Model selection theory: a tutorial with applications to learning

Pascal Massart
Université Paris-Sud, Orsay

ALT 2012, October 29

- Asymptotic approach to model selection

- Idea of using some penalized empirical criterion goes back to the seminal works of Akaike (’70).
- Akaike celebrated criterion (AIC) suggests to penalize the log-likelihood by the number of parameters of the parametric model.
- This criterion is based on some asymptotic approximation that essentially relies on Wilks’ Theorem

Wilks' Theorem: under some proper regularity conditions the log-likelihood $L_n(\theta)$ based on n i.i.d. observations with distribution belonging to a parametric model with D parameters obeys to the following weak convergence result

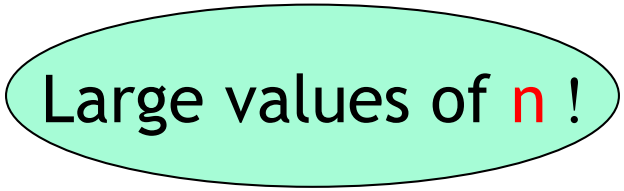
$$2\left(L_n\left(\hat{\theta}\right) - L_n\left(\theta_0\right)\right) \rightarrow \chi^2(D)$$

where $\hat{\theta}$ denotes the MLE and θ_0 is the true value of the parameter.

- Non asymptotic Theory

  In many situations, it is usefull to make the size of the models tend to infinity or make the list of models depend on n. In these situations, classical asymptotic analysis breaks down and one needs to introduce an alternative approach that we call non asymptotic.

  We still like

  But the size of the models as well the size of the list of models should be authorized to be large too.

Functional estimation

- The basic problem

Construct estimators of some function **s**, using as few prior information on **s** as possible. Some typical frameworks are the following.

- Density estimation

$(X_1,\ldots,X_n)$ i.i.d. sample with unknown density **s** with respect to some given measure μ .

- Regression framework

One observes $(X_1,Y_1),\ldots,(X_n,Y_n)$

With $Y_i = s(X_i)+\varepsilon_i$

The explanatory variables X_i are fixed or i.i.d.

The errors ε_i are i.i.d. with $E\left[\varepsilon_i \middle| X_i\right]=0$

- Binary classification

We consider an i.i.d. regression framework where the response variable Y is a « label » :0 or 1. A basic problem in *statistical learning* is to estimate the best classifier $s(x)=1_{\eta(x)\geq 1/2}$, where η denotes the regression function

$$\eta(x)=E\left[Y\middle|X=x\right]$$

- Gaussian white noise

Let s be a numerical function on $[0,1]$. One observes the process $Y^{(n)}$ on $[0,1]$ defined by

$$dY^{(n)}(x)=s(x)dx+\frac{1}{\sqrt{n}}dB(x), Y^{(n)}(0)=0$$

Where B is a Brownian motion. The level of noise is written as $1/\sqrt{n}$ by allow an easy comparison.

Empirical Risk Minimization (ERM)

A classical strategy to estimate s consists of taking a set of functions S (a « model ») and consider some *empirical criterion* (based on the data) γ_n such that

$$t \to E\left[\gamma_n(t)\right]$$

achieves a minimum at point $t = s$. The ERM estimator $\hat{s}$ of s minimizes γ_n over S. One can hope that $\hat{s}$ is close to s , if the target s belongs to model S (or at least is not far from S). This approach is most popular in the parametric case (i.e. when S is defined by a finite number of parameters and one assumes that $s \in S$).

- Maximum likelihood estimation (MLE)

Context:density estimation (i.i.d. setting to be simple) $(X_1,\ldots,X_n)$ i.i.d. sample with distribution $sd\,\mu$

with

$$\gamma_n(t) = -\frac{1}{n}\sum_{i=1}^{n}\log t(X_i)$$

$$E\left[\gamma_n(t)-\gamma_n(s)\right] = K(s,t) \geq 0$$

Kullback Leibler information

- Least squares

Regression

with

$$\gamma_n(t) = \frac{1}{n}\sum_{i=1}^{n}\left(Y_i - t(X_i)\right)^2$$

White noise

$$E\left[\gamma_n(t) - \gamma_n(s)\right] = \frac{1}{n}\sum_{i=1}^{n} E\left[(t-s)^2(X_i)\right] \geq 0$$

with

$$\gamma_n(t) = \|t\|^2 - 2\int t(x)\, dY^{(n)}(x)$$

Density

$$E\left[\gamma_n(t) - \gamma_n(s)\right] = \|t-s\|^2 \geq 0$$

with

$$\gamma_n(t) = \|t\|^2 - \frac{2}{n}\sum_{i=1}^{n} t(X_i)$$

$$E\left[\gamma_n(t) - \gamma_n(s)\right] = \|t-s\|^2 \geq 0$$

Exact calculations in the linear case

In the white noise or the density frameworks, when S is a finite dimensional subspace of $L_2(\mu)$ (where μ denotes the Lebesgue measure in the white noise case), the LSE can be explicitly computed. Let $(\phi_\lambda)_{\lambda\in\Lambda}$ be some orthonormal basis of S, then

$$\hat{s} = \sum_{\lambda\in\Lambda} \hat{\beta}_\lambda \phi_\lambda$$

$$\hat{\beta}_\lambda = \int \phi_\lambda(x)\, dY^{(n)}(x) \quad \text{or} \quad \hat{\beta}_\lambda = \frac{1}{n}\sum_{i=1}^{n} \phi_\lambda(X_i)$$

White noise

Density

- The model choice paradigm
- If a model S is defined by a « small » number of parameters (as compared to n), then the target s can happen to be far from the model.
- If the number of parameters is taken too large then $\hat{s}$ will be a poor estimator of s *even* if s truly belongs to S.

  Illustration (white noise)

  One takes S as a linear space with dimension D, the expected *quadratic risk* of the LSE can be easily computed

$$E\left\|\hat{s}-s\right\|^2 = d^2\left(s,S\right)+\frac{D}{n}$$

  Of course, since we do not know $d^2\left(s,S\right)$ the quadratic risk cannot be used as a model choice criterion but just as a benchmark.

- First Conclusions
- It is safer to play with several possible models rather than with *a single* one given in advance.
- The notion of expected risk allows to compare the candidates and can serve as a benchmark.
- According to the risk minimization criterion, S is a « good » model *does not mean* that the target s belongs to S.
- Since the minimization of the risk cannot be used as a selection criterion, one needs to introduce some empirical version of it.

Model selection via penalization

Consider some empirical criterion γ_n .

- Framework: Consider some (at most countable) collection of models $(S_m)_{m\in\mathfrak{M}}$. Represent each model S_m by the ERM $\hat{s}_m$ on S_m .
- Purpose: select the « best » estimator among the collection $(\hat{s}_m)_{m\in\mathfrak{M}}$.
- Procedure: Given some penalty function $\mathrm{pen}:\mathfrak{M}\to\mathbb{R}_+$, we take $\hat{m}$ minimizing

$$\gamma_n(\hat{s}_m)+\mathrm{pen}(m)$$

over $\mathfrak{M}$ and define

$$\tilde{s}=\hat{s}_{\hat{m}}.$$

- The classical asymptotic approach

Origin: Akaike (log-likelihood), Mallows (least squares)

① The penalty function is proportional to the number of parameters D_m of the model S_m.

Akaike : D_m / n

Mallows' C_p : $2D_m / n$,

where the variance of the errors of the regression framework is assumed to be equal to 1 by the sake of simplicity.

② The heuristics (Akaike ('73)) leading to the choice of the penalty function D_m / n relies on the assumption: the dimensions and the number of the models are bounded w.r.t. n and n tends to infinity.

BIC (log-likelihood) criterion Schwartz (‘78) :

- aims at selecting a « true » model rather than mimicking an oracle

- also asymptotic, with a penalty which is proportional to the number of parameters:

$$\ln(n) D_m / n$$

- The non asymptotic approach

Barron,Cover (’91) for discrete models, Birgé, Massart (‘97) and Barron, Birgé, Massart (’99)) for general models. Differs from the asymptotic approach on the following points

- The number as well as the dimensions of the models may depend on n.
- One can choose a list of models because of its *approximation properties*:

  wavelet expansions, trigonometric or piecewise polynomials, artificial neural networks etc

It may perfectly happen that many models of the list have the same dimension and in our view, the « complexity » of the list of models is typically taken into account. Shape of the penalty

$$C_1 \frac{D_m}{n} + C_2 \frac{x_m}{n}$$

with $\sum_{m \in \mathfrak{M}} e^{-x_m} \leq \Sigma$.

Data driven penalization

Practical implementation requires some data-driven calibration of the penalty.

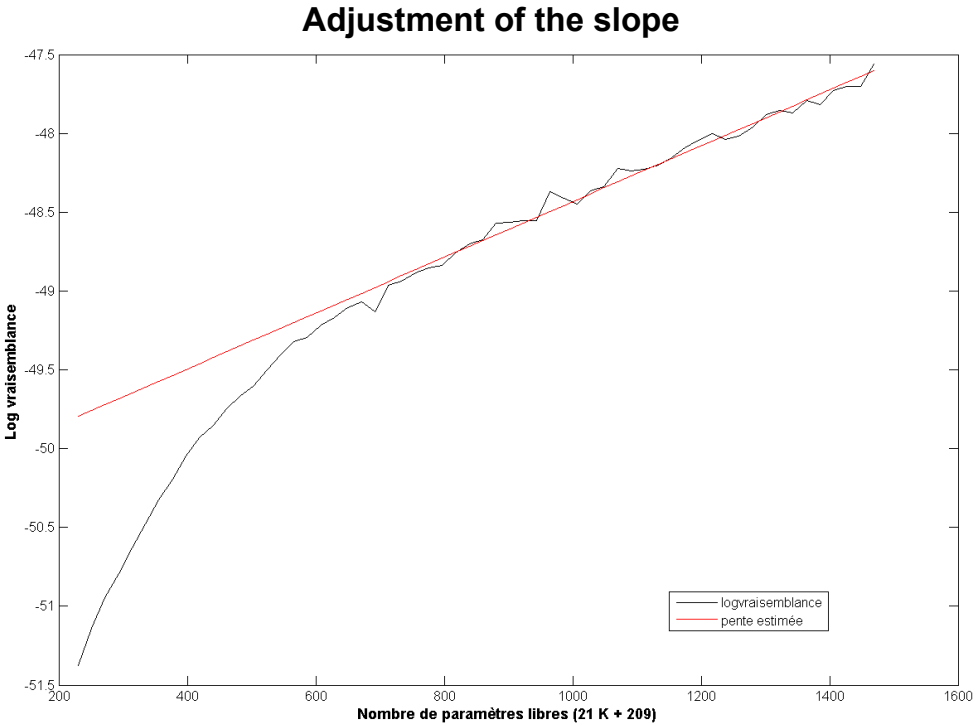
« Recipe »

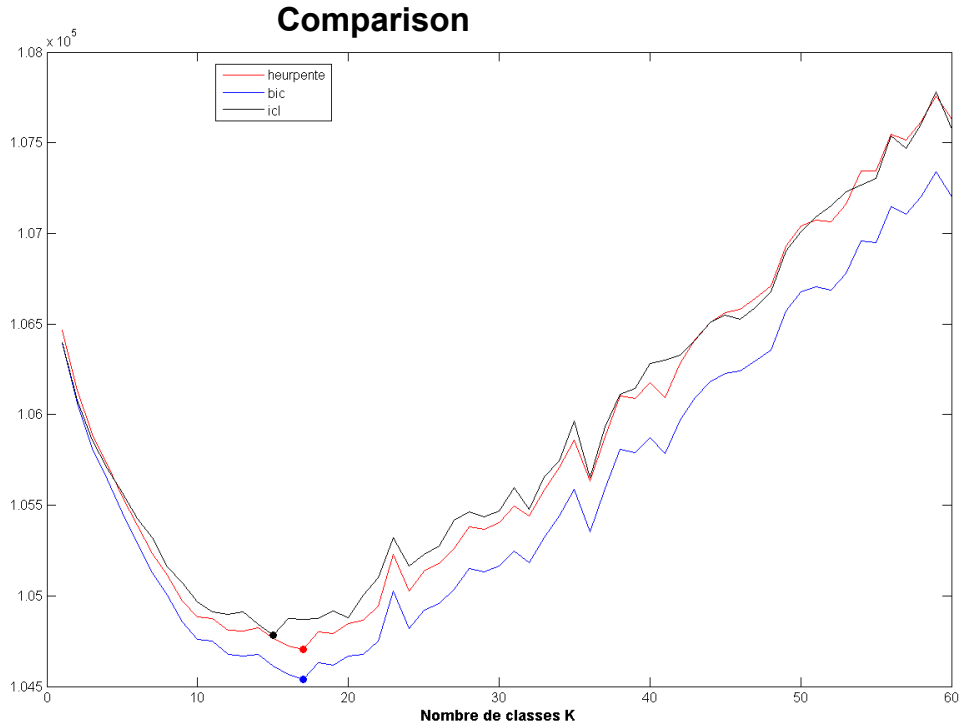
1. Compute the ERM $\hat{s}_D$ on the union of models with D parameters
2. Use theory to guess the shape of the penalty pen(D), typically pen(D)=aD (but aD(2+ln(n/D)) is another possibility)
3. Estimate a from the data by multiplying by 2 the smallest value for which the penalized criterion explodes.

Implemented first by Lebarbier ('05) for multiple change points detection

Celeux, Martin, Maugis '07

- Gene expression data: 1020 genes and 20 experiments
- Mixture models $$\mathcal{S}_K = \left\{ x \in \mathbb{R}^{20} \mapsto \sum_{k=1}^{K} p_k \Phi(x|\mu_k, \Sigma) \right\}$$
- Choice of K ? Slope heuristics: K=17 BIC: K=17 ICL: K=15





Akaike's heuristics revisited

The main issue is to remove the asymptotic approximation argument in Akaike's heuristics

$$\gamma_n\left(\hat{s}_D\right) = \gamma_n\left(s_D\right) - \underbrace{\left[\gamma_n\left(s_D\right) - \gamma_n\left(\hat{s}_D\right)\right]}_{\hat{v}_D}$$

variance term

minimizing $\gamma_n\left(\hat{s}_D\right) + \mathrm{pen}\left(D\right)$, is equivalent to minimizing

$$\underbrace{\gamma_n\left(s_D\right) - \gamma_n\left(s\right)}_{} - \hat{v}_D + \mathrm{pen}\left(D\right)$$

Fair estimate of $\ell(s, s_D)$

Ideally: $\text{pen}_{id}\left(D\right) = \hat{v}_D + \ell\left(s_D, \hat{s}_D\right)$

In order to (approximately) minimize

$$\ell(s, \hat{s}_D) = \ell(s, s_D) + \ell\left(s_D, \hat{s}_D\right)$$

The key : Evaluate the excess risks

$$\hat{v}_D = \gamma_n\left(s_D\right) - \gamma_n\left(\hat{s}_D\right)$$

$$\ell\left(s_D, \hat{s}_D\right)$$

This the very point where the various approaches diverge. Akaike's criterion relies on the asymptotic approximation

$$\ell(s_D, \hat{s}_D) \approx \hat{v}_D \approx \frac{D}{2n}$$

The method initiated in Birgé, Massart ('97) relies on upper bounds for the sum of the excess risks which can be written as

$$\hat{v}_D + \ell\left(s_D, \hat{s}_D\right) = \left[\bar{\gamma}_n\left(s_D\right) - \bar{\gamma}_n\left(\hat{s}_D\right)\right]$$

where $\bar{\gamma}_n$ denotes the empirical process

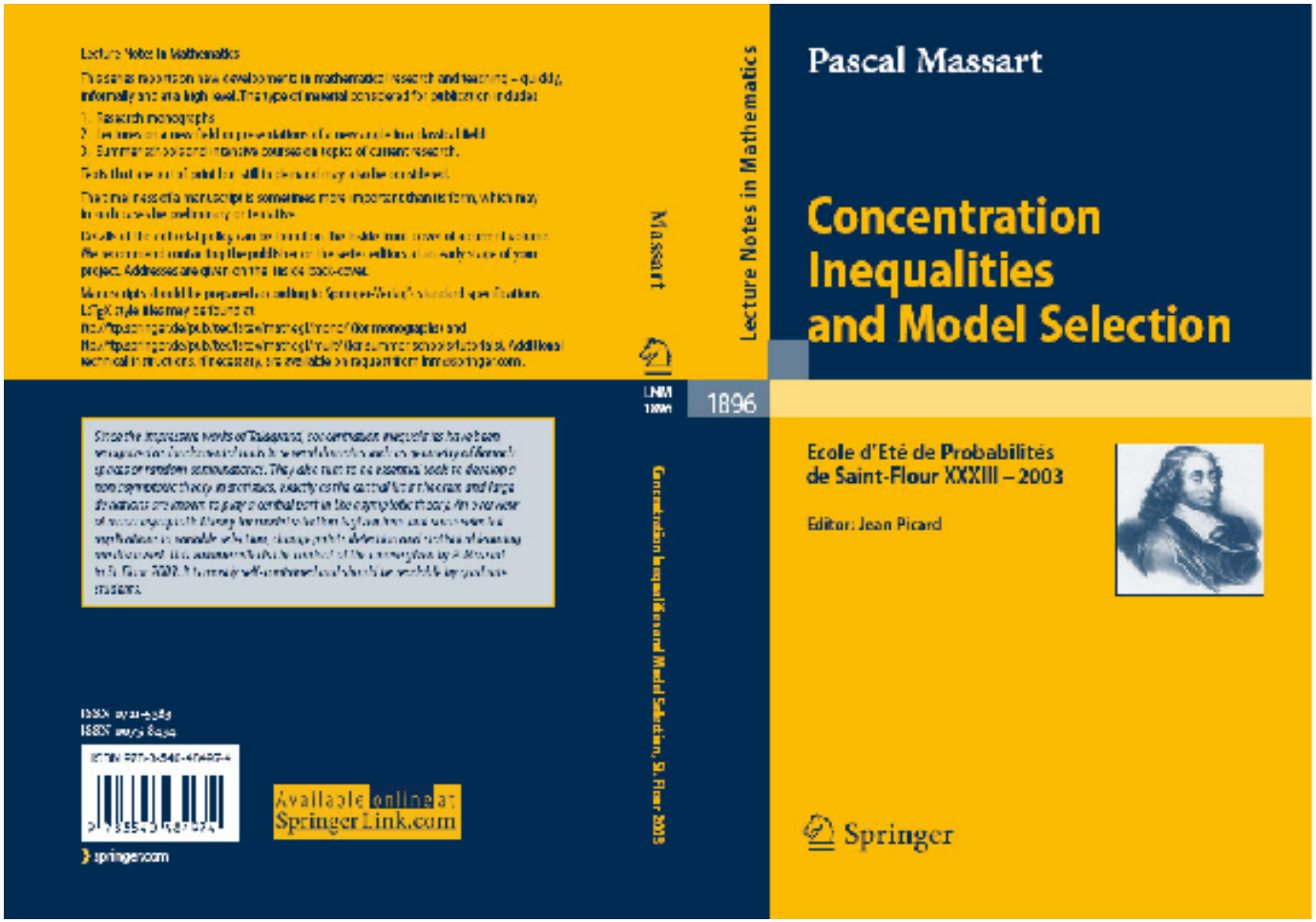
$$\bar{\gamma}_n(t) = \gamma_n(t) - E\left[\gamma_n(t)\right]$$

These bounds derive from concentration inequalities for the supremum of the appropriately weighted empirical process

$$\frac{\bar{\gamma}_n(t) - \bar{\gamma}_n(u)}{\omega(t,u)}, t \in S_D$$

The prototype being Talagrand's inequality ('96) for empirical processes.

This approach has been fruitfully used in several works. Among others: Baraud ('00) and ('03) for least squares in the regression framework, Castellan ('03) for log-splines density estimation, Patricia Reynaud ('03) for poisson processes, etc...



Main drawback: typically involve some unkown multiplicative constante which may depend on the unknown distribution (variance of the regression errors, supremum of the density, classification noise etc…).

Needs to be calibrated…

Slope heuristics : one looks for some approximation of $\hat{v}_D$ (typically) of the form aD with a unknown. When D is large, $\gamma_n\left(s_D\right)$ is almost constant, it suffices to « read » a as a slope on the graph of $\gamma_n\left(\hat{s}_D\right)$. On chooses the final penalty as

$$\mathrm{pen}\left(D\right) = 2 \times aD$$

In fact $\mathrm{pen}_{\min}\left(D\right)=\hat{v}_D$ is a minimal penalty and the slope heuristics provides a way of approximating it. The factor 2 which is finally used reflects our hope that the excess risks

$$\hat{v}_D=\gamma_n\left(s_D\right)-\gamma_n\left(\hat{s}_D\right)$$

$$\ell\left(s_D,\hat{s}_D\right)$$

are of the same order of magnitude. If this the case then

« optimal » penalty=2 * « minimal » penalty

Recent advances

- Justification of the slope heuristics:

  Arlot and Massart (JMLR'08) for histograms in the regression framework. Phd of Saumard (2010) regular parametric models.

  Boucheron and Massart (PTRF'11) for concentration of the empirical excess risk (Wilks phenomenon)

- Calibration of regularization

  Linear estimators Arlot and Bach (2010)

  Lasso type algorithms. Thesis: Connault (2010)

  and Meynet (work in progress…)

High dimensional Wilks' phenomenon

Wilks' Theorem: under some proper regularity conditions the log-likelihood $L_n(\theta)$ based on n i.i.d. observations with distribution belonging to a parametric model with D parameters obeys to the following weak convergence result

$$2\left(L_n\left(\hat{\theta}\right) - L_n\left(\theta_0\right)\right) \rightarrow \chi^2(D)$$

where $\hat{\theta}$ denotes the MLE and θ_0 is the true value of the parameter.

Question: what's left if we consider possibly irregular empirical risk minimization procedures and let the dimension of the model tend to infinity?

Obviously one cannot expect similar asymptotic results. However it is still possible to exhibit some kind of Wilks' phenomenon.

Motivation: modification of Akaike's heuristics for model selection

Data-driven penalties

- A statistical learning framework

We consider the i.i.d. framework where one observes independent copies $\xi_1,\dots,\xi_n$ of a random variable ξ with distribution P. We have in mind the *regression framework* for which $\xi=(X,Y)$. X is an explanatory variable and Y is the response variable.

Let s be some target function to be estimated. For instance, if η denotes the *regression function* $\eta(x)=E\left[Y\middle|X=x\right]$

The function of interest s may be the regression function η itself.

In the *binary classification* case where the response variable takes only the two values 0 and 1, it may be the Bayes classifier

$$s = 1\mathrm{I}_{\{\eta \geq 1/2\}}$$

We consider some criterion γ, such that the target function s achieves the minimum of

$$t \to P\gamma(t,.)$$

over some set $\mathcal{S}$. For example

$$\gamma(t,(X,Y)) = (Y - t(X))^2$$

- with $\mathcal{S} = L_2$ leads to the regression function as a minimizer
- with $\mathcal{S} = \{t : \mathcal{X} \to \{0,1\}\}$ leads to the Bayes classifier

Introducing the empirical criterion

$$\gamma_n(t) = P_n(\gamma(t,.)) = \frac{1}{n}\sum_{i=1}^{n}\gamma(t,\xi_i)$$

in order to estimate s one considers some subset S of $\mathcal{S}$ (a « model ») and defines the empirical risk minimizer $\hat{s}$ as a minimizer of γ_n over S.

This commonly used procedure includes LSE and also MLE for density estimation.

In this presentation we shall assume that $s \in S$ (makes life simpler but not necessary) and also that $0 \le \gamma \le 1$ (boundedness is necessary).

Introducing the natural loss function

$$\ell(s,t) = P\gamma(t,.) - P\gamma(s,.)$$

We are dealing with two « dual » estimation errors

- the *excess loss*

$$\ell(s,\hat{s}) = P\gamma(\hat{s},.) - P\gamma(s,.)$$

- the *empirical excess loss*

$$\ell_n(s,\hat{s}) = P_n\gamma(s,.) - P_n\gamma(\hat{s},.)$$

Note that for MLE $\gamma(t,.) = -\log t(.)$ and Wilks' theorem provides the asymptotic behavior of $\ell_n(s,\hat{s})$ when S is a regular parametric model.

Crucial issue:

Concentration of the empirical excess loss: connected to empirical processes theory because

$$\ell_n\left(s,\hat{s}\right)=\sup_{t\in S}P_n\left(\gamma\left(s,.\right)-\gamma\left(t,.\right)\right)$$

Difficult problem: Talagrand's inequality does not make directly the job (the $1/n$ rate is hard to gain).

Let us begin with the related but easier question:

What is the order of magnitude of the excess loss and the empirical excess loss?

Risk bounds for the excess loss

We need to relate the variance of $\gamma(t,.)-\gamma(s,.)$ with the excess loss

$$\ell(s,t)=P\left(\gamma(t,.)-\gamma(s,.)\right)$$

Introducing some pseudo-metric d such that

$$P\left(\gamma(t,.)-\gamma(s,.)\right)^2 \le d^2(s,t)$$

We assume that for some convenient function w

$$d(s,t)\le w\left(\sqrt{\ell(s,t)}\right)$$

In the regression or the classification case d is simply the L_2 distance and w is either identity for regression or is related to a *margin condition* for classification.

- Tsybakov's margin condition (AOS 2004)

$$\ell(s,t) \geq h d^{2\kappa}(s,t)$$

where $\kappa \geq 1$ and $h \in [0,1]$ with

$$d^2(s,t) = E\left[\left|s(X) - t(X)\right|\right]$$

Since for binary classification

$$\ell(s,t) = E\left[\left|2\eta(X) - 1\right| \left|s(X) - t(X)\right|\right]$$

this condition is closely related to the behavior of $\eta(X)$ around $1/2$. For example margin condition $\kappa = 1$ is achieved whenever

$$\left|2\eta - 1\right| \geq h$$

- Heuristics

Let us introduce

$$\overline{\gamma_n}(t) = (P_n - P)\gamma(t,.)$$

Then

$$\ell(s,\hat{s}) + \ell_n(s,\hat{s}) = \overline{\gamma_n}(s) - \overline{\gamma_n}(\hat{s}) \quad \Rightarrow \quad \ell(s,\hat{s}) \leq \overline{\gamma_n}(s) - \overline{\gamma_n}(\hat{s})$$

Now the variance of $\overline{\gamma_n}(s) - \overline{\gamma_n}(t)$ is bounded by $d^2(s,t)/n$ hence empirical process theory tells you that the uniform fluctuation of

$$\sqrt{n}\left[\overline{\gamma_n}(s) - \overline{\gamma_n}(t)\right]$$

remains under control in the ball $d(s,t) \leq \sigma$

(Massart, Nédélec AOS 2006)

Theorem : *Let* ϕ, $w \nearrow$ *such that* $\phi(x)/x, w(x)/x \searrow$ *with* $\phi(1) \geq 1$ *and* $w(1) \geq 1$. *Assume that*

$$d(s,t) \leq w\left(\sqrt{\ell(s,t)}\right)$$

and

$$\mathbb{E}\left[\sup_{t \in S, d(s,t) \leq \sigma} \sqrt{n}\left(\overline{\gamma_n}(s) - \overline{\gamma_n}(t)\right)\right] \leq \phi(\sigma)$$

for every σ *such that* $\sqrt{n}\sigma^2 \geq \phi(\sigma)$. *Then, defining* ε_* *as*

$$\sqrt{n}\varepsilon_*^2 = \phi \circ w(\varepsilon_*)$$

one has

$$\mathbb{E}\left[\ell\left(s, \hat{s}\right)\right] \leq C\varepsilon_*^2$$

where C *is an absolute constant.*

Application to classification

- Tsybakov's framework

Tsybakov's margin condition means that $w(\varepsilon) \approx \varepsilon^{1/\kappa}$
An entropy with bracketing condition implies that one can take $\phi(\sigma) \simeq \sigma^{1-\rho}$ and we recover Tsybakov's rate $\varepsilon_*^2 \simeq n^{-\kappa/(2\kappa+\rho-1)}$

- VC-classes under margin condition $|2\eta - 1| \geq h$ one has $w(\varepsilon) = \varepsilon / \sqrt{h}$. If S is a VC-class with VC-dimension D

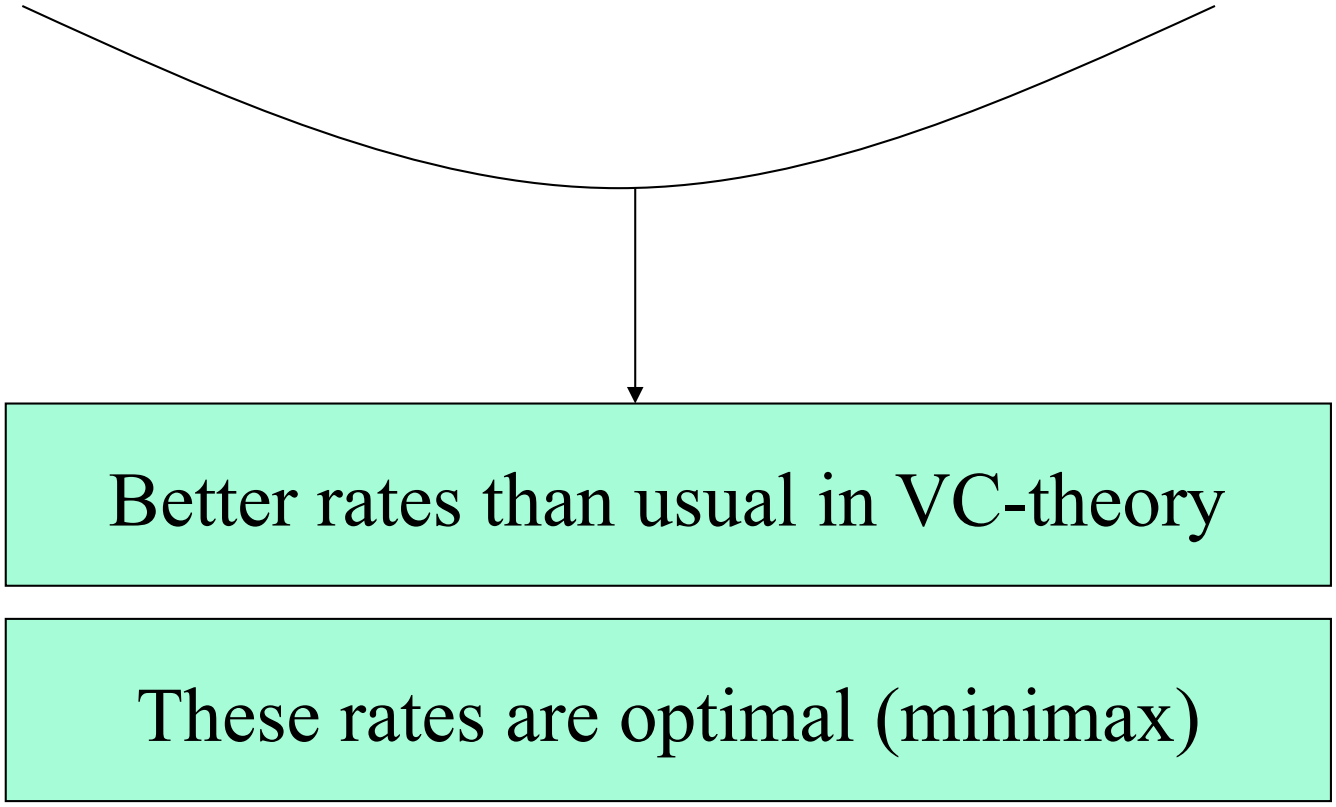
$$\phi(\sigma) \approx \sigma \sqrt{D(1 + \log(1/\sigma))}$$

so that (whenever $nh^2 \geq D$)

$$\varepsilon_*^2 = \frac{C}{nh} D\left(1 + \log\left(\frac{nh^2}{D}\right)\right)$$

Main points

- Local behavior of the empirical process
- Connection between ℓ and d .



Concentration of the empirical excess loss

Joint work with Boucheron (PTRF 2011). Since

$$\ell\left(s,\hat{s}\right)+\ell_n\left(s,\hat{s}\right)=\overline{\gamma_n}\left(s\right)-\overline{\gamma_n}\left(\hat{s}\right)$$

the proof of the above Theorem also leads to an upper bound for the empirical excess risk for the same price. In other words $\mathbb{E}\left[\overline{\gamma_n}\left(s\right)-\overline{\gamma_n}\left(\hat{s}\right)\right]$ is also bounded by ε_*^2 up to some absolute constant. Concentration: start from identity

$$\ell_n\left(s,\hat{s}\right)=\sup_{t\in S}P_n\left(\gamma\left(s,.\right)-\gamma\left(t,.\right)\right)$$

- Concentration tools

Let $\xi_1',\dots\xi_n'$ be some independent copy of $\xi_1,\dots\xi_n$. Defining $Z_i' = \zeta\left(\xi_1,\dots,\xi_{i-1},\xi_i',\xi_{i+1},\dots\xi_n\right)$ and setting

$$V^+ = \mathbb{E}\left[\sum_{i=1}^{n}\left(Z - Z_i'\right)_+^2 \middle| \xi\right]$$

Efron-Stein's inequality asserts that

$$\mathrm{Var}[Z] \le \mathbb{E}\left[V^+\right]$$

A Burkholder-type inequality (BBLM, AOP2005)

For every $q \ge 2$ such that $|Z|^q$ is integrable, one has

$$\left\|\left(Z - \mathbb{E}[Z]\right)_+\right\|_q \le \sqrt{3q\left\|V^+\right\|_{q/2}}$$

Comments

This inequality can be compared to Burkholder's martingale inequality

$$\|Z-\mathbb{E}[Z]\|_q \le (q-1)\sqrt{\|\langle Z\rangle\|_{q/2}}$$

where $\langle Z\rangle = \sum_{k=1}^{n}\left(\mathbb{E}\left[Z|\mathbf{F}_k\right]-\mathbb{E}\left[Z|\mathbf{F}_{k-1}\right]\right)^2$ denotes the quadratic variation w.r.t. Doob's filtration $\mathbf{F}_k=\sigma(X_1,\ldots,X_k)$, $1\le k\le n$ and $\mathbf{F}_0$ trivial σ-field.

It can also be compared with Marcinkiewicz Zygmund's inequality which asserts that in the case where $Z=X_1+\ldots+X_n$

$$\|Z-\mathbb{E}[Z]\|_q \le \sqrt{q\|\langle Z\rangle\|_{q/2}}$$

Note that the constant $q-1$ in Burkhölder's inequality cannot generally be improved. Our inequality is therefore somewhere « between » Burkholder's and Marcinkiewicz-Zygmund's inequalities. We always get the $\sqrt{q}$ factor instead of $q-1$ which turns out to be a crucial gain if one wants to derive sub-Gaussian inequalities. The price is to make the substitution

$$V^+ \Longleftrightarrow \langle Z \rangle$$

which is absolutely painless in the Marcinkiewicz-Zygmund case. More generally, for the examples that we have in view, V^+ will turn to be a quite manageable quantity and explicit moment inequalities will be obtained by applying iteratively the preceding one.

- Fundamental example

The supremum of an empirical process

$$Z = \sup_{t \in S} \sum_{i=1}^{n} f_t(\xi_i),$$

provides an important example, both for theory and applications. Assuming that $\sup_{t \in T} \|f_t\|_\infty \leq 1$, Talagrand's inequality (Inventiones 1996) ensures the existence of some absolute positive contants K and η , such that

$$\mathbb{P}\left[Z - \mathbb{E}[Z] \geq x\right] \leq K \exp\left(-\eta\left(\frac{x^2}{\mathbb{E}[W]} \wedge x\right)\right)$$

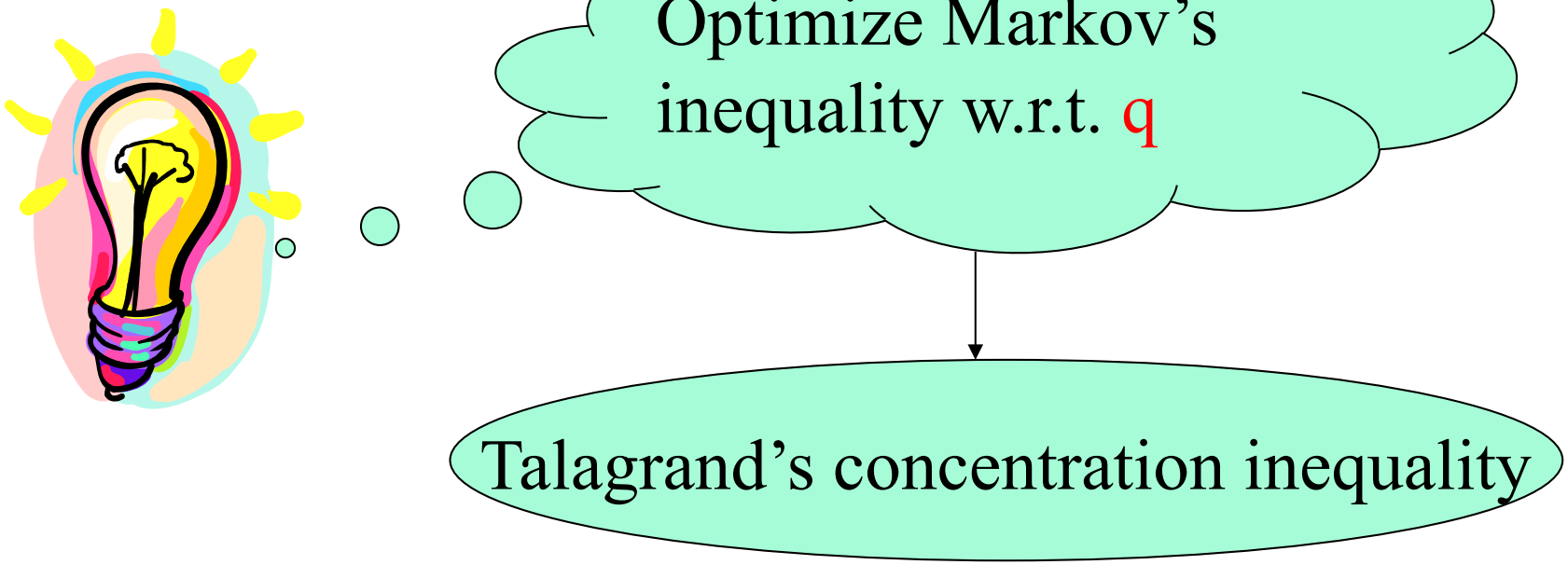
where $$W = \sup_{t \in S} \sum_{i=1}^{n} f_t^2(\xi_i)$$

- Moment inequalities in action

Why?

Example: for empirical processes, one can prove that for some absolute positive constant

$$\left\| Z - \mathbb{E}[Z] \right\|_q \le C\left[\sqrt{q\mathbb{E}[W]} \vee q \right] \quad \forall q \ge 2$$



Refining Talagrand's inequality

The key: In the process of recovering Talagrand's inequality via the moment method above, we may improve on the variance factor. Indeed, setting $f_t = \gamma(s,.) - \gamma(t,.)$ and

$$Z = \sup_{t \in S} \sum_{i=1}^{n} f_t(\xi_i) = \sum_{i=1}^{n} f_{\hat{s}}(\xi_i) = n\ell_n(s,\hat{s})$$

we see that

$$Z - Z_i^{'} \leq f_{\hat{s}}(\xi_i) - f_{\hat{s}}(\xi_i^{'})$$

and therefore

$$V^+ = \mathbb{E}\left[\sum_{i=1}^{n} \left(Z - Z_i^{'}\right)_+^2 | \xi\right] \leq 2\sum_{i=1}^{n} \left(Pf_{\hat{s}}^2 + f_{\hat{s}}^2(\xi_i)\right)$$

at this stage instead of using the crude bound

$$\frac{V^+}{n} \le 2\left(\sup_{t \in S} P f_t^2 + \sup_{t \in S} P_n f_t^2 \right)$$

we can use the refined bound

$$\frac{V^+}{n} \le 2Pf_{\hat{s}}^2 + 2P_n\left(f_{\hat{s}}^2\right) \le 4Pf_{\hat{s}}^2 + 2\left(P_n - P\right)\left(f_{\hat{s}}^2\right)$$

Now the point is that on the one hand

$$P\left(f_{\hat{s}}^2\right) \le w^2\left(\sqrt{\ell\left(s,\hat{s}\right)}\right)$$

and on the other hand we can handle the second term $\left(P_n - P\right)\left(f_{\hat{s}}^2\right)$ by using some kind of *square root trick*. $\left(P_n - P\right)\left(f_{\hat{s}}^2\right)$ can indeed shown to behave not worse than

$$\left(P_n - P\right)\left(f_{\hat{s}}\right) = \ell_n\left(s,\hat{s}\right) + \ell\left(s,\hat{s}\right)$$

So finally it can be proved that

$$\mathrm{Var}\left[Z\right] \le \mathbb{E}\left[V^+\right] \le Cnw^2\left(\varepsilon_*\right)$$

and similar results for higher moments.

Illustration 1

In the (bounded) regression case. If we consider the regressogram estimator on some partition with D pieces, it can be proved that

$$n\left\|\ell_n\left(s,\hat{s}\right)-\mathbb{E}\left[\ell_n\left(s,\hat{s}\right)\right]\right\|_q \leq C\left[\sqrt{qD}+q\right]$$

In this case $n\mathbb{E}\left[\ell_n\left(s,\hat{s}\right)\right]$ can be shown to be approximately proportional to D. This exemplifies the high dimensional Wilks phenomenon.

Application to model selection with adaptive penalties: Arlot and Massart, JMLR 2009.

.

Illustration 2

It can be shown that in the classification case,

If S is a VC-class with VC-dimension D, under the margin condition $|2\eta - 1| \geq h$

$$nh \left\| \ell_n(s, \hat{s}) - \mathbb{E}\left[\ell_n(s, \hat{s}) \right] \right\|_q \leq C \left[\sqrt{qD\left(1 + \log\left(\frac{nh^2}{D} \right) \right)} + q \right]$$

provided that $nh^2 \geq D$.

Application to model selection: work in progress with Saumard and Boucheron.

Thanks for your attention!