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

n° 2005-34

Empirical φ^*-Discrepancies and Quasi-Empirical Likelihood : Exponential Bounds

P. BERTAIL* – E. GAUTHERAT*
H. HARARI-KERMADEC*

* CREST-LS, Timbre J340, 3 avenue Pierre Larousse, 92245 Malakoff Cedex. France.

Empirical φ^*-Discrepancies and quasi-empirical likelihood : exponential bounds.

BERTAIL Patrice, GAUTHERAT Emmanuelle
& HARARI-KERMADEC Hugo
CREST-LS, Timbre J340, Malakoff

17th January 2006

Abstract

We study some extensions of the empirical likelihood method, when the Kullback distance is replaced by some general convex divergence or I_{φ^*} discrepancy. 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, explicit and exponential bounds for the confidence intervals that may be deduced by using this method. These bounds are derived via the study of self-normalized sums in the multivariate case. The results on self-normalized sums are of interest by themselves.

1 Introduction

Empirical likelihood is now a useful and classical method for testing or constructing confidence regions for the value of some parameters in non-parametric or semi-parametric models. It has been introduced and studied by Owen (1988, 1990), see Owen (2001) for a complete overview and exhaustive references. The now well-known idea of empirical likelihood consists in maximizing a profile likelihood supported by the data, under some model constraints. It can be seen as an extension of "model based likelihood" used in survey sampling when some marginal constraints are available (see Hartley & Rao, 1968, Deville & Sarndal, 1992). Owen and many followers have shown that one can get a useful and automatic non-parametric version of Wilks' theorem (stating the convergence of the log-likelihood ratio 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 moment constraints (see Qin & Lawless, 1994, Newey & Smith, 2003). It can now be considered as an alternative to the generalized method of moments (GMM, see Smith, 1997). 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 DiCiccio *et al.*, 1991).

A possible interpretation of empirical log-likelihood ratio is to see it as the minimization of the Kullback divergence, say K, between the empirical distribution of the data $\mathbb{P}_n$ and a measure (or a probability measure) $\mathbb{Q}$ dominated by $\mathbb{P}_n$, under linear or non-linear constraints imposed on $\mathbb{Q}$ by the model (see Bertail, 2005). The use of other pseudo-metrics instead of the Kullback divergence K has been suggested by Owen (1990) and many other authors. For example, the choice of relative entropy has been investigated by DiCiccio & Romano (1990), Jing & Wood (1995) and led to "Entropy econometrics" in the econometric field (see Golan *et al.*, 1996). Related results may be found in the probabilistic literature about divergence or the method of entropy in mean (see Csiszár, 1967, Liese & Vajda, 1987, Gamboa & Gassiat, 1997, Leonard, 2001, Broniatowski & Keziou, 2005). More recently, some generalizations of the empirical likelihood method have also been obtained by using Cressie-Read discrepancies (Baggerly 1998, Corcoran 1998 and led to some econometric extensions known as "generalized empirical likelihood" (Newey & Smith, 2003), even if the "likelihood" properties and in particular the Bartlett-correctability in these cases are lost (Jing

& Wood, 1995). Bertail, Harari & Ravaille (2005) have recently shown that Owen's (1988) 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. We call this method "empirical energy minimizers" by reference to the theoretical probabilistic literature on the subject (see Leonard, 2001 and references therein).

However, the previous results (including Bartlett-correction) 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.

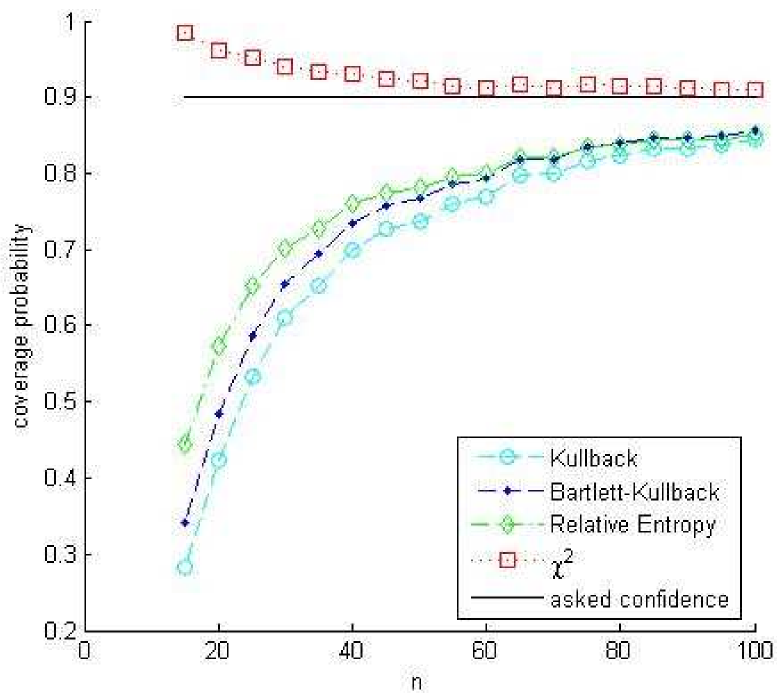


Figure 1: Coverage probability for different discrepancies

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 $\mathbb{R}^6$. The 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, Harari & Ravaille, 2005). 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. For very small sample size, Tsao (2004) 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.

One goal of this paper is to introduce and 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. The basic idea is to consider a family of barycenters of the Kullback divergence and the χ^2 divergence, called quasi-Kullback, defined by $(1-\varepsilon)K + \varepsilon\chi^2$ for $\varepsilon \in [0,1]$ and to minimize the dual expression of this divergence on the constraints. It can be seen as a quasi-empirical likelihood or a penalized empirical likelihood. 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. 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 (2004). On the other hand, for any fixed value of ε, it is possible to use the self-normalizing properties of the χ^2 divergence to obtain non-asymptotic exponential bounds for the error of the confidence intervals.

Exponential bounds for self-normalized sums have been obtained by several authors in the unidimensional case or can be derived from non-uniform Berry-Esséen or Cramer type bounds (see Shao, 1997, Wang & Jing, 1999, Christiakov & Götze, 1999, Jing, Shao & Wang, 2003). However, to our knowledge, non-asymptotic exponential bounds with **explicit constants** are only available in the unidimensional framework with symmetric distribution (Hoeffding, 1963 and Efron, 1969). In this paper, we obtain a generalization of this kind of bounds by using the symmetrization method developed by Panchenko (2003) as well as arguments taken from the literature on self-normalized process (see Bercu, Gassiat & Rio, 2002). Our bounds hold for any value of the parameter size q: one technical difficulty in this case is to obtain an explicit exponential bound for the smallest eigenvalue of the empirical variance. For this, we use chaining arguments from Barbe & Bertail (2004). These bounds are of interest in our quasi-empirical likelihood framework but also for self-normalized sums.

The layout of this paper is the following. In Part 2, we first recall some basic facts about convex integral functionals and their dual representation. As a consequence, we briefly state the asymptotic validity of the corresponding "empirical energy minimizers" in the case of M-estimators. We then focus in part 3 on a specific family of discrepancies, that we call quasi-Kullback divergences. These pseudo-distances enjoy several interesting convex duality and Lipschitz properties. This makes them an alternative method to empirical likelihood, easier to handle in practical situations. Moreover, for adequate choices of the weight ε, the corresponding empirical energy minimizers are shown to be Bartlett-correctable. In part 4, our main result claims that, for these discrepancies, it is possible to obtain exact asymptotic exponential bounds in a multivariate framework. A data-driven method for choosing the weight ε is also proposed. Part 5 gives some small sample simulation results and compares the confidence regions and their level for different discrepancies. The proofs of the main theorems are postponed to the appendix. There, some lemmas are also of interest for self-normalized sums.

2 Empirical φ^*-discrepancy minimizers

In this part, we extend the empirical likelihood method to a large class of φ^*-discrepancies, including Kullback and Cressie-Read discrepancies. We show that Owen's results (1990), stating a generalized version of Wilk's theorem and recent results of the econometric literature, are essentially linked to the convexity of the φ^*-discrepancies say I_{φ^*}.

2.1 Notations : φ^*-discrepancies and convex duality

We recall here a few notions on φ^*-discrepancies or divergences (Csiszár 1967). For more details on these metrics and some historical comments, see Rockafellar (1968, 1970 and 1971), Liese & Vajda (1987), Léonard (2001).

We consider a measured space $(\mathcal{X}, \mathcal{A}, \mathcal{M})$ where $\mathcal{M}$ is a space of signed measures.

For simplicity, $\mathcal{X}$ is a finite dimensional space endowed with $\mathcal{A}$ the Borel σ-algebra but general spaces may be considered at the price of additional technical measurability assumptions. It will be essential for applications to work with signed measures. 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 $\mu f = \int f d\mu$.

In the following, we consider φ, a convex function whose support $d(\varphi)$, defined as $\{x \in \mathbb{R},\ \varphi(x) < \infty\}$, is assumed to be non-void (φ is said to be proper).

We denote respectively $\inf d(\varphi)$ and $\sup d(\varphi)$, the extremes of this support. For every convex function φ, its convex dual or Fenchel-Legendre transform is given by

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

Recall that φ^* is then a semi-continuous inferiorly (s.c.i.) convex function. We define by $\varphi^{(i)}$ the derivative of order i of φ when it exists. From now on, we will assume the following assumptions for the function φ.

H1 φ is strictly convex and $d(\varphi)$ contains a neighborhood of 0 ;

H2 φ is twice differentiable on a neighborhood of 0 ;

H3 $\varphi(0) = 0$ and $\varphi^{(1)}(0) = 0$, $\varphi^{(2)}(0) > 0$, which implies that φ has an unique minimum at zero ;

H4 φ is differentiable on $d(\varphi)$, that is to say differentiable on $int\{d(\varphi)\}$, with 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 $m > 0$.

The assumptions in **H3** on the value of φ and $\varphi^{(1)}$ at 0 are simply normalization properties. Notice that the boundedness in **H5** hold as soon as $\varphi^{(1)}$ is itself convex (then $\varphi^{(2)}(x)$ is increasing and then on $\mathbb{R}^+$ by hypothesis, $\varphi^{(2)}(x) \geq \varphi^{(2)}(0) > 0$).

Let φ satisfies the hypotheses **H1, H2, H3**. Then, the Fenchel dual transform φ^* of φ also satisfies these hypotheses. The φ^*-discrepancy I_{φ^*} between $\mathbb{Q}$ and $\mathbb{P}$, where $\mathbb{Q}$ is a signed measure and $\mathbb{P}$ a signed positive measure, is defined as follows :

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

Under **H1-H3**, $I_{\varphi^*}(.,.)$ is a pseudo-metric (it is not symmetric in general).

It is easy to check that Cressie-Read discrepancies (Cressie & Read, 1988) fulfill these assumptions with, for $\kappa \in \mathbb{R}$,

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

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 ($\kappa \to 0$). The use of Cressie-Read in the framework of empirical likelihood goes back to Baggerly (1998) (see also Newey & Smith, 2003)

For us, the main interest of φ^*-discrepancies lies on the following duality representation, which follows from results of Borwein & Lewis (1991) on convex functional integrals (see also Léonard, 2001).

Theorem 1 *Let $\mathbb{P} \in \mathcal{M}$ be a probability 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 following qualification constraint holds,*

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

then, we have the dual equality :

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

*If φ 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*

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

Remark 1 *We obtain the results for a probability with a finite support for our applications. This clearly simplifies the statement of the dual equality but a similar result holds for general $\mathbb{P}$ under additional assumptions. It may be trivially (or easily) checked, provided that we work with signed measures. If φ is finite everywhere (that is $d(\varphi) = \mathbb{R}$) then (2) holds without Qual($\mathbb{P}$) for general $\mathbb{P}$, see Borwein & Lewis (1991), Bertail (2004) or Broniatowski & Keziou (2005).*

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 $\mathbb{P} \in\mathcal{M}$. 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 function at X_i. We will here consider that the parameter of interest $\theta \in \mathbb{R}^q$ is the solution of some M-estimation problem $E_{\mathbb{P}} f(X, \theta) = 0$, where f is now a regular differentiable function from $\mathcal{X}\times\mathbb{R}^q \to \mathbb{R}^r$. For simplicity, we now assume that f takes its value in $\mathbb{R}^q$, that is $r = q$ and that there is no over-identification problem. The over-identified case can be treated similarly by first reducing the problem to the strictly identified case (see Qin & Lawless, 1993).

For a given φ, we define, by analogy to Owen (1990, 2001), the quantity $\beta_n(\theta)$ as the minimum of the empirical φ^*-discrepancy, under the constraint $E_{\mathbb{Q}} f(X, \theta) = 0$, over all the measures $\mathbb{Q}$ dominated by $\mathbb{P}_n$ ($\mathbb{Q} \ll \mathbb{P}_n$). We define $\mathscr{C}_n(\eta)$ the corresponding random confidence region, where $\eta = \eta(\alpha)$ is a quantity such that

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

More precisely, consider

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

In the following, we denote

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

Considering this set of measures, instead of a set of probabilities, can be partially explained by Theorem 1. It establishes the existence of the solution of the dual problem for general signed measures, but in general not for probability measures.

The underlying idea of empirical likelihood and its extensions is actually a plug-in rule. Consider the functional defined by

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

that is, the minimization of a contrast under the constraints imposed by the model. 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 $E_{\mathbb{P}} f(X,\theta)=0$ at the true underlying probability $\mathbb{P}$, then clearly $M(\mathbb{P},\theta)=0$. A natural estimator of $M(\mathbb{P},\theta)$ for fixed θ is given by the plug-in estimator $M(\mathbb{P}_n,\theta)$, which is $\beta_n(\theta)/n$. This estimator can then be used to test $M(\mathbb{P},\theta)=0$ or, in a dual approach, to build confidence region for θ by inverting the test.

For $\mathbb{Q}$ in $\mathcal{M}_n$, the constraints can be rewritten as

$$(\mathbb{Q}-\mathbb{P}_n)f(.,\theta)=-\mathbb{P}_n f(.,\theta).$$

Using Theorem 1, we get the dual representation

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

Notice that $-x-\varphi(x)$ is a strictly concave function and that the function $\lambda \to \lambda' f$ is also concave. The parameter λ can be simply interpreted as the Kuhn & Tucker coefficient associated to the original optimization problem. From this representation of $\beta_n(\theta)$, we can now derive the usual properties of the empirical likelihood and its generalization. In the following, we will also use the notations

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

The following theorem states that generalized empirical likelihood essentially behaves asymptotically like a self-normalized sum. Links to self-normalized sum for finite n will be investigated in paragraph 4.

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 $E_{\mathbb{P}} f(X,\theta)=0$. Assume that $S=E_{\mathbb{P}} f(X,\theta)f(X,\theta)'$ is of rank q and that φ satisfies the hypotheses **H1-H4**. Assume that the qualification constraints Qual($\mathbb{P}_n$) hold. For any α in $]0,1[$, set $\eta = \frac{\varphi^{(2)}(0)\chi_q^2(1-\alpha)}{2}$, where $\chi_q^2(.)$ is the χ^2 distribution quantile. Then*

$\mathscr{C}_n(\eta)$ *is a convex asymptotic confidence region with*

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

The proof of this theorem starts from the convex dual-representation and follows the main arguments of Bertail, Harari-Kermadec & Ravaille (2005) and Owen (2001) for the case of the mean. It is left to the reader.

Remark 2 *As noticed earlier, 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 Two basic examples

We illustrate Theorem 1 by reexamiming the case of the Kullback and χ^2 discrepancies, which lead respectively to the empirical likelihood method and the Generalized Method of Moments (GMM).

2.3.1 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 $K(\mathbb{Q}, \mathbb{P}) = -\int \log(\frac{d\mathbb{Q}}{d\mathbb{P}})d\mathbb{P}$, the dual program obtained in (3) becomes, for the admissible θ,

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

As Bertail (2003, 2004) points out, this quantity is itself a parametric log-likelihood ratio indexed by the parameter λ (to test $\lambda = 0$). It can also be seen as a dual likelihood in the sense of Mykland (1995). It is then easy to show that $2\beta_n(\theta)$ is asymptotically $\chi^2(q)$ when $n \to \infty$, if the variance of $f(X, \theta)$ is definite. As a parametric likelihood indexed by λ, it is also Bartlett-correctable (DiCiccio *et al.*, 1991). Using a duality point of view, the proof of the Bartlett-correctability is almost immediate, see Mykland (1995) and Bertail (2004). For a general discrepancy, the dual form is not a likelihood and may not be Bartlett-correctable, see DiCiccio *et al.* (1991) and Jing & Wood (1995). We will latter propose a family of discrepancies, the Quasi-Kullback indexed by some smoothing parameter ε, which still have this property for some specific choice of ε.

Moreover, we necessarily have the $q_i's > 0$ and $\sum_{i=1}^{n} q_i = 1$, so that the qualification constraint essentially means that 0 belongs to the convex hull of the $f(X_i, \theta)$. Only the θ's which satisfy this constraint are of interest to us ; asymptotically, this is by no mean a restriction, unless we have for some specific configuration of the realization $int\{\theta \backslash\ 0 \in conv(f(X_1, \theta), ..., f(X_n, \theta))\} = \emptyset$, where $conv(., ., .)$ is the convex hull of the points.

2.3.2 GMM and χ^2 discrepancy

The particular case of the χ^2 corresponds to $\varphi_2(x) = \varphi_2^*(x) = \frac{x^2}{2}$. The Kuhn & Tucker multiplier λ, and consequently the value of $\beta_n(\theta)$ at any point θ, can be explicitly calculated. Indeed, we get easily that $\lambda_n = S_n^{-2}\overline{f}_n$ so that, by Theorem 1, the minimum is attained at $\mathbb{Q}_n^* = \sum_{i=1}^{n} q_{i,n}\delta_{X_i}$ with

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

and

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

which is exactly the square of a self-normalized sum which typically appears in the Generalized Method of Moments (GMM). Notice that $\mathbb{Q}_n^*$ is a signed measure, not a probability.

This short calculus also shows that if we want to force our measure $\mathbb{Q} \in \mathcal{M}_n$ to be a probability measure, then the qualification constraints Qual($\mathbb{P}_n$) of Theorem 1 can not be fulfilled. Indeed, imposing the additional constraints $q_i \geq 0$, $\sum_{i=1}^{n} q_i = 1$, implies that the dual problem has no solution. This explains why, for some discrepancies, we have to work with signed measure and not probability measure. The drawback is that, in opposition to the Kullback discrepancy, we may charge positively some region outside of the convex hull of the points, yielding bigger (that is too conservative) confidence region. See the simulation results of Bertail *et al.* (2004). However, as noticed in the introduction, the results of Tsao (2004) shows 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.

Remark 3 *If S_n^2 is of rank $l < q$, notice that we still have the duality relationship :*

$$\beta_n(\theta) = n \sup_{\lambda \in ^q} \left\{ -\lambda' \frac{1}{n} \sum_{i=1}^{n} f(X_i, \theta) - \frac{1}{2n} \lambda' S_n^2 \lambda \right\}.$$

Write $S_n^2 = R' \begin{pmatrix} \Delta_n & 0 \\ 0 & 0 \end{pmatrix} R$, where Δ_n is inversible 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 $R_b \in \mathfrak{M}_{q-l,q}(\mathbb{R})$. We denote $\overline{f}_n = (\overline{f}_{n,1}, \cdots, \overline{f}_{n,q})'$. Since for all $j = 1, \cdots, q-l$, we can write

$$0 \le (R_b \overline{f}_n)_j^2 \le \frac{1}{n} \sum_{i=1}^{n} \left(\sum_{k=1}^{q} R_{b_{j,k}} f_k(X_i, \theta) \right)^2 \le \left(R_b S_n^2 R_b \right)_{l+j,l+j} = 0.$$

We deduce that $R_b \overline{f}_n = 0$. Then, the duality relationship becomes

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

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-Kullbacks and Bartlett-correctability

The main underlying idea for considering these functions is that we want to keep the good properties of Kullback discrepancy and to avoid some algorithmic problems linked with the behavior of the log of Kullback discrepancy in the neighborhood of 0. This kind of discrepancies is actually currently used in the convex optimization literature (see for instance Auslender et al., 1999) because the resulting optimization algorithm leads to efficient tractable interior point solutions.

3.1 Quasi-Kullback: definitions

For $\varepsilon \in]0;1]$ and $x \in]-\infty;1[$ let,

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

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 $\mathbb{R}$:

$$K^*_\varepsilon(x) = -\frac{1}{2} + \frac{(2\varepsilon - x - 1)\sqrt{1 + x(x + 2 - 4\varepsilon)} + (x+1)^2}{4\varepsilon} - (\varepsilon - 1)\log\frac{2\varepsilon - x - 1 + \sqrt{1 + x(x + 2 - 4\varepsilon)}}{2\varepsilon}.$$

The second order derivative of K_ε is bounded from below : $K^{(2)}_\varepsilon(x) \geq \varepsilon$. Moreover, the second order derivative of K^*_ε is bounded both from below and above : $0 \leq K^{*(2)}_\varepsilon(x) \leq 1/\varepsilon$. These controls ensure a quick and regular convergence of the algorithms based on such discrepancies.

In addition, another algorithmic improvement is obtained in comparison with empirical likelihood. The Kullback must be approached for practical optimization, for instance by replacing the log by a pseudo log, see section 12.3 in Owen (2001). Since the domain of K^*_ε is $\mathbb{R}$, the quasi-Kullback discrepancy can be used exactly. Thus, the use of quasi-Kullback discrepancy in the empirical likelihood method, the "quasi-empirical likelihood" may be seen as a "regularized" empirical likelihood.

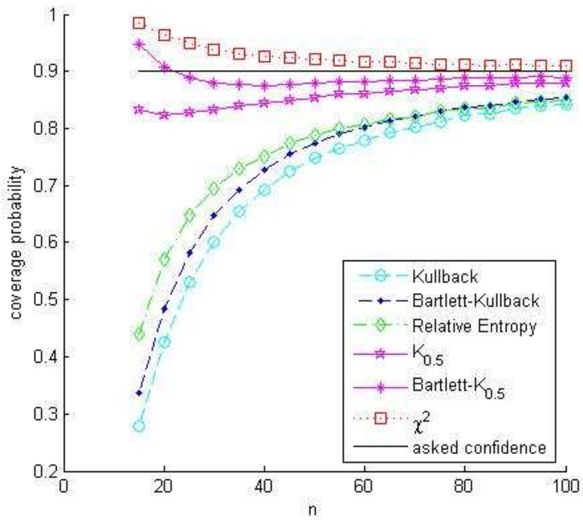


Figure 2: Cover probabilities and Quasi-Kullback

Figure 2 illustrates the improvements coming from the use of Quasi-Kullbacks. 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.

3.2 Bartlett-correctability

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

This choice of ε is probably not optimal but considerably simplifies the proof. An attentive reading of Corcoran (2001) shows that, if ε is small enough, the statistic is Bartlett-correctable. Unfortunately, as our discrepancy depend 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 $o(n^{-1})$.

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

Another interesting feature of quasi-Kullback discrepancies is that the control of the second order derivatives allows the behavior of $\beta_n(\theta)$ to be linked to that of self-normalized sums. We thus can get exponential bounds for the quantities of interest. Some of the bounds that we propose here for self-normalized sums are new and of interest by themselves. These bounds may be quite easily obtained in the symmetric case (that is for random variables having a symmetric distribution) and are well-known in the unidimensional case.

Self-normalized sums have recently given rise to an important literature : see for instance Jing & Wang (1999), Götze & Chistyakov (2003) or Bercu, Gassiat & Rio (2002) for self-normalized processes. Unfortunately, except in the unidimensional symmetric case, these bounds are not universal and depend on higher order moments, $\tilde{\gamma}_3 = E_{\mathbb{P}}|S^{-1}f(X_i,\theta)|^3$ or even an higher moment condition : $\tilde{\gamma}_{10/3} = E_{\mathbb{P}}|S^{-1}f(X_i,\theta)|^{10/3}$. Actually, uniform bounds in $\mathbb{P}$ are impossible to obtain, otherwise this would contradict Bahadur & Savage (1956)'s result on the non-existence of uniform confidence region over large class of probabilities, see Romano & Wolf (2000) for related results.

In the general non-symmetric case, for $q = 1$, if $\tilde{\gamma}_{10/3} < \infty$, for some $A \in \mathbb{R}$ and some $a \in]0,1[$, the result of Jing & Wang (1999) lead to

$$\Pr\left(\frac{n}{2}\overline{f}_n^2/S_n^2 \geq \varepsilon\eta\right) = \chi_1^2(\varepsilon\eta) + A\gamma_{10/3}n^{-1/2}e^{-a\varepsilon\eta}. \tag{4}$$

However the constants A and a are not explicit and the bound is of no practical use. In the non-symmetric case our bounds are worse than (4) as far as the power in the exponential and the control of the approximation by a χ^2 distribution are concerned, but entirely explicit.

Theorem 4 *Let* $(Z_i)_{i=1,\cdots,n}$ *be i.i.d. sample in* $\mathbb{R}^q$ *with probability* $\mathbb{P}$*. Note that* $\overline{Z}_n = \frac{1}{n}\sum_{i=1}^n Z_i$, $S_n^2 = \frac{1}{n}\sum_{i=1}^n Z_i Z_i'$ *and* $S^2 = E_{\mathbb{P}} Z_1 Z_1'$ *is of rank* q*. Then the following inequalities hold,* ***for finite*** $n > q$ *and for* $u \le nq$*,*

a) *if* Z_1 *has a symmetric distribution, without any moment assumption,*

$$\Pr\left(n\overline{Z}_n S_n^{-2}\overline{Z}_n \ge u\right) \le 2qe^{-\frac{u}{2q}}; \tag{5}$$

b) *for general distribution of* Z_1 *with kurtosis* $\tilde{\gamma}_4 < \infty$*,*

$$\begin{aligned}\Pr\left(n\overline{Z}_n S_n^{-2}\overline{Z}_n \ge u\right) &\le \inf_{a>1}\left\{2qe^{1-\frac{u}{2q(1+a)}} + C(q)\, n^{3\widetilde{q}}\tilde{\gamma}_4^{-\widetilde{q}} e^{-\frac{n}{\gamma_4(q+1)}\left(1-\frac{1}{a}\right)^2}\right\} \\ &\le \inf_{a>1}\left\{2qe^{1-\frac{u}{2q(1+a)}} + C(q)\, n^{3\widetilde{q}} e^{-\frac{n}{\tilde{\gamma}_4(q+1)}\left(1-\frac{1}{a}\right)^2}\right\}\end{aligned} \tag{6}$$

with $\widetilde{q} = \frac{q-1}{q+1}$, $\tilde{\gamma}_4 = E_{\mathbb{P}}(\|S^{-1}Z_1\|_2^4)$ *and* $C(q) = \frac{(2e\pi)^{2\widetilde{q}}(q+1)}{2^{2/(q+1)}(q-1)^{3\widetilde{q}}} \le \frac{(2e\pi)^2(q+1)}{(q-1)^{3\widetilde{q}}} \le 18.$

Moreover for $nq \le u$*, we have*

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

The proof is postponed to Appendix A.3. The exponential inequality (5) is classical in the unidimensional case. This bound is universal for symmetric laws. We generalize it to the multidimensional case by using simple diagonalization arguments leading to a sum of q self-normalized sums. In the general multidimensional framework, the main difficulty is actually to keep the self-normalized structure when symmetrizing the original sum. For this we use a multidimensional extension of a symmetrization lemma by Panchenko (2003). 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 (6) is essentially due to this control.

Remark 4 *In the best case, past studies give 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 is 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* $n \geq \frac{2\varepsilon\eta}{q}$*,*

$$\Pr(\theta \notin \mathscr{C}_n(\eta)) = \Pr(\beta_n(\theta) \geq \eta) \leq \Pr\left(n\overline{f}_n S_n^{-2}\overline{f}_n \geq 2\varepsilon\eta\right). \tag{7}$$

Else if $n > \frac{2\varepsilon\eta}{q}$, $\Pr(\theta \notin \mathscr{C}_n(\eta)) = 0$.
Then bounds (5) and (6) may be used with $u = 2\varepsilon\eta$ *and* $Z_i = f(X_i, \theta)$*.*

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

$$\begin{aligned}\beta_n(\theta) &= \sup_{\lambda \in \mathbb{R}^q}\left\{-n\lambda'\overline{f}_n - \frac{1}{2}\sum_{i=1}^{n}(\lambda' f(X_i,\theta))^2 K_\varepsilon^{(2)}(t_{i,n})\right\} \\ &\leq \sup_{\lambda \in \mathbb{R}^q}\left\{-n\lambda'\overline{f}_n - \frac{1}{2}\sum_{i=1}^{n}(\lambda' f(X_i,\theta))^2 \varepsilon\right\}. \end{aligned} \tag{8}$$

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 (8) 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

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

◊

Remark 5 *In Hjort, McKeague, and Van Keilegom (2004), convergence of empirical likelihood is investigated when* q *is allowed to increase with* n*. They show that convergence to a Chi-square distribution still holds when* $q = O(n^{\frac{1}{3}})$ *as* n *tends to infinity. Our bounds shows that even if* $q = o\left(n/log(n)\right)$*, 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 on the following graph.*

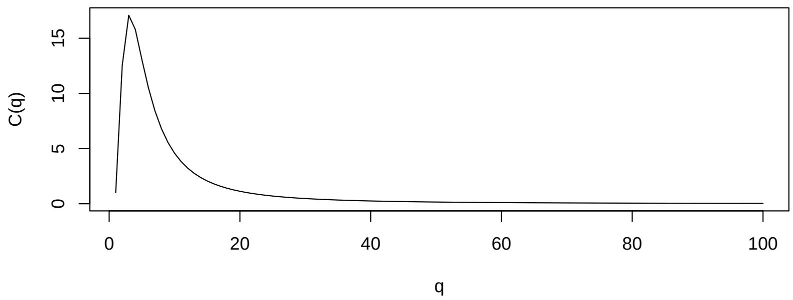


Figure 3: Value of $C(q)$ as a function of q

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 (2004), which gives a lower bound for the attainable levels.

5 Discussion and simulation results

5.1 Non-asymptotic comparisons

For $\varepsilon = 1$, that is for the χ^2 discrepancy, the inequality (7) becomes an equality. In the following table 1, we tabulate some values of η corresponding to a given confidence level $1-\alpha$, for different γ_4 and n for a unidimensional model ($q = 1$). The values of η corresponding to K_ε with $\varepsilon \neq 1$ are easily obtained by multiplying the values in the table by $1/\epsilon$.

- The column "Asymptotic" corresponds to η equal to the $(1-\alpha)$-quantile of the χ^2_1 distribution.

- The column "Symmetric bound" corresponds to η obtained by inverting the exponential inequality in the symmetric case, that is $\eta = -q\ln(\frac{\alpha}{2q})$.

- The next column NS, for "Non-symmetric", is obtained by inverting the general exponential bound for $\gamma_4 = 3$ (that corresponds to the kurtosis of standard gaussian distribution) and for two values of n, 50 and 200.

- The last column is similar to the third, but for $\gamma_4 = 5.4$ (that corresponds to our gaussian scale mixture).

	Asymptotic	Symmetric	NS $\gamma_4 = 3$		NS $\gamma_4 = 5.4$	
Confidence	χ^2	bound	$n = 50$	$n = 200$	$n = 50$	$n = 200$
50%	0.46	1.4	7.99	6.05	9.86	6.62
90%	**2.71**	**3.0**	16.3	10.6	**26.7**	12.0
95%	3.84	3.7	21.5	12.7	44.1	14.7
99%	6.64	5.3	40.4	18.0	10^4	21.7

Table 1: Values of η for $q = 1$.

We notice 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 graphics, we build confidence intervals for the mean of unidimensional data. We simulated 50 i.i.d. centered gaussian scale mixture r.v.'s : 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. The figure shows the profile quasi-likelihood $\beta_n(\theta)$ for different values of ε, the bottom right graphic correspond to $\varepsilon = 1$. In addition to the profile quasi-likelihood, we indicate the bounds corresponding to 90% confidence intervals ($1 - \alpha = 0.9$) using respectively the asymptotic approximation, the symmetric bound and the general bound (NS) with the true kurtosis (5.4) and an estimated kurtosis.

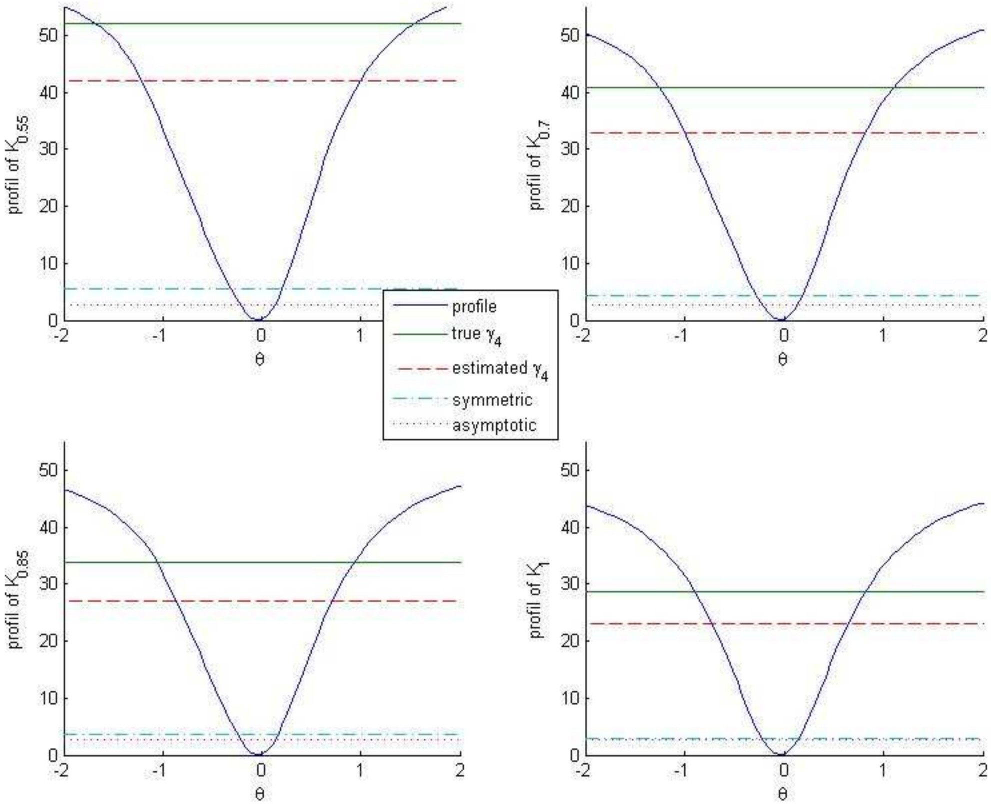


Figure 4: Discrepancy profile and 90% confidence levels

For a given sample size n and confidence $1-\alpha$, 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 interval as ε increases is more delicate to understand. The profile likelihood gets wider but the η's corresponding to the symmetric and NS bounds decrease like $1/\varepsilon$. These two behaviors have contradictory effects on the confidence intervals $\mathscr{C}_n(\eta)$. On the figure 4, for $\alpha = 0.1$, $q = 1$ and our simulated data, the effect of the decrease of η dominates : the confidence intervals get smaller when ε increases. In higher dimension or for a smaller α, the two contradictory effects could be balanced.

In figure 5, we build confidence regions for the mean of multi-dimensional ($\mathbb{R}^2$) data, for 2 sizes (500 and 2000) and 2 distributions : a couple of independent gaussian scale mixtures and the distribution $0.01 \cdot \delta_{(10,10)} + \frac{0.81}{4} \cdot \big(\delta_{(-1,-1)} + \delta_{(-1,1)} + \delta_{(1,-1)} + \delta_{(1,1)}\big) +$

$\frac{0.09}{2} \cdot \left(\delta_{(-1,10)} + \delta_{(1,10)} + \delta_{(10,-1)} + \delta_{(10,1)}\right)$, that will be referred as discrete distribution. We give in figure 5 the corresponding 90% confidence regions, using respectively the asymptotic approximation, the symmetric bound and the general bound (NS) with the true kurtosis.

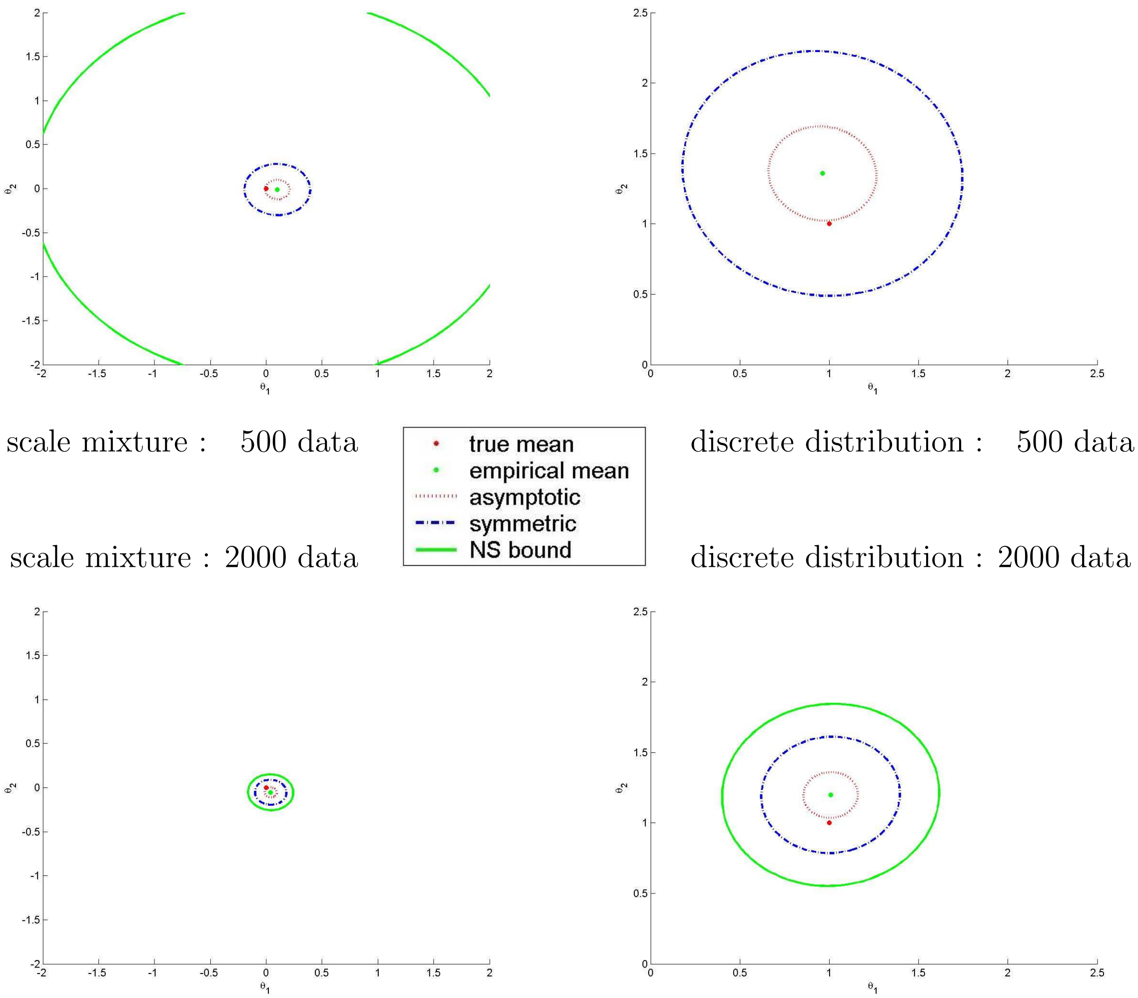


Figure 5: Confidence regions, for 2 distributions and 2 data sizes.

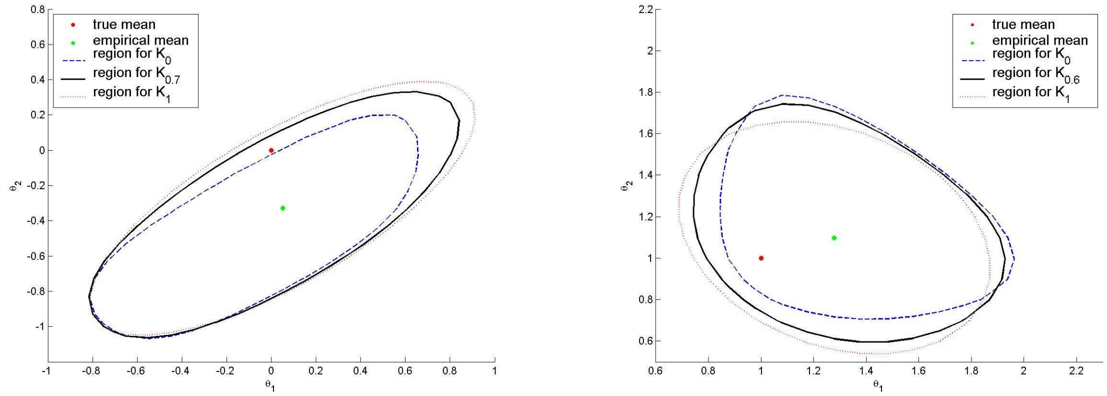
For small sample size, as expected, the confidence region obtained with NS bound is quite large (for our discrete data and $n = 500$, the region is too large to be represented on the figure) with a coverage probability close to 1. On the contrary, the asymptotic confidence regions are small but when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

We conclude from these simulations that, on the one hand, if the asymptotic and NS confidence regions are not too far from each other then we may trust the asymptotic behavior for a coverage point of view. On the other hand, to protect oneself against exotic distributions, the use of NS bound is justified.

5.2 Adaptative 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 bigger 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 too conservative. 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 & Smith (2002) for argument 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. Moreover, 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. The figure 6 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 $\mathbb{R}^2$.



scale mixture : 15 data exponential distribution : 25 data

Figure 6: Asymptotic confidence regions for data driven K_ε.

The figure 7 represents the coverage probability obtained by Monte-Carlo simulations (10 000 repetitions) for K_ε with data driven ε and different sample sizes n. Some curves from figure 1 giving the coverage probability of previously available methods are recalled for comparison.

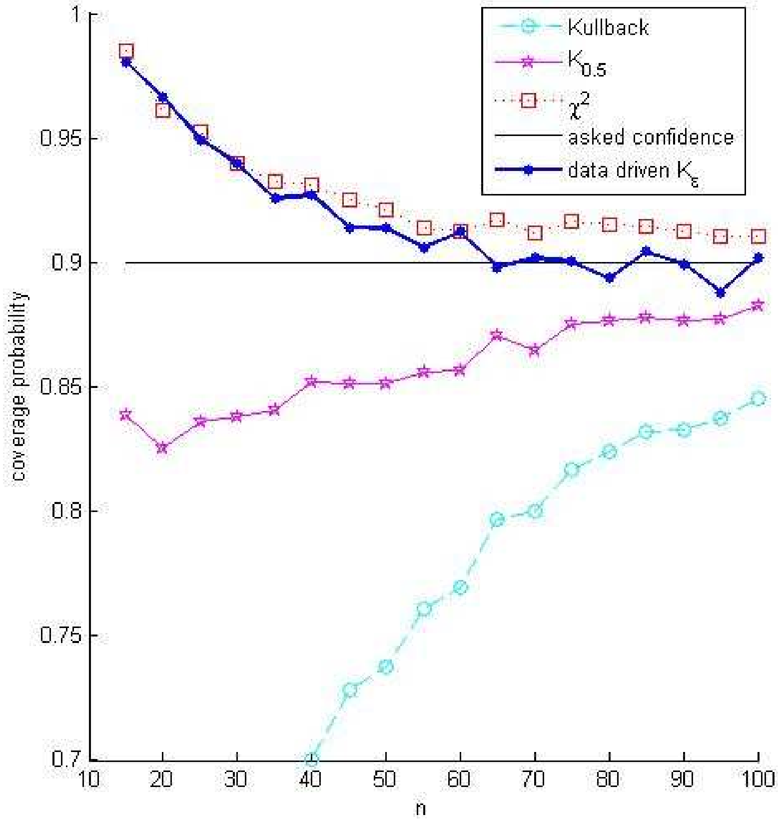


Figure 7: Coverage probability for different data sizes n for data-driven ε.

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 $n = 20$. 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.

Note that when, for all values of ε, the cross-validation estimate of the coverage probability is smaller than the targeted confidence, the distribution may be "exotic". In such a case, the NS bound should be considered.

The simulations and graphics have been computed with Matlab : algorithms are available from the authors on request . The Monte-Carlo simulations of figure 7 have been carried out simulatively on 18 computers with 2.5 GHz processors and took

18*200 hours of computation time.

A Proofs of the main results

A.1 Proof of theorem 3

Write $\beta_n^\varepsilon(\theta)$ for the 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, Hall & Romano [16] (see also Bertail, 2005), 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 $\bar{F}_Z(.) = \mathbb{P}(Z > .)$, when $Z \sim \mathbb{P}$. This equality implies in particular that

$$\bar{F}_{T_n^0}(\eta - n^{-\frac{3}{2}}) = \bar{F}_{\chi^2(q)}(\eta) + \mathcal{O}(n^{-\frac{3}{2}}). \tag{9}$$

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 with (9)

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

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 (4) 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 Some bounds for self-normalized sums

Lemma 1 (Extension of Panchenko, 2003 Corollary 1) *Let Γ be the unit circle of $\mathbb{R}^q$, $\Gamma = \{\lambda \in \mathbb{R}^q,\ \|\lambda\|_{2,q} = 1\}$. Let $(Z_i)_{1\leq i\leq n}$ and $(Y_i)_{1\leq i\leq n}$ be i.i.d. centered random vectors in $\mathbb{R}^q$ with $(Z_i)_{1\leq i\leq n}$ independent of $(Y_i)_{1\leq i\leq n}$. We denote for all random vector W with probability $\mathbb{P}$: $S_n^2(W) = \frac{1}{n}\sum_i^n W_i W_i'$ and $S^2 = E_{\mathbb{P}}(WW')$.*
If there exists $D > 0$ and $d > 0$ such that, for all $u \geq 0$,

$$\Pr\left(\sup_{\lambda\in\Gamma}\left(\frac{\sqrt{n}\lambda'(\overline{Z}_n - \overline{Y}_n)}{\sqrt{\lambda' S_n^2(Z-Y)\lambda}}\right) \geq \sqrt{u}\right) \leq De^{-du},$$

then, for all $u \geq 0$,

$$\Pr\left(\sup_{\lambda\in\Gamma}\frac{\sqrt{n}\lambda'\overline{Z}_n}{\sqrt{\lambda' S_n^2(Z)\lambda + \lambda' S^2\lambda}} \geq \sqrt{u}\right) \leq De^{1-du}. \tag{10}$$

Proof. In the unidimensional case, this result reduces to Corollary 1 of Panchenko (2003) [32]. In the multidimensional case, this is an extension of Panchenko (2003)'s Lemma 1[32]. Denote

$$A_n(Z) = \sup_{\lambda\in\Gamma}\sup_{b>0} E_Y^Z\left\{4b(\lambda'(\overline{Z}_n - \overline{Y}_n) - b\lambda' S_n^2(Z-Y)\lambda)\right\}$$
$$C_n(Z,Y) = \sup_{\lambda\in\Gamma}\sup_{b>0}\left\{4b(\lambda'(\overline{Z}_n - \overline{Y}_n) - b\lambda' S_n^2(Z-Y)\lambda)\right\}.$$

By Jensen inequality, we have Pr-almost surely

$$A_n(Z) \leq E_Y[C_n(Z,Y)|Z]$$

and, for any convex function Φ, by Jensen inequality, we also get

$$\Phi(A_n(Z)) \leq E_Y[\Phi(C_n(Z,Y))|Z].$$

We obtain

$$E_Z(\Phi(A_n(Z))) \leq E(\Phi(C_n(Z,Y))). \tag{11}$$

Now remark that

$$\begin{aligned} A_n(Z) &= \sup_{\lambda\in\Gamma}\sup_{b>0}\left\{4b\left(\lambda'\overline{Z}_n - b\lambda' S_n^2(Z)\lambda - b\lambda' S^2\lambda\right)\right\} \\ &= \sup_{\lambda\in\Gamma}\frac{\lambda'\overline{Z}_n}{\sqrt{\lambda' S_n^2(Z)\lambda + \lambda' S^2\lambda}} \end{aligned}$$

and

$$C_n(Z,Y) = \sup_{\lambda\in\Gamma}\frac{\lambda'(\overline{Z}_n - \overline{Y}_n)}{\sqrt{\lambda' S_n^2(Z-Y)\lambda}}.$$

Now, notice that $\sup_{\lambda\in\Gamma}\frac{\lambda'\overline{Z}_n}{\sqrt{\lambda' S_n^2\lambda}} > 0$ and apply the same arguments as Corollary 1's proof of Panchenko [32] applied to inequality (11) to obtain the result.

$\diamondsuit$

We now extend a result of [5], which controls the behaviour of the smallest eigenvalue of the empirical variance. In the following, for a given symmetric matrix A, we denote $\mu_1(A)$ its smallest eigenvalue.

Lemma 2 *Let $(Z_i)_{i=1,\cdots n}$ be i.i.d. random vectors in $\mathbb{R}^q$ with common mean 0. Denote $S^2 = E(Z_1 Z_1')$, $0 < \gamma_4 = E(\|Z_1\|_2^4) < +\infty$ and $\widetilde{q} = \frac{q-1}{q+1}$. Then, for any $1 \leq q < n$ and $0 < u \leq \mu_1(S^2)$,*

$$\Pr\left(\mu_1(S_n^2) \leq u\right) \leq C(q)\,\frac{n^{3\widetilde{q}}\mu_1(S^2)^{2\widetilde{q}}}{\gamma_4^{\widetilde{q}}}e^{-\frac{n(\mu_1(S^2)-u)^2}{\gamma_4(q+1)}} \wedge 1,$$

with

$$C(q) = \pi^{2\widetilde{q}}(q+1)e^{2\widetilde{q}}(q-1)^{-3\widetilde{q}}2^{2\widetilde{q}-\frac{2}{q+1}} \tag{12}$$

$$\leq 4\pi^2(q+1)e^2(q-1)^{-3\widetilde{q}}. \tag{13}$$

Remark 6 *The value of $C(q)$ could certainly be improved. The term $\pi^{2\widetilde{q}}$ essentially comes from a basic bound for the number of caps of diameter ε needed to cover a half unit-sphere S_{q-1}, say $N(S_{q-1},\varepsilon)$. We use the bound $N(S_{q-1},\varepsilon) \leq \pi^{q-1}\varepsilon^{-(q-1)}$. There is a huge bibliography in convex geometry about covering numbers of the sphere. For instance Boroczky & Wintsche (2003) give a bound on the number of sphere (for the euclidian geometry on the sphere) needed to cover the sphere. We can deduce from*

Boroczky & Wintsche (2003) the following bound for $N(S_{q-1},\varepsilon)$: when $\varepsilon \leq Arcos(\frac{1}{\sqrt{q}})$, for $q \geq 2$, there exists c an universal constant such that

$$\varepsilon^{-(q-1)} \leq N(S_{q-1},\varepsilon) \leq c\cos(\varepsilon)\sin(\varepsilon)^{-(q-1)}(q-1)^{3/2}\ \log(1+(q-1)\cos(\varepsilon)^2)$$

Using the fact that for $x>0$, $\frac{2}{\pi}x \leq \sin(x) \leq x$ and $\cos(\varepsilon)^2 \leq \frac{1}{q}$, we get the more friendly bound

$$\varepsilon^{-(q-1)} \leq N(S_{q-1},\varepsilon) \leq c\left(\frac{\pi}{2\varepsilon}\right)^{q-1}(q-1)^{3/2}\ \log\left(1+\frac{q-1}{q}\right).$$

However, an explicit value for c is not clear to us.

Proof. This proof is adapted from the proof of [5] and makes use of some idea of Bercu-Gassiat-Rio [3]. In the following, we denote by $\mathcal{S}_{q-1}$ the northern hemisphere of the sphere.

We first have by a truncation argument and applying Markov's inequality on the last term in the inequality (see the proof of Barbe and Bertail [5], lemma 4), for every $M>0$, $\Pr\left(\mu_1(\sum_{i=1}^n Z_iZ_i') \leq t\right)$ is less than

$$\Pr\left(\inf_{v\in\mathcal{S}_{q-1}}\sum_{i=1}^n(v'Z_i)^2 \leq t,\ \sup_{i=1,\dots,n}||Z_i||_2 \leq M\right)+n\frac{\gamma_4}{M^4} \tag{14}$$

We call the first term on right side I.
Notice that by symmetry of the sphere, we can always work with the northern hemisphere of the sphere rather than the sphere. Notice first, that, if $\sup_{i=1,\dots,n}||Z_i||_2 \leq M$, then for u,v in $\mathcal{S}_{q-1}$, we have

$$|\sum_{i=1}^n(v'Z_i)^2-\sum_{i=1}^n(u'Z_i)^2| \leq 2n||u-v||M^2.$$

Thus if u and v are apart of $t\eta/(2nM^2)$ then $|\sum_{i=1}^n(v'Z_i)^2-\sum_{i=1}^n(u'Z_i)^2| \leq \eta t$. Now let $N(\mathcal{S}_{q-1},\varepsilon)$ be the smallest number of caps of radius ε centered at some points on S_{q-1} (for the $||.||_2$ norm) needed to cover $\mathcal{S}_{q-1}$ (the half sphere). Following the same arguments as [5], we have, for any $\eta>0$,

$$I \leq N(\mathcal{S}_{q-1},\frac{t\eta}{2nM^2})\max_{u\in S_{q-1}}\Pr\left(\sum_{i=1}^n(u'Z_i)^2 \leq (1+\eta)t\right).$$

The proof is now divided in three steps, i) control of $N(\mathcal{S}_{q-1},\frac{t\eta}{2nM^2})$ ii) control of the maximum over $\mathcal{S}_{q-1}$ of the last expression in I, iii) optimization over all the free parameters.

i) On the one hand, we have

$$N(\mathcal{S}_{q-1}, \varepsilon) \leq b(q)\varepsilon^{-(q-1)} \vee 1, \tag{15}$$

with, for instance, $b(q) \leq \pi^{q-1}$. Indeed, following [5], the northern hemisphere can be parameterized in polar coordinates, realizing a diffeomorphism with $S^{q-2} \times [0, \pi]$. Now proceed by induction, notice that for $q = 2$, $\mathcal{S}_{q-1}$, the half circle can be covered by $[\pi/2\varepsilon] \vee 1 + 1 \leq 2([\pi/2\varepsilon] \vee 1) \leq \pi/\varepsilon \vee 1$ caps of diameter 2ε, that is, we can choose the caps with their center on a $\varepsilon-$grid on the circle. Note that this is not a good bound for q=2 since in that case the overlapping of the caps is ε. Now, by induction we can cover the cylinder $S^{q-2} \times [0, \pi]$ with $[\pi/2\varepsilon\ (\pi)^{q-2}/\varepsilon^{q-2}] \vee 1 + 1 \leq \pi^{q-1}/\varepsilon^{q-1}$ intersecting cylinders which in turn can be mapped to region belonging to caps of radius ε, covering the whole sphere (this is still a covering because the mapping from the cylinder to the sphere is contractive).

ii) On the other hand, for all $t > 0$, we have by exponentiation and Markov's inequality, and independence of (Z_i), for any $\lambda > 0$

$$\max_{u \in \mathcal{S}_{q-1}} \Pr\left(\sum_{i=1}^{n} u' Z_i Z_i' u \leq t\right) \leq e^{\lambda t} \max_{u \in \mathcal{S}_{q-1}} (\mathbb{E}(e^{-\lambda u' Z_1 Z_1' u}))^n.$$

Now, using the classical inequalities, $\log(x) \leq x - 1$ and $e^{-x} - 1 \leq -x + x^2/2$, both valid for $x > 0$, we have

$$\begin{aligned}
\max_{u \in \mathcal{S}_{q-1}} (E(e^{-\lambda u' Z_1 Z_1' u}))^n &\leq \max_{u \in \mathcal{S}_{q-1}} \exp n(E(e^{-\lambda u' Z_1 Z_1' u} - 1)) \\
&\leq \max_{u \in \mathcal{S}_{q-1}} \exp n \left(E(-\lambda u' Z_1 Z_1' u) + \frac{\lambda^2}{2} E\left(u' Z_1 Z_1' u)^2\right)\right) \\
&\leq \max_{u \in \mathcal{S}_{q-1}} \exp n \left(-\lambda u' S^2 u + \frac{\lambda^2}{2}\gamma_4\right) \\
&\leq e^{\frac{\lambda^2}{2} n\gamma_4} e^{-\lambda n \min_{u \in \mathcal{S}_{q-1}} u' S^2 u} \\
&= e^{\frac{\lambda^2}{2} n\gamma_4 - \lambda n \mu_1(S^2)}.
\end{aligned} \tag{16}$$

iii) From (16) and (15), we deduce that, for any $t > 0, \lambda > 0,\ \eta > 0$,

$$I \leq b(q)(\frac{2nM^2}{t\eta})^{q-1} e^{\lambda(1+\eta)t + \frac{\lambda^2}{2} n\gamma_4 - \lambda n \mu_1(S^2)}.$$

Optimizing the expression $\exp(-(q-1)log(\eta) + \lambda\eta t)$ in $\eta > 0$, yields immediately , for any $t > 0$, any $M > 0$, any $\lambda > 0$

$$I \leq b(q)\left(\frac{2enM^2\lambda}{q-1}\right)^{q-1} e^{\lambda\left(t - n\mu_1(S^2)\right) + n\lambda^2\gamma_4/2}.$$

The infimum in λ in the exponential term is attained at $\lambda = \frac{\mu_1(S^2)-\frac{t}{n}}{\gamma_4}$, provided that $0 < t < n\,\mu_1(S^2)$. Therefore, for these t and all $M > 0$, we get $\Pr(\mu_1(\sum_{i=1}^n Z_i Z_i') \le t)$ is less than

$$b(q)\left(\frac{2enM^2\mu_1(S^2)}{\gamma_4(q-1)}\right)^{q-1}\exp\left(-\frac{n}{2\gamma_4}\left(\mu_1(S^2)-\frac{t}{n}\right)^2\right)+n\frac{\gamma_4}{M^4}.$$

We now optimize in $M^2 > 0$ and the optimum is attained at

$$M_*^2 = \left(\frac{2n\gamma_4}{(q-1)b(q)}\right)^{\frac{1}{q+1}}\left(\frac{2en}{q-1}\frac{\mu_1(S^2)}{\gamma_4}\right)^{-\frac{(q-1)}{q+1}}\exp\left(\frac{n(\mu_1(S^2)-\frac{t}{n})^2}{2\gamma_4(q+1)}\right),$$

yielding the bound

$$\Pr\left(\mu_1\left(\sum_{i=1}^n Z_i Z_i'\right)\le t\right)\le \tilde{C}(q)\, n^{3\frac{q-1}{q+1}}\mu_1(S^2)^{\frac{2(q-1)}{q+1}}\gamma_4^{-\frac{q-1}{q+1}}\exp\left(-\frac{n\left(\mu_1(S^2)-\frac{t}{n}\right)^2}{\gamma_4(q+1)}\right),$$

with

$$\tilde{C}(q) = b(q)^{\frac{2}{q+1}}(q+1)e^{\frac{2(q-1)}{q+1}}(q-1)^{-3\frac{q-1}{q+1}}2^{\frac{2q-4}{q+1}}.$$

Using $b(q) \le \pi^{q-1}$ we obtained $C(q)$, which is majorized by the simpler bound (for large q this bound will be sufficient) $4\pi^2(q+1)e^2(q-1)^{-3\frac{q-1}{q+1}}$, using the fact that $\gamma_4 \ge 1$. The result of the Lemma follows by applying this inequality on inequation 14 with $t = nu$.

◊

A.3 Proof of Theorem 4

Notice that we have always $\bar{Z}_n' S_n^{-2}\bar{Z}_n \le q$. Indeed, there exists an orthogonal transformation O_n and a diagonal matrix $\Lambda_n^2 := diag[\hat{\mu}_j]_{1\le j\le q}$ with $\hat{\mu}_j > 0$ being the eigenvalues of S_n^2, such that $S_n^2 = O_n'\Lambda_n^2 O_n$. Now put $Y_{i,n} := [Y_{i,j,n}]_{1\le j\le q} = O_n Z_i$. It is easy to see that by construction the empirical variance of the $Y_{i,n}$ is

$$\frac{1}{n}\sum_{i=1}^n Y_{i,n}Y_{i,n}' = \frac{1}{n}\sum_{i=1}^n O_n Z_i Z_i' O_n' = O_n S_n^2 O_n' = \Lambda_n^2.$$

It also follows from this equality that, for all $j = 1,\cdots,q$, $\frac{1}{n}\sum_{i=1}^n Y_{i,j,n}^2 = \hat{\mu}_j$, and

$$\bar{Z}_n' S_n^{-2}\bar{Z}_n = \bar{Y}_n'\Lambda_n^{-2}\bar{Y}_n = \sum_{j=1}^q\left(\frac{1}{n}\sum_{i=1}^n Y_{i,j,n}\right)^2/\hat{\mu}_j \le q.$$

by Cauchy-Schwartz. So, for all $u > qn$

$$P(\bar{Z}_n' S_n^{-2} \bar{Z}_n \geq \frac{u}{n}) = 0.$$

a) In the symmetric and unidimensional framework ($q = 1$), this bound easily follows from Hoeffding inequality (see Efron, 1969). For completeness and to fix the notations, we recall the following simple proof. First of all

$$\Pr(n \frac{\overline{Z}_n^2}{S_n^2} \geq u) = 2 \Pr(\frac{\sum_{i=1}^n Z_i}{\sum_{i=1}^n Z_i^2} \geq \sqrt{u}).$$

With $q = 1$ inequality $Pr(\sqrt{n}\bar{Z}_n \geq \sqrt{u} S_n)$ becomes

$$\Pr\left(\frac{\sum_{i=1}^n Z_i}{(\sum_{i=1}^n Z_i^2)^{\frac{1}{2}}} > \sqrt{u} \right) \leq e^{-\frac{u}{2}}. \tag{17}$$

Let $\sigma_i, 1 \leq i \leq n$ be Rademacher random variables, independent from $(Z_i)_{1 \leq i \leq n}$, $\mathbb{P}(\sigma_i = -1) = \mathbb{P}(\sigma_i = 1) = 1/2$. We denote $\sigma_n(Z) = \left(\frac{1}{\sqrt{n}} \sum_{i=1}^n \sigma_i Z_i \right)$ and remark that $S_n^2 = \frac{1}{n} \sum_{i=1}^n \sigma_i Z_i Z_i' \sigma_i$.
Then we have by independence and symmetry of the Z_i's

$$\Pr\left(\frac{\bar{Z}_n}{S_n} \geq \sqrt{u} \right) = \int \Pr\left(\frac{\sigma_n(Z)}{S_n} \geq \sqrt{u} \middle| \bigcap_{i=1}^n Z_i = z_i \right) \Pi_{i=1}^n \mathbb{P}(dz_i).$$

But by Hoeffding inequality, we have

$$\Pr\left(\frac{\sigma_n(Z)}{S_n} \geq \sqrt{u} \middle| \bigcap_{i=1}^n Z_i = z_i \right) \leq e^{-u/2} \tag{18}$$

and the result follows by integration.

In the symmetric multidimensional framework ($q > 1$), the result is based on the inequality (18). Since the Z_i's have a symmetric distribution meaning, $-Z_i$ has the same distribution as Z_i. Then using a first symmetrization step we have,

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

Now,

$$\begin{aligned}
\sigma_n(Z)' S_n^{-2} \sigma_n(Z) &= \sigma_n(Y)' \Lambda_n^{-2} \sigma_n(Y) \\
&= \sum_{j=1}^q \left(\frac{1}{\sqrt{n}} \sum_{i=1}^n \sigma_i Y_{i,j,n} \right)^2 / \hat{\mu}_j \\
&= \sum_{j=1}^q \left(\sum_{i=1}^n \sigma_i Y_{i,j,n} \right)^2 / \sum_{i=1}^n Y_{i,j,n}^2.
\end{aligned}$$

It follows that

$$\Pr(\sigma_n(Z)' S_n^{-2} \sigma_n(Z) \geq u) \leq \sum_{j=1}^q \Pr\left(\frac{|\sum_{i=1}^n \sigma_i Y_{i,j,n}|}{\sqrt{\sum_{i=1}^n Y_{i,j,n}^2}} \geq \sqrt{u/q} \right)$$
$$\leq 2 \sum_{j=1}^q \mathbb{E} \Pr\left(\frac{\sum_{i=1}^n \sigma_i Y_{i,j,n}}{\sqrt{\sum_{i=1}^n Y_{i,j,n}^2}} \geq \sqrt{u/q} \,\middle|\, (Z_i)_{1\leq i \leq n} \right).$$

Apply now (18) to each self-normalized term in this sum to conclude.

b) The non-symmetric framework requires further investigations.

Our goal is to control $\Pr(n\bar{Z}_n' S_n^{-2} \bar{Z}_n \geq t)$. Define

$$B_n = \sup_{\substack{\|\lambda\|_{2,q}=1 \\ \lambda'\overline{Z}_n \geq 0}} \left\{ \frac{\lambda' \overline{Z}_n}{\sqrt{\lambda' S_n^2 \lambda}} \right\} \text{ and } D_n = \sup_{\substack{\|\lambda\|_{2,q}=1 \\ \lambda'\overline{Z}_n \geq 0}} \left\{ \sqrt{1 + \frac{\lambda' S^2 \lambda}{\lambda' S_n^2 \lambda}} \right\}.$$

First of all, remark that the following events are equivalent

$$\left\{ n\overline{Z}_n' S_n^{-2} \overline{Z}_n \geq t \right\} = \left\{ B_n \geq \sqrt{\frac{t}{n}} \right\}.$$

The final control is obtain by the control of two terms since

$$\Pr\left(B_n \geq \sqrt{\frac{t}{n}} \right) \leq \inf_{a > -1} \left\{ \Pr\left(B_n D_n^{-1} \geq \sqrt{\frac{t}{n(1+a)}} \right) + \Pr(D_n \geq \sqrt{1+a}) \right\}.$$

The control of the first term on the right side is obtained by applying part a) of Theorem 2 to $n^{1/2} \sup_{\substack{\|\lambda\|_{2,q}=1 \\ \lambda' \in \Gamma}} \frac{\lambda' \overline{Z}_n - \overline{Y}_n}{\sqrt{\lambda' S_n^2 (Z-Y) \tilde{\lambda}}}$ to obtain the control $2qe^{-\frac{t}{2q}}$. Then, by application of the Lemma 1 and the previous remark, we get $\sqrt{n} B_n D_n^{-1} \leq n^{1/2} \sup_{\substack{\|\lambda\|_{2,q}=1 \\ \lambda'\overline{Z}_n \geq 0}} \frac{\lambda' \overline{Z}_n}{\sqrt{\lambda' S_n^2 \tilde{\lambda} + \lambda' S^2 \tilde{\lambda}}}$, we have for all $t > 0$,

$$\Pr\left(B_n D_n^{-1} \geq \sqrt{\frac{t}{n}} \right) \leq 2qe^{1-\frac{t}{2q}}.$$

The control of the second term is trivial and useless for $a \leq 0$. Whereas, for all $a > 0$, and all $t > 0$ we have

$$\left\{ D_n \geq \sqrt{a+1} \right\} = \left\{ \sup_{\substack{\|\lambda\|_{2,q}=1 \\ \lambda'\overline{Z}_n \geq 0}} \left(1 + \frac{\lambda' S^2 \lambda}{\lambda' S_n^2 \lambda} \right) \geq 1 + a \right\} = \left\{ \inf_{\substack{\|\lambda\|_{2,q}=1 \\ \lambda'\overline{Z}_n \geq 0}} \left(\lambda' S^{-1} S_n^2 S^{-1} \lambda \right) \leq \frac{1}{a} \right\}$$
$$= \left\{ \mu_1(S^{-1} S_n^2 S^{-1}) \leq \frac{1}{a} \right\}.$$

We now use Lemma 2 applied to the r.v.'s $(S^{-1}Z_i)_{i=1,\cdots,n}$. Note that here we have $\tilde{\gamma}_4 = \mathbb{E}\|S^{-1}Z_1\|_2^4$, $\tilde{S}^2 = Id_q$, $\mu_1(\tilde{S}^2) = 1$ and $u = \frac{1}{a}$. For all $1 < a$, we have,

$$\Pr(D_n > \sqrt{1+a}) \le C(q)\left(\frac{n^3}{\tilde{\gamma}_4}\right)^{\tilde{q}} e^{-\frac{n}{(q+1)\gamma_4}(1-\frac{1}{a})^2}.$$

Since $\inf_{a>-1} \le \inf_{a>1}$, we conclude that, for any $t > n$,

$$\Pr\left(B_n > \sqrt{\frac{t}{n}}\right) \le \inf_{a>1}\left\{2qe\, e^{-\frac{t}{2q(1+a)}} + C(q)\left(\frac{n^3}{\tilde{\gamma}_4}\right)^{\tilde{q}} e^{-\frac{n}{(q+1)\tilde{\gamma}_4}(1-\frac{1}{a})^2}\right\}.$$

References

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

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

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

[4] Bertail, P. *Empirical likelihood in non and semi-parametric models.* M. Nikulin, 2003. to appear in *Semi-parametric models and applications.*

[5] Bertail, P., and Barbe, P. Testing the global stability of a linear model. Working Paper at CREST, 2004.

[6] Bertail, P., Harari-Kermadec, H., and Ravaille, D. $\gamma-$Divergence empirique et vraisemblance empirique généralisée. Submitted, 2005.

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

[8] Broniatowski, M., and Kéziou, A. *Parametric estimation and tests through divergences.* PhD thesis, L.S.T.A., 2003.

[9] Broniatowski, M., and Kéziou, A. Optimization of phi-divergences on sets of signed measures. to appear in Studia Mathematicarum Huncaricarum, 2005.

[10] Chen, S., and Cui, H. On the second order properties of empirical likelihood with moment restrictions. preprint at Iowa State University, 2005.

[11] Chistyakov, G. P., and Gtze, F. Moderate deviations for Student's statistic. *Theory of Probability & Its Applications 47*, 3 (2003), 415–428.

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

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

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

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

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

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

[18] Efron, B. Student's t-test under symmetry conditions. *Journal of american statistical society 64* (1969), 1278–1302.

[19] Embrechts, P., Lindskog, F., and McNeil, A. J. *Handbook of heavy tailed distributions in finance.* Elsevier, 2003, ch. Modelling dependence with copulas and applications to risk management. edited by Rachev ST.

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

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

[22] Jing, B.-Y., and Wang, Q. An exponential nonuniform Berry-Esseen bound for self-normalized sums. *Annals of Probability 27*, 4 (1999), 2068–2088.

[23] Jing, B. Y., and Wood, A. T. A. Exponential empirical likelihood is not bartlett correctable. *Annals of Statistics 24* (1996), 365–369.

[24] Liese, F., and Vajda, I. *Convex Statistical distance.* Teubner, Leipzig, 1987.

[25] Lonard, C. Minimization of energy functionals applied to some inverse problems. *Applied mathematics and optimization 44*, 3 (2001), 273–297.

[26] Mykland, P. A. Bartlett type of identities. *Annals of Statistics 22* (1994), 21–38.

[27] Mykland, P. A. Dual likelihood. *Annals of Statistics 23* (1995), 396–421.

[28] Newey, W. K., and Smith, R. J. Higher order properties of GMM and generalized empirical likelihood estimators. *Econometrica 72*, 1 (2004), 219–255.

[29] Owen, A. B. Empirical likelihood ratio confidence intervals for a single functional. *Biometrika 75*, 2 (1988), 237–249.

[30] Owen, A. B. Empirical likelihood ratio confidence regions. *Annals of Statistics 18* (1990), 90–120.

[31] Owen, A. B. *Empirical Likelihood.* Chapman and Hall/CRC, Boca Raton, 2001.

[32] Panchenko, D. Symmetrization approach to concentration inequalities for empirical processes. *Annals of Probability 31*, 4 (2003), 2068–2081.

[33] Qin, J., and Lawless, J. Empirical likelihood and general estimating equations. *Annals of Statistics 22*, 1 (1994), 300–325.

[34] Rao, M. M., and Ren, Z. D. *Theory of Orlicz Spaces.* Marcel Dekker, New York, 1991.

[35] Rockafellar, R. T. Integrals which are convex functionals. *Pacific Journal of Mathematics 24* (1968), 525–539.

[36] Rockafellar, R. T. *Convex Analysis.* Princeton University Press, Princeton, NJ, 1970.

[37] Rockafellar, R. T. Integrals which are convex functionals (II). *Pacific Journal of Mathematics 39* (1971), 439–469.