Generalized empirical likelihood for large parameters

Patrice BERTAIL

LS, CREST-INSEE
and MODAL'X Université Paris-Ouest-Nanterre-La Défense
FRANCE

Joint works with *Emmanuelle GAUTHERAT*, CREST and Université Reims

Outline

We observe $X_1, \cdots, X_n$ random variable of a separable Banach space $\mathbb{B}$ i.i.d. P in $\mathbb{P}$.

- Goal : Construct confidence regions for some multi-dimensional parameter $\theta = T(P)$ in $\mathbb{R}^q$, with a finite number of observations n (with n/q small but only the case $q << n$.
- Particular case : θ satisfies some moment constraints (including margin constraints). There exists $f \in \mathbb{R}^q$ with

$$\mathbb{E}_P f(X, \theta) = 0.$$

.

- General case of interest $\theta = T(P)$ is a functional parameter defined on a **space of signed measure**, Hadamard differentiable (tangentially to well-chosen sets of function satisfying some uniform entropy conditions), with first order gradient (influence function) $T^{(1)}(X, P)$ such that

- Main idea of empirical likelihood (Owen,1988, 1990, 2001, Chapman & Hall) and its generalization (see Judge, Golan, Miller, 1996, Bertail, 2003, Newey and Smith, 2004)= project the empirical measure

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

where the δ_{X_i}'s are Dirac measures, on the space of **signed measure**satisfying the constraints $\{Q, \mathbb{E}_Q T^{(1)}(., P) = 0\}$.

- with respect to a convex pseudo-distance I (typically $-$divergence) defined on a space of **signed measures (FUNDAMENTAL)**,

$$\begin{array}{cc} \inf & I(Q_n, P_n) \\ Q_n << P_n & \\ \mathbb{E}_{Q_n} T^{(1)}(., P) = 0 & \end{array}$$

- Confidence region for $\theta = T(P)$:

$$\mathcal{C}_n(\eta) = \{T(Q_n/Q_n(\mathbb{B})), n \inf_{Q_n << P_n} I(Q_n, P_n) < \eta)$$

Goal : asymptotic confidence interval and non-asymptotic control of this kind of region.

$$Pr\left(\theta \in \mathcal{C}_n(\eta)\right) \approx Pr\left(n \inf_{\substack{Q_n << P_n \\ \mathbb{E}_{Q_n} T^{(1)}(., P) = 0}} I(Q_n, P_n) < \eta \right)$$

Because of the constraint $Q_n << P_n$, Q_n belong to the set

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

In case $I = K$, the Kullback distance yields empirical likelihood and the solution of the minimization problem, $p_{i,n}$ are strictly positive and $\sum_{i=1}^{n} p_{i,n} = 1$.

For other choice of I, one can not impose these constraints, else there might be no solution to the minimization problem, because constraint qualifications may not be satisfied. Ex : χ^2 type divergence. Even for the Kullback defined for signed measure, this constraint may be dropped (the weight automatically sum to 1).

Empirical φ^*-Discrepancies (Bertail, 2003, Harari, 2005, B., Harari, Ravaillé, 2007, Kéziou, 2003, Broniatowski, Kéziou, 2006, Kitamura, 2006, Peletier, 2010, Rochet, 2011,...)
A family of convex pseudo-distance I is call φ^*-divergence (or φ^*-discrepancies).

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

where φ^* is the convex conjugate (Fenchel transform) of a function φ

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

Hypotheses

* φ function satisfying assumptions A_1:
It is convex, twice differentiable on its (non-void) domain containing 0, non negative,$\varphi(0) = 0, \varphi^{(1)}(0) = 0, \varphi^{(2)}(0) = 1...$

* The second order derivative $\varphi^{(2)}$ is lower bounded par $m > 0$ on $d(\varphi) \cap R^+(\neq 0)$.

For details on φ^*-discrepancies or divergences see Csiszar, 1967, Rockafellar, 1970, Princeton U. P., 1971, Pacific M. J.

It is easy to check that Cressie-Read discrepancies (Cressie-Read 1984) leading to the so called generalized empirical likelihood in the Econometric literature (see Newey and Smith, 2004) fulfill assumptions A_1

For $\kappa \in \mathbb{R}$,

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

then

$$\varphi_\kappa(x) = \frac{\left[(\kappa-1)x+1\right]^{\frac{\kappa}{\kappa-1}} - \kappa x - 1}{\kappa}$$

This family contains all the usual discrepancies, such as

- Relative Entropy $(\kappa \to 1)$, MEM (maximum entropy in Mean), Csiszar(1985), Gamboa, Gassiat (1997), entropy econometrics (Judge, Golan, Miller (1996), Exponential tilting, Schennach (2007) (robustness properties)
- Hellinger distance $(\kappa = 1/2)$,
- "Kullback" divergence $(\kappa \to 0)$ corresponding to the useful and classical empirical likelihood method (Owen 1990, 2001). Low coverage for small sample size n/q small, see Tsao, 2004, Ann. Stat., upper bounds for the coverage rate). easy to understand : for a real mean, the largest confidence interval is the convex envelopp $[min(Xi), max(X_i)]$ whose coverage accuracy is bounded (cannot attain any level).
  Computational problems when the number of constraints is very large (a well known fact in the convex programming literature: semi-infinite programming).

This family contains all the usual discrepancies, such as

- χ^2 ($\kappa = 2$), leading to exact calculation (square of self-nomalized sums), GMM, and large confidence regions (too conservative) (the solution of the primal problem is not a probability). Not Bartlett correctable. But enjoy the self normalized properties thus very robust. There exists exact exponential bound, Pinelis (1994), Bertail, Gautherat, Harari
- but a lot of other discrepancies also available (polylogarithm, some convex combinations of discrepancies) for which asymptotic results and exact exponential control also apply

Generalized empirical likelihood works because of duality theory for convex integral functionals (see Rockafellar, 1970, Borwein and Lewis, 1991, SIAM J. Comp. Opt., Leonard, 2003, Math. Hung., Bertail, 2003, B. H.R., 2007, Broniatowski, Keziou, 2009) Main hypotheses for the existence of the dual program,

constraints qualification :
There exists a measure R, dominated by P_n), satisfying the constraints such that

$$\inf d(\varphi^*) \ < \ \inf_{\Omega} \frac{dR}{d\mathbb{P}_n} \ \leq \ \sup_{\Omega} \frac{dR}{d\mathbb{P}_n} \ < \ \sup d(\varphi^*),$$

then we have a duality representation. Remark : also valid in

infinite dimension (Leonard, 2003) up to a duality gap, depending on the shape of the constraints, this duality gap may be controlled by adding penalities.

Duality on space of signed measure

$$\begin{array}{c} \inf \qquad\qquad nI_{\varphi^*}(Q_n, P_n) \\ Q_n << P_n \\ Q_n T^{(1)}(X,P) = 0 \end{array}$$

$$= n \sup_{\lambda \in \mathbb{R}^q} \left\{ -\lambda' P_n T^{(1)}(X,P) - P_n \varphi(\lambda' T^{(1)}(X,P)) \right\}$$

If $Var(T^{(1)}(X,P))$ is definite-positive then this quantity is asymptotically $\chi^2(q)/2$, (almost obvious since φ behaves like $x^2/2$ in the neighborhood of 0...) . The solution in λ (Kuhn-Tucker coefficient) is then asymptotically close to the square of a self-normalized sum. For the χ^2 divergence, exact computation.

Theorem

- For Hadamard differentiable functionals, tangentially to some Donsker classes of functions (satisfying some uniform entropy condition for the $L2$ norm) then generalized empirical likelihood is valid.
- The image by T of the ball centered at P_n for I_{φ^*} with radius $\chi^2_{1-\alpha}(q)/n$ is asymptotically a $1-\alpha$ confidence region for $T(P)$

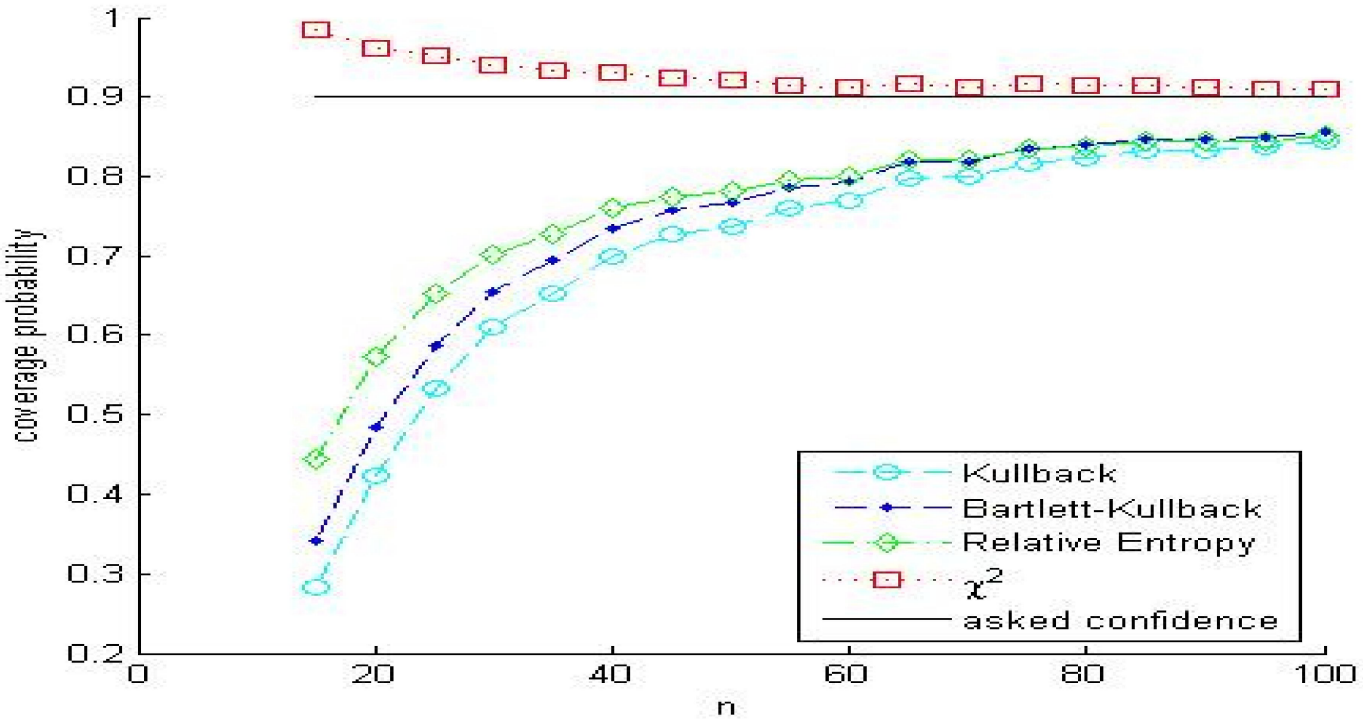
The proof relies on empirical process theory, by controlling that the weights belonging to the ball centered at P_n belong to some compact sets and establishing the uniform convergence of the corresponding weighted empirical process. Then applying Hadamard differentiability.

On the constraints qualification :

- Problems if the domain of φ^* is not $\mathbb{R}$, in particular, Kullback, Hellinger. CQ is not satisfied if 0 not is in the convex envelop of the $T(X_i, P)$.
For (moderate or very) large dimension q or small n, several solutions

- enlarge the constraint (add points or allow for a margin) AEL, BAEL
- penalize directly the dual form either with a L^2 norm (leading exactly to Tikhonov regularization of the original linearized inverse problem) or a L^1 norm (reduce the dimension of $T^{(1)}$).
- Mikland(1995) 's approach more interesting, because dual likelihood is a likelihood... All results on penalized likelihood apply directly to the dual program!
- Penalizing the dual with $pen^{-1}/4 * ||\lambda||_{\Sigma_n} = \lambda' \Sigma_n \lambda$ is strictly equivalent to the penalized likelihood of Bartolucci (2007) and Lahiri, Mukhopadhyay(2010) (according to the choice of Σ_n,
- For scad penalisation of the dual (but with righ definition of the Kullback divergence for **measure** not ensuring summation to 1), see related works in Tang and Lang(2010) (unfortunately not the correct likelihood for measure).

Problem : How to choose the dicrepancies for finite n?
Asymptotically all equivalents... Even if Kullback preferable from a large deviation point of view + Bartlett correctability, maybe not a good choice for small n (small n/q, see Tsao, 2004).



Simulation: Gaussian scale mixture with q=6

Quasi empirical-likelihood. Notion introduced in B., Harari, Ravaille(2007) but already used in convex optimisation, Auslender, Teboulle, Ben-Tiba (1999) log-proximal methods
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.
Efficient optimization algorithm in the optimization literature even with a large number of constraints (see interior point methods, log-proximal methods, semi-infinite programming), Teboulle(1997), Auslender, Teboulle, Ben-Tiba (1999) .

How to choose ε ?

K_ε^* has an explicit expression

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

and satisfy nice properties
(i) the domain $d(K_\varepsilon^*) = \mathbb{R}$
(ii) the second order derivative of k_ε is bounded from below:
$K_\varepsilon^{(2)}(x) \geq \varepsilon$.
(iii) $0 \leq K_\varepsilon^{*(2)}(x) \leq 1/\varepsilon$.

Theorem

- If $\varepsilon = O(n^{-3/2})$ then the quasi-empirical likelihood is Bartlett correctable. Even if q is large, exact computation (measure not probability) of the dual problem can be obtained. Efficient algorithm from the convex analysis litterature (log-proximal methods).
- Under the hypotheses A1, for all $n > q$, for any $\alpha > 0$, for any $n \geq \frac{2\varepsilon\alpha}{q}$, then

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

where $\overline{T_n}$ is the mean of the IF's and S_n the (uncentered) empirical variance.
Remark : useless for $\epsilon = 0$ (empirical likelihood).

Theorem

The following inequalities hold, **for finite** $n > q$ *and for* $t < nq$*:*

a) *(Pinelis 1994) if* $T^{(1)}(X, P)$ *has a symmetric distribution, without any moment assumption, denote then we have*

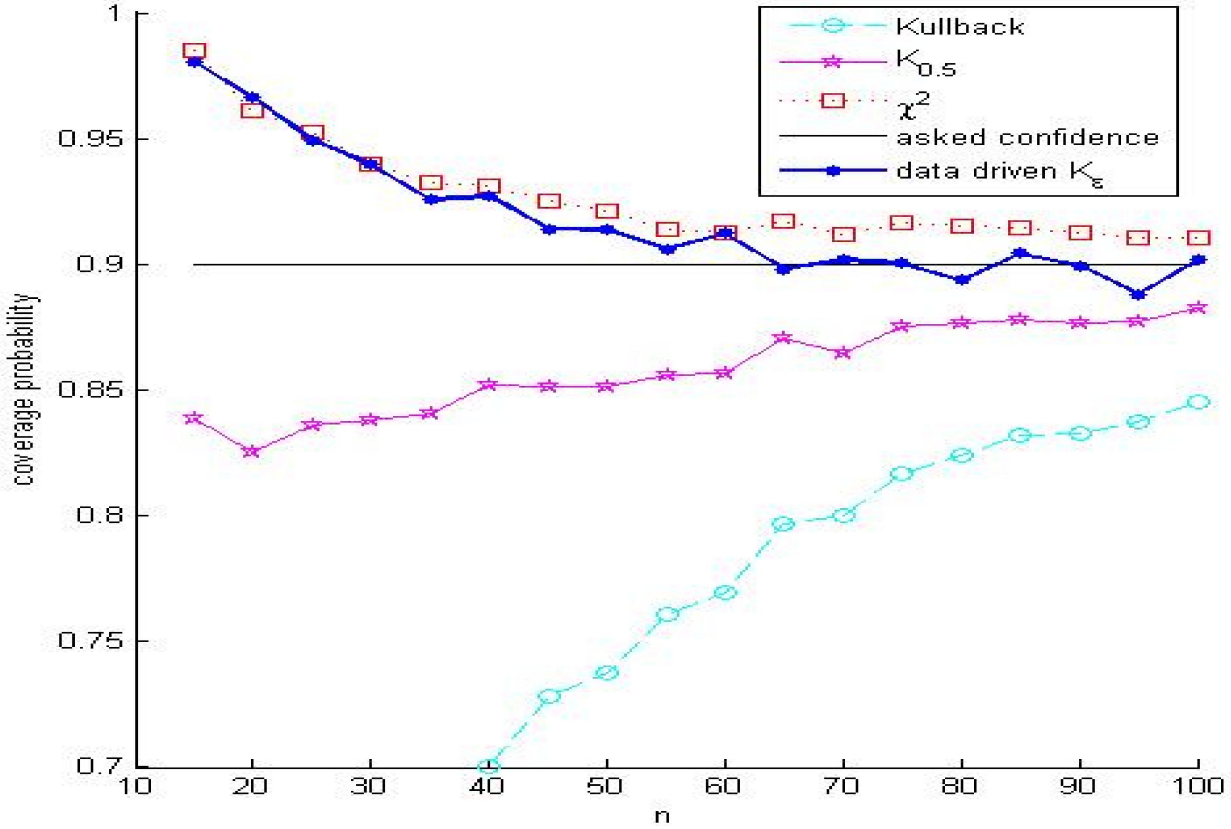
$$\Pr\left(n\overline{T}'_n S_n^{-2}\overline{T}_n \geq t\right) \leq \frac{2e^3}{9}\overline{F}_q(t), \tag{1}$$

b) *for general distribution of* $T^{(1)}(X, P)$ *with kurtosis* $\gamma_4 < \infty$*, for any* $a > 1$ *and for* $t \geq 2q(1+a)$ *and* $\tilde{q} = \frac{q-1}{q+1}$ *we have*

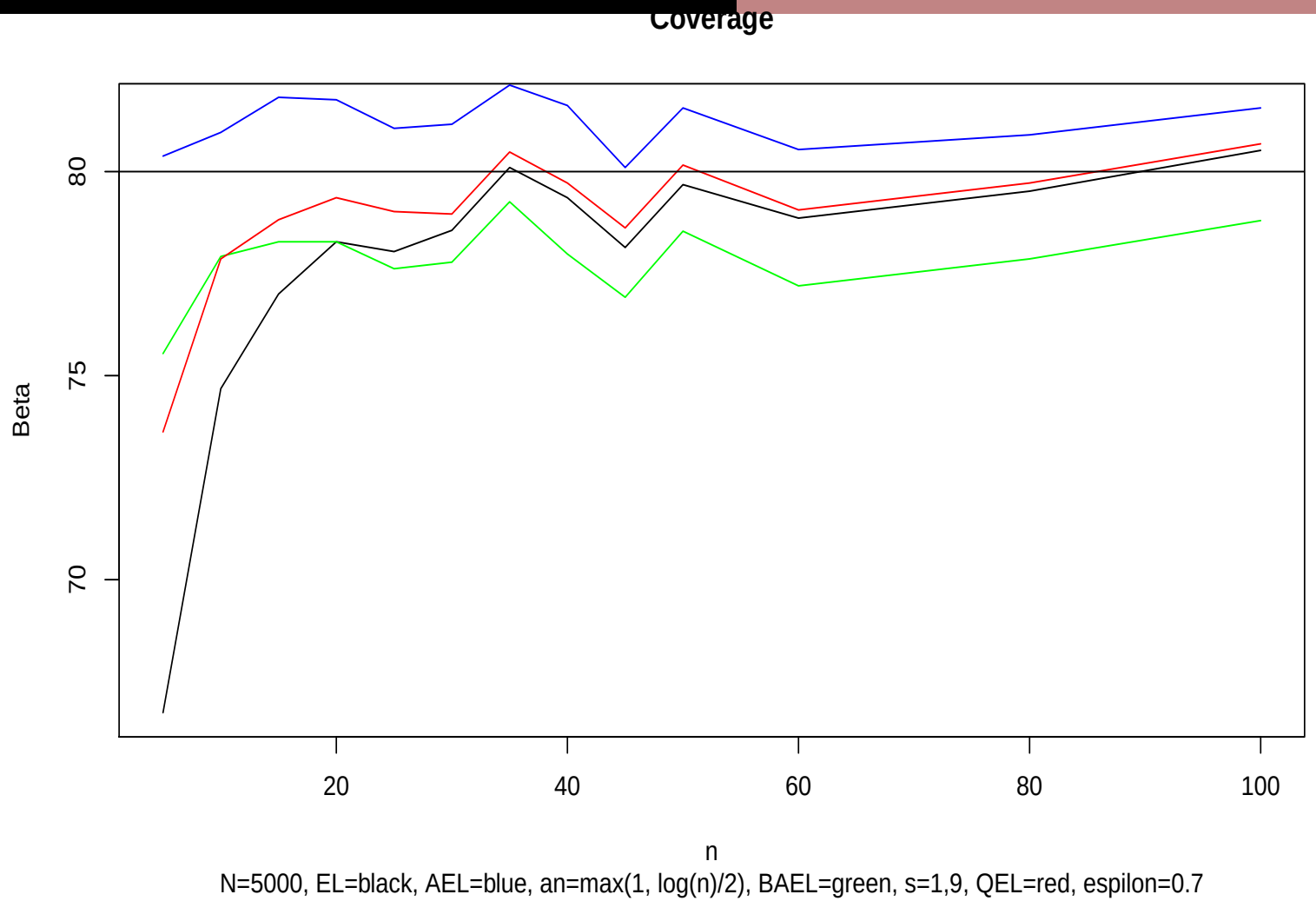
$$\Pr\left(n\overline{T}'_n S_n^{-2}\overline{T}_n \geq t\right)$$

$$\leq \frac{2e^3}{9\Gamma(\frac{q}{2}+1)}\left(\frac{t-q(1+a)}{2(1+a)}\right)^{\frac{q}{2}} e^{-\frac{t-q(1+a)}{2(1+a)}} + C(q)\left(\frac{n^3}{\gamma_4}\right)^{\tilde{q}} e^{-\frac{n\left(1-\frac{1}{a}\right)^2}{\gamma_4(q+1)}} \tag{2}$$

- If $\theta = T(P)$ unique solution of an M-equation $f(X, \theta) = 0$ then quasi-empirical likelihood may be controlled, for any n and $q = O(n/log(n)^2)$, by essentially the tail of a $\chi^2(q)$. Notice the recentering by q in the bounds.
- If T is robust and the influence function bounded by some known value, the bound is distribution free.
- Parcular case : quantile satisfying $E_P(I_{X \leq \theta} - \alpha) = 0$



Coverage probabilities: q=6, scale mixture, for $\alpha = 10$.
Choice of ε: calibrate to minimize the estimated coverage probability error (by using bootstrap for instance).



Comparison between Ajusted likelihood and quasilikelihood: q=2, normal, for $\alpha = 10$.

- Bertail "Empirical Likelihood in Some Non-parametric and Semi-Parametric Models". in Semiparametric estimation, ed Nikulin, 2003
- Bertail "Empirical Likelihood in Some Semi-Parametric Models". *Bernoulli*, 12(2) : 2999-331, 2006.
- Bertail, Gautherat, Harari-Kermadec " Exponential Bounds for multivariate self-normalized sums". Electronic Communication in Probability 13, 2009, pages 628-640.
- Hjort, N. L., McKeague, I., and Van Keilegom, I. (2009). Extending the scope of empirical likelihood. The Annals of Statistics 37, 1079-1111.
- Owen *"Empirical Likelihood"*, Chapman and Hall/CRC, 2001.
- Rockafellar "Integrals which are Convex Functionals", *Pacific Journal of Mathematics*, 24, 525-539, 1971.
- Rockafellar *"Convex Analysis"*, Princeton University Press", 1970.
- Tsao "Bounds on coverage probabilities of the empirical likelihood ratio confidence regions".*Annals of Statistics*. 32(3) : 1215-1221, 2004.