A weighted bootstrap procedure for divergence minimization problems

Michel Broniatowski

A weighted bootstrap procedure for divergence minimization problems

Michel Broniatowski

Abstract Sanov type results hold for some weighted versions of empirical measures, and the rates for those Large Deviation principles can be identified as divergences between measures, which in turn characterize the form of the weights. This correspondence is considered within the range of the Cressie-Read family of statistical divergences, which covers most of the usual statistical criterions. We propose a weighted bootstrap procedure in order to estimate these rates. To any such rate we produce an explicit procedure which defines the weights, therefore replacing a variational problem in the space of measures by a simple Monte Carlo procedure.

1 The scope of this paper

Recall that a sequence of random elements X_n with values in a measurable space $(T, \mathscr{T})$ satisfies a Large Deviation Principle with rate Φ whenever, for all measurable set $\Omega \subset T$ it holds

$$\begin{aligned}\Phi\left(int\left(\Omega\right)\right) &\leq -\liminf_{n\to\infty}\frac{1}{n}\log P\left(X_n\in\Omega\right)\\ &\leq -\limsup_{n\to\infty}\frac{1}{n}\log P\left(X_n\in\Omega\right)\leq\Phi\left(cl\left(\Omega\right)\right)\end{aligned}$$

where $int\left(\Omega\right)$ (resp. $cl\left(\Omega\right)$) denotes the interior (resp. the closure) of Ω in T and $\Phi(\Omega) := \inf\left\{\Phi(t); t\in\Omega\right\}$. The σ-field $\mathscr{T}$ is the Borel one defined by a given basis on T. For subsets Ω in T such that

$$\Phi\left(int\left(\Omega\right)\right) = \Phi\left(cl\left(\Omega\right)\right) \tag{1}$$

Michel Broniatowski
Laboratoire de Statistique Théorique et Appliquée, Université Pierre et Marie Curie, Boîte Courrier 158, 4 place Jussieu, 75252 Paris Cedex 05, e-mail: michel.broniatowski@courriel.upmc.fr

it follows by inclusion that

$$-\lim_{n\to\infty}\frac{1}{n}\log P\left(X_n\in\Omega\right)=\Phi\left(int\left(\Omega\right)\right) \tag{2}$$
$$=\Phi\left(cl\left(\Omega\right)\right)=\inf_{t\in\Omega}\Phi(t)=\Phi(\Omega).$$

Assume that we are given such a family of random elements $X_1,X_2,\ldots$ together with a set $\Omega\subset T$ which satisfies (1). Suppose that we are interested in estimating $\Phi\left(\Omega\right)$. Then, whenever we are able to simulate a family of replicates $X_{n,1},\ldots,X_{n,K}$ such that $P\left(X_n\in\Omega\right)$ can be approximated by the frequency of those $X_{n,i}$'s in Ω, say

$$f_{n,K}\left(\Omega\right):=\frac{1}{K}card\left(i:X_{n,i}\in\Omega\right) \tag{3}$$

a natural estimator of $\Phi\left(\Omega\right)$ writes

$$\Phi_{n,K}\left(\Omega\right):=-\frac{1}{n}\log f_{n,K}\left(\Omega\right). \tag{4}$$

The rationale for this proposal is that visits of Ω by the random elements $X_{n,j}$'s tend to concentrate on the most favorable domain in Ω, namely where Φ assumes its minimal value in Ω, since $\left(\exp-n\Phi(x)\right)dx$ is a good first order approximation for the probability that X_n belongs to a neighborhood of x with volume dx. We have substituted the approximation of the variational problem $\Phi\left(\Omega\right):=\inf\left(\Phi\left(\omega\right),\omega\in\Omega\right)$ by a much simpler one, namely a Monte Carlo one, defined by (3). Notice further that we do not need to identify the set of points ω in Ω which minimize Φ; indeed there may be no such points even. Condition (1) provides an easy way to get statement (2), which yields to our estimates (4). Sometimes we may obtain (2) bypassing (1).

This program can be realized whenever we can identify the sequence of random elements X_i's for which, given the criterion Φ and the set Ω, the limit statement (2) holds. The present paper explores this approach in the case when the X_i's are empirical measures of some kind, and $\Phi(\Omega)$ writes $\phi\left(\Omega,P\right)$ which is the infimum of a divergence between some reference probability measure P and a class of probability measures Ω. This technique may lead to inferential procedures: for example assuming that $\Omega=\left\{Q_\theta\in\mathscr{M}_1,\theta\in\Theta\right\}$ is a statistical model such that $d(Q_\theta,P)\geq\varepsilon$ for some given distance d and some $\varepsilon>0$ and all θ in Θ, then minimizing a proxy of $\phi\left(\Omega_\theta,P\right)$ as obtained in this paper over θ provides minimum distance estimators of P within Ω.

The present paper presents estimators of $\phi\left(\Omega,P\right)$, focusing on their construction. We denote (P) the problem of finding an estimator for

$$\phi\left(\Omega,P\right) \tag{5}$$

where Ω is defined according to the context. But for simple convergence result of the proposed estimators, we do not provide finite sample or asymptotic properties of the estimators, which is postponed to future work; as seen later the method which we propose holds for rather general sets Ω; henceforth specific limit results of the estimator depend on the peculiar nature of the problem. Also the definition of the estimator through (4) may be changed using a better estimator of $P(X_n \in \Omega)$ than $f_n(\Omega)$,the naive one, which may have poor statistical performances and which may require a long runtime for calculation, since $(X_n \in \Omega)$ is a rare event; Importance Sampling procedures should be used. This is also out of the scope of this paper.

1.1 Existing solutions for similar problems

Minimizing a divergence between an empirical measure pertaining to a data set and a class of distributions is somehow synonymous as estimating the parent distribution of the data (although other methods exist); for example the maximum likelihood method amounts to minimize the likelihood (or modified Kullback-Leibler) divergence between P_n and a parametrized model. Inspired by the celebrated Empirical Likelihood approach, empirical divergence methods aim at finding solutions of the minimization of the divergence between P_n and all distributions in Ω which are supported by the data points; see [4]. Those may exist or not, yielding (or not yielding) to the estimation of the minimum value of the divergence. Besides the fact that Ω may consists in distributions which cannot have the data as supporting points, the resulting equations for the solution of the problem may be intractable. Also there may be an infinity of solutions for this problem. The case when Ω is defined by conditions on moments of some $L-$statistics is illuminating in this respect; indeed the direct approach fails, and leads to a new problem, defining divergences between quantile measures (see [6]). Instead, looking first for some estimator of the infimum value of the divergence leads to a well posed problem of finding the set of minimizers, an algorithmic problem for which a solution can be obtained along the lines of the present paper. Once obtained the minimal value of the divergence, minimizers may sometimes be obtained by dichotomous search; this depends on the context.

2 Divergences

Let $(\mathscr{X},\mathscr{B})$ be a measurable space and P be a given reference probability measure (p.m.) on $(\mathscr{X},\mathscr{B})$. The set $\mathscr{X}$ is assumed to be a Polish space. Denote $\mathscr{M}$ the real vector space of all signed finite measures on $(\mathscr{X},\mathscr{B})$ and $\mathscr{M}(P)$ the vector subspace of all signed finite measures absolutely continuous (a.c) with respect to (w.r.t.) P. Denote also $\mathscr{M}_1$ the set of all p.m.'s on $(\mathscr{X},\mathscr{B})$ and $\mathscr{M}_1(P)$ the subset of all p.m.'s a.c w.r.t. P. Let φ be a proper[1] closed[2] convex function from $]-\infty,+\infty[$ to $[0,+\infty]$ with $\varphi(1)=0$ and such that its domain $\mathrm{dom}\varphi := \{x \in \mathbb{R} \text{ such that } \varphi(x) < \infty\}$ is an interval with endpoints $a_\varphi < 1 < b_\varphi$ (which may be finite or infinite).

[1] We say a function is proper if its domain is non void.

[2] The closedness of φ means that if a_φ or b_φ are finite numbers then $\varphi(x)$ tends to $\varphi(a_\varphi)$ or $\varphi(b_\varphi)$ when $x \downarrow a_\varphi$ or $x \uparrow b_\varphi$, respectively.

For any signed finite measure Q in $\mathscr{M}(P)$, a classical definition for the ϕ-divergence between Q and P is defined by

$$\phi(Q,P) := \int_{\mathscr{X}} \varphi\left(\frac{dQ}{dP}(x)\right) dP(x). \tag{6}$$

When Q is not a.c. w.r.t. P, we set $\phi(Q,P) = +\infty$; see [34]. The first definition of ϕ-divergences between p.m.'s were introduced by I.Csiszar in [7] as "f-divergences". Csiszar's definition of ϕ-divergences between p.m.'s requires a common dominating σ-finite measure λ for Q and P. Note that the two definitions of $\phi-$divergences coincide on the set of all p.m.'s a.c w.r.t. P and dominated by λ. The ϕ-divergences between any signed finite measure Q and a p.m. P were introduced by [20] which proposes the following definition

$$\phi(Q,P) := \int \varphi(q)\, dP + b_{\varphi^*}\sigma_Q^+(\mathscr{X}) - a_{\varphi^*}\sigma_Q^-(\mathscr{X}), \tag{7}$$

where

$$a_{\varphi^*} = \lim_{y\to-\infty} \frac{\varphi(y)}{y}, \quad b_{\varphi^*} = \lim_{y\to+\infty} \frac{\varphi(y)}{y}. \tag{8}$$

and

$$Q = qP + \sigma_Q, \quad \sigma_Q = \sigma_Q^+ - \sigma_Q^-$$

is the Lebesgue decomposition of Q, and the Jordan decomposition of the singular part σ_Q, respectively. Definitions (6) and (7) coincide when Q is a.c. w.r.t. P or when $a_\varphi = -\infty$ or $b_\varphi = +\infty$. Since we will consider optimization of $Q \mapsto \phi(Q,P)$ on sets of signed finite measures a.c. w.r.t. P, it is more adequate for our sake to use the definition (7).

For all p.m. P, the mappings $Q \in \mathscr{M} \mapsto \phi(Q,P)$ are convex and take nonnegative values. When $Q = P$ then $\phi(Q,P) = 0$. Furthermore, if the function $x \mapsto \varphi(x)$ is strictly convex on a neighborhood of $x = 1$, then the following basic property holds

$$\phi(Q,P) = 0 \text{ if and only if } Q = P. \tag{9}$$

All these properties are presented in [7], [21], [22] and [24] Chapter 1, for ϕ-divergences defined on the set of all p.m.'s $\mathscr{M}_1$. When the ϕ-divergences are defined on $\mathscr{M}$, then the same properties hold making use of definition (7); see also [2].

When defined on $\mathscr{M}_1$, the Kullback-Leibler (KL), modified Kullback-Leibler (KL_m), χ^2, modified χ^2 (χ_m^2), Hellinger (H), and L_1 divergences are respectively associated to the convex functions $\varphi(x) = x\log x - x + 1$, $\varphi(x) = -\log x + x - 1$, $\varphi(x) = \frac{1}{2}(x-1)^2$, $\varphi(x) = \frac{1}{2}(x-1)^2/x$, $\varphi(x) = 2(\sqrt{x}-1)^2$ and $\varphi(x) = |x-1|$. All those divergences except the L_1 one, belong to the class of power divergences introduced in [19] (see also [24] chapter 2). They are defined through the class of convex functions

$$x \in]0,+\infty[\mapsto \varphi_\gamma(x) := \frac{x^\gamma - \gamma x + \gamma - 1}{\gamma(\gamma-1)} \tag{10}$$

if $\gamma \in \mathbb{R}\setminus\{0,1\}$, $\varphi_0(x) := -\log x + x - 1$ and $\varphi_1(x) := x\log x - x + 1$. (For all $\gamma \in \mathbb{R}$, we define $\varphi_\gamma(0) := \lim_{x\downarrow 0}\varphi_\gamma(x)$). So, the $KL-$divergence is associated to φ_1, the KL_m to φ_0, the χ^2 to φ_2, the χ^2_m to φ_{-1} and the Hellinger distance to $\varphi_{1/2}$.
Those divergence functions defined in (10) are the Cressie-Read divergence functions; see [19].
The Kullback-Leibler divergence (KL-divergence) is sometimes called Boltzmann Shannon relative entropy. It appears in the domain of large deviations and it is frequently used for reconstruction of laws, and in particular in the classical moment problem (see e.g. [20] and the references therein). The modified Kullback-Leibler divergence (KL_m-divergence) is sometimes called Burg relative entropy. It is frequently used in Statistics and it leads to efficient methods in statistical estimation and tests problems; in fact, the celebrate "maximum likelihood" method can be seen as an optimization problem of the KL_m-divergence between the discrete or continuous parametric model and the empirical measure associated to the data; see [26] and [3]. On the other hand, the recent "empirical likelihood" method can also be seen as an optimization problem of the KL_m-divergence between some set of measures satisfying some linear constraints and the empirical measure associated to the data; see [30] and the references therein, [18] and [4]. The Hellinger divergence is also used in Statistics, it leads to robust statistical methods in parametric and semi-parametric models; see [17], [29], [9] and [4].

The power divergences functions $Q \in \mathscr{M}_1 \mapsto \phi_\gamma(Q,P)$ can be defined on the whole vector space of signed finite measures $\mathscr{M}$ via the extension of the definition of the convex functions φ_γ : For all $\gamma \in \mathbb{R}$ such that the function $x \mapsto \varphi_\gamma(x)$ is not defined on $]-\infty,0[$ or defined but not convex on whole $\mathbb{R}$, we extend its definition as follows

$$x \in]-\infty,+\infty[\mapsto \begin{cases} \varphi_\gamma(x) & \text{if } x \in [0,+\infty[, \\ +\infty & \text{if } x \in]-\infty,0[. \end{cases} \tag{11}$$

Note that for the χ^2-divergence for instance, $\varphi_2(x) := \frac{1}{2}(x-1)^2$ is defined and convex on whole $\mathbb{R}$. This extension of the domain of the divergence functions φ_γ to $]-\infty,+\infty[$ implies that (8) is well defined, with $a_{\varphi^*} = +\infty$.

The conjugate (or Fenchel-Legendre transform) of φ will be denoted φ^*,

$$t \in \mathbb{R} \mapsto \varphi^*(t) := \sup_{x\in\mathbb{R}} \{tx - \varphi(x)\},$$

and the endpoints of domφ^* (the domain of φ^*) are a_{φ^*} and b_{φ^*} with $a_{\varphi^*} \leq b_{\varphi^*}$. Note that φ^* is a proper closed convex function. In particular, $a_{\varphi^*} < 0 < b_{\varphi^*}$, $\varphi^*(0) = 0$. By the closedness of φ, the conjugate φ^{**} of φ^* coincides with φ, i.e.,

$$\varphi^{**}(t) := \sup_{x\in\mathbb{R}} \{tx - \varphi^*(x)\} = \varphi(t), \text{ for all } t \in \mathbb{R}.$$

For proper convex functions defined on $\mathbb{R}$ (endowed with the usual topology), the lower semi-continuity[3] and the closedness properties are equivalent.

We say that φ (resp. φ^*) is differentiable if it is differentiable on $]a_\varphi, b_\varphi[$ (resp. $]a_{\varphi^*}, b_{\varphi^*}[$), the interior of its domain. We say also that φ (resp. φ^*) is strictly convex if it is strictly convex on $]a_\varphi, b_\varphi[$ (resp. $]a_{\varphi^*}, b_{\varphi^*}[$).

The strict convexity of φ is equivalent to the condition that its conjugate φ^* is essentially smooth, i.e., differentiable with

$$\begin{array}{l} \lim_{t\downarrow a_{\varphi^*}} \varphi^{*\prime}(t) = -\infty \text{ if } a_{\varphi^*} > -\infty, \\ \lim_{t\uparrow b_{\varphi^*}} \varphi^{*\prime}(t) = +\infty \text{ if } b_{\varphi^*} < +\infty. \end{array}$$

Conversely, φ is essentially smooth if and only if φ^* is strictly convex; see e.g. [33] section 26 for the proofs of these properties.

If φ is differentiable, we denote φ' the derivative function of φ, and we define $\varphi'(a_\varphi)$ and $\varphi'(b_\varphi)$ to be the limits (which may be finite or infinite) $\lim_{x\downarrow a_\varphi} \varphi'(x)$ and $\lim_{x\uparrow b_\varphi} \varphi'(x)$, respectively. We denote $\mathrm{Im}\varphi'$ the set of all values of the function φ', i.e., $\mathrm{Im}\varphi' := \left\{\varphi'(x) \text{ such that } x \in [a_\varphi, b_\varphi]\right\}$. If additionally the function φ is strictly convex, then φ' is increasing on $[a_\varphi, b_\varphi]$. Hence, it is a one-to-one function from $[a_\varphi, b_\varphi]$ onto $\mathrm{Im}\varphi'$; we denote in this case ${\varphi'}^{-1}$ the inverse function of φ' which is defined from $\mathrm{Im}\varphi'$ onto $[a_\varphi, b_\varphi]$.

Note that if φ is differentiable, then for all $x \in]a_\varphi, b_\varphi[$,

$$\varphi^*\left(\varphi'(x)\right) = x\varphi'(x) - \varphi(x). \tag{12}$$

If additionally φ is strictly convex, then for all $t \in \mathrm{Im}\varphi'$ we have

$$\varphi^*(t) = t{\varphi'}^{-1}(t) - \varphi\left({\varphi'}^{-1}(t)\right) \quad \text{and} \quad \varphi^{*\prime}(t) = {\varphi'}^{-1}(t).$$

On the other hand, if φ is essentially smooth, then the interior of the domain of φ^* coincides with that of $\mathrm{Im}\varphi'$, i.e., $\left(a_{\varphi^*}, b_{\varphi^*}\right) = \left(\varphi'(a_\varphi), \varphi'(b_\varphi)\right)$.

The domain of the ϕ-divergence will be denoted $\mathrm{dom}\phi$, i.e.,

$$\mathrm{dom}\phi := \{Q \in \mathscr{M} \text{ such that } \phi(Q,P) < \infty\}.$$

In the present paper we will deal with essentially smooth divergence functions, so that all the above properties are fulfilled.

[3] We say a function φ is lower semi-continuous if the level sets $\{x \text{ such that } \varphi(x) \leq \alpha\}$, $\alpha \in \mathbb{R}$ are closed.

3 Large deviations for the bootstrapped empirical measure

The present Section aims at providing a solution to Problem (P) when Ω is a subset of $\mathscr{M}_1$, the class of all probability measures on $(\mathscr{X},\mathscr{B})$. Such a goal will be achieved in two cases of interest, namely the Kullback-Leibler and the Likelihood divergence.

We first push forwards some definition.

Let $Y, Y_1, Y_2, \ldots$ denote a sequence of non negative independent real valued random variables with expectation 1. We assume that Y satisfies the so-called Cramer condition, namely that the set

$$\mathscr{N} := \left\{ t \in \mathbb{R} \text{ such that } \Lambda(t) := \log E e^{tY} < \infty \right\}$$

contains a neighborhood of 0 with non void interior. By its very definition, $\mathscr{N}$ is an interval, say $\mathscr{N} := (a,b)$ which we assume to be open. We also assume that the strictly convex function Λ is a steep function, namely that $\lim_{t\to a} \Lambda(t) = \lim_{t\to b} \Lambda(t) = +\infty$. It will also be assumed that $t \to \Lambda'(t)$ parametrizes the convex hull of the support of the distribution of Y. We refer to [16] for those notions and conditions.

Consider now the weights $W_i^n, 1 \leq i \leq n$ defined through

$$W_i^n := \frac{Y_i}{(1/n)\sum_{i=1}^n Y_i}$$

which define a vector of exchangeable variables $(W_1^n, \ldots, W_n^n)$ for all $n \geq 1$.

Define further the Legendre transform of Λ, say Λ^* which is a strictly convex function defined on $\operatorname{Im}\Lambda'$ by

$$\Lambda^*(x) := \sup_t tx - \Lambda(t).$$

We assume that we are given an array of observations $(x_i^n)_{i=1,\ldots,n,n\geq 1}$ in $\mathscr{X}$ which we assume to be "fair", meaning that there exists a probability measure P defined on $(\mathscr{X},\mathscr{B})$ such that

$$\lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^n \delta_{x_i^n} = P. \tag{13}$$

When the observations are sampled under P we assume that the above condition (13) holds almost surely. We define the bootstrapped empirical measure of $(x_1^n, \ldots, x_n^n)$ by

$$P_n^W := \frac{1}{n} \sum_{i=1}^n W_i^n \delta_{x_i^n}.$$

Note that P_n^W is random due to the weights $W_1^n, \ldots, W_n^n$ and that the data set $x_1^n, \ldots, x_n^n$ is considered as non random. The following result provides a Sanov type LDP statement conditionally upon the array (x_i^n) $1 \leq i \leq n, n \geq 1$. Assuming that Y has no atom at 0 and that $t \to \Lambda_Y(t)$ is steep at point

$$t^{+} := \sup\{t : \Lambda_Y(t) < +\infty\}$$

with $t^{+} > 0$, it holds

Theorem 1. *Under the above hypotheses and notation the sequence P_n^W obeys a LDP on the space of all probability measures on X equipped with the weak convergence topology with good rate function*

$$\phi\left(Q,P\right) := \begin{array}{ll} inf_{m>0} \int \Lambda^{*}\left(m\frac{dQ}{dP}(x)\right) dP(x) & if\, Q << P \\ +\infty & otherwise \end{array} \tag{14}$$

Remark 1. This Theorem is a variation on Corollary 3.3 in [35]. Indeed it holds

$$\lim_{x \to -\infty} \Lambda_Y'(t) = \lim_{x \to -\infty} \left((\Lambda^{*})'\right)^{-1}(x) = 0$$

and

$$\lim_{x \to +\infty} \Lambda_Y'(t) = \lim_{x \to +\infty} \left((\Lambda^{*})'\right)^{-1}(x) = +\infty$$

The above Theorem does not meet our requirement that the rate should be a divergence between probability measures. Two cases of upmost interest however fulfill our quest.

We make use of independent copies of P_n^W, obtained as follows: consider

$$\left(Y_{1,1}, \ldots, Y_{1,n}\right), \ldots, \left(Y_{,1}, \ldots, Y_{K,n}\right)$$

where all the $Y_{i,j}$ are i.i.d. copies of Y, and

$$W_i^k := \frac{Y_{k,i}}{\sum_{i=1}^{k} Y_{k,i}},$$

$$P_{k,n}^W := \sum_{i=1}^{n} W_i^k \delta_{x_i^n}.$$

and for any set Ω in $\mathscr{M}_1$ define

$$P_{n,K}(\Omega) := \frac{1}{K} card\left(k \in \{1, \ldots, K\} : P_{k,n}^W \in \Omega\right) \tag{15}$$

and denote

$$L_{n,K}\left(\Omega\right) := -\frac{1}{n} \log P_{n,K}\left(\Omega\right). \tag{16}$$

3.1 Minimizing the Kullback-Leibler divergence

Assume that the random variable Y is Poisson distributed with mean 1. Then

$$\Lambda^*(x) = x\log x - x + 1$$

which is the Kullback-Leibler divergence function. For any couple of probability measures (Q,P) it readily follows that the infimum upon m in (14) is reached at $m = \exp -KL(Q,P)$, which yields

$$inf_{m>0} \int \Lambda^* \left(m\frac{dQ}{dP}(x) \right) dP(x) = 1 - \exp -KL(Q,P). \tag{17}$$

It follows that the rate (14) takes the form

$$\phi\,(Q,P) = 1 - \exp -KL(Q,P)$$

and that

$$\phi\,(\Omega,P) = 1 - \exp -KL(\Omega,P)$$

Proposition 1. *Consider any set Ω of probability measures which satisfies*

$$KL\,(int\Omega,P) = KL\,(cl\Omega,P)\,,$$

where $\mathscr{M}_1$ is endowed with the weak topology. Consider Y a r.v. with Poisson distribution with mean 1. Then the following expression

$$\widehat{KL}\,(\Omega,P) := -\log\left[1 - L_{n,K}\,(\Omega)\right]$$

estimates $KL\,(\Omega,P)$.

3.2 Minimizing the Likelihood divergence

Let the r.v. Y have an exponential distribution with mean 1. Then

$$\Lambda^*(x) = -\log x + x - 1$$

which is the divergence function which defines the modified Kullback-Leibler divergence, also named as Likelihood divergence, since its minimization in statistically relevant contexts yields the celebrated maximum likelihood divergence estimators.

For all P and Q in $\mathscr{M}_1$ such that $KL_m(Q,P)$ is finite, the function $(0,1) \ni m \to \int \varphi \left(m\frac{dQ}{dP}(x) \right) dP(x)$ is decreasing. Therefore the (14) takes the form

$$\phi\,(Q,P) = KL_m(Q,P)$$

and

$$\phi\,(\Omega,P) = KL_m(\Omega,P)$$

This yields an analogue of Proposition 1, namely

Proposition 2. *With the same notation and hypotheses as in Proposition 1 , with Y a random variable with Exponential(*1*) distribution, the following expression*

$$\widehat{KL_m}\left(\Omega,P\right):=L_{n,K}\left(\Omega\right)$$

estimates $KL_m\left(\Omega,P\right)$.

Remark 2. When Y is exponentially distributed with expectation 1 then by Pyke's Theorem, the vector $\left(W_1^n,\ldots,W_n^n\right)$ coincides in distribution with

$$\left(U_{1,n},U_{2,n}-U_{1,n},\ldots,U_{n,n}-U_{n-1,n}\right),$$

the spacings of the ordered statistics $\left(U_{1,n},U_{2,n},\ldots,U_{n,n}\right)$ of n i.i.d. uniformly distributed r.v's on $(0,1)$, with uniform distribution. This is indeed the simplest weighted bootstrap variation of P_n based on exchangeable weights.

4 Wild bootstrap

We now consider other random elements whose visits in Ω will define estimators of minimum divergence between P and Ω for other useful divergence function, as the Chi-square, the Hellinger, etc.

We may consider some wild bootstrap versions, defining the wild empirical measure by

$$P_n^{Wild}:=\frac{1}{n}\sum_{i=1}^{n}Y_i\delta_{x_{i,n}}$$

where the r.v's $Y_1,Y_2,\ldots$ are i.i.d. with common expectation 1. The use of the word "wild" is relevant: P_n^{Wild} is not merely a probability measure; it can even put negative masses on some points of its support, since the r.'s Y_i may assume negative values. We will be able to solve Problem (P) when Ω is a subset of $\mathscr{M}$, the class of all signed finite measures on $\left(\mathscr{X},\mathscr{B}\right)$. Thus the estimator of $\phi\left(\Omega,P\right)$ is typically smaller than the estimator of $\phi\left(\Omega\cap\mathscr{M}_1,P\right)$, which cannot be estimated using the results of this Section, in contrast with just obtained in the previous Section. Also we will need $\mathscr{X}$ to be a compact set.

We assume that the Cramer condition holds for Y and define, as above,

$$\Lambda_Y(t):=\log E\exp tY.$$

4.1 A conditional LDP for the wild bootstrapped empirical measure

In this case we make use of the following result (see [23]) which holds when $\mathscr{X}$ is compact.

Theorem 2. *The wild empirical measure P_n^{Wild} obeys a LDP in the class of all signed finite measures endowed by the weak topology with good rate function $\phi\left(Q,P\right)$ defined in (7), where the function φ is defined by*

$$\varphi(x) := \Lambda^*(x) = \sup_t tx - \Lambda_Y(t).$$

Remark 3. Making use of the results in [23],we may consider the constant a_{φ^*} and b_{φ^*} in (7); by convexity, $\varphi^*(x) := \Lambda_Y(x)$. The LDP rate (7) writes

$$\phi\left(Q,P\right) := \int_{\mathscr{X}} \Lambda^* \left(\frac{dQ_a}{dP}\right) dP + \int_{\mathscr{X}} \rho\left(\frac{dQ_s}{d\theta}\right) d\theta$$

where

$$\rho(z) := \sup\left\{\lambda z : \lambda \in Dom\Lambda_Y\right\}$$

and θ is any real valued non negative measure with respect to which Q_s is absolutely continuous. Choosing

$$\theta = \left|Q_s^+ - Q_s^-\right|$$

yields

$$\phi\left(Q,P\right) := \int_{\mathscr{X}} \Lambda^* \left(\frac{dQ_a}{dP}\right) dP + \rho(-1)Q_s^-\left(\mathscr{X}\right) + \rho(+1)Q_s^+\left(\mathscr{X}\right)$$

so that $a_{\varphi^*} = \inf\left\{t : \Lambda_Y(t) < \infty\right\}$ and $b_{\varphi^*} = \sup\left\{t : \Lambda_Y(t) < \infty\right\}$.

Remark 4. Theorem 2 has been proved by numerous authors, under various regularity conditions; see e.g. [15], [23], [5]. A strong result is as follows:

When Ω is a subset in $\mathscr{M}$ such that $\phi\left(cl\left(\Omega\right),P\right) = \phi\left(int\left(\Omega\right),P\right)$ holds in the $\tau-$topology, then

$$\lim_{n\to\infty} -\frac{1}{n}\log P\left(P_n^{Wild} \in \Omega\right) = \phi\left(\Omega,P\right). \tag{18}$$

However that $\tau-$open (resp. $\tau-$closed sets) are not necessarily weakly open (resp weakly closed); thus this latest result (18) is merely useful when Ω is defined as the pre-image of some open (closed) set by some $\tau-$continuous mapping from $\left(\mathscr{X},\mathscr{B}\right)$ onto some topological space; see Section 6.

4.2 Cressie-Read divergences and exponential families

In this Section we consider a reciprocal statement to Theorem 2. We first prove that any Cressie-Read divergence function as defined in (11) is the Fenchel-Legendre transform of some cumulant generating function Λ_Y for some r.v. Y. Henceforth we state a one to one correspondence between the class of Cressie-Read divergence functions and the distribution of some Y which can be used in order to build a bootstrap empirical measure of the form P_n^{Wild}.

4.3 Natural Exponential families and their variance functions

We turn to some results due to Letac and Mora; see [12].

For μ a positive $\sigma-$finite measure on $\mathbb{R}$ define $\phi_\mu(t) := \int e^{tx} d\mu(x)$ and its domain $\mathscr{D}_\mu$, the set of all values of t such that $\phi_\mu(t)$ is finite, which is a convex (possibly void) subset of $\mathbb{R}$. Denote $k_\mu(t) := \log \phi_\mu(t)$ and let $m_\mu(t) := (d/dt) k_\mu(t)$ and $s_\mu^2(t) := \left(d^2/dt^2\right) k_\mu(t)$. Associated with μ is the Natural Exponential Family NEF(μ) of distributions

$$dP_t^\mu(x) := \frac{e^{tx} d\mu(x)}{\phi_\mu(t)}$$

which is indexed by t. It is a known fact that, denoting X_t a r.v. with distribution P_t^μ it holds $EX_t = m_\mu(t)$ and $VarX_t = s_\mu^2(t)$.The mapping $t \to m_\mu(t) := EX_t$ is a strictly increasing homeomorphism from $\mathbb{R}^+$ onto $\mathbb{R}^+$, with inverse $m_\mu^{\leftarrow}$.

The NEF(μ) is said to be *generated* by μ. The NEF(ν) generated by ν defined through

$$d\nu(x) = \exp(ax+b) d\mu(x) \tag{19}$$

coincides with NEF(μ), which yields to the definition of the NEF generated by the class of positive measures ν satisfying (19) for some constants a and b.Following [12] for the notation and main results the class of such measures will be denoted $\mathscr{B}$ and be called a *base* for NEF(μ), hence denoted NEF($\mathscr{B}$). Also it can be checked that the range of m_ν does not depend on the very choice of ν in $\mathscr{B}$, although its domain depends on ν. The range $\operatorname{Im} m_{\mathscr{B}}$ of m_ν, which is the same for all ν in $\mathscr{B}$, is called the mean range of $\mathscr{B}$ since it depends only on the class of generating measures $\mathscr{B}$.

Defined on $\operatorname{Im} m_{\mathscr{B}}$, the function

$$x \to V(x) := s_\mu^2 o m_\mu^{\leftarrow}(x)$$

is independent of the peculiar choice of μ in $\mathscr{B}$ (see [12]) and is therefore called the variance function of the NEF($\mathscr{B}$). It can be proved that the variance function characterizes the NEF. From the statistician point of view the functional form of the function V is of relevant interest: it corresponds to models for which regression of the variance on the mean is considered, which is a common feature in heteroscedastic models; see the seminal paper [14] which is at the origin of models characterized by V, and [10].

Starting with [11], a wide effort has been developed in order to characterize the basis of a NEF with given variance function.

4.4 Power variance functions and the corresponding natural exponential families

Power variance functions have been explored by various authors; see e.g. [1], [12], etc. Summarizing it holds (see [1]) the NEF with variance function $V(x) = Cx^{\alpha}$; for sake of brevity with respect to the sequel we denote $\alpha = 2 - \gamma$. NEF with variance function V are obtained through integration and identification of the resulting moment generating function. They are generated as follows.

- For $\gamma < 0$ by stable distributions on $\mathbb{R}^+$ with characteristic exponent in $(0, 1)$. The resulting distributions define the Tweedie scale family which we briefly describe in the next paragraph.
- For $\gamma = 0$ by the exponential distribution
- For $0 < \gamma < 1$ by Compound Gamma-Poisson distributions
- For $\gamma = 1$ by the Poisson distribution
- For $\gamma = 2$ by the normal distribution

Other values of γ do not yield NEF's.

4.4.1 The Tweedie scale

Let Z be a r.v. with stable distribution on $\mathbb{R}^+$ with exponent τ, $0 < \tau < 1$. Denote p its density and $f(t) = E \exp itZ$ its characteristic function, which satisfies

$$f(t) = \exp\left\{iat - c\,|t|^{\tau}\,(1 + i\beta sign\,(t)\,\omega\,(t, \tau))\right\}$$

where $a \in \mathbb{R}$, $c > 0$ and $\omega\,(t, \tau) = \tan\left(\frac{\pi\tau}{2}\right)$.

We consider the case when $\beta = 1$. It then holds:

For $Z_1, \ldots, Z_n$ n i.i.d. copies of Z ,

$$\frac{Z_1 + \ldots + Z_n}{n^{1/\tau}} =_d Z$$

where the equality holds in distribution. The Laplace transform of p satisfies

$$\varphi(t) := \int_0^{\infty} e^{-tx} p(x)dx = e^{-t^{\tau}}$$

for all non negative value of t; see [27].

Associated with p is the Natural Exponential family (NEF) with basis p namely the densities defined for non negative t through

$$p_t(x) := e^{-tx} p(x)/e^{-t^{\tau}}$$

with support $\mathbb{R}^+$. For positive t , a r.v. X_t with density p_t has a moment generating function $E \exp \lambda X_t$ which is finite in a non void neighborhood of 0 and therefore has moments of any order.

Consider the density $p_1(x) = e^{-x+1}p(x)$ with finite m.g.f. in $(-\infty, 1)$, expectation $\mu = \tau$ and variance $\sigma^2 = \tau(1-\tau)$. Finally set for all non negative x

$$q(x) := \sqrt{\tau(1-\tau)}p_1\left(x\sqrt{\tau(1-\tau)} + \tau - 1\right)$$

which for all $0 < \tau < 1$ is the density of some r.v. Y with expectation 1 and variance 1. The m.g.f. of Y is

$$E\exp\lambda Y = e\exp\left[1 - \frac{\tau}{\sqrt{\tau(1-\tau)}}\right]\exp - \left[1 - \frac{\lambda}{\sqrt{\tau(1-\tau)}}\right]^{\tau}.$$

For $\tau = 1/2$,Y has the Inverse Gaussian distribution with parameters $(1,1)$ and m.g.f

$$E\exp\lambda Y = e\left(\exp - [1-2\lambda]^{1/2}\right).$$

The variance function of the NEF generated by a stable distribution with index τ in $(0,1)$ writes

$$V(x) = C_\tau x^{\frac{2-\tau}{1-\tau}}$$

with

$$C_\tau := \left(\frac{1-\tau}{\tau}\right)^{\frac{2-\tau}{2(1-\tau)}}.$$

4.4.2 Compound Gamma Poisson distributions

We briefly characterize this compound distribution and the resulting weight Y. Let μ denote the distribution of $S_N := \sum_{i=0}^{N} \Gamma_i$ where $S_0 := 0$, N is a Poisson (p) r.v. independent of the independent family $(\Gamma_i)_{i\geq 1}$ where the Γ_i's are distributed with Gamma distribution with scale parameter $1/\lambda$ and shape parameter $-\rho$. Here

$$\rho := \frac{\gamma - 1}{\gamma}$$

$$\lambda := \rho$$

$$p := (\gamma - 1)^{-1/\gamma}$$

where we used the results in [1] p1516. Consider the family of distributions NEF(μ) generated by μ, which has power variance function $V(x) = x^{\gamma+1}$ defined on $\mathbb{R}^+$. The r.v. Y has distribution in NEF(μ) with expectation and variance 1. Its density is of the form

$$f_W(x) := \exp(ax+b)\,f(x)$$

where $f(x) := (d\mu(x)/dx)$ is the density of S_N. The values of the parameters a and b are

$$a := -1$$

$$b := -(\gamma-1)^{-1/\gamma}\left[\left(1-\frac{\gamma}{\gamma-1}\right)^{\rho}-1\right].$$

4.5 Cressie Read divergences, weights and variance functions

For

$$\varphi_\gamma(x) := C\frac{x^\gamma-\gamma x+\gamma-1}{\gamma(\gamma-1)} \tag{20}$$

with $\gamma \neq 0,1$, the convex function φ_γ satisfies $\varphi_\gamma(1) = \varphi'_\gamma(1) = 0$ and $\varphi^{"}_\gamma(1) = C$, being therefore a divergence function; it is customary to assume that the positive constant C satisfies $C = 1$, a condition which we will not consider, still denoting this class of functions the Cressie-Read family of divergence functions. Set $\varphi_0(x) = -\log x + x - 1$ and $\varphi_1(x) = x\log x - x + 1$, the likelihood divergence function and the Kullback-Leibler one, noting that $\lim_{\gamma\to 0}\varphi_\alpha(x) = \varphi_0(x)$ and $\lim_{\gamma\to 1}\varphi_\gamma(x) = \varphi_1(x)$. The Cressie-Read family defined through (20) is the simplest system of non negative convex functions satisfying the requirements for a divergence function.

We prove that any Cressie-Read divergence function is the Fenchel Legendre transform of a moment generating function of a random variable with expectation 1 and variance $1/C$ in a specific NEF, depending upon the divergence. Indeed we identify such a r.v. Y as follows: let Y be a r.v. with a cumulant generating function. $\Lambda(t) := \log E\exp tY$ such that

$$\varphi_\gamma(x) = \Lambda^*(x) = \sup_t tx - \psi(t); \tag{21}$$

then

$$\frac{1}{\frac{d^2}{dx^2}\varphi_\gamma(x)} = \frac{1}{C}x^\alpha = V(x) \tag{22}$$

with $\alpha = 2-\gamma$ for $x \to V(x)$ the variance function of the NEF generated by the distribution of Y. Since the differential equation $\frac{d^2}{dx^2}\varphi_\gamma(x) = Cx^{-\alpha}$ defines $\varphi_\gamma(x)$ through (20) in a unique way we have proved the one to one correspondence between Cressie-Read divergences and NEF's with power Variance functions.

Remark 5. Reproductible NEF's with power variance functions and power normalizing factors are infinitely divisible (see [1]); reciprocally all reproductible NEF's with power normalizing factors are infinitely divisible. The Cressie Read family of divergences possesses therefore a quite peculiar property : they are the only ones which are the Legendre transform of cumulant generating functions of reproductible infinitely divisible distributions with power normalizing constants. Reciprocally any wild empirical measure with reproductible infinitely divisible weights with power normalizing factors and with expectation 1 has LDP rate in the Cressie Read family.

4.6 Examples

For example the Tweedie scale of distributions defines random variables Y with expectation 1 and variance C_τ corresponding to Cressie Read divergences with negative index $\gamma = -\tau/(1-\tau)$.

For $\gamma = -1$, the resulting divergence is

$$\varphi_{-1}(x) = \frac{1}{2}\frac{(x-1)^2}{x}$$

which is the modified χ^2 divergence (or Neyman χ^2). The associated r.v. Y has an Inverse Gaussian distribution with expectation 1 and variance 1.

For $\gamma = 2$ it holds

$$\varphi_2(x) = \frac{1}{2}(x-1)^2$$

which is the Spearman χ^2 divergence. The resulting r.v. Y has a Gaussian distribution with expectation 1 and variance 1. Note that in this case, Y is not a positive random variable.

For $\gamma = 1/2$ we get

$$\varphi_{1/2}(x) = 2\left(\sqrt{x}-1\right)^2$$

which is the Hellinger divergence. The associated random variable Y has a Compound Gamma-Poisson distribution with $\rho = -1, \lambda = -1, p = 4, a = -1$ and $b = 4$.

When $\gamma = 3/2$ the distribution of Y belongs to the NEF generated by the stable law μ on $\mathbb{R}^+$ with characteristic exponent $1/3$, hence with density the Modified Bessel type distribution

$$f(x) = (d\mu(x)/dx) = (2\pi)^{-1}\lambda K_{1/2}\left(\lambda x^{1/2}\right)\exp\left(-px + 3\left(\lambda^2 p/4\right)^{1/3}\right)$$

where λ and p are positive and $K_{1/2}(z)$ is the modified Bessel function of order $1/2$ with argument z.

When $\gamma = 1$ then

$$\varphi_0(x) = x\log x - x + 1,$$

the Kullback-Leibler divergence function, and Y has a Poisson distribution with parameter 1. Since the rate of the corresponding LDP coincides with the rate of the LDP for the empirical distribution of the data (unconditionally), and since the variance function characterizes the distribution of the weights, this is the only wild bootstrap which is LDP efficient.

When $\gamma = 0$ then

$$\varphi_0(x) = -\log x + x - 1,$$

the Likelihood divergence and Y has an exponential with parameter 1.

The L_1 divergence function $\varphi(x) = |x-1|$ does not yield to any weighted sampling; indeed $\varphi^*(t) = t1_{(-1,1)}(t) + \infty 1_{(-1,1)^c}(t)$ which is not a cumulant generating function.

5 Monte Carlo minimization of a Cressie read divergence through Wild bootstrap

Due to the preceding correspondence between the minimization problem (P) and Large Deviation rates, we propose the following procedures for the estimation of $\phi\left(\Omega,P\right)$.

Simulate nK i.i.d. random variables $Y,Y_{1,i},Y_{2,i},\ldots,Y_{K,i}$,$1\leq i\leq n$ with common distribution in correspondence with the divergence function φ, namely such that

$$\varphi(x)=\Lambda^{*}(x)$$

for $x\in Dom\varphi$ where $\Lambda^{*}(x):=\sup_{t}tx-\Lambda\left(t\right)$ and $\Lambda\left(t\right)=\log E\exp tY$. Define

$$P_{n,K}(\Omega):=\frac{1}{K}card\left(j\in\{1,\ldots,K\}:P_{n,j}^{Wild}\in\Omega\right)$$

where

$$P_{n,j}^{Wild}:=\frac{1}{n}\sum_{i=1}^{n}Y_{j,i}\delta_{x_{i}}$$

$1\leq j\leq K$.

Define

$$\phi_{n,K}^{Wild}\left(\Omega,P\right):=-\frac{1}{n}\log P_{n,K}\left(\Omega\right).$$

6 Sets of measures for which the Monte Carlo minimization technique applies

We explore cases when

$$\phi\left(int\left(\Omega\right),P\right)=\phi\left(cl\left(\Omega\right),P\right)\tag{23}$$

in the weak topology on $\mathscr{M}$. Two conditions are derived; in the first case we make use of convexity arguments; we make use of a similar argument as used in [28], Corollary 3.1. For Ω a subset of $\mathscr{M}$ denote $cl_{w}\left(\left(\Omega\right)\right)$, resp. $int_{w}\left(\Omega\right)$, the weak closure (resp.) the weak interior of Ω in $\mathscr{M}$.

A convex set Ω in $\mathscr{M}$ is strongly $w-$convex if for all Q in $cl_{w}\left(\left(\Omega\right)\right)$ and each R in $int_{w}\left(\Omega\right)$ it holds that

$$\left\{\alpha Q+(1-\alpha)R;0<\alpha<1\right\}\subset int_{w}\left(\Omega\right).$$

It holds

Proposition 3. *Let $P\in\mathscr{M}_{1}$ and let $\Omega_{1},\ldots,\Omega_{J}$ be subsets of $\mathscr{M}$. Set $\Omega:=\Omega_{1}\cup\ldots\cup\Omega_{J}$. Then when all Ω_{j} s are strongly $w-$convex and $\phi\left(int_{w}\left(\Omega_{j}\right),P\right)<\infty$ for all j, (23) holds.*

Proof. For any $j = 1, \ldots, J$, fix $\varepsilon > 0$. Let $Q \in cl_w((\Omega_j))$ be such that

$$\phi(Q,P) < \phi(cl_w(\Omega_j), P) + \varepsilon$$

and $R \in int_w(\Omega_j)$ be such that $\phi(R,P) < \infty$. Define $Q_\alpha := \alpha Q + (1-\alpha)R, 0 < \alpha < 1$. Then $Q_\alpha \in int_w(\Omega_j)$ and the convexity of $Q' \to \phi(Q',P)$ implies

$$\begin{aligned}\phi(int_w(\Omega j), P) &\leq \lim_{\alpha \uparrow 1} \{\alpha \phi(Q,P) + (1-\alpha)\phi(R,P)\} \\ &= \phi(Q,P) < \phi(cl_w((\Omega_j)), P) + \varepsilon.\end{aligned}$$

Hence $\phi(int_w(\Omega_j), P) = \phi(cl_w((\Omega_j)), P)$. Therefore (23) holds for the finite union of the Ω_j's, as sought.

Some other class of sets $\Omega \subset \mathscr{M}$ for which (23) holds are defined as pre-images of continuous linear functions defined from $\mathscr{X}$ onto some Hausdorff topological space E. Adapting Theorem 4.1 in [28] we may state

Proposition 4. *Let $P \in \mathscr{M}_1$ and E be a real Hausdorff topological space; let $B_1 \subset B_2 \subset \ldots$ be an increasing sequence of Borel sets in supp(P) such that*

$$\lim_{m \to \infty} P(B_m) = 1.$$

Let $\Psi_m := \{Q \in \mathscr{M} : |Q|(B_m) = 1\}$ for all $m \in \mathbb{N}$ and $\mathscr{M}^* := \cup_m \Psi_m$. Let $T : \mathscr{M}^* \to E$ a function such that its restriction $T_{|\Psi_m}$ is linear and weakly continuous at each Q in $\mathscr{M}^*$ such that $\phi(Q,P) < \infty$ for each m. Let A be a convex set in E with $\phi\left(T^{-1}(intA), P\right) < \infty$. Then

$$\phi\left(T^{-1}(intA), P\right) = \phi\left(T^{-1}(clA), P\right). \tag{24}$$

Proof. It proceeds following nearly verbatim the Proof of Theorem 4.1 in [28]. Convexity arguments similar to the one in the Proof of Proposition 3 provide a version of (24) for sets $T^{-1}_{|\Psi_m}(A)$. Making use of Theorem 2, which substitutes Theorem 3.1 in [28] concludes the proof.

7 A simple convergence result and some perspectives

All estimators of $\phi(\Omega, P)$ considered in this paper converge strongly to $\phi(\Omega, P)$ as n tends to infinity, as does K. Indeed going back to the general setting presented in Section 1, for fixed n it clearly holds that

$$\lim_{K \to \infty} \frac{1}{K} card(i : X_{n,i} \in \Omega) = \Pr(X_n \in \Omega)$$

a.s.

When

$$\lim_{n\to\infty}\frac{1}{n}\log\Pr\left(X_n\in\Omega\right)=-\Phi\left(\Omega\right)$$

it follows that

$$\lim_{n\to\infty}\lim_{K\to\infty}\frac{1}{n}\log f_{n,K}=-\Phi\left(\Omega\right)\quad\text{a.s.}$$

as sought. Since the estimators of $KL\left(\Omega,P\right)$ and $KL_m\left(\Omega,P\right)$ considered in Section 3, as well as the estimators of $\phi_\gamma\left(\Omega,P\right)$ considered in Section 5 are obtained through continuous transformations of the former estimates, all estimators considered in the present article converge strongly to their respective limits as K tends to infinity and n tends to infinity. This leaves a large field of investigations wide open, such as the choice of some sequence $K=K_n$ which would lead to a single limit procedure. Also the resulting rate of convergence of these estimators as well as their distributional limit would be of interest.

Also Importance Sampling (IS) techniques should be investigated in order to reduce the calculation burden caused by the fact that any of the weighted empirical measures considered in this article would visit the set Ω quite rarely, if P does not belong to Ω. The hit rate can be increased substantially using some ad hoc modification of the weights, resulting from an IS strategy.

Once estimated the minimum value of the divergence, one may be interested in the identification of the measures Q which achieve this minimum in Ω. Dichotomous methods can be used, iterating the evaluation of the minimum divergence between P and subsets of Ω where the global infimum on Ω coincides with the local ones, leaving apart the subsets where they do not coincide, and iterating this routine.

References

1. Bar-Lev, S. K.; Enis, P. Reproducibility and natural exponential families with power variance functions. Ann. Statist. 14 (1986), no. 4, 1507–1522.
2. Broniatowski, M.; Keziou, A. Minimization of φ-divergences on sets of signed measures. Studia Sci. Math. Hungar. 43 (2006), no. 4, 403–442.
3. Broniatowski, M. Keziou, A. Parametric estimation and tests through divergences and the duality technique. J. Multivariate Anal. 100 (2009), no. 1, 16–36.
4. Broniatowski, M.; Keziou, A. Divergences and duality for estimation and test under moment condition models. J. Statist. Plann. Inference 142 (2012), no. 9, 2554–2573.
5. Broniatowski, M. Weighted sampling, maximum likelihood and minimum divergence estimators. Geometric science of information, 467–478, Lecture Notes in Comput. Sci., 8085, Springer, Heidelberg, 2013
6. Broniatowski, M; Decurninge, A. Estimation for models defined by conditions on their L-moments, to appear IEEE Transactions on Information Theory DOI:10.1109/TIT.2016.2586085
7. Csiszár, I. Eine informationstheoretische Ungleichung und ihre Anwendung auf den Beweis der Ergodizität von Markoffschen Ketten. (German) Magyar Tud. Akad. Mat. Kutató Int. Közl. 8 1963 85–108.
8. Read, T. R. C.; Cressie, N. A. C. Goodness-of-fit statistics for discrete multivariate data. Springer Series in Statistics. Springer-Verlag, New York, 1988. xii+211 pp. ISBN: 0-387-96682-X

9. Jiménez, R.; Shao, Y. On robustness and efficiency of minimum divergence estimators. Test 10 (2001), no. 2, 241–248.
10. Jørgensen, B. Exponential dispersion models. With discussion and a reply by the author. J. Roy. Statist. Soc. Ser. B 49 (1987), no. 2, 127–162.
11. Morris, C. N. Natural exponential families with quadratic variance functions. Ann. Statist. 10 (1982), no. 1, 65–80.
12. Letac, G.; Mora, M. Natural real exponential families with cubic variance functions. Ann. Statist. 18 (1990), no. 1, 1–37.
13. Mason D.M.; Newton M. A., A rank statistic approach to the consistency of a general bootstrap, *Ann. Statist.* Vol. 20 (1992), pp. 1611-1624.
14. Tweedie, M. C. K. Functions of a statistical variate with given means, with special reference to Laplacian distributions. Proc. Cambridge Philos. Soc. 43, (1947). 41–49.
15. Barbe, P., Bertail, P., “The Weighted Bootstrap,” Lecture Notes in Statistics (1995), Springer-Verlag, New York.
16. Barndorff-Nielsen, O. Information and exponential families in statistical theory. Wiley Series in Probability and Mathematical Statistics. John Wiley & Sons, Ltd., Chichester, 1978. ix+238 pp.
17. Beran, R. Minimum Hellinger distance estimates for parametric models. Ann. Statist. 5 (1977), no. 3, 445–463.
18. Bertail, P. Empirical likelihood in some semiparametric models. Bernoulli 12 (2006), no. 2, 299–331
19. Cressie, N.; Read, TR. C.Multinomial goodness-of-fit tests.J. Roy. Statist. Soc. Ser. B 46 (1984), no. 3, 440–464.
20. Csiszár, I.; Gamboa, F.; Gassiat, E. MEM pixel correlated solutions for generalized moment and interpolation problems. IEEE Trans. Inform. Theory 45 (1999), no. 7, 2253–2270.
21. Csiszár, I. Information-type measures of difference of probability distributions and indirect observations. Studia Sci. Math. Hungar. 2 1967 299–318.
22. Csiszár, I. On topology properties of f-divergences. Studia Sci. Math. Hungar. 2 1967 329–339.
23. Najim, J., “A Cramer type theorem for weighted random variables,” *Electron. J. Probab.,* Vol. 7 (2002).
24. Liese, F.; Vajda, I. Convex statistical distances. With German, French and Russian summaries. Teubner-Texte zur Mathematik [Teubner Texts in Mathematics], 95. BSB B. G. Teubner Verlagsgesellschaft, Leipzig, 1987. 224 pp. ISBN: 3-322-00428-7
25. Hoadley, A. B. On the probability of large deviations of functions of several empirical cdf’s. Ann. Math. Statist. 38 1967 360–381.
26. Keziou, A. Dual representation of φ-divergences and applications. C. R. Math. Acad. Sci. Paris 336 (2003), no. 10, 857–862.
27. Feller, W. An introduction to probability theory and its applications. Vol. II. Second edition John Wiley & Sons, Inc., New York-London-Sydney 1971 xxiv+669 pp
28. Groeneboom, P.; Oosterhoff, J.; Ruymgaart, F. H.Large deviation theorems for empirical probability measures. Ann. Probab. 7 (1979), no. 4, 553–586.
29. Lindsay, B. G. Efficiency versus robustness: the case for minimum Hellinger distance and related methods. Ann. Statist. 22 (1994), no. 2, 1081–1114.
30. Owen, A.Empirical likelihood ratio confidence regions. Ann. Statist. 18 (1990), no. 1, 90–120.
31. Withers, C. S.; Nadarajah, S. On the compound Poisson-gamma distribution. Kybernetika (Prague) 47 (2011), no. 1, 15–37.
32. Several applications of divergence criteria in continuous families, to appear Kybernetika (2012).
33. Rockafellar, R. Tyrrell Convex analysis. Princeton Mathematical Series, No. 28 Princeton University Press, Princeton, N.J. 1970 xviii+451 pp.
34. Ruschendorf, L. Projections of probability measures.Statistics 18 (1987), no. 1, 123-129.
35. Trashorras, J.; Wintenberger, O. Large deviations for bootstrapped empirical measures. Bernoulli 20 (2014), no. 4, 1845–1878.