Continuously updated extremum estimators*

Valentin PATILEA† Eric RENAULT‡

August 1997

Abstract

An important class of structural econometric models (nonlinear rational expectations, option pricing, auction models, ...) characterize observable variables as highly nonlinear transformations of some latent variables. These transformations are one-to-one, but they depend on the unknown distribution of the latent variables through the equilibrium of the game and/or the learning process. Therefore numerical complexity of the equilibrium definition generates substantial obstacles for the direct implementation of maximum likelihood inference. This is a particular case of argmax estimators based on a untractable sample based criterion $Q_T\left[\theta, \lambda(\theta)\right]$ contaminated by the occurrences of θ in a 'nuisance function' $\lambda(\theta)$ to which corresponds a simple criterion $Q_T\left[\theta, \lambda(\theta^0)\right]$ with θ^0 the true, unknown value of the parameter. The natural idea is to replace the unknown value $\lambda(\theta^0)$ by some 'good proxy', say $\lambda(\theta^1)$, to maximize $Q_T\left[\theta, \lambda(\theta^1)\right]$ with respect to θ and to get a new, updated estimate which, in turn, can be used for approximating $\lambda(\theta^0)$,... Such steps can be considered for a fixed sample size (iterative $M-$estimators), or each time new data arrive (recursive estimators). In this paper we present and analyze several iterative and recursive (Robbins-Monro type) estimation procedures of this kind which, at the end, are applied to the class of structural econometric models which motivated this work.

Keywords: Iterative procedures, $M-$estimators, Robbins-Monro procedures, structural models, latent variables

*The authors thank Halbert White for drawing their attention on the recursive estimation procedures relevant for the framework considered in this paper.

†Institut de Statistique, Université Catholique de Louvain. The first author gratefully acknowledges research support from the contract "Projet d'Actions de Recherche Concertées" (PARC No.93/98-164) of the Belgium Government.

‡GREMAQ-IDEI, Université de Toulouse I et Institut Universitaire de France.

1 Introduction

P. Young (1985) distinguishes two types of data processing for inference purposes in dynamic models. On one hand, iterative data processing used to maximize a given criterion as, for example, in "iterative least squares or some other relaxation procedures". Such kind of procedures involve, at each stage, the batch processing of the full block of data, but the highly nonlinear computation of the required estimator is decomposed in a sequence of simpler computations of intermediate estimators. On the other hand, the recursive approach which "introduces an extra dimension to estimation : in addition to the en bloc estimates based on complete data set (...), the analyst is also able to obtain estimates of the parameters for (...) subsets of the data, in a computationally elegant and efficient manner. This approach "also opens up the way for the estimation of (...) states in stochastic-dynamic systems".

In this paper we show how the two aspects, iterative/recursive, are both useful for extremum estimation in a large class of nonlinear dynamic systems. The basic framework we have in mind is a case where a natural sample based criterion, say, $Q_T\,[\theta,\, \lambda(\theta)]$ for extremum (argmax) estimation of a structural parameter θ is 'contaminated' by other occurrences of θ in a 'nuisance function' $\lambda(\theta)$. From the consistency and/or simplicity point of view, maximizing $Q_T\,[\theta,\, \lambda(\theta)]$ with respect to θ does not produce a satisfactory procedure. However, maximizing the criterion $Q_T\,[\theta,\, \lambda(\theta^0)]$ with respect to θ should provide a convenient (consistent, easy to compute, precise,...) estimator of θ^0, the true unknown value of the parameter, provided that we are able to compute this criterion. The problem is, of course, that we are not able to compute it. Below we will exhibit a class of structural econometric models where the inference issue reveals, in a natural way, a pattern as described above. While standard efficient inference on θ is infeasible, or at least unpalatable, some structural features of the model suggest a simple, but uncomputable, criterion $Q_T\,[\cdot,\, \lambda(\theta^0)]$ which, roughly speaking, would also allow efficient inference. This motivates us to look for continuously updated proxies of this criterion, either by 'nearly efficient' iterative procedures, or by less computationally demanding recursive approaches.

We argue that in many econometric models leading to a sample based criterion $Q_T\,[\theta,\, \lambda(\theta)]$ there exists a duality between the 'nuisance function' $\lambda(\theta)$ and unobserved state variables. In order to understand this duality we refer to two alternative interpretations of a standard recursive procedure of estimation proposed by Hannan and Rissanen (1982) for ARMA models. Consider a ARMA(p, q) process

$$Y_t - \sum_{j=1}^{p} \alpha_j Y_{t-j} = e_t - \sum_{i=1}^{q} \beta_i e_{t-i},$$

$E[e_t] = 0$, $E[e_s e_t] = \delta_{st}\sigma_e^2$, as considered by Hannan and Rissanen. The first T realizations of the ARMA process are observed. Assume that σ_e^2 is known and denote $\theta = (\alpha_1, ..., \alpha_p, \beta_1, ..., \beta_s)$. The simple, uncomputable criterion for estimating θ is the least squares criterion

$$-\frac{1}{T}\sum_{t=1}^{T}\left[Y_t - \sum_{j=1}^{p} \alpha_j Y_{t-j} + \sum_{i=1}^{q} \beta_i e_{t-i}\right]^2.$$

The innovations e_t, $t \geq 1$, play the role of (unobserved) latent variables. Define $\lambda(\theta) = (\lambda_h(\theta))_{1\leq h\leq n}$ as being the coefficients of an AR(n) representation of the observed process Y_t, $t \geq 1$ (typically $n = \log T$). A 'guess' of $\lambda(\theta)$, say $\hat{\lambda}$, can be obtained from the Yule-Walker equations. This guess provides proxies $\hat{e}_t$ of the innovations and thus it allows to build

$$-\frac{1}{T}\sum_{t=1}^{T}\left[Y_t - \sum_{j=1}^{p} \alpha_j Y_{t-j} + \sum_{i=1}^{q} \beta_i \hat{e}_{t-i}\right]^2$$

as proxy of the simple least squares criterion.[1] First, Hannan and Rissanen (1982) method can be interpreted as a procedure for providing an approximation of the simple least squares criterion. On the other hand, it can be understood as a procedure for obtaining good approximations of the latent process of innovations which, in turn, allow for simpler inference on the parameters of interest. The extension of both interpretations to nonlinear models is the main interest of this paper.

The paper is organized as follows. In section 2 we describe the general framework and we analyze the identification issue. In section 3 we present an extension of an iterative procedure proposed by Renault and Touzi (1996) and we prove its convergence. Section 3 ends with a recursive argmax estimator which represents the bridge towards the recursive estimators presented in section 4. We argue that the recursive techniques have a wide potential of application in the general framework considered in this paper, particularly in the presence of an unobserved state variable process

[1] Note that this example asks for a slightly more general framework than that considered in the sequel of the paper. More precisely, in this example the simple criterion, say $Q_T\left[\cdot, \overline{\lambda}(\theta^0)\right]$ depends on an unknown function $\overline{\lambda}(\cdot)$ and thus an approximation of it cannot be directly obtained as below in this paper, that is by plugging-in a proxy of θ^0 in the function $\overline{\lambda}(\cdot)$.

that enters the structural model nonlinearly. Three recursive (Robbins-Monro type) methods are proposed in section 4. The asymptotic behavior of the iterative and recursive estimators are deduced in sections 5 and 6, respectively. In section 7 we apply the general procedures developed in the previous section to the case of structural econometric models where the observable variables are defined as highly nonlinear functions of some latent variables. These functions are one-to-one, but they depend on the unknown distribution of the latent variables. This kind of models represent the principal motivation of this paper.

2 The general framework and the identitification problem

Consider that we observe a sequence of random vectors $(Y_t)_{t\geq 1}$ defined on some abstract probability space $(\Omega, \mathcal{F}, \mathbb{P})$ and taking values in a subset of $\mathbb{R}^J$, $J \geq 1$. Say that the structural econometric model we consider impose a sample based criterion (objective function)

$$Q_T\left[\theta,\, \lambda(\theta)\right] = Q_T\left[\theta,\, \lambda(\theta), (Y_t)_{1\leq t\leq T}\right]$$

for the estimation of θ. The parameter of interest θ lies in Θ, a subset of $\mathbb{R}^p$. The function $\lambda(\cdot)$ is known and it is defined on Θ and it takes values in Γ, a subset of $\mathbb{R}^q$. Typically, $\lambda(\Theta) = \Gamma$. Let us remark that for convergence issues the sets Θ and Γ could be quite general (see Proposition 3.1 below). However, most of the time we consider them subsets of Euclidian spaces.

As announced in the introduction, we have in mind the case where we cannot, or we should not, maximize Q_T with respect to all the occurrences of the parameter of interest. However, $Q_T\left[\theta,\, \lambda(\theta^0)\right]$ represents a simple, but impossible to compute, criterion for estimating θ. The natural idea is to replace the unknown value $\lambda(\theta^0)$ by some 'good proxy' and to maximize the simple criterion obtained in this way. We call this a two stages estimation strategy. The 'good proxies' of $\lambda(\theta^0)$ will be obtained from the previous steps of the estimation procedure. In other words, once we have, say, θ^1 an estimate of θ^0, $\lambda(\theta^1)$ will be a proxy for $\lambda(\theta^0)$.[2] Maximizing $Q_T\left[\theta,\, \lambda(\theta^1)\right]$ with respect to θ we get a new (updated) estimate of θ^0 which can be

[2] A more general case where the function $\lambda(\cdot)$ is unknown or practically uncomputable, and thus it should also be approximated, will be considered in future work.

used in the next step. The steps could be considered for a fixed sample (iterative, or cross-sectional procedures), or, each time that a new data arrive, a new step is performed (recursive procedures).

These facts lead us to the definition of a two parameter criterion

$$Q_T\left[\theta,\, \lambda\right] = Q_T\left[\theta,\, \lambda, (Y_t)_{1\leq t\leq T}\right], \quad (\theta, \lambda) \in \Theta \times \Gamma,$$

deduced from the original one. We consider a first set of assumptions similar to those usually imposed for $M-$estimation in the presence of nuisance parameters (see, e.g., section 4 of Wooldridge (1994)).

Assumption 2.1 *i) For any $T \geq 1$, $Q_T(\cdot\,,\,\cdot)$ satisfies the standard measurability and continuity conditions, i.e., it is measurable as function of observations and it is continuous as a function of parameters (θ, λ).*

ii) There exists a limit criterion $Q_\infty(\cdot\,,\,\cdot) : \Theta \times \Gamma \to I\!R$ such that

$$\lim_{T\to\infty} E\left[Q_T(\theta, \lambda)\right] = Q_\infty(\theta, \lambda), \quad \forall(\theta, \lambda) \in \Theta \times \Gamma,$$

where the expectation is computed with respect to the true distribution of the observations which depends on θ^0.

A problem to be addressed is, of course, the identifiability of θ^0. A usual identification condition for $M-$estimation based on the full criterion $Q_T\left[\theta, \lambda(\theta)\right]$ would state that θ^0 is the unique maximizer of $Q_\infty\left[\theta, \lambda(\theta)\right]$. However, the starting point for the estimation procedures we consider herein is the simpler criterion $Q_T\left[\theta, \lambda(\theta^0)\right]$. Therefore, a basic identification condition to be imposed in our framework is that θ^0 is the unique maximizer of the simplified limit criterion $Q_\infty\left[\theta, \lambda(\theta^0)\right]$.[3] This condition is quite similar with the usual identification condition that would have been imposed if $\lambda = \lambda(\theta)$ was a nuisance parameter (see, e.g., Wooldridge (1994), Theorem 4.3). The basic identification condition should be strengthened in order to take into account our two stages estimation strategy, i.e., the fact that we maximize $Q_T\left[\theta, \lambda(\theta^0)\right]$ after getting an approximation of $\lambda(\theta^0)$.

In order to get an idea about how we should reinforce the basic identifiability condition, assume that, for $\lambda \in \Gamma$, the limit criterion $Q_\infty(\theta, \lambda)$, viewed as a function

[3] It is clear that, in general, the identification condition based on the simplified limit criterion neither implies, nor is implied by the identification condition based on the full limit criterion.

of θ, admits a unique maximizer $\overline{\theta}(\theta^0, \lambda) \in \Theta$. This unique maximizer depends on θ^0, since the limit criterion depends on θ^0. Note that the basic identification condition can be written under the form

$$\overline{\theta}(\theta^0, \lambda(\theta^0)) = \theta^0. \tag{2.1}$$

Intuitively, the strengthened identification condition should state that the only fixed point of the function $\overline{\theta}(\theta^0, \lambda(\cdot))$ is θ^0. Otherwise, it may happen that the statistician will not be able to reject a 'bad guess' $\theta^1 \neq \theta^0$ used to build the simplified criterion $Q_T\left[\theta, \lambda(\theta^1)\right]$. A justification for the unique fixed point condition could go as follows. Let $\tilde{\theta}$ and θ^0 be two different parameters corresponding to the same distribution of the observations. Maximizing the corresponding limit criteria (see Assumption 2.1 *ii)*) for $\theta^1 = \theta^0$ and using (2.1), we deduce

$$\overline{\theta}(\tilde{\theta}, \lambda(\tilde{\theta})) = \overline{\theta}(\theta^0, \lambda(\theta^0)) = \theta^0.$$

Finally, the unique fixed point condition applied to $\overline{\theta}(\tilde{\theta}, \lambda(\cdot))$ implies $\tilde{\theta} = \theta^0$. These remarks lead us to the following identification assumption.

Assumption 2.2 *i) For any $\theta^1 \in \Theta$, the function $\theta \to Q_\infty\left[\theta, \lambda(\theta^1)\right]$ admits a unique maximizer denoted by $\overline{\theta}(\theta^0, \lambda(\theta^1))$.*

ii) If θ^0 is the value of the parameter corresponding to the distribution governing the observations, then θ^0 is the unique fixed point of the map $\overline{\theta}(\theta^0, \lambda(\cdot))$.

Imagine now that we want to estimate the parameter of interest using the first order conditions associated to the given maximization problem instead of maximizing the criterion in hand. Assuming the necessary regularity conditions define

$$M(\theta, \theta^1) = \frac{\partial Q_\infty}{\partial \theta}\left[\theta, \lambda(\theta^1)\right] \tag{2.2}$$

and assume that, for any $\theta^1 \in \Theta$, $\overline{\theta}(\theta^0, \lambda(\theta^1))$ is the unique solution of the equation $M(\theta, \theta^1) = 0$. Then, the unique fixed point property of $\overline{\theta}(\theta^0, \lambda(\cdot))$ is equivalent with the condition that $M(\theta^1, \theta^1) = 0$ implies $\theta^1 = \theta^0$. In other words, Assumption 2.2 could be replaced in this case by the following one.

Assumption 2.3 *Let θ^0 be the true unknown value of the parameter and assume that, for any $\theta^1 \in \Theta$, $\overline{\theta}(\theta^0, \lambda(\theta^1))$ defined in Assumption 2.2 i) is the unique solution of the equation $M(\theta, \theta^1) = 0$. Moreover,*

$$M(\theta^1, \theta^1) = 0 \Longrightarrow \theta^1 = \theta^0.$$

Assumption 2.2 will be used for iterative estimation procedures, while Assumption 2.3 will be invoked for recursive (Robbins-Monro type) estimation procedures.

3 Iterative $M-$estimation

In order to analyze the estimation strategy described above, let us define[4], for any $\lambda \in \Gamma$,

$$\overline{\theta}_T(\lambda) = \arg\max_{\theta \in \Theta} Q_T(\theta, \lambda). \tag{3.1}$$

Note that $\overline{\theta}_T(\lambda(\theta^0))$ is nothing else than the $M-$estimator defined through the simple criterion $Q_T(\theta, \lambda(\theta^0))$. Moreover, remark that $\overline{\theta}_T(\lambda)$ is the sample based counterpart of $\overline{\theta}(\theta^0, \lambda)$ defined in section 2. We need the following uniform convergence assumption on $Q_T(\theta, \lambda)$ in order to obtain the uniform convergence of $\overline{\theta}_T(\lambda)$ towards $\overline{\theta}(\theta^0, \lambda)$ (see also the usual uniform assumptions imposed for $M-$estimation in the presence of nuisance parameters; e.g., Wooldridge (1994)).

Assumption 3.1 *Assume that $Q_T(\cdot, \cdot)$, $T \geq 1$, satisfies the measurability condition of Assumption 2.1 i). Moreover,*

$$\sup_{(\theta,\lambda) \in \Theta \times \Gamma} |Q_T(\theta, \lambda) - Q_\infty(\theta, \lambda)| \xrightarrow{p} 0.$$

with Q_∞ defined as in Assumpton 2.1 ii).

The convergence results of this section rely on the following proposition which we state for a quite general parameter space.

Proposition 3.1 *Assume that Θ and Γ are compact subsets of some metric spaces $(\tilde{\Theta}, d_1)$ and $(\tilde{\Gamma}, d_2)$, respectively. Moreover, the map $\overline{\theta}(\theta^0, \cdot) : \Gamma \to \Theta$ is continuous. If Assumption 2.1 and Assumption 3.1 hold and $\overline{\theta}(\theta^0, \lambda)$ is the unique maximizer of $\theta \to Q_\infty(\theta, \lambda)$, then*

$$\sup_{\lambda \in \Gamma} d_1 \left(\overline{\theta}_T(\lambda), \, \overline{\theta}(\theta^0, \lambda) \right) \xrightarrow{p} 0.$$

[4] Remark that $\overline{\theta}_T(\lambda)$ may not be uniquely defined. The asymptotic results hold for any version of $\overline{\theta}_T(\lambda)$, $T \geq 1$.

Proof See appendix

Let us now present two iterative $M-$estimation procedures. First, inspired by Renault and Touzi (1996), Patilea, Ravoteur and Renault (1995) (PRR hereafter) and Renault (1995) proposed the following estimator : consider an arbitrary sequence $p(T)$, $T \geq 1$, of positive integers such that $\lim_{T\to\infty} p(T) = \infty$ and define

$$\theta_T = \theta_T^{(p(T)+1)}, \ \ T \geq 1, \tag{3.2}$$

where, for any $p \geq 1$,

$$\theta_T^{(p+1)} = \overline{\theta}_T \left(\lambda(\theta_T^{(p)}) \right)$$

and $\theta_T^{(1)} \in \Theta$ is some starting value. This estimator is an extension and a practical version of the Renault and Touzi's estimator defined for a log-likelihood type criterion and $p(T) \equiv \infty$. PRR and Renault (1995) impose the following contracting assumption, a reinforcement of the unique fixed point condition states in Assumption 2.2 *ii)*.

Assumption 3.2 *Let $(\tilde{\Theta}, d_1)$ be a metric spaces containing the set Θ. If θ^0 is the true unknown value of the parameters, the mapping*

$$\overline{\theta}(\theta^0, \lambda(\cdot)) : \Theta \longrightarrow \Theta$$

is strongly contracting, that is there exists $k = k(\theta^0) \in (0,1)$ such that, $\forall\, \theta_1,\ \theta_2 \in \Theta$,

$$d_1 \left(\overline{\theta}(\theta^0, \lambda(\theta^1))\, , \overline{\theta}(\theta^0, \lambda(\theta^2)) \right) \leq k\, d_1 \left(\theta^1 \, , \, \theta^2 \right).$$

We are able now to state the weak consistency result for the estimator considered by PRR and Renault. Again, Θ and Γ could be quite general parameter sets.

Proposition 3.2 *If Θ is as in Proposition 3.1 and the Assumptions 2.1, 2.2 i), 3.1*[5] *and 3.2 hold, then θ_T defined in (3.2) is weakly consistent.*

[5]Note that we only need the uniform convergence of $Q_T[\cdot, \lambda(\cdot)]$, in probability. By abuse, from now on, any time we invoke Assumption 3.1 we think at the convergence of $Q_T[\cdot, \lambda(\cdot)]$, in probability, uniformly with respect to θ, $\theta^1 \in \Theta$.

Proof See appendix.

Given the computational burden implied by the estimator presented above we try to relax the condition $\lim_{T\to\infty} p(T) = \infty$. Moreover, we are interested to define 'on-line' procedures providing estimators which can be easily updated when new data arrive. Motivated by these aspects we propose in this paper several recursive procedures. A first such procedure is introduced in this section, while the rest will be presented in the next section.

The first recursive estimator we propose is built as follows: fix some[6] $t_0 < T$ and an initial value $\hat{\theta}_{t_0}$ and, for any $t_0 \le t \le T$, take

$$\hat{\theta}_{t+1} = \overline{\theta}_t \left(\lambda(\hat{\theta}_t) \right). \tag{3.3}$$

The recursive estimator we propose is exactly $\hat{\theta}_{T+1}$. Remark that this recursive estimator is an estimator as defined in (3.2) with $p(T) = 1$ and initial value the estimate obtained without the last observation, estimate which, in turn, is obtained for $p(T-1) = 1$ and initial value the estimate obtained without the last two observations, ... We end this section with the consistency of this recursive estimator.

Proposition 3.3 *Under the assumptions of Proposition 3.2 the estimator defined in (3.3) is weakly consistent.*

Proof See appendix.

Even if we are not able to provide the limit distribution for our first recursive estimator, it can be used for producing initial values for other estimation procedures. Moreover, its simple form could become even more interesting when Θ is an infinite-dimensional parameter space.

4 Robbins-Monro type estimators

In this subsection, instead of applying the analogy principle (see, e.g, Wooldridge (1994)) using a maximization problem solved by the population parameters, we

[6] Let us remark that the choice of t_0 does not influence the asymptotic results. For the sake of simplicity, for all the recursive estimators presented in the sequel of this paper we will choose $t_0 = 1$.

rather use this principle starting from first order conditions. For this purpose let us assume that the sample based criterion $Q_T(\theta,\lambda)$ defined in section 2 has the form

$$Q_T(\theta,\lambda) = \frac{1}{T}\sum_{t=1}^{T} q_t(\theta,\lambda), \quad (\theta,\lambda) \in \Theta \times \Gamma,$$

with Θ a subset of an $I\!R^p$ and

$$q_t(\theta,\lambda) = q_t(w_t,\theta,\lambda)$$

where $(w_t)_{t\geq 1}$ is defined through a deterministic function of the observed process $(Y_t)_{t\geq 1}$. Typically, if the observations are i.i.d. then $w_t = Y_t$ while if Y_t, $t \geq 1$, is Markovian of order $p \geq 1$, $w_t = Y_t, Y_{t-1}, ..., Y_{t-p}$. Moreover, $q_t(\cdot,\cdot)$ is continuously differentiable as function of (θ,λ). Consider

$$M(\theta^1) = M(\theta^1,\theta^1) = \frac{\partial Q_\infty}{\partial \theta}\left[\theta, \lambda(\theta^1)\right]\Bigg|_{\theta=\theta^1}, \tag{4.1}$$

where $Q_\infty(\theta,\lambda)$ is the limit criterion from Assumption 3.1. Typically, we also have

$$M(\theta^1) = \lim_{T\to\infty} E\left[\frac{\partial Q_T}{\partial \theta}\left[\theta,\lambda(\theta^1)\right]\Bigg|_{\theta=\theta^1}\right]$$

Recall that under Assumption 2.3

$$M(\theta^1) = 0 \Longrightarrow \theta^1 = \theta^0,$$

i.e., θ^0 is the unique solution of the population problem defined by $M(\cdot)$. In this section we estimate this solution in a recursive way using *stochastic approximation* or Robbins-Monro procedures. For a description and the properties of such procedures see Robbins and Monro (1951) and, among many others, Ljung (1977), Kushner and Clark (1978), Chen and White (1992), Sargent (1993) and Kuan and White (1994a, b). For our purposes, a Robbins-Monro (RM hereafter) procedure could be presented as follows: given a population problem $M(\eta) = 0$, defined by some $M(\cdot) : \Delta \to \Delta$ and having a unique solution η^0, for any $1 \leq t \leq T$, write a sample counterpart of $M(\cdot)$ as

$$M_t(\eta) = M(\eta) + U_t(\eta),$$

with $(U_t(\eta))_{t\geq 1}$ satisfying some conditions. Afterwards, starting from some initial value η_1, define an estimator for η^0 as being the last value obtained from the recursion

$$\eta_{t+1} = \eta_t + a_t M_t(\eta_t), \tag{4.2}$$

$1 \leq t \leq T$, where $(a_t)_{t\geq 1}$ is a sequence of positive real numbers decreasing to zero. Basically a_t should tend to zero as fast as $t^{-\alpha}$ for some $\alpha \in (1/2, 1]$. For the purposes of this paper $a_t = ct^{-1}$, with $c > 0$ some scaling factor. Moreover, as the results of this subsection are not influenced by c we consider it for the moment as being equal to one. The (random) function $M_t(\cdot)$ can be seen as a measurement of $M(\cdot)$ at time t (or a 'learning update function') and a_t as a 'learning rate'. Usually it is assumed that the averaged 'errors' $U_t(\eta_t)$, $t \geq 1$, vanish almost surely, more precisely $a_T \sum_{t=1}^{T} U_t(\eta_t) \to 0$, a.s.

As it has been mentioned by Kuan and White (1994a), "the RM algorithm has two significant advantages: (i) its recursive nature places few demands on computer resources; and (ii) in theory, just one pass through a sufficiently large data set can yield a consistent estimate. The RM algorithm is therefore particularly appealing for estimating parameters in nonlinear models in large data sets."

The convergence results of RM procedures are usually obtained via the ordinary differential equation (ODE) method proposed by Ljung (1977). The idea of Ljung was to show that the sequence $(\eta_t)_{t\geq 1}$ defined in (4.2) asymptotically follow a trajectory (solution) of a deterministic ODE

$$\frac{d\eta}{dt}(t) = M(\eta(t)) \tag{4.3}$$

associated to the population problem $M(\eta) = 0$. If any trajectory of this ODE converges (as $t \to \infty$) to η^0, that is if the ODE satisfies a (asymptotic) stability condition, then the consistency of the recursive estimates is ensured. The stability condition is ensured if $M(\cdot)$ is a differentiable function and the matrix $(\partial/\partial\eta')M(\eta^0)$ is negative stable, that is if all eigenvalues of this matrix have (strictly) negative real parts (see, e.g., Kuan and White (1994a), p 40, Horn and Johnson (1991), p 90, or Rouche and Mawhin (1980), ch. 1). We recall that the Lyapunov theorem (see, e.g., Horn and Johnson (1991), p.96) tells that a square matrix A, with real elements, is positive stable, i.e., $-A$ is negative stable, if and only if there exists a positive definite G such that $GA + A'G$ is positive definite.

We apply the RM procedure to the population problem $M(\theta^1) = 0$ with $M(\cdot)$ defined in (4.1). Its unique solution θ^0 is assumed to belong to the interior of Θ.

First of all we check the stability condition. Assuming the necessary regularity conditions we write

$$\frac{\partial M}{\partial \theta^{1\prime}}(\theta^1) = \frac{\partial^2 Q_\infty}{\partial \theta \partial \theta'}\left[\theta, \lambda(\theta^1)\right]\Big|_{\theta=\theta^1} + \frac{\partial^2 Q_\infty}{\partial \theta \partial \lambda'}\left[\theta, \lambda(\theta^1)\right]\Big|_{\theta=\theta^1} \frac{\partial \lambda}{\partial \theta^{1\prime}}(\theta^1). \tag{4.4}$$

Let us differentiate the identity

$$\frac{\partial Q_\infty}{\partial \theta}\left[\theta, \lambda(\theta^1)\right]\Big|_{\theta=\overline{\theta}(\theta^0,\lambda(\theta^1))} = 0, \qquad \theta^1 \in \Theta,$$

with respect to θ^1. For $\theta^1 = \theta^0$ and under the unique fixed point property for $\overline{\theta}(\theta^0, \lambda(\cdot))$, we get

$$\frac{\partial^2 Q_\infty}{\partial \theta \partial \theta'}\left[\theta, \lambda(\theta^0)\right]\Big|_{\theta=\theta^0} \frac{\partial \overline{\theta}}{\partial \theta^{1\prime}}(\theta^0, \lambda(\theta^0)) + \frac{\partial^2 Q_\infty}{\partial \theta \partial \lambda'}\left[\theta, \lambda(\theta^1)\right]\Big|_{\theta=\theta^0} \frac{\partial \lambda}{\partial \theta^{1\prime}}(\theta^0), \tag{4.5}$$

where

$$\frac{\partial \overline{\theta}}{\partial \theta^{1\prime}} = \frac{\partial \overline{\theta}}{\partial \lambda'} \frac{\partial \lambda}{\partial \theta^{1\prime}}.$$

From (4.4) and (4.5) we deduce

$$\frac{\partial M}{\partial \theta^{1\prime}}(\theta^0) = \frac{\partial^2 Q_\infty}{\partial \theta \partial \theta'}\left[\theta, \lambda(\theta^0)\right]\Big|_{\theta=\theta^0} \left(I_p - \frac{\partial \overline{\theta}}{\partial \theta^{1\prime}}(\theta^0, \lambda(\theta^0))\right) \tag{4.6}$$

(we denote by I_p the $p-$dimensional identity matrix). A usual assumption for $M-$estimation with limit criterion $\theta \to Q_\infty[\theta, \lambda(\theta^0)]$ is that its Hessian matrix is negative definite (see, e.g., Wooldridge (1994), section 4). At this stage it seems natural to impose the assumption below (for a matrix A we denote by $\|A\|_\lambda$ the spectral norm of the matrix A, that is the square root of the largest eigenvalue of $A'A$).

Assumption 4.1 *Let θ^0 be the true unknown value of the parameters and assume that the function $\overline{\theta}(\theta^0, \lambda(\cdot))$ defined in Assumption 2.2 i) is continuously differentiable and*

$$\left\| \frac{\partial \overline{\theta}}{\partial \theta^{1\prime}}(\theta^0, \lambda(\theta^1))\Big|_{\theta^1=\theta^0} \right\|_\lambda < 1.$$

Assumption 4.1 has two important consequences. First, since it implies Assumption 3.2, it ensures the local identifiability of θ^0 from the observed variables. Secondly, it guarantees the stability condition for the ODE associated to the population problem defined by $M(\cdot)$ defined in (4.1) . Our second statement is a consequence of the following lemma.

Lemma 4.1 *If A_1 and A_2 are p-dimensional squared matrices with real elements such that A_1 is (symmetric) positive definite and $\|A_2\|_\lambda < 1$, then $A = A_1(I_p - A_2)$ is positive stable.*

Proof We apply Lyapunov theorem for $G = A_1^{-1}$. We have

$$\frac{1}{2}(G\,A + A'\,G) = I_p - \frac{1}{2}(A + A').$$

Now remark that $\|A_2\|_\lambda = \|A_2'\|_\lambda$ (see, e.g., Horn and Johnson (1985), exercise 11, p 312). Thus $(A_2 + A_2')/2$ is a symmetric matrix with spectral norm smaller than one. As a consequence $I_p - (A + A')/2$ is positive definite. ∎

We consider below a RM procedure based on a natural measurement of $M(\cdot)$, that is

$$M_t(\theta^1) = \frac{\partial q_t}{\partial \theta}(\theta, \lambda(\theta^1))\bigg|_{\theta=\theta^1}, \tag{4.7}$$

$t \geq 1$. Since there is no reason to believe that the RM procedure based on $M_t(\cdot)$ defined above will remain in Θ, we propose a kind of randomly truncated RM algorithm (see also Yin and Zhu (1990), Chen and White (1992)) : for any $t \geq 1$, $M_t(\cdot)$ is defined as in (4.7) and

$$\theta_{t+1} = \begin{cases} \theta_t + t^{-1}\,M_t(\theta_t) & \text{on } B_{t+1} \\ \overline{\theta}_t\,(\lambda(\theta_t)) & \text{on } B_{t+1}^c \end{cases}, \tag{4.8}$$

with

$$B_{t+1} = \{\theta_t + t^{-1}\,M_t(\theta_t) \in \Theta\} \subset \Omega$$

and $\overline{\theta}_t(\cdot)$ as in (3.1). For any $T \geq 1$, let θ_T^{RM} denote the value θ_{T+1} obtained from the previous recursive procedure and let us call θ_T^{RM} a simple RM estimator. Let

$D(T)$ count the number of times the second branch of (4.8) was used for defining θ_T^{RM}. That is, for any $1 \leq t \leq T$,

$$D(t+1) = D(t) + \mathbb{1}(B_{t+1}^c),$$

with $D(1) = 0$. For proving the convergence of the procedure (4.8) we use the assumption below. Let us recall that convergence results for RM estimators are usually obtained under 'almost surely' type hypotheses. Clark (1984) showed that in a RM algorithm the strong LLN for the errors $(U_t(\theta_t))_{t\geq 1}$ is a minimal condition for the almost surely convergence of the algorithm. On the other hand, he showed that the weak LLN for $(U_t(\theta_t))_{t\geq 1}$ is neither necessary nor sufficient for the weak convergence of $(\theta_t)_{t\geq 1}$. Moreover, if we want to ensure that, with probability one, the truncation procedure described in (4.8) is used finitely often, we need almost sure convergence.

Assumption 4.2 *The true unknown value of the parameters θ^0 belongs to the interior of Θ and Θ is assumed convex. The limit criterion $Q_\infty(\cdot,\cdot)$ admits continuous second order derivatives and $\lambda(\cdot)$ is continuously differentiable. Moreover, $\lambda(\theta^0)$ is an interior point of Γ. If $M(\cdot)$ is defined as in (4.1), then*

$$\sup_{\theta^1\in\Theta} \left\| \frac{\partial Q_T}{\partial \theta}\left[\theta, \lambda(\theta^1)\right]\Big|_{\theta=\theta^1} - M(\theta^1) \right\| \longrightarrow 0, \quad \boldsymbol{P} - \text{almost surely.}$$

We have the following proposition[7] (see Chen and White (1992)). See also Kuan and White (1994a) for a slightly different result.

Proposition 4.2 *Let $(\theta_T^{RM})_{T\geq 1}$ be a sequence of simple RM estimators obtained as in (4.8), starting with some initial value $\theta_1 \in \Theta$. Consider that $\Theta \subset I\!R^p$ is compact and Assumptions 2.1, 2.3, 4.1 and 4.2 hold. Moreover, the Hessian matrix*

$$\frac{\partial^2 Q_\infty}{\partial\theta\partial\theta'}\left[\theta, \lambda(\theta^0)\right]\Big|_{\theta=\theta^0} \tag{4.9}$$

of the simple limit criterion $Q_\infty\left[\cdot, \lambda(\theta^0)\right]$ is negative definite. Then, there exists a positive integer-valued random variable V such that

[7] The convergence results of RM procedures could be formulated under very general conditions on Θ and/or the properties of the observed process (see, e.g., Yin and Zhu (1990), Chen and White (1992)). For instance, we can take Θ a Hilbert space and transpose the assumptions above in terms of Fréchet differentiability and bounded operators. However, herein we confine ourselves to work with finite-dimensional parameter space.

$$\mathbb{P}(\sup_{T\geq 1} D(T) \leq V < \infty) = 1 \tag{4.10}$$

and $(\theta_T^{RM})_{T\geq 1}$ *is strongly consistent.*

Proof See appendix.

The previous proposition says that, almost surely, the second branch of (4.8) will be used only a finite number of times and thus, from the asymptotic point of view, its definition is not important. Let us remark that it may happen Θ to be only a part of the parameter space to which θ^0 is hoped to belong. If the statistician considers that the prior information on the location of θ^0 is not trustful, she/he may consider a sequence of compact sets $(\Theta_t)_{t\geq 1}$ growing to a 'trustful' set of parameters Θ_∞ and define any B_t with Θ replaced by Θ_t. If θ^0 is an interior point of Θ_∞, then the same convergence result could be obtained (see Yin and Zhu (1990), Chen and White (1992)). The same remark is also valid for other of the estimators presented above. However, we do not pursue this idea herein.

The simple RM estimator defined above has a very attractive feature, namely is a recursive estimator with a very simple updating scheme (at least on its first branch). A disadvantage of this estimator (see Kuan and White (1994a), KW hereafter, section II.3 and the references mentioned therein) is that it may converge quite slowly. A natural modification to be done in order to improve the speed of convergence is to take an approximate Newton-Raphson step at each stage. This yields a 'modified RM procedure', also called a 'stochastic Newton method' (see, e.g., KW or White (1989)). However, the derivative we need for this purpose has an unfriendly form since it involves the second derivative of $Q_\infty(\cdot,\cdot)$ with respect to θ and λ (see (4.4)). Given (4.6), a natural idea is to use the tractable part of $(\partial M/\partial\theta^{1'})(\theta^0)$, i.e., the Hessian matrix of the simple limit criterion $Q_\infty[\cdot,\lambda(\theta^0)]$.[8] For easier notation, let us define

$$\Sigma(\theta^1) = -\left.\frac{\partial^2 Q_\infty}{\partial\theta\partial\theta'}\left[\theta,\lambda(\theta^1)\right]\right|_{\theta=\theta^1}, \qquad \theta^1 \in \Theta. \tag{4.11}$$

For any $t \geq 1$, we define

[8] When $Q_\infty\left[\cdot,\lambda(\theta^0)\right]$ is a log-likelihood criterion (see section 7 , a perhaps easier procedure is obtained if we replace the Hessian matrix by minus score variance matrix.

$$\theta_{t+1}^M = \begin{cases} \theta_t^M + t^{-1}\, G_{t+1}^{-1}(\theta_t^M) M_t(\theta_t^M) & \text{on } B_{t+1}^M \\ \overline{\theta}_t\,(\lambda(\theta_t^M)) & \text{on } (B_{t+1}^M)^c \end{cases}, \tag{4.12}$$

where

$$G_{t+1}(\theta^1) = -\frac{1}{t}\sum_{s=1}^{t} \frac{\partial^2 q_s}{\partial\theta\partial\theta'}(\theta, \lambda(\theta^1))\bigg|_{\theta=\theta^1} \tag{4.13}$$

and $M_t(\cdot)$ is defined as in (4.7). Let $U^M(\theta^0)$ be a neighborhood of $\Sigma(\theta^0)$:

$$U^M(\theta^0) = \{G,\ \|vec(\Sigma(\theta^0) - G)\| \leq \rho\}, \quad \rho > 0,$$

($vec(G)$ stands for the p^2-dimensional vector obtained from the lines of a $p-$ dimensional square matrix G). We assume that $U^M(\theta^0)$ contains only invertible matrices. The events B_{t+1}^M, $t \geq 1$, are defined as

$$B_{t+1}^M = \left\{G_{t+1}(\theta_t^M) \in U^M(\theta^0) \ \text{ and } \theta_t^M + t^{-1}\, G_{t+1}^{-1}(\theta_t^M) M_t(\theta_t^M) \in \Theta\right\} \subset \Omega.$$

For any T we denote by θ_T^{MRM} the value θ_{T+1}^M obtained from the recursive procedure above. Let us call θ_T^{MRM} a modified RM estimator. Let $D^M(T)$ count the number of times the second branch of (4.12) was used for the definition of θ_T^{MRM}, that is, for any $1 \leq t \leq T$,

$$D^M(t+1) = D^M(t) + 1\!\!1(\,(B_{t+1}^M)^c).$$

with $D^M(1) = 0$.

Remark that the procedure (4.12) is a part of a randomly truncated RM procedure. The population problem is given by the function $M^M(\delta)' = (M_1^M(\delta)', M_2^M(\delta)')$ where

$$\delta' = (vec(G)', \theta^{1'}) \in I\!R^{p^2} \times \Theta \ \subset I\!R^{p^2+p}$$

and

$$M_1^M(\delta) = vec\left(\Sigma(\theta^1) - G\right) \tag{4.14}$$

$$M_2^M(\delta) = G^{-1} M(\theta^1), \tag{4.15}$$

with $M(\cdot)$ defined as in (4.1). The unique root of the equation $M^M(\delta) = 0$ is $(vec(\Sigma(\theta^0)', \theta^{0'})'$. Finally, remark that (4.8) can be obtained from (4.12) by replacing all $G_{t+1}^{-1}(\cdot)$ with the identity matrix.

We have the following convergence result.

Proposition 4.3 *Let $(\theta_T^{MRM})_{T\geq 1}$ be a sequence of modified RM estimators obtained from (4.12) starting from some initial value $\theta_1 \in \Theta$ and using an enough small neighborhood $U^M(\theta^0)$. Assume that the hypotheses of Proposition 4.2 hold and that $(G_t(\theta^1))_{t\geq 2}$ satisfies a strong LLN, uniformly in θ^1. Moreover, the function $\Sigma(\cdot)$ is continuously differentiable on Θ. Then, there exists a positive integer-valued random variable V such that*

$$\boldsymbol{P}(\sup_{T\geq 1} D^M(T) \leq V < \infty) = 1. \tag{4.16}$$

and the modified RM estimator is strongly consistent.

Proof See appendix.

We may simplify the estimator given in (4.12) by considering G_{t+1} a diagonal matrix. In this way we avoid a possibly difficult matrix inversion. In particular, for any t, we may take $G_{t+1} = e_{t+1} I_p$, with e_{t+1} a positive scalar. More precisely, we may define the following 'quick and dirty' procedure (see KW, p 48) : $\forall t \geq 1$,

$$\theta_{t+1}^Q = \begin{cases} \theta_t^Q + t^{-1}\, e_{t+1}^{-1}\, M_t(\theta_t^Q) & \text{on } B_{t+1}^Q \\ \overline{\theta}_t\,(\lambda(\theta_t^Q)) & \text{on } (B_{t+1}^Q)^c \end{cases}, \tag{4.17}$$

with

$$e_{t+1} = -tr\left(\frac{\partial^2 Q_t}{\partial\theta\partial\theta'} \left[\theta, \lambda(\theta_t^Q)\right] \Big|_{\theta=\theta_t^Q} \right) = tr\left(G_{t+1}(\theta_t^Q)\right).$$

Let $U^Q(\theta^0)$ be an interval of positive numbers centered at $tr\,(\Sigma(\theta^0))$. The event B_{t+1}^Q is defined as

$$B_{t+1}^Q := \left\{ e_{t+1} \in U^Q(\theta^0) \text{ and } \theta_t^Q + t^{-1}\, e_{t+1}^{-1} M_t(\theta_t^Q) \in \Theta \right\} \subset \Omega.$$

For any T we denote by θ_T^{QRM} the value θ_{T+1}^Q obtained from the previous recursion. As before, let $D^Q(T)$ count the number of times the algorithm (4.17) passed through its second branch in order to compute θ_T^{QRM}, that is, for any $1 \leq t \leq T$,

$$D^Q(t+1) = D^Q(t) + \mathbb{1}(\,(B_{t+1}^Q)^c).$$

with $D^Q(1) = 0$.

The function giving the population problem corresponding to the 'quick and dirty' RM algorithm (4.17) is $M^Q(e,\theta^1)' = (M_1^Q(e,\theta^1),\, M_2^Q(e,\theta^1)')$ where

$$M_1^Q(e,\theta^1) = tr\left(\Sigma(\theta^0)\right) - e \tag{4.18}$$

and

$$M_2^Q(e,\theta^1) = \frac{1}{e}\, M(\theta^1), \tag{4.19}$$

with $M(\cdot)$ as defined in (4.1). The unique root of $M(\cdot,\cdot)$ above is $(\Sigma(\theta^0),\, \theta^0)$. We state below the convergence result of the 'quick and dirty' estimator.

Proposition 4.4 *Let $(\theta_T^{QRM})_{T\geq 1}$ be a sequence of estimators obtained from (4.17) starting from some initial value $\theta_1 \in \Theta$ and using an enough small neighborhood $U^Q(\theta^0)$. Assume that the hypotheses of Proposition 4.2 hold and that*

$$\sup_{\theta^1\in\Theta}\left| tr\left(\frac{\partial^2 Q_t}{\partial\theta\partial\theta'}\left[\theta,\lambda(\theta^1)\right]\Bigg|_{\theta=\theta^1} - \frac{\partial^2 Q_t}{\partial\theta\partial\theta'}\left[\theta,\lambda(\theta^1)\right]\Bigg|_{\theta=\theta^1} \right)\right|, \quad \boldsymbol{P}\text{- almost surely.}$$

Moreover, the function $tr\,(\Sigma(\cdot))$ is continuously differentiable on Θ. Then, there exists a positive integer-valued random variable V such that

$$\boldsymbol{P}(\sup_{T\geq 1} D^Q(T) \leq V < \infty) = 1, \tag{4.20}$$

and the 'quick and dirty' RM estimator θ_T^{QRM} is strongly consistent.

Proof See appendix

The 'quick and dirty' RM estimator is a compromise between the simple RM estimator and the modified RM estimator in that it uses some local curvature information through a scaling factor. The simple RM procedure can be seen as an extreme case of 'quick and dirty' RM estimator which corresponds to a scaling factor fixed at one. As a consequence this last estimator ought to concentrate more quickly than the simple RM estimator, but more slowly than the modified RM estimator, around the unknown value of the parameters. This is for instance the case for the nonlinear least squares problem considered by KW. However, it is clear that in a more general framework we cannot immediately draw such a conclusion. The comparisons of the performances of the RM estimators we presented above should heavily depend on the model that is considered.

5 The asymptotic distribution of the iterative $M-$estimators

In section 3 we presented two estimators : an extension of Renault an Touzi (1996)'s estimator (see (3.2)) and a recursive $M-$estimator (see (3.3)). We believe that, under some conditions, the rate of convergence of the recursive estimator is $1/\sqrt{T}$. We do not know its asymptotic distribution. It should be less concentrated than that of the first estimator. Below in this section we (re)analyze the limit distribution of the estimator defined in (3.2). Before stating our results we need some hypotheses (see, e.g., Wooldridge (1994), section 4). The sets Θ and Γ are assumed to be subsets of Euclidian spaces.

Assumption 5.1 *If θ^0 is the true unknown value of the parameters, then*

$$\sqrt{T}\ \frac{\partial Q_T}{\partial \theta}\left[\theta, \lambda(\theta^0)\right]\Big|_{\theta=\theta^0} \xrightarrow{d} N_p\left(0,\ B(\theta^0)\right)$$

with

$$B(\theta^0) = \lim_{T\to\infty} Var\left(\sqrt{T}\ \frac{\partial Q_T}{\partial \theta}\left[\theta, \lambda(\theta^0)\right]\Big|_{\theta=\theta^0}\right)$$

which is supposed to be positive definite.

The asymptotic normality of the score vector of the simple criterion $Q_T[\cdot, \lambda(\theta^0)]$ is as usual hypotheses for the $M-$estimators. Another usual assumption is the uniform convergence, in probability, of the Hessian matrix of the criterion to be maximized. In our framework this corresponds to the following assumption.

Assumption 5.2 *The matrix*

$$\Sigma(\theta,\theta^1) = -\frac{\partial^2 Q_\infty}{\partial\theta\partial\theta'}\left[\theta,\lambda(\theta^1)\right],$$

exists for any $\theta,\theta^1 \in \Theta$ *and its components are continuous functions of* (θ,θ^1). $\Sigma(\theta^0,\theta^0)$ *is asumed positive definite. Moreover, when* $T\to\infty$,

$$\sup_{\theta,\,\theta^1\in\Theta}\left\|\frac{\partial^2 Q_T}{\partial\theta\partial\theta'}\left[\theta,\lambda(\theta^1)\right]-\Sigma(\theta,\theta^1)\right\|_\lambda \longrightarrow 0, \tag{5.1}$$

in probability.

Note that $\Sigma(\theta^1,\theta^1)$ coincides with $\Sigma(\theta^1)$ written in (4.11). As it has been remarked in section 4, $-\Sigma(\theta^0,\theta^0)$ is nothing else that the Hessian matrix, considered for $\theta=\theta^0$, of the simple limit criterion $Q_\infty[\cdot,\lambda(\theta^0)]$. This matrix is usually assumed negative definite. The next assumption is more specific to our framework.

Assumption 5.3 *For any* $\theta^1\in\Theta$, *define*

$$H(\theta^1) = \frac{\partial^2 Q_\infty}{\partial\theta\partial\lambda'}\left[\theta,\lambda(\theta^1)\right]\Big|_{\theta=\theta^1}\frac{\partial\lambda}{\partial\theta^{1\prime}}(\theta^1),$$

$$H_T(\theta^1) = \frac{\partial^2 Q_T}{\partial\theta\partial\lambda'}\left[\theta,\lambda(\theta^1)\right]\Big|_{\theta=\theta^1}\frac{\partial\lambda}{\partial\theta^{1\prime}}(\theta^1),$$

and assume that $H(\cdot)$ *is continuous. Then*

$$\sup_{\theta^1\in\Theta}\left\|H_T(\theta^1)-H(\theta^1)\right\|_\lambda \longrightarrow 0,$$

in probability.

Let us remark that (4.4) and (4.6) imply

$$H(\theta^0) = -\Sigma(\theta^0,\theta^0)\,\frac{\partial\overline{\theta}}{\partial\theta^{1\prime}}(\theta^0,\lambda(\theta^0))\Big|_{\theta^1=\theta^0}.$$

Moreover, under Assumption 4.1, $H(\theta^0)+\Sigma(\theta^0,\theta^0)$ is invertible. We have the following asymptotic result.

Proposition 5.1 *Assume that θ^0, the true unknown value of the parameters, is an interior point of $\Theta \subset I\!R^p$. Consider that Assumptions 2.1, 2.2 i), 3.1, 3.2 and 5.1 to 5.3 hold and that $Q_\infty[\cdot, \lambda(\cdot)]$ is continuous on $\Theta \times \Theta$. If, in addition, the sequence $p(T)$, $T \geq 1$, considered in (3.2) is such that*

$$\sqrt{T}\,\left(\theta_T - \theta_T^{(p(T))}\right) = \sqrt{T}\,\left(\theta_T^{(p(T)+1)} - \theta_T^{p(T)}\right) \longrightarrow 0,$$

in probability, then the θ_T is asymptotically normal with asymptotic variance matrix

$$V(\theta^0) = A(\theta^0)^{-1}\Sigma(\theta^0,\theta^0)^{-1}B(\theta^0)\Sigma(\theta^0,\theta^0)^{-1}A(\theta^0)'^{-1}.$$

where

$$A(\theta^0) = I_p - \frac{\partial\overline{\theta}}{\partial\theta^{1\prime}}(\theta^0,\theta^1)\bigg|_{\theta^1=\theta^0}.$$

Proof See appendix.

Let us remark that the estimator θ_T has an asymptotic variance quite similar with the asymptotic variance of the $M-$estimator obtained from the simple criterion $Q_T[\cdot, \lambda(\theta^0)]$. A natural question arises : $V(\theta^0)$ is greater or smaller than

$$\Sigma(\theta^0,\theta^0)^{-1}B(\theta^0)\Sigma(\theta^0,\theta^0)^{-1},$$

the asymptotic variance of the estimator obtained from the simpler $M-$estimation? Using an example, we show in section 7 that there is no generally valid order relationship between those two matrices. Formally, we can easily imagine this if we remark that, for instance, for an unidimensional parameter the derivative of $\overline{\theta}(\theta^0, \lambda(\cdot))$ could be positive or negative. Intuitively, the full criterion $Q_T[\cdot, \lambda(\cdot)]$ could be 'more, or less informative' than the simplified criterion $Q_T[\cdot, \lambda(\theta^0)]$ and this may still be the case even when we adopt a two stages estimation strategy as in this paper.

6 The asymptotic distribution of the recursive $M-$estimators

The asymptotic normality of RM type estimators as considered in this paper could be obtained from Theorem 2 of Kushner and Huang (1979) (KH hereafter) or Theorem II.2.4 of KW which represents a version of the result obtained by KH[9]. Herein we

[9]For other references treating limit behavior of general RM procedures see Chen and White (1992) and the references therein.

rather use the latter result. Let us sketch, under a general form, the basic facts we use from KH and KW: let $M(\cdot) : \Delta \to \Delta$, $\Delta \subset I\!R^p$, $p \geq 1$, define a population problem with unique solution η^0. Consider $M_t(\eta)$, $t \geq 1$, a sequence of stationary[10] measurements of $M(\eta)$, $\eta \in \Delta$. Let $(\eta_T)_{T\geq 1}$ be an almost surely convergent sequence of estimators obtained from a RM procedure as in (4.2) with $a_t = ct^{-1}$, $c > 0$. Assume that the matrix

$$\bar{H} := c\,\frac{\partial M}{\partial \eta'}(\eta^0) + \frac{1}{2} I_p \tag{6.1}$$

is negative stable (that is the eigenvalues of $\bar{H}$ have negative real parts). It can be proved that, under some conditions, $\sqrt{T}(\eta_T - \eta^0)$ converges in distribution to a normal random variable with zero mean and variance matrix

$$c^2 \int_0^\infty \exp(\bar{H}u) R \exp(\bar{H}'u) du \tag{6.2}$$

where

$$R = \sum_{t=0}^{\infty} R_t + \sum_{t=1}^{\infty} R_t' \tag{6.3}$$

with

$$R_t = E\left[M_1(\eta^0) M_{t+1}(\eta^0)'\right], \quad t \geq 0.$$

A set of general conditions needed for the asymptotic normality of RM estimators is presented below (see Theorem II.2.4 of KW and their Assumptions B1 to B3 and B5). Let $(w_t)_{t\geq 0}$ be a stationary process with values in $\mathcal{W}$, some subset of an Euclidian space.

Assumption 6.1 *Let Δ a compact subset of $I\!R^p$, $p \geq 1$, and η^0 be an interior point of Δ. Consider*

$$M(\eta) = E\left[\psi\left(w_t\,;\,\eta\right)\right], \tag{6.4}$$

with $\psi(\cdot\,;\,\cdot)$ some function defined on $\mathcal{W} \times \Delta$ with values in $I\!R^p$, continuously differentiable w.r.t. η, such that the equation $M(\eta) = 0$ has as unique solution η^0. In addition the function ψ satisfies

$$E\left[\left\|\psi\left(w_t\,;\,\eta^0\right)\right\|^6\right] < \infty$$

[10] In this section, for the sake of a simpler presentation, we impose the stationarity of the measurements $M_t(\eta)$.

and

$$E\left[\left\|\frac{\partial\psi}{\partial\eta'}\left(w_t\,;\,\eta^0\right)\right\|_\lambda^2\right]<\infty.$$

Moreover, assume that there exist functions $\rho : \mathbb{R}_+ \to \mathbb{R}_+$ and $h : \mathcal{W} \to \mathbb{R}_+$ such that $\rho(u) \to 0$ as $u \to 0$, $h(\cdot)$ is measurable and for any η in an open neighborhood of η^0 and $w \in \mathcal{W}$

$$\left\|\frac{\partial\psi}{\partial\eta'}\left(w\,;\,\eta\right)-\frac{\partial\psi}{\partial\eta'}\left(w\,;\,\eta^0\right)\right\|_\lambda\leq\rho(\|\eta-\eta^0\|)h(w).$$

Assumption 6.2 *i) For some $\tau_1 \in \mathbb{R}$*

$$\sum_{t=1}^{T}\frac{1}{t}\left(h(w_{t-1})-\tau_1\right),$$

with h appearing in Assumption 6.1, converges $\boldsymbol{P}-$almost surely.

ii) For $\eta \in \Delta$, let

$$M_t(\eta)=\psi\left(w_{t-1}\,;\,\eta\right). \tag{6.5}$$

The series

$$\sum_{t=1}^{T}\frac{1}{t}\left(\frac{\partial M_t}{\partial\eta'}(\eta^0)-\frac{\partial M}{\partial\eta'}(\eta^0)\right)$$

and

$$\sum_{t=1}^{T}\frac{1}{t}\left(\left\|\frac{\partial M_t}{\partial\eta'}(\eta^0)\right\|_\lambda-E\left[\left\|\frac{\partial M_t}{\partial\eta'}(\eta^0)\right\|_\lambda\right]\right)$$

converge $\boldsymbol{P}-$almost surely.

Assumption 6.3 *i) Let $\kappa_t^2=E\left[\|E\left(M_t(\eta^0)\mid w_0\right)\|^2\right]$ with $M_t(\cdot)$ defined in (6.5). Then $\sum_{t\geq1}\kappa_t^{1/2}<\infty$.*

ii) Let R_j, $j \geq 0$, be defined as in (6.3) and

$$\xi_t^2=\sup_{j\geq0}E\left[\left\|E\left(M_t(\eta^0)M_{t+j}(\eta^0)'\mid w_0\right)-R_j\right\|^2\right].$$

Then $\sum_{t\geq1}\xi_t^{1/2}<\infty$.

KW show in their Theorem II.2.4 that the conclusions of Theorem 2 of KH can be obtained under the three assumptions above. Assumption 6.1 imposes some integrability conditions and puts some restrictions on the local behavior of $(\partial\psi/\partial\eta')(z,\eta)$. Assumption 6.3 asks that the dependence of the measurements $M_t(\eta^0)$ on the starting value w_0 becomes very weak as t grows to infinity (see KW, pp 45-46, for some examples). KW prove (see pp 63) that Assumption 6.3 and the integrability conditions on $\psi(\cdot,\cdot)$ ensure that R appearing in (6.2) is finite. Perhaps the most surprising is Assumption 6.2. Based on Theorem 2.2.2 of Kushner and Clark (1978), KH give some examples which show that this assumption is far less restrictive than it may seem to be (see also KW, pp 63, where it is proved that the Assumption 6.3 implies (A5) of KH).

When we consider the result of KW and KH for our RM estimators we have to ensure that the corresponding $\bar{H}$'s are negative stable. We will see from the proof of Proposition 6.1 that all we need is to ensure $\bar{H}$ negative stable for the simple RM estimator defined in (4.8). This condition depends on the Hessian matrix of the simple limit criterion and/or on the norm of the derivative of $\overline{\theta}(\theta^0,\cdot)$ (see (4.6)). If the real parts of the eigenvalues of $(\partial M/\partial\theta^{1'})(\theta^0)$, with $M(\cdot)$ as in (4.1), are negative but too close to zero in such way that $\bar{H}$ is not negative stable, a scaling factor $c>1$ in the learning rate may solve the problem. The choice of this factor could be done after a preliminary estimation of θ^0 (for instance using a simple RM procedure as in (4.8)) and of the eigenvalues of $(\partial M/\partial\theta^{1'})(\theta^0)$. We may also use a scaled learning rate for the other two RM procedures proposed in section 4.

For our RM estimators the function ψ from the assumptions of this section is as on the right hand side of (4.7). The matrix R defined in (6.3) is nothing else than the matrix $B(\theta^0)$ defined in Assumption 5.1. We provide now the asymptotic behavior of the three RM estimators we defined in subsection 4.

Proposition 6.1 *Assume that w_t, $t\geq 0$, defined in section 4 is stationary*[11]*. Moreover, assume that θ^0 is an interior point of Θ. Let $B(\theta^0)$ be the matrix defined in Assumption 5.1.*

i) Consider $M(\cdot)$ and $M_t(\cdot)$ be as in (4.1) and (4.7), respectively, and $c_1>0$ such that

$$\bar{H}_1 = c_1\,\frac{\partial M}{\partial\theta^{1'}}(\theta^0)+\frac{1}{2}\,I_p$$

[11]Some extensions to nonstationary observations can be considered following the arguments of Kushner and Huang (1981). However, for a simpler exposure we impose the stationarity condition.

is negative stable. Assume that θ_T^{RM}, $T \geq 1$, *is defined as in (4.8) with learning rate* $a_t = c_1 t^{-1}$, *and the conditions of Proposition 4.2 are fulfilled. If Assumptions 6.1 to 6.3 hold, then*

$$\sqrt{T}\,\left(\theta_T^{RM} - \theta^0\right) \xrightarrow{d} N_p(0, V_1(\theta^0))$$

with

$$V_1(\theta^0) = c_1^2 \int_0^\infty \exp(\bar{H}_1 u) B(\theta^0) \exp(\bar{H}_1' u) du. \tag{6.6}$$

ii) Consider $\Delta = U^M(\theta^0) \times \Theta \subset I\!R^{p^2} \times I\!R^p$ *where* $U^M(\theta^0)$ *is a small enough neighborhood of* $vec(\Sigma(\theta^0))$. *Let* $M(\cdot)' = (M_1^M(\cdot)', M_2^M(\cdot)')$ *be defined as in (4.14)-(4.15) and* $c_2 > 1/2$ *such that*

$$F_2 = F_2(\theta^0, c_2) = \left(\frac{1}{2} - c_2\right) I_p + c_2\, \frac{\partial \overline{\theta}}{\partial \theta^{1'}}(\theta^0, \theta^1)\bigg|_{\theta^1 = \theta^0}$$

is negative stable. Assume that θ_T^{MRM}, $T \geq 1$, *is defined as in (4.12) with learning rate* $a_t = c_2 t^{-1}$ *and the conditions of Proposition 4.3 are fulfilled. If Assumptions 6.1 to 6.3 hold, then*

$$\sqrt{T}\,\left(\theta_T^{MRM} - \theta^0\right) \xrightarrow{d} N_p(0, V_2(\theta^0))$$

with

$$V_2(\theta^0) = c_2^2 \int_0^\infty \exp\left(F_2 u\right) \Sigma(\theta^0, \theta^0)^{-1} B(\theta^0) \Sigma(\theta^0, \theta^0)^{-1} \exp(F_2' u)\, du \tag{6.7}$$

and $\Sigma(\theta^0, \theta^0)$ *as in Assumption 5.2.*

iii) Let $\Delta = U^Q(\theta^0) \times \Theta \subset I\!R \times I\!R^p$ *with* U^Q *is a small enough neighborhood of* $tr(\Sigma(\theta^0))$. *Let* $M(\cdot)' = (M_1^Q(\cdot), M_2^Q(\cdot)')$ *be defined as in (4.18)-(4.19) and* $c_3 > 1/2$ *such that*

$$F_3 = F_3(\theta^0, c_3) = -\frac{c_3}{tr(\Sigma(\theta^0))}\, \Sigma(\theta^0, \theta^0) \left(I_p - \frac{\partial \overline{\theta}}{\partial \theta^{1'}}(\theta^0, \theta^1)\bigg|_{\theta^1 = \theta^0}\right) + \frac{1}{2} I_p$$

is negative stable. Assume that θ_T^{QRM}, $T \geq 1$, *is defined as in (4.17) with learning rate* $a_t = c_3 t^{-1}$ *and the conditions of Proposition 4.4 are fulfilled. If Assumptions 6.1 to 6.3 hold, then*

$$\sqrt{T}\,\left(\theta_T^{QRM} - \theta^0\right) \xrightarrow{d} N_p(0, V_3(\theta^0))$$

with

$$V_3(\theta^0) = \frac{c_3^2}{tr(\Sigma(\theta^0, \theta^0))^2} \int_0^\infty \exp\left(F_3 u\right) B(\theta^0) \exp(F_3' u)\, du. \tag{6.8}$$

Before closing this section let us make some comments on the asymptotic variances of the RM estimators. First remark that it is difficult to make a general statement on how depend these variances on the constants c_1, c_2 and c_3. If we look at (6.2) we observe that a multiplicative constant in the learning rate does not only change the factor in front of the integral but also affects in a nonlinear way the integrand. After this remark we may immediately question the existence of an efficiency loss for the modified RM estimator due to the replacement of $-(\partial M/\partial\theta^{1'})(\theta^1)$ by a more feasible matrix in the definition of its population problem. If we think at the particular case of a derivative of $\overline{\theta}(\theta^0,\cdot)$ locally constant and equal to a multiple of the identity matrix, we realize that it is hard, if not impossible, to prove such an efficiency loss due to the simplification we considered.

On the other hand, when we may take $c_1 = c_2 = c_3$, one may expect the following increasing order in the efficiencies of the three RM estimators we considered: simple RM, 'quick and dirty' and modified RM. However, the asymptotic variances of these estimators depend intrinsicly on the derivative of $\overline{\theta}(\theta^0,\cdot)$ at θ^0. In our opinion this fact prevents us from obtaining generally valid relationships between the RM estimators we considered herein.

Finally, one may be willing to know which of the classes of estimators we presented in this paper, iterative or RM, is preferable from the asymptotic variance point of view. Again, the answer heavily depends on the form of the derivative of the function $\overline{\theta}(\theta^0,\cdot)$ and thus it seems that no general statement could be made. Empirical studies should complete the analysis of the methods we propose in this paper.

7 Example : a class of structural econometric models

Many structural econometric models (nonlinear rational expectations, option pricing, auction models, ...) characterize observable variables as highly nonlinear functions of some latent variables. These functions are one-to-one, but they depend on the unknown distribution of the latent variables through the equilibrium of the game or the learning process. Therefore numerical complexity of the equilibrium definition generates substantial obstacles for the direct implementation of maximum likelihood inference. Motivated by the fact that the law of motion of the latent variables is often defined in a fairly simpler way, simulation-based strategies (e.g., indirect inference) have been developed recently. The methodology we proposed above allows for alternative estimation procedures based on learning on the latent variables in order to perform approximated MLE directly inside the more tractable latent model. We briefly present these alternative estimation procedures below.

7.1 The latent statistical model

Let Y_t^*, $t = 0, 1, 2, ...,$ be a sequence of homogeneous Markovian, of order one, $J \times 1$ random vectors defined on some abstract probability space $(\Omega, \mathcal{F}, \boldsymbol{P})$ and taking values in some $\mathcal{Y}^* \subset \mathbb{R}^J$. We specify a $p-$dimensional parametric model for the conditional probability distribution of the process Y^*, given the initial value Y_0^*, through the family of transition densities

$$\mathcal{M}^* := \{f^*(\cdot \mid \cdot\ ; \theta),\ \ \theta \in \Theta \subset \mathbb{R}^p\} \tag{7.1}$$

defined with respect to the Lebesgue measure on $\mathbb{R}^J$. The model $\mathcal{M}^*$ is such that for any $\theta \in \Theta$ the transition $f^*(\cdot \mid \cdot\ ; \theta)$ allows for a unique stationary initial distribution. Moreover, we assume that the model we consider is correctly specified, that is for some $\theta^0 \in \Theta$, called the true unknown value of the parameters,

$$\prod_{t=1}^{T} f^*(y_t^* \mid y_{t-1}^*\ ; \theta^0)$$

is a correct description (in terms of densities) of the DGP providing the latent data $Y^*{}_1^T = (Y_t^*)_{1 \leq t \leq T}$, given that $Y_0^* = y_0^*$.

The maximum likelihood estimator in the latent (unobservable) world, denoted by $\hat{\theta}_T^*$, is defined as

$$\hat{\theta}_T^* = \arg\max_{\theta \in \Theta} \sum_{t=1}^{T} l^*(Y_t^* \mid Y_{t-1}^*\ ; \theta) \tag{7.2}$$

where, for any θ,

$$l^*(\cdot \mid \cdot\ ; \theta) := \log f^*(\cdot \mid \cdot\ ; \theta).$$

This estimator is not computable since we do not observe the process Y^*.

7.2 The observable statistical model

As it has been stressed above, the specificity of the framework we consider lies in the observation scheme. The observed data $Y_1^T = (Y_t)_{1\leq t\leq T}$ and the observed initial value Y_0 belong to some $\mathcal{Y} \subset I\!R^J$ and are given by a known one-to-one transformation $g(\cdot, \theta^0)$ (depending on the true unknown parameter[12]) of the latent data ${Y^*}_1^T$ and the latent initial value, respectively. That is, for any $t = 0, 1, 2, ...,$

$$Y_t = g(Y_t^*, \theta^0) \Longleftrightarrow Y_t^* = g^{-1}(Y_t, \theta^0). \tag{7.1}$$

In this framework, a maximum likelihood procedure in the observable world would maximize the criterion

$$\tilde{Q}_T[\theta, \lambda(\theta)] = \frac{1}{T}\sum_{t=1}^{T} l^*\left(g^{-1}(Y_t, \theta) \,|\, g^{-1}(Y_{t-1}, \theta)\ ; \theta\right) + \frac{1}{T}\sum_{t=1}^{T} \log\left|J_y g^{-1}(Y_t, \theta)\right|. \tag{7.2}$$

We denote by $|J_y g^{-1}(\cdot, \theta)|$ the absolute value of the Jacobian with respect to y of the one-to-one mapping $y \to g^{-1}(y, \theta)$. We have in mind the case where this criterion is very complicated. This happens, for instance, in the option pricing model considered by Christensen (1992). In such a framework, at any time t, Y_t^* represents a vector of unobservable state variables while Y_t is a vector of observed option prices. The relationship $g(\cdot, \theta^0)$ is provided by arbitrage-based derivative asset pricing à la Harrison and Kreps (1979). The observable model behind (7.2) was called by Christensen 'the empirical martingale model'. Renault and Touzi (1996) focused on the at-the-money (ATM) option case. They considered Y^* as being the (latent) volatility process of the underlying asset and Y_1^T a time series of prices of ATM European options (with fixed maturity period) written on this underlying asset. Moreover, they used the Hull and White (1987) option pricing formula for the passage $g(\cdot, \theta)$ between the two worlds, latent and observable. Typically, the dynamics of the latent process Y^* considered by Renault and Touzi can be described by a simple diffusion process (for instance the exponential of an Ornstein-Uhlenbeck process) in such a way that the latent log-likelihood is "nice". Meanwhile, their observable log-likelihood is "not nice", that is cumbersome to maximize, since it involves highly nonlinear functions of the unknown parameters. Facing such kind of problems, Renault and Touzi (1996) (see also Pastorello *et al.* (1993)) have proposed a two stage estimation procedure which represented the starting point of this paper.

[12]In this framework, the function $\lambda(\cdot)$ appearing in the previous sections is the identity function. A more general case can be also considered.

In the framework of this section, the criteria appearing in the previous sections become

$$Q_T[\theta, \lambda(\theta^1)] = Q_T[\theta, \theta^1] = \frac{1}{T}\sum_{t=1}^{T} l^* \left(g^{-1}(Y_t, \theta^1) \,|\, g^{-1}(Y_{t-1}, \theta^1) \;; \theta\right)$$

and

$$Q_\infty[\theta, \lambda(\theta^1)] = Q_\infty[\theta, \theta^1] = E_{\theta^0}\left[l^* \left(g^{-1}(Y_1, \theta^1) \,|\, g^{-1}(Y_0, \theta^1) \;; \theta\right)\right]$$

(in this section E_{θ^0} denotes the expectation considered w.r.t. the stationary distribution of (Y_1, Y_0), characterized by θ^0).

Let us notice that, despite what happens in usual partial observation schemes in econometrics (as those corresponding to Probit, Tobit,..., models), in our framework there is no automatic loss of information when passing from the latent to observable world. In order to justify this statement we recall that $g(\cdot, \theta^0)$ is a one-to-one transformation and we remark that the observable and latent models may swap their roles. Thus our estimation procedures could not be immediately be justified by a necessary richer information about θ^0 contained in the latent variables. However, we have in mind some structural models written in a simple and informative way in what concerns the structural parameters θ. Meanwhile, the observable likelihood is clearly more complicated, even intractable, and *intuitively* less informative. The example below reveals some information and identifiability aspects specific for our framework.

Example 1 (*About the loss of information and identifiability*)

Let us assume that the variables Y_t^*, $t = 1, ..., T$, are i.i.d., normal with mean θ and variance one. Assume that $Y_t = Y_t^* - \theta/2$ and remark that the Cramer-Rao (CR) bound in the observable world equals four while in the latent world it is equal to one. In other words, there is an information loss in estimating θ when passing from the latent world to the observable one.[13] Let us remark that in our framework we may even loose the whole information on (a part of) the parameters, that is (a part of) the parameters may become unidentifiable. To see this let us consider Y_t^*, $t = 1, ..., T$, as above and take $Y_t = Y_t^* - \theta$ which makes θ unidentifiable from the observables. We can also exemplify the identification problems in a dynamic model by extending the specification above to a AR(1) Gaussian process:

[13] Of course, if we swap the roles of the variables, that is if we consider Y_t^*, $t = 1, ..., T$, normally distributed with mean $\theta/2$ and variance one and $Y_t = Y_t^* + \theta/2$, we find a reverse result: the latent variables are less informative then the observable ones.

$$Y_t^* = \rho Y_{t-1}^* + \varepsilon_t, \quad t \in \mathbb{Z},$$

with $|\rho| < 1$ and ε_t, $t \in \mathbb{Z}$, i.i.d. $\mathcal{N}(0, \sigma^2)$ r.v.'s. In this case we remark that a transformation like $Y_t = Y_t^*/\sigma$ does not allow to identify σ, even though ρ is still identifiable. $\diamondsuit$

The previous example warns us that in our framework, due to (apparently innocent) one-to-one transformation, the structural parameters may easily become partially or totally unidentifiable in the observable world, even if they are identifiable in the latent world.[14] When the parameters are not identifiable in the observable world, it is clear that the two stage estimation strategy we adopt will fail to consistently estimate the true parameters.

Following the facts presented in section 2, let us denote

$$\begin{aligned} \overline{\theta}(\theta^0, \theta^1) &:= \arg\max_{\theta \in \Theta} E_{\theta^0} \left[l^* \left(g^{-1}(Y_1, \theta^1) \,|\, g^{-1}(Y_0, \theta^1) \,; \theta \right) \right], \qquad (7.3) \\ &= \arg\max_{\theta \in \Theta} E_{\theta^0} \left[l^* \left(g^{-1}(Y_1, \theta^1) \,|\, g^{-1}(Y_0, \theta^1) \,; \theta \right) + \log \left| J_y g^{-1}(Y_1, \theta^1) \right| \right]. \end{aligned}$$

Recall that the basic identification condition in the latent model $\mathcal{M}^*$ can be rewritten as

$$\overline{\theta}(\theta^0, \theta^0) = \theta^0.$$

Moreover, the strengthened identification condition for our framework where we consider a two stage estimation strategy was that the unique fixed point of $\overline{\theta}(\theta^0, \cdot)$ is θ^0 (see Assumption 2.2). Clearly, the reinforcement of the basic identification condition, is stronger than the identification in the observable world. The remark above on the possible loss of indentifiability when passing from one world to the other gives an idea about how restrictive our reinforced identification condition could be.

The intuitions we sketched above and in section 2, leading us to state Assumption 2.2, may raise the question if the unique fixed point property for the function $\overline{\theta}(\theta^0, \cdot)$ is necessary for ensuring the identification in the observable model

[14] If we swap the roles of Y and Y^* we can also deduce from Example 1 above that identification could hold in the observable model without necessarily being ensured in the latent one. Due to the fact that herein we favor the latent world where we build our structural model, we will always look for identification and consistency assumptions for the observable model starting from assumptions on the latent one.

$$\mathcal{M} = \left\{ f^* \left(g^{-1}(\cdot,\theta) \,|\, g^{-1}(\cdot,\theta) \,;\theta \right) \left| J_y g^{-1}(\cdot,\theta) \right|, \;\; \theta \in \Theta \right\}.$$

The counterexample below shows that the answer is negative.

Example 2 (*Counterexample*)

Let us consider a sequence of i.i.d. r.v.'s $Y^* = (Y_t^*)_{t\geq 1}$ and a family of exponential probability density functions

$$f^*(y_t^*;\theta) = \theta \exp(-\theta y_t^*)\, \mathbb{1}_{(0,\infty)}(y_t^*),$$

with $\theta \in \Theta = [1/2,\, 3/2]$. We assume that the distribution of Y_1^* is given by $f^*(\cdot;\theta^0)$, for some $\theta^0 \in \Theta$. The observable variables are obtained via the transformation

$$Y_t = g(Y_t^*, \theta^0) = Y_t^* + \theta^0.$$

In this case

$$E_{\theta^0} \left[l^* \left(g^{-1}(Y_1,\theta^1);\theta \right) \right] = \log\theta - \theta \left(\frac{1}{\theta^0} + \theta^0 - \theta^1 \right),$$

and

$$\overline{\theta}(\theta^0,\theta^1) = \left(\frac{1}{\theta^0} + \theta^0 - \theta^1 \right)^{-1}.$$

As θ^0 and $1/\theta^0$ are both fixed points for $\overline{\theta}(\theta^0,\cdot)$, we obtain that Assumption 2.2 is violated, provided that $\theta^0 \neq 1$. However, θ^0 is clearly identifiable in the observable world since the support of the variables $(Y_t)_{t\geq 1}$ is $[\theta^0,\infty)$. $\diamondsuit$

The previous example shows that it may happen that the two stages strategy we adopt in this paper cannot be applied, while the ML estimation based on the full criterion (7.2) could, *theoretically*, provide consistent estimation. However, we argue that there exists an important class of structural econometric models where the ML estimator based on the full criterion is not computable. Meanwhile, the two stages strategy we considered herein applies. For example, Renault and Touzi (1996) reported that the unique fixed point property of $\overline{\theta}(\theta^0,\cdot)$ seems to be verified in the Hull and White (1987) option pricing model.

7.3 Iterative and recursive estimators

In the previous subsection we stressed our preference for the latent world and, guided by our choice, we gave additional motivations for the general identification condition written in section 2. Now remark that in this framework, definition (3.1) becomes

$$\overline{\theta}_T(\theta^1) = \arg\max_{\theta\in\Theta} \frac{1}{T}\sum_{t=1}^{T} l^*\left(g^{-1}(Y_t,\theta^1)\,|\,g^{-1}(Y_{t-1},\theta^1)\,;\theta\right), \qquad \theta^1 \in \Theta. \tag{7.1}$$

The motivation for this definition is that, once we have in hand a proxy value of θ^0, we will always look for bringing the observations in the latent world where the structural model is defined. The estimation strategy of transforming data via an approximating value of θ^0 and performing ML estimation in the latent model $\mathcal{M}^*$ has some common features with the strategy of the EM algorithms introduced by Dempster *et al.* (1977). Thus we can consider that the argmax estimators obtained using $\overline{\theta}_T(\cdot)$ are EM type estimators. However, as it has been remarked by Renault and Touzi (1996), the general theory of EM algorithms can not be used in this framework since, for any t, the distribution of Y_t^* given Y_t is a degenerate one the support of which depends on the unknown value of the parameters.

We may remark that $\overline{\theta}_T(\theta^0)$ coincides with the MLE in the latent model! Using $\overline{\theta}_T(\cdot)$ written in (7.1) we may rewrite the two estimators of section 3 (see (3.2) and (3.3)).

Note that in the framework considered in this section the matrix $B(\theta^0)$ defined in Assumption 5.1 represents the inverse of the Fisher information in the latent model and coincides with the matrix $\Sigma(\theta^0,\theta^0)$ of Assumption 5.2. Proposition 5.1 tells that, under some conditions, the Renault and Touzi (1996) estimator and the iterative estimator presented in section 3 are asymptotically normal with the same asymptotic variance

$$V(\theta^0) = \left(I_p - \frac{\partial\overline{\theta}}{\partial\theta^{1\prime}}(\theta^0,\theta^1)\bigg|_{\theta^1=\theta^0}\right)^{-1} \Sigma(\theta^0,\theta^0)^{-1} \left(I_p - \frac{\partial\overline{\theta}\,'}{\partial\theta^{1}}(\theta^0,\theta^1)\bigg|_{\theta^1=\theta^0}\right)^{-1}.$$

In the limit case where, for any $\theta^1\in\Theta$, $g^{-1}(\cdot,\theta^1)$ is the identity function, that is when one observes directly the latent variables, as expected, $V(\theta^0)$ becomes equal to $B(\theta^0)^{-1}$, the CR bound in the latent model. Let us recall that in subsection 7.2 we have remarked that in our framework there is no general rule stating that the latent world should reveal more information than the observable one does. On the other hand, it is clear that $V(\theta^0)$ should be greater than the asymptotic CR bound corresponding to the observable model $\mathcal{M}$. In other words, it is obvious that the iterative estimator given in (3.2) should be asymptotically less concentrated than the intractable MLE obtained from the observable log-likelihood (7.2). It remains a natural question : is $V(\theta^0)$ greater or smaller than $B(\theta^0)^{-1}$? In the example below we show that there is no generally valid order relationship between these two matrices.

Example 3

For the sake of simplicity, let us consider, as in Example 2, that the variables Y_t^*, $t = 1, ..., T$, are i.i.d. as normals of mean θ and variance 1. Moreover, $Y_t = Y_t^* + \theta/2$, thus $g^{-1}(y, \theta) = y - \theta/2$. In this case $\overline{\theta}(\theta^0, \theta^1) = 3\theta^0/2 - \theta^1/2$ and $(d\overline{\theta}/d\theta^1)(\theta^0, \theta^1) = -1/2$, for any $\theta^1 \in \Theta$. As a consequence $V(\theta^0) = 4/9$ and we remark that $V(\theta^0)$ coincides with the CR bound in the observable world, that is the two argmax estimators mentioned above are asymptotically equivalent with the MLE obtained from the observable log-likelihood. Moreover, $V(\theta^0)$ is smaller than $B(\theta^0)^{-1}$ which equals one. Swapping the roles of the variables (as we have done in Example 2) we obtain a reverse conclusion: $B(\theta^0)^{-1}$ is smaller than $V(\theta^0)$. It is clear that there exists a close link between the form of the matrix

$$I_p - \frac{\partial \overline{\theta}}{\partial \theta^{1\prime}}(\theta^0, \theta^1) \Big|_{\theta^1 = \theta^0}$$

and the relationship between the information bounds in $\mathcal{M}^*$ and $\mathcal{M}$. This relationship is to be examined in future work. $\diamondsuit$

Despite what we may learn from Example 3, as we have stressed from the beginning of this section, we have in mind latent variables which we believe containing more information on the parameters than their transformations. As a consequence we expect a loss of efficiency for our iterative estimator compared with the (infeasible) latent MLE written in (7.2). More contracting the function $\overline{\theta}(\theta^0, \cdot)$ is, smaller this loss (expressed as a part of CR bound in the latent world) is expected to be.

For the Robbins-Monro type estimators the $M(\cdot)$ function defined in (4.1) becomes

$$M(\theta^1) = E_{\theta^0}\left[\frac{\partial l^*}{\partial \theta}\left(g^{-1}(Y_1, \theta^1) \,|\, g^{-1}(Y_0, \theta^1) \,;\, \theta\right)\right] \Big|_{\theta = \theta^1},$$

while the counterpart of $M_t(\cdot)$ defined in (4.7) is

$$M_t(\theta^1) = \frac{\partial l^*}{\partial \theta} \left(g^{-1}(Y_t, \theta^1) \,|\, g^{-1}(Y_{t-1}, \theta^1) \; ; \; \theta \right) \Big|_{\theta=\theta^1} .$$

The matrix $G_{t+1}(\cdot)$ defined in (4.13) becomes

$$G_{t+1}(\theta^1) = -\frac{1}{t} \sum_{s=1}^{t} \frac{\partial^2 l^*}{\partial \theta \partial \theta'} \left(g^{-1}(Y_s, \theta^1) \,|\, g^{-1}(Y_{s-1}, \theta^1) \; ; \; \theta \right) \Big|_{\theta=\theta^1} .$$

A perhaps easier modified RM procedure is obtained for

$$\begin{aligned} \tilde{G}_{t+1}(\theta^1) \quad = \quad & -\frac{1}{t} \sum_{s=1}^{t} \Bigg[\frac{\partial l^*}{\partial \theta} \left(g^{-1}(Y_s, \theta^1) \,|\, g^{-1}(Y_{s-1}, \theta^1) \; ; \; \theta \right) \Big|_{\theta=\theta^1} \\ \times \quad & \frac{\partial l^*}{\partial \theta'} \left(g^{-1}(Y_s, \theta^1) \,|\, g^{-1}(Y_{s-1}, \theta^1) \; ; \; \theta \right) \Big|_{\theta=\theta^1} \Bigg] . \end{aligned}$$

The asymptotic results for the modified RM procedures defined with $G_{t+1}(\cdot)$ or $\tilde{G}_{t+1}(\cdot)$ are the same. The same remarks remain valid for the 'quick and dirty' RM procedure. In (4.17) we may use the trace of $G_{t+1}(\cdot)$ or that of $\tilde{G}_{t+1}(\cdot)$ without changing the asymptotics.

As it has been mentioned at the end of section 6, it seems very difficult to compare the different recursive estimators we proposed above. Moreover, it seems very difficult to compare their asymptotic variances with that of the iterative estimator we propose herein. Intuitively, the recursive estimators should be less precise than the iterative one. This issue is to be studied in future work.

References

[1] **Chen, X. and H. White (1992)**, "Asymptotic Properties of Some Projection-based Robbins-Monro Procedures in a Hilbert Space", UCSD Department of Economics Discussion Paper.

[2] **Chen, X. and H. White (1995)**, "Nonparametric Adaptive Learning with Feedback", UCSD Department of Economics Discussion Paper.

[3] **Christensen, B.J. (1992)**, "Asset Prices and the Empirical Martingale Model", Working Paper, New-York University.

[4] **Clark, D. (1984)**, "Necessary and Sufficient Conditions for the Robbins-Monro Method", *Stochastic Processes and their Applications*, 17, 359-367.

[5] **Davidson, J (1994)** *Stochastic Limit Theory*, Advanced Texts in Econometrics, Oxford Univ. Press.

[6] **Hannan, E.J. and J. Rissanen (1982)** Recursive estimation of mixed autoregressive-moving average order, *Biometrika*, 69, 1, pp 81-94.

[7] **Harrison, J. and D. Kreps (1979)**, "Martingales and Arbitrage in Multiperiod Security Markets", *Journal of Economic Theory*, 20, 381-408.

[8] **Horn, R. and C. Johnson (1985)**, *Matrix Analysis*, Cambridge University Press.

[9] **Horn, R. and C. Johnson (1991)**, *Topics in Matrix Analysis*, Cambridge University Press.

[10] **Hull, J. and White A. (1987)**, "The Pricing of Options on Assets with Stochastic Volatilities", *Journal of Finance*, 42, 281-300.

[11] **Kuan, C.M. and H. White (1994a)**, "Artificial Neural Networks: an Econometric Perspective", *Econometric Reviews*, 13(1), 1-91.

[12] **Kuan, C.M. and H. White (1994b)**, "Adaptive Learning with Nonlinear Dynamics Driven by Dependent Processes", *Econometrica*, 62, 1087-1114.

[13] **Kushner, H.J. and D. Clark (1978)**, *Stochastic Approximation Methods for Constrained and Unconstrained Systems*, Springer-Verlag, New-York.

[14] **Kushner, H.J. and H. Huang (1979)**, "Rates of convergence for stochastic approximation type algorithms", *SIAM Journal of Control and Optimization*, 17, No. 5, 607-617.

[15] **Kushner, H.J. and H. Huang (1981)** "On the weak convergence of a sequence of general stochastic difference equations to a diffusion", *SIAM Journal of Applied Mathematics*, 40, No. 3, 528-541.

[16] **Ljung, L. (1977)**, "Analysis of Recursive Stochastic Algorithms", *IEEE Transactions on Automatic Control*, AC-22, 551-575.

[17] **Pastorello S., E. Renault and N. Touzi (1993)**, "Statistical Inference for Random Variance Option Pricing", Southern European Economics Discussion Series.

[18] **Patilea, V (1997)** *Ph.D. dissertation*, Institute of Statistics, Louvain-la-Neuve.

[19] **Patilea, V., M.P. Ravoteur and E. Renault (1995)**, "Multivariate Time Series Analysis of Option Prices", Mimeo, GREMAQ, Toulouse.

[20] **Renault, E. (1995)**, "Econometric Models of Option Pricing Errors", *Proceedings of the* 7^{th} *WCES*,

[21] **Renault, E and N. Touzi (1996)**, "Option Hedging and Implied Volatilities in a Stochastic Volatility Model", *Mathematical Finance*, Vol. 6, No. 3 , 279-302.

[22] **Robbins, H. and S. Monro (1951)**, "A Stochastic Approximation Method", *Annals of Mathematical Statistics*, 22, 400-407.

[23] **Rouche, N. and J. Mawhin (1980)**, *Ordinary Differential Equations*, Pitman Advanced Publishing Program.

[24] **Sargent, T. (1993)**, *Bounded Rationality in Macroeconomics*, Oxford: Clarendon Press.

[25] **White, H. (1989)**, "Some Asymptotic Results for Learning in Single Hidden-Layer Feedforward Network Models", *Journal of the American Statistical Association*, Vol. 84, No. 408, 1003-1013.

[26] **Wooldridge, J. (1994)**, "Estimation and Inference for Dependent Processes", *Handbook of Econometrics, Vol IV*, R.F. Engle and D. McFadden eds., 2641-2739.

[27] **Yin, G. and Y.M. Zhu (1990)**, "On H-Valued Robbins-Monro Process", *Journal of Multivariate Analysis*, 34, 116-140.

[28] **Young, P (1985)** Recursive Identification, Estimation and Control, *Handbook of Statistics, Vol. 5*, Elsevier Science Publishers.

Appendix

Proof of Proposition 3.1

The result is true under the stated assumptions. However, for a simpler presentation, herein we prove it under the assumption that there exists $k \in (0,1)$ such that

$$d_1\left(\overline{\theta}(\theta^0,\lambda),\overline{\theta}(\theta^0,\lambda')\right) \leq k\, d_2(\lambda,\lambda'), \quad \lambda,\lambda' \in \Gamma. \tag{A.1}$$

It is well-known that $X_T \to X$, in probability, iff every subsequence X_{T_k}, $k \geq 1$, contains a further subsequence $X_{T_{k(j)}}$, $j \geq 1$, which converges almost surely to X (see, e.g., Davidson (1994), Theorem 18.6). Therefore, it suffices to prove that

$$\sup_{(\theta,\lambda)\in\Theta\times\Gamma} |Q_T(\theta,\lambda) - Q_\infty(\theta,\lambda)| \longrightarrow 0, \quad \boldsymbol{P}-\text{almost surely}, \tag{A.2}$$

implies

$$\sup_{\lambda\in\Gamma} d_1\left(\overline{\theta}_T(\lambda),\overline{\theta}_\infty(\theta^0,\lambda)\right) \longrightarrow 0, \quad \boldsymbol{P}-\text{almost surely}. \tag{A.3}$$

Using standard arguments, for any fixed $\lambda \in \Gamma$, from the almost sure uniform (with respect to θ) convergence of $Q_T(\theta,\lambda)$ we deduce $\overline{\theta}_T(\lambda) \to \overline{\theta}_\infty(\theta^0,\lambda)$, $\boldsymbol{P}-$almost surely. In order to obtain (A.3) it remains to prove that $\overline{\theta}_T(\lambda)$, $T \geq 1$, is strongly asymptotically uniformly stochastically equicontinuous (see, e.g., Theorem 21.8 of Davidson (1994)). Let $N \subset \Omega$ denote the exceptional set on which the convergence in (A.2) fails. For any $\omega \in N^c$, we show that $\forall\varepsilon > 0$, $\exists\delta > 0$ and $\exists T_0 \geq 1$ such that $\forall T \geq T_0$,

$$d_2(\lambda,\lambda') < \delta \quad \Longrightarrow \quad d_1\left(\overline{\theta}_T(\omega,\lambda),\,\overline{\theta}_T(\omega,\lambda')\right) < \varepsilon.$$

Assume that this is not true, i.e., there exists $\varepsilon > 0$ such that $\forall\delta > 0$ and $\forall T_0 \geq 1$, $\exists T \geq T_0$ and $\exists\lambda,\lambda' \in \Gamma$ with $d_2(\lambda,\lambda') < \delta$ and

$$d_1\left(\overline{\theta}_T(\omega,\lambda),\,\overline{\theta}_T(\omega,\lambda)\right) \geq \varepsilon.$$

Taking $\delta = \varepsilon/2$ we can build a subsequence T_i, $i \geq 1$, such that, for any i, there exists $\lambda_i, \lambda_i' \in \Gamma$ with $d_2(\lambda_i,\lambda_i') < \varepsilon/2$ and

$$d_1\left(\overline{\theta}_{T_i}(\omega,\lambda_i),\,\overline{\theta}_{T_i}(\omega,\lambda_i')\right) \geq \varepsilon. \tag{A.4}$$

Each of the sequences $(\lambda_i)_{i\geq 1}$ and $(\lambda_i')_{i\geq 1}$ contain convergent subsequences. Say, without loss of generality, that $\lambda_i \to \lambda$ and $\lambda_i' \to \lambda'$ with $\lambda,\lambda' \in \Gamma$. Necessarily,

$d_2(\lambda_i, \lambda_i') \le \varepsilon/2$. On the other hand, $\overline{\theta}_{T_i}(\omega, \lambda_i)$, $i \ge 1$, contains a convergent subsequence $\overline{\theta}_{T_{i(j)}}(\omega, \lambda_{i(j)})$, $j \ge 1$, with limit, say, $\theta_\infty \in \Theta$. For any $\theta \in \Theta$ we may write

$$Q_{T_{i(j)}}\left(\overline{\theta}_{T_{i(j)}}(\omega, \lambda_{i(j)}),\ \lambda_{i(j)}\right) \ge Q_{T_{i(j)}}\left(\theta,\ \lambda_{i(j)}\right),$$

and the uniform convergence assumption gives

$$Q_\infty(\theta_\infty, \lambda) \ge Q_\infty(\theta, \lambda).$$

Then, necessarily, $\theta_\infty = \overline{\theta}(\theta^0, \lambda)$. With the same arguments we obtain that there exists a subsequence of $\overline{\theta}_{T_i}(\omega, \lambda_i')$, $i \ge 1$, converging to $\overline{\theta}(\theta^0, \lambda')$. From (A.4) we deduce $d_1\left(\overline{\theta}(\theta^0, \lambda), \overline{\theta}(\theta^0, \lambda')\right) \ge \varepsilon$ while $d_2(\lambda, \lambda') \le \varepsilon/2$. As this is in contradiction with A.1 we may conclude the proof. ■

Proof of Proposition 3.2

For simpler notation we write p instead of $p(T)$. From the contracting mapping property we get, for any $T \ge 1$,

$$\begin{aligned} d_1\left(\theta_T,\, \theta^0\right) &\le d_1\left(\overline{\theta}_T\left(\lambda(\theta_T^{(p)})\right),\, \overline{\theta}\left(\theta^0, \lambda(\theta_T^{(p)})\right)\right) + d_1\left(\overline{\theta}\left(\theta^0, \lambda(\theta_T^{(p)})\right),\, \overline{\theta}(\theta^0, \lambda(\theta^0))\right) \\ &\le \sup_{\theta^1 \in \Theta} d_1\left(\overline{\theta}_T(\lambda(\theta^1)),\, \overline{\theta}(\theta^0, \lambda(\theta^1))\right) + k\, d_1\left(\theta_T^{(p)},\, \theta^0\right) \\ &\le \left(1 + \ldots + k^{p-1}\right) \sup_{\theta^1 \in \Theta} d_1\left(\overline{\theta}_T(\lambda(\theta^1)) - \overline{\theta}(\theta^0, \lambda(\theta^1))\right) + k^p d_1\left(\theta_T^{(1)} - \theta^0\right). \end{aligned}$$

Finally, we may restate Proposition 3.1 for $\overline{\theta}(\theta^0, \lambda(\cdot)) : \Theta \to \Theta$, $Q_T[\cdot, \lambda(\cdot)]$ and $Q_\infty[\cdot, \lambda(\cdot)]$ and we obtain the uniform convergence of $\overline{\theta}_T(\lambda(\cdot))$. Recalling that $k \in (0, 1)$, $p = p(T) \to \infty$ and Θ bounded set, we may close the proof. ■

Lemma A.1 *Let $(X_T)_{T \ge 1}$, $(A_T)_{T \ge 1}$ be two sequences of nonnegative, r.v.'s. Assume that $E(A_T) \to 0$, as $T \to \infty$, and that there exists $T_0 \ge 1$ such that*

$$E(X_{T_0}) < \infty. \tag{A.5}$$

If, for any $T \ge T_0$,

$$X_{T+1} \le A_T + k X_T, \tag{A.6}$$

for some $k \in (0, 1)$, then $X_T \to 0$, in probability.

Proof For simplicity we assume $T_0 = 1$. Let us first remark that the result is quite simple for real numbers. That is, if $(x_T)_{T\geq 1}$ and $(a_T)_{T\geq 1}$ are two sequences of nonnegative real numbers such that $a_T \to 0$ and , for any $T \geq 1$,

$$x_{T+1} \leq a_T + k x_T,$$

with $k \in (0,1)$, then $x_T \to 0$. To see this fix $\varepsilon > 0$ arbitrary. There exists $T_1 = T_1(\varepsilon)$ such that, $\forall T \geq T_1$, $a_T \leq \varepsilon$. Then, for any $S \geq 1$,

$$x_{T_1+S} \leq \varepsilon(1 + ... + k^{S-1}) + k^S x_{T_1}.$$

We deduce that

$$\limsup_{T\to\infty} x_T = \limsup_{S\to\infty} x_{T_1+S} \leq \varepsilon(1-k)^{-1}.$$

Now we take the expectations in (A.6) and we get $E(X_T) < \infty$, for any $T \geq 1$ and $E(X_T) \to 0$, which in turn implies $X_T \to 0$ in probability. ■

Proof of Proposition 3.3

For any $T > t_0$ we can write

$$d_1\left(\hat{\theta}_{T+1},\, \theta^0\right) = d_1\left(\overline{\theta}_{T+1}\left(\lambda(\hat{\theta}_T)\right),\, \overline{\theta}\left(\theta^0, \lambda(\theta^0)\right)\right) \tag{A.7}$$

$$\leq d_1\left(\overline{\theta}_{T+1}\left(\lambda(\hat{\theta}_T)\right),\, \overline{\theta}\left(\theta^0, \lambda(\hat{\theta}_T)\right)\right) + d_1\left(\overline{\theta}\left(\theta^0, \lambda(\hat{\theta}_T)\right), \overline{\theta}\left(\theta^0, \lambda(\theta^0)\right)\right)$$

$$\leq \sup_{\theta^1\in\Theta} d_1\left(\overline{\theta}_{T+1}\left(\lambda(\theta^1)\right),\, \overline{\theta}\left(\theta^0, \lambda(\theta^1)\right)\right) + k\, d_1\left(\hat{\theta}_T,\, \theta^0\right).$$

Apply Lemma A.1 *i)* for $A_T = \sup_{\theta^1\in\Theta} d_1\left(\overline{\theta}_{T+1}(\lambda(\theta^1)),\, \overline{\theta}(\theta^0, \lambda(\theta^1))\right)$ and $X_T = d_1\left(\hat{\theta}_T,\, \theta^0\right)$. The needed hypotheses are ensured by Proposition 3.1 and the boundedness of Θ. ■

Proof of Proposition 4.2 First of all let us recall a classic result used in the theory of (Lyapunov) stability in ODE (see, e.g., Rouche and Mahwin (1980), ch. 1): if $-A$ is a positive stable matrix, to each negative definite matrix U there corresponds one and only one positive definite matrix Q such that

$$\langle\, Q\,\theta,\; A\,\theta\,\rangle = \theta'\,U\,\theta. \tag{A.8}$$

(by $\langle \cdot, \cdot \rangle$ we denote the inner product in $I\!R^p$). For any $\theta^1 \in \Theta$, we can write

$$M(\theta^1) = \frac{\partial M}{\partial \theta^{1'}}(\theta^0)\,(\theta^1 - \theta^0) + o(\|\theta^1 - \theta^0\|).$$

Assumption 4.1 ensures that the matrix $-(\partial M/\partial \theta^{1'})(\theta^0)$ is positive stable. As a consequence, there exists Q positive definite such that, for any $\theta^1 \neq \theta^0$,

$$\left\langle Q(\theta^1 - \theta^0),\ \frac{\partial M}{\partial \theta^{1'}}(\theta^0)\,(\theta^1 - \theta^0) \right\rangle = -\|\theta^1 - \theta^0\|^2.$$

Define now

$$V(\theta^1) = \frac{1}{2}\left\langle \theta^1 - \theta^0,\ Q(\theta^1 - \theta^0) \right\rangle + f(\|\theta^1 - \theta^0\|^2), \tag{A.9}$$

with $f(\cdot)$ a real function specified below. Then, $\forall\, \theta^1 \in \Theta$,

$$\begin{aligned}\left\langle \frac{\partial V}{\partial \theta^{1'}}(\theta^1),\ M(\theta^1) \right\rangle &= \left\langle Q(\theta^1 - \theta^0),\ \frac{\partial M}{\partial \theta^{1'}}(\theta^0)\,(\theta^1 - \theta^0) \right\rangle \\ &\quad + 2f'(\|\theta^1 - \theta^0\|^2)\left\langle \theta^1 - \theta^0,\ \frac{\partial M}{\partial \theta^{1'}}(\theta^0)\,(\theta^1 - \theta^0) \right\rangle + o(\|\theta^1 - \theta^0\|^2) \\ &\quad + 2f'(\|\theta^1 - \theta^0\|^2)\, o(\|\theta^1 - \theta^0\|^2).\end{aligned}$$

Take $f(\cdot)$ differentiable such that $f(0) = 0$, $V(\theta^1) > 0$, $\forall \theta^1 \neq \theta^0$, and the sum of the last three terms on the r.h.s. of the last inequality are non positive. The function V defined in (A.9) is usually called a Lyapunov function (see, e.g., condition (A3) of Yin and Zhu (1990)). Now we can proceed as in Yin and Zhu (1990) or Chen and White (1992). ■

Proof of Proposition 4.3 As in the proof of Proposition 4.2, first we have to find a suitable Lyapunov function and afterwards we may invoke the results of Yin and Zhu (1990) or Chen and White (1992). We consider

$$V_1(\delta) = V_1(vec(G)) = \frac{1}{2}\left\| vec(\Sigma(\theta^0) - G)\right\|^2,$$

with $\Sigma(\cdot)$ as in (4.11). We have

$$\left\langle \frac{\partial V_1}{\partial vec(G)}(\delta),\ M_1^M(\delta) \right\rangle = -\left\langle vec(\Sigma(\theta^0) - G),\ vec(\Sigma(\theta^1) - G) \right\rangle$$

$$= -\|vec(\Sigma(\theta^0) - G)\|^2 - \left\langle vec(\Sigma(\theta^0) - G),\ \frac{\partial vec(\Sigma)}{\partial \theta^{1\prime}}(\theta^0)(\theta^1 - \theta^0)\right\rangle$$

$$- \left\langle vec(\Sigma(\theta^0) - G),\ h\right\rangle,$$

where h is the rest of the Taylor expansion of $vec(\Sigma(\cdot))$ around θ^0. As a consequence there exists $C_1 > 0$ such that, for any $\theta^1 \in \Theta$

$$\left\langle \frac{\partial V_1}{\partial vec(G)}(\delta),\ M_1^M(\delta)\right\rangle \leq -\|vec(\Sigma(\theta^0) - G)\|^2 + C_1\|vec(\Sigma(\theta^0) - G)\|\,\|\theta^1 - \theta^0\|.$$

On the other hand, we have

$$\Sigma^{-1}(\theta^0)\frac{\partial M}{\partial \theta^{1\prime}}(\theta^0) = I_p - \frac{\partial \overline{\theta}}{\partial \theta^{1\prime}}(\theta^0, \theta^1)\bigg|_{\theta^1 = \theta^0}, \tag{A.10}$$

with $M(\cdot)$ as in (4.1). Under Assumption 4.1 we may apply Lemma 4.1 for $A_1 = I_p$. We deduce that the matrix in (A.10) is positive stable. For easier writings let us define

$$M_2^0(\theta^1) = M_2^M((vec(\Sigma(\theta^0))', \theta^{1\prime})')$$

and remark that, under the stated assumptions, $-(\partial M_2^0/\partial \theta^{1\prime})(\theta^0)$ coincides with the matrix written in (A.10). If in (A.8) we take $U = -I_p$, there exists Q_2 positive definite such that, for any $\theta^1 \in \Theta$,

$$\left\langle Q_2(\theta^1 - \theta^0),\ \frac{\partial M_2^0}{\partial \theta^{1\prime}}(\theta^0)\,(\theta^1 - \theta^0)\right\rangle = -\|\theta^1 - \theta^0\|^2.$$

Let us remark that, for any G,

$$M_2^M((vec(G)', \theta^{0\prime})') = 0,$$

Consequently, from a Taylor expansion we may deduce, for δ in some (bounded) neighborhood of $\delta^0 = (vec(\Sigma(\theta^0)', \theta^{0\prime})'$, that

$$M_2^M(\delta) = \frac{\partial M_2^0}{\partial \theta^{1\prime}}(\theta^0)\,(\theta^1 - \theta^0) + r_2 + o(\|\theta^1 - \theta^0\|),$$

with $\|r_2\| \leq C'\|vec(\Sigma(\theta^0) - G)\|\|\theta^1 - \theta^0\|$ and C' some positive constant. Moreover, we have

$$\left\langle Q_2(\theta^1-\theta^0),\ M_2^M(\delta)\right\rangle = -\|\theta^1-\theta^0\|^2 + \left\langle Q_2(\theta^1-\theta^0),\ r_2\right\rangle + o(\|\theta^1-\theta^0\|^2).$$

Now we may take as Lyapunov function

$$V(\delta) = V_1(vec(G)) + \frac{C_2}{2}\left\langle \theta^1-\theta^0,\ Q_2(\theta^1-\theta^0)\right\rangle + g(\|\theta^1-\theta^0\|^2), \tag{A.11}$$

with $C_2 > 0$ and $g(\cdot)$ to be specified and V_1, Q_2 as above. We have

$$\left\langle \frac{\partial V}{\partial \delta}(\delta),\ (M_1^M(\delta), M_2^M(\delta))\right\rangle$$

$$\leq -\|vec(\Sigma(\theta^0)-G)\|^2 + C_1\|vec(\Sigma(\theta^0)-G)\|\,\|\theta^1-\theta^0\| - C_2\|\theta^1-\theta^0\|^2$$

$$+ C_2\left\langle Q_2(\theta^1-\theta^0),\ r_2\right\rangle + o(\|\theta^1-\theta^0\|^2) + r$$

where r is a rest in which appears the derivative of the function $g(\cdot)$. For C_2 a suitably chosen positive constant we have

$$-\|vec(\Sigma(\theta^0)-G)\|^2 + C_1\|vec(\Sigma(\theta^0)-G)\|\,\|\theta^1-\theta^0\| - \frac{C_2}{2}\|\theta^1-\theta^0\|^2 < 0,$$

for any $\theta^1 \neq \theta^0$. Moreover, for any choice of $C_2 > 0$, considering a small enough neighborhood of $vec(\Sigma(\theta^0))$ and a suitable choice of $g(\cdot)$, we may assume that, for any $\theta^1 \in \Theta$,

$$-\frac{C_2}{2}\|\theta^1-\theta^0\|^2 + C_2\left\langle Q_2(\theta^1-\theta^0),\ r_2\right\rangle + o(\|\theta^1-\theta^0\|^2) + r \leq 0.$$

As a consequence, after good choices, we may obtain, for any $\delta \neq \delta^0$, $V(\delta) > 0$,

$$\left\langle \frac{\partial V}{\partial \delta}(\delta),\ (M_1^M(\delta), M_2^M(\delta))\right\rangle < 0,$$

and for $\delta^0 = (vec(\Sigma(\theta^0))', \theta^{0'})'$, $V(\delta^0) = 0$. In other words we have found a (local) Lyapunov function for our problem. The rest of the proof can be continued as in Yin and Zhu (1990) or Chen and White (1992). ■

Proof of Proposition 4.4 This proof is a simplified version of the proof of the Proposition 4.3. We look for a suitable (local) Lyapunov function. For easier writings let us denote

$$G(\theta^1) = tr\left(\Sigma(\theta^1)\right).$$

Let us consider

$$V_1(e) = \frac{1}{2}\left(G(\theta^0) - e\right)^2$$

and remark that there exists $C_1 > 0$ such that, for any $\theta^1 \in \Theta$,

$$\left\langle \frac{\partial V_1}{\partial e}(e),\, M_1^Q(e,\theta^1) \right\rangle \leq -\left(G(\theta^0) - e\right)^2 + C_1 \left|G(\theta^0) - e\right| \, \|\theta^1 - \theta^0\|$$

On the other hand, let us remark that $eM_2^Q(e,\theta^1) = M(\theta^1)$ with $M(\cdot)$ defined in (4.1). Therefore, under the given assumptions, $G(\theta^0)\,(\partial M_2^Q/\partial \theta^{1\prime})(G(\theta^0),\theta^0)$ is negative stable. Thus there exists Q_2 positive definite such that

$$\left\langle\, Q_2(\theta^1 - \theta^0),\ \frac{\partial M_2^Q}{\partial \theta^{1\prime}}(G(\theta^0),\theta^0)(\theta^1 - \theta^0) \,\right\rangle = -\|\theta^1 - \theta^0\|^2.$$

We deduce that in a (bounded) neighborhood of $G(\theta^0)$

$$\left\langle\, Q_2(\theta^1 - \theta^0),\, M_2^Q(e,\theta^1) \right\rangle = -\|\theta^1 - \theta^0\|^2 + \left\langle\, Q_2(\theta^1 - \theta^0),\ r_2 \right\rangle + o(\|\theta^1 - \theta^0\|^2),$$

with $\|r_2\| \leq C'|G(\theta^0) - e|\,\|\theta^1 - \theta^0\|$, where C' is some positive constant.

Now, for a small enough neighborhood of $G(\theta^0)$, we may define a (local) Lyapunov function

$$V(\delta) = V_1(e) + \frac{C_2}{2}\left\langle\, \theta^1 - \theta^0,\, Q_2(\theta^1 - \theta^0) \right\rangle + g(\|\theta^1 - \theta^0\|^2),$$

with $C_2 > 0$ some suitable constant, V_1 and Q_2 as above in this proof and $g(\cdot)$ some suitable function. The remaining facts allowing to close the proof are as in the proof of Proposition 4.3. ■

Proof of Proposition 5.1 For easier notation define

$$s_T(\theta,\theta^1) = \frac{\partial Q_T}{\partial \theta}\left[\theta, \lambda(\theta^1)\right].$$

Consider a Taylor expansion of the ith component of $s_T(\cdot, \theta_T^{(p(T))})$, $i \in \{1,...,p\}$, in a convex neighborhood of θ^0. We obtain

$$0 = \sqrt{T}\, s_T^i(\theta_T, \theta_T^{(p(T))}) = \sqrt{T} s_T^i(\theta^0, \theta_T^{(p(T))}) + \frac{\partial s_T^i}{\partial \theta'}(\theta, \theta_T^{(p(T))})\Big|_{\theta=\tilde{\theta}_T^i} \sqrt{T}(\theta_T - \theta^0),$$

with $\tilde{\theta}_T^i$ lying between θ^0 and θ_T. On the other hand

$$s_T^i(\theta^0, \theta_T^{(p(T))}) = s_T^i(\theta^0, \theta^0) + \frac{\partial s_T^i}{\partial \theta^{1\prime}}(\theta^0, \theta^1)\Big|_{\theta^1 = \tilde{\tilde{\theta}}_T^i} \sqrt{T}(\theta_T^{(p(T))} - \theta^0)$$

with $\tilde{\tilde{\theta}}_T^i$ lying between θ^0 and $\theta_T^{(p(T))}$. Thus we get

$$\begin{aligned} 0 &= \sqrt{T} s_T^i(\theta^0, \theta^0) + \left(\frac{\partial s_T^i}{\partial \theta'}(\theta, \theta_T^{(p(T))})\Big|_{\theta=\tilde{\theta}_T^i} + \frac{\partial s_T^i}{\partial \theta^{1\prime}}(\theta^0, \theta^1)\Big|_{\theta^1 = \tilde{\tilde{\theta}}_T^i} \right) \sqrt{T}(\theta_T - \theta^0) \\ &+ \frac{\partial s_T^i}{\partial \theta^{1\prime}}(\theta^0, \theta^1)\Big|_{\theta^1 = \tilde{\tilde{\theta}}_T^i} \sqrt{T}(\theta_T^{(p(T))} - \theta_T). \end{aligned}$$

Looking at the stated assumptions we get that

$$\sqrt{T}(\theta_T - \theta^0) \xrightarrow{d} N_p(0, V(\theta^0))$$

with $V(\theta^0)$ as written in the proposition. ■

Proof of Proposition 6.1 *i*) Simple application of Theorem II.2.4 of KW.

ii) First we compute $\bar{H}_2$ as in (6.1) for $M(\cdot)' = (M_1^M(\cdot)', M_2^M(\cdot)')$. We get

$$\bar{H}_2 = \begin{pmatrix} \left(-c_2 + \frac{1}{2}\right) I_{p^2} & c_2 \frac{\partial vec(\Sigma(\theta^1))}{\partial \theta^{1\prime}}\Big|_{\theta^1=\theta^0} \\ c_2 E_{\theta^0}\left[M_1(\theta^0)' \otimes I_p\right] \frac{\partial vec(G^{-1})}{\partial vec(G)}\Big|_{G=\Sigma(\theta^0)} & F_2(\theta^0, c_2) \end{pmatrix}$$

$$= \begin{pmatrix} \left(-c_2 + \frac{1}{2}\right) I_{p^2} & c_2 \frac{\partial vec(\Sigma(\theta^1))}{\partial \theta^{1\prime}}\Big|_{\theta^1=\theta^0} \\ 0 & F_2(\theta^0, c_2) \end{pmatrix},$$

with $M_1(\cdot)$ defined as in (4.7) for $t = 1$. Under the stated assumptions $\bar{H}_2$ is negative stable. The lower right $p \times p$ bloc of the matrix R defined as in (6.3) with $M(\cdot)$ considered in this proof is

$$\begin{aligned}
R_2 &= \sum_{t\geq 0} E\left[\,(\Sigma(\theta^0)^{-1}M_1(\theta^0)\,)\,(\,\Sigma(\theta^0)^{-1}M_{t+1}(\theta^0)\,)'\,\right] \\
&+ \sum_{t\geq 1} E\left[\,(\Sigma(\theta^0)^{-1}M_{t+1}(\theta^0)\,)\,(\,\Sigma(\theta^0)^{-1}M_1(\theta^0)\,)'\,\right] \\
&= \Sigma(\theta^0)^{-1}\left(\sum_{t\geq 0} E\left[M_1(\theta^0)M_{t+1}(\theta^0)'\right] + \sum_{t\geq 1} E\left[M_{t+1}(\theta^0)M_1(\theta^0)'\right]\right)\Sigma(\theta^0)^{-1} \\
&= \Sigma(\theta^0)^{-1}B(\theta^0)\Sigma(\theta^0)^{-1},
\end{aligned}$$

with $M_t(\cdot)$, $t \geq 1$, as in (4.7). Due to the block triangularity of $\bar{H}_2$, the lower right $p \times p$ bloc of the corresponding matrix (6.2) is exactly as in (6.7). It remains to invoke Theorem II.2.4 of KW.

iii) If we compute $\bar{H}_3$ as in (6.1) for $M(\cdot)$ as in (4.18)-(4.19) we obtain

$$\begin{aligned}
\bar{H}_3 &= \begin{pmatrix} -c_3 + \frac{1}{2} & c_3 \left.\frac{\partial tr(\Sigma(\theta^1))}{\partial \theta^{1\,\prime}}\right|_{\theta^1=\theta^0} \\ -c_3 \frac{M(\theta^0)}{tr(\Sigma(\theta^0))^2} & F_3(\theta^0, c_3) \end{pmatrix} \\
&= \begin{pmatrix} -c_3 + \frac{1}{2} & c_3 \left.\frac{\partial tr(\Sigma(\theta^1))}{\partial \theta^{1\,\prime}}\right|_{\theta^1=\theta^0} \\ 0 & F_3(\theta^0, c_3) \end{pmatrix}.
\end{aligned}$$

Thus $\bar{H}_3$ is block triungular and negative stable. The lower $p \times p$ bloc of the corresponding matrix R obtained as in (6.3) is

$$\begin{aligned}
R_3 &= \sum_{t\geq 0} E\left[\left(\frac{1}{tr(\Sigma(\theta^0))}M_1(\theta^0)\right)\left(\frac{1}{tr(\Sigma(\theta^0))}M_{t+1}(\theta^0)\right)'\right] \\
&+ \sum_{t\geq 1} E\left[\left(\frac{1}{tr(\Sigma(\theta^0))}M_{t+1}(\theta^0)\right)\left(\frac{1}{tr(\Sigma(\theta^0))}M_1(\theta^0)\right)'\right] \\
&= \frac{1}{tr(\Sigma(\theta^0))^2}B(\theta^0),
\end{aligned}$$

with $M_t(\cdot)$, $t \geq 1$, as in (4.7). Now we apply Theorem II.2.4 of KW and we obtain the stated result. ■