Model selection via penalization, resampling and cross-validation, with application to change-point detection

Sylvain Arlot

[1] CNRS

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

Cergy, January 30 – February 2, 2012

Outline of the lectures

1. Model selection via penalization, with application to change-point detection
2. Resampling methods for penalization, and robustness to heteroscedasticity in regression
3. Cross-validation for model/estimator selection, with application to detecting changes in the mean of a signal

Part II

Resampling methods for penalization, and robustness to heteroscedasticity in regression

Outline

Outline

Heteroscedastic regression framework

- Random design: $(X_1, Y_1), \ldots, (X_n, Y_n) \in \mathcal{X} \times \mathbb{R}$ i.i.d.

$$Y_i = \eta(X_i) + \varepsilon_i$$

$$\mathbb{E}\left[\varepsilon_i \mid X_i\right] = 0 \quad \text{and} \quad \mathbb{E}\left[\varepsilon_i^2 \mid X_i\right] = \sigma^2(X_i)$$

Heteroscedastic regression framework

- Random design: $(X_1, Y_1), \ldots, (X_n, Y_n) \in \mathcal{X} \times \mathbb{R}$ i.i.d.

$$Y_i = \eta(X_i) + \varepsilon_i$$

$$\mathbb{E}\left[\varepsilon_i \mid X_i\right] = 0 \quad \text{and} \quad \mathbb{E}\left[\varepsilon_i^2 \mid X_i\right] = \sigma^2(X_i)$$

- Quadratic loss:

$$P\gamma(t) = \mathbb{E}_{(X,Y)\sim P}\left[\gamma(t;(X,Y))\right] = \mathbb{E}_{(X,Y)\sim P}\left[\left(t(X) - Y\right)^2\right]$$

- Excess loss: $\eta = s^\star$ and

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

Regressograms

For any finite partition m of $\mathcal{X}$

$$S_m := \left\{ \sum_{\lambda \in m} \alpha_\lambda \mathbb{1}_\lambda \text{ s.t. } \alpha \in \mathbb{R}^m \right\}$$

$\Rightarrow$ least-squares estimator over S_m (regressogram):

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

Regressograms

For any finite partition m of $\mathcal{X}$

$$S_m := \left\{ \sum_{\lambda \in m} \alpha_\lambda \mathbb{1}_\lambda \text{ s.t. } \alpha \in \mathbb{R}^m \right\}$$

$\Rightarrow$ least-squares estimator over S_m (regressogram):

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

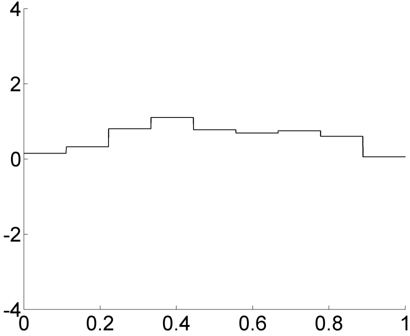
If for every $\lambda \in m$

$$\widehat{p}_\lambda = \widehat{p}_\lambda(D_n) = \frac{1}{n} \operatorname{Card} \left\{ i \text{ s.t. } X_i \in \lambda \right\} > 0$$

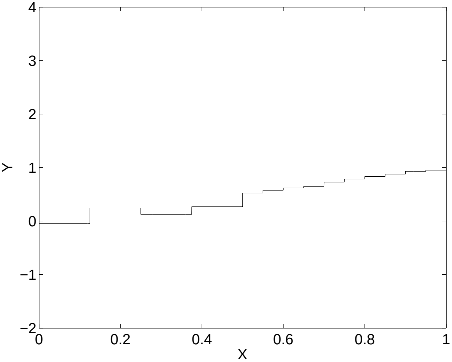
$$\widehat{s}_m = \sum_{\lambda \in m} \widehat{\beta}_\lambda \mathbb{1}_\lambda \qquad \widehat{\beta}_\lambda := \frac{1}{n \widehat{p}_\lambda} \sum_{i \text{ s.t. } X_i \in \lambda} Y_i$$

Regressograms: examples ($\mathcal{X} = [0,1]$)

$\mathcal{M}_n^{(\mathrm{reg})}$ (regular partitions)



$\mathcal{M}_n^{(\mathrm{reg},1/2)}$ (regular partitions on $[0,1/2]$ and on $[1/2,1]$)



Regressograms: bias, ideal penalty

Regressograms: bias, ideal penalty

$$s_m^\star = \sum_{\lambda \in m} \beta_\lambda \mathbb{1}_\lambda \qquad \beta_\lambda := \mathbb{E}_{(X,Y)\sim P}\left[Y \mid X \in \lambda \right]$$

$$\ell\left(s^\star, s_m^\star\right) = \sum_{\lambda \in m} p_\lambda \left(\sigma_\lambda^{(d)}\right)^2 \qquad \left(\sigma_\lambda^{(d)}\right)^2 := \mathbb{E}\left[\left(\beta_\lambda - s^\star(X)\right)^2 \,\middle|\, X \in \lambda \right]$$

Regressograms: bias, ideal penalty

$$s_m^\star = \sum_{\lambda \in m} \beta_\lambda \mathbb{1}_\lambda \qquad \beta_\lambda := \mathbb{E}_{(X,Y)\sim P}\left[Y \mid X \in \lambda \right]$$

$$\ell\left(s^\star, s_m^\star\right) = \sum_{\lambda \in m} p_\lambda \left(\sigma_\lambda^{(d)}\right)^2 \qquad \left(\sigma_\lambda^{(d)}\right)^2 := \mathbb{E}\left[\left(\beta_\lambda - s^\star(X)\right)^2 \,\middle|\, X \in \lambda \right]$$

$$\mathrm{pen}_{\mathrm{id}}(m) = p_1(m) + p_2(m) - \delta(m)$$

$$p_1(m) = P\left(\gamma\left(\widehat{s}_m\right) - \gamma\left(s_m^\star\right)\right) = \sum_{\lambda \in m} p_\lambda \left(\widehat{\beta}_\lambda - \beta_\lambda\right)^2$$

$$p_2(m) = P_n\left(\gamma\left(s_m^\star\right) - \gamma\left(\widehat{s}_m\right)\right) = \sum_{\lambda \in m} \widehat{p}_\lambda \left(\widehat{\beta}_\lambda - \beta_\lambda\right)^2$$

$$\delta(m) = (P_n - P)\gamma\left(s_m^\star\right)$$

Regressograms: conditional expectations

$$\mathcal{P}_m := \left(\mathbb{1}_{X_i \in \lambda} \right)_{1 \leq i \leq n, \lambda \in m}$$

$$\mathbb{E}\left[p_1(m) \mid \mathcal{P}_m \right] = \frac{1}{n} \sum_{\lambda \in m} \frac{p_\lambda}{\widehat{p}_\lambda} \sigma_\lambda^2$$

$$\mathbb{E}\left[p_2(m) \mid \mathcal{P}_m \right] = \frac{1}{n} \sum_{\lambda \in m} \sigma_\lambda^2$$

$$\sigma_\lambda^2 := \mathbb{E}_{(X,Y)\sim P}\left[\left(Y - \beta_\lambda \right)^2 \middle| \, X \in \lambda \right] = \left(\sigma_\lambda^{(d)} \right)^2 + \left(\sigma_\lambda^{(r)} \right)^2$$

$$\left(\sigma_\lambda^{(r)} \right)^2 := \mathbb{E}_{(X,Y)\sim P}\left[\left(\sigma(X) \right)^2 \middle| \, X \in \lambda \right]$$

Regressograms: expectations

$$\mathbb{E}\left[p_1(m) \right] = \frac{1}{n} \sum_{\lambda \in m} \sigma_\lambda^2 \left(1 + \delta_{n,p_\lambda} \right)$$

$$\mathbb{E}\left[p_2(m) \right] = \frac{1}{n} \sum_{\lambda \in m} \sigma_\lambda^2$$

$$\delta_{n,p_\lambda} := \mathbb{E}\left[\frac{p_\lambda}{\widehat{p}_\lambda} \,\middle|\, \widehat{p}_\lambda > 0 \right] - 1$$

$$-\exp(-np) \leq \delta_{n,p} \leq \min\left\{ 1 + \frac{\kappa_1}{(np)^{1/4}} \ , \ \kappa_2 \right\}$$

Regressograms: risk, expectation of the ideal penalty

$$\mathbb{E}\left[\ell\left(s^{\star}, \widehat{s}_m\right)\right] = \sum_{\lambda \in m} p_{\lambda} \left(\sigma_{\lambda}^{(d)}\right)^2 + \frac{1}{n} \sum_{\lambda \in m} \left(1 + \delta_{n,p_{\lambda}}\right) \sigma_{\lambda}^2$$

$$\mathbb{E}\left[\mathrm{pen}_{\mathrm{id}}(m)\right] = \frac{1}{n} \sum_{\lambda \in m} \left(2 + \delta_{n,p_{\lambda}}\right) \sigma_{\lambda}^2$$

Outline

Drawbacks of pen = pen(D_m)

$$Y = s^\star(X) + \varepsilon \quad \text{with} \quad X \sim \mathcal{U}([0,1])$$

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

$m \in \mathcal{M}_n^{(\mathrm{reg},1/2)}$: $D_{m,1}$ pieces on $\left[0, \frac{1}{2}\right]$

$D_{m,2}$ pieces on $\left[\frac{1}{2}, 1\right]$

Drawbacks of pen = pen(D_m)

$$Y = s^{\star}(X) + \varepsilon \quad \text{with} \quad X \sim \mathcal{U}([0,1])$$

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

$$m \in \mathcal{M}_n^{(\mathrm{reg},1/2)}: \begin{array}{l} D_{m,1} \text{ pieces on } \left[0, \frac{1}{2}\right] \\ D_{m,2} \text{ pieces on } \left[\frac{1}{2}, 1\right] \end{array}$$

$$\mathbb{E}\left[\mathrm{pen}_{\mathrm{id}}(m)\right] \approx \frac{4}{n}\left[D_{m,1} \int_0^{1/2} \left(\sigma(x)\right)^2 dx + D_{m,2} \int_{1/2}^{1} \left(\sigma(x)\right)^2 dx\right]$$

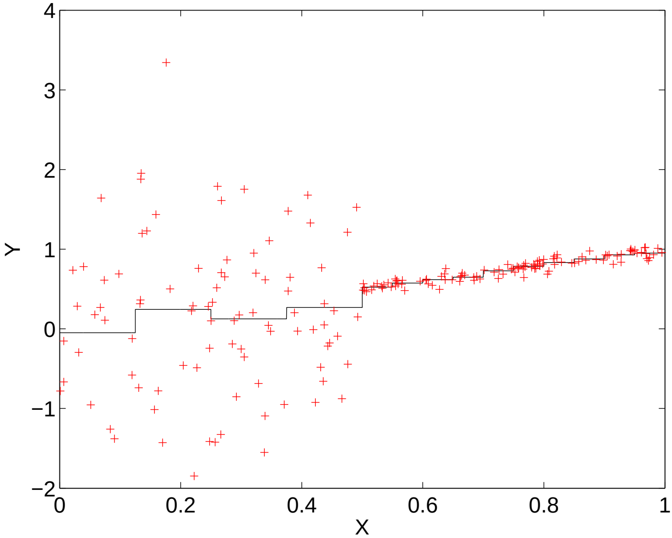
Drawbacks of pen = pen(D_m): an example

$$Y = s^\star(X) + \varepsilon \quad \text{with} \quad X \sim \mathcal{U}([0,1])$$
$$\mathcal{L}(\varepsilon \mid X) = \mathcal{N}\left(0, \sigma(X)^2\right)$$

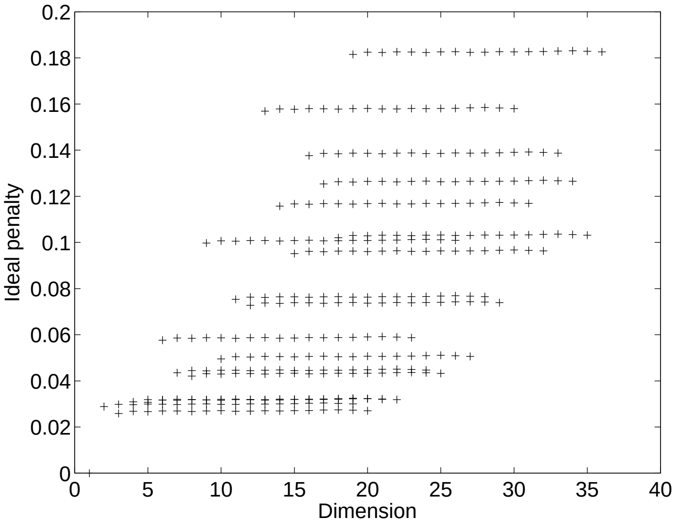
$$s^\star(X) = X \qquad \sigma(X) = \mathbb{1}_{X \leq \frac{1}{2}} + \frac{1}{20}\mathbb{1}_{X > 1/2}$$

$$\mathbb{E}\left[\mathrm{pen}_{\mathrm{id}}(m)\right] \approx \frac{2}{n}\left[D_{m,1} + \frac{D_{m,2}}{400}\right]$$

Example: data and oracle ($n = 200$)



Example: $\mathrm{pen}_{\mathrm{id}}(m)$ as a function of D_m



Penalties function of the dimension

Lemma

For any $D \in \mathcal{D}_n = \{ D_m \text{ s.t. } m \in \mathcal{M}_n \}$

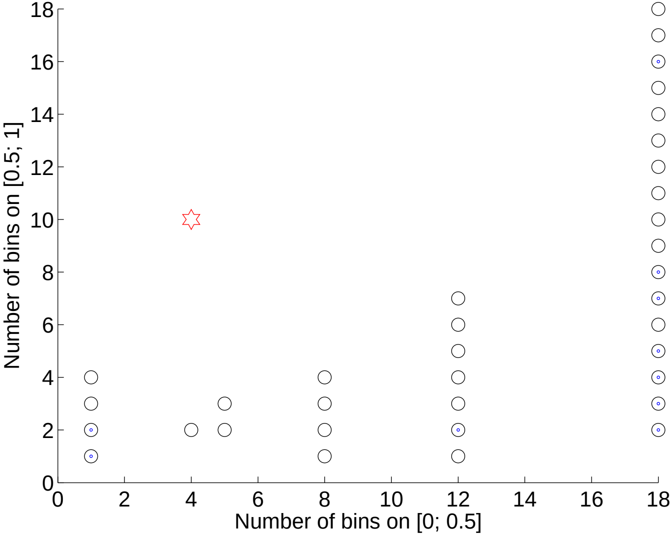
$$\mathcal{M}_{\dim}(D) := \operatorname{argmin}_{m \in \mathcal{M}_n \text{ s.t. } D_m = D} \{ P_n \gamma (\widehat{s}_m) \}$$

$$\mathcal{M}_{\dim} := \bigcup_{D \in \mathcal{D}_n} \mathcal{M}_{\dim}(D)$$

Then, $\forall F : \mathcal{M}_n \mapsto \mathbb{R} \; \forall (X_i, Y_i)_{1 \leq i \leq n}$

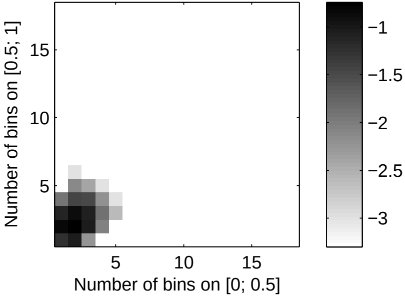
$$\operatorname{argmin}_{m \in \mathcal{M}_n} \{ P_n \gamma (\widehat{s}_m) + F(D_m) \} \subset \mathcal{M}_{\dim}$$

Models that can be selected with pen(D_m)

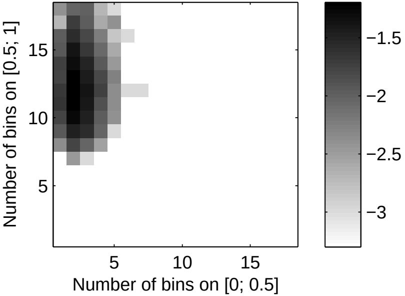


Drawbacks of pen = pen(D_m): $\widehat{m}(D^\star) \neq m^\star$

Densities of $(D_{\widehat{m}(D^\star),1}, D_{\widehat{m}(D^\star),2})$ and $(D_{m^\star,1}, D_{m^\star,2})$ over $N = 1000$ samples



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



$m^\star$

Towards a proof: concentration of $\mathrm{pen}_{\mathrm{id}}$

Assumption: $\|Y\|_\infty \leq A < \infty$ and $\sigma(\cdot) \geq \sigma_{\min} > 0$

- Concentration of p_1 and p_2:
  if $\min_{\lambda \in m} \{ np_\lambda \} \geq \Diamond \ln(n)$, with probability at least $1 - Ln^{-\gamma}$, for $i = 1, 2$

$$\left| p_i(m) - \mathbb{E}\left[p_i(m) \right] \right| \leq \frac{L_{A,\sigma_{\min},\gamma} \left(\ln(n) \right)^2}{\sqrt{D_m}} \mathbb{E}\left[p_2(m) \right]$$

- Bernstein's inequality: with probability at least $1 - 2e^{-x}$,

$$\forall \theta \in (0,1] \quad , \quad \left| (P_n - P) \left(\gamma\left(s_m^\star \right) - \gamma\left(s^\star \right) \right) \right| \leq \theta \ell\left(s^\star, s_m^\star \right) + \frac{6A^2 x}{\theta n}$$

Heuristical proof: expectations

$$\mathbb{E}\left[p_1(m) \right] \approx \mathbb{E}\left[p_2(m) \right] \approx \frac{\beta_1 D_{m,1}}{n} + \frac{\beta_2 D_{m,1}}{n}$$

$$\beta_1 = 2 \int_0^{1/2} \sigma^2 \qquad \beta_2 = 2 \int_{1/2}^{1} \sigma^2$$

$$\ell\left(s^{\star}, s_m^{\star} \right) \approx \frac{\alpha_1}{D_{m,1}^2} + \frac{\alpha_2}{D_{m,2}^2}$$

$$\alpha_1 = \frac{1}{48} \int_0^{1/2} \left(s^{\star\prime} \right)^2 \qquad \alpha_2 = \frac{1}{48} \int_{1/2}^{1} \left(s^{\star\prime} \right)^2$$

Heuristical proof: expectations

$$\beta_1 = 2 \int_0^{1/2} \sigma^2 \qquad \beta_2 = 2 \int_{1/2}^{1} \sigma^2$$

$$\alpha_1 = \frac{1}{48} \int_0^{1/2} \left(s^{\star\prime} \right)^2 \qquad \alpha_2 = \frac{1}{48} \int_{1/2}^{1} \left(s^{\star\prime} \right)^2$$

$$P_n\gamma\left(\widehat{s}_m\right) - P\gamma\left(s^\star\right) \approx \frac{\alpha_1}{D_{m,1}^2} + \frac{\alpha_2}{D_{m,2}^2} - \frac{\beta_1 D_{m,1}}{n} - \frac{\beta_2 D_{m,1}}{n}$$

$$\ell\left(s^\star, \widehat{s}_m\right) \approx \frac{\alpha_1}{D_{m,1}^2} + \frac{\alpha_2}{D_{m,2}^2} + \frac{\beta_1 D_{m,1}}{n} + \frac{\beta_2 D_{m,1}}{n}$$

Heuristical proof: expectations

$$\beta_1 = 2 \int_0^{1/2} \sigma^2 \quad > \quad \beta_2 = 2 \int_{1/2}^1 \sigma^2$$

$$\alpha_1 = \frac{1}{48} \int_0^{1/2} \left(s^{\star\prime} \right)^2 \qquad \alpha_2 = \frac{1}{48} \int_{1/2}^1 \left(s^{\star\prime} \right)^2$$

$$P_n \gamma \left(\widehat{s}_m \right) - P \gamma \left(s^\star \right) \approx \frac{\alpha_1}{D_{m,1}^2} + \frac{\alpha_2}{D_{m,2}^2} - \frac{\beta_1 D_{m,1}}{n} - \frac{\beta_2 D_{m,1}}{n}$$

$$\ell \left(s^\star, \widehat{s}_m \right) \approx \frac{\alpha_1}{D_{m,1}^2} + \frac{\alpha_2}{D_{m,2}^2} + \frac{\beta_1 D_{m,1}}{n} + \frac{\beta_2 D_{m,1}}{n}$$

$$m^\star \approx \left(\left(\frac{2\alpha_1 n}{\beta_1} \right)^{1/3} , \left(\frac{2\alpha_2 n}{\beta_2} \right)^{1/3} \right)$$

Drawbacks of pen $=$ pen(D_m): theory

$$Y = s^\star(X) + \varepsilon \quad \text{with} \quad X \sim \mathcal{U}([0,1]) \ , \qquad \mathbb{E}\left[\varepsilon^2 \,\middle|\, X \right] = \sigma(X)$$

$$\text{and} \quad \sigma_a^2 = \int_0^{1/2} (\sigma(x))^2 \, dx \neq \int_{1/2}^{1} (\sigma(x))^2 \, dx = \sigma_b^2$$

Theorem (A. 2008)

If $\mathcal{M} = \mathcal{M}_n^{(\mathrm{reg},1/2)}$, under "reasonable" assumptions on $(s^\star, \varepsilon, \sigma)$, $\exists \eta(\sigma_a^2 / \sigma_b^2) > 0$ such that with probability at least $1 - C(\lVert\varepsilon\rVert_\infty, \sigma_a^2, \sigma_b^2, \lVert s^{\star\prime}\rVert_\infty, \lVert s^{\star\prime\prime}\rVert_\infty) n^{-2}$

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

$$\ell\left(s^\star, \widehat{s}_{\widehat{m}_F}\right) \geq \left(1 + \eta\left(\frac{\sigma_a^2}{\sigma_b^2} \right) \right) \inf_{m \in \mathcal{M}_n} \left\{ \ell\left(s^\star, \widehat{s}_m\right) \right\}$$

Why should we estimate the shape of the penalty?

- $\mathrm{pen}(D) = F(D) \Rightarrow$ loss of a factor $(1 + \eta) > 1$

Why should we estimate the shape of the penalty?

- $\mathrm{pen}(D) = F(D) \Rightarrow$ loss of a factor $(1 + \eta) > 1$
- $\mathrm{pen}(m) = 2\mathbb{E}\left[\sigma(X)^2\right] D_m / n \Rightarrow$ possible burst of the risk

Why should we estimate the shape of the penalty?

- $\mathrm{pen}(D) = F(D) \Rightarrow$ loss of a factor $(1 + \eta) > 1$
- $\mathrm{pen}(m) = 2\mathbb{E}\left[\sigma(X)^2\right] D_m/n \Rightarrow$ possible burst of the risk
- $\mathrm{pen}(m) = 2\left\|\sigma\right\|_{\infty}^2 D_m/n \Rightarrow$ oracle-inequality with constant $\mathcal{O}(\max \sigma^2 / \min \sigma^2)$

Why should we estimate the shape of the penalty?

- $\mathrm{pen}(D) = F(D) \Rightarrow$ loss of a factor $(1+\eta) > 1$
- $\mathrm{pen}(m) = 2\mathbb{E}\left[\sigma(X)^2\right] D_m/n \Rightarrow$ possible burst of the risk
- $\mathrm{pen}(m) = 2\left\|\sigma\right\|_\infty^2 D_m/n \Rightarrow$ oracle-inequality with constant $\mathcal{O}(\max\sigma^2/\min\sigma^2)$

$\Rightarrow$ must estimate $\mathbb{E}[\mathrm{pen}_{\mathrm{id}}(m)]$ for an oracle inequality with constant $(1+o(1))$ and for avoiding overfitting

Outline

Resampling heuristics (bootstrap, Efron 1979)

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

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

Resampling heuristics (bootstrap, Efron 1979)

$$
\begin{tikzcd}
\text{Real world :} \arrow[d, rightsquigarrow] & P \arrow[r, "\text{sampling}"] & P_n \arrow[r, Rightarrow] & \widehat{s}_m \\
\text{Bootstrap world :} & P_n & &
\end{tikzcd}
$$

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

Resampling heuristics (bootstrap, Efron 1979)

$$\begin{tikzcd} \text{Real world :} \arrow[d, squiggly] & P \arrow[r, "\text{sampling}"] & P_n \arrow[r, Rightarrow] & \widehat{s}_m \\ \text{Bootstrap world :} & P_n \arrow[r, "\text{resampling}"] & P_n^W \arrow[r, Rightarrow] & \widehat{s}_m^W \end{tikzcd}$$

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

Exchangeable weighted resampling

$$P_n^W := \frac{1}{n} \sum_{i=1}^{n} W_i \delta_{\xi_i}$$

- Bootstrap:

$$W \sim \mathcal{M}\left(n; \frac{1}{n}, \ldots, \frac{1}{n}\right)$$

Exchangeable weighted resampling

$$P_n^W := \frac{1}{n} \sum_{i=1}^{n} W_i \delta_{\xi_i}$$

- Efron(m) or m out of n bootstrap:

$$\frac{m}{n} W \sim \mathcal{M}\left(m; \frac{1}{n}, \ldots, \frac{1}{n} \right)$$

Exchangeable weighted resampling

$$P_n^W := \frac{1}{n} \sum_{i=1}^{n} W_i \delta_{\xi_i} \quad \text{or} \quad \frac{1}{\sum_k W_k} \sum_{i=1}^{n} W_i \delta_{\xi_i} = \frac{1}{n} \sum_{i=1}^{n} \frac{W_i}{\overline{W}} \delta_{\xi_i}$$

- Efron(m) or m out of n bootstrap:

$$\frac{m}{n} W \sim \mathcal{M}\left(m; \frac{1}{n}, \dots, \frac{1}{n} \right)$$

- Subsampling:
  - Random-hold out(q), $q \in \{1, \dots, n-1\}$:

$$W_i = \frac{n}{q} \mathbb{1}_{i \in I} \quad \text{with} \quad I \sim \mathcal{U}\left(\mathfrak{P}_q(\{1, \dots, n\})\right)$$

  - Rademacher(p) or Bernoulli:

$$pW_1, \dots, pW_n \text{ i.i.d. } \sim \mathcal{B}(p)$$

Theoretical justification: asymptotics

Theorem (van der Vaart & Wellner, 1996)

Let $(W_{n,1}, \ldots, W_{n,n}) \in \mathbb{R}^n$ be a non-negative random vector, exchangeable, independent from $\xi_{1\ldots n}$, bounded and such that

$$n^{-1} \sum_{i=1}^{n} \left(W_{n,i} - \overline{W}_n \right)^2 \xrightarrow{(p)} c^2 > 0 \enspace .$$

Then, as n goes to infinity,

$$\sup_{h \in BL_1} \left| \mathbb{E}_W \left[h \left(\sqrt{n} \left(P_n^W - \overline{W}_n P_n \right) \right) \right] - \mathbb{E} \left[h \left(c \mathbb{G} \right) \right] \right| \xrightarrow{(p)} 0$$

where $\mathbb{G}$ is a Gaussian process, limit of $\sqrt{n}(P_n - P)$, with zero mean and covariance function $\mathrm{cov}(f, g) = P(fg) - P(f)P(g)$.

Classical uses of resampling

- estimating a variance, a quadratic risk
- estimation and/or bias correction
- confidence intervals, p-values
- estimation of prediction error, model selection
- stabilization (bagging, random forests)
- ...

A resampling-based estimator of variance

Framework:

$$\xi_1, \ldots, \xi_n \text{ i.i.d. } \sim P \qquad \mathbb{E}[\xi_i] = \mu \quad \mathbb{E}\left[(\xi_i - \mu)^2\right] = \sigma^2$$

$$\sigma^2 = n\mathbb{E}\left[\left(\frac{1}{n}\sum_{i=1}^{n}\xi_i - \mu\right)^2\right] = n\mathbb{E}\left[\left(\mathbb{E}_{\xi\sim P_n}\xi - \mathbb{E}_{\xi\sim P}\xi\right)^2\right]$$

$$= n\mathbb{E}\left[F(P, P_n)\right]$$

$\Rightarrow$ resampling-based estimator

$$\widehat{\sigma^2_W} = n\mathbb{E}_W\left[F(P_n, P_n^W)\right]$$

A resampling-based estimator of variance

$$\widehat{\sigma^2_W} = n\mathbb{E}_W \left[F(P_n, P_n^W) \right]$$

$$\widehat{\sigma^2_W} = \frac{R_V^{(W)}}{n} \left[\sum_{i=1}^{n} (\xi_i - \mu)^2 - \frac{1}{n-1} \sum_{i \neq j} (\xi_i - \mu)(\xi_j - \mu) \right]$$

$$R_V^{(W)} := \mathbb{E}_W \left[\left(\frac{nW_i}{\sum_{k=1}^{n} W_k} - 1 \right)^2 \right]$$

Comparison with the classical variance estimator

Classical unbiased estimator of variance:

$$\widehat{\sigma^2} = \frac{1}{n-1} \sum_{i=1}^{n} \left(\xi_i - \frac{1}{n} \sum_{k=1}^{n} \xi_k \right)^2$$

Comparison with the classical variance estimator

Classical unbiased estimator of variance:

$$\widehat{\sigma^2} = \frac{1}{n-1} \sum_{i=1}^{n} \left(\xi_i - \frac{1}{n} \sum_{k=1}^{n} \xi_k \right)^2$$

$$\widehat{\sigma^2_W} = R_V^{(W)} \widehat{\sigma^2}$$

$$\Rightarrow \qquad \mathbb{E}\left[\widehat{\sigma^2_W} \right] = R_V^{(W)} \sigma^2$$

Resampling and structure

- Properties of $F(P, P_n) = (\mathbb{E}_{\xi \sim P_n} \xi - \mathbb{E}_{\xi \sim P} \xi)^2$:
  - exchangeable
  - translation-invariance
  - homogeneity
  - polynomial function of ξ_i and $\mathbb{E}_{\xi \sim P} \xi$

$\Rightarrow$ $\mathbb{E}_W [F(P_n, P_n^W)]$ has similar properties

Resampling and structure

- Properties of $F(P, P_n) = (\mathbb{E}_{\xi \sim P_n} \xi - \mathbb{E}_{\xi \sim P} \xi)^2$:
  - exchangeable
  - translation-invariance
  - homogeneity
  - polynomial function of ξ_i and $\mathbb{E}_{\xi \sim P} \xi$

$\Rightarrow$ $\mathbb{E}_W[F(P_n, P_n^W)]$ has similar properties

$$\Rightarrow \qquad \mathbb{E}_W \left[F(P_n, P_n^W) \right] \propto \widehat{\sigma^2}$$

Resampling and concentration

Over-concentration phenomenon for the resampling-based estimator:

$$\operatorname{var}\left(n \left(\frac{1}{n} \sum_{i=1}^{n} \xi_i - \mu \right)^2 \right) = 2\sigma^4 + \frac{\mathbb{E}\left[(\xi_1 - \mu)^4 \right] - 3\sigma^4}{n}$$

$$\operatorname{var}\left(\frac{1}{R_V^{(W)}} \widehat{\sigma_W^2} \right) = \frac{1}{n} \left(\mathbb{E}\left[(\xi_1 - \mu)^4 \right] - \sigma^4 \right) + \frac{2}{n(n-1)} \sigma^4$$

Computation of the multiplicative factor

$$R_V^{(W)} := \mathbb{E}_W \left[\left(\frac{nW_i}{\sum_{k=1}^n W_k} - 1 \right)^2 \right]$$

Computation of the multiplicative factor

$$R_V^{(W)} := \mathbb{E}_W \left[\left(\frac{nW_i}{\sum_{k=1}^n W_k} - 1 \right)^2 \right]$$

Efron(m): $R_V^{(W)} = \frac{n-1}{m}$

Rademacher(p): $R_V^{(W)} = \frac{1+\delta_{n,p}}{p} - 1 \approx \frac{1}{p} - 1$

Random hold-out(q): $R_V^{(W)} = \frac{n}{q} - 1$

Leave-one-out = Rho($n-1$): $R_V^{(W)} = \frac{1}{n-1}$

Resampling-based estimator of $\mathrm{pen}_{\mathrm{id}}(m)$

- Ideal penalty:

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

- Resampling-based estimator of $\mathbb{E}\left[F(P, P_n) \right]$:

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

- bootstrap (Efron, 1983; Shibata, 1997), m out of n bootstrap for identification (Shao, 1996), general exchangeable weights (A. 2009)

- Multiplicative factor C_W: why? how can we estimate it?

Rademacher penalties

- Global penalties:

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

Rademacher penalties

- Global penalties:

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

- Global Rademacher penalties in classification (Koltchinskii & Panchenko, 2001; Bartlett, Boucheron & Lugosi, 2002), exchangeable weights (Fromont, 2004)

$$\mathbb{E}\left[\sup_{t \in S_m} \left\{ \frac{1}{n} \sum_{i=1}^{n} \varepsilon_i \gamma\left(t; \xi_i\right) \right\} \,\middle|\, P_n \right]$$

$$\text{with} \quad \varepsilon_1, \ldots, \varepsilon_n \text{ i.i.d. } \sim \mathcal{U}(\{-1, +1\})$$

- Local Rademacher complexities (Bartlett, Bousquet & Mendelson, 2004; Koltchinskii, 2006)

Outline

Reminder

$$\widehat{s}_m = \sum_{\lambda \in m} \widehat{\beta}_\lambda \mathbb{1}_\lambda \quad \text{with} \quad \widehat{\beta}_\lambda := \frac{1}{n \widehat{p}_\lambda} \sum_{i \text{ s.t. } X_i \in \lambda} Y_i$$

$$\widehat{p}_\lambda = \widehat{p}_\lambda(D_n) = \frac{1}{n} \operatorname{Card} \{ i \text{ s.t. } X_i \in \lambda \}$$

$$\operatorname{pen}_{\mathrm{id}}(m) = p_1(m) + p_2(m) - \delta(m)$$

$$p_1(m) = P\left(\gamma\left(\widehat{s}_m\right) - \gamma\left(s_m^\star\right)\right) = \sum_{\lambda \in m} \left[p_\lambda \left(\widehat{\beta}_\lambda - \beta_\lambda \right)^2 \right]$$

$$p_2(m) = P_n\left(\gamma\left(s_m^\star\right) - \gamma\left(\widehat{s}_m\right)\right) = \sum_{\lambda \in m} \left[\widehat{p}_\lambda \left(\widehat{\beta}_\lambda - \beta_\lambda \right)^2 \right]$$

$$\delta(m) = (P_n - P)\gamma\left(s_m^\star\right)$$

Resampling-based penalty

$$\mathrm{pen}_W(m) = \frac{C_W}{n} \sum_{\lambda \in m} \frac{R_{1,W} + R_{2,W}}{n\widehat{p}_\lambda - 1} \left(S_{\lambda,2} - \frac{1}{n\widehat{p}_\lambda} S_{\lambda,1}^2 \right) \mathbb{1}_{n\widehat{p}_\lambda \geq 2}$$

with $\quad S_{\lambda,1} := \sum_{X_i \in \lambda} (Y_i - \beta_\lambda) \qquad S_{\lambda,2} := \sum_{X_i \in \lambda} (Y_i - \beta_\lambda)^2$

$$R_{1,W}(n, \widehat{p}_\lambda) := \mathbb{E}\left[\frac{(W_1 - \widehat{W}_\lambda)^2}{\widehat{W}_\lambda^2} \,\middle|\, X_1 \in \lambda \,,\, \widehat{W}_\lambda > 0 \right]$$

and $\quad R_{2,W}(n, \widehat{p}_\lambda) := \mathbb{E}\left[\frac{(W_1 - \widehat{W}_\lambda)^2}{\widehat{W}_\lambda} \,\middle|\, X_1 \in \lambda \right]$

Value of R_1 and R_2: examples

$$R_{1,W}(n, \widehat{p}_\lambda) \sim R_{2,W}(n, \widehat{p}_\lambda) \text{ as } n\widehat{p}_\lambda \to \infty$$

$$C_{W,\infty}(n) := \lim_{n\widehat{p}_\lambda \to \infty} \frac{1}{R_{2,W}(n, \widehat{p}_\lambda)}$$

Efron(m): $R_{2,W}(n, \widehat{p}_\lambda) = \dfrac{n}{m}\left(1 - \dfrac{1}{n\widehat{p}_\lambda}\right)$ $\quad C_{W,\infty} = \dfrac{m}{n}$

Rademacher(p): $R_{2,W}(n, \widehat{p}_\lambda) = \dfrac{1}{p} - 1$ $\quad C_{W,\infty} = \dfrac{p}{1-p}$

Random hold-out(q): $R_{2,W}(n, \widehat{p}_\lambda) = \dfrac{n}{q} - 1$ $\quad C_{W,\infty} = \dfrac{q}{n-q}$

Leave-one-out: $R_{2,W}(n, \widehat{p}_\lambda) = \dfrac{1}{n-1}$ $\quad C_{W,\infty} = n - 1$

Expectations

$$\mathbb{E}\left[\, Y_i - \beta_\lambda \mid X_i \in \lambda \right] = 0 \quad \text{and} \quad \mathbb{E}\left[\left(Y_i - \beta_\lambda \right)^2 \middle| \, X_i \in \lambda \right] = \sigma_\lambda^2$$

$$\mathbb{E}\left[\mathrm{pen}_W(m) \mid \mathcal{P}_m \right] = \frac{C_W}{n} \sum_{\lambda \in m} \left(R_{1,W} + R_{2,W} \right) \sigma_\lambda^2 \mathbb{1}_{n\widehat{p}_\lambda \geq 2}$$

$$\mathbb{E}\left[\mathrm{pen}_W(m) \right] = \frac{C_W}{C_{W,\infty}} \frac{1}{n} \sum_{\lambda \in m} \left(2 + \overline{\delta}_{n,p_\lambda}^{(\mathrm{penW})} \right) \sigma_\lambda^2$$

with $\overline{\delta}_{n,p_\lambda}^{(\mathrm{penW})} \to 0$ quand $np_\lambda \to +\infty$

$\Rightarrow$ adaptation to heteroscedasticity

Concentration

Proposition (A. 2009)

- *Bounded data*: $\|Y_i\|_\infty \leq A < \infty$
- *Lower-bounded noise*: $\sigma(X_i) \geq \sigma_{\min} > 0$
- $\mathcal{L}(W)$ *among Efr*(n), *Rad*$(1/2)$, *Rho*$(n/2)$, *Loo*

For every $A_n \geq 2$, *with probability at least* $1 - L_1 n^{-\gamma}$,

$$\left|\mathrm{pen}_W(m) - \mathbb{E}\left[\mathrm{pen}_W(m) \mid \mathcal{P}_m\right]\right| \mathbb{1}_{\min_{\lambda \in m}\{ n\widehat{p}_\lambda \} \geq A_n} \leq \frac{C_W}{C_{W,\infty}} \frac{L_2(A/\sigma_{\min}, \gamma) \ln(n)}{\sqrt{A_n D_m}} \mathbb{E}\left[p_2(m)\right]$$

Pathwise non-asymptotic oracle inequality

- $\mathcal{L}(W)$ among Efr(n), Rad($1/2$), Rho($n/2$), Loo
- $C_W \approx C_{W,\infty}$

Pathwise non-asymptotic oracle inequality

- $\mathcal{L}(W)$ among Efr(n), Rad($1/2$), Rho($n/2$), Loo
- $C_W \approx C_{W,\infty}$

- $\operatorname{Card}(\mathcal{M}_n) \leq C_{\mathcal{M}} n^{\alpha_{\mathcal{M}}}$
- Bounded data: $\|Y_i\|_\infty \leq A < \infty$
- Lower-bounded noise: $\sigma(X_i) \geq \sigma_{\min} > 0$

Pathwise non-asymptotic oracle inequality

- $\mathcal{L}(W)$ among Efr(n), Rad($1/2$), Rho($n/2$), Loo
- $C_W \approx C_{W,\infty}$

- $\mathrm{Card}(\mathcal{M}_n) \leq C_{\mathcal{M}} n^{\alpha_{\mathcal{M}}}$
- Bounded data: $\|Y_i\|_\infty \leq A < \infty$
- Lower-bounded noise: $\sigma(X_i) \geq \sigma_{\min} > 0$

- $s^\star \in \mathcal{H}(\alpha, R)$ non-constant
- Pre-selected models: $\forall m \in \mathcal{M} \quad , \quad \min_{\lambda \in m} n\widehat{p}_\lambda \geq 3$

Pathwise non-asymptotic oracle inequality

- $\mathcal{L}(W)$ among Efr(n), Rad($1/2$), Rho($n/2$), Loo
- $C_W \approx C_{W,\infty}$

- $\mathrm{Card}(\mathcal{M}_n) \leq C_{\mathcal{M}} n^{\alpha_{\mathcal{M}}}$
- Bounded data: $\|Y_i\|_\infty \leq A < \infty$
- Lower-bounded noise: $\sigma(X_i) \geq \sigma_{\min} > 0$

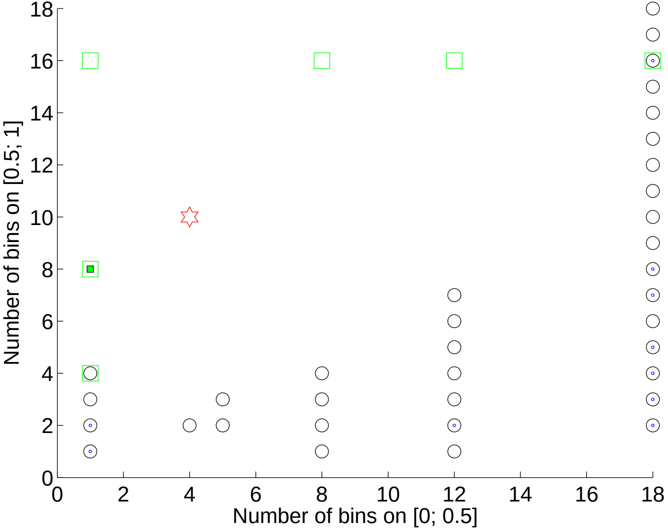
- $s^\star \in \mathcal{H}(\alpha, R)$ non-constant
- Pre-selected models: $\forall m \in \mathcal{M}$, $\min_{\lambda \in m} n\widehat{p}_\lambda \geq 3$

Theorem (A. 2009)

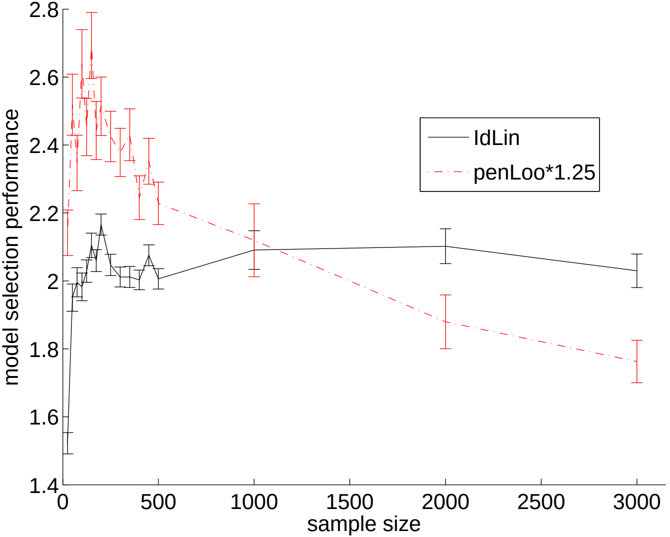
With probability at least $1 - \Diamond n^{-2}$,

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

Models that can be selected: penLoo better than pen(D_m)



Simulations: $1.25 \times \mathrm{pen}_{\mathrm{Loo}}(m)$ vs. $K^\star D_m$



Adaptation

$$\widetilde{s} := \widehat{s}_{\widehat{m}} \quad \text{with} \quad \widehat{m} \in \underset{\substack{m \in \mathcal{M}_n^{(\mathrm{reg})} \\ \min_{\lambda \in m}\{ n\widehat{p}_\lambda \} \geq 3}}{\operatorname{argmin}} \left\{ P_n \gamma\left(\widehat{s}_m\right) + \mathrm{pen}_W(m) \right\}$$

Assumptions:

- Bounded data: $\left\| Y_i \right\|_\infty \leq A < \infty$
- Lower-bounded noise: $\sigma(X_i) \geq \sigma_{\min} > 0$
- Lower-bounded density of X: $\forall I \subset \mathcal{X}$, $\mathbb{P}\left(X \in I \right) \geq c_X^{\min} \mathrm{Leb}(I)$
- $s^\star = \eta \in \mathcal{H}(\alpha, R)$ with $\alpha \in (0, 1]$:

$$\forall x_1, x_2 \in \mathcal{X} \ , \ \left| s^\star(x_1) - s^\star(x_2) \right| \leq R \left\| x_1 - x_2 \right\|_\infty^\alpha$$

Adaptation

$$\widetilde{s} := \widehat{s}_{\widehat{m}} \quad \text{with} \quad \widehat{m} \in \operatorname*{argmin}_{\substack{m \in \mathcal{M}_n^{(\mathrm{reg})} \\ \min_{\lambda \in m}\{ n\widehat{p}_\lambda \} \geq 3}} \left\{ P_n \gamma\left(\widehat{s}_m\right) + \mathrm{pen}_W(m) \right\}$$

$$\mathbb{E}\left[\ell\left(s^{\star}, \widetilde{s}\right)\right] \leq K_2 R^{\frac{2d}{2\alpha+d}} n^{\frac{-2\alpha}{2\alpha+d}} \left\lVert \sigma \right\rVert_{\infty}^{\frac{4\alpha}{2\alpha+d}} + \frac{K_3 A^2}{n^2}$$

Adaptation

$$\widetilde{s} := \widehat{s}_{\widehat{m}} \quad \text{with} \quad \widehat{m} \in \operatorname*{argmin}_{\substack{m \in \mathcal{M}_n^{(\mathrm{reg})} \\ \min_{\lambda \in m} \{ n \widehat{p}_\lambda \} \geq 3}} \left\{ P_n \gamma \left(\widehat{s}_m \right) + \mathrm{pen}_W(m) \right\}$$

$$\mathbb{E} \left[\ell \left(s^\star, \widetilde{s} \right) \right] \leq K_2 R^{\frac{2d}{2\alpha+d}} n^{\frac{-2\alpha}{2\alpha+d}} \left\| \sigma \right\|_\infty^{\frac{4\alpha}{2\alpha+d}} + \frac{K_3 A^2}{n^2}$$

and if $\sigma(\cdot)$ is K_σ-Lipschitz with at most J_σ jumps:

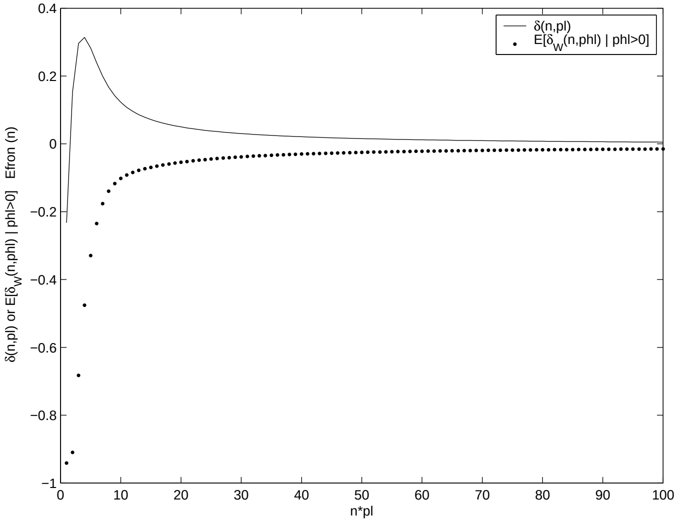
$$\mathbb{E} \left[\ell \left(s^\star, \widetilde{s} \right) \right] \leq K_2 R^{\frac{2d}{2\alpha+d}} n^{\frac{-2\alpha}{2\alpha+d}} \left\| \sigma \right\|_{L^2(\mathrm{Leb})}^{\frac{4\alpha}{2\alpha+d}} + \frac{K_4 A^2}{n^2}$$

Theoretical comparison of weights: reminder

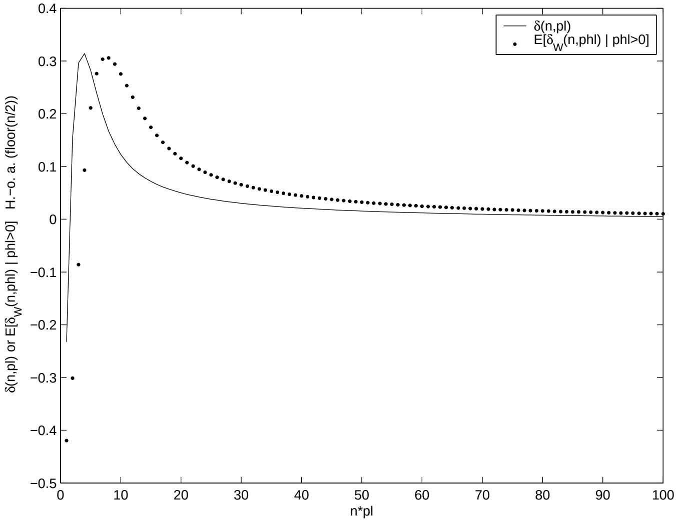
$$\mathbb{E}\left[\mathrm{pen}_{\mathrm{id}}(m)\right] = \frac{1}{n} \sum_{\lambda \in m} \left(2 + \delta_{n,p_\lambda} \right) \sigma_\lambda^2$$

and $$\mathbb{E}\left[\mathrm{pen}_W(m)\right] = \frac{1}{n} \sum_{\lambda \in m} \left(2 + \overline{\delta}_{n,\widehat{p}_\lambda}^{(\mathrm{penW})} \right) \sigma_\lambda^2$$

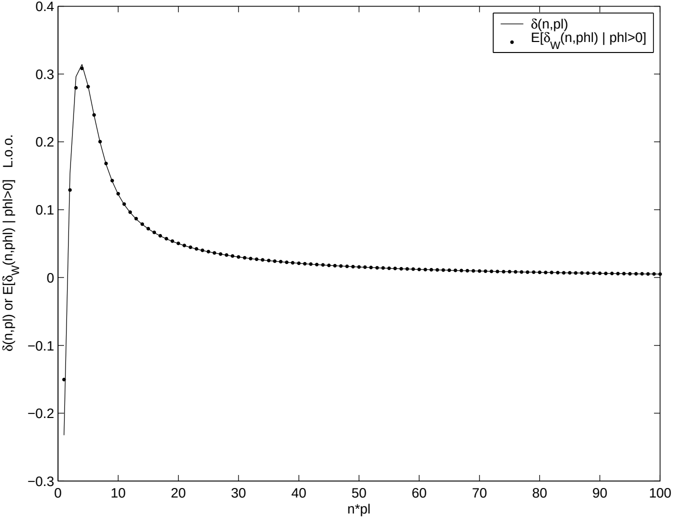
$\overline{\delta}^{(\mathrm{penW})}_{n,\widehat{p}_\lambda}$ vs. δ_{n,p_λ}: Efron(n) $\approx$ Poisson(1)



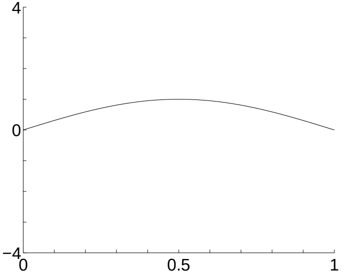
$\overline{\delta}^{(\mathrm{penW})}_{n,\widehat{p}_\lambda}$ vs. δ_{n,p_λ}: Rho$(n/2) \approx$ Rad$(1/2)$

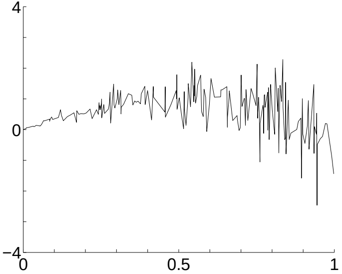


$\overline{\delta}^{(\mathrm{penW})}_{n,\widehat{p}_\lambda}$ vs. δ_{n,p_λ}: Leave-one-out



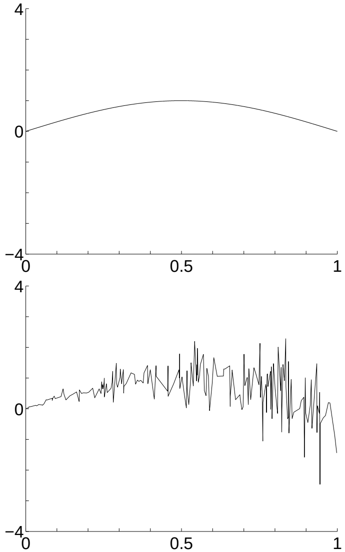
$s^\star(x) = \sin(2\pi x)$ $n = 200$ $\sigma(x) = x$ $\mathcal{M} = \mathcal{M}^{(\mathrm{reg},1/2)}$





Mallows	3.69 ± 0.07
$\mathbb{E}[\mathrm{pen}_{\mathrm{id}}]$	2.30 ± 0.05
pen Efr	3.15 ± 0.07
pen Loo	2.59 ± 0.06
pen Rho	2.50 ± 0.06
pen Rad	2.49 ± 0.06

$s^\star(x) = \sin(2\pi x) \quad n = 200 \quad \sigma(x) = x \quad \mathcal{M} = \mathcal{M}^{(\mathrm{reg},1/2)}$



Mallows	3.69 ± 0.07
$\mathbb{E}[\mathrm{pen}_{\mathrm{id}}]$	2.30 ± 0.05
pen Efr	3.15 ± 0.07
pen Loo	2.59 ± 0.06
pen Rho	2.50 ± 0.06
pen Rad	2.49 ± 0.06
Mallows $\times 1.25$	3.17 ± 0.07
$\mathbb{E}[\mathrm{pen}_{\mathrm{id}}] \times 1.25$	2.03 ± 0.04
pen Efr $\times 1.25$	2.60 ± 0.06
pen Loo $\times 1.25$	2.22 ± 0.05
pen Rho $\times 1.25$	2.14 ± 0.05
pen Rad $\times 1.25$	2.14 ± 0.05

Outline

Least-squares density estimation

- μ reference measure on Ξ
- $f = dP/d\mu \in \mathbb{S} = L^2(\mu)$

Least-squares density estimation

- μ reference measure on Ξ
- $f = dP/d\mu \in \mathbb{S} = L^2(\mu)$

- $\gamma(t;\xi) = \|t\|^2_{L^2(\mu)} - 2t(\xi)$
  $\Rightarrow P\gamma(t) = \|t\|^2_{L^2(\mu)} - 2\langle t, f\rangle_{L^2(\mu)}$
  $\Rightarrow s^\star = f$ and $\ell(s^\star, t) = \|t - s^\star\|^2_{L^2(\mu)}$

Least-squares density estimation

- μ reference measure on Ξ
- $f = dP/d\mu \in \mathbb{S} = L^2(\mu)$

- $\gamma(t;\xi) = \|t\|^2_{L^2(\mu)} - 2t(\xi)$
  $\Rightarrow P\gamma(t) = \|t\|^2_{L^2(\mu)} - 2\left\langle t,\, f\right\rangle_{L^2(\mu)}$
  $\Rightarrow s^\star = f$ and $\ell\left(s^\star, t\right) = \|t - s^\star\|^2_{L^2(\mu)}$

- $\left(\psi_\lambda\right)_{\lambda\in m}$ orthonormal basis of S_m
  $\Rightarrow s^\star_m = \sum_{\lambda\in m}(P\psi_\lambda)\psi_\lambda$ and $\widehat{s}_m = \sum_{\lambda\in m}(P_n\psi_\lambda)\psi_\lambda$

Least-squares density estimation

- μ reference measure on Ξ
- $f = dP/d\mu \in \mathbb{S} = L^2(\mu)$

- $\gamma(t;\xi) = \|t\|^2_{L^2(\mu)} - 2t(\xi)$
  $\Rightarrow P\gamma(t) = \|t\|^2_{L^2(\mu)} - 2\left\langle t,\, f\right\rangle_{L^2(\mu)}$
  $\Rightarrow s^\star = f$ and $\ell\left(s^\star, t\right) = \|t - s^\star\|^2_{L^2(\mu)}$

- $(\psi_\lambda)_{\lambda\in m}$ orthonormal basis of S_m
  $\Rightarrow s^\star_m = \sum_{\lambda\in m}(P\psi_\lambda)\psi_\lambda$ and $\widehat{s}_m = \sum_{\lambda\in m}(P_n\psi_\lambda)\psi_\lambda$

$$\begin{aligned}\mathrm{pen}_{\mathrm{id}}(m) &= (P - P_n)\gamma\left(\widehat{s}_m\right) = 2(P_n - P)(\widehat{s}_m)\\ &= 2\left\|s^\star_m - \widehat{s}_m\right\|^2_{L^2(\mu)} + 2(P_n - P)(s^\star_m)\end{aligned}$$

I.i.d. framework (Lerasle 2009)

$$\mathrm{pen}_{\mathrm{id}}(m) = 2(P_n - P)(\widehat{s}_m)$$

$$\mathrm{pen}_W(m) = C_W \mathbb{E}_W \left[2(P_n^W - \overline{W} P_n)(\widehat{s}_m^W) \right]$$

I.i.d. framework (Lerasle 2009)

$$\mathrm{pen}_W(m) = C_W \mathbb{E}_W \left[2(P_n^W - \overline{W} P_n)(\widehat{s}_m^W) \right]$$

$\Rightarrow$ $\mathrm{pen}_W(m)$ only depends on W through a multiplicative factor

I.i.d. framework (Lerasle 2009)

$$\mathrm{pen}_W(m) = C_W \mathbb{E}_W \left[2(P_n^W - \overline{W} P_n)(\widehat{s}_m^W) \right]$$

$\Rightarrow$ $\mathrm{pen}_W(m)$ only depends on W through a multiplicative factor

$\Rightarrow$ $\mathbb{E}\left[\mathrm{pen}_W(m)\right] = C_W \operatorname{var}\left(W_1 - \overline{W} \right) \mathbb{E}\left[\mathrm{pen}_{\mathrm{id}}(m)\right]$

$+$ concentration of $\mathrm{pen}_W(m)$ around its expectation (faster than $\mathrm{pen}_{\mathrm{id}}(m)$)

I.i.d. framework (Lerasle 2009)

$$\mathrm{pen}_W(m) = C_W \mathbb{E}_W \left[2(P_n^W - \overline{W} P_n)(\widehat{s}_m^W) \right]$$

$\Rightarrow$ $\mathrm{pen}_W(m)$ only depends on W through a multiplicative factor

$\Rightarrow$ $\mathbb{E}\left[\mathrm{pen}_W(m)\right] = C_W \operatorname{var}\left(W_1 - \overline{W}\right) \mathbb{E}\left[\mathrm{pen}_{\mathrm{id}}(m)\right]$
$+$ concentration of $\mathrm{pen}_W(m)$ around its expectation (faster than $\mathrm{pen}_{\mathrm{id}}(m)$)

$\Rightarrow$ oracle inequality with constant $1 + o(1)$ under well-chosen assumptions on P and $\mathcal{M}_n$

Dependent case (Lerasle 2010)

- Mixing (β or τ)

- Split the data into several blocks $\Rightarrow$ keep one every two blocks
- Resample the blocks (which are almost independent)

$\Rightarrow$ Oracle inequality (with an oracle only based on part of the original sample)

Outline

Limits of resampling

- Computational complexity

$\Rightarrow$ alternative: non-exchangeable weights (e.g., V-fold)

- Non-asymptotic results: can we have some without closed-form expressions?