INSTITUT NATIONAL DE LA STATISTIQUE ET DES ETUDES ECONOMIQUES
Série des Documents de Travail du CREST
(Centre de Recherche en Economie et Statistique)

n° 2003-12

Empirical Likelihood in Some Semiparametric Models

P. BERTAIL[1]

Les documents de travail ne reflètent pas la position de l'INSEE et n'engagent que leurs auteurs.

Working papers do not reflect the position of INSEE but only the views of the authors.

[1] CREST-Laboratoire de Statistique.

Empirical likelihood in some semiparametric models

by

Patrice Bertail

Crest, Laboratoire de Statistique

Abstract

This paper studies the properties of empirical likelihood for Hadamard differentiable functionals tangentially to a well chosen set and gives some extensions in more general semiparametric models. We give a straightforward proof of its asymptotic validity and Bartlett correctability, essentially based on two ingredients convex duality and LAN properties of the empirical likelihood ratio in its dual form. Extensions to semiparametric problems with estimated infinite dimensional parameters are also considered. We give some applications to confidence intervals for the position parameter of a symmetric model and general functionals in biased sampling models.

Résumé

Dans cet article nous montrons comment les idées de vraisemblance empirique se généralisent à des fonctionnelles Hadamard différentiables tangentiellement à un ensemble de fonctions bien choisies et donnons des extensions à des modèles semiparamétriques plus généraux. Des arguments de dualité convexe et les propriétés localement asymptotiquement normales du rapport de vraisemblance empirique sous sa forme duale permettent de montrer la validité asymptotique de la méthode et la validité des corrections de type Bartlett. Nous proposons des extensions à des modèles semiparamétriques avec paramètres de nuisance estimés. Ces résultats sont illustrés par plusieurs exemples dont l'estimation d'un paramètre de position dans un modèle symétrique et l'estimation de paramètres fonctionnels dans des modèles avec biais de sélection.

Key Words and Phrases: Empirical likelihood, Hadamard differentiability, Empirical process, Donsker class, Semiparametric models, Bartlett correction, Bias sampling models.

AMS 1991 Subject Classifications. Primary: 62E20; **Secondary:** 62D05,G2N01.

1 Introduction

The likelihood principle has been one of the major tools in parametric statistics. Owen (1988, 1990, 2001) introduced an "empirical likelihood" ratio and obtained in a non-parametric setting a generalization of Wilk's (1936) theorem, stating that twice the log-likelihood ratio is asymptotically distributed as a χ^2 distribution. The idea of empirical likelihood goes back to Thomas and Grunkemeier (1975), but also in some sense to Hartley and Rao (1968) in the context of survey sampling, where it is known as model based likelihood. It is closely connected to the notion of non-parametric maximum likelihood introduced by Kiefer and Wolfowitz (1956). Nice accounts of empirical likelihood may be found in Hall and La Scala (1990) and Owen (2001)'s recent book, where one may find a huge bibliography.

For independent, identically distributed (i.i.d.) data, empirical likelihood ratio allows to build confidence regions for smooth parameters, mainly Fréchet differentiable parameters with respect to the Kolmogorov metric, including M-robust estimates (see Owen (1988)). A more precise description of the method is recalled in section 2 in the general framework of Hadamard differentiable functionals. We give a short proof of the validity and Bartlett correctability of empirical likelihood, extending results of Qin and Lawless (1994) (and removing third order moment conditions). It relies on the existence of a convex dual representation of the empirical likelihood, which may itself be seen as the log-likelihood ratio associated to a least favorable parametric family passing through the true model. This representation is closely connected to the important notion of dual likelihood introduced by Mykland (1995). It immediately leads to a Wilk's type theorem and the validity of the Bartlett correctability of empirical likelihood ratio, provided that this family satisfies the LAN property, which may be checked for instance by showing that it is quadratically differentiable, see Le Cam (1986). We also show that Hadamard differentiability (according to a well chosen set of functions) is sufficient to validate the use of empirical likelihood of general functionals, extending some results of Owen (1988) (2001). Part 3 discusses extensions to a more general semiparametric framework with nuisance parameters of infinite dimension. Our approach (based on derivatives of functionals) is different from the one considered in Murphy and van der Vaart (1997), in which the semiparametric likelihood incorporates the knowledge contained in the likelihood of the model. We give a general formulation of empirical likelihood in our framework. The idea is essentially based on using an estimated version of the efficient influence function which serves as the basis for the empirical likelihood procedure : we prove its validity under weak assumptions. In Part 4, we give some examples and applications to semiparametric models including confidence intervals for the position parameter of a symmetric distribution, a problem discussed in chapter 10 of Owen (2001) and reexamine the results of Qin (1993) in biased sampling models under weaker conditions. We do not discuss here the algorithmic problems, which may appear practically and refer to Owen (2001), chapter 12 for some propositions. The technical proofs of the lemmas and theorems are deferred to section 5.

2 Empirical likelihood for Hadamard differentiable functionals.

2.1 Empirical likelihood for a functional parameter

Let $X_1, ..., X_n, ...$ be i.i.d. random variables defined on a probability space (Ω, A, P_Ω) with common probability measure P belonging to a convex set $\wp$ of signed measures (containing the Dirac measure). Denote $(x_1, ..., x_n)$ a realization of $(X_1, ..., X_n)$ taking their value in $\mathcal{X}^n$. In the following, we are interested in constructing a confidence region for the functional parameter $\theta = T(P)$ (see Von Mises (1936)) defined on $\wp$, taking its value in $\mathbb{R}^q$. The empirical probability measure defined by

$$P_n = n^{-1} \sum_{i=1}^{n} \delta_{X_i}$$

is known to be the non-parametric maximum likelihood estimator (NPMLE) of P (see Gill (1989), Owen (1988), (1990)) The non-parametric maximum likelihood estimator (NPMLE) of $\theta = T(P)$ is then its empirical counterpart

$$\widehat{\theta}_n = T(P_n).$$

Many statisticians since von Mises have been interested in deriving the asymptotic properties of $\widehat{\theta}_n$ using differentiability assumptions on T (see Gill (1989)) via Taylor expansion (the delta method). Under some regularity assumptions, it is then possible to build confidence intervals or regions for the parameter θ. The approach of Owen (1988) is dual to this approach: the idea is to profile an "empirical likelihood" supported by the data so as to built directly a confidence region without relying on previous estimations.

The empirical likelihood ratio evaluated at θ is defined by

$$R_{E,n}(\theta) = \sup_{Q_n \in \mathcal{P}_n} \left\{ \Pi_{i=1}^{n} \frac{dQ_n}{dP_n}(X_i),\ T(Q_n) = \theta \right\}$$

where $\mathcal{P}_n$ is the set of discrete probability measures dominated by P_n that is

$$\mathcal{P}_n = \{ \widetilde{P}_n = \sum_{i=1}^{n} p_{i,n} \delta_{X_i}\ ,\ p_{i,n} \geq 0,\ \sum_{i=1}^{n} p_{i,n} = 1 \}.$$

Actually it should be noticed that for certain values of θ, $R_{E,n}(\theta)$ may not have any solution (consider for instance $\theta = E_P X$, then for values of θ outside the convex hull of the X_i there is no solution to the maximization problem). In that case we arbitrarily put $R_{E,n}(\theta) = 0$. This does not have any consequence in the construction because we will essentially be interested in the value of θ for which $R_{E,n}(\theta) > 0$. The empirical log-likelihood ratio is thus

$$\log(R_{E,n}(\theta)) = \sup_{p_{i,n}\ , i=1,...n} \left\{ \sum_{i=1}^{n} \log(p_{i,n}/(1/n)),\ T(\sum_{i=1}^{n} p_{i,n} \delta_{X_i}) = \theta,\ p_{i,n} \geq 0,\ \sum_{i=1}^{n} p_{i,n} = 1 \right\}$$

A better way to see this problem from a probabilistic point of view is to consider $r_n(\theta)$ as the minimization of the Kullback distance that we define here as

$$K(Q,P)=\begin{cases} -\int \log(\frac{dQ}{dP})dP \ , \ if\ Q << P \\ \\ \infty \ , \ else \end{cases}$$

between Q and P_n, over all the probability Q dominated by the empirical distribution P_n, with Q satisfying the constraint $T(Q)=\theta$, that is

$$(1) \qquad -\log(R_{E,n}(\theta))/n = \inf_{Q_n\in\mathcal{P}_n}(K(Q_n,P_n),\ T(Q_n)=\theta,\ \int dQ_n=1)$$

This thus may be seen as the empirical minimization of a particular distance to solve the inverse problem $T(Q)=\theta$. Other distances, which are all particular cases of convex distance or I-divergence (see Liese and Vajda (1986)) have been suggested in place of the Kullback distance. This has given rise to what is called in econometrics, "maximum entropy econometrics" (see for instance Golan, Judge and Miller (1996)). Most of the (first order) asymptotic that we discuss here may be obtained in the much more general framework of I-divergence for which a dual representation holds. We will however focus here on the particular case of Kullback distance and empirical likelihood because of its interesting third order properties.

Owen (1990) showed that if $T(P)=E_PX$ is the mean of a q-variate random variable with a covariance matrix $\Sigma=Var(X)$ of rank q then $-\log(R_{E,n}(\theta))$ converges in distribution to a $\chi^2(q)$ distribution, a result which is Wilk's (1936) non-parametric analog.

This yields a confidence region asymptotically of level $1-\alpha$

$$(2) \qquad \Re_{n,1-\alpha}=\left\{\theta,\ \Lambda_n(\theta)=-2\log(R_{E,n}(\theta))\leq\chi^2_{1-\alpha}(q)\right\}$$

It is easy to show by reciprocal inclusion that **in the case of a linear functional** that

$$(3) \qquad \Re_{n,1-\alpha}=\{T(\widetilde{P}_n),\ \widetilde{P}_n\in\overline{\mathcal{P}_{n,1-\alpha}}\}$$

with

$$\begin{aligned}\overline{\mathcal{P}_{n,1-\alpha}} &= \{\sum_{i=1}^n p_{i,n}\delta_{X_i}\ ,\ \sum_{i=1}^n p_{i,n}=1, p_{i,n}\geq 0, -2\sum_{i=1}^n \log(p_{i,n}/(1/n))\leq\chi^2_{1-\alpha}(q)\} \\ &= \{Q\in\mathcal{P},\ K(Q,P_n)\leq\frac{\chi^2_{1-\alpha}(q)}{2n},\ \int dQ=1,\ Q\geq 0\}\subset\mathcal{P}_n\end{aligned}$$

This equality which plays an important role in our analysis fails for non linear statistics for which we simply have $\{T(\widetilde{P}_n),\ \widetilde{P}_n\in\overline{\mathcal{P}_{n,1-\alpha}}\}\subset\Re_{n,1-\alpha}$. Notice that for any fixed

n, the set $\overline{\mathcal{P}_{n,1-\alpha}}$ contains P_n for any fixed value of $0 < \alpha \leq 1$. One purpose of this paper is to show that asymptotically for Hadamard differentiable functionals $\{T(\widetilde{P}_n),\ \widetilde{P}_n \in \overline{\mathcal{P}_{n,1-\alpha}}\}$ is still an asymptotically valid $1-\alpha$ confidence region for $T(P) \in \mathbb{R}^q$

Owen(1988) showed that in the case of the mean

$$P(\theta \in \Re_{n,1-\alpha}) = 1 - \alpha + O(n^{-1}). \tag{4}$$

This is actually the error rate for a two-sided confidence interval based on the normal asymptotic distribution in regular cases for instance smooth functions of the means. DiCiccio, Hall and Romano (1989, 1990) (see also DiCiccio and Romano (1989) and Hall (1990)) proved that empirical likelihood ratios (like in the parametric case) are Bartlett correctable. The Bartlett correction aims at fixing the expectation of $\Lambda_n(\theta) = -2\log\left(R_{E,n}(\theta)\right)$ exactly to q, the expectation of the limiting distribution. Because the first term in the Edgeworth expansion of $\Lambda_n(\theta)$ is of order n^{-1} multiplied by a polynomial of degree 1, a simple and explicit correction of the form $q\Lambda_n(\theta)/E_P\Lambda_n(\theta)$ allows to obtain a confidence region with a coverage error of size $O(n^{-2})$ (see Bickel and Ghosh (1990)). Notice that this is also the rate that can be obtained with the bootstrap in the case of two sided confidence intervals in smooth cases. In practice $E_P\Lambda_n(\theta)$ is generally unknown but may be replaced by a suitable estimator : if the estimator is chosen adequately the accuracy up to $O(n^{-2})$ still holds (Barndorff-Nielsen and Hall (1988)). Thus empirical likelihood does not require intensive computations in opposition to the Bootstrap distribution, which, in most cases, needs to be approximated by Monte-Carlo simulations. It should be noted that a "corrected version" of the weighted bootstrap has been proposed in Barbe and Bertail(1995) to improve over the usual bootstrap. Adequate choice of the weights, depending on the data (which may be seen as an attempt to invert the Edgeworth expansion of the bootstrap distribution thanks to the weights) typically leads to an accuracy of $O(n^{-5/2})$, for symmetric statistics, under regularity assumptions on the functional of interest (see also Guillou(1999)). However this requires strong knowledge on the functional of interest (the gradients up to order 6!) whereas as we will see, empirical likelihood may lead to an accuracy of order $O(n^{-2})$ in a quite automatic way.

Computational problems may however arise in the algorithms used to built the empirical likelihood regions if the parameter is very complicated : see Owen (2001) for tricks and algorithms to improve the computational aspects. In the case of smooth functions of a (possibly vector) mean, the confidence region is convex and the problem is to find the boundary of the confidence region. This may be done by solving a system of simultaneous equations and is achieved practically, for instance, via standard multivariate Newton algorithms. These results have been generalized by Hall and La Scala (1990) for smooth functions of multivariate mean.

The case of a Fréchet differentiable functional with respect to the Kolmogorov metric has been studied by Owen (1988, 1990) and of M-estimators by Qin and Lawless (1994). These results may be generalized to more general functionals, Fréchet differentiable with respect to an adequate metric which shares the same properties as the Kolmogorov metric, for instance a metric indexed by a class of functions (see Dudley (1993), Barbe and Bertail (1995)). In the following paragraph, we show that Hadamard differentiability is sufficient to obtain such generalizations.

2.2 Asymptotic validity of empirical likelihood for Hadamard differentiable functionals

We will establish our results for Hadamard differentiable functionals with an explicit canonical gradient (see for instance Pfanzagl (1981)). Hadamard differentiability is a notion of differentiability in which the remainder is controlled over compact neighborhoods. It is well suited for studying functionals of asymptotically tight random sequences (see Gill (1989)). Moreover Hadamard differentiability is the weakest form of differentiability for which the chain rule holds and preserves asymptotic efficiency which makes it a privileged tool in semiparametric analysis (see van der Vaart (1998)). The main problem to apply this notion in statistical applications lies in the choice of the metric or the topology which ensures both the convergence of the empirical process and the Hadamard differentiability.

For sake of generality, we will consider the following abstract empirical process framework. Assume that the functional T is defined on $\mathcal{P}$ considered as a subset of $\mathcal{L}_\infty(\mathcal{F})$. $\mathcal{F}$ is a subset of functions of a normed space of function here $L^2(P) = \{h,\ E_P h^2 < \infty\}$ endowed with $||f||_{2,P} = (E_P(f)^2)^{1/2}.\mathcal{L}_\infty(\mathcal{F})$ is equipped with the uniform convergence norm (or equivalently Zolotarev metric)

$$||P-Q||_{\mathcal{F}} = d_{\mathcal{F}}(P,Q) = \sup_{h\in\mathcal{F}} |\int hdP - \int hQ|$$

To avoid any measurability problem, we assume that expectations (resp. probability) are outer expectations (resp. outer probability) so that weak convergence is interpreted as Hoffman-Jørgensen convergence (see Van der Vaart and Wellner (1996) for details). For the same reason, we will also assume that $\mathcal{F}$ is image admissible Suslin. This ensures that the classes of the square functions and difference of square functions are P-measurable (see Dudley(1984)). In the following, it is assumed that $\mathcal{F}$ is a Donsker Class of functions with envelop H satisfying

$$0 < \int H^2 dP < \infty \tag{5}$$

so that the $\mathcal{F}$ indexed empirical process $n^{1/2}(P_n - P)$ converges (as an element of $\mathcal{L}_\infty(\mathcal{F})$) to a limit G_P, a tight Borel measurable element of $\mathcal{L}_\infty(\mathcal{F})$.with uniformly $||\ ||_{2,P}$ continuous sample paths $f \to G_P(f)$. Extensive references and results on empirical processes indexed by class of functions and conditions on $\mathcal{F}$ to be Donsker may be found in Van der Vaart and Wellner (1996).

More precisely, denote the covering number (the minimal number of ball of radius ε for the seminorm $||.||$ needed to cover $\mathcal{F}$) by $\mathcal{N}(\varepsilon, \mathcal{F}, ||.||)$. We will assume the following usual uniform entropy condition

$$\int_0^\infty \sup_{Q \in \mathcal{D}} \sqrt{\log(N(\varepsilon ||H||_{2,Q}, \mathcal{F}, ||.||_{2,Q})} d\varepsilon < \infty \tag{6}$$

where $\mathcal{D}$ is the set of all discrete probability measures Q with $\infty > \int H^2 dQ > 0$. Notice that if H is an envelop of the class then $H+1$ is also an envelop so that we may assume without loss of generality that $H \geq 1$.

The following lemma shows that the set $\overline{\mathcal{P}_{n,1-\alpha}}$ is small and contained in a band around P_n. It implies that the associated weighted empirical process indexed by $\mathcal{F}$ correctly standardized is asymptotically converging in $\mathcal{L}_\infty(\mathcal{F})$ uniformly over $\overline{\mathcal{P}_{n,1-\alpha}}$.

Lemma 2.1 . *For any $\alpha \in [0,1[$, there exists non negative constants $a(\alpha) < 1 < b(\alpha)$ such that for any $\widetilde{P}_n = \sum_{i=1}^n p_{i,n} \delta_{X_i}$ in $\overline{\mathcal{P}_{n,1-\alpha}}$ we have*

$$\frac{a(\alpha)}{n} \leq p_{i,n} \leq \frac{b(\alpha)}{n}$$

where $b(\alpha) \to 1$ when $\alpha \to 1$ (and $b(\alpha) \to \infty$ when $\alpha \to 0$).

For any fixed $\alpha \in]0,1[$, if $\mathcal{F}$ is a (Suslin) Donsker class of functions satisfying (5) and (6) then,

$$\left(\sum_{i=1}^n p_{i,n}^2 \right)^{-1/2} (\widetilde{P}_n - P) \underset{n \to \infty}{\overset{w}{\to}} G_P \text{ in } \mathcal{L}_\infty(\mathcal{F})$$

uniformly over $\overline{\mathcal{P}_{n,1-\alpha}}$, where G_P is a gaussian process in $\mathcal{L}_\infty(\mathcal{F})$.

Define now $B(\mathcal{F}, P)$, the subset of $\mathcal{L}_\infty(\mathcal{F})$ (seen as application (or path) $f \to \mu f = \int f d\mu$ from $\mathcal{F} \to \mathbb{R}$) which are $// //_{2,P}-$uniformly continuous and bounded (which is the smallest natural space in which G_P lies). We recall the following definition of Hadamard differentiability tangentially to $B(\mathcal{F})$ which is adapted from Pons and Turckeim (1991). Notice that the fact that differentiation is taken tangentially to $B(\mathcal{F}, P)$ (and not to $\mathcal{L}_\infty(\mathcal{F})$ which is too large) weakens the notion of differentiation and makes it easier to check in statistical problems (see examples in Gill (1989), Pons and Turckheim (1991), Van der Vaart (1998), Chap. 20.3, Van der Vaart and Wellner (1996), chap. 3.9).

Definition 2.1 *The functional T from $\mathcal{P} \subset \mathcal{L}_\infty(\mathcal{F})$ to $\mathbb{R}^q$ (or any Banach space $(\mathcal{B}_1, ||.||_{\mathcal{B}_1})$ is said to be Hadamard (or Compact) differentiable at $P \in \mathcal{P}$ tangentially* *to $B(\mathcal{F},P)$, say T is $HDT_{\mathcal{F}}-P$, iff there exists a continuous linear mapping dT_P (defined on $\mathcal{P}$), such that for every sequence $h_n \to h \in B(\mathcal{F},P)$, for every sequence $t_n \to 0$ such that $P+t_nh \in \mathcal{P}$,*

$$\frac{T((P+t_nh_n))-T(P)}{t_n}-dT_P.h \to 0 \quad \textit{as } t_n \to 0.$$

For a Hadamard differentiable functional, we call canonical (or first) gradient $T^{(1)}(.,P)$ any function from $\mathcal{X}$ to $\mathcal{B}_1$ such that

$$dT_P(Q-P)=\int T^{(1)}(x,P)(Q-P)(dx)$$

with the normalization

$$E_PT^{(1)}(X,P)=0$$

In the robustness terminology $T^{(1)}(x,P)$ is the influence function of the parameter $T(P)$ and is defined by $lim_{t\to 0}\left(\frac{T((1-t)P+t\delta_x)-T(P)}{t}\right)$ (see Hampel (1974)). Notice that, in a semiparametric framework, in which the parameter is defined implicitly by the model, the canonical gradient may not be unique. In the following we will assume that such a gradient exists and is non degenerated that is the covariance operator associated to $T^{(1)}(X,P)$ has full rank.

Assume that T is Hadamard differentiable ($HDT_{\mathcal{F}}-P$) with canonical gradient $T^{(1)}(.,P)$ then we have

$$T(\widetilde{P}_n)-\theta=\int T^{(1)}(x,P)(\widetilde{P}_n-P)(dx)+R_n(\widetilde{P}_n,P).$$

The preceding lemma implies that the solutions $\widetilde{P}_n$ in $\overline{\mathcal{P}_{n,1-\alpha}}$ are close to P typically up to $O_P(n^{-1/2})$ in $\mathcal{L}_\infty(\mathcal{F})$. Thus we expect the Delta method for Hadamard differentiable functionals to yield $R_n(\widetilde{P}_n,P)=o_P(\left(\sum p_{i,n}^2\right)^{1/2})=o_P(n^{-1/2})$ uniformly over all the admissible $\widetilde{P}_n$ in $\overline{\mathcal{P}_{n,1-\alpha}}$. These arguments suggest that the empirical likelihood ratio may be replaced by a linearized version

(7)
$$\begin{aligned}&\overline{R}_{E,n}^{L}(P)\\ &=\sup_{\widetilde{P}_n\in\mathcal{P}_n}\left\{\Pi_{i=1}^n nd\widetilde{P}_n(X_i),\ E_{\widetilde{P}_n}T^{(1)}(X,P)=0\right\}\\ &=\sup_{p_{i,n}\ i=1,\dots,n}\left\{\Pi_{i=1}^n np_{i,n},\ \sum_{i=1}^n p_{i,n}T^{(1)}(X_i,P)=0,\ p_{i,n}\geq 0,\ \sum_{i=1}^n p_{i,n}=1\right\}\end{aligned}$$

yielding an asymptotically (but intractable) asymptotic confidence region for $T(P)$ of the form

$$\Re^{\xi}_{n,1-\alpha} = \{T(P) + \int T^{(1)}(.,P)d\widetilde{P}_n,\ \widetilde{P}_n \in \overline{\mathcal{P}_{n,1-\alpha}}\}$$

But the analog of (3),

$$\Re^{T}_{n,1-\alpha} = \{T(\widetilde{P}_n),\ \widetilde{P}_n \in \overline{\mathcal{P}_{n,1-\alpha}}\}$$

is "close" up to $o_P((\sum p_{i,n}^2)^{-1/2}) = o_P(n^{-1/2})$ to the linearized confidence region which may be deduced from (7) so that the use of $\Re^{T}_{n,1-\alpha}$ is asymptotically justified.

The following theorem states that such approximations are asymptotically valid and establishes the validity of empirical likelihood for Hadamard differentiable functionals.

Theorem 2.1 *Assume that P is dominated by a measure μ. Assume that there exists a (Suslin) Donsker class of function $\mathcal{F}$ with envelop H, satisfying (6) such that T defined on $\mathcal{P}$ is $HDT_{\mathcal{F}} - P$ with gradient $T^{(1)}(.,P)$. If $Var(T^{(1)}(X,P)) < \infty$ is of rank q, then we have*

$$-2\log\left(\overline{R}^{L}_{E,n}(P)\right) \underset{n\to\infty}{\to} \chi^2(q)$$

and

$$P(\theta \in \Re^{\xi}_{n,1-\alpha}) = 1 - \alpha + O(n^{-1})$$

which implies

$$P(T(\theta) \in \Re^{T}_{n,1-\alpha}) \underset{n\to\infty}{\to} 1 - \alpha$$

If in addition $\overline{R}^{L}_{E,n}(P) \equiv \overline{R}^{L}_{E,n}(\theta)$ only depends on θ through $T^{(1)}(x,P) \equiv T^{(1)}(x,\theta)$, assume that the following Cramer condition holds

(8) $$\overline{\lim_{t\to\infty}} |E\exp(itT^{(1)}(X_i,P))| < \infty$$

and that

(9) $$E||T^{(1)}(X_i,P)||^s < \infty \textit{ for } s \geq 8 + \varepsilon,\ \varepsilon > 0,$$

then the Bartlett corrected confidence region

$$\Re^{B}_{1-\alpha} = \left\{\theta,\ q\log\left(\overline{R}^{L}_{E,n}(\theta)\right)/E(\log\left(\overline{R}^{L}_{E,n}(\theta)\right)) \leq \chi^2_{1-\alpha}(q)\right\}$$

is a third order correct confidence region for $T(P)$ that is

$$P(\theta \in \Re^B_{1-\alpha}) = 1 - \alpha + O(n^{-2})$$

Proof: For a better understanding of these results which are quite straightforward, we do not defer the proof to the appendix and give here a commented proof of the result.

Recall that using standard variational calculus (see Owen(2001)), the solution of the maximization problem (7) is given by

$$p_{i,n}(\lambda) = \frac{1}{n(1 + \lambda' T^{(1)}(X_i, P))} > 0$$

where λ, the Kuhn and Tucker coefficient, satisfies

$$\sum_{i=1}^n p_{i,n}(\lambda) = 1 \ , \ 0 \le p_{i,n}(\lambda) \le 1 \text{ and } \sum_{i=1}^n p_{i,n}(\lambda) T^{(1)}(X_i, P) = 0$$

By standard Kuhn and Tucker duality theory, we have

$$(10) \qquad -2\log\left(\overline{R}^L_{E,n}(P)\right) = 2 \sup_{\lambda \in R^q} \sum_{i=1}^n \log(1 + \lambda' T^{(1)}(X_i, P)) := 2 \sup_{\lambda \in R^q} L_n(\lambda)$$

We may see L_n as the log-likelihood ratio of a worst parametric family of distribution parameterized by λ, which passes through the true model at $\lambda = 0$. Indeed since $E_P T^{(1)}(X_i, P) = 0$,

$$p_\lambda(.) = \frac{dP}{d\mu}(.)(1 + \lambda' T^{(1)}(., P))\mathbb{I}_{\{1+\lambda' T^{(1)}(.,P)>0\}}$$

is a density defined for any λ (notice that we may also choose $\mu = P$). The log likelihood ratio in this parametric family at 0 is exactly $L_n(\lambda)$. In some sense empirical likelihood generates a least favorable model (see Bickel and al. (1993)) indexed by the Kuhn and Tucker parameters.

This interpretation of empirical likelihood ratio as the likelihood ratio associated to a least favorable family will be particularly useful in semiparametric models. Since $L_n(0) = 0$, $L_n(\lambda)$ may also be seen exactly as a dual log-likelihood in the sense of Mykland (1995) that is, in his terminology, a log-likelihood such that

$$\left[\frac{\partial L_n(\lambda)}{\partial \lambda}\right]_{\lambda=0} = \sum_{i=1}^n T^{(1)}(X_i, P)$$

$L_n(\lambda)$ is well defined, strictly concave thus admitting an unique maximum. Moreover by concavity of the log,

$$E_P(\log(1 + \lambda' T^{(1)}(X_i, P)) \le \log(1 + \lambda' E_P T^{(1)}(X_i, P)) = 0$$

Thus $E_P(\log(1+\lambda' T^{(1)}(X_i,P))$ has an unique maximum at $\lambda=0$ and the m.l.e. converges to 0. Notice that since $Var(T^{(1)}(X,P))$ exists and is strictly positive, the family $\{p_\lambda,\ \lambda\in\mathbb{R}^q\}$ is differentiable in quadratic mean and thus the associate log-likelihood ratio is Locally Asymptotically Normal (LAN) (see Le Cam(1986)). Indeed, the differentiability in quadratic mean follows from lemma 7.6 of Van der Vaart (1998), p. 95. $p_\lambda(x)$ is continuously differentiable in λ everywhere except on the set $\{x,\ 1+\lambda' T^{(1)}(x,P)=0\}$. But it is easy to see that this set has probability 0 if $Var(T^{(1)}(X,P))>0$ (see also the direct proof of Owen(2001), lemma 11.1 p.217). Thus the empirical likelihood ratio is simply a likelihood ratio for testing $\lambda=0$ in the LAN model $\{p_\lambda,\ \lambda\in\mathbb{R}^q\}$ and it follows that

$$-2\log\left(\overline{R}^L_{E,n}(P)\right)\to\chi^2(q).$$

Because $L_n(\lambda)$ is itself a parametric log-likelihood ratio (as a function of vector parameter λ), it is Bartlett correctable under (8) and (9). These conditions are sufficient to ensure the validity of the Edgeworth expansion of the standardized version of $n^{-1}\sum T^{(1)}(X_i,P)$ up to order $O(n^{-2})$, which is needed for the Bartlett correction to hold. Thus if $\overline{R}^L_{E,n}(P)$ depends only on θ, Bartlett corrected empirical likelihood can be used to construct confidence region with improved accuracy.

Now define the linear parameter for $Q\in\mathcal{P}$,

$$\xi(Q)=\theta(P)+\int T^{(1)}(x,P)Q(dx)$$

then a $1-\alpha$ empirical likelihood based confidence region for this parameter is

$$\begin{aligned}\Re^{\xi}_{n,1-\alpha} &= \{\xi(\widetilde{P}_n),\ \text{with } \widetilde{P}_n\in\overline{\mathcal{P}_{n,1-\alpha}}\}\} \\ &= \left\{\xi(P),\ -2log(\overline{R}^L_{E,n}(P))\le\chi^2_{1-\alpha}(q)\right\} \\ &= \{\theta(P),\ -2log(\overline{R}^L_{E,n}(P))\le\chi^2_{1-\alpha}(q)\}\end{aligned}$$

with $P(\theta(P)\in\Re^{\xi}_{n,1-\alpha})=1-\alpha\ +O(n^{-1})$ (since the parameter $\xi(Q)$ is linear).

Now we have by Hadamard differentiability

$$\begin{aligned}T(\widetilde{P}_n) &= \xi(\widetilde{P}_n)+R_n(\widetilde{P}_n,P) \\ &= \theta(P)+\int T^{(1)}(x,P)(\widetilde{P}_n-P)(dx)+R_n(\widetilde{P}_n,P)\end{aligned}$$

The results now follow from similar arguments as in Th. 20.8 of Van der Vaart (1998). Take $t=\left(\sum p^2_{i,n}\right)^{-1/2}$ which is of order $o(n^{-1/2})$ uniformly over $\overline{\mathcal{P}_{n,1-\alpha}}$ by lemma 2.1 and $h_n=\left(\sum p^2_{i,n}\right)^{-1/2}(\widetilde{P}_n-P)\ \in\mathcal{L}_\infty(\mathcal{F})$ in the definition Hadamard differentiability then by definition $h=G_P\in B(\mathcal{F},P)$. We deduce that $R(\widetilde{P}_n,P)=\ o(n^{-1/2})$ uniformly over $\overline{\mathcal{P}_{n,1-\alpha}}$. It follows that $\Re^{\xi}_{n,1-\alpha}$ and $\Re^{T}_{n,1-\alpha}$ are asymptotically equivalent up to $o_P(n^{-1/2})$. □

Remark 2.1 : The proof essentially relies on a convex duality argument which allows to write the empirical likelihood as a true parametric likelihood ratio indexed by the Kuhn and Tucker coefficient. Duality is used in a constructive way in Mykland(1995) : here duality is actually a consequence of the fact that Kullback distance is a convex statistical distance (see Liese and Vajda(1986)). The duality principle generates a least favorable family which may be checked to be locally asymptotically normal, under $Var(T^{(1)}(X,P)) < \infty$. Hadamard differentiability is essentially needed to show that $\Re^{\xi}_{n,1-\alpha}$ and $\Re^{T}_{n,1-\alpha}$ are asymptotically equivalent. Such arguments may be used in a large number of applications to obtain the asymptotic distribution of the empirical log-likelihood ratio as well as its Bartlett correctability (see example 4 in the last section). It also may be used to prove (first order) asymptotic results when the Kullback distance in (1) is replaced by another convex statistical distance for instance the entropy or actually any convex statistical distance (I-divergence) for which a convex duality principle holds. In the case of empirical likelihood, Bartlett correctability follows from the fact that the dual function is itself a likelihood, which is not the case for more general convex statistical distance.

Remark 2.2 : Owen (1990), Qin and Lawless (1994) showed how that kind of results may be used for M estimates: indeed in that case the influence function depends only on θ and $R^{L}_{E,n}(\theta)$ may be quite easy to calculate. Notice that in a semiparametric model the choice of the influence function is left to the statistician. Of course if the efficient influence function (in the sense of Bickel, Klaassen, Ritov, Wellner (1993)) is known and independent of nuisance parameter (see the work of Amari and Kawanabe (1997) for the existence of general estimating equation) then this would be the best candidate for $T^{(1)}$. However many problems may appear :

-the efficient influence function is not always easy to obtain since most of the time it involves the projection into an infinite dimensional space

-it is not clear whether this expression may be used in practice for $T^{(1)}(.,P)$ may have a very complicated form and depend on some nuisance parameter. This kind of problems typically appears in the "challenges" exposed in chap. 10 of Owen (2001) We shall further examine these points in the next paragraph.

Remark 2.3 : The preceding arguments mainly rely on the existence of a dual form for the likelihood ratio and it is interesting to investigate and use the special structure of this dual representation. At 0, the information matrix with respect to λ is given by $V_P(T^{(1)}(X,P))$ and the m.l.e. for λ in this LAN family is such that

$$\widehat{\lambda}_n(P) = \left(\frac{1}{n}\sum T^{(1)}(X_i,P)T^{(1)}(X_i,P)'\right)^{-1}\sum_{i=1}^{n} T^{(1)}(X_i,P))(1+o_p(1)) = O_P(\frac{1}{\sqrt{n}})$$

and we have easily by the strong law of large number for the first term and the central limit theorem for the second term in this expression

$$\sqrt{n}\widehat{\lambda}_n(P) \underset{n\rightarrow\infty}{\rightarrow} N(0, V(T^{(1)}(X_i,P))^{-1})$$

It follows that at the m.l.e. $\widehat{\lambda}_n(P)$, $L_n(\widehat{\lambda}_n(P))$ that is the empirical likelihood ratio behaves asymptotically like the usual GMM (generalized method of moments) objective function

$$\left(\frac{1}{\sqrt{n}}\sum_{i=1}^{n}T^{(1)}(X_i,P)\right)^{\prime}\left(\frac{1}{n}\sum T^{(1)}(X_i,P)T^{(1)}(X_i,P)\right)^{-1}\left(\frac{1}{\sqrt{n}}\sum_{i=1}^{n}T^{(1)}(X_i,P)\right).$$

which may be seen as the square of the norm of an autonormalized sum (which is asymptotically $\chi^2(q)$). This auto-normalization, carried out internally by the optimization procedure as noticed by several authors is essentially due to the LAN structure of the dual likelihood ratio.

Remark 2.4 : Because $L_n(\lambda)$ is itself a likelihood (as a function of λ), a rather interesting property of this approximate linearized empirical likelihood is that even though we do not take into account the second order terms in the Taylor expansion of $T(P_n)$, it shares the same properties as the empirical likelihood of the mean that is, it is Bartlett correctable. This is a rather amazing fact which plays in favor of empirical likelihood against other third order correct methods for constructing confidence intervals such as iterative inversion of Edgeworth expansion or weighted bootstrap approximations. Indeed in these cases, the structure of the statistics (its Hoeffding decomposition in term of orthogonal U-statistics) plays a fundamental role for implementing these methods (see Barbe and Bertail (1995)). However it should be noticed that this is only possible when the influence function is simple and does not depend on additional nuisance parameter. If $T^{(1)}(.,P)$ does not depend only on θ, it is in general not possible to control the error induced by the linearization of $T(P)$, so that the Bartlett properties will not hold.

Remark 2.5 Hadamard differentiability seems to be the weakest form of differentiability which ensures that we may approximate the exact interval by its linearized form. Lemma 1 is the key point for showing the validity of the approximation. Of course other type of conditions may be used to obtain a similar result, for instance by using bracketing entropy.

3 Empirical likelihood in semiparametric models.

3.1 Semiparametric extensions

A model is said semiparametric if $\wp_{\Theta,H}=\{P_{\theta,G},\ \theta\in\Theta,\ G\in H\}$ is a set of probability measures indexed both by a parameter of interest in a set Θ of $\mathbb{R}^k$ and a nuisance parameter G in a space H possibly of infinite dimension. Such models and generalization to the infinite dimensional case for Θ are studied at length in Bickel, Klaassen, Ritov, Wellner (1993). One of the main problem which appears in semiparametric model is that generally the parameter of interest $\theta=T(P_{\theta,G})$ is defined on a smaller set $\wp_{\Theta,H}$ than the one considered before, so that $T(P_n)$ or the gradient $T^{(1)}(x,Q)$ at $Q=P_n$ may not be defined properly.

To illustrate and motivate a generalization of empirical likelihood in semiparametric models, we first begin by examining some examples. When dealing with confidence intervals for quantiles in the real valued case,

$$T_\alpha(P)=\inf(t,\ P(]-\infty,t])>\alpha)$$

with $\wp_{\Theta,H}=\{P,\ \text{with } P<<\mu \text{ and } f=dP/d\mu,\ f\in H\}$, where H is a restricted set of densities Chen and Hall (1993) have shown that the standard empirical likelihood method as presented before leads to confidence intervals with coverage error of size $O(n^{-1/2})$ (which is actually the error size of the asymptotic approximation). However a version based on the smoothed influence function of the fractiles leads to an error of size $O(n^{-1})$. Recall that the influence function of $T_\alpha(P)$ is lattice and given by

$$T^{(1)}(x,P)=\frac{\delta_{\{x<T_\alpha(P)\}}-\alpha}{f(T_\alpha(P))}.$$

Moreover an exact Bartlett correction leads to an error of size $O(n^{-2})$. This result shows that, when the functional involves the density f of the underlying probability, it is clearly preferable to smooth the gradient. A similar reasoning also holds when the parameter of interest is the density itself (see Chen (1996)).

This suggests the following generalization for functional parameters defined on a restricted set of probability. Consider a semiparametric model

$$\wp_{\Theta,H}=\{P_{\theta,G}\in\wp\ ,\ \eta\in I\!R^k,\ G\in H\}$$

where H is a infinite dimensional space. We are interested in estimating the functional

$$\theta=T(P_{\theta,G})$$

Then it is clear that there is no reason for $\widetilde{P}_n=\sum_{i=1}^n p_{i,n}\delta_{X_i}$ to belong to $\wp_{\Theta,H}$. The semiparametric approach generally used in such a context is to extend the functional $T(.)$ to a more general space. For this, one generally introduces a pseudo metric d on $\wp$ and defines a pseudo projection Π (not necessarily unique) into the model of any $P\in\wp$ to be

$$\Pi(P)=Arg\min_{Q\in\wp_{\Theta,H}}(d(P,Q))$$

Then the functional

$$\widetilde{T}(P)=T\circ\Pi(P)$$

extends T defined on $\wp_{\Theta,H}$ to $\wp$. $\widetilde{T}(P_n)$ defines a minimum distance estimator (see Bickel, Klaassen, Ritov, Wellner (1989)). More generally we may choose any function which extends the functional to $\wp$, the set of all signed measures. For instance, in the case of $\wp_{\Theta,G}$ being the set of probability measures with continuous density with respect to the Lebesgue measure λ, we may choose Π as being the convolution of P with a continuous kernel if P does not have a continuous density with respect to λ and $\Pi(P)=\frac{dP}{d\lambda}$ else. In that case because of the linearity of the convolution operator the influence function of $\widetilde{T}$ will be exactly the smoothed version of the influence function of T as considered in Chen and Hall (1993). If such extension exists then we may define the empirical likelihood ratio in the semiparametric model as

$$(11)\quad R_{E,n}(\theta)=\sup_{p_{i,n}\ ,\ i=1,\ldots,n}\left\{\Pi_{i=1}^{n}np_{i,n},\ \widetilde{T}(\sum_{i=1}^{n}p_{i,n}\delta_{X_i})=\theta,\ p_{i,n}\geq 0,\ \sum_{i=1}^{n}p_{i,n}=1\right\}$$

However, in many problems, the right "efficient" choice of Π (in the sense of Bickel, Klaasen, Ritov and Wellner(1993)) depends on the geometry of the problem. Ideally it should be chosen in such a way that the gradient at P of $\widetilde{T}$ coincides with the efficient influence function of T in the original semiparametric problem. In many problems, it may be however easier to work directly with the efficient influence function or a non efficient but tractable one. Let $\widetilde{T}^{(1)}(.,P_{\theta,G})$ be such quantity. The linearized version of the original problem is thus

$$(12)\qquad \sup_{p_{i,n}\ ,\ i=1,\ldots,n}\left\{\Pi_{i=1}^{n}np_{i,n},\ \sum_{i=1}^{n}p_{i,n}\widetilde{T}^{(1)}(X_i,P_{\theta,G})=0,\ p_{i,n}\geq 0,\ \sum_{i=1}^{n}p_{i,n}=1\right\}$$

Of course, since in practice G is unknown this also depends on the nuisance parameter G. However one may in many situations , for any fixed θ, find a smooth estimator $\widehat{G}_{\theta,n}$ of G. Assume that such a consistant estimator exists, then we may use as semiparametric empirical likelihood

$$(13)\qquad \widetilde{R}_{E,n}(\theta)=\sup\left\{\begin{array}{c}\Pi_{i=1}^{n}np_{i,n},\ \sum_{i=1}^{n}p_{i,n}\widetilde{T}^{(1)}(X_i,P_{\theta,\widehat{G}_{\theta,n}})=0,\\ p_{i,n}\geq 0,\ \sum_{i=1}^{n}p_{i,n}=1\end{array}\right\}.$$

It should be noticed that in general the solution of the original problem (11) and that of (13) are different but asymptotically equivalent (for instance if $\widetilde{T}$ Hadamard differentiable with a gradient continuous in P).

Another possible definition which would ease the technical difficulties that we will encounter later, when studying the asymptotic properties of this approximate empirical likelihood, would be to use the splitting trick frequently used in the semiparametric literature. For this define $G^{(1)}_{\theta,n/2}$ and $G^{(2)}_{\theta,n/2}$be the estimators of G, based respectively on the first half ($[n/2]$ first values) and second half of the sample. Then we may define the approximate semiparametric empirical likelihood by

(14)

$$\widetilde{\widetilde{R}}_{E,n}(\theta)=\sup\left\{\begin{array}{c}\Pi_{i=1}^{n}np_{i,n},\ p_{i,n}\geq 0,\ \ \sum_{i=1}^{n}p_{i,n}=1\\ \sum_{i=1}^{[n/2]}p_{i,n}\widetilde{T}^{(1)}(X_i,P_{\theta,G^{(2)}_{\theta,n/2}})+\sum_{i=[n/2]+1}^{n}p_{i,n}\widetilde{T}^{(1)}(X_i,P_{\theta,G^{(1)}_{\theta,n/2}})=0\end{array}\right\}.$$

However from a practical point of view the splitting trick is less than satisfactory (the loss in using only half of the sample for the estimation of the nuisance parameter, for instance a density may have disastrous effects on the semiparametric estimators for fixed n), so that we will not pursue this analysis. The results that we have obtained may be carried out in this case too using the same kind of arguments.

3.2 Asymptotic validity under weak assumptions

Consider the optimization program (11), similar arguments as in the preceding part yield the dual equality

$$\begin{aligned}&-2\log(\widetilde{R}_{E,n}(\theta))\\ =\ &2\sup_{\lambda\in\mathbb{R}^q}\left\{\sum_{i=1}^{n}\log\left(1+\lambda'\widetilde{T}^{(1)}(X_i,P_{\theta,\widehat{G}_{\theta,n}})\right)\right\}\equiv 2\sup_{\lambda\in\mathbb{R}^q}\widetilde{L}_n(\lambda).\end{aligned}$$

Notice however that $\widetilde{L}_n(\lambda)$ cannot be seen directly as a log-likelihood ratio because of the dependencies in $\widehat{G}_{\theta,n}$ and the absence of recentering. Indeed, there is no reason that $E_{P_{\theta,G}}\widetilde{T}^{(1)}(X_i,P_{\theta,\widehat{G}_{\theta,n}})=0$. However a result similar to Theorem 2.1 may be proved by combining martingale and empirical process arguments, provided that $\widehat{G}_{\theta,n}$ is chosen adequately.

To prove the validity of empirical likelihood in this framework, we will assume the following hypotheses :

H_1 : Assume that the sequence of estimators $\widehat{G}_{\theta,n}$ is a symmetric statistic of the observations $X_1,...X_n$ converging to G with probability.one.

The following condition is the usual one ensuring that the bias of the estimated influence function is small compared to the rate of convergence, that we expect. This implies that $l_{n,E}(\theta)$ is close to a sequence of likelihood ratio.

H_2 : The estimator $\widehat{G}_{\theta,n}$ is such that

$$E_{P_{\theta,G}}\widetilde{T}^{(1)}(X_i,P_{\theta,\widehat{G}_{\theta,n}})=o(n^{-1/2}).$$

H_3 : $\widetilde{T}^{(1)}(.,P_{\theta,G})$ is a continuous function of G (with respect to a metric metrizing convergence of $\widehat{G}_{\theta,n}$ to G).

The last condition implies a uniform control of the approximation of $\widetilde{T}^{(1)}(.,P_{\theta,G})$

by $\widetilde{T}^{(1)}(., P_{\theta,\widehat{G}_{\theta,n}})$.

H_4 : For every θ and n the functions $\widetilde{T}^{(1)}(., P_{\theta,\widehat{G}_{\theta,n}})$ belong to a Donsker-class of functions with probability one. The class has an envelop $H(.) > 0$ may be depending on θ with

$$E_{P_{\theta,G}} H(X)^2 < \infty.$$

Actually these conditions are weaker than the conditions that one usually assumes in the framework of semiparametric models (see for instance Bickel, Klaassen, Ritov and Wellner (1993) or Van der Vaart (1998), chap. 25, see his theorem 25.54). The main reason for this is that we just want to give here conditions for the asymptotic validity of the empirical likelihood principle. Moreover, the fact that we choose an estimator $\widehat{G}_{\theta,n}$ which is symmetric of the observations allows to weaken the usual hypotheses thanks to backward martingale arguments (see lemma 5.1). Nevertheless, if we want to obtain efficient estimators by minimizing the resulting asymptotically χ^2 statistics, additional assumptions (uniformity conditions in the neighborhood of the true value θ) as required by Van der Vaart (1998) seem to be needed.

Theorem 3.1 *Assume that $H_1 - H_4$ holds then, if $Var(\widetilde{T}^{(1)}(X, P_{\theta,G}))$ is of rank q,*

$$-2\log(\widetilde{R}_{E,n}(\theta)) \rightarrow \chi^2(q) \text{ as } n \rightarrow +\infty$$

yielding asymptotically correct confidence intervals of level $1-\alpha$ of the form

$$\{\theta, -2\log(\widetilde{R}_{E,n}(\theta)) \leq \chi^2_{1-\alpha}(q)\}$$

Remark 3.1 Another way to prove this result is to consider the sequence of approximate least favorable models (notice the recentering factor which ensures that we have a density)

$$\begin{aligned} p_{\lambda,n}(.) &= \frac{dP}{d\mu}(.)\left[1 + \lambda'\left(\widetilde{T}^{(1)}(., P_{\theta,\widehat{G}_{\theta,n}}) - E_{P_{\theta,G}}\widetilde{T}^{(1)}(X_1, P_{\theta,\widehat{G}_{\theta,n}})\right)\right] \\ &\quad \mathbb{I}_{\left\{1+\lambda'\left(\widetilde{T}^{(1)}(.,P_{\theta,\widehat{G}_{\theta,n}}) - E_{P_{\theta,G}}\widetilde{T}^{(1)}(X_1,P_{\theta,\widehat{G}_{\theta,n}})\right)>0\right\}} \end{aligned}$$

Even if it may be possible to check the quadratic differentiability, conditions which ensure that the maximum likelihood estimator of λ in this family has a good behavior in the presence of the estimated parameter $\widehat{G}_{\theta,n}$ may be less easy to check. However it is interesting to see that for the Bartlett correctability of the approximate empirical likelihood (13) to hold, the behavior of $E_{P_{\theta,G}}\widetilde{T}^{(1)}(X_1, P_{\theta,\widehat{G}_{\theta,n}})$ is of great importance. In many situations, for instance convex linear models (see Bickel, Klaassen, Ritov, Wellner (1993)) we have

$$E_{P_{\theta,G}}\widetilde{T}^{(1)}(X_i, P_{\theta,\widehat{G}_{\theta,n}}) = 0 \ , \tag{15}$$

so that it may be easier to show the Bartlett correctability (at least up to order $O(n^{-3/2})$) in that case (see Chen(1996), Chen and Hall (1993) for some examples).

Remark 3.2

Although the splitting trick used to construct (14) is not very satisfactory from a practical point of view, it may be used to weaken the hypotheses of the preceding theorem. Indeed in that case, we do not even have to assume that the class is Donsker (provided that we still have a square integrable envelop). If we assume instead

$$H_5 \;\; E_{P_{\theta,G}}||\widetilde{T}^{(1)}(X_1,P_{\theta,\widehat{G}^{(i)}_{\theta,n}})-E_{P_{\theta,G}}\widetilde{T}^{(1)}(X,P_{\theta,G})||^2\rightarrow 0 \text{ as } n\rightarrow\infty, \text{ for } i=1,2$$

then the result of theorem 3.1 still holds. Indeed, the Donsker property is only needed to show the uniformity (24) in the proof. To obtain a similar theorem for (14), we have to check that

$$n^{-1/2}\left(\sum_{i=1}^{n}\widetilde{T}^{(1)}(X_i,P_{\theta,G})-\sum_{i=1}^{[n/2]}\widetilde{T}^{(1)}(X_i,P_{\theta,\widehat{G}^{(2)}_{\theta,n}})-\sum_{i=[n/2]+1}^{n}\widetilde{T}^{(1)}(X_i,P_{\theta,\widehat{G}^{(1)}_{\theta,n}})\right)=o_P(1)$$

which is a consequence of H_5. Condition H_5 may be sometimes easier to check than the Donsker property.

4 Examples

Most of the examples of Hadamard differentiable functionals considered in Pons and Turckeim (1991), Van der Vaart (1998) enter the framework of part 2. There is nothing really new in detailing these examples : provided that we consider finite dimensional parameter (this includes constructing confidence intervals at several points for a density or a hazard rate with censored data), $\Re^{T}_{n,1-\alpha}=\{T(\widetilde{P}_n),\ \widetilde{P}_n\in\overline{\mathcal{P}_{n,1-\alpha}}\}$ is asymptotically a valid confidence region. We rather illustrate our results and remarks by some examples taken from the semiparametric literature. In Example 1 and 2, the efficient influence function is known (up to a unidimensional parameter in example 1). We show also how empirical likelihood internally calculates this value. Many problems quoted as challenging in Chapter 10 of Owen (2001) are actually semiparametric problems which may be treated as example 3. Finally we show in example 4 how extensions and Bartlett correctability may be obtained quite straightforwardly in the case of bias sampling models studied by Qin (1994).

Example 1 : *Third order correct confidence interval for a P constrained mean*

Consider the example 3 p. 68 of Bickel and al. (1993) in which one is interested in estimating the mean $\theta=E_PX\neq 0$ with $T^{(1)}(x,P):=T^{(1)}(x,\theta)=x-\theta$, on the set of probability with a fixed coefficient of variation $\{P$ such that $E_PX^8<\infty$ and $\gamma(P)=E_PX^2-(1+c_0)(E_PX)^2=0,\ c_0\neq 0\}$. Let $\gamma^{(1)}(.,P)$ be the influence function of $\gamma(.)$ at P. By a straightforward calculus, it is given by

$$\gamma^{(1)}(x,P) := \gamma^{(1)}(x,\theta) = x^2 - 2(1+c_0)\theta(x-\theta) - (1+c_0)\theta^2$$

The efficient influence function which is given by the projection on the nuisance tangent space $\{h \in L^2(P),\ E_P h = 0 \text{ and } E_P h\gamma^{(1)} = 0\}$ has the following expression (this is simply the residual of the regression of $T^{(1)}$ on $\gamma^{(1)}$, see Bickel and al. (1993) p.55)

$$\widetilde{T}^{(1)}(x,P) = T^{(1)}(x,P) - \frac{cov_P(T^{(1)}(X,P)\gamma^{(1)}(X,P))}{Var_P(\gamma^{(1)}(X,P))}\gamma^{(1)}(x,P) \tag{16}$$

and has variance

$$V_P\widetilde{T}^{(1)}(X,P) = V_P T^{(1)}(X,P) - \frac{cov_P(T^{(1)}(X,P)\gamma^{(1)}(X,P))^2}{Var_P(\gamma^{(1)}(X,P))}.$$

However the regression coefficient $\alpha = \frac{cov_P(T^{(1)}(X,P)\gamma^{(1)}(X,P))}{Var_P(\gamma^{(1)}(X,P))}$ is unknown and must be estimated for instance by

$$\widehat{\alpha}(\theta) = \frac{\frac{1}{n}\sum T^{(1)}(X_i,\theta)\gamma^{(1)}(X_i,\theta)}{\frac{1}{n}\sum \gamma^{(1)}(X_i,\theta)^2}$$

which is a symmetric function of the observations. However $\widetilde{T}^{(1)}(x,P)$ is clearly continuous in α (which plays here the role of the nuisance parameter G) and we have by lemma 2, $\widehat{\alpha}(\theta) \to \alpha(\theta)$ a.s.. We then may use the "estimated" estimating function

$$\sum_{i=1}^n T^{(1)}(X_i,\theta) - \widehat{\alpha}(\theta)\gamma^{(1)}(X_i,\theta) = 0$$

It is easy to check that

$$\begin{aligned} & E_P(T^{(1)}(X_i,\theta) - \widehat{\alpha}(\theta)\gamma^{(1)}(X_i,\theta)) \\ = \quad & -E_P\left\{\frac{\frac{1}{n}\sum_{j=1}^n T^{(1)}(X_j,\theta)\gamma^{(1)}(X_j,\theta)\gamma^{(1)}(X_i,\theta)}{\frac{1}{n}\sum_{j=1}^n \gamma^{(1)}(X_j,\theta)^2}\right\} = O(n^{-1}). \end{aligned}$$

Moreover, H_4 is satisfied with

$$H(x) = |T^{(1)}(x,\theta)| + 2V_P(T^{(1)}(X,P))^{1/2}V_P(\gamma^{(1)}(X,P))^{-1/2}|\gamma^{(1)}(x,\theta)|.$$

Moreover under $EX^8 < \infty$, we have

$$\begin{aligned} & E_P\left\{T^{(1)}(X_i,\theta) - \widehat{\alpha}(\theta)\gamma^{(1)}(X_i,\theta) - T^{(1)}(X_i,\theta) - \alpha(\theta)\gamma^{(1)}(X_i,\theta)\right\}^2 \\ \leq \quad & \left(E_P\gamma^{(1)}(X_i,\theta)^4\right)^{1/2}\left(E_P(\widehat{\alpha}(\theta) - \alpha(\theta))^4\right)^{1/2} \to 0 \; as \; n \to \infty \end{aligned}$$

so that H_5 is satisfied. We may then apply Theorem 3.1 to obtain an empirical likelihood based confidence intervals.

In this case, it is simpler to consider this semiparametric model as a problem in which there are two estimating functions corresponding respectively to $E_Q(T^{(1)}(X,P))=0$ and $E_Q\ \gamma^{(1)}(X,P)=0$. Notice that at P, these two estimating functions only depend on θ so that the results of Qin and Lawless (1994) apply in this framework. This result may be explained by the fact that the optimization problem internally computes (up to a constant) the efficient influence function. Indeed if one tries to solve directly the dual optimization problem

$$\sup_{\lambda,\mu} n^{-1}\sum_{i=1}^{n}\log\left(1+\lambda' T^{(1)}(X_i,P)+\mu'\gamma^{(1)}(X_i,P)\right)$$

straightforward calculus based on Taylor expansion (see remark 3) yields

$$\Omega\begin{pmatrix}\widehat{\lambda}\\ \\ \widehat{\mu}\end{pmatrix}=\begin{pmatrix}\frac{1}{n}\sum_{i=1}^{n}T^{(1)}(X_i,P)\\ \\ \frac{1}{n}\sum_{i=1}^{n}\gamma^{(1)}(X_i,P)\end{pmatrix}+o_P(1)$$

with

$$\Omega=\begin{pmatrix}V_P(T^{(1)}(X,P)) & Cov_P(T^{(1)}(X,P)\gamma^{(1)}(X,P))\\ \\ Cov_P(T^{(1)}(X,P)\gamma^{(1)}(X,P)) & V_P(\gamma^{(1)}(X,P))\end{pmatrix}$$

that is

$$\begin{pmatrix}\widehat{\lambda}\\ \\ \widehat{\mu}\end{pmatrix}=\begin{pmatrix}\frac{1}{n}\sum_{i=1}^{n}\widetilde{T}^{(1)}(X_i,P)/V_P(\widetilde{T}^{(1)}(X_i,P))\\ \\ \frac{1}{n}\sum_{i=1}^{n}\widetilde{\gamma}^{(1)}(X_i,P)/V_P(\widetilde{\gamma}^{(1)}(X_i,P))\end{pmatrix}$$

where $\widetilde{\gamma}^{(1)}$ is the residual of the regression of $\gamma^{(1)}$ on $T^{(1)}$

$$\widetilde{\gamma}^{(1)}(x,P)=\gamma^{(1)}(x,P)-\frac{cov_P(T^{(1)}(X,P)\gamma^{(1)}(X,P))}{Var_P(T^{(1)}(X,P))}T^{(1)}(x,P)$$

that is the efficient influence function when estimating $\gamma(P)$ with a known mean θ. Thus the m.l.e. of λ is exactly proportional to the efficient estimating function given by (16). As a by product, this suggests that we may use the solution of the estimated KT coefficient seen as function of the parameter θ, $\widehat{\lambda}=\widehat{\lambda}(\theta)$ to obtain an efficient estimator of θ by solving $\widehat{\lambda}(\theta)=0$ (which may be done practically by discretizing $\widehat{\lambda}(\theta)$), without any preliminary estimation (of α) as in the first method.

Moreover the likelihood ratio behaves like

$$G_n = n \begin{pmatrix} \frac{1}{n}\sum_{i=1}^n T^{(1)}(X_i, P) \\ \\ \frac{1}{n}\sum_{i=1}^n \gamma^{(1)}(X_i, P) \end{pmatrix}' \Omega^{-1} \begin{pmatrix} \frac{1}{n}\sum_{i=1}^n T^{(1)}(X_i, P) \\ \\ \frac{1}{n}\sum_{i=1}^n \gamma^{(1)}(X_i, P) \end{pmatrix}$$

which is no more than the usual General Method of Moments (GMM) objective function. Because of the likelihood structure of the least favorable family parameterized by μ and λ, we may straightforwardly modify Theorem 2.1 and obtain the Bartlett correctability of the empirical likelihood. Notice however that the two methods, that is, on one hand calculating first the efficient influence function and then applying the empirical likelihood method or on the other hand, applying the empirical likelihood to the constraints seen as estimating functions lead to different objective functions. The first method somehow amounts in estimating $'\Omega^{-1}$, which is actually internally computed in the second method.

Example 2 *Mixture models (see Chap. 4.5 of Bickel and al (1993) and Amari and Kawanabe (1997))*

Let $g(\eta)$ be an unknown positive density on $\mathbb{R}$ and $\{f(x, \theta, \eta), \theta \in R,\ \eta \in R\}$ be a regular parametric exponential family of density

$$f(x, \theta, \eta) = C(\eta, \theta) \exp(\eta T_1(x, \theta) + T_2(x, \theta))$$

where T_1 and T_2 are measurable functions not depending on η, differentiable in θ such that $\frac{\partial T_1(.,\theta)}{\partial \theta}$ is a function of T_1.

The observations $(X_1, X_2, ...X_n)$ are taken from

$$p(x, \theta) = \int f(x, \theta, \eta) g(\eta) d\eta$$

then the efficient influence function of θ is given by

$$T^{(1)}(X, \theta, P) = \frac{\partial T_2(X, \theta)}{\partial \theta} - E_P\{\frac{\partial T_2(X, \theta)}{\partial \theta} | T_1(X, \theta)\}$$

and is independent of the nuisance density g (see Amari and Kawanabe (1997)) for details on the existence of an estimating function in this case).so that we may directly apply theorem 2.1.

Example 3 : *Confidence region for the center of symmetry of a semiparametric family*

Assume that the model is given by $\wp_{\theta,G} = \{P_{\theta,\eta} << \mu$ (any dominating measure) with $\frac{dP_{\theta,\eta}}{d\mu} = \eta(x - \theta)$ and η symmetric about 0, $\eta \in G\}$. To avoid technical difficulties, we assume that the densities are bounded and strictly positive on the whole support. We will also assume some conditions (Lipschitz or Sobolev type conditions) to ensure that the class $\{\frac{\dot{\eta}(x-\theta)^2}{\eta(x-\theta)},\ \eta \in G\}$ is a Donsker class (see for instance Van der Vaart and Wellner(1996)). It is known that θ may be estimated adaptively. An efficient influence function for the parameter θ is given by $-I(\eta)^{-1}\ \frac{\dot{\eta}(x-\theta)}{\eta(x-\theta)}$ with $I(\eta) = \int \frac{\dot{\eta}(x-\theta)^2}{\eta(x-\theta)}dx$. Let $\widehat{\eta}_\theta$ be a symmetrized kernel density estimator of η based on the recentered observations $\{X_i - \theta,\ -X_i - \theta\}$. Consider for instance the construction in van der Vaart (1998), p. 397, then all the conditions of Theorem 3.1 are satisfied (H_1 follows by construction, H_2 is implied by the bounding hypotheses on the family of densities, H_3 follows from the symmetry). Thus the semiparametric empirical log-likelihood given by

$$2 \sup_{\lambda} n^{-1} \sum_{i=1}^{n} \log\left(1 + \lambda' \frac{\dot{\widehat{\eta}}_\theta(X_i - \theta)}{\widehat{\eta}_\theta(X_i - \theta)}\right)$$

is asymptotically $\chi^2(1)$. Bartlett correctability essentially depends on the choice of the smoothing parameter for constructing $\widehat{\eta}_\theta$ and will be investigated elsewhere.

Example 4 : *Empirical likelihood in biased sampling model revisited.*

We refer to chap 6 of Owen(2001) for complete references and give only a few arguments showing how our approach can lead directly to the validity of empirical likelihood for general parameters. In biased sampling problems, we have s-independent samples generated by s biased distributions defined by nonnegative weight functions w_i

$$Q_i(dy) = \frac{w_i(y)}{W_i(P)} P(dy)$$

$$W_i(P) = \int w_i(y) P(dy).$$

We do not assume here that there is a preliminary selection of a "stratum" with known probabilities : this case may be handled quite similarly. We assume for simplicity that P is dominated by a measure μ.

Let

$$X_{1,i}, ...X_{n_i,i}\ i.i.d.\ Q_i\ ,\ i = 1, ..., s$$

and denote $n = \sum_{i=1}^{s} n_i$ the total sample size. We use in the following the dominating measure

$$P_n = n^{-1} \sum_{i=1}^{s} \sum_{j=1}^{n_i} \delta_{X_{j,i}}$$

Notice that this is not the non parametric maximum likelihood estimator for P.

Let us give some examples :

Example 4.1: Stratified sampling

Let X be a random variable taking its value in R^k. And let $S_1, S_2, ..., S_s$ be a partition of the space: $\underset{i=1}{\overset{s}{U}} S_i = R^k$, $S_j \cap S_i = \emptyset$. Then the weight functions are $w_i(x) =$

$I_{S_i}\{x\}$ where $I_A\{.\}$ is the indicator of set A. It is known that, unless auxiliary (transverse) informations are available, the probability P is not identifiable.

Example 4.2: Enriched sample

It is more frequent that a sample obtained by sampling in the population is completed by $s-1$ biased samples (this is for instance the case when a survey is first based on a random sampling scheme and then completed by some additional biased sample), in that case $S_1 = R^k$ and $S_2, ...S_s$ do not form a partition and we have simply in that case $w_1(x) = 1$. It is generally assumed that the biasing scheme i.e. the w_i are known. Then the likelihood of the data is given by

$$L_n(P,\mu) = \Pi_{i=1}^{s}\Pi_{j=1}^{n_i}\frac{dQ_i}{d\mu}(X_{j,i}) = \Pi_{i=1}^{s}\Pi_{j=1}^{n_i}\frac{w_i(X_{j,i})}{W_i(P)}\frac{dP}{d\mu}(X_{j,i}) \tag{17}$$

Example 4.3: Length biased sampling

It happens sometimes that the bias of the sampling scheme is related to the length of the variable (see Vardi (1982)). In survey sampling this often happens when the inclusion probability is proportional to a positive measure of size. In that case the weight is typically of the form $w(x) = x$.

Vardi (1982, 1985), Gill, Vardi and Wellner (1988) have given conditions for the identifiability of P and for the existence and unicity of the non-parametric maximum likelihood estimator (NPMLE) of P say $P_{w,n}$. If one is interested in a functional of P, then the von Mises' principle (known as the delta method) yields asymptotically convergent (and often Gaussian) estimators. The NPMLE of $T(P)$ is no-more than $T(P_{w,n})$.Qin (1993) has generalized the approach of Owen (1988) in the case of example 2 (enriched sample with s=2). We think that it is easier to understand his work in our framework : most of his results may be obtained and generalized in a more straightforward way by using convex duality arguments provided that an adequate (LAN) least favorable family is constructed. The empirical likelihood in a biased sampling model evaluated at θ is defined here similarly to (17) by considering only probability dominated by P_n

$$L_{w,n}(\theta) = \sup_{Q}\left\{\ L_n(Q,P_n),\ Q << P_n,\ T(Q)=\theta,\int dQ = 1\right\}$$

$$= \sup_{\substack{p_{j,i,n}\\ i=1,\dots,s\\ j=1,\dots,n_i}}\left\{\begin{array}{c}\Pi_{i=1}^{s}\Pi_{j=1}^{n_i}\frac{w_i(X_{j,i})}{\sum_{k=1}^{s}\sum_{l=1}^{n_i} w_i(X_{l,k})p_{l,k,n}}np_{j,i,n},\\ T(\sum_{i=1}^{s}\sum_{j=1}^{n_i}p_{j,i,n}\delta_{X_{j,i}})=\theta,\ p_{j,i,n}>0,\ \sum_{k=1}^{s}\sum_{l=1}^{n_i}p_{j,i,n}=1\end{array}\right\}$$

$$= \sup_{\substack{p_{j,i,n},W_i\\ i=1,\dots,s\\ j=1,\dots,n_i}}\left\{\begin{array}{c}\Pi_{i=1}^{s}\Pi_{j=1}^{n_i}\frac{w_i(X_{j,i})}{W_i}np_{j,i,n},\ T(\sum_{i=1}^{s}\sum_{j=1}^{n_i}p_{j,i,n}\delta_{X_{j,i}})=\theta,\ p_{j,i,n}>0\ ,\\ \sum_{k=1}^{s}\sum_{l=1}^{n_i}p_{j,i,n}=1,\sum_{k=1}^{s}\sum_{l=1}^{n_i}w_i(X_{l,k})p_{l,k,n}=W_i\end{array}\right\}$$

which we approximate by the linearized version

$$\widetilde{L}_{w,n}(\theta)=\sup_{\substack{p_{j,i,n},W_i\\ i=1,...,s\\ j=1,...,n_i}}\left\{\begin{array}{c}\Pi_{i=1}^{s}\Pi_{j=1}^{n_i}\frac{w_i(X_{j,i})}{W_i}np_{j,i,n},\ \sum_{i=1}^{s}\sum_{j=1}^{n_i}p_{j,i,n}T^{(1)}(X_{j,i},\theta)=0,\\ p_{j,i,n}\geq 0\ ,\sum_{k=1}^{s}\sum_{l=1}^{n_i}p_{j,i,n}=1,\sum_{k=1}^{s}\sum_{l=1}^{n_i}w_i(X_{l,k})p_{l,k,n}=W_i\end{array}\right\}$$

We assume for simplicity that the gradient (or the estimating function) $T^{(1)}(X_{j,i},\theta)$ only depends on θ. We will also assume the following conditions (see Vardi (1985), Owen (2001)) which ensures the existence of a non parametric maximum likelihood estimator. This condition essentially means that we are not in the situation of example 4.1, that is that we have transverse informations or "linking" observations.

(H_1) For every proper subset B of $\{1,...,s\}$,

$$\left(\underset{i\in B}{\cup}\{X_{1,i},...,X_{n_i,i}\}\right)\cap\left(\underset{i\notin B}{\cup}\{X,\ w_i(X)>0\}\right)\neq\emptyset.$$

Actually this condition plays the role of a qualification constraint ensuring that the original and dual solutions have a finite solution so that the set equivalent to $\overline{\mathcal{P}_n}$ in this framework is non empty.

The following condition also appears in Qin (1993)[1]. It ensures somehow that the sampling bias is not proportional to $T^{(1)}(X,P)$. This hypothesis thus excludes example 4.3 in the case of the mean with $T^{(1)}(x,P)=x-\theta$ and $w(x)=x$, a case which can be considered by itself by minor modification.

(H_2)

$$Var\begin{pmatrix}T^{(1)}(X,P)\\ \\ w(X)\end{pmatrix}\text{ is of rank } q+s$$

Under the condition H_1, the value of the empirical likelihood calculated at Vardi's non-parametric maximum likelihood is

$$\begin{aligned}L_{w,n}&=\sup_{\substack{p_{j,i,n},W_i\\ i=1,...,s\\ j=1,...,n_i}}\left\{\begin{array}{c}\Pi_{i=1}^{s}\Pi_{j=1}^{n_i}\frac{w_i(X_{j,i})}{W_i}p_{j,i,n},\ p_{j,i,n}\geq 0\ ,\\ \sum_{i=1}^{n}p_{j,i,n}=1,\sum_{k=1}^{s}\sum_{l=1}^{n_i}w_i(X_{l,k})p_{l,k,n}=W_i\end{array}\right\}\\ &=\sup_{Q}\left\{L_n(Q,P_n),\ Q<<P_n,\ \int dQ=1\right\}\end{aligned}$$

[1] Notice that Qin (1993) makes the assumption on p.1183 that w(x) ($w_2(x)$ in our notation) is not proportional to x. See also his comment after his Theorem 1.

that is the non-parametric maximum likelihood estimator $P_{w,n}$ of P is the solution of the unconstrained empirical likelihood. The empirical log-likelihood ratio for θ is then

$$R_{E,w,n}(\theta) = \widetilde{L}_{w,n}(\theta)/L_{w,n}$$

Define $w(x) = (w_1(x), ..., w_s(x))$ and $W = (W_1, ..., W_s)$. As in part 2, we may now define the least favorable model

$$\begin{aligned}
(18) \qquad & p_{\lambda,\gamma}(x) \\
= \; & \frac{dP}{d\mu}(x)(1+\lambda' T^{(1)}(x,\theta) + \gamma'(w(x)-W)) \\
& \mathbb{I}\left\{1+\lambda' T^{(1)}(x,\theta) + \gamma'(w(x)-W) \; > 0\right\} \\
= \; & \frac{dQ_i}{d\mu}(x)(1+\lambda' T^{(1)}(x,\theta) + \gamma'(w(x)-W))\frac{W_i}{w_i(x)}\mathbb{I}\{x \in S_i\} \\
& \mathbb{I}\left\{1+\lambda' T^{(1)}(x,\theta) + \gamma'(w(x)-W) \; > 0\right\}
\end{aligned}$$

where the family is indexed by the parameter $(\lambda, \gamma, W) \in \mathbb{R}^q \times \mathbb{R}^s \times \mathbb{R}^s$.

The convex duality arguments of part 1 (used twice) imply that the empirical likelihood ratio is

$$\begin{aligned}
(19) \quad & -2\log(R_{E,w,n}(\theta)) \\
= \; & 2\left(\sup_{W,\gamma}\left(\sum_{k=1}^{s}\sum_{l=1}^{n_k}\log(1+\gamma'(w(X_{l,k})-W)) + \sum_{k=1}^{s} n_k \log(W_k))\right)\right. \\
& \left. - \sup_{W,\lambda,\gamma}\left(\sum_{k=1}^{s}\sum_{l=1}^{n_i}\log(1+\lambda' T^{(1)}(X_{l,k},\theta) + \gamma'(w(X_{l,k})-W) + \sum_{k=1}^{s} n_k \log(W_k)\right)\right)
\end{aligned}$$

which is exactly the log likelihood ratio for testing $\lambda = 0$ in model (18). Compare with Qin (1993). Now under H_1 and H_2, (18) is quadratically differentiable (using the same arguments as in Th. 1) (if H_2 does not hold then $P(1+\lambda' T^{(1)}(X,\theta)+\mu'(w(X)-W) = 0) \neq 0$ and the quadratic differentiability may fail). It follows immediately that (19) is asymptotically $\chi^2(q)$, yielding a confidence region of the form

$$\Re_{1-\alpha} = \left\{\theta, -2\log(R_{E,w,n}(\theta)) \leq \chi^2_{1-\alpha}(q)\right\},$$

asymptotically of level $1-\alpha$. Under additional moments on $(T^{(1)}(X,P), w(X))$ Bartlett correctability also follows from representation (19) as a log-likelihood ratio for testing $\lambda = 0$ in the family (18).

5 Technical details

5.1 Proof of lemma 2.1

Put $r_0 = \exp(-\frac{1}{2}\chi^2_{1-\alpha}(q)) < 1$ for $\alpha \in]0,1[$ and $p_* = \min_i(p_{i,n}) \leq \frac{1}{n} \leq \max(p_{i,n})$. Consider j such that $p_* = p_{j,n}$ then the constraint on the likelihood implies

$$(20) \qquad \begin{aligned} r_0 &\leq \frac{p_* \prod_{i=1,i\neq j}^n p_{i,n}}{\left(\frac{1}{n}\right)^n} \leq \frac{p_* \max \prod_{i=1,i\neq j}^n p_{i,n}}{\left(\frac{1}{n}\right)^n} \text{ with },\sum_{i\neq j} p_{i,n} = 1 - p_* \\ &= n\, p_*(1-p_*)^{n-1}(n/(n-1))^{n-1} \end{aligned}$$

because $(n/(n-1))^{n-1}$ is an sequence converging and increasing to e. This yields the inequality

$$\frac{1}{n}\frac{r_0}{e} \leq p_* \leq \frac{1}{n}$$

Now we have that

$$d_H(Q,P) \leq K(Q,P),$$

where

$$d_H(Q,P) = \int (\left(\frac{dQ}{dP}\right)^{1/2} - 1)^2 dP$$

is the Hellinger distance between Q and P when Q is dominated by P.

It follows that on $\overline{\mathcal{P}_{n,1-\alpha}}$

$$d_H(Q,P_n) \leq \frac{\chi^2_{1-\alpha}(q)}{2n}$$

implies

$$n^{-1}\sum_{i=1}^n \left((np_{i,n})^{1/2} - 1\right)^2 \leq \frac{\chi^2_{1-\alpha}(q)}{2n}$$

and particularly

$$p^* \leq n^{-1}\left(1 + \left(\frac{\chi^2_{1-\alpha}(q)}{2}\right)^{1/2}\right)^2$$

Notice that when $\alpha \to 1$, then the bound converges to $\frac{1}{n}$, that is at the limit all the $p'_{i,n}s$ are equal to $\frac{1}{n}$.

Now notice that $\left(\sum_{i=1}^n p_{i,n}^2\right)^{1/2}(\widetilde{P}_n - P) = \left(\sum_{i=1}^n p_{i,n}^2\right)^{1/2}\left(\sum p_{i,n}(\delta_{X_i} - P)\right)$ is nothing else than a weighted empirical process with deterministic weights $p_{i,n}/\left(\sum_{i=1}^n p_{i,n}^2\right)^{1/2}$. First check that

$$\max_{1\leq i\leq n} \frac{p_{i,n}}{\left(\sum_{i=1}^n p_{i,n}^2\right)^{1/2}} \to 0,$$

since each $p_{i,n}$ is of order $\frac{1}{n}$by the first part of the lemma. Since the X_i are i.i.d. and $\mathcal{F}$ is Donsker and satisfy the uniform entropy condition (6), it follows (see Van der Vaart and Wellner (1996) p. 210 and Koul (1992) Th. 2.2 for the real multidimensional case), that

$$\frac{1}{(\sum_{i=1}^{n} p_{i,n}^2)^{1/2}} \sum p_{i,n}(\delta_{X_i} - P) \rightarrow G_P \ in \ L_\infty(\mathcal{F})$$

where G_P is a gaussian process with covariance operator independent of the weights. Now for $p_n = (p_{1,n}, ..., p_{n,n})$
constrained by $\overline{\mathcal{P}_{n,1-\alpha}}$ (we will use the notation $p_n \Subset \overline{\mathcal{P}_{n,1-\alpha}}$), put for $f \in \mathcal{F}$

$$G_{n,p_n}(f) = \frac{1}{(\sum_{i=1}^{n} p_{i,n}^2)^{1/2}} \sum p_{i,n}(\delta_{X_i} - P)(f).$$

To prove the uniform convergence over $\overline{\mathcal{P}_{n,1-\alpha}}$, it is sufficient to check the uniform equicontinuity condition

$$\lim_{\delta \rightarrow 0} \lim \sup_{n \rightarrow \infty} \sup_{p_n \Subset \overline{\mathcal{P}_{n,1-\alpha}}} \Pr(|| \sup_{||f-g||_{2,P} < \delta} |G_{n,p_n}(f) - G_{n,p_n}(g)| > \varepsilon) \rightarrow 0.$$

where

$$||f - g||_{2,p_n}^2 = \sum_{i=1}^{n} \frac{p_{i,n}^2}{\sum_{i=1}^{n} p_{i,n}^2} (f(X_i) - g(X_i))^2.$$

Using the first part of lemma 2.1, there exists non negative constants A and B such that for any $p_n \Subset \overline{\mathcal{P}_{n,1-\alpha}}$

$$A||f - g||_{2,P_n}^2 \leq ||f - g||_{2,p_n}^2 \leq B||f - g||_{2,P_n}^2 \tag{21}$$

Thus $||f - g||_{2,p_n}^2$ is uniformly equivalent over $\mathcal{P}_{n,1-\alpha}$ to $||f - g||_{2,P_n}^2$.Define also

$$\mathcal{F}_{\delta,P} = \{f - g, \ f \in \mathcal{F}, \ g \in \mathcal{F}, \ ||f - g||_{2,P} < \delta\}$$

is a measurable class of function by the Suslin hypothesis.

Now using standard empirical process arguments, subgaussiannity of $G_{n,p_n}(f)$ (for the seminorm $||f - g||_{2,p_n}^2$), symmetrization and Markov inequality (see the proofs of Th 2.5.2 and 2.8.3 in Van der Vaart and Wellner (1996)), we have for any sequence $\delta \rightarrow 0$, there exists a constant C such that

$$\begin{aligned} \Delta_n &= P(\sup_{||f-g||_{2,p_n} < \delta} |G_{n,p_n}(f) - G_{n,p_n}(g)| > \varepsilon) \\ &\leq CE_P \int_0^{\theta_n/||H||_{2,p_n}} \sqrt{\log(N(\varepsilon||H||_{2,p_n}, \mathcal{F}_{\delta,P}, ||.||_{2,p_n}^2)} d\varepsilon \ ||H||_{2,p_n} \end{aligned}$$

with

$$\theta_n = \sup_{f \in \mathcal{F}_{\delta,P}} ||f(X_i)||_{2,p_n} \leq B \sup_{f \in \mathcal{F}_{\delta,P}} ||f(X_i)||_{2,P_n} = \theta_n^*$$

Now (21) implies that there exists a constant C such that for all $p_n \Subset \overline{\mathcal{P}_{n,1-\alpha}}$,

$$N(\varepsilon||H||_{2,p_n}, \mathcal{F}_{\delta,P}, ||.||_{2,p_n}^2) \leq CN(\varepsilon||H||_{2,P_n}, \mathcal{F}_{\delta,P}, ||.||_{2,P_n}^2).$$

Since we have $||H||_{2,P_n} \geq 1$ and $E_P||H||_{2,P_n}^2 = E_P H^2$, it follows that by Cauchy-Schwartz inequality that

$$\begin{aligned}
\Delta_n^2 &\leq C_1 E_P \left(\int_0^{\theta_n^{'}/A} \sqrt{\sup_Q \log(N(\varepsilon||H||_{2,Q}, \mathcal{F}_{\delta,P}, ||.||_{2,Q})} d\varepsilon \; ||H||_{2,P_n} \right) \\
(22) &\leq C_2 \left(E_P \left(\int_0^{\theta_n^{'}/A} \sqrt{\sup_Q \log(N(\varepsilon||H||_{2,Q}, \mathcal{F}_{\delta,P}, ||.||_{2,Q})} d\varepsilon \right)^2 \right)^{1/2} \left(E_P H^2\right)^{1/2} \\
&\leq C_3 \left(\int_0^{\eta} \sqrt{\sup_Q \log(N(\varepsilon||H||_{2,Q}, \mathcal{F}_{\delta,P}, ||.||_{2,Q})} d\varepsilon + P(\theta_n^{*}/A > \eta) \right)^{1/2} \left(E_P H^2\right)^{1/2}
\end{aligned}$$

Under the uniform entropy condition, the right hand side of (22) does not depend on p_n and may be made as small as we want provided that $\theta_n^{*} \to 0$.This a consequence of Th 2.5.2 in Van der Vaart and Wellner(1996) and follows from the uniform laws of large number over the class $\{f - g,\ f \in \mathcal{F},\ g \in \mathcal{F}\}$, which is measurable in our case because $\mathcal{F}$ is admissible Suslin. Taking the supremum over $\overline{\mathcal{P}_{n,1-\alpha}}$ on the right hand side of (22) yields the result. □

5.2 Proof of Theorem 3.1

We recall the following lemma taken from Bertail and Lo (1996). This result may also be useful in semiparametric applications (when one wants to avoid the splitting trick). For seek of completeness, we give a short proof of this result.

Lemma 5.1 *Assume $X_1, X_2, ...X_n$ are i.i.d. random variables and for each n, let G_n be a symmetric statistic of the observations. Let $\omega(x,t)$ be a function of two variables such that (i) $||\omega(x,t)|| \leq H(x)$ with $EH(X) < \infty$ and (ii) $\omega(x,t)$ is continuous in t. Then $G_n \overset{a.s.}{\to} G$ implies that*

$$S_n^{\omega} = \frac{1}{n} \sum_{i=1}^{n} \omega(X_i, G_n) \overset{a.s.}{\to} E(\omega(X_i, G))$$

Proof : see also Bertail and Lo (1996). It is sufficient to write

$$S_n^{\omega} = E\left(\omega(X_1, G_n)|\mathcal{S}^n\right)$$

where $\mathcal{S}^n$ is the symmetric field containing all the symmetric functions of $X_1, X_2, ..., X_n$. By the extended backward martingale convergence of Blackwell and Dubins (1965), S_n^ω converges with probability one to $E\left(\omega(X_1, G)|\mathcal{S}^\infty\right)$. But by the Hewitt-Savage zero-one law, $\mathcal{S}^\infty$ is non trivial and therefore $E\left(\omega(X_1, G)|\mathcal{S}^\infty\right)$ is constant equal to $E\left(\omega(X_1, G)\right)$.□

This implies the convergence of the estimated efficiency bound to the true one stated in the following lemma.

Lemma 5.2 *Under H_1 and H_3,*

$$I_n(\theta) = n^{-1}\sum_{i=1}^{n} \widetilde{T}^{(1)}(X_i, P_{\theta,\widehat{G}_{\theta,n}})\widetilde{T}^{(1)}(X_i, P_{\theta,\widehat{G}_{\theta,n}})' \to I(\theta, G) \ \ a.s.$$

with

$$I(\theta, G) = E_{P_{\theta,G}}\widetilde{T}^{(1)}(X_i, P_{\theta,G})\widetilde{T}^{(1)}(X_i, P_{\theta,G})'$$

Proof : Apply Lemma 5.1 with $\omega(X_i, G) = \widetilde{T}^{(1)}(X_i, P_{\theta,G})'\widetilde{T}^{(1)}(X_i, P_{\theta,G})$.Under H_3, $||\omega(X_i, G)|| \leq H(X)^2$. Since $P_{\theta,\widehat{G}_{\theta,n}}$ is symmetric of the observations and $EH(X)^2 < \infty$, $I_n \to I(\theta, G)$ as $n \to \infty$. □

We also use the following useful and straightforward result which may be found in Le Cam (1986): p. 188. This simple lemma allows to avoid the assumptions on the existence of third order moments generally made in the literature.

Lemma 5.3 *Let $Y_{k,n}$ be an array of r.v. such that*

(i) $max_k(Y_{k,n}) \to 0$ in probability

(ii) $\sum_{k=1}^n Y_{k,n}^2$ is bounded in probability

and let $\phi(x)$ be a measurable and second order (Peano) differentiable function at 0 with $\phi(0) = 0$ then

$$\sum_{k=1}^{n}\phi(Y_{k,n}) - \phi^{'}(0)\sum_{k=1}^{n}Y_{k,n} - \phi^{''}(0)/2\sum_{k=1}^{n}Y_{k,n}^2 = o_P(1)$$

Proof : Taylor expansion.

Proof of Theorem 3.1 :

The proof is now on the same line as Owen (1990). Notice first that by lemma 5.1, for each fixed λ, $L_n(\lambda) = n^{-1}\sum_{i=1}^n \log\left(1 + \lambda'\widetilde{T}^{(1)}(X_i, P_{\theta,\widehat{G}_{\theta,n}})\right)$ converges to

$$E_{P_{\theta,G}}\log\left(1 + \lambda'\widetilde{T}^{(1)}(X, P_{\theta,G})\right) \leq \log\left(1 + \lambda' E_{P_{\theta,G}}\widetilde{T}^{(1)}(X, P_{\theta,G})\right) = 0$$

by Jensen inequality. Thus the unique maximum of the limit is 0. Because of the strict concavity of $L_n(\lambda)$, the supremum is attained at $\widehat{\lambda}$ which is the unique solution of the equation

$$\frac{1}{n}\sum_{i=1}^{n}\frac{\widetilde{T}^{(1)}(X_i,P_{\theta,\widehat{G}_{\theta,n}})}{1+\lambda'\widetilde{T}^{(1)}(X_i,P_{\theta,\widehat{G}_{\theta,n}})}=0 \tag{23}$$

Since $\sup_n(E_{P_{\theta,G}}||\widetilde{T}^{(1)}(X_i,P_{\theta,\widehat{G}_{\theta,n}})||^2) < E_{P_{\theta,G}}H^2(X) < \infty$,following Owen (2001) p. 220, we obtain $\widehat{\lambda}=O_P(n^{-1/2})$ (use his arguments as well as lemma 5.2 to control the moments uniformly) and that we have by direct Taylor expansion of (23)

$$\left(\sum_{i=1}^{n}\widetilde{T}^{(1)}(X_i,P_{\theta,\widehat{G}_{\theta,n}})\right)-\sum_{i=1}^{n}\widetilde{T}^{(1)}(X_i,P_{\theta,\widehat{G}_{\theta,n}})\widetilde{T}^{(1)}(X_i,P_{\theta,\widehat{G}_{\theta,n}})'\widehat{\lambda}=o_P(1).$$

Under H_2 and H_4, we get that

$$n^{-1/2}\sum_{i=1}^{n}\widetilde{T}^{(1)}(X_i,P_{\theta,G})-n^{-1/2}\sum_{i=1}^{n}\widetilde{T}^{(1)}(X_i,P_{\theta,\widehat{G}_{\theta,n}})=o_P(1). \tag{24}$$

(24) and lemma 5.2 implies

$$\sqrt{n}\widehat{\lambda}=I(\theta,G)^{-1}\left(n^{-1/2}\sum_{i=1}^{n}\widetilde{T}^{(1)}(X_i,P_{\theta,G})\right)+o_P(1)\to N(0,I(\theta,G)^{-1}).$$

Now put $Y_{k,n}=\widehat{\lambda}'\widetilde{T}^{(1)}(X_i,P_{\theta,\widehat{G}_{\theta,n}})$, then we can check

$$\max_{1\le i\le n}(Y_{k,n})=\max_{1\le i\le n}(\widetilde{T}^{(1)}(X_i,P_{\theta,\widehat{G}_{\theta,n}}))\ O_P(n^{-1/2})=o_P(1)$$

and, using lemma 5.2

$$\sum Y_{k,n}^2=n^{1/2}\widehat{\lambda}'\left(n^{-1}\sum_{i=1}^{n\prime}\widetilde{T}^{(1)}(X_i,P_{\theta,\widehat{G}_{\theta,n}})\widetilde{T}^{(1)}(X_i,P_{\theta,\widehat{G}_{\theta,n}})'\right)n^{1/2}\widehat{\lambda}=O_P(1)$$

Thus applying lemma 5.3 with $\phi(x)=\log(1+x)$ and using lemma 5.2 we get

$$\begin{aligned}&\sum_{i=1}^{n}\log\left(1+\widehat{\lambda}'\widetilde{T}^{(1)}(X_i,P_{\theta,\widehat{G}_{\theta,n}})\right)\\=&\left(n^{-1/2}\sum_{i=1}^{n}\widetilde{T}^{(1)}(X_i,P_{\theta,G})\right)'I(\theta,G)^{-1}\left(n^{-1/2}\sum_{i=1}^{n}\widetilde{T}^{(1)}(X_i,P_{\theta,G})\right)+o_P(1)\end{aligned}$$

and the result follows.□

References

Amari, S.I. and Kawanabe, M. (1997). Information geometry of estimating functions in semiparametric statistical models. *Bernoulli*, **3**, 29-54.

Barbe, Ph. and Bertail, P. (1995). *The Weighted Bootstrap.* Lecture Notes in Statistics, **98**. Springer Verlag, N. Y.

Barndorff-Nielsen, O.E. and Hall, P (1988). On the Level-Error after Bartlett adjustment of the likelihood ratio. *Biometrika*, 75, 374-378.

Bertail, P. and Lo A. (1996). Accurate posterior approximations, preprint.

Blackwell, D., Dubins, L. (1962). Merging of opinion with increasing information, *Ann. Math. Statist.*, **33**, 882-886.

Bickel, P.J., Klaasen, C.A.J., Ritov, Y. and Wellner, J.A. (1993). *Efficient Estimation for semiparametric Models.* Johns Hopkins Univ. Press.

Chen, S.X. and Hall, P. (1993). Smoothed Empirical Likelihood Confidence Intervals for Quantiles, *Ann. Statist.*, **21**, 1166-1181

Chen, S.X. (1996). Empirical likelihood confidence intervals for nonparametric density estimation, *Biometrika,* **83**, 2, 329-341.

DiCiccio,T., Hall, P. and Romano, J.(1988). Empirical Likelihood is Bartlett Correctable, *Ann. Statist.*, **19**, 1053-1061.

Gill, R.D. (1989). Non- and semiparametric Maximum Likelihood Estimators and the von Mises Method, *Scand. J. Statist.,* **16**, 97-128.

Gill, R.D., Vardi, Y. and Wellner, J.A. (1988). Large Sample Theory of Empirical Distributions in Biased Sampling Models. *Ann. Statist.*, **16**, 1069-1112.

Gillou A. (1999). Efficient weighted bootstraps for the mean, *J.Statist.Plann.Inference*, **77**, 1, p. 11-35.

Golan, A., Judge, G. and Miller, D.(1996). *Maximum Entropy Econometrics: Robust Estimation with Limited Data*, New York: John Wiley & Sons Inc.

Hall, P. (1990). Pseudo-likelihood Theory for Empirical Likelihood, *Ann. Statist.*, **18**, 121-140.

Hall, P. and La Scala, B. (1990). Methodology and Algorithms of Empirical Likelihood, *Int. Statist. Rev.*, **58**, 109-127.

Hampel, F.R. (1974). The Influence Curve and its Role in Robust Estimation, *J. Amer. Statist. Assoc.*, **69**, 383-393.

Hartley, H.O. and Rao, J.N.K. (1968). A New estimation Theory for Sample Survey, *Biometrika*, **55**, 547-57.

Kiefer, J. and Wolfowitz, J. (1959). Asymptotic Minimax Character of the Sample Distribution Function for Vector Chance Variables, *Ann. Math. Statist.*, **30**, 463-489.

Koul, H.L. (1992). *Weighted Empiricals and Linear Models*, IMS Lecture Notes, **21**.

Le Cam , L. (1986). *Asymptotic methods in statistical decision theory.* Springer Verlag.

Liese, F., Vajda, I. (1987). *Convex Statistical distances.* Teubner, Leipzig.

von Mises, R. (1936). Les lois de Probabilités pour les Fonctions Statistiques, *Ann. Inst. H. Poincaré*, **6**, 185-212.

Murphy, S.A., and van der Vaart, (2000). Semiparametric likelihood ratio inference, *Ann. Statist.* , **25**, 1471-1509.

Mykland, P. (1995). Dual likelihood. *Ann. Statist.*, **23**, 396-421.

Owen, A.B. (1988). Empirical Likelihood Ratio Confidence Intervals for a Single Functional, *Biometrika*, **75**, 2, 237-249.

Owen, A.B. (1990). Empirical Likelihood Ratio Confidence Regions. *Ann. Statist.*, **18**, 90-120.

Owen, A.B. (2001). *Empirical Likelihood.* Chapman and Hall/CRC.

Pons, O., Turckheim E. (1991). Von Mises method, Bootstrap and Hadamard differentiability, Statistics, **22**, 205-214.

Thomas, D. R. and Grunkemeier, G.L. (1975). Confidence Interval Estimation of Survival Probabilities for Censored Data, *J. Amer. Statist. Assoc.* , **70**, 865-871.

Qin, J. and Lawless, J. (1994). Empirical Likelihood and General Estimating Equations, *Ann. Statist.*, **22**, 300-325.

Quin, J. (1993). Empirical Likelihood in Biased Sample Problems. *Ann. Statist.*, **21**, 1182-1196.

van der Vaart, A.W. (1998). *Asymptotic Statistics.* Cambridge series in Statistical and Probabilistic Mathematics.

Vardi, Y. (1982). Non-parametric Estimation in the Presence of Length Bias, *Ann. Statist.* , **10**, 616-620.

Vardi, Y.(1985). Empirical Distributions in Selection Bias Models, *Ann. Statist.*, **13**, 178-203.

Wilks, S.S. (1938). The Large Sample Distribution of the Likelihood Ratio for Testing Composite Hypothesis, *Ann. Math. Statist.*, **9**, 60-62.

Aknowledgement : I would like to thank Emmanuelle Gautherat, Université de Reims and Crest, as well as France Caillavet , INRA-Corela for their carefull reading of the manuscript and all their comments which greatly improved the presentation of this paper. All errors are mine.

Adress for correspondance : Patrice Bertail, CREST, Timbre J340, 3, ave Pierre Larousse, 94205 Malakoff Cedex, France, email: Patrice.Bertail@ensae.fr