Data-driven calibration of linear estimators with minimal penalties, with an application to multi-task regression

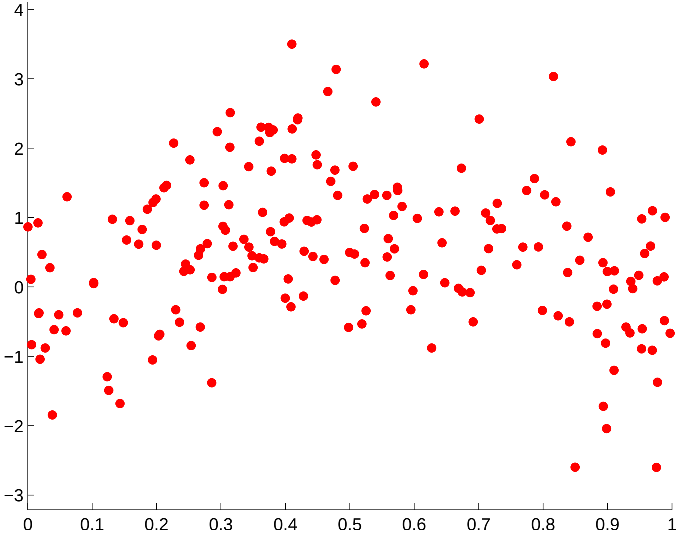
Sylvain Arlot (joint works with F. Bach & M. Solnon)

[1]CNRS

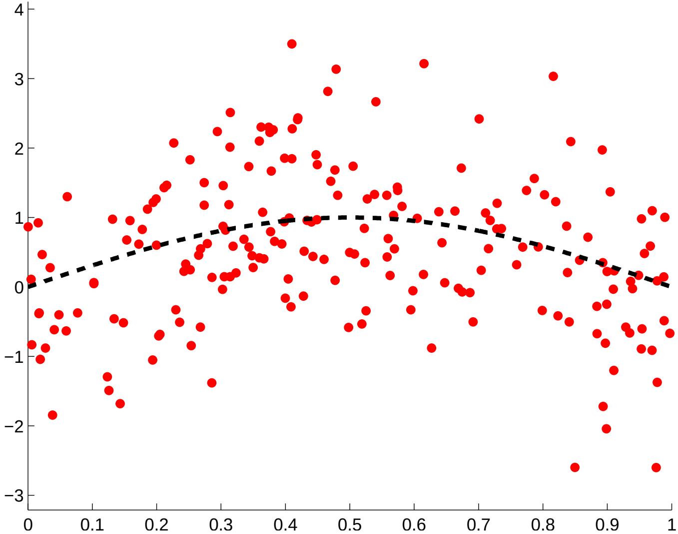
[2]École Normale Supérieure (Paris), LIENS, Équipe SIERRA

Cambridge, November, 4th 2011

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



Goal: find the signal (denoising)



Statistical framework: regression, least-squares loss

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

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

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

Statistical framework: regression, least-squares loss

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

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

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

- Fixed design: x_i deterministic

Statistical framework: regression, least-squares loss

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

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

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

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

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

Statistical framework: regression, least-squares loss

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

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

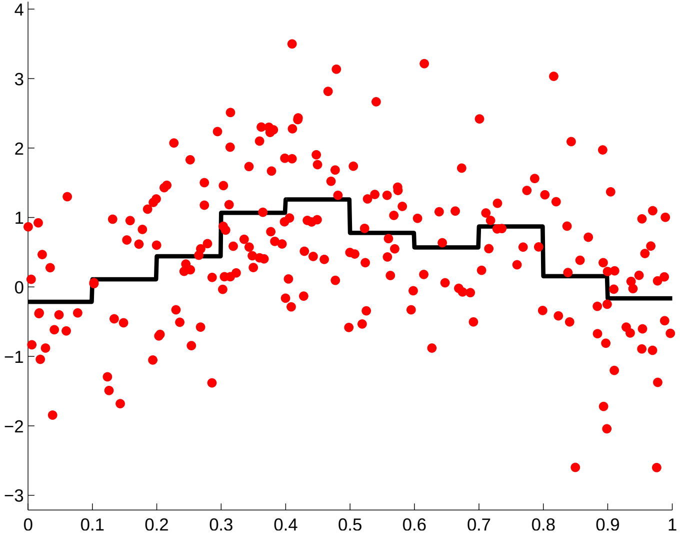
with $Y_i \in \mathbb{R}$, $(\varepsilon_i)_{1 \le i \le n}$ i.i.d.

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

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

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

Estimators: example: regressogram



Least-squares estimators

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

Least-squares estimators

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

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

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

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:

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

so that

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

Model examples

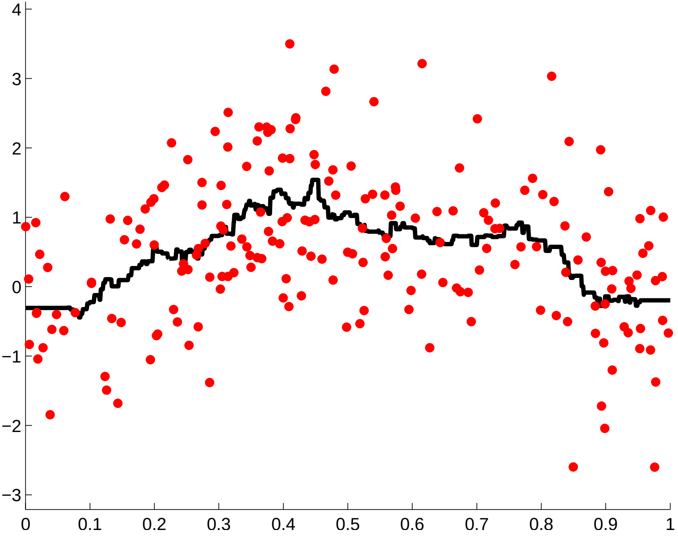
- histograms on some partition Λ of $\mathcal{X}$
  $\Rightarrow$ the least-squares estimator (regressogram) can be written

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

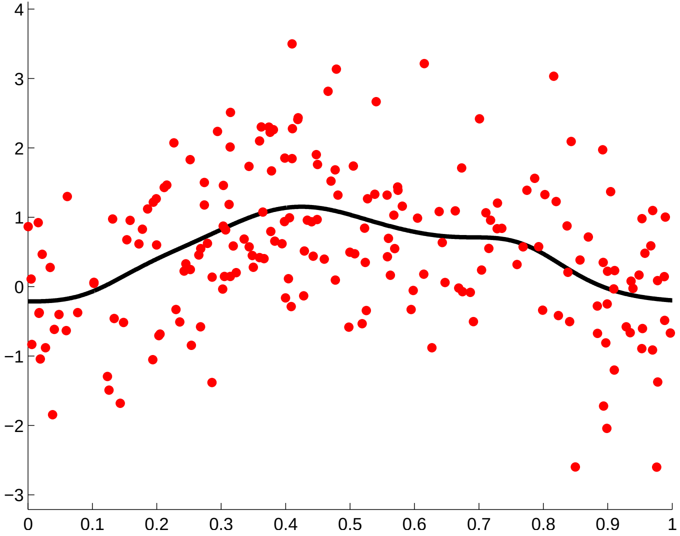
- subspace generated by a subset of an orthogonal basis of $L^2(\mu)$ (Fourier, wavelets, and so on)
- variable selection: $x_i = \left(x_i^{(1)}, \ldots, x_i^{(p)} \right) \in \mathbb{R}^p$ gathers p variables that can (linearly) explain Y_i

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

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:

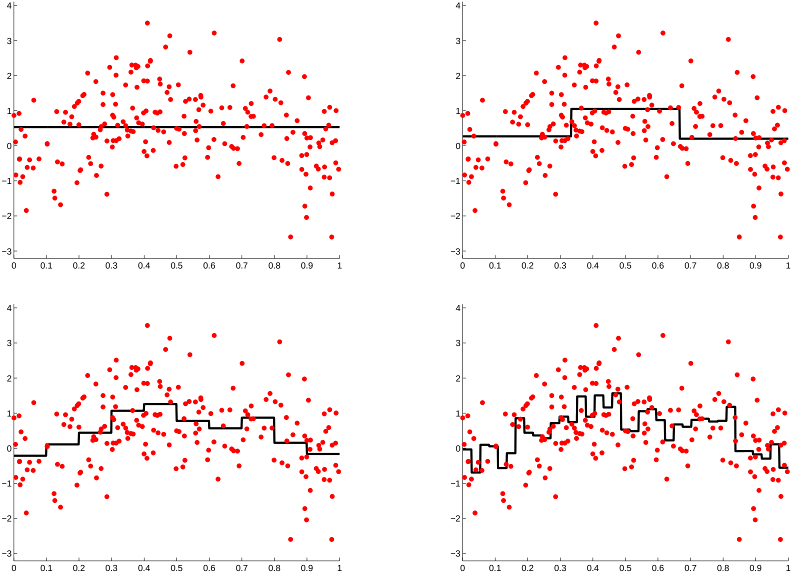
$$\widehat{F}_i = \widehat{f}(x_i) \quad \text{with} \quad \widehat{f} \in \arg\min_{f \in \mathcal{F}_K} \left\{ \frac{1}{n} \sum_{i=1}^{n} \left(Y_i - f(x_i) \right)^2 + \lambda \left\| f \right\|_{\mathcal{F}_K}^2 \right\}$$

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

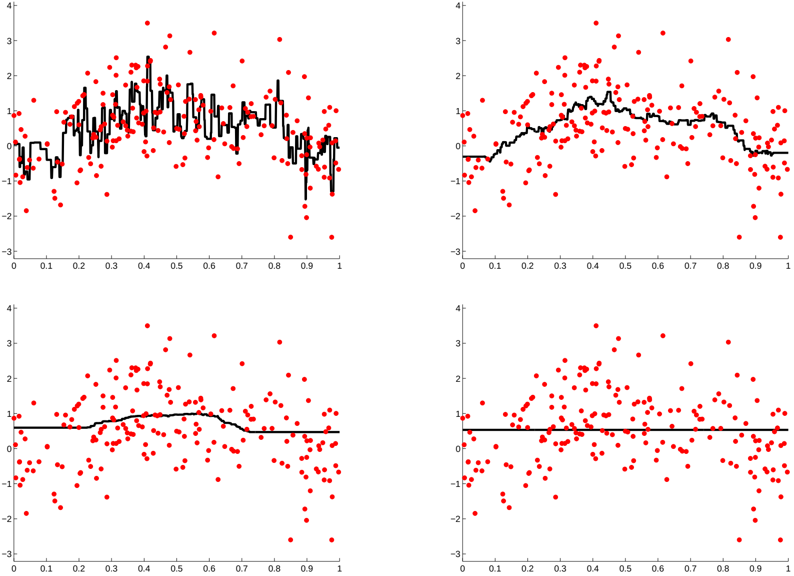
- k-nearest neighbours
- Nadaraya-Watson estimators

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

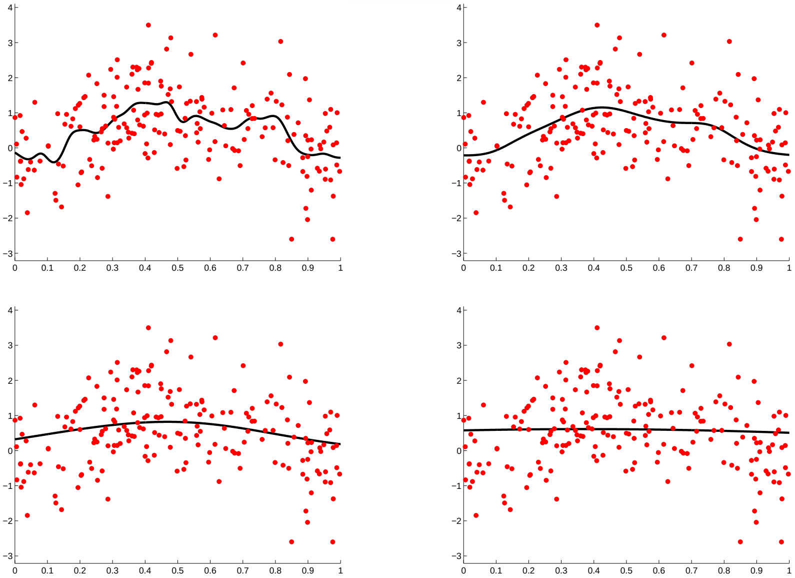
Estimator selection: regular regressograms



Estimator selection: k nearest neighbours



Estimator selection: Nadaraya-Watson



Estimator selection

- Estimator collection $(\widehat{F}_m)_{m \in \mathcal{M}} \Rightarrow \widehat{m}(Y)$?
  Example: $\widehat{F}_m = A_m Y$

Estimator selection

- Estimator collection $(\widehat{F}_m)_{m \in \mathcal{M}} \Rightarrow \widehat{m}(Y)$?
  Example: $\widehat{F}_m = A_m Y$

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

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

Estimator selection

- Estimator collection $(\widehat{F}_m)_{m \in \mathcal{M}} \Rightarrow \widehat{m}(Y)$?
  Example: $\widehat{F}_m = A_m Y$
- Goal: minimize the risk, i.e.,
  Oracle inequality (in expectation or with a large probability):

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

- Examples:
  - model selection
  - calibration (choosing k or the distance for k-NN, choice of a regularization parameter, choice of a kernel, etc.)
  - choice between methods different in nature
    ex.: k-NN vs. smoothing splines?

Bias-variance trade-off

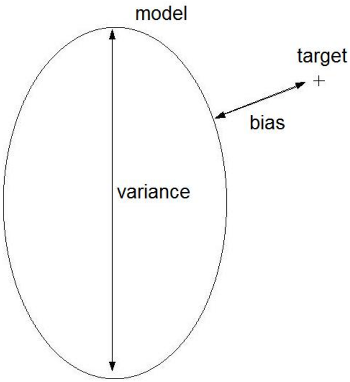
$$\mathbb{E}\left[\frac{1}{n}\left\|\widehat{F}_m - F\right\|^2\right] = \text{Bias} + \text{Variance}$$

Bias or Approximation error

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

Variance or Estimation error

$$\frac{\sigma^2 \operatorname{tr}\left(A_m^\top A_m\right)}{n} \qquad \text{OLS:} \quad \frac{\sigma^2 \dim(S_m)}{n}$$



Bias-variance trade-off

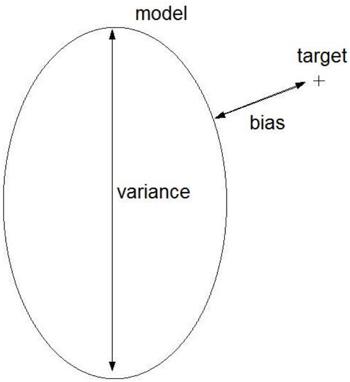
$$\mathbb{E}\left[\frac{1}{n}\left\|\widehat{F}_m - F\right\|^2\right] = \text{Bias} + \text{Variance}$$

Bias or Approximation error

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

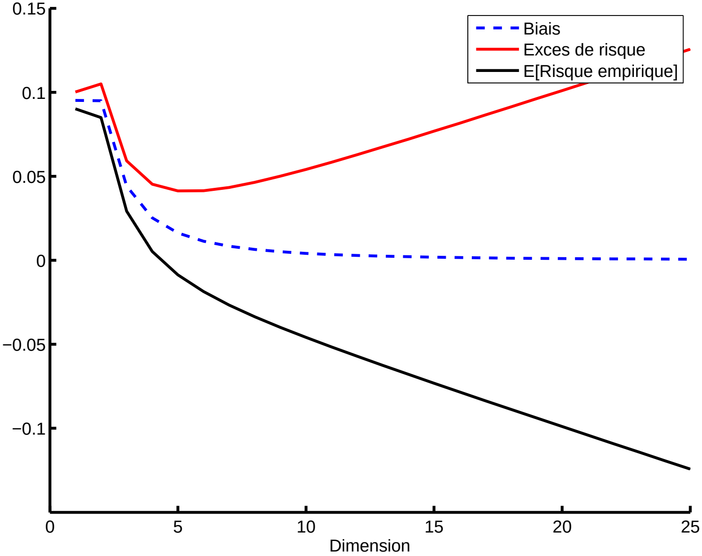
Variance or Estimation error

$$\frac{\sigma^2 \operatorname{tr}\left(A_m^\top A_m\right)}{n} \qquad \text{OLS:} \quad \frac{\sigma^2 \dim(S_m)}{n}$$



Bias-variance trade-off
$\Leftrightarrow$ avoid overfitting and underfitting

Why should the empirical risk be penalized?



Penalization

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

Penalization

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

- 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

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

- 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)

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

$\Rightarrow$ Linear estimators: C_L : $2\sigma^2 \operatorname{tr}(A_m)/n$ (Mallows, 1973)

Oracle inequality

Theorem (Birgé & Massart 2007, A. & Bach 2009–2011)

Assumptions:

- $\mathrm{pen}(m) = \frac{2C \operatorname{tr}(A_m)}{n}$ *with* $\left| C\sigma^{-2} - 1 \right| \leq L_0 \sqrt{\frac{\ln(n)}{n}}$

Then, with probability at least $1 - 3 \operatorname{Card}(\mathcal{M}) n^{-\delta}$, *if* $n \geq n_0$, *for every* $\eta \in (0,1)$,

$$\frac{1}{n} \left\| \widehat{F}_{\widehat{m}} - F \right\|^2 \leq (1+\eta) \inf_{m \in \mathcal{M}} \left\{ \frac{1}{n} \left\| \widehat{F}_m - F \right\|^2 \right\} + K(\delta, L_0, L_1) \frac{\ln(n) \sigma^2}{\eta n}$$

Oracle inequality

Theorem (Birgé & Massart 2007, A. & Bach 2009–2011)

Assumptions:

- $\operatorname{pen}(m) = \frac{2C \operatorname{tr}(A_m)}{n}$ *with* $\left| C\sigma^{-2} - 1 \right| \leq L_0 \sqrt{\frac{\ln(n)}{n}}$

Then, with probability at least $1 - 3 \operatorname{Card}(\mathcal{M}) n^{-\delta}$*, if* $n \geq n_0$*, for every* $\eta \in (0,1)$*,*

$$\frac{1}{n} \left\| \widehat{F}_{\widehat{m}} - F \right\|^2 \leq (1+\eta) \inf_{m \in \mathcal{M}} \left\{ \frac{1}{n} \left\| \widehat{F}_m - F \right\|^2 \right\} + K(\delta, L_0, L_1) \frac{\ln(n) \sigma^2}{\eta n}$$

Note: ridge, $\lambda \in [0, +\infty] \;\Leftrightarrow\; \operatorname{Card}(\mathcal{M}) \propto n$

Oracle inequality

Theorem (Birgé & Massart 2007, A. & Bach 2009–2011)

Assumptions:

- $\mathrm{pen}(m) = \frac{2C\,\mathrm{tr}(A_m)}{n}$ *with* $\left|C\sigma^{-2} - 1\right| \leq L_0 \sqrt{\frac{\ln(n)}{n}}$
- $\forall m \in \mathcal{M}$, $\|A_m\| \leq L_1$ *and* $\mathrm{tr}(A_m^\top A_m) \leq \mathrm{tr}(A_m) \leq n$
- *Gaussian noise:* $\varepsilon_i \sim \mathcal{N}(0, \sigma^2)$

Then, with probability at least $1 - 3\,\mathrm{Card}(\mathcal{M}) n^{-\delta}$, *if* $n \geq n_0$, *for every* $\eta \in (0,1)$,

$$\frac{1}{n}\left\|\widehat{F}_{\widehat{m}} - F\right\|^2 \leq (1+\eta) \inf_{m \in \mathcal{M}} \left\{ \frac{1}{n}\left\|\widehat{F}_m - F\right\|^2 \right\} + K(\delta, L_0, L_1)\frac{\ln(n)\sigma^2}{\eta n}$$

Note: ridge, $\lambda \in [0, +\infty] \Leftrightarrow \mathrm{Card}(\mathcal{M}) \propto n$

Motivation (1): penalties known up to a constant factor

Ex.: $\mathrm{pen}_{\mathrm{Cp}}(m) = \dfrac{2\sigma^2 D_m}{n}$ $\qquad$ $\mathrm{pen}_{\mathrm{CL}}(m) = \dfrac{2\sigma^2 \operatorname{tr}(A_m)}{n}$

Motivation (1): penalties known up to a constant factor

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

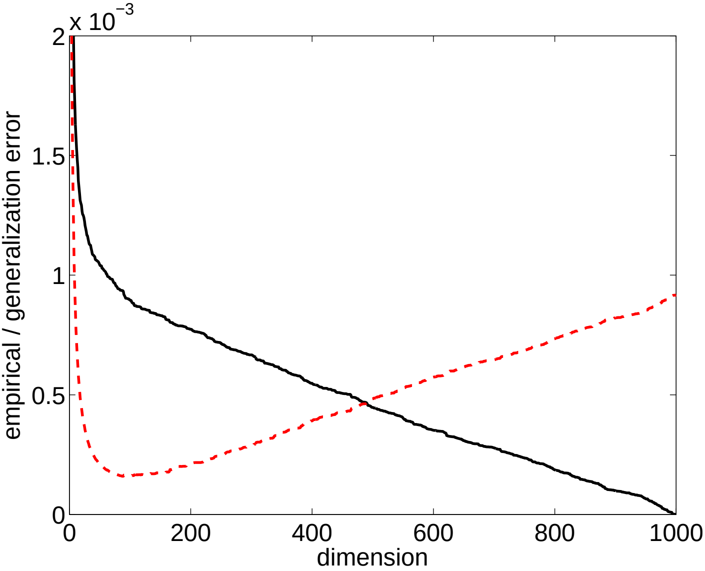
- Classical estimators of σ^2 :
  - $\widehat{\sigma}^2_{m_0} := \| Y - \widehat{F}_{m_0} \|^2 / (n - D_{m_0})$ (OLS)
    problem: choice of m_0 ?
  - $\widehat{\sigma}^2_m \Rightarrow$ FPE (Akaike, 1970) and GCV (Craven & Wahba, 1979)
    problem: avoiding the largest models

Motivation (1): penalties known up to a constant factor

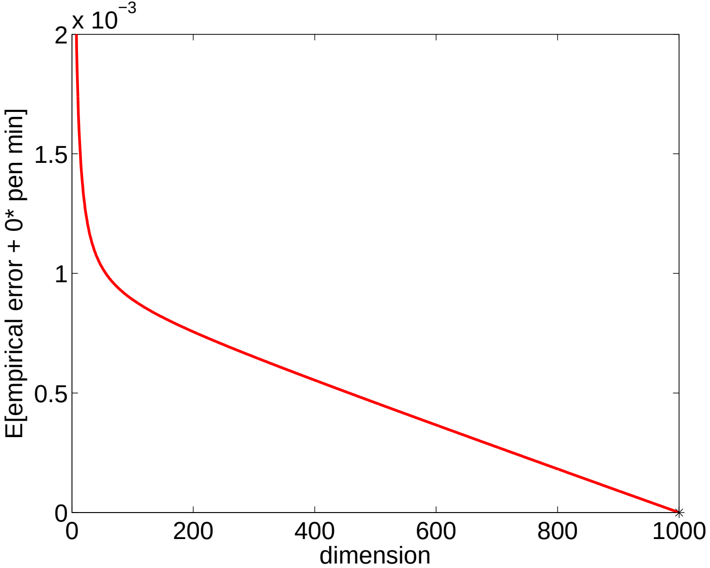
Ex.: $\mathrm{pen}_{\mathrm{Cp}}(m) = \dfrac{2\sigma^2 D_m}{n}$ $\qquad \mathrm{pen}_{\mathrm{CL}}(m) = \dfrac{2\sigma^2 \operatorname{tr}(A_m)}{n}$

- Classical estimators of σ^2 :
  - $\widehat{\sigma}^2_{m_0} := \| Y - \widehat{F}_{m_0} \|^2 / (n - D_{m_0})$ (OLS)
    problem: choice of m_0 ?
  - $\widehat{\sigma}^2_m \Rightarrow$ FPE (Akaike, 1970) and GCV (Craven & Wahba, 1979)
    problem: avoiding the largest models
- Goals: estimation of σ^2 for model selection, under minimal assumptions, without overfitting

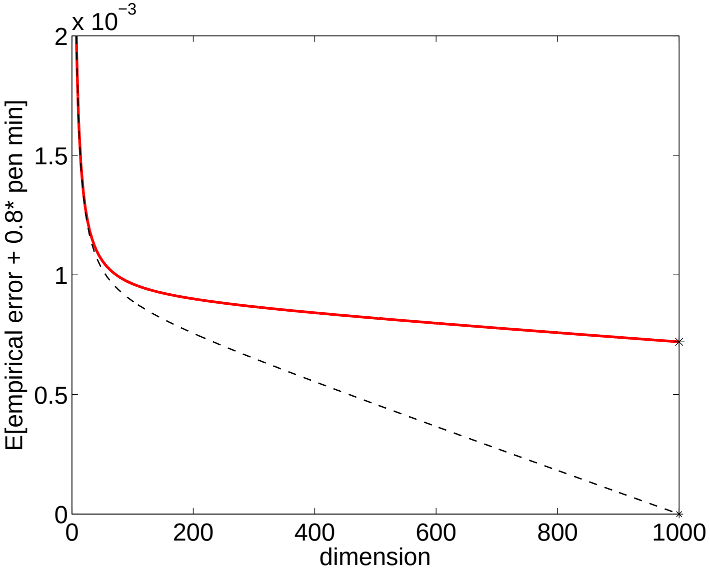
Motivation (2): “L-curve” and elbow heuristics?



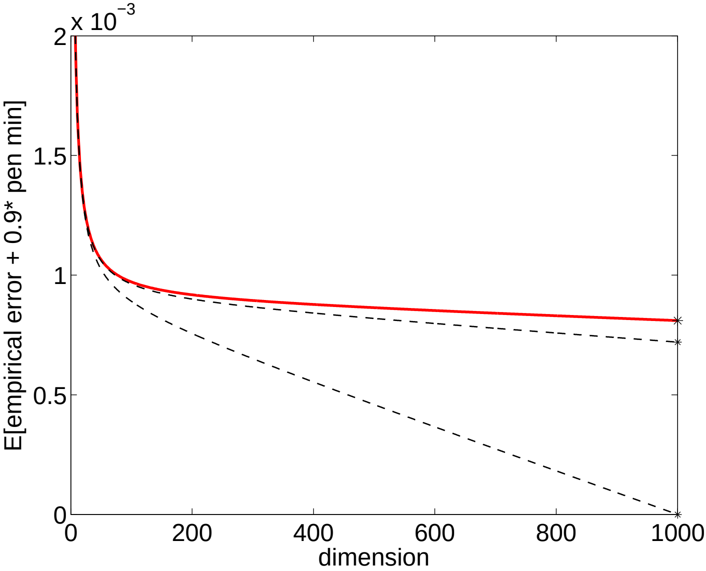
$\mathbb{E}[\text{ Empirical risk }] + 0 \times \sigma^2 D_m n^{-1}$ (OLS)



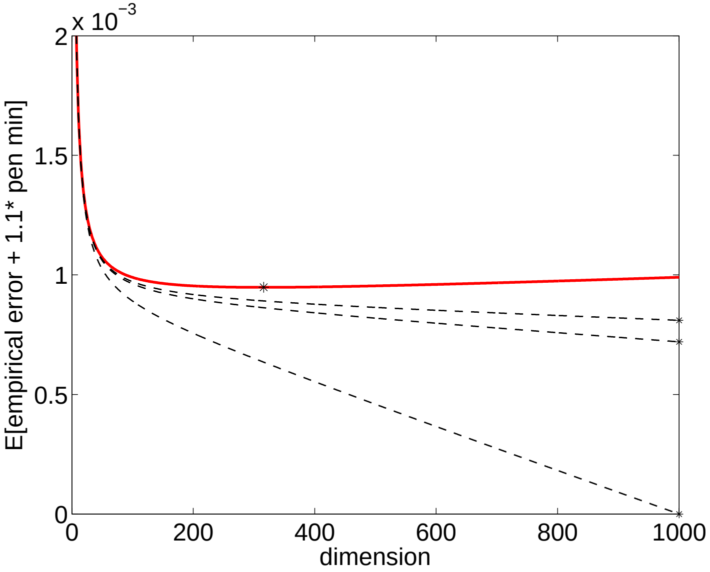
$\mathbb{E}[$ Empirical risk $] + 0.8 \times \sigma^2 D_m n^{-1}$ (OLS)



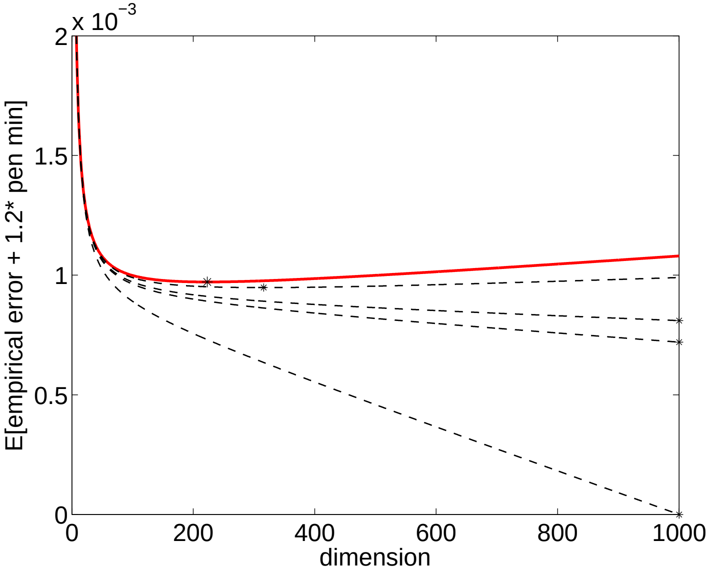
$\mathbb{E}[$ Empirical risk $] + 0.9 \times \sigma^2 D_m n^{-1}$ (OLS)



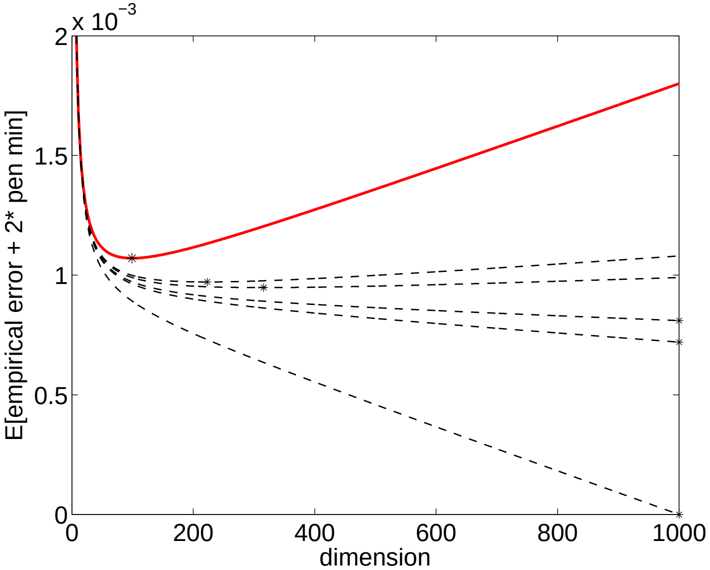
$\mathbb{E}[$ Empirical risk $] + 1.1 \times \sigma^2 D_m n^{-1}$ (OLS)



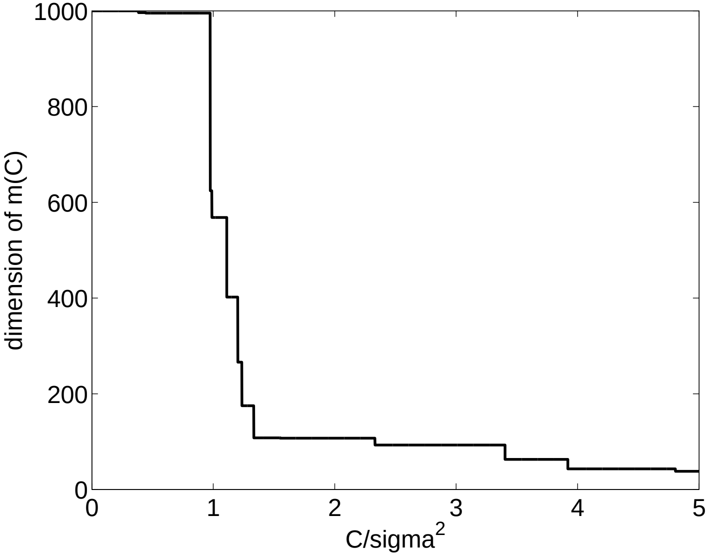
$\mathbb{E}$[Empirical risk] + $1.2 \times \sigma^2 D_m n^{-1}$ (OLS)



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



OLS: Dimension jump



OLS: algorithm (Birgé & Massart 2007)

1. for every $C > 0$, compute

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

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

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

Practical use: CAPUSHE package (Baudry, Maugis & Michel, 2010)
`http://www.math.univ-toulouse.fr/~maugis/CAPUSHE.html`

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

Theorem (1): Dimension jump / Minimal penalty

Theorem (Birgé & Massart 2007, A. & Bach 2009–2011)

Assumptions:

- $\exists m_0 \in \mathcal{M}$, $D_{m_0} \leq \sqrt{n}$ *and* $\frac{1}{n} \left\| F_{m_0} - F \right\|^2 \leq \sigma^2 \sqrt{\ln(n)/n}$.
- *Gaussian noise:* $\varepsilon_i \sim \mathcal{N}(0, \sigma^2)$

Then, with probability at least $1 - 3 \operatorname{Card}(\mathcal{M}) n^{-\delta}$, *if* $n \geq n_0(\delta)$,

$$\forall C < \left(1 - L_3 \delta \sqrt{\frac{\ln(n)}{n}} \right) \sigma^2 , \quad D_{\widehat{m}(C)} \geq \frac{n}{3}$$

$$\forall C > \left(1 + L_3 \delta \sqrt{\frac{\ln(n)}{n}} \right) \sigma^2 , \quad D_{\widehat{m}(C)} \leq \frac{n}{10}$$

and in the first case, $\frac{1}{n} \| \widehat{F}_{\widehat{m}(C)} - F \|^2 \gg \inf_{m \in \mathcal{M}_n} \{ \frac{1}{n} \| \widehat{F}_m - F \|^2 \}$.

Theorem (2): Oracle inequality

Theorem (Birgé & Massart 2007, A. & Bach 2009–2011)

Assumptions:

- $\widehat{m} \in \arg\min_{m \in \mathcal{M}} \{ \frac{1}{n} \| \widehat{F}_m - Y \|^2 + 2 \widehat{C}_{\min} \frac{D_m}{n} \}$
- $\exists m_0 \in \mathcal{M}$, $D_{m_0} \leq \sqrt{n}$ *and* $\frac{1}{n} \| F_{m_0} - F \|^2 \leq \sigma^2 \sqrt{\ln(n)/n}$.
- *Gaussian noise:* $\varepsilon_i \sim \mathcal{N}(0, \sigma^2)$

Then, with probability at least $1 - 3 \operatorname{Card}(\mathcal{M}) n^{-\delta}$*, if* $n \geq n_0(\delta)$*, for every* $\eta > 0$,

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

Generalization of minimal penalties to linear estimators?

OLS

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

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

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

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

Generalization of minimal penalties to linear estimators?

OLS

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

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

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

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

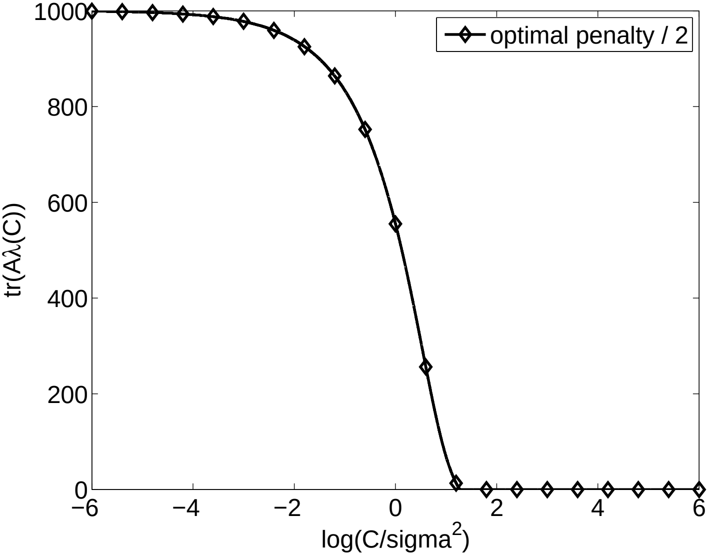
Linear estimators

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

Generalization of minimal penalties to linear estimators?

OLS	Linear estimators
$\mathrm{pen}_{\mathrm{Cp}}(m) = \dfrac{2\sigma^2 D_m}{n}$	$\mathrm{pen}_{\mathrm{CL}}(m) = \dfrac{2\sigma^2 \operatorname{tr}(A_m)}{n}$
$\arg\min_{m \in \mathcal{M}} \left\{ \dfrac{1}{n} \lVert \widehat{F}_m - Y \rVert^2 + C \dfrac{D_m}{n} \right\}$	$\arg\min_{m \in \mathcal{M}} \left\{ \dfrac{1}{n} \lVert \widehat{F}_m - Y \rVert^2 + C \dfrac{\operatorname{tr}(A_m)}{n} \right\}$
$\Rightarrow D_{\widehat{m}(C)}$ "jumps" at $\widehat{C}_{\min} \approx \sigma^2$	Does $\operatorname{tr}(A_{\widehat{m}(C)})$ jump at $\widehat{C}_{\min} \approx \sigma^2$?
$\Rightarrow$ optimal choice with $\widehat{m}(2\widehat{C}_{\min})$	optimal choice with $\widehat{m}(2\widehat{C}_{\min})$?

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

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

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

Minimal penalties for linear estimators

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

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

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

Minimal penalties for linear estimators

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

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

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

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

Minimal penalties for linear estimators

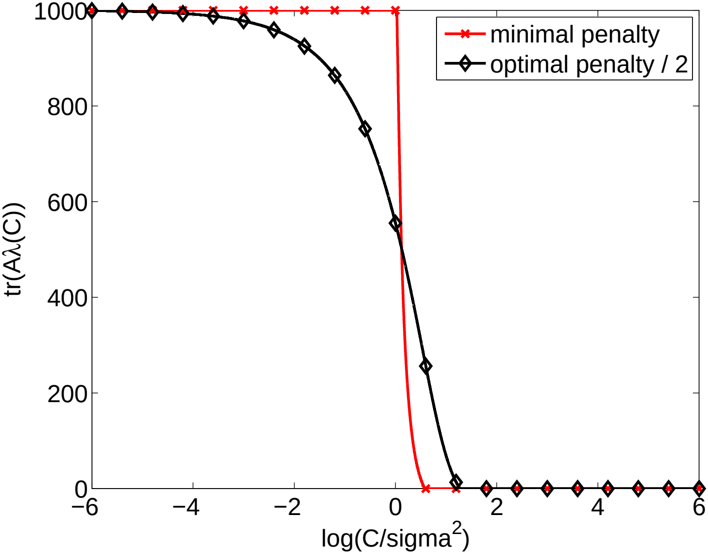
$$\mathbb{E}\left[\frac{1}{n}\left\|\widehat{F}_m - F\right\|^2\right] = \frac{1}{n}\|(I - A_m)F\|^2 + \frac{\operatorname{tr}(A_m^\top A_m)\sigma^2}{n} = \text{bias} + \text{variance}$$

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

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

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

"Dimension" jump (ridge regression)



Penalty calibration algorithm (A. & Bach 2009)

1. for every $C > 0$, compute

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

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

3. select

$$\widehat{m} \in \arg\min_{m \in \mathcal{M}} \left\{ \frac{1}{n} \left\lVert Y - \widehat{F}_m \right\rVert^2 + \frac{2 \widehat{C}_{\min} \operatorname{tr}(A_m)}{n} \right\}$$

Penalty calibration algorithm (A. & Bach 2009)

1. for every $C > 0$, compute
$$\widehat{m}_{\min}(C) \in \arg\min_{m \in \mathcal{M}} \left\{ \frac{1}{n} \left\| Y - \widehat{F}_m \right\|^2 + \frac{C \left(2 \operatorname{tr}(A_m) - \operatorname{tr}(A_m^\top A_m) \right)}{n} \right\}$$

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

3. select
$$\widehat{m} \in \arg\min_{m \in \mathcal{M}} \left\{ \frac{1}{n} \left\| Y - \widehat{F}_m \right\|^2 + \frac{2 \widehat{C}_{\min} \operatorname{tr}(A_m)}{n} \right\}$$

$\Rightarrow \left| \widehat{C}_{\min} \sigma^{-2} - 1 \right| \leq L_4 \sqrt{\ln(n)/n}$ and oracle inequality (same assumptions as before).

Comparison with least-squares

- Linear estimators:

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

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

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

Comparison with least-squares

- Linear estimators:

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

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

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

- 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

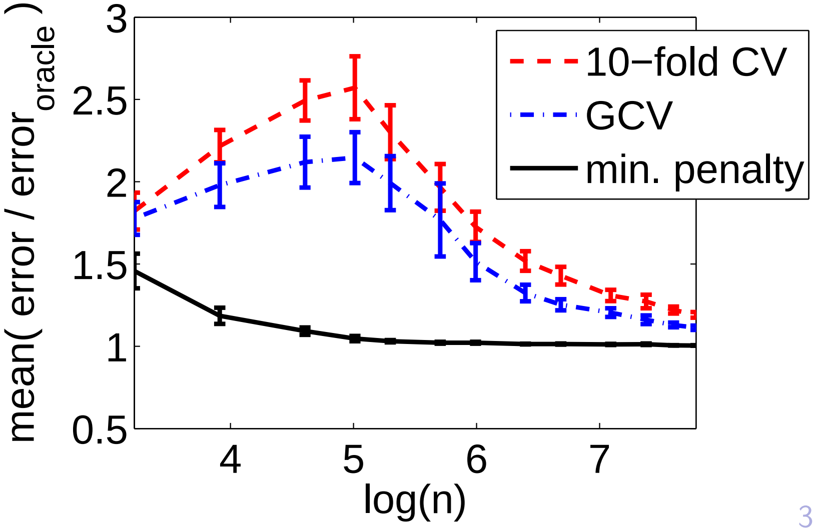
$$\forall i,j \in \{1, \ldots, n\} \, , \quad A_{i,j} \in \left\{ 0, \frac{1}{k} \right\}$$

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

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

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

Simulation study (ridge regression, choice of λ)



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 $t \in \mathbb{R}^{np}$ ("$t_i^j = t^j(x_i)$"):

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

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

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:

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

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

Multi-task estimator selection

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

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

- 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 $A_M = \left(M^{-1} \otimes K \right) \left(\left(M^{-1} \otimes K \right) + npI_{np} \right)^{-1}$

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

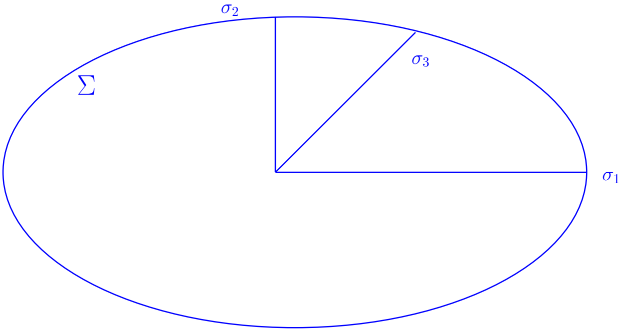
with $\widehat{F}_M = A_M Y$ and $A_M = \left(M^{-1} \otimes K \right) \left(\left(M^{-1} \otimes K \right) + npI_{np} \right)^{-1}$

- 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:

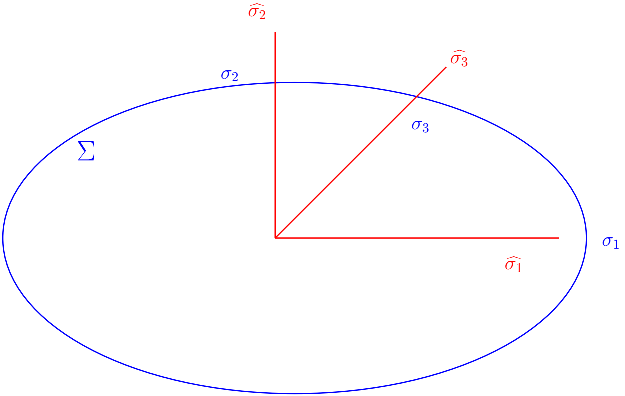
$$\mathbb{E}\left[\mathrm{pen}_{\mathrm{id}}(M) \right] = \frac{2}{np} \operatorname{tr}\left(A_M \left(\Sigma \otimes I_n \right) \right)$$

- Problem: How to estimate Σ ?

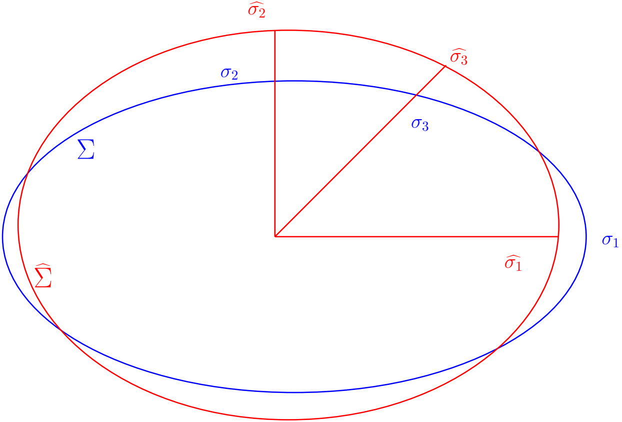
Estimating the covariance matrix: idea ($p = 2$)



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 $(Y_i^j)_{1 \leq i \leq n}$
  $\Rightarrow$ 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 $\Sigma_{j,j} + \Sigma_{\ell,\ell} + 2\Sigma_{j,\ell}$

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)

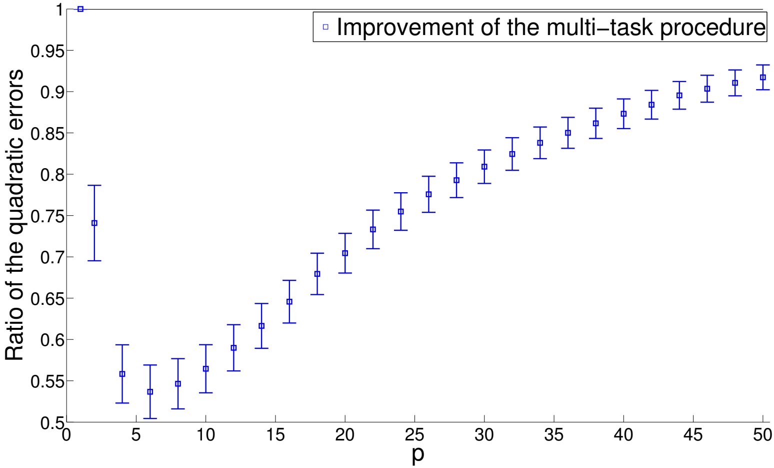
If moreover matrices $M \in \mathcal{M}$ can be diagonalized in the same orthogonal basis, and if

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

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
- $\mathrm{pen}_{\mathrm{opt}} \, / \, \mathrm{pen}_{\mathrm{min}} \approx 2$ for least-squares estimators

Summary

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

Summary

- Minimal penalties: efficient for data-driven calibration of multiplicative constants in penalties
- $\mathrm{pen}_{\mathrm{opt}} / \mathrm{pen}_{\mathrm{min}} \approx 2$ for least-squares estimators
- $\mathrm{pen}_{\mathrm{opt}} / \mathrm{pen}_{\mathrm{min}} \in (1; 2]$ for linear estimators
- Can be applied with a data-driven shape of penalty (e.g., if data are heteroscedastic): V-fold/resampling penalties (OLS: A. 2008, 2009; Lerasle, 2009)

Minimal penalties: which frameworks?

- Theoretical results:
  - OLS, homoscedastic Gaussian regression (Birgé & Massart, 2007)
  - regressograms, heteroscedastic (A. & Massart, 2009)
  - Least-squares density estimation, i.i.d. (Lerasle, 2009) or mixing (Lerasle, 2010) data
  - Linear estimators, regression (A. & Bach, 2009–2011)
  - Minimum contrast estimator, regular contrast (Saumard, 2010)
  - Multitask regression (Solnon, A. & Bach, 2011)
  - ...
- Empirical results:
  - Change-point detection (Lebarbier, 2005)
  - Gaussian mixture models (Maugis & Michel, 2008–2010)
  - Unsupervised classification (Baudry, 2009)
  - Computational geometry (Caillerie & Michel, 2009)
  - Lasso (Connault, 2011)
  - ... (see Baudry, Maugis & Michel, 2011)