Generalized empirical likelihood in high dimension

Patrice BERTAIL

Laboratory of Statistics, CREST-INSEE
and MODAL'X University Paris X, Nanterre
FRANCE

Joint work with *Emmanuelle GAUTHERAT, Crest*
+some on going works with A. Chaterjee and S. Lahiri

Outline

1. Empirical φ^*-Discrepancies and Quasi Empirical Likelihood
   - The statistical model and Empirical φ^*-Discrepancies
   - The key property : Duality
2. $q << n/log(n)^2$: exponential bounds for quasi-empirical Likelihood
   - Properties of Quasi Empirical-Likelihood
   - Exponential bounds and conservative confidence regions
3. Large process valued parameter, without penalization
   - Uniform entropy conditions
   - link to self-normalized process
4. Large parameter : Penalizing the dual likelihood
   - Regularized generalized emp. divergence and Hotteling T^2
   - Dimension of the problem in the Hilbert space case

We observe $X_1, \cdots, X_n$ random vectors of $\mathbb{R}^r$ or a separable Banach space i.i.d. $\mathbb{P}$.

- Goal : Construct confidence regions for some multi-dimensional parameter θ of "large dimension" q, with a finite number of observations n (with n/q small...).

- Model :θ satisfies some moment constraints (including eventually some additional margin constraints).

$$\mathbb{E}_{\mathbb{P}}\xi(X, \theta) = 0.$$

where $\xi \in R^{\mathcal{F}}$, $\mathcal{F}$ is some finite set or some general class of real functions . For instance semiparametric model may be seen as infinite dimensional M-parameter (see van der Vaart (1995), Stat. Neerland.)

- In this talk
$$\theta = \{E_{\mathbb{P}} f,\ f \in \mathcal{F}\}$$
and
$$\xi(X, \theta) = \{f(X) - E_{\mathbb{P}} f\}_{f \in \mathcal{F}}.$$
Not : $E_P f = \mathbb{P} f$
- What really matters: the structure of the space of the constraints, its dimension in term of metric entropy and/or in term of the covariance operator of $\xi(X, \theta)$ (the eigenvalue).

- Main idea of empirical likelihood (Owen, 01, Chapman & Hall) and its generalization (see Newey and Smith, 03, Econometrica)= project the empirical measure

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

with δ_{X_i} is Dirac measure, on the space of "probability" or a signed measure satisfying the constraints $\{\mathbb{Q}, \mathbb{Q}(f - \mathbb{P}f) = 0\; f \in \mathcal{F}\}$.
- with respect to a convex pseudo-distance I (defined on a space of signed measures),

$$\inf_{\substack{Q_n << P_n \\ \mathbb{E}_{Q_n}(f(X) - \mathbb{P}f) = 0 \\ f \in \mathcal{F}}} I(Q_n, P_n)$$

- Confidence region :

$$Pr\left(\theta \in \mathcal{C}_n(\eta_n)\right) = Pr\left(\inf_{\substack{Q_n << P_n \\ \mathbb{E}_{Q_n}\xi(X,\theta)=0,\, f\in\mathcal{F}}} I(Q_n, P_n) < \eta_n \right)$$

Empirical likelihood (Kullback) Min(-

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 and imposing $\sum_{i=1}^{n} p_{i,n} = 1$ yields empirical likelihood. Largest region = CONVEX HULL of the $Z_i = \{f(X_i), f \in \mathcal{F}\} \in \mathbb{R}^{\mathcal{F}}$.

Very strong constraint (see Tsao, 2004, Ann. Stat.) : geometry of the region determined by the extreme points. Very few extremal points, when we have fat tail and even in the gaussian case : very bad finite sample properties in large dimension.

Several solutions have been proposed to solve this problem : Chen, Varyath, Abraham (2008), Emmerson and Owen (2009) or penalizing the original likelihood, Bartolucci (2007), Lahiri and Mukhopadhyay (2012) etc... : relaxing the constraints, either by adding a perturbation on the constraints.

For other choice of I , one can not impose these constraints else there might be no solution to the minimization problem. Ex : χ^2 type distance.

Work with general divergence but on space of signed measures rather than probability : escape the convex hull!

Empirical φ^*-Discrepancies on space of measures

Define the φ^*-divergence (or φ^*-discrepancies) between two (signed) measures.

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

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

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

φ function satisfying assumptions A_1: convex, twice differentiable on its (non-void) domain containing 0, non negative, $\varphi(0) = 0, \varphi^{(1)}(0) = 0, \varphi^{(2)}(0) = 1$, its second order derivative is lower bounded on R^+ (intersected with its domain) ...

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

Empirical φ^*-Discrepancies (Newey and Smith, 2003, Bertail, 2003, A. Kéziou, 2003, H. Harari, 2005, B., Harari, Ravaille, 2007, Broniatowski, Kéziou, 2006, Kitamura, 2006, Peletier, 2010, Rochet, 2011,...)

Key results : duality theory for convex integral on space of signed measures. In the finite dimensional case $card(\mathcal{F}) = q$, Borwein and Lewis, 1991, SIAM J. Comp. Opt., in the infinite dimensional case , Leonard, 2003, Math. Hung.

Duality on space of signed measure, q fixed

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

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

$$\inf_{\substack{Q_n << P_n \\ Q_n \xi(X,\theta) = 0}} nI_{\varphi^*}(Q_n, P_n)$$

$$= n \sup_{\lambda \in \mathbb{R}^q} \left\{ -\lambda' P_n \xi(X,\theta) - P_n \varphi(\lambda' \xi(X,\theta)) \right\}$$

Quasi empirical-likelihood : B., Harari, Ravaille (2007), Ann. Stat. Eco.

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

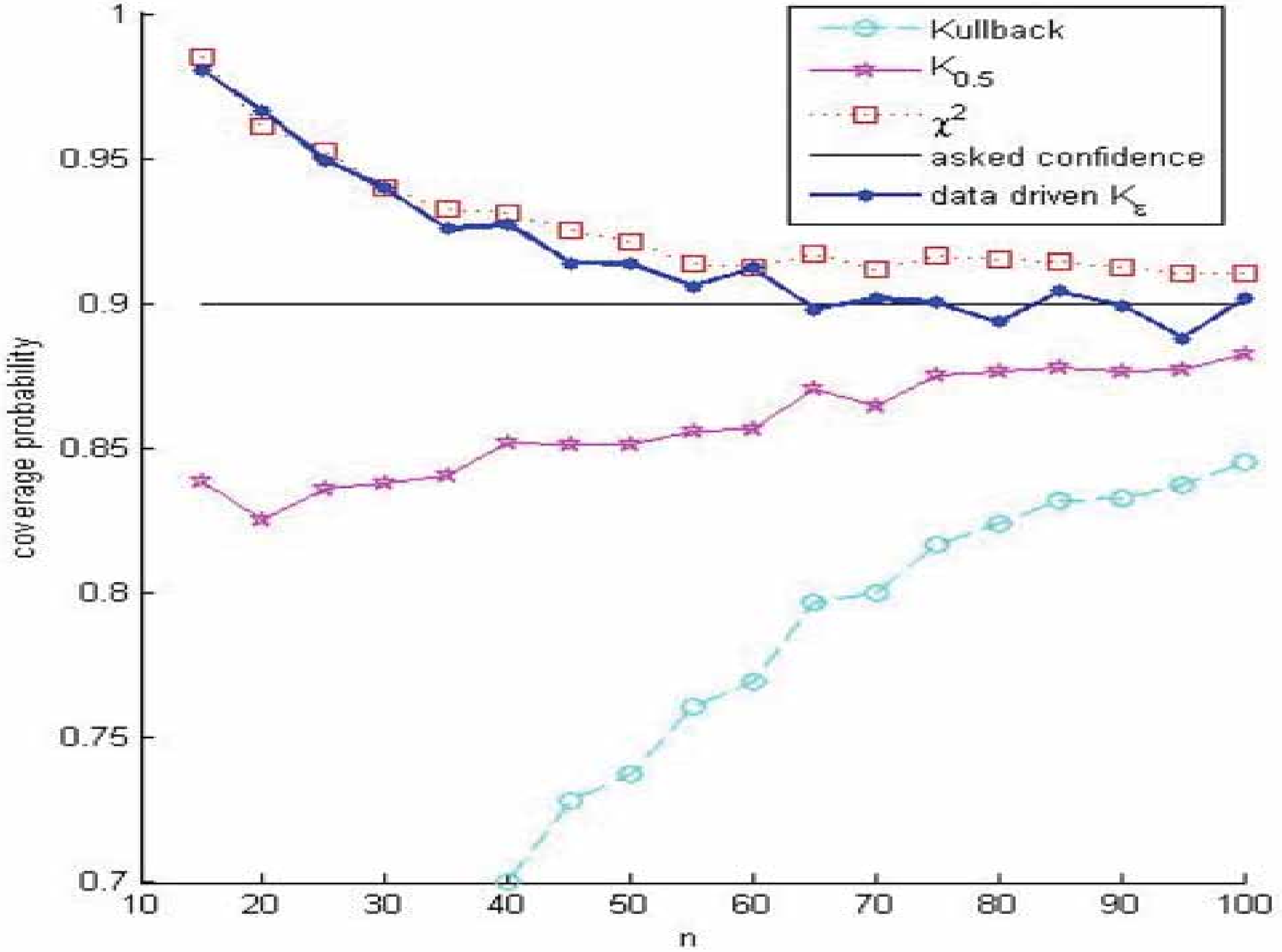
$$K_\varepsilon(x) = \varepsilon\, x^2/2 + (1-\varepsilon)(-x - \log(1-x)).$$

We call the corresponding K_ε^*-discrepancy, the quasi-Kullback discrepancy.
Efficient optimization algorithm are available in the optimization literature for this regularized discrepancy even with a large number of constraints (see log-proximal methods, Auslender, Teboulle, Ben-Tiba (1999) and the semi-infinite programming literature in convex analysis).

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) $0 \leq K_\varepsilon^{*(2)}(x) \leq 1/\varepsilon$.


Kullback
$K_{0.5}$
χ^2
asked confidence
data driven K_ε
coverage probability
n
1
0.95
0.9
0.85
0.8
0.75
0.7
10
20
30
40
50
60
70
80
90
100

Theorem

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

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

with $S_n^2 = n^{-1}P_n(\xi(X,\theta)\xi(X,\theta)^t)$
Remark : useless for $\epsilon = 0$ (empirical likelihood).

Exponential bounds, B., Gautherat, Harari (2008), E.C.P.

1. (Pinelis). Symmetric distribution
   $\Pr\left(n\overline{\xi}_n' S_n^{-2}\overline{\xi}_n \geq u\right) \leq \frac{2e^3}{9}\overline{F}_q(u).$
2. General distribution with kurtosis $\gamma_4 = E(||\xi||_2^4) < \infty$, for any $a > 1$ and for $2q(1+a) \leq u$,

$$\begin{aligned}\Pr\left(n\overline{\xi}_n' S_n^{-2}\overline{\xi}_n \geq u\right) &\leq K_1(q)P_{q,a}(u)e^{-\frac{u}{2(1+a)}} \\ &+K(q)\, n^{3\widetilde{q}}e^{-\frac{n}{\gamma_4(q+1)}\left(1-\frac{1}{a}\right)^2}\end{aligned}$$

with $K_1(q) = \frac{2e^{3+\frac{q}{2}}}{9\Gamma(\frac{q}{2}+1)}$ and $P_{q,a}(u) = \left(\frac{e^{2(1+a)}(u-q(1+a))}{2(1+a)}\right)^{\frac{q}{2}}$.

Proof : Pinelis(1994), Ann. Stat + generalization Panchenko(2003), Ann. Stat's symmetrization + Control smallest eigenvalue, Barbe and Bertail (2003). For all $n > q$ and for $u < nq$

Outlines

Large process valued parameter, without penalization

No need for relaxation or penalization if the parameter lies in a smooth structured space, see a recent work by Davit Varron (2014).

A1 : Envelop of the class. There exists a measurable function $H : \mathcal{X} \rightarrow \mathbb{R}$ such that there exists $\eta > 0$ such that $H(x) > \eta$ for every x and $\int_{x \in \mathcal{X}} H^{2+\eta}(x)P(dx) < \infty$ for some η and $|f(x)| \leq H(x)$ for all $x \in \mathcal{X}$ and any $f \in \mathcal{F}$.

A2 : Donsker classes. $\mathcal{F}$ is a Donsker class of function (Hoffmann-Jorgensen weak convergence). The set of probability measures P may be considered as a subset of $l_\infty(\mathcal{F})$, *i.e.* the space of all maps $\Phi : \mathcal{F} \rightarrow \mathbb{R}$ such that $||\Phi||_{\mathcal{F}} = \sup_{f \in \mathcal{F}} |\Phi(f)| < +\infty$) equipped with the uniform convergence norm $||P - Q||_{\mathcal{F}} = \sup_{h \in \mathcal{F}} |\int hdP - \int hQ|$.

Covering number at some probability Q

$N(\varepsilon, \mathcal{F}, ||.||_{2,Q})$ = number of balls of size ε needed to cover the class $\mathcal{F}$ with the $L^2(Q)$ norm

A3 : Uniform entropy number condition

$$\int_0^1 \sup_{Q \in \mathcal{D}} \sqrt{log(N(\varepsilon||H||_{2,Q}, \mathcal{F}, ||.||_{2,Q}))} d\varepsilon < \infty$$

,

where $\mathcal{D}$ is the set of all discrete probability measures Q such that $0 < \int H^2 dQ < \infty$

Then for each f it is possible to construct a quasi-empirical likelihood region. Then uniformly, over the class of function, we have the following result

Proposition

Assume that the divergence satisfies the constraints qualification, that the class of function satisfy A1-A3. If in in addition the smallest eigenvalue of the covariance operator $S^2 = (cov(f(X), g(X)))_{f\in\mathcal{F}, g\in\mathcal{F}}$ is strictly positive then

$$\sup_{f\in\mathcal{F}}\Big(\inf_{\substack{Q_n << P_n \\ Q_n f(X,\theta)=0}} nI_{\varphi^*}(Q_n, P_n)\Big)\Big)$$

$$= n \sup_{\lambda\in\mathbb{R}} \left\{-\lambda P_n f(X,\theta) - P_n \varphi(\lambda' f(X,\theta))\right\}$$

converges weakly to $\sup_{f\in\mathcal{F}}(G(f)' S^{-2}(f) G(f))$ where $G(f)$ is a gaussian process indexed by $\mathcal{F}$.

- Holds also true for empirical likelihood under additional assumptions ensuring the existence of a solution for every f (too strong) (see Davit Varron(2014))
- The limiting distribution is not distribution free... BUT we also can get some finite sample exponential bounds, using bounds for self-normalized process, under some sign coherency assumptions, using for instance Bercu, Gassiat, Rio (2002), Ann. Probab. Need for improvements (constants)...

- Unfortunately the eigenvalue assumption can not be removed. Consider for instance the case of $f(X) = 1_{X<x}, x \in \mathbb{R}$ then it is known (see Jaeschke, 1979, Ann. Stat.) that $sup_{x\in(R)}(\sqrt{n}(F_n - F(x))/(F_n(x)(1 - F_n(x)))^{1/2}) = \infty$ as well as $sup_{x\in(R)}(\sqrt{n}(F_n - F(x))/(F(x)(1 - F(x)))^{1/2}) = \infty$

The internal self-normalization of general empirical likelihood may be bad in some cases.

Outlines

Large parameter : Penalizing the dual likelihood

Solving the problem simultaneously for all f = semi-infinite optimisation problem (a lot of literature in convex optimization). Important to choose a divergence different from the Kullback.

Rather than relaxing the original likelihood, penalize the dual program (which is a likelihood for empirical likelihood).

Let λ be in the dual space of the constraints . To simplify the exposition we assume that the constraints lie in an Hilbert space. We introduce the norm

$$||\lambda||^2_{R,2} = \frac{1}{2}\lambda^* R_n \lambda$$

where R_n is some given self adjoint bounded (invertible) positive operator (eventually depending on n but it may be the identity). Define the penalized dual generalized empirical minimizer as

$$L_n(\lambda) = -\lambda^* P_n \xi(X,\theta) - P_n \varphi(\lambda^* \xi(X,\theta)) + \delta_n ||\lambda||^2_{R_n,\ 2}$$

By standard convex duality theory (see Boyd and Vanderburghe, 2009, E. 3.26), this is the dual problem of a relaxed primal problem

$$L_n = \inf_{\substack{Q_n << P_n \\ ||R_n^{-1/2}\mathbb{E}_{Q_n}\xi(X,\theta)||_2^2 \leq 2/\delta_n}} I_{\varphi^*}(Q_n, P_n)$$

This is very close to the proposal of Bartolucci (2007) (with $R = S_n^2$) and Lahiri and Mukhopadhyay (2012) (with $R = diag(S_n^2)$) except that the relaxation is directly integrated to the likelihood in their papers.
Leads to the same solution for an adequate choice of the penalisation, but not exactly equal to the value of the dual version.

Main Idea : under the preceeding conditions on φ, $L_n(\lambda)$ behaves like

$$\widetilde{L}_n(\lambda) = -\lambda^* P_n \xi(X,\theta) - \lambda^* (P_n \xi(X,\theta)\xi(X,\theta)^* + \delta_n R)\lambda/2$$

Denote

$$S_n^2 = P_n \xi(X,\theta)\xi(X,\theta)' = (\frac{1}{n}\sum_{i=1}^{n}(f(X_i) - Pf)(g(X_i) - Pf))$$

the empirical covariance operator (recentered at the true expectation). Assuming that the operator $S_n^2 + \delta_n R_n$ is invertible then, the solution in λ is given by

$$\lambda = (S_n^2 + \delta_n R_n)^{-1} P_n \xi(X,\theta)$$

The penalized generalized empirical likelihood should behave like

$$\widetilde{L}_n(\lambda) = \frac{1}{2} P_n \xi(X,\theta)^* ((S_n^2 + \delta_n R_n)^{-1}) P_n \xi(X,\theta)$$

that is, penalizing the dual leads to a Tikhonov or Ridge like regularization of the empirical covariance metric (which is calculated internally by the optimisation procedure).

Rq : a pb with Bartolucci(2007), it does not really regularize (it enlarges the margin and thus escape the convex hull of $\xi(X_i,\theta)'s$ but, for large dimension does not solve the problem).

The problem reduces to studying the behaviour of regularized quadratic forms or Hotteling T^2.

See also Peng and Schick(2013, 2014), for empirical likelihood with a growing number of constraints (extending Hjort, McKeague and Keilegom(2009) and Cheng, Peng and Qing(2009) : need to ensure the existence of a solution (control of the smallest eigenvalue). No need for quasi-empirical likelihood.

Other interesting references in the normal case : Srivastava Fujikoshi (2006), Srivastava(2007), J. Japan Stat. Soc, Chen Debashis, Ross, Prentice and Wang (2011) (with R=Id under normality assumptions and $q/n \to \gamma > 0$).

More precise results in a particular case.
Consider that ξ take its value in an Hilbert space H_2 with scalar product $< .,. >$, with countable basis $f_j, .j = 1, ...\infty$.

Define $\xi = \sum_{i=1}^{\infty} \xi_j f_j \ \xi_j =< \xi, f_j >$. Assume that $E||\xi||^4 < \infty$.

It is known that $\frac{1}{\sqrt{n}} \sum \xi(X_i, \theta) \underset{n\to\infty}{\overset{w}{\to}} G$, Gaussian r.v. in H_2 with covariance operator $S^2 = E(\xi(X_i, \theta) \odot \xi(X_i, \theta))$

Assume that S^2 is bounded and positive.
Let $\lambda_1 \geq \lambda_2 \geq ... \geq ...$ be eigenvalues of this operator (they are such that $\sum_{i=1}^{\infty} |\lambda_i| \leq \infty$) . Denote $\widehat{\lambda}_1 \geq \widehat{\lambda}_2 \geq ... \geq ...$ the estimated eigenvalues of S_n^2 and $\widehat{\mu}_1 \geq \widehat{\mu}_2 \geq ... \geq ...$ the eigenvalues of the regularizing operator R_n . R_n is assumed to commute with S_n^2 and to have bounded eigenvalues.
Define, for $l \in \mathbb{N}$, the dimensions

$$c(S^2, R, l, \delta) = \sum_{i=1}^{l} \frac{\lambda_i}{\lambda_i + \delta\mu_i}$$

$$d(S^2, R, l, \delta) = \sum_{i=1}^{l} \left(\frac{\lambda_i}{\lambda_i + \delta\mu_i}\right)^2$$

Theorem

Then, their always exists some sequence $\delta_n \to 0$ and $l_n \to \infty$ such that $l_n = \gamma\sqrt{n}\delta_n$ for some $\gamma > 0$, we have

$$\frac{\sup_{\lambda\in H_2}(L_n(\lambda)) - c(S_n^2, R_n, l_n, \delta_n)}{\sqrt{2d(S_n^2, R_n, l_n, \delta_n)}} \xrightarrow[n\to\infty]{d} N(0,1)$$

Proof : control the truncation l_n ensuring that $\sum_{i=l_n}^{\infty} \frac{\lambda_i}{\lambda_i+\delta_n\mu_i} \to 0$ and show that $\sup_{\lambda\in H_2}(L_n(\lambda))$ behaves like the weighted sum of i.i.d $\chi^2(1)$ r.v. $\sum_{i=1}^{l} \frac{\lambda_i}{\lambda_i+\delta_n\mu_i}(< G, f_j >^2 -1)$. Then apply Hajek(1961)'s, Ann. Math. stat. Theorem.

Moral : convergence of penalized generalized empirical likelihood (or of quadratic forms) to a distribution essentially depends on the "true" dimension of the problem measured by c and d or more precidely on $c/\sqrt{(2d)}$, not really on q.

- 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, 2008, pages 628-640.
- Panchenko "Symmetrization approach to concentration inequalities for empirical processes". *Annals of Statistics*. 31(4) : 2068-2081, 2003.
- Owen *"Empirical Likelihood"*, Chapman and Hall/CRC, 2001.
- Pinelis "Probabilitic Problem and Hotelling's t^2 test under a symmetry condition".*Annals of Statistics*. 22(1) : 357-368, 1994.
- 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.