Model selection via penalization, resampling and cross-validation, with application to change-point detection

Sylvain Arlot

[1] CNRS

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

Cergy, January 30 – February 2, 2012

Outline of the lectures

1. Model selection via penalization, with application to change-point detection
2. Resampling methods for penalization, and robustness to heteroscedasticity in regression
3. Cross-validation for model/estimator selection, with application to detecting changes in the mean of a signal

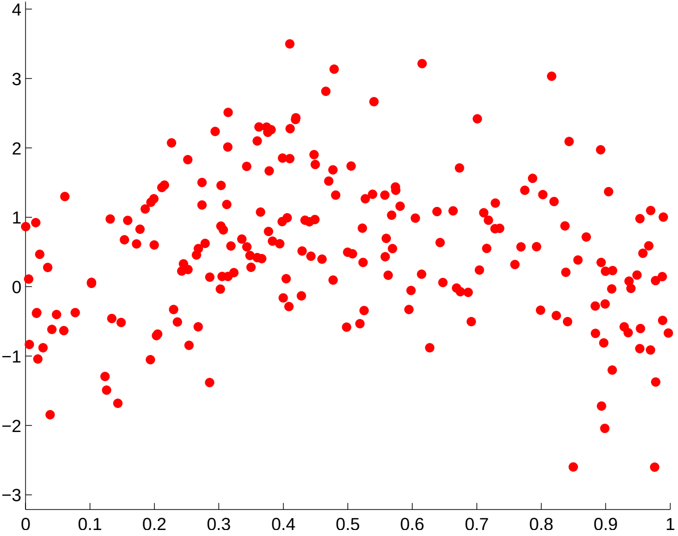
Part I

Model selection via penalization, with application to change-point detection

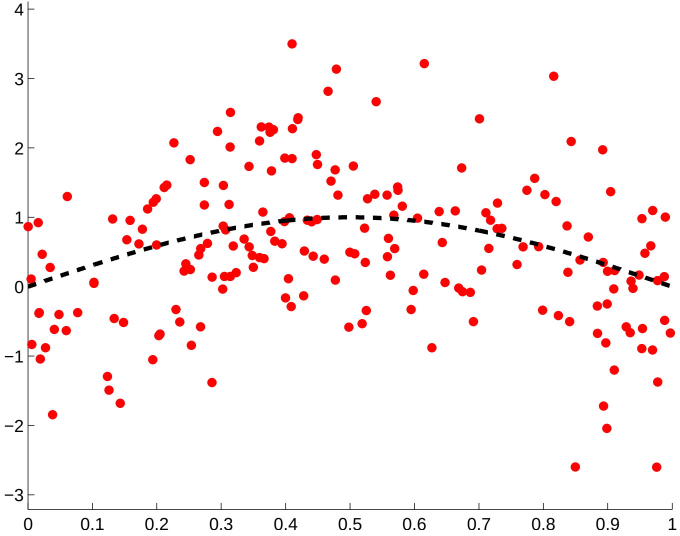
Outline

Outline

Regression: data $(x_1, Y_1), \ldots, (x_n, Y_n)$



Goal: find the signal (denoising)



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

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

  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

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

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

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

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

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

with $\forall x \in \mathbb{R}^n$, $\left\Vert x \right\Vert^2 = \sum_{i=1}^{n} x_i^2$

Example: regression: fixed vs. random design

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

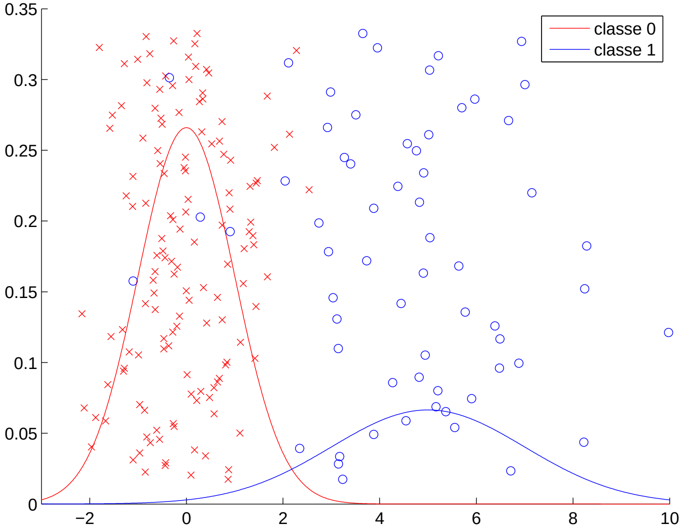
$$\text{with} \quad \forall x \in \mathbb{R}^n \ , \qquad \left\Vert x\right\Vert^2 = \sum_{i=1}^{n} x_i^2$$

Example: regression: fixed vs. random design

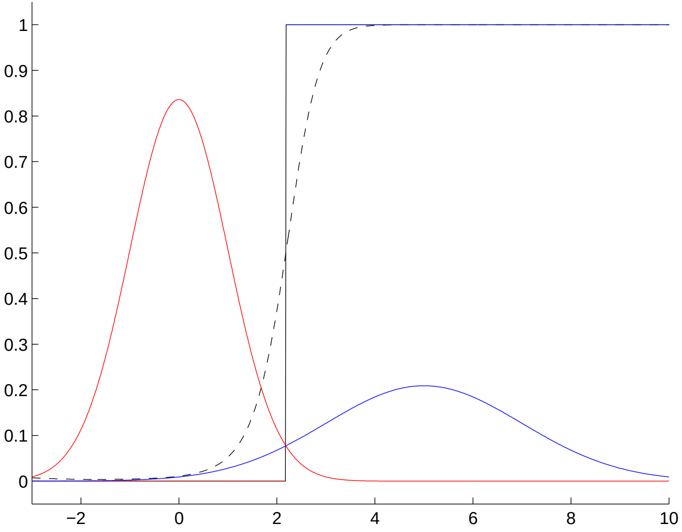
	Random design	Fixed design
D_n	$(X_i, Y_i)_{1 \leq i \leq n}$ i.i.d. $\sim P$	$Y = F + \varepsilon \in \mathbb{R}^n$
	$(X_{n+1}, Y_{n+1}) \sim P$	$X_{n+1} \sim \mathcal{U}(x_1, \ldots, x_n)$
$\mathbb{S}$	$t : \mathcal{X} \to \mathbb{R}$	$t \in \mathbb{R}^n$
$P\gamma(t)$	$\mathbb{E}_{(X,Y)\sim P}\left[(Y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

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

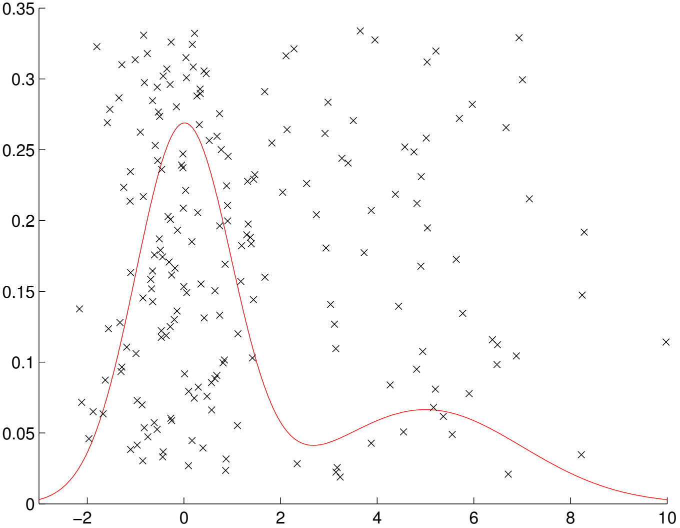
Example: classification (prediction, $\mathcal{X} = \mathbb{R}$, $\mathcal{Y} = \{0, 1\}$)



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

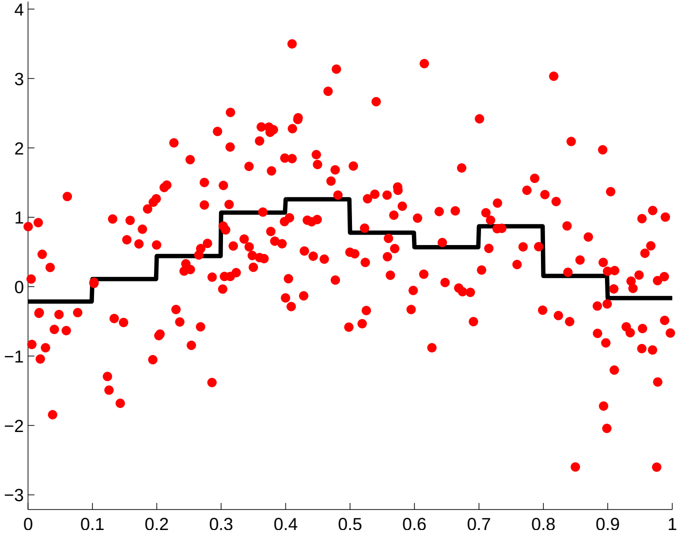


Example: density estimation ($\Xi = \mathbb{R}$): data and target



Outline

Estimators: example: regressogram



Least-squares estimators

- Natural idea: minimize an estimator of the risk $\frac{1}{n} \left\| F - t \right\|^2$

Least-squares estimators

- Natural idea: minimize an estimator of the risk $\frac{1}{n} \|F - t\|^2$
- Least-squares criterion:

$$\frac{1}{n} \|t - Y\|^2 = \frac{1}{n} \sum_{i=1}^{n} (t_i - Y_i)^2$$

$$\forall t \in \mathbb{S} \ , \qquad \mathbb{E}\left[\frac{1}{n} \|t - Y\|^2 \right] = \frac{1}{n} \|F - t\|^2 + \frac{1}{n} \mathbb{E}\left[\|\varepsilon\|^2 \right]$$

Least-squares estimators

- Natural idea: minimize an estimator of the risk $\frac{1}{n} \| F - t \|^2$
- Least-squares criterion:

$$\frac{1}{n} \| t - Y \|^2 = \frac{1}{n} \sum_{i=1}^{n} (t_i - Y_i)^2$$

$$\forall t \in \mathbb{S} \ , \qquad \mathbb{E} \left[\frac{1}{n} \| t - Y \|^2 \right] = \frac{1}{n} \| F - t \|^2 + \frac{1}{n} \mathbb{E} \left[\| \varepsilon \|^2 \right]$$

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

$$\widehat{F}_S \in \arg\min_{t \in S} \left\{ \frac{1}{n} \| t - Y \|^2 \right\} = \arg\min_{t \in S} \left\{ \frac{1}{n} \sum_{i=1}^{n} (t_i - Y_i)^2 \right\}$$

so that

$$\widehat{F}_S = \Pi_S(Y) \qquad \text{(orthogonal projection)}$$

Model examples

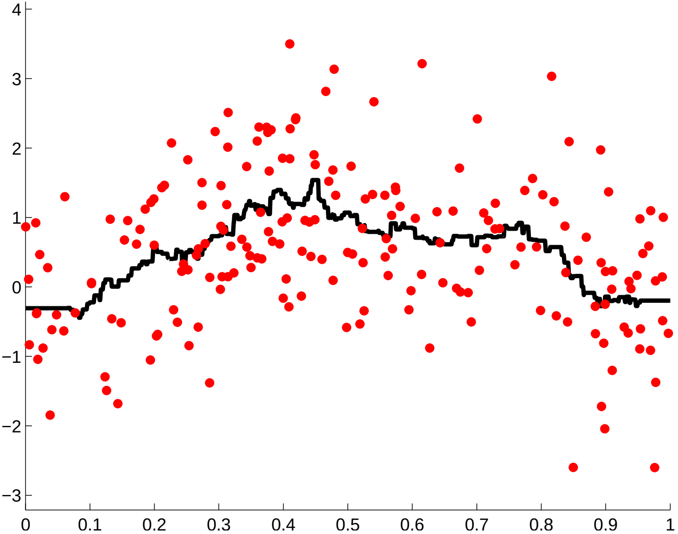
- histograms on some partition Λ of $\mathcal{X}$
  $\Rightarrow$ the least-squares estimator (regressogram) can be written

$$\widehat{F}_{\Lambda}(x_i) = \sum_{\lambda \in \Lambda} \widehat{\beta}_{\lambda} \mathbb{1}_{x_i \in \lambda} \qquad \widehat{\beta}_{\lambda} = \frac{1}{\mathrm{Card}\left\{ x_i \in \lambda \right\}} \sum_{x_i \in \lambda} Y_i$$

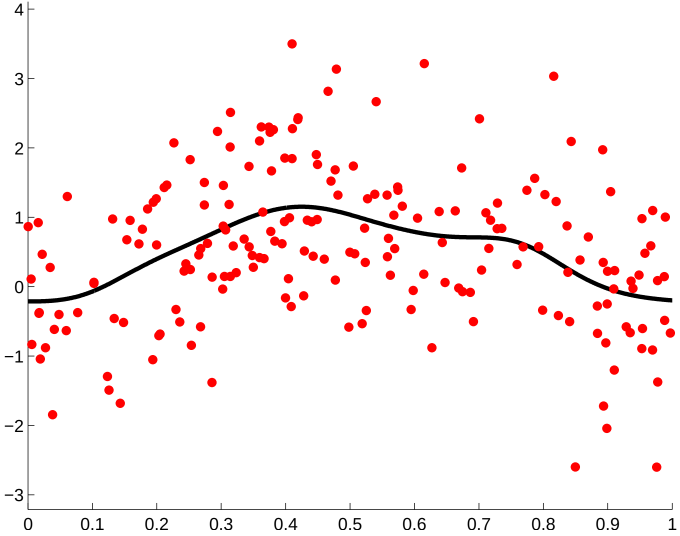
- subspace generated by a subset of an orthogonal basis of $L^2(\mu)$ (Fourier, wavelets, and so on)
- variable selection: $x_i = \left(x_i^{(1)}, \ldots, x_i^{(p)} \right) \in \mathbb{R}^p$ gathers p variables that can (linearly) explain Y_i

$$\forall m \subset \{ 1, \ldots, p \} \quad , \quad S_m = \mathrm{vect} \left\{ x^{(j)} \text{ s.t. } j \in m \right\}$$

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

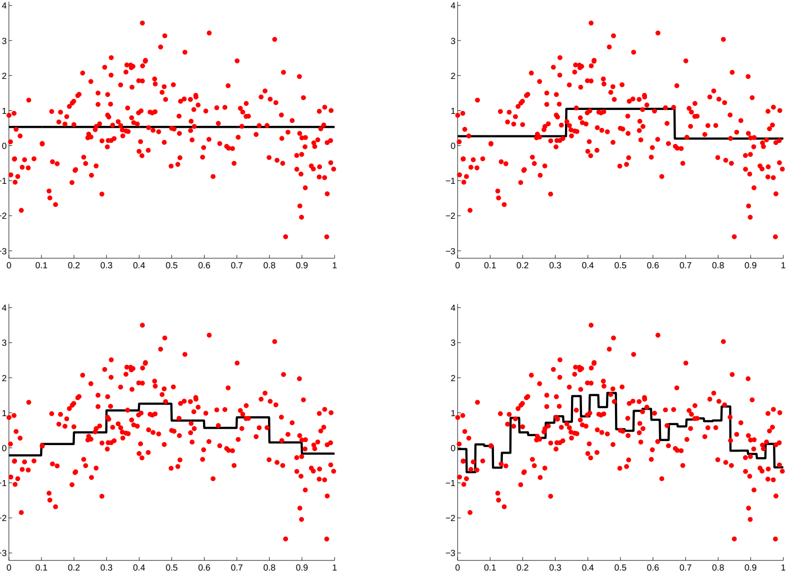


Nadaraya-Watson estimator ($\sigma = 0.01$)



Outline

Model selection: regular regressograms



Model selection problem

- Collection of candidate models: $(S_m)_{m \in \mathcal{M}}$
- Problem: choosing among $(\widehat{F}_m)_{m \in \mathcal{M}}$

$$\text{with} \quad \widehat{F}_m = \widehat{F}_{S_m} = \Pi_{S_m}(Y) = \Pi_m(Y) \enspace .$$

Goal: estimation or prediction

- Main goal: find $\widehat{m}$ minimizing $\frac{1}{n} \left\lVert F - \widehat{F}_{\widehat{m}} \right\rVert^2$
- Oracle: $m^\star \in \arg\min_{m \in \mathcal{M}_n} \left\{ \frac{1}{n} \left\lVert F - \widehat{F}_m \right\rVert^2 \right\}$

Goal: estimation or prediction

- Main goal: find $\widehat{m}$ minimizing $\frac{1}{n} \left\| F - \widehat{F}_{\widehat{m}} \right\|^2$
- Oracle: $m^\star \in \arg\min_{m \in \mathcal{M}_n} \left\{ \frac{1}{n} \left\| F - \widehat{F}_m \right\|^2 \right\}$
- Oracle inequality (in expectation or with high probability):

$$\frac{1}{n} \left\| F - \widehat{F}_{\widehat{m}} \right\|^2 \leq C \inf_{m \in \mathcal{M}_n} \left\{ \frac{1}{n} \left\| F - \widehat{F}_m \right\|^2 \right\} + R_n$$

- 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 $\frac{1}{n} \left\| F - \widehat{F}_{\widehat{m}} \right\|^2$
- Oracle: $m^\star \in \arg\min_{m \in \mathcal{M}_n} \left\{ \frac{1}{n} \left\| F - \widehat{F}_m \right\|^2 \right\}$
- Oracle inequality (in expectation or with high probability):

$$\frac{1}{n} \left\| F - \widehat{F}_{\widehat{m}} \right\|^2 \leq C \inf_{m \in \mathcal{M}_n} \left\{ \frac{1}{n} \left\| F - \widehat{F}_m \right\|^2 \right\} + R_n$$

- 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 F, and so on (if $(S_m)_{m \in \mathcal{M}_n}$ is well chosen)

Goal: identification

- Additional assumption (model selection case): $F \in S_{m_0}$ 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): $F \in S_{m_0}$ 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)

Decomposition of the risk

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

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

$\Rightarrow$ Bias-variance decomposition of the risk

Decomposition of the risk

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

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

$\Rightarrow$ Bias-variance decomposition of the risk

$$F_m := \arg\min_{t \in S_m} \left\{ \frac{1}{n} \|F - t\|^2 \right\} = \Pi_m F$$

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

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

Model selection: bias and variance

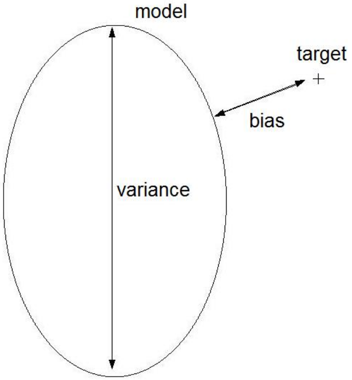
$$\mathbb{E}\left[\frac{1}{n}\left\|F-\widehat{F}_m\right\|^2\right] = \text{Bias} + \text{Variance}$$

Bias or Approximation error

$$\frac{1}{n}\|F-F_m\|^2 := \inf_{t\in S_m}\left\{\frac{1}{n}\|F-t\|^2\right\}$$

Variance or Estimation error

$$\frac{\sigma^2 D_m}{n}$$



Model selection: bias and variance

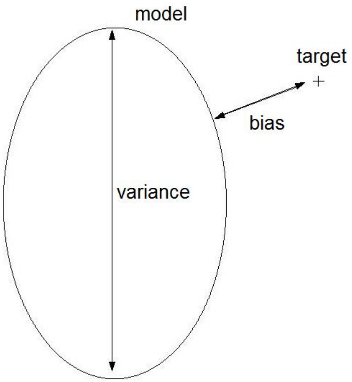
$$\mathbb{E}\left[\frac{1}{n}\left\|F - \widehat{F}_m\right\|^2\right] = \text{Bias} + \text{Variance}$$

Bias or Approximation error

$$\frac{1}{n}\|F - F_m\|^2 := \inf_{t \in S_m}\left\{\frac{1}{n}\|F - t\|^2\right\}$$

Variance or Estimation error

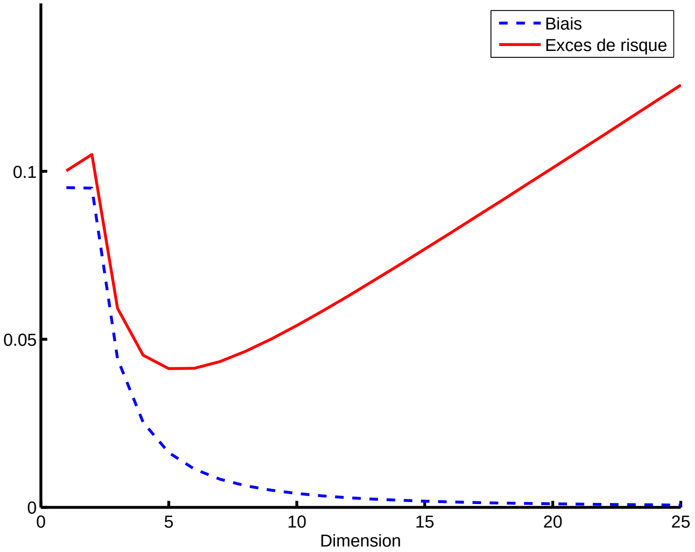
$$\frac{\sigma^2 D_m}{n}$$



Bias-variance trade-off

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

Bias-variance trade-off



Unbiased risk estimation principle

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

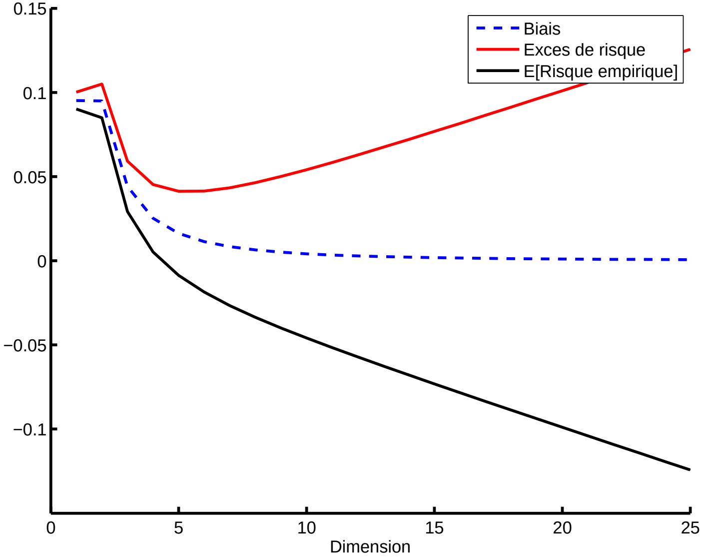
$$\mathrm{crit}_{\mathrm{id}}(m) = \frac{1}{n} \left\| F - \widehat{F}_m \right\|^2$$

Heuristics:

$$\mathrm{crit}(m) \approx \mathbb{E}\left[\frac{1}{n} \left\| F - \widehat{F}_m \right\|^2 \right]$$

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

Why should the empirical risk be penalized?



Penalization

- Penalization: $\operatorname{crit}(m) = \frac{1}{n} \left\| \widehat{F}_m - Y \right\|^2 + \operatorname{pen}(m)$

$$\widehat{m} \in \arg\min_{m \in \mathcal{M}_n} \left\{ \frac{1}{n} \left\| \widehat{F}_m - Y \right\|^2 + \operatorname{pen}(m) \right\}$$

Penalization

- Penalization: $\operatorname{crit}(m) = \frac{1}{n} \left\| \widehat{F}_m - Y \right\|^2 + \operatorname{pen}(m)$

$$\widehat{m} \in \arg \min_{m \in \mathcal{M}_n} \left\{ \frac{1}{n} \left\| \widehat{F}_m - Y \right\|^2 + \operatorname{pen}(m) \right\}$$

- Ideal penalty:

$$\operatorname{pen}_{\mathrm{id}}(m) = \frac{1}{n} \left\| F - \widehat{F}_m \right\|^2 - \frac{1}{n} \left\| \widehat{F}_m - Y \right\|^2$$

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

Computation of the ideal penalty and its expectation

Recall that

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

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

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

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

Computation of the ideal penalty and its expectation

Recall that

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

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

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

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

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

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

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):

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

Unbiased risk estimation principle

Heuristics:

$$\mathbb{E}\left[\mathrm{crit}(m)\right] \approx \mathbb{E}\left[\frac{1}{n}\left\|F - \widehat{F}_m\right\|^2\right] \quad \Leftrightarrow \quad \mathbb{E}\left[\mathrm{pen}(m)\right] \approx \mathbb{E}\left[\mathrm{pen}_{\mathrm{id}}(m)\right]$$

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

A key lemma

Lemma

Let $\mathrm{pen} : \mathcal{M}_n \mapsto \mathbb{R}$ *some penalty (possibly data-dependent).*
On the event Ω *on which for every* $m, m' \in \mathcal{M}_n$,

$$\Big(\mathrm{pen}(m) - \mathrm{pen}_{\mathrm{id}}(m) \Big) - \Big(\mathrm{pen}(m') - \mathrm{pen}_{\mathrm{id}}(m') \Big) \\ \leq A(m) + B(m')$$

$$\textit{we have} \quad \forall \widehat{m} \in \arg\min_{m \in \mathcal{M}_n} \left\{ \frac{1}{n} \left\| \widehat{F}_m - Y \right\|^2 + \mathrm{pen}(m) \right\}$$

$$\frac{1}{n} \left\| \widehat{F}_{\widehat{m}} - F \right\|^2 - B(\widehat{m}) \leq \inf_{m \in \mathcal{M}_n} \left\{ \frac{1}{n} \left\| \widehat{F}_m - F \right\|^2 + A(m) \right\}$$

Oracle inequality for Gaussian regression (1)

Assumptions:

- Fixed design regression, least-squares contrast
- Gaussian homoscedastic noise: $\varepsilon \sim \mathcal{N}(0, \sigma^2)$
- Model collection of polynomial complexity: $\mathrm{Card}(\mathcal{M}_n) \leq C n^{\alpha}$
- For all $m \in \mathcal{M}_n$, $\widehat{F}_m = \Pi_m Y$ (least-squares estimator)
- Penalty

$$\mathrm{pen}(m) = \frac{K \sigma^2 D_m}{n} \qquad \text{with} \quad K > 1$$

Oracle inequality for Gaussian regression (2)

$$-B(m) \leq \mathrm{pen}(m) - \mathrm{pen}_{\mathrm{id}}(m) \leq A(m)$$

$$\Rightarrow \qquad \frac{1}{n} \left\| \widehat{F}_{\widehat{m}} - F \right\|^2 - B(\widehat{m}) \leq \inf_{m \in \mathcal{M}_n} \left\{ \frac{1}{n} \left\| \widehat{F}_m - F \right\|^2 + A(m) \right\}$$

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

First term has expectation $\frac{2\sigma^2 D_m}{n}$, the second term is centered.

Oracle inequality for Gaussian regression (3)

Two Gaussian concentration results (see Massart 2007):

Proposition

Let ξ be some standard Gaussian vector in $\mathbb{R}^n$, $\alpha \in \mathbb{R}^n$, $M \in \mathcal{M}_n(\mathbb{R})$. Then, for every $x \geq 0$,

$$\mathbb{P}\left(\left| \langle \xi, \alpha \rangle \right| \leq \sqrt{2x} \left\| \alpha \right\|_2 \right) \geq 1 - 2e^{-x}$$

$$\mathbb{P}\left(\left| \langle \xi, M\xi \rangle - \operatorname{tr}(M) \right| \leq 2\sqrt{x \operatorname{tr}\left(M^\top M \right)} + 2 \left\| M \right\| x \right) \geq 1 - 2e^{-x}$$

Oracle inequality for Gaussian regression (4)

Sketch of the proof:

- For all $m \in \mathcal{M}_n$,
  concentrate $\langle \Pi_m \varepsilon, \varepsilon \rangle$ around $\sigma^2 D_m$
  and $\langle (\Pi_m - I_n) F, \varepsilon \rangle$ around 0
- Apply the Lemma on the intersection of these $\mathrm{Card}(\mathcal{M}_n)$ events
- Control the remainder terms

Oracle inequality for Gaussian regression (5)

Theorem (Birgé & Massart, 2001–2007)

For every $x \geq 0$, *with probability at least* $1 - 4\,\mathrm{Card}(\mathcal{M}_n) e^{-x}$, *for every*

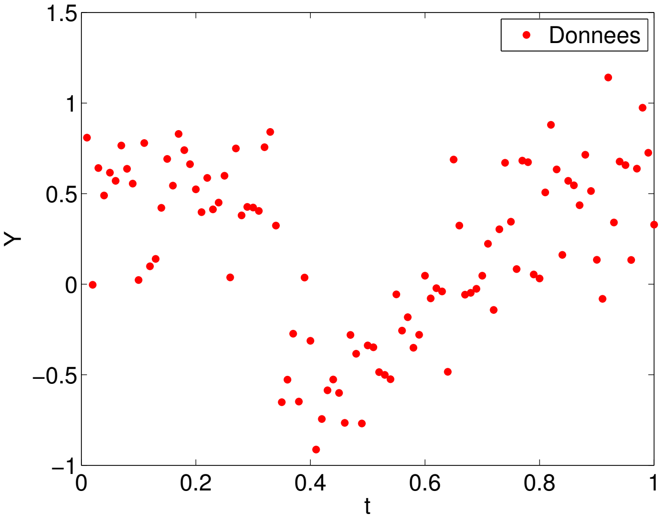
$$\widehat{m} \in \arg\min_{m \in \mathcal{M}_n} \left\{ \frac{1}{n} \left\| Y - \widehat{F}_m \right\|^2 + \frac{K \sigma^2 D_m}{n} \right\} \quad ,$$

we get the oracle inequality $\forall \delta > 0$,

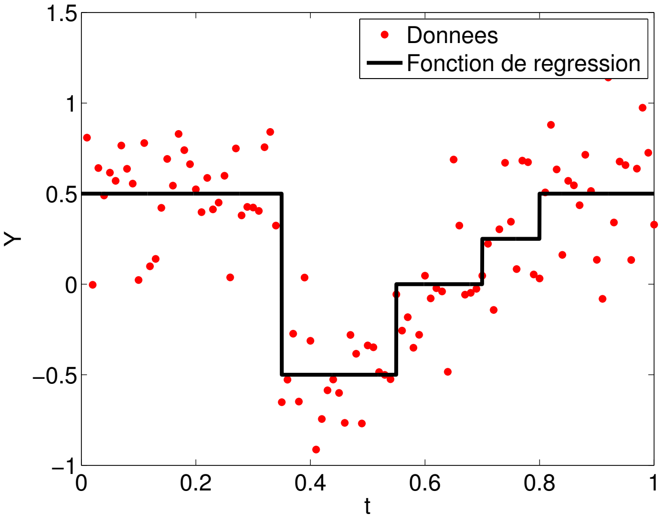
$$\frac{1}{n} \left\| \widehat{F}_{\widehat{m}} - F \right\|^2 \leq \left(\frac{1 + (K-2)_+}{1 - (2-K)_+} + \delta \right) \inf_{m \in \mathcal{M}_n} \left\{ \frac{1}{n} \left\| \widehat{F}_m - F \right\|^2 \right\} + \frac{C(K) x \sigma^2}{\delta n}$$

Outline

Change-point detection: data



Change-point detection: target function



Change-point detection and model selection

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

- 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 into intervals)

$$\text{with} \quad F_i = \eta(t_i)$$

The previous oracle inequality is not sufficient

$$\textit{Problem}: \quad \mathrm{Card}(\mathcal{M}_n) = 2^{n-1}$$

Theorem (Birgé & Massart, 2001–2007)

For every $x \geq 0$, *with probability at least* $1 - 4\,\mathrm{Card}(\mathcal{M}_n) e^{-x}$, *for every*

$$\widehat{m} \in \arg\min_{m \in \mathcal{M}_n} \left\{ \frac{1}{n} \left\| Y - \widehat{F}_m \right\|^2 + \frac{K\sigma^2 D_m}{n} \right\} ,$$

we get the oracle inequality $\forall \delta > 0$,

$$\frac{1}{n} \left\| \widehat{F}_{\widehat{m}} - F \right\|^2 \leq \left(\frac{1 + (K-2)_+}{1 - (2-K)_+} + \delta \right) \inf_{m \in \mathcal{M}_n} \left\{ \frac{1}{n} \left\| \widehat{F}_m - F \right\|^2 \right\} + \frac{C(K) x \sigma^2}{\delta n}$$

A general oracle inequality

Theorem (Birgé & Massart, 2001)

Let $K > 1$ *and* $(L_m)_{m \in \mathcal{M}_n}$ *be nonnegative weights such that* $\sum_{m \in \mathcal{M}_n} e^{-L_m D_m} = \Sigma < +\infty$. *For every*

$$\widehat{m} \in \arg\min_{m \in \mathcal{M}_n} \left\{ \frac{1}{n} \left\| Y - \widehat{F}_m \right\|^2 + \frac{K \sigma^2 D_m}{n} \left(1 + \sqrt{2 L_m} \right)^2 \right\} \quad ,$$

we get the oracle inequality

$$\mathbb{E}\left[\frac{1}{n} \left\| F - \widehat{F}_{\widehat{m}} \right\|^2 \right] \leq C(K) \inf_{m \in \mathcal{M}_n} \left\{ \frac{1}{n} \left\| F - F_m \right\|^2 + \mathrm{pen}(m) \right\} + \frac{C'(K) \Sigma \sigma^2}{n}$$

Weights for change-point detection

If $L_m = L(D_m)$,

Weights for change-point detection

If $L_m = L(D_m)$,

$$\sum_{m \in \mathcal{M}_n} e^{-L_m D_m} = \sum_{D \geq 1} \operatorname{Card} \{ m \in \mathcal{M}_n \text{ s.t. } D_m = D \} e^{-DL(D)}$$
$$= \sum_{D \geq 1} \exp \left[-DL(D) + \ln \operatorname{Card} \{ m \in \mathcal{M}_n \text{ s.t. } D_m = D \} \right]$$

is finite by taking (for instance)

$$L(D) = \ln \left(\operatorname{Card} \{ m \in \mathcal{M}_n \text{ s.t. } D_m = D \} \right) + \alpha D \quad \text{with} \quad \alpha > 0 \ ,$$

Weights for change-point detection

If $L_m = L(D_m)$,

$$\sum_{m \in \mathcal{M}_n} e^{-L_m D_m} = \sum_{D \geq 1} \operatorname{Card}\{ m \in \mathcal{M}_n \text{ s.t. } D_m = D \} e^{-DL(D)}$$

$$= \sum_{D \geq 1} \exp\left[-DL(D) + \ln \operatorname{Card}\{ m \in \mathcal{M}_n \text{ s.t. } D_m = D \} \right]$$

is finite by taking (for instance)

$$L(D) = \ln\left(\operatorname{Card}\{ m \in \mathcal{M}_n \text{ s.t. } D_m = D \} \right) + \alpha D \quad \text{with} \quad \alpha > 0 \ ,$$

and for change-point detection

$$\ln\left(\operatorname{Card}\{ m \in \mathcal{M}_n \text{ s.t. } D_m = D \} \right) = \ln \binom{n-1}{D-1} \leq D \ln\left(\frac{en}{D} \right) \quad .$$

Resulting penalty

- Birgé-Massart theory + simulation experiments for optimizing the constants (Lebarbier, 2005):

$$\widehat{m} \in \operatorname{argmin}_{m \in \mathcal{M}_n} \left\{ \frac{1}{n} \left\| \widehat{F}_m - Y \right\|^2 + \frac{C \sigma^2 D_m}{n} \left(5 + 2 \ln \left(\frac{n}{D_m} \right) \right) \right\}$$

- Equivalent to aggregating models of the same dimension:

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

$$\widehat{F}_D \in \operatorname{argmin}_{t \in \widetilde{S}_D} \left\{ \frac{1}{n} \left\| t - Y \right\|^2 \right\}$$

$$\widehat{D} \in \operatorname{argmin}_{1 \leq D \leq n} \left\{ \frac{1}{n} \left\| \widehat{F}_D - Y \right\|^2 + \frac{C \sigma^2 D}{n} \left(5 + 2 \ln \left(\frac{n}{D} \right) \right) \right\}$$

Computational complexity

- Dynamic programming algorithm (Bellman & Dreyfus, 1962)
- Key remark: $\widehat{F}_D = \widehat{F}_{\widehat{m}(D)}$ with

$$\widehat{m}(D) \in \operatorname*{argmin}_{m \in \mathcal{M}_n \, , \, D_m = D} \left\{ \sum_{\lambda \in m} f(\lambda) \right\} \quad \text{with} \quad f(\lambda) = \operatorname{var}\left(\left(Y_i \right)_{i \in \lambda} \right)$$

Computational complexity

- Dynamic programming algorithm (Bellman & Dreyfus, 1962)
- Key remark: $\widehat{F}_D = \widehat{F}_{\widehat{m}(D)}$ with

$$\widehat{m}(D) \in \operatorname*{argmin}_{m \in \mathcal{M}_n \, , \, D_m = D} \left\{ \sum_{\lambda \in m} f(\lambda) \right\} \quad \text{with} \quad f(\lambda) = \operatorname{var}\left(\left(Y_i \right)_{i \in \lambda} \right)$$

- Algorithm:
  1. Compute $f(\lambda)$ for all possible λ ($n(n+1)/2$ possible segments):

Computational complexity

- Dynamic programming algorithm (Bellman & Dreyfus, 1962)
- Key remark: $\widehat{F}_D = \widehat{F}_{\widehat{m}(D)}$ with

$$\widehat{m}(D) \in \operatorname*{argmin}_{m \in \mathcal{M}_n \, , \, D_m = D} \left\{ \sum_{\lambda \in m} f(\lambda) \right\} \quad \text{with} \quad f(\lambda) = \operatorname{var}\left(\left(Y_i \right)_{i \in \lambda} \right)$$

- Algorithm:
  1. Compute $f(\lambda)$ for all possible λ ($n(n+1)/2$ possible segments):
  2. For $1 \leq i \leq k \leq n$, let $\widehat{m}_k(i)$ be a minimizer of the empirical risk over segmentations of $\{1, \ldots, k\}$ into i segments, and $R_k(i)$ the corresponding empirical risk.
     Then, $\{\widehat{m}_k(1), \ldots, \widehat{m}_k(k)\}$ and $\{R_k(1), \ldots, R_k(k)\}$ can be computed sequentially from $k = 1$ to $k = n$.

Computational complexity

- Dynamic programming algorithm (Bellman & Dreyfus, 1962)
- Key remark: $\widehat{F}_D = \widehat{F}_{\widehat{m}(D)}$ with

$$\widehat{m}(D) \in \operatorname*{argmin}_{m \in \mathcal{M}_n \, , \, D_m = D} \left\{ \sum_{\lambda \in m} f(\lambda) \right\} \quad \text{with} \quad f(\lambda) = \operatorname{var}\left((Y_i)_{i \in \lambda} \right)$$

- Algorithm:
  1. Compute $f(\lambda)$ for all possible λ ($n(n+1)/2$ possible segments):
  2. For $1 \le i \le k \le n$, let $\widehat{m}_k(i)$ be a minimizer of the empirical risk over segmentations of $\{ 1, \ldots, k \}$ into i segments, and $R_k(i)$ the corresponding empirical risk.
     Then, $\{ \widehat{m}_k(1), \ldots, \widehat{m}_k(k) \}$ and $\{ R_k(1), \ldots, R_k(k) \}$ can be computed sequentially from $k = 1$ to $k = n$.

$\Rightarrow$ complexity $\mathcal{O}(n^2)$

- Remark: can be done faster with pruning (Rigaill, 2011)

Illustration: empirical risk minimizer with $D = 4$

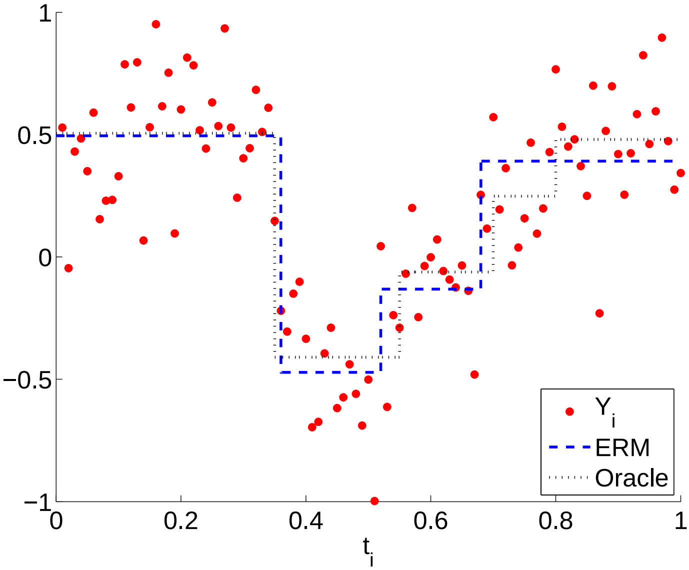
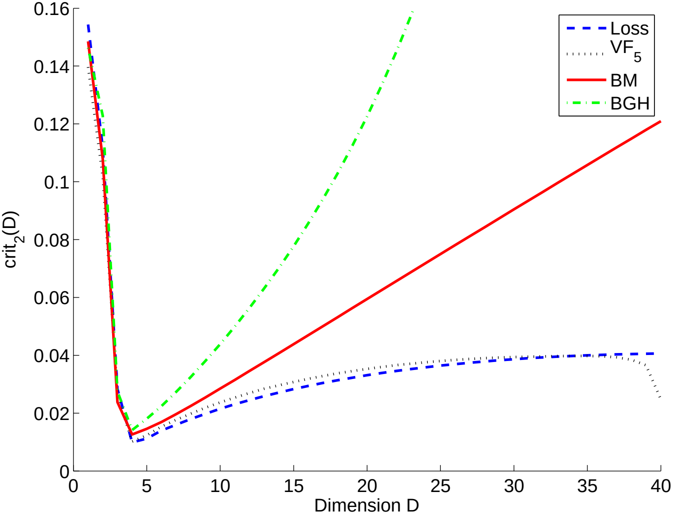


Illustration: estimate of the loss as a function of D



(Some) other model selection approaches

- penalization:
  - Baraud, Giraud & Huet 2009: multiplicative penalty, Gaussian noise
  - Zhang & Siegmund, 2007: modified BIC
  - see also Lavielle, 2005
- cross-validation (third lecture; A. & Celisse, 2010)
- Picard *et al.*, 2005: penalized maximum likelihood, looks for change-points of (η, σ), assuming a Gaussian model

Outline

Conclusion

- bias-variance trade-off for model selection (overfitting vs. underfitting)
- model selection via penalization: $\mathbb{E}\left[\mathrm{pen}_{\mathrm{id}}(m)\right]$ leads to an oracle inequality for polynomial collection of models
- possible extension to exponential collections, with larger penalties
  example: change-point detection
- related problem: data-driven calibration of constants in the penalty (σ^2): slope heuristics

http://www.di.ens.fr/~arlot/2012cergy.htm