

Model Selection and Geometry

Pascal Massart
Université Paris-Sud, Orsay

Leipzig, February 22

Purpose of the talk

- Concentration of measure plays a fundamental role in the theory of model selection
- Model selection may be useful for geometric inference
- Focus on the Gaussian framework: convenient to provide the main ideas and intuitions

Concentration of measure

Isoperimetry

The **Gaussian isoperimetric Theorem**: let P be the standard Gaussian measure on \mathbb{R}^N . Given θ , among the sets A with $P(A) = \theta$

$$P\{d(\cdot, A) < \varepsilon\}$$

is minimal for a half-space

$$\mathbb{R}^{N-1} \times (-\infty, \varepsilon_\theta]$$

with $Q(\varepsilon_\theta) = \theta$ (Q : standard Gaussian tail function, d : Euclidean distance)

Gaussian concentration inequality

and therefore if $\theta \geq 1/2$, then $\varepsilon_\theta \geq 0$ and

$$P\{d(\cdot, A) \geq \varepsilon\} \leq Q(\varepsilon + \varepsilon_\theta) \leq Q(\varepsilon) \leq e^{-\varepsilon^2/2}$$

Given some L -Lipschitz function f , applying this inequality with $\varepsilon = \sqrt{2x}$ to the set

$$A = \{f \leq Mf\}$$

where Mf is a median of f leads to

$$P\{f - Mf \geq L\sqrt{2x}\} \leq e^{-x}$$

True (not obvious) with Ef instead of Mf

Chi-square distribution

If we apply this inequality to the Euclidean norm itself, i.e. $f : z \rightarrow \|z\|$, then $L=1$ and since $Ef^2 = N$

$$Ef \leq \sqrt{N}$$

and therefore

$$P\left\{f \geq \sqrt{N} + \sqrt{2x}\right\} \leq e^{-x}$$

Of course f^2 follows a chi-square distribution $\chi^2(N)$.

The Gaussian concentration inequality also applies to suprema of Gaussian processes.

Let $\{G(t), t \in T\}$ be some centered Gaussian process. Assume that T is equipped with the covariance pseudo-metric d defined by $d(u, t) = \sqrt{\text{Var}(G(t) - G(u))}$

If (T, d) is separable and if $\{G(t), t \in T\}$ is a.s. continuous on (T, d) , setting $Z = \sup_{t \in T} G(t)$ one has

$$P\left[Z \geq E[Z] + \sigma \sqrt{2x}\right] \leq e^{-x}$$

where $\sigma^2 = \sup_{t \in T} \text{Var}[G(t)]$

Outside the Gaussian world

- Concentration of measure for product measures has been studied for years by **M. Talagrand** in a series of remarkable papers.
- **Talagrand's** concentration inequality for empirical processes plays a fundamental role in statistics in general and for model selection in particular.
- Gaussian life is easier!

Model Selection

What are we talking about?

- Statistical inference

One observes X (random vector, random process...) with unknown distribution P .

Purpose: take a decision about some quantity $\theta(P)$ related to P . Predict with a level of confidence.

What to do: design a genuine estimation procedure $\hat{\theta}(X)$ of θ and get some idea of how far it is from the target θ .

- The role of probability theory

Problem: the exact distribution of $\hat{\theta}$ is generally unknown.

Solution: make some approximation or evaluation based on Probability Theory.

- Asymptotic Theory

Typically, when $X = (X_1, \dots, X_n)$, $n \rightarrow \infty$ and X_1, \dots, X_n are independent. *Asymptotic theory* in statistics uses limit theorems (Central limit Theorems, Large Deviation Principles) as approximation tools.

Example: when P belongs to some given model (*which does not depend of n*), behavior of maximum likelihood estimators.

- Model selection

Designing a genuine $\hat{\theta}(X)$ requires some prior knowledge on P . Choosing a proper model for P is a major problem for the statistician.

Aim of model selection : construct *data-driven* criteria to select a model among a given list.

- 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

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

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

- **Non asymptotic Theory**

In many situations, it is useful 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

Large values of n !

But the size of the models as well the size of the list of models should be authorized to be large too. This approach is based on

Concentration Inequalities

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...



Gaussian Model Selection

The Gaussian framework

We consider the generalized linear Gaussian model. This means that, given some separable Hilbert space \mathbb{H} , one observes

$$Y_\varepsilon(t) = \langle s, t \rangle + \varepsilon W(t), \quad t \in \mathbb{H}$$

where W is some centered Gaussian isonormal process, i.e. maps isometrically \mathbb{H} onto some Gaussian subspace of $\mathbb{L}_2(\Omega)$. We have in mind that writing the level of noise as $\varepsilon = 1/\sqrt{n}$ allows an easy comparison with other frameworks involving n observations.

This framework is convenient to cover both the infinite dimensional white noise model and the finite dimensional linear model for which $\mathbb{H} = \mathbb{R}^n$ and $W(t) = \langle \xi, t \rangle$, where ξ is a standard Gaussian random vector.

Consider some collection $(S_m)_{m \in \mathcal{M}}$ of (*typically closed and convex*) subsets of \mathbb{H} .

We also consider the least squares criterion

$$L_\varepsilon(t) = \|t\|^2 - 2Y_\varepsilon(t)$$

and, for each S_m , \hat{s}_m minimizing L_ε over S_m . The quality of model S_m is measured by the quadratic risk

$$E \|s - \hat{s}_m\|_2^2$$

Does the risk reflect the model choice paradigm?

Let us compute it in the simplest situation where S_m is a linear model with dimension D_m . Then

$$E \left\| s - \hat{s}_m \right\|_2^2 = \varepsilon^2 D_m + d^2(s, S_m)$$

The best (oracle) model according to this criterion should make a trade-off between its size and its quality of approximation to the truth. The aim is now to mimic it.

Our purpose is to analyze the penalized LSE $\hat{s}_{\hat{m}}$, where

$$\hat{m} = \arg \min_{m \in \mathcal{M}} \left\{ L_\varepsilon(\hat{s}_m) + \text{pen}(m) \right\}$$

Selecting linear models

Let us begin with model selection among linear models and review some results obtained in joint works with **Lucien Birgé (JEMS' 01 and PTRF' 07)**.

Each model S_m is assumed to be linear with dimension D_m and represented by the least squares estimator \hat{s}_m on S_m . Then

$$E \left\| s - \hat{s}_m \right\|_2^2 = \varepsilon^2 D_m + d^2(s, S_m)$$

Variance

Bias

Oracle : Ideal model achieving $\inf_{m \in \mathfrak{M}} E \left\| s - \hat{s}_m \right\|_2^2$

Aim: mimic the oracle by estimating the risk

- Mallows' C_p heuristics

Let s_m denote the orthogonal projection of s on S_m .

An « ideal » model should minimize the quadratic risk

$$\|s - s_m\|_2^2 + \varepsilon^2 D_m = \|s\|^2 - \|s_m\|^2 + \varepsilon^2 D_m \quad \text{Pythagore}$$

or equivalently $-\|s_m\|^2 + \varepsilon^2 D_m$

Substituting to $\|s_m\|^2$ its natural unbiased estimator $\|\hat{s}_m\|^2 - \varepsilon^2 D_m$ leads to Mallows' C_p criterion

$$-\|\hat{s}_m\|^2 + 2\varepsilon^2 D_m = L_\varepsilon(\hat{s}_m) + 2\varepsilon^2 D_m$$

Issue : look the way $\|\hat{s}_m\|^2$ concentrates around $\varepsilon^2 D_m$ uniformly w.r.t. $m \in \mathfrak{M}$.

Theorem (Birgé, Massart' 01)

Let $(x_m)_{m \in \mathfrak{M}}$ be a family of non negative weights such that $\sum_{m \in \mathfrak{M}} e^{-x_m} = \Sigma < \infty$

Let $K > 1$. Assume that

$$\text{pen}(m) \geq K^2 \varepsilon^2 \left(\sqrt{D_m} + \sqrt{2x_m} \right)^2$$

and take

$$\hat{m} = \operatorname{argmin}_{m \in \mathfrak{M}} \left[-\|\hat{s}_m\|^2 + \text{pen}(m) \right]$$

Then

$$E \|\hat{s}_{\hat{m}} - s\|^2 \leq C(K) \left\{ \inf_{m \in \mathfrak{M}} \left[d^2(s, S_m) + \text{pen}(m) \right] + \Sigma \varepsilon^2 \right\}$$

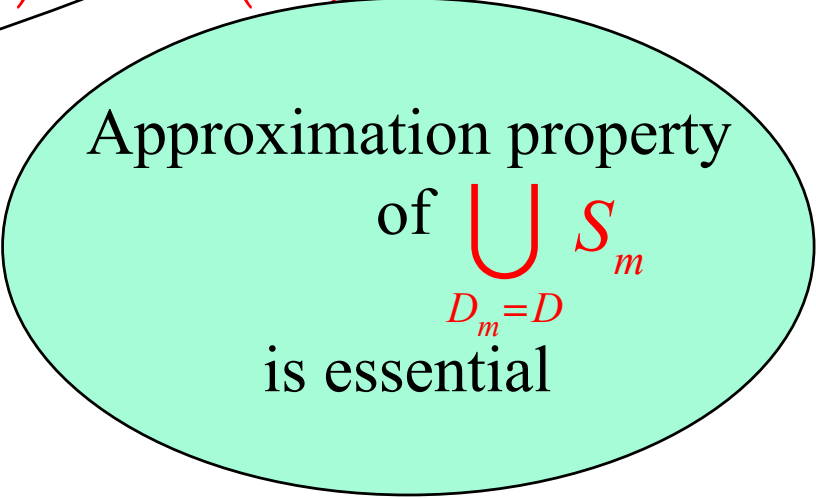
Choice of the weights

A typical choice is $x_m = x(D_m)$. Then if one defines

$$x(D) = \alpha D + \log \# \{m \in \mathfrak{M} : D_m = D\}$$

the weights actually appear as the price to pay for redundancy (i.e. many models with the same dimension). The penalty becomes $\text{pen}(m) = \text{pen}(D_m)$ and the upper bound (up to constant) writes

$$\left\{ \inf_D \left\{ \inf_{D_m=D} d^2(s, S_m) + \text{pen}(D) \right\} \right\}$$



Approximation property
of $\bigcup_{D_m=D} S_m$
is essential

Comparison with the oracle

Since

$$\mathbb{E} \left\| \hat{s}_m - s \right\|_2^2 = d^2(s, S_m) + \varepsilon^2 D_m$$

If the weights $(x_m)_{m \in \mathfrak{M}}$ are such that $x_m \leq LD_m$ and $\sum_{m \in \mathfrak{M}} e^{-x_m} \leq 1$ then the upper bound provided

by the Theorem becomes

$$\inf_{m \in \mathfrak{M}} \mathbb{E} \left\| \hat{s}_m - s \right\|_2^2$$

(up to some multiplicative constant).

Conclusion : The selected estimator performs (almost) as well as an oracle.

Variable Selection

Let $\{\varphi_j, j \leq N\}$ be a family of elements in \mathbb{H} . Let \mathfrak{M} be some collection of subsets of $\{1, \dots, N\}$ and define $S_m = \langle \varphi_j, j \in m \rangle, m \in \mathfrak{M}$.

a. Ordered variable selection

\mathfrak{M} is the collection of subsets of the form $\{1, \dots, D\}$ with $D \leq N$ (possibly $N = \infty$)

Then one can take

$$\text{pen}(m) = K' \varepsilon^2 |m|$$

$K' > 1$
sharp

Comparable to the oracle (linearly independent case)

$$\mathbb{E} \left\| s - \hat{s}_{\hat{m}} \right\|^2 \leq C(K') \inf_{m \in \mathfrak{M}} \mathbb{E} \left\| s - \hat{s}_m \right\|^2$$

b. Complete variable selection

Let \mathfrak{M} be the collection of all subsets of $\{1, \dots, N\}$.

One can take $x_m = |m| \ln(N)$ which leads to

$$\Sigma = \sum_{m \in \mathfrak{M}} e^{-x_m} = \sum_{D \leq N} C_N^D e^{-D \ln(N)} \leq e.$$

and

$$\text{pen}(m) = K^2 \varepsilon^2 |m| \left(1 + \sqrt{2 \ln(N)}\right)^2$$

with $K > 1$. In the orthonormal case $\hat{s}_{\hat{m}}$ can be explicitated. It is simply hard thresholding (Donoho, Johnstone, Kerkyacharian, Picard)

$$\hat{s}_{\hat{m}} = \sum_{j=1}^N \hat{\beta}_j \mathbf{1}_{|\hat{\beta}_j| \geq T} \varphi_j, \quad \text{with } \hat{\beta}_j = Y(\varphi_j)$$

with $T = K\varepsilon\left(1 + \sqrt{2\ln(N)}\right)$. Then

$$E\|s - \hat{s}_{\hat{m}}\|_2^2 \leq C'(K) \inf_{D \leq N} \left\{ \inf_{|m|=D} \|s - s_m\|^2 + \varepsilon^2 D \log(N) \right\}$$

$K > 1$
sharp

Role of the basis

Refinement : One can choose $x_m = D(\ln(N/D) + 2)$ whenever $D = |m|$, which leads to an optimal risk bound (minimax sense).

- Link with non parametric adaptation

Let $\{\varphi_j, j \geq 1\}$ be the Fourier basis with its natural ordering then $\|s - s_D\|_2^2 \leq \kappa \|s^{(\alpha)}\|_2^2 D^{-2\alpha}$.

Up to constant $E_s \|s - \hat{s}_{\hat{m}}\|_2^2$ is bounded by

$$R^{\frac{2}{2\alpha+1}} \varepsilon^{\frac{4\alpha}{2\alpha+1}}$$

uniformly w.r.t s satisfying $\|s^{(\alpha)}\|_2 \leq R$.

Optimal in minimax sense since

$$\inf_{\hat{s}} \sup_{\|s^{(\alpha)}\|_2 \leq R} E_s \|s - \hat{s}\|_2^2 \geq \kappa' R^{\frac{2}{2\alpha+1}} \varepsilon^{\frac{4\alpha}{2\alpha+1}}$$

Conclusion Minimax, up to constant over all the Sobolev balls $\|s^{(\alpha)}\|_2 \leq R$ simultaneously.

- **Conclusions**
- Mallows' criterion can underpenalize.
- Condition $K > 1$ is sharp.
- What penalty should be recommended? One can try to optimize the oracle inequality. The result is that twice the minimal value is a good choice (Birgé, Massart (2007))
- **Practical use** One does not know the level of noise but one can retain from the theory that
 - « optimal » penalty = 2 « minimal » penalty
 - *data-driven penalty*

Back to Geometry

Theorem (M. 2007)

Assume that for every $m \in \mathcal{M}$, there exists some function $\phi_m \nearrow$ such that $\phi_m(x)/x \searrow$ on $(0, +\infty)$ and

$$2E \left[\sup_{t \in S_m} \left[\frac{W(t) - W(u)}{\|t - u\|^2 + x^2} \right] \right] \leq x^{-2} \phi_m(x)$$

for any positive x and any point u in S_m .

Let us define D_m such that $\phi_m(\varepsilon \sqrt{D_m}) = \varepsilon D_m$ and consider some family of weights $(x_m)_{m \in \mathcal{M}}$ such that

$$\sum_{m \in \mathcal{M}} e^{-x_m} = \Sigma < \infty$$

Let K be some constant with $K > 1$ and take

$$\text{pen}(m) \geq K^2 \varepsilon^2 \left(\sqrt{D_m} + \sqrt{2x_m} \right)^2.$$

Then

$$E \left\| \hat{s}_{\hat{m}} - s \right\|^2 \leq C(K) \left\{ \inf_{m \in \mathfrak{M}} \left[d^2(s, S_m) + \text{pen}(m) \right] + \Sigma \varepsilon^2 \right\}$$

Comments

If S_m is finite dimensional with dimension D'_m we can take $D'_m = D_m$ which shows that the above Theorem strictly implies the linear model selection Theorem (Birgé, Massart ('01)).
Indeed since $2x\|t - u\| \geq x^2 + \|t - u\|^2$

if $(\psi_j)_{1 \leq j \leq D'_m}$ denotes some orthonormal basis of S_m

$$2 \sup_{t \in S_m} \left[\frac{W(t) - W(u)}{\|t - u\|^2 + x^2} \right] \leq x^{-1} \sup_{t \in S_m} \left[\frac{W(t)}{\|t\|} \right] = \sqrt{\sum_{j=1}^{D'_m} W^2(\psi_j)}$$

therefore
$$2E \left[\sup_{t \in S_m} \left[\frac{W(t) - W(u)}{\|t - u\|^2 + x^2} \right] \right] \leq x^{-1} \sqrt{D'_m}$$

So we may take
$$\phi_m(x) = x\sqrt{D'_m} \longrightarrow D'_m = D_m$$

More generally, the function ϕ_m should be understood as a modulus of continuity of W over model S_m . One can indeed use a *peeling device* to ensure that if for every $u \in S_m$ and every $\sigma > 0$

$$E \left[\sup_{t \in S_m, \|t-u\| \leq \sigma} (W(t) - W(u)) \right] \leq \phi_m(\sigma)$$

where φ_m is a non decreasing and concave function with $\varphi_m(0) = 0$, then one can take

$$\phi_m(x) = 5\varphi_m(x)$$

One possibility is to use covering numbers

$$\varphi_m(\sigma) = \kappa \int_0^\sigma \sqrt{\log N(\delta, S_m)} d\delta$$

Simplicial approximation

Here is an illustration of what we are promoting with [Frédéric Chazal](#) and our students.

[C. Caillerie and B. Michel \(2011\)](#) have applied the Gaussian model selection Theorem above to simplicial approximation.

Consider that one observes

$$x_i = s_i + \sigma \xi_i \quad \text{with } 1 \leq i \leq n$$

The unobserved deterministic points s_i 's belong to \mathbb{R}^p and the variables ξ_i 's are i.i.d. standard normal random vectors. One has in mind that the points s_i 's are sampled on some geometrical object that one wants to learn.

They consider some collection of k -homogenous simplicial complexes $\{C_m, m \in \mathfrak{M}\}$ in \mathbb{R}^p . They use covering numbers to define a proper penalty term and derive a risk bound on the selected estimator

$$E \left\| s - \hat{s}_{\hat{m}} \right\|_2^2$$

where S simply denotes the vector of \mathbb{R}^{np} « built » from the S_i 's.

As expected (at least if the collection is not too rich) the penalty recommended by the Theorem is found to be proportionnal to

$$\log |C_m|_k$$

where $|C_m|_k$ is measuring the size of the simplicial complex. More precisely

$$|C_m|_k = \left(\sum_{\omega \in C_m^+} \Delta_{\omega}^k \right)^{1/k}$$

where the sum is extended to the set of simplices with maximal dimension k and Δ_{ω} denotes the diameter of the smallest Euclidean ball containing the simplex ω . The bad news are that the penalty involves nasty constant (and the level of noise that we do not know in practice).

They use a heuristics to solve this problem.

They use the penalized criterion

$$\|X - \hat{s}_m\|^2 + 2\hat{\beta} \log|C_m|_k$$

where $\hat{\beta}$ is estimated from the data by using the following phase transition which is empirically observed: when β is too small ($\beta < \hat{\beta}$) the criterion

$$\|X - \hat{s}_m\|^2 + \beta \log|C_m|_k$$

chooses a simplicial complex with a size which is close to the maximal one in the collection.

Many remaining issues...

Among which: escaping from the square loss,
building an adaptive estimation theory...

Thanks for your attention!