Maximization by parts in extremum estimation

Maximization by Parts in Extremum Estimation*

Yanqin Fan[†] Sergio Pastorello[‡] and Eric Renault[§]

First version: March 2006
This version: January 2015

Abstract

In this paper, we present various iterative algorithms for extremum estimation in cases where direct computation of the extremum estimator or via the Newton-Raphson algorithm is difficult, if not impossible. While the Newton-Raphson algorithm makes use of the full Hessian matrix which may be difficult to evaluate, our algorithms use parts of the Hessian matrix only, the parts that are easier to compute. We establish consistency and asymptotic efficiency of our iterative estimators under regularity and information dominance conditions. We argue that the economic interpretation of a structural econometric model will often allow us to give credibility to a well-suited information dominance condition. We apply our algorithms to the estimation of Merton's structural credit risk model and to Heston's stochastic volatility option pricing model.

JEL codes: C13, C58, G13

Keywords: Extremum Estimators, Maximum Likelihood, Generalized Method of Moments, Structural Econometrics, Iterative Backfitting Methods, Implied States GMM

*Previous versions of this paper have been presented at Free University at Amsterdam in 2006, the 2007 ESEM in Budapest, 2009 SoFiE Annual Conference in Geneva, Conference in tribute of Russell Davidson, Marseille 2009, Conference in tribute of Adrian Pagan, Sydney 2009, and the Department of Statistics at the University of Chicago in 2009. We are grateful to participants of these conferences and seminars for helpful discussions and comments that have led to a much improved paper. Fan acknowledges financial support from the National Science Foundation. Part of the work in this paper was done when Fan visited SAMSI whose financial support and hospitality are gratefully acknowledged. Pastorello acknowledges financial support from the PRIN 2005 program.

[†]Department of Economics, Vanderbilt University, VU Station B #351819, 2301 Vanderbilt Place, Nashville TN 37235-1819, USA.

[‡]Dipartimento di Scienze Economiche, Università di Bologna, Piazza Scaravilli, 2, 40126 Bologna, Italy.

[§]Department of Economics, Robinson Hall, Brown University, Providence, RI 02912, USA.

1 Introduction

Most econometric/statistical estimators can be defined as extremum estimators obtained from optimizing a sample objective function. They include maximum likelihood estimators, generalized method of moments estimators, empirical likelihood and minimum distance estimators. Under certain conditions including smoothness of the sample objective function and interior solution, the extremum estimator is a solution to the first order condition (FOC). In many interesting cases, solving the FOC condition directly with respect to all the occurrences of the parameters of interest may be numerically cumbersome or impractical. This paper provides a unified theory for efficient estimation algorithms that avoid solving the FOC directly by iteratively solving much simpler problems.

Specifically consider a generic objective function $L_T(\theta)$, where θ is a p-dimensional vector of parameters. In the regular case, the parameter estimate $\hat{\theta}_T$ can be computed by solving the first order conditions: $\partial L_T(\theta)/\partial\theta = 0$. Since analytical solutions are generally not available, the estimates are usually computed using iterative methods such as Newton-Raphson, quasi-Newton methods or derivative-free methods, such as the simplex method or simulated annealing. It is well known that in a neighborhood of $\hat{\theta}_T$, the Newton-Raphson method converges quadratically, but at the cost of requiring the computation of the Hessian matrix of $L_T(\theta)$. When the objective function is very complicated, analytical expressions of the second derivatives are difficult or impossible to obtain. Numerical approximations based on finite differences are possible, but can be costly, especially when p is large and/or evaluating $L_T(\theta)$ is time-consuming. Moreover, these approximations can be very sensitive to the choice of the intervals used for differencing, and the results they provide are frequently unreliable. For these reasons, the use of approximated second derivatives can worsen significantly the performance of the Newton-Raphson algorithm, both in terms of convergence speed and of computational cost. Alternatively, quasi-Newton algorithms (e.g., BFGS or DFP) can be used. These methods exploit only the information in the gradient of $L_T(\theta)$ and do not require second derivatives, but their performance can be significantly worse than Newton-Raphson's in terms of speed, and tends to depend on problem scaling and algorithm parameter selection. Finally, in moderate-to large-scale problems derivative-free methods usually make the search for the optimum very slow and inefficient.

In this paper we introduce a class of alternative iterative optimization algorithms that may be considered as intermediate between Newton-Raphson and quasi-Newton methods. In a nutshell, our algorithms do not require to evaluate the full Hessian matrix of $L_T(\theta)$, but only

the portion of it associated to the "simple" occurrences of θ in the objective function. This should allow to exploit at least in part the information about the curvature of the objective criterion, thus providing a computationally simple and yet effective method to compute the efficient estimator $\hat{\theta}_T$. The precise meaning of "simple" occurrence likely depends on the problem under scrutiny. In this paper, we assume that the objective function $L_T(\theta)$ can be written as $Q_T[\theta, \nu(\theta)]$, where $\nu(\cdot)$ is a vector of functions that collect all the problematic occurrences of θ – i.e., difficult or heavy to evaluate and/or hard to differentiate. Pastorello, Patilea and Renault (2003) (PPR hereafter) notice that such a structure naturally arises in many examples in Economics and Finance. In this framework, computing the efficient estimator $\hat{\theta}_T$ may be difficult because of the presence of $\nu(\theta)$; for this reason PPR develop an alternative iterative estimator of the parameters that does not require to maximize w.r.t. the second occurrence of θ in $Q_T[\theta, \nu(\theta)]$. If some conditions are satisfied, the PPR estimator is consistent and much easier to compute than $\hat{\theta}_T$, but generally inefficient w.r.t. to it because it is based only on the "simple" occurrence of θ in the objective function, thus neglecting the information contained in the second occurrence.

In this paper we present simple modifications of the PPR iterations which produce iterative algorithms that, upon convergence, yield the efficient estimator $\hat{\theta}_T$. We do this by extending to the general extremum estimation problem the idea of Maximization-By-Parts (MBP) put forth by Song, Fan and Kalbfleisch (2005) (SFK hereafter) in the context of Maximum Likelihood with a convenient separability property of the log-likelihood function. It must be stressed that in terms of computational cost each iteration of our new algorithms is not more complicated than a PPR iteration. The modification we consider is not unique; in the text we will focus on the one closest to the PPR algorithm, but an alternative is illustrated in an Appendix. This flexibility of our algorithms allows to tailor our approach to the specific details of a large variety of estimation criteria. Moreover, since the convergence to $\hat{\theta}_T$ of the two algorithms imposes different requirements on Q_T, one is free to switch between them if iterations fail to converge in reasonable time. We present conditions under which our algorithms converge and establish asymptotic properties of the corresponding estimators. We discuss the modifications that arise in these properties when the algorithm's iterations are started at a consistent estimator, and we also devote special attention to the case of a GMM objective function.

The rest of this paper is organized as follows. In section 2, we present the main idea underlying our new algorithms and highlight their connections with the backfitting estimator of PPR and the MBP algorithm of SFK. Section 3 illustrates one example, and section 4

discusses asymptotic results, including conditions that guarantee the convergence of the algorithms for both consistent and inconsistent initial values. In section 5 we discuss the implementation of the new algorithms for a GMM objective function and present a new algorithm specific to it. A numerical example in the GMM context is studied in Section 6. The last section concludes. Technical proofs and additional details are relegated to two appendices.

2 The Framework and a New Efficient Algorithm

Consider the objective function $L_T(\theta) = Q_T[\theta, \nu(\theta)]$ and the corresponding extremum estimator defined by

$$\hat{\theta}_T = \arg\max_{\theta \in \Theta} Q_T[\theta, \nu(\theta)].$$

In this paper we focus on the case that $\hat{\theta}_T$ is the only solution to the first order condition:

$$\frac{\partial L_T(\theta)}{\partial \theta} = \frac{\partial Q_T[\theta, \nu(\theta)]}{\partial \theta} + \frac{\partial \nu(\theta)'}{\partial \theta} \frac{\partial Q_T[\theta, \nu(\theta)]}{\partial \nu} = 0. \tag{1}$$

We assume that solving directly (1) w.r.t. θ is difficult. Typically this occurs because the first order condition depends on θ in several places and some terms, collected in $\nu(\theta)$, are much more complicated than others. PPR and SFK develop iterative algorithms that are easier to implement than solving (1) directly. However, the algorithms in PPR may not produce efficient estimators upon convergence and those in SFK only apply to the separable likelihood framework motivating the current paper.

In the rest of this section, we first present a brief review of PPR and SFK and then introduce a new efficient iterative algorithm.

2.1 Existing Iterative Estimators in PPR and SFK

To avoid maximization w.r.t. the second occurrence of θ in the objective function $Q_T[\theta, \nu(\theta)]$, PPR observed that if the value $\nu(\theta^0)$ was known where θ^0 denotes the true parameter value, then an alternative and simpler estimator could be computed by solving $\max_{\theta \in \Theta} Q_T[\theta, \nu(\theta^0)]$. Of course this is not directly feasible, but an estimator can nevertheless be computed as the limit of an iterative backfitting procedure, $\hat{\theta}_T^{PPR} = \lim_{k \to \infty} \hat{\theta}_T^{(k)}$, where $\hat{\theta}_T^{(k)}$ is implicitly defined by:

$$\frac{\partial Q_T[\hat{\theta}_T^{(k)}, \nu(\hat{\theta}_T^{(k-1)})]}{\partial \theta} = 0.$$

To compare $\hat{\theta}_T$ and $\hat{\theta}_T^{PPR}$, let us introduce the limit objective function $L_\infty(\theta) = Q_\infty[\theta, \nu(\theta)] = \text{plim}_{T\to\infty} Q_T[\theta, \nu(\theta)]$. The consistency of $\hat{\theta}_T$ and $\hat{\theta}_T^{PPR}$ requires that the true value θ^0 is the only solution of $\max_{\theta\in\Theta} Q_\infty[\theta, \nu(\theta)]$ and $\max_{\theta\in\Theta} Q_\infty[\theta, \nu(\theta^0)]$, respectively. Under these assumptions, the asymptotic variances of the two estimators will be determined by sample counterparts of the first order conditions:

$$\frac{\partial Q_\infty[\theta^0, \nu(\theta^0)]}{\partial \theta} + \frac{\partial \nu(\theta^0)'}{\partial \theta} \frac{\partial Q_\infty[\theta^0, \nu(\theta^0)]}{\partial \nu} = 0 \text{ for } \hat{\theta}_T \text{ and} \tag{2}$$

$$\frac{\partial Q_\infty[\theta^0, \nu(\theta^0)]}{\partial \theta} = 0 \text{ for } \hat{\theta}_T^{PPR}. \tag{3}$$

Taken jointly, (2) and (3) imply that

$$\frac{\partial \nu(\theta^0)'}{\partial \theta} \frac{\partial Q_\infty[\theta^0, \nu(\theta^0)]}{\partial \nu} = 0. \tag{4}$$

Given that by assumption, (3) identifies θ^0, we can think of (4) as an additional moment condition, and look at (2) as the optimal way to combine these two sets of moment conditions. In general, given that it uses (3) only, $\hat{\theta}_T^{PPR}$ is asymptotically less efficient than $\hat{\theta}_T$, except if by chance (2) and (3) are equivalent. This case occurs when the function $\nu(\theta)$ is (at least asymptotically) the result of a preliminary concentration stage.[1]

SFK consider the special case in which the objective function is a loglikelihood function taking the additive form: $Q_T[\theta, \nu(\theta)] = Q_{1,T}(\theta_1) + Q_{2,T}[\nu(\theta_1), \theta_2]$, where $\theta = (\theta_1', \theta_2')'$. In this case, $\hat{\theta}_T$ is the MLE. SFK provide several examples for which the log-likelihood function is of this form and the full MLE may be difficult to compute directly, as $Q_{2,T}[\nu(\theta_1), \theta_2]$ depends on θ_1 in a complicated way. For separable estimation criteria, the PPR estimator can be computed as the solution to

$$\frac{\partial Q_{1,T}(\hat{\theta}_{1,T}^{PPR})}{\partial \theta_1} = 0 \quad \text{and} \quad \frac{\partial Q_{2,T}[\nu(\hat{\theta}_{1,T}^{PPR}), \hat{\theta}_{2,T}^{PPR}]}{\partial \theta_2} = 0,$$

where in the first step, $\hat{\theta}_{1,T}^{PPR}$ is computed and then fixing $\hat{\theta}_{1,T}^{PPR}$, $\hat{\theta}_{2,T}^{PPR}$ is computed in the second step. This kind of two-step estimation procedures are extremely popular in Economics and Finance because of their simplicity, but they are in general inefficient w.r.t. $\hat{\theta}_T$.

To avoid the efficiency loss associated with $\hat{\theta}_T^{PPR}$, SFK proposed MBP, an iterative algorithm which produces an estimator asymptotically equivalent to the MLE upon convergence.

[1] When $\nu(\theta)$ derives from a concentration step, $\frac{\partial Q_\infty[\theta, \nu(\theta)]}{\partial \nu} = 0$, for all θ. Also, note that even in this case the assumption that $\nu(\cdot)$ is known is not really restrictive since any sample dependence could always be absorbed in $Q_T(\cdot)$.

The generic iteration of MBP is based on the structure of the first order condition for $\hat{\theta}_T$, and is implicitly defined as:

$$\frac{\partial Q_{1,T}(\hat{\theta}_1^{(k)})}{\partial \theta_1} = -\frac{\partial Q_{2,T}[\nu(\hat{\theta}_1^{(k-1)}), \hat{\theta}_2^{(k-1)}]}{\partial \theta_1}, \tag{5}$$

$$\frac{\partial Q_{2,T}[\nu(\hat{\theta}_1^{(k)}), \hat{\theta}_2^{(k)}]}{\partial \theta_2} = 0. \tag{6}$$

The key observation is that, for each single iteration, running MBP is not more involved than running the PPR two-step estimator. The only difference is that the expression on the right hand side in (5) is no longer stuck to zero. To start the algorithm, we can use the two-step estimator above as an initial value for θ. SFK showed that, under an information dominance condition, iterating from $\hat{\theta}_T^{PPR}$ via (5)-(6) yields an estimator asymptotically equivalent to the MLE upon convergence.

Computationally, each step in MBP is no more difficult than maximizing the first term $Q_{1,T}(\theta_1)$ (as well as the second term for given θ_1) and hence is well suited to examples in which $Q_{1,T}(\theta_1)$ is of a much simpler form than the second term $Q_{2,T}[\nu(\theta_1), \theta_2]$ (as a function of θ_1). In the next subsection we will show that a natural extension of the MBP idea to the case of general, non-separable estimation criteria of the form $Q_T[\theta, \nu(\theta)]$ allows to overcome the inefficiency of $\hat{\theta}_T^{PPR}$ and provides simple iterative algorithms approaching the efficient estimator $\hat{\theta}_T$ upon convergence.

2.2 An Efficient Iterative Backfitting Algorithm

The main contribution of this paper is to provide some simple iterative algorithms that upon convergence yield the efficient extremum estimator, solution of (1) at the same computational cost as the inefficient PPR algorithm reviewed in the previous section. Instead of trying to solve the complete FOC (1), our algorithms iterate on the more complicated terms. In what follows, we assume that iterations are started at $\hat{\theta}_T^{(0)}$, which may or may not be a consistent estimator of θ. Consider the following iterative scheme.

Algorithm I. Given $\hat{\theta}_T^{(k-1)}$, let $\hat{\theta}_T^{(k)}$ be the solution of:

$$\frac{\partial Q_T[\theta, \nu(\hat{\theta}_T^{(k-1)})]}{\partial \theta} = -\frac{\partial \nu(\hat{\theta}_T^{(k-1)})'}{\partial \theta} \frac{\partial Q_T[\hat{\theta}_T^{(k-1)}, \nu(\hat{\theta}_T^{(k-1)})]}{\partial \nu}, \quad k = 1, 2, \ldots. \tag{7}$$

Let $\overline{\theta}_T(\cdot)$ denote a function such that the solution of (7) obeys

$$\hat{\theta}_T^{(k)} = \overline{\theta}_T(\hat{\theta}_T^{(k-1)}), \quad k = 1, 2, \ldots \tag{8}$$

Following Dominitz and Sherman (2005), if $\overline{\theta}_T(\cdot)$ is an asymptotic contraction mapping, then there exists a fixed point that necessarily coincides with the efficient estimator $\hat{\theta}_T$, and the sequence defined by (8) converges to it as $k \to \infty$. With a scalar θ, it can be checked that $\overline{\theta}_T(\cdot)$ is an asymptotic contraction mapping if the absolute value of its derivative is asymptotically bounded between 0 and 1 with probability one. If $\hat{\theta}_T^{(0)}$ is a consistent estimator, the contraction mapping condition is a local one, in that it must hold only in θ^0; otherwise, it must hold globally, i.e. over the whole parameter space Θ.

An interesting variant of algorithm I is defined by the following updating rule.

Algorithm I (Newton version). Given $\hat{\theta}_T^{(k-1)}$, let $\hat{\theta}_T^{(k)}$ be given by:

$$\hat{\theta}_T^{(k)} = \hat{\theta}_T^{(k-1)} - \left[\frac{\partial^2 Q_T[\hat{\theta}_T^{(k-1)}, \nu(\hat{\theta}_T^{(k-1)})]}{\partial\theta\partial\theta'}\right]^{-1} \left[\frac{\partial L_T(\hat{\theta}_T^{(k-1)})}{\partial\theta}\right], \quad k = 1, 2, \ldots \tag{9}$$

We will refer to this iterative rule as the Newton version of algorithm I, because it corresponds to a single iteration of the Newton algorithm associated to the multivariate nonlinear equations (7). The proof of Theorem 4.2 below shows that, under the same contraction mapping condition required for (7), the sequence $\hat{\theta}_T^{(k)}$ defined by (9) converges as $k \to \infty$ to an estimator asymptotically equivalent (for large T) to $\hat{\theta}_T$.[2] Note, however, that implementing (9) is computationally much cheaper than (7).

It is also instructive to compare (9) with the Newton-Raphson iterates used to compute $\hat{\theta}_T$ given by:

$$\tilde{\theta}_T^{(k)} = \tilde{\theta}_T^{(k-1)} - \left[D^2 L_T(\tilde{\theta}_T^{(k-1)})\right]^{-1} \left[\frac{\partial L_T(\tilde{\theta}_T^{(k-1)})}{\partial\theta}\right], \quad k = 1, 2, \ldots \tag{10}$$

where

$$\begin{aligned} D^2 L_T(\theta) \;=\; & \frac{\partial^2 Q_T[\theta, \nu(\theta)]}{\partial\theta\partial\theta'} + \frac{\partial^2 Q_T[\theta, \nu(\theta)]}{\partial\theta\partial\nu'}\frac{\partial\nu(\theta)}{\partial\theta'} + \frac{\partial\nu(\theta)'}{\partial\theta}\frac{\partial^2 Q_T[\theta, \nu(\theta)]}{\partial\nu\partial\theta'} \\ & + \frac{\partial\nu(\theta)'}{\partial\theta}\frac{\partial^2 Q_T[\theta, \nu(\theta)]}{\partial\nu\partial\nu'}\frac{\partial\nu(\theta)}{\partial\theta'} + \sum_{j=1}^{\dim(\nu)} \frac{\partial Q_T[\theta, \nu(\theta)]}{\partial\nu_j}\frac{\partial^2\nu_j(\theta)}{\partial\theta\partial\theta'}. \end{aligned} \tag{11}$$

Under standard regularity conditions, it is known that starting from a consistent estimator, one-step iteration of (10) yields asymptotically efficient estimator. Robinson (1988) establishes higher order properties of $\tilde{\theta}_T^{(k)}$ for fixed k in terms of its stochastic difference with the efficient estimator $\hat{\theta}_T$. For GMM criterion function and criterion functions that can be

[2] Even if they share the same limit, the iterates originated by (7) and (9) will generally be different.

written as sample averages, Andrews (2002) studies higher order efficiency properties of $\tilde{\theta}_T^{(k)}$ for fixed k.

Instead of using the full Hessian as (10), the Newton version of algorithm I exploits the first term on the right hand side of (11) only. This explains why, when started from a consistent estimator, (9) will not deliver an asymptotically efficient estimator in a single step; its advantage, however, lies in its computational ease, since only second derivatives w.r.t. the first occurrence of θ are required.

In section 4 we detail the contraction mapping condition needed when the number of parameters is larger than one, and we show that this condition can be alternatively represented in terms of information dominance (ID). The ID condition requires that the part of the Hessian that is used in (9) dominates the part that is ignored.

It should be noticed that even if the limit criterion $Q_\infty[\cdot, \nu(\cdot)]$ satisfies the contraction mapping condition, the finite sample criterion $Q_T[\cdot, \nu(\cdot)]$ may not. In this case the convergence of the iterations (7) or (9) is not guaranteed. We present in appendix B.1 an alternative algorithm based on different choices of the occurrences of θ in the full FOC (1) on which to iterate. Using its Newton variant, this algorithm can be interpreted as exploiting an additional component of the full Hessian in (11). Upon convergence, it provides an asymptotically efficient estimator but the required contraction mapping condition differs from algorithm I. Depending on the specific application, one of the algorithms may be computationally more convenient or exhibits superior small sample performance in terms of convergence rate.

The lack of small sample convergence of a given algorithm can alternatively be addressed by letting the number of iterations k grow to infinity with T at a sufficiently fast rate. Let us denote by $k(T)$ the number of iterations to highlight the dependence of k on T, and by $\hat{\theta}_T^{(k(T))}$ the associated estimator. Proposition 3 in PPR implies that this estimator is consistent and asymptotically equivalent to $\hat{\theta}_T$ if, in addition to the contraction mapping condition,

$$\sqrt{T}\left(\hat{\theta}_T^{(k(T))} - \hat{\theta}_T^{(k(T)-1)}\right) \to 0 \quad \text{in probability.}$$

A sufficient condition for this property to hold is the "uniform contraction mapping assumption" put forward by Dominitz and Sherman (2005) with $k(T) = T^\delta$, for some $\delta > 0$.

3 An Example—The Merton Credit Risk Model

In this section we use the Merton Credit Risk Model to demonstrate the flexibility of our new algorithms and to establish their connections with well-known existing estimation approaches. In its simplest version, the Merton model assumes that the firm's market value is

a latent variable Y_t^* whose dynamics is described by a Geometric Brownian Motion:

$$\frac{dY_t^*}{Y_t^*} = \mu dt + \sigma dW_t,$$

where (W_t) is a standard Brownian motion. Suppose that the firm's debt consists of a zero-coupon bond with face value K expiring in τ, and let Y_t denote the firm's equity price. Since at the debt's maturity $Y_\tau = \max[Y_\tau^* - K, 0]$, we can interpret the shares offered on the firm's value as European call options written on this value with strike K and maturity τ. Hence, the observed equity prices $Y_0, Y_1, Y_2, \ldots, Y_T$ can be seen as option prices written on latent values $Y_0^*, Y_1^*, Y_2^*, \ldots, Y_T^*$.

Let us denote with $g_t(Y_t, \sigma^2)$ the inverse of the Black and Scholes pricing formula:

$$Y_t^* = g_t(Y_t, \sigma^2) \quad \Leftrightarrow \quad Y_t = g_t^{-1}(Y_t^*, \sigma^2) = Y_t^* \Phi(d_t(\sigma^2)) - K\, e^{-r(\tau - t)} \Phi(d_t(\sigma^2) - \sigma\sqrt{\tau - t}), \tag{12}$$

where $\Phi(\cdot)$ is the cdf of the standard normal distribution, $d_t(\sigma^2) = [\log(Y_t^*/K) + (r + \frac{\sigma^2}{2})(\tau - t)]/(\sigma\sqrt{\tau - t})$, and r is the risk-free interest rate assumed to be deterministic and time-invariant. Notice that the link function $g_t(\cdot, \sigma^2)$ relating latent and observed variables depends on t through the residual time to maturity $(\tau - t)$ of the implicit option contract. It is also worth reminding that the Black-Scholes price is continuously differentiable with respect to the underlying stock price, and the derivative is the so-called delta of the option contract given by (see e.g. Hull (1999) p 312):

$$\frac{\partial Y_t}{\partial Y_t^*} = \Phi[d_t(\sigma^2)]. \tag{13}$$

Maximum likelihood inference about unknown parameters (μ, σ^2) would be straightforward if we had observed the time series Y_t^*, $t = 0, 1, \ldots, T$ of latent firm values. By Ito's lemma, we know that the latent log-returns $R_t^* = \log Y_t^* - \log Y_{t-1}^*$, $t = 1, \ldots, T$, are independent, identically normally distributed with mean $(\mu - \sigma^2/2)$ and variance σ^2 so that, conditional on the first observation, the likelihood function is simply:

$$l_T(Y_1^*, \ldots, Y_T^* | Y_0^*; \mu, \sigma^2) = \prod_{t=1}^{T} (2\pi\sigma^2)^{-1/2} \exp\left\{ -\frac{1}{2\sigma^2} \left(R_t^* - \mu + \frac{\sigma^2}{2} \right)^2 \right\} \prod_{t=1}^{T} \frac{1}{Y_t^*}. \tag{14}$$

Since we only observe $Y_t = g_t^{-1}(Y_t^*, \sigma^2)$, $t = 0, 1, ..., T$, the Jacobian formula gives us:

$$l_T(Y_1, \ldots, Y_T | Y_0; \mu, \sigma^2) = \tag{15}$$
$$\prod_{t=1}^{T} (2\pi\sigma^2)^{-1/2} \exp\left\{ -\frac{1}{2\sigma^2} \left(R_t(\sigma^2) - \mu + \frac{\sigma^2}{2} \right)^2 \right\} \left[\prod_{t=1}^{T} \Phi(d_t(\sigma^2)) \right]^{-1} \prod_{t=1}^{T} \frac{1}{g_t(Y_t, \sigma^2)},$$

where implicit returns $R_t(\sigma^2)$ that can be backed out from option prices for each value of σ^2 are given by:

$$R_t(\sigma^2) = \ln g_t(Y_t, \sigma^2) - \ln g_{t-1}(Y_{t-1}, \sigma^2).$$

Since μ is not a parameter of interest, it is simpler to concentrate w.r.t. μ. Let:

$$\bar{R}_T(\sigma^2) = \frac{1}{T}\sum_{t=1}^{T} R_t(\sigma^2)$$

and, by maximization of (15), $\mu_T(\sigma^2) = \bar{R}_T(\sigma^2) + \frac{\sigma^2}{2}$. The concentrated likelihood is then given by:

$$l_T^c(Y_1, \ldots, Y_T | Y_0; \sigma^2) = \tag{16}$$
$$\prod_{t=1}^{T}(2\pi\sigma^2)^{-1/2} \exp\left\{-\frac{1}{2\sigma^2}\left(R_t(\sigma^2) - \bar{R}_T(\sigma^2)\right)^2\right\}\left[\prod_{t=1}^{T}\Phi(d_t(\sigma^2))\right]^{-1}\prod_{t=1}^{T}\frac{1}{g_t(Y_t, \sigma^2)}.$$

Obviously, direct maximization of the concentrated likelihood function (16) with respect to σ^2 does not deliver a user-friendly estimator, neither for computation nor for interpretation. By contrast, the infeasible estimator of σ^2 obtained by maximization of the latent likelihood (14) would have coincided with the natural idea of historical volatility. To overcome this difficulty, let us dub problematic occurence of σ^2 all the occurences that show up in (16) and not in (14). These problematic occurences, due to Jacobian formula and/or inversion of Black-Scholes formula, will be singled out thanks to a specific notation $\nu(\sigma^2)$ (with $\nu(\sigma^2) = \sigma^2$). We end up with the objective function:

$$\begin{aligned} Q_T[\sigma^2, \nu(\sigma^2)] &= -\frac{1}{2}\log(2\pi\sigma^2) - \frac{1}{2T}\sum_{t=1}^{T}\frac{\left[R_t[\nu(\sigma^2)] - \overline{R}_T[\nu(\sigma^2)]\right]^2}{\sigma^2} \\ &\quad -\frac{1}{T}\sum_{t=1}^{T}\log g_t[Y_t, \nu(\sigma^2)] - \frac{1}{T}\sum_{t=1}^{T}\log\Phi\left[d_t(\nu(\sigma^2))\right]. \end{aligned} \tag{17}$$

It is then worth comparing in this context the k-th PPR iteration with the k-th iteration of our efficient algorithm I. Starting from an estimator at the $(k-1)$-th iteration stage denoted as $\hat{\sigma}_T^{2(k-1)}$, the PPR iteration will update it by picking as value for the k-th iteration the solution σ^2 of:

$$\frac{\partial Q_T[\sigma^2, \nu(\hat{\sigma}_T^{2(k-1)})]}{\partial\sigma^2} = 0,$$

that is:

$$-\frac{1}{2\sigma^2} + \frac{1}{2\sigma^4}\tilde{\sigma}_T^2[\hat{\sigma}_T^{2(k-1)}] = 0, \tag{18}$$

where $\tilde{\sigma}_T^2[\sigma^2]$ stands for the infeasible historical (squared) volatility estimator that one would deduce from the prior knowledge of σ^2:

$$\tilde{\sigma}_T^2[\sigma^2] = \frac{1}{T}\sum_{t=1}^{T}\left[R_t[\sigma^2] - \overline{R}_T[\sigma^2]\right]^2.$$

By contrast, the efficient algorithm I will update the $(k-1)$-th iteration stage, still denoted as $\hat{\sigma}_T^{2(k-1)}$, by picking as value for the k-th iteration the solution σ^2 of:

$$\frac{\partial Q_T[\sigma^2, \nu(\hat{\sigma}_T^{2(k-1)})]}{\partial \sigma^2} + \frac{\partial Q_T[\hat{\sigma}_T^{2(k-1)}, \nu(\hat{\sigma}_T^{2(k-1)})]}{\partial \nu} = 0,$$

that is:

$$-\frac{1}{2\sigma^2} + \frac{1}{2\sigma^4}\tilde{\sigma}_T^2[\hat{\sigma}_T^{2(k-1)}] + \frac{\partial Q_T[\hat{\sigma}_T^{2(k-1)}, \nu(\hat{\sigma}_T^{2(k-1)})]}{\partial \nu} = 0.$$

It is convenient to rewrite the latter equation as a second degree equation w.r.t. σ^2:

$$A_T(\hat{\sigma}_T^{2(k-1)})\,\sigma^4 - \sigma^2 + \tilde{\sigma}_T^2[\hat{\sigma}_T^{2(k-1)}] = 0, \quad \text{with } A_T(\sigma^2) = 2\frac{\partial Q_T[\sigma^2, \nu(\sigma^2)]}{\partial \nu}. \tag{19}$$

Obviously (see solution of (18)) the k-th PPR iteration is given by

$$\hat{\sigma}_T^{2(k)} = \tilde{\sigma}_T^2[\hat{\sigma}_T^{2(k-1)}] = \frac{1}{T}\sum_{t=1}^{T}\left[R_t[\hat{\sigma}_T^{2(k-1)}] - \overline{R}_T[\hat{\sigma}_T^{2(k-1)}]\right]^2. \tag{20}$$

This expression implies that the PPR algorithm coincides with the popular KMV iterative technique developed by Kealhofer, Mcquown and Vasicek (see Crouhy, Galai and Mark, 2000 for more details on this method). It can actually be seen as a version of the EM algorithm (see Duan, Gauthier and Simonato, 2005) since it amounts to maximizing the latent likelihood (14) for the predicted values $g_t(Y_t, \hat{\sigma}_T^{2(k-1)})$ of the latent variables Y_t^*. It is unfortunately a case where the EM algorithm does not deliver an estimate asymptotically equivalent to MLE (see PPR and references therein). This is the reason why the efficient algorithm I involves a correction of the naive updating rule (20) by solving instead the second degree equation (19). Note however that the PPR/KMV estimator is consistent because the correction term $A_T(\sigma^2)$ vanishes for large T. Therefore, the discriminant of (19) is positive for large T. The intuition of the efficiency gain when moving from PPR/KMV to algorithm I is as follows: when $A_T(\hat{\sigma}_T^{2(k-1)}) > 0$, the two solutions of (19) are both larger than the naive estimator $\tilde{\sigma}_T^2[\hat{\sigma}_T^{2(k-1)}]$. By just computing the empirical variance of (implied) returns, the naive PPR/KMV estimator overlooks that the implied returns actually depend on the input $\hat{\sigma}_T^{2(k-1)}$, coming from the previous iteration. Since at this level, the partial derivative $\partial Q_T/\partial \nu$ is positive, there is some likelihood gain to take an estimator of σ^2 larger than the

naive one. When $A_T(\hat{\sigma}_T^{2(k-1)}) < 0$, it is the other way around. Note that in this case equation (19) will have only one positive solution.

The bottom line is that, even though the Black and Scholes is the simplest possible option pricing formula, everybody would prefer, both for sake of computation and of interpretation, to keep the simple sequence of empirical variances (20) as a benchmark, rather than to maximize directly (17). Our algorithm I allows us to do so, by just correcting the naive PPR/KMV estimator (which coincides exactly with the empirical variance) through the solution of a simple second degree equation.

To illustrate the usefulness of the efficient algorithms in the context of a Merton's structural credit risk model, we set up a couple of Monte Carlo experiments comparing the KMV/PPR estimator and the full MLE estimator computed using the Newton version of the efficient algorithms I and II described in section 2.2 and in appendix B.1. The results of the first experiment are detailed in table 2, and are based on 5,000 synthetic samples of 500 time series observations of daily returns. Each firm's value trajectory was initialized at 10^4, and the debt face value was fixed at $K = 9000$. We directly focused on the concentrated loglikelihood function, and set the parameters at $\mu = 0.1$ and $\sigma^2 = 0.09$. The estimates obtained using the KMV/PPR algorithm were used as starting values for the efficient algorithms iterations. The average number of iterations needed to attain convergence to the MLE ranged from 34 to 38, depending on the specific technique used.

The results in table 2 show that the KMV/PPR and the efficient estimators are almost equivalent, both in terms of bias and of dispersion, with, as expected, a slight advantage to MLE as far as the latter is concerned.

These results refer to a univariate model, which is clearly a very simple set up, not only because in real life applications the correlations between the values of different firms play a crucial role in evaluating the credit riskiness of a portfolio return, but also because it features a single parameter. Intuitively, this crucially simplifies the search for the MLE using the efficient algorithms, because there is essentially only one possible search direction, and for an algorithm to work, it is sufficient that it points the updating rule to the correct direction. To check the contracting behavior of the efficient algorithms in a more complicated model we set up a second Monte Carlo experiment based on 5,000 samples of 500 time series of daily returns for 2 firms. In this model the concentrated loglikelihood contains 3 parameters, the instantaneous variances of the two firms, and their correlation: $\theta = (\sigma_1^2, \sigma_2^2, \rho)'$. We fixed both variances to 0.09, and ρ to 0.5. The results of this experiment are illustrated in table 3.

Again, the overall performance of the inefficient KMV/PPR estimates and the MLE is quite close, with some slight edge for MLE. Algorithms I and II converged to the MLE in every replication and needed on average slightly less than 17 iterations to converge.

4 Asymptotics

4.1 Consistency

In this section we establish the consistency of the estimators defined via the algorithms presented in section 2.2 and appendix B.1. To provide a unified treatment, we introduce for each algorithm a score function $S_T(\theta, \theta_1)$ that depends not only on the parameter of interest θ through its own occurrence but also through the other occurrence θ_1 treated as a nuisance parameter. For algorithm I, we have:

$$S_T(\theta, \theta_1) = \frac{\partial Q_T[\theta, \nu(\theta_1)]}{\partial \theta} + \frac{\partial \nu(\theta_1)'}{\partial \theta} \frac{\partial Q_T[\theta_1, \nu(\theta_1)]}{\partial \nu}.$$

The expression for $S_T(\theta, \theta_1)$ for algorithm II introduced in appendix B.1 is provided in the same appendix. Let $\hat{\theta}_T^{(k)}$ be the iterative estimator obtained from step k in either algorithm I or II. Then it satisfies

$$\hat{\theta}_T^{(k)} = \arg\max_{\theta \in \Theta} \left[- \left\| S_T(\theta, \hat{\theta}_T^{(k-1)}) \right\| \right].$$

To establish the consistency of $\hat{\theta}_T^{(k)}$, we first make the following standard assumptions for proving consistency of extremum estimators.

Assumption C1. [*Uniform convergence of criterion function*] a) For any $T \geq 1$, $S_T(\theta, \theta_1)$ satisfies the standard measurability and continuity conditions; that is, it is measurable as a function of observations and it is continuous as a function of parameters (θ, θ_1);

b) There exists a limit function $S_\infty(\theta, \theta_1)$ such that $\sup_{\theta, \theta_1 \in \Theta} \|S_T(\theta, \theta_1) - S_\infty(\theta, \theta_1)\| \xrightarrow{p} 0$.

Assumption C2. [*Identification*] a) For any $\theta_1 \in \Theta$, the function $\theta \to ||S_\infty(\theta, \theta_1)||$ admits a unique minimizer $\overline{\theta}_\infty(\theta_1)$;

b) Θ is a compact subset of $\mathbb{R}^p$ and the map $\overline{\theta}_\infty(\cdot) : \Theta \to \Theta$ is continuous on Θ;

c) θ^0 is a fixed point of this map: $\theta^0 = \overline{\theta}_\infty(\theta^0)$.

Under standard regularity and concavity conditions, $\theta^0 = \overline{\theta}_\infty(\theta^0)$ can be interpreted from the limit first order conditions (2). The true unknown value θ^0 is the unique solution of $\max_{\theta \in \Theta} Q_\infty[\theta, \nu(\theta)]$. Since this problem is not solved directly but only through an iterative algorithm, consistency of the proposed estimator may take an additional assumption about

uniqueness of fixed point of the map $\overline{\theta}_\infty(\cdot)$. To see that, it is worth considering the following triangle inequality:

$$\left\|\hat{\theta}_T^{(k)} - \theta^0\right\| \leq \left\|\overline{\theta}_T(\hat{\theta}_T^{(k-1)}) - \overline{\theta}_\infty(\hat{\theta}_T^{(k-1)})\right\| + \left\|\overline{\theta}_\infty(\hat{\theta}_T^{(k-1)}) - \overline{\theta}_\infty(\theta^0)\right\| \tag{21}$$

where $\overline{\theta}_T(\theta_1)$ is the minimizer of $\|S_T(\theta, \theta_1)\|$ over $\theta \in \Theta$ defined by (8). With different, albeit similar in spirit, functions $\overline{\theta}_T(.)$ and $\overline{\theta}_\infty(\cdot)$, PPR are able to show (see their proposition 1 on page 463) that the former converges in probability (when $T \to \infty$) to the latter uniformly on Θ. This result remains obviously valid in our context, based on regularity conditions C1 and C2 above. Then we see from the inequality (21) that:

(i) The consistency (when $T \to \infty$) of $\hat{\theta}_T^{(k)}$ as an estimator of θ^0 results from the consistency of $\hat{\theta}_T^{(k-1)}$ by virtue of the continuity of the map $\overline{\theta}_\infty(\cdot)$. In other words, an iteration with a starting point $\hat{\theta}_T^{(0)}$ that is already a consistent estimator would deliver another consistent estimator $\hat{\theta}_T^{(k)}$ at each step $k = 1, 2, ...$ of the algorithm.

(ii) By contrast, but still with an argument of proof put forward by PPR (see their proposition 2 on page 463) and not reproduced here, starting the algorithm with an arbitrary value $\hat{\theta}_T^{(0)}$ will in general deliver a consistent estimator $\hat{\theta}_T^{(k)}$ only if the number of iterations is pushed to infinity with the sample size ($k \equiv k(T) \to \infty$ when $T \to \infty$) and a contraction mapping argument allows us to write:

$$\left\|\overline{\theta}_\infty(\hat{\theta}_T^{(k)}) - \overline{\theta}_\infty(\theta^0)\right\| \leq c \left\|\hat{\theta}_T^{(k)} - \theta^0\right\|, \quad 0 < c < 1.$$

We can then state the following assumptions and theorem, without need of any additional proof.

Assumption C3. [*Consistent starting point*] $\hat{\theta}_T^{(0)}$ is a weakly consistent estimator of θ^0.

Assumption C4. [*Contraction mapping*] $\overline{\theta}_\infty(\cdot)$ is contracting on Θ, that is, there is a constant $c \in]0, 1[$ such that, for any $\theta_1, \theta_2 \in \Theta$:

$$\left\|\overline{\theta}_\infty(\theta_1) - \overline{\theta}_\infty(\theta_2)\right\| \leq c \left\|\theta_1 - \theta_2\right\|.$$

Theorem 4.1 *If assumptions C1 and C2 hold, then:*
(i) under assumption C3, $\hat{\theta}_T^{(k)}$ is weakly consistent for any $k = 1, 2, ...$;
(ii) under assumption C4, $\hat{\theta}_T^{(k)}$ is weakly consistent if $k \equiv k(T) \to \infty$ when $T \to \infty$.

We emphasize here that Assumption C4, the contraction mapping condition on $\overline{\theta}_\infty(\cdot)$, is only required when Assumption C3 does not hold, i.e. the initial estimator $\hat{\theta}^{(0)}$ is not a

consistent estimator. Otherwise, if $\hat{\theta}^{(0)}$ is consistent (e.g., because it is provided by the PPR algorithm), Assumption C4 is not needed and in addition, we get the stronger result that $\hat{\theta}_T^{(k)}$ is consistent for any k.

A couple of remarks are in order here. First, using a similar approach, we can show that the results in Theorem 4.1 hold for the Newton versions of the algorithms. Second, consistency will likely need a contraction mapping condition at some level. We document in subsection 4.3 that all relevant contraction mapping conditions are implied, at least locally, by an Information Dominance (ID) condition. We use this terminology to stress that the relevant conditions are not ad hoc high level assumptions but on the contrary are implied by the structure of the inference problem at hand. In a particular context, it is natural to assume that one part of the model is more informative than another one about the unknown structural parameters. This natural assumption will provide the necessary ID condition. Additional interpretations are provided by the GMM framework as discussed in section 5 below.

4.2 Asymptotic Distribution

For the asymptotic distribution of $\hat{\theta}_T$, we adopt the following assumptions.

Assumption E1.

$$\sqrt{T}\,\frac{\partial L_T(\theta^0)}{\partial \theta} \xrightarrow{d} \mathcal{N}_p[0, \Omega(\theta^0)], \quad \text{with } \Omega(\theta^0) = \lim_{T\to\infty} \text{Var}\left[\sqrt{T}\,\frac{\partial L_T(\theta^0)}{\partial \theta}\right],$$

which is supposed to be positive definite.

Assumption E2.

$$\sup_{\theta\in\Theta}\left|\frac{\partial^2 L_T(\theta)}{\partial\theta\partial\theta'} - \frac{\partial^2 L_\infty(\theta)}{\partial\theta\partial\theta'}\right| \xrightarrow{p} 0.$$

It is known that under assumptions E1 and E2, $\hat{\theta}_T$ is asymptotically normally distributed with asymptotic variance given by

$$\left[\frac{\partial^2 L_\infty(\theta^0)}{\partial\theta\partial\theta'}\right]^{-1} \Omega(\theta^0) \left[\frac{\partial^2 L_\infty(\theta^0)}{\partial\theta\partial\theta'}\right]^{-1}.$$

We now state our main result about the asymptotic distribution of the efficient iterative estimators.

Theorem 4.2 *Suppose assumptions E1, E2, C1, and C2 hold. Then (i) under assumption C3 and if $\overline{\theta}_\infty(\cdot)$ is contracting in the neighborhood of θ^0, we obtain: $\sqrt{T}(\hat{\theta}_T^{(k)} - \hat{\theta}_T^{(k-1)}) \to 0$ in*

probability and $\hat{\theta}_T^{(k)}$ *has the same asymptotic distribution as* $\hat{\theta}_T$ *as long as* $k = k(T) \geq \log T$*; (ii) under assumption C4, we obtain:* $\sqrt{T}(\hat{\theta}_T^{(k)} - \hat{\theta}_T^{(k-1)}) \to 0$ *in probability and* $\hat{\theta}_T^{(k)}$ *has the same asymptotic distribution as* $\hat{\theta}_T$ *as long as* $k = k(T) \geq T^{\delta}$ *for some* $\delta > 0$*.*

Theorem 4.2 implies that among other things, the asymptotic efficiency of our iterative estimators relies on a contraction mapping condition. If the initial estimator is inconsistent, then this condition is the global contraction mapping condition C4; on the other hand, if the initial estimator is $\sqrt{T}$-consistent, then this condition is local and it requires $\overline{\theta}_\infty(\cdot)$ to be contracting in the neighborhood of θ^0.

Efficiency of our algorithms does not come without cost. The original PPR algorithm requires only the evaluation of the function $\nu(\theta)$. To attain efficiency, we need to make use of the full "score function", which requires the evaluation of the derivative $\partial\nu(\theta)/\partial\theta'$. This can be cumbersome in some cases, but there are important applications for which one can evaluate $\partial\nu(\theta)/\partial\theta'$ relatively easily.

4.3 Information Dominance

For each efficient algorithm, there exists a representation of the local contraction mapping condition in terms of information dominance. We present Information Dominance I for algorithm I below. Information Dominance II for algorithm II is provided in appendix B.1.

Recall that $\overline{\theta}_\infty(\theta_1)$ satisfies $S_\infty\left[\overline{\theta}_\infty(\theta_1), \theta_1\right] = 0$, implying:

$$\frac{\partial S_\infty\left[\overline{\theta}_\infty(\theta_1), \theta_1\right]}{\partial \theta'} \frac{\partial \overline{\theta}_\infty(\theta_1)}{\partial \theta_1'} + \frac{\partial S_\infty\left[\overline{\theta}_\infty(\theta_1), \theta_1\right]}{\partial \theta_1'} = 0,$$

where $\frac{\partial S_\infty[\cdot,\cdot]}{\partial \theta'}$ and $\frac{\partial S_\infty[\cdot,\cdot]}{\partial \theta_1'}$ are the partial derivatives of $S_\infty[\cdot,\cdot]$ with respect to the first and second argument, respectively. Thus, the local contraction mapping condition is equivalent to

$$\left\| \left[\frac{\partial S_\infty[\theta^0, \theta^0]}{\partial \theta'}\right]^{-1} \frac{\partial S_\infty[\theta^0, \theta^0]}{\partial \theta_1'} \right\| < 1. \tag{22}$$

(22) immediately yields the ID conditions I and II suitable for algorithms I and II. In the first one, the ID condition is given by:

Information Dominance I.

$$\left\| \left[\frac{\partial^2 Q_\infty[\theta^0, \nu(\theta^0)]}{\partial \theta \partial \theta'}\right]^{-1} \left[D^2 L_\infty(\theta^0) - \frac{\partial^2 Q_\infty[\theta^0, \nu(\theta^0)]}{\partial \theta \partial \theta'}\right] \right\| < 1.$$

Heuristically, the ID condition requires that the part of the Hessian that is used in the algorithm must dominate the part that is ignored. As a result, eventually, the impact of the part of the Hessian that is not used is negligible and the algorithm, upon convergence, produces an asymptotically efficient estimator of θ^0. If the algorithm starts from a $\sqrt{T}$-consistent estimator, the contraction mapping condition is only imposed locally at θ^0 and there is no requirement on the rate of divergence of $k(T)$, provided that it increases to ∞ with T; otherwise, the global contraction mapping condition is required and $k(T) \geq T^\delta$ for some small $\delta > 0$.

5 Implied-States GMM

The term Implied-States (IS) GMM was coined by Pan (2002) in the context of an option pricing model with latent variables. As in the example of section 3, the idea is to back out from observed option price data the latent state variables that summarize stochastic time variations in volatility, jump size and intensity through a one-to-one relationship analogous to (12). In general, the true parameter value θ^0 is identified by a set of moment conditions about the law of motion of the state variables:

$$\mathrm{E}[\psi^*(Y_t^*, \theta)] = 0 \quad \Leftrightarrow \quad \theta = \theta^0. \tag{23}$$

To implement a GMM estimator, it is natural to consider the moment conditions obtained by substituting in (23) the latent states Y_t^* with the "implied states" $g_t[Y_t, \nu(\theta)]$ computed from the option pricing model:

$$E\left[\psi[Y_t, \theta, \nu(\theta)]\right] = 0, \quad \text{where } \psi[Y_t, \theta, \nu(\theta)] = \psi^*[g_t(Y_t, \nu(\theta)), \theta]. \tag{24}$$

Note that in (24), θ occurs twice: from the latent moment (23) and from the option pricing model through $\nu(\theta)$. We expect the former occurrence to be much more user-friendly than the latter.

It is interesting to observe that, when θ is identified, the iterative procedures outlined in section 2.2 improve on the PPR approach in terms of efficiency only in the overidentified case, when the dimension H of $\psi^*(\cdot, \cdot)$ (and of $\psi(\cdot, \cdot, \cdot)$ as well) is strictly larger than the dimension p of θ. When $H = p$, $\hat{\theta}_T$ can be computed as the solution of

$$\frac{1}{T} \sum_{t=1}^{T} \psi[Y_t, \hat{\theta}_T, \nu(\hat{\theta}_T)] = 0,$$

and such an estimator can always be computed as the limit of the PPR algorithm:

$$\frac{1}{T}\sum_{t=1}^{T}\psi[Y_t,\hat{\theta}_T^{(k)},\nu(\hat{\theta}_T^{(k-1)}))]=0.$$

The efficiency issue is quite different in the overidentified case, because a subset of estimating equations could be selected in a suboptimal way. Consider the quadratic form for the two-step efficient GMM estimator of Hansen (1982):

$$Q_T[\theta,\nu(\theta)]=-\bar{\psi}_T[\theta,\nu(\theta)]'W_T(\widetilde{\theta}_T)^{-1}\bar{\psi}_T[\theta,\nu(\theta)], \tag{25}$$

where $\bar{\psi}_T[\theta,\nu(\theta)]=\frac{1}{T}\sum_{t=1}^{T}\psi[Y_t,\theta,\nu(\theta)]$ and $W_T(\widetilde{\theta}_T)$ with $\widetilde{\theta}_T$ being a preliminary consistent estimator of θ^0 is a consistent estimator of $W(\theta^0)$, the long term variance matrix of $\bar{\psi}_T[\theta^0,\nu(\theta^0)]$. Recall from section 2.1 that the PPR estimator $\hat{\theta}_T^{PPR}$ solves $\partial Q_T[\theta,\nu(\theta)]/\partial\theta=0$. For an objective function like (25), this is equivalent to solving

$$\frac{\partial\bar{\psi}_T[\theta,\nu(\theta)]'}{\partial\theta}W_T(\widetilde{\theta}_T)^{-1}\bar{\psi}_T[\theta,\nu(\theta)]=0$$

In this GMM context, the inefficiency of PPR is caused by a suboptimal selection of p linear combinations of estimating equations. The selection matrix used by PPR is a consistent estimator of:

$$\mathrm{E}\left[\frac{\partial\psi[Y_t,\theta^0,\nu(\theta^0)]'}{\partial\theta}\right]W(\theta^0)^{-1}.$$

By contrast, the efficient estimator $\hat{\theta}_T$ is based on a consistent estimator of the optimal selection matrix:

$$\Gamma(\theta^0)=\left\{\mathrm{E}\left[\frac{\partial\psi[Y_t,\theta^0,\nu(\theta^0)]'}{\partial\theta}\right]+\mathrm{E}\left[\frac{\partial\nu(\theta^0)'}{\partial\theta}\frac{\partial\psi[Y_t,\theta^0,\nu(\theta^0)]'}{\partial\nu}\right]\right\}W(\theta^0)^{-1}.$$

A similar inefficiency issue appears when it comes to the choice of optimal instruments for conditional moment restrictions. See Pan (2002, section 3.2).

Any one of the efficient algorithms introduced in section 2.2 and appendix B.1 can be used to define estimators that are asymptotically equivalent to $\hat{\theta}_T$; we provide some details in appendix B.2. The special structure of the GMM objective function, however, naturally suggests a new algorithm that exhibits significant implementation advantages. Its iteration scheme is defined as follows.

Algorithm III. Given $\hat{\theta}_T^{(k-1)}$, let $\hat{\theta}_T^{(k)}$ be the solution of:

$$\Gamma_T(\hat{\theta}_T^{(k-1)})\bar{\psi}_T[\hat{\theta}_T^{(k)},\nu(\hat{\theta}_T^{(k-1)})]=0, \tag{26}$$

where

$$\Gamma_T(\theta) = \left[\frac{\partial \bar{\psi}_T[\theta, \nu(\theta)]'}{\partial \theta} + \frac{\partial \nu(\theta)'}{\partial \theta}\frac{\partial \bar{\psi}_T[\theta, \nu(\theta)]'}{\partial \nu}\right] W_T(\theta)^{-1}$$

is the sample counterpart of the optimal selection matrix $\Gamma(\theta^0)$. As for algorithms I and II, algorithm III delivers upon convergence an efficient estimator without paying the computational price of solving w.r.t. θ equations involving $\nu(\theta)$. To check this, note that the benchmark estimator $\hat{\theta}_T$ must solve:

$$\Gamma_T(\hat{\theta}_T)\bar{\psi}_T[\hat{\theta}_T, \nu(\hat{\theta}_T)] = 0.$$

As we have seen, all efficient algorithms admit an equivalent Newton version. In the case of a GMM estimation criterion, however, these alternative iterative schemes are complicated by the need to differentiate w.r.t. θ in not only the sample moments $\bar{\psi}_T[\theta, \nu(\theta)]$, but also the selection matrix $\Gamma_T(\theta)$.[3] The main advantage of algorithm III over algorithms I and II is that in (26) the updated estimator $\hat{\theta}_T^{(k)}$ does not occur in the Jacobian matrix, but only in the sample moment conditions. This feature simplifies significantly the evaluation of the weighting matrix, yielding the following Newton type algorithm:

$$\hat{\theta}_T^{(k)} = \hat{\theta}_T^{(k-1)} - \left[\Gamma_T(\hat{\theta}_T^{(k-1)})\frac{\partial \bar{\psi}_T[\hat{\theta}_T^{(k-1)}, \nu(\hat{\theta}_T^{(k-1)})]}{\partial \theta}\right]^{-1} \Gamma_T(\hat{\theta}_T^{(k-1)})\bar{\psi}_T[\hat{\theta}_T^{(k-1)}, \nu(\hat{\theta}_T^{(k-1)})], \quad k = 1, 2, \ldots$$

The asymptotic theory of algorithm III can be studied in the same way as that of algorithms I and II; further details are provided in appendix B.2.

6 A GMM Example

In this section we illustrate an application of Algorithm III to the GMM estimation of a stochastic volatility (SV) option pricing model.

6.1 The Framework

We consider the same framework analyzed in Garcia et al. (2011), in which the information contained in realized volatilities (computed using high frequency observations of the underlying price) and option prices is combined to jointly estimate by GMM the parameters

[3] It should be noted that in $\Gamma_T(\cdot)$ the vector θ occurs in many places. Depending on the algorithm, some (but not all) of these are set to $\hat{\theta}_T^{(k-1)}$ in the k-th iteration, and hence it is not necessary to differentiate w.r.t. to these instances. Nevertheless, some degree of differentiation of $\Gamma_T(\cdot)$ w.r.t. θ is required for any one of algorithms I and II.

of a continuous time SV option pricing model. We now briefly review the model and the estimator; we defer the reader to the paper for a detailed description.

Let S_t and V_t be the date t price and volatility of the underlying asset. The objective process is given by the bivariate affine SDE

$$d\begin{bmatrix} S_t \\ V_t \end{bmatrix} = \begin{bmatrix} \mu_t S_t \\ \kappa(\bar{V} - V_t) \end{bmatrix} dt + \sqrt{V_t} \begin{bmatrix} S_t & 0 \\ \gamma\rho & \gamma\sqrt{1-\rho^2} \end{bmatrix} \begin{bmatrix} dW_t^1 \\ dW_t^2 \end{bmatrix} \tag{27}$$

and the risk-neutral process by the bivariate affine SDE

$$d\begin{bmatrix} S_t \\ V_t \end{bmatrix} = \begin{bmatrix} r_t S_t \\ \kappa^*(\bar{V}^* - V_t) \end{bmatrix} dt + \sqrt{V_t} \begin{bmatrix} S_t & 0 \\ \gamma\rho & \gamma\sqrt{1-\rho^2} \end{bmatrix} \begin{bmatrix} d\tilde{W}_t^1 \\ d\tilde{W}_t^2 \end{bmatrix} \tag{28}$$

The relationship between the objective and the risk-neutral parameters is given by $\kappa^* = \kappa - \lambda$ and $\kappa\bar{V} = \kappa^*\bar{V}^*$. The vector of parameters is $\theta = (\kappa, \bar{V}, \gamma, \rho, \lambda)'$[4].

To estimate θ we use a GMM approach based on two distinct sets of moment conditions. The first set exploits the information contained in high frequency measures of returns which can be used to consistently approximate $\mathcal{V}_{t,t+1}$, the integrated volatility of the process V_t between date t and $t+1$, using the quadratic variation estimator. These estimates can be plugged in the following set of orthogonality conditions:

$$\psi_{\mathcal{V}t}(\theta) = \begin{bmatrix} \mathcal{V}_{t+1,t+2}^k - \mathrm{E}(\mathcal{V}_{t+1,t+2}^k|\mathcal{F}_t),\ k=1,2,3 \\ (\mathcal{V}_{t+1,t+2}^k - \mathrm{E}(\mathcal{V}_{t+1,t+2}^k|\mathcal{F}_t))\mathcal{V}_{t-1,t}^k,\ k=1,2,3 \\ R_{t+1}\mathcal{V}_{t+1,t+2} - \mathrm{E}(R_{t+1}\mathcal{V}_{t+1,t+2}|\mathcal{F}_t) \\ (R_{t+1}\mathcal{V}_{t+1,t+2} - \mathrm{E}(R_{t+1}\mathcal{V}_{t+1,t+2}|\mathcal{F}_t))R_{t-1}\mathcal{V}_{t-1,t} \\ R_{t+1}^2\mathcal{V}_{t+1,t+2} - \mathrm{E}(R_{t+1}^2\mathcal{V}_{t+1,t+2}|\mathcal{F}_t) \\ (R_{t+1}^2\mathcal{V}_{t+1,t+2} - \mathrm{E}(R_{t+1}^2\mathcal{V}_{t+1,t+2}|\mathcal{F}_t))R_{t-1}^2\mathcal{V}_{t-1,t} \\ R_{t+1}\mathcal{V}_{t+1,t+2}^2 - \mathrm{E}(R_{t+1}\mathcal{V}_{t+1,t+2}^2|\mathcal{F}_t) \\ (R_{t+1}\mathcal{V}_{t+1,t+2}^2 - \mathrm{E}(R_{t+1}\mathcal{V}_{t+1,t+2}^2|\mathcal{F}_t))R_{t-1}\mathcal{V}_{t-1,t}^2 \end{bmatrix}$$

where $R_t = \log S_t - \log S_{t-1}$, $\mathcal{F}_t$ is the discrete filtration generated by the observed prices and realized volatilities and the conditional expectations in $\psi_{\mathcal{V}t}(\theta)$ are known functions of θ.

The second set of orthogonality conditions exploits the information contained in option prices using the approach proposed in Pan (2002). The idea is to invert the option pricing formula to retrieve the implied spot volatility as a function of θ, S_t and the option characteristics. The implied volatilities $V_t(\theta)$ can then be plugged in the following set of orthogonality

[4]We leave unspecified the drift term μ_t because it doesn't matter for option pricing and the inference method is robust to its specification.

conditions:

$$\psi_{Vt}(\theta) = \begin{bmatrix} V_{t+1}(\theta)^k - \mathrm{E}(V_{t+1}^k|\mathcal{F}_t),\ k=1,2,3 \\ (V_{t+1}(\theta)^k - \mathrm{E}(V_{t+1}^k|\mathcal{F}_t))V_t(\theta)^k,\ k=1,2,3 \\ R_{t+1}V_{t+1}(\theta) - \mathrm{E}(R_{t+1}V_{t+1}|\mathcal{F}_t) \\ (R_{t+1}V_{t+1}(\theta) - \mathrm{E}(R_{t+1}V_{t+1}|\mathcal{F}_t))R_{t-1}V_t(\theta) \\ R_{t+1}^2V_{t+1}(\theta) - \mathrm{E}(R_{t+1}^2V_{t+1}|\mathcal{F}_t) \\ (R_{t+1}^2V_{t+1}(\theta) - \mathrm{E}(R_{t+1}^2V_{t+1}|\mathcal{F}_t))R_{t-1}^2V_t(\theta) \\ R_{t+1}V_{t+1}(\theta)^2 - \mathrm{E}(R_{t+1}V_{t+1}^2|\mathcal{F}_t) \\ (R_{t+1}V_{t+1}(\theta)^2 - \mathrm{E}(R_{t+1}V_{t+1}^2|\mathcal{F}_t))R_{t-1}V_t(\theta)^2 \end{bmatrix}$$

where again the conditional expectations in $\psi_{Vt}(\theta)$ are known functions of θ. To simplify the option price inversion step, Garcia et al. (2011) rewrite the pricing formula as a power series w.r.t. γ around zero, for which the model can be analytically solved, and then invert it to retrieve a series expansion for the implied volatility. We adopt the same approach in the Monte Carlo experiment; additional details are available in Garcia et al. (2011).

Finally, let $\psi[Y_t, \theta, \nu(\theta)] = [\psi_{\mathcal{V}t}(\theta)', \psi_{Vt}(\theta)']'$ denote the joint vector of 24 orthogonality conditions, where Y_t contains the option price, the integrated volatility and the observed option prices at date t. The GMM estimation is conducted using a Newey-West kernel with a lag length of two, to take into account the autocorrelation structure of the moment conditions.

In implementing Algorithm III we must specify the partition of the occurrences of θ in the orthogonality conditions between the "easy" (θ) and the "nasty" ($\nu(\theta)$) ones. It would be tempting to separate the occurrences on the basis of whether they come from the analytical expressions of the conditional moments in $\psi_{\mathcal{V}T}$ or ψ_{Vt} ("easy") or from the implied spot volatilities ("nasty"). Such a partition, however, does not allow to identify the parameters on the basis of the easy occurrences only, since the conditional moments do not depend on the risk premium parameter λ. To overcome this, we also include in the easy subset all the occurrences of κ, $\bar{V}$ and λ in the implied spot volatilities.

6.2 Results

To assess the performance of GMM estimation using the MBP algorithm III in section 5 for the parameters of the SV option pricing model we conducted a Monte Carlo study using the same design adopted in Garcia et al. (2011). We generated 5,000 independent sets of $T = 960$ daily observations, corresponding to approximately 4 years. Each day is subdivided in 80 five-minute periods over which the quadratic variations are aggregated; each 5-min interval is subdivided in ten 30 seconds subintervals, used to simulate the SDE. The stock and option prices are observed at the beginning of the period over which the quadratic

variation is computed (opening prices). We assume that each date 15 options are observed, corresponding to 5 different strike prices and 3 maturities. The date t implied volatility $V_t(\theta)$ is the average over the 15 implied volatilities associated to the individual options. All the computation were conducted in Fortran using analytical first derivatives on a Dell Precision T5600 workstation with two Intel Xeon CPUs running at 2.40 Ghz. No parallelization was used.

Garcia et al. (2011) consider four parameters sets, but for the sake of simplicity we report here only the results relative to the first one, $(\kappa, \bar{V}, \gamma, \rho, \lambda) = (0.10, 0.25, 0.10, -0.50, 0.05)$ (daily percent values). Table 4 reports summary statistics for the two-step and iterated GMM estimates computed using a state-of-the-art Quasi-Newton optimizer (Liu and Nocedal, 1989); on all Monte Carlo samples, however, our Algorithm III provides exactly the same estimates. We will later compare the performance of the two algorithms in terms of CPU time required to attain convergence. The estimation results of the two-step procedure are broadly in line with those reported in Table 2, first panel of Garcia et al. (2011) (iterated GMM were not considered in that paper). The estimates seem to be almost unbiased and quite accurate, as witnessed by their standard deviations and RMSE, with the only partial exception of ρ. From the Table it is also apparent that iterated GMM performs only marginally better than two-step GMM in terms of accuracy, mostly because the distribution of the latter seems to be quite concentrated around the true values. Table 4 also reports summary statistics for the parameter estimates obtained using the inefficient PPR algorithm based on the same partition of the occurrences of θ in the moment conditions underlying the efficient Algorithm III; as for the latter, we again consider the two-step and the iterated version of the PPR estimator. In both cases, the results show that the inefficient estimator behaves significantly worse than the efficient one. The performance difference is especially apparent in the correlation parameter ρ, but even for the remaining parameters the PPR approach yields estimates with a MC RMSE which is at least 20% higher, but in some cases the increase is much larger. Overall, the Table suggests that in this example Algorithm III brings compelling improvements w.r.t. PPR.

Table 5 reports some selected percentiles of the CPU time required to compute the GMM estimates. For two-step and iterated GMM we compared the results obtained using a state-of-the-art Quasi-Newton method (L-BFGS) with those based on the Newton version of Algorithm III. Both algorithms used analytical first derivatives. As far as the standard Quasi-Newton method (L-BFGS) is concerned, it is not surprising to observe that iterated GMM takes a significantly larger computation time, required by the need to repeatedly

minimize the GMM quadratic form for a sequence of weighting matrices. For L-BFGS, the CPU time seems to grow linearly with the number of these "outer" iterations. However, this is not true for Algorithm III. The latter is almost always faster than L-BFGS even for two-step GMM, but the speed difference becomes striking for iterated GMM: with Algorithm III, the CPU time required by iterated GMM is only marginally larger than that required by two-step GMM, and is an order of magnitude smaller than the CPU time required by L-BFGS to complete iterated GMM. The reason for this result is that in Algorithm III the updating of the weighting matrix (the "outer" cycle) is absorbed in the iterations leading to the parameter estimates. Of course this could also be done with L-BFGS, which would essentially be equivalent to consider Continuous Updating GMM instead of Iterated GMM, but this would require to differentiate the weighting matrix w.r.t. θ. Such derivatives are difficult to compute analytically and their numerical evaluation would take time and provide less accurate results (see Hansen, Heaton and Yaron, 1996 for a numerical analysis of alternative GMM strategies based on MC experiments and for a discussion of the numerical issues typical of Continuous Updating GMM).

7 Conclusions

The development of non-linear structural econometrics, especially in relation to intertemporal optimization of representative agents, has put on the table estimation problems that are often viewed as computationally difficult. They typically involve optimality or equilibrium restrictions that may be computationally unpalatable when repeated at each step of an estimation algorithm (see also Su and Judd, 2011). This potential computational burden has led to the development of computationally light estimators like simulation based estimators (like indirect inference of Gouriéroux, Monfort and Renault, 1993) or two-step estimators as the so-called method of inference functions for margins (see the copula example in Joe, 1997, chap. 10, or the DCC example in Engle, 2002) or more generally inefficient iterative procedures studied in a systematic way in PPR (see also Dominitz and Sherman, 2005). The nonlinear dynamic models with latent state variables that are popular in modern asset pricing have led to the so-called implied states estimation procedures that may display the same kind of inefficiency. In this paper, we have shown that this inefficiency is always due to the overlook of the information content of some awkward occurrences of parameters in the criterion function.

Popular iterative (or two-step) procedures are precisely devised to allow this kind of over-

looking, possibly at the cost of efficiency loss. The goal of this paper was to propose efficient iterative estimation procedures whose computational cost, at each step of the iteration, is not higher than that of popular inefficient inference procedures. This is made possible by the fact that our algorithms iterate on the occurrences of the parameters that people would like to overlook. In this way, their informational content is not ignored anymore. In this respect, the efficient estimation procedures in this paper generalize the algorithms put forward in SFK in the special case of a separable loglikelihood criterion.

While we relax in this paper the separability conditions of SFK, we do not fully address yet the issue of making efficient all procedures considered in PPR. First, we do not explicitly consider the case of some nuisance parameters popping up due to error terms (respectively risk premiums) in financial asset prices due to market frictions (respectively market incompleteness). Second, recursive procedures also studied in PPR are not considered in this paper. Such procedures would be especially useful for inference on asset pricing models where investors are simultaneously learning about latent state variables. As recently emphasized by Hansen, Polson and Sargent (2010), investors of the asset pricing model are themselves faced with an implied-state inference problem where they must solve simultaneously filtering and estimation problems given signal histories. While Hansen, Polson and Sargent (2010) put forward particle filtering methods as attractive alternatives to quasi-analytical recursive algorithms based on Kalman filtering or discretization of the state space, Robbins-Monroe procedures considered in PPR could be coupled with the efficient estimation algorithms devised in the present paper.

References

Andrews, D.W.K. (2002), "Equivalence of the Higher Order Asymptotic Efficiency of k-step and Extremum Statistics",*Econometric Theory*, 18, 1040-1085.

Crouhy M., Galai D. and Mark R. (2000), "A comparative analysis of current credit risk models", *Journal of Banking and Finance*, 56(5), 24, 59-117.

Dominitz J. and Sherman R. P. (2005), "Some Convergence Theory for Iterative Estimation Procedures With an Application to Semiparametric Estimation", *Econometric Theory* 21, 838-863.

Duan J.-C., Gauthier G. and Simonato J.-G. (2005), "On the Equivalence of the KMV and Maximum Likelihood Methods for Structural Credit Risk Models", mimeo, HEC Montréal.

Engle R.(2002), "Dynamic Conditional Correlation: A Simple Class of Multivariate Generalized Autoregressive Conditional Heteroskedasticity Models", *Journal of Business & Economic Statistics* 20, 339-350.

Engle R. and Sheppard K. (2001), "Theoretical and Empirical Properties of Dynamic Conditional Correlation Multivariate GARCH", Discussion Paper 2001-15, UCSD.

Garcia, R., Lewis M.-A., Pastorello S. and Renault E. (2011), "Estimation of Objective and Risk-Neutral Distributions Based on Moments of Integrated Volatility", *Journal of Econometrics*, 160(1), 22-32.

Gouriéroux C., Monfort A. and Renault E. (1993), "Indirect Inference" , *Journal of Applied Econometrics*, 8(S1), S85-S118.

Hansen, L.P. (1982), "Large Sample Properties of Generalized Method of Moments Estimators," *Econometrica*, 50, 1029-1054.

Hansen L. P., Heaton J. and Yaron A. (1996), "Finite -Sample Properties of Some Alternative GMM Estimators", *Journal of Business & Economic Statistics*, 14(3), 262-280.

Hansen L. P., Polson N. and Sargent T. J. (2010), "Nonlinear Filtering and Robust Learning", *Journal of Business & Economic Statistics* Invited Lecture, ASSA Winter Meetings, Atlanta.

Joe H. (1997), *Multivariate Models and Dependence Concepts*, Chapman&Hall, London.

Liao J. G. and Qaqish B. F. (2005), "Comment to Song, Fan and Kalbfleisch (2005)", *Journal of the American Statistical Association* 100(472), 1160-1161.

Liu D. C. and Nocedal, J. (1989), "On the Limited Memory Method for Large Scale Optimization", *Mathematical Programming B* 45(3), 503–528.

Merton R. (1974), "On the Pricing of Corporate Debt: the Risk Structure of Interest Rates", *Journal of Finance* 28, 449-470.

Pan J. (2002), "The Jump-Risk Premia Implicit in Options: Evidence From an Integrated Time-Series Study", *Journal of Financial Economics*, 63(1), 3-50.

Pastorello S., Patilea V. and Renault E. (2003), "Iterative and Recursive Estimation in Structural Nonadaptive Models", *Journal of Business & Economic Statistics* 21, 449-482.

Robinson, P.M. (1988), "The Stochastic Difference Between Econometric Statistics," *Econometrica*, 56, 531-548.

Song P., Fan Y. and Kalbfleisch J. (2005), "Maximization by Parts in Likelihood Inference", *Journal of the American Statistical Association* 100, 1145-1158.

Su C.-L. and Judd K. L. (2011), "Constrained Optimization Approaches to Estimation of Structural Models", *Econometrica*, forthcoming.

A Technical Proofs

Theorem 4.2 The proof involves two steps: first, we show that $\sqrt{T}(\hat{\theta}_T^{(k)} - \hat{\theta}_T^{(k-1)}) \xrightarrow{p} 0$; second, we establish the asymptotic distribution of $\hat{\theta}_T^{(k)}$.

We provide a sketch of the proof for algorithm I only. Theorem 4.1 implies the consistency of $\hat{\theta}_T^{(k)}$ under the conditions in (i) and (ii) in Theorem 4.2. Applying Taylor expansion to both sides of

$$\frac{\partial Q_T[\hat{\theta}_T^{(k)}, \nu(\hat{\theta}_T^{(k-1)})]}{\partial \theta} = -\frac{\partial \nu(\hat{\theta}_T^{(k-1)})'}{\partial \theta} \frac{\partial Q_T[\hat{\theta}_T^{(k-1)}, \nu(\hat{\theta}_T^{(k-1)})]}{\partial \nu},$$

at θ^0, collecting terms, and ignoring higher order terms, we obtain:

$$\hat{\theta}_T^{(k)} - \theta^0 = f_T(\theta^0) + F(\theta^0)\left(\hat{\theta}_T^{(k-1)} - \theta^0\right),$$

where:

$$\begin{aligned} f_T(\theta^0) &= \left[-\frac{\partial^2 Q_\infty[\theta_0, \nu(\theta_0)]}{\partial\theta\partial\theta'}\right]^{-1}\left[\frac{\partial L_T(\theta^0)}{\partial\theta}\right], \\ F(\theta^0) &= \left[-\frac{\partial^2 Q_\infty[\theta_0, \nu(\theta_0)]}{\partial\theta\partial\theta'}\right]^{-1}\left[D^2 L_\infty(\theta^0) - \frac{\partial^2 Q_\infty[\theta_0, \nu(\theta_0)]}{\partial\theta\partial\theta'}\right]. \end{aligned}$$

Iterating the above equation, we get:

$$\hat{\theta}_T^{(k)} - \theta^0 = \sum_{j=0}^{k-1} F(\theta^0)^j f_T(\theta^0) + F(\theta^0)^k(\hat{\theta}_T^{(0)} - \theta^0).$$

The local contraction mapping condition of $\overline{\theta}_\infty(\cdot)$ at θ^0 implies that $\left\|F(\theta^0)\right\| < 1$, see section 4.3. Thus, we obtain that

$$\sqrt{T}(\hat{\theta}_T^{(k)} - \hat{\theta}_T^{(k-1)}) = F(\theta^0)^{k-1}\sqrt{T}f_T + \left\{\sqrt{T}F(\theta^0)^{k-1}\right\}\left[F(\theta^0) - I\right](\hat{\theta}_T^{(0)} - \theta^0) = o_p(1)$$

under the respective conditions in (i) and (ii) in Theorem 4.2.

For the second step, by expanding terms involving $\hat{\theta}_T^{(k)}$ in its definition and collecting terms, we get

$$\hat{\theta}_T^{(k)} = \hat{\theta}_T^{(k-1)} - \left[G_T(\hat{\theta}_T^{(*1)}, \hat{\theta}_T^{(*2)})\right]^{-1} \left[\frac{\partial L_T(\hat{\theta}_T^{(k-1)})}{\partial \theta}\right],$$

where the expression of $G_T(\theta, \theta_1)$ for algorithm I is:

$$G_T(\theta, \theta_1) = \frac{\partial^2 Q_T[\theta, \nu(\theta_1)]}{\partial \theta \partial \theta'}$$

and $\hat{\theta}_T^{(*1)}, \hat{\theta}_T^{(*2)}$ lie between $\hat{\theta}_T^{(k)}$ and $\hat{\theta}_T^{(k-1)}$. The expression of $G_T(\theta, \theta_1)$ for algorithm II is given in appendix B.1.

Using

$$\frac{\partial L_T(\hat{\theta}_T^{(k-1)})}{\partial \theta} = \frac{\partial L_T(\theta^0)}{\partial \theta} + \frac{\partial^2 L_T(\hat{\theta}_T^{(*0)})}{\partial \theta \partial \theta'}(\hat{\theta}_T^{(k-1)} - \theta^0),$$

we get

$$\hat{\theta}_T^{(k)} = \hat{\theta}_T^{(k-1)} - \left[G_T(\hat{\theta}_T^{(*1)}, \hat{\theta}_T^{(*2)})\right]^{-1} \left[\frac{\partial L_T(\theta^0)}{\partial \theta} + \frac{\partial^2 L_T(\hat{\theta}_T^{(*0)})}{\partial \theta \partial \theta'}(\hat{\theta}_T^{(k-1)} - \theta^0)\right].$$

Hence, ignoring the higher order terms:

$$\begin{aligned}
&\hat{\theta}_T^{(k)} - \theta^0 \\
&= -\left[G_T(\theta^0, \theta^0)\right]^{-1} \left[\frac{\partial L_T(\theta^0)}{\partial \theta}\right] - \left[G_T(\theta^0, \theta^0)\right]^{-1} [D^2 L_T(\theta^0) - G_T(\theta^0, \theta^0)](\hat{\theta}_T^{(k-1)} - \theta^0) \\
&= -\left[G_T(\theta^0, \theta^0)\right]^{-1} \left[\frac{\partial L_T(\theta^0)}{\partial \theta}\right] \\
&\qquad - \left[G_T(\theta^0, \theta^0)\right]^{-1} [D^2 L_T(\theta^0) - G_T(\theta^0, \theta^0)](\hat{\theta}_T^{(k)} - \theta^0) + o_p(T^{-1/2}),
\end{aligned}$$

because $\sqrt{T}(\hat{\theta}_T^{(k)} - \hat{\theta}_T^{(k-1)}) \to 0$.

Rearranging the above equation leads to

$$\left[D^2 L_T(\theta^0)\right](\hat{\theta}_T^{(k)} - \theta^0) = -\frac{\partial L_T(\theta^0)}{\partial \theta} + o_p(T^{-1/2})$$

and hence the conclusion that the asymptotic distribution of $\hat{\theta}_T^{(k)}$ is the same as that of $\hat{\theta}_T$.□

B Additional Efficient Algorithms

In this appendix, we discuss some additional efficient algorithms.

B.1 General Extremum Estimation Criterion

For the case of an estimation criterion $L_T(\theta) = Q_T[\theta, \nu(\theta)]$, let us introduce algorithm II using the following iterative scheme. We assume that iterations are started at some initial value $\hat{\theta}_T^{(0)}$, that may or may not be a consistent estimator of θ^0.

Algorithm II. For $k = 1, 2, 3, \ldots$, let $\hat{\theta}_T^{(k)}$ solve:

$$\frac{\partial Q_T[\theta, \nu(\hat{\theta}_T^{(k-1)})]}{\partial \theta} = -\frac{\partial \nu(\hat{\theta}_T^{(k-1)})'}{\partial \theta} \frac{\partial Q_T[\theta, \nu(\hat{\theta}_T^{(k-1)})]}{\partial \nu}.$$

The algorithm admits an alternative and computationally convenient Newton version. It is listed below. To simplify the notation, let

$$H_T(\theta, \theta_1) = \frac{\partial^2 Q_T[\theta, \nu(\theta_1)]}{\partial \theta \nu'} \frac{\partial \nu(\theta_1)}{\partial \theta'} \quad \text{and} \quad \Sigma_T(\theta, \theta_1) = \frac{\partial^2 Q_T[\theta, \nu(\theta_1)]}{\partial \theta \partial \theta'}.$$

To simplify the comparison between the algorithms, we also rewrite here the Newton version of algorithm I (formula (9)) using the simplified notation.

Algorithm I (Newton version). For $k = 1, 2, 3, \ldots$, compute $\hat{\theta}_T^{(k)}$ as:

$$\hat{\theta}_T^{(k)} = \hat{\theta}_T^{(k-1)} - \left[\Sigma_T(\hat{\theta}_T^{(k-1)}, \hat{\theta}_T^{(k-1)})\right]^{-1} \left[\frac{\partial L_T(\hat{\theta}_T^{(k-1)})}{\partial \theta}\right].$$

Algorithm II (Newton version). For $k = 1, 2, 3, \ldots$, compute $\hat{\theta}_T^{(k)}$ as:

$$\hat{\theta}_T^{(k)} = \hat{\theta}_T^{(k-1)} - \left[\Sigma_T(\hat{\theta}_T^{(k-1)}, \hat{\theta}_T^{(k-1)}) + H_T(\hat{\theta}_T^{(k-1)}, \hat{\theta}_T^{(k-1)})'\right]^{-1} \left[\frac{\partial L_T(\hat{\theta}_T^{(k-1)})}{\partial \theta}\right].$$

The asymptotic behavior of algorithm II is the same as algorithm I. To show this, it is sufficient to apply Theorems 4.1 and 4.2 in section 4 with the following definition of the score function $S_T(\theta, \theta_1)$ and weighting matrix $G_T(\theta, \theta_1)$:

$$\begin{aligned} S_T(\theta, \theta_1) &= \frac{\partial Q_T[\theta, \nu(\theta_1)]}{\partial \theta} + \frac{\partial \nu(\theta_1)'}{\partial \theta} \frac{\partial Q_T[\theta, \nu(\theta_1)]}{\partial \nu}, \\ G_T(\theta, \theta_1) &= \Sigma_T(\theta, \theta_1) + H_T(\theta, \theta_1)' \end{aligned}$$

Finally, the information dominance condition for algorithm II is as follows. To simplify the notation, let $\Sigma(\theta, \theta)$ and $H(\theta, \theta)$ denote the population counterparts of $\Sigma_T(\theta, \theta)$ and $H_T(\theta, \theta)$, respectively. Again, to facilitate comparisons we also include the Information Dominance I condition introduced in section 4.3.

Information Dominance I.

$$\left\| \left[\Sigma(\theta^0,\theta^0)\right]^{-1} \left[D^2 L_\infty(\theta^0) - \Sigma(\theta^0,\theta^0)\right] \right\| < 1.$$

Information Dominance II.

$$\left\| \left[\Sigma(\theta^0,\theta^0) + H(\theta^0,\theta^0)'\right]^{-1} \left[D^2 L_\infty(\theta^0) - \Sigma(\theta^0,\theta^0) - H(\theta^0,\theta^0)'\right] \right\| < 1.$$

B.2 GMM Estimation Criterion

In section 5 we considered an estimation criterion given by the quadratic form for efficient GMM estimation. The following expressions describe the iterative schemes characterizing the two efficient algorithms introduced in section 2.2 and appendix B.1 in this case. As usual, we assume that the iterations are started at some initial value $\hat{\theta}_T^{(0)}$, that may or may not be a consistent estimator of θ^0.

Algorithm I (GMM criterion). For $k = 1, 2, 3, \ldots$, let $\hat{\theta}_T^{(k)}$ solve:

$$\frac{\partial \bar{\psi}_T[\hat{\theta}_T^{(k)}, \nu(\hat{\theta}_T^{(k-1)})]'}{\partial \theta} W_T(\hat{\theta}_T^{(k-1)})^{-1} \bar{\psi}_T[\hat{\theta}_T^{(k)}, \nu(\hat{\theta}_T^{(k-1)})] =$$
$$- \frac{\partial \nu(\hat{\theta}_T^{(k-1)})'}{\partial \theta} \frac{\partial \bar{\psi}_T[\hat{\theta}_T^{(k-1)}, \nu(\hat{\theta}_T^{(k-1)})]'}{\partial \nu} W_T(\hat{\theta}_T^{(k-1)})^{-1} \bar{\psi}_T[\hat{\theta}_T^{(k-1)}, \nu(\hat{\theta}_T^{(k-1)})]$$

Algorithm II (GMM criterion). For $k = 1, 2, 3, \ldots$, let $\hat{\theta}_T^{(k)}$ solve:

$$\frac{\partial \bar{\psi}_T'[\hat{\theta}_T^{(k)}, \nu(\hat{\theta}_T^{(k-1)})]}{\partial \theta} W_T(\hat{\theta}_T^{(k-1)})^{-1} \bar{\psi}_T[\hat{\theta}_T^{(k)}, \nu(\hat{\theta}_T^{(k-1)})] =$$
$$- \frac{\partial \nu(\hat{\theta}_T^{(k-1)})'}{\partial \theta} \frac{\partial \bar{\psi}_T'[\hat{\theta}_T^{(k)}, \nu(\hat{\theta}_T^{(k-1)})]}{\partial \nu} W_T(\hat{\theta}_T^{(k-1)})^{-1} \bar{\psi}_T[\hat{\theta}_T^{(k)}, \nu(\hat{\theta}_T^{(k-1)})]$$

The asymptotic behavior of the efficient algorithms I-II applied to a GMM estimation criterion is simply a special case of the general results provided by Theorems 4.1 and 4.2 in section 4. The expressions of the score function $S_T(\theta, \theta_1)$ and weighting matrix $G_T(\theta, \theta_1)$ are tedious but straightforward to derive. Finally, using (22), it is easy to show that the information dominance conditions for algorithms I and II are described by the following expressions. To simplify the notation, let $\Psi(\theta^0) = \Psi_0(\theta^0) + \Psi_1(\theta^0)$, where

$$\Psi_0(\theta^0) = \mathrm{E}\left[\frac{\partial \psi[Y_t, \theta^0, \nu(\theta^0)]}{\partial \theta'}\right] \quad \text{and} \quad \Psi_1(\theta^0) = \mathrm{E}\left[\frac{\partial \psi[Y_t, \theta^0, \nu(\theta^0)]}{\partial \nu'} \frac{\partial \nu(\theta^0)}{\partial \theta'}\right].$$

Information Dominance I (GMM criterion).

$$\left\| \left[\Psi_0(\theta^0)' W(\theta^0)^{-1} \Psi_0(\theta^0)\right]^{-1} \left[\Psi(\theta^0)' W(\theta^0)^{-1} \Psi(\theta^0) - \Psi_0(\theta^0)' W(\theta^0)^{-1} \Psi_0(\theta^0)\right] \right\| < 1.$$

Information Dominance II (GMM criterion).

$$\left\| \left[\Psi(\theta^0)' W(\theta^0)^{-1} \Psi_0(\theta^0) \right]^{-1} \left[\Psi(\theta^0)' W(\theta^0)^{-1} \Psi_1(\theta^0) \right] \right\| < 1.$$

Finally, to study the asymptotic theory of algorithm III, it is sufficient to introduce the score function:

$$S_T^V(\theta, \theta_1) = \Gamma_T(\theta_1) \bar{\psi}_T[\theta, \nu(\theta_1)].$$

Consistency and asymptotic efficiency then follow from Theorems 4.1 and 4.2. It is also straightforward to show that the Information Dominance III condition coincides with Information Dominance II (GMM criterion) above.

Table 1: Descriptive statistics of the results of a Monte Carlo experiment comparing two estimators of the parameters of a DCC model: the two steps estimator proposed by Engle (2002), and the MLE computed using the Newton version of the MBP algorithm. The results are based on 5,000 synthetic samples of 1,000 time series observations of 2 daily returns. In terms of the MBP algorithm described in section 2.1, $\theta_1 = (\omega_1, \kappa_1, \lambda_1, \omega_2, \kappa_2, \lambda_2)'$ is the subvector of GARCH parameters, and $\theta_2 = (a, b)'$ is the subvector of the correlation parameters. The estimates obtained with the two-step procedure were used as starting values for the MBP iterations. The average number of MBB iterations needed to attain convergence was 12.58, with a standard deviation of 5.11.

		Two steps estimator			MLE		
	True	Mean	Std. Dev.	RMSE	Mean	Std. Dev.	RMSE
ω_1	1.000	1.031	0.299	0.301	1.038	0.295	0.298
κ_1	0.050	0.050	0.014	0.014	0.050	0.013	0.013
λ_1	0.940	0.930	0.030	0.032	0.933	0.026	0.028
ω_2	1.667	1.670	0.135	0.135	1.667	0.128	0.128
κ_2	0.200	0.200	0.046	0.046	0.200	0.040	0.040
λ_2	0.500	0.483	0.117	0.119	0.485	0.101	0.103
a	0.050	0.050	0.012	0.012	0.050	0.012	0.012
b	0.940	0.933	0.024	0.026	0.933	0.019	0.020

Table 2: Descriptive statistics of the results of a Monte Carlo experiment comparing two estimators of the σ^2 parameter of the Merton's structural credit risk model with one firm: the KMV/PPR estimator outlined in subsection 2.1, and the MLE computed using the Newton versions of the efficient algorithms developed in subsection 2.2 and appendix B.1 in the text. The results are based on 5,000 synthetic samples of 500 time series observations of daily returns (see the text for further details on the MC experiment setup). The KMV/PPR estimates were used as starting values for the efficient algorithms' iterations. The average number of iterations needed to attain convergence to MLE ranged from 34 to 38, depending on the specific algorithm.

		KMV/PPR			MLE		
Parameter	True	Mean	Std. Dev.	RMSE	Mean	Std. Dev.	RMSE
σ^2	0.0900	0.0903	0.0895	0.0126	0.0902	0.0894	0.0122

Table 3: Descriptive statistics of the results of a Monte Carlo experiment comparing two estimators of the parameters of the Merton's structural credit risk model with two firms: the KMV/PPR estimator outlined in subsection 3, and the full MLE computed using the Newton versions of the efficient algorithms developed in subsection 2.2 and appendix B.1 in the text. The results are based on 5,000 synthetic samples of 500 time series observations of daily returns (see the text for further details on the MC experiment setup). The KMV/PPR estimates were used as starting values for the efficient algorithms' iterations. Of the two efficient algorithms, only algorithm I converged to the MLE in every replication; algorithm II converged to the MLE in 89.92% of the replications. The former needed on average slightly less than 17 iterations, whereas the latter needed roughly 44 iterations.

		KMV/PPR			MLE		
Parameter	True	Mean	Std. Dev.	RMSE	Mean	Std. Dev.	RMSE
σ_1^2	0.0900	0.0905	0.0128	0.0128	0.0903	0.0121	0.0121
σ_2^2	0.0900	0.0903	0.0125	0.0125	0.0902	0.0120	0.0120
ρ	0.5000	0.5001	0.0336	0.0336	0.5000	0.0334	0.0334

Table 4: Descriptive statistics of the results of a Monte Carlo experiment comparing two GMM (two-step and iterated) and two PPR estimators of the parameters of the SV option pricing model model considered in Garcia et al. (2011) and outlined in section 6.1. The results are based on 5,000 synthetic samples of 960 observations of the daily return, quadratic variation and 15 option prices (see the text for further details on the MC experiment setup). The table reports summary statistics of the parameter estimates.

	Parameter	κ	$\bar{V}$	γ	ρ	λ
	True	0.100	0.250	0.100	-0.500	0.050
Two-step GMM	Mean	0.100	0.246	0.097	-0.501	0.051
	Median	0.100	0.245	0.097	-0.501	0.051
	Std. Dev.	0.008	0.016	0.006	0.032	0.006
	RMSE	0.008	0.016	0.007	0.032	0.006
Two-step PPR	Mean	0.098	0.247	0.096	-0.480	0.051
	Median	0.100	0.246	0.098	-0.429	0.051
	Std. Dev.	0.012	0.023	0.008	0.123	0.008
	RMSE	0.012	0.023	0.009	0.125	0.008
Iterated GMM	Mean	0.101	0.247	0.098	-0.501	0.051
	Median	0.101	0.247	0.099	-0.501	0.051
	Dev. Std.	0.007	0.014	0.005	0.030	0.006
	RMSE	0.007	0.014	0.005	0.030	0.006
Iterated PPR	Mean	0.100	0.247	0.098	-0.485	0.051
	Median	0.100	0.247	0.098	-0.438	0.050
	Dev. Std.	0.010	0.016	0.006	0.118	0.007
	RMSE	0.010	0.016	0.006	0.119	0.007

Table 5: Descriptive statistics of the results of a Monte Carlo experiment comparing two GMM (two-step and iterated) estimators of the parameters of the SV option pricing model model considered in Garcia et al. (2011) and outlined in section 6.1. The results are based on 5,000 synthetic samples of 960 observations of the daily return, quadratic variation and 15 option prices (see the text for further details on the MC experiment setup). The table reports selected percentiles of the CPU times (in seconds) required to attain convergence by two algorithms, L-BFGS (a state-of-the-art Quasi-Newton method; see Liu and Nocedal, 1989) and Algorithm III (see section 5). Both algorithms converged in all samples.

	Two-step GMM		Iterated GMM	
Percentile	L-BFGS	Algorithm III	L-BFGS	Algorithm III
0.5%	6.0	2.3	42.1	3.1
2.5%	8.6	3.2	54.4	4.1
25%	11.6	5.1	90.3	6.4
50%	16.0	6.2	114.6	7.7
75%	23.3	7.9	145.3	9.6
97.5%	34.1	18.5	249.6	20.0
99.5%	39.6	50.8	418.8	34.1