf operator applied, we have for each j,

$$\begin{aligned}
&P\left(\inf_{\theta\in S_{j,T}} T\left[\bar{\psi}_T(\hat{\theta}_T)'W_T\bar{\psi}_T(\hat{\theta}_T) - \bar{\psi}_T(\theta_T^*)'W_T\bar{\psi}_T(\theta_T^*)\right] \leq 0\right) \\
&\leq P\left(-\inf_{\theta\in S_{j,T}}\sum_{i=1}^{l}\left[X_{iT}(\theta) - X_{iT}(\theta_T^*)\right] \geq \inf_{\theta\in S_{j,T}}\sum_{i=1}^{l}\frac{1}{J_{i,T}}\left[\rho_{iT}(\theta)'W_{iT}\rho_{iT}(\theta) - \rho_{iT}(\theta_T^*)'W_{iT}\rho_{iT}(\theta_T^*)\right]\right) \\
&\leq P\left(\sup_{\theta\in S_{j,T}}\left|\sum_{i=1}^{l} X_{iT}(\theta) - X_{iT}(\theta_T^*)\right| \geq \inf_{\theta\in S_{j,T}}\sum_{i=1}^{l}\frac{1}{J_{i,T}}\left[\rho_{iT}(\theta)'W_{iT}\rho_{iT}(\theta) - \rho_{iT}(\theta_T^*)'W_{iT}\rho_{iT}(\theta_T^*)\right]\right).
\end{aligned}$$

Choose η small enough to ensure that the first condition of the theorem holds for all θ such that $\|\theta - \theta_T^*\| < \eta$ and the second condition holds for every $\epsilon < \eta$. Then, for each j, the above probability is bounded by

$$P\left(\sup_{\theta\in S_{j,T}}\left|\sum_{i=1}^{l} r_T X_{iT}(\theta) - r_T X_{iT}(\theta_T^*)\right| \geq C r_T 2^{2(j-1)} r_T^{-2}\right)$$

for some finite C. By Markov inequality and the second condition,

$$P\left(\sup_{\theta\in S_{j,T}}\left|\sum_{i=1}^{l} r_T X_{iT}(\theta) - r_T X_{iT}(\theta_T^*)\right| \geq C 2^{2(j-1)} r_T^{-1}\right) \leq C^{-1} r_T 2^{-2(j-1)} r_T^{-1} 2^j = C^{-1} 2^2 2^{-j}.$$

Therefore, the first term on the right of (C.4) is eventually bounded by $C^{-1}2^2\sum_{j\geq M} 2^{-j}$ which

converges to 0 as $M \to \infty$. This proves that

$$r_T \|\hat{\theta}_T - \theta_T^*\| = O_p(1).$$

We now remain to show

$$r_T \|\theta^0 - \theta_T^*\| = O_p(1).$$

By assumption 5, $\|\rho_{iT}(\theta^0)\| = O(1/\lambda_{i,T})$. By the condition (i) in this theorem,

$$\sum_{i=1}^{l} \frac{1}{J_{i,T}} \rho_{iT}(\theta_T^*)' W_{iT} \rho_{iT}(\theta_T^*) \leq \sum_{i=1}^{l} \frac{1}{J_{i,T}} \rho_{iT}(\theta^0)' W_{iT} \rho_{iT}(\theta^0) = O_p\left(\frac{1}{\lambda_{l,T}^2 J_{l,T}}\right).$$

Hence,

$$\begin{aligned} r_T^2 \|\theta_T^* - \theta^0\|^2 &\leq C r_T^2 \sum_{i=1}^{l} \frac{1}{J_{i,T}} \left[\rho_{iT}(\theta^0)' W_{iT} \rho_{iT}(\theta^0) - \rho_{iT}(\theta_T^*)' W_{iT} \rho_{iT}(\theta_T^*)\right] \\ &= O_p(1). \end{aligned}$$

QED

Lemma 2

Under assumptions 1-11,

$$\sqrt{T} \tilde{\Lambda}_T^{-1} [R^0]' \frac{\partial \bar{\psi}_T(\hat{\theta}_T)'}{\partial \theta} W_T \sqrt{T} \bar{\psi}_T(\theta^0) \xrightarrow{d} \mathcal{N}(B, V)$$

where V is defined in assumption 11 and $B = \underset{T \to \infty}{lim} \left(D_T^0\right)' W_T \Lambda_T \rho_T(\theta^0)$.

Proof of Lemma 2

Note that

$$\begin{aligned} \sqrt{T} \tilde{\Lambda}_T^{-1} [R^0]' \frac{\partial \bar{\psi}_T(\hat{\theta}_T)'}{\partial \theta} W_T \sqrt{T} \bar{\psi}_T(\theta^0) = &\left(\sqrt{T} \tilde{\Lambda}_T^{-1} [R^0]' \frac{\partial \bar{\psi}_T(\hat{\theta}_T)'}{\partial \theta} - \left(D_T^0\right)'\right) W_T \sqrt{T} \bar{\psi}_T(\theta^0) \\ &+ \left(D_T^0\right)' W_T \sqrt{T} \bar{\psi}_T(\theta^0) \end{aligned}$$

Let $A_T = \sqrt{T}\tilde{\Lambda}_T^{-1}[R^0]' \frac{\bar{\psi}_T(\hat{\theta}_T)'}{\partial\theta} - \left(D_T^0\right)'$. Note that

$$\begin{aligned} A_T W_T \sqrt{T}\bar{\psi}_T(\theta^0) &= A_T W_T \left((\Phi_T(\theta^0) + \Lambda_T \rho_T(\theta^0)\right) \\ &= A_T W_T \Phi_T(\theta^0) + A_T W_T \Lambda_T \rho_T(\theta^0) \\ &= \Psi_{4T} + \Psi_{5T} \end{aligned}$$

Since $\frac{\left\|\Phi_T(\theta^0)\right\|}{\sqrt{J_T}} = O_p(1)$ by assumption 2, $\left\|\Lambda_T \rho_T(\theta^0)\right\| = O_p(1)$ by assumption 5, and $\|A_T\| = o_p(1)$ by assumption 10, we can deduce

$$\|\Psi_{4T}\| = o_p(1), \quad \|\Psi_{5T}\| = o_p(1).$$

Also,

$$\begin{aligned} \left(D_T^0\right)' W_T \sqrt{T}\bar{\psi}_T(\theta^0) &= \left(D_T^0\right)' W_T \Big(\Phi_T(\theta^0) + \Lambda_T \rho_T(\theta^0)\Big) \\ &= \left(D_T^0\right)' W_T \Phi_T(\theta^0) + \left(D_T^0\right)' W_T \Lambda_T \rho_T(\theta^0). \end{aligned}$$

By assumptions 5 and 11, $\left(D_T^0\right)' W_T \Lambda_T \rho_T(\theta^0) = O(1)$. Now we will prove that

$$\left(D_T^0\right)' W_T \Phi_T(\theta^0) \xrightarrow{d} \mathcal{N}(0, V),$$

by showing that the Lyapunov conditions for central limit theorem hold. Let

$$v_{t,T} = (D_T^0)' W_T \frac{1}{\sqrt{T}} (\psi_{t,T}(\theta^0) - E[\psi_{t,T}(\theta^0]).$$

Let α is a p-dimensional vector with $\|\alpha\| = 1$. Then we see that $E[\alpha' v_{t,T}] = 0$ and

$$\sigma_T^2 = \sum_{t=1}^{T} E\left[\alpha' v_{t,T} v_{t,T}' \alpha\right] = \alpha'(D_T^0)' W_T S_T W_T D_T^0 \alpha \to \alpha' V \alpha.$$

Since V is positive definite by assumption 11(i), $\sigma_T^{-4} = O(1)$. In addition,

$$\begin{aligned} (\alpha' v_{t,T})^4 &= T^{-2}\left(\alpha'(D_T^0)' W_T (\psi_{t,T}(\theta^0) - E[\psi_{t,T}(\theta^0])\right)^4 \\ &\le T^{-2} \|\alpha\|^4 \left\|D_T^0\right\|^4 \|W_T\|^4 \left\|\psi_{t,T}(\theta^0) - E[\psi_{t,T}(\theta^0])\right\|^4 \\ &\le C T^{-2} \left\|\psi_{t,T}(\theta^0) - E[\psi_{t,T}(\theta^0])\right\|^4, \end{aligned}$$

where C is some generic constant. Then, as $T \to \infty$,

$$\frac{1}{\sigma_T^4} \sum_{t=1}^{T} E\left[(\alpha' v_{t,T})^4\right] \leq \frac{1}{\sigma_T^4} \frac{1}{T^2} \sum_{t=1}^{T} CE\left[\left\|\psi_{t,T}(\theta^0) - E[\psi_{t,T}(\theta^0])\right\|^4\right] \to 0,$$

by assumption 11(iv). Hence we obtain the desired asymptotic normality result by central limit theorem. Then, we have

$$\left(D_T^0\right)' W_T \sqrt{T} \bar{\psi}_T(\theta^0) \xrightarrow{d} \mathcal{N}(B, V).$$

QED

Proof of Theorem 3

By mean-value theorem,

$$\hat{\theta}_T - \theta^0 = \left(\frac{\partial \bar{\psi}_T(\hat{\theta}_T)'}{\partial \theta} W_T \frac{\partial \bar{\psi}_T(\bar{\theta}_T)}{\partial \theta'} \right)^{-1} \frac{\partial \bar{\psi}_T(\hat{\theta}_T)'}{\partial \theta} W_T \bar{\psi}_T(\theta^0)$$

for some $\bar{\theta}_T$ between $\hat{\theta}_T$ and θ^0. Then

$$\begin{aligned} \tilde{\Lambda}_T [R^0]^{-1} (\hat{\theta}_T - \theta^0) = & \left(\sqrt{T} \tilde{\Lambda}_T^{-1} [R^0]' \frac{\partial \bar{\psi}_T(\hat{\theta}_T)'}{\partial \theta} W_T \frac{\partial \bar{\psi}_T(\hat{\theta}_T)}{\partial \theta'} [R^0] \tilde{\Lambda}_T^{-1} \sqrt{T} \right)^{-1} \\ & \times \sqrt{T} \tilde{\Lambda}_T^{-1} [R^0]' \frac{\partial \bar{\psi}_T(\hat{\theta}_T)'}{\partial \theta} W_T \sqrt{T} \bar{\psi}_T(\theta^0). \end{aligned}$$

By assumption 10,

$$\sqrt{T} \tilde{\Lambda}_T^{-1} [R^0]' \frac{\partial \bar{\psi}_T(\hat{\theta}_T)'}{\partial \theta} W_T \frac{\partial \bar{\psi}_T(\hat{\theta}_T)}{\partial \theta'} [R^0] \tilde{\Lambda}_T^{-1} \sqrt{T} \xrightarrow{p} \lim_{T=\infty} \left(D_T^0\right)' W_T D_T^0 = D$$

and by lemma 2, the result follows.

QED

References

Antoine, B. and E. Renault (2009): "Efficient GMM with nearly-weak instruments," *The Econometrics Journal*, 12, S135–S171.

——— (2012): "Efficient minimum distance estimation with multiple rates of convergence," *Journal of Econometrics*, 170, 350–367.

Cattaneo, M. D., M. Jansson, and W. K. Newey (2018): "Alternative asymptotics and the partially linear model with many regressors," *Econometric Theory*, 32, 277–301.

Chao, J. C. and N. R. Swanson (2005): "Consistent estimation with a large number of weak instruments," *Econometrica*, 73, 1673–1692.

Chao, J. C., N. R. Swanson, J. A. Hausman, W. K. Newey, and T. Woutersen (2012): "Asymptotic distribution of JIVE in a heteroskedastic IV regression with many instruments," *Econometric Theory*, 28, 42–86.

Domínguez, M. A. and I. N. Lobato (2004): "Consistent estimation of models defined by conditional moment restrictions," *Econometrica*, 72, 1601–1615.

Gagliardini, P., C. Gourieroux, and E. Renault (2011): "Efficient derivative pricing by the extended method of moments," *Econometrica*, 79, 1181–1232.

Han, C. and P. C. B. Phillips (2006): "GMM with many moment conditions," *Econometrica*, 74, 147–192.

Newey, W. K. and F. Windmeijer (2009): "Generalized Method of Moments with many weak moment conditions," *Econometrica*, 77, 687–719.

van der Vaart, A. W. (1998): *Asymptotic Statistics*, Cambridge, MA: Cambridge University Press.