Sélection d'estimateurs par validation croisée

Sylvain Arlot (collaborations avec Alain Celisse, Matthieu Lerasle, Nelo Magalhães)

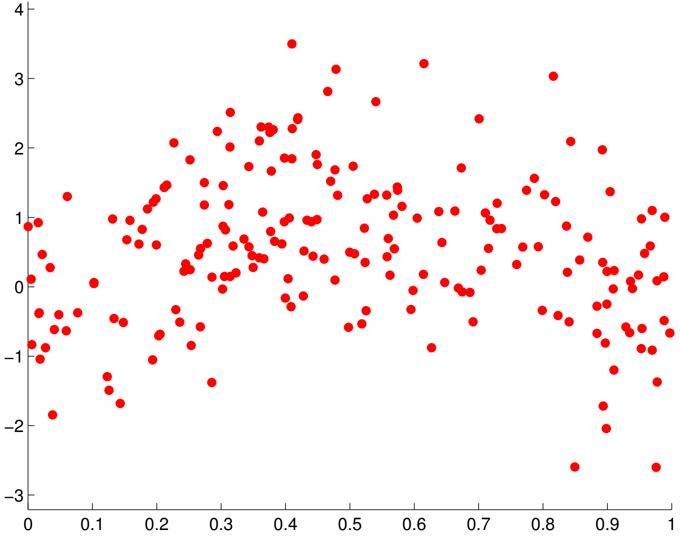
Université Paris-Sud

Colloquium du MAP5, Paris
13 novembre 2015

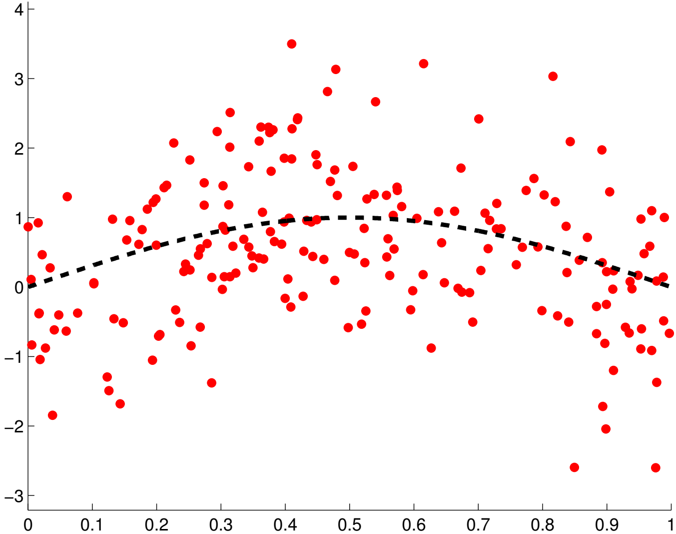
Référence principale (article de survol): arXiv:0907.4728

Outline

Regression: data $(X_1, Y_1), \ldots, (X_n, Y_n)$



Goal: predict Y given X, i.e., denoising



Prediction problem / regression

- Data D_n : $(X_1, Y_1), \ldots, (X_n, Y_n) \in \mathcal{X} \times \mathcal{Y}$ (i.i.d. $\sim P$)
- Contrast $\gamma(t; (x, y))$ measures how well $t(x)$ "predicts" y
- Goal: learn $t \in \mathbb{S} = \{$ measurable functions $\mathcal{X} \to \mathcal{Y}\}$ s.t. $\mathbb{E}_{(X,Y)\sim P}[\gamma(t; (X, Y))] =: P\gamma(t)$ is minimal.

Prediction problem / regression

- Data D_n: $(X_1, Y_1), \ldots, (X_n, Y_n) \in \mathcal{X} \times \mathcal{Y}$ (i.i.d. $\sim P$)

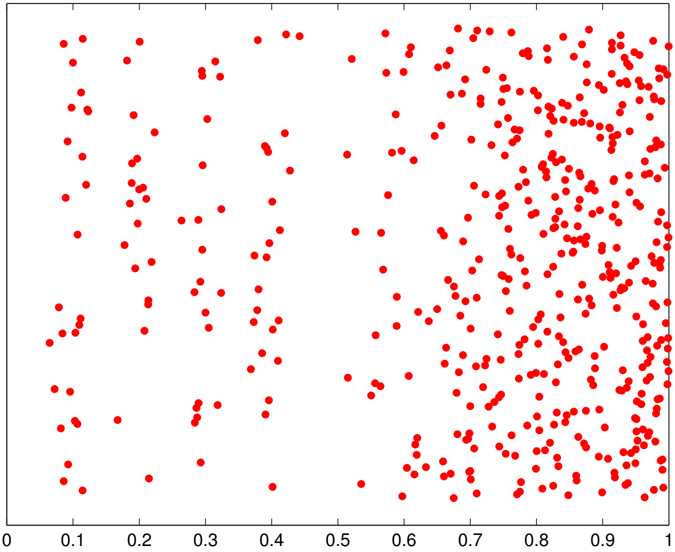
- Contrast $\gamma(t; (x,y))$ measures how well $t(x)$ "predicts" y
- Goal: learn $t \in \mathbb{S} = \{$ measurable functions $\mathcal{X} \to \mathcal{Y}\}$ s.t. $\mathbb{E}_{(X,Y)\sim P}\left[\gamma(t;(X,Y))\right] =: P\gamma(t)$ is minimal.

- Example: regression $\mathcal{Y} = \mathbb{R}$,
  least-squares contrast $\gamma(t;(x,y)) = (t(x) - y)^2$
  $s^\star \in \operatorname{argmin}_{t \in \mathbb{S}} P\gamma(t)$ is the regression function:
  $s^\star(X) = \mathbb{E}[Y \mid X]$

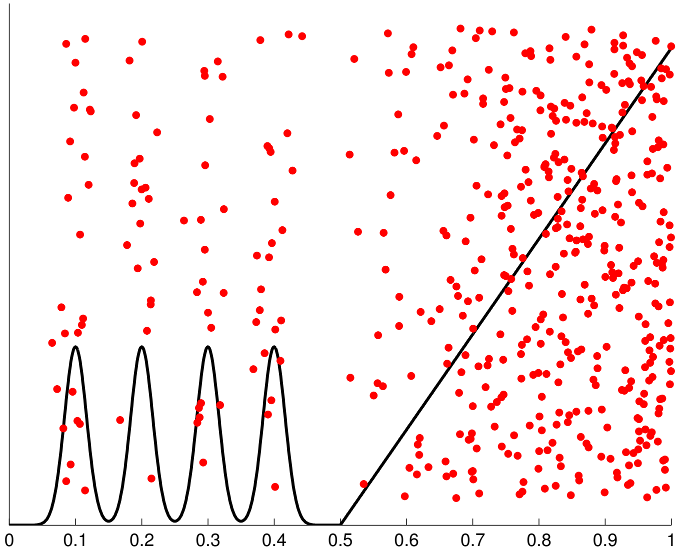
$\Rightarrow$ excess loss

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

Density estimation: data $\xi_1, \ldots, \xi_n$



Goal: estimate the common density $s^\star$ of ξ_i



Problem: density estimation

- Data D_n : $\xi_1, \ldots, \xi_n \in \Xi$ (i.i.d. $\sim P$, density $s^\star$ w.r.t. μ)
- Least-squares contrast $\gamma(t, \xi) = \|t\|^2_{L^2(\mu)} - 2t(\xi)$
- Goal: learn $t \in \mathbb{S} = \{$ measurable functions $\Xi \to \mathbb{R}\}$ s.t. $\mathbb{E}_{\xi \sim P}[\gamma(t;\xi)] =: P\gamma(t)$ is minimal.

Problem: density estimation

- Data D_n : $\xi_1, \ldots, \xi_n \in \Xi$ (i.i.d. $\sim P$, density $s^\star$ w.r.t. μ)
- Least-squares contrast $\gamma(t, \xi) = \|t\|_{L^2(\mu)}^2 - 2t(\xi)$
- Goal: learn $t \in \mathbb{S} = \{$ measurable functions $\Xi \to \mathbb{R}\}$ s.t. $\mathbb{E}_{\xi \sim P}[\gamma(t;\xi)] =: P\gamma(t)$ is minimal.

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

$\Rightarrow$ the true density $s^\star \in \mathrm{argmin}_{t \in \mathbb{S}} P\gamma(t)$ and the excess loss is

$$\ell(s^\star, t) := P\gamma(t) - P\gamma(s^\star) = \|t - s^\star\|_{L^2(\mu)}^2$$

General setting

- Data $\xi_1, \ldots, \xi_n \in \Xi$ i.i.d. with distribution P

  prediction: $\xi_i = (X_i, Y_i) \in \mathcal{X} \times \mathcal{Y}$

- Goal: Estimate some feature $s^\star \in \mathbb{S}$ of P

  density, regression function, Bayes predictor...

- Contrast function $\gamma : \mathbb{S} \times \Xi \to \mathbb{R}$ such that

$$s^\star \in \operatorname*{argmin}_{t \in \mathbb{S}} \left\{ P\gamma(t) \right\} \qquad \text{with} \qquad P\gamma(t) := \mathbb{E}_{\xi \sim P}\left[\gamma(t;\xi)\right]$$

- Excess loss

$$\ell(s^\star, t) := P\gamma(t) - P\gamma(s^\star) \geqslant 0$$

Examples

- Prediction: $\xi_i = (X_i, Y_i)$
  $X_{n+1} \quad \rightsquigarrow \quad$ "predict" Y_{n+1} with $t(X_{n+1})$?
  $\gamma(t;(x,y))$ quantifies the "distance" between $t(x)$ and y
- Regression ($\mathcal{Y} = \mathbb{R}$), least squares:

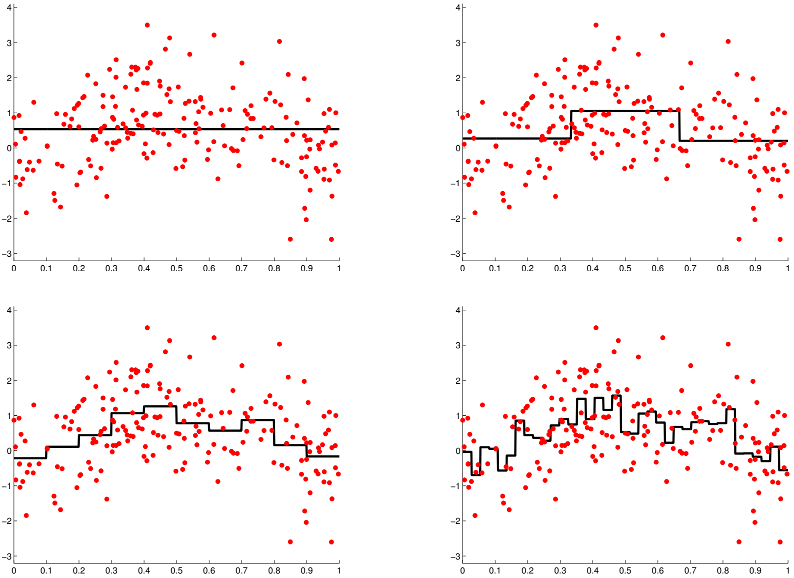
$$\gamma(t;(x,y)) = (t(x) - y)^2 \qquad s^\star(X) = \mathbb{E}[Y|X]$$

- Binary classification ($\mathcal{Y} = \{0,1\}$), 0–1 contrast:

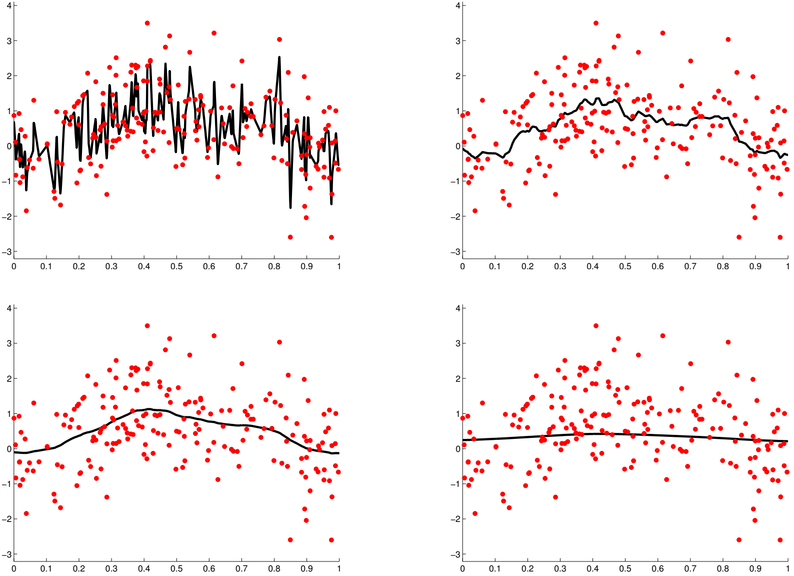
$$\gamma(t;(x,y)) = 1_{t(x) \neq y}$$

- Density estimation (reference measure μ):
  least squares: $\gamma(t;\xi) = \|t\|^2_{L^2(\mu)} - 2t(\xi)$
  log-likelihood: $\gamma(t;\xi) = -\log(t(\xi))$

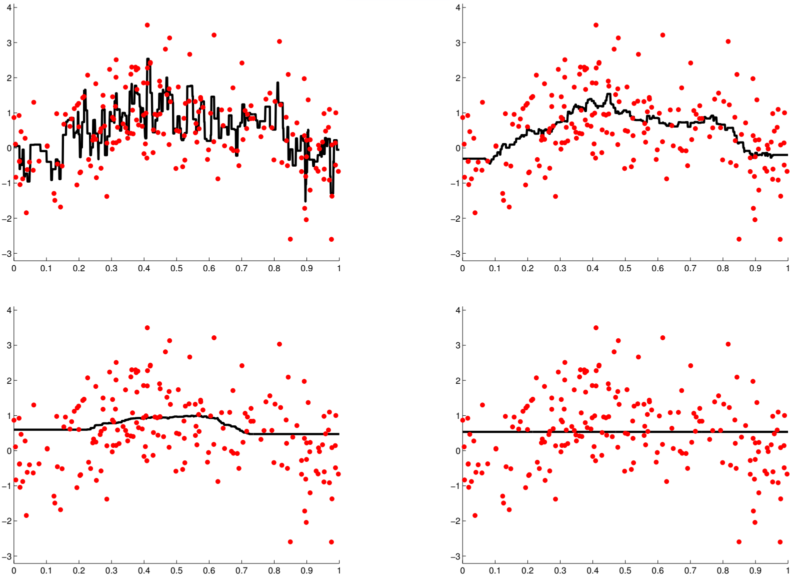
Estimator selection (regression): regular regressograms



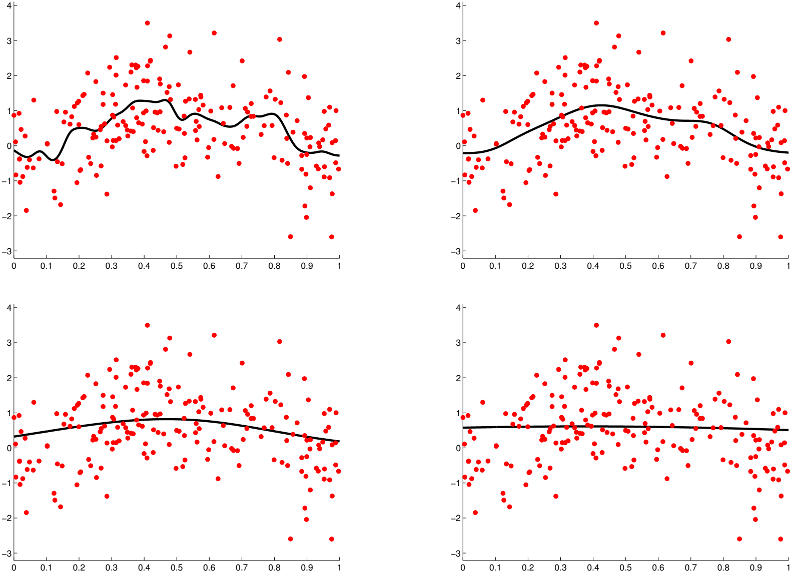
Estimator selection (regression): kernel ridge



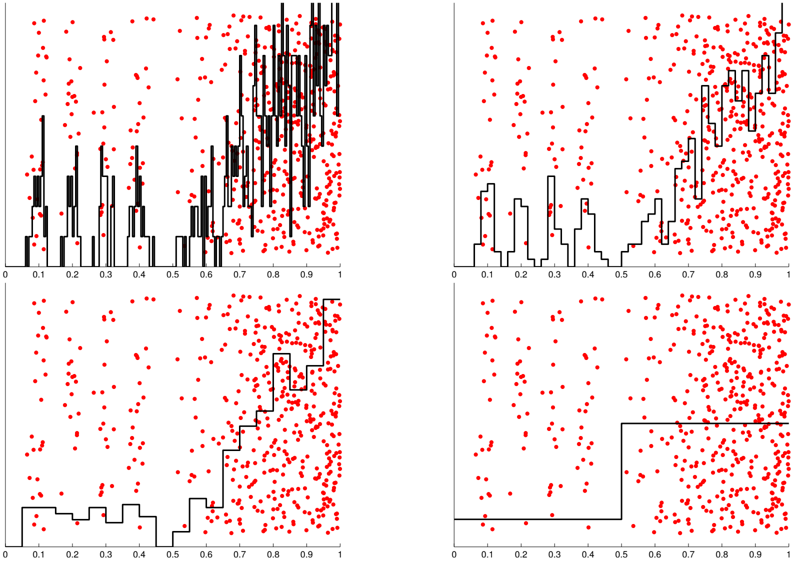
Estimator selection (regression): k nearest neighbours



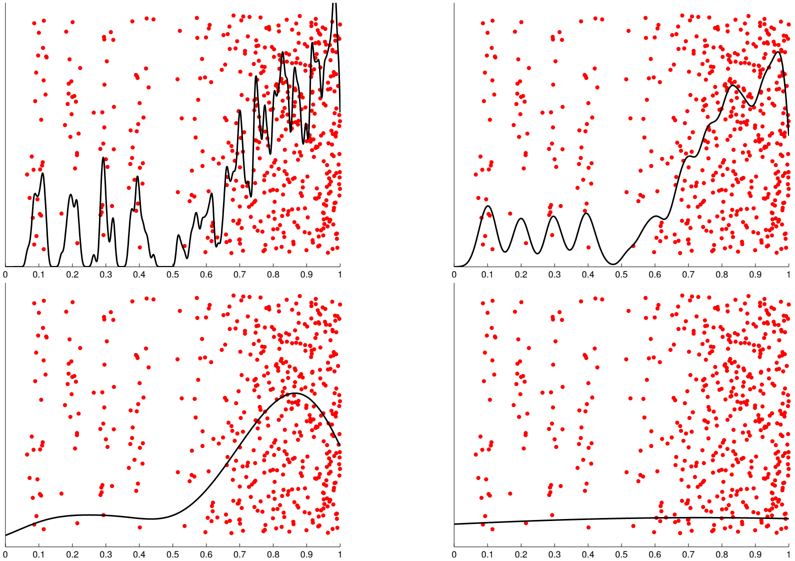
Estimator selection (regression): Nadaraya-Watson



Estimator selection (density): regular histograms



Estimator selection (density): Parzen, Gaussian kernel



Estimator selection

- Estimator/Learning algorithm: $\widehat{s} : D_n \mapsto \widehat{s}(D_n) \in \mathbb{S}$
- Example: least-squares estimator on some model $S_m \subset \mathbb{S}$

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

Examples of models: histograms, $\mathrm{span}\{\varphi_1, \ldots, \varphi_D\}$

Estimator selection

- Estimator/Learning algorithm: $\widehat{s} : D_n \mapsto \widehat{s}(D_n) \in \mathbb{S}$
- Example: least-squares estimator on some model $S_m \subset \mathbb{S}$

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

Examples of models: histograms, $\operatorname{span}\{\varphi_1, \ldots, \varphi_D\}$

- Estimator collection $(\widehat{s}_m)_{m \in \mathcal{M}} \Rightarrow$ choose $\widehat{m} = \widehat{m}(D_n)$?

Estimator selection

- Estimator/Learning algorithm: $\widehat{s} : D_n \mapsto \widehat{s}(D_n) \in \mathbb{S}$
- Example: least-squares estimator on some model $S_m \subset \mathbb{S}$

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

Examples of models: histograms, $\operatorname{span}\{\varphi_1, \ldots, \varphi_D\}$

- Estimator collection $(\widehat{s}_m)_{m \in \mathcal{M}} \Rightarrow$ choose $\widehat{m} = \widehat{m}(D_n)$?
- Examples:
  - model selection
  - calibration of tuning parameters (choosing k or the distance for k-NN, choice of a regularization parameter, choice of a kernel, etc.)
  - choice between different methods
    ex.: k-NN vs. smoothing splines?

Estimator selection: two possible goals

- Estimation goal: minimize the risk of the final estimator, i.e., Oracle inequality (in expectation or with a large probability):

$$\ell\left(s^{\star}, \widehat{s}_{\widehat{m}}\right) \leqslant C \inf_{m \in \mathcal{M}} \left\{ \ell\left(s^{\star}, \widehat{s}_{m}\right) \right\} + R_n$$

Estimator selection: two possible goals

- Estimation goal: minimize the risk of the final estimator, i.e., Oracle inequality (in expectation or with a large probability):

$$\ell\left(s^{\star}, \widehat{s}_{\widehat{m}}\right) \leqslant C \inf_{m \in \mathcal{M}} \left\{ \ell\left(s^{\star}, \widehat{s}_{m}\right) \right\} + R_n$$

- Identification goal: select the (asymptotically) best model/estimator, assuming it is well-defined, i.e., Selection consistency:

$$\mathbb{P}(\widehat{m}(D_n) = m^{\star}) \xrightarrow[n \to \infty]{} 1 \, .$$

Equivalent to estimation in the parametric setting.

Estimator selection: two possible goals

- Estimation goal: minimize the risk of the final estimator, i.e., Oracle inequality (in expectation or with a large probability):

$$\ell\left(s^{\star}, \widehat{s}_{\widehat{m}}\right) \leqslant C \inf_{m \in \mathcal{M}} \left\{ \ell\left(s^{\star}, \widehat{s}_{m}\right) \right\} + R_n$$

- Identification goal: select the (asymptotically) best model/estimator, assuming it is well-defined, i.e., Selection consistency:

$$\mathbb{P}(\widehat{m}(D_n) = m^{\star}) \xrightarrow[n \to \infty]{} 1 \, .$$

  Equivalent to estimation in the parametric setting.

- Both goals with the same procedure (AIC-BIC dilemma)? No in general (Yang, 2005). Sometimes possible.

Estimation goal: Bias-variance trade-off

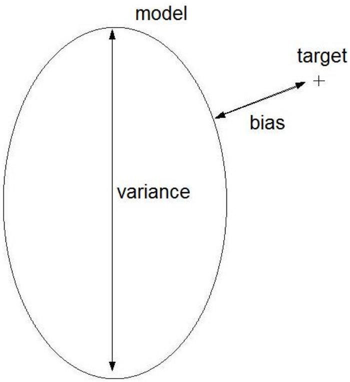
$$\mathbb{E}\left[\ell\left(s^\star, \widehat{s}_m\right)\right] = \text{Bias} + \text{Variance}$$

Bias or Approximation error

$$\ell\left(s^\star, s_m^\star\right) = \inf_{t \in S_m} \ell\left(s^\star, t\right)$$

Variance or Estimation error

$$\text{OLS in regression:} \quad \frac{\sigma^2 \dim(S_m)}{n}$$



Estimation goal: Bias-variance trade-off

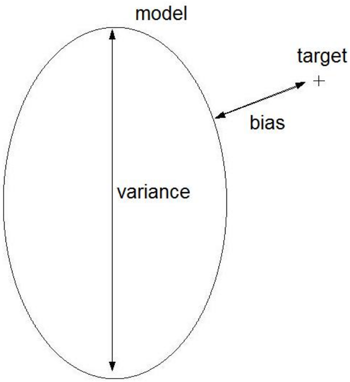
$$\mathbb{E}\left[\ell\left(s^\star, \widehat{s}_m\right)\right] = \text{Bias} + \text{Variance}$$

Bias or Approximation error

$$\ell\left(s^\star, s_m^\star\right) = \inf_{t \in S_m} \ell\left(s^\star, t\right)$$

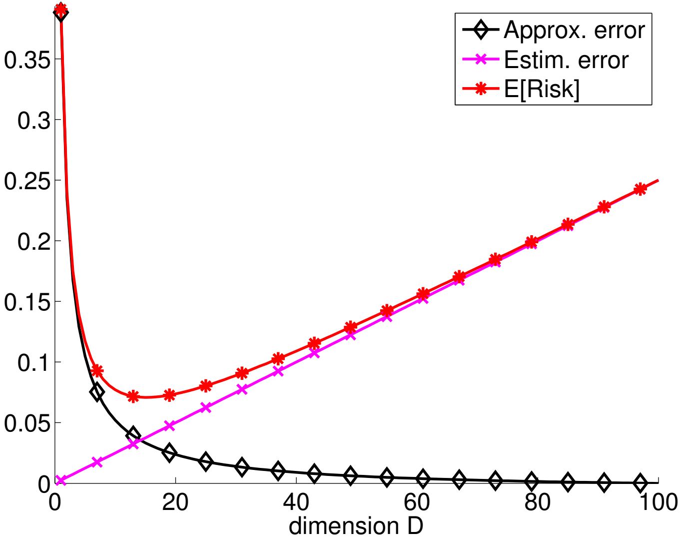
Variance or Estimation error

$$\text{OLS in regression:} \quad \frac{\sigma^2 \dim(S_m)}{n}$$



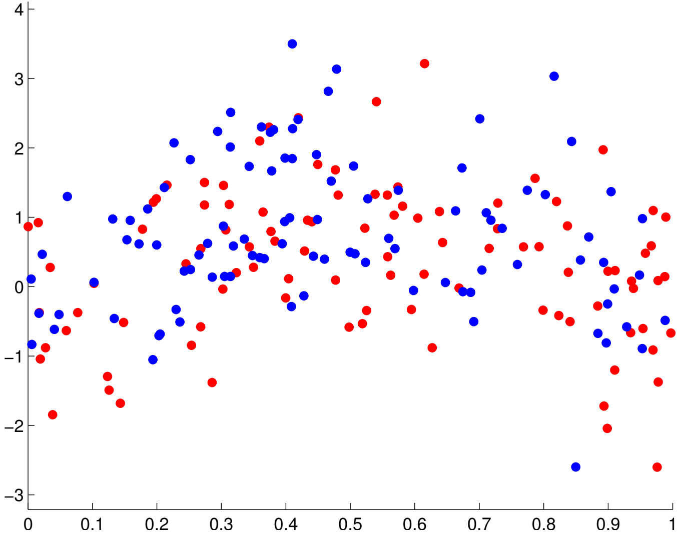
Bias-variance trade-off
⇔ avoid overfitting and underfitting

Estimation goal: Bias-variance trade-off

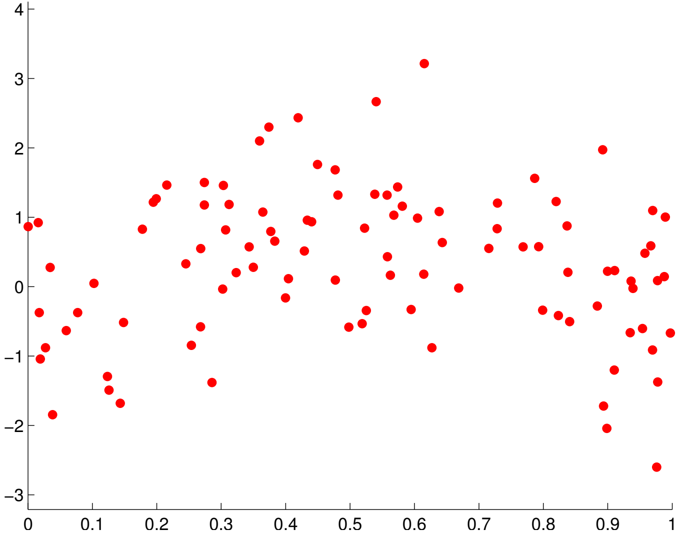


Outline

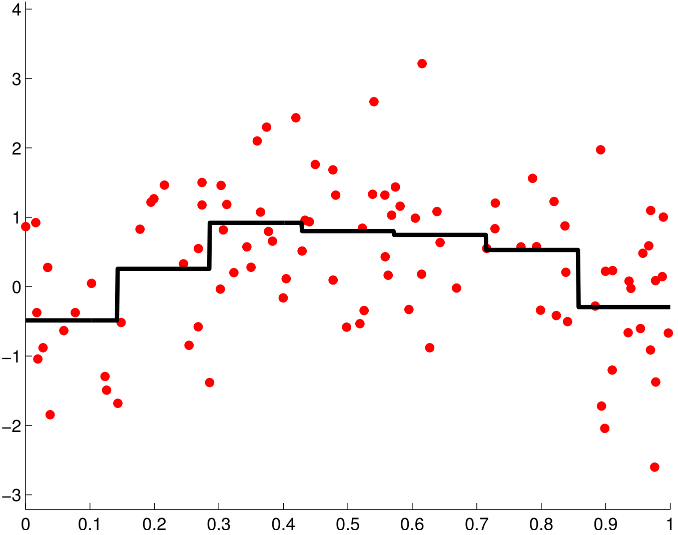
Validation principle



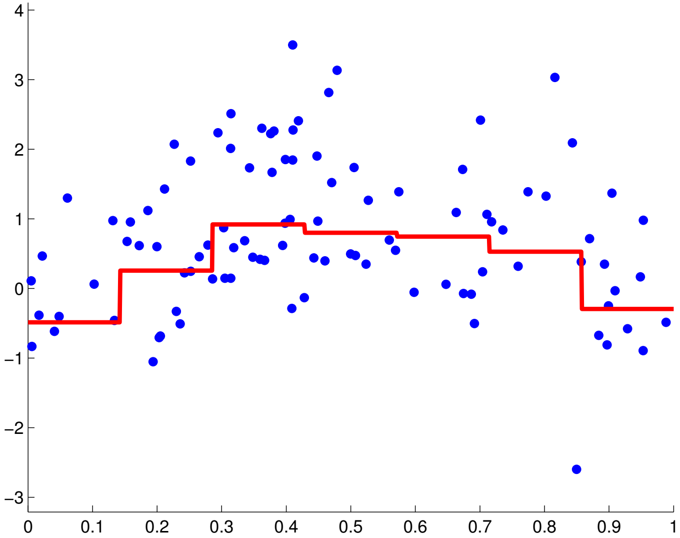
Validation principle: learning sample



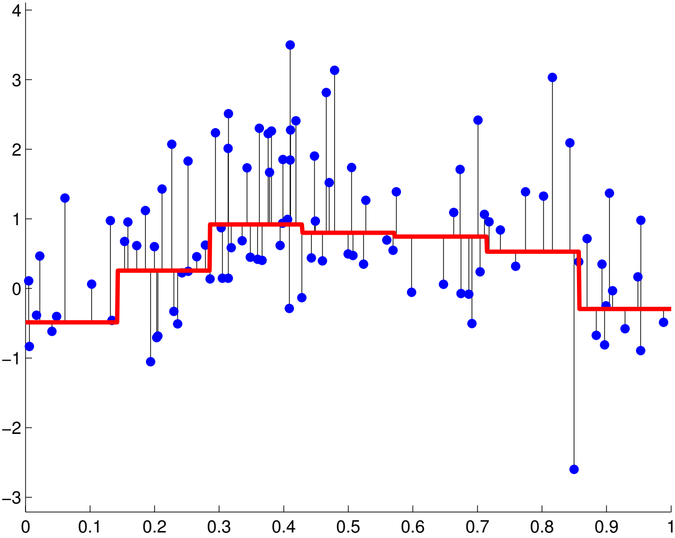
Validation principle: learning sample



Validation principle: validation sample



Validation principle: validation sample



Cross-validation

$$\underbrace{(X_1, Y_1), \ldots, (X_{n_t}, Y_{n_t})}_{\text{Training set } D_n^{(t)} \Rightarrow \widehat{s}_m^{(t)} = \widehat{s}_m(D_n^{(t)})} \quad \underbrace{(X_{n_t+1}, Y_{n_t+1}), \ldots, (X_n, Y_n)}_{\text{Validation set } D_n^{(v)} \Rightarrow \text{evaluate risk}}$$

Cross-validation

$$\underbrace{(X_1, Y_1), \ldots, (X_{n_t}, Y_{n_t})}_{\text{Training set } D_n^{(t)} \Rightarrow \widehat{s}_m^{(t)} = \widehat{s}_m(D_n^{(t)})} \quad \underbrace{(X_{n_t+1}, Y_{n_t+1}), \ldots, (X_n, Y_n)}_{\text{Validation set } D_n^{(v)} \Rightarrow \text{evaluate risk}}$$

- hold-out estimator of the risk:

$$P_n^{(v)} \gamma \left(\widehat{s}_m^{(t)} \right) = \frac{1}{n_v} \sum_{\xi \in D_n^{(v)}} \gamma \left(\widehat{s}_m^{(t)} ; \xi \right) \qquad n_v = |D_n^{(v)}| = n - n_t$$

Cross-validation

$$\underbrace{(X_1, Y_1), \ldots, (X_{n_t}, Y_{n_t})}_{\text{Training set } D_n^{(t)} \Rightarrow \widehat{s}_m^{(t)} = \widehat{s}_m(D_n^{(t)})} \quad \underbrace{(X_{n_t+1}, Y_{n_t+1}), \ldots, (X_n, Y_n)}_{\text{Validation set } D_n^{(v)} \Rightarrow \text{ evaluate risk}}$$

- hold-out estimator of the risk:

$$P_n^{(v)} \gamma\left(\widehat{s}_m^{(t)}\right) = \frac{1}{n_v} \sum_{\xi \in D_n^{(v)}} \gamma\left(\widehat{s}_m^{(t)}; \xi\right) \qquad n_v = |D_n^{(v)}| = n - n_t$$

- cross-validation: average several hold-out estimators

$$\widehat{\mathcal{R}}^{\mathrm{cv}}\left(\widehat{s}_m; D_n; (I_j^{(t)})_{1 \leqslant j \leqslant B}\right) = \frac{1}{B} \sum_{j=1}^{B} P_n^{(v,j)} \gamma\left(\widehat{s}_m^{(t,j)}\right) \qquad D_n^{(t,j)} = (\xi_i)_{i \in I_j^{(t)}}$$

Cross-validation

$$\underbrace{(X_1, Y_1), \ldots, (X_{n_t}, Y_{n_t})}_{\text{Training set } D_n^{(t)} \Rightarrow \widehat{s}_m^{(t)} = \widehat{s}_m(D_n^{(t)})} \quad \underbrace{(X_{n_t+1}, Y_{n_t+1}), \ldots, (X_n, Y_n)}_{\text{Validation set } D_n^{(v)} \Rightarrow \text{ evaluate risk}}$$

- hold-out estimator of the risk:

$$P_n^{(v)} \gamma\left(\widehat{s}_m^{(t)}\right) = \frac{1}{n_v} \sum_{\xi \in D_n^{(v)}} \gamma\left(\widehat{s}_m^{(t)}; \xi\right) \qquad n_v = |D_n^{(v)}| = n - n_t$$

- cross-validation: average several hold-out estimators

$$\widehat{\mathcal{R}}^{\mathrm{cv}}\left(\widehat{s}_m; D_n; (I_j^{(t)})_{1 \leqslant j \leqslant B}\right) = \frac{1}{B} \sum_{j=1}^{B} P_n^{(v,j)} \gamma\left(\widehat{s}_m^{(t,j)}\right) \qquad D_n^{(t,j)} = (\xi_i)_{i \in I_j^{(t)}}$$

- estimator selection:

$$\widehat{m} \in \operatorname*{argmin}_{m \in \mathcal{M}} \left\{ \widehat{\mathcal{R}}^{\mathrm{cv}}\left(\widehat{s}_m; D_n\right) \right\}$$

Cross-validation: examples

- Exhaustive data splitting: all possible subsets of size n_t
  $\Rightarrow$ leave-one-out ($n_t = n - 1$)

$$\widehat{\mathcal{R}}^{\mathrm{loo}}\left(\widehat{s}_m; D_n\right) = \frac{1}{n}\sum_{j=1}^{n}\gamma\left(\widehat{s}_m^{(-j)}; \xi_j\right)$$

  $\Rightarrow$ leave-p-out ($n_t = n - p$)

Cross-validation: examples

- Exhaustive data splitting: all possible subsets of size n_t
  $\Rightarrow$ leave-one-out ($n_t = n - 1$)

$$\widehat{\mathcal{R}}^{\mathrm{loo}}\left(\widehat{s}_m; D_n\right) = \frac{1}{n}\sum_{j=1}^{n}\gamma\left(\widehat{s}_m^{(-j)}; \xi_j\right)$$

  $\Rightarrow$ leave-p-out ($n_t = n - p$)

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

$$\Rightarrow \widehat{\mathcal{R}}^{\mathrm{vf}}\left(\widehat{s}_m; D_n; \mathcal{B}\right) = \frac{1}{V}\sum_{j=1}^{V} P_n^j \gamma\left(\widehat{s}_m^{(-j)}\right)$$

Cross-validation: examples

- Exhaustive data splitting: all possible subsets of size n_t
  $\Rightarrow$ leave-one-out ($n_t = n - 1$)

$$\widehat{\mathcal{R}}^{\mathrm{loo}}\left(\widehat{s}_m; D_n\right) = \frac{1}{n}\sum_{j=1}^{n} \gamma\left(\widehat{s}_m^{(-j)}; \xi_j\right)$$

  $\Rightarrow$ leave-p-out ($n_t = n - p$)

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

$$\Rightarrow \widehat{\mathcal{R}}^{\mathrm{vf}}\left(\widehat{s}_m; D_n; \mathcal{B}\right) = \frac{1}{V}\sum_{j=1}^{V} P_n^j \gamma\left(\widehat{s}_m^{(-j)}\right)$$

- Monte-Carlo CV / Repeated learning testing:

$$I_1^{(t)}, \ldots, I_B^{(t)} \text{ i.i.d. uniform}$$

Outline

Bias of cross-validation

- In this talk, we always assume: $\forall j$, $\mathrm{Card}(D_n^{(t,j)}) = n_t$
  For V-fold CV: $\mathrm{Card}(B_j) = n/V$.
- Ideal criterion: $P\gamma(\widehat{s}_m(D_n))$

Bias of cross-validation

- In this talk, we always assume: $\forall j$, $\operatorname{Card}\left(D_n^{(t,j)}\right) = n_t$
  For V-fold CV: $\operatorname{Card}(B_j) = n/V$.
- Ideal criterion: $P\gamma(\widehat{s}_m(D_n))$
- General analysis for the bias:

$$\mathbb{E}\left[\widehat{\mathcal{R}}^{\mathrm{cv}}\left(\widehat{s}_m; D_n; \left(I_j^{(t)}\right)_{1 \leqslant j \leqslant B}\right)\right] = \mathbb{E}\left[P\gamma(\widehat{s}_m(D_{n_t}))\right]$$

Bias of cross-validation

- In this talk, we always assume: $\forall j$, $\mathrm{Card}\left(D_n^{(t,j)}\right) = n_t$
  For V-fold CV: $\mathrm{Card}(B_j) = n/V$.
- Ideal criterion: $P\gamma(\widehat{s}_m(D_n))$
- General analysis for the bias:

$$\mathbb{E}\left[\widehat{\mathcal{R}}^{\mathrm{cv}}\left(\widehat{s}_m; D_n; \left(I_j^{(t)}\right)_{1 \leqslant j \leqslant B}\right)\right] = \mathbb{E}\left[P\gamma\left(\widehat{s}_m\left(D_{n_t}\right)\right)\right]$$

$\Rightarrow$ everything depends on $n \to \mathbb{E}\left[P\gamma(\widehat{s}_m(D_n))\right]$

Bias of cross-validation

- In this talk, we always assume: $\forall j$, $\mathrm{Card}\left(D_n^{(t,j)}\right) = n_t$
  For V-fold CV: $\mathrm{Card}(B_j) = n/V$.
- Ideal criterion: $P\gamma(\widehat{s}_m(D_n))$
- General analysis for the bias:

$$\mathbb{E}\left[\widehat{\mathcal{R}}^{\mathrm{cv}}\left(\widehat{s}_m; D_n; \left(I_j^{(t)}\right)_{1\leqslant j\leqslant B}\right)\right] = \mathbb{E}\left[P\gamma(\widehat{s}_m(D_{n_t}))\right]$$

$\Rightarrow$ everything depends on $n \to \mathbb{E}\left[P\gamma(\widehat{s}_m(D_n))\right]$

- Note: bias can be corrected in some settings (Burman, 1989).

Bias of cross-validation

- In this talk, we always assume: $\forall j$, $\text{Card}\left(D_n^{(t,j)}\right) = n_t$
  For V-fold CV: $\text{Card}(B_j) = n/V$.
- Ideal criterion: $P\gamma(\widehat{s}_m(D_n))$
- General analysis for the bias:

$$\mathbb{E}\left[\widehat{\mathcal{R}}^{\mathrm{cv}}\left(\widehat{s}_m; D_n; \left(I_j^{(t)}\right)_{1 \leqslant j \leqslant B}\right)\right] = \mathbb{E}\left[P\gamma\left(\widehat{s}_m(D_{n_t})\right)\right]$$

$\Rightarrow$ everything depends on $n \to \mathbb{E}\left[P\gamma\left(\widehat{s}_m(D_n)\right)\right]$

- Note: bias can be corrected in some settings (Burman, 1989).
- Note: $D_n \to \widehat{s}_m(D_n)$ must be fixed before seeing any data; otherwise, stronger bias.

Bias of cross-validation: generic example

Assume

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

(e.g., LS/ridge/k-NN regression, LS/kernel density estimation).

Bias of cross-validation: generic example

Assume

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

(e.g., LS/ridge/k-NN regression, LS/kernel density estimation).

$$\Rightarrow \qquad \mathbb{E}\left[\widehat{\mathcal{R}}^{\mathrm{cv}}\left(\widehat{s}_m; D_n; \left(I_j^{(t)}\right)_{1 \leqslant j \leqslant B}\right)\right] = \alpha(m) + \frac{n}{n_t}\frac{\beta(m)}{n}$$

Bias of cross-validation: generic example

Assume

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

(e.g., LS/ridge/k-NN regression, LS/kernel density estimation).

$$\Rightarrow \qquad \mathbb{E}\left[\widehat{\mathcal{R}}^{\mathrm{cv}}\left(\widehat{s}_m; D_n; \left(I_j^{(t)}\right)_{1\leqslant j\leqslant B}\right)\right] = \alpha(m) + \frac{n}{n_t}\frac{\beta(m)}{n}$$

$\Rightarrow$ Bias:

- decreases as a function of n_t,
- minimal for $n_t = n - 1$,
- negligible if $n_t \sim n$.

Bias of cross-validation: generic example

Assume

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

(e.g., LS/ridge/k-NN regression, LS/kernel density estimation).

$$\Rightarrow \qquad \mathbb{E}\left[\widehat{\mathcal{R}}^{\mathrm{cv}}\left(\widehat{s}_m; D_n; \left(I_j^{(t)}\right)_{1 \leqslant j \leqslant B}\right)\right] = \alpha(m) + \frac{n}{n_t}\frac{\beta(m)}{n}$$

$\Rightarrow$ Bias:

- decreases as a function of n_t,
- minimal for $n_t = n - 1$,
- negligible if $n_t \sim n$.

$\Rightarrow$ V-fold: bias decreases when V increases, vanishes as $V \to +\infty$.

Variance of cross-validation

- Hold-out (Nadeau & Bengio, 2003):

$$\operatorname{var}\left(P_n^{(v)}\gamma\left(\widehat{s}_m^{(t)}\right)\right) = \frac{1}{n_v}\mathbb{E}\left[\operatorname{var}\left(\gamma(u;\xi) \,\middle|\, u = \widehat{s}_m^{(t)}\right)\right] + \operatorname{var}\left(P\gamma(\widehat{s}_m(D_{n_t}))\right)$$

Variance of cross-validation

- Hold-out (Nadeau & Bengio, 2003):

$$\operatorname{var}\left(P_n^{(v)}\gamma\left(\widehat{s}_m^{(t)}\right)\right) = \frac{1}{n_v}\mathbb{E}\left[\operatorname{var}\left(\gamma(u;\xi) \,\middle|\, u = \widehat{s}_m^{(t)}\right)\right] + \operatorname{var}\left(P\gamma(\widehat{s}_m(D_{n_t}))\right)$$

- Monte-Carlo CV and number of splits: $(p = n - n_t)$

$$\operatorname{var}\left(\widehat{\mathcal{R}}^{\mathrm{cv}}\left(\widehat{s}_m; D_n; \left(I_j^{(t)}\right)_{1\leqslant j\leqslant B}\right)\right) = \operatorname{var}\left(\widehat{\mathcal{R}}^{\ell\mathrm{po}}(\widehat{s}_m; D_n)\right) + \frac{1}{B}\underbrace{\mathbb{E}\left[\operatorname{var}_{I^{(t)}}\left(P_n^{(v)}\gamma\left(\widehat{s}_m^{(t)}\right) \,\middle|\, D_n\right)\right]}_{\text{permutation variance}}$$

Variance of cross-validation

- Hold-out (Nadeau & Bengio, 2003):

$$\operatorname{var}\left(P_n^{(v)}\gamma\left(\widehat{s}_m^{(t)}\right)\right) = \frac{1}{n_v}\mathbb{E}\left[\operatorname{var}\left(\gamma(u;\xi)\,\middle|\, u = \widehat{s}_m^{(t)}\right)\right] + \operatorname{var}\left(P\gamma(\widehat{s}_m(D_{n_t}))\right)$$

- Monte-Carlo CV and number of splits: $(p = n - n_t)$

$$\operatorname{var}\left(\widehat{\mathcal{R}}^{\mathrm{cv}}\left(\widehat{s}_m; D_n; \left(I_j^{(t)}\right)_{1\leqslant j\leqslant B}\right)\right) = \operatorname{var}\left(\widehat{\mathcal{R}}^{\ell\mathrm{po}}(\widehat{s}_m; D_n)\right) + \frac{1}{B}\underbrace{\mathbb{E}\left[\operatorname{var}_{I^{(t)}}\left(P_n^{(v)}\gamma\left(\widehat{s}_m^{(t)}\right)\,\middle|\, D_n\right)\right]}_{\text{permutation variance}}$$

- V-fold CV: B, n_t, n_v related
  leave-one-out: related to stability? (empirical results)

Variance of the V-fold CV criterion

- Least-squares density estimation (A. & Lerasle 2012), exact computation (non-asymptotic):

$$\operatorname{var}\left(\widehat{\mathcal{R}}^{\mathrm{vf}}\left(\widehat{s}_m; D_n; \mathcal{B}\right)\right) = \frac{1+\mathcal{O}(1)}{n} \operatorname{var}_P(s_m^\star) + \frac{2}{n^2}\left[1 + \frac{4}{V-1} + \mathcal{O}\left(\frac{1}{V} + \frac{1}{n}\right)\right] A(m)$$

(simplified formula, histogram model with bin size d_m^{-1}, $A(m) \approx d_m$)

- Linear regression, asymptotic formula (Burman, 1989):

$$\operatorname{var}\left(\widehat{\mathcal{R}}^{\mathrm{vf}}\left(\widehat{s}_m; D_n; \mathcal{B}\right)\right) = \frac{2\sigma^2}{n} + \frac{4\sigma^4}{n^2}\left[4 + \frac{4}{V-1} + \frac{2}{(V-1)^2} + \frac{1}{(V-1)^3}\right] + \mathrm{o}\left(n^{-2}\right)$$

$\Rightarrow$ decreasing with V, dependence only in second order terms.

Outline

Risk estimation and estimator selection are different goals

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

- For any Z (deterministic or random),

$$\widehat{m} \in \operatorname*{argmin}_{m \in \mathcal{M}} \left\{ \widehat{\mathcal{R}}^{\mathrm{cv}}\left(\widehat{s}_m\right) + Z \right\}$$

$\Rightarrow$ bias and variance meaningless.

Risk estimation and estimator selection are different goals

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

- For any Z (deterministic or random),

$$\widehat{m} \in \operatorname*{argmin}_{m \in \mathcal{M}} \left\{ \widehat{\mathcal{R}}^{\mathrm{cv}} \left(\widehat{s}_m \right) + Z \right\}$$

  $\Rightarrow$ bias and variance meaningless.

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

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

Risk estimation and estimator selection are different goals

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

- For any Z (deterministic or random),

$$\widehat{m} \in \operatorname*{argmin}_{m \in \mathcal{M}} \left\{ \widehat{\mathcal{R}}^{\mathrm{cv}}\left(\widehat{s}_m\right) + Z \right\}$$

  $\Rightarrow$ bias and variance meaningless.

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

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

$\Rightarrow$ $\mathbb{E}\left[\widehat{\mathcal{R}}^{\mathrm{cv}}\left(\widehat{s}_m\right) - \widehat{\mathcal{R}}^{\mathrm{cv}}\left(\widehat{s}_{m'}\right)\right]$ should be of the good sign (unbiased risk estimation heuristic: AIC, C_p, leave-one-out...)

Risk estimation and estimator selection are different goals

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

- For any Z (deterministic or random),

$$\widehat{m} \in \operatorname{argmin}_{m \in \mathcal{M}} \left\{ \widehat{\mathcal{R}}^{\mathrm{cv}}(\widehat{s}_m) + Z \right\}$$

$\Rightarrow$ bias and variance meaningless.

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

$$\operatorname{sign}(\widehat{\mathcal{R}}^{\mathrm{cv}}(\widehat{s}_m) - \widehat{\mathcal{R}}^{\mathrm{cv}}(\widehat{s}_{m'})) = \operatorname{sign}(P\gamma(\widehat{s}_m) - P\gamma(\widehat{s}_{m'}))$$

$\Rightarrow$ $\mathbb{E}\left[\widehat{\mathcal{R}}^{\mathrm{cv}}(\widehat{s}_m) - \widehat{\mathcal{R}}^{\mathrm{cv}}(\widehat{s}_{m'})\right]$ should be of the good sign (unbiased risk estimation heuristic: AIC, C_p, leave-one-out...)

$\Rightarrow$ $\operatorname{var}\left(\widehat{\mathcal{R}}^{\mathrm{vf}}(\widehat{s}_m) - \widehat{\mathcal{R}}^{\mathrm{vf}}(\widehat{s}_{m'})\right)$ should be minimal (detailed heuristic: A. & Lerasle 2012)

Bias and estimator selection: generic example

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

- Assume

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

(e.g., LS/ridge/k NN regression, LS/kernel density estimation).

Bias and estimator selection: generic example

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

- Assume

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

(e.g., LS/ridge/k NN regression, LS/kernel density estimation).

- Key quantities:

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

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

$\Rightarrow$ CV favours m with smaller complexity $\beta(m)$, more and more as n_t decreases.

CV with an estimation goal: the big picture ($\mathcal{M}$ "small")

- At first order, the bias drives the performance of:
  leave-p-out, V-fold CV,
  Monte-Carlo CV if $B \gg n^2$
  or if n_v large enough (including hold-out)
- CV performs similarly to

$$\operatorname*{argmin}_{m \in \mathcal{M}} \left\{ \mathbb{E}\left[P\gamma\left(\widehat{s}_m(D_{n_t})\right)\right] \right\}$$

CV with an estimation goal: the big picture ($\mathcal{M}$ "small")

- At first order, the bias drives the performance of:
  leave-p-out, V-fold CV,
  Monte-Carlo CV if $B \gg n^2$
  or if n_v large enough (including hold-out)
- CV performs similarly to

$$\operatorname*{argmin}_{m \in \mathcal{M}} \left\{ \mathbb{E}\left[P\gamma\left(\widehat{s}_m(D_{n_t})\right)\right]\right\}$$

$\Rightarrow$ first-order optimality if $n_t \sim n$

$\Rightarrow$ suboptimal otherwise
e.g., V-fold CV with V fixed.

- Theoretical results for least-squares regression and density estimation at least.

Bias-corrected VFCV / V-fold penalization

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

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

Bias-corrected VFCV / V-fold penalization

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

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

$$= P_n\gamma\left(\widehat{s}_m\right) + \underbrace{\mathrm{pen}_{\mathrm{VF}}(\widehat{s}_m; D_n; \mathcal{B})}_{V\text{-fold penalty (A. 2008)}}$$

Bias-corrected VFCV / V-fold penalization

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

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

$$= P_n\gamma\left(\widehat{s}_m\right) + \underbrace{\mathrm{pen}_{\mathrm{VF}}(\widehat{s}_m; D_n; \mathcal{B})}_{V\text{-fold penalty (A. 2008)}}$$

- In least-squares density estimation (A. & Lerasle, 2012):

$$\widehat{\mathcal{R}}^{\mathrm{vf}}(\widehat{s}_m; D_n; \mathcal{B}) = P_n\gamma(\widehat{s}_m(D_n)) + \underbrace{\left(1 + \frac{1}{2(V-1)}\right)}_{\text{overpenalization factor}} \mathrm{pen}_{\mathrm{VF}}(\widehat{s}_m; D_n; \mathcal{B})$$

$$\widehat{\mathcal{R}}^{\ell\mathrm{po}}(\widehat{s}_m; D_n; \mathcal{B}) = P_n\gamma(\widehat{s}_m(D_n)) + \overbrace{\left(1 + \frac{1}{2\left(\frac{n}{p} - 1\right)}\right)} \mathrm{pen}_{\mathrm{VF}}(\widehat{s}_m; D_n; \mathcal{B}_{\mathrm{loo}})$$

Variance and estimator selection

$$\Delta(m, m', V) = \widehat{\mathcal{R}}^{\mathrm{vf,corr}}(\widehat{s}_m) - \widehat{\mathcal{R}}^{\mathrm{vf,corr}}(\widehat{s}_{m'})$$

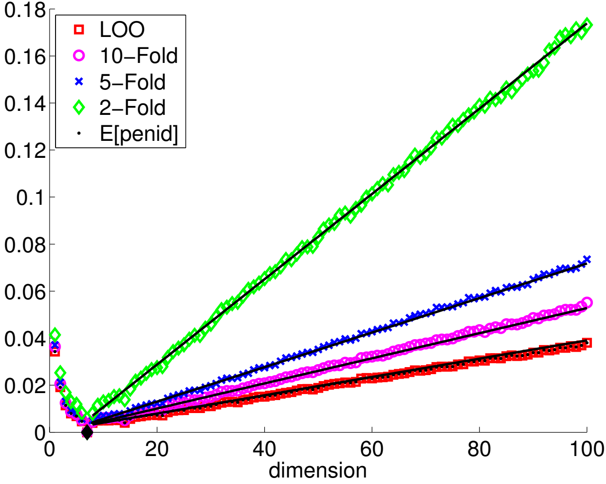
Theorem (A. & Lerasle 2012, least-squares density estimation)

$$\operatorname{var}(\Delta(m, m', V)) = 4\left(1 + \frac{2}{n} + \frac{1}{n^2}\right)\frac{\operatorname{var}_P(s_m^\star - s_{m'}^\star)}{n} + 2\left(1 + \frac{4}{V-1} - \frac{1}{n}\right)\underbrace{\frac{B(m, m')}{n^2}}_{\geqslant 0}$$

If $S_m \subset S_{m'}$ are two histogram models with constant bin sizes $d_m^{-1}, d_{m'}^{-1}$, then, $B(m, m') \propto \|s_m^\star - s_{m'}^\star\| \, d_m$.

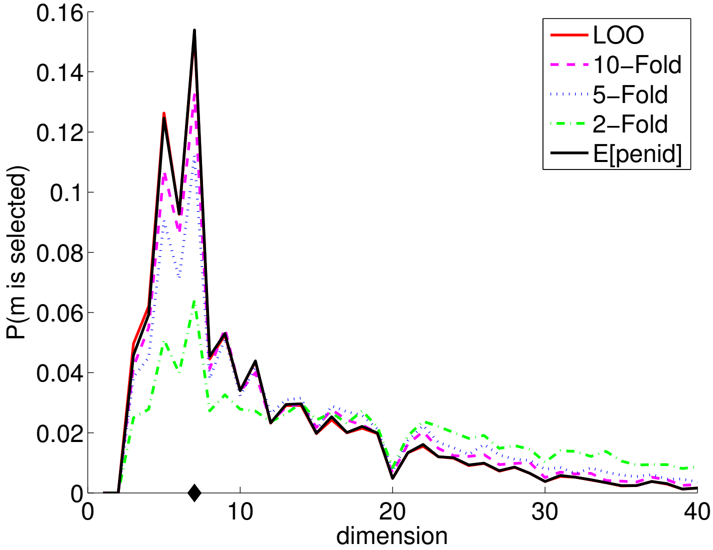
The two terms are of the same order if $\|s_m^\star - s_{m'}^\star\| \approx d_m / n$.

Variance of $\widehat{\mathcal{R}}^{\mathrm{vf,corr}}(\widehat{s}_m) - \widehat{\mathcal{R}}^{\mathrm{vf,corr}}(\widehat{s}_{m^\star})$ vs. (d_m, V)

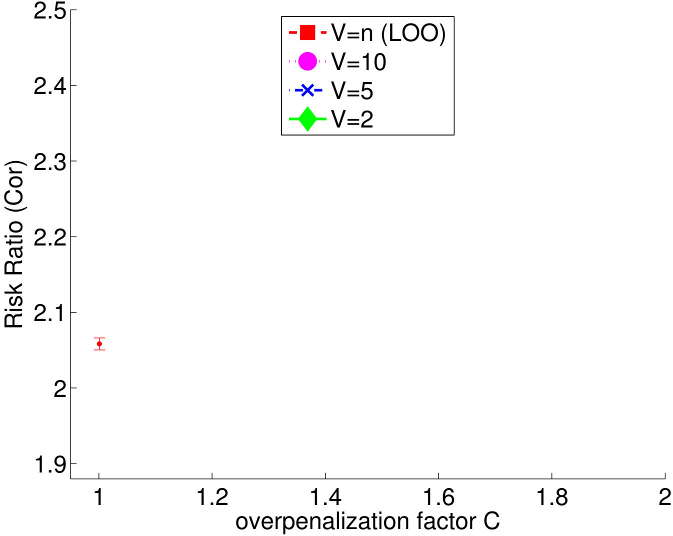


$$\mathrm{var}(\Delta(m, m', V)) \approx n^{-2}\left[29\left(1 + \frac{0.8}{V-1}\right) + 3.7\left(1 + \frac{3.8}{V-1}\right)(d_m - d_{m^\star})\right]$$

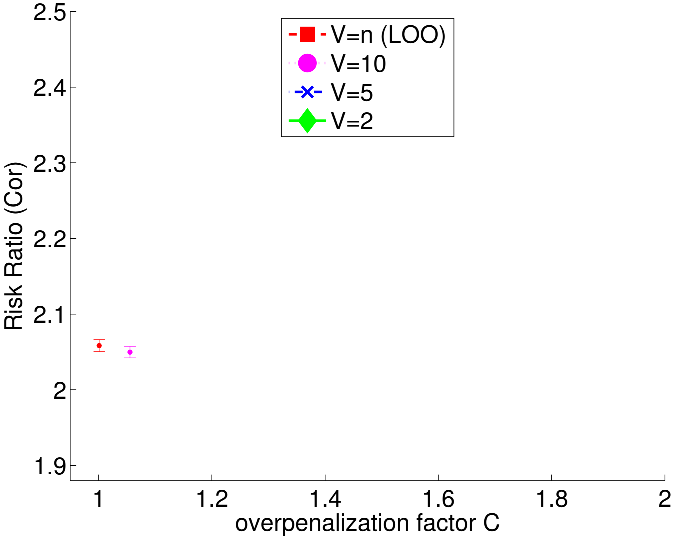
Probability of selection of every m



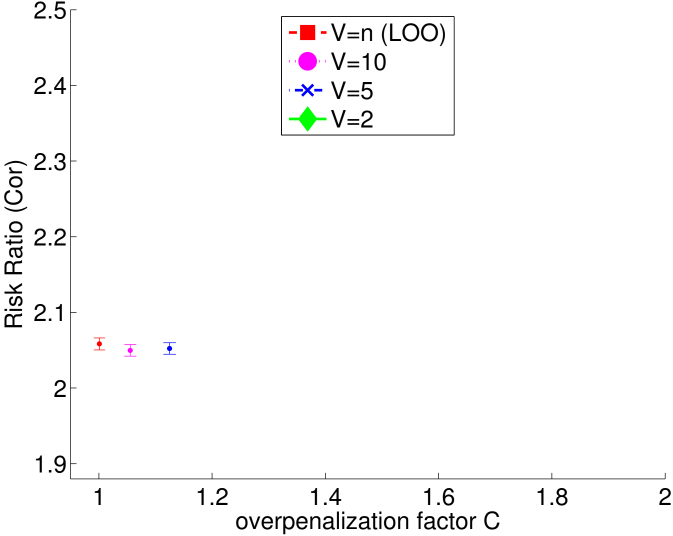
Experiment (LS density estimation): V-fold CV



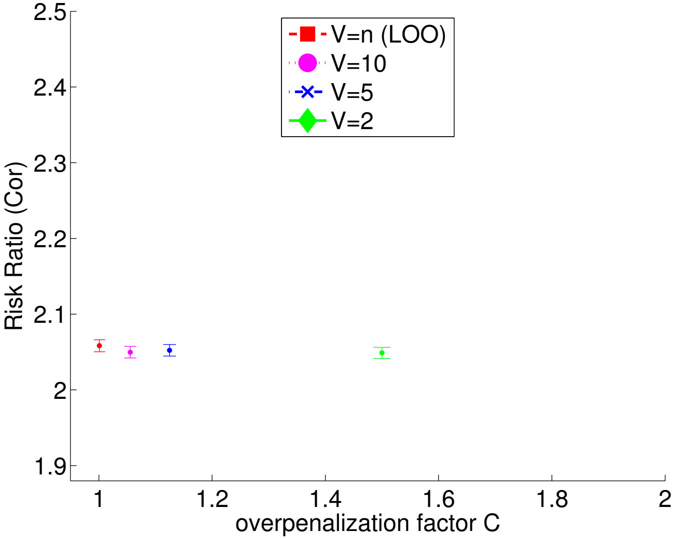
Experiment (LS density estimation): V-fold CV



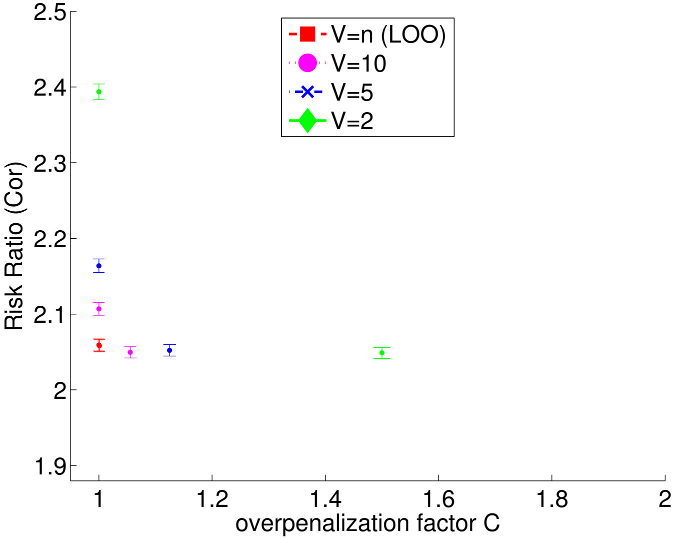
Experiment (LS density estimation): V-fold CV



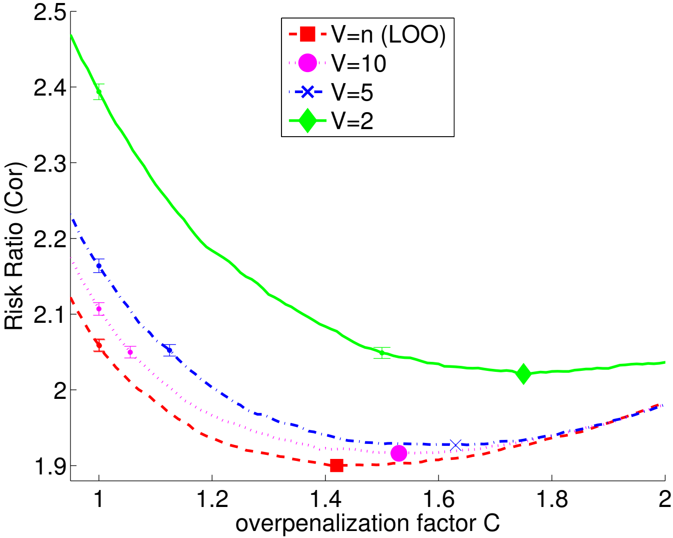
Experiment (LS density estimation): V-fold CV



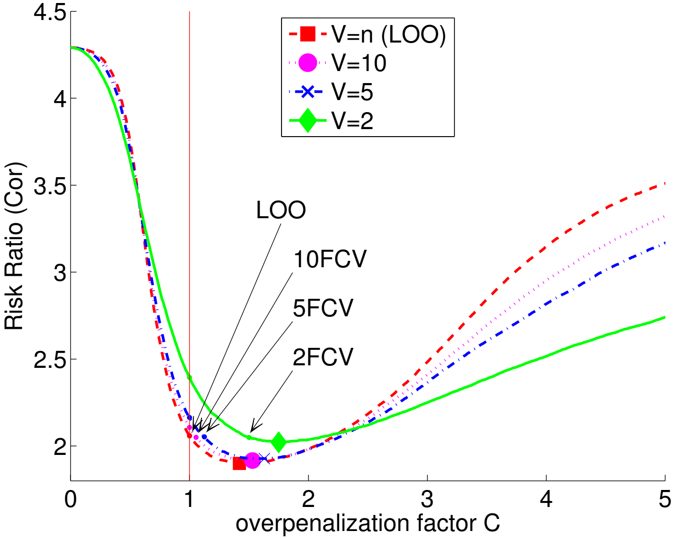
Experiment (LS density estimation): V-fold penalization



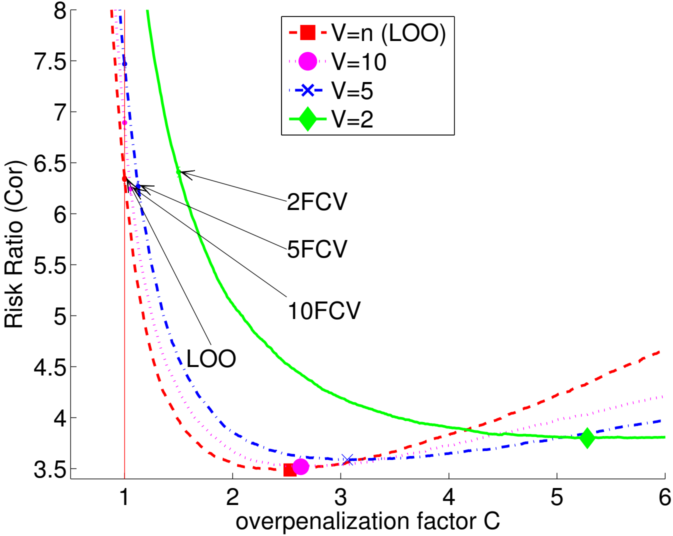
Experiment (LS density estimation): overpenalization



Experiment (LS density estimation): conclusion



Experiment (LS density estimation): other setting



Estimator selection with V-fold: conclusion

- Computational complexity: $\mathcal{O}(V)$ in general

Estimator selection with V-fold: conclusion

- Computational complexity: $\mathcal{O}(V)$ in general
- V-fold cross-validation:
  - Bias: decreases with V / can be removed
  - Variance: decreases with V / almost minimal with $V \in [5, 10]$

$\Rightarrow$ best performance for the largest V and almost optimal with $V = 10$...

Estimator selection with V-fold: conclusion

- Computational complexity: $\mathcal{O}(V)$ in general
- V-fold cross-validation:
  - Bias: decreases with V / can be removed
  - Variance: decreases with V / almost minimal with $V \in [5, 10]$

  $\Rightarrow$ best performance for the largest V and almost optimal with $V = 10$...

  ... if optimal overpenalization factor $C^\star \approx 1$ (various behaviours possible).

Estimator selection with V-fold: conclusion

- Computational complexity: $\mathcal{O}(V)$ in general
- V-fold cross-validation:
    - Bias: decreases with V / can be removed
    - Variance: decreases with V / almost minimal with $V \in [5, 10]$

  $\Rightarrow$ best performance for the largest V and almost optimal with $V = 10$...
  ... if optimal overpenalization factor $C^\star \approx 1$ (various behaviours possible).
- V-fold penalization:
    - Decoupling of bias and variance $\Rightarrow$ easier to understand.
    - Bias: chosen directly through C, without any constraint.
    - Variance: decreases with V / almost minimal with $V \in [5, 10]$.

Outline

Generality of the results

- At least valid for least-square regression / density estimation, kernel density estimation.

Generality of the results

- At least valid for least-square regression / density estimation, kernel density estimation.
- Bias-correction / V-fold penalization: valid if

$$\mathbb{E}\Big[(P - P_n)\gamma(\widehat{s}_m)\Big] \approx \frac{\gamma(m)}{n} \quad .$$

Otherwise: use repeated V-fold or Monte-Carlo CV with a well-chosen n_t.

Generality of the results

- At least valid for least-square regression / density estimation, kernel density estimation.
- Bias-correction / V-fold penalization: valid if

$$\mathbb{E}\Big[(P - P_n)\gamma\left(\widehat{s}_m\right)\Big] \approx \frac{\gamma(m)}{n} \enspace .$$

  Otherwise: use repeated V-fold or Monte-Carlo CV with a well-chosen n_t.
- Variance: different behaviours can occur in other settings (experiments).

Generality of the results

- At least valid for least-square regression / density estimation, kernel density estimation.
- Bias-correction / V-fold penalization: valid if

$$\mathbb{E}\left[(P - P_n)\gamma(\widehat{s}_m)\right] \approx \frac{\gamma(m)}{n} \quad .$$

  Otherwise: use repeated V-fold or Monte-Carlo CV with a well-chosen n_t.
- Variance: different behaviours can occur in other settings (experiments).
- Everything can be checked on synthetic data: plot

$$n \to \mathbb{E}\left[P\gamma(\widehat{s}_m(D_n))\right] \qquad \text{and} \qquad m \to \operatorname{var}\left(\widehat{\mathcal{R}}^{\mathrm{cv}}(\widehat{s}_m) - \widehat{\mathcal{R}}^{\mathrm{cv}}(\widehat{s}_{m^\star})\right) .$$

Large collection of estimators/models

- Estimator/model selection with an "exponential" collection (implicitly excluded in all results above).
  $\Rightarrow$ Expectations do not drive the first order!

Large collection of estimators/models

- Estimator/model selection with an "exponential" collection (implicitly excluded in all results above).
  $\Rightarrow$ Expectations do not drive the first order!

- Examples: variable selection with $p \geqslant n$ variables, change-point detection.

Large collection of estimators/models

- Estimator/model selection with an "exponential" collection (implicitly excluded in all results above).
  $\Rightarrow$ Expectations do not drive the first order!

- Examples: variable selection with $p \geqslant n$ variables, change-point detection.

- Solution: group the models $\Rightarrow$ one estimator per dimension (e.g., empirical risk minimizer)
  works for change-point detection (A. & Celisse, 2010).

Cross-validation with an identification goal

- Main change: value of the optimal overpenalization factor $C^{\star}$, often $C^{\star} \to +\infty$ when $n \to +\infty$.

Cross-validation with an identification goal

- Main change: value of the optimal overpenalization factor $C^\star$, often $C^\star \to +\infty$ when $n \to +\infty$.

$\Leftrightarrow$ Cross-validation paradox (Yang, 2006, 2007): $n_t \ll n$ can be necessary!

- Why? Smaller $n_t \Rightarrow$ easier to distinguish the two best procedures...

Cross-validation with an identification goal

- Main change: value of the optimal overpenalization factor $C^\star$, often $C^\star \to +\infty$ when $n \to +\infty$.

$\Leftrightarrow$ Cross-validation paradox (Yang, 2006, 2007): $n_t \ll n$ can be necessary!

- Why? Smaller $n_t \Rightarrow$ easier to distinguish the two best procedures... if n_t large enough (asymptotic regime).

Cross-validation with an identification goal

- Main change: value of the optimal overpenalization factor $C^\star$, often $C^\star \to +\infty$ when $n \to +\infty$.

$\Leftrightarrow$ Cross-validation paradox (Yang, 2006, 2007): $n_t \ll n$ can be necessary!

- Why? Smaller $n_t \Rightarrow$ easier to distinguish the two best procedures... if n_t large enough (asymptotic regime).

- Remark: estimation goal, parametric setting $\Rightarrow$ similar behaviour.

Dependent data

- $D_n^{(t)}, D_n^{(v)}$ dependent $\Rightarrow$ CV heuristic fails!

$\Rightarrow$ possible troubles for risk estimation (Hart & Wehrly, 1986; Opsomer et al., 2001).

Dependent data

- $D_n^{(t)}, D_n^{(v)}$ dependent $\Rightarrow$ CV heuristic fails!

$\Rightarrow$ possible troubles for risk estimation (Hart & Wehrly, 1986; Opsomer et al., 2001).

- Solution for short-term dependence:
  remove some data at each split $\Rightarrow$ gap between training and validation samples.

Questions?