Model selection and estimator selection for statistical learning

Sylvain Arlot

[1] CNRS

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

Scuola Normale Superiore di Pisa, 14–23 February 2011

Outline of the 5 lectures

1. Statistical learning
2. Model selection for least-squares regression
3. Linear estimator selection for least-squares regression
4. Resampling and model selection
5. Cross-validation and model/estimator selection

Part V

Cross-validation and model/estimator selection

Outline

Outline

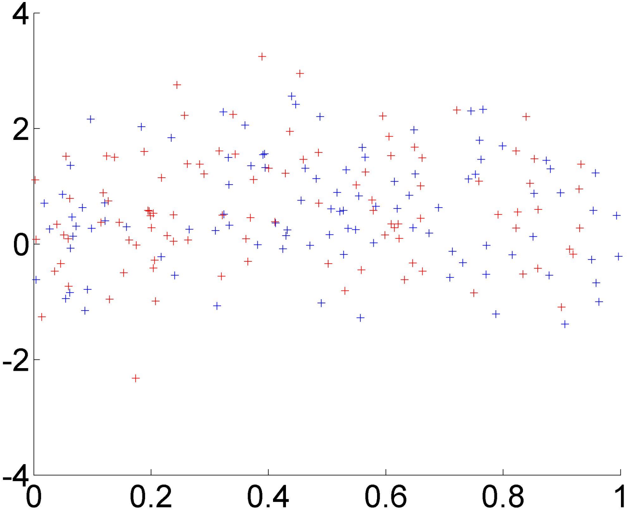
Reminder

- Data: $D_n = (\xi_1, \ldots, \xi_n) \in \Xi^n$, $D_n \sim P^{\otimes n}$
- Excess loss

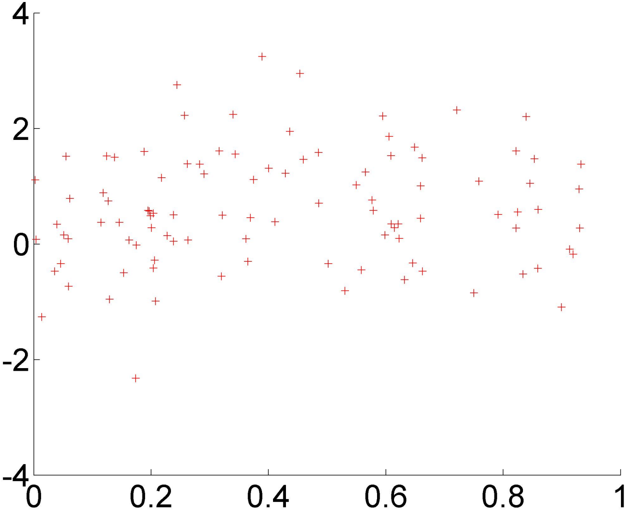
$$\ell\left(s^\star, t\right) = P\gamma(t) - P\gamma(s^\star)$$

- Statistical algorithms: $\forall m \in \mathcal{M}_n$, $\mathcal{A}_m$: $\bigcup_{n \in \mathbb{N}} \Xi^n \mapsto \mathbb{S}$
  $\mathcal{A}_m(D_n) = \widehat{s}_m(D_n) \in \mathbb{S}$ is an estimator of $s^\star$
- Estimation/prediction goal: find $\widehat{m}(D_n) \in \mathcal{M}$ such that $\ell\left(s^\star, \widehat{s}_{\widehat{m}(D_n)}(D_n)\right)$ is minimal

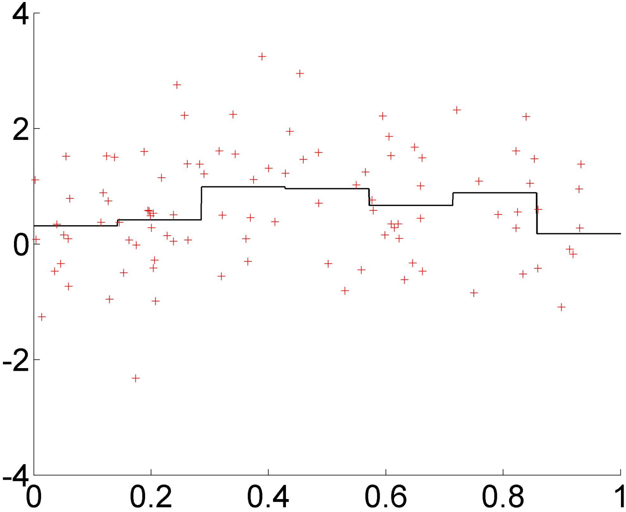
Hold-out



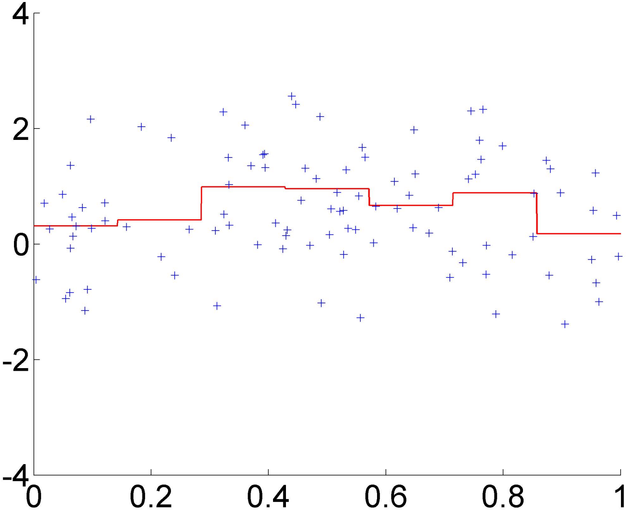
Hold-out: training sample



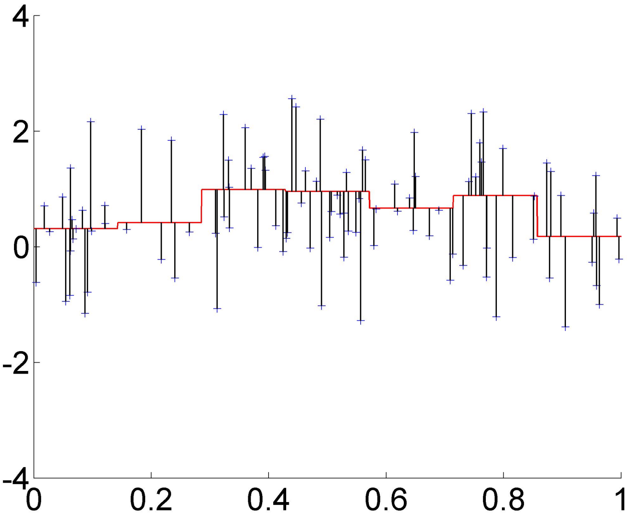
Hold-out: training sample



Hold-out: validation sample



Hold-out: validation sample



Cross-validation heuristics: hold-out

$$\underbrace{\xi_1 \; , \; \cdots \; , \; \xi_{n_t}}_{\text{Training } (I^{(t)})} \quad , \quad \underbrace{\xi_{n_t+1} \; , \; \cdots \; , \; \xi_n}_{\text{Validation } (I^{(v)})}$$

$$\widehat{s}_m^{(t)} := \mathcal{A}_m \left(D_n^{(t)} \right) \quad \text{where} \quad D_n^{(t)} := (\xi_i)_{i \in I^{(t)}}$$

$$P_n^{(v)} = \frac{1}{n_v} \sum_{i \in I^{(v)}} \delta_{\xi_i} \qquad n_v := n - n_t$$

$$\Rightarrow \widehat{\mathcal{R}}^{\mathrm{val}} \left(\mathcal{A}_m ; D_n ; I^{(t)} \right) = P_n^{(v)} \gamma \left(\widehat{s}_m^{(t)} \right) = \frac{1}{n_v} \sum_{i \in I^{(v)}} \gamma \left(\mathcal{A}_m \left(D_n^{(t)} \right) ; \xi_i \right)$$

General definition of cross-validation

- $B \geq 1$ training sets:

$$I_1^{(t)}, \ldots, I_B^{(t)} \subset \{1, \ldots, n\}$$

- Cross-validation estimator of the risk of $\mathcal{A}_m$:

$$\widehat{\mathcal{R}}^{\mathrm{vc}} \left(\mathcal{A}_m; D_n; \left(I_j^{(t)} \right)_{1 \leq j \leq B} \right) := \frac{1}{B} \sum_{j=1}^{B} \widehat{\mathcal{R}}^{\mathrm{val}} \left(\mathcal{A}_m; D_n; I_j^{(t)} \right)$$

- Chosen algorithm:

$$\widehat{m} \in \operatorname{argmin}_{m \in \mathcal{M}_n} \left\{ \widehat{\mathcal{R}}^{\mathrm{vc}} \left(\mathcal{A}_m; D_n; \left(I_j^{(t)} \right)_{1 \leq j \leq B} \right) \right\}$$

- Usually, $\forall j$, $\mathrm{Card}(I_j^{(t)}) = n_t$

Example: exhaustive data splitting

- Leave-one-out (LOO), or delete-one CV, or ordinary cross-validation:

$$n_t = n - 1 \qquad B = n$$

(Stone, 1974; Allen, 1974; Geisser, 1975)

- Leave-p-out (LPO), or delete-p CV:

$$n_t = n - p \qquad B = \binom{n}{p}$$

Examples: partial data-splitting

- *V*-fold cross-validation (VFCV, Geisser, 1975):
  $\mathcal{B} = (B_j)_{1 \leq j \leq V}$ partition of $\{1, \ldots, n\}$

$$\widehat{\mathcal{R}}^{\mathrm{vf}}\left(\mathcal{A}_m; D_n; \mathcal{B}\right) = \frac{1}{V} \sum_{j=1}^{V} \widehat{\mathcal{R}}^{\mathrm{val}}\left(\mathcal{A}_m; D_n; B_j^c\right)$$

- Repeated Learning-Testing (RLT, Breiman *et al.*, 1984):
  $I_1^{(t)}, \ldots, I_B^{(t)} \subset \{1, \ldots, n\}$ of cardinality n_t, sampled uniformly without replacement
- Monte-Carlo cross-validation (MCCV, Picard & Cook, 1984):
  same with $I_1^{(t)}, \ldots, I_B^{(t)}$ of cardinality n_t, sampled uniformly with replacement (i.i.d.)

Related procedures

- Generalized cross-validation (GCV): rotation-invariant version of LOO for linear regression, closer to C_p and C_L than to cross-validation (Efron, 1986, 2004)
- Analytical approximation to leave-p-out (Shao, 1993)
- Leave-one-out bootstrap (Efron, 1983):
  stabilized version of leave-one-out
  heuristical bias-correction $\Rightarrow$ .632 bootstrap
  $\Rightarrow$ .632+ bootstrap (Efron & Tibshirani, 1997)

Bias of the cross-validation estimator

- Target: $P\gamma\left(\mathcal{A}_m(D_n)\right)$
- Bias: if $\forall j$, $\mathrm{Card}(I_j^{(t)}) = n_t$

Bias of the cross-validation estimator

- Target: $P\gamma\left(\mathcal{A}_m(D_n)\right)$
- Bias: if $\forall j$, $\mathrm{Card}(I_j^{(t)}) = n_t$

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

$\Rightarrow$ bias $\mathbb{E}\left[P\gamma\left(\mathcal{A}_m\left(D_{n_t}\right)\right)\right] - \mathbb{E}\left[P\gamma\left(\mathcal{A}_m\left(D_n\right)\right)\right]$

- Smart rule (Devroye, Györfi & Lugosi, 1996):
  $n \mapsto \mathbb{E}\left[P\gamma\left(\mathcal{A}_m\left(D_n\right)\right)\right]$ non-increasing
  $\Rightarrow$ the bias is non-negative, minimal for $n_t = n - 1$
- Example: regressogram:

$$\mathbb{E}\left[P\gamma(\widehat{s}_m(D_n))\right] \approx P\gamma(s_m^\star) + \frac{1}{n}\sum_{\lambda \in m} \sigma_\lambda^2$$

Bias-correction

- Corrected V-fold cross-validation (Burman, 1989, 1990):

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

$+$ the same for Repeated Learning-Testing

- Asymptotical result: bias $= \mathcal{O}(n^{-2})$ (Burman, 1989)

Variability of the cross-validation estimator

$$\operatorname{var}\left[\widehat{\mathcal{R}}^{\mathrm{vc}}\left(\mathcal{A}_m; D_n; \left(I_j^{(t)}\right)_{1 \le j \le B}\right)\right]$$

Variability sources:

Variability of the cross-validation estimator

$$\operatorname{var}\left[\widehat{\mathcal{R}}^{\mathrm{vc}}\left(\mathcal{A}_m; D_n; \left(I_j^{(t)}\right)_{1\leq j\leq B}\right)\right]$$

Variability sources:

- (n_t, n_v) : hold-out case (Nadeau & Bengio, 2003)

$$\begin{aligned}
&\operatorname{var}\left[\widehat{\mathcal{R}}^{\mathrm{val}}\left(\mathcal{A}_m; D_n; I^{(t)}\right)\right] \\
&= \mathbb{E}\left[\operatorname{var}\left(P_n^{(v)}\gamma\left(\mathcal{A}_m(D_n^{(t)})\right) \,\middle|\, D_n^{(t)}\right)\right] + \operatorname{var}\left[P\gamma\left(\mathcal{A}_m(D_{n_t})\right)\right] \\
&= \frac{1}{n_v}\mathbb{E}\left[\operatorname{var}\left(\gamma\left(\widehat{s},\xi\right) \mid \widehat{s} = \mathcal{A}_m(D_n^{(t)})\right)\right] + \operatorname{var}\left[P\gamma\left(\mathcal{A}_m(D_{n_t})\right)\right]
\end{aligned}$$

Variability of the cross-validation estimator

$$\operatorname{var}\left[\widehat{\mathcal{R}}^{\mathrm{vc}} \left(\mathcal{A}_m ; D_n ; \left(I_j^{(t)} \right)_{1 \le j \le B} \right) \right]$$

Variability sources:

- (n_t, n_v) : hold-out case (Nadeau & Bengio, 2003)

$$\begin{aligned} & \operatorname{var}\left[\widehat{\mathcal{R}}^{\mathrm{val}} \left(\mathcal{A}_m ; D_n ; I^{(t)} \right) \right] \\ & = \mathbb{E}\left[\operatorname{var}\left(P_n^{(v)} \gamma \left(\mathcal{A}_m(D_n^{(t)}) \right) \,\middle|\, D_n^{(t)} \right) \right] + \operatorname{var}\left[P\gamma\left(\mathcal{A}_m(D_{n_t}) \right) \right] \\ & = \frac{1}{n_v} \mathbb{E}\left[\operatorname{var}\left(\gamma\left(\widehat{s}, \xi \right) \mid \widehat{s} = \mathcal{A}_m(D_n^{(t)}) \right) \right] + \operatorname{var}\left[P\gamma\left(\mathcal{A}_m(D_{n_t}) \right) \right] \end{aligned}$$

- Stability of $\mathcal{A}_m$ (Bousquet & Elisseff, 2002)
- Number of splits B

- Problem: B, n_t, n_v linked for VFCV and LPO

Results on variability

- Linear regression, least-squares, special case (Burman, 1989):

$$\frac{2\sigma^2}{n} + \frac{4\sigma^4}{n^2}\left[4 + \frac{4}{V-1} + \frac{2}{(V-1)^2} + \frac{1}{(V-1)^3}\right] + o\left(n^{-2}\right)$$

Results on variability

- Linear regression, least-squares, special case (Burman, 1989):

$$\frac{2\sigma^2}{n} + \frac{4\sigma^4}{n^2}\left[4 + \frac{4}{V-1} + \frac{2}{(V-1)^2} + \frac{1}{(V-1)^3}\right] + o\left(n^{-2}\right)$$

- Explicit quantification in regression (LPO) and density estimation (VFCV, LPO): Celisse (2008)
- LOO quite variable when $\mathcal{A}_m$ is unstable (e.g., k-NN or CART), much less when $\mathcal{A}_m$ is stable (e.g., least-squares estimators; see Molinaro *et al.*, 2005)
- Data-driven estimation of the variability of cross-validation difficult: no universal unbiased estimator (RLT, Nadeau & Bengio, 2003; VFCV, Bengio & Grandvalet, 2004), several estimators proposed (ibid.; Markatou *et al.*, 2005; Celisse & Robin, 2008)

Outline

Link between risk estimation and estimator selection

- Unbiased risk estimation principle
  $\Rightarrow$ the important quantity (asymptotically) is the bias
- What is the best criterion?
  In principle, the best $\widehat{m}$ is the minimizer of the best risk estimator.
- Sometimes more tricky (Breiman & Spector, 1992):
  - Only m "close" to the oracle $m^{\star}$ really count
  - Overpenalization sometimes necessary (many models or small signal-to-noise ratio)

Reminder: key lemma

Lemma

On the event Ω where for every $m, m' \in \mathcal{M}_n$,

$$\left(\mathrm{crit}(m) - P\gamma\left(\widehat{s}_m(D_n)\right)\right) - \left(\mathrm{crit}(m') - P\gamma\left(\widehat{s}_{m'}(D_n)\right)\right) \\ \leq A(m) + B(m')$$

$$\forall \widehat{m} \in \mathrm{argmin}_{m \in \mathcal{M}_n} \left\{ \mathrm{crit}(m) \right\}$$

$$\ell\left(s^\star, \widehat{s}_{\widehat{m}}(D_n)\right) - B(\widehat{m}) \leq \inf_{m \in \mathcal{M}_n} \left\{ \ell\left(s^\star, \widehat{s}_m(D_n)\right) + A(m) \right\}$$

Cross-validation for prediction: key role of n_t

Linear regression framework (Shao, 1997) representative of the general behaviour of cross-validation:

- If $n_t \sim n$, asymptotic optimality (CV $\sim C_p$)
- If $n_t \sim \kappa n$, $\kappa \in (0, 1)$, CV $\sim \mathrm{GIC}_{1+\kappa^{-1}}$ (i.e., overpenalizes from a factor $(1 + \kappa^{-1})/2 \Rightarrow$ asymptotically sub-optimal)

$\Rightarrow$ valid for LPO (Shao, 1997), RLT (if $B \gg n^2$, Zhang, 1993)

Sub-optimality of V-fold cross-validation

- $Y = X + \sigma\varepsilon$ with ε bounded and $\sigma > 0$
- $\mathcal{M} = \mathcal{M}_n^{(\mathrm{reg})}$ (regular histograms over $\mathcal{X} = [0,1]$)
- $\widehat{m}$ obtained by V-fold cross-validation with a fixed V as n increases

Theorem (A., 2008)

With probability $1 - Ln^{-2}$,

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

where $\kappa(V) > 0$

Oracle inequalities for cross-validation

- If $n_v \to \infty$ fast enough, one can "easily" prove the hold-out performs at least as well as

$$\operatorname{argmin}_{m \in \mathcal{M}_n} \{ P\gamma (\mathcal{A}_m(D_{n_t})) \}$$

- van der Laan, Dudoit & van der Vaart (2006): same property for LPO, VFCV and MCCV in a fairly general setting

Oracle inequalities for cross-validation

- If $n_v \to \infty$ fast enough, one can "easily" prove the hold-out performs at least as well as

$$\operatorname{argmin}_{m \in \mathcal{M}_n} \{ P\gamma (\mathcal{A}_m(D_{n_t})) \}$$

- van der Laan, Dudoit & van der Vaart (2006): same property for LPO, VFCV and MCCV in a fairly general setting
- Regressograms: VFCV suboptimal, but still adaptive to heteroscedasticity (up to a multiplicative factor $C(V) > 1$)
- LPO in regression and density estimation when $p/n \in [a, b]$, $0 < a < b < 1$ (Celisse, 2008)

Oracle inequalities for cross-validation

- If $n_v \to \infty$ fast enough, one can "easily" prove the hold-out performs at least as well as

$$\operatorname{argmin}_{m \in \mathcal{M}_n} \left\{ P\gamma \left(\mathcal{A}_m(D_{n_t}) \right) \right\}$$

- van der Laan, Dudoit & van der Vaart (2006): same property for LPO, VFCV and MCCV in a fairly general setting
- Regressograms: VFCV suboptimal, but still adaptive to heteroscedasticity (up to a multiplicative factor $C(V) > 1$)
- LPO in regression and density estimation when $p/n \in [a, b]$, $0 < a < b < 1$ (Celisse, 2008)

- Open problem: theoretical comparison taking B into account (hence the variability of cross-validation)

Cross-validation for identification: problem

- Collection of algorithms $(\mathcal{A}_m)_{m \in \mathcal{M}}$
- Goal: identify the best one for analyzing a new sample of size $n' \to \infty$

$$m_0 \in \lim_{n' \to \infty} \operatorname{argmin}_{m \in \mathcal{M}} \left\{ \mathbb{E}\left[P\gamma\left(\mathcal{A}_m(D'_{n'}) \right) \right] \right\}$$

- Consistency:

$$\mathbb{P}\left(\widehat{m}(D_n) = m_0 \right) \xrightarrow[n \to \infty]{} 1$$

- Examples:
  - identification of the true model in model selection
  - parametric vs. non-parametric algorithm?
  - $\widehat{k}$-NN or SVM?
  - ...

Cross-validation with voting (Yang, 2006)

Two algorithms $\mathcal{A}_1$ and $\mathcal{A}_2$

- For $m = 1, 2$

$$\left(\widehat{\mathcal{R}}^{\mathrm{val}} \left(\mathcal{A}_m; D_n; I_j^{(t)} \right) \right)_{1 \leq j \leq B}$$

$\Rightarrow$ majority vote

$$\mathcal{V}_1(D_n) = \operatorname{Card} \left\{ j \text{ s.t. } \widehat{\mathcal{R}}^{\mathrm{val}} \left(\mathcal{A}_1; D_n; I_j^{(t)} \right) < \widehat{\mathcal{R}}^{\mathrm{val}} \left(\mathcal{A}_2; D_n; I_j^{(t)} \right) \right\}$$

$$\widehat{m} = \begin{cases} 1 & \text{if} \quad \mathcal{V}_1(D_n) > n/2 \\ 2 & \text{otherwise} \end{cases}$$

- Usual cross-validation: averaging before comparison

Cross-validation for identification: regression

- "Cross-validation paradox" (Yang, 2007)
- $r_{n,m}$: asymptotics of $\mathbb{E}\|\mathcal{A}_m(D_n) - s^\star\|_2$
- Goal: recover $\operatorname{argmin}_{m \in \mathcal{M}} r_{n,m}$
- Assumption: at least a factor $C > 1$ between $r_{n,1}$ and $r_{n,2}$

Cross-validation for identification: regression

- "Cross-validation paradox" (Yang, 2007)
- $r_{n,m}$: asymptotics of $\mathbb{E}\|\mathcal{A}_m(D_n) - s^\star\|_2$
- Goal: recover $\operatorname{argmin}_{m \in \mathcal{M}} r_{n,m}$
- Assumption: at least a factor $C > 1$ between $r_{n,1}$ and $r_{n,2}$
- VFCV, RLT, LPO (with voting) are (model) consistent if

$$n_v, n_t \to \infty \quad \text{and} \quad \sqrt{n_v} \max_{m \in \mathcal{M}} r_{n_t,m} \to \infty$$

under some conditions on $\left(\|\mathcal{A}_m(D_n) - s^\star\|_p\right)_{p=2,4,\infty}$

Cross-validation for identification: regression

- Parametric vs. parametric ($r_{n,m} \propto n^{-1/2}$)
  $\Rightarrow$ the condition becomes $n_v \gg n_t \to \infty$
- Non-parametric vs. (non-)parametric ($\max_{m \in \mathcal{M}} r_{n,m} \gg n^{-1/2}$)
  $\Rightarrow n_t / n_v = \mathcal{O}(1)$ is sufficient, and we can have $n_t \sim n$ (not too close)
- Intuition:
  - risk estimated with precision $\propto n_v^{-1/2}$
  - difference between risks of order $\max_{m \in \mathcal{M}} r_{n_t,m}$
    $\Rightarrow$ easier to distinguish algorithms with n_t small because the difference between the risks is larger (questionable in practice)

Cross-validation in practice: computational complexity

- Naive implementation: complexity proportional to B
  $\Rightarrow$ LPO untractable, LOO sometimes tractable
  $\Rightarrow$ VFCV, RLT and MCCV often better

Cross-validation in practice: computational complexity

- Naive implementation: complexity proportional to B
  $\Rightarrow$ LPO untractable, LOO sometimes tractable
  $\Rightarrow$ VFCV, RLT and MCCV often better

- Closed-form formulas for LPO in (least-squares) density estimation and regression (projection or kernel estimators): Celisse & Robin (2008), Celisse (2008)
  $\Rightarrow$ can be used for instance in change-point detection (with dynamic programming)
- Generalized cross-validation: generalization of a formula for LOO in linear regression

Cross-validation in practice: computational complexity

- Naive implementation: complexity proportional to B
  $\Rightarrow$ LPO untractable, LOO sometimes tractable
  $\Rightarrow$ VFCV, RLT and MCCV often better
- Closed-form formulas for LPO in (least-squares) density estimation and regression (projection or kernel estimators): Celisse & Robin (2008), Celisse (2008)
  $\Rightarrow$ can be used for instance in change-point detection (with dynamic programming)
- Generalized cross-validation: generalization of a formula for LOO in linear regression
- Without closed-form formulas, smart algorithms for LOO (linear discriminant analysis, Ripley, 1996; k-NN, Daudin & Mary-Huard, 2008): uses results obtained for previous data splits in order to avoid doing again part of the computations

Choosing among cross-validation methods

Trade-off between bias, variability and computational cost:

- Bias: increases as n_t decreases (except for bias-corrected methods)
  large SNR: the bias must be minimized
  small SNR: a small amount of bias is better ($\Rightarrow n_t = \kappa n$ for some $\kappa \in (0, 1)$)
- Variability: usually a decreasing function of B and with n_v, but it depends on the nature of algorithms considered (stability)
- Computational cost: proportional to B, except in some cases

VFCV: B and n_t functions of $V \Rightarrow$ complex problem ($V = 10$ is not always a good choice)

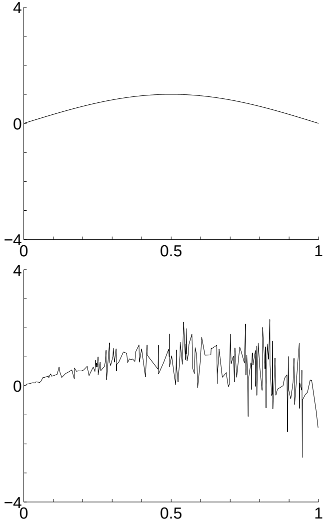
Choosing the training samples

- Usual advice: take into account a possible stratification of data, e.g.,
  - distribution of the X_i in the feature space (regression)
  - distribution of the Y_i among the classes (classification)
  - ...

  but no clear theoretical result (simulations by Breiman & Spector, 1992: unsignificant difference).

- Dependency between the $I_j^{(t)}$?
  Intuitively, better to give similar roles to all data in the training and validation tasks $\Rightarrow$ VFCV
  But no clear comparison between VFCV (strong dependency), RLT (weak dependency) and MCCV (independence).

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



Models: $\mathcal{M}_n = \mathcal{M}_n^{(\mathrm{reg},1/2)}$

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

computed with $N = 1000$ samples

Mallows	3.69 ± 0.07
2-fold	2.54 ± 0.05
5-fold	2.58 ± 0.06
10-fold	2.60 ± 0.06
20-fold	2.58 ± 0.06
leave-one-out	2.59 ± 0.06

Universality of cross-validation?

- Almost universal heuristics (i.i.d. data, no other explicit assumption)
- But $D_n \mapsto \mathcal{A}_{\widehat{m}(D_n)}$ still is a learning rule
  $\Rightarrow$ No Free Lunch Theorems apply
- Implicit assumptions of cross-validation:
  - generalization error well estimated from a finite number of points n_v
  - behaviour of the algorithm with n_t points representative from its behaviour with n points

  $+$ assumptions of the unbiased risk estimation principle

Dependent data

- cross-validation wrong in principle (assumes i.i.d.)
- Stationary Markov process ⇒ CV still works (Burman & Nolan, 1992)
- Positive correlations ⇒ can overfit (Hart & Wehrly, 1986; Opsomer *et al.*, 2001)

Dependent data

- cross-validation wrong in principle (assumes i.i.d.)
- Stationary Markov process $\Rightarrow$ CV still works (Burman & Nolan, 1992)
- Positive correlations $\Rightarrow$ can overfit (Hart & Wehrly, 1986; Opsomer *et al.*, 2001)
- Answer: for short range dependencies, choose $I^{(t)}$ and $I^{(v)}$ such that

$$\min_{i \in I^{(t)}, j \in I^{(v)}} |i - j| \geq h > 0$$

$\Rightarrow$ modified CV (Chu & Marron, 1991), h-block CV (can be bias-corrected, Burman *et al.*, 1994), and so on

Large collections of models

- Model selection in regression, exponential number of models per dimension $\Rightarrow$ minimal penalty of order $\ln(n) D_m / n$ (Birgé & Massart, 2007)

  $\Rightarrow$ cross-validation overfits (except maybe if $n_t \ll n$)

Large collections of models

- Model selection in regression, exponential number of models per dimension $\Rightarrow$ minimal penalty of order $\ln(n) D_m / n$ (Birgé & Massart, 2007)

  $\Rightarrow$ cross-validation overfits (except maybe if $n_t \ll n$)

- Wegkamp (2003): penalized hold-out
- A. & Celisse (2009): gather models of the same dimension, with application to change-point detection

Outline

Change-point detection and model selection

$$Y_i = \eta(t_i) + \sigma(t_i)\varepsilon_i \quad \text{with} \quad \mathbb{E}[\varepsilon_i] = 0 \quad \mathbb{E}\left[\varepsilon_i^2\right] = 1$$

- Goal: detect the change-points of the mean η of the signal Y

$\Rightarrow$ Model selection, collection of regressograms with $\mathcal{M}_n = \mathfrak{P}_{\mathrm{interv}}(\{ t_1, \ldots, t_n \})$ (partitions of $\mathcal{X}$ into intervals)

- Here: no assumption on the variance $\sigma(t_i)^2$

Classical approach (Lebarbier, 2005; Lavielle, 2005)

- "Birgé-Massart" penalty (assumes $\sigma(t_i) \equiv \sigma$):

$$\widehat{m} \in \operatorname{argmin}_{m \in \mathcal{M}_n} \left\{ P_n \gamma \left(\widehat{s}_m \right) + \frac{C \sigma^2 D_m}{n} \left(5 + 2 \ln \left(\frac{n}{D_m} \right) \right) \right\}$$

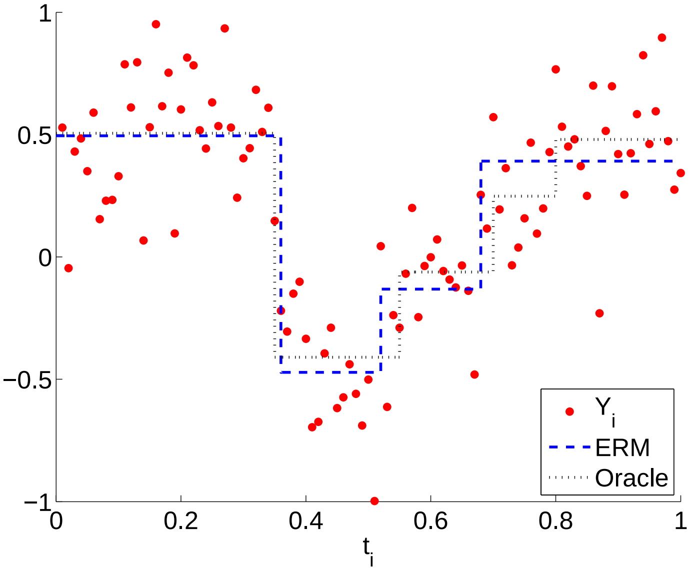
- Equivalent to aggregating models of the same dimension:

$$\widetilde{S}_D := \bigcup_{m \in \mathcal{M}_n \, , \, D_m = D} S_m$$

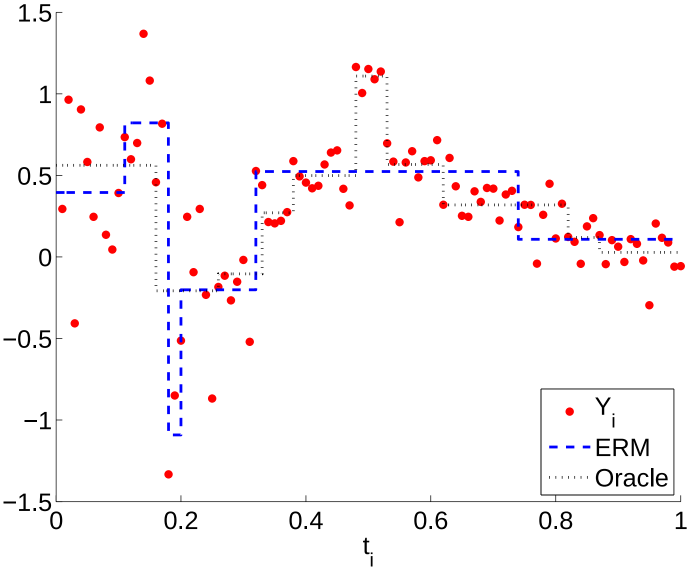
$$\widehat{s}_D \in \operatorname{argmin}_{t \in \widetilde{S}_D} \left\{ P_n \gamma \left(t \right) \right\} \qquad \text{dynamic programming}$$

$$\widehat{D} \in \operatorname{argmin}_{1 \leq D \leq n} \left\{ P_n \gamma \left(\widehat{s}_D \right) + \frac{C \sigma^2 D}{n} \left(5 + 2 \ln \left(\frac{n}{D} \right) \right) \right\}$$

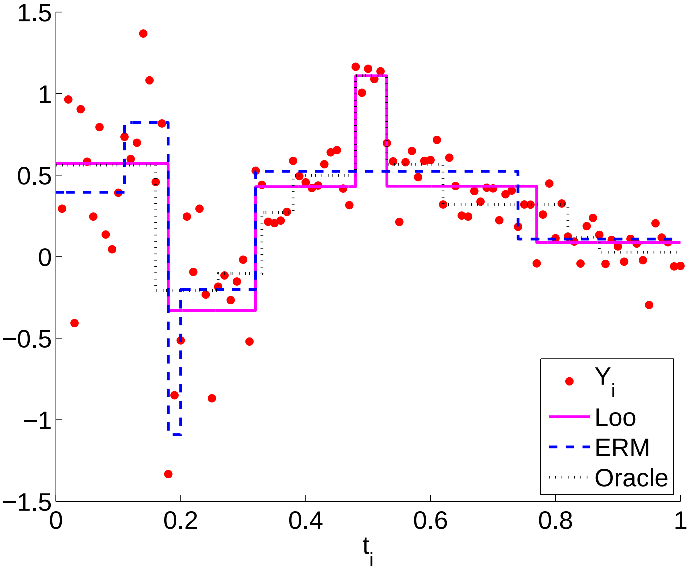
$D = 4$, homoscedastic; $n = 100$, $\sigma = 0.25$



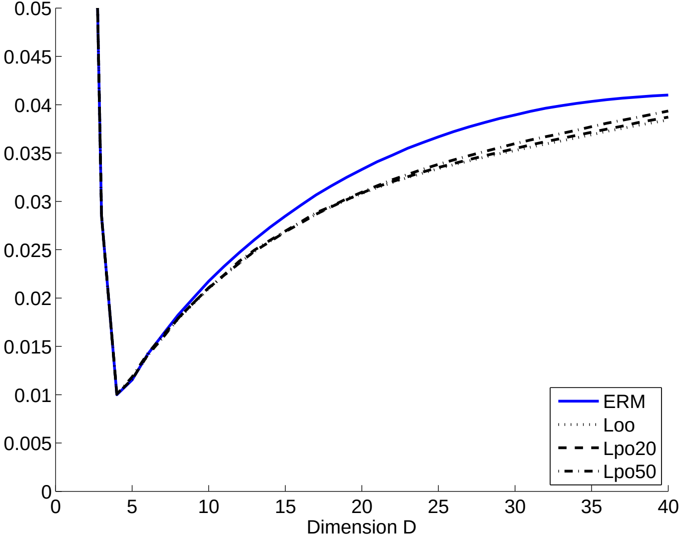
$D = 6$, heteroscedastic; $n = 100$, $\|\sigma\| = 0.30$



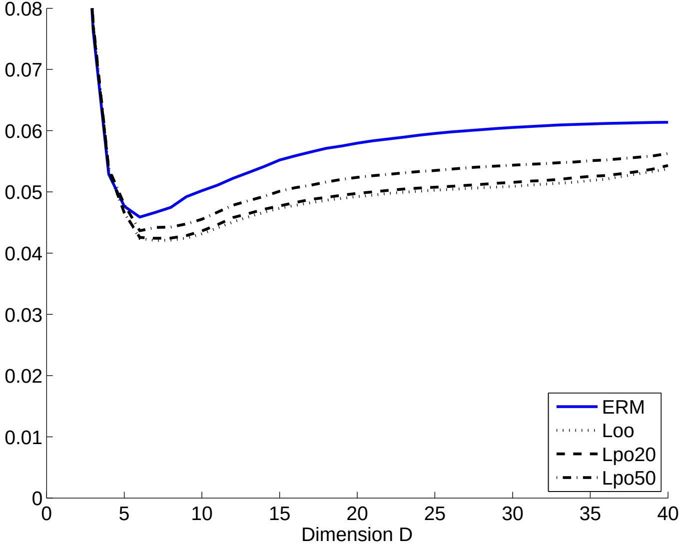
$D = 6$, heteroscedastic; $n = 100$, $\|\sigma\| = 0.30$



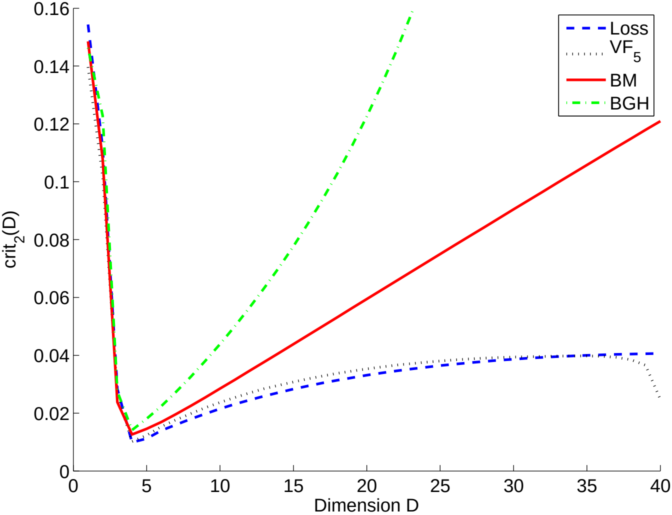
Homoscedastic: loss as a function of D



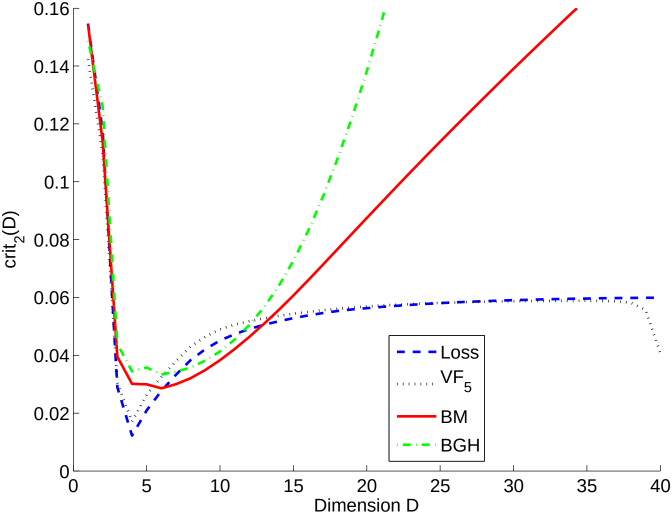
Heteroscedastic: loss as a function of D



Homoscedastic: estimate of the loss as a function of D



Heteroscedastic: estimate of the loss as a function of D



Change-point detection algorithms (A. & Celisse, 2010)

1. $\forall D \in \{1, \ldots, D_{\max}\}$, select

$$\widehat{m}(D) \in \operatorname{argmin}_{m \in \mathcal{M}_n,\, D_m = D} \left\{ \mathrm{crit}_1(m; (t_i, Y_i)_i) \right\}$$

Examples for crit_1: empirical risk, or leave-p-out or V-fold estimators of the risk (dynamic programming)

Change-point detection algorithms (A. & Celisse, 2010)

1. $\forall D \in \{1, \ldots, D_{\max}\}$, select

$$\widehat{m}(D) \in \operatorname{argmin}_{m \in \mathcal{M}_n,\, D_m = D} \left\{ \operatorname{crit}_1(m; (t_i, Y_i)_i) \right\}$$

   Examples for crit_1: empirical risk, or leave-p-out or V-fold estimators of the risk (dynamic programming)

2. Select

$$\widehat{D} \in \operatorname{argmin}_{D \in \{1, \ldots, D_{\max}\}} \left\{ \operatorname{crit}_2(D; (t_i, Y_i)_i; \operatorname{crit}_1(\cdot)) \right\}$$

   Examples for crit_2: penalized empirical criterion, V-fold estimator of the risk

Competitors

- $[\mathrm{Emp}, \mathrm{BM}]$: assume $\sigma(\cdot) \equiv \sigma$

$$\operatorname{argmin}_{m \in \mathcal{M}_n} \left\{ P_n \gamma \left(\widehat{s}_m \right) + \frac{C \widehat{\sigma}^2 D_m}{n} \left(5 + 2 \log \left(\frac{n}{D_m} \right) \right) \right\}$$

Competitors

- [Emp, BM]: assume $\sigma(\cdot) \equiv \sigma$

$$\operatorname{argmin}_{m \in \mathcal{M}_n} \left\{ P_n \gamma\left(\widehat{s}_m\right) + \frac{C \widehat{\sigma}^2 D_m}{n} \left(5 + 2 \log\left(\frac{n}{D_m} \right) \right) \right\}$$

- BGH (Baraud, Giraud & Huet 2009): multiplicative penalty, $\sigma(\cdot) \equiv \sigma$

$$\operatorname{argmin}_{m \in \mathcal{M}_n} \left\{ P_n \gamma\left(\widehat{s}_m\right) \left[1 + \frac{\mathrm{pen}_{\mathrm{BGH}}(m)}{n - D_m} \right] \right\}$$

Competitors

- [Emp, BM]: assume $\sigma(\cdot) \equiv \sigma$

$$\operatorname{argmin}_{m \in \mathcal{M}_n} \left\{ P_n \gamma \left(\widehat{s}_m \right) + \frac{C \widehat{\sigma}^2 D_m}{n} \left(5 + 2 \log \left(\frac{n}{D_m} \right) \right) \right\}$$

- BGH (Baraud, Giraud & Huet 2009): multiplicative penalty, $\sigma(\cdot) \equiv \sigma$

$$\operatorname{argmin}_{m \in \mathcal{M}_n} \left\{ P_n \gamma \left(\widehat{s}_m \right) \left[1 + \frac{\mathrm{pen}_{\mathrm{BGH}}(m)}{n - D_m} \right] \right\}$$

- ZS (Zhang & Siegmund, 2007): modified BIC, $\sigma(\cdot) \equiv \sigma$

Competitors

- [Emp, BM]: assume $\sigma(\cdot) \equiv \sigma$

$$\operatorname{argmin}_{m \in \mathcal{M}_n} \left\{ P_n \gamma \left(\widehat{s}_m \right) + \frac{C \widehat{\sigma}^2 D_m}{n} \left(5 + 2 \log \left(\frac{n}{D_m} \right) \right) \right\}$$

- BGH (Baraud, Giraud & Huet 2009): multiplicative penalty, $\sigma(\cdot) \equiv \sigma$

$$\operatorname{argmin}_{m \in \mathcal{M}_n} \left\{ P_n \gamma \left(\widehat{s}_m \right) \left[1 + \frac{\mathrm{pen}_{\mathrm{BGH}}(m)}{n - D_m} \right] \right\}$$

- ZS (Zhang & Siegmund, 2007): modified BIC, $\sigma(\cdot) \equiv \sigma$
- PML (Picard *et al.*, 2005): penalized maximum likelihood, looks for change-points of (η, σ), assuming a Gaussian model

$$\operatorname{argmin}_{m \in \mathcal{M}_n} \left\{ \sum_{\lambda \in m} n \widehat{p}_\lambda \log \left(\frac{1}{n \widehat{p}_\lambda} \sum_{t_i \in \lambda} \left(Y_i - \widehat{s}_m(t_i) \right)^2 \right) + \widehat{C}'' D_m \right\}$$

Simulations: comparison to the oracle (quadratic risk)

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

$N = 10\,000$ sample

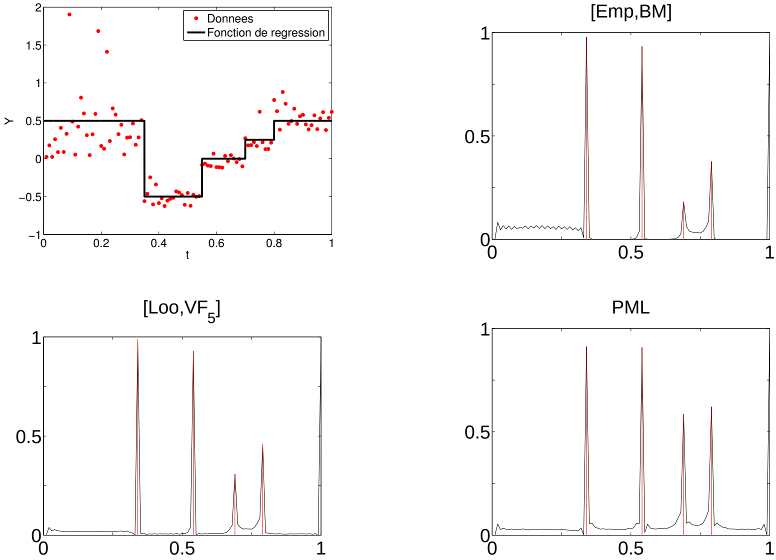
$\mathcal{L}(\varepsilon)$ / $\sigma(\cdot)$ / η	Gaussian homosc. s_2	Gaussian heterosc. s_2	Gaussian heterosc. s_3
[Loo, VF_5]	4.02 ± 0.02	4.95 ± 0.05	5.59 ± 0.02
[Emp, VF_5]	3.99 ± 0.02	5.62 ± 0.05	6.13 ± 0.02
[Emp, BM]	3.58 ± 0.02	9.25 ± 0.06	6.24 ± 0.02
BGH	3.52 ± 0.02	10.13 ± 0.07	6.31 ± 0.02
ZS	3.62 ± 0.02	6.50 ± 0.05	6.61 ± 0.02
PML	4.34 ± 0.02	2.73 ± 0.03	4.99 ± 0.03

Simulations: comparison to the oracle (quadratic risk)

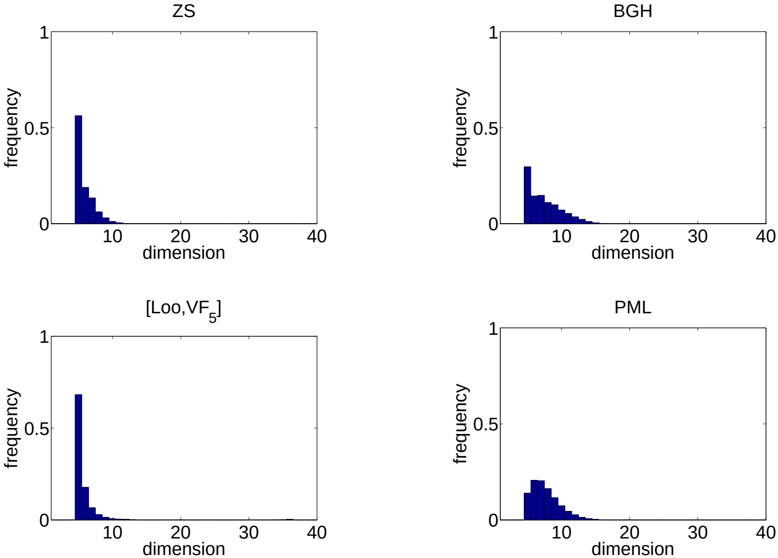
$$\frac{\mathbb{E}\left[\ell\left(s^{\star}, \widehat{s}_{\widehat{m}}\right)\right]}{\mathbb{E}\left[\inf_{m \in \mathcal{M}_n}\left\{\ell\left(s^{\star}, \widehat{s}_m\right)\right\}\right]} \qquad N = 10\,000 \text{ sample}$$

$\mathcal{L}(\varepsilon)$ $\sigma(\cdot)$ η	Gaussian homosc. s_2	Exponential heterosc. s_2	Exponential heterosc. s_3
$[\mathrm{Loo}, \mathrm{VF}_5]$	4.02 ± 0.02	4.47 ± 0.05	5.11 ± 0.03
$[\mathrm{Emp}, \mathrm{VF}_5]$	3.99 ± 0.02	5.98 ± 0.07	6.22 ± 0.04
$[\mathrm{Emp}, \mathrm{BM}]$	3.58 ± 0.02	10.81 ± 0.09	6.45 ± 0.04
BGH	3.52 ± 0.02	11.67 ± 0.09	6.42 ± 0.04
ZS	3.62 ± 0.02	9.34 ± 0.09	6.83 ± 0.04
PML	4.34 ± 0.02	5.04 ± 0.06	5.40 ± 0.03

Simulations: position of the change-points



Simulations: selected dimension ($D_0 = 5$)



Outline

Resampling heuristics (bootstrap, Efron 1979)

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

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

Resampling heuristics (bootstrap, Efron 1979)

$$
\begin{array}{lcl}
\text{Real world :} & P \xrightarrow{\text{sampling}} P_n \Longrightarrow \widehat{s}_m \\
\Big\downarrow & \\
\text{Bootstrap world :} & P_n
\end{array}
$$

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

Resampling heuristics (bootstrap, Efron 1979)

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

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

Resampling heuristics (bootstrap, Efron 1979)

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

$$\text{Bootstrap world :} \qquad P_n \xrightarrow{\text{sub-sampling}} P_n^W \Longrightarrow \widehat{s}_m^W$$

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

V-fold: $\quad P_n^W = \dfrac{1}{n - \mathrm{Card}(B_J)} \displaystyle\sum_{i \notin B_J} \delta_{(X_i, Y_i)} \quad$ with $\quad J \sim \mathcal{U}(1, \ldots, V)$

V-fold penalties (A. 2008)

- Ideal penalty:

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

- V-fold penalty (A., 2008):

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

$$\widehat{s}_m^{(-B_j)} = \widehat{s}_m \left(D_n^{(-B_j)} \right)$$

- Selected model:

$$\widehat{m} \in \mathrm{argmin}_{m \in \mathcal{M}_n} \{ P_n \gamma(\widehat{s}_m) + \mathrm{pen}(m) \}$$

Computing expectations

Assumptions:

$$\left.\begin{array}{r}\mathcal{B} = (B_j)_{1 \leq j \leq V} \text{ partition of } \{1, \ldots, n\} \\ \text{and} \quad \forall j \in \{1, \ldots, V\} \ , \qquad \operatorname{Card}(B_j) = \dfrac{n}{V}\end{array}\right\} \qquad \textbf{(RegPart)}$$

$$\forall 1 \leq N \leq n \ , \qquad \mathbb{E}\left[\operatorname{pen}_{\mathrm{id}}(m; D_N)\right] = \frac{\gamma_m}{N} \qquad \textbf{(Epenid)}$$

Computing expectations

Assumptions:

$$\left.\begin{array}{r} \mathcal{B} = (B_j)_{1 \le j \le V} \text{ partition of } \{1, \ldots, n\} \\ \text{and} \quad \forall j \in \{1, \ldots, V\} \;, \qquad \operatorname{Card}(B_j) = \dfrac{n}{V} \end{array}\right\} \qquad \textbf{(RegPart)}$$

$$\forall 1 \le N \le n \;, \qquad \mathbb{E}\left[\mathrm{pen}_{\mathrm{id}}(m; D_N)\right] = \frac{\gamma_m}{N} \qquad \textbf{(Epenid)}$$

Proposition (A. 2011)

$$\mathbb{E}\left[\mathrm{pen}_{\mathrm{VF}}(m; D_n; C; \mathcal{B})\right] = \frac{C}{V-1}\mathbb{E}\left[\mathrm{pen}_{\mathrm{id}}(m; D_n)\right]$$

Concentration: additional assumptions

For all $N \in \{1, \ldots, n\}$,

$$\mathbb{P}\left(\left| p_1(m; D_N) - \mathbb{E}\left[p_1(m; D_N) \right] \right| \leq w_N \mathbb{E}\left[p_1(m; D_N) \right] \right) \geq 1 - q_N \quad (\mathbf{C} p_1)$$

$$\mathbb{P}\left(\left| p_2(m; D_N) - \mathbb{E}\left[p_2(m; D_N) \right] \right| \leq w_N \mathbb{E}\left[p_2(m; D_N) \right] \right) \geq 1 - q_N \quad (\mathbf{C} p_2)$$

$$\left. \begin{array}{l} \exists S_m \subset \mathbb{S} \text{ s.t. } s_m^\star \in S_m \;,\; \widehat{s}_m(D_N) \in S_m \quad \text{a.s.} \\ \text{and} \quad \forall t \in S_m \;,\; \forall x \geq 0 \;, \\ \mathbb{P}\left(\left| \delta(t; D_N) - \delta(s^\star; D_N) \right| \leq \inf_{\eta \in (0,1]} \left\{ \eta \ell\left(s^\star, t \right) + \dfrac{K_\delta x}{\eta N} \right\} \right) \\ \quad \geq 1 - 2e^{-x} \end{array} \right\} \quad (\mathbf{C}\delta)$$

$$p_1(m; D_N) = P\gamma\left(\widehat{s}_m(D_N) \right) - P\gamma\left(s_m^\star \right)$$

$$p_2(m; D_N) = P_N\gamma\left(s_m^\star \right) - P_N\gamma\left(\widehat{s}_m(D_N) \right)$$

$$\delta(t; D_N) = (P_N - P)\gamma\left(t \right)$$

Concentration: result

Proposition (A. 2011)

Assume: $V \geq 2$, (**RegPart**), (**Epenid**), ($\mathbf{C}p_1$), ($\mathbf{C}p_2$) *and* ($\mathbf{C}\delta$) *with* $\gamma_m \geq 0$, $K_\delta > 0$ *and* $(w_k), (q_k)$ *non-increasing non-negative.* *Then,* $\forall C > 0, x \geq 0$, *with probability* $1 - 2V \left(q_{\frac{n(V-1)}{V}} + 2e^{-x} \right)$, $\forall \eta \in (0, 1]$,

$$\begin{aligned} &\left| \mathrm{pen}_{\mathrm{VF}}(m; D_n; C; \mathcal{B}) - \mathbb{E}\left[\mathrm{pen}_{\mathrm{VF}}(m; D_n; C; \mathcal{B})\right] - \mathcal{Z} \right| \\ &\qquad \leq \frac{4C}{V} \left(\eta + 2 w_{\frac{n(V-1)}{V}} \right) \mathbb{E}\left[\mathrm{pen}_{\mathrm{id}}(m; D_n)\right] \\ &\qquad\quad + \frac{C}{V} \left(2 \eta \ell(s^\star, s^\star_m) + \frac{4 K_\delta x V}{\eta n} \right) \end{aligned}$$

where $\mathcal{Z} = \mathcal{Z}(D_n; C; \mathcal{B}) = \dfrac{C}{V} \displaystyle\sum_{j=1}^{V} \left(\delta\left(s^\star; D_n^{(B_j)} \right) - \delta\left(s^\star; D_n^{(-B_j)} \right) \right)$

Oracle inequality for V-fold penalization

Theorem (A. 2008–2011)

Assume also that $w_k \to 0$, $C = V - 1$ *and* $\exists (\kappa_k)_{k \geq 1}$ *non-increasing,*

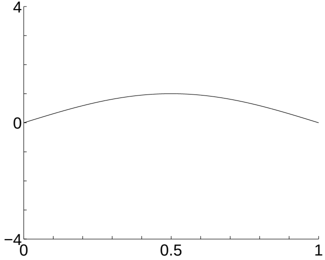
$$\forall N \geq 1 \ , \qquad 0 \leq \mathbb{E}\left[\mathrm{pen}_{\mathrm{id}}(m; D_N) \right] \leq \kappa_N \mathbb{E}\left[\ell\left(s^{\star}, \widehat{s}_m\left(D_N \right) \right) \right]$$

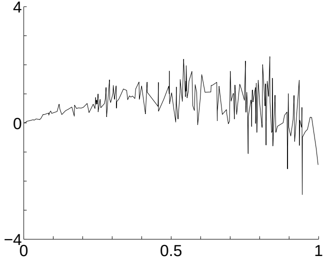
Then, with probability at least $1 - L_1 V \operatorname{Card}(\mathcal{M}_n)(q_{\frac{n(V-1)}{V}} + e^{-x})$, *for every* $\eta_k \to 0$,

$$\ell\left(s^{\star}, \widehat{s}_{\widehat{m}_{\mathrm{pen}_{\mathrm{VF}}}(D_n)} \right) \leq \left[1 + L_2 \left(\eta_n + \frac{1}{n} + w_{\frac{n(V-1)}{V}} \right) \right] \times \inf_{m \in \mathcal{M}_n} \left\{ \ell\left(s^{\star}, \widehat{s}_m(D_n) \right) \right\} + \frac{L_3 K_\delta x V}{\eta_n n}$$

Example: regressograms under reasonable assumptions on $(\|Y\|_\infty \leq A, \ \sigma(\cdot) \geq \sigma_{\min} > 0, \ \ldots)$

Simulations: sin, $n = 200$, $\sigma(x) = x$, $\mathcal{M}_n = \mathcal{M}_n^{(\mathrm{reg},1/2)}$





Mallows	3.69 ± 0.07
2-fold	2.54 ± 0.05
5-fold	2.58 ± 0.06
10-fold	2.60 ± 0.06
20-fold	2.58 ± 0.06
leave-one-out	2.59 ± 0.06
pen 2-f	3.06 ± 0.07
pen 5-f	2.75 ± 0.06
pen 10-f	2.65 ± 0.06
pen Loo	2.59 ± 0.06
Mallows $\times 1.25$	3.17 ± 0.07
pen 2-f $\times 1.25$	2.75 ± 0.06
pen 5-f $\times 1.25$	2.38 ± 0.06
pen 10-f $\times 1.25$	2.28 ± 0.05
pen Loo $\times 1.25$	2.21 ± 0.05

Choice of V: density estimation (A. & Lerasle, 2011)

- Least-squares density estimation: assuming (**RegPart**),

$$\operatorname{var}\left((\mathrm{pen}_{\mathrm{VF}}(m) - \mathrm{pen}_{\mathrm{id}}(m)) - (\mathrm{pen}_{\mathrm{VF}}(m') - \mathrm{pen}_{\mathrm{id}}(m')) \right)$$
$$= \frac{8}{n^2} \left[1 + \frac{1}{V-1} \right] F\left(m, m' \right) + \frac{4}{n} \operatorname{var}_P \left(s_m^\star - s_{m'}^\star \right)$$

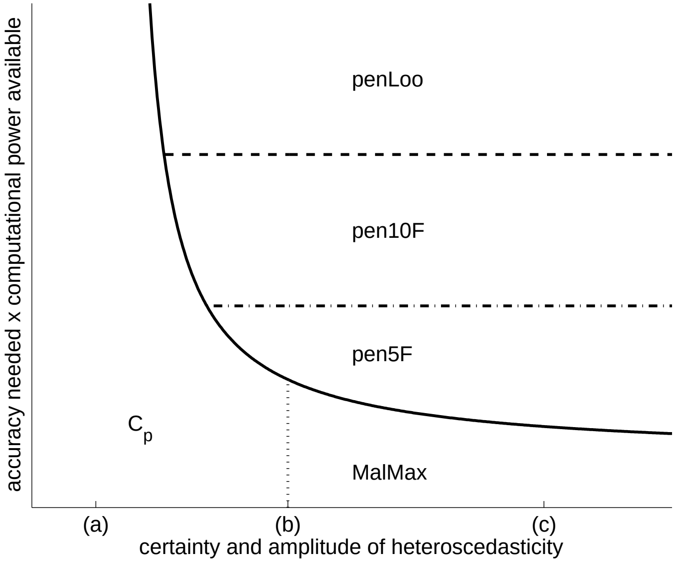
with $F\left(m, m' \right) > 0$.

- For regular histograms,

$$F\left(m, m' \right) \leq \left(D_m + D_{m'} \right) \left\| s^\star \right\|^2 + 2 \left\| s^\star \right\|^4$$

Outline

Choice of an estimator selection procedure



Conclusion

- guarantees for practical procedures:
  - "elbow" heuristics on the L-curve, slope heuristics
  - resampling(-based penalties)
  - cross-validation

Conclusion

- guarantees for practical procedures:
  - "elbow" heuristics on the L-curve, slope heuristics
  - resampling(-based penalties)
  - cross-validation
- use theory for designing new procedures:
  - minimal penalties for linear estimators
  - V-fold penalties for correcting the bias of VFCV

Conclusion

- guarantees for practical procedures:
  - “elbow” heuristics on the L-curve, slope heuristics
  - resampling(-based penalties)
  - cross-validation
- use theory for designing new procedures:
  - minimal penalties for linear estimators
  - V-fold penalties for correcting the bias of VFCV
- theory precise enough for explaining differences observed experimentally:
  - compare resampling weights
  - influence of V on V-fold methods

Conclusion

- guarantees for practical procedures:
  - "elbow" heuristics on the L-curve, slope heuristics
  - resampling(-based penalties)
  - cross-validation
- use theory for designing new procedures:
  - minimal penalties for linear estimators
  - V-fold penalties for correcting the bias of VFCV
- theory precise enough for explaining differences observed experimentally:
  - compare resampling weights
  - influence of V on V-fold methods
- "non-asymptotic" results

Open problems

- guarantees for practical procedures:
  - cross-validation and resampling penalties outside "toy frameworks" (regressograms, least-squares density estimation)?
  - minimal penalties without the least-squares contrast (SVM, Lasso, and so on)?

Open problems

- guarantees for practical procedures:
  - cross-validation and resampling penalties outside "toy frameworks" (regressograms, least-squares density estimation)?
  - minimal penalties without the least-squares contrast (SVM, Lasso, and so on)?
- theory precise enough for explaining differences observed experimentally:
  - choice of a resampling scheme / a cross-validation method?
  - explain the (non-systematic) variability of leave-one-out?

Open problems

- guarantees for practical procedures:
  - cross-validation and resampling penalties outside "toy frameworks" (regressograms, least-squares density estimation)?
  - minimal penalties without the least-squares contrast (SVM, Lasso, and so on)?
- theory precise enough for explaining differences observed experimentally:
  - choice of a resampling scheme / a cross-validation method?
  - explain the (non-systematic) variability of leave-one-out?
- "non-asymptotic" results:
  - overpenalization phenomenon?