10ISSN 1183-1057

SIMON FRASER UNIVERSITY

Department of Economics

Working Papers

12-03

"Efficient Minimum Distance Estimation with Multiple Rates of Convergence"

Bertille Antoine
and
Eric Renault

April, 2010

SFU Economics

Efficient Minimum Distance Estimation with Multiple Rates of Convergence *

Bertille Antoine† and Eric Renault‡

April 7, 2010

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) 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; Kernel estimation; 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 the following: the minimum distance estimator $\hat{\theta}_T$ of the vector θ of parameters (with θ^0 as true unknown value),

$$\hat{\theta}_T = \arg\min_{\theta} \left[m_T'(\theta) \Omega m_T(\theta) \right] \tag{1.1}$$

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

$$\sqrt{T}(\hat{\theta}_T - \theta^0) = - \left[\frac{\partial m_T'(\theta^0)}{\partial \theta} \Omega \frac{\partial m_T(\theta^0)}{\partial \theta'} \right]^{-1} \frac{\partial m_T'(\theta_0)}{\partial \theta} \Omega \sqrt{T} m_T(\theta^0) + o_P(1) \tag{1.2}$$

$$\text{while} \quad \texttt{Plim}\left[m_T(\theta) \right] = 0 \Longleftrightarrow \theta = \theta^0 \tag{1.3}$$

There are many cases (see section 2 for examples and a literature review), including local smoothing, trimming, infill asymptotic, or any kind of non-root T asymptotics, where the asymptotic normality of $m_T(\theta^0)$ comes at a non-standard rate of convergence: $\left[T^{\alpha} m_T(\theta^0) \right]$ is asymptotically a non-degenerated gaussian variable for some $\alpha \neq 1/2$. This does not invalidate the first-order expansion argument (1.2) since $\left[T^{\alpha}(\hat{\theta}_T - \theta^0) \right]$ is asymptotically equivalent to $\left[T^{\alpha} m_T(\theta^0) \right]$. 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). As a result, similarly to the Hill estimator, asymptotic normality is reached at a rate different from square-root of T, while standard GMM formulas for asymptotic covariance matrices hold.

This paper focuses on the more involved case where identification of θ comes from moment-based pieces of information, possibly coming at different rates of convergence. In this case, no exponent α allows the characterization of a non-degenerate asymptotic distribution for $\left[T^{\alpha}(\hat{\theta}_T - \theta^0) \right]$ 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 (2008) has addressed this issue in the general setting of extremum estimation, the specificity of GMM (or minimum distance) estimation yields to more explicit results for point estimates and corresponding power of Wald-type tests and confidence sets.

The contribution of this paper is threefold. First, an empirical process approach allows us to prove the consistency of estimators of structural parameters at the slowest available rate of convergence. Second, special linear combinations of the parameters associated with specific rates of convergence are identified and efficiently estimated. Third, we show that inference procedures, like the overidentification test and Wald-type test, remain valid 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 to get efficient estimators. Kitamura (1996) and, to some extent, Sims, Stock and Watson (1990) develop a testing strategy with 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 since we 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. Two identification approaches and associated consistent estimators of the structural parameters are also discussed. Section 3 proves the consistency of the two GMM estimators of structural parameters θ associated with the two former identifying approaches. Only one of them enables the fast convergence of directions in the parameter space, after a convenient reparametrization. In section 4, 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. The issue of inference of functions of the structural parameters is addressed in section 5. Section 6 illustrates the above theory with a Monte Carlo study of a stochastic volatility model with option price data. Section 7 concludes. All the proofs are gathered in the appendix.

2 Examples and Identification

2.1 Two motivating examples

We start with two econometric models where different rates of convergence must be considered simultaneously for asymptotic identification of the vector of structural parameters.

Example 1 *(Kernel smoothing)*

Assume we observe a time series $(X_t, Y_t), t = 1, ..., T$ on a stationary process with stationary distribution denoted as distribution of (X, Y). Consider a Nadaraya-Watson estimator of the conditional expectation $E[g(Y, \theta)|X = x)$ of a known function of some unknown parameters θ. 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 an (under-smoothing) bandwidth sequence considered with a suitable exponent, the kernel estimator $m_T(\theta)$ of $E[g(Y, \theta)|X = x)$ will be such that $\sqrt{Th_T}\{m_T(\theta) - E[g(Y, \theta)|X = x]\}$ is asymptotically gaussian with zero mean.

Assume now that for inference about the true unknown value θ^0 of θ, the estimating equation $E[g(Y, \theta^0)|X = x] = 0$ is valid for a given value x of the conditioning variable while it may not be uniformly valid over all the support of X. For instance, Gagliardini, Gouriéroux and Renault (2009) consider such conditional expectations produced by Euler optimality conditions on an asset pricing model, where the pricing kernel is parameterized by θ. In their case $E[g(Y_t, \theta)|X_t = x_t)$ stands for the price a time t of some financial asset and the lack of uniformity over all the values of X comes from the fact that such a price is observed at only one given date (see section 6 below for a more explicit example). Then,

$$\sqrt{T}\left[\overline{\phi}_T(\theta) - \frac{\lambda_T}{\sqrt{T}}\rho(\theta)\right]$$

is asymptotically gaussian, where $\rho(\theta) = E[g(Y, \theta)|X = x)$, $\lambda_T = \sqrt{Th_T} \overset{T}{\to} \infty$, but slower than $\sqrt{T}$, and $\overline{\phi}_T(\theta) \equiv \sqrt{h_T} m_T(\theta)$. Euler optimality conditions are fulfilled for the true unknown value θ^0 of θ: $\rho(\theta^0) = 0$. Note that in this example $\overline{\phi}_T(\theta)$ is a sample mean of a double array:

$$\overline{\phi}_T(\theta) = \frac{1}{T}\sum_{t=1}^{T} \phi_{t,T}(\theta) \quad \text{where} \quad \phi_{t,T}(\theta) = \sqrt{h_T}\left[\sum_{s=1}^{T} K\left(\frac{x_s - x}{h_T}\right)\right]^{-1} K\left(\frac{x_t - x}{h_T}\right) g(y_t, \theta)$$

Suppose now that several conditional expectations are informative about θ. 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 $\overline{\phi}_T(\theta)$ and $\rho(\theta)$ such that, for each component i:

$$\sqrt{T}\left[\overline{\phi}_{iT}(\theta) - \frac{\lambda_{iT}}{\sqrt{T}}\rho_i(\theta)\right]$$

is asymptotically gaussian, where $\lambda_{iT} = \sqrt{Th_{iT}}$ are heterogeneous due to different bandwidths choices h_{iT}. Inference about the unknown value θ^0 of θ can be performed from the maintained assumption $\rho_i(\theta^0) = 0$ for all i.

In the asset pricing example of Gagliardini, Gouriéroux and Renault (2009), 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, estimating equations $\rho_i(\theta^0) = 0$ are defined through unconditional moments with square-root T consistent sample counterparts. Hence, the associated rate is simply $\lambda_{iT} = \sqrt{T}$. However, due to market incompleteness, these assets are not sufficient to identify θ^0 and additional estimating functions estimated only with nonparametric rates of convergence are needed. These estimating functions are provided by conditional moment restrictions only valid at a finite set of dates.

Example 2 *(Nearly-weak instruments)*

For GMM with nearly-weak instruments, as introduced by Caner (2008) for a non-linear extension of Hahn and Kuersteiner (2002), the correlation between the instruments and the first-order conditions declines at a rate slower than root-T. Both Caner (2008) and Antoine and Renault (2009) 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 (2009) consider the case where both strong and nearly-weak instruments are simultaneously needed to identify two groups of directions in the

parameter space at rates root-T and a slower one respectively. 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). More formally, Antoine and Renault (2009) consider a set of moments computed as a sample mean of a double array:

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

while an assumed drifting DGP provides population moments as:

$$E\left[\overline{\phi}_T(\theta)\right] = \frac{\Lambda_T}{\sqrt{T}}\rho(\theta)$$

with $\rho(\theta^0) = 0$, and Λ_T a diagonal matrix whose diagonal coefficients λ_{iT} all go to infinity with T, but possibly slower than $\sqrt{T}$.

2.2 A new approach to identification

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

Assumption 1 *(Identifying equations)*
$\theta \longrightarrow \rho(\theta)$ is a continuous function from a compact parameter space $\Theta \subset \mathbb{R}^p$ into $\mathbb{R}^K$ such that: $\rho(\theta) = 0 \iff \theta = \theta^0$.

Assumption 1 implies that θ^0 is a well-separated zero of the above equation, that is:

$$\forall \epsilon > 0 \quad \inf_{\|\theta-\theta^0\|\geq\epsilon} \|\rho(\theta)\| > 0 \tag{2.1}$$

This is all we need to prove consistency of a classical minimum distance estimator of θ (see e.g. chapter 5 in van der Vaart (1998))[1], when we have at our disposal some sample counterparts

[1]The standard distinction between global assumptions for consistency and local assumptions for asymptotic distributional theory (see e.g. Pakes and Pollard (1989)) could also be 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 θ^0.

$\overline{\zeta}_T(\theta)$ of the estimating equations that converge in probability uniformly on the set Θ of parameters:

$$\texttt{Plim}\left[\sup_{\theta\in\Theta}\left|\overline{\zeta}_T(\theta)-\rho(\theta)\right|\right]=0 \tag{2.2}$$

A classical minimum distance consistent estimator of θ is then defined as solution of:

$$\min_{\theta}\left[\overline{\zeta}_T(\theta)'\Omega_T\overline{\zeta}_T(\theta)\right] \tag{2.3}$$

where Ω_T is a sequence of symmetric positive definite random matrices which converges in probability towards a positive definite matrix Ω.

The key new insight of this paper is that the estimator (2.3), albeit consistent for any choice of the limit weighting matrix Ω, cannot in general efficiently use the informational content of the identifying assumption we have in mind. As in Examples 1 and 2 above, this identification assumption rests upon an empirical process approach[2]:

Assumption 2 *(Functional CLT)*
(i) The empirical process $(\Psi_T(\theta))_{\theta\in\Theta}$ *obeys a functional central limit theorem:*

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

where $\Psi(\theta)$ *is a gaussian stochastic process on* Θ *with mean zero and* $\Rightarrow$ *denotes weak convergence for the sup-norm on* Θ*.*
(ii) Λ_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:*

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

We explain now why, in general, Assumption 2 makes any estimator (2.3) inefficient. The suitable framework for (2.3) requires a preliminary rescaling of the moment conditions in order to satisfy the consistency condition (2.2):

$$\overline{\zeta}_T(\theta)=\sqrt{T}\Lambda_T^{-1}\overline{\phi}_T(\theta) \tag{2.4}$$

[2]The standard minimum distance theory corresponds to the special case where all diagonal coefficients of the matrix Λ_T are equal to $T^{1/2}$. In the context of Example 1, some diagonal coefficients are like $(Th_T)^{1/2}$, where the bandwidth parameter h_T goes to zero when T goes to infinity; however Th_T still goes to infinity. In the nearly-weak instruments case of Example 2, the fact that even $\underline{\lambda}_T$ goes to infinity makes the difference with the "actual" weak instruments setting of Stock and Wright (2000).

The above rescaling may be unfeasible since it requires the prior knowledge of the matrix Λ_T of rates of convergence (for instance, in the nearly-weak case as in Example 2, Λ_T is unknown). In addition, this rescaling may not be appropriate. Consider Example 1 when the moment conditions mix standard $\sqrt{T}$ consistent sample means with $\sqrt{Th_T}$ kernel smoothing estimators:

$$\begin{aligned} \overline{\phi}_{1T}(\theta) &= \frac{1}{T}\sum_{t=1}^{T}\phi_1(\theta) \quad \text{and} \quad \overline{\phi}_{2T}(\theta) = \frac{1}{T}\sum_{t=1}^{T}\phi_{2t,T}(\theta) \\ \text{where} \quad \phi_{2t,T}(\theta) &= \sqrt{h_T}\left[\frac{1}{T}\sum_{s=1}^{T}K\left(\frac{x_s - x}{h_T}\right)\right]^{-1} K\left(\frac{x_t - x}{h_T}\right) y_t \end{aligned}$$

The two components of $\overline{\zeta}_T(\theta)$ are respectively $\overline{\phi}_{1T}(\theta)$ and $\left[\frac{1}{\sqrt{h_T}}\overline{\phi}_{2T}(\theta)\right]$. Both converge respectively (possibly uniformly with respect to θ) towards an unconditional, respectively conditional, mathematical expectation $\rho_1(\theta)$, respectively $\rho_2(\theta)$. Then, the shortcoming of (2.3) is to give weights with the same order of asymptotic magnitude (through the fixed weighting matrix Ω) to $\sqrt{T}\overline{\phi}_{1T}(\theta^0)$, which is asymptotically normal, and to $\left[\frac{\sqrt{T}}{\sqrt{h_T}}\overline{\phi}_{2T}(\theta^0)\right]$, which blows up asymptotically.

Intuitively, the nonparametric component, associated with a slower rate of convergence, should rather be downplayed with respect to the standard sample mean which brings relevant information at a parametric rate. This is exactly what the class of minimum distance estimators studied in this paper does:

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

$$\hat{\theta}_T = \arg\min_{\theta\in\Theta}\left[Q_T(\theta)\right] \qquad \textit{where} \qquad Q_T(\theta) = \overline{\phi}_T'(\theta)\Omega_T\overline{\phi}_T(\theta) \tag{2.5}$$

It is important to realize that, in general, the above class of minimum distance estimators is not nested into the class (2.3). Let us for instance focus on the special GMM case, where, as in Examples 1 and 2, $\overline{\phi}_T(\theta)$ is a sample mean and $\overline{\zeta}_T(\theta) = \sqrt{T}\Lambda_T^{-1}\overline{\phi}_T(\theta)$ is the sample mean of suitably rescaled moments. Clearly, due to this rescaling, the difference between (2.3) and (2.5) only goes through the set of acceptable sequences of weighting matrices Ω_T. Since these sequences are immaterial in the just-identified case ($\dim(\phi) = \dim(\theta)$), there is no difference

between (2.3) and (2.5) for just-identified GMM[3]. Interestingly enough, there is no difference either when the matrix Λ_T features only one rate of convergence, $\Lambda_T = \lambda_T Id$, since:

$$\arg\min_{\theta\in\Theta}\left[\overline{\phi}_T'(\theta)\Omega_T\overline{\phi}_T(\theta)\right] = \arg\min_{\theta\in\Theta}\left\{\left[\frac{\sqrt{T}}{\lambda_T}\overline{\phi}_T(\theta)\right]'\Omega_T\left[\frac{\sqrt{T}}{\lambda_T}\overline{\phi}_T(\theta)\right]\right\}$$

As already stressed in the Introduction, non-trivial cases occur when at least two different rates of convergence are involved: in such cases, the above equivalence does not hold anymore. Of course, one can always write:

$$\arg\min_{\theta\in\Theta}\left[\overline{\phi}_T'(\theta)\Omega_T\overline{\phi}_T(\theta)\right] = \arg\min_{\theta\in\Theta}\left\{\left[\sqrt{T}\Lambda_T^{-1}\overline{\phi}_T(\theta)\right]'\Lambda_T\Omega_T\Lambda_T\left[\sqrt{T}\Lambda_T^{-1}\overline{\phi}_T(\theta)\right]\right\}$$

but, when $\texttt{Plim}[\Omega_T] = \Omega$, the sequence of rescaled matrices $[\Lambda_T\Omega_T\Lambda_T]$ does not admit a limit anymore. In this paper, we show that the efficient choice of Ω is the inverse of the asymptotic variance matrix of $\sqrt{T}\overline{\phi}_T(\theta^0)$; however, the corresponding efficient GMM estimator does not belong to the class (2.3). As explained above, (2.3) does not allow efficient estimation because it gives too much weight to some low quality information. The price to pay for our new approach is that identification cannot be reached by an argument *à la* Jennrich (1969) based on some uniform law of large numbers. By contrast with $\overline{\zeta}_T(\theta)$, the probability limit of $\overline{\phi}_T(\theta)$ does not bring identification in general since, for each component i such that $\frac{\lambda_{iT}}{\sqrt{T}}$ goes to zero, $\overline{\phi}_{iT}(\theta)$ goes to zero for all $\theta \in \Theta$. The identification power of the estimating equations $\rho(\theta) = 0$ appears to be wasted by the low quality of their sample counterparts. Fortunately, the following consistency result holds:

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

Since existence of a consistent estimator of θ means in particular that θ is identified, we claim that Assumption 2 brings a new approach to identification. Of course, Assumption 2 implies (2.2) which is a sufficient condition for identification *à la* Jennrich (1969). However the proof of our consistency result does not rely at all on such an argument since it does not work

[3]Since M-estimators can always be interpreted as just-identified GMM through first order conditions, the distinction between (2.3) and (2.5) is immaterial also in the case of M-estimators with non standard rates of convergence (see van der Vaart (1998), Chapter 5.8)

with the same class of weighting matrices. On the contrary, the proof relies heavily upon the empirical process approach of Assumption 2. Actually, we claim that identification is reached even without the knowledge of the (multiple) rates of convergence at play in the matrix Λ_T, while the argument based on (2.2) would be unfeasible in this case.

2.3 More examples

The following examples provide possible applications where several rates of convergence may be considered simultaneously. The formal treatment of these examples is not provided here.

Example 3 *(Trimmed-mean estimation)*

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

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

and the standard asymptotic distributional theory does not work when $\text{Var}[y_{it}(\theta)]$ is infinite. Hill and Renault (2009) 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 $\overline{Y}_{iT}(\theta)$ with:

$$m_{iT}(\theta) \;=\; \frac{1}{T}\sum_{t=1}^{T} m_{it,T}(\theta) \quad \text{where} \quad m_{it,T}(\theta) = \begin{cases} y_{it}(\theta) & \text{if } |y_{it}(\theta)| < c_{iT} \\ 0 & \text{otherwise} \end{cases}$$

The truncation threshold c_{iT} is such that $c_{iT} \overset{T}{\to} \infty$ to get asymptotic unbiased moments. In practice we do not choose the truncation threshold but rather the proportion k_{iT} of observations we trim: we replace the residual $y_{it}(\theta)$ by zero every time its absolute value exceeds the k_{iT}/T sample quantile. In other words, we have an implicit trimming threshold c_{iT} such that:

$$\lim_{T\to\infty} \frac{T}{k_{iT}} P\left[\left|y_{iT}(\theta^0)\right| > c_{iT}\right] = 1$$

Taking for instance $k_{iT} = T^{\lambda_i}, 0 < \lambda_i < 1$, may allow to control for infinite variance. Then the eventual rates of convergence of the components of the GMM estimator will depend on both

the truncation parameters λ_i and the tail parameter of the variable $\left|y_{iT}(\theta^0)\right|$. It is shown in Hill and Renault (2009) that by such tail trimming, GMM can achieve both asymptotic normality and nonstandard rates of convergence. Yet the rates of convergence will be dampened precisely due to trimming. Moroever, 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[4]. Note that inference about θ will not require knowledge of the rates of convergence since the GMM weighting matrix computed from trimmed moment conditions will self-normalize.

Example 4 *(Mean excess function)*

In a way somewhat symmetric to Example 3, 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 to estimate the parameters of a bivariate extreme copula. To apply this idea in a dimension larger than 2, one may have to consider different selection rates $[n_{iT}/T]$ 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 5 *(Infill asymptotic)*

In the above examples, rates of convergence slower than square-root of T show up because only part of the sample is actually used for estimation. Such rates may also occur because asymptotic theory is based on increasingly dense observations in a fixed bounded region. In fixed-domain asymptotic (or infill asymptotic), it 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

[4]The asymptotic theory developed in this paper requires the differentiability of the sample counterparts of the estimating equations (m_T here). Example 3 violates this smoothness condition, even asymptotically, since c_{iT} goes to infinity. This non-trivial issue, addressed in Hill and Renault (2009), is beyond the scope of the present paper.

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 but 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 (2007) 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 setting, 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 lead to MLE of some parameters associated with a slower rate of convergence.

Example 6 *(Social interactions)*

A social interaction model considers 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 (2010) 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 (2010) 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.

3 Rate of Convergence

3.1 Minimum rate of convergence

Following van der Vaart (1998) (see Theorem 5.52), the rate of convergence of an extremum estimator depends on the combined behavior of two maps involving respectively the regularity of the limit (deterministic) criterion function in the neighborhood of the true value, and the regularity of the asymptotic approximation: "if the deterministic map changes rapidly as θ moves away from the point of minimum and the random fluctuations are small, then $\hat{\theta}_T$ has a high rate of convergence". Since we focus on the weakness of some asymptotic approximations, we do not introduce any singularity issue 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* Θ.
(ii) θ^0 *belongs to the interior of* Θ.
(iii) The $(K \times p)$*-matrix* $[\partial\rho(\theta)/\partial\theta']$ *has full column rank* p *for all* $\theta \in \Theta$.
(iv) $T^{1/2}\Lambda_T^{-1}\frac{\partial\overline{\phi}_T(\theta)}{\partial\theta}$ *converges in probability towards* $\frac{\partial\rho'(\theta)}{\partial\theta}$ *uniformly on* $\theta \in \Theta$.

The rate of convergence assumed for the Jacobian matrix of moment conditions in Assumption 3 corresponds to the one assumed for moment conditions in Assumption 2. While automatic in the linear case, we need to assume that rates of convergence are maintained after differentiation with respect to the parameters. Under the maintained Assumption 3, the respective pros and cons of the two alternative minimum distance approaches (2.3) and (2.5) can easily be characterized. The rescaling (2.4) ensures that the second-derivative of the limit objective function corresponding to (2.3) is non-singular at the true value θ^0:

$$\begin{aligned} \texttt{Plim}\left[\overline{\zeta}_T(\theta)'\Omega_T\overline{\zeta}_T(\theta)\right] &= Q_\infty^*(\theta) \\ \frac{\partial^2 Q_\infty^*(\theta^0)}{\partial\theta\partial\theta'} &= 2\frac{\partial\rho'(\theta^0)}{\partial\theta}\Omega\frac{\partial\rho(\theta^0)}{\partial\theta'} \quad \text{non-singular} \end{aligned}$$

However, as already mentioned, this non-singularity comes at the cost of missing the suitable downplaying of low quality information. By contrast, while renouncing to a non-singular second-derivative of the limit criterion function, our preferred estimator (2.5) ensures that, at least in some directions, "the random fluctuations are small". More precisely, for components

i such that $\lambda_{iT}/T^{1/2}$ goes to zero, $\overline{\phi}_{iT}(\theta)$ (and also $\partial\overline{\phi}_{iT}(\theta)/\partial\theta$ goes to zero for all $\theta \in \Theta$: this obviously implies singularities in the second-derivative matrix of $\texttt{Plim}\left[\overline{\phi}_T(\theta)'\Omega_T\overline{\phi}_T(\theta)\right]$. To better understand why Assumption 2 ensures a minimum rate of convergence for $\hat{\theta}_T$ in spite of these singularities, we rewrite the minimization program (2.5) as follows:

$$\hat{\theta}_T = \arg\min_{\theta\in\Theta}\left[\frac{\Psi_T(\theta)}{T^{1/2}} + \frac{\Lambda_T}{T^{1/2}}\rho(\theta)\right]' \Omega_T \left[\frac{\Psi_T(\theta)}{T^{1/2}} + \frac{\Lambda_T}{T^{1/2}}\rho(\theta)\right] \tag{3.1}$$

When some diagonal coefficients of Λ_T go to infinity slower than $T^{1/2}$, the corresponding components of $\rho(\theta)$ are squeezed to zero in the optimization problem (3.1): their identifying power might then be lost. This explains why we need the empirical process approach[5]. The functional CLT (see Assumption 2) controls $\Psi_T(\theta)$ uniformly on Θ, and takes advantage of the identifying Assumption 1 in the minimization problem (3.1). More precisely, while $\Psi_T(\theta)$ is uniformly $\mathcal{O}_P(1)$, we show (see Lemma A.1 in the appendix) that:

$$\|\rho(\hat{\theta}_T)\| = \mathcal{O}_P\left(\frac{1}{\underline{\lambda}_T}\right) \tag{3.2}$$

And the (minimum) rate of convergence of $\hat{\theta}_T$ follows by Assumption 3:

Theorem 3.1 *Under Assumptions 1 to 3, we have:*

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

where $\underline{\lambda}_T$ has been defined in Assumption 2 as the minimal coefficient of Λ_T.

A special case of our result has been stated by Lee (2005): in his setting, $\Psi_T(\theta)$ does not depend on θ and, thus, tightness is no longer an issue. A similar simplification happens in the case of instrumental variables estimation of a linear regression model with weak instruments, as in Staiger and Stock (1997), and Hahn and Kuersteiner (2002). However, Theorem 3.1 is quite a poor result since it assigns the slowest possible rate of convergence to all components of the structural parameters. We now identify faster directions in the parameter space.

[5]This has already been pointed out: see Stock and Wright (2000).

3.2 What about faster convergence?

Without loss of generality, we consider Λ_T, the diagonal matrix with the following blocks[6]:

$$\Lambda_T = \begin{pmatrix} \lambda_{1T} I_{k_1} & & & \\ & \lambda_{2T} I_{k_2} & & \bigcirc \\ & & \ddots & \\ & \bigcirc & & \lambda_{lT} I_{k_l} \end{pmatrix} \quad \text{with} \quad \begin{cases} i) & \sum_{i=1}^{l} k_i = K \\ ii) & \lim\limits_{T\to\infty} \lambda_{iT} = \infty \quad \text{for} \quad i = 1, ..., l \\ iii) & \lambda_{i+1,T} = o(\lambda_{i,T}) \quad \text{for} \quad i = 1, ..., l-1 \end{cases}$$

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

$$\rho(\theta) = [\rho_1'(\theta)\ \rho_2'(\theta)\ \cdots\ \rho_l'(\theta)]' \quad \text{with} \quad \dim[\rho_i(\theta)] = k_i \quad \text{for } i = 1, ..., l \tag{3.3}$$

$$\overline{\phi}_T(\theta) = [\overline{\phi}_{1T}'(\theta)\ \overline{\phi}_{2T}'(\theta)\ \cdots\ \overline{\phi}_{lT}'(\theta)]' \quad \text{with} \quad \dim[\overline{\phi}_{iT}(\theta)] = k_i \quad \text{for } i = 1, ..., l \tag{3.4}$$

Throughout the paper we assume the correct partition of the K moment conditions between subsets of respective sizes $(k_1, k_2, \cdots, k_l)$ to be known. Typically, this is the case when they correspond to different smoothing or trimming schemes. Denote by $\mathtt{rk}[M]$ the rank of any square matrix M. Assumption 3(iii) is now reinforced as follows:

Assumption 4 *(Reinforced Assumption 3(iii))*
There exist non-negative integers s_i, for $i = 1, ..., l$, such that for all θ in the interior of Θ:

$$\mathtt{rk}\left[J_i(\theta)\right] = s_1 + s_2 + \cdots + s_i$$

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

The various ranks $s_1, s_2, \cdots, s_l$ are assumed to be known. Since 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 λ_{1T}. These first k_1 equations can be used in a standard way. Unfortunately, in general, the rank of the associated Jacobian $J_1(\theta^0)$ is lower than the dimension

[6] I_k represents the identity matrix of size k. When there is no ambiguity, the index k might be dropped.

of the parameter space ($s_1 < p$). Thus, these estimating equations are not sufficient to identify the entire parameter θ. Intuitively, they only identify the s_1 directions in the p-dimensional space of parameters which correspond to $\texttt{col}\left[J_1(\theta^0)\right]$.[7]

(ii) Similarly, $\texttt{col}\left[J_2(\theta^0)\right]$ characterizes the $(s_1 + s_2)$ directions in the parameter space that can be estimated at least at rate λ_{2T}. However, since s_1 directions (out of the former) can be estimated faster (at rate λ_{1T}), it is crucial for efficient estimation to disentangle them.

(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 parameter space will 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. To characterize these subspaces, we define recursively a sequence of matrices R_i, $i = 1, \cdots, l$, as follows:

(i) First, the orthogonal space of $\texttt{col}\left[J_{l-1}(\theta^0)\right]$ in $\mathbb{R}^p$ entails s_l directions that can only be estimated at the slowest rate λ_{lT}. Consequently, we display a basis[8] of this space as the columns of a matrix R_l of size (p, s_l), such that:

$$\frac{\partial \rho_i(\theta^0)}{\partial \theta'} R_l = 0 \quad \text{for } i < l \quad \text{with} \quad \texttt{rk}[R_l] = s_l$$

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

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

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

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

(iv) Finally, we choose R_1, of size (p, s_1), such that $\texttt{rk}\left[R_1\ R_2\ \cdots\ R_l\right] = p$.

We do not formally preclude that $s_i = 0$ for some i. If it is the case, the construction of the matrix R_i is skipped. In any case, we consider the matrix $R^0 = [R_1\ R_2\ \cdots\ R_l]$. By

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

[8] We could consider, more generally, any set of s_l linearly independent vectors of $\mathbb{R}^p$ which are not in the column space col $\left[J_{l-1}(\theta^0)\right]$. We choose to focus on orthogonal directions for the sake of expositional simplicity.

construction, R^0 is a (p,p) non-singular matrix of change of the basis in $\mathbb{R}^p$. Consider the associated new parametrization:

$$\eta = \left[R^0\right]^{-1}\theta = [\eta_i]_{1\leq i\leq l} \quad \text{with } \dim(\eta_i) = s_i \text{ for } i = 1,\cdots,l. \tag{3.5}$$

The exponent "0" remains in R^0 to stress that this matrix is defined as a function of the true unknown value θ^0. Even if the above reparametrization (3.5) is not feasible in practice, it helps to disentangle the various rates of convergence. More precisely, we must keep in mind that there is no hope, in general, to ensure that 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, $\hat{\theta}_T$ is generally asymptotically equivalent to some linear transformation of $\overline{\phi}_T(\theta)$, which likely mixes up all the components of $\overline{\phi}_T(\theta)$: as a result, components of $\hat{\theta}_T$ are contaminated by the slow rates of convergence. The above reparametrization conveniently isolates the various rates[9]. Consider now the reparametrized estimating equations: $\rho^*(\eta) = \rho(R^0\eta)$. First-order identification of η comes through the matrix $[\partial\rho^*(\eta)/\partial\eta'] = \left[\partial\rho(R^0\eta)/\partial\theta'\right] R^0$. It is lower triangular for $\eta^0 = [R^0]^{-1}\theta^0$ since,

$$\frac{\partial\rho_i(\theta^0)}{\partial\theta'}R_j = 0 \quad \text{for } i,j = 1,2,3,\cdots,l \text{ and } i < j \tag{3.6}$$

Under convenient assumptions, we show in Section 4 that this lower triangularity ensures:

$$\lambda_{iT}[\hat{\eta}_{iT} - \eta_i^0] = \mathcal{O}_P(1) \quad \text{for } i = 1,\cdots,l, \quad \text{where} \quad \hat{\eta}_T = [\hat{\eta}_{iT}]_{1\leq i\leq l} = [R^0]^{-1}\hat{\theta}_T \tag{3.7}$$

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 λ_{jT} for $j > i$) of the sample counterpart $\overline{\phi}_{iT}(\theta)$ of the estimating equation $\rho_i(\theta)$.

[9]The change of basis matrix R^0 has been built with orthogonal subspaces only for a matter of convenience in the interpretation of new parameters η. It is worth realizing that only the column spaces matter. For instance, in the just-identified case, $J_l(\theta)$ is a square non-singular matrix of size p, and nothing prevents us from running the change of basis with $R^0 = J_l(\theta^0)$. However, as mentioned in Section 2, the focus of this paper on a new minimum distance approach for identification is irrelevant in the just-identified case.

4 Asymptotic Distribution Theory

4.1 Asymptotic normality and efficiency

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

Assumption 5 *For all* $i = 1, \cdots, l$*:*
(i) $\left[\frac{T^{1/2}}{\lambda_{iT}}\frac{\partial\overline{\phi}'_{iT}(\theta)}{\partial\theta}\right]$ *converges in probability towards* $\frac{\partial\rho'_i(\theta)}{\partial\theta}$ *uniformly on* $\theta \in \Theta$*.*
(ii)

$$\frac{\partial\Psi'_{iT}(\theta^0)}{\partial\theta} = T^{1/2}\left[\frac{\partial\overline{\phi}'_{iT}(\theta^0)}{\partial\theta} - \frac{\lambda_{iT}}{T^{1/2}}\frac{\partial\rho'_i(\theta^0)}{\partial\theta}\right] = \mathcal{O}_P(1)$$

The above assumption would be ensured by an empirical process approach on $\left[\frac{\partial\overline{\phi}'_T(\theta)}{\partial\theta}\right]$, similar to the one adopted on $\left[\overline{\phi}_T(\theta)\right]$ in Assumption 2. In this respect, Assumption 5 is akin to assuming that Assumption 2 is maintained after differentiation with respect to θ.[10]

For sake of expositional simplicity, our asymptotic distributional theory focuses on the situation where parameters η_j for $j > i$ (estimated at slower rates than η_i) can be treated as nuisance parameters, without any impact on the asymptotic distribution of the estimator of η_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 orthogonality between the different parameters[11].

More precisely, consider the unfeasible minimum distance estimation problem:

$$\min_{\eta}\left[\overline{\phi}'_T(R^0\eta)\Omega_T\overline{\phi}_T(R^0\eta)\right] \tag{4.1}$$

The associated first-order conditions can be written as:

$$R^{0\prime}\frac{\partial\overline{\phi}'_T(R^0\hat{\eta}_T)}{\partial\theta}\Omega_T\overline{\phi}_T(R^0\hat{\eta}_T) = 0 \tag{4.2}$$

[10] Kleibergen (2005) also maintains the same kind of assumptions in the context of weak identification.
[11] 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 $\left[T^{1/2}\overline{\phi}_T(R^0\hat{\eta}_T)\right]$ in (4.2) by its first-order Taylor expansion:

$$T^{1/2}\overline{\phi}_T(R^0\eta^0) + T^{1/2}\frac{\partial\phi_T(R^0\eta_T^*)}{\partial\theta'}R^0\left[\hat{\eta}_T - \eta^0\right]$$

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

$$T^{1/2}\overline{\phi}_{iT}(R^0\eta^0) + \sum_{j=1}^{l}\frac{T^{1/2}}{\lambda_{jT}}\frac{\partial\phi_{iT}(R^0\eta_T^*)}{\partial\theta'}R_j\lambda_{jT}\left[\hat{\eta}_{jT} - \eta_j^0\right]$$

Since $\lambda_{iT}[\hat{\eta}_{iT} - \eta_i^0] = \mathcal{O}_P(1)$ ($i = 1, \cdots, l$) (see equation (3.7)), we need to ensure the following to avoid the contamination of the distribution of fast converging parameters by the slow ones:

$$\frac{T^{1/2}}{\lambda_{jT}}\frac{\partial\overline{\phi}_{iT}(R^0\eta_T^*)}{\partial\theta'}R_j \xrightarrow{P} 0 \quad \text{when } T \to \infty \quad \text{for all} \quad j > i \tag{4.3}$$

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

Assumption 6 *(Orthogonality condition)*
(i) If θ_T^* is such that $\|\theta_T^* - \theta^0\| = \mathcal{O}(1/\lambda_{lT})$ then for $i = 1, \cdots, l$

$$\frac{T^{1/2}}{\lambda_{jT}}\frac{\partial\overline{\phi}_{iT}(\theta_T^*)}{\partial\theta'}R_j \xrightarrow{P} 0 \quad \textit{when} \quad T \to \infty \quad \textit{for all} \quad j > i$$

(ii) For all $i = 1, \cdots, l$ and each component $k = 1, \cdots, k_i$: $\dfrac{T^{1/2}}{\lambda_{iT}}\left[\dfrac{\partial^2\overline{\phi}_{iT,k}(\theta)}{\partial\theta\partial\theta'}\right]$ 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 $\left[\partial\rho^*(\eta^0)/\partial\eta'\right] = \left[\partial\rho(\theta^0)/\partial\eta'\right]R^0$. Actually:

$$\texttt{Plim}\left[\frac{T^{1/2}}{\lambda_{jT}}\frac{\partial\overline{\phi}_{iT}(\theta_T^*)}{\partial\theta'}R_j\right] = \texttt{Plim}\left[\frac{\lambda_{iT}}{\lambda_{jT}}\left(\frac{T^{1/2}}{\lambda_{iT}}\frac{\partial\overline{\phi}_{iT}(\theta_T^*)}{\partial\theta'} - \frac{\partial\rho_i(\theta^0)}{\partial\theta}\right)R_j\right] \tag{4.4}$$

The difficulty is that, due to θ_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 $[\partial\overline{\phi}_i(\theta_T^*)/\partial\theta']$ in the neighborhood of θ^0. Hence, the required orthogonality condition follows only if we know that:

$$\frac{\lambda_{iT}}{\lambda_{jT}} \times \frac{1}{\lambda_{lT}} \to 0 \ \ \forall\ j > i, \qquad \text{or} \quad \lambda_{iT} = o(\lambda_{lT}^2) \tag{4.5}$$

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

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

(*ii*) *For all* $i = 1, \cdots, l$ *and each component* $k = 1, \cdots, k_i$, $\dfrac{T^{1/2}}{\lambda_{iT}} \left[\dfrac{\partial^2 \overline{\phi}_{iT,k}(\theta)}{\partial\theta\partial\theta'} \right]$ *converges in probability uniformly on* $\theta \in \Theta$ *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 bandwidths, or trimming parameters that one can consider in the examples of section 2. For instance, in the case of one dimensional kernel smoothing, $\lambda_{2T} = \sqrt{Th_T}$ fulfills the required condition (with respect to $\lambda_{1T} = \sqrt{T}$) only if $h_T\sqrt{T} \overset{T}{\to} \infty$. Interestingly enough, the case of first-order under-identification (Sargan (1983), Dovonon and Renault (2009)) is the limit case where the slow rate (namely $T^{1/4}$) is just sufficiently slow to violate the condition[12]. Technically, maintaining Assumptions 5 and 6 (or 6*) warrants the following well-suited block-diagonality property for the limit Jacobian matrix:

Lemma 4.1 *Under Assumptions 1 to 6 (or 6***), if* θ_T^* *is such that* $\|\theta_T^* - \theta^0\| = \mathcal{O}_P(1/\lambda_{lT})$,

$$T^{1/2}\frac{\partial\overline{\phi}_T(\theta_T^*)}{\partial\theta'}R^0\tilde{\Lambda}_T^{-1} \overset{P}{\to} J^0 \quad \textit{when } T \to \infty$$

[12]In the context of weak instruments, Antoine and Renault (2009) define nearly-strong instruments as instruments featuring some degree of weakness, but still conformable to Assumption 6*.

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

$$\tilde{\Lambda}_T = \begin{pmatrix} \lambda_{1T}I_{s_1} & & & \\ & \lambda_{2T}I_{s_2} & & O \\ & & \ddots & \\ & O & & \lambda_{lT}I_{s_l} \end{pmatrix} \quad with \quad \begin{cases} i) & \sum_{i=1}^{l} s_i = p \\ ii) & \lim_{T\to\infty} \lambda_{iT} = \infty \quad for \;\; i = 1,...,l \\ iii) & \lambda_{i+1,T} = o(\lambda_{i,T}) \quad for \;\; i = 1,...,l-1 \end{cases}$$

Thanks to the aforementioned block-diagonality, we get a standard asymptotic normal distribution for the new parameters $\eta = [R^0]^{-1}\theta$, albeit with non standard rates of convergence:

Theorem 4.2 *(Asymptotic Normality)*
Under Assumptions 1 to 6 (or 6), the minimum distance estimator $\hat{\theta}_T$ (2.5) is such that:*

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

where S^0 denotes the covariance matrix of the asymptotic gaussian distribution of $\sqrt{T}\,\overline{\phi}_T(\theta^0)$.

This result is similar to the partial identification results discussed in Phillips (1989). In addition, it also has strong similarities with Hansen (1982) classical result about the asymptotic distribution of GMM. At first sight, the matrix J^0 may almost be interpreted as $\left[\partial\rho(\theta^0)/\partial\theta' R^0\right] = \left[\partial\rho^*(\eta^0)/\partial\eta'\right]$ where $\rho^*(\eta) = \rho(R^0\theta)$. However this simple interpretation is not fully correct. While $\left[\partial\rho^*(\eta^0)/\partial\eta'\right]$ 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 4.1, in such a way that J^0 is block-diagonal. However, seeing J^0 as $\left[\partial\rho^*(\eta^0)/\partial\eta'\right]$ would allow us to interpret the asymptotic variance in Theorem 4.2 as the standard asymptotic variance of a minimum distance estimator computed from the (unfeasible) minimization problem (4.1). In particular, the cancelation of upper-diagonal blocks does not invalidate the standard argument to get the optimal weighting matrix:

Theorem 4.3 *Let S^0 denote the covariance matrix of the asymptotic gaussian distribution of $\sqrt{T}\,\overline{\phi}_T(\theta^0)$. Under Assumptions 1 to 6 (or 6*), the asymptotic variance displayed in Theorem 4.2 is minimal when the minimum distance estimator $\hat{\theta}_T$ is defined by (2.5) while using a consistent estimator of $[S^0]^{-1}$ as the weighting matrix Ω_T. Then,*

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

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 θ. 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 η.

4.2 Feasible asymptotic distributions

In general, the focus of interest is not the vector η (new parameters) but the vector θ (structural parameters). As far as inference about θ is concerned, several practical implications of Theorem 4.3 are worth mentioning. From Lemma 4.1, a consistent estimator of the asymptotic covariance matrix $\left[J^{0\prime}[S^0]^{-1}J^0\right]^{-1}$ is:

$$\begin{aligned} & T^{-1}\left[\tilde{\Lambda}_T^{-1}R^{0\prime}\frac{\partial\overline{\phi}_T'(\hat{\theta}_T)}{\partial\theta}S_T^{-1}\frac{\partial\overline{\phi}_T(\hat{\theta}_T)}{\partial\theta'}R^0\tilde{\Lambda}_T^{-1}\right]^{-1} \qquad (4.6)\\ = \; & T^{-1}\tilde{\Lambda}_T[R^0]^{-1}\left[\frac{\partial\overline{\phi}_T'(\hat{\theta}_T)}{\partial\theta}S_T^{-1}\frac{\partial\overline{\phi}_T(\hat{\theta}_T)}{\partial\theta'}\right]^{-1}[R^{0\prime}]^{-1}\tilde{\Lambda}_T \end{aligned}$$

Since from Theorem 4.3, for large T, $\left[\tilde{\Lambda}_T[R^0]^{-1}(\hat{\theta}_T-\theta^0)\right]$ behaves like a gaussian random variable with mean zero and variance (4.6), one might be tempted to deduce that $\left[\sqrt{T}(\hat{\theta}_T-\theta^0)\right]$ behaves like a gaussian with mean zero and variance

$$\left[\frac{\partial\overline{\phi}_T'(\hat{\theta}_T)}{\partial\theta}S_T^{-1}\frac{\partial\overline{\phi}_T(\hat{\theta}_T)}{\partial\theta'}\right]^{-1} \qquad (4.7)$$

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 is theoretically misleading for several reasons. In general, all components of $\hat{\theta}_T$ converge slowly towards θ^0 and thus $\left[\sqrt{T}(\hat{\theta}_T-\theta^0)\right]$ has no limit distribution. When we say that it is approximately a gaussian with variance (4.7), one must realize that since

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

we actually have

$$\frac{\partial\overline{\phi}_{iT}(\hat{\theta}_T)}{\partial\theta'} \xrightarrow{P} 0 \quad \text{for } i>1$$

In other words, considering the asymptotic variance (4.7) is akin to considering the inverse of an asymptotically singular matrix: (4.7) is not an estimator of the standard population matrix

$$\left[\frac{\partial \rho'(\theta_0)}{\partial \theta}\left[S^0\right]^{-1}\frac{\partial \rho(\theta^0)}{\partial \theta'}\right]^{-1} \tag{4.8}$$

Typically, beyond the above singularity, the population matrix (4.8) will not display, in general, the right block-diagonality structure. Inference about θ is actually more involved than one may believe at first sight, from the apparent similarity with standard GMM formulas. This will be discussed in section 5. At least, the seemingly standard asymptotic distribution theory allows us to perform an overidentification test as usual:

Theorem 4.4 *(J-test)*
Under Assumptions 1 to 6 (or 6), if Ω_T is a consistent estimator of $\left[S^0\right]^{-1}$, $TQ_T(\hat{\theta}_T)$ is asymptotically distributed as a chi-square with $(K-p)$ degrees of freedom.*

The asymptotic distribution of estimators given in theorems 4.2 and 4.3 is not feasible , since based on the unknown true matrix of change of basis R^0. In section 5, we show that valid asymptotic inference about θ does not require the estimation of R^0. However, we study in this section the consistent estimation of R^0 and corresponding plugging in asymptotics for the parameters of interest[13]. 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. Then, we would get a consistent estimator R-hat of R-0 withThe 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

$$J_{l-1}(\theta^0) = \left[\frac{\partial \rho_1'(\theta^0)}{\partial \theta}\ \frac{\partial \rho_2'(\theta^0)}{\partial \theta}\ \cdots\ \frac{\partial \rho_{l-1}'(\theta^0)}{\partial \theta}\right]$$

By virtue of Assumption 5, we have for $i = 1, \cdots, l$:

$$\frac{\partial \rho_i'(\theta^0)}{\partial \theta} = \texttt{Plim}\left[\frac{T^{1/2}}{\lambda_{iT}}\frac{\partial \overline{\phi}_{iT}'(\hat{\theta}_T)}{\partial \theta}\right]$$

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 $\lambda_{1T}, \lambda_{2T}, \cdots, \lambda_{l-1,T}$. It is the case,

[13]Recall that a maintained assumption is the knowledge of the right dimensions (k_i, s_i), $i = 1, 2, \cdots, l$.

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 λ_{lT} is not required. This is important, since it solves, at least, all the examples with only two rates of convergence ($l = 2$) with 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:

$$\left\|\hat{R} - R^0\right\| = \mathcal{O}_P\left(\frac{1}{\lambda_{lT}}\right)$$

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

Theorem 4.5 *(Feasible asymptotic normality)*
Under Assumptions 1 to 6, if* $\left\|\hat{R} - R^0\right\| = \mathcal{O}_P\left(1/\lambda_{lT}\right)$
then the asymptotic distribution of $\left[\tilde{\Lambda}_T \hat{R}^{-1}\left(\hat{\theta}_T - \theta^0\right)\right]$ *coincides with the gaussian asymptotic distribution of* $\left[\tilde{\Lambda}_T \left[R^0\right]^{-1}\left(\hat{\theta}_T - \theta^0\right)\right]$ *as characterized in Theorems 4.2 and 4.3.*

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

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

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 λ_{lT}^2 at worst, which is still faster than λ_{1T} by Assumption 6*. To summarize, Theorem 4.5 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 $\left[\hat{R}^{-1}\hat{\theta}_T - \hat{R}^{-1}\theta^0\right]$ instead of $\left[\hat{R}^{-1}\hat{\theta}_T - \left[R^0\right]^{-1}\theta^0\right]$. It is, by definition, impossible to take advantage of the fast rates of convergence in the estimation of $\left[R^0\right]^{-1}\theta^0$, 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 components of θ 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 (2009) develop such an example for option pricing with a prior partition between the different 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 θ. Section 5 below shows 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. Moreover, as stressed above, the price to pay for this consistent estimation is the knowledge (or consistent estimation) of the various rates of convergence which, in turn, determines the rates of relevant sequences of local alternatives.

5 Inference on functions of θ

This section is dedicated to testing the null hypothesis, $H_0 : g(\theta) = 0$, where the function $g(.)$, from Θ to $\mathbb{R}^q$, is continuously differentiable on the interior of Θ. A couple of preliminary remarks are in order. First, working under the null may lead to revisit significantly the reparametrization defined in section 4. Typically, with additional information, the linear combinations of θ estimated respectively at the different rates of convergence may be defined differently. To circumvent this difficulty, we focus on Wald-type test[14]. 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: the coexistence of different rates of convergence may introduce (asymptotically) some multicolinearity between the q estimated constraints, and the delta-theorem may be invalidated as illustrated in the example below.

Example 7 *Consider two groups of moment conditions (associated respectively with the rates of convergence λ_{1T} and λ_{2T}) and the null hypothesis $H_0 : g(\theta) = 0$ with $g(\theta) = [g_j(\theta)]_{1\leq j\leq q}$ where none of the q vectors $\left[\partial g_j(\theta^0)/\partial\theta\right]$, $j = 1, \cdots, q$ belongs to $\mathtt{col}[\partial\rho_1'(\theta^0)/\partial\theta]$. If we extent the standard argument for Wald test, we have: under the null, $\left[\lambda_{2T}g(\hat{\theta}_T)\right]$ is asymptotically*

[14]Caner (2008) derives the standard asymptotic equivalence results for the trinity of tests. However, he only considers testing when all parameters converge at the same nearly-weak rate.

like $\left[(\partial g(\theta^0)/\partial\theta')\lambda_{2T}(\hat{\theta}_T-\theta^0)\right]$*, that is, for large* T*,* $\left[\lambda_{2T}g(\hat{\theta}_T)\right]$ *behaves like a gaussian*

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

Consider now that, for some nonzero vector α*,*

$$\frac{\partial g'(\theta^0)}{\partial\theta}\alpha=\sum_{j=1}^{q}\alpha_j\frac{\partial g_j(\theta^0)}{\partial\theta}\quad \textit{belongs to}\quad \texttt{col}[\partial\rho'_1(\theta^0)/\partial\theta]$$

Then, under the null, $\left[\lambda_{1T}\alpha' g(\hat{\theta}_T)\right]$ *is asymptotically gaussian and thus*

$$\lambda_{2T}\alpha' g(\hat{\theta}_T)=\frac{\lambda_{2T}}{\lambda_{1T}}\lambda_{1T}\alpha' g(\hat{\theta}_T)\xrightarrow{P}0$$

In other words, even if the q *constraints are locally linearly independent (or a full rank assumption is maintained),* $\left[\lambda_{2T}g(\hat{\theta}_T)\right]$ *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.*

In spite of the aforementioned singularity problem, standard Wald-type inference is valid without additional regularity assumption as stated in Theorem 5.1 below: this is related to the well-known fact that the finite sample performance of the Wald test depends on the way the null hypothesis is formulated[15]. Consider a fictitious situation where the range of $\left[\partial\rho'_1(\theta^0)/\partial\theta\right]$ is known. Then, it is always possible to define a (q,q) non-singular 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 (in the simpler case with only two different rates) for some q_1 such that $1\leq q_1\leq q$, we have:

- for $j=1,\cdots,q_1$: $\left[\partial h_j(\theta^0)/\partial\theta\right]$ belongs to $\texttt{col}\left[\partial\rho'_1(\theta^0)/\partial\theta\right]$
- for $j=q_1+1,\cdots,q$: $\left[\partial h_j(\theta^0)/\partial\theta\right]$ does not belong to $\texttt{col}\left[\partial\rho'_1(\theta^0)/\partial\theta\right]$ and no linear combinations of them do.

Then, the asymptotic singularity of example 7 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

[15] In some respect, our approach complements the higher-order expansions of Phillips and Park (1988).

ones $g(\theta) = 0$ (since the matrix H depends on θ). It turns out that, for all practical purposes, by treating H as known, the Wald-type test statistics written with $h(.)$ or $g(.)$ are numerically equal. The formal proof of Theorem 5.1 is provided in the Appendix.

Theorem 5.1 *(Wald test)*
Under the Assumptions 1 to 6 (or 6) and if $g(.)$ is twice continuously differentiable, the Wald test statistic ζ_T^W,*

$$\zeta_T^W = Tg'(\hat{\theta}_T) \left\{ \frac{\partial g(\hat{\theta}_T)}{\partial \theta'} \left[\frac{\partial \overline{\phi}_T'(\hat{\theta}_T)}{\partial \theta} S_T^{-1} \frac{\partial \overline{\phi}_T(\hat{\theta}_T)}{\partial \theta'} \right]^{-1} \frac{\partial g'(\hat{\theta}_T)}{\partial \theta} \right\}^{-1} g(\hat{\theta}_T) \tag{5.1}$$

for testing $H_0 : g(\theta) = 0$ is asymptotically chi-square with q degrees of freedom under the null.

Regarding the size of the test, considering 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 we consider only $q = 1$ restriction to test (for simplicity), and two rates of convergence, we get:

Theorem 5.2 *(Local alternatives)*
Under Assumptions 1 to 6 (or 6), the Wald test of $H_0 : g(\theta) = 0$ (with $g(.)$ one dimensional continuously differentiable) is consistent under the sequence of local alternatives $H_{1T} : g(\theta) = 1/\delta_T$ if and only if either*

$$\frac{\partial g(\theta^0)}{\partial \theta} \in \texttt{col} \left[\frac{\partial \rho_1'(\theta^0)}{\partial \theta} \right] \quad \textit{and} \quad \delta_T = o(\lambda_{1T})$$

or

$$\frac{\partial g(\theta^0)}{\partial \theta} \notin \texttt{col} \left[\frac{\partial \rho_1'(\theta^0)}{\partial \theta} \right] \quad \textit{and} \quad \delta_T = o(\lambda_{2T})$$

The proof of Theorem 5.2 is rather straightforward. A nonlinear function $g(.)$ of θ, interpreted as $\left[g(\theta^0) + \frac{\partial g(\theta^0)}{\partial \theta'}(\theta - \theta^0) \right]$, is identified at the fast rate λ_{1T} if and only if

$$\frac{\partial g(\theta^0)}{\partial \theta'} \in \texttt{col} \left[\frac{\partial \rho_1'(\theta^0)}{\partial \theta} \right]$$

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

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

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 4.1 (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 $[\partial \overline{\phi}_T(.)/\partial \theta']$. 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 resort to the local alternatives approach considered in Theorem 5.2.

In a somewhat related framework, Lee (2005) puts forward some high-level assumptions (see his Assumptions (R) and (G)) to deal with the aforementioned asymptotic singularity.

Lee's (2005) Assumption:
There exists a sequence of (q,q) invertible matrices D_T such that for any $\theta \in \Theta$

$$\texttt{Plim} \left[D_T \frac{\partial g(\theta^0)}{\partial \theta'} R^0 [\tilde{\Lambda}_T]^{-1} \right] = B_0$$

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

Lee's (2005) Assumption clearly implies the standard rank condition: $\texttt{rk}\,[\partial g(\theta)/\partial \theta'] = q$, for all θ in the interior of Θ, or at least in a neighborhood of θ^0. However, the converse is not true as it can be shown from the counterexample above[16]. And, this is all we need 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 $\left[B^0 \tilde{\Lambda}_T [R^0]^{-1} (\hat{\theta}_T - \theta^0) \right]$. From Theorem 4.2, we know that $\left[\tilde{\Lambda}_T [R^0]^{-1} (\hat{\theta}_T - \theta^0) \right]$ behaves asymptotically like a gaussian. 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.

[16] By contrast, in the case with only $q = 1$ constraint, Lee's Assumption is trivially fulfilled.

6 A Monte Carlo Illustration

In this section, we illustrate our inference procedure with the Monte Carlo study of an application of the kernel smoothing example announced in section 2.1. This application closely follows the option pricing example put forward by Gagliardini, Gouriéroux and Renault (2009, hereafter GGR) for their extended method of moments. The data generating process is a discrete time version of the Stochastic Volatility (SV) option pricing model of Heston (1993).

6.1 Framework

We consider a market with a risk-free asset with a zero risk-free rate and a risky asset with geometric return $r_t = \log(p_t/p_{t-1})$ at time t such that: $r_t = \gamma\sigma_t^2 + \sigma_t\epsilon_t$ where ϵ_t is a standard Gaussian white noise, σ_t denotes the volatility process and γ measures the magnitude of the risk premium in the expected return. The volatility process is independent of the standardized innovation process ϵ_t. Moreover σ_t^2 follows an autoregressive gamma (ARG) process, a discretized version of the square-root process: see Gouriéroux and Jasiak (2006). Its transition distribution is characterized by the following conditional Laplace transform:

$$E\left[\exp(-u\sigma_{t+1}^2)|\sigma_t^2\right] = \exp\left[-a(u)\sigma_t^2 - b(u)\right] \quad \text{with } a(u) = \frac{\rho u}{(1+cu)}\ ,\ b(u) = \delta\log(1+cu)$$

The positive parameter ρ is the first-order autocorrelation of the variance process, the non-negative parameter δ describes its (conditional) over-/under-dispersion, and the positive parameter c is a scale parameter. In this model, the Markov process of observed state variables is $Y_t = (r_t, \sigma_t^2)$. In GGR's application, σ_t^2 is the daily realized volatility computed from 30-minute S&P returns. As in Heston (1993), we specify an exponential affine stochastic discount factor (SDF):

$$M_{t,t+1}(\theta) = \exp\left[\theta_1 - \theta_2\sigma_{t+1}^2 - \theta_3\sigma_t^2 - \theta_4 r_{t+1}\right]$$

where $\theta = [\theta_1, \theta_2, \theta_3, \theta_4]'$ is the vector of structural parameters. The two restrictions implied by no-arbitrage opportunity for pricing of the risk-free rate and the underlying asset are given by: $E[M_{t,t+1}(\theta)|Y_t] = 1$ and $E[M_{t,t+1}(\theta)\exp(r_{t+1})|Y_t] = 1$. It can easily be checked that the above SDF is compatible with these restrictions if and only if:

$$\begin{aligned}\theta_1 &= \theta_1(\theta_2) = -\delta \log\left[1 + c(\theta_2 + \gamma^2/2 - 1/8)\right] \\ \theta_3 &= \theta_3(\theta_2) = -\rho \frac{\theta_2 + \gamma^2/2 - 1/8}{1 + c(\theta_2 + \gamma^2/2 - 1/8)} \\ \theta_4 &= \gamma + 1/2\end{aligned}$$

Therefore, no-arbitrage restrictions are written as a couple of conditional moment restrictions:

$$E\left[G_1(Y_{t+1}, \theta)|Y_t\right] = 0 \text{ with } G_1(Y_{t+1}, \theta) = \begin{bmatrix} M_{t,t+1}(\theta) - 1 \\ M_{t,t+1}(\theta)\exp(r_{t+1}) - 1 \end{bmatrix}$$

These restrictions are fulfilled not only at the true value θ^0, but also all over the curve $\theta = h(\theta_2) = [\theta_1(\theta_2),\ \theta_2,\ \theta_3(\theta_2),\ \gamma + 1/2]'$. Thus,

$$\frac{\partial E\left[G_1(Y_{t+1}, h(\theta_2))|Y_t\right]}{\partial \theta_2} = E\left[\frac{\partial G_1(Y_{t+1}, h(\theta_2))}{\partial \theta'}\frac{\partial h(\theta_2)}{\partial \theta_2}|Y_t\right] = 0 \ \ \forall\ \theta_2 \in \mathbb{R}$$

However, it can easily be checked that the rank of the matrix $E[\partial G_1(Y_{t+1}, \theta^0)/\partial\theta'|Y_t]$ is not smaller than 3. We have then proved that the dimension of the subspace where strong identification is lacking is spanned by the vector:

$$R_2 = \frac{\partial h(\theta_2^0)}{\partial \theta_2} = \left[-\frac{db(\lambda^0)}{d\lambda},\ 1,\ -\frac{da(\lambda^0)}{d\lambda},\ 0\right]' \text{ with } \lambda^0 = \theta_2^0 + \gamma^2/2 - 1/8 \qquad (6.1)$$

In financial terms, it means that observing prices of the risk-free asset and the underlying asset is not sufficient to elicit a unique risk-neutral probability measure. The change of probability measure, to be defined from $M_{t,t+1}(\theta^0)$, is not unique because the true unknown value θ^0 is not identified. For a given θ_2^0, $\theta^0 = h(\theta_2^0)$ and one can check that the risk-neutral probability measure Q is characterized again by ARG volatility dynamics with conditional log-normality of stock returns and risk neutral parameters given by:

$$\gamma^* = -1/2\ ,\ \delta^* = \delta\ ,\ \rho^* = \frac{\rho}{(1 + c\lambda^0)^2}\ ,\ c^* = \frac{c}{1 + c\lambda^0}$$

Interestingly enough, the above direction R_2 of lack of identification is tightly related to risk-neutral prediction errors on volatility. We deduce from the Laplace transform of the risk-neutral ARG process that:

$$\frac{\partial M_{t,t+1}(h(\theta_2^0))}{\partial \theta'} R_2 = M_{t,t+1}(\theta^0)\left(\sigma_{t+1}^2 - E^Q\left[\sigma_{t+1}^2|Y_t\right]\right)$$

It implies that for any risky asset with payoff ξ_{t+1} at time $t+1$, we have:

$$E\left[\frac{\partial M_{t,t+1}(h(\theta_2))}{\partial \theta'}\frac{\partial h(\theta_2)}{\partial \theta_2}\xi_{t+1}|Y_t\right] = Cov^Q\left[\sigma^2_{t+1}, E^Q(\xi_{t+1}|Y_t,\sigma^2_{t+1})|Y_t\right]$$

This result shows what kind of asset pricing condition may alleviate the aforementioned lack of identification. We need to observe the price at time t of an asset with payoff ξ_{t+1} such that:

$$Cov^Q\left[\sigma^2_{t+1}, E^Q(\xi_{t+1}|Y_t,\sigma^2_{t+1})|Y_t\right] \neq 0$$

The price of a derivative written on the primitive return r_{t+1} is then well-suited. Let us consider for instance the payoff of a European call option with unit maturity at time t:

$$\xi_{t+1} = [\exp(r_{t+1}) - k]^+ \quad \text{with} \quad [u]^+ \equiv \max(u, 0)\ \forall u \in \mathbb{R}$$

Taking advantage of the risk-neutral conditional log-normality of r_{t+1} given (Y_t, σ^2_{t+1}), we have

$$E^Q(\xi_{t+1}|Y_t,\sigma^2_{t+1})] = BS(k,\sigma^2_{t+1})$$

where $BS(k,\sigma^2)$ denotes the Black-Scholes price of a European call written on r_{t+1} with relative strike k, time-to-maturity 1 and constant volatility σ^2. As expected, adding the pricing equation of any European call solves the aforementioned lack of identification, since:

$$E\left[\frac{\partial M_{t,t+1}(h(\theta_2))}{\partial \theta'}\frac{\partial h(\theta_2)}{\partial \theta_2}\left[\exp(r_{t+1}) - k\right]^+|Y_t\right] = \texttt{Cov}^Q\left[\sigma^2_{t+1}, BS(k,\sigma^2_{t+1})|Y_t\right] > 0$$

The covariance is strictly positive since the BS formula is strictly increasing in volatility.

As pointed out by GGR, the fact that option prices are necessary to identify the volatility risk premium parameter θ_2 paves the way for the simultaneous consideration of two different rates of convergence in the line of Example 1. On one hand, the observation of time series $Y_t = (r_t, \sigma^2_t)$ for $t = 1, \cdots, T$ of the state variable allows us to transform the conditional moment restrictions $E\left[G_1(Y_{t+1},\theta)|Y_t\right] = 0$ into unconditional ones for any choice of (strong) instruments. These unconditional moment restrictions admit root-T consistent sample counterparts. Moreover, the above spanning argument shows that three out of four dimensions of the structural parameter space (unknown parameter θ) can be identified from these restrictions: for any given value of θ_2, we should be able to estimate all parameters with a root-T rate of convergence. On the other hand, option prices needed to identify θ_2 are rarely available as time series. We rather observe an option price at a given moneyness, say k, and maturity

1 at one date, say τ. Thus, we can only use the information from an option price observed at time τ, coming as local conditional restriction for a given observed price $c_\tau(k)$:

$$E\left[M_{\tau,\tau+1}(\theta)\left[\exp(r_{\tau+1})-k\right]^+ - c_\tau(k)|Y_\tau\right]=0$$

The only sample counterpart available for such a local moment restriction is a kernel estimator consistent at non-parametric rate $\sqrt{Th_T}$ after choosing the bandwidth parameter h_T.

6.2 Monte Carlo results

We assume that we have T observations of the state variable $Y_t=(r_t,\sigma_t^2)$ for $t=\tau-T+1,\cdots,\tau$ and 3 option prices at date τ with respective strikes $k_1=0.95$, $k_2=0.97$, and $k_3=0.99$. The state variables are generated by the DGP described in the previous section where the parameters values are set as in Gagliardini, Gouriéroux and Renault[17] (2009): $\gamma=0.360$, $\rho=0.960$, $\delta=1.047$, $c=3.65.10^{-6}$, $\theta=[0.456.10^{-6},\ -0.059,\ 0.114,\ 0.860]'$. The option prices are computed by simulation as explained in Appendix B. We use two instruments, the constant and the (normalized) lagged asset return, to transform the two uniform conditional moment restrictions into four unconditional ones. As a result, we have four standard moment restrictions based on a sample mean $\overline{\phi}_{1T}(\theta)$ and three local ones based on $\overline{\phi}_{2T}(\theta)$ with:

$$\begin{aligned}
\phi_{1t}(\theta) &= \begin{pmatrix} M_{t-1,t}(\theta)-1 \\ M_{t-1,t}(\theta)e^{r_t}-1 \end{pmatrix} \otimes \begin{pmatrix} 1 \\ r_{t-1} \end{pmatrix} \\
\phi_{2t}^{(j)}(\theta) &= \left[M_{t-1,t}[e^{r_t}-k_j]^+ - c_\tau(k_j)\right]\omega_{t,T} \quad \text{for } j=1,\cdots,3 \\
\text{and } \omega_{t,T} &= \sqrt{h_T}K\left(\frac{Y_t-Y_\tau}{h_T}\right)\left[\frac{1}{T}\sum_{s=\tau-T+1}^{\tau}K\left(\frac{Y_s-Y_\tau}{h_T}\right)\right]^{-1}
\end{aligned}$$

$K(.)$ denotes the Gaussian product kernel of dimension 2, and $h_T=c_h/T^{(1/6+\epsilon)}$ is the bandwidth chosen to get an unbiased estimator with ϵ some (small) positive real number and c_h some positive real number[18].

We compare two inference procedures: our proposed estimator $\hat{\theta}_T$ (2.5), and the classic estimator $\tilde{\theta}_T^{(N)}$ (2.3) defined through the naive rescaling (2.4). It is important to realize that the

[17] The ARG parameters are chosen to match the stationary mean, variance, and first-order autocorrelation of the realized volatility of the S&P500 index from June, 1, 2001 to May, 31, 2005.

[18] ϵ is set as "eps", the spacing of floating point numbers in Matlab: eps$\sim 2.22.10^{-16}$; several values of c_h have been used and do not affect the results much.

classic estimator is feasible only because in the example we consider here the slower (nonparametric) rate of convergence is known after choosing the bandwidth.

To identify the standard directions in the parameter space, we now define the matrix of the change of basis R. In the previous section, we have determined that the space of nonstandard (ie slow) direction is spanned by the vector R_2 (6.1). To get R, R_2 is normalized and is completed by the orthogonal $(4,3)$-matrix R_1 that is orthogonal to R_2. For this Monte Carlo exercise, we choose to consider R_2 as known and to compute R accordingly. Introducing estimation error into R would not change the results significantly. As a result, R is not data-dependent in this example and the new parameter η is defined as: $\theta = R\eta$. R is given in Appendix B. We conclude that $\theta_1 \sim \eta_1$, $(\theta_2 + \theta_3) \propto \eta_2$, $(\theta_2 - \theta_3) \propto \eta_4$ and $\theta_4 = \eta_3$, keeping in mind that the first three components of η are estimated at the standard rate $\sqrt{T}$ while the last one is estimated at the nonparametric rate $\sqrt{Th_T}$.

According to our asymptotic results, the asymptotic variance of the estimator of the new parameters η_1, η_2, η_3 should decrease faster with the sample size than the one of η_4; as for the structural parameter, we can see from the matrix of change of basis that the asymptotic variance of the estimator of parameters θ_1, θ_2 and θ_3 should decrease at the same rate. Figures 1 and 2 plot the evolution of ratios of Monte Carlo variances of components of η and θ respectively with the sample size: left panels for our estimator and right ones for the naive estimator. Ratios related to our estimator behave as expected: ratios of variances of components of η, $\text{Var}(\hat{\eta}_{4T})/\text{Var}(\hat{\eta}_{iT})$ $i = 1, 2, 3$, increase with T, which means that $\hat{\eta}_{iT}$ converges at a faster rate than $\hat{\eta}_{4T}$; ratios of variances of components of θ are relatively constant with respect to the sample size, which means that $\hat{\theta}_{1T}$, $\hat{\theta}_{2T}$ and $\hat{\theta}_{3T}$ converge at the same rate. No similar conclusions can be drawn from the ratios associated with the naive estimators: in other words, faster directions cannot be captured.

7 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, 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. We stress in particular that the ability of GMM to improve the estimators of parameters of interest by taking advantage of additional estimating equations as control variables is not destroyed by the occurrence of heterogeneous rates of convergence. We show that even estimating equations with slow rates of convergence may not be redundant with respect to GMM estimation of fast identified directions.

Finally, we demonstrated the validity of usual inference procedures with standard formulas, like the overidentification test and Wald test. In addition, both estimation and testing work without requiring the knowledge of the various rates. However, 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 (2010).

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 (2009), as well as in Gagliardini, Gouriéroux and Renault (2009).

References

[1] Y. Aït-Sahalia and J. Jacod, *Fisher's Information for Discretely Sampled Lévy Processes*, Econometrica **76** (2008), 727–761.

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

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

[4] B. Antoine and E. Renault, *Efficient GMM with Nearly-weak Instruments*, Econometric Journal **12** (2009), 135–171.

[5] F.M. Bandi and P.C.B. Phillips, *A simple approach to the parametric estimation of potentially nonstationary diffusions*, Journal of Econometrics **137** (2007), 354–395.

[6] P. Bertail, D.N. Politis, and J.P. Romano, *On Subsampling Estimators with unknown rates of convergence*, Journal of the American Statistical Association **94** (1999), 569–579.

[7] M. Caner, *Near-singular Design in GMM and Generalized Empirical Likelihood Estimators*, Journal of Econometrics **144** (2008), 511–523.

[8] ———, *Testing, Estimation in GMM and CUE with Nearly-Weak Instruments*, Econometrics Reviews **29** (2010), 330–363.

[9] P. Dovonon and E. Renault, *GMM Overidentification Test with First-Order Unidentification*, Working Paper, UNC-CH (2009).

[10] P. Gagliardini, C. Gouriéroux, and E. Renault, *Efficient Derivative Pricing by Extended Methods of Moments*, Working Paper, UNC-CH (2009).

[11] C. Gouriéroux and J. Jasiak, *Autoregressive Gamma Processes*, Journal of Forecasting **25** (2006), 129–152.

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

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

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

[15] S. Heston, *A Closed Form Solution for Options with Stochastic Volatility with Applications to Bond and Currency Options*, Review of Financial Studies **6** (1993), 327–343.

[16] B.M. Hill, *A simple general approach to inference about tails of a distribution*, Annals of Statistics **3** (1975), 1163–1174.

[17] J. Hill and E. Renault, *Generalized Method of Moments with Tail Trimming*, Working Paper, UNC-CH (2009).

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

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

[20] Y. Kitamura, *Hypotheses Testing in models with possibly Nonstationary Processes*, Working Paper (1996).

[21] Y. Kitamura and P.C.B. Phillips, *Fully Modified IV, GIVE and GMM estimation with possibly non-stationary regressors and instruments*, Journal of Econometrics **80** (1997), 85–123.

[22] F. Kleibergen, *Testing Parameters in GMM without assuming that they are identified*, Econometrica **73** (2005), 1103–1123.

[23] Y. Kotlyarova and V. Zinde-Walsh, *Non and Semi-parametric Estimation in Models with unknown Smoothness*, Economic Letters **93** (2006), 369–386.

[24] L. Lee, *Asymptotic Distributions of Quasi-Maximum Likelihood Estimators for Spatial Econometric Models*, Econometrica **72** (2004), 1899–1926.

[25] ———, *Classical Inference with ML and GMM Estimates with Various Rates of Convergence*, Working Paper, Ohio State University (2005).

[26] ———, *Pooling Estimators with Different Rates of Convergence - A minimum χ^2 Approach: with an emphasis on a Social Interaction Model*, Econometric Theory **26** (2010), 260–299.

[27] A. Pakes and D. Pollard, *Simulation and the Asymptotics of Optimization Estimators*, Econometrica **57** (1989), 1027–1058.

[28] P.C.B. Phillips, *Partially Identified Econometric Models*, Econometric Theory **5** (1989), 181–240.

[29] P.C.B. Phillips and J.Y. Park, *On the formulation of Wald tests of nonlinear restrictions*, Econometrica **56** (1988), 1065–1083.

[30] P. Prescott, *Selection of Trimming Proportions for Robust Adaptive Trimmed Means*, Journal of the American Statistical Association **73** (1978), 133–140.

[31] P. Radchenko, *Mixed-Rates Asymptotics*, Annals of Statistics **36** (2008), 287–309.

[32] C.Y. Robert, *Estimation paramétrique d'une copule extrêmale bivariée à l'aide de la méthode des moments*, Working Paper, CREST-Paris (2006).

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

[34] C.A. Sims, J. Stock, and M. Watson, *Inference in Linear Time Series Models with some Unit-roots*, Econometrica **58** (1990), 113–144.

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

[36] S.M. Stigler, *The Asymptotic Distribution of the Trimmed Mean*, Annals of Statistics **1** (1973), 472–477.

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

[38] A.W. van der Vaart, *Asymptotic Statistics*, Cambridge Series in Statistical and Probabilistic Mathematics, Cambridge University Press, 1998.

Appendix A: Proofs of the main results

Proof of Equation (2.1): (*Stronger identification property*)
Let us denote by S_ϵ the set of $\theta \in \Theta$ such that $\|\theta - \theta^0\| \geq \epsilon$. Since it is compact, the identification Assumption 1 with the continuity of $\rho(.)$ implies that the minimum of $\|\rho(\theta)\|$ on this set is $\alpha > 0$. ■

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

Lemma A.1 $\|\rho(\hat{\theta}_T)\| = \mathcal{O}_P\left(1/\underline{\lambda}_T\right)$

Proof of Lemma A.1: From (3.1), 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:

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

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 μ_T be the smallest eigenvalue of Ω_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, $x_T = \|\Lambda_T\rho(\hat{\theta}_T)\|$ solves the inequality:

$$x_T^2 - \frac{2\|\Omega_T\Psi_T(\hat{\theta}_T)\|}{\mu_T}x_T + \frac{d_T}{\mu_T} \leq 0 \quad \text{and thus with} \quad \Delta_T = \frac{\|\Omega_T\Psi_T(\hat{\theta}_T)\|^2}{\mu_T^2} - \frac{d_T}{\mu_T} \quad \text{we have}$$

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

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

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

which amounts to show that:

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

Denote by $\det(M)$ the determinant of any square matrix M. Since $\det(\Omega_T) \xrightarrow{P} \det(\Omega) > 0$, no subsequence of μ_T can converge in probability towards zero and thus we can assume (for T sufficiently large) that μ_T remains lower bounded away from zero with asymptotic probability one. Therefore, we just have to show that: $\|\Omega_T \Psi_T(\hat{\theta}_T)\| = \mathcal{O}_P(1)$ and $d_T = \mathcal{O}_P(1)$.
Denote by $\mathtt{tr}(M)$ the trace of any square matrix M. Since $\mathtt{tr}(\Omega_T) \xrightarrow{P} \mathtt{tr}(\Omega)$ and the sequence $\mathtt{tr}(\Omega_T)$ is upper bounded in probability, so are all the eigenvalues of Ω_T. Therefore the required boundedness in probability just follows from our functional CLT Assumption 2. We then have:

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

The proof of Lemma A.1 is completed. Let us then deduce the weak consistency of $\hat{\theta}_T$ by a contradiction argument. If $\hat{\theta}_T$ was not consistent, there would exist some positive ϵ such that $P\left[\|\hat{\theta}_T - \theta^0\| > \epsilon\right]$ does not converge to zero. Then we can define a subsequence $(\hat{\theta}_{T_n})_{n\in\mathbb{N}}$ such that, for some positive η: $P\left[\|\hat{\theta}_{T_n} - \theta^0\| > \epsilon\right] \geq \eta \quad \text{for} \quad n \in \mathbb{N}$.
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 3.1: *(Rate of convergence)*
From Lemma A.1, $\|\rho(\hat{\theta}_T)\| = \|\rho(\hat{\theta}_T) - \rho(\theta^0)\| = \mathcal{O}_P(1/\underline{\lambda}_T)$ and by application of the mean-value theorem, for some $\tilde{\theta}_T$ between $\hat{\theta}_T$ and θ^0 (component by component), we get:

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

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: $z_T \equiv \left[\partial \rho(\theta^0)/\partial \theta'\right](\hat{\theta}_T - \theta^0)$. Since $\left[\partial \rho(\theta^0)/\partial \theta'\right]$

is full column rank,

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

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

$$\tilde{z}_T = z_T + \left(\frac{\partial \rho(\tilde{\theta}_T)}{\partial \theta'} - \frac{\partial \rho(\theta^0)}{\partial \theta'}\right)(\hat{\theta}_T - \theta^0) \tag{A.1}$$

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

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

We then conclude from the above and equation (A.1) that $\|z_T\| = \mathcal{O}_P(1/\underline{\lambda}_T)$. ■

Proof of Lemma 4.1: To get the results, we have to show the following:

i) (diagonal terms) $\dfrac{T^{1/2}}{\lambda_{iT}}\dfrac{\partial \overline{\phi}_{iT}(\theta_T^*)}{\partial \theta'} \xrightarrow{P} \dfrac{\partial \rho_i(\theta^0)}{\partial \theta'}$ for $i = 1, \cdots, l$

ii) (lower diagonal) $\dfrac{T^{1/2}}{\lambda_{jT}}\dfrac{\partial \overline{\phi}_{iT}(\theta_T^*)}{\partial \theta'} \xrightarrow{P} 0$ for $i = 2, \cdots, l$; with $1 \le j < i$

iii) (upper diagonal) $\dfrac{T^{1/2}}{\lambda_{jT}}\dfrac{\partial \overline{\phi}_{iT}(\theta_T^*)}{\partial \theta'} R_j \xrightarrow{P} 0$ for $i = 1, \cdots, l-1$; with $l \ge j > i$

i) From Assumption 5(ii):

$$T^{1/2}\frac{\partial \overline{\phi}'_{iT}(\theta^0)}{\partial \theta} - \lambda_{iT}\frac{\partial \rho'_i(\theta^0)}{\partial \theta} = \mathcal{O}_P(1) \quad \Rightarrow \quad \frac{T^{1/2}}{\lambda_{iT}}\frac{\partial \overline{\phi}_{iT}(\theta_T^*)}{\partial \theta'} - \frac{\partial \rho_i(\theta^0)}{\partial \theta'} \xrightarrow{P} 0 \quad \text{since} \quad \lambda_{iT} \xrightarrow{T} \infty$$

The mean-value theorem applied to the k-th component of $\left[\partial \overline{\phi}_{iT}/\partial \theta'\right]$, $1 \le k \le k_i$, for $\tilde{\theta}_T$ between θ^0 and θ_T^*:

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

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

ii) It directly follows from the proof of i), since for $i > j$ $\lambda_{iT} = o(\lambda_{jT})$:

$$\frac{T^{1/2}}{\lambda_{iT}}\left[\frac{\partial\overline{\phi}_{iT}(\theta_T^*)}{\partial\theta'}\right] \xrightarrow{P} \frac{\partial\rho_i(\theta^0)}{\partial\theta'} \;\Rightarrow\; \frac{T^{1/2}}{\lambda_{jT}}\frac{\partial\overline{\phi}_{iT}(\theta_T^*)}{\partial\theta'} = \frac{\lambda_{iT}}{\lambda_{jT}} \times \frac{T^{1/2}}{\lambda_{iT}}\frac{\partial\overline{\phi}_{iT}(\theta_T^*)}{\partial\theta'} \xrightarrow{P} 0$$

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

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

Recall $\lambda_{lT}\|(\theta_T^* - \theta^0)\| = \mathcal{O}_P(1)$; $\frac{T^{1/2}}{\lambda_{iT}}\frac{\partial^2\overline{\phi}_{iT,k}(\theta)}{\partial\theta\partial\theta'} \xrightarrow{P} H(\theta)$; and by Assumption 6 (or 6*), $\lambda_{iT}/(\lambda_{jT}\lambda_{lT}) = \lambda_{iT}/\lambda_{lT}^2 \times \lambda_{lT}/\lambda_{jT} \xrightarrow{T} 0$. We now prove that the first element of the RHS converges to 0 in probability. From Assumption 5(ii), we have:

$$\frac{T^{1/2}}{\lambda_{jT}}\left[\frac{\partial\overline{\phi}_{iT,k}(\theta^0)}{\partial\theta'}R_j - \frac{\lambda_{iT}}{T^{1/2}}\frac{\partial\rho_i'(\theta^0)}{\partial\theta'}R_j\right] = \mathcal{O}_P\left(\frac{1}{\lambda_{jT}}\right)$$

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

Proof of Theorem 4.2: *(Asymptotic normality)*
From the optimization problem (2.5), the first-order conditions for $\hat{\theta}_T$ are written as:

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

A mean-value expansion yields to:

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

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

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

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

From Theorem 3.1 and Lemma 4.1: $\texttt{Plim}\tilde{J}_T = J^0$ and $\texttt{Plim}\hat{J}_T = J^0$. Hence,

$$\hat{J}_T'\Omega_T\tilde{J}_T \xrightarrow{P} J^{0\prime}\Omega J^0 \quad \text{non-singular by assumption}$$

Recall now that by Assumption 2(i), $\Psi_T(\theta^0) = \sqrt{T}\overline{\phi}_T(\theta^0)$ converges to a normal distribution with mean 0 and variance S^0. We then get the announced result. ■

Proof of Theorem 4.3:
Directly follows from Theorem 4.2 and the discussion in the main text. ■

Proof of Theorem 4.4: *(J-test)*
A Taylor expansion of order 1 of the moment conditions gives:

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

with $\hat{J}_T = \sqrt{T}\left[\partial\overline{\phi}_T(\hat{\theta}_T)/\partial\theta'\right]R^0\tilde{\Lambda}_T^{-1}$. A Taylor expansion of the FOC gives:

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

with S_T a consistent estimator of the asymptotic covariance matrix of the process $\Psi(\theta)$. Combining the 2 above results leads to:

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

Use the previous result to rewrite the criterion function:

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

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

Proof of Theorem 4.5: *(Feasible asymptotic normality)*
From Theorem 4.2, $\tilde{\Lambda}_T[R^0]^{-1}(\hat{\theta}_T - \theta^0)$ is asymptotically normally distributed. We now show that the above convergence is not altered when R^0 is replaced by $\hat{R}$, some λ_{lT}-consistent estimator. To simplify the calculations, rewrite $[R^0]^{-1}$ and $\hat{R}^{-1}$ as follows:

$$\left[R^0\right]'^{-1} = \left(\begin{array}{cccc} R^{1'} & R^{2'} & \cdots & R^{l'} \end{array} \right) \quad \text{and} \quad \left[\hat{R}\right]'^{-1} = \left(\begin{array}{cccc} \hat{R}^{1'} & \hat{R}^{2'} & \cdots & \hat{R}^{l'} \end{array} \right)$$

Then: $\tilde{\Lambda}_T[R^0]^{-1}(\hat{\theta}_T - \theta^0) = \left[\lambda_{iT} R^i(\hat{\theta}_T - \theta^0)\right]_{1 \leq i \leq l}$ and $\tilde{\Lambda}_T \hat{R}^{-1}(\hat{\theta}_T - \theta^0) = \left[\lambda_{iT} \hat{R}^i(\hat{\theta}_T - \theta^0)\right]_{1 \leq i \leq l}$
We need to show that, for any component i: $\lambda_{iT}\hat{R}^i(\hat{\theta}_T - \theta^0) = \lambda_{iT} R^i(\hat{\theta}_T - \theta^0) + o_P(1)$.
- For $i = l$, we have:

$$\begin{aligned} \lambda_{lT}\hat{R}^l(\hat{\theta}_T - \theta^0) &= \lambda_{lT} R^l(\hat{\theta}_T - \theta^0) + \lambda_{lT}(\hat{R}^l - R^l)(\hat{\theta}_T - \theta^0) \\ &= \lambda_{lT} R^l(\hat{\theta}_T - \theta^0) + (\hat{R}^l - R^l)\lambda_{lT}(\hat{\theta}_T - \theta^0) \end{aligned}$$

From Theorem 3.1, $\lambda_{lT}(\hat{\theta}_T - \theta^0) = \mathcal{O}_P(1)$. 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:

$$\begin{aligned} \lambda_{iT}\hat{R}^i(\hat{\theta}_T - \theta^0) &= \lambda_{iT} R^i(\hat{\theta}_T - \theta^0) + \lambda_{iT}(\hat{R}^i - R^i)(\hat{\theta}_T - \theta^0) \\ &= \underbrace{\lambda_{iT} R^i(\hat{\theta}_T - \theta^0)}_{(1)} + \underbrace{\frac{\lambda_{iT}}{\lambda_{lT}}(\hat{R}^i - R^i)\lambda_{lT}(\hat{\theta}_T - \theta^0)}_{(2)} \end{aligned}$$

From Theorem 4.2, $(1) = \mathcal{O}_P(1)$ and from Theorem 3.1, $\lambda_{lT}(\hat{\theta}_T - \theta^0) = \mathcal{O}_P(1)$. We need to show that (2) is negligible in front of (1) for any i:

$$\begin{aligned} (2) \prec (1) \ \ \forall\, i \quad &\Leftrightarrow \quad \hat{R}^i - R^i = o_P\left(\frac{\lambda_{lT}}{\lambda_{iT}}\right) \ \ \forall\, i \quad \Leftrightarrow \quad \frac{1}{\lambda_{lT}} = o\left(\frac{\lambda_{lT}}{\lambda_{iT}}\right) \ \ \forall\, i \\ &\Leftrightarrow \quad \lambda_{iT} = o(\lambda_{lT}^2) \ \ \forall\, i \\ &\Leftarrow \quad \text{Assumption } 6^*(i) \ \blacksquare \end{aligned}$$

Proof of Theorem 5.1: *(Wald test)*
To simplify the exposition, the proof is performed with only 2 groups of moment conditions associated with 2 rates. There are two steps: in 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 λ_{1T}, while the remaining ones are identified at the slow rate λ_{2T} without any

linear combinations of the latter being identified at the fast rate; in 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:

$$I^0(g) = \left[\text{col}\frac{\partial g'(\theta^0)}{\partial \theta}\right] \cap \left[\text{col}\frac{\partial \rho_1'(\theta^0)}{\partial \theta}\right]$$

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 ϵ_j $(j = 1, \cdots, q)$ such that $\left[(\partial g'(\theta^0)/\partial\theta) \times \epsilon_j\right]_{j=1}^{q_1}$ is a basis of $I^0(g)$ and $\left[(\partial g'(\theta^0)/\partial\theta) \times \epsilon_j\right]_{j=q_1+1}^{q}$ is a basis of

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

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,

$$\texttt{Plim}\left[D_T \frac{\partial h(\theta^0)}{\partial \theta'} R^0 \left[\tilde{\Lambda}_T\right]^{-1}\right] = B^0$$

with D_T a (q, q) invertible diagonal matrix with its first $n^0(g)$ coefficients equal to λ_{1T} and the $(p - n^0(g))$ remaining ones equal to λ_{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.

$$\begin{aligned}
\zeta_T^W(g) &= Tg'(\hat{\theta}_T)\left\{\frac{\partial g(\hat{\theta}_T)}{\partial \theta'}\left[\frac{\partial \overline{\phi}_T'(\hat{\theta}_T)}{\partial \theta} S_T^{-1} \frac{\partial \overline{\phi}_T(\hat{\theta}_T)}{\partial \theta'}\right]^{-1} \frac{\partial g'(\hat{\theta}_T)}{\partial \theta}\right\}^{-1} g(\hat{\theta}_T) \\
&= TH'g'(\hat{\theta}_T)\left\{H\frac{\partial g(\hat{\theta}_T)}{\partial \theta'}\left[\frac{\partial \overline{\phi}_T'(\hat{\theta}_T)}{\partial \theta} S_T^{-1} \frac{\partial \overline{\phi}_T(\hat{\theta}_T)}{\partial \theta'}\right]^{-1} \frac{\partial g'(\hat{\theta}_T)}{\partial \theta} H'\right\}^{-1} Hg(\hat{\theta}_T) \\
&= \zeta_T^W(h)
\end{aligned}$$

Then we show $\zeta_T^W(h)$ 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 θ^0 is replaced by a λ_{2T}-consistent estimator θ_T^*:

$$\texttt{Plim}\left[D_T\frac{\partial h(\theta_T^*)}{\partial \eta'}[\tilde{\Lambda}_T]^{-1}\right] = B^0$$

The proof, very similar to Lemma 4.1, is not reproduced here: $g(.)$ needs to be twice continuously differentiable. The Wald test statistic on $h(.)$ now writes:

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

where $\hat{J}_T \equiv \sqrt{T}\frac{\partial \overline{\phi}_T(\hat{\theta}_T)}{\partial \theta'}R^0\tilde{\Lambda}_T^{-1}$ with $\hat{J}_T \xrightarrow{P} J^0$ and $\hat{J}_T'S_T^{-1}\hat{J}_T \xrightarrow{P} J^{0\prime}\left[S(\theta^0)\right]^{-1}J^0 \equiv \Sigma$.
Now from the mean-value theorem under H_0 we deduce:

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

$$\text{with}\quad \left[D_T\frac{\partial h(\theta_T^*)}{\partial \theta'}R^0\tilde{\Lambda}_T^{-1}\right] \xrightarrow{P} B^0 \quad\text{and}\quad \tilde{\Lambda}_T\left[R^0\right]^{-1}\left(\hat{\theta}_T - \theta^0\right) \xrightarrow{d} \mathcal{N}(0, \Sigma^{-1})$$

Finally we get

$$\xi_T^W(h) = \left[\tilde{\Lambda}_T[R^0]^{-1}(\hat{\theta}_T - \theta^0)\right]' B_0'(B_0\Sigma B_0')^{-1}B_0\left[\tilde{\Lambda}_T[R^0]^{-1}(\hat{\theta}_T - \theta^0)\right] + o_P(1)$$

Following the proof of Theorem 4.4 we get the expected result. ■

Appendix B: Monte Carlo study

Simulation of the observables:

(1) The state variable is $Y_t = (r_t, \sigma_t^2)$ such that
(a) $r_{t+1} = \gamma\sigma_{t+1} + \sigma_{t+1}\epsilon_{t+1}$ with ϵ_t iid gaussian with mean 0 and variance 1.
(b) σ_t^2 follows ARG with parameters ρ, δ, c. Following Gouriéroux and Jasiak (2006), we simulate it in 2 steps through an intermediate latent variable Z_t:

i) $\sigma_t^2|Z_t \sim \text{Gamma}(\delta + Z_t, c)$; ii) $Z_t|\sigma_{t-1}^2 \sim \text{Poisson}(\rho\sigma_{t-1}^2/c)$.

Then, for a given vector of structural parameters θ and observation Y_t, we deduce the SDF: $M_{t,t+1} = \exp[-\theta_1 - \theta_2\sigma_{t+1}^2 - \theta_3\sigma_t^2 - \theta_4 r_{t+1}]$.
The numerical values for the risk premium γ, and ARG parameters ρ, δ, c, θ are provided in the main text.

(2) The standardized price of a European call option with strike k and residual maturity 1 at time τ is:

$$c_\tau(1) = E_\tau[M_{\tau,\tau+1}(\theta)[\exp(r_{\tau+1}) - k]^+] = E_\tau^Q[BS(k, \sigma_{\tau+1}^2)]$$
$$\text{with } BS(k, \sigma_{\tau+1}^2) = \mathcal{N}(d_1) - k\mathcal{N}(d_2);\ d_1 = \frac{\log(1/k) + \sigma_{\tau+1}^2/2}{\sqrt{\sigma_{\tau+1}^2}};\ d_2 = d_1 - \sqrt{\sigma_{\tau+1}^2}$$

where $\mathcal{N}$ is the cumulative standard normal distribution. We simulate the prices of 3 options at time T (final date) with respective strikes k =0.95, 0.97, 0.99. Under the risk-neutral probability, state variable still follow the above model with risk premium parameter $\gamma^* = -1/2$ and ARG parameters:

$$\rho^* = \frac{\rho}{[1 + c(\theta_2 + \gamma^2/2 - 1/8)]^2}\ ,\ \delta^* = \delta\ ,\ c^* = \frac{c}{1 + c(\theta_2 + \gamma^2/2 - 1/8)}$$

$E_\tau^Q[BS(k, \sigma_{\tau+1}^2)]$ is evaluated by Monte Carlo: we simulate 500 paths, calculate the BS price for each path, and the option price corresponds corresponds to the mean over all these paths.

The matrix of the change of basis is given by:

$$R = \begin{pmatrix} 0.999999999992077 & -0.000002871682674 & 0 & -0.000002756817766 \\ 0 & 0.692532141792440 & 0 & 0.721387019972203 \\ -0.000003980779518 & 0.721387019966488 & 0 & -0.692532141786953 \\ 0 & 0 & 1 & 0 \end{pmatrix}$$

Simulations are performed in MATLAB with the solver KNITRO developed by TOMLAB Optimization.

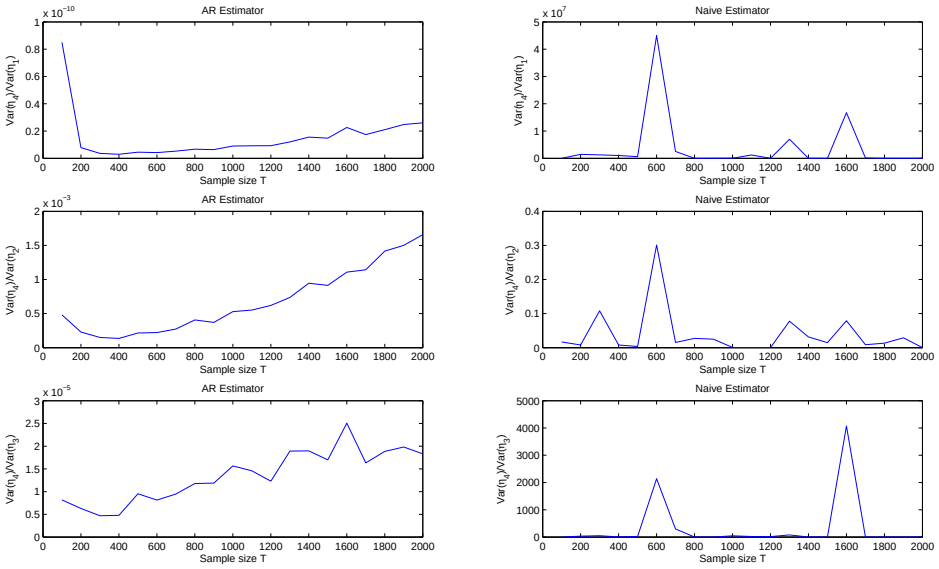

Figure 1: Ratio of Monte Carlo variances of components of $\hat{\eta}_T$ as a function of the sample size, respectively $\mathrm{Var}(\hat{\eta}_{4T})/\mathrm{Var}(\hat{\eta}_{1T})$ (top), $\mathrm{Var}(\hat{\eta}_{4T})/\mathrm{Var}(\hat{\eta}_{2T})$ (middle), $\mathrm{Var}(\hat{\eta}_{4T})/\mathrm{Var}(\hat{\eta}_{3T})$ (bottom): left panel for our estimator; right panel for the naive estimator.

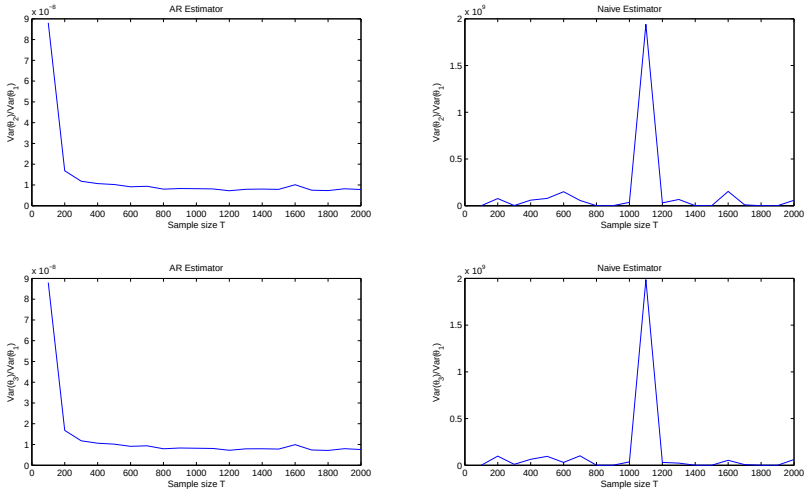

Figure 2: Ratio of Monte Carlo variances of components of $\hat{\theta}_T$ as a function of the sample size, respectively $\mathrm{Var}(\hat{\theta}_{2T})/\mathrm{Var}(\hat{\theta}_{1T})$ (top), $\mathrm{Var}(\hat{\theta}_{3T})/\mathrm{Var}(\hat{\theta}_{1T})$ (bottom): left panel for our estimator; right panel for the naive estimator.