Bernoulli **12**(2), 2006, 299–331

Empirical likelihood in some semiparametric models

PATRICE BERTAIL

Laboratoire de Statistique, CREST, Timbre J340, 3 ave. Pierre Larousse, 94205 Malakoff Cedex, France. E-mail: patrice.bertail@ensae.fr

We study the properties of empirical likelihood for Hadamard differentiable functionals tangentially to a well chosen set and give 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 local asymptotic normality 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 location parameter of a symmetric model, M-estimators with some nuisance parameters and general functionals in biased sampling models.

Keywords: Bartlett correction; bias sampling models; Donsker class; empirical likelihood; empirical process; Hadamard differentiability; semiparametric models

1. Introduction

Likelihood inference has been one of the major tools of parametric statistics. Owen (1988, 1990, 2001) introduced the 'empirical likelihood' ratio in a nonparametric setting and obtained a generalization of Wilks's theorem, stating that twice the log-likelihood ratio asymptotically has a χ^2 distribution. The idea of empirical likelihood goes back to Thomas and Grunkemeier (1975), but also to some extent to Hartley and Rao (1968) in the context of survey sampling, where it is known as 'model-based likelihood'. It is closely related to the notion of nonparametric maximum likelihood. For a good review of empirical likelihood, see Hall and La Scala (1990) and Owen (2001) which includes a large bibliography.

For independent and identically distributed (i.i.d.) data, the empirical likelihood ratio allows the construction of confidence regions for smooth parameters – mainly Fréchet differentiable parameters with respect to the Kolmogorov metric, including M-robust parameters (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). This relies on the existence of a convex dual representation of the empirical likelihood, which may itself be seen as the log-likelihood ratio associated with a least favourable parametric family. This representation is closely connected to the important notion of dual likelihood introduced by

1350–7265

Mykland (1995). It leads to a Wilks-type theorem and the validity of the Bartlett correctability of the empirical likelihood ratio, provided that this family satisfies the local asymptotic normality property. This may be checked by showing that it is quadratically differentiable; see Le Cam (1986). We show that Hadamard differentiability (according to a well-chosen set of functions) is sufficient to validate the use of the empirical likelihood of general statistical functionals, extending some results of Owen (1988, 2001).

Section 3 discusses extensions to a more general semiparametric framework with infinite-dimensional nuisance parameters. Our approach (based on derivatives of functionals) is different from that 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 the empirical likelihood in this framework. The idea is essentially based on using an estimated version of the (efficient, when it is available) influence function which serves as the basis for the empirical likelihood procedure. This type of construction is already implicit in many recent studies of empirical likelihood in specific semiparametric models (see Chen and Hall 1993; Chen 1996; Qin and Jing 2001; and the examples given in Owen 2001, Chapters 5 and 6). We give here some generic arguments for studying such models and to prove the asymptotic validity of the empirical likelihood under weak assumptions.

In Section 4, we give examples and applications to semiparametric models, including confidence intervals for the location parameter of a symmetric distribution, a problem discussed in Chapter 10 of Owen (2001), M-estimators with nuisance parameters, and parameter estimation in some mixture models. We also re-examine the results of Qin (1993) in biased sampling models under weaker conditions. We do not discuss algorithmic problems here, in spite of their importance: see 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 $X_1, \ldots, X_n, \ldots$ be i.i.d. random variables, taking values on a space $\mathcal{X}$, defined on a probability space (Ω, A, P_Ω) with common probability measure P belonging to a convex set ϱ of signed measures (containing the Dirac measure). In the following, we are interested in constructing a confidence region for the functional parameter $\theta = T(P)$ defined on ϱ, taking values in $\mathbb{R}^q$ (see von Mises 1936). The empirical probability measure defined by

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

is known to be the nonparametric maximum likelihood estimator (NPMLE) of P (see Gill 1989; Owen 1988; 1990) The NPMLE of $T(P)$ is then its empirical counterpart,

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

Many statisticians since von Mises have been interested in deriving the asymptotic properties of $\hat{\theta}_n$ using differentiability assumptions on T (see Gill 1989) via Taylor expansion (the delta method). Under some regularity conditions, it is then possible to construct 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 construct directly a confidence region without relying on previous estimations.

The empirical likelihood ratio evaluated at θ is defined by

$$R_{E,n}(\theta) = \sup_{Q_n \in \mathcal{P}_n} \left\{ \prod_{i=1}^{n} \frac{\mathrm{d}Q_n}{\mathrm{d}P_n}(X_i),\ T(Q_n) = \theta \right\},$$

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

$$\mathcal{P}_n = \left\{ \tilde{P}_n = \sum_{i=1}^{n} p_{i,n}\delta_{X_i},\ p_{i,n} \geqslant 0,\ \sum_{i=1}^{n} p_{i,n} = 1 \right\}.$$

Actually it should be noticed that for certain values of θ, $R_{E,n}(\theta)$ may not have any solution (consider for instance $\theta = \mathrm{E}_P X$; 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 is of no 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

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

A better way to see this problem from a probabilistic point of view is to consider (1) as the minimization of the Kullback distance

$$K(Q, P) = \begin{cases} -\int \log\left(\dfrac{\mathrm{d}Q}{\mathrm{d}P}\right)\mathrm{d}P, & \text{if } Q \ll P \\ \infty, & \text{otherwise,} \end{cases}$$

between Q and P_n, over all probabilities Q dominated by the empirical distribution P_n, where Q satisfies the constraint $T(Q) = \theta$. That is,

$$-\log(R_{E,n}(\theta))/n = \inf_{Q_n \in \mathcal{P}_n} \left(K(Q_n, P_n),\ T(Q_n) = \theta,\ \int \mathrm{d}Q_n = 1 \right). \tag{2}$$

This 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 Rockafeller 1968; Liese and Vajda 1987), have been suggested in place of the Kullback distance. This has given rise to what econometricians call 'maximum entropy econometrics' (see, for instance, Golan *et al.* 1996) when choosing the relative entropy. The Cressie–Read distance has also been suggested by many authors (see

Owen 2001; Baggerly 1998; Corcoran 1998; and references therein). Most of the (first-order) asymptotic results that we discuss here can be obtained in the much more general framework of I-divergences for which a dual representation holds (see Borwein and Lewis 1991; Bertail 2004). 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 $T(P) = \mathrm{E}_P X$ is the mean of a q-variate random variable with a covariance matrix $\Sigma = \mathrm{var}(X)$ of rank q then $-2\log(R_{E,n}(\theta))$ converges in distribution to $\chi^2(q)$, a result which is the nonparametric analogue to Wilks's (1938) result.

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

$$\mathfrak{R}_{n,1-\alpha} = \{\theta,\, \Lambda_n(\theta) = -2\log(R_{E,n}(\theta)) \leqslant \chi^2_{1-\alpha}(q)\}. \tag{3}$$

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

$$\mathfrak{R}_{n,1-\alpha} = \{T(\tilde{P}_n),\, \tilde{P}_n \in \overline{\mathcal{P}_{n,1-\alpha}}\}, \tag{4}$$

with

$$\overline{\mathcal{P}_{n,1-\alpha}} = \left\{ \sum_{i=1}^n p_{i,n}\delta_{X_i},\, \sum_{i=1}^n p_{i,n} = 1,\, p_{i,n} \geqslant 0,\, -2\sum_{i=1}^n \log\left(\frac{p_{i,n}}{1/n}\right) \leqslant \chi^2_{1-\alpha}(q) \right\}$$

$$= \left\{ Q \in \mathcal{P},\, K(Q,\, P_n) \leqslant \frac{\chi^2_{1-\alpha}(q)}{2n},\, \int \mathrm{d}Q = 1,\, Q \geqslant 0 \right\} \subset \mathcal{P}_n.$$

This equality, which plays an important role in our analysis, fails for nonlinear statistics, for which we simply have $\{T(\tilde{P}_n),\, \tilde{P}_n \in \overline{\mathcal{P}_{n,1-\alpha}}\} \subset \mathfrak{R}_{n,1-\alpha}$. Notice that for any fixed n, the set $\overline{\mathcal{P}_{n,1-\alpha}}$ contains P_n for any fixed value of $0 < \alpha \leqslant 1$. One purpose of this paper is to show that, for Hadamard differentiable functionals, $\{T(\tilde{P}_n),\, \tilde{P}_n \in \overline{\mathcal{P}_{n,1-\alpha}}\}$ is still an asymptotically valid $1-\alpha$ confidence region for $T(P) \in \mathbb{R}^q$. This means that, for smooth functionals, the image by T of the ball (constructed with the Kullback distance) centred at P_n with diameter $\chi^2_{1-\alpha}(q)/n$ is an asymptotically correct confidence interval for $T(P)$.

Owen (1988) and DiCiccio *et al.* (1991) showed in the case of the mean (under moment and Cramér conditions) that

$$P(\theta \in \mathfrak{R}_{n,1-\alpha}) = 1 - \alpha + O(n^{-1}). \tag{5}$$

This is actually the error rate for a two-sided confidence interval based on the normal asymptotic distribution in regular cases such as smooth functions of the means. DiCiccio *et al.* (1991) proved that empirical likelihood ratios (as in the parametric case) are Bartlett correctable (under some regular assumptions); see also DiCiccio and Romano (1990) and Hall (1990). The Bartlett correction aims to fix the expectation of $\Lambda_n(\theta) = -2\log(R_{E,n}(\theta))$ exactly at q, the expectation of the limiting distribution. Since the first term in the Edgeworth expansion of $\Lambda_n(\theta)$ is of order n^{-1} multiplied by a polynomial function of degree 1, a simple and explicit correction of the form $q\Lambda_n(\theta)/\mathrm{E}_P\Lambda_n(\theta)$ results in a confidence region with a coverage error of order $O(n^{-2})$ (see Bickel and Ghosh 1990). Notice that this is also the rate that can be obtained with the bootstrap for two-sided confidence intervals in smooth cases. In practice $\mathrm{E}_P\Lambda_n(\theta)$ is unknown but can be replaced by a suitable estimator: if the estimator is

chosen adequately, then accuracy up to $O(n^{-2})$ still holds (Barndorff-Nielsen and Hall 1988). For instance, one may use a jackknife estimator or a bootstrap estimator of this mean. Thus empirical likelihood does not require intensive computations (at least theoretically), contrary to the bootstrap distribution, which in most cases needs to be approximated by Monte Carlo simulation. It should be noted that a 'corrected version' of the weighted bootstrap has been proposed in Barbe and Bertail (1995) to improve on 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), typically leads to an accuracy of order $O(n^{-5/2})$, for symmetric statistics, under regularity assumptions on the functional of interest. However this requires strong knowledge of the functional of interest (the gradients up to order 6), whereas, as we will see, the empirical likelihood leads to an accuracy of order $O(n^{-2})$ in a quite automatic way.

Computational problems may, however, arise in the algorithms used to build the empirical likelihood regions if the parameter is very complicated; see Owen (2001) for algorithms and tricks to alleviate these. 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 the multivariate mean.

The case of a Fréchet differentiable functional with respect to the Kolmogorov metric has been studied by Owen (1988, 1990) and the case 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 1990; Barbe and Bertail 1995). In the next subsection, 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. Hadamard differentiability is a notion of differentiability in which the remainder is controlled over compact neighbourhoods. 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. It also preserves asymptotic efficiency, which makes it a privileged tool in semiparametric analysis (see van der Vaart 1998). The main problem in using this notion in statistical applications lies in the choice of metric or topology to ensure both the convergence of the empirical process and the Hadamard differentiability of the functional.

For the 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 functions, here $L^2(P) = \{h, \mathrm{E}_P h^2 < \infty\}$,

endowed with $\|f\|_{2,P} = (\mathrm{E}_P(f^2))^{1/2}$. $\mathcal{L}_\infty(\mathcal{F})$ is equipped with the uniform convergence norm (or equivalently Zolotarev metric)

$$\|P - Q\|_{\mathcal{F}} = d_{\mathcal{F}}(P, Q) = \sup_{h \in \mathcal{F}} \left| \int h \mathrm{d}P - \int h \mathrm{d}Q \right|.$$

To avoid measurability problems, we assume that expectations (probabilities) are outer expectations (outer probabilities) 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 envelope H satisfying

$$0 < \int H^2 \, \mathrm{d}P < \infty, \tag{6}$$

so that the empirical process $n^{1/2}(P_n - P)$ indexed by $\mathcal{F}$ converges (as an element of $\mathcal{L}_\infty(\mathcal{F})$) to a limit G_P, which is a tight Borel measurable element of $\mathcal{L}_\infty(\mathcal{F})$ with uniformly $\|\cdot\|_{2,P}$ continuous sample paths $f \to G_P(f)$. Extensive references and results on empirical processes indexed by class of functions and conditions for $\mathcal{F}$ to be Donsker may be found in van der Vaart and Wellner (1996). Denote the covering number (the minimal number of balls of radius ε for the seminorm $\|\cdot\|$ needed to cover $\mathcal{F}$) by $N(\varepsilon, \mathcal{F}, \|\cdot\|)$. We will assume the usual uniform entropy condition

$$\int_0^\infty \sup_{Q \in \mathcal{D}} \sqrt{\log(N(\varepsilon \|H\|_{2,Q}, \mathcal{F}, \|\cdot\|_{2,Q}))} \mathrm{d}\varepsilon < \infty, \tag{7}$$

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

The following lemma shows that the set $\overline{\mathcal{P}_{n,1-\alpha}}$ is small and contained in a band around P_n. This 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 exist non-negative constants, $a(\alpha) < 1 < b(\alpha)$, such that for any $\tilde{P}_n = \sum_{i=1}^n p_{i,n} \delta_{X_i}$ in $\overline{\mathcal{P}_{n,1-\alpha}}$ we have*

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

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 (6) *and* (7)*, then*

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

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

Now define $B(\mathcal{F}, P)$, the subset of $\mathcal{L}_\infty(\mathcal{F})$ (seen as functions (or paths) $f \to \mu f = \int f d\mu$ from $\mathcal{F} \to \mathbb{R}$) which are $\|\cdot\|_{2,P}$-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 $B(\mathcal{F})$ adapted from Pons and Turckheim (1991). Notice that differentiation 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 and Wellner, 1996, Chapter 3.9; van der Vaart, 1998, Chapter 20).

Definition 2.1. *The functional* T *from* $\mathcal{P} \subset \mathcal{L}_\infty(\mathcal{F})$ *to* $\mathbb{R}^q$ (*or any Banach space* $(\mathcal{B}_1, \|\cdot\|_{\mathcal{B}_1})$) *is said to be Hadamard (or compact) differentiable at* $P \in \mathcal{P}$ *tangentially to* $B(\mathcal{F}, P)$ – *or* T *is* $HDT_{\mathcal{F}} - P$ *for short* – *if and only if 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* $P + t_n h \in \mathcal{P}$,

$$\frac{T((P + t_n h_n)) - T(P)}{t_n} - dT_P.h \to 0, \qquad \text{as } t_n \to 0.$$

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

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

with the normalization

$$\mathrm{E}_P T^{(1)}(X, P) = 0.$$

In the terminology of robustness, $T^{(1)}(x, P)$ 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-degenerate (that is to say, the covariance operator associated with $T^{(1)}(X, P)$ has full rank).

Assuming that T is $HDT_{\mathcal{F}} - P$ with canonical gradient $T^{(1)}(\cdot, P)$, then we have

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

Lemma 2.1 implies that the solutions $\tilde{P}_n$ 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

$R_n(\tilde{P}_n, P) = o_P((\sum_{i,n}^2)^{1/2}) = o_P(n^{-1/2})$ uniformly over all admissible $\tilde{P}_n$ in $\overline{\mathcal{P}_{n,1-\alpha}}$. These arguments suggest that the empirical likelihood ratio may be replaced by a linearized version

$$\begin{aligned}\bar{R}^L_{E,n}(P) &= \sup_{\tilde{P}_n \in \mathcal{P}_n} \left\{ \prod_{i=1}^n n d\tilde{P}_n(X_i),\ \mathrm{E}_{\tilde{P}_n} T^{(1)}(X, P) = 0 \right\} \\ &= \sup_{p_{i,n}, i=1,\ldots,n} \left\{ \prod_{i=1}^n np_{i,n},\ \sum_{i=1}^n p_{i,n} T^{(1)}(X_i, P) = 0,\ p_{i,n} \geqslant 0,\ \sum_{i=1}^n p_{i,n} = 1 \right\}, \quad (8)\end{aligned}$$

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

$$\mathfrak{R}^{\xi}_{n,1-\alpha} = \left\{ T(P) + \int T^{(1)}(\cdot, P) \mathrm{d}\tilde{P}_n,\ \tilde{P}_n \in \overline{\mathcal{P}_{n,1-\alpha}} \right\}.$$

But the analogue of (4),

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

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 (8) so that the use of $\mathfrak{R}^{T}_{n,1-\alpha}$ is asymptotically justified.

The following theorem states that these 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 functions $\mathcal{F}$ with envelope H, satisfying* (7) *such that T defined on $\mathcal{P}$ is $HDT_{\mathcal{F}} - P$ with gradient $T^{(1)}(\cdot, P)$. If $\mathrm{var}(T^{(1)}(X, P)) < \infty$ is of rank q, then we have*

$$-2 \log\left(\bar{R}^L_{E,n}(P) \right) \underset{n\to\infty}{\rightarrow} \chi^2(q) \quad (9)$$

and

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

If, in addition, $\bar{R}^L_{E,n}(P) \equiv \bar{R}^L_{E,n}(\theta)$ only depends on θ through $T^{(1)}(x, P) \equiv T^{(1)}(x, \theta)$, assume the Cramér condition

$$\overline{\lim_{\|t\|\to\infty}} |\mathrm{E}_p \exp(\mathrm{i}t' T^{(1)}(X_i, P))| < \infty, \quad (10)$$

and

$$\mathrm{E}_p \| T^{(1)}(X_i, P) \|^s < \infty, \qquad \text{for } s \geqslant 8 + \varepsilon,\ \varepsilon > 0, \quad (11)$$

and then the Bartlett corrected confidence region,

$$\mathfrak{R}^{B}_{1-\alpha} = \left\{ \theta,\ q \log\left(\bar{R}^L_{E,n}(\theta) \right) / \mathrm{E}\left(\log\left(\bar{R}^L_{E,n}(\theta) \right) \right) \leqslant \chi^2_{1-\alpha}(q) \right\},$$

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

$$P(\theta \in \mathfrak{R}^{B}_{1-\alpha}) = 1 - \alpha + O(n^{-2}).$$

Proof. For a better understanding of these results, we give here a short proof. Recall that using standard variational calculus (see Owen 2001), the solution of the maximization problem (8) is given by

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

where λ, the Kuhn–Tucker coefficient, satisfies

$$\sum_{i=1}^{n} p_{i,n}(\lambda) = 1, \qquad 0 \leqslant p_{i,n}(\lambda) \leqslant 1,$$

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

By standard Kuhn–Tucker duality theory, we have

$$-2 \log\left(\overline{R}^{L}_{E,n}(P)\right) = 2 \sup_{\lambda \in \mathbb{R}^q} \sum_{i=1}^{n} \log(1 + \lambda' T^{(1)}(X_i, P)) := 2 \sup_{\lambda \in \mathbb{R}^q} L_n(\lambda). \tag{12}$$

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

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

is a density defined for any λ in the neighbourhood of 0 (notice that we may 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 favourable model (see Bickel *et al.* 1993; DiCiccio and Romano 1990) indexed by the Kuhn–Tucker parameters. This interpretation of the empirical likelihood ratio as the likelihood ratio associated with a least favourable family, which is already present in DiCiccio and Romano (1990), will be particularly useful in semiparametric models. Since $L_n(0) = 0$, $L_n(\lambda)$ may also be seen precisely 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 and strictly concave. Thus it admits a unique maximum. Moreover, by concavity of the log,

$$\mathrm{E}_P(\log(1 + \lambda' T^{(1)}(X_i, P))) \leqslant \log(1 + \lambda' \mathrm{E}_P T^{(1)}(X_i, P)) = 0.$$

Thus $\mathrm{E}_P(\log(1 + \lambda' T^{(1)}(X, P)))$ has a unique maximum at $\lambda = 0$ and the maximum likelihood estimator (MLE) converges to 0. Notice that since $\mathrm{var}(T^{(1)}(X, P))$ exists and is strictly positive, the family $\{p_\lambda,\ \lambda \in \mathbb{R}^q\}$ is differentiable in quadratic mean and the associate log-likelihood ratio is locally asymptotically normal (see Le Cam 1986). Indeed, the differentiability in quadratic mean follows from Lemma 7.6 of van der Vaart (1998: 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 $\mathrm{var}(T^{(1)}(X,\ P))>0$ (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 $\lambda=0$ and, as in Mykland (1995), (9) follows.

Because $L_n(\lambda)$ is itself a parametric log-likelihood ratio (as a function of the parameter λ), it is Bartlett correctable under (10) and (11). 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 efficient confidence region.

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

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

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

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

with $P(\theta(P)\in\mathfrak{R}^{\xi}_{n,1-\alpha})\to 1-\alpha$, since the parameter $\xi(Q)$ is linear. More precisely, we have $P(\theta(P)\in\mathfrak{R}^{\xi}_{n,1-\alpha})=1-\alpha+O(n^{-1})$ if (10) and (11) hold for $s>4$.

Now we have by Hadamard differentiability

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

The results now follow by similar arguments to those in Theorem 20.8 of van der Vaart (1998). Take $t=(\sum p^2_{i,n})^{-1/2}$, which is of order $o(n^{-1/2})$ uniformly over $\overline{\mathcal{P}_{n,1-\alpha}}$ by Lemma 2.1, and $h_n=(\sum p^2_{i,n})^{-1/2}(\tilde{P}_n-P)\in\mathcal{L}_\infty(\mathcal{F})$ in the definition of Hadamard differentiability. Then, by definition, $h=G_P\in B(\mathcal{F},\ P)$. We deduce that $R(\tilde{P}_n,\ P)=o(n^{-1/2})$ uniformly over $\overline{\mathcal{P}_{n,1-\alpha}}$. It follows that $\mathfrak{R}^{\xi}_{n,1-\alpha}$ and $\mathfrak{R}^{T}_{n,1-\alpha}$ are asymptotically equivalent. □

Remark 2.1. The proof essentially relies on a convex duality argument that allows us to write the empirical likelihood as a true parametric likelihood ratio indexed by the Kuhn–Tucker coefficient. Duality is used in a constructive way in Mykland (1995): here duality is actually a consequence of the fact that the Kullback distance is a convex statistical distance (see Liese and Vajda 1987). The duality principle generates a least favourable family, which may be verified to be locally asymptotically normal, under the condition $\mathrm{var}(T^{(1)}(X,\ P))<\infty$. This dual representation is easy to obtain when there is a finite number of constraints in the

empirical likelihood (here if the parameter is of finite dimension). Actually, even if there is an infinite number of constraints (which is actually the case in many semiparametric problems), a dual representation still holds (see Remark 3.1).

Hadamard differentiability is needed to show that $\mathfrak{R}^{\xi}_{n,1-\alpha}$ and $\mathfrak{R}^{T}_{n,1-\alpha}$ 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 5 below). It also may be used to prove (first-order) asymptotic results, when the Kullback distance in (2) is replaced by another convex statistical distance, the entropy or even any convex statistical distance (I-divergence) for which a convex duality principle holds (see Borwein and Lewis 1991; Leonard 2001; Bertail 2004). In the case of the 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 distances (see Baggerly 1998; Corcoran 1998). Of course this result is theoretical in that $\mathrm{E}_P(\log(\overline{R}^{L}_{E,n}(\theta)))$ is generally unknown and must be estimated. This can be done by using jackknife or even bootstrap procedures, even though this may be computer-intensive. The fact that accuracy up to $O(n^{-2})$ may still hold even with an estimated value is supported by results of Barndorff-Nielsen and Hall (1988).

Remark 2.2. Owen (1990) and Qin and Lawless (1994) showed how results of this kind may be used for M-estimates. Indeed, in this 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 *et al.* 1993) is known and independent of the nuisance parameters – see Amari and Kawanabe (1997) for the existence of general estimating equations – then this would be the best candidate for $T^{(1)}$. However, many problems may appear. First, the efficient influence function is not always easy to obtain since most of the time it involves the projection into an infinite-dimensional space. Second, it is not clear whether this expression may be used in practice, for $T^{(1)}(\cdot, P)$ may have a very complicated form and depend on some nuisance parameter. This kind of problem typically appears in the 'challenges' exposed in Chapter 10 of Owen (2001). We shall further examine these points in the next section.

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 MLE for λ in this locally asympotically normal family is such that

$$\hat{\lambda}_n(P) = S_n^{-2} \sum_{i=1}^{n} T^{(1)}(X_i, P))(1 + o_p(1)) = O_P\left(\frac{1}{\sqrt{n}}\right),$$

with

$$S_n^2 = \frac{1}{n} \sum T^{(1)}(X_i, P) T^{(1)}(X_i, P)'.$$

Apply the strong law of large numbers to the first term and the central limit theorem to the second term to obtain

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

It follows that at the MLE $\hat{\lambda}_n(P)$, the empirical likelihood ratio $L_n(\hat{\lambda}_n(P))$ behaves asymptotically like the usual generalized method of moments (see Hansen 1982) objective function

$$\left(\frac{1}{\sqrt{n}}\sum_{i=1}^{n} T^{(1)}(X_i,\ P)\right)' (S_n^2)^{-1} \left(\frac{1}{\sqrt{n}}\sum_{i=1}^{n} T^{(1)}(X_i,\ P)\right),$$

which may be seen as the square of the norm of a self-normalized sum (which is asymptotically $\chi^2(q)$). This self-normalization, carried out internally by the optimization procedure as noticed by several authors (see Owen 2001), is essentially due to the locally asymptotically normal 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 speaks in favour of empirical likelihood against other third-order correction methods for constructing confidence intervals such as iterative inversion of Edgeworth expansion or weighted bootstrap approximations. Indeed, in these cases, the structure of the statistic (its Hoeffding decomposition in terms of orthogonal U-statistics) plays a fundamental role in implementing these methods (see Barbe and Bertail 1995). However, it should be noted that this is only possible when the influence function is simple and does not depend on additional nuisance parameters. If $T^{(1)}(\cdot,\ P)$ 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 2.1 is the key point for showing the validity of the approximation. Of course other types 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 called semiparametric if $\varrho_{\Theta,H} = \{P_{\theta,G} \in \varrho,\ \theta \in \Theta,\ G \in H\}$ is a set of probability measures indexed both by a parameter of interest θ in a subset Θ of $\mathbb{R}^k$ and a nuisance parameter G in a space H, possibly of infinite dimension. Such models and the generalization to the infinite-dimensional case for Θ are studied at length in Bickel *et al.*

(1993). One of the main problems which appears in semiparametric models is that generally the parameter of interest $\theta = T(P_{\theta,G})$ is defined on a set $\varrho_{\Theta,H}$ smaller than ϱ considered above, so that $T(P_n)$, $T(\tilde{P}_n)$ with $\tilde{P}_n = \sum_{i=1}^n p_{i,n}\delta_{X_i}$, or the gradient $T^{(1)}(x, Q)$ at $Q = P_n$ may not be defined properly and may depend on some nuisance parameter.

One semiparametric approach generally used in such a context is to extend the functional $T(\cdot)$ to a more general space. For this purpose, one generally introduces a pseudo-metric d on ϱ and defines a pseudo-projection Π (not necessarily unique, or even a sequence of pseudo-projectors Π_m, $m \to \infty$) into the model of any $P \in \varrho$ such that

$$\Pi(P) = \arg \min_{Q \in \varrho_{\Theta,H}} (d(P, Q)).$$

Then the functional

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

extends T defined on $\varrho_{\Theta,H}$ to ϱ. $\tilde{T}(P_n)$ defines a minimum distance estimator (see Bickel *et al.* 1993). More generally, we may choose Π to be any function (or sequence of functions) which extends the functional $T(\cdot)$ to ϱ, the set of all signed measures. For instance, if $\varrho_{\Theta,G}$ is the set of probability measures with continuous density with respect to the Lebesgue measure λ, we may choose Π to be the convolution of P with a continuous kernel if P does not have a continuous density with respect to λ and $\Pi(P) = P$ otherwise. In that case, because of the linearity of the convolution operator, the influence function of $\tilde{T}$ will be the smoothed version of the influence function of T – see Chen and Hall (1993) and Chen (1996) for such use in the context of empirical likelihood. If such an extension exists then we may define the empirical likelihood ratio in the semiparametric model as

$$R_{E,n}(\theta) = \sup_{p_{i,n}, i=1,\ldots,n} \left\{ \prod_{i=1}^n np_{i,n},\ \tilde{T}\left(\sum_{i=1}^n p_{i,n}\delta_{X_i} \right) = \theta,\ p_{i,n} \geqslant 0,\ \sum_{i=1}^n p_{i,n} = 1 \right\}. \tag{13}$$

However, in many problems, the right 'efficient' choice of the extension Π (in the sense of Bickel *et al.* 1993) depends on the geometry of the problem. Ideally it should be chosen in such a way that the gradient at P of $\tilde{T}$ coincides with the efficient influence function of T in the original semiparametric problem. However, in many problems it may be easier to work directly with the efficient influence function or a non-efficient but tractable one.

In the following $\tilde{T}^{(1)}(\cdot, P_{\theta,G})$ is the efficient influence function when it is available or a particular tractable influence function of $\tilde{T}$. $\tilde{T}^{(1)}(\cdot, P_{\theta,G})$ may be simply proportional to an influence function, as is the case in Example 4 below. The linearized version of the original problem is thus

$$\sup_{p_{i,n}, i=1,\ldots,n} \left\{ \prod_{i=1}^n np_{i,n},\ \sum_{i=1}^n p_{i,n}\tilde{T}^{(1)}(X_i, P_{\theta,G}) = 0,\ p_{i,n} \geqslant 0,\ \sum_{i=1}^n p_{i,n} = 1 \right\}. \tag{14}$$

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 $\hat{G}_{\theta,n}$ of G.

Assuming that such a consistent estimator exists, then we may use the approximate semiparametric empirical likelihood

$$\tilde{R}_{E,n}(\theta) = \sup\left\{ \prod_{i=1}^{n} np_{i,n},\ \sum_{i=1}^{n} p_{i,n}\tilde{T}^{(1)}(X_i,\ P_{\theta,\hat{G}_{\theta,n}}) = 0,\ p_{i,n} \geqslant 0,\ \sum_{i=1}^{n} p_{i,n} = 1 \right\}. \tag{15}$$

It should be noted that in general the solution of the original problem (13) and that of (15) are different but asymptotically equivalent (for instance, if $\tilde{T}$ is Hadamard differentiable with a gradient continuous in P).

Another possible definition which will ease the technical difficulties that we will encounter later, when studying the asymptotic properties of this approximate empirical likelihood, is to rely on the splitting trick frequently used in the semiparametric literature. For this purpose, define two estimators of G, $G^{(1)}_{\theta,n/2}$ and $G^{(2)}_{\theta,n/2}$, 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

$$\tilde{R}_{E,n}(\theta) = \sup_{p_{i,n},i=1,\ldots,n} \left\{ \prod_{i=1}^{n} np_{i,n},\ p_{i,n} \geqslant 0,\ \sum_{i=1}^{n} p_{i,n} = 1, \right.$$
$$\left. \sum_{i=1}^{[n/2]} p_{i,n}\tilde{T}^{(1)}(X_i,\ P_{\theta,G^{(2)}_{\theta,n/2}}) + \sum_{i=[n/2]+1}^{n} p_{i,n}\tilde{T}^{(1)}(X_i,\ P_{\theta,G^{(1)}_{\theta,n/2}}) = 0 \right\}. \tag{16}$$

The asymptotic results that we obtain may be derived in this case too by using the same kind of arguments. 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.

Remark 3.1. As mentioned by a referee and an associate editor, computing the efficient influence function may be a difficult task in many semiparametric models, so that the attractive automatic implementation of empirical likelihood may be lost. Of course the choice of an inefficient influence function is always possible, but even in that case the influence function may depend on some nuisance parameter. Empirical likelihood then has to be adapted to the semiparametric problem under consideration.

A possible solution is to profile the empirical likelihood according to the nuisance parameter and to maximize it or to integrate it over the nuisance parameter (see Leblanc and Crowley 1995; Owen, 2001, Section 3.5). Another solution is simply to try to estimate the nuisance parameter by using some plug-in estimators. This idea is used in many recent constructions – see Qin and Jing (2001) as well as the examples given in Owen (2001, Chapters 5 and 6) and our previous construction. This solution is only possible when the model may be reduced to a problem with a finite number of constraints indexed both by the parameter of interest and the nuisance parameter. Moreover, the choice of the estimator of the nuisance parameter estimator may have some drastic effects on the limiting distribution of the empirical likelihood ratio which

may asymptotically converge to a mixture of χ^2 distributions. For an example, see the study of Qin and Jing (2001). In their paper, the problem is essentially due to the choice of the constraints (choice of the influence function) and of the estimator of the nuisance parameter (which are not 'orthogonal' in some sense). In the following, we will show how to choose the estimator of the nuisance parameter to avoid this kind of problem. The split empirical likelihood introduced before is also a solution in other complicated situations.

A second solution is to note that the semiparametric problem generates a large number (even an infinity) of linear constraints. This suggests the use of a growing number of constraints or moments in the empirical likelihood maximization program – see, for instance, Chen and Cui (2003) for generalized linear models. Such a construction is also used in many applications of empirical likelihood in econometrics. Indeed, as mentioned by van der Vaart (1995) and Bickel *et al.* (1993), a semiparametric problem is somehow asymptotically equivalent to an infinite-dimensional M-estimation problem, with an associated operator A. To see this, one may write all the scores with respect to the parameter of interest and to all finite-dimensional parametrizations of the nuisance parameter and write that, under the true model, they all have expectations equal to 0. This approach leads to several issues. How should one choose the constraints which make sense for the problem? How many constraints should be retained and is there any efficiency loss in retaining only some constraints? The first issue is closely related to the choice of an adequate base for the tangent space of the semiparametric model (see Bickel *et al.*, 1993). The possibility of asymptotically reducing the problem to a finite number of efficient constraints is actually closely linked to semiparametric analysis and the existence of estimating equations (see Amari and Kawanabe 1997). This is the approach that we have taken in this paper. However, since the operator A may be very complicated (and projections into the tangent space difficult to carry out), it may still be interesting to look at the empirical likelihood under an infinite number of constraints. In this case, it is still possible to obtain a dual version of the empirical likelihood problem, using, for instance, the results of Leonard (2001). Of course, the dual likelihood (indexed by a parameter of infinite dimension) may be more difficult to control in that case, unless one has some control over the operator A (and/or the size of the functions inducing the constraints as measured by some entropy index). As far as the number of constraints and the practical implementation of such problems are concerned, some elements are already given in the literature on nonlinear convex analysis (in particular, semi-infinite and semi-definite programming). We will not pursue this analysis here; it is currently under study.

3.2. Asymptotic validity under weak assumptions

Consider the optimization program (13). Similar arguments to those in Section 2 yield the dual equality

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

Notice, however, that $\tilde{L}_n(\lambda)$ cannot be seen directly as a log-likelihood ratio because of the dependence on $\hat{G}_{\theta,n}$ and the absence of recentring. Indeed, there is no reason why $\mathrm{E}_{P_{\theta,G}}\tilde{T}^{(1)}(X_i,\, P_{\theta,\hat{G}_{\theta,n}}) = 0$. However, a result similar to Theorem 2.1 may be proved by combining martingale and empirical process arguments, provided that $\hat{G}_{\theta,n}$ 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 $\hat{G}_{\theta,n}$ is a symmetric statistic of the observations $X_1, \ldots X_n$ and that it converges 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 which we expect. This implies that $l_{n,\mathrm{E}}(\theta)$ is close to a sequence of likelihood ratios.

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

$$\mathrm{E}_{P_{\theta,G}}\tilde{T}^{(1)}(X_i,\, P_{\theta,\hat{G}_{\theta,n}}) = o(n^{-1/2}).$$

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

The last condition implies a uniform control of the approximation of $\tilde{T}^{(1)}(\cdot,\, P_{\theta,G})$ by $\tilde{T}^{(1)}(\cdot,\, P_{\theta,\hat{G}_{\theta,n}})$.

H_4 For every θ and n, the functions $\tilde{T}^{(1)}(\cdot,\, P_{\theta,\hat{G}_{\theta,n}})$ belong to a Donsker class of functions with probability one. The class has an envelope $H(\cdot) > 0$, possibly dependent on θ, with

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

Actually these conditions are weaker than the conditions that one usually assumes in the framework of semiparametric models – see, for instance, Bickel *et al.* (1993) or van der Vaart (1998, 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 $\hat{G}_{\theta,n}$ which is a symmetric function of the observations allows us to weaken the usual hypotheses thanks to extended 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 neighbourhood of the true value θ) close to those assumed by van der Vaart (1998) seem to be needed.

Theorem 3.1. *Assume that hypotheses H_1–H_4 hold. If* $\mathrm{var}(\tilde{T}^{(1)}(X,\, P_{\theta,G}))$ *is of rank q,*

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

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

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

Remark 3.2. Another way to prove this result is to consider the sequence of approximate least favourable models (notice the recentring factor which ensures that we have a density, for λ in the neighbourhood of 0)

$$p_{\lambda,n}(\cdot) = \frac{\mathrm{d}P}{\mathrm{d}\mu}(\cdot)[1 + \lambda'(\tilde{T}^{(1)}(\cdot,\ P_{\theta,\hat{G}_{\theta,n}}) - \mathrm{E}_{P_{\theta,G}}\tilde{T}^{(1)}(X_1,\ P_{\theta,\hat{G}_{\theta,n}}))]$$

$$\mathbb{I}_{\{1+\lambda'(\tilde{T}^{(1)}(\cdot,P_{\theta,\hat{G}_{\theta,n}})-\mathrm{E}_{P_{\theta,G}}\tilde{T}^{(1)}(X_1,P_{\theta,\hat{G}_{\theta,n}}))>0\}}.$$

Even if it may be possible to check the quadratic differentiability, conditions which ensure that the MLE of λ in this family behaves well in the presence of the estimated parameter $\hat{G}_{\theta,n}$ may be more difficult to check. However, it is interesting to see that for the Bartlett correctability of the approximate empirical likelihood (15) to hold, the behaviour of $\mathrm{E}_{P_{\theta,G}}\tilde{T}^{(1)}(X_1,\ P_{\theta,\hat{G}_{\theta,n}})$ is of great importance. In many situations, for instance convex models (see Bickel *et al.* 1993), we have

$$\mathrm{E}_{P_{\theta,G}}\tilde{T}^{(1)}(X_i,\ P_{\theta,\hat{G}_{\theta,n}}) = 0, \tag{17}$$

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) and Chen and Hall (1993) for some examples.

Remark 3.3. Although the splitting trick used to construct (16) 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 envelope). If we assume instead

$$H_5 \quad \mathrm{E}_{P_{\theta,G}}\|\tilde{T}^{(1)}(X_1,\ P_{\theta,\hat{G}^{(i)}_{\theta,n}}) - \mathrm{E}_{P_{\theta,G}}\tilde{T}^{(1)}(X,\ P_{\theta,G})\|^2 \to 0, \qquad \text{as } n \to \infty, \text{ for } i = 1, 2,$$

then the result of Theorem 3.1 still holds. Indeed, the Donsker property is only needed to show the uniformity (26) in the proof. To obtain a similar theorem for (16), we have to check that

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

which is a consequence of H_5. In some models, condition H_5 may be easier to check than the Donsker property.

4. Examples

Most of the examples of Hadamard differentiable functionals considered in Pons and Turckheim (1991) and van der Vaart (1998) can be studied within the framework of Section 2. There is nothing really new in detailing these examples. Provided that we consider finite-

dimensional parameters, which are Hadamard differentiable with smooth gradients (this includes constructing confidence intervals at several points of hazard rates with censored data, or copulas at some given points, which have applications in finance), $\Re^T_{n,1-\alpha} = \{T(\tilde{P}_n),\ \tilde{P}_n \in \overline{\mathcal{P}_{n,1-\alpha}}\}$ is asymptotically a valid confidence region. We therefore illustrate our results and remarks with some examples taken from the semiparametric literature. In Examples 1 and 2, the efficient influence function is known (up to a unidimensional parameter in Example 1). These two examples show that the first and second (nonparametric and semiparametric) approaches are not equivalent and may be more or less adequate according to the problem at hand. Examples 3 and 4 propose a way to handle nuisance parameters in semiparametric models. Many semiparametric problems quoted as challenging in Chapter 10 of Owen (2001) are actually semiparametric problems which may be treated as in Examples 3 and 4. Finally, we show in Example 5 how extensions and Bartlett correctability may be obtained quite straightforwardly in the case of bias sampling models studied by Qin (1993).

The following example illustrates the two different points of view taken in this paper, that is, the nonparametric and semiparametric approaches. It shows that, when there is a finite number of constraints (with no nuisance parameter), it is easier to use the arguments of Section 2 to obtain third-order correct confidence intervals. The efficient influence function is actually computed internally by the program. In that case estimating the efficient influence function is not really necessary.

Example 1 *Third-order correct confidence interval for a P-constrained mean: a comparison of the nonparametric and semi-parametric approaches.* Consider Example 3 of Bickel *et al.* (1993: 68), in which one is interested in estimating the mean $\theta = \mathrm{E}_P X \neq 0$ 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 $\mathrm{E}_P X^8 < \infty$ and $\gamma(P) = \mathrm{E}_P X^2 - (1 + c_0)(\mathrm{E}_P X)^2 = 0,\ c_0 \neq 0\}$. Let $\gamma^{(1)}(\cdot, P)$ be the influence function of $\gamma(\cdot)$ at P. By a straightforward calculus, this 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 $\{h \in L^2(P), \mathrm{E}_P h = 0$ and $\mathrm{E}_P h\gamma^{(1)} = 0\}$ is expressed as

$$\tilde{T}^{(1)}(x, P) = T^{(1)}(x, P) - \frac{\mathrm{cov}_P(T^{(1)}(X, P)\gamma^{(1)}(X, P))}{\mathrm{var}_P(\gamma^{(1)}(X, P))}\gamma^{(1)}(x, P) \tag{18}$$

(this is simply the residual of the regression of $T^{(1)}$ on $\gamma^{(1)}$; see Bickel *et al.* (1993: 55), with variance

$$V_P\tilde{T}^{(1)}(X, P) = V_P T^{(1)}(X, P) - \frac{\mathrm{cov}_P(T^{(1)}(X, P)\gamma^{(1)}(X, P))^2}{\mathrm{var}_P(\gamma^{(1)}(X, P))}.$$

However, the regression coefficient

$$\alpha = \frac{\mathrm{cov}_P(T^{(1)}(X, P)\gamma^{(1)}(X, P))}{\mathrm{var}_P(\gamma^{(1)}(X, P))}$$

is unknown and must be estimated, for instance, by

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

which is a symmetric function of the observations. However, $\tilde{T}^{(1)}(x, P)$ is clearly continuous in α (which plays the role of the nuisance parameter G) and we have $\hat{\alpha}(\theta) \to \alpha(\theta)$ almost surely. We may then use the 'estimated' estimating function

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

It is easy to check that

$$\mathrm{E}_P(T^{(1)}(X_i, \theta) - \hat{\alpha}(\theta)\gamma^{(1)}(X_i, \theta)) = O(n^{-1}).$$

H_4 is satisfied with an envelope given by

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

Moreover, when $\mathrm{E}X^8 < \infty$, we have

$$\mathrm{E}_P\big\{T^{(1)}(X_i, \theta) - \hat{\alpha}(\theta)\gamma^{(1)}(X_i, \theta) - T^{(1)}(X_i, \theta) - \alpha(\theta)\gamma^{(1)}(X_i, \theta)\big\}^2$$

$$\leqslant \big(\mathrm{E}_P\gamma^{(1)}(X_i, \theta)^4\big)^{1/2}\big(\mathrm{E}_P(\hat{\alpha}(\theta) - \alpha(\theta))^4\big)^{1/2} \to 0 \quad \text{as } n \to \infty,$$

so that H_5 is satisfied. We may then apply Theorem 3.1 to obtain empirical likelihood based confidence intervals. However, in this case, it is simpler to consider this semiparametric model as a problem in which there are two estimating functions corresponding respectively to $\mathrm{E}_Q(T^{(1)}(X, P)) = 0$ and $\mathrm{E}_Q\, \gamma^{(1)}(X, P) = 0$. 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

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

straightforward calculus based on Taylor expansions (see Remark 2.3) yields

$$\Omega\begin{pmatrix}\hat{\lambda}\\ \hat{\mu}\end{pmatrix} = \begin{pmatrix}\dfrac{1}{n}\displaystyle\sum_{i=1}^{n} T^{(1)}(X_i, P)\\[2ex] \dfrac{1}{n}\displaystyle\sum_{i=1}^{n} \gamma^{(1)}(X_i, P)\end{pmatrix} + o_P(1),$$

with

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

that is,

$$\begin{pmatrix} \hat{\lambda} \\ \hat{\mu} \end{pmatrix} = \begin{pmatrix} \dfrac{1}{n}\displaystyle\sum_{i=1}^{n} \tilde{T}^{(1)}(X_i, P)/V_P(\tilde{T}^{(1)}(X_i, P)) \\ \dfrac{1}{n}\displaystyle\sum_{i=1}^{n} \tilde{\gamma}^{(1)}(X_i, P)/V_P(\tilde{\gamma}^{(1)}(X_i, P)) \end{pmatrix},$$

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

$$\tilde{\gamma}^{(1)}(x, P) = \gamma^{(1)}(x, P) - \frac{\operatorname{cov}_P(T^{(1)}(X, P)\gamma^{(1)}(X, P))}{\operatorname{var}_P(T^{(1)}(X, P))} T^{(1)}(x, P).$$

This is the efficient influence function when estimating $\gamma(P)$ with a known mean θ. Thus the MLE of λ is exactly proportional to the efficient estimating function given by (18). As a by-product, this suggests that we may use the solution of the estimated Kuhn–Tucker coefficient seen as a function of the parameter θ, $\hat{\lambda} = \hat{\lambda}(\theta)$, to obtain an efficient estimator of θ by solving $\hat{\lambda}(\theta) = 0$ (which may be done practically by discretizing $\hat{\lambda}(\theta)$), without any preliminary estimation (of α) as in the first method.

Moreover, the likelihood ratio behaves as

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

which is the usual general method of moments objective function (Hansen 1982). Because of the likelihood structure of the least favourable family parametrized by μ and λ, we may modify Theorem 2.1 straightforwardly and obtain the Bartlett correctability of the empirical likelihood. Notice, however, that the two methods – on the 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 in some sense amounts to estimating Ω^{-1}, which is actually internally computed in the second method.

Conversely, when the model is semiparametric, there may be some clear advantage in computing the efficient influence function (especially if it does not depend on the nuisance parameter). In the following example dealing with mixture models (see Bickel *et al*. 1993, pp. 126–133; Amari and Kawanabe 1997) the mixing distribution g introduces an infinite-dimensional nuisance parameter, so that the classical empirical likelihood approach cannot be directly implemented: we would need to write an infinite number of moment constraints (just as in the examples given in Owen 2001, Chapter 10). The approach of Section 3.1 is easier to implement, all the more so as the efficient influence function does not depend on the nuisance parameter, so that the second part of Theorem 1 also applies and leads to Bartlett correctability.

Example 2 *Mixture models.* Let $g(\eta)$ be an unknown positive density on $\mathbb{R}$ and let $\{f(x, \theta, \eta), \theta \in \mathbb{R}, \eta \in \mathbb{R}\}$ 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 dependent on η, differentiable in θ such that $\partial T_1(\cdot, \theta)/\partial\theta$ is a function of T_1.

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

$$p(x, \theta) = \int f(x, \theta, \eta)g(\eta)\mathrm{d}\eta;$$

then the efficient influence function of θ is given by

$$T^{(1)}(X, \theta, P) = \frac{\partial T_2(X, \theta)}{\partial\theta} - \mathrm{E}_P\left\{\frac{\partial T_2(X, \theta)}{\partial\theta}|T_1(X, \theta)\right\}$$

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). In some models $E_P\{(\partial T_2(X, \theta)/\partial\theta)|T_1(X, \theta)\}$ may be calculated explicitly (see Bickel *et al.* 1993, Section 4.5). In other general models, this quantity may be seen as a nuisance parameter that may also be estimated by some kernel smoothing method (see Example 4 below).

The following example is considered in Chapter 10 of Owen (2001) and illustrates the fact that a direct approach of empirical likelihood is not possible here. The result follows directly from an application of Theorem 3.1.

Example 3 *Confidence region for the centre of symmetry of a semiparametric family.* Assume that the model is given by $\varrho_{\theta,G} = \{P_{\theta,\eta} \ll \mu$ (any dominating measure) with $\mathrm{d}P_{\theta,\eta}/\mathrm{d}\mu = \eta(x - \theta)$ 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 $\{\dot{\eta}(x - \theta)^2/\eta(x - \theta), \eta \in G\}$ is a Donsker class (see 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

$$-I(\eta)^{-1}\frac{\dot{\eta}(x - \theta)}{\eta(x - \theta)} \qquad \text{with } I(\eta) = \int \frac{\dot{\eta}(x - \theta)^2}{\eta(x - \theta)}\mathrm{d}x.$$

Let $\hat{\eta}_\theta$ be a symmetrized kernel density estimator of η based on the recentred observations $\{X_i - \theta, \ -X_i - \theta\}$. Consider, for instance, the construction in van der Vaart (1998: 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

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

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

The following example illustrates the fact that even if we do not have an efficient influence function but the parameter is the solution of some moment equations, with some nuisance parameter, it is still possible to use Theorem 3.1 to construct empirical likelihood based confidence region.

Example 4 *M-estimators with nuisance parameters.* In many applications and many econometric models, the parameter of interest is the solution of an equation of the form

$$\mathrm{E}_P m(X, \theta, g) = 0,$$

where m is a function from (χ, Θ, H) to $\mathbb{R}^k$, $\Theta \subset \mathbb{R}^k$, and g belongs to H which is of infinite dimension. We assume for the sake of simplicity that $m(\cdot, \theta, g)$ is C^1 in θ, continuous in g and bounded (but less restrictive hypotheses may be considered on the class of functions $m(\cdot, \theta, g)$, indexed by θ and g, to cover specific applications). Typically, in many applications, g is a density or an unknown regression function. Then the problem reduces to exhibiting an estimator of g (possibly dependent on θ) which is invariant, by permutation of the X_i. In many applications, a kernel estimator or a Nadaraya–Watson estimator (possibly dependent on θ as in Example 3) of g may be used. To check H_2 and control the bias, it may be necessary to use higher-order kernels, a technique which is now standard. The result of Theorem 3.1 can then be applied directly. There is actually no need to compute an influence function here. To recast this example in our framework, simply notice that the parameter $\theta = T(P)$ can be extended to the entire space $\mathcal{P}$. This is precisely the principle of generalized estimating equations. A (generally) non-efficient influence function for this parameter is given by

$$\left(-\mathrm{E}_P\left(\frac{\partial m(X, \theta, g)}{\partial \theta}\right)\right)^{-1} m(X, \theta, g),$$

which is proportional to $m(X, \theta, g)$. Of course there is no reason for the procedure to be efficient, but it is easy to implement.

Our final example shows how the theoretical arguments of the preceding section (mainly convex duality theory) can be extended in specific situations by interpreting empirical likelihood ratio as an adequate dual likelihood ratio.

Example 5 *Empirical likelihood in biased sampling model, revisited.* We refer to Chapter 6 of Owen (2001) for complete references and give brief 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 non-negative weight functions w_i:

$$Q_i(\mathrm{d}y) = \frac{w_i(y)}{W_i(P)} P(\mathrm{d}y),$$

$$W_i(P) = \int w_i(y) P(\mathrm{d}y), \qquad i = 1, \ldots, s.$$

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

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

and denote by $n = \sum_{i=1}^{s} n_i$ 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 NPMLE for P. Let us give some specific cases.

Case 1: Stratified sampling. Let X be a random variable taking values in $\mathbb{R}^k$. Let $S_1, S_2, \ldots, S_s$ be a partition of the space: $\cup_{i=1}^{s} S_i = \mathbb{R}^k$, $S_j \cap S_i = \varnothing$. Then the weight functions are $w_i(x) = I_{S_i}\{x\}$, where $I_A\{\cdot\}$ is the indicator of set A. It is known that, unless auxiliary (transverse) information is available, the probability P is not identifiable.

Case 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 which case $S_1 = \mathbb{R}^k$ and $S_2, \ldots, S_s$ do not form a partition and we have simply $w_1(x) = 1$. It is generally assumed that the biasing scheme, that is, the w_i, are known. Then the likelihood of the data is given by

$$L_n(P, \mu) = \prod_{i=1}^{s} \prod_{j=1}^{n_i} \frac{\mathrm{d}Q_i}{\mathrm{d}\mu}(X_{j,i}) = \prod_{i=1}^{s} \prod_{j=1}^{n_i} \frac{w_i(X_{j,i})}{W_i(P)} \frac{\mathrm{d}P}{\mathrm{d}\mu}(X_{j,i}). \tag{19}$$

Case 3: Length-biased sampling. It sometimes happens 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) and Gill *et al.* (1988) have given conditions for the identifiability of P and for the existence and uniqueness of the 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) for case 2 (enriched sample with $s = 2$). We think that it is easier to understand his work within 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 (locally asymptotically normal) least favourable family is constructed explicitly. The empirical likelihood in a biased sampling model evaluated at θ is defined here similarly to (19) by considering only probability dominated by P_n,

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

$$= \sup_{\substack{p_{j,i,n} \\ i=1,\dots,s \\ j=1,\dots,n_i}} \left\{ \prod_{i=1}^{s}\prod_{j=1}^{n_i} \frac{w_i(X_{j,i})}{\sum_{k=1}^{s}\sum_{l=1}^{n_i} w_i(X_{l,k}) p_{l,k,n}} np_{j,i,n},\ T\left(\sum_{i=1}^{s}\sum_{j=1}^{n_i} p_{j,i,n}\delta_{X_{j,i}}\right) = \theta, \right.$$

$$\left. p_{j,i,n} > 0, \sum_{k=1}^{s}\sum_{l=1}^{n_i} p_{j,i,n} = 1 \right\}$$

$$= \sup_{\substack{p_{j,i,n}, W_i \\ i=1,\dots,s \\ j=1,\dots,n_i}} \left\{ \prod_{i=1}^{s}\prod_{j=1}^{n_i} \frac{w_i(X_{j,i})}{W_i} np_{j,i,n},\ T\left(\sum_{i=1}^{s}\sum_{j=1}^{n_i} p_{j,i,n}\delta_{X_{j,i}}\right) = \theta,\ p_{j,i,n} > 0, \right.$$

$$\left. \sum_{k=1}^{s}\sum_{l=1}^{n_i} p_{j,i,n} = 1, \sum_{k=1}^{s}\sum_{l=1}^{n_i} w_i(X_{l,k}) p_{l,k,n} = W_i \right\},$$

which we approximate by the linearized version

$$\tilde{L}_{w,n}(\theta) = \sup_{\substack{p_{j,i,n}, W_i \\ i=1,\dots,s \\ j=1,\dots,n_i}} \left\{ \prod_{i=1}^{s}\prod_{j=1}^{n_i} \frac{w_i(X_{j,i})}{W_i} np_{j,i,n},\ \sum_{i=1}^{s}\sum_{j=1}^{n_i} p_{j,i,n} T^{(1)}(X_{j,i}, \theta) = 0,\ p_{j,i,n} \geqslant 0, \right.$$

$$\left. \sum_{k=1}^{s}\sum_{l=1}^{n_i} p_{j,i,n} = 1, \sum_{k=1}^{s}\sum_{l=1}^{n_i} w_i(X_{l,k}) p_{l,k,n} = W_i \right\}.$$

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 condition (see Vardi 1985; Owen 2001), which ensures the existence of an NPMLE. This condition essentially means that we are not in the situation of case 1, that is, we have some transverse information or 'linking' observations.

(H_1) For every proper subset B of $\{1, \dots, s\}$, there exists at least one point X^* in $\cup_{i\in B}\{X_{1,i}, \dots, X_{n_i,i}\}$, for which we have $w_j(X^*) > 0$, for some j in B^c.

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 that the sampling bias is

not proportional to $T^{(1)}(X, P)$. This hypothesis thus excludes case 3, when $T^{(1)}(x, P) = x - \theta$ and $w(x) = x$. This case can be considered by itself after minor modifications.

$$(H_2) \quad \operatorname{var}\begin{pmatrix} T^{(1)}(X, P) \\ w(X) \end{pmatrix} \text{ is of rank } q + s.$$

Under condition (H_1), the value of the empirical likelihood calculated at Vardi's nonparametric maximum likelihood is

$$L_{w,n} = \sup_{\substack{p_{j,i,n}, W_i \\ i=1,\ldots,s \\ j=1,\ldots,n_i}} \left\{ \prod_{i=1}^{s} \prod_{j=1}^{n_i} \frac{w_i(X_{j,i})}{W_i} p_{j,i,n},\ p_{j,i,n} \geqslant 0,\ \sum_{i=1}^{n} p_{j,i,n} = 1,\ \sum_{k=1}^{s} \sum_{l=1}^{n_i} w_i(X_{l,k}) p_{l,k,n} = W_i \right\}$$

$$= \sup_{Q} \left\{ L_n(W, P_n),\ Q \ll P_n,\ \int \mathrm{d}Q = 1 \right\}.$$

The NPMLE $P_{w,n}$ of P is precisely the solution of this unconstrained empirical likelihood. The empirical log-likelihood ratio at θ is then

$$R_{E,w,n}(\theta) = \frac{\tilde{L}_{w,n}(\theta)}{L_{w,n}}.$$

Define $w(x) = (w_1(x), \ldots, w_s(x))$ and $W = (W_1, \ldots, W_s)$. As in Section 2, we may now define the least favourable model:

$$p_{\lambda,\gamma,W}(x) \tag{20}$$

$$= \frac{\mathrm{d}P}{\mathrm{d}\mu}(x)(1 + \lambda' T^{(1)}(x, \theta) + \gamma'(w(x) - W))\mathbb{I}\{1 + \lambda' T^{(1)}(x, \theta) + \gamma'(w(x) - W) > 0\}$$

$$= \frac{\mathrm{d}Q_i}{\mathrm{d}\mu}(x)(1 + \lambda' T^{(1)}(x, \theta) + \gamma'(w(x) - W)) \frac{W_i}{w_i(x)} \mathbb{I}\{x \in S_i\}$$

$$\times \mathbb{I}\{1 + \lambda' T^{(1)}(x, \theta) + \gamma'(w(x) - W) > 0\},$$

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 Section 1 (used twice) imply that the empirical likelihood ratio is

[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.

$$-2\log(R_{E,w,n}(\theta)) \tag{21}$$

$$= 2\left(\sup_{W,\gamma}\left(\sum_{k=1}^{s}\sum_{l=1}^{n_k}\log(1+\gamma'(w(X_{l,k})-W))+\sum_{k=1}^{s}n_k\log(W_k)\right)\right.$$

$$\left.-\sup_{W,\lambda,\gamma}\left(\sum_{k=1}^{s}\sum_{l=1}^{n_i}\log(1+\lambda' T^{(1)}(X_{l,k},\theta)+\gamma'(w(X_{l,k})-W)+\sum_{k=1}^{s}n_k\log(W_k)\right)\right).$$

This is exactly the log-likelihood ratio for testing $\lambda = 0$ in model (20); compare with Qin (1993). Now under (H_1) and (H_2), (20) is quadratically differentiable (using the same arguments as in Theorem 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). Hence using the same argument as Mykland (1995), (21) is asymptotically $\chi^2(q)$, yielding an asymptotically $(1-\alpha)$ confidence region

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

Under additional moments on $(T^{(1)}(X,\theta), w(X))$, Bartlett correctability also follows from representation (21) seen as a log-likelihood ratio for testing $\lambda = 0$ in the family (20).

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 let $p_* = \min_{i=1,\ldots,n}(p_{i,n}) \leqslant 1/n \leqslant \max_{i=1,..,n}(p_{i,n}) = p^*$. Consider j such that $p_* = p_{j,n}$. Then the constraint on the likelihood implies

$$r_0 \leqslant \frac{p_*\Pi^n_{i=1,i\neq j}\,p_{i,n}}{(1/n)^n} \leqslant \frac{p_*\max\Pi^n_{i=1,i\neq j}\,p_{i,n}}{(1/n)^n}, \qquad \text{with } \sum_{i\neq j}p_{i,n} = 1-p_* \tag{22}$$

$$= np_*(1-p_*)^{n-1}(n/(n-1))^{n-1}.$$

Because $(n/(n-1))^{n-1}$ is a sequence that converges to e from below, this yields the inequality

$$\frac{1}{n}\frac{r_0}{\mathrm{e}} \leqslant p_* \leqslant \frac{1}{n}.$$

Now we have that

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

where

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

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

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

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

which implies

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

and, in particular,

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

Notice that when $\alpha \to 1$, the bound converges to $1/n$, that is, at the limit, all the $p_{i,n}$s are equal to $1/n$.

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

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

since each $p_{i,n}$ is of order $1/n$ by the first part of the lemma. Since the X_i are i.i.d. and $\mathcal{F}$ is Donsker and satisfies the uniform entropy condition (7), it follows – see van der Vaart and Wellner (1996: 210) or Koul (1992, Theorem 2.2) for the real multidimensional case – that

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

where G_P is a Gaussian process with covariance operator independent of the weights. Now for $p_n = (p_{1,n}, \ldots, p_{n,n})$ 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

$$\lim_{\delta \to 0} \limsup_{n \to \infty} \sup_{p_n \Subset \overline{\mathcal{P}_{n,1-\alpha}}} \mathrm{P}\left(\sup_{\|f-g\|_{2,P} < \delta} |G_{n,p_n}(f) - G_{n,p_n}(g)| > \varepsilon \right) \to 0,$$

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 exist some non-negative constants A and B such that, for any $p_n \Subset \overline{\mathcal{P}_{n,1-\alpha}}$,

$$A\|f-g\|_{2,P_n}^2 \leqslant \|f-g\|_{2,p_n}^2 \leqslant B\|f-g\|_{2,P_n}^2. \tag{23}$$

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

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

this is a measurable class of functions by the 'Suslin hypothesis'.

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

$$\begin{aligned}\Delta_n &= P\left(\sup_{\|f-g\|_{2,p_n}<\delta}|G_{n,p_n}(f)-G_{n,p_n}(g)|>\varepsilon\right)\\ &\leqslant C\mathrm{E}_P\int_0^{\theta_n/\|H\|_{2,p_n}}\sqrt{\log(N(\varepsilon\|H\|_{2,p_n},\mathcal{F}_{\delta,P},\|\cdot\|_{2,p_n}^2))}\mathrm{d}\varepsilon\ \|H\|_{2,p_n},\end{aligned}$$

with

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

Now (23) implies that there exists a constant C such that for all $p_n \Subset \overline{\mathcal{P}_{n,1-\alpha}}$,

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

Since we have $\|H\|_{2,P_n}\geqslant 1$ and $\mathrm{E}_P\|H\|_{2,P_n}^2 = \mathrm{E}_P H^2$, it follows, by the Cauchy–Schwarz inequality, that

$$\begin{aligned}\Delta_n^2 &\leqslant C_1\mathrm{E}_P\left(\int_0^{\theta_n'/A}\sqrt{\sup_Q\log(N(\varepsilon\|H\|_{2,Q},\mathcal{F}_{\delta,P},\|\cdot\|_{2,Q}))}\mathrm{d}\varepsilon\ \|H\|_{2,P_n}\right) \qquad (24)\\ &\leqslant C_2\left(\mathrm{E}_P\left(\int_0^{\theta_n'/A}\sqrt{\sup_Q\log(N(\varepsilon\|H\|_{2,Q},\mathcal{F}_{\delta,P},\|\cdot\|_{2,Q}))}\mathrm{d}\varepsilon\right)^2\right)^{1/2}\left(\mathrm{E}_P H^2\right)^{1/2}\\ &\leqslant C_3\left(\int_0^{\eta}\sqrt{\sup_Q\log(N(\varepsilon\|H\|_{2,Q},\mathcal{F}_{\delta,P},\|\cdot\|_{2,Q}))}\mathrm{d}\varepsilon + P(\theta_n^*/A>\eta)\right)^{1/2}\left(\mathrm{E}_P H^2\right)^{1/2}.\end{aligned}$$

Under the uniform entropy condition, the right-hand side of (24) does not depend on p_n and may be made as small as we wish provided that $\theta_n^*\to 0$. This is a consequence of

Theorem 2.5.2 in van der Vaart and Wellner (1996) and follows from the uniform law of large numbers 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 (24) yields the result.

5.2. Proof of Theorem 3.1

The following lemma and its short proof are taken from Bertail and Lo (1996). This result may also be useful in semiparametric applications (when one wishes to avoid the splitting trick).

Lemma 5.1. *Assume $X_1, X_2, \ldots, X_n$ 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) $\|\omega(x, t)\| \leqslant H(x)$ *with* $\mathrm{E}H(X) < \infty$ *and* (ii) $\omega(x, t)$ *is continuous in t. Then* $G_n \xrightarrow{a.s.} G$ *implies that*

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

Proof. It is sufficient to write

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

where $\mathcal{S}^n$ is the symmetric field containing all the symmetric functions of $X_1, X_2, \ldots, X_n$. By the extended backward martingale convergence of Blackwell and Dubins (1962), S_n^{ω} converges with probability one to $\mathrm{E}(\omega(X_1, G)|\mathcal{S}^{\infty})$. But by the Hewitt–Savage zero–one law, $\mathcal{S}^{\infty}$ is non-trivial and therefore $\mathrm{E}(\omega(X_1, G)|\mathcal{S}^{\infty})$ is a constant equal to $\mathrm{E}(\omega(X_1, G))$. □

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

Lemma 5.2. *Under H_1 and H_3,*

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

with

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

Proof. Apply Lemma 5.1 with $\omega(X_i, G) = \tilde{T}^{(1)}(X_i, P_{\theta,G})'\tilde{T}^{(1)}(X_i, P_{\theta,G})$. Under H_3, $\|\omega(X_i, G)\| \leqslant H(X)^2$. Since $P_{\theta,\hat{G}_{\theta,n}}$ is symmetric in the observations and $\mathrm{E}H(X)^2 < \infty$, $I_n \to I(\theta, G)$ as $n \to \infty$. □

We also need the following useful and straightforward result, which may be found in Le

Cam (1986: 188). It allows us 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 random variables 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

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

Proof. Taylor expansion. □

Proof of Theorem 3.1. The proof is now along the same lines as Owen (1990). Notice first that by Lemma 5.1, for each fixed λ, $L_n(\lambda) = n^{-1}\sum_{i=1}^n \log(1 + \lambda' \tilde{T}^{(1)}(X_i, P_{\theta,\hat{G}_{\theta,n}}))$ converges to

$$\mathrm{E}_{P_{\theta,G}} \log\big(1 + \lambda' \tilde{T}^{(1)}(X, P_{\theta,G})\big) \leqslant \log\big(1 + \lambda' \mathrm{E}_{P_{\theta,G}} \tilde{T}^{(1)}(X, P_{\theta,G})\big) = 0,$$

by the 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 $\hat{\lambda}$, which is the unique solution of the equation

$$\frac{1}{n}\sum_{i=1}^n \frac{\tilde{T}^{(1)}(X_i, P_{\theta,\hat{G}_{\theta,n}})}{1 + \lambda' \tilde{T}^{(1)}(X_i, P_{\theta,\hat{G}_{\theta,n}})} = 0. \tag{25}$$

Since $\sup_n(\mathrm{E}_{P_{\theta,G}} \|\tilde{T}^{(1)}(X_i, P_{\theta,\hat{G}_{\theta,n}})\|^2) < \mathrm{E}_{P_{\theta,G}} H^2(X) < \infty$, following Owen (2001: 220), we obtain $\hat{\lambda} = O_P(n^{-1/2})$ (using his arguments as well as Lemma 5.2 to control the moments uniformly) and we have by direct Taylor expansion of (25),

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

Under H_2 and H_4, we obtain that

$$n^{-1/2}\sum_{i=1}^n \tilde{T}^{(1)}(X_i, P_{\theta,G}) - n^{-1/2}\sum_{i=1}^n \tilde{T}^{(1)}(X_i, P_{\theta,\hat{G}_{\theta,n}}) = o_P(1). \tag{26}$$

Equation (26) and Lemma 5.2 imply that

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

Now put $Y_{i,n} = \hat{\lambda}'\tilde{\mathrm{T}}^{(1)}(\mathrm{X_i}, \mathrm{P}_{\theta,\hat{\mathrm{G}}_{\theta,\mathrm{n}}})$. Then we can check that

$$\max_{1\leqslant i\leqslant n}(Y_{i,n}) = \max_{1\leqslant i\leqslant n}(\tilde{T}^{(1)}(X_i, P_{\theta,\hat{G}_{\theta,n}}))\ O_P(n^{-1/2}) = o_P(1)$$

and, using Lemma 5.2,

$$\sum_{i=1}^{n} Y_{i,n}^2 = n^{1/2}\hat{\lambda}'\left(n^{-1}\sum_{i=1}^{n}\tilde{T}^{(1)}(X_i, P_{\theta,\hat{G}_{\theta,n}})\tilde{T}^{(1)}(X_i, P_{\theta,\hat{G}_{\theta,n}})'\right)n^{1/2}\hat{\lambda} = O_P(1).$$

Thus applying Lemma 5.3 with $\phi(x) = \log(1+x)$ and using Lemma 5.2, we obtain

$$\sum_{i=1}^{n}\log\left(1+\hat{\lambda}'\tilde{T}^{(1)}(X_i, P_{\theta,\hat{G}_{\theta,n}})\right)$$

$$= \left(n^{-1/2}\sum_{i=1}^{n}\tilde{T}^{(1)}(X_i, P_{\theta,G})\right)' I(\theta, G)^{-1}\left(n^{-1/2}\sum_{i=1}^{n}\tilde{T}^{(1)}(X_i, P_{\theta,G})\right) + o_P(1)$$

and the result follows.

Acknowledgements

The first version of this paper, entitled 'Empirical likelihood in i.i.d. and bias sampling problems', was written while I was working at INRA-Corela on a research contract with the 'Observatoire des Consommations Alimentaires'. I would like to thank both institutions, as well as CREST, for their past and current financial support. My gratitude also goes to Emmanuelle Gautherat, Université de Reims and CREST-LS, to Judith Rousseau, Université Paris 9 and CREST-LS, to France Caillavet, INRA-Corela, and to Albert Lo, HKUST, for their careful reading of the manuscript and their comments, which greatly improved the results and the presentation of this paper. I would also like to thank a referee, an associate editor as well as the copy editor of *Bernoulli*, for all their suggestions and corrections. All errors remaining are my sole responsibility.

References

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

Baggerly, K.A. (1998) Empirical likelihood as a goodness-of-fit measure. *Biometrika*, **85**, 535–547

Barbe, Ph. and Bertail, P. (1995) *The Weighted Bootstrap*, Lecture Notes in Statist. 98. Berlin: Springer-Verlag.

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

Bertail P. (2004) Empirical likelihood in some nonparametric and semiparametric models. In M.S. Nikulin, N. Balakrishnan, M. Mesbah, and N. Limnios (eds), *Parametric and Semiparametric Models with Applications to Reliability, Survival Analysis, and Quality of Life*. Boston: Birkhäuser.

Bertail, P. and Lo, A. (1996) Accurate posterior approximations. Preprint, INRA-Corela.

Bickel, P.J. and Ghosh, J.K. (1990) A decomposition for the likelihood ratio statistic and the Bartlett correction – A Bayesian argument. *Ann. Statist.*, **18**, 1070–1090.

Bickel, P.J., Klaasen, C.A.J., Ritov, Y. and Wellner, J.A. (1993) *Efficient and Adaptive Estimation for Semiparametric Models*. Baltimore, MD: Johns Hopkins University Press.

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

Borwein, J.M. and Lewis, A.S. (1991) Duality relationships for entropy-like minimization problem. *SIAM J. Control Optim*., **29**, 325–338.

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

Chen, S.X. and Cui, H. (2003) An extended empirical likelihood for generalized linear models. *Statist. Sinica,* **13**, 69–81.

Chen, S.X. and Hall, P. (1993) Smoothed empirical likelihood confidence intervals for quantiles. *Ann. Statist.*, **21**, 1166–1181.

Corcoran, S. (1998) Bartlett adjustment of empirical discrepancy statistics. *Biometrika*, **85**, 967–972.

DiCiccio, T. and Romano, J. (1990). Nonparametric confidence limits by resampling methods and least favourable families. *Internat. Statist. Rev.*, **58**, 59–76.

DiCiccio,T., Hall, P. and Romano, J. (1991) Empirical likelihood is Bartlett correctable. *Ann. Statist.*, **19**, 1053–1061.

Dudley, R.M. (1984) A course on empirical processes. In P.L. Hennequin (ed.), *Ecole d'Été de Probabilités de Saint Flour XII – 1982*, Lecture Notes in Math. 1097, pp. 2–241. Berlin: Springer-Verlag.

Dudley, R.M. (1990) Nonlinear functionals of empirical measures and the bootstrap. In E. Eberlein, J. Kuelbs and M.B. Marcus (eds) *Probability in Banach Spaces 7*, Progr. Probab. 21, pp. 63–82. Boston: Birkhäuser.

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.

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

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. *Internat. 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.

Hansen, L.P. (1982) Large sample properties of generalized method of moments estimators. *Econometrica*, **50**, 1029–1054.

Hartley, H.O. and Rao, J.N.K. (1968) A new estimation theory for sample survey. *Biometrika*, **55**, 547–557.

Koul, H.L. (1992) *Weighted Empiricals and Linear Models*, IMS Lecture Notes, Monogr. Ser. 21. Hayward, CA: Institute of Mathematical Statistics.

Le Cam, L. (1986) *Asymptotic Methods in Statistical Decision Theory.* New York: Springer-Verlag.

Leblanc, M. and Crowley, J. (1995) Semiparametric regression functionals. *J. Amer. Statist. Assoc.*, **90**, 95–105.

Leonard, C. (2001) Minimizers of energy functionals. *Acta Math. Hungar.*, **93**, 281–325.

Liese, F. and Vajda, I. (1987) *Convex Statistical Distances*. Leipzig: Teubner.
Murphy, S.A., and van der Vaart, A.W. (1997) 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**, 237–249.
Owen, A.B. (1990) Empirical likelihood ratio confidence regions. *Ann. Statist.*, **18**, 90–120.
Owen, A.B. (2001) *Empirical Likelihood*. Boca Raton, FL: Chapman & Hall/CRC.
Pons, O. and Turckheim E. (1991) Von Mises method, bootstrap and Hadamard differentiability. *Statistics*, **22**, 205–214.
Qin, G.S. and Jing, B.Y. (2001) Empirical likelihood for censored linear regression. *Scand. J. Statist.*, **28**, 661–673.
Qin, J. (1993). Empirical likelihood in biased sample problems. *Ann. Statist.*, **21**, 1182–1196.
Qin, J. and Lawless, J. (1994) Empirical likelihood and general estimating equations. *Ann. Statist.*, **22**, 300–325.
Rockafeller, R. (1968) Integrals which are convex functionals. *Pacific J. Math.*, **24**, 525–339.
Thomas, D.R. and Grunkemeier, G.L. (1975) Confidence interval estimation of survival probabilities for censored data. *J. Amer. Statist. Assoc.*, **70**, 865–871.
van der Vaart, A.W. (1995) Efficiency of infinite dimensional M-estimators. *Statist. Neerlandica*, **49**, 9–30.
van der Vaart, A.W. (1998) *Asymptotic Statistics*. Cambridge: Cambridge University Press.
van der Vaart, A.W. and Wellner, J.A. (1996) *Weak Convergence and Empirical Process: With Applications to Statistics*. New York: Springer-Verlag.
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.
von Mises, R. (1936) Les lois de probabilités pour les fonctions statistiques. *Ann. Inst. H. Poincaré*, **6**, 185–212.
Wilks, S.S. (1938) The large-sample distribution of the likelihood ratio for testing composite hypotheses. *Ann. Math. Statist.*, **19**, 60–62

Received March 2003 and revised June 2005.