Efficient Minimum Distance Estimation
with Multiple Rates of Convergence *

Bertille Antoine† and Eric Renault‡

February 19, 2008

Abstract:
This paper extends the asymptotic theory of GMM inference to allow sample counterparts of the estimating equations to converge at (multiple) rates, different from the usual square-root of the sample size. In this setting, we provide consistent estimation of the structural parameters. In addition, we define a convenient rotation in the parameter space (or reparametrization) which permits to disentangle the different rates of convergence. More precisely, we identify special linear combinations of the structural parameters associated with a specific rate of convergence. Finally, we demonstrate the validity of usual inference procedures, like the overidentification test and Wald test, with standard formulas. It is important to stress that both estimation and testing work without requiring the knowledge of the various rates. However, the assessment of these rates is crucial for (asymptotic) power considerations. Possible applications include econometric problems with two dimensions of asymptotics, due to trimming, tail estimation, infill asymptotic, social interactions, kernel smoothing or any kind of regularization.

JEL classification: C32; C12; C13; C51.

Keywords: GMM; Mixed-rates asymptotics; Set estimation; Control variables; Rotation in the coordinate system.

*We would like to thank M. Carrasco, A. Guay, J. Jacod, Y. Kitamura, P. Lavergne, L. Magee, A. Shaikh, V. Zinde-Walsh, and seminar participants at University of British Columbia, University of Montreal, and Yale University for helpful discussions.
†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

The cornerstone of GMM asymptotic distribution theory is that when an estimator $\hat{\theta}_T$ of some vector $\theta$ of parameters (with $\theta^0$ as true unknown value) is defined through a minimum distance problem:

$$\hat{\theta}_T = \arg \min_{\theta} [m_T'(\theta)\Omega m_T(\theta)]$$  \hspace{1cm} (1.1)

$$\sqrt{T}(\hat{\theta}_T - \theta^0)$$ inherits the asymptotic normality of $[\sqrt{T}m_T(\theta^0)]$ by a first-order expansion argument:

$$\sqrt{T}(\hat{\theta}_T - \theta^0) = - \left[ \frac{\partial m_T'(\theta^0)}{\partial \theta} \Omega \frac{\partial m_T(\theta^0)}{\partial \theta'} \right]^{-1} \frac{\partial m_T'(\theta^0)}{\partial \theta} \Omega \sqrt{T}m_T(\theta^0) + o_P(1)$$  \hspace{1cm} (1.2)

while

$$\lim_{T \to \infty} [m_T(\theta)] = 0 \iff \theta = \theta^0$$  \hspace{1cm} (1.3)

It turns out that, for many reasons (see section 2 for a list of examples and a review of the literature), including local smoothing, trimming, infill asymptotic, or any kind of non-root $T$ asymptotics, the asymptotic normality of $m_T(\theta^0)$ may come at a non-standard rate of convergence: $[T^\alpha m_T(\theta^0)]$ is asymptotically a non-degenerated gaussian variable for some $\alpha \neq 1/2$. This does not invalidate the first-order expansion argument (1.2) when realizing that $[T^\alpha(\hat{\theta}_T - \theta^0)]$ is asymptotically equivalent to $[T^\alpha m_T(\theta^0)]$. For instance, Robert (2006) has recently used this argument to estimate extreme copulas. The copulas parameters are backed out from the joint behavior of the tails through a Hill’s type approach (1975). Therefore, as for the Hill estimator, asymptotic normality is reached at a rate different from square-root of $T$, while standard GMM formulas for asymptotic covariance matrices remain.

This paper focuses on the more involved case where identification of $\theta$ comes from different pieces of moment-based information, each of them possibly coming with a different rate of convergence. Then, no exponent $\alpha$ allows the characterization of a non-degenerate asymptotic distribution for $[T^\alpha(\hat{\theta}_T - \theta^0)]$ as in (1.2). We need to resort to mixed-rates asymptotics, where the asymptotic behavior of the minimum distance estimator $\hat{\theta}_T$ is deduced from uniform limit theorems for rescaled and reparametrized estimating equations. While Radchenko (2005) has addressed this issue in the general setting of extremum estimation, the specificity of GMM (or minimum distance) estimation allows us to be more explicit about asymptotic efficiency.
of point estimates and corresponding power of Wald-type testing procedures and confidence sets.

The contribution of this paper is threefold. First, we prove the consistency and get the minimum rate of convergence for estimators of structural parameters, through an empirical process approach. Second, we identify special linear combinations of the parameters associated with specific rates of convergence and we efficiently estimate these directions. Third, we provide inference procedures, like the overidentification test and Wald-type test, with standard formulas. Both estimation and testing work without requiring the knowledge of the various rates. However, the value of these rates and corresponding directions in the parameter space characterize the relevant sequences of local alternatives for asymptotic power analysis.

In econometrics, a related approach can be found in the unit-root literature. Kitamura and Phillips (1997) develop a GMM estimation theory for which the integration properties of the regressors and the corresponding heterogeneous rates of convergence do not need to be known in order to get efficient estimators. Kitamura (1996) and, to some extent, Sims, Stock and Watson (1990) develop a testing strategy which has a standard limit distribution no matter where unit-roots are located. Although similar in spirit, our minimum distance estimation theory does not encompass the above examples because we rather focus on standard gaussian asymptotic distributions where the various rates of convergence are typically not larger than square-root of $T$.

The paper is organized as follows. Section 2 provides a number of motivating examples in modern econometrics where sample counterparts of estimating equations converge at different rates, albeit with gaussian limit distributions. Section 3 precisely defines our framework and proves consistency of GMM estimators of structural parameters $\theta$. We also show how to disentangle and estimate the directions with different rates of convergence, and we prove asymptotic normality of well-suited linear combinations of the structural parameters. Asymptotic efficiency can only be defined about these linear combinations, while estimators of the structural parameters may all be slowly consistent. We show that, on top of a standard issue of efficient choice of the weighting matrix for minimum distance estimation, estimators of fast linear combinations of the structural parameters may be improved through control variables. By contrast with standard GMM, this is not automatically achieved. The issue of Wald-type set estimation is addressed in section 4 while section 5 concludes. All the proofs are gathered in the appendix.
2 Motivating examples

We describe in this section a family of econometric models where different rates of convergence must be considered simultaneously for asymptotic identification of the same vector $\theta$ of structural parameters. Of course, one may even imagine that inference about $\theta$ resorts to several of these examples together. Then, our mixed-rates asymptotic theory is a fortiori needed.

Example 2.1 (Kernel smoothing)

Consider a Nadaraya-Watson estimator of some conditional expectation $E(Y|X = x)$. Depending on the dimension of $X$, and on the combination of bandwidth and kernel, convergence rates to a gaussian limit may differ. With a generic notation $h_T$ for a bandwidth sequence considered with a suitable exponent, the kernel estimator $m_T$ of $E(Y|X = x)$ will be such that $\sqrt{T h_T} [m_T - E(Y|X = x)]$ is asymptotically gaussian.

Assume now that the value of $E(Y|X = x)$ is informative for some structural parameters $\theta$. For instance, Gagliardini, Gouriéroux and Renault (2007) consider such conditional expectations produced by Euler optimality conditions on an asset pricing model, where the pricing kernel is parametrized by $\theta$. Then, $\sqrt{T} \left[ \phi_T(\theta) - \frac{\lambda_T}{\sqrt{T}} \rho(\theta) \right]$ is asymptotically gaussian, where $\rho(\theta) = E(Y|X = x)$, $\lambda_T = \sqrt{T h_T} \to \infty$, but slower than $\sqrt{T}$, and $\phi_T(\theta) = \sqrt{T} h_T m_T$.

Suppose now that several conditional expectations are informative about $\theta$. It may be the case that the different regression functions of interest display different degrees of smoothness, and then lead to choosing heterogeneous rates of convergence for corresponding optimal bandwidths (see Kotlyarova and Zinde-Walsh (2006)). Then, we end up with vectorial functions $\phi_T(\theta)$ and $\rho(\theta)$ such that, for each component $i$:

$\sqrt{T} \left[ \phi_{iT}(\theta) - \frac{\lambda_{iT}}{\sqrt{T}} \rho_i(\theta) \right]$ is asymptotically gaussian, where $\lambda_{iT} = \sqrt{T h_{iT}}$ are heterogeneous due to different bandwidth choices $h_{iT}$. In the asset pricing example of Gagliardini, Gouriéroux and Renault (2007), some
assets are sufficiently liquid to be observed at each date. The associated Euler conditions, written at each date, provide time series of conditional moment restrictions which can be replaced by unconditional ones (thanks to convenient choices of instruments). For such assets, we only have unconditional moments with square-root of $T$ consistent sample counterparts. Hence, the associated rate is simply $\sqrt{T}$.

**Example 2.2 (Trimmed-mean estimation)**

In presence of population moment conditions, $E[y_{it}(\theta)] = 0$ ($i = 1, \ldots, l$) with real-valued $y_{it}(\cdot)$, standard GMM is based on sample counterparts:

$$\bar{Y}_{iT}(\theta) = \frac{1}{T} \sum_{t=1}^{T} y_{it}(\theta)$$

and the standard asymptotic distributional theory does not work when $\text{Var}[y_{it}(\theta)]$ is infinite. Hill and Renault (2008) propose to resort to the concept of trimmed-mean as studied in the statistics literature by Stigler (1973) and Prescott (1978) among others. The key input for minimum distance estimation is $m_{iT}(\theta)$ rather than $Y_{iT}(\theta)$ with:

$$m_{iT}(\theta) = \frac{1}{T} \sum_{t=1}^{T} m_{it,T}(\theta)$$

where $m_{it,T}(\theta) = \begin{cases} y_{it}(\theta) & \text{if } |y_{it}(\theta)| < c_{iT} \\ 0 & \text{otherwise} \end{cases}$

The truncation threshold $c_{iT}$ is such that $c_{iT} \overset{T}{\to} \infty$ (to get asymptotic unbiased moments) and $c_{iT}/\sqrt{T} \overset{T}{\to} 0$ (to control for infinite variance). The rate of convergence to normality for $m_{iT}(\theta)$ is typically slower than $\sqrt{T}$, since an asymptotically non-negligible part of the observations is discarded. Moreover, different moment conditions $E[y_{it}(\theta)] = 0$ with different tail behaviors induce different rates of convergence to normality. Minimum distance estimation based on a vector $m_{T}(\theta) = [m_{iT}(\theta)]_{1 \leq i \leq K}$ typically displays mixed-rates asymptotics.

**Example 2.3 (Mean excess function)**

In a way somewhat symmetric to example 2.2, a mean excess function sets the focus on the $n_T$ largest observations. Typically, the Hill estimator (1975) of a tail index is based on the log-likelihood function of a Pareto distribution considered only for the $n_T$ largest observations,
where \( n_T \overset{T}{\to} \infty \) and \( n_T/T \overset{T}{\to} 0 \).

In a GMM setting, this idea has been revisited by Robert (2006) to estimate the parameters of a bivariate extreme copula. In order to apply the same idea in a dimension larger than 2, one may have to consider different selection rates \([n_T/T]\) to accommodate different tail behaviors. Since the rate of convergence to an asymptotic Gaussian distribution of a Hill-type estimator is given by the number \( n_T \) of included observations, mixed-rates asymptotics show up.

**Example 2.4 (Infill asymptotic)**

In the above examples, rates of convergence slower than square-root of \( T \) show up because only a part of the sample is actually used for estimation. Such rates may also occur because asymptotic is based on increasingly dense observations in a fixed and bounded region. In this case, called fixed-domain asymptotic (or infill asymptotic), this is not the number of useful observations that increases infinitely slower than the sample size, but the effective number of observations: when the sample size increases, new observations represent less and less independent pieces of information. For statistical estimation of diffusion processes, it is well-known (see for instance Kessler (1997)) that infill asymptotic does provide a consistent estimator of the diffusion term and not, in general, of the drift term. Joint increasing-domain asymptotic and fixed-domain asymptotic may provide consistent asymptotically Gaussian estimators of both the drift and the diffusion terms, but at a slower rate for the former. Bandi and Phillips (2003) embed this joint increasing/fixed-domain asymptotic in a minimum distance problem where sample counterparts of both the drift and the diffusion terms are obtained by kernel smoothing. A parametric model of the diffusion process is estimated by matching it against these kernel counterparts. Hence, non-standard rates of convergence show up both due to infill asymptotic and to kernel smoothing. In a more general context, without a natural partition of the set of structural parameters between the drift and the diffusion coefficients, mixed-rates asymptotics would be relevant. Aït-Sahalia and Jacod (2008) show that considering more generally Levy-stable processes introduces even more non-standard rates for jumps components and tails parameters. Lee (2004) considers infill asymptotic for spatial data where a unit can be influenced by many neighbors. For the same reason, irregularity of the information matrix may occur and MLE of some parameters come with a slower rate of convergence.

**Example 2.5 (Social interactions)**
A social interaction model is about economic effects due to individual interactions in a group setting. If $n$ is the total number of individuals under consideration, distributed among $R$ groups with $m$ standing for the average size of a group, Lee (2005) studies the asymptotic properties of estimators of parameters of an interaction model, when both $n$ and $m$ go to infinity, but $m$ is asymptotically infinitely small in front of square-root of $n$. Then, while some parameters estimates are asymptotically gaussian with the standard rate root-$n$, some others only converge at the slower rate $[n^{1/2}/m]$. Lee (2005) stresses that estimation of the structural parameters of interest involves a minimum distance problem where the various components of the matched instrumental parameters may have different rates. It is actually a special case of the general issue we address throughout the paper.

**Example 2.6** *(Nearly-weak instruments)*

In the nearly-weak GMM as introduced by Caner (2007) as a non-linear extension of Hahn and Kuersteiner (2002), the correlation between the instruments and the first-order conditions decline at a rate slower than square-root of $T$. Both Caner (2007) and Antoine and Renault (2008) show that this setting is significantly different from the weak identification case (as in Stock and Wright (2000)): in the latter, since the correlation declines as fast as root-$T$, there is no asymptotic accumulation of information that would allow consistent estimation of all the parameters. In the nearly-weak case, both moments and parameters are asymptotically gaussian, but at rates slower than root-$T$ in proportion of the corresponding degree of near-weakness. Antoine and Renault (2008) set the focus on the case where both strong and nearly-weak instruments are simultaneously at stake for identification of different directions in the parameter space, at respective rates root-$T$ and a slower one. The goal is then to apply the tools of the present paper to revisit a large literature on weak instruments, and, in particular, to reconsider the issue of testing parameters without assuming that they are identified as in Kleibergen (2005).


3 Efficient estimation

3.1 Identification and consistency of a minimum distance estimator

The starting point of minimum distance estimation of an unknown vector \( \theta \) of \( p \) parameters is generally given by \( K \) estimating equations, \( \rho(\theta) = 0 \). These equations are assumed to identify the true unknown value \( \theta^0 \) of the parameter \( \theta \), thanks to the following maintained assumption:

**Assumption 1 (Identifying equations)**

\( \theta \rightarrow \rho(\theta) \) is a continuous function from a compact parameter space \( \Theta \subset \mathbb{R}^p \) into \( \mathbb{R}^K \) such that:

\[
\rho(\theta) = 0 \iff \theta = \theta^0
\]  

(3.1)

Note that assumption 1 implies that \( \theta^0 \) is a well-separated zero of the above equation:

\[
\forall \epsilon > 0 \quad \inf_{\|\theta - \theta^0\| \geq \epsilon} \|\rho(\theta)\| > 0
\]  

(3.2)

This is all we need to prove consistency\(^1\) (see e.g. chapter 5 in van der Vaart (1998)), when we have at our disposal some sample counterpart \( \phi_T(\theta) \) of the estimating equations. More precisely, with time series notations, consider a sample of size \( T \), corresponding to observations at dates \( t = 1, ..., T \). For any possible value \( \theta \in \Theta \) of the parameters, we can compute a \( K \)-dimensional sample-based vector \( \bar{\phi}_T(\theta) \). In many cases, minimum distance estimation can be seen as GMM because \( \bar{\phi}_T(\theta) \) is the sample mean of a double array:

\[
\bar{\phi}_T(\theta) = \frac{1}{T} \sum_{t=1}^{T} \phi_{t,T}(\theta)
\]  

(3.3)

The minimum distance estimator is defined as usual by:

**Definition 3.1** Let \( \Omega_T \) be a sequence of symmetric positive definite random matrices of size \( K \) which converges in probability towards a positive definite matrix \( \Omega \). A minimum distance estimator is defined as:

\[
\hat{\theta}_T = \text{argmin}_{\theta \in \Theta} \left\{ \sum_{t=1}^{T} \phi_{t,T}(\theta) \Omega_T^{-1} \phi_{t,T}(\theta) : \rho(\theta) = 0 \right\}
\]  

where \( \Omega_T \) is a sequence of symmetric positive definite matrices.

---

\(^{1}\)The standard distinction between global assumptions for consistency and local assumptions for asymptotic distributional assumption (see e.g. Pakes and Pollard (1989)) could be also used in our framework, at the cost of longer exposition. Assumption of a compact parameter space is only maintained to simplify the exposition of uniform convergence. Uniform convergence is only needed on a compact neighborhood of \( \theta^0 \).
estimator $\hat{\theta}_T$ of $\theta^0$ is then defined as:

$$
\hat{\theta}_T = \arg \min_{\theta \in \Theta} Q_T(\theta) \quad \text{where} \quad Q_T(\theta) = \tilde{\phi}_T'(\theta)\Omega_T\tilde{\phi}_T(\theta)
$$

(3.4)

Standard minimum distance asymptotic theory assumes that $\tilde{\phi}_T(\theta)$ converges uniformly in probability towards $\rho(\theta)$, thanks to some uniform law of large numbers. Here, we consider more generally the situation where $\tilde{\phi}_T(\theta)$ may converge towards zero even for $\theta \neq \theta^0$: identification is however maintained through higher-order asymptotics. More precisely, we have

$$
T^{-1/2}\left[ \tilde{\phi}_T(\theta) - \frac{\Lambda_T}{T^{1/2}}\rho(\theta) \right] = O_P(1)
$$

(3.5)

where $\Lambda_T$ is a diagonal matrix whose coefficients converge to infinity, but possibly at slower rates than $T^{1/2}$. For sake of simplicity, we always see $\Lambda_T$ as a deterministic sequence, but extensions to random sequences (with associated convergence in probability of diagonal terms towards infinity) would be straightforward. The key point is the following: when a given diagonal coefficient of $\Lambda_T$ goes to infinity strictly slower than $T^{1/2}$ for all $\theta$ in some subset $\Theta^* \subseteq \Theta$, the corresponding component $\rho_i(\theta)$ of $\rho(\theta)$ is squeezed to zero and $\text{Plim} \ \tilde{\phi}_{iT}(\theta) = 0$ for all $\theta \in \Theta^*$. Thus, the probability limit of $\tilde{\phi}_T(\theta)$ does not allow to discriminate between the true unknown value $\theta^0$ and any other $\theta \in \Theta^*$. Identification is then recovered through a central limit theorem about (3.5):

**Assumption 2** (Functional CLT)

(i) The empirical process $(\Psi_T(\theta))_{\theta \in \Theta}$ obeys a functional central limit theorem:

$$
\Psi_T(\theta) \equiv T^{-1/2}\left[ \tilde{\phi}_T(\theta) - \frac{\Lambda_T}{T^{1/2}}\rho(\theta) \right] \Rightarrow \Psi(\theta)
$$

where $\Psi(\theta)$ is a gaussian stochastic process on $\Theta$ with mean zero.

(ii) $\Lambda_T$ is a deterministic diagonal matrix with positive coefficients, such that its minimal and maximal coefficients, respectively denoted as $\underline{\Lambda}_T$ and $\overline{\Lambda}_T$, verify:

$$
\lim_{T \to \infty} \underline{\Lambda}_T = +\infty \quad \text{and} \quad \lim_{T \to \infty} \frac{\overline{\Lambda}_T}{T^{1/2}} < \infty
$$

As suggested by a standard central limit theorem, we mainly focus on the case where the fastest rate is the standard rate $T^{1/2}$. More generally, we could consider any real number $\gamma$
such that equation (3.5) is replaced by
\[
T^\gamma \left[ \theta_T(\theta) - \frac{\Lambda_T}{T^{\gamma}} \rho(\theta) \right] = O_P(1)
\]
and the associated functional CLT is replaced by a tightness assumption. This would allow us to accommodate, for instance, a case with possible unit-roots and normal/mixed normal asymptotics, or with cube-root asymptotics and convergence in distribution towards the maximum of a gaussian process, like Manski’s maximum score estimator (1985). This is beyond the scope of this paper.

The main consistency result of this paper follows:

**Theorem 3.1 (Consistency of \( \hat{\theta}_T \))** Under assumptions 1 and 2, any minimum distance estimator \( \hat{\theta}_T \) like (3.4) is weakly consistent.

As already explained, identification is not ensured is a standard way, and the usual proof of consistency for M-estimators (Jenrich (1969), Amemyia (1985)) does not work in this setting. It is worth rewriting the minimization problem (3.4) to get some intuition on why consistency still holds:
\[
\hat{\theta}_T = \arg \min_{\theta \in \Theta} \left[ \Psi_T(\theta) + \frac{\Lambda_T}{T^{1/2}} \rho(\theta) \right]'
\Omega_T \left[ \Psi_T(\theta) + \frac{\Lambda_T}{T^{1/2}} \rho(\theta) \right] (3.6)
\]
When some diagonal coefficients of \( \Lambda_T \) go to infinity slower than \( T^{1/2} \), the corresponding components of \( \rho(\theta) \) are squeezed to zero in the optimization problem (3.6): their identifying power might then be lost. This explains why we need the empirical process approach\(^2\). The functional CLT (see assumption 2) is helpful to control \( \Psi_T(\theta) \) uniformly on \( \Theta \), and takes advantage of the identifying assumption 1 in the minimization problem (3.6). More precisely, while \( \Psi_T(\theta) \) is uniformly \( O_P(1) \), we show (see lemma A.1 in the appendix) that:
\[
\| \rho(\hat{\theta}_T) \| = O_P \left( \frac{1}{\Lambda_T} \right) (3.7)
\]
And the consistency of \( \hat{\theta}_T \) follows. A special case of our consistency result (theorem 3.1) has been stated by Lee (2005): in his case, \( \Psi_T(\theta) \) does not depend on \( \theta \) and, thus, tightness is no longer an issue. A similar simplification happens in the case of instrumental variables

\(^2\)This has already been pointed out: see Stock and Wright (2000).
estimation of a linear regression model with weak instruments, as in Staiger and Stock (1997), and Hahn and Kuersteiner (2002).

As already explained, our focus of interest is the multiplicity of rates of convergence induced by the asymptotic behavior of the sample moments $\varphi_T(\theta)$, 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 local identification assumption:

**Assumption 3 (Local identification)**

(i) $\rho(.)$ is continuously differentiable on the interior of $\Theta$, $\text{int}(\Theta)$.

(ii) $\theta^0 \in \text{int}(\Theta)$.

(iii) The $(K \times p)$-matrix $[\partial \rho(\theta)/\partial \theta']$ has full column rank $p$ for all $\theta \in \Theta$.

We directly deduce from equation (3.7) and assumption 3 the slowest possible rate of convergence of our estimators:

**Theorem 3.2** Under assumptions 1 to 3, we have:

$$\|\hat{\theta}_T - \theta^0\| = O_P\left(\frac{1}{\lambda_T}\right)$$

where $\lambda_T$ has been defined in assumption 2 as the minimal coefficient of $\Lambda_T$.

Theorem 3.2 is quite a poor result since it assigns the slowest possible rate of convergence to all components of the structural parameters. In the next section, we identify faster directions in the parameter space.

### 3.2 Disentangling the rates of convergence

Without loss of generality, let us consider $\Lambda_T$, the diagonal matrix with the following blocks$^3$:

$$\Lambda_T = \begin{pmatrix} \lambda_{1T} \text{Id}_{k_1} & 0 & \ldots & 0 \\ 0 & \lambda_{2T} \text{Id}_{k_2} & \ldots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \ldots & \lambda_{lT} \text{Id}_{k_l} \end{pmatrix}$$

$^3\text{Id}_k$ represents the identity matrix of size $(k, k)$.
with

\[ i \sum_{i=1}^{l} k_i = K \]

\[ \lim_{T \to \infty} \lambda_{iT} = \infty \quad \text{for } i = 1, \ldots, l \]

\[ \lambda_{i+1,T} = o(\lambda_{i,T}) \quad \text{for } i = 1, \ldots, l - 1 \]

Accordingly, we consider a partition of the estimating equations and of their sample counterparts:

\[ \rho(\theta) = [\rho_1'(\theta) \rho_2'(\theta) \cdots \rho_l'(\theta)]' \quad \text{with } \dim[\rho_i(\theta)] = k_i \quad \text{for } i = 1, \ldots, l \] (3.9)

\[ \overline{\phi}_T(\theta) = [\overline{\phi}_{1T}(\theta) \overline{\phi}_{2T}(\theta) \cdots \overline{\phi}_{lT}(\theta)]' \quad \text{with } \dim[\overline{\phi}_{iT}(\theta)] = k_i \quad \text{for } i = 1, \ldots, l \] (3.10)

A maintained assumption throughout the paper is the prior knowledge of the right partition of the K moment conditions between subsets of respective sizes \((k_1, k_2, \ldots, k_l)\). Typically, this is the case when they correspond to different smoothing or trimming schemes.

Assumption 3(iii) is also reinforced as follows:

**Assumption 4 (Reinforced assumption 3(iii))**

There exist non-negative integers \(s_i\), for \(i = 1, \ldots, l\), such that for all \(\theta\) in the interior of \(\Theta\):

\[ \text{Rank } J_i(\theta) = s_1 + s_2 + \cdots + s_i \]

with the \([p, (k_1 + k_2 + \cdots + k_i)]\)-matrix \(J_i(\theta) = \begin{bmatrix} \frac{\partial \rho_1(\theta)}{\partial \theta} & \frac{\partial \rho_2(\theta)}{\partial \theta} & \cdots & \frac{\partial \rho_l(\theta)}{\partial \theta} \end{bmatrix}\) and \(\sum_{j=1}^{l} s_j = p\).

We assume throughout the paper that the various ranks \(s_1, s_2, \cdots, s_l\) are known. Since the matrices \(J_i(\theta^0)\) are consistently estimated by their sample counterparts (see assumption 5 below), these ranks could be estimated in practice. This extension is beyond the scope of this paper.

We are then faced with the following situation:

(i) Only \(k_1\) estimating equations (defined by \(\rho_1(\theta)\)) have a sample counterpart converging at the fastest available rate \(\lambda_{1T}\). These first \(k_1\) equations can be used in a standard way. Unfortunately, in general, the rank of the associated Jacobian \(J_1(\theta^0)\) is lower than the dimension of the parameter space \((s_1 < p)\). Thus, these estimating equations are not sufficient to identify
the entire parameter $\theta$. Intuitively, they only identify the $s_1$ directions in the $p$-dimensional space of parameters corresponding to $\text{Im} \left[ J_1(\theta^0) \right]$.\(^4\)

(ii) For the same reason, $\text{Im} \left[ J_2(\theta^0) \right]$ characterizes the $(s_1 + s_2)$ directions in the parameter space that can be estimated at least at rate $\lambda_{2T}$. However, since $s_1$ directions (out of the former) can be estimated faster (at rate $\lambda_{1T}$), it is important to disentangle them, for the purpose of efficient estimation.

(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 recursively a sequence of matrices $R_i$, $i = 1, \ldots, l$, as follows:

(i) First, since the orthogonal space of $\text{Im} \left[ J_{l-1}(\theta^0) \right]$ in $\mathbb{R}^p$ entails $s_l$ directions that could be estimated only at the slowest rate $\lambda_{lT}$, we display a basis\(^5\) of this space as the columns of a matrix $R_l$ of size $(p, s_l)$ and rank $s_l$, such that:

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

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

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

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

$$\frac{\partial \rho_i(\theta^0)}{\partial \theta^\prime} R_j = 0 \quad \text{for } i < j \quad \text{with} \quad \text{Rank}[R_j R_{j+1} \cdots R_l] = s_j + s_{j+1} + \cdots + s_l$$

---

\(^4\)For any $(n \times m)$-matrix, $\text{Im}[M]$ represents the subspace of $\mathbb{R}^n$ generated by the column vectors of $M$. In the literature, it is also referred to as $\text{Col}[M]$ and $\text{Range}[M]$.

\(^5\)We could consider, more generally, any set of $s_l$ linearly independent vectors of $\mathbb{R}^p$ which are not in the column space $\text{Im} \left[ J_{l-1}(\theta^0) \right]$. We choose to focus on orthogonal directions for the sake of expositional simplicity.
(iv) Finally, we choose \( R_1 \), of size \( (p, s_1) \), such that \( \text{Rank} \left[ R_1 \ R_2 \ \cdots \ R_l \right] = p \).

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 \( R_i \). In any case, \( R^0 \) 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:

\[
\eta = [R^0]^{-1} \theta = [\eta_i]_{1 \leq i \leq l} \tag{3.11}
\]

with \( \dim(\eta_i) = s_i \) for \( i = 1, \cdots, l \). The exponent "0" remains in \( R^0 \) to stress that this matrix is defined as a function of the true unknown value \( \theta^0 \). For a hypothetical knowledge of \( R^0 \), we define the new parametrization (3.11). Of course, this reparametrization is not feasible in practice. However, it is worth contemplating it 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 induce fast converging estimators of some components of the minimum distance estimator \( \hat{\theta}_T \). In fact, \( \hat{\theta}_T \) will generally be asymptotically equivalent to some linear transformation of \( \bar{\phi}_T(\theta) \), which is likely to mix up all the components of \( \bar{\phi}_T(\theta) \) in such a way that the components of \( \hat{\theta}_T \) are contaminated by the slow rates of convergence. The advantage of the above reparametrization is precisely to isolate the various rates. Let us consider the reparametrized estimating equations:

\[
\rho^*(\eta) = \rho(R^0 \eta) \tag{3.12}
\]

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

\[
\text{for } j = 2, 3, \cdots, l \quad \frac{\partial \rho_j(\theta^0)}{\partial \theta^t} R_j = 0 \text{ with } i < j \tag{3.13}
\]

Under some convenient assumptions, we show that this lower triangularity ensures:

\[
\lambda_{iT}[\hat{\eta}_{iT} - \eta_i^0] = O_P(1) \quad \text{for } i = 1, \cdots, l \tag{3.14}
\]

In other words, the \( s_i \) components of the minimum distance estimator \( \hat{\eta}_{iT} \) inherit the fast rate of convergence (in the sense faster than \( \lambda_{jT} \) for \( j > i \)) of the sample counterpart \( \bar{\phi}_{iT}(\theta) \) of the estimating equation \( \rho_i(\theta) \).
3.3 Asymptotic distribution theory

The following assumption naturally accounts for heterogeneous rates of convergence for the Jacobian matrix \( \partial \rho(\theta)/\partial \theta \):

**Assumption 5** For all \( i = 1, \ldots, l \):

(i) \[ \frac{T^{1/2} \partial \phi_T^i(\theta)}{\sqrt{T}} \]

converges in probability towards \( \frac{\partial \rho_i(\theta)}{\partial \theta} \) uniformly on \( \theta \in \Theta \).

(ii) \[ \frac{\partial \Psi_T^i(\theta^0)}{\partial \theta} = T^{1/2} \left[ \frac{\partial \phi_T^i(\theta^0)}{\partial \theta} - \frac{\lambda_T}{T^{1/2}} \frac{\partial \rho_i'(\theta^0)}{\partial \theta} \right] = O_P(1) \]

Note that both above assumptions would be ensured by an empirical process approach about \( \partial \phi_T^i(\theta) \), similar to the one adopted about \( \bar{\phi}_T(\theta) \) in assumption 2. In this respect, assumption 5 is akin to assuming that assumption 2 is maintained after differentiation with respect to \( \theta \).

For sake of expositional simplicity, our asymptotic distribution theory focuses on the situation where parameters \( \eta_j \) for \( j > i \) (estimated at slower rates than \( \eta_i \)) can be treated as nuisance parameters, without any impact on the asymptotic distribution of the estimator of \( \eta_i \). This issue of interest is similar to Andrews (1994) study of MINPIN estimators, or estimators defined as MINimizing a criterion function that might depend on a Preliminary Infinite dimensional Nuisance parameter estimator. Infinite dimensional or not, we want to avoid the contamination of the asymptotic distribution of the parameters of interest by the nuisance parameters (estimated at slower rates). As Andrews (1994), we also need to ensure some kind of orthogonality between the different parameters.

More precisely, consider the unfeasible minimum distance estimation problem:

\[
\min_{\eta} \left[ \bar{\phi}_T^i(R^0(\eta^T)\Omega_T^{-1}R^0(\eta)) \right] \tag{3.15}
\]

The associated first-order conditions can be written as:

\[
R^0_\theta \frac{\partial \phi_T^i'(R^0(\eta^T))}{\partial \theta} \Omega_T^{-1} \bar{\phi}_T(R^0(\eta^T)) = 0 \tag{3.16}
\]

\(^{6}\text{Kleibergen (2005) also maintains the same kind of assumptions in the context of weak identification.}\)

\(^{7}\text{This is also related to the block-diagonality of the information matrix in maximum likelihood contexts.}\)
and the asymptotic distribution of the estimator $\hat{\eta}_T$ is derived by replacing $[T^{1/2} \varphi_T(R^0 \hat{\eta}_T)]$ in (3.16) by its first-order Taylor expansion:

$$T^{1/2} \varphi_T(R^0 \eta^0) + T^{1/2} \frac{\partial \varphi_T(R^0 \eta^0)}{\partial \eta} R^0 (\hat{\eta}_T - \eta^0)$$

for some $\eta^*_T$ defined component by component between $\eta^0$ and $\hat{\eta}_T$. Then, for the $i$-th group of components ($i = 1, \cdots, l$), this expansion writes:

$$T^{1/2} \varphi_{iT}(R^0 \eta^0) + \sum_{j=1}^l \frac{T^{1/2}}{\lambda_{iT}} \frac{\partial \varphi_{iT}(R^0 \eta^0)}{\partial \theta} R_j \lambda_{iT} [\hat{\eta}_{iT} - \eta^0_j]$$

Since $\lambda_{iT}[\hat{\eta}_{iT} - \eta^0] = \mathcal{O}(1)$ ($i = 1, \cdots, l$) (see equation (3.14)), we need to ensure the following to avoid the contamination of the distribution of fast converging parameters by the slow ones:

$$\frac{T^{1/2}}{\lambda_{iT}} \frac{\partial \varphi_{iT}(R^0 \eta^0)}{\partial \theta} R_j \xrightarrow{P} 0 \text{ when } T \to \infty \text{ for all } j > i$$

(3.17)

The difficulty is that, in general, $\theta^*_T = R^0 \eta^*_T$ mixes all rates of convergence, and may be estimated as slowly as $\lambda_{iT}$. This is the reason why we need to maintain the following assumption:

**Assumption 6 (Orthogonality condition)**

(i) If $\theta^*_T$ is such that $\|\theta^*_T - \theta^0\| = \mathcal{O}(1/\lambda_{iT})$ then for $i = 1, \cdots, l$

$$\frac{T^{1/2}}{\lambda_{iT}} \frac{\partial \varphi_{iT}(\theta^*_T)}{\partial \theta} R_j \xrightarrow{P} 0 \text{ when } T \to \infty \text{ for all } j > i$$

(ii) For all $i = 1, \cdots, l$ and each component $k = 1, \cdots, k_i$: $T^{1/2} \lambda_{iT} \left[ \frac{\partial^2 \varphi_{iT,k}(\theta)}{\partial \theta \partial \theta'} \right]$ converges in probability uniformly on $\theta \in \Theta$ towards some well-defined matrix $H_{ik}(\theta)$.

This orthogonality condition is strikingly similar to condition (2.12) p49 in Andrews (1994). Of course, it is also tightly related to the lower triangularity of the matrix $[\partial \rho^*(\eta^0)/\partial \eta'] = [\partial \rho(\theta^0)/\partial \theta'] R^0$. Actually:

$$\text{Plim} \left[ \frac{T^{1/2}}{\lambda_{iT}} \frac{\partial \varphi_{iT}(\theta^*_T)}{\partial \theta} R_j \right] = \text{Plim} \left[ \frac{\lambda_{iT}}{\lambda_{iT}} \left( \frac{T^{1/2}}{\lambda_{iT}} \frac{\partial \varphi_{iT}(\theta^*_T)}{\partial \theta} - \frac{\partial \rho(\theta^0)}{\partial \theta} \right) R_j \right]$$

(3.18)
The difficulty is that, due to $\theta^*_T$, the term within parenthesis is not of order $(1/\lambda iT)$ (as it would be if $\theta^*_T = \theta^0$) but only $(1/\lambda lT)$, at least if a uniform mean-value theorem can be applied to $[\partial \phi_i(\theta^*_T)/\partial \theta']$ in the neighborhood of $\theta^0$. Hence, the required orthogonality condition follows only if we know that:

$$
\frac{\lambda_{ijT}}{\lambda_{ijT}} \times \frac{1}{\lambda_{ijT}} \to 0 \quad \forall \ j > i, \quad \text{or} \quad \lambda_{ijT} = o(\lambda_{ijT}^2) \quad (3.19)
$$

In other words, we can get assumption 6 if we maintain:

**Assumption 6** *(Sufficient condition for assumption 6)*

(i) $\lambda_{1T} = o(\lambda_{1T}^2)$.

(ii) For all $i = 1, \ldots, l$ and each component $k = 1, \ldots, k_i$: 
$$
\frac{T^{1/2}}{\lambda_{1T}} \left[ \frac{\partial^2 \phi_{IT,k}(\theta)}{\partial \theta \partial \theta'} \right] \text{ converges in probability uniformly on } \theta \in \Theta \text{ towards some well-defined matrix } H_{ik}(\theta).
$$

Assumption 6* states that, even though the sample counterparts of the estimating equations converge at different rates, the discrepancy of these rates cannot be too large. For instance, if the fast rate is $T^{1/2}$, the slowest rate must be faster than $T^{1/4}$. This is typically a sufficient condition that Andrews (1995, e.g. p563) considers to illustrate in what circumstances MINPIN estimators are well-behaved. It has of course strong implications on the range of bandwidth or trimming parameters that one can consider in the examples of section 2.

For instance, in the case of one dimensional kernel smoothing, $\lambda_{2T} = \sqrt{T\eta_T}$ fulfills the required condition (with respect to $\lambda_{1T} = \sqrt{T}$) only if $h_T \sqrt{T} \to \infty$. Interestingly enough, the case of first-order underidentification (Sargan (1983), Dovonon and Renault (2007)) is the limit case where the slow rate (namely $T^{1/4}$) is just sufficiently slow to violate the condition. Technically, maintaining assumptions 5 and 6 (or 6*) is actually useful for proving the following lemma:

**Lemma 3.3** *Under assumptions 1 to 6 (or 6*), if $\theta^*_T$ is such that $||\theta^*_T - \theta^0|| = O_P(1/\lambda_{1T})$, then

$$
T^{1/2} \frac{\partial \phi_T(\theta^*_T)}{\partial \theta'} R^0 \tilde{\Lambda}_T^{-1} \to P \quad \text{when } T \to \infty
$$

*In the context of weak instruments, Antoine and Renault (2008) define nearly-strong instruments as instruments featuring some degree of weakness, but still conformable to assumption 6*.

17
where \( J^0 \) is the \((K,p)\) block-diagonal matrix with diagonal blocks \( [(\partial \rho_t(\theta^0))/\partial \theta^t)R_t] \) and \( \tilde{\Lambda}_T \) is the \((p,p)\) diagonal matrix defined as

\[
\tilde{\Lambda}_T = \begin{pmatrix}
\lambda_{1T} I_{s_1} & 0 & \cdots & 0 \\
0 & \lambda_{2T} I_{s_2} & \cdots & 0 \\
\vdots & \vdots & \ddots & \vdots \\
0 & 0 & \cdots & \lambda_{lT} I_{s_l}
\end{pmatrix}
\]

with \( \sum_{i=1}^{l} s_i = p \)

\[
\lim_{T \to \infty} \lambda_{iT} = \infty \quad \text{for} \quad i = 1, \ldots, l
\]

\[
\lambda_{i+1,T} = o(\lambda_{i,T}) \quad \text{for} \quad i = 1, \ldots, l - 1
\]

Up to unusual rates of convergence, we get a standard asymptotically normal distribution for the new parameters \( \eta = [R^0]^{-1}\theta \):

**Theorem 3.4 (Asymptotic Normality)**

Under assumptions 1 to 6 (or 6*), the minimum distance estimator \( \hat{\theta}_T \) defined by (3.4) is such that:

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

where \( S^0 \) denotes the covariance matrix of the asymptotic gaussian distribution of \( \sqrt{T} \phi_T(\theta^0) \).

It is worth noting that this result has strong similarities with Hansen (1982) classical result about the asymptotic distribution of GMM. First, the matrix \( J^0 \) may almost be interpreted as \( [(\partial \rho^*(\eta^0))/\partial \eta^t)R^0] \) where \( \rho^*(\eta) = \rho(R^0\eta) \). This simple interpretation is not fully correct. While \( [(\partial \rho^*(\eta^0))/\partial \eta^t] \) is a lower-triangular matrix (due to the discrepancy between rates of convergence), the upper-diagonal blocks also cancel out in the limit considered in lemma 3.3, in such a way that \( J^0 \) is block-diagonal. However, seeing \( J^0 \) as \( [(\partial \rho^*(\eta^0))/\partial \eta^t] \) would allow us to interpret the asymptotic variance in theorem 3.4 as the standard asymptotic variance of a minimum distance estimator computed from the (unfeasible) minimization problem (3.15). In particular, the cancelation of upper-diagonal blocks does not invalidate the standard argument that the optimal weighting matrix is a consistent estimator of the inverse of the long-term variance.
Theorem 3.5 Let $S^0$ denote the covariance matrix of the asymptotic gaussian distribution of $\sqrt{T} \bar{\phi}_T(\theta^0)$. Under assumptions 1 to 6 (or 6$^*$), the asymptotic variance displayed in theorem 3.4 is minimal when the minimum distance estimator $\hat{\theta}_T$ is defined by (3.4) while using a consistent estimator of $[S^0]^{-1}$ as the weighting matrix $\Omega_T$. Then,

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

A consistent estimator $S_T$ of the long-term covariance matrix $S^0$ can be constructed in the standard way (see e.g. Hall (2005)) from a preliminary inefficient GMM estimator of $\theta$. Then, up to the block-diagonality of the matrix $J^0$, we get the standard formula for the asymptotic distribution of an efficient minimum distance estimator of $\eta$.

In general, the focus of interest is not the vector $\eta$ (new parameters) but the vector $\theta$ (structural parameters). As far as inference about $\theta$ is concerned, several practical implications of theorem 3.5 are worth mentioning. From the lemma 3.3, a consistent estimator of the asymptotic covariance matrix $[J^0[S^0]^{-1}J^0]^{-1}$ is:

$$T^{-1} \left[ \tilde{\Lambda}_T[R^0]^{-1} \left[ \frac{\partial \bar{\phi}_T(\hat{\theta}_T)}{\partial \theta} S_T^{-1} \frac{\partial \bar{\phi}_T(\hat{\theta}_T)}{\partial \theta'} R^0 \tilde{\Lambda}_T^{-1} \right] \right]^{-1}$$

Note that we do not address the estimation of the matrix $R^0$ at this stage: its knowledge is actually not really necessary. From theorem 3.5, 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 (3.20): informally, we can say that $\sqrt{T}(\hat{\theta}_T - \theta^0)$ behaves like a gaussian with mean zero and variance$

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

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 required for inference. However, the above intuition (albeit practically relevant) is theoretically misleading for several reasons. First, in general, all components of $\hat{\theta}_T$ converge.
slowly towards \( \theta^0 \) and thus \( \sqrt{T}(\hat{\theta}_T - \theta^0) \) has no limit distribution. When we say that it is approximately a gaussian with variance (3.21), one must realize that since

\[
\sqrt{T} \frac{\partial \phi_T(\hat{\theta}_T)}{\partial \theta'} \xrightarrow{p} \frac{\partial \rho(\theta^0)}{\partial \theta'}
\]

we actually have

\[
\frac{\partial \phi_T(\hat{\theta}_T)}{\partial \theta'} \xrightarrow{p} 0 \quad \text{for } i > 1
\]

In other words, considering the asymptotic variance (3.21) is akin to considering the inverse of an asymptotically singular matrix: (3.21) is not an estimator of the standard population matrix

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

Typically, beyond the above singularity, the population matrix (3.22) will not display, in general, the right block-diagonality structure. Inference about \( \theta \) is actually more involved than one may believe at first sight, from the apparent similarity with standard GMM formulas: section 4 is devoted to inference issues.

At least, the seemingly standard asymptotic distribution theory allows us to perform an over-identification test as usual:

**Theorem 3.6 (J-test)**

*Under assumptions 1 to 6 (or 6*), if \( \Omega_T \) is a consistent estimator of \( S^0 \), \( TQ_T(\hat{\theta}_T) \) is asymptotically distributed as a chi-square with \((K - p)\) degrees of freedom.*

### 3.4 Control variables

In a standard GMM setting, the efficient choice of the weighting matrix implicitly implements a control variables strategy (see Back and Brown (1993), Antoine, Bonnal and Renault (2007)). When a moment condition:

\[
E[Y_t] = \mu
\]

is augmented by an overidentified set of moment conditions,

\[
E[\phi_t(\theta)] = \rho(\theta)
\]

20
the resulting efficient GMM estimator \((\hat{\theta}_T, \hat{\mu}_T)\) provides, in general, an estimator \(\hat{\mu}_T\) of \(\mu\) more efficient than the naive sample mean \(\bar{Y}_T\), because it takes advantage of the asymptotically valid control variable \(\left[ \phi_t(\hat{\theta}_T) - \rho(\hat{\theta}_T) \right] \).

The situation is quite dramatically different when the sample counterparts of (3.23) and (3.24) do not converge at the same rate. First, if the sample counterpart of (3.24) converges at a faster rate than the sample counterpart of (3.23), there is no hope to improve the estimator of \(\mu\). As usual, nuisance parameters estimated at a fast rate do not play any role in the asymptotic distribution of slowly converging estimators of parameters of interest: the asymptotic distribution is the same as if nuisance parameters were known. Second, if the sample counterpart of (3.24) converges at a slower rate than the sample counterpart of (3.23), there is room for improvement of the estimator of \(\mu\). However, this is not automatically done by efficient GMM defined in the previous sections. Consider the informational content of estimating equations with respect to the new parameters \(\eta\) (as defined in former sections). In general,

\[
\frac{\partial \rho_i(\theta^0)}{\partial \theta^f} R_j \neq 0 \quad \text{for} \ i \geq j
\]

Therefore, with the notations introduced in equation (3.12),

\[
\frac{\partial \rho^*_i(\eta^0)}{\partial \eta_j} \neq 0 \quad \text{for} \ i \geq j
\]

This means that the \(i\)-th set of estimating equations contains some information about \(\eta_j\) (\(i \geq j\)), or about all parameters estimated at a rate \(\lambda iT\) or faster. However, the informational content of estimating equations \(\rho_i(\eta) = 0\) (about parameters \(\eta_j, i > j\)) is basically wasted by our efficient GMM procedure since:

\[
\sqrt{T} \frac{\partial \phi^* T(\theta^0)}{\partial \theta^f} \sim 0 \quad \text{for} \ i > j
\]

In other words, as already mentioned, the matrix \(J^0\) is block-diagonal even though the matrix \(\left[ \frac{\partial \rho^*(\eta^0)}{\partial \eta} \right]\) is not. In the implicit selection of estimating equations done by efficient GMM, a zero-weight is given to the dependence of the \(i\)-th group of estimating equations on parameters estimated faster. Therefore, we do not take advantage of these equations for accurate estimation of the former parameters.

In this section, we show how to use the above relevant information; however, it requires a slight generalization of our former setting. Assume you want to improve the estimator of

\[21\]
(fast) parameters (say $\eta_j$), by some estimating equations with a slower empirical counterpart $\rho_i^*(\eta) = 0$ (for $i > j$). A control variables principle amounts to replacing the moment vector $\bar{\phi}_{IT}(\theta^0)$ by the residual $\left[\bar{\phi}_{jT}(\theta^0) - A_j\bar{\phi}_{up(j),T}(\theta^0)\right]$ of its asymptotic regression on $\bar{\phi}_{up(j),T}(\theta^0)\equiv \left[\bar{\phi}_{iT}(\theta^0)\right]_{j<i\leq l}$, where:

$$A_j = \lim_{T \to \infty} \left\{ \text{Cov} \left[\bar{\phi}_{jT}(\theta^0), \bar{\phi}_{up(j),T}(\theta^0)\right] \cdot \left[\text{Var}\bar{\phi}_{up(j),T}(\theta^0)^{-1}\right] \right\}$$

However, while standard GMM asymptotic distributional theory is invariant to linear combinations of moment equations, it is no longer the case with heterogeneous rates of convergence. To see this, let us just consider the first group of equations. For a given matrix $B$, we introduce:

$$\begin{cases}
\tilde{\phi}_{1T}(\theta) = \bar{\phi}_{1T}(\theta) - B\bar{\phi}_{up(1),T}(\theta) \\
\tilde{\phi}_{iT}(\theta) = \bar{\phi}_{iT}(\theta), \quad \forall \ i > 1 \\
\tilde{\phi}_T(\theta) = \left[\tilde{\phi}_{iT}(\theta)\right]_{1\leq i \leq l}
\end{cases}$$

First, we need a slight extension of our empirical process approach to consistent estimation. Starting from assumption 2,

$$T^{1/2} \left[\bar{\phi}_T(\theta) - \frac{\Lambda_T}{T^{1/2}}\rho(\theta)\right] \Rightarrow \Psi(\theta)$$

where $\Psi(\theta)$ is a gaussian stochastic process on $\Theta$ with mean zero, we have:

$$T^{1/2} \left[\tilde{\phi}_{1T}(\theta) - \frac{\lambda_1T}{T^{1/2}}\rho_1(\theta)\right] = T^{1/2} \left[\tilde{\phi}_{1T}(\theta) - \frac{\lambda_1T}{T^{1/2}}\tilde{\rho}_1T(\theta)\right] - B\Lambda_{up(1),T}\rho_{up(1)}(\theta)$$

with $\tilde{\rho}_1T(\theta) = \rho_1(\theta) - B\frac{\Lambda_{up(1),T}}{\lambda_1T}\rho_{up(1)}(\theta)$, $\rho_{up(j)}(\theta) = [\rho_i(\theta)]_{j<i\leq l}$, and

$$\Lambda_{up(j),T} = \begin{pmatrix}
\lambda_{j+1,T}Id_{k_{j+1}} & & \\
& \lambda_{j+2,T}Id_{k_{j+2}} & \\
& & \ddots \\
& & & \lambda_{l,T}Id_{k_l}
\end{pmatrix}$$

Extension to any subset of estimating equations would be straightforward.
For the first term, we have:

\[
T^{1/2} \left[ \tilde{\phi}_{IT}(\theta) - \frac{\lambda_{IT}}{T^{1/2}} \tilde{\rho}_{IT}(\theta) \right] = T^{1/2} \left[ \tilde{\phi}_{IT}(\theta) - \frac{\lambda_{IT}}{T^{1/2}} \rho_{1}(\theta) \right] - BT^{1/2} \left[ \tilde{\phi}_{up(1),T} - \frac{\Lambda_{up(1),T}}{T^{1/2}} \rho_{up(1)}(\theta) \right] \Rightarrow \Psi_{1}(\theta) - B\Psi_{up(1)}(\theta)
\]

The second term introduces a bias component. However, since \(\lambda_{iT} = o(\lambda_{1T})\) (for all \(i > 1\)), \(\tilde{\rho}_{IT}(\theta)\) converges towards \(\rho_{1}(\theta)\). The same asymptotic theory (see appendix) can then be derived for an estimator of \(\theta\) solution of:

\[
\min_{\theta} \left[ \tilde{\phi}_{T}^{T}(\theta) \Omega \tilde{\phi}_{T}(\theta) \right]
\]

as if,

\[
T^{1/2} \left[ \tilde{\phi}_{T}(\theta) - \frac{\Lambda_{T}}{T^{1/2}} \rho(\theta) \right] \Rightarrow \tilde{\Psi}(\theta)
\]

with

\[
\begin{align*}
\tilde{\Psi}_{1}(\theta) &= \Psi_{1}(\theta) - B\Psi_{up(1)}(\theta) \\
\tilde{\Psi}_{i}(\theta) &= \Psi_{i}(\theta) \quad \forall \ i > 1
\end{align*}
\]

The intuition is the following. First, equation (3.28) does not hold, but only (with obvious notations):

\[
T^{1/2} \left[ \tilde{\phi}_{T}(\theta) - \frac{\Lambda_{T}}{T^{1/2}} \rho_{T}(\theta) \right] \Rightarrow \tilde{\Psi}(\theta)
\]

This difference does not matter for consistency result, since \(\tilde{\rho}_{T}(\theta) \overset{T}{\to} \rho(\theta)\). Second, the asymptotic distributional theory is not modified by the replacement of \(\rho(\theta)\) by \(\tilde{\rho}_{T}(\theta)\), since for all \(T\): \(\tilde{\rho}_{T}(\theta^{0}) = \rho(\theta^{0}) = 0\). As a result, we can state:

**Theorem 3.7 (Asymptotic normality in the extended case)**

For \(B\) given \((k_{1}, K - k_{1})\)-matrix, define:

\[
\begin{align*}
\tilde{\phi}_{1T}(\theta) &= \tilde{\phi}_{IT}(\theta) - B\tilde{\phi}_{up(1),T}(\theta) \\
\tilde{\phi}_{iT}(\theta) &= \tilde{\phi}_{iT}(\theta), \quad \forall \ i \geq 2 \\
\tilde{\phi}_{T}(\theta) &= \left[ \tilde{\phi}_{iT}(\theta) \right]_{1 \leq i \leq l}
\end{align*}
\]

and

\[
\begin{align*}
\tilde{\Psi}_{1}(\theta) &= \Psi_{1}(\theta) - B\Psi_{up(1)}(\theta) \\
\tilde{\Psi}_{i}(\theta) &= \Psi_{i}(\theta), \quad \forall \ i \geq 2 \\
\tilde{\Psi}(\theta) &= \left[ \tilde{\Psi}_{i}(\theta) \right]_{1 \leq i \leq l}
\end{align*}
\]
and $S(B) = \text{Var} \left[ \hat{\Psi}(\theta^0) \right]$. Under assumptions 1 to 6 (or 6∗), if $\hat{\theta}(B)$ is the minimum estimator solution of (3.27), with $\Omega_T$ consistent estimator of $[S(B)]^{-1}$, then:

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

For each choice of the matrix $B$, theorem 3.7 provides different estimators with different asymptotic variances. It is worth reminding that this result is strikingly different from the standard GMM asymptotic theory. The estimators considered in the previous sections correspond to $B = 0$. However, alternative values of $B$ may be preferred. Consider the estimator

$$\hat{\eta}_T(B) = \left[ \hat{\eta}_i(B) \right]_{1 \leq i \leq l} = [R^0]^{-1} \hat{\theta}(B)$$

that clearly disentangles the various rates of convergence. As far as the asymptotic variance of $\hat{\eta}_1T$ is concerned, the optimal choice of $B$ corresponds to the control variables strategy:

**Theorem 3.8 (Optimal choice for $B$)**

Define $\text{AVar} [\hat{\eta}_{1T}(B)]$ as the asymptotic variance of $\left[ \lambda_{1T} \hat{\eta}_{1T}(B) \right]$ as in theorem 3.7. Assume that $S^0 = \lim_{T \to \infty} \text{Var} \left[ \sqrt{T} \phi_T(\theta^0) \right]$. Under the assumptions of theorem 3.7, $\text{AVar} [\hat{\eta}_{1T}(A_1)] \leq \text{AVar} [\hat{\eta}_{1T}(B)]$ for any matrix $B$, where $A_1$ corresponds to the long-term regression coefficients (or control variables strategy):

$$A_1 = \lim_{T \to \infty} \left\{ \text{Cov} \left[ \phi_{1T}(\theta^0), \phi_{up(1),T}(\theta^0) \right] \left[ \text{Var} \phi_{up(1),T}(\theta^0) \right]^{-1} \right\}$$

where $\leq$ denotes the comparison between matrices.

Moreover, the asymptotic variances of $\hat{\eta}_{1T}(A_1)$ and $\hat{\eta}_{1T}$ (when $B = 0$) are different, as long as the matrix $A_1$ is nonzero, that is when the asymptotic covariance between $\phi_{1T}(\theta^0)$ and $\phi_{1T}(\theta^0)$ is nonzero for some $i > 1$. In other words, the above linear combination of the moment conditions allows us to improve the asymptotic variance of the (fast) estimated direction by using more efficiently the informational content of the estimating equations with a slower rate of convergence. However, there is no such thing as a free lunch as shown in the example below.
Example 3.1 Assume that we have two groups of moment conditions. We want to compare the two competing estimators:

\[ \hat{\eta}_T = [\hat{\eta}_{iT}]_{1 \leq i \leq 2} = [R^0]^{-1}\hat{\theta}_T \quad \text{and} \quad \hat{\eta}^{(A_1)}_T = [\hat{\eta}^{(A_1)}_{iT}]_{1 \leq i \leq 2} = [R^0]^{-1}\hat{\theta}^{(A_1)}_T \]

As shown in the appendix, under the assumptions of theorem 3.8,

\[ \text{AVar} \left[ \hat{\eta}^{(A_1)}_T \right] \geq \text{AVar} \left[ \hat{\eta}_T \right] \]

and the two matrices are in general different when

\[ \lim_{T \to \infty} \{ \text{Cov} \left[ \phi_1^T(\theta^0), \phi_2^T(\theta^0) \right] \} \neq 0 \]

In other words, in order to improve the accuracy of the estimator of \( \eta_1 \), we pay the price of deteriorating the accuracy of the estimator of \( \eta_2 \).

3.5 Feasible asymptotic distribution

The asymptotic theory developed so far is not feasible, since based on the unknown true matrix of change of basis \( R^0 \). The purpose of this section is to study the consistent estimation of \( R^0 \) and corresponding plugging in asymptotics for the parameters of interest\(^{10}\). Since \( R^0 \) is a matrix of change of basis in \( \mathbb{R}^p \), we choose it as an orthogonal matrix. Thus, it is consistently estimated by a sequence of moment-based estimation.

(i) First, \( R_l \) is the only squared matrix of size \( s_l \) (up to a rotation of columns and sign changes) such that:

\[ \begin{cases} R'_l R_l = I_{s_l} \\ \frac{\partial \rho}{\partial \theta} R_l = 0 \quad \forall \ 1 \leq i < l \end{cases} \]  

(3.29)

For a convenient normalization, the system (3.29) defines a unique solution for the \( s_l^2 \) coefficients of \( R_l \). Note that it can be written:

\[ [J_{l-1}(\theta^0) \ R_l]' R_l = \begin{bmatrix} 0 \\ I_{s_l} \end{bmatrix} \]

\(^{10}\)Recall that a maintained assumption is the knowledge of the right dimensions \( (k_i, s_i) \), \( i = 1, 2, \ldots, l \).
with $J_{l-1}(\theta^0)$ matrix of rank $(p - s_l)$ and, by construction, with $\text{Im} [J_{l-1}(\theta^0)]$ orthogonal to $\text{Im}[R_l]$. Therefore, a consistent estimator $\hat{J}_{l-1}$ of $J_{l-1}(\theta^0)$ provides a consistent moment estimator $\hat{R}_l$ of $R_l$ defined as the (asymptotically) unique solution of the equations

$$[\hat{J}_{l-1} \hat{R}_l] \hat{R}_l = \begin{bmatrix} 0 \\ \text{Id}_{s_l} \end{bmatrix}$$

(ii) Then, we can define a unique $R_{l-1}$ (still with a convenient normalization) solution of:

$$\begin{cases} R_{l-1}' R_{l-1} = \text{Id}_{s_{l-1}} \\ R_{l-1}' R_{l-1} = 0 \\ \frac{\partial \rho_i(\theta^0)}{\partial \theta} R_{l-1} = 0 \quad \forall \ 1 \leq i \leq l - 2 \end{cases}$$

that is:

$$[J_{l-2}(\theta^0) R_l R_{l-1}]' R_{l-1} = \begin{bmatrix} 0 \\ 0 \\ \text{Id}_{s_{l-1}} \end{bmatrix}$$

Adding these equations to the former ones provides a consistent moment estimator $[\hat{R}_{l-1} \hat{R}_l]$ of $[R_{l-1} R_l]$.

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

$$[J_{j-1}(\theta^0) R_l \cdots R_j]' R_{j-1} = \begin{bmatrix} 0 \\ 0 \\ \vdots \\ 0 \\ \text{Id}_{s_j} \end{bmatrix}$$

The above recursion provides a consistent moment-based estimator $\hat{R}$ of $R^0$ with, as only input, a consistent estimator $\hat{J}_{l-1}$ of

$$J_{l-1}(\theta^0) = \begin{bmatrix} \frac{\partial \rho_1(\theta^0)}{\partial \theta} & \frac{\partial \rho_2(\theta^0)}{\partial \theta} & \cdots & \frac{\partial \rho_{l-1}(\theta^0)}{\partial \theta} \end{bmatrix}$$

By virtue of assumption 5, we have for $i = 1, \ldots, l$:

$$\frac{\partial \rho_i(\theta^0)}{\partial \theta} = \text{P} \lim \frac{T^{1/2}}{\lambda_{iT}} \frac{\partial \hat{\pi}^T(\hat{\theta}_T)}{\partial \theta}$$

26
A consistent estimator of $J_{l-1}(\theta^0)$ (and in turn of $R^0$) is then easy to derive from sample counterparts, insofar as we know the rates of convergence $\lambda_1T, \lambda_2T, \ldots, \lambda_{l-1}T$. It is the case, for instance, in a kernel smoothing based problem where $\lambda_jT = \sqrt{Th_jT}$ for a given bandwidth sequence $h_jT$. Moreover, it is worth noting that the knowledge of the slowest rate $\lambda_lT$ is not required. This is important, since it solves, at least, all the examples with only two rates of convergence ($l = 2$) with a standard square-root of $T$ as the fast rate of convergence ($\lambda_1T = \sqrt{T}$). In any case, the resulting estimators $\hat{J}_{l-1}$ and $\hat{R}_l$ (unfortunately) inherit the slowest rate of convergence, as $\hat{\theta}_T$ itself:

$$\|\hat{R} - R^0\| = O_P\left(\frac{1}{\lambda_lT}\right)$$

Surprisingly enough, this slow rate of convergence does not prevent us from feasible asymptotic distribution theory (for the estimation of relevant directions in the parameter space), including for the fast estimated directions.

**Theorem 3.9** *(Feasible asymptotic normality)*

Under assumptions 1 to 6*, if

$$\|\hat{R} - R^0\| = O_P\left(\frac{1}{\lambda_lT}\right)$$

then the asymptotic distribution of $[\tilde{\Lambda}_T\hat{R}^{-1}\left(\hat{\theta}_T - \theta^0\right)]$ coincides with the gaussian asymptotic distribution of $[\tilde{\Lambda}_T [R^0]^{-1}\left(\hat{\theta}_T - \theta^0\right)]$ as characterized in theorems 3.4 and 3.5.

The key intuition for the proof of this theorem consists in the following decomposition:

$$\hat{R}^{-1}\left(\hat{\theta}_T - \theta^0\right) = [R^0]^{-1}\left(\hat{\theta}_T - \theta^0\right) + \left(\hat{R}^{-1} - [R^0]^{-1}\right)\left(\hat{\theta}_T - \theta^0\right)$$

The (potentially) slow rates of convergence in the second term of the RHS do not deteriorate the (potentially) fast rates in the directions $[R^0]^{-1}(\hat{\theta}_T - \theta^0)$, since these slow rates show up as $\lambda_lT$ at worst, which is still faster than $\lambda_1T$ by assumption 6*.

To summarize, theorem 3.9 provides feasible estimation of the directions $[\hat{R}^{-1}\theta^0]$. The estimator $[\hat{R}^{-1}\hat{\theta}_T]$ preserves the hierarchy of the rates of convergence in the different directions of the parameter space. It must be acknowledged, however, that the estimation error is endowed with the desirable rates of convergence when computed as $[\hat{R}^{-1}\hat{\theta}_T - \hat{R}^{-1}\theta^0]$ instead
of \( \hat{R}^{-1}\hat{\theta}_T - [R^0]^{-1}\theta^0 \). It is, by definition, impossible to take advantage of the fast rates of convergence in the estimation of \([R^0]^{-1}\theta^0\), since the identification of \(R^0\) involves, in general, all the directions of the parameter space, including the poorly identified ones. However, it may be the case that the structural parametrization is such that some component of \(\theta\) do not enter some specific moment conditions. Then, more may be known about the relevant directions in the parameter space. In the framework of kernel smoothing, Gagliardini, Gouriéroux, and Renault (2007) have developed such an example for an option pricing application with a prior partition between several kinds of preference parameters.

In any case, the asymptotic distributional theory for \([R^0]^{-1}\hat{\theta}\) paves the way for the design of Wald-type confidence sets for any function of \(\theta\). The main purpose of section 4 below is to show that these confidence sets and their confidence levels can be computed exactly as usual, without considering mixed-rates asymptotics. The estimation of the relevant rotation in the coordinate system only matters when one wants to assess the power against different sequences of local alternatives.

### 4 Set estimation

In this section, we focus on the test for a null hypothesis, \(H_0 : g(\theta) = 0\) or equivalently on set estimation of \(g(\theta)\), where the function \(g(\cdot)\), from \(\Theta\) to \(\mathbb{R}^q\), is continuously differentiable on the interior of \(\Theta\). A couple of preliminary remarks are in order.

First, working under the null may lead to revisit significantly the reparametrization \(\eta = [R^0]^{-1}\theta\) defined in section 3. Typically, additional information may conduct us to define differently the linear combinations of \(\theta\) estimated respectively at the different rates of convergence. To circumvent this difficulty, we do not consider any constrained estimator and we focus on Wald-type set estimation\(^{11}\) and test.

Second, as already explained, this paper specifically considers the simultaneous treatment of different rates of convergence. This more general point of view comes at a price when one wants to build up confidence sets. More precisely, we may face singularity issues when some

\(^{11}\)Recently, Caner (2007) overlooks this complication and derives the standard asymptotic equivalence results for the trinity of tests. However, he only considers testing when all the parameters converge at the same nearly-weak rate.
restrictions, estimated at slower rates, can be linearly combined so as to be estimated at a faster rate. Lee (2005) puts forward some high-level assumptions (see his assumptions (R) and (G)) to deal with the asymptotic singularity problem. We show that our setting allows us to perform standard Wald-type set estimation, even without maintaining Lee’s (2005) high-level assumptions.

From previous discussion and findings, we can guess that the Wald test statistic for \( H_0 \) can actually be written with a standard formula:

\[
\zeta^W_T = T g'(\hat{\theta}_T) \left\{ \frac{\partial g(\hat{\theta}_T)}{\partial \theta'} \left[ \frac{\partial^2 g(\theta)}{\partial \theta \partial \theta'} S^{-1}_T \frac{\partial g(\hat{\theta}_T)}{\partial \theta} \right]^{-1} \frac{\partial g(\hat{\theta}_T)}{\partial \theta} \right\}^{-1} g(\hat{\theta}_T)
\]

Recall the standard rank assumption ensuring that the Wald test statistic is asymptotically chi-square with \( q \) degrees of freedom:

\[
\text{Rank} \left[ \frac{\partial g(\theta)}{\partial \theta'} \right] = q
\]

for all \( \theta \) in the interior of \( \Theta \), or at least in a neighborhood of \( \theta^0 \). As well known, this condition is not restrictive, since it is akin to say that, at least locally, the \( q \) restrictions are linearly independent. Unfortunately, the coexistence of different rates of convergence may introduce (asymptotically) some perverse multicolinearity between the \( q \) estimated constraints.

The counterexample below points out the key issue.

**Example 4.1 (Counterexample)**

Consider only two groups of moment conditions (corresponding respectively to two rates of convergence \( \lambda_{1T} \) and \( \lambda_{2T} \)), and assume that we want to test \( H_0 : g(\theta) = 0 \) with \( g(\theta) = [g_j(\theta)]_{1 \leq j \leq q} \) and none of the \( q \) vectors \( \frac{\partial g_j(\theta^0)}{\partial \theta} \), \( j = 1, \ldots, q \) belongs to \( \text{Im}\left[ \frac{\partial \phi}{\partial \theta} \right] \).

Then the extension of the standard argument for Wald test would be to say that, under the null, \( \left[ \lambda_{2T} g(\hat{\theta}_T) \right] \) behaves asymptotically like \( \left[ (\frac{\partial g(\theta^0)}{\partial \theta}) \lambda_{2T} (\hat{\theta}_T - \theta^0) \right] \), that is, for large \( T \), \( \left[ \lambda_{2T} g(\hat{\theta}_T) \right] \) behaves like a gaussian

\[
\mathcal{N} \left( 0, \frac{\partial g(\theta^0)}{\partial \theta'} \left[ \frac{\partial^2 g(\theta^0)}{\partial \theta \partial \theta'} S^{-1}_T \frac{\partial g(\theta^0)}{\partial \theta} \right]^{-1} \frac{\partial g(\theta^0)}{\partial \theta} \right)
\]

\[\text{For any } (n \times m)\text{-matrix, } \text{Im} [M] \text{ represents the subspace of } \mathbb{R}^n \text{ generated by the column vectors of } M.\]

29
Imagine however that for some nonzero vector $\alpha$,

$$\alpha' \frac{\partial g(\theta_0)}{\partial \theta} = \sum_{j=1}^{q} \alpha_j \frac{\partial g_j(\theta_0)}{\partial \theta}$$

belongs to $\text{Im}[\partial \rho_1(\theta_0)/\partial \theta]$. Then, under the null, $\left[\lambda_1' \alpha' g(\hat{\theta}_T)\right]$ is asymptotically gaussian and thus

$$\lambda_2' \alpha' g(\hat{\theta}_T) = \frac{\lambda_2'}{\lambda_1} \lambda_1' \alpha' g(\hat{\theta}_T) \xrightarrow{P} 0$$

In other words, even if the $q$ constraints are locally linearly independent (or a full rank assumption is maintained), $\left[\lambda_2' g(\hat{\theta}_T)\right]$ does not behave asymptotically like a gaussian with a non-singular variance matrix. This is the reason why deriving an asymptotically chi-square distribution with $q$ degrees of freedom for the Wald test statistic is more involved than usual.

Lee (2005) avoids this asymptotic singularity by maintaining the following assumption:

**Lee’s (2005) assumption:**

There exists a sequence of $(q, q)$ invertible matrices $D_T$ such that for any $\theta \in \Theta$

$$\text{Plim}_{T \to \infty} \left[D_T \frac{\partial g(\theta)}{\partial \theta} R_0 [\tilde{\Lambda}_T]^{-1}\right] = B_0$$

where $B_0$ is a $(q, p)$ deterministic finite matrix of full row rank.

Lee’s (2005) assumption clearly implies the standard rank condition (4.2). However, the converse is not true as it can be shown from the counterexample above. And this is actually what is needed to justify the construction of a Wald-type confidence set, through the usual delta-theorem approach. The above assumption implies that, under the null, $D_T g(\hat{\theta}_T)$ behaves like $\left[D_T(\partial g(\theta^0)/\partial \theta)(\hat{\theta}_T - \theta^0)\right]$, that is like $\left[B_0^\top [\tilde{\Lambda}_T]^{-1} (\hat{\theta}_T - \theta^0)\right]$. From theorem 3.4, we know that $\left[\tilde{\Lambda}_T [R_0]^{-1} (\hat{\theta}_T - \theta^0)\right]$ nicely behaves as an asymptotic gaussian distribution. In other words, the matrix $D_T$ provides us with the right scaling to get asymptotic normality of $\left[(\partial g(\theta^0)/\partial \theta')(\hat{\theta}_T - \theta^0)\right]$. However, standard Wald-type confidence sets are valid even without Lee’s assumption:

**Theorem 4.1 (Wald test)**

Under the assumptions 1 to 6 (or 6') and if $g(.)$ is twice continuously differentiable, the Wald

---

13By contrast, in the case of only $q = 1$ constraint, Lee’s assumption is trivially fulfilled.
test statistic $\zeta_T^W$ (4.1) for testing $H_0: g(\theta) = 0$ is asymptotically distributed as a chi-square with $q$ degrees of freedom under the null.

While a detailed proof of theorem 4.1 is provided in the appendix, it is worth explaining why it works, in spite of the aforementioned singularity problem. The key intuition is related to the well-known phenomenon that the finite sample performance of the Wald test depends on the way the null hypothesis is formulated.\(^{14}\)

Consider a fictitious situation where the range of $[\partial \rho_1'(\theta^0)/\partial \theta]$ is known. Then, it is always possible to define a $(q,q)$-nonsingular matrix $H$ and $q$-dimensional function $h(\theta) = Hg(\theta)$ to ensure a genuine disentangling of the directions to be tested. By genuine disentangling, we mean that, for instance in the simpler case with only two different rates, for some $q_1$ such that $1 \leq q_1 \leq q$, we have:

- for $j = 1, \cdots, q_1$: $[\partial h_j(\theta^0)/\partial \theta]$ belongs to $Im[\partial \rho_1'(\theta^0)/\partial \theta]$
- for $j = q_1 + 1, \cdots, q$: $[\partial h_j(\theta^0)/\partial \theta]$ does not belong to $Im[\partial \rho_1'(\theta^0)/\partial \theta]$ and no linear combinations of them do.

Then, the perverse asymptotic singularity of example 4.1 is clearly avoided. Of course, at a deeper level, the new restrictions $h(\theta) = 0$ should be interpreted as a nonlinear transformation of the initial ones $g(\theta) = 0$ (since the matrix $H$ depends on $\theta$). It turns out that, for all practical purposes, by fictitiously seeing $H$ as known, the Wald-type test statistics written with $h(.)$ or $g(.)$ are numerically equal. The proof of theorem 4.1 shows that this is the key reason why standard Wald-type set estimation always works (despite appearing invalid at first sight).

As far as the size of the test is concerned, the existence of several rates of convergence does not modify the standard Wald result. Of course, the power of the test heavily depends on the strength of identification of the various constraints to test. More precisely, if, for the sake of notational simplicity, we consider only $q = 1$ restriction to test, and two rates of convergence, we get:

**Theorem 4.2 (Local alternatives)**

*Under assumptions 1 to 6 (or 6*), the Wald test of $H_0: g(\theta) = 0$ (with $g(.)$ one dimensional*.

\(^{14}\)In some respect, our approach complements the higher-order expansions of Phillips and Park (1988).
continuously differentiable) is consistent under the sequence of local alternatives $H_{1T}$: $g(\theta) = 1/\delta_T$ if and only if either

$$\frac{\partial g(\theta^0)}{\partial \theta} \in \text{Im} \left[ \frac{\partial \rho'_1(\theta^0)}{\partial \theta} \right] \quad \text{and} \quad \delta_T = o(\lambda_{1T})$$

or

$$\frac{\partial g(\theta^0)}{\partial \theta} \notin \text{Im} \left[ \frac{\partial \rho'_1(\theta^0)}{\partial \theta} \right] \quad \text{and} \quad \delta_T = o(\lambda_{2T})$$

The proof of theorem 4.2 is rather straightforward. A nonlinear function $g(.)$ of $\theta$, interpreted as $\left[g(\theta^0) + \frac{\partial g(\theta^0)}{\partial \theta}(\theta - \theta^0)\right]$, is identified at the fast rate $\lambda_{1T}$ if and only if

$$\frac{\partial g(\theta^0)}{\partial \theta^r} \in \text{Im} \left[ \frac{\partial \rho'_1(\theta^0)}{\partial \theta} \right]$$

As far as set estimation is concerned, it is worth realizing that we can compute, in a standard way, a confidence set with asymptotic level $(1 - \alpha)$, by considering the set of values $g \in \mathbb{R}^q$, such that:

$$T[g - g(\hat{\theta}_T)]' \left\{ \frac{\partial g(\hat{\theta}_T)}{\partial \theta^r} \left[ \frac{\partial \phi_T(\hat{\theta}_T)}{\partial \theta} S^{-1}_T \frac{\partial \phi_T(\hat{\theta}_T)}{\partial \theta^r} \right]^{-1} \frac{\partial g'(\hat{\theta}_T)}{\partial \theta} \right\}^{-1} \leq \chi^2_{1-\alpha}(q)$$

where $\chi^2_{1-\alpha}(q)$ is the $(1 - \alpha)$-quantile of a chi-square distribution with $q$ degrees of freedom. As already explained in section 3.3 (in the special case $g(\theta) = \theta$), for all practical purposes, the underlying singularities do not matter. They are actually hidden within the asymptotic behavior of $[\partial \phi_T(.)/\partial \theta^r]$. This result is strikingly reminiscent of the "Testing parameters in GMM without assuming that they are identified" of Kleibergen (2005). However, when one really wants to know the accuracy of information about a given component of $g(\theta)$, one needs to answer the kind of question asked in theorem 4.2.

5 Conclusion

This paper extends the asymptotic theory of GMM inference to allow sample counterparts of the estimating equations to converge at (multiple) rates, different from the usual square-root of the sample size. Many econometrics models consider simultaneously several rates of
convergence for the asymptotic identification of the structural parameters: our mixed-rates asymptotic theory is then well-suited. Some examples were detailed in section 2, including kernel smoothing, trimmed-mean estimation, mean excess function, infill asymptotic, social interactions, and nearly-weak identification.

In such a setting, we provided consistent estimation of the structural parameters. We actually stressed that such GMM estimators of the structural parameters are likely to be only slowly consistent. Then, we were able to disentangle and estimate the directions associated with the different rates of convergence. These well-suited linear combinations of the structural parameters were defined through a convenient and feasible rotation in the coordinate system. This is only with respect to these linear combinations that the issue of (asymptotic) efficiency can be considered. In addition to the standard efficient selection of the weighting matrix for minimum distance estimation, mixed-rates may allow the improvement of the fast directions through a control variable approach. And, in sharp contrast with standard GMM, this is not automatically achieved. We provided some results on how to improve fast directions, by using more efficiently the information contained in slower estimating equations.

Finally, we demonstrated the validity of usual inference procedures, like the overidentification test and Wald test, with standard formulas. It is important to stress that both estimation and testing work without requiring the knowledge of the various rates. However, as emphasized in the main text, their assessment is crucial for (asymptotic) power considerations. As suggested in an earlier draft of this paper, the subsampling approach of Bertail, Politis and Romano (1999) may be helpful to assess these rates. An adaptation of this idea has recently been developed by Caner (2008).

Overall, two main motivations may lead an applied econometrician to resort to the techniques developed in this paper. First, common inference procedures (J-test and Wald-test) are validated in quite uncommon, albeit empirically relevant, settings. Second, and even more importantly, this paper helps identify which directions in the parameter space are more or less accurately estimated. Examples where some economically meaningful directions are estimated at specific rates are put forward in the companion paper Antoine and Renault (2008), as well as in Gagliardini, Gouriéroux and Renault (2007).
References


Appendix

Proof of Equation (3.2): (Stronger identification property)
Let us denote by $S_\epsilon$ 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 $\alpha > 0$. ■

Proof of Theorem 3.1: (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

$$\|\rho(\hat{\theta}_T)\| = O_p \left( \frac{1}{\Lambda_T} \right)$$

Proof of lemma A.1: From (3.6), the objective function is written as follows

$$Q_T(\theta) = \left[ \frac{\Psi_T(\theta)}{T^{1/2}} + \frac{\Lambda_T}{T^{1/2}} \rho(\theta) \right]' \Omega_T \left[ \frac{\Psi_T(\theta)}{T^{1/2}} + \frac{\Lambda_T}{T^{1/2}} \rho(\theta) \right]$$

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

$$Q_T(\hat{\theta}_T) \leq Q_T(\theta^0)$$

$$\Rightarrow \left[ \frac{\Psi_T(\hat{\theta}_T)}{T^{1/2}} + \frac{\Lambda_T}{T^{1/2}} \rho(\hat{\theta}_T) \right]' \Omega_T \left[ \frac{\Psi_T(\hat{\theta}_T)}{T^{1/2}} + \frac{\Lambda_T}{T^{1/2}} \rho(\hat{\theta}_T) \right] \leq \frac{\Psi_T(\theta^0)}{T^{1/2}} \Omega_T \frac{\Psi_T(\theta^0)}{T^{1/2}}$$

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:

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

Let $\mu_T$ be the smallest eigenvalue of $\Omega_T$. The former inequality implies:

$$\mu_T \|\Lambda_T \rho(\hat{\theta}_T)\|^2 - 2 \|\Lambda_T \rho(\hat{\theta}_T)\| \times \|\Omega_T \Psi_T(\hat{\theta}_T)\| + d_T \leq 0$$
In other words, \( x_T = \| \Delta_T \rho(\hat{\theta}_T) \| \) 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 \( x_T \geq (\Delta_T) \| \rho(\hat{\theta}_T) \| \) we want to show that \( x_T = \mathcal{O}_P(1) \), that is
\[
\frac{\| \Omega_T \psi_T(\hat{\theta}_T) \|}{\mu_T} = \mathcal{O}_P(1) \quad \text{and} \quad \Delta_T = \mathcal{O}_P(1)
\]
which amounts to show that:
\[
\frac{\| \Omega_T \psi_T(\hat{\theta}_T) \|}{\mu_T} = \mathcal{O}_P(1) \quad \text{and} \quad \frac{d_T}{\mu_T} = \mathcal{O}_P(1)
\]
Note that since \( \text{det}(\Omega_T) \xrightarrow{P} \text{det}(\Omega) > 0 \), no subsequence of \( \mu_T \) can converge in probability towards zero and thus we can assume (for \( T \) sufficiently large) that \( \mu_T \) remains lower bounded away from zero with asymptotic probability one. Therefore, we just have to show that:
\[
\| \Omega_T \psi_T(\hat{\theta}_T) \| = \mathcal{O}_P(1) \quad \text{and} \quad d_T = \mathcal{O}_P(1)
\]
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 \( \Omega_T \). Therefore the required boundedness in probability just results from our functional CLT 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 \( \epsilon \) 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 \( \eta \):
\[
P \left[ \| \hat{\theta}_{T_n} - \theta^0 \| > \epsilon \right] \geq \eta \quad \text{for} \quad n \in \mathbb{N}
\]
Let us denote

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

by assumption 1. Then for all \(n \in \mathbb{N}\):

$$P\left[\|\rho(\hat{\theta}_T^n)\| \geq \alpha\right] \geq \eta > 0$$

This last inequality contradicts lemma A.1. This completes the proof of consistency. ■

**Proof of Theorem 3.2: (Rate of convergence)**

From lemma A.1, \(\|\rho(\hat{\theta}_T)\| = \|\rho(\hat{\theta}_T) - \rho(\theta^0)\| = O_P(1/\lambda T)\) and by application of the mean-value theorem, for some \(\tilde{\theta}_T\) between \(\hat{\theta}_T\) and \(\theta^0\) (component by component), we get:

$$\|\tilde{z}_T\| = \left\|\frac{\partial \rho(\hat{\theta}_T)}{\partial \theta'}(\hat{\theta}_T - \theta^0)\right\| = O_P\left(\frac{1}{\lambda T}\right)$$

Note that, by a common abuse of notation, we omit to stress that \(\tilde{\theta}_T\) actually depends on the component of \(\rho(.)\). Define now \(z_T\) as follows:

$$z_T \equiv \frac{\partial \rho(\theta^0)}{\partial \theta'}(\hat{\theta}_T - \theta^0) \quad (A.1)$$

and since \(\left[\partial \rho(\theta^0)/\partial \theta'\right]\) is full column rank:

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

Hence, we only need to prove that \(\|z_T\| = O_P(1/\lambda T)\) to get the desired result. By definition of \(z_T\) and \(\tilde{z}_T\), we have the following:

$$\tilde{z}_T = z_T + \left(\frac{\partial \rho(\hat{\theta}_T)}{\partial \theta'} - \frac{\partial \rho(\theta^0)}{\partial \theta'}\right)(\hat{\theta}_T - \theta^0) \quad (A.2)$$

with \(\|\tilde{z}_T\| = O_P(1/\lambda T)\). Moreover, since \(\rho(.)\) is continuously differentiable and \(\hat{\theta}_T\) (as well as \(\hat{\theta}_T\)) converges in probability towards \(\theta^0\), we also have:

$$\frac{\partial \rho(\hat{\theta}_T)}{\partial \theta'} \xrightarrow{P} \frac{\partial \rho(\theta^0)}{\partial \theta'} \quad \Rightarrow \quad \left\|\left(\frac{\partial \rho(\hat{\theta}_T)}{\partial \theta'} - \frac{\partial \rho(\theta^0)}{\partial \theta'}\right)(\hat{\theta}_T - \theta^0)\right\| = \epsilon_T \|z_T\| \quad \text{with} \quad \epsilon_T \to 0$$

We then conclude from the above and equation (A.2) that \(\|z_T\| = O_P(1/\lambda T)\). ■
Proof of Lemma 3.3: To get the results, we have to show the following:

i) (diagonal terms) \( \frac{T^{1/2} \partial \tilde{\phi}_{iT}(\theta^*_T)}{\lambda_{iT}} \xrightarrow{p} \frac{\partial \rho_i(\theta^0)}{\partial \theta'} \) for \( i = 1, \ldots, l \)

ii) lower diagonal \( \frac{T^{1/2} \partial \tilde{\phi}_{iT}(\theta^*_T)}{\lambda_{iT}} \xrightarrow{p} 0 \) for \( i = 2, \ldots, l \); with \( 1 \leq j < i \)

iii) upper diagonal \( \frac{T^{1/2} \partial \tilde{\phi}_{iT}(\theta^*_T)}{\lambda_{iT}} R_i \xrightarrow{p} 0 \) for \( i = 1, \ldots, l - 1 \); with \( l \geq j > i \)

i) From assumption 5(ii)
\[
T^{1/2} \frac{\partial \tilde{\phi}_{iT}(\theta^0)}{\partial \theta} \xrightarrow{p} \frac{\partial \rho_i(\theta^0)}{\partial \theta'} = O_P(1)
\]

A fortiori since \( \lambda_{iT} \rightarrow \infty \):
\[
\frac{T^{1/2} \partial \tilde{\phi}_{iT}(\theta^*_T)}{\lambda_{iT}} \xrightarrow{p} 0
\]

The mean-value theorem applied to the \( k \)-th component of \( [\partial \tilde{\phi}_{iT} / \partial \theta'] \), \( 1 \leq k \leq k_i \), for \( \tilde{\theta}_T^* \)
between \( \theta^0 \) and \( \theta^*_T \):
\[
\frac{T^{1/2}}{\lambda_{iT}} \left( \frac{\partial \tilde{\phi}_{iT,k}(\theta^*_T)}{\partial \theta'} - \frac{\partial \tilde{\phi}_{iT,k}(\theta^0)}{\partial \theta'} \right) = \frac{T^{1/2}}{\lambda_{iT}} (\theta^*_T - \theta^0)' \frac{\partial^2 \tilde{\phi}_{iT,k}(\tilde{\theta}_T)}{\partial \theta \partial \theta'} = o_P(1)
\]
because by assumption \( \|\theta^*_T - \theta^0\| = O_P(1/\lambda_{iT}) \) and by assumption 6 (or 6*),
\( (T^{1/2} / \lambda_{iT}) (\partial^2 \tilde{\phi}_{iT,k}(\theta)/\partial \theta \partial \theta') \xrightarrow{p} H(\theta) \). Hence we get the announced result i).

ii) This result directly follows from the proof of the above result i) \( \frac{T^{1/2}}{\lambda_{iT}} \left[ \frac{\partial \tilde{\phi}_{iT}(\theta^*_T)}{\partial \theta'} \right] \xrightarrow{p} \frac{\partial \rho_i(\theta^0)}{\partial \theta'} \)
and \( \lambda_{iT} = o(\lambda_{iT}) \):
\[
\frac{T^{1/2}}{\lambda_{iT}} \frac{\partial \tilde{\phi}_{iT}(\theta^*_T)}{\partial \theta'} \xrightarrow{p} 0
\]

iii) Again, we apply the mean-value theorem to the \( k \)-th component of \( [\left( \partial \tilde{\phi}_{iT,k}(\cdot)/\partial \theta' \right) R_j] \)
for \( 1 \leq k \leq k_i \), with \( \tilde{\theta}_T^* \) between \( \theta^0 \) and \( \theta^*_T^* \):
\[
\frac{T^{1/2}}{\lambda_{iT}} \frac{\partial \tilde{\phi}_{iT,k}(\theta^*_T)}{\partial \theta'} R_j = \frac{1}{\lambda_{iT}} \left[ \frac{T^{1/2}}{\lambda_{iT}} \frac{\partial \tilde{\phi}_{iT,k}(\theta^0)}{\partial \theta'} R_j \right] + \lambda_{iT} (\theta^*_T - \theta^0)' \frac{\partial^2 \tilde{\phi}_{iT,k}(\tilde{\theta}_T)}{\partial \theta \partial \theta'} R_j
\]

40
Now recall that \( \lambda_{IT} \| (\hat{\theta}_T - \theta^0) \| = \mathcal{O}_P(1); \frac{T^{1/2} \partial^2 \tilde{\phi}_{iT,k}(\theta)}{\partial \theta' \partial \theta'} \overset{P}{\to} H(\theta); \)

and also \( \lambda_{IT}/(\lambda_{IT} \lambda_{IT}) = \lambda_{IT}/\lambda_{IT}^2 \times \lambda_{IT}/\lambda_{IT} T \to 0 \) by assumption 6 (or 6*).

We just have to prove that the first element of the RHS converges to 0 in probability. From assumption 5(ii), we have:

\[
\frac{T^{1/2}}{\lambda_{IT}} \left[ \frac{\partial \tilde{\phi}_{iT,k}(\theta^0)}{\partial \theta'} R_j - \frac{\partial \hat{\rho}_l(\theta^0)}{\partial \theta'} R_j \right] = \mathcal{O}_P \left( \frac{1}{\lambda_{IT}} \right)
\]

and we get the result because \( (\partial \hat{\rho}_l(\theta^0)/\partial \theta') R_j = 0 \) by definition of \( R^0 \). ■

**Proof of Theorem 3.4: (Asymptotic normality)**

From the optimization problem (3.4), the first-order conditions for \( \hat{\theta}_T \) are written as:

\[
\frac{\partial \tilde{\phi}_{iT}(\hat{\theta}_T)}{\partial \theta} \Omega_T \hat{\theta}_T = 0
\]

A mean-value expansion yields to:

\[
\frac{\partial \tilde{\phi}_{iT}(\hat{\theta}_T)}{\partial \theta} \Omega_T \hat{\theta}_T = \frac{\partial \tilde{\phi}_{iT}(\hat{\theta}_T)}{\partial \theta} \Omega_T \hat{\theta}_T + \frac{\partial \tilde{\phi}_{iT}(\hat{\theta}_T)}{\partial \theta} \Omega_T \frac{\partial \tilde{\phi}_{iT}(\hat{\theta}_T)}{\partial \theta} \times (\hat{\theta}_T - \theta^0) = 0
\]

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

\[
\hat{J}_T^' \Omega_T \left[ \sqrt{T} \tilde{\phi}_{iT}(\theta^0) \right] + \hat{J}_T^' \Omega_T \hat{J}_T \times \hat{\Lambda}_T [R^0]^{-1} (\hat{\theta}_T - \theta^0) = 0
\]

after defining:

\[
\hat{J}_T = \sqrt{T} \frac{\partial \tilde{\phi}_{iT}(\hat{\theta}_T)}{\partial \theta'} R^0 \hat{\Lambda}_T^{-1} \quad \text{and} \quad \hat{J}_T = \sqrt{T} \frac{\partial \tilde{\phi}_{iT}(\hat{\theta}_T)}{\partial \theta'} R^0 \hat{\Lambda}_T^{-1}
\]

From theorem 3.2 and lemma 3.3, we can deduce that:

\[
P \lim \hat{J}_T = J^0 \quad \text{and} \quad P \lim \hat{J}_T = J^0
\]

Hence,

\[
\hat{J}_T^' \Omega_T \hat{J}_T \overset{P}{\to} J^0 \Omega_J^0 \quad \text{nonsingular by assumption}
\]

Recall now that by assumption 2(i), \( \Psi_T(\theta^0) = \sqrt{T} \tilde{\phi}_{iT}(\theta^0) \) converges to a normal distribution with mean 0 and variance \( S^0 \). We then get the announced result. ■
Proof of Theorem 3.5:
Directly follows from theorem 3.4 and the discussion in the main text. □

Proof of Theorem 3.6: (J-test)
A Taylor expansion of order 1 of the moment conditions gives:

$$\sqrt{T} \Theta_T(\theta_T) = \sqrt{T} \Theta_T(\theta^0) + \sqrt{T} \frac{\partial \Theta_T(\theta_T)}{\partial \theta_T} (\hat{\theta}_T - \theta^0) + o_p(1)$$

$$= \sqrt{T} \Theta_T(\theta^0) + J_T \Lambda_T [R^0]^{-1} (\hat{\theta}_T - \theta^0) + o_p(1)$$

with \( J_T = \sqrt{T} \left[ \partial \Theta_T(\hat{\theta}_T) / \partial \theta_T \right] R^0 \Lambda_T^{-1} \).

A Taylor expansion of the FOC gives:

$$\Lambda_T [R^0]^{-1} (\hat{\theta}_T - \theta^0) = - \left[ \left( \sqrt{T} \frac{\partial \Theta_T(\theta_T)}{\partial \theta_T} R^0 \Lambda_T^{-1} \right)' S_T^{-1} \left( \sqrt{T} \frac{\partial \Theta_T(\theta_T)}{\partial \theta_T} R^0 \Lambda_T^{-1} \right) \right]^{-1} \times \left( \sqrt{T} \frac{\partial \Theta_T(\hat{\theta}_T)}{\partial \theta_T} R^0 \Lambda_T^{-1} \right)' S_T^{-1} \sqrt{T} \Theta_T(\theta^0) + o_p(1)$$

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

Combining the 2 above results leads to:

$$\sqrt{T} \Theta_T(\hat{\theta}_T) = \sqrt{T} \Theta_T(\theta^0) - J_T \left[ J_T S_T^{-1} J_T \right]^{-1} J_T S_T^{-1} \sqrt{T} \Theta_T(\theta^0) + o_p(1)$$

Use the previous result to rewrite the criterion function:

$$T Q_T(\hat{\theta}_T) = \left[ \sqrt{T} \Theta_T(\hat{\theta}_T) \right]' S_T^{-1} \sqrt{T} \Theta_T(\hat{\theta}_T)$$

$$= \left[ \sqrt{T} \Theta_T(\theta^0) - J_T \left[ J_T S_T^{-1} J_T \right]^{-1} J_T S_T^{-1} \sqrt{T} \Theta_T(\theta^0) \right]' S_T^{-1} \times \left[ \sqrt{T} \Theta_T(\theta^0) - J_T \left[ J_T S_T^{-1} J_T \right]^{-1} J_T S_T^{-1} \sqrt{T} \Theta_T(\theta^0) \right] + o_p(1)$$

$$= \left[ \sqrt{T} \Theta_T(\theta^0) \right]' S_T^{-1} \sqrt{T} \Theta_T(\theta^0)$$

$$- \sqrt{T} \Theta_T(\theta^0) S_T^{-1} J_T \left[ J_T S_T^{-1} J_T \right]^{-1} J_T S_T^{-1} \sqrt{T} \Theta_T(\theta^0) + o_p(1)$$

$$= \sqrt{T} \Theta_T(\theta^0)' S_T^{-1/2} \left[ I - M \right]^{-1} S_T^{1/2} \sqrt{T} \Theta_T(\theta^0) + o_p(1)$$

42
where $S_T^{1/2}$ is such that $S_T = S_T^* S_T^{-1/2}$ and $M = S_T^{-1/2} J_T \left[ J_T' S_T^{-1} J_T \right]^{-1} J_T' S_T^{-1/2}$ which is a projection matrix of rank $(K - p)$. The expected result follows. ■

Proof of Theorem 3.7: (Asymptotic normality in the extended case)
Recall:
\[
\tilde{\rho}_T(\theta) = \rho_1(\theta) - B \frac{A_{\text{up}(1),T}}{\lambda_{1T}} \rho_{\text{up}(1)}(\theta) \quad \text{and} \quad \tilde{\rho}_T(\theta) = [\rho_1'(\theta) \rho_2'(\theta) \cdots \rho_l'(\theta)]'
\]
where $B$ is a deterministic matrix. Since $\lambda_{1T} = o(\lambda_{1T})$, and $\rho(.)$ is continuous on $\Theta$ compact,
\[
\tilde{\rho}_T(\theta) \xrightarrow{T} \rho(\theta) \quad \text{uniformly in } \theta
\]
Note also that $\tilde{\rho}_T(\theta_0)$ is the deterministic sequence of 0. To prove asymptotic normality in the extended case, we prove first theorems 3.1, 3.2 and lemma 3.3 in the extended case.

Theorem A.2 (Consistency in the extended case)
Any minimum distance estimator $\hat{\theta}_T^{(B)}$ like (3.27) is weakly consistent.

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

Lemma A.3
\[
\| \rho_T(\hat{\theta}_T^{(B)}) \| = O_P \left( \frac{1}{\Delta_T} \right)
\]

Proof of lemma A.3:
The proof of the above lemma works very similarly to the one of lemma A.1, precisely because $\tilde{\rho}_T(\theta_0) = 0$ for any $T$. Let us then deduce the weak consistency of $\hat{\theta}_T^{(B)}$ by a contradiction argument. If $\hat{\theta}_T^{(B)}$ was not consistent, there would exist some positive $\epsilon$ such that:
\[
P \left[ \| \hat{\theta}_T^{(B)} - \theta_0 \| > \epsilon \right]
\]
does not converge to zero. Then we can define a subsequence $(\hat{\theta}_T^{(B)})_{n \in N}$ such that, for some positive $\eta$:
\[
P \left[ \| \hat{\theta}_T^{(B)} - \theta_0 \| > \epsilon \right] \geq \eta \quad \text{for } n \in N.
\]
Let us denote
\[
\alpha = \inf_{\| \theta - \theta_0 \| > \epsilon} \| \rho(\theta) \| > 0 \quad \text{by assumption 1}
\]
Then, we have:
\[
0 < \alpha \leq \| \rho(\theta) \| \leq \| \tilde{\rho}_T(\theta) \| + \| \rho(\theta) - \tilde{\rho}_T(\theta) \| \Rightarrow \inf_{\| \theta - \theta_0 \| > \epsilon} \| \tilde{\rho}_T(\theta) \| > 0
\]
because, by uniform convergence, for \( T \) large enough, \( \| \rho(\theta) - \tilde{\rho}_T(\theta) \| < \alpha/2 \). Then for all \( n \in \mathbb{N} \): \( P \left[ \| \tilde{\rho}_T(\hat{\theta}_{T_n}) \| \geq \alpha \right] > 0 \). This last inequality contradicts lemma A.3. This completes the proof of consistency. ■

**Theorem A.4** (Rate of convergence in the extended case)

\[
\left\| \hat{\theta}^{(B)}_T - \theta^0 \right\| = \mathcal{O}_P \left( \frac{1}{\Delta T} \right)
\]

**Proof:**

Equation (A.2) now becomes:

\[
\tilde{z}_{2T} = z_T + \left( \frac{\partial \tilde{\rho}_T(\hat{\theta}_T)}{\partial \theta'} - \frac{\partial \tilde{\rho}_T(\theta^0)}{\partial \theta'} + \frac{\partial \tilde{\rho}_T(\theta^0)}{\partial \theta'} - \frac{\partial \rho(\theta^0)}{\partial \theta'} \right) (\hat{\theta}_T - \theta^0)
\]

where \( z_T \) is defined as in equation (A.1) and \( \tilde{z}_{2T} \) is such that:

\[
\left\| \tilde{z}_{2T} \right\| = \left\| \frac{\partial \tilde{\rho}_T(\hat{\theta}_T)}{\partial \theta'} (\hat{\theta}_T - \theta^0) \right\| = \mathcal{O}_P \left( \frac{1}{\Delta T} \right)
\]

Here again, we only need to show that \( \| z_T \| = \mathcal{O}_P(1/\Delta T) \) to get the desired result. By combining the uniform convergence of \( \partial \tilde{\rho}_T(\cdot)/\partial \theta' \) and a method similar to the original proof, we can also show that:

\[
\left\| \left( \frac{\partial \tilde{\rho}_T(\hat{\theta}_T)}{\partial \theta'} - \frac{\partial \tilde{\rho}_T(\theta^0)}{\partial \theta'} + \frac{\partial \tilde{\rho}_T(\theta^0)}{\partial \theta'} - \frac{\partial \rho(\theta^0)}{\partial \theta'} \right) (\hat{\theta}_T - \theta^0) \right\| = \epsilon_{2T} \| z_T \| \text{ with } \epsilon_{2T} \to 0
\]

We then conclude from the above that \( \| z_T \| = \mathcal{O}_P(1/\Delta T) \). ■

**Lemma A.5** (Lemma 3.3 in the extended case)

If \( \hat{\theta}_T^* \) is such that \( \| \hat{\theta}_T^* - \theta^0 \| = \mathcal{O}_P(1/\lambda_T) \) then

\[
T^{1/2} \frac{\partial \tilde{\phi}_T(\theta^*_T)}{\partial \theta'} \mathbb{R}^0 \tilde{\Lambda}^{-1}_T \xrightarrow{P} \mathcal{J}^0
\]

If we compare the result in the extended case and the original one, we only need to prove one convergence result (case (i) when \( i = 1 \) in the existing proof), that is:

\[
\frac{T^{1/2} \partial \tilde{\phi}_1(\theta^*_T)}{\lambda_1 T} \xrightarrow{P} \frac{\partial \rho_1(\theta^0)}{\partial \theta'}
\]
We have:

\[
\frac{T^{1/2}}{\lambda_{1T}} \partial \tilde{\phi}_{1T}(\theta_T^*) = \frac{T^{1/2}}{\lambda_{1T}} \partial \tilde{\phi}_{1T}(\theta_T^*) - B \frac{T^{1/2}}{\lambda_{1T}} \partial \tilde{\phi}_{up(1),T}(\theta_T^*)
\]

by definition

by applying lemma 3.3. ■

We now come back to the prove of theorem 3.7. We can simply mimic the proof of theorem 3.4. We just replace \( \phi_T(\cdot) \) by \( \tilde{\phi}_T(\cdot) \) and \( \hat{\theta}_T \) by \( \hat{\theta}_T(\cdot) \), because all the required intermediate results (theorems 3.1, 3.2, and lemma 3.3) have been proved in the extended case. ■

**Proof of Theorem 3.8:** (Optimal choice for \( B \))

The proof is decomposed into two steps: in step 1, we show that any set of valid moment conditions like (3.25) leads to the same orthogonalized set of moment conditions (with \( B = A_1 \)); in step 2, we show that \( AVar(\hat{\eta}_{1T}^{(A_1)}) \leq AVar(\hat{\eta}_{1T}) \). The desired result directly follows from steps 1 and 2.

- Step 1: Consider any matrix \( B \) that leads to \( \tilde{\phi}_{1T}(\theta^0) \) as in theorem 3.7. The associated orthogonalized moment conditions are:

\[
\tilde{\phi}_T(\theta^0) = \begin{bmatrix} \tilde{\phi}_{1T}^{(B)}(\theta^0) - \hat{A}_1 \tilde{\phi}_{up(1),T}(\theta^0) \\ \tilde{\phi}_{up(1),T}(\theta^0) \end{bmatrix}
\]

where

\[
\hat{A}_1 = \lim_{T \to \infty} \left\{ Cov[\tilde{\phi}_{1T}^{(B)}(\theta^0), \tilde{\phi}_{up(1),T}(\theta^0)] \left[ Var(\tilde{\phi}_{up(1),T}(\theta^0)) \right]^{-1} \right\}
\]

\[
= \lim_{T \to \infty} \left\{ Cov[\tilde{\phi}_{1T}(\theta^0) - B \tilde{\phi}_{up(1),T}(\theta^0), \tilde{\phi}_{up(1),T}(\theta^0)] \left[ Var(\tilde{\phi}_{up(1),T}(\theta^0)) \right]^{-1} \right\}
\]

\[
= \lim_{T \to \infty} \left\{ Cov[\tilde{\phi}_{1T}(\theta^0), \tilde{\phi}_{up(1),T}(\theta^0)] \left[ Var(\tilde{\phi}_{up(1),T}(\theta^0)) \right]^{-1} \right\} - B
\]

Hence, we have:

\[
\tilde{\phi}_{1T}(\theta^0) = \begin{bmatrix} \tilde{\phi}_{1T}(\theta^0) - B \tilde{\phi}_{up(1),T}(\theta^0) \\ \tilde{\phi}_{1T}(\theta^0) - A_1 \tilde{\phi}_{up(1),T}(\theta^0) \end{bmatrix} - [A_1 - B] \tilde{\phi}_{up(1),T}(\theta^0)
\]

\[
= \tilde{\phi}_{1T}(\theta^0) - A_1 \tilde{\phi}_{up(1),T}(\theta^0)
\]

\[
= \tilde{\phi}_{1T}^{(A_1)}(\theta^0)
\]
We need to compare: $\phi_T(\theta) = \lim_{T \to \infty} [V ar(\sqrt{T}\phi_T(\theta^0))]$ and the partition introduced in section 3.4, $\phi_T(\theta^0) = [\phi_{1T}(\theta^0) \phi_{up(1),T}(\theta^0)]'$. We consider accordingly the appropriate partition of $S^0$:

$$S^0 = \begin{pmatrix} S^0_1 & S^0_{1,up(1)} \\ S^0_{up(1),1} & S^0_{up(1)} \end{pmatrix}$$

Recall also the inverse formulas:

$$[S^0]^{-1} = \begin{bmatrix} [S^0]^{-1} & \left[ S^0_{1,1}^{-1} \right]^{-1} \\ -P^{-1}S_{up(1),1} & [S^0]^{-1} \end{bmatrix}$$

with $Q = S^0_1 - S^0_{1,1}^{-1}S^0_{up(1),1}^{-1}S^0_{up(1),1}$ and $P = S^0_{up(1)} - S^0_{up(1),1}[S^0]^{-1}S^0_{1,1}^{-1}$. Recall the matrix $J^0$ as defined in lemma 3.3. We also consider its appropriate partition:

$$J^0 = \begin{pmatrix} \frac{\partial \rho_1(\theta^0)}{\partial \theta} R_1 & 0 \\ 0 & \frac{\partial \rho_{up(1)}(\theta^0)}{\partial \theta} R_{up(1)} \end{pmatrix}$$

Recall from theorems 3.5 and 3.7:

$$AV ar(\hat{\eta}_{IT}) = \left\{ J^0 \left[ S^0 \right]^{-1} J^0 \right\}^{-1}$$

$$AV ar(\hat{\gamma}_{IT}^{(A)}) = \left\{ J^0 \left[ S^{(A)} \right]^{-1} J^0 \right\}^{-1}$$

where $S^{(A)} = V ar[\tilde{\Psi}(\theta^0)] = \left[ \begin{array}{c} \Psi_1(\theta^0) - A_1 \Psi_{up(1)}(\theta^0) \\ \Psi_{up(1)}(\theta^0) \end{array} \right]$. We also consider its appropriate partition:

$$S^{(A)} = \begin{pmatrix} S^{(A)_1} & 0 \\ 0 & S^{(A)_{up(1)}} \end{pmatrix}$$

where

$$S^{(A)_1} = S^0_1 - A_1 S^0_{up(1),1} A_1' = S^0_1 - S^0_{1,1}^{-1} S^0_{up(1),1}^{-1} S^0_{up(1),1}$$

$$S^{(A)_{up(1)}} = S^0_{up(1)}$$

We need to compare: $AV ar(\hat{\eta}_{IT})$ and $AV ar(\hat{\gamma}_{IT}^{(A)})$. Straightforward calculations lead to:

$$\left[ J^0 \left[ S^{(A)_1} \right]^{-1} J^0 \right]^{-1} = \begin{pmatrix} \tilde{R}_1 Q^{-1} \tilde{R}_1 & 0 \\ 0 & \tilde{R}_{up(1)} S^0_{up(1),1}^{-1} \tilde{R}_{up(1)} \end{pmatrix}^{-1}$$
The last 2 terms of the RHS can be rewritten as follows:

\[ \tilde{R}_{1} = \frac{\partial \rho_{1}(\theta^{0})}{\partial \theta^{0}} - R_{1} \]

\[ \tilde{R}_{up(1)} = \frac{\partial \rho_{up(1)}(\theta^{0})}{\partial \theta^{0}} - R_{up(1)} \]

\[ Q = S_{1}^{0} - S_{1,up(1)}^{0}[S_{up(1)}^{0}]^{-1}S_{up(1),1}^{0} \]

On the other hand, we have:

\[ \tilde{R}_{1} = \frac{\partial \rho_{1}(\theta^{0})}{\partial \theta^{0}} - R_{1} \]

\[ \tilde{R}_{up(1)} = \frac{\partial \rho_{up(1)}(\theta^{0})}{\partial \theta^{0}} - R_{up(1)} \]

\[ Q = S_{1}^{0} - S_{1,up(1)}^{0}[S_{up(1)}^{0}]^{-1}S_{up(1),1}^{0} \]

Hence:

\[ AVar(\hat{\eta}_{1T}) = A^{-1} + A^{-1}B(D - B'A^{-1}B)^{-1}B'A^{-1} \]

and \( AVar(\hat{\eta}_{1T}^{(A_{1})}) = [\tilde{R}_{1}Q^{-1}\tilde{R}_{1}]^{-1} \)

where \( A^{-1} = [\tilde{R}_{1}Q^{-1}\tilde{R}_{1}] \). We only have to study the 2nd term of the RHS of \( AVar(\hat{\eta}_{1T}) \).

After realizing that \( AVar(\hat{\eta}_{up(1),T}) = (D - B'A^{-1}B)^{-1} \geq 0 \),

we conclude that \( [A^{-1}B(D - B'A^{-1}B)^{-1}B'A^{-1}] \) is positive semi-definite.

Hence: \( AVar(\hat{\eta}_{1T}) \geq AVar(\hat{\eta}_{1T}^{(A_{1})}) \).

**Proof of Example 3.1:**

Recall the notations introduced in the proof of theorem 3.8. Now, we have:

\[ \tilde{\varphi}_{T}(\theta^{0}) = [\tilde{\varphi}_{1T}(\theta^{0}) \tilde{\varphi}_{up(1),T}(\theta^{0})]' = [\tilde{\varphi}_{1T}(\theta^{0}) \tilde{\varphi}_{2T}(\theta^{0})]' \]

We also get:

\[ AVar(\hat{\eta}_{1T}^{(A_{1})}) = [\tilde{R}_{1}^{0}[S_{2}^{0}]^{-1}\tilde{R}_{2}]^{-1} \]

and \( AVar(\hat{\eta}_{2T}) = (D - B'A^{-1}B)^{-1} \)

\[ D - B'A^{-1}B = \tilde{R}_{2}^{0}[S_{2}^{0}]^{-1}\tilde{R}_{2} + \tilde{R}_{2}^{0}[S_{2}^{0}]^{-1}S_{21}^{0}Q^{-1}S_{12}^{0}[S_{2}^{0}]^{-1}\tilde{R}_{2} 
\]

\[ -\tilde{R}_{2}^{0}[S_{2}^{0}]^{-1}S_{21}^{0}Q^{-1}\tilde{R}_{1} \]

\[ [\tilde{R}_{1}Q^{-1}\tilde{R}_{1}]^{-1} \]

\[ \tilde{R}_{1}Q^{-1}S_{12}^{0}[S_{2}^{0}]^{-1}\tilde{R}_{2} \]

The last 2 terms of the RHS can be rewritten as follows:

\[ \tilde{R}_{1}^{0}[S_{2}^{0}]^{-1}S_{21}^{0}Q^{-1}S_{12}^{0}[S_{2}^{0}]^{-1}\tilde{R}_{2} - \tilde{R}_{1}^{0}[S_{2}^{0}]^{-1}S_{21}^{0}Q^{-1}\tilde{R}_{1} \]

\[ [\tilde{R}_{1}Q^{-1}\tilde{R}_{1}]^{-1} \]

\[ \tilde{R}_{1}Q^{-1}S_{12}^{0}[S_{2}^{0}]^{-1}\tilde{R}_{2} \]

\[ = \tilde{R}_{2}^{0}[S_{2}^{0}]^{-1}S_{21}^{0} \left\{ Q^{-1} - Q^{-1}\tilde{R}_{1} \left[ \tilde{R}_{1}Q^{-1}\tilde{R}_{1} \right]^{-1} \right\} \tilde{R}_{1}Q^{-1}S_{12}^{0}[S_{2}^{0}]^{-1}\tilde{R}_{2} \]

47
Finally, we have:

$$Q^{-1} - Q^{-1} \tilde{R}_1 \left[ \tilde{R}_1' Q^{-1} \tilde{R}_1 \right]^{-1} \tilde{R}_1' Q^{-1}$$

$$= Q^{-1/2} \left\{ I - Q^{-1/2} \tilde{R}_1 \left[ \tilde{R}_1' Q^{-1/2} Q^{-1/2} \tilde{R}_1 \right]^{-1} \tilde{R}_1' Q^{-1/2} \right\} Q^{-1/2}$$

$$= Q^{-1/2} \left\{ I - X(X'X)^{-1} X' \right\} Q^{-1/2}$$

$$= Q^{-1/2} M_X Q^{-1/2}$$

with $Q^{-1} \equiv Q^{-1/2} Q^{-1/2}$, $X \equiv Q^{-1/2} \tilde{R}_1$ and $M_X \equiv I - X(X'X)^{-1} X'$.

Finally, we have:

$$D - B'A^{-1} B = \tilde{R}_2 S_2^{-1} \tilde{R}_2 + \left( Q^{-1/2} S_1^{0} S_2^{-1} \tilde{R}_2 \right)' M_X \left( Q^{-1/2} S_1^{0} S_2^{-1} \tilde{R}_2 \right)$$

because by definition, $M_X$ is a projection matrix. Hence it is positive semi-definite as well as $H'M_XH$ for any matrix $H$. We can then conclude: $AVar(\eta_{2T}) \leq AVar(\hat{\eta}(A_1))$.

**Proof of Theorem 3.9:** (Feasible asymptotic normality)

From theorem 3.4, $\hat{\lambda}_T [R^0]^{-1}(\hat{\theta}_T - \theta^0)$ is asymptotically normally distributed. We now show that the above convergence is not altered when $R^0$ is replaced by $\tilde{R}$, some $\lambda_{IT}$-consistent estimator. To simplify the calculations, rewrite $[R^0]^{-1}$ and $\tilde{R}^{-1}$ as follows:

$$[R^0]^{-1} = \begin{pmatrix} R^1 \\ R^2 \\ \vdots \\ R^l \end{pmatrix} \quad \text{and} \quad \begin{pmatrix} \tilde{R}^1 \\ \tilde{R}^2 \\ \vdots \\ \tilde{R}^l \end{pmatrix}^{-1}$$

Then,

$$\tilde{\lambda}_T [R^0]^{-1}(\hat{\theta}_T - \theta^0) = \left[ \lambda_{IT} R^i(\hat{\theta}_T - \theta^0) \right]_{1 \leq i \leq l} \quad \text{and} \quad \tilde{\lambda}_T \tilde{R}^{-1}(\hat{\theta}_T - \theta^0) = \left[ \lambda_{IT} \tilde{R}^i(\hat{\theta}_T - \theta^0) \right]_{1 \leq i \leq l}$$

We need to show that, for any component $i$:

$$\lambda_{IT} \tilde{R}^i(\hat{\theta}_T - \theta^0) = \lambda_{IT} R^i(\hat{\theta}_T - \theta^0) + o_P(1)$$

- For $i = l$, we have:

$$\lambda_{IT} \tilde{R}^l(\hat{\theta}_T - \theta^0) = \lambda_{IT} R^l(\hat{\theta}_T - \theta^0) + \lambda_{IT} (\tilde{R}^l - R^l)(\hat{\theta}_T - \theta^0)$$

$$= \lambda_{IT} R^l(\hat{\theta}_T - \theta^0) + (\tilde{R}^l - R^l)\lambda_{IT}(\hat{\theta}_T - \theta^0)$$
From theorem 3.2, $\lambda_{IT}(\hat{\theta}_T - \theta^0) = O_P(1)$. Hence, the second term of the RHS is negligible is front of the first one and we get the desired result.
- For $1 \leq i \leq l - 1$, we have:
  \[
  \lambda_{IT} \hat{R}^i(\hat{\theta}_T - \theta^0) = \lambda_{IT} R^i(\hat{\theta}_T - \theta^0) + \frac{\lambda_{IT}}{\lambda_{IT}} (\hat{R} - R) \lambda_{IT} (\hat{\theta}_T - \theta^0)
  \]

From theorem 3.4, $(1) = O_P(1)$ and from theorem 3.2, $\lambda_{IT}(\hat{\theta}_T - \theta^0) = O_P(1)$. We need to show that $(2)$ is negligible in front of $(1)$ for any $i$:

\[
(2) \prec (1) \quad \forall \ i \quad \Leftrightarrow \quad \frac{\lambda_{IT}}{\lambda_{IT}} (\hat{R} - R) = o_P(1) \quad \forall \ i
\]

\[
\Leftrightarrow \quad \hat{R} - R = o_P \left( \frac{\lambda_{IT}}{\lambda_{IT}} \right) \quad \forall \ i
\]

\[
\Leftrightarrow \quad \frac{1}{\lambda_{IT}} = o \left( \frac{\lambda_{IT}}{\lambda_{IT}} \right) \quad \forall \ i
\]

\[
\Leftrightarrow \quad \lambda_{IT} = o(\lambda_{IT}) \quad \forall \ i
\]

\[
\Leftarrow \quad \text{Assumption 6*(i)}
\]

\[\blacksquare\]

**Proof of Theorem 4.1:** *(Wald test)*

To simplify the exposition, the proof is performed with only 2 groups of moment conditions associated with 2 rates. The proof is divided into two steps:

- step 1: we define an algebraically equivalent formulation of $H_0 : g(\theta) = 0$ as $H_0 : h(\theta) = 0$ such that its first components are identified at the fast rate $\lambda_{1T}$, while the remaining ones are identified at the slow rate $\lambda_{2T}$ without any linear combinations of the latter being identified at the fast rate.

- step 2: we show that the Wald test statistic on $H_0 : h(\theta) = 0$ asymptotically converges to the proper chi-square distribution with $q$ degrees of freedom and that it is numerically equal to the Wald test statistic on $H_0 : g(\theta) = 0$.

- Step 1: The space of fast directions to be tested is:
  \[
  I^0(g) = \left[ Im \frac{\partial g'(\theta^0)}{\partial \theta} \right] \cap \left[ Im \frac{\partial \rho^2_{1}(\theta^0)}{\partial \theta} \right]
  \]
Denote $n^0(g)$ the dimension of $I^0(g)$. Then, among the $q$ restrictions to be tested, $n^0(g)$ are identified at the fast rate and the $(q - n^0(g))$ remaining ones are identified at the slow rate.

Define $q$ vectors of $\mathbb{R}^q$ denoted as $\epsilon_j (j = 1, \cdots, q)$ such that $[(\partial g'(\theta^0)/\partial \theta) \times \epsilon_j]_{j=1}^{q}$ is a basis of $I^0(g)$ and $[(\partial g'(\theta^0)/\partial \theta) \times \epsilon_j]_{j=1}^{q}$ is a basis of

$$[I^0(g)]^\perp \cap \left[1m \frac{\partial g'(\theta^0)}{\partial \theta} \right]$$

We can then define a new formulation of the null hypothesis $H_0 : g(\theta) = 0$ as, $H_0 : h(\theta) = 0$

where $h(\theta) = Hg(\theta)$ with $H$ invertible matrix such that $H' = [\epsilon_1 \cdots \epsilon_q]$. The two formulations are algebraically equivalent since $h(\theta) = 0 \iff g(\theta) = 0$. Moreover,

$$\lim_{T \to \infty} DT \frac{\partial h(\theta^0)}{\partial \theta'} R^0 [\hat{\Lambda}_T]^{-1} = B^0$$

with $DT$ a $(q, q)$ invertible diagonal matrix with its first $n^0(g)$ coefficients equal to $\lambda_{1T}$ and the $(p - n^0(g))$ remaining ones equal to $\lambda_{2T}$ and $B^0$ a $(q, p)$ matrix with full column rank.

- Step 2: first we show that the 2 induced Wald test statistics are numerically equal.

$$\zeta^W_T (g) = T g'(\hat{\theta}_T) \left\{ \frac{\partial g(\hat{\theta}_T)}{\partial \theta'} \left[ \frac{\partial \varphi_T (\hat{\theta}_T)}{\partial \theta} S_T^{-1} \frac{\partial \varphi_T (\hat{\theta}_T)}{\partial \theta'} \right]^{-1} \frac{\partial g'(\hat{\theta}_T)}{\partial \theta} \right\}^{-1} g(\hat{\theta}_T)$$

$$= T H' g'(\hat{\theta}_T) \left\{ H \frac{\partial g(\hat{\theta}_T)}{\partial \theta'} \left[ \frac{\partial \varphi_T (\hat{\theta}_T)}{\partial \theta} S_T^{-1} \frac{\partial \varphi_T (\hat{\theta}_T)}{\partial \theta'} \right]^{-1} \frac{\partial g'(\hat{\theta}_T)}{\partial \theta} H' \right\}^{-1} H g(\hat{\theta}_T)$$

$$= \zeta^W_T (h)$$

Then we show $\zeta^W_T (h)$ is asymptotically distributed as a chi-square with $q$ degrees of freedom. First we need a preliminary result which naturally extends the above convergence towards $B^0$ when $\hat{\theta}^0$ is replaced by a $\lambda_{2T}$-consistent estimator $\hat{\theta}_T$:

$$\lim_{T \to \infty} DT \frac{\partial h(\hat{\theta}_T)}{\partial \eta'} [\hat{\Lambda}_T]^{-1} = B^0$$

The proof is very similar to lemma 3.3 and is not reproduced here. The fact that $g(\cdot)$ is twice continuously differentiable is needed for this proof.
The Wald test statistic on \( h(.) \) can be written as follows:

\[
\xi^W_T(h) = T \left[ D_T h(\hat{\theta}_T) \right]' \left\{ D_T \frac{\partial h(\hat{\theta}_T)}{\partial \theta'} \left[ \frac{\partial^2 \phi_T(\hat{\theta}_T)}{\partial \theta^2} S_T^{-1} \frac{\partial \phi_T(\hat{\theta}_T)}{\partial \theta} \right]^{-1} \frac{\partial h'(\hat{\theta}_T)}{\partial \theta} D_T \right\}^{-1} \left[ D_T h(\hat{\theta}_T) \right]
\]

where \( \hat{J}_T \equiv \sqrt{T} \frac{\partial \phi_T(\hat{\theta}_T)}{\partial \theta} \Lambda_T^{-1} \) with \( \hat{J}_T \stackrel{P}{\rightarrow} J^0 \) and \( \hat{J}_T S_T^{-1} \hat{J}_T \stackrel{P}{\rightarrow} J^0 [S(\theta^0)]^{-1} J^0 \equiv \Sigma \).

Now from the mean-value theorem under \( H_0 \) we deduce:

\[
D_T h(\hat{\theta}_T) = D_T \frac{\partial h(\hat{\theta}_T^*)}{\partial \theta'} (\hat{\theta}_T - \theta^0) = \left[ D_T \frac{\partial h(\hat{\theta}_T^*)}{\partial \theta'} R^0 \Lambda_T^{-1} \right] \Lambda_T [R^0]^{-1} (\hat{\theta}_T - \theta^0)
\]

with \( \left[ D_T \frac{\partial h(\hat{\theta}_T^*)}{\partial \theta'} R^0 \Lambda_T^{-1} \right] \stackrel{P}{\rightarrow} B^0 \) and \( \Lambda_T [R^0]^{-1} (\hat{\theta}_T - \theta^0) \stackrel{d}{\rightarrow} N(0, \Sigma^{-1}) \)

Finally we get

\[
\xi^W_T(h) = \left[ \Lambda_T [R^0]^{-1} (\hat{\theta}_T - \theta^0) \right]' B_0^0 (B_0 \Sigma B_0^0)^{-1} B_0 \left[ \Lambda_T [R^0]^{-1} (\hat{\theta}_T - \theta^0) \right] + o_P(1)
\]

Following the proof of theorem 3.6 we get the expected result.■