Cross-validation for estimator selection or aggregation

Sylvain Arlot (joint works with Alain Celisse, Matthieu Lerasle, Nelo Magalhães, Guillaume Maillard)

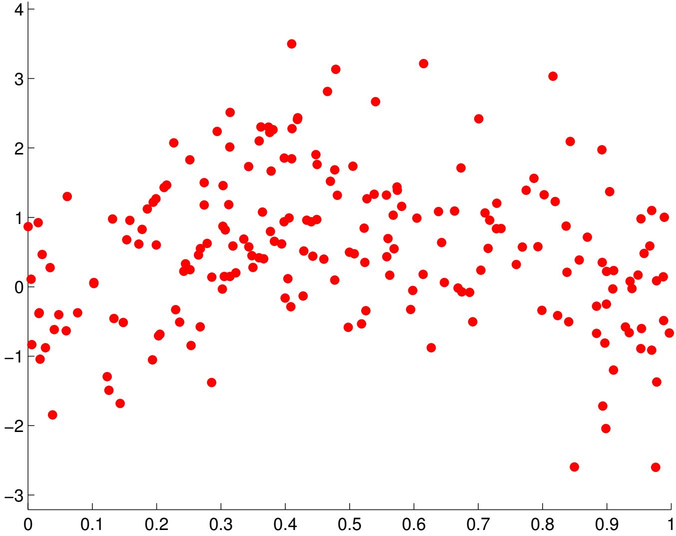
Université Paris-Sud, Laboratoire de Mathématiques d'Orsay

Data and Analytics for Short-Term Operations, Cambridge
February 27, 2019

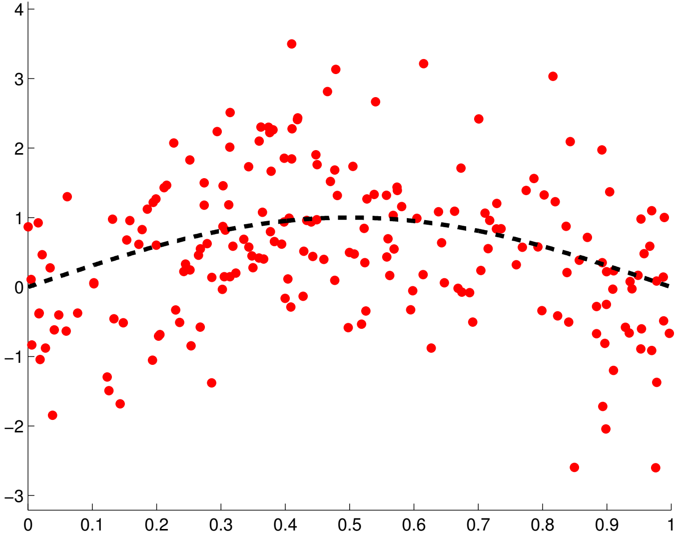
Survey: arXiv:0907.4728 (& arXiv:1703.03167)
CV in L^2 density estimation: arXiv:1210.5830
Aggregated hold-out: arXiv:1709.03702

Outline

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



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

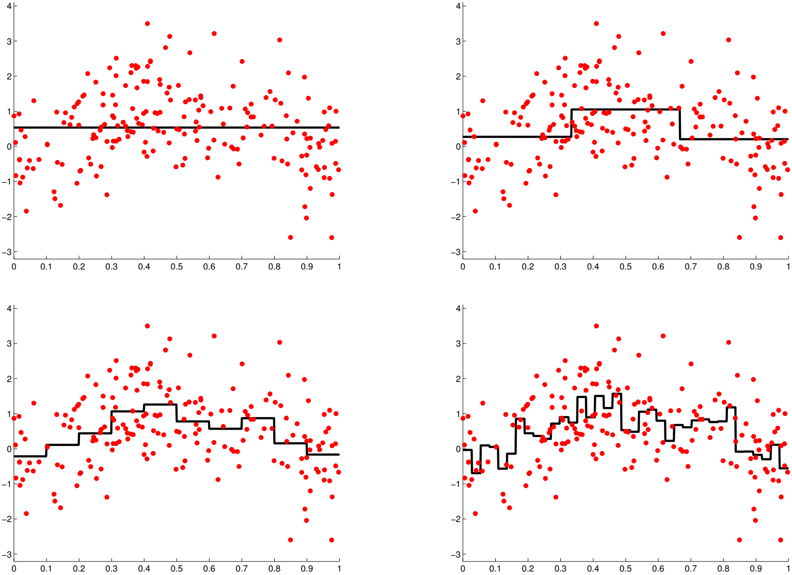


General setting: prediction

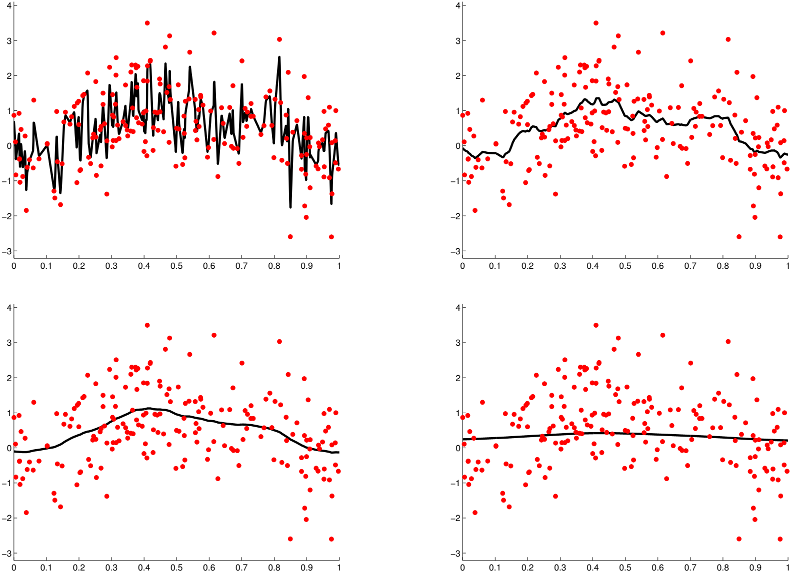
- Data: $D_n = (X_i, Y_i)_{1 \leqslant i \leqslant n} \in (\mathcal{X} \times \mathcal{Y})^n$ assumed i.i.d. $\sim P$
- Predictor: $f : \mathcal{X} \to \mathcal{Y}$ ($\mathcal{F}$: set of all predictors)
- Risk (prediction error): $\mathcal{R}(f) = \mathbb{E}\left[c(f(X), Y)\right]$
  minimal for $f = f^\star$

  LS regression: $c(y, y') = (y - y')^2$, $f^\star(X) = \mathbb{E}[Y | X]$ and
  $\mathcal{R}(f) - \mathcal{R}(f^\star) = \mathbb{E}\left[(f(X) - f^\star(X))^2\right]$
- Goal: from D_n only, find $f \in \mathcal{F}$ with $\mathcal{R}(f)$ minimal.
- Examples: regression, classification
- More general setting possible, including density estimation with LS or KL risk.

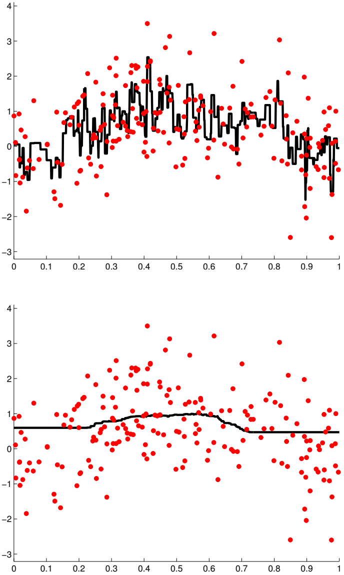
Estimator selection (regression): regular regressograms



Estimator selection (regression): kernel ridge



Estimator selection (regression): k nearest neighbours



Estimator selection

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

$$\widehat{f}_m \in \operatorname*{argmin}_{f \in S_m} \left\{ \widehat{\mathcal{R}}_n(f) \right\} \quad \text{where} \quad \widehat{\mathcal{R}}_n(f) := \frac{1}{n} \sum_{(X_i, Y_i) \in D_n} c\left(f(X_i), Y_i\right)$$

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

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

Estimator selection

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

$$\widehat{f}_m \in \operatorname*{argmin}_{f \in S_m} \left\{ \widehat{\mathcal{R}}_n(f) \right\} \quad \text{where} \quad \widehat{\mathcal{R}}_n(f) := \frac{1}{n} \sum_{(X_i,Y_i) \in D_n} c\big(f(X_i), Y_i\big)$$

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

- Estimator collection $(\widehat{f}_m)_{m \in \mathcal{M}} \Rightarrow$ 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, etc.)
  - choice between different methods
    ex.: random forests vs. SVM?

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):

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

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):

$$\mathcal{R}(\widehat{f}_{\widehat{m}}) - \mathcal{R}(f^{\star}) \leqslant C \inf_{m \in \mathcal{M}} \{ \mathcal{R}(\widehat{f}_m) - \mathcal{R}(f^{\star}) \} + R_n$$

- 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):

$$\mathcal{R}(\widehat{f}_{\widehat{m}}) - \mathcal{R}(f^\star) \leqslant C \inf_{m \in \mathcal{M}} \{\mathcal{R}(\widehat{f}_m) - \mathcal{R}(f^\star)\} + R_n$$

- 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

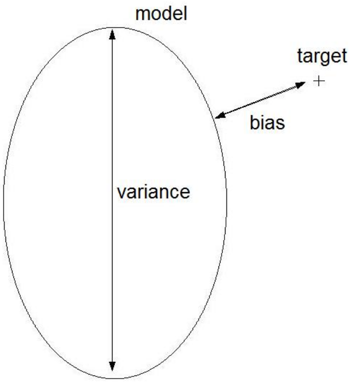
$$\mathbb{E}\left[\mathcal{R}(\widehat{f}_m)\right] - \mathcal{R}(f^\star) = \text{Bias} + \text{Variance}$$

Bias or Approximation error

$$\mathcal{R}(f_m^\star) - \mathcal{R}(f^\star) = \inf_{f \in S_m} \mathcal{R}(f) - \mathcal{R}(f^\star)$$

Variance or Estimation error

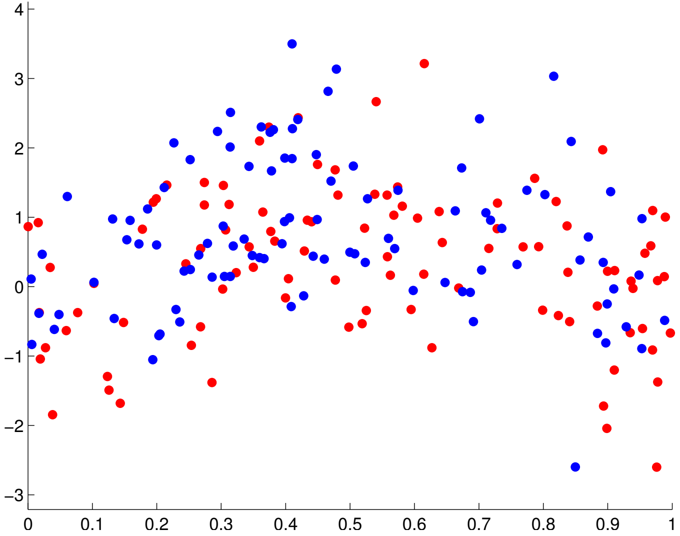
OLS in regression: $\dfrac{\sigma^2 \dim(S_m)}{n}$



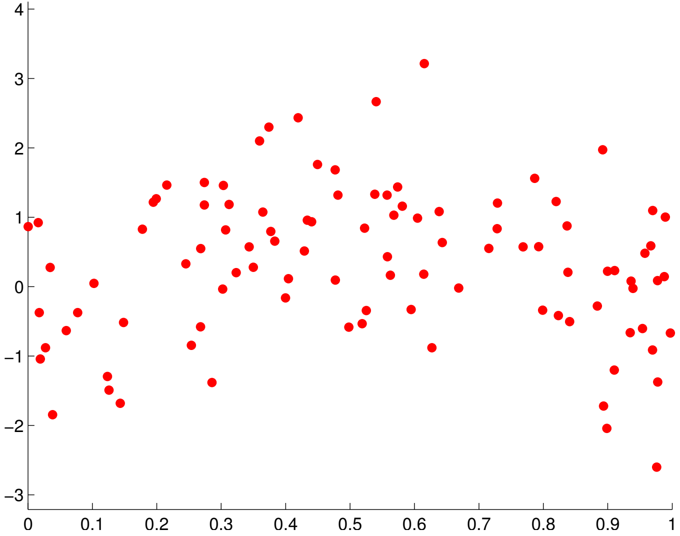
Bias-variance trade-off
⇔ avoid overfitting and underfitting

Outline

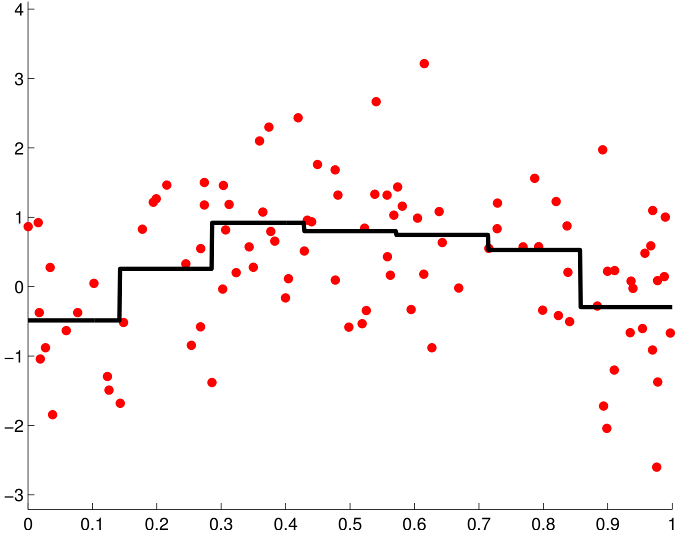
Validation principle: data splitting



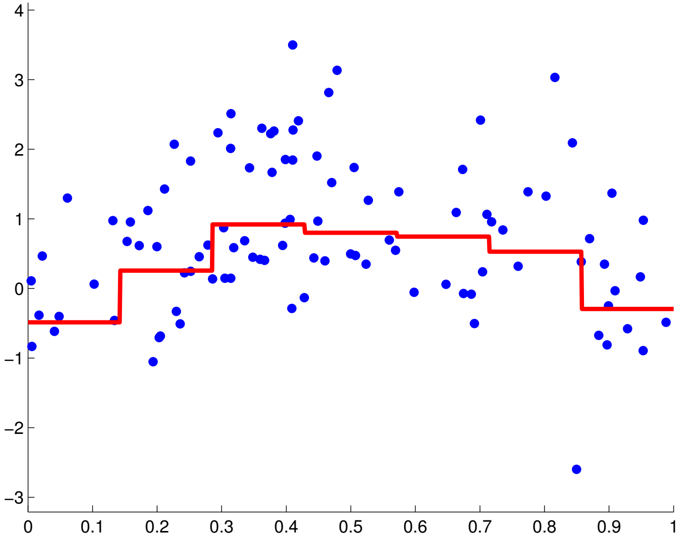
Validation principle: training/learning sample



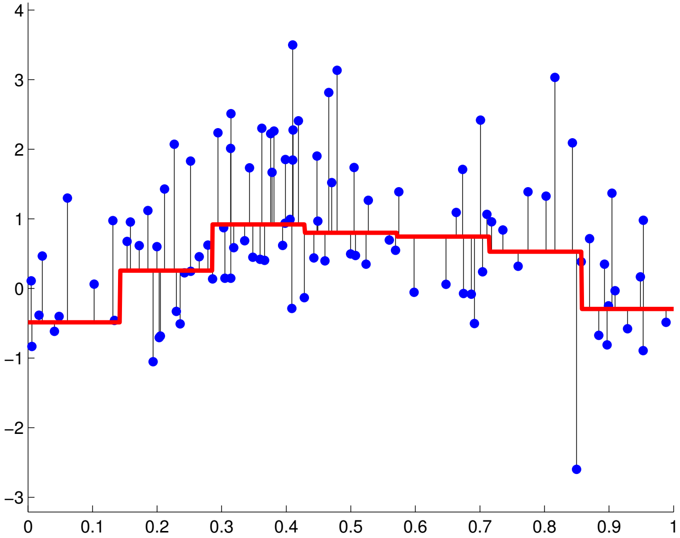
Validation principle: training/learning sample



Validation principle: validation sample



Validation principle: validation sample



Cross-validation

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

- hold-out estimator of the risk:

$$\widehat{\mathcal{R}}_n^{(v)}\left(\widehat{f}_m^{(t)}\right) = \frac{1}{n_v} \sum_{(X_i, Y_i) \in D_n^{(v)}} c\left(\widehat{f}_m^{(t)}(X_i); Y_i\right) \qquad n_v = |D_n^{(v)}| = n - n_t$$

Cross-validation

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

- hold-out estimator of the risk:

$$\widehat{\mathcal{R}}_n^{(v)}\left(\widehat{f}_m^{(t)}\right) = \frac{1}{n_v} \sum_{(X_i,Y_i)\in D_n^{(v)}} c\left(\widehat{f}_m^{(t)}(X_i); Y_i\right) \qquad n_v = |D_n^{(v)}| = n - n_t$$

- cross-validation: average several hold-out estimators

$$\widehat{\mathcal{R}}^{\mathrm{cv}}\left(\widehat{f}_m; D_n; (I_j^{(t)})_{1\leqslant j\leqslant V}\right) = \frac{1}{V}\sum_{j=1}^{V} \widehat{\mathcal{R}}_n^{(v,j)}\left(\widehat{f}_m^{(t,j)}\right) \qquad D_n^{(t,j)} = (X_i, Y_i)_{i\in I_j^{(t)}}$$

Cross-validation

$$\underbrace{(X_1, Y_1), \dots, (X_{n_t}, Y_{n_t})} \qquad \underbrace{(X_{n_t+1}, Y_{n_t+1}), \dots, (X_n, Y_n)}$$

Training set $D_n^{(t)} \Rightarrow \widehat{f}_m^{(t)} = \widehat{f}_m(D_n^{(t)})$ Validation set $D_n^{(v)} \Rightarrow$ evaluate risk

- hold-out estimator of the risk:

$$\widehat{\mathcal{R}}_n^{(v)}\left(\widehat{f}_m^{(t)}\right) = \frac{1}{n_v} \sum_{(X_i, Y_i) \in D_n^{(v)}} c\left(\widehat{f}_m^{(t)}(X_i); Y_i\right) \qquad n_v = |D_n^{(v)}| = n - n_t$$

- cross-validation: average several hold-out estimators

$$\widehat{\mathcal{R}}^{\mathrm{cv}}\left(\widehat{f}_m; D_n; (I_j^{(t)})_{1 \leqslant j \leqslant V}\right) = \frac{1}{V} \sum_{j=1}^{V} \widehat{\mathcal{R}}_n^{(v,j)}\left(\widehat{f}_m^{(t,j)}\right) \qquad D_n^{(t,j)} = (X_i, Y_i)_{i \in I_j^{(t)}}$$

- estimator selection:

$$\widehat{m}^{cv}\left(D_n; (I_j^{(t)})_{1 \leqslant j \leqslant V}\right) \in \operatorname*{argmin}_{m \in \mathcal{M}} \left\{ \widehat{\mathcal{R}}^{\mathrm{cv}}\left(\widehat{f}_m; D_n\right) \right\} \quad \Rightarrow \quad \widehat{f}_{\widehat{m}^{cv}(D_n; (I_j^{(t)})_{1 \leqslant j \leqslant V})}(D_n)$$

Cross-validation: examples

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

$$\widehat{\mathcal{R}}^{\mathrm{loo}}\left(\widehat{f}_m; D_n\right) = \frac{1}{n}\sum_{j=1}^{n} c\left(\widehat{f}_m^{(-j)}(X_j); Y_j\right)$$

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

Cross-validation: examples

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

$$\widehat{\mathcal{R}}^{\mathrm{loo}}\left(\widehat{f}_m; D_n\right) = \frac{1}{n}\sum_{j=1}^{n} c\left(\widehat{f}_m^{(-j)}(X_j); Y_j\right)$$

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

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

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

Cross-validation: examples

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

$$\widehat{\mathcal{R}}^{\mathrm{loo}}\left(\widehat{f}_m; D_n\right) = \frac{1}{n}\sum_{j=1}^{n} c\left(\widehat{f}_m^{(-j)}(X_j); Y_j\right)$$

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

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

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

- Monte-Carlo CV / Repeated learning testing:

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

Outline

Bias of cross-validation

- In this talk, we always assume: $\forall j$, $\mathrm{Card}\big(D_n^{(t,j)}\big) = n_t$
  For V-fold CV: $\mathrm{Card}(B_j) = n/V \Rightarrow n_t = n(V-1)/V$.
- Ideal criterion: $\mathcal{R}\big(\widehat{f}_m(D_n)\big)$

Bias of cross-validation

- In this talk, we always assume: $\forall j$, $\mathrm{Card}\big(D_n^{(t,j)}\big) = n_t$
  For V-fold CV: $\mathrm{Card}(B_j) = n/V \Rightarrow n_t = n(V-1)/V$.
- Ideal criterion: $\mathcal{R}\big(\widehat{f}_m(D_n)\big)$
- General analysis for the bias:

$$\mathbb{E}\left[\widehat{\mathcal{R}}^{\mathrm{cv}}\left(\widehat{f}_m; D_n; \left(I_j^{(t)}\right)_{1 \leqslant j \leqslant V}\right)\right] = \mathbb{E}\Big[\mathcal{R}\big(\widehat{f}_m(D_{n_t})\big)\Big]$$

Bias of cross-validation

- In this talk, we always assume: $\forall j$, $\text{Card}(D_n^{(t,j)}) = n_t$
  For V-fold CV: $\text{Card}(B_j) = n/V \Rightarrow n_t = n(V-1)/V$.
- Ideal criterion: $\mathcal{R}(\widehat{f}_m(D_n))$
- General analysis for the bias:

$$\mathbb{E}\left[\widehat{\mathcal{R}}^{\text{cv}}\left(\widehat{f}_m; D_n; \left(I_j^{(t)}\right)_{1 \leqslant j \leqslant V}\right)\right] = \mathbb{E}\left[\mathcal{R}(\widehat{f}_m(D_{n_t}))\right]$$

$\Rightarrow$ everything depends on $n \to \mathbb{E}\left[\mathcal{R}(\widehat{f}_m(D_n))\right]$

Bias of cross-validation

- In this talk, we always assume: $\forall j$, $\text{Card}(D_n^{(t,j)}) = n_t$
  For V-fold CV: $\text{Card}(B_j) = n/V \Rightarrow n_t = n(V-1)/V$.
- Ideal criterion: $\mathcal{R}(\widehat{f}_m(D_n))$
- General analysis for the bias:

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

$\Rightarrow$ everything depends on $n \to \mathbb{E}\left[\mathcal{R}(\widehat{f}_m(D_n))\right]$

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

Bias of cross-validation: generic example

Assume:

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

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

Bias of cross-validation: generic example

Assume:

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

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

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

$\Rightarrow$ 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: general case

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

$$\operatorname{var}\left(\widehat{\mathcal{R}}_n^{(v)}\left(\widehat{f}_m^{(t)}\right)\right) = \frac{1}{n_v}\mathbb{E}\left[\operatorname{var}\left(c(f(X),Y) \,\middle|\, f = \widehat{f}_m^{(t)}\right)\right] + \operatorname{var}\left(\mathcal{R}(\widehat{f}_m(D_{n_t}))\right)$$

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

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

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

Variance of V-fold CV criterion

- Least-squares density estimation (A. & Lerasle, 2016), exact computation (non-asymptotic):

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

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

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

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

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

Outline

Risk estimation and estimator selection are different goals

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

- For any Z (deterministic or random),

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

$\Rightarrow$ bias and variance meaningless.

Risk estimation and estimator selection are different goals

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

- For any Z (deterministic or random),

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

$\Rightarrow$ bias and variance meaningless.

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

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

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

$\Rightarrow$ $\operatorname{var}\left(\widehat{\mathcal{R}}^{\mathrm{cv}}(\widehat{f}_m) - \widehat{\mathcal{R}}^{\mathrm{cv}}(\widehat{f}_{m'}) \right)$ should be minimal (detailed heuristic: A. & Lerasle, 2016)

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 $V \gg n^2$
  or if n_v large enough (including hold-out)
- CV performs similarly to

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

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 $V \gg n^2$
  or if n_v large enough (including hold-out)
- CV performs similarly to

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

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

⇒ 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):

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

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

Bias-corrected VFCV / V-fold penalization

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

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

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

- In least-squares density estimation (A. & Lerasle, 2016):

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

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

Variance and estimator selection

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

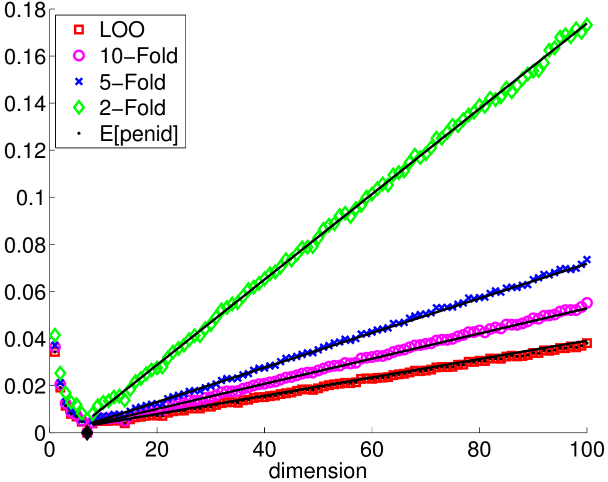
Theorem (A. & Lerasle, 2016, least-squares density estimation)

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

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

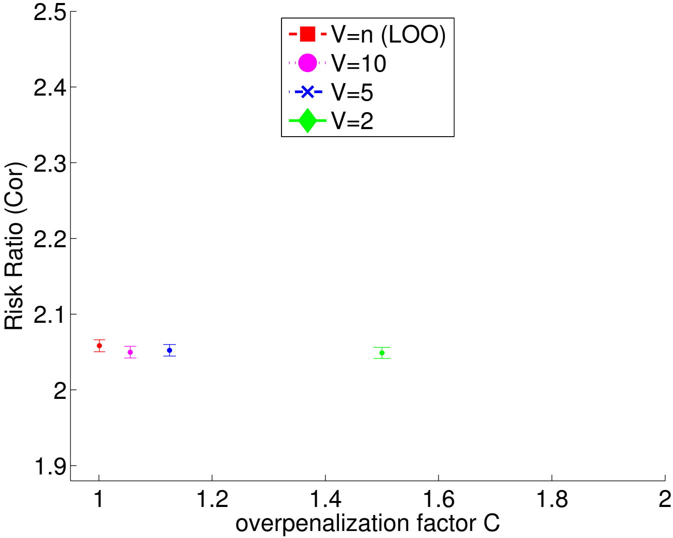
The two terms are of the same order if $\left\| f_m^\star - f_{m'}^\star \right\| \approx d_m / n$.

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

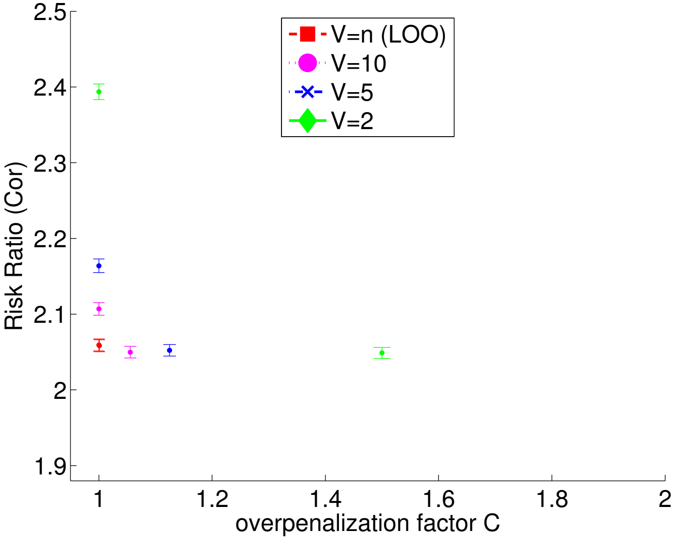


$$\mathrm{var}(\Delta(m, m', V)) \approx n^{-2}\left[29\left(1 + \frac{0.8}{V-1}\right) + 3.7\left(1 + \frac{3.8}{V-1}\right)(d_m - d_{m^\star})\right]$$

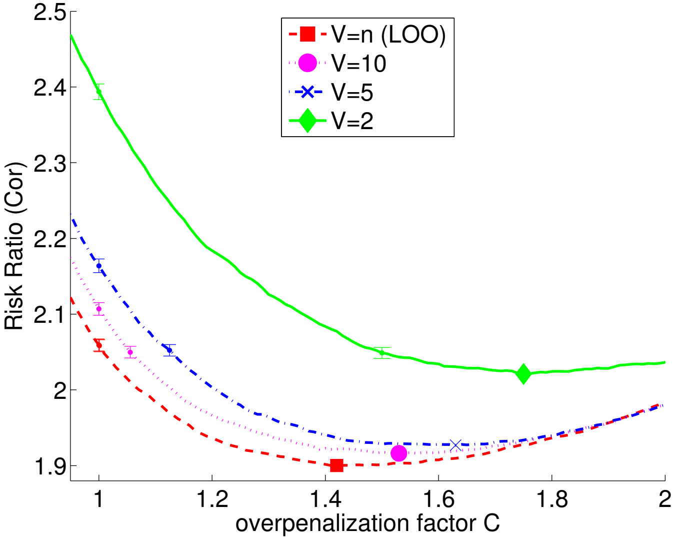
Experiment (LS density estimation): V-fold CV



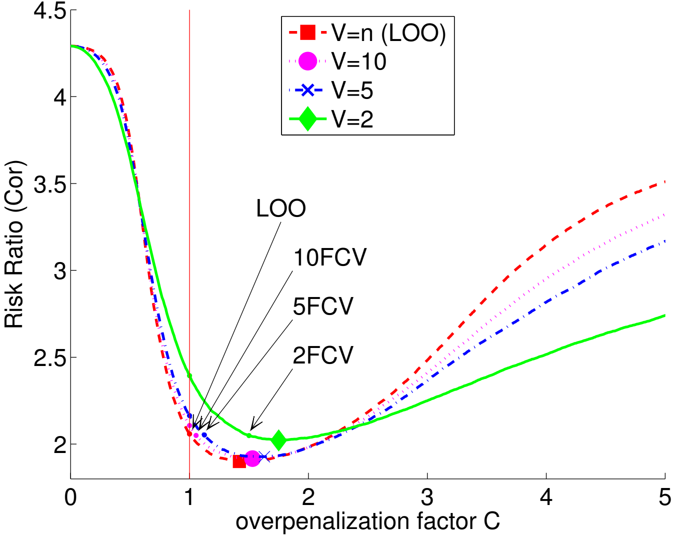
Experiment (LS density estimation): V-fold penalization



Experiment (LS density estimation): overpenalization



Experiment (LS density estimation): conclusion



Outline

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 $V \in [5, 10]$.

How general are these conclusions? (i.i.d. case)

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

$$\mathbb{E}\Big[(\mathcal{R} - \widehat{\mathcal{R}}_n)(\widehat{f}_m)\Big] \approx \frac{\gamma(m)}{n} \enspace .$$

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

$$n \to \mathbb{E}\Big[\mathcal{R}\big(\widehat{f}_m(D_n)\big)\Big] \qquad \text{and} \qquad m \to \operatorname{var}\Big(\widehat{\mathcal{R}}^{\mathrm{cv}}(\widehat{f}_m) - \widehat{\mathcal{R}}^{\mathrm{cv}}(\widehat{f}_{m^\star})\Big) \enspace .$$

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.

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.

Large collection of estimators/models

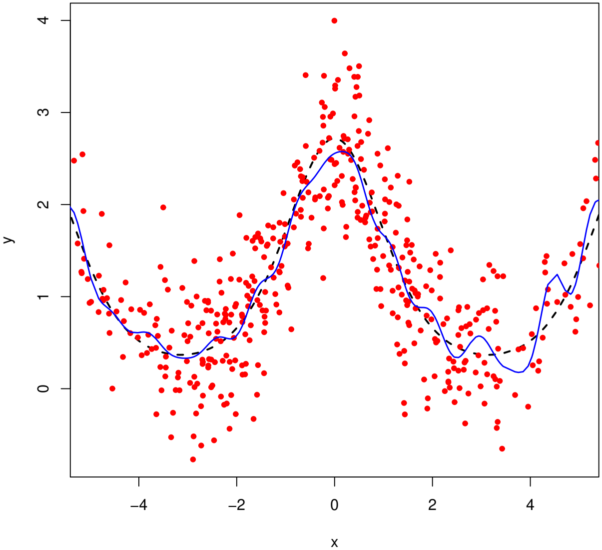
- Estimator/model selection with an "exponential collection" (implicitly excluded in all results above).
$\Rightarrow$ Expectations do not drive the first order!

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

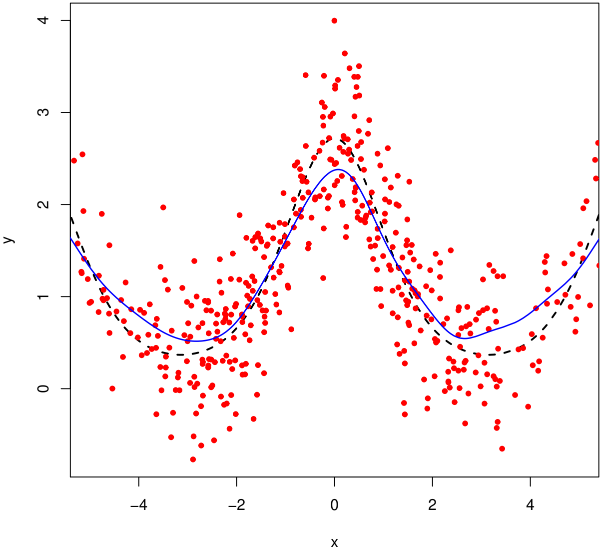
- Solution: group the models $\Rightarrow$ one estimator per "dimension" (e.g., empirical risk minimizer)
works for change-point detection (A. & Celisse, 2010).

Outline

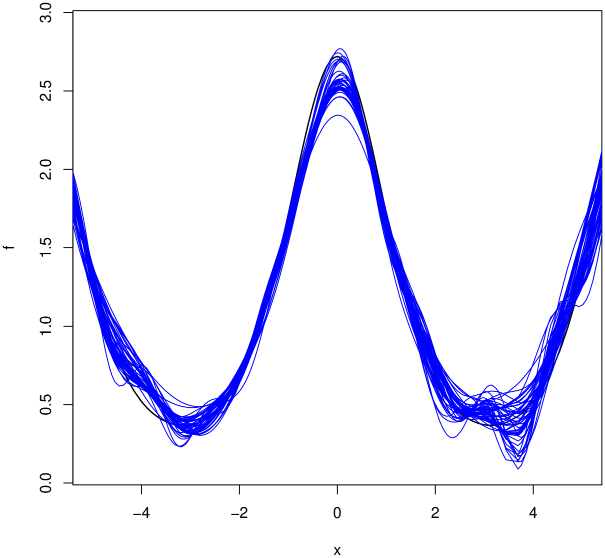
Example: regression, ϵ-SVM estimator (undersmoothed)



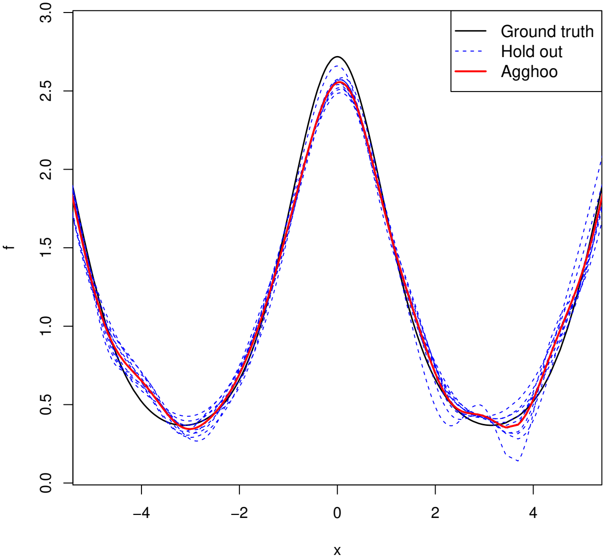
Example: regression, ϵ-SVM estimator (oversmoothed)



Example: regression, ϵ-SVM: hold-out estimators



Example: regression, ϵ-SVM: aggregated hold-out



Aggregated hold-out (Agghoo): definition

- Idea: aggregate several hold-out estimators.
- If $\mathcal{Y}$ is convex (e.g., regression):

$$\widehat{f}^{agghoo} = \frac{1}{V} \sum_{j=1}^{V} \widehat{f}_{\widehat{m}^{ho}(I_j^{(t)})} \left(D_n^{(t,j)} \right)$$

Aggregated hold-out (Agghoo): definition

- Idea: aggregate several hold-out estimators.
- If $\mathcal{Y}$ is convex (e.g., regression):

$$\widehat{f}^{agghoo} = \frac{1}{V} \sum_{j=1}^{V} \widehat{f}_{\widehat{m}^{ho}(I_j^{(t)})} \left(D_n^{(t,j)} \right)$$

- If $\mathcal{Y}$ is finite (classification):

$$\widehat{f}^{agghoo} : x \mapsto \text{majority vote among } \left\{ \widehat{f}_{\widehat{m}^{ho}(I_j^{(t)})} \left(x; D_n^{(t,j)} \right) / j = 1, \ldots, V \right\}$$

Aggregated hold-out (Agghoo): definition

- Idea: aggregate several hold-out estimators.
- If $\mathcal{Y}$ is convex (e.g., regression):

$$\widehat{f}^{agghoo} = \frac{1}{V} \sum_{j=1}^{V} \widehat{f}_{\widehat{m}^{ho}(I_j^{(t)})}\left(D_n^{(t,j)}\right)$$

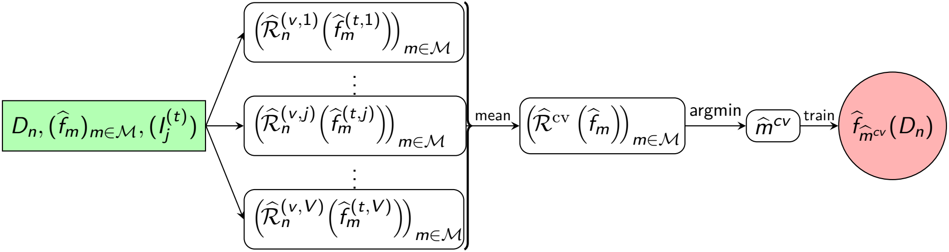
- If $\mathcal{Y}$ is finite (classification):

$$\widehat{f}^{agghoo} : x \mapsto \text{majority vote among } \left\{ \widehat{f}_{\widehat{m}^{ho}(I_j^{(t)})}\left(x; D_n^{(t,j)}\right) \,/\, j = 1, \ldots, V \right\}$$

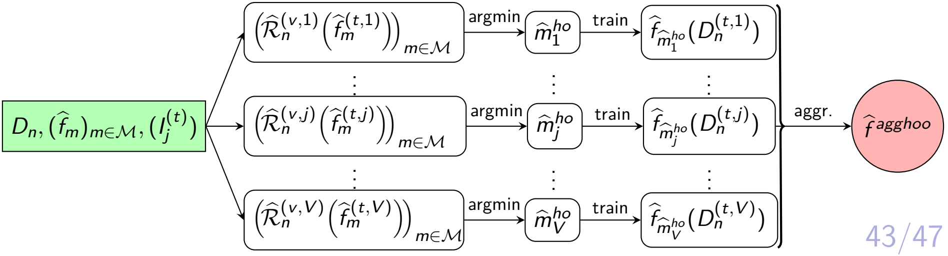
- Usual assumption: $\forall j \in \{1, \ldots, V\}$, $\mathrm{Card}(I_j^{(t)}) = \tau n$.
- Remark: $V = 1 \Rightarrow \widehat{f}^{agghoo} = \widehat{f}_{\widehat{m}^{ho}}(D_n^t) \approx$ hold-out estimator

Agghoo and cross-validation

Cross-validation:



Agghoo:



Related procedures

- CV bagging (data science folklore)
  - hold-out + subagging $\neq$ agghoo: for $j = 1, \ldots, V$,

hold-out + subagging = hold-out on subsamples

$$\underbrace{(X_{i_1}, Y_{i_1}), \ldots, (X_{i_k}, Y_{i_k})}_{\text{train}}, \underbrace{(X_{i_{k+1}}, Y_{i_{k+1}}), \ldots, (X_{i_\ell}, Y_{i_\ell})}_{\text{validation}}, \underbrace{(X_{i_{\ell+1}}, Y_{i_{\ell+1}}), \ldots, (X_{i_n}, Y_{i_n})}_{\text{unused}}$$

agghoo = hold-out on different splits

$$\underbrace{(X_{i_1}, Y_{i_1}), \ldots, (X_{i_k}, Y_{i_k}), (X_{i_{k+1}}, Y_{i_{k+1}}), \ldots, (X_{i_\ell}, Y_{i_\ell})}_{\text{train}}, \underbrace{(X_{i_{\ell+1}}, Y_{i_{\ell+1}}), \ldots, (X_{i_n}, Y_{i_n})}_{\text{validation}}$$

  - "CV bagging" also used for procedures close to agghoo

- Averaging of the chosen *parameters* $\widehat{m}_j^{ho}$:
  - K-fold averaging cross-validation (ACV; Jung and Hu, 2015)
  - efficient K-fold cross-validation (EKCV; Jung, 2016)

Performance of agghoo: theory (sanity check)

- Regression with c convex: if $\forall j$, $\mathrm{Card}(I_j^{(t)}) = \tau n$,

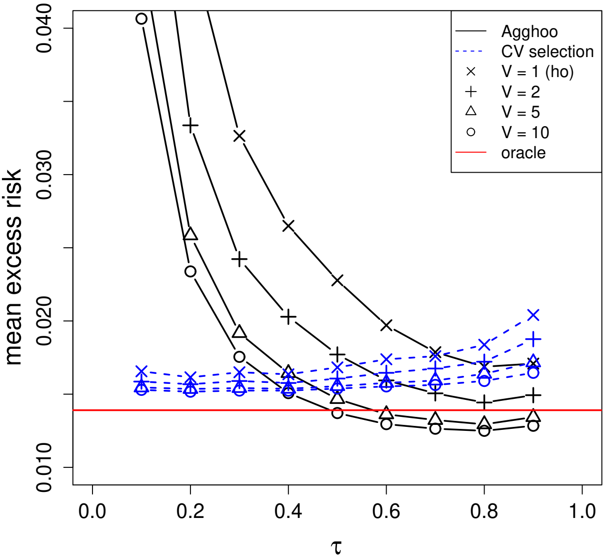
$$\forall V \geqslant 1, \qquad \mathbb{E}\left[\mathcal{R}\left(\widehat{f}^{agghoo}\left((\widehat{f}_m)_{m\in\mathcal{M}}; D_n; (I_j^{(t)})_{1\leqslant j\leqslant V}\right)\right)\right] \leqslant \mathbb{E}\left[\mathcal{R}\left(\widehat{f}_{\widehat{m}^{ho}\left((\widehat{f}_m)_{m\in\mathcal{M}}; D_n; I_1^{(t)}\right)}(D_n^{t,1})\right)\right]$$

Corollary: oracle inequalities for the hold-out $\Rightarrow$ oracle inequalities for agghoo

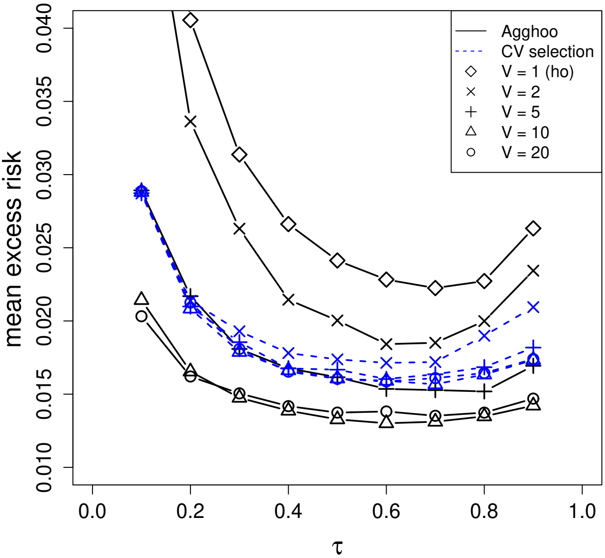
- Binary classification, 0–1 risk (Maillard, A. & Lerasle, 2017):

$$\forall V \geqslant 1, \qquad \mathbb{E}\left[\mathcal{R}\left(\widehat{f}^{agghoo}\left((\widehat{f}_m)_{m\in\mathcal{M}}; D_n; (I_j^{(t)})_{1\leqslant j\leqslant V}\right)\right) - \mathcal{R}(f^\star)\right] \leqslant 2\mathbb{E}\left[\mathcal{R}\left(\widehat{f}_{\widehat{m}^{ho}\left((\widehat{f}_m)_{m\in\mathcal{M}}; D_n; I_1^{(t)}\right)}(D_n^{t,1})\right) - \mathcal{R}(f^\star)\right]$$

Numerical experiments: regression, L^1 loss, ϵ-SVM



Numerical experiments: 0–1 binary classification, k-NN



Questions?