Model selection and estimator selection for statistical learning

Sylvain Arlot

[1] Cnrs

[2] École Normale Supérieure (Paris), Liens, Équipe Sierra

Scuola Normale Superiore di Pisa, 14–23 February 2011

Outline of the 5 lectures

1. Statistical learning
2. Model selection for least-squares regression
3. Linear estimator selection for least-squares regression
4. Resampling and model selection
5. Cross-validation and model/estimator selection

Part II

Model selection for least-squares regression

Outline

Outline

A key lemma

Lemma

Let $\mathrm{pen} : \mathcal{M}_n \mapsto \mathbb{R}$ *some penalty (possibly data-dependent).*
On the event Ω *on which for every* $m, m' \in \mathcal{M}_n$,

$$\big(\mathrm{pen}(m) - \mathrm{pen}_{\mathrm{id}}(m, D_n) \big) - \big(\mathrm{pen}(m') - \mathrm{pen}_{\mathrm{id}}(m', D_n) \big) \leq A(m) + B(m')$$

we have $\forall \widehat{m} \in \arg\min_{m \in \mathcal{M}_n} \left\{ \frac{1}{n} \left\| \widehat{F}_m - Y \right\|^2 + \mathrm{pen}(m) \right\}$

$$\frac{1}{n} \left\| \widehat{F}_{\widehat{m}} - F \right\|^2 - B(\widehat{m}) \leq \inf_{m \in \mathcal{M}_n} \left\{ \frac{1}{n} \left\| \widehat{F}_m - F \right\|^2 + A(m) \right\}$$

Oracle inequality for Gaussian regression (1)

Assumptions:

- Fixed design regression, least-squares contrast
- Gaussian homoscedastic noise: $\varepsilon \sim \mathcal{N}(0, \sigma^2)$
- Model collection of polynomial complexity: $\mathrm{Card}(\mathcal{M}_n) \leq C n^{\alpha}$
- For all $m \in \mathcal{M}_n$, $\widehat{F}_m = A_m Y = \Pi_{S_m} Y$ (least-squares estimator)
- Penalty

$$\mathrm{pen}(m) = \frac{K \sigma^2 \dim(S_m)}{n} \qquad \text{with} \quad K > 1$$

Oracle inequality for Gaussian regression (2)

$$-B(m) \leq \mathrm{pen}(m) - \mathrm{pen}_{\mathrm{id}}(m, D_n) \leq A(m)$$

$$\Rightarrow \qquad \frac{1}{n} \left\| \widehat{F}_{\widehat{m}} - F \right\|^2 - B(\widehat{m}) \leq \inf_{m \in \mathcal{M}_n} \left\{ \frac{1}{n} \left\| \widehat{F}_m - F \right\|^2 + A(m) \right\}$$

$$\mathrm{pen}_{\mathrm{id}}(m, D_n) = \frac{2}{n} \langle A_m \varepsilon, \varepsilon \rangle + \frac{2}{n} \langle (A_m - I_n) F, \varepsilon \rangle$$

First term has expectation $\frac{2\sigma^2 \dim(S_m)}{n}$, the second term is centered.

Oracle inequality for Gaussian regression (3)

Two Gaussian concentration results (see Massart 2007):

Proposition

Let ξ be some standard Gaussian vector in $\mathbb{R}^n$, $\alpha \in \mathbb{R}^n$, $M \in \mathcal{M}_n(\mathbb{R})$. Then, for every $x \geq 0$,

$$\mathbb{P}\left(\left| \langle \xi , \alpha \rangle \right| \leq \sqrt{2x} \left\| \alpha \right\|_2 \right) \geq 1 - 2e^{-x}$$

$$\mathbb{P}\left(\left| \langle \xi , M\xi \rangle - \operatorname{tr}(M) \right| \leq 2\sqrt{x \operatorname{tr}\left(M^\top M \right)} + 2 \left\| M \right\| x \right) \geq 1 - 2e^{-x}$$

Oracle inequality for Gaussian regression (4)

Sketch of the proof:

- For all $m \in \mathcal{M}_n$,
  concentrate $\langle A_m \varepsilon, \varepsilon \rangle$ around $\sigma^2 \dim(S_m)$
  and $\langle (A_m - I_n) F, \varepsilon \rangle$ around 0
- Apply the Lemma on the intersection of these $\mathrm{Card}(\mathcal{M}_n)$ events
- Control the remainder terms

Oracle inequality for Gaussian regression (5)

Theorem (Birgé & Massart 2007)

For every $x \geq 0$, *with probability at least* $1 - 4\,\mathrm{Card}(\mathcal{M}_n) e^{-x}$, *for every*

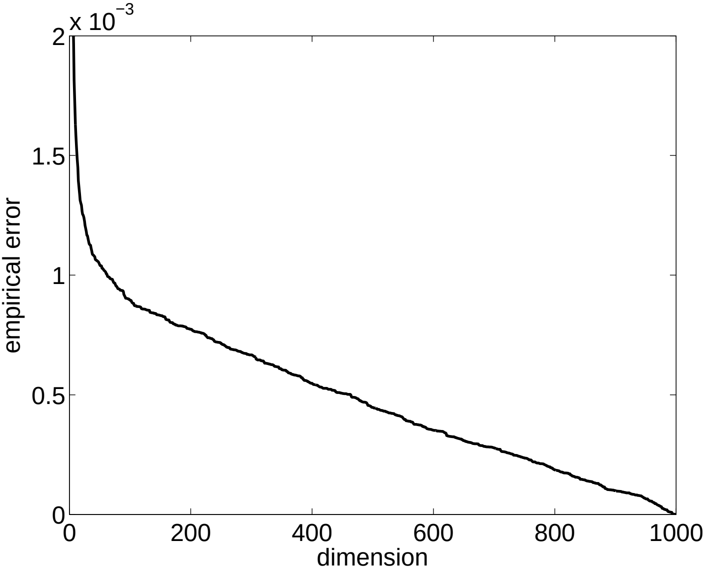
$$\widehat{m} \in \arg\min_{m \in \mathcal{M}_n} \left\{ \frac{1}{n} \left\| Y - \widehat{F}_m \right\|^2 + \frac{K \sigma^2 \dim(S_m)}{n} \right\} \enspace ,$$

we get the oracle inequality $\forall \delta > 0$,

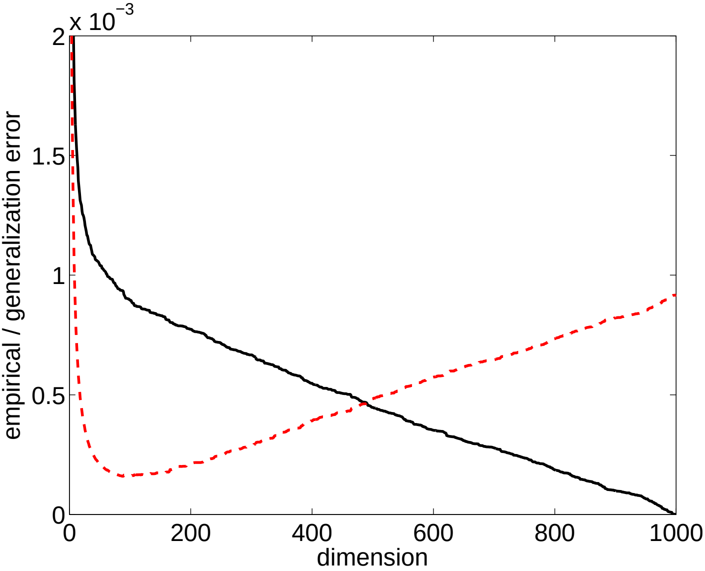
$$\frac{1}{n} \left\| \widehat{F}_{\widehat{m}} - F \right\|^2 \leq \left(\frac{1 + (K-2)_+}{1 - (2-K)_+} + \delta \right) \inf_{m \in \mathcal{M}_n} \left\{ \frac{1}{n} \left\| \widehat{F}_m - F \right\|^2 \right\} + \frac{C(K) x \sigma^2}{\delta n}$$

Outline

Motivation (1): L-curve and elbow heuristics?



Motivation (1): L-curve and elbow heuristics?



Motivation (2): what if $K \leq 1$?

Theorem (Birgé & Massart 2007)

If $K > 1$, for every $x \geq 0$, with probability $1 - 4\,\mathrm{Card}(\mathcal{M}_n) e^{-x}$, for every

$$\widehat{m} \in \arg\min_{m \in \mathcal{M}_n} \left\{ \frac{1}{n} \left\| Y - \widehat{F}_m \right\|^2 + \frac{K\sigma^2 \dim(S_m)}{n} \right\} \enspace ,$$

we get the concentration inequality $\forall \delta > 0$,

$$\frac{1}{n} \left\| \widehat{F}_{\widehat{m}} - F \right\|^2 \leq \left(\frac{1 + (K-2)_+}{1 - (2-K)_+} + \delta \right) \inf_{m \in \mathcal{M}_n} \left\{ \frac{1}{n} \left\| \widehat{F}_m - F \right\|^2 \right\} + \frac{C(K) x \sigma^2}{\delta n}$$

Motivation (3): penalty calibration

- C_p and C_L (Mallows, 1973):

$$\mathrm{pen}(m) = \frac{2\sigma^2 D_m}{n}$$

$$\mathrm{pen}(m) = \frac{2\sigma^2 \operatorname{tr}(A_m)}{n}$$

- Penalties proportional to D_m with the optimal multiplying factor unknown: change-point detection (Birgé & Massart, 2001; Lebarbier, 2005), mixture models (Maugis & Michel, 2008), and so on
- Rademacher penalties

$$\mathrm{pen}(m) = 2 \times \mathbb{E}\left[\sup_{t \in S_m} \left\{ \frac{1}{n} \sum_{i=1}^{n} \left(\varepsilon_i \gamma\left(t; \xi_i\right)\right) \middle| \, D_n \right\} \right]$$

- ...

Naive estimator of σ^2

Example: homoscedastic regression on a fixed design
Computation of the empirical risk

Naive estimator of σ^2

Example: homoscedastic regression on a fixed design

$$\mathbb{E}\left[\frac{1}{n} \left\| Y - \widehat{F}_m \right\|^2 \right] = \frac{1}{n} \left\| (I_n - A_m) F \right\|^2 + \frac{\sigma^2 (n - D_m)}{n}$$

Naive estimator of σ^2:

$$\widehat{\sigma}_m^2 := \frac{1}{n - D_m} \left\| Y - \widehat{F}_m \right\|^2$$

Bias of this estimator:

$$\mathbb{E}\left[\widehat{\sigma}_m^2 \right] = \sigma^2 + \frac{1}{n - D_m} \left\| (I_n - A_m) F \right\|^2$$

Naive estimator of σ^2

Example: homoscedastic regression on a fixed design

$$\mathbb{E}\left[\frac{1}{n}\left\|Y - \widehat{F}_m\right\|^2\right] = \frac{1}{n}\left\|(I_n - A_m)F\right\|^2 + \frac{\sigma^2(n - D_m)}{n}$$

Naive estimator of σ^2:

$$\widehat{\sigma}_m^2 := \frac{1}{n - D_m}\left\|Y - \widehat{F}_m\right\|^2$$

Bias of this estimator:

$$\mathbb{E}\left[\widehat{\sigma}_m^2\right] = \sigma^2 + \frac{1}{n - D_m}\left\|(I_n - A_m)F\right\|^2$$

$\Rightarrow$ Using it inside the penalty $2\sigma^2 D_m / n$?

Naive estimator of σ^2

Naive estimator of σ^2:

$$\widehat{\sigma}_m^2 := \frac{1}{n - D_m} \left\| Y - \widehat{F}_m \right\|^2$$

Bias of this estimator:

$$\mathbb{E}\left[\widehat{\sigma}_m^2 \right] = \sigma^2 + \frac{1}{n - D_m} \left\| (I_n - A_m) F \right\|^2$$

$\Rightarrow$ Using it inside the penalty $2\sigma^2 D_m / n$?
First idea:

$$\operatorname{crit}(m) = \frac{1}{n} \left\| Y - \widehat{F}_m \right\|^2 + \frac{2 \widehat{\sigma}_{m_0}^2 D_m}{n}$$

Drawbacks: we have to know/choose m_0, overpenalization by an unknown factor

FPE (Akaike, 1970) and GCV (Craven & Wahba, 1979)

$$\mathrm{crit}_{\mathrm{FPE}}(m) = \frac{1}{n} \left\| Y - \widehat{F}_m \right\|^2 + \frac{2 \widehat{\sigma}_m^2 D_m}{n} = \frac{1}{n} \left\| Y - \widehat{F}_m \right\|^2 \left(1 + \frac{2 D_m}{n - D_m} \right)$$

(Akaike, 1970; see also Baraud, Giraud & Huet, 2009)

FPE (Akaike, 1970) and GCV (Craven & Wahba, 1979)

$$\mathrm{crit}_{\mathrm{FPE}}(m) = \frac{1}{n} \left\| Y - \widehat{F}_m \right\|^2 + \frac{2 \widehat{\sigma}_m^2 D_m}{n} = \frac{1}{n} \left\| Y - \widehat{F}_m \right\|^2 \left(1 + \frac{2 D_m}{n - D_m} \right)$$

(Akaike, 1970; see also Baraud, Giraud & Huet, 2009)

Generalized cross-validation (GCV, Craven & Wahba, 1979)

$$\mathrm{crit}_{\mathrm{GCV}}(m) = \frac{1}{n} \frac{\left\| Y - \widehat{F}_m \right\|^2}{\left(1 - \frac{D_m}{n} \right)^2}$$

FPE (Akaike, 1970) and GCV (Craven & Wahba, 1979)

$$\mathrm{crit}_{\mathrm{FPE}}(m) = \frac{1}{n} \left\| Y - \widehat{F}_m \right\|^2 + \frac{2\widehat{\sigma}_m^2 D_m}{n} = \frac{1}{n} \left\| Y - \widehat{F}_m \right\|^2 \left(1 + \frac{2D_m}{n - D_m} \right)$$

(Akaike, 1970; see also Baraud, Giraud & Huet, 2009)

Generalized cross-validation (GCV, Craven & Wahba, 1979)

$$\mathrm{crit}_{\mathrm{GCV}}(m) = \frac{1}{n} \frac{\left\| Y - \widehat{F}_m \right\|^2}{\left(1 - \frac{D_m}{n} \right)^2}$$

If $D_m \ll n$,

$$\mathrm{crit}_{\mathrm{GCV}}(m) \approx \frac{1}{n} \left\| Y - \widehat{F}_m \right\|^2 \frac{n + D_m}{n - D_m} = \frac{1}{n} \left\| Y - \widehat{F}_m \right\|^2 \left(1 + \frac{2D_m}{n - D_m} \right)$$

FPE (Akaike, 1970) and GCV (Craven & Wahba, 1979)

$$\mathrm{crit}_{\mathrm{FPE}}(m) = \frac{1}{n}\left\|Y - \widehat{F}_m\right\|^2 + \frac{2\widehat{\sigma}_m^2 D_m}{n} = \frac{1}{n}\left\|Y - \widehat{F}_m\right\|^2 \left(1 + \frac{2D_m}{n - D_m}\right)$$

(Akaike, 1970; see also Baraud, Giraud & Huet, 2009)

Generalized cross-validation (GCV, Craven & Wahba, 1979)

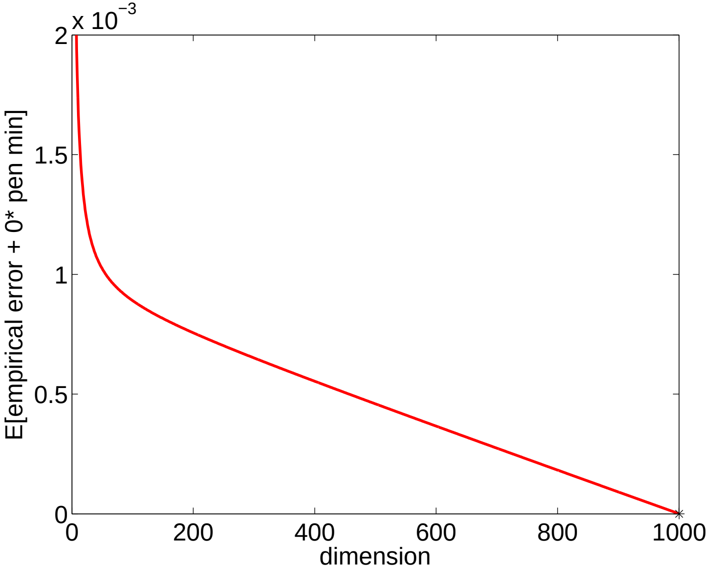
$$\mathrm{crit}_{\mathrm{GCV}}(m) = \frac{1}{n}\frac{\left\|Y - \widehat{F}_m\right\|^2}{\left(1 - \frac{D_m}{n}\right)^2}$$

If $D_m \ll n$,

$$\mathrm{crit}_{\mathrm{GCV}}(m) \approx \frac{1}{n}\left\|Y - \widehat{F}_m\right\|^2 \frac{n + D_m}{n - D_m} = \frac{1}{n}\left\|Y - \widehat{F}_m\right\|^2 \left(1 + \frac{2D_m}{n - D_m}\right)$$

Drawbacks: for the largest models $\left\|Y - \widehat{F}_m\right\|^2 \approx 0$

$\mathbb{E}[\|Y - \widehat{F}_m\|^2] = n\sigma^2 + \|F - F_m\|^2 - \sigma^2 \times D_m$



Outline

Minimal penalty: heuristics

For all $C > 0$,

$$\widehat{m}(C) \in \arg\min_{m \in \mathcal{M}_n} \left\{ \frac{1}{n} \left\| Y - \widehat{F}_m \right\|^2 + \frac{C D_m}{n} \right\}$$

$\Rightarrow \exists C_{\min}$ s.t.: for $C < C_{\min}$, $\widehat{F}_{\widehat{m}(C)}$ overfits

for $C > C_{\min}$, oracle-inequality for $\widehat{F}_{\widehat{m}(C)}$?

Minimal penalty: heuristics

For all $C > 0$,

$$\widehat{m}(C) \in \arg\min_{m \in \mathcal{M}_n} \left\{ \frac{1}{n} \left\| Y - \widehat{F}_m \right\|^2 + \frac{C D_m}{n} \right\}$$

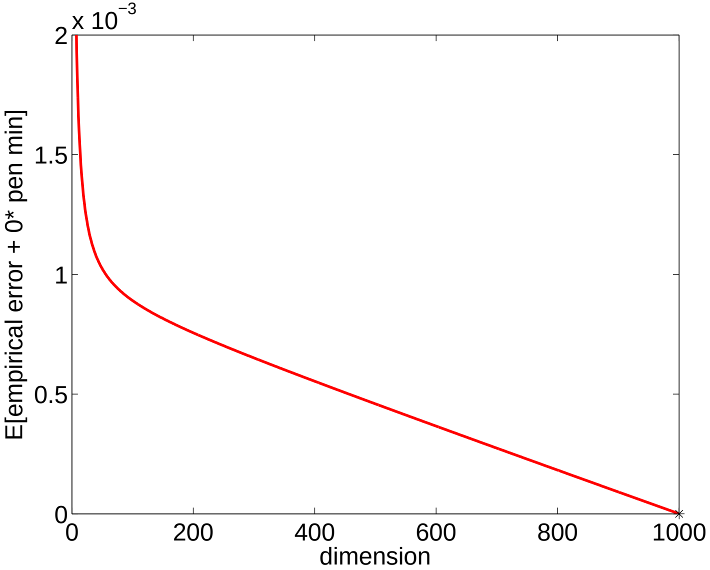
$\Rightarrow \exists C_{\min}$ s.t.: for $C < C_{\min}$, $\widehat{F}_{\widehat{m}(C)}$ overfits

for $C > C_{\min}$, oracle-inequality for $\widehat{F}_{\widehat{m}(C)}$?

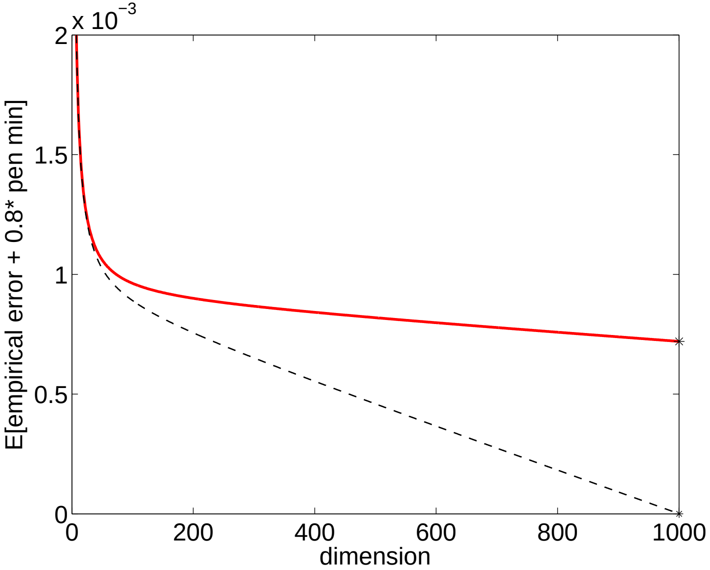
$$m^\star(C) \in \arg\min_{m \in \mathcal{M}_n} \left\{ \mathbb{E}\left[\frac{1}{n} \left\| Y - \widehat{F}_m \right\|^2 + \frac{C D_m}{n} \right] \right\}$$

$$= \arg\min_{m \in \mathcal{M}_n} \left\{ \frac{1}{n} \left\| F - F_m \right\|^2 + (C - \sigma^2) \frac{D_m}{n} \right\}$$

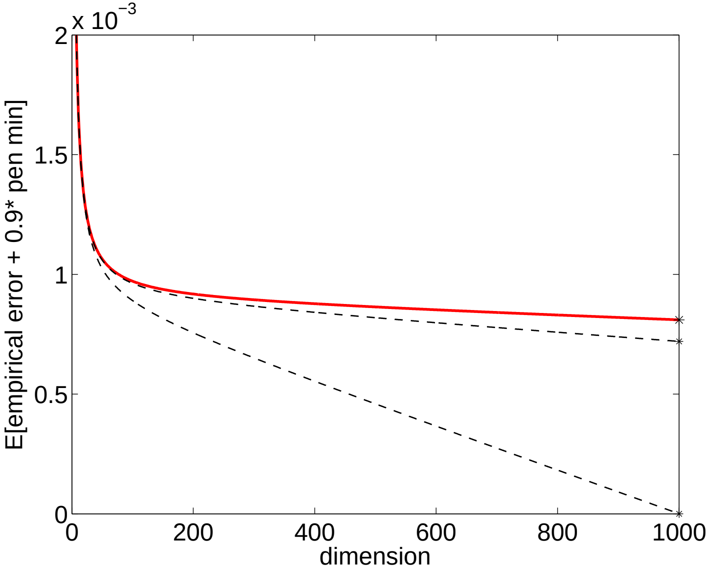
$\mathbb{E}[n^{-1}\|Y - \widehat{F}_m\|^2] + 0 \times \sigma^2 D_m n^{-1}$



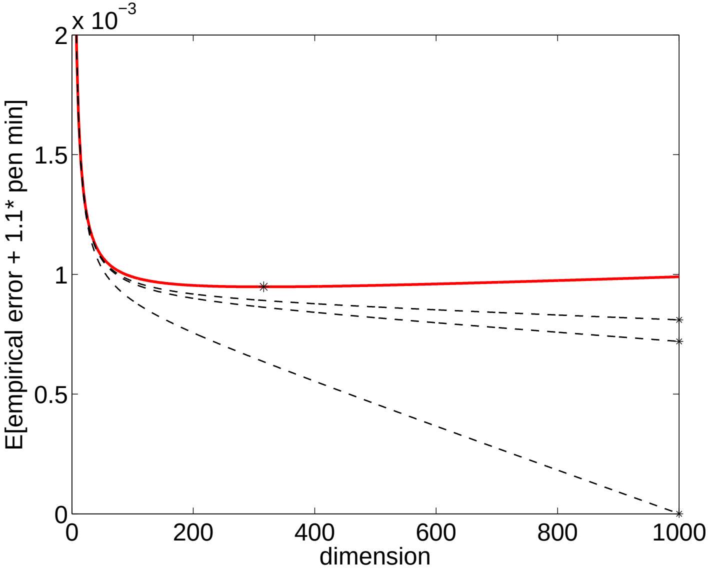
$\mathbb{E}[n^{-1}\|Y - \widehat{F}_m\|^2] + 0.8 \times \sigma^2 D_m n^{-1}$



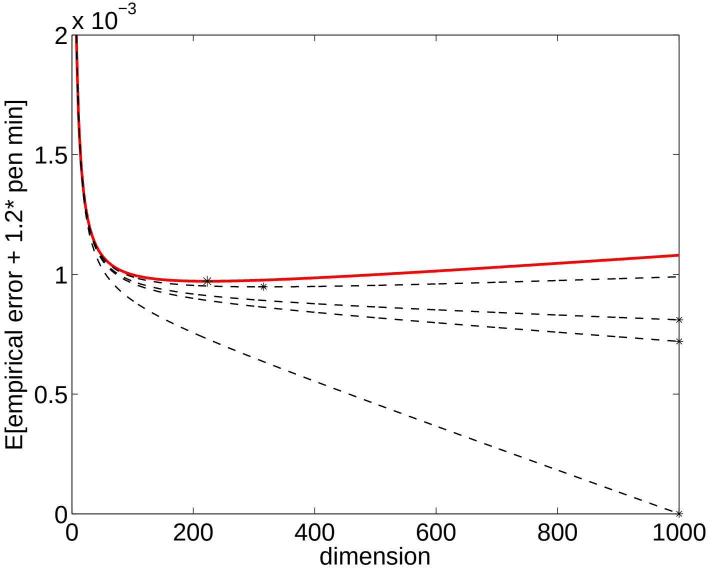
$\mathbb{E}[n^{-1}\|Y - \widehat{F}_m\|^2] + 0.9 \times \sigma^2 D_m n^{-1}$



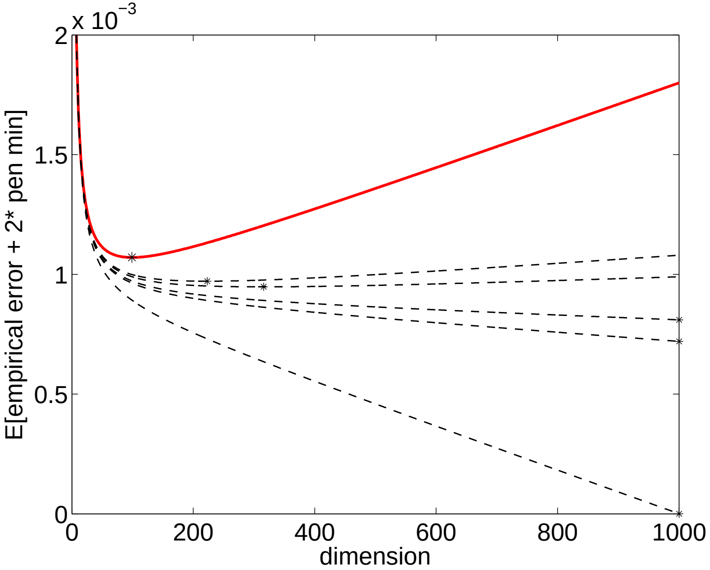
$\mathbb{E}[n^{-1}\|Y - \widehat{F}_m\|^2] + 1.1 \times \sigma^2 D_m n^{-1}$



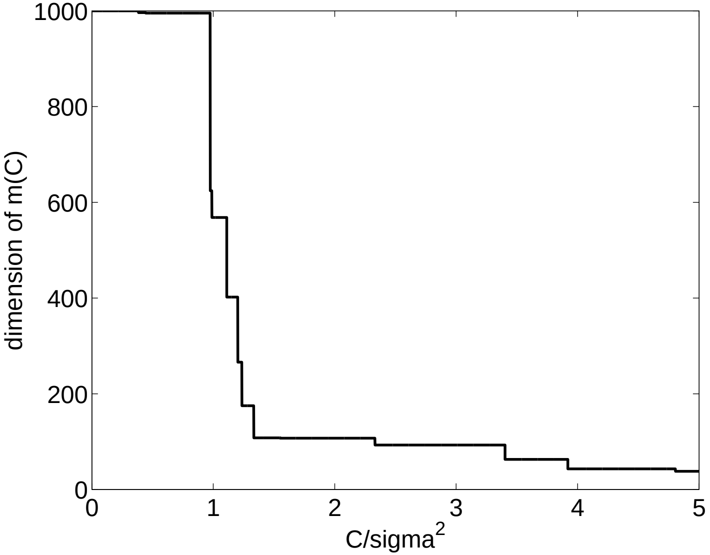
$\mathbb{E}[n^{-1}\|Y - \widehat{F}_m\|^2] + 1.2 \times \sigma^2 D_m n^{-1}$



$\mathbb{E}[n^{-1}\|Y - \widehat{F}_m\|^2] + 2 \times \sigma^2 D_m n^{-1}$



Dimension jump



Calibration of penalties (Birgé & Massart 2007)

1. for all $C > 0$, compute

$$\widehat{m}(C) \in \arg\min_{m \in \mathcal{M}_n} \left\{ \frac{1}{n} \left\| Y - \widehat{F}_m \right\|^2 + C \frac{D_m}{n} \right\}$$

2. find $\widehat{C}_{\min}$ such that $D_{\widehat{m}(C)}$ is "too large" when $C < \widehat{C}_{\min}$ and "reasonably small" when $C > \widehat{C}_{\min}$

3. select $\widehat{m} = \widehat{m}\left(2\widehat{C}_{\min}\right)$

Proof: assumptions and concentration inequalities

Assumptions:

- polynomial complexity: $\mathrm{Card}(\mathcal{M}_n) \leq C_{\mathcal{M}} n^{\alpha}$
- homoscedastic Gaussian noise, fixed design
- $\exists m_1, \ m_2 \in \mathcal{M}_n$ s.t. $D_{m_1} \geq n/2$, $D_{m_2} \leq \sqrt{n}$ and $\forall i \in \{1, 2\}$, $n^{-1} \| F - F_{m_i} \|^2 \leq \sigma^2 \sqrt{\ln(n)/n}$

Proposition

If $\xi \sim \mathcal{N}(0, I_n)$, $\alpha \in \mathbb{R}^n$, $M \in \mathcal{M}_n(\mathbb{R})$, for all $x \geq 0$,

$$\mathbb{P}\left(\left| \langle \xi, \, \alpha \rangle \right| \leq \sqrt{2x} \, \|\alpha\|_2 \right) \geq 1 - 2e^{-x}$$

$$\mathbb{P}\left(\left| \langle \xi, \, M\xi \rangle - \mathrm{tr}(M) \right| \leq 2\sqrt{x \, \mathrm{tr}\left(M^{\top} M \right)} + 2 \, \|M\| \, x \right) \geq 1 - 2e^{-x}$$

Theorem (1): Minimal penalty / Dimension jump

Theorem (Birgé & Massart 2007, A. & Bach 2009)

With probability at least $1 - 4C_{\mathcal{M}} n^{-2}$, *si* $n \geq n_0(\alpha)$,

$$\forall C < \left(1 - 42 \sqrt{\frac{(\alpha+2)\ln(n)}{n}} \right) \sigma^2 \ , \qquad D_{\widehat{m}(C)} \geq \frac{n}{3}$$

$$\forall C > \left(1 + 8 \frac{\sqrt{(\alpha+2)\ln(n)}}{n^{1/4}} \right) \sigma^2 \ , \qquad D_{\widehat{m}(C)} \leq n^{3/4}$$

and in the first case,

$$\left\| F - \widehat{F}_{\widehat{m}(C)} \right\|^2 \geq \ln(n) \inf_{m \in \mathcal{M}_n} \left\{ \left\| F - \widehat{F}_m \right\|^2 \right\}$$

Theorem (2): Oracle inequality

Theorem (Birgé & Massart 2007)

For every $x \geq 0$, *with probability* $1 - 4\,\mathrm{Card}(\mathcal{M}_n) e^{-x}$, *for every* $K > 1$, $\delta > 0$, *and every*

$$\widehat{m}\left(K\sigma^2\right) \in \arg\min_{m \in \mathcal{M}_n} \left\{ \frac{1}{n} \left\| Y - \widehat{F}_m \right\|^2 + \frac{K\sigma^2 \dim(S_m)}{n} \right\} ,$$

$$\frac{1}{n} \left\| \widehat{F}_{\widehat{m}(K\sigma^2)} - F \right\|^2 \leq \left(\frac{1 + (K-2)_+}{1 - (2-K)_+} + \delta \right) \inf_{m \in \mathcal{M}_n} \left\{ \frac{1}{n} \left\| \widehat{F}_m - F \right\|^2 \right\} + \frac{C(K) x \sigma^2}{\delta n}$$

Theorem (2): Oracle inequality

Theorem (Birgé & Massart 2007)

If $\mathbb{P}(2\widehat{C} \in \left[(1-\eta_-)2\sigma^2, (1+\eta_+)2\sigma^2\right]) \geq 1 - 4C_{\mathcal{M}} n^{-2}$.

For every $x \geq 0$, *with probability* $1 - 4\operatorname{Card}(\mathcal{M}_n) e^{-x} - 4C_{\mathcal{M}} n^{-2}$, *for every* $\delta > 0$, *and every*

$$\widehat{m}\left(2\widehat{C}\right) \in \arg\min_{m \in \mathcal{M}_n} \left\{ \frac{1}{n}\left\| Y - \widehat{F}_m \right\|^2 + \frac{2\widehat{C}\dim(S_m)}{n} \right\} ,$$

$$\frac{1}{n}\left\| \widehat{F}_{\widehat{m}(2\widehat{C})} - F \right\|^2 \leq \left(\frac{1+\eta_+}{1-\eta_-} + \delta \right) \inf_{m \in \mathcal{M}_n} \left\{ \frac{1}{n}\left\| \widehat{F}_m - F \right\|^2 \right\} + \frac{\max\left\{ C(2-\eta_-), C(2+\eta_+) \right\} x \sigma^2}{\delta n}$$

Theorem (2): Oracle inequality

Theorem (Birgé & Massart 2007)

We take $x = (\alpha + 2)\ln(n)$ *and assume* $n \geq n_0(\alpha)$.

With probability $1 - 4C_{\mathcal{M}} n^{-2}$, *for every*

$$\widehat{m}\left(2\widehat{C}\right) \in \arg\min_{m \in \mathcal{M}_n} \left\{ \frac{1}{n} \left\| Y - \widehat{F}_m \right\|^2 + \frac{2\widehat{C}\dim(S_m)}{n} \right\} \quad ,$$

$$\frac{1}{n}\left\| \widehat{F}_{\widehat{m}(2\widehat{C})} - F \right\|^2 \leq \left(1 + \frac{L_\alpha \sqrt{\ln(n)}}{n^{1/4}} + \delta \right) \inf_{m \in \mathcal{M}_n} \left\{ \frac{1}{n} \left\| \widehat{F}_m - F \right\|^2 \right\} + \frac{L_\alpha \ln(n) \sigma^2}{\delta n}$$

Outline

The slope heuristics (Birgé & Massart, 2007)

1. existence of a minimal penalty $\mathrm{pen}_{\mathrm{min}}(m)$:

$$\widehat{m}_{\mathrm{min}}(C) \in \arg\min_{m \in \mathcal{M}_n} \left\{ P_n \gamma\left(\widehat{s}_m\right) + C \,\mathrm{pen}_{\mathrm{min}}(m) \right\}$$

$$\frac{\ell\left(s^{\star}, \widehat{s}_{\widehat{m}_{\mathrm{min}}(C)}\right)}{\inf_{m \in \mathcal{M}_n} \left\{ \ell\left(s^{\star}, \widehat{s}_m\right) \right\}} \quad \text{jumps at } C = 1$$

The slope heuristics (Birgé & Massart, 2007)

1. existence of a minimal penalty $\mathrm{pen}_{\min}(m)$:

$$\widehat{m}_{\min}(C) \in \arg\min_{m \in \mathcal{M}_n} \left\{ P_n \gamma\left(\widehat{s}_m\right) + C \,\mathrm{pen}_{\min}(m) \right\}$$

$$\frac{\ell\left(s^\star, \widehat{s}_{\widehat{m}_{\min}(C)}\right)}{\inf_{m \in \mathcal{M}_n} \left\{ \ell\left(s^\star, \widehat{s}_m\right) \right\}} \quad \text{jumps at } C = 1$$

2. the minimal penalty can be detected:
   $\mathcal{C}_{\widehat{m}_{\min}(C)}$ "jumps" around $C = 1$

The slope heuristics (Birgé & Massart, 2007)

1. existence of a minimal penalty $\mathrm{pen}_{\min}(m)$:

$$\widehat{m}_{\min}(C) \in \arg\min_{m \in \mathcal{M}_n} \left\{ P_n \gamma\left(\widehat{s}_m\right) + C \, \mathrm{pen}_{\min}(m) \right\}$$

$$\frac{\ell\left(s^{\star}, \widehat{s}_{\widehat{m}_{\min}(C)}\right)}{\inf_{m \in \mathcal{M}_n} \left\{ \ell\left(s^{\star}, \widehat{s}_m\right) \right\}} \quad \text{jumps at } C = 1$$

2. the minimal penalty can be detected:
$\mathcal{C}_{\widehat{m}_{\min}(C)}$ "jumps" around $C = 1$

3. link between minimal and optimal penalty:

$$\mathrm{pen}_{\mathrm{opt}}(m) \approx 2 \, \mathrm{pen}_{\min}(m)$$

Data-driven penalties with the slope heuristics

Inputs: $\left(\mathrm{pen}_0(m)\right)_{m \in \mathcal{M}_n}$ $\left(\mathcal{C}_m\right)_{m \in \mathcal{M}_n}$

Assumption: $\mathrm{pen}_0(m) \propto \mathrm{pen}_{\min}(m)$

1. for every $C > 0$, compute
$$\widehat{m}(C) \in \arg\min_{m \in \mathcal{M}_n} \left\{ P_n \gamma\left(\widehat{s}_m\right) + C \, \mathrm{pen}_0(m) \right\}$$
2. find $\widehat{C}_{\min}$ such that $\mathcal{C}_{\widehat{m}(C)}$ is "too large" when $C < \widehat{C}_{\min}$ and "reasonably small" when $C > \widehat{C}_{\min}$
3. select $\widehat{m} = \widehat{m}\left(2\widehat{C}_{\min}\right)$.

Slope heuristics recipe

$$p_2(m) = P_n \left(\gamma \left(s_m^\star \right) - \gamma \left(\widehat{s}_m \right) \right) \qquad \mathrm{pen}_{\min}(m) = \mathbb{E} \left[p_2(m) \right]$$

Slope heuristics recipe

$$p_2(m) = P_n\left(\gamma\left(s_m^\star\right) - \gamma\left(\widehat{s}_m\right)\right) \qquad \mathrm{pen}_{\min}(m) = \mathbb{E}\left[p_2(m)\right]$$

$$p_1(m) = P\left(\gamma\left(\widehat{s}_m\right) - \gamma\left(s_m^\star\right)\right) \qquad \delta(m) = (P - P_n)\gamma\left(s_m^\star\right)$$

$$\mathrm{pen}_{\mathrm{id}}(m) = p_1(m) + p_2(m) - \delta(m)$$

$$\mathbb{E}\left[\mathrm{pen}_{\mathrm{id}}(m)\right] = \mathrm{pen}_{\mathrm{opt}}(m) = \mathbb{E}\left[p_1(m)\right] + \mathbb{E}\left[p_2(m)\right]$$

Slope heuristics recipe

$$p_2(m) = P_n \left(\gamma \left(s_m^\star \right) - \gamma \left(\widehat{s}_m \right) \right) \qquad \mathrm{pen}_{\min}(m) = \mathbb{E} \left[p_2(m) \right]$$

$$p_1(m) = P \left(\gamma \left(\widehat{s}_m \right) - \gamma \left(s_m^\star \right) \right) \qquad \delta(m) = (P - P_n) \gamma \left(s_m^\star \right)$$

$$\mathrm{pen}_{\mathrm{id}}(m) = p_1(m) + p_2(m) - \delta(m)$$

$$\mathbb{E} \left[\mathrm{pen}_{\mathrm{id}}(m) \right] = \mathrm{pen}_{\mathrm{opt}}(m) = \mathbb{E} \left[p_1(m) \right] + \mathbb{E} \left[p_2(m) \right]$$

Heuristics: $p_1(m) \approx p_2(m)$

- concentration of p_1, p_2, δ
- $\mathbb{E} \left[p_1(m) \right] \approx \mathbb{E} \left[p_2(m) \right]$
- increase of the expectation for compensating the bias

Known results

- Least-squares, regression, homoscedastic Gaussian noise (Birgé & Massart, 2007)
- Heteroscedastic regressograms (A. & Massart, 2009)
- Least-squares density estimation, i.i.d. (Lerasle, 2009) or mixing data (Lerasle, 2010)
- Minimum contrast estimators, regular contrast (Saumard, 2010)

Minimizer of a regular contrast (Saumard, 2010)

- Regular contrast on some convex model S_m:
  - $s_m^\star \in \arg\min_{t \in S_m} P\gamma(t)$ exists
  - $t \in S_m \mapsto P\gamma(t)$ strictly convex
  - $\exists c > 0, \ t \in B_\infty(s_m^\star, c) \mapsto \gamma(t; \cdot) \in L_\infty(P)$ is $\mathcal{C}^3$

- Concentration of $p_1(m)$ and $p_2(m)$ around the same deterministic quantity $D_m \mathcal{K}_m^2 / (4n)$ (unobservable in general)

$+$ control of $\|\widehat{s}_m - s_m^\star\|_\infty$

$\Rightarrow$ validates the slope heuristics for:

  - heteroscedastic regression (histograms, piecewise polynomials)
  - least-squares density estimation
  - log-likelihood density estimation on histograms

Experimental results

- Change-point detection (Lebarbier, 2005)
- Gaussian mixture models (Maugis & Michel, 2008)
- Unsupervised classification (choice of the number of clusters) (Baudry, 2009)
- Computational geometry (Caillerie & Michel, 2009)
- Lasso (Connault, 2011)
- ...

for a complete list, see Baudry, Maugis & Michel, 2010

Outline

Practical qualities of the algorithm

- visual checking of existence of a jump
- calibration independent from the choice of some m_0
- too strong overfitting almost impossible
- one remaining parameter: how to localize the jump

How to localize the jump in practice?

- Complexity jump: largest jump? largest relative jump? complexity threshold?

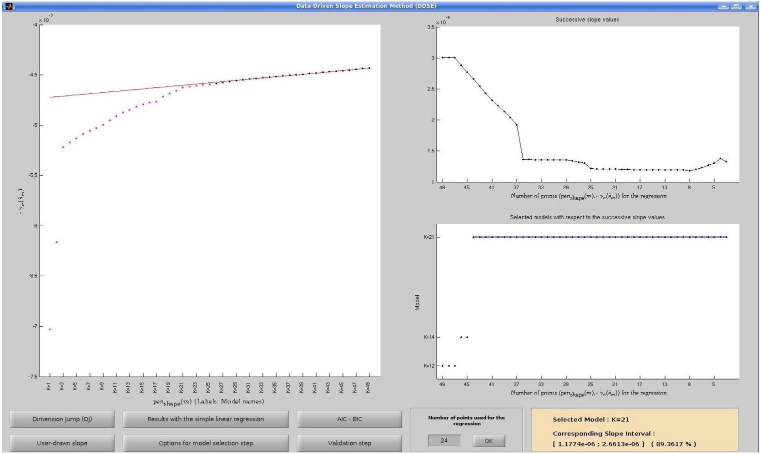
How to localize the jump in practice?

- Complexity jump: largest jump? largest relative jump? complexity threshold?
- Estimation of the slope of the empirical risk as a function of the complexity:
  computed with which models? robust regression?

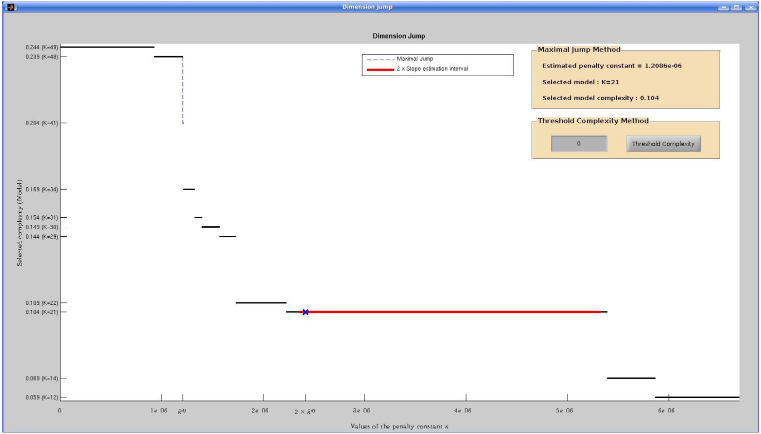
How to localize the jump in practice?

- Complexity jump: largest jump? largest relative jump? complexity threshold?
- Estimation of the slope of the empirical risk as a function of the complexity:
  computed with which models? robust regression?
- Jump vs. slope? Take both!
  $\Rightarrow$ package CAPUSHE (Baudry, Maugis & Michel, 2010)
  `http://www.math.univ-toulouse.fr/~maugis/CAPUSHE.html`

CAPUSHE (Baudry, Maugis & Michel, 2010): slope



CAPUSHE (Baudry, Maugis & Michel, 2010): jump



Outline

Overpenalization