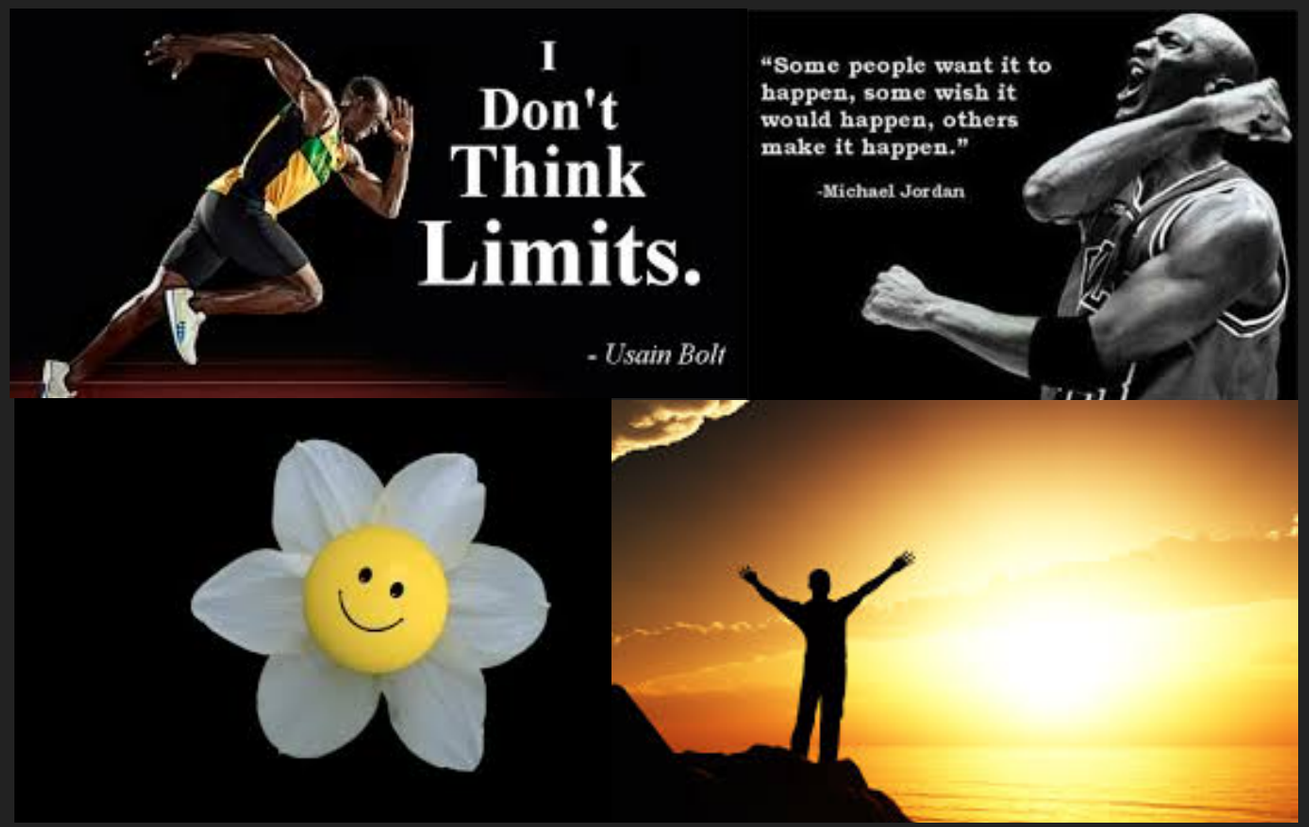
MATTHIEU LERASLE

LEARNING FROM MOM'S PRINCIPLES

OVERVIEW

- Introduction
- Basic ideas
- Toward learning theory
- Learning in small dimension
- Extension to high dimension
- Some perspectives



INTRODUCTION

OUR GOAL

- We are interested in sub-Gaussian estimators, that is estimators satisfying a sub-Gaussian deviation inequality :

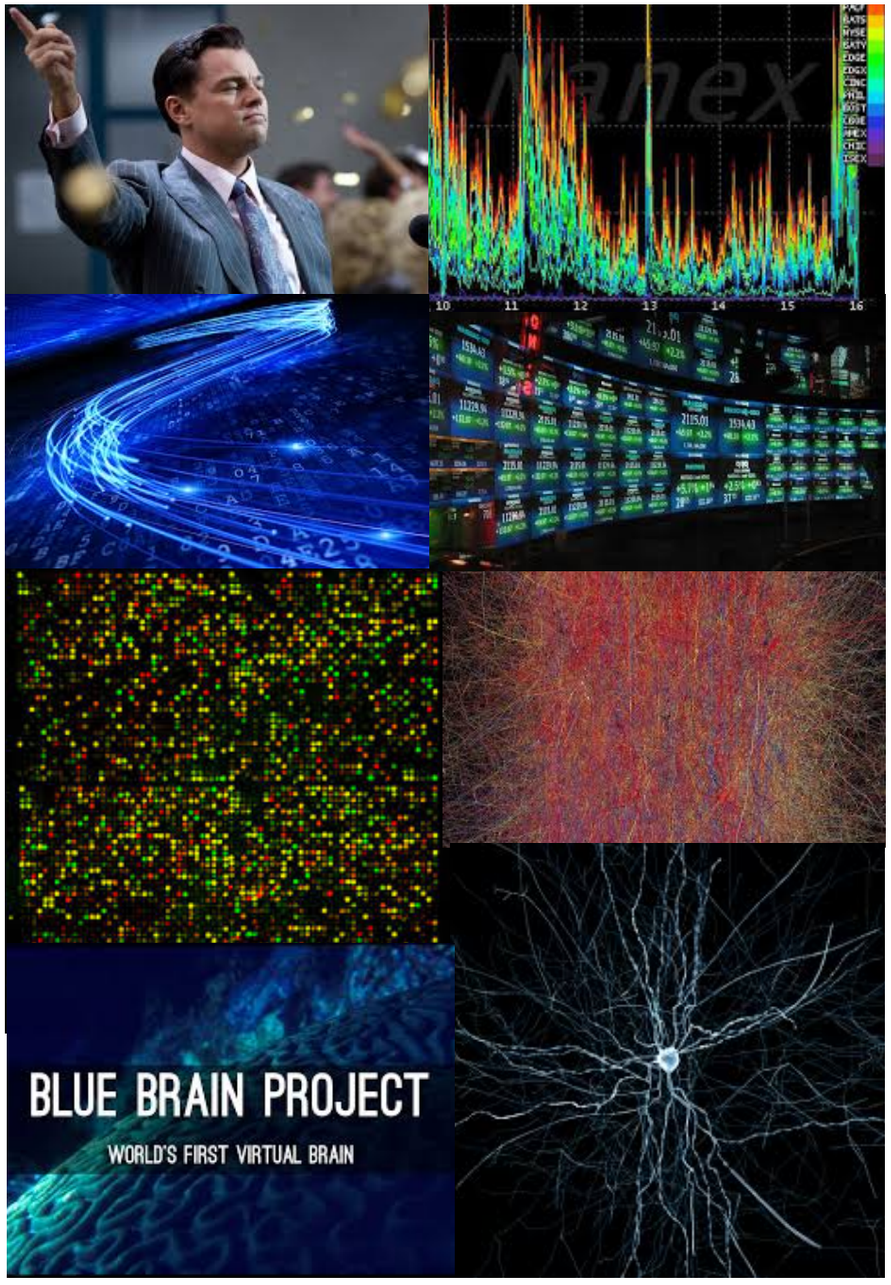
$$\forall x \leq x_n, \qquad \mathbb{P}\left(|\hat{E} - \mu| > C\sigma\sqrt{\frac{1+x}{n}}\right) \leq e^{-x} \quad .$$

- The empirical mean satisfies such inequalities under several assumptions like independence and identical distributions of the data and sub-Gaussian tails

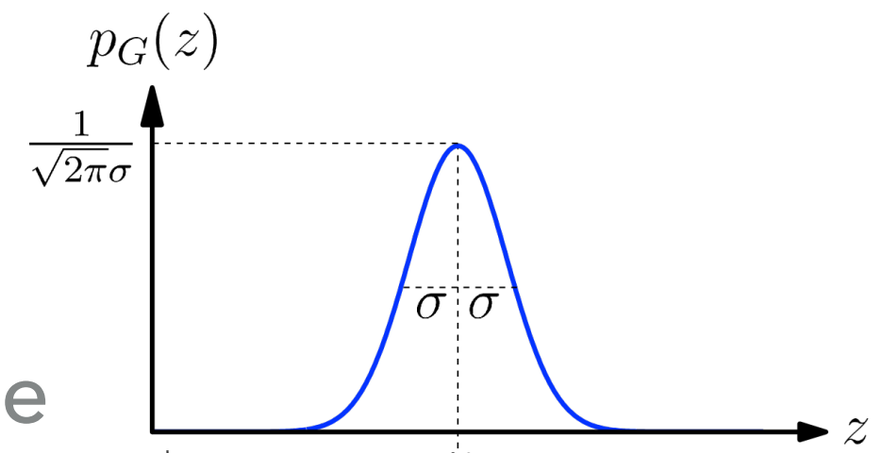
$$\forall \lambda > 0, \qquad \mathbb{E}[e^{\lambda X}] \leq e^{\frac{\lambda^2 \sigma^2}{C}} \quad .$$

MOTIVATION

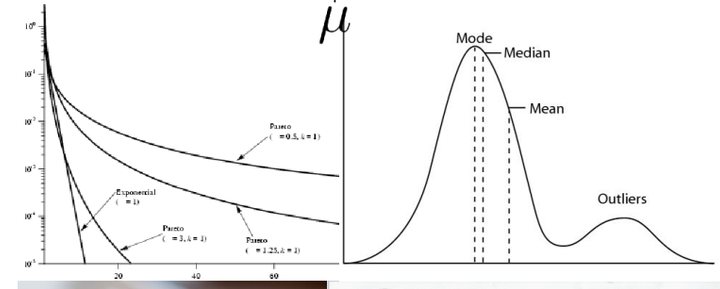
- Non independent, heavy-tailed data are common in high frequency trading.
- Data are corrupted in many applications in biology : micro-array analysis, neuro imaging.
- Robustness is a central issue in various modern applications!

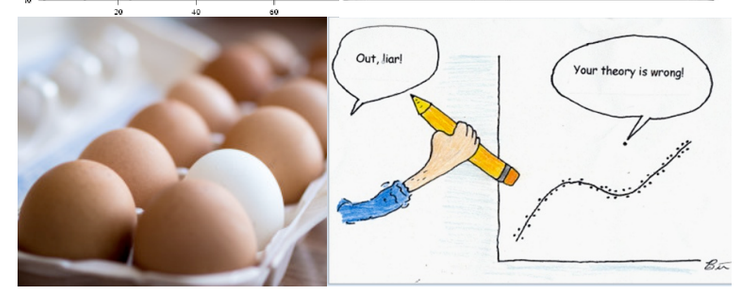


GOAL : BUILD ESTIMATORS (OF THE MEAN)



- with sub-Gaussian deviations up to exponentially low levels of confidence
- robust to « heavy-tailed » data
- robust to (a few) outliers…





WE CANNOT USE THE EMPIRICAL MEAN BECAUSE

Annales de l'Institut Henri Poincaré – Probabilités et Statistiques
2012, Vol. 48, No. 4, 1148–1185
DOI: 10.1214/11-AIHP454
© Association des Publications de l'Institut Henri Poincaré, 2012

www.imstat.org/aihp

Challenging the empirical mean and empirical variance: A deviation study

Olivier Catoni[a]

[a]CNRS – UMR 8553, Département de Mathématiques et Applications, École Normale Supérieure, 45 rue d'Ulm, F75230 Paris cedex 05, and INRIA Paris-Rocquencourt – CLASSIC team. E-mail: olivier.catoni@ens.fr

Received 13 September 2010; revised 12 August 2011; accepted 18 August 2011

Abstract. We present new M-estimators of the mean and variance of real valued random variables, based on PAC-Bayes bounds. We analyze the non-asymptotic minimax properties of the deviations of those estimators for sample distributions having either a bounded variance or a bounded variance and a bounded kurtosis. Under those weak hypotheses, allowing for heavy-tailed distributions, we show that the worst case deviations of the empirical mean are suboptimal. We prove indeed that for any confidence level, there is some M-estimator whose deviations are of the same order as the deviations of the empirical mean of a Gaussian statistical sample, even when the statistical sample is instead heavy-tailed. Experiments reveal that these new estimators perform even better than predicted by our bounds, showing deviation quantile functions uniformly lower at all probability levels than the empirical mean for non-Gaussian sample distributions as simple as the mixture of two Gaussian measures.

Résumé. Nous présentons de nouveaux M-estimateurs de la moyenne et de la variance d'une variable aléatoire réelle, fondés sur des bornes PAC-Bayésiennes. Nous analysons les propriétés minimax non-asymptotiques des déviations de ces estimateurs pour des distributions de l'échantillon soit de variance bornée, soit de variance et de kurtosis bornées. Sous ces hypothèses faibles, permettant des distributions à queue lourde, nous montrons que les déviations de la moyenne empirique sont dans le pire des cas sous-optimales. Nous prouvons en effet que pour tout niveau de confiance, il existe un M-estimateur dont les déviations sont du même ordre que les déviations de la moyenne empirique d'un échantillon Gaussien, même dans le cas où la véritable distribution de l'échantillon a une queue lourde. Le comportement expérimental de ces nouveaux estimateurs est du reste encore meilleur que ce que les bornes théoriques laissent prévoir, montrant que la fonction quantile des déviations est constamment en dessous de celle de la moyenne empirique pour des échantillons non Gaussiens aussi simples que des mélanges de deux distributions Gaussiennes.

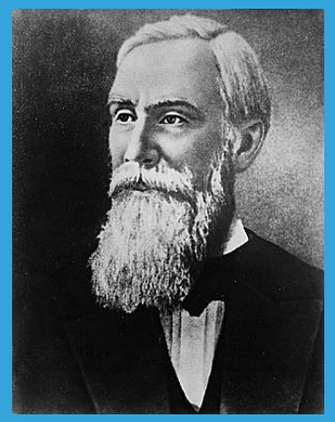
MSC: 62G05; 62G35

Keywords: Non-parametric estimation; M-estimators; PAC-Bayes bounds

1. Introduction

This paper is devoted to the estimation of the mean and possibly also of the variance of a real random variable from an independent identically distributed sample. While the most traditional way to deal with this question is to focus on the mean square error of estimators, we will instead focus on their deviations. Deviations are related to the estimation of confidence intervals which are of importance in many situations. While the empirical mean has an optimal minimax mean square error among all mean estimators in all models including Gaussian distributions, its deviations tell a different story. Indeed, as far as the mean square error is concerned, Gaussian distributions represent already the worst case, so that in the framework of a minimax mean least square analysis, no need is felt to improve estimators for non-Gaussian sample distributions. On the contrary, the deviations of estimators, and especially of the empirical mean, are worse for non-Gaussian samples than for Gaussian ones. Thus a deviation analysis will point

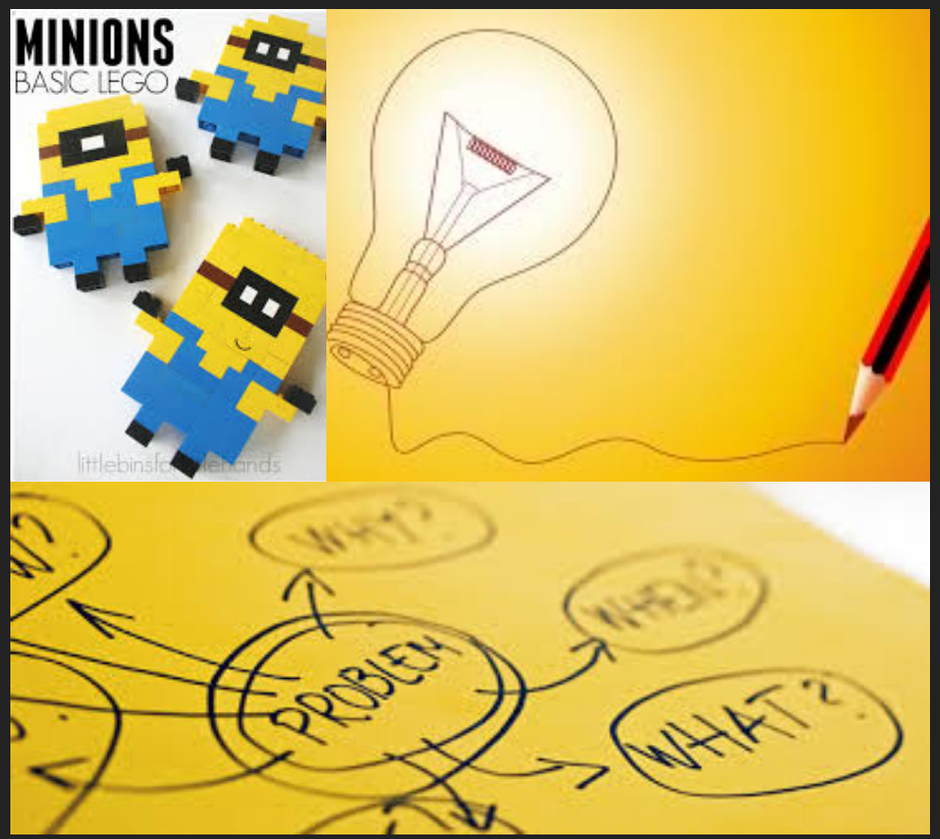
CHEBYCHEV'S INEQUALITY IS SHARP



[Catoni (2012)]

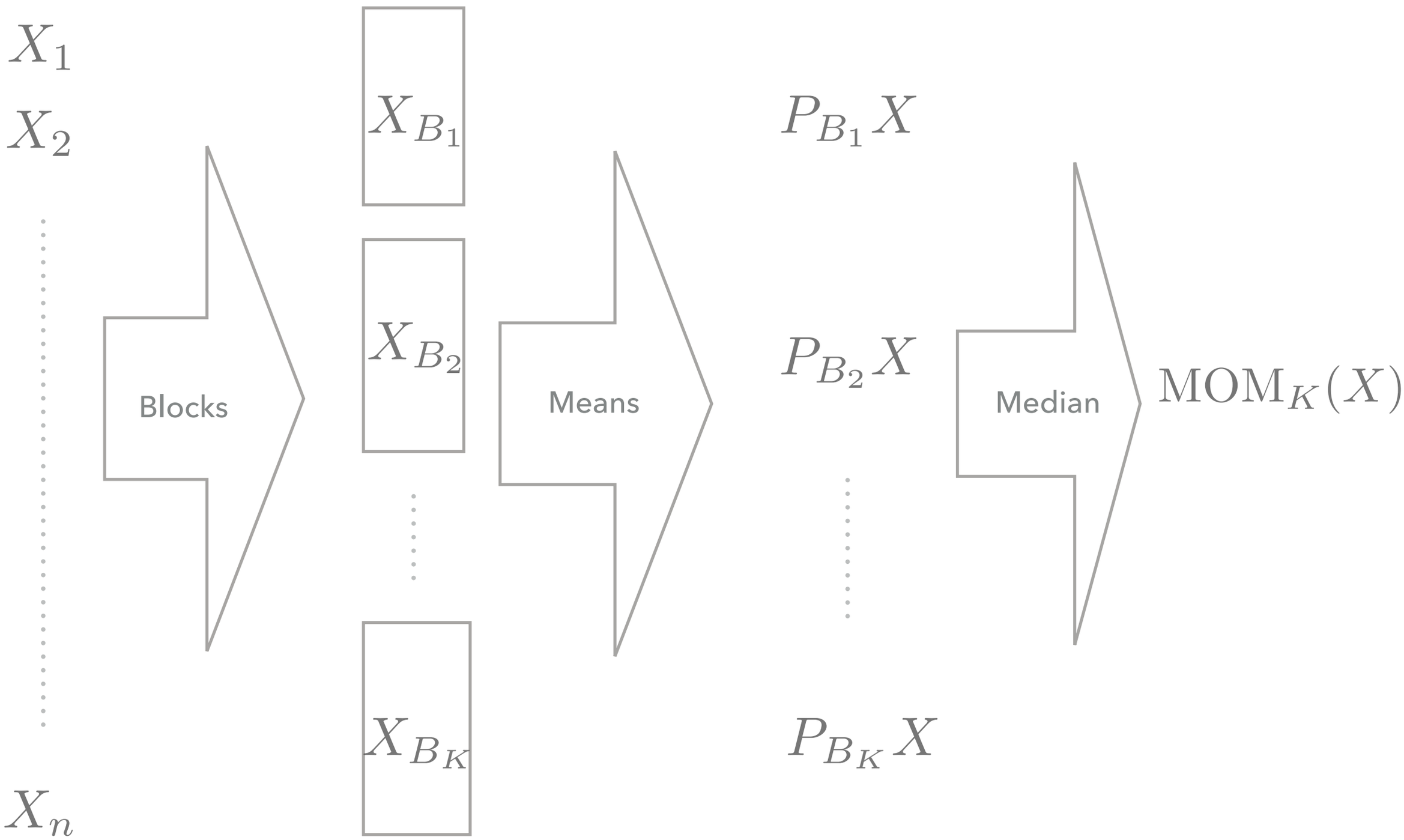
REPLACING THE EMPIRICAL MEAN

- « Trimming » ideas : getting rid of extremal points and take the empirical mean of the remaining data.
- « Quantile » regression.
- « mixing » L^2 and L^1 losses : Huber loss.
- Using robust tests : T-estimation, ρ-estimation.
- « smoothing » the loss : Catoni's loss.



BASIC IDEAS

THE MEDIAN OF MEANS ESTIMATORS



THE 2-LINES PROOF

- First, by definition :

$$\mathbb{P}\left(|\mathrm{MOM}_K(X) - PX| > C\sigma\sqrt{\frac{K}{n}}\right) \leq \mathbb{P}\left(\left|\left\{k : |(P_{B_k} - P)X| > C\sigma\sqrt{\frac{K}{n}}\right\}\right| \geq \frac{K}{2}\right)$$

$$\leq \sum_{k=K/2}^{K} \underbrace{\binom{K}{k}}_{\leq 2^K} \underbrace{p_K^k}_{\geq K/2} \underbrace{(1-p_K)^{K-k}}_{\leq 1} \leq 2^K p_K^{K/2} .$$

- Next, by Markov's inequality

$$p_K = \mathbb{P}\left(|P_{B_k}X - PX| > C\sigma\sqrt{\frac{K}{n}}\right) \leq \frac{1}{C^2} \leq \left(\frac{2}{C}\right)^K$$

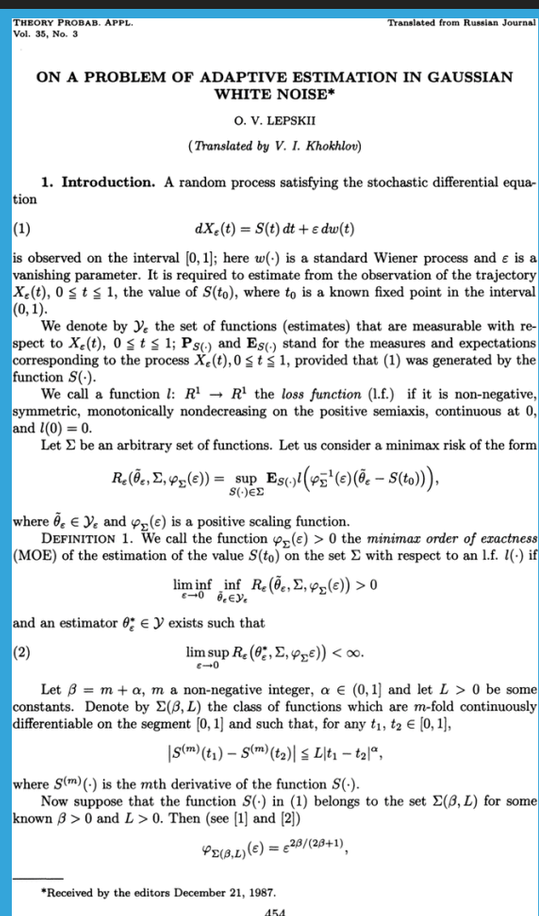
A PARTIAL RESULT

- taking $C = 2e$, we get

$$\mathbb{P}\left(|\mathrm{MOM}_K(X) - PX| > C\sigma\sqrt{\frac{K}{n}}\right) \leq e^{-K} \quad .$$

- Sub-Gaussian deviation inequality!!
- Need only 2 moments
- Extension to non-id data easy, robustness to (few) outliers
- BUT : need to fix the level to build the estimator

FORTUNATELY



THEORY PROBAB. APPL. Translated from Russian Journal
Vol. 35, No. 3

ON A PROBLEM OF ADAPTIVE ESTIMATION IN GAUSSIAN WHITE NOISE*

O. V. LEPSKII

(*Translated by V. I. Khokhlov*)

1. Introduction. A random process satisfying the stochastic differential equation

$$dX_\varepsilon(t) = S(t)\,dt + \varepsilon\,dw(t) \tag{1}$$

is observed on the interval $[0,1]$; here $w(\cdot)$ is a standard Wiener process and ε is a vanishing parameter. It is required to estimate from the observation of the trajectory $X_\varepsilon(t)$, $0 \leqq t \leqq 1$, the value of $S(t_0)$, where t_0 is a known fixed point in the interval $(0,1)$.

We denote by $\mathcal{Y}_\varepsilon$ the set of functions (estimates) that are measurable with respect to $X_\varepsilon(t)$, $0 \leqq t \leqq 1$; $\mathbf{P}_{S(\cdot)}$ and $\mathbf{E}_{S(\cdot)}$ stand for the measures and expectations corresponding to the process $X_\varepsilon(t), 0 \leqq t \leqq 1$, provided that (1) was generated by the function $S(\cdot)$.

We call a function l: $R^1 \to R^1$ the *loss function* (l.f.) if it is non-negative, symmetric, monotonically nondecreasing on the positive semiaxis, continuous at 0, and $l(0) = 0$.

Let Σ be an arbitrary set of functions. Let us consider a minimax risk of the form

$$R_\varepsilon(\tilde\theta_\varepsilon, \Sigma, \varphi_\Sigma(\varepsilon)) = \sup_{S(\cdot)\in\Sigma} \mathbf{E}_{S(\cdot)} l\Big(\varphi_\Sigma^{-1}(\varepsilon)\big(\tilde\theta_\varepsilon - S(t_0)\big)\Big),$$

where $\tilde\theta_\varepsilon \in \mathcal{Y}_\varepsilon$ and $\varphi_\Sigma(\varepsilon)$ is a positive scaling function.

DEFINITION 1. We call the function $\varphi_\Sigma(\varepsilon) > 0$ the *minimax order of exactness* (MOE) of the estimation of the value $S(t_0)$ on the set Σ with respect to an l.f. $l(\cdot)$ if

$$\liminf_{\varepsilon\to 0} \inf_{\tilde\theta_\varepsilon \in \mathcal{Y}_\varepsilon} R_\varepsilon\big(\tilde\theta_\varepsilon, \Sigma, \varphi_\Sigma(\varepsilon)\big) > 0$$

and an estimator $\theta_\varepsilon^* \in \mathcal{Y}$ exists such that

$$\limsup_{\varepsilon\to 0} R_\varepsilon\big(\theta_\varepsilon^*, \Sigma, \varphi_\Sigma \varepsilon)\big) < \infty. \tag{2}$$

Let $\beta = m + \alpha$, m a non-negative integer, $\alpha \in (0,1]$ and let $L > 0$ be some constants. Denote by $\Sigma(\beta, L)$ the class of functions which are m-fold continuously differentiable on the segment $[0,1]$ and such that, for any $t_1, t_2 \in [0,1]$,

$$\big|S^{(m)}(t_1) - S^{(m)}(t_2)\big| \leqq L|t_1 - t_2|^\alpha,$$

where $S^{(m)}(\cdot)$ is the mth derivative of the function $S(\cdot)$.

Now suppose that the function $S(\cdot)$ in (1) belongs to the set $\Sigma(\beta, L)$ for some known $\beta > 0$ and $L > 0$. Then (see [1] and [2])

$$\varphi_{\Sigma(\beta,L)}(\varepsilon) = \varepsilon^{2\beta/(2\beta+1)},$$

*Received by the editors December 21, 1987.

ADAPTIVE ESTIMATORS?

USE MY METHOD!

[Lepski (1990)]

LEPSKI'S METHOD

- Start from a collection of confidence intervals $(\hat{I}_K)_{K=1,\ldots,n/2}$.
- Consider the smallest index

$$\hat{K} = \inf\left\{K : \cap_{J=K}^{n/2} \hat{I}_J \neq \emptyset\right\} \quad .$$

- Pick as an estimator the mid-point

$$\hat{E} \in \cap_{J=\hat{K}}^{n/2} \hat{I}_J \quad .$$

- $\hat{I}_K = \left[\mathrm{MOM}_K(X) \pm C\sigma\sqrt{\frac{K}{n}}\right]$ are confidence intervals for PX.

If the variance is known!!

THE 2-LINES PROOF

- Let $x < n/2 - 2$ and $K = [x] + 2$. Consider the event

$$\Omega_{\text{good}} = \left\{ PX \in \cap_{J=K}^{n/2} \hat{I}_J \right\} \quad .$$

- Ω_{good} has probability larger than

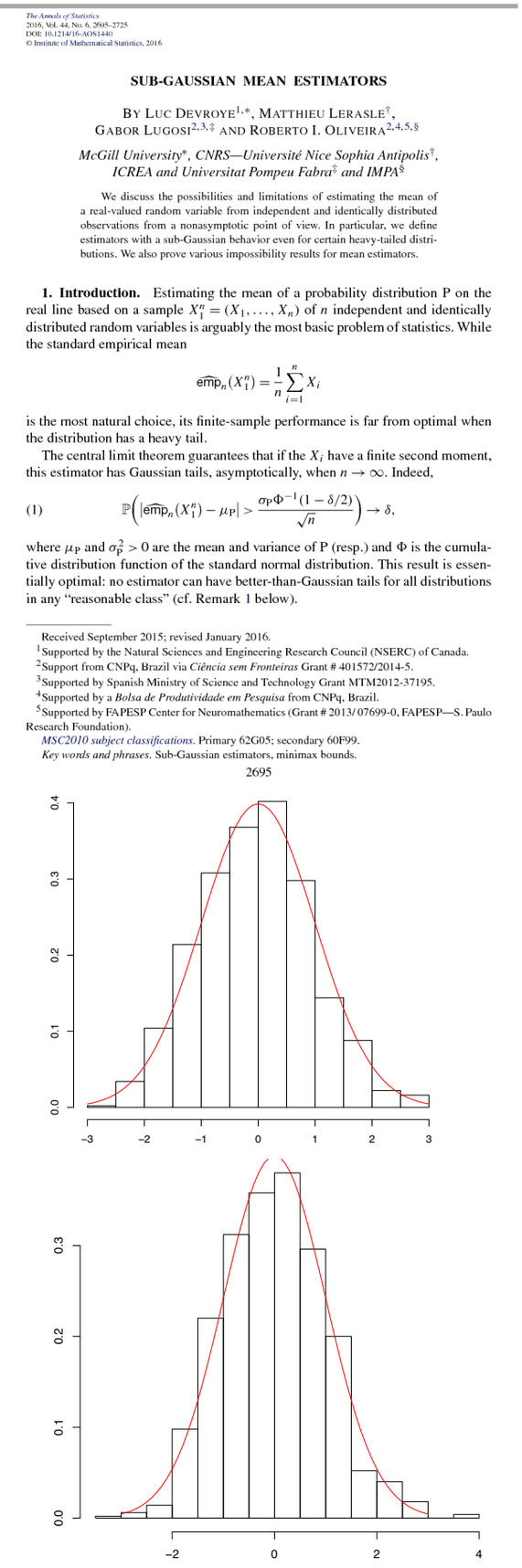
$$1 - \sum_{J=K}^{n/2} e^{-J} \geq 1 - e^{1-K} \geq 1 - e^{-x} \quad .$$

- On this event $\hat{K} \leq K$ thus $PX, \hat{E} \in \cap_{J=K}^{n/2} \hat{I}_J$, so

$$|PX - \hat{E}| \leq C\sigma\sqrt{\frac{K}{n}} \leq C\sigma\sqrt{\frac{2+x}{n}} \quad .$$

FOR THE ESTIMATION OF THE MEAN

- We have built sub-Gaussian estimators of the mean of a real valued random variables.
- The method only requires finite second moment.
- The exponentially small confidence level cannot be beaten.
- Building Sub-Gaussian estimators not depending on the confidence level is impossible without extra knowledge.
- The optimal constant $C = \sqrt{2}$ [Catoni 2012] requires more work.



The Annals of Statistics
2016, Vol. 44, No. 6, 2695–2725
DOI: 10.1214/16-AOS1440

SUB-GAUSSIAN MEAN ESTIMATORS

BY LUC DEVROYE[1,*], MATTHIEU LERASLE[†], GABOR LUGOSI[2,3,‡] AND ROBERTO I. OLIVEIRA[2,4,5,§]

McGill University, CNRS—Université Nice Sophia Antipolis[†], ICREA and Universitat Pompeu Fabra[‡] and IMPA[§]*

We discuss the possibilities and limitations of estimating the mean of a real-valued random variable from independent and identically distributed observations from a nonasymptotic point of view. In particular, we define estimators with a sub-Gaussian behavior even for certain heavy-tailed distributions. We also prove various impossibility results for mean estimators.

1. Introduction. Estimating the mean of a probability distribution P on the real line based on a sample $X_1^n = (X_1, \dots, X_n)$ of n independent and identically distributed random variables is arguably the most basic problem of statistics. While the standard empirical mean

$$\widehat{\mathsf{emp}}_n(X_1^n) = \frac{1}{n}\sum_{i=1}^n X_i$$

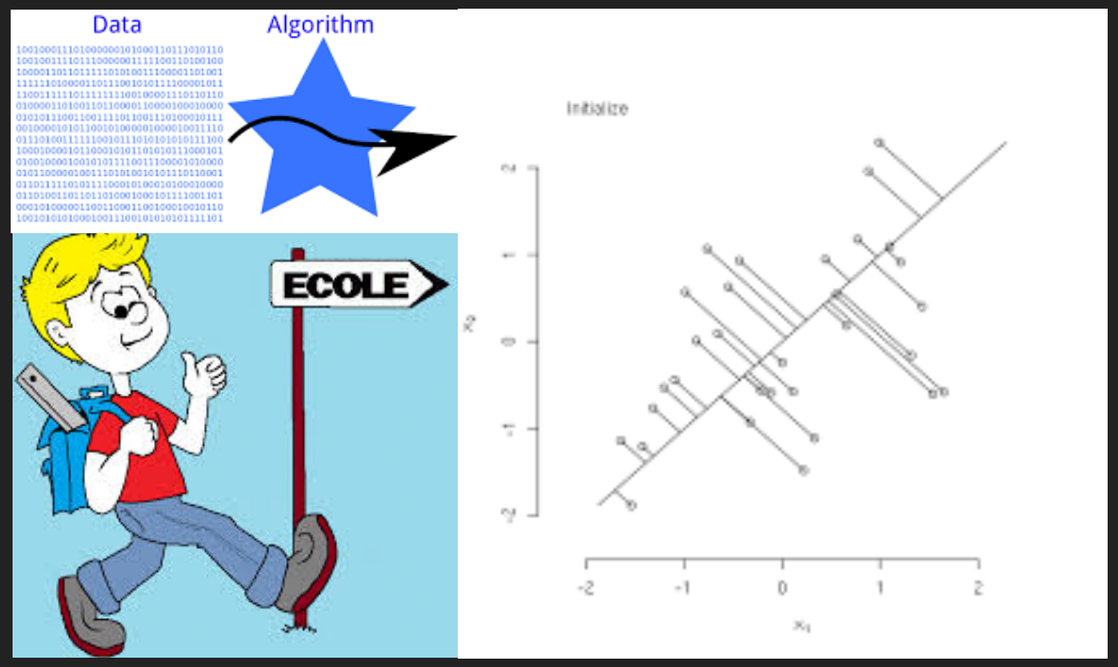
is the most natural choice, its finite-sample performance is far from optimal when the distribution has a heavy tail.

The central limit theorem guarantees that if the X_i have a finite second moment, this estimator has Gaussian tails, asymptotically, when $n \to \infty$. Indeed,

$$\mathbb{P}\left(|\widehat{\mathsf{emp}}_n(X_1^n) - \mu_{\mathrm{P}}| > \frac{\sigma_{\mathrm{P}}\Phi^{-1}(1-\delta/2)}{\sqrt{n}}\right) \to \delta, \tag{1}$$

where μ_{P} and $\sigma_{\mathrm{P}}^2 > 0$ are the mean and variance of P (resp.) and Φ is the cumulative distribution function of the standard normal distribution. This result is essentially optimal: no estimator can have better-than-Gaussian tails for all distributions in any "reasonable class" (cf. Remark 1 below).

Received September 2015; revised January 2016.
[1]Supported by the Natural Sciences and Engineering Research Council (NSERC) of Canada.
[2]Support from CNPq, Brazil via *Ciência sem Fronteiras* Grant # 401572/2014-5.
[3]Supported by Spanish Ministry of Science and Technology Grant MTM2012-37195.
[4]Supported by a *Bolsa de Produtividade em Pesquisa* from CNPq, Brazil.
[5]Supported by FAPESP Center for Neuromathematics (Grant # 2013/ 07699-0, FAPESP—S. Paulo Research Foundation).
MSC2010 subject classifications. Primary 62G05; secondary 60F99.
Key words and phrases. Sub-Gaussian estimators, minimax bounds.



TOWARD LEARNING THEORY

ESTIMATE THE MINIMUM OF A INTEGRATED CONTRAST

- We can write

$$PX = \arg\min_{\mu \in \mathbb{R}} P[(X-\mu)^2] \ .$$

- We use the contrast to compare candidates μ and ν. Ideally, we would prefer ν if

$$P[(X-\mu)^2 - (X-\nu)^2] > 0 \ .$$

- This ideal comparison can be replaced by a test using MOM's estimators.

T-AGGREGATED MOM'S TESTS

- We shall say that μ beats ν any time

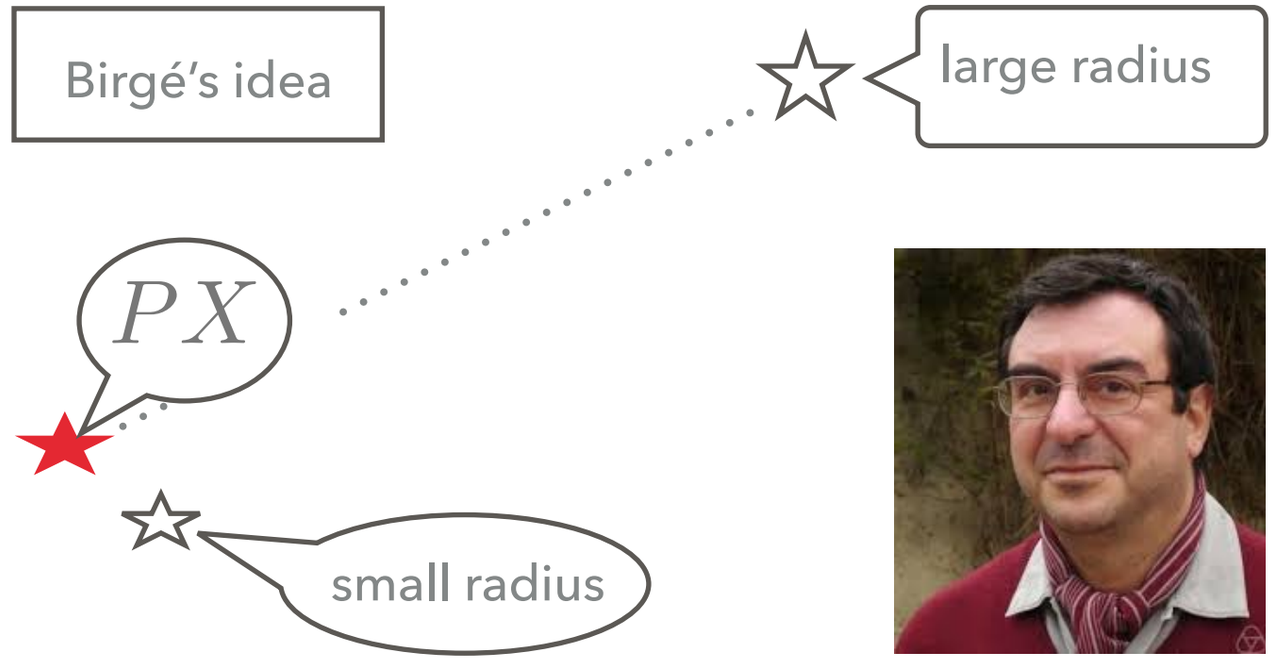
$$\mathrm{MOM}_K[(X-\nu)^2-(X-\mu)^2]>0 \ .$$

- To any real μ we associate the set $\mathcal{B}_K(\mu)$ of reals ν that beat μ and define

$$C_K(\mu)=\max_{\nu\in\mathcal{B}_K(\mu)}|\mu-\nu|$$

$$\hat{E}_K\in\arg\min_{\mu\in\mathbb{R}}C_K(\mu)$$

Birgé's idea



Risk minimization by median-of-means tournaments *

Gábor Lugosi†‡ Shahar Mendelson §

August 3, 2016

Abstract

We consider the classical statistical learning/regression problem, when the value of a real random variable Y is to be predicted based on the observation of another random variable X. Given a class of functions $\mathcal{F}$ and a sample of independent copies of (X,Y), one needs to choose a function $\hat{f}$ from $\mathcal{F}$ such that $\hat{f}(X)$ approximates Y as well as possible, in the mean-squared sense. We introduce a new procedure, the so-called median-of-means tournament, that achieves the optimal tradeoff between accuracy and confidence under minimal assumptions, and in particular outperforms classical methods based on empirical risk minimization.

1 Introduction

Estimation and prediction problems are of central importance in statistics and learning theory. In the standard regression setup, (X,Y) is a pair of random variables: X takes its values in some (measurable) set $\mathcal{X}$ and is distributed according to an unknown probability measure μ, while Y is real valued that is also unknown. Given a class $\mathcal{F}$ of real-valued functions defined on $\mathcal{X}$, one wishes to find $f\in\mathcal{F}$ for which $f(X)$ is a good prediction of Y. Although one may consider various notions of 'a good prediction', we restrict our attention to the–perhaps most commonly used–*squared error*: the learner is penalized by $(f(X)-Y)^2$ for predicting $f(X)$ instead of Y. Thus, one would like to find a function $f\in\mathcal{F}$ for which the expected loss $\mathbb{E}(f(X)-Y)^2$, known as the *risk*, is as small as possible. Naturally, the best performance one may hope for is of the risk minimizer in the class, that is, that of

$$f^*=\operatorname*{argmin}_{f\in\mathcal{F}}\mathbb{E}(f(X)-Y)^2 \ .$$

We assume in what follows that the minimum is attained and $f^*\in\mathcal{F}$ exists and is unique, as is the case when $\mathcal{F}\subset L_2(\mu)$ is a closed, convex set.

One may formulate two natural goals in estimation and prediction problems. One of them is to find a function $f\in\mathcal{F}$ whose $L_2(\mu)$ distance to f^*

$$\left(\mathbb{E}\left(f(X)-f^*(X)\right)^2\right)^{1/2} \qquad (1.1)$$

BIRGE'S ESTIMATOR IS MOM'S ESTIMATOR

- μ beats ν any time

$$\begin{aligned} 0 &< \nu^2 - \mu^2 - 2\mathrm{MOM}_K(X)(\nu - \mu) \\ &= (\nu - \mathrm{MOM}_K(X))^2 - (\mu - \mathrm{MOM}_K(X))^2 \ . \end{aligned}$$

- If $\mu = \mathrm{MOM}_K(X)$, this inequality is equivalent to

$$(\nu - \mathrm{MOM}_K(X))^2 > 0 \ .$$

- In other words, $\mathrm{MOM}_K(X)$ beats everyone, so

$$\hat{E}_K = \mathrm{MOM}_K(X) \ .$$

- MOM's estimators can be extended to learning problems !

BARAUD-BIRGE-SART PROCEDURE

- We consider Birgé's T-estimator based on MOM's tests. We could also at this stage consider the ρ-estimator :

$$\hat{E}_K \in \arg\min_{\nu} \left\{ \sup_{\mu} \mathrm{MOM}_K[(X-\nu)^2 - (X-\mu)^2] \right\} \quad .$$

- It is easy to check that we also have $\hat{E}_K = \mathrm{MOM}_K(X)$.

Model selection via testing:
an alternative to (penalized) maximum likelihood estimators

Lucien Birgé

UMR 7599 "Probabilités et modèles aléatoires", Laboratoire de Probabilités, boîte 188, Université Paris VI, 4, place Jussieu, 75252 Paris cedex 05, France

Received 9 July 2003; received in revised form 28 February 2005; accepted 12 April 2005

Available online 18 November 2005

Abstract

This paper is devoted to the definition and study of a family of model selection oriented estimators that we shall call T-estimators ("T" for tests). Their construction is based on former ideas about deriving estimators from some families of tests due to Le Cam [L.M. Le Cam, Convergence of estimates under dimensionality restrictions, Ann. Statist. 1 (1973) 38–53 and L.M. Le Cam, On local and global properties in the theory of asymptotic normality of experiments, in: M. Puri (Ed.), Stochastic Processes and Related Topics, vol. 1, Academic Press, New York, 1975, pp. 13–54] and Birgé [L. Birgé, Approximation dans les espaces métriques et théorie de l'estimation, Z. Wahrscheinlichkeitstheorie Verw. Gebiete 65 (1983) 181–237, L. Birgé, Sur un théorème de minimax et son application aux tests, Probab. Math. Statist. 3 (1984) 259–282 and L. Birgé, Stabilité et instabilité du risque minimax pour des variables indépendantes équidistribuées, Ann. Inst. H. Poincaré Sect. B 20 (1984) 201–223] and about complexity based model selection from Barron and Cover [A.R. Barron, T.M. Cover, Minimum complexity density estimation, IEEE Trans. Inform. Theory 37 (1991) 1034–1054].

It is well-known that maximum likelihood estimators and, more generally, minimum contrast estimators do suffer from various weaknesses, and their penalized versions as well. In particular they are not robust and they require restrictive assumptions on both the models and the underlying parameter set to work correctly. We propose an alternative construction, which derives an estimator from many simultaneous tests between some probability balls in a suitable metric space. In many cases, although not in all, it results in a penalized M-estimator restricted to a suitable countable set of parameters.

On the one hand, this construction should be considered as a theoretical rather than a practical tool because of its high computational complexity. On the other hand, it solves many of the previously mentioned difficulties provided that the tests involved in our construction exist, which is the case for various statistical frameworks including density estimation from i.i.d. variables or estimating the mean of a Gaussian sequence with a known variance. For all such frameworks, the robustness properties of our estimators allow to deal with minimax estimation and model selection in a unified way, since bounding the minimax risk amounts to performing our method with a single, well-chosen, model. This results, for those frameworks, in simple bounds for the minimax risk solely based on some metric properties of the parameter space. Moreover the method applies to various statistical frameworks and can handle essentially all types of models, linear or not, parametric and non-parametric, simultaneously. It also provides a simple way of aggregating preliminary estimators.

From these viewpoints, it is much more flexible than traditional methods and allows to derive some results that do not presently seem to be accessible to them.

A NEW METHOD FOR ESTIMATION AND MODEL SELECTION: ρ-ESTIMATION

Y. BARAUD, L. BIRGÉ, AND M. SART

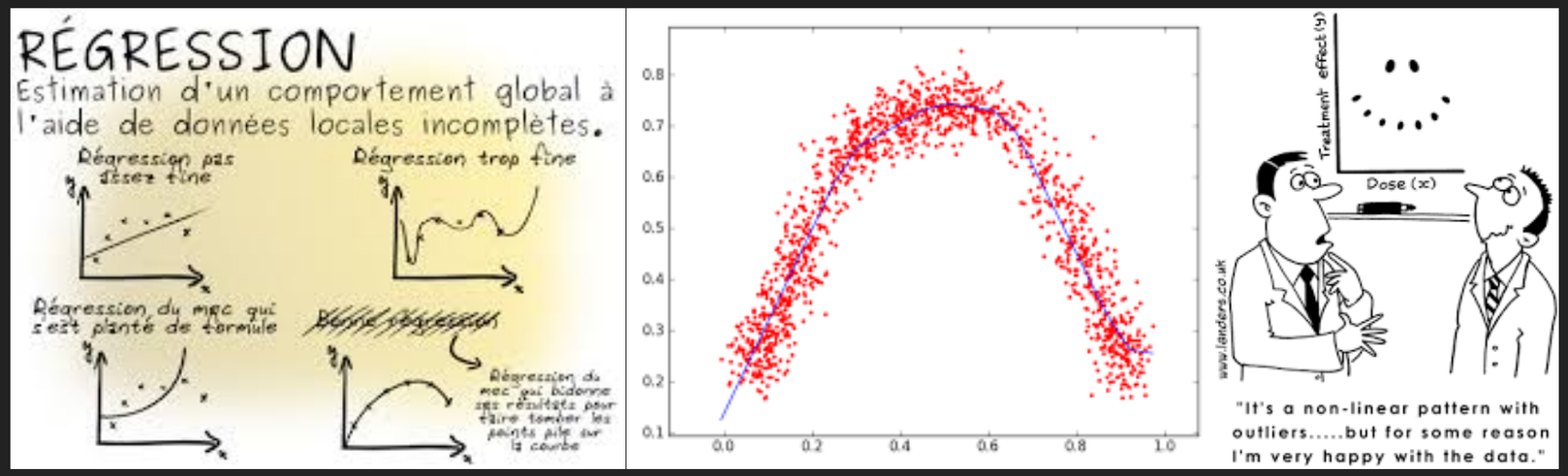
Abstract. The aim of this paper is to present a new estimation procedure that can be applied in various statistical frameworks including density and regression and which leads to both robust and optimal (or nearly optimal) estimators. In density estimation, they asymptotically coincide with the celebrated maximum likelihood estimators at least when the statistical model is regular enough and contains the true density to estimate. For very general models of densities, including non-compact ones, these estimators are robust with respect to the Hellinger distance and converge at optimal rate (up to a possible logarithmic factor) in all cases we know. In the regression setting, our approach improves upon the classical least squares in many respects. In simple linear regression for example, it provides an estimation of the coefficients that are both robust to outliers and simultaneously rate-optimal (or nearly rate-optimal) for a large class of error distributions including Gaussian, Laplace, Cauchy and uniform among others.

1. Introduction

The primary scope of this paper was to design a new and more or less universal estimation method for the regression framework where we observe n independent real random variables $X_1, \ldots, X_n$ of the form $X_i = f_i + \varepsilon_i$ where the f_i are the unknown parameters of interest and the ε_i i.i.d. real random errors with a partially unknown distribution which may be quite different from the usual Gaussian one. The problem arose from a question by Oleg Lepski to the first author during his visit to Nice in January 2012. This question was about the regression framework when the errors have rather unusual distributions, in which case the classical least squares method can be far from optimal. That was the starting point of our study which finally resulted in a much broader approach and the design of a new class of estimators with several remarquable and partly unexpected properties.

LUGOSI AND MENDELSON'S APPROACH

- Compute an upper bound r^* on $C_K(PX)$for a good choice of K=K* on an event of large probability.
- Call « champion » any μ such that $C_K(\mu) \leq r^*$.
- Estimate PX by a champion.
- Of course, the T-estimator is a champion on the same event, but it is also always defined and its definition does not require the knowledge of r^* or K*.



LEARNING IN SMALL DIMENSION

SETTING

- Let $(X, Y), (X_i, Y_i)_{i=1,\ldots,n}$ denote i.i.d. observations taking values in $\mathbb{R}^p \times \mathbb{R}$ with common unknown distribution P. Let

$$\beta^* \in \arg\min_{\beta \in \mathbb{R}^p} P(Y - X^T\beta)^2, \qquad \xi_i = Y_i - X_i^T\beta^* \ .$$

- To compare β and β', we use the following decomposition of the least-squares contrast

$$(Y - X^T\beta)^2 - (Y - X^T\beta')^2 = [X^T(\beta - \beta')]^2 - 2(Y - X^T\beta')X^T(\beta - \beta') \ .$$

- We assume to simplify that the distribution of X is known.

$$\|\beta\|^2_{L^2(P^X)} = P[(X^T\beta)^2] = \beta^T P(XX^T)\beta \ .$$

EXTENSION OF MOM'S PRINCIPLES

- Consider, for any β, the set $\mathcal{B}_K(\beta)$ of those β' such that

$$\mathrm{MOM}_K\left[(X^T(\beta-\beta'))^2 - 2(Y - X^T\beta')X^T(\beta-\beta')\right] \geq 0 \ .$$

- Then define as a criterion the diameter of this set :

$$C_K(\beta) = \sup_{\beta' \in \mathcal{B}_K(\beta)} \|\beta - \beta'\|^2_{L^2(P^X)} \ .$$

- Select finally the estimator minimizing this criterion :

$$\hat{\beta}_K \in \arg\min_{\beta \in \mathbb{R}^p} C_K(\beta) \ .$$

MAIN RESULT [LECUÉ, L., SAUMARD]

Learning from MOM's principles

G. Lecué and M. Lerasle

January 10, 2017

Abstract

We obtain estimation error rates and sharp oracle inequalities for a Birgé's T-estimator using a regularized median of mean principle as based tests. The results hold with exponentially large probability – the same one as in the gaussian framework with independent noise– under only weak moments assumption like a L_4/L_2 assumption and without assuming independence between the noise and the design X. The obtained rates are minimax optimal. The regularization norm we used can be any norm. When it has some sparsity inducing power we recover sparse rates of convergence and sparse oracle inequalities. As in [29], the size of the sub-differential of the regularization norm in a neighborhood of the oracle plays a central role in our analysis.

Moreover, the procedure allows for robust estimation in the sense that a large part of the data may have nothing to do with the oracle we want to reconstruct. The number of such irrelevant data (which can be seen as outliers) may be as large as *(sample size)×(rate of convergence)* as long as the quantity of useful data is larger than a proportion of the number of observations.

- Assume that $r^2 = \mathbb{E}[\xi^2 \|X\|^2] < \infty$. Let $C_\xi \geq \max_{\beta \in \mathbb{S}(0,1)} \mathrm{Var}(\xi X^T \beta)$, For any $K \gtrsim \frac{r^2}{C_\xi}$,

$$\mathbb{P}\left(\|\hat{\beta}_K - \beta^*\|_{L^2(P^X)} \gtrsim \sqrt{C_\xi \frac{K}{n}} \right) \leq 2e^{-K/C} \quad .$$

- One can apply Lepski's method to derive an estimator satisfying this bound simultaneously for all K, in particular :

$$\mathbb{P}\left(\|\tilde{\beta} - \beta^*\|_{L^2(P^X)} \gtrsim \sqrt{\frac{r^2}{n}} \right) \lesssim e^{-\frac{r^2}{CC_\xi}} \quad .$$

MAIN STEPS

- First remark that β^* beats any β such that

$$Q_{1/4,K}[(X^T(\beta-\beta^*))^2] - 2Q_{3/4,K}[\xi X^T(\beta-\beta^*)] \geq 0 \ .$$

- Then it beats any β such that $\|\beta-\beta^*\|_{L^2(P^X)} \geq r_K$, where

$$r_K = 2\frac{\sup_{\beta\in S(\beta^*,1)}(Q_{3/4,K}-P)[\xi X^T(\beta-\beta^*)]}{\inf_{\beta\in S(\beta^*,1)} Q_{1/4,K}[(X^T(\beta-\beta^*))^2]} \ .$$

- We deduce from this result that $\|\hat{\beta}_K-\beta^*\|_{L^2(P^X)} \leq r_K$.
- We use empirical process theory to bound r_K.

MAIN IDEAS

$$\|Q_{\alpha,K}[Z_t]\|_T \leq x \quad \Leftarrow \left\| \frac{1}{K} \sum_{k=1}^{K} I_{\{P_{B_k} Z_t > x\}} \right\|_T \leq 1-\alpha$$

$$\Leftarrow p_T + \left\| \frac{1}{K} \sum_{k=1}^{K} I_{\{P_{B_k} Z_t > x\}} - p_t \right\|_T \leq 1-\alpha$$

Bounded difference inequality

$$\Leftarrow p_T + \mathbb{E} \left\| \frac{1}{K} \sum_{k=1}^{K} I_{\{P_{B_k} Z_t > x\}} - p_t \right\|_T + u \leq 1-\alpha$$

symetrization + contraction principle

$$\Leftarrow p_T + \frac{4}{x} \mathbb{E} \left\| \frac{1}{n} \sum_{i=1}^{n} \epsilon_i Z_{t,i} \right\|_T + u \leq 1-\alpha$$

with probability $1 - e^{-\frac{u^2}{2}K}$

BOUNDING QUANTILE/MEAN PROCESSES

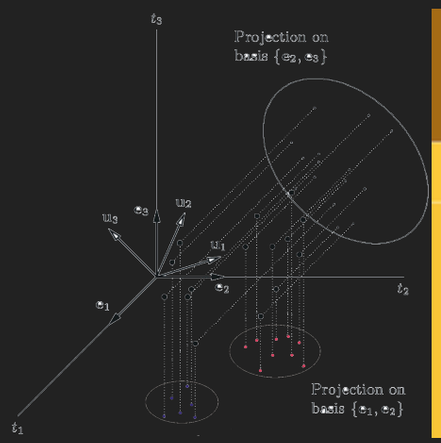
$$\text{if} \quad x \gtrsim \sqrt{\frac{K}{n}\|\mathbb{E}[Z_{t,i}^2]\|_T} \vee \mathbb{E}\left\|\frac{1}{n}\sum_{i=1}^n \epsilon_i Z_{t,i}\right\|_T$$

$$\mathbb{P}\left(\|Q_{\alpha,K}[Z_t]\|_T > x\right) \leq e^{-K/C} \quad .$$

ON THE MARGIN CONDITION

- C_ξ can be bounded by $2\left(|\mathrm{Var}(Y|X)|_\infty + \left|\mathbb{E}[Y|X] - X^T\beta^*\right|_\infty^2\right)$ provided that these quantities are finite.
- This example covers the case where the linear model is correct and the variance of the noise is bounded.
- More generally, if $\mathrm{Var}(Y|X)$ and $\mathbb{E}[Y|X] - X^T\beta^*$ have finite moment of order $2+\alpha$ and $\Psi = \sup_{\beta \in B_2(\beta^*,1), x \in \mathrm{supp}(P^X)} |x^T(\beta - \beta^*)|$, then $C_\xi \lesssim \Psi^{\frac{1}{1+\alpha/2}}$.
- In particular, for Fourier basis, Wavelet basis, histograms

$$\Psi \lesssim \sqrt{p} \qquad \text{thus} \qquad C_\xi \lesssim p^{\frac{1}{2+\alpha}} \quad .$$



EXTENSION TO HIGH DIMENSION

ADDING PENALTIES TO TESTS

- As usual, we have to add a penalty to the tests to deal with the large dimension setting. One can use for example, the ℓ_1 penalty and define the penalized test statistics

$$\mathrm{MOM}_K[(X^T(\beta-\beta'))^2 - 2(Y - X^T\beta')X^T(\beta-\beta')] + \lambda(|\beta|_1 - |\beta'|_1)$$

- Birge's aggregation procedure has to be slightly extended to benefit from the penalty.

$$C_K^{(1)}(\beta) = \sup_{\beta' \in \mathcal{B}_K(\beta)} \{|\beta - \beta'|_1\} \qquad C_K^{(2)}(\beta) = \sup_{\beta' \in \mathcal{B}_K(\beta)} \{|\beta - \beta'|_{L^2(P^X)}\} \ .$$

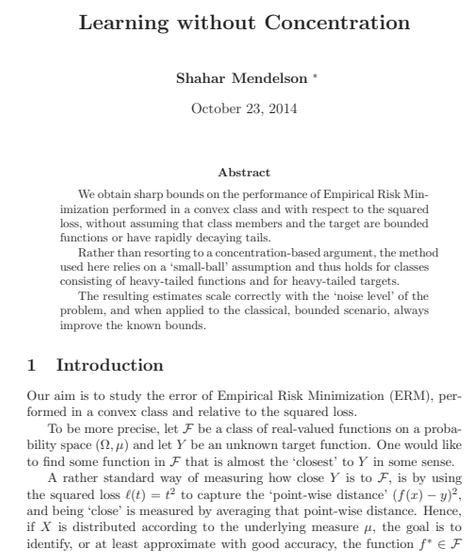
$$\mathcal{C}_K(\beta) = \min\{\rho \geq 0 : C_K^{(1)}(\beta) \leq \rho, \qquad C_K^{(2)}(\beta) \leq r(\rho)\} \ .$$

IDEAS UNDERLYING THE PROOF

- Using the approach of Mendelson, we reduce the problem to the control of a localized MOM_K process.

- The concentration of this terms reduces to the study of $\mathbb{E}\left[\sup_{\beta\in B_2(\beta^*,r)\cap B_1(\beta_*,\rho)} \frac{1}{n}\sum_{i=1}^{n}\epsilon_i\xi_i X^T(\beta-\beta^*)\right]$.

- This last term can be bounded at least in the linear regression framework.



Learning without Concentration

Shahar Mendelson *

October 23, 2014

Abstract

We obtain sharp bounds on the performance of Empirical Risk Minimization performed in a convex class and with respect to the squared loss, without assuming that class members and the target are bounded functions or have rapidly decaying tails.

Rather than resorting to a concentration-based argument, the method used here relies on a 'small-ball' assumption and thus holds for classes consisting of heavy-tailed functions and for heavy-tailed targets.

The resulting estimates scale correctly with the 'noise level' of the problem, and when applied to the classical, bounded scenario, always improve the known bounds.

1 Introduction

Our aim is to study the error of Empirical Risk Minimization (ERM), performed in a convex class and relative to the squared loss.

To be more precise, let $\mathcal{F}$ be a class of real-valued functions on a probability space (Ω,μ) and let Y be an unknown target function. One would like to find some function in $\mathcal{F}$ that is almost the 'closest' to Y in some sense.

A rather standard way of measuring how close Y is to $\mathcal{F}$, is by using the squared loss $\ell(t)=t^2$ to capture the 'point-wise distance' $(f(x)-y)^2$, and being 'close' is measured by averaging that point-wise distance. Hence, if X is distributed according to the underlying measure μ, the goal is to identify, or at least approximate with good accuracy, the function $f^*\in\mathcal{F}$

MAIN RESULTS (LECUÉ, L. AND LUGOSI,MENDELSON, 2017)

- Assume that $\xi \in L_{2+\epsilon}$ and let $\sigma \geq \|\xi\|_{2+\epsilon}$, suppose also

$$\forall \beta \in \mathbb{R}^p, \qquad \|\beta\|^2_{L^2(P^X)} = \|\beta\|^2,$$

design isotropic

$$\forall k \in [C_0 \log p], \forall i \in [p], \qquad \|X_i\|_{L_k(P^X)} \leq L\sqrt{k},$$

subGaussian moments

- Then, with probability larger than $1 - 3e^{-K/C}$,

$$\forall K \geq \sigma^2 s \log \frac{ep}{s} \ , \qquad |\hat{\beta}_K - \beta^*|_1 \leq \frac{K}{\sigma}\sqrt{\frac{1}{N}\log^{-1}\left(\frac{\sigma^2 p}{K}\right)} \ ,$$

$$\|\hat{\beta}_K - \beta^*\|_{L_2(P^X)} \lesssim_{C_\xi} \sqrt{\frac{K}{n}} \ .$$

COMPARISON/DISCUSSION

- Compared to [Lugosi and Mendelson 2017], Lepski's approach allows to remove the dependency of the estimator in an upper bound of $r(\rho_{K^*})$.
- Regarding « robustness » properties, we prove that the previous result is not affected by the presence of K_o outliers, provided that $K_o \leq \delta r^2(\rho_{K^*})$. (see also [Baraud, Birgé 2016]).
- The « informative data » may not be i.i.d., the procedure just requires close L^1, L^2 moments.

RHO-ESTIMATORS REVISITED: GENERAL THEORY AND APPLICATIONS

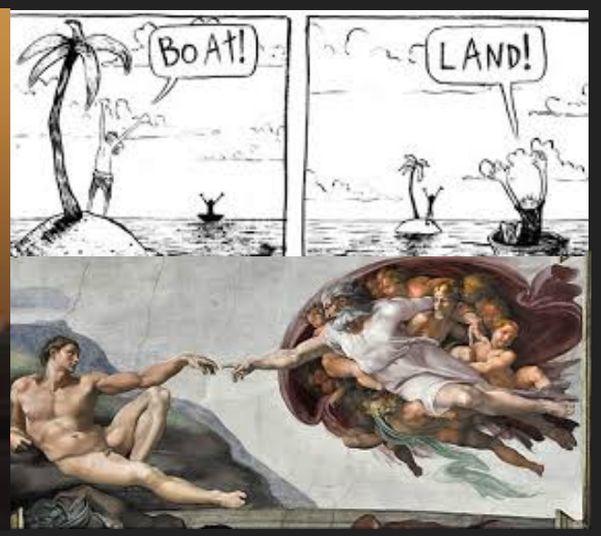
Y. BARAUD AND L. BIRGÉ

Abstract. Following Baraud, Birgé and Sart (2016), we pursue our attempt to design a universal and robust estimation method based on independent (but not necessarily i.i.d.) observations. Given such observations with an unknown joint distribution $\mathbf{P}$ and a dominated model $\mathscr{Q}$ for $\mathbf{P}$, we build an estimator $\widehat{\mathbf{P}}$ based on $\mathscr{Q}$ and measure its risk by an Hellinger-type distance. When $\mathbf{P}$ does belong to the model, this risk is bounded by some new notion of dimension which relies on the local complexity of the model in a vicinity of $\mathbf{P}$. In most situations this bound corresponds to the minimax risk over the model (up to a possible logarithmic factor). When $\mathbf{P}$ does not belong to the model, its risk involves an additional bias term proportional to the distance between $\mathbf{P}$ and $\mathscr{Q}$, whatever the true distribution $\mathbf{P}$. From this point of view, this new version of ρ-estimators improves upon the previous one described in Baraud, Birgé and Sart (2016) which required that $\mathbf{P}$ be absolutely continuous with respect to some known reference measure. Further additional improvements have been brought as compared to the former construction. In particular, it provides a very general treatment of the regression framework with random design as well as a computationally tractable procedure for aggregating estimators. Finally, we consider the situation where the Statistician has at disposal many different models and we build a penalized version of the ρ-estimator for model selection and adaptation purposes. In the regression setting, this penalized estimator not only allows to estimate the regression function but also the distribution of the errors.

1. Introduction

In a previous paper, namely Baraud, Birgé and Sart (2016), we introduced a new class of estimators that we called ρ-estimators for estimating the distribution $\mathbf{P}$ of a random variable $\boldsymbol{X} = (X_1, \ldots, X_n)$ with values in some measurable space $(\mathscr{X}, \mathscr{B})$ under the assumption that the X_i are independent but not necessarily i.i.d. These estimators are based on density models, a *density model* being a family of densities $\mathbf{t}$ with respect to some reference measure $\boldsymbol{\mu}$ on $\mathscr{X}$. We also assumed that $\mathbf{P}$ was absolutely continuous with respect to $\boldsymbol{\mu}$ with density $\mathbf{s}$ and we measured the performance of an estimator $\widehat{\mathbf{s}}$ in terms of $\mathbf{h}^2(\mathbf{s}, \widehat{\mathbf{s}})$, where $\mathbf{h}$ is a Hellinger-type distance to be defined later. Originally, the motivations for this construction were to design an estimator $\widehat{\mathbf{s}}$ of $\mathbf{s}$ with the following properties.

— Given a density model $\mathbf{S}$, the estimator $\widehat{\mathbf{s}}$ should be nearly optimal over $\mathbf{S}$ from the minimax point of view, which means that it is possible to bound the risk of the estimator $\widehat{\mathbf{s}}$ over $\mathbf{S}$ from above by some quantity $CD(\mathbf{S})$ which is approximately of the order of the



SOME PERSPECTIVES

FURTHER DEVELOPMENTS

- Efficient algorithm in learning problems : using a ρ-aggregation procedure, reduction to a saddle-point detection.
- Remove the « small-ball » assumption to allow more general designs : no need to lower bound the quadratic process for all data.
- More general learning problems (density estimation, non-quadratic losses, …)

THANKS FOR YOUR ATTENTION!!

THANKS FOR YOUR ATTENTION.

R. I. Oliveira

G. Lugosi

L. Devroye

G. Lecué

S. Mendelson

A. Saumard