

Model/estimator selection

Sylvain Arlot^{1,2,3}

¹Université Paris-Saclay

²Inria Saclay, Celeste project-team

³Institut Universitaire de France

CEA/EDF/INRIA Summer School 2021
June 18, 2021

Main references: [arXiv:1901.07277](https://arxiv.org/abs/1901.07277) (parts 1 & 2)
[arXiv:0907.4728](https://arxiv.org/abs/0907.4728) or [hal-01485508](https://hal.archives-ouvertes.fr/hal-01485508) (part 3)

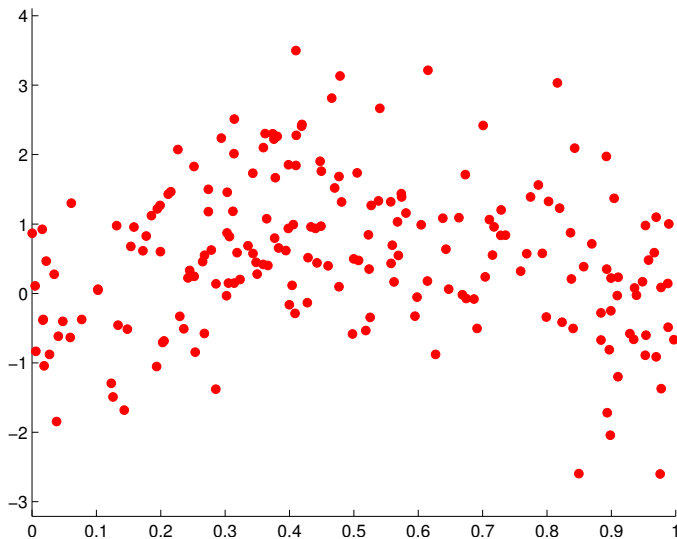
- Part I: Model selection for fixed-design regression
- Part II: Model/estimator selection in the prediction setting
- Part III: Cross-validation for estimator selection/aggregation

Part I

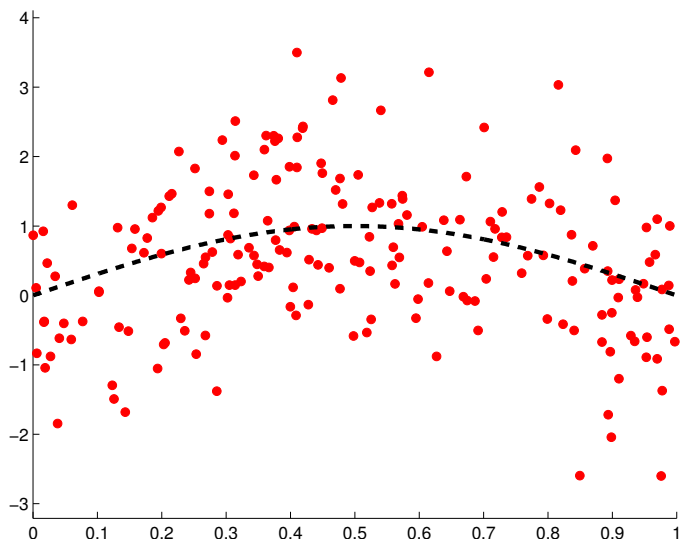
Model selection for fixed-design regression

Outline

- 1 Statistical framework
- 2 Model selection
- 3 Minimal penalties

Regression: data $(X_1, Y_1), \dots, (X_n, Y_n)$ 

Goal: find the signal (denoising)



Statistical framework: regression, least-squares risk

- Observations:

$$Y_i = f^*(x_i) + \varepsilon_i \in \mathbb{R}$$

with $\varepsilon_1, \dots, \varepsilon_n$ i.i.d. $\sim \mathcal{N}(0, \sigma^2)$, f^* and σ^2 unknown

- Fixed design: $x_i \in \mathcal{X}$ deterministic

- Notation: $Y = F + \varepsilon$ with

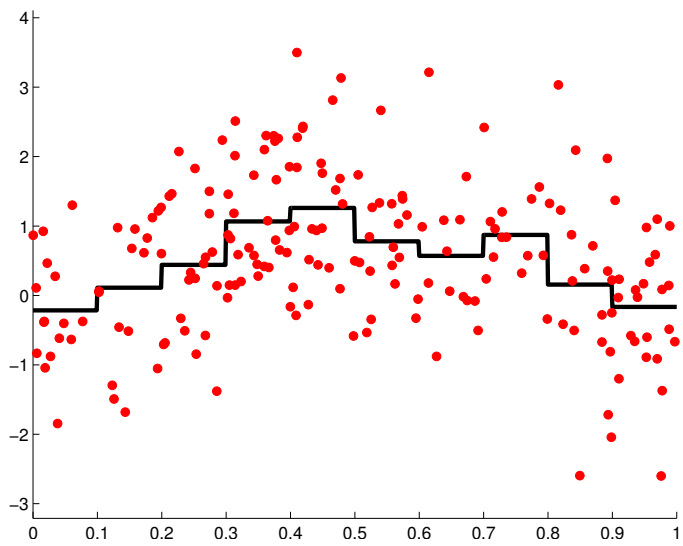
$$Y = (Y_i)_{1 \leq i \leq n}, \quad F = (f^*(x_i))_{1 \leq i \leq n}, \quad \varepsilon = (\varepsilon_i)_{1 \leq i \leq n} \in \mathbb{R}^n$$

- **Least-squares risk** of a predictor $t \in \mathbb{R}^n$ (" $t_i = t(x_i)$ "):

$$\frac{1}{n} \|t - F\|^2 = \frac{1}{n} \sum_{i=1}^n (t_i - F_i)^2$$

⇒ **Estimator** $\hat{F}(Y) \in \mathbb{R}^n$?

Estimators: example: regressogram



Least-squares estimators

- Natural idea: minimize an estimator of the risk $\frac{1}{n} \|t - F\|^2$

Least-squares estimators

- Natural idea: minimize an estimator of the risk $\frac{1}{n} \|t - F\|^2$
- **Least-squares criterion:**

$$\frac{1}{n} \|t - Y\|^2 = \frac{1}{n} \sum_{i=1}^n (t_i - Y_i)^2$$

$$\forall t \in \mathbb{R}^n, \quad \mathbb{E} \left[\frac{1}{n} \|t - Y\|^2 \right] = \frac{1}{n} \|t - F\|^2 + \frac{1}{n} \mathbb{E} \left[\|\varepsilon\|^2 \right]$$

Least-squares estimators

- Natural idea: minimize an estimator of the risk $\frac{1}{n} \|t - F\|^2$
- Least-squares criterion:

$$\frac{1}{n} \|t - Y\|^2 = \frac{1}{n} \sum_{i=1}^n (t_i - Y_i)^2$$

$$\forall t \in \mathbb{R}^n, \quad \mathbb{E} \left[\frac{1}{n} \|t - Y\|^2 \right] = \frac{1}{n} \|t - F\|^2 + \frac{1}{n} \mathbb{E} \left[\|\varepsilon\|^2 \right]$$

- Model: $S \subset \mathbb{R}^n \Rightarrow$ **Least-squares estimator** on S :

$$\hat{F}_S \in \operatorname{argmin}_{t \in S} \left\{ \frac{1}{n} \|t - Y\|^2 \right\} = \operatorname{argmin}_{t \in S} \left\{ \frac{1}{n} \sum_{i=1}^n (t_i - Y_i)^2 \right\}$$

so that

$$\hat{F}_S = \Pi_S(Y) \quad (\text{orthogonal projection})$$

Model examples

- **histograms** on some partition m of \mathcal{X}
 $\Rightarrow S_m = \text{vect}\{(\mathbf{1}_{x_i \in \lambda})_{1 \leq i \leq n} / \lambda \in m\}$
 \Rightarrow the least-squares estimator (regressogram) can be written

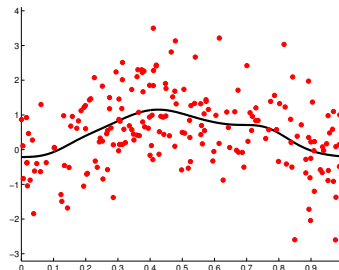
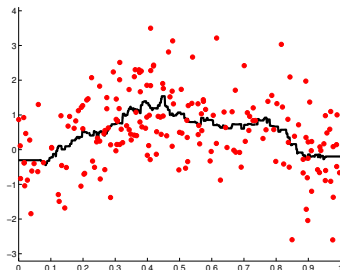
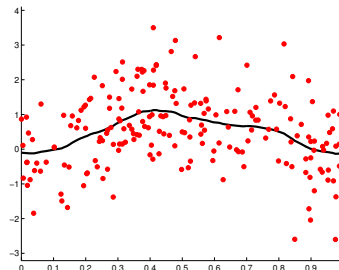
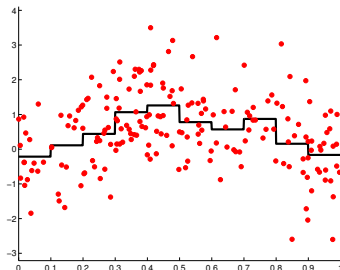
$$\widehat{F}_{S_m}(x_i) = \sum_{\lambda \in m} \widehat{\beta}_\lambda \mathbf{1}_{x_i \in \lambda} \quad \widehat{\beta}_\lambda = \frac{1}{\text{Card}\{x_i \in \lambda\}} \sum_{x_i \in \lambda} Y_i$$

- **variable selection**: $x_i = (x_i^{(1)}, \dots, x_i^{(p)}) \in \mathbb{R}^p$ gathers p variables that can (linearly) explain Y_i

$$\forall m \subset \{1, \dots, p\} \quad , \quad S_m = \text{vect}\{x^{(j)} \text{ s.t. } j \in m\}$$

- S_m subspace generated by a subset of an orthogonal basis of \mathbb{R}^n (**Fourier, wavelets, ...**)

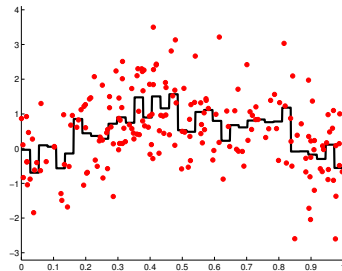
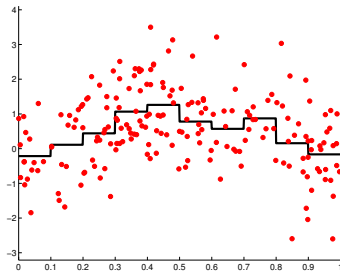
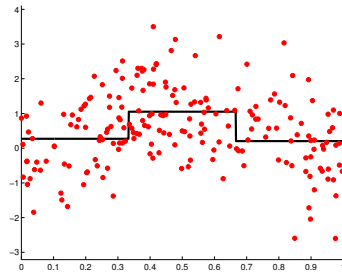
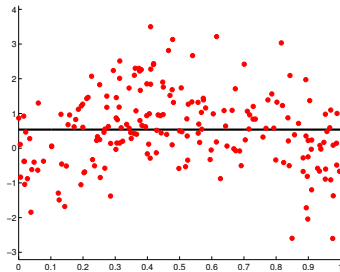
Estimators: regressogram, ridge, k -NN, Nadaraya-Watson



Outline

- 1 Statistical framework
- 2 **Model selection**
- 3 Minimal penalties

Model selection: regular regressograms, choose D ?



Model selection

- Model collection $(S_m)_{m \in \mathcal{M}} \Rightarrow (\hat{F}_m)_{m \in \mathcal{M}} \Rightarrow \hat{m}(Y)$?

$$\hat{F}_m = \Pi_m Y = \Pi_{S_m} Y$$

- Goal: minimize the risk, i.e.,
Oracle inequality (in expectation or with a large probability):

$$\frac{1}{n} \left\| \hat{F}_{\hat{m}} - F \right\|^2 \leq C \inf_{m \in \mathcal{M}} \left\{ \frac{1}{n} \left\| \hat{F}_m - F \right\|^2 \right\} + R_n$$

Approximation and estimation error

Fixed model S_m , linear subspace of \mathbb{R}^n , dimension $D_m = \dim(S_m)$.

- **Approximation error** of S_m :

$$\inf_{t \in S_m} \frac{1}{n} \|t - F\|^2 = \frac{1}{n} \|F_m - F\|^2 = \frac{1}{n} \|(I_n - \Pi_m)F\|^2$$

where $F_m = \Pi_m F$ orthogonal projection of F onto S_m

- $\hat{F}_m \in S_m \Rightarrow \frac{1}{n} \|\hat{F}_m - F\|^2 \geq \frac{1}{n} \|F_m - F\|^2$

Approximation and estimation error

Fixed model S_m , linear subspace of \mathbb{R}^n , dimension $D_m = \dim(S_m)$.

- **Approximation error** of S_m :

$$\inf_{t \in S_m} \frac{1}{n} \|t - F\|^2 = \frac{1}{n} \|F_m - F\|^2 = \frac{1}{n} \|(I_n - \Pi_m)F\|^2$$

where $F_m = \Pi_m F$ orthogonal projection of F onto S_m

- $\hat{F}_m \in S_m \Rightarrow \frac{1}{n} \|\hat{F}_m - F\|^2 \geq \frac{1}{n} \|F_m - F\|^2$

- **Estimation error** of S_m :

$$\frac{1}{n} \|\hat{F}_m - F\|^2 - \frac{1}{n} \|F_m - F\|^2 \geq 0$$

Expectation of the estimation error

$$\begin{aligned}\|\hat{F}_m - F\|^2 &= \|\Pi_m(F + \varepsilon) - F\|^2 \\ &= \|\Pi_m F - F\|^2 + 2 \underbrace{\langle \Pi_m F - F, \Pi_m \varepsilon \rangle}_{=0} + \|\Pi_m \varepsilon\|^2\end{aligned}$$

Expectation of the estimation error

$$\begin{aligned}\|\widehat{F}_m - F\|^2 &= \|\Pi_m(F + \varepsilon) - F\|^2 \\ &= \|\Pi_m F - F\|^2 + 2 \underbrace{\langle \Pi_m F - F, \Pi_m \varepsilon \rangle}_{=0} + \|\Pi_m \varepsilon\|^2\end{aligned}$$

⇒ **Estimation error** (of \widehat{F}_m):

$$\frac{1}{n} \|\widehat{F}_m - F\|^2 - \frac{1}{n} \|F_m - F\|^2 = \frac{1}{n} \|\Pi_m \varepsilon\|^2 = \frac{1}{n} \langle \Pi_m \varepsilon, \varepsilon \rangle$$

Expectation of the estimation error

$$\begin{aligned}\|\widehat{F}_m - F\|^2 &= \|\Pi_m(F + \varepsilon) - F\|^2 \\ &= \|\Pi_m F - F\|^2 + 2 \underbrace{\langle \Pi_m F - F, \Pi_m \varepsilon \rangle}_{=0} + \|\Pi_m \varepsilon\|^2\end{aligned}$$

⇒ **Estimation error** (of \widehat{F}_m):

$$\frac{1}{n} \|\widehat{F}_m - F\|^2 - \frac{1}{n} \|F_m - F\|^2 = \frac{1}{n} \|\Pi_m \varepsilon\|^2 = \frac{1}{n} \langle \Pi_m \varepsilon, \varepsilon \rangle$$

⇒ **Expectation** of the estimation error (of \widehat{F}_m):

$$\frac{1}{n} \mathbb{E}[\langle \Pi_m \varepsilon, \varepsilon \rangle] = \frac{\sigma^2 \operatorname{tr}(\Pi_m)}{n} = \frac{\sigma^2 D_m}{n}$$

Bias-variance trade-off

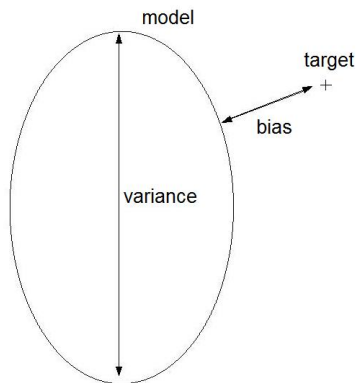
$$\mathbb{E} \left[\frac{1}{n} \left\| \widehat{F}_m - F \right\|^2 \right] = \frac{1}{n} \left\| F_m - F \right\|^2 + \frac{\sigma^2 D_m}{n}$$

Approximation error or **Bias**:

$$\frac{1}{n} \left\| F_m - F \right\|^2 = \frac{1}{n} \left\| \Pi_m F - F \right\|^2$$

\mathbb{E} [Estimation error] or **Variance**:

$$\frac{\sigma^2 D_m}{n}$$



Bias-variance trade-off

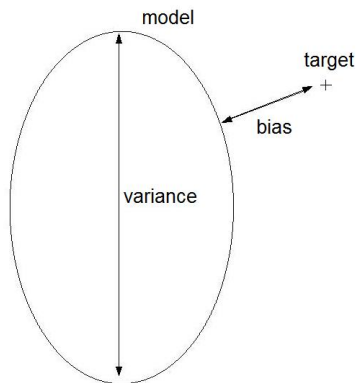
$$\mathbb{E} \left[\frac{1}{n} \left\| \widehat{F}_m - F \right\|^2 \right] = \frac{1}{n} \left\| F_m - F \right\|^2 + \frac{\sigma^2 D_m}{n}$$

Approximation error or Bias:

$$\frac{1}{n} \left\| F_m - F \right\|^2 = \frac{1}{n} \left\| \Pi_m F - F \right\|^2$$

\mathbb{E} [Estimation error] or Variance:

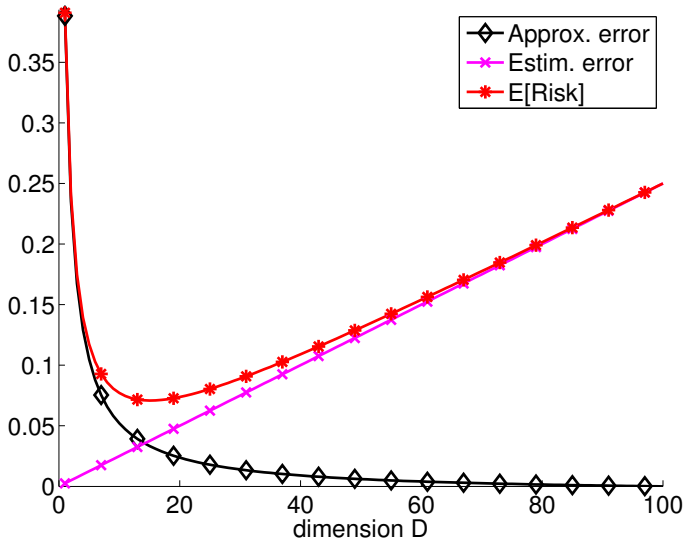
$$\frac{\sigma^2 D_m}{n}$$



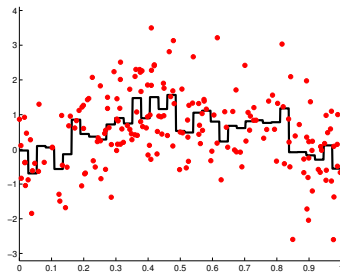
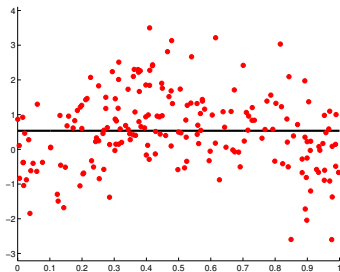
Bias-variance trade-off

⇔ avoid **overfitting** and **underfitting**

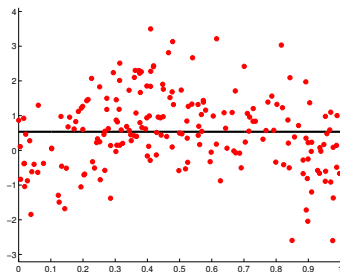
Bias-variance trade-off



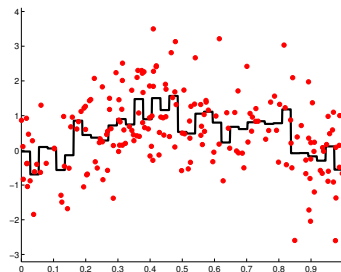
Overfitting/underfitting: regressograms



Overfitting/underfitting: regressograms

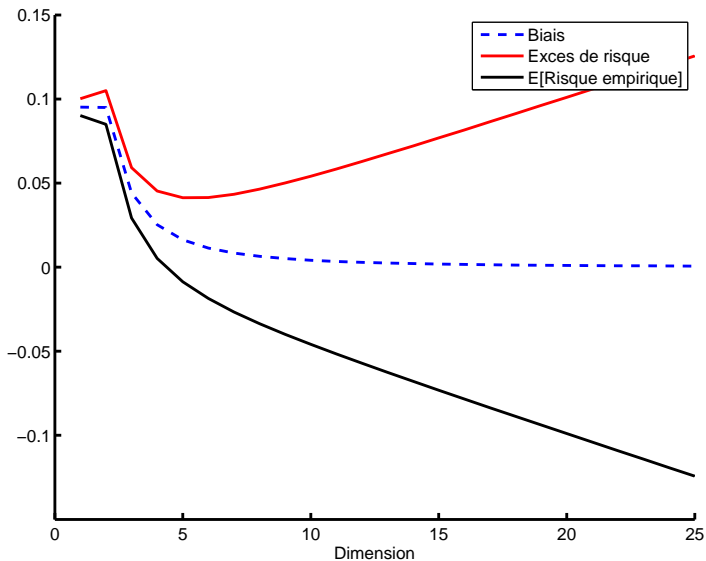


Underfitting
 $D = 1$ (too small)



Overfitting
 $D = 37$ (too large)

Why should the empirical risk be penalized?



Penalization

$$\hat{m} \in \operatorname{argmin}_{m \in \mathcal{M}} \left\{ \frac{1}{n} \|\hat{F}_m - Y\|^2 + \operatorname{pen}(m) \right\}$$

Penalization

$$\hat{m} \in \operatorname{argmin}_{m \in \mathcal{M}} \left\{ \frac{1}{n} \left\| \hat{F}_m - Y \right\|^2 + \operatorname{pen}(m) \right\}$$

- Ideal penalty:

$$\operatorname{pen}_{\text{id}}(m) := \frac{1}{n} \left\| \hat{F}_m - F \right\|^2 - \frac{1}{n} \left\| \hat{F}_m - Y \right\|^2 = \text{Risk} - \text{Empirical risk}$$

- **Mallows' heuristic:** $\operatorname{pen}(m) \approx \mathbb{E} [\operatorname{pen}_{\text{id}}(m)]$
⇒ oracle inequality if $\operatorname{Card}(\mathcal{M})$ not too large
(+ concentration inequalities)

Ideal penalty and its expectation

$$\begin{aligned}\|\widehat{F}_m - Y\|^2 &= \|\widehat{F}_m - F - \varepsilon\|^2 \\ &= \|\widehat{F}_m - F\|^2 + \|\varepsilon\|^2 - 2\langle \widehat{F}_m - F, \varepsilon \rangle \\ &= \|\widehat{F}_m - F\|^2 + \|\varepsilon\|^2 - 2\langle \Pi_m F + \Pi_m \varepsilon - F, \varepsilon \rangle \\ &= \|\widehat{F}_m - F\|^2 + \|\varepsilon\|^2 - 2\langle \Pi_m F - F, \varepsilon \rangle - 2\langle \Pi_m \varepsilon, \varepsilon \rangle\end{aligned}$$

Ideal penalty and its expectation

$$\begin{aligned}\|\widehat{F}_m - Y\|^2 &= \|\widehat{F}_m - F - \varepsilon\|^2 \\ &= \|\widehat{F}_m - F\|^2 + \|\varepsilon\|^2 - 2\langle \widehat{F}_m - F, \varepsilon \rangle \\ &= \|\widehat{F}_m - F\|^2 + \|\varepsilon\|^2 - 2\langle \Pi_m F + \Pi_m \varepsilon - F, \varepsilon \rangle \\ &= \|\widehat{F}_m - F\|^2 + \|\varepsilon\|^2 - 2\langle \Pi_m F - F, \varepsilon \rangle - 2\langle \Pi_m \varepsilon, \varepsilon \rangle\end{aligned}$$

⇒ Ideal penalty

$$\begin{aligned}\text{pen}_{\text{id}}(m) &= \frac{1}{n} \|\widehat{F}_m - F\|^2 - \frac{1}{n} \|\widehat{F}_m - Y\|^2 \\ &= \frac{2}{n} \langle \Pi_m F - F, \varepsilon \rangle + \frac{2}{n} \langle \Pi_m \varepsilon, \varepsilon \rangle - \frac{1}{n} \|\varepsilon\|^2\end{aligned}$$

Ideal penalty and its expectation

⇒ Ideal penalty

$$\begin{aligned} \text{pen}_{\text{id}}(m) &= \frac{1}{n} \left\| \widehat{F}_m - F \right\|^2 - \frac{1}{n} \left\| \widehat{F}_m - Y \right\|^2 \\ &= \frac{2}{n} \langle \Pi_m F - F, \varepsilon \rangle + \frac{2}{n} \langle \Pi_m \varepsilon, \varepsilon \rangle - \frac{1}{n} \|\varepsilon\|^2 \end{aligned}$$

⇒ Expectation of the ideal penalty

$$\begin{aligned} \mathbb{E}[\text{pen}_{\text{id}}(m)] &= \frac{2}{n} \mathbb{E}[\langle \Pi_m F - F, \varepsilon \rangle] + \frac{2}{n} \mathbb{E}[\langle \Pi_m \varepsilon, \varepsilon \rangle] - \frac{1}{n} \mathbb{E}[\|\varepsilon\|^2] \\ &= \underbrace{\frac{2\sigma^2 D_m}{n}}_{\text{Mallows' } C_p} - \sigma^2 \end{aligned}$$

Towards theoretical guarantees: a key lemma

Lemma

Let $\text{crit} : \mathcal{M} \rightarrow \mathbb{R}$ be any function (possibly data-dependent).
On the event Ω on which, $\forall m, m' \in \mathcal{M}$

$$\left[\text{crit}(m) - \frac{1}{n} \|\widehat{F}_m - F\|^2 \right] - \left[\text{crit}(m') - \frac{1}{n} \|\widehat{F}_{m'} - F\|^2 \right] \leq A(m) + B(m'),$$

we have $\forall \widehat{m} \in \arg \min_{m \in \mathcal{M}} \{ \text{crit}(m) \}$,

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

Proof of the key lemma

For any $m \in \mathcal{M}$, we have

$$\begin{aligned} & \frac{1}{n} \left\| \widehat{F}_{\widehat{m}} - F \right\|^2 \\ &= \text{crit}(\widehat{m}) + \frac{1}{n} \left\| \widehat{F}_{\widehat{m}} - F \right\|^2 - \text{crit}(\widehat{m}) \end{aligned}$$

Proof of the key lemma

For any $m \in \mathcal{M}$, we have

$$\begin{aligned} & \frac{1}{n} \left\| \widehat{F}_{\widehat{m}} - F \right\|^2 \\ &= \text{crit}(\widehat{m}) + \frac{1}{n} \left\| \widehat{F}_{\widehat{m}} - F \right\|^2 - \text{crit}(\widehat{m}) \\ &\leq \text{crit}(m) + \frac{1}{n} \left\| \widehat{F}_{\widehat{m}} - F \right\|^2 - \text{crit}(\widehat{m}) \end{aligned}$$

Proof of the key lemma

For any $m \in \mathcal{M}$, we have

$$\begin{aligned} & \frac{1}{n} \left\| \widehat{F}_{\widehat{m}} - F \right\|^2 \\ &= \text{crit}(\widehat{m}) + \frac{1}{n} \left\| \widehat{F}_{\widehat{m}} - F \right\|^2 - \text{crit}(\widehat{m}) \\ &\leq \text{crit}(m) + \frac{1}{n} \left\| \widehat{F}_{\widehat{m}} - F \right\|^2 - \text{crit}(\widehat{m}) \\ &= \frac{1}{n} \left\| \widehat{F}_m - F \right\|^2 + \text{crit}(m) - \frac{1}{n} \left\| \widehat{F}_m - F \right\|^2 + \frac{1}{n} \left\| \widehat{F}_{\widehat{m}} - F \right\|^2 - \text{crit}(\widehat{m}) \end{aligned}$$

Proof of the key lemma

For any $m \in \mathcal{M}$, we have

$$\begin{aligned} & \frac{1}{n} \left\| \widehat{F}_{\widehat{m}} - F \right\|^2 \\ &= \text{crit}(\widehat{m}) + \frac{1}{n} \left\| \widehat{F}_{\widehat{m}} - F \right\|^2 - \text{crit}(\widehat{m}) \\ &\leq \text{crit}(m) + \frac{1}{n} \left\| \widehat{F}_{\widehat{m}} - F \right\|^2 - \text{crit}(\widehat{m}) \\ &= \frac{1}{n} \left\| \widehat{F}_m - F \right\|^2 + \text{crit}(m) - \frac{1}{n} \left\| \widehat{F}_m - F \right\|^2 + \frac{1}{n} \left\| \widehat{F}_{\widehat{m}} - F \right\|^2 - \text{crit}(\widehat{m}) \\ &\leq \frac{1}{n} \left\| \widehat{F}_m - F \right\|^2 + A(m) + B(\widehat{m}) \end{aligned}$$

Proof of the key lemma

For any $m \in \mathcal{M}$, we have

$$\begin{aligned} & \frac{1}{n} \left\| \widehat{F}_{\widehat{m}} - F \right\|^2 \\ &= \text{crit}(\widehat{m}) + \frac{1}{n} \left\| \widehat{F}_{\widehat{m}} - F \right\|^2 - \text{crit}(\widehat{m}) \\ &\leq \text{crit}(m) + \frac{1}{n} \left\| \widehat{F}_{\widehat{m}} - F \right\|^2 - \text{crit}(\widehat{m}) \\ &= \frac{1}{n} \left\| \widehat{F}_m - F \right\|^2 + \text{crit}(m) - \frac{1}{n} \left\| \widehat{F}_m - F \right\|^2 + \frac{1}{n} \left\| \widehat{F}_{\widehat{m}} - F \right\|^2 - \text{crit}(\widehat{m}) \\ &\leq \frac{1}{n} \left\| \widehat{F}_m - F \right\|^2 + A(m) + B(\widehat{m}) \end{aligned}$$

hence

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

Key lemma (reformulated)

Lemma

Let $\text{pen} : \mathcal{M} \rightarrow \mathbb{R}$ be any penalty (possibly data-dependent).
On the event Ω on which, $\forall m, m' \in \mathcal{M}$

$$\begin{aligned} & [\text{pen}(m) - \text{pen}_{\text{id}}(m)] - [\text{pen}(m') - \text{pen}_{\text{id}}(m')] \\ & \leq A(m) + B(m'), \end{aligned}$$

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

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

Proof: take $\text{crit}(m) = \frac{1}{n} \left\| \hat{F}_m - Y \right\|^2 + \text{pen}(m)$. □

How to use the key lemma here?

- 1 Take $\text{pen}(m) = \mathbb{E}[\text{pen}_{\text{id}}(m)]$ (up to a translation).
- 2 Prove that $\forall m \in \mathcal{M}$, $\text{pen}_{\text{id}}(m)$ concentrates around its expectation.
- 3 Union bound over $m \in \mathcal{M}$ (if $\text{Card}(\mathcal{M})$ “small”).
- 4 Apply the key lemma with $A(m) \propto$ deviations of $\text{pen}_{\text{id}}(m)$.

Ideal penalty (reminder)

$$\begin{aligned}
 \text{pen}_{\text{id}}(m) &= \frac{1}{n} \left\| \widehat{F}_m - F \right\|^2 - \frac{1}{n} \left\| \widehat{F}_m - Y \right\|^2 \\
 &= \frac{2}{n} \langle \Pi_m F - F, \varepsilon \rangle + \frac{2}{n} \langle \Pi_m \varepsilon, \varepsilon \rangle - \frac{1}{n} \|\varepsilon\|^2
 \end{aligned}$$

Ideal penalty (reminder)

$$\begin{aligned}\text{pen}_{\text{id}}(m) &= \frac{1}{n} \left\| \widehat{F}_m - F \right\|^2 - \frac{1}{n} \left\| \widehat{F}_m - Y \right\|^2 \\ &= \frac{2}{n} \langle \Pi_m F - F, \varepsilon \rangle + \frac{2}{n} \langle \Pi_m \varepsilon, \varepsilon \rangle - \frac{1}{n} \|\varepsilon\|^2\end{aligned}$$

- Linear term:

$$\frac{2}{n} \langle \Pi_m F - F, \varepsilon \rangle \quad \Rightarrow \quad \text{mean} = 0, \quad \text{Gaussian distribution}$$

Ideal penalty (reminder)

$$\begin{aligned}\text{pen}_{\text{id}}(m) &= \frac{1}{n} \left\| \widehat{F}_m - F \right\|^2 - \frac{1}{n} \left\| \widehat{F}_m - Y \right\|^2 \\ &= \frac{2}{n} \langle \Pi_m F - F, \varepsilon \rangle + \frac{2}{n} \langle \Pi_m \varepsilon, \varepsilon \rangle - \frac{1}{n} \|\varepsilon\|^2\end{aligned}$$

- Linear term:

$$\frac{2}{n} \langle \Pi_m F - F, \varepsilon \rangle \quad \Rightarrow \quad \text{mean} = 0, \quad \text{Gaussian distribution}$$

- Quadratic term:

$$\frac{2}{n} \langle \Pi_m \varepsilon, \varepsilon \rangle \quad \Rightarrow \quad \text{mean} = \frac{2\sigma^2 D_m}{n}, \quad \chi^2 \text{ distribution}$$

Ideal penalty (reminder)

$$\begin{aligned}\text{pen}_{\text{id}}(m) &= \frac{1}{n} \left\| \widehat{F}_m - F \right\|^2 - \frac{1}{n} \left\| \widehat{F}_m - Y \right\|^2 \\ &= \frac{2}{n} \langle \Pi_m F - F, \varepsilon \rangle + \frac{2}{n} \langle \Pi_m \varepsilon, \varepsilon \rangle - \frac{1}{n} \|\varepsilon\|^2\end{aligned}$$

- Linear term:

$$\frac{2}{n} \langle \Pi_m F - F, \varepsilon \rangle \quad \Rightarrow \quad \text{mean} = 0, \quad \text{Gaussian distribution}$$

- Quadratic term:

$$\frac{2}{n} \langle \Pi_m \varepsilon, \varepsilon \rangle \quad \Rightarrow \quad \text{mean} = \frac{2\sigma^2 D_m}{n}, \quad \chi^2 \text{ distribution}$$

- Constant term:

$$\frac{1}{n} \|\varepsilon\|^2 \quad \Rightarrow \quad \text{can be discarded}$$

Concentration of the ideal penalty (1): linear term

Proposition (Gaussian concentration)

If $\varepsilon \sim \mathcal{N}(0, \sigma^2 I_n)$ and $\alpha \in \mathbb{R}^n$, for every $x \geq 0$,

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

Concentration of the ideal penalty (1): linear term

Proposition (Gaussian concentration)

If $\varepsilon \sim \mathcal{N}(0, \sigma^2 I_n)$ and $\alpha \in \mathbb{R}^n$, for every $x \geq 0$,

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

⇒ with probability $\geq 1 - 2e^{-x}$, for every $\theta > 0$,

$$\frac{2}{n} |\langle \Pi_m F - F, \varepsilon \rangle| \leq \frac{2\sigma\sqrt{2x}}{n} \|\Pi_m F - F\| \leq \frac{\theta}{n} \|\Pi_m F - F\|^2 + \frac{2x\sigma^2}{\theta n}$$

since $2ab \leq \theta a^2 + \theta^{-1} b^2 \quad \forall a, b \geq 0, \theta > 0$.

Concentration of the ideal penalty (1): linear term

Proposition (Gaussian concentration)

If $\varepsilon \sim \mathcal{N}(0, \sigma^2 I_n)$ and $\alpha \in \mathbb{R}^n$, for every $x \geq 0$,

$$\mathbb{P}(|\langle \varepsilon, \alpha \rangle| \leq \sigma \sqrt{2x} \|\alpha\|) \geq 1 - 2e^{-x}.$$

⇒ with probability $\geq 1 - 2e^{-x}$, for every $\theta > 0$,

$$\frac{2}{n} |\langle \Pi_m F - F, \varepsilon \rangle| \leq \frac{2\sigma\sqrt{2x}}{n} \|\Pi_m F - F\| \leq \frac{\theta}{n} \|\Pi_m F - F\|^2 + \frac{2x\sigma^2}{\theta n}$$

since $2ab \leq \theta a^2 + \theta^{-1} b^2 \forall a, b \geq 0, \theta > 0$.

- Can be generalized to **sub-Gaussian noise**, see arXiv:1901.07277 (Remark 1).

Concentration of the ideal penalty (2): quadratic term

Proposition (see A. & Bach 2011 (arXiv:0909.1884), Proposition 6)

If $\varepsilon \sim \mathcal{N}(0, \sigma^2 I_n)$ and $M \in \mathcal{M}_n(\mathbb{R})$, for every $x \geq 0$,

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

Concentration of the ideal penalty (2): quadratic term

Proposition (see A. & Bach 2011 (arXiv:0909.1884), Proposition 6)

If $\varepsilon \sim \mathcal{N}(0, \sigma^2 I_n)$ and $M \in \mathcal{M}_n(\mathbb{R})$, for every $x \geq 0$,

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

\Rightarrow with probability $\geq 1 - 2e^{-x}$, for every $\theta > 0$,

$$\frac{2}{n} \left| \langle \Pi_m \varepsilon, \varepsilon \rangle - \sigma^2 D_m \right| \leq \frac{4\sigma^2}{n} \sqrt{x D_m} + \frac{4\sigma^2 x}{n}$$

Concentration of the ideal penalty (2): quadratic term

Proposition (see A. & Bach 2011 (arXiv:0909.1884), Proposition 6)

If $\varepsilon \sim \mathcal{N}(0, \sigma^2 I_n)$ and $M \in \mathcal{M}_n(\mathbb{R})$, for every $x \geq 0$,

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

⇒ with probability $\geq 1 - 2e^{-x}$, for every $\theta > 0$,

$$\begin{aligned} \frac{2}{n} \left| \langle \Pi_m \varepsilon, \varepsilon \rangle - \sigma^2 D_m \right| &\leq \frac{4\sigma^2}{n} \sqrt{x D_m} + \frac{4\sigma^2 x}{n} \\ &\leq \frac{\theta \sigma^2 D_m}{n} + \left(4 + \frac{4}{\theta} \right) \frac{\sigma^2 x}{n}. \end{aligned}$$

Concentration of the ideal penalty (2): quadratic term

Proposition (see A. & Bach 2011 (arXiv:0909.1884), Proposition 6)

If $\varepsilon \sim \mathcal{N}(0, \sigma^2 I_n)$ and $M \in \mathcal{M}_n(\mathbb{R})$, for every $x \geq 0$,

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

⇒ with probability $\geq 1 - 2e^{-x}$, for every $\theta > 0$,

$$\begin{aligned} \frac{2}{n} \left| \langle \Pi_m \varepsilon, \varepsilon \rangle - \sigma^2 D_m \right| &\leq \frac{4\sigma^2}{n} \sqrt{x D_m} + \frac{4\sigma^2 x}{n} \\ &\leq \frac{\theta \sigma^2 D_m}{n} + \left(4 + \frac{4}{\theta} \right) \frac{\sigma^2 x}{n}. \end{aligned}$$

- Can be generalized to **sub-Gaussian noise**, see arXiv:1901.07277 (Remark 1).

Concentration of the ideal penalty: summary

- With probability $\geq 1 - 4e^{-x}$, for every $\theta > 0$,

$$\begin{aligned} & \left| \text{pen}_{\text{id}}(m) + \frac{1}{n} \|\varepsilon\|^2 - \mathbb{E} \left[\text{pen}_{\text{id}}(m) + \frac{1}{n} \|\varepsilon\|^2 \right] \right| \\ & \leq \frac{\theta}{n} \|\Pi_m F - F\|^2 + \frac{2x\sigma^2}{\theta n} + \frac{\theta\sigma^2 D_m}{n} + \left(4 + \frac{4}{\theta}\right) \frac{\sigma^2 x}{n} \end{aligned}$$

Concentration of the ideal penalty: summary

- With probability $\geq 1 - 4e^{-x}$, for every $\theta > 0$,

$$\begin{aligned}
 & \left| \text{pen}_{\text{id}}(m) + \frac{1}{n} \|\varepsilon\|^2 - \mathbb{E} \left[\text{pen}_{\text{id}}(m) + \frac{1}{n} \|\varepsilon\|^2 \right] \right| \\
 & \leq \frac{\theta}{n} \|\Pi_m F - F\|^2 + \frac{2x\sigma^2}{\theta n} + \frac{\theta\sigma^2 D_m}{n} + \left(4 + \frac{4}{\theta}\right) \frac{\sigma^2 x}{n} \\
 & = \theta \mathbb{E} \left[\frac{1}{n} \|\hat{F}_m - F\|^2 \right] + \left(6 + \frac{4}{\theta}\right) \frac{x\sigma^2}{n}
 \end{aligned}$$

Concentration of the ideal penalty: summary

- With probability $\geq 1 - 4e^{-x}$, for every $\theta > 0$,

$$\begin{aligned}
 & \left| \text{pen}_{\text{id}}(m) + \frac{1}{n} \|\varepsilon\|^2 - \mathbb{E} \left[\text{pen}_{\text{id}}(m) + \frac{1}{n} \|\varepsilon\|^2 \right] \right| \\
 & \leq \frac{\theta}{n} \|\Pi_m F - F\|^2 + \frac{2x\sigma^2}{\theta n} + \frac{\theta\sigma^2 D_m}{n} + \left(4 + \frac{4}{\theta}\right) \frac{\sigma^2 x}{n} \\
 & = \theta \mathbb{E} \left[\frac{1}{n} \|\hat{F}_m - F\|^2 \right] + \left(6 + \frac{4}{\theta}\right) \frac{x\sigma^2}{n} \\
 & \leq \frac{\theta}{1 - \theta} \frac{1}{n} \|\hat{F}_m - F\|^2 + L(\theta) \frac{x\sigma^2}{n} =: A(m) = B(m).
 \end{aligned}$$

Concentration of the ideal penalty: summary

- With probability $\geq 1 - 4e^{-x}$, for every $\theta > 0$,

$$\begin{aligned} & \left| \text{pen}_{\text{id}}(m) + \frac{1}{n} \|\varepsilon\|^2 - \mathbb{E} \left[\text{pen}_{\text{id}}(m) + \frac{1}{n} \|\varepsilon\|^2 \right] \right| \\ & \leq \frac{\theta}{n} \|\Pi_m F - F\|^2 + \frac{2x\sigma^2}{\theta n} + \frac{\theta\sigma^2 D_m}{n} + \left(4 + \frac{4}{\theta}\right) \frac{\sigma^2 x}{n} \\ & = \theta \mathbb{E} \left[\frac{1}{n} \|\widehat{F}_m - F\|^2 \right] + \left(6 + \frac{4}{\theta}\right) \frac{x\sigma^2}{n} \\ & \leq \frac{\theta}{1-\theta} \frac{1}{n} \|\widehat{F}_m - F\|^2 + L(\theta) \frac{x\sigma^2}{n} =: A(m) = B(m). \end{aligned}$$

⇒ end of step 2 (concentration of $\text{pen}_{\text{id}}(m)$)

Concentration of the ideal penalty: summary

- With probability $\geq 1 - 4e^{-x}$, for every $\theta > 0$,

$$\begin{aligned} & \left| \text{pen}_{\text{id}}(m) + \frac{1}{n} \|\varepsilon\|^2 - \mathbb{E} \left[\text{pen}_{\text{id}}(m) + \frac{1}{n} \|\varepsilon\|^2 \right] \right| \\ & \leq \frac{\theta}{n} \|\Pi_m F - F\|^2 + \frac{2x\sigma^2}{\theta n} + \frac{\theta\sigma^2 D_m}{n} + \left(4 + \frac{4}{\theta}\right) \frac{\sigma^2 x}{n} \\ & = \theta \mathbb{E} \left[\frac{1}{n} \|\widehat{F}_m - F\|^2 \right] + \left(6 + \frac{4}{\theta}\right) \frac{x\sigma^2}{n} \\ & \leq \frac{\theta}{1-\theta} \frac{1}{n} \|\widehat{F}_m - F\|^2 + L(\theta) \frac{x\sigma^2}{n} =: A(m) = B(m). \end{aligned}$$

- Step 3: **union bound** \Rightarrow with probability $\geq 1 - 4 \text{Card}(\mathcal{M})e^{-x}$, $\forall \theta > 0$, $\forall m, m' \in \mathcal{M}$,

$$\begin{aligned} & \left[\frac{2\sigma^2 D_m}{n} - \text{pen}_{\text{id}}(m) \right] - \left[\frac{2\sigma^2 D_{m'}}{n} - \text{pen}_{\text{id}}(m') \right] \\ & \leq A(m) + B(m'). \end{aligned}$$

Application of the key lemma (step 4)

- With probability $\geq 1 - 4 \text{Card}(\mathcal{M})e^{-x}$, $\forall \theta > 0$,

$$\forall \hat{m} \in \operatorname{argmin}_{m \in \mathcal{M}} \left\{ \frac{1}{n} \|\hat{F}_m - Y\|^2 + \frac{2\sigma^2 D_m}{n} \right\},$$

$$\begin{aligned} & (1 - \theta) \mathbb{E} \left[\frac{1}{n} \|\hat{F}_{\hat{m}} - F\|^2 \right] - \left(6 + \frac{4}{\theta} \right) \frac{x\sigma^2}{\theta n} \\ & \leq (1 + \theta) \inf_{m \in \mathcal{M}} \mathbb{E} \left[\frac{1}{n} \|\hat{F}_m - F\|^2 \right] + \left(6 + \frac{4}{\theta} \right) \frac{x\sigma^2}{\theta n}, \end{aligned}$$

Application of the key lemma (step 4)

- With probability $\geq 1 - 4 \text{Card}(\mathcal{M})e^{-x}$, $\forall \theta > 0$,

$$\forall \hat{m} \in \operatorname{argmin}_{m \in \mathcal{M}} \left\{ \frac{1}{n} \|\hat{F}_m - Y\|^2 + \frac{2\sigma^2 D_m}{n} \right\},$$

$$\begin{aligned} & (1 - \theta) \mathbb{E} \left[\frac{1}{n} \|\hat{F}_{\hat{m}} - F\|^2 \right] \\ & \leq (1 + \theta) \inf_{m \in \mathcal{M}} \mathbb{E} \left[\frac{1}{n} \|\hat{F}_m - F\|^2 \right] + 2 \left(6 + \frac{4}{\theta} \right) \frac{x\sigma^2}{\theta n}, \end{aligned}$$

Application of the key lemma (step 4)

- With probability $\geq 1 - 4 \text{Card}(\mathcal{M})e^{-x}$, $\forall \theta \in (0, 1)$,

$$\forall \hat{m} \in \operatorname{argmin}_{m \in \mathcal{M}} \left\{ \frac{1}{n} \|\hat{F}_m - Y\|^2 + \frac{2\sigma^2 D_m}{n} \right\},$$

$$\begin{aligned} & \mathbb{E} \left[\frac{1}{n} \|\hat{F}_{\hat{m}} - F\|^2 \right] \\ & \leq \frac{1 + \theta}{1 - \theta} \inf_{m \in \mathcal{M}} \mathbb{E} \left[\frac{1}{n} \|\hat{F}_m - F\|^2 \right] + \frac{2}{1 - \theta} \left(6 + \frac{4}{\theta} \right) \frac{x\sigma^2}{\theta n}. \end{aligned}$$

Conclusion: oracle inequality for C_p

Theorem (Birgé & Massart 2007, reformulated = Theorem 1 in arXiv:1901.07277)

Assumptions: $\text{pen}(m) = \frac{2\sigma^2 D_m}{n}$, *i.i.d. Gaussian noise.*

Then, for every $x \geq 0$, with probability at least $1 - 4 \text{Card}(\mathcal{M})e^{-x}$, for every $\theta \in (0, 1/3)$,

$$\frac{1}{n} \left\| \widehat{F}_{\widehat{m}} - F \right\|^2 \leq (1 + 3\theta) \underbrace{\inf_{m \in \mathcal{M}} \left\{ \frac{1}{n} \left\| \widehat{F}_m - F \right\|^2 \right\}}_{\text{oracle risk}} + \frac{L(\theta)\sigma^2 x}{n}.$$

Conclusion: oracle inequality for C_p

Theorem (Birgé & Massart 2007, reformulated = Theorem 1 in arXiv:1901.07277)

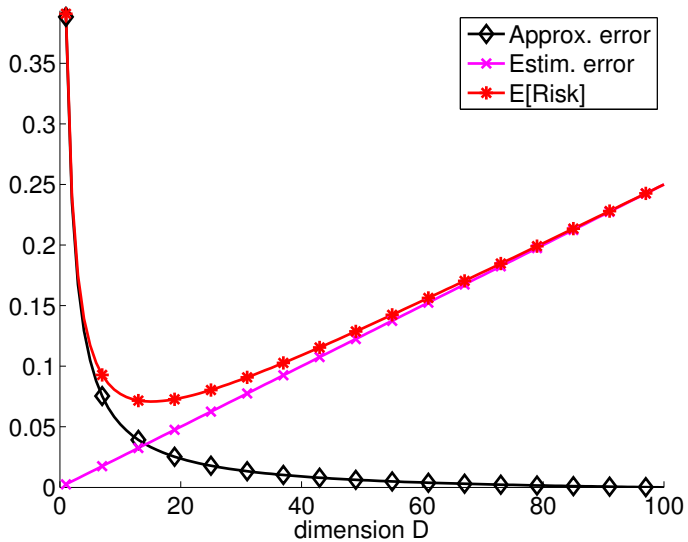
Assumptions: $\text{pen}(m) = \frac{2\sigma^2 D_m}{n}$, i.i.d. Gaussian noise.

Then, for every $x \geq 0$, with probability at least $1 - 4 \text{Card}(\mathcal{M})e^{-x}$, for every $\theta \in (0, 1/3)$,

$$\frac{1}{n} \left\| \widehat{F}_{\widehat{m}} - F \right\|^2 \leq (1 + 3\theta) \underbrace{\inf_{m \in \mathcal{M}} \left\{ \frac{1}{n} \left\| \widehat{F}_m - F \right\|^2 \right\}}_{\text{oracle risk}} + \frac{L(\theta)\sigma^2 x}{n}.$$

Generalization (arXiv:1901.07277): sub-Gaussian noise.

Meaning of an oracle inequality



Outline

- 1 Statistical framework
- 2 Model selection
- 3 Minimal penalties

Practical and theoretical questions about C_p

Mallows' C_p penalty:

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

- σ^2 unknown in general

$$\Rightarrow \text{pen}(m) = \frac{CD_m}{n}$$

with C (hopefully) close to $2\sigma^2$.

Practical and theoretical questions about C_p

Mallows' C_p penalty:

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

- σ^2 unknown in general

$$\Rightarrow \text{pen}(m) = \frac{CD_m}{n}$$

with C (hopefully) close to $2\sigma^2$.

- Oracle inequality for any $C > 0$? Minimal value for C ?

Practical and theoretical questions about C_p

Mallows' C_p penalty:

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

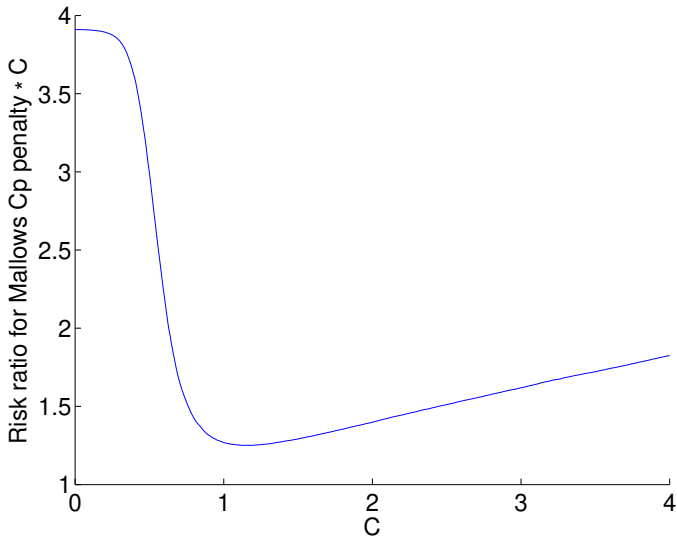
- σ^2 unknown in general

$$\Rightarrow \text{pen}(m) = \frac{CD_m}{n}$$

with C (hopefully) close to $2\sigma^2$.

- Oracle inequality for any $C > 0$? Minimal value for C ?
- How to estimate σ^2 ?

Performance of the penalty $K \times \sigma^2 D_m/n, K > 0$



Oracle inequality for the penalty CD_m/n

Theorem (Birgé & Massart 2007, reformulated = Theorem 1 in arXiv:1901.07277)

Assumptions: $\text{pen}(m) = \frac{CD_m}{n}$, $C > \sigma^2$, i.i.d. Gaussian noise.

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

$$\frac{1}{n} \|\hat{F}_{\hat{m}} - F\|^2 \leq L_1 \left(\frac{C}{\sigma^2} \right) \inf_{m \in \mathcal{M}} \left\{ \frac{1}{n} \|\hat{F}_m - F\|^2 \right\} + \frac{20\sigma^2 x}{n}.$$

Proof: same arguments as when $C = 2\sigma^2$.

Generalization to sub-Gaussian noise.

Oracle inequality for the penalty CD_m/n

Theorem (Birgé & Massart 2007, reformulated = Theorem 1 in arXiv:1901.07277)

Assumptions: $\text{pen}(m) = \frac{CD_m}{n}$, $C > \sigma^2$, i.i.d. Gaussian noise.

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

$$\frac{1}{n} \|\hat{F}_{\hat{m}} - F\|^2 \leq L_1 \left(\frac{C}{\sigma^2} \right) \inf_{m \in \mathcal{M}} \left\{ \frac{1}{n} \|\hat{F}_m - F\|^2 \right\} + \frac{20\sigma^2 x}{n}.$$

Proof: same arguments as when $C = 2\sigma^2$.

Generalization to sub-Gaussian noise.

Remark: $L_1(h) \xrightarrow{h \rightarrow 1^+} +\infty$, $L_1(h) \xrightarrow{h \rightarrow +\infty} +\infty$.

\Rightarrow overfitting when $C < \sigma^2$?

Expectation of the penalized criterion

$$\|\widehat{F}_m - Y\|^2 = \|\widehat{F}_m - F\|^2 + \|\varepsilon\|^2 - 2\langle \Pi_m F - F, \varepsilon \rangle - 2\langle \Pi_m \varepsilon, \varepsilon \rangle$$

implies that $\forall C \in \mathbb{R}$,

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

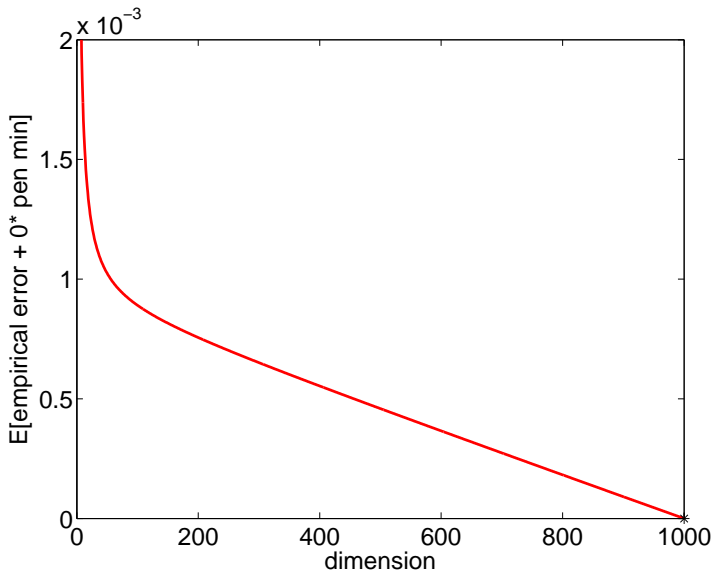
Expectation of the penalized criterion

$$\|\widehat{F}_m - Y\|^2 = \|\widehat{F}_m - F\|^2 + \|\varepsilon\|^2 - 2 \langle \Pi_m F - F, \varepsilon \rangle - 2 \langle \Pi_m \varepsilon, \varepsilon \rangle$$

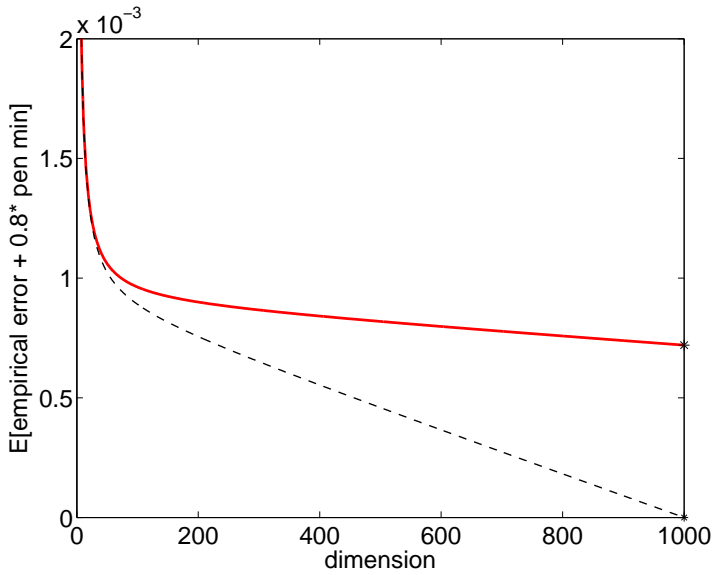
implies that $\forall C \in \mathbb{R}$,

$$\begin{aligned} \mathbb{E} \left[\|\widehat{F}_m - Y\|^2 + \frac{CD_m}{n} \right] &= \mathbb{E} \left[\|\widehat{F}_m - F\|^2 \right] + \sigma^2 + \frac{(C - 2\sigma^2)D_m}{n} \\ &= \mathbb{E} \left[\|F_m - F\|^2 \right] + \frac{(C - \sigma^2)D_m}{n} + \sigma^2. \end{aligned}$$

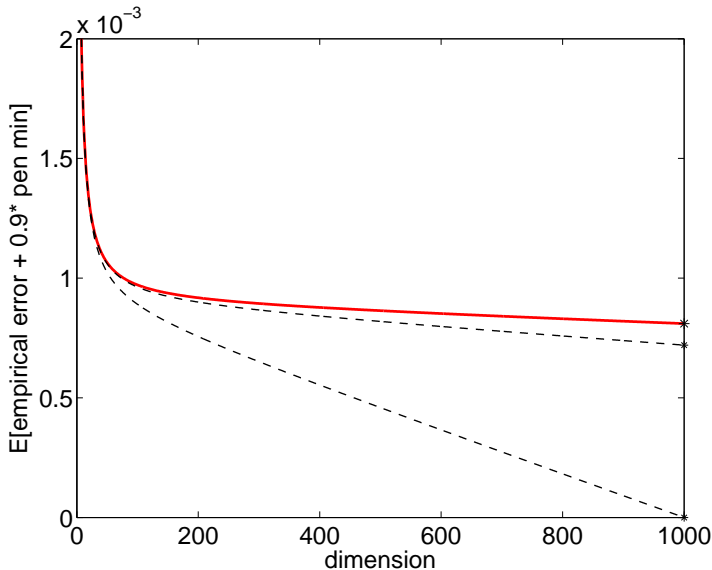
$$\mathbb{E}[\text{Empirical risk}] + 0 \times \sigma^2 D_m n^{-1} \text{ (OLS)}$$



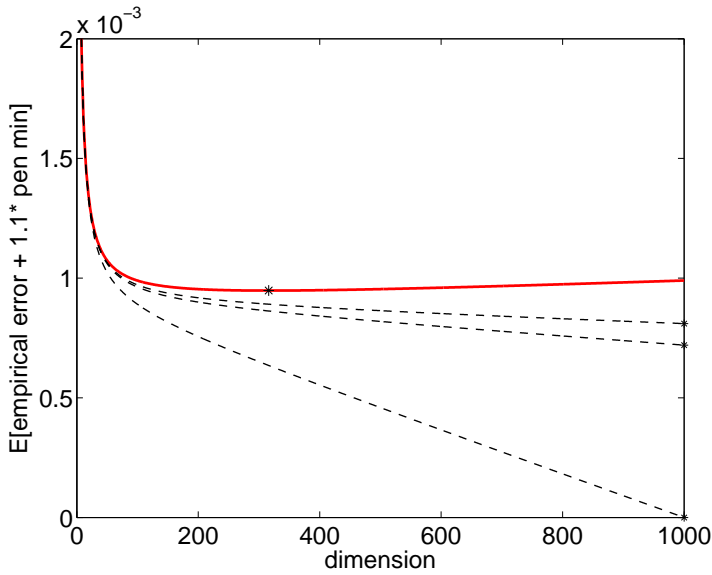
$$\mathbb{E}[\text{ Empirical risk }] + 0.8 \times \sigma^2 D_m n^{-1} \text{ (OLS)}$$



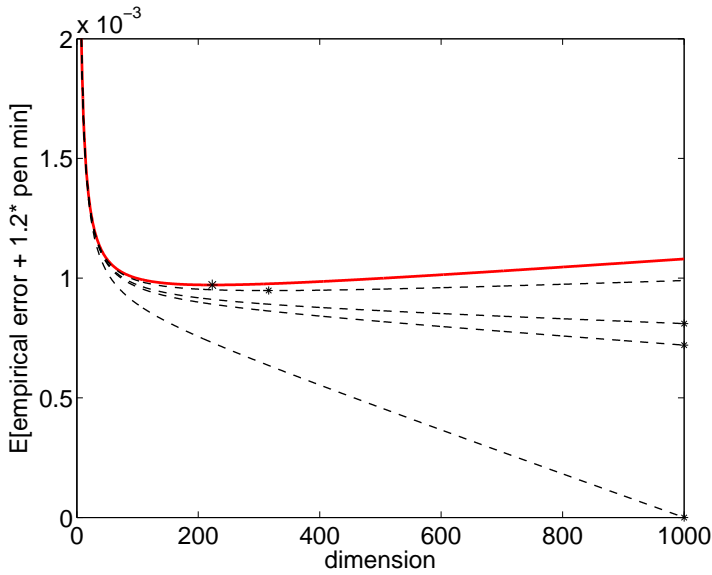
$$\mathbb{E}[\text{Empirical risk}] + 0.9 \times \sigma^2 D_m n^{-1} \text{ (OLS)}$$



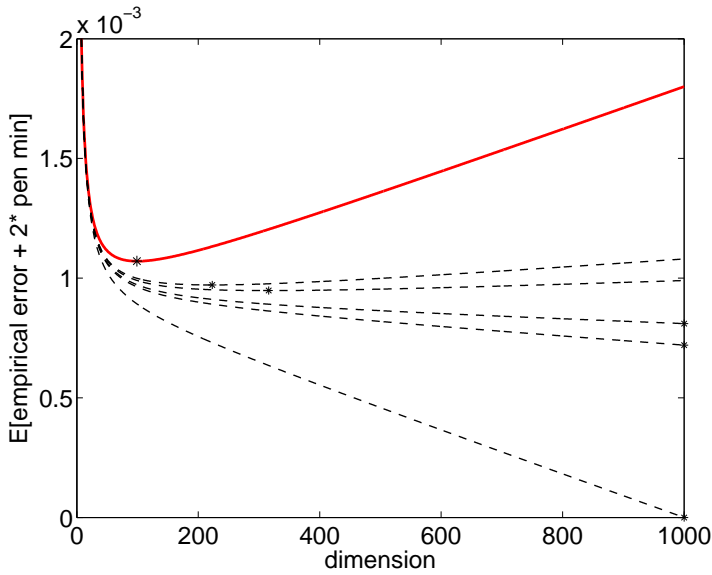
$$\mathbb{E}[\text{Empirical risk}] + 1.1 \times \sigma^2 D_m n^{-1} \text{ (OLS)}$$



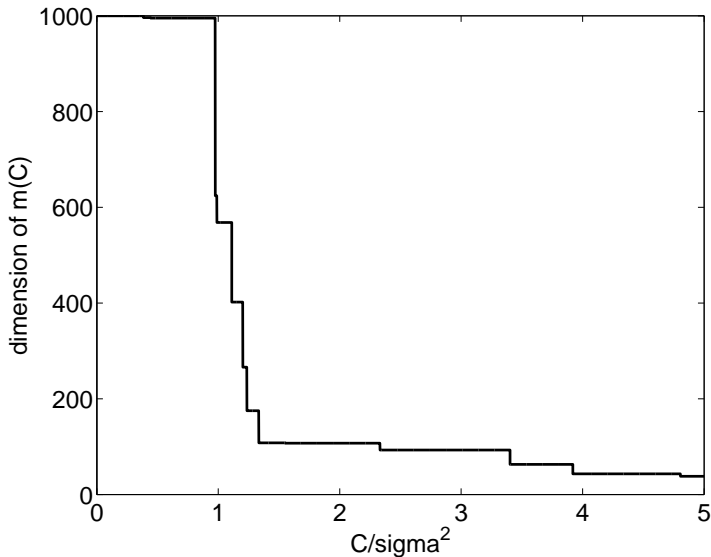
$$\mathbb{E}[\text{Empirical risk}] + 1.2 \times \sigma^2 D_m n^{-1} \text{ (OLS)}$$



$\mathbb{E}[\text{Empirical risk}] + 2 \times \sigma^2 D_m n^{-1}$ (OLS)



Dimension jump around $C = \sigma^2$



Slope heuristics algorithm (Birgé & Massart 2007)

- 1 For every $C > 0$, compute

$$\hat{m}(C) \in \operatorname{argmin}_{m \in \mathcal{M}_n} \left\{ \frac{1}{n} \left\| \hat{F}_m - Y \right\|^2 + C \frac{D_m}{n} \right\}.$$

- 2 Find \hat{C}_{jump} such that $D_{\hat{m}(C)}$ is “very large” when $C < \hat{C}_{\text{jump}}$ and “reasonably small” when $C > \hat{C}_{\text{jump}}$.
- 3 Select $\hat{m} = \hat{m}(2\hat{C}_{\text{jump}})$.

Practical use: CAPUSHE package (Baudry, Maugis & Michel, 2011)

<http://www.math.univ-toulouse.fr/~maugis/CAPUSHE.html>

Dimension jump / Minimal penalty: theory

Theorem (A. & Bach 2011, reformulated as Theorem 1 of arXiv:1901.07277)

Assumptions: i.i.d. Gaussian noise and $\exists m_1 \in \mathcal{M}$, $S_{m_1} = \mathbb{R}^n$, i.e., $\hat{F}_{m_1} = Y$. Then, $\forall x \in [0, n]$, with probability at least $1 - 4 \text{Card}(\mathcal{M})e^{-x}$,

$$\forall C < \left(1 - 100\sqrt{\frac{x}{n}}\right) \sigma^2, \quad D_{\hat{m}(C)} \geq \frac{9n}{10}$$

$$\forall C > \left(1 + 240\sqrt{\frac{x}{n}}\right) \sigma^2 + 40 \inf_{m / D_m \leq n/20} \left\{ \frac{1}{n} \|F_m - F\|^2 \right\}, \quad D_{\hat{m}(C)} \leq \frac{n}{10}$$

In the first case, $\frac{1}{n} \left\| \hat{F}_{\hat{m}(C)} - F \right\|^2 \geq \frac{7\sigma^2}{8}$ on the same event.

Remark: generalization to sub-Gaussian noise (arXiv:1901.07277).

Dimension jump / Minimal penalty: theory

Theorem (A. & Bach 2011, reformulated as Theorem 1 of arXiv:1901.07277)

Assumptions: i.i.d. Gaussian noise and $\exists m_1 \in \mathcal{M}$, $S_{m_1} = \mathbb{R}^n$, i.e., $\widehat{F}_{m_1} = Y$. Then, $\forall x \in [0, n]$, with probability at least $1 - 4 \text{Card}(\mathcal{M})e^{-x}$,

$$\forall C < \left(1 - 100\sqrt{\frac{x}{n}}\right) \sigma^2, \quad D_{\widehat{m}(C)} \geq \frac{9n}{10}$$

$$\forall C > \left(1 + 240\sqrt{\frac{x}{n}}\right) \sigma^2 + 40 \inf_{m / D_m \leq n/20} \left\{ \frac{1}{n} \|F_m - F\|^2 \right\}, \quad D_{\widehat{m}(C)} \leq \frac{n}{10}$$

Consequences: $\widehat{C}_{\text{jump}}$ estimates well σ^2 & oracle inequality for $\widehat{m}(2\widehat{C}_{\text{jump}})$ (see arXiv:1901.07277).

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?

- **Dimension jump**: largest jump? jump on a geometrical window? complexity threshold?

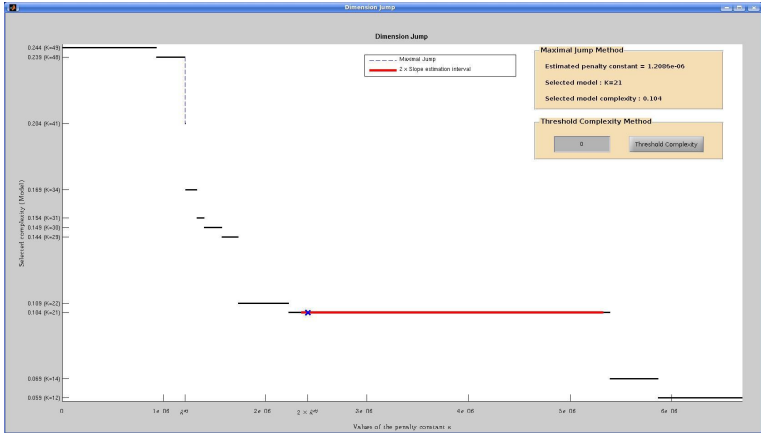
How to localize the jump in practice?

- Dimension jump: largest jump? jump on a geometrical window? complexity threshold?
- Estimation of the slope of the empirical risk as a function of the dimension:
computed with which models? robust regression?

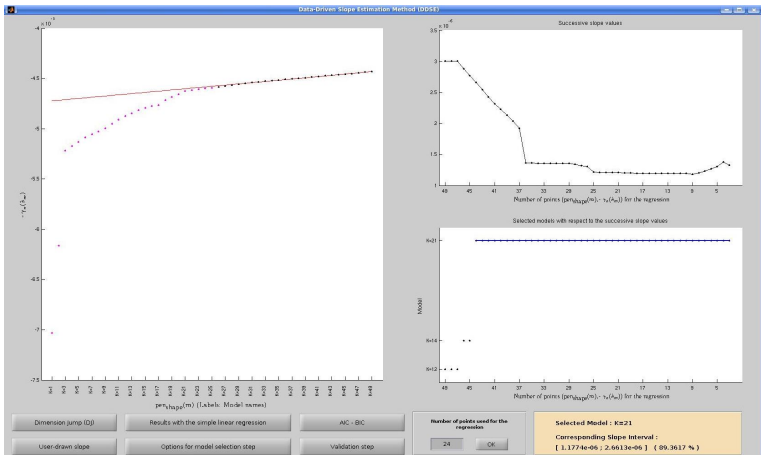
How to localize the jump in practice?

- Dimension jump: largest jump? jump on a geometrical window? complexity threshold?
- Estimation of the slope of the empirical risk as a function of the dimension:
computed with which models? robust regression?
- **Jump vs. slope? Take both!**
⇒ package CAPUSHE (Baudry, Maugis & Michel, 2011)
<http://www.math.univ-toulouse.fr/~maugis/CAPUSHE.html>

CAPUSHE (Baudry, Maugis & Michel, 2011): jump



CAPUSHE (Baudry, Maugis & Michel, 2011): slope



Part II

Model/estimator selection in the
prediction setting

Outline

- 1 Framework
- 2 Estimator selection
- 3 Minimal penalties

General prediction setting

- **Data:** $D_n = (X_i, Y_i)_{1 \leq i \leq n} \in (\mathcal{X} \times \mathcal{Y})^n$ assumed i.i.d. $\sim P$

General prediction setting

- **Data:** $D_n = (X_i, Y_i)_{1 \leq i \leq n} \in (\mathcal{X} \times \mathcal{Y})^n$ assumed i.i.d. $\sim P$
- **Predictor:** $t : \mathcal{X} \rightarrow \mathcal{Y}$
new data $X_{n+1} \rightsquigarrow t(X_{n+1})$ “predicts” Y_{n+1}

General prediction setting

- **Data:** $D_n = (X_i, Y_i)_{1 \leq i \leq n} \in (\mathcal{X} \times \mathcal{Y})^n$ assumed i.i.d. $\sim P$
- **Predictor:** $t : \mathcal{X} \rightarrow \mathcal{Y}$
new data $X_{n+1} \rightsquigarrow t(X_{n+1})$ “predicts” Y_{n+1}
- **Risk (prediction error):** $\mathcal{R}(t) = \mathbb{E}[c(t(X), Y)]$ where
 $(X, Y) \sim P$
minimal for $t = f^*$ (Bayes predictor)
 \Rightarrow **Excess risk** $\ell(t, f^*) := \mathcal{R}(t) - \mathcal{R}(f^*) \geq 0$.

General prediction setting

- **Data:** $D_n = (X_i, Y_i)_{1 \leq i \leq n} \in (\mathcal{X} \times \mathcal{Y})^n$ assumed i.i.d. $\sim P$
- **Predictor:** $t : \mathcal{X} \rightarrow \mathcal{Y}$
new data $X_{n+1} \rightsquigarrow t(X_{n+1})$ “predicts” Y_{n+1}
- **Risk (prediction error):** $\mathcal{R}(t) = \mathbb{E}[c(t(X), Y)]$ where
 $(X, Y) \sim P$
minimal for $t = f^*$ (Bayes predictor)
 \Rightarrow **Excess risk** $\ell(t, f^*) := \mathcal{R}(t) - \mathcal{R}(f^*) \geq 0$.
- **Goal:** from D_n only, find t with $\mathcal{R}(t)$ minimal.

General prediction setting

- **Data:** $D_n = (X_i, Y_i)_{1 \leq i \leq n} \in (\mathcal{X} \times \mathcal{Y})^n$ assumed i.i.d. $\sim P$
- **Predictor:** $t : \mathcal{X} \rightarrow \mathcal{Y}$
new data $X_{n+1} \rightsquigarrow t(X_{n+1})$ “predicts” Y_{n+1}
- **Risk (prediction error):** $\mathcal{R}(t) = \mathbb{E}[c(t(X), Y)]$ where
 $(X, Y) \sim P$
minimal for $t = f^*$ (Bayes predictor)
 \Rightarrow **Excess risk** $\ell(t, f^*) := \mathcal{R}(t) - \mathcal{R}(f^*) \geq 0$.
- **Goal:** from D_n only, find t with $\mathcal{R}(t)$ minimal.
- More general setting possible, including density estimation with LS or KL risk.

Prediction setting: examples

- **Regression:** $\mathcal{Y} = \mathbb{R}$

Prediction setting: examples

- **Regression:** $\mathcal{Y} = \mathbb{R}$
 - least squares: $c(y, y') = (y - y')^2$
 $\Rightarrow f^*(X) = \mathbb{E}[Y|X]$ and $\ell(t, f^*) = \mathbb{E}[(t(X) - f^*(X))^2]$

Prediction setting: examples

- **Regression:** $\mathcal{Y} = \mathbb{R}$
 - least squares: $c(y, y') = (y - y')^2$
 $\Rightarrow f^*(X) = \mathbb{E}[Y|X]$ and $\ell(t, f^*) = \mathbb{E}[(t(X) - f^*(X))^2]$
 - L^p loss: $c(y, y') = |y - y'|^p$, $p \geq 1$

Prediction setting: examples

- **Regression:** $\mathcal{Y} = \mathbb{R}$
 - least squares: $c(y, y') = (y - y')^2$
 $\Rightarrow f^*(X) = \mathbb{E}[Y|X]$ and $\ell(t, f^*) = \mathbb{E}[(t(X) - f^*(X))^2]$
 - L^p loss: $c(y, y') = |y - y'|^p$, $p \geq 1$
 - Huber loss (robustness):

$$c(y, y') = \begin{cases} \frac{1}{2}(y - y')^2 & \text{if } |y - y'| \leq \delta \\ \delta(|y - y'| - \frac{\delta}{2}) & \text{otherwise} \end{cases}$$

Prediction setting: examples

- **Regression:** $\mathcal{Y} = \mathbb{R}$
 - least squares: $c(y, y') = (y - y')^2$
 $\Rightarrow f^*(X) = \mathbb{E}[Y|X]$ and $\ell(t, f^*) = \mathbb{E}[(t(X) - f^*(X))^2]$
 - L^p loss: $c(y, y') = |y - y'|^p$, $p \geq 1$
 - Huber loss (robustness):

$$c(y, y') = \begin{cases} \frac{1}{2}(y - y')^2 & \text{if } |y - y'| \leq \delta \\ \delta(|y - y'| - \frac{\delta}{2}) & \text{otherwise} \end{cases}$$

- **Binary classification:** $\mathcal{Y} = \{0, 1\}$

Prediction setting: examples

- **Regression:** $\mathcal{Y} = \mathbb{R}$
 - least squares: $c(y, y') = (y - y')^2$
 $\Rightarrow f^*(X) = \mathbb{E}[Y|X]$ and $\ell(t, f^*) = \mathbb{E}[(t(X) - f^*(X))^2]$
 - L^p loss: $c(y, y') = |y - y'|^p$, $p \geq 1$
 - Huber loss (robustness):

$$c(y, y') = \begin{cases} \frac{1}{2}(y - y')^2 & \text{if } |y - y'| \leq \delta \\ \delta(|y - y'| - \frac{\delta}{2}) & \text{otherwise} \end{cases}$$

- **Binary classification:** $\mathcal{Y} = \{0, 1\}$
 - 0-1 loss: $c(y, y') = \mathbb{1}_{t(x) \neq y}$
 $\Rightarrow f^*(X) = \mathbb{1}_{\mathbb{E}[Y|X] \geq 1/2}$

Prediction setting: examples

- **Regression:** $\mathcal{Y} = \mathbb{R}$
 - least squares: $c(y, y') = (y - y')^2$
 $\Rightarrow f^*(X) = \mathbb{E}[Y|X]$ and $\ell(t, f^*) = \mathbb{E}[(t(X) - f^*(X))^2]$
 - L^p loss: $c(y, y') = |y - y'|^p$, $p \geq 1$
 - Huber loss (robustness):

$$c(y, y') = \begin{cases} \frac{1}{2}(y - y')^2 & \text{if } |y - y'| \leq \delta \\ \delta(|y - y'| - \frac{\delta}{2}) & \text{otherwise} \end{cases}$$

- **Binary classification:** $\mathcal{Y} = \{0, 1\}$
 - 0-1 loss: $c(y, y') = \mathbb{1}_{t(x) \neq y}$
 $\Rightarrow f^*(X) = \mathbb{1}_{\mathbb{E}[Y|X] \geq 1/2}$
 - convex losses (hinge, logistic, exponential, ...)

Prediction setting: examples

- **Regression:** $\mathcal{Y} = \mathbb{R}$
 - least squares: $c(y, y') = (y - y')^2$
 $\Rightarrow f^*(X) = \mathbb{E}[Y|X]$ and $\ell(t, f^*) = \mathbb{E}[(t(X) - f^*(X))^2]$
 - L^p loss: $c(y, y') = |y - y'|^p$, $p \geq 1$
 - Huber loss (robustness):

$$c(y, y') = \begin{cases} \frac{1}{2}(y - y')^2 & \text{if } |y - y'| \leq \delta \\ \delta(|y - y'| - \frac{\delta}{2}) & \text{otherwise} \end{cases}$$

- **Binary classification:** $\mathcal{Y} = \{0, 1\}$
 - 0-1 loss: $c(y, y') = \mathbb{1}_{t(x) \neq y}$
 $\Rightarrow f^*(X) = \mathbb{1}_{\mathbb{E}[Y|X] \geq 1/2}$
 - convex losses (hinge, logistic, exponential, ...)
- **Multi-class classification:** $\mathcal{Y} = \{0, \dots, M - 1\}$

Link with part I (fixed-design regression)

- Random-design regression, least-squares loss:
 X_1, \dots, X_n **i.i.d.**

- Fixed-design regression:
 X_1, \dots, X_n **deterministic**

Link with part I (fixed-design regression)

- Random-design regression, least-squares loss:

X_1, \dots, X_n **i.i.d.**

target: $f^*(X) = \mathbb{E}[Y|X]$

- Fixed-design regression:

X_1, \dots, X_n **deterministic**

target $(f^*(X_i))_{1 \leq i \leq n} = (\mathbb{E}[Y_i|X_i])_{1 \leq i \leq n}$

Link with part I (fixed-design regression)

- Random-design regression, least-squares loss:

X_1, \dots, X_n **i.i.d.**

target: $f^*(X) = \mathbb{E}[Y|X]$

excess risk: $\ell(t, f^*) = \mathcal{R}(t) - \mathcal{R}(f^*) = \mathbb{E}[(t(X) - f^*(X))^2]$

- Fixed-design regression:

X_1, \dots, X_n **deterministic**

target $(f^*(X_i))_{1 \leq i \leq n} = (\mathbb{E}[Y_i|X_i])_{1 \leq i \leq n}$

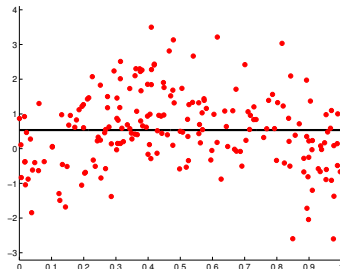
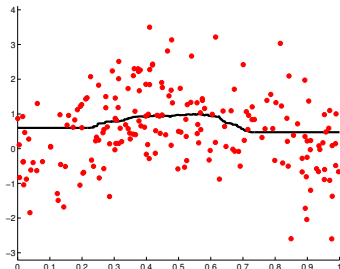
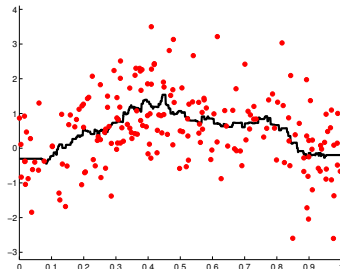
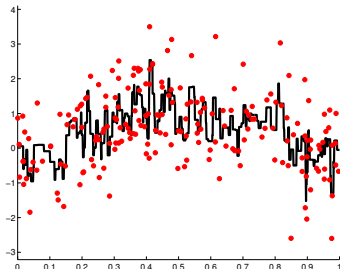
excess risk: $\frac{1}{n} \sum_{i=1}^n (t(X_i) - f^*(X_i))^2$

\leftrightarrow “ $X \sim \mathcal{U}(\{X_1, \dots, X_n\})$ ”

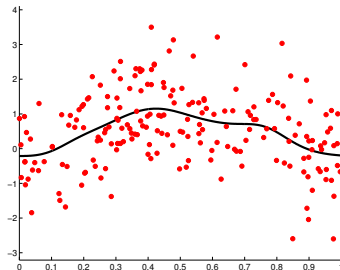
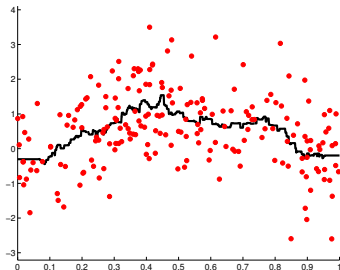
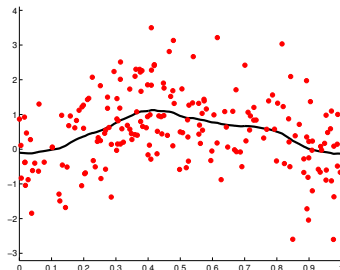
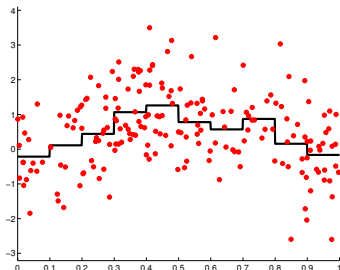
Outline

- 1 Framework
- 2 Estimator selection
- 3 Minimal penalties

Estimator selection: k nearest neighbours



Estimators: regressogram, ridge, k -NN, Nadaraya-Watson



Estimator selection

- Estimator collection $(\hat{f}_m)_{m \in \mathcal{M}} \Rightarrow \hat{m}(D_n)$?

Estimator selection

- Estimator collection $(\hat{f}_m)_{m \in \mathcal{M}} \Rightarrow \hat{m}(D_n)$?
- Examples:
 - model selection
 - parameter tuning (choosing k or the distance for k -NN, choice of a regularization parameter, choice of a kernel, etc.)
 - choice between different methods
ex.: k -NN vs. kernel ridge?

Estimator selection

- Estimator collection $(\hat{f}_m)_{m \in \mathcal{M}} \Rightarrow \hat{m}(D_n)$?
- Examples:
 - model selection
 - parameter tuning (choosing k or the distance for k -NN, choice of a regularization parameter, choice of a kernel, etc.)
 - choice between different methods
ex.: k -NN vs. kernel ridge?
- Goal: minimize the risk $\mathcal{R}(\hat{f}_{\hat{m}(D_n)}(D_n))$
- Other possible goal: identify the “best” estimator (or the “true” model)

Approximation / estimation error decomposition?

- No general decomposition of the risk between approximation and estimation error

Approximation / estimation error decomposition?

- No general decomposition of the risk between approximation and estimation error
- Sometimes possible:
 - empirical risk minimizer: $\hat{f}_m \in \operatorname{argmin}_{t \in \mathcal{S}_m} \hat{\mathcal{R}}_n(t)$
 - “linear” estimators (fixed-design regression)
 - local averaging estimators

Approximation / estimation error decomposition?

- No general decomposition of the risk between approximation and estimation error
- Sometimes possible:
 - empirical risk minimizer: $\hat{f}_m \in \operatorname{argmin}_{t \in \mathcal{S}_m} \hat{\mathcal{R}}_n(t)$
 - “linear” estimators (fixed-design regression)
 - local averaging estimators
- Always have to **avoid overfitting and underfitting**

Estimator selection: methods

- Classical approach:

$$\hat{m}(D_n) \in \underset{m \in \mathcal{M}}{\operatorname{argmin}} \{ \operatorname{crit}(m) \}$$

Estimator selection: methods

- Classical approach:

$$\hat{m}(D_n) \in \underset{m \in \mathcal{M}}{\operatorname{argmin}} \{ \operatorname{crit}(m) \}$$

- Examples:
 - **penalization** (Mallows' C_p , AIC, BIC, structural risk minimization, ...)
 - **cross-validation**
 - FPE, GCV, ...

Estimator selection: methods

- Classical approach:

$$\hat{m}(D_n) \in \underset{m \in \mathcal{M}}{\operatorname{argmin}} \{ \operatorname{crit}(m) \}$$

- Examples:
 - **penalization** (Mallows' C_p , AIC, BIC, structural risk minimization, ...)
 - **cross-validation**
 - FPE, GCV, ...
- How to choose crit?

Estimator selection: methods

- Classical approach:

$$\hat{m}(D_n) \in \underset{m \in \mathcal{M}}{\operatorname{argmin}} \{ \operatorname{crit}(m) \}$$

- Examples:
 - **penalization** (Mallows' C_p , AIC, BIC, structural risk minimization, ...)
 - **cross-validation**
 - FPE, GCV, ...
- How to choose crit?
Idea: use the **key lemma**

The key lemma (revisited)

Lemma

Let $\text{crit} : \mathcal{M} \rightarrow \mathbb{R}$ be any function (possibly data-dependent).
On the event Ω on which, $\forall m, m' \in \mathcal{M}$

$$\begin{aligned} & \left[\text{crit}(m) - \mathcal{R}(\hat{f}_m) \right] - \left[\text{crit}(m') - \mathcal{R}(\hat{f}_{m'}) \right] \\ & \leq A(m) + B(m'), \end{aligned}$$

we have $\forall \hat{m} \in \arg \min_{m \in \mathcal{M}} \{ \text{crit}(m) \}$,

$$\ell(\hat{f}_{\hat{m}}, f^*) - B(\hat{m}) \leq \inf_{m \in \mathcal{M}} \left\{ \ell(\hat{f}_m, f^*) + A(m) \right\}$$

Proof of the key lemma

For any $m \in \mathcal{M}$, we have

$$\begin{aligned} & \mathcal{R}(\widehat{f}_{\widehat{m}}) \\ &= \text{crit}(\widehat{m}) + \mathcal{R}(\widehat{f}_{\widehat{m}}) - \text{crit}(\widehat{m}) \\ &\leq \text{crit}(m) + \mathcal{R}(\widehat{f}_{\widehat{m}}) - \text{crit}(\widehat{m}) \\ &= \mathcal{R}(\widehat{f}_m) + \text{crit}(m) - \mathcal{R}(\widehat{f}_m) + \mathcal{R}(\widehat{f}_{\widehat{m}}) - \text{crit}(\widehat{m}) \\ &\leq \mathcal{R}(\widehat{f}_m) + A(m) + B(\widehat{m}) \end{aligned}$$

hence

$$\mathcal{R}(\widehat{f}_{\widehat{m}}) - B(\widehat{m}) \leq \inf_{m \in \mathcal{M}} \left\{ \mathcal{R}(\widehat{f}_m) + A(m) \right\}$$

Proof of the key lemma

For any $m \in \mathcal{M}$, we have

$$\begin{aligned} & \mathcal{R}(\widehat{f}_{\widehat{m}}) \\ &= \text{crit}(\widehat{m}) + \mathcal{R}(\widehat{f}_{\widehat{m}}) - \text{crit}(\widehat{m}) \\ &\leq \text{crit}(m) + \mathcal{R}(\widehat{f}_{\widehat{m}}) - \text{crit}(\widehat{m}) \\ &= \mathcal{R}(\widehat{f}_m) + \text{crit}(m) - \mathcal{R}(\widehat{f}_m) + \mathcal{R}(\widehat{f}_{\widehat{m}}) - \text{crit}(\widehat{m}) \\ &\leq \mathcal{R}(\widehat{f}_m) + A(m) + B(\widehat{m}) \end{aligned}$$

hence

$$\begin{aligned} & \mathcal{R}(\widehat{f}_{\widehat{m}}) - B(\widehat{m}) \leq \inf_{m \in \mathcal{M}} \left\{ \mathcal{R}(\widehat{f}_m) + A(m) \right\} \\ \text{and} \quad & \ell(\widehat{f}_{\widehat{m}}, f^*) - B(\widehat{m}) \leq \inf_{m \in \mathcal{M}} \left\{ \ell(\widehat{f}_m, f^*) + A(m) \right\}. \quad \square \end{aligned}$$

Application 1: unbiased risk estimation

$$\forall m \in \mathcal{M}, \quad \mathbb{E}[\text{crit}(m; D_n)] \approx \mathbb{E}[\mathcal{R}(\hat{f}_m(D_n))]$$

- **Examples:** C_p , AIC, cross-validation, FPE, ...

Application 1: unbiased risk estimation

$$\forall m \in \mathcal{M}, \quad \mathbb{E}[\text{crit}(m; D_n)] \approx \mathbb{E}[\mathcal{R}(\hat{f}_m(D_n))]$$

- **Examples:** C_p , AIC, cross-validation, FPE, ...
- **Concentration inequalities + \mathcal{M} “not too large”**

⇒ with a large probability, $\forall m, m' \in \mathcal{M}$,

$$\left[\text{crit}(m) - \mathcal{R}(\hat{f}_m) \right] - \left[\text{crit}(m') - \mathcal{R}(\hat{f}_{m'}) \right] \leq A(m) + B(m'),$$

with $A(m) = B(m) \leq \epsilon_1 \ell(\hat{f}_m, f^*) + \epsilon_2.$

Application 1: unbiased risk estimation

$$\forall m \in \mathcal{M}, \quad \mathbb{E}[\text{crit}(m; D_n)] \approx \mathbb{E}[\mathcal{R}(\hat{f}_m(D_n))]$$

- **Examples:** C_p , AIC, cross-validation, FPE, ...
- **Concentration inequalities + \mathcal{M} “not too large”**
⇒ with a large probability, $\forall m, m' \in \mathcal{M}$,

$$\left[\text{crit}(m) - \mathcal{R}(\hat{f}_m) \right] - \left[\text{crit}(m') - \mathcal{R}(\hat{f}_{m'}) \right] \leq A(m) + B(m'),$$

$$\text{with} \quad A(m) = B(m) \leq \epsilon_1 \ell(\hat{f}_m, f^*) + \epsilon_2.$$

- If $\epsilon_1 < 1$, by the key lemma

$$\ell(\hat{f}_{\hat{m}}, f^*) \leq \frac{1 + \epsilon_1}{1 - \epsilon_1} \inf_{m \in \mathcal{M}} \left\{ \ell(\hat{f}_m(D_n), f^*) \right\} + \frac{2\epsilon_2}{1 - \epsilon_1}$$

- ⇒ **oracle inequality**, first-order optimal if $\epsilon_1 \ll 1$ and $\epsilon_2 \ll \inf_{m \in \mathcal{M}} \{ \ell(f^*, \hat{f}_m(D_n)) \}$.

Application 2: upper bound on the risk

$$\forall m \in \mathcal{M}, \quad \text{crit}(m; D_n) \geq \mathcal{R}(\hat{f}_m(D_n)) \quad \left(\text{or } \ell(\hat{f}_m(D_n), f^*) \right)$$

(with a large probability)

- **Examples:** BIC, structural risk minimization, ...

Application 2: upper bound on the risk

$$\forall m \in \mathcal{M}, \quad \text{crit}(m; D_n) \geq \mathcal{R}(\hat{f}_m(D_n)) \quad (\text{or } \ell(\hat{f}_m(D_n), f^*))$$

(with a large probability)

- **Examples:** BIC, structural risk minimization, ...
- Then, $\forall m, m' \in \mathcal{M}$,

$$\left[\text{crit}(m) - \mathcal{R}(\hat{f}_m) \right] - \left[\text{crit}(m') - \mathcal{R}(\hat{f}_{m'}) \right] \leq A(m) + B(m'),$$

with $A(m) = 0$ and $B(m) = \text{crit}(m) - \mathcal{R}(\hat{f}_m(D_n))$.

Application 2: upper bound on the risk

$$\forall m \in \mathcal{M}, \quad \text{crit}(m; D_n) \geq \mathcal{R}(\hat{f}_m(D_n)) \quad (\text{or } \ell(\hat{f}_m(D_n), f^*))$$

(with a large probability)

- **Examples:** BIC, structural risk minimization, ...
- Then, $\forall m, m' \in \mathcal{M}$,

$$\left[\text{crit}(m) - \mathcal{R}(\hat{f}_m) \right] - \left[\text{crit}(m') - \mathcal{R}(\hat{f}_{m'}) \right] \leq A(m) + B(m'),$$

with $A(m) = 0$ and $B(m) = \text{crit}(m) - \mathcal{R}(\hat{f}_m(D_n))$.

- By the key lemma

$$\ell(\hat{f}_{\hat{m}}, f^*) \leq \inf_{m \in \mathcal{M}} \left\{ \ell(\hat{f}_m(D_n), f^*) + B(m) \right\}$$

⇒ **oracle inequality**, interesting if $B(m)$ small enough.

Estimator selection by penalization

$$\forall m \in \mathcal{M}, \quad \text{crit}(m; D_n) = \widehat{\mathcal{R}}_n(\widehat{f}_m(D_n)) + \text{pen}(m)$$

$$\text{where} \quad \widehat{\mathcal{R}}_n(t) := \frac{1}{n} \sum_{i=1}^n c(t(X_i), Y_i) \quad (\text{empirical risk})$$

Estimator selection by penalization

$$\forall m \in \mathcal{M}, \quad \text{crit}(m; D_n) = \widehat{\mathcal{R}}_n(\widehat{f}_m(D_n)) + \text{pen}(m)$$

$$\text{where} \quad \widehat{\mathcal{R}}_n(t) := \frac{1}{n} \sum_{i=1}^n c(t(X_i), Y_i) \quad (\text{empirical risk})$$

- Ideal penalty:

$$\text{pen}_{\text{id}}(m; D_n) := \mathcal{R}(\widehat{f}_m(D_n)) - \widehat{\mathcal{R}}_n(\widehat{f}_m(D_n)).$$

Estimator selection by penalization

$$\forall m \in \mathcal{M}, \quad \text{crit}(m; D_n) = \widehat{\mathcal{R}}_n(\widehat{f}_m(D_n)) + \text{pen}(m)$$

$$\text{where} \quad \widehat{\mathcal{R}}_n(t) := \frac{1}{n} \sum_{i=1}^n c(t(X_i), Y_i) \quad (\text{empirical risk})$$

- Ideal penalty:

$$\text{pen}_{\text{id}}(m; D_n) := \mathcal{R}(\widehat{f}_m(D_n)) - \widehat{\mathcal{R}}_n(\widehat{f}_m(D_n)).$$

- Unbiased risk estimation

$$\Leftrightarrow \forall m \in \mathcal{M}, \quad \mathbb{E}[\text{pen}(m; D_n)] = \mathbb{E}[\text{pen}_{\text{id}}(m; D_n)].$$

Estimator selection by penalization

$$\forall m \in \mathcal{M}, \quad \text{crit}(m; D_n) = \widehat{\mathcal{R}}_n(\widehat{f}_m(D_n)) + \text{pen}(m)$$

$$\text{where} \quad \widehat{\mathcal{R}}_n(t) := \frac{1}{n} \sum_{i=1}^n c(t(X_i), Y_i) \quad (\text{empirical risk})$$

- Ideal penalty:

$$\text{pen}_{\text{id}}(m; D_n) := \mathcal{R}(\widehat{f}_m(D_n)) - \widehat{\mathcal{R}}_n(\widehat{f}_m(D_n)).$$

- Unbiased risk estimation

$$\Leftrightarrow \forall m \in \mathcal{M}, \quad \mathbb{E}[\text{pen}(m; D_n)] = \mathbb{E}[\text{pen}_{\text{id}}(m; D_n)].$$

- Upper bound on the risk

$$\Leftrightarrow \forall m \in \mathcal{M}, \quad \text{pen}(m; D_n) \geq \text{pen}_{\text{id}}(m; D_n).$$

Outline

- 1 Framework
- 2 Estimator selection
- 3 Minimal penalties

Motivation: Penalties known up to a constant factor

$$\hat{m}(D_n) \in \operatorname{argmin}_{m \in \mathcal{M}} \left\{ \hat{\mathcal{R}}_n(\hat{f}_m) + \operatorname{pen}(m) \right\}$$

- Optimal penalties depending on the noise level σ^2 (Mallows, 1973):

$$\operatorname{pen}_{\text{Cp}}(m) = \frac{2\sigma^2 D_m}{n} \quad \operatorname{pen}_{\text{CL}}(m) = \frac{2\sigma^2 \operatorname{tr}(A_m)}{n}$$

Rk: various methods for estimating σ^2 or avoiding its estimation (FPE, Akaike, 1970; GCV, Craven & Wahba, 1978; Baraud, Giraud & Huet, 2009).

Motivation: Penalties known up to a constant factor

$$\hat{m}(D_n) \in \operatorname{argmin}_{m \in \mathcal{M}} \left\{ \hat{\mathcal{R}}_n(\hat{f}_m) + \operatorname{pen}(m) \right\}$$

- Optimal penalties depending on the noise level σ^2 (Mallows, 1973):

$$\operatorname{pen}_{\text{CP}}(m) = \frac{2\sigma^2 D_m}{n} \quad \operatorname{pen}_{\text{CL}}(m) = \frac{2\sigma^2 \operatorname{tr}(A_m)}{n}$$

- Optimal penalty known **asymptotically** (AIC; Akaike, 1973)
- Resampling-based penalties
- Optimal constant unknown even in theory (change-point detection, mixture models, global/local Rademacher complexities, ...)

Motivation: Penalties known up to a constant factor

$$\hat{m}(D_n) \in \operatorname{argmin}_{m \in \mathcal{M}} \left\{ \hat{\mathcal{R}}_n(\hat{f}_m) + \operatorname{pen}(m) \right\}$$

- Optimal penalties depending on the noise level σ^2 (Mallows, 1973):

$$\operatorname{pen}_{\text{CP}}(m) = \frac{2\sigma^2 D_m}{n} \quad \operatorname{pen}_{\text{CL}}(m) = \frac{2\sigma^2 \operatorname{tr}(A_m)}{n}$$

- Optimal penalty known **asymptotically** (AIC; Akaike, 1973)
- Resampling-based penalties
- Optimal constant unknown even in theory (change-point detection, mixture models, global/local Rademacher complexities, ...)

Goals: estimation of the optimal constant (e.g., σ^2) for estimator selection, under minimal assumptions, without overfitting

Motivation: what is the minimal penalization level?

$$\hat{m}(D_n) \in \operatorname{argmin}_{m \in \mathcal{M}} \left\{ \hat{\mathcal{R}}_n(\hat{f}_m) + \operatorname{pen}(m) \right\}$$

- **Optimal penalization level:**

$$\mathbb{E}[\operatorname{pen}_{\text{id}}(m)] = \mathbb{E} \left[\mathcal{R}(\hat{f}_m(D_m)) - \hat{\mathcal{R}}_n(\hat{f}_m(D_n)) \right]$$

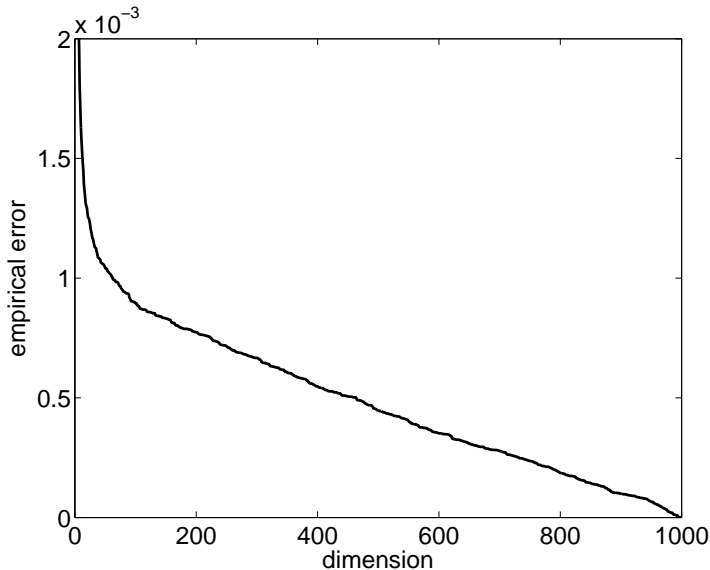
- **Minimal penalization level?**

$$\operatorname{pen}_{\min}(m) = C_{\min} \times \mathbb{E}[\operatorname{pen}_{\text{id}}(m)]?$$

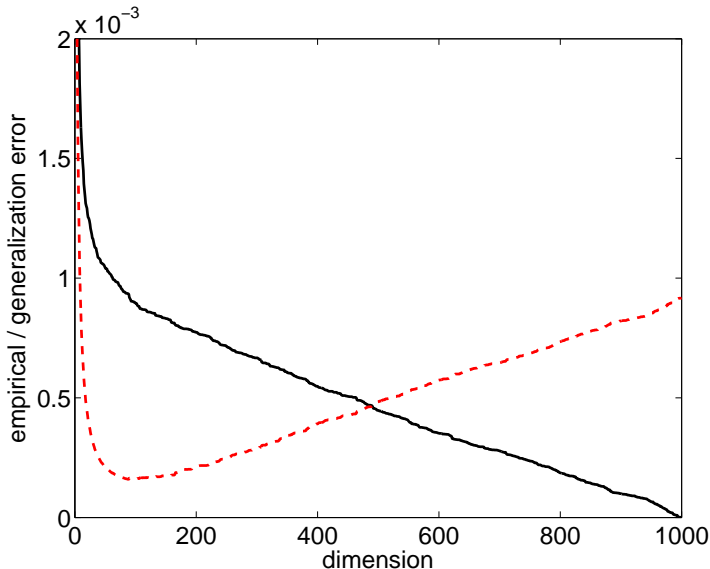
other penalty shape?

⇒ natural candidate: $\mathbb{E}[\hat{\mathcal{R}}_n(f_m^*) - \hat{\mathcal{R}}_n(\hat{f}_m)]$ (up to the definition of f_m^*)

Motivation: “L-curve” and elbow heuristics?



Motivation: “L-curve” and elbow heuristics?



General slope heuristics algorithm

Input: $\forall m \in \mathcal{M}$, $\widehat{\mathcal{R}}_n(\widehat{s}_m)$, $\text{pen}_0(m)$ and \mathcal{C}_m

- 1 For every $C > 0$, compute

$$\widehat{m}_{\min}(C) \in \underset{m \in \mathcal{M}}{\text{argmin}} \left\{ \widehat{\mathcal{R}}_n(\widehat{s}_m) + C \text{pen}_0(m) \right\}.$$

- 2 Find $\widehat{C}_{\text{jump}}$ such that $C_{\widehat{m}_{\min}(C)}$ is “too large” when $C < \widehat{C}_{\text{jump}}$ and “reasonably small” when $C > \widehat{C}_{\text{jump}}$.

- 3 Select

$$\widehat{m} \in \underset{m \in \mathcal{M}}{\text{argmin}} \left\{ \widehat{\mathcal{R}}_n(\widehat{s}_m) + 2\widehat{C}_{\text{jump}} \text{pen}_0(m) \right\}.$$

Slope heuristics: ideas for a proof

- $\exists C^* > 0$, C^* pen₀ minimal penalty, C^* pen₁ optimal penalty.
- Decomposition of the ideal penalty:

$$\begin{aligned} \text{pen}_{\text{id}}(m) &= \mathcal{R}(\hat{s}_m) - \hat{\mathcal{R}}_n(\hat{s}_m) \\ &= \underbrace{\mathcal{R}(\hat{s}_m) - \mathcal{R}(f_m^*)}_{p_1(m)} + \underbrace{\mathcal{R}(f_m^*) - \hat{\mathcal{R}}_n(f_m^*)}_{\delta(m)} + \underbrace{\hat{\mathcal{R}}_n(f_m^*) - \hat{\mathcal{R}}_n(\hat{s}_m)}_{p_2(m)}. \end{aligned}$$

- Good candidate for the minimal penalty: p_2 or $\mathbb{E}[p_2]$.
- Good candidate for the optimal penalty: $p_1 + p_2$ or $\mathbb{E}[p_1 + p_2] = \mathbb{E}[\text{pen}_{\text{id}}]$.
- Slope heuristics: $\mathbb{E}[p_1] \approx \mathbb{E}[p_2]$.

More details in:

S. A. *Minimal penalties and the slope heuristics: a survey*. Journal de la Société Française de Statistique, 2019, Vol 160, No 3. 1–106.

arxiv:1901.07277

Slope heuristics: theoretical results

- OLS, fixed-design regression, homoscedastic Gaussian noise (Birgé & Massart, 2007) \Rightarrow part I

Slope heuristics: theoretical results

- OLS, fixed-design regression, homoscedastic Gaussian noise (Birgé & Massart, 2007) \Rightarrow part I
- OLS, **random-design** regression, **heteroscedastic** noise (regressograms, A. & Massart, 2009; piecewise polynomials, Saumard, 2013; localized basis, Navarro & Saumard 2017)
- Least-squares **density estimation**, i.i.d. (Lerasle, 2012) or **mixing data** (Lerasle, 2011)

Slope heuristics: theoretical results

- OLS, fixed-design regression, homoscedastic Gaussian noise (Birgé & Massart, 2007) \Rightarrow part I
- OLS, **random-design** regression, **heteroscedastic** noise (regressograms, A. & Massart, 2009; piecewise polynomials, Saumard, 2013; localized basis, Navarro & Saumard 2017)
- Least-squares **density estimation**, i.i.d. (Lerasle, 2012) or **mixing data** (Lerasle, 2011)
- Density estimation, **Kullback risk**, **maximum-likelihood estimators** on histograms (Saumard, 2010)
- **Minimum contrast estimator**, **regular contrast** (Saumard, 2010)
- **Specification probabilities in general random fields**, least-squares/Kullback risks, empirical contrast minimizers (Lerasle & Takahashi, 2016)

Slope heuristics: theoretical results

- OLS, fixed-design regression, homoscedastic Gaussian noise (Birgé & Massart, 2007) \Rightarrow part I
- OLS, **random-design** regression, **heteroscedastic** noise (regressograms, A. & Massart, 2009; piecewise polynomials, Saumard, 2013; localized basis, Navarro & Saumard 2017)
- Least-squares **density estimation**, i.i.d. (Lerasle, 2012) or **mixing data** (Lerasle, 2011)
- Density estimation, **Kullback risk**, **maximum-likelihood estimators** on histograms (Saumard, 2010)
- **Minimum contrast estimator**, **regular contrast** (Saumard, 2010)
- **Specification probabilities in general random fields**, least-squares/Kullback risks, empirical contrast minimizers (Lerasle & Takahashi, 2016)
- and many partial results, see arxiv:1901.07277

Slope heuristics: Empirical results

- Binary (supervised) classification (Zwald & Blanchard, 2005)
- Model-based clustering (Maugis & Michel 2011, Gallopin & Devijver 2018, and many others)
- Lasso (Connault, 2011)
- Large collection of models: Change-point detection (Lebarbier, 2005)

Slope heuristics: Empirical results

- Binary (supervised) classification (Zwald & Blanchard, 2005)
- Model-based clustering (Maugis & Michel 2011, Gallopin & Devijver 2018, and many others)
- Lasso (Connault, 2011)
- Large collection of models: Change-point detection (Lebarbier, 2005)
- and many others, see Baudry, Maugis & Michel (2011) and [arXiv:1901.07277](https://arxiv.org/abs/1901.07277)

General minimal penalty algorithm

Input: $\forall m \in \mathcal{M}$, $\widehat{\mathcal{R}}_n(\widehat{s}_m)$, $\text{pen}_0(m)$, $\text{pen}_1(m)$ and \mathcal{C}_m .

- 1 For every $C > 0$, compute

$$\widehat{m}_{\min}(C) \in \underset{m \in \mathcal{M}}{\text{argmin}} \left\{ \widehat{\mathcal{R}}_n(\widehat{s}_m) + C \text{pen}_0(m) \right\}.$$

- 2 Find $\widehat{C}_{\text{jump}}$ such that $\widehat{C}_{\widehat{m}_{\min}(C)}$ is “too large” when $C < \widehat{C}_{\text{jump}}$ and “reasonably small” when $C > \widehat{C}_{\text{jump}}$.
- 3 Select

$$\widehat{m} \in \underset{m \in \mathcal{M}}{\text{argmin}} \left\{ \widehat{\mathcal{R}}_n(\widehat{s}_m) + \widehat{C}_{\text{jump}} \text{pen}_1(m) \right\}.$$

General minimal penalty algorithm

Input: $\forall m \in \mathcal{M}$, $\widehat{\mathcal{R}}_n(\widehat{s}_m)$, $\text{pen}_0(m)$, $\text{pen}_1(m)$ and \mathcal{C}_m .

- 1 For every $C > 0$, compute

$$\widehat{m}_{\min}(C) \in \underset{m \in \mathcal{M}}{\text{argmin}} \left\{ \widehat{\mathcal{R}}_n(\widehat{s}_m) + C \text{pen}_0(m) \right\}.$$

- 2 Find $\widehat{C}_{\text{jump}}$ such that $\widehat{C}_{\widehat{m}_{\min}(C)}$ is “too large” when $C < \widehat{C}_{\text{jump}}$ and “reasonably small” when $C > \widehat{C}_{\text{jump}}$.
- 3 Select

$$\widehat{m} \in \underset{m \in \mathcal{M}}{\text{argmin}} \left\{ \widehat{\mathcal{R}}_n(\widehat{s}_m) + \widehat{C}_{\text{jump}} \text{pen}_1(m) \right\}.$$

Example (slope heuristics): $\text{pen}_1 = 2 \text{pen}_0$.

General algorithm: results

Full theoretical results:

- Linear estimators, regression (OLS, k -NN, Nadaraya-Watson, kernel ridge, ...): (A. & Bach, 2009–2011)

Fixed design regression, $\hat{F}_m = A_m Y$

$$\text{pen}_0(m) = [2 \text{tr}(A_m) - \text{tr}(A_m^\top A_m)]/n$$

$$\text{pen}_1(m) = 2 \text{tr}(A_m)/n$$

- Linear estimators, least-squares density estimation (OLS, weighted least-squares, Parzen): Lerasle, Magalhães & Reynaud-Bouret (2016)

General algorithm: results

Full theoretical results:

- **Linear estimators, regression (OLS, k -NN, Nadaraya-Watson, kernel ridge, ...):** (A. & Bach, 2009–2011)

Fixed design regression, $\hat{F}_m = A_m Y$

$$\text{pen}_0(m) = [2 \text{tr}(A_m) - \text{tr}(A_m^\top A_m)]/n$$

$$\text{pen}_1(m) = 2 \text{tr}(A_m)/n$$

- **Linear estimators, least-squares density estimation (OLS, weighted least-squares, Parzen):** Lerasle, Magalhães & Reynaud-Bouret (2016)

Partial theory and empirical results for **large collection of models** (eg, change-point detection): Lebarbier (2005), Birgé & Massart (2007), Sorba (2017)...

$\Rightarrow \text{pen}_1(m) = \kappa \text{pen}_0(m)$ with $\kappa \neq 2$.

see arXiv:1901.07277 for details

More generalization?

- Idea: use a **visible phase transition** (dimension/complexity jump) for making an optimal estimator selection
- Related procedures:
 - L-curve / **elbow heuristics** (no theory)
 - Scree test (no theory)
 - Thresholding under the null (requires to know null distribution)
 - **Goldenshluger-Lepski's method, penalized comparison to overfitting (PCO)**: with theory!
 - ... (see arXiv:1901.07277)

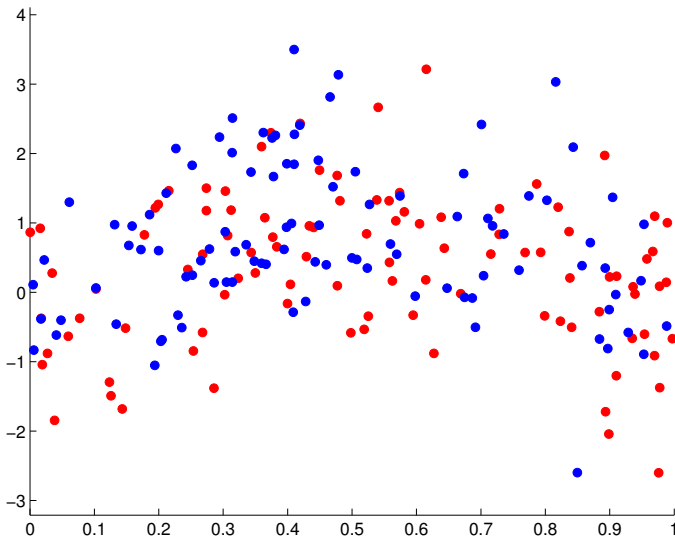
Part III

Cross-validation for estimator selection/aggregation

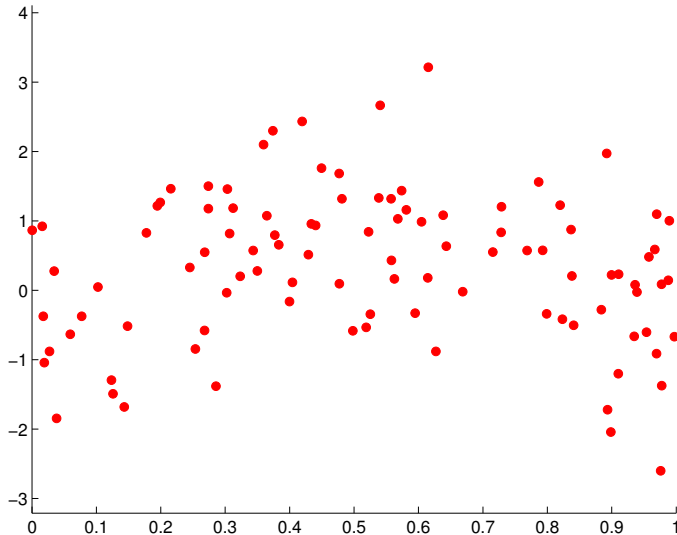
Outline

- 1 Definition and Examples
- 2 Cross-validation for risk estimation
- 3 Cross-validation for estimator selection
- 4 Conclusion on CV
- 5 Combining cross-validation with aggregation

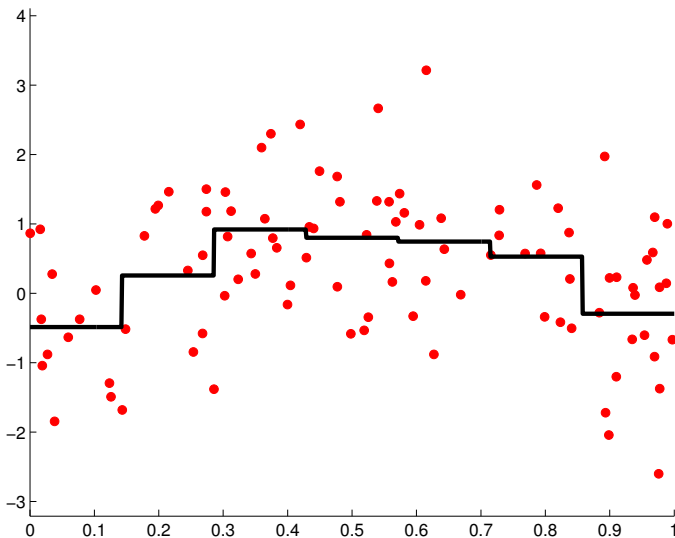
Validation principle: data splitting



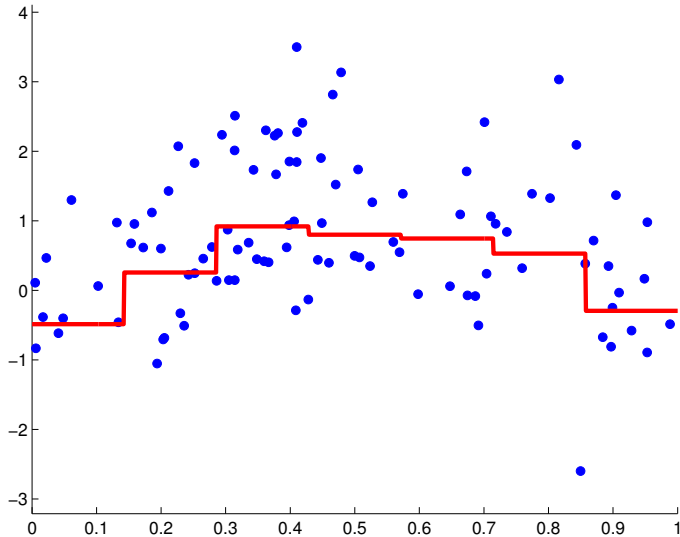
Validation principle: training/learning sample



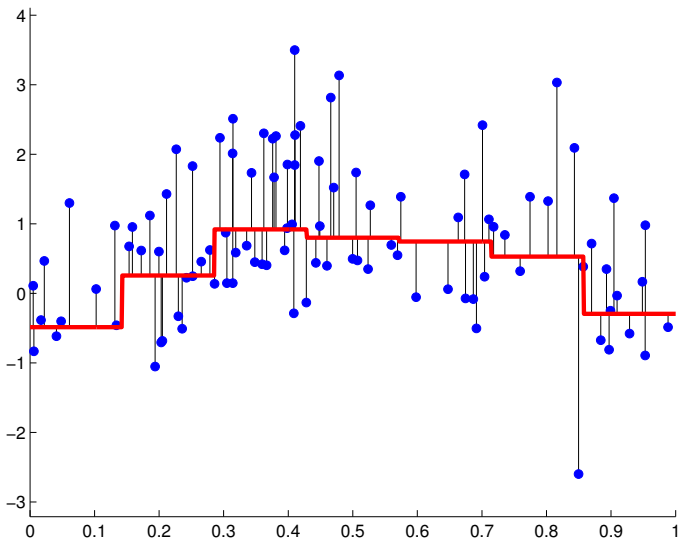
Validation principle: training/learning sample



Validation principle: validation sample



Validation principle: validation sample



Cross-validation

$$\underbrace{(X_1, Y_1), \dots, (X_{n_t}, Y_{n_t})}_{\text{Training set}}$$

Training set $D_n^{(t)} \Rightarrow \hat{f}_m^{(t)} = \hat{f}_m(D_n^{(t)})$

$$\underbrace{(X_{n_t+1}, Y_{n_t+1}), \dots, (X_n, Y_n)}_{\text{Validation set}}$$

Validation set $D_n^{(v)} \Rightarrow$ evaluate risk

- hold-out estimator of the risk:

$$\hat{\mathcal{R}}_n^{(v)}(\hat{f}_m^{(t)}) = \frac{1}{n_v} \sum_{(X_i, Y_i) \in D_n^{(v)}} c(\hat{f}_m^{(t)}(X_i); Y_i) \quad n_v = |D_n^{(v)}| = n - n_t$$

Cross-validation

$$\underbrace{(X_1, Y_1), \dots, (X_{n_t}, Y_{n_t})}_{\text{Training set}}$$

$$\underbrace{(X_{n_t+1}, Y_{n_t+1}), \dots, (X_n, Y_n)}_{\text{Validation set}}$$

Training set $D_n^{(t)} \Rightarrow \hat{f}_m^{(t)} = \hat{f}_m(D_n^{(t)})$ Validation set $D_n^{(v)} \Rightarrow$ evaluate risk

- **hold-out** estimator of the risk:

$$\hat{\mathcal{R}}_n^{(v)}(\hat{f}_m^{(t)}) = \frac{1}{n_v} \sum_{(X_i, Y_i) \in D_n^{(v)}} c(\hat{f}_m^{(t)}(X_i); Y_i) \quad n_v = |D_n^{(v)}| = n - n_t$$

- **cross-validation**: average several hold-out estimators

$$\hat{\mathcal{R}}^{cv}(\hat{f}_m; D_n; (I_j^{(t)})_{1 \leq j \leq V}) = \frac{1}{V} \sum_{j=1}^V \hat{\mathcal{R}}_n^{(v,j)}(\hat{f}_m^{(t,j)}) \quad D_n^{(t,j)} = (X_i, Y_i)_{i \in I_j^{(t)}}$$

Cross-validation

$(X_1, Y_1), \dots, (X_{n_t}, Y_{n_t})$

$(X_{n_t+1}, Y_{n_t+1}), \dots, (X_n, Y_n)$

Training set $D_n^{(t)} \Rightarrow \hat{f}_m^{(t)} = \hat{f}_m(D_n^{(t)})$ Validation set $D_n^{(v)} \Rightarrow$ evaluate risk

- **hold-out** estimator of the risk:

$$\hat{\mathcal{R}}_n^{(v)}(\hat{f}_m^{(t)}) = \frac{1}{n_v} \sum_{(X_i, Y_i) \in D_n^{(v)}} c(\hat{f}_m^{(t)}(X_i); Y_i) \quad n_v = |D_n^{(v)}| = n - n_t$$

- **cross-validation**: average several hold-out estimators

$$\hat{\mathcal{R}}^{cv}(\hat{f}_m; D_n; (I_j^{(t)})_{1 \leq j \leq v}) = \frac{1}{v} \sum_{j=1}^v \hat{\mathcal{R}}_n^{(v,j)}(\hat{f}_m^{(t,j)}) \quad D_n^{(t,j)} = (X_i, Y_i)_{i \in I_j^{(t)}}$$

- **estimator selection**:

$$\hat{m}^{cv}(D_n; (I_j^{(t)})_{1 \leq j \leq v}) \in \operatorname{argmin}_{m \in \mathcal{M}} \left\{ \hat{\mathcal{R}}^{cv}(\hat{f}_m; D_n) \right\} \Rightarrow \hat{f}_{\hat{m}^{cv}(D_n; (I_j^{(t)})_{1 \leq j \leq v})}^{(t)}(D_n)$$

Cross-validation: examples

- Exhaustive data splitting: all possible subsets of size n_t
 \Rightarrow leave-one-out ($n_t = n - 1$)

$$\widehat{\mathcal{R}}^{\text{loo}}(\widehat{f}_m; D_n) = \frac{1}{n} \sum_{j=1}^n c(\widehat{f}_m^{(-j)}(X_j); Y_j)$$

\Rightarrow leave- p -out ($n_t = n - p$)

Cross-validation: examples

- Exhaustive data splitting: all possible subsets of size n_t
 \Rightarrow leave-one-out ($n_t = n - 1$)

$$\widehat{\mathcal{R}}^{\text{loo}}(\widehat{f}_m; D_n) = \frac{1}{n} \sum_{j=1}^n c(\widehat{f}_m^{(-j)}(X_j); Y_j)$$

\Rightarrow leave- p -out ($n_t = n - p$)

- V -fold cross-validation: $\mathcal{B} = (B_j)_{1 \leq j \leq V}$ partition of $\{1, \dots, n\}$

$$\Rightarrow \widehat{\mathcal{R}}^{\text{vf}}(\widehat{f}_m; D_n; \mathcal{B}) = \frac{1}{V} \sum_{j=1}^V \widehat{\mathcal{R}}_n^j(\widehat{f}_m^{(-j)})$$

Cross-validation: examples

- Exhaustive data splitting: all possible subsets of size n_t
 \Rightarrow leave-one-out ($n_t = n - 1$)

$$\widehat{\mathcal{R}}^{\text{loo}}(\widehat{f}_m; D_n) = \frac{1}{n} \sum_{j=1}^n c(\widehat{f}_m^{(-j)}(X_j); Y_j)$$

\Rightarrow leave- p -out ($n_t = n - p$)

- V -fold cross-validation: $\mathcal{B} = (B_j)_{1 \leq j \leq V}$ partition of $\{1, \dots, n\}$

$$\Rightarrow \widehat{\mathcal{R}}^{\text{vf}}(\widehat{f}_m; D_n; \mathcal{B}) = \frac{1}{V} \sum_{j=1}^V \widehat{\mathcal{R}}_n^j(\widehat{f}_m^{(-j)})$$

- Monte-Carlo CV / Repeated learning testing:

$$I_1^{(t)}, \dots, I_V^{(t)} \text{ i.i.d. uniform}$$

Outline

- 1 Definition and Examples
- 2 Cross-validation for risk estimation
- 3 Cross-validation for estimator selection
- 4 Conclusion on CV
- 5 Combining cross-validation with aggregation

Bias of cross-validation

- In this talk, we always assume: $\forall j, \text{Card}(D_n^{(t,j)}) = n_t$
For V -fold CV: $\text{Card}(B_j) = n/V \Rightarrow n_t = n(V-1)/V$.
- Ideal criterion: $\mathcal{R}(\hat{f}_m(D_n))$

Bias of cross-validation

- In this talk, we always assume: $\forall j, \text{Card}(D_n^{(t,j)}) = n_t$
For V -fold CV: $\text{Card}(B_j) = n/V \Rightarrow n_t = n(V-1)/V$.
- Ideal criterion: $\mathcal{R}(\hat{f}_m(D_n))$
- General analysis for the bias:

$$\mathbb{E} \left[\hat{\mathcal{R}}^{\text{cv}} \left(\hat{f}_m; D_n; \left(I_j^{(t)} \right)_{1 \leq j \leq V} \right) \right] = \mathbb{E} \left[\mathcal{R}(\hat{f}_m(D_{n_t})) \right]$$

Bias of cross-validation

- In this talk, we always assume: $\forall j, \text{Card}(D_n^{(t,j)}) = n_t$
For V -fold CV: $\text{Card}(B_j) = n/V \Rightarrow n_t = n(V-1)/V$.
- Ideal criterion: $\mathcal{R}(\hat{f}_m(D_n))$
- General analysis for the bias:

$$\mathbb{E} \left[\hat{\mathcal{R}}^{\text{cv}} \left(\hat{f}_m; D_n; \left(I_j^{(t)} \right)_{1 \leq j \leq V} \right) \right] = \mathbb{E} \left[\mathcal{R}(\hat{f}_m(D_{n_t})) \right]$$

\Rightarrow everything depends on $n \rightarrow \mathbb{E} \left[\mathcal{R}(\hat{f}_m(D_n)) \right]$

Bias of cross-validation

- In this talk, we always assume: $\forall j, \text{Card}(D_n^{(t,j)}) = n_t$
For V -fold CV: $\text{Card}(B_j) = n/V \Rightarrow n_t = n(V-1)/V$.
- Ideal criterion: $\mathcal{R}(\hat{f}_m(D_n))$
- General analysis for the bias:

$$\mathbb{E} \left[\hat{\mathcal{R}}^{\text{cv}} \left(\hat{f}_m; D_n; \left(I_j^{(t)} \right)_{1 \leq j \leq V} \right) \right] = \mathbb{E} \left[\mathcal{R}(\hat{f}_m(D_{n_t})) \right]$$

\Rightarrow everything depends on $n \rightarrow \mathbb{E} \left[\mathcal{R}(\hat{f}_m(D_n)) \right]$

- Note: **bias can be corrected** in some settings (Burman, 1989).
- Note: $D_n \rightarrow \hat{f}_m(D_n)$ must be fixed **before seeing any data**; otherwise (e.g., data-driven model m), stronger bias.

Bias of cross-validation: generic example

Assume:

$$\mathbb{E}\left[\mathcal{R}(\hat{f}_m(D_n))\right] = \alpha(m) + \frac{\beta(m)}{n}$$

(e.g., LS/ridge/ k -NN regression, LS/kernel density estimation).

Bias of cross-validation: generic example

Assume:

$$\mathbb{E} \left[\mathcal{R}(\hat{f}_m(D_n)) \right] = \alpha(m) + \frac{\beta(m)}{n}$$

(e.g., LS/ridge/ k -NN regression, LS/kernel density estimation).

$$\Rightarrow \mathbb{E} \left[\hat{\mathcal{R}}^{\text{cv}} \left(\hat{f}_m; D_n; \left(I_j^{(t)} \right)_{1 \leq j \leq V} \right) \right] = \alpha(m) + \frac{n}{n_t} \frac{\beta(m)}{n}$$

\Rightarrow Bias:

- decreases as a function of n_t ,
- minimal for $n_t = n - 1$,
- negligible if $n_t \sim n$.

\Rightarrow V -fold: bias decreases when V increases, vanishes as $V \rightarrow +\infty$.

Variance of cross-validation: general case

- **Hold-out** (Nadeau & Bengio, 2003):

$$\begin{aligned} \text{var} \left(\widehat{\mathcal{R}}_n^{(v)} \left(\widehat{f}_m^{(t)} \right) \right) &= \frac{1}{n_v} \mathbb{E} \left[\text{var} \left(c(f(X), Y) \mid f = \widehat{f}_m^{(t)} \right) \right] \\ &\quad + \text{var} \left(\mathcal{R} \left(\widehat{f}_m(D_{n_t}) \right) \right) \end{aligned}$$

- **Monte-Carlo CV and number of splits:** ($p = n - n_t$)

$$\begin{aligned} \text{var} \left(\widehat{\mathcal{R}}^{\text{cv}} \left(\widehat{f}_m; D_n; \left(I_j^{(t)} \right)_{1 \leq j \leq v} \right) \right) &= \text{var} \left(\widehat{\mathcal{R}}^{\text{lp}} \left(\widehat{f}_m; D_n \right) \right) \\ &\quad + \underbrace{\frac{1}{V} \mathbb{E} \left[\text{var}_{I^{(t)}} \left(\widehat{\mathcal{R}}_n^{(v)} \left(\widehat{f}_m^{(t)} \right) \mid D_n \right) \right]}_{\text{permutation variance}} \end{aligned}$$

- **V-fold CV:** V , n_t , n_v related
leave-one-out: related to stability? (empirical results)

Variance of V -fold CV criterion

- **Least-squares density estimation** (A. & Lerasle, 2016), exact computation (non-asymptotic):

$$\begin{aligned} \text{var} \left(\widehat{\mathcal{R}}^{\text{vf}} \left(\widehat{f}_m; D_n; \mathcal{B} \right) \right) &= \frac{1 + \mathcal{O}(1)}{n} \text{var}_P(f_m^*) \\ &+ \frac{2}{n^2} \left[1 + \frac{4}{V-1} + \mathcal{O}\left(\frac{1}{V} + \frac{1}{n}\right) \right] A(m) \end{aligned}$$

(simplified formula, histogram model with bin size d_m^{-1} , $A(m) \approx d_m$)

- Linear regression, asymptotic formula (Burman, 1989):

$$\text{var} \left(\widehat{\mathcal{R}}^{\text{vf}} \left(\widehat{f}_m; D_n; \mathcal{B} \right) \right) = \frac{2\sigma^2}{n} + \frac{4\sigma^4}{n^2} \left[4 + \frac{4}{V-1} + \frac{2}{(V-1)^2} + \frac{1}{(V-1)^3} \right] + o(n^{-2})$$

⇒ decreasing with V , dependence only in second order terms.

Outline

- 1 Definition and Examples
- 2 Cross-validation for risk estimation
- 3 **Cross-validation for estimator selection**
- 4 Conclusion on CV
- 5 Combining cross-validation with aggregation

Risk estimation and estimator selection are different goals

$$\hat{m} \in \operatorname{argmin}_{m \in \mathcal{M}} \left\{ \widehat{\mathcal{R}}^{\text{cv}}(\hat{f}_m) \right\} \quad \text{vs.} \quad m^* \in \operatorname{argmin}_{m \in \mathcal{M}} \left\{ \mathcal{R}(\hat{f}_m(D_n)) \right\}$$

- For any Z (deterministic or random),

$$\hat{m} \in \operatorname{argmin}_{m \in \mathcal{M}} \left\{ \widehat{\mathcal{R}}^{\text{cv}}(\hat{f}_m) + Z \right\}$$

⇒ **bias and variance meaningless.**

Risk estimation and estimator selection are different goals

$$\hat{m} \in \operatorname{argmin}_{m \in \mathcal{M}} \left\{ \widehat{\mathcal{R}}^{\text{cv}}(\hat{f}_m) \right\} \quad \text{vs.} \quad m^* \in \operatorname{argmin}_{m \in \mathcal{M}} \left\{ \mathcal{R}(\hat{f}_m(D_n)) \right\}$$

- For any Z (deterministic or random),

$$\hat{m} \in \operatorname{argmin}_{m \in \mathcal{M}} \left\{ \widehat{\mathcal{R}}^{\text{cv}}(\hat{f}_m) + Z \right\}$$

⇒ bias and variance meaningless.

- Perfect ranking among $(\hat{f}_m)_{m \in \mathcal{M}} \Leftrightarrow \forall m, m' \in \mathcal{M},$

$$\operatorname{sign}(\widehat{\mathcal{R}}^{\text{cv}}(\hat{f}_m) - \widehat{\mathcal{R}}^{\text{cv}}(\hat{f}_{m'})) = \operatorname{sign}(\mathcal{R}(\hat{f}_m) - \mathcal{R}(\hat{f}_{m'}))$$

⇒ $\mathbb{E}[\widehat{\mathcal{R}}^{\text{cv}}(\hat{f}_m) - \widehat{\mathcal{R}}^{\text{cv}}(\hat{f}_{m'})]$ should be of the good sign (unbiased risk estimation heuristic: AIC, C_p , leave-one-out...)

⇒ $\operatorname{var}(\widehat{\mathcal{R}}^{\text{cv}}(\hat{f}_m) - \widehat{\mathcal{R}}^{\text{cv}}(\hat{f}_{m'}))$ should be minimal (detailed heuristic: A. & Lerasle, 2016)

CV with an estimation goal: the big picture (\mathcal{M} “small”)

- At first order, the **bias drives the performance** of:
 - leave- p -out, V -fold CV,
 - Monte-Carlo CV if $V \gg n^2$
or if n_V large enough (including hold-out)
- CV performs similarly to

$$\operatorname{argmin}_{m \in \mathcal{M}} \left\{ \mathbb{E} \left[\mathcal{R}(\hat{f}_m(D_{n_t})) \right] \right\}$$

CV with an estimation goal: the big picture (\mathcal{M} “small”)

- At first order, the bias drives the performance of:
 - leave- p -out, V -fold CV,
 - Monte-Carlo CV if $V \gg n^2$
or if n_V large enough (including hold-out)
- CV performs similarly to

$$\operatorname{argmin}_{m \in \mathcal{M}} \left\{ \mathbb{E} \left[\mathcal{R}(\hat{f}_m(D_{n_t})) \right] \right\}$$

⇒ first-order optimality if $n_t \sim n$

⇒ suboptimal otherwise

e.g., V -fold CV with V fixed.

- Theoretical results for least-squares regression and density estimation at least.

Bias-corrected VFCV / V-fold penalization

- Bias-corrected V-fold CV (Burman, 1989):

$$\begin{aligned}\widehat{\mathcal{R}}^{\text{vf,corr}}(\widehat{f}_m; D_n; \mathcal{B}) &:= \widehat{\mathcal{R}}^{\text{vf}}(\widehat{f}_m; D_n; \mathcal{B}) + \widehat{\mathcal{R}}_n(\widehat{f}_m) - \frac{1}{V} \sum_{j=1}^V \widehat{\mathcal{R}}_n(\widehat{f}_m^{(-j)}) \\ &= \widehat{\mathcal{R}}_n(\widehat{f}_m(D_n)) + \underbrace{\text{pen}_{\text{VF}}(\widehat{f}_m; D_n; \mathcal{B})}_{\text{V-fold penalty (A. 2008)}}.\end{aligned}$$

Bias-corrected VFCV / V-fold penalization

- Bias-corrected V-fold CV (Burman, 1989):

$$\begin{aligned}\widehat{\mathcal{R}}^{\text{vf,corr}}(\widehat{f}_m; D_n; \mathcal{B}) &:= \widehat{\mathcal{R}}^{\text{vf}}(\widehat{f}_m; D_n; \mathcal{B}) + \widehat{\mathcal{R}}_n(\widehat{f}_m) - \frac{1}{V} \sum_{j=1}^V \widehat{\mathcal{R}}_n(\widehat{f}_m^{(-j)}) \\ &= \widehat{\mathcal{R}}_n(\widehat{f}_m(D_n)) + \underbrace{\text{pen}_{\text{VF}}(\widehat{f}_m; D_n; \mathcal{B})}_{\text{V-fold penalty (A. 2008)}}.\end{aligned}$$

- In least-squares density estimation (A. & Lerasle, 2016):

$$\begin{aligned}\widehat{\mathcal{R}}^{\text{vf}}(\widehat{f}_m; D_n; \mathcal{B}) &= \widehat{\mathcal{R}}_n(\widehat{f}_m(D_n)) + \underbrace{\left(1 + \frac{1}{2(V-1)}\right)}_{\text{overpenalization factor}} \text{pen}_{\text{VF}}(\widehat{f}_m; D_n; \mathcal{B}) \\ \widehat{\mathcal{R}}^{\text{lp}}(\widehat{f}_m; D_n; \mathcal{B}) &= \widehat{\mathcal{R}}_n(\widehat{f}_m(D_n)) + \underbrace{\left(1 + \frac{1}{2\left(\frac{n}{p} - 1\right)}\right)}_{\text{overpenalization factor}} \text{pen}_{\text{VF}}(\widehat{f}_m; D_n; \mathcal{B}_{\text{loo}}).\end{aligned}$$

Variance and estimator selection

$$\Delta(m, m', V) = \widehat{\mathcal{R}}^{\text{vf,corr}}(\widehat{f}_m) - \widehat{\mathcal{R}}^{\text{vf,corr}}(\widehat{f}_{m'})$$

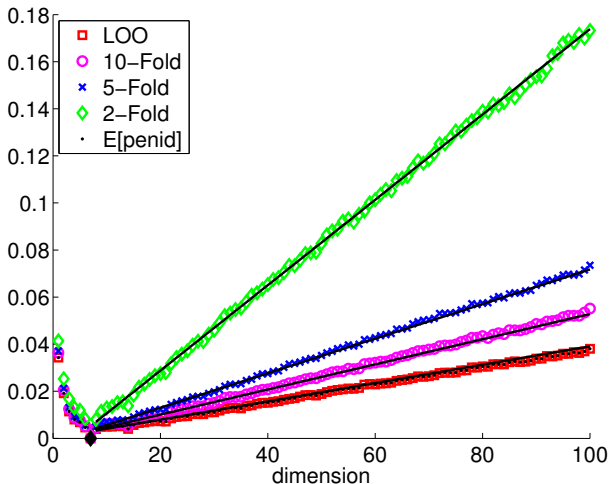
Theorem (A. & Lerasle, 2016, least-squares density estimation)

$$\begin{aligned} \text{var}(\Delta(m, m', V)) &= 4 \left(1 + \frac{2}{n} + \frac{1}{n^2} \right) \frac{\text{var}_P(f_m^* - f_{m'}^*)}{n} \\ &\quad + 2 \left(1 + \frac{4}{V-1} - \frac{1}{n} \right) \underbrace{\frac{B(m, m')}{n^2}}_{\geq 0} \end{aligned}$$

If $S_m \subset S_{m'}$ are two histogram models with constant bin sizes $d_m^{-1}, d_{m'}^{-1}$, then, $B(m, m') \propto \|f_m^* - f_{m'}^*\| d_m$.

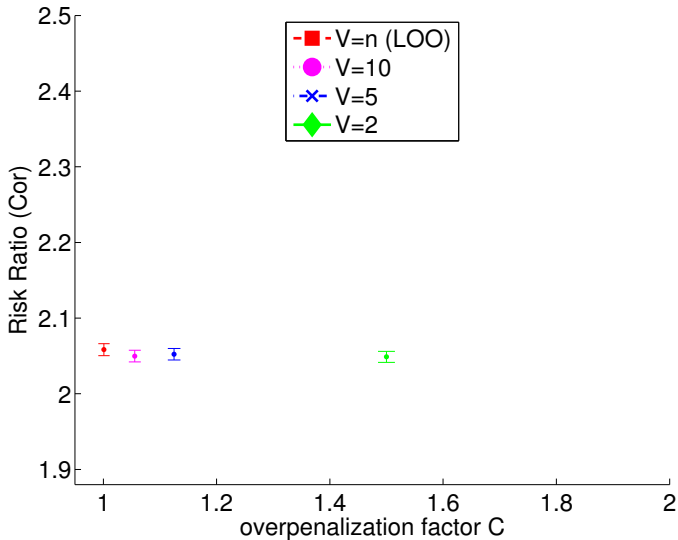
The two terms are of the same order if $\|f_m^* - f_{m'}^*\| \approx d_m/n$.

Variance of $\widehat{\mathcal{R}}^{\text{vf,corr}}(\widehat{f}_m) - \widehat{\mathcal{R}}^{\text{vf,corr}}(\widehat{f}_{m^*})$ vs. (d_m, V)

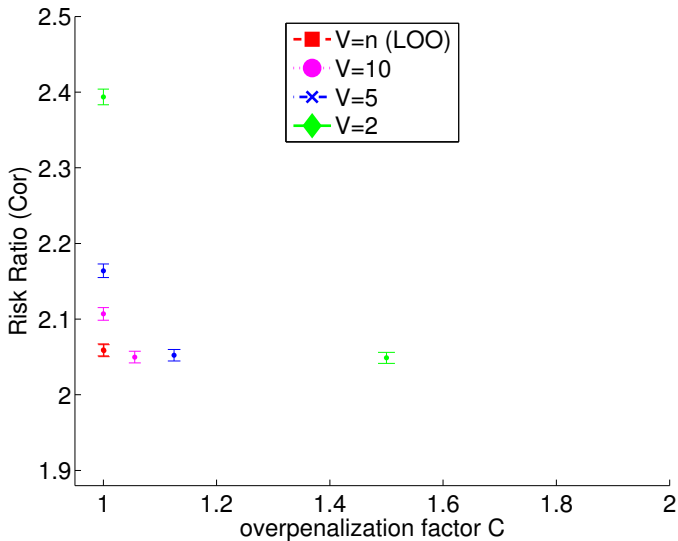


$$\text{var}(\Delta(m, m', V)) \approx n^{-2} \left[29 \left(1 + \frac{0.8}{V-1} \right) + 3.7 \left(1 + \frac{3.8}{V-1} \right) (d_m - d_{m^*}) \right]$$

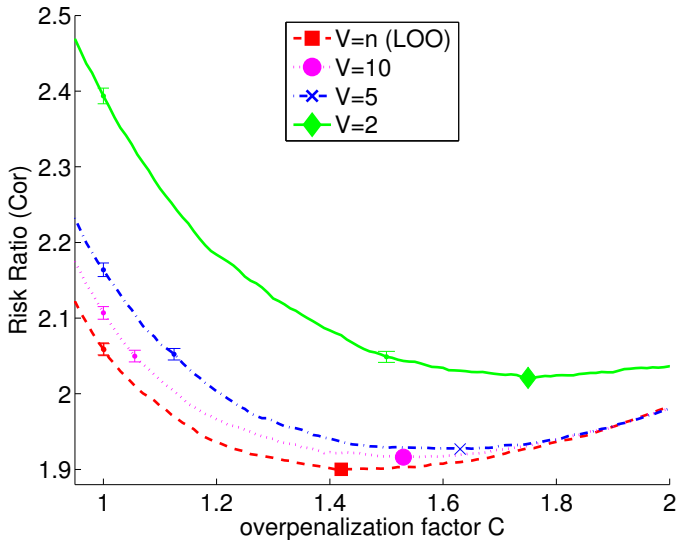
Experiment (LS density estimation): V -fold CV



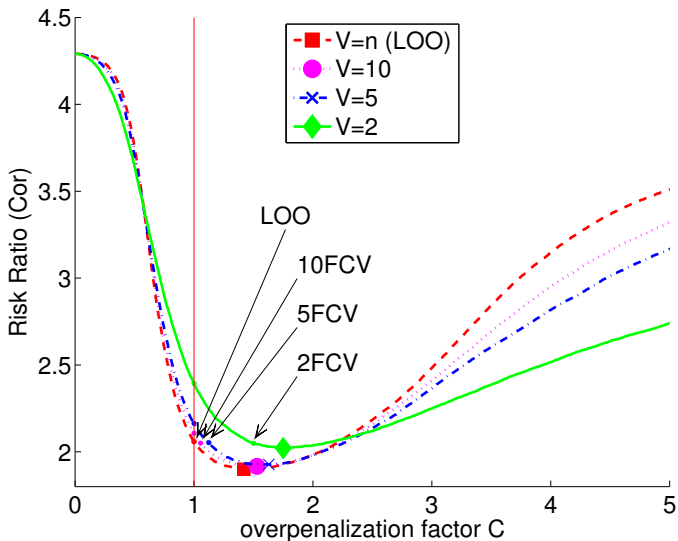
Experiment (LS density estimation): V -fold penalization



Experiment (LS density estimation): overpenalization



Experiment (LS density estimation): conclusion



Outline

- 1 Definition and Examples
- 2 Cross-validation for risk estimation
- 3 Cross-validation for estimator selection
- 4 Conclusion on CV
- 5 Combining cross-validation with aggregation

Estimator selection with V -fold: conclusion

- **Computational complexity:** $\mathcal{O}(V)$ in general

Estimator selection with V -fold: conclusion

- **Computational complexity:** $\mathcal{O}(V)$ in general
 - **V -fold cross-validation:**
 - **Bias:** decreases with V / can be removed
 - **Variance:** decreases with V / almost minimal with $V \in [5, 10]$
- ⇒ best performance for the largest V and **almost optimal with $V = 10$...**

Estimator selection with V -fold: conclusion

- **Computational complexity:** $\mathcal{O}(V)$ in general
 - **V -fold cross-validation:**
 - Bias: decreases with V / can be removed
 - Variance: decreases with V / almost minimal with $V \in [5, 10]$
- ⇒ best performance for the largest V and **almost optimal with $V = 10$...**
- ... **if optimal overpenalization factor $C^* \approx 1$ (various behaviours possible).**

Estimator selection with V -fold: conclusion

- **Computational complexity:** $\mathcal{O}(V)$ in general
- **V -fold cross-validation:**
 - Bias: decreases with V / can be removed
 - Variance: decreases with V / almost minimal with $V \in [5, 10]$ \Rightarrow best performance for the largest V and almost optimal with $V = 10 \dots$
... if optimal overpenalization factor $C^* \approx 1$ (various behaviours possible).
- **V -fold penalization:**
 - **Decoupling** of bias and variance \Rightarrow **easier to understand.**
 - Bias: **chosen directly** through C , **without any constraint.**
 - Variance: decreases with V / **almost minimal with $V \in [5, 10]$.**

How general are these conclusions? (i.i.d. case)

- At least valid for least-square regression / density estimation, kernel density estimation.
- **Bias-correction** / V -fold penalization: valid if

$$\mathbb{E}\left[(\mathcal{R} - \widehat{\mathcal{R}}_n)(\widehat{f}_m)\right] \approx \frac{\gamma(m)}{n} .$$

Otherwise: use repeated V -fold or Monte-Carlo CV with a well-chosen n_t .

- **Variance**: different behaviours can occur in other settings (experiments).
- Everything can be **checked on synthetic data**: plot

$$n \rightarrow \mathbb{E}\left[\mathcal{R}(\widehat{f}_m(D_n))\right] \quad \text{and} \quad m \rightarrow \text{var}\left(\widehat{\mathcal{R}}^{\text{cv}}(\widehat{f}_m) - \widehat{\mathcal{R}}^{\text{cv}}(\widehat{f}_{m^*})\right) .$$

Dependent data

- $D_n^{(t)}, D_n^{(v)}$ dependent \Rightarrow CV heuristic fails!

\Rightarrow possible troubles for risk estimation (Hart & Wehrly, 1986; Opsomer et al., 2001).

Dependent data

- $D_n^{(t)}, D_n^{(v)}$ dependent \Rightarrow CV heuristic fails!

\Rightarrow possible troubles for risk estimation (Hart & Wehrly, 1986; Opsomer et al., 2001).

- **Solution for short-term dependence:**
remove some data at each split \Rightarrow gap between training and validation samples.

Cross-validation with an identification goal

- **Main change**: value of the optimal overpenalization factor C^* , often $C^* \rightarrow +\infty$ when $n \rightarrow +\infty$.
- ⇒ **Cross-validation paradox** (Yang, 2006, 2007): $n_t \ll n$ can be necessary!
 - Why? Smaller $n_t \Rightarrow$ easier to distinguish the two best procedures... **if** n_t large enough (asymptotic regime).
 - Remark: **estimation goal, parametric setting** \Rightarrow similar behaviour.

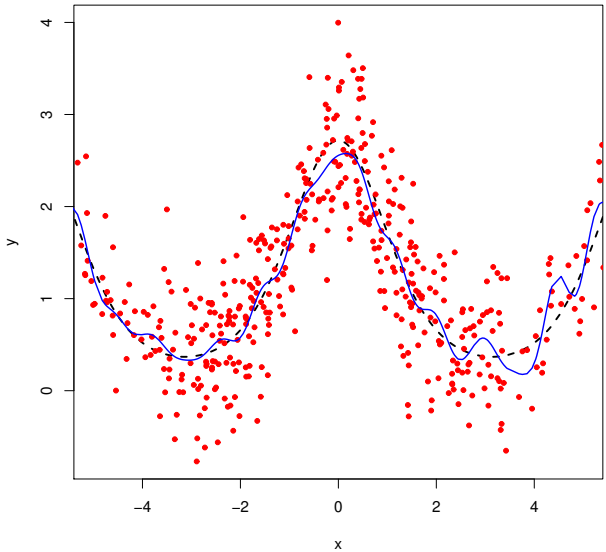
Large collection of estimators/models

- Estimator/model selection with an “**exponential** collection” (implicitly excluded in all results above).
⇒ Expectations do not drive the first order!
- Examples: variable selection with $p \geq n$ variables, change-point detection.
- **Solution: group the models** ⇒ one estimator per “dimension” (e.g., empirical risk minimizer)
works for change-point detection (A. & Celisse, 2010).

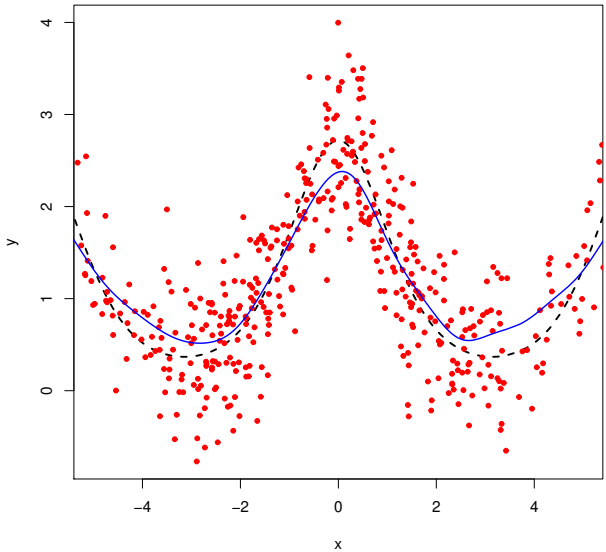
Outline

- 1 Definition and Examples
- 2 Cross-validation for risk estimation
- 3 Cross-validation for estimator selection
- 4 Conclusion on CV
- 5 Combining cross-validation with aggregation

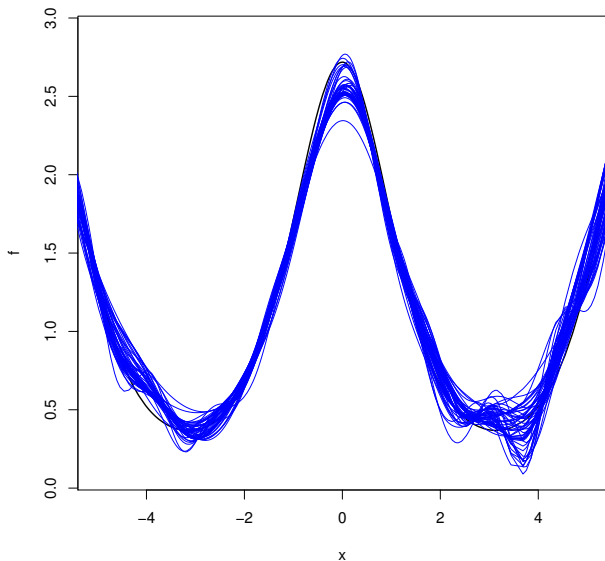
Example: regression, ϵ -SVM estimator (undersmoothed)



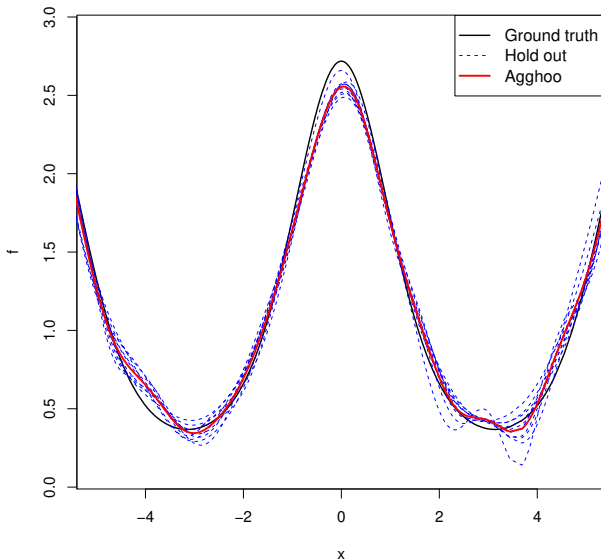
Example: regression, ϵ -SVM estimator (oversmoothed)



Example: regression, ϵ -SVM: several hold-out estimators



Example: regression, ϵ -SVM: aggregated hold-out



Aggregated hold-out (Agghoo): definition

- Idea: **aggregate several hold-out estimators.**
- If \mathcal{Y} is convex (e.g., regression):

$$\hat{f}^{agghoo} = \frac{1}{V} \sum_{j=1}^V \hat{f}_{\hat{m}^{ho}(I_j^{(t)})} (D_n^{(t,j)})$$

Aggregated hold-out (Agghoo): definition

- Idea: **aggregate several hold-out estimators.**
- If \mathcal{Y} is convex (e.g., regression):

$$\hat{f}^{agghoo} = \frac{1}{V} \sum_{j=1}^V \hat{f}_{\hat{m}^{ho}(I_j^{(t)})} (D_n^{(t,j)})$$

- If \mathcal{Y} is finite (classification):

$$\hat{f}^{agghoo} : x \mapsto \text{majority vote among } \left\{ \hat{f}_{\hat{m}^{ho}(I_j^{(t)})} (x; D_n^{(t,j)}) / j = 1, \dots, V \right\}$$

Aggregated hold-out (Agghoo): definition

- Idea: **aggregate several hold-out estimators.**
- If \mathcal{Y} is convex (e.g., regression):

$$\hat{f}^{agghoo} = \frac{1}{V} \sum_{j=1}^V \hat{f}_{\hat{m}^{ho}(I_j^{(t)})} (D_n^{(t,j)})$$

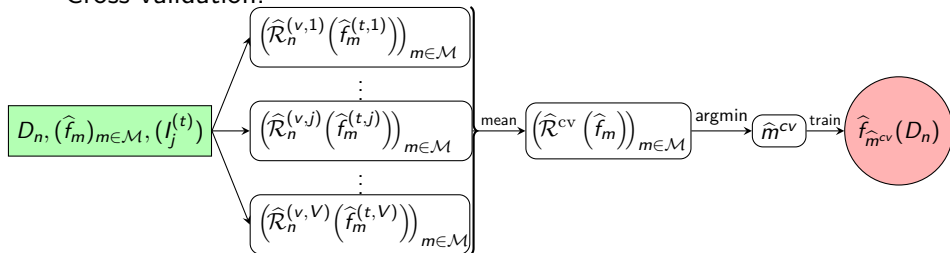
- If \mathcal{Y} is finite (classification):

$$\hat{f}^{agghoo} : x \mapsto \text{majority vote among } \left\{ \hat{f}_{\hat{m}^{ho}(I_j^{(t)})} (x; D_n^{(t,j)}) / j = 1, \dots, V \right\}$$

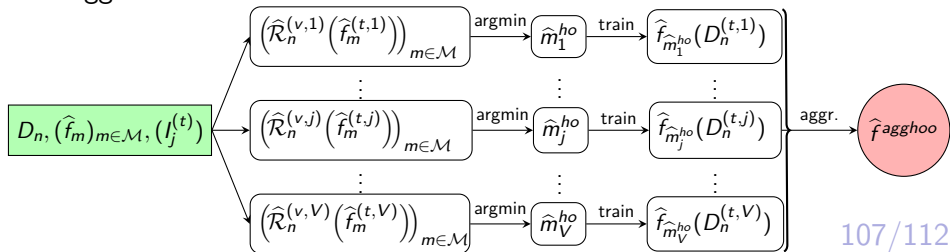
- Usual assumption: $\forall j \in \{1, \dots, V\}, \text{Card}(I_j^{(t)}) = \tau n$.
- Remark: $V = 1 \Rightarrow \hat{f}^{agghoo} = \hat{f}_{\hat{m}^{ho}}(D_n^t) \approx \text{hold-out estimator}$

Agghoo and cross-validation

Cross-validation:



Agghoo:



Related procedures

- **CV bagging** (data science folklore)
 - hold-out + subbagging \neq agghoo: for $j = 1, \dots, V$,

hold-out + subbagging = hold-out on subsamples

$$\underbrace{(X_{i_1}, Y_{i_1}), \dots, (X_{i_k}, Y_{i_k})}_{\text{train}}, \underbrace{(X_{i_{k+1}}, Y_{i_{k+1}}), \dots, (X_{i_\ell}, Y_{i_\ell})}_{\text{validation}}, \underbrace{(X_{i_{\ell+1}}, Y_{i_{\ell+1}}), \dots, (X_{i_n}, Y_{i_n})}_{\text{unused}}$$

agghoo = hold-out on different splits

$$\underbrace{(X_{i_1}, Y_{i_1}), \dots, (X_{i_k}, Y_{i_k}), (X_{i_{k+1}}, Y_{i_{k+1}}), \dots, (X_{i_\ell}, Y_{i_\ell})}_{\text{train}}, \underbrace{(X_{i_{\ell+1}}, Y_{i_{\ell+1}}), \dots, (X_{i_n}, Y_{i_n})}_{\text{validation}}$$

- “CV bagging” also used for procedures close to agghoo
- **Averaging of the chosen parameters \hat{m}_j^{ho} :**
 - K -fold averaging cross-validation (ACV; Jung and Hu, 2015)
 - efficient K -fold cross-validation (EKCV; Jung, 2016)

Performance of agghoo: theory

- **Regression with c convex**: if $\forall j, \text{Card}(I_j^{(t)}) = \tau n$,

$$\forall V \geq 1, \quad \mathbb{E} \left[\mathcal{R} \left(\hat{f}^{\text{agghoo}} \left((\hat{f}_m)_{m \in \mathcal{M}}; D_n; (I_j^{(t)})_{1 \leq j \leq V} \right) \right) \right] \\ \leq \mathbb{E} \left[\mathcal{R} \left(\hat{f}_{\hat{m}^{\text{ho}}} \left((\hat{f}_m)_{m \in \mathcal{M}}; D_n; I_1^{(t)} \right) (D_n^{t,1}) \right) \right]$$

Corollary: oracle inequalities for the hold-out \Rightarrow oracle inequalities for agghoo (sanity check)

- **Binary classification, 0–1 risk** (Maillard, A. & Lerasle, 2017): same with an **additional factor 2**

Performance of agghoo: theory

- **Regression with c convex**: if $\forall j, \text{Card}(I_j^{(t)}) = \tau n$,

$$\forall V \geq 1, \quad \mathbb{E} \left[\mathcal{R} \left(\hat{f}^{\text{agghoo}} \left((\hat{f}_m)_{m \in \mathcal{M}}; D_n; (I_j^{(t)})_{1 \leq j \leq V} \right) \right) \right] \\ \leq \mathbb{E} \left[\mathcal{R} \left(\hat{f}_{\hat{m}^{\text{ho}}} \left((\hat{f}_m)_{m \in \mathcal{M}}; D_n; I_1^{(t)} \right) (D_n^{t,1}) \right) \right]$$

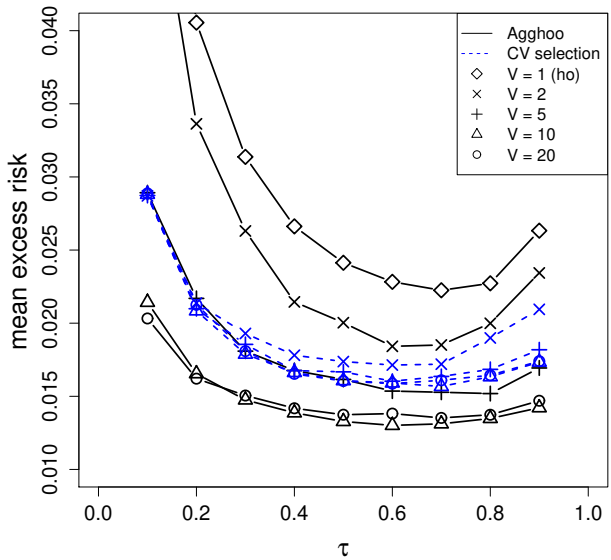
Corollary: oracle inequalities for the hold-out \Rightarrow oracle inequalities for agghoo (sanity check)

- **Binary classification, 0–1 risk** (Maillard, A. & Lerasle, 2017): same with an **additional factor 2**
- Maillard (2020): in a specific setting (projection estimators on a trigonometric basis, density estimation), proof that

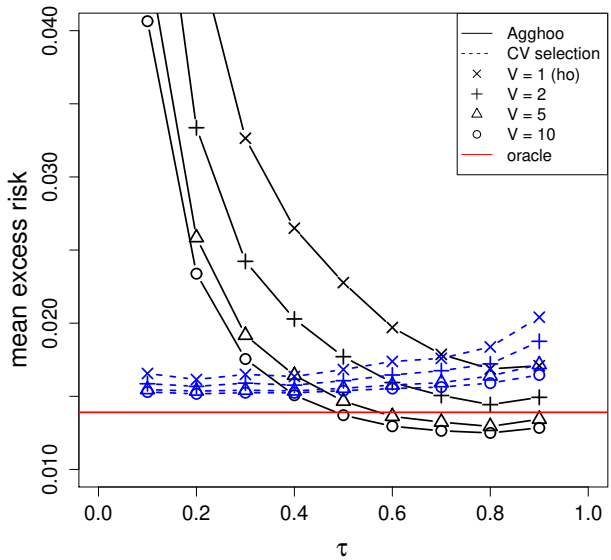
$$\ell(\hat{f}^{\text{agghoo}}, f^*) \leq C \inf_{m \in \mathcal{M}} \left\{ \ell(\hat{f}_m, f^*) \right\} \quad \text{with } C < 1$$

\Rightarrow better than **any** selection procedure, even the oracle!

Numerical experiments: 0–1 binary classification, k -NN



Numerical experiments: regression, L^1 loss, ϵ -SVM



Questions?