# Estimator selection: the calibration issue

Pascal Massart
Université Paris-Sud, Orsay

Toulouse
June 8 2017

# 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 $\mu$ .

- 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 $\varepsilon_i$ are i.i.d. with $E[\varepsilon_i | X_i]=0$

# Estimator selection

A versatile approach to functional estimation consists of considering some (possibly huge) collection of estimators $\{\hat{s}_m, m \in \mathfrak{M}\}$ and define some genuine selection rule $\hat{m}$ such that $\mathbb{E}\left[\ell(s, \hat{s}_{\hat{m}})\right]$ is as close as possible to the oracle

$$\inf_{m \in \mathfrak{M}} \mathbb{E}\left[\ell(s, \hat{s}_m)\right]$$

where $\ell$ is some given loss function.

In many cases the selection procedure involves some hyperparemeter $\lambda$ and most of the positive results ensuring that the selected estimator behaves approximately like an oracle are proved under the constraint that $\lambda$ is larger than some quantity which is more or less precisely known. The choice of $\lambda$ is left to the user…

Message: Negative results can be helpful to choose $\lambda$ from the data.

# Model Selection

# Empirical Risk Minimization (ERM)

Consider some *empirical criterion* (based on the data) $\gamma_n$ such that

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

achieves a minimum at point $t = s$ and the related loss $\ell(s,t) = E\left[\gamma_n(t) - \gamma_n(s)\right]$.

The ERM estimator $\hat{s}_m$ minimizes $\gamma_n$ over some « model » $S_m$. In this case the expected risk

$$\mathbb{E}\left[\ell(s,\hat{s}_m)\right]$$

reflects the quality of model $S_m$.

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

$$\ell(s,t) = K(s,t) \geq 0$$

Kullback Leibler information

- Least squares estimation (LSE)

## Regression

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

with

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

## Density

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

with

$$\ell(s,t) = \left\|t-s\right\|^2 \geq 0$$

# Model selection via penalization

Consider some empirical criterion $\gamma_n$ .

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

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

over $\mathfrak{M}$ and one defines the selected estimator

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

- 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' $C_p$ : $2D_m\sigma^2 / n$ ,

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

# How to penalize?

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

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

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

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

In order to (approximately) minimize

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

The key : Evaluate the excess risks

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

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

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

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

while for Mallows' $C_p$

$$\mathbb{E}\left[\ell(s_D, \hat{s}_D)\right] = \mathbb{E}\left[\hat{v}_D\right] = \frac{D\sigma^2}{n}$$

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

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

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

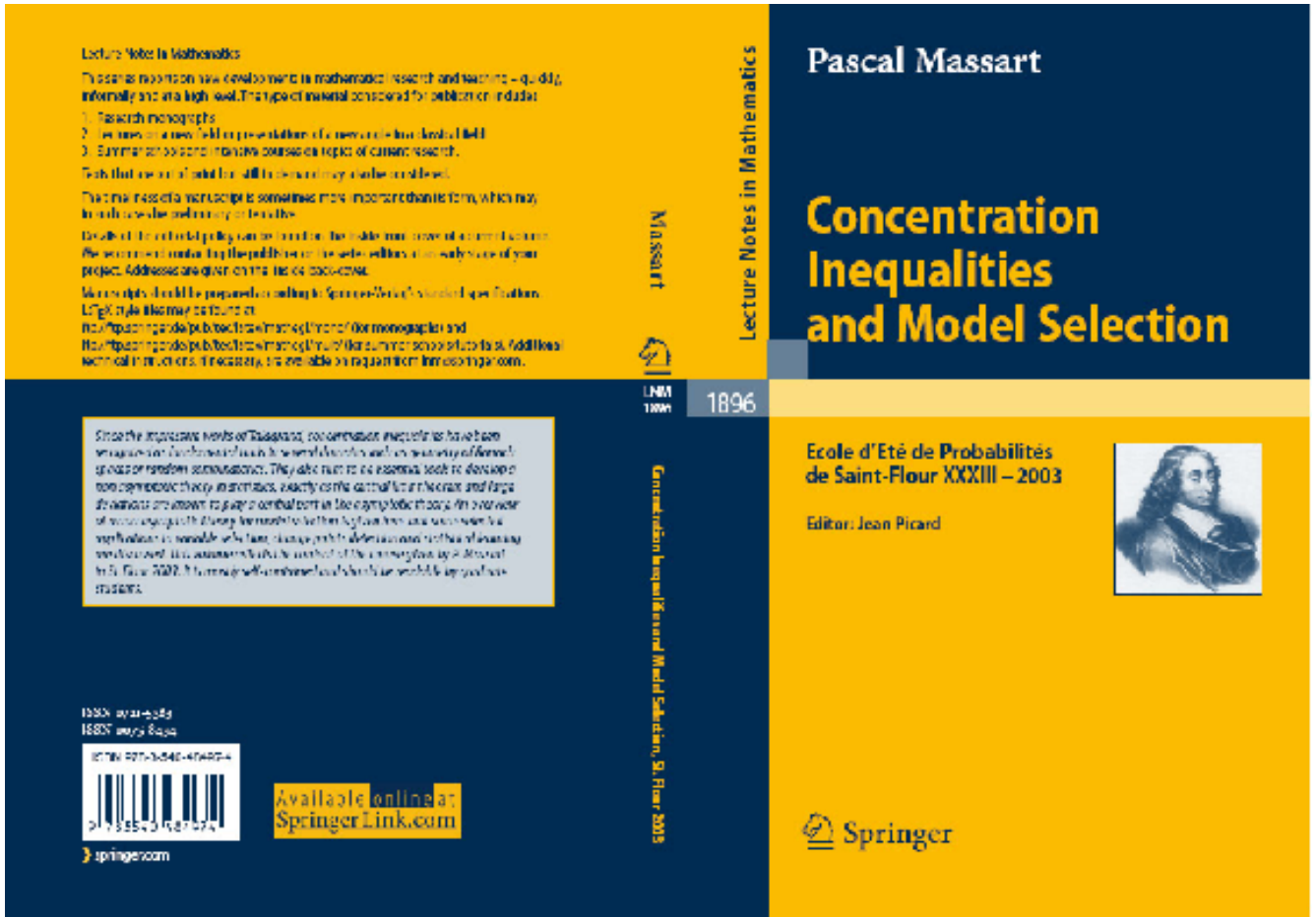
$$\bar{\gamma}_n(t) = \gamma_n(t) - E\left[\gamma_n(t)\right]$$

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 constant 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 $\hat{\lambda} D$ . When $D$ is large, $\gamma_n\left(s_D\right)$ is almost constant, it suffices to « read » $\hat{\lambda}$ as a slope on the graph of $\gamma_n\left(\hat{s}_D\right)$ . One chooses the final penalty as

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

The factor 2 which is finally used reflects the hope that the excess risks

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

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

« optimal » penalty=2 * « minimal » penalty

# Data driven penalization

## « 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 $\mathrm{pen}(D)$, typically $\mathrm{pen}(D)=\lambda D$ (but $\lambda D(2+\ln(n/D))$ is another possibility)
3. Estimate $\lambda$ from the data by multiplying by 2 the smallest value for which the penalized criterion explodes.

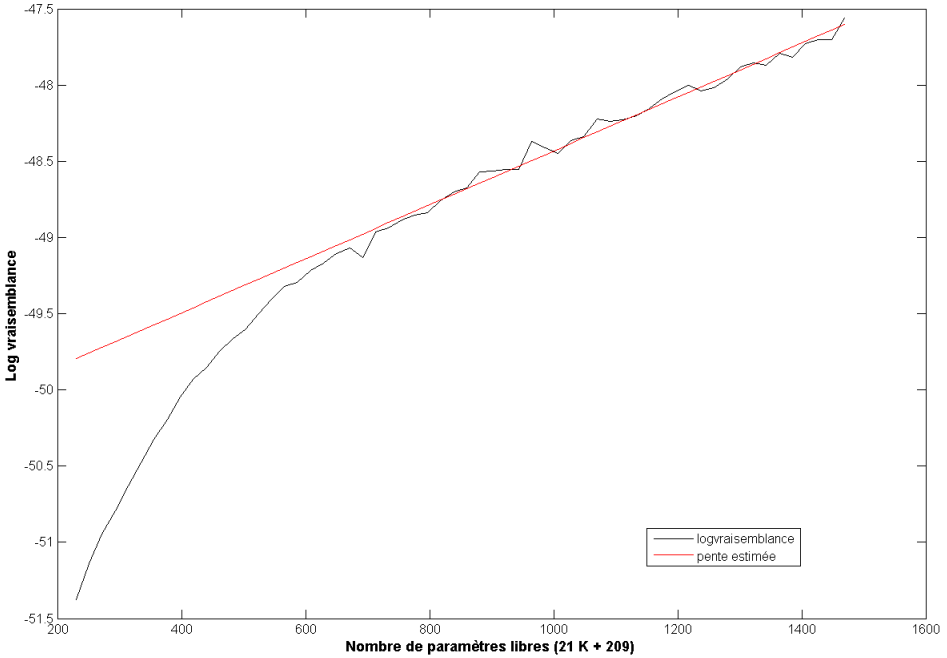
Theoretical validation in Birgé and M. (’07) for some Gaussian model selection issues.

Implemented by Lebarbier ( ‘05) for multiple change points detection.

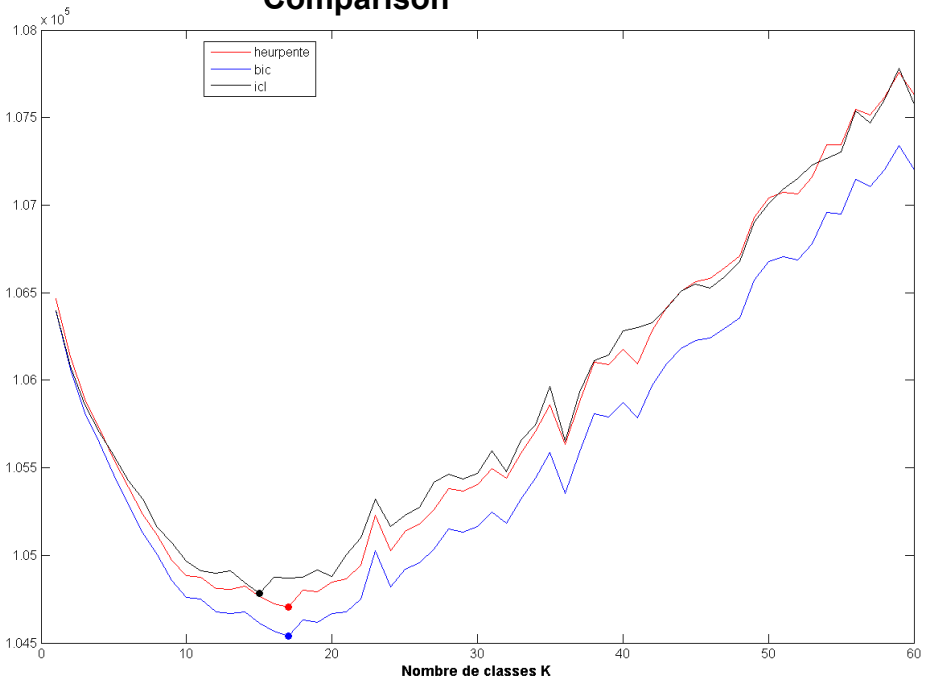
# Celeux, Martin, Maugis

- 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

**Adjustment of the slope**




**Comparison**



# Mathematical results

# Birgé, M. ('07)

One of the results that we proved for Gaussian model selection is that for a family of nested linear models taking a penalty of the form $\mathrm{pen}(D) = \lambda D$, if $\lambda$ is *below the critical value*

$$\lambda^* = \sigma^2 / n$$

then the *criterion* « *explodes* » while the choice $\lambda = 2\lambda^*$ leads to some optimal oracle inequality (asymptotically).

# Concentration inequalities

The proofs in the Gaussian case relie on concentration inequalities for chi-square distributions that are sharp enough to capture the behavior of the empirical excess loss when the dimension becomes large.

In particular, the explosion phenomenon derives from *lower tails inequalities*.

In the general case such inequalities are much more difficult to establish.

# Crucial issue: behavior of the excess losses

Concentration of the empirical excess loss:

connected to empirical processes theory because

$$\ell_n\left(s_D,\hat{s}_D\right)=\sup_{t\in S_D}\gamma_n\left(s_D,.\right)-\gamma_n\left(t,.\right)$$

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

The main task (Boucheron and M. PTRF '11): prove that the empirical excess loss concentrates around its expectation at a rate which is of order of the variance of $\gamma_n\left(s_D,.\right)-\gamma_n\left(t,.\right)$ at point $t=\hat{s}_D$ (typically smaller than the maximal variance).

- Concentration tools

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

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

A Burkholder-type inequality (BBLM, AOP2005)

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

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

## Illustration

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

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

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

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

.

# Saumard (EJS '13)

Extension to piecewise polynomials:

linear expansion+control of the remainder term via empirical process technics (Koltchinskii and Giné (2006)).

Conclusion: empirical excess loss and expected loss are of the same order and (approximately) proportional to dimension.

# Selection of linear estimators

# Arlot and Bach (2011)

Gaussian fixed design regression

$$\hat{s}_m = A_m Y = A_m s + A_m \varepsilon$$

(example: the $m$-nearest neighbours estimator). It is still relevent to consider the penalized least square criterion to select among some collection of such estimators by minimizing

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

One can extend the above analysis in this context. This time the minimal penalty writes

$$\mathrm{pen}_{\min}(m) = \frac{\sigma^2}{n}\left(2Tr\left(A_m^T A_m\right) - Tr\left(A_m\right)\right)$$

while the optimal penalty has a possibly different shape

$$\mathrm{pen}_{opt}(m) = \frac{2\sigma^2}{n} Tr\left(A_m\right)$$

For projection estimators one recovers the same formulas as before

$$\mathrm{pen}_{\min}(m) = \frac{\sigma^2 D_m}{n} \qquad \mathrm{pen}_{opt}(m) = 2\,\mathrm{pen}_{\min}(m)$$

And the same holds true for $m$-nearest neighbours estimators if one sets $D_m = n/m$

But of course, it may perfectly happen (for ridge regression for instance) that the optimal and minimal penalties are *no longer linked within a factor 2*.

Nevertheless one can still use the explosion property of selection criterion

$$\mathrm{pen}_{\lambda}(m) = \lambda\left(2Tr\left(A_m^T A_m\right) - Tr\left(A_m\right)\right)$$

below the critical level $\lambda^* = \sigma^2 / n$ , to estimate $\lambda^*$ from the data by $\hat{\lambda}$ and finally use as a penalty

$$\mathrm{pen}(m) = \hat{\lambda} Tr\left(A_m\right)$$

# Lepski's method

Lepski's method is an alternative selection method to penalized empirical risk minimization. It has the advantage to be usable *whatever the loss function*. Several versions are available, here is one due to Goldensluger and Lepski. Assume the collection to be ordered and define

$$B(m) = \sup_{m' \geq m} \left( \ell(\hat{s}_m, \hat{s}_{m'}) - \mathrm{pen}(m') \right)_+$$

and then select $\hat{m}$ minimizing

$$B(m) + \mathrm{pen}(m)$$

Is the concept of minimal penalty still relevent for this method?

An answer can be provided in the context which widely used by Lepski and his co-authors: kernel density estimation

## Kernel density estimation

Let us consider the density (with respect to Lebesgue measure on the real line) estimation framework. Given some convolution « kernel » $K$ on the real line and some collection (grid) of bandwidths $\{h_m, m \in \mathfrak{M}\}$ the corresponding collection of kernel density estimators is defined by

$$\hat{s}_m(x) = \sum_{i=1}^{n} K_{h_m}(x - X_i)$$

with $K_h := \frac{1}{h} K\left(\frac{\cdot}{h}\right)$

Joint work with Claire Lacour (SPA'15). For bandwith selection of kernel density estimators, for the squared $L_2$-loss, we can prove that a minimal penalty does exist

$$\mathrm{pen}_{\min}(m) = \frac{\left\| K_{h_m} \right\|_2^2}{n}$$

The existence of an optimal penalty is not that clear and we decided to shift to some close in spirit but different method.

# PCO: A new selection method

In the process of understanding how to calibrate Goldenshluger-Lepski's method we discovered that it can be much simplified. Assume that there exists some « worse/best » estimator $\hat{s}_N$ as far as the « variance/bias » trade off is concerned and consider this time

$$B(m) = \ell\left(\hat{s}_m, \hat{s}_N\right)$$

We introduce the Penalized Comparison to Overfitting which consists of selecting $\hat{m}$ minimizing

$$B(m)+\text{pen}(m)$$

We have been able to show (joint work with C. Lacour and V. Rivoirard (Sankhya'17)) that for bandwith selection of kernel density estimators, for the square $L_2$ -loss, taking the penalty as

$$\text{pen}_\lambda(m)=\lambda\frac{\|K_{h_m}\|_2^2}{n}-\frac{\|K_{h_N}-K_{h_m}\|_2^2}{n}$$

leads to a minimal penalty for the critical value $\lambda = 0$ , while the value $\lambda = 1$ corresponds to an optimal choice of the penalty. Simulations are confirming the theory (which remains valid in the multivariate case).

# Open problems

- Several algorithms (CART, stepwise variable selection, Lasso) have been designed to oversome this difficulty, leading to the natural issue of selecting a model among a data dependent list. Empirical studies indicate that the slope heuristics behaves well (Thesis: Meynet (2012), Devijver (2015)).
- We do not know yet what are the limitations of PCO (different loss functions, different kind of estimators).