V-fold selection of kernel estimators

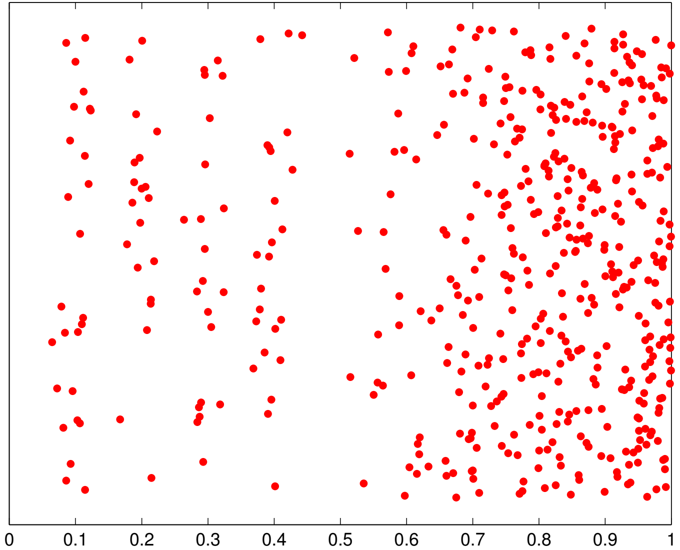
Sylvain Arlot (joint work with Matthieu Lerasle & Nelo Magalhães)

[1] Cnrs

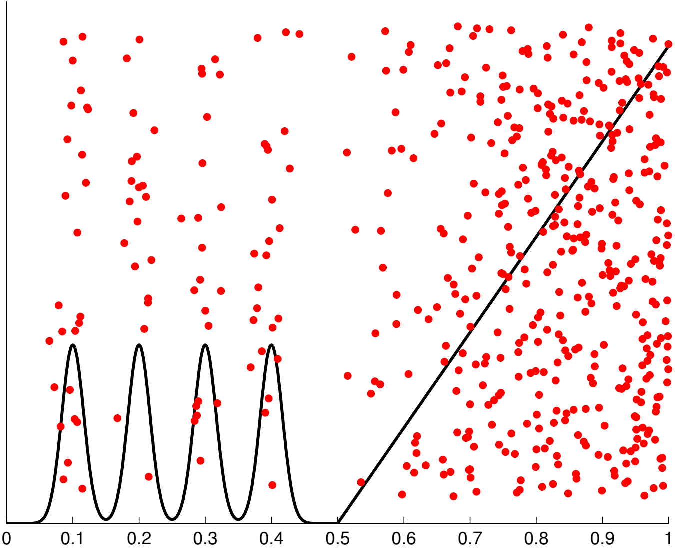
[2] École Normale Supérieure (Paris), DI/ENS, Équipe Sierra

EMS 2015, Amsterdam, July 6th, 2015

Density estimation: data $X_1, \ldots, X_n$



Goal: estimate the common density $f^\star$ of X_i



Problem: density estimation

- Data D_n : $X_1, \ldots, X_n \in \mathbb{X}$ (i.i.d. $\sim P$, density $f^\star$ w.r.t. μ)
  Assumption: $f^\star \in L^\infty$
- Least-squares loss $\gamma(t,x) = \|t\|^2_{L^2(\mu)} - 2t(x)$
- Goal: learn $t \in \mathbb{S} = \{\text{measurable functions } \mathbb{X} \to \mathbb{R}\}$ s.t. $\mathbb{E}_{X\sim P}[\gamma(t;X)] =: P\gamma(t)$ is minimal.

Problem: density estimation

- Data D_n : $X_1, \ldots, X_n \in \mathbb{X}$ (i.i.d. $\sim P$, density $f^\star$ w.r.t. μ)
  Assumption: $f^\star \in L^\infty$

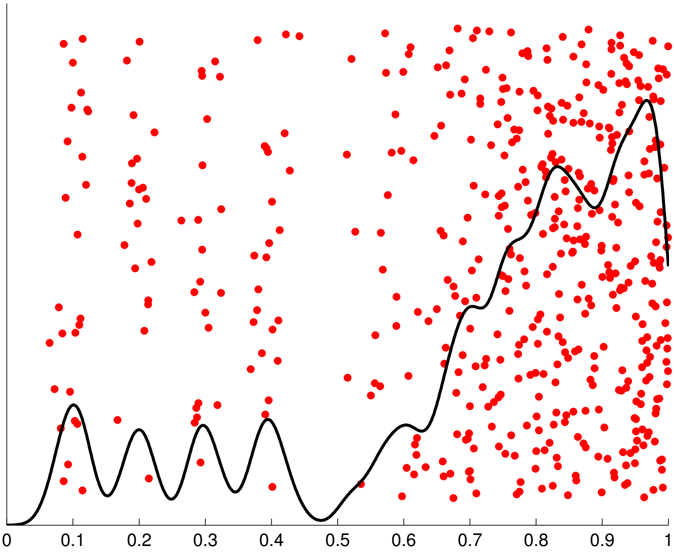
- Least-squares loss $\gamma(t,x) = \|t\|^2_{L^2(\mu)} - 2t(x)$
- Goal: learn $t \in \mathbb{S} = \{$measurable functions $\mathbb{X} \to \mathbb{R}\}$ s.t. $\mathbb{E}_{X\sim P}[\gamma(t;X)] =: P\gamma(t)$ is minimal.

$$P\gamma(t) = \int t^2 \,\mathrm{d}\mu - 2\int t f^\star \,\mathrm{d}\mu = \int (t - f^\star)^2 \,\mathrm{d}\mu - \|f^\star\|^2_{L^2(\mu)}$$

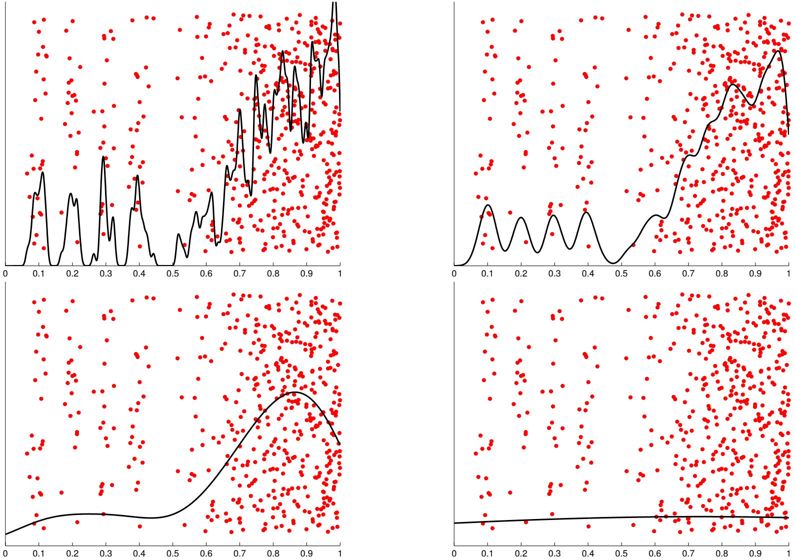
$\Rightarrow$ the true density $f^\star \in \operatorname{argmin}_{t\in\mathbb{S}} P\gamma(t)$ and the excess risk is

$$P\gamma(t) - P\gamma(f^\star) = \|t - f^\star\|^2_{L^2(\mu)}$$

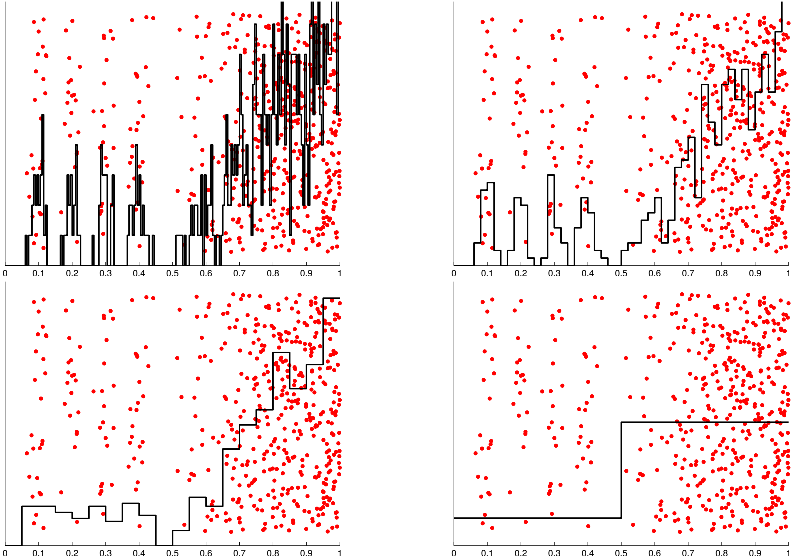
Estimators: example: Parzen, Gaussian kernel



Estimator selection: Parzen, Gaussian kernel



Estimator selection: regular histograms



Selection of linear estimators

- Linear estimator (a.k.a. additive / delta-sequence estimator): $P_n \mapsto \widehat{f}_m$ linear, i.e.,

$$\widehat{f}_m(D_n) : x \mapsto \frac{1}{n} \sum_{i=1}^{n} K_m(x, X_i) \enspace .$$

$K_m : \mathbb{X} \times \mathbb{X} \to \mathbb{R}$ symmetric (the kernel).
Assumptions in this talk: K_m bounded and $\left(K_m(x, \cdot)\right)_{x \in \mathbb{X}}$ bounded in $L^2(\mu)$.

Selection of linear estimators

- Linear estimator (a.k.a. additive / delta-sequence estimator): $P_n \mapsto \widehat{f}_m$ linear, i.e.,

$$\widehat{f}_m(D_n) : x \mapsto \frac{1}{n} \sum_{i=1}^{n} K_m(x, X_i) \enspace .$$

$K_m : \mathbb{X} \times \mathbb{X} \to \mathbb{R}$ symmetric (the kernel).
Assumptions in this talk: K_m bounded and $\left(K_m(x, \cdot)\right)_{x \in \mathbb{X}}$ bounded in $L^2(\mu)$.

- Estimator collection $(\widehat{f}_m)_{m \in \mathcal{M}} \Rightarrow$ choose $\widehat{m} = \widehat{m}(D_n)$?
i.e., family of kernels $(K_m)_{m \in \mathcal{M}} \Rightarrow$ family of linear estimators

Selection of linear estimators

- Linear estimator (a.k.a. additive / delta-sequence estimator): $P_n \mapsto \widehat{f}_m$ linear, i.e.,

$$\widehat{f}_m(D_n) : x \mapsto \frac{1}{n} \sum_{i=1}^{n} K_m(x, X_i) \enspace .$$

  $K_m : \mathbb{X} \times \mathbb{X} \to \mathbb{R}$ symmetric (the kernel).
  Assumptions in this talk: K_m bounded and $\bigl(K_m(x, \cdot)\bigr)_{x \in \mathbb{X}}$ bounded in $L^2(\mu)$.

- Estimator collection $(\widehat{f}_m)_{m \in \mathcal{M}} \Rightarrow$ choose $\widehat{m} = \widehat{m}(D_n)$?
  i.e., family of kernels $(K_m)_{m \in \mathcal{M}} \Rightarrow$ family of linear estimators

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

$$\left\| \widehat{f}_{\widehat{m}} - f^{\star} \right\|^2 \leqslant C \inf_{m \in \mathcal{M}} \left\{ \left\| \widehat{f}_m - f^{\star} \right\|^2 \right\} + R_n$$

Linear estimators: examples

- Projection on $S_m = \mathrm{span}(\psi_\lambda)_{\lambda \in \Lambda_m}$ (orthonormal in $L^2(\mu)$)

$$\widehat{f}_m \in \operatorname*{argmin}_{t \in S_m} \{ P_n \gamma(t) \} \quad \text{where} \quad P_n\gamma(t) := \frac{1}{n} \sum_{x \in D_n} \gamma(t; x)$$

$$K_m(x, y) = \sum_{\lambda \in \Lambda_m} \psi_\lambda(x) \psi_\lambda(y) \ .$$

Linear estimators: examples

- Projection on $S_m = \mathrm{span}(\psi_\lambda)_{\lambda \in \Lambda_m}$ (orthonormal in $L^2(\mu)$)

$$\widehat{f}_m \in \operatorname*{argmin}_{t \in S_m} \{ P_n \gamma(t) \} \quad \text{where} \quad P_n\gamma(t) := \frac{1}{n} \sum_{x \in D_n} \gamma(t; x)$$

$$K_m(x,y) = \sum_{\lambda \in \Lambda_m} \psi_\lambda(x) \psi_\lambda(y) \enspace .$$

- Weighted projection estimator (e.g., Pinsker):

$$K_m(x,y) = \sum_{\lambda \in \Lambda_m} m_\lambda \psi_\lambda(x) \psi_\lambda(y) \qquad \text{with} \qquad m_\lambda \in [0,1] \enspace .$$

Linear estimators: examples

- Projection on $S_m = \mathrm{span}(\psi_\lambda)_{\lambda \in \Lambda_m}$ (orthonormal in $L^2(\mu)$)

$$\widehat{f}_m \in \operatorname*{argmin}_{t \in S_m} \{ P_n \gamma(t) \} \quad \text{where} \quad P_n \gamma(t) := \frac{1}{n} \sum_{x \in D_n} \gamma(t; x)$$

$$K_m(x,y) = \sum_{\lambda \in \Lambda_m} \psi_\lambda(x) \psi_\lambda(y) \ .$$

- Weighted projection estimator (e.g., Pinsker):

$$K_m(x,y) = \sum_{\lambda \in \Lambda_m} m_\lambda \psi_\lambda(x) \psi_\lambda(y) \qquad \text{with} \qquad m_\lambda \in [0,1] \ .$$

- Kernel estimator on $\mathbb{X} = \mathbb{R}$ (or $\mathbb{R}^d$), Parzen-Rosenblatt:

$$K_{(k,h)}(x,y) = \frac{1}{h} k \left(\frac{x-y}{h} \right) \qquad \text{with } k \in L^2(\mu) \text{ symmetric and } h > 0 \ .$$

Example: Gaussian kernel $\qquad k(x) = \frac{1}{\sqrt{2\pi}} \exp\left(\frac{x^2}{2} \right)$

Expectation of the loss

$$\mathbb{E}\left[\left\|\widehat{f}_m - f^\star\right\|^2\right] = \underbrace{\|f_m^\star - f^\star\|^2}_{\text{approximation error}} + \underbrace{\frac{1}{n}\mathbb{E}\left[A_m(X,X) - A_m(X,Y)\right]}_{\text{estimation error}}$$

$$f_m^\star(x) = \mathbb{E}\left[K_m(x,X)\right]$$

$$A_m(x,y) = \int_{\mathbb{X}} K_m(x,z) K_m(y,z) \, \mathrm{d}\mu(z)$$

Expectation of the loss

$$\mathbb{E}\left[\left\|\widehat{f}_m - f^\star\right\|^2\right] = \underbrace{\|f_m^\star - f^\star\|^2}_{\text{approximation error}} + \underbrace{\frac{1}{n}\mathbb{E}\left[A_m(X,X) - A_m(X,Y)\right]}_{\text{estimation error}}$$

$$f_m^\star(x) = \mathbb{E}\left[K_m(x,X)\right]$$

$$A_m(x,y) = \int_{\mathbb{X}} K_m(x,z) K_m(y,z) \, \mathrm{d}\mu(z)$$

- Projection estimators:

$$f_m^\star \in \operatorname*{argmin}_{t \in S_m} \|t - f^\star\|^2 \qquad \text{and} \qquad A_m = K_m \ .$$

Regular histograms on $\mathbb{R}$ with bin size d_m^{-1}: estim. error $\approx \frac{d_m}{n}$

Expectation of the loss

$$\mathbb{E}\left[\left\|\widehat{f}_m - f^\star\right\|^2\right] = \underbrace{\|f_m^\star - f^\star\|^2}_{\text{approximation error}} + \underbrace{\frac{1}{n}\mathbb{E}[A_m(X,X) - A_m(X,Y)]}_{\text{estimation error}}$$

$$f_m^\star(x) = \mathbb{E}[K_m(x,X)]$$

$$A_m(x,y) = \int_{\mathbb{X}} K_m(x,z) K_m(y,z) \, \mathrm{d}\mu(z)$$

- Projection estimators:

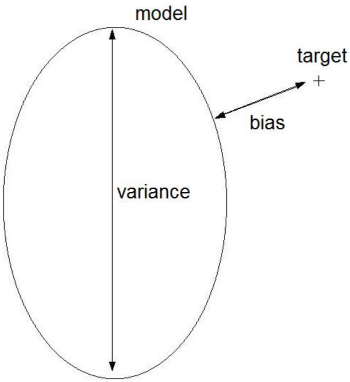
$$f_m^\star \in \operatorname*{argmin}_{t \in S_m} \|t - f^\star\|^2 \qquad \text{and} \qquad A_m = K_m \enspace .$$

Regular histograms on $\mathbb{R}$ with bin size d_m^{-1}: estim. error $\approx \frac{d_m}{n}$

- Parzen: $A_{(k,h)}(x,y) = \frac{1}{h}(k * k)\left(\frac{x-y}{h}\right)$; estim. error $\approx \frac{\|k\|^2_{L^2(\mu)}}{nh}$

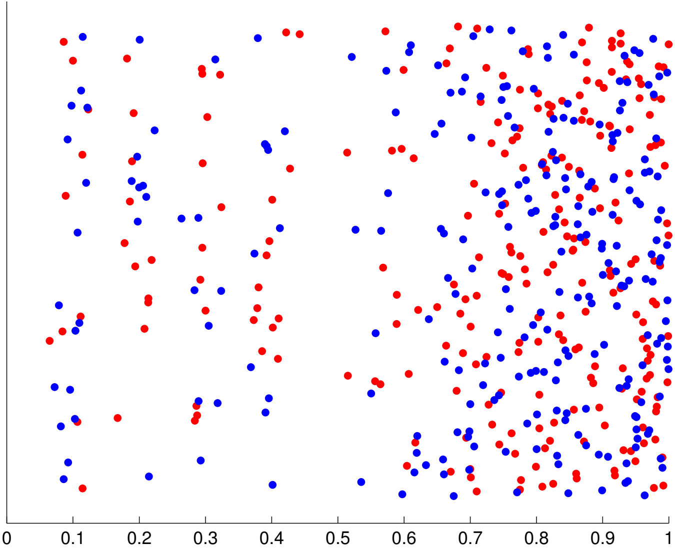
Bias-variance trade-off

$$\mathbb{E}\left[\left\|\widehat{f}_m - f^{\star}\right\|^2\right] = \text{Approx. error} + \text{Estim. error} = \text{Bias} + \text{Variance}$$

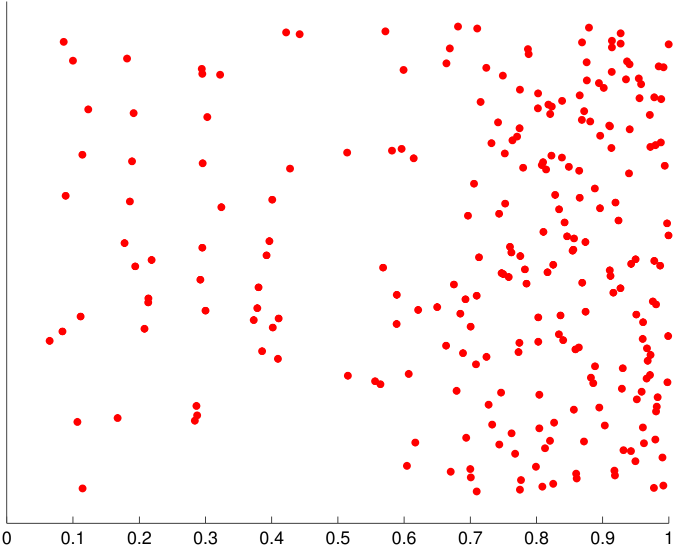


Bias-variance trade-off
$\Leftrightarrow$ avoid overfitting and underfitting

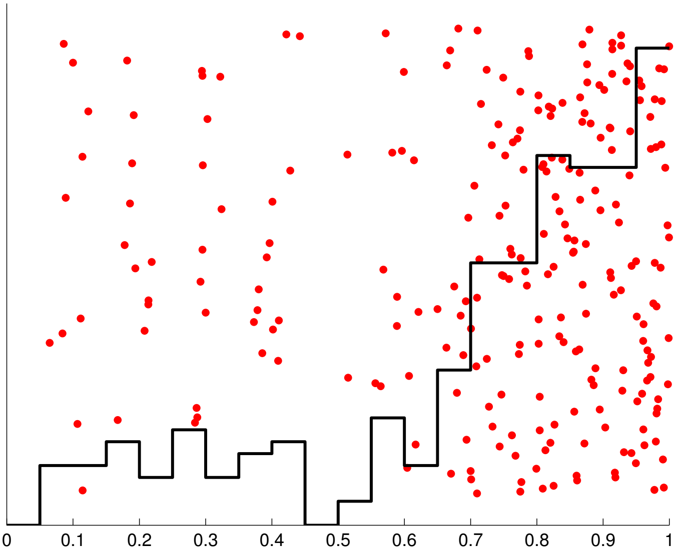
Validation principle



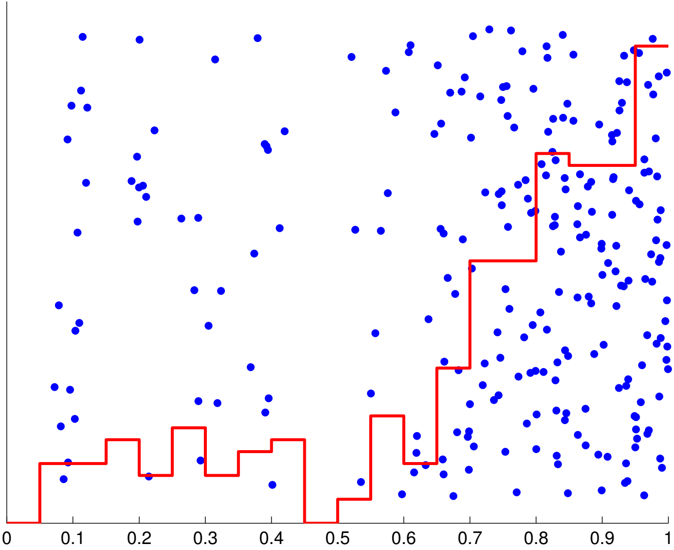
Validation principle: training sample



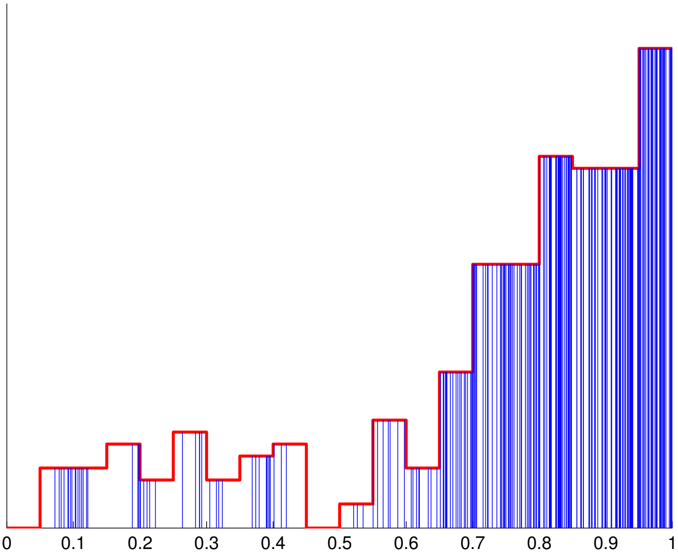
Validation principle: training sample



Validation principle: validation sample



Validation principle: validation sample



V-fold cross-validation

$$\underbrace{X_1, \ldots, X_q}_{\text{Training}}, \underbrace{X_{q+1}, \ldots, X_n}_{\text{Validation}}$$

$$\widehat{f}_m^{(t)} = \widehat{f}_m((X_i)_{1 \leqslant i \leqslant q}) = \frac{1}{q} \sum_{i=1}^{q} K_m(\cdot, X_i)$$

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

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

$$\Rightarrow \widehat{\mathcal{R}}^{\mathrm{vf}} \left(\widehat{f}_m; D_n; \mathcal{B} \right) = \frac{1}{V} \sum_{j=1}^{V} P_n^j \gamma \left(\widehat{f}_m^{(-j)} \right) \qquad \widehat{m} \in \arg\min_{m \in \mathcal{M}} \left\{ \widehat{\mathcal{R}}^{\mathrm{vf}} \left(\widehat{f}_m \right) \right\}$$

Expectation of cross-validation criteria

- In this talk, we always assume $\mathrm{Card}(B_j) = n/V$ for all j.

Expectation of cross-validation criteria

- In this talk, we always assume $\operatorname{Card}(B_j) = n/V$ for all j.
- Ideal criterion: $P\gamma(\widehat{f}_m)$
- General analysis for the bias:

$$\mathbb{E}\left[P\gamma\left(\widehat{f}_m(D_n)\right)\right] = \alpha(m) + \frac{\beta(m)}{n}$$

$$\Rightarrow \quad \mathbb{E}\left[\widehat{\mathcal{R}}^{\mathrm{vf}}\left(\widehat{f}_m; D_n; \mathcal{B}\right)\right] = \mathbb{E}\left[P_n^{(j)}\gamma\left(\widehat{f}_m^{(-j)}\right)\right] = \mathbb{E}\left[P\gamma\left(\widehat{f}_m^{(-j)}\right)\right] = \alpha(m) + \frac{V}{V-1}\frac{\beta(m)}{n}$$

Expectation of cross-validation criteria

- In this talk, we always assume $\operatorname{Card}(B_j) = n/V$ for all j.
- Ideal criterion: $P\gamma(\widehat{f}_m)$
- General analysis for the bias:

$$
\begin{aligned}
& \mathbb{E}\left[P\gamma\left(\widehat{f}_m(D_n)\right)\right] = \alpha(m) + \frac{\beta(m)}{n} \\
\Rightarrow \quad & \mathbb{E}\left[\widehat{\mathcal{R}}^{\mathrm{vf}}\left(\widehat{f}_m; D_n; \mathcal{B}\right)\right] = \mathbb{E}\left[P_n^{(j)}\gamma\left(\widehat{f}_m^{(-j)}\right)\right] = \mathbb{E}\left[P\gamma\left(\widehat{f}_m^{(-j)}\right)\right] \\
& = \alpha(m) + \frac{V}{V-1}\frac{\beta(m)}{n}
\end{aligned}
$$

$\Rightarrow$ bias decreases with V, vanishes as $V \to \infty$

Expectation of cross-validation criteria

- In this talk, we always assume $\mathrm{Card}(B_j) = n/V$ for all j.
- Ideal criterion: $P\gamma(\widehat{f}_m)$
- General analysis for the bias:

$$\mathbb{E}\left[P\gamma\left(\widehat{f}_m(D_n)\right)\right] = \alpha(m) + \frac{\beta(m)}{n}$$

$$\Rightarrow \quad \mathbb{E}\left[\widehat{\mathcal{R}}^{\mathrm{vf}}\left(\widehat{f}_m; D_n; \mathcal{B}\right)\right] = \mathbb{E}\left[P_n^{(j)}\gamma\left(\widehat{f}_m^{(-j)}\right)\right] = \mathbb{E}\left[P\gamma\left(\widehat{f}_m^{(-j)}\right)\right] = \alpha(m) + \frac{V}{V-1}\frac{\beta(m)}{n}$$

$\Rightarrow$ bias decreases with V, vanishes as $V \to \infty$

- Same result for the leave-p-out with V replaced by n/p.

Bias and model selection

$$\widehat{m} \in \operatorname*{argmin}_{m \in \mathcal{M}} \left\{ \widehat{\mathcal{R}}^{\mathrm{vf}} \left(\widehat{f}_m \right) \right\} \quad \text{vs.} \quad m^{\star} \in \operatorname*{argmin}_{m \in \mathcal{M}} \left\{ P\gamma \left(\widehat{f}_m(D_n) \right) \right\}$$

- Perfect ranking among $(\widehat{f}_m)_{m \in \mathcal{M}} \Leftrightarrow \forall m, m' \in \mathcal{M}$,

$$\operatorname{sign} \left(\widehat{\mathcal{R}}^{\mathrm{vf}} \left(\widehat{f}_m \right) - \widehat{\mathcal{R}}^{\mathrm{vf}} \left(\widehat{f}_{m'} \right) \right) = \operatorname{sign} \left(P\gamma \left(\widehat{f}_m \right) - P\gamma \left(\widehat{f}_{m'} \right) \right)$$

Bias and model selection

$$\widehat{m} \in \operatorname*{argmin}_{m \in \mathcal{M}} \left\{ \widehat{\mathcal{R}}^{\mathrm{vf}}\left(\widehat{f}_m\right) \right\} \quad \text{vs.} \quad m^{\star} \in \operatorname*{argmin}_{m \in \mathcal{M}} \left\{ P\gamma\left(\widehat{f}_m(D_n)\right) \right\}$$

- Perfect ranking among $(\widehat{f}_m)_{m \in \mathcal{M}} \Leftrightarrow \forall m, m' \in \mathcal{M}$,

$$\operatorname{sign}\left(\widehat{\mathcal{R}}^{\mathrm{vf}}\left(\widehat{f}_m\right) - \widehat{\mathcal{R}}^{\mathrm{vf}}\left(\widehat{f}_{m'}\right)\right) = \operatorname{sign}\left(P\gamma\left(\widehat{f}_m\right) - P\gamma\left(\widehat{f}_{m'}\right)\right)$$

- Key quantities:

$$\mathbb{E}\left[P\gamma\left(\widehat{f}_m\right) - P\gamma\left(\widehat{f}_{m'}\right)\right] = \alpha(m) - \alpha(m') + \frac{\beta(m) - \beta(m')}{n}$$

$$\mathbb{E}\left[\widehat{\mathcal{R}}^{\mathrm{vf}}\left(\widehat{f}_m\right) - \widehat{\mathcal{R}}^{\mathrm{vf}}\left(\widehat{f}_{m'}\right)\right] = \alpha(m) - \alpha(m') + \frac{V}{V-1}\frac{\beta(m) - \beta(m')}{n}$$

- V-fold CV favours m with smaller complexity $\beta(m)$

Bias-corrected VFCV / V-fold penalization

- Bias-corrected V-fold CV (Burman, 1989):

$$\widehat{\mathcal{R}}^{\mathrm{vf,corr}}\left(\widehat{f}_m; D_n; \mathcal{B}\right) := \widehat{\mathcal{R}}^{\mathrm{vf}}\left(\widehat{f}_m; D_n; \mathcal{B}\right) + P_n\gamma\left(\widehat{f}_m\right) - \frac{1}{V}\sum_{j=1}^{V} P_n\gamma\left(\widehat{f}_m^{(-j)}\right)$$

- Resampling heuristics (Efron, 1983), V-fold subsampling and penalization principle ⇒ V-fold penalty (A. 2008)

$$\mathrm{pen}_{\mathrm{VF}}\left(\widehat{f}_m; D_n; \mathcal{B}\right) := \frac{V-1}{V}\sum_{j=1}^{V}\left(P_n - P_n^{(-j)}\right)\gamma\left(\widehat{f}_m^{(-j)}\right)$$

Bias-corrected VFCV / V-fold penalization

- Bias-corrected V-fold CV (Burman, 1989):

$$\widehat{\mathcal{R}}^{\mathrm{vf,corr}}\left(\widehat{f}_m; D_n; \mathcal{B}\right) := \widehat{\mathcal{R}}^{\mathrm{vf}}\left(\widehat{f}_m; D_n; \mathcal{B}\right) + P_n\gamma\left(\widehat{f}_m\right) - \frac{1}{V}\sum_{j=1}^{V} P_n\gamma\left(\widehat{f}_m^{(-j)}\right)$$

- Resampling heuristics (Efron, 1983), V-fold subsampling and penalization principle $\Rightarrow$ V-fold penalty (A. 2008)

$$\mathrm{pen}_{\mathrm{VF}}\left(\widehat{f}_m; D_n; \mathcal{B}\right) := \frac{V-1}{V}\sum_{j=1}^{V}\left(P_n - P_n^{(-j)}\right)\gamma\left(\widehat{f}_m^{(-j)}\right)$$

- $\widehat{\mathcal{R}}^{\mathrm{vf,corr}}\left(\widehat{f}_m; D_n; \mathcal{B}\right) = P_n\gamma\left(\widehat{f}_m(D_n)\right) + \mathrm{pen}_{\mathrm{VF}}\left(\widehat{f}_m; D_n; \mathcal{B}\right)$
- Projection estimators:
$\widehat{\mathcal{R}}^{\mathrm{vf}}\left(\widehat{f}_m; D_n; \mathcal{B}\right) = P_n\gamma\left(\widehat{f}_m(D_n)\right) + \frac{V-1/2}{V-1}\mathrm{pen}_{\mathrm{VF}}\left(\widehat{f}_m; D_n; \mathcal{B}\right)$

Expectations

Bias-corrected V-fold CV:

$$\mathbb{E}\left[\widehat{\mathcal{R}}^{\mathrm{vf,corr}}\left(\widehat{f}_m; D_n; \mathcal{B}\right)\right]$$
$$= \mathbb{E}\left[P\gamma\left(\widehat{f}_m(D_n)\right)\right]$$
$$= \left\|f_m^\star - f^\star\right\|^2 + \frac{1}{n}\mathbb{E}[A_m(X,X) - A_m(X,Y)] - \left\|f^\star\right\|^2$$

Expectations

Bias-corrected V-fold CV:

$$\mathbb{E}\left[\widehat{\mathcal{R}}^{\mathrm{vf,corr}}\left(\widehat{f}_m; D_n; \mathcal{B}\right)\right]$$

$$= \mathbb{E}\left[P\gamma\left(\widehat{f}_m(D_n)\right)\right]$$

$$= \left\| f_m^{\star} - f^{\star} \right\|^2 + \frac{1}{n}\mathbb{E}[A_m(X,X) - A_m(X,Y)] - \left\| f^{\star} \right\|^2$$

V-fold penalization:

$$\mathbb{E}\left[P_n\gamma\left(\widehat{f}_m(D_n)\right) + C\,\mathrm{pen}_{\mathrm{VF}}(\widehat{f}_m; D_n; \mathcal{B})\right]$$

$$= \mathbb{E}\left[P\gamma(\widehat{f}_m(D_n))\right] + \frac{C-1}{n}\mathbb{E}[K_m(X,X) - K_m(X,Y)]$$

Oracle inequality for bias-corrected V-fold

Theorem (A., Lerasle & Magalhães 2015)

Under some "reasonable" assumptions, with probability at least $1 - e^{-x}$, for all $\varepsilon \in (0,1)$, for any
$\widehat{m} \in \operatorname{argmin}_{m \in \mathcal{M}} \{ P_n \gamma \left(\widehat{f}_m(D_n) \right) + \mathrm{pen}_{\mathrm{VF}}(\widehat{f}_m; D_n; \mathcal{B}) \}$,

$$\left\| \widehat{f}_{\widehat{m}} - f^\star \right\|^2 \leqslant \frac{1+\varepsilon}{1-\varepsilon} \inf_{m \in \mathcal{M}} \left\{ \left\| \widehat{f}_m - f^\star \right\|^2 \right\} + \frac{L \left(\log \mathrm{Card}(\mathcal{M}) + x \right)^2}{(1-\varepsilon)\varepsilon^3 \, n} \enspace .$$

Related result: A. & Lerasle (2012) for projection estimators.

Oracle inequality for CV procedures

Theorem (A., Lerasle & Magalhães 2015)

Under some "reasonable" assumptions, with probability at least $1 - e^{-x}$, for all $\varepsilon > 0$, for any $\widehat{m}$ selected by V-fold penalization ($C \times \mathrm{pen}_{\mathrm{VF}}$), V-fold CV or Leave-p-out, for any $m \in \mathcal{M}$,

$$\frac{1 - \delta^-_{\widehat{m}}}{1 + \delta^+_m} \left\| \widehat{f}_{\widehat{m}} - f^\star \right\|^2 \leqslant \inf_{m \in \mathcal{M}} \left\{ \left\| \widehat{f}_m - f^\star \right\|^2 \right\} + R_{n,x,\varepsilon}$$

where $\gamma_m := \frac{\mathbb{E}[K_m(X,X)]}{A_m(X,X)}$, $R_{n,x,\varepsilon} = \frac{L\,(C^2 \vee 1)(\log \mathrm{Card}(\mathcal{M}) + x)^2}{\varepsilon^3 \, n}$ *and*

	$C \times \mathrm{pen}_{\mathrm{VF}}$	VFCV	LPO
$\delta^+_m =$	$2(C-1)_+ \gamma_m + \varepsilon$	$2/(V-1) + \varepsilon$	$2p/(n-p) + \varepsilon$
$\delta^-_{\widehat{m}} =$	$2(C-1)_- \gamma_{\widehat{m}} + \varepsilon$	ε	ε

Related results: van der Laan, Dudoit & Keles (2004); Celisse (2014) for the leave-p-out; A. & Lerasle (2012).

Assumptions

General assumption set for linear estimators, holds true for instance if, for some numerical constant $\kappa > 0$,

- projection estimators $(\gamma_m = 1)$:

$$\kappa L \geqslant 1 \vee \frac{\left\| \sum_{\lambda \in \Lambda_m} \psi_\lambda^2 \right\|_\infty}{n} \vee \left\| f_m^\star \right\|_\infty \quad .$$

Assumptions

General assumption set for linear estimators, holds true for instance if, for some numerical constant $\kappa > 0$,

- projection estimators ($\gamma_m = 1$):

$$\kappa L \geqslant 1 \vee \frac{\left\| \sum_{\lambda \in \Lambda_m} \psi_\lambda^2 \right\|_\infty}{n} \vee \left\| f_m^\star \right\|_\infty \quad .$$

- kernel (Parzen) estimators ($\gamma_m = k(0) / \left\| k \right\|^2$):

$$h \geqslant \frac{\left\| k \right\|_{L^2(\mu)}^2 \vee \left\| k \right\|_\infty}{n} \qquad \text{and} \qquad \kappa L \geqslant 1 \vee \frac{k(0)}{\left\| k \right\|_{L^2(\mu)}^2} \vee \left\| f^\star \right\|_\infty \left\| k \right\|_{L^1(\mu)}^2 \quad .$$

Assumptions

General assumption set for linear estimators, holds true for instance if, for some numerical constant $\kappa > 0$,

- projection estimators ($\gamma_m = 1$):

$$\kappa L \geqslant 1 \vee \frac{\left\| \sum_{\lambda \in \Lambda_m} \psi_\lambda^2 \right\|_\infty}{n} \vee \left\| f_m^\star \right\|_\infty \quad .$$

- kernel (Parzen) estimators ($\gamma_m = k(0) / \left\| k \right\|^2$):

$$h \geqslant \frac{\left\| k \right\|_{L^2(\mu)}^2 \vee \left\| k \right\|_\infty}{n} \qquad \text{and} \qquad \kappa L \geqslant 1 \vee \frac{k(0)}{\left\| k \right\|_{L^2(\mu)}^2} \vee \left\| f^\star \right\|_\infty \left\| k \right\|_{L^1(\mu)}^2 \quad .$$

Also possible to get sharp minimax Sobolev adaptivity with V-fold calibrated Pinsker estimators.

Variance of the (corrected)-VFCV criterion

- Exact non-asymptotic formula for

$$\mathrm{var}(\mathcal{C}(m)) \qquad \text{and} \qquad \mathrm{var}(\mathcal{C}(m) - \mathcal{C}(m'))$$

with

$$\mathcal{C}(m) \in \left\{ \widehat{\mathcal{R}}^{\mathrm{vf,corr}} \left(\widehat{f}_m; D_n; \mathcal{B} \right), \widehat{\mathcal{R}}^{\mathrm{vf}} \left(\widehat{f}_m; D_n; \mathcal{B} \right), \widehat{\mathcal{R}}^{\ell\mathrm{po}} \left(\widehat{f}_m; D_n \right) \right\} \quad .$$

Variance of the (corrected)-VFCV criterion

- Exact non-asymptotic formula for

$$\operatorname{var}(\mathcal{C}(m)) \qquad \text{and} \qquad \operatorname{var}(\mathcal{C}(m) - \mathcal{C}(m'))$$

with

$$\mathcal{C}(m) \in \left\{ \widehat{\mathcal{R}}^{\mathrm{vf,corr}} \left(\widehat{f}_m; D_n; \mathcal{B} \right), \widehat{\mathcal{R}}^{\mathrm{vf}} \left(\widehat{f}_m; D_n; \mathcal{B} \right), \widehat{\mathcal{R}}^{\ell\mathrm{po}} \left(\widehat{f}_m; D_n \right) \right\} \quad .$$

- For bias-corrected V-fold (and V-fold penalization), the variance always decreases with V.

Variance of the (corrected)-VFCV criterion

- Exact non-asymptotic formula for

$$\operatorname{var}(\mathcal{C}(m)) \qquad \text{and} \qquad \operatorname{var}(\mathcal{C}(m) - \mathcal{C}(m'))$$

with

$$\mathcal{C}(m) \in \left\{ \widehat{\mathcal{R}}^{\mathrm{vf,corr}}\left(\widehat{f}_m; D_n; \mathcal{B}\right), \widehat{\mathcal{R}}^{\mathrm{vf}}\left(\widehat{f}_m; D_n; \mathcal{B}\right), \widehat{\mathcal{R}}^{\ell\mathrm{po}}\left(\widehat{f}_m; D_n\right) \right\} \quad .$$

- For bias-corrected V-fold (and V-fold penalization), the variance always decreases with V.
- For projection estimators (A. & Lerasle 2012), the variance decreases by a constant factor (at most).

Conclusion

- Non-asymptotic oracle inequality for V-fold cross-validation and the leave-p-out, for any V, p.
  Optimal up to a constant factor (bias).

Conclusion

- Non-asymptotic oracle inequality for V-fold cross-validation and the leave-p-out, for any V, p.
  Optimal up to a constant factor (bias).
- Bias-corrected V-fold (or V-fold penalization) $\Rightarrow$ first-order optimal non-asymptotic oracle inequality, for any V.

Conclusion

- Non-asymptotic oracle inequality for V-fold cross-validation and the leave-p-out, for any V, p.
  Optimal up to a constant factor (bias).
- Bias-corrected V-fold (or V-fold penalization) $\Rightarrow$ first-order optimal non-asymptotic oracle inequality, for any V.
- Result holds for various settings, such as: model selection, bandwidth and kernel choice (Parzen), Pinsker estimators, mix of various kinds of these.

Conclusion

- Non-asymptotic oracle inequality for V-fold cross-validation and the leave-p-out, for any V, p.
  Optimal up to a constant factor (bias).
- Bias-corrected V-fold (or V-fold penalization) $\Rightarrow$ first-order optimal non-asymptotic oracle inequality, for any V.
- Result holds for various settings, such as: model selection, bandwidth and kernel choice (Parzen), Pinsker estimators, mix of various kinds of these.
- Choice of V for V-fold?

$\Rightarrow$ must take into account second-order terms:

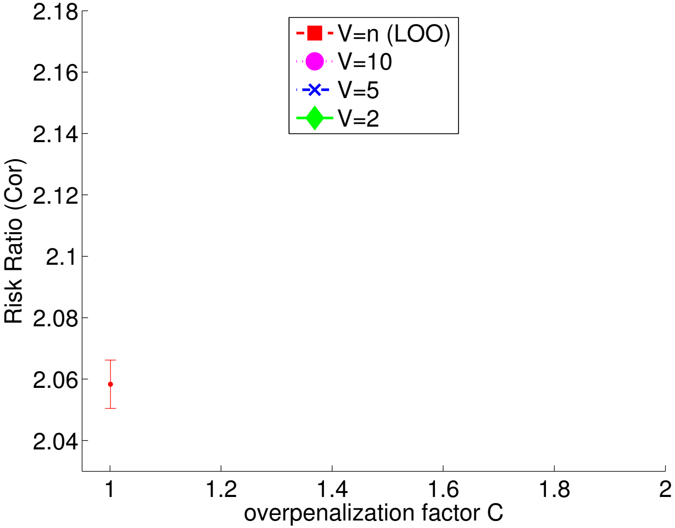
- variance: decreases with V
- small bias can be benefic (open problem).

https://tel.archives-ouvertes.fr/tel-01164581 Chap. 3–4

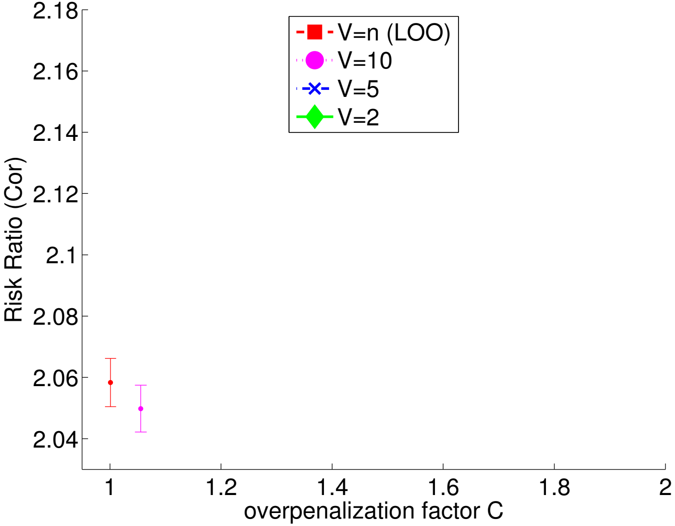
Part I

Appendix

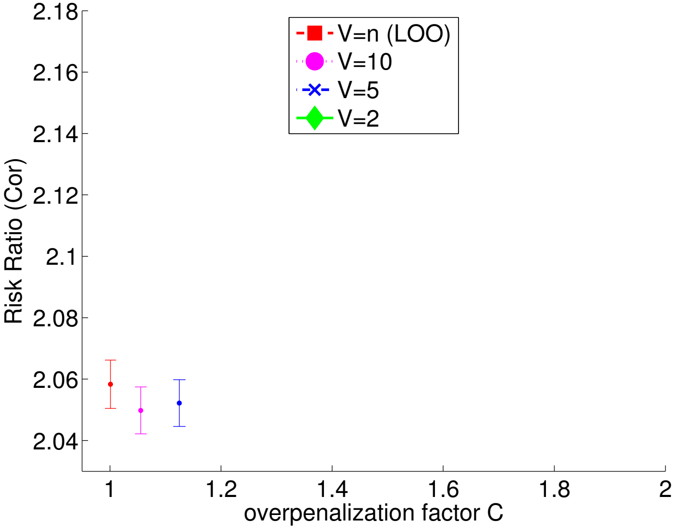
Experiment: V-fold CV



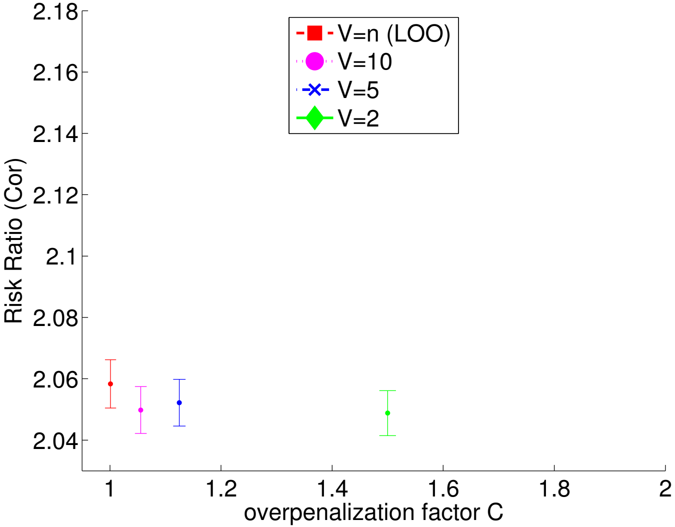
Experiment: V-fold CV



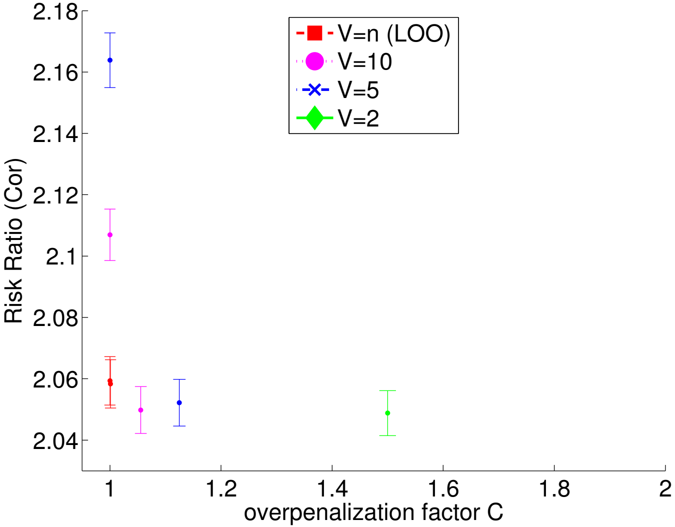
Experiment: V-fold CV



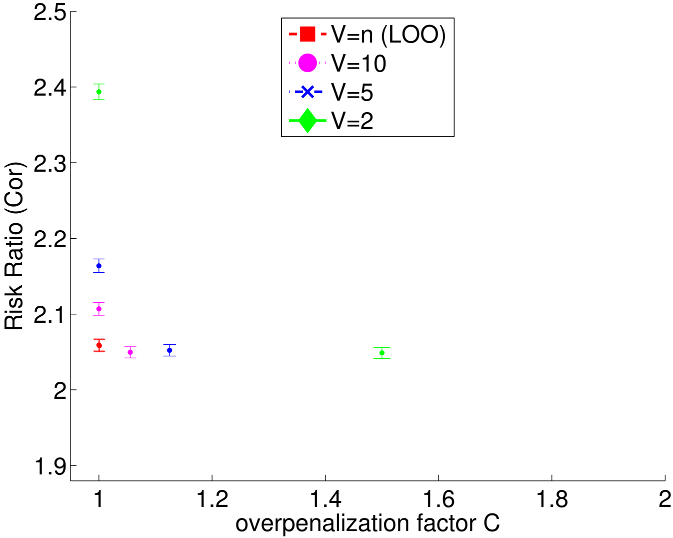
Experiment: V-fold CV



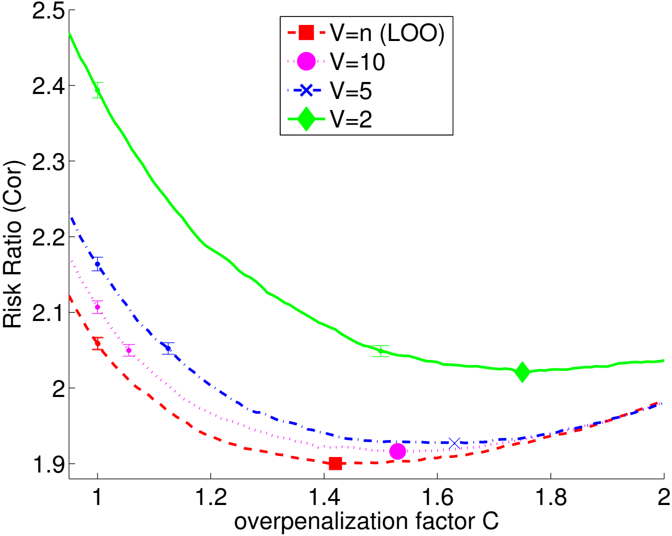
Experiment: V-fold penalization



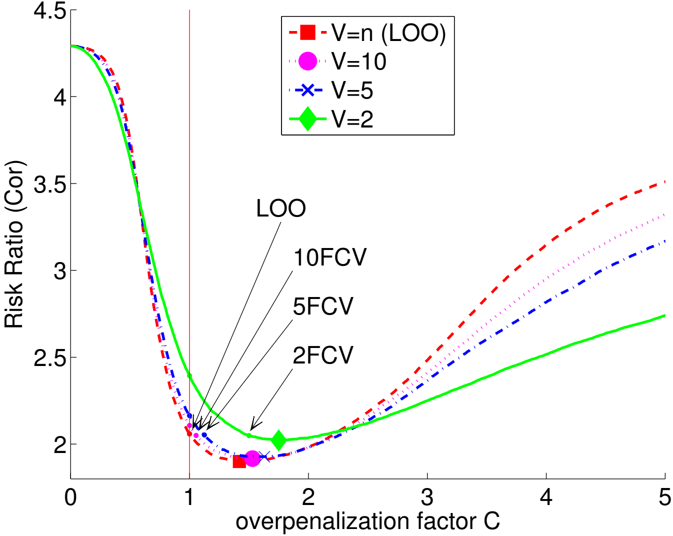
Experiment: V-fold penalization



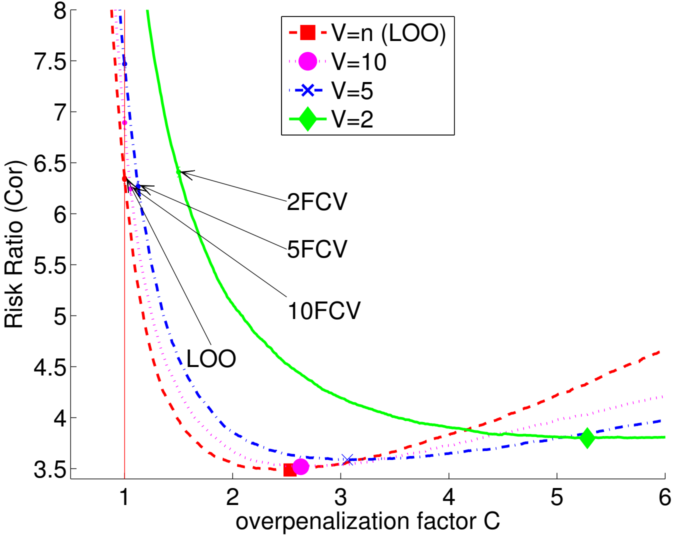
Experiment: overpenalization



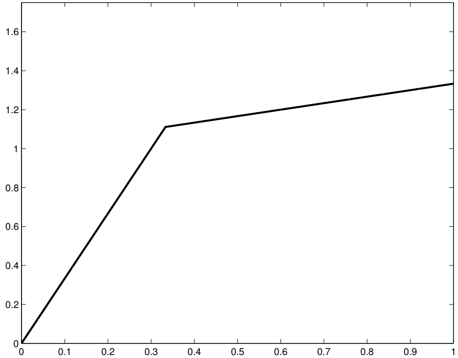
Experiment: conclusion (setting S)



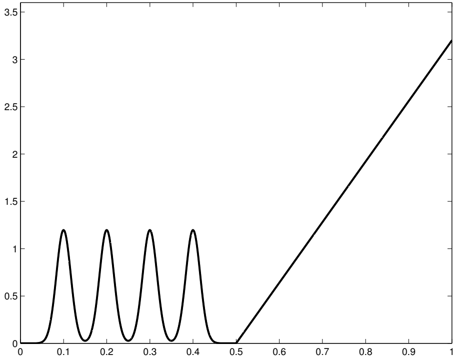
Experiment: "practically parametric" setting (L)



Simulation setting: densities

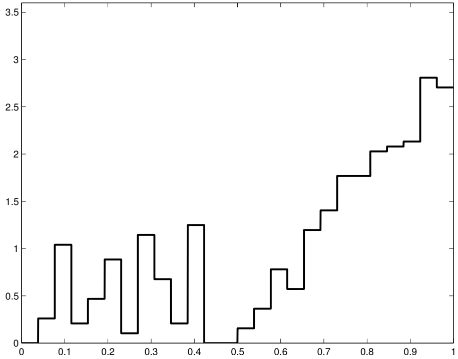


L

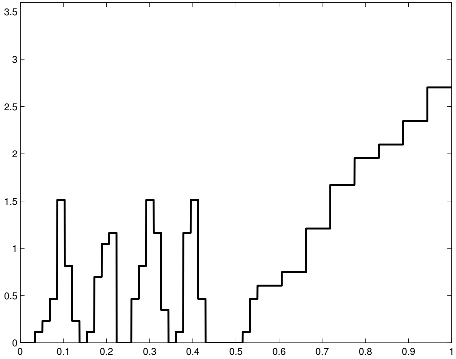


S

Simulation setting: model families



Regu



Dya2