Efficient Two-Step Estimation via Targeting

David T. Frazier*and Eric Renault†

April 24, 2015

Abstract

The standard description of two-step extremum estimation amounts to plugging in a first step estimator of nuisance parameters in order to simplify the optimization problem and then to deduce a user friendly estimator for the parameters of interest. This two-step procedure often induces an efficiency loss with respect to estimation of the parameters of interest. In this paper, we consider a more general setting where we do not necessarily have such thing as nuisance parameters but rather awkward occurrences of the parameters of interest. By awkward, we mean that within the estimating equations for a vector of unknown parameters of interest θ , some occurrences of θ, encapsulated by a vector $\nu(\theta)$, may be computationally tricky. Then, it is still the case that prior knowledge of the unknown auxiliary parameters $\nu = \nu(\theta)$ would make inference on θ much simpler, and it is this fact that motivates the two-step approach developed in this paper. The efficiency problem is more difficult than for the case of standard nuisance parameters since even the (infeasible) approach of plugging in the true unknown value of $\nu = \nu(\theta)$ may not allow efficiency, since it overlooks the information about θ contained in the awkward occurrences $\nu(\theta)$. Moreover, we stress that standard ways to restore efficiency for two-step procedures may not work due to a consistency issue; when setting the focus on a first step estimator for only some of the occurrences $\nu = \nu(\theta)$ of the unknown parameters θ, global identification may be lost. To alleviate this issue, we develop a targeting strategy that enforces consistency and achieves efficiency. Such difficult occurrences $\nu(\theta)$ of the parameters, which are a nuisance when it comes to solving estimating equations, are present in many financial econometrics applications, often handled by indirect inference. Leading examples are asset pricing models with latent variables (their observation would make estimation much simpler), models where it is simpler to first set the focus of inference on marginal distributions (multivariate GARCH, copulas), models with highly nonlinear objective functions, etc. Based on targeting and penalization of the auxiliary parameters, we propose a new two-step estimation procedure that leads to stable and user-friendly computations. Moreover, estimators delivered in the second step of the estimation procedure are asymptotically efficient. We compare this new method with existing iterative methods in the framework of copula models and asset pricing models. Simulation results illustrate that this new method performs better than existing iterative procedures and is (nearly) computationally equivalent.

*Department of Econometrics and Business Statistics, Monash University. email: david.frazier@monash.edu
†Department of Economics, Brown University. email: eric_renault@brown.edu

Keywords: Targeting, Penalization, Multivariate Time Series Models, Asset Pricing, Implied States.

1 Introduction

The standard treatment of two-stage estimation (see e.g. Pagan, 1986 or Newey and McFadden, 1994, section 6) is generally motivated by the following sequence of arguments as coined by Pagan (1986):

(i) Econometricians are often faced with the troublesome problem that "in order to estimate the parameters they are ultimately interested in, it becomes necessary to quantify a number of nuisance parameters (...) it is the presence of these parameters which converts a relatively simple computational problem into a very complex one".

(ii) "Because estimation would generally be easy if the nuisance parameter were known, a very common strategy for dealing with them has emerged: they are replaced by a nominated value which is estimated from the data". Then, the key issue for asymptotic theory is to assess the effect of first-step estimators on second-step standard errors (see Newey and McFadden, 1994, subsection 6.2) and the most favorable situation is when ignoring the first step would be valid: the asymptotic distribution on the second-step estimator for the parameters of interest does not depend on the first step estimator for the nuisance parameters and would have been the same whether the nuisance parameters had been known upfront.

Our focus of interest in this paper is germane to the above one but more general. The main difference is that we do not necessarily have such thing as nuisance parameters but rather awkward occurrences of the parameters of interest. By awkward, we mean that within the estimating equations for a vector of unknown parameters of interest θ, some occurrences of θ may be computationally tricky, either due to the complexity of the relationship, or numerical instability, or both. In order to disentangle these unpleasant occurrences from user-friendly ones, we denote the sample-based estimating functions as $q_T[\theta, \nu(\theta)]$, where $\nu(\theta)$ encapsulates all the occurrences of θ considered as somewhat awkward while T stands for the sample size. Generally speaking, our estimator of interest is $\hat{\theta}_T$ defined as a zero of the vector function $f_T(\theta) = q_T[\theta, \nu(\theta)]$.

Note that, this general framework obviously encompasses the standard nuisance parameter setting described above. If, within the vector θ of unknown parameters, we distinguish some parameters of interest, denoted by θ_1, and some nuisance parameters, denoted by θ_2, such that $\theta = (\theta_1', \theta_2')'$ and $\nu(\theta) = \theta_2$, we are back to the standard case as far as efficient estimation of θ_1 is concerned. Note that, up to a slight change of notation, our setup nests the case where the function $\nu(\theta)$ would be a sample dependent one $\nu_T(\theta)$, for instance because $\nu(\theta)$ shows up after some nuisance parameters have been profiled out. Up to a specific discussion on how to accommodate this case (see the Appendix), the simpler notation $\nu(\theta)$ will be kept throughout.

Our leading example will be the case of an extremum estimator

$$\hat{\theta}_T = \arg\max_{\theta} Q_T[\theta, \nu(\theta)], \tag{1}$$

so that the estimating equations correspond to first order conditions:

$$q_T[\theta, \nu(\theta)] = \frac{\partial Q_T[\theta, \nu(\theta)]}{\partial \theta} + \frac{\partial \nu'(\theta)}{\partial \theta} \frac{\partial Q_T[\theta, \nu(\theta)]}{\partial \nu}. \tag{2}$$

$\hat{\theta}_T$ may be the MLE if the function $Q_T[\theta, \nu(\theta)]$ is a well-specified (log)likelihood function. More

generally, we will see $\hat{\theta}_T$ throughout as our benchmark estimator for the purpose of asymptotic efficiency.

We highlight two important classes of examples in this paper. First, in Section 4, we consider a class of additively separable log-likelihood functions that are usually encountered in the so-called"estimation from likelihood of margins" (see e.g. Joe, 1997). In this setting, the unknown parameters, components of θ, can be split into two parts $\theta = (\theta_1', \theta_2')'$, where θ_1 characterizes the likelihood of the margins and θ_2 characterizes the dependence between components, let's say the "cross-dependence", through some link functions (typically linear correlations or copulas). However, the link function describing the cross-dependence applies to data components that have been first standardized using the knowledge of θ_1. In other words, the part of the likelihood capturing cross-dependence also involves the parameters θ_1 that describe the marginal distributions. Such occurrences of θ_1 are an example of the awkward occurrences mentioned earlier, in that these situations can be difficult to deal with in practice; i.e., $\nu(\theta) = \theta_1$ corresponds to the occurrences of θ_1 in the cross-dependence portion of the log-likelihood. Fortunately, a consistent user friendly estimator of θ_1 is available from the likelihood of the margins and can be plugged into the cross-dependence portion in order to estimate θ_2. This approach is actually popular for the estimation of nonlinear multivariate time series models like multivariate GARCH or copulas models. However, as explained below, the simplicity obviously entails an efficiency loss since the information in the cross-dependence model about the margin parameters θ_1 has been overlooked.

In Section 5, we consider nonlinear models in which observable variables are viewed as functions of some latent variables. Typically, the latent model, which is characterized by a vector of unknown parameters θ, specifies a Markov process for the state variables and defines their (possibly nonlinear) transition equation. Such an approach becomes difficult when the measurement equation of this non-linear state space model, which is the function that relates observable variables to latent ones, also depends on the same unknown parameters through a vector $\nu(\theta)$. While it would have been relatively easy to estimate θ from the observations on the latent variables, inference using available observations is complicated by the additional awkward occurrence of θ, namely $\nu(\theta)$, in the transformation from latent to observable variables. It is worth noting that the issue we have in mind is not really about filtering latent variables because we actually consider a case where the relationship latent-observable is one-to-one. Hence, backing out the latent variables from observations would have been easy if not polluted by the additional occurrence of unknown parameters in the measurement equation. This kind of situation is common in modern arbitrage-based asset pricing models with hedging of various sources of risk defined by an underlying model of state variables. Latent state variables are common factors, the dynamics of which characterizes the dynamics of observed yields or derivative asset prices. Since the measurement equation, typically an arbitrage-based asset pricing formula, is one-to-one, we can, following Pan (2002), dub "Implied States" the value of latent variables that can be backed out from observations for a given value of parameters $\nu(\theta)$. From this general intuition, Pan (2002) has extended the approach put forward by Renault and Touzi (1996) (and later revisited by Pastorello, Patilea and Renault (2003)) to devise the so-called "Implied States GMM" estimator. Again, the simplicity of this strategy also comes at the cost of an efficiency loss since the information content about θ brought by its awkward occurrence $\nu(\theta)$ is overlooked in this procedure. As Pan (2002) put it, "the efficiency

of this "optimal instrument"scheme is limited in that (...) we sacrifice efficiency by ignoring the dependence of $\sigma(t)$ on θ, " spot volatility $\sigma(t)$ backed out from option price being for her the implied state.

We are then generally faced with the following trade off between asymptotic efficiency and computational cost (both in terms of computational complexity and stability). On the one hand, we still contemplate that estimation would be easy if the awkward part $\nu(\theta)$ were known. Therefore, there is still some rationale to estimate it in a first stage, that is, if θ^0 stands for the true unknown value of θ, to replace $\nu(\theta^0)$ by a consistent sample counterpart $\tilde{\nu}_T$.

On the other hand, it is well known (see Newey and McFadden, 1994 for a discussion) that the two-step estimator obtained by plugging in the first-step consistent estimator $\tilde{\nu}_T$ of the nuisance parameters would be inefficient in general. However, we want to stress that in our more general case where ν is not necessarily a nuisance parameter but may be a known function $\nu(\theta)$ of parameters of interest, there is even no reason to believe that we would get a more accurate estimator by computing the infeasible estimator $\breve{\theta}_T$, the solution of

$$q_T[\breve{\theta}_T, \nu(\theta^0)] = 0. \tag{3}$$

On the contrary, there are many circumstances (see Pastorello et al., 2003 and references therein) in which the infeasible estimator $\breve{\theta}_T$ is actually less accurate than $\hat{\theta}_T$. The efficiency loss is due to the fact that the computation of $\breve{\theta}_T$ disregards the information about θ contained in the function $\nu(\theta)$ (see also Crepon et al., 1997 for a similar remark in a GMM context). More precisely, the efficient estimator $\hat{\theta}_T$ is asymptotically equivalent to:

$$\left[\frac{\partial q_T}{\partial \theta'}[\theta^0, \nu(\theta^0)] + \frac{\partial q_T}{\partial \nu'}[\theta^0, \nu(\theta^0)]\frac{\partial \nu}{\delta \theta'}(\theta^0)\right]^{-1} q_T[\theta^0, \nu(\theta^0)],$$

while the infeasible estimator $\breve{\theta}_T$ is asymptotically equivalent to:

$$\left[\frac{\partial q_T}{\partial \theta'}[\theta^0, \nu(\theta^0)]\right]^{-1} q_T[\theta^0, \nu(\theta^0)].$$

Two standard strategies are available in the literature to address this efficiency issue. A first possibility, as recently developed by Fan, Pastorello and Renault (2015) (hereafter, FPR) is to devise a sequence of estimators $\hat{\theta}_T^{(k)}, k = 1, 2...$ from a feasible counterpart of (3)

$$q_T[\hat{\theta}_T^{(k+1)}, \nu(\hat{\theta}_T^{(k)})] = 0, \tag{4}$$

with, for instance, the aforementioned consistent first-step estimator $\tilde{\theta}_T$ as the initial value ($\hat{\theta}_T^{(1)} = \tilde{\theta}_T$). Following a seminal paper by Song, Fan and Kalbfleich (2005) (hereafter, SFK) who had proposed a simplified version of this strategy in the particular case of separable log-likelihood functions (see Section 4 below), with the algorithm (4) being dubbed "Maximization by Parts" (MBP hereafter). The nice thing with (4) is that each step of the iteration to compute $\hat{\theta}_T^{(k+1)}$ from $\hat{\theta}_T^{(k)}$ is no more computationally demanding than the solution of (3). Moreover, by contrast with (3), this iterative procedure may allow us to reach efficiency since, when the iterative procedure (4) has a limit $\hat{\theta}_T^{(\infty)}$, this limit must coincide with the efficient estimator

$\hat{\theta}_T$. However, it is worth realizing that the required contraction mapping property to secure convergence of (4) is not in general fulfilled in finite samples. Therefore, a feasible efficient estimator relies upon the choice of a tuning parameter $k(T)$, going to infinity at a sufficient rate with the sample size T, in order to obtain an estimator $\hat{\theta}_T^{(k(T))}$ that is asymptotically equivalent to $\hat{\theta}_T$. This may obviously come with the computational cost of a large number $k(T)$ of iterations, especially when the required population contraction mapping property is hardly fulfilled. Needless to say, the situation is even worse when it is not fulfilled at all, as illustrated in Section 4 below.

The main goal of this paper is to promote a new efficient two-step procedure that does not require the contraction mapping property. We will argue that even though its second-step may be more computationally involved than each step of MBP, it keeps some of its simplicity, in particular by comparison with the brute force computation of the efficient estimator $\hat{\theta}_T$. Our efficient two-step procedure is actually an extension of a two-step extremum estimator first proposed by Trognon and Gourieroux (1990). The key intuition is to correct the naive two-step objective function $Q_T[\theta, \tilde{\nu}_T]$ to compensate for the inefficiency caused by plugging in the first-step consistent estimator $\tilde{\nu}_T$. Our proposed extremum estimator would then be

$$\hat{\theta}_T^{ext} = \arg\max_{\theta} \tilde{Q}_T[\theta, \tilde{\nu}_T], \tag{5}$$

with

$$\tilde{Q}_T[\theta, \tilde{\nu}_T] = Q_T[\theta, \tilde{\nu}_T] + \frac{\partial Q_T[\theta, \tilde{\nu}_T]}{\partial \nu'}.[\nu(\theta) - \tilde{\nu}_T] - \frac{1}{2}[\nu(\theta) - \tilde{\nu}_T]' J_T(\theta)[\nu(\theta) - \tilde{\nu}_T] \tag{6}$$

and

$$P \lim_{T=\infty} \left[J_T(\theta^0) + \frac{\partial^2 Q_T[\theta^0, \nu(\theta^0)]}{\partial \nu \partial \nu'} \right] = 0.$$

We show that, when consistent, the estimator $\hat{\theta}_T^{ext}$ is asymptotically equivalent to the efficient estimator $\hat{\theta}_T$. The main intuition for this result is that, up to the occurrence of unknown θ inside the matrix $J_T(\theta)$, the first order conditions of the maximization program (5) can be seen as a linearization of first order conditions (2) of the efficient program (1), namely, linearization with respect to ν in the neighborhood of the first-step estimator $\tilde{\nu}_T$. Then, the efficiency argument will be based on a generalization of an argument extensively studied by Robinson (1988). In this seminal paper, general efficiency comparisons are led between roots of rival estimating equations, in particular, as provided by local linearizations. However, we point out a difficulty that seems to have been overlooked in the literature so far. When linearization around a preliminary consistent estimator is applied to a vector of estimating equations, like, $f_T(\theta) = q_T[\theta, \nu(\theta)]$, but linearization is performed only with respect to the second set of occurrences of θ (the so-called awkward occurrences within $\nu(\theta)$), the fact that $f_T(\theta)$ may also depend nonlinearly on θ through first occurrences, say $\theta = \theta^*$ in $q_T[\theta^*, \nu(\theta)]$, can impair the consistency of the estimator defined as the root of this (partially) linearized estimating equation. More precisely, local identification is granted but not global identification.

Our proposed hedge against this risk is the addition of a penalty term $\alpha_T \|\nu(\theta) - \tilde{\nu}_T\|^2$ to the (partially linearized) estimating equations, with a tuning parameter α_T going to infinity

slower than the rate of convergence of our initial estimator $\tilde{\nu}_T$. In other words, both the MBP approach and our new two-step procedure with penalized partial linearization come with the cost of a tuning parameter. While the MBP approach requires choosing the number $k(T)$ of iterations, instead we will have to choose the rate of divergence $\alpha(T)$ of the penalty weight. We will see, for instance, that in the standard case where all estimators are root-T consistent, a rate $T^{1/4}$ is well suited. Moreover, we will propose two different two-step procedures, both based on partial linearization, depending upon whether we have a first-step consistent estimator $\tilde{\nu}_T$, only of the unpleasant parameters, or we have at our disposal an initial consistent estimator $\tilde{\theta}_T$ for the whole parameter vector θ (and then $\tilde{\nu}_T = \nu(\tilde{\theta}_T)$). Of course, in the latter case, the penalty term could instead be $\alpha_T \left\| \theta - \tilde{\theta}_T \right\|^2$ and should be able to enforce global identification in even more general circumstances.

The paper is organized as follows. The proposed extension of the Trognon and Gourieroux (1990) efficient two-step procedure is studied in Section 2. Our general result explains why some well known two-step estimators are efficient, in spite of the appearance to the contrary: Hatanaka (1974) for a dynamic regression model, Gourieroux, Monfort and Renault (1996) for a GMM estimator. It is worth stressing that efficiency is warranted in these two specific examples because consistency is not an issue. However, we also point out other examples, such as, nonlinear least squares and GMM, where consistency is not warranted, except if one uses the penalty strategy that we have devised through first order conditions. Robinson's (1988) comparison of estimators is developed in Section 3. It allows us to propose two different penalized two-step estimators, depending on whether one has at her disposal a first-step consistent estimator of θ^0 or only of $\nu(\theta^0)$. Section 4 sets the focus on the separable estimation problem with a detailed comparison with MBP, both analytically and through Monte Carlo experiments in the framework of a copula example. Section 5 addresses the general implied states issue, both in the context of maximum likelihood and GMM as well. Again, we are able to provide a detailed comparison with MBP, both analytically and through Monte Carlo experiments, in the simple framework of Merton's credit risk model. Concluding remarks are given in Section 6.

Mathematical proofs, regularity conditions and detailed Monte Carlo evidence are all gathered in the Appendix.

2 An efficient two-step extremum estimator

2.1 General framework

Let $\Theta \subset \mathbb{R}^p$ be a compact parameter space, and θ^0 the true unknown value of θ. Additional parameters ν are defined by some continuous function $\nu(.)$ from Θ to some subset Γ of $\mathbb{R}^q$. We assume that the extremum estimator $\hat{\theta}_T$ of θ, defined by (1), is a consistent asymptotically normal estimator of θ^0. In addition, we assume the following standard regularity conditions are satisfied.

Assumption A1: There is a real-valued deterministic function $Q_\infty[.,.]$, continuous on $\Theta \times \Gamma$

and such that:

$$\text{(i) } \text{Plim}_{T=\infty} \sup_{\theta\in\Theta} |Q_\infty[\theta,\nu(\theta)] - Q_T[\theta,\nu(\theta)]| = 0 \text{ and}$$
$$\text{(ii) } \theta^0 = \arg\max_{\theta\in\Theta} Q_\infty[\theta,\nu(\theta)].$$

Assumption A2: The following are satisfied

(i) $\nu(.)$ is twice continuously differentiable on the interior of Θ.

(ii) $\theta^0 \in Int(\Theta)$, interior set of Θ, and $\nu^0 = \nu(\theta^0) \in Int(\Gamma)$, interior set of Γ.

(iii) The function $Q_T[\theta,\nu]$ is twice continuously differentiable on $\mathring{\Theta}\times\ \mathring{\Gamma}$ and with $q_T[\theta,\nu(\theta)]$ defined by (2)

(1) $\sqrt{T}q_T[\theta^0,\nu(\theta^0)] \to_d \aleph[0,I_0]$

(2) $\text{Plim}_{T=\infty}\left\{\frac{\partial q_T[\theta^0,\nu(\theta^0)]}{\partial\theta'} + \frac{\partial q_T[\theta^0,\nu(\theta^0)]}{\partial\nu'}.\frac{\partial\nu(\theta^0)}{\partial\theta'}\right\} = H_0$

In addition, we maintain the following high-level assumptions.

Assumption A3: $(\hat{\theta}_T', \tilde{\nu}_T')'$ is a root-T consistent asymptotically normal estimator of $(\theta^{0\prime}, \nu^{0\prime})'$and

$$\text{Plim}_{T=\infty} \sup_{\theta\in\Theta} \left| J_T(\theta) + \frac{\partial^2 Q_T[\theta,\tilde{\nu}_T)]}{\partial\nu\partial\nu'} \right| = 0.$$

The focus of interest in this section is the comparison of the efficient estimator $\hat{\theta}_T$, with the two-step alternative $\hat{\theta}_T^{ext}$ defined in the introduction. For the sake of interpretation, it is worth comparing $\hat{\theta}_T$ and $\hat{\theta}_T^{ext}$ with the infeasible estimator

$$\hat{\theta}_T^* = \arg\max_{\theta\in\Theta} \tilde{Q}_T^0[\theta,\tilde{\nu}_T],$$

where

$$\tilde{Q}_T^0[\theta,\tilde{\nu}_T] = Q_T[\theta,\tilde{\nu}_T] + \frac{\partial Q_T[\theta,\tilde{\nu}_T]}{\partial\nu'}.[\nu(\theta)-\tilde{\nu}_T] - \frac{1}{2}[\nu(\theta)-\tilde{\nu}_T]' J_T(\theta^0)[\nu(\theta)-\tilde{\nu}_T].$$

We are then able to prove the following result.

Theorem 2.1: Under the maintained assumption that they are all root-T consistent, the three estimators $\hat{\theta}_T$, $\hat{\theta}_T^{ext}$ and $\hat{\theta}_T^*$ are asymptotically equivalent.

We stress that, as announced in the introduction, it is only the careful analysis of the first order conditions (see section 3 below) that will allow us to devise a proper penalty to ensure consistency of our two-step estimators. It is, however, worth interpreting them further to discern the reason why the two-step approach is not responsible for any efficiency loss. We

can do that at least in the only case considered by Trognon and Gourieroux (1990), namely the case of a genuine nuisance parameter

$$\theta = (\theta_1', \theta_2')', \nu(\theta) = \theta_2.$$

Then, the modified objective function becomes

$$\tilde{Q}_T[\theta, \tilde{\nu}_T] = Q_T[\theta, \tilde{\nu}_T] + \frac{\partial Q_T[\theta, \tilde{\nu}_T]}{\partial \nu'}.[\theta_2 - \tilde{\nu}_T] - \frac{1}{2}[\theta_2 - \tilde{\nu}_T]' J_T(\theta)[\theta_2 - \tilde{\nu}_T],$$

and the parameters of interest for efficient estimation are included in the sub-vector θ_1. With this point in mind, we can set the focus on an even simpler two-step estimator obtained as the maximizer of the following simplified objective function, where for sake of avoiding confusion about partial derivatives, we use two different notations for the same first-step estimator

$$\begin{aligned} \tilde{\theta}_{2,T} &= \tilde{\nu}_T \\ \breve{Q}_T[\theta, \tilde{\nu}_T] &= Q_T[\theta_1, \tilde{\theta}_{2,T}, \tilde{\nu}_T] + \frac{\partial Q_T[\theta_1, \tilde{\theta}_{2,T}, \tilde{\nu}_T]}{\partial \nu'}.[\theta_2 - \tilde{\nu}_T] - \frac{1}{2}[\theta_2 - \tilde{\nu}_T]' J_T(\theta_1, \tilde{\theta}_{2,T})[\theta_2 - \tilde{\nu}_T] \end{aligned}$$

Then, it is easy to profile θ_2 out of $\breve{Q}_T[\theta, \tilde{\nu}_T]$

$$\frac{\partial \breve{Q}_T[\theta, \tilde{\nu}_T]}{\partial \theta_2} = 0 \Leftrightarrow \theta_2 = \tilde{\nu}_T + \left[J_T(\theta_1, \tilde{\theta}_{2,T})\right]^{-1} \frac{\partial Q_T[\theta_1, \tilde{\theta}_{2,T}, \tilde{\nu}_T]}{\partial \nu}.$$

Plugging the above value of θ_2 into $\breve{Q}_T[\theta, \tilde{\nu}_T]$, we can concentrate the objective function with respect to the nuisance parameters $\nu(\theta) = \theta_2$ and obtain the following profile objective function

$$\breve{Q}_{c,T}[\theta_1, \tilde{\nu}_T] = Q_T[\theta_1, \tilde{\theta}_{2,T}, \tilde{\nu}_T] + \frac{1}{2}\frac{\partial Q_T[\theta_1, \tilde{\theta}_{2,T}, \tilde{\nu}_T]}{\partial \nu'}\left[J_T(\theta_1, \tilde{\theta}_{2,T})\right]^{-1} \frac{\partial Q_T[\theta_1, \tilde{\theta}_{2,T}, \tilde{\nu}_T]}{\partial \nu}.$$

For sake of interpretation, let us consider instead the infeasible objective function and its profile counterpart. Then, the concentrated score vector is

$$\frac{\partial \breve{Q}^0_{c,T}[\theta_1, \tilde{\nu}_T]}{\partial \theta_1} = \frac{\partial Q_T[\theta_1, \tilde{\theta}_{2,T}, \tilde{\nu}_T]}{\partial \theta_1} + \frac{\partial^2 Q_T[\theta_1, \tilde{\theta}_{2,T}, \tilde{\nu}_T]}{\partial \theta_1 \partial \nu'}\left[J_T(\theta^0_1, \tilde{\theta}_{2,T})\right]^{-1} \frac{\partial Q_T[\theta_1, \tilde{\theta}_{2,T}, \tilde{\nu}_T]}{\partial \nu}.$$

From the definition of the matrix $J_T(\theta^0)$, we can then deduce that

$$\text{Plim}_{T=\infty} \frac{\partial \breve{Q}^0_{c,T}[\theta^0_1, \nu^0]}{\partial \theta_1 \partial \nu'} = 0. \tag{7}$$

Equation (7) is precisely the standard condition (see e.g Newey and McFadden, 1994, formula (6.6) p 2179) to ensure that the asymptotic distribution of the estimator of the parameters of interest θ_1 does not depend on the asymptotic distribution of the estimator for the nuisance parameters ν. This provides clear intuition as to why Theorem 2.1. works, at least in the particular case considered by Trognon and Gourieroux (1990): the modification of the objective

function in (6) has been devised precisely to restore the asymptotic independence between the two kinds of parameters. However, the main contribution of the paper is to provide a much more general setup for efficient two-step estimation through the use of penalized estimating equations. This penalty amounts to a slight twist (via targeting) on the two-step estimator $\hat{\theta}_T^{ext}$, in order to ensure its consistency. Then, its asymptotic equivalence with $\hat{\theta}_T$, as stated in Theorem 2.1., ensures its asymptotic efficiency. As far as the equivalence between $\hat{\theta}_T^{ext}$ and $\hat{\theta}_T^*$ is concerned, note that this is germane to the equivalence between two-step efficient GMM and continuously updated GMM, as first put forward by Hansen et al. (1996).

2.2 Application to nonlinear regression

In this subsection, we consider the example of nonlinear least squares. Note that while we consider only ordinary least squares, weighted least squares would not introduce any specific difficulty. Joint estimation of models for conditional mean and variance using Gaussian QMLE (Bollerslev and Wooldridge, 1992) would also fit in this class of examples. Thus, for sake of notational simplicity, let us just consider the following objective function

$$Q_T[\theta, \nu(\theta)] = -\frac{1}{T}\sum_{t=1}^{T}\left[y_t - g(x_t, \theta, \nu(\theta))\right]^2,$$

where $g(.,.,.)$ is a known function such that

$$g(x_t, \theta^0, \nu(\theta^0)) = E[y_t \,|x_t]\,. \tag{8}$$

Hence, the maintained identification assumption is

$$E[y_t - g(x_t, \theta, \nu(\theta))\,|x_t] = 0 \Leftrightarrow \theta = \theta^0. \tag{9}$$

Then,

$$\begin{aligned}
\frac{\partial Q_T[\theta, \nu(\theta)]}{\partial \nu} &= \frac{2}{T}\sum_{t=1}^{T}\frac{\partial g(x_t, \theta, \nu(\theta))}{\partial \nu}\left[y_t - g(x_t, \theta, \nu(\theta)\right], \\
\frac{\partial^2 Q_T[\theta, \nu(\theta)]}{\partial \nu \partial \nu'} &= -\frac{2}{T}\sum_{t=1}^{T}\frac{\partial g(x_t, \theta, \nu(\theta))}{\partial \nu}.\frac{\partial g(x_t, \theta, \nu(\theta))}{\partial \nu'} \\
&\quad + \frac{2}{T}\sum_{t=1}^{T}\frac{\partial^2 g(x_t, \theta, \nu(\theta))}{\partial \nu \partial \nu'}\left[y_t - g(x_t, \theta, \nu(\theta)\right].
\end{aligned}$$

However, by applying (8), we can choose the following consistent estimator for the Hessian matrix with respect to the parameters ν

$$J_T(\theta) = \frac{2}{T}\sum_{t=1}^{T}\frac{\partial g(x_t, \theta, \tilde{\nu}_T)}{\partial \nu}.\frac{\partial g(x_t, \theta, \tilde{\nu}_T)}{\partial \nu'}.$$

With this choice, the modified extremum estimator is obtained as the maximizer of

$$\begin{aligned}\tilde{Q}_T[\theta,\tilde{\nu}_T] &= Q_T[\theta,\tilde{\nu}_T]+\frac{\partial Q_T[\theta,\tilde{\nu}_T]}{\partial\nu'}.[\nu(\theta)-\tilde{\nu}_T]-\frac{1}{2}[\nu(\theta)-\tilde{\nu}_T]'J_T(\theta)[\nu(\theta)-\tilde{\nu}_T] \quad (10)\\ &= -\frac{1}{T}\sum_{t=1}^{T}\left[y_t-g(x_t,\theta,\tilde{\nu}_T)-\frac{\partial g(x_t,\theta,\tilde{\nu}_T)}{\partial\nu'}.[\nu(\theta)-\tilde{\nu}_T]\right]^2.\end{aligned}$$

In other words, while the estimator defined as the solution to

$$\min_{\theta}\sum_{t=1}^{T}[y_t-g(x_t,\theta,\tilde{\nu}_T)]^2$$

is not efficient in general, we can restore efficiency by the additional term in (10). The fact that a nonlinear regression model can be efficiently estimated after linearization of the regression function around a first-step consistent estimator has been known since Hartley (1961). However, it must be kept in mind that the case (10) is more general because we consider only a partial linearization so as to deal with the nasty occurrences $\nu(\theta)$ in $g(\cdot)$. As a result, efficiency is warranted only when consistency is enforced, which may take the penalty strategy developed in Section 3. To see this, note that the identification assumption (9) does not say that

$$E[y_t\,|x_t]=g(x_t,\theta,\nu(\theta^0))-\frac{\partial g(x_t,\theta,\nu(\theta^0))}{\partial\nu'}.\left[\nu(\theta)-\nu(\theta^0)\right]\Rightarrow\theta=\theta^0.$$

The role of targeting will be to enforce the equality $\nu(\theta)=\nu(\theta^0)$ so that the implication above becomes a consequence of the identification assumption (9). Fortunately, there are cases where penalty/targeting is not needed because consistency is directly implied. Trognon and Gourieroux (1990) point out the example of Hatanaka's (1974) two-step estimator for a dynamic adjustment model with autoregressive errors. With obvious notations, the model is

$$\begin{aligned}y_t &= \alpha_1 y_{t-1}+\alpha_2 z_t+u_t\\ u_t &= \beta u_{t-1}+\varepsilon t\end{aligned}$$

and is generally rewritten as

$$y_t-\beta y_{t-1}=\alpha_1(y_{t-1}-\beta y_{t-2})+\alpha_2(z_t-\beta z_{t-1})+\varepsilon_t.$$

Thus, we end up with a nonlinear regression model that can be rewritten in the notational system of (8)

$$\begin{aligned}y_t &= g(x_t,\alpha_1,\alpha_2,\nu(\theta))+\varepsilon_t\\ x_t &= (z_t,y_{t-1}),\theta=(\alpha_1,\alpha_2,\beta)',\nu(\theta)=\beta.\end{aligned}$$

However, a key remark is that the regression function, albeit nonlinear, is linear with respect to ν when the friendly occurrence of θ is fixed. Therefore, this partial linearization with respect to ν does not cause consistency to break down. Theorem 2.1 can be directly applied to confirm

that Hatanaka's (1974) two-step estimator is efficient.

2.3 Application to GMM

We now contemplate the case of a parameter identified through H moment restrictions with two kinds of occurrences for the parameters:

$$E[\varphi_t(\theta, \nu(\theta))] = 0 \Leftrightarrow \theta = \theta^0. \tag{11}$$

Moment restrictions of the form in (11), and their possible applications, are discussed in more details in Section 5 within the setting of implied states GMM. In this section, however, we give a general discussion in the context of modified two-step extremum estimators.

When working with (11), we typically have in mind estimators defined from the criterion function

$$Q_T[\theta, \nu(\theta)] = -\bar{\varphi}_T(\theta, \nu(\theta))' W_T \bar{\varphi}_T(\theta, \nu(\theta))$$

where

$$\bar{\varphi}_T(\theta, \nu(\theta)) = \frac{1}{T}\sum_{t=1}^{T} \varphi_t(\theta, \nu(\theta))$$

and W_T is some positive definite sequence of matrices. Note that, in order to obtain an estimator $\hat{\theta}_T$, defined by (1), that reaches the semiparametric efficiency bound, the sequence W_T should provide a consistent estimator for the inverse of the long term variance matrix $\lim_{T=\infty} Var\left[\sqrt{T}\bar{\varphi}_T(\theta^0, \nu(\theta^0))\right]$. However, this issue is irrelevant for us as we only discuss how to obtain estimators that are asymptotically equivalent to $\hat{\theta}_T$, irrespective of its efficiency.

From the definition of $Q_T[\theta, \nu(\theta)]$,

$$\begin{aligned}
\frac{\partial Q_T[\theta, \nu(\theta)]}{\partial \nu} &= -2\frac{\partial \bar{\varphi}_T(\theta, \nu(\theta))'}{\partial \nu} W_T \bar{\varphi}_T(\theta, \nu(\theta)) \\
\frac{\partial^2 Q_T[\theta, \nu(\theta)]}{\partial \nu \partial \nu'} &= -2\frac{\partial \bar{\varphi}_T(\theta, \nu(\theta))'}{\partial \nu} W_T \frac{\partial \bar{\varphi}_T(\theta, \nu(\theta))}{\partial \nu} \\
&\quad -2\sum_{h=1}^{H} \frac{\partial^2 \bar{\varphi}_{h;T}(\theta, \nu(\theta))}{\partial \nu \partial \nu'} . W_{h.;T} \bar{\varphi}_T(\theta, \nu(\theta))
\end{aligned}$$

where $W_{h.;T}$ stands for the h^{th} row of W_T. Then, we can choose the following consistent estimator for the Hessian matrix with respect to the parameters ν

$$J_T(\theta) = 2\frac{\partial \bar{\varphi}_T(\theta, \nu(\theta))'}{\partial \nu} W_T \frac{\partial \bar{\varphi}_T(\theta, \nu(\theta))}{\partial \nu'}.$$

With this choice, the modified extremum estimator is obtained as the maximizer of

$$\begin{aligned}\tilde{Q}_T[\theta,\tilde{\nu}_T] &= Q_T[\theta,\tilde{\nu}_T]+\frac{\partial Q_T[\theta,\tilde{\nu}_T]}{\partial\nu'}.\left[\nu(\theta)-\tilde{\nu}_T\right]-\frac{1}{2}\left[\nu(\theta)-\tilde{\nu}_T\right]'J_T(\theta)\left[\nu(\theta)-\tilde{\nu}_T\right] \qquad (12)\\ &= -\left[\bar{\varphi}_T(\theta,\tilde{\nu}_T)+\frac{\partial\bar{\varphi}_T(\theta,\tilde{\nu}_T)}{\partial\nu'}.\left[\nu(\theta)-\tilde{\nu}_T\right]\right]'W_T\left[\bar{\varphi}_T(\theta,\tilde{\nu}_T)+\frac{\partial\bar{\varphi}_T(\theta,\tilde{\nu}_T)}{\partial\nu'}.\left[\nu(\theta)-\tilde{\nu}_T\right]\right]\end{aligned}$$

In other words, while the solution of

$$\min_{\theta}\left[\bar{\varphi}_T(\theta,\tilde{\nu}_T)\right]'W_T\left[\bar{\varphi}_T(\theta,\tilde{\nu}_T)\right] \qquad (13)$$

would not be equivalent to $\hat{\theta}_T$ in general, we can restore equivalence (and efficiency in the sense of $\hat{\theta}_T$) by using the additional term in (12). However, since (similarly to the former subsection) it is only a partial linearization of the moment conditions, consistency may not be warranted.

To see why consistency may be an issue, note that the identification assumption (11) does not say that

$$E\left[\varphi_t(\theta,\nu(\theta^0))+\frac{\partial\varphi_t(\theta,\nu(\theta^0))}{\partial\nu'}.\left[\nu(\theta)-\nu(\theta^0)\right]\right]=0\Longrightarrow\theta=\theta^0. \qquad (14)$$

The role of targeting in this context is to enforce the equality $\nu(\theta)=\nu(\theta^0)$ so that the implication in (14) becomes a consequence of the identification assumption (11).

Fortunately, there are cases where the penalty/targeting is not needed because consistency is directly implied. Gourieroux et al. (1996) consider the case where the vector of moment conditions can be split in two parts, with only the second one depending on ν:

$$\varphi_t(\theta,\nu(\theta))=\left[\varphi_{1t}(\theta)',\varphi_{2t}(\theta,\nu(\theta))'\right]'. \qquad (15)$$

Then, the implication (14) is obviously warranted when the first set of moment conditions is sufficient to identify θ, which is typically the case considered by Gourieroux et al. (1996). In this case, Theorem 2.1. ensures efficiency of the modified two-step estimator.

Interestingly enough, the efficient two-step estimator proposed by Gourieroux et al. (1996) may be different from $\hat{\theta}_T^{ext}$. It is only when the first set of moment conditions $\varphi_{1t}(\theta)$ is linear with respect to θ that they will numerically coincide (see section 2.6 in Gourieroux et al., 1996). In the general case, their two-step efficient estimator is not based on a (partial) linearization but on minimizing the norm of the moment vector $\bar{\varphi}_T(\theta,\tilde{\nu}_T)$, where the weighted matrix is a suitably twisted version of the estimator for the inverse of the long term variance matrix. Note that, we know from (13) that efficiency cannot be met without such a twist. Moreover, our equivalence result is more general since not only does it apply to general moment conditions (not only in the form (15)) but it does not assume that W_T is a consistent estimator of the inverse of the long term variance matrix.

3 Stochastic differences for linearized estimating equations

We first state our general result concerning roots of linearized estimating equations, which extends Theorem 2 of Robinson (1988). Then, in a second subsection, we provide two more user friendly versions of our two-step estimator, depending on whether one want to use a first-step consistent estimator of θ^0 or only of $\nu(\theta^0)$. Note that, this section is generally valid for estimating equations and their linearizations, irrespective of the fact that these estimating equations can be seen as first order conditions provided by an extremum estimator, as in our leading example studied in Section 2.

3.1 The general result

Linear approximations will be considered in some neighborhood $\aleph(\varepsilon), \varepsilon > 0$, of the true unknown value

$$\aleph(\varepsilon) = \left\{\theta \in \mathbb{R}^p : \left\|\theta - \theta^0\right\| < \varepsilon\right\} \subset \Theta.$$

Note that, the existence of such ε is tantamount to the maintained assumption that the true unknown value θ^0 belongs to the interior of the parameter space.

In order to extend the results of Robinson (1988), we first characterize our benchmark estimator $\hat{\theta}_T$ as the solution of some just-identified estimating equations. For sake of generality, we maintain some high level assumptions about these estimating equations, although they would in general be implied by more primitive assumptions, as seen in our leading example of extremum estimation in section 2.

Assumption B1: $f_T(\theta) = q_T[\theta, \nu(\theta)]$ is a p-vector valued random variable such that:

(i) f_T has a zero $\hat{\theta}_T = \theta^0 + o_P(1)$,

(ii) For some $\varepsilon > 0$, the functions of θ: $\nu(\theta)$, $f_T(\theta)$ and $\frac{\partial q_T}{\partial \nu'}[\theta, \nu(\theta^*)]$ are continuously differentiable on $\aleph(\varepsilon)$, for any given θ^* in $\aleph(\varepsilon)$.

(iii) $F_T(\theta^0) = F + o_P(1)$, where $F_T(\theta) = \frac{\partial f_T(\theta)}{\partial \theta'}$ and F is non-singular.

Under standard regularity conditions (see appendix), the non-singular matrix F can obviously be written as

$$F = \frac{\partial q_\infty[\theta^0, \nu(\theta^0)]}{\partial \theta'} + \frac{\partial q_\infty}{\partial \nu'}[\theta^0, \nu(\theta^0)]\frac{\partial \nu}{\partial \theta'}(\theta^0),$$

for some population estimating equations $q_\infty[\theta, \nu(\theta)]$ with θ^0 the only zero of $q_\infty[\theta, \nu(\theta)]$.

Assumption B2: $q_\infty[\theta, \nu(\theta)] = 0 \Leftrightarrow \theta = \theta^0$.

We are interested in partially linear approximations of the estimating function around some consistent initial estimator $\tilde{\theta}_T$. Thus, let us define

$$\tilde{h}_T(\theta) = q_T[\theta, \nu(\tilde{\theta}_T)] + \frac{\partial q_T}{\partial \nu'}[\theta, \nu(\tilde{\theta}_T)]\frac{\partial \nu}{\partial \theta'}(\tilde{\theta}_T)(\theta - \tilde{\theta}_T).$$

Note that $\tilde{h}_T(\theta)$ provides alternative estimating equations that also locally identify θ since, with obvious notations (and under standard regularity conditions), a solution $\theta = \theta_T^*$ of $g_T(\theta) = 0$ will converge towards a solution $\theta = \bar{\theta}$ of the population equations

$$q_\infty[\theta, \nu(\theta^0)] + \frac{\partial q_\infty}{\partial \nu'}[\theta, \nu(\theta^0)] \frac{\partial \nu}{\partial \theta'}(\theta^0)(\theta - \theta^0) = 0.$$

With a genuine linearization of $f_T(\theta)$ (not only a partial one), Robinson's Theorem 2 shows that a zero of $\tilde{h}_T(\theta)$ is, in a sense, asymptotically equivalent to $\hat{\theta}_T$. With a partial linearization, we cannot maintain such a claim since we may only have local identification and not global identification. That is, there may exist some $\bar{\theta} \neq \theta^0$ such that, with obvious notations,

$$q_\infty[\bar{\theta}, \nu(\theta^0)] + \frac{\partial q_\infty}{\partial \nu'}[\bar{\theta}, \nu(\theta^0)] \frac{\partial \nu}{\partial \theta'}(\theta^0)(\bar{\theta} - \theta^0) = 0$$

even though $\theta = \theta^0$ is the only solution of

$$q_\infty[\theta, \nu(\theta)] = 0.$$

To avoid such a perverse situation, we have to slightly penalize our (partially) linearized sequence by defining:

$$\tilde{h}_T^P(\theta) = q_T[\theta, \nu(\tilde{\theta}_T)] + \frac{\partial q_T}{\partial \nu'}[\theta, \nu(\tilde{\theta}_T)] \frac{\partial \nu}{\partial \theta'}(\tilde{\theta}_T)(\theta - \tilde{\theta}_T) + \alpha_T \left\| \theta - \tilde{\theta}_T \right\|^2 e_p, \tag{16}$$

for a real sequence α_T going slowly to infinity, where e_p stands for the p-dimensional vector whose components all equal 1. More precisely, our extension of Robinson's result can be stated as follows:

Proposition 3.1: Under standard regularity conditions detailed in the appendix: under Assumption B1, if $\tilde{\theta}_T$ is a consistent estimator of θ^0 such that $\left\| \tilde{\theta}_T - \theta^0 \right\| = o_P(1/\alpha_T)$ with $\lim_{T=\infty} \alpha_T = \infty$, then for any zero $\tilde{\theta}_T^P$ of $\tilde{h}_T^P(\theta)$ in (16)

$$\hat{\theta}_T - \tilde{\theta}_T^P = O_P\left(\alpha_T \left\| \hat{\theta}_T - \tilde{\theta}_T \right\|^2 \right).$$

Proposition 3.1 is a generalization of Theorem 2 in Robinson (1988). When there is no function ν, our Assumptions B1 and B2 exactly match those of Robinson (1988). However, there is a price to pay for a linearization that is only partial and thus takes a penalty term α_T going to infinity. Fortunately, the penalty term $\alpha_T \left\| \hat{\theta}_T - \tilde{\theta}_T \right\|^2$, which shows up in the rate of convergence, will more often than not have a very minor impact for the use of Proposition 3.1. As a matter of fact, Proposition 3.1. will often be applied to state that, when the initial estimator $\tilde{\theta}_T$ is root-T consistent, the two estimators $\hat{\theta}_T$ and $\tilde{\theta}_T^P$ are first order asymptotically equivalent. This conclusion is indeed warranted insofar as we pick a penalty rate α_T going to infinity slower than $\sqrt{T}$. However, the choice of the tuning parameter is more constrained if one wants to use an initially consistent estimator $\tilde{\theta}_T$ converging slower than $\sqrt{T}$. Exactly as in

the case of Robinson (1988), the conclusion of asymptotic equivalence between $\hat{\theta}_T$ and $\tilde{\theta}_T^P$ takes anyway an initial estimator $\tilde{\theta}_T$ converging faster than $T^{1/4}$. But, on top of that, if the rate of convergence of $\tilde{\theta}_T$ is, say, $T^{(1/4)+\varepsilon}$, $\varepsilon > 0$, (resp $T^{(1/4)} \log(T)$), the wished asymptotic equivalence will be warranted only for a slowly diverging penalty rate α_T like T^{ε} (resp. $\log[\log(T)]$). It is worth noting a tight similarity between the choice of this tuning parameter α_T and the choice of the number $k(T)$ of iterations in iterative procedures like generalized backfitting in Pastorello et al. (2003) or MBP in Fan et al. (2015). As it can be seen, for instance, on the bottom of page 465 in Pastorello et al. (2003), $k(T)$ must go to infinity faster than $\log(T)$ and, in finite samples, the size of the needed $k(T)$ is inversely related to the strength of the contraction in the contraction mapping argument at stake for convergences of the iterations.

3.2 A couple of two-step efficient estimators

Our two-step efficient estimator is a direct extension of Robinson (1988), replacing the complete linearization by a partial one, and is defined as a zero $\tilde{\theta}_T^P$ of the estimating equations

$$\tilde{h}_T^P(\theta) = q_T[\theta, \nu(\tilde{\theta}_T)] + \frac{\partial q_T}{\partial \nu'}[\theta, \nu(\tilde{\theta}_T)] \frac{\partial \nu}{\partial \theta'}(\tilde{\theta}_T)(\theta - \tilde{\theta}_T) + \alpha_T \left\| \theta - \tilde{\theta}_T \right\|^2 e_p.$$

These estimating equations provide an efficient estimator by contrast with the naive two-step strategy that would only solve the equations

$$q_T[\theta, \nu(\tilde{\theta}_T)] = 0.$$

with iteration on these equations a possibility. Up to the penalty term (only used to enforce consistency), the difference between $q_T[\theta, \nu(\tilde{\theta}_T)]$ and $\tilde{h}_T^P(\theta)$ is the introduction of the first-order correction through partial linearization. However, there is an obvious way to make the estimating equations $\tilde{h}_T^P(\theta)$ even more computationally friendly by making the correction term linear in the unknown parameters θ; that is, rather, by solving the following estimating equations:

$$h_T^{(1)}(\theta) = q_T[\theta, \nu(\tilde{\theta}_T)] + \frac{\partial q_T}{\partial \nu'}[\tilde{\theta}_T, \nu(\tilde{\theta}_T)] \frac{\partial \nu}{\partial \theta'}(\tilde{\theta}_T)(\theta - \tilde{\theta}_T) + \alpha_T \left\| \theta - \tilde{\theta}_T \right\|^2 e_p. \tag{17}$$

The difference between $\tilde{h}_T^P(\theta)$ and $h_T^{(1)}(\theta)$ is that we have also plugged in the first-step consistent estimator $\tilde{\theta}_T$ to replace the non-awkward occurrence of the parameters θ in the complete Jacobian matrix. The great thing with this simplifying modification is that it does not impair the general equivalence result of proposition 3.1

Theorem 3.1: Under standard regularity conditions detailed in appendix: Under assumptions B1 and B2, if $\tilde{\theta}_T$ is a consistent estimator of θ^0 such that $\left\| \tilde{\theta}_T - \theta^0 \right\| = o_P(1/\alpha_T)$ with $\lim_{T=\infty} \alpha_T = \infty$, then for any zero $\theta_T^{(1)}$ of $h_T^{(1)}(\theta)$ in (17)

$$\hat{\theta}_T - \theta_T^{(1)} = O_P\left(\alpha_T \left\| \hat{\theta}_T - \tilde{\theta}_T \right\|^2 \right).$$

Theorem 3.1 implies that the previous discussion about the asymptotic efficiency of $\tilde{\theta}_T^P$, deduced from Proposition 3.1, applies similarly to efficiency of $\theta_T^{(1)}$.

While $\theta_T^{(1)}$ is obviously the most computationally friendly two-step estimator when we have at our disposal a first-step consistent estimator $\tilde{\theta}_T$, it may be a shame to require the use of such an estimator when, after all, our only trouble is to properly deal with the awkward parameter occurrences $\nu(\theta)$. We now propose an alternative two-step efficient estimator $\theta_T^{(2)}$ that only requires knowledge of a first-step consistent estimator $\tilde{\nu}_T$ of $\nu(\theta^0)$, and not the knowledge of a consistent estimator $\tilde{\theta}_T$ of the complete parameter vector θ^0. In applications, it may be typically the case that only a sub-vector of the parameters of interest θ can be consistently estimated in a first-step. Of course, the price to pay for this additional extension of Robinson (1988) will be to give up the computational simplification brought by the change from the estimating equations $\tilde{h}_T^P(\theta)$ to $h_T^{(1)}(\theta)$ (change from Proposition 3.1 to Theorem 3.1). By definition, if we don't have such thing as a first-step estimator $\tilde{\theta}_T$, we cannot plug it in to simplify the equations.

However, we will be able to derive an alternative two-step efficient estimator through the estimating equations defined by

$$h_T^{(2)}(\theta) = q_T[\theta, \tilde{\nu}_T\ ] + \frac{\partial q_T}{\partial \nu'}[\theta, \tilde{\nu}_T\ ].(\nu(\theta) - \tilde{\nu}_T\ ) + \alpha_T \left\|\nu(\theta) - \tilde{\nu}_T\ \right\|^2 e_p. \tag{18}$$

Theorem 3.2: Under standard regularity conditions detailed in the appendix: under assumptions B1 and B2, if $\tilde{\nu}_T$ is a consistent estimator of $\nu(\theta^0)$ such that $\|\tilde{\nu}_T - \nu(\theta^0)\| = o_P(1/\alpha_T)$ with $\lim_{T=\infty} \alpha_T = \infty$, then for any zero $\theta_T^{(2)}$ of $h_T^{(2)}(\theta)$ in (18)

$$\hat{\theta}_T - \theta_T^{(2)} = O_P\left(\alpha_T \left\|\nu(\hat{\theta}_T) - \tilde{\nu}_T\right\|^2\right).$$

Theorem 3.2 implies that the previous discussion about the asymptotic efficiency of $\tilde{\theta}_T^P$ and $\theta_T^{(1)}$, deduced from Proposition 3.1 and Theorem 3.1, respectively, applies similarly to efficiency of $\theta_T^{(2)}$. The main difference is that the leading rate of convergence is now the one of the estimator $\tilde{\nu}_T$. It is also worth noting that the idea of the proof of Theorem 3.2 can be applied even when plugging in a first-step consistent estimator $\tilde{\theta}_T$ to replace part of or all components of the first occurrence of θ in the Jacobian term. In particular, the two simplifying ideas of Theorems 3.1 and 3.2 can be used simultaneously.

3.3 Practical implications:

In this subsection, we set the focus on the simplest case where both the benchmark efficient estimator $\hat{\theta}_T$ and the initial estimators $\tilde{\theta}_T$ or $\tilde{\nu}_T$ are all root-T consistent.

When $\hat{\theta}_T$ is defined as the solution of

$$q_T[\hat{\theta}_T, \nu(\hat{\theta}_T)] = 0,$$

we propose two more user-friendly estimators, both associated with a sequence α_T of tuning parameters.

First, $\theta_T^{(1)}$ defined as solution of:

$$q_T[\theta_T^{(1)}, \nu(\tilde{\theta}_T)] + \frac{\partial q_T}{\partial \nu'}[\tilde{\theta}_T, \nu(\tilde{\theta}_T)] \frac{\partial \nu}{\partial \theta'}(\tilde{\theta}_T)(\theta_T^{(1)} - \tilde{\theta}_T) + \alpha_T \left\| \theta_T^{(1)} - \tilde{\theta}_T \right\|^2 e_p = 0. \tag{19}$$

Second, $\theta_T^{(2)}$ defined as solution of:

$$q_T[\theta_T^{(2)}, \tilde{\nu}_T\ ] + \frac{\partial q_T}{\partial \nu'}[\theta_T^{(2)}, \tilde{\nu}_T\ ].(\nu(\theta_T^{(2)}) - \tilde{\nu}_T\ ) + \alpha_T \left\| \nu(\theta_T^{(2)}) - \tilde{\nu}_T\ \right\|^2 e_p = 0. \tag{20}$$

Recall that several variants are possible, depending upon what part of the first-step estimator is used for the penalty term and/or for computing the derivative $\partial q_T / \partial \nu'$.

For practical choice of the tuning parameter sequence α_T, the two golden rules are as follows. First, for sake of asymptotic efficiency, α_T must go to infinity strictly slower than $\sqrt{T}$; Second, the fact that α_T goes to infinity is only useful to ensure consistency (see Step 1 in the proof of Proposition 3.1).

In many circumstances, consistency will be warranted even without the penalty, that is, with $\alpha_T = 0$. This, in particular, paves the way for many efficient two-step extremum estimators as exemplified in sections 2.2 and 2.3. Generally speaking, when consistency is not an issue, Theorem 2.1 states asymptotic efficiency of two-step extremum estimators $\hat{\theta}_T^{ext}$ computed as solutions of

$$\hat{\theta}_T^{ext} = \arg\max_{\theta} \left\{ Q_T[\theta, \tilde{\nu}_T] + \frac{\partial Q_T[\theta, \tilde{\nu}_T]}{\partial \nu'}.[\nu(\theta) - \tilde{\nu}_T] - \frac{1}{2}[\nu(\theta) - \tilde{\nu}_T]' J_T(\theta)[\nu(\theta) - \tilde{\nu}_T] \right\}. \tag{21}$$

Moreover, the proof of Theorem 2.1 shows that the dependence on θ of the weighting matrix $J_T(\theta)$ can be overlooked in computing the first order conditions and then, up to the penalty term, the first order conditions for (21) are very similar to (19). Sections 2.2 and 2.3 display user friendly closed form formulas for the weighting matrix $J_T(\theta)$ that do not involve any second derivatives. However, it is important to keep in mind that consistency is not always warranted, and then, the only solution is the introduction of the penalty term in first order conditions leading to (19) or (20).

4 Additive decomposition of Extremum Criterion

4.1 Efficient two-step Estimation via Margin Targeting

There exist many interesting situations in economics and finance where the extremum criterion takes the additively separable form

$$Q_T[\theta, \nu(\theta)] \quad = \quad Q_{1T}[\theta_1] + Q_{2T}[\theta_2, \nu(\theta)], \tag{22}$$

where $\theta = (\theta_1', \theta_2')', \nu(\theta) = \theta_1 \in \mathbb{R}^{p_1}, \theta_2 \in \mathbb{R}^{p_2}$ and $p_1 + p_2 = p$. This particular structure for $Q_T[\theta, \nu(\theta)]$ includes many nonlinear time series models, such as, the Dynamic Conditional Correlations (DCC-GARCH) model of Engle (2002), the rotated ARCH model of Noureldin et al. (2014), and many copula models. In these multivariate models θ_1 generally represents the

parameters that govern the marginal distributions and θ_2 represent the parameters that govern the dependence between the different components. In this framework, $\nu(\theta) = \theta_1$ represents the additional occurrences of θ_1 that show up in the dependence structure and complicate estimation of θ.

In this setting, a common way of estimating $\theta = (\theta_1', \theta_2')'$ is the so-called inference from the margins, where a root-T consistent estimator $\tilde{\theta}_T$ is obtained by first maximizing $Q_{1T}[\theta_1]$ to obtain $\tilde{\theta}_{1T}$, which is equivalent to solving the estimating equations

$$\frac{\partial Q_{1T}[\tilde{\theta}_{1T}]}{\partial \theta_1} = 0, \tag{23}$$

$\tilde{\theta}_{1T}$ then replaces the unknown θ_1 in $Q_{2T}[\theta_2, \theta_1]$ and $Q_{2T}[\theta_2, \tilde{\theta}_{1T}]$ is maximized to obtain $\tilde{\theta}_{2T}$, which is equivalent to solving

$$\frac{\partial Q_{2T}[\tilde{\theta}_{2T}, \tilde{\theta}_{1T}]}{\partial \theta_2} = 0. \tag{24}$$

If (23) and (24) are unbiased estimating equations for θ^0, in the sense that,

$$\begin{aligned} \lim_{T=\infty} \frac{\partial Q_{1T}[\theta_1]}{\partial \theta_1} &= 0 \Longleftrightarrow \theta_1 = \theta_1^0, \\ \lim_{T=\infty} \frac{\partial Q_{2T}[\theta_2, \theta_1^0]}{\partial \theta_2} &= 0 \Longleftrightarrow \theta_2 = \theta_2^0, \end{aligned}$$

$\tilde{\theta}_T = (\tilde{\theta}_{1T}', \tilde{\theta}_{2T}')'$ is generally a root-T consistent estimator of θ^0.

While computationally simple, the estimator $\tilde{\theta}_T$ is inefficient, which is seen by noting that the efficient estimator $\hat{\theta}_T$, the maximizer of $Q_T[\theta, \nu(\theta)]$, solves the estimating equations

$$q_T[\theta, \nu(\theta)] = \begin{bmatrix} q_{1T}[\theta, \nu(\theta)] \\ q_{2T}[\theta, \nu(\theta)] \end{bmatrix},$$

where

$$\begin{aligned} q_{1T}[\theta, \nu(\theta)] &= \frac{\partial Q_{1T}[\theta_1]}{\partial \theta_1} + \frac{\partial Q_{2T}[\theta_2, \nu(\theta)]}{\partial \nu}, \\ q_{2T}[\theta, \nu(\theta)] &= \frac{\partial Q_{2T}[\theta_2, \nu(\theta)]}{\partial \theta_2}. \end{aligned}$$

Computationally simple and efficient estimators can be obtained in this setting using the two-step estimators $\theta_T^{(1)}$ and $\theta_T^{(2)}$, defined in Section 3.2 as the solutions to the estimating equations $0 = h_T^{(1)}(\theta)$ and $0 = h_T^{(2)}(\theta)$ (with $h_T^{(1)}(\theta)$ and $\tilde{h}_T^{(2)}(\theta)$ given in (19) and (20) respectively), and first-step estimator $\tilde{\theta}_T$ defined by estimating equations (23) and (24).

Obtaining $\theta_T^{(1)}$ and $\theta_T^{(2)}$ when $Q_T[\theta, \nu(\theta)]$ is additively separable following (22) then requires

specializing the definitions of $h_T^{(1)}(\theta)$ and $h_T^{(2)}(\theta)$. To this end, for

$$h_T^{(1)}(\theta) = \begin{bmatrix} h_{1T}^{(1)}(\theta) \\ h_{2T}^{(1)}(\theta) \end{bmatrix}, \; h_T^{(2)}(\theta) = \begin{bmatrix} h_{1T}^{(2)}(\theta) \\ h_{2T}^{(2)}(\theta) \end{bmatrix},$$

we have that $\theta_T^{(1)}$, defined as the solution to $0 = h_T^{(1)}(\theta)$, solves

$$0 = h_{1T}^{(1)}(\theta_T^{(1)}) = \frac{\partial Q_{1T}[\theta_{1T}^{(1)}]}{\partial \theta_1} + \frac{\partial Q_{2T}[\theta_{2T}^{(1)}, \tilde{\theta}_{1T}]}{\partial \nu} + \frac{\partial^2 Q_{2T}[\tilde{\theta}_{2T}, \tilde{\theta}_{1T}]}{\partial \nu \partial \nu'}(\theta_{1T}^{(1)} - \tilde{\theta}_{1T}) + \alpha_T \left\| \theta_T^{(1)} - \tilde{\theta}_T \right\|^2 e_{p_1} \quad (25)$$

$$0 = h_{2T}^{(1)}(\theta_T^{(1)}) = \frac{\partial Q_{2T}[\theta_{2T}^{(1)}, \tilde{\theta}_{1T}]}{\partial \theta_2} + \frac{\partial^2 Q_{2T}[\tilde{\theta}_{2T}, \tilde{\theta}_{1T}]}{\partial \theta_2 \partial \nu'}(\theta_{1T}^{(1)} - \tilde{\theta}_{1T}) + \alpha_T \left\| \theta_T^{(1)} - \tilde{\theta}_T \right\|^2 e_{p_2} \quad (26)$$

and $\theta_T^{(2)}$, defined as the solution to $0 = h_T^{(2)}(\theta)$, solves

$$0 = h_{1T}^{(2)}(\theta_T^{(2)}) \frac{\partial Q_{1T}[\theta_{1T}^{(2)}]}{\partial \theta_1} + \frac{\partial Q_{2T}[\, \theta_{2T}^{(2)}, \tilde{\theta}_{1T}]}{\partial \nu} - \frac{\partial^2 Q_{2T}[\theta_{2T}^{(2)}, \tilde{\theta}_{1T}]}{\partial \nu \partial \nu'}(\, \theta_{1T}^{(2)} - \tilde{\theta}_{1T}) + \alpha_T \left\| \, \theta_{1T}^{(2)} - \tilde{\theta}_{1T} \right\|^2 e_{p_1} \quad (27)$$

$$0 = h_{2T}^{(2)}(\theta_T^{(2)}) = \frac{\partial Q_{2T}[\theta_{2T}^{(2)}, \tilde{\theta}_{1T}]}{\partial \theta_2} + \frac{\partial^2 Q_{2T}[\theta_{2T}^{(2)}, \tilde{\theta}_{1T}]}{\partial \theta_2 \partial \nu'}(\theta_{1T}^{(2)} - \tilde{\theta}_{1T}) + \alpha_T \left\| \theta_{1T}^{(2)} - \tilde{\theta}_{1T} \right\|^2 e_{p_2}, \quad (28)$$

for some sequence α_T going to infinity slower than $\sqrt{T}$.

Obviously, solving (25) and (26) (respectively, (27) and (28)) to obtain $\theta_T^{(1)}$ (respectively, $\theta_T^{(2)}$) is more computationally involved than the estimator $\tilde{\theta}_T$. However, both $\theta_T^{(1)}$ and $\theta_T^{(2)}$ share with $\tilde{\theta}_T$ the convenient feature that the cumbersome occurrence of θ_1 in $Q_{2T}[\theta_2, \theta_1]$ never shows up as an unknown parameter in the estimating equations, which makes our two-step efficient estimator computationally friendly in comparison with the brute force efficient estimator $\hat{\theta}_T$.

This simplification of the estimating equations is also shared by the MBP estimator proposed in SFK. When $Q_T[\theta, \nu(\theta)]$ is additively separable following (22), the MBP algorithm takes as its starting value $\tilde{\theta}_T$ and defines a sequence of iterative estimators $\hat{\theta}_T^{(k)}$, $k > 1$, by solving

$$0 = \frac{\partial Q_{1T}[\hat{\theta}_{1T}^{(k+1)}]}{\partial \theta_1} + \frac{\partial Q_{2T}[\hat{\theta}_{2T}^{(k)}, \hat{\theta}_{1T}^{(k)}]}{\partial \theta_1},$$

$$0 = \frac{\partial Q_{2T}[\hat{\theta}_{2T}^{(k+1)}, \hat{\theta}_{1T}^{(k)}]}{\partial \theta_2}.$$

While each iteration of the MBP procedure is computationally simpler than the second-step of the penalized two-step estimators, the price to pay for this simplicity is two-fold: one, to achieve efficiency we require $k \to \infty$, possibly according to a tuning parameter $k = k(T)$, and two, convergence of the MBP iterations requires the existence of a local contraction mapping condition, often called an information dominance condition.

If the information dominance condition is nearly unsatisfied, the MBP iterations converge very slowly, and if this condition is not satisfied $\hat{\theta}_T^{(k)}$ does not converge. To deal with such situations FPR propose a modification of the MBP estimator in SFK that regains a portion of the information associated with the occurrence of θ_2 in $Q_{2T}[\theta_2, \theta_1]$ neglected by the original

MBP scheme. Consequently, FPR define this alternative MBP estimator $\tilde{\theta}_T^{(k)}$ as the solution to the following estimating equations,

$$0 = \frac{\partial Q_{1T}[\tilde{\theta}_{1T}^{(k+1)}]}{\partial \theta_1} + \frac{\partial Q_{2T}[\tilde{\theta}_{2T}^{(k+1)}, \tilde{\theta}_{1T}^{(k)}]}{\partial \theta_1}, \tag{29}$$

$$0 = \frac{\partial Q_{2T}[\tilde{\theta}_{2T}^{(k+1)}, \tilde{\theta}_{1T}^{(k)}]}{\partial \theta_2}. \tag{30}$$

Note that this estimator is nothing but the MBP estimator conformable to the general definition (4).

It is straightforward to compare the computational burden associated with the MBP estimator in (29), (30) and the two-step penalized estimator $\theta_T^{(1)}$ (dubbed P-TS$_1$), as well as the additional two-step estimator $\theta_{T(P)}^{(1)}$ (dubbed TS$_1$) that arises from neglecting the penalty terms; i.e., the TS$_1$ estimator $\theta_{T(P)}^{(1)}$ solves the estimating equations (25, 26), but with $\alpha_T = 0$.[1] Firstly, comparing the MBP estimator and TS$_1$ (respectively, P-TS$_1$), the only difference between the two estimators is that TS$_1$ (respectively, P-TS$_1$) entails some minor computational burden associated with the introducing of a linear function of $\theta_{1T}^{(1)}$ (this statement holds up to the penalty term for P-TS$_1$). This tiny additional complexity is the price to pay to get efficiency in two steps instead of fishing for the limit of an iterative procedure, which, as stated above, may require many iterations depending on the strength of the local-contraction mapping.

However, when the local contraction mapping is strong, the MBP procedure of SFK is the simplest from a computational standpoint. As the required contraction mapping condition becomes weaker, the MBP estimator becomes more computationally burdensome.[2] In contrast, the two-step procedures discussed herein do not require a contraction mapping condition and can therefore yield consistent and efficient estimators in situations where this condition is violated.

In comparison with the aforementioned estimators, the penalized two-step estimator $\theta_T^{(2)}$ (dubbed P-TS$_2$), and the corresponding version $\theta_{T(P)}^{(2)}$ (dubbed TS$_2$) that neglects the penalty function, incurs additional computational complexity because θ_2 occurs within the partial Hessian term in the estimating equations. However, in this setting, the P-TS$_2$ (and TS$_2$) estimator is unique in that it only requires a consistent first-step estimator for θ_1^0, and not for θ_2^0. In the framework of estimation from the margins, this advantageous property of TS$_2$ (and P-TS$_2$) can be interpreted as follows. In many multivariate models, θ_1 can simply be estimated from the margins and is numerical stable. In contrast, estimation of the dependence parameters θ_2 is often tricky and numerically unstable. Indeed, this is a primary reason why (unconditional) variance targeting, as initially proposed by Engle and Mezrich (1996), became popular in the estimation of multivariate GARCH models, with similar reasoning leading researchers to contemplated correlation targeting in estimation of GARCH-DCC models. From a targeting standpoint, the P-TS$_2$ (TS$_2$) estimator first obtains a simple estimate $\tilde{\theta}_{1T}$ of θ_1^0 from the margins, then uses $\tilde{\theta}_{1T}$ via a "margin targeting" procedure whereby the second-step of the

[1] Note that, from Proposition 3.1 and Theorems 3.1 and 3.2, when consistent the two-step estimators that disregards the penalty term will also be asymptotically efficient.

[2] This statement also holds for the MBP estimator proposed in FPR.

estimation procedure is stabilized by targeting the consistent marginal parameter estimates.

In contrast to (unconditional) variance targeting, P-TS$_2$ (and TS$_2$) does not incur an efficiency loss associated with margin targeting. More importantly, P-TS$_2$ (and TS$_2$) need not maintain the problematic assumption in unconditional variance targeting on the existence of higher order unconditional moments, which is required in order to for variance targeting to yield an asymptotically normal estimator of the unconditional variance.

4.2 Bivariate Gaussian Copula Models

In the following subsection, we illustrate the above discussion between the different estimation procedures using a Gaussian Copula model. The Bivariate Gaussian copula model has been extensively studied in statistics and economics, see, e.g., Joe (1997), Song (2000), among others, and is often used in empirical analysis.

Assume our goal is to estimate the parameters governing the distribution of $\mathbf{y}_i = (y_{i,1}, y_{i,2})'$. Denoting the marginal distribution of $y_{i,j}$ as $F_j(\cdot;\alpha_j)$, where α_j is a vector of unknown parameters, the joint distribution can be constructed using a copula function $C(u_1, u_2; \rho)$, where ρ denotes the copula dependence parameter. In what follows, we assume $\mathbf{y}_i = (y_{i,1}, y_{i,2})'$ follows a bivariate Gaussian copula with cumulative distribution function (CDF)

$$C(F_1(y_{i,1};\alpha_1), F_2(y_{i,2};\alpha_2); \rho) = \Phi_\rho(\Phi^{-1}(F_1(y_{i,1};\alpha_1)), \Phi^{-1}(F_2(y_{i,2};\alpha_2))), \tag{31}$$

where $\Phi_\rho(\cdot)$ is the bivariate Gaussian cumulative distribution function with correlation parameter ρ and $\Phi(\cdot)$ is the standard normal CDF. Denote by $c(F_1(y_{i,1};\alpha_1), F_2(y_{i,2};\alpha_2); \rho)$ the copula density derived from equation (31). For $(u_1, u_2)' \in (0,1)^2$, Song (2000) demonstrates that the density of the bivariate Gaussian copula is

$$c(u_1, u_2; \rho) = \frac{1}{\sqrt{1-\rho^2}} \exp\left(-\frac{\rho(z_1^2 + z_2^2) - 2\rho(z_1 \cdot z_2)}{2(1-\rho^2)}\right),$$

where $z_j = \Phi^{-1}(u_j)$ for $j = 1, 2$.

Let $f_j(y_{i,j};\alpha_j)$ denote the marginal density of $y_{i,j}$ and define $\theta_1 = (\alpha_1', \alpha_2')'$, $\theta_2 = \rho$, with $\theta = (\theta_1', \theta_2)'$. Inference for θ in the Bivariate Gaussian copula model can be carried out using maximum likelihood, with corresponding log-likelihood function

$$Q_T[\theta, \nu(\theta)] = \sum_{i=1}^{T}\sum_{j=1}^{2} \log(f_j(y_{i,j};\alpha_j)) - \frac{T}{2}\log(1-\rho^2) - \frac{\rho}{2(1-\rho^2)}(\rho A(\theta_1) - 2B(\theta_1)). \tag{32}$$

Herein, $A(\theta_1) = \sum_{i=1}^{T}[z_{i,1}(\alpha_1)^2 + z_{i,2}(\alpha_2)^2]$, $B(\theta_1) = \sum_{i=1}^{T} z_{i,1}(\alpha_1) z_{i,2}(\alpha_2)$, and $z_{i,j}(\alpha_j) = \Phi^{-1}(F_j(y_{i,j};\alpha_j))$ for $j = 1, 2$. The likelihood in (32) is separable and we denote the two pieces

$$Q_{1T}[\theta_1] = \sum_{i=1}^{T}\sum_{j=1}^{2} \log(f_j(y_{i,j};\alpha_j)), \text{ and } Q_{2T}[\theta_2, \nu(\theta)] = -\frac{T}{2}\log(1-\rho^2) - \frac{\rho}{2(1-\rho^2)}(\rho A(\theta_1) - 2B(\theta_1)),$$

where, again, $\nu(\theta) = \theta_1$.

4.2.1 Estimators of θ

Depending on the specification of the marginals $f_j(\cdot;\alpha_j)$, maximizing $Q_T[\theta,\nu(\theta)]$ to obtain the Maximum Likelihood estimator (MLE) $\hat{\theta}_T$ can be difficult. In these cases a simple two-step estimation approach, the so-called inference from margins (IFM) approach, is often used to estimate θ (see, e.g., Shih and Louis (1995), Joe (1997) and Patton (2009) for examples and discussion). The IFM approach first maximizes $Q_{1T}[\theta_1] = \sum_{i=1}^{T}\sum_{j=1}^{2}\log(f_j(y_{i,j};\alpha_j))$ to obtain $\tilde{\theta}_{1T} = (\tilde{\alpha}'_{1T}, \tilde{\alpha}'_{2T})'$, defined as the solution to

$$0 = \frac{\partial Q_{1T}[\theta_1]}{\partial \theta_1} = \begin{pmatrix} \sum_{i=1}^{n} \frac{1}{f_1(y_{i,1};\alpha_1)} \frac{\partial f_1(y_{i,1};\alpha_1)}{\partial \alpha_1} \\ \sum_{i=1}^{n} \frac{1}{f_2(y_{i,2};\alpha_2)} \frac{\partial f_2(y_{i,2};\alpha_2)}{\partial \alpha_2} \end{pmatrix}.$$

Next, the unknown θ_1 in $Q_{2T}[\theta_2,\theta_1]$ is replaced with $\tilde{\theta}_{1T}$ and $Q_{2T}[\theta_2,\tilde{\theta}_{1T}] = -\frac{T}{2}\log(1-\rho^2) - \frac{\rho}{2(1-\rho^2)}(\rho A(\tilde{\theta}_{1T}) - 2B(\tilde{\theta}_{1T}))$ is maximized to obtain $\tilde{\theta}_{2T} = \tilde{\rho}_T$, defined as the solution to

$$0 = \frac{\partial Q_{2T}[\tilde{\theta}_{1T},\tilde{\theta}_{2T}]}{\partial \theta_2} = \frac{T\rho}{1-\rho^2} - \frac{1}{(1-\rho^2)^2}(\rho A(\tilde{\theta}_{1T}) - (1+\rho^2)B(\tilde{\theta}_{1T})).$$

It is clear from this decomposition that the IMF estimator disregard the information about θ_1 contained in

$$\frac{\partial Q_{2T}[\theta_2,\theta_1]}{\partial \theta_1} = -\sum_{i=1}^{n} \frac{\rho}{1-\rho^2} \begin{pmatrix} \rho\frac{\partial A(\theta_1)}{\partial \alpha_1} - 2\frac{\partial B(\theta_1)}{\partial \alpha_1} \\ \rho\frac{\partial A(\theta_1)}{\partial \alpha_2} - 2\frac{\partial B(\theta_1)}{\partial \alpha_2} \end{pmatrix}.$$

From the above definitions, we see that the efficient MBP and penalized two-step estimators obtain efficiency by adding back, in differing combinations, terms associated with $\partial Q_{2T}[\theta_2,\theta_1]/\partial\theta_1$. MBP accomplishes this task by adding back $\partial Q_{2T}[\theta_2,\theta_1]/\partial\theta_1$ to the estimating equations for θ_1 and iterating over the cumbersome occurrences of θ_1 (and θ_2, depending on the precise MBP method). On the other hand, the penalized two-step estimator $\theta_T^{(1)}$ (previously dubbed P-TS$_1$) linearizes $\partial Q_{2T}[\theta_2,\theta_1]/\partial\theta_1$, with respect to the cumbersome occurrence of θ_1, around the consistent estimator $\tilde{\theta}_{1T}$, and targets the second-step estimators using the initially consistent $\tilde{\theta}_T$. The penalized two-step estimator $\theta_T^{(2)}$ (previously dubbed P-TS$_2$) is similar to P-TS$_1$ but only penalizes the estimating equations with respect to the margins estimator $\tilde{\theta}_{1T}$. Both two-step approaches have the same asymptotic distribution, but can behave differently in finite samples.

In comparison with the two-step procedures, the critical regularity condition needed for the MBP estimator to be efficient is the satisfaction of a local contraction mapping condition, also termed the information dominance condition. However, in the bivariate Gaussian copula model, simulation evidence in SFK and Liu and Luger (2009) demonstrate that the MBP approach can behave poorly if there is even moderate correlation. Intuitively, this phenomena is present because as ρ increases the portions of the estimating equations that MBP iterates over become more informative for estimating the parameters. For ρ large enough the MBP algorithm neglects too much information and yields an inconsistent estimator.

4.2.2 Example: Exponential Marginals

In this subsection we compare the finite sample properties of the MBP approach of SFK and four different efficient two-step procedures: the penalized two-step estimator P-TS$_1$, the non-penalized counterpart to P-TS$_1$ given by TS$_1$, the partially penalized two-step estimator P-TS$_2$, and the non-penalized counterpart to P-TS$_2$ given by TS$_2$. Data for the exercise is generated from the Gaussian copula in the situation where the marginal densities are exponential: $f_j(y_{i,j};\alpha_j) = \alpha_j \exp(-\alpha_j y_{i,j}),\ \alpha_j > 0,\ j = 1, 2$.

In particular, the simulation study compares the effects of the correlation parameter and sample size on the various estimators. For the simulation study we set $\alpha_1 = .1$, $\alpha_2 = 1$ and consider three different values for the correlation parameter $\rho = .75, .95, .985$. Across the three values of ρ we consider three different sample sizes $T = 100, 200, 300$. For each T and ρ combination we create 1,000 synthetic samples.

It is important to note that for ρ greater than approximately .95 the information dominance condition associated with the proposed MBP procedure is no longer satisfied. Therefore, at high levels of correlation we expect the finite sample properties of the MBP estimator to be poor in comparison with the various two-stage estimators.

The estimators are compared in terms of their means, mean squared error (MSE) and mean absolute error (MAE), across the different sample sizes. We define convergence for the MBP algorithm as the maximum absolute difference across the parameters being less than $1.0e^{-05}$ for two or more successive iterations. Tables 1 to 3 report the averages over the 1,000 synthetic samples for the mean, MSE and MAE across the three correlation values $\rho = \{.75, .95, .985\}$. For the penalized two-step estimators the penalty term is taken proportional to $T^{1/4}$.

For low values of the correlation parameter the MBP algorithm and the efficient two-step estimators are very similar. However, as the correlation parameter increases, the penalized two-step methods give smaller MSEs and MAEs than the MBP estimator and non-penalized two-step estimator. With high correlation values and larger sample sizes the MBP algorithm encounters difficulty in estimation since the matrix driving the updates does not fulfill the IDC. It is important to point out that the same behavior is not found in the two-stage and penalized two-stage estimates, which perform well even for $\rho = .985$.

The various combinations of penalized and non-penalized two-step estimators all deliver stable parameter estimates with good finite sample properties. However, the fully penalized estimator P-TS$_1$ does seem to have a slight edge over the other estimators in terms of performance. The small impact of the penalty term in this situation is very easy to interpret: for copula models the IFM procedure often provides accurate starting values, and therefore the need to penalize is drastically reduced.[3] In other words, in the copula case, the two-step procedure merely ensures efficiency and penalization seems not to be required.

5 Efficient two-step estimation with Implied States

In this section we analyze situations where θ^0 is determined by the law of motion governing a latent stochastic process of interest $\{Y_t^* : t \geq 1\}$. The latent state variables Y_t^* are unobservable

[3]Recall, the penalty term is needed to rule out any perverse solutions to the estimating equations, which can exist because of the partial linearization.

to the econometrician, but are related to observed data Y_t through a function $h[\cdot, \nu^0]$, known up to the unknown parameters $\nu^0 = \nu(\theta^0)$, according to the relationship

$$Y_t = h[Y_t^*, \nu^0].$$

We are only interested in situations where $Y^* \mapsto g[Y^*, \nu]$ is one-to-one for any ν, which implies that, if ν^0 was known, Y_t^* could be directly obtained by inverting $h[\cdot, \nu^0]$; i.e.,

$$Y_t = h[Y_t^*, \nu^0] \iff Y_t^* = g[Y_t, \nu^0]. \tag{33}$$

When $\nu(\theta^0)$ is unknown, equation (33) defines the implied state (variable) $Y_t^*(\theta) = g[Y_t, \nu(\theta)]$.

As has been noted by several authors, such as, e.g., Renault and Touzi (1996), and Pastorello et al. (2003), the setup in (33) covers many interesting applications in economics and finance. However, estimation of θ^0 is often complicated by the nature of the function $h[\cdot, \nu]$ and the difficulties encountered when transforming the estimation problem from one based on latent states Y_t^*, to one based on implied states $g[Y_t, \nu(\theta)]$.

In what follows, we demonstrate that the efficient penalized two-step estimator can often be used to obtain consistent and efficient estimators for θ^0 in models with implied states. In particular, we focus on the use of implied states in GMM, so-called, Implied States GMM, and in likelihood models with latent states. A comparison with existing estimation approaches in these settings is also given.

5.1 Implied States GMM

Pan (2002) uses the terminology Implied States GMM (IS-GMM) to describe GMM estimation in the context of option pricing models with latent variables. More specifically, the IS-GMM estimator of Pan (2002) uses observed option price data to back-out, through an option pricing formula, the latent state variables driving the price process.

Formally, we are interested in analyzing a model with true parameter θ^0, defined as the unique zero of a vector of moment conditions derived from the law of motion for Y_t^*:

$$E[\Psi^*(Y_t^*, \theta)] = 0 \iff \theta = \theta^0. \tag{34}$$

Clearly, GMM estimation from (34) is not feasible since Y_t^* is unobservable. Implementation of GMM in this setting can, however, be carried out by substituting the implied states, say, $g[Y_t, \theta]$, which are obtained by inverting $Y_t = h[Y_t^*, \theta]$ to get $Y_t^* = g[Y_t, \theta]$, into (34). The existence of the one-to-one relationship in (33) is common in many arbitrage-based asset pricing models. For instance, in options pricing Y_t may be the observed option price, Y_t^* can represent the latent variables driving the price process and $h[Y_t^*, \theta]$ will be the pricing formula linking Y_t and Y_t^*.

Plugging the implied states $g[Y_t, \theta]$ into (34) yields

$$E[\Psi(Y_t, \theta, \nu(\theta))] = 0, \text{ with } \Psi(Y_t, \theta, \nu(\theta)) = \Psi^*(g[Y_t, \nu(\theta)], \theta). \tag{35}$$

The first occurrence of θ within (35) represents the original occurrences of θ in the moment conditions represented by the latent data, while $\nu(\theta)$ represents the occurrences of θ in the

implied states. This later occurrence of $\nu(\theta)$ in $\Psi(\cdot)$ is generally computationally cumbersome in comparison with the former occurrence of θ in $\Psi(\cdot)$.

When the moment conditions in (35) are overidentified, we take as our extremum criterion $Q_T[\theta, \nu(\theta)]$ the efficient two-step GMM criterion:

$$Q_T[\theta, \nu(\theta)] = -\bar{\Psi}_T[\theta, \nu(\theta)]' W_T^{-1}(\tilde{\theta}_T) \bar{\Psi}_T[\theta, \nu(\theta)],$$

where $\bar{\Psi}_T[\theta, \nu(\theta)] = \frac{1}{T}\sum_{t=1}^{T} \Psi(Y_t, \theta, \nu(\theta))$, $\tilde{\theta}_T$ is a preliminary consistent estimator, and $W_T^{-1}(\tilde{\theta}_T)$ is a consistent estimator for the long-run variance matrix of $\sqrt{T}\bar{\Psi}_T[\theta^0, \nu(\theta^0)]$. The efficient estimator $\hat{\theta}_T$ can then be defined as the (unique) zero of the estimating equations

$$0 = \left[\frac{\partial \bar{\Psi}_T[\theta, \nu(\theta)]}{\partial \theta} + \frac{\partial \nu'(\theta)}{\partial \theta}\frac{\partial \bar{\Psi}_T[\theta, \nu(\theta)]}{\partial \nu}\right]' W_T^{-1}(\tilde{\theta}_T)\bar{\Psi}_T[\theta, \nu(\theta)]. \tag{36}$$

Note that, the estimator $\hat{\theta}_T$ uses a consistent estimator for the selection matrix

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

In contrast, the simpler IS-GMM estimator θ_T^{IS} defined as the solution of

$$\max_{\theta} -\bar{\Psi}_T[\theta, \nu(\tilde{\theta}_T)]' W_T^{-1}(\tilde{\theta}_T)\bar{\Psi}_T[\theta, \nu(\tilde{\theta}_T)]$$

solves the estimating equations

$$0 = \left[\frac{\partial \bar{\Psi}_T[\theta, \nu(\tilde{\theta}_T)]}{\partial \theta}\right]' W_T^{-1}(\tilde{\theta}_T)\bar{\Psi}_T[\theta, \nu(\tilde{\theta}_T)], \tag{37}$$

and therefore employs a consistent estimator for the selection matrix

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

When $dim(\Psi) > \dim(\Theta)$, the selection matrix $\widetilde{\Gamma}(\theta^0)$ selects p linear combinations of the estimating equations in a suboptimal manner, and so θ_T^{IS} will be inefficient in general. Intuitively, the inefficiency of θ_T^{IS} is a direct consequence of the estimators disregard for the impact of the awkward occurrences $\nu(\theta)$ of θ on the selection matrix, through the Jacobian matrix.

Unlike the unconditional moment setting described herein, Pan (2002) considers the application of IS-GMM in the context of conditional moment restrictions. However, when it comes to optimal instruments, the same inefficiency issue will be faced if we overlook components of the Jacobian matrix associated with the occurrences of θ in the implied states.

Besides the above IS-GMM estimators, Pastorello et al. (2003) propose an iterative latent backfitting estimator that defines estimates $\tilde{\theta}_T^k$ through the iterations

$$\tilde{\theta}_T^{k+1} = \arg\max_{\theta} Q_T[\theta, \nu(\tilde{\theta}_T^k)].$$

Upon convergence, $\tilde{\theta}_T^k$ solves the estimating equations

$$0 = \left[\frac{\partial \bar{\Psi}_T[\theta, \nu(\theta)]}{\partial \theta}\right]' W_T^{-1}(\tilde{\theta}_T)\bar{\Psi}_T[\theta, \nu(\theta)], \tag{38}$$

and therefore, similar to θ_T^{IS}, the latent backfitting estimator of Pastorello et al. (2003) is inefficient when $dim(\Psi) > \dim(\Theta)$.

An alternative to directly solving (36) and the inefficient estimators that solve (37), (38), is the penalized two-step estimator developed herein. Clearly, we have at our disposal an initial consistent estimator $\tilde{\theta}_T$ of θ^0. Moreover, $\tilde{\theta}_T$ can also be used to consistently estimate the optimal instruments via

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

The existence of a consistent estimator for $\Gamma(\theta^0)$ allows us to define a new two-step estimator that utilizes $\Gamma_T(\tilde{\theta}_T)$ to simplify the existing two-step estimator $\theta_T^{(1)}$ defined in equation (19).

To this end, defining $q_T[\theta, \nu(\tilde{\theta}_T)] = \Gamma(\tilde{\theta}_T)\bar{\Psi}_T[\theta, \nu(\tilde{\theta}_T)]$, we can obtain a simplified efficient two-step estimator θ_T^{*IS}, in the spirit of $\theta_T^{(1)}$, by solving

$$0 = h_T^{*IS}(\theta) = q_T[\theta, \nu(\tilde{\theta}_T)] + \Gamma_T(\tilde{\theta}_T)\frac{\partial \bar{\Psi}_T[\tilde{\theta}_T, \nu(\tilde{\theta}_T)]}{\partial \nu}\frac{\partial \nu(\tilde{\theta}_T)}{\partial \theta'}(\theta - \tilde{\theta}_T) + \alpha_T\|\theta - \tilde{\theta}_T\|^2.$$

The main difference between θ_T^{*IS} and $\theta_T^{(1)}$, is that $\theta_T^{(1)}$ requires differentiating $\bar{\Psi}_T[\theta, \nu(\theta)]$ with respect to $\nu(\theta)$ and also the occurrences of $\nu(\theta)$ in $\Gamma_T(\theta)$. In other words, for the estimator θ_T^{*IS}, $\Gamma_T(\theta)$ is calculated once and is not altered thereafter. Efficiency of θ_T^{*IS} can be shown by a direct application of Theorem 3.1.

The two-step estimator θ_T^{*IS} is similar to the IS-GMM estimator developed in FPR, and defined by the sequence of estimators $\hat{\theta}_T^{(k)}$, the solutions of

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

In comparison, neither the two-step or MBP estimator actively search over the cumbersome occurrences of θ in $\nu(\theta)$. In this way, both approaches share some of the computational simplicity associated with the inefficient estimators θ_T^{IS} and $\tilde{\theta}_T^k$, however, in contrast both estimators retain efficiency (under certain conditions). The main difference between the two approaches is that the two-step approach directly corrects the information loss associated with not optimizing over the occurrences of θ due to $\nu(\theta)$ by forming a consistent estimator of these quantities, the FPR approach on the other hand only offers this correction as $k \to \infty$, and only if the required contraction condition is satisfied.[4] The price to pay for this two-step procedure is the additional computational cost induced by the linear term in $h_T^{*IS}(\theta)$ and the addition of a penalty to guarantee consistency.

[4]See FPR for a precise statement of this contraction mapping condition.

5.2 Implied States in Latent Likelihood

Let us now consider the case where the unobservable stochastic process $\{Y_t^* : t \geq 1\}$ is drawn from a transition density that is known up to the unknown θ^0, and let

$$\mathcal{P} = \{f(\cdot|\cdot;\theta) : \theta \in \Theta\}$$

denote the family of transition densities indexed by θ. Denoting the log-likelihood based on the unobservable latent state variables Y_t^* by

$$Q_T^*[\theta] = \frac{1}{T}\sum_{t=1}^{T} \ell(Y_t^*|Y_{t-1}^*;\theta), \text{ where } \ell(Y_t^*|Y_{t-1}^*;\theta) = \log(f(Y_t^*|Y_{t-1}^*;\theta)),$$

the implied states framework utilizes the relationship $Y_t = h[Y_t^*, \nu^0]$ to transform the estimation problem from one based on Y_t^* and $Q_T^*[\theta]$ to one based on Y_t. Using the implied states $g[Y_t, \nu(\theta)]$, obtained by inverting (33) at the value θ, and the Jacobian formula, the infeasible log-likelihood $Q_T^*[\theta]$ is transformed into the feasible log-likelihood

$$Q_T[\theta, \nu(\theta)] = \frac{1}{T}\sum_{t=1}^{T} \ell(g[Y_t, \nu(\theta)]|g[Y_{t-1}, \nu(\theta)];\theta) + \frac{1}{T}\sum_{t=1}^{T} \log|H_y g[Y_t, \nu(\theta)]|.$$

$|H_y g[Y_t, \nu(\theta)]|$ is the absolute value of the Jacobian for Y associated with the map $Y \mapsto g[Y, \nu(\theta)]$.

Estimation of θ^0 from $Q_T[\theta, \nu(\theta)]$ is often encountered in estimation of option pricing models, see, e.g., Renault and Touzi (1996), as well as credit risk models, see, e.g., Duan (1994). Maximization of $Q_T[\theta, \nu(\theta)]$ is generally much more difficult than would be maximization of $Q_T^*[\theta]$, if such maximization were indeed feasible.

It is clear that directly solving

$$0 = q_T[\theta, \nu(\theta)] = \frac{\partial Q_T[\theta, \nu(\theta)]}{\partial \theta} + \frac{\partial \nu'(\theta)}{\partial \theta}\frac{\partial Q_T[\theta, \nu(\theta)]}{\partial \nu}$$

can be cumbersome, as θ shows up in several places within $Q_T[\theta, \nu(\theta)]$ and in highly nonlinear ways. While the two-step procedures discussed herein can be applied to such settings, it is perhaps more informative to consider precise implementation of these estimators in a relatively simple example.

5.2.1 Example: Merton Credit Risk Model

To demonstrate the penalized two-step methodology in the situation of implied states likelihood estimation, we now consider estimation of the parameters in the structural credit risk model of Merton (1974).

Suppose that the firm's debt consists of a zero coupon bond with face value B and maturity date δ. Letting V_t denote the firm's unobservable market value at time-t, the firm's observable equity price can be interpreted as an European call option written on the firm's market value

with strike price B and maturity δ; i.e.,

$$S_\delta \equiv \max[V_\delta - B, 0]. \tag{39}$$

From (39) the observed equity prices $S_0, ..., S_T$ can be interpreted as option prices written on the firm's unobservable market values $V_0, ..., V_T$.

In the simplest case, the firm's unobservable market value is described as a Geometric Brownian Motion:

$$\frac{dV_t}{V_t} = \mu dt + \sigma dW_t, \tag{40}$$

where W_t is a standard Brownian motion. Equation (40) allows us to write the conditional likelihood of the sample path $(V_1, V_2, ..., V_T)$ given some initial value V_0 and historical parameters (μ, σ). The conditional log-likelihood function of the unobserved asset values is then given by

$$Q_T^*[\mu, \sigma^2] = -\frac{1}{2}\ln(2\pi\sigma^2) - \frac{1}{2T}\sum_{t=1}^{T}\frac{\left(\ln(V_t/V_{t-1}) - (\mu - \frac{1}{2}\sigma^2)\right)^2}{\sigma^2} - \frac{1}{T}\sum_{t=1}^{n}\ln\ V_t,$$

see, e.g., Duan (1994, 2000) and FPR for a discussion. Unfortunately, maximum likelihood estimation of (μ, σ) from $Q_T^*[\mu, \sigma^2]$ is not feasible since the sample path $(V_1, V_2, ..., V_T)$ is unobserved.

However, when the dynamics of the firm's market value are described by (40), the observable equity values can be related to the unobservable firm values through the Black and Scholes option pricing formula:

$$S_t = V_t\Phi(d_t) - B\exp(-r(\delta - t))\Phi(d_t - \sigma\sqrt{\delta - t}), \tag{41}$$

where $d_t(\sigma^2) = \ln(V_t/B) + (r + \frac{1}{2}\sigma^2)(\delta - t)/\sigma\sqrt{\delta - t}$, $\Phi(\cdot)$ is the standard normal CDF and r is the risk-free interest rate assumed to be deterministic and time-invariant. Letting $g[\cdot, \sigma^2]$ denote the inverse of the Black and Scholes option pricing formula, the unobserved firm values are related to the observed equity prices through

$$V_t = g[S_t, \sigma^2],$$

which can be obtained, at least numerically, from equation (41) and a given value of σ^2. Technically $g[\cdot, \sigma^2]$ depends on t through the time-to-maturity $(\delta - t)$, however, we eschew this dependence in favor of notational simplicity.

Therefore, even though V_t is unobserved, if σ^2 were known its value could be imputed from $V_t = g[S_t, \sigma^2]$ for each $t = 1, ..., T$. Given this fact, using $V_t = g[S_t, \sigma^2]$ and the Jacobian formula, we transform the log-likelihood from one based on V_t to one based on S_t. Following arguments in Duan (1994), the conditional log-likelihood based on observable equity values is given by

$$Q_T[\mu, \sigma^2] = -\frac{1}{2}\ln(2\pi\sigma^2) - \frac{1}{2T}\sum_{t=1}^{T}\frac{\left(R_t(\sigma^2) - (\mu - \frac{1}{2}\sigma^2)\right)^2}{\sigma^2} - \frac{1}{T}\sum_{t=1}^{n}\ln\ g(S_t, \sigma^2) - \frac{1}{T}\sum_{t=1}^{T}\ln\left(\Phi\left(d_t(\sigma^2)\right)\right),$$

where implicit returns

$$R_t(\sigma^2) = \ln(g[S_t, \sigma^2]) - \ln(g[S_{t-1}, \sigma^2]),$$

can be obtained using the Black and Scholes formula and a given value of σ^2. Estimation of (μ, σ^2) then proceeds by maximizing $Q_T[\mu, \sigma^2]$.

Since estimation of μ is not a priority the first-step is often to concentrate out μ, which yields

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

and the log-likelihood based on the observable equity values becomes

$$Q_T[\sigma^2] = -\frac{1}{2}\log(2\pi\sigma^2) - \frac{1}{2T}\sum_{t=1}^{T}\frac{\left(R_t(\sigma^2) - \bar{R}_T(\sigma^2)\right)^2}{\sigma^2} - \frac{1}{T}\sum_{t=1}^{T}\log\left(g[S_t, \sigma^2]\right) - \frac{1}{T}\sum_{j=1}^{T}\log\Phi\left(d_t(\sigma^2)\right).$$

$Q_T[\sigma^2]$ depends, in several places, on the structural relationship $g[S_t, \sigma^2]$, which makes directly maximizing $Q_T[\sigma^2]$ numerically unstable. As in section Section 5.1, we denote the problematic occurrences of σ^2 in $Q_T[\sigma^2]$ due to the structural relationship $g[S_t, \sigma^2]$ by $\nu(\sigma^2)$; note, $\nu(\sigma^2) = \sigma^2$ and the difference between the two occurrences of σ^2 is for notational purposes. The concentrated log-likelihood function then becomes

$$\begin{aligned} Q_T[\sigma^2, \nu(\sigma^2)] = &-\frac{1}{2}\ln(2\pi\sigma^2) - \frac{1}{2T}\sum_{t=1}^{T}\frac{\left(R_t(\nu(\sigma^2)) - \bar{R}_T(\nu(\sigma^2))\right)^2}{\sigma^2} - \frac{1}{T}\sum_{t=1}^{T}\ln\left(g[S_t, \nu(\sigma^2)]\right) \\ &- \frac{1}{T}\sum_{t=1}^{T}\ln\Phi\left(d_t(\nu(\sigma^2))\right). \end{aligned}$$

Defining

$$\tilde{\sigma}_T^2[\nu(\sigma^2)] = \frac{1}{T}\sum_{j=1}^{T}(R_j(\nu(\sigma^2)) - \bar{R}_T(\nu(\sigma^2)))^2 \text{ and } A_T[\nu(\sigma^2)] = 2\frac{\partial Q_T[\sigma^2, \nu(\sigma^2)]}{\partial \nu}\frac{\partial \nu(\sigma^2)}{\partial \sigma^2},$$

an estimator of σ^2 can be obtained as the solution to the log-likelihood first-order conditions

$$0 = -\frac{1}{\sigma^2} + \frac{1}{\sigma^4}\tilde{\sigma}_T^2[\nu(\sigma^2)] + A_T[\nu(\sigma^2)].$$

Solving the above equation is equivalent to solving the estimating equation $0 = q_T[\sigma^2, \nu(\sigma^2)]$, where

$$q_T[\sigma^2, \nu(\sigma^2)] = \sigma^4 A_T[\nu(\sigma^2)] - \sigma^2 + \tilde{\sigma}_T^2[\nu(\sigma^2)].$$

Directly solving $0 = q_T[\sigma^2, \nu(\sigma^2)]$ to estimate σ^2 can be cumbersome, and a popular alternative, due to Kealhofer, Mcquown and Vasicek and dubbed the KMV iterative method, is to

base estimation of σ^2 on

$$\tilde{\sigma}_T^2[\nu(\sigma^2)] = \frac{1}{T}\sum_{j=1}^{T}(R_j(\nu(\sigma^2)) - \bar{R}_T(\nu(\sigma^2)))^2.$$

Given a starting value $\hat{\sigma}^{2(1)}$, for $k > 1$, the KMV iterative method updates its estimates of σ^2 by calculating

$$\hat{\sigma}^{2(k)} = \tilde{\sigma}_T^2[\nu(\hat{\sigma}^{2(k-1)})] = \frac{1}{T}\sum_{t=1}^{T}(R_t(\nu(\hat{\sigma}^{2(k-1)})) - \bar{R}_T(\nu(\hat{\sigma}^{2(k-1)})))^2,$$

and iterating till convergence. This iterative procedure is often much simpler than one based on solving $q_T[\sigma^2, \nu(\sigma^2)] = 0$ since it completely neglects the influence of $A_T[\nu(\sigma^2)]$ on the estimates of σ^2. FPR demonstrate that the iterative KMV approach coincides with the latent backfitting estimator proposed by Pastorello et al. (2003) (hereafter, PPR).

While much simpler than maximum likelihood, the KMV/PPR estimator does not utilize all of the information in the estimating equation $q_T[\sigma^2, \nu(\sigma^2)]$ and therefore is not asymptotically equivalent to the MLE. To this end, FPR use the MBP approach to obtain an estimator that maintains some of the computational advantages of the KMV/PPR iterative strategy yet still delivers an estimator that is asymptotically equivalent to the MLE. Given an initial estimator $\hat{\sigma}^{2(1)}$, at the k-th iteration ($k > 1$) the MBP estimator solves the following second-order equation in σ^2:

$$\sigma^4 A_T[\nu(\hat{\sigma}^{2(k-1)})] - \sigma^2 + \tilde{\sigma}_T^2[\nu(\hat{\sigma}^{2(k-1)})] = 0. \tag{42}$$

An alternative to the KMV/PPR and MBP approaches is the two-step approach discussed herein. In this context, the two-step approach linearizes the estimating equations $q_T[\sigma^2, \nu(\sigma^2)]$, with respect to the cumbersome occurrences of $\nu(\sigma^2) = \sigma^2$, around an initially consistent estimator. For $\hat{\sigma}^{2(1)}$ an initial estimator of σ^2, the non-penalized two-step approach estimates σ^2 by solving

$$\begin{aligned} 0 = \sigma^4 A_T[\nu(\hat{\sigma}^{2(1)})] - \sigma^2 + \tilde{\sigma}_T^2[\nu(\hat{\sigma}^{2(1)})] + [\partial A_T[\nu(\hat{\sigma}^{2(1)})]/\partial\nu]\sigma^4(\sigma^2 - \hat{\sigma}^{2(1)}) \\ + [\partial\tilde{\sigma}_T^2[\nu(\hat{\sigma}^{2(1)})]/\partial\nu](\sigma^2 - \hat{\sigma}^{2(1)}), \end{aligned} \tag{43}$$

and, for α_T a penalty term, the penalized two-step approach estimates σ^2 by solving

$$\begin{aligned} 0 = \sigma^4 A_T[\nu(\hat{\sigma}^{2(1)})] - \sigma^2 + \tilde{\sigma}_T^2[\nu(\hat{\sigma}^{2(1)})] + [\partial A_T[\nu(\hat{\sigma}^{2(1)})]/\partial\nu]\sigma^4(\sigma^2 - \hat{\sigma}^{2(1)}) \\ + [\partial\tilde{\sigma}_T^2[\nu(\hat{\sigma}^{2(1)})]/\partial\nu](\sigma^2 - \hat{\sigma}^{2(1)}) + \alpha_T(\sigma^2 - \hat{\sigma}^{2(1)})^2. \end{aligned} \tag{44}$$

Note that the two-step estimators in equations (43) and (44) require solving a third-order equation in σ^2, whereas the MBP estimator in equation (42) solves a second-order equation. However, the two-step estimators solve *only one* third-order equation in σ^2, whereas the MBP estimator requires solving (potentially) *many* second-order equations in σ^2. The computational merits of both approaches will depend on the quality of the first-step estimator $\hat{\sigma}^{2(1)}$ and, in the

case of MBP, the strength of the contraction mapping guiding the iterations. For both estimation procedures a convenient starting value can be obtained using the KMV/PPR estimation procedure.

5.2.2 Simulation Example

To illustrate the usefulness of the efficient two-stage method in the context of the Merton credit risk model we devise a small Monte Carlo experiment comparing the MBP estimator with the penalized (respectively, non-penalized) two-step estimator. We construct 1,000 synthetic samples of 250 and 500 time series observations for daily returns. The firm's value trajectory is initialized at $10,000$ and the face value of the firm's debt is fixed at $B = 9,000$. The parameters are set to $\mu = .01$ and $\sigma^2 = .09$. We focus on estimation of σ^2 only and so we work directly with the concentrated log-likelihood function for both estimators.[5]

The MBP estimator is obtained using a Newton-Raphson approach to solve equation (42). The penalized (respectively, non-penalized) two-step estimator is obtained using a mix of bisection and interpolation and the penalty term satisfies $\alpha_T \propto T^{1/4}$. Both methods use starting values obtained from the KMV/PPR method. Across the 1,000 synthetic samples we calculate the mean, median, root mean squared error (RMSE) and mean absolute error (MAE) for the MBP estimator and the two-step estimators.

The results of the Monte Carlo experiment are contained in Table 5. Table 5 demonstrates that the two-step estimators and the MBP estimator have similar finite sample properties, with the penalized two-step estimator having significantly smaller RMSE and MAE. It is also important to point out that, as with the copula example in Section 4.2.2, the finite sample properties of the penalized and non-penalized two-step estimator are very similar.

6 Conclusion

The development of nonlinear dynamic models in financial econometrics has given rise to estimation problems that are often viewed as computationally difficult. This potential computational burden has led to the development of computationally light estimators whose starting point is often a simple consistent estimator of some instrumental parameters. This first step estimator can be used either for targeting the structural parameters (Indirect Inference a la Gourieroux, Monfort and Renault (1993)) or for simplifying estimating equations for the parameters of interest. More often than not, this simplification comes at the price of some loss in efficiency. Not only do two-step estimators have, in general, an asymptotic distribution that depends on the distribution of the first step estimator, but even iterations may not be able to restore efficiency (see PPR and references therein).

FPR demonstrate that the aforementioned inefficiency is caused by disregarding the information contained in (some of) the awkward occurrences of the parameters in the criterion function. Popular iterative (or two-step) procedures are devised precisely to allow us to overlook these awkward occurrences, possibly at the cost of efficiency loss. The goal of FPR was to propose efficient iterative estimation procedures whose computational cost, at each step of

[5]The proposed simulation design is similar to that of FPR.

the iteration, is no higher than those of popular inefficient inference procedures. This goal was made possible by the fact that their algorithms iterate on the occurrences of the parameters that researchers would like to overlook. In this way, the informational content of these occurrences was no longer ignored, at least in the limit of the iterative procedure.

In the present paper, we replace the method of iteration by a partial linearization of the estimating equations around a first step consistent estimator for the parameters that are difficult to deal with. With respect to the efficient iterative procedure of FPR, the pros and cons of our two step procedure are as follows.

On the one hand, our approach is not required to compute a sequence of estimators but only a second step estimator. Our second step, in general, maintains the computational simplicity associated with each step of the FPR iterations. Moreover, while consistency of the FPR iterations may break down when their so-called Information Dominance condition is not fulfilled, our approach does not require such a condition.

On the other hand, linearization, when it is only partial, may be a risky exercise because it may deliver a solution for the non-linearized portion that is biased, even asymptotically, by the approximated linearization. Then the consistency property of the estimator may be lost. In order to hedge against this risk, we develop a strategy of targeting first step consistent estimators, in the spirit of indirect inference. However, in contrast with indirect inference, targeting is for us only a complementary tool for enforcing consistency. In particular, we don't want the asymptotic variance of our second step estimator to be inefficiently driven by the first step estimator used for targeting. This is the reason why we must elicit a tuning parameter (the penalty weight) that goes to infinity, in order to enforce consistency, but not too fast in order to avoid the efficiency loss that would be produced by contamination of the second step estimator by the inefficiency of the first step estimator.

Finally, it is worth noting that the strategy developed in this paper may be of more general interest. While indirect inference has demonstrated the usefulness of targeting instrumental parameters for simple identification of structural parameters of interest, the recent literature on multivariate GARCH has stressed that targeting some unconditional moments may be a safe way to hedge against the risk of numerical instability associated with supposedly efficient estimators, at least in the presence of high dimensional and/or highly nonlinear optimization problems. In a companion paper, we demonstrate that for multivariate GARCH models, in contrast to existing targeting strategies, our penalization/targeting approach can deliver numerically stable estimates with good finite sample properties without the need to sacrifice efficiency. Moreover, as pointed out in our copula example, in addition to unconditional moments, the relatively simple and robust estimators of the marginal distributions can often provide a useful target.

References

T. Bollerslev and J. M. Wooldridge. Quasi-maximum likelihood estimation and inference in dynamic models with time-varying covariances. *Econometric Reviews*, 11(2):143–172, 1992.

B. Crepon, F. Kramarz, and A. Trognon. Parameters of interest, nuisance parameters and

orthogonality conditions an application to autoregressive error component models. *Journal of Econometrics*, 82(1):135 – 156, 1997.

J.-C. Duan. Maximum likelihood estimation using price data of the derivative contract. *Mathematical Finance*, 4(2):155–167, 1994.

J.-C. Duan. Correction: Maximum likelihood estimation using price data of the derivative contract. *Mathematical Finance*, 10(4):461–462, 2000.

R. Engle. Dynamic conditional correlation. *Journal of Business an Economic Statistics*, 20(3): 339–350, 2002.

R. Engle and J. Mezrich. Garch for groups. *RISK*, 9(8):36 – 40, 1996.

Y. Fan, S. Pastorello, and E. Renault. Maximization by parts in extremum estimation. *The Econometrics Journal*, 2015. Forthcoming.

C. Gourieroux, A. Monfort, and E. Renault. Indirect inference. *Journal of Applied Econometrics*, 8:S85–S118, 1993. Special Issue on Econometric Inference Using Simulation Techniques.

C. Gourieroux, A. Monfort, and E. Renault. Two-stage generalized moment method with applications to regressions with heteroscedasticity of unknown form. *Journal of Statistical Planning and Inference*, 50(1):37 – 6193, 1996. Econometric Methodology, Part III.

L. P. Hansen, J. Heaton, and A. Yaron. Finite-sample properties of some alternative gmm estimators. *Journal of Business an Economic Statistics*, 14(3):pp. 262–280, 1996.

H. O. Hartley. The modified gauss-newton method for the fitting of non-linear regression functions by least squares. *Technometrics*, 3(2):269–280, 1961.

M. Hatanaka. An efficient two-step estimator for the dynamic adjustment model with autoregressive errors. *Journal of Econometrics*, 2(3):199–220, 1974.

H. Joe. *Multivariate models and dependence concepts*, volume 73. London: Chapman an Hall, 1997.

Y. Liu and R. Luger. Efficient estimation of copula-garch models. *Computational Statistics an Data Analysis*, 53(6):2284 – 2297, 2009. The Fourth Special Issue on Computational Econometrics.

R. C. Merton. On the pricing of corporate debt: The risk structure of interest rates. *The Journal of Finance*, 29(2):449–470, 1974.

W. K. Newey and D. L. McFadden. Large sample estimation and hypothesis testing. In R. F. Engle and D. L. McFadden, editors, *Handbook of Econometrics*, volume 4, pages 2111 – 2245. 1994.

D. Noureldin, N. Shephard, and K. Sheppard. Multivariate rotated ARCH models. *Journal of Econometrics*, 179(1):16 – 30, 2014.

A. Pagan. Two stage and related estimators and their applications. *The Review of Economic Studies*, 53(4):517–538, 1986.

A. Pakes and D. Pollard. Simulation and the asymptotics of optimization estimators. *Econometrica*, 57(5):pp. 1027–1057, 1989.

J. Pan. The jump-risk premia implicit in options: evidence from an integrated time-series study. *Journal of Financial Economics*, 63(1):3 – 50, 2002.

S. Pastorello, V. Patilea, and E. Renault. Iterative and recursive estimation in structural nonadaptive models [with comments, rejoinder]. *Journal of Business an Economic Statistics*, 21(4):449–482, 2003.

A. J. Patton. Copula-based models for financial time series. In T. Andersen, R. Davis, J.-P. Kreiss, and T. Mikosch, editors, *Handbook of Financial Time Series*, pages 767 – 785. Springer Verlag, 2009.

E. Renault and N. Touzi. Option hedging and implied volatilities in a stochastic volatility model1. *Mathematical Finance*, 6(3):279–302, 1996.

P. M. Robinson. The stochastic difference between econometric statistics. *Econometrica*, 56 (3):531–548, 1988.

J. H. Shih and T. A. Louis. Inferences on the association parameter in copula models for bivariate survival data. *Biometrics*, 51(4):1384–1399, 1995.

P. X.-K. Song. Multivariate dispersion models generated from gaussian copula. *Scandinavian Journal of Statistics*, 27(2):305–320, 2000.

P. X.-K. Song, Y. Fan, and J. D. Kalbfleisch. Maximization by parts in likelihood inference [with comments, rejoinder]. *Journal of the American Statistical Association*, 100(472):1145–1167, 2005.

A. Trognon and C. Gourieroux. A note on the efficiency of two-step estimation methods. In *Essays in Honor of Edmond Malinvaud*, pages 233–248. MIT Press, 1990.

A Regularity Conditions for Extremum Estimators

In all the applications considered in this paper, the estimating equations $f_T(\theta) = q_T[\theta, \nu(\theta)]$ of interest are obtained as first order conditions of some extremum estimation program:

$$\hat{\theta}_T = \arg\max_{\theta \in \Theta} Q_T[\theta, \nu(\theta)] \tag{45}$$

so that

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

It is worth noting that by contrast with the possibly more general framework mentioned in the introduction, we have introduced a simplification by considering only a fixed known function $\nu(\theta)$ instead of a more general sample dependent function $\nu_T(\theta)$. This may look restrictive since the function $\nu_T(\theta)$ may typically show up when profiling out some specific occurrences of some components of θ and thus be computed as data dependent. However, it must be kept in mind that the difference is more notational than real since a general objective function $Q_T^*[\theta, \nu_T(\theta)]$ may always be rewritten $Q_T[\theta, \theta]$ with a new function defined from $Q_T^*[.,.]$ and $\nu_T(.)$ by:

$$Q_T[\theta, \theta^*] = Q_T^*[\theta, \nu_T(\theta^*)] \tag{46}$$

This remark actually shows that we could always choose $\nu(\theta) = \theta$. We prefer to keep the notation $\nu(\theta)$ for the sake of notational transparency. In most cases, $\nu(\theta)$ will be nothing but a sub-vector of θ. However, while we will keep in mind that (45) is actually not less general than (46), we will make explicit how the regularity conditions must be interpreted when $\nu_T(\theta)$ is actually a sample-dependent consistent estimator of some underlying unknown true $\nu^0(\theta)$.

In the simple set up of (45),the maintained regularity conditions are the following.

R1. The following are satisfied:

(1) $\Theta \subset \mathbb{R}^p$ and $\Gamma \subset \mathbb{R}^q$ are two compact parameters spaces.

(2) $\nu(.)$ is a continuous function from Θ to Γ, twice continuously differentiable on the interior of Θ.

(3) $\theta^0 \in Int(\Theta)$, interior set of Θ, and $\nu^0 = \nu(\theta^0) \in Int(\Gamma)$, interior set of Γ.

R2. $Q_T[\theta, \nu]$ converges in probability towards a nonstochastic function $Q_\infty[\theta, \nu]$ uniformly on $(\theta, \nu) \in \Theta \times \Gamma$.

R3. The function $\theta \longmapsto Q_\infty[\theta, \nu(\theta)]$ attains a unique global maximum on Θ at $\theta = \theta^0$, unique solution of the equations $q_\infty[\theta, \nu(\theta)] = 0$, where

$$q_\infty[\theta, \nu(\theta)] = \frac{\partial Q_\infty[\theta, \nu(\theta)]}{\partial \theta} + \frac{\partial \nu'(\theta)}{\partial \theta} \frac{\partial Q_\infty[\theta, \nu(\theta)]}{\partial \nu}$$

R4. The function $Q_T[\theta, \nu]$ is twice continuously differentiable on $\mathring{\Theta} \times \mathring{\Gamma}$.

R5. The following are satisfied

(1) With $\lambda' = (\theta', \nu')$, the second derivative $\frac{\partial^2 Q_T(\lambda)}{\partial \lambda \partial \lambda'}$ converges uniformly on $\lambda \in \mathring{\Theta} \times \mathring{\Gamma}$ towards a non stochastic matrix $D(\lambda)$.

(2) The matrix $D_{\theta\theta}(\lambda^0) = P\lim_{T=\infty} \frac{\partial^2 Q_T(\lambda^0)}{\partial \theta \partial \theta'}$ (where $\lambda^{0\prime} = (\theta^{0\prime}, \nu^{0\prime})$) is negative definite.

R6. $\sqrt{T}\left[\frac{\partial Q_T(\lambda^0)}{\partial \theta} + \frac{\partial \nu'}{\partial \theta}(\theta^0).\frac{\partial Q_T(\lambda^0)}{\partial \nu}\right]$ converges in distribution towards a normal distribution with zero mean and variance Ω.

It is worth reinterpreting these regularity conditions when the objective function Q_T is actually deduced from another function Q_T^* as in (46). Note that in this case, $\nu(.)$ is just the identity function ($\nu(\theta) = \theta, \Theta = \Gamma$), making trivial all maintained assumptions about ν. However, it must be kept in mind that the role of the data dependent function $\nu_T(.)$ will typically be the consistent estimation of some true unknown function $\nu^0(.)$. Then, the above regularity conditions can be rewritten identical by only replacing the functions $Q_T[\theta, \nu]$ and $\nu(.)$ by the functions $Q_T^*[\theta, \nu]$ and $\nu^0(.)$. Only the limit arguments involving the function $\nu^0(.)$ have to be revisited to take into account its consistent estimation. We will basically rewrite condition R2 and R6 as follows:

R2*. The following are satisfied:

(1) $Q_T^*[\theta, \nu]$ converges in probability towards a non-stochastic function $Q_\infty^*[\theta, \nu]$ uniformly on $(\theta, \nu) \in \Theta \times \Gamma$.

(2) $\nu_T(\theta)$ converges in probability towards $\nu^0(\theta)$ uniformly on $\theta \in \Theta$.

R6*. $\sqrt{T}\left[\frac{\partial Q_T(\theta^0,\theta^0)}{\partial \theta} + \frac{\partial Q_T(\theta^0,\theta^0)}{\partial \theta^*}\right]$, where $Q_T[\theta, \theta^*] = Q_T^*[\theta, \nu_T(\theta^*)]$,converges in distribution towards a normal distribution with zero mean and variance Ω.

Obviously, a more primitive condition for R6* should be based of an assumption of joint asymptotic normality, involving not only the score function but also $\sqrt{T}\left(\nu_T(\theta^0) - \nu^0(\theta^0)\right)$ whose impact on the asymptotic distribution would be deduced from a Taylor expansion

$$\sqrt{T}\frac{\partial Q_T^*(\theta^0, \nu_T(\theta^0))}{\partial \lambda} = \sqrt{T}\frac{\partial Q_T^*(\theta^0, \nu^0(\theta^0))}{\partial \lambda} + \frac{\partial^2 Q_T^*(\theta^0, \nu^0(\theta^0))}{\partial \lambda \partial \nu'}.\sqrt{T}\left(\nu_T(\theta^0) - \nu^0(\theta^0)\right)$$

While this more specific set up would not introduce any theoretical complication, we omit it throughout for sake of expositional simplicity.

B Proofs

B.1 Proof of Theorem 2.1

Part (i) Asymptotic equivalence between $\hat{\theta}_T$ and $\hat{\theta}_T^*$:

The first order conditions that characterize $\hat{\theta}_T^*$ can be written:

$$q_T^*[\hat{\theta}_T^* \, , \tilde{\nu}_T] = 0$$

with

$$q_T^*[\theta, \tilde{\nu}_T] = \frac{\partial Q_T[\theta, \tilde{\nu}_T]}{\partial \theta} + \frac{\partial^2 Q_T[\theta, \tilde{\nu}_T]}{\partial \theta \partial \nu'}.\left[\nu(\theta) - \tilde{\nu}_T\right] + \frac{\partial \nu'(\theta)}{\partial \theta}.\frac{\partial Q_T[\theta, \tilde{\nu}_T]}{\partial \nu'} - \frac{\partial \nu'(\theta)}{\partial \theta} J_T(\theta^0)\left[\nu(\theta) - \tilde{\nu}_T\right]$$

Thus, by comparing with (2):

$$q_T^*[\theta, \tilde{\nu}_T] = q_T[\theta, \tilde{\nu}_T] + \frac{\partial q_T[\theta, \tilde{\nu}_T]}{\partial \nu'}.\left[\nu(\theta) - \tilde{\nu}_T\right] - \xi_T(\theta)$$

with

$$\xi_T(\theta) = \frac{\partial \nu'(\theta)}{\partial \theta} \left[\frac{\partial^2 Q_T[\theta, \tilde{\nu}_T]}{\partial \nu \partial \nu'} + J_T(\theta^0) \right] [\nu(\theta) - \tilde{\nu}_T]$$

Hence,

$$0 = q_T[\hat{\theta}_T^*, \tilde{\nu}_T] + \frac{\partial q_T[\hat{\theta}_T^*, \tilde{\nu}_T]}{\partial \nu'} . \left[\nu(\hat{\theta}_T^*) - \tilde{\nu}_T \right] - \xi_T(\hat{\theta}_T^*)$$

with

$$\xi_T(\hat{\theta}_T^*) = o_P\left(1/\sqrt{T}\right)$$

by virtue of Assumption A3, since $\hat{\theta}_T^*$ is root-T consistent. We will see in section 3 that, whenever consistent, an estimator $\mathring{\theta}_T$ solution of

$$h_T(\mathring{\theta}_T) = 0$$

with

$$h_T(\theta) = q_T[\theta, \tilde{\nu}_T] + \frac{\partial q_T[\theta, \tilde{\nu}_T]}{\partial \nu'} . [\nu(\theta) - \tilde{\nu}_T]$$

being asymptotically equivalent to $\hat{\theta}_T$. Thus, by application of Theorem 3.3 of Pakes and Pollard (1989), it is also the case for a solution $\hat{\theta}_T^*$ of

$$h_T(\hat{\theta}_T^*) = o_P\left(1/\sqrt{T}\right).$$

Note that we can apply Theorem 3.3 of Pakes and Pollard (1989) in particular because, by virtue of Assumptions A2 and A3, $\sqrt{T} h_T(\theta^0)$ is asymptotically normal.

Part (ii): Asymptotic equivalence between $\hat{\theta}_T^*$ and $\hat{\theta}_T^{ext}$

By definition, $\hat{\theta}_T^{ext}$ is solution of first order conditions:

$$g_T(\hat{\theta}_T^{ext}) = 0$$

such that

$$g_T(\hat{\theta}_T^*) = o_P\left(1/\sqrt{T}\right),$$

since $g_T(\hat{\theta}_T^*)$ is a p-dimensional vector whose component $j = 1, ..., p$ is

$$\left[\nu(\hat{\theta}_T^*) - \tilde{\nu}_T \right]' J_{jT}(\hat{\theta}_T^*) \left[\nu(\hat{\theta}_T^*) - \tilde{\nu}_T \right].$$

where $J_{j;T}(\theta)$ stands for the matrix of partial derivatives with respect to θ_j of all the coefficients of the matrix $J_T(\theta)$. Then, the announced asymptotic equivalence follows again by application of Theorem 3.3 of Pakes and Pollard (1989).

B.2 Proof of Proposition 3.1

Step 1: We show that $\tilde{\theta}_T^P$ is a consistent estimator of θ^0.

By definition:

$$0 = q_T[\tilde{\theta}_T^P, \nu(\tilde{\theta}_T)] + \frac{\partial q_T}{\partial \nu'}[\tilde{\theta}_T^P, \nu(\tilde{\theta}_T)]\frac{\partial \nu}{\partial \theta'}(\tilde{\theta}_T)(\tilde{\theta}_T^P - \tilde{\theta}_T) + \alpha_T \left\|\tilde{\theta}_T^P - \tilde{\theta}_T\right\|^2 e_p \tag{47}$$

Since the parameter space is compact, we only have to show that for any subsequence of $\tilde{\theta}_T^P$ that converges in probability towards some limit value $\bar{\theta}$, we necessarily have $\bar{\theta} = \theta^0$. By the regularity conditions (continuity and uniform convergence) we deduce from (47) that

$$0 = q_\infty[\bar{\theta}, \nu(\theta^0)] + \frac{\partial q_\infty}{\partial \nu'}[\bar{\theta}, \nu(\theta^0)]\frac{\partial \nu}{\partial \theta'}(\theta^0)(\bar{\theta} - \theta^0) + P\lim_{T=\infty} \alpha_T \left\|\tilde{\theta}_T^P - \tilde{\theta}_T\right\|^2 e_p. \tag{48}$$

Since $\text{Plim}_{T=\infty}\tilde{\theta}_T = \theta^0$ and $\lim_{T=\infty} \alpha_T = \infty$, (48) implies that $\text{Plim}_{T=\infty}\tilde{\theta}_T^P = \theta^0$.

Step 2: We show that

$$\hat{\theta}_T - \tilde{\theta}_T^P = O_P\left(\left\|f_T(\hat{\theta}_T) - \tilde{h}_T^P(\hat{\theta}_T)\right\|\right) = O_P\left(\left\|\tilde{h}_T^P(\hat{\theta}_T)\right\|\right)$$

This result is a direct consequence of Robinson (1988) Theorem 1 if we can show that the function $\tilde{h}_T^P(\theta)$ is conformable to Robinson's Assumption A2. We have

$$\begin{aligned}\frac{\partial \tilde{h}_T^P(\theta)}{\partial \theta'} &= \frac{\partial q_T[\theta, \nu(\tilde{\theta}_T)]}{\partial \theta'} + \frac{\partial}{\partial \theta'}\left[\frac{\partial q_T}{\partial \nu'}[\theta, \nu(\tilde{\theta}_T)]\right]\left[\frac{\partial \nu}{\partial \theta'}(\tilde{\theta}_T)(\theta - \tilde{\theta}_T) \otimes Id_p\right] \\ &\quad + \frac{\partial q_T}{\partial \nu'}[\theta, \nu(\tilde{\theta}_T)]\frac{\partial \nu}{\partial \theta'}(\tilde{\theta}_T) + 2\alpha_T\left(\theta - \tilde{\theta}_T\right)'\end{aligned}$$

where, for a $(p \times q)$ matrix A whose coefficients are functions of θ, we define $\partial A/\partial \theta'$ as the $(p \times qp)$ matrix

$$\left[\begin{array}{cccc} \frac{\partial A^1}{\partial \theta'} & \frac{\partial A^2}{\partial \theta'} & \ldots & \frac{\partial A^q}{\partial \theta'} \end{array}\right]$$

where $A^1, A^2, ..., A^q$ stands for the q columns of the matrix A. Since , by assumption, $\left\|\tilde{\theta}_T - \theta^0\right\| = o_P(1/\alpha_T)$, we deduce that, under regularity conditions

$$P\lim \frac{\partial h_T(\theta^0)}{\partial \theta'} = \frac{\partial q_\infty[\theta^0, \nu(\theta^0)]}{\partial \theta'} + \frac{\partial q_\infty}{\partial \nu'}[\theta^0, \nu(\theta^0)]\frac{\partial \nu}{\partial \theta'}(\theta^0) = F,$$

that is by assumption a non-singular matrix. Therefore, we get Assumption A2 of Robinson (1988) under standard regularity conditions.

Step 3: We show that

$$\hat{\theta}_T - \tilde{\theta}_T^P = O_P\left(\alpha_T \left\|\hat{\theta}_T - \tilde{\theta}_T\right\|^2\right).$$

We have

$$\begin{aligned} f_T(\hat{\theta}_T) &= q_T[\hat{\theta}_T, \nu(\hat{\theta}_T)] \\ &= q_T[\hat{\theta}_T, \nu(\tilde{\theta}_T)] + \frac{\partial q_T}{\partial \nu'}[\hat{\theta}_T, \nu(\tilde{\theta}_T)]\frac{\partial \nu}{\partial \theta'}(\tilde{\theta}_T)(\hat{\theta}_T - \tilde{\theta}_T) + O_P\left(\left\|\hat{\theta}_T - \tilde{\theta}_T\right\|^2\right) \\ &= \tilde{h}_T^P(\hat{\theta}_T) + O_P\left(\left\|\hat{\theta}_T - \tilde{\theta}_T\right\|^2\right) - \alpha_T\left\|\hat{\theta}_T - \tilde{\theta}_T\right\|^2 e_p. \end{aligned}$$

Therefore,

$$\left\|f_T(\hat{\theta}_T) - \tilde{h}_T^P(\hat{\theta}_T)\right\| = O_P\left(\alpha_T\left\|\hat{\theta}_T - \tilde{\theta}_T\right\|^2\right),$$

which gives the announced result by using the result of Step 2.

B.3 Proof Theorem 3.1

We just show that the proof of Proposition 3.1. will go through with very minor changes. The proof of consistency (**Step 1**) is the same except that equation (48) must now be replaced by

$$0 = q_\infty[\bar{\theta}, \nu(\theta^0)] + \frac{\partial q_\infty}{\partial \nu'}[\theta^0, \nu(\theta^0)]\frac{\partial \nu}{\partial \theta'}(\theta^0)(\bar{\theta} - \theta^0) + \text{Plim}_{T=\infty}\alpha_T\left\|\theta_T^{**} - \tilde{\theta}_T\right\|^2 e_p \qquad (49)$$

Obviously, the same consistency argument is a fortiori still valid. Since $\text{Plim}_{T=\infty}\tilde{\theta}_T = \theta^0$ and $\lim_{T=\infty}\alpha_T = \infty$, (49) implies that $\text{Plim}_{T=\infty}\theta_T^{(1)} = \theta^0$. With this new way to partially linearize, the Jacobian of the estimating equation is simplified as follows

$$\frac{\partial h_T^{(1)}(\theta)}{\partial \theta'} = \frac{\partial q_T[\theta, \nu(\tilde{\theta}_T)]}{\partial \theta'} + \frac{\partial q_T}{\partial \nu'}[\tilde{\theta}_T, \nu(\tilde{\theta}_T)]\frac{\partial \nu}{\partial \theta'}(\tilde{\theta}_T) + 2\alpha_T\left(\theta - \tilde{\theta}_T\right)'$$

Thus, we still have

$$\lim \frac{\partial h_T^{(1)}(\theta^0)}{\partial \theta'} = \frac{\partial q_\infty[\theta^0, \nu(\theta^0)]}{\partial \theta'} + \frac{\partial q_\infty}{\partial \nu'}[\theta^0, \nu(\theta^0)]\frac{\partial \nu}{\partial \theta'}(\theta^0) = F$$

and thus we can prove a **Step 2** exactly as in Proposition 3.1. This Step 2 will tell us that

$$\hat{\theta}_T - \theta_T^{**} = O_P\left(\left\|f_T(\hat{\theta}_T) - h_T^*(\hat{\theta}_T)\right\|\right).$$

We already know from Proposition 3.1 that

$$\left\|f_T(\hat{\theta}_T) - \tilde{h}_T^P(\hat{\theta}_T)\right\| = O_P\left(\alpha_T\left\|\hat{\theta}_T - \tilde{\theta}_T\right\|^2\right)$$

Thus, the triangle inequality will give the result if we can also show that

$$\left\|\tilde{h}_T^P(\hat{\theta}_T) - h_T^{(1)}(\hat{\theta}_T)\right\| = O_P\left(\alpha_T\left\|\hat{\theta}_T - \tilde{\theta}_T\right\|^2\right)$$

We have

$$\tilde{h}_T^P(\hat{\theta}_T) - h_T^{(1)}(\hat{\theta}_T) = \left[\frac{\partial q_T}{\partial \nu'}[\hat{\theta}_T, \nu(\tilde{\theta}_T)] - \frac{\partial q_T}{\partial \nu'}[\tilde{\theta}_T, \nu(\tilde{\theta}_T)]\right] \frac{\partial \nu}{\partial \theta'}(\tilde{\theta}_T)(\hat{\theta}_T - \tilde{\theta}_T)$$

Assuming that the initial estimating equations $q_T[\theta, \nu]$ are twice continuously differentiable on the interior of the compact set $\Theta \times \Gamma$ (see regularity conditions in appendix), we know that:

$$\left\| \frac{\partial q_T}{\partial \nu'}[\hat{\theta}_T, \nu(\tilde{\theta}_T)] - \frac{\partial q_T}{\partial \nu'}[\tilde{\theta}_T, \nu(\tilde{\theta}_T)] \right\| = O_P\left(\left\|\hat{\theta}_T - \tilde{\theta}_T\right\|\right)$$

Therefore

$$\left\|\tilde{h}_T^P(\hat{\theta}_T) - h_T^{(1)}(\hat{\theta}_T)\right\| = O_P\left(\left\|\hat{\theta}_T - \tilde{\theta}_T\right\|^2\right) = O_P\left(\alpha_T \left\|\hat{\theta}_T - \tilde{\theta}_T\right\|^2\right)$$

since α_T goes to infinity.

B.4 Proof of Theorem 3.2

We just show that the proof of Proposition 3.1. will go through with some suitable changes.

The proof of consistency (**Step 1**) is the same except that equation (48) must now be replaced by:

$$0 = q_\infty[\bar{\theta}, \nu(\theta^0)] + \frac{\partial q_\infty}{\partial \nu'}[\bar{\theta}, \nu(\theta^0)](\nu(\bar{\theta}) - \nu(\theta^0) + P \lim_{T=\infty} \alpha_T \left\|\nu(\theta_T^{(2)}) - \tilde{\nu}_T\right\|^2 e_p \tag{50}$$

Obviously, the same kind of consistency argument is still valid. Since $\text{Plim}_{T=\infty}\tilde{\nu}_T = \nu(\theta^0)$ and $\lim_{T=\infty} \alpha_T = \infty$, (50) implies that $\text{Plim}_{T=\infty}\nu(\theta_T^{(2)}) = \nu(\bar{\theta}) = \nu(\theta^0)$. Therefore we must have

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

from which we deduce $\bar{\theta} = \theta^0$ by virtue of Assumption B2.

To get **Step 2**, we now compute the Jacobian of the estimating equations

$$\begin{aligned}\frac{\partial h_T^{(2)}(\theta)}{\partial \theta'} &= \frac{\partial q_T[\theta, \tilde{\nu}_T]}{\partial \theta'} + \frac{\partial}{\partial \theta'}\left[\frac{\partial q_T}{\partial \nu'}[\theta, \tilde{\nu}_T]\right][(\nu(\theta) - \tilde{\nu}_T) \otimes Id_p] \\ &\quad + \frac{\partial q_T}{\partial \nu'}[\theta, \tilde{\nu}_T]\frac{\partial \nu}{\partial \theta'}(\theta) + 2\alpha_T\left[[\nu(\theta) - \tilde{\nu}_T]'\frac{\partial \nu}{\partial \theta'}(\theta)e_p\right]e_p\end{aligned}$$

Thus, we still have

$$\lim \frac{\partial h_T^{(2)}(\theta^0)}{\partial \theta'} = \frac{\partial q_\infty[\theta^0, \nu(\theta^0)]}{\partial \theta'} + \frac{\partial q_\infty}{\partial \nu'}[\theta^0, \nu(\theta^0)]\frac{\partial \nu}{\partial \theta'}(\theta^0) = F$$

and thus we can prove a **Step 2** exactly as in Proposition 3.1. This **Step 2** will tell us that

$$\hat{\theta}_T - \theta_T^{(2)} = O_P\left(\left\|f_T(\hat{\theta}_T) - h_T^{(2)}(\hat{\theta}_T)\right\|\right)$$

To get the announced result, we now (**Step 3**) need to show that

$$\left\| f_T(\hat{\theta}_T) - h_T^{(2)}(\hat{\theta}_T) \right\| = O_P\left(\alpha_T \left\| \nu(\hat{\theta}_T) - \tilde{\nu}_T \right\|^2 \right)$$

We have

$$\begin{aligned} f_T(\hat{\theta}_T) &= q_T[\hat{\theta}_T, \nu(\hat{\theta}_T)] \\ &= q_T[\hat{\theta}_T, \tilde{\nu}_T\ ] + \frac{\partial q_T}{\partial \nu'}[\hat{\theta}_T, \tilde{\nu}_T\ ].(\nu(\hat{\theta}_T) - \tilde{\nu}_T\ ) + O_P\left(\left\| \nu(\hat{\theta}_T) - \tilde{\nu}_T \right\|^2 \right) \\ &= h_T^{(2)}(\hat{\theta}_T) + O_P\left(\left\| \nu(\hat{\theta}_T) - \tilde{\nu}_T \right\|^2 \right) - \alpha_T \left\| \nu(\hat{\theta}_T) - \tilde{\nu}_T\ \right\|^2 e_p \end{aligned}$$

which gives the announced result.

C Tables

The following section details the Monte Carlo results from the simulation experiments in Section 4.2.2 and Section 5.2.1. In the tables below, MBP stands for the maximization by parts estimator, P-TS_1 is the fully penalized two-step estimator, the two-step estimator with partial penalization is P-TS_2, the two-stage estimator without penalization is TS_1 and the simplified two-step estimator is TS_2.

Table 1: μ_{100} is the estimated Mean, times 100, for the different estimator, MAE_{100} is the Monte Carlo mean squared error, times 1000, and MAE_{100} is the Monte Carlo mean absolute error, times 1000. The penalization parameter was taken proportional to $T^{1/4}$. The parameters α_1, α_2 were set equal to .1 and 1 across all sample sizes. The table below fixed $\rho = .75$.

$\rho = .75$		μ_{100}	MSE_{100}	MAE_{100}	μ_{200}	MSE_{200}	MAE_{200}	μ_{300}	MSE_{300}	MAE_{300}
MBP	α_1	10.0366	1.0429	81.3085	9.9641	0.5694	59.8332	10.0058	0.3432	46.3782
	α_2	100.1881	103.3290	801.1743	99.6738	54.0119	605.4414	99.6216	34.5963	469.3129
	ρ	75.3606	399.0850	1963.9320	74.9131	410.2502	2008.6820	75.0615	402.3241	1993.8420
P-TS_1	α_1	10.03615	1.0367	81.1981	9.9634	0.56507	59.6130	10.0047	0.34109	46.2839
	α_2	100.1919	103.4473	801.3737	99.6749	54.0004	605.3785	99.6255	34.5726	469.1240
	ρ	75.3727	398.5865	1962.7233	74.9204	409.9491	2007.9506	75.0674	402.0887	1993.2528
P-TS_2	α_1	10.0387	1.0375	81.2172	9.9657	0.5655	59.6264	10.0071	0.3413	46.2821
	α_2	100.2106	103.4881	801.4460	99.6922	54.0372	605.6657	99.6434	34.5842	469.1822
	ρ	74.6152	430.1701	2038.4766	74.2100	439.7422	2078.9956	74.3475	431.9187	2065.2499
TS_1	α_1	10.03671	1.0369	81.2033	9.9636	0.5650	59.6107	10.0048	0.34109	46.2837
	α_2	100.1963	103.4614	801.4120	99.6764	54.0001	605.3864	99.6264	34.5723	469.1192
	ρ	75.3633	398.9833	1963.6687	74.9160	410.1344	2008.3948	75.0646	402.2031	1993.5322
TS_2	α_1	10.0389	1.0375	81.2189	9.9658	0.5655	59.8332	10.0071	0.3413	46.2821
	α_2	100.2121	103.4928	801.4590	99.6928	54.0369	605.4414	99.6437	34.5841	469.1803
	ρ	75.3633	430.1764	2038.4899	74.9160	439.7444	2008.6820	75.0646	431.9198	2065.2523

Table 2: μ_{100} is the estimated Mean, times 100, for the different estimator, MAE_{100} is the Monte Carlo mean squared error, times 1000, and MAE_{100} is the Monte Carlo mean absolute error, times 1000. The penalization parameter was taken proportional to $T^{1/4}$. The parameters α_1, α_2 were set equal to .1 and 1 across all sample sizes. The table below fixed $\rho = .95$.

$\rho = .95$		μ_{100}	MSE_{100}	MAE_{100}	μ_{200}	MSE_{200}	MAE_{200}	μ_{300}	MSE_{300}	MAE_{300}
MBP	α_1	10.0368	1.0436	81.2541	9.9679	0.5843	60.4423	10.0609	0.5784	61.1738
	α_2	100.3326	101.6088	803.7984	99.5895	57.6078	608.1178	99.8313	59.6184	615.5324
	ρ	95.0740	0.5900	61.2780	94.9604	0.2925	43.7780	94.5256	0.5025	56.1719
P-TS_1	α_1	10.0361	1.03802	81.2689	9.9624	0.5655	59.6429	10.0041	0.3419	46.3435
	α_2	100.3307	101.3133	802.8277	99.5938	56.7603	604.8641	99.8505	35.0822	468.8306
	ρ	95.0799	0.5908	61.3661	94.9895	0.2909	43.5954	95.0280	0.2124	37.6353
P-TS_2	α_1	10.0381	1.0382	81.2629	9.9645	0.5656	59.6420	10.0062	0.3420	46.3441
	α_2	100.3489	101.3216	802.9570	99.6126	56.7682	604.96135	99.8697	35.0885	468.8434
	ρ	94.8122	0.8194	72.4880	94.7308	0.4688	54.3549	94.7739	0.3468	46.7765
TS_1	α_1	10.0361	1.0380	81.2689	9.9624	0.5655	59.6428	10.0041	0.3419	46.3434
	α_2	100.3309	101.3135	802.8293	99.5939	56.7602	604.8638	99.8505	35.0822	468.8304
	ρ	95.0797	0.5909	61.3679	94.9894	0.2910	43.5963	95.0280	0.2124	37.6357
TS_2	α_1	10.0381	1.0382	81.2629	9.9645	0.5656	60.4423	10.0062	0.34205	46.3440
	α_2	100.3489	101.3217	802.9575	99.6126	56.7682	608.1178	99.8697	35.0885	468.8434
	ρ	95.0797	0.8194	72.4880	94.9894	0.4688	43.7780	95.0280	0.3468	46.7766

Table 3: μ_{100} is the estimated Mean, times 100, for the different estimator, MAE_{100} is the Monte Carlo mean squared error, times 1000, and MAE_{100} is the Monte Carlo mean absolute error, times 1000. The penalization parameter was taken proportional to $T^{1/4}$. The parameters α_1, α_2 were set equal to .1 and 1 across all sample sizes. The table below fixed $\rho = .985$.

$\rho = .985$		μ_{100}	MSE_{100}	MAE_{100}	μ_{200}	MSE_{200}	MAE_{200}	μ_{300}	MSE_{300}	MAE_{300}
MBP	α_1	10.0372	1.2314	88.7035	10.1446	1.7055	112.0659	10.2592	2.4063	143.4284
	α_2	100.8034	123.5412	879.2445	100.1246	165.4135	1108.4347	01.5723	242.56481	1439.3493
	ρ	98.0858	0.2934	43.3899	96.4156	4.6093	208.4360	94.8641	13.5292	363.5840
P-TS_1	α_1	10.0357	1.0392	81.2881	9.9621	0.5662	59.6819	10.0039	0.3421	46.3557
	α_2	100.3488	102.2060	804.7207	99.5960	57.0553	600.34733	99.9360	34.8534	468.4711
	ρ	98.5239	0.0541	18.6253	98.4977	0.0263	13.1028	98.5098	0.0193	11.3651
P-TS_2	α_1	10.0367	1.0393	81.2839	9.9631	0.5662	59.6806	10.0049	0.3421	46.3560
	α_2	100.3582	102.2073	804.7324	99.6055	57.0552	600.3720	99.9456	34.8571	468.5121
	ρ	98.4358	0.0801	22.7917	98.4122	0.0457	16.9308	98.42705	0.0330	14.5021
TS_1	α_1	10.0357	1.0392	81.2881	9.96217	0.5662	59.6819	10.0039	0.3421	46.3557
	α_2	100.3488	102.2061	804.7207	99.5960	57.0553	600.34731	99.9361	34.8534	468.471
	ρ	99.8524	0.0541	18.6253	98.4977	0.0263	13.1029	98.5098	0.0193	11.3652
TS_2	α_1	10.0367	1.0393	81.2839	9.9631	0.5662	59.6806	10.0049	0.3421	46.3561
	α_2	100.3582	102.2073	804.7325	99.6056	57.0553	600.3720	99.9456	34.8572	468.5121
	ρ	98.5239	0.0801	22.7917	98.4977	0.0457	16.9301	98.5098	0.0331	14.5022

Table 4: Relative computing time, in seconds and number of iterations for MBP (in brackets).

		$\rho = .75$		$\rho = .95$		$\rho = .985$	
T=100	MBP	0.0119	[3]	0.0260	[6]	0.0922	[43]
	P-TS	0.0203		0.0166		0.0137	
	TS	0.0207		0.0164		0.0142	
T=200	MBP	0.0152	[4]	0.0450	[16]	0.1005	[44]
	P-TS	0.0193		0.0141		0.0138	
	TS	0.0199		0.0167		0.0148	
T=300	MBP	0.0169	[4]	0.1001	[42]	0.1018	[45]
	P-TS	0.0184		0.0143		0.0145	
	TS	0.0187		0.0160		0.0149	

Table 5: Results for two-step (TS_1, penalized two-step (P-TS_1) and MBP estimators in the Merton credit Risk model. MAE is the median absolute error across the simulations multiplied by 100, and RMSE is the root mean squared error across the simulation multiplied by 100.

TS					
T	Parameter	Median	Mean.	MAE	RMSE
T=250	$\sigma = 0.09$	0.0889	0.0888	6.1099	7.7331
T=500	$\sigma = 0.09$	0.0897	0.0895	4.7301	5.8404
P-TS					
T	Parameter	Median	Mean.	MAE	RMSE
T=250	$\sigma = 0.09$	0.0895	0.0890	5.9292	7.5341
T=500	$\sigma = 0.09$	0.0898	0.0895	4.6715	5.6284
MBP					
T	Parameter	Median	Mean	MAE	RMSE
T=250	$\sigma = 0.09$	0.0892	0.0888	8.1746	9.8406
T=500	$\sigma = 0.09$	0.0894	0.0898	6.7727	6.6129