Oberwolfach March 1996



Saint-Flour July 1997



Aussois June 2006



Leiden July 2006



Leiden July 2006



The search for a "universal" estimator

Lucien Birgé

Laboratoire de Probabilités, Statistique et Modélisation
Sorbonne Université – Paris

STATISTICS CONFERENCE IN HONOR OF AAD VAN DER VAART'S 60TH BIRTHDAY

June 20, 2019 - Leiden

Some historical considerations I

The search for some form of "universal estimator" is not new and, as far as I know, goes back to Fisher and his work on the MLE (maximum likelihood estimator). Since then, this estimator has been analyzed by thousands of authors and used in many situations. It has become a standard method although it is far from being "universal" as showed by numerous counterexamples, unfortunately not so well-known. In particular, it may not exist, or not converge to the true unknown parameter and it is definitely not robust and may be strongly affected by the presence of a small number of outliers.

Well aware of these difficulties, Le Cam (1973 - 1975) designed a new estimator based on tests between Hellinger balls which solved part of these problems and, during the period 1977 - 1984, I worked at modifying, improving, and extending this estimator. Under some (not too restrictive) conditions this estimator could be proved to have some optimality and robustness properties. I then left the subject.

Lucien Le Cam – Oberwolfach December 1983



Some historical considerations II

It is on the occasion of the statistical semester organized in Paris in 2001, at Institut Henri Poincaré, to which some of the colleagues who are present here today took an active part, that I was pushed by the questions of some of them to go back to the subject. This resulted in the construction of the T-estimators (2006), a penalized and adaptive version of the estimators I worked on in the 70's. The "T" stands for "test" since all the estimators that I mentioned are based on tests, as is the MLE.

T-estimators were definitely closer to being universal than the MLE but with some quite serious limitations for their use in regression (mainly boundedness of the parameter space).

A new idea

I have been involved in various collaborations with Yannick Baraud for more than 15 years and the idea of an alternative to T-estimators came from him and is based on a new construction of robust tests between two probabilities that appears in his 2010 paper. Given n i.i.d. observations with unknown distribution $P^\star$ and a known distribution P it is unfortunately not possible to correctly estimate the Hellinger (or $\mathbb{L}_1$) distance between P and $P^\star$ but Baraud showed that if one has two probabilities P and Q, it is possible to get a reasonable estimator of $h^2(P, P^\star) - h^2(Q, P^\star)$ and therefore to decide which of P or Q is closer to $P^\star$, at least when the difference between these two distances is large enough.

That was the first step towards the construction of ρ-estimators that I shall now sketch. The ρ-estimators are not quite "universal" but probably more than the previous ones. All that I shall now explain is joint work with Yannick and can be found in our recent publications (to be found in the Bibliography).

Yannick Baraud – IHES June 2018



Back to the MLE

Although not typically presented this way, the MLE can be viewed as the result of many simultaneous tests. To make things simple, let us work within the following simple framework. One observes $\boldsymbol{X} = (X_1, \ldots, X_n)$ an n-sample with distribution $P_s = s \cdot \mu$ and density s with respect to μ. The model for s is a subset S of all probability densities with respect to μ and the log-likelihood $L_{\boldsymbol{X}}$ is the function on S given by

$$L_{\boldsymbol{X}}(t) = \sum_{i=1}^{n} \log\big(t(X_i)\big).$$

When we want to perform a test between u and v, we can look at the sign of the log-likelihood ratio

$$\mathbf{L}(\boldsymbol{X}, u, v) = \sum_{i=1}^{n} \log\left(\frac{v(X_i)}{u(X_i)}\right) = L_{\boldsymbol{X}}(v) - L_{\boldsymbol{X}}(u),$$

and we know that this is a good test if the true density is either u or v and P_u and P_v are far enough from one another with respect to the Hellinger distance h. If we perform all the tests of this form for all pairs $(u, v) \in S^2$ we see that there is a single winner of all tests which is the MLE: $\widehat{t}_n = \operatorname{argmax}_{t \in S} L_{\boldsymbol{X}}(t)$. Note that maximizing $L_{\boldsymbol{X}}$ amounts to minimizing with respect to t

$$\left[\sup_{u \in S} L_{\boldsymbol{X}}(u)\right] - L_{\boldsymbol{X}}(t) = \sup_{u \in S} \left[L_{\boldsymbol{X}}(u) - L_{\boldsymbol{X}}(t)\right] = \sup_{u \in S} \mathbf{L}(\boldsymbol{X}, t, u) = \sup_{u \in S} \sum_{i=1}^{n} \log\left(\frac{u(X_i)}{t(X_i)}\right).$$

An alternative to the MLE - 1

This finally leads to

$$\widehat{t}_n = \operatorname*{argmin}_{t\in S} \sup_{u\in S} \sum_{i=1}^{n} \log\left(\frac{u(X_i)}{t(X_i)}\right).$$

Once one has recognized that many difficulties connected with the MLE are due to the unboundedeness of the log function, a natural idea is to try to replace, in the last formula, the log function by a bounded function φ such that the sign of

$$\sum_{i=1}^{n} \varphi\left(\frac{u(X_i)}{t(X_i)}\right)$$

provides a reasonable alternative to the likelihood ratio test given by the sign of $\mathbf{L}(\boldsymbol{X}, u, t)$. It happens that such a function does exist, namely

$$\varphi(x) = 4\frac{\sqrt{x}-1}{\sqrt{x}+1} \quad \text{for all } x \geqslant 0,$$

a function which, like the log function, is increasing, concave and satisfies $\varphi(1/x) = -\varphi(x)$. In fact, these two functions (φ and log) coincide at $x = 1$, their first and second derivatives as well and

$$0.99 < \frac{\varphi(x)}{\log x} \leqslant 1 \quad \text{and} \quad |\varphi(x) - \log x| \leqslant 0.055|x-1|^3 \quad \text{for all } x \in [1/2, 2].$$

An alternative to the MLE - 2

Replacing the log by φ we see that the analogue of

$$\widehat{t}_n = \operatorname*{argmin}_{t\in S} \sup_{u\in S} \sum_{i=1}^{n} \log\left(\frac{u(X_i)}{t(X_i)}\right) = \operatorname*{argmin}_{t\in S} \sup_{u\in S} \sum_{i=1}^{n} \frac{1}{4}\log\left(\frac{u(X_i)}{t(X_i)}\right)$$

can be written as

$$\widetilde{t}_n = \operatorname*{argmin}_{t\in S} \sup_{u\in S} \sum_{i=1}^{n} \frac{1}{4}\varphi\left(\frac{u(X_i)}{t(X_i)}\right) = \operatorname*{argmin}_{t\in S} \sup_{u\in S} \mathbf{T}(\boldsymbol{X}, u, t),$$

with

$$\mathbf{T}(\boldsymbol{X}, u, t) = \sum_{i=1}^{n} \psi\left(\sqrt{\frac{u(X_i)}{t(X_i)}}\right) \quad \text{and} \quad \psi(x) = \frac{x-1}{x+1}.$$

Note that $\widetilde{t}_n$ does not coincide with $\operatorname{argmax}_{t\in S} \sum_{i=1}^{n} \psi\left(t(X_i)\right)$ since, when replacing the log by ψ, we have lost the magic property of the logarithm which says that $\log(a/b) = \log a - \log b$,

Why does this work?

The sign of $\mathbf{T}(\boldsymbol{X}, u, t)$ actually provides a good test between t and u because it is a good estimator of $\rho(u,s) - \rho(t,s) = h^2(t,s) - h^2(u,s)$, where s denotes the true density and $\rho = 1 - h^2$ and h are respectively the Hellinger affinity and distance (hence the name of ρ-estimators):

$$h^2(t,u) = \frac{1}{2}\int_{\mathscr{X}} \left(\sqrt{t} - \sqrt{u}\right)^2 d\mu \quad \text{and} \quad \rho(t,u) = \int_{\mathscr{X}} \sqrt{tu}\, d\mu.$$

Just pretend for a minute that

$$\mathbf{T}(\boldsymbol{X}, u, t) = h^2(t,s) - h^2(u,s) \tag{1}$$

exactly. Then

$$\operatorname*{argmin}_{t\in S} \sup_{u\in S} \mathbf{T}(\boldsymbol{X}, u, t) = \operatorname*{argmin}_{t\in S} \sup_{u\in S} \left[h^2(t,s) - h^2(u,s)\right] = \operatorname*{argmin}_{t\in S} h(t,s)$$

and if $s \in S$, one gets $\widetilde{t}_n = s$; otherwise $\widetilde{t}_n$ is a point in S which is closest to s. Of course this is not true since (1) does not hold but one can derive precise results about the performance of $\widetilde{t}_n$ from a study of the process $\mathbf{T}(\boldsymbol{X}, u, t) - h^2(t,s) + h^2(u,s)$ which can actually be done even if $s \notin S$, resulting in the robustness properties of the ρ-estimators.

A basic result about ρ-estimators

Theorem

Given the dominated statistical model $\mathscr{P} = \{P_t = t \cdot \mu, , t \in S\}$ *and whatever the true density* s *with respect to* μ, *the* ρ-estimator $\widetilde{t}_n$ *satisfies the following error bound whatever* $\xi > 0$*:*

$$\mathbb{P}_s \left[h^2(s, \widetilde{t}_n) \leqslant C \left(h^2(s, S) + \frac{D_n(\mathscr{P})}{n} + \frac{\xi}{n} \right) \right] \geqslant 1 - e^{-\xi},$$

where $D_n(\mathscr{P})$ *stands for the complexity of the model, defined in a suitable way.*

The term $h^2(s, S)$ accounts for the robustness properties of the estimator. As to the complexity $D_n(\mathscr{P})$, it can typically be bounded via metric- or VC-dimension arguments. In all the cases we analyzed, this bound results in optimal rates of convergence.

Extensions

The previous theorem is a very simplified version of our general result and the method does work for more general situations.

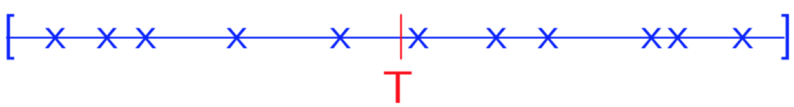
- Our results actually deal with independent but not necessarily i.i.d. observations in order to apply to regression frameworks.
- The model $\mathscr{P}$ has to be dominated by μ but the true distribution can be anything and is not necessarily absolutely continuous with respect to μ.
- As is the case for T-estimators, one can actually work with many models simultaneously, by adding a penalty term to the criterion we minimize. This leads to adaptive estimators and also allows to deal with fixed design regression without precisely knowing the distribution of the errors.
- The way the estimator is built also allows to perform estimation in the regression framework with random design without knowing the distribution of the design.

- In favourable situations (nice parametric models), if the model is correct, the ρ-estimator asymptotically coincides with the maximum likelihood estimator.
- One can build a Bayesian version of the ρ-estimators with the advantage that it is robust with respect to Hellinger deviations for the same reasons as before.

Main drawback: the price to pay for all these nice theoretical properties is the high level of complexity required to effectively compute the estimator in most situations.

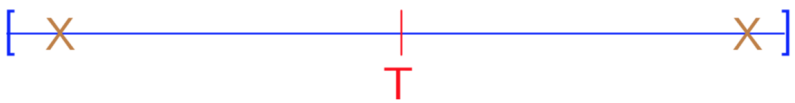
A very simple illustrative example

We have at hand n independent observations which are (theoretically) uniformly distributed in an interval $[T-1, T+1]$ with an unknown parameter $T \in \mathbb{R}$ but, since our observations are not quite reliable, there may be some outliers. We want nevertheless to design an accurate estimator of T from the observations. What accuracy can we expect of an estimator of T?



Mean or median of the data

Accuracy of order $1/\sqrt{n}$ plus robustness in the case of the median



Mean of the extreme points

It is one of the maximum likelihood estimators. Accuracy of order $1/n$

The second strategy reaches the best possible accuracy. However...

Instability/Stability

What if the set of data contains one remote outlier?



the average of the extremal points leads to a poor estimation of T. This strategy is unstable/not robust to the presence of a single outlier. The mean suffers from the same defect. Not the median but, in the absence of outliers it leads to an error of order $1/\sqrt{n}$ rather than n.

Can we get both optimality and robustness?

Yes: a ρ-estimator $\widehat{T}$ of T is such that the interval $\left[\widehat{T}-1, \widehat{T}+1\right]$ contains most of the data. As a consequence, it is robust and will still have a risk of order $1/n$ if the number of outliers is small enough. Note that when the model is correct (no outlier) it is a maximum likelihood.



Bibliography

Baraud, Y. (2011). Estimator selection with respect to Hellinger-type risks. *Probab. Theory Related Fields*, 151(1-2):353–401.

Baraud, Y. and Birgé, L. (2016). Rho-estimators for shape restricted density estimation. *Stochastic Process. Appl.*, 126(12):3888–3912.

Baraud, Y. and Birgé, L. (2017). Robust Bayes-like estimation: Rho-Bayes estimation. Technical report, arXiv:1711.08328v1.

Baraud, Y. and Birgé, L. (2018). Rho-estimators revisited: General theory and applications. *Ann. Statist.*, 46(6B):3767–3804.

Baraud, Y., Birgé, L., and Sart, M. (2017). A new method for estimation and model selection: ρ-estimation. *Invent. Math.*, 207(2):425–517.

Barron, A., Birgé, L., and Massart, P. (1999). Risk bounds for model selection via penalization. *Probab. Theory Related Fields*, 113(3):301–413.

Birgé, L. (1983). Approximation dans les espaces métriques et théorie de l'estimation. *Z. Wahrsch. Verw. Gebiete*, 65(2):181–237.

Birgé, L. (2006). Model selection via testing: an alternative to (penalized) maximum likelihood estimators. *Ann. Inst. H. Poincaré Probab. Statist.*, 42(3):273–325.

Le Cam, L. (1973). Convergence of estimates under dimensionality restrictions. *Ann. Statist.*, 1:38–53.

Le Cam, L. (1975). On local and global properties in the theory of asymptotic normality of experiments. In *Stochastic processes and related topics (Proc. Summer Res. Inst. Statist. Inference for Stochastic Processes, Indiana Univ., Bloomington, Ind., 1974, Vol. 1; dedicated to Jerzy Neyman)*, pages 13–54. Academic Press, New York.

Le Cam, L. (1990). Maximum likelihood: An introduction. *Inter. Statist. Review*, 58(2):153–171.

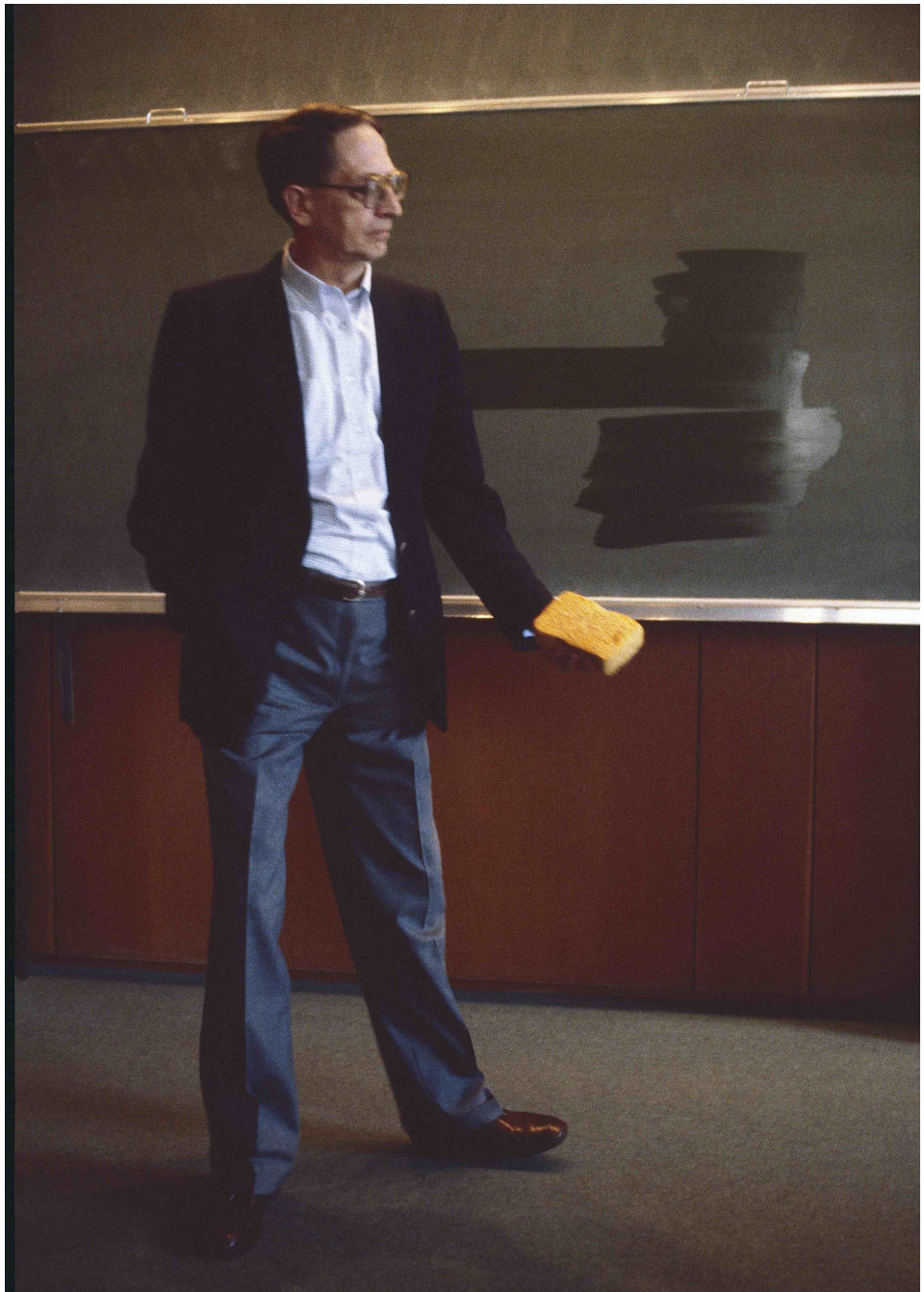
Cambridge June 2014



Cambridge June 2014



Oberwolfach March 1987



Oberwolfach March 1987



Saint-Flour July 1994



Saint-Flour July 1994



Saint-Flour July 1994



Orsay June 2001



Orsay June 2001



IHP Paris June 2001



Aussois June 2003



Aussois June 2003



Aussois June 2003



Leiden July 2006



Leiden July 2006



Leiden July 2006



CIRM Marseille November 2006



CIRM Marseille December 2012



Cambridge June 2014



Cambridge June 2014



Heidelberg July 2014

