2003s-08

Iterative and Recursive Estimation in Structural Non-Adaptive Models

Sergio Pastorello, Valentin Patilea, Éric Renault

Série Scientifique
Scientific Series

Montréal
Avril 2003

CIRANO
Centre interuniversitaire de recherche
en analyse des organisations

Les cahiers de la série scientifique (CS) visent à rendre accessibles des résultats de recherche effectuée au CIRANO afin de susciter échanges et commentaires. Ces cahiers sont écrits dans le style des publications scientifiques. Les idées et les opinions émises sont sous l'unique responsabilité des auteurs et ne représentent pas nécessairement les positions du CIRANO ou de ses partenaires.

This paper presents research carried out at CIRANO and aims at encouraging discussion and comment. The observations and viewpoints expressed are the sole responsibility of the authors. They do not necessarily represent positions of CIRANO or its partners.

ISSN 1198-8177

Iterative and Recursive Estimation in Structural Non-Adaptive Models*

Sergio Pastorello[†], *Valentin Patilea*[‡], *Éric Renault*[†]

Résumé / *Abstract*

Nous proposons une méthode d'inférence appelée «latent backfitting». Cette méthode est spécialement conçue pour les modèles économétriques dans lesquels les relations structurelles d'intérêt définissent les variables endogènes observées comme une fonction connue des variables d'états non observées et des paramètres inconnus. Cette spécification espace-état non linéaire ouvre la voie à des stratégies itératives ou récursives de type EM. Dans l'étape E, les variables d'état sont prédites à partir des observations et des valeurs des paramètres. Dans l'étape M, ces prévisions sont utilisées pour déduire des estimateurs des paramètres inconnus à partir du modèle statistique des variables latentes. L'estimation itérative/récursive proposée est particulièrement utile pour les modèles avec équation de régression latente et les modèles dynamiques d'équilibre utilisant des variables d'état latentes. Les questions relatives à l'application de ces méthodes sont analysées à travers l'exemple des modèles de structure par termes des taux d'intérêt.

Mots clés : Modèles d'évaluation d'actifs financiers, variables latentes, estimation, algorithmes itératifs ou récursifs.

An inference method, called latent backfitting is proposed. It appears well suited for econometric models where the structural relationships of interest define the observed endogenous variables as a known function of unobserved state variables and unknown parameters. This nonlinear state space specification paves the way for iterative or recursive EM-like strategies. In the E-steps the state variables are forecasted given the observations and a value of the parameters. In the M-steps these forecasts are used to deduce estimators of the unknown parameters from the statistical model of latent variables. The proposed iterative/recursive estimation is particularly useful for latent regression models and for dynamic equilibrium models involving latent state variables. Practical implementation issues are discussed through the example of term structure models of interest rates.

Keywords: *Asset Pricing Models, Latent Variables, Estimation, Iterative or Recursive Algorithms.*

* A previous version of this paper was circulated under the title "Continuously Updated M-estimators" (CORE, Louvain La Neuve, Working Paper # 9776). We would like to thank David Bates, Xiaohong Chen, John Geweke, Eric Jacquier, Nicholas Polson and seminar participants at the University of British Columbia, Chicago School of Business, Iowa University, Princeton University, University of North Carolina, MIT, Harvard University and Queens University for helpful discussions. Special thanks are due to Halbert White for drawing our attention on the recursive estimation procedures and to Eric Ghysels for details comments. All remaining errors are our responsibility. The current version of this paper was written while Eric Renault visited the Bendheim Center for Finance, Princeton University. The hospitality of this institution is grateful acknowledged.

[†] Dipartimento di Scienze Economiche, Università di Bologna, Italy.

[‡] LEO, Université d'Orléans, France.

[†] CIRANO and CIREQ, Université de Montréal. Email: Eric.Renault@UMontreal.ca

1 Introduction

The goal of this paper is to provide an unified theory for a variety of estimation algorithms that can be used as part of a likelihood analysis of a non linear state space model or as part of a conditional moment-based inference in a structural econometric model. Structural econometric models are challenging because the quantities of interest are not directly related to sample moments of the observations. Typically, these quantities may not only be some unknown parameters but also latent Markov states which enter the observation in a non-analytical form, as the solution to differential equations.

Although this econometric setting arises in many empirical applications, we are going to consider as a leading example for this paper the econometrics of financial market models. Beginning with the Generalized Method of Moments (GMM) as applied by Hansen and Singleton (1982), econometric analysis of asset pricing models is mainly dedicated to an inversion problem: recovering the pricing kernel or stochastic discount factor (SDF) from the observed asset prices.

Empirical asset pricing theory takes as given a set of J discretely observed asset prices Y_t ($Y_t \in \mathsf{R}^J$) and tries to recover the SDF m_{t+1} which relates them to the vector G_{t+1} of their payoffs by the asset pricing model:

$$Y_t = E[m_{t+1} \cdot G_{t+1} | I_t] \tag{1.1}$$

where I_t is the relevant conditioning information at time t. Note that (1.1) implicitly assumes that prices and payoffs are observed at equally spaced intervals, although this assumption is not important.

In the simplest cases, the asset pricing model (1.1) is fully specified by the definition of the joint conditional probability distribution of (m_{t+1}, G_{t+1}) given the observed values at time t of some state variables (which summarize the relevant conditioning information I_t) and the value of a vector λ of q unknown parameters. The two most famous examples of such asset pricing models are the CAPM of Sharpe-Lintner (1964/1965) and the Black-Scholes-Merton (1973) option pricing model.

It has since been widely documented that both the standard CAPM and Black-Scholes (BS) models generate more often than not empirically untenable results. Typically, there is no hope that a fixed volatility parameter in the BS option pricing formula or some fixed shares of the wealth invested in the CAPM market portfolio characterize the SDF in a satisfactory way. This has been acknowledged since the mid-seventies through both the Roll critique and the practitioners' rather bizarre uses of the BS formula. First, the Roll critique pointed out that the wealth portfolio needed for CAPM pricing might not be observable. Second, practitioners have considered the BS pricing formula as a measurement equation to recover an implied volatility parameter. In complete contradiction to the original BS model, this BS implied volatility is seen as a stochastic and time varying summary of the relevant information conveyed by option prices. Similarly, the Roll critique revisited as the "Hansen-Richard critique" (see Cochrane (2001)) stresses that the conditioning information of agents may not be observable, and that one cannot omit it for inference regarding asset pricing models.

This observation does not invalidate the methodology of statistical inference in the context of structural models of financial markets equilibrium. On the contrary, when one imagines that agents on the market observe some state variables which are not those an econometrician can measure directly but the values of which are precisely incorporated in the observed asset prices, the addition of state variables is an elegant way of reconciling the equilibrium paradigm and the statistical data on market prices. While the use of the BS option pricing formula will surely lead to a logical contradiction (different values of a volatility parameter are used, while its constancy is a maintained assumption), the introduction of a convenient number of state variables will avoid this logical contradiction. This strategy of structural econometric modeling of option pricing errors (see Renault (1997) for a survey) maintains the no-arbitrage principle while even very small pricing errors would open the door (up to transaction costs) for huge arbitrage opportunities.

The resulting statistical specification of the economic model (1.1) will appear as a relationship:

$$Y_t = g[Y_t^*, \lambda] \tag{1.2}$$

where g is a given function completely defined by the economic model, possibly in a non-analytic form, while Y_t^* is a vector of possibly latent state variables which summarize the conditioning information I_t needed to compute the conditional expectation (1.1) when the relevant distributional characteristics are specified up to a vector λ of unknown parameters. Typically, the stochastic process $(m, G) = \{m_{t+1}, G_{t+1}\}_{t=1}^{T}$ will be seen as a function of the path over the lifetime of the asset of the possibly continuous-time stochastic process Y^* of state variables, the probability distribution of which is described by a vector θ of p unknown parameters. Then, the joint definition of the SDF and the vector of payoffs will characterize the vector λ of "pricing parameters" as a given function of the vector θ of statistical parameters:

$$\lambda = \lambda(\theta), \theta \in \Theta \subset \mathsf{R}^p. \tag{1.3}$$

This "data augmentation" modeling strategy is often considered as a difficult challenge for empirical finance and has generated a variety of new simulation-based minimum chi-square methods surveyed in Tauchen (1997). However, these methods often use a fully parametric model as a simulator while neither the parametric efficiency nor the semiparametric robustness are guaranteed (see Dridi and Renault (2000) for a discussion). Still the Bayesian literature and in particular the explosion of papers in the area of Markov chain Monte Carlo (MCMC) in the past ten years told us that data augmentation should not amount to "difficulty augmentation".

As nicely summarized by Tanner (1996) (see introduction of chapter 4 about EM algorithm) "all these data augmentation algorithms share a common approach to problems: rather than performing a complicated maximization or simulation, one augments the observed data with "stuff" (latent data) that simplifies the calculation and subsequently performs a series of simple maximizations or simulations. This "stuff" can be the "missing" data or parameters values". Typically, in our case, this will include both the data augmentation from Y to Y^* and the parameter augmentation from λ to θ. And, instead of increasing the difficulty, it will allow for simpler maximization-based estimation procedures (M-estimation or minimum distance) as applied to latent data.

The basic idea of this paper can be described by analogy with the MCMC methodology. Consider the following algorithm: given an initial estimator $\theta^{(1)}$ of the parameters of interest, compute a proxy $Y_t^{*(1)}$ of the latent variables Y_t^* from the structural relationships: $Y_t = g[Y_t^{*(1)}, \lambda(\theta^{(1)})], t = 1, \cdots, T$. This is the analog of the MCMC step of drawing latent variables given the observables and an initial draw of the parameters. Then, using the simplicity of the latent model enables one to compute an estimator $\theta^{(2)}$ of the parameters from the "draw" $Y_t^{*(1)} t = 1, \cdots, T$ of the latent data. This is the analog of the posterior mean (or a random draw) of the augmented posterior distribution which has precisely been used to obtain simplicity. Typically, the latent statistical model which characterizes the stochastic latent data generator is much simpler than the statistical model which characterizes the law of motion of the observed process Y. Continuing in this fashion, the algorithm generates a sequence of random variables $[Y_t^{*(p)}, \theta^{(p)}]$ which are consistent with both the observed data and the structural model:

$$Y_t = g[Y_t^{*(p)}, \lambda(\theta^{(p)})],\ t = 1, \cdots, T. \tag{1.4}$$

Moreover, in the same way that a fixed point argument implies the convergence of the Markov chain produced by a MCMC algorithm, a fixed point argument is going to ensure that the limit $[Y_t^{*(\infty)}, \theta^{(\infty)}] = Lim_{p\to\infty}[Y_t^{*(p)}, \theta^{(p)}]$ of the above algorithm will exist. Then, $\theta^{(\infty)}$ is going to be a consistent estimator (for a sample size T going to infinity) of the true value θ^0 of the parameters while $Y_t^{*(\infty)}$ consistently "estimates" the best proxy of Y_t^* one could deduce from the exact pricing relationship:

$$Y_t = g[Y_t^*, \lambda(\theta^0)] \tag{1.5}$$

We will show that it is even not necessary to iterate the above algorithm to infinity. A number $p(T)$ of iterations going to infinity with T at a sufficient rate will ensure the consistency and root-T normality of the resulting estimators.

Therefore, while in a recent survey of MCMC methods for financial econometrics, Johannes and Polson(2001) claim that "in contrast to all other estimation methods, MCMC is a unified estimation procedure, estimating both parameters and state variables simultaneously", the iterative methodology put forward in this paper can be seen as the frequentist analog of MCMC and shares most of its advantages. Among the three main advantages of MCMC methods highlighted by Johannes and Polson (2001), our iterative procedure shares two of them. Besides the aforementioned simultaneous estimation of parameters and state variables, it is also true that it "exclusively uses conditional simulation, and it therefore avoids numerical maximization and long unconditional state variable simulation. Because of this, (...) estimation is typically extremely fast in terms of computing time". Of course, when one reads "conditional simulation", one should understand in the context of classic statistics "estimation" of Y_t^* and θ^0, irrespective of the fact that this estimation is simulation-based or not. The important point, in particular for computational time, is that it is "conditional" that is it takes advantage of a previous estimator of the parameters or of the state variables to work directly with the "stuff" (latent data) that simplifies the calculation.

Of course, the third advantage of MCMC methods, namely to "deliver exact finite sample inference" is not shared by our frequentist methodology which replaces the Bayesian paradigm by an asymptotic one. However, the alleged exact finite sample performance of the Bayesian approach rests upon the validity of the specification of a prior, while our asymptotic approach may be much less demanding in terms of parametric specification. The likelihood based inference the closest to our methodology is the EM algorithm (Dempster et al. (1977)). In some particular cases, our approach can be interpreted as an EM algorithm. However, it is more general for at least two reasons:

First, it is not limited to a likelihood framework and can be useful in a number of conditional moment-based inference procedures as popular in modern semiparametric structural econometrics. The analog of the M-step of the EM algorithm (computation of the "estimator" $\theta^{(p+1)}$ from the "data" $Y_t^{*(p)}, t = 1, \ldots, T$) will then be performed by any extremum or minimum distance principle.

Second, even in its maximum-likelihood type applications, our approach does not resort to EM for the main examples of this paper. In these examples, EM theory cannot be applied since the support of the conditional probability distribution of the latent variables given the observable ones depend on the unknown parameters via the "inversion" of the measurement equation (1.5). The analog of the E-step of the EM algorithm (computation of the "data" $Y_t^{*(p+1)}$ from the new estimator $\theta^{(p+1)}$) will resort more to an "implied-state" approach as popularized in finance by the practitioners' use of BS implied volatility parameters.

Typically, as explained above, it is often the case that a number of state variables has only been introduced to get rid of observed pricing errors with respect to the theoretical asset pricing model. In this case, the measurement equation (1.5) will provide, for a given value of the parameters, a one-to-one relationship between observed prices or returns Y_t and latent state variables Y_t^* . Then, implied states Y_t^* can simply be recovered from observations by:

$$Y_t = g[Y_t^*, \lambda] \Leftrightarrow Y_t^* = g^{-1}[Y_t, \lambda]. \tag{1.6}$$

Of course, this does not make the inference issue as trivial as it may appear at first sight since implied states can be recovered only for a given value of the unknown parameters. Even in the simplest case of a nonlinear state space model defined by the measurement equation (1.5) (conformable to (1.6) with $\lambda = \lambda(\theta)$) jointly with a transition equation which specifies the dynamics of the latent process Y^* as a parametric model indexed by the vector θ of unknown parameters, it turns out that the identification and efficiency issues for asymptotic statistical inference about θ have been largely ignored by the literature until now.

In the context of multifactor equilibrium models of the term structure of interest rates, Chen and Scott (1993), Pearson and Sun (1994) and Duan (1994) have been the first papers to propose a "maximum likelihood estimation using price data of the derivative contract". These papers have stressed in particular that, since the observed data are the transformed values of some underlying state variables and "since the transformations in finance generally involve unknown parameters and the inverse transformations may not have analytical solutions, the likelihood functions can be difficult to derive" and in particular "it will be better to avoid the direct computation of the Jacobian for the inverse transformation" (Duan (1994)).

Affine term structure models (Duffie and Kan (1996), Dai and Singleton (2000)) are very popular nowadays precisely because the bond prices can be calculated quickly as a function of state variables solving a system of ordinary differential equations. It is usually argued that, given the closed-form expressions for bond prices, one can invert any n bond prices into the n state variables and use the implied state variables in the estimation as if they were directly observable.

This common belief is actually wrong on theoretical grounds. The key point is that, since transformations in finance generally involve unknown parameters, these unknown parameters will appear in the likelihood function, not only through the parametric model of the latent state variables, but also as inputs of the inverse transformation and its Jacobian. Besides the aforementioned computational difficulties, the main consequence of that is a modification of the statistical information, as measured by the Fisher information matrix in the likelihood setting.

The first contribution of this paper is a theoretical study of the difference between the statistical information conveyed by the state variables and the one conveyed by observed data. This study will be carried out not only in the fully parametric setting (implied states maximum likelihood and Fisher information matrix) but also in more general semiparametric contexts defined either by a set of conditional moment restrictions (implied states GMM and semiparametric efficiency bound) or by some quasi-maximum likelihood or any extremum estimation (robustified Fisher information matrix).

Besides its theoretical drawbacks, the aforementioned confusion between infeasible efficient estimation with hypothetically observed state variables (we will say hereafter "latent" estimation) and actual "implied-states" estimation is misleading for the definition of well-suited estimation algorithms. The other contribution of this paper is various algorithmic estimation procedures, first with an iterative viewpoint, second with a recursive approach. A general asymptotic theory of the proposed estimators is developed.

As far as iterative estimation is concerned, the asymptotic properties of our EM-type estimator cannot properly be assessed without a clear characterization of its two main ingredients: asymptotic probability distribution of the infeasible latent estimator on the one hand, and asymptotic contracting feature of the mapping $\bar{\theta}_T$ which generates the sequences of random variables $\theta^{(p)}$ on the other hand:

$$\theta^{(p+1)} = \bar{\theta}_T\left(\theta^{(p)}\right) \quad p = 1, \cdots, p(T). \tag{1.7}$$

The general asymptotic theory of the estimator $\theta^{p(T)}$ proposed in this paper nests several estimators previously proposed in the literature. Renault and Touzi (1996) had considered, for the purpose of option pricing with stochastic volatility, the case of maximum likelihood with $p(T) = \infty$. Renault (1997) sketched the case of more general extremum estimators and other fields of applications. Pan (2002) focuses on an implied states method of moments which is implicitly considered with $p(T) = \infty$.

Moreover, the structural econometric model of interest (1.5) may not only be any asset pricing model but also other equilibrium models, as produced in particular by game theory. Florens, Protopopescu and Richard (2001) consider the case of auction markets. Our estimator also nests the iterative least square (ILS) estimator as proposed by Gouriéroux, Monfort, Renault and Trognon (1987) and extensively studied by Dominitz and Sherman (2001) in the context of semiparametric binary choice models.

We will show that in this latter context too it is worth disentangling the semiparametric efficiency concepts associated respectively with latent estimation and implied-states estimation.

It is often the case that the filtering of the latent variables (computation of the implied states from a given value of the parameters) involves computer intensive calculations and/or simulations like in the E-step of simulated SEM algorithms (see *e.g.*, Ruud (1991)). Therefore, we also propose some recursive procedures which, contrary to the iterative ones, do not involve the batch processing of the full block of data at each step of the algorithm but only a simple updating of the estimator each time new data arrives. Moreover, the updating scheme does not imply an optimization procedure (*e.g.*, Young (1985), Kuan and White (1994), Kushner and Yin (1997)). Generally speaking, the advantage of a recursive approach is twofold. First, as long as the guess on the parameters is far from the true value, the updating of the guess may be performed using only a small part of the sample which allows much faster nonlinear filtering procedures. Once the sequence of recursive guesses stabilize one may use these values as starting values for an iterative algorithm. Second, the quick and simple updating formula provided by the recursive approach is well-suited for on-line estimation.

The price to pay for time saving through recursive procedures which are directly focused on the implied-states latent moment conditions is, in general, some efficiency loss with respect to iterative procedures. We also characterize this efficiency loss in terms of the asymptotic contracting feature of the mapping $\bar{\theta}_T$. This efficiency loss should be much smaller than the one produced by less specific recursive procedures like the Kalman filter, which do not take advantage of the exact non-linear structure of the model and introduce too much error. For the simplest examples of affine term structure models, we put forward some Monte-Carlo results, which show the better computational and statistical performance of our approach with respect to traditional Kalman filter ones (Duan and Simonato (1995), De Jong(2000)). The superior performance of exact implied-states recursive procedures proposed here should be even more striking in more non-linear asset pricing models where the Kalman filter no longer admits any theoretical justification.

The paper is organized as follows. In section 2, we study the general identification problem for our implied-states approach and we compare it with some classical issues of non-adaptivity in econometrics. The regression example allows us to interpret our iterative approach as an extension of classical backfitting. The identification condition is further interpreted in the framework of latent regression models and iterative least squares through generalized residuals. In section 3, we specialize the implied states approach to the case (1.6) where the observed variables are one-to-one functions of the latent state variables. We show how our approach can apply to GMM and likelihood type criteria. Section 4 is devoted to asymptotic theory of the iterative estimators while this theory is extended in section 5 to a recursive version of the same algorithms. The efficiency loss of the recursive implied-states estimator with respect to the iterative one is also analyzed. The practical issues associated with the implementation of the various competing approaches are sketched in section 6 through a short empirical illustration in the context of a model of the term structure of interest rates. Miscellaneous proposals for further empirical research and concluding remarks appear in section 7. The technical proofs are gathered in the appendix.

2 Identification and non-adaptivity

In this section, we motivate our iterative approach as a natural way to address an issue of non-adaptivity, that is the occurrence of some nuisance parameters which prevent one from obtaining directly a consistent estimator of the parameters of interest. The price to pay for the proposed iterative strategy is that the required identification condition may be stronger than the usual one. Though, we will argue that the strengthening is rather weak.

In order to support this claim, we consider some benchmark examples of non-adaptivity in econometrics and we show that our iterative strategy nests some already well-documented procedures. The first example is the partially parametric regression model, when only some parameters of the parametric part are of interest. Our iterative estimation procedure is then tantamount to standard backfitting and the required identification condition appears to be, at least locally, not stronger than the identification condition of the regression model.

The second example is iterative least squares through generalized residuals of a latent regression model. In this case, the iterative procedure can be interpreted as an extension of standard backfitting that we term latent backfitting. The relevant identification conditions have already been extensively characterized by Dominitz and Sherman (2001) and we just revisit their argument to illustrate that the needed identification condition is not much stronger than the usual one.

2.1 The general framework

The focus of interest is a class of inference problems that can be described via the more general issue of non-adaptivity. We consider a semiparametric statistical model specified by a family $\mathcal{P}$ of probability measures P on the sampling space and two mappings $\boldsymbol{\theta}(\cdot) : \mathcal{P} \to \Theta \subset \mathsf{R}^p$ and $\boldsymbol{\lambda}(\cdot) : \mathcal{P} \to \Gamma \subset \mathsf{R}^q$. The vector $\theta = \boldsymbol{\theta}(P)$ contains the p parameters representing the features of interest of the probability P which has hypothetically governed the sampling of the observations. The vector $\lambda = \boldsymbol{\lambda}(P)$ contains q nuisance parameters. The model is assumed to be well specified in the sense that there exists a true unknown probability $P^0 \in \mathcal{P}$ which defines the Data Generating Process (DGP) and $\theta^0 = \boldsymbol{\theta}(P^0)$ and $\lambda^0 = \boldsymbol{\lambda}(P^0)$ are the corresponding true unknown values of the parameters of interest and of the nuisance parameters, respectively.

We are interested in an extremum estimator of θ deduced from a sample based criterion (objective function)

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

which depends on both vectors $\theta \in \Theta$ and $\lambda \in \Gamma$. Let us consider a first set of assumptions similar to those usually imposed for extremum 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 $Q_\infty[\cdot,\cdot] = Q_{\infty,P^0}[\cdot,\cdot]$ such that

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

The specificity of the general framework considered in this paper is the mode of occurrence of the nuisance parameters λ into the objective function Q_T. As noted in the introduction, θ is associated with the probability distribution of some state variables process Y^* while λ appears into the measurement equation:

$$Y_t = g(Y_t^*, \lambda). \tag{2.1}$$

In other words, $Q_T\left[\theta, \lambda^0\right]$ will typically be an objective function provided by some standard latent extremum estimation principle while λ appears due to the need to recover implied states from the observations $\{Y_t\}_{1\leq t\leq T}$ and the measurement equation (2.1).

Since this measurement equation is defined by an equilibrium model stemming from for example option pricing or game theory, it is itself tightly related to the DGP of the state variables. Therefore the true unknown value λ^0 of the nuisance parameters is a function of the true unknown value θ^0 of the parameters of interest. To highlight this point, we will write $\lambda(P) = \lambda(\theta(P))$. Note that $\lambda(P)$ may also depend on some other nuisance parameters insofar as we get a consistent estimator of them. The important point is that there is no hope in general to obtain a preliminary consistent estimator of λ^0 to be able to deduce a consistent estimator of θ^0 since the former is a function of the latter.

In other words, we consider that there is no reason to assume that an extremum estimator $\bar{\theta}_T(\bar{\lambda})$ built as

$$\bar{\theta}_T(\lambda) = \arg \max_{\theta} Q_T\left[\theta, \lambda\right] \tag{2.2}$$

from an arbitrarily fixed value λ of the nuisance parameters will provide a consistent estimation of the true value θ^0 of the parameters of interest. Typically, only $\bar{\theta}_T\left(\lambda^*\right)$ computed from some particular value λ^* of the nuisance parameters may provide a consistent estimator. This is the general issue of non-adaptivity.

Assumption 2.2 *i) For any $\lambda \in \Gamma$, the function $\theta \to Q_\infty\left[\theta, \lambda\right]$ admits a unique maximizer $\overline{\theta}(P^0, \lambda)$, where P^0 is the probability measure governing the observations.*

ii) There exists some $\lambda^* \in \Gamma$ such that $\overline{\theta}(P^0, \lambda^*) = \theta^0$.

Perhaps, the best known example of non-adaptivity in econometrics is the linear regression model with two sets of explanatory variables which we shall consider now for illustration purposes. With a slight change of notations to introduce exogenous variables, let $\left(Y_t, X_t\right), t = 1, ..., T$ be i.i.d. random vectors such that $Y_t \in \mathsf{R}$, $X_t = \left(X_{1t}', X_{2t}'\right)' \in \mathsf{R}^p \times \mathsf{R}^q$ and

$$\begin{cases} E_P\left(X_t X_t'\right) \text{ is a non singular matrix,} \\ \\ E_P\left[Y_t - X_{1t}'\boldsymbol{\theta}(P) - X_{2t}'\boldsymbol{\lambda}(P) \mid X_{1t}, X_{2t}\right] = 0, \\ \\ \boldsymbol{\theta}(\mathcal{P}) = \Theta = \mathsf{R}^p \text{ and } \boldsymbol{\lambda}(\mathcal{P}) = \Gamma = \mathsf{R}^q, \end{cases} \tag{2.3}$$

where E_P denotes the expectation with respect to the probability P; in particular, $E_{P^0} = E^0$ stands for the expectation with respect to the DGP.

The ordinary least-squares principle leads to the maximization criterion

$$Q_\infty\left[\theta, \lambda\right] = -E^0\left(Y_t - X_{1t}'\theta - X_{2t}'\lambda\right)^2. \tag{2.4}$$

Therefore, the set of values λ^* of the nuisance parameters λ such that

$$\bar{\theta}\left(P^0, \lambda^*\right) = \theta^0$$

is characterized by a translation of the kernel space of the matrix $E^0\left(X_{1t}X_{2t}'\right)$:

$$\lambda^* - \lambda^0 \in Ker\ E^0\left(X_{1t}X_{2t}'\right).$$

Of course, we are faced with the non-adaptivity problem when the kernel subspace does not coincide with the whole space R^q. Therefore, except for specific choices of the function $\lambda(\cdot)$, we have

$$\theta^0 \neq \arg \max_{\theta} Q_\infty\left[\theta, \lambda(\theta)\right] \tag{2.5}$$

In the simple linear model (2.3), the Frisch-Waugh theorem (see Frisch and Waugh (1933); see also Davidson and Mackinnon (1993), page 19) provides a particular function $\lambda^0(\cdot)$ to avoid the 'problem' (2.5). Indeed, if we define $\lambda^0(\cdot)$ by

$$E^0\left[X_{2t}X_{2t}'\ \lambda^0(\theta)\right] = E^0\left[X_{2t}\left(Y_t - X_{1t}'\theta\right)\right], \tag{2.6}$$

then

$$\theta^0 = \arg\ \max_\theta\ Q_\infty\left[\theta, \lambda^0(\theta)\right]. \tag{2.7}$$

In practice, it may happen that the function $\lambda^0(\cdot)$ is unknown, but it can be consistently estimated by, say, $\lambda_T(\cdot)$. In the Frisch-Waugh example

$$\lambda_T(\theta) = \left(\mathsf{X}_2'\mathsf{X}_2\right)^{-1}\mathsf{X}_2'\left[\mathsf{Y} - \mathsf{X}_1\theta\right],$$

is the empirical counterpart of (2.6); $\mathsf{Y}, \mathsf{X}_1, \mathsf{X}_2$ are the matrices corresponding to the first T observations of the variables Y, X_1 and X_2, respectively. The finite sample criterion associated with (2.4) where λ is replaced by the function λ_T equals:

$$\begin{aligned} Q_T\left[\theta, \lambda_T(\theta^1)\right] &= -T^{-1}\sum_{i=1}^{T}\left[Y_t - X_{1t}'\theta - X_{2t}'\lambda_T(\theta^1)\right]^2 \\ &= -T^{-1}\left\|(Id - \mathsf{P}_{\mathsf{X}_2})\mathsf{Y} - \mathsf{X}_1\theta + \mathsf{P}_{\mathsf{X}_2}\mathsf{X}_1\theta^1\right\|^2, \end{aligned}$$

where $\mathsf{P}_{\mathsf{X}_2} = \mathsf{X}_2\left(\mathsf{X}_2'\mathsf{X}_2\right)^{-1}\mathsf{X}_2'$ is the orthogonal projection matrix onto the subspace of R^T spanned by the columns of X_2.

The focus of interest of this paper is a class of structural econometric models where only a function $\lambda(\cdot)$ such that

$$\theta^0 = \arg\max_\theta Q_\infty\left[\theta, \lambda(\theta^0)\right] \tag{2.8}$$

is known, or is available by estimation, but (2.5) may happen. Therefore, by analogy with standard backfitting, natural estimation strategies based on the criterion Q_T should distinguish the two occurrences of θ. The criterion is then written $Q_T\left[\theta, \lambda_T\left(\theta^1\right)\right]$ and the estimations are defined through iterative steps:

$$\theta^{(p+1)} = \arg\max_\theta Q_T\left[\theta, \lambda(\theta^{(p)})\right], \qquad p = 1, 2, ... \tag{2.9}$$

Note that for the sake of notational simplicity we always denote by $Q_T\left[\theta, \lambda(\theta^1)\right]$ the finite sample criterion while, in some cases, $\lambda(\theta^1)$ should be replaced by a consistent estimator $\lambda_T(\theta^1)$. In other words, by a slight abuse, all the data dependence is summarized by the notation Q_T.

It is important to keep in mind in this respect that all the iterative/recursive estimation strategies considered in this paper do not change when the criterion Q_T is replaced by $\tilde{Q}_T$ conformable to

$$\tilde{Q}_T\left[\theta, \lambda(\theta^{(p)})\right] = Q_T\left[\theta, \lambda(\theta^{(p)})\right] + \varphi_T(\theta^{(p)}),$$

for an arbitrary numerical function $\varphi_T(\cdot)$ defined on Θ. Thus, we are definitely unable to check a condition like (2.7).

Moreover, this 'limited information' use of the criterion Q_T requires that we strengthen the standard identification condition. Since, by (2.8) we have in mind an infeasible extremum estimator which would be defined from the criterion $Q_T\left[\theta, \lambda(\theta^0)\right]$, a basic identification condition to be imposed should be that θ^0 is the unique maximizer of the limit criterion $\theta \to Q_\infty\left[\theta, \lambda(\theta^0)\right]$, that is

$$\bar{\theta}\left[P^0, \lambda(\theta^0)\right] = \theta^0. \tag{2.10}$$

This condition would be quite similar to the usual identification condition that would have been imposed if the limit criterion was considered as a function $(\theta, \lambda) \to Q_\infty[\theta, \lambda]$ and λ were nuisance parameters for which a consistent estimator is available (see, e.g. Wooldridge (1994), Theorem 4.3). But, given the non-adaptivity problem and the need to build an estimation procedure from the steps (2.9), we will have to maintain a stronger identification condition which imposes that the only fixed point of the function $\bar{\theta}\left[P^0, \lambda(\cdot)\right]$ is θ^0. Otherwise, it may happen that the statistician will not be able to reject a "bad guess" $\theta^{(p)} \neq \theta^0$ used to build the criterion $Q_T\left[\cdot, \lambda(\theta^{(p)})\right]$. This remark leads to the following identification condition required by our estimation strategy.

Assumption 2.3 $\theta^0 = \boldsymbol{\theta}(P^0)$ *is the unique fixed point of the map* $\bar{\theta}\left[P^0, \lambda(\cdot)\right].$

Note that this assumption implies that θ^0 is well identified from the observations as the unique fixed point of the function $\bar{\theta}\left[P^0, \lambda(\cdot)\right]$ where, recall, $\bar{\theta}\left[P^0, \lambda\right]$ is the unique maximizer of $p\lim Q_T[\cdot, \lambda]$.

Now, let us analyze some conditions ensuring Assumption 2.3. Consider that the function $\bar{\theta}\left[P^0, \lambda(\cdot)\right]$ is differentiable with continuous partial derivatives. If $\bar{\theta}\left[P^0, \lambda(\theta^0)\right] = \theta^0$, by a Taylor expansion argument we can deduce that (at least locally) θ^0 is the unique fixed point of $\theta_1 \to \bar{\theta}\left[P^0, \lambda(\theta_1)\right]$ provided that the matrix

$$I_p - \frac{\partial \bar{\theta}}{\partial \theta^{1\prime}}\left[P^0, \lambda\left(\theta^0\right)\right]$$

is invertible. In terms of the first derivatives of the function $\bar{\theta}\left[P^0, \lambda(\cdot)\right]$, this appears to be the minimal condition for ensuring the unique fixed point property locally. However, θ^0 will be estimated by iterations which requires a stronger condition for ensuring the convergence of the estimator.

Suppose that the function $\bar{\theta}\left[P^0, \lambda(\cdot)\right]$ is known. Then, its unique fixed point could be approximated through the iterations $\theta^{(k)} = \bar{\theta}[P^0, \lambda(\theta^{(k-1)})]$, $k = 1, 2, ...$ Note that, in fact, the iterative estimation strategy methodology proposes to replace $\bar{\theta}\left[P^0, \lambda(\cdot)\right]$ with its sample counterpart (2.2) and to perform the corresponding iterations. A suitable way to ensure the convergence of the iterations $\theta^{(k)}$ is to assume that $\bar{\theta}\left[P^0, \lambda(\cdot)\right]$ is a contracting mapping, that is there exists $c \in [0, 1)$ such that

$$\left\|\bar{\theta}\left[P^0, \lambda(\theta')\right] - \bar{\theta}\left[P^0, \lambda(\theta'')\right]\right\| \leq c\left\|\theta' - \theta''\right\|, \qquad \theta', \theta'' \in \Theta.$$

When $\bar{\theta}\left[P^0, \lambda(\cdot)\right]$ admits continuous first order derivatives and θ^0 is an interior point of Θ, the contracting property is tantamount to the condition

$$\left\| \frac{\partial \bar{\theta}}{\partial \theta^{1\prime}} \left[P^0, \lambda(\theta^0)\right] \right\| < 1, \tag{2.11}$$

at least after reducing Θ; here and for the rest of the paper, $\|A\|$ denotes the spectral norm of the matrix A and it is defined by $\|A\|^2 = \rho(A'A)$ where $\rho(B)$ stands for the spectral radius of the squared matrix B, that is its largest eigenvalue in absolute value. Clearly, if (2.11) holds, then $I_p - \partial\bar{\theta}/\partial\theta^{1\prime}\left[P^0, \lambda(\theta^0)\right]$ is invertible and thus the unique fixed point property is guaranteed locally.

In conclusion, the contracting condition, or its local version (2.11) , will represent key assumptions for proving convergence results for our estimators in a general framework. However, as far as identification is concerned, the relevant condition is the weaker Assumption 2.3.

Let us now illustrate the hierarchy of the identification hypotheses – the basic condition $\bar{\theta}\left[P^0, \lambda(\theta^0)\right] = \theta^0$, the unique fixed point property as well as the model identification condition – in the context of the standard backfitting algorithm.

2.2 Classical backfitting revisited

Consider the backfitting algorithm identification (see, e.g., Hastie and Tibshirani (1990)) in the context of a general partially parametric regression model (PPR hereafter). Such a model is characterized by a regression function split into two parts: a parametric one, possibly nonlinear, and a nonparametric one (see Andrews (1994a), page 52). In general, one writes

$$\begin{cases} Y_t = h(X_{1t}, \theta) + \lambda(X_{2t}) + u_t, & \theta \in \Theta \subset \mathsf{R}^p, \quad \lambda \in \mathcal{G}, \\ E(u_t \mid X_t) = 0, \qquad X_t = (X_{1t}', X_{2t}')', & t = 1, ..., T, \end{cases}$$

with $h(\cdot,\cdot)$ known and $\mathcal{G}$ an infinite dimensional set of real-valued functions. For the sake of simplicity, in this example we assume that the process $\{u_t, X_t'\}$ is stationary and ergodic. The quantities of interest, namely the true unknown values θ^0 and λ^0 of the Euclidean parameter θ and the function λ, respectively, are defined as the unique minimizers of the mean squared error, that is the unique maximizers of

$$Q_\infty\left[\theta, \lambda\right] = -E^0\left[Y_t - h(X_{1t}, \theta) - \lambda(X_{2t})\right]^2.$$

The identification assumption expressed by this condition can be made even more precise using the following decomposition:

$$\begin{aligned} E\left[Y_t - h(X_{1t}, \theta) - \lambda(X_{2t})\right]^2 &= E\left[Y_t - h(X_{1t}, \theta) - E\left(Y_t - h(X_{1t}, \theta) \mid X_{2t}\right)\right]^2 \\ &\quad + E\left[E\left(Y_t - h(X_{1t}, \theta) \mid X_{2t}\right) - \lambda(X_{2t})\right]^2. \end{aligned} \tag{2.12}$$

Indeed, since for any value of θ one can define a function $\lambda(\cdot)$ which renders the second term of the right hand side (2.12) equal to zero, the identification condition for the Euclidean parameter amounts to

$$h(X_{1t},\theta^1) - E^0\left(h(X_{1t},\theta^1) \mid X_{2t}\right) = h(X_{1t},\theta^0) - E^0\left(h(X_{1t},\theta^0) \mid X_{2t}\right) \Longrightarrow \theta^1 = \theta^0.$$

Given a guess $\lambda_T^{(1)}$, a natural way to estimate θ is to consider the empirical counterpart $\theta_T^{(2)}$ of

$$\overline{\theta}(P^0,\lambda) = \arg\max_{\theta\in\Theta} - E^0\left[Y_t - h(X_{1t},\theta) - \lambda(X_{2t})\right]^2$$

with λ replaced by $\lambda_T^{(1)}$. With this estimate $\theta_T^{(2)}$, one can get an improved guess $\lambda_T^{(2)}$ of the function λ by smoothing $Y_t - h(X_{1t},\theta_T^{(2)})$ on X_{2t}. We can continue this iterative smoothing process thus obtaining an example of classical backfitting. If it exists, the limit of the backfitting algorithm can be explicitly characterized as solution of the following system of equations

$$\begin{aligned} \widehat{\lambda}_T &= K\left[\mathsf{Y} - \mathsf{h}(\mathsf{X}_1,\widehat{\theta}_T)\right] \\ \widehat{\theta}_T &= \arg\min_{\theta\in\Theta}\sum_{t=1}^{T}\left[Y_t - h(X_{1t},\theta) - \widehat{\lambda}_T(X_{2t})\right]^2, \end{aligned} \tag{2.13}$$

where $\mathsf{h}(\mathsf{X}_1,\theta) = (h(X_{11},\theta),...,h(X_{1T},\theta))'$ and K is the smoothing matrix used to estimate the conditional expectation of $Y - h(X_1,\theta)$ given X_2. Such a fixed point may not exist in finite samples but only asymptotically when T goes to infinity. This issue will be addressed later in our general theory (see section 4) but is omitted here for the sake of expositional simplicity. Hence, $\widehat{\theta}_T$ is obtained by solving the first order conditions

$$\sum_{t=1}^{T}\frac{\partial h}{\partial\theta}(X_{1t},\widehat{\theta}_T)\left[Y_t - h(X_{1t},\widehat{\theta}_T) - \widehat{\lambda}_T(X_{2t})\right] = 0,$$

which corresponds to the limit first order conditions

$$E^0\left[\frac{\partial h}{\partial\theta}(X_{1t},\theta^{(p+1)})\left[Y_t - h(X_{1t},\theta^{(p+1)}) - \lambda^{(p)}(X_{2t})\right]\right] = 0, \tag{2.14}$$

where the function $\lambda^{(p)}(\cdot)$ is defined by

$$\lambda^{(p)}(X_{2t}) = \lambda(\theta^{(p)})(X_{2t}) = E^0\left[Y_t - h(X_{1t},\theta^{(p)}) \mid X_{2t}\right].$$

Therefore, the backfitting algorithm will possibly provide a consistent estimator of θ^0 insofar as the following identification condition is fulfilled:

$$E^0\left[\frac{\partial h}{\partial\theta}(X_{1t},\theta^1)\left(Y_t - h(X_{1t},\theta^1) - E^0\left[Y_t - h(X_{1t},\theta^1) \mid X_{2t}\right]\right)\right] = 0 \Longrightarrow \theta^1 = \theta^0.$$

When maintaining the assumption that the first order conditions characterize a unique maximizer, the backfitting identification condition coincides with the unique fixed point condition of Assumption 2.3 where the function $\bar{\theta}$ is defined by

$$\bar{\theta}\left[P^0, \lambda\left(\theta^1\right)\right] = \arg\max_{\theta} - E^0\left[Y_t - h\left(X_1, \theta\right) - \lambda\left(\theta^1\right)\left(X_{2t}\right)\right]^2.$$

Using the model equations

$$Y_t = h(X_{1t}, \theta^0) + \lambda^0(X_{2t}) + u_t, \qquad E^0(u_t \mid X_t) = 0,$$

we rewrite this backfitting identification condition as:

Condition **C1** (backfitting identification)

$$\begin{aligned} & E^0\left[\frac{\partial h}{\partial \theta}(X_{1t}, \theta^1)\left(h(X_{1t}, \theta^1) - E^0\left[h(X_{1t}, \theta^1) \mid X_{2t}\right]\right)\right] \\ & = E^0\left[\frac{\partial h}{\partial \theta}(X_{1t}, \theta^1)\left(h(X_{1t}, \theta^0) - E^0\left[h(X_{1t}, \theta^0) \mid X_{2t}\right]\right)\right] \quad \Longrightarrow \theta^1 = \theta^0. \end{aligned}$$

As already noted, the PPR model identification condition can be written in the following way.

Condition **C2** (identification in the PPR model)

$$h(X_{1t}, \theta^1) - E^0\left(h(X_{1t}, \theta^1) \mid X_{2t}\right) = h(X_{1t}, \theta^0) - E^0\left(h(X_{1t}, \theta^0) \mid X_{2t}\right) \Longrightarrow \theta^1 = \theta^0.$$

Finally, let us note that the basic identification condition

$$\bar{\theta}\left[P^0, \lambda\left(\theta^0\right)\right] = \theta^0$$

is tantamount to the following standard identification condition in the "latent" regression model.

Condition **C3** (identification in the latent regression model)

$$h(X_{1t}, \theta^1) = h(X_{1t}, \theta^0) \Longrightarrow \theta^1 = \theta^0;$$

It is clear that

$$\mathsf{C1} \Longrightarrow \mathsf{C2} \Longrightarrow \mathsf{C3}.$$

As previously mentioned, the difficulty in PPR models appears when considering nonlinear regression functions h. If h is linear (this is the case of partially linear regression model), the conditions **C1** and **C2** are equivalent and mean that the residual variance

$Var^0\left(X_{1t}-E\left[X_{1t}\mid X_{2t}\right]\right)$ is positive definite. **C3** means that $E^0\left[X_{1t}X_{1t}'\right]$ is positive definite.

Let us now study to what extent, in the general PPR, the backfitting identification condition **C1** is more restrictive than the standard identification condition **C2**. Define

$$\varphi(X_t,Y_t,\theta,\lambda\left(\theta^1\right))=\frac{\partial h}{\partial\theta}(X_{1t},\theta)\left(Y_t-h(X_{1t},\theta)-E^0\left[Y_t-h(X_{1t},\theta^1)\mid X_{2t}\right]\right)$$

and assume that the function $\bar\theta\left[P^0,\lambda\left(\cdot\right)\right]$ can be also defined as the implicit solution of

$$E^0\left[\varphi(X_t,Y_t,\bar\theta\left[P^0,\lambda\left(\theta^1\right)\right],\lambda\left(\theta^1\right))\right]=0,\qquad\qquad \theta^1\in\Theta.$$

Assuming the needed regularity conditions, differentiate this identity with respect to θ^1, take $\theta^1=\theta^0$ (recall that $\bar\theta\left[P^0,\lambda\left(\theta^0\right)\right]=\theta^0$) and deduce that

$$\frac{\partial\bar\theta}{\partial\theta^{1\prime}}\left[P^0,\,\lambda\left(\theta^0\right)\right]=M^{-1}N,\tag{2.15}$$

where

$$M=-E^0\left[\left.\frac{\partial\varphi'}{\partial\theta}(X_t,Y_t,\theta,\lambda\left(\theta^0\right))\right|_{\theta=\theta^0}\right]=E^0\left[\frac{\partial h}{\partial\theta}(X_{1t},\theta^0)\frac{\partial h}{\partial\theta'}(X_{1t},\theta^0)\right]\tag{2.16}$$

and

$$\begin{aligned}N&=E^0\left[\left.\frac{\partial\varphi'}{\partial\theta^1}(X_t,Y_t,\theta^0,\lambda\left(\theta^1\right))\right|_{\theta^1=\theta^0}\right]\\&=E^0\left[\frac{\partial h}{\partial\theta}(X_{1t},\theta^0)E^0\left[\frac{\partial h}{\partial\theta'}(X_{1t},\theta^0)\mid X_{2t}\right]\right]\\&=E^0\left[E^0\left[\frac{\partial h}{\partial\theta}(X_{1t},\theta^0)\mid X_{2t}\right]E^0\left[\frac{\partial h}{\partial\theta'}(X_{1t},\theta^0)\mid X_{2t}\right]\right].\end{aligned}\tag{2.17}$$

Note that M and N are symmetric, positive semidefinite matrices and

$$\begin{aligned}M-N&=Var^0\left[\frac{\partial h}{\partial\theta}(X_{1t},\theta^0)\right]-Var^0\left[E^0\left[\frac{\partial h}{\partial\theta}(X_{1t},\theta^0)\mid X_{2t}\right]\right]\\&=Var^0\left[\frac{\partial h}{\partial\theta}(X_{1t},\theta^0)-E^0\left[\frac{\partial h}{\partial\theta}(X_{1t},\theta^0)\mid X_{2t}\right]\right].\end{aligned}$$

Thus, $M-N$ is a positive semidefinite matrix, which implies that the eigenvalues of the matrix $M^{-1}N$ lie in $[0,1]$. Moreover, it is sufficient to assume that $M-N$ is positive definite to ensure that all these eigenvalues are even smaller than one (see Lemma A.1 in Appendix 7). In other words, the backfitting identification condition is tantamount to the very natural assumption that no linear combination of $\partial h/\partial\theta(X_{1t},\theta^0)$ is (almost surely) a function of X_{2t}. Using this observation and a Taylor expansion argument for the function $h(X_{1t},\theta^1)-E^0\left(h(X_{1t},\theta^1)\mid X_{2t}\right)$, we argue that, at least locally, the backfitting identification condition is not much stronger than the standard identification **C2** of the PPR model.

It is worthwhile to note that the linearity of the regression function h makes unnecessary the backfitting iterations since the equations (2.13) admit the closed form solution

$$\widehat{\theta}_T = \left[\mathsf{X}_1(I_T - K)\mathsf{X}_1'\right]^{-1} \mathsf{X}_1(I_T - K)\mathsf{Y}, \tag{2.18}$$

provided that $\mathsf{X}_1(I_T - K)\mathsf{X}_1'$ is invertible. This formula has been introduced by Green, Jennison and Seheult (1985) and Speckman (1988). In the particular case where the λ function itself is specified as linear, the previous equation amounts to the Frisch-Waugh theorem. Note that in this particular case where the regression function is linear we deduce $E^0\left[\partial\varphi'/\partial\theta\right] = -E^0\left[X_{1t}X_{1t}'\right]$ and

$$E^0\left[\frac{\partial\varphi'}{\partial\theta^1}\right] = E^0\left[X_{1t}X_{2t}'\right]\left(E^0\left[X_{2t}X_{2t}'\right]\right)^{-1} E^0\left[X_{2t}X_{1t}'\right]$$

and therefore

$$\frac{\partial\bar{\theta}}{\partial\theta^{1\prime}}\left[P^0,\, \lambda\left(\theta^0\right)\right] = \left(E^0\left[X_{1t}'X_{1t}\right]\right)^{-1} E^0\left[X_{1t}X_{2t}'\right]\left(E^0\left[X_{2t}X_{2t}'\right]\right)^{-1} E^0\left[X_{2t}X_{1t}'\right].$$

That is, in the Frisch-Waugh case $\partial\overline{\theta}/\partial\theta^{1\prime}$ is what is sometimes called the matricial correlation coefficient between X_1 and X_2. It's eigenvalues represent the squared canonical correlations between X_1 and X_2 (see, e.g., Muirhead (1982), section 11.3).

Designed for a more general framework than the PPR model with linear function h, the latent backfitting methodology which is our focus of interest can be useful in cases where closed form formulae like (2.18) are not available.

For the sake of its extension, we can interpret the standard backfitting setting in terms of latent variables. Let us define the 'latent regression model'

$$Y_t^* = h(X_{1t}, \theta) + u_t.$$

If $\{Y_t^*\} = \left\{Y_t - \lambda^0(X_{2t})\right\}$ were observed, that is if the function λ^0 were known, we could use the latent model and obtain a more precise estimator of θ^0 than in the PPR model. We can say that the classical backfitting looks for a sample analogue of the unique fixed point of the function $\overline{\theta}\left[P^0, \lambda(\cdot)\right]$ defined by

$$\overline{\theta}\left[P^0, \lambda(\theta^1)\right] = \arg\max_{\theta\in\Theta} - E^0\left[Y_t^*(\theta^1) - h(X_{1t}, \theta)\right]^2,$$

where $Y_t^*(\theta^1) = Y_t - \lambda(\theta^1)(X_{2t})$ is the natural guess of the 'latent' variable $Y_t^* = h\left(X_t, \theta^0\right) + u_t$ for a given value θ^1 of the parameters. Indeed, the classical backfitting step for updating $\widehat{\theta}_T$ can be interpreted as a nonlinear regression in the latent model where Y_t^* has been replaced by its guess $Y_t - \widehat{\lambda}_T(X_{2t})$. Below, we will extend this idea to genuine latent variables models.

2.3 Latent regression models and generalized residuals

Consider a latent regression model

$$\begin{cases} Y_t^* = h\left(X_t;\theta\right) + u_t, & E\left[u_t \mid X_t\right] = 0, \\ Y_t = r\left(Y_t^*, X_t;\theta\right), & \theta \in \Theta, \end{cases} \tag{2.19}$$

where, for expositional simplicity, it is assumed that there is only one set of explanatory variables denoted by X_t which are independent of the error term u_t. Moreover, the process $\{u_t, X_t'\}$ is supposed to be stationary. If not stated differently, there is no additional assumption on the joint law of error terms.

A popular example of latent regression model we will consider throughout this section is the binary response model where

$$Y_t = \mathbb{1}_{\{Y_t^*>0\}};$$

($\mathbb{1}_A$ denotes the indicator function of the set A).

In the context of latent regression models, a natural guess of the latent variable Y_t^* deduced from the observations (Y_t, X_t) and a given value θ^1 of the parameters is

$$Y_t^*\left(\theta^1\right) = h\left(X_t, \theta^1\right) + E^0\left[u_t \mid Y_t, X_t; \theta^1\right]. \tag{2.20}$$

Following Gouriéroux *et al.* (1987), the notation $E^0\left[u_t \mid Y_t, X_t; \theta^1\right] = \tilde{u}_t\left(\theta^1\right)$ termed 'generalized residual' means that, given the observation (Y_t, X_t), we compute a forecast of the latent error term u_t from the knowledge of the measurement equation

$$Y_t = r\left[h\left(X_t, \theta^0\right) + u_t, X_t; \theta^0\right] = g\left(u_t, X_t, \theta^0\right)$$

evaluated at the value θ^1 of the unknown parameters. That is,

$$\tilde{u}_t\left(\theta^1\right) = E^0\left[u_t \mid Y_t = y_t\ ,\ X_t = x_t;\ \theta^1\right] = E^0\left[u_t \mid g\left(u_t, x_t, \theta^1\right) = y_t\right]. \tag{2.21}$$

Note that $\tilde{u}_t\left(\theta^1\right)$ and $Y_t^*(\theta^1)$ also depend also on P^0 through the true conditional probability distribution of u_t involved in the last expectation. In the binary response model one has

$$\begin{aligned} \tilde{u}_t\left(\theta^1\right) &= E^0\left[u_t \mid Y_t, X_t; \theta^1\right] \\ &= Y_t E^0\left[u_t \mid u_t > -h\left(X_t; \theta^1\right), X_t\right] + \left(1 - Y_t\right) E^0\left[u_t \mid u_t \leq -h\left(X_t; \theta^1\right), X_t\right], \end{aligned} \tag{2.22}$$

where the two conditional expectations are computed with respect to the true probability distribution of u_t. For the sake of expositional simplicity the true marginal distribution of u is considered as known. See Dominitz and Sherman (2001) for an elegant extension of the generalized residuals to semiparametric binary choice models where this distribution has to be estimated nonparametrically.

Let us note that the forecast provided by the generalized residuals is optimal insofar as the variables (X', u) are serially independent. Nevertheless, we can also consider the case of autoregressive dynamics for the error term u_t as in Robinson (1982) and Gouriéroux, Monfort and Trognon (1985). Following Gouriéroux, Monfort and Trognon (1985), we will always compute the generalized residuals as if there were no serial dependence. This because, even for the simplest models of dependence, the density function of the observables will take the form of a multidimensional integral whose dimension is T the number of observations. This creates serious problems for computing the likelihood function associated to the observations and the optimal forecasts of the latent variables as well. The use of generalized residuals $\tilde{u}_t\left(\theta^1\right)$ and associated latent backfitting has been precisely conceived to overcome these difficulties. The formula (2.21) computing the generalized residuals depends only on the *marginal distribution* of the errors, as well as the fixed point property which ensures consistency of the latent backfitting estimator proposed below.

By analogy with the above interpretation of the classical backfitting, we define the latent backfitting for latent regression models as a search for a sample counterpart of the alleged unique fixed point of the function $\overline{\theta}\left[P^0, \lambda\left(\theta^1\right)\right]$ defined by

$$\overline{\theta}\left[P^0, \lambda\left(\theta^1\right)\right] = \arg\max_{\theta\in\Theta} - E^0\left[Y_t^*(\theta^1) - h(X_t, \theta)\right]^2. \tag{2.23}$$

In other words, we define the latent backfitting by the iterative procedure

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

where

$$\overline{\theta}_T\left(\lambda\left(\theta^{(p)}\right)\right) = \arg\max_{\theta\in\Theta} - \sum_{t=1}^{T}\left[Y_t^*(\theta^{(p)}) - h(X_t, \theta)\right]^2 \qquad p = 1, 2, ...$$

In terms of first order conditions, $\theta_T^{(p+1)}$ is then defined from $\theta_T^{(p)}$ by

$$\sum_{t=1}^{T}\frac{\partial h}{\partial\theta}(X_t, \theta_T^{(p+1)})\left[Y_t^*(\theta_T^{(p)}) - h(X_t, \theta_T^{(p+1)})\right] = 0.$$

In the particular case of a linear latent regression model, that is $h\left(X_t, \theta^1\right) = X_t'\theta^1$, one computes $\theta_T^{(p+1)}$ from the orthogonality conditions between explanatory variables and 'generalized residuals' as first considered by Gouriéroux *et al.* (1987). Dominitz and Sherman (2001) revisited this procedure which they term 'iterative least squares' and characterized its consistency and asymptotic behavior under more general assumptions on the errors distribution.

According to (2.23), the identification condition for this latent backfitting, that is the unique fixed point property of the function $\overline{\theta}\left[P^0, \lambda\left(\cdot\right)\right]$ can be written as

$$\theta^1 = \arg\max_{\theta\in\Theta} - E^0\left[Y_t^*(\theta^1) - h(X_t, \theta)\right]^2 \Longleftrightarrow \theta^1 = \theta^0 \quad .$$

Using the above definition of generalized residuals, we rewrite it as follows.

Condition **C1'** (Backfitting identification)

$$\theta^1 = \arg\max_{\theta\in\Theta} - E^0\left[h(X_t,\theta^1) - h(X_t,\theta) + \tilde{u}_t\left(\theta^1\right)\right]^2 \Longleftrightarrow \theta^1 = \theta^0 .$$

In the case of a linear latent regression function and provided that $E\left[X_t X_t'\right]$ is invertible, we deduce from (2.23)

$$\overline{\theta}(\theta^0, \lambda(\theta^1)) = \left(E\left[X_t X_t'\right]\right)^{-1} E\left[X_t Y_t^*(\theta^1)\right] = \theta^1 + \left(E\left[X_t X_t'\right]\right)^{-1} E\left[X_t \widetilde{u}_t(\theta^1)\right]. \qquad (2.24)$$

In this case the unique fixed point property means

$$E\left[X_t \widetilde{u}_t(\theta^1)\right] = 0 \Longrightarrow \theta^1 = \theta^0. \qquad (2.25)$$

Note that the weaker condition $\theta^0 = \bar{\theta}\left[P^0, \lambda\left(\theta^0\right)\right]$ only means that

$$\theta^0 = \arg\max_{\theta\in\Theta} - E^0\left[h(X_t,\theta^0) - h(X_t,\theta) + \tilde{u}_t(\theta^0)\right]^2,$$

that is

$$\theta^0 = \arg\max_{\theta\in\Theta} - E^0\left[h(X_t,\theta^0) - h(X_t,\theta)\right]^2$$

since

$$E^0\left[\tilde{u}_t\left(\theta^0\right) \mid X_t\right] = E^0\left[E\left[u_t \mid Y_t, X_t, \theta^0\right] \mid X_t\right] = E^0\left[u_t \mid X_t\right] = 0.$$

In other words, while backfitting identification requires that θ^0 is the only fixed point of $\bar{\theta}\left[P^0, \lambda\left(\cdot\right)\right]$, the fact that θ^0 is a fixed point is tantamount to identification in the latent regression model:

Condition **C3'** (Identification in the latent regression model)

$$E^0\left[u_t \mid X_t\right] = 0,$$

and

$$h\left(X_t, \theta^1\right) = h\left(X_t, \theta^0\right) \Longleftrightarrow \theta^1 = \theta^0.$$

Moreover, one can check that, as for standard backfitting, identification in the observable model corresponds to a condition **C2'** which is intermediate between **C1'** and **C3'**. Under very mild technical assumptions on the support of the error distribution and on the form of $h(X_t, \theta)$ (see Appendix 7), the condition **C2'** below amounts to identification from binary observations.

Condition **C2'** (Identification in the observable model)

$$E^0\left[\tilde{u}_t\left(\theta^1\right) \mid X_t\right] = 0 \Longleftrightarrow \theta^1 = \theta^0$$

and

$$h\left(X_t, \theta^1\right) = h\left(X_t, \theta^0\right) \Longleftrightarrow \theta^1 = \theta^0.$$

The condition **C2'** is usually weaker than **C1'**, as it can be seen from the equation (2.25). A more general argument can be the following: if **C1'** holds and one were able to find $\theta^1 \neq \theta^0$ such that

$$E^0\left[\tilde{u}_t\left(\theta^1\right) \mid X_t\right] = 0,$$

then $\theta \to -E^0\left[h\left(X_t, \theta^1\right) - h\left(X_t, \theta\right) + \tilde{u}_t\left(\theta^1\right)\right]^2$ would be maximum for $\theta = \theta^1 \neq \theta^0$ and this contradicts **C1'.**

Generally speaking,

$$\mathsf{C1'} \Longrightarrow \mathsf{C2'} \Longrightarrow \mathsf{C3'},$$

and, by contrast with the standard backfitting case, the linearity of the latent regression model is not sufficient to ensure that **C1'** and **C2'** are equivalent (see also (2.25)). However, the two conditions may be equivalent, for instance under the quite restrictive assumption that, for any θ^1, there exists θ^2 such that

$$h(X_t, \theta^2) = h(X_t, \theta^1) + E\left[\tilde{u}_t\left(\theta^1\right) \mid X_t\right].$$

This additional assumption is satisfied, for example, in the case of a binary choice model with constant latent regression function h. Nevertheless, it is no longer satisfied in the case of a linear latent regression function, that is $E\left[\tilde{u}_t\left(\theta^1\right) \mid X_t\right]$ is not necessarily linear in X_t even if $h(X_t, \theta) = X_t'\theta$.

Let us now study to what extent the backfitting identification condition **C1'** is in this latter case more restrictive than the standard identification condition **C2'**. Define:

$$\varphi(X_t, Y_t, \theta, \lambda\left(\theta^1\right)) = X_t\left[Y_t^*(\theta^1) - X_t'\theta\right]$$

and assume that the function $\bar{\theta}\left[P^0, \lambda\left(\cdot\right)\right]$ can be also defined as the implicit solution of

$$E^0\left[\varphi(X_t, Y_t, \bar{\theta}\left[P^0, \lambda\left(\theta^1\right)\right], \lambda\left(\theta^1\right))\right] = 0, \qquad \theta^1 \in \Theta.$$

Differentiate this identity with respect to θ^1, take $\theta^1 = \theta^0$ and deduce that

$$\frac{\partial\bar{\theta}}{\partial\theta^{1\prime}}\left[P^0, \lambda\left(\theta^0\right)\right] = M^{-1}N,$$

where

$$M = -E^0\left[\left.\frac{\partial\varphi'}{\partial\theta}(X_t, Y_t, \theta, \lambda\left(\theta^0\right))\right|_{\theta=\theta^0}\right] = E^0\left[X_t X_t'\right]$$

and

$$\begin{aligned} N &= E^0\left[\left.\frac{\partial\varphi'}{\partial\theta^1}(X_t, Y_t, \theta^0, \lambda\left(\theta^1\right))\right|_{\theta^1=\theta^0}\right] \\ &= E^0\left[X_t X_t'\right] + E^0\left[X_t \left.\frac{\partial\tilde{u}_t}{\partial\theta^{1\prime}}\left(\theta^1\right)\right|_{\theta^1=\theta^0}\right]. \end{aligned}$$

Next, (2.22) yields:

$$\begin{aligned} & E^{0}\left[\left.\frac{\partial \tilde{u}_{t}}{\partial \theta^{1\prime}}\left(\theta^{1}\right)\right|_{\theta^{1}=\theta^{0}}\right] \\ = & -\left[p^{0} m_{1}^{\prime}\left(-X_{t}^{\prime} \theta^{0}\right)+\left(1-p^{0}\right) m_{2}^{\prime}\left(-X_{t}^{\prime} \theta^{0}\right)\right] X_{t}^{\prime} \end{aligned}$$

where

$$p^{0}=p^{0}\left(X_{t}\right)=E^{0}\left[Y_{t}\left|X_{t}\right.\right], m_{1}(s)=E^{0}\left[u_{t}\left|u_{t}>s, X_{t}\right.\right] \text{ and } m_{2}(s)=E^{0}\left[u_{t}\left|u_{t} \leq s, X_{t}\right.\right].$$

Note that by definition the functions $m_1(s)$ and $m_2(s)$ are non-decreasing and therefore their derivatives m_1' and m_2' are non-negative.

Consequently:

$$N=E^{0}\left\{X_{t}\left[1-g\left(X_{t}, \theta^{0}\right)\right] X_{t}^{\prime}\right\}$$

with:

$$g\left(X_{t}, \theta^{0}\right)=p^{0} m_{1}^{\prime}\left(-X_{t}^{\prime} \theta^{0}\right)+\left(1-p^{0}\right) m_{2}^{\prime}\left(-X_{t}^{\prime} \theta^{0}\right) \geq 0.$$

Thus:

$$\begin{aligned} \frac{\partial \bar{\theta}}{\partial \theta^{1\prime}}\left[P^{0}, \lambda\left(\theta^{0}\right)\right] &= M^{-1} N \\ &= \left\{E^{0}\left[X_{t} X_{t}^{\prime}\right]\right\}^{-1} E^{0}\left\{X_{t}\left[1-g\left(X_{t}, \theta^{0}\right)\right] X_{t}^{\prime}\right\}. \end{aligned}$$

If we assume, without loss of generality, $E^0\left(X_t X_t'\right) = I_p$ then $\partial\bar{\theta}/\partial\theta^{1\prime}\left[P^0, \lambda\left(\theta^0\right)\right]$ is a symmetric matrix. It is well known that if u_t has a strictly log-concave distribution, then $0 \leq m_1'(s)$, $m_2'\ (s) \leq 1$ and the inequalities are strict, except possibly on the boundary of the support of u_t (see Heckman and Honoré (1990)). Therefore, in such a case, $\partial\bar{\theta}/\partial\theta^{1\prime}\left[P^0, \lambda\left(\theta^0\right)\right] = M^{-1}N$ is a positive semidefinite matrix with the norm smaller than one.

In other words, the required contraction mapping condition is satisfied at least for error terms whose marginal distribution belongs to the class of strictly log-concave distribution functions. This class is quite large, including many of the common distributions (normal, logistic, beta, gamma...). Log-concavity is tightly related to well-behaved maximum likelihood equations in the particular case of serial independence of the error terms. Moreover, as stressed by Dominitz and Sherman (2001), the class of marginal distributions of the error terms for which $\partial\bar{\theta}/\partial\theta^{1\prime}\left[P^0, \lambda\left(\theta^0\right)\right] = M^{-1}N$ has a norm smaller than one is even larger than the class of strictly log-concave distributions.

Therefore, we conclude similarly to the standard backfitting case, that the backfitting identification condition (C1') is not much more restrictive than the standard identification condition (C2') of the binary choice model.

Moreover, it is important to realize that the formulation of **C2'** in terms of conditional moment restrictions on the generalized residuals should not lead to a one-step least squares procedure on these generalized residuals:

$$\min_{\theta} \sum_{t=1}^{T} \tilde{u}_{t}^{2}(\theta). \tag{2.26}$$

Note that, (2.26) is akin to the direct maximization of the sample counterpart of $\theta \rightarrow Q_\infty [\theta, \lambda(\theta)]$, where, according to (2.23),

$$Q_\infty \left[\theta, \lambda(\theta^1)\right] = -E^0 \left[Y_t^*(\theta^1) - h(X_t, \theta)\right]^2.$$

We show in Appendix A2 that the binary response model provides a counter-example for the consistency of such a one-step least squares estimator. Therefore, the latent backfitting is really relevant in this setting.

3 Implied state backfitting

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. 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, see Gouriéroux, Monfort and Renault (1993)) have been developed recently. The general latent backfitting methodology we develop in this paper is well-suited for providing appealing alternative estimation procedures based on the latent variables and perform inference directly using the more tractable latent model. We will call *implied state backfitting* the latent backfitting applied to econometric models as described in this section.

3.1 Basic motivations

The vector $\theta = \boldsymbol{\theta}(P)$ of the parameters of interest is now defined by the law of motion P of, say, a stationary and ergodic Markovian of order one process $\{Y_t^*\}$.

The components of Y_t^* are considered as state variables which are not directly observed. The observations, denoted by Y_t, are defined through a known function of Y_t^* and $\lambda^0 \in \Gamma$, the true unknown value of some nuisance parameters. That is, we can write

$$Y_t = g(Y_t^*, \lambda^0), \qquad t \geq 1.$$

What we shall call implied state backfitting applies to versions of the model defined by P and the mapping above which are specific in two respects. First, according to the focus of interest of the paper, we consider that the nuisance parameters are defined as a known function of the parameters of interest, that is $\lambda = \lambda(\theta)$. Second, we assume that the state variables have been defined such that the mapping $g(\cdot, \lambda)$ is one-to-one for any $\lambda \in \Gamma$. Hence,

$$Y_t = g(Y_t^*, \lambda^0) \Longleftrightarrow Y_t^* = g^{-1}(Y_t, \lambda^0), \qquad t \geq 1.$$

It appears that these two characteristics are shared by many structural econometric models. A first example is provided by the option pricing literature. As argued by

Renault (1997), latent state variables such as stochastic volatility, jumps or unobserved short rates represent a convenient way to introduce pricing errors in arbitrage based models. Typically, the known functions $g(\cdot,\cdot)$ and $\lambda(\cdot)$ are provided by the option pricing formula. Another class of applications has been proposed recently by the econometrics of game theoretical models. As stressed by Florens, Protopopescu and Richard (2001), the observed actions of the agents are typically modeled as a functional transformation g of unobserved variables (often referred to as signals or types). Moreover, due the equilibrium form of the game, such transformations (strategies) depend on the distribution of the unobservables.

As in the examples from the previous section, the nuisance parameters λ will be linked to the 'learning problem': starting from an extremum estimation principle in the latent world, one has to replace the latent value Y_t^* in the corresponding criterion function by a guess $g^{-1}(Y_t,\lambda)$ computed from the observation Y_t for a given value of λ. However, the key difference with respect to the previous examples is related to the nature of the loss of statistical information. Until now this information loss when passing from the latent 'statistic' Y_t^* to the observable one Y_t was due, essentially, to a genuine missing data problem, that is the mapping between Y_t^* and Y_t was not one-to-one. This time the value Y_t^* is not really 'missing' since it is uniquely defined from the observations as $g^{-1}(Y_t,\lambda^0)$. The problem is, of course, that this formula depends upon unknown parameters and therefore the informational contents of the latent and observable worlds are different by nature.

For illustration purposes, let us compute the statistical information associated with a set of moment restrictions. Consider H moment conditions in the latent world defined by the $H-$dimensional function $\psi\left(Y_t^*,\theta\right)$, $\theta\in\Theta\subset\mathsf{R}^p$, that is

$$E^0\left[\psi\left(Y_t^*,\theta\right)\right]=0\Longleftrightarrow\theta=\theta^0 \tag{3.1}$$

with

$$Rank\,E^0\left[\frac{\partial\psi}{\partial\theta'}\left(Y_t^*,\theta\right)\Big|_{\theta=\theta^0}\right]=p.$$

If one wants to exploit these moment restrictions to perform inference about θ^0 from the observations Y_t, $t\geq 1$, one has to define implied states $Y_t^*=g^{-1}(Y_t,\lambda\left(\theta^0\right))$ and consider moment conditions of the form

$$E^0\left[\varphi\left(Y_t,\theta,\lambda\left(\theta\right)\right)\right]=0, \tag{3.2}$$

where

$$\varphi\left(Y_t,\theta,\lambda\left(\theta\right)\right)=\psi\left(g^{-1}(Y_t,\lambda\left(\theta\right));\theta\right).$$

Pan (2002) termed *implied states GMM* (IS-GMM) the method of moments estimation of θ deduced from (3.2). The same approach was previously considered by Florens, Protopopescu and Richard (2001).

First of all, it is important to realize that the identification condition (3.1) does not imply that θ^0 can be automatically recovered from the moment conditions (3.2). To emphasize this, one can even imagine extreme examples as below where the probability distribution of the function $\varphi\left(Y_t;\theta,\lambda\left(\theta\right)\right)$ no longer depends on θ. In other words, it may happen that there is complete information loss when passing from the latent world to the observable one.

Example 1 (*An impossibility example for implied state GMM*)
Let

$$\psi\left(Y_t^*;\theta\right)=Y_t^*-\theta=g(Y_t^*,\lambda\left(\theta\right)).$$

Then

$$\varphi\left(Y_t;\theta,\lambda\left(\theta\right)\right)=g^{-1}(Y_t,\lambda\left(\theta\right))-\theta=\left(Y_t+\theta\right)-\theta=Y_t$$

and

$$E^0\left[\varphi\left(Y_t;\theta,\lambda\left(\theta\right)\right)\right]=E^0\left[Y_t\right]=0,$$

for any value of θ.

Another closely related aspect to observe is that there is no general relationship between the informational contents of the two worlds. In the case of the moment conditions (3.1) the corresponding semiparametric efficiency bound is the inverse of the matrix

$$I^*\left(\theta^0\right)=E^0\left[\frac{\partial\psi'}{\partial\theta}\left(Y_t^*;\theta\right)\bigg|_{\theta=\theta^0}\right]\Omega\left(\theta^0\right)^{-1}E^0\left[\frac{\partial\psi}{\partial\theta'}\left(Y_t^*;\theta\right)\bigg|_{\theta=\theta^0}\right],$$

where

$$\Omega\left(\theta^0\right)=\lim_{T\to\infty}Var^0\left[\frac{1}{\sqrt{T}}\sum_{t=1}^{T}\psi\left(Y_t^*;\theta^0\right)\right].$$

On the other hand, if implied state GMM can be performed, that is θ^0 can be identified from the equations (3.2), the associated semiparametric efficiency bound is the inverse of

$$I\left(\theta^0\right)=E^0\left[\partial_\theta\varphi'\left(Y_t;\theta,\lambda\left(\theta\right)\right)|_{\theta=\theta^0}\right]\Omega\left(\theta^0\right)^{-1}E^0\left[\partial_\theta\varphi\left(Y_t;\theta,\lambda\left(\theta\right)\right)|_{\theta=\theta^0}\right],\qquad(3.3)$$

where

$$\partial_\theta\varphi'=\left(\partial_\theta\varphi\right)'=\frac{\partial\varphi'}{\partial\theta}+\frac{\partial\varphi'}{\partial\lambda}\frac{\partial\lambda'}{\partial\theta}=\frac{\partial\psi'}{\partial\theta}+\frac{\partial\varphi'}{\partial\lambda}\frac{\partial\lambda'}{\partial\theta}.$$

In view of the form of $\partial_\theta\varphi$, it is clear that there is no predetermined order between $I\left(\theta^0\right)$ and $I^*\left(\theta^0\right)$. Moreover, one has to remember that the respective roles of the latent world and the observable one could be swapped, since the mapping linking Y_t to Y_t^* is one-to-one. These ideas are easily supported by the following example.

Example 2 (*About the information loss*)
Let

$$\psi\left(Y_t^*;\theta\right)=Y_t^*-\theta,\qquad Y_t^*\in\mathsf{R},\quad\theta\in\mathsf{R},$$

and

$$g(Y_t^*,\lambda\left(\theta\right))=Y_t^*-\lambda\left(\theta\right)$$

with $\lambda(\cdot)$ some differentiable function. Then we have

$$\varphi\left(Y_t;\theta,\lambda\left(\theta^1\right)\right)=Y_t+\lambda\left(\theta^1\right)-\theta,$$

$$I^*(\theta)=Var^0(Y_t^*)^{-1}$$

and

$$I(\theta)=I^*(\theta)\left[\lambda'(\theta)-1\right]^2,$$

where λ' denotes here the derivative of λ. It is clear that $I^*(\theta)>I(\theta)$ if and only if $0<\lambda'(\theta)<2$.

To summarize, there is no universally valid argument to prefer the latent "data" to the observable ones. Nevertheless, a deep motivation behind the data augmentation paradigm in econometrics is that it allows for characterizations of the probability distributions that are preferable, both in terms of interpretation and computation, to the 'observable models' conceived as direct descriptions of the DGP. This is the reason why, when a one-to-one relationship between the two statistics exists, one will typically prefer to use this coming and going vehicle for learning about the state variables Y_t^* and to base inference on this learning. In the following, two classes of inference methodologies will illustrate this general principle: generalized method of moments and maximum likelihood. First, we revisit the implied state GMM as a methodology closely related to our latent backfitting and we propose an alternative implied state backfitting estimator.

3.2 From IS-GMM to implied state backfitting for GMM

The latent moment restrictions (3.2) define the extremum estimation criterion

$$Q_T[\theta,\lambda]=-\left[\frac{1}{T}\sum_{t=1}^{T}\varphi(Y_t;\theta,\lambda)\right]W_T\left[\frac{1}{T}\sum_{t=1}^{T}\varphi(Y_t;\theta,\lambda)\right].$$

For the sake of expositional simplicity, we will always consider hereafter that W_T is a consistent estimator of the weighting matrix $\Omega\left(\theta^0\right)^{-1}$, the optimal weighting matrix for the case $\lambda=\lambda^0$. Let

$$\widehat{\theta}_T^*=\arg\max_{\theta\in\Theta}Q_T\left[\theta,\lambda^0\right]$$

define the oracle (infeasible) estimator of θ that would be obtained if latent variables were observed. Its asymptotic variance is the inverse of the matrix

$$I^*\left(\theta^0\right)=\left.\frac{\partial Q_\infty}{\partial\theta\partial\theta'}\left[\theta,\lambda^0\right]\right|_{\theta=\theta^0},$$

where

$$Q_\infty[\theta,\lambda]=-E^0\left[\varphi(Y_t;\theta,\lambda)\right]'\Omega\left(\theta^0\right)^{-1}E^0\left[\varphi(Y_t;\theta,\lambda)\right]. \tag{3.4}$$

The aforementioned guiding principle leads one to define, as Florens, Protopopescu and Richard (2001) and Pan (2002) have already done, an implied states GMM estimator

$$\widehat{\theta}_T^{IS} = \arg\max_{\theta \in \Theta} Q_T \left[\theta, \lambda(\theta)\right].$$

Under standard regularity conditions, this estimator is consistent insofar as the moment conditions (3.2) do identify θ^0. In other words, the non-adaptivity problem in its form (2.5) is not an issue in the GMM setting.

The asymptotic variance of $\widehat{\theta}_T^{IS}$ is given by the inverse of $I\left(\theta^0\right)$ written in (3.3). We argue that this estimator can be fruitfully interpreted as a particular application of the general latent backfitting methodology.

As explained in subsection 2.1, a central piece of our inference approach is the function $\overline{\theta}\left[P^0, \cdot\right]$ defined here as

$$\overline{\theta}\left[P^0, \lambda\right] = \arg\max_{\theta \in \Theta} Q_\infty \left[\theta, \lambda\right]$$

with $Q_\infty\left[\theta, \lambda\right]$ defined as in (3.4). We will maintain in this subsection the following additional assumption, a strengthened version of the general Assumption 2.2 stating the existence of $\overline{\theta}\left[P^0, \cdot\right]$.

Assumption 3.1 *For any* $\lambda \in \Gamma$,

$$E^0\left[\varphi\left(Y_t; \bar{\theta}\left[P^0, \lambda\right], \lambda\right)\right] = E^0\left[\psi\left(g^{-1}(Y_t, \lambda); \bar{\theta}\left[P^0, \lambda\right]\right)\right] = 0.$$

Note that this assumption is innocuous in the case of a just identified GMM estimator as, for instance, considered in Pan (2002). In view of the Assumption 3.1, the identification condition imposed by the Assumption 2.3 means

$$E^0\left[\psi\left(g^{-1}(Y_t, \lambda\left(\theta^1\right)); \theta^1\right)\right] = 0 \implies \theta^1 = \theta^0.$$

This is nothing but the identification condition ensuring that the implied state GMM estimator makes sense.

Since, for any θ^1, we have by definition

$$E^0\left[\varphi\left(Y_t; \bar{\theta}\left[P^0, \lambda\left(\theta^1\right)\right], \lambda\left(\theta^1\right)\right)\right] = 0,$$

assuming the needed regularity conditions, we can derive this identity with respect to θ^1. For $\theta^1 = \theta^0$ this yields

$$E^0\left[\frac{\partial \varphi}{\partial \theta'}\left(Y_t; \theta^0, \lambda\left(\theta^0\right)\right)\right] \frac{\partial \bar{\theta}}{\partial \theta^{1\prime}}\left[P^0, \lambda\left(\theta^0\right)\right] + E^0\left[\frac{\partial \varphi}{\partial \theta^{1\prime}}\left(Y_t; \theta^0, \lambda\left(\theta^0\right)\right)\right] = 0. \tag{3.5}$$

This allows us to interpret the difference between the asymptotic variances I^{*-1} and I^{-1} of the oracle GMM and the implied state GMM, respectively. Since, computed for $\theta = \theta^0$,

$$E^0\left[\partial_\theta \varphi\right] = E^0\left[\frac{\partial \varphi}{\partial \theta'} + \frac{\partial \varphi}{\partial \theta^{1\prime}}\right] = E^0\left[\frac{\partial \varphi}{\partial \theta'}\right]\left(I_p - \frac{\partial \bar{\theta}}{\partial \theta^{1\prime}}\right),$$

we have from (3.3)

$$I^{-1} = \left(I_p - \frac{\partial \bar{\theta}}{\partial \theta^{1\prime}} \right)^{-1} I^{*-1} \left(I_p - \frac{\partial \bar{\theta}'}{\partial \theta^{1}} \right)^{-1}. \tag{3.6}$$

Note that in order to get the asymptotic distribution of the IS-GMM estimator, one needs to assume $Rank\, E^0 \left[\partial_\theta \varphi\right] = p$, which implies that $I_p - \partial\bar{\theta}/\partial\theta^{1\prime}$ is a nonsingular matrix. As already mentioned, this is akin to the unique fixed property of $\bar{\theta}\left[P^0,\ \lambda\left(\cdot\right)\right]$ in a neighborhood of θ^0.

Remark that the interpretation of IS-GMM estimators through the latent backfitting approach allows also for better understanding of the definition of some estimators of this type proposed in the literature. Consider a set of H conditional moment restrictions that are directly formulated in terms of latent variables:

$$E^0\left[\Psi^c\left(Y_t^*, \theta\right) \left| Y_{t-1}^*\right.\right] = 0.$$

Thanks to the one-to-one relationship between Y_t^* and Y_t, these restrictions are tantamount to H conditional moment restrictions about observable variables:

$$E^0\left[\varphi^c\left(Y_t, \theta, \lambda(\theta)\right) \left| Y_{t-1}\right.\right] = 0.$$

To perform efficient GMM estimation, one would like to use a set of optimal instruments the computation of which involves the derivative of $\varphi^c\left(Y_t, \theta, \lambda(\theta)\right)$ with respect to both occurrences of θ. However, as already noted by Duan (1994), "it will be better to avoid the direct computation of the Jacobian for the inverse transformation", a natural idea suggested by the latent backfitting methodology is to compute the optimal instruments corresponding to the oracle GMM conditions

$$E^0\left[\varphi^c\left(Y_t, \theta, \lambda(\theta^0)\right) \left| Y_{t-1}\right.\right] = 0$$

and use them as "almost optimal" instruments for the original moment restrictions

$$E^0\left[\varphi^c\left(Y_t, \theta, \lambda(\theta)\right) \left| Y_{t-1}\right.\right] = 0.$$

Pan (2002) acknowledges this tension by noting that "the efficiency loss of this "optimal-instrument" scheme is limited in that (...) we sacrifice efficiency by ignoring the dependence of $\lambda\left(\theta\right)$ on θ". However, since the above examples 1 and 2 have shown that informational and related efficiency issues may be highly sensitive to the dependence of $\lambda\left(\theta\right)$ on θ, it is more cautious to keep in mind the two ingredients of the accuracy of any implied states estimator. As shown in (3.6) , the contracting feature of the mapping $\bar{\theta}$ may matter even more than the semiparametric efficiency bound of the latent model. In other words, the so-called "limited optimality" that is alleged by reference to the optimal instruments computed from the latent moment conditions cannot be the only criterion to assess the accuracy of the implied states estimator. It is worth interpreting it as the limit of a backfitting algorithm and to assess its accuracy through the strength of the contraction at play in this algorithm.

In the context of Assumption 3.1, the IS-GMM estimator $\widehat{\theta}_T^{IS}$ can be interpreted as the unique fixed point of the finite sample analogue of $\bar{\theta}\left[P^0,\ \lambda\left(\cdot\right)\right]$, that is

$$\bar{\theta}_T\ \left(\lambda\left(\theta^1\right)\right) = \arg\max_{\theta}\ Q_T\left[\theta, \lambda\left(\theta^1\right)\right]. \tag{3.7}$$

Starting from this interpretation, we propose a competitor GMM estimator provided by the general latent backfitting methodology. Consider $\varphi\left(Y_t; \theta, \lambda\left(\theta\right)\right)$ a R^H-valued function defining moment restrictions as in (3.2) and let $Q_\infty\left[\theta, \lambda\left(\theta\right)\right]$ be as in (3.4). Moreover, $Q_T\left[\theta, \lambda\left(\theta\right)\right]$ denotes its empirical version of the GMM criterion. The implied-state GMM backfitting estimator we propose is the estimator $\theta^{p(T)+1}$ obtained from a finite number of iterations

$$\theta^{(p+1)} = \arg\max\ Q_T\left[\theta, \lambda(\theta^{(p)})\right] = \bar{\theta}_T\ \left(\lambda\left(\theta^{(p)}\right)\right), \qquad p = 1, 2, ...p(T).$$

Typically, if the function $\bar{\theta}_T\ \left(\lambda\left(\cdot\right)\right)$ admits a unique fixed point, the IS-GMM estimator $\widehat{\theta}_T^{IS}$ coincides with the limit of the iterations ($p(T) = \infty$).

We argue that our estimator has some interesting features. It can be easily deduced from (3.6) and the general asymptotic results below that it has the same asymptotic variance as the IS-GMM estimator. The IS-GMM estimator is computed by optimization from a criterion $\theta \to Q_T\left[\theta, \lambda(\theta)\right]$ which may be very flat. Our estimator can be computed by iteration from objective functions of the form $\theta \to Q_T\left[\theta, \lambda(\theta^1)\right]$, with θ^1 fixed, which can be much easier to work with. Moreover, we only maintain the contraction mapping property for the limit function $\bar{\theta}\left[P_0, \lambda\left(\cdot\right)\right]$.

The basic assumption for latent backfitting estimation $\left\|\partial\bar{\theta}/\partial\theta^{1\prime}\left[P^0,\ \lambda\left(\theta^0\right)\right]\right\| < 1$ can be analyzed through the identity (3.5) which fully characterizes $\partial\bar{\theta}/\partial\theta^{1\prime}$ since, in general, it is assumed that the matrix $E^0\left[\partial\varphi/\partial\theta'(Y_t, \theta^0, \lambda^0)\right]$ is of full column-rank. Our basic assumption means, intuitively, that the moment restrictions are more informative about θ through its first occurrence than through the second one.

Before closing this subsection let us point out that the interpretation of the IS-GMM and the definition of our implied state backfitting estimator is valid for any oracle moment estimator based on oracle moment conditions

$$E^0\left[\varphi\left(Y_t; \theta, \lambda\left(\theta^0\right)\right)\right] = 0$$

considered as restrictions about θ unknown (for $\lambda\left(\theta^0\right)$ given). These restrictions are not necessarily underpinned by a one-to-one mapping related to a latent data set.

3.3 Implied state backfitting for latent likelihood

Recall that $\{Y_t^*\} \subset \mathsf{R}^J$ denotes a sequence of homogeneous Markovian of order one random vectors. We specify a $p-$dimensional parametric model for the probability distribution of the process $\{Y_t^*\}$ through the family of transition densities

$$\mathcal{M}^* = \{f^*(\cdot \mid \cdot\,; \theta), \quad \theta \in \Theta \subset \mathsf{R}^p\}$$

defined with respect to the Lebesgue measure on 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 $\theta^0 \in \Theta$

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

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

The infeasible maximum likelihood estimator in the latent world, denoted by $\widehat{\theta}_T^*$, is:

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

where $l^*(\cdot \mid \cdot\,; \theta) = \log f^*(\cdot \mid \cdot\,; \theta)$, $\theta \in \Theta$.

As already noted, the particularity of the implied state framework lies in the observation scheme. The observed data Y_t, $1 \leq t \leq T$ are given by a known one-to-one transformation $Y_t = g(Y_t^*, \lambda(\theta^0))$ depending on the true unknown parameter. The maximum likelihood procedure in the observable world would maximize the criterion

$$\tilde{Q}_T[\theta, \lambda(\theta)] = \frac{1}{T}\sum_{t=1}^{T} l^*\left(g^{-1}(Y_t, \lambda(\theta)) \mid g^{-1}(Y_{t-1}, \lambda(\theta))\ ; \theta\right) + \frac{1}{T}\sum_{t=1}^{T} \log\left|J_y g^{-1}(Y_t, \lambda(\theta))\right|. \tag{3.8}$$

We denote by $|J_y g^{-1}(\cdot, \lambda(\theta))|$ the absolute value of the Jacobian with respect to y of the one-to-one mapping $y \to g^{-1}(y, \lambda(\theta))$. We have again in mind, following Duan (1994), the case where this criterion is very complicated. This happens for instance in stock option pricing models with stochastic volatility and jump components in the stock dynamics. In such a framework, Y_t^* represents a vector of unobservable state variables while Y_t is a vector of observed option prices, at time t. The relationship $g(\cdot, \lambda(\theta^0))$ is provided by arbitrage-based derivative asset pricing à la Harrison and Kreps (1979). The observable model behind (3.8) was called by Christensen (1992) "the empirical martingale model". Renault and Touzi (1996) focused on the at-the-money (ATM) option case. They considered $\{Y_t^*\}$ as being the (latent) volatility process of the underlying asset and $\{Y_t\}$ 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 $g(\cdot, \lambda(\theta))$ between the two worlds, latent and observable. Typically, the dynamics of the latent process $\{Y_t^*\}$ 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 well behaved. Meanwhile, the observable log-likelihood is cumbersome to maximize, since it involves highly nonlinear functions of the unknown parameters. Facing such issues, Renault and Touzi (1996) (see also Renault (1997)) proposed an iterative estimation procedure which represented the starting point of this paper.

In this framework, we define

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

the criterion to be used for defining the ML based implied state backfitting estimation. Its population counterpart is

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

where the expectation is considered with respect to the stationary distribution of (Y_1, Y_0), characterized by θ^0 and

$$\bar{\theta} \left[P^0, \lambda(\theta^1) \right] = \arg\max_{\theta \in \Theta} E^0 \left[l^* \left(g^{-1}(Y_1, \lambda(\theta^1)) \mid g^{-1}(Y_0, \lambda(\theta^1)) \, ; \theta \right) \right].$$

Using the general arguments presented above in a GMM framework, we remark that the implied state backfitting estimation procedure could not be immediately 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. We can actually interpret our estimator as

$$\bar{\theta} \left[P^0, \lambda(\theta^1) \right] = \arg\max_{\theta \in \Theta} E^0 \left[l^* \left(g^{-1}(Y_1, \lambda(\theta^1)) \mid g^{-1}(Y_0, \lambda(\theta^1)) \, ; \theta \right) + \log \left| J_y g^{-1}(Y_1, \lambda(\theta^1)) \right| \right].$$

for fixed θ^1 while observable maximum likelihood would consist in writing $\theta^1 = \theta$ and maximizing simultaneously with respect to the four occurrences of θ.

As in the examples of the previous section, let us investigate the relationship between the unique fixed point property for the function $\bar{\theta} \left[P^0, \lambda(\cdot) \right]$ and the identification in the observable model

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

The example below shows that, in general, the implied state latent backfitting identification condition is not necessary for ensuring the identification in the observable model (similar to the regression examples of section 2).

Example 3

Consider a family of exponential probability density functions

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

with $\theta \in \Theta = [1/2, 3/2]$. Let $\{Y_t^*\}$ be a sequence of i.i.d. random variables and 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

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

In this case

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

and

$$\bar{\theta} \left[P^0, \lambda(\theta^1) \right] = \left(\frac{1}{\theta^0} + \theta^0 - \theta^1 \right)^{-1}.$$

As θ^0 and $1/\theta^0$ are both fixed points for $\bar{\theta}\left[P^0, \lambda(\cdot)\right]$, 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 $\{Y_t\}$ is $[\theta^0, \infty)$.

The previous example shows that the estimation strategy we propose may not be applicable, whereas the ML estimation based on the full criterion (3.8) 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 practically not computable. Meanwhile, the implied state backfitting we considered above applies. For example, Renault and Touzi (1996) reported simulations results showing that the unique fixed point property of $\bar{\theta}\left[P^0, \lambda(\cdot)\right]$ should be satisfied in the Hull and White (1987) option pricing model.

4 Iterative extremum estimation

In this section we present a general iterative latent backfitting estimator and we study its asymptotic properties (see also Patilea and Renault (1997) and Renault (1997)). For this purpose, recall that $\overline{\theta}_T\left(\lambda\left(\theta^1\right)\right)$, $T \geq 1$ is defined as the sample counterpart of $\overline{\theta}\left[P^0, \lambda\left(\theta^1\right)\right]$, that is the maximizer of the criterion $Q_T\left[\cdot, \lambda\left(\theta^1\right)\right]$ introduced in section 2.1 (see (3.7)). Note that $\overline{\theta}_T(\lambda(\theta^0))$ is nothing else than the oracle extremum (argmax) estimator defined through the simple criterion $Q_T\left[\theta, \lambda(\theta^0)\right]$.

Given the criterion $Q_T\left[\cdot, \lambda\left(\cdot\right)\right], T \geq 1$ and a sequence $\{p(T)\}$ of positive integers such that $p(T) \to \infty$, the corresponding iterative latent backfitting estimator we consider is

$$\widehat{\theta}_T = \theta_T^{(p(T)+1)}, \;\; T \geq 1, \tag{4.1}$$

where, for any $p \geq 1$

$$\theta_T^{(p+1)} = \overline{\theta}_T\left(\lambda(\theta_T^{(p)})\right) \tag{4.2}$$

and $\theta_T^{(1)} \in \Theta$ is some starting value.

The iterative latent backfitting estimator represents an extension of the estimator introduced by Renault and Touzi (1996) in the case of a log-likelihood type criterion and $p(T) \equiv \infty$.

4.1 Consistency

To proving consistency, first we have to ensure the uniform convergence of $\overline{\theta}_T\left(\lambda\left(\theta^1\right)\right)$ towards $\overline{\theta}\left[P^0, \lambda\left(\theta^1\right)\right]$. For this we impose the following uniform convergence assumption on $Q_T\left[\theta, \lambda\left(\theta^1\right)\right]$.

Assumption 4.1 *If $Q_T\left[\cdot, \lambda(\cdot)\right]$, $T \geq 1$ and $Q_\infty\left[\cdot, \lambda(\cdot)\right]$ are defined as in section 2.1, then*

$$\sup_{\theta, \theta^1 \in \Theta} \left|Q_T\left[\theta, \lambda\left(\theta^1\right)\right] - Q_\infty\left[\theta, \lambda\left(\theta^1\right)\right]\right| \xrightarrow{p} 0.$$

Such a uniform convergence property can be obtained under quite general conditions on the parameter space Θ and the data generating process (see, *e.g.,* Davidson (1994), Andrews (1994b), van de Geer (2000)). The convergence results of this section rely on the following proposition which we state for a general parameter space. For the sake of expositional simplicity, here and in the rest of the paper, we assume there is no problem of measurability with the quantities we manipulate.

Proposition 4.2 *Assume that Θ is a compact subset of a normed space $\left(\widetilde{\Theta}, \|\cdot\|\right)$. Moreover, $\overline{\theta}\left[P^0, \lambda(\cdot)\right] : \Theta \longrightarrow \Theta$ is continuous. If Assumption 2.1, 2.2 i) and 4.1 hold, then*

$$\sup_{\theta^1 \in \Theta} \left\|\overline{\theta}_T\left(\lambda\left(\theta^1\right)\right) - \overline{\theta}\left[P^0, \lambda\left(\theta^1\right)\right]\right\| \xrightarrow{p} 0.$$

Following Patilea and Renault (1997) and Renault (1997) (see also Dominitz and Sherman (2001)) we impose a contracting assumption on the $\overline{\theta}$ function, a reinforcement of the unique fixed point condition stated in Assumption 2.3.

Assumption 4.3 *The parameter set Θ is a subset of a normed space $\left(\widetilde{\Theta}, \|\cdot\|\right)$. The mapping*

$$\overline{\theta}\left[P^0, \lambda(\cdot)\right] : \Theta \longrightarrow \Theta$$

is contracting on Θ, that is there exists a constant $c \in [0, 1)$ such that, for any $\theta^1, \theta^2 \in \Theta$

$$\left\|\overline{\theta}\left[P^0, \lambda\left(\theta^1\right)\right] - \overline{\theta}\left[P^0, \lambda\left(\theta^2\right)\right]\right\| \leq c\left\|\theta^1 - \theta^2\right\|.$$

Note that the basic identification condition $\overline{\theta}\left[P^0, \lambda\left(\theta^0\right)\right] = \theta^0$ together with the contracting property ensures the unique fixed point property for θ^0. We are now able to state the weak (in probability) consistency result for the iterative estimator we propose.

Proposition 4.4 *Assume that Θ is a compact subset of a normed space $\left(\widetilde{\Theta}, \|\cdot\|\right)$. If Assumption 2.1, 2.2 i), 4.1 and 4.3 hold, then $\widehat{\theta}_T$, $T \geq 1$ defined in (4.1)-(4.2) with $P(T) \to \infty$ is weakly consistent.*

Let us discuss the particular case of affine mappings $\overline{\theta}_T\left(\lambda\left(\cdot\right)\right), T \geq 1$ and $\bar{\theta}\left[P^0, \lambda(\cdot)\right]$. For simplicity, consider the case of Euclidean parameters. If $\overline{\theta}_T\left(\lambda\left(\theta^1\right)\right) = B_T\theta^1 + a_T$, then the fixed points of this function verify $\left(I_p - B_T\right)\theta = a_T$. There exists only one fixed point if $I_p - B_T$ is invertible. For a given sample, the iterations in (4.2) converge to $\left(I_p - B_T\right)^{-1} a_T$ if and only if B_T is convergent. If $\bar{\theta}\left[P^0, \lambda(\theta^1)\right] = B\theta^1 + a$, the unique fixed point property is tantamount to $I_p - B$ invertible and $\theta^0 = \left(I_p - B\right)^{-1} a$. Note that the uniform convergence of $\overline{\theta}_T\left(\lambda\left(\cdot\right)\right)$ to $\bar{\theta}\left[P^0, \lambda(\cdot)\right]$ means here $B_T \to B$ and $a_T \to a$. In this particular case, the estimator $\widehat{\theta}_T$ can be defined as a fixed point of $\overline{\theta}_T\left(\lambda\left(\cdot\right)\right)$ that is with $p(T) = \infty$ for given T. It converges to θ^0 in probability provided that Assumption 4.1 holds and $I_p - B$ is invertible. For an example where $\overline{\theta}_T\left(\lambda\left(\cdot\right)\right), T \geq 1$ and $\bar{\theta}\left[P^0, \lambda(\cdot)\right]$ are affine see the classical backfitting in partially parametric models (see section 2.2). We will discuss in the next subsection the gain in generality in non-linear contexts resulting from the possibility of choosing $p(T)$ finite for any given sample size T.

4.2 Asymptotic distribution

In in this section we derive the limit distribution of a latent backfitting estimator as defined in (4.1). Before stating our results let us introduce some additional hypotheses (see, e.g., Newey and McFadden (1994), Wooldridge (1994)). The set Θ is assumed to be a subset of some Euclidean space R^p, $p \geq 1$ and θ^0 is an interior point of Θ.

Assumption 4.5 *If θ^0 is the true unknown value of the parameters, then*

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

with

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

which is supposed to be positive definite.

The asymptotic normality of the score of the simplified criterion $Q_T[\cdot, \lambda(\theta^0)]$ is an usual assumption for extremum 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 assumption below.

Assumption 4.6 *For any $\theta^1 \in \Theta$, $Q_T[\cdot, \lambda(\theta^1)], T \geq 1$ and $Q_\infty[\cdot, \lambda(\theta^1)]$ are twice continuously differentiable. For any $\theta^1, \theta \in \Theta$, define the matrix*

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

and assume that $\Sigma(\cdot, \cdot)$ is continuous and $\Sigma(\theta^0, \theta^0)$ is positive definite. Moreover,

$$\sup_{\theta,\, \theta^1 \in \Theta} \left\| \frac{\partial^2 Q_T}{\partial\theta\partial\theta'}\left[\theta, \lambda(\theta^1)\right] + \Sigma(\theta, \theta^1) \right\| \longrightarrow 0 \tag{4.3}$$

in probability when $T \to \infty$.

Note that $-\Sigma(\theta^0, \theta^0)$ is nothing else than the Hessian matrix, considered for $\theta = \theta^0$, of the simple limit criterion $Q_\infty[\cdot, \lambda(\theta^0)]$. This matrix is usually assumed to be negative definite. The next assumption is more specific to nuisance parameters framework.

Assumption 4.7 *The functions $Q_\infty[\cdot, \lambda(\cdot)]$ and $Q_T[\cdot, \lambda(\cdot)], T \geq 1$ are twice continuously differentiable. Define the matrices*

$$H(\theta^1) = \frac{\partial^2 Q_\infty}{\partial\theta\partial\theta^{1\prime}}\left[\theta^0, \lambda(\theta^1)\right], \qquad H_T(\theta^1) = \frac{\partial^2 Q_T}{\partial\theta\partial\theta^{1\prime}}\left[\theta^0, \lambda(\theta^1)\right], \quad T \geq 1$$

$\theta^1 \in \Theta$. Then,

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

in probability, as $T \to \infty$.

The respective "orders of magnitude" of the two matrices $H(\theta^0)$ and $\Sigma\left(\theta^0, \theta^0\right)$ determine to what extent the non-adaptivity problem matters, or equivalently, the strength of the contraction feature of the mapping $\bar{\theta}\left[P^0, \lambda(\cdot)\right]$. To see this, let us differentiate the identity:

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

and deduce:

$$\begin{aligned}\frac{\partial\overline{\theta}}{\partial\theta^{1\prime}}\left[P^0,\lambda(\theta^0)\right] &= \Sigma(\theta^0,\theta^0)^{-1}H(\theta^0) \qquad (4.4)\\ &= -\left[\frac{\partial^2 Q_\infty}{\partial\theta\partial\theta'}\left[\theta^0,\lambda(\theta^0)\right]\right]^{-1}\frac{\partial^2 Q_\infty}{\partial\theta\partial\lambda'}\left[\theta^0,\lambda^0\right]\frac{\partial\lambda}{\partial\theta^{1\prime}}\left(\theta^0\right).\end{aligned}$$

In other words, our maintained contraction mapping assumption 4.3, which is locally equivalent to the condition:

$$\left\|\frac{\partial\overline{\theta}}{\partial\theta^{1\prime}}\left[P^0,\lambda(\theta^0)\right]\right\| < 1 \qquad (4.5)$$

puts an upper bound on the non-adaptivity problem we can be faced with. It is well known (see, *e.g.* Wooldridge (1994)) that the case where the cross-derivative matrix $\partial^2 Q_\infty/\partial\theta\partial\lambda'$ vanishes at (θ^0,λ^0) is precisely the case where an extremum estimator

$$\widetilde{\theta}_T = \arg\max_{\theta\in\Theta} Q_T\left[\theta,\widetilde{\lambda}_T\right]$$

has an asymptotic probability distribution which does not depend upon the choice of a $\sqrt{T}-$consistent estimator $\widetilde{\lambda}_T$ of λ^0. We are now saying that, perhaps the cross-derivative matrix is not zero but it is "sufficiently small" to be sure that:

$$\left\|\Sigma\left(\theta^0,\theta^0\right)^{-1}H\left(\theta^0\right)\right\| < 1 \qquad (4.6)$$

In some sense, this means that the occurrence of θ in the latent model (the transition equation of the state variables) carries more information about the unknown parameters of interest than their occurrence in the measurement equation. This is conformable to the spirit of standard asset pricing models and auction models as well.

Of course, smaller is this matrix, more contracting is the mapping $\overline{\theta}\left[P^0,\lambda(\cdot)\right]$, smaller is the efficiency loss of our backfitting estimator with respect to the oracle estimator. This is the main message of proposition 4.8 below:

Proposition 4.8 *Assume that θ^0, the true unknown value of the parameters, is an interior point of $\Theta\subset \mathsf{R}^p$. Consider that Assumptions 2.1, 2.2 i), 4.1, and 4.5 to 4.7 hold. Moreover, suppose that $\overline{\theta}\left[P^0,\lambda(\cdot)\right]$ is contracting on Θ. If, in addition, the sequence $p(T)$, $T\geq 1$, considered in (4.1) is such that*

$$\sqrt{T}\left(\widehat{\theta}_T-\theta_T^{(p(T))}\right)=\sqrt{T}\left(\theta_T^{(p(T)+1)}-\theta_T^{p(T)}\right)\longrightarrow 0, \qquad (4.7)$$

in probability, then $\widehat{\theta}_T$ is asymptotically normal with asymptotic variance matrix

$$V(\theta^0)=A(\theta^0)^{-1}\Sigma(\theta^0,\theta^0)^{-1}B(\theta^0)\Sigma(\theta^0,\theta^0)^{-1}A(\theta^0)'^{-1}, \qquad (4.8)$$

where

$$A(\theta^0)=I_p-\frac{\partial\overline{\theta}}{\partial\theta^{1\prime}}\left[P^0,\lambda\left(\theta^0\right)\right].$$

Proposition 4.8 extends to a general framework a corresponding result of Renault and Touzi (1996) on maximum likelihood type estimators (see also Renault (1997) and Dominitz and Sherman (2001) for similar results). Now, let us comment this result and the assumptions used to obtain it.

Remark 1 (*about the asymptotic variance*) The asymptotic variance $V(\theta^0)$ is closely related to the standard asymptotic variance

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

of the hypothetical extremum estimator associated to the criterion $Q_T\left[\cdot, \lambda(\theta^0)\right]$. Typically, we expect that $V(\theta^0)$ will be larger than $W(\theta^0)$ due to the factor $[A(\theta^0)]^{-1}$, and one should not hope for any general result in this respect since we have shown in section 3 that the latent "data" could even be less informative than the observable ones. However, the superiority of the latent data is implicit in our backfitting strategy, like in any data augmentation approach. It was clearly the case in our two examples of standard and latent backfitting considered in section 2. For instance, the matrix $\partial\overline{\theta}/\partial\theta^{1\prime}\left[P^0, \lambda(\theta^0)\right]$ could be symmetric and positive. Its eigenvalues ρ_j, $j = 1, \cdots p$ are smaller than one. Then, in a orthonormal basis of eigenvectors, $[A(\theta^0)]^{-1}$ is a diagonal matrix with diagonal coefficients:

$$\left[1 - \rho_j\right]^{-1} > 1 + \rho_j > 1.$$

In other words, the backfitting efficiency loss (with respect to the infeasible oracle estimator), as measured by the eigenvalues ρ_j, is inversely related to the strength of the contraction.

Remark 2 (*about the EM algorithm*)

Proposition 4.8 can be interpreted as a generalization of a result stated by Nielsen (2000) about simulated EM algorithm. Nielsen (2000) nicely explains that when the E-step is performed with only one random draw that is reused for each iteration, the simulated EM algorithm can also be interpreted as looking for a fixed point by the method of successive substitution. In this case, the asymptotic variance matrix of the SEM estimator (Nielsen (2000)), theorem 2 p. 273) can be written:

$$\left[I_p - F\left(\theta^0\right)\right]^{-1} B\left(\theta^0\right)^{-1} \left[I_p - F\left(\theta^0\right)\right]^{-1}$$

Note that $B\left(\theta^0\right)^{-1} = \Sigma\left(\theta^0, \theta^0\right)$ is the Fisher information matrix in the latent model and $F\left(\theta^0\right)$ is the so-called "fraction of missing information".

$$F\left(\theta^0\right) = \left[B\left(\theta^0\right) - K\left(\theta^0\right)\right] B\left(\theta^0\right)^{-1}$$

where $K(\theta^0)$ is the Fisher information matrix in the observable model.

In other words, $F\left(\theta^0\right)$ corresponds to $\partial\bar{\theta}/\partial\theta_1'\left[P^0, \lambda\left(\theta^0\right)\right]$ as expressed by (4.4).

Remark 3 (*more on non-adaptivity*)

Now, let $\widetilde{\theta}_T$ denote the oracle, asymptotically normal estimator of θ^0 obtained by maximizing $\theta \to Q_T\left[\theta, \lambda(\theta^0)\right]$. Consider the infeasible estimator

$$\widehat{\widetilde{\theta}}_T = \arg\max_{\theta} Q_T\left[\theta, \lambda(\widetilde{\theta}_T)\right]$$

Then, under suitable regularity conditions, we have

$$\begin{aligned}
0 &= \frac{\partial Q_T}{\partial \theta}\left[\widehat{\widetilde{\theta}}_T, \lambda(\widetilde{\theta}_T)\right] \\
&\approx \frac{\partial Q_T}{\partial \theta}\left[\widetilde{\theta}_T, \lambda(\theta^0)\right] + \frac{\partial^2 Q_T}{\partial \theta \partial \theta'}\left[\theta^0, \lambda(\theta^0)\right]\left(\widehat{\widetilde{\theta}}_T - \widetilde{\theta}_T\right) \\
&\quad + \frac{\partial^2 Q_T}{\partial \theta \partial \theta^{1\prime}}\left[\theta^0, \lambda(\theta^0)\right]\left(\widetilde{\theta}_T - \theta^0\right) \\
&= \frac{\partial^2 Q_T}{\partial \theta \partial \theta'}\left[\theta^0, \lambda(\theta^0)\right]\left(\widehat{\widetilde{\theta}}_T - \theta^0\right) \\
&\quad - \frac{\partial^2 Q_T}{\partial \theta \partial \theta'}\left[\theta^0, \lambda(\theta^0)\right]\left(I_p + \frac{\partial \overline{\theta}}{\partial \theta^{1\prime}}\left[P^0, \lambda\left(\theta^1\right)\right]\right)\left(\widetilde{\theta}_T - \theta^0\right),
\end{aligned}$$

and we obtain that $\sqrt{T}(\widehat{\widetilde{\theta}}_T - \theta^0)$ is asymptotically normal with asymptotic variance

$$\left(I_p + \frac{\partial \overline{\theta}}{\partial \theta^{1\prime}}\left[P^0, \lambda\left(\theta^1\right)\right]\right)\Sigma(\theta^0, \theta^0)^{-1}B(\theta^0)\Sigma(\theta^0, \theta^0)^{-1}\left(I_p + \frac{\partial \overline{\theta}'}{\partial \theta^1}\left[P^0, \lambda\left(\theta^1\right)\right]\right).$$

As already observed in Remark 1, at least if the matrix $\partial\overline{\theta}/\partial\theta^{1\prime}\left[P^0, \lambda(\theta^0)\right]$ is symmetric and positive, we have:

$$I_p \ll I_p + \frac{\partial \overline{\theta}}{\partial \theta^{1\prime}}\left[P^0, \lambda(\theta^0)\right] \ll \left[I_p - \frac{\partial \overline{\theta}}{\partial \theta^{1\prime}}\left[P^0, \lambda(\theta^0)\right]\right]^{-1}$$

While the factor I_p is obtained for the oracle estimator and $[I_p - \partial\overline{\theta}/\partial\theta^{1\prime}\left[P^0, \lambda(\theta^0)\right]]^{-1}$ corresponds to our latent backfitting estimator, the infeasible estimator $\widehat{\widetilde{\theta}}_T$ is in between with the factor $I_p + \partial\overline{\theta}/\partial\theta^{1\prime}\left[P^0, \lambda(\theta^0)\right]$. This is typical of non-adaptivity. Even if we have at our disposal the best estimator of the nuisance parameters $\lambda^0 = \lambda\left(\theta^0\right)$, we would get an estimator less accurate than the oracle estimator. Of course, it would be more accurate than our latent backfitting estimator. In some respect, the backfitting estimator cumulates two sources of efficiency loss, both inversely related to the strength of the contraction.

Remark 4 (*the contracting property for finite samples*) Renault and Touzi (1996) proved the asymptotic normality of the latent backfitting in the particular case of MLE under a stronger hypothesis than (4.8). More precisely, they assumed that, except for a negligeable set, the norm $\left\|\partial\overline{\theta}_T/\partial\theta^1\left(\lambda(\theta^0)\right)\right\|$ becomes smaller than a constant $c \in [0,1)$ when $T \to \infty$. Then, with probability tending to one as $T \to \infty$, the function $\overline{\theta}_T(\lambda(\cdot))$ has a unique fixed point and the iterative estimator is defined by $\widehat{\theta}_T = \theta_T^{(\infty)}$. Dominitz and Sherman (2001)) weakened Renault and Touzi's assumption requiring only that $\overline{\theta}_T(\lambda(\cdot))$ is a uniform contracting mapping, which means there exists $c \in [0,1)$ independent of T and the sample such that, with probability tending to one as $T \to \infty$,

$$\left\|\overline{\theta}_T(\lambda(\theta')) - \overline{\theta}_T(\lambda(\theta''))\right\| \leq c\left\|\theta' - \theta''\right\|, \qquad \theta', \theta'' \in \Theta.$$

Dominitz and Sherman also observed that it suffices to define the iterative estimator using only a finite number of iterations provided that this number is greater than a power of T. Indeed, we can write

$$\sqrt{T}\left(\widehat{\theta}_T - \theta_T^{(p(T))}\right) \leq \sqrt{T}\,c\left(\theta_T^{(p(T))} - \theta_T^{(p(T)-1)}\right) ... \leq \sqrt{T}c^{p(T)}\left(\theta_T^{(1)} - \theta_T^{(0)}\right),$$

given that $\overline{\theta}_T(\lambda(\cdot))$ is contracting, and thus $\sqrt{T}c^{p(T)} \to 0$ if $p(T) \geq T^\delta$, $\delta > 0$. This shows that our Condition (4.8) is implied by the uniform contracting mapping condition for $\overline{\theta}_T(\lambda(\cdot))$. Even if in the examples the uniform contracting property may appear as more convenient to check, the condition (4.8) provides a more transparent rule for the choice of the sequence $\{p(T)\}$ in practice.

A general additional condition which implies the uniform contracting mapping condition for $\overline{\theta}_T(\lambda(\cdot))$ is

$$\sup_{\theta,\,\theta^1 \in \Theta}\left\|\frac{\partial^2 Q_T}{\partial\theta\partial\theta^{1\prime}}\left[\theta, \lambda(\theta^1)\right] - \frac{\partial^2 Q_\infty}{\partial\theta\partial\theta^{1\prime}}\left[\theta, \lambda(\theta^1)\right]\right\| \longrightarrow 0,$$

in probability as $T \to \infty$, which represents a more stringent version of Assumption 4.7. A quick way to see that this stronger uniform convergence condition implies the uniform contracting mapping condition for $\overline{\theta}_T(\lambda(\cdot))$ is to consider the identity

$$\left.\frac{\partial Q_T}{\partial\theta}\left[\theta, \lambda(\theta^1)\right]\right|_{\theta = \overline{\theta}_\mathrm{T}(\lambda(\theta^1))} = 0, \qquad \theta^1 \in \Theta,$$

and to differentiate it with respect to θ^1. This yields

$$\frac{\partial}{\partial\theta^{1\prime}}\overline{\theta}_T(\lambda(\theta^1)) = -\left[\frac{\partial^2 Q_T}{\partial\theta\partial\theta'}\left[\overline{\theta}_T(\lambda(\theta^1)), \lambda(\theta^1)\right]\right]^{-1}\frac{\partial^2 Q_T}{\partial\theta\partial\theta^{1\prime}}\left[\overline{\theta}_T(\lambda(\theta^1)), \lambda(\theta^1)\right]$$

and from the uniform convergence in probability of the second-order derivatives $\partial^2 Q_T/\partial\theta\partial\theta'$ and $\partial^2 Q_T/\partial\theta\partial\theta^{1\prime}$ we deduce that

$$\left\|\frac{\partial}{\partial\theta^{1\prime}}\overline{\theta}_T(\lambda(\theta^1))\right\| - \left\|\frac{\partial\overline{\theta}}{\partial\theta^{1\prime}}\left[P^0, \lambda(\theta^1)\right]\right\| = o_P(1),$$

uniformly in θ^1.

5 Robbins-Monro type estimators

Now, we focus on the case where the maximization problem characterizing the true values of the population parameters can be solved through the first order conditions. For this purpose, let us assume that the sample based criterion $Q_T[\theta, \lambda(\theta^1)]$ defined in section 2.1 is under the form

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

with $q_t(\theta, \lambda(\theta^1)) = q(\theta, \lambda(\theta^1); Y_t, X_t)$. Moreover, denote

$$M(\theta^1) = \frac{\partial Q_\infty}{\partial \theta}\left[\theta, \lambda(\theta^1)\right]\Big|_{\theta=\theta^1}. \tag{5.1}$$

Assuming the necessary regularity, let us restate the latent backfitting identification condition (see Assumption 2.3) in terms of first order conditions.

Assumption 5.1 *For any $\theta^1 \in \Theta$, $\overline{\theta}\left[P^0, \lambda(\theta^1)\right]$ is the unique solution of the equation*

$$\frac{\partial Q_\infty}{\partial \theta}\left[\theta, \lambda(\theta^1)\right] = 0.$$

Moreover, if θ^0 is the true unknown value of the parameter, then

$$M(\theta^1) = 0 \implies \theta^1 = \theta^0.$$

The previous assumption states that θ^0 is the unique solution of the just-identified moment problem defined by the function $M(\cdot)$. 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, amongst others, Ljung (1977), Kushner and Clark (1978), Kuan and White (1994a, b) and Kushner and Yin (1997). For our purposes, a Robbins-Monro (RM hereafter) procedure could be presented as follows: consider a population problem $M(\theta) = 0$, defined by some $M : \Theta \to \Theta$ and having a unique solution θ^0. The set Θ could be a subset of some Hilbert space, or for simplicity, $\Theta \subset \mathsf{R}^p$. Define finite sample counterparts of $M(\cdot)$ as

$$M_t(\theta) = M(\theta) + U_t(\theta), \qquad t \geq 1,$$

with $\{U_t(\theta)\}$ satisfying some technical conditions. Afterwards, for a sample size $T \geq 1$ and starting from some initial value θ_1, define an estimator for θ^0 as being the value θ_{T+1} obtained from the recursion

$$\theta_{t+1} = \theta_t + a_t M_t(\theta_t), \qquad 1 \leq t \leq T, \tag{5.2}$$

where $\{a_t\}$ 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 $a_t = c\, t^{-1}$, with $c > 0$ some scaling factor. This scaling factor does not influence the consistency results but it might be quite important for ensuring the asymptotic normality assumptions. The (random) function $M_t(\cdot)$ and the scalar a_t can be interpreted as a measurement of $M(\cdot)$ at time t (or 'learning update function') and a 'learning rate', respectively. Basically, the technical conditions required for $\{U_t(\theta)\}$ ensure, in general, that the averaged 'errors' $U_t(\theta_t)$, $t \geq 1$ vanish almost surely, more precisely $a_T \sum_{t=1}^{T} U_t(\theta_t) \to 0$, a.s.

The convergence results for RM procedures are commonly obtained via the ordinary differential equation (ODE) method proposed by Ljung (1977). The idea of Ljung was to show that the sequence $\{\theta_t\}$ defined in (5.2) asymptotically follow a trajectory (solution) $t \to \theta\,(t)$ of a deterministic ODE

$$\frac{d\theta}{dt}(t) = M(\theta(t))$$

associated with the population problem $M(\theta) = 0$. If the ODE satisfies a suitable stability condition, the trajectories of this ODE converge to θ^0 (as $t \to \infty$) and thus the consistency of the recursive estimates is ensured. Basically, the required stability condition is ensured locally if $M(\cdot)$ is a differentiable function and the matrix $\partial M/\partial \theta'(\theta^0)$ is negative stable, that is if all eigenvalues of this matrix have (strictly) negative real parts (see, e.g., Kuan and White (1994a), page 40, Horn and Johnson (1991), page 90, or Rouche and Mawhin (1980), ch. 1). We recall that the Lyapunov theorem (see, e.g., Horn and Johnson (1991), page 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.

In subsection 5.1 we introduce a recursive latent backfitting procedure for which we prove consistency. The asymptotic normality of this procedure is obtained in subsection 5.2. We will also show that our recursive latent backfitting estimator is, in general, less efficient than the iterative estimator studied in the previous section. However, the recursive estimation involve much less computer resources and it may be particularly appealing for estimating parameters in nonlinear models when large data sets are available.

5.1 Recursive latent backfitting: consistency

Let us investigate how the RM methodology applies to the population problem $M(\theta^1) = 0$ with $M(\cdot)$ defined in (5.1). Its unique solution θ^0 is assumed to belong to the interior of Θ. First, we check the negative stable condition. From (4.4), we derive

$$\begin{aligned} \frac{\partial M}{\partial \theta^{1\prime}}(\theta^0) &= \frac{\partial^2 Q_\infty}{\partial \theta \partial \theta'}\left[\theta^0, \lambda(\theta^0)\right] + \frac{\partial^2 Q_\infty}{\partial \theta \partial \theta^{1\prime}}\left[\theta^0, \lambda(\theta^0)\right] \qquad (5.3) \\ &= \frac{\partial^2 Q_\infty}{\partial \theta \partial \theta'}\left[\theta^0, \lambda(\theta^0)\right] \left(I_p - \frac{\partial \overline{\theta}}{\partial \theta^{1\prime}}\left[P^0, \lambda(\theta^1)\right] \right). \end{aligned}$$

As previously mentioned , a typical assumption for argmax estimation with limit criterion $\theta \to Q_\infty \left[\theta, \lambda(\theta^0)\right]$ is that its Hessian matrix is negative definite. Given this property, the condition

$$\left\| \frac{\partial \overline{\theta}}{\partial \theta^{1\prime}} \left[P^0, \lambda(\theta^0)\right] \right\| < 1, \tag{5.4}$$

guarantees the negative stable condition for $\partial M / \partial \theta^{1\prime}(\theta^0)$ (and thus the local stability for the ODE associated to the population problem defined by $M(\cdot)$), as it is shown in the following lemma.

Lemma 5.2 *If A_1 and A_2 are p-dimensional squared matrices with real elements such that A_1 is (symmetric) positive definite and $\|A_2\| < 1$, then $A = A_1(I_p - A_2)$ is positive stable.*

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

$$\frac{1}{2}(G\,A + A'\,G) = I_p - \frac{1}{2}(A_2 + A_2').$$

Now remark that $\|A_2\| = \|A_2'\|$ (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 $I_p - (A_2 + A_2')/2$ is positive definite. ■

A simple idea is to consider a RM procedure

$$\theta_{t+1} = \theta_t + ct^{-1}\, M_t(\theta_t)$$

based on the natural measurement of $M(\cdot)$, that is

$$M_t(\theta^1) = \left.\frac{\partial q_t}{\partial \theta}(\theta, \lambda(\theta^1))\right|_{\theta=\theta^1} = \left.\frac{\partial q}{\partial \theta}(\theta, \lambda(\theta^1); Y_t, X_t)\right|_{\theta=\theta^1}, \qquad t \geq 1. \tag{5.5}$$

The asymptotic properties of such a recursive procedure can easily be derived using general results on RM algorithms (see, e.g., Kuan and White (1994a)). We skip the details here.

Even if such a RM procedure seems quite attractive due to its very simple updating scheme, it may converge quite slowly (see Kuan and White (1994a), section II.3). 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., Kuan and White (1994a) or White (1989)). The basic idea for such a procedure is to replace the p equations $M(\theta^1) = 0$ by the $p^2 + p$ restrictions

$$\left\{ \begin{array}{l} vec\left(\partial M / \partial \theta^{1\prime}(\theta^1) - G\right) = 0 \\ \\ G^{-1} M(\theta^1) = 0 \end{array} \right. \tag{5.6}$$

with variables $vec(G)$ and θ^1 and to define the corresponding updating steps ($vec(G)$ stands for the p^2-dimensional vector obtained from the lines of a $p \times p$ matrix G).

By simple calculus, it can be noted that the negative stable condition for the derivative of the function defining a population problem as (5.6) automatically holds. Indeed, the derivative is a matrix of the form

$$\begin{pmatrix} -I_{p^2} & C \\ O & -I_p \end{pmatrix}$$

and it is clearly negative stable. Moreover, extending the arguments of Kuan and White (1994a) (see also the proof of Proposition 5.4 below), it can be proved that the RM estimator based on these population conditions has the same asymptotic variance as the (just-identified) GMM estimator for the restrictions $M(\theta^1) = 0$. In view of the equation (3.6) and the facts discussed in section 4.2, the asymptotic variance of such a GMM estimator is equal to the asymptotic variance of the iterative latent backfitting estimator built from $Q_\infty \left[\theta, \lambda(\theta^1)\right]$. However, the derivative of $M(\cdot)$ we need for building a recursive counterpart of the iterative latent backfitting estimator without loosing precision has an unfriendly form since it involves the second order cross-derivative of $Q_\infty[\cdot, \lambda(\cdot)]$. If one is ready to give up some precision for more tractability, a natural idea is to use the 'simple' part of $\partial M/\partial \theta^{1\prime}(\theta^0)$, that is the Hessian matrix of the simple limit criterion $Q_\infty \left[\cdot, \lambda(\theta^0)\right]$. This leads us to define a new modified RM procedure, conformable to the general idea of the latent backfitting. Since, in general, it may happen that the RM steps do not remain in a given domain for the parameter space, our procedure has to include a truncation (projection) device (see Kuan and White (1994a), Chen and White (1992)).

Denote

$$\Sigma(\theta^1) = - \frac{\partial^2 Q_\infty}{\partial\theta\partial\theta'} \left[\theta, \lambda(\theta^1)\right] \Big|_{\theta=\theta^1}, \qquad \theta^1 \in \Theta, \tag{5.7}$$

and let

$$B\left(\theta^0, r\right) = \{\theta, \quad \|\theta - \theta^0\| < r\} \subset \Theta, \qquad U(\theta^0, \rho) = \{G, \quad \left\|vec\left(G - \Sigma(\theta^0)\right)\right\| < \rho\},$$

$r, \rho > 0$ be neighborhoods of θ^0 and $vec\left(\Sigma(\theta^0)\right)$, respectively. Assume that $U(\theta^0, \rho)$ contains only invertible matrices. Consider

$$\overline{\theta}_t\left(\lambda(\theta^1); r\right) = \arg \max_{\theta \in B\left(\theta^0, r\right)} Q_T\left[\theta, \lambda(\theta^1)\right].$$

By a slight abuse of notation, this maximum is well-defined, at least locally.
Fix $c > 0$ and define

$$\theta_{t+1} = \begin{cases} \theta_t + \frac{c}{t} G_{t+1}^{-1} M_t(\theta_t) & on \quad B_{t+1} \\ \\ \overline{\theta}_t\left(\lambda(\theta_t); r\right) & on \quad \Omega \setminus B_{t+1} \end{cases}, \qquad t \geq 1, \tag{5.8}$$

where

$$G_{t+1} = \begin{cases} G_t - \frac{c}{t}\left[\partial^2 q_s/\partial\theta\partial\theta'(\theta, \lambda(\theta_t))|_{\theta=\theta_t} + G_t\right] & on \quad C_{t+1} \\ \\ \overline{G} & on \quad \Omega \setminus C_{t+1} \end{cases} \qquad t \geq 1, \tag{5.9}$$

and $M_t(\cdot)$ is defined in (5.5); $\overline{G}$ is some fixed invertible $p \times p$ matrix, for example the identity matrix if it belongs to $U(\theta^0, \rho)$. The events B_{t+1}, C_{t+1}, $t \geq 1$ are defined as

$$\begin{aligned} B_{t+1} &= \left\{\theta_t + ct^{-1} G_{t+1}^{-1} M_t(\theta_t) \in B\left(\theta^0, r\right)\right\}, \\ C_{t+1} &= \left\{G_t - \frac{c}{t}\left[\left.\frac{\partial^2 q_s}{\partial\theta\partial\theta'}(\theta, \lambda(\theta_t))\right|_{\theta=\theta_{\mathrm{t}}} + G_t\right] \in U(\theta^0, \rho)\right\}, \end{aligned}$$

for some $r, \rho > 0$. The starting values are some $\theta_1 \in B\left(\theta^0, r\right)$ and G_1 invertible matrix with $vec(G_1) \in U\left(\theta^0, \rho\right)$. Note that the recursive procedure above depend on the neighborhoods $B\left(\theta^0, r\right)$ and $U\left(\theta^0, \rho\right)$ through the initial values and the truncation device.

We denote by θ_T^{RM}, $T \geq 1$ the value θ_{T+1} obtained from the last recursive procedure above, sometimes called a bounded truncated RM procedure with $\theta-$dependent errors. The population problem is given by the function $M^{RM}(\delta) = (M_1(\delta)', M_2(\delta)')'$ with

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

and

$$M_1(\delta) = vec\left(\Sigma(\theta^1) - G\right), \qquad M_2(\delta) = G^{-1} M(\theta^1). \tag{5.10}$$

Under the Assumption 5.1, the unique root of the equation $M^{RM}(\delta) = 0$ is the vector $\delta^0 = (vec(\Sigma(\theta^0))', \theta^{0\prime})'$. The measurements of $M^{RM}(\delta)$ are $M_t^{RM}(\delta) = (M_{1t}(\delta)', M_{2t}(\delta)')'$, $t \geq 1$ with

$$M_{1t}(\delta) = -vec\left[\left.\frac{\partial^2 q_t}{\partial\theta\partial\theta'}(\theta, \lambda(\theta^1))\right|_{\theta=\theta^1} + G\right], \qquad M_{2t}(\delta) = G^{-1}\left.\frac{\partial q_t}{\partial\theta}(\theta, \lambda(\theta^1))\right|_{\theta=\theta^1}. \tag{5.11}$$

Simple algebra and the equation (5.3) yield

$$\frac{\partial M^{RM}}{\partial\delta}(\delta^0) = \begin{pmatrix} -I_{p^2} & \partial vec(\Sigma(\theta^0))/\partial\theta^{1\prime} \\ 0 & -\left(I_p - \partial\overline{\theta}/\partial\theta^{1\prime}\left[P^0, \lambda(\theta^0)\right]\right) \end{pmatrix}. \tag{5.12}$$

Clearly, this matrix is negative stable provided that the blocs on the diagonal are negative stable matrices. This leads us to the following assumption.

Assumption 5.3 *The matrix*

$$I_p - \frac{\partial\overline{\theta}}{\partial\theta^{1\prime}}\left[P^0, \lambda(\theta^0)\right]$$

is positive stable.

Note that Assumption 5.3 is weaker than the contraction mapping assumption needed for the convergence of the iterative algorithm.

Also, observe that the positive stable condition implies the invertibility of the matrix $I_p - \partial\overline{\theta}/\partial\theta^{1\prime}\left[P^0, \lambda(\theta^0)\right]$. Therefore, the last assumption above ensures, locally, the uniqueness of θ^0 as solution of the equations $M(\theta^1) = 0$ (see Assumption 5.1).

The remaining conditions for ensuring the almost sure (strong) convergence of our recursive latent backfitting can be stated as in Kuan and White (1994a), section II.2 (see Appendix 7 for a precise description of these conditions). Having all the ingredients, we can prove the following asymptotic result. In particular, it turns out that, with probability one, the truncation in the definition of $\left\{\theta_T^{RM}\right\}$ is invoked only a finite number of times.

Proposition 5.4 *Assume that the function $M(\cdot)$ defined in (5.1) is continuously differentiable on Θ for which θ^0 is an interior point. Suppose that Assumptions 5.1 and 5.3 hold. Moreover, assume that Assumption A.0.1 to A.0.3 in Appendix 7 are satisfied. Then there exist r, $\rho > 0$ such that, if $\left\{\theta_T^{RM}\right\}$ is the corresponding RM sequence obtained from (5.8)-(5.9) starting from some initial values θ_1 and vec(G_1), then*

$$\theta_T^{RM} \to \theta^0, \qquad \textit{almost} \quad \textit{surely}.$$

5.2 The asymptotic distribution of the recursive latent backfitting estimators

The asymptotic normality of a RM type estimator as proposed above could be obtained from Theorem 2 of Kushner and Huang (1979) or Theorem II.2.4 of Kuan and White (1994a). Let us recall the basic facts we use here: consider a population problem $M(\theta) = 0$ defined by some $M : \Theta \subset \mathsf{R}^p \to \Theta$ and assume that θ^0 is the unique solution. Let $\{M_t(\theta)\}$ be a sequence of stationary measurements of $M(\theta)$ and $\{\theta_T\}$ be an almost surely convergent sequence of estimators obtained from a RM procedure as in (5.2) with $a_t = ct^{-1}$, $c > 0$. Suppose that the matrix

$$\overline{H} = c\,\frac{\partial M}{\partial \theta'}(\theta^0) + \frac{1}{2}\,I_p \tag{5.13}$$

is negative stable. It can be shown that, under the conditions stated in the Appendix 7, $\sqrt{T}(\theta_T - \theta^0)$ converges in distribution to a normal random variable with zero mean and variance matrix C solution of the equation

$$\overline{H}C + C\overline{H}' = -c^2 R, \tag{5.14}$$

where

$$R = \sum_{t=0}^{\infty} R_t + \sum_{t=1}^{\infty} R_t' \tag{5.15}$$

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

$$C = c^2 \int_0^\infty \exp(\overline{H}u) R \exp(\overline{H}'u) du. \tag{5.16}$$

Recall that an usual condition for the consistency of the RM estimators is that the matrix $\partial M/\partial\theta'(\theta^0)$ is negative stable. If the real parts of the eigenvalues of this matrix are negative but too close to zero in such way that $\overline{H}$ 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 and of the eigenvalues of $\partial M/\partial\theta'(\theta^0)$.

In the case of our recursive latent backfitting estimator the population problem is given by the function $M^{RM}(\cdot) = (M_1(\cdot)', M_2(\cdot)')'$ defined in (5.10) and we have

$$\overline{H} = \begin{pmatrix} \left(\frac{1}{2} - c\right) I_{p^2} & \partial vec(\Sigma(\theta^0))/\partial\theta^{1\prime} \\ & \\ 0 & \left(\frac{1}{2} - c\right) I_p + c\partial\overline{\theta}/\partial\theta^{1\prime}\left[P^0, \lambda(\theta^0)\right] \end{pmatrix}.$$

Proposition 5.5 *Suppose that θ^0 is an interior point of Θ and that Assumption 5.1 is satisfied. Let $M^{RM}(\cdot) = (M_1(\cdot)', M_2(\cdot)')'$ be defined as in (5.10). Assume that Assumptions (A.0.1) to (A.0.5) in Appendix 7 hold. Let $c > 1/2$ such that*

$$F = F(\theta^0, c) = \left(\frac{1}{2} - c\right) I_p + c\frac{\partial\overline{\theta}}{\partial\theta^{1\prime}}\left[P^0, \lambda\left(\theta^0\right)\right]$$

is negative stable. Consider $U(\theta^0, \rho) \times B\left(\theta^0, r\right) \subset \mathsf{R}^{p^2} \times \Theta$ such that the corresponding RM sequence $\left\{\theta_T^{RM}\right\}$ defined in (5.8)-(5.9) with learning rate $a_t = ct^{-1}$ converges almost surely. Then

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

where

$$V^{RM}(\theta^0) = c^2 \int_0^\infty \exp\left(Fu\right)\Sigma(\theta^0)^{-1}B(\theta^0)\Sigma(\theta^0)^{-1}\exp(F'u)\,du,$$

with $B(\theta^0)$ and $\Sigma(\theta^0)$ defined in Assumption 4.5 and equation (5.7), respectively.

We can prove that our RM estimator is generally less efficient than the iterative latent backfitting estimator analyzed in section 4. Indeed, let

$$W = \Sigma(\theta^0)^{-1}B(\theta^0)\Sigma(\theta^0)^{-1} \qquad \textit{and} \qquad A = I_p - \frac{\partial\overline{\theta}}{\partial\theta^{1\prime}}\left[P^0, \lambda\left(\theta^1\right)\right],$$

where W is assumed positive definite. Recall that the variance of the iterative latent backfitting estimator is $V(\theta^0) = A^{-1}WA^{-1\prime}$ (see Proposition 4.8). On the other hand, the variance $V^{RM}(\theta^0)$ of the RM estimator is the solution of the matrix equation

$$FV + VF' = -c^2W,$$

where $F = (1/2)I_p - cA$. Since F is supposed to be negative stable, if we prove that

$$F\left(V^{RM}(\theta^0) - V(\theta^0)\right) + \left(V^{RM}(\theta^0) - V(\theta^0)\right)F'$$

is negative definite, then, using the converse Lyapunov theorem (Horn and Johnson (1991), Th. 2.2.3, page 98), we can show that $V^{RM}(\theta^0) - V(\theta^0)$ is positive definite. By simple algebra we obtain

$$\begin{aligned} & F\left(V^{RM}(\theta^0) - V(\theta^0)\right) + \left(V^{RM}(\theta^0) - V(\theta^0)\right)F' \\ = \; & -c^2W + cWA^{-1\prime} + cA^{-1}W - A^{-1}WA^{-1\prime} \\ = \; & -\left(cI_p - A^{-1}\right)W\left(cI_p - A^{-1}\right)'. \end{aligned}$$

Thus, $V^{RM}(\theta^0) - V(\theta^0)$ is positive definite except for the case $cI_p = A^{-1}$ when $V^{RM}(\theta^0) = V(\theta^0)$.

6 Empirical implementation issues

6.1 The framework

We focus in this section on the application of the implied state backfitting methodology to the estimation of asset pricing models. Typically, such a pricing model explains an observed stationary process Y_t of n asset "prices" as a known function of the current value X_t of K latent state variables and p unknown parameters θ:

$$Y_t = \{h_i[X_t, \theta]\}_{1 \le i \le n} \tag{6.1}$$

Note that when one loosely says asset "prices", one should rather understand "yields" in the case of bonds or "option premium by unit of spot price" in case of options on equity or any other transformation well-suited to build a n-dimensional stationary time series Y_t from the observation of time series of asset prices, likely to be non-stationary. In the context of options on equity, one may also replace (see e.g. Renault and Touzi (1996) and Pastorello, Renault and Touzi (2000)) option prices by the corresponding Black-Scholes implied volatilities.

With respect to the most general formulation of empirical asset pricing models presented in the introduction, we focus here on a more specific approach that is more common in the arbitrage-free asset pricing literature:

First, the pricing kernel is not explicitly included in the list Y_t^* of latent state variables. Instead, it is defined as a known function of a collection X_t of relevant risk factors as instantaneous risk free rate, diffusive return shocks, volatility shocks and jump events as well as a collection of risk premium parameters θ_2 that define the compensation for the various risk factors. Typically, we will view X_t as a subset of the relevant vector Y_t^* of state variables. Then, the dynamics of the latent risk factors X_t only identify a set θ_1 of unknown "statistical" parameters while the risk premium parameters θ_2 must be added to define the complete vector θ of structural parameters of interest for asset pricing:

$$\theta = [\theta_1', \theta_2']' \tag{6.2}$$

For empirical option pricing on equity, the above approach is typically the one followed by Heston (1993), Bates (2000), Chernov and Ghysels (2000), and Pan(2002) among others. For term structure modeling, this approach is particularly well-suited to capture through K explanatory latent factors of the yield curve the relationships between n observed yields in cross-section. A large strand of literature, initiated in particular by Chen and Scott (1993), Pearson and Sun (1994) and Duan (1994), uses this indirect empirical modeling of bond yields through underlying latent factors. In contrast, explicit dynamic modeling of the joint stochastic process of asset returns and pricing kernel can be found in the consumption-based equilibrium asset pricing literature (see e.g. Aït-Sahalia and Lo (2000), Jackwerth (2000), Rosenberg and Engle (2000) for applications to option pricing) or, in an even more general way in Constantinides (1992) and Garcia, Luger and Renault (2002).

The second important difference between the asset pricing model (6.1) and the general framework appearing in the introduction is the fact that we do not maintain at this stage the assumption of a one-to-one relationship between the vector Y_t of observed prices and the vector X_t of structural latent state variables.We even do not consider that the two dimensions n and K of these two vectors should necessarily coincide.

Of course, the simplest approach to estimating a K factors model is to select $n = K$ asset prices in the cross section and to exploit the one-to-one relationship between prices and factors to get either the exact likelihood (Chen and Scott (1993), Pearson and Sun (1994), Duan (1994)) or an expansion of it (Aït-Sahalia and Kimmel (2002)) or implied moments (Pan(2002)) or a simulated score (Dai and Singleton(2000)). This approach leads unmistakably to neglect the potentially useful information conveyed by a number of observed related prices in the cross section. For instance Pan (2002) estimates a stochastic volatility model for option pricing on the S&P 500 index from the joint time series of the index and one near-the-money short dated option on it. One option price is sufficient to get a one-to-one relationship with the volatility factor, yet (see e.g. Dumas, Fleming and Whaley (1998)), by taking into account the various possible moneynesses and maturities, the number of fairly liquid option prices on S&P 500 that can be observed at any given date may be about ten or even more. Similarly, while common models of the yield curve involve $K = 1$, 2 or 3 factors, the number n of available maturities in the cross section (see e.g. the McCullogh and Kwon data set used in the empirical part of this section) is about thirty or even more.

Since asset pricing based on ten, twenty or more structural factors is not very appealing, the only way to reconcile an orthodox view of statistical information (the n cross sectional observations must be used in the inference process) and structural asset pricing with a reduced number K of latent structural factors is to include $(n - K)$ error terms. In other words, the price to pay to incorporate all the available statistical information is to assume that, due to some frictions in the financial markets, some degrees of freedom remains possible around the theoretical no-arbitrage based prices. Therefore, the retained empirical specification of the asset pricing model (6.1) will be:

$$\begin{aligned} Z_t &= (Y_{it})_{1\leq i\leq K} = h[X_t,\theta] = [h_i(X_t,\theta)]_{1\leq i\leq K} \\ V_t &= (Y_{it})_{K+1\leq i\leq n} = e[X_t,\theta] + u_t = [h_i(X_t,\theta)]_{K+1\leq i\leq n} + [u_{it}]_{K+1\leq i\leq n}. \end{aligned} \tag{6.3}$$

Note that we consider at this stage that the n assets prices have been relabeled in order to get zero pricing errors for the K first ones while the $(n-K)$ other ones differ from their theoretical values by error terms u_{it}. Hence, we do not really maintain the arbitrary assumption that exactly K prices coincide with their theoretical values while error terms may be added to the other ones. We just say that, since the structural model already involves K latent factors, there is no reason to introduce more than $(n-K)$ error terms, while at least K independent linear combinations should be observed without error. Of course, such a specification needs to know a priori what are the K prices (or the K linear combinations of prices) that are observed without error. This empirical issue will be discussed below.

For sake of expositional simplicity, we limit the comparison of the implied state backfitting methodology with competitors for inference on (6.3) to the context of maximum likelihood-based inference strategies. A maintained assumption will be that the error terms u_{it} have a zero unconditional mean and that the first K equations provide a one-to-one relationship between the vector Z_t of the K prices observed without error and the vector X_t of structural state variables:

$$Z_t = (Y_{it})_{1\leq i\leq K} = h[X_t,\theta] \Leftrightarrow X_t = h^{-1}[Z_t,\theta] \tag{6.4}$$

6.2 Maximum Likelihood based inference

The conditional likelihood associated to a data set $\{Y_t, t=1,\ldots,T\}$ (and an initial conditioning value Y_0) must be derived, through the Jacobian formula, from the latent one associated with the "latent data" set $\{Y_t^*, t=1,\ldots,T\}$ produced by the latent realizations of a Markov process Y^* one-to-one function of Y:

$$Y_t = g[Y_t^*,\theta] \Leftrightarrow Y_t^* = g^{-1}[Y_t,\theta] \tag{6.5}$$

Typically, (6.5) must be defined by n equations, thanks to $(n-K)$ equations that complete the K equations (6.4). A natural idea would be to define the state vector Y_t^* by augmenting the vector X_t of K structural factors with the vector u_t of $(n-K)$ error terms. However, we certainly do not want to do this for two reasons.

First, the parameters η that would define the probability distribution of the error term u_t are not the focus of interest. Of course, their consistent estimation may be useful for improving the accuracy of the estimation of the parameters of interest θ. We do want to ensure, however, that even if η is not consistently estimated, we obtain a consistent estimator of θ. Typically, in case of Gaussian errors, the vector of nuisance parameters η consists of the unconditional covariance matrix Ω of the $(n-K)$ error terms u_t and possibly the parameters defining the conditional mean and variance dynamics. The mere fact that these error terms are added ex post and not rationalized within a structural asset pricing model with additional state variables implies that we have no structural information about their dynamics. Since from (6.3) we note that the estimation of the dynamics of the error terms may contaminate the estimation of the dynamics of the structural factors, it is important to define a backfitting procedure that focuses only on the structural parameters θ and not on the augmented vector (θ, η).

Second, the backfitting identification condition for θ would not be fulfilled is we defined the latent state vector Y_t^* as $Y_t^* = (X_t, u_t)$. The empirical asset pricing model (6.3) provides a one-to-one relationship between observed prices Y_t and latent variables (X_t, u_t) but the risk premium parameters θ_2 are identified only by the relationship itself and not by the probability distribution of the latent process (X_t, u_t). In other words, by defining $Y_t^* = (X_t, u_t)$, we would be faced with the exact opposite situation of the one described in the comments following Assumption 4.7. As it was stressed at this earlier stage, the philosophy of our backfitting methodology is precisely to assume that the latent model (the transition equation of the state variables) carries more information about the unknown parameters of interest than their occurrence in the measurement equation. To remain true to this philosophy, a better strategy is to define the latent vector Y_t^* and the associated function $g[Y_t^*, \theta]$ in the following way:

$$\begin{aligned} Y_t^* &= [X_t', V_t']', Y_t = [Z_t', V_t']' \\ \text{with:} & \\ Y_t &= g[X_t, V_t, \theta] = [h'(X_t, \theta), V_t']'. \end{aligned} \tag{6.6}$$

Note that $(n - K)$ among the n so-called latent variables Y_t^* are actually observed but this does not prevent us from applying the general backfitting methodology. In this context, the transition density function of the Markov process Y_t^*:

$$l[Y_t^* \left| Y_{t-1}^* \right.] = l[X_t \left| Y_{t-1}^* \right.] \, l[V_t \left| X_t, Y_{t-1}^* \right.] \tag{6.7}$$

will be specified under the maintained common assumption that error terms do not cause structural factors, neither in the Granger sense nor instantaneously. This assumption is natural since, if one imagines its violation, one implicitly endows the error terms with some structural interpretation. Then, by the no-Granger causality assumption:

$$l[X_t \left| Y_{t-1}^* \right.] = l[X_t \left| X_{t-1} \right.] = l[X_t \left| X_{t-1}, \theta_1 \right.] \tag{6.8}$$

where the last expression stresses the fact that this density function depends on the value of the unknown parameters only through θ_1. By the no instantaneous causality assumption, $l[V_t \left| X_t, Y_{t-1}^* \right.]$ is simply obtained by a translation of size $e[X_t, \theta]$ applied to the conditional probability distribution l$[u_t \left| Y_{t-1}^*, \eta \right.]$ of the error terms given the past. This probability density function depends on the value of the unknown parameters only through the nuisance parameters η.

Since we maintain the assumption that all the structural content of the model is captured by the factors X_t, we do not really want to specify the dynamics of the error terms and we will carry out inference about structural parameters through a latent quasi-likelihood, written as the likelihood of a latent model where the error terms would be i.i.d. Gaussian with a covariance matrix specified as a function $\Omega(\eta)$:

$$l[u_t \left| Y_{t-1}^*, \eta \right] = l[u_t \left| \eta \right] = (2\pi)^{-(n-K)/2}[det\Omega(\eta)]^{-1/2}exp[-\frac{1}{2}u_t'\Omega^{-1}(\eta)u_t] \tag{6.9}$$

Several remarks are in order about the use of this quasi-likelihood. First, it is well-suited only if the scale Y_t used to measure asset prices is consistent with conditional normality like for instance log-returns or log-implied volatilities. Second, we should not forget that the quasi-likelihood may differ from the true likelihood and that we just want to plug it into (6.7) to get a consistent estimator of the structural parameters of interest θ. The nuisance parameters η are likely to be poorly defined and not consistently estimated. However, a general specification of the covariance matrix $\Omega(\eta)$ should at least allow us to take into account the obvious strong cross sectional patterns of correlation and heteroskedasticity among error terms (see Renault (1997) for a general discussion).

A third important remark is that, in contrast with standard estimation/filtering strategies, the inference approach must take the hierarchy between θ and η into account and, in particular, it does not make sense to filter on an equal basis simultaneously the artificial state variables u_t and the state variable of interest X_t. Besides the gain in computing time (we will document below a significant difference between computing times for implied states backfitting and Kalman filtering), the hierarchy of the two filtering issues should avoid to contaminate the filtered current values of essential pricing factors like X_t with the noise caused by error terms. This contamination at the filtering level has been avoided by imposing a priori the nullity of a subset of K error terms, rather than specifying a more general model:

$$Y_t = (Y_{it})_{1\leq i\leq n} = [h_i(X_t, \theta)]_{1\leq i\leq n} + \Sigma[\varepsilon_{it}]_{K+1\leq i\leq n}, \tag{6.10}$$

with Σ matrix of size $n \times (n-K)$ to be estimated. Our trials of estimation/filtering procedures on general specifications like (6.10) of term structure models lead us to the conclusion that the resulting filtered factors X_t are highly unstable, due the instability of the error terms. In contrast, by imposing the specification like (6.3) that the first K rows of the matrix Σ are zero, we get satisfactory estimation and filtering results, conditionally to a preliminary statistical procedure (see discussion in subsection 6.3. below) to decide what are the K zero rows (to be relabeled as the first K rows). At the estimation level, the hierarchy is ensured by computing first an estimator

$$\Omega_T = \Omega(\eta_T)$$

of $\Omega(\eta)$, and then, plugging it into (6.7) to define the latent criterion for extremum estimation of the structural parameters θ:

$$Q_T^*(\theta) = \Sigma_{t=2}^T \; Log \; l \; [X_t \left| X_{t-1}, \theta_1 \right] - \frac{1}{2}\Sigma_{t=1}^T[V_t - e(X_t, \theta)]'\Omega_T^{-1}[V_t - e(X_t, \theta)] \tag{6.11}$$

Up to recursive refinements, the backfitting methodology amounts defining a sequence $\theta^{(p)}$ of estimators in the following way:

a) Start from an estimator $\theta^{(1)}$ provided by a quick procedure.

b) For $\theta^{(p)}$ given, replace in (6.11) the unknown factor values X_t by $X_t(\theta^{(p)}) = h^{-1}[Z_t, \theta^{(p)}]$. This defines a sample based criterion $Q_T(\theta, \theta^{(p)})$.

c) Compute the estimator $\theta^{(p+1)}$ as $arg\ max_\theta Q_T(\theta, \theta^{(p)})$.

Since the nuisance parameters η have been introduced in a way that preserves adaptivity, the resulting asymptotic probability distribution of the backfitting estimator of θ will only depend upon the probability limit of Ω_T and not upon its accuracy as estimator of the (pseudo) true unknown value of $\Omega(\eta)$. However, at least in case where the conditional distribution of the error terms would be well-specified, the most accurate backfitting estimator would be obtained when Ω_T is a consistent estimator of the true value of $\Omega(\eta)$. This is the reason why it is natural to think to a "quasi-generalized" version of backfitting in the following way.

Start from an arbitrary Ω_T (e.g. the identity matrix) and compute the corresponding backfitting estimator θ_T of θ. Then, use it to compute "estimated error terms":

$$u_t(\theta_T) = V_t - e[X_t(\theta_T), \theta_T] \tag{6.12}$$

and to derive a consistent estimator $\eta(\theta_T)$ of the pseudo true value of η and in turn, a consistent estimator $\Omega_T^* = \Omega[\eta(\theta_T)]$ of the pseudo true value of Ω. Then, perform a second backfitting estimation of θ based on the criterion (6.11) where Ω_T has been replaced by W_T^*. Of course, such a procedure is costly since it implies several backfittiting estimations. Fortunately, there exists a much faster procedure that is, in terms of estimation of θ, asymptotically equivalent to quasi-generalized backfitting, but in terms of computing time, equivalent to a simple backfitting.

This procedure, that we term "extended backfitting" amounts to using each step $\theta^{(p)}$ of the backfitting iteration to compute a new estimator $\Omega[\eta(\theta^{(p)})]$ of the matrix Ω and to plug it into (6.11) in place of Ω_T to derive the next step estimator $\theta^{(p+1)}$ of θ. At first sight, extended backfitting is similar to standard backfitting applied to the augmented vector (θ, η) of unknown parameters. However, we do not refer to our general backfitting theory (in terms of an augmented vector of parameters) to justify this procedure. There is little hope to get a sequence that is contracting with respect to the nuisance parameters η and this is the reason why the convergence criterion of the approximation sequence that we will use in applications will only be based on the norm $||\theta^{(p+1)} - \theta^{(p)}||$.

The relevant argument is the following. Irrespective of the choice of the weighting matrix Ω_T in (6.11), the backfitting estimator is a consistent estimator of the true unknown value of θ. Therefore, it is clear that the limit of the sequence $\theta^{(p)}$ produced by the extended backfitting algorithm also provides a consistent estimator of θ and, in turn, the limit of the sequence $\Omega[\eta(\theta^{(p)})]$ provides a consistent estimator of the true unknown value of $\Omega[\eta]$. Since the asymptotic probability distribution of the backfitting estimator of θ only depends on the probability limit of Ω_T, it is then clear that we get an estimator asymptotically equivalent to the quasi-generalized backfitting. This procedure will be illustrated in subsection 6.3 below for the estimation of an affine term structure model of interest rates.

Before going into the details of an application, let us first briefly sketch a comparison with the maximum likelihood based competitors also well-suited for inference on such empirical asset pricing models with latent factors.

A first competitor is the Kalman filter based quasi maximum likelihood. The most popular strategy is to introduce n error terms instead of $(n - K)$ to avoid the instability properties already mentioned about (6.10). This has been first proposed in the context of affine models of the yield curve by Duan and Simonato (1999) and systematically developed by De Jong (2000). Of course, severe nonlinearities or non-normality of the structural model are likely to alter the validity of the Kalman filter. Generally speaking, the Kalman filter should not be used for highly nonlinear models and our backfitting filtering strategy should be much better suited. However, in the context of return dynamics that are not too far to be linear as in the case of affine models of the yield curve, the two approaches may be competitors and we are going to compare their performance in the empirical application below. Roughly speaking, the Kalman filtering approach can be seen as a quick and dirty procedure to check the validity of our possibly more accurate but also more risky approach. Typically, the backfitting approach seeks to get more efficient estimators and filters by taking the risk to specify exact nonlinear relationships between prices and factors with K zero error terms.

Another quasi maximum likelihood approach for factor models of the yield curve has been applied by Fisher and Gilles (1996) and Duffee (2002). Their idea is quite simple. Even though the latent model is conceived to be simpler than the observable one, the hard part of the latent log-likelihood (6.11) is the transition density function of the structural factors X_t. This function is in general produced by a continuous time model and may be hard to compute or simply unknown. However, consistent (albeit inefficient) estimates can still be obtained if we substitute the true theoretical transition density with a Gaussian one, provided that the first two conditional moments of X_t are correctly specified. Besides its potential inefficiency, this alternative QML approach also suffers from a risk of mispecification bias in case of a nonlinear mapping g between the latent variables and the observables. In such a case, the Jacobian formula applied to a latent Gaussian quasi-likelihood may not yield a correct quasi-likelihood for observables. This drawback is not detrimental in the case of affine (Fisher and Gilles (1996)) or essentially affine (Duffee (2002)) term structure models but would be an issue in the case of option prices on equity with stochastic volatility.

Moreover, as neatly put forward by Duffee (2002), "another advantage of QML (which it shares with maximum likelihood and related techniques) is that $(\cdots)$ a model estimated with QML will guarantee that the time-t state vector implied by time-t yields is in the state vector's admissible space (to avoid a likelihood zero). By contrast, $(\cdots)$ techniques such as EMM $(\cdots)$ do not require that the estimated term structure model be sufficiently flexible to reproduce the term structure shapes in the data. The parameters of the model in Dai and Singleton (2000), which were estimated with EMM, illustrate this point."

This point is actually our main motivation to focus on implied-states based likelihood methodologies. Besides the Kalman filter approach, two likelihood based methodologies will be empirically compared in subsection 6.3 below. For a fair comparison between our backfitting approach and its two main competitors, we focus on the class of affine term structure models since it gives a chance to Kalman based strategies and also to exact likelihood strategies that are not too cumbersome in this case.

6.3 An application to one-factor affine models of the yield curve

In this subsection, we will outline the results of the estimation of two one-factor affine term structure models on a widely used data set of U.S. zero coupon yields. By doing this, our goal is definitely not to prove that an affine model with only one factor is able to capture all the relevant empirical features of the yield curve. We just want to exploit the analytical tractability of this model to make more explicit our comparison between competing maximum likelihood based approaches. We will be in particular able to show that an additional advantage of the backfitting strategy of working directly on the latent likelihood rather than on the observable one is to allow us to apply the Aït-Sahalia (2002) closed-form likelihood expansions. By contrast, when one wants to work directly on such expansions about the likelihood associated to observed bond prices as proposed recently by Aït-Sahalia and Kimmel (2002), the maximization of such expansion will be shown to be highly misleading.

Of course, the comparison of several competing methodologies on a given data set may be flawed by some misspecification of the structural model. Such a misspecification is even more likely when using a very simple model as we do. An extensive Monte Carlo study that allows to control for misspecification bias in our conclusions is work in progress. However, we can already assert that the proposed implied states extended backfitting approach will not be contaminated by specification errors on the ad hoc model of pricing errors. This is clearly not the case with direct likelihood strategies on observed asset prices since, for instance, some omitted dynamics in the error terms is likely to bias the maximum likelihood estimation of factor dynamics.

Models and data

Our database is the extended McCulloch dataset (see McCulloch (1975), (1990) and Kwon (1992)). The whole dataset consists of monthly zero-coupon rates that were calculated using McCulloch's interpolation method.

Following De Jong (2000), we use a restricted sample starting in January 1970 and ending in February 1991, for a total of T = 254 observation dates.

There are 56 fixed maturities available, ranging from 1 month to 40 years (the number of maturities actually available at each date depends on the number of outstanding bonds). However, the estimation of longer maturity yields is based on very few coupon bonds, and hence they are usually dropped from the sample. The same is usually done with the shortest maturity yields (1 and 2 months), as they exhibit some extremely large variations over short intervals.

For sake of comparability with the Kalman approach of De Jong (2000), we estimate the various models by using only $n = 4$ maturities: 3 months, 1, 5 and 10 years. But, by contrast with the Kalman approach, our model (6.3) stipulates that $K = 1$ among the four yields is observed without error. However, we are going to show that the estimation results about the structural parameters θ are highly sensitive to the choice of the component supposed to be observed without error. This leads us to follow Collin-Dufresne, Goldstein and Jones (2002) to consider that it might be preferable to work on the time series of the principal components of the observed yields, instead of directly on the time series of the yields. The first principal component is defined by the following four weights (in increasing order of maturities of the underlying zero yields): (0.5488, 0.5436, 0.4687, 0.4285). We obtain our most reliable estimation results by assuming that it is precisely this first estimated principal component that is observed without error.

In affine term structure models, zero coupon yields are affine functions of the unobservable factor X_t. By taking log of the price $P_t(\tau)$ in t of a zero coupon bond with maturity $t+\tau$ and dividing by $(-\tau)$, we get:

$$y_t(\tau) = -\alpha(\tau) + \beta(\tau) X_t + u_t(\tau) \tag{6.13}$$

where:

$$\alpha(\tau) = \frac{A(\tau)}{\tau} \quad , \quad \beta(\tau) = \frac{B(\tau)}{\tau},$$

$A(\tau)$ and $B(\tau)$ are model specific known functions of τ and θ, and $u_t(\tau)$ is the error term on the observation of the zero coupon rate $y_t(\tau)$. In matrix notation:

$$Y_t = -\alpha + \beta\ X_t + u_t \tag{6.14}$$

where:

$$\begin{aligned} Y_t &= [y_t(\tau_1), \cdots, y_t(\tau_n)]' \\ \alpha &= [\alpha(\tau_1), \cdots, \alpha(\tau_n)]' \\ \beta &= [\beta(\tau_1), \cdots, \beta(\tau_n)]' \end{aligned}$$

We consider the estimation of a one-factor term structure model, when the short term interest rate X_t follows either an Ornstein-Uhlenbeck process (Vasicek model) or a square root process (CIR model). In the two cases, the vector θ_1 of parameters about the factor dynamics can be written $\theta_1 = (k, c, \sigma)'$, but with two different interpretations for the volatility parameter σ:

$$\begin{aligned} &\text{Vasicek model} \quad : \\ dX_t &= k(c - X_t)\,dt + \sigma dW_t \end{aligned} \tag{6.15}$$

$$\begin{aligned} &\text{CIR model} \quad : \\ dX_t &= k(c - X_t)\,dt + \sigma\sqrt{X_t}dW_t \end{aligned} \tag{6.16}$$

In the two cases, there is one risk premium parameter $\theta_2 = \lambda$ which allows to compute the coefficients $A(\tau)/\tau$ and $B(\tau)/\tau$ of the column matrices α and β as known functions of θ:
(i) Vasicek model

$$\begin{cases} A(\tau) = x_\infty \left[\beta(\tau) - \tau\right] - \dfrac{\sigma^2}{4k} B^2(\tau) \\ B(\tau) = \dfrac{1 - \exp(-k\tau)}{k} \end{cases} \tag{6.17}$$

where:

$$x_\infty = c - \frac{\lambda}{k} - \frac{\sigma^2}{2k^2}$$

(ii) CIR model:

$$\begin{aligned} A(\tau) &= \frac{2kc}{\sigma^2} \ln \frac{2\gamma \exp\left[\frac{1}{2}(k + \lambda + \gamma)\tau\right]}{(k + \lambda + \gamma)\left[\exp(\gamma\tau) - 1\right] + 2\gamma} \\ B(\tau) &= \frac{2\left[\exp(\gamma\tau) - 1\right]}{(k + \lambda + \gamma)\left[\exp(\gamma\tau) - 1\right] + 2\gamma} \\ \gamma &= \sqrt{(k + \lambda)^2 + 2\sigma^2} \end{aligned}$$

Empirical Results

We first apply the Kalman filter approach of De Jong (2000), by considering that the covariance matrix Ω of the vector u_t of four Gaussian error terms is unconstrained, and defined by a vector η of ten parameters (specification 3 in Tables 1 and 2 below). We also apply the Kalman estimation technique to two models of constrained matrices Ω:

- In a first model (specification 1), we exclude any cross-correlation or heteroskedasticity of the error terms, then defining Ω by the common value η of the diagonal coefficients.
- In a second model (specification2), we allow for heteroskedasticity but still exclude any cross correlation. Then Ω is a diagonal matrix defined by a vector η of four diagonal coefficients.

Tables 1 and 2 respectively give the Kalman-based estimation results for the Vacisek and CIR one-factor term structure model.

Table 1. Kalman estimation results of the Vacisek model

Parameter	Specification of Ω		
	1	2	3
k	0.0612	0.0303	0.0213
	(0.0098)	(0.0046)	(0.0031)
c	0.0699	0.0741	0.0686
	(0.0048)	(0.0031)	(0.0025)
σ	0.0162	0.0153	0.0124
	(0.0015)	(0.0011)	(0.0011)
λ	-0.0070	-0.0082	-0.0074
	(0.0008)	(0.0010)	(0.0013)
ℓ_{K}	4378.72	4568.94	5004.32
dim η	1	4	10

Note: This table reports Kalman-based QML estimates and standard errors for the parameters of one-factor term structure model with short rate process $dX_{\mathrm{t}} = k\,(c - X_{\mathrm{t}})\,dt + \sigma dW_{\mathrm{t}}$. The table also reports the log-likelihood value and the dimension of the vector η of free parameters defining the observation errors covariance matrix Ω.

Table 2. Kalman estimation results of the CIR model

Parameter	Specification of Ω		
	1	2	3
k	0.1162	0.0640	0.0422
	(0.0140)	(0.0151)	(0.0092)
c	0.0634	0.0559	0.0587
	(0.0080)	(0.0136)	(0.0131)
σ	0.0646	0.0652	0.0462
	(0.0055)	(0.0044)	(0.0024)
λ	-0.0694	-0.0526	-0.0308
	(0.0143)	(0.0158)	(0.0091)
ℓ_{K}	4414.76	4645.71	5030.30
dim η	1	4	10

Note: This table reports Kalman-based QML estimates and standard errors for the parameters of one-factor term structure model with short rate process $dX_{\mathrm{t}} = k\,(c - X_{\mathrm{t}})\,dt + \sigma\sqrt{X_{\mathrm{t}}}dW_{\mathrm{t}}$. The table also reports the log-likelihood value and the dimension of the vector η of free parameters defining the observation errors covariance matrix Ω.

Columns 3 of Tables 1 and 2 replicate almost exactly the results reported by De Jong (2000) in Table 3, p. 305. It is worth noting that he uses a slightly different parameterization from the one adopted here and the observed discrepancy between estimation results may be due to numerical roundoff errors. By comparison, columns 1 and 2 of Tables 1 and 2 show that, as expected, the specification of the observation errors covariance matrix Ω does matter in term of estimation of the structural parameters θ. While the results are not that much different between cases 1, 2 and 3 for estimation of c, σ and λ, the mean reversion parameter k is highly sensitive to the specification of Ω. Less Ω is restricted, less mean reverting appears to be the short rate process.

Maximum Likelihood with $(n-1)$ measurement errors

We focus here on the exact maximum likelihood estimation of the structural parameters θ when a zero yield or a linear combination of zero yields is observed without error. Let us assume that the zero yield with maturity τ_1 is observed without error. In this case, (6.14) can be decomposed in two equations, up to a slight change in notations:

$$\begin{aligned} Z_t &= -\alpha_1 + \beta_1 X_t \\ V_t &= -\alpha_2 + \beta_2 X_t + u_t \end{aligned} \tag{6.18}$$

where u_t is now a $(n-1) \times 1$ random measurement error assumed to be i.i.d $\mathcal{N}(0, \Omega)$. Since the specification of Ω matters for the estimation of θ, we explore the same three cases previously encountered. Note that the number (dim η) of free parameters in η is now smaller since Ω has lost a row and a column. Namely, under case 1 of homoskedastic and uncorrelated measurement errors, dim $\eta = 1$; under case 2 of heteroskedastic uncorrelated errors dim $\eta = 3$; and, finally, under case 3 of heteroskedastic and correlated errors, dim $\eta = 6$.

Then, the aforementioned non-causality assumptions from error terms to the short rate process give rise to a loglikelihood function which is made up of two terms:

(i) The log-likelihood of the latent factor computed using the first zero yield, evaluated using the Jacobian formula

(ii) The log-likelihood of the $(n-1)$ measurement errors.

More precisely, let us denote with $\ell_x [X_t | X_{t-1}; \theta_1]$ the transition density function between consecutive observations of the short rate deduced from the continuous time model (Vacisek or CIR). The total sample loglikelihood is given by:

$$Q_T(\theta, \eta) = Q_{1T}(\theta) + Q_{2T}(\theta, \eta)$$

where:

$$Q_{1T}(\theta) = \sum_{t=1}^{T} \log \ \ell_x \left[\beta_1^{-1}(Z_t + \alpha_1) | \beta_1^{-1}(Z_{t-1} + \alpha_1); \theta_1\right] - T \log \beta_1 \tag{6.19}$$

is the sample loglikelihood of the latent factor, and:

$$Q_{2T}(\theta, \eta) = -\frac{T(n-1)}{2} \log(2\Pi) - \frac{T}{2} \log \ (\det(\Omega)) \tag{6.20}$$

$$-\frac{1}{2} \sum_{t=1}^{T} \left[V_t + \alpha_2 - \beta_2 \beta_1^{-1}(Z_t + \alpha_1)\right]' \Omega^{-1} \left[V_t + \alpha_2 - \beta_2 \beta_1^{-1}(Z_t + \alpha_1)\right]$$

is the sample loglikelihood of the measurement errors.

The transition density function $\ell_x [X_t | X_{t-1}, \theta_1]$ is Gaussian for the Vasicek model and non central chi square for the CIR model. In the latter case, it is also possible to compute a proxy of ℓ_x using a high order Aït-Sahalia (2002) analytical approximation technique.

The resulting expression is extremely precise and much faster to evaluate than the true theoretical non central chi square density.

However, in the present context of implied states, a few warnings are in order here. The approximation is not reliable for values of both the backward variable X_{t-1} and the forward one X_t that are close to the boundary of the admissible space. To some extent, this is acknowledged by Aït-Sahalia (2002), footnote 17, by considering more generally a time interval Δ between two consecutive discrete time observations: "[...] The expansion is known to deliver an approximation of the density function $x \longrightarrow p_X (\Delta, x \,|x_0; \theta)$ for a fixed value of the backward (conditioning) variable x_0. Therefore, except in the limit where Δ becomes infinitely small, it is not designed to reproduce the limiting behavior of p_X in the limit where x_0 tends to the boundaries.[...]".

To better understand what this means, it is useful to inspect the loglikelihood expansion of the CIR model (see also Aït-Sahalia and Kimmel (2002) for a two factors case). The fundamental difference with the Gaussian case is that the square root term leads to an expansion of $\log p_X (\Delta, x x_0; \theta)$ where the coefficients of Δ, Δ^2 and Δ^3 have powers of x and x_0 in the denominator. When either one of these variables tend to zero (the boundary), the approximation diverges either to plus or minus infinity.

If the factor realizations are available, and thus fixed in the expanded loglikelihood maximization, this is not an issue. In contrast, when X_t and X_{t-1} are not observable and are obtained as a function of the observable variable Z_t and Z_{t-1} and of the unknown parameters θ, this feature is devastating. Inevitably, any serious maximization algorithm will end up with an infinite value of the approximate log-likelihood for a value θ which sets to zero one or more implied values of the factors. The latent backfitting algorithm, however, will deliver implied values $X_t \left(\theta^{(p)}\right)$ and $X_{t-1} \left(\theta^{(p)}\right)$ for a given value $\theta^{(p)}$ of the structural parameters and allow to maximize with respect to θ the expansion of the latent log-likelihood obtained from the expansion of $\log p_X \left[\Delta, X_t \left(\theta^{(p)}\right) |X_{t-1} \left(\theta^{(p)}\right), \theta\right]$. There is no more problem with boundary values of X_t or X_{t-1}.

We first apply the standard maximum likelihood using (6.19) and (6.20). We get rid of the large number of local maxima by choosing the best maximization results obtained among 25 trials associated with different starting values of the parameters. These starting points where random draws from a multivariate uniform distribution with plausible upper and lower bounds.

Tables 3 and 4 report respectively the maximum estimation results for the Vasicek and CIR one factor term structure model. The first four columns (denoted with $y(\tau_i)$, $i = 1, 2, 3, 4$) assume that zero yields with maturity τ_i are observed without error. Column PC assumes that the first principal component (in descending order of the corresponding eigenvalues) computed on the yields with the usual four maturities is observed without error. Specifications 1,2,3 of the errors covariance matrix Ω, corresponding respectively to 1,3 and 6 free parameters η, give rise to the three sets of maximum likelihood estimation results.

Table 3. Maximum likelihood estimation results for the Vasicek model

	Yield (or combination of yields) without error				
Parameter	y(τ_1)	y(τ_2)	y(τ_3)	y(τ_4)	PC
			Specification 1, dim $\eta = 1$		
k	0.1047	0.0780	0.0304	0.0168	0.0584
	(0.0126)	(0.0090)	(0.0102)	(0.0089)	(0.0094)
c	0.0678	0.0706	0.0746	0.0990	0.0710
	(0.0546)	(0.0804)	(0.0658)	(0.2943)	(0.0816)
σ	0.0270	0.0263	0.0196	0.0150	0.0219
	(0.0028)	(0.0026)	(0.0015)	(0.0011)	(0.0021)
λ	-0.0065	-0.0050	-0.0041	-0.0026	-0.0046
	(0.0059)	(0.0064)	(0.0020)	(0.0050)	(0.0048)
$\ell_{\mathsf{ML}}\left(\hat{\theta}_{\mathsf{ML}}\right)$	4247.39	4428.02	4395.21	4329.32	4420.55
			Specification 2, dim $\eta = 3$		
k	0.1105	0.0801	0.0294	0.0219	0.0278
	(0.0145)	(0.0096)	(0.0044)	(0.0045)	(0.0030)
c	0.0682	0.0708	0.0771	0.0984	0.0651
	(0.0525)	(0.0869)	(0.1124)	(0.2037)	(0.2428)
σ	0.0276	0.0268	0.0206	0.0160	0.0207
	(0.0030)	(0.0028)	(0.0016)	(0.0011)	(0.0018)
λ	-0.0068	-0.0051	-0.0031	-0.0019	-0.0033
	(0.0059)	(0.0070)	(0.0033)	(0.0045)	(0.0069)
$\ell_{\mathsf{ML}}\left(\hat{\theta}_{\mathsf{ML}}\right)$	4302.34	4443.22	4629.00	4574.34	4929.85
			Specification 3, dim $\eta = 6$		
k	0.0329	0.0319	0.0277		0.0292
	(0.0030)	(0.0037)	(0.0028)		(0.0030)
c	0.0505	0.0579	0.0784		0.0676
	(0.3540)	(2.3778)	(0.3647)		(0.0809)
σ	0.0258	0.0248	0.0185		0.0204
	(0.0021)	(0.0020)	(0.0012)		(0.0017)
λ	-0.0048	-0.0044	-0.0024		-0.0031
	(0.0114)	(0.0759)	(0.0101)		(0.0024)
$\ell_{\mathsf{ML}}\left(\hat{\theta}_{\mathsf{ML}}\right)$	4871.18	4882.30	4957.89		4934.30

Note: This table reports maximum likelihood estimations and asymptotic robust standard errors for the parameters of one-factor term structure model with short rate process $dX_{\mathrm{t}} = k\left(c - X_{\mathrm{t}}\right)dt + \sigma dW_{\mathrm{t}}$. The table also reports the log likelihood value and the dimension of the vector η of free parameters defining the observation errors covariance matrix Ω. The first four columns (denoted with $y(\tau_i), i = 1, 2, 3, 4)$ assume respectively that zero yields with maturity $\tau = 3$ months, 1, 5 or 10 years are observed without error. Column PC assumes instead that the first principal component among the yields with these four maturities is observed without error.

Table 4. Maximum likelihood estimation results for the CIR model

	Yield (or combination of yields) without error				
Parameter	y(τ_1)	y(τ_2)	y(τ_3)	y(τ_4)	PC
			Specification 1, dim $\eta = 1$		
k	0.1711	0.1185	0.0714	0.0403	0.1012
	(0.0504)	(0.0227)	(0.1313)	(0.0480)	(0.0248)
c	0.0711	0.0740	0.0776	0.0872	0.0742
	(0.0216)	(0.0130)	(0.1394)	(0.1042)	(0.0165)
σ	0.0871	0.0834	0.0650	0.0507	0.0707
	(0.0074)	(0.0064)	(0.0044)	(0.0034)	(0.0053)
λ	-0.0809	-0.0575	-0.0498	-0.0314	-0.0550
	(0.0532)	(0.0199)	(0.1287)	(0.0486)	(0.0214)
$\ell_{\mathsf{ML}}\left(\hat{\theta}_{\mathsf{ML}}\right)$	4046.11	4222.04	4180.02	4114.13	4213.85
			Specification 2, dim $\eta = 3$		
k	0.1831	0.1234	0.0451	0.0326	0.0406
	(0.0316)	(0.0156)	(0.0230)	(0.0182)	(0.0148)
c	0.0714	0.0742	0.0792	0.0927	0.0704
	(0.0104)	(0.0071)	(0.0414)	(0.0498)	(0.0248)
σ	0.0885	0.0846	0.0649	0.0520	0.0638
	(0.0077)	(0.0066)	(0.0043)	(0.0031)	(0.0042)
λ	-0.0853	-0.0595	-0.0036	-0.0220	-0.0345
	(0.0248)	(0.0104)	(0.0234)	(0.0171)	(0.0126)
$\ell_{\mathsf{ML}}\left(\hat{\theta}_{\mathsf{ML}}\right)$	4100.32	4236.86	4410.79	4358.00	4716.35
			Specification 3, dim $\eta = 6$		
k	0.0533	0.0479		0.0262	0.0420
	(0.1919)	(0.0200)		(0.0262)	(0.0040)
c	0.0603	0.0655		0.0873	0.0708
	(0.2167)	(0.0413)		(0.0237)	(0.0067)
σ	0.0787	0.0753		0.0495	0.0631
	(0.0050)	(0.0046)		(0.0026)	(0.0042)
λ	-0.0527	-0.0456		-0.0154	-0.0035
	(0.1894)	(0.0313)		(0.0064)	(0.0025)
$\ell_{\mathsf{ML}}\left(\hat{\theta}_{\mathsf{ML}}\right)$	4653.22	4664.11		4777.18	4720.97

Note: This table reports maximum likelihood estimations and asymptotic robust standard errors for the parameters of one-factor term structure model with short rate process $dX_t = k\left(c - X_t\right)dt + \sigma\sqrt{X_t}dW_t$. The table also reports the log likelihood value and the dimension of the vector η of free parameters defining the observation errors covariance matrix Ω. The first four columns (denoted with $y(\tau_i), i = 1, 2, 3, 4)$ assume respectively that zero yields with maturity $\tau = 3$ months, 1, 5 or 10 years are observed without error. Column PC assumes instead that the first principal component among the yields with these four maturities is observed without error.

In two cases it has not been possible to obtain acceptable results in terms of multiple trials converging to the same estimates. These cases are left empty in the tables. These cases correspond to the assumption that the yield without measurement error has maturity either 5 or 10 years and to the richest structure of the observation errors covariance matrix Ω (dim $\eta = 6$). It is possible that by augmenting the number of starting points, or by changing the way they are chosen, an acceptable solution can be found even in those cases. However, it is clear that these specifications lend themselves to numerical instability, at least compared with those stemming from simpler models (i.e. with a simpler structure for Ω) and/or based on the assumption that the yields observed without measurement error are associated with the shortest maturities.

In all the tables, the results depend clearly on the yield assumed to be observed without error.

If we denote with τ^* its maturity, a clear pattern emerges: k, σ and λ decrease monotonically with τ^*, whereas c increases. To decide what are the most plausible values, it is useful to compare the results with the corresponding ones obtained with the Kalman filter. As a matter of fact, since the Kalman filter does not impose any zero error term, its results may be considered as more robust. A striking result for the three specifications of Ω and the two models is that the maximum likelihood estimates the closest to Kalman results are obtained in the column PC. In other words, it appears to be sensible to assume that the combination of yields observed without error is the one provided by the first principal component. Note that in both Tables 3 and 4, the column PC follows closely the pattern highlighted above, that is its results are placed somewhere between those in columns $y(\tau_2)$ and $y(\tau_3)$ (respectively corresponding to maturities 1 and 5 years) while the weighted average maturity for the principal component is 3.67 years.

Overall, the relationship between estimates and maturity of the yield without error is strongest with the most parsimonious specifications of Ω (specifications 1 and 2), and is weaker when no restrictions are placed on Ω (specification 3). This last case is also clearly to be preferred on the basis of the value of the maximized loglikelihood.

Generally speaking, the maximum likelihood estimation results with the principal component assumed to be observed without pricing error are quite close to the estimates based on Kalman filter. Not surprisingly, the maximum likelihood standard errors are rather smaller, particularly in the case of the mean reversion parameter k. The challenge for the implied states backfitting methodology is then, besides its computational advantages, to deliver estimators with an accuracy similar to maximum likelihood.

Tables 5 and 6 respectively report the extended backfitting estimation results for the Vasicek and CIR one-factor term structure model. The definition of columns and rows of these tables mimics the one used for Tables 3 and 4. However, when the longest maturity (10 years) yield is assumed to be free of measurement error, it has not been possible to obtain acceptable results in terms of multiple trials converging to the same estimates. Of course, increasing the number of trials would may be solve the problem. Following the same strategy as for maximum likelihood, we have preferred not to give results in this case. This is the reason why Tables 5 and 6 do not include the column $y(\tau_4)$. The results obtained are very similar to maximum likelihood and we follow the same argument to consider that the most reliable estimation results are obtained when the principal component is assumed to be free of measurement error.

Table 5. Extended backfitting estimation results for the Vasicek model

	Yield (or combination of yields) without error			
Parameter	y(τ_1)	y(τ_2)	y(τ_3)	PC
	Specification 1, dim $\eta = 1$			
k	0.1052	0.0790	0.0301	0.0573
	(0.0128)	(0.0088)	(0.0032)	(0.0096)
c	0.0680	0.0711	0.0796	0.0712
	(0.0490)	(0.0722)	(0.1660)	(0.0960)
σ	0.0272	0.0253	0.0177	0.0219
	(0.0028)	(0.0024)	(0.0012)	(0.0021)
λ	-0.0065	-0.0047	-0.0022	-0.0045
	(0.0053)	(0.0058)	(0.0050)	(0.0055)
	Specification 2, dim $\eta = 3$			
k	0.1122	0.0812	0.0292	0.0278
	(0.0153)	(0.0092)	(0.0053)	(0.0031)
c	0.0684	0.0713	0.0869	0.0696
	(0.0478)	(0.0658)	(0.3334)	(0.1212)
σ	0.0281	0.0252	0.0178	0.0207
	(0.0032)	(0.0023)	(0.0012)	(0.0018)
λ	-0.0070	-0.0047	-0.0022	-0.0031
	(0.0055)	(0.0054)	(0.0097)	(0.0035)
	Specification 3, dim $\eta = 6$			
k	0.0327	0.0362	0.0259	0.0294
	(0.0030)	(0.0034)	(0.0027)	(0.0030)
c	0.0572	0.0643	0.0955	0.0667
	(0.1542)	(0.3061)	(0.3075)	(0.0067)
σ	0.0267	0.0223	0.0186	0.0204
	(0.0022)	(0.0016)	(0.0014)	(0.0017)
λ	-0.0249	-0.0035	-0.0020	-0.0031
	(0.0053)	(0.0112)	(0.0080)	(0.0087)

Note: This table reports extended backfitting estimations and asymptotic robust standard errors for the parameters of one-factor term structure model with short rate process $dX_{\rm t} = k\,(c - X_{\rm t})\,dt + \sigma dW_{\rm t}$. The table also reports the dimension of the vector η of free parameters defining the observation errors covariance matrix Ω. The first three columns (denoted with $y(\tau_i), i = 1, 2, 3)$ assume respectively that zero yields with maturity $\tau = 3$ months, 1 or 5 years are observed without error. Column PC assumes instead that the first principal component among the four yields is observed without error.

Table 6. Extended backfitting estimation results for the CIR model

	Yield (or combination of yields) without error			
Parameter	y(τ_1)	y(τ_2)	y(τ_3)	PC
	Specification 1, dim $\eta = 1$			
k	0.1722	0.1175	0.0378	0.0992
	(0.0116)	(0.0722)	(0.0055)	(0.0352)
c	0.0711	0.0739	0.0800	0.0741
	(0.0468)	(0.0463)	(0.0126)	(0.0278)
σ	0.0876	0.0806	0.0566	0.0706
	(0.0076)	(0.0558)	(0.0033)	(0.0054)
λ	-0.0816	-0.0550	-0.0243	-0.0548
	(0.1150)	(0.0743)	(0.0054)	(0.0381)

	Specification 2, dim $\eta = 3$			
k	0.1870	0.1206	0.0410	0.0406
	(0.0427)	(0.0100)	(0.0134)	(0.0061)
c	0.0715	0.0741	0.0796	0.0705
	(0.0172)	(0.0062)	(0.0256)	(0.0103)
σ	0.0896	0.0804	0.0570	0.0636
	(0.0081)	(0.0057)	(0.0034)	(0.0042)
λ	-0.0872	-0.0051	-0.0271	-0.0343
	(0.0460)	(0.0096)	(0.0130)	(0.0044)
	Specification 3, dim $\eta = 6$			
k	0.0531	0.0522	0.0332	0.0419
	(0.8427)	(0.0083)	(0.0038)	(0.0074)
c	0.0604	0.0670	0.0793	0.0709
	(0.9546)	(0.0108)	(0.0087)	(0.0123)
σ	0.0805	0.0682	0.0592	0.0629
	(0.0051)	(0.0037)	(0.0039)	(0.0041)
λ	-0.0543	-0.0391	-0.0267	-0.0031
	(0.8427)	(0.0071)	(0.0033)	(0.0062)

Note: This table reports extended backfitting estimations and asymptotic robust standard errors for the parameters of one-factor term structure model with short rate process $dX_{\mathrm{t}} = k\left(c - X_{\mathrm{t}}\right) dt + \sigma\sqrt{X_{\mathrm{t}}} dW_{\mathrm{t}}$. The table also reports the dimension of the vector η of free parameters defining the observation errors covariance matrix Ω. The first three columns (denoted with $y(\tau_i), i = 1, 2, 3)$ assume respectively that zero yields with maturity $\tau = 3$ months, 1 or 5 years are observed without error. Column PC assumes instead that the first principal component among the four yields is observed without error.

About the content of Tables 5 and 6, a couple of comments are in order.

First, as in Tables 3 and 4, each estimated parameter vector is chosen among 25 different trials corresponding to an equal number of starting points. Contrary to the maximum likelihood strategy, however, the backfitting approach does not provide an easy way to discriminate between two or more alternative sample fixed points, which are sometimes observed in these explorations. After all, even if the uniqueness of the fixed point of the asymptotic map $\bar{\theta}\left[P^0, \lambda\left(\cdot\right)\right]$ is guaranteed by assumption 2.3., it may very well be the case that for finite T the sample map exhibits multiple fixed points, each one associated with plausible estimates for θ. The results outlined in the Tables 5 and 6 correspond to those with highest value of the latent criterion.

Second, to detect convergence to a fixed point, we used the following criterion:

$$\begin{aligned} & \left\|\theta^{(p+1)} - \theta^{(p)}\right\| \\ = \; & Max\left\{\left|k^{(p+1)} - k^{(p)}\right|, \left|c^{(p+1)} - c^{(p)}\right|, \left|\sigma^{(p+1)} - \sigma^{(p)}\right|, \left|\lambda^{(p+1)} - \lambda^{(p)}\right|\right\} < 10^{-5}. \end{aligned}$$

Note that, following the philosophy of extended backfitting, this criterion does not check the convergence of η. As already explained, there is no reason to assume that the contraction mapping property remains fulfilled when η is included in the list of structural parameters. We repeatedly observed that convergence was significantly slowed down by including η in the criterion, although θ had already settled down to a constant between iterations. Also, note that the critical threshold (10^{-5}) is rather probing, since it is equal to the critical threshold used to detect convergence in the estimation step of the backfitting approach. Stated otherwise, $\theta^{(p)}$ is itself precise to the fifth decimal. Moreover, each estimation step is started at the previous estimated parameter values. In other words, to estimate $\theta^{(p+1)}$, iterations are started at $\theta^{(p)}$.

6.4 Lessons from the application

The first striking conclusion of the empirical results reported above is that the efficiency loss resulting from the replacement of genuine maximum likelihood estimation by extended backfitting is negligible. Moreover, it is worthwhile to realize that the backfitting procedure is more robust than its competitors with respect to specification errors in the pricing errors dynamics and with respect to boundary problems in the domain of state variables.

Note also that one could even push the backfitting principle further by including the principal components strategy (to decide the combination of yields that is free of measurement errors) inside the backfitting iteration. In other words, for a given value $\theta^{(p)}$ of the structural parameters, one could look for the most negligible combination of pricing errors. This extension is left for future research.

As far as computing times are concerned, several comments are in order.

First, we have chosen the affine model to allow for a sensible comparison between backfitting and Kalman filtering. Of course, the computational advantages of the backfitting strategy would be much more striking in more severely nonlinear asset pricing models like models of options on equity with stochastic volatility. As explained in section 3, these computational advantages explain that Pan (2002) had chosen an IS-GMM approach asymptotically equivalent to implied state backfitting.

Second, we have renounced to report in the tables of results above any consideration about the relative and absolute speed, in terms of number of iterations and time to convergence, or about the number of trials which actually converged of the estimation strategies. Yet, we did observe that simple applications of the backfitting strategy are a matter of seconds while even the Kalman filter approach needs a computing time several times larger.

7 Concluding remarks

In this article, we have developed a new inference method, called latent backfitting. We argued that it is in general much more efficient, both on computational and statistical grounds, than standard filtering or simulation-based methods, when applied to asset pricing models with latent state variables. The main reason of this efficiency gain is that the implied-states methodology fully takes advantage of the one-to-one relationship between latent state variables and observed asset prices.

Moreover, we have shown how the performance assessment of such implied-states inference must disentangle the two key ingredients of the inference strategy: informational content of the latent model on the one hand, contracting feature of the backfitting mapping on the other hand. In a very simple empirical example, we report some evidence of the practical advantages of the latent backfitting method with respect to more common likelihood based methods.

However, a lot of work remains to be done to fully exploit the advantages of approach we propose. The two main directions for future empirical research are the following. First, the Jacobian matrix $\partial\bar{\theta}/\partial\theta^{1\prime}$ of the backfitting function is the crucial ingredient of any rigorous statistical inference based on our iterative or recursive theories. The role of this matrix has been overlooked in the literature when people think that we are allowed to use the implied state variables in the estimation as if they were directly observable. On the other hand, this is precisely because we consider that it will be better to avoid the direct computation of the Jacobian for the inverse transformation that we propose the backfitting approach. Therefore, it is the spirit of our approach to look for the value of this matrix only in a second stage, when estimation has been performed. A numerical assessment of this Jacobian matrix is a byproduct of our iterative algorithm. An additional issue to address would be a performance assessment of various possible extended backfitting strategies that may lead to rather different contracting properties of the backfitting function.

The second direction for future research is the definition of practical guidelines for implementation of the recursive approach. While the iterative approach is largely sufficient for a simple affine one-factor model as considered in section 6, the recursive approach should be tremendously advantagous in less user-friendly models. The basic idea is that it does not make sense to filter all the state values as long as the estimation algorithm has not reached sensible values. While the Kalman filter has proven its usefulness in Gaussian linear state-space models, the recursive strategies we propose here should be the best approaches to non-linear state space structural models of asset pricing. However, people familiar with Robbins-Monro type algorithms know that the empirical performance of such algorithms are highly dependent on key parameters like the learning rate, the Newton-Raphson type devices and the truncation scheme. Therefore, there is still a need of work regarding recursive approaches in asset pricing to get user-friendly methodologies that are fully reliable in practice.

Appendix

A.1 Identifiability in binary response models

Let us note that, typically, the identification condition **C2'** (see subsection 2.3) amounts to the identification of θ from the observed variables. Let us consider, for example, a general binary response model without any specific assumption on the latent regression function $h(X_t, \theta^1)$. We have

$$\begin{aligned} E\left[u_t \mid Y_t, X_t; \theta^1\right] = {} & Y_t E\left[u_t \mid u_t > -h(X_t, \theta^1),\, X_t\right] \\ & + (1 - Y_t)\, E\left[u_t \mid u_t \leq -h(X_t, \theta^1),\, X_t\right], \end{aligned}$$

where

$$E\left[u_t \mid u_t > -h(X_t, \theta^1),\, X_t\right] = \frac{E\left[u_t\, \mathbf{1}\left\{u_t > -h(X_t, \theta^1)\right\} \mid X_t\right]}{E\left[\mathbf{1}\left\{u_t > -h(X_t, \theta^1)\right\} \mid X_t\right]} \overset{not}{=} \frac{\Psi\left(X_t; \theta^1\right)}{p\left(X_t; \theta^1\right)}$$

and

$$E\left[u_t \mid u_t \leq -h(X_t, \theta^1),\, X_t\right] = -\frac{\Psi\left(X_t; \theta^1\right)}{1 - p\left(X_t; \theta^1\right)}.$$

Since, by definition

$$E\left[Y_t \mid X_t\right] = \mathsf{P}\left[Y_t = 1 \mid X_t\right] = p\left(X_t; \theta^0\right),$$

we obtain

$$\begin{aligned} E\left[E\left[u_t \mid Y_t, X_t; \theta^1\right] \mid X_t\right] &= E\left[Y_t \mid X_t\right] \frac{\Psi\left(X_t; \theta^1\right)}{p\left(X_t; \theta^1\right)} - \left(1 - E\left[Y_t \mid X_t\right]\right) \frac{\Psi\left(X_t; \theta^1\right)}{1 - p\left(X_t; \theta^1\right)} \\ &= \Psi\left(X_t; \theta^1\right) \left[\frac{p\left(X_t; \theta^0\right)}{p\left(X_t; \theta^1\right)} - \frac{1 - p\left(X_t; \theta^0\right)}{1 - p\left(X_t; \theta^1\right)}\right]. \end{aligned}$$

Thus, condition **C2'** is equivalent to the observable model identification condition

$$p\left(X_t; \theta^1\right) = p\left(X_t; \theta^0\right) \quad \Longrightarrow \quad \theta^1 = \theta^0,$$

provided that, for any $\theta^1 \in \Theta$, the range of $h(X_t, \theta^1)$ is contained in the interior of the support of u_t and thus $\Psi\left(X_t; \theta^1\right) \neq 0$.

A.2 About minimizing the squares of the generalized residuals

Let us note that the binary choice model provides an example of criterion $Q_T[\theta, \lambda\left(\theta\right)]$ which has not to be maximized with respect to both occurrences of θ. Indeed,

$$\begin{aligned} Q_\infty[\theta, \lambda\left(\theta^1\right)] &= -E\left[\left[Y_t^*(\theta^1) - h(X_t, \theta)\right]^2\right] \\ &= -E\left[\left[h(X_t, \theta^1) - h(X_t, \theta)\right]^2\right] - E\left[E\left[u_t \mid Y_t, X_t; \theta^1\right]^2\right] \\ &= -E\left[\left[h(X_t, \theta^1) - h(X_t, \theta)\right]^2\right] - E\left[\tilde{u}_t\left(\theta^1\right)^2\right]. \end{aligned}$$

Since for $\theta = \theta^1$ the first term after the second equality vanishes, it remains to check that $\theta^1 = \theta^0$ does not necessarily yield the smaller value for $E\left[\tilde{u}_t\left(\theta^1\right)^2\right]$. Recall that we have

$$\begin{aligned} E\left[u_t \mid Y_t, X_t; \theta^1\right] &= Y_t E\left[u_t \mid u_t > -h(X_t, \theta^1),\, X_t\right] \\ &\quad + (1 - Y_t)\, E\left[u_t \mid u_t \leq -h(X_t, \theta^1),\, X_t\right], \end{aligned}$$

where

$$E\left[u_t \mid u_t > -h(X_t, \theta^1),\, X_t\right] = \frac{E\left[u_t\, \mathbf{1}\left\{u_t > -h(X_t, \theta^1)\right\} \mid X_t\right]}{E\left[\mathbf{1}\left\{u_t > -h(X_t, \theta^1)\right\} \mid X_t\right]} \overset{not}{=} \frac{\Psi\left(X_t; \theta^1\right)}{p\left(X_t; \theta^1\right)}$$

and

$$E\left[u_t \mid u_t \leq -h(X_t, \theta^1),\, X_t\right] = -\frac{\Psi\left(X_t; \theta^1\right)}{1 - p\left(X_t; \theta^1\right)}.$$

Thus

$$E\left[u_t \mid Y_t, X_t; \theta^1\right]^2 = Y_t\left(\frac{\Psi\left(X_t; \theta^1\right)}{p\left(X_t; \theta^1\right)}\right)^2 + (1 - Y_t)\left(\frac{\Psi\left(X_t; \theta^1\right)}{1 - p\left(X_t; \theta^1\right)}\right)^2.$$

Let $p_0 = E\left[Y_t \mid X_t\right]$, $p_1 = p\left(X_t; \theta^1\right)$ and $\Psi_1 = \Psi\left(X_t; \theta^1\right)$. Then,

$$E\left[\tilde{u}_t\left(\theta^1\right)^2\right] = E\left[\Psi_1^2\left(\frac{p_0}{p_1^2} + \frac{1 - p_0}{(1 - p_1)^2}\right)\right]$$

and there is no reason to believe that the last expectation is minimized for $\theta^1 = \theta^0$.

A.3 A matrix algebra lemma

Lemma A.1 *Let A and B symmetric matrices such that $A \gg B \gg 0$ and A is invertible. Then,*

a) the eigenvalues of $A^{-1}B$ lie in $[0, 1]$; if $A - B$ is positive definite, then the eigenvalues are smaller than one.

b) if, in addition, $A^2 - B^2$ is positive definite, then $\|A^{-1}B\| < 1$.

Proof a) For any $\lambda > 1$, $\lambda A - B$ is positive definite and thus

$$\det\left(\lambda I - A^{-1}B\right) = \det\left(A^{-1}\right)\det\left(\lambda A - B\right) > 0.$$

If $A - B$ is positive definite the inequality still holds for $\lambda = 1$. On the other hand,

$$\det\left(\lambda I - A^{-1}B\right) = \det\left(A^{-1}\right)\det\left(\lambda A - B\right) = \det\left(\lambda I - A^{-1/2}BA^{-1/2}\right).$$

In other words, the eigenvalue of $A^{-1}B$ coincide with those of the positive semidefinite matrix $A^{-1/2}BA^{-1/2}$. We can conclude that the eigenvalues of $A^{-1}B$ lie in $[0, 1]$ if $A \gg B$ and they are certainly smaller than one if $A - B$ is positive definite.

b) The square of the norm of $A^{-1}B$ is the largest eigenvalue of $BA^{-2}B$. Since $A^2 - B^2$ positive definite implies $B^{-2} - A^{-2}$ positive definite, for any $\lambda \geq 1$ we have $\lambda B^{-2} - A^{-2}$ positive definite and therefore we can write

$$\det\left(\lambda I - BA^{-2}B\right) = \det(B)^2 \det\left(\lambda B^{-2} - A^{-2}\right) > 0.$$

That is, all the eigenvalues of $BA^{-2}B$ lie in $[0, 1)$. ■

A.4 Consistency of the latent backfitting

Proof of Proposition 4.2 The proof follows the steps of the usual weak consistency proof for argmax estimators (see, *e.g.,* Newey and McFadden (1994), page 2121). If $\theta^1 \in \Theta$, then

$$Q_T\left[\overline{\theta}\left[P^0, \lambda(\theta^1)\right], \lambda(\theta^1)\right] \leq Q_T\left[\overline{\theta}_T(\lambda(\theta^1)), \lambda(\theta^1)\right]. \tag{.1}$$

Let $\eta > 0$. Then,

$$\lim_{T\to\infty} \mathsf{P}\left(\sup_{\theta^1\in\Theta}\left\{Q_T\left[\overline{\theta}_T(\lambda(\theta^1)),\, \lambda(\theta^1)\right] - Q_\infty\left[\overline{\theta}_T(\lambda(\theta^1)), \lambda(\theta^1)\right]\right\} < \eta/2\right) = 1 \tag{.2}$$

and

$$\lim_{T\to\infty} \mathsf{P}\left(\sup_{\theta^1\in\Theta}\left\{Q_\infty\left[\overline{\theta}\left[P^0, \lambda(\theta^1)\right], \lambda(\theta^1)\right] - Q_T\left[\overline{\theta}\left[P^0, \lambda(\theta^1)\right], \lambda(\theta^1)\right]\right\} < \eta/2\right) = 1 \tag{.3}$$

(see Assumption 4.1). From (.1) to (.3) we obtain that, for any $\eta > 0$

$$\lim_{T\to\infty} \mathsf{P}\left(\sup_{\theta^1\in\Theta}\left\{Q_\infty\left[\overline{\theta}\left[P^0, \lambda(\theta^1)\right], \lambda(\theta^1)\right] - Q_\infty\left[\overline{\theta}_T(\lambda(\theta^1)), \lambda(\theta^1)\right]\right\} < \eta\right) = 1. \tag{.4}$$

Let

$$Q(V) = \inf_{(\theta,\theta^1)\in V}\left\{Q_\infty\left[\overline{\theta}\left[P^0, \lambda(\theta^1)\right], \lambda(\theta^1)\right] - Q_\infty\left[\theta,\, \lambda(\theta^1)\right]\right\}, \qquad V \subset \Theta^2 = \Theta \times \Theta,$$

and, for $\varepsilon > 0$, define the set

$$N_\varepsilon = \left\{(\theta, \theta^1) \in \Theta^2; \quad \left\|\theta - \overline{\theta}\left[P^0, \lambda(\theta^1)\right]\right\| < \varepsilon\right\}$$

which is supposed not to be empty. The continuity of $\overline{\theta}\left[P^0, \lambda(\cdot)\right]$ implies that $\Theta^2 \setminus N_\varepsilon$ is a compact set. Moreover, since the function

$$\left(\theta, \theta^1\right) \to Q_\infty\left[\overline{\theta}\left[P^0, \lambda(\theta^1)\right], \lambda(\theta^1)\right] - Q_\infty\left[\theta,\, \lambda(\theta^1)\right]$$

is continuous and $\overline{\theta}\left(P^0, \lambda\right)$ is the unique of $Q_\infty\left[\theta,\, \lambda\right]$, necessarily

$$Q\left(\Theta^2 \setminus N_\varepsilon\right) > 0. \tag{.5}$$

Now, define the set

$$A_T = \left\{ \left(\overline{\theta}_T(\lambda(\theta^1)), \theta^1 \right) ; \ \theta^1 \in \Theta \right\} \subset \Theta^2.$$

Note that the uniform convergence in probability of $\overline{\theta}_T(\lambda(\cdot))$ is equivalent to

$$\mathsf{P}\left(A_T \subset N_\varepsilon\right) \to 1, \quad T \to \infty, \tag{.6}$$

for any $\varepsilon > 0$. If, for some $\varepsilon' > 0$, A_T is not included in $N_{\varepsilon'}$, then there exists $\theta_T^1 \in \Theta$ (depending on ε' and the sample) such that $\left\|\overline{\theta}\left[P^0, \lambda(\theta_T^1)\right] - \overline{\theta}_T(\lambda(\theta_T^1))\right\| > \varepsilon'$. Since then $\left(\overline{\theta}_T(\lambda(\theta_T^1)), \theta_T^1\right) \in \Theta^2 \setminus N_{\varepsilon'}$, invoke (.5), take $0 < \eta < Q(\Theta^2 \setminus N_{\varepsilon'})$ and derive that

$$Q_\infty\left[\overline{\theta}\left[P^0, \lambda(\theta_T^1)\right], \lambda(\theta_T^1)\right] - Q_\infty\left[\overline{\theta}_T(\lambda(\theta_T^1)), \lambda(\theta_T^1)\right] > \eta,$$

with η depending only on ε'. Consequently,

$$\{A_T \subset N_{\varepsilon'}\}^c \subset \left\{ \sup_{\theta^1 \in \Theta} \left\{ Q_\infty\left[\overline{\theta}\left[P^0, \lambda(\theta^1)\right], \lambda(\theta^1)\right] - Q_\infty\left[\overline{\theta}_T(\lambda(\theta^1)), \lambda(\theta^1)\right] \right\} > \eta \right\}.$$

Therefore, assuming that (.6) fails is contradicted by (.4) and thus, the proof is complete. ■

Proof of Proposition 4.4 From the contracting property stated in Assumption 4.3 we get, for any $T \geq 1$

$$\begin{aligned} \left\|\widehat{\theta}_T - \theta^0\right\| &\leq \left\|\overline{\theta}_T\left(\lambda(\theta_T^{(p(T))})\right) - \overline{\theta}\left[P^0, \lambda(\theta_T^{(p(T))})\right]\right\| \\ &\quad + \left\|\overline{\theta}\left[P^0, \lambda(\theta_T^{(p(T))})\right] - \overline{\theta}\left[P^0, \lambda\left(\theta^0\right)\right]\right\| \\ &\leq \sup_{\theta^1 \in \Theta} \left\|\overline{\theta}_T(\lambda(\theta^1)) - \overline{\theta}\left[P^0, \lambda(\theta^1)\right]\right\| + k\left\|\theta_T^{(p(T))} - \theta^0\right\| \\ &\leq \left(1 + ... + k^{p(T)-1}\right) \sup_{\theta^1 \in \Theta} \left\|\overline{\theta}_T(\theta^1) - \overline{\theta}\left[P^0, \lambda(\theta^1)\right]\right\| + k^{p(T)}\left\|\theta_T^{(1)} - \theta^0\right\|. \end{aligned}$$

Finally, use Proposition 4.2, $k \in (0, 1)$, $p(T) \to \infty$ and the fact that Θ is a bounded set in order to complete the proof. ■

A.5 Asymptotic normality for the latent backfitting

Proof of Proposition 4.8 For simpler writings denote

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

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

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

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

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

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

$$0 = \sqrt{T} S_T^i(\theta^0, \theta^0) + \left(\frac{\partial S_T^i}{\partial \theta'}(\theta, \theta_T^{(p(T))})\bigg|_{\theta=\tilde{\theta}^{\mathrm{i}}_{\mathrm{T}}} + \frac{\partial S_T^i}{\partial \theta^{1\prime}}(\theta^0, \theta^1)\bigg|_{\theta^1 = \tilde{\tilde{\theta}}{}^{\mathrm{i}}_{\mathrm{T}}} \right) \sqrt{T}(\widehat{\theta}_T - \theta^0)$$

$$+ \frac{\partial S_T^i}{\partial \theta^{1\prime}}(\theta^0, \theta^1)\bigg|_{\theta^1 = \tilde{\tilde{\theta}}{}^{\mathrm{i}}_{\mathrm{T}}} \sqrt{T}(\theta_T^{(p(T))} - \widehat{\theta}_T).$$

We can derive

$$\sqrt{T}(\widehat{\theta}_T - \theta^0) = \left[\Sigma(\theta^0, \theta^0) - H(\theta^0)\right]^{-1} \sqrt{T} S_T(\theta^0, \theta^0) + o_P(1)$$

which implies

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

with $V(\theta^0)$ as in (4.8). ■

A.6 Consistency of the recursive latent backfitting

Kuan and White (1994a) (see also Kuan and White (1994b)) provide a set of general assumptions for proving the almost sure convergence of stochastic approximation (Robbins-Monro) procedures; see their Assumptions A.1 to A.5. Assumption A.1 introduces the data generating process denoted by $\{Z_t\}$ and taking values in R^s, $s \geq 1$, while the Assumption A.4 automatically hold for a learning rate $a_t = ct^{-1}$, $t = 1, 2, ...$ with c a positive constant. Let us recall the remaining three assumptions of Kuan and White (1994a). The general RM procedure they consider is

$$\widehat{\theta}_{t+1} = \widehat{\theta}_t + a_t \psi\left(Z_t, \widehat{\theta}_t\right), \qquad t = 1, 2, ..., \tag{.7}$$

and it approximates θ^0, the zero of a function $\Psi(\theta)$ defined on a compact set $\Theta \subset \mathsf{R}^k$ with values in R^k.

The Euclidean spaces R^k, $k \geq 1$ are endowed with the Borel $\sigma-$fields and the function $\psi : \mathsf{R}^s \times \Theta \rightarrow \mathsf{R}^p$ is measurable. The following three assumptions correspond, respectively, to Assumption A.2, A.3 and A.5 of Kuan and White (1994a). For our purposes $\{Z_t\} = \{Y_t, X_t\}$, θ becomes $\left(vec(G)', \theta^{1\prime}\right)'$, θ^0 is transformed in $\left(vec(\Sigma\left(\theta^0\right))', \theta^{0\prime}\right)'$, while the parameters set Θ is replaced by $\Theta \times \overline{U}$, where $\overline{U} = \{vec(G),\ G\,\text{non singular}\} \subset \mathsf{R}^{p^2+p}$ is a compact set including $vec(\Sigma(\theta^0))$. Moreover,

$$\psi\left(Z_t, \left(vec(G)', \theta^{1\prime}\right)'\right) = M_t^{RM}\left(\left(vec(G)', \theta^{1\prime}\right)'\right),$$

$$\Psi\left(\left(vec(G)', \theta^{1\prime}\right)'\right) = M^{RM}\left(\left(vec(G)', \theta^{1\prime}\right)'\right)$$

with M_t^{RM} and M^{RM} defined in (5.11) and (5.10), respectively.

Assumption A.0.1 *i) There exist functions $b : \Theta \rightarrow \mathsf{R}^+$ continuous and $h_i : \mathsf{R}^s \rightarrow \mathsf{R}^+$, $i = 1, 2$ measurables such that*

$$\|\psi(z,\theta)\| \leq b\,(\theta)\,h_1(z) + h_2(z), \qquad\qquad (z,\theta) \in \mathsf{R}^s \times \Theta.$$

ii) There exist functions $\rho_1 : \mathsf{R}^+ \rightarrow \mathsf{R}^+$ and $h_3 : \mathsf{R}^s \rightarrow \mathsf{R}^+$ such that $\rho_1\,(u) \rightarrow 0$ as $u \rightarrow 0$, h_3 is measurable, and for each (z, θ_1, θ_2) in $\mathsf{R}^s \times \Theta \times \Theta$

$$\|\psi(z,\theta_1) - \psi(z,\theta_2)\| \leq \rho_1\,(\|\theta_1 - \theta_2\|)\,h_3(z).$$

Assumption A.0.2 *$E\left[\psi(Z_t,\theta)\right] < \infty$ for each $\theta \in \Theta$, and $\Psi(\theta) = \lim_{t\rightarrow\infty} E\left[\psi(Z_t,\theta)\right]$. Moreover, $\Psi(\cdot)$ is continuous on Θ.*

Assumption A.0.3 *a) For each $\theta \in \Theta$, $\sum_{t=1}^T a_t\left(\psi(Z_t,\theta) - E\left[\psi(Z_t,\theta)\right]\right)$ converges almost surely.*

b) For $j = 1, 2, 3$, there exist bounded non-stochastic sequences $\left\{\eta_{jt}\right\}$ such that the sum $\sum_{t=1}^T a_t\left(h_j(Z_t) - \eta_{jt}\right)$ converges almost surely.

Assumption A.0.1 imposes some mild restrictions on the growth and smoothness properties of the measurement function ψ, while Assumption A.0.2 imposes a mild stationary requirement. As proved by Kuan and White, there exists a large class of data generating processes satisfying Assumption A.0.3, in particular the class of mixingales.

Proof of Proposition 5.4 Under the stated assumptions, there exists a convex neighborhood of $\delta^0 = (vec(\Sigma(\theta^0))', \theta^{0\prime})'$ on which $M^{RM}\,(\cdot)$ is continuously differentiable. Consequently, we can write

$$M^{RM}\,(\delta) = \frac{\partial M^{RM}}{\partial \delta'}\left(\delta^0\right)\left(\delta - \delta^0\right) + R\,(\delta)\,,$$

with $\left(\partial M^{RM}/\partial\delta'\right)\left(\delta^0\right)$ negative stable and $R\left(\delta\right)/\left\|\delta-\delta^0\right\|\to 0$ as $\left\|\delta-\delta^0\right\|\to 0$. Classical results from ODE theory show that this is sufficient to ensure the (local) asymptotic stability of δ^0 for the ODE corresponding to our recursive procedure, that is

$$\frac{\partial\delta}{\partial t}\left(t\right)=M^{RM}\left(\delta\left(t\right)\right),$$

(see, e.g., Rouche and Mahwin (1980), ch. 1). Now, we can apply Theorem II.2.1 (b) of Kuan and White (1994a). ■

A.7 Asymptotic normality for the recursive latent backfitting

Below, we present a set of general conditions ensuring the asymptotic normality of an almost sure convergent RM estimator defined as (.7) with $a_t=ct^{-1}$, $t=1,2,...$, $c>0$. See Theorem II.2.4 of Kuan and White (1994a) (KW hereafter) and their Assumptions B1 to B3 and B5. See also Assumptions (A1)-(A5) and Theorem 2 of Kushner and Huang (1979). The notation is that used in the previous appendix and the function $\psi:\mathsf{R}^s\times\Theta\to\mathsf{R}^p$ is still supposed measurable. Moreover, θ^0, the zero of $\Psi\left(\theta\right)$, is an interior point of Θ.

Assumption A.0.1 *$\{Z_t\}=\{Z_1,Z_2,...\}$ is stationary process and $\Psi\left(\theta\right)=E\left[\psi\left(Z_t,\theta\right)\right]$.*

Assumption A.0.2 *For each $z\in\mathsf{R}^s$, $\psi\left(z,\cdot\right)$ is continuously differentiable such there exists functions $\rho_2:\mathsf{R}^+\to\mathsf{R}^+$ and $h_4:\mathsf{R}^s\to\mathsf{R}^+$ such that $\rho_2\left(u\right)\to 0$ as $u\to 0$, h_4 is measurable, and for each θ in an open neighborhood in Θ of θ^0 and $z\in\mathsf{R}^s$,*

$$\left\|\frac{\partial\psi}{\partial\theta}(z,\theta)-\frac{\partial\psi}{\partial\theta}(z,\theta^0)\right\|\leq\rho_2\left(\left\|\theta-\theta^0\right\|\right)h_4(z).$$

Assumption A.0.3 $E\left[\left\|\psi\left(Z_t,\theta^0\right)\right\|^6\right]<\infty$ *and* $E\left[\left\|\frac{\partial\psi}{\partial\theta'}\left(Z_t,\theta^0\right)\right\|^2\right]<\infty$.

Assumption A.0.4 *i)* $E\left[h_4\left(Z_t\right)\right]<\infty$ *and*

$$\sum_{t=1}^{T}\frac{1}{t}\left(h_4(Z_t)-E\left[h_4\left(Z_t\right)\right]\right),$$

converges almost surely.

ii) Let $H^*=E\left[\frac{\partial\psi}{\partial\theta'}\left(Z_t,\theta^0\right)\right]$ and $h^*=E\left[\left\|\frac{\partial\psi}{\partial\theta'}\left(Z_t,\theta^0\right)\right\|\right]$. Then

$$\sum_{t=1}^{T}\frac{1}{t}\left[\frac{\partial\psi}{\partial\theta'}\left(Z_t,\theta^0\right)-H^*\right],\qquad\sum_{t=1}^{T}\frac{1}{t}\left[\left\|\frac{\partial\psi}{\partial\theta'}\left(Z_t,\theta^0\right)\right\|-h^*\right]$$

converge almost surely.

Assumption A.0.5 *i) If* $\kappa_t = E\left[\left\|E\left[\psi\left(Z_t,\theta^0\right) \mid Z_1\right]\right\|^2\right]^{1/2}$*,* $t \geq 1$*, then* $\sum_{t=1}^{\infty} \kappa_t^{1/2} < \infty$*.*

ii) $\sum_{t=1}^{\infty} \xi_t^{1/2} < \infty$*, where*

$$\xi_t^2 = \sup_{j\geq 0} E\left[\left\|E\left(\psi\left(Z_t,\theta^0\right)\psi\left(Z_{t+j},\theta^0\right)' \mid Z_1\right) - \sigma_j\right\|^2\right]$$

and $\sigma_j = E\left(\psi\left(Z_t,\theta^0\right)\psi\left(Z_{t+j},\theta^0\right)'\right)$*.*

KW and Kushner and Huang (1979) provide detailed remarks on these conditions. It is shown that such kind of assumptions are far less restrictive than it may seem at the first sight. Note that the previous assumptions imply, in particular, that $\partial\Psi/\partial\theta(\theta) = E\left[\partial\psi/\partial\theta\left(Z_t,\theta\right)\right]$ and thus $H^* = \partial\Psi/\partial\theta(\theta^0)$.

Proof of Proposition 5.5 For the application of the general result of KW to our recursive procedure, the same identifications as in the previous appendix have to be done ($\{Z_t\} = \{Y_t, X_t\}$, the parameters are $\left(vec(G)', \theta^{1\prime}\right)'$, ...).

By simple algebra we obtain

$$\bar{H} = c\frac{\partial M^{RM}}{\partial\delta'}(\delta^0) + \frac{1}{2}I_{p^2+p} = \begin{pmatrix} \left(-c+\frac{1}{2}\right)I_{p^2} & c\,\partial vec(\Sigma(\theta^0))/\partial\theta^{1\prime} \\ 0 & F(\theta^0,c) \end{pmatrix},$$

which is negative stable under the stated assumptions. The lower-right $p\times p$ bloc of the matrix R defined in (5.15) is

$$\begin{aligned} R &= \sum_{t\geq 0} E\left[M_{21}(\delta^0)M_{2t+1}(\delta^0)'\right] + \sum_{t\geq 1} E\left[M_{2t+1}(\delta^0)M_{21}(\delta^0)'\right] \\ &= \sum_{t\geq 0} E\left[\,(\Sigma(\theta^0)^{-1}M_1(\theta^0)\,)\,(\,\Sigma(\theta^0)^{-1}M_{t+1}(\theta^0)\,)'\,\right] \\ &+ \sum_{t\geq 1} E\left[\,(\Sigma(\theta^0)^{-1}M_{t+1}(\theta^0)\,)\,(\,\Sigma(\theta^0)^{-1}M_1(\theta^0)\,)'\,\right] \\ &= \Sigma(\theta^0)^{-1}\left(\sum_{t\geq 0} E\left[M_1(\theta^0)M_{t+1}(\theta^0)'\right] + \sum_{t\geq 1} E\left[M_{t+1}(\theta^0)M_1(\theta^0)'\right]\right)\Sigma(\theta^0)^{-1} \\ &= \Sigma(\theta^0)^{-1}B(\theta^0)\Sigma(\theta^0)^{-1}, \end{aligned}$$

with $M_t(\cdot)$, $t\geq 1$, defined in (5.5). Due to the block triangularity of $\bar{H}$, the lower right $p\times p$ bloc of the corresponding matrix (5.16) is exactly $V^{RM}(\theta^0)$. It remains to invoke Theorem II.2.4 of KW. ■

References

[1] **Aït-Sahalia, Y. (2002)**, "Maximum-Likelihood Estimation of Discretely-Sampled Diffusions: A Closed-Form Approach", *Econometrica*, 70, 223-262.

[2] **Aït-Sahalia, Y. and R. Kimmel (2002)**, "Estimating Affine Multifactor Term Structure Models Using Closed-Form Likelihood Expansions", Working Paper, Princeton University.

[3] **Aït-Sahalia, Y., and A.W. Lo (2000)**, "Nonparametric Risk Management and Implied Risk Aversion", *Journal of Econometrics*, 94, 9-51

[4] **Andrews, D.W.K. (1994a)**, "Asymptotics for semiparametric econometric models via stochastic equicontinuity", *Econometrica*, 62, 43-72.

[5] **Andrews, D.W.K. (1994b)**, "Empirical Process Methods in Econometrics", *Handbook of Econometrics, Vol IV*, R.F. Engle and D. McFadden eds., 2247 - 2294.

[6] **Bates, D. (2000),** "Post-'87 Crash Fears in the S&P 500 Futures Option", *Journal of Econometrics*, 94, 181-238.

[7] **Black, F. and M. Scholes (1973),** "The Pricing of Options and Corporate Liabilities", *Journal of Political Economy*, 81, 637-659.

[8] **Chen, R. and L. Scott (1993),** "ML Estimation for a Multifactor Equilibrium Model of the Term Structure of Interest Rates", *Journal of Fixed Income*, 3, 14-31.

[9] **Chen, X. and H. White (1998)**, "Nonparametric Adaptive Learning with Feedback", *Econometric Theory*, 82, 190-222.

[10] **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.

[11] **Chernov, M. and E. Ghysels (2000)**, "A Study towards a Unified Approach to the Joint Estimation of Objective and Risk Neutral Measures for the Purpose of Options Valuation", *Journal of Financial Economics*, 56, 407-458.

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

[13] **Cochrane, J. (2001),** *Asset Pricing*, Princeton University Press.

[14] **Collin-Dufresne, P., R. S. Goldstein and C.S. Jones (2002)**, "Identification and Estimation of 'Maximal' Affine Term Strucuture Models: An Application to Stochastic Volatility", Working Paper, University of Rochester.

[15] **Constantinides, G. (1992)**, "A Theory of the Nominal Term Structure of Interest Rates", *Review of Financial Studies*, 5, 531-552.

[16] **Cox, J., J. Ingersoll and S. Ross (1985)**, "A Theory of the Term Structure of Interest Rates", *Econometrica*, 53, 385-407.

[17] **Dai, Q. and K.J. Singleton (2000),** "Specification Analysis of Affine Term Structure Models", *Journal of Finance*, 55, 1943-1978.

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

[19] **Davidson, R. and J. G. MacKinnon (1993)**, *Estimation and inference in econometrics*, Oxford University Press.

[20] **De Jong, F. (2000),** "Time Series and Cross-section Information in Affine Term-Structure Models", *Journal of Business and Economic Statistics*, 18, 300-314.

[21] **Dempster, A., Laird, N. and D. Rubin (1977)**, "Maximum Likelihood from Incomplete Data via the EM Algorithm (with discussion)", *Journal of the Royal. Statistical Society, B*, 39, 1-38.

[22] **Dominitz, J. and R.P. Sherman (2001)**, "Convergence theory for stochastic iterative procedures with an application to semiparametric estimation", Working Paper, California Institute of Technology.

[23] **Dridi R. and E. Renault (2001),** "Semi-Parametric Indirect Inference", Working Paper, Université de Montréal.

[24] **Duan, J.C. (1994)**, "Maximum Likelihood Estimation using Price Data of the Derivative Contract", *Mathematical Finance*, 4, 155-167.

[25] **Duan, J. and J. Simonato (1999),** "Estimating and Testing Exponential-Affine Term Structure Models by Kalman Filter", *Review of Quantitative Finance and Accounting*, 13, 111-135.

[26] **Duffee, G.R. (2002)**, "Term Premia and Interest Rate Forecasts in Affine Models", *Journal of Finance*, 57, 405-443.

[27] **Duffie, D. and R. Kan (1996),** "A Yield-factor Model of Interest Rates", *Mathematical Finance*, 6, 379-406.

[28] **Dumas, B., J. Fleming and R. E. Whaley (1998)**, "Implied Volatility Functions: Empirical Tests", *Journal of Finance*, 53, 2059-2106.

[29] **Fisher, M. and C. Gilles (1996)**, "Estimating Exponential-Affine Models of the Term Structure", Working Paper, Board of Governors, Federal Reserve System.

[30] **Florens, J.P., C. Protopopescu and J.F. Richard (2001),** "Identification and Estimation of a Class of Game Theory Models", Working Paper, GREMAQ, Université de Toulouse.

[31] **Frisch, W.R. and F.V. Waugh (1933)**, "Partial time regression as compared with individual trends", *Econometrica*, 1, 387-401.

[32] **Gallant, R. A. and G. Tauchen (1996)**, "Which moments to Match", *Econometric Theory*, 12, 657-681.

[33] **Garcia, R., Luger, R. and Renault, E. (2002)**, "Empirical Assessment of an Intertemporal Option Pricing Model with Latent Variables", forthcoming in *Journal of Econometrics*

[34] **Gouriéroux, C., Monfort, A. and E. Renault (1993)**, "Indirect inference", *Journal of Applied Econometrics*, 8, S85-S118.

[35] **Gouriéroux, C., A. Monfort, E. Renault and A. Trognon (1987)**, "Generalized Residuals", *Journal of Econometrics* 34, 5-32.

[36] **Gouriéroux, C., Monfort A. and A. Trognon (1985)**, "A general Approach to Serial Correlation", *Econometric Theory* 1, 315-340.

[37] **Green, P., Jennison, C. and A. Seheult (1985)**, "Analysis of field experiments by least squares smoothing", *Journal of the Royal Statistical Society, B*, 47, 299-315.

[38] **Hansen L. and K. Singleton (1982),** "Generalized Instrumental Variables Estimation of Nonlinear Rational Expectations Models", *Econometrica*, 50, 1269-1288.

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

[40] **Hastie, T. J. and R. J. Tibshirani (1990)**, *Generalized Additive Models*, Chapman and Hall, London.

[41] **Heckman, J.J. and B. Honoré (1990),** "The Empirical Content of the Roy Model", *Econometrica*, 58, 1121-1149.

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

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

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

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

[46] **Jackwerth, J. (2000)**, "Recovering Risk Aversion from Option Prices and Realized Returns", *Review of Financial Studies*, 13, 433-451.

[47] **Johannes, M. and N. Polson (2001)**, "MCMC Methods for Financial Econometrics", prepared for the Handbook of Financial Econometrics, North-Holland, Y. Aït-Sahalia and L.P. Hansen Eds.

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

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

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

[51] **Kushner, H.J. and H. Huang (1979)**, "Rates of convergence for stochastic approximation type algorithms", *SIAM Journal of Control and Optimization*, 17, 607-617.

[52] **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, 528-541.

[53] **Kushner, H.J. and G.G. Yin (1997)**, *Stochastic Approximation Algorithms and Applications*, Springer-Verlag, New York.

[54] **Kwon, H.-C. (1992)**, *The Time Varying Term Premium of Interest Rates*, Ohio State University, Economics Department , Ph.D. dissertation.

[55] **Lintner, J. (1965a),** "Security Prices, Risk and Maximal Gains from Diversification", *Journal of Finance*, 20, 587-615.

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

[57] **McCulloch, J.H. (1975)**, "The Tax-Adjusted Yield Curve", *Journal of Finance*, 30, 811-830.

[58] **McCulloch, J.H. (1990)**, "U.S. Term Structure Data, 1946-87", *Handbook of Monetary Economics, North-Holland, Vol. I,* B. M. Friedman and F. H. Hahn eds., 672-715.

[59] **Merton, R. (1973),** "Rational Theory of Option Pricing", *Bell Journal of Economics and Management Science*, 4, 141-183.

[60] **Muirhead, R.J. (1982),** "Aspects of Multivariate Statistical Theory", John Wiley & Sons.

[61] **Nielsen, S.F. (2000)**, "On Simulated EM Algorithms", *Journal of Econometrics*, 96, 267-292.

[62] **Newey, W. and D. McFadden (1994)**, "Large sample estimation and hypothesis testing", *Handbook of Econometrics, Vol IV*, R.F. Engle and D. McFadden eds., 2111 - 2245.

[63] **Pakes, A. and D. Pollard (1989)**, "Simulation and the Asymptotics of the Optimization Estimators", *Econometrica*, 57, 1027-1057.

[64] **Pan, J. (2002),** "The Jump-Risk Premia Implicit in Options: Evidence from an Integrated Time-Series Study", *Journal of Financial Economics*, 63, 3-50.

[65] **Pastorello, S. Renault, E. and Touzi, N. (2000)**, "Statistical Inference for Random-Variance Option Pricing", *Journal of Business and Economic Statistics*, 18, 358-367.

[66] **Patilea, V. and E. Renault (1997)**, "Continuously updated estimators", CORE DP 9776.

[67] **Pearson, N. and T.S. Sun (1994),** "Exploiting the Conditional Density in Estimating the Term Structure: An Application to the CIR Model", *Journal of Finance*, 46, 1279-1304.

[68] **Renault, E. (1997)**, "Econometric Models of Option Pricing Errors", *Advances in economics and econometrics: theory and applications, Seventh World Congress, vol III, chapter 8*, D. Kreps and K. Wallis (eds.), Cambridge University Press.

[69] **Renault, E and N. Touzi (1996)**, "Option Hedging and Implied Volatilities in a Stochastic Volatility Model", *Mathematical Finance*, 6, 279-302.

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

[71] **Robinson, P. (1982)**, "On the Asymptotic Properties of Estimators of Models Containing Limited Dependent Variables", *Econometrica,* 50, 1, 27-41.

[72] **Rosenberg, J.V. and R.F. Engle (2000)**, "Empirical Pricing Kernels", Working Paper, NYU Stern School of Business.

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

[74] **Ruud, P. (1991)**, "Extensions of Estimations using the EM Algorithm", *Journal of Econometrics*, 49, 305-341.

[75] **Sharpe, W. (1964),** "Capital Asset Prices: A Theory of Market Equilibrium under Conditions of Risk", *Journal of Finance*, 19, 425-442.

[76] **Speckman, P. (1988)**, "Kernel Smoothing in Partial Linear Models", *Journal of the Royal Statistical Society, B*, 50, 413-436.

[77] **Tanner, M.A. (1996),** "Tools for Statistical Inference, Methods for the Exploration of Posterior Distributions and Likelihood Functions", Springer Series in Statistics.

[78] **Tauchen, G. (1997)**, "New Minimum Chi-Square Methods in Empirical Finance", in *Advances in economics and econometrics: theory and applications, Seventh World Congress, vol III, chapter 7*, D. Kreps and K. Wallis (eds.), Cambridge University Press.

[79] **van de Geer, S. (2000)**, *Empirical Process in M-estimation*, Cambridge University Press.

[80] **Vasicek, O. (1977)**, "An Equilibrium Characterization of the Term Structure", *Journal of Financial Economics*, 5, 177-188.

[81] **White, H. (1989)**, "Some Asymptotic Results for Learning in Single Hidden-Layer Feedforward Network Models", *Journal of the American Statistical Association*, 84, 1003-1013.

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