CREST
CENTER FOR RESEARCH
IN ECONOMICS AND STATISTICS

Série des Documents de Travail

n° 2017-10

Consistent Pseudo-Maximum Likelihood Estimators

C. GOURIEROUX[1]
A.MONFORT[2]
E.RENAULT[3]

Les documents de travail ne reflètent pas la position du CREST et n'engagent que leurs auteurs.
Working papers do not reflect the position of CREST but only the views of the authors.

[1] CREST, University of Toronto. E-mail : Christian.gourieroux@ensae.fr

[2] CREST. E-mail : Alain.monfort@ensae.fr

[3] Brown university. E-mail : Eric_Renault@brown.edu

Consistent Pseudo-Maximum Likelihood Estimators

C., Gourieroux [(1)], A., Monfort [(2)], and E., Renault [(3)]

(September, 2016)

This research has benefited of the financial support of the chair LCL : ”New Challenges for New Data”, and of the chair ACPR : ”Regulation and Systemic Risks”.

[1] CREST and University of Toronto.

[2] CREST

[3] Brown University.

Consistent Pseudo-Maximum Likelihood Estimators
Abstract

The development of the literature on the pseudo maximum likelihood (PML) estimators would not have been so efficient without the modern proof of consistency of extremum estimators introduced at the end of the sixties by E. Malinvaud and R. Jennrich. We discuss this proof and replace it in an historical perspective. In this paper we also provide a survey of the literature on consistent (PML) estimators. We emphasize the role of the white noise assumptions on the set of pseudo distributions leading to consistent estimators. The stronger these assumptions, the larger the set of consistent PML estimators. We also illustrate the importance of these PML approaches in big data environment.

Keywords : Pseudo-Likelihood, Composite Pseudo-Likelihood, Consistency, Big Data, ARCH Model, Normalized Data, Lie Group.

Estimateurs du Pseudo Maximum de Vraisemblance Convergents
Résumé

Le développement de la littérature sur les estimateurs du pseudo maximum de vraisemblance (PMV) n'aurait pas été aussi important sans la preuve de la convergence des estimateurs extrémaux introduite par E. Malinvaud et R. Jennich à la fin des années soixante. On discute cette preuve et on la replace dans son contexte historique. Dans cet article on propose également une revue de la litterature sur la convergence des estimateurs PMV. On souligne le rôle de l'hypothèse de bruit blanc apparaissant dans les pseudo lois fournissant des estimateurs convergents. Plus cette hypothèse est forte plus l'ensemble des estimateurs convergents est grand. On illustre aussi l'importance des ces estimateurs dans un contexte de données massives.

1 Introduction

The paper by Edmond Malinvaud about consistency of nonlinear regressions [Malinvaud (1970)] has triggered a large literature dealing with consistency of various kinds of estimators, in particular the M-estimators. A M-estimator of a parameter is obtained by optimizing sums of parametrized functions of the data. The class of M-estimators [4] is broad: it includes the maximum likelihood (ML) estimators, and the pseudo maximum likelihood [5] (PML) estimators, when the distributions used are misspecified. In particular, for regression models this class includes estimation methods such as (nonlinear) least squares and least absolute deviation methods. When the number of observations is large, the asymptotic properties of M-estimators, that are their existence, consistency and asymptotic distribution can be derived.

In this paper, we focus on (existence and) consistency of such estimators, considering PML estimators for expository purpose. Different approaches have been proposed in the early literature for proving the consistency of M-estimators.

The first approach, sometimes dubbed the Wald's consistency proof, was initially developed for independently and identically distributed (i.i.d) variables and the maximum likelihood approach. This proof and its extensions [see e.g. Cramer (1946), Wald (1949), Huber (1967)] are based on regularity conditions that may be restrictive.

This may explain the success of the modern proof for the asymptotic existence of a M-estimator and its consistency that appears at the end of the sixties [see Jennrich (1969), Malinvaud (1966),(1970)], with initially in mind the nonlinear least squares estimators, but is nowadays popular for any kind of extremum estimator.

For the two classes of proof, the basic idea is always to replace the finite sample objective function to be optimized by its limit after an appropriate standardization. Then the M-estimator is hopefully consistent if the as-

[4] Following [Huber (1974)] M is for Maximizing. However, following Huber's idea of "maximum likelihood like" (see also [Wooldridge (1994)]), we do not include in the class of M-estimators all the extremum estimators. The objective function must be a sum over the observations by contrast with other extremum estimators, like minimum distance Asymptotic Least Squares (ALS), or Generalized Method of Moments (GMM) [Hansen (1982)]. However, the consistency arguments are quite general.

[5] also called quasi-maximum likelihood (QML) estimators, especially for Gaussian pseudo-distribution.

sociated asymptotic objective function is optimal at the true value of the parameter. Consistency may be either a weak one (convergence in probability), or a strong one (almost sure convergence) (see also [Kim and Pollard (1990)]). The main difference between the Wald approach on the one hand and the Jennrich-Malinvaud approach on the other hand is the convergence theorem for the objective function that underpins the consistency argument for the M-estimator. The Wald approach refers to the monotone convergence theorem; the price to pay is the need to assume that the supremum of the Log Likelihood Ratio (LLR) over a sufficiently small ball around the true value is integrable. The Jennrich-Malinvaud approach resorts to a convergence of the sample objective function towards the asymptotic one that, because it is uniform with respect to the unknown parameters, allows to consider the consistency of the sequence of maximizers.

The paper is organized as follows :

We provide in Section 2 a brief historical perspective of the two main strategies, global uniformity assumption vs dominance, to prove consistency of M-estimators. The uniformity approach is illustrated in Section 3 through classical examples of PML estimators in the modern econometric literature. Up to regularity conditions to apply a suitable uniform Law of Large Numbers, the key trick is to figure out why a pseudo-log-likelihood, albeit being a misspecified log-likelihood, is asymptotically maximized at a pseudo true value that coincides with the true value. Important historical references in this respect are [Huber (1967), Gourieroux-Monfort-Trognon (1984b), Bollerslev-Wooldridge (1992)] . In particular, with a special emphasis on pseudo-likelihood functions that are suggested by conditional moment restrictions, we can compare the PML approach with the Generalized Method of Moments (GMM) based on a choice of instruments. A complete characterization is provided for a dynamic model of conditional mean and variance, like an AutoRegressive model with ARCH-type errors ([Meddahi-Renault (1998)].

The aforementioned case of regression type models with an error satisfying "conditional moment restrictions" paves the way for a possible characterization of all PML approaches, that is to say, all pseudo-distributions of the error, for which the PML estimators are consistent. We illustrate these results in Section 4 for mean regressions [Gourieroux, Monfort, Trognon (1984) b] and quantile regressions [Gourieroux, Monfort, Renault (1987)].

While illustrations of PML provided in Sections 3 and 4 can be traced back to the eighties and nineties, more recent developments of the PML strategy are described in Sections 5 and 6. Section 5 considers PML approaches directly derived from a likelihood with partial information. These methods are known as composite likelihood methods [see e.g. Cox, Reid (2004), Varin, Reid, Firth (2011)]. They are especially useful in big data contexts and nonlinear dynamic models [see e.g. Gourieroux, Monfort, Trognon (1984a), Gourieroux, Monfort (2016)].

Stronger assumptions can be introduced on the errors of the regression type models such as the assumption of i.i.d. errors. We explain in Section 6 that in this framework the PML estimators of the parameters defining the sensitivities of the explanatory variable are consistently estimated by any PML approach if additional parameters are introduced in the model. This consistency result is valid for nonlinear mean and quantile regression models as well as for ARCH type models [Newey, Steigerwald (1997), Berkes, Horvath (2004), Francq, Lepage, Zakoian (2011), Fan, Qi, Xiu (2014). They can be extended to regression models based on a Lie group of linear transformations [Gourieroux, Monfort, Zakoian (2016)b]. Section 7 concludes. Technical material is provided in the appendices.

2 A Historical Perspective

[Wald (1949)] considers a sequence $X_1, X_2, ..., X_n$ of i.i.d. real valued random variables whose probability distribution depends on some unknown vector θ of parameters:

$$\Pr(X_i < x) = F(x, \theta).$$

It is assumed that for any θ, "$F(x, \theta)$ admits an elementary probability law $f(x, \theta)$" meaning that $f(x, \theta)$ stands for the probability density function of the distribution $F(x, \theta)$, either with respect to the Lebesgue measure (absolute continuous case), or with respect to the counting measure (discrete case). In all cases, it is first noted that, if θ_0 stands for the true unknown value of θ, then :

$$E_0 [\log f(X, \theta)] < E_0 [\log f(X, \theta_0)], \forall \theta \neq \theta^0, \tag{2.1}$$

where X is a variable with true distribution $F_0(x) = F(x, \theta_0)$. [Huber (1967)] will generalize Wald's proof of consistency towards θ_0 to the PML case, that is to the case where we want a consistent estimator of the pseudo-true value

θ_0 of θ defined by (2.1), even though we acknowledge that there may be no such thing as a true value because no distribution $F(x,\theta)$ (for any θ) coincides with the true unknown distribution F_0 of X in (2.1). In all case, the standard identification assumption is implicitly maintained:

$$f(x,\theta) =_{as} f(x,\theta_0) \Rightarrow \theta = \theta_0,$$

where $=_{as}$means equal almost surely, for the true unknown probability distribution of X. Then, one defines the supremum of the density function in the neighborhood of any possible value θ as follows:

$$f(x,\theta,\rho) = \sup \left\{ f(x,\theta'); \|\theta - \theta'\| < \rho \right\},$$

where $\mathcal{U}(\rho) = \{\theta' : ||\theta - \theta'|| < \rho\}$ is a ball of radius ρ centered at θ.

Then, [Wald (1949)] maintains the dominance assumption:

$$E_0 \left[Max_{\theta \in \mathcal{U}(\rho)} \left\{ \log \left(f(X,\theta,\rho) \right), 0 \right\} \right] < \infty, \tag{2.2}$$

for ρ sufficiently small , which allows him to apply the monotone convergence theorem to show that:

$$\lim_{\rho \to 0} E_0 \left[\log f(X,\theta,\rho) \right] = E_0 \left[\log f(X,\theta) \right].$$

This result provides in turn the consistency of maximum likelihood estimator by considering the behavior of the likelihood ratio on closed subsets of the parameter space, which do not contain the true parameter point θ_0.

For expository purpose, let us assume that we have an i.i.d. sample $X_1, ..., X_n$, a compact parameter space Θ, and we want to prove that the maximizer $\hat{\theta}_n$ over Θ of the process:

$$M_n(\theta) = \frac{1}{n} \sum_{i=1}^{n} m(X_i, \theta),$$

converges in probability to a point θ_0 maximum of the function:

$$M(\theta) = E_0 \left[m(X,\theta) \right].$$

The key assumption is, as in (2.2), the dominance assumption:

$$E_0 \left[\sup_{\theta \in \mathcal{U}(\rho)} m(X,\theta) \right] < \infty, \tag{2.3}$$

for every sufficiently small ball $\mathcal{U}(\rho) \subset \Theta$. Morever, it is assumed that the function $\theta \longmapsto m(x,\theta)$ is upper-semicontinuous for almost all x:

$$\limsup_{\theta_n \to \theta} m(x,\theta_n) \leq_{as} m(x,\theta).$$

We shall allow multiple points of maximum by defining the set:

$$\Theta_0 = \left\{ \theta_0 \in \Theta; M(\theta_0) = \sup_{\theta \in \Theta} M(\theta) \right\}.$$

The set Θ_0 is assumed not empty. We are then able to prove ([van der Vaart (1998)], Theorem 5.14):

Theorem 1.: For every estimator $\hat{\theta}_n$ such that for some $\theta_0 \in \Theta_0$:

$$M_n\left(\hat{\theta}_n\right) \geq M_n(\theta_0) - o_P(1), \tag{2.4}$$

we have, for every $\varepsilon > 0$:

$$\lim_{n\to\infty} \Pr\left[d(\hat{\theta}_n, \Theta^0) > \varepsilon \right] = 0,$$

where d is a distance on the parameter space and $o_p(1)$ denotes a negligible term in probability.

Proof : see Appendix 1

Theorem 1 shows in particular that, if the limit objective function $M(.)$ has a unique maximum at $\theta = \theta_0$, the estimator $\hat{\theta}_n$ maximizing $M_n(\theta)$ is weakly consistent for θ_0. As already mentioned, the weakness of the Wald's consistency proof is to need the dominance assumption (2.3). The contribution of Jennrich-Malinvaud has been to replace this assumption by an assumption of uniform convergence in probability:

$$\sup_{\theta\in\Theta} |M_n(\theta) - M(\theta)| \longrightarrow_P 0. \tag{2.5}$$

Note that, in the context of M-estimators , (2.5) is a uniform weak Law of Large Numbers, but this approach is even more general for an extremum estimator. This approach based on uniformity (2.5) has become so popular that, when refereing to it, [Kim and Pollard (1990)] write that "the argument for consistency has become quite standard, almost to the point of cliché. It is enough to assume that...(i) $\hat{\theta}_n$ comes within $o_P(1)$ of maximizing (assumption (2.4) above), (ii) the uniform convergence condition (2.5) is fulfilled, and there is a unique maximum at θ_0 in the sense that:

$$\forall \varepsilon > 0, \sup\left\{M(\theta); \|\theta - \theta_0\| \geq \varepsilon\right\} < M(\theta_0). \tag{2.6}$$

To summarize, we prove the following result:

Theorem 2.:

Under Assumptions (2.5) and (2.6), a M-estimator in the sense of (2.4) is weakly consistent for θ_0.

Proof: By virtue of (2.4) and (2.5) respectively:

$$M_n\left(\hat{\theta}_n\right) \geq M_n(\theta_0) - o_P(1) = M(\theta_0) - o_P(1).$$

Therefore:

$$\begin{aligned} M(\theta_0) - M(\hat{\theta}_n) &\leq M_n\left(\hat{\theta}_n\right) - M(\hat{\theta}_n) + o_P(1) \qquad (2.7) \\ &\leq \sup_{\theta\in\Theta} |M_n(\theta) - M(\theta)| + o_P(1) = o_P(1). \end{aligned}$$

We want to show that for any given $\varepsilon > 0$:

$$\lim_{n\to\infty} \Pr\left[\left\|\hat{\theta}_n - \theta_0\right\| \geq \varepsilon\right] = 0. \tag{2.8}$$

But, by (2.6), for this particular ε:

$$M(\theta_0) - \sup\left\{M(\theta); \|\theta - \theta_0\| \geq \varepsilon\right\} = \eta > 0.$$

In other words:

$$\left\|\hat{\theta}_n - \theta_0\right\| \geq \varepsilon \Rightarrow M(\theta_0) - M(\hat{\theta}_n) \geq \eta,$$

and we know from (2.7) that:

$$\lim_{n\to\infty} \Pr\left[M(\theta_0) - M(\hat{\theta}_n) \geq \eta\right] = 0.$$

So we get (2.8). □

Several remarks are in order to fully realize the high degree of generality of the Jennrich-Malinvaud approach to consistency of M-estimators.

First, even though Theorem 2 is stated in terms of convergence in probability, one can instantaneously translate it in terms of almost sure convergence thanks to the following lemma (see e.g. [Bierens (2004)], Theorem 6.B3. page 168):

Lemma 1.:

$X_n \to_P X$ if and only if every subsequence n_m of $n = 1, 2, 3...$ contains a further subsequence $n_m(k)$ such that for $k \to \infty$, $X_{n_m(k)} \to_{as} X$.

By straightforward application of this lemma, we realize that an immediate corollary of Theorem 2. is that we get a strongly consistent estimator for θ_0 ($\hat{\theta}_n \to_{as} \theta_0$), if we reinforce the uniformity assumption (2.5) by assuming instead:

$$\sup_{\theta \in \Theta} |M_n(\theta) - M(\theta)| \longrightarrow_{as} 0.$$

Second, the above arguments may not be fully correct since some expressions on display may be non-measurable. However, [Jennrich (1969)] had addressed directly the measurability issue by proving the following lemma:

Lemma 2.: If the function $m(x, \theta)$ is a measurable function of x for any given θ, and a continuous function of θ for any given x, then there exists a measurable function $\hat{\theta}_n$ (as function of $(x_{1,}x_2, ..., x_n)$) such that:

$$M_n(\hat{\theta}_n) = \sup_{\theta \in \Theta} M_n(\theta). \tag{2.9}$$

The proof given in [Jennrich (1969)] is actually very general and valid for any extremum estimator. However, this result does not guarantee the measurability for any solution of extremum problems when the estimators solution of (2.9) are multiple. Of course, they may be multiple even though the limit criterion admits a unique global maximum. [Hansen (2012)] extends Jennrich's result, while assuming that the function $x \to m(x, .)$ is measurable as a function taking its value in the set of continuous functions on Θ (with the sup norm and associated Borel sets). Then, it can be shown that the function:

$$(x_{1,}x_2,...,x_n) \to \left\{\bar{\theta} \in \Theta; M_n(\bar{\theta}) \geq M_n(\theta), \forall \theta \in \Theta\right\},$$

is measurable when the sets of parts of Θ is endowed with the Hausdorff metric (and the corresponding Borel sets).

Third, from an historical point of view, it is worth knowing that [Jennrich (1969)] and [Malinvaud(1970)], albeit tightly related, have been written independently. [Malinvaud (1970)] was actually a follow up on [Malinvaud (1966)], Chapter 9. [Malinvaud (1970)] compares itself to [Jennrich (1969)] (see its footnote 3 page 957) as giving more primitive assumptions and an "elementary, but tedious proof" of consistency instead of a "more elegant", but more involved proof in [Jennrich (1969)].

Fourth, the distinction put forward in this section between the Wald's proof of consistency on the one hand, and the Jennrich-Malinvaud's proof on the other hand is relevant for the history of econometrics but, according to [van der Vaart (1998)], "the two approaches can be unified by replacing the uniform convergence by "one-sided uniform convergence"".

3 Marginal and Pairwise Pseudo-Likelihood Methods

In parametric models the maximum likelihood approach is frequently difficult to implement due to its computational complexity. This problem arises in models with unobservable (latent) variables, where the likelihood function involves multidimensional integrals of a large dimension, which increases with the number of observations. This problem also arises when the number of data and parameters is huge, the so-called big-data framework. Indeed the standard algorithms to compute the ML estimates require the inversion of square matrices with a dimension equal to the number of parameters. Such an inversion may be inaccurate, or even untractable, if the number of parameters is more than one hundred, say.

In such frameworks, it has been proposed to replace the likelihood function by approximations, that neglect some dependence between the observations. This dependence can be entirely neglected in marginal pseudo-likelihood [6] methods. Pairwise dependence only is captured in pairwise

[6] sometimes referred as the independence likelihood [Chandler, Bate (2007)].

pseudo-likelihood methods.[7] The two approaches are mixed in composite likelihood methods. These approaches are becoming standard in financial econometrics [see e.g. Engle, Pakel, Shephard, Sheppard (2014)], as well as in network analysis [see e.g. Besag (1974), Stein et al. (2004)], in statistical genetics [see Larribe, Fearnhead (2011) for a survey], or in bioinformatics [Mardia et al. (2008)].

3.1 The principle

For expository purpose, we describe the principle and discuss the consistency of the associated PML estimator, when the observations $y_1, \ldots, y_T$ correspond to a strictly stationary process, with transition p.d.f. denoted by $f(y_t|\underline{y_{t-1}};\theta)$, where $\underline{y_{t-1}} = (y_{t-1}, y_{t-2}\ldots)$ is the information available at date t, and θ_0 denotes the true value of the parameter. The log-likelihood function is [8] :

$$L_T(\theta) = \sum_{t=1}^{T} \log f(y_t|\underline{y_{t-1}};\theta). \tag{3.1}$$

As seen in the examples of Section 3.2, this log-likelihood can be difficult to compute and/or to optimize numerically. Other summary statistics of the joint distribution of the observations may be much easier to compute. For instance, we can consider the marginal p.d.f. of y_t, denoted by $f_1(y_t;\theta)$, or the pairwise joint p.d.f. of (y_t, y_{t-h}), denoted by $f_{2,h}(y_t, y_{t-h};\theta)$, where h is a given lag. They can be used to define :

i) the marginal pseudo-likelihood :

$$L_{1,T}(\theta) = \sum_{t=1}^{T} \log f_1(y_t;\theta), \tag{3.2}$$

ii) the pairwise pseudo-likelihood at horizon h :

$$L_{2,h,T}(\theta) = \sum_{t=1}^{T} \log f_{2,h}(y_t, y_{t-h};\theta). \tag{3.3}$$

[7] Appropriate conditional likelihoods can also be considered [see e.g. Cox (1975) for an application to proportional hazard models].

[8] In practice the information set is truncated to account for observations available after $t = 1$ only.

The marginal pseudo-likelihood (3.2) is a mispecified likelihood in which all serial dependencies are ignored. It depends on a subset of parameter α, that are the parameters characterizing the marginal distributions, but does not involve the parameters β, say, characterizing the serial dependence.

The pairwise pseudo-likelihoods generally depend on both types of parameters α and β. These pairwise pseudo-likelihoods are not likelihood functions corresponding to a misspecified model. More precisely, let us consider the case $h = 1$.

$$\sum_{p=1}^{[T/2]} \log f_{2,1}(y_{2p}, y_{2p-1}; \alpha, \beta) \equiv L^{(1)}_{2,1,T}(\alpha, \beta),$$

is the log-likelihood of a misspecified model assuming that the successive pairs $(y_{2p}, y_{2p-1}), p$ varying, are independent. A similar interpretation exists for the sum:

$$\sum_{p=1}^{(T/2]} \log f_{2,1}(y_{2p+1}, y_{2p}; \alpha, \beta), \equiv L^{(2)}_{2,1,T}(\alpha, \beta).$$

The pairwise pseudo- log-likelihood is equal to : $L_{2,1,T}(\alpha, \beta) = L^{(1)}_{2,1,T}(\alpha, \beta) + L^{(2)}_{2,1,T}(\alpha, \beta)$; it is not the log-likelihood of a misspecified model, but a composite pseudo-likelihood obtained by adding different pseudo log-likelihoods.

The marginal and pairwise pseudo-likelihoods can be used to define PML estimators of $\theta = (\alpha', \beta)'$. For instance, we can introduce :

i) the marginal PML estimator of α as :

$$\hat{\alpha}_T = \arg\max_{\alpha} L_{1,T}(\alpha); \tag{3.4}$$

ii) the two-step PML estimator of β as :

$$\hat{\beta}_T = \arg\max_{\beta} \sum_{h=1}^{H} L_{2,h,T}(\hat{\alpha}_T, \beta), \tag{3.5}$$

where $\hat{\alpha}_T$ is deduced from (4.4).

We have the following consistency property :

Theorem 3.: Under standard regularity conditions and appropriate identifiability assumptions, $\hat{\alpha}_T$ (resp. $\hat{\beta}_T$) is a consistent estimator of α_0 (resp. β_0).

Proof : Let us for instance consider the PML estimator $\hat{\alpha}_T$. If the process (y_t) satisfies ergodicity properties, the standardized marginal pseudo log-likelihood $\frac{1}{T} L_{1,T}(\alpha)$ tends to :

$$E_0 \log f(y_t; \alpha) = \int \log f(y; \alpha) f(y; \alpha_0) dy.$$

By the properties of the Kullback Information divergence, this quantity is maximum for $f(y;\alpha) = f(y;\alpha_0)$, a.s. in y, and thus for $\alpha = \alpha_0$ only, if α_0 is identifiable from the marginal p.d.f.

The proof for the two-step PML estimator of β is similar. It assumes the identification of β for pairwise dependence and known α :

$$f_{2,h}(y_t, y_{t-h}; \alpha_0, \beta) = f_{2,h}(y_t, y_{t-h}; \alpha_0, \beta), \text{a.s. in } y_t, y_{t-h}, \text{ for } h = 1, \dots, H$$
$$\Rightarrow \beta = \beta_0.$$

QED

Remark 1 :

When the number of parameters is large, we may still encounter the curse of dimensionality problem at the second-step optimization. In such a case, it is sometimes possible to disentangle the parameters β_1 characterizing the dependence at horizon 1, the parameters β_2 characterizing the dependence at horizon 2, once the dependence at horizon 1 is known, and so on. Then we can apply the following sequence of optimizations (with clear notations) :

$$\left\{ \begin{array}{lcl} \hat{\beta}_{1T} & = & \arg\max_{\beta_1} L_{2,1,T}(\hat{\alpha}_T; \beta_1), \\ & & \\ \hat{\beta}_{2T} & = & \arg\max_{\beta_2} L_{2,2,T}(\hat{\alpha}_T; \hat{\beta}_{1,T}, \beta_2), ... \end{array} \right. \tag{3.6}$$

Remark 2 :

The pairwise pseudo-likelihoods can be mixed in different ways. For instance, we could consider the composite pseudo log-likelihood :

$$\tilde{L}_{2,T,\gamma}(\alpha, \beta) = \sum_{h=1}^{H} \gamma^h L_{2,h,T}(\alpha, \beta), \tag{3.7}$$

where we downweight observations that are far apart in time with γ defining the discounting. The corresponding PML estimator solution of the maximization of $\tilde{L}_{2,T,\gamma}(\hat{\alpha}_T, \beta)$ is still consistent, but with another asymptotic accuracy depending on the selected weights. The search of optimal weighting in composite likelihood approaches is based on the analysis of the joint asymptotic distribution of these composite PML estimators .

Remark 3 :

The recursive PML approach using marginal and pairwise pseudo likelihoods is especially appropriate for spatial data [9] and network analysis. Let us consider individual data $(x_i, y_i), i = 1, \dots, n$, where y_i (resp. x_i) are the observations of the endogenous (resp. exogenous) variables, and a joint conditional model with p.d.f. : $f(y_1, \dots, y_n | x_1, \dots, x_n; \theta)$, say. Let us also introduce a "distance" between individuals $d(i,j)$, where $d(i,j) = 1$, if i and j are friends, $d(i,j) = 2$, if i and j have a common friend without being friends.[10] Then we can consider a sequence of pseudo log-likelihoods :

$$\begin{aligned}
L_{1,n}(\alpha) &= \sum_{i=1}^{n} \log f_1(y_i | x_i; \alpha), \\
L_{2,1,n}(\alpha, \beta_1) &= \sum_{i,j=1; d(i,j)=1}^{n} \log f_{2,1}(y_i, y_j | x_i, x_j; \alpha, \beta_1), \\
L_{2,2,n}(\alpha, \beta_1, \beta_2) &= \sum_{i,j=1, d(i,j)=2}^{n} \log f_{2,2}(y_i, y_j | x_i, x_j; \alpha, \beta_1, \beta_2), \dots,
\end{aligned}$$

in which the effects of the explanatory variables is also made either marginal, or pairwise.

3.2 Parametric model with complicated likelihood

Let us now illustrate the practical interest of the marginal and pairwise pseudo-likelihood methods by providing examples of models with compli-

[9] See e.g. Besag (1974), Vecchia (1988), Stein et al. (2004) for application to geostatistics, such as the spatial analysis of water levels.

[10] The approach is easily extended to a non symmetric measure d, since j can be a friend of i without i being really a friend of j.

cated likelihood functions.

Example 3.1 : Dynamic Probit model.

One of the first examples developed in the literature is a dynamic probit model [see Gourieroux, Monfort, Trognon (1984)a, p 338].[11] This type of model is especially relevant for the structural model of corporate default used in the Basel regulation for Finance and Insurance [see e.g. Crouhy et al. (2000)]. We consider an homogenous population of firms and their situations either alive $y_{i,t} = 0$, or defaulted $y_{i,t} = 1, i = 1, \dots, n, t = 1, \dots, T$. In the structural model [Merton (1974)] this indicator variable is defined from the latent asset/liability ratio, up to the time of default y_i^*, say, as :

$$y_{it} = 1, \text{if } \log(A_{i,t}|L_{i,t}) < 0, y_{it} = 0, \text{ otherwise}, \tag{3.8}$$

and

$$y_i^* = \inf\{t; y_{i,t} = 1\}, \tag{3.9}$$

where $A_{i,t}$ (resp. $L_{i,t}$) denotes the asset (resp. liability) component of the balance sheet. The model is completed by a state equation providing the dynamic of the unobserved asset/liability ratios :

$$\log(A_{i,t}/L_{i,t}) = \theta_1 + \theta_2 \log(A_{i,t-1}/L_{i,t-1}) + \theta_3 u_{i,t}, \tag{3.10}$$

where the errors $u_{i,t}$ are i.i.d. standard normal.

The likelihood function based on the time-to-default is complicated. Indeed the distribution of the time-to-default y_i^* is such that :

$$P(y_i^* = H) = P[\log(A_{i,h}/L_{i,h}) > 0, h = 1, \dots, H-1, \log(A_{i,H}/L_{i,H}) < 0],$$

and involves a H-dimensional integral of a multivariate Gaussian distribution, where H can be large, up to y_i^*. In the marginal/pairwise pseudo likelihood approach based on the marginal distribution of $y_{i,t}$ and $(y_{i,t}, y_{i,t-1})$ this dimension is at most equal to 2 and one integration can be done analytically. The marginal parameter is $\alpha = [\theta_1/(1-\theta_2), \theta_3/\sqrt{1-\theta_2^2}]$, and the serial dependence parameter is $\beta = \theta_2$.

[11] See also Czado, Varin (2010) for a more recent example.

Example 3.2 : Stochastic volatility model with common factor.

Multidimensional integrals also appear in the likelihood expression of a stochastic volatility model with a dynamic factor. An example of such model is the following one, written on asset returns $y_{i,t}, i = 1, \ldots, n, t = 1, \ldots, T$ [12] :

$$\begin{aligned} y_{it} &= a_i + b_i F_t + \sigma_{i,t} u_{i,t}, i = 1, \ldots, n, \\ F_t &= \rho F_{t-1} + \eta v_t,\ |\rho| < 1, \\ \log \sigma_{i,t} &= \mu + \rho \log \sigma_{i,t-1} + \omega_{i,t}, i = 1, \ldots, n, \end{aligned}$$

where the errors $u_{i,t}, v_t, \omega_{i,t}$, are independent standard normal.

The model above contains the following unobserved state variables :

F_t is the factor with a common linear effect on the asset returns;

$\sigma_{i,t}, i = 1, \ldots, n$, are the stochastic specific volatilities of the assets.

As in Example 3.1, the interest of marginal/pairwise pseudo-likelihoods constructed from $y_t = (y_{1t}, \ldots, y_{nt})'$ is to substitute awkward high dimensional integration involved in the full likelihood with low dimensional integrals. Indeed the joint marginal distribution of $(F_t, \log \sigma_{i,t}, i = 1, \ldots, n)$ has a closed form, since these variables are marginally independent Gaussian.

Example 3.3 : Kriging or binary spatial model.

If s is a localization in $I\!R^d$ (usually $d = 2$ or 3,) a spatial process is a sequence of random variables $y^*(s)$, indexed by s, $s \in I\!R^d$. A spatial process is Gaussian if all the joint distributions of $y^*(s_1), \ldots, y^*(s_K)$ for any $s_1, \ldots, s_K$ are Gaussian. Such a process is characterized by the mean function $\mu(s) = E[y^*(s)]$ and the covariance function $cov[y^*(s_1), y^*(s_2)] \equiv \gamma(s_1, s_2)$. It is stationary if $\mu(s)$ does not depend on s, and if $\gamma(s_1, s_2)$ only depends on $s_2 - s_1$. It is isotropic if $\gamma(s_1, s_2)$ is function of the Euclidean norm $\| s_2 - s_1 \|$ of $s_2 - s_1$ denoted $c(\| s_2 - s_1 \|)$, and, in this case, the correlation function is $\rho(\| s_2 - s_1 \|) = \dfrac{c(\| s_2 - s_1 \|)}{c(0)}$.

[12] See also Engle et al. (2014) for other examples of time varying volatility-covolatility matrices.

The covariance function $\gamma(s_1, s_2)$ must be positive definite, i.e. it must satisfy :

$$\sum_{i=1}^{K}\sum_{j=1}^{K} a_i a_j \gamma(s_i, s_j) > 0,$$

for any n, any collection $\{s_1, \ldots, s_n\}$ and any numbers $a_1, \ldots, a_n$. This implies restrictions on function $\rho(r)$ in the isotropic case. The more frequent correlation functions are :

the power exponential function : $\rho(r) = \exp\left[-\left(\frac{r}{\alpha}\right)^{\beta}\right], \alpha > 0, 0 < \beta \leq 2;$

the Cauchy function : $\rho(r) = \frac{1}{\left[1 + \left(\frac{r}{\alpha}\right)^{2}\right]^{\beta}}, \alpha > 0, \beta > 0.$

The power exponential function is called exponential function if $\beta = 1$ and Gaussian function if $\beta = 2$.

Suppose for instance that $d = 2$ and that the observed process is the binary process :

$$\begin{aligned} y(s) &= 1, \text{ if } x'(s)\theta + y^*(s) > 0, \\ &= 0, \text{ otherwise,} \end{aligned}$$

where $y^*(s)$ is an isotropic Gaussian process with $\mu(s) = 0$, unit variance and correlation function $\rho(r)$, whereas $x(s)$ is a vector of exogenous variable and θ a parameter. The likelihood function is clearly untractable since it involves high dimensional integrals. However the marginal and pairwise distributions of $y(s)$ are easily computed since :

$$P[y(s) = 1] = \Phi[x'(s)\theta],$$

and, for instance,

$$P[y(s_1) = 1, y(s_2) = 1] = \psi[x'(s_1)\theta, x'(s_2)\theta, \rho(\| s_2 - s_1 \|)],$$

where $\psi(z_1, z_2, \rho)$ is the c.d.f of $N\left[\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}\right]$.

Therefore the marginal/pairwise pseudo likelihood approach is easily implemented [see Heagerty Lele (1998)].

4 Pseudo-Maximum Likelihood and Conditional Estimating Equations

4.1 Conditional estimating equation

In semi-parametric econometric models the parameter of interest is frequently defined by means of conditional estimating equations [13] :

$$E_0[\psi(y_t, x_t; \theta_0)|x_t] = 0, \tag{4.1}$$

where θ_0 is the true value of the parameter of interest, ψ a known function, x_t explanatory variables, which can include exogenous variables as well as lagged endogenous variables. We assume that the true value θ_0 is identifiable from this set of restrictions, that is :

$$E_0[\psi(y_t, x_t; \theta)|x_t] = 0, \text{ a.s. in } x_t \Rightarrow \theta = \theta_0. \tag{4.2}$$

The system (4.1) is not sufficient to characterize the complete distribution of the observable variables, since both the distribution of x_t and the part of the conditional distribution of y_t given x_t, which is not function of θ, are left unspecified. This is why such a model is semi-parametric.

Well-known examples of semi-parametric models are given below.

i) The mean regression model is written as :

$$y_t = a(x_t; \theta_0) + u_t, \text{ where } E_0(u_t|x_t) = 0. \tag{4.3}$$

Equivalently the true value is characterized by the set of moment restrictions :

$$E_0[y_t - a(x_t; \theta_0)|x_t] = 0, \text{ a.s. in } x_t, \tag{4.4}$$

that is, we have :

$$\psi(y_t, x_t; \theta) = y_t - a(x_t; \theta). \tag{4.5}$$

ii) The quantile regression model is written as :

$y_t = a(x_t; \theta_0) + u_t$, where the conditional α-quantile of u_t is equal to zero. This condition on the distribution of the error can be written as :

[13] see Godambe (1960) for the first introduction of the notion of estimating equation.

$$E_0[\mathbb{1}_{u_t<0} - \alpha | x_t] = 0 \text{ a.s in } x_t, \tag{4.6}$$

that is, we have :

$$\psi(y_t, x_t; \theta) = \mathbb{1}_{y_t - a(x_t;\theta)<0} - \alpha, \tag{4.7}$$

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

4.2 Consistent PML estimators

Let us assume that the joint process (x_t, y_t) is strongly stationary and consider a PML estimator :

$$\hat{\theta}_T = \arg\max_\theta \sum_{t=1}^{T} \log g(y_t | x_t; \theta), \tag{4.8}$$

where $g(y|x, \theta)$ is a pseudo density for y_t given x_t. When function g is continuous with a right derivative, the asymptotic first-order conditions for the pseudo true value[14] θ_0^* are :

$$E_0 \frac{\partial \log g^+}{\partial \theta}(y|x; \theta_0^*) = 0, \tag{4.9}$$

where $\dfrac{\partial \log g^+}{\partial \theta}$ denotes the right derivative with respect to θ.

By applying an appropriate version of Farkas lemma [see Gourieroux, Monfort, Renault (1987), Gourieroux, Monfort (1995), Chapter 8), we get the following proposition :

Theorem 4.: Under standard regularity conditions and identification assumption (4.2), the PML estimator of θ is a consistent estimator of θ_0, if and only if :

$$\frac{\partial \log g^+}{\partial \theta}(y|x; \theta) = \wedge(x; \theta)\psi(y, x; \theta), \text{ a.s. in } x, y,$$

where $\wedge(x; \theta)$ is a matrix of size $dim\theta \times dim\psi$, with full column rank.

Let us now apply this proposition to mean and quantile regression models.

[14] Note that we now denote differently the pseudo-true value θ_0^* from the true value.

i) **Mean regression model.**

Let us consider the regression model (4.3), a pseudo-distribution family $f(y;m)$ for y indexed by the mean m and a well-specified regression function a. We have : $g(y_t|x_t;\theta) = f[y_t;a(x_t;\theta)]$. When f is differentiable with respect to m, the condition of Theorem 4 becomes :

$$\frac{\partial \log g^+}{\partial \theta}(y|x;\theta) = \frac{\partial \log f}{\partial m}(y;a(x_t;\theta))\frac{\partial a(x_t;\theta)}{\partial \theta} = \wedge(x_t;\theta)[y - a(x_t;\theta)],$$

a.e. in x_t, y_t, θ.

This implies :

$$\frac{\partial \log f}{\partial m}(y;m) \equiv \lambda(m)(y-m), \text{ say,}$$

and, after integrating with respect to m for fixed y, a pseudo distribution of the type :

$$f(y;m) = \exp[c(m)y + b(m) + d(y))], \text{ say,} \tag{4.10}$$

with $\frac{dc}{dm}(m) = \lambda(m), \frac{db(m)}{dm} = -m\lambda(m)$, belonging to the so-called exponential linear family [Gourieroux, Monfort, Trognon (1984)b]. We deduce the following Corollary :

Corollary 1 : The PML estimator of a mean regression is consistent if the pseudo distribution indexed by mean m is chosen in an exponential linear family.

ii) **Quantile regression model**

Let us now consider the quantile regression

$$y_t = a(x_i;\theta) + u_t,$$

where the conditional α-quantile of u_t is equal to zero, and a family $f(y,m)$ of pseudo distributions indexed by their α-quantile m.

When f admits a right derivative with respect to m, the condition of Theorem 4, with $g(y_t|x_t;\theta) = f[y_t;a(x_t,\theta)]$, becomes :

$$\frac{\partial \log f^+}{\partial m}[y; a(x_t;\theta)]\frac{\partial a}{\partial \theta}(x_t;\theta) = \wedge(x_t;\theta)[\mathbb{1}_{y_t - a(x_t;\theta)<0} - \alpha].$$

This implies :

$$\frac{\partial \log f^+(y;m)}{\partial m} = \lambda(m)[\mathbb{1}_{(y_t - m<0)} - \alpha],$$

and

$$f(y;m) = \exp[-\alpha[A(y)-A(m)]^+ - (1-\alpha)[A(y)-A(m)]^- + d(y)],$$

where A is the cumulative function of a nonnegative measure on $\mathbb{R}$. A member of this family is obtained by taking $A(y) = \beta y, \beta > 0$ (Lebesgue measure), that is the skewed Laplace distribution with density :

$$f(y;m) = \alpha(1-\alpha)\beta \exp\{-\beta[\alpha(y-m)^+ + (1-\alpha)(y-m)^-]\}.$$

4.3 Pseudo maximum likelihood vs moment method.

The conditional estimating equations are the basis of other consistent estimation methods, that are the methods of moments [see Godambe, Thompson (1978), Hansen (1982)]. The principle of such methods consists in replacing the conditional estimating equations by unconditional ones. More precisely let us introduce a matrix Z_t of size $dim\theta \times dim\psi$, such that the elements of Z_t are functions of the conditioning variables x_t. They are called instrumental variables, or instruments. Then we deduce from conditional restrictions (4.1) the unconditional restrictions :

$$E_0[Z_t \psi(y_t, x_t; \theta)] = 0. \tag{4.11}$$

The associated instrumental variable estimator is the solution of the associated sample moments conditions[15] :

[15] When this system admits a solution. Otherwise, $\hat{\theta}_T$ is defined as :
$\hat{\theta}_T = \arg\min_\theta ||\frac{1}{T}\sum_{t=1}^{T} Z_t \psi(y_t, x_t; \theta)||^2$.

$$\frac{1}{T}\sum_{t=1}^{T} Z_t \psi(y_t, x_t; \tilde{\theta}_T) = 0. \tag{4.12}$$

The PML estimators introduced in Section 4.2 are moment estimators corresponding to a special choice of instruments. For expository purpose let us focus on a mean regression model (4.3). The asymptotic FOC corresponding to the linear exponential family are :

$$E_0\left[\frac{\partial a(x_t;\theta_0)}{\partial\theta}\lambda[a(x_t;\theta_0)][y_t - a(x_t;\theta_0)]\right] = 0. \tag{4.13}$$

The associated set of instruments is :

$$Z_t = \frac{\partial a(x_t;\theta_0)}{\partial\theta}\lambda[a(x_t;\theta_0)]. \tag{4.14}$$

These instruments involve two components, that are the sensitivities of the regression w.r.t. the parameters, i.e. $\frac{\partial a(x_t;\theta_0)}{\partial\theta}$, and another component $\lambda(m)$ giving the sensitivity of the pseudo log p.d.f. with respect to the mean parameter.

It is known that there exist an optimal choice of instruments Z_t for the method of moment [see Hansen, Singleton (1982), Godambe, Thompson (1984), Liang, Zeger (1986)]. However this optimal choice depends on the true distribution and is difficult to implement in practice. An appropriate choice of the pseudo distribution, that is of λ, can partly circumvent this difficulty. Indeed the interpretation of the endogenous variable can suggest, which type of exponential linear family has likely to be chosen, as illustrated below.

i) The (pseudo) **Poisson regression model** [Gourieroux, Monfort, Trognon (1984)c].

Let us consider independent observations (x_i, y_i) such that :

$$y_i \sim \mathcal{P}[\exp(x_i'\theta_0 + v_i)],$$

where v_i is an unobserved heterogeneity, independent of x_i such that $E_0(\exp v_i) = 1$. Since the Poisson family is an exponential linear family with :

$$f(y;m) = \exp(y \log m - m - \log y!),$$

the associated PML estimator corresponds to the misspecified model :

$$y_i \sim \mathcal{P}(\exp(x_i'\theta)),$$

in which the omitted heterogeneity is set to 0. Even, if the pseudo-model is misspecified, the feature of a count variable is still taken into account and the efficiency loss not so large, at least if v_i is not too large.

ii) The (pseudo) **autoregressive conditional duration model** [Engle, Russell (1998)].

This dynamic model has been introduced for analyzing the intertrade durations in financial markets. The model for a given asset is :

$$y_t \sim \gamma(1, \exp(\theta_0 + \theta_1 y_{t-1} + v_t)),$$

where y_t is the intertrade duration between trades t and $t+1$, and v_t an omitted variable such that $E_0(\exp v_t | \underline{y_{t-1}}) = 1$. Since the family of exponential distributions is a linear exponential family, with :

$$f(y; m) = \exp(-y/m - \log m),$$

the associated PML estimator corresponds to the misspecified model :

$$y_t \sim \gamma(1, \exp(\theta_0 + \theta_1 y_{t-1})),$$

in which the omitted variable is set to zero. In this framework, the pseudo exponential distribution is appropriate for a duration variable.

5 Transformation Models

5.1 The model

In a transformation model, the observations of the endogenous variables are written as functions of explanatory variables and i.i.d. errors :

$$y_t = H(x_t, u_t; \theta), \tag{5.1}$$

where y_t are the n endogenous variables, x_t the explanatory variables, which are exogenous or lagged endogenous variables, u_t n-dimensional errors and θ a vector of parameters. Moreover we make the following assumptions :

Assumption A.1 : i) The errors u_t are independent identically distributed (i.i.d.).

ii) The current and past values of the explanatory variables $\underline{x_t} = (x_t, x_{t-1}, \ldots)$ are independent of the current and future values of the errors $\bar{u}_t = (u_t, u_{t+1}, \ldots)$.

iii) The transformation H is known and a one-to-one function from u to y, for any given x, θ.

We will discuss the consistency of PML estimators of θ, or of a subvector of θ, in such a transformation model. We will see that the PML estimators are consistent for a large set of pseudo distributions of the errors. Compared to the analysis of Section 4, this is a consequence of the stronger assumption made on the errors. Typically, if the errors are such that $E_0 u_t = 0$, we have automatically : $E_0(u_t | \underline{x_t}) = 0$ under Assumption A1 i), ii), that is a condition of martingale difference sequence used for instance in the analysis of PML applied to a nonlinear regression model (see Section 4.2). But we have also $E_0[h(u_t)|\underline{x_t}] = 0$, for any nonlinear integrable transformation of u_t.

However, more structure has to be introduced on the transformation model. In fact the transformation models considered in this section are of the form :

$$y_t^* = H_1\{\Pi_{j=1}^J \exp[a_j^*(x_t;\theta)C_j] H_2(u_t^*)\}, \tag{5.2}$$

where H_1, H_2 are known one-to-one transformations on $I\!R^n, C_j, j = 1, \ldots, J$ are (n,n) matrices, $a_j^*(.,.), j = 1, \ldots, J$ are scalar index functions, x_t, u_t^* satisfy Assumption A1 i), ii).

After appropriate transformations, it is equivalent to consider the transformation model :

$$y_t = \Pi_{j=1}^J \exp[a_j^*(x_t;\theta)C_j] u_t, \tag{5.3}$$

where : $y_t = H_1^{-1}(y_t^*), u_t = H_2(u_t^*)$.

Model (5.3) involves linear transformations of the errors constructed from exponential of matrices. Let us briefly review the definition of such an exponential and its main properties. The exponential of a matrix C is defined by :

$$\exp(aC) = \sum_{h=0}^{\infty} \frac{a^h C^h}{h!}. \tag{5.4}$$

This matrix is well defined (that is the series exists) for any scalar a and matrix C.

When a varies, we get a Lie group of linear transformations such that :

$$\exp(aC)\exp(bC) = \exp[(a+b)C], \tag{5.5}$$

$$[\exp(aC)]^{-1} = \exp(-aC), \tag{5.6}$$

$$\left[\frac{d}{da}\exp(aC)\right]_{a=0} = C. \tag{5.7}$$

In Lie group theory, C is the infinitesimal generator of the group, a is the velocity, and $a \to \exp(aC)$ is called a geodesic.

The Lie groups can be combined, that is, we can also consider the transformations :

$$(a_1, a_2) \to \exp(a_1 C_1)\exp(a_2 C_2).$$

When C_1 and C_2 commute, we can write :

$$\exp(a_1C_1)\exp(a_2C_2) = \exp(a_2C_2)\exp(a_1C_1) = \exp(a_1C_1 + a_2C_2),$$

and this set of compound transformations $\exp(a_1C_1 + a_2C_2), a_1, a_2$ varying, defines another Lie group with dimension 2.

When C_1 and C_2 do not commute, the set of compound transformations :

$\exp(a_1C_1)\exp(a_2C_2), a_1, a_2$ varying, is no longer a Lie group, but can still have interesting properties.

5.2 Consistent PML Approach

5.2.1 Single Index Model (j=1)

Let us consider the transformation model (5.3) with $J = 1$, in which we introduce the velocity intercept parameter α_0 :

$$y_t = \exp([a(x_t;\beta_0) + \alpha_0]C)u_t, \tag{5.8}$$

with $\theta_0 = (\alpha_0, \beta_0)', a^*(x_t;\theta_0) = a(x_t;\beta_0) + \alpha_0, \alpha_0, \beta_0$ denoting the true values of the parameters.

Let us also introduce a pseudo-distribution g for the error. The PML estimator is defined by :

$$\begin{aligned}(\hat{\alpha}_T, \hat{\beta}_T) &= \arg\max_{\alpha,\beta} \sum_{t=1}^{T} \{\log g[\exp(-[a(x_t;\beta)+\alpha]C)y_t] \\ &- (a(x_t;\beta)+\alpha)TrC\}. \qquad (5.9)\end{aligned}$$

Note the special form of the log-Jacobian. Indeed we have :

$$\begin{aligned} & \log|\det(\exp[-[a(x_t;\beta)+\alpha]C]) \\ = & \log|\Pi_{i=1}^{n}\lambda_i(x_t;\alpha,\beta,c)| \\ = & \log \Pi_{i=1}^{n} \exp[-(a(x_t;\beta)+\alpha)\tilde{\lambda}_i] \\ = & -(a(x_t;\beta)+\alpha)\sum_{i=1}^{n}\tilde{\lambda}_i = -[a(x_t;\beta)+\alpha]TrC, \end{aligned}$$

where $\lambda_i(x_t;\alpha,\beta,c)$ (resp. $\tilde{\lambda}_i$) are the eigenvalues of $\exp(-[a(x_t;\beta)+\alpha]C)$ (resp. C).[16]

The following result has been derived in Gourieroux, Monfort, Zakoian (2016) :

Theorem 5: i) The PML estimator $\hat{\beta}_T$ is a consistent estimator of β_0 for any choice of the pseudo-distribution g.

ii) The PML estimator $\hat{\alpha}_T$ is a consistent estimator of α_0, if moreover : $Tr\{C[E_0\left(u\dfrac{\partial \log g(u)}{\partial u'}\right) + Id]\} = 0$,where E_0 denotes the expectation w.r.t. the true distribution of the error.

If the initial econometric model already contains a velocity intercept parameter, we get the consistency of all parameters characterizing the sensitivity of the index function w.r.t. the explanatory variables. If the initial econometric model contains no velocity intercept, such a parameter has to

[16] The derivation of the log-Jacobian is performed for real eigenvalues, but is easily extended to the case of conjuguate complex eigenvalues.

be artificially introduced to capture all the asymptotic bias due to the misspecification of the error distribution and to recover the consistency of the PML estimators of the sensitivity parameters.

For a given pseudo-distribution g, the additional restriction in Theorem 5 ii) defines the set of true distributions such that $\hat{\theta}_T = (\hat{\alpha}_T, \hat{\beta}'_T)'$ converges to $\theta_0 = (\alpha_0, \beta'_0)'$. This is a set of unconditional moment restrictions on u. For instance, for a standard multivariate Gaussian pseudo-distribution : $\log g(u) = -\frac{n}{2}\log(2\pi) - \frac{1}{2}u'u$, the additional restrictions become :

$$Tr\{C(E_0(u_t u'_t) - Id)\} = 0.$$

The possibility to estimate consistently by PML the sensitivity parameters β_0 is due to an interpretation of the first-order conditions of the PML optimization, as covariance conditions. At the true value β_0 these conditions turn out to involve covariances between functions of x_t and functions of u_t. They are satisfied because of the independence assumption A1 i), ii).

The PML estimator of β is asymptotically normal with an asymptotic variance-covariance matrix given by the sandwich formula [see Appendix 2]. In the transformation model of interest, it becomes :

$$V_{as}[\sqrt{T}(\hat{\beta}_T - \beta_0)) = \frac{i^*}{j^{*2}} \left(V_0 \left[\frac{\partial a(x_t;\beta_0)}{\partial \beta} \right] \right)^{-1}, \tag{5.10}$$

where i^*, j^* depend on the difference between the pseudo and true values of the velocity intercept $\alpha_0^* - \alpha_0$ and on the pseudo and true distributions of u only. In particular these asymptotic variance-covariance matrices for different pseudo-distributions are proportional, and also proportional to the inverse of the variance-covariance matrix of the informative underlying explanatory variables.

5.2.2 Multi Index Model

The consistency result for the PML estimator of β_0 (resp. α_0, β_0) can be extended to the multi-index framework. For expository purpose we consider $J = 2$ [see Gourieroux, Monfort, Zakoian (2016) for the general case]. The transformation model is defined by :

$$y_t = \exp[a_1(x_t;\beta_0)C_1] \exp[a_2(x_t;\beta_0)C_2] \exp(\alpha_{10}C_1) \exp(\alpha_{20}C_2) u_t. \tag{5.11}$$

An assumption on generators C_1, C_2 is required.

Assumption A.1 : Closure under commutation.

$$\exp(a_1 C_1) \exp(a_2 C_2) = \exp[\gamma_2(a_1, a_2) C_2] \exp[\gamma_1(a_1, a_2) C_2],$$

for a one-to-one mapping $\gamma = (\gamma_1, \gamma_2)$.

Thus, up to a change of velocities, the exponential operators can be commuted. The assumption is in particular satisfied if the generators themselves commute : $C_1 C_2 = C_2 C_1$.

Then we have the following Proposition :

Theorem 6.: Under closure under commutation, the PML estimator of β_0 is consistent for any pseudo-distribution.

Thus we have to introduce a number of velocity intercepts α_1, α_2 equal to the number of index functions and put them at the right places in the pseudo likelihood in order to capture all the bias due to the misspecification of the error distribution.

5.2.3 Application to GARCH Type Models

The consistent PML approach can in particular be used for consistent estimation of univariate as well as multivariate GARCH models. Examples are described below with the econometric model and their associated generators.

Example 5.1 : One-dimensional GARCH.

The model is :

$$y_t = \exp(a(x_t; \beta_0) + \alpha_0) u_t \equiv \sigma_0 \exp a(x_t; \beta_0) u_t.$$

We have $n = J = 1$; the Lie group is the group of homotheties and the additional parameter is (after transformation) a scale parameter. The consistency of the PML estimator in this special case has been derived in Berkes, Horvath (2004) [see also Fan et al. (2014)].

Example 5.2 : One-dimensional ARCH-in-mean model.

This in an example with $n = 1, J = 2$, constructed from the group of translations : $u \to u + a_1$, and the group of homotheties : $u \to \exp(a_2)u$.

These two groups do not commute, but they satisfy the closure under commutation, since :

$$u \to u + a_1 \to \exp a_2(u + a_1) = u \exp a_2 + a_1 \exp a_2,$$

$$u \to \exp a_2 u \to u \exp a_2 + a_1,$$

provide similar result up to a change of velocities.

The econometric model is :

$$\begin{aligned} y_t &= \exp a_2(x_t; \beta_0)[\alpha_{10} + \exp(\alpha_{20})u_t] + a_1(x_t; \beta_0) \\ &= a_1(x_t; \beta_0) + \alpha_{10} \exp[a_2(x_t; \beta_0)] + \exp[a_2(x_t; \beta_0) + \alpha_{20}]u_t. \end{aligned}$$

This is the framework studied in Newey, Steigerwald (1997). We need two adjustement velocity intercepts, that are a scale effect for the volatility, and a term including the volatility effect, that is a risk premium, for the drift, respectively.

Example 5.3 : Conditional Rotation.

The econometric model is :

$$\begin{aligned} y_t &= \begin{pmatrix} \cos[a(x_t; \beta_0) + \alpha_0] & \sin[a(x_t; \beta_0) + \alpha_0] \\ -\sin[a(x_t; \beta_0) + \alpha_0] & \cos[a(x_t; \beta_0) + \alpha_0] \end{pmatrix} u_t \\ &\equiv R[a(x_t; \beta_0) + \alpha_0]u_t. \end{aligned}$$

It is based on the group of rotations with generator $C = \begin{pmatrix} 0 & -1 \\ 1 & 0 \end{pmatrix}$.

If the components of u_t are uncorrelated, the model has a volatility-covolatility matrix of the form :

$$V_0(y_t|x_t) = R[a(x_t; \beta_0) + \alpha_0] \begin{pmatrix} \lambda_{10} & 0 \\ 0 & \lambda_{20} \end{pmatrix} R'[a(x_t; \beta_0) + \alpha_0].$$

Thus the focus is on the time varying direction of largest (resp. smallest) risk, that is the eigenvector associated with the largest eigenvalue (resp. the smallest eigenvalue).

Example 5.4 : Cholesky GARCH model.

This example corresponds to the two following Lie groups :
First Lie group with commuting generators :

$$C_1 = \begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix}, C_2 = \begin{pmatrix} 0 & 0 \\ 0 & 1 \end{pmatrix},$$

and geodesic : $\exp(a_1 C_1 + a_2 C_2) = \begin{pmatrix} \exp a_1 & 0 \\ 0 & \exp a_2 \end{pmatrix}.$

Second Lie group with generator $C_3 = \begin{pmatrix} 0 & 1 \\ 0 & 0 \end{pmatrix}$ and geodesic :
$\exp(a_3 C_3) = \begin{pmatrix} 1 & a_3 \\ 0 & 1 \end{pmatrix}.$

These Lie groups do not commute, but are closed under commutation. Indeed we have :

$$\begin{aligned} \begin{pmatrix} \exp a_1 & 0 \\ 0 & \exp a_2 \end{pmatrix} \begin{pmatrix} 1 & a_3 \\ 0 & 1 \end{pmatrix} &= \begin{pmatrix} \exp a_1 & a_3 \exp a_1 \\ 0 & \exp a_2 \end{pmatrix}, \\ \begin{pmatrix} 1 & a_3 \\ 0 & 1 \end{pmatrix} \begin{pmatrix} \exp a_1 & 0 \\ 0 & \exp a_2 \end{pmatrix} &= \begin{pmatrix} \exp a_1 & a_3 \exp a_2 \\ 0 & \exp a_2 \end{pmatrix}. \end{aligned}$$

with the same form up to a one-to-one change on velocity parameters.

The associated econometric model is a Cholesky-ARCH model [See Dellaortas, Pourahmadi (2012)] :

$$y_t = \begin{pmatrix} \exp[a_1(x_t;\beta_0)+\alpha_{10}] & \exp a_1(x_t;\beta_0)[\alpha_{30}\exp\alpha_{10} + a_3(x_t;\beta_0)\exp\alpha_{20}] \\ 0 & \exp[a_2(x_t;\beta_0)+\alpha_{20}] \end{pmatrix} u_t.$$

We have scale adjustments for the diagonal elements and a bias adjustment à la Newey, Steigerwald (1997), via the introduction of a risk premium for the nonzero off diagonal element.

5.2.4 Normalized data

The consistency of the PML estimator of β is due to the simple expression of the pseudo log-likelihood : $\log g[\exp(-aC)y] - aTrC$, to the possibility to (partly) commute the Lie groups within the log(.), and to the application of the Jacobian formula. The Jacobian formula is valid when the Lie group is applied to an invariant manifold of $I\!R^n$, even with a dimension strictly smaller than n. Thus the consistency results apply to normalized data with different normalizations.

Example 5.5 : Lie group on the simplex

Let us denote $y_t^* = (y_{t1}^*, \ldots, y_{tn}^*)'$ the expenditures of a given household on month t. The associated budget shares $y_t = (y_{t1}, \ldots, y_{tn})'$, where $y_{tj} = y_{tj}^* / \sum_{j=1}^{n} y_{tj}^*$, take values in the simplex :

$$U = \{y : y_j \geq 0, \forall j, \sum_{j=1}^{n} y_j = 1\}.$$

An econometric model defined on the simplex is :

$$y_t = \exp[a^*(x_t;\theta)C]u_t,$$

where u_t is valued in the simplex, the generator is such that : $c_{ij} \geq 0, \forall i \neq j, c_{ii} = -\Sigma_{j\neq i} c_{ij}$, and generates stochastic matrices $\exp(aC)$, that are matrices with nonnegative elements, with rows summing up to one.

Example 5.6 : Lie group on the unit sphere

Let us now consider the allocation of an arbitrage portfolio : $y_t^* = (y_{t1}^*, \ldots, y_{tn}^*)'$, where $j = 1, \ldots, n$, are the financial assets in which the portfolio is invested. Since $\sum_{j=1}^{n} y_{tj}^* = 0$ (arbitrage), the normalization of Example 5.5 can no longer be used. The practice is to normalize by $||y_t^*|| = (\sum_{j=1}^{n} y_{tj}^{*2})^{1/2}$, which measures the magnitude of the financial leverage [17]. The normalized portfolio allocation : $y_t = (y_{t1}, \ldots, y_{tn})'$, where $y_{tj} = y_{tj}/||y_t^*||$, is on the unit sphere.

An econometric model on the unit sphere is :

$$y_t = \exp[a^*(x_t; v)C]u_t,$$

where u_t is valued in the unit sphere generator C is antisymetric : $C' = -C$, and generates a set of orthogonal matrices $\exp(aC)$.

6 Concluding Remarks

Our paper illustrates the interest of the modern approach used to analyse the consistency of estimators defined by optimizing a criterion function. This modern approach was first introduced for nonlinear least squares estimation method in Jennrich (1969), Malinvaud (1966), (1970). We apply this approach to estimators defined by optimizing a misspecified likelihood function, that are the pseudo maximum likelihood estimators, or a combination of such misspecified likelihoods, that are composite pseudo maximum likelihood estimators. We emphasize the variety of consistent PML approaches and how they depend on the assumptions on the error terms and the choice of the pseudo-likelihoods.

Three directions of research emerged from the literature on pseudo maximum likelihood.

i) There can exist other situations in which PML estimators are consistent and do not enter in the examples discussed in our paper. A typical example is the literature on Independent Component Analysis (ICA) in which consistent

[17] Typically, if $n = 2$, we get : $y_{t2}^* = -y_{t1}^*$, and $||y_t^*|| = \sqrt{2}|y_{t1}^*|$. Larger $|y_{t1}^*|$, larger is the leverage in this arbitrage portfolio.

PML requires a group of orthogonal transformations used in an appropriate way [see e.g. Gourieroux, Monfort, Renne (2016)].

ii) When the PML estimators are not consistent, the asymptotic bias can be adjusted by indirect inference. This allows to also consider all these bias adjusted PML approaches. This leads to the class of composite indirect inference estimators [see e.g. Gourieroux, Monfort (2016)].

iii) As seen from the examples provided in the paper, it is easy in practice to find a lot of PML estimators, which are consistent for a given econometric model. Their comparison can be the basis of tests and/or diagnostic tools for considering the validity of the initial econometric model.

References

[1] Amemiya, T. (1985) : "Advanced Econometrics", Harvard University Press.

[2] Basawa, I., Feignin, P., and C., Heyde (1976) : "Asymptotic Properties of Maximum Likelihood Estimators for Stochastic Processes", Sankhya, The Indian Journal of Statistics, 38, 259-270.

[3] Bassett, G., and R., Koenker (1978) : "Regression Quantiles", Econometrica, 46, 33-50.

[4] Berkes, I., and L., Horvath (2004) : "The Efficiency of the Estimators of the Parameters in GARCH Processes", The Annals of Statistics, 32, 633-645.

[5] Besag, J. (1974) : "Spatial Interaction and the Statistical Analysis of Lattice Systems", JRSS B, 36, 192-236.

[6] Bierens H.,(2004): "Introduction to the Mathematical and Statistical Foundations of Econometrics", Cambridge Univ. Press.

[7] Bollerslev, T., and J., Woolridge (1992) : "Quasi-Maximum Likelihood Estimators and Inference in Dynamic Models with Time-Varying Covariances", Econometric Reviews, 11, 143-172.

[8] Chandler, R., and S., Bate (2007) : "Inference for Clustered Data Using the Independence Log-Likelihood", Biometrika, 94, 167-183.

[9] Chen, Y. (2015) : "Semi-Parametric Time Series Models with Log-Concave Innovations : Maximum Likelihood Estimation and its Consistency", Scandinavian Journal of Statistics, 42, 1-31.

[10] Cox, D. (1975) : "Partial Likelihood", Biometrika, 62, 269-276.

[11] Cox, D., and N., Reid (2004) : "A Note on Pseudolikelihood Constructed from Marginal Densities", Biometrika, 91, 729-737.

[12] Cramer, H. (1946) : "Mathematical Methods of Statistics", Cambridge Univ. Press, 500-504.

[13] Crouhy, M., Galai, D., and R., Mark (2000) : "A Comparative Analysis of Current Credit Risk Models", Journal of Banking and Finance, 29, 59-117.

[14] Czado, C., and C., Varin (2010) : "A Mixed Autoregressive Probit Model for Ordinal Longitudinal Data", Biostatistics, 11, 127-138.

[15] Dellaportas, P., and M., Pourahmadi (2012) : "Cholesky-GARCH Models with Application to Finance", Stat. Comput., 22, 849-855.

[16] Engle, R., Lilien, D., and R., Robins (1987) : "Estimating Time Varying Risk Premia in the Term Structure : the ARCH-M Model", Econometrica, 55, 391-407.

[17] Engle, R., and J., Russell (1998) : "Autoregressive Conditional Duration: A New Model for Irregularly Spaced Transaction Data", Econometrica, 66, 1127-1162.

[18] Engle, R., Pakel, C., Shephard, N., and K., Sheppard (2014) : "Fitting and Testing Vast Dimensional Time Varying Covariance Models", DP NYU.

[19] Fan, J., Qi., L., and D., Xiu (2014) : "Quasi Maximum Likelihood Estimation of GARCH Models with Heavy-Tailed Likelihood", Journal of Business and Economic Statistics, 32, 178-191.

[20] Francq, C., Lepage, G., and J.M., Zakoian (2011) : "Two State Non Gaussian QML Estimation of GARCH Models and Testing the Efficiency of the Gaussian QMLE", Journal of Econometrics, 165, 246-257.

[21] Godambe, V. (1960) : "An Optimum Property of Regular Maximum Likelihood Estimation", Ann. Math. Statist., 31, 1208-1211.

[22] Godambe, V., and M., Thompson (1978) : "Some Aspects of the Theory of Estimating Equations", J. Stat. Planning and Inference, 2, 95-104.

[23] Godambe, V., and M., Thompson (1984) : "Robust Estimation Through Estimating Equations", Biometrika, 71, 115-125.

[24] Gourieroux, C., and A., Monfort (1995) : "Statistics and Econometric Models", Cambridge Univ. Press, French version (1989), Economica.

[25] Gourieroux, C., and A., Monfort (2016) : ”Composite Indirect Inference with Application to Corporate Risks”, CREST-DP.

[26] Gourieroux, C., Monfort, A., and E., Renault (1987) : ”Consistent M-Estimators in a Semi-Parametric Model”, INSEE DP.

[27] Gourieroux, C., Monfort, A., and J.P. Renne (2016) : ”Statistical Inference for Independent Component Analysis with Application to Structural VAR Models”, forthcoming Journal of Econometrics.

[28] Gourieroux, C., Monfort, A., and A., Trognon (1984)a : ”A General Approach to Serial Correlation”, Econometric Theory, 1, 315-340.

[29] Gourieroux, C., Monfort, A., and A., Trognon (1984) b : ”Pseudo Maximum Likelihood Methods : Theory”, Econometrica, 52, 681-700.

[30] Gourieroux, C., Monfort, A., and A., Trognon (1984)c : ”Pseudo Maximum Likelihood Methods : Application to Poisson Models”, Econometrica, 52, 700-720.

[31] Gourieroux, C., Monfort, A., and J.M., Zakoian (2016) : ”Pseudo Maximum Likelihood and Lie Groups of Linear Transformations”, CREST DP.

[32] Hansen, L. (1982) : ”Large Sample Properties of Generalized Method of Moments Estimators”, Econometrica, 50, 1029-1054.

[33] Hansen, L. (2012): ”Proofs for Large Sample Properties of Generalized Method of Moment Estimators”, Journal of Econometrics, 170, 325-330.

[34] Hansen, L., and K., Singleton (1982) : ”Generalized Instrumental Variables Estimation of Nonlinear Rational Expectations Models”, Econometrica, 50, 1269-1286.

[35] Heyde, C. (1997) : ”Quasi-Likelihood and its Application : A General Approach to Optimal Parameter Estimation”, Springer Series in Statistic, New-York, Springer.

[36] Heagerthy P. and S., Lele (1998) : ”A Composite Likelihood Approach to Binary Data”, JASA, 93, 443, 1099-1111.

[37] Huber, P. (1964) : ”Robust Estimation of a Location Parameter”, Annals of Mathematical Statistics, 35,

[38] Huber, P. (1967) : "The Behavior of Maximum Likelihood Estimators Under Nonstandard Conditions", Proc. Fifth Berkeley Symposium, Math. Stat. Prob., 1, 221-233.

[39] Huber, P. (1974) : "Robust Statistics", New-York, Wiley, p43.

[40] Huzurbazar, V. (1948) : "The Likelihood Equation, Consistency and the Maxima of the Likelihood Function", Annals of Eugenics, 14.

[41] Jennrich, R. (1969) : "Asymptotic Properties of Nonlinear Least Squares Estimation", The Annals of Mathematical Statistics, 40, 633-643.

[42] Kiefer, J., and J., Wolfowitz (1956) : "Consistency of the Maximum Likelihood Estimator in the Presence of Infinitely Many Incidental Parameters", Annals of Mathematical Statistics, 27 887-906.

[43] Kim, J., and D., Pollard (1990): "Cube Root Asymptotics", The Annals of Statistics,18, 191-219.

[44] Larribe, F., and P., Fearnhead (2011) : "On Composite Likelihoods in Statistical Genetics", Statist. Sinica, 21, 43-69.

[45] Liang, K., and S., Zeger (1986) : "Longitudinal Data Analysis Using Generalized Linear Models", Biometrika, 69, 503-512.

[46] Lindsay, B. (1982) : "Conditional Score Functions : Some Optimality Results", Biometrika, 69, 503-512.

[47] Malinvaud, E. (1966) : "Statistical Methods of Econometrics", North-Holland, translation of a book first published in French in 1964 by Dunod.

[48] Malinvaud, E. (1970) : "The Consistency of Nonlinear Regressions", Annals of Mathematical Statistics, 41, 956-969.

[49] Mardia, K., Hughes, G., Taylor, C., and H., Singh (2008) : "A Multivariate von Mises Distribution with Applications to Bioinformatics", Canadian Journal of Statistics, 36, 99-109.

[50] McCullagh, P. (1984) : "Quasi-Likelihood Functions", Annals of Statistics, 11, 59-67.

[51] Meddahi, N., and E., Renault (1998): "Quadratic M-estimators for ARCH-Type Processes", WP CIRANO 98-29.

[52] Merton, R. (1983) : "On the Pricing of Corporate Debt : The Risk Structure of Interest Rates", Journal of Finance, 20, 449-470.

[53] Neyman, J., and E., Scott (1948) : "Consistent Estimates Based on Partially Consistent Observations", Econometrica, 16, 1-32.

[54] Newey, W., and D., Steigerwald (1997) : "Asymptotic Bias for Quasi-Maximum Likelihood Estimators in Conditional Heteroscedastic Models", Econometrica, 65, 587-599.

[55] Rubin, H. (1950) : "Consistency of Maximum Likelihood Estimates in the Explosive Case", in T.C. Koopmans, ed.

[56] Stein, M., Chi, Z., and L., Welty (2004) : "Approximating Likelihoods for Large Spatial Data Sets", JRSS B, 66, 275-296.

[57] Van der Vaart, A.W. (1998): "Asymptotic Statistics", Cambridge Univ. Press.

[58] Varin, C., Reid, N., and D., Firth (2011) : "An Overview of Composite Likelihood Methods", Statistica Sinica, 21,5-42.

[59] Varin, C., and P., Vidoni (2005) : "A Note on Composite Likelihood Inference and Model Selection", Biometrika, 92, 519-528.

[60] Vecchia, A. (1988) : "Estimation and Model Identification for Continuous Spatial Processes", JRSS B, 50, 297-312.

[61] Wald, A. (1949) : " Note on the Consistency of the Maximum Likelihood Estimate", Annals of Mathematical Statistics, 20, 595-601.

[62] Wald, A. (1950) : "Statistical Decision Functions", Wiley.

[63] Wedderburn, R. (1974) : "Quasi Likelihood Functions, Generalized Linear Models and the Gauss-Newton Method", Biometrika, 61, 439-447.

[64] White, H. (1982) : "Maximum Likelihood Estimation of Misspecified Models", Econometrica, 50, 1-26.

[65] Wolfowitz, J. (1954) : "Estimation by the Minimum Distance Method in Nonparametric Difference Equations", Annals of Mathematical Statistics, 25, 203-217.

[66] Wooldridge J. (1994): "Estimation and Inference for Dependent Processes"? in Handbook of Econometrics, Vol 4, Chap 45, eds R., Engle and D., McFadden, North-Holland, 2641-2738.

Appendix 1

Proof of Theorem 1

If the function $M(.)$ is identically $(-\infty)$, then $\Theta_0 = \Theta$ and there is nothing to prove. Hence, for $\theta_0 \in \Theta_0$, we can assume $M(\theta_0) > -\infty$, and then (2.3) implies that:

$$E_0\left[|m(X,\theta_0)|\right] < \infty$$

Fix some $\theta \in \Theta$ and let $U_l, l = 1, 2, ..$ be a decreasing sequence of open balls around θ of radius converging to zero. Let us write $m_U(x)$ for $\sup_{\theta\in U} m(x,\theta)$. The sequence of functions $m_{U_l}(.)$ is decreasing and lower bounded by $m(.,\theta)$. Therefore, since by virtue of (2.3) the function $m_{U_l}(.)$ must be integrable for l sufficiently large, we can conclude by the monotone convergence theorem that:

$$\lim_{l\to\infty} E_0\left[m_{U_l}(X)\right] = E_0\left[m(X,\theta)\right]. \tag{5.12}$$

Let us consider the compact set:

$$B_\varepsilon = \left\{\theta \in \Theta; d(\theta,\Theta_0) \geq \varepsilon\right\}.$$

Fix some $\theta \in B_\varepsilon$ and let $U_l, l = 1, 2, ..$ be a decreasing sequence of open balls around θ of radius converging to zero. Since by definition:

$$\theta \in B_\varepsilon \Longrightarrow E_0\left[m(X,\theta)\right] < M(\theta_0)$$

we deduce from (5.12) that for l sufficiently large:

$$E_0\left[m_{U_l}(X)\right] < M(\theta_0)$$

Therefore, for each $\theta \in B_\varepsilon$, we are able to find an open ball V_θ around θ such that:

$$E_0\left[m_{V_\theta}(X)\right] < M(\theta_0).$$

Since the compact set B_ε is covered by the open balls $V_\theta, \theta \in B$, it is also covered by a finite subfamily $V_{\theta_j}, j = 1, ..., J$. Then by the Law of Large Numbers:

$$\sup_{\theta \in B_\varepsilon} M_n(\theta) \leq \sup_{j=1,..,J} \frac{1}{n} \sum_{i=1}^{n} m_{V_{\theta_j}}(X_i) \longrightarrow_{as} \sup_{j=1,..,J} E\left[m_{V_{\theta_j}}(X)\right] < M(\theta_0). \tag{5.13}$$

However, by definition of $\hat{\theta}_n$:

$$\hat{\theta}_n \in B_\varepsilon \Longrightarrow M_n(\theta_0) - o_P(1) \leq M_n\left(\hat{\theta}_n\right) \leq \sup_{\theta \in B_\varepsilon} M_n(\theta). \tag{5.14}$$

Comparing (5.13) and (5.14) , we conclude that:

$$\lim_{n \to \infty} \Pr\left[\hat{\theta}_n \in B_\varepsilon\right] = 0,$$

which is the announced result.
QED

Appendix 2

Asymptotic Normality

Even if this survey focuses on the consistency of PML estimators, it is important to mention its speed of convergence and the form of its asymptotic distribution. Under ergodicity conditions for (y_t, x_t), we have [see White (1982)] :

$$\sqrt{T}(\hat{\theta}_T - \theta_0^*) \simeq N[0, J^{-1} I J^{-1}],$$

where :

$$\begin{aligned} J &= E_0 \left(\frac{-\partial^2 \log g(y_t|x_t;\theta_0^*)}{\partial\theta\partial\theta'} \right), \\ I &= \sum_{h=-\infty}^{+\infty} Cov_0 \left(\frac{\partial \log g(y_t|x_t;\theta_0^*)}{\partial\theta} \frac{\partial \log g(y_{t-h}|x_{t-h};\theta_0^*)}{\partial\theta'} \right). \end{aligned}$$

When the pseudo-distribution is well-specified, the two matrices I and J are equal and define the average Fisher information matrix for one observation. Otherwise, they generally differ and the asymptotic variance-covariance matrix of the PML estimator is given by $J^{-1}IJ^{-1}$ (neither by J^{-1}, nor by I^{-1}). This is the so-called "sandwich" formula [Godambe (1960)].

When the pseudo score is a martingale difference sequence, matrix I can be simplified into $I = V_0 \left(\frac{\partial \log g(y_t|x_t;\theta_0^*)}{\partial\theta} \right)$. This condition is often fulfilled in the applications.