Optimal data-driven estimator selection with minimal penalties

Sylvain Arlot (joint works with F. Bach, P. Massart & M. Solnon)

[1] CNRS

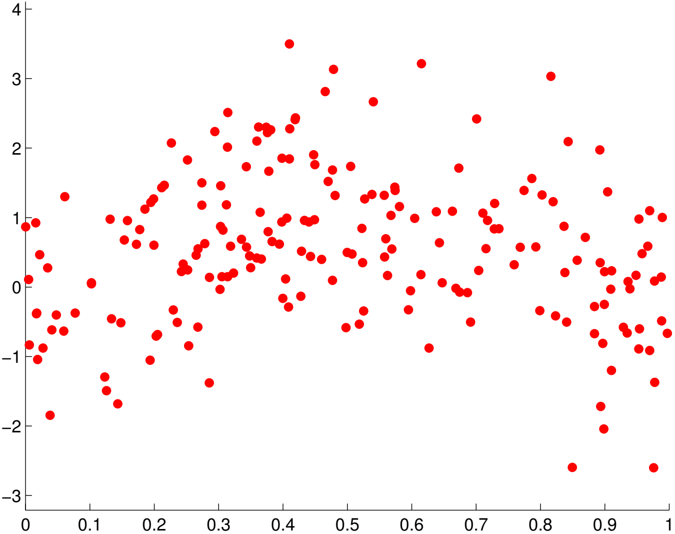
[2] École Normale Supérieure (Paris), DI/ENS, Équipe SIERRA

CIRM, December, 12–13, 2013

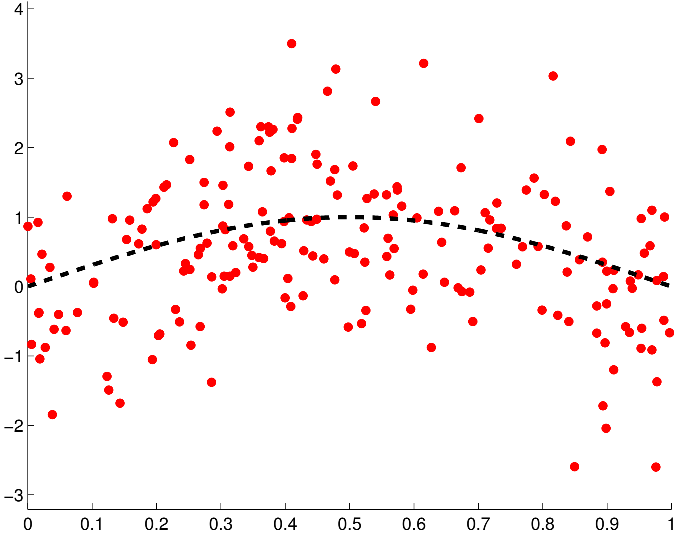
Plan

Outline

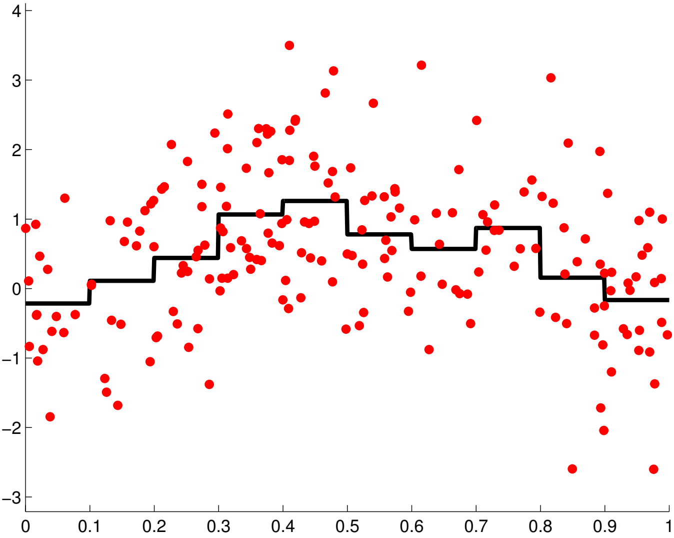
Regression: data $(x_1, Y_1), \ldots, (x_n, Y_n)$



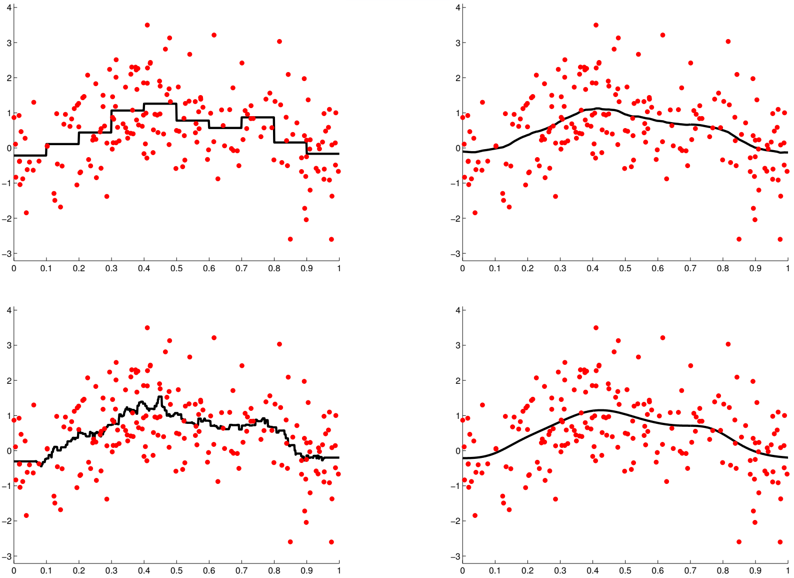
Goal: find the signal (denoising)



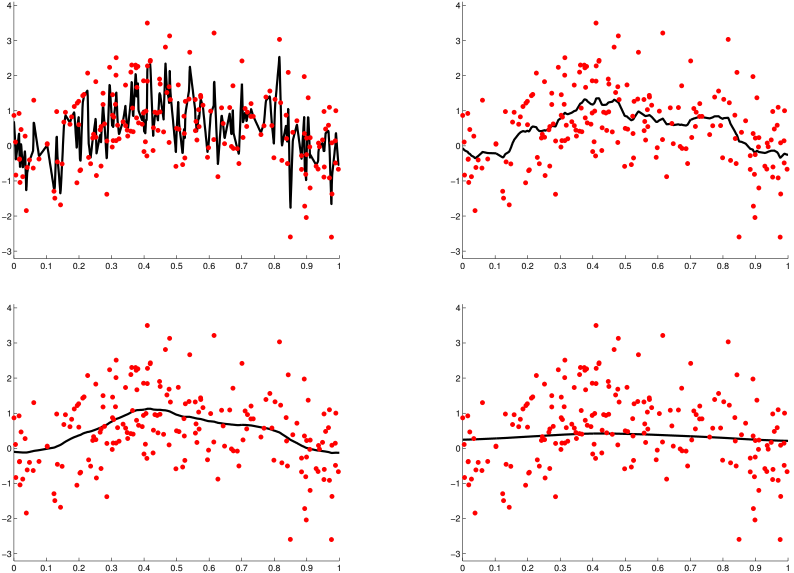
Estimators: example: regressogram



Estimators: regressogram, ridge, k-NN, NW



Estimator selection: kernel ridge



Estimator selection

- Estimator collection $(\widehat{F}_m)_{m \in \mathcal{M}} \Rightarrow \widehat{m}(Y)$?
- Goal: minimize the risk

Estimator selection

- Estimator collection $(\widehat{F}_m)_{m \in \mathcal{M}} \Rightarrow \widehat{m}(Y)$?
- Goal: minimize the risk
- 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 $(\widehat{F}_m)_{m \in \mathcal{M}} \Rightarrow \widehat{m}(Y)$?
- Goal: minimize the risk
- 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?
- Classical approaches and their limitations:
  - cross-validation: computational cost
  - penalization: unknown constants
  - elbow heuristics: no clear definition/justification

Penalties known up to a constant factor

$$\widehat{m}(Y) \in \operatorname*{argmin}_{m \in \mathcal{M}} \left\{ \text{Emp. risk}(\widehat{F}_m) + \text{pen}(m) \right\}$$

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

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

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

Penalties known up to a constant factor

$$\widehat{m}(Y) \in \operatorname*{argmin}_{m \in \mathcal{M}} \left\{ \text{Emp. risk}(\widehat{F}_m) + \text{pen}(m) \right\}$$

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

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

- 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, ...)

Penalties known up to a constant factor

$$\widehat{m}(Y) \in \operatorname*{argmin}_{m \in \mathcal{M}} \left\{ \text{Emp. risk}(\widehat{F}_m) + \text{pen}(m) \right\}$$

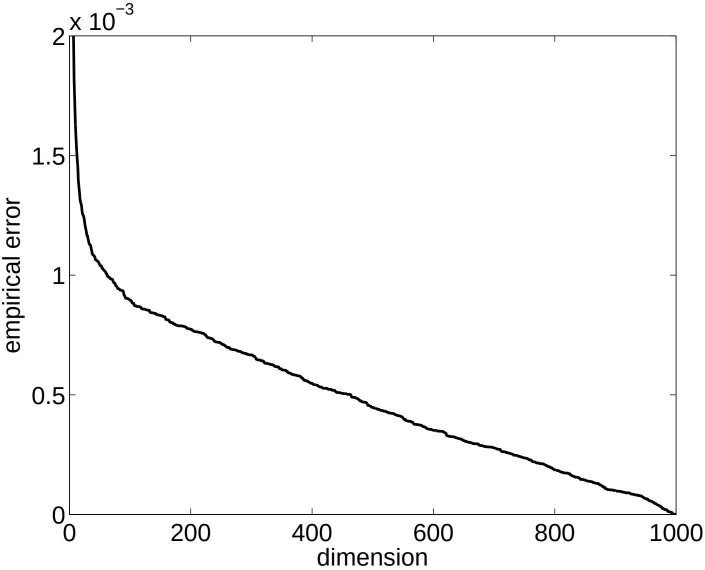
- Optimal penalties depending on the noise level σ^2 (Mallows, 1973):

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

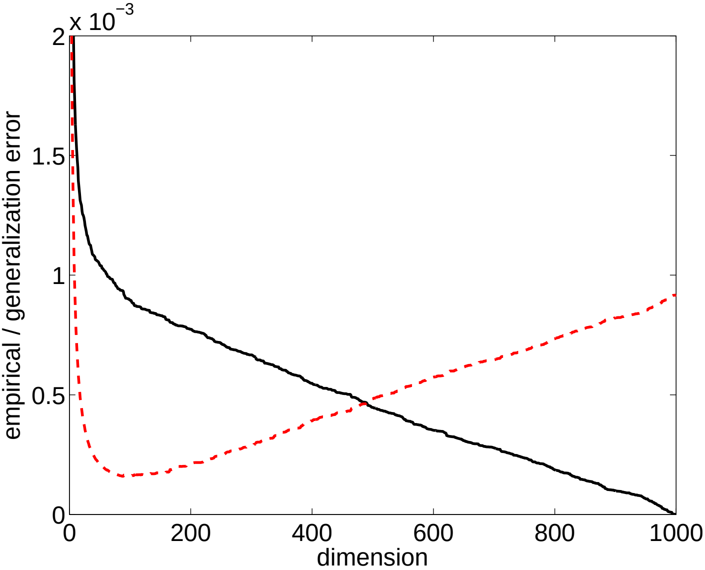
- Optimal penalty known asymptotically (AIC; Akaike, 1973)
- Resampling-based penalties
- Optimal constant unknown even in theory (change-point detection, mixture models, global/local Rademacher complexities, ...)

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

"L-curve" and elbow heuristics?



"L-curve" and elbow heuristics?



Outline

Statistical framework: regression, least-squares loss

- Observations: $Y = (Y_1, \ldots, Y_n) \in \mathbb{R}^n$

$$Y_i = F_i + \varepsilon_i \qquad \text{(e.g., } F_i = F(x_i)\text{)}$$

with $Y_i \in \mathbb{R}$, $(\varepsilon_i)_{1 \leq i \leq n}$ i.i.d.

Statistical framework: regression, least-squares loss

- Observations: $Y = (Y_1, \ldots, Y_n) \in \mathbb{R}^n$

$$Y_i = F_i + \varepsilon_i \qquad (\text{e.g., } F_i = F(x_i))$$

with $Y_i \in \mathbb{R}$, $(\varepsilon_i)_{1 \le i \le n}$ i.i.d.

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

Statistical framework: regression, least-squares loss

- Observations: $Y = (Y_1, \ldots, Y_n) \in \mathbb{R}^n$

$$Y_i = F_i + \varepsilon_i \qquad \text{(e.g., } F_i = F(x_i)\text{)}$$

with $Y_i \in \mathbb{R}$, $(\varepsilon_i)_{1 \le i \le n}$ i.i.d.

- Fixed design: $x_i \in \mathcal{X}$ deterministic
- Least-squares loss of a predictor $t \in \mathbb{R}^n$ ("$t_i = t(x_i)$"):

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

Statistical framework: regression, least-squares loss

- Observations: $Y = (Y_1, \ldots, Y_n) \in \mathbb{R}^n$

$$Y_i = F_i + \varepsilon_i \qquad (\text{e.g., } F_i = F(x_i))$$

with $Y_i \in \mathbb{R}$, $(\varepsilon_i)_{1 \le i \le n}$ i.i.d.

- Fixed design: $x_i \in \mathcal{X}$ deterministic
- Least-squares loss of a predictor $t \in \mathbb{R}^n$ ("$t_i = t(x_i)$"):

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

$\Rightarrow$ Estimator $\widehat{F}(Y) \in \mathbb{R}^n$?

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 $\frac{1}{n}\left\lVert t - F \right\rVert^2$
- Least-squares criterion:

$$\frac{1}{n}\left\lVert t - Y \right\rVert^2 = \frac{1}{n}\sum_{i=1}^{n}\left(t_i - Y_i \right)^2$$

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

Least-squares estimators

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

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

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

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

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

so that

$$\widehat{F}_S = \Pi_S(Y) \qquad \text{(orthogonal projection)}$$

Model examples

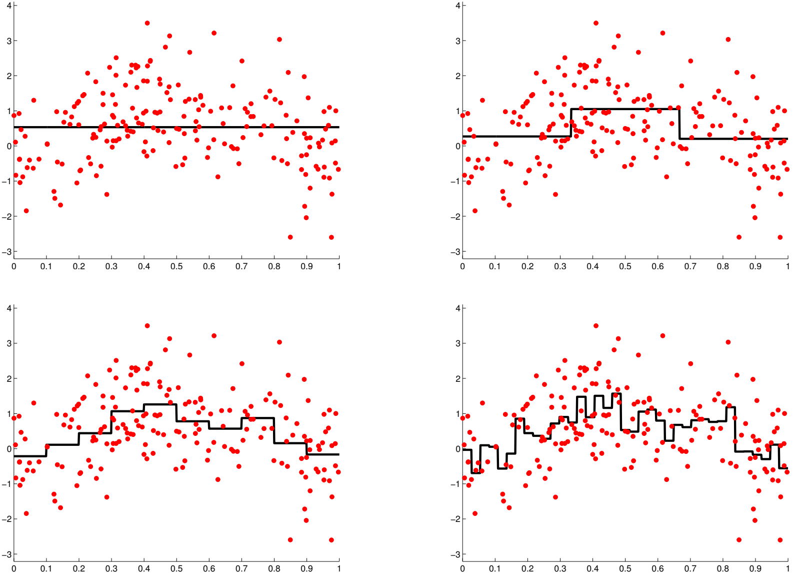
- histograms on some partition Λ of $\mathcal{X}$
  $\Rightarrow$ the least-squares estimator (regressogram) can be written

$$\widehat{F}_m(x_i) = \sum_{\lambda \in \Lambda} \widehat{\beta}_\lambda \mathbb{1}_{x_i \in \lambda} \qquad \widehat{\beta}_\lambda = \frac{1}{\operatorname{Card}\{x_i \in \lambda\}} \sum_{x_i \in \lambda} Y_i$$

- subspace generated by a subset of an orthogonal basis of $L^2(\mu)$ (Fourier, wavelets, ...)
- variable selection: $x_i = \left(x_i^{(1)}, \ldots, x_i^{(p)}\right) \in \mathbb{R}^p$ gathers p variables that can (linearly) explain Y_i

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

Model selection: regular regressograms



Model selection

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

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

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

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

Bias-variance trade-off

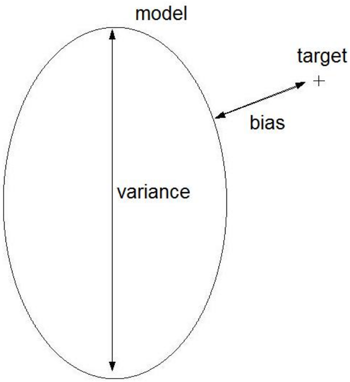
$$\mathbb{E}\left[\frac{1}{n}\left\|\widehat{F}_m - F\right\|^2\right] = \text{Bias} + \text{Variance}$$

Bias or Approximation error

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

Variance or Estimation error

$$\frac{\sigma^2 \dim(S_m)}{n}$$



Bias-variance trade-off

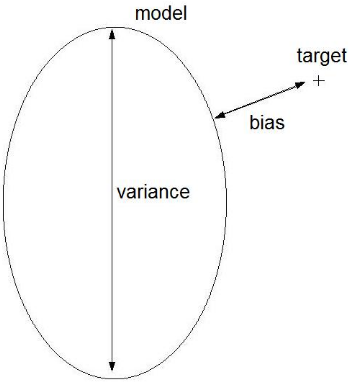
$$\mathbb{E}\left[\frac{1}{n}\left\|\widehat{F}_m - F\right\|^2\right] = \text{Bias} + \text{Variance}$$

Bias or Approximation error

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

Variance or Estimation error

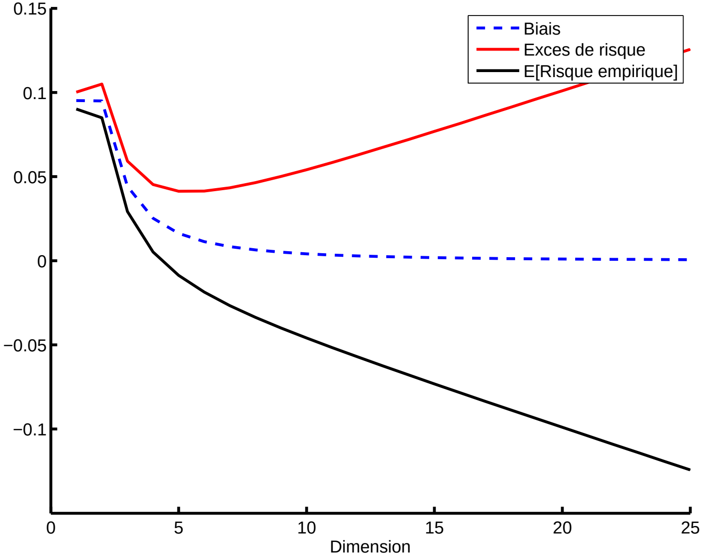
$$\frac{\sigma^2 \dim(S_m)}{n}$$



Bias-variance trade-off

⇔ avoid overfitting and underfitting

Why should the empirical risk be penalized?



Penalization

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

Penalization

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

- Ideal penalty:

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

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

Penalization

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

- Ideal penalty:

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

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

$\Rightarrow$ C_p : $2\sigma^2 D_m / n$ (Mallows, 1973)

Oracle inequality

Theorem (Birgé & Massart 2007, reformulated)

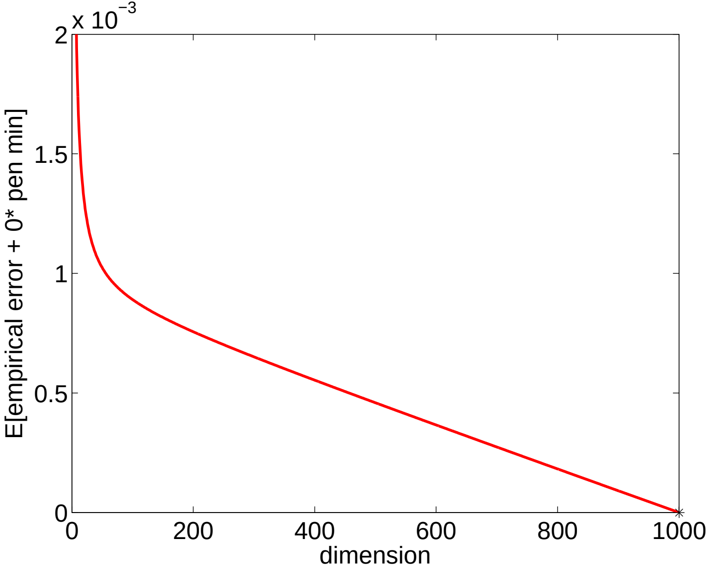
Assumptions:

- $\operatorname{pen}(m) = \frac{2\sigma^2 D_m}{n}$
- *Gaussian noise:* $\varepsilon_i \sim \mathcal{N}(0, \sigma^2)$

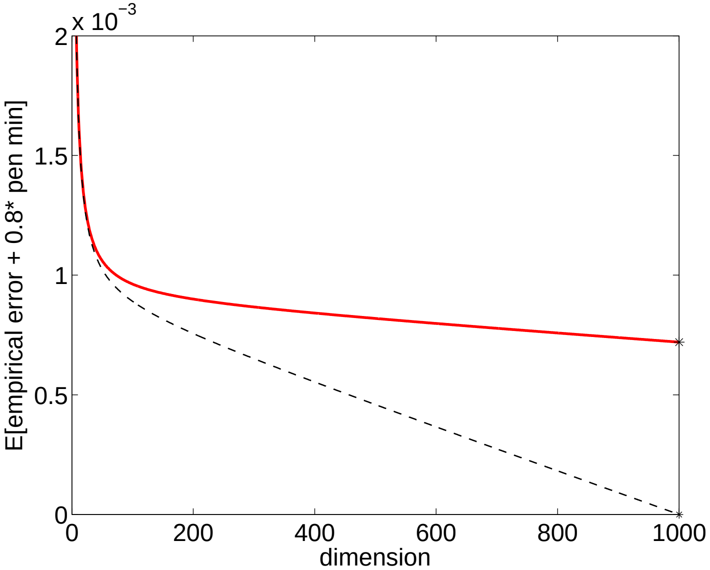
Then, for every $\gamma > 0$*, with probability at least* $1 - 4 \operatorname{Card}(\mathcal{M}) n^{-\gamma}$*, if* $n \geq n_0(\gamma)$*, for every* $\eta \in (0,1)$*,*

$$\frac{1}{n} \left\| \widehat{F}_{\widehat{m}} - F \right\|^2 \leq (1+\eta) \inf_{m \in \mathcal{M}} \left\{ \frac{1}{n} \left\| \widehat{F}_m - F \right\|^2 \right\} + \frac{80 \gamma \log(n) \sigma^2}{\eta n}$$

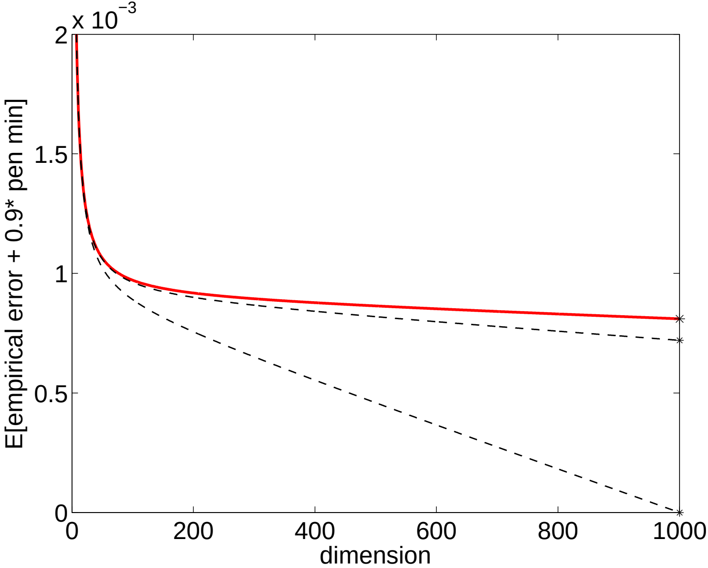
$\mathbb{E}$[Empirical risk] $+ 0 \times \sigma^2 D_m n^{-1}$ (OLS)



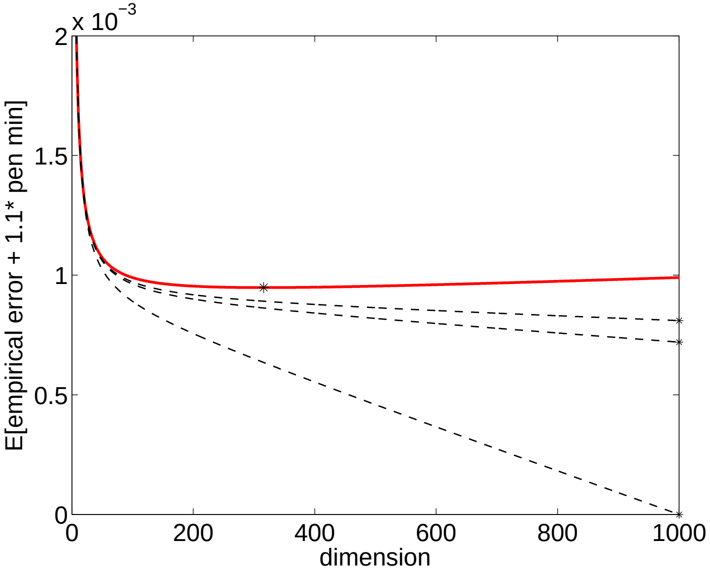
$\mathbb{E}$[Empirical risk] $+ 0.8 \times \sigma^2 D_m n^{-1}$ (OLS)



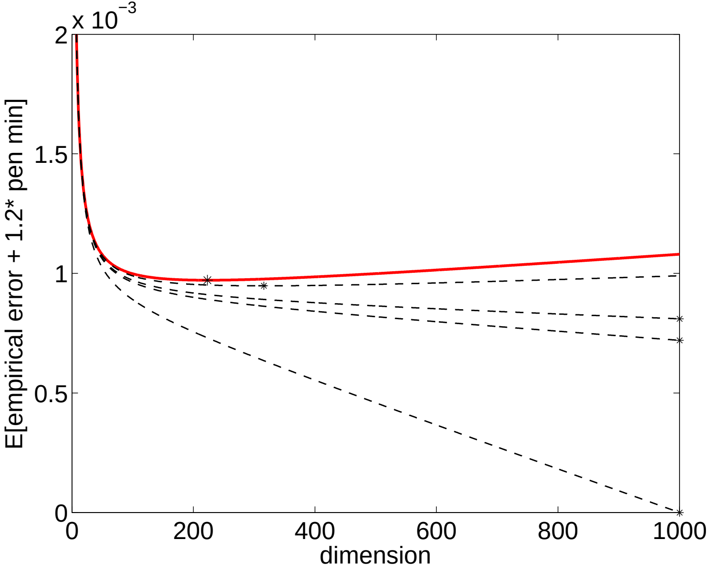
$\mathbb{E}[$ Empirical risk $] + 0.9 \times \sigma^2 D_m n^{-1}$ (OLS)



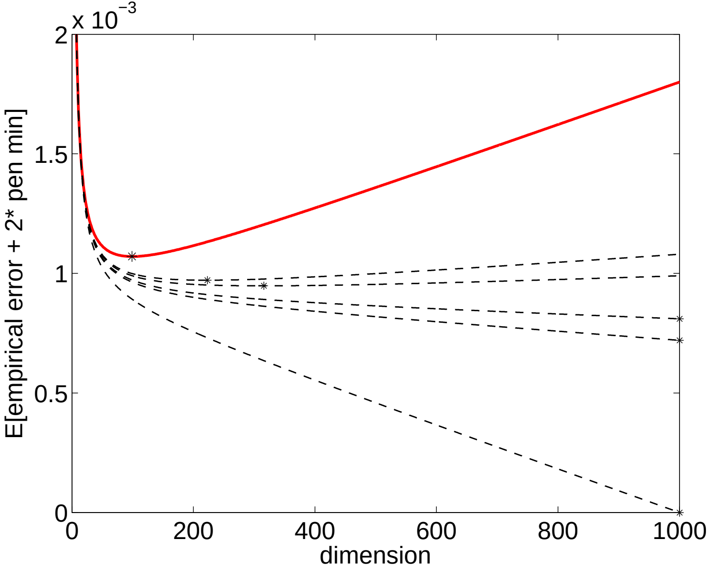
$\mathbb{E}$[Empirical risk] $+ 1.1 \times \sigma^2 D_m n^{-1}$ (OLS)



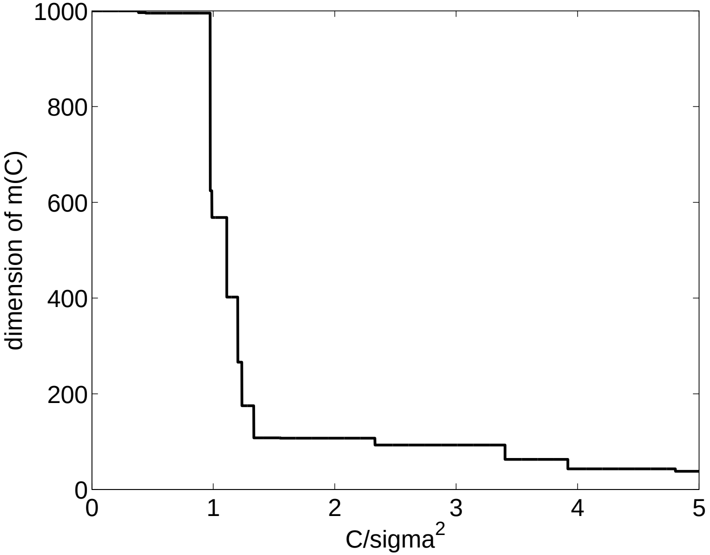
$\mathbb{E}[$ Empirical risk $] + 1.2 \times \sigma^2 D_m n^{-1}$ (OLS)



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



OLS: Dimension jump



OLS: slope heuristics algorithm (Birgé & Massart 2007)

1. for every $C > 0$, compute

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

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

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

Practical use: CAPUSHE package (Baudry, Maugis & Michel, 2011)
`http://www.math.univ-toulouse.fr/~maugis/CAPUSHE.html`

Theorem (1): Dimension jump / Minimal penalty

Theorem (Birgé & Massart 2007, reformulated)

Assumptions:

- $\exists m_0 \in \mathcal{M}$, $S_{m_0} = \mathbb{R}^n$, *i.e.*, $\widehat{F}_{m_0} = Y$,
- $\inf_{m \in \mathcal{M}} \{ \mathbb{E}[\frac{1}{n} \| \widehat{F}_m - F \|^2] \} \leq \sigma^2 \delta_n$, $\delta_n \leq 1/20$,
- *Gaussian noise:* $\varepsilon_i \sim \mathcal{N}(0, \sigma^2)$.

Then, $\forall \gamma > 0, n \geq n_0(\gamma)$, *w.p. at least* $1 - 4 \operatorname{Card}(\mathcal{M}) n^{-\gamma}$,

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

$$\forall C > \left(1 + \eta_n^+\right) \sigma^2 , \quad D_{\widehat{m}(C)} \leq \frac{n}{10}$$

with $\eta_n^- = 81 \sqrt{\frac{\gamma \log(n)}{n}}$, $\eta_n^+ = \eta_n^- + 20 \delta_n$. *In the first case,* $\frac{1}{n} \| \widehat{F}_{\widehat{m}(C)} - F \|^2 \geq \frac{7 \sigma^2}{8} \gg \inf_{m \in \mathcal{M}_n} \{ \frac{1}{n} \| \widehat{F}_m - F \|^2 \}$.

Theorem (1'): Dimension jump / Minimal penalty

Under the same assumptions, on the same event, $\forall a_n, b_n$ such that

$$2n\delta_n + 16.2\sqrt{\gamma \log(n) n} < b_n < a_n < n \quad ,$$
$$\forall C < \left(1 - \eta_n^-\right)\sigma^2 \, , \quad D_{\widehat{m}(C)} \geq a_n$$
$$\forall C > \left(1 + \eta_n^+\right)\sigma^2 \, , \quad D_{\widehat{m}(C)} \leq b_n$$

with

$$\eta_n^- = \left(1 - \frac{a_n}{n}\right)^{-1} \sqrt{\frac{\gamma \log(n)}{n}}$$
$$\eta_n^+ = \frac{n}{b_n - n\delta_n} \left(\delta_n + 8.1\sqrt{\frac{\gamma \log(n)}{n}}\right)$$

Increasing γ, a_n, decreasing $b_n \Rightarrow$ larger window for C

Larger $\delta_n \Rightarrow$ larger upper bound for C & lower bound for b_n

Theorem (2): Oracle inequality

Theorem (Birgé & Massart 2007, reformulated)

Assumptions:

- $\widehat{m} \in \operatorname{argmin}_{m \in \mathcal{M}} \left\{ \frac{1}{n} \| \widehat{F}_m - Y \|^2 + 2 \widehat{C}_{\mathrm{jump}} \frac{D_m}{n} \right\}$
- $\exists m_0 \in \mathcal{M}$, $S_{m_0} = \mathbb{R}^n$, *i.e.*, $\widehat{F}_{m_0} = Y$,
- $\inf_{m \in \mathcal{M}} \left\{ \mathbb{E}\left[\frac{1}{n} \left\| \widehat{F}_m - F \right\|^2 \right] \right\} \leq \sigma^2 \delta_n$, $\delta_n \leq 1/20$,
- *Gaussian noise:* $\varepsilon_i \sim \mathcal{N}(0, \sigma^2)$.

Then, with probability at least $1 - 4 \operatorname{Card}(\mathcal{M}) n^{-\gamma}$, *if* $n \geq n_0(\gamma)$, *for every* $\eta \geq 2 \max\{ \eta_n^-, \eta_n^+ \}$,

$$\frac{1}{n} \left\| \widehat{F}_{\widehat{m}} - F \right\|^2 \leq (1 + 3\eta) \inf_{m \in \mathcal{M}} \left\{ \frac{1}{n} \left\| \widehat{F}_m - F \right\|^2 \right\} + \frac{880 \sigma^2 \gamma \log(n)}{\eta n}$$

Variance estimation

- Slope heuristics: with probability $1 - 4\,\mathrm{Card}(\mathcal{M}) n^{-\gamma}$,

$$1 - 81\sqrt{\frac{\gamma \log(n)}{n}} \leq \frac{\widehat{C}_{\mathrm{jump}}}{\sigma^2} \leq 1 + 20\delta_n + 81\sqrt{\frac{\gamma \log(n)}{n}}$$

Variance estimation

- Slope heuristics: with probability $1 - 4\,\mathrm{Card}(\mathcal{M}) n^{-\gamma}$,

$$1 - 81\sqrt{\frac{\gamma \log(n)}{n}} \leq \frac{\widehat{C}_{\mathrm{jump}}}{\sigma^2} \leq 1 + 20\delta_n + 81\sqrt{\frac{\gamma \log(n)}{n}}$$

- Naive estimator:

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

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

Variance estimation

- Slope heuristics: with probability $1 - 4\,\mathrm{Card}(\mathcal{M}) n^{-\gamma}$,

$$1 - 81\sqrt{\frac{\gamma \log(n)}{n}} \leq \frac{\widehat{C}_{\mathrm{jump}}}{\sigma^2} \leq 1 + 20\delta_n + 81\sqrt{\frac{\gamma \log(n)}{n}}$$

- Naive estimator:

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

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

- Best variance estimator for $\mathbb{E}[(\widehat{\sigma} - \sigma)^2]$ is not necessarily the best for model selection.

Data-driven penalties

- Naive estimator with some fixed m_0 + plug in:

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

Drawbacks: choice of m_0 ? unknown bias (overpenalization)

Data-driven penalties

- Naive estimator with some fixed m_0 + plug in:

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

Drawbacks: choice of m_0 ? unknown bias (overpenalization)

- FPE (Akaike, 1970; Baraud, Giraud & Huet, 2009)

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

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

Drawbacks: deal carefully with the largest models
oracle inequalities (Baraud, Giraud & Huet, 2009) hold assuming an upper bound on $\max_{m \in \mathcal{M}} D_m$ (FPE) or for a new penalty, very large for the largest models (BGH)

Generalized cross-validation (Craven & Wahba, 1978)

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

If $D_m \ll n$,

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

Drawbacks: deal carefully with the largest models
$\Rightarrow$ e.g., for smoothing splines, oracle inequality assumes the effective dimension is $\leq n/5$ for all m (Cao & Golubev, 2006)

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?

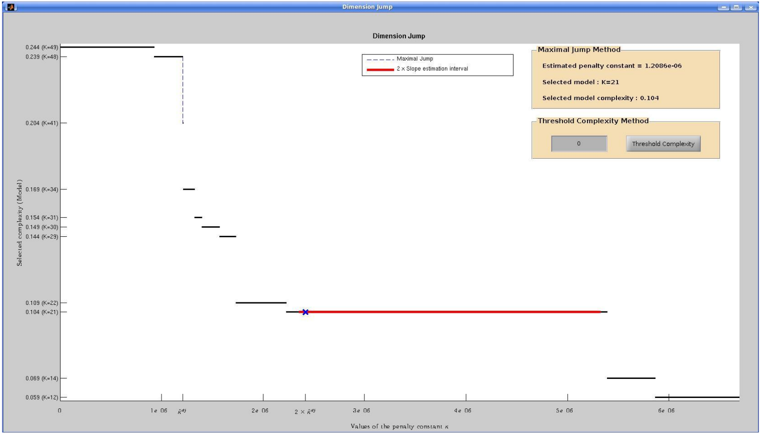
How to localize the jump in practice?

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

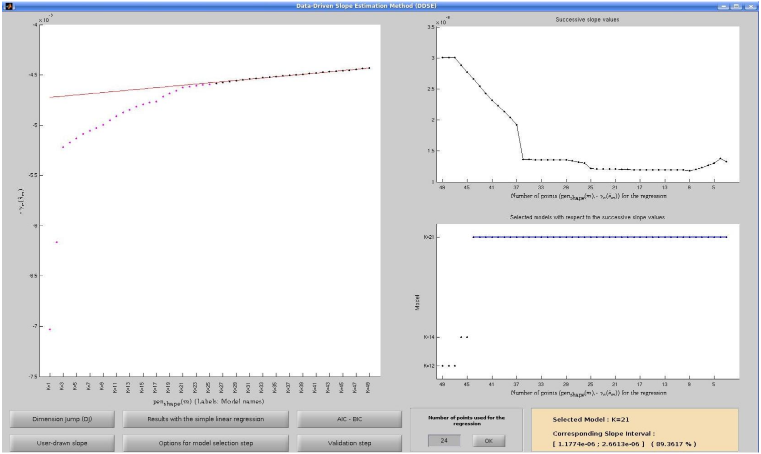
- Estimation of the slope of the empirical risk as a function of the dimension:
  computed with which models? robust regression?

- Jump vs. slope? Take both!
  $\Rightarrow$ package CAPUSHE (Baudry, Maugis & Michel, 2011)
  `http://www.math.univ-toulouse.fr/~maugis/CAPUSHE.html`

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

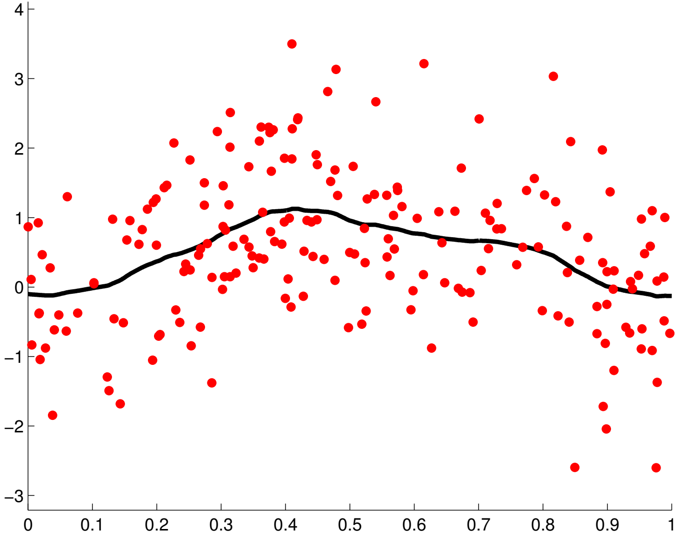


CAPUSHE (Baudry, Maugis & Michel, 2011): slope

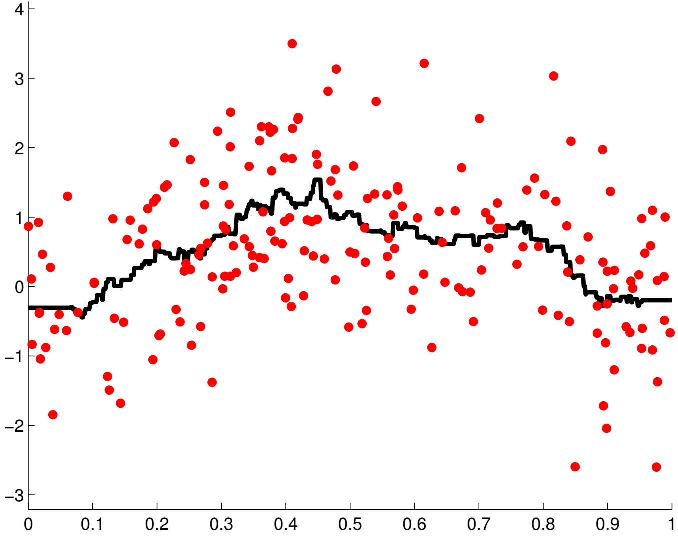


Outline

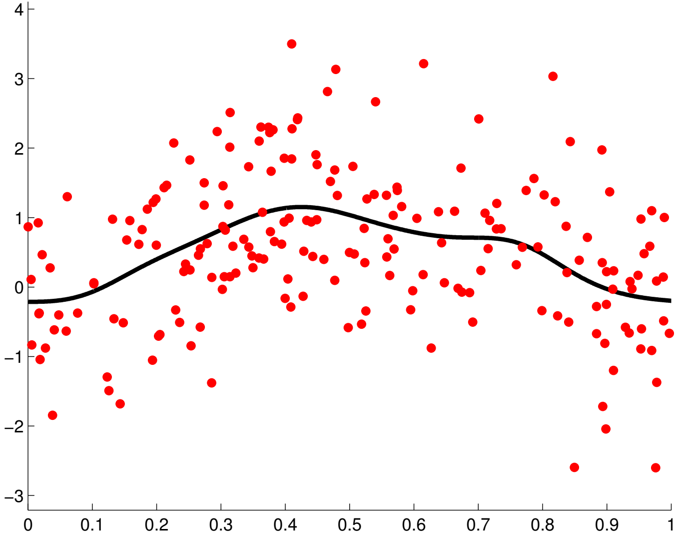
Kernel ridge estimator ($\lambda = 0.01$)



k-nearest-neighbours estimator ($k = 20$)



Nadaraya-Watson estimator ($\sigma = 0.01$)



Linear estimators

- OLS: $\widehat{F}_m = \Pi_{S_m} Y$ (projection onto S_m)
- (kernel) ridge regression, spline smoothing (Wahba, 1990):

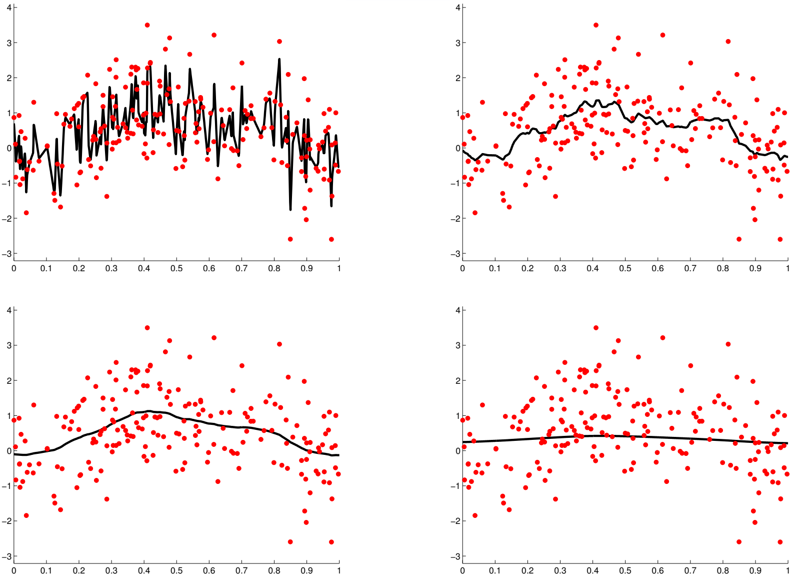
$$\widehat{F}_i = \widehat{f}(x_i) \quad \text{with} \quad \widehat{f} \in \operatorname*{argmin}_{f \in \mathcal{F}_K} \left\{ \frac{1}{n} \sum_{i=1}^{n} \left(Y_i - f(x_i) \right)^2 + \lambda \left\| f \right\|_{\mathcal{F}_K}^2 \right\}$$

$$\Rightarrow \quad \widehat{F}_{\lambda,K} = K(K + \lambda I)^{-1} Y \quad \text{where} \quad K = (K(x_i, x_j))_{1 \le i,j \le n}$$

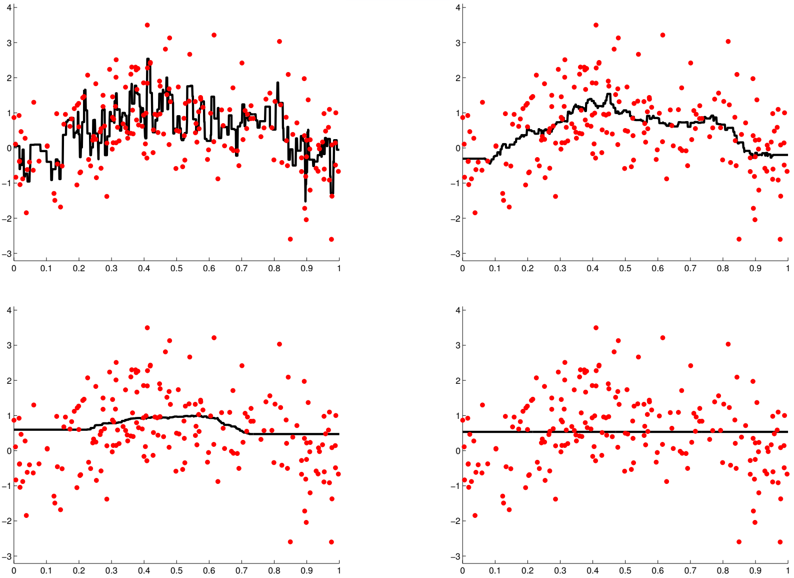
- k-nearest neighbours
- Nadaraya-Watson estimators

$$\widehat{F} = AY \qquad \text{where } A \text{ does not depend on } Y$$

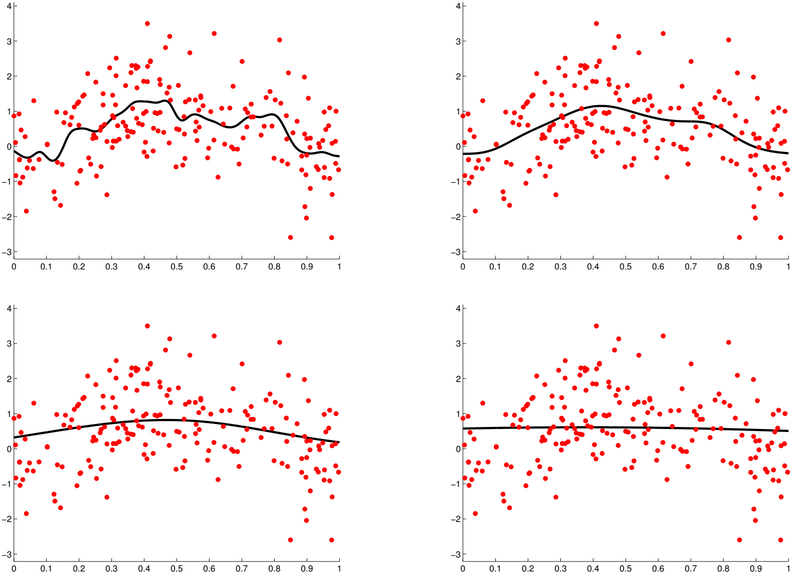
Estimator selection: kernel ridge



Estimator selection: k nearest neighbours



Estimator selection: Nadaraya-Watson



Slope heuristics for linear estimators?

OLS

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

$$\operatorname*{argmin}_{m \in \mathcal{M}} \left\{ \frac{1}{n} \| \widehat{F}_m - Y \|^2 + C \frac{D_m}{n} \right\}$$

$\Rightarrow D_{\widehat{m}(C)}$ "jumps" at $\widehat{C}_{\mathrm{jump}} \approx \sigma^2$

$\Rightarrow$ optimal choice with $\widehat{m}(2 \widehat{C}_{\mathrm{jump}})$

Slope heuristics for linear estimators?

OLS

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

$$\operatorname*{argmin}_{m \in \mathcal{M}} \left\{ \frac{1}{n} \| \widehat{F}_m - Y \|^2 + C \frac{D_m}{n} \right\}$$

$\Rightarrow D_{\widehat{m}(C)}$ “jumps” at $\widehat{C}_{\mathrm{jump}} \approx \sigma^2$

$\Rightarrow$ optimal choice with $\widehat{m}(2\widehat{C}_{\mathrm{jump}})$

Linear estimators

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

Slope heuristics for linear estimators?

OLS

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

$$\operatorname*{argmin}_{m \in \mathcal{M}} \left\{ \frac{1}{n} \| \widehat{F}_m - Y \|^2 + C \frac{D_m}{n} \right\}$$

$\Rightarrow D_{\widehat{m}(C)}$ "jumps" at $\widehat{C}_{\mathrm{jump}} \approx \sigma^2$

$\Rightarrow$ optimal choice with $\widehat{m}(2\widehat{C}_{\mathrm{jump}})$

Linear estimators

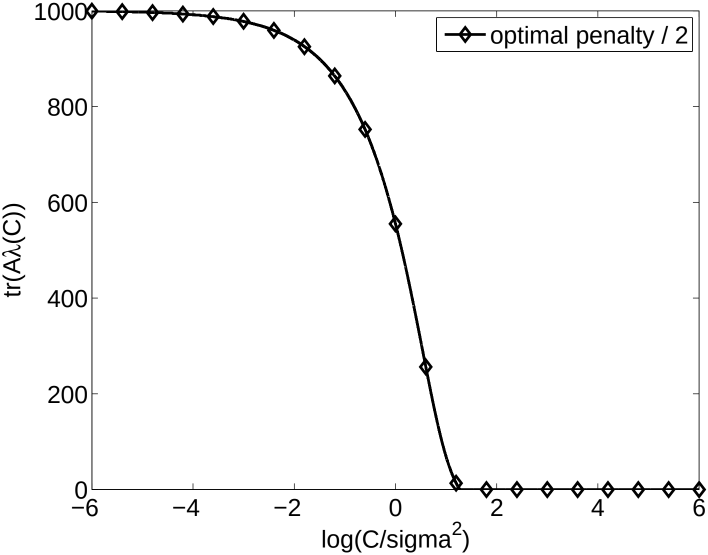
$$\mathrm{pen}_{\mathrm{CL}}(m) = \frac{2\sigma^2 \operatorname{tr}(A_m)}{n}$$

$$\operatorname*{argmin}_{m \in \mathcal{M}} \left\{ \frac{1}{n} \| \widehat{F}_m - Y \|^2 + C \frac{\operatorname{tr}(A_m)}{n} \right\}$$

Does $\operatorname{tr}(A_{\widehat{m}(C)})$ jump at $\widehat{C}_{\mathrm{jump}} \approx \sigma^2$?

optimal choice with $\widehat{m}(2\widehat{C}_{\mathrm{jump}})$?

No dimension jump with a penalty $\propto \mathrm{tr}(A_m)$



Minimal penalties for linear estimators

$$\mathbb{E}\left[\frac{1}{n}\left\|\widehat{F}_m - F\right\|^2\right] = \frac{1}{n}\left\|(I - A_m)F\right\|^2 + \frac{\operatorname{tr}(A_m^\top A_m)\sigma^2}{n} = \text{bias} + \text{variance}$$

Minimal penalties for linear estimators

$$\mathbb{E}\left[\frac{1}{n}\left\|\widehat{F}_m - F\right\|^2\right] = \frac{1}{n}\|(I - A_m)F\|^2 + \frac{\mathrm{tr}(A_m^\top A_m)\sigma^2}{n} = \text{bias} + \text{variance}$$

$$\mathbb{E}\left[\frac{1}{n}\left\|\widehat{F}_m - Y\right\|^2\right] = \sigma^2 + \frac{1}{n}\|(I - A_m)F\|^2 - \frac{\left(2\,\mathrm{tr}(A_m) - \mathrm{tr}(A_m^\top A_m)\right)\sigma^2}{n}$$

Minimal penalties for linear estimators

$$\mathbb{E}\left[\frac{1}{n}\left\|\widehat{F}_m - F\right\|^2\right] = \frac{1}{n}\|(I - A_m)F\|^2 + \frac{\operatorname{tr}(A_m^\top A_m)\sigma^2}{n} = \text{bias} + \text{variance}$$

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

$$\Rightarrow \quad \text{optimal penalty} \quad \frac{\left(2\operatorname{tr}(A_m)\right)\sigma^2}{n}$$

Minimal penalties for linear estimators

$$\mathbb{E}\left[\frac{1}{n}\left\|\widehat{F}_m - F\right\|^2\right] = \frac{1}{n}\|(I - A_m)F\|^2 + \frac{\operatorname{tr}(A_m^\top A_m)\sigma^2}{n} = \text{bias} + \text{variance}$$

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

$$\Rightarrow \quad \text{optimal penalty} \quad \frac{\left(2\operatorname{tr}(A_m)\right)\sigma^2}{n}$$

$$\Rightarrow \quad \text{minimal penalty} \quad \frac{\left(2\operatorname{tr}(A_m) - \operatorname{tr}(A_m^\top A_m)\right)\sigma^2}{n}$$

Minimal penalties for linear estimators

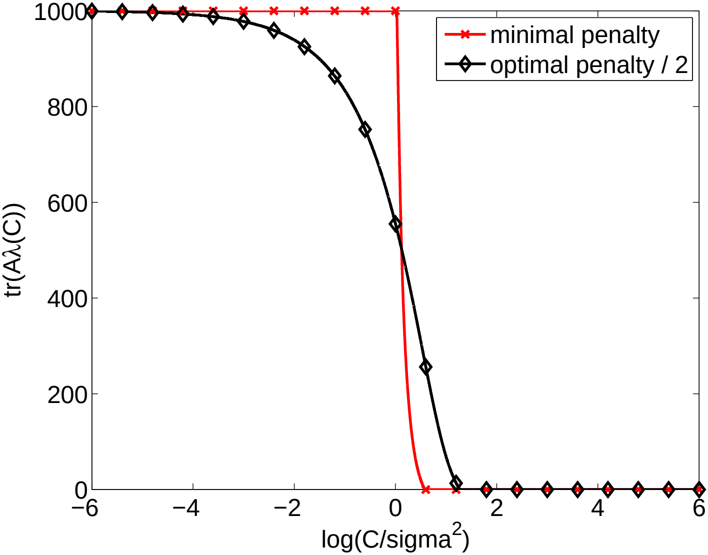
$$\mathbb{E}\left[\frac{1}{n}\left\|\widehat{F}_m - F\right\|^2\right] = \frac{1}{n}\left\|(I - A_m)F\right\|^2 + \frac{\operatorname{tr}(A_m^\top A_m)\sigma^2}{n} = \text{bias} + \text{variance}$$

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

$$\Rightarrow \quad \text{optimal penalty} \quad \frac{\left(2\operatorname{tr}(A_m)\right)\sigma^2}{n}$$

$$\widehat{m}(C) \in \operatorname*{argmin}_{\lambda \in \Lambda}\left\{\frac{1}{n}\left\|\widehat{F}_m - Y\right\|^2 + C \times \frac{2\operatorname{tr}(A_m) - \operatorname{tr}(A_m^\top A_m)}{n}\right\}$$

"Dimension" jump (ridge regression)



Penalty calibration algorithm (A. & Bach 2009)

1. for every $C > 0$, compute

$$\widehat{m}_{\min}(C) \in \operatorname*{argmin}_{m\in\mathcal{M}} \left\{ \frac{1}{n}\left\| Y - \widehat{F}_m \right\|^2 + \frac{C\left(2\operatorname{tr}(A_m) - \operatorname{tr}(A_m^\top A_m)\right)}{n} \right\}$$

2. find $\widehat{C}_{\mathrm{jump}}$ such that $\operatorname{tr}(A_{\widehat{m}_{\min}(C)})$ is "too large" when $C < \widehat{C}_{\mathrm{jump}}$ and "reasonably small" when $C > \widehat{C}_{\mathrm{jump}}$,

3. select

$$\widehat{m} \in \operatorname*{argmin}_{m\in\mathcal{M}} \left\{ \frac{1}{n}\left\| Y - \widehat{F}_m \right\|^2 + \frac{2\widehat{C}_{\mathrm{jump}}\operatorname{tr}(A_m)}{n} \right\}$$

Theorem for linear estimators

Theorem (A. & Bach 2009–2011)

Assumptions:

- $\forall m \in \mathcal{M}, \ \|A_m\| \leq L_1$ *and* $\operatorname{tr}(A_m^\top A_m) \leq \operatorname{tr}(A_m) \leq n$
- $\exists m_0, m_1 \in \mathcal{M}$, $A_{m_1} = I_n$, $D_{m_0} \leq \sqrt{n}$ *and* $\frac{1}{n} \left\| F_{m_0} - F \right\|^2 \leq \sigma^2 \sqrt{\log(n)/n}$.
- *Gaussian noise:* $\varepsilon_i \sim \mathcal{N}(0, \sigma^2)$

Then, $\forall \gamma > 0, n \geq n_0(\gamma)$, *w.p. at least* $1 - 6 \operatorname{Card}(\mathcal{M}) n^{-\gamma}$,

$$\forall C < \left(1 - \eta_n^-\right) \sigma^2 , \quad D_{\widehat{m}(C)} \geq \frac{n}{3}$$

$$\forall C > \left(1 + \eta_n^+\right) \sigma^2 , \quad D_{\widehat{m}(C)} \leq \frac{n}{10}$$

with $\eta_n^- = \eta_n^+ = L_2 \delta \sqrt{\frac{\log(n)}{n}}$.

Theorem for linear estimators

Theorem (A. & Bach 2009–2011)

Assumptions:

- $\forall m \in \mathcal{M}$, $\|A_m\| \leq L_1$ *and* $\mathrm{tr}(A_m^\top A_m) \leq \mathrm{tr}(A_m) \leq n$
- $\exists m_0, m_1 \in \mathcal{M}$, $A_{m_1} = I_n$, $D_{m_0} \leq \sqrt{n}$ *and* $\frac{1}{n} \|F_{m_0} - F\|^2 \leq \sigma^2 \sqrt{\log(n)/n}$.
- *Gaussian noise:* $\varepsilon_i \sim \mathcal{N}(0, \sigma^2)$

Then, $\forall \gamma > 0, n \geq n_0(\gamma)$, *w.p. at least* $1 - 6\,\mathrm{Card}(\mathcal{M}) n^{-\gamma}$, $\forall C \in (1 - \eta_n^-, 1 + \eta_n^+)$, $\eta \in (0,2)$,

$$\forall \widehat{m} \in \operatorname*{argmin}_{m \in \mathcal{M}} \left\{ \frac{1}{n} \left\| Y - \widehat{F}_m \right\|^2 + \frac{2C\,\mathrm{tr}(A_m)}{n} \right\} ,$$

$$\frac{1}{n} \left\| \widehat{F}_{\widehat{m}} - F \right\|^2 \leq (1 + \eta) \inf_{m \in \mathcal{M}} \left\{ \frac{1}{n} \left\| \widehat{F}_m - F \right\|^2 \right\} + \frac{L_3 \gamma^2 \log(n) \sigma^2}{\eta n}$$

Comparison with least-squares

- Linear estimators:

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

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

$$\frac{\mathrm{pen}_{\mathrm{opt}}(m)}{\mathrm{pen}_{\min}(m)} = \frac{2 \operatorname{tr}(A_m)}{2 \operatorname{tr}(A_m) - \operatorname{tr}(A_m^\top A_m)} \in (1, 2]$$

Comparison with least-squares

- Linear estimators:

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

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

$$\frac{\mathrm{pen}_{\mathrm{opt}}(m)}{\mathrm{pen}_{\mathrm{min}}(m)} = \frac{2\,\mathrm{tr}(A_m)}{2\,\mathrm{tr}(A_m) - \mathrm{tr}(A_m^\top A_m)} \in (1, 2]$$

- Least-squares case:

$$A_m^\top A_m = A_m \quad \Rightarrow \quad \frac{\mathrm{pen}_{\mathrm{opt}}(m)}{\mathrm{pen}_{\mathrm{min}}(m)} = 2 \quad \Rightarrow \quad \text{Slope heuristics}$$

The k-nearest neighbours case

$$\forall i,j \in \{1, \ldots, n\} , \quad A_{i,j} \in \left\{ 0, \frac{1}{k} \right\}$$

$$\forall i \in \{1, \ldots, n\} , \quad A_{i,i} = \frac{1}{k} \quad \text{and} \quad \sum_{j=1}^{n} A_{i,j} = 1$$

The k-nearest neighbours case

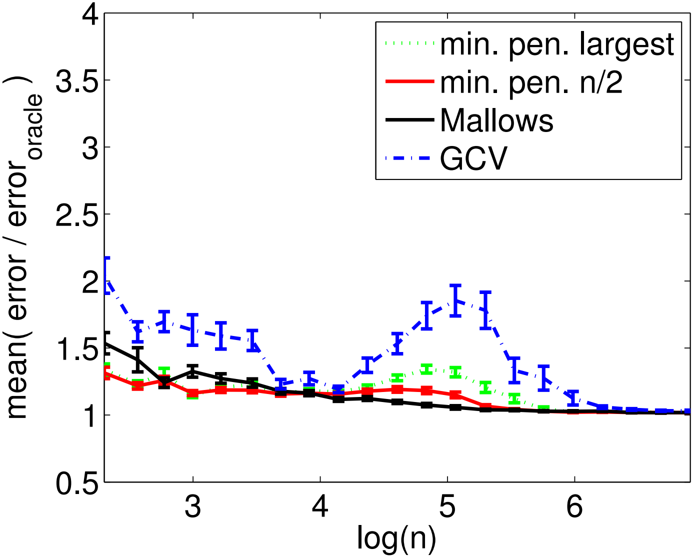
$$\forall i,j \in \{1, \ldots, n\} \ , \quad A_{i,j} \in \left\{ 0, \frac{1}{k} \right\}$$

$$\forall i \in \{1, \ldots, n\} \ , \quad A_{i,i} = \frac{1}{k} \quad \text{and} \quad \sum_{j=1}^{n} A_{i,j} = 1$$

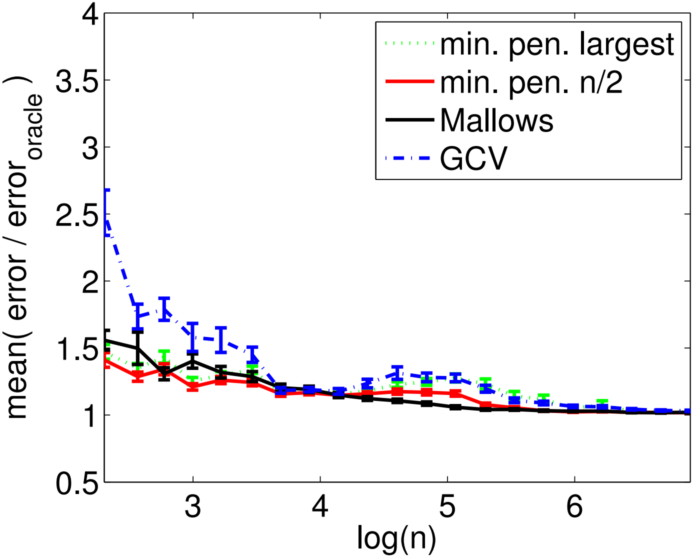
$$\Rightarrow \quad \operatorname{tr}(A) = \frac{n}{k} = \operatorname{tr}(A^\top A)$$

$$\Rightarrow \qquad \mathrm{pen}_{\mathrm{opt}} = 2 \, \mathrm{pen}_{\mathrm{min}}$$

Simulations: N-W, $F_i = \sin(25\pi x_i^3)$, $n = 200$



Simulations: ridge, $F_i = \sin(25\pi x_i^3)$, $n = 200$



General framework

- Goal: find from data $t \in \mathbb{S}$ with $\mathcal{R}(t)$ minimal.
- Empirical risk $\widehat{\mathcal{R}}_n(t)$
- Collection of estimators $(\widehat{s}_m)_{m \in \mathcal{M}}$
- Oracle inequality:

$$\mathcal{R}(\widehat{s}_{\widehat{m}}) - \mathcal{R}(s^{\star}) \leq K_n \inf_{m \in \mathcal{M}} \{\mathcal{R}(\widehat{s}_m) - \mathcal{R}(s^{\star})\} + R_n$$

where $\mathcal{R}(s^{\star}) := \inf_{t \in \mathbb{S}} \mathcal{R}(t)$.

General algorithm

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

1. for every $C > 0$, compute

$$\widehat{m}_{\min}(C) \in \operatorname*{argmin}_{m \in \mathcal{M}} \left\{ \widehat{\mathcal{R}}_n(\widehat{s}_m) + C\,\mathrm{pen}_0(m) \right\}$$

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

3. select

$$\widehat{m} \in \operatorname*{argmin}_{m \in \mathcal{M}} \left\{ \widehat{\mathcal{R}}_n(\widehat{s}_m) + \widehat{C}_{\mathrm{jump}}\,\mathrm{pen}_1(m) \right\}$$

General algorithm

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

1. for every $C > 0$, compute

$$\widehat{m}_{\min}(C) \in \operatorname*{argmin}_{m \in \mathcal{M}} \left\{ \widehat{\mathcal{R}}_n(\widehat{s}_m) + C \, \mathrm{pen}_0(m) \right\}$$

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

3. select

$$\widehat{m} \in \operatorname*{argmin}_{m \in \mathcal{M}} \left\{ \widehat{\mathcal{R}}_n(\widehat{s}_m) + \widehat{C}_{\mathrm{jump}} \, \mathrm{pen}_1(m) \right\}$$

Example (slope heuristics): $\mathrm{pen}_1 = 2 \, \mathrm{pen}_0$

Ideas for a proof

- $\exists C^{\star} > 0,\ C^{\star} \operatorname{pen}_0$ minimal penalty, $C^{\star} \operatorname{pen}_1$ optimal penalty

Ideas for a proof

- $\exists C^\star > 0$, $C^\star \operatorname{pen}_0$ minimal penalty, $C^\star \operatorname{pen}_1$ optimal penalty
- Decomposition of the ideal penalty:

$$
\begin{aligned}
\operatorname{pen}_{\mathrm{id}}(m) &= \mathcal{R}\left(\widehat{s}_m\right) - \widehat{\mathcal{R}}_n\left(\widehat{s}_m\right) \\
&= \underbrace{\mathcal{R}\left(\widehat{s}_m\right) - \mathcal{R}\left(s_m^\star\right)}_{} + \underbrace{\mathcal{R}\left(s_m^\star\right) - \widehat{\mathcal{R}}_n\left(s_m^\star\right)}_{} + \underbrace{\widehat{\mathcal{R}}_n\left(s_m^\star\right) - \widehat{\mathcal{R}}_n\left(\widehat{s}_m\right)}_{} \\
&= \quad p_1(m) \quad + \quad \delta(m) \quad + \quad p_2(m)
\end{aligned}
$$

Ideas for a proof

- $\exists C^\star > 0$, $C^\star \operatorname{pen}_0$ minimal penalty, $C^\star \operatorname{pen}_1$ optimal penalty
- Decomposition of the ideal penalty:

$$\begin{aligned} \operatorname{pen}_{\mathrm{id}}(m) &= \mathcal{R}(\widehat{s}_m) - \widehat{\mathcal{R}}_n(\widehat{s}_m) \\ &= \underbrace{\mathcal{R}(\widehat{s}_m) - \mathcal{R}(s_m^\star)}_{p_1(m)} + \underbrace{\mathcal{R}(s_m^\star) - \widehat{\mathcal{R}}_n(s_m^\star)}_{\delta(m)} + \underbrace{\widehat{\mathcal{R}}_n(s_m^\star) - \widehat{\mathcal{R}}_n(\widehat{s}_m)}_{p_2(m)} \\ &= p_1(m) + \delta(m) + p_2(m) \end{aligned}$$

- Good candidate for the minimal penalty: p_2 or $\mathbb{E}[p_2]$

Ideas for a proof

- $\exists C^\star > 0$, $C^\star \operatorname{pen}_0$ minimal penalty, $C^\star \operatorname{pen}_1$ optimal penalty
- Decomposition of the ideal penalty:

$$\begin{aligned} \operatorname{pen}_{\mathrm{id}}(m) &= \mathcal{R}(\widehat{s}_m) - \widehat{\mathcal{R}}_n(\widehat{s}_m) \\ &= \underbrace{\mathcal{R}(\widehat{s}_m) - \mathcal{R}(s_m^\star)} + \underbrace{\mathcal{R}(s_m^\star) - \widehat{\mathcal{R}}_n(s_m^\star)} + \underbrace{\widehat{\mathcal{R}}_n(s_m^\star) - \widehat{\mathcal{R}}_n(\widehat{s}_m)} \\ &= \qquad p_1(m) \qquad + \qquad \delta(m) \qquad + \qquad p_2(m) \end{aligned}$$

- Good candidate for the minimal penalty: p_2 or $\mathbb{E}[p_2]$
- Good candidate for the optimal penalty: $p_1 + p_2$ or $\mathbb{E}[p_1 + p_2] = \mathbb{E}[\operatorname{pen}_{\mathrm{id}}]$

Ideas for a proof

- $\exists C^\star > 0$, $C^\star \operatorname{pen}_0$ minimal penalty, $C^\star \operatorname{pen}_1$ optimal penalty
- Decomposition of the ideal penalty:

$$\begin{aligned} \operatorname{pen}_{\mathrm{id}}(m) &= \mathcal{R}(\widehat{s}_m) - \widehat{\mathcal{R}}_n(\widehat{s}_m) \\ &= \underbrace{\mathcal{R}(\widehat{s}_m) - \mathcal{R}(s_m^\star)}_{p_1(m)} + \underbrace{\mathcal{R}(s_m^\star) - \widehat{\mathcal{R}}_n(s_m^\star)}_{\delta(m)} + \underbrace{\widehat{\mathcal{R}}_n(s_m^\star) - \widehat{\mathcal{R}}_n(\widehat{s}_m)}_{p_2(m)} \\ &= p_1(m) + \delta(m) + p_2(m) \end{aligned}$$

- Good candidate for the minimal penalty: p_2 or $\mathbb{E}[p_2]$
- Good candidate for the optimal penalty: $p_1 + p_2$ or $\mathbb{E}[p_1 + p_2] = \mathbb{E}[\operatorname{pen}_{\mathrm{id}}]$

Key tools: concentration inequalities for $\delta(m) - \delta(m')$, $p_2(m)$ (Wilks phenomenon: Boucheron & Massart, 2010; Spokoiny, 2012; Andresen & Spokoiny, 2013) and $p_1(m)$ (Saumard, 2010–2012)

Theoretical results (1)

- OLS, fixed-design regression, homoscedastic Gaussian noise (Birgé & Massart, 2007)

Theoretical results (1)

- OLS, fixed-design regression, homoscedastic Gaussian noise (Birgé & Massart, 2007)
- OLS, random-design regression, heteroscedastic noise (regressograms, A. & Massart, 2009; piecewise polynomials, Saumard, 2013)

Theoretical results (1)

- OLS, fixed-design regression, homoscedastic Gaussian noise (Birgé & Massart, 2007)
- OLS, random-design regression, heteroscedastic noise (regressograms, A. & Massart, 2009; piecewise polynomials, Saumard, 2013)
- Least-squares density estimation, i.i.d. (Lerasle, 2009) or mixing data (Lerasle, 2010)

Theoretical results (1)

- OLS, fixed-design regression, homoscedastic Gaussian noise (Birgé & Massart, 2007)
- OLS, random-design regression, heteroscedastic noise (regressograms, A. & Massart, 2009; piecewise polynomials, Saumard, 2013)
- Least-squares density estimation, i.i.d. (Lerasle, 2009) or mixing data (Lerasle, 2010)
- Density estimation, Kullback risk, maximum-likelihood estimators on histograms (Saumard, 2010)

Theoretical results (1)

- OLS, fixed-design regression, homoscedastic Gaussian noise (Birgé & Massart, 2007)
- OLS, random-design regression, heteroscedastic noise (regressograms, A. & Massart, 2009; piecewise polynomials, Saumard, 2013)
- Least-squares density estimation, i.i.d. (Lerasle, 2009) or mixing data (Lerasle, 2010)
- Density estimation, Kullback risk, maximum-likelihood estimators on histograms (Saumard, 2010)
- Minimum contrast estimator, regular contrast (Saumard, 2010)

Theoretical results (1)

- OLS, fixed-design regression, homoscedastic Gaussian noise (Birgé & Massart, 2007)
- OLS, random-design regression, heteroscedastic noise (regressograms, A. & Massart, 2009; piecewise polynomials, Saumard, 2013)
- Least-squares density estimation, i.i.d. (Lerasle, 2009) or mixing data (Lerasle, 2010)
- 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, 2011)

Theoretical results (2)

- Linear estimators, regression (A. & Bach, 2009–2011)
- Fixed-design regression, complete variable selection (many models), homoscedastic Gaussian noise (Birgé & Massart, 2007)
- Context tree estimation, Kullback risk, maximum-likelihood estimators, mixing data (Garivier & Lerasle, 2011)
- Partial proofs in other settings (Baraud, Giraud & Huet, 2009; Verzelen, 2010; Giraud, 2011)

Resampling and minimal penalties

Problem: some of the theoretical results work for

$$\mathrm{pen}_0(m) \propto \mathbb{E}\left[\widehat{\mathcal{R}}_n (s_m^\star) - \widehat{\mathcal{R}}_n (\widehat{s}_m) \right] \quad \text{or} \quad \widehat{\mathcal{R}}_n (s_m^\star) - \widehat{\mathcal{R}}_n (\widehat{s}_m)$$

that is unknown.

Resampling and minimal penalties

Problem: some of the theoretical results work for

$$\mathrm{pen}_0(m) \propto \mathbb{E}\left[\widehat{\mathcal{R}}_n\left(s_m^\star\right) - \widehat{\mathcal{R}}_n\left(\widehat{s}_m\right) \right] \quad \text{or} \quad \widehat{\mathcal{R}}_n\left(s_m^\star\right) - \widehat{\mathcal{R}}_n\left(\widehat{s}_m\right)$$

that is unknown.

⇒ Resampling-based estimator:

$$C_W \mathbb{E}\left[\widehat{\mathcal{R}}_n^W\left(\widehat{s}_m\right) - \widehat{\mathcal{R}}_n^W\left(\widehat{s}_m^W\right) \,\middle|\, \xi_1, \ldots, \xi_n \right]$$

Heteroscedastic regression (A., 2008–09), density estimation (Lerasle, 2009)

C_W often unknown (or known only asymptotically) ⇒ estimate it with the minimal penalty algorithm

Generalization: phase transition and parameter tuning

- Idea: $(\widetilde{s}_\gamma)_{\gamma > 0}$ family of estimators, (observable) phase transition around $\gamma = \gamma_{\mathrm{min}}$, relationship between γ_{opt} and γ_{min}.

Generalization: phase transition and parameter tuning

- Idea: $(\widetilde{s}_\gamma)_{\gamma>0}$ family of estimators, (observable) phase transition around $\gamma = \gamma_{\mathrm{min}}$, relationship between γ_{opt} and γ_{min}.

- Slope heuristics: $\gamma_{\mathrm{opt}} = 2\gamma_{\mathrm{min}}$ with

$$\widetilde{s}_\gamma = \widehat{s}_{\widehat{m}(\gamma)} \qquad \widehat{m}(\gamma) \in \operatorname*{argmin}_{m \in \mathcal{M}} \left\{ \widehat{\mathcal{R}}_n \left(\widehat{s}_m \right) + \gamma \, \mathrm{pen}_0(m) \right\}$$

for some well-chosen pen_0.

Generalization: phase transition and parameter tuning

- Idea: $(\widetilde{s}_\gamma)_{\gamma>0}$ family of estimators, (observable) phase transition around $\gamma = \gamma_{\min}$, relationship between γ_{opt} and $\gamma_{\min}$.

- Slope heuristics: $\gamma_{\mathrm{opt}} = 2\gamma_{\min}$ with

$$\widetilde{s}_\gamma = \widehat{s}_{\widehat{m}(\gamma)} \qquad \widehat{m}(\gamma) \in \operatorname*{argmin}_{m \in \mathcal{M}} \left\{ \widehat{\mathcal{R}}_n\left(\widehat{s}_m\right) + \gamma \operatorname{pen}_0(m) \right\}$$

  for some well-chosen pen_0.

- Partial theoretical results in least-squares density estimation: hard thresholding estimators (Reynaud-Bouret & Rivoirard, 2010; Reynaud-Bouret, Rivoirard & Tuleau-Malot, 2011), Dantzig estimator (Bertin, Le Pennec & Rivoirard, 2011)

Empirical results

- Large collection of models: Change-point detection (Lebarbier, 2005)
- Gaussian mixture models (Maugis & Michel, 2008–2010)
- Binary (supervised) classification (Zwald & Blanchard, 2005)
- Unsupervised classification (Baudry, 2009)
- Computational geometry (Caillerie & Michel, 2009)
- Lasso (Connault, 2011)
- ...

(see Baudry, Maugis & Michel, 2011)

Outline

Multi-task regression

- We want to solve $p \geq 2$ regression problems simultaneously

Multi-task regression

- We want to solve $p \geq 2$ regression problems simultaneously
- Observations: $Y_1, \ldots, Y_n \in \mathbb{R}^p$

$$Y_i^j = F_i^j + \varepsilon_i^j \qquad j = 1, \ldots, p \qquad (\text{e.g., } F_i^j = F^j(x_i))$$

with $(\varepsilon_i)_{1 \leq i \leq n}$ i.i.d. $\mathcal{N}(0, \Sigma)$, $\Sigma \in \mathcal{S}_p^+(\mathbb{R})$

Multi-task regression

- We want to solve $p \geq 2$ regression problems simultaneously
- Observations: $Y_1, \ldots, Y_n \in \mathbb{R}^p$

$$Y_i^j = F_i^j + \varepsilon_i^j \qquad j = 1, \ldots, p \qquad (\text{e.g., } F_i^j = F^j(x_i))$$

with $(\varepsilon_i)_{1 \leq i \leq n}$ i.i.d. $\mathcal{N}(0, \Sigma)$, $\Sigma \in \mathcal{S}_p^+(\mathbb{R})$

- Implicit assumption: the p problems are "similar"

Multi-task regression

- We want to solve $p \geq 2$ regression problems simultaneously
- Observations: $Y_1, \ldots, Y_n \in \mathbb{R}^p$

$$Y_i^j = F_i^j + \varepsilon_i^j \qquad j = 1, \ldots, p \qquad (\text{e.g., } F_i^j = F^j(x_i))$$

with $(\varepsilon_i)_{1 \leq i \leq n}$ i.i.d. $\mathcal{N}(0, \Sigma)$, $\Sigma \in \mathcal{S}_p^+(\mathbb{R})$

- Implicit assumption: the p problems are "similar"
- Least-squares loss of a predictor $t \in \mathbb{R}^{np}$ ("$t_i^j = t^j(x_i)$"):

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

Multi-task regression

- We want to solve $p \geq 2$ regression problems simultaneously
- Observations: $Y_1, \ldots, Y_n \in \mathbb{R}^p$

$$Y_i^j = F_i^j + \varepsilon_i^j \qquad j = 1, \ldots, p \qquad (\text{e.g., } F_i^j = F^j(x_i))$$

with $(\varepsilon_i)_{1 \leq i \leq n}$ i.i.d. $\mathcal{N}(0, \Sigma)$, $\Sigma \in \mathcal{S}_p^+(\mathbb{R})$

- Implicit assumption: the p problems are "similar"
- Least-squares loss of a predictor $t \in \mathbb{R}^{np}$ ("$t_i^j = t^j(x_i)$"):

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

$\Rightarrow$ Estimator $\widehat{s}(Y_1, \ldots, Y_n) \in \mathbb{R}^{np}$?

Ridge multi-task regression

$$\widehat{F} = (\widehat{F}_i^j)_{1 \leq i \leq n\,,\, 1 \leq j \leq p} \quad \text{with} \quad \widehat{F}_i^j = \widehat{f}^j(x_i) \quad \text{and} \quad \widehat{f} \quad \text{defined by:}$$

- If we consider the tasks separately:

$$\arg\min_{f \in \mathcal{F}_K^p} \left\{ \frac{1}{np} \sum_{i=1}^{n} \sum_{j=1}^{p} \left(Y_i^j - f^j(x_i) \right)^2 + \sum_{j=1}^{p} \lambda^j \left\| f^j \right\|_{\mathcal{F}_K}^2 \right\}$$

Ridge multi-task regression

$$\widehat{F} = (\widehat{F}_i^j)_{1 \le i \le n \, , \, 1 \le j \le p} \quad \text{with} \quad \widehat{F}_i^j = \widehat{f}^j(x_i) \quad \text{and} \quad \widehat{f} \quad \text{defined by:}$$

- If we consider the tasks separately:

$$\arg\min_{f \in \mathcal{F}_K^p} \left\{ \frac{1}{np} \sum_{i=1}^{n} \sum_{j=1}^{p} \left(Y_i^j - f^j(x_i) \right)^2 + \sum_{j=1}^{p} \lambda^j \left\| f^j \right\|_{\mathcal{F}_K}^2 \right\}$$

- A possible multi-task approach (Evgeniou *et al.*, 2005):

$$\arg\min_{f \in \mathcal{F}_K^p} \left\{ \frac{1}{np} \sum_{i=1}^{n} \sum_{j=1}^{p} \left(Y_i^j - f^j(x_i) \right)^2 + \lambda \sum_{j=1}^{p} \left\| f^j \right\|_{\mathcal{F}_K}^2 + \mu \sum_{j \neq \ell} \left\| f^j - f^\ell \right\|_{\mathcal{F}_K}^2 \right\}$$

Ridge multi-task regression

$$\widehat{F} = (\widehat{F}_i^j)_{1 \le i \le n \, , \, 1 \le j \le p} \quad \text{with} \quad \widehat{F}_i^j = \widehat{f}^j(x_i) \quad \text{and} \quad \widehat{f} \quad \text{defined by:}$$

- If we consider the tasks separately:

$$\arg\min_{f \in \mathcal{F}_K^p} \left\{ \frac{1}{np} \sum_{i=1}^{n} \sum_{j=1}^{p} \left(Y_i^j - f^j(x_i) \right)^2 + \sum_{j=1}^{p} \lambda^j \left\| f^j \right\|_{\mathcal{F}_K}^2 \right\}$$

- A possible multi-task approach (Evgeniou *et al.*, 2005):

$$\arg\min_{f \in \mathcal{F}_K^p} \left\{ \frac{1}{np} \sum_{i=1}^{n} \sum_{j=1}^{p} \left(Y_i^j - f^j(x_i) \right)^2 + \lambda \sum_{j=1}^{p} \left\| f^j \right\|_{\mathcal{F}_K}^2 + \mu \sum_{j \neq \ell} \left\| f^j - f^\ell \right\|_{\mathcal{F}_K}^2 \right\}$$

- More generally: for $M \in \mathcal{S}_p^+(\mathbb{R})$,

$$\arg\min_{f \in \mathcal{F}_K^p} \left\{ \frac{1}{np} \sum_{i=1}^{n} \sum_{j=1}^{p} \left(Y_i^j - f^j(x_i) \right)^2 + \sum_{j,\ell} M_{j,\ell} \left\langle f^j , \, f^\ell \right\rangle_{\mathcal{F}_K} \right\}$$

Multi-task estimator selection

$\Rightarrow$ Estimators collection $(\widehat{F}_M)_{M \in \mathcal{M}}$, $\mathcal{M} \subset \mathcal{S}_p^+(\mathbb{R})$,

with $\widehat{F}_M = A_M Y$ and $A_M = \left(M^{-1} \otimes K\right)\left(\left(M^{-1} \otimes K\right) + npI_{np}\right)^{-1}$

Multi-task estimator selection

$\Rightarrow$ Estimators collection $(\widehat{F}_M)_{M\in\mathcal{M}}$, $\mathcal{M} \subset \mathcal{S}_p^+(\mathbb{R})$,

with $\widehat{F}_M = A_M Y$ and $A_M = \left(M^{-1} \otimes K\right)\left(\left(M^{-1} \otimes K\right) + npI_{np}\right)^{-1}$

- Goal: select $\widehat{M} \in \mathcal{M}$ such that $\frac{1}{np}\|\widehat{F}_{\widehat{M}} - F\|^2$ is minimal

Multi-task estimator selection

$\Rightarrow$ Estimators collection $(\widehat{F}_M)_{M \in \mathcal{M}}$, $\mathcal{M} \subset \mathcal{S}_p^+(\mathbb{R})$,

with $\widehat{F}_M = A_M Y$ and $A_M = \left(M^{-1} \otimes K \right) \left(\left(M^{-1} \otimes K \right) + np I_{np} \right)^{-1}$

- Goal: select $\widehat{M} \in \mathcal{M}$ such that $\frac{1}{np} \| \widehat{F}_{\widehat{M}} - F \|^2$ is minimal

- Expectation of the ideal penalty:

$$\mathbb{E}\left[\mathrm{pen}_{\mathrm{id}}(M) \right] = \frac{2}{np} \operatorname{tr}\left(A_M \left(\Sigma \otimes I_n \right) \right)$$

Multi-task estimator selection

$\Rightarrow$ Estimators collection $(\widehat{F}_M)_{M \in \mathcal{M}}$, $\mathcal{M} \subset \mathcal{S}_p^+(\mathbb{R})$,

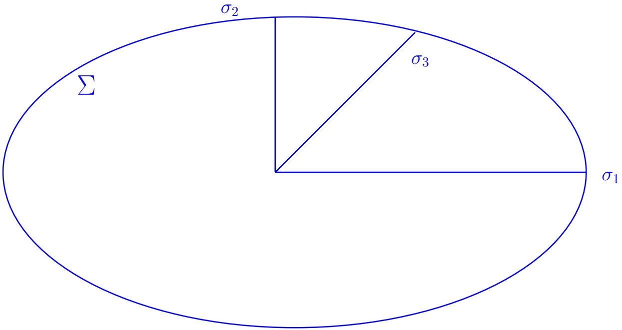
with $\widehat{F}_M = A_M Y$ and $A_M = \left(M^{-1} \otimes K \right) \left(\left(M^{-1} \otimes K \right) + np I_{np} \right)^{-1}$

- Goal: select $\widehat{M} \in \mathcal{M}$ such that $\frac{1}{np} \| \widehat{F}_{\widehat{M}} - F \|^2$ is minimal
- Expectation of the ideal penalty:

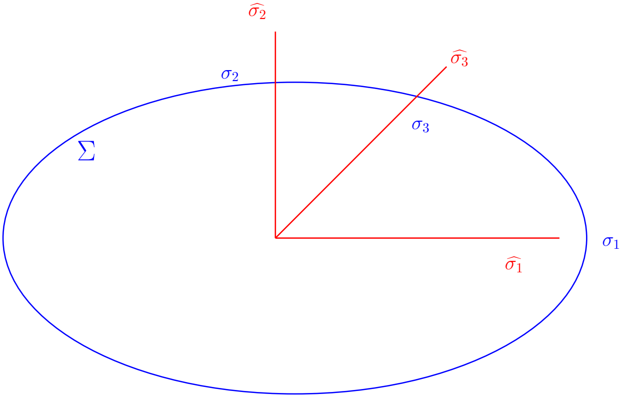
$$\mathbb{E} \left[\mathrm{pen}_{\mathrm{id}}(M) \right] = \frac{2}{np} \mathrm{tr} \left(A_M \left(\Sigma \otimes I_n \right) \right)$$

- Problem: How to estimate Σ ?

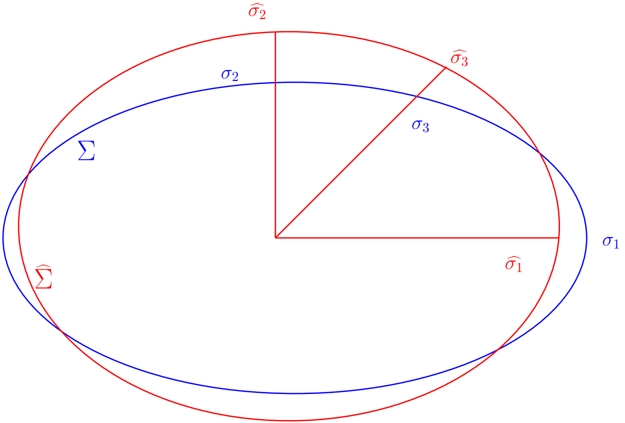
Estimating the covariance matrix: idea ($p = 2$)



Estimating the covariance matrix: idea ($p = 2$)



Estimating the covariance matrix: idea ($p = 2$)



Estimating the covariance matrix: algorithm

- for every $j \in \{1, \ldots, p\}$, apply the "minimal penalties" algorithm to the data set $(Y_i^j)_{1 \le i \le n}$
  $\Rightarrow$ estimator $a(e_j)$ of $\Sigma_{j,j}$
- for every $j \neq \ell \in \{1, \ldots, p\}$, apply the "minimal penalties" algorithm to the data set $(Y_i^j + Y_i^\ell)_{1 \le i \le n}$
  $\Rightarrow$ estimator $a(e_j + e_\ell)$ of $\Sigma_{j,j} + \Sigma_{\ell,\ell} + 2\Sigma_{j,\ell}$

Estimating the covariance matrix: algorithm

- for every $j \in \{1, \ldots, p\}$, apply the "minimal penalties" algorithm to the data set $(Y_i^j)_{1 \leq i \leq n}$
  $\Rightarrow$ estimator $a(e_j)$ of $\Sigma_{j,j}$
- for every $j \neq \ell \in \{1, \ldots, p\}$, apply the "minimal penalties" algorithm to the data set $(Y_i^j + Y_i^\ell)_{1 \leq i \leq n}$
  $\Rightarrow$ estimator $a(e_j + e_\ell)$ of $\Sigma_{j,j} + \Sigma_{\ell,\ell} + 2\Sigma_{j,\ell}$

- Recover an estimator $\widehat{\Sigma}$ of Σ:

$$\widehat{\Sigma} = J\left(a(e_1), \ldots, a(e_p), a(e_1 + e_2), \ldots, a(e_{p-1} + e_p)\right)$$

where J is the unique linear application $R^{p(p+1)/2} \mapsto \mathcal{S}_p(\mathbb{R})$ such that

$$\Sigma = J\left(\Sigma_{1,1}, \ldots, \Sigma_{p,p}, \Sigma_{1,1} + \Sigma_{2,2} + 2\Sigma_{1,2}, \ldots, \Sigma_{p-1,p-1} + \Sigma_{p,p} + 2\Sigma_{p-1,p}\right)$$

Theorem: Estimating the covariance matrix

Theorem (Solnon, A. & Bach, 2011)

If for every $j = 1, \ldots, p$, some $\lambda_j > 0$ exists such that $\operatorname{tr}(A_{\lambda_j}) \leq \sqrt{n}$ *and*

$$\frac{1}{n} \left\| (I_n - A_{\lambda_j}) F^j \right\|^2 \leq \Sigma_{j,j} \sqrt{\frac{\ln(n)}{n}} \quad \text{where} \quad A_{\lambda_j} = K(K + n\lambda_j I_n)^{-1} \ ,$$

Then, with probability $1 - L_5 p^2 n^{-\delta}$, if $n \geq n_0(\delta)$,

$$(1 - \eta)\Sigma \preceq \widehat{\Sigma} \preceq (1 + \eta)\Sigma \quad \text{with} \quad \eta := L(2 + \delta) c(\Sigma)^2 p \sqrt{\frac{\ln(n)}{n}}$$

where $c(\Sigma) = \max(\operatorname{Sp}(\Sigma)) / \min(\operatorname{Sp}(\Sigma))$.

$\Rightarrow$ sufficient condition for consistency

Theorem: Oracle inequality

Theorem (Solnon, A. & Bach, 2011)

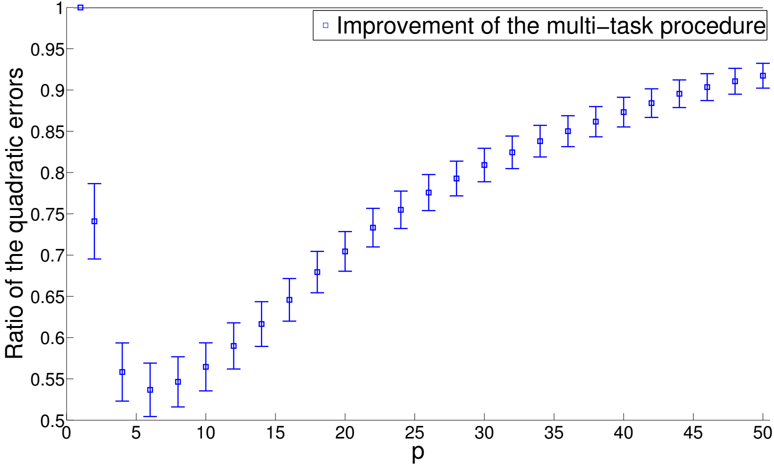
If moreover matrices $M \in \mathcal{M}$ can be diagonalized in the same orthogonal basis, and if

$$\widehat{M} \in \arg\min_{M \in \mathcal{M}} \left\{ \frac{1}{np} \left\| \widehat{F}_M - Y \right\|^2 + \frac{2}{np} \operatorname{tr}\left(A_M \left(\widehat{\Sigma} \otimes I_n \right) \right) \right\} ,$$

Then, with probability $1 - L_5 p^2 n^{-\delta}$, if $n \geq n_0(\delta)$,

$$\frac{1}{np} \left\| \widehat{F}_{\widehat{M}} - F \right\|^2 \leq \left(1 + \frac{1}{\ln(n)} \right)^2 \inf_{M \in \mathcal{M}} \left\{ \frac{1}{np} \left\| \widehat{F}_M - F \right\|^2 \right\} + L(2+\delta)^2 c(\Sigma)^4 \frac{\operatorname{tr}(\Sigma)}{p} \frac{p^3 \ln(n)^3}{n}$$

Simulations: $n = 100$, $2 \leq p \leq 50$, $1.1 \leq c(\Sigma) \leq 22.5$



Summary

- Minimal penalties: efficient for data-driven calibration of multiplicative constants in penalties

Summary

- Minimal penalties: efficient for data-driven calibration of multiplicative constants in penalties
- "Slope heuristics": $\mathrm{pen}_{\mathrm{opt}} \,/\, \mathrm{pen}_{\mathrm{min}} \approx 2$

Summary

- Minimal penalties: efficient for data-driven calibration of multiplicative constants in penalties
- "Slope heuristics": $\mathrm{pen}_{\mathrm{opt}} \,/\, \mathrm{pen}_{\mathrm{min}} \approx 2$
- $\mathrm{pen}_{\mathrm{opt}} \,/\, \mathrm{pen}_{\mathrm{min}} \in (1; 2]$ for linear estimators

Summary

- Minimal penalties: efficient for data-driven calibration of multiplicative constants in penalties
- "Slope heuristics": $\mathrm{pen}_{\mathrm{opt}} / \mathrm{pen}_{\mathrm{min}} \approx 2$
- $\mathrm{pen}_{\mathrm{opt}} / \mathrm{pen}_{\mathrm{min}} \in (1;2]$ for linear estimators

$\Rightarrow$ Extend/adapt results to other estimators/learning algorithms (e.g., SVM or Lasso-type)? Other losses?
Key question: compute/estimate the expectation and prove concentration inequalities for

$$p_1(m) = \mathcal{R}(\widehat{s}_m) - \mathcal{R}(s^\star_m) \quad \text{and} \quad p_2(m) = \widehat{\mathcal{R}}_n(s^\star_m) - \widehat{\mathcal{R}}_n(\widehat{s}_m)$$

Summary

- Minimal penalties: efficient for data-driven calibration of multiplicative constants in penalties
- "Slope heuristics": $\mathrm{pen}_{\mathrm{opt}} / \mathrm{pen}_{\mathrm{min}} \approx 2$
- $\mathrm{pen}_{\mathrm{opt}} / \mathrm{pen}_{\mathrm{min}} \in (1; 2]$ for linear estimators

$\Rightarrow$ Extend/adapt results to other estimators/learning algorithms (e.g., SVM or Lasso-type)? Other losses?
Key question: compute/estimate the expectation and prove concentration inequalities for

$$p_1(m) = \mathcal{R}(\widehat{s}_m) - \mathcal{R}(s_m^\star) \quad \text{and} \quad p_2(m) = \widehat{\mathcal{R}}_n(s_m^\star) - \widehat{\mathcal{R}}_n(\widehat{s}_m)$$

$\Rightarrow$ What if $\mathcal{M}$ is "large" (e.g., variable selection with $p \geq n$ explanatory variables)?