ESAIM: PROCEEDINGS AND SURVEYS, October 2015, Vol. 51, p. 212-231
A. Garivier et al, Editors

EMPIRICAL PHI-DISCREPANCIES AND QUASI-EMPIRICAL LIKELIHOOD: EXPONENTIAL BOUNDS

Patrice Bertail[1], Emmanuelle Gautherat[2] and Hugo Harari-Kermadec[3]

Abstract. We review some recent extensions of the so-called generalized empirical likelihood method, when the Kullback distance is replaced by some general convex divergence. We propose to use, instead of empirical likelihood, some regularized form or quasi-empirical likelihood method, corresponding to a convex combination of Kullback and χ^2 discrepancies. We show that for some adequate choice of the weight in this combination, the corresponding quasi-empirical likelihood is Bartlett-correctable. We also establish some non-asymptotic exponential bounds for the confidence regions obtained by using this method. These bounds are derived via bounds for self-normalized sums in the multivariate case obtained in a previous work by the authors. We also show that this kind of results may be extended to process valued infinite dimensional parameters. In this case some known results about self-normalized processes may be used to control the behavior of generalized empirical likelihood.

AMS 2000 subject classifications. Primary 62G15; Secondary 62E17, 62H15.
Key words and phrases. Empirical likelihood, convex duality, discrepancy, confidence region, Bartlett correction, exponential bound, empirical processes indexed by class of functions

1. Introduction

Empirical likelihood has been introduced and studied by Owen [34, 35], see Owen [36] for a complete overview and important references. The main idea underlying empirical likelihood consists of maximizing a profile likelihood or multinomial likelihood supported by the data, under some linear constraints induced by the model. It can also be seen as an extension of "model based likelihood" used in survey sampling when some marginal constraints are available (see Deville and Sarndal [15], Hartley and Rao [23]). Owen and many followers (see Owen [36]) have shown that one can get a useful and automatic non-parametric version of Wilks' theorem, stating that the log-likelihood ratio convergences to a χ^2 distribution.

Generalizations of empirical likelihood methods are available for many statistical and econometric models as soon as the parameter of interest is defined by some linear moment constraints (see Newey and Smith [33], Qin and Lawless [39]). It can now be considered as an alternative to the generalized method of moments (GMM, see Smith [43]). Moreover just like in the parametric case, this log-likelihood ratio is Bartlett-correctable. This means that an explicit correction leads to confidence regions with third order properties. The asymptotic error on the level is then of order $\mathcal{O}(n^{-2})$ instead of $\mathcal{O}(n^{-1})$ under some regularity assumptions (see Bertail [5], DiCiccio et al. [17]).

[1] MODAL'X, Université Paris-Ouest-Nanterre-La Défense
[2] CREST-LS et Laboratoire REGARDS, Université de Reims Champagne Ardennes
[3] Ecole Normale Supérieure de Cachan

The empirical log-likelihood ratio may also be naturally seen as the minimization of the Kullback divergence between a measure $\mathbb{Q}$, dominated by the empirical distribution of the data $\mathbb{P}_n$, and $\mathbb{P}_n$, under linear or non-linear constraints imposed on $\mathbb{Q}$ by a model to be tested (see Bertail [5], Bertail et al. [9]). Then if the model is true, asymptotically, this distance is zero and the inversion of the test provides a natural confidence region for the parameter of interest. The use of other pseudo-metrics instead of the Kullback divergence K has been suggested by Owen [35] and many other authors. For example, the choice of relative entropy has been investigated by DiCiccio and Romano [16], Jing and Wood [27] and led to "Entropy econometrics" in the econometric field (see Golan et al. [22]). Related results may be found in the probabilistic literature about divergence or the method of entropy in mean (see Broniatowski and Kéziou [11], Csiszár [14], Gamboa and Gassiat [21], Léonard [29], Liese and Vajda [30]). Some generalizations of the empirical likelihood method have also been obtained by using Cressie-Read discrepancies (see Baggerly [2], Corcoran [12]) and led to some econometric extensions known as "generalized empirical likelihood" (see Newey and Smith [33]), even if the "likelihood" properties and in particular the Bartlett-correctability in these cases are lost (see Jing and Wood [27]). Bertail et al. [7] have shown that Owen's original method in the case of the mean can be extended to any regular convex statistical divergence or φ^*-discrepancy (where φ^* is a regular convex function) under weak assumptions (see also Bertail et al. [9]). We also call this method "empirical energy minimizers" by reference to the theoretical probabilistic literature on the subject (see Léonard [29] and references therein).

One goal of this paper is to study a family of discrepancies for which we have a non-asymptotic control of the level of the confidence regions -a lower bound for the coverage probability- for any parameter size, including process valued parameters. The basic idea is to consider a family of divergences consisting in a linear combination (with rate $\varepsilon \in [0,1]$) of the Kullback divergence and the χ^2 divergence, called quasi-Kullback. Then we minimize the dual expression of this divergence under the constraints of the model. It can be seen as a quasi-empirical likelihood or a regularized log-proximal empirical likelihood (see for instance Ausslender et al. [1], for the use of such divergences in the convex literature). The domain of the corresponding divergence is the whole real line making the algorithmic aspects of the problem much more tractable than for empirical likelihood when the number of constraints is large.
Moreover, this approach allows us to keep the interesting properties of both discrepancies. On the one hand, from an asymptotic point of view, we show that this method is still Bartlett-correctable for an adequate choice of ε, typically depending on n. Regions are still automatically shaped by the sample, as in the empirical likelihood case without the limitation stressed by Tsao [44]. On the other hand, for any fixed value of ε, it is possible to use the self-normalizing properties of the empirical divergence to obtain non-asymptotic exponential bounds for the error of the confidence intervals.

The layout of the paper is the following : in part 2, we present the main notation and duality concepts related to empirical energy minimizers or φ^*-discrepancy minimizers. In part 3, we introduce the notion of quasi-Kullback divergences and prove that the corresponding generalized quasi-empirical likelihood problem is Bartlett correctable for an adequate choice of the parameter ε. In part 4, we show that generalized quasi-empirical likelihood may be controlled by the square of a self-normalized sum, allowing for a precise control of the the coverage probability of confidence regions, even for large parameter (with size smaller than $n/log(n)$). This paves the way for studying more general process valued parameters in part 5. A short sample simulation study, showing the advantage (robustness) and weakness of the corresponding confidence regions, is provided in part 6. Last section presents proofs of Theorem 3 and Corollary 1.

2. Empirical φ^*-discrepancy minimizers

Let $(\mathcal{X}, \mathcal{A}, \mathcal{M})$ be a general measurable space endowed with a space of signed measures. Working with signed measures is very important for establishing the existence of solutions for the generalized empirical likelihood problem (see also Bertail et al. [9]). Let f be a measurable function defined from $\mathcal{X}$ to $\mathbb{R}^r$, $r \geq 1$. For any measure $\mu \in \mathcal{M}$, we write $\mu f = \int f d\mu$ and if μ is a density of probability, $\mu f = \mathbb{E}_\mu(f(X))$.

2.1. Notation: φ^*-discrepancies and convex duality

In the following, we use the same notations as in Bertail et al. [9]. Let φ be a convex function. Its support $d(\varphi)$ is defined as $\{x \in \mathbb{R},\ \varphi(x) < \infty\}$. We assume that it is non-void. We denote respectively $\inf d(\varphi)$ and $\sup d(\varphi)$, the extremes of this support. For every convex function φ, the Fenchel-Legendre transform is given by

$$\varphi^*(y) = \sup_{x \in \mathbb{R}} \{xy - \varphi(x)\},\ \forall\ y \in \mathbb{R}.$$

φ^* is then a semi-continuous inferiorly convex function. We denote by $\varphi^{(i)}$ the derivative of order i of φ (when it exists).

The following assumptions for the function φ are classical in the convex literature.

- **H1** φ is strictly convex and $d(\varphi)$ contains a neighborhood of 0 ;
- **H2** φ is twice differentiable on a neighborhood of 0 ;
- **H3** φ is normalized so that $\varphi(0) = 0$, $\varphi^{(1)}(0) = 0$ and $\varphi^{(2)}(0) > 0$.
- **H4** φ is differentiable on $d(\varphi)$, that is to say φ differentiable on $int\{d(\varphi)\}$, with φ' has right and left limits on the respective endpoints of the support of $d(\varphi)$, where $int\{.\}$ is the topological interior.
- **H5** φ is twice differentiable on $d(\varphi) \cap \mathbb{R}^+$ and, on this domain, the second order derivative of φ is bounded from below by some constant $m > 0$.

Under the hypotheses **H1, H2, H3**, the Fenchel dual transform φ^* of φ also satisfies these hypotheses.

The φ^*-discrepancy (see Csiszár [14]) I_{φ^*} between $\mathbb{Q}$ and $\mathbb{P}$, where $\mathbb{Q}$ is a signed measure and $\mathbb{P}$ a positive measure, is defined by

$$I_{\varphi^*}(\mathbb{Q}, \mathbb{P}) = \begin{cases} \int_{\mathcal{X}} \varphi^* \left(\frac{d\mathbb{Q}}{d\mathbb{P}} - 1\right) d\mathbb{P} & \text{if } \mathbb{Q} \ll \mathbb{P} \\ +\infty & \text{else.} \end{cases} \tag{1}$$

Its properties are studied at length for instance by Léonard [28], Liese and Vajda [30], Rockafellar [40, 41, 42], where references may also be found. For general φ^*-discrepancies, the following duality representation is a consequence of Borwein and Lewis [10] on convex functional integrals (see also Broniatowski and Kéziou [11], Léonard [29]).

We denote λ' the transposed vector of λ.

Theorem 1. *Let $\mathbb{P} \in \mathcal{M}$ be a probability measure with a finite support and f be a measurable function on $(\mathcal{X}, \mathcal{A}, \mathcal{M})$. Let φ be a convex function satisfying assumptions **H1-H3**. If the constraints are qualified, that is, if*

$$Qual(\mathbb{P}) : \begin{cases} \exists \mathbb{T} \in \mathcal{M}, \mathbb{T}f = b_0 \text{ and} \\ \inf d(\varphi^*) \ < \ \inf_{\mathcal{X}} \frac{d\mathbb{T}}{d\mathbb{P}} \ \leq \ \sup_{\mathcal{X}} \frac{d\mathbb{T}}{d\mathbb{P}} \ < \ \sup d(\varphi^*) \ \ \mathbb{P} - a.s., \end{cases}$$

then, we have the dual equality

$$\inf_{\{\mathbb{Q} \in \mathcal{M}, (\mathbb{Q} - \mathbb{P})f = b_0\}} \{I_{\varphi^*}(\mathbb{Q}, \mathbb{P})\} = \sup_{\lambda \in \mathbb{R}^r} \left\{ \lambda' b_0 - \int_{\mathcal{X}} \varphi(\lambda' f) d\mathbb{P} \right\}. \tag{2}$$

*If φ also satisfies **H4**, then the supremum on the right hand side of (2) is achieved at a point λ^* and the infimum on the left hand side at $\mathbb{Q}^*$ is given by*

$$\mathbb{Q}^* = (1 + \varphi^{(1)}(\lambda^{*\prime} f))\mathbb{P}.$$

This theorem essentially states that if the constraints are satisfied at least by a measure, which support belongs to the domain of the convex function φ^* then the primal and the dual problem are equal (with no gap). It is fundamental to work not on set of probabilities but rather of measures. This also explains what is called the empty set problem in the empirical likelihood literature (see Tsao [44]).

Let $X_1, ...X_n$ be random vectors defined from a probability space $(\Pr, \mathfrak{A}, \Omega)$ on $\mathcal{X} = \mathbb{R}^p$ with common probability measure $\mathbb{P} \in \mathcal{M}$. We now consider the empirical probability measure $\mathbb{P}_n = \frac{1}{n}\sum_{i=1}^n \delta_{X_i}$, where δ_{X_i} is the Dirac measure at X_i. We remark that, on the set defined by

$$\mathcal{M}_n = \{\mathbb{Q} \in \mathcal{M} \text{ with } \mathbb{Q} \ll \mathbb{P}_n\} = \left\{ \mathbb{Q} = \sum_{i=1}^n q_i \, \delta_{X_i}, \ (q_i)_{1 \le i \le n} \in \mathbb{R}^n \right\},$$

there always exists a signed measure $\mathbb{Q}$ satisfying the first constraint in $Qual(\mathbb{P})$.

2.2. **Empirical optimization of φ^*-discrepancies**

Let $X_1, ...X_n$ be i.i.d. r.v.'s defined on $\mathcal{X} = \mathbb{R}^p$ with common probability measure $\mathbb{P} \in \mathcal{M}$. The parameter of interest $\theta \in \mathbb{R}^q$ is supposed to be the solution of some M-estimation problem $\mathbb{E}_{\mathbb{P}} f(X, \theta) = 0$, where f is a regular differentiable function from $\mathcal{X} \times \mathbb{R}^q \to \mathbb{R}^r$. We assume that f takes its values in $\mathbb{R}^q$, that is $r = q$. The over-identified case $r > q$ can be treated similarly by first reducing the problem to the strictly identified case (see below and Qin and Lawless [39]).
Empirical likelihood and its extensions may actually be seen as a plug-in rule. For a given φ, we define

$$\beta_{\mathbb{P}}(\theta) = \inf_{\{\mathbb{Q} \in \mathcal{M}, \ \mathbb{Q} \ll \mathbb{P}, \ \mathbb{Q} f(.,\theta) = 0\}} \{I_{\varphi^*}(\mathbb{Q}, \mathbb{P})\}.$$

This can be seen as a projection of $\mathbb{P}$ on the model of interest for the given pseudo-metric I_{φ^*}. If the model is true at $\mathbb{P}$, that is, if $\mathbb{E}_{\mathbb{P}} f(X, \theta) = 0$, then $\beta_{\mathbb{P}}(\theta) = 0$.

The plug-in estimator of $\beta_{\mathbb{P}}(\theta)$ for fixed θ is given by $\beta_{\mathbb{P}_n}(\theta)$, denoted by $\beta_n(\theta)$.

$$\beta_n(\theta) = \inf_{\{\mathbb{Q} \in \mathcal{M}_n, \ \mathbb{Q} f(.,\theta) = 0\}} \{I_{\varphi^*}(\mathbb{Q}, \mathbb{P}_n)\}$$

This quantity may be seen as a test of $\beta_{\mathbb{P}}(\theta) = 0$ and may be inverted to build a confidence region for θ.
The corresponding random confidence region is simply defined by

$$\mathscr{C}_n(\eta) = \{\theta \in \mathbb{R}^q \ | \exists \mathbb{Q} \ll \mathbb{P}_n \text{ with } \mathbb{Q} f(.,\theta) = 0 \text{ and } n I_{\varphi^*}(\mathbb{Q}, \mathbb{P}_n) \le \eta\},$$

where $\eta = \eta(\alpha)$ is a quantity such that

$$\Pr(\theta \in \mathscr{C}_n(\eta)) = 1 - \alpha + o(1).$$

For $\mathbb{Q}$ in $\mathcal{M}_n$, the constraints can be rewritten as $(\mathbb{Q} - \mathbb{P}_n) f(.,\theta) = -\mathbb{P}_n f(.,\theta)$. By Theorem 1, we get the following dual representation which transforms the original program into a much simpler empirical process problem. This representation is at the core of most properties obtained on generalized empirical likelihood:

$$\begin{aligned} \beta_n(\theta) &:= \inf_{\{\mathbb{Q} \in \mathcal{M}_n, (\mathbb{Q} - \mathbb{P}_n) f(.,\theta) = -\mathbb{P}_n f(.,\theta)\}} \{I_{\varphi^*}(\mathbb{Q}, \mathbb{P}_n)\} \\ &= \sup_{\lambda \in \mathbb{R}^q} \left\{ \mathbb{P}_n \left(-\lambda' f(.,\theta) - \varphi(\lambda' f(.,\theta)) \right) \right\}. \end{aligned} \tag{3}$$

The parameter λ is simply the Kuhn & Tucker coefficient associated to the original optimization problem. By remembering that under the assumptions **H1-H4**, φ typically behaves like $x^2/2$ in the neighborhood of 0, the dual representation essentially behaves like a quadratic program in λ. This explains why generalized empirical likelihood essentially behaves asymptotically like the square of a self-normalized sum. In the following, we will also use the notations

$$\overline{f}_n = \frac{1}{n}\sum_{i=1}^n f(X_i, \theta), \ S_n^2 = \frac{1}{n}\sum_{i=1}^n f(X_i, \theta) f(X_i, \theta)' \text{ and } S_n^{-2} = (S_n^2)^{-1}.$$

We state the following theorem which is a minor variation of Bertail et al. [7] :

Theorem 2. *Let $X,\ X_1, ...,\ X_n$ be in $\mathbb{R}^p$, i.i.d. with probability $\mathbb{P}$ and $\theta \in \mathbb{R}^q$ such that $\mathbb{E}_{\mathbb{P}} f(X,\theta) = 0$. Assume that $S^2 = \mathbb{E}_{\mathbb{P}} f(X,\theta) f(X,\theta)'$ is of rank q and that φ satisfies the hypotheses **H1-H4**. Assume that the qualification constraints* **Qual** $(\mathbb{P}_{\mathbf{n}})$ *hold. For any α in $]0,1[$, set $\eta = \frac{\varphi^{(2)}(0)\chi^2_q(1-\alpha)}{2}$, where $\chi^2_q(.)$ is the χ^2 distribution quantile. Then $\mathscr{C}_n(\eta)$ is a convex asymptotic confidence region with*

$$\begin{aligned}\lim_{n\to\infty} \Pr(\theta \notin \mathscr{C}_n(\eta)) &= \lim_{n\to\infty} \Pr(n\beta_n(\theta) \geq \eta) \\ &= \lim_{n\to\infty} \Pr\left(n\overline{f}'_n S_n^{-2} \overline{f}_n \geq \chi^2_q(1-\alpha)\right) \\ &= \alpha.\end{aligned}$$

The proof of this theorem starts from the convex dual-representation and follows the same arguments as Bertail et al. [7] and Owen [36] for the case of the mean.

Remark 1. *If φ is finite everywhere then the qualification constraints are not needed (this is for instance the case for the χ^2 divergence). However, in the case of empirical likelihood or the generalized empirical method introduced below, this actually simply puts some restriction on the θ which are of interest as noticed in the following examples.*

2.3. **Basic examples**

It is easy to check that Cressie-Read discrepancies (see Cressie and Read [13]) fulfill the assumptions **H1-H4** so that Theorem 2 applies for this kind of divergence. Indeed, a Cressie-Read discrepancy can be seen as a φ^*-discrepancy, with φ^* given by:

$$\varphi^*_\kappa(x) = \frac{(1+x)^\kappa - \kappa x - 1}{\kappa(\kappa-1)},\ \varphi_\kappa(x) = \frac{[(\kappa-1)x+1]^{\frac{\kappa}{\kappa-1}} - \kappa x - 1}{\kappa}$$

for some $\kappa \in \mathbb{R}$. This family contains all the usual discrepancies, such as Relative Entropy ($\kappa \to 1$), Hellinger distance ($\kappa = 1/2$), the χ^2 ($\kappa = 2$) and the Kullback distance ($\kappa \to 0$).

We give two examples to illustrate the impact of the choice of φ^* on the generalized empirical likelihood program.

Empirical likelihood and the Kullback discrepancy. In the particular case $\varphi_0(x) = -x - \log(1-x)$ and $\varphi^*_0(x) = x - \log(1+x)$ corresponding to the Kullback divergence for measures

$$K(\mathbb{Q},\mathbb{P}) = -\int \log(\frac{d\mathbb{Q}}{d\mathbb{P}}) d\mathbb{P} + \int (d\mathbb{Q} - d\mathbb{P}),$$

the dual program obtained in (3) becomes, for the admissible θ,

$$\beta_n(\theta) = \sup_{\lambda \in \mathbb{R}^q} \left(\frac{1}{n} \sum_{i=1}^n \log(1 + \lambda' f(X_i,\theta)) \right).$$

As a parametric likelihood indexed by λ, it is easy to show that $2n\beta_n(\theta)$ is asymptotically $\chi^2(q)$ when $n \to \infty$, if the variance of $f(X,\theta)$ is definite. It is also Bartlett-correctable DiCiccio et al. [17]. Using a duality point of view, the proof of the Bartlett-correctability is almost immediate, see Mykland [32] and Bertail [4, 5]. For a general discrepancy, the dual form is not a likelihood and may not be Bartlett-correctable, see DiCiccio et al. [17] and Jing and Wood [27] for the relative entropy E defined by $E(\mathbb{Q},\mathbb{P}) = K(\mathbb{P},\mathbb{Q})$.

Moreover, we necessarily have the $q'_i s > 0$ and the optimization program implies in this case that $\sum_{i=1}^n q_i = 1$, that is the solution is a probability, so that the qualification constraint essentially means that 0 belongs to the convex hull of the $f(X_i,\theta)$. This is in particular the reason why we may obtain very bad coverage probability for empirical likelihood as explained in Tsao [44]. Indeed the results of Tsao [44] show that taking the convex hull of the points (the largest confidence region for empirical likelihood) may yield too narrow confidence regions, when n is small compared to q.

GMM and χ^2 discrepancy. The particular case of the χ^2 discrepancy corresponds to $\varphi_2(x) = \varphi_2^*(x) = \frac{x^2}{2}$. $\beta_n(\theta)$ can be explicitly calculated. Indeed, we get easily that $\lambda = S_n^{-2}\overline{f}_n$ so that, by Theorem 1, the minimum is attained at $\mathbb{Q}^* = \sum_{i=1}^n q_i \delta_{X_i}$ with

$$q_i = \frac{1}{n}(1 + \overline{f}_n' S_n^{-2} f(X_i, \theta))$$

and

$$I_{\varphi_2^*}(\mathbb{Q}^*, \mathbb{P}_n) = \sum_{i=1}^n \frac{(nq_{i,n} - 1)^2}{2n} = \frac{1}{2}\overline{f}_n' S_n^{-2}\overline{f}_n,$$

which is exactly the square of a self-normalized multivariate sum which typically appears in the Generalized Method of Moments (GMM), see also Bertail et al. [8].

Notice that, unlike to the Kullback discrepancy, we may charge positively some region outside of the convex hull of the points, yielding larger (that is too conservative) confidence region.

Remark 2. *If S_n^2 is of rank $l < q$, write $S_n^2 = R' \begin{pmatrix} \Delta_n & 0 \\ 0 & 0 \end{pmatrix} R$, where Δ_n is invertible of rank l, $R = \begin{pmatrix} R_a \\ R_b \end{pmatrix}$ is an orthogonal matrix with $R_a \in \mathfrak{M}_{l,q}(\mathbb{R})$ and $R_b \in \mathfrak{M}_{q-l,q}(\mathbb{R})$. By straightforward arguments, the duality relationship still holds and becomes*

$$\beta_n(\theta) = \sup_{\lambda \in \mathbb{R}^l} \left\{ -\lambda' R_a \overline{f}_n - \frac{1}{2}\lambda' \Delta_n \lambda \right\} = \frac{1}{2n}(R_a \overline{f}_n)' \Delta_n^{-1} (R_a \overline{f}_n).$$

Notice that $(R_a\overline{f}_n)(R_a\overline{f}_n)' = \Delta_n$. This means that if S_n^2 has rank $l < q$ we can always reduce the problem to the study of a self-normalized sum in $\mathbb{R}^l$ and that, from an algorithmic point of view, this reduction is carried out internally by the optimization program. From now on, we will assume that S_n^2 is of rank $l = q$.

3. Quasi-Kullback and Bartlett-correctability

3.1. How to choose the divergence

The previous results are all asymptotic results. A natural statistical issue is how the choice of φ^* influences the corresponding confidence regions and their coverage probability, for finite sample size n, in a multivariate setting.

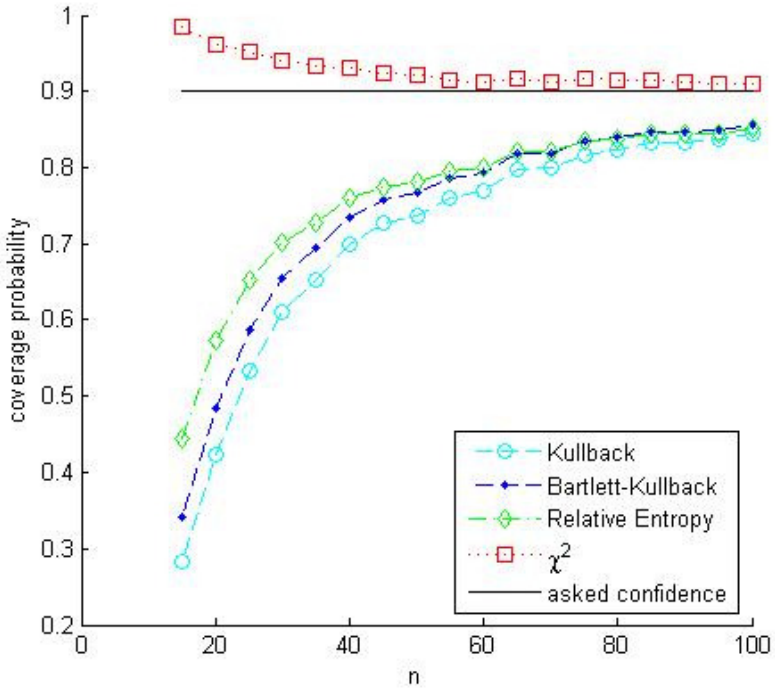


FIGURE 1. Coverage probability for different discrepancies

To illustrate this fact, we use different discrepancies to build confidence intervals for the mean of the product of a uniform r.v. with an independent standard Gaussian r.v. (a scale mixture) on $\mathbb{R}^6$. Figure 1 represents the coverage probability obtained by Monte-Carlo simulations (100 000 repetitions) for different divergences and different sample sizes n. Asymptotically, all these empirical energy minimizers are theoretically equivalent in the case of the mean Bertail et al. [7]. However, this simulation clearly stresses their distinct behavior for small sample sizes. Empirical likelihood corresponding to K performs very badly for small sample size, even with a Bartlett-correction. However, the χ^2 divergence (leading to GMM type of estimators) tends to be too conservative. These problems tend to increase with the dimension of the parameter of interest a well known fact in the empirical likelihood literature. For very small sample size, Tsao [44] obtained an exact upper bounds for the coverage probability of empirical likelihood for q, the parameter size, less than 2, which confirms our simulation results. It also sheds some doubt on the relevance of empirical likelihood when n is small compared to q.

3.2. **Quasi-Kullback or log-proximal divergence**

The main underlying idea of this section is that we want to keep the good properties of the Kullback discrepancy and to avoid some algorithmic problems linked with the behavior of the log of the Kullback discrepancy in the neighborhood of 0. For this, we will introduce family of divergences, the quasi-Kullback. For $\varepsilon \in]0;1]$ and $x \in]-\infty;1[$ let,

$$K_\varepsilon(x) = \varepsilon\, x^2/2 + (1-\varepsilon)(-x - \log(1-x)).$$

This kind of discrepancies is actually currently used in the convex optimization literature and may be seen a regularized log-proximal method (see for instance Ausslender et al. [1]). The resulting optimization algorithm leads to efficient tractable interior point solutions even when the number of constraint is large.

We call the corresponding K_ε^*-discrepancy, the quasi-Kullback discrepancy. The parameter $\varepsilon > 0$ may be interpreted as a regularization parameter (proximal in term of convex optimization). This family fulfills our hypotheses **H1-H5**. Its Fenchel-Legendre transform K_ε^* has the following explicit expression, for all x in $\mathbb{R}$:

$$\begin{aligned} K_\varepsilon^*(x) &= -\frac{1}{2} + \frac{(2\varepsilon - x - 1)\sqrt{1 + x(x + 2 - 4\varepsilon)} + (x+1)^2}{4\varepsilon} \\ &\quad -(\varepsilon - 1)\log \frac{2\varepsilon - x - 1 + \sqrt{1 + x(x + 2 - 4\varepsilon)}}{2\varepsilon}. \end{aligned}$$

Note that the second order derivative of K_ε is bounded from below: $K_\varepsilon^{(2)}(x) \geq \varepsilon$. Moreover, the second order derivative of K_ε^* is bounded both from below and above: $0 \leq K_\varepsilon^{*(2)}(x) \leq 1/\varepsilon$. These controls ensure a quick and regular convergence of the algorithms based on such discrepancies. The corresponding "quasi-empirical likelihood" may be seen as a "regularized" empirical likelihood.

The following theorem establishes sufficient conditions on the regularization parameter ε to obtain the Bartlett-correctability of quasi-empirical likelihood.

Theorem 3. *Under the assumptions of Theorem 2, assume that $f(X,\theta)$ satisfies the Cramer condition: $\overline{\lim}_{||t||\to\infty} |\mathbb{E}_\mathbb{P} \exp(it' f(X,\theta))| < 1$, as well as the moment condition $\mathbb{E}_\mathbb{P} ||f(X,\theta)||^s < \infty$, for $s > 8$. If $\varepsilon = \varepsilon_n = \mathcal{O}\left(n^{-3/2}/\log(n)\right)$ then the quasi-empirical likelihood is Bartlett-correctable up to $\mathcal{O}(n^{-3/2})$.*

The proof is postponed in last section.

This choice of ε is probably not optimal but considerably simplifies the proof. An attentive reading of Corcoran [12] shows that, if ε is small enough, the statistic is still Bartlett-correctable. Unfortunately, as our discrepancy depends on n, Corcoran's result cannot be applied directly and does not allow ε to be precisely calibrated. We conjecture that, at the cost of tedious calculations, the rate of ε_n in $o(n^{-1})$ is enough, at least to get Bartlett-correctability up to $o(n^{-1})$.

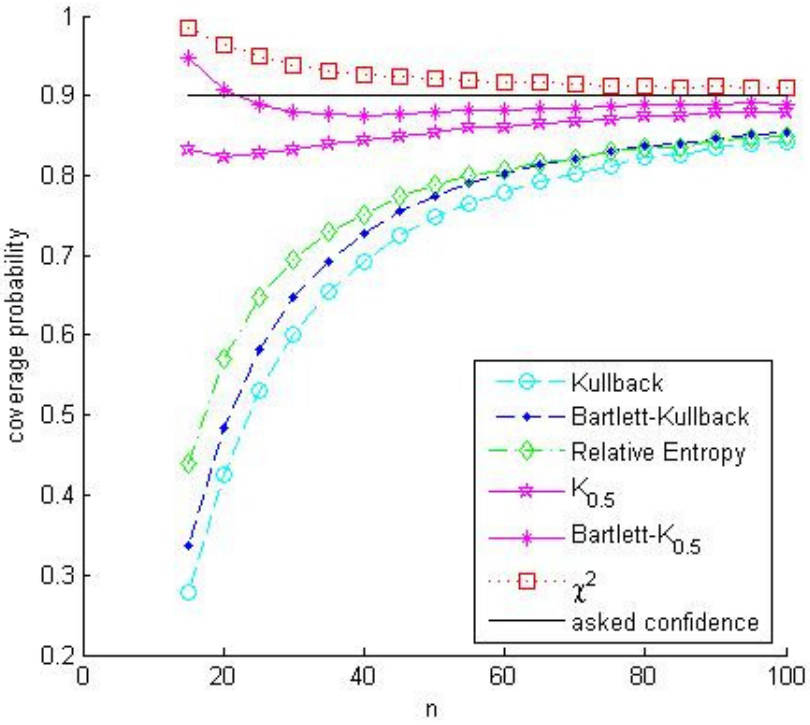


FIGURE 2. Cover probabilities and Quasi-Kullback

Figure 2 illustrates the improvements coming from the use of Quasi-Kullback. It presents the coverage probabilities of the usual discrepancies given in the introduction, as well as the ones for Quasi-Kullback discrepancy (for a given value of $\varepsilon = 0.5$) on the same data. As expected, the Quasi-Kullback discrepancy leads to a confidence region with a coverage probability much closer to the targeted one, especially with a Bartlett adjustment even for an ε which is not close to 0.

4. Exponential bounds for self-normalized sums and quasi-empirical likelihood

As seen in theorem 2, the behavior of generalized empirical likelihood is asymptotically governed by the behavior of the square of a self-normalized multivariate sum. We are now interested in controlling this behavior of quasi-empirical likelihood for finite n. We will show that indeed it is possible to relate it to self normalized vector. For this, we use some exponential bounds obtained by Bertail et al. [8]. For the sake of completeness, we recall two types of bounds, which will allow us to control the finite behavior of quasi-empirical likelihood.

Theorem 4. *Suppose that S^2 is of rank q. Then the following inequalities hold, for finite $n > q$ and for $u < nq$,*

a) *if $f(X_1, \theta)$ has a symmetric distribution, without any additional moment assumption,*

$$\Pr\left(n\overline{f}_n' S_n^{-2} \overline{f}_n \geq u\right) \leq 2qe^{-\frac{u}{2q}}; \tag{4}$$

b) *for general distribution of $f(X_1,\theta)$ with kurtosis $\tilde{\gamma}_4 < \infty$, for any $a > 1$,*

$$\Pr\left(n\overline{f}'_n S_n^{-2}\overline{f}_n \geq u\right) \leq 2qe^{1-\frac{u}{2q(1+a)}} + C(q)\, n^{3\widetilde{q}} e^{-\frac{n}{\tilde{\gamma}_4(q+1)}\left(1-\frac{1}{a}\right)^2} \tag{5}$$

with $\widetilde{q} = \frac{q-1}{q+1}$, $\tilde{\gamma}_4 = \mathbb{E}_{\mathbb{P}}(\|S^{-1}f(X_1,\theta)\|_2^4)$ *and* $C(q) = \frac{(2e\pi)^{2\widetilde{q}}(q+1)}{2^{2/(q+1)}(q-1)^{3\widetilde{q}}} \leq \frac{(2e\pi)^2(q+1)}{(q-1)^{3\widetilde{q}}} \leq 18$.

Moreover for $nq \leq u$, we have

$$\Pr\left(n\overline{f}'_n S_n^{-2}\overline{f}_n \geq u\right) = 0.$$

Part a) in the symmetric multidimensional case follows from an easy but crude extension of Hoeffding [25] (or Eaton and Efron [19], Efron [20]). The exponential inequality (4) is classical in the unidimensional case. Other type of inequalities with suboptimal rate in the exponential have also been obtained by Major [31].

In the general multidimensional framework, the main difficulty is actually to keep the self-normalized structure when symmetrizing the original sum. Another difficulty is to have a precise control of the behavior of the smallest eigenvalue of the normalizing empirical variance. The second term in the right hand side of inequality (5) is essentially due to this control. The crude bound obtained in part a) allows us to use a multidimensional extension of a symmetrization lemma by Panchenko [37]. However for $q > 1$, the bound of part a) is clearly not optimal. A better bound, which has not exactly an exponential form, has been obtained by Pinelis [38]. It essentially says that in the symmetric case the tail of the self-normalized sum can essentially be bounded by the tail of a χ^2 distribution (up to a constant equal to $2e^3/9$). This bounds gives the right behavior of the tail (in q) when n grows, which is not the case for a). However, in the unidimensional case a) still gives a better approximation than Pinelis [38]. It can still be used in the multidimensional case to get crude but exponential bounds as prove in Bertail et al. [8]. For these reasons, we also recall some results of Theorem 4 when using a χ^2 type of control. This essentially consists in extending lemma 1 of Panchenko [37] to non exponential bound.

In the following, we denote f_q the density function of a $\chi^2(q)$ distribution, which is given by $f_q(x) = \dfrac{1}{2^{q/2}\Gamma(q/2)} x^{q/2-1}e^{-\frac{x}{2}}$, with $\Gamma(p) = \int_0^{+\infty} x^{p-1}e^{-x}dx$ and let $\overline{F}_q$ denote the survival function ($\overline{F}_q(x) = \int_x^{+\infty} f_q(y)dy$).

Theorem 5. *We use the same notation as in the Theorem 4. Then the following inequalities hold, for finite $n > q$ and for $u < nq$,*

a) *(Pinelis 1994) if $f(X_1,\theta)$ has a symmetric distribution, without any additional moment assumption,*

$$\Pr\left(n\overline{f}'_n S_n^{-2}\overline{f}_n \geq u\right) \leq \frac{2e^3}{9}\overline{F}_q(u), \tag{6}$$

b) *for general distribution of $f(X_1,\theta)$ with kurtosis $\tilde{\gamma}_4 < \infty$, for any $a > 1$ and for $2q(1+a) \leq u$,*

$$\Pr\left(n\overline{f}'_n S_n^{-2}\overline{f}_n \geq u\right) \leq \frac{2e^3}{9\Gamma(\frac{q}{2}+1)}\left(\frac{u-q(1+a)}{2(1+a)}\right)^{\frac{q}{2}} e^{-\frac{u-q(1+a)}{2(1+a)}} + C(q)\, n^{3\widetilde{q}} e^{-\frac{n\left(1-\frac{1}{a}\right)^2}{\tilde{\gamma}_4(q+1)}} \tag{7}$$

Moreover, for $nq \leq u$, $\Pr\left(n\overline{f}'_n S_n^{-2}\overline{f}_n \geq u\right) = 0$.

Remark 3. *In the best case, past studies gave some bounds for n sufficiently large, without an exact value for "sufficiently large". Here, the bounds are valid for any n. All the constants are also explicit. This bound may also be used to give some ideas on the sample size needed to reach a given confidence level (as a function of q and $\tilde{\gamma}_4$).*

The following corollary implies that, for the whole class of quasi-Kullback discrepancies, the finite sample behavior of the corresponding empirical energy minimizers can be reduced to the study of a self-normalized sum.

Corollary 1. *Under the hypotheses of Theorem 2, the following inequalities hold, for finite $n > q$, for any $\eta > 0$, for any $n > \frac{2\varepsilon\eta}{q}$,*

$$\Pr(\theta \notin \mathscr{C}_n(\eta)) = \Pr(n\beta_n(\theta) \geq \eta) \leq \Pr\left(n\overline{f}_n S_n^{-2}\overline{f}_n \geq 2\varepsilon\eta\right). \tag{8}$$

Else if $n \leq \frac{2\varepsilon\eta}{q}$, $\Pr(\theta \notin \mathscr{C}_n(\eta)) = 0$.
Then, for $n > 2q$, bounds (4-7) may be used with $u = 2\varepsilon\eta$.

Remark 4. *In Hjort et al. [24], convergence of empirical likelihood is investigated when q is allowed to increase with n. They show that convergence to a χ^2 distribution still holds when $q = O(n^{\frac{1}{3}})$ as n tends to infinity.*

Our bounds show that even if $q = o(n/log(n))$, it is still possible to get asymptotically valid confidence intervals with our bounds. Notice that the constant $C(q)$ does not increase with q as can be seen in Figure 3.

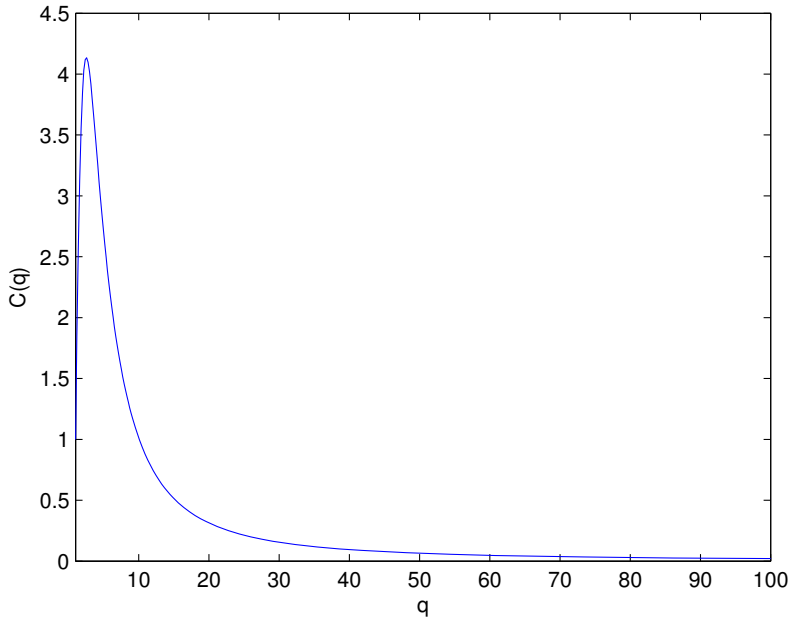


FIGURE 3. Value of $C(q)$ as a function of q

A close examination of the bounds shows that essentially $q\tilde{\gamma}_4$ has to be small compared to n for practical use of these bounds. Of course practically $\tilde{\gamma}_4$ is not known, however one may use an estimator or an upper bound for this quantity to get some insight on a given estimation problem.

Notice that the bounds are non-informative when $\varepsilon \to 0$, which corresponds to empirical likelihood. Actually, it is not possible to establish an exponential bound for this case. If we were able to do so, for a sufficiently large η, we could control the confidence region built with empirical likelihood for any level $1 - \alpha$. This would contradict the statements of Tsao [44], which gives a lower bound for the attainable levels.

5. Generalization to process valued parameters

In this section we will consider the setting where the parameter of interest is itself a process value parameter indexed by some class of functions or may be approximated by such parameter. Applications

of this type appear naturally in many semi-parametric models, which may be seen as infinite M-parameters. Another typical example is the case of confidence interval of cumulative distribution functions on a compact set.

For this we consider the following abstract empirical process framework (see van der Vaart and Wellner [45] for details). $\mathcal{F}$ is a subset of functions of a normed space of function here $L^2(\mathbb{P}) = \{h,\ \mathbb{E}_{\mathbb{P}}(h^2) < \infty\}$ endowed with $||h||_{2,\mathbb{P}} = \left(\mathbb{E}_{\mathbb{P}}(h^2)\right)^{1/2}$. We assume that $\mathcal{L}_\infty(\mathcal{F})$ is equipped with the uniform norm

$$\|\mathbb{P}\|_{\mathcal{F}} = \sup_{h\in\mathcal{F}} \left| \int h d\mathbb{P} \right|.$$

To avoid measurability problems, we assume that expectations (resp. probabilities) are outer expectations (resp. outer probabilities) so that weak convergence is interpreted as Hoffman-Jørgensen convergence (see van der Vaart and Wellner [45] for details). For the same reason, we will also assume that $\mathcal{F}$ is image admissible Suslin (for the definition of the image admissible Suslin property see Dudley [18], sections 10.3 and 11.1.). This ensures that the classes of the square functions and difference of square functions are) $\mathbb{P}$-measurable (see Dudley [18]). In the following, it is assumed that $\mathcal{F}$ is a Donsker Class of functions with envelop H satisfying

$$0 < \int H^2 dP < \infty. \tag{9}$$

Notice that if H is an envelop of the class then $H + 1$ is also an envelop, so that we may assume without loss of generality that $H \geq 1$. The empirical process $n^{1/2}(\mathbb{P}_n - \mathbb{P})$ indexed by $\mathcal{F}$ converges (as an element of $\mathcal{L}_\infty(\mathcal{F})$) to a limit $G_{\mathbb{P}}$, which is a tight Borel measurable element of $\mathcal{L}_\infty(\mathcal{F})$ with uniformly $||\ ||_{2,P}$ continuous sample paths $f \to G_P(f)$. Extensive references and results on empirical processes indexed by class of functions and conditions on $\mathcal{F}$ to be Donsker may be found in van der Vaart and Wellner [45].

We are now interested in the infinite value parameter $\theta = (\mathbb{P}f)_{f\in\mathcal{F}} = (\mathbb{E}_{\mathbb{P}}f)_{f\in\mathcal{F}}$ and write $\theta(f) = \mathbb{P}f$ for the component of θ on a particular f.

For an infinite parameter, the initial problem would be to solve the semi-infinite optimisation problem

$$\beta_n(\theta) = \inf_{\{\mathbb{Q}\in\mathcal{M}_n, \mathbb{Q}f=\mathbb{E}_{\mathbb{P}}f,\ \ f\in\mathcal{F}\}} \{I_{\varphi^*}(\mathbb{Q},\mathbb{P}_n)\}.$$

However this not possible directly. Several approaches have been recently proposed in the literature to handle such problems. A first one is to discretize the problem and to retain a reasonable number of constraints (typically of order $n^{1/3}$, see Hjort et al. [24]) before applying the empirical likelihood procedure. Another way to discretize the problem is to consider what is called in the statistical-learning literature a skeleton class approximating $\mathcal{F}$. One problem is of course the choice of the constraints and the loss of efficiency induced by this choice. Another simpler proposition is to try to maximize the empirical likelihood program component by component and to try to catch the worst possible matching (see Varron [46]). In this case, we can still obtain non asymptotic bounds for the empirical likelihood program using previous results on self-normalized processes.

Instead of solving the initial problem, we try to solve the programs defined by

$$\forall f \in \mathcal{F}, \quad \beta_n(f) = \inf_{\{\mathbb{Q}\in\mathcal{M}_n,\ \mathbb{Q}f=E_{\mathbb{P}}f\}} \{I_{\varphi^*}(\mathbb{Q},\mathbb{P}_n)\}.$$

and to consider the concentrated empirical likelihood function

$$\widetilde{\beta}_n(\theta) = \sup_{f\in\mathcal{F}} \{\beta_n(f)\}.$$

Then, using the duality principle of the first part, we have the following result provided that the constraints qualification are satisfied for any f. In practice this may be very difficult to check and this is one of the reason why we advocate the use of a divergence of the quasi Kullback type. In this case we simply have by using the same arguments as before

$$\begin{aligned}\widetilde{\beta}_n(\theta) &= \sup_{f\in\mathcal{F}}\left\{\sup_{\lambda\in\mathbb{R}}\left\{\mathbb{P}_n\left(-\lambda'(f-\mathbb{P}f)-\varphi(\lambda'(f-\mathbb{P}f))\right)\right\}\right\}\\ &\leq \sup_{f\in\mathcal{F}}\left\{\frac{(\mathbb{P}_n f-\mathbb{P}f)^2}{2\varepsilon\mathbb{P}_n(f-\mathbb{P}f)^2}\right\}=\frac{1}{2n\varepsilon}\sup_{f\in\mathcal{F}}\frac{\left(\sum_{i=1}^n f(X_i)-\mathbb{P}f\right)^2}{\sum_{i=1}^n(f(X_i)-\mathbb{P}f)^2}\end{aligned}$$

As a consequence, the concentrated empirical likelihood is now controlled by a self-normalized process. Such control has been obtained under different type of hypothesis in the probabilistic literature : see for instance Bercu et al. [3].

Under the hypotheses that $\mathcal{F}$ be a permissible class (in the sense of Pollard (1984)) of real, measurable, centered and normalized functions. If it is assumed that

(1) $\mathcal{F}$ is a countable class with finite bracketing numbers in $L^2(\mathbb{P})$
(2) $E=\sup_{n>0}\mathbb{E}\left[\sup_{f\in\mathcal{F}}\max(\sqrt{n}\mathbb{P}_n(f);0)\right]<+\infty$,
(3) the class $\mathcal{F}$ is symmetric in the sense that if $f\in\mathcal{F}$ then $-f\in\mathcal{F}$

then we obtain, using theorem 1.1 of Bercu et al. [3], that for any $\epsilon\in[0,1],\delta>0$ and $\alpha>\sqrt{2}$, there exists γ and n_0 depending on $\mathcal{F}$, ε, α and δ such that, for $n\geq n_0$ and for any x in $[0,\gamma\sqrt{n}]$,

$$Pr\left(\widetilde{\beta}_n(\theta)\geq 2\varepsilon(x+\alpha E)^2\right)\leq 4\exp\left(-\frac{x^2}{4\alpha^2(1+\delta)}\right).$$

However since this control depends on (unknown) constants (themselves depending on ε and the class of functions) there is still some room for improvements.
Moreover it should be mentioned that a precise control of the eigenvalues of the covariance function is needed for a valid bound to exist. This is by no mean astonishing and it is implicitly assumed in condition $E<+\infty$.

Indeed if we consider the particular case where $\mathcal{F}=\{f(x)=\mathbb{1}_{x<y}, y\in\mathbb{R}\}$ then the parameter $\theta=F$ is the cumulative distribution function and in that case we have the control

$$\widetilde{\beta}_n(\theta)\leq\frac{1}{2\varepsilon}\sup_{y\in\mathbb{R}}\left[\frac{\left(F_n(y)-F(y)\right)^2}{F_n(y)(1-F_n(y))}\right]$$

and it is known from Jaeschke [26] that $n\beta_n(\theta)$ is always equal to ∞, as well as $E=+\infty$. It follows that in the case of a cumulative repartition function that we can only control the empirical likelihood over a compact set in the interior domain of the support of F.

6. Simulations and calibration

6.1. **Non-asymptotic regions**

Previous simulations studies (see Bertail et al. [6]) show that, for small values of n, the values of η are quite high, leading to confidence regions that may be too conservative but that are very robust. In the following

- "Symmetric bound" corresponds to η obtained by inverting the Pinelis inequality in the symmetric case, that is the quantile of a χ^2.
- "NS", for "Non-symmetric", corresponds the η obtained by inverting the general exponential bounds, from Theorems 4b) or 5b).

Numerical results show that the profile quasi-likelihood gets wider as ε increases. As a consequence, the asymptotic confidence intervals become wider. With the non-asymptotic bounds, the behavior

of the corresponding confidence region as ε increases is more delicate to understand. The profile likelihood gets wider but the η's corresponding to the symmetric bound and NS bounds decrease like $1/\varepsilon$. These two behaviors have contradictory effects on the confidence regions $\mathscr{C}_n(\eta)$. It seems that the effect of the decrease of η dominates: the confidence regions get smaller when ε increases. For higher dimension or for a smaller α, the two contradictory effects could be balanced.

In Figure 4, we build confidence regions for the mean of a two-dimensional data, for two sample sizes ($n = 500$ and 2000) and two distributions :

1) a Gaussian scale mixture, that is realizations of $U * N$, where U and N are respectively independent uniform r.v.'s on [0,1] and standard gaussian r.v.'s;
2) the discrete distribution d_1 defined by

$$\frac{1}{100} \cdot \delta_{(10,10)} + \frac{0.81}{4} \sum_{a,a'=\pm 1} \delta_{(a,a')} + \frac{0.09}{2} \sum_{a=\pm 1} \left(\delta_{(a,10)} + \delta_{(10,a)} \right).$$

We give in Figure 4 the corresponding 90% confidence regions, using respectively the asymptotic approximation from Theorem 2, the symmetric bound from Theorem 5a) and the general bounds (NS) from Theorems 4b) and 5b) with the true kurtosis.

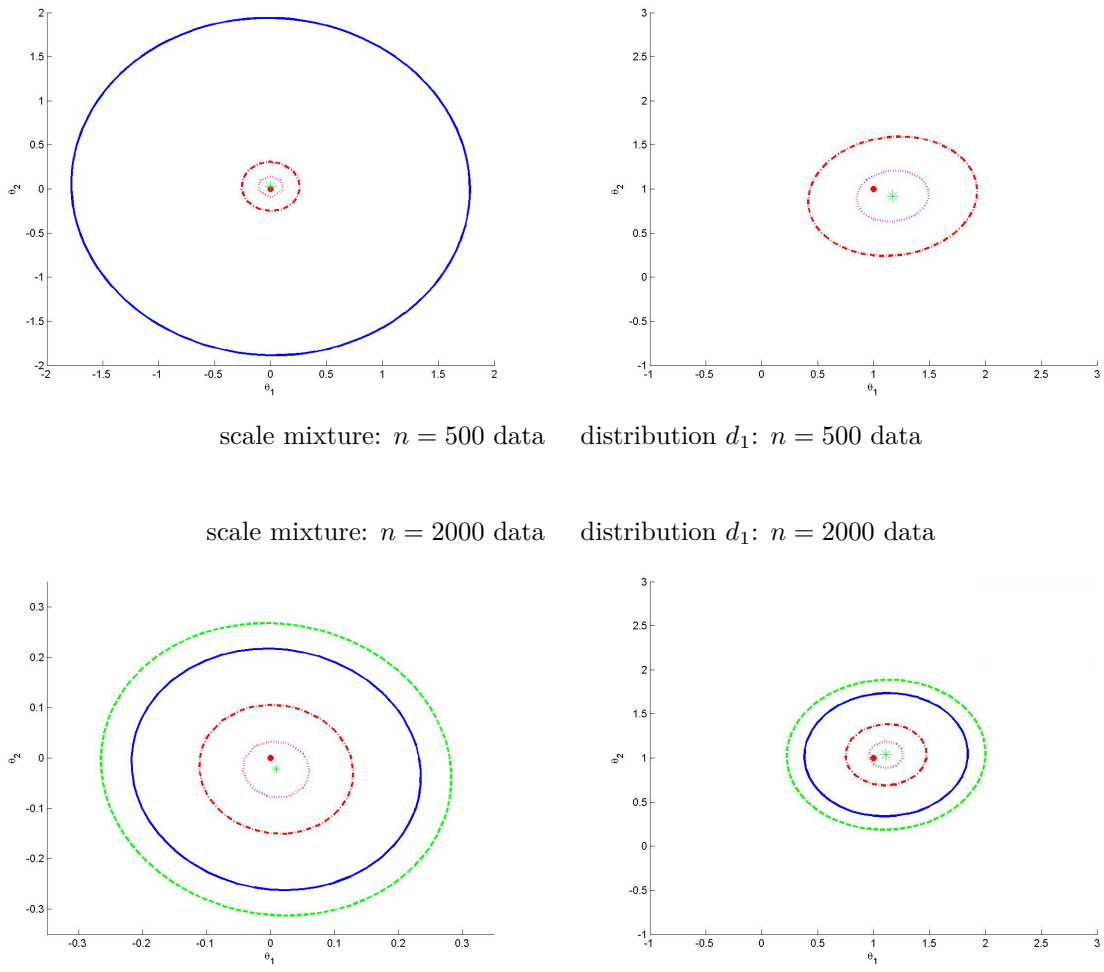


FIGURE 4. Confidence regions, for 2 distributions and 2 data sizes. Legend : red point true mean, green star empirical mean, pink doted line asymptotic confidence region (Theorem 2), red doted line confidence region based on the exponential bound under symmetry (Theorem 5.a), blue continuous line confidence region based on the exponential bound under non symmetry distribution (Theorem 5.b), green doted line confidence region based on the exponential bound under non symmetry distribution (Theorem 4.b)

For small sample size, as expected, the confidence regions obtained with NS bounds are quite large (for our discrete data and $n = 500$, the regions are too large to be represented on the figure) with a coverage probability close to 1. On the contrary, the asymptotic confidence regions are small but when the distribution has a large $\tilde{\gamma}_4$, the coverage probability can be significantly smaller than the targeted level $1 - \alpha$. Thus the use of NS bounds are essentially justified to protect oneself against exotic distributions.

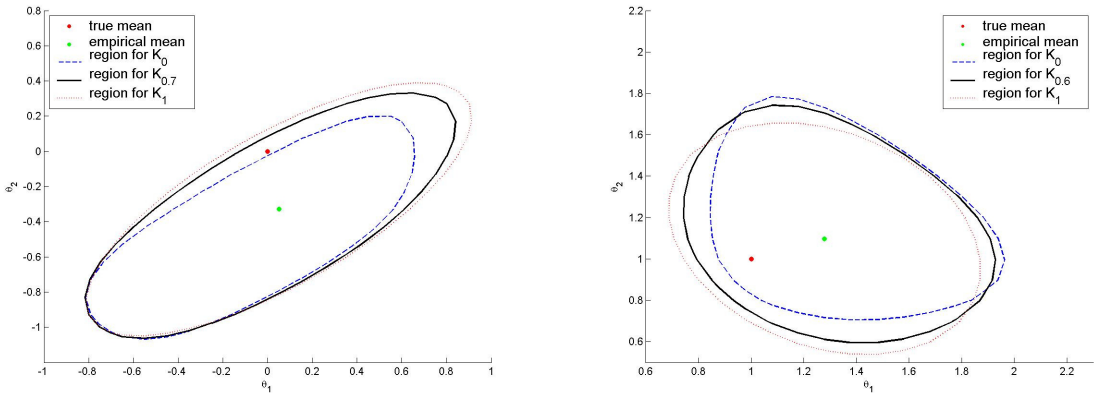
6.2. **Calibration of asymptotic confidence regions**

Corollary 1 does not allow for a precise calibration of ε for finite sample size. Indeed, the finite exponential bounds essentially say that the larger ε is (close to 1), the better the bound. This clearly advocates that, in term of our bound sizes, the χ^2 discrepancy leads to the best results. This is partially true in the sense that the χ^2 leads immediately to a self-normalized sum which

has quite robust properties. However, it can be argued that, for regular enough distributions, the χ^2 discrepancy leads to confidence regions that are to too conservative confidence intervals. The result on Bartlett-correctability suggests that the bias of the empirical minimizer for quasi-Kullback is smaller for very small values of ε (see also Newey and Smith [33] for arguments in that direction). Choosing adequately ε could result in a better equilibrium and a compromise between coverage probability and the adaptation to the data.

From a practical point of view, several choices are possible for calibrating ε. A simple solution is simply to use cross-validation (either bootstrap, leave one-out or K-fold methods). Of course, this is very computationally-expensive but the use of a quasi-Kullback distance eases the convergence of the algorithms. It is not clear how the use of cross-validation and thus the use of an ε depending on the data will deteriorate the finite sample bounds.

Figure 5 allows us to compare the asymptotic confidence regions built with the Kullback discrepancy (K_0), the χ^2 (K_1) and the Quasi-Kullback (K_ε) with ε chosen by cross-validation, for a parameter in $\mathbb{R}^2$. The algorithm leads to $\varepsilon \simeq 0.7$ for the scale mixture example and $\varepsilon \simeq 0.6$ for a standard exponential distribution .



scale mixture: 15 data exponential distribution: 25 data

FIGURE 5. Asymptotic confidence regions for data driven K_ε.

Figure 6 represents the coverage probability obtained by Monte-Carlo (25 000 repetitions) simulations of scale-mixture distribution with $q = 6$ for K_ε with data driven ε and some specific choice of ε, for different sample sizes n.

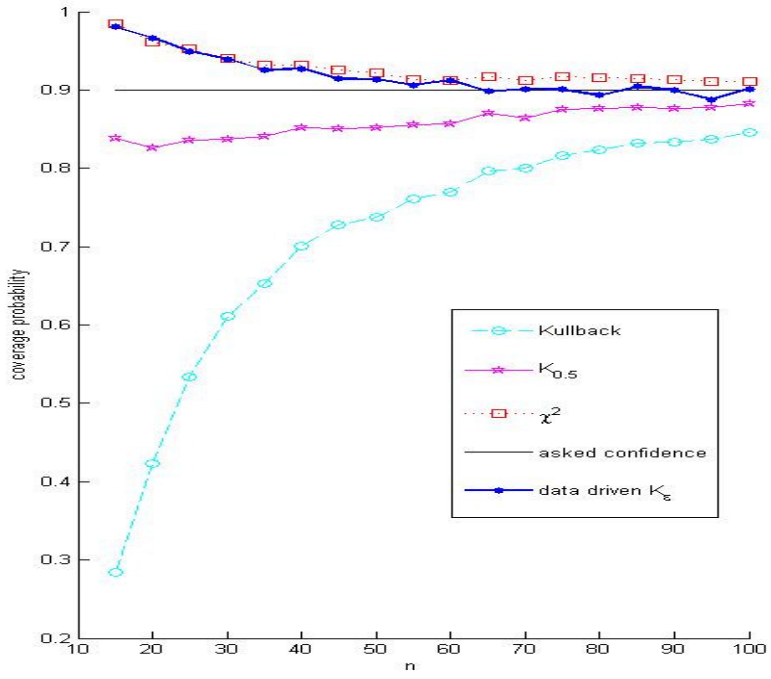


FIGURE 6. Coverage probability for different values of ε in terms of n.

In the multidimensional case ($q > 1$), with n finite, the volume of the confidence region for the quasi-Kullback divergence remains closed to the volume of the ellipsoid corresponding to the χ^2 divergence with a better coverage probability.

The "adaptative" value of ε decreases with n. Over our 25 000 Monte-Carlo repetitions, the mean value of ε is 1 for $n = 15$ and $n = 20$ (that is cross-validation automatically select the χ^2 divergence for small n. It decreases to 0.7 for $n = 100$.

For smooth distributions like our scale mixture, the coverage probability of the confidence region constructed with the calibrated K_ε is close to the targeted one. Moreover, the region is small and adapts to the data.

A. Technical details

A.1. **Proof of Theorem 3**

Write $\beta_n^\varepsilon(\theta)$ for the value of n times the sup in the dual program (3) when $\varphi = K_\varepsilon$. $\beta_n^0(\theta)$ corresponds to the log likelihood ratio for Kullback discrepancy $\varphi = K_0$ and $\beta_n^1(\theta)$ corresponds to the minimization of the χ^2-divergence $\varphi = K_1$. Let $\mathbb{E}_n$ be either the true value of $\mathbb{E}[\beta_n^0(\theta)]/q$ or an estimator of this quantity such that empirical likelihood is Bartlett-correctable when standardized by this quantity. We denote

$$T_n^\varepsilon = \frac{2\beta_n^\varepsilon(\theta)}{\mathbb{E}_n}.$$

Then, using DiCiccio et al. [17] (see also Bertail [5]), under the Cramer condition and assuming $\mathbb{E}_{\mathbb{P}}||f(X,\theta)||^8 < \infty$, the Bartlett-correctability of T_n^0 implies that

$$\Pr\left(\frac{2\beta_n^0(\mu)}{\mathbb{E}_n} \geq x\right) = \bar{F}_{\chi^2}(x) + \mathcal{O}(n^{-2}),$$

where we denote $\bar{F}_Z(.) = \int_0^{+\infty} d\mathbb{P}(z)$, when $Z \sim \mathbb{P}$. This equality implies in particular that

$$\bar{F}_{T_n^0}(\eta - n^{-\frac{3}{2}}) = \bar{F}_{\chi^2(q)}(\eta) + \mathcal{O}(n^{-\frac{3}{2}}). \tag{10}$$

Now, we can write

$$\begin{aligned} T_n^\varepsilon &= \frac{2}{\mathbb{E}_n} \sup_{\lambda \in \mathbb{R}^q} \left\{ \sum_{i=1}^n \lambda' f(X_i, \theta) - \sum_{i=1}^n K_\varepsilon(\lambda' f(X_i, \theta)) \right\} \\ &\leq \frac{2}{\mathbb{E}_n} \left\{ \varepsilon \beta_n^1(\theta) + (1-\varepsilon)\beta_n^0(\theta) \right\}. \end{aligned}$$

In other words

$$T_n^\varepsilon \leq T_n^0 + \varepsilon \left[T_n^1 - T_n^0 \right].$$

This implies

$$\bar{F}_{T_n^\varepsilon}(\eta) \leq \bar{F}_{T_n^0 + \varepsilon[T_n^1 - T_n^0]}(\eta).$$

We also have from (10)

$$\begin{aligned} \bar{F}_{T_n^0 + \varepsilon[T_n^1 - T_n^0]}(\eta) &\leq \Pr(T_n^0 + n^{-\frac{3}{2}} \geq \eta) + \Pr(|T_n^1 - T_n^0| \geq \varepsilon^{-1} n^{-\frac{3}{2}}) \\ &= \bar{F}_{T_n^0}(\eta - n^{-\frac{3}{2}}) + \Pr(|T_n^1 - T_n^0| \geq \varepsilon^{-1} n^{-\frac{3}{2}}) \\ &= \bar{F}_{\chi^2}(\eta) + \mathcal{O}(n^{-\frac{3}{2}}) + \Pr(|T_n^1 - T_n^0| \geq \varepsilon^{-1} n^{-\frac{3}{2}}). \end{aligned}$$

If we take ε of order $n^{-3/2}\log(n)^{-1}$, the last term in the right hand side of this inequality is of order $\mathcal{O}(n^{-3/2})$. This can be shown by using for example the moderate deviation inequality for T_n^1 and the fact that T_n^0 is already Bartlett-correctable. It follows that the corresponding discrepancy is still Bartlett-correctable, at least up to the order $\mathcal{O}(n^{-3/2})$.

A.2. **Proof of corollary 1**

Following the arguments of the remark of Theorem 2, we use the dual form and expand K_ε near 0. Then we get

$$\begin{aligned} \sup_{\lambda \in \mathbb{R}^q} \left\{ -n\lambda' \overline{f}_n - \frac{1}{2} \sum_{i=1}^n (\lambda' f(X_i, \theta))^2 K_\varepsilon^{(2)}(t_{i,n}) \right\} & \\ \leq \sup_{\lambda \in \mathbb{R}^q} \left\{ -n\lambda' \overline{f}_n - \frac{1}{2} \sum_{i=1}^n (\lambda' f(X_i, \theta))^2 \varepsilon \right\}. & \end{aligned} \tag{11}$$

Indeed, by construction of the quasi-Kullback, we have $K_\varepsilon^{(2)} \geq \varepsilon$. If we write $l = -\varepsilon\lambda$, the right hand side of inequality (11) becomes

$$\frac{n}{\varepsilon} \sup_{l \in \mathbb{R}^q} \left\{ l' \overline{f}_n - \frac{1}{2} l' S_n^2 l \right\} = \frac{n}{2\varepsilon} \overline{f}_n' S_n^{-2} \overline{f}_n.$$

Thus we immediately get

$$\Pr(\theta \notin \mathscr{C}_n(\eta)) \leq \Pr\left(\frac{n}{2} \overline{f}_n' S_n^{-2} \overline{f}_n \geq \eta\varepsilon \right).$$

References

[1] A. Ausslender, M. Teboulle, and S. Ben-Tiba. Logarithm-quadratic proximal method for variational inequalities. *Computational Optimization and Applications*, 12:31–40, 1999.

[2] K. A. Baggerly. Empirical likelihood as a goodness of fit measure. *Biometrika*, 85:535–547, 1998.

[3] B. Bercu, E. Gassiat, and E. Rio. Concentration inequalities, large and moderate deviations for self-normalized empirical processes. *Annals of Probability*, 30(4):1576–1604, 2002.

[4] P. Bertail. Empirical likelihood in some nonparametric and semiparametric models. In M. S. Nikulin, N. Balakrishnan, M. Mesbah, and N. Limnios, editors, *Parametric and Semiparametric Models with Applications to Reliability, Survival Analysis, and Quality of Life*, Statistics for Industry and Technology. Birkhauser, 2004.

[5] P. Bertail. Empirical likelihood in some semi-parametric models. *Bernoulli*, 12(2):299–331, 2006.

[6] P. Bertail, E. Gautherat, and H. Harari-Kermadec. Exponential bounds for quasi-empirical likelihood. Working Paper nï£¡34, CREST, 2005.

[7] P. Bertail, H. Harari-Kermadec, and D. Ravaille. φ-divergence empirique et vraisemblance empirique généralisée. *Annales d'Économie et de Statistique*, 85:131–158, 2007.

[8] P. Bertail, E. Gautherat, and H. Harari-Kermadec. Exponential bounds for multivariate self-normalized sums. *Electronic communication in probability*, 13:628–640, 2008.

[9] P. Bertail, E. Gautherat, and H. Harari-Kermadec. Empirical $\phi^\star$-divergence minimizers for hadamard differentiable functionals. In M. Akritas, S. Lahiri, and D. Politis, editors, *Topics in Nonparametric Statistics*, volume 74, pages 21–32. Springer Proceedings in Mathematics & Statistics, 2014.

[10] J. M. Borwein and A. S. Lewis. Duality relationships for entropy-like minimization problem. *SIAM Journal on Control and Optimization*, 29(2):325–338, 1991.

[11] M. Broniatowski and A. Kéziou. Parametric estimation and tests through divergences and the duality technique. *Journal of Multivariate Analysis*, 100:16–36, 2009.

[12] S. A. Corcoran. Bartlett adjustment of empirical discrepancy statistics. *Biometrika*, 85(4):967–972, 1998.

[13] N. Cressie and T. R. C. Read. Multinomial goodness-of-fit tests. *Journal of the Royal Statistical Society, Series B*, 46(3):440–464, 1984.

[14] I. Csiszár. Information-type measures of difference of probability distributions and indirect observations. *Studia Scientiarum Mathematicarum Hungarica*, 2:299–318, 1967.

[15] J. C. Deville and C. E. Sarndal. Calibration estimators in survey sampling. *Journal of the American Statistical Association*, 87:376–382, 1992.

[16] T. DiCiccio and J. Romano. Nonparametric confidence limits by resampling methods and least favorable families. *International Statistical Review*, 58:59–76, 1990.

[17] T. DiCiccio, P. Hall, and J. Romano. Empirical likelihood is bartlett-correctable. *Annals of statistics*, 19(2):1053–1061, 1991.

[18] R. Dudley. *A course on empirical processes*. Ecole d'été de probabilités de Saint Flour. Springer-Verlag, N.Y., 1984.

[19] M. L. Eaton and B. Efron. Hotelling's t^2 test under symmetry conditions. *Journal of American statistical society*, 65:702–711, 1970.

[20] B. Efron. Student's t-test under symmetry conditions. *Journal of American statistical society*, 64:1278–1302, 1969.

[21] F. Gamboa and E. Gassiat. Bayesian methods and maximum entropy for ill-posed inverse problems. *Annals of Statistics*, 25(1):328–350, 1996.

[22] A. Golan, G. Judge, and D. Miller. *Maximum Entropy Econometrics*. Wiley, New York, 1996.

[23] H. O. Hartley and J. N. K. Rao. A new estimation theory for sample surveys. *Biometrika*, 55: 547–557, 1968.

[24] N. L. Hjort, I. W. McKeague, and I. Van Keilegom. Extending the scope of empirical likelihood. *Annals of Statistics*, 37:1079–1111, 2009.

[25] W. Hoeffding. Probability inequalities for sums of bounded variables. *Journal of the American Statistical Association*, 58:13–30, 1963.

[26] D. Jaeschke. The asymptotic distribution of the supremum of the standardized empirical distribution function on subintervals. *Annals of Statistics*, 7:108–115, 1979.
[27] B. Y. Jing and A. T. A. Wood. Exponential empirical likelihood is not bartlett correctable. *Annals of Statistics*, 24:365–369, 1996.
[28] C. Léonard. Convex conjugates of integral functionals. *Acta Mathematica Hungarica*, 93(4): 253–280, 2001.
[29] C. Léonard. Minimization of energy functionals applied to some inverse problems. *Applied mathematics and optimization*, 44(3):273–297, 2001.
[30] F. Liese and I. Vajda. *Convex Statistical distance*. Teubner, Leipzig, 1987.
[31] P. Major. A multivariate generalization of Hoeffding's inequality. *Electronic Communication in Probability*, 2:220–229, 2006.
[32] P. A. Mykland. Dual likelihood. *Annals of Statistics*, 23:396–421, 1995.
[33] W. K. Newey and R. J. Smith. Higher order properties of GMM and generalized empirical likelihood estimators. *Econometrica*, 72(1):219–255, 2004.
[34] A. B. Owen. Empirical likelihood ratio confidence intervals for a single functional. *Biometrika*, 75(2):237–249, 1988.
[35] A. B. Owen. Empirical likelihood ratio confidence regions. *Annals of Statistics*, 18:90–120, 1990.
[36] A. B. Owen. *Empirical Likelihood*. Chapman and Hall/CRC, Boca Raton, 2001.
[37] D. Panchenko. Symmetrization approach to concentration inequalities for empirical processes. *Annals of Probability*, 31(4):2068–2081, 2003.
[38] I. Pinelis. Extremal probabilistic problems and Hotelling's t^2 test under a symmetry condition. *Annals of Statistics*, 22(1):357–368, 1994.
[39] Y. S. Qin and J. Lawless. Empirical likelihood and general estimating equations. *Annals of Statistics*, 22(1):300–325, 1994.
[40] R. T. Rockafellar. Integrals which are convex functionals. *Pacific Journal of Mathematics*, 24: 525–539, 1968.
[41] R. T. Rockafellar. *Convex Analysis*. Princeton University Press, Princeton, NJ, 1970.
[42] R. T. Rockafellar. Integrals which are convex functionals (II). *Pacific Journal of Mathematics*, 39:439–469, 1971.
[43] R. J. Smith. Alternative semi-parametric likelihood approaches to generalised method of moments estimation. *Economic Journal*, 107(441):503–519, 1997.
[44] M. Tsao. Bounds on coverage probabilities of the empirical likelihood ratio confidence regions. *Annals of Statistics*, 32(3):1215–1221, 2004.
[45] A. van der Vaart and J. Wellner. *Weak Convergence and Empirical Process: With Applications to Statistics*. Springer Verlag, 1996.
[46] D. Varron. Empirical likelihood confidence bands for functional parameters in plug-in estimation. Working Paper, 2014.

Acknowledgments : The authors thank two anonymous referees for their suggestions who improved the presentation of the paper. This work was partially developed within the MME-DII Center of Excellence (ANR-11-LABEX-0023-01). The first author acknowledges the support of the French Agence Nationale de la Recherche(ANR) under grant ANR-13-BS-01-0005 (project SPADRO).