Optimal model selection with V-fold cross-validation: how should V be chosen?

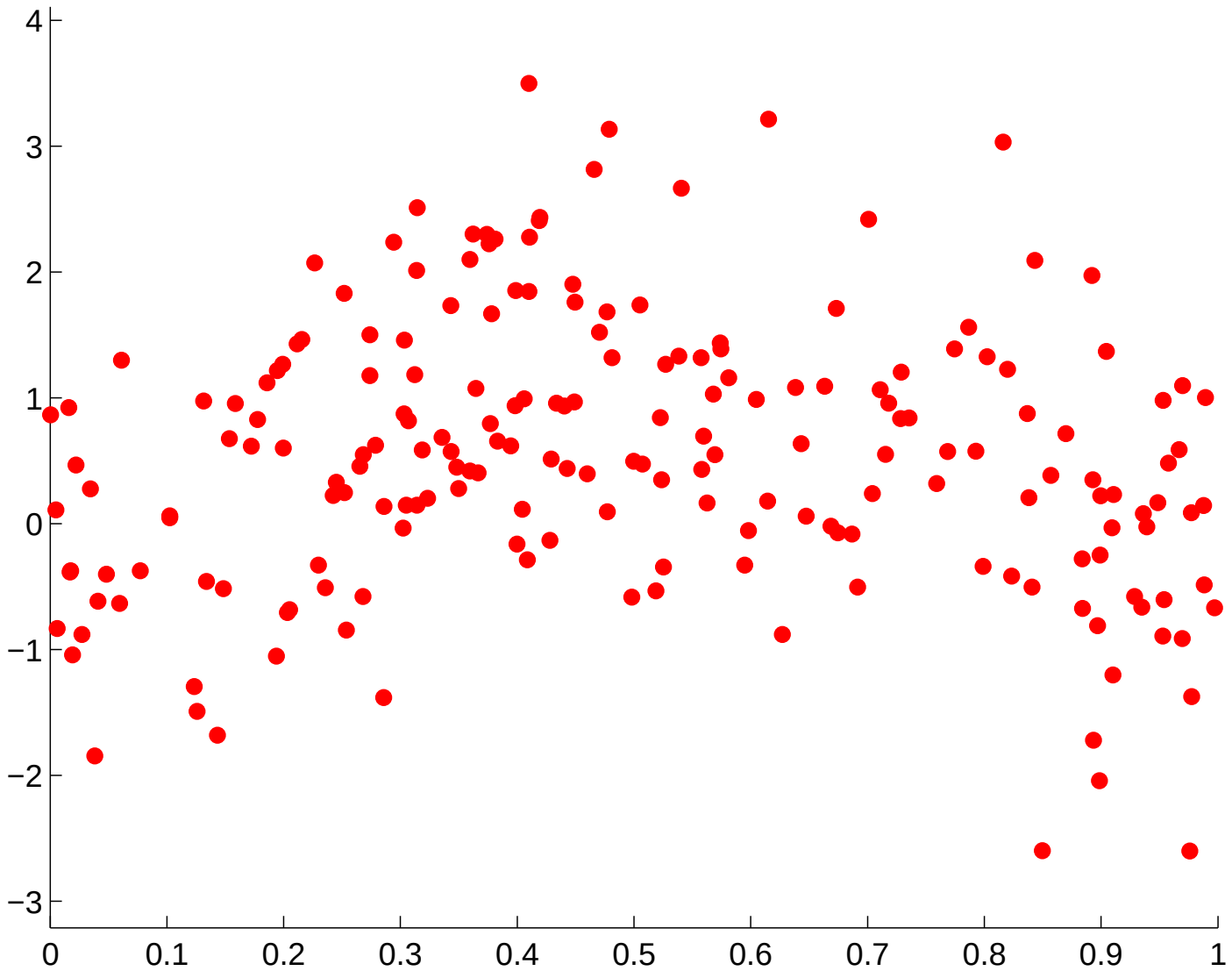
Sylvain Arlot (joint work with M. Lerasle)

[1]CNRS

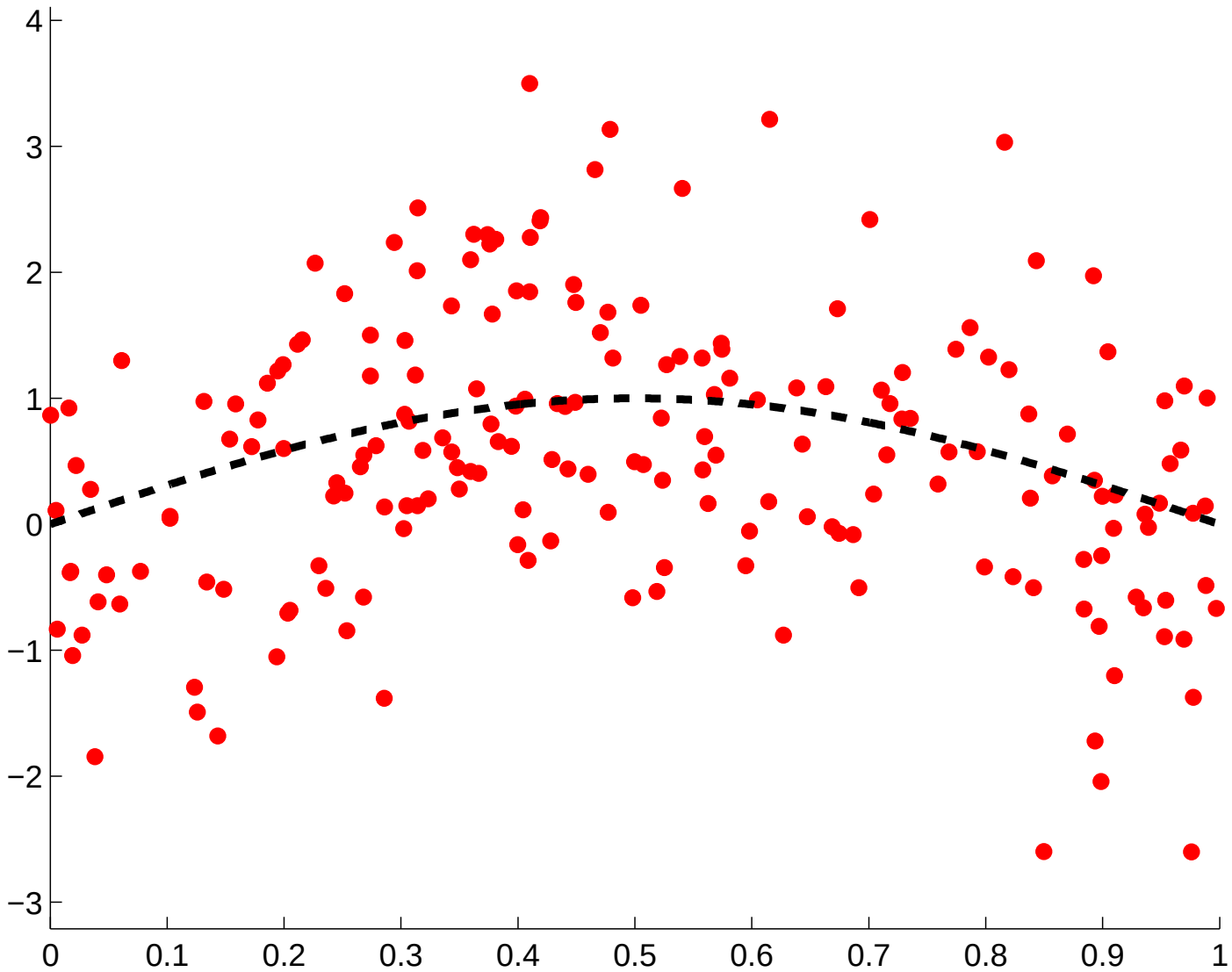
[2]École Normale Supérieure (Paris), DI/ENS, Équipe SIERRA

IHES, March 20th, 2013

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



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



Learning problem (1): prediction/regression

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

Learning problem (1): prediction/regression

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

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

$\Rightarrow$ excess risk

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

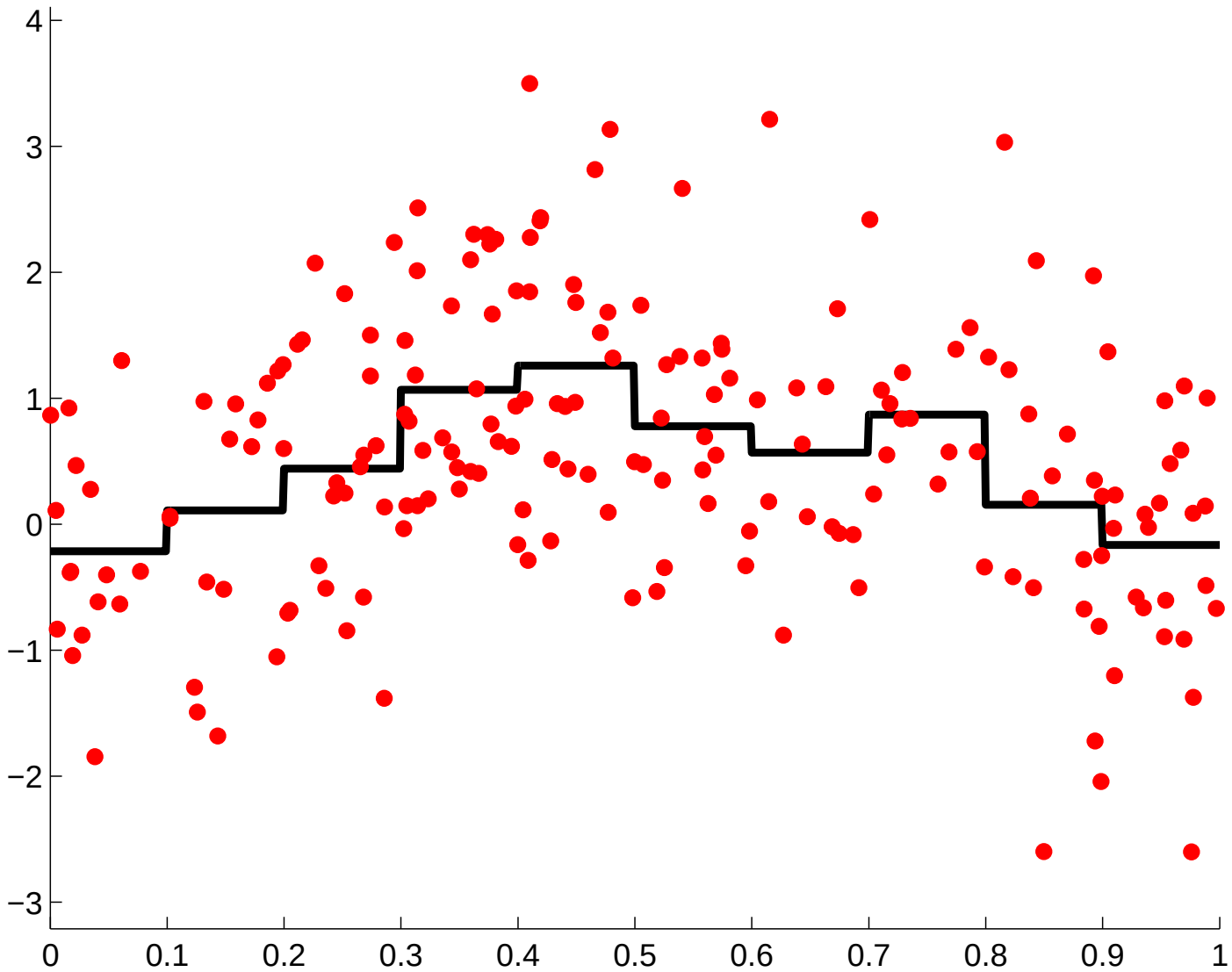
Learning problem (2): density estimation

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

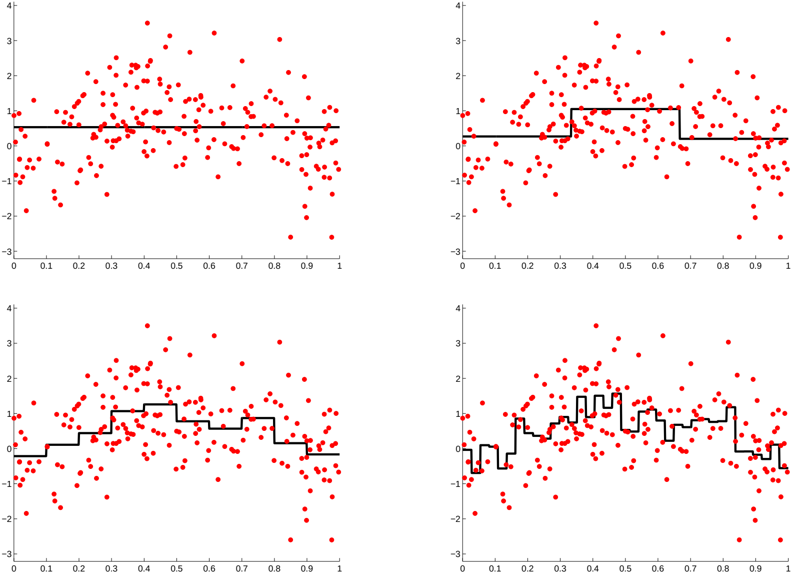
the true density $s^\star \in \operatorname{argmin}_{t \in S} P\gamma(t)$ and the excess risk is

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

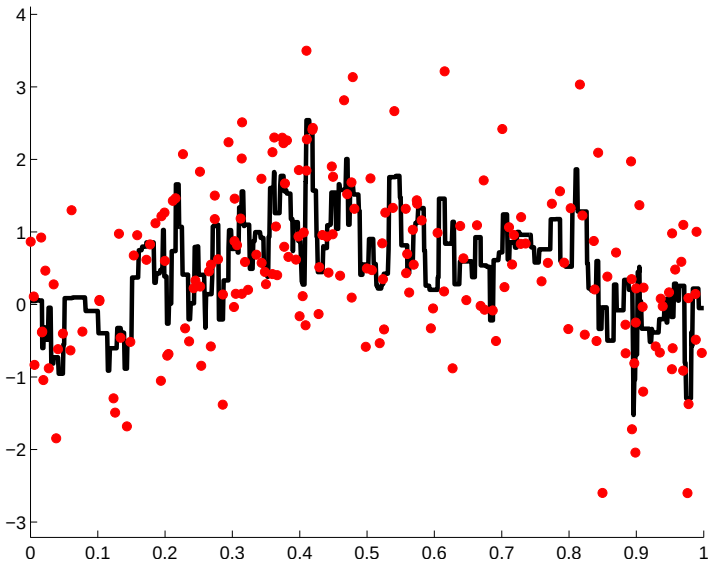
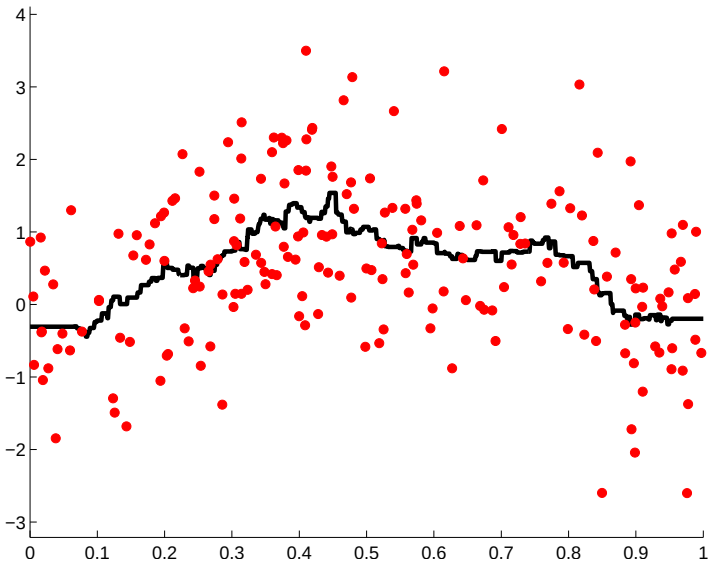
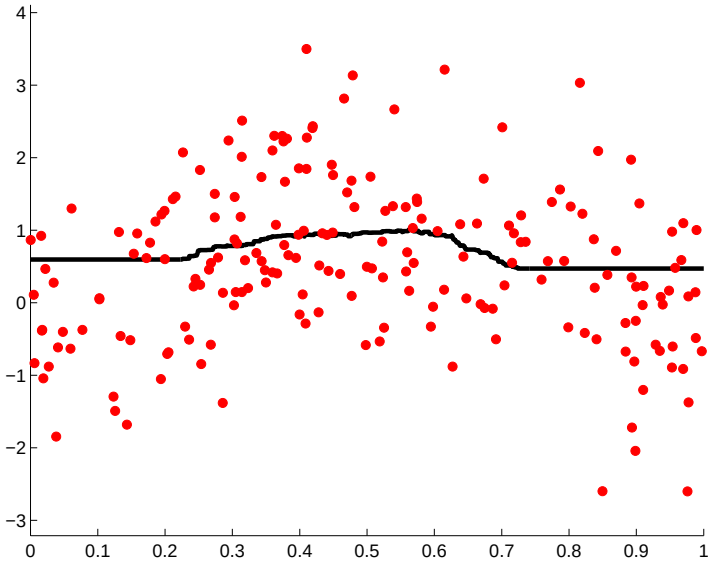
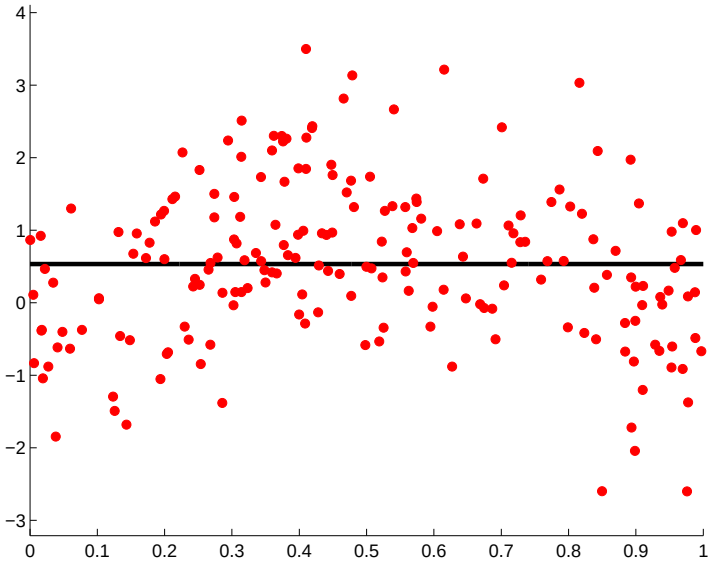
Estimators: example: regressogram



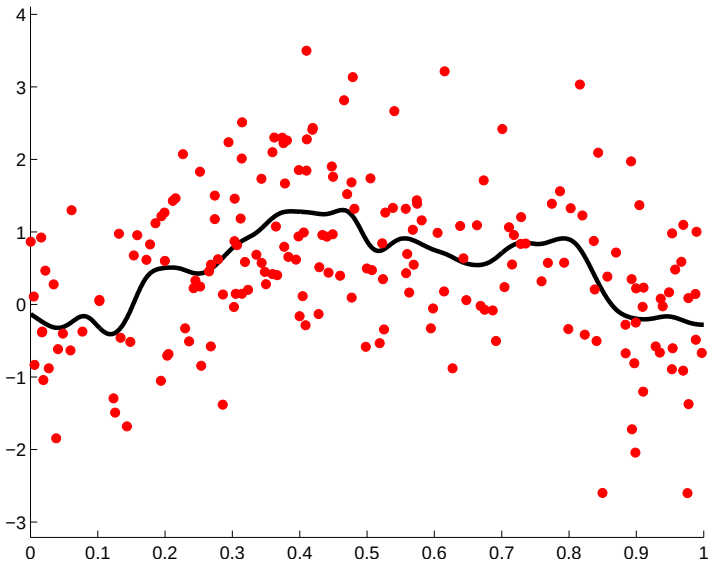
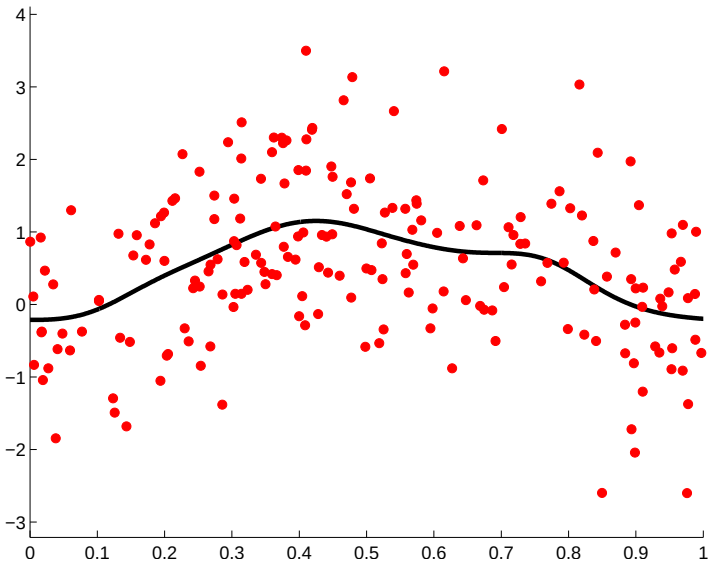
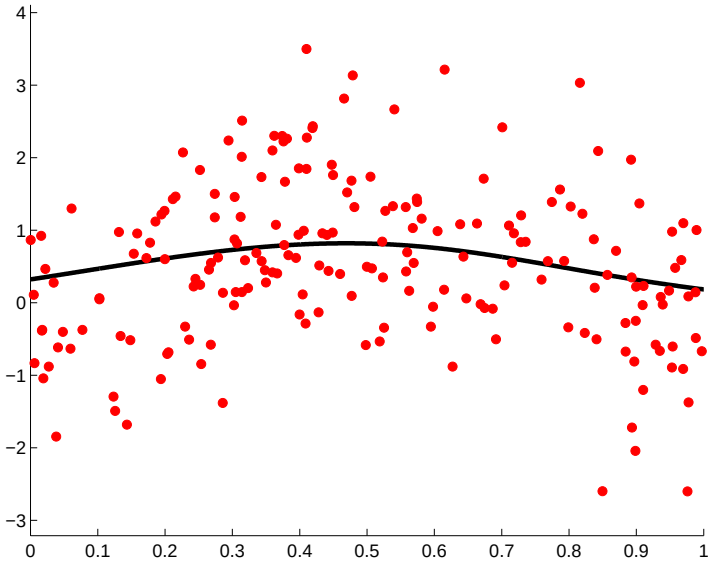
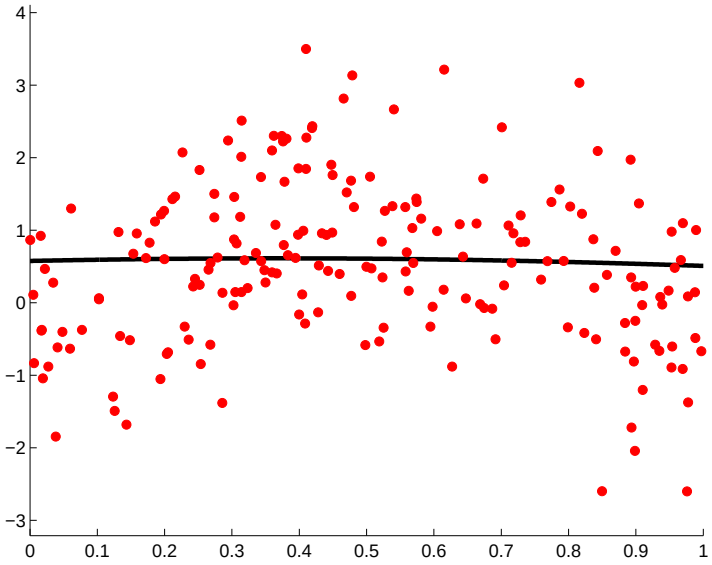
Estimator selection: regular regressograms



Estimator selection: k nearest neighbours

Estimator selection: Nadaraya-Watson

Estimator selection

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

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

Example of model: histograms ($\Rightarrow$ regressograms in regression)

Estimator selection

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

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

  Example of model: histograms ($\Rightarrow$ regressograms in regression)

- Estimator collection $(\widehat{s}_m)_{m \in \mathcal{M}} \Rightarrow$ choose $\widehat{m} = \widehat{m}(D_n)$?
  e.g., family of models $(S_m)_{m \in \mathcal{M}} \Rightarrow$ family of least-squares estimators

Estimator selection

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

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

  Example of model: histograms ($\Rightarrow$ regressograms in regression)

- Estimator collection $(\widehat{s}_m)_{m \in \mathcal{M}} \Rightarrow$ choose $\widehat{m} = \widehat{m}(D_n)$?
  e.g., family of models $(S_m)_{m \in \mathcal{M}} \Rightarrow$ family of least-squares estimators

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

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

Bias-variance trade-off

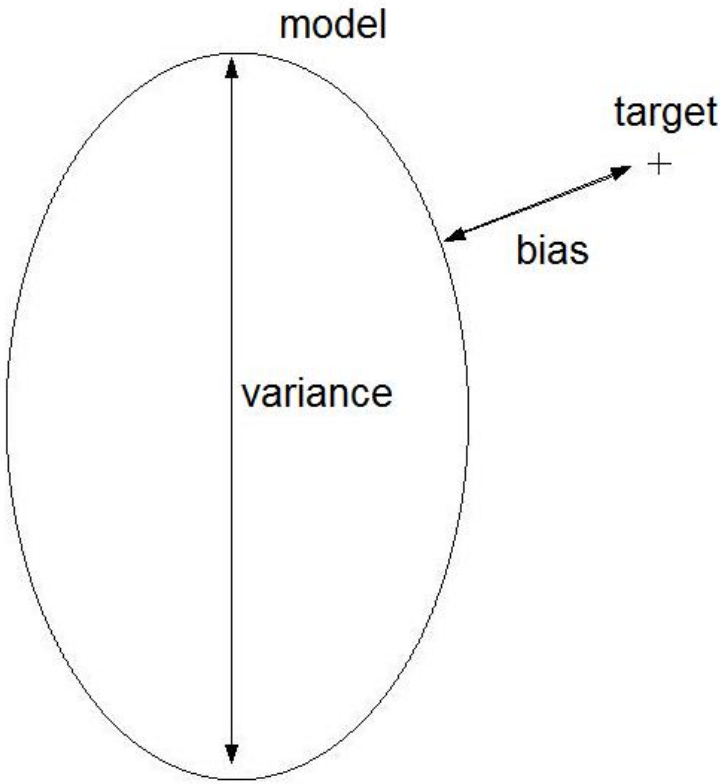
$$\mathbb{E}\left[\ell\left(s^\star, \widehat{s}_m\right)\right] = \text{Bias} + \text{Variance}$$

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

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



Bias-variance trade-off

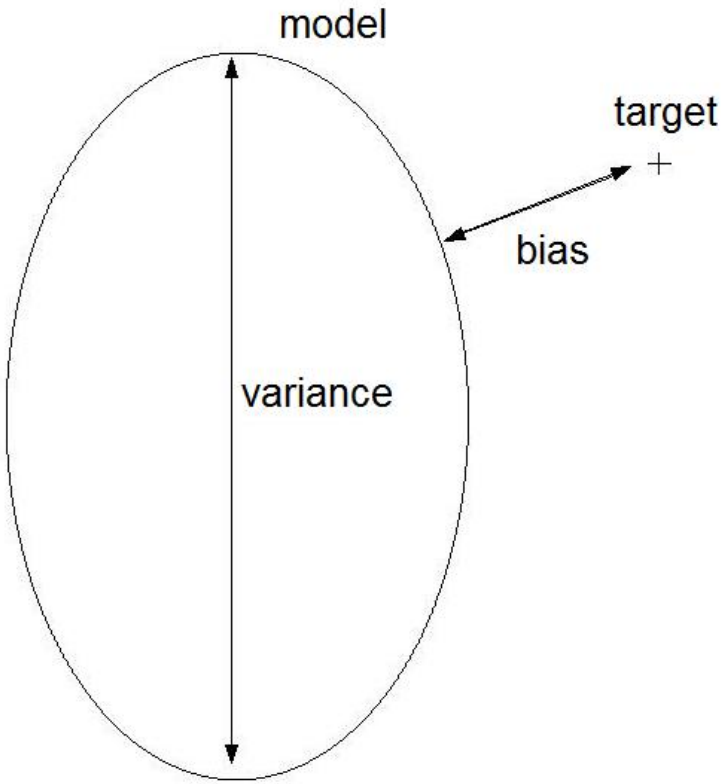
$$\mathbb{E}\left[\ell\left(s^\star, \widehat{s}_m\right)\right] = \text{Bias} + \text{Variance}$$

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

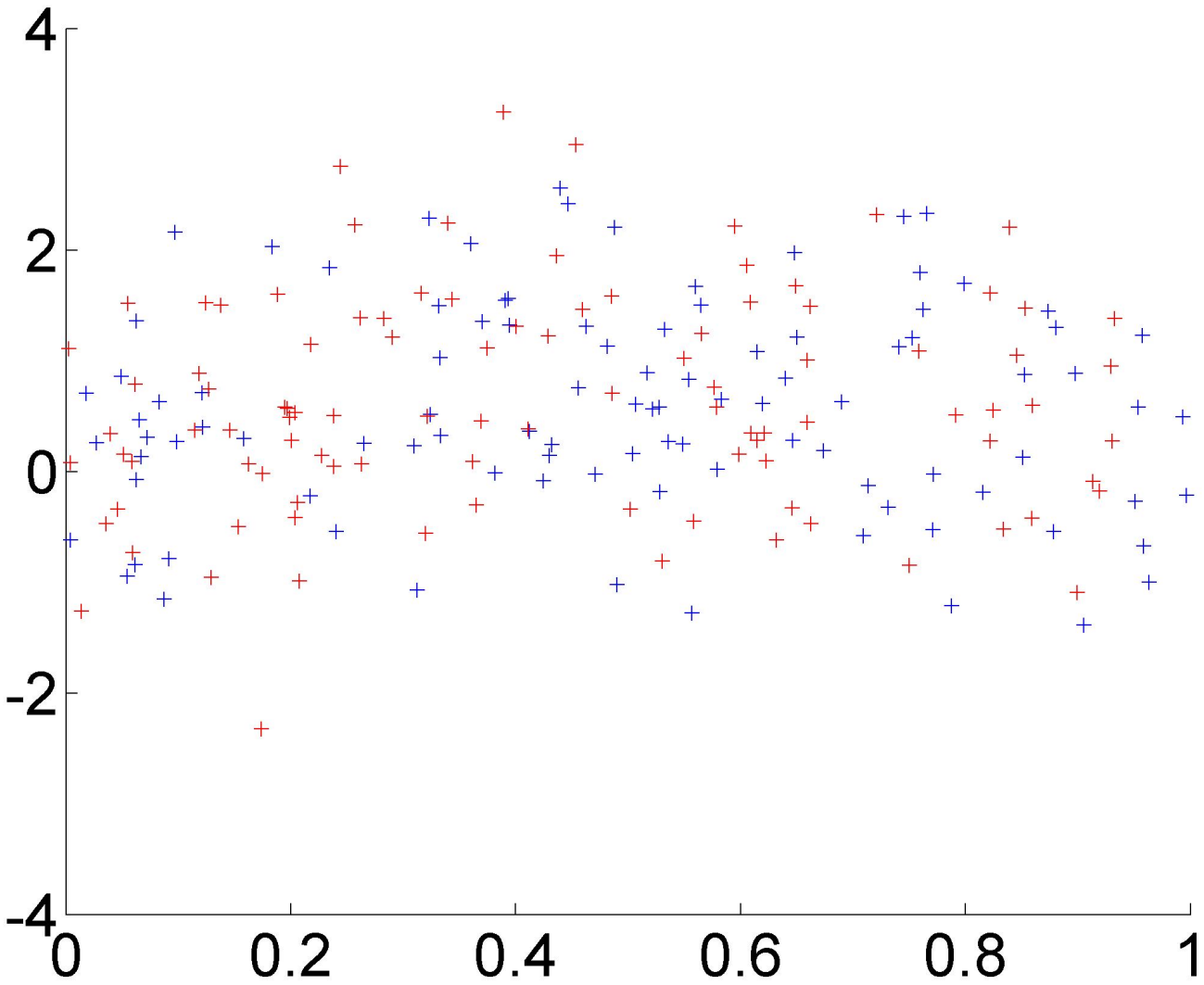
OLS in regression: $\dfrac{\sigma^2 \dim(S_m)}{n}$



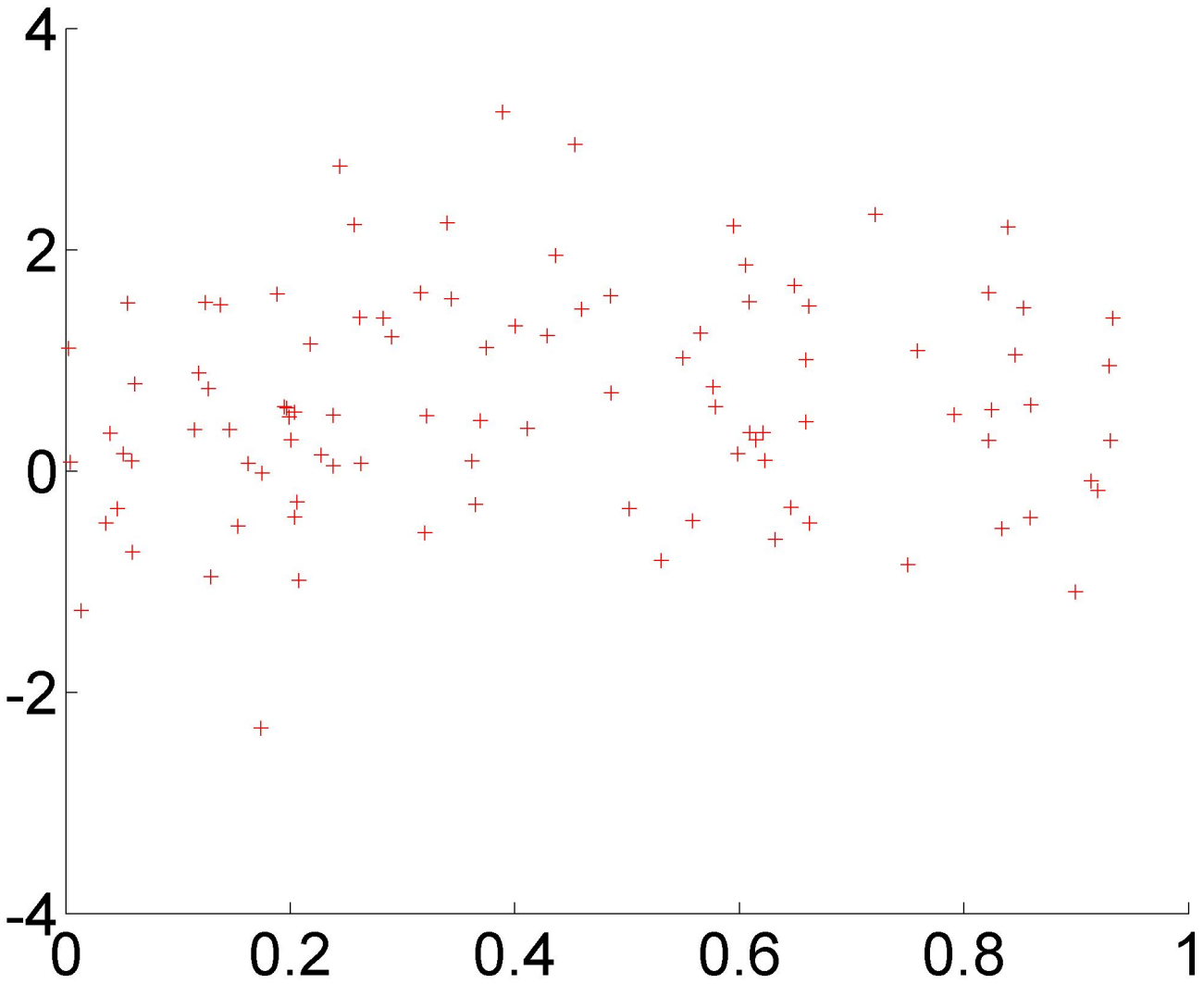
Bias-variance trade-off

$\Leftrightarrow$ avoid overfitting and underfitting

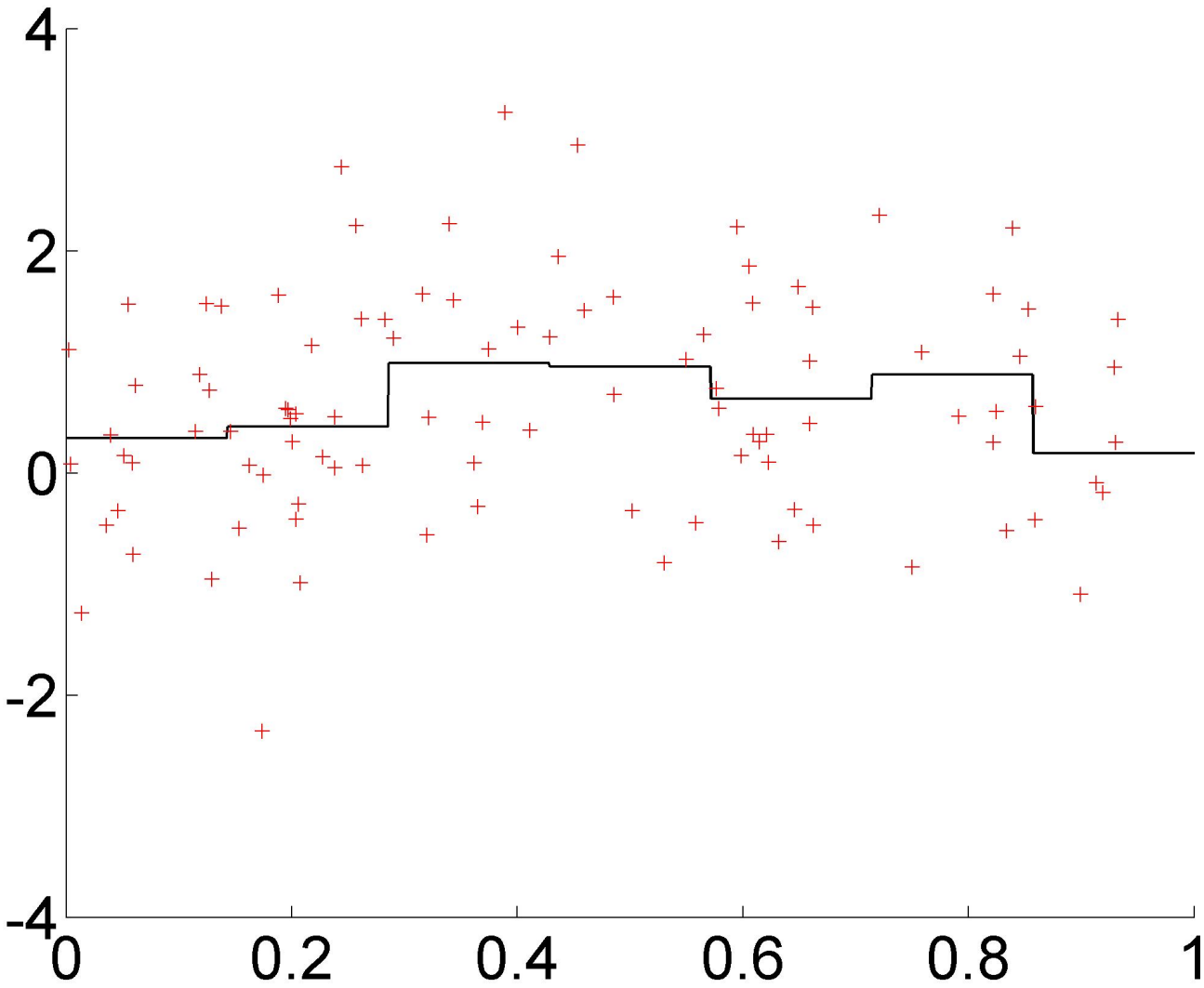
Validation principle



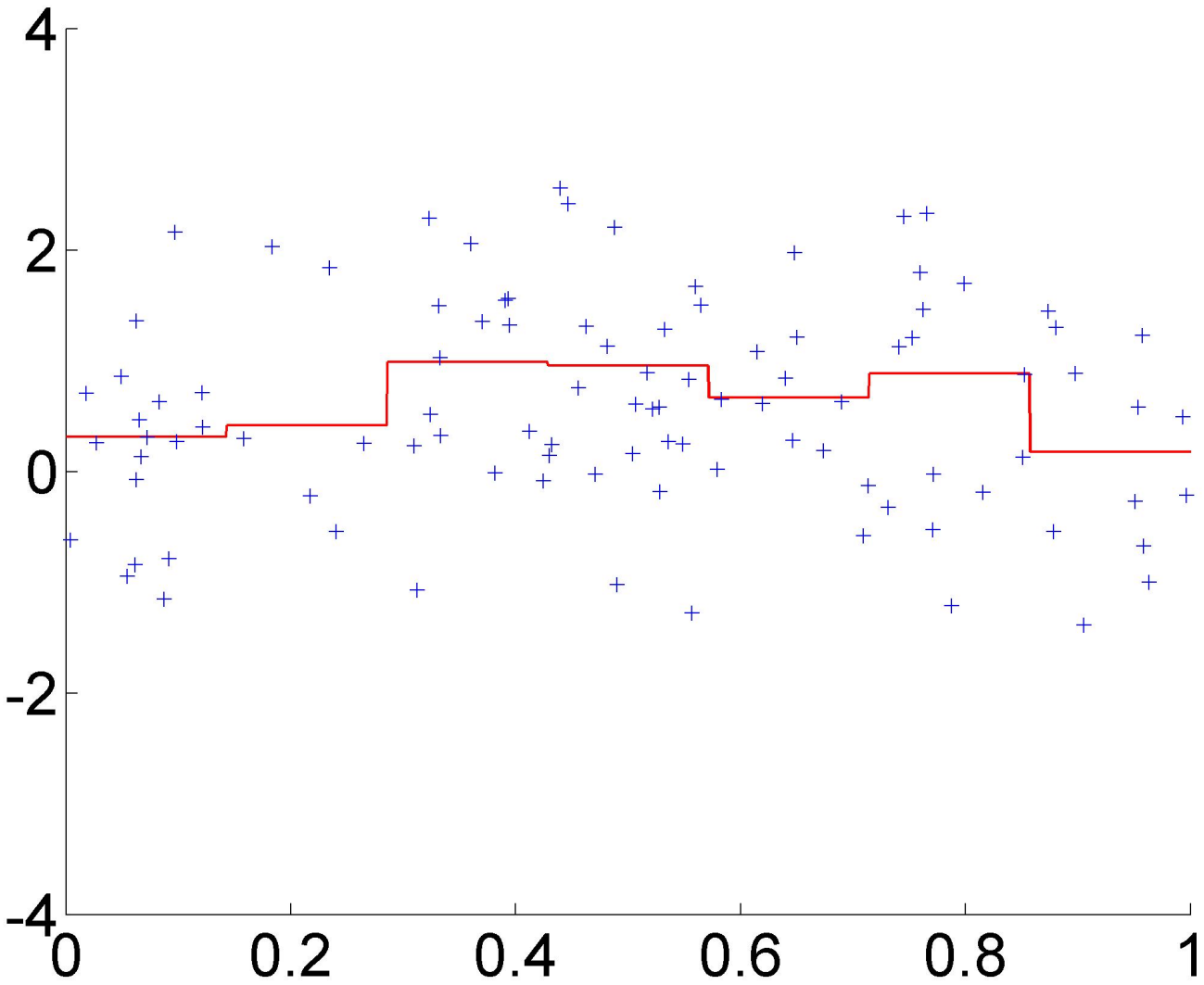
Validation principle: learning sample



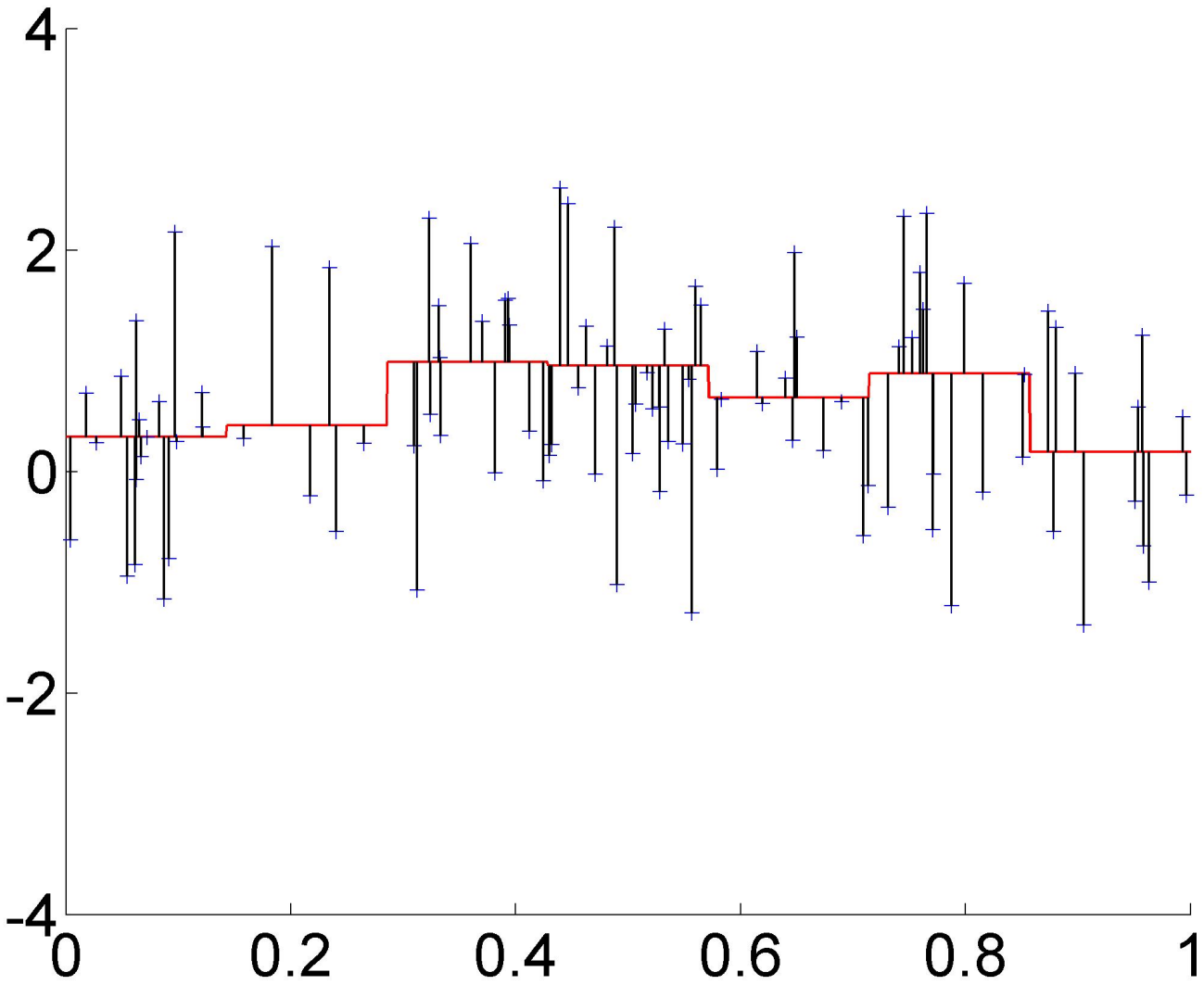
Validation principle: learning sample



Validation principle: validation sample



Validation principle: validation sample



Cross-validation

$$\underbrace{\xi_1, \ldots, \xi_q}_{\text{Learning}}, \underbrace{\xi_{q+1}, \ldots, \xi_n}_{\text{Validation}}$$

$$\widehat{s}_m^{(\ell)} \in \arg\min_{t \in S_m} \left\{ \sum_{i=1}^{q} \gamma(t, \xi_i) \right\}$$

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

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

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

Bias of cross-validation

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

Bias of cross-validation

- In this talk, we always assume $\mathrm{Card}(B_j) = n/V$ for all j.
- Ideal criterion: $P\gamma(\widehat{s}_m)$
- For regressograms and for least-squares density estimation:

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

$$\Rightarrow \quad \mathbb{E}\left[\widehat{\mathcal{R}}^{\mathrm{vf}}\left(\widehat{s}_m; D_n; \mathcal{B}\right)\right] = \mathbb{E}\left[P_n^{(j)}\gamma\left(\widehat{s}_m^{(-j)}\right)\right] = \mathbb{E}\left[P\gamma\left(\widehat{s}_m^{(-j)}\right)\right]$$

$$\approx \alpha(m) + \frac{V}{V-1}\frac{\beta(m)}{n}$$

Bias of cross-validation

- In this talk, we always assume $\mathrm{Card}(B_j) = n/V$ for all j.
- Ideal criterion: $P\gamma(\widehat{s}_m)$
- For regressograms and for least-squares density estimation:

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

$$\Rightarrow \quad \mathbb{E}\left[\widehat{\mathcal{R}}^{\mathrm{vf}}\left(\widehat{s}_m; D_n; \mathcal{B}\right)\right] = \mathbb{E}\left[P_n^{(j)}\gamma\left(\widehat{s}_m^{(-j)}\right)\right] = \mathbb{E}\left[P\gamma\left(\widehat{s}_m^{(-j)}\right)\right]$$

$$\approx \alpha(m) + \frac{V}{V-1}\frac{\beta(m)}{n}$$

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

Bias and model selection

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

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

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

Bias and model selection

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

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

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

- Key quantities:

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

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

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

Suboptimality of V-fold cross-validation

- $Y = X + \sigma\varepsilon$ with ε bounded and $\sigma > 0$
- $\mathcal{M}$: family of regular histograms on $\mathcal{X} = [0,1]$
- $\widehat{m}$ selected by V-fold cross-validation with V fixed as n grows

Theorem (A. 2008, arXiv:0802.0566)

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

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

with $\kappa(V) > 0$.

Bias-corrected VFCV / V-fold penalization

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

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

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

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

Bias-corrected VFCV / V-fold penalization

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

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

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

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

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

Oracle inequality with V-fold penalization (1): regression

- Least-squares regression, regressogram estimators, heteroscedastic noise
- $\mathrm{Card}(\mathcal{M}) \leq n^{\Diamond}$
- technical assumptions

Theorem (A. 2008, arXiv:0802.0566)

Some $\delta > 0$ exists such that, with probability at least $1 - \Diamond n^{-2}$, for any $\widehat{m}$ selected by V-fold penalization,

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

Oracle inequality with V-fold penalization (2): density

- Least-squares density estimation, histograms on $\mathbb{R}$
- technical assumption: $\forall m \in \mathcal{M}, \forall \lambda \in m, \mathrm{Leb}(\lambda) \geq 1/n$.

Theorem (A. & Lerasle 2012, arXiv:1210.5830)

With probability at least $1 - 10e^{-x}$, *for all* $\varepsilon > 0$, *if* $\delta_- + \varepsilon < 1$,

$$\ell\left(s^\star, \widehat{s}_{\widehat{m}}\right) \leq \frac{1 + \delta_+ + \varepsilon}{1 - \delta_- - \varepsilon} \inf_{m \in \mathcal{M}} \left\{ \ell\left(s^\star, \widehat{s}_m\right) \right\} + R_{n,x}$$

where $\delta := 2\left(\frac{C}{V-1} - 1\right)$ *and* $R_{n,x} := \frac{\kappa(V, s^\star, \delta, \varepsilon)}{\varepsilon^3 n} \left(x + \ln\left(\mathrm{Card}(\mathcal{M})\right)\right)^2$.

Similar result valid for more general models (with other technical assumptions).
Related results: van der Laan, Dudoit & Keles (2004); Celisse (2012) for the Lpo case.

Variance of the (corrected)-VFCV criterion

- Least-squares density estimation
- Histogram model with constant bin size d_m^{-1} (to simplify)

Theorem (A. & Lerasle 2012, arXiv:1210.5830)

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

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

$$\text{where} \quad A(m) \approx d_m$$

Variance and model selection

- $\widehat{m} \in \operatorname{argmin}_{m \in \mathcal{M}} \{ \operatorname{crit}(m; D_n) \}$ invariant by any random translation of crit $\Rightarrow$ $\operatorname{var}(\operatorname{crit}(m; D_n))$ meaningless.

Variance and model selection

- $\widehat{m} \in \operatorname{argmin}_{m \in \mathcal{M}} \{ \operatorname{crit}(m; D_n) \}$ invariant by any random translation of crit $\Rightarrow$ $\operatorname{var}(\operatorname{crit}(m; D_n))$ meaningless.

- $m = \widehat{m} \Leftrightarrow \forall m' \in \mathcal{M}$, $\operatorname{crit}(m) - \operatorname{crit}(m') \leq 0$ which is likely to happen only if

$$\forall m' \in \mathcal{M}, \quad \mathbb{E}\left[\operatorname{crit}(m) - \operatorname{crit}(m')\right] - t\sqrt{\operatorname{var}\left(\operatorname{crit}(m) - \operatorname{crit}(m')\right)} \leq 0$$

for some "small" $t > 0$.

Variance and model selection

- $\widehat{m} \in \operatorname{argmin}_{m \in \mathcal{M}} \{ \operatorname{crit}(m; D_n) \}$ invariant by any random translation of crit $\Rightarrow$ $\operatorname{var}(\operatorname{crit}(m; D_n))$ meaningless.

- $m = \widehat{m} \Leftrightarrow \forall m' \in \mathcal{M}$, $\operatorname{crit}(m) - \operatorname{crit}(m') \leq 0$ which is likely to happen only if

$$\forall m' \in \mathcal{M}, \quad \mathbb{E}\left[\operatorname{crit}(m) - \operatorname{crit}(m')\right] - t\sqrt{\operatorname{var}\left(\operatorname{crit}(m) - \operatorname{crit}(m')\right)} \leq 0$$

  for some "small" $t > 0$.

- Constant bias among $(\widehat{\mathcal{R}}^{\mathrm{vf,corr}})_{2 \leq V \leq n}$ $\Rightarrow$ compare

$$\operatorname{var}\left(\widehat{\mathcal{R}}^{\mathrm{vf,corr}}\left(\widehat{s}_m; D_n; \mathcal{B}\right) - \widehat{\mathcal{R}}^{\mathrm{vf,corr}}\left(\widehat{s}_{m'}; D_n; \mathcal{B}\right) \right)$$

Variance and model selection

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

Theorem (A. & Lerasle 2012, arXiv:1210.5830)

If $S_m \subset S_{m'}$ are two histogram models with constant bin sizes $d_m, d_{m'}$, then,

$$\mathrm{var}\left(\Delta(m, m', V)\right) = \left(4 + \frac{2}{n} + \frac{1}{n^2}\right) \frac{\mathrm{var}_P\left(s^\star_m - s^\star_{m'}\right)}{n} + 2\left(1 + \frac{4}{V-1} - \frac{1}{n}\right) \frac{B(m, m')}{n^2}$$

where $B(m, m') \propto \left\| s^\star_m - s^\star_{m'} \right\| d_m$.

The two terms are of the same order if $\left\| s^\star_m - s^\star_{m'} \right\| \approx d_m / n$.
Similar formulas for general models $S_m, S_{m'}$.

Intuition: how to choose V?

- Bias: decreases with V / can be removed (V-fold penalization)
- Variance: decreases with V / close to its minimum with $V = 5$ or 10

$\Rightarrow$ best performance for the largest V (not necessarily valid in general in learning)

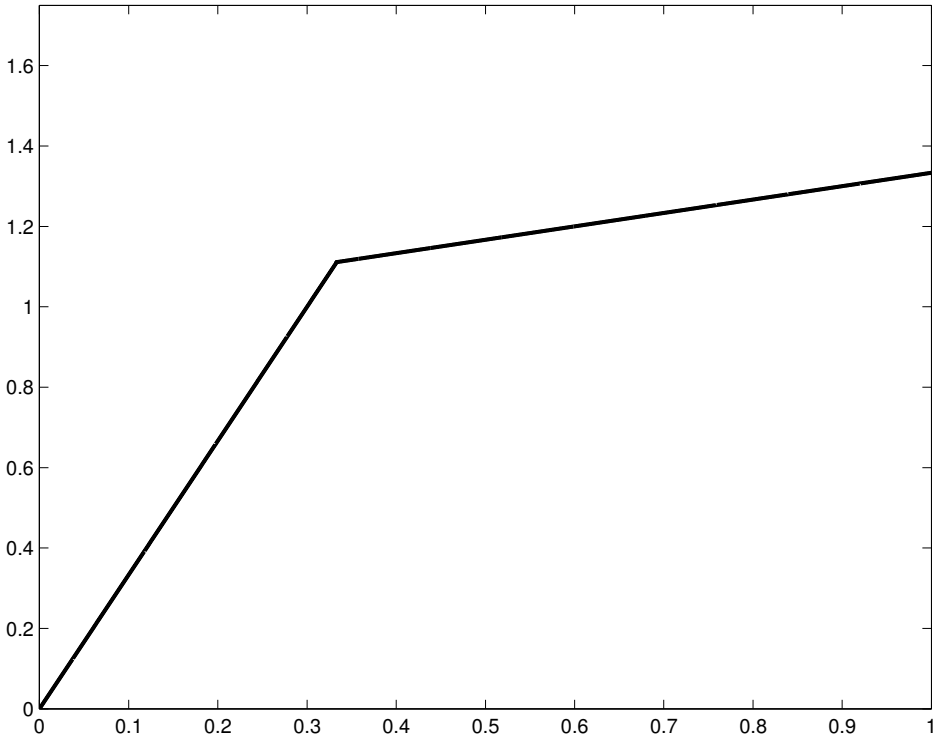
Intuition: how to choose V?

- Bias: decreases with V / can be removed (V-fold penalization)
- Variance: decreases with V / close to its minimum with $V = 5$ or 10

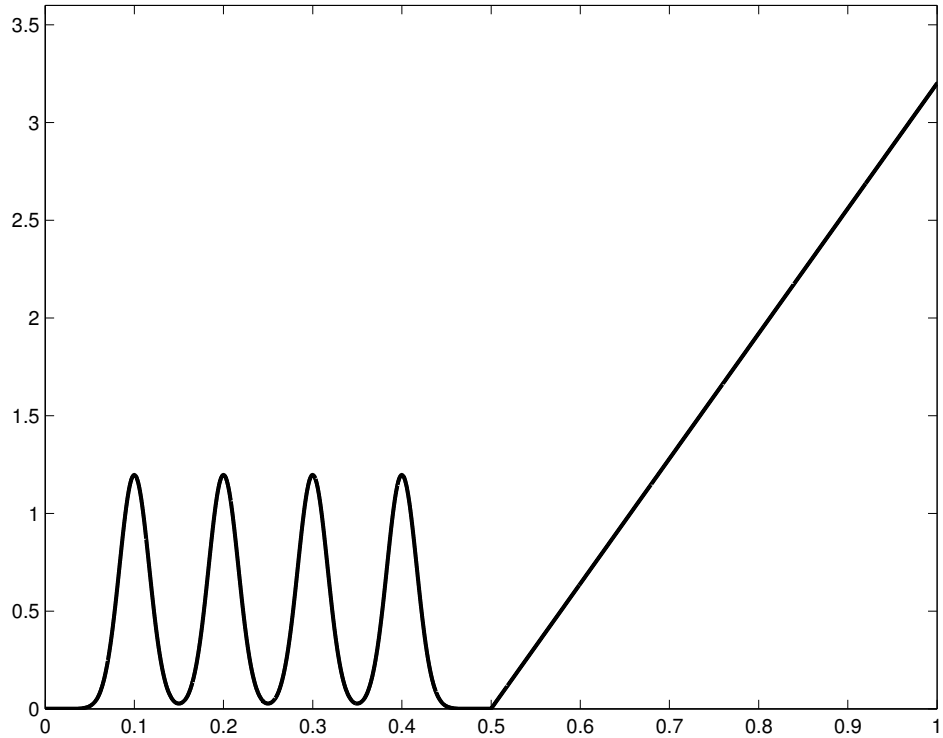
$\Rightarrow$ best performance for the largest V (not necessarily valid in general in learning)

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

Simulation setting: densities

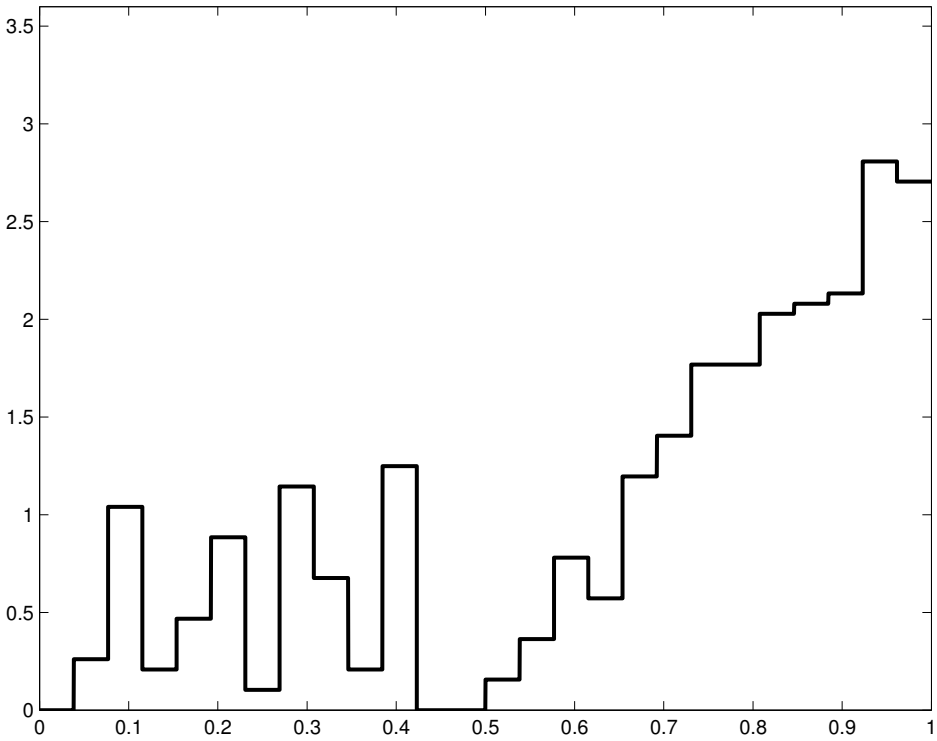


L

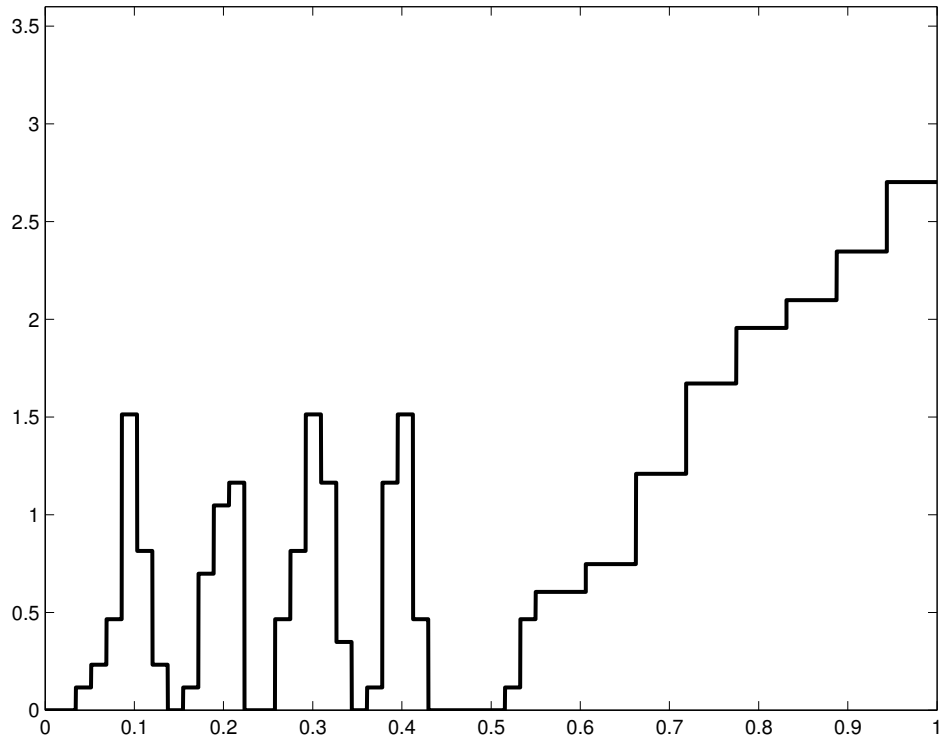


S

Simulation setting: model families



Regu



Dya2

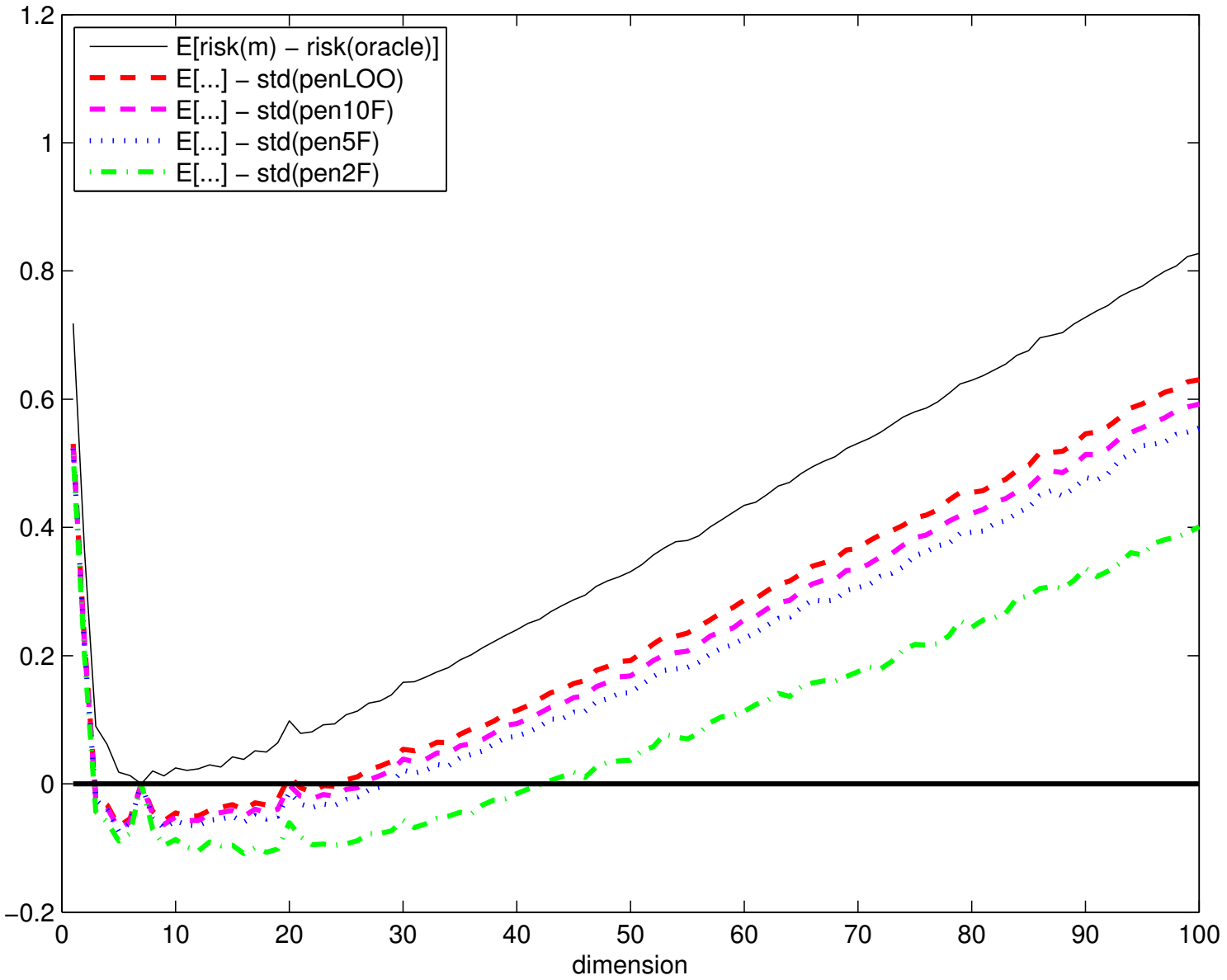
Simulation: $\mathbb{E}[\ell(s^\star, \widehat{s}_{\widehat{m}}) / \inf_{m \in \mathcal{M}} \ell(s^\star, \widehat{s}_m)]$

Procedure	L–Dya2	S–Dya2
C_p	8.52 ± 0.24	3.26 ± 0.04
pen2F	10.27 ± 0.24	2.46 ± 0.03
pen5F	7.53 ± 0.19	2.21 ± 0.03
pen10F	6.76 ± 0.17	2.14 ± 0.03
penLOO	6.41 ± 0.18	2.08 ± 0.03
2FCV	6.41 ± 0.16	2.10 ± 0.02
5FCV	6.27 ± 0.16	2.09 ± 0.03
10FCV	6.25 ± 0.16	2.07 ± 0.03
LOO	6.41 ± 0.18	2.08 ± 0.03
$\mathbb{E}[\mathrm{pen}_{\mathrm{id}}]$	6.62 ± 0.18	2.09 ± 0.03

Simulation: $\mathbb{E}[\ell(s^\star, \widehat{s}_{\widehat{m}}) / \inf_{m \in \mathcal{M}} \ell(s^\star, \widehat{s}_m)]$

Procedure	L–Dya2	S–Dya2
$C^\star \times C_p$	4.38 ± 0.09	3.01 ± 0.04
$C^\star \times$ pen2F	5.12 ± 0.12	2.10 ± 0.02
$C^\star \times$ pen5F	3.80 ± 0.07	1.95 ± 0.02
$C^\star \times$ pen10F	3.66 ± 0.06	1.91 ± 0.02
$C^\star \times$ penLOO	3.61 ± 0.06	1.91 ± 0.02
2FCV	6.41 ± 0.16	2.10 ± 0.02
5FCV	6.27 ± 0.16	2.09 ± 0.03
10FCV	6.25 ± 0.16	2.07 ± 0.03
LOO	6.41 ± 0.18	2.08 ± 0.03
$C^\star \times \mathbb{E}[\mathrm{pen}_{\mathrm{id}}]$	3.66 ± 0.06	1.93 ± 0.02
$C^\star$	2	1.5

Simulation: $\mathbb{E}[\ell(s^\star, \widehat{s}_m) - \ell(s^\star, \widehat{s}_{m^\star})] - \mathrm{std}(\Delta(m, m^\star))$



Simulation: Selectable models