

Maximization by Parts in Extremum Estimation*

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

March 2006

This version: May 2009

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-Ralphson algorithm is difficult, if not impossible. While the Newton-Ralphson 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 convergence and asymptotic properties of our algorithms under regularity conditions including the information dominance conditions. We apply our algorithms to the estimation of Merton's structural credit risk model.

*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, Gardner Hall, University of North Carolina at Chapel Hill, Chapel Hill, NC 27599-3305, 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. In cases where the dimension of the parameter is large, the method of concentration is often used to reduce the dimension of the parameter over which optimization is carried out, see e.g., Amemiya (1985) or Ruud (2000) for concentrating the likelihood function. In general, let parameter vector ξ be partitioned into two sub-vectors θ and ν as $\xi = (\theta', \nu')'$. Let $Q_T(\theta, \nu)$ denote the sample objective function. Given any θ , one can find the optimal value of ν as a function of θ by solving:

$$\nu_T(\theta) = \arg \max_{\nu} Q_T(\theta, \nu). \quad (1)$$

The concentrated objective function (or profile objective function) is then defined as $Q_T[\theta, \nu_T(\theta)]$ and its maximization with respect to θ obviously gives the extremum estimator of interest:

$$\theta_T = \arg \max_{\theta} Q_T[\theta, \nu_T(\theta)] \Rightarrow [\theta_T, \nu_T(\theta_T)] = \arg \max_{(\theta, \nu)} Q_T(\theta, \nu). \quad (2)$$

Note that the concentrated objective function $Q_T[\theta, \nu_T(\theta)]$ depends on θ both directly and indirectly through $\nu_T(\theta)$, where the latter arises from concentration. Pastorello, Patilea and Renault (2003) (PPR hereafter) note, however, that such a structure arise naturally in many examples in Economics and Finance. In contrast to concentration, such a structure in these examples stems from economic theory. This difference between a concentrated objective function and a general objective function of the same structure has major implications on efficient estimation of the parameter θ . To see this, let the parameter of interest be some unknown p -dimensional parameter $\theta^0 \in \Theta \subset \mathbb{R}^p$. The extremum estimator of θ^0 is denoted by $\hat{\theta}_T$ defined as:

$$\hat{\theta}_T = \arg \max_{\theta \in \Theta} Q_T[\theta, \nu(\theta)] \quad (3)$$

for some known function $\nu(\theta)$ and some given criterion function:

$$Q_T(\theta, \nu) = Q_{(T)}[\theta, \nu, (Y_t)_{1 \leq t \leq T}]$$

associated with a sample $(Y_t)_{1 \leq t \leq T}$ of observations. Note that the assumption that the function $\nu(\cdot)$ is known and not dependent on the sample is not really restrictive since the

sample dependence can always be incorporated within the definition of the function $Q_T(\cdot)$. We maintain this assumption in order to simplify the exposition. However, in order to encompass the case of a preliminary concentration step giving rise to a sample dependent function $\nu_T(\theta)$ the definition of which depends on some ex-ante specified function $Q_T(\theta, \nu)$, we study more carefully the asymptotic distributions of resulting estimators in the second part of Section 3 below.

In general, the extremum estimator $\hat{\theta}_T$ satisfies the first order condition,

$$\frac{\partial Q_T(\theta, \nu(\theta))}{\partial \theta} + \frac{\partial Q_T(\theta, \nu(\theta))}{\partial \nu'} \frac{\partial \nu(\theta)}{\partial \theta} = 0. \quad (4)$$

If, instead, $Q_T[\theta, \nu(\theta)]$ is a concentrated objective function, then the second term in the above expression is exactly zero and the extremum estimator satisfies

$$\frac{\partial Q_T(\theta, \nu(\theta))}{\partial \theta} = 0. \quad (5)$$

The presence of the second term on the left hand side of (4) makes solving $\hat{\theta}_T$ more difficult, in particular, given the fact that in most examples, either the objective function is numerically cumbersome to maximize with respect to the second occurrence of θ and/or it is the source of some explosive behavior of the objective function. The latter includes the case of ill-behaved likelihood functions produced by models with latent variables.

To avoid maximization with respect to the second occurrence of θ in the objective function $Q_T[\theta, \nu(\theta)]$, PPR devised an extension of the idea of backfitting. They first note that if the true unknown value $\nu^0 = \nu(\theta^0)$ were known, we could estimate θ by solving

$$\max_{\theta \in \Theta} Q_T[\theta, \nu^0]. \quad (6)$$

By doing this, we would avoid having to play with the nasty occurrence of the function $\nu(\theta)$ within the criterion function. Of course, this is not feasible but gives the idea of an iterative estimation algorithm defined by the sequence:

$$\hat{\theta}_T^{(k+1)} = \arg \max_{\theta \in \Theta} Q_T[\theta, \nu(\hat{\theta}_T^{(k)})]. \quad (7)$$

Upon convergence, this algorithm will provide a consistent estimator of θ^0 , denoted as $\hat{\theta}_{B,T} = \lim_{k \rightarrow \infty} \hat{\theta}_T^{(k)}$. Note that $\hat{\theta}_{B,T}$ satisfies (5), but not (4) in general and hence is not as efficient as $\hat{\theta}_T$ unless the function $\nu(\theta)$ comes from a concentration step like (1).

To better understand the connection between the two extremum problems (3) and (6) and hence the estimators $\hat{\theta}_T$ and $\hat{\theta}_{B,T}$, let us introduce the limit objective function $Q_\infty[\theta, \nu(\theta)] = P \lim_{T \rightarrow \infty} Q_T[\theta, \nu(\theta)]$, where the limit is in probability uniformly with respect to $\theta \in \Theta$. Typically, a consistency argument for $\hat{\theta}_T$ would be based on the maintained assumption that the true unknown value θ^0 is the only solution of:

$$\theta^0 = \arg \max_{\theta \in \Theta} Q_\infty[\theta, \nu(\theta)], \quad (8)$$

while consistency of $\hat{\theta}_{B,T}$ should result (see PPR for more discussion) from identification of θ^0 as the only solution of:

$$\theta^0 = \arg \max_{\theta \in \Theta} Q_\infty[\theta, \nu^0]. \quad (9)$$

Under these maintained identification assumptions, the asymptotic variances of these consistent asymptotically normal estimators will be determined from sample counterparts of the first order conditions, respectively written as:

$$\frac{\partial Q_\infty(\theta^0, \nu(\theta^0))}{\partial \theta} + \frac{\partial Q_\infty(\theta^0, \nu(\theta^0))}{\partial \nu'} \frac{\partial \nu(\theta^0)}{\partial \theta} = 0 \quad (10)$$

for $\hat{\theta}_T$ and as:

$$\frac{\partial Q_\infty(\theta^0, \nu(\theta^0))}{\partial \theta} = 0 \quad (11)$$

for $\hat{\theta}_{B,T}$. Note that we can deduce from (10) and (11) that θ^0 must also satisfy

$$\frac{\partial Q_\infty(\theta^0, \nu(\theta^0))}{\partial \nu'} \frac{\partial \nu(\theta^0)}{\partial \theta} = 0. \quad (12)$$

Given that, for validity of PPR, one must assume that (11) identifies θ^0 , one can think of (12) as additional moment conditions satisfied by θ^0 and (10) provides an optimal way to combine these two sets of moment conditions. In general, since $\hat{\theta}_T$ makes use of both sets of moment conditions optimally, we see it as asymptotically efficient. On the other hand, since $\hat{\theta}_{B,T}$ makes use of one of the two sets of moment conditions (11) only, it is asymptotically less efficient than $\hat{\theta}_T$, except if by chance (10) and (11) are equivalent. This is precisely the case when the function $\nu(\theta)$ is (at least asymptotically) the result of a preliminary concentration stage since we would have then:

$$\frac{\partial Q_\infty(\theta, \nu(\theta))}{\partial \nu} = 0 \text{ for all } \theta.$$

To compute the extremum estimator $\hat{\theta}_T$, we need to solve the first order condition (4). Our focus is on cases where the occurrence of $\nu(\theta)$ in the objective function is of complicated form so that the second term on the left hand side of (4) renders solving (4) directly for $\hat{\theta}_T$ difficult if not impossible. However the first term on the left hand side of (4) may be of a simple form. The log-likelihood function for the observables in Merton's (1974) credit risk model provides one such example, see Section 4 for details.

Song, Fan and Kalbfleisch (2005) (SFK hereafter) considered a special case of (3) in which the objective function $Q_T(\cdot, \nu(\cdot))$ is a log-likelihood function, taking the following additive form:

$$Q_T[\theta, \nu(\theta)] = Q_{T1}(\theta) + Q_{T2}(\nu(\theta)). \quad (13)$$

In this case, the estimator $\hat{\theta}_T$ defined in (3) is the Maximum Likelihood Estimator (MLE) of θ^0 . SFK provides numerous examples for which the log-likelihood function is of this form and the full MLE may be difficult to compute directly, as $Q_{T2}(\nu(\theta))$ depends on θ in a complicated way. Here, we point out another important example: the Dynamic Conditional Correlation MV-GARCH model of Engle (2002) in which the first part $Q_{T1}(\theta)$ is the volatility term and the second is the correlation term. To simplify the estimation, Engle (2002) proposed a two-step estimator and noted that it is consistent, but not efficient in general. To solve the computational issue involved in computing the full MLE in these cases, SFK proposed an iterative algorithm, maximization by parts (MBP), which produces the MLE upon convergence. Computationally, each step in MBP is no more difficult than maximizing the first term $Q_{T1}(\theta)$ and hence is well suited to examples in which $Q_{T1}(\theta)$ is of a much simpler form than the second term $Q_{T2}(\nu(\theta))$. We provide a brief review of MBP in Section 2. The backfitting estimator of PPR when applied to the additive function $Q_T[\theta, \nu(\theta)]$ defined in (13) leads to $\hat{\theta}_{B,T} = \operatorname{argmax}_{\theta \in \Theta} Q_{T1}(\theta)$. Under regularity conditions, $\hat{\theta}_{B,T}$ is a consistent estimator of θ^0 , but is less efficient than $\hat{\theta}_T$, the MLE.

The objective of this paper is to extend the algorithm in SFK to the general extremum estimation problem characterized by the condition (8). We provide a number of different algorithms which, upon convergence, all approach the efficient estimator $\hat{\theta}_T$. These algorithms differ from each other in terms of which θ in the three different places that $\nu(\theta)$ appears in (4) to iterate on. In cases where the condition (9) also holds, we can start our efficient algorithms from the consistent PPR estimator $\hat{\theta}_{B,T}$ or the modified PPR estimators presented in Section 3. We present conditions under which our algorithms converge and establish asymptotic

properties of the corresponding estimators.

The rest of this paper is organized as follows. In Section 2, we apply the iterative algorithm of SFK to estimating the DCC MV-GARCH model of Engle (2002). In Section 3, we first review the backfitting estimator of PPR and present two of its modifications. Then we provide four iterative algorithms grouped in Algorithms I and II, which, upon convergence, provide estimators that have the same asymptotic distribution as the efficient estimator $\hat{\theta}_T$. We provide conditions that guarantee their convergence for both consistent and inconsistent initial values. The classical KMV iterative approach to estimating Merton's credit risk model provides a convenient example to compare the backfitting algorithms with the full MLE. This is done in Section 4. Typically, the backfitting algorithms, including the classical KMV iterative procedure, entail some efficiency loss that is avoided by the new algorithms proposed in Section 3. Section 5 considers the application of the algorithms in Section 4 to the implied state GMM and develops a new algorithm making use of the special structure of the IS-GMM objective function. The last section concludes. Technical proofs are relegated to Appendix A. Algorithms III and IV are presented in Appendix B.

2 Two Versions of MBP and an Application to DCC MV-GARCH Model

2.1 MBP

SFK developed an iterative algorithm, referred to as MBP, for an extension of (13):

$$Q_T[\theta, \nu(\theta)] = Q_{T1}(\theta_1) + Q_{T2}(\nu(\theta)), \quad (14)$$

where $\theta' = (\theta'_1, \theta'_2)$ and $\nu(\theta) = \theta$. The algorithm is based on the structure of the first order condition for $\hat{\theta}_T$:

$$\begin{aligned} \frac{\partial Q_{T1}(\theta_1)}{\partial \theta_1} &= -\frac{\partial Q_{T2}(\theta_1, \theta_2)}{\partial \theta_1}, \\ \frac{\partial Q_{T2}(\theta_1, \theta_2)}{\partial \theta_2} &= 0. \end{aligned}$$

The j th step in the iteration of MBP is:

$$\begin{aligned} \frac{\partial Q_{T1}(\theta_1)}{\partial \theta_1} &= -\frac{\partial Q_{T2}(\hat{\theta}_{1,j-1}, \hat{\theta}_{2,j-1})}{\partial \theta_1}, \\ \frac{\partial Q_{T2}(\hat{\theta}_{1,j}, \theta_2)}{\partial \theta_2} &= 0. \end{aligned} \quad (15)$$

To start the algorithm, we need an initial value for θ_1 . In SFK, $Q_T[\theta, \nu(\theta)]$ is a log-likelihood function in which the first term $Q_{T1}(\theta_1)$ is the log-likelihood function ignoring the possible dependence of the observations and the second term $Q_{T2}(\theta) = Q_{T2}(\theta_1, \theta_2)$ accounts for the dependence of the observations. For example, the log-likelihood function of a copula-based model has this structure in which θ_1 denotes parameters in the marginal distributions and θ_2 denotes parameters in the copula. A consistent estimator for θ_1 is given by the solution to:

$$\frac{\partial Q_{T1}(\theta_1)}{\partial \theta_1} = 0$$

which serves as a candidate for the initial value for θ_1 . Denote this as $\hat{\theta}_{1,1}$. Solving

$$\frac{\partial Q_{T2}(\hat{\theta}_{1,1}, \theta_2)}{\partial \theta_2} = 0$$

leads a consistent estimator of θ_2 denoted as $\hat{\theta}_{2,1}$. We note that for copula-based models, the estimator $(\hat{\theta}_{1,1}, \hat{\theta}_{2,1})$ is often used as an alternative to MLE due to its computational ease. In fact, this is also Engle's two-step estimator for the DCC MV-GARCH model, see below for details on this. SFK showed that under their information dominance condition, iterating from $(\hat{\theta}_{1,1}, \hat{\theta}_{2,1})$ via (15) yields MLE upon convergence.

In their discussion of SFK, Liao and Qaqish suggest the use of a one-step Newton-Raphson update leading to another version of MBP:

$$\begin{aligned} \hat{\theta}_{1,j} &= \hat{\theta}_{1,j-1} - \left[\frac{\partial^2 Q_{T1}(\hat{\theta}_{1,j-1})}{\partial \theta_1 \partial \theta_1'} \right]^{-1} \left[\frac{\partial Q_T(\hat{\theta}_{1,j-1}, \hat{\theta}_{2,j-1})}{\partial \theta_1} \right], \\ \hat{\theta}_{2,j} &= \hat{\theta}_{2,j-1} - \left[\frac{\partial^2 Q_{T2}(\hat{\theta}_{1,j}, \hat{\theta}_{2,j-1})}{\partial \theta_2 \partial \theta_2'} \right]^{-1} \left[\frac{\partial Q_{T2}(\hat{\theta}_{1,j}, \hat{\theta}_{2,j-1})}{\partial \theta_2} \right]. \end{aligned} \quad (16)$$

Following the proof of Theorem 3 in SFK, one can show that under the conditions of Theorem 3 in SFK, the above modified version of SFK has the same asymptotic properties as the original MBP and yet has the advantage of being computationally less costly.

2.2 The DCC MV-GARCH model

The DCC MV-GARCH model of Engle (2002) can be formulated as follows:

$$\begin{aligned}
r_t | \mathcal{F}_{t-1} &\sim N(0, D_t R_t D_t), \\
D_t^2 &= \text{diag} \{ \omega_i \} + \text{diag} \{ \kappa_i \} \circ r_{t-1} r'_{t-1} + \text{diag} \{ \lambda_i \} \circ D_{t-1}^2, \\
\epsilon_t &= D_t^{-1} r_t, \\
Q_t &= S \circ (\iota \iota' - A - B) + A \circ \epsilon_{t-1} \epsilon'_{t-1} + B \circ Q_{t-1}, \\
R_t &= \text{diag} \{ Q_t \}^{-1/2} Q_t \text{diag} \{ Q_t \}^{-1/2}.
\end{aligned}$$

Engle (2002) proposed a two-step approach for estimating the parameters in DCC. Let the parameters in D be denoted as θ_1 and the additional parameters in R be denoted as θ_2 . Let $\theta' = (\theta'_1, \theta'_2)$. The log likelihood function for θ can be written as the sum of a volatility part and a correlation part:

$$L(\theta) = L_V(\theta_1) + L_C(\theta_1, \theta_2), \quad (17)$$

where the volatility term is

$$\begin{aligned}
L_V(\theta_1) &= -\frac{1}{2} \sum_{t=1}^T (n \log(2\pi) + \log |D_t|^2 + r'_t D_t^{-2} r_t) \\
&= -\frac{1}{2} \sum_{t=1}^T \sum_{i=1}^n \left(\log(2\pi) + \log(h_{i,t}) + \frac{r_{i,t}^2}{h_{i,t}} \right),
\end{aligned}$$

and the correlation component is

$$L_C(\theta_1, \theta_2) = -\frac{1}{2} \sum_{t=1}^T (\log |R_t| + \epsilon'_t R_t^{-1} \epsilon_t - \epsilon'_t \epsilon_t).$$

We note that (17) is of the additive form (14) and can be efficiently estimated by MBP. The two-step approach of Engle (2002) is to find

$$\hat{\theta}_1 = \arg \max_{\theta_1} [L_V(\theta_1)]$$

and then take this value as given in the second stage:

$$\hat{\theta}_2 = \arg \max_{\theta_2} [L_C(\hat{\theta}_1, \theta_2)].$$

Engle (2002) and Engle and Sheppard (2001) established the asymptotic properties of $\hat{\theta} = (\hat{\theta}_1, \hat{\theta}_2)$. Compared with the full MLE of θ , they noted that the two-step estimator is easier

to compute, but asymptotically less efficient. It is interesting to observe that Engle's two-step estimator is exactly the initial value $(\hat{\theta}'_{1,1}, \hat{\theta}'_{2,1})$ from which SFK iterates their algorithm. At each step j , the computation of the correlation parameter $\hat{\theta}_{2,j}$ is the same as that of Engle's two-step estimator so any existing program for DCC can be used for this purpose. The computation of $\hat{\theta}_{1,j}$ is slightly different because of the need to evaluate $\frac{\partial L_C(\hat{\theta}_{1,j-1}, \hat{\theta}_{2,j-1})}{\partial \theta_1}$.

We may also relate Engle's two-step estimator to PPR by noting that from the following expression for $L(\theta)$:

$$L(\theta) = L_V(\theta_1) + L_C(\nu(\theta), \theta_2),$$

where $\nu(\theta) = \theta_1$, applying PPR from an arbitrary initial value $(\theta_1^{(1)}, \theta_2^{(1)})$ will lead to Engle's two step estimator after two iterations:

$$\begin{aligned} \hat{\theta}_1 &= \theta_1^{(k)} \text{ for all } k > 1, \\ \hat{\theta}_2 &= \theta_2^{(k)} \text{ for all } k > 2. \end{aligned}$$

2.3 Numerical Results

To illustrate the usefulness of the MBP algorithms in the context of a DCC model, we set up a Monte Carlo experiment comparing Engle's (2002) two step estimator and the full MLE estimator computed using the Newton-Raphson version of MBP described by (16). The results are detailed in table 1, and they are based on 5,000 synthetic samples of 1,000 time series observations of 2 daily returns. The values of the GARCH parameters coincide with those used by Engle (2002, section 5); they imply two processes characterized by a fairly different degree of persistence (the first one is highly persistent, and the second is not). The values of the correlation parameters are roughly coherent with those reported by Engle and Sheppard (2001, table 1), and they imply high persistency in the correlation between the two 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 iterations of the two steps estimator were started at the true parameter values; the resulting estimates were used as starting values for the MBP iterations. The average number of MBP iterations needed to attain convergence was 12.58, with a standard deviation of 5.11.

The two estimators are roughly equivalent in terms of bias; their averages over the Monte

Carlo replications are indeed extremely close to each other, and to the true values of the parameters. As expected, however, they differ in terms of precision, with the ML estimator characterized by a lower RMSE. The reduction in variance is highest for the autocorrelation parameters λ_1 , λ_2 and b , but also for κ_2 . Overall, the gain in precision ranges from 0 to 30%, depending on the parameter under scrutiny. MBP allows to attain this improvement without the need to resort to the full Newton-Raphson iterates required by standard ML estimation, thus significantly reducing the computational burden associated to efficient estimation of the parameters.

3 Efficient Algorithms for Extremum Estimation

In this section, we develop several iterative procedures for solving the first order condition (4) which upon convergence approach $\hat{\theta}_T$, the extremum estimator. We first review the backfitting algorithm of PPR and present two modifications, and then describe our algorithms and their asymptotic properties including consistency and asymptotic distribution.

The backfitting estimator and its modified versions are consistent estimators of θ^0 under the identifying assumption B1, see below. Under our maintained assumption that θ^0 solves (10), the backfitting estimator and its modified ones are in general asymptotically less efficient than $\hat{\theta}_T$. In addition to being more efficient than the backfitting estimator of PPR and its modified versions, our new algorithms also have the advantage of not requiring the identification condition B1. As discussed in PPR, neither condition B1 nor (10) imply each other. However, since the estimator of interest to us in this paper is $\hat{\theta}_T$ defined through (3), (10) is our maintained identification condition for θ^0 . If, in addition, the identifying assumption B1 also holds, we can start each of our algorithms from either the backfitting estimator or its modified versions; otherwise, we start each algorithm from an arbitrary value. We discuss conditions that guarantee convergence in each case.

3.1 The Backfitting Algorithm of PPR and two Modifications

The main focus of PPR is on providing iterative algorithms to estimate θ^0 based on (9) only. They first note that this requires a specific identification assumption that, under standard regularity conditions making optimization equivalent to first order conditions, can be written as:

Assumption B1.

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

We refer the reader to PPR for many examples and counterexamples concerning the validity of such an identification assumption, which is maintained throughout this section.

The purpose of this section is to show that the issue of estimation of θ^0 based on (9) under the maintained assumption B1 can be addressed in at least three ways which, when they work, provide asymptotically equivalent estimators of θ^0 . One of them is the initial PPR method. It turns out that, according to our implementation experience, one method or the other may be preferred depending on the context of application and the shape of the function $Q_\infty[\theta, \nu(\theta)]$.

3.1.1 First Modified PPR Estimator

The most natural way to use identifying assumption B1 for the purpose of consistent estimation is to look for $\hat{\theta}_{B,T}$, solution of:

$$\frac{\partial Q_T}{\partial \theta}[\hat{\theta}_{B,T}, \nu(\hat{\theta}_{B,T})] = 0. \quad (18)$$

In all applications to M-estimators (see Section 5 below for a slightly different setting), $Q_T[\theta, \nu(\theta)]$ is a sample mean and $Q_\infty[\theta, \nu(\theta)]$ is the corresponding population expectation. In other words, (18) is nothing but a just-identified GMM estimator that can in practice be obtained by minimizing an arbitrary norm of the vector $\frac{\partial Q_T}{\partial \theta}[\theta, \nu(\theta)]$. Thus, under standard regularity conditions justifying the application of GMM asymptotic theory, the first modified PPR estimator $\hat{\theta}_{B,T}$ is consistent and asymptotically normal, with an asymptotic variance simply deduced from standard GMM formula. In order to state the result, we assume that the empirical moment functions are asymptotically normal:

Assumption B2. $\sqrt{T} \frac{\partial Q_T}{\partial \theta}(\theta, \nu(\theta^0))|_{\theta=\theta^0} \xrightarrow{d} N_p(0, B(\theta^0))$, with

$$B(\theta^0) = \lim_{T \rightarrow \infty} \text{Var}(\sqrt{T} \frac{\partial Q_T}{\partial \theta}(\theta, \nu(\theta^0))|_{\theta=\theta^0}),$$

which is supposed to be positive definite.

In addition, it is worth realizing that the Jacobian matrix of the moment conditions is actually the difference $H(\theta, \theta) - \Sigma(\theta, \theta)$ of two matrices defined as:

$$\Sigma(\theta, \theta^1) = -\frac{\partial^2 Q_\infty}{\partial \theta \partial \theta'}[\theta, \nu(\theta^1)] \text{ and } H(\theta, \theta^1) = \frac{\partial^2 Q_\infty}{\partial \theta \partial \nu'}[\theta, \nu(\theta^1)] \frac{\partial \nu(\theta^1)}{\partial \theta}. \quad (19)$$

Assume that θ^0 is an interior point of the compact set $\Theta \subset R^p$. Suppose that Assumptions B1 and B2 hold. Then under standard regularity conditions, we have straightforwardly :

Proposition 3.1 $\hat{\theta}_{B,T}$ is asymptotically normal with asymptotic variance matrix

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

where $A(\theta^0) = I_p - \Sigma(\theta^0, \theta^0)^{-1}H(\theta^0, \theta^0)$.

For practical implementation and for the sake of comparison between this estimator and other estimators put forward in this paper, it is worth considering a Newton algorithm based version of it, that is to see $\hat{\theta}_{B,T}$ as limit of an iterative scheme: $\hat{\theta}_{B,T} = \lim_{k \rightarrow \infty} \hat{\theta}_T^{(k)}$, where $\{\hat{\theta}_T^{(k)}\}$ are obtained from:

$$\hat{\theta}_T^{(k)} = \hat{\theta}_T^{(k-1)} - \left[H_{2T}(\hat{\theta}_T^{(k-1)}) \right]^{-1} \frac{\partial Q_T(\hat{\theta}_T^{(k-1)}, \nu(\hat{\theta}_T^{(k-1)}))}{\partial \theta}, \quad (20)$$

in which

$$H_{2T}(\theta) = \frac{\partial^2 Q_T(\theta, \nu(\theta))}{\partial \theta \partial \theta'} + \frac{\partial^2 Q_T[\theta, \nu(\theta)]}{\partial \theta \partial \nu'} \frac{\partial \nu(\theta)}{\partial \theta}. \quad (21)$$

3.1.2 PPR Backfitting Estimator Revisited

The starting point of PPR backfitting is an interpretation of the identifying assumption B1 as defining the true unknown value θ^0 as the only fixed point of a contraction mapping. They assume more precisely that for any $\theta^1 \in \Theta$, the equation

$$\frac{\partial Q_\infty}{\partial \theta}[\theta, \nu(\theta^1)] = 0$$

admits a unique solution $\theta = \bar{\theta}(\theta_1)$ and that the function $\bar{\theta}(\theta_1)$ is differentiable and contractive as implied by their maintained assumption:

$$\left\| \frac{\partial \bar{\theta}}{\partial \theta'}(\cdot) \right\| < 1.$$

Then the function $\bar{\theta}(\cdot)$ admits a unique fixed point and it is sufficient that this fixed point is indeed the true unknown value θ^0 to imply the identification assumption B1. As noted by PPR, the identity:

$$\frac{\partial Q_\infty}{\partial \theta}[\bar{\theta}(\theta^0), \nu(\theta^0)] = 0$$

implies by differentiation that:

$$-\Sigma(\theta^0, \theta^0) \frac{\partial \bar{\theta}}{\partial \theta'}(\theta^0) + H(\theta^0, \theta^0) = 0.$$

This identity sheds more light on both the PPR contraction mapping condition and on the asymptotic variance of GMM estimator of Proposition 3.1.

As noted by PPR, their contraction mapping condition at θ^0 is equivalent to

$$\left\| \frac{\partial \bar{\theta}}{\partial \theta'}(\theta^0) \right\| = \left\| \Sigma(\theta^0, \theta^0)^{-1} H(\theta^0, \theta^0) \right\| < 1, \quad (22)$$

and thus can be interpreted in terms of the relative size of various blocks of the Hessian matrix. Moreover, the asymptotic variance of GMM characterized in Proposition 3.1 above can be rewritten as:

$$V(\theta^0) = \left[I_p - \frac{\partial \bar{\theta}}{\partial \theta'}(\theta^0) \right]^{-1} \Sigma(\theta^0, \theta^0)^{-1} B(\theta^0) \Sigma(\theta^0, \theta^0)^{-1} \left[I_p - \frac{\partial \bar{\theta}}{\partial \theta'}(\theta^0) \right]^{-1}.$$

Note that the inner term $\Sigma(\theta^0, \theta^0)^{-1} B(\theta^0) \Sigma(\theta^0, \theta^0)^{-1}$ is nothing but the asymptotic variance of the extremum estimator associated with the infeasible criterion $Q_T[\cdot, \nu(\theta^0)]$. As extensively discussed by PPR, one may expect that more often than not the additional factor $\left[I_p - \frac{\partial \bar{\theta}}{\partial \theta'}(\theta^0) \right]^{-1}$ in the sandwich formula, with $\left\| \frac{\partial \bar{\theta}}{\partial \theta'}(\cdot) \right\| < 1$, implies that the asymptotic variance of “efficient” GMM $\hat{\theta}_{B,T}$ is larger than the one of this infeasible extremum estimator. It is worth reminding however that, in contrast with PPR, the benchmark for efficiency in this paper is the asymptotic variance of the feasible extremum estimator associated with the criterion $Q_T[\cdot, \nu(\cdot)]$.

The main benefit of the contraction mapping assumption in PPR is to allow them to define an alternative estimator $\hat{\theta}_T^{(k)}$ from an iterated approximation scheme defined by the recursion (7). Noting that $\hat{\theta}_T^{(k)}$ satisfies

$$\frac{\partial Q_T(\hat{\theta}_T^{(k)}, \nu(\hat{\theta}_T^{(k-1)}))}{\partial \theta} = 0, \quad (23)$$

the original PPR backfitting algorithm can be regarded as an iterative approach to solving (18), or to computing $\hat{\theta}_{B,T}$. However, PPR stress that, while the contraction property (22) is about the asymptotic criterion $Q_\infty[\cdot, \nu(\cdot)]$, it might be the case that the finite sample criterion $Q_T[\cdot, \nu(\cdot)]$ never gives rise to a similar contraction property. In other words, the convergence of (23) when k goes to infinity is not guaranteed. This is the reason why PPR

propose a slightly different estimator defined as $\hat{\theta}_T^{(k(T))}$, where $k(T)$ goes to infinity with T at a sufficiently fast rate. To understand under what conditions this estimator is asymptotically equivalent to the backfitting estimator $\hat{\theta}_{B,T}$, it is worth comparing equation (18) with the following equation defining $\hat{\theta}_T^{(k(T))}$:

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

It is then clear that the two estimators are going to have the same first order asymptotic expansions insofar as one can neglect the additional term:

$$\frac{\partial Q_T(\theta^0, \nu(\theta^0))}{\partial \nu'} \frac{\partial \nu(\theta^0)}{\partial \theta'} \sqrt{T} (\hat{\theta}_T^{(k(T))} - \hat{\theta}_T^{(k(T)-1)}).$$

This remark allows PPR to prove the following result:

Proposition 3.2 *Assume that θ^0 is an interior point of the compact set $\Theta \subset R^p$. Suppose that in addition to assumptions B1 and B2, the contraction mapping condition (22) and other conditions in Proposition 3 in PPR hold. Then, insofar as $\sqrt{T}(\hat{\theta}_T^{(k(T))} - \hat{\theta}_T^{(k(T)-1)}) \rightarrow 0$ in probability, $\hat{\theta}_T^{(k(T))}$ is a consistent estimator of θ^0 , asymptotically normal and asymptotically equivalent to $\hat{\theta}_{B,T}$.*

In their Remark 4, PPR also show that their maintained assumption $\sqrt{T}(\hat{\theta}_T^{(k(T))} - \hat{\theta}_T^{(k(T)-1)}) \rightarrow 0$ in probability encompasses the case put forward by Dominitz and Sherman (2005) where a so-called ‘‘asymptotic contraction mapping’’ assumption is maintained with $k(T) = T^\delta$ for some $\delta > 0$. Note that the contraction mapping condition in PPR and that in Dominitz and Sherman (2005) basically require the mapping to be asymptotically contractive in the whole parameter space¹ Θ , i.e., (22) holds for all $\theta \in \Theta$, which is necessary for the algorithm to converge from an arbitrary initial value for θ .

3.1.3 Another Modified PPR Estimator

An alternative way of implementing the backfitting algorithm is by the following updating rule:

$$\hat{\theta}_T^{*(k)} = \hat{\theta}_T^{*(k-1)} - \left[\frac{\partial^2 Q_T(\hat{\theta}_T^{*(k-1)}, \nu(\hat{\theta}_T^{*(k-1)}))}{\partial \theta \partial \theta'} \right]^{-1} \frac{\partial Q_T(\hat{\theta}_T^{*(k-1)}, \nu(\hat{\theta}_T^{*(k-1)}))}{\partial \theta}. \quad (24)$$

¹Or at least the mapping is contractive on the closed ball of radius $\|\hat{\theta}^{(0)} - \theta^0\|$ about θ^0 , where $\hat{\theta}^{(0)}$ is some initial value.

The updating rule (24) is obtained by applying the one-step Newton algorithm to (23). At first glance, it is not obvious at all why a Newton-Raphson principle applied once to the maximization problem (7) should be sufficient, given that for a fixed iteration k , $\hat{\theta}_T^{*(k-1)}$ may not be a consistent estimator of θ^0 . However, it is rather clear that when considering on the one hand an estimator $\hat{\theta}_T^{(k(T))}$ defined from iterations (23) and on the other hand an estimator $\hat{\theta}_T^{*(k(T))}$ defined from iterations (24) with the same maintained condition $\sqrt{T}(\hat{\theta}^{(k(T))} - \hat{\theta}^{(k(T)-1)}) \rightarrow 0$ in probability, we get in both cases consistent estimators the asymptotic distribution of which are characterized by the same Taylor expansion. In other words the three PPR estimators $\hat{\theta}_{B,T}$, $\hat{\theta}_T^{(k(T))}$ and $\hat{\theta}_T^{*(k(T))}$ are asymptotically equivalent. Obviously, iterating the updating rule (24) is computationally less costly than the backfitting algorithm in PPR, but it requires that $Q_T(\theta, \nu(\theta^1))$ be at least twice differentiable with respect to θ . To shed even more light on the proximity of these two estimators, note that (24) is nothing but a finite sample counterpart of a population recursion defined by:

$$\theta^{(k+1)} = \bar{\theta}^*(\theta^{(k)})$$

where

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

is a local approximation of the function $\bar{\theta}(\cdot)$ considered in the previous subsection. In particular, under the additional assumption of existence of relevant third derivatives of the function $Q_\infty(\theta, \nu(\theta))$, the contraction mapping assumption (22) about the function $\bar{\theta}(\cdot)$ also works for the function $\bar{\theta}^*(\cdot)$ since:

$$\begin{aligned} & \frac{\partial \bar{\theta}^*}{\partial \theta'}(\theta^0) \\ &= I_p - \left[\frac{\partial^2 Q_\infty(\theta^0, \nu(\theta^0))}{\partial \theta \partial \theta'} \right]^{-1} \left[\frac{\partial^2 Q_\infty(\theta^0, \nu(\theta^0))}{\partial \theta \partial \theta'} + \frac{\partial^2 Q_\infty}{\partial \theta \partial \nu'}[\theta^0, \nu(\theta^0)] \frac{\partial \nu(\theta^0)}{\partial \theta} \right] \\ &= - \left[\frac{\partial^2 Q_\infty(\theta^0, \nu(\theta^0))}{\partial \theta \partial \theta'} \right]^{-1} \left[\frac{\partial^2 Q_\infty}{\partial \theta \partial \nu'}[\theta^0, \nu(\theta^0)] \frac{\partial \nu(\theta^0)}{\partial \theta} \right] = \frac{\partial \bar{\theta}}{\partial \theta'}(\theta^0) \end{aligned}$$

where the first equality is obtained by direct differentiation of $\bar{\theta}^*(\theta)$ defined above, while noting that occurrences of third order derivatives are deleted by the factor $\frac{\partial Q_\infty(\theta^0, \nu(\theta^0))}{\partial \theta}$ which is zero by virtue of the maintained identification assumption B1. Note also that the recursion (24) which defines the estimator $\hat{\theta}_T^{*(k(T))}$ can be seen as a simplification of the recursion (20)

that we had proposed to compute $\hat{\theta}_{B,T}$. Of course, this simplification is convenient if and only if the required contraction property is fulfilled.

The general message of this section is that we have at our disposal three versions of a PPR kind of estimator that are asymptotically equivalent. These estimators are implemented via updating rules (15), (23), and (24) respectively and can be regarded as different iterative methods for computing $\hat{\theta}_{B,T}$ defined in (18). As already explained, these three estimators are not efficient in general, except if by chance (10) and (11) are equivalent, which is the case when the function $\nu(\theta)$ is the result of a preliminary concentration stage. Indeed, in this case, we have by definition:

$$\frac{\partial Q_T(\theta, \nu_T(\theta))}{\partial \nu} = 0$$

for all θ , and thus

$$\frac{\partial^2 Q_T(\theta, \nu_T(\theta))}{\partial \theta \partial \nu'} = 0.$$

Obviously, in this case, the two modified PPR algorithms coincide when the first one is computed with the recursion (20). Moreover, under standard regularity conditions, including the existence of a limit $\nu(\theta)$ for the function $\nu_T(\theta)$, the matrix $H(\theta, \theta)$ should be identically zero. In other words the recursions defining the PPR backfitting estimator and its second modification are sample counterparts of population recursions defined by functions $\bar{\theta}(\cdot)$ and $\bar{\theta}^*(\cdot)$ respectively both of which have a zero derivative at the true value θ^0 . Therefore, the contraction mapping condition required by the PPR estimator and the second modified PPR estimator is expected to hold in at least some neighborhood of θ^0 .

It is also worth noticing that all the above considerations still apply when the “efficient” criterion $Q_T[\theta, \nu(\theta)]$ is itself the result of a preliminary concentration step:

$$Q_T[\theta, \nu(\theta)] = Q_T^*[\theta, \nu(\theta), \alpha_T(\theta)],$$

where

$$\alpha_T(\theta) = \arg \max_{\alpha} Q_T^*[\theta, \nu(\theta), \alpha].$$

3.2 New algorithms

The first order condition (4) for the asymptotically efficient estimator $\hat{\theta}_T$ is of a complicated nonlinear form, which depends on θ in several places. Depending on the specific applications, some terms in (4) may be more complicated than others. In this section, we present four

algorithms, grouped in Algorithm I and Algorithm II, for solving (4), which iterate on the more complicated terms. Two additional sets of algorithms are provided in Appendix B.

For notational compactness, we let $L_T(\theta) = Q_T(\theta, \nu(\theta))$. Then

$$\frac{\partial L_T(\theta)}{\partial \theta} = \frac{\partial Q_T(\theta, \nu(\theta))}{\partial \theta} + \frac{\partial Q_T(\theta, \nu(\theta))}{\partial \nu} \frac{\partial \nu(\theta)}{\partial \theta},$$

and the efficient estimator $\hat{\theta}_T$ satisfies:

$$\frac{\partial L_T(\hat{\theta}_T)}{\partial \theta} = 0.$$

In addition to simplifying the computation of $\hat{\theta}_T$, our algorithms may also be useful when the objective function $L_T(\theta)$ is flat around its maximum.

Algorithm I.

Step 1. We start out our algorithm from an initial estimator denoted as $\hat{\theta}_T^{(0)}$;

Step k. Let $\hat{\theta}_T^{(k)}$ solve ($k = 1, 2, 3, \dots$):

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

Step k'. Let $\hat{\theta}_T^{(k)}$ solve ($k = 1, 2, 3, \dots$):

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

Implementing Step k' will be computationally less costly than implementing Step k. Moreover, if the algorithm starts from a consistent initial estimator such as the backfitting estimator of PPR or its modified versions, we can further simplify the updating rule in Step k' as follows:

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

Step k' makes it clear that Algorithm I is similar to the second modified backfitting algorithm (24) except that Algorithm I makes use of the full score in its iterations, while the second modified backfitting algorithm (24) uses part of the score only. This is the reason why in contrast to the modified backfitting algorithm (24), upon convergence, Algorithm I delivers an asymptotically efficient estimator.

It is also instructive to compare Algorithm I with the Newton-Raphson algorithm for computing $\hat{\theta}_T$. The Newton-Raphson algorithm takes the following form:

$$\tilde{\theta}_T^{(k)} = \tilde{\theta}_T^{(k-1)} - [D^2 L_T(\tilde{\theta}_T^{(k-1)})]^{-1} \left[\frac{\partial L_T(\tilde{\theta}_T^{(k-1)})}{\partial \theta} \right], \quad (25)$$

where

$$\begin{aligned} D^2 L_T(\theta) = & \frac{\partial^2 Q_T(\theta, \nu(\theta))}{\partial \theta \partial \theta'} + \frac{\partial^2 Q_T}{\partial \theta \partial \nu'}[\theta, \nu(\theta)] \frac{\partial \nu(\theta)}{\partial \theta} + \frac{\partial Q_T(\theta, \nu(\theta))}{\partial \nu} \nu''(\theta) \\ & + \left[\frac{\partial^2 Q_T(\theta, \nu(\theta))}{\partial \nu \partial \theta'} + \frac{\partial^2 Q_T(\theta, \nu(\theta))}{\partial \nu \partial \nu'} \nu'(\theta) \right] \nu'(\theta). \end{aligned} \quad (26)$$

Instead of using the full Hessian as the Newton-Raphson algorithm, Algorithm I makes use of the first term in the expression for the Hessian matrix only. This explains why in sharp contrast to the Newton-Raphson algorithm, one-step iteration of Algorithm I will not deliver an asymptotically efficient estimator even when it starts from a consistent estimator. The advantage of Algorithm I over the Newton-Raphson, however, lies in the computational ease.

The above discussion suggests the possibility of developing other algorithms by using different terms in the expression for Hessian in (26). Algorithm II below makes use of the first two terms in the expression for Hessian in (26) and can be regarded as the efficient version of the first modified backfitting algorithm (18). Two additional algorithms are provided in Appendix B. For a specific application, one can choose any of these algorithms depending on their computational ease.

Algorithm II.

Step 1. We start out our algorithm from an initial estimator denoted as $\hat{\theta}_T^{(0)}$;

Step k. Let $\hat{\theta}_T^{(k)}$ solve ($k = 1, 2, 3, \dots$):

$$\frac{\partial Q_T(\theta, \nu(\theta))}{\partial \theta} = - \frac{\partial Q_T(\hat{\theta}_T^{(k-1)}, \nu(\hat{\theta}_T^{(k-1)}))}{\partial \nu} \frac{\partial \nu}{\partial \theta}(\hat{\theta}_T^{(k-1)}); \quad (27)$$

Step k'. Let $\hat{\theta}_T^{(k)}$ solve ($k = 1, 2, 3, \dots$):

$$\hat{\theta}_T^{(k)} = \hat{\theta}_T^{(k-1)} - [H_{2T}(\hat{\theta}_T^{(k-1)})]^{-1} \left[\frac{\partial L_T(\hat{\theta}_T^{(k-1)})}{\partial \theta} \right], \quad (28)$$

where $H_{2T}(\theta)$ is defined in (21).

Again, if the algorithm starts from a consistent initial estimator such as the backfitting estimator of PPR or its modified versions, we can further simplify Step k' as follows:

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

Remark 1. When $Q_T(\cdot, \cdot)$ is of the additive form (13), both algorithms reduce to that of SFK. That is, $\hat{\theta}_T^k$ solves:

$$Q'_{T1}(\theta) = -Q'_{T2}(\nu(\hat{\theta}_T^{(k-1)})) \frac{\partial \nu}{\partial \theta}(\hat{\theta}_T^{(k-1)}).$$

SFK show that under regularity conditions, upon convergence, $\hat{\theta}_T^k$ is asymptotically as efficient as $\hat{\theta}_T$. We will extend their results to our general objective function in the next subsection.

3.3 Asymptotic theory

In this subsection, we establish consistency and asymptotic distribution of the estimators defined via the algorithms presented in Subsection 3.2 and Appendix B. To provide a unified treatment, we introduce $S_T(\theta, \theta_1)$, where for the original algorithms defined via Step k,

$$\begin{aligned} S_T(\theta, \theta_1) &= \frac{\partial Q_T(\theta, \nu(\theta_1))}{\partial \theta} + \frac{\partial Q_T(\theta_1, \nu(\theta_1))}{\partial \nu} \frac{\partial \nu}{\partial \theta}(\theta_1) \text{ for Algorithm I} \\ &= \frac{\partial Q_T(\theta, \nu(\theta))}{\partial \theta} + \frac{\partial Q_T(\theta_1, \nu(\theta_1))}{\partial \nu} \frac{\partial \nu}{\partial \theta}(\theta_1) \text{ for Algorithm II.} \end{aligned}$$

The expressions for $S_T(\theta, \theta_1)$ for Algorithms III and IV are provided in Appendix B. Let $\hat{\theta}_T^{(k)}$ be an iterative estimator obtained from Step k in any of the algorithms described above. Then it satisfies

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

To establish the consistency of $\hat{\theta}_T^{(k)}$, we make the following assumptions.

Assumption C1. 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 $S_T(\theta, \theta_1) \xrightarrow{p} S_\infty(\theta, \theta_1)$ for any $(\theta, \theta_1) \in \Theta \times \Theta$.

Assumption C2. a) For any $\theta_1 \in \Theta$, the function $\theta \rightarrow \|S_\infty(\theta, \theta_1)\|$ admits a unique minimizer $\bar{\theta}(P^0, \theta_1)$, where P^0 is the probability measure governing the observations; b) $\theta^0 = \bar{\theta}(P^0, \theta^0)$.

Assumption C3. θ^0 is the unique fixed point of the map $\bar{\theta}(P^0, \cdot)$.

Assumption C4. $\sup_{\theta, \theta_1 \in \Theta} \|S_T(\theta, \theta_1) - S_\infty(\theta, \theta_1)\| \xrightarrow{P} 0$.

Assumption C5. $\bar{\theta}(P^0, \cdot)$ is contracting on Θ .

Theorem 3.3 *Assume that Θ is a compact subset of R^p . (i) If assumptions C1, C2a), C4, and C5 hold, then $\hat{\theta}_T^{(k)}$ with $k = k(T) \rightarrow \infty$, is consistent; (ii) If assumptions C1, C2, and C4 hold and $\hat{\theta}_T^{(0)}$ is consistent, then $\hat{\theta}_T^{(k)}$ is consistent for any k .*

We emphasize here that the contracting mapping condition on $\bar{\theta}(P^0, \cdot)$ is only required when the initial estimator $\hat{\theta}^{(0)}$ is not a consistent estimator. Otherwise, this condition is not needed and in addition, we get the stronger result that $\hat{\theta}_T^{(k)}$ is consistent for any k .

Remark 2. As mentioned in the Introduction, the assumption that $\nu(\cdot)$ is known and does not depend on the sample is not as restrictive as it looks, since we can always incorporate the sample dependence in the objective function. For example, suppose the original objective function is $\tilde{Q}_T(\theta, \nu_T(\theta))$. We can define the new objective function $Q_T(\theta, \nu(\theta)) \equiv \tilde{Q}_T(\theta, \nu_T(\nu(\theta)))$, where $\nu(\theta) = \theta$. Under suitable conditions on $\tilde{Q}_T(\cdot, \cdot)$ and $\nu_T(\cdot)$, Theorem 4.1 still holds.

We now consider the modified algorithms with Step k'. Let $L_\infty(\theta) = Q_\infty(\theta, \nu(\theta))$. Define

$$\bar{\theta}_T(\nu(\theta_1)) = \theta_1 - [G_T(\theta_1, \theta_1)]^{-1} \left[\frac{\partial L_T(\theta_1)}{\partial \theta} \right],$$

where

$$\begin{aligned} G_T(\theta, \theta_1) &= \frac{\partial^2 Q_T(\theta, \nu(\theta_1))}{\partial \theta \partial \theta'} \text{ for Algorithm I,} \\ &= \frac{\partial^2 Q_T(\theta, \nu(\theta_1))}{\partial \theta \partial \theta'} + H_T(\theta, \theta_1) \text{ for Algorithm II.} \end{aligned} \tag{29}$$

Similarly, define

$$\bar{\theta}(P^0, \nu(\theta_1)) = \theta_1 - [G_\infty(\theta_1)]^{-1} \frac{\partial L_\infty(\theta_1)}{\partial \theta},$$

where

$$\begin{aligned} G_\infty(\theta) &= -\Sigma(\theta, \theta) \text{ for Algorithm I,} \\ &= -\Sigma(\theta, \theta) + H(\theta, \theta) \text{ for Algorithm II.} \end{aligned}$$

With these notations, the same proof as that of Theorem 3.3 leads to:

Theorem 3.4 *Assume that Θ is a compact subset of R^p . Let $k = k(T)$. If $\sup_{\theta_1 \in \Theta} \|G_T(\theta_1, \theta_1) - G_\infty(\theta_1)\| \xrightarrow{p} 0$ and $\sup_{\theta_1 \in \Theta} \left\| \frac{\partial L_T(\theta_1)}{\partial \theta} - \frac{\partial L_\infty(\theta_1)}{\partial \theta} \right\| \xrightarrow{p} 0$, then (i) $\hat{\theta}_T^{(k)}$ defined in Step k' of Algorithms I-IV with $k(T) \rightarrow \infty$, is consistent, provided that the mapping $\bar{\theta}(\nu(\theta_1))$ is contracting on Θ ; (ii) If $\hat{\theta}_T^{(0)}$ is a consistent estimator, then $\hat{\theta}_T^{(k)}$ is consistent for any k .*

The following conditions will be used to establish the asymptotic distribution of $\hat{\theta}_T^{(k)}$.

Assumption E1. $\sqrt{T} \frac{\partial L_T(\theta^0)}{\partial \theta} \xrightarrow{d} N_p(0, \Omega(\theta^0))$, with

$$\Omega(\theta^0) = \lim_{T \rightarrow \infty} \text{Var}(\sqrt{T} \frac{\partial L_T}{\partial \theta}(\theta^0)),$$

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 $[\frac{\partial^2 L_\infty(\theta^0)}{\partial \theta \partial \theta'}]^{-1} \Omega(\theta^0) [\frac{\partial^2 L_\infty(\theta^0)}{\partial \theta \partial \theta'}]^{-1}$. Assumptions E1 and E2 include the case discussed earlier where $Q_T(\theta, \nu(\theta)) \equiv \tilde{Q}_T(\theta, \nu_T(\nu(\theta)))$ with $\nu(\theta) = \theta$ (say). In this case, let

$$\tilde{S}_T \left(\theta^0, \nu_T(\theta^0), \frac{\partial \nu_T(\theta^0)}{\partial \theta} \right) \equiv \frac{\partial L_T(\theta^0)}{\partial \theta} = \frac{\partial \tilde{Q}_T(\theta^0, \nu_T(\theta^0))}{\partial \theta} + \frac{\partial \tilde{Q}_T(\theta^0, \nu_T(\theta^0))}{\partial \nu} \frac{\partial \nu_T(\theta^0)}{\partial \theta}.$$

Suppose $p \lim \nu_T(\theta^0) = \nu_\infty(\theta^0)$ and $p \lim \frac{\partial \nu_T(\theta^0)}{\partial \theta} = \frac{\partial \nu_\infty(\theta^0)}{\partial \theta}$. Then ignoring higher order terms, we get

$$\begin{aligned} \tilde{S}_T \left(\theta^0, \nu_T(\theta^0), \frac{\partial \nu_T(\theta^0)}{\partial \theta} \right) &= \tilde{S}_T \left(\theta^0, \nu_\infty(\theta^0), \frac{\partial \nu_\infty(\theta^0)}{\partial \theta} \right) \\ &\quad + \tilde{S}_{T2} \left(\theta^0, \nu_\infty(\theta^0), \frac{\partial \nu_\infty(\theta^0)}{\partial \theta} \right) (\nu_T(\theta^0) - \nu_\infty(\theta^0)) \\ &\quad + \tilde{S}_{T3} \left(\theta^0, \nu_\infty(\theta^0), \frac{\partial \nu_\infty(\theta^0)}{\partial \theta} \right) \left(\frac{\partial \nu_T(\theta^0)}{\partial \theta} - \frac{\partial \nu_\infty(\theta^0)}{\partial \theta} \right), \end{aligned}$$

where \tilde{S}_{T2} and \tilde{S}_{T3} denote respectively the partial derivatives of \tilde{S}_T with respect to its second and third arguments. Typically, the first order asymptotic distributions of $\tilde{S}_T \left(\theta^0, \nu_\infty(\theta^0), \frac{\partial \nu_\infty(\theta^0)}{\partial \theta} \right)$, $(\nu_T(\theta^0) - \nu_\infty(\theta^0))$, and $\left(\frac{\partial \nu_T(\theta^0)}{\partial \theta} - \frac{\partial \nu_\infty(\theta^0)}{\partial \theta} \right)$ are given by some sample means and hence one can collect the three terms on the right hand side of the above equation and verify Assumption E1 by a CLT. Similarly, Assumption E2 is expected to hold under appropriate conditions.

Theorem 3.5 Assume that θ^0 is an interior point of Θ . Under assumptions E1 and E2, if $\hat{\theta}_T^{(k)}$ is consistent and the sequence $k(T)$, $T \geq 1$, is such that $\sqrt{T}(\hat{\theta}_T^{(k)} - \hat{\theta}_T^{(k-1)}) \rightarrow 0$ in probability, then $\hat{\theta}_T^{(k)}$ has the same asymptotic distribution as $\hat{\theta}_T$.

Efficiency of our algorithms does not come without cost. The original backfitting algorithm of PPR requires the evaluation of the function $\nu(\theta)$ only. To improve 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.

Remark 3. All the algorithms and their asymptotic properties apply to objective functions of the form: $Q_T[\theta, \nu(\theta_1)]$ or $Q_T[\theta_1, \nu(\theta)]$, where $\theta = (\theta_1^T, \theta_2^T)^T$. In some cases, θ_2 is such that consistent estimation of θ_1 is feasible at any value of θ_2 . In this case, one possible choice of the initial value for θ_1 is an extension of the backfitting estimator of PPR. That is, we fix a value for θ_2 , $\hat{\theta}_2^{(0)}$ (say), then let $\hat{\theta}_1^{(0)}$ be the backfitting estimator or its modifications proposed in Section 2 of the following optimization problem:

$$\max_{\theta_1 \in \Theta_1} Q_T((\theta_1, \hat{\theta}_2^{(0)}), \nu(\theta_1)) \text{ or } \max_{\theta_1 \in \Theta_1} Q_T(\theta_1, \nu(\theta_1, \hat{\theta}_2^{(0)}))$$

Then under regularity conditions, $\hat{\theta}_1^{(0)}$ is a consistent estimator of θ_1^0 and one can take $\hat{\theta}_T^{(0)}$ as $(\hat{\theta}_1^{(0)}, \hat{\theta}_2^{(0)})$.

3.4 Information dominance

The asymptotic normality result for $\hat{\theta}_T^{(k)}$ in Theorem 3.5 depends on the condition: $\sqrt{T}(\hat{\theta}_T^{(k)} - \hat{\theta}_T^{(k-1)}) \rightarrow 0$ in probability. If the initial estimator is consistent, then this condition is implied by a different information dominance condition for each algorithm. Corresponding to Algorithms I-IV, there are Information Dominance I-IV. The first two are provided in the theorem below and the last two are provided in Appendix B.

Theorem 3.6 Suppose the conditions of Theorem 3.5 except $\sqrt{T}(\hat{\theta}_T^{(k)} - \hat{\theta}_T^{(k-1)}) \rightarrow 0$ in probability hold. If the initial estimator $\hat{\theta}_T^{(0)}$ is consistent, then the condition: $\sqrt{T}(\hat{\theta}_T^{(k)} - \hat{\theta}_T^{(k-1)}) \rightarrow 0$ in probability is implied by the following information dominance conditions as long as $k \rightarrow \infty$ as $T \rightarrow \infty$:

Information Dominance I.

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

Information Dominance II.

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

The information dominance conditions are also closely related to the contraction mapping conditions in Assumption C5. To see this, recall that $\bar{\theta}(P^0, \theta_1)$ satisfies: $S_\infty(\bar{\theta}(P^0, \theta_1), \theta_1) = 0$ implying:

$$\frac{\partial}{\partial \theta} S_\infty(\bar{\theta}(P^0, \theta_1), \theta_1) \frac{\partial \bar{\theta}(P^0, \theta_1)}{\partial \theta'_1} + \frac{\partial}{\partial \theta_1} S_\infty(\bar{\theta}(P^0, \theta_1), \theta_1) = 0,$$

where $\frac{\partial}{\partial \theta} S_\infty(\cdot, \cdot)$ is the partial derivative of $S_\infty(\cdot, \cdot)$ with respect to the first argument and $\frac{\partial}{\partial \theta_1} S_\infty(\cdot, \cdot)$ is the partial derivative of $S_\infty(\cdot, \cdot)$ with respect to the second argument. Thus, Assumption C5 at θ_0 is equivalent to

$$\left\| \left[\frac{\partial}{\partial \theta} S_\infty(\theta_0, \theta_0) \right]^{-1} \frac{\partial}{\partial \theta_1} S_\infty(\theta_0, \theta_0) \right\| < 1. \quad (30)$$

It is easy to show that (30) corresponds to Information Dominance I-IV for Algorithms I-IV respectively. For example, for Algorithm I, we have

$$S_\infty(\theta, \theta_1) = \frac{\partial Q_\infty(\theta, \nu(\theta_1))}{\partial \theta} + \frac{\partial Q_\infty(\theta_1, \nu(\theta_1))}{\partial \nu} \nu'(\theta_1)$$

so that

$$\begin{aligned} \frac{\partial}{\partial \theta} S_\infty(\theta_0, \theta_0) &= -\Sigma(\theta^0, \theta^0), \\ \frac{\partial}{\partial \theta_1} S_\infty(\theta_0, \theta_0) &= \frac{\partial^2 Q_\infty}{\partial \theta \partial \nu'}[\theta_0, \nu(\theta_0)] \nu'(\theta_0) + \frac{\partial Q_\infty(\theta_0, \nu(\theta_0))}{\partial \nu} \nu''(\theta_0) \\ &\quad + \left[\frac{\partial^2 Q_\infty(\theta_0, \nu(\theta_0))}{\partial \nu \partial \theta'} + \frac{\partial^2 Q_\infty(\theta_0, \nu(\theta_0))}{\partial \nu \partial \nu'} \nu'(\theta_0) \right] \nu'(\theta_0) \\ &= D^2L_\infty(\theta^0) - \Sigma(\theta^0, \theta^0). \end{aligned}$$

Heuristically, each information dominance 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 information dominance condition is only imposed locally at θ^0 and there is no requirement on the rate of divergence of $k(T)$; it is only required

to increase to ∞ as T goes to ∞ . If the initial estimator $\hat{\theta}_T^{(0)}$ is not consistent, then the condition: $\sqrt{T}(\hat{\theta}_T^{(k(T))} - \hat{\theta}_T^{(k(T)-1)}) \rightarrow 0$ in probability is implied by the asymptotic contracting mapping condition on $\bar{\theta}_T(\nu(\theta_1))$ and the condition: $k(T) \geq T^\delta$ for a small $\delta > 0$, see the discussion following Proposition 3.2.

4 Iterative Estimation of Structural Credit Risk Models

Structural credit-risk models, also dubbed firm-value models, characterize the occurrence of default by the fact that firm value falls below a threshold representing liabilities (see e.g., McNeil, Frey and Embrechts (2005) for a general exposition). The key quantity of interest, the so-called expected default frequency (EDF) is the probability that the firm value V_h , at some given horizon h , falls below the liabilities threshold, given that the firm value today is V_0 . This approach raises two statistical issues. First, computing the theoretical EDF amounts to the computation of a conditional probability distribution of V_h given V_0 . This can be done from a specified parametric model of this distribution. An alternative route is a nonparametric assessment of an empirical EDF. These EDF computations will not be detailed here. Our focus is on the second statistical issue, namely the filtering of the current firm value V_0 , which is in any case a key input for computing an empirical EDF. It turns out that the firm market value V_0 is not something one can easily observe. It must generally be indirectly inferred from the observed value S_0 of firm's equity.

4.1 Merton Model

Following Merton's (1974) seminal firm value credit risk model, firm's market value V_t at time t and firm's equity price S_t at the same date are in a one-to-one correspondence through an option pricing model. As a template, let us consider the simplest case put forward by Merton (1974) in which the firm's debt consists of a zero-coupon bond with face value B and maturity τ . Then firm's equity can be interpreted as an European call option written on firm's value with strike price B and exercise date τ . To see this, it is sufficient to realize that firm's equity price at the maturity date τ of the debt must be:

$$S_\tau = \text{Max}[V_\tau - B, 0].$$

Consequently, with frictionless markets, the firm's equity value S_t at any time $t, 0 \leq t \leq \tau$, should be the market price of this call option at time t . Let us assume for simplicity that the process (V_t) follows a univariate diffusion and that the risk-free interest rate is deterministic and equal to $r \geq 0$. The great advantage of a univariate diffusion model with deterministic interest rate is the completeness of markets which allows us to characterize the risk-neutral dynamics of the firm's market value through a diffusion equation with a known drift:

$$\frac{dV_t}{V_t} = rdt + \sigma(V_t, \nu)dW_t^Q \quad (31)$$

where (W_t^Q) is a standard Brownian motion for the risk neutral probability measure Q . For the purpose of statistical inference, we specify the volatility of the process (V_t) as a known function $\sigma(V_t, \cdot)$ of some vector ν of unknown parameters. Then, standard risk-neutral valuation provides the key relationship between firm's market value V_t and firm's equity price S_t :

$$S_t = \exp[-r(\tau - t)]E_t^Q\{Max(0, V_\tau - B)\}, \quad (32)$$

where the conditional risk-neutral expectation operator $E_t^Q\{\cdot\}$ is defined by the conditional probability distribution of V_τ given V_t as characterized by (31). In other words, after computation of a conditional expectation, we can rewrite (32) as a structural relationship:

$$S_t = g(V_t, \nu), \quad (33)$$

where $g(\cdot, \cdot)$ is a known function, at least theoretically, up to computational issues. Note that any standard option pricing model will deliver a function which is strictly increasing with respect to its first occurrence, allowing us to rewrite (33) equivalently as, with obvious notations:

$$V_t = g^{-1}(S_t, \nu). \quad (34)$$

The filtering problem of latent firm's market value V_t from observed equity price S_t would thus be trivial, up to computational issues, if the pricing parameters ν were known. In general they are not and an estimation step is needed. One natural way to address the estimation issue is to bridge the gap between historical and risk-neutral probability distributions, by specifying that the dynamics of the firm's value process (V_t) under the real-world or historical probability measure P is given by a parametric diffusion model:

$$\frac{dV_t}{V_t} = \mu(V_t, \alpha)dt + \sigma(V_t, \theta)dW_t^P \quad (35)$$

where (W_t^P) is a standard Brownian motion under P . Note that by a standard Girsanov argument, the two equivalent probability distributions P and Q must have the same diffusion coefficient and thus:

$$\nu = \nu(\theta) = \theta.$$

It is however worth denoting differently on the one hand the occurrence of θ in the historical distribution (35) of (V_t) and on the other hand the occurrence of ν in the option pricing formula (33). Up to computational issues, the diffusion equation (35) allows us to write the conditional likelihood of a sample path $(V_{t_1}, V_{t_2}, \dots, V_{t_n})$ given some initial value V_{t_0} and a value (α, θ) of unknown historical parameters. Note that to deal with a finite sample of values drawn from a continuous time process (V_t) , a slight change in notation is convenient to avoid any a priori constraint about the length of time between consecutive observations. In sample dates are now denoted as $t_j, j = 0, 1, \dots, n$.

However, firm's market values V_{t_j} are not directly observed but recovered from the inverse option pricing formula $V_{t_j}[\nu(\theta)] = g^{-1}[S_{t_j}, \nu(\theta)]$. A simple application of the Jacobian formula for change in variables then provides the conditional log-likelihood function, denoted as $Q_n^*[\theta, \nu(\theta), \alpha]$. When concentrating out the drift parameters α , we get a profile objective function $Q_n[\theta, \nu(\theta)]$ the maximization of which would provide an efficient estimator of θ .

An example illustrating the main issues can actually be obtained from the simplest Black and Scholes option pricing model, that is, an historical distribution defined as a geometric Brownian motion:

$$\frac{dV_t}{V_t} = \mu dt + \sigma dW_t^P.$$

In other words, $\mu(V_t, \alpha) = \alpha = \mu$ and $\sigma(V_t, \theta) = \sqrt{\theta}$ (with $\theta = \sigma^2$) are viewed as constant parameters. Therefore, adopting notations from Duan, Gauthier, and Simonato (2004) and considering in particular equally spaced sample points ($t_j = jh$ for $j = 0, 1, \dots, n$), we can write the conditional log-likelihood function for the unobserved asset values as:

$$L^V(\mu, \sigma^2; V_h, \dots, V_{nh} | V_0) = -\frac{n}{2} \ln(2\pi\sigma^2 h) - \frac{1}{2} \sum_{j=1}^n \frac{\left(R_j - \left(\mu - \frac{\sigma^2}{2}\right)h\right)^2}{\sigma^2 h} - \sum_{j=1}^n \ln V_{jh},$$

where $R_j = \ln(V_{jh}/V_{(j-1)h})$ is the asset continuously compounded rate of return over the time interval $[t_{j-1}, t_j]$. The observable equity values $S_0, S_h, S_{2h}, \dots, S_{nh}$ are related to the asset values via the Black and Scholes option pricing formula:

$$S_t = V_t \Phi(d_t) - B e^{-r(\tau-t)} \Phi(d_t - \sigma \sqrt{\tau-t}) \equiv g(V_t; \sigma^2)$$

in which

$$d_t = \frac{\ln(V_t/B) + (r + \frac{1}{2}\sigma^2)(\tau - t)}{\sigma\sqrt{\tau - t}}.$$

The conditional log-likelihood function for the observables is given by

$$\begin{aligned} L^S(\mu, \sigma^2; \nu(\sigma^2); S_h, \dots, S_{nh} | S_0) &= nQ_n^*[\sigma^2, \nu(\sigma^2), \mu]. \\ &= L^V[\mu, \sigma^2; V_h(\nu(\sigma^2)), \dots, V_{nh}(\nu(\sigma^2)) | V_0(\nu(\sigma^2))] - \sum_{j=1}^n \ln[\Phi(d_{jh}(\nu(\sigma^2))], \end{aligned} \quad (36)$$

where $V_{jh}(\nu(\sigma^2)) = g^{-1}(S_{jh}; \nu(\sigma^2))$ with $\nu(\sigma^2) = \sigma^2$ and

$$d_{jh}(\nu(\sigma^2)) = \frac{\ln(V_{jh}(\nu(\sigma^2))/B) + (r + \frac{1}{2}\nu(\sigma^2))(\tau - jh)}{\sqrt{\nu(\sigma^2)}\sqrt{\tau - jh}}.$$

Writing out the complete expression for $L^S(\mu, \sigma^2; \nu(\sigma^2); S_h, \dots, S_{nh} | S_0)$, we have

$$\begin{aligned} &nQ_n^*[\sigma^2, \nu(\sigma^2), \mu] \\ &= -\frac{n}{2} \ln(2\pi\sigma^2 h) - \frac{1}{2} \sum_{j=1}^n \frac{\left(R_j(\nu(\sigma^2)) - (\mu - \frac{\sigma^2}{2})h\right)^2}{\sigma^2 h} - \sum_{j=1}^n \ln V_{jh}(\nu(\sigma^2)) \\ &\quad - \sum_{j=1}^n \ln \Phi \left(\frac{\ln(V_{jh}(\nu(\sigma^2))/F) + (r + \frac{1}{2}\nu(\sigma^2))(\tau - jh)}{\sqrt{\nu(\sigma^2)}\sqrt{\tau - jh}} \right), \end{aligned} \quad (37)$$

where $R_j(\nu(\sigma^2)) = \ln(V_{jh}(\nu(\sigma^2))/V_{(j-1)h}(\nu(\sigma^2)))$.

4.2 On the Non-Equivalence of PPR and Maximum Likelihood: the KMV Iterative Technique

This section revisits a popular application of a PPR kind of algorithm in the field of credit risk modeling. In this application, the algorithm is usually introduced by the name of KMV, borrowing the name of the private company founded by Kealhofer, McQuown and Vacicek which has developed this algorithm. Interestingly enough, in the simple context of a Black and Scholes option pricing model, this algorithm displays all the characteristics described above: preliminary concentration step with respect to the irrelevant drift parameter α , complicated resulting efficient criterion to characterize full maximum likelihood, user-friendly PPR kind of iteration which comes with the price of efficiency loss. To the best of our knowledge, the fact that the KMV iterative procedure bears the efficiency loss of PPR with respect to full maximum likelihood has not been realized yet in the credit risk

literature. As explained below, the pitfall comes from the fact that in some contexts of models with latent variables, PPR amounts to an EM (Expectation-Maximization) algorithm, but precisely in a situation where the EM theory does not deliver the full MLE.

It is interesting to note that the volatility parameter σ appears in (37) in several places, while the expected return parameter μ appears in (37) only once. Therefore, there are two different ways to apply the backfitting algorithm from this point. We can either deal with μ together with σ^2 in backfitting or concentrate out μ before applying backfitting. For this example, it turns out that the second approach is easier than the first, because it is simple to concentrate out μ . We thus follow this approach. We have straightforwardly:

$$\mu_n(\nu(\sigma^2)) = \frac{1}{n} \sum_{j=1}^n R_j(\nu(\sigma^2)) + \frac{\sigma^2}{2}. \quad (38)$$

Then, if we define $\bar{R}_n(\nu(\sigma^2)) = \frac{1}{n} \sum_{j=1}^n R_j(\nu(\sigma^2))$, the concentrated log-likelihood takes the following form:

$$\begin{aligned} & nQ_n[\sigma^2, \nu(\sigma^2)] \quad (39) \\ = & -\frac{n}{2} \ln(2\pi\sigma^2 h) - \frac{1}{2} \sum_{j=1}^n \frac{[R_j(\nu(\sigma^2)) - \bar{R}_n(\nu(\sigma^2))]^2}{\sigma^2 h} - \sum_{j=1}^n \ln V_{jh}(\nu(\sigma^2)) \\ & - \sum_{j=1}^n \ln \Phi \left(\frac{\ln(V_{jh}(\nu(\sigma^2))/B) + (r + \frac{1}{2}\nu(\sigma^2))(\tau - jh)}{\sqrt{\nu(\sigma^2)}\sqrt{\tau - jh}} \right). \quad (40) \end{aligned}$$

From (7), it follows that the backfitting estimator $\hat{\sigma}_n^{(k)}$ solves the following maximization problem:

$$(\hat{\sigma}_n^{(k)})^2 = \arg \max_{\sigma} Q_n[\sigma^2, \nu((\hat{\sigma}_n^{(k-1)})^2)].$$

The solution solves the first order condition:

$$0 = -\frac{n}{2\sigma^2} + \frac{1}{2h\sigma^4} \sum_{j=1}^n \left[R_j(\nu((\hat{\sigma}_n^{(k-1)})^2)) - \bar{R}_n(\nu((\hat{\sigma}_n^{(k-1)})^2)) \right]^2$$

and is given by

$$(\hat{\sigma}_n^{(k)})^2 = \frac{1}{nh} \sum_{j=1}^n [R_j(\nu(\hat{\sigma}_n^{(k-1)})) - \bar{R}_n(\nu(\hat{\sigma}_n^{(k-1)}))]^2. \quad (41)$$

We recognize immediately that the KMV algorithm is exactly the backfitting algorithm of PPR. This actually illustrates the computational advantages of PPR backfitting with respect to full MLE. The occurrences of the unknown parameters θ in the criterion function

$Q_n[\theta, \nu(\theta)]$ are rather nasty when they come through the function $\nu(\theta)$ because they involve the inversion of the option pricing formula. This is the reason why we may prefer to keep them fixed at the maximization stage and just maximizing with respect to the first occurrences. The KMV model is rather striking in this respect. Even though the Black and Scholes option pricing formula is the simplest possible one may imagine, nobody wants to maximize directly (39) and it is much more pleasant to compute a sequence of empirical variances as in (41) above. Once one has obtained the limit $\hat{\sigma}_{B,n}^2$ of this sequence (with possibly an earlier stopping rule as considered in section 2), it is time to go to the inversion of the option pricing formula to get the filtered value of the initial firm's market value:

$$\hat{V}_0 = g^{-1}(S_0, \hat{\sigma}_{B,n}^2).$$

This is exactly what is done by the KMV approach: all along the recursion, the inversion step (to compute the implied returns $R_j(\nu((\hat{\sigma}_n^{(k-1)})^2))$) is disentangled from the maximization step (to deduce $\hat{\sigma}_n^{(k)}$). While Pastorello, Patilea and Renault (2003) had noticed that generally speaking, the PPR algorithm in such a setting of a one-to-one relationship $S_t = g(V_t, \nu)$ between latent variables and observables is tantamount to an EM algorithm (where the expectation step is akin to the inversion step to compute the implied returns), Duan, Gauthier, and Simonato (2004) have recently rediscovered this result in the particular case of the KMV approach, which, as explained above, is nested in PPR. But Duan, Gauthier, and Simonato (2004) wrongly deduce from this remark that KMV coincides with maximum likelihood estimation. This conclusion is erroneous, simply because, as already stressed by PPR, the standard EM theory does not apply in this setting. The deep explanation of this failure is that the existence of a one to one relationship $S_t = g(V_t, \nu)$ between latent variables and observables implies that the support of the latent variable V_t given the observation S_t is simply the singleton $g^{-1}(S_t, \nu)$ (making degenerate the expectation step) and thus depends on the unknown parameters $\nu = \theta$.

Indeed KMV is on the contrary a striking counter-example to show that PPR is not in general as efficient as full MLE. To see this, we have just to apply to the KMV algorithm the general asymptotic theory of the backfitting algorithm established in PPR. Let us first check that the required conditions for application of the general PPR theory hold in the KMV example. It is easy to see that:

$$Q_\infty(\sigma^2, \nu(\sigma^2)) = -\frac{1}{2} \ln(2\pi\sigma^2 h) - \frac{1}{2\sigma^2 h} \text{Var} [R_j(\nu(\sigma^2))] - E [\ln V_{jh}(\nu(\sigma^2))] \\ - E \left[\ln \Phi \left(\frac{\ln(V_{jh}(\nu(\sigma^2))/B) + (r + \frac{1}{2}\nu(\sigma^2))(\tau - jh)}{\sqrt{\nu(\sigma^2)}\sqrt{\tau - jh}} \right) \right].$$

Hence:

$$\frac{\partial Q_\infty(\sigma^2, \nu(\sigma^2))}{\partial \sigma^2} = -\frac{1}{2\sigma^2} + \frac{1}{2h\sigma^4} \text{Var} [R_j(\nu(\sigma^2))]. \quad (42)$$

The root of the above function must satisfy:

$$\sigma^2 = \frac{1}{h} \text{Var} [R_j(\nu(\sigma^2))]. \quad (43)$$

Obviously the true value of σ^2 satisfies (43). But the PPR identification condition requires that there exists a unique solution to (43).

To gain insight into the PPR identification condition and the contraction mapping condition, for any $\sigma_1^2 > 0$, define

$$\bar{\sigma}^2(\sigma_1^2) = \frac{1}{h} \text{Var} [R_j(\nu(\sigma_1^2))].$$

Then PPR assume that the function $\bar{\sigma}^2(\sigma_1^2)$ is differentiable and contracting so that

$$\left| \frac{\partial \bar{\sigma}^2(\sigma_1^2)}{\partial \sigma_1^2} \right| < 1.$$

To see why this property is fulfilled, it is worth realizing that $\frac{1}{h} \text{Var} [R_j(\nu(\sigma_1^2))]$ is actually a function $V[\sigma^2, \sigma_1^2]$. It depends on the true volatility parameter σ which defines the Data Generating Process (DGP) for the underlying geometric Brownian motion (V_t). The required contraction property is:

$$\left| \frac{\partial V(\sigma^2, \sigma_1^2)}{\partial \sigma_1^2} \Big|_{\sigma_1^2 = \sigma^2} \right| < 1.$$

But by definition, we have:

$$V(\sigma^2, \sigma^2) = \sigma^2$$

and thus for all σ :

$$\frac{\partial V(\sigma^2, \sigma_1^2)}{\partial \sigma_1^2} \Big|_{\sigma_1^2 = \sigma^2} + \frac{\partial V(\sigma^2, \sigma_1^2)}{\partial \sigma^2} \Big|_{\sigma_1^2 = \sigma^2} = 1.$$

Therefore, a sufficient condition to ensure the required contraction is to have the conjunction of the two following inequalities:

$$\frac{\partial V(\sigma^2, \sigma_1^2)}{\partial \sigma_1^2} \Big|_{\sigma_1^2 = \sigma^2} \geq 0$$

and

$$\frac{\partial V(\sigma^2, \sigma_1^2)}{\partial \sigma^2} \Big|_{\sigma_1^2 = \sigma^2} > 0.$$

In other words, the PPR contraction property amounts to say:

(i) For a given DGP of the underlying returns, the higher our guess σ_1 on volatility is to compute implied returns $R_j(\nu(\sigma_1^2))$, the more volatile these implied returns are;

(ii) For a given guess σ_1 on volatility, the more volatile the true underlying return process is, the more volatile the implied returns $R_j(\nu(\sigma_1^2))$ are.

To put it differently: volatility of implied returns is an increasing function of the volatility of the true DGP but it does not overreact: the volatility of implied returns cannot increase faster than the true volatility. This property is intuitively obvious and more extensively discussed in the appendix.

Therefore the asymptotic theory of PPR applies and the asymptotic distribution of the KMV estimator can be derived from Theorem 2.1 or Theorem 2.2 in Section 2. More specifically, one can easily verify from (42) the following expressions:

$$\begin{aligned} -\Sigma(\sigma^2, \sigma_1^2) &= \frac{1}{2\sigma^4} - \frac{1}{h\sigma^6} \text{Var}[R_j(\nu(\sigma_1^2))] \\ H(\sigma^2, \sigma_1^2) &= \frac{1}{2h\sigma^4} \frac{\partial \text{Var}[R_j(\nu(\sigma_1^2))]}{\partial \nu}. \end{aligned}$$

In addition, some algebra shows that the variance matrix $B(\sigma^2)$ in Assumption B2 takes the following form:

$$B(\sigma^2) = \frac{1}{4h^2\sigma^8} \text{Var}\{[R_j - (\mu - \frac{\sigma^2}{2})h]^2\} = \frac{1}{2\sigma^4}$$

Noting that $\Sigma(\sigma^2, \sigma^2) = \frac{1}{2\sigma^4}$, we conclude that the KMV estimator of the volatility parameter is root- n normally distributed with variance given by

$$\omega(\sigma^2) = \frac{1}{2\sigma^4} \left[1 - \frac{\partial V(\sigma^2, \sigma_1^2)}{\partial \sigma_1^2} \Big|_{\sigma_1^2 = \sigma^2} \right]^{-2}.$$

Note that by straightforward application of the delta theorem, the asymptotic variance of the quantity of interest, namely the estimated firm's market value \widehat{V}_0 , is proportional to the variance $\omega(\sigma^2)$ of the KMV estimator of σ^2 . Several remarks are in order to understand what makes this estimator more or less accurate. First we observe that an efficiency bound that cannot be reached is the variance of $[R_{jh} - (\mu - \frac{\sigma^2}{2})h]^2$ (scaled by $1/h^2$), that is $\frac{1}{2\sigma^4}$. This is the variance of the estimator that we would have obtained from direct observations

of the latent firm's market value process (V_t). The factor $\left[1 - \frac{\partial V(\sigma^2, \sigma_1^2)}{\partial \sigma_1^2} \Big|_{\sigma_1^2 = \sigma^2}\right]^{-2}$ is the price to pay for having been obliged to iteratively estimate underlying returns from a sequence of guesses ($\widehat{\sigma}_n^{(k)}$) of the true volatility. More elastic to errors in these guesses the volatility of implied returns is, higher is the price to pay.

However, the key issue is to know to what extent the KMV estimator exploits efficiently the information available, that is observation of firm's equity (S_t) instead of firm's market value (V_t). Duan, Gauthier, and Simonato (2004) argue via EM algorithm that the KMV algorithm is asymptotically equivalent to MLE and hence is asymptotically efficient. Is it? To answer this question, we look at the full score function corresponding to the concentrated log-likelihood function:

$$\frac{dL_C^S(\sigma^2; \nu(\sigma^2); S_h, \dots, S_{nh} \mid S_0)}{d\sigma^2} = \left\{ -\frac{n}{2\sigma^2} + \frac{1}{2h\sigma^4} \sum_{j=1}^n [R_{jh}(\nu(\sigma^2)) - \bar{R}_n(\nu(\sigma^2))]^2 \right\} - s_2(\sigma^2),$$

where

$$s_2(\sigma^2) = -\sum_{j=0}^n \frac{\partial L_C^V(\sigma^2; V_h(\nu(\sigma^2)), \dots, V_{nh}(\nu(\sigma^2)) \mid S_0)}{\partial \nu} + \sum_{j=1}^n \frac{\partial \ln(\Phi(d_{jh}(\nu(\sigma^2))))}{\partial \nu},$$

in which

$$\begin{aligned} & L_C^V(\sigma^2; V_h(\nu(\sigma^2)), \dots, V_{nh}(\nu(\sigma^2)) \mid S_0) \\ &= -\frac{n}{2} \ln(2\pi\sigma^2 h) - \frac{1}{2} \sum_{j=1}^n \frac{(R_{jh}(\nu(\sigma^2)) - \bar{R}_n(\nu(\sigma^2)))^2}{\sigma^2 h} - \sum_{j=1}^n \ln V_{jh}(\nu(\sigma^2)). \end{aligned}$$

Since $V_{jh}(\nu(\sigma^2)) = g^{-1}(S_{jh}; \nu(\sigma^2))$ comes from inversion of the Black and Scholes option pricing formula, there is no reason to expect that $s_2(\widehat{\sigma}_n^2) = 0$ at MLE. As a result, the backfitting algorithms including the KMV algorithm are not efficient in general. To put it in another way, the (erroneous) belief in the efficiency of the KMV algorithm is, as explained in Section 1, tantamount to the belief that the function $\nu(\sigma^2) = \sigma^2$ corresponds, at least asymptotically, to genuine concentration of the likelihood function with respect to ν . Let us have a look at the asymptotic (concentrated with respect to the drift parameter) log-likelihood function $Q_\infty(\sigma^2, \nu)$ given above. There is no way to imagine that, for all σ^2 , $\frac{\partial Q_\infty(\sigma^2, \nu)}{\partial \nu}$ is zero precisely for $\nu = \sigma^2$.

4.3 Efficient Estimation of Merton Model

All four efficient algorithms presented in Section 3 and Appendix B can be applied to compute MLE at convergence. Algorithm I turns out to be the most convenient for this example. In fact, one can easily show that Step K for σ^2 in Algorithm I is equivalent to solving

$$\sigma^4 s_2(\sigma^{2(k-1)}) + \frac{n\sigma^2}{2} - \frac{1}{2h} \sum_{j=1}^n (R_j^{(k-1)} - \bar{R}^{(k-1)})^2 = 0, \quad (44)$$

which has a closed-form solution:

$$\sigma^{2(k)} = \frac{-\frac{n}{2} + \sqrt{\frac{n^2}{4} + 4s_2(\sigma^{2(k-1)}) \sum_{j=1}^n (R_j^{(k-1)} - \bar{R}^{(k-1)})^2}}{2s_2(\sigma^{2(k-1)})}.$$

The updating rule for μ is:

$$\mu^{(k)} = \frac{1}{h} \bar{R}^{(k-1)} + \frac{\sigma^{2(k)}}{2}.$$

Obviously, KMV or PPR ignores the first term on the left hand side of (44), while our algorithm makes use of it leading to an asymptotically efficient estimator upon convergence.

4.4 Numerical Results

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-Raphson version of the efficient algorithms described in the previous subsections and in Appendix B. The results of the first experiment are detailed in table 2, and they are based on 5,000 synthetic samples of 500 time series observations of daily returns. We directly focused on the concentrated loglikelihood function, and set the single parameter σ^2 at 0.09. The iterations of the inefficient KMV/PPR estimator were started at the true parameter value; the resulting estimates 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 algorithm used.

The results in table 2 show that the two 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. However, there are important differences in the behavior of the four efficient algorithms. Only algorithms I and IV converged to the MLE in every replication; algorithms II and III converged to the MLE only in the 89.92% and 16.72%, respectively, of the replications. The former two algorithms needed on average slightly less than 17 iterations, whereas the latter two needed roughly 44 and 29 iteration, respectively. Apparently, the contracting behavior of the four algorithms may be significantly different. Notice that these success rates should be intended as lower bounds, because in real applications we could try several starting point, which is not possible in a Monte Carlo experiment due to computational constraints. Nevertheless, it seems clear that the requirements imposed by the Information Dominance conditions differ across algorithms.

5 Implied States GMM

5.1 Background

In this section, we assume that the vector θ of the parameters of interest is defined by the law of motion of, say, a stationary and ergodic process Y_t^* . However, the components of Y_t^* are considered as state variables that are not directly observed. The observations, denoted by Y_t , are defined through a known function $Y_t = g(Y_t^*, \nu(\theta^0))$ of latent variables Y_t^* and the known function $\nu(\theta^0)$ of true unknown parameters values as well. Moreover, we assume

that the state variables have been defined such that the mapping $g(\cdot, \nu(\theta))$ is one-to-one for any $\theta \in \Theta$:

$$Y_t = g(Y_t^*, \nu(\theta)) \iff Y_t^* = g^{-1}(Y_t, \nu(\theta)) \quad (45)$$

Statistical identification of θ from observations (Y_t) will then come from the joint knowledge of this one-to-one mapping and of a set of H moment conditions in the latent world. More precisely, following Hansen (1982), we consider a H -dimensional function $\psi(Y_t^*, \theta^0), \theta \in \Theta \subset \mathbb{R}^p$ such that:

$$E\psi(Y_t^*, \theta) = 0 \iff \theta = \theta^0 \quad (46)$$

Statistical identification of θ must then come from the observable moment conditions:

$$E\phi[Y_t, \theta, \nu(\theta)] = 0$$

where:

$$\phi[Y_t, \theta, \nu(\theta)] = \psi[g^{-1}(Y_t, \nu(\theta)), \theta]. \quad (47)$$

Remark.

By contrast with common models with incomplete data which have motivated the vast literature on data augmentation algorithms, including the EM algorithm, we are in a very particular setting where latent and observable data are one-to-one related, albeit through a function which depends on unknown parameters. Therefore, as extensively discussed in PPR, there is no universally valid argument to explain why one should prefer the latent "data" to the observable ones. As a toy example, let us assume that the (Y_t^*) are i.i.d real random variables, with , for some known numbers a and b , $E(Y_t^*) = a\theta^0$ and $Y_t = Y_t^* - b\theta^0$. Then θ is identifiable from latent variables (Y_t^*) (resp. from observable variables (Y_t)) if and only if $a \neq 0$ (resp $b \neq a$). When identification is granted, the variance of the efficient estimator of θ based on the latent variables (Y_t^*) (resp. on observable variables (Y_t)) is inversely proportional to a^2 (resp. $(a-b)^2$). Therefore, there is no such thing as an ubiquitous information loss when going from latent world to observable one. However, we will in general advocate some economic argument to imagine that it is the case. In this toy example, we will typically set the focus on cases like $a > b > 0$.

Then, while it not necessarily implied by the identification condition (46) in the latent world, we maintain the assumption that θ^0 is also identified by the observable moment

conditions:

$$E\phi[Y_t, \theta, \nu(\theta)] = 0 \iff \theta = \theta^0 \quad (48)$$

For a given weight matrix W , define

$$Q_T(\theta, \nu(\theta)) = -\bar{\phi}'_T(\theta, \nu(\theta))W\bar{\phi}_T(\theta, \nu(\theta)), \quad (49)$$

where $\bar{\phi}_T(\theta, \nu) = T^{-1} \sum_{t=1}^T \phi(Y_t, \theta, \nu)$. Then, following Pan (2002), the IS-GMM estimator $\hat{\theta}_T^{IS}$ is defined as

$$\hat{\theta}_T^{IS} = \arg \max_{\theta \in \Theta} Q_T(\theta, \nu(\theta)). \quad (50)$$

The motivation for considering this estimator in Pan (2002) paper was the following. Latent moment conditions (46) characterize the dynamics of a latent stochastic spot volatility process. Moreover, we have option price data and an option pricing model that specifies a one-to-one relationship (45) between option prices Y_t and latent spot volatility Y_t^* . Then the GMM estimator (50) matches the latent moments by using the "implied state variables" $Y_t^* = g^{-1}(Y_t, \nu(\theta))$. The standard asymptotic distributional theory of GMM applies in this context when the observable moment conditions fulfill the local identification condition:

$$E \left[\frac{\partial \phi(\theta, \nu(\theta))}{\partial \theta'} \right]_{\theta=\theta^0} + E \left[\frac{\partial \phi(\theta, \nu(\theta))}{\partial \nu'} \cdot \frac{\partial \nu(\theta)}{\partial \theta'} \right]_{\theta=\theta^0} \text{ is of rank } p \quad (51)$$

Under this maintained assumption, we know that the optimal choice of the weighting matrix W is (a consistent estimator of) the inverse of the asymptotic variance matrix of $\sqrt{T} \bar{\phi}_T(\theta^0, \nu(\theta^0))$. We assume throughout that this efficient choice of W has been made (possibly based on a first step consistent estimator of θ) to define the criterion function (49).

Let

$$\Gamma(\theta) = \left[\frac{\partial \bar{\phi}_T(\theta, \nu(\theta))}{\partial \theta'} + \frac{\partial \bar{\phi}_T(\theta, \nu(\theta))}{\partial \nu'} \frac{\partial \nu(\theta)}{\partial \theta'} \right]$$

Then, the first order condition for the IS-GMM is:

$$\Gamma'(\theta)W\bar{\phi}_T(\theta, \nu(\theta)) = 0. \quad (52)$$

5.2 The PPR version of IS-GMM

We consider the PPR estimator defined from the iteration:

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

From the above definition of the objective function $Q_T(\theta, \nu(\theta))$, the PPR estimator is then defined by the first order conditions:

$$\frac{\partial \bar{\phi}'_T(\theta^{(k)}, \nu(\theta^{(k-1)}))}{\partial \theta} W \bar{\phi}_T(\theta^{(k)}, \nu(\theta^{(k-1)})) = 0$$

5.3 Algorithms I-IV

The algorithms introduced in the previous section can be used to compute IS-GMM $\hat{\theta}_T^{IS}$. For simplicity, we present Step k only. The updating rules for Algorithms I-IV are respectively

$$\begin{aligned} \frac{\partial \bar{\phi}'_T(\theta_I^{(k)}, \nu(\theta_I^{(k)}))}{\partial \theta} W \bar{\phi}_T(\theta_I^{(k)}, \nu(\theta_I^{(k)})) &= - \frac{\partial \bar{\phi}'_T(\theta_I^{(k-1)}, \nu(\theta_I^{(k-1)}))}{\partial \nu} \frac{\partial \nu(\theta_I^{(k-1)})}{\partial \theta} W \bar{\phi}_T(\theta_I^{(k-1)}, \nu(\theta_I^{(k-1)})); \\ \frac{\partial \bar{\phi}'_T(\theta_{II}^{(k)}, \nu(\theta_{II}^{(k-1)}))}{\partial \theta} W \bar{\phi}_T(\theta_{II}^{(k)}, \nu(\theta_{II}^{(k-1)})) &= - \frac{\partial \bar{\phi}'_T(\theta_{II}^{(k-1)}, \nu(\theta_{II}^{(k-1)}))}{\partial \nu} \frac{\partial \nu(\theta_{II}^{(k-1)})}{\partial \theta} W \bar{\phi}_T(\theta_{II}^{(k-1)}, \nu(\theta_{II}^{(k-1)})); \end{aligned} \quad (53)$$

$$\frac{\partial \bar{\phi}'_T(\theta_{III}^{(k)}, \nu(\theta_{III}^{(k)}))}{\partial \theta} W \bar{\phi}_T(\theta_{III}^{(k)}, \nu(\theta_{III}^{(k)})) = - \frac{\partial \bar{\phi}'_T(\theta_{III}^{(k)}, \nu(\theta_{III}^{(k-1)}))}{\partial \nu} \frac{\partial \nu(\theta_{III}^{(k-1)})}{\partial \theta} W \bar{\phi}_T(\theta_{III}^{(k)}, \nu(\theta_{III}^{(k-1)}));$$

$$\frac{\partial \bar{\phi}'_T(\theta_{IV}^{(k)}, \nu(\theta_{IV}^{(k-1)}))}{\partial \theta} W \bar{\phi}_T(\theta_{IV}^{(k)}, \nu(\theta_{IV}^{(k-1)})) = - \frac{\partial \bar{\phi}'_T(\theta_I^{(k)}, \nu(\theta_I^{(k-1)}))}{\partial \nu} \frac{\partial \nu(\theta_I^{(k-1)})}{\partial \theta} W \bar{\phi}_T(\theta_I^{(k)}, \nu(\theta_I^{(k-1)})).$$

The updating rule for the IS-GMM backfitting estimator of PPR is

$$\frac{\partial \bar{\phi}'_T(\hat{\theta}^{(p)}, \nu(\hat{\theta}^{(p-1)}))}{\partial \theta} W \bar{\phi}_T(\hat{\theta}^{(p)}, \nu(\hat{\theta}^{(p-1)})) = 0 \quad (54)$$

Comparing (53) with (54) indicates that the IS-GMM backfitting estimator of PPR ignores the right hand side of (53) and hence part of the first order condition for the IS-GMM $\hat{\theta}_T^{IS}$. If the model is exactly identified, the IS-GMM $\hat{\theta}_T^{IS}$ satisfies

$$\bar{\phi}_T(\hat{\theta}_T^{IS}, \nu(\hat{\theta}_T^{IS})) = 0$$

and in this case, the IS-GMM backfitting estimator of PPR satisfies $\bar{\phi}_T(\hat{\theta}^{(p)}, \nu(\hat{\theta}^{(p-1)})) = 0$ and hence is asymptotically efficient. In general, Algorithms I-IV are asymptotically efficient.

5.4 Algorithm V

The first order condition for the IS-GMM (52) suggests another algorithm, Algorithm V, for computing $\hat{\theta}_T^{IS}$, i.e.,

$$\Gamma'(\hat{\theta}_V^{(k-1)})W\bar{\phi}_T(\hat{\theta}_V^{(k)}, \nu(\hat{\theta}_V^{(k-1)})) = 0, \quad (55)$$

or equivalently

$$\hat{\theta}_V^{(k)} = \hat{\theta}_V^{(k-1)} - \left[\Gamma'(\hat{\theta}_V^{(k-1)})W\frac{\partial \bar{\phi}_T}{\partial \theta}(\hat{\theta}_V^{(k-1)}, \nu(\hat{\theta}_V^{(k-1)})) \right]^{-1} \Gamma'(\hat{\theta}_V^{(k-1)})W\bar{\phi}_T(\hat{\theta}_V^{(k-1)}, \nu(\hat{\theta}_V^{(k-1)})). \quad (56)$$

If the model is exactly identified, then Algorithm V reduces to the IS-GMM backfitting algorithm of PPR. We now show that in the over-identified case, Algorithm V remains to be asymptotically efficient.

6 Appendix A. Technical Proofs

Proof of Theorem 3.3: (i) follows from Proposition 2 in PPR.

(ii) Let $\bar{\theta}_T(\theta_1)$ denote the minimizer of $\|S_T(\theta, \theta_1)\|$ over $\theta \in \Theta$. Proposition 1 in PPR implies that $\sup_{\theta_1 \in \Theta} \|\bar{\theta}_T(\theta_1) - \bar{\theta}(P^0, \theta_1)\| \xrightarrow{p} 0$. Hence

$$|\hat{\theta}_T^{(1)} - \theta^0| \leq |\bar{\theta}_T(\hat{\theta}_T^{(0)}) - \bar{\theta}(P^0, \hat{\theta}_T^{(0)})| + |\bar{\theta}(P^0, \hat{\theta}_T^{(0)}) - \bar{\theta}(P^0, \theta^0)| \xrightarrow{p} 0.$$

Similarly, one can show the consistency of $\hat{\theta}_T^{(k)}$ for any k .

□

Proof of Theorem 3.4: It is the same as that of Theorem 3.3 and thus omitted.

Proof of Theorem 3.5: 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 $G_T(\theta, \theta_1)$ is defined in (29) and $\hat{\theta}_T^{(*1)}, \hat{\theta}_T^{(*2)}$ lie between $\hat{\theta}_T^{(k)}$ and $\hat{\theta}_T^{(k-1)}$.

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

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

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

Rearranging the above equation leads to

$$D^2 L_T(\theta_0) (\hat{\theta}_T^{(k)} - \theta^0) = -\frac{\partial L_T(\theta^0)}{\partial \theta}.$$

□

Proof of Theorem 3.6: We provide a sketch of a proof for Algorithm I only. Applying Taylor expansion to both sides of the following equality at θ^0 , collecting terms, and ignoring higher order terms,

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

we obtain:

$$\hat{\theta}_T^{(k)} - \theta^0 = [-\Sigma(\theta^0, \theta^0)]^{-1} \left[\frac{\partial L_T(\theta^0)}{\partial \theta} \right] + [-\Sigma(\theta^0, \theta^0)]^{-1} [D^2 L_\infty(\theta^0) - \Sigma(\theta^0, \theta^0)] (\hat{\theta}_T^{(k-1)} - \theta^0).$$

Let

$$B_T = [-\Sigma(\theta^0, \theta^0)]^{-1} \left[\frac{\partial L_T(\theta^0)}{\partial \theta} \right], \quad A = [-\Sigma(\theta^0, \theta^0)]^{-1} [D^2 L_\infty(\theta^0) - \Sigma(\theta^0, \theta^0)].$$

Iterating the above equation, we get:

$$\hat{\theta}_T^{(k)} - \theta^0 = \sum_{j=0}^{k-2} A^j B_T + A^{k-1} (\hat{\theta}_T^{(1)} - \theta^0).$$

Thus, $\|A\| < 1$ implies:

$$\sqrt{T} (\hat{\theta}_T^{(k)} - \hat{\theta}_T^{(k-1)}) = A^{k-2} \sqrt{T} B_T + \sqrt{T} A^{k-2} (A - I) (\hat{\theta}_T^{(1)} - \theta^0) = o_p(1).$$

7 Appendix B. Algorithms III and IV

In this appendix, we provide two more algorithms, the corresponding expressions for $S_T(\theta, \theta_1)$, $G_T(\theta, \theta_1)$, and $G_\infty(\theta_1)$ introduced in Subsection 4.2, and the information dominance conditions used in Theorem 4.4.

Algorithm III.

Step 1. We start out our algorithm from an initial estimator denoted as $\hat{\theta}_T^{(0)}$;

Step k. Let $\hat{\theta}_T^{(k)}$ solve ($k = 1, 2, 3, \dots$)

$$\frac{\partial Q_T(\theta, \nu(\theta))}{\partial \theta} = - \frac{\partial Q_T(\theta, \nu(\hat{\theta}_T^{(k-1)}))}{\partial \nu} \frac{\partial \nu(\hat{\theta}_T^{(k-1)})}{\partial \theta}; \quad (57)$$

Step k'. Let $\hat{\theta}_T^{(k)}$ solve ($k = 1, 2, 3, \dots$)

$$\hat{\theta}_T^{(k)} = \hat{\theta}_T^{(k-1)} - \left[H_{2T}(\hat{\theta}_T^{(k-1)}) + \frac{\partial^2 Q_T}{\partial \nu \partial \theta'}[\hat{\theta}_T^{(k-1)}, \nu(\hat{\theta}_T^{(k-1)})] \frac{\partial \nu(\hat{\theta}_T^{(k-1)})}{\partial \theta} \right]^{-1} \left[\frac{\partial L_T(\hat{\theta}_T^{(k-1)})}{\partial \theta} \right]. \quad (58)$$

Algorithm IV.

Step 1. We start out our algorithm from an initial estimator denoted as $\hat{\theta}_T^{(0)}$;

Step k. Let $\hat{\theta}_T^{(k)}$ solve ($k = 1, 2, 3, \dots$)

$$\frac{\partial Q_T(\theta, \nu(\hat{\theta}_T^{(k-1)}))}{\partial \theta} = - \frac{\partial Q_T(\theta, \nu(\hat{\theta}_T^{(k-1)}))}{\partial \nu} \frac{\partial \nu(\hat{\theta}_T^{(k-1)})}{\partial \theta}; \quad (59)$$

Step k'. Let $\hat{\theta}_T^{(k)}$ solve ($k = 1, 2, 3, \dots$)

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

We now provide expressions for $S_T(\theta, \theta_1)$, $G_T(\theta, \theta_1)$, and $G_\infty(\theta_1)$ corresponding to Algorithms III and IV:

$$\begin{aligned} S_T(\theta, \theta_1) &= \frac{\partial Q_T(\theta, \nu(\theta))}{\partial \theta} + \frac{\partial Q_T(\theta, \nu(\theta_1))}{\partial \nu} \frac{\partial \nu}{\partial \theta}(\theta_1) \text{ for Algorithm III} \\ &= \frac{\partial Q_T(\theta, \nu(\theta_1))}{\partial \theta} + \frac{\partial Q_T(\theta, \nu(\theta_1))}{\partial \nu} \frac{\partial \nu}{\partial \theta}(\theta_1) \text{ for Algorithm IV;} \end{aligned}$$

$$\begin{aligned}
G_T(\theta, \theta_1) &= \frac{\partial^2 Q_T(\theta, \nu(\theta_1))}{\partial \theta \partial \theta'} + H_T(\theta, \theta_1) + \frac{\partial \nu}{\partial \theta'}(\theta_1) \frac{\partial^2 Q_T(\theta_1, \nu(\theta))}{\partial \nu \partial \theta'} \text{ for Algorithm III,} \\
&= \frac{\partial^2 Q_T(\theta, \nu(\theta_1))}{\partial \theta \partial \theta'} + \frac{\partial \nu}{\partial \theta'}(\theta_1) \frac{\partial^2 Q_T(\theta_1, \nu(\theta))}{\partial \nu \partial \theta'} \text{ for Algorithm IV;} \tag{60}
\end{aligned}$$

$$\begin{aligned}
G_\infty(\theta) &= -\Sigma(\theta, \theta) + H(\theta, \theta) + H(\theta, \theta)^T \text{ for Algorithm III,} \\
&= -\Sigma(\theta, \theta) + H(\theta, \theta)^T \text{ for Algorithm IV.}
\end{aligned}$$

The information dominance conditions required in Theorem 4.4 for Algorithms III and IV are:

Information Dominance III.

$$\| [-\Sigma(\theta^0, \theta^0) + H(\theta^0, \theta^0) + H(\theta^0, \theta^0)^T]^{-1} [D^2 L_\infty(\theta^0) + \Sigma(\theta^0, \theta^0) - H(\theta^0, \theta^0) - H(\theta^0, \theta^0)^T] \| < 1;$$

Information Dominance IV.

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

References

- [1] Amemiya, T. (1985), *Advanced Econometrics*, Havard University Press.
- [2] Dominitz, J. and R. P. Sherman (2005), "Some Convergence Theory for Iterative Procedures With an Application to Semiparametric Estimation," *Econometric Theory* 21, 838-863.
- [3] Duan, J.-C., G. Gauthier, and J.-G. Simonato (2004), "On the Equivalence of the KMV and Maximum Likelihood Methods for Structural Credit Risk Models," Manuscript.
- [4] Engle, R. (2002), "Dynamic Conditional Correlation: A Simple Class of Multivariate Generalized Autoregressive Conditional Heteroskedasticity Models," *Journal of Business & Economic Statistics* 20, 339-350.
- [5] Engle, R. and K. Sheppard (2001), "Theoretical and Empirical Properties of Dynamic Conditional Correlation Multivariate GARCH," Discussion Paper 2001-15, UCSD.
- [6] Kiefer, N. M. (1978), "Discrete Parameter Variation: Efficient Estimation of a Switching Regression Model," *Econometrica* 46, 427-434.

- [7] McNeil, A. J., R. Frey, and P. Embrechts (2005). *Quantitative Risk Management*, Princeton Series in Finance.
- [8] Merton, R. (1974), "On the Pricing of Corporate Debt: the Risk Structure of Interest Rates," *Journal of Finance* 28, 449-470.
- [9] Pastorello, S., V. Patilea, and E. Renault (2003), "Iterative and Recursive Estimation in Structural Nonadaptive Models," *Journal of Business & Economic Statistics* 21, 449-482.
- [10] Ruud, P. A. (2000), *An Introduction to Classical Econometric Theory*, Oxford University Press.
- [11] Song, P., Y. Fan, and J. Kalbfleisch (2005), "Maximization by Parts in Likelihood Inference," *Journal of the American Statistical Association* 100, 1145-1158.

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 full MLE computed using the Newton version of the MBP algorithm outlined by formula (16) in the text. 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 iterations of the two steps estimator were started at the true parameter values; the resulting estimates 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.

	True	Two steps estimator			MLE		
		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 4.2, and the full MLE computed using the Newton versions of the efficient algorithms developed in subsection 3.2 and appendix B in the text. The results are based on 5,000 synthetic samples of 500 time series observations of daily returns. The iterations of the KMV/PPR estimator were started at the true parameter values; the resulting 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.

Parameter	True	KMV/PPR			MLE		
		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 4.2, and the full MLE computed using the Newton versions of the efficient algorithms developed in subsection 3.2 and appendix B in the text. The results are based on 5,000 synthetic samples of 500 time series observations of daily returns. The iterations of the KMV/PPR estimator were started at the true parameter values; the resulting estimates were used as starting values for the efficient algorithms’ iterations. Of the four efficient algorithms, only algorithms I and IV converged to the MLE in every replication; algorithms II and III converged to the MLE only in the 89.92% and 16.72%, respectively, of the replications. The former two algorithms needed on average slightly less than 17 iterations, whereas the latter two needed roughly 44 and 29 iteration, respectively.

Parameter	True	KMV/PPR			MLE		
		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