Journal of Econometrics 201 (2017) 212–227

Contents lists available at ScienceDirect

Journal of Econometrics

journal homepage: www.elsevier.com/locate/jeconom

Efficient two-step estimation via targeting

David T. Frazier [a], Eric Renault [b,*]

[a] Department of Econometrics and Business Statistics, Monash University, Australia
[b] Department of Economics, Brown University, United States

ARTICLE INFO

Article history:
Available online 19 August 2017

Keywords:
Targeting
Penalization
Multivariate time series models
Asset pricing

ABSTRACT

The standard description of two-step extremum estimation amounts to plugging-in a first-step estimator of nuisance parameters to simplify the optimization problem and then deducing a user friendly, but potentially inefficient, estimator for the parameters of interest. In this paper, we consider a more general setting of two-step estimation where we do not necessarily have 'nuisance parameters' but rather awkward occurrences of the parameters of interest. The efficiency problem associated with two-step estimators in this context is more difficult than with standard nuisance parameters as even if the true unknown value of the parameters were plugged-in to alleviate the awkward occurrences of the parameters, the resulting second-step estimator may not be efficient. In addition, standard approaches to restore efficiency for two-step procedures may not work due to a consistency issue. To alleviate this potential issue, we propose a new computationally simple two-step estimation procedure that relies on targeting and penalization to enforce consistency, with the second-step estimators maintaining asymptotic efficiency. 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.

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 a 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

* Corresponding author.
E-mail addresses: david.frazier@monash.edu (D.T. Frazier), eric_renault@brown.edu (E. Renault).

http://dx.doi.org/10.1016/j.jeconom.2017.08.004
0304-4076/

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. 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 components of θ can be split into two parts $\theta = (\theta_1', \theta_2')'$: θ_1 characterizes the likelihood of the margins and θ_2 characterizes the dependence between components, let us 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 log-likelihood portion 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 popular in 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 is overlooked.

In Section 5, we consider nonlinear models in which observable variables are viewed as functions of some latent state 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 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. The issue we have in mind is not about filtering latent variables since we only consider cases where the latent-to-observable relationship is one-to-one. Hence, backing out the latent variables from the observations would have been easy if not polluted by the additional awkward 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 for the state variables. Since this measurement equation, which could for instance be an arbitrage-based asset pricing formula, is one-to-one, we follow Pan (2002) and dub "Implied States" the value of latent variables that can be backed out from observations for a given value of $\nu(\theta)$.[1]

In this setting we are then 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 $\hat{\theta}_T$. This inefficiency is due to the estimator $\breve{\theta}_T$ disregarding the information about θ contained in the function $\nu(\theta)$ (see Crepon et al., 1997 for a similar remark in a GMM context). More precisely, (under standard conditions) the efficient estimator $\hat{\theta}_T$ in (1) satisfies

$$\sqrt{T}(\hat{\theta}_T - \theta^0) = -\left[\frac{\partial q_T[\theta^0, \nu(\theta^0)]}{\partial \theta'} + \frac{\partial q_T[\theta^0, \nu(\theta^0)]}{\partial \nu'}\frac{\partial \nu(\theta^0)}{\delta \theta'}\right]^{-1} \times \sqrt{T} q_T[\theta^0, \nu(\theta^0)] + o_P(1),$$

whereas the infeasible estimator $\breve{\theta}_T$ satisfies

$$\sqrt{T}(\hat{\theta}_T - \theta^0) = -\left[\frac{\partial q_T[\theta^0, \nu(\theta^0)]}{\partial \theta'}\right]^{-1} \sqrt{T} q_T[\theta^0, \nu(\theta^0)] + o_P(1).$$

Two standard strategies are available in the literature to address this efficiency issue. A first possibility, as recently developed by Fan et al. (2015) (hereafter, FPR) is to devise a sequence of estimators $\hat{\theta}_T^{(k)}$, $k = 1, 2, \ldots$, 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 ($\hat{\theta}_T^{(1)} = \tilde{\theta}_T$). In the case of separable log-likelihood functions (see Section 4) a simplified version of (4) was proposed in a seminal paper by Song et al. (2005) (hereafter, SFK), with their proposed algorithm being dubbed "Maximization by Parts" (hereafter, MBP). The key feature of (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, in 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 $\hat{\theta}_T$.[2] However, it is worth realizing that the required contraction mapping property to secure convergence of (4) need not be fulfilled in finite samples.[3] Therefore, a feasible efficient estimator relies upon the choice of a tuning parameter $k(T)$, going to infinity at a sufficiently fast 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

[1] Extending the approach put forward by Renault and Touzi (1996) (and later revisited by Pastorello et al., 2003), Pan (2002) used this structure to devise the so-called "Implied States GMM" estimator.

[2] A similar results is obtained for the iterative estimators studied in Dominitz and Sherman (2005).

[3] It is well known that consistent estimation of a function does not imply consistent estimation of its derivative. Therefore, a contraction mapping condition, stated as the norm of the derivative for the limit function being smaller than unity, may not be satisfied in finite sample.

computational cost of a large number $k(T)$ of iterations, especially when the required population contraction mapping property is nearly unfulfilled. Needless to say, the situation is even worse when it is not fulfilled at all, as illustrated in Section 4.

The main goal of this paper is to promote a new efficient two-step procedure that does not require a 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

$$\operatorname*{plim}_{T\to\infty} \left[J_T(\theta^0) + \frac{\partial^2 Q_T[\theta^0, \nu(\theta^0)]}{\partial \nu \partial \nu'} \right] = 0.$$

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 the unknown θ inside the matrix $J_T(\theta)$, the first-order conditions of the maximization program (5) can be seen as a linearization of the 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 consistency of estimators defined as roots 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 penalized two-step procedure come with the cost of a tuning parameter. While MBP requires choosing the number $k(T)$ of iterations, our approach must choose the rate of divergence $\alpha(T)$ for the penalty weight. We will see, for instance, that in the standard case where all estimators are $\sqrt{T}$-consistent, a rate $T^{1/4}$ is well suited. Moreover, we propose two simplified versions of this approach, both based on partial linearization, depending upon whether one has at her disposal a first-step consistent estimator of θ^0 (the full vector) or only of $\nu(\theta^0)$ (only the awkward occurrences).

Heuristically, the two-step estimators we propose "target", through the use of a penalty term, simple and consistent first-step estimators to deduce a computationally simple and efficient estimate of θ^0. In this way, our two-step approach bears a resemblance to so-called variance "targeting" estimators in GARCH models, which make use of computationally simple estimates of the unconditional variance as a means of simplifying the optimization problem. However, in contrast with variance "targeting" GARCH estimators, our two-step approach delivers efficient estimators.

The remainder of 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 et al. (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 and allows us to derive two simplified penalized two-step estimators. 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 applications of this approach in the setting of 'so-called' implied states. For brevity we focus on the case of Maximum likelihood estimation, while noting that our results can easily be applied to GMM as well. Again, in this section 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. All proofs are 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 that 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

(i) $\operatorname*{plim}_{T\to\infty} \left\{ \sup_{\theta\in\Theta} |Q_\infty[\theta, \nu(\theta)] - Q_T[\theta, \nu(\theta)]| \right\} = 0$ and

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

Assumption A2. The following conditions are satisfied:

(i) $\nu(.)$ is twice continuously differentiable on $\text{Int}(\Theta)$, the interior of Θ.

(ii) $\theta^0 \in \text{Int}(\Theta)$ and $\nu^0 = \nu(\theta^0) \in \text{Int}(\Gamma)$.

(iii) The function $Q_T[\theta, \nu]$ is twice continuously differentiable on $\text{Int}(\Theta) \times \text{Int}(\Gamma)$. For $q_T[\theta, \nu(\theta)]$ defined by (2)

(1) $\sqrt{T}\left(\frac{\partial Q_T[\theta^0, \nu(\theta^0)]}{\partial \theta'}, \frac{\partial Q_T[\theta^0, \nu(\theta^0)]}{\partial \nu'} \right)' \to_d N[0, W_0]$ and $\sqrt{T} q_T[\theta, \nu(\theta)] \to_d N[0, I_0]$.

(2) $\operatorname*{plim}_{T\to\infty} \left\{ \frac{\partial q_T[\theta^0, \nu(\theta^0)]}{\partial \theta'} + \frac{\partial q_T[\theta^0, \nu(\theta^0)]}{\partial \nu'} . \frac{\partial \nu(\theta^0)}{\partial \theta'} \right\} = F$ and non-singular

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

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

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

The examples in Sections 2.2 and 2.3 demonstrate that, for simplicity, in practice one may want to define the matrix $-J_T$ different from (albeit asymptotically equivalent to) the Hessian matrix.

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. To this end, we have the following result.

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

It is worth elaborating on this result to discern the reason why the two-step approach is not responsible for any efficiency loss. Consider the case of genuine nuisance parameter discussed in Trognon and Gourieroux (1990):

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

Then, the modified objective function becomes

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

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. Namely,

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

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:

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

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

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

so that, under Assumption A2 and from the definition of J_T, we deduce

$$\operatorname*{plim}_{T\to\infty}\frac{\partial\breve{Q}^0_{c,T}[\theta_1^0,\nu^0]}{\partial\theta_1\partial\nu'}=0. \tag{7}$$

Eq. (7) is precisely the standard condition (see e.g Newey and McFadden, 1994, formula (6.6) pp. 2179) to ensure that the asymptotic distribution of the estimated parameters θ_1 do not depend on the asymptotic distribution of the estimated nuisance parameters ν. This provides clear intuition as to why Theorem 2.1 works in the particular case considered by Trognon and Gourieroux (1990): the modified objective function in (6) restores the asymptotic independence between the two kinds of parameters.

We stress that it is only the careful analysis of the first-order conditions (see Section 3) that will allow us to devise a penalized estimation strategy that ensures consistency of such two-step estimators. This approach amounts to a slight twist (via targeting and penalization) of the two-step estimator $\hat{\theta}_T^{ext}$, to ensure its consistency, with efficiency then guaranteed by Theorem 2.1. The following examples illustrate why such a penalized two-step approach may be required to ensure consistency.

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:

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

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 can be stated as

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

Noting,

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

and by applying (8), we can choose the following sample counterpart 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

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

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 is crucial to note that the linearization in (10) is only partial, as it only deals with the nasty occurrences $\nu(\theta)$ in $g(\cdot)$, and so efficiency is warranted only when consistency is satisfied. 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 penalization/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

$$y_t = \alpha_1 y_{t-1} + \alpha_2 z_t + u_t$$
$$u_t = \beta u_{t-1} + \varepsilon_t$$

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)

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

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. Hence, modifying the objective function as in Eq. (6) delivers the correct first-order conditions, since this quadratic approximation amounts to partial linearization of the first-order conditions w.r.t ν. Therefore, this approximation does not jeopardize consistency and 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 (11) and their possible applications are, for instance, described in the literature on Implied States GMM (see, e.g., Pan, 2002, Pastorello et al., 2003, and Fan et al., 2015). 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 $\lim_{T\to\infty}\text{Var}[\sqrt{T}\bar{\varphi}_T(\theta^0, \nu(\theta^0))]$. From the definition of $Q_T[\theta, \nu(\theta)]$, we have

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

where $W_{h.;T}$ stands for the h^{th} row of W_T. Then, we can choose the following sample counterpart 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 minimizer of

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

In other words, while the solution of

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

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 (12) constitutes only a partial linearization of the moment conditions, consistency may not be warranted.

Herein, consistency may be an issue since the identification assumption (11) does not ensure

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

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:

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

Then, the implication (14) is obviously warranted when the first set of moment conditions is sufficient to identify θ^0, a condition maintained by Gourieroux et al. (1996), since (14) becomes

$$E\begin{bmatrix}\varphi_{1t}(\theta)\\ \varphi_{2t}(\theta, \nu(\theta))\end{bmatrix} + E\begin{bmatrix}0\\ \dfrac{\partial \varphi_{2t}(\theta, \nu(\theta^0))}{\partial \nu'}\end{bmatrix}\left[\nu(\theta) - \nu(\theta^0)\right]$$
$$= 0 \iff \theta = \theta^0. \tag{16}$$

In this case, Theorem 2.1 ensures efficiency of the modified two-step estimator.[4]

[4] Interestingly, the estimator proposed by Gourieroux et al. (1996) will only numerically coincide with $\hat{\theta}_T^{ext}$ when the moment conditions $\varphi_{1t}(\theta)$ are linear in θ (see Section 2.6 in Gourieroux et al., 1996).

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 wants to use a first-step consistent estimator of θ^0 or only of $\nu(\theta^0)$.

3.1. The general estimator and 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. We remind the reader that for the purpose of efficiency our benchmark estimator is $\hat{\theta}_T$. For sake of generality, we maintain some high level assumptions about these estimating equations.

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 $\tilde{h}_T(\theta) = 0$ will converge towards a solution $\theta = \bar{\theta}$ of the population equations:

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

If $\tilde{h}_T(\theta)$ were a genuine linearization of $f_T(\theta)$ (not only a partial one), Robinson's Theorem 2 shows that the zeros of $\tilde{h}_T(\theta)$ and $f_T(\theta)$ are, in a sense, asymptotically equivalent. 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

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

for a real sequence α_T going slowly to infinity, where e_p stands for a fixed p-dimensional vector with at least one non-zero component. More precisely, our extension of Robinson's result can be stated as follows:

Proposition 3.1. *Under the standard regularity conditions detailed in the appendix, and under Assumption* B1, *if $\tilde{\theta}_T$ is a consistent estimator of θ^0 such that $\|\tilde{\theta}_T - \theta^0\| = o_P(1/\alpha_T)$ with $\alpha_T \to \infty$ as $T \to \infty$, then for any zero $\tilde{\theta}_T^P$ of $\tilde{h}_T^P(\theta)$ in (17), and $\hat{\theta}_T$ as in (1)*

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

Proposition 3.1 is a generalization of Theorem 2 in Robinson (1988). However, unlike Robinson (1988), our partial linearization requires a penalty term $\alpha_T\|\theta - \tilde{\theta}_T\|^2$, with $\alpha_T \to \infty$ to ensure consistency. Fortunately, the penalty term will only have a very minor impact for Proposition 3.1: when the initial estimator $\tilde{\theta}_T$ is $\sqrt{T}$-consistent, $\hat{\theta}_T$ and $\tilde{\theta}_T^P$ are first-order asymptotically equivalent if $\alpha_T \to \infty$ 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), asymptotic equivalence between $\hat{\theta}_T$ and $\tilde{\theta}_T^P$ requires $\tilde{\theta}_T$ to converge faster than $T^{1/4}$.[5] However, 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. $\log[\log(T)]$).[6] 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).[7]

3.2. Simplified two-step efficient estimators

In the following subsections we present two simplified estimators depending on whether one has an initially consistent estimator of the partial parameters ν or the full vector θ.

3.2.1. Initially consistent $\tilde{\theta}_T$

Our general two-step efficient estimator $\tilde{\theta}_T^P$ is obtained by a direct extension of Robinson (1988), replacing the complete linearization by a partial one. The estimating equations $\tilde{h}_T^P(\theta)$, in (17), can be made even more computationally friendly by making the correction term linear in the unknown parameters θ; that is, rather, by solving the following estimating equations:

$$h_T^{(1)}(\theta) = q_T[\theta, \nu(\tilde{\theta}_T)] + \frac{\partial q_T}{\partial \nu'}[\tilde{\theta}_T, \nu(\tilde{\theta}_T)]\frac{\partial \nu}{\partial \theta'}(\tilde{\theta}_T)(\theta - \tilde{\theta}_T) + \alpha_T \left\|\theta - \tilde{\theta}_T\right\|^2 e_p. \tag{18}$$

[5] Note that, such an instance can occur even when θ is finite dimensional. Examples include, among many others, estimation of a tail index, local GMM, nearly weak identification, etc. For discussion, see Antoine and Renault (2012) and references therein.

[6] For the examples considered in this paper, $\alpha_T \propto T^{1/4}$ is always asymptotically valid. Of course in any particular implementation, one may use some cross-validation procedure to find a suitable coefficient of proportionality. However, this is beyond the scope of this paper.

[7] From page 465 in Pastorello et al. (2003), $k(T)$ must go to infinity faster than $\log(T)$, with $k(T)$ inversely related to the strength of the contraction mapping argument required for convergences.

The difference between $\tilde{h}_T^P(\theta)$ and $h_T^{(1)}(\theta)$ is the use of the first-step consistent estimator $\tilde{\theta}_T$ to replace the non-awkward occurrence of the parameters θ in the complete Jacobian matrix. This simplification does not impair the general equivalence result of Proposition 3.1

Theorem 3.1. *Under the standard regularity conditions detailed in appendix, and under Assumptions* B1 *and* B2, *if* $\tilde{\theta}_T$ *is a consistent estimator of* θ^0 *such that* $\|\tilde{\theta}_T - \theta^0\| = o_P(1/\alpha_T)$ *with* $\alpha_T \to \infty$ *as* $T \to \infty$ *then for any zero* $\theta_T^{(1)}$ *of* $h_T^{(1)}(\theta)$ *in* (18), *and* $\hat{\theta}_T$ *as in* (1)

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

3.2.2. Initially consistent $\tilde{\nu}_T$

In applications, it may be the case that only a sub-vector of the parameters of interest θ can be consistently estimated in a first-step. In this case, an alternative two-step efficient estimator that only requires knowledge of a first-step consistent estimator $\tilde{\nu}_T$ of $\nu(\theta^0)$ can also be obtained. 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).

The alternative two-step estimator $\theta_T^{(2)}$ is defined as a zero of the estimating equations

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

with the following result as a consequence.

Theorem 3.2. *Under standard regularity conditions detailed in the appendix, and 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* $\alpha_T \to \infty$ *as* $T \to \infty$, *then for any zero* $\theta_T^{(2)}$ *of* $h_T^{(2)}(\theta)$ *in* (19), *and* $\hat{\theta}_T$ *as in* (1)

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

Theorem 3.2 implies that the previous asymptotic efficiency of $\tilde{\theta}_T^P$ and $\theta_T^{(1)}$, deduced from Proposition 3.1 and Theorem 3.1, respectively, applies to $\theta_T^{(2)}$. The main difference is that the leading rate of convergence is now that 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.

3.2.3. Practical implications

The penalized two-step estimators, in particular the simplified ones, $\theta_T^{(1)}$ defined as the solution to $0 = h_T^{(1)}(\theta)$ and $\theta_T^{(2)}$ defined as the solution to $0 = h_T^{(2)}(\theta)$, require a particular choice for the penalty term α_T.

When the benchmark estimator $\hat{\theta}_T$, the solution to $q_T[\hat{\theta}_T, \nu(\hat{\theta}_T)] = 0$, and the initial estimators $\tilde{\theta}_T$ or $\tilde{\nu}_T$, are $\sqrt{T}$-consistent, the two rules to which the penalty term must adhere are as follows: one, for sake of asymptotic efficiency, α_T must go to infinity strictly slower than $\sqrt{T}$; two, $\alpha_T \to \infty$ to enforce consistency (see Step 1 in the proof of Proposition 3.1).

As exemplified in Sections 2.2 and 2.3, if consistency is not an issue, Theorem 2.1 demonstrates the asymptotic efficiency of $\hat{\theta}_T^{ext}$, the solution to

$$\hat{\theta}_T^{ext} = \arg\max_\theta \left\{ Q_T[\theta, \tilde{\nu}_T] + \frac{\partial Q_T[\theta, \tilde{\nu}_T]}{\partial \nu'}.[\nu(\theta) - \tilde{\nu}_T] - \frac{1}{2}[\nu(\theta) - \tilde{\nu}_T]' J_T(\theta)[\nu(\theta) - \tilde{\nu}_T] \right\}. \tag{20}$$

Overlooking the dependence on θ within $J_T(\theta)$, up to the penalty term, the first-order conditions for (20) are very similar to (19).[8] However, it is important to keep in mind that consistency is not always warranted, and then the only solution is the introduction of a penalty term in the first-order conditions as in (19).

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

$$Q_T[\theta, \nu(\theta)] = Q_{1T}[\theta_1] + Q_{2T}[\theta_2, \nu(\theta)], \tag{21}$$

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 $\sqrt{T}$-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

$$\frac{\partial Q_{1T}[\tilde{\theta}_{1T}]}{\partial \theta_1} = 0, \tag{22}$$

$\tilde{\theta}_{1T}$ 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

$$\frac{\partial Q_{2T}[\tilde{\theta}_{2T}, \tilde{\theta}_{1T}]}{\partial \theta_2} = 0. \tag{23}$$

If (22) and (23) are unbiased estimating equations for θ^0, in the sense that,

$$\lim_{T\to\infty} \frac{\partial Q_{1T}[\theta_1]}{\partial \theta_1} = 0 \Longleftrightarrow \theta_1 = \theta_1^0,$$

$$\lim_{T\to\infty} \frac{\partial Q_{2T}[\theta_2, \theta_1^0]}{\partial \theta_2} = 0 \Longleftrightarrow \theta_2 = \theta_2^0,$$

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

While computationally simple, the estimator $\tilde{\theta}_T$ is inefficient. 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. Obtaining $\theta_T^{(1)}$ and $\theta_T^{(2)}$ with $Q_T[\theta, \nu(\theta)]$ as in (21) only requires specializing the definitions of $h_T^{(1)}(\theta)$ and $h_T^{(2)}(\theta)$. To this end, for

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

[8] See the proof of Theorem 2.1 as to why overlooking the dependence of θ in $J_T(\theta)$ will not alter the asymptotic distribution of $\hat{\theta}_T^{ext}$.

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

$$0 = h_{1T}^{(1)}(\theta_T^{(1)}) = \frac{\partial Q_{1T}[\theta_{1T}^{(1)}]}{\partial \theta_1} + \frac{\partial Q_{2T}[\theta_{2T}^{(1)}, \tilde{\theta}_{1T}]}{\partial \nu} + \frac{\partial^2 Q_{2T}[\tilde{\theta}_{2T}, \tilde{\theta}_{1T}]}{\partial \nu \partial \nu'}(\theta_{1T}^{(1)} - \tilde{\theta}_{1T}) + \alpha_T \left\| \theta_T^{(1)} - \tilde{\theta}_T \right\|^2 e_{p_1} \quad (24)$$

$$0 = h_{2T}^{(1)}(\theta_T^{(1)}) = \frac{\partial Q_{2T}[\theta_{2T}^{(1)}, \tilde{\theta}_{1T}]}{\partial \theta_2} + \frac{\partial^2 Q_{2T}[\tilde{\theta}_{2T}, \tilde{\theta}_{1T}]}{\partial \theta_2 \partial \nu'}(\theta_{1T}^{(1)} - \tilde{\theta}_{1T}) + \alpha_T \left\| \theta_T^{(1)} - \tilde{\theta}_T \right\|^2 e_{p_2} \quad (25)$$

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

$$0 = h_{1T}^{(2)}(\theta_T^{(2)}) = \frac{\partial Q_{1T}[\theta_{1T}^{(2)}]}{\partial \theta_1} + \frac{\partial Q_{2T}[\,\theta_{2T}^{(2)}, \tilde{\theta}_{1T}]}{\partial \nu} + \frac{\partial^2 Q_{2T}[\theta_{2T}^{(2)}, \tilde{\theta}_{1T}]}{\partial \nu \partial \nu'}(\theta_{1T}^{(2)} - \tilde{\theta}_{1T}) + \alpha_T \left\| \theta_{1T}^{(2)} - \tilde{\theta}_{1T} \right\|^2 e_{p_1} \quad (26)$$

$$0 = h_{2T}^{(2)}(\theta_T^{(2)}) = \frac{\partial Q_{2T}[\theta_{2T}^{(2)}, \tilde{\theta}_{1T}]}{\partial \theta_2} + \frac{\partial^2 Q_{2T}[\theta_{2T}^{(2)}, \tilde{\theta}_{1T}]}{\partial \theta_2 \partial \nu'}(\theta_{1T}^{(2)} - \tilde{\theta}_{1T}) + \alpha_T \left\| \theta_{1T}^{(2)} - \tilde{\theta}_{1T} \right\|^2 e_{p_2}, \quad (27)$$

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

Obviously, solving (24) and (25) (respectively, (26) and (27)) 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 brute force efficient estimation.

This simplification of the estimating equations is also shared by the MBP estimator proposed in SFK. With $Q_T[\theta, \nu(\theta)]$ as in Eq. (21), 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 $k = k(T)$; 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. If this condition is not satisfied $\hat{\theta}_T^{(k)}$ will 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:

$$0 = \frac{\partial Q_{1T}[\tilde{\theta}_{1T}^{(k+1)}]}{\partial \theta_1} + \frac{\partial Q_{2T}[\tilde{\theta}_{2T}^{(k+1)}, \tilde{\theta}_{1T}^{(k)}]}{\partial \theta_1}, \quad (28)$$

$$0 = \frac{\partial Q_{2T}[\tilde{\theta}_{2T}^{(k+1)}, \tilde{\theta}_{1T}^{(k)}]}{\partial \theta_2}. \quad (29)$$

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 (28), (29) 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 Eqs. (24), (25) but with $\alpha_T = 0$.[9] Firstly, comparing the MBP estimator and TS$_1$ (resp., P-TS$_1$), the only difference between the two estimators is that TS$_1$ (resp., 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, from a computational standpoint, when the local contraction mapping is strong the MBP procedure of SFK is the simplest. As the required contraction mapping condition becomes weaker, the MBP estimator becomes more computationally burdensome.[10] 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 MBP, the penalized two-step estimator $\theta_T^{(2)}$ (dubbed P-TS$_2$) and the corresponding non-penalized version $\theta_{T(P)}^{(2)}$ (dubbed TS$_2$) incurs additional computational complexity because θ_2 occurs within the partial Hessian term in the estimating equations. However, 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. Similar reasoning has even led 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 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 for variance targeting to yield an asymptotically normal estimator of the unconditional variance.

In the following subsection, we illustrate the above discussion between the different estimation procedures using a Gaussian Copula model; see, e.g., Joe (1997), Song (2000),

4.2. Bivariate Gaussian Copula models

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

[9] 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.

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

$C(u_1, u_2; \rho)$, 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)

$$C(F_1(y_{i,1}; \alpha_1), F_2(y_{i,2}; \alpha_2); \rho) = \Phi_\rho(\Phi^{-1}(F_1(y_{i,1}; \alpha_1)), \Phi^{-1}(F_2(y_{i,2}; \alpha_2))), \tag{30}$$

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 Eq. (30). 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

$$Q_T[\theta, \nu(\theta)] = \sum_{i=1}^{T}\sum_{j=1}^{2} \log(f_j(y_{i,j}; \alpha_j)) - \frac{T}{2}\log(1-\rho^2) - \frac{\rho}{2(1-\rho^2)}(\rho A(\theta_1) - 2B(\theta_1)). \tag{31}$$

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 (31) 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.

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

We now compare the finite sample properties of the MBP approach of SFK and two different efficient two-step procedures: the penalized two-step estimator P-TS$_1$ and the partially penalized two-step estimator P-TS$_2$.[11] 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 1000 synthetic samples. Note that for ρ greater than approximately .95 the information dominance condition associated with the proposed MBP procedure is no longer satisfied and we expect the finite sample properties of the MBP estimator to be poor in comparison with the 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. Table 1 reports the averages over the 1000 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 $T^{1/4}$.

[11] Additional non-penalized versions of P-TS$_1$ and P-TS$_2$ were also considered. However, results for these additional estimators are not reported for brevity but are available from the authors upon request.

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.

		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
$\rho = .75$	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
$\rho = .75$	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
$\rho = .75$	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
MBP	α_1	10.0368	1.0436	81.2541	9.9679	0.5843	60.4423	10.0609	0.5784	61.1738
$\rho = .95$	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
$\rho = .95$	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
$\rho = .95$	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
MBP	α_1	10.0372	1.2314	88.7035	10.1446	1.7055	112.0659	10.2592	2.4063	143.4284
$\rho = .985$	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
$\rho = .985$	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
$\rho = .985$	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

For $\rho = .75$ the MBP algorithm and the two-step estimators are very similar. However, for larger ρ the penalized two-step methods give smaller MSEs and MAEs than the MBP estimator. With high correlation values and larger sample sizes the MBP algorithm encounters difficulty since the matrix driving the updates does not fulfill the IDC. The same behavior is not in evidence for the two-stage and penalized two-stage estimates, which perform well even for $\rho = .985$.[12]

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 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.,

$$Y_t = h[Y_t^*, \nu^0] \iff Y_t^* = g[Y_t, \nu^0]. \tag{32}$$

When $\nu(\theta^0)$ is unknown, Eq. (32) 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 (32) 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 likelihood models with latent states. While the approach is equally relevant to so-called Implied stated GMM, we do not present any formal analysis in the name of space.

5.1. 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

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

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 (32) at the value θ, and the Jacobian formula, the infeasible log-likelihood $Q_T^*[\theta]$ is transformed into the feasible log-likelihood

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

[12] At lower sample sizes and smaller values of ρ, the computational speed of MBP, P-TS_1 and P-TS_2 are all roughly equivalent. However, as T increases and/or as ρ increases, the computational speed of P-TS_1 and P-TS_2 stays the same but increase for MBP.

$\left|H_y g[Y_t, \nu(\theta)]\right|$ is the determinant 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 structural 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 procedure discussed herein can be applied in this general setting, it is perhaps more informative to consider precise implementation in a relatively simple example.

5.2. Example: Merton (1974) Credit risk model

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.,

$$S_\delta \equiv \max[V_\delta - B, 0]. \tag{33}$$

From (33) the observed equity prices $S_0, \ldots, S_T$ can be interpreted as option prices written on the firm's unobservable market values $V_0, \ldots, V_T$.

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

$$\frac{dV_t}{V_t} = \mu dt + \sigma dW_t, \tag{34}$$

where W_t is a standard Brownian motion. Eq. (34) allows us to write the conditional likelihood of the sample path $(V_1, V_2, \ldots, V_T)$ 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

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

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

However, when the dynamics of the firm's market value are described by (34), the observable equity values can be related to the unobservable firm values through the Black and Scholes option pricing formula:

$$S_t = V_t\Phi(d_t) - B\exp(-r(\delta - t))\Phi(d_t - \sigma\sqrt{\delta - t}), \tag{35}$$

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 from Eq. (35) 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 V_t could be imputed from $V_t = g[S_t, \sigma^2]$ for each $t = 1, \ldots, T$. 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

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

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 and work with the concentrated log-likelihood

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

$Q_T[\sigma^2]$ is complicated, in several places, by the structural relationship $g[S_t, \sigma^2]$. We denote these problematic occurrences of σ^2 in $Q_T[\sigma^2]$ due to $g[S_t, \sigma^2]$ 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

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

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 MLE, the KMV/PPR estimator is inefficient. To this end, FPR propose a MBP estimator that maintains the computational advantages of KMV/PPR but that is asymptotically equivalent to the MLE. Given an initial estimator $\tilde{\sigma}^2$, at the kth iteration ($k > 1$) the MBP estimator solves the following second-order equation in σ^2:

$$\sigma^4 A_T[\nu(\hat{\sigma}^{2(k-1)})] - \sigma^2 + \tilde{\sigma}_T^2[\nu(\hat{\sigma}^{2(k-1)})] = 0. \tag{36}$$

An alternative to the KMV/PPR and MBP approaches is the general two-step approach discussed in Section 3.1. 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 $\tilde{\sigma}^2$ an initial estimator of σ^2 and, for α_T a penalty term, the general penalized two-step approach estimates σ^2 by solving[13]

$$\begin{aligned} 0 = {} & \sigma^4 A_T[\nu(\tilde{\sigma}^2)] - \sigma^2 + \tilde{\sigma}_T^2[\nu(\tilde{\sigma}^2)] \\ & + [\partial A_T[\nu(\tilde{\sigma}^2)]/\partial\nu]\sigma^4(\sigma^2 - \tilde{\sigma}^2) \\ & + [\partial\tilde{\sigma}_T^2[\nu(\tilde{\sigma}^2)]/\partial\nu](\sigma^2 - \tilde{\sigma}^2) + \alpha_T(\sigma^2 - \tilde{\sigma}^2)^2. \end{aligned} \tag{37}$$

Note that the two-step estimator in Eq. (37) requires solving a third-order equation in σ^2, whereas the MBP estimator in Eq. (36) solves a second-order equation. However, the two-step estimator solves *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 $\tilde{\sigma}^2$ 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.

5.3. 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 two-step estimator in (37). We construct 1000 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 $B = 9000$. 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.

[13] A non-penalized version of this estimator displays similar performance. The results are available from the authors upon request.

Table 2
Results for 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.

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

The MBP estimator is obtained using a Newton–Raphson approach to solve Eq. (36). The 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 1000 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 estimator.

The results of the Monte Carlo experiments are contained in Table 2. Table 2 demonstrates that the two-step estimator and the MBP estimator have similar finite sample properties, with the penalized two-step estimator having significantly smaller RMSE and MAE.

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 Gourieroux et al. (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 an asymptotic distribution that depends (in general) on the distribution of the first step estimator but even iterations may not be able to restore efficiency.

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 often devised precisely to allow us to overlook these awkward occurrences (possibly) at the cost of efficiency. The goal of FPR was to propose efficient iterative estimation procedures whose computational cost, at each step of 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 is 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. On the one hand, our approach is not required to compute a sequence of estimators but only a second step estimator, which generally maintains the computational simplicity associated with each step of the FPR iterative estimators. 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 the solution of the (partly) linearized estimating equation may be inconsistent. 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 do not 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. 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.

Acknowledgments

We thank the editors and two anonymous referees for their helpful comments, which greatly improved the paper.

Appendix 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:

$$\hat{\theta}_T = \arg\max_{\theta\in\Theta} Q_T[\theta, \nu(\theta)] \tag{38}$$

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

$$Q_T[\theta, \theta^*] = Q_T^*[\theta, \nu_T(\theta^*)] \tag{39}$$

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 (38) is actually not less general than (39), 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 (38), 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 non-stochastic 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 $\mathrm{Int}(\Theta) \times \mathrm{Int}(\Gamma)$.

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 $\lambda \in \mathrm{Int}(\Theta) \times \mathrm{Int}(\Gamma)$ towards a non stochastic matrix $D(\lambda)$.
(2) The matrix $D_{\theta\theta}(\lambda^0) = \operatorname{plim}_{T\to\infty} \frac{\partial^2 Q_T(\lambda^0)}{\partial\theta\partial\theta'}$ (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 (39). 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 exposition simplicity.

Appendix B. Proofs

B.1. Proof of Theorem 2.1

First, define the infeasible estimator $\hat{\theta}_T^* = \arg\max_{\theta\in\Theta} \tilde{Q}_T^0[\theta, \tilde{\nu}_T]$, where

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

The proof proceeds in two parts.

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

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

Adding and subtracting

$$\frac{\partial \nu'(\theta)}{\partial \theta}\frac{\partial^2 Q_T[\theta, \tilde{\nu}_T]}{\partial \nu \partial \nu'}[\nu(\theta) - \tilde{\nu}_T]$$

within the definition of $q_T^*[\theta, \nu(\theta)]$, using the definition of $q_T[\theta, \nu(\theta)]$ in (2), and grouping terms yields the following equivalent definition:

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

with

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

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 $\sqrt{T}$-consistent. Whenever consistent, an estimator $\mathring{\theta}_T$ solution of

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

with

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

is 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 the 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 $j = 1, \ldots, p$ 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:

$$0 = q_T[\tilde{\theta}_T^P, \nu(\tilde{\theta}_T)] + \frac{\partial q_T}{\partial \nu'}[\tilde{\theta}_T^P, \nu(\tilde{\theta}_T)]\frac{\partial \nu}{\partial \theta'}(\tilde{\theta}_T)(\tilde{\theta}_T^P - \tilde{\theta}_T) + \alpha_T \left\|\tilde{\theta}_T^P - \tilde{\theta}_T\right\|^2 e_p \tag{40}$$

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 (40) that

$$0 = q_\infty[\bar{\theta}, \nu(\theta^0)] + \frac{\partial q_\infty}{\partial \nu'}[\bar{\theta}, \nu(\theta^0)]\frac{\partial \nu}{\partial \theta'}(\theta^0)(\bar{\theta} - \theta^0) + \operatorname*{plim}_{T\to\infty} \alpha_T \left\|\tilde{\theta}_T^P - \tilde{\theta}_T\right\|^2 e_p. \tag{41}$$

Since $\operatorname*{plim}_{T\to\infty}\left\{\tilde{\theta}_T\right\} = \theta^0$ and $\lim_{T=\infty}\alpha_T = \infty$, (41) implies that $\operatorname*{plim}_{T\to\infty}\left\{\tilde{\theta}_T^P\right\} = \theta^0$.

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

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

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

$$\left[\frac{\partial A^1}{\partial \theta'} \quad \frac{\partial A^2}{\partial \theta'} \quad \cdots \quad \frac{\partial A^q}{\partial \theta'}\right]$$

where $A^1, A^2, \ldots, A^q$ stand for the q columns of 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

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

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 show that the proof of Proposition 3.1goes through with minor changes. The proof of consistency (**Step 1**) is the same except that Eq. (41) must now be replaced by

$$\begin{aligned} 0 = q_\infty[\bar{\theta}, \nu(\theta^0)] + \frac{\partial q_\infty}{\partial \nu'}[\theta^0, \nu(\theta^0)]\frac{\partial \nu}{\partial \theta'}(\theta^0)(\bar{\theta} - \theta^0) \\ + \operatorname*{plim}_{T\to\infty}\left\{\alpha_T\left\|\theta_T^{**} - \tilde{\theta}_T\right\|^2 e_p\right\} \end{aligned} \tag{42}$$

Obviously, the same consistency argument is a fortiori still valid. Since $\operatorname{plim}_{T\to\infty}\{\tilde{\theta}_T\} = \theta^0$ and $\alpha_T \to \infty$, (42) implies that $\operatorname{plim}_{T\to\infty}\{\theta_T^{(1)}\} = \theta^0$. 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

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

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 show that the proof of Proposition 3.1goes through with suitable changes. The proof of consistency (**Step 1**) is the same except that Eq. (41) must now be replaced by

$$\begin{aligned} 0 = q_\infty[\bar{\theta}, \nu(\theta^0)] + \frac{\partial q_\infty}{\partial \nu'}[\bar{\theta}, \nu(\theta^0)]\nu(\bar{\theta}) - \nu(\theta^0) \\ + \operatorname*{plim}_{T\to\infty}\left\{\alpha_T\left\|\nu(\theta_T^{(2)}) - \tilde{\nu}_T\right\|^2 e_p\right\} \end{aligned} \tag{43}$$

Obviously, the same kind of consistency argument is still valid. Since $\operatorname{plim}_{T\to\infty}\{\tilde{\nu}_T\} = \nu(\theta^0)$ and $\lim_{T\to\infty}\alpha_T = \infty$, (43) implies that $\operatorname{plim}_{T\to\infty}\left\{\nu(\theta_T^{(2)})\right\} = \nu(\bar{\theta}) = \nu(\theta^0)$. 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:

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

Thus, we still have

$$\lim_{T\to\infty}\frac{\partial h_T^{(2)}(\theta^0)}{\partial \theta'} = \frac{\partial q_\infty[\theta^0, \nu(\theta^0)]}{\partial \theta'} + \frac{\partial q_\infty}{\partial \nu'}[\theta^0, \nu(\theta^0)]\frac{\partial \nu}{\partial \theta'}(\theta^0) = F$$

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 then have the following, which gives the announced result,

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

References

Antoine, B., Renault, E., 2012. Efficient minimum distance estimation with multiple rates of convergence. J. Econometrics 170 (2), 350–367. Thirtieth Anniversary of Generalized Method of Moments.

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

Crepon, B., Kramarz, F., Trognon, A., 1997. Parameters of interest, nuisance parameters and orthogonality conditions an application to autoregressive error component models. J. Econometrics 82 (1), 135–156.

Dominitz, J., Sherman, R.P., 2005. Some convergence theory for iterative estimation procedures with an application to semiparametric estimation. Econometric Theory 21 (4), 838–863.

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

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

Engle, R., 2002. Dynamic conditional correlation. J. Bus. Econom. Statist. 20 (3), 339–350.

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

Fan, Y., Pastorello, S., Renault, E., 2015. Maximization by parts in extremum estimation. Econom. J. 18, 147–171.
Gourieroux, C., Monfort, A., Renault, E., 1993. Indirect inference. J. Appl. Econometrics 8, S85–S118. Special issue on econometric inference using simulation techniques.
Gourieroux, C., Monfort, A., Renault, E., 1996. Two-stage generalized moment method with applications to regressions with heteroscedasticity of unknown form. J. Statist. Plann. Inference 50 (1), 37–6193. Econometric methodology, Part III.
Hartley, H.O., 1961. The modified Gauss-Newton method for the fitting of non-linear regression functions by least squares. Technometrics 3 (2), 269–280.
Hatanaka, M., 1974. An efficient two-step estimator for the dynamic adjustment model with autoregressive errors. J. Econometrics 2 (3), 199–220.
Joe, H., 1997. Multivariate Models and Dependence Concepts, Vol. 73. Chapman an Hall, London.
Liu, Y., Luger, R., 2009. Efficient estimation of copula-garch models. Comput. Stat. Data Anal. 53 (6), 2284–2297. The Fourth Special Issue on Computational Econometrics.
Merton, R.C., 1974. On the pricing of corporate debt: The risk structure of interest rates. J. Finance 29 (2), 449–470.
Newey, W.K., McFadden, D.L., 1994. Large sample estimation and hypothesis testing. In: Engle, R.F., McFadden, D.L. (Eds.), Handbook of Econometrics, Vol. 4. pp. 2111–2245.
Noureldin, D., Shephard, N., Sheppard, K., 2014. Multivariate rotated ARCH models. J. Econometrics 179 (1), 16–30.
Pagan, A., 1986. Two stage and related estimators and their applications. Rev. Econom. Stud. 53 (4), 517–538.
Pakes, A., Pollard, D., 1989. Simulation and the asymptotics of optimization estimators. Econometrica 57 (5), 1027–1057.
Pan, J., 2002. The jump-risk premia implicit in options: evidence from an integrated time-series study. J. Financ. Econ. 63 (1), 3–50.
Pastorello, S., Patilea, V., Renault, E., 2003. Iterative and recursive estimation in structural nonadaptive models [with comments, rejoinder]. J. Bus. Econom. Statist. 21 (4), 449–482.
Patton, A.J., 2009. Copula-Based models for financial time series. In: Andersen, T., Davis, R., Kreiss, J.-P., Mikosch, T. (Eds.), Handbook of Financial Time Series. Springer Verlag, pp. 767–785.
Renault, E., Touzi, N., 1996. Option hedging and implied volatilities in a stochastic volatility model 1. Math. Finance 6 (3), 279–302.
Robinson, P.M., 1988. The stochastic difference between econometric statistics. Econometrica 56 (3), 531–548.
Shih, J.H., Louis, T.A., 1995. Inferences on the association parameter in copula models for bivariate survival data. Biometrics 51 (4), 1384–1399.
Song, P.X.-K., 2000. Multivariate dispersion models generated from gaussian copula. Scand. J. Stat. 27 (2), 305–320.
Song, P.X.-K., Fan, Y., Kalbfleisch, J.D., 2005. Maximization by parts in likelihood inference [with comments, rejoinder]. J. Amer. Statist. Assoc. 100 (472), 1145–1167.
Trognon, A., Gourieroux, C., 1990. A note on the efficiency of two-step estimation methods. In: Essays in Honor of Edmond Malinvaud. MIT Press, pp. 233–248.