Efficient Minimum Distance Estimation with multiple rates of convergence

Bertille Antoine*and Eric Renault†

July, 16 2007

PRELIMINARY and INCOMPLETE

* *Simon Fraser University. Email: bertille_antoine@sfu.ca.*
† *University of North Carolina at Chapel Hill, CIRANO and CIREQ. Email: renault@email.unc.edu*

1 Introduction

Extension of GMM with several groups of moment conditions associated with different rates of convergence.

2 Theoretical results

2.1 Identification and consistency of a minimum distance estimator

The starting point of minimum distance estimation of an unknown vector θ of p parameters is generally given by K estimating equations:

$$\rho(\theta) = 0 \tag{2.1}$$

These estimating equations are assumed to identify the true unknown value θ^0 of θ thanks to the following maintained assumption:

Assumption 1 *(Identifying equations)*
$\theta \longrightarrow \rho(\theta)$ is a continuous function from a compact parameter space $\Theta \subset \mathbb{R}^p$ into $\mathbb{R}^K$ such that:

$$\rho(\theta) = 0 \iff \theta = \theta^0 \tag{2.2}$$

Note that continuity and compacity altogether actually imply an identification property even stronger than equation (2.2) and crucial in the following:

$$\forall \epsilon > 0 \inf_{\|\theta - \theta^0\| \geq \epsilon} \|\rho(\theta)\| > 0 \tag{2.3}$$

The above result paves the way for consistent minimum distance estimation as soon as we have at our disposal some sample counterpart $\overline{\phi}_T(\theta)$ of the estimating equations. More precisely, with time series notations, we consider that with a sample of size T, corresponding to observations at dates $t = 1, ..., T$ and for any possible value $\theta \in \Theta$ of the parameters, we can compute a K-dimensional sample-based vector $\overline{\phi}_T(\theta)$. In most cases, minimum distance estimation can be seen as GMM because $\overline{\phi}_T(\theta)$ is the sample mean of a double array:

$$\overline{\phi}_T(\theta) = \frac{1}{T} \sum_{t=1}^{T} \phi_{t,T}(\theta) \tag{2.4}$$

In all cases, a minimum distance estimator is defined as usual by:

Definition 2.1 *Let Ω_T be a sequence of symmetric positive definite random matrices of size K which converges in probability towards a positive definite matrix Ω. A minimum distance estimator $\hat{\theta}_T$ of θ^0 is then defined as:*

$$\hat{\theta}_T = \arg\min_{\theta \in \Theta} Q_T(\theta) \tag{2.5}$$

where $Q_T(\theta) = \overline{\phi}'_T(\theta)\Omega_T\overline{\phi}_T(\theta)$

While standard minimum distance asymptotic theory would assume that thanks to some uniform law of large numbers, $\overline{\phi}_T(\theta)$ converges in probability towards $\rho(\theta)$, we will consider more generally a situation where $\overline{\phi}_T(\theta)$ may converge towards zero even for $\theta \neq \theta^0$: however identification is maintained through higher order asymptotics. Let us imagine more precisely that for some positive real number γ:

$$T^\gamma \left[\overline{\phi}_T(\theta) - \frac{\Lambda_T(\theta)}{T^\gamma}\rho(\theta)\right] = \mathcal{O}_P(1) \tag{2.6}$$

where $\Lambda_T(\theta)$ is a diagonal matrix whose coefficients converge to infinity but possibly at slower rates than T^γ. For sake of simplicity, we always see $\Lambda_T(\theta)$ as a deterministic sequence but extension to random sequences with convergence in probability of diagonal terms towards infinity would be straightforward.

The key point is that, when a given diagonal coefficient of $\Lambda_T(\theta)$ goes to infinity strictly slower than T^γ for all θ in some subset $\Theta^* \subset \Theta$, the corresponding component of $\rho(\theta)$ is squeezed to zero and $Plim\ \overline{\phi}_T(\theta) = 0$ for all $\theta \in \Theta^*$. Thus, the probability limit of $\overline{\phi}_T(\theta)$ does not allow to discriminate between θ^0 and any $\theta \in \Theta^*$. Identification will then be recovered thanks to a uniform tightness assumption about (2.6):

Assumption 2 *(Tightness of the error process for the sup norm)*
For some deterministic sequence of diagonal matrices $\Lambda_T(\theta)$ and some positive number γ, the family of random functions:

$$\Psi_T(\theta) = T^\gamma \left[\overline{\phi}_T(\theta) - \frac{\Lambda_T(\theta)}{T^\gamma}\rho(\theta)\right]$$

is such that:

$$\forall \epsilon > 0 \ \exists M \ / \ P\left[\sup_{\theta\in\Theta} \sup_{T\in\mathbb{N}} \|\Psi_T(\theta)\| < M\right] > 1-\epsilon$$

As suggested by a standard central limit theorem, our main focus of interest will be the case $\gamma = 1/2$. More generally, the tightness condition implies that the coefficients of $\Lambda_T(\theta)$ do not go to infinity faster than T^γ. A maintained assumption will then be:

Assumption 3 *$\Lambda_T(\theta)$ is a diagonal matrix with positive coefficients, such that the minimal coefficient, denoted as $\underline{\lambda}_T(\theta)$, verifies:*

$$\lim_{T\to\infty} \inf_{\theta\in\Theta} \underline{\lambda}_T(\theta) = +\infty$$

We can now state our main consistency result:

Theorem 2.1 *Under assumptions 1, 2 and 3, any minimum distance estimator $\hat{\theta}_T$ like (2.5) is weakly consistent.*

As already explained, identification is not ensured is a standard way and the standard proof of consistency of M-estimators (Jenrich (1969), Amemyia (1985)) will not work in this setting. It is worth rewriting the minimization problem (2.5) as follows in order to get some intuition on why we still get consistency:

$$\hat{\theta}_T = \arg\min_{\theta\in\Theta} \left[\frac{\Psi_T(\theta)}{T^\gamma} + \frac{\Lambda_T(\theta)}{T^\gamma}\rho(\theta)\right]' \Omega_T \left[\frac{\Psi_T(\theta)}{T^\gamma} + \frac{\Lambda_T(\theta)}{T^\gamma}\rho(\theta)\right] \tag{2.7}$$

Let us consider for now a standard set of asymptotically normal sample moments, that is $\gamma = 1/2$. The problem when some diagonal coefficients of $\Lambda_T(\theta)$ go to infinity slower than $T^{1/2}$ is that the corresponding components of $\rho(\theta)$ are squeezed to zero in the optimization problem (2.7): hence their identifying power might be lost. This is the reason why we need to refer to the empirical process literature[1]. More precisely, the tightness assumption 2 is helpful to control $\Psi_T(\theta)$ uniformly on Θ to be able to take advantage of the identifying assumption 1 in the minimization problem (2.7). More precisely, while $\Psi_T(\theta)$ is uniformly $\mathcal{O}_P(1)$, we are able to show (see lemma A.1 in the appendix) that:

$$\|\rho(\hat{\theta}_T)\| = \mathcal{O}_P\left(\frac{1}{\inf_{\theta\in\Theta} \underline{\lambda}_T(\theta)}\right) \tag{2.8}$$

[1]This has already been pointed out. See for instance Stock and Wright (2000).

This leads to the consistency of $\hat{\theta}_T$ thanks to the identification property (2.3). Note that a special case of our consistency result (theorem 2.1) has been stated by Lee (2004). However, the case he considers is akin to assuming that $\Psi_T(\theta)$ does not depend on θ and, thus, tightness is no longer an issue. A similar simplification happens in the case of instrumental variables estimation of a linear regression model as in Staiger and Stock (1997) and Hahn and Kuersteiner (2002).

2.2 Disentangling the rates of convergence

As already explained, the focus of interest is multiplicity of rates of convergence induced by the asymptotic behavior of the sample moments $\overline{\phi}_T(\theta)$ and not by singularity issues in the estimating functions $\rho(\theta)$. In this respect, we differ from Sargan (1983) since we maintain the first order identification assumption:

Assumption 4 *(First-order identification)*
(i) $\rho(.)$ *is continuously differentiable on the interior of* Θ, $int(\Theta)$.
(ii) $\theta^0 \in int(\Theta)$.
(iii) The $(K \times p)$*-matrix* $\partial\rho(\theta)/\partial\theta'$ *has full column rank* p *for all* $\theta \in \Theta$.

Thanks to (2.8), we directly deduce from assumption 4 the slowest possible rate of convergence of our estimators:

Theorem 2.2 *Under assumptions 1 to 4, we have:*

$$\left\|\hat{\theta}_T - \theta^0\right\| = \mathcal{O}_P\left(\frac{1}{\inf_{\theta\in\Theta} \underline{\lambda}_T(\theta)}\right)$$

While it may help in some circumstances to consider that the various rates of convergence depend on θ (see e.g. Antoine (2007)), we will simplify the exposition by assuming that for given T, the diagonal matrix Λ_T is fixed. Then, without loss of generality, we will write

throughout the rest of the paper:

$$\Lambda_T = \begin{pmatrix} \lambda_{1T} Id_{k_1} & & & \\ & \lambda_{2T} Id_{k_2} & & \bigcirc \\ \bigcirc & & \ddots & \\ & & & \lambda_{lT} Id_{k_l} \end{pmatrix} \tag{2.9}$$

$$\text{with} \qquad \begin{array}{l} \sum_{i=1}^{l} k_i = K \\ \lim_{T\to\infty} \lambda_{iT} = \infty \quad \text{for} \quad i = 1, ..., l \\ \lambda_{i+1,T} = o(\lambda_{i,T}) \quad \text{for} \quad i = 1, ..., l-1 \end{array}$$

Let us consider accordingly a partition of the estimating equations:

$$\rho(\theta) = [\rho_1'(\theta)\ \rho_2'(\theta)\ \cdots\ \rho_l'(\theta)]' \quad \text{with} \quad dim[\rho_i(\theta)] = k_i \quad \text{for } i = 1, ..., l \tag{2.10}$$

and of their sample counterparts

$$\overline{\phi}_T(\theta) = [\overline{\phi}_{1T}'(\theta)\ \overline{\phi}_{2T}'(\theta)\ \cdots\ \overline{\phi}_{lT}'(\theta)]' \quad \text{with} \quad dim[\overline{\phi}_{iT}(\theta)] = k_i \quad \text{for } i = 1, ..., l \tag{2.11}$$

Then we reinforce assumption 4 as follows:

Assumption 5 *(Reinforced assumption 4)*
There exist non-negative integers s_i, for $i = 1, ..., l$, such that for all θ in the interior of Θ:

$$Rank\ J_i(\theta) = s_1 + s_2 + \cdots + s_i$$

with $J_i(\theta)$ the $[p, (k_1 + k_2 + ... + k_i)]$-matrix $J_i(\theta) = [\partial\rho_1'(\theta)/\partial\theta\ \partial\rho_2'(\theta)/\partial\theta\ \cdots\ \partial\rho_i'(\theta)/\partial\theta]$ and $\sum_{i=1}^{l} s_i = p$.

We are then faced with the following situation:

(i) Only k_1 estimating equations (defined by $\rho_1(\theta)$) have a sample counterpart which converges at the fastest available rate of convergence λ_{1T}. These first k_1 equations can be used in a standard way. Unfortunately, in general the rank of the associated Jacobian J_1 is lower

than the dimension of the parameter space ($s_1 < p$). Thus these estimating equations are not sufficient to identify the entire parameter θ. Intuitively, they only identify s_1 directions in the p-dimensional space of parameters.

(ii) The second group of k_2 estimating equations (defined by $\rho_2(\theta)$) should be used to identify s_2 additional directions. However, this additional identification will come with a slower rate of convergence since $\lambda_{2T} = o(\lambda_{1T})$.

(iii) Now if the total number of identified directions is still lower than the dimension of the parameter space ($s_1 + s_2 < p$), then the third group of estimating equations (defined by $\rho_3(\theta)$) should be used. And so on...

The above discussion helps us understand that the parameter space is going to be separated into several subspaces (as many as the number of groups of moment conditions), each of them collecting directions that will be estimated at a specific rate of convergence. In order to characterize these subspaces, it is natural to define a sequence of matrices $R_i(\theta)$, $i = 1, \cdots, l$, of respective sizes (p, s_i), which will ultimately yield to a change of basis, or in other words a reparametrization.
Before providing their formal definitions, we start with some intuition. The key element consists in realizing that, for a given rate of convergence, the corresponding directions cannot have been identified by a faster group of estimating equations. Going back to the standard GMM theory, the directions identified by a group of estimating equations $\rho_i(\theta)$ correspond to the column space of the associated jacobian $Im[\frac{\partial \rho_i(\theta)}{\partial \theta'}]$. And from some basic matrix theory, we know that the directions not identified by such a group are orthogonal to its column space. The matrices $R_i(\theta)$, $i = 1, \cdots, l$ are then defined by the following backward recursion:

(i) First, since $Rank\ J_{l-1}(\theta) = s_1 + s_2 + \cdots + s_{l-1}$, we can define a matrix $R_l(\theta)$, of size (p, s_l) and of rank s_l, such that its s_l columns are orthogonal to the $(p - s_l)$ columns of $J_{l-1}(\theta)$, that is:

$$\frac{\partial \rho_i(\theta)}{\partial \theta'} R_l(\theta) = 0 \quad \text{for } i < l$$

(ii) Second, since $Rank\ J_{l-2}(\theta) = s_1 + s_2 + \cdots + s_{l-2}$, we can define a matrix $R_{l-1}(\theta)$, of size (p, s_{l-1}) and such that $Rank\,[R_{l-1}(\theta), R_l(\theta)] = s_{l-1} + s_l$, with:

$$\frac{\partial \rho_i(\theta)}{\partial \theta'} R_{l-1}(\theta) = 0 \quad \text{for } \ i < l - 1$$

(iii) And so on. For $j = 2, \cdots, l$, we have:

$$\frac{\partial \rho_i(\theta)}{\partial \theta'} R_j(\theta) = 0 \quad \text{for} \quad i < j$$

$$\text{And with} \quad Rank\left[R_j(\theta)\ R_{j+1}(\theta)\ \cdots\ R_l(\theta)\right] = s_j + s_{j+1} + \cdots + s_l$$

(iv) Finally, we choose $R_1(\theta)$, of size (p, s_1), such that:

$$Rank\left[R_1(\theta)\ R_2(\theta)\ \cdots\ R_l(\theta)\right] = p$$

Note that we do not formally preclude that $s_i = 0$ for some i. If it is the case, just skip the construction of the matrix $R_i(\theta)$. For expositional simplicity, let us consider in all the rest of this section that the true unknown value θ^0 of θ is given and let us note: $R_i(\theta^0) = R_i^0$ and $R^0 = [R_1^0\ R_2^0\ \cdots\ R_l^0]$.

By construction, R^0 is a (p, p) non-singular matrix that we can use for a change of basis in $\mathbb{R}^p$, that is for a new parametrization:

$$\eta = \left[R^0\right]^{-1} \theta = [\eta_i]_{1 \leq i \leq l}, \tag{2.12}$$

with $dim(\eta_i) = s_i$ for $i = 1, \cdots, l$. Of course, this reparametrization is not feasible in practice when we do not know the matrix R^0 but it is worthwhile to contemplate it in order to disentangle the various rates of convergence. More precisely, we must keep in mind that there is no hope in general to ensure that the fast convergence of some components of the estimating equations will produce fast converging estimators of some components of the minimum distance estimator $\hat{\theta}_T$ of θ^0. The reason for this negative result is that, in general, $\hat{\theta}_T$ will be asymptotically equivalent to some linear transformation of $\overline{\phi}_T(\theta)$ and this linear transformation is likely to mix up all the components of $\overline{\phi}_T(\theta)$ in such a way that all the components of $\hat{\theta}_T$ are contaminated by the slow rates of convergence. The advantage of the reparametrization is precisely to isolate the various rates. More precisely, let us consider the reparametrized estimating equations:

$$\rho^*(\eta) = \rho(R^0 \eta) \tag{2.13}$$

First order identification of η comes through the matrix $\partial \rho^*(\eta)/\partial \eta' = \partial \rho(R^0\eta)/\partial \theta' R^0$. This matrix is lower triangular for $\eta = \eta_0 = [R^0]^{-1}\theta^0$ since we have:

$$\text{For} \quad j = 2, 3, \cdots, l \quad \frac{\partial \rho_i(\theta)}{\partial \theta'} R_j(\theta) = 0 \quad \text{for} \quad i < j \tag{2.14}$$

The key idea will be to show that, under some convenient assumptions, this lower triangular property does ensure that for $i = 1, \cdots, l$ $\lambda_{iT}[\hat{\eta}_{iT} - \eta_i^0] = \mathcal{O}_P(1)$: that is the s_i components of the minimum distance estimator $\hat{\eta}_{iT}$ inherit the fast rate of convergence (in the sense faster than λ_{jT} for $j > i$) of the sample-counterparts $\overline{\phi}_{iT}(\theta)$ of the estimating equation $\rho_i(\theta)$.

In this respect, parameters η_j for $j > i$ will be treated as nuisance parameters for the estimation of parameters η_i since they are estimated at slower rates. The situation we are faced with is then rather similar to the one studied in Andrews (1994) for the so-called MINPIN estimators, that are estimators defined as MINimizing a criterion function that might depend on a Preliminary Infinite dimensional Nuisance parameter estimator. Infinite dimensional or not, the nuisance parameters are estimated at slower rates and we want to avoid that their distributions contaminate the asymptotic distribution of the parameters of interest. As Andrews (1994) we then need to ensure some kind of orthogonality between the different kinds of parameters that is analogous to the block diagonality of the information matrix in maximum likelihood contexts. In our framework, the required orthogonality condition will precisely come from the aforementioned lower triangularity of the matrix $\partial\rho^*(\eta^0)/\partial\eta'$. We need however to maintain some additional assumptions to ensure that the relevant sample counterparts of this matrix are well-behaved. To see this, let us consider the infeasible minimum distance problem:

$$\min_{\eta} \overline{\phi}_T'(R^0\eta)\Omega_T\overline{\phi}_T(R^0\eta) \tag{2.15}$$

The first order conditions can be written as:

$$R^{0\prime}\frac{\partial\overline{\phi}_T'(R^0\hat{\eta}_T)}{\partial\theta}\Omega_T\overline{\phi}_T(R^0\eta_T) = 0 \tag{2.16}$$

and the asymptotic distribution of the estimator $\hat{\eta}_T$ can be derived by replacing $T^\gamma\overline{\phi}_T(R^0\hat{\eta}_T)$ in (2.16) by its first order Taylor expansion:

$$T^\gamma\overline{\phi}_T(R^0\eta^0) + T^\gamma\frac{\partial\phi_T(R^0\eta_T^*)}{\partial\theta'}R^0\left[\hat{\eta}_T - \eta^0\right]$$

for some η_T^* defined component by component to be between η^0 and $\hat{\eta}_T$. Then, for the i-th group of components, $i = 1, \cdots, l$, we can write this expansion as:

$$T^\gamma\overline{\phi}_{iT}(R^0\eta^0) + \sum_{j=1}^{l}\frac{T^\gamma}{\lambda_{jT}}\frac{\partial\phi_{iT}(R^0\eta_T^*)}{\partial\theta'}R_j^0\lambda_{jT}\left[\hat{\eta}_{jT} - \eta_j^0\right]$$

It is then clear that when $\lambda_{iT}[\hat{\eta}_{iT} - \eta_i^0] = \mathcal{O}_P(1)$ for $i = 1, \cdots, l$, in order to avoid the contamination of the distribution of fast converging parameters by one of slowly converging ones, we have to ensure that:

$$\frac{T^\gamma}{\lambda_{jT}} \frac{\partial \overline{\phi}_{iT}(R^0 \eta_T^*)}{\partial \theta'} R_j^0 \xrightarrow{P} 0 \text{ when } T \to \infty \quad \text{for all } \; j > i \tag{2.17}$$

Since $\hat{\theta}_T = R^0 \hat{\eta}_T$ is a minimum distance estimator conformable to definition 2.1, we know from theorem 2.2 that $\|\hat{\theta}_T - \theta^0\| = \mathcal{O}(1/\inf_\theta \underline{\lambda}_T(\theta)) = \mathcal{O}(1/\lambda_{lT})$. Hence we also have $\|(R^0 \eta_T^* - R^0 \eta^0\| = \mathcal{O}(1/\lambda_{lT})$. Therefore, the orthogonality condition we do need is:

Orthogonality condition: If θ_T^* is such that $\|\theta_T^* - \theta^0\| = \mathcal{O}(1/\lambda_{lT})$ then for $i = 1, \cdots, l$

$$\frac{T^\gamma}{\lambda_{jT}} \frac{\partial \overline{\phi}_{iT}(\theta_T^*)}{\partial \theta'} R_j^0 \xrightarrow{P} 0 \text{ when } \; T \to \infty \; \text{ for all } \; j > i$$

Note that this orthogonality condition is strikingly similar to the condition (2.12), p49, written by Andrews (1994). Moreover, it is worth noting that this condition is tightly related to the aforementioned lower triangularity of the matrix $\partial \rho^*(\eta^0)/\partial \eta'$. To see this, note that, from assumption 2, we know that:

$$\lambda_{iT} \left[\frac{T^\gamma}{\lambda_{iT}} \overline{\phi}_{iT}(\theta^0) - \rho_i(\theta^0) \right] = \mathcal{O}_P(1)$$

and thus, it is rather natural to assume that:

$$\lambda_{iT} \left[\frac{T^\gamma}{\lambda_{iT}} \frac{\partial \overline{\phi}_{iT}(\theta^0)}{\partial \theta'} - \frac{\partial \rho_i(\theta^0)}{\partial \theta'} \right] = \mathcal{O}_P(1)$$

and thus, for all $j = 1, \cdots, l$:

$$\lambda_{iT} \left[\frac{T^\gamma}{\lambda_{iT}} \frac{\partial \overline{\phi}_{iT}(\theta^0)}{\partial \theta'} R_j^0 - \frac{\partial \rho_i(\theta^0)}{\partial \theta'} R_j^0 \right] = \mathcal{O}_P(1)$$

Hence, the lower triangularity of $\partial \rho^*(\eta^0)/\partial \eta'$ implies that for all $j > i$:

$$\lambda_{iT} \left[\frac{T^\gamma}{\lambda_{iT}} \frac{\partial \overline{\phi}_{iT}(\theta^0)}{\partial \theta'} R_j^0 \right] = \mathcal{O}_P(1),$$

that is:

$$T^\gamma \frac{\partial \overline{\phi}_{iT}(\theta^0)}{\partial \theta'} R_j^0 = \mathcal{O}_P(1) \tag{2.18}$$

A fortiori, since $\lambda_{jT} \to \infty$ when $T \to \infty$,

$$\frac{T^\gamma}{\lambda_{jT}} \frac{\partial \overline{\phi}_{iT}(\theta^0)}{\partial \theta'} R_j^0 \xrightarrow{P} 0 \quad \text{when} \quad T \to \infty \quad \text{for all } j > i \tag{2.19}$$

Note however that this does not exactly give the orthogonality condition we want since the Jacobian matrix should be allowed to be computed at any point θ_T^* such that $\|\theta_T^* - \theta^0\| = \mathcal{O}_P(1/\lambda_{lT})$ and not only at the true value θ^0. In order to control for that, we have basically to ensure that the Jacobian matrix is itself continuously differentiable ($\overline{\phi}_T(\theta)$ is twice continuously differentiable) and that higher order terms coming from the slowest rate (λ_{lT}) are not detrimental because at least λ_{lT}^2 goes to infinity faster than the fastest rate, that is (λ_{1T}). All these considerations lead us to maintain the following assumption:

Assumption 6 *(i) For all $i = 1, \cdots, l$:*

$$\frac{\partial \Psi'_{iT}(\theta^0)}{\partial \theta} = T^\gamma \left[\frac{\partial \overline{\phi}'_{iT}(\theta^0)}{\partial \theta} - \frac{\lambda_{iT}}{T^\gamma} \frac{\partial \rho'_i(\theta^0)}{\partial \theta} \right] = \mathcal{O}_P(1)$$

(ii) $\overline{\phi}_T(\theta)$ is twice continuously differentiable on the interior of Θ and for all $i = 1, \cdots, l$ and for each component $1 \leq k \leq k_i$:

$$\frac{T^\gamma}{\lambda_{iT}} \frac{\partial^2 \overline{\phi}_{iT,k}(\theta)}{\partial \theta \partial \theta'} \xrightarrow{P} H(\theta)$$

uniformly on θ in some neighborhood of θ^0, for some (p, p) matricial function $H(\theta)$.
(iii) $\lambda_{1T} = o(\lambda_{lT}^2)$

The third condition of assumption 6 basically says that even though the sample counterparts of the estimating equations converge at different rates, the discrepancy of the rates is not so large that the squared of the slowest rate would not exceed the fastest rate. This condition makes our approach radically different from two potential applications: first, from the standard context of weak identification. See also the companion paper by Antoine and Renault (2007) which considers specifically the application with nearly-weak identification; second, from the concept of under-identification *à la* Sargan (1983). In the latter case, slow rate means $T^{1/4}$ while fast rate is the usual $T^{1/2}$ one. Then it can be shown (see also Dovonon and Renault (2006)) that even though the slower directions do not slow down the faster ones ($T^{1/4}$ squared is not smaller

than $T^{1/2}$), we have however a contamination of the distribution of the fast estimator by the one of the slow estimator, precisely because $T^{1/4}$ squared is not strictly larger than $T^{1/2}$. As already mentioned, our context of interest is much more analogous to Andrews MINPIN estimators. Interestingly enough, Andrews (1995) provided for his necessary orthogonality conditions a set of sufficient conditions (see Andrews (1995) p563) precisely saying that when the parameters of interest are estimated at rate $T^{1/2}$, the nuisance parameters must be estimated at a rate faster than $T^{1/4}$.

As explained above, assumption 6 has been conceived to keep the lower triangularity of the matrix $\partial\rho^*(\eta^0)/\partial\eta' = \partial\rho(\theta^0)/\partial\theta' R^0$ when computing it from sample counterparts. We can actually show:

Lemma 2.3 *Under assumptions 1 to 6, if θ_T^* is such that $\|\theta_T^* - \theta^0\| = \mathcal{O}_P(1/\lambda_{lT})$, then*

$$T^\gamma \frac{\partial \overline{\phi}_T(\theta_T^*)}{\partial \theta'} R^0 \tilde{\Lambda}_T^{-1} \xrightarrow{P} J^0 \quad \textit{when } T \to \infty$$

Where J^0 is the (K,p) block diagonal matrix with diagonal blocks $\partial\rho_i(\theta^0)/\partial\theta' R_i^0$ and $\tilde{\Lambda}_T$ is the (p,p) diagonal matrix defined as

$$\tilde{\Lambda}_T = \begin{pmatrix} \lambda_{1T} Id_{s_1} & & & \\ & \lambda_{2T} Id_{s_2} & & O \\ O & & \ddots & \\ & & & \lambda_{lT} Id_{s_l} \end{pmatrix}$$

$$\textit{with} \qquad \begin{array}{l} \sum_{i=1}^{l} s_i = p \\ \lim_{T\to\infty} \lambda_{iT} = \infty \quad \textit{for} \quad i = 1, ..., l \\ \lambda_{i+1,T} = o(\lambda_{i,T}) \quad \textit{for} \quad i = 1, ..., l-1 \end{array}$$

2.3 Efficient GMM estimation

Even though a more general theory would be easy to write down, we are going to specialize throughout the rest of the paper the asymptotic theory of the minimum distance estimator

$\hat{\theta}_T$ defined by (2.5) to the context of square root-consistent asymptotically normal sample counterparts of estimating equations:

Assumption 7 *For θ^0 true unknown value of θ, $\Psi_T(\theta^0) = T^{1/2}\overline{\phi}_T(\theta^0)$ converges in distribution towards a normal distribution with mean zero and variance $S(\theta^0)$.*

Note that assumption 7 implies that γ introduced in assumption 2 will be considered with the special value $(1/2)$ hereafter. Up to unusual rates of convergence, we then get a standard asymptotic normality result for the new parameters $\eta = [R^0]^{-1}\theta$:

Theorem 2.4 *(Asymptotic Normality)*
Under assumptions 1 to 7, the minimum distance estimator $\hat{\theta}_T$ defined by (2.5) is such that:

$$\tilde{\Lambda}_T[R^0]^{-1}\left(\hat{\theta}_T - \theta^0\right) \xrightarrow{d} \mathcal{N}\left(0, \left[J^{0\prime}\Omega J^0\right]^{-1} J^{0\prime}\Omega S(\theta^0)\Omega J^0 \left[J^{0\prime}\Omega J^0\right]^{-1}\right)$$

It is worth noting that this result has strong similarities with the Hansen (1982) classical result about the asymptotic distribution of GMM. To see this, first note that the matrix J^0 may almost be interpreted as $\partial\rho(\theta^0)/\partial\theta' R^0 = \partial\rho^*(\eta^0)/\partial\eta'$ where $\rho^*(\eta) = \rho(R^0\theta)$. This simple interpretation is actually not fully correct since, while $\partial\rho^*(\eta^0)/\partial\eta'$ is a lower-triangular matrix: due to the discrepancy between rates of convergence, the upper diagonal blocks are canceled out in the limit considered in lemma 2.3, in such a way that J^0 is a block-diagonal matrix. However, seeing J^0 as $\partial\rho^*(\eta^0)/\partial\eta'$ would allow one to interpret the asymptotic variance in theorem 2.4 as the standard asymptotic variance of a minimum distance estimator computed from the minimization problem:

$$\min_{\eta}\left[\phi_T'(R^0\eta)\Omega_T\phi_T(R^0\eta)\right]$$

In particular, the cancelation of upper-diagonal blocks does not invalidate the standard argument that the optimal weighting matrix should be a consistent estimator of the inverse of the long term variance matrix $S(\theta^0)$:

Theorem 2.5 *Under assumptions 1 to 7, the asymptotic variance displayed in theorem 2.4 is minimal when the minimum distance estimator $\hat{\theta}_T$ is defined by (2.5) with a weighting matrix Ω_T consistent estimator of $\Omega = [S(\theta^0)]^{-1}$. Then,*

$$\tilde{\Lambda}_T[R^0]^{-1}\left(\hat{\theta}_T - \theta^0\right) \xrightarrow{d} \mathcal{N}\left(0, \left[J^{0\prime}[S(\theta^0)]^{-1}J^0\right]^{-1}\right)$$

Note that a consistent estimator S_T of the long-term covariance matrix $S(\theta^0)$ can be constructed in the standard way (see in general Hall (2005)) from a preliminary inefficient GMM estimator θ.

Then, up to the block-diagonality of the matrix J^0 as mentioned above, we get the standard formula for the asymptotic distribution of an efficient minimum distance estimator of η. For the same reason, the standard overidentification test can be performed in the usual way.

Theorem 2.6 *(J-test)*
Under assumptions 1 to 7, if Ω_T is a consistent estimator of $\left[S(\theta^0)\right]^{-1}$,

$$TQ_T(\hat{\theta}_T) \xrightarrow{d} \chi^2_{K-p}$$

By analogy with standard GMM, the block-diagonality of J^0 means that, as far as the new subsets of parameters η_i, $1 \leq i \leq l$, are concerned, all the GMM-like formulas look like if each subset of estimating equations

$$Plim_{T=\infty}\left[\frac{\sqrt{T}}{\lambda_{iT}}\overline{\phi}_{iT}(R^0\eta) - \rho_i(R^0\eta)\right] = 0 \quad i = 1, \cdots, l$$

depends on the unknown parameters η only through η_i. As well known for GMM theory, this does not imply that efficient estimation of the various subsets can be performed independently. This would be the case only when the long-term covariance matrix $S(\theta^0)$ is itself block-diagonal with the same partition of indices. This means that we would have zero long-term covariance between sample counterparts of estimating equations $\overline{\phi}_{iT}(\theta^0)$ and $\overline{\phi}_{jT}(\theta^0)$ with $i \neq j$. As it will be confirmed by several important examples in section 4 below, this is a quite natural situation. To see this, let us imagine that for some $i < j$, $\overline{\phi}_{iT}(\theta^0)$ and $\overline{\phi}_{jT}(\theta^0)$ have a non-zero long-term covariance. For sake of notational simplicity, let us just consider them as univariate. Then the non-zero covariance allows us to find number a such that

$$\overline{\phi}^*_{iT}(\theta^0) = \overline{\phi}_{iT}(\theta^0) + a\overline{\phi}_{jT}(\theta^0)$$

has a long-term covariance matrix $S^*_i(\theta^0) < S_i(\theta^0)$. However, $\overline{\phi}^*_{iT}(\theta)$ has the same identifying power about $\rho_i(\theta)$ than $\overline{\phi}_{iT}(\theta^0)$ since the translation by $[a(\lambda_{jT}/T^{1/2})\rho_j(\theta)]$ is negligible in front of the main term $[(\lambda_{iT}/T^{1/2})\rho_i(\theta)]$. In other words, we find ourselves in a rather paradoxical situation where the slowly converging moments bring relevant information about

the estimating equations whose sample counterparts have a faster rate. This is the reason why we expect in practice the above block-diagonality of $S(\theta^0)$ to be fulfilled more often than not, at least when the estimating equations and their sample counterparts are endowed with some structural interpretation. The theorem 2.5 actually confirms the intuition above: when $S(\theta^0)$ is block-diagonal, the asymptotic covariance of the minimum distance estimator $\hat{\eta}_T = [R^0]^{-1}\hat{\theta}_T$ is block-diagonal in such a way that the fast and slowly consistent directions have asymptotically independent estimators.

In general, our focus of interest is not the vector η of new parameters but the vector θ of initial ones. As far as inference about θ is concerned, several practical implications of theorem 2.5 are worth mentioning. From the lemma 2.3 a consistent estimator of the asymptotic covariance matrix $\left[J^{0\prime}[S(\theta^0)]^{-1}J^0\right]^{-1}$ is:

$$\begin{aligned} & T^{-1}\left[\tilde{\Lambda}_T^{-1}R^{0\prime}\frac{\partial\overline{\phi}_T'(\hat{\theta}_T)}{\partial\theta}S_T^{-1}\frac{\partial\overline{\phi}_T(\hat{\theta}_T)}{\partial\theta'}R^0\tilde{\Lambda}_T^{-1}\right]^{-1} \qquad (2.20)\\ = \;& T^{-1}\tilde{\Lambda}_T[R^0]^{-1}\left[\frac{\partial\overline{\phi}_T'(\hat{\theta}_T)}{\partial\theta}S_T^{-1}\frac{\partial\overline{\phi}_T(\hat{\theta}_T)}{\partial\theta'}\right]^{-1}[R^{0\prime}]^{-1}\tilde{\Lambda}_T \end{aligned}$$

where S_T is a standard consistent estimator of the long-term covariance matrix. Note that we do not address the estimation of the matrix R^0 at this stage. It is actually worth realizing the knowledge of R^0 is not really necessary due to the following intuition. Since theorem 2.5 tells us that for large T, $\tilde{\Lambda}_T[R^0]^{-1}(\hat{\theta}_T - \theta^0)$ behaves like a gaussian random variable with mean zero and variance (2.20), we can informally say that $\sqrt{T}(\hat{\theta}_T - \theta^0)$ behaves like a gaussian with mean zero and variance

$$\left[\frac{\partial\overline{\phi}_T'(\hat{\theta}_T)}{\partial\theta}S_T^{-1}\frac{\partial\overline{\phi}_T(\hat{\theta}_T)}{\partial\theta'}\right]^{-1} \qquad (2.21)$$

Saying this gives the feeling that we are back to standard GMM formulas of Hansen (1982). This intuition is correct for all practical purposes; in particular the knowledge of the change of basis R^0 is not necessary for inference. It must however be stressed that the above intuition, albeit practically relevant, is theoretically a bit misleading for several reasons.
First, in general, all components of $\hat{\theta}_T$ converge slowly towards θ^0 and thus $\sqrt{T}(\hat{\theta}_T - \theta^0)$ has no limit distribution. When we say that it is approximately a gaussian with variance (2.21) it must be realized that since

$$\frac{\sqrt{T}}{\lambda_{iT}}\frac{\partial\overline{\phi}_{iT}(\hat{\theta}_T)}{\partial\theta'} \xrightarrow{P} \frac{\partial\rho_i(\theta^0)}{\partial\theta'}$$

we actually have

$$\frac{\partial \overline{\phi}_{iT}(\hat{\theta}_T)}{\partial \theta'} \xrightarrow{P} 0$$

as soon as $i > 1$. In other words, considering the asymptotic variance (2.21) is akin to consider the inverse of an asymptotically singular matrix.
Second, for the same reason, it must also be realized that (2.21) is not an estimator of the standard population matrix

$$\left[\frac{\partial \rho'(\theta^0)}{\partial \theta}[S(\theta^0)]^{-1}\frac{\partial \rho(\theta^0)}{\partial \theta'}\right] \tag{2.22}$$

Typically, beyond the singularity issue mentioned above, the population matrix (2.22) will not display in general the right block-diagonality structure.

For all these reasons, inference about θ is actually more involved than one may believe at first sight from the apparent similarity with standard GMM formulas.

3 Inference

- Results on the equivalence of tests
- Constrained estimation
- Estimation of the subspaces (when more than 2 speeds)

TO BE COMPLETED

4 Examples

4.1 Conditional moment restrictions

Gagliardini, Gouriéroux and Renault (2004) propose the extended method of moments where they consider two kinds of moment restrictions. First, an unconditional set (possibly produced by some conditional restrictions with a choice of instruments) where the sample counterpart, $\overline{\phi}_{1T}(\theta)$, is such that:

$$\sqrt{T}\left[\overline{\phi}_{1T}\theta - \rho_1(\theta)\right] \Longrightarrow \Psi_1(\theta) \quad \text{a Gaussian process.}$$

Second, a conditional set defined at a specific point in time (that is a specific conditioning environment) where the "sample (rescaled kernel) counterpart", $\overline{\phi}_{2T}(\theta)$, is such that:

$$\sqrt{T}\left[\overline{\phi}_{2T}(\theta) - h_T^{1/2}\rho(\theta)\right] \Longrightarrow \Psi_2(\theta) \quad \text{a Gaussian process.}$$

where h_T is the bandwidth parameter. Our framework can easily incorporate such a setting by choosing:

$$\Lambda_T = \begin{bmatrix} \sqrt{T}Id_1 & 0 \\ 0 & \sqrt{Th_T}Id_2 \end{bmatrix}$$

and defining $\overline{\phi}_{2T}$ as root-h_T times the kernel estimator.

4.2 Other examples

Lee (2004) considers a social interactions model to motivate the following minimum distance estimator:

$$\min_{\theta} \left[\alpha_T - f(\theta)\right]' \Omega_T \left[\alpha_T - f(\theta)\right]$$

with $\Lambda_T\left(\alpha_T - \alpha^0\right) \xrightarrow{d} \mathcal{N}(0, V)$ for some positive definite matrix V. Here again, there is no uniformity issue with respect to θ since θ does not enter the sample dependent counterpart of the estimating equations: $\overline{\phi}_T(\theta)$ corresponds here to $\frac{\Lambda_T}{\sqrt{T}}\alpha_T$.

Kotlyarova and Zinde-Walsh (2006) consider pooling kernel-based estimators corresponding to different choices of bandwidths. This allows them to propose a combined estimator with possibly the best available rate without knowledge of density smoothness.

TO BE COMPLETED

- Example with 3 speeds
- Unit-root example
- Extreme values example (Robert)

References

[1] D. Andrews, *Asymptotics for Semiparametric Econometric Models via Stochastic Equicontinuity*, Econometrica **62** (1994), 43–72.

[2] ———, *Nonparametric Kernel Estimation for Semiparametric Econometric Models*, Econometric Theory **11** (1995), 560–596.

[3] B. Antoine, *On the Inference on Parameter Ratios with Applications to Weak Identification*, Working Paper (2006).

[4] M. Caner, *Testing, Estimation and Higher Order Expansions in GMM with Nearly-Weak Instruments*, Working Paper (2005).

[5] P. Gagliardini, C. Gouriéroux, and E. Renault, *Efficient Derivative Pricing by Extended Methods of Moments*, Working Paper (2005).

[6] J. Hahn and G. Kuersteiner, *Discontinuities of Weak Instruments limiting Distributions*, Economics Letters **75** (2002), 325–331.

[7] A.R. Hall, *Generalized Method of Moments*, Advanced Texts in Econometrics, Oxford University Press, 2005.

[8] L.P. Hansen, *Large Sample Properties of Generalized Method of Moments Estimators*, Econometrica **50** (1982), no. 4, 1029–1054.

[9] R.I. Jennrich, *Asymptotic Properties of Non-linear Least Squares Estimators*, Annals of Mathematical Statistics **40** (1969), 633–43.

[10] M. Kessler, *Estimation of an Ergodic Diffusion from Discrete Observations*, Scandinavian Journal of Statistics **24** (1997), 211–29.

[11] Y. Kotlyarova and V. Zinde-Walsh, *Non and Semi-parametric Estimation in Models with unknown Smoothness*, forthcoming Economic Letters (2006).

[12] L. Lee, *Pooling Estimators with Different Rates of Convergence - A minimum χ^2 Approach with an emphasis on a Social Interaction Model*, Working Paper (2004).

[13] P.Dovonon and E. Renault, *GMM Overidentification Test with First-Order Unidentification*, Working Paper (2006).

[14] J.D. Sargan, *Identification and Lack of Identification*, Econometrica **51** (1983), no. 6, 1605–1634.

[15] D. Staiger and J. Stock, *Instrumental Variables Regression with Weak instruments*, Econometrica **65** (1997), 557–586.

[16] J.H. Stock and J.H. Wright, *GMM with Weak Identification*, Econometrica **68** (2000), no. 5, 1055–1096.

Appendix

A Proofs of the main results

Proof of Equation (2.3): (*Stronger identification property*)
Let us denote by S_ϵ the set of $\theta \in \Theta$ such that $\|\theta - \theta^0\| \geq \epsilon$. Since it is compact, the identification assumption 1 with the continuity of $\rho(.)$ implies that the minimum of $\|\rho(\theta)\|$ on this set is $\delta > 0$. ■

Proof of Theorem 2.1: *(Consistency)*
The consistency of the minimum distance estimator $\hat{\theta}_T$ is a direct implication of the identification assumption 1 jointly with the following lemma:

Lemma A.1

$$\|\rho(\hat{\theta}_T)\| = \mathcal{O}_P\left(\frac{1}{\inf_{\theta\in\Theta} \underline{\lambda}_T(\theta)}\right)$$

Proof of lemma A.1: From (2.7), the objective function is written as follows

$$Q_T(\theta) = \left[\frac{\Psi_T(\theta)}{T^\gamma} + \frac{\Lambda_T(\theta)}{T^\gamma}\rho(\theta)\right]' \Omega_T \left[\frac{\Psi_T(\theta)}{T^\gamma} + \frac{\Lambda_T(\theta)}{T^\gamma}\rho(\theta)\right]$$

Since $\hat{\theta}_T$ is the minimizer of $Q(.)$ we have in particular:

$$\begin{aligned} & Q_T(\hat{\theta}_T) \leq Q(\theta^0) \\ \Longrightarrow \quad & \left[\frac{\Psi_T(\hat{\theta}_T)}{T^\gamma} + \frac{\Lambda_T(\hat{\theta}_T)}{T^\gamma}\rho(\hat{\theta}_T)\right]' \Omega_T \left[\frac{\Psi_T(\hat{\theta}_T)}{T^\gamma} + \frac{\Lambda_T(\hat{\theta}_T)}{T^\gamma}\rho(\hat{\theta}_T)\right] \leq \frac{\Psi_T'(\theta^0)}{T^\gamma}\Omega_T\frac{\Psi_T(\theta^0)}{T^\gamma} \end{aligned}$$

Denoting $d_T = \Psi_T'(\hat{\theta}_T)\Omega_T\Psi_T(\hat{\theta}_T) - \Psi_T'(\theta^0)\Omega_T\Psi_T(\theta^0)$, we get:

$$\left[\Lambda_T(\hat{\theta}_T)\rho(\hat{\theta}_T)\right]' \Omega_T \left[\Lambda_T(\hat{\theta}_T)\rho(\hat{\theta}_T)\right] + 2\left[\Lambda_T(\hat{\theta}_T)\rho(\hat{\theta}_T)\right]' \Omega_T\Psi_T(\hat{\theta}_T) + d_T \leq 0$$

Let μ_T be the smallest eigenvalue of Ω_T. The former inequality implies:

$$\mu_T\|\Lambda_T(\hat{\theta}_T)\rho(\hat{\theta}_T)\|^2 - 2\|\Lambda_T(\hat{\theta}_T)\rho(\hat{\theta}_T)\| \times \|\Omega_T\Psi_T(\hat{\theta}_T)\| + d_T \leq 0$$

In other words, $x_T = \|\Lambda_T(\hat{\theta}_T)\rho(\hat{\theta}_T)\|$ solves the inequality:

$$x_T^2 - \frac{2\|\Omega_T\Psi_T(\hat{\theta}_T)\|}{\mu_T}x_T + \frac{d_T}{\mu_T} \leq 0$$

and thus with

$$\Delta_T = \frac{\|\Omega_T\Psi_T(\hat{\theta}_T)\|^2}{\mu_T^2} - \frac{d_T}{\mu_T}$$

we have:

$$\frac{\|\Omega_T\Psi_T(\hat{\theta}_T)\|}{\mu_T} - \sqrt{\Delta_T} \leq x_T \leq \frac{\|\Omega_T\Psi_T(\hat{\theta}_T)\|}{\mu_T} + \sqrt{\Delta_T}$$

Since $x_T \geq (\inf_{\theta\in\Theta} \underline{\lambda}_T(\theta)) \|\rho_T(\hat{\theta}_T)\|$ we want to show that $x_T = \mathcal{O}_P(1)$ that is

$$\frac{\|\Omega_T\Psi_T(\hat{\theta}_T)\|}{\mu_T} = \mathcal{O}_P(1) \text{ and } \Delta_T = \mathcal{O}_P(1)$$

which amounts to show that:

$$\frac{\|\Omega_T\Psi_T(\hat{\theta}_T)\|}{\mu_T} = \mathcal{O}_P(1) \text{ and } \frac{d_T}{\mu_T} = \mathcal{O}_P(1)$$

Note that since $det(\Omega_T) \xrightarrow{P} det(\Omega) > 0$ no subsequence of Ψ_T can converge in probability towards zero and thus we can assume (for T sufficiently large) that μ_T remains lower bounded away from zero with asymptotic probability one. Therefore, we just have to show that:

$$\|\Omega_T\Psi_T(\hat{\theta}_T)\| = \mathcal{O}_P(1) \text{ and } d_T = \mathcal{O}_P(1)$$

We note that since $Tr(\Omega_T) \xrightarrow{P} Tr(\Omega)$ and the sequence $Tr(\Omega_T)$ is upper bounded is probability and so are all the eigenvalues of Ω_T. Therefore the required boundedness in probability just results from our tightness assumption 2 ensuring that:

$$\sup_{\theta\in\Theta} \|\Psi_T(\theta)\| = \mathcal{O}_P(1)$$

The proof of lemma A.1 is completed. Let us then deduce the weak consistency of $\hat{\theta}_T$ by a contradiction argument. If $\hat{\theta}_T$ was not consistent, there would exist some positive ϵ such that:

$$P\left[\|\hat{\theta}_T - \theta^0\| > \epsilon\right]$$

does not converge to zero. Then we can define a subsequence $(\hat{\theta}_{T_n})_{n\in\mathbb{N}}$ such that, for some positive η:

$$P\left[\|\hat{\theta}_T - \theta^0\| > \epsilon\right] \geq \eta \text{ for } n \in \mathbb{N}$$

Let us denote

$$\alpha = \inf_{\|\theta-\theta^0\|>\epsilon} \|\rho(\theta)\| > 0 \text{ by assumption1}$$

Then for all $n \in \mathbb{N}$:

$$P\left[\|\rho(\hat{\theta}_{T_n})\| \geq \alpha\right] > 0$$

When considering the identification assumption 3, this last inequality contradicts lemma $A.1$. This completes the proof of consistency. ■

Proof of Theorem 2.2: *(Rate of convergence)*
From (2.8) $\|\rho(\hat{\theta}_T)\| = \|\rho(\hat{\theta}_T) - \rho(\theta^0)\| = \mathcal{O}_P(1/\inf_{\theta\in\Theta} \underline{\lambda}_T(\theta))$ and by application of the Mean-Value theorem, we get for some $\tilde{\theta}_T$ between $\hat{\theta}_T$ and θ^0 we get:

$$\left\| \frac{\partial \rho(\tilde{\theta}_T)}{\partial \theta'} \left(\hat{\theta}_T - \theta^0\right) \right\| = \mathcal{O}_P\left(\frac{1}{\inf_{\theta\in\Theta} \underline{\lambda}_T}\right)$$

Note that, by a common abuse of notation, we omit to stress that $\tilde{\theta}_T$ actually depends on the cpmponent of $\rho(.)$. The key point is that since $\rho(.)$ is continuously differentiable and $\tilde{\theta}_T$, as $\hat{\theta}_T$, converges in probability towards θ_0, we have:

$$\frac{\partial \rho(\tilde{\theta}_T)}{\partial \theta'} \xrightarrow{P} \frac{\partial \rho(\theta^0)}{\partial \theta'}$$

and thus:

$$\frac{\partial \rho(\theta^0)}{\partial \theta'} \times \left(\hat{\theta}_T - \theta^0\right) = z_T$$

with $\|z_T\| = \mathcal{O}_P(1/\inf_{\theta\in\Theta} \underline{\lambda}_T(\theta))$. Since $\partial\rho(\theta^0)/\partial\theta'$ is full column rank, we deduce that:

$$\left(\hat{\theta}_T - \theta^0\right) = \left[\frac{\partial \rho'(\theta^0)}{\partial \theta} \frac{\partial \rho(\theta^0)}{\partial \theta'}\right]^{-1} \frac{\partial \rho'(\theta^0)}{\partial \theta} z_T$$

also fulfills:

$$\left\|\hat{\theta}_T - \theta^0\right\| = \mathcal{O}_P\left(\frac{1}{\inf_{\theta\in\Theta} \underline{\lambda}_T(\theta)}\right)$$

■

Proof of Lemma 2.3:

To get the results, we have to show the following:

$$
\begin{aligned}
&i) \quad \text{(diagonal terms)} \ \frac{T^\gamma}{\lambda_{iT}} \frac{\partial \overline{\phi}_{iT}(\theta_T^*)}{\partial \theta'} \xrightarrow{P} \frac{\partial \rho_i(\theta^0)}{\partial \theta'} \quad i = 1, \cdots, l \\
&ii) \quad \text{lower diagonal} \ \frac{T^\gamma}{\lambda_{jT}} \frac{\partial \overline{\phi}_{iT}(\theta_T^*)}{\partial \theta'} \xrightarrow{P} 0 \quad i = 2, \cdots, l; j = 1, \cdots, l-1; j < i \\
&iii) \quad \text{upper diagonal} \ \frac{T^\gamma}{\lambda_{jT}} \frac{\partial \overline{\phi}_{jT}(\theta_T^*)}{\partial \theta'} R_i^0 \xrightarrow{P} 0 \quad i = 1, \cdots, l-1; j = 2, \cdots, l; j > i
\end{aligned}
$$

i) From Assumption 6i)

$$
T^\gamma \frac{\partial \overline{\phi}'_{iT}(\theta^0)}{\partial \theta} - \lambda_{iT} \frac{\partial \rho'_i(\theta^0)}{\partial \theta} = \mathcal{O}_P(1)
$$

A fortiori since $\lambda_{iT} \xrightarrow{T} \infty$:

$$
\frac{T^\gamma}{\lambda_{iT}} \frac{\partial \overline{\phi}_{iT}(\theta_T^*)}{\partial \theta'} - \frac{\partial \rho_i(\theta^0)}{\partial \theta'} \xrightarrow{P} 0
$$

The mean-value theorem applied to the kth component of $\partial \overline{\phi}_{iT}/\partial \theta'$, $1 \leq k \leq k_i$ gives, for $\tilde{\theta}_T$ in between θ^0 and θ_T^*:

$$
\frac{T^\gamma}{\lambda_{iT}} \left(\frac{\partial \overline{\phi}_{iT,k}(\theta_T^*)}{\partial \theta'} - \frac{\partial \overline{\phi}_{iT,k}(\theta^0)}{\partial \theta'} \right) = \frac{T^\gamma}{\lambda_{iT}} (\theta^* - \theta^0)' \frac{\partial^2 \overline{\phi}_{iT,k}(\tilde{\theta}_T)}{\partial \theta \partial \theta'} = o_P(1)
$$

because by assumption $\|\theta_T^* - \theta^0\| = \mathcal{O}_P(1/\lambda_{lT})$ and by assumption 6(ii) $T^\gamma/\lambda_{iT} \partial^2 \overline{\phi}_{iT,k}(\theta)/\partial \theta \partial \theta' \xrightarrow{P} H(\theta)$. Hence we get the announced result i).

ii)

$$
i > j \quad \frac{T^\gamma}{\lambda_{jT}} \frac{\partial \overline{\phi}_{iT}(\theta_T^*)}{\partial \theta'} R_j = \frac{\lambda_{iT}}{\lambda_{jT}} \times \frac{T^\gamma}{\lambda_{iT}} \frac{\partial \overline{\phi}_{iT}(\theta_T^*)}{\partial \theta'} R_j^0 \xrightarrow{P} 0
$$

because $\lambda_{iT} = o_P(\lambda_{jT})$ and $\partial \overline{\phi}_{iT}(\theta_T^*)/\partial \theta' \xrightarrow{P} \partial \rho_i(\theta^0)/\partial \theta'$.

iii) Again let apply the mean-value theorem to the kth component of $\partial \overline{\phi}_{jT(.)}/\partial \theta' R_j^0$ for $1 \leq k \leq k_i$, with $\tilde{\theta}_T$ between θ^0 and θ_T^*:

$$
\begin{aligned}
& \frac{T^\gamma}{\lambda_{jT}} \frac{\partial \overline{\phi}_{iT,k}(\theta_T^*)}{\partial \theta'} R_j^0 \\
= \ & \frac{1}{\lambda_{jT}} \times \left[\frac{T^\gamma \partial \overline{\phi}_{iT,k}(\theta^0)}{\partial \theta'} R_j^0 \right] + \lambda_{lT} (\theta_T^* - \theta^0)' \frac{\lambda_{iT}}{\lambda_{jT} \lambda_{lT}} \frac{T^\gamma}{\lambda_{iT}} \frac{\partial^2 \overline{\phi}_{iT,k}(\tilde{\theta}_T)}{\partial \theta \partial \theta'} R_j^0
\end{aligned}
$$

Now recall that $\lambda_{lT}\|(\theta_T^*-\theta^0)\| = \mathcal{O}_P(1)$; $\frac{T^\gamma}{\lambda_{iT}}\partial^2\overline{\phi}_{iT,k}(\theta)/\partial\theta\partial\theta' \xrightarrow{P} H(\theta)$; and also $\lambda_{iT}/(\lambda_{jT}\lambda_{lT}) = \lambda_{iT}/\lambda_{lT}^2 \times \lambda_{lT}/\lambda_{jT} \xrightarrow{T} 0$ by assumption 6iii).
We just have to prove that the first element of the RHS converges to 0 in probability. From assumption 6i), we have:

$$\frac{T^\gamma}{\lambda_{jT}}\left[\frac{\partial\overline{\phi}_{iT,k}(\theta^0)}{\partial\theta'}R_j^0 - \frac{\partial\rho_i'(\theta^0)}{\partial\theta'}R_j^0\right] = \mathcal{O}_P\left(\frac{1}{\lambda_{iT}}\right)$$

and we get the result because $\partial\rho_i(\theta^0)/\partial\theta' R_j^0 = 0$ by definition of R^0.

Proof of Theorem 2.4: *(Asymptotic Normality)*
From the optimization problem (2.5), the first order conditions for $\hat{\theta}_T$ are written as:

$$\frac{\partial\overline{\phi}_T'(\hat{\theta}_T)}{\partial\theta}\Omega\overline{\phi}_T(\hat{\theta}_T) = 0$$

A mean-value expansion yields to:

$$\frac{\partial\overline{\phi}_T'(\hat{\theta}_T)}{\partial\theta}\Omega\overline{\phi}_T(\theta^0) + \frac{\partial\overline{\phi}_T'(\hat{\theta}_T)}{\partial\theta}\Omega\frac{\partial\overline{\phi}_T(\tilde{\theta}_T)}{\partial\theta'} \times \left(\hat{\theta}_T - \theta^0\right) = 0$$

where $\tilde{\theta}_T$ is between $\hat{\theta}_T$ and θ^0. Premultiplying the above equation by the non-singular matrix $T\tilde{\Lambda}_T^{-1}R^{0\prime}$ yields to an equivalent set of equations:

$$\hat{J}_T'\Omega\left[\sqrt{T}\overline{\phi}_T(\theta^0)\right] + \hat{J}_T'\Omega\tilde{J}_T \times \Lambda_T R^{-1}\left(\hat{\theta}_T - \theta^0\right) = 0$$

after defining:

$$\hat{J}_T = \sqrt{T}\frac{\partial\overline{\phi}_T(\hat{\theta}_T)}{\partial\theta'}R^0\Lambda_T^{-1} \quad \text{and} \quad \tilde{J}_T = \sqrt{T}\frac{\partial\overline{\phi}_T(\tilde{\theta}_T)}{\partial\theta}R^0\Lambda_T^{-1}$$

From theorem 2.2 and lemma 2.3, we can deduce that:

$$plim\tilde{J}_T = J^0 \quad \text{and} \quad plim\hat{J}_T = J^0$$

Hence,

$$\hat{J}_T'\Omega\tilde{J}_T \xrightarrow{P} J^{0\prime}\Omega J^0 \quad \text{nonsingular by assumption}$$

Recall now that by assumption 7), $\Psi_T(\theta^0) = \sqrt{T}\left[\overline{\phi}_T(\theta^0)\right]$ converges to a normal distribution with mean 0 and variance $S(\theta^0)$. We then get the announced result.■

Proof of Theorem 2.6: *(Overidentifying test)*
A Taylor expansion of order 1 of the moment conditions gives:

$$\begin{aligned}\sqrt{T}\overline{\phi}_T(\hat{\theta}_T) &= \sqrt{T}\overline{\phi}_T(\theta^0) + \sqrt{T}\frac{\partial\overline{\phi}_T(\hat{\theta}_T)}{\partial\theta'}(\hat{\theta}_T - \theta^0) + o_P(1) \\ &= \sqrt{T}\overline{\phi}_T(\theta^0) + \hat{J}_T\tilde{\Lambda}_T[R^0]{-}1(\hat{\theta}_T - \theta^0) + o_P(1)\end{aligned}$$

with $\hat{J}_T = \sqrt{T}\partial\overline{\phi}_T(\hat{\theta}_T)/\partial\theta' R^0\tilde{\Lambda}_T^{-1}$.

A Taylor expansion of the FOC gives:

$$\begin{aligned}\tilde{\Lambda}_T[R^0]^{-1}(\hat{\theta}_T - \theta^0) &= -\left[\left(\sqrt{T}\frac{\partial\overline{\phi}_T(\hat{\theta}_T)}{\partial\theta'}R^0\tilde{\Lambda}_T^{-1}\right)' S_T^{-1}\left(\sqrt{T}\frac{\partial\overline{\phi}_T(\hat{\theta}_T)}{\partial\theta'}R^0\tilde{\Lambda}_T^{-1}\right)\right]^{-1} \\ &\quad\times\left(\sqrt{T}\frac{\partial\overline{\phi}_T(\hat{\theta})}{\partial\theta'}R^0\tilde{\Lambda}_T^{-1}\right)' S_T^{-1}\sqrt{T}\overline{\phi}_T(\theta^0) + o_P(1)\end{aligned}$$

with S_T a consistent estimator of the asymptotic covariance matrix of the process $\Psi(\theta)$.

Combining the 2 above results leads to:

$$\sqrt{T}\overline{\phi}_T(\hat{\theta}_T) = \sqrt{T}\overline{\phi}_T(\theta^0) - \hat{J}_T\left[\hat{J}_T' S_T^{-1}\hat{J}_T\right]^{-1}\hat{J}_T' S_T^{-1}\sqrt{T}\overline{\phi}_T(\theta^0) + o_P(1)$$

Use the previous result to rewrite the criterion function:

$$\begin{aligned}TQ_T(\hat{\theta}_T) &= \left[\sqrt{T}\overline{\phi}_T(\hat{\theta}_T)\right]' S_T^{-1}\sqrt{T}\overline{\phi}_T(\hat{\theta}_T) \\ &= \left[\sqrt{T}\overline{\phi}_T(\theta^0) - \hat{J}_T\left[\hat{J}_T' S_T^{-1}\hat{J}_T\right]^{-1}\hat{J}_T' S_T^{-1}\sqrt{T}\overline{\phi}_T(\theta^0)\right]' S_T^{-1} \\ &\quad\times\left[\sqrt{T}\overline{\phi}_T(\theta^0) - \hat{J}_T\left[\hat{J}_T' S_T^{-1}\hat{J}_T\right]^{-1}\hat{J}_T' S_T^{-1}\sqrt{T}\overline{\phi}_T(\theta^0)\right] + o_P(1) \\ &= \left[\sqrt{T}\overline{\phi}_T(\theta^0)\right]' S_T^{-1}\sqrt{T}\overline{\phi}_T(\theta^0) \\ &\quad -\sqrt{T}\overline{\phi}_T(\theta^0) S_T^{-1}\hat{J}_T\left[\hat{J}_T' S_T^{-1}\hat{J}_T\right]^{-1}\hat{J}_T' S_T^{-1}\sqrt{T}\overline{\phi}_T(\theta^0) + o_P(1) \\ &= \sqrt{T}\overline{\phi}_T(\theta^0)'\hat{S}_T'^{-1/2}\left[I - M\right]^{-1} S_T^{1/2}\sqrt{T}\overline{\phi}_T(\theta^0) + o_P(1)\end{aligned}$$

where $S_T^{1/2}$ is such that $S_T = S_T'^{-1/2}S_T^{-1/2}$ and $M = S_T^{-1/2}\hat{J}_T\left[J_T' S_T^{-1}\hat{J}_T\right]^{-1}\hat{J}_T' S_T'^{-1/2}$ which is a projection matrix, hence idempotent and of rank $K - p$. The expected result follows. ■