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. Monday 14, 14:00–16:00: Statistical learning
2. Tuesday 15, 9:00–11:00: Model selection for least-squares regression
3. Thursday 17, 14:00–16:00: Linear estimator selection for least-squares regression
4. Tuesday 22, 14:00–16:00: Resampling and model selection
5. Wednesday 23, 9:00–11:00: Cross-validation and model/estimator selection

Part I

Statistical learning

Outline

Outline

General framework

- Data: $\xi_1, \ldots, \xi_n \in \Xi$ i.i.d. $\sim P$
- Goal: estimate a feature $s^\star \in \mathbb{S}$ of P
- Quality measure: loss function

$$\forall t \in \mathbb{S} \enspace , \qquad \mathcal{L}_P (t) = \mathbb{E}_{\xi \sim P} \left[\gamma(t;\xi) \right] = P\gamma(t)$$

minimal at $t = s^\star$
Contrast function: $\gamma : \mathbb{S} \times \Xi \mapsto [0, +\infty)$

- Excess loss

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

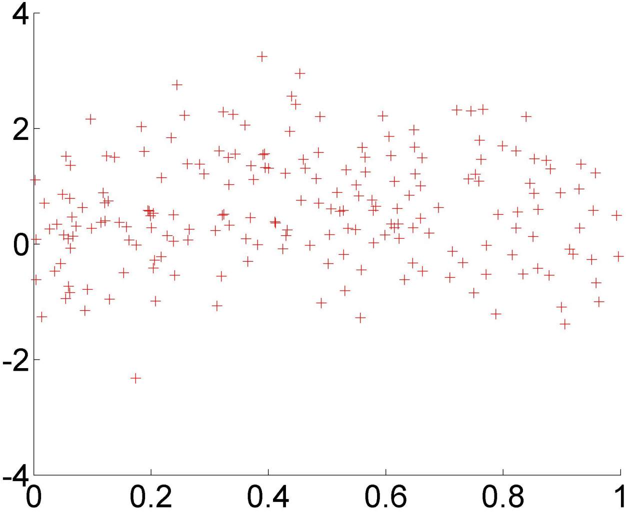
Example: prediction

- Data: $(X_1, Y_1), \ldots, (X_n, Y_n) \in \Xi = \mathcal{X} \times \mathcal{Y}$
- Goal: predict Y given X with $(X, Y) = \xi \sim P$
- $s^\star(X)$ is the "best predictor" of Y given X, i.e., $s^\star$ minimizes the loss function

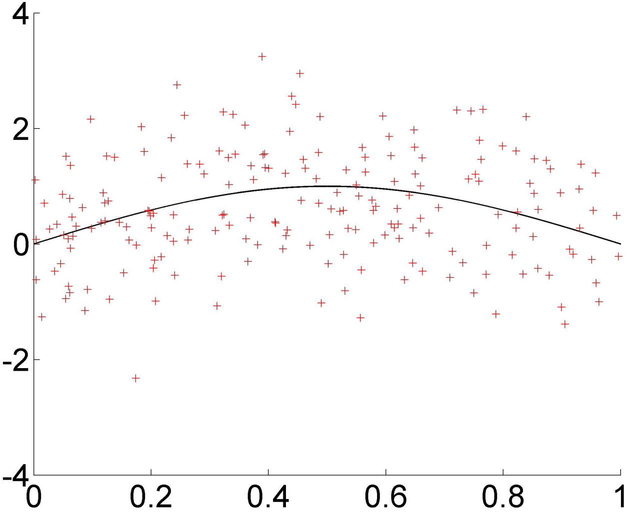
$$P\gamma(t) \quad \text{with} \quad \gamma(t; (x, y)) = d(t(x), y)$$

measuring some "distance" between y and the prediction $t(x)$.

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



Goal: find the signal (denoising)



Example: regression

- prediction with $\mathcal{Y} = \mathbb{R}$
- Data: $(X_1, Y_1), \ldots, (X_n, Y_n)$ i.i.d.

$$Y_i = \eta(X_i) + \varepsilon_i \quad \text{with} \quad \mathbb{E}\left[\varepsilon_i \mid X_i\right] = 0$$

Example: regression

- prediction with $\mathcal{Y} = \mathbb{R}$
- Data: $(X_1, Y_1), \ldots, (X_n, Y_n)$ i.i.d.

$$Y_i = \eta(X_i) + \varepsilon_i \quad \text{with} \quad \mathbb{E}[\varepsilon_i \mid X_i] = 0$$

- least-squares contrast: $\gamma(t; (x, y)) = (t(x) - y)^2$

$$\Rightarrow \quad s^\star = \eta \quad \text{and} \quad \ell(s^\star, t) = \| t - \eta \|_2^2 = \mathbb{E}\left[\left(t(X) - \eta(X) \right)^2 \right]$$

Example: regression on a fixed design

- $(X_1, \ldots, X_n) = (x_1, \ldots, x_n)$ deterministic

$$Y = F + \varepsilon \in \mathbb{R}^n \quad \text{with} \quad F = (\eta(x_1), \ldots, \eta(x_n)) \in \mathbb{R}^n$$

and $\varepsilon_1, \ldots, \varepsilon_n$ centered and independent.

Example: regression on a fixed design

- $(X_1, \ldots, X_n) = (x_1, \ldots, x_n)$ deterministic

$$Y = F + \varepsilon \in \mathbb{R}^n \quad \text{with} \quad F = (\eta(x_1), \ldots, \eta(x_n)) \in \mathbb{R}^n$$

  and $\varepsilon_1, \ldots, \varepsilon_n$ centered and independent.
- Homoscedastic case: $\varepsilon_1, \ldots, \varepsilon_n$ i.i.d.

Example: regression on a fixed design

- $(X_1, \ldots, X_n) = (x_1, \ldots, x_n)$ deterministic

$$Y = F + \varepsilon \in \mathbb{R}^n \quad \text{with} \quad F = (\eta(x_1), \ldots, \eta(x_n)) \in \mathbb{R}^n$$

and $\varepsilon_1, \ldots, \varepsilon_n$ centered and independent.

- Homoscedastic case: $\varepsilon_1, \ldots, \varepsilon_n$ i.i.d.
- Quadratic loss of $t \in \mathbb{S} = \mathbb{R}^n$:

$$\mathcal{L}_P(t) = \mathbb{E}_Y\left[\frac{1}{n}\|Y - t\|^2\right] = \mathbb{E}_Y\left[\frac{1}{n}\sum_{i=1}^{n}(Y_i - t_i)^2\right]$$

$$\Rightarrow \quad s^\star = F \quad \text{and} \quad \ell(s^\star, t) = \frac{1}{n}\|F - t\|^2 = \frac{1}{n}\sum_{i=1}^{n}(\eta(x_i) - t_i)^2$$

Example: regression: fixed vs. random design

	Random design	Fixed design
D_n	$(X_i, Y_i)_{1 \le i \le n}$ i.i.d. $\sim P$	$Y = F + \varepsilon \in \mathbb{R}^n$
	$(X_{n+1}, Y_{n+1}) \sim P$	$X_{n+1} \sim \mathcal{U}(x_1, \ldots, x_n)$
$\mathbb{S}$	$t : \mathcal{X} \to \mathbb{R}$	$t \in \mathbb{R}^n$
$P\gamma(t)$	$\mathbb{E}_{(X,Y)\sim P}\left[(Y - t(X))^2 \right]$	$E_Y\left[\frac{1}{n} \lVert Y - t \rVert^2 \right]$
$s^\star$	$\eta : x \to \mathbb{E}[Y \mid X = x]$	$F = (\eta(x_1), \ldots, \eta(x_n))$
$\ell(s^\star, t)$	$\mathbb{E}_{(X,Y)\sim P}\left[(t(X) - \eta(X))^2 \right]$	$\frac{1}{n} \lVert F - t \rVert^2$

with $\forall x \in \mathbb{R}^n$, $\lVert x \rVert^2 = \sum_{i=1}^{n} x_i^2$

Example: regression: fixed vs. random design

	Random design	Fixed design
D_n	$(X_i, Y_i)_{1 \leq i \leq n}$ i.i.d. $\sim P$	$Y = F + \varepsilon \in \mathbb{R}^n$
	$(X_{n+1}, Y_{n+1}) \sim P$	$X_{n+1} \sim \mathcal{U}(x_1, \ldots, x_n)$
$\mathbb{S}$	$t : \mathcal{X} \to \mathbb{R}$	$t \in \mathbb{R}^n$
$P\gamma(t)$	$\mathbb{E}_{(X,Y)\sim P}\left[(Y - t(X))^2 \right]$	$E_Y \left[\frac{1}{n} \lVert Y - t \rVert^2 \right]$
$s^\star$	$\eta : x \to \mathbb{E}[Y \mid X = x]$	$F = (\eta(x_1), \ldots, \eta(x_n))$
$\ell(s^\star, t)$	$\mathbb{E}_{(X,Y)\sim P}\left[(t(X) - \eta(X))^2 \right]$	$\frac{1}{n} \lVert F - t \rVert^2$

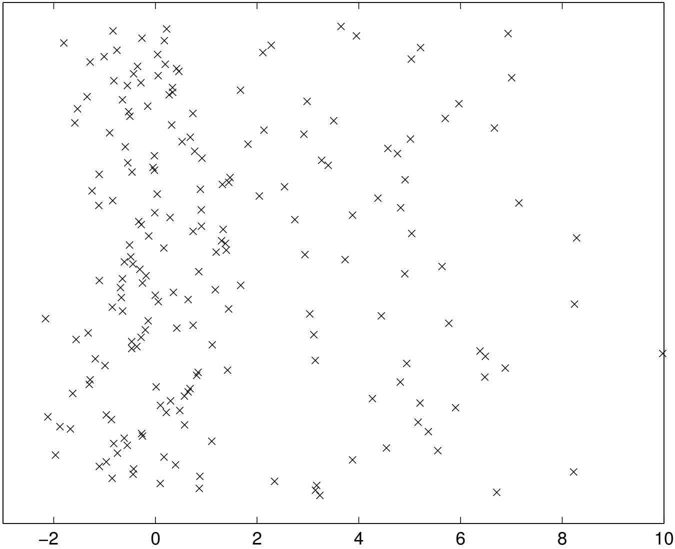
with $\forall x \in \mathbb{R}^n$, $\lVert x \rVert^2 = \sum_{i=1}^{n} x_i^2$

Example: regression: fixed vs. random design

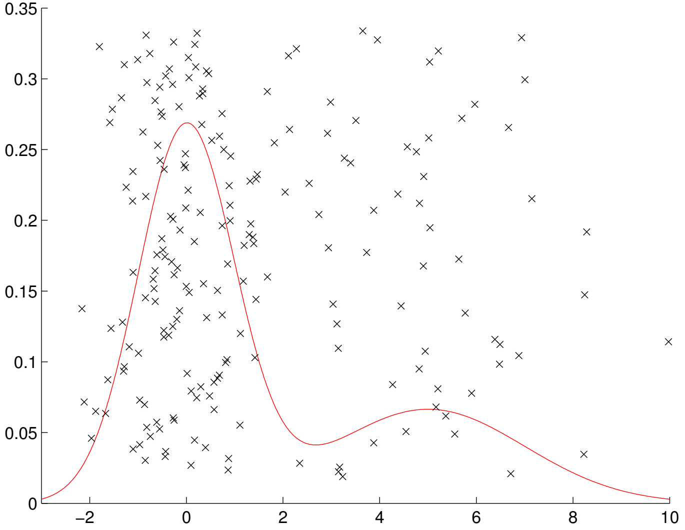
	Random design	Fixed design
D_n	$(X_i, Y_i)_{1 \leq i \leq n}$ i.i.d. $\sim P$	$Y = F + \varepsilon \in \mathbb{R}^n$
	$(X_{n+1}, Y_{n+1}) \sim P$	$X_{n+1} \sim \mathcal{U}(x_1, \ldots, x_n)$
$\mathbb{S}$	$t : \mathcal{X} \to \mathbb{R}$	$t \in \mathbb{R}^n$
$P\gamma(t)$	$\mathbb{E}_{(X,Y)\sim P}\left[(Y - t(X))^2\right]$	$E_Y\left[\frac{1}{n}\|Y - t\|^2\right]$
$s^\star$	$\eta : x \to \mathbb{E}[Y \mid X = x]$	$F = (\eta(x_1), \ldots, \eta(x_n))$
$\ell(s^\star, t)$	$\mathbb{E}_{(X,Y)\sim P}\left[(t(X) - \eta(X))^2\right]$	$\frac{1}{n}\|F - t\|^2$

$$\text{with} \quad \forall x \in \mathbb{R}^n \; , \qquad \|x\|^2 = \sum_{i=1}^{n} x_i^2$$

Example: density estimation ($\Xi = \mathbb{R}$): data



Example: density estimation ($\Xi = \mathbb{R}$): data and target



Density estimation

- μ reference measure on Ξ
- f density of P w.r.t. μ

Density estimation

- μ reference measure on Ξ
- f density of P w.r.t. μ

- $\gamma(t;\xi) = -\ln(t(\xi))$
  $\Rightarrow s^\star = f$ and $\ell(s^\star, t)$ Kullback-Leibler distance from $s^\star$ to t

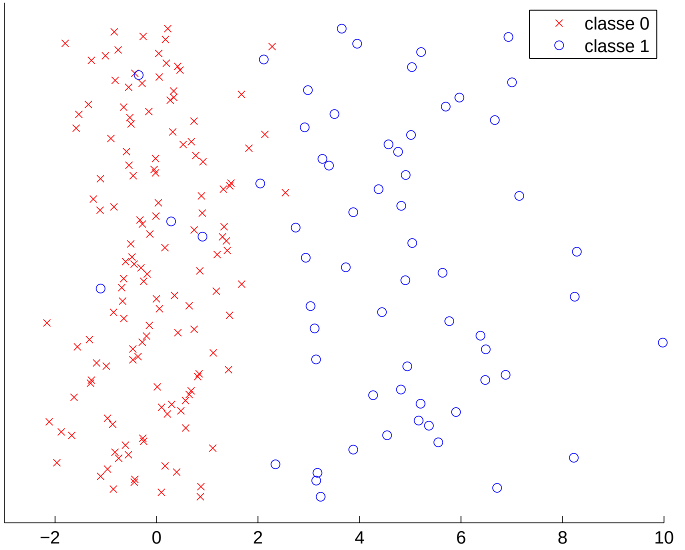
Density estimation

- μ reference measure on Ξ
- f density of P w.r.t. μ

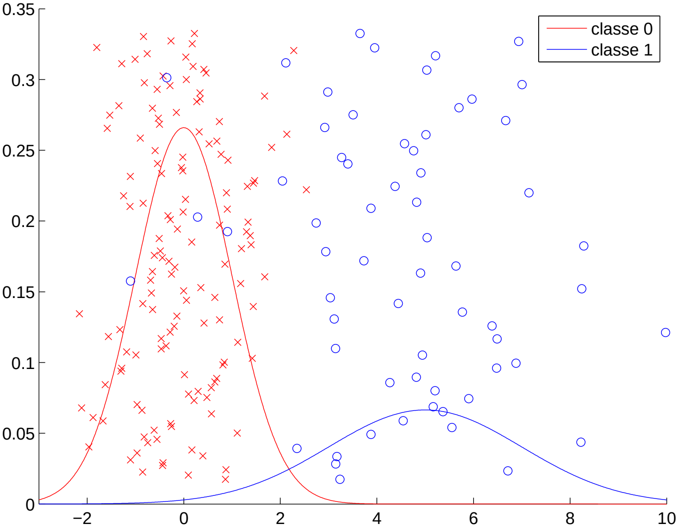
- $\gamma(t;\xi) = -\ln(t(\xi))$
  $\Rightarrow s^\star = f$ and $\ell(s^\star, t)$ Kullback-Leibler distance from $s^\star$ to t

- $\gamma(t;\xi) = \|t\|^2_{L^2(\mu)} - 2t(\xi)$
  $\Rightarrow s^\star = f$ and $\ell(s^\star, t) = \|t - s^\star\|^2_{L^2(\mu)}$

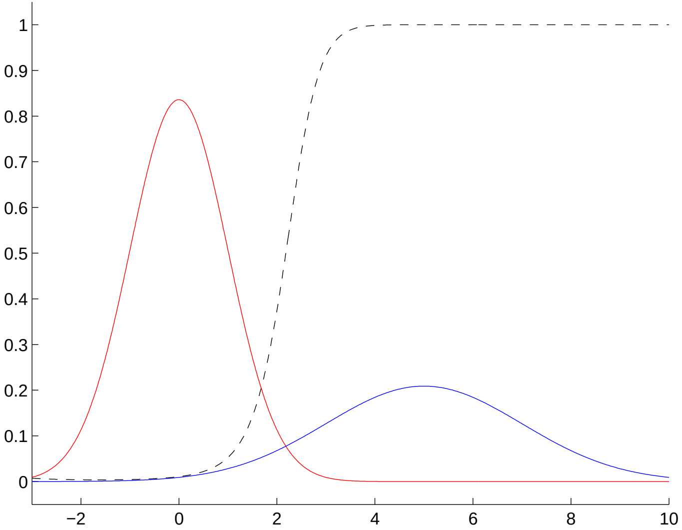
Example: classification (prediction, $\mathcal{X} = \mathbb{R}$, $\mathcal{Y} = \{0, 1\}$)



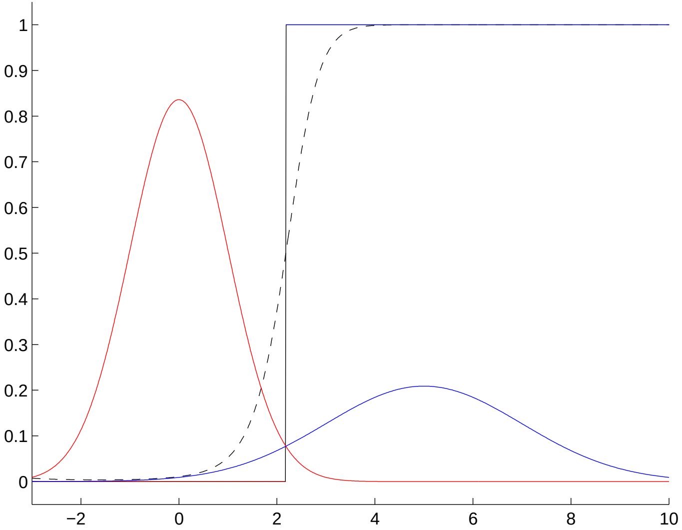
Example: classification (prediction, $\mathcal{X} = \mathbb{R}$, $\mathcal{Y} = \{0, 1\}$)



Example: classification (prediction, $\mathcal{X} = \mathbb{R}$, $\mathcal{Y} = \{0, 1\}$)



Example: classification (prediction, $\mathcal{X} = \mathbb{R}$, $\mathcal{Y} = \{0, 1\}$)



Example: binary supervised classification

- Prediction, $\mathcal{X} = \mathbb{R}$ and $\mathcal{Y} = \{0, 1\}$
- If $\mathbb{S} = \{$measurable mappings $\mathcal{X} \mapsto \mathcal{Y}\}$
  0–1 loss: $\gamma(t; (x, y)) = \mathbb{1}_{t(x) \neq y}$

Example: binary supervised classification

- Prediction, $\mathcal{X} = \mathbb{R}$ and $\mathcal{Y} = \{0, 1\}$
- If $\mathbb{S} = \{$measurable mappings $\mathcal{X} \mapsto \mathcal{Y}\}$
  0–1 loss: $\gamma(t; (x, y)) = \mathbb{1}_{t(x) \neq y}$

- If $t \in \mathbb{S} = \{$measurable mappings $\mathcal{X} \mapsto [0, 1]\}$,
  Convex losses: $\gamma(t; (x, y)) = \varphi(t(x)(1 - 2y))$ with $\varphi : \mathbb{R} \mapsto \mathbb{R}$ convex, non-negative, non-increasing.

Outline

What is an estimator?

- Statistical algorithm or Learning rule:
$$\mathcal{A}\colon \bigcup_{n\in\mathbb{N}} \Xi^n \mapsto \mathbb{S}$$
$$\text{sample } D_n = (\xi_1, \ldots, \xi_n) \mapsto \mathcal{A}(D_n)$$
- $\mathcal{A}(D_n) = \widehat{s}^{\mathcal{A}}(D_n) = \widehat{s}(D_n) \in \mathbb{S}$ is an estimator of $s^{\star}$

What is an estimator?

- Statistical algorithm or Learning rule:

$$\mathcal{A}\colon \bigcup_{n \in \mathbb{N}} \Xi^n \mapsto \mathbb{S}$$

$$\text{sample } D_n = (\xi_1, \ldots, \xi_n) \mapsto \mathcal{A}(D_n)$$

- $\mathcal{A}(D_n) = \widehat{s}^{\mathcal{A}}(D_n) = \widehat{s}(D_n) \in \mathbb{S}$ is an estimator of $s^\star$

- Remark: $P\gamma\left(\widehat{s}^{\mathcal{A}}(D_n)\right)$ and $\ell\left(s^\star, \widehat{s}^{\mathcal{A}}(D_n)\right)$ are random

What is an estimator?

- Statistical algorithm or Learning rule:

$$\mathcal{A}\colon \bigcup_{n \in \mathbb{N}} \Xi^n \mapsto \mathbb{S}$$
$$\text{sample } D_n = (\xi_1, \ldots, \xi_n) \mapsto \mathcal{A}(D_n)$$

- $\mathcal{A}(D_n) = \widehat{s}^{\mathcal{A}}(D_n) = \widehat{s}(D_n) \in \mathbb{S}$ is an estimator of $s^\star$
- Remark: $P\gamma\left(\widehat{s}^{\mathcal{A}}(D_n)\right)$ and $\ell\left(s^\star, \widehat{s}^{\mathcal{A}}(D_n)\right)$ are random
- Risk of $\widehat{s}^{\mathcal{A}}$:

$$\mathbb{E}_{D_n \sim P^{\otimes n}}\left[P\gamma\left(\widehat{s}^{\mathcal{A}}(D_n)\right)\right] = \mathcal{R}(\mathcal{A}, n)$$

- Excess risk of $\widehat{s}^{\mathcal{A}}$:

$$\mathbb{E}_{D_n \sim P^{\otimes n}}\left[\ell\left(s^\star, \widehat{s}^{\mathcal{A}}(D_n)\right)\right] = \mathcal{R}(\mathcal{A}, n) - P\gamma\left(s^\star\right)$$

(Universal) consistency, learning rates

- Consistency (P fixed): $\ell\left(s^{\star}, \widehat{s}^{\mathcal{A}}(D_n)\right) \rightarrow 0$ as $n \rightarrow +\infty$

(Universal) consistency, learning rates

- Consistency (P fixed): $\ell\left(s^\star, \widehat{s}^{\mathcal{A}}(D_n)\right) \to 0$ as $n \to +\infty$
- Universal consistency:
  $\sup_P \left\{ \overline{\lim}_{n\to\infty} \mathbb{E}_{D_n \sim P^{\otimes n}} \left[\ell\left(s^\star, \widehat{s}^{\mathcal{A}}(D_n)\right) \right] \right\} = 0$

(Universal) consistency, learning rates

- Consistency (P fixed): $\ell\left(s^\star, \widehat{s}^{\mathcal{A}}(D_n)\right) \to 0$ as $n \to +\infty$
- Universal consistency:
  $\sup_P \left\{ \overline{\lim}_{n \to \infty} \mathbb{E}_{D_n \sim P^{\otimes n}} \left[\ell\left(s^\star, \widehat{s}^{\mathcal{A}}(D_n)\right) \right] \right\} = 0$
- Uniform universal consistency:
  $\overline{\lim}_{n \to \infty} \sup_P \left\{ \mathbb{E}_{D_n \sim P^{\otimes n}} \left[\ell\left(s^\star, \widehat{s}^{\mathcal{A}}(D_n)\right) \right] \right\} = 0$ (uniform learning rate over all distributions).

(Universal) consistency, learning rates

- Consistency (P fixed): $\ell\left(s^\star, \widehat{s}^{\mathcal{A}}(D_n)\right) \to 0$ as $n \to +\infty$
- Universal consistency:
  $\sup_P \left\{ \overline{\lim}_{n\to\infty} \mathbb{E}_{D_n \sim P^{\otimes n}} \left[\ell\left(s^\star, \widehat{s}^{\mathcal{A}}(D_n)\right) \right] \right\} = 0$
- Uniform universal consistency:
  $\overline{\lim}_{n\to\infty} \sup_P \left\{ \mathbb{E}_{D_n \sim P^{\otimes n}} \left[\ell\left(s^\star, \widehat{s}^{\mathcal{A}}(D_n)\right) \right] \right\} = 0$ (uniform learning rate over all distributions).
- "No Free Lunch" (cf. Devroye, Györfi & Lugosi, 1996):

(Universal) consistency, learning rates

- Consistency (P fixed): $\ell\left(s^\star, \widehat{s}^{\mathcal{A}}(D_n)\right) \to 0$ as $n \to +\infty$

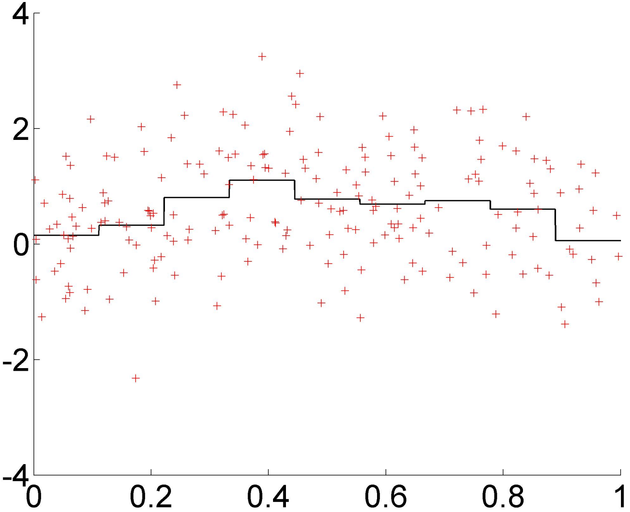
- Universal consistency:
  $\sup_P \left\{ \overline{\lim}_{n\to\infty} \mathbb{E}_{D_n \sim P^{\otimes n}} \left[\ell\left(s^\star, \widehat{s}^{\mathcal{A}}(D_n)\right) \right] \right\} = 0$
- Uniform universal consistency:
  $\overline{\lim}_{n\to\infty} \sup_P \left\{ \mathbb{E}_{D_n \sim P^{\otimes n}} \left[\ell\left(s^\star, \widehat{s}^{\mathcal{A}}(D_n)\right) \right] \right\} = 0$ (uniform learning rate over all distributions).

- "No Free Lunch" (cf. Devroye, Györfi & Lugosi, 1996):
  In binary classification with $\mathcal{X}$ infinite, $\forall \mathcal{A}$, $\forall n \geq 1$,

$$\sup_P \left\{ \mathbb{E}_{D_n \sim P^{\otimes n}} \left[\ell\left(s^\star, \widehat{s}^{\mathcal{A}}(D_n)\right) \right] \right\} = \frac{1}{2}$$

$\Rightarrow$ assumptions on P are necessary for having uniform learning rates

Least-squares estimator: regressogram



Least-squares estimator

- Framework: Regression, least-squares contrast
  $\gamma(t; (x, y)) = (t(x) - y)^2$
- Natural idea: minimize an estimator of
  $P\gamma(t) = \mathbb{E}\left[(t(X) - Y)^2 \right]$

Least-squares estimator

- Framework: Regression, least-squares contrast $\gamma(t;(x,y)) = (t(x) - y)^2$
- Natural idea: minimize an estimator of $P\gamma(t) = \mathbb{E}\left[(t(X) - Y)^2 \right]$
- Least-squares criterion:

$$P_n\gamma(t) = \frac{1}{n} \sum_{i=1}^{n} (t(X_i) - Y_i)^2 \qquad \text{with} \quad P_n = \frac{1}{n} \sum_{i=1}^{n} \delta_{\xi_i}$$

$$\forall t \in \mathbb{S} \ , \qquad \mathbb{E}\left[P_n\gamma(t) \right] = P\gamma(t)$$

Least-squares estimator

- Framework: Regression, least-squares contrast $\gamma(t;(x,y)) = (t(x) - y)^2$
- Natural idea: minimize an estimator of $P\gamma(t) = \mathbb{E}\left[(t(X) - Y)^2 \right]$
- Least-squares criterion:

$$P_n\gamma(t) = \frac{1}{n}\sum_{i=1}^{n} (t(X_i) - Y_i)^2 \qquad \text{with} \quad P_n = \frac{1}{n}\sum_{i=1}^{n} \delta_{\xi_i}$$

$$\forall t \in \mathbb{S} \ , \qquad \mathbb{E}\left[P_n\gamma(t) \right] = P\gamma(t)$$

- Model: $S \subset \mathbb{S} \Rightarrow$ Least-squares estimator on S:

$$\widehat{s}_S \in \arg\min_{t \in S} \{ P_n\gamma(t) \} = \arg\min_{t \in S} \left\{ \frac{1}{n}\sum_{i=1}^{n} (t(X_i) - Y_i)^2 \right\}$$

Model examples in regression

- histograms on some partition Λ of $\mathcal{X}$
  $\Rightarrow$ the least-squares estimator (regressogram) can be written

$$\widehat{s}_m = \sum_{\lambda \in \Lambda} \widehat{\beta}_\lambda \mathbb{1}_\lambda \qquad \widehat{\beta}_\lambda = \frac{1}{\operatorname{Card}\left\{ X_i \in \lambda \right\}} \sum_{X_i \in \lambda} Y_i$$

- subspace generated by a subset of an orthogonal basis of $L^2(\mu)$ (Fourier, wavelets, and so on)
- variable selection: $X_i = \left(X_i^{(1)}, \ldots, X_i^{(p)} \right) \in \mathbb{R}^p$ gathers p variables that can (linearly) explain Y

$$\forall m \subset \left\{ 1, \ldots, p \right\} \quad , \quad S_m = \left\{ t : x \in \mathcal{X} \mapsto \sum_{j \in m} \beta_j x^{(j)} \text{ s.t. } \beta \in \mathbb{R}^m \right\}$$

Regression: fixed vs. random design

	Random design	Fixed design
D_n	$(X_i, Y_i)_{1 \leq i \leq n}$ i.i.d. $\sim P$	$Y = F + \varepsilon \in \mathbb{R}^n$
	$(X_{n+1}, Y_{n+1}) \sim P$	$X_{n+1} \sim \mathcal{U}(x_1, \ldots, x_n)$
$\mathbb{S}$	$t : \mathcal{X} \to \mathbb{R}$	$t \in \mathbb{R}^n$
$P\gamma(t)$	$\mathbb{E}_{(X,Y)\sim P}\left[(Y - t(X))^2 \right]$	$E_Y \left[\frac{1}{n} \lVert Y - t \rVert^2 \right]$
$s^\star$	$\eta : x \to \mathbb{E}[Y \mid X = x]$	$F = (\eta(x_1), \ldots, \eta(x_n))$
$\ell(s^\star, t)$	$\mathbb{E}_{(X,Y)\sim P}\left[(t(X) - \eta(X))^2 \right]$	$\frac{1}{n} \lVert F - t \rVert^2$
	$P_n\gamma(t) = \frac{1}{n} \sum_{i=1}^n (Y_i - t(X_i))^2$	$\frac{1}{n} \lVert Y - t \rVert^2$

with $\forall x \in \mathbb{R}^n$, $\lVert x \rVert^2 = \sum_{i=1}^{n} x_i^2$

Minimum contrast estimators

- Empirical risk (or empirical contrast)

$$P_n \gamma(t) = \frac{1}{n} \sum_{i=1}^{n} \gamma\left(t ; \xi_i \right)$$

- $\forall t \in \mathbb{S}$, $\mathbb{E}\left[P_n \gamma(t) \right] = P\gamma(t)$
- Minimum contrast estimator (empirical risk minimizer) on some model $S \subset \mathbb{S}$:

$$\widehat{s}_S \in \arg\min_{t \in S} P_n \gamma(t) \quad \text{with} \quad P_n = \frac{1}{n} \sum_{i=1}^{n} \delta_{\xi_i}$$

- Another example: maximum-likelihood in density estimation: $\gamma(t;\xi) = -\ln(t(\xi))$

Regularized estimator: kernel ridge regression

- Idea: control the estimator norm in some functional space $\mathcal{F}$

Regularized estimator: kernel ridge regression

- Idea: control the estimator norm in some functional space $\mathcal{F}$
- $\mathcal{F} \subset \mathbb{S}$ is the Reproducing Kernel Hilbert Space (RKHS) associated with a positive definite kernel $k : \mathcal{X} \times \mathcal{X} \mapsto \mathbb{R}$

$$\widehat{f} \in \arg\min_{f \in \mathcal{F}} \left\{ \frac{1}{n} \sum_{i=1}^{n} \left(Y_i - f(X_i) \right)^2 + \lambda \left\| f \right\|_{\mathcal{F}}^2 \right\}$$

Regularized estimator: kernel ridge regression

- Idea: control the estimator norm in some functional space $\mathcal{F}$
- $\mathcal{F} \subset \mathbb{S}$ is the Reproducing Kernel Hilbert Space (RKHS) associated with a positive definite kernel $k : \mathcal{X} \times \mathcal{X} \mapsto \mathbb{R}$

$$\widehat{f} \in \arg\min_{f \in \mathcal{F}} \left\{ \frac{1}{n} \sum_{i=1}^{n} \left(Y_i - f(X_i) \right)^2 + \lambda \left\| f \right\|_{\mathcal{F}}^2 \right\}$$

- Representer theorem $\Rightarrow \widehat{f} = \sum_{i=1}^{n} \widehat{\alpha}_i k(X_i, \cdot)$
- Fixed design: $(\widehat{f}(x_i))_{1 \leq i \leq n} = \widehat{F} = K(K + n\lambda I_n)^{-1} Y$

Regularized estimator: kernel ridge regression

- Idea: control the estimator norm in some functional space $\mathcal{F}$
- $\mathcal{F} \subset \mathbb{S}$ is the Reproducing Kernel Hilbert Space (RKHS) associated with a positive definite kernel $k : \mathcal{X} \times \mathcal{X} \mapsto \mathbb{R}$

$$\widehat{f} \in \arg\min_{f \in \mathcal{F}} \left\{ \frac{1}{n} \sum_{i=1}^{n} \left(Y_i - f(X_i) \right)^2 + \lambda \left\| f \right\|_{\mathcal{F}}^2 \right\}$$

- Representer theorem $\Rightarrow \widehat{f} = \sum_{i=1}^{n} \widehat{\alpha}_i k(X_i, \cdot)$
- Fixed design: $(\widehat{f}(x_i))_{1 \leq i \leq n} = \widehat{F} = K(K + n\lambda I_n)^{-1} Y$
- An example of linear estimator $\widehat{F} = AY$
  Other examples: least-squares, k-nearest-neighbours (in regression), Nadaraya-Watson, and so on

Other regularized estimators

- Support Vector Machines (SVM) in classification:

$$\arg\min_{f \in \mathcal{F}} \left\{ P_n \gamma_{\mathrm{hinge}}(f) + \lambda \left\| f \right\|_{\mathcal{F}}^2 \right\}$$

- Lasso (Tibshirani 1996): regression, $\mathcal{X} = \mathbb{R}^p$

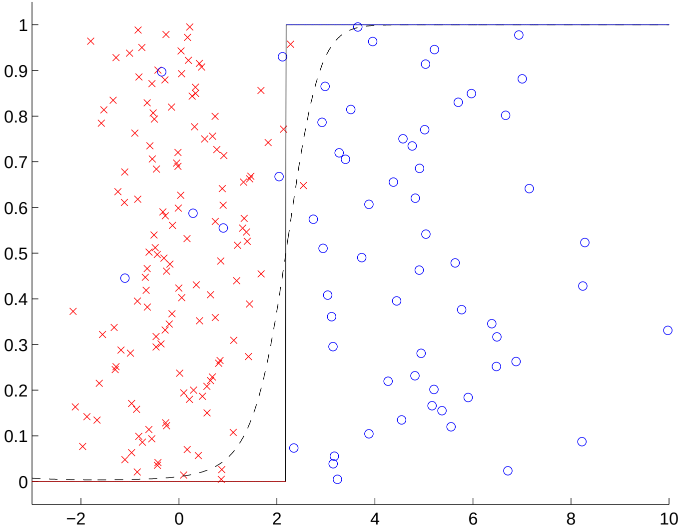
$$\arg\min_{w \in \mathbb{R}^p} \left\{ \frac{1}{2} \sum_{i=1}^{n} \left(Y_i - w^\top X_i \right)^2 + \lambda \left\| w \right\|_1 \right\}$$

- Structured Lasso

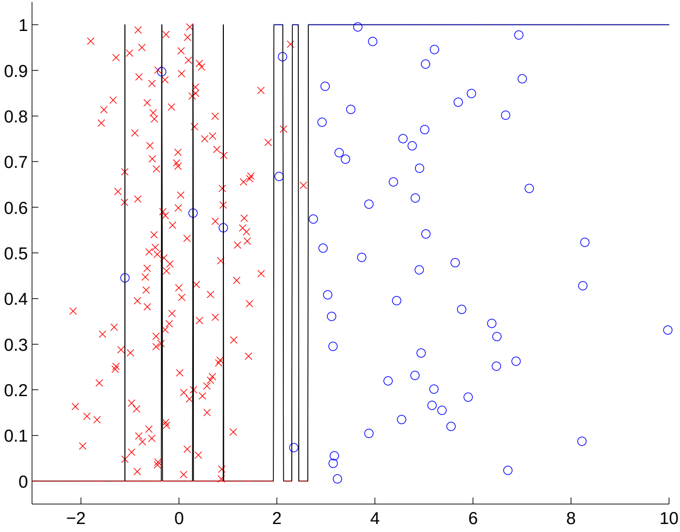
$$\arg\min_{w \in \mathbb{R}^p} \left\{ \frac{1}{2} \sum_{i=1}^{n} \left(Y_i - w^\top X_i \right)^2 + \lambda \Omega(w) \right\}$$

e.g., group Lasso (Yuan & Lin 2006): $\Omega(w) = \sum_{g \in \mathcal{G}} \left\| w_g \right\|_2$

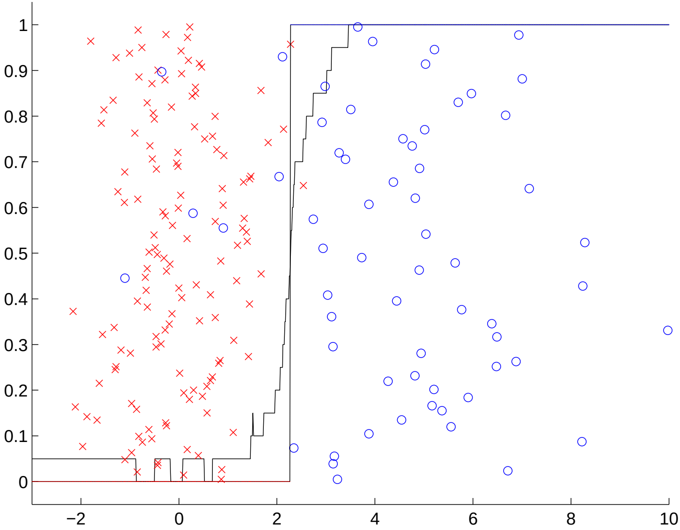
Classification ($\mathcal{X} = \mathbb{R}$)



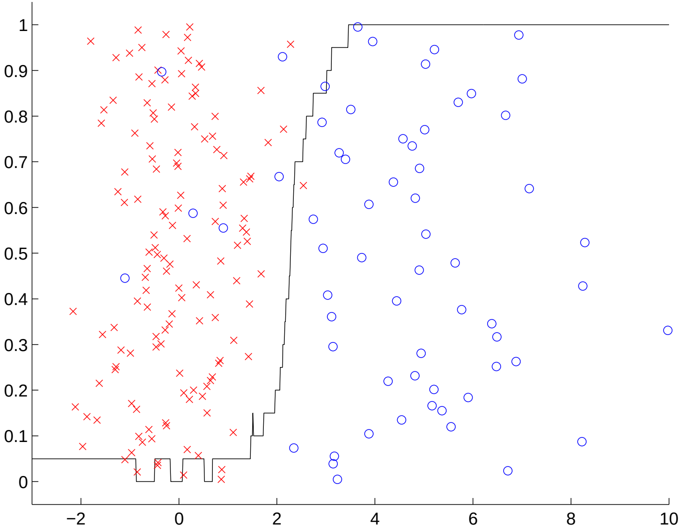
Nearest neighbour rule



k-nearest neighbours rule ($k = 20$)

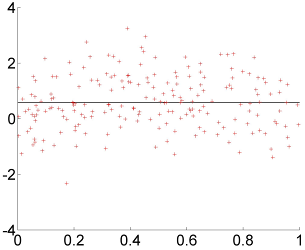


20-nearest neighbours rule: regression

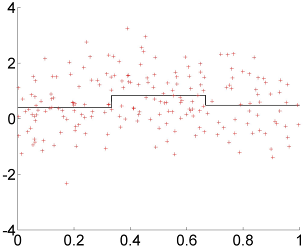


Outline

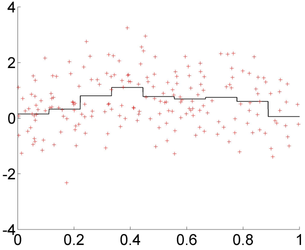
How to choose the dimension D?



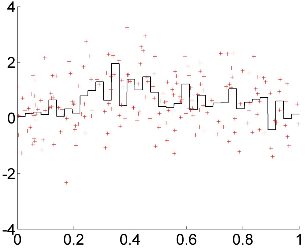
$D = 1$



$D = 3$

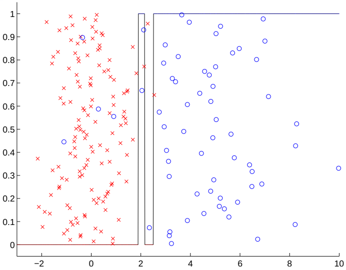


$D = 9$

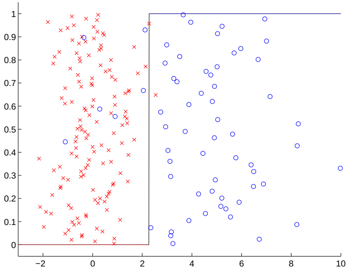


$D = 36$

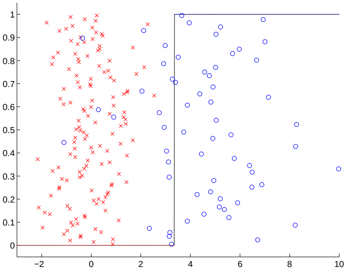
How to choose the number k of neighbours?



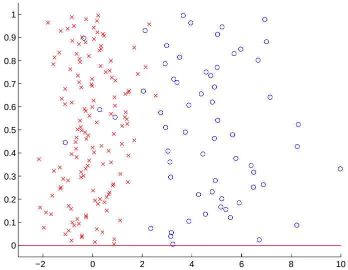
$k = 3$



$k = 20$



$k = 100$



$k = 200$

Estimator selection problem

- Collection of statistical algorithms given: $(\mathcal{A}_m)_{m \in \mathcal{M}}$
- Problem: choosing among $(\mathcal{A}_m(D_n))_{m \in \mathcal{M}} = (\widehat{s}_m(D_n))_{m \in \mathcal{M}}$

Estimator selection problem

- Collection of statistical algorithms given: $(\mathcal{A}_m)_{m\in\mathcal{M}}$
- Problem: choosing among $(\mathcal{A}_m(D_n))_{m\in\mathcal{M}} = (\widehat{s}_m(D_n))_{m\in\mathcal{M}}$

- Examples:
  - model selection
  - calibration (choice of k or of the distance for k-NN, choice of the regularization parameter, choice of some kernel, and so on)
  - choosing among algorithms of different nature, e.g., k-NN and SVM

Goal: estimation or prediction

- Main goal: find $\widehat{m}$ minimizing $\ell\left(s^{\star}, \widehat{s}_{\widehat{m}(D_n)}(D_n)\right)$
- Oracle: $m^{\star} \in \arg\min_{m \in \mathcal{M}_n} \left\{ \ell\left(s^{\star}, \widehat{s}_m(D_n)\right) \right\}$

Goal: estimation or prediction

- Main goal: find $\widehat{m}$ minimizing $\ell\left(s^\star, \widehat{s}_{\widehat{m}(D_n)}(D_n)\right)$
- Oracle: $m^\star \in \arg\min_{m \in \mathcal{M}_n} \left\{ \ell\left(s^\star, \widehat{s}_m(D_n)\right) \right\}$

- Oracle inequality (in expectation or with high probability):

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

- Non-asymptotic: all parameters can vary with n, in particular the collection $\mathcal{M} = \mathcal{M}_n$

Goal: estimation or prediction

- Main goal: find $\widehat{m}$ minimizing $\ell\left(s^\star, \widehat{s}_{\widehat{m}(D_n)}(D_n)\right)$
- Oracle: $m^\star \in \arg\min_{m \in \mathcal{M}_n} \left\{ \ell\left(s^\star, \widehat{s}_m(D_n)\right) \right\}$
- Oracle inequality (in expectation or with high probability):

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

- Non-asymptotic: all parameters can vary with n, in particular the collection $\mathcal{M} = \mathcal{M}_n$
- Adaptation (e.g., in the minimax sense) to the regularity of $s^\star$, to variations of $\mathbb{E}\left[\varepsilon^2 \,\middle|\, X\right]$, and so on (if $(\mathcal{A}_m)_{m \in \mathcal{M}_n}$ is well chosen)

Goal: identification

- Additional assumption (model selection case): $s^\star \in S_{m_0}$ for some $m_0 \in \mathcal{M}_n$
- Additional goal: select $\widehat{m} = m_0$ with a maximal probability
- Consistency:

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

Goal: identification

- Additional assumption (model selection case): $s^\star \in S_{m_0}$ for some $m_0 \in \mathcal{M}_n$
- Additional goal: select $\widehat{m} = m_0$ with a maximal probability
- Consistency:

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

- Estimation and identification (AIC-BIC dilemma)?
  Contradictory goals in general (Yang, 2005)
  Sometimes possible to share the strengths of both approaches (e.g., Yang, 2005; van Erven et al., 2008)

Model selection: bias and variance

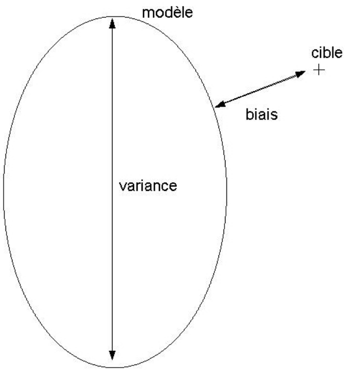
$$\mathbb{E}\left[\ell\left(s^{\star}, \widehat{s}_m(D_n)\right)\right] = \text{Bias} + \text{Variance}$$

Bias or Approximation error

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

Variance or Estimation error

$$\mathbb{E}\left[P\gamma\left(\widehat{s}_m(D_n)\right)\right] - P\gamma\left(s_m^{\star}\right)$$



Model selection: bias and variance

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

Bias or Approximation error

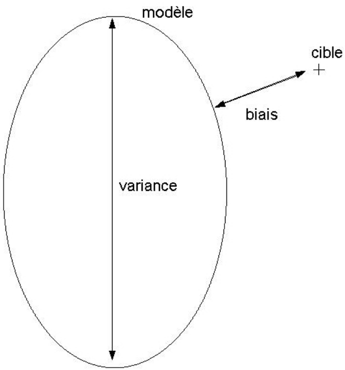
$$\ell\left(s^\star, s_m^\star\right) := \inf_{t \in S_m} \left\{ \ell\left(s^\star, t\right) \right\}$$

Variance or Estimation error

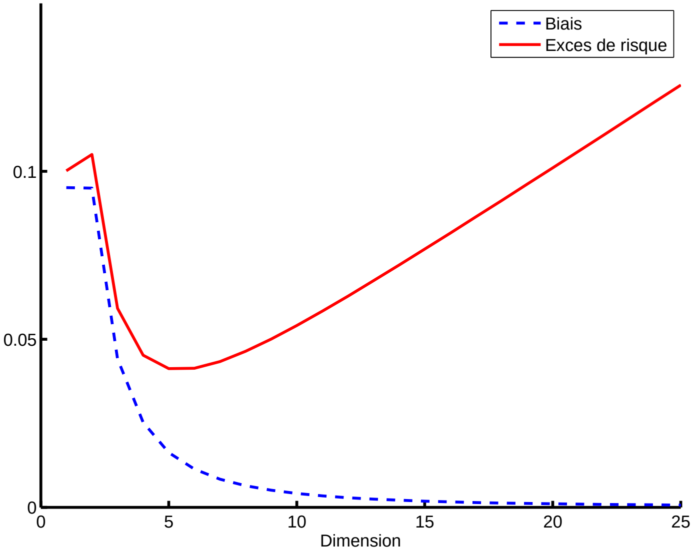
$$\mathbb{E}\left[P\gamma\left(\widehat{s}_m(D_n)\right)\right] - P\gamma\left(s_m^\star\right)$$

Bias-variance trade-off

$\Rightarrow$ avoid over-fitting and under-fitting



Bias-variance trade-off



Example: homoscedastic regression on a fixed design

$$Y = F + \varepsilon \quad \text{with} \quad \mathbb{E}\left[\varepsilon_i^2\right] = \sigma^2$$

$$\widehat{F}_m = A_m Y \quad \text{with} \quad A_m = A_m^\top = A_m^2 \quad \text{and} \quad \operatorname{tr}(A_m) = \dim(S_m)$$

$\Rightarrow$ Bias-variance decomposition of the risk

Example: homoscedastic regression on a fixed design

$$Y = F + \varepsilon \quad \text{with} \quad \mathbb{E}\left[\varepsilon_i^2\right] = \sigma^2$$

$$\widehat{F}_m = A_m Y \quad \text{with} \quad A_m = A_m^\top = A_m^2 \quad \text{and} \quad \operatorname{tr}(A_m) = \dim(S_m)$$

$\Rightarrow$ Bias-variance decomposition of the risk

$$F_m = \arg\min_{t \in S_m} \left\{ \|t - F\|^2 \right\} = A_m F$$

$$\mathbb{E}\left[\frac{1}{n}\left\|\widehat{F}_m - F\right\|^2\right] = \frac{1}{n}\|(A_m - I)F\|^2 + \frac{\sigma^2 \dim(S_m)}{n}$$

$$= \text{Bias} + \text{Variance}$$

Unbiased risk estimation principle

$$\widehat{m} \in \arg\min_{m \in \mathcal{M}_n} \{ \mathrm{crit}(m) \}$$

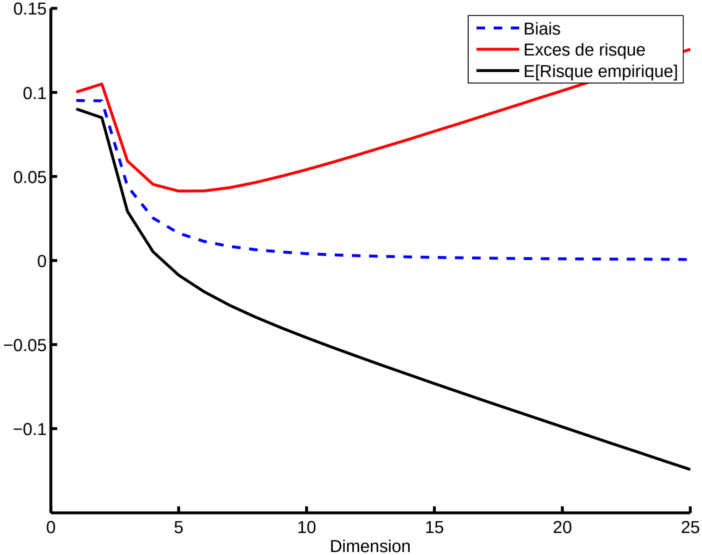
$$\mathrm{crit}_{\mathrm{id}}(m) = \ell\left(s^\star, \widehat{s}_m(D_n) \right)$$

Heuristics:

$$\mathrm{crit}(m) \approx \mathbb{E}\left[\ell\left(s^\star, \widehat{s}_m(D_n) \right) \right]$$

$\Rightarrow$ valid if $\mathrm{Card}(\mathcal{M}_n)$ is not too large
(+ concentration inequalities)

Why should the empirical risk be penalized?



Penalization

- Penalization: $\operatorname{crit}(m) = P_n \gamma\left(\widehat{s}_m\right) + \operatorname{pen}(m)$

$$\widehat{m} \in \arg\min_{m \in \mathcal{M}_n} \left\{ P_n \gamma\left(\widehat{s}_m\right) + \operatorname{pen}(m) \right\}$$

Penalization

- Penalization: $\mathrm{crit}(m) = P_n \gamma\left(\widehat{s}_m\right) + \mathrm{pen}(m)$

$$\widehat{m} \in \arg\min_{m \in \mathcal{M}_n} \left\{ P_n \gamma\left(\widehat{s}_m\right) + \mathrm{pen}(m) \right\}$$

- Ideal penalty:

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

- Mallows' heuristics:
  $\mathrm{pen}(m) \approx \mathbb{E}\left[\mathrm{pen}_{\mathrm{id}}(m)\right] \Rightarrow$ oracle inequality

Example: homoscedastic regression on a fixed design

Recall that

$$Y = F + \varepsilon \quad \text{with} \quad \mathbb{E}\left[\varepsilon_i^2\right] = \sigma^2$$

$$\widehat{F}_m = A_m Y \quad \text{with} \quad A_m = A_m^\top = A_m^2 \quad \text{and} \quad \operatorname{tr}(A_m) = \dim(S_m)$$

$$\mathbb{E}\left[\frac{1}{n}\left\|\widehat{F}_m - F\right\|^2\right] = \frac{1}{n}\|(A_m - I)F\|^2 + \frac{\sigma^2 \dim(S_m)}{n}$$

$\Rightarrow$ Empirical risk? Ideal penalty? Expectations?

Example: homoscedastic regression on a fixed design

Recall that

$$Y = F + \varepsilon \quad \text{with} \quad \mathbb{E}\left[\varepsilon_i^2\right] = \sigma^2$$

$$\widehat{F}_m = A_m Y \quad \text{with} \quad A_m = A_m^\top = A_m^2 \quad \text{and} \quad \operatorname{tr}(A_m) = \dim(S_m)$$

$$\mathbb{E}\left[\frac{1}{n} \left\| \widehat{F}_m - F \right\|^2 \right] = \frac{1}{n} \left\| (A_m - I) F \right\|^2 + \frac{\sigma^2 \dim(S_m)}{n}$$

$\Rightarrow$ Empirical risk? Ideal penalty? Expectations?

$$\mathrm{pen}_{\mathrm{id}}(m) = \frac{2}{n} \langle A_m \varepsilon, \varepsilon \rangle + \frac{2}{n} \langle (A_m - I_n) F, \varepsilon \rangle$$

$$\mathbb{E}\left[\mathrm{pen}_{\mathrm{id}}(m)\right] = \frac{2\sigma^2 D_m}{n} \qquad \Rightarrow \qquad C_p \text{ (Mallows, 1973)}$$

Classical penalties

- C_p (Mallows, 1973; regression, least-squares estimator):

$$2\sigma^2 D_m / n$$

- C_L (Mallows, 1973; regression, linear estimator $\widehat{F}_m = A_m Y$):

$$2\sigma^2 \operatorname{tr}(A_m) / n$$

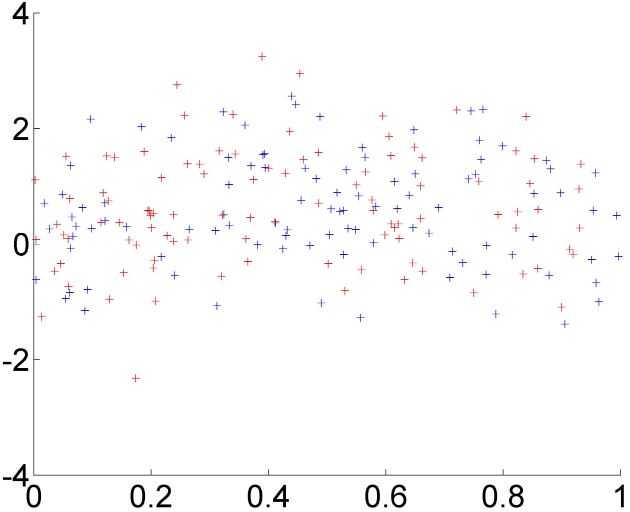
- AIC (Akaike, 1973; log-likelihood, p degrees of freedom):

$$2p/n$$

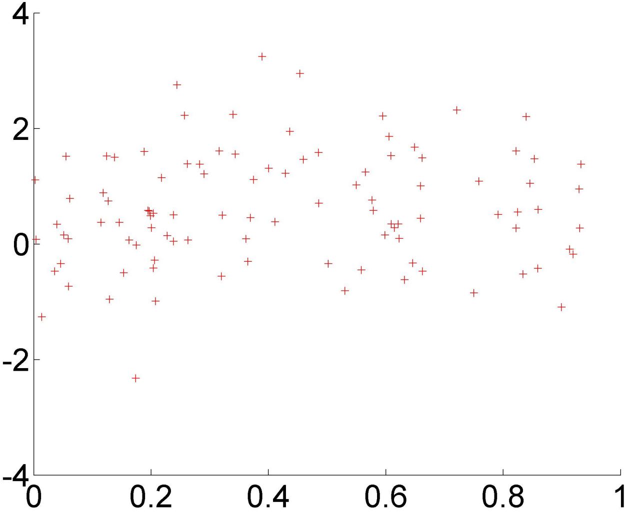
- BIC (Schwarz, 1978; log-likelihood, identification goal):

$$\ln(n) p / n$$

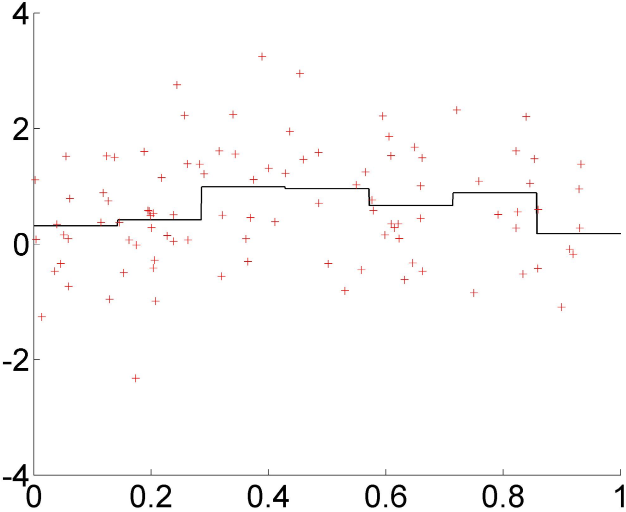
Hold-out



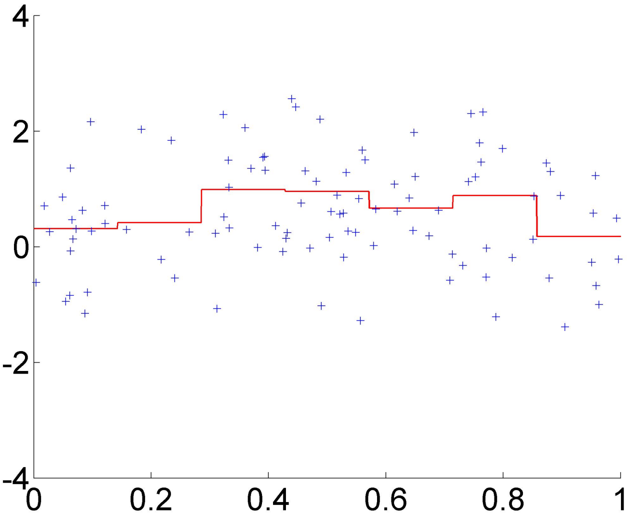
Hold-out: training sample



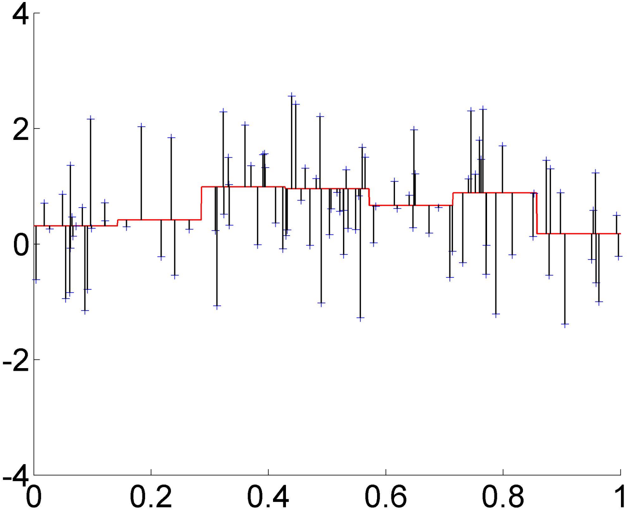
Hold-out: training sample



Hold-out: validation sample



Hold-out: validation sample



Unbiased risk estimation principle

Heuristics:

$$\mathbb{E}[\text{crit}(m)] \approx \mathbb{E}[P\gamma(\widehat{s}_m)] \quad \Leftrightarrow \quad \mathbb{E}[\text{pen}(m)] \approx \mathbb{E}[\text{pen}_{\text{id}}(m)]$$

Examples:

- FPE (Akaike, 1970), SURE (Stein, 1981)
- some kinds of cross-validation (e.g., leave-p-out, $p \ll n$)
- log-likelihood: AIC (Akaike, 1973), AICc (Sugiura, 1978; Hurvich & Tsai, 1989)
- least-squares: C_p, C_L (Mallows, 1973), GCV (Craven & Wahba, 1979)
- covariance penalties (Efron, 2004)
- bootstrap penalty (Efron, 1983), resampling (A., 2009)
- ...

Outline

Probability theory: measure concentration

- Empirical processes:

$$(P_n - P)\gamma(t) \quad \text{or} \quad \sup_{t \in S} \{ (P_n - P)\gamma(t) \}$$

- Concentration of quadratic terms, $\|M\varepsilon\|^2$, χ^2-type statistics (writting them as a sup, or through the general problem of concentration of U-statistics)

- More complex quantities, such as the "ideal penalty"

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

Probability theory

- Exact computation or upper bounds on expectations:

$$\mathbb{E}\left[\sup_{t \in S} \{ (P_n - P)\gamma(t) \} \right]$$

$$\mathbb{E}\left[(P - P_n)\gamma\left(\widehat{s}_m(D_n) \right) \right]$$

- Understanding the risk as a function of n

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

- Resampling process
- Control of remainder terms (variance, deviations, ...) compared to expectations
- ...

Approximation theory

- Bias term $\ell (s^\star, S_m)$
- Necessary to control it for deducing an adaptation result from an oracle inequality
- Conversely, how should we choose $(S_m)_{m \in \mathcal{M}_n}$ knowing that $P \in \mathcal{P}$?
- Control of $\ell (s^\star, S_m)$ (upper and lower bound) useful for controlling $\dim(S_{\widehat{m}})$ and $\dim(S_{m^\star})$

Optimization: for practical reasons

- $\widehat{s}_m(D_n)$ often defined as an arg min

$\Rightarrow$ Computing $\widehat{s}_m(D_n)$ for every m (approximately or not)?

$\Rightarrow$ Direct computation of $\left(\widehat{s}_m(D_n)\right)_{m \in \mathcal{M}_n}$ (regularization path, e.g. LARS-Lasso)?

Optimization: for practical reasons

- $\widehat{s}_m(D_n)$ often defined as an arg min

$\Rightarrow$ Computing $\widehat{s}_m(D_n)$ for every m (approximately or not)?

$\Rightarrow$ Direct computation of $(\widehat{s}_m(D_n))_{m \in \mathcal{M}_n}$ (regularization path, e.g. LARS-Lasso)?

- Computing $\widehat{m} \in \arg\min_{m \in \mathcal{M}_n} \{\mathrm{crit}(m)\}$ without going through all $m \in \mathcal{M}_n$? (e.g., dynamic programming for change-point detection: Bellman & Dreyfus, 1962; Rigaill, 2010)

Optimization: for practical reasons

- $\widehat{s}_m(D_n)$ often defined as an arg min

$\Rightarrow$ Computing $\widehat{s}_m(D_n)$ for every m (approximately or not)?

$\Rightarrow$ Direct computation of $(\widehat{s}_m(D_n))_{m \in \mathcal{M}_n}$ (regularization path, e.g. LARS-Lasso)?

- Computing $\widehat{m} \in \arg\min_{m \in \mathcal{M}_n} \{\mathrm{crit}(m)\}$ without going through all $m \in \mathcal{M}_n$? (e.g., dynamic programming for change-point detection: Bellman & Dreyfus, 1962; Rigaill, 2010)

- The most interesting procedures to study are the ones for which efficient algorithms exist.

Optimization: for theoretical reasons

- $\widehat{s}_m(D_n)$ often defined as an arg min

$\Rightarrow$ KKT conditions can caracterize it

- Ex: ideal penalty for the Lasso (Efron et al. 2004; Zou, Hastie & Tibshirani 2007)

- RKHS and kernel methods: representer theorem
- ...

Outline

Results we are looking for

- guarantees for practical procedures
- theory precise enough for explaining differences observed experimentally
- "non-asymptotic" results
- use theory for designing new procedures, that do not have the drawbacks of existing procedures

Results we are looking for

- guarantees for practical procedures
- theory precise enough for explaining differences observed experimentally
- "non-asymptotic" results
- use theory for designing new procedures, that do not have the drawbacks of existing procedures

http://www.di.ens.fr/~arlot/2011pisa.htm