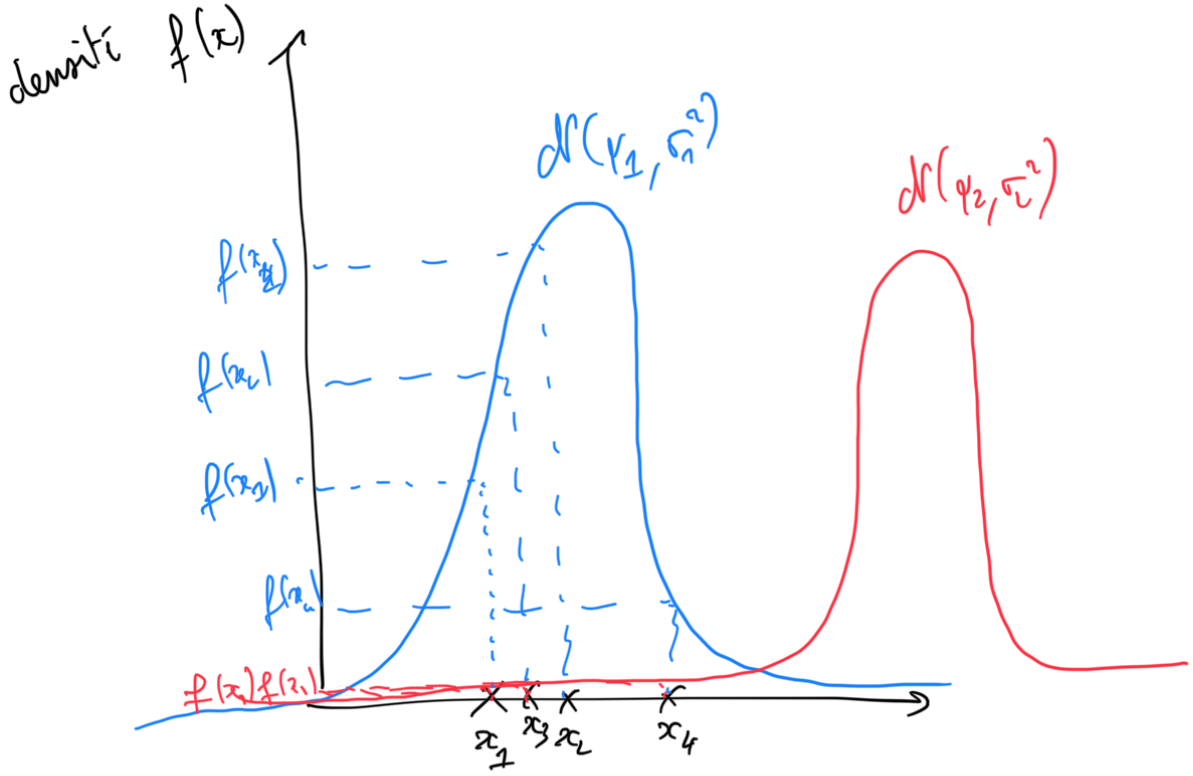
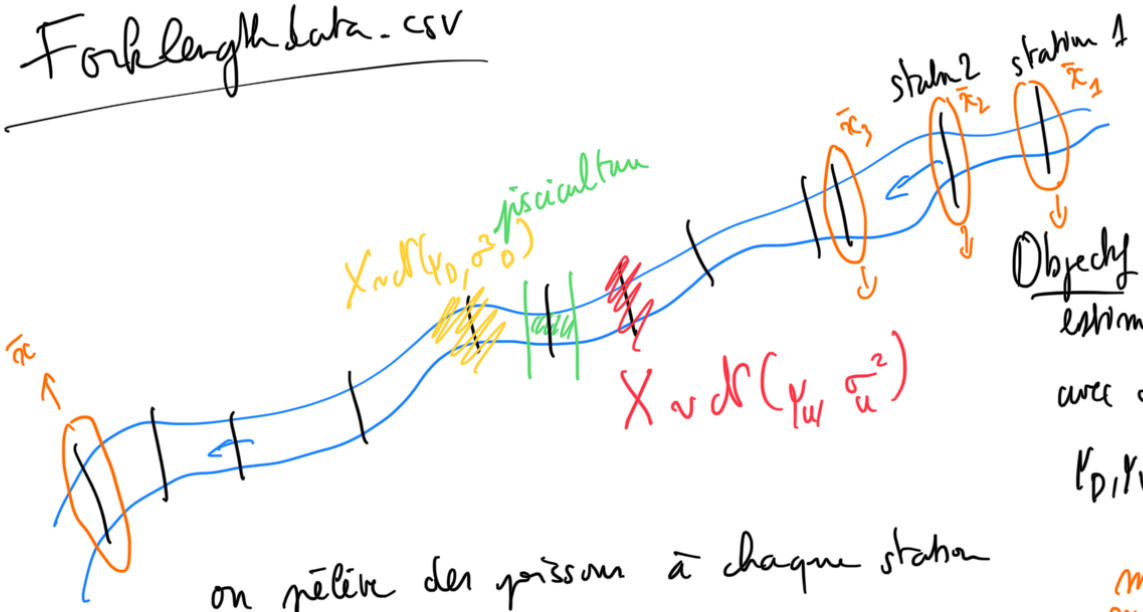
Vraisemblance



$f(x_1) \times \dots \times f(x_n) > f(x_1) \times \dots \times f(x_n)$

vraisemblance

Forklength data.csv



on prélève des poissons à chaque station
on mesure leur longueur

data - n° de station
- longueur des poissons

estimer μ_D et μ_M
avec un prior
$\mu_D, \mu_M \sim \mathcal{N}(\quad, \quad)$

moyenne empirique des poissons de la station
variance de $(\bar{x}_1, \bar{x}_4, \dots)$
other.

- longitude
- info sur la station : location
  - "other"
  - by the farm
  - downstream
  - upstream

En appliquant la formule pour la loi a posteriori, on obtient

$$\begin{cases} X_D \sim \mathcal{N}(\gamma_D, \sigma_D^2) \\ X_U \sim \mathcal{N}(\gamma_U, \sigma_U^2) \\ \gamma_D, \gamma_U \sim \mathcal{N}(m, s^2) \end{cases}$$

avec $m \simeq 97$, $s^2 \simeq 70$ } issu des données "other"

taille de données D

$$\gamma_D \mid \underline{X} \sim \mathcal{N}\left(\frac{s^2 \times \bar{x}_D + m \times \frac{\sigma_D^2}{n}}{s^2 + \frac{\sigma_D^2}{n}} , \frac{s^2 \frac{\sigma_D^2}{n}}{s^2 + \frac{\sigma_D^2}{n}} \right)$$

Calcul de la loi a posteriori

Th de Bayes

$$p(\theta \mid \underline{X}) = \frac{p(\underline{X} \mid \theta)\, p(\theta)}{p(\underline{X})}$$

$p(\underline{X} \mid \theta)\, p(\theta)$: $\mathcal{N}$ $\mathcal{N}$

$p(\underline{X}) = ?$ = proba des données "quelque soit" la valeur de θ

$$\begin{cases} p(\underline{x}|\theta) \\ p(\underline{x},\theta) = p(\underline{x}|\theta)\,p(\theta) \end{cases}$$

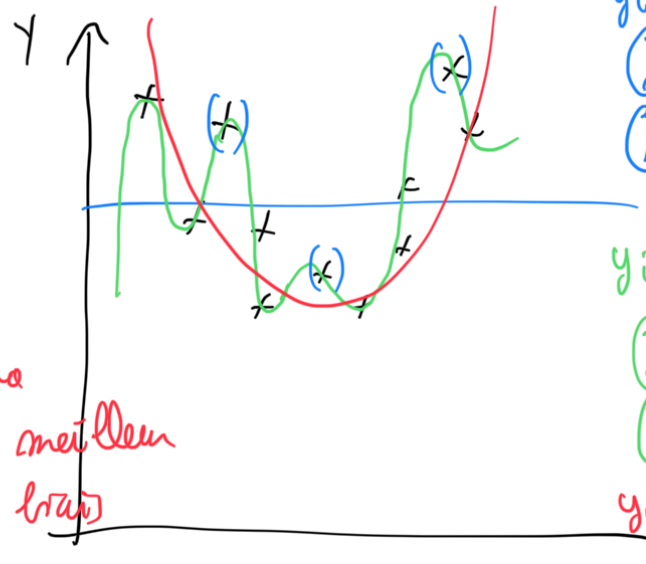
$$p(\underline{x}) = \int_{(F)} p(\underline{x},\theta)\, d\theta$$

$$= \int_{(H)} p(\underline{x}|\theta)\, p(\theta)\, d\theta$$

$\mathcal{N}$ $\mathcal{N}$ $\mathcal{N}$

Comment comparer 2 modèles en statistique ?

ex : intégration



On recherche toujours le meilleur compromis variance biais

$y_i = a + b x_i + \varepsilon_i$

- grand biais (ε_i trop grand)
- faible variance : peu sensible aux fluctuations d'échantillon

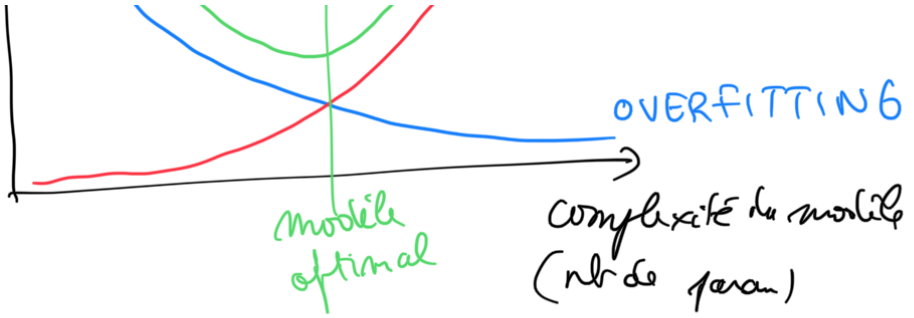
$y_i = a + \sum_{j=1}^{m} b_j x_i^j + \varepsilon_i$

OVERFITTING

- biais faible
- forte variance

$y_i = a + b_1 x_i + b_2 x_i^2 + \varepsilon_i$

- faible variance
- faible biais



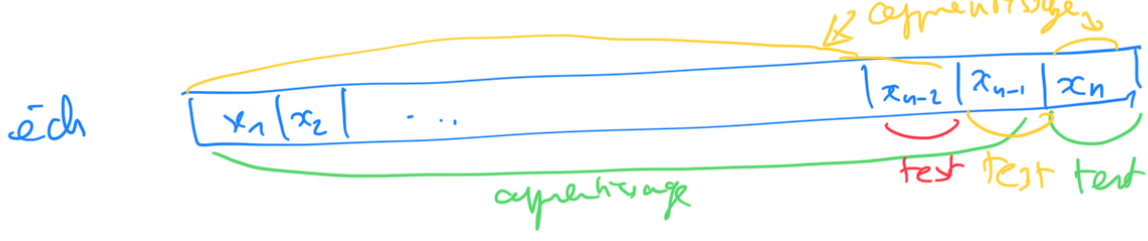
Quel critère

- vraisemblance : NON choisira toujours le modèle le + complexe
- somme des carrés résiduels sur l'éch d'apprentissage : NON
- test du rapport de vraisemblance maximale : pour 2 modèles emboîtés ex : $Y = \alpha X_1 + \beta X_2$
- vraisemblance pénalisée : AIC OUI
- somme des carrés résiduels sur un éch test : OUI

(testé les modèles proba / universel)

↳ on découpe notre base de données en une partie "apprentissage" ($\frac{2}{3}$ des données) et une partie "test"
on estime le modèle sur la partie apprentissage et on l'évalue sur la partie test.

→ quand on a peu de données, l'alternative est la validation croisée :



on boucle sur toutes les données de sorte que chacune soit servi une et une seule fois de test

VALIDATION CROISÉE LEAVE-ONE-OUT

(validation croisée ≠ fold, ...)

Modèle de régression linéaire

$$\begin{cases} y_i = \alpha + \sum_{j=1}^{p} \beta_j x_{ij} + \varepsilon_i \\ \text{avec } \varepsilon_i \sim \mathcal{N}(0, \sigma^2) \end{cases}$$

$$\Rightarrow \boxed{y_i | x_i \sim \mathcal{N}(\alpha + \textstyle\sum \beta_j x_{ij}, \sigma^2)}$$

$$\begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix} \sim \mathcal{N}\left(\begin{pmatrix} \alpha + \sum_j \beta_j x_{1j} \\ \vdots \\ \alpha + \sum_j \beta_j x_{nj} \end{pmatrix}, \begin{pmatrix} \sigma^2 & & 0 \\ & \ddots & \\ 0 & & \sigma^2 \end{pmatrix} \right)$$

$$Y \sim \mathcal{N}\left(\alpha \underbrace{\begin{pmatrix} 1 \\ \vdots \\ 1 \end{pmatrix}}_{1_n} + \underbrace{\begin{pmatrix} x_{11} & \cdots & x_{1p} \\ \vdots & & \vdots \\ x_{n1} & \cdots & x_{np} \end{pmatrix}}_{X} \underbrace{\begin{pmatrix} \beta_1 \\ \vdots \\ \beta_p \end{pmatrix}}_{\beta}, \sigma^2 \underbrace{\begin{pmatrix} 1 & & 0 \\ & \ddots & \\ 0 & & 1 \end{pmatrix}}_{I_n} \right)$$

$$\boxed{Y \sim \mathcal{N}(\alpha 1_n + X\beta, \sigma^2 I_n)}$$

Th Si X et Y sont gaussiens, la meilleure* fonction $f(x)$ qui permet d'expliquer Y est

$f(x) = aX+b$

* au sens $E[(Y - f(x))^2]$ soit minimale

Regression linéaire en grande dimension

- corrélation dans les X $\Rightarrow$ $V(\hat{\beta})$ très grande $\Rightarrow$ non significativité
- $n \approx p$ (ou $n < p$) $\Rightarrow$ "faire" corrélation qui cependant

Solutions

(1) algorithme de sélection de variables de type backward / forward / stepwise (step en R)
marche bien pour quelques dizaines de variables ($\leq 20, 30$...)

(2) construire de nouvelles "méta" variables qui résume l'information contenu dans X, tout en étant non corrélées entre elles
→ régression en composantes principales
→ régression PLS : Partial least Square

efficace en prédiction, mais on perd en interprétabilité du modèle (pcr / pls en R)

(3) méthodes de régression pénalisée

$Y = X\beta + \varepsilon$

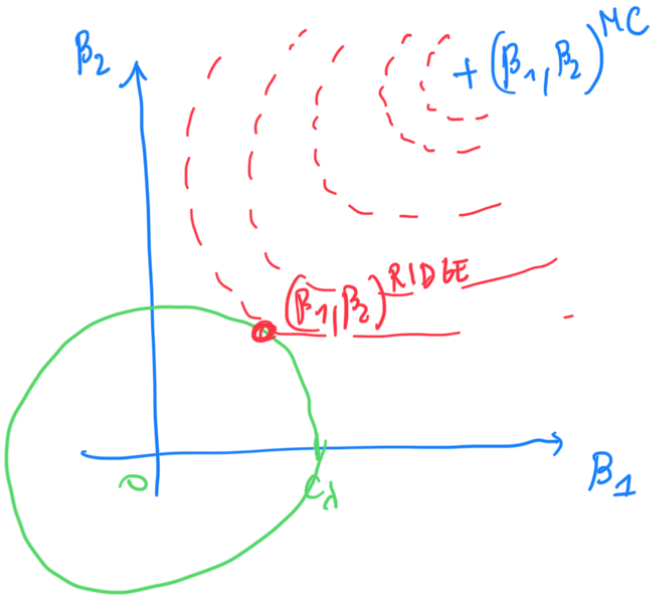
$\hat{\beta}$ maximum de vraisemblance $\Leftrightarrow$ moindres carrés

$\hat{\beta}$ minimisent $\sum_i \left(y_i - \left(\beta_0 + \sum_j \beta_j x_{ij} \right) \right)^2$

RIDGE on cherche $\hat{\beta}$ qui minimisent $\sum (y_i - (\dots))^2$ sous la contrainte $\sum_j \beta_j^2 \leq c_\lambda$

$$\Leftrightarrow \sum_i \left(y_i - \left(\beta_0 + \sum \beta_j x_{ij} \right) \right)^2 + \lambda \sum_j \beta_j^2$$

$p=2$



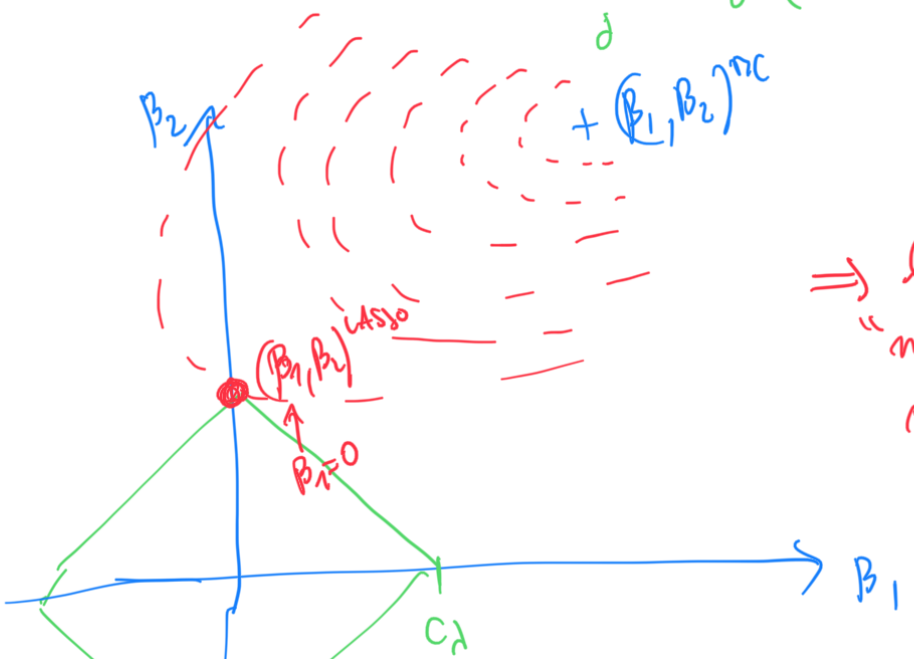
$\Leftrightarrow$ modèle bayésien avec $\beta \sim \mathcal{N}(0, \frac{\sigma^2}{\lambda} I)$

$$\hat{\beta} = (X^t X + \lambda I)^{-1} X^t y$$

LASSO on cherche $\hat{\beta}$ minimisant

$$\sum_i \left(y_i - \left(\beta_0 + \sum_j \beta_j x_{ij} \right) \right)^2 \text{ sous la contrainte}$$

$$\sum_j |\beta_j| < c_\lambda$$



$\Rightarrow$ les β_j des variables "non significatives" sont mis à 0.

meilleur alternative pour la régression en dimension

Rq c'est une régression bayesienne avec

$\beta \sim$ Laplace (..)