Big data, heterogenity and empirical likelihood, with some applications to text simplification and nutritional logo selection.

P. Bertail*, T. Dumont*., E. Gautherat**, E. Issouani*, N. Rhomari+,M. Badaoui+, M. Zetlaoui*

* Université Paris-Nanterre, contact : melanie.zetlaoui@gmail.com

** Université de Reims, contact : emmanuelle.gautherat@gmail.com

+ Université d'Oujda, Maroc, contact: nrhomari@yahoo.fr

1 Context and main objectives

Context: This long term project involves several laboratories around the theory and the applications of the empirical likelihood and its generalisation, MODAL'X, University Paris-Nanterre , the university of Reims and the university of Oujda. Some participants have already worked on the subject together and would like to develop some tools in the field of big data, text mining and econometric analysis. Our group is also opened to collaborations with other members of the Labex on this subject.We would like to organize some working seminars and two workshops around these problems during the two coming years.

Main objectives

The increasing capacity to collect data has improved much faster than our ability to process and analyze big datasets. The availability of massive information in the big data era (on which statistical tools or machine-learning procedures could theoretically now rely on) strongly suggests to use subsampling techniques (Politis and Romano, 2001) as a remedy to the apparent intractability of learning from datasets of explosive size, in order to break the current computational barriers. Such an approach has been for instance developed in Kleiner et al. (2014). It is also at the core of some recent developments on survey sampling method in the framework of big data (see Bertail et al. (2014)(2015a)(2015b), Zetlaoui et al. (2017)). However one of the main difficulties with big data sets is the problem of heterogeneity of data. Heterogeneity of sources or of the data can make difficult or even dubious the use of subsamples if they are not controlled by some macro-variables. Indeed for many studies the data at hand is somehow not representative of the population of interest. For instance, data collected on internet do not respect the structure of the whole population. On the other hand we often have concurrent datasets of moderate or smaller sizes or exhaustive information which should allow to correct the big data structure.

The purpose of this project is to explore these possibilities with the help of empirical likelihood, a flexible tools to incorporate extra-information. The project is thus at the cross-road between statistics, survey sampling techniques and optimization problems.

Empirical likelihood is now a useful and classical method for constructing confidence regions for the value of some parameters in non-parametric or semi-parametric models, which allows to incorporate additional (margin) information. It has been introduced and studied by Owen (1988)(1990), see Owen(2001) for a complete overview and exhaustive references (until 2001). The now well-known idea of empirical likelihood consists in maximizing a profile likelihood supported by the data, under some model constraints and margin constraints. It can be seen as an extension of "model based likelihood" used in survey sampling when some marginal constraints are available (see Hartley and Rao(1968), Thomas, D. R. and Grunkemeier(1975)). 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), which enjoy the same Bartlett correctability as parametric likelihood (DiCiccio et al., 1991).

The purpose of this project is to develop these methods in four directions :

- study the validity of the method for infinite dimensional parameter or Banach valued parameters, when a lot of marginal constraints are taken into account in the optimization program. In particular we will establish dual representation of the optimization program related to empirical likelihood and its generalization (see part 3). This is particularly important for the applications to text simplification that we are dealing with.
- propose penalization methods when large parameter values (with dimension p bigger than n the sample size) are involved (see part 4). The choice of the divergence and of the penalization should be studied in details.
- obtain exact exponential bounds for penalized empirical likelihood (see part 4). This tools will allow to build finite sample confidence region eventually for large parameters.
- propose feasible gradient methods for big data eventually based on subsampling techniques (see part 4) . Since all the statistical aspects of the method actually rely on an internal optimization procedure, it is of prime importance to propose adequate interior point methods based on adequate divergence and subsampling techniques for the method to be practically implemented. This aspect will be particular sensible for big data.

The proposed methods will be applied to real data on two subjects

- text simplification via information theoretical tools (including empirical likelihood) to facilitate the access to web pages for the deaf people population
- the impact of nutritional logos on nutritional and consumption data.

2 Empirical likelihood for general parameters : a short review

Let $\mathcal{B}$ be some separable Banach space. Consider the framework of estimating a general functional $T(P)$ (see Von Mises,1936) on some probability measure convex space $\mathcal{P}$ on $\mathcal{B}$, large enough to contain the Dirac measures. Let $X_1, ..., X_n, ...$ be i.i.d. random variables, taking their value in $\mathcal{B}$ with common probability measure P in $\mathcal{P}$. The empirical probability measure is defined by

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

where the δ_{X_i} are Dirac measures at $X_i's$. The empirical likelihood ratio evaluated at $\theta = T(P)$ is defined by

$$R_{E,n}(\theta) = \sup_{Q_n \in \mathcal{P}_n} \left\{ \Pi_{i=1}^{n} \frac{dQ_n}{dP_n}(X_i),\ T(Q_n) = \theta \right\},$$

where $\mathcal{P}_n$ is the set of discrete probability measures dominated by P_n that is

$$\mathcal{P}_n = \{\widetilde{P}_n = \sum_{i=1}^{n} p_{i,n} \delta_{X_i}\ , p_{i,n} \geq 0,\ \sum_{i=1}^{n} p_{i,n} = 1\}.$$

This can be seen as a convex optimization under some possibly non convex constraints.

$T(P)$ may be the unique solution of some estimating equations $E_P f(X, T(P)) = 0$ (see Qin and Lawless, 1994). These equations will also include marginal constraints in this project : see an application on large datasets of this kind of idea in Crepet et al. (2009). In this case, the constraint becomes $E_{Q_n} f(X, \theta) = 0 = \sum p_{i,n} f(X_i, \theta)$ and the empirical likelihood boils down to the convex maximization program.

$$R_{E,n}(\theta) = \sup_{p_{i,n}, i=1,...,n} \left\{ \begin{array}{c} \frac{\Pi_{i=1}^{n} p_{i,n}}{1/n^n} \text{ under } \sum_{i=1}^{n} p_{i,n} f(X_i, \theta) = 0 \\ \sum_{i=1}^{n} p_{i,n} = 1,\ p_{i,n} \geq 0 \end{array} \right\}.$$

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 and Lawless, 1994). It can now be considered as an alternative to the generalized method of moments (GMM, see Hansen, 1982). 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 the empirical log-likelihood ratio is to interpretate 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. 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 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 Leonard, 2001a,b,c, Gamboa and Gassiat, 1996). Some generalizations of the empirical likelihood method have also been obtained by using Cressie-Read discrepancies. This has led to some econometric extensions known as "generalized empirical likelihood" (see Newey and Smith, 2004), even if the "likelihood" properties and in particular the Bartlett-correctability in these cases are lost (see DiCiccio et al., 1991). Bertail et al. (2014) have shown that Owen's original method in the case of the mean can be extended to any regular convex statistical divergence or φ^*-discrepancy (where φ^* is a regular convex function) under weak assumptions, for general Hadamard differentiable functionals. We call this method "empirical energy minimizers" by analogy to the theoretical probabilistic literature on the subject (see Leonard, 2001a,b,c and the references therein). A goal of the project will be to further explore these generalization.

3 A general view of empirical likelihood : generalization to process value parameters or Banach value parameters

We consider a measured space $(\mathcal{X}, \mathcal{A}, \mathcal{M})$ where $\mathcal{M}$ is a space of signed measures. 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 $\mu f = \int f d\mu$ and if μ is a density of probability, $\mu f = \mathbb{E}_\mu(f(X))$. 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 inferior (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 (renormalization) $\varphi(0) = 0$ 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 $m > 0$.

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 positive measure, is defined as follows:

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

For details on φ^*-discrepancies or divergences and some historical comments, see Liese and Vajda(1987), Leonard (2001). It is easy to check that Cressie-Read discrepancies fulfill these assumptions. Indeed, a Cressie-Read discrepancy can be seen as a φ^*-discrepancy, with φ^* given by:

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

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

For us, the main interest of φ^*-discrepancies lies on the following duality representation, which follows from results of Borwein and Lewis (1991) on convex functional integrals (see also Rockafellar, 1968).

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

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

then, we have the dual equality:

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

*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*

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

The same kind of results also holds when the number of constraints goes to infinity or even is infinite (see Leonard, 2001abc).

One purpose of this project will be to explore these kinds of duality theorem when the data has a very big dimension in comparison to the data size. In particular we will determine under which conditions on the constraints it is still possible to obtain such a dual representation (with no dual gap) when the number of constraints is infinite or belongs to some Banach space. A particular case of interest, is when the parameter of interest or the constraints belong to a set of function with a finite Vapnik dimension. We will consider in particular the case when $\mathbb{P}$ has discrete support which is the case of interest for general empirical discrepancy minimization as seen below.

3.1 Empirical optimization of φ^*-discrepancies when p is large

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 measure 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 $\mathbb{E}_{\mathbb{P}} f(X, \theta) = 0$, where f is now on 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 and Lawless, 1994).

For a given φ, we define, by analogy to the empirical likelihood problem, the quantity

$$\beta_n(\theta) = n \inf_{\{\mathbb{Q} \ll \mathbb{P}_n,\ \mathbb{E}_{\mathbb{Q}} f(X,\theta)=0\}} \{I_{\varphi^*}(\mathbb{Q}, \mathbb{P}_n)\}$$

We define the corresponding random confidence region

$$C_n(1-\alpha) = \left\{\theta \in \mathbb{R}^q \mid \exists \mathbb{Q} \ll \mathbb{P}_n \text{ with } \mathbb{E}_{\mathbb{Q}} f(X, \theta) = 0 \text{ and } n I_{\varphi^*}(\mathbb{Q}, \mathbb{P}_n) \leq \eta(\alpha)\right\},$$

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

$$\Pr(\theta \in C_n(1-\alpha)) = 1 - \alpha + o(1).$$

Define $\mathcal{M}_n = \{\mathbb{Q} \in \mathcal{M} \text{ with } \mathbb{Q} \ll \mathbb{P}_n\} = \{\mathbb{Q} = \sum_{i=1}^n q_i\ \delta_{X_i},\ (q_i)_{1 \leq i \leq n} \in \mathbb{R}^n\}$. Considering this set of measures, instead of a set of probabilities, can be partially explained by Theorem 1.

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

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

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

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

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

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

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

then, under an empirical qualification constraint, $C_n(1-\alpha)$ is a convex asymptotic confidence region with

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

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

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

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

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

As a parametric likelihood indexed by λ, it is easy to show that $2\beta_n(\theta)$ is asymptotically $\chi^2(q)$ when $n\to\infty$, if the variance of $f(X,\theta)$ is definite. It is also Bartlett-correctable since it is a likelihood in λ in its dual form (see Bertail, 2006). For a general discrepancy, the dual form is not a likelihood and may not be Bartlett-correctable, see DiCiccio et al. (1991).

Moreover, we necessarily have the $q_i's > 0$ and the optimisation program implies in this case that $\sum_{i=1}^n q_i = 1$, that is the solution is a probability, so that the qualification constraint essentially means that 0 belongs to the convex hull of the $f(X_i, \theta)$. This is in particular the reason which one may obtain very bad coverage probability for empirical likelihood

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

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

and

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

which is exactly the square of a self-normalized sum which typically appears in the Generalized Method of Moments (GMM).

Notice 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. Notice that in this case it is possible to get exact exponential bounds for this quantity as shown in Bertail et al. (2008). As a consequence even if $p << n$ but is of the same orderand grows with n,it is still possible to get an automatic confidence region, by just relying on the internal optimization problem, whithout having to invert the empirical covariance matrix (which may be complicated) for big datasets with a lot of variables. One purpose of this project we focus on what happens when the dimension in p is very large.

4 Generalized Empirical likelihood for big data

big data may be big according to different features. They can be tall because the number of individuals are important or they can be large because the dimension of the variables p is very big ($p >> n$) (see Bühlmann, van de Geer, S., 2011). Most of the time the tools to treat these two problems are totally different . In one case, one tries to reduce the size of the individuals either by subsampling or using some parallelized procedures : in these aspects the computational and optimization problems are of prime importance.. On the other case one generally makes some "sparsity assumptions" and try to select only a few components which are of importance. Most of the time, such tasks is performed by using some penalization procedure for instance the LASSO procedure (see the recent book by Hastie, T. , Tibshirani, T., Wainwright, 2016 and their references.).

The very flexible form of empirical likelihood and its generalization (in particular with the χ^2 discrepancy) makes it an ideal tool to treat the problem in an

unified way, without any parametric assumptions. It is indeed an ideal tools for incoporating extra-information when one wants to deal with large datasets on both dimensions, even if the data may not be representative of the population. The idea is simply to write the constraints induced by the model and to add the additional macro-constraints brought by the marginal information and to penalize the constraints to select the most important ones. Of course the optimization procedure may be very time consuming on very big datasets. For this reasons we propose to explore the feasibility and the validity of minibatch or subsampling techniques in the optimization problem as emphasized by Bertail et al (2016), Zetlaoui et al. (2017).

4.1 Penalizing the dual likelihood in large dimension

Several propositions have emerged to treat big data (large dimension) with empirical likelihood. We may classified them into too classes which will be studied precisely in this project.

i) **Enlarge-the-margin methods** : by this, we mean that instead of the original empirical likelihood problem, allow for some flexibility or some pertubations of the original constraints. This can be done either by adding one or several points to the data which do not have exactly the correct mean (see Chen, Variyath and Abraham (2008), Emerson and Owen (2009)). Or this can be done by replacing the original constraints by some inequality constraints with respect to some norm $||.||_R$ where R is possibly random allowing for some flexibility in the constraints. This leads to relaxed empirical likelihood version

$$R_{E,n}^{pen}(\theta) = \sup_{p_{i,n}, i=1,...,n} \left\{ \begin{array}{c} n^n \Pi_{i=1}^n p_{i,n} \text{ under } ||\sum_{i=1}^n p_{i,n} f(X_i,\theta)||_R \leq \delta_n \\ \sum_{i=1}^n p_{i,n} = 1,\ p_{i,n} \geq 0 \end{array} \right\} \quad (4)$$

where δ_n is a margin to be calibrated (possibly depending on the data).

ii) **Penalize empirical likelihood either on the primal form or the dual form.** It is well known in the convex literature that program 4 may also be rewritten

$$log(R_{E,n}^{pen}(\theta)) = \sup_{p_{i,n}, i=1,...,n} \left\{ \begin{array}{c} \sum_{i=1}^n \log(p_{i,n}) - \ C_n(\delta_n)||\sum_{i=1}^n p_{i,n} f(X_i,\theta)||_R \\ \sum_{i=1}^n p_{i,n} = 1,\ p_{i,n} \geq 0 \end{array} \right\}$$

which may be interpreted as a penalized version of the original program. Such penalizations have been studied in Bartollucci(2007) and Lahiri and Mukhopadhyay(2011) when $f(X_i,\theta) = X_i - \theta,\ X_i = (X_i^1, ...X_i^q) \in R^q$. The proposition of Bartolluci (2007) corresponds to the choice $R = \widehat{S}_n^{-2}$ and $||x||_R = x'\widehat{S}_n^{-2}x$, $C(\delta_n) = n/2h^2$, where $\widehat{S}_n^2$ is the sample covariance matrix

$$\widehat{S}_n^2 = \frac{1}{n}\sum_{i=1}^n (X_i - \overline{X}_n)(X_i - \overline{X}_n)'.$$

The proposition of Lahiri and Mukhopadhyay (2011) in a more general dependent framework corresponds to $R = diag(\sigma_{j_n}^{-2})_{i=1,...,q}$, $||x||_R = x'Rx$, $C(\delta_n) = \lambda$

where the

$$\sigma_{j_n}^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i^j - \overline{X}_n^j)^2$$

are the marginal empirical variances. One purpose of this project is to further explore the effect of the choice of the penalization by allowing $||.||_R$ to be some general norm.

Another proposition is to penalize the empirical likelihood in its dual form (see Mikland for an introduction to dual likelihood). Consider the penalized program

$$P_n(\theta,\lambda) = \mathbb{P}_n\left(-\lambda' f(.,\theta) - \varphi(\lambda' f(.,\theta))\right) - \frac{1}{2}||\lambda||_R^2.$$

which is clearly linked to the original penalized problem by duality consideration

$$\sup_{\lambda\in R^q}(P_n(\theta,\lambda)) = Inf_{p_i}(I_{\varphi^*}(\mathbb{Q},\mathbb{P}_n) + \frac{1}{2}||\mathbb{Q}f(.,\theta)||_{R^{-1}}^2).$$

We will further investigate the relations between these different dual formulations in our project in particular when one use the $L1$ or the L_∞ or a combination of these norms with L2 (elastic net).

iii) **Choose another divergence** (on space of signed measure). Another proposition is to use a different criterion than the likelihood criterion, arising from the choice of measuring the distance beetween Q_n and P_n with the Kullback-Leibner divergence say $KL(Q,P) = -\int \log\left(\frac{dQ}{dP}\right)$. It is known for instance that the choice of $\chi^2(Q,P) = \int(\frac{dQ}{dP}-1)^2$ leads to exact computation of the generalized empirical likelihood version provided that one works with signed measures $Q_n$dominated by P_n rather than probability measure (that is, one does not impose $p_{i,n}\geq 0$ and $\sum_{i=1}^n p_{i,n} = 1$). In this case, the maximization problem becomes

$$\begin{aligned} R_{\chi^2,n}(\theta) &= \sup_{p_{i,n},i=1,\ldots,n}\left\{\sum(\frac{p_{i,n}}{1/n}-1)^2 \text{ under}\sum_{i=1}^n p_{i,n}(X_i-\theta) = 0\right\}, \\ &= \frac{1}{2}(\overline{X}_n-\theta)' S_n^{2-}(\overline{X}_n-\theta) \end{aligned}$$

where $S_n^2 = \frac{1}{n}\sum(X_i-\theta)(X_i-\theta)'$ and S_n^{2-} is its Moore-Penrose generalized inverse. For general constraints, the solution is close to the so called GMM program.

Note that in the χ^2 case the dual problem when $R = C^{-1}.I$ for some constant C, and with the choice of à L2 penalization then the optimization program becomes

$$D = \sup_{b^*\in L_{\mathcal{F}}^*}\mathbb{P}_n\left\{-\lambda' f - (\lambda' f)^2/2 - C\lambda'\lambda\right\}$$

and the solution of this program is simply the regularized Hotteling statistics

$$\frac{1}{2}P_n f\,'(P_n f f' + CI)^{-1}P_n f$$

which is a regularized form of the T^2 Hottelling statistics (with no centering).

When the dimension $p << n$, Bertail et al. (2008) have shown that one can choose C=0 and can get some exact exponential bounds for this quantity. We would like to investigate conditions to obtain exponential bounds in this large dimension framework by choosing $C = C_n \to 0$. The GMM case when there is an infinite number (or a continuum) of constraints has been treated by several authors in the econometric literature, see for instance Carasco and Florens (2000), using some Tikhonov regularization of the operator S_n^2 (which somehow is the version (ii)). We want to show that this a special case of L_2 penalization of generalized empirical likelihood.

4.2 An optimization challenge

From an optimization point of view, according to the choice of the divergence and penalization, there exists very efficient interior point methods to solve the empirical likelihood optimization problem. However, since access to the data, may be time consuming we will explore the validity of subsampling techniques as used in Politis and Romano (1992). The idea underlying subsampling which is currently used for big data is to use the universal validity of the subsampling method as proved in Politis and Romano (1992), to extrapolate inferential methods to bigger size (see Bickel et al, 2008). Such ideas are not new and have also been put forward in some earlier works by Bickel and Yahav (1988) about bootstrap and Richardson extrapolation, when the computer capacities were not sufficient to treat even moderate sample size. Such methods are themselves related to well known numerical methods. In any case these optimization problems are very CPU expensive and time consumming : we will use GPU solutions and explore other alternatives.

5 Some applications

5.1 An application to text simplication

People with hearing loss have a vocabulary in sign language that does not evolve as quickly as French language. Hearing impairment impoverishes vocabulary, leading to mild illiteracy. In many cases, the syntax of websites (especially administrative sites that are essential to citizens) is inappropriate for people with hearing loss and inappropriate voice assistance services. The aim of this application is to contribute to the democratization of access to the web service for people with severe hearing loss and to allow the development of web accessibility through linguistic and semantic analysis using techniques from information

theory and statistical learning. A first step is to try to evaluate the complexity of web-sites. The most commonly adopted measure is based on the cross entropy between a language model seen as a tree connecting words and the actual unknown distribution of observed data, assuming that the data can be modelled by a stationary and ergodic phenomenon (typically a Markov chain or a hidden Markov chain of fixed length (n-gram). Such approach can be rephrased in term of empirical likelihood problems which are more robust to the distributional hypotheses as the sample size increases. The second step is to propose a technique for automatically reducing the content of web pages to make them accessible to the hearing impaired. Such task is also performed either by using information theoretical tools under a large number of constraints (on the features of texts) or neural networks tools. We would like to investigate the use of penalized empirical likelihood methods in this framework.

5.2 An application to nutritional logos

Although there is a growing body of evidence about consumer understanding, attention, and purchasing intention to front-of-pack (FOP) nutrition labels, comparative studies of their effectiveness on actual food purchases are scarce. Our objective is to provide empirical evidence on which FOP format is the more effective in improving the nutritional quality of food purchases in real-life shopping conditions. This study aims at taking into account all the informations and eventually the selection biais problems that could have appeared in the Ministry of Agriculture survey (this survey was controlled by the Fond Francais pour l'Alimentation et la Santé, done in 2016 under the statistical supervision of P. Bertail and P. Dubois). The conclusions of the study by Dubois et al (2017) shows that the Nutriscore is the preferred logo both from a perception and efficiency point of view when modeling the impact of logos on the FSA (Food Standard agency score), using first and double difference models taking into account prices and some socio-economic factors. However such models are very sensitive to biais selection problems and we would like to investigate the possible impact of such biases on the final results. A solution is to incoporate all the external information (eventually modelling bias) into the empirical likelihood as proposed for instance in Bertail (1997). Since we expect a lot of constraints to appear in the maximum empirical likelihood estimation problems, penalization will be required and we want to evaluate on this specific example the feasibility of our methods.

References

[1] Bartolucci, F. (2007). A penalized version of the empirical likelihood ratio for the population mean. *Statist. Probab. Lett.*. 77, 104–110.

[2] Bertail, P. (1997). Second Order Properties of an Extrapolated Bootstrap withoutreplacement under weak assumptions : the i.i.d. and strong mixing case. *Bernouilli*, 3, 149-179.

[3] Bertail, P., Politis, D., Romano, J. (1999). Undersampling with unknown rate of convergence. *Journal of the American. Statistical Association*, **94**(446), 569-579.

[4] Bertail, P., Politis, D., Rhomari N. (2000). Undersampling continuous random fields and a Bernstein inequality, *Statistics*, 33, 367-392.

[5] Bertail P., Haeffke, C., Politis D., White H. (2004). A subsampling approach to estimating the distribution of diverging statistics with applications to assessing financial market risks, *Journal of Econometrics*, **120**, 295-326.

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

[7] Bertail, P. and Gautherat, E. and Harari-Kermadec, H. (2008). Exponential bounds for self-normalized sums", *Electronic communications in probability*, 13, paper no. 57, 628-640.

[8] Bertail, P. and Gautherat, E. and Harari-Kermadec, H. (2014). Empirical phi-Divergence Minimizers for Hadamard Differentiable Functionals, In : *Topics in Nonparametric Statistics : Proceedings of the First Conference of the International Society for Nonparametric Statistics*. New York (USA) : Editions Springer (Springer Proceedings in Mathematics and Statistics, 74), 2014. 367 p.

[9] Bertail P. , E. Chautru, S. Clémençon (2014). Scaling-up M-estimation via sampling designs: the Horvitz-Thompson stochastic gradient descent. In the *Proceedings of the 2014 IEEE International Conference on big data*, Washington (USA).

[10] Bertail P., Chautru E., Clémençon S. (2015). Tail Index Estimation Based on Survey Data. *ESAIM Probability & Statistics*, 19, 28-59 .

[11] Bertail P., Chautru E., Clémençon S. (2015). Empirical processes in survey sampling. *Scandinavian Journal of Statistics*, 44(1), 97–111.

[12] Bickel, P. J. and J. A. Yahav. Richardson Extrapolation and the Bootstrap. *Journal of the American Statistical Association*. 83, No. 402, pp. 387-393, 1988.

[13] Bickel, P.J. , Boley, N. , Brown, J.B. , Huang, H., and N.R. Zhang (2010). Subsampling Methods for genomic inference, *Annals of Applied Statistics*, **4**, 1660-1697.

[14] Borwein, J.M. and Lewis, A.S. (1991). Duality relationships for entropy like minimization problem, *SIAM J. Optim.*, **29**, 325-338.

[15] Bühlmann, P. , van de Geer, S. (2011). *Statistics for High-Dimensional Data*, Springer-Verlag Berlin Heidelberg.

[16] Carrasco, M. and J. P. Florens (2000) Generalization of GMM to a continuum of moment conditions, *Econometric Theory,* 16, 797-834.

[17] Chen, J., Variath, A. M. & Abraham, B. (2008). Adjusted empirical likelihood and its properties. *J. Comput.Graph. Statist.,* 3, 426–443.

[18] Crépet, A. and Harari-Kermadec, H. and Tressou, Jessica (2009). Using empirical likelihood to combine data, Application to food risk assessment, *Biometrics,* 65(1), 257-266..

[19] DiCiccio,T., Hall, P. and Romano, J.(1991). Empirical Likelihood is Bartlett Correctable, *Ann. Statist.*, **19**, 1053-1061.

[20] Dubois, P., Allais, O., Albuquerque P., Bertail P., Bonnet P., Chandon P., Combris P., Lemmens A., Renaudin, N., and Ruffieux B. (2017). Impact of Different Front-of-pack Nutrition Labels on the Nutritional Quality of Food Purchases: Evidence from the French Randomized Control Experiment, working paper , submitted.

[21] Emerson S. C. and Owen, A. B. (2009). Calibration of the empirical likelihood method for a vector mean. Electron.J. Statist. 3, 1161–1192.

[22] Golan, A., Judge, G. and Miller, D.(1996). *Maximum Entropy Econometrics: Robust Estimation with Limited Data*, New York: John Wiley & Sons Inc.

[23] Hansen, L. (1982) Large Sample Properties of Generalized Method of Moments Estimators,Econometrica 50, 1029-1054.

[24] Hartley, H.O. and Rao, J.N.K. (1968). A New estimation Theory for Sample Survey, *Biometrika*, **55**, 547-57.

[25] Hastie, T. , Tibshirani, T., Wainwright, W (2016). Statistical Learning with Sparsity: The Lasso and Generalizations, CRC Monographs on Statistics & Applied Probability, Chapman & Hall.

[26] Hjort, N. L. and McKeague, I. W. and Van Keilegom, I. (2004). Extending the scope of empirical likelihood. Annals of statistics.

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

[28] Gill, R.D., Vardi, Y. and Wellner, J.A. (1988). Large Sample Theory of Empirical Distributions in Biased Sampling Models. *Ann. Statist.*, **16**, 1069-1112.

[29] Kleiner A., A. Talwalkar, P. Sarkar and M. I. Jordan (2014) A scalable bootstrap for massive data. Journal of the Royal Statistical Society: Series B, 76(4), 795–816.

[30] Lahiri,S. N. and Mukhopadhyay, S. (2011). A penalized empirical likelihood method in high dimensions. Ann.Statist. 40, 2511-2540.

[31] Léonard, C. (2001a). Convex conjugates of integral functionals. Acta Mathematica Hungarica, 93, 253-280.

[32] Léonard, C. (2001b). Minimization of energy functionals applied to some inverse problems, Applied mathematics and optimization, 44, 273-297.

[33] Léonard, C. (2001c). Minimizers of energy functionals, Acta Mathematica Hungarica, 93, 281-325.

[34] Liese F. and Vajda I. (1987). Convex Statistical Distances, Volume 95 de Teubner-Texte zur Mathematik, ed. Teubner.

[35] Mykland, P. (1995). Dual likelihood. *Ann. Statist.*, **23**, 396-421.

[36] Newey, W. K. and Smith, R. J. (2004). Higher Order Properties of {GMM} and Generalized Empirical Likelihood Estimators, Econometrica, 72, 219-255.

[37] Owen, A. B.(1988). Empirical Likelihood Ratio Confidence intervals for a Single Functional. Biometrika, 75, 237-249.

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

[39] Owen, A. B. (2001). Empirical likelihood. Chapman and Hall/CRC, Boca Raton

[40] Politis, D., Romano, J.P. and Wolf ,M. (1999). *Subsampling*. Springer Verlag.

[41] Qin, Y. S. and Lawless, J. (1994). Empirical likelihood and General Estimating Equations. Annals of Statistics, 22, 300-325

[42] Rockafeller, R. (1968). Integrals which are Convex Functionals, *Pacific J. Math.*, **24**, 525-339.

[43] Thomas, D. R. and Grunkemeier, G.L. (1975). Confidence Interval Estimation of Survival Probabilities for Censored Data, *J. Amer. Statist. Assoc.* , **70**, 865-871..

[44] von Mises, R. (1936). Les lois de Probabilités pour les Fonctions Statistiques, *Ann. Inst. H. Poincaré*, **6**, 185-212.

[45] Zetlaoui, M., Bertail, P., Jelassi, O., Tressou, J. (2017). Scaling by subsampling for big data, to appear ISNPS, recent advances in Non-parametric Statistics.