Data-driven penalties for model selection

Sylvain Arlot

[1] CNRS

[2] École Normale Supérieure (Paris), LIENS, WILLOW Team

Mathematical Statistics Seminar, WIAS, Berlin, 22/04/2009

Statistical framework: regression on a random design

$$(X_1, Y_1), \ldots, (X_n, Y_n) \in \mathcal{X} \times \mathcal{Y} \quad \text{i.i.d.} \qquad (X_i, Y_i) \sim P \text{ unknown}$$

$$Y = s(X) + \sigma(X)\epsilon \qquad X \in \mathcal{X} \subset \mathbb{R}^d, \quad Y \in \mathcal{Y} = [0;1] \text{ or } \mathbb{R}$$

$$\text{noise } \epsilon : \qquad \mathbb{E}[\epsilon | X] = 0 \quad \mathbb{E}\left[\epsilon^2 | X\right] = 1 \qquad \text{noise level} \quad \sigma(X)$$

$$\text{predictor} \qquad t : \mathcal{X} \mapsto \mathcal{Y} \qquad ?$$

Statistical framework: regression on a random design

$$(X_1, Y_1), \ldots, (X_n, Y_n) \in \mathcal{X} \times \mathcal{Y} \quad \text{i.i.d.} \qquad (X_i, Y_i) \sim P \text{ unknown}$$

$$Y = s(X) + \sigma(X)\epsilon \qquad X \in \mathcal{X} \subset \mathbb{R}^d, \quad Y \in \mathcal{Y} = [0; 1] \text{ or } \mathbb{R}$$

$$\text{noise } \epsilon : \qquad \mathbb{E}[\epsilon | X] = 0 \quad \mathbb{E}\left[\epsilon^2 | X\right] = 1 \qquad \text{noise level} \quad \sigma(X)$$

$$\text{predictor} \qquad t : \mathcal{X} \mapsto \mathcal{Y} \qquad ?$$

Statistical framework: regression on a random design

$$(X_1, Y_1), \ldots, (X_n, Y_n) \in \mathcal{X} \times \mathcal{Y} \quad \text{i.i.d.} \qquad (X_i, Y_i) \sim P \text{ unknown}$$

$$Y = s(X) + \sigma(X)\epsilon \qquad X \in \mathcal{X} \subset \mathbb{R}^d, \quad Y \in \mathcal{Y} = [0;1] \text{ or } \mathbb{R}$$

$$\text{noise } \epsilon : \qquad \mathbb{E}[\epsilon | X] = 0 \quad \mathbb{E}\left[\epsilon^2 | X\right] = 1 \qquad \text{noise level} \quad \sigma(X)$$

$$\text{predictor} \qquad t : \mathcal{X} \mapsto \mathcal{Y} \qquad ?$$

Statistical framework: regression on a random design

$$(X_1, Y_1), \ldots, (X_n, Y_n) \in \mathcal{X} \times \mathcal{Y} \quad \text{i.i.d.} \qquad (X_i, Y_i) \sim P \text{ unknown}$$

$$Y = s(X) + \sigma(X)\epsilon \qquad X \in \mathcal{X} \subset \mathbb{R}^d, \quad Y \in \mathcal{Y} = [0; 1] \text{ or } \mathbb{R}$$

$$\text{noise } \epsilon : \qquad \mathbb{E}\left[\epsilon | X\right] = 0 \quad \mathbb{E}\left[\epsilon^2 | X\right] = 1 \qquad \text{noise level} \quad \sigma(X)$$

$$\text{predictor} \qquad t : \mathcal{X} \mapsto \mathcal{Y} \qquad ?$$

Loss function, least-square estimator

- Least-square risk:

$$\mathbb{E}\gamma(t,(X,Y)) = P\gamma(t,\cdot)$$
$$\text{with} \quad \gamma(t,(x,y)) = (t(x)-y)^2$$

- Empirical risk minimizer on S_m (= model):

$$\widehat{s}_m \in \arg\min_{t \in S_m} P_n\gamma(t,\cdot) = \arg\min_{t \in S_m} \frac{1}{n}\sum_{i=1}^{n}(t(X_i)-Y_i)^2 \ .$$

- *e.g.*, histograms on a partition $(I_\lambda)_{\lambda\in\Lambda_m}$ of $\mathcal{X}$.

$$\widehat{s}_m = \sum_{\lambda\in\Lambda_m}\widehat{\beta}_\lambda \mathbb{1}_{I_\lambda} \qquad \widehat{\beta}_\lambda = \frac{1}{\mathrm{Card}\{X_i \in I_\lambda\}}\sum_{X_i\in I_\lambda} Y_i \ .$$

Loss function, least-square estimator

- Loss function:

$$\ell(s,t) = P\gamma(t,\cdot) - P\gamma(s,\cdot) = \mathbb{E}\left[\left(t(X) - s(X)\right)^2 \right]$$

with $\quad \gamma(t,(x,y)) = (t(x) - y)^2$

- Empirical risk minimizer on S_m (= model):

$$\widehat{s}_m \in \arg\min_{t \in S_m} P_n \gamma(t,\cdot) = \arg\min_{t \in S_m} \frac{1}{n} \sum_{i=1}^{n} (t(X_i) - Y_i)^2 \enspace .$$

- *e.g.*, histograms on a partition $(I_\lambda)_{\lambda \in \Lambda_m}$ of $\mathcal{X}$.

$$\widehat{s}_m = \sum_{\lambda \in \Lambda_m} \widehat{\beta}_\lambda \mathbb{1}_{I_\lambda} \qquad \widehat{\beta}_\lambda = \frac{1}{\operatorname{Card}\{X_i \in I_\lambda\}} \sum_{X_i \in I_\lambda} Y_i \enspace .$$

Loss function, least-square estimator

- Loss function:

$$\ell(s,t) = P\gamma(t,\cdot) - P\gamma(s,\cdot) = \mathbb{E}\left[(t(X) - s(X))^2 \right]$$

with $\quad \gamma(t,(x,y)) = (t(x) - y)^2$

- Empirical risk minimizer on S_m ($=$ model):

$$\widehat{s}_m \in \arg\min_{t \in S_m} P_n \gamma(t,\cdot) = \arg\min_{t \in S_m} \frac{1}{n} \sum_{i=1}^{n} (t(X_i) - Y_i)^2 \enspace .$$

- *e.g.*, histograms on a partition $(I_\lambda)_{\lambda \in \Lambda_m}$ of $\mathcal{X}$.

$$\widehat{s}_m = \sum_{\lambda \in \Lambda_m} \widehat{\beta}_\lambda \mathbb{1}_{I_\lambda} \qquad \widehat{\beta}_\lambda = \frac{1}{\mathrm{Card}\{X_i \in I_\lambda\}} \sum_{X_i \in I_\lambda} Y_i \enspace .$$

Loss function, least-square estimator

- Loss function:

$$\ell(s,t) = P\gamma(t,\cdot) - P\gamma(s,\cdot) = \mathbb{E}\left[(t(X) - s(X))^2\right]$$

with $\quad \gamma(t,(x,y)) = (t(x) - y)^2$

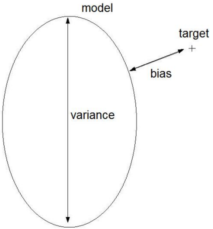
- Empirical risk minimizer on S_m ($=$ model):

$$\widehat{s}_m \in \arg\min_{t \in S_m} P_n\gamma(t,\cdot) = \arg\min_{t \in S_m} \frac{1}{n}\sum_{i=1}^{n} (t(X_i) - Y_i)^2 \ .$$

- *e.g.*, histograms on a partition $(I_\lambda)_{\lambda\in\Lambda_m}$ of $\mathcal{X}$.

$$\widehat{s}_m = \sum_{\lambda\in\Lambda_m} \widehat{\beta}_\lambda \mathbb{1}_{I_\lambda} \qquad \widehat{\beta}_\lambda = \frac{1}{\mathrm{Card}\{X_i \in I_\lambda\}} \sum_{X_i \in I_\lambda} Y_i \ .$$

Model selection



$$(S_m)_{m\in\mathcal{M}} \longrightarrow (\widehat{s}_m)_{m\in\mathcal{M}} \longrightarrow \widehat{s}_{\widehat{m}} \quad ???$$

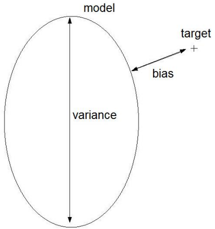
Goals:

- Oracle inequality (in expectation, or with a large probability):

$$\ell(s, \widehat{s}_{\widehat{m}}) \leq C \inf_{m\in\mathcal{M}} \left\{ \ell(s, \widehat{s}_m) + R(m,n) \right\}$$

- Adaptivity (provided $(S_m)_{m\in\mathcal{M}_n}$ is well chosen), e.g., to the smoothness of s or to the variations of σ

Model selection



$$(S_m)_{m\in\mathcal{M}} \longrightarrow (\widehat{s}_m)_{m\in\mathcal{M}} \longrightarrow \widehat{s}_{\widehat{m}} \quad ???$$

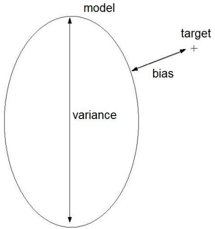
Goals:

- Oracle inequality (in expectation, or with a large probability):

$$\ell(s,\widehat{s}_{\widehat{m}}) \leq C \inf_{m\in\mathcal{M}} \{\ell(s,\widehat{s}_m) + R(m,n)\}$$

- Adaptivity (provided $(S_m)_{m\in\mathcal{M}_n}$ is well chosen), e.g., to the smoothness of s or to the variations of σ

Model selection



$$(S_m)_{m\in\mathcal{M}} \longrightarrow (\widehat{s}_m)_{m\in\mathcal{M}} \longrightarrow \widehat{s}_{\widehat{m}} \quad ???$$

Goals:

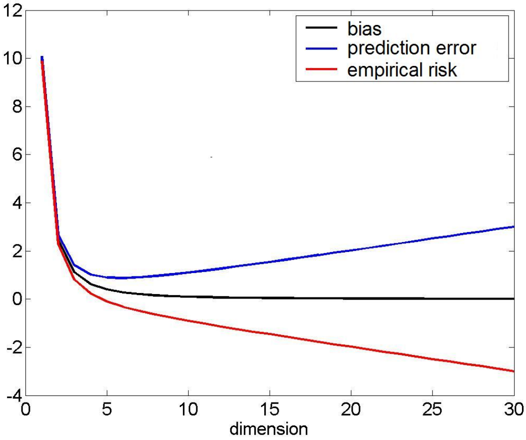
- Oracle inequality (in expectation, or with a large probability):

$$\ell(s, \widehat{s}_{\widehat{m}}) \leq C \inf_{m\in\mathcal{M}} \left\{ \ell(s, \widehat{s}_m) + R(m,n) \right\}$$

- Adaptivity (provided $(S_m)_{m\in\mathcal{M}_n}$ is well chosen), e.g., to the smoothness of s or to the variations of σ

Penalization

$$\widehat{m} \in \arg\min_{m \in \mathcal{M}} \{ P_n \gamma(\widehat{s}_m) + \mathrm{pen}(m) \}$$



Penalization

$$\widehat{m} \in \arg\min_{m \in \mathcal{M}} \{ P_n \gamma(\widehat{s}_m) + \mathrm{pen}(m) \}$$

$$\mathrm{pen}(m) = \frac{2\sigma^2 D_m}{n} \qquad \text{(Mallows 1973)}$$

$$\mathrm{pen}(m) = \frac{2\widehat{\sigma}^2 D_m}{n} \quad \text{or} \quad \widehat{K} D_m$$

And several other penalties (global or local Rademacher complexities, bootstrap or resampling penalties, *etc.*)

$$\Rightarrow \text{Ideal penalty:} \quad \mathrm{pen}_{\mathrm{id}}(m) = (P - P_n)(\gamma(\widehat{s}_m, \cdot))$$

Penalization

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

$$\mathrm{pen}(m) = \frac{2\sigma^2 D_m}{n} \qquad \text{(Mallows 1973)}$$

$$\mathrm{pen}(m) = \frac{2\widehat{\sigma}^2 D_m}{n} \quad \text{or} \quad \widehat{K} D_m$$

And several other penalties (global or local Rademacher complexities, bootstrap or resampling penalties, *etc.*)

$\Rightarrow$ Ideal penalty: $\mathrm{pen}_{\mathrm{id}}(m) = (P - P_n)(\gamma(\widehat{s}_m, \cdot))$

Data-driven calibration of the penalty

Assume that we know (or have estimated) pen_0 such that

$$K^{\star} \, \mathrm{pen}_0(m) \approx \mathbb{E}\left[\mathrm{pen}_{\mathrm{id}}(m)\right] \qquad (K^{\star} \text{ unknown})$$

Examples: $\mathrm{pen}_0(m) = D_m$, Rademacher complexity, *etc.*

$$\widehat{m}(K) \in \arg\min_{m \in \mathcal{M}_n} \left\{ P_n \gamma\left(\widehat{s}_m\right) + K \, \mathrm{pen}_0(m) \right\}$$

$\Rightarrow$ how to choose K?

Data-driven calibration of the penalty

Assume that we know (or have estimated) pen_0 such that

$$K^\star \, \mathrm{pen}_0(m) \approx \mathbb{E}\left[\mathrm{pen}_{\mathrm{id}}(m)\right] \qquad (K^\star \text{ unknown})$$

Examples: $\mathrm{pen}_0(m) = D_m$, Rademacher complexity, *etc.*

$$\widehat{m}(K) \in \arg\min_{m \in \mathcal{M}_n} \left\{ P_n \gamma\left(\widehat{s}_m\right) + K \, \mathrm{pen}_0(m) \right\}$$

$\Rightarrow$ how to choose K?

Data-driven calibration of the penalty

Assume that we know (or have estimated) pen_0 such that

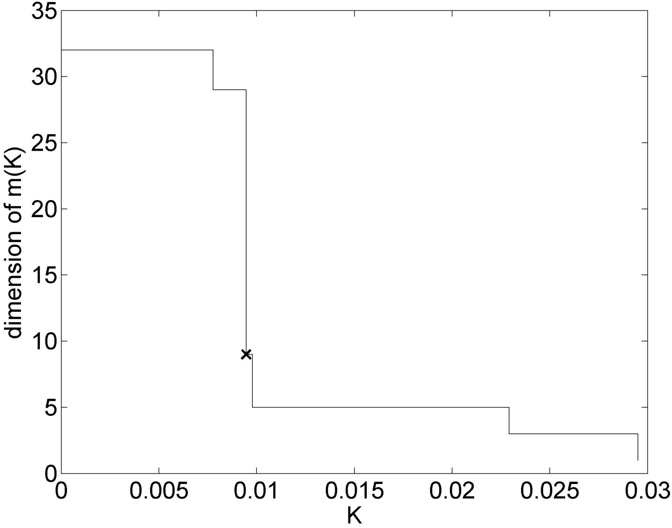
$$K^{\star} \, \mathrm{pen}_0(m) \approx \mathbb{E}\left[\mathrm{pen}_{\mathrm{id}}(m)\right] \qquad (K^{\star} \text{ unknown})$$

Examples: $\mathrm{pen}_0(m) = D_m$, Rademacher complexity, *etc.*

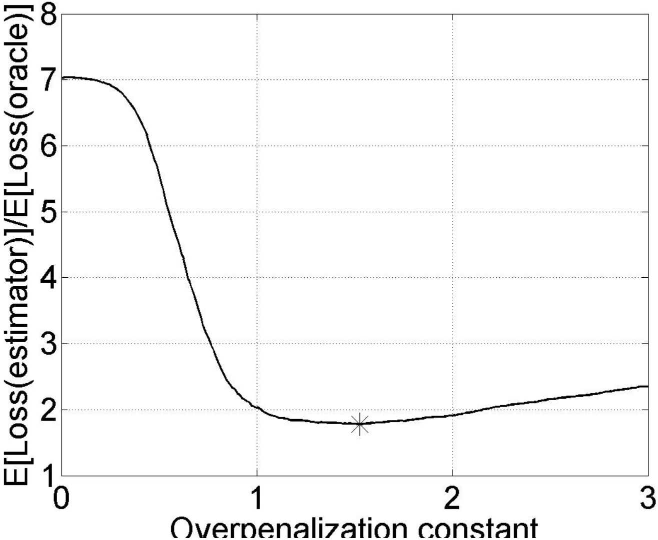
$$\widehat{m}(K) \in \arg\min_{m \in \mathcal{M}_n} \left\{ P_n \gamma\left(\widehat{s}_m\right) + K \, \mathrm{pen}_0(m) \right\}$$

$\Rightarrow$ how to choose K?

Dimension jump



Efficiency as a function of K



Algorithm (Birgé, Massart 2007; A., Massart, JMLR 2009)

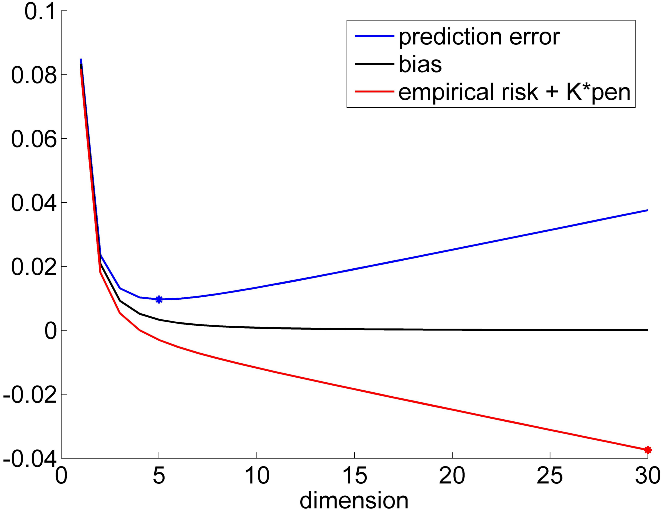
1. for every $K > 0$, compute

$$\widehat{m}(K) \in \arg\min_{m \in \mathcal{M}_n} \left\{ P_n \gamma\left(\widehat{s}_m\right) + K \operatorname{pen}_0(m) \right\}$$

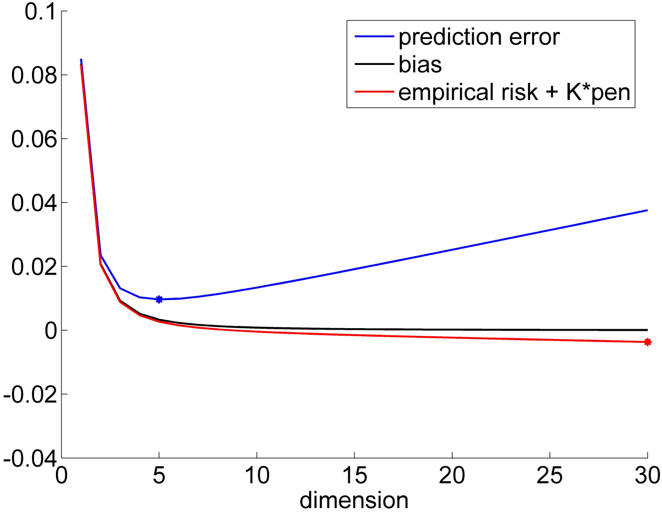
2. find $\widehat{K}_{\min}$ such that $D_{\widehat{m}(K)}$ is "very large" when $K < \widehat{K}_{\min}$ and "reasonably small" when $K > \widehat{K}_{\min}$

3. choose the model $\widehat{m} = \widehat{m}\left(2\widehat{K}_{\min}\right)$

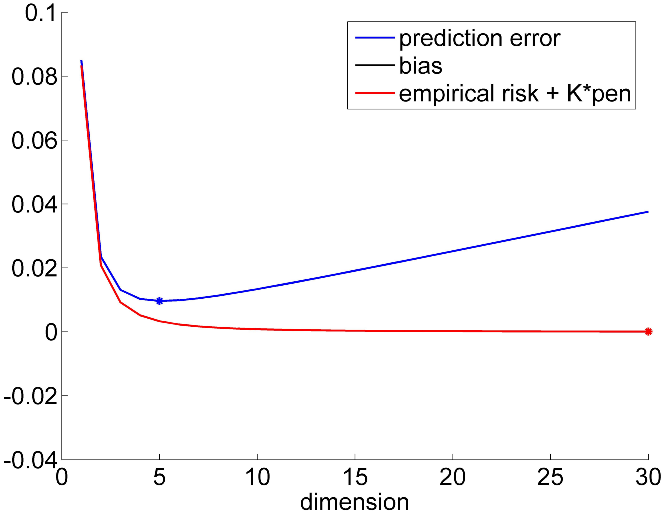
The slope heuristics $K = 0$



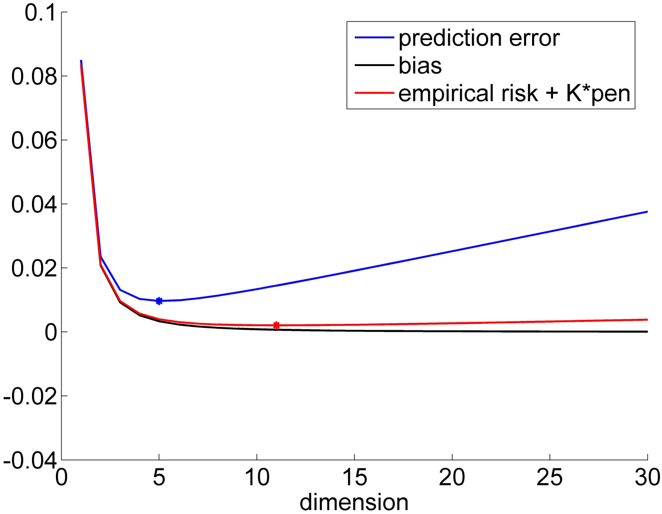
The slope heuristics $K = 0.45K^{\star}$



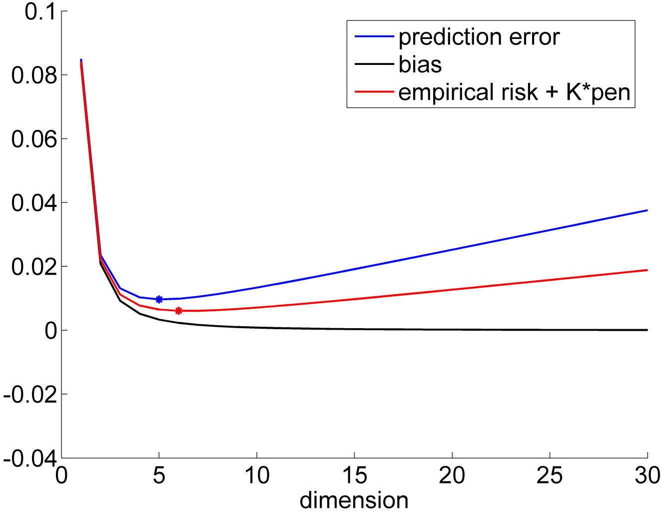
The slope heuristics $K = 0.5K^{\star}$



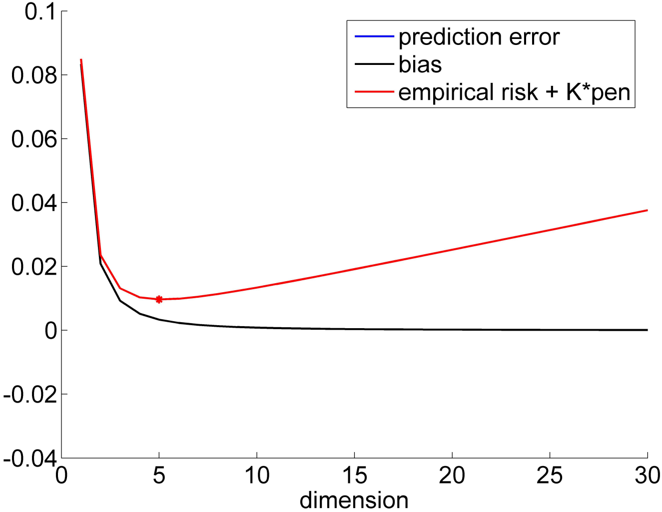
The slope heuristics $K = 0.55K^{\star}$



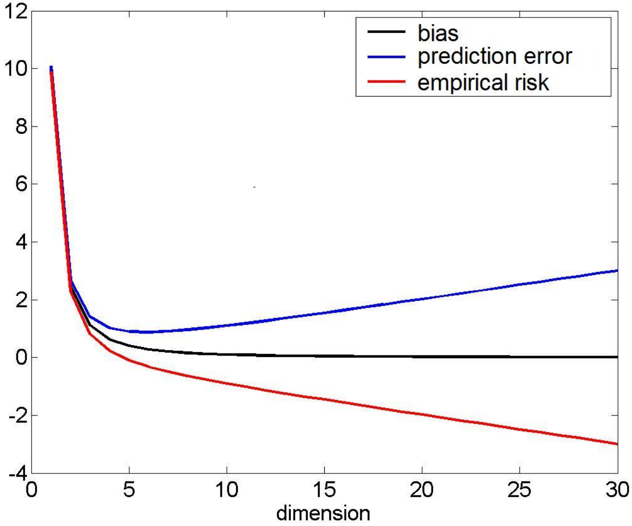
The slope heuristics $K = 0.75 K^{\star}$



The slope heuristics $K = K^{\star}$



The slope heuristics: informal argument



Two theorems

- Histograms; "small" number of models ($\operatorname{Card}(\mathcal{M}_n) \leq \Diamond n^{\Diamond}$)
- Bounded data: $\|Y\|_\infty \leq A < \infty$
- Noise-level lower bounded: $0 < \sigma_{\min} \leq \sigma(X)$
- Smooth s: non-constant, α-hölderian

Theorem (Minimal penalty; A. and Massart, JMLR 2009)

If $0 \leq K < K^\star / 2$, with probability at least $1 - \Diamond n^{-2}$,

$$\ell(s, \widehat{s}_{\widehat{m}(K)}) \geq \ln(n) \inf_{m \in \mathcal{M}} \{\ell(s, \widehat{s}_m)\} \quad \textit{and} \quad D_{\widehat{m}(K)} \geq \frac{\Diamond n}{\ln(n)}$$

Two theorems

Theorem (Optimal penalty; A. and Massart, JMLR 2009)

If $K > K^\star/2$, with probability at least $1 - \Diamond n^{-2}$,

$$\ell(s, \widehat{s}_{\widehat{m}(K)}) \leq C_n(K) \inf_{m \in \mathcal{M}} \{\ell(s, \widehat{s}_m)\} \quad \text{and} \quad D_{\widehat{m}(K)} \leq n^{1-\eta}$$

where $C_n(K) \leq C(K)$, $C_n(K^\star) \leq 1 + \ln(n)^{-1/5}$ and $\eta > 0$ may depend on the smoothness of s.

Theorem (Minimal penalty; A. and Massart, JMLR 2009)

If $0 \leq K < K^\star/2$, with probability at least $1 - \Diamond n^{-2}$,

$$\ell(s, \widehat{s}_{\widehat{m}(K)}) \geq \ln(n) \inf_{m \in \mathcal{M}} \{\ell(s, \widehat{s}_m)\} \quad \text{and} \quad D_{\widehat{m}(K)} \geq \frac{\Diamond n}{\ln(n)}$$

The slope heuristics: sketch of proof

prediction error $$P\gamma\left(\widehat{s}_m\right) = P\gamma\left(s_m\right) + P\left(\gamma\left(\widehat{s}_m\right) - \gamma\left(s_m\right)\right)$$

empirical risk $$P_n\gamma\left(\widehat{s}_m\right) = P_n\gamma\left(s_m\right) - \left(P_n\left(\gamma\left(s_m\right) - \gamma\left(\widehat{s}_m\right)\right)\right)$$

$$P_n\left(\gamma\left(s_m\right) - \gamma\left(\widehat{s}_m\right)\right) \approx P\left(\gamma\left(\widehat{s}_m\right) - \gamma\left(s_m\right)\right)$$

Ingredients of the proof:

- estimation of the expectations
- concentration inequalities

The slope heuristics: sketch of proof

prediction error $P\gamma(\widehat{s}_m) = P\gamma(s_m) + P(\gamma(\widehat{s}_m) - \gamma(s_m))$

empirical risk $P_n\gamma(\widehat{s}_m) = P_n\gamma(s_m) - (P_n(\gamma(s_m) - \gamma(\widehat{s}_m))$

$$P_n(\gamma(s_m) - \gamma(\widehat{s}_m)) \approx P(\gamma(\widehat{s}_m) - \gamma(s_m))$$

Ingredients of the proof:

- estimation of the expectations
- concentration inequalities

The slope heuristics: sketch of proof

prediction error $P\gamma(\widehat{s}_m) = P\gamma(s_m) + P(\gamma(\widehat{s}_m) - \gamma(s_m))$

empirical risk $P_n\gamma(\widehat{s}_m) = P_n\gamma(s_m) - (P_n(\gamma(s_m) - \gamma(\widehat{s}_m))$

$$P_n(\gamma(s_m) - \gamma(\widehat{s}_m)) \approx P(\gamma(\widehat{s}_m) - \gamma(s_m))$$

Ingredients of the proof:

- estimation of the expectations
- concentration inequalities

The slope heuristics: sketch of proof

prediction error $$P\gamma(\widehat{s}_m) = P\gamma(s_m) + P(\gamma(\widehat{s}_m) - \gamma(s_m))$$

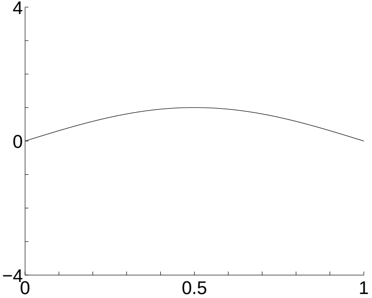
empirical risk $$P_n\gamma(\widehat{s}_m) = P_n\gamma(s_m) - (P_n(\gamma(s_m) - \gamma(\widehat{s}_m))$$

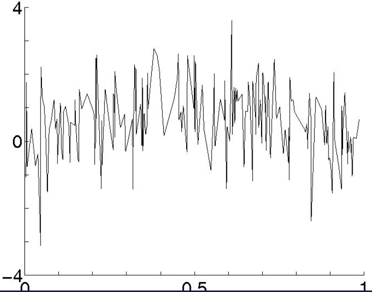
$$P_n(\gamma(s_m) - \gamma(\widehat{s}_m)) \approx P(\gamma(\widehat{s}_m) - \gamma(s_m))$$

Ingredients of the proof:

- estimation of the expectations
- concentration inequalities

Illustration: $s(x) = \sin(\pi x)$, $n = 200$, $\sigma \equiv 1$





$$\mathrm{pen}_0(m) = D_m$$

$$\frac{\mathbb{E}\left[\ell(s, \widehat{s}_{\widehat{m}})\right]}{\mathbb{E}\left[\inf_{m \in \mathcal{M}} \left\{\ell(s, \widehat{s}_m)\right\}\right]}$$

computed over 1000 samples.

Model selection method	Efficiency
Mallows (σ)	2.03 ± 0.04
Mallows ($\widehat{\sigma}$)	1.93 ± 0.04
Slope (threshold)	1.88 ± 0.03
Slope (maximal jump)	2.01 ± 0.04

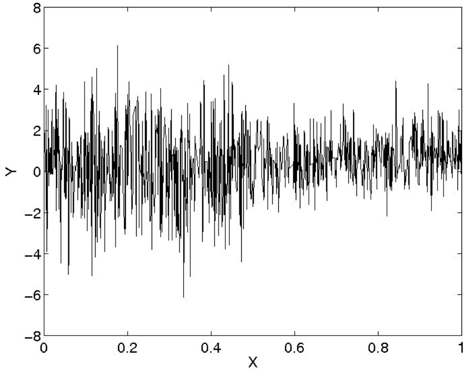
Related results

- Birgé and Massart (2007): similar theoretical results when the noise is Gaussian homoscedastic (either polynomial or exponential collections of models).
  Successfully applied to change-point detection (Lebarbier, 2005).
- The slope heuristics experimentally works in several other frameworks:
  - mixture models (Maugis and Michel, 2008),
  - clustering (Baudry, 2007),
  - spatial statistics (Verzelen, 2008),
  - estimation of oil reserves (Lepez, 2002),
  - genomics (Villers, 2007).

Limitations of linear penalties: illustration

$$Y = X + \left(1 + \mathbb{1}_{X \leq 1/2}\right) \epsilon \qquad n = 1000 \text{ data points}$$

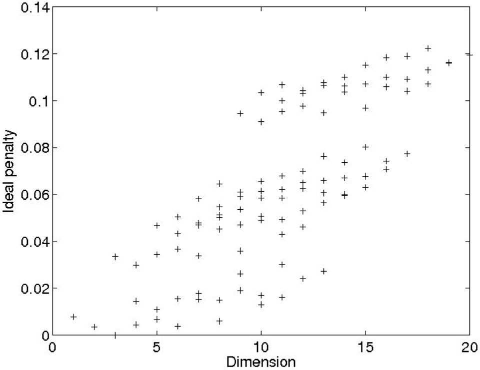
Regular histograms on $\left[0; \frac{1}{2}\right]$ ($D_{m,1}$ pieces), then regular histograms on $\left[\frac{1}{2}; 1\right]$ ($D_{m,2}$ pieces).



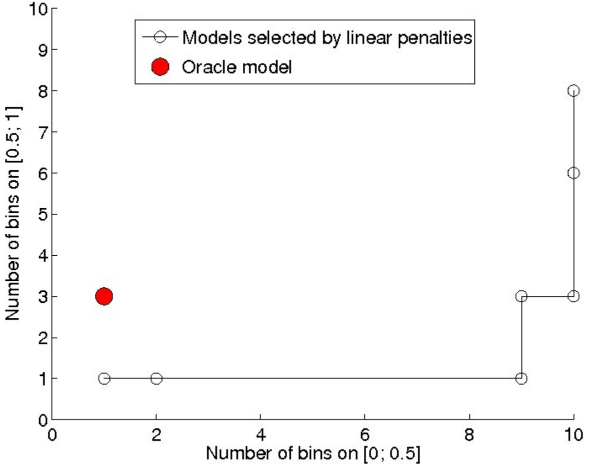
Limitations of linear penalties: illustration

$$Y = X + \left(1 + \mathbb{1}_{X \leq 1/2}\right) \epsilon \qquad n = 1000 \text{ data points}$$

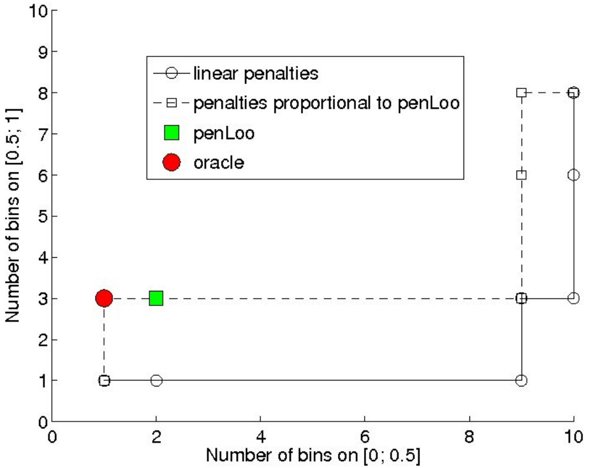
The ideal penalty is not a linear function of the dimension.



Limitations of linear penalties: illustration

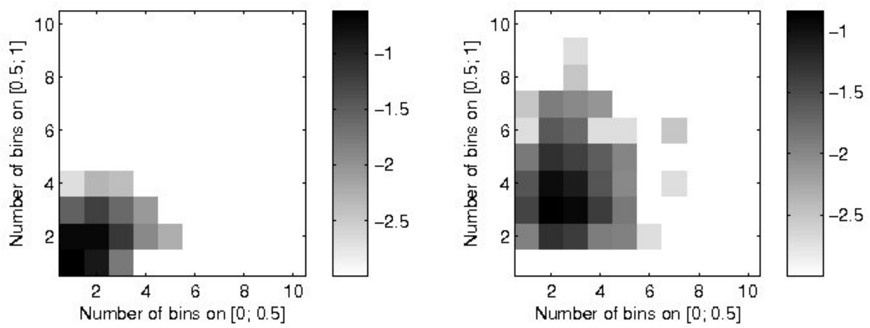


Limitations of linear penalties: illustration



Limitations of linear penalties: $\widehat{m}(K^\star) \neq m^\star$

Density of $(D_{\widehat{m}(K^\star),1}, D_{\widehat{m}(K^\star),2})$ and $(D_{m^\star,1}, D_{m^\star,2})$ according to $N = 1000$ samples



$\widehat{m}(K^\star)$ $m^\star$

Limitations of linear penalties: theory

$$Y = X + \sigma(X)\epsilon \qquad \text{with} \quad X \sim \mathcal{U}([0;1]) \enspace ,$$

$$\mathbb{E}\left[\epsilon | X\right] = 0 \quad \mathbb{E}\left[\epsilon^2 | X\right] = 1 \quad \text{and} \quad \int_0^{1/2} (\sigma(x))^2 \, dx \neq \int_{1/2}^{1} (\sigma(x))^2 \, dx$$

Regular histograms on $\left[0; \frac{1}{2}\right]$ $(1 \leq D_{m,1} \leq n/(2\ln(n)^2)$ pieces),
then regular histograms on $\left[\frac{1}{2}; 1\right]$ $(1 \leq D_{m,2} \leq n/(2\ln(n)^2)$ pieces).

Theorem (A. 2008, arXiv:0812.3141)

There exist absolute constants $C, \eta > 0$ and an event of probability at least $1 - Cn^{-2}$ on which

$$\forall K > 0, \forall \widehat{m}(K) \in \arg\min_{m \in \mathcal{M}_n} \left\{ P_n \gamma\left(\widehat{s}_m\right) + K D_m \right\} \enspace ,$$

$$\ell(s, \widehat{s}_{\widehat{m}(K)}) \geq (1 + \eta) \inf_{m \in \mathcal{M}_n} \left\{ \ell(s, \widehat{s}_m) \right\} \enspace .$$

Resampling heuristics (bootstrap, Efron 1979)

Real world : $P \xrightarrow{\text{sampling}} P_n \Longrightarrow \widehat{s}_m$

$$\mathrm{pen}_{\mathrm{id}}(m) = (P - P_n)\gamma(\widehat{s}_m) = F(P, P_n)$$

Resampling heuristics (bootstrap, Efron 1979)

Real world : $P \xrightarrow{\text{sampling}} P_n \Longrightarrow \widehat{s}_m$

Bootstrap world : P_n

$$\mathrm{pen}_{\mathrm{id}}(m) = (P - P_n)\gamma\left(\widehat{s}_m\right) = F(P, P_n)$$

Resampling heuristics (bootstrap, Efron 1979)

Real world : $P \xrightarrow{\text{sampling}} P_n \Longrightarrow \widehat{s}_m$

$\rightsquigarrow$

Bootstrap world : $P_n \xrightarrow{\text{resampling}} P_n^W \Longrightarrow \widehat{s}_m^W$

$$(P - P_n)\gamma\left(\widehat{s}_m\right) = F(P, P_n) \rightsquigarrow F(P_n, P_n^W) = (P_n - P_n^W)\gamma\left(\widehat{s}_m^W\right)$$

Resampling heuristics (bootstrap, Efron 1979)

$$\text{Real world :} \qquad P \xrightarrow{\text{sampling}} P_n \Longrightarrow \widehat{s}_m$$

$$\rightsquigarrow$$

$$\text{Bootstrap world :} \qquad P_n \xrightarrow{\text{subsampling}} P_n^W \Longrightarrow \widehat{s}_m^W$$

$$(P - P_n)\gamma(\widehat{s}_m) = F(P, P_n) \rightsquigarrow F(P_n, P_n^W) = (P_n - P_n^W)\gamma\left(\widehat{s}_m^W\right)$$

V-fold: $$P_n^W = \frac{1}{n - \mathrm{Card}(B_J)} \sum_{i \notin B_J} \delta_{(X_i, Y_i)} \quad \text{with } J \sim \mathcal{U}(1, \ldots, V)$$

V-fold penalization

- Ideal penalty:

$$(P - P_n)(\gamma(\widehat{s}_m))$$

- V-fold penalty:

$$\operatorname{pen}(m) = \frac{C}{V} \sum_{j=1}^{V} \left[(P_n - P_n^{(-j)})(\gamma(\widehat{s}_m^{(-j)})) \right]$$

$$\widehat{s}_m^{(-j)} \in \arg\min_{t \in S_m} P_n^{(-j)} \gamma(t)$$

with $C \geq V - 1$ to be chosen
$C = V - 1$ for estimating (almost) unbiasedly the ideal penalty

- The final estimator is $\widehat{s}_{\widehat{m}}$ with

$$\widehat{m} \in \arg\min_{m \in \mathcal{M}} \{ P_n \gamma(\widehat{s}_m) + \operatorname{pen}(m) \}$$

V-fold penalization

- Ideal penalty:

$$(P - P_n)(\gamma(\widehat{s}_m))$$

- V-fold penalty:

$$\mathrm{pen}(m) = \frac{C}{V} \sum_{j=1}^{V} \left[(P_n - P_n^{(-j)})(\gamma(\widehat{s}_m^{(-j)})) \right]$$

$$\widehat{s}_m^{(-j)} \in \arg\min_{t \in S_m} P_n^{(-j)} \gamma(t)$$

with $C \geq V - 1$ to be chosen
$C = V - 1$ for estimating (almost) unbiasedly the ideal penalty

- The final estimator is $\widehat{s}_{\widehat{m}}$ with

$$\widehat{m} \in \arg\min_{m \in \mathcal{M}} \{ P_n \gamma(\widehat{s}_m) + \mathrm{pen}(m) \}$$

V-fold penalization

- Ideal penalty:

$$(P - P_n)(\gamma(\widehat{s}_m))$$

- V-fold penalty:

$$\mathrm{pen}(m) = \frac{C}{V} \sum_{j=1}^{V} \left[(P_n - P_n^{(-j)})(\gamma(\widehat{s}_m^{(-j)})) \right]$$

$$\widehat{s}_m^{(-j)} \in \arg\min_{t \in S_m} P_n^{(-j)} \gamma(t)$$

with $C \geq V - 1$ to be chosen
$C = V - 1$ for estimating (almost) unbiasedly the ideal penalty

- The final estimator is $\widehat{s}_{\widehat{m}}$ with

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

Non-asymptotic pathwise oracle inequality

- Fixed V or $V = n$
- $C \approx V - 1$
- Histograms; "small" number of models ($\mathrm{Card}(\mathcal{M}_n) \leq \Diamond n^{\Diamond}$)
- Bounded data: $\|Y\|_{\infty} \leq A < \infty$
- Noise-level lower bounded: $0 < \sigma_{\min} \leq \sigma(X)$
- Smooth s: non-constant, α-hölderian

Theorem (A. 2008, arXiv:0802.0566)

Under a "reasonable" set of assumptions on P, with probability at least $1 - \Diamond n^{-2}$,

$$\ell(s, \widehat{s}_{\widehat{m}}) \leq \left(1 + \ln(n)^{-1/5}\right) \inf_{m \in \mathcal{M}} \{\ell(s, \widehat{s}_m)\}$$

Non-asymptotic pathwise oracle inequality

- Fixed V or $V = n$
- $C \approx V - 1$
- Histograms; "small" number of models ($\mathrm{Card}(\mathcal{M}_n) \leq \Diamond n^{\Diamond}$)
- Bounded data: $\|Y\|_\infty \leq A < \infty$
- Noise-level lower bounded: $0 < \sigma_{\min} \leq \sigma(X)$
- Smooth s: non-constant, α-hölderian

Theorem (A. 2008, arXiv:0802.0566)

Under a "reasonable" set of assumptions on P, with probability at least $1 - \Diamond n^{-2}$,

$$\ell(s, \widehat{s}_{\widehat{m}}) \leq \left(1 + \ln(n)^{-1/5}\right) \inf_{m \in \mathcal{M}} \{\ell(s, \widehat{s}_m)\}$$

Non-asymptotic pathwise oracle inequality

- Fixed V or $V = n$
- $C \approx V - 1$
- Histograms; "small" number of models ($\mathrm{Card}(\mathcal{M}_n) \leq \Diamond n^{\Diamond}$)
- Bounded data: $\|Y\|_{\infty} \leq A < \infty$
- Noise-level lower bounded: $0 < \sigma_{\min} \leq \sigma(X)$
- Smooth s: non-constant, α-hölderian

Theorem (A. 2008, arXiv:0802.0566)

Under a "reasonable" set of assumptions on P, with probability at least $1 - \Diamond n^{-2}$,

$$\ell(s, \widehat{s}_{\widehat{m}}) \leq \left(1 + \ln(n)^{-1/5}\right) \inf_{m \in \mathcal{M}} \left\{ \ell(s, \widehat{s}_m) \right\}$$

Non-asymptotic pathwise oracle inequality

- Fixed V or $V = n$
- $C \approx V - 1$
- Histograms; "small" number of models ($\mathrm{Card}(\mathcal{M}_n) \leq \diamondsuit n^{\diamondsuit}$)
- Bounded data: $\|Y\|_\infty \leq A < \infty$
- Noise-level lower bounded: $0 < \sigma_{\min} \leq \sigma(X)$
- Smooth s: non-constant, α-hölderian

Theorem (A. 2008, arXiv:0802.0566)

Under a "reasonable" set of assumptions on P, with probability at least $1 - \diamondsuit n^{-2}$,

$$\ell(s, \widehat{s}_{\widehat{m}}) \leq \left(1 + \ln(n)^{-1/5}\right) \inf_{m \in \mathcal{M}} \{\ell(s, \widehat{s}_m)\}$$

Simulation framework

$$Y_i = s(X_i) + \sigma(X_i)\epsilon_i \qquad X_i \sim^{\text{i.i.d.}} \mathcal{U}([0;1]) \qquad \epsilon_i \sim^{\text{i.i.d.}} \mathcal{N}(0,1)$$

$\mathcal{M}_n$: histograms regular on $[0, 1/2]$ (D_1 pieces), and on $[1/2, 1]$ (D_2 pieces), with $1 \leq D_1, \, D_2 \leq \frac{n}{2\log(n)}$.

$\Rightarrow$ Benchmark:

$$C_{\text{classical}} = \frac{\mathbb{E}[\ell(s, \widehat{s}_{\widehat{m}})]}{\mathbb{E}[\inf_{m \in \mathcal{M}} \ell(s, \widehat{s}_m)]}$$ computed with $N = 1000$ samples

Simulation framework

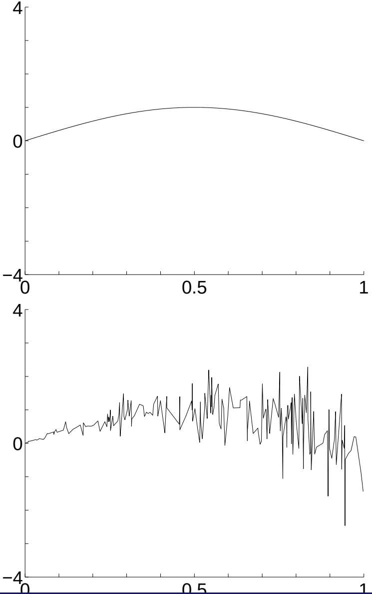
$$Y_i = s(X_i) + \sigma(X_i)\epsilon_i \qquad X_i \sim^{\text{i.i.d.}} \mathcal{U}([0;1]) \qquad \epsilon_i \sim^{\text{i.i.d.}} \mathcal{N}(0,1)$$

$\mathcal{M}_n$: histograms regular on $[0,1/2]$ (D_1 pieces), and on $[1/2,1]$ (D_2 pieces), with $1 \leq D_1,\ D_2 \leq \frac{n}{2\log(n)}$.

$\Rightarrow$ Benchmark:

$$C_{\text{classical}} = \frac{\mathbb{E}[\ell(s,\widehat{s}_{\widehat{m}})]}{\mathbb{E}[\inf_{m\in\mathcal{M}} \ell(s,\widehat{s}_m)]}$$ computed with $N = 1000$ samples

Simulations: sin, $n = 200$, $\sigma(x) = x$, 2 bin sizes



Mallows	3.69 ± 0.07
2-fold	2.54 ± 0.05
5-fold	2.58 ± 0.06
10-fold	2.60 ± 0.06
20-fold	2.58 ± 0.06
leave-one-out	2.59 ± 0.06
pen 2-f	3.06 ± 0.07
pen 5-f	2.75 ± 0.06
pen 10-f	2.65 ± 0.06
pen Loo	2.59 ± 0.06
Mallows $\times 1.25$	3.17 ± 0.07
pen 2-f $\times 1.25$	2.75 ± 0.06
pen 5-f $\times 1.25$	2.38 ± 0.06
pen 10-f $\times 1.25$	2.28 ± 0.05
pen Loo $\times 1.25$	2.21 ± 0.05

Other resampling-based penalties

- Efron's bootstrap penalties (Efron 1983, Shibata 1997):

$$\mathrm{pen}(m) = \mathbb{E}\left[(P_n - P_n^W)(\gamma(\widehat{s}_m^W)) \middle| (X_i, Y_i)_{1 \leq i \leq n} \right]$$

- General resampling penalties (A. 2008, hal-00262478)
- Rademacher complexities (Koltchinskii 2001 ; Bartlett, Boucheron, Lugosi 2002): subsampling

$$\mathrm{pen}_{\mathrm{id}}(m) \leq \mathrm{pen}_{\mathrm{id}}^{\mathrm{glo}}(m) = \sup_{t \in S_m} (P - P_n)\gamma(t, \cdot)$$

- idem with general exchangeable weights (Fromont 2004)
- Local Rademacher complexities (Bartlett, Bousquet, Mendelson 2004 ; Koltchinskii 2004)

Cross-validation procedures

- Hold-out, Cross-validation, Leave-one-out, V-fold cross-validation:
  $I \subset \{1, \ldots, n\}$ random sub-sample of size q (VFCV: $q = \frac{n(V-1)}{V}$).
- V-fold cross-validation is biased
  $\Rightarrow$ suboptimal model selection when V is fixed as $n \longrightarrow \infty$ (A. 2008, arXiv:0802.0566)
- V-fold penalization with $C = V - 1$
  $\Leftrightarrow$ Burman's corrected V-fold cross-validation (1989).

Cross-validation procedures

- Hold-out, Cross-validation, Leave-one-out, V-fold cross-validation:
  $I \subset \{1, \ldots, n\}$ random sub-sample of size q (VFCV: $q = \frac{n(V-1)}{V}$).
- V-fold cross-validation is biased
  $\Rightarrow$ suboptimal model selection when V is fixed as $n \to \infty$ (A. 2008, arXiv:0802.0566)
- V-fold penalization with $C = V - 1$
  $\Leftrightarrow$ Burman's corrected V-fold cross-validation (1989).

Cross-validation procedures

- Hold-out, Cross-validation, Leave-one-out, V-fold cross-validation:
  $I \subset \{1, \ldots, n\}$ random sub-sample of size q (VFCV: $q = \frac{n(V-1)}{V}$).
- V-fold cross-validation is biased
  $\Rightarrow$ suboptimal model selection when V is fixed as $n \to \infty$ (A. 2008, arXiv:0802.0566)
- V-fold penalization with $C = V - 1$
  $\Leftrightarrow$ Burman's corrected V-fold cross-validation (1989).

Conclusion

- Shape of the penalty: estimated by resampling (V-fold, bootstrap, exchangeable bootstrap...)
  $\Rightarrow$ adaptation to unknown variations of the noise-level
- Multiplying constant: estimated thanks to the slope heuristics (model-selection based estimator)
  $\Rightarrow$ oracle inequalities with constant $1 + \epsilon_n$,
  even when $\mathrm{pen}_0(m)$ is a V-fold or resampling penalty, inside the slope heuristics algorithm
- Cross-validation and resampling penalties can also be used for change-point detection, i.e., for detecting changes in the mean of an heteroscedastic sequence (joint work with A. Celisse, arXiv:0902.3977)

Conclusion

- Shape of the penalty: estimated by resampling (V-fold, bootstrap, exchangeable bootstrap...)
  $\Rightarrow$ adaptation to unknown variations of the noise-level
- Multiplying constant: estimated thanks to the slope heuristics (model-selection based estimator)
  $\Rightarrow$ oracle inequalities with constant $1 + \epsilon_n$,
  even when $\text{pen}_0(m)$ is a V-fold or resampling penalty, inside the slope heuristics algorithm
- Cross-validation and resampling penalties can also be used for change-point detection, i.e., for detecting changes in the mean of an heteroscedastic sequence (joint work with A. Celisse, arXiv:0902.3977)

Conclusion

- Shape of the penalty: estimated by resampling (V-fold, bootstrap, exchangeable bootstrap...)
  $\Rightarrow$ adaptation to unknown variations of the noise-level
- Multiplying constant: estimated thanks to the slope heuristics (model-selection based estimator)
  $\Rightarrow$ oracle inequalities with constant $1 + \epsilon_n$,
  even when $\mathrm{pen}_0(m)$ is a V-fold or resampling penalty, inside the slope heuristics algorithm
- Cross-validation and resampling penalties can also be used for change-point detection, i.e., for detecting changes in the mean of an heteroscedastic sequence (joint work with A. Celisse, arXiv:0902.3977)

Change-point detection via cross-validation

$$\forall 1 \leq i \leq n, \qquad Y_i = s(t_i) + \sigma(t_i)\epsilon_i \qquad \text{with} \quad \mathbb{E}[\epsilon_i] = 0 \quad \mathbb{E}\left[\epsilon_i^2\right] = 1$$

- Goal: detect changes in the mean s of the signal Y
  $\Rightarrow$ model selection
- No assumption on the variance $\sigma(t_i)^2$
- Birgé and Massart's penalty (assumes $\sigma(t_i) \equiv \sigma$):

$$\operatorname{pen}(m) = \frac{C D_m}{n} \left(5 + 2 \log \left(\frac{n}{D_m} \right) \right)$$

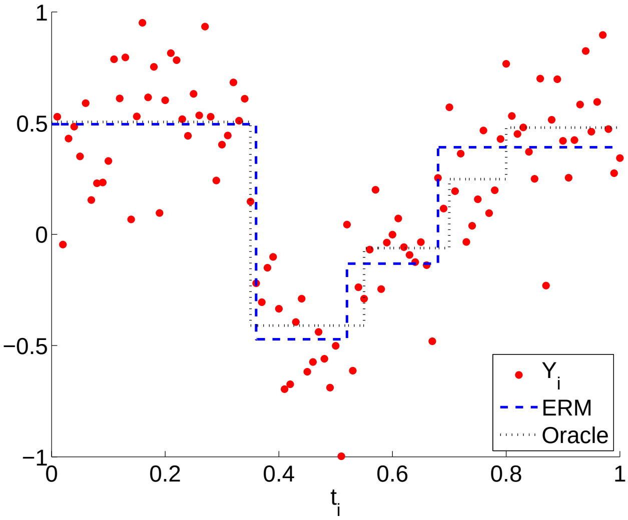
Change-point detection via cross-validation

$$\forall 1 \le i \le n, \qquad Y_i = s(t_i) + \sigma(t_i)\epsilon_i \qquad \text{with} \quad \mathbb{E}\left[\epsilon_i\right] = 0 \quad \mathbb{E}\left[\epsilon_i^2\right] = 1$$

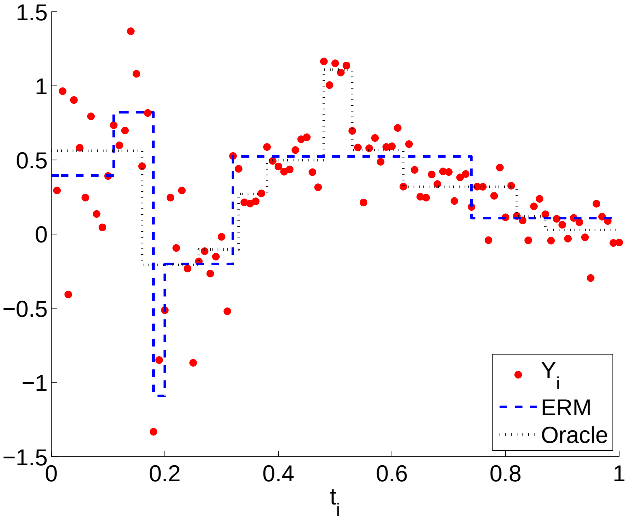
- Goal: detect changes in the mean s of the signal Y
  $\Rightarrow$ model selection
- No assumption on the variance $\sigma(t_i)^2$
- Birgé and Massart's penalty (assumes $\sigma(t_i) \equiv \sigma$):

$$\mathrm{pen}(m) = \frac{C D_m}{n}\left(5 + 2\log\left(\frac{n}{D_m}\right)\right)$$

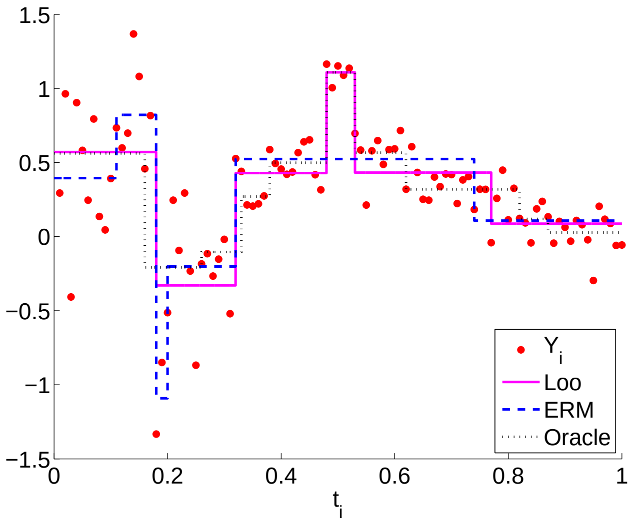
Fixed D, Homoscedastic data; $n = 100$, $\sigma = 0.25$, $D = 4$



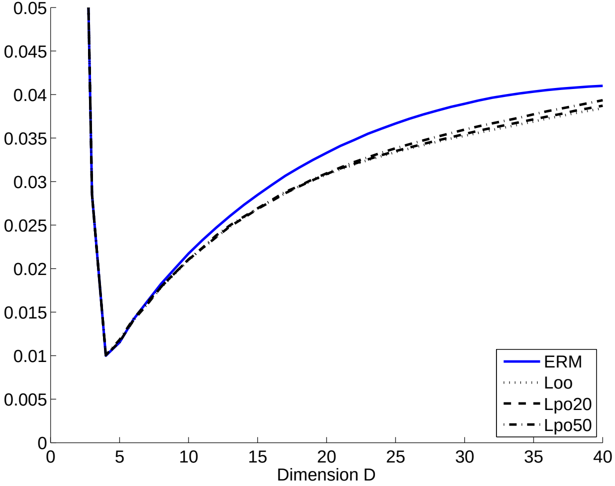
Fixed D, Heteroscedastic; $n = 100$, $\|\sigma\| = 0.30$, $D = 6$



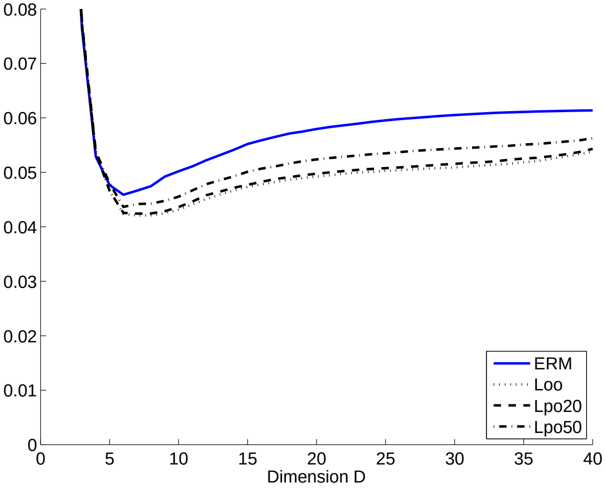
Fixed D, Heteroscedastic; $n = 100$, $\|\sigma\| = 0.30$, $D = 6$



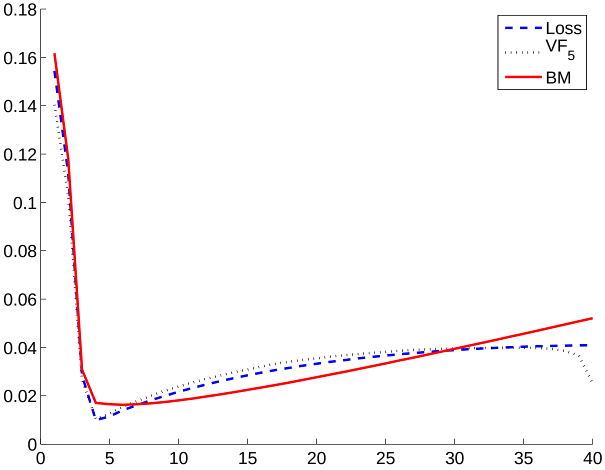
Homoscedastic data: loss as a function of D



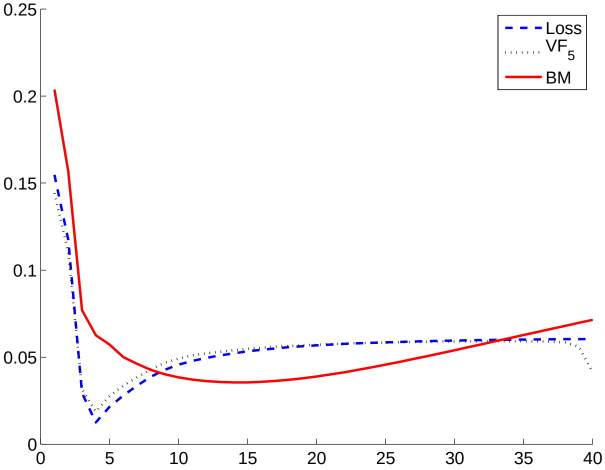
Heteroscedastic data: loss as a function of D



Homoscedastic data: estimation of the loss for every D



Heteroscedastic data: estimation of the loss for every D



A family of two-steps change-point detection algorithms

1. $\forall D \in \{1, \ldots, D_{\max}\}$, select a model $\widehat{m}(D)$ of dimension D:

$$\widehat{m}(D) \in \arg \min_{m \in \mathcal{M}_n,\, D_m = D} \{\mathrm{crit}_1(m; (t_i, Y_i)_i)\}$$

   Examples of crit_1: empirical risk, leave-p-out or V-fold estimators of the risk

2. Select $\widehat{D}$

$$\widehat{D} \in \arg \min_{D \in \{1, \ldots, D_{\max}\}} \{\mathrm{crit}_2(D; (t_i, Y_i)_i; \mathrm{crit}_1(\cdot))\}$$

   Examples of crit_2: penalized empirical criterion, V-fold cross-validation estimator of the risk

Simulation results

Deterministic (s, σ):

σ	$[Emp, VF_5]$	$[Loo, VF_5]$	$[Lpo_{20}, VF_5]$	$[Emp, BM]$
cst	4.41 ± 0.02	4.54 ± 0.02	4.62 ± 0.02	**4.39** ± 0.01
p-c	6.32 ± 0.02	**5.74** ± 0.02	5.81 ± 0.02	8.47 ± 0.03
sine	5.97 ± 0.02	**5.72** ± 0.02	5.86 ± 0.02	7.59 ± 0.03

Random (s, σ):

σ	$[Emp, VF_5]$	$[Loo, VF_5]$	$[Lpo_{20}, VF_5]$	$[Emp, BM]$
A	4.78 ± 0.03	**4.65** ± 0.03	4.78 ± 0.03	6.82 ± 0.03
B	5.09 ± 0.03	**4.88** ± 0.03	**4.91** ± 0.03	7.21 ± 0.04
C	7.17 ± 0.05	6.61 ± 0.05	**6.49** ± 0.05	13.49 ± 0.07

Bias of cross-validation

Ideal criterion: $P\gamma(\widehat{s}_m)$

Regression on a model of histograms with D_m pieces ($\sigma(X) \equiv \sigma$ for simplicity):

$$\mathbb{E}\left[P\gamma(\widehat{s}_m)\right] \approx P\gamma(s_m) + \frac{D_m \sigma^2}{n}$$

$$\mathbb{E}\left[P_n^{(j)}\gamma\left(\widehat{s}_m^{(-j)}\right)\right] = \mathbb{E}\left[P\gamma\left(\widehat{s}_m^{(-j)}\right)\right] \approx P\gamma(s_m) + \frac{V}{V-1}\frac{D_m \sigma^2}{n}$$

$\Rightarrow$ bias if V is fixed ("overpenalization")

Bias of cross-validation

Ideal criterion: $P\gamma(\widehat{s}_m)$

Regression on a model of histograms with D_m pieces ($\sigma(X) \equiv \sigma$ for simplicity):

$$\mathbb{E}\left[P\gamma(\widehat{s}_m)\right] \approx P\gamma(s_m) + \frac{D_m \sigma^2}{n}$$

$$\mathbb{E}\left[P_n^{(j)}\gamma\left(\widehat{s}_m^{(-j)}\right)\right] = \mathbb{E}\left[P\gamma\left(\widehat{s}_m^{(-j)}\right)\right] \approx P\gamma(s_m) + \frac{V}{V-1}\frac{D_m \sigma^2}{n}$$

$\Rightarrow$ bias if V is fixed ("overpenalization")

Bias of cross-validation

Ideal criterion: $P\gamma(\widehat{s}_m)$

Regression on a model of histograms with D_m pieces ($\sigma(X) \equiv \sigma$ for simplicity):

$$\mathbb{E}\left[P\gamma(\widehat{s}_m)\right] \approx P\gamma(s_m) + \frac{D_m \sigma^2}{n}$$

$$\mathbb{E}\left[P_n^{(j)}\gamma\left(\widehat{s}_m^{(-j)}\right)\right] = \mathbb{E}\left[P\gamma\left(\widehat{s}_m^{(-j)}\right)\right] \approx P\gamma(s_m) + \frac{V}{V-1}\frac{D_m \sigma^2}{n}$$

$\Rightarrow$ bias if V is fixed ("overpenalization")

Suboptimality of V-fold cross-validation

- $Y = X + \sigma\epsilon$ with ϵ bounded and $\sigma > 0$
- $\mathcal{M}$: family of regular histograms on $\mathcal{X} = [0,1]$
- V fixed

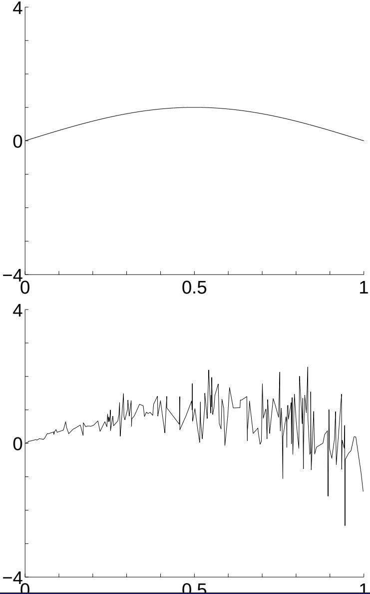
Theorem (A. 2008)

With probability at least $1 - \lozenge n^{-2}$,

$$\ell(s, \widehat{s}_{\widehat{m}}) \geq (1 + \kappa(V)) \inf_{m \in \mathcal{M}} \{\ell(s, \widehat{s}_m)\}$$

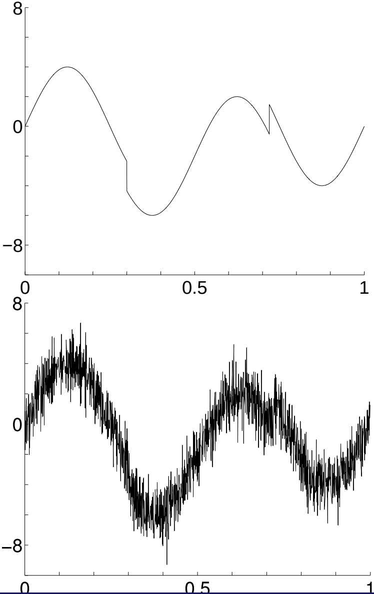
with $\kappa(V) > 0$.

Simulations: sin, $n = 200$, $\sigma(x) = x$, 2 bin sizes



Mallows	3.69 ± 0.07
2-fold	2.54 ± 0.05
5-fold	2.58 ± 0.06
10-fold	2.60 ± 0.06
20-fold	2.58 ± 0.06
leave-one-out	2.59 ± 0.06

Simulations: HeaviSine, $n = 2048$, $\sigma \equiv 1$



Models: dyadic regular histograms

2-fold	1.002 ± 0.003
5-fold	1.014 ± 0.003
10-fold	1.021 ± 0.003
20-fold	1.029 ± 0.004
leave-one-out	1.034 ± 0.004

Choice of V

- optimal performance when $V = V^\star$: trade-off variability–bias (difficult to find $V^\star$ from the data)

- SNR large:
  $\Rightarrow V^\star \to \infty$ when $n \to \infty$ (suboptimality result if V fixed)
  $\Rightarrow V^\star$ too large for computations

- SNR small:
  $\Rightarrow V^\star = 2$ is possible
  $\Rightarrow$ unsatisfactory (highly variable)

- V should be chosen according to computation time also

Choice of V

- optimal performance when $V = V^\star$: trade-off variability–bias (difficult to find $V^\star$ from the data)
- SNR large:
  $\Rightarrow$ $V^\star \to \infty$ when $n \to \infty$ (suboptimality result if V fixed)
  $\Rightarrow$ $V^\star$ too large for computations
- SNR small:
  $\Rightarrow$ $V^\star = 2$ is possible
  $\Rightarrow$ unsatisfactory (highly variable)
- V should be chosen according to computation time also

Choice of V

- optimal performance when $V = V^\star$: trade-off variability–bias (difficult to find $V^\star$ from the data)
- SNR large:
  $\Rightarrow V^\star \to \infty$ when $n \to \infty$ (suboptimality result if V fixed)
  $\Rightarrow V^\star$ too large for computations
- SNR small:
  $\Rightarrow V^\star = 2$ is possible
  $\Rightarrow$ unsatisfactory (highly variable)
- V should be chosen according to computation time also

Choice of V

- optimal performance when $V = V^\star$: trade-off variability–bias (difficult to find $V^\star$ from the data)
- SNR large:
  $\Rightarrow V^\star \to \infty$ when $n \to \infty$ (suboptimality result if V fixed)
  $\Rightarrow V^\star$ too large for computations
- SNR small:
  $\Rightarrow V^\star = 2$ is possible
  $\Rightarrow$ unsatisfactory (highly variable)
- V should be chosen according to computation time also