Optimal model selection

Sylvain Arlot

[1] Cnrs

[2] École Normale Supérieure (Paris), Liens, Willow Team

EMS 2009, Toulouse, 23/07/2009

Statistical framework: regression on a random design

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

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

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

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

Statistical framework: regression on a random design

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

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

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

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

Statistical framework: regression on a random design

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

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

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

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

Statistical framework: regression on a random design

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

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

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

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

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:

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

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

- 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:

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

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

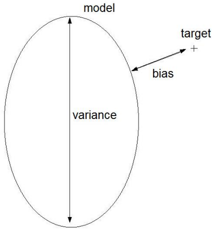
- Empirical risk minimizer on S_m (= model):

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

- *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



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

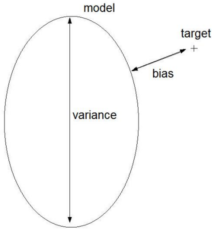
Goals:

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

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

- 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



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

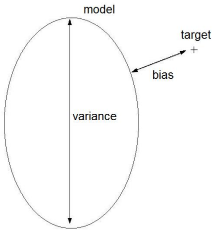
Goals:

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

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

- 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:

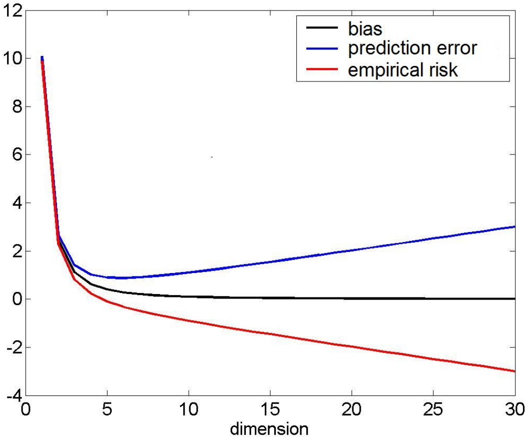
- Oracle inequality (in expectation, or with a large probability):

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

- 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

Penalization

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



Penalization

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

Unbiased risk estimation principle

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

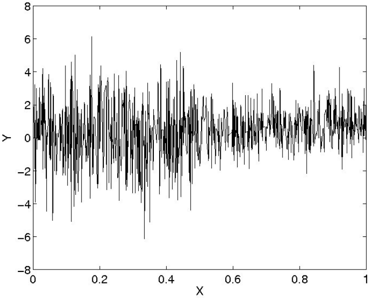
$$\mathrm{pen}(m) = \frac{2\sigma^2 D_m}{n} \qquad \text{(Mallows 1973)}$$

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

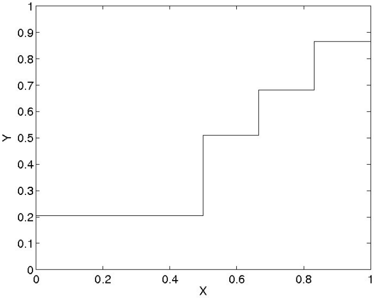
Limitations of linear penalties: illustration

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

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



data sample

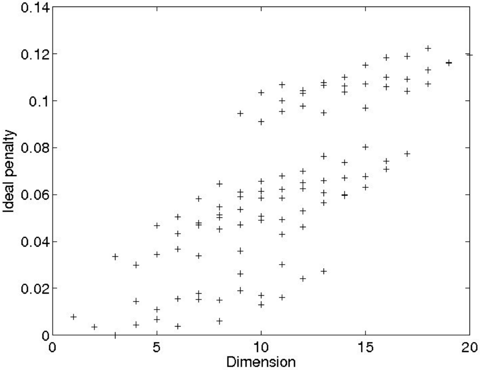


oracle: $D_{m,1} = 1$, $D_{m,2} = 3$

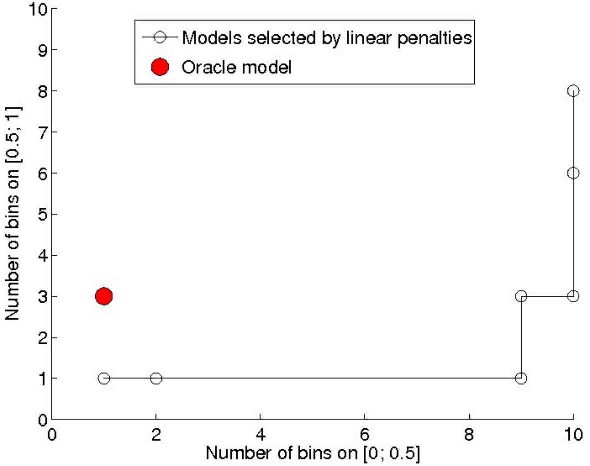
Limitations of linear penalties: illustration

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

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

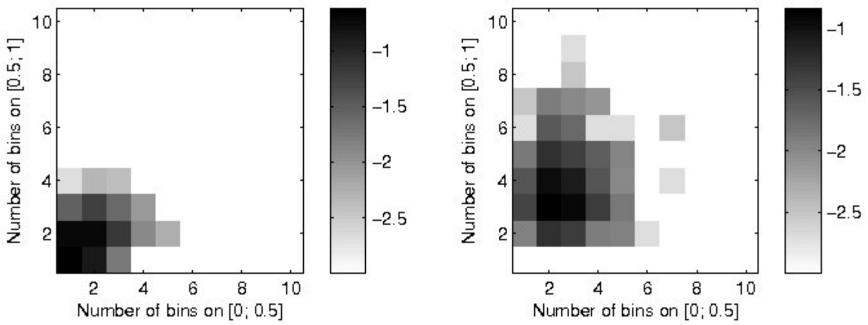


Limitations of linear penalties: illustration



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

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

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

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

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

Theorem (A. 2008, arXiv:0812.3141)

There exist constants $C, \eta > 0$ (only depending on $\sigma(\cdot)$) and an event of probability at least $1 - Cn^{-2}$ on which

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

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

Cross-validation

$$\underbrace{(X_1, Y_1), \ldots, (X_q, Y_q)}_{\text{Training}}, \underbrace{(X_{q+1}, Y_{q+1}), \ldots, (X_n, Y_n)}_{\text{Validation}}$$

$$\widehat{s}_m^{(e)} \in \arg\min_{t \in S_m} \left\{ \sum_{i=1}^{q} \gamma(t, (X_i, Y_i)) \right\}$$

$$P_n^{(v)} = \frac{1}{n-q} \sum_{i=q+1}^{n} \delta_{(X_i, Y_i)} \qquad \Rightarrow P_n^{(v)} \gamma \left(\widehat{s}_m^{(e)} \right)$$

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

$$\Rightarrow \widehat{m} \in \arg\min_{m \in \mathcal{M}} \left\{ \frac{1}{V} \sum_{j=1}^{V} P_n^j \gamma \left(\widehat{s}_m^{(-j)} \right) \right\} \qquad \widetilde{s} = \widehat{s}_{\widehat{m}}$$

Bias of cross-validation

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

Regression on a model of histograms with D_m bins ($\sigma(X) \equiv \sigma$ 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 bins ($\sigma(X) \equiv \sigma$ 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 bins ($\sigma(X) \equiv \sigma$ 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

- $Y = X + \sigma\varepsilon$ with ε bounded and $\sigma > 0$
- $\mathcal{M}$: family of regular histograms on $\mathcal{X} = [0,1]$
- $\widehat{m}$ selected by V-fold cross-validation with V fixed as n grows

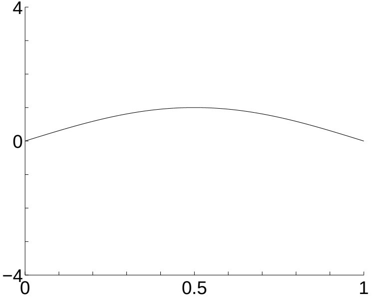
Theorem (A. 2008, arXiv:0802.0566)

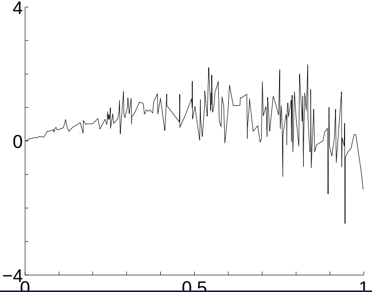
With probability at least $1 - \Diamond n^{-2}$,

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

with $\kappa(V) > 0$.

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





Models: regular histograms on $\left[0; \frac{1}{2}\right]$, then regular histograms on $\left[\frac{1}{2}; 1\right]$.

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

computed over 1000 samples.

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

Resampling heuristics (bootstrap, Efron 1979)

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

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

Resampling heuristics (bootstrap, Efron 1979)

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

$\rightsquigarrow$

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

$\rightsquigarrow$

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

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

where $\quad P_n^W = n^{-1} \sum_{i=1}^{n} W_i \delta_{(X_i, Y_i)} \quad .$

Resampling penalization

- Ideal penalty:

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

- Resampling penalty:

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

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

with $C \geq C_W$ to be chosen (no bias if $C = C_W$)

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

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

Resampling penalization

- Ideal penalty:

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

- Resampling penalty:

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

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

with $C \geq C_W$ to be chosen (no bias if $C = C_W$)

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

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

Resampling penalization

- Ideal penalty:

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

- Resampling penalty:

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

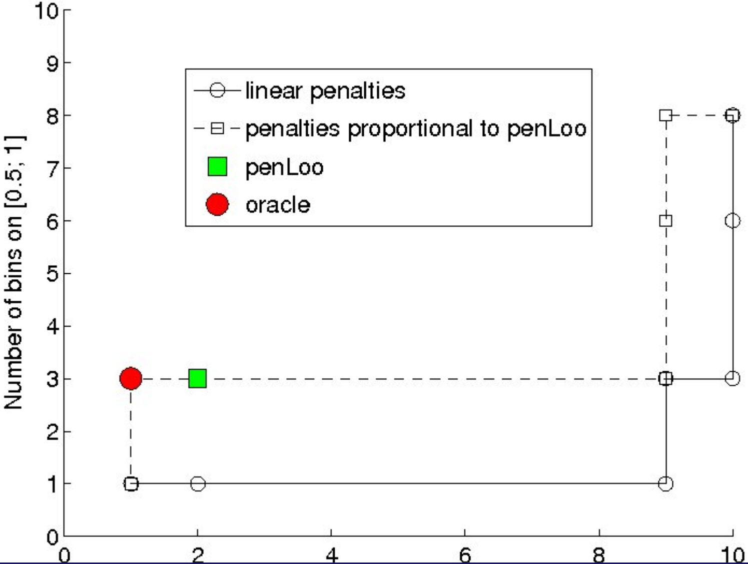
$$\widehat{s}_m^W \in \arg\min_{t \in S_m} P_n^W \gamma(t)$$

with $C \geq C_W$ to be chosen (no bias if $C = C_W$)

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

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

Resampling penalization with heteroscedastic data



Other resampling-based penalties

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

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

- Rademacher complexities (Koltchinskii 2001; Bartlett, Boucheron and Lugosi, 2002): subsampling

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

- idem with general exchangeable weights (Fromont, 2004)
- Local Rademacher complexities (Bartlett, Bousquet and Mendelson, 2004; Koltchinskii, 2006)
- ...

Non-asymptotic pathwise oracle inequality

- W exchangeable (e.g., bootstrap or subsampling)
- $C \approx C_W$
- 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. 2009, EJS)

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

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

Similar result in density estimation recently (Lerasle, 2009)

Non-asymptotic pathwise oracle inequality

- W exchangeable (e.g., bootstrap or subsampling)
- $C \approx C_W$
- 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. 2009, EJS)

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

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

Similar result in density estimation recently (Lerasle, 2009)

Non-asymptotic pathwise oracle inequality

- W exchangeable (e.g., bootstrap or subsampling)
- $C \approx C_W$
- 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. 2009, EJS)

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

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

Similar result in density estimation recently (Lerasle, 2009)

Non-asymptotic pathwise oracle inequality

- W exchangeable (e.g., bootstrap or subsampling)
- $C \approx C_W$
- 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. 2009, EJS)

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

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

Similar result in density estimation recently (Lerasle, 2009)

Non-asymptotic pathwise oracle inequality

- W exchangeable (e.g., bootstrap or subsampling)
- $C \approx C_W$
- 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. 2009, EJS)

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

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

Similar result in density estimation recently (Lerasle, 2009)

V-fold penalization

- V-fold penalty:

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

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

with $C \geq V - 1$ to be chosen (no bias if $C = V - 1$, see also Burman, 1989)

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

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

$\Rightarrow$ oracle inequality with constant $1 + \ln(n)^{-1/5}$ if $V = \mathcal{O}(1)$ or $V = n$ (A. 2008, arXiv:0802.0566)

V-fold penalization in the general framework

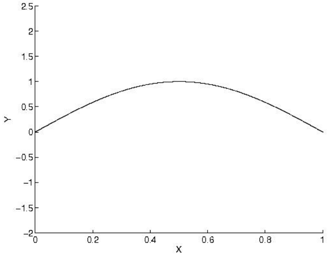
- Resampling and V-fold penalization are well-defined in the general framework
- Constant C_W or $V-1$: could be estimated with the slope heuristics (A. and Massart, JMLR 2009)
- Constant $V-1$ for V-fold penalization:

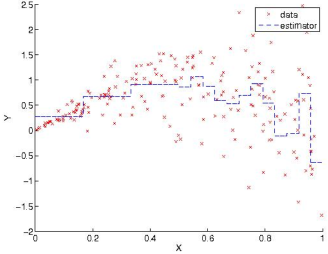
$$\operatorname{pen}_{\mathrm{VF}}(m,n) = \frac{C}{V}\left(P_n^{(j)} - P_n^{(-j)}\right)\gamma\left(\widehat{s}_m^{(-j)}\right)$$

$$\Rightarrow \quad \mathbb{E}\left[\operatorname{pen}_{\mathrm{VF}}(m,n)\right] = \frac{C\mathbb{E}\left[\operatorname{pen}_{\mathrm{id}}\left(m, \frac{n(V-1)}{V}\right)\right]}{V}$$

$$= \frac{C\mathbb{E}\left[\operatorname{pen}_{\mathrm{id}}(m,n)\right]}{V-1} \quad \text{if} \quad \mathbb{E}\left[\operatorname{pen}_{\mathrm{id}}(m,n)\right] \approx \frac{\alpha(m)}{n}$$

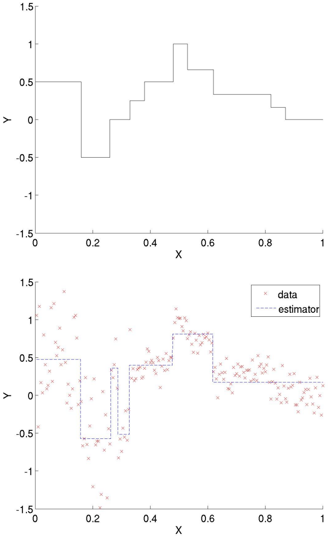
Simulations: sin, $n = 200$, $\sigma(x) = x$, 2 bin sizes





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

Simulations: change-point detection, $n = 200$



$N = 5000$ samples generated

5-fold	1.436 ± 0.008
10-fold	1.400 ± 0.008
20-fold	1.372 ± 0.008
pen 5-f	1.615 ± 0.011
pen 10-f	1.444 ± 0.009
pen 20-f	1.390 ± 0.008
pen 5-f $\times 1.25$	1.462 ± 0.008
pen 10-f $\times 1.25$	1.379 ± 0.008
pen 20-f $\times 1.25$	1.315 ± 0.007

Conclusion

- Usual model selection procedures (C_p, V-fold cross-validation) are suboptimal in some realistic frameworks
- Resampling and V-fold penalties are (first order) optimal and robust to unknown variations of the noise-level
- Theoretical results for regressograms (and recently in density estimation by Lerasle, see CPS 49),
  but these procedures are well-defined in the general framework, rely on a widely valid heuristics, and experimentally perform well.