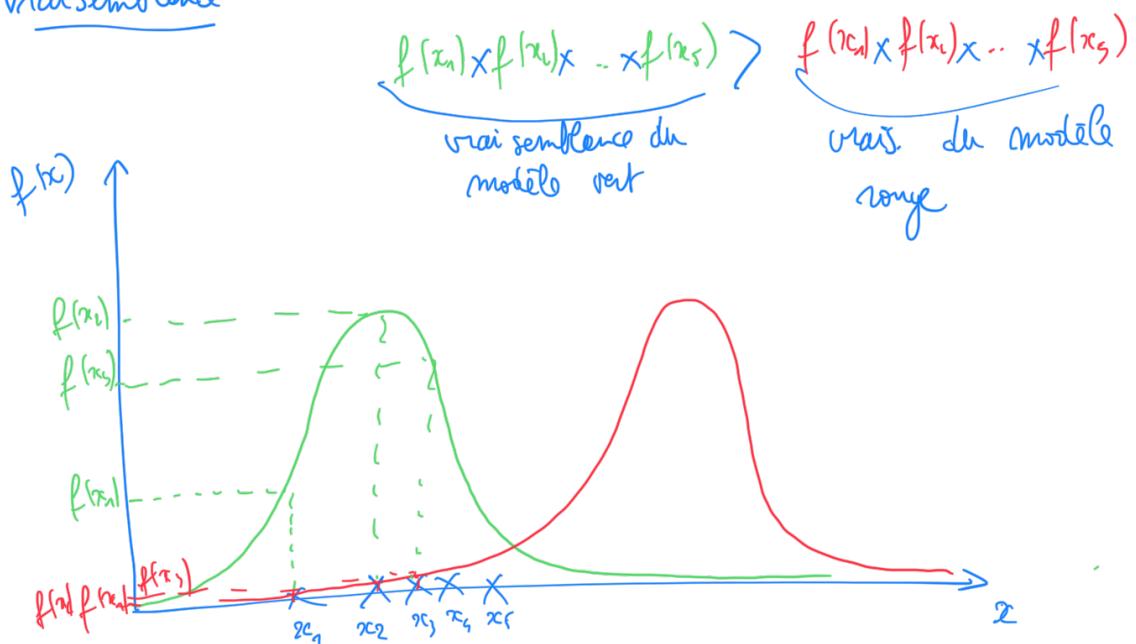
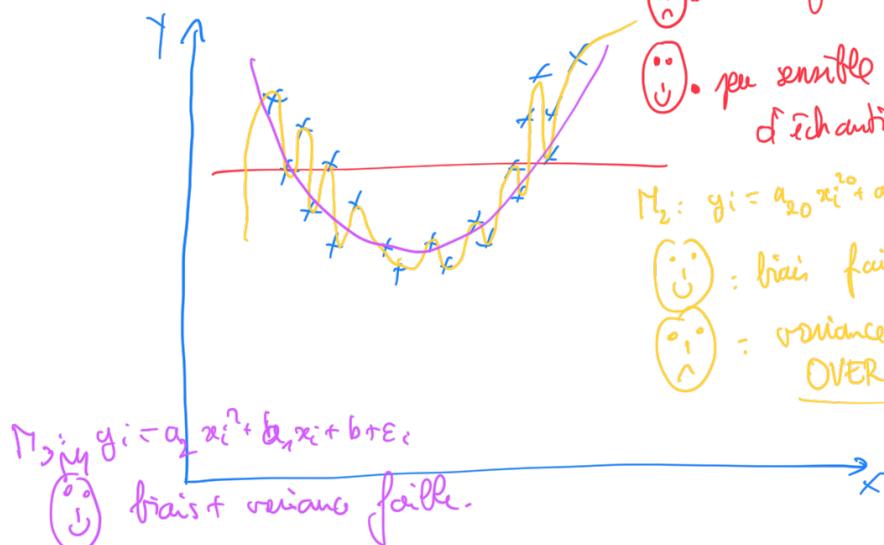


Vraisemblance



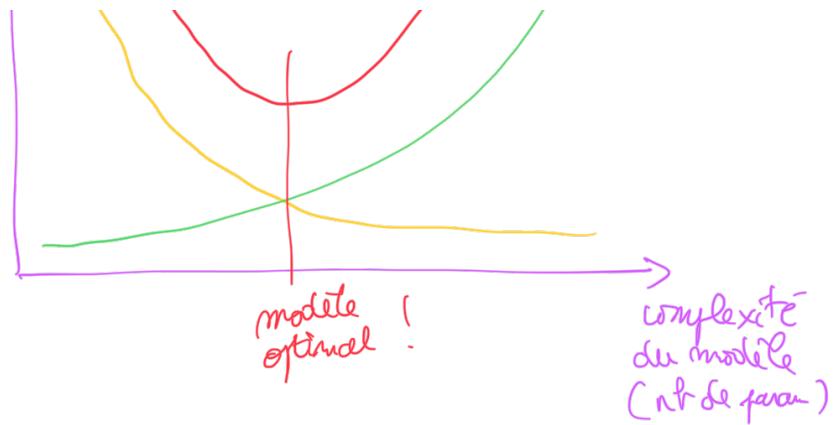
Régression (y, x)

$$(x_i, y_i)_{i=1 \dots n}$$



On cherchera toujours un modèle avec le meilleur compromis bias variance





En pratique, pour choisir le modèle optimal,

on regarde :

- R^2 ajusté
 - (ΔR^2 classique croît avec la complexité du modèle)
- vraisemblance généralisée : AIC, BIC, ...
 - (Δ vraisemblance croît avec la complexité...)
- erreur quadratique RSS = $\sum (y_i - \hat{y}_i)^2$ sur des données indépendantes (test, validation) des données d'entraînement (celles qui servent à estimer les paramètres du modèle)
 - (Δ sur les données d'entraînement, RSS croît avec la complexité, ...)

Régression linéaire

$y_{i1}, x_{i1}, \dots, x_{ip}$

réponse variable explicative

modèle : $y_i = \alpha + \beta_1 x_{i1} + \dots + \beta_p x_{ip} + \varepsilon_i$

residu $\varepsilon_i \sim \mathcal{N}(0, \sigma^2)$

paramètres du modèle: $\alpha, \beta = (\beta_1, \dots, \beta_p), \sigma^2$

$$\text{. . . } \sim \mathcal{N}(\alpha + \beta_1 x_{i1} + \dots + \beta_p x_{ip}, \sigma^2)$$

$$\Rightarrow y_i \sim N(\alpha + \beta_1 x_{i1} + \dots + \beta_p x_{ip}, \sigma^2)$$

On a observé $(y_i, x_{i1}, \dots, x_{ip})_{i=1,n}$ supposé indép.

$$\begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix} = \begin{pmatrix} \alpha \\ \vdots \\ \alpha \end{pmatrix} + \underbrace{\begin{pmatrix} x_{11} & \dots & x_{1p} \\ \vdots & \ddots & \vdots \\ x_{n1} & \dots & x_{np} \end{pmatrix}}_{\text{matrice de } X \text{ de taille } n \times p} \begin{pmatrix} \beta_1 \\ \vdots \\ \beta_p \end{pmatrix} + \begin{pmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_n \end{pmatrix}$$

$$y = \alpha 1_n + X \beta + \varepsilon$$

\uparrow vecteur de 1 de taille n \uparrow
 $(\beta_1, \dots, \beta_p)$

le modèle s'écrit matriciellement

$$y = \alpha 1_n + X \beta + \varepsilon$$

$$y | \alpha, \beta, \sigma^2 \sim \mathcal{N}\left(\underbrace{\alpha 1_n + X \beta}_{\text{modèle linéaire}}, \sigma^2 \underbrace{I_n}_{\text{matrice d'inconnues}} \right)$$

J2

- La statistique bayésienne permet d'introduire de l'information dont on dispose sur le phénomène étudié en plus de celle contenue dans les données

oui non
- Je peux choisir le prior que je souhaite en fonction du résultat que je veux obtenir

oui non

lorsque la taille de l'échantillon est grande, l'influence

- L'unique au niveau de la modélisation, l'apriori est faible
oui non
- Je peux utiliser une approche bayésienne sans connaissance a priori
oui non
- En statistique bayésienne, je compare des modèles pour tester mes hypothèses
oui non

• Le terme $p(\theta | \underline{x}) = \frac{l(\underline{x} | \theta) p(\theta)}{p(\underline{x})}$ est :

$\underline{x} = (x_1, \dots, x_n)$ $l(\underline{x} | \theta) p(\theta)$ $p(\underline{x})$
 la vraisemblance \underline{x} la loi a priori θ la loi a posteriori.

- Pour choisir le meilleur parmi deux modèles de complexité (nombre de paramètres) différentes, je peux utiliser la vraisemblance D **NON**, la vraisemblance D avec le critère de l'erreur de prédiction sur un échantillon test
 le critère AIC } vraisemblance pénalisée par la complexité
 le critère BIC }
 le critère R^2 **NON**, il croît avec la complexité
 le R^2 ajusté

Régression linéaire multivariée bayésienne

$$\begin{cases} y = \alpha \mathbf{1}_n + \mathbf{x} \beta + \epsilon \\ \epsilon \sim \mathcal{N}(0, \sigma^2 I_n) \end{cases}$$

→ il faut ajouter un prior sur les paramètres :
 α, β, σ^2

... D. Lois a posteriori Approximation NCPG

→ on estime ensuite les α, β, σ^2

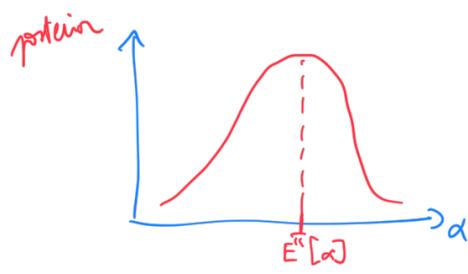
$$p(\alpha | x)$$

$$p(\beta | x)$$

$$p(\sigma^2 | x)$$

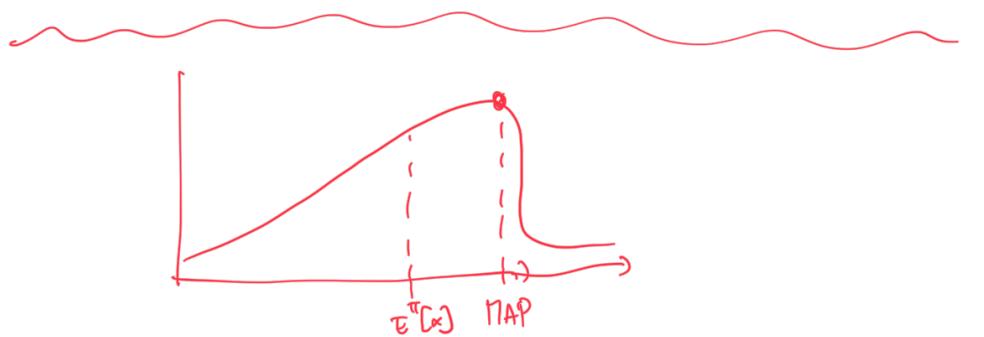
on peut de ces lois a posteriori tirer soit les MMAP soit l'espérance pour obtenir une estimation pointuelle des α, β, σ^2 .

On va choisir ici l'espérance, que l'on notera $E^\pi[\cdot]$ pour dire que c'est l'espérance de la loi a posteriori.



et on utilise comme estimation de α dans le modèle de régression logistique $E^\pi[\alpha]$!

idem pour $\beta, \sigma^2 \dots$



Fixe la loi a priori sur α, β, σ^2

$\alpha \in \mathbb{R}$ pas trop informatif comme paramètre ...

$\sigma^2 \in \mathbb{R}^+$ pas évident d'avoir un a priori ...

$\beta = \begin{pmatrix} \beta_1 \\ \vdots \\ \beta_p \end{pmatrix}$ sont les paramètres posteur du sens
on va traduire notre connaissance
sur le prior sur β .

$$\beta \in \mathbb{R}^T$$

$$p(\beta) \sim \mathcal{N}_p(\underbrace{\beta}_{\text{vecteur moyen}}, \underbrace{\Sigma}_{\text{covariance}})$$

que simple de fixer Σ ?

Si Y et X sont 2 variables aléatoires gaussiennes, alors la meilleure fonction $f(X)$ qui permet d'approcher Y , au

$$\text{seulement } E[\| Y - f(X) \|^2]$$

est la fonction linéaire $f(X) = aX + b$

\Rightarrow En pratique, on demandera à transformer Y pour qu'il soit le plus gaussien possible

$$y_i = x_{1i} + 3x_{2i} + \varepsilon_i$$

x_1 et x_2 très corélés
 $x_{1i} \approx x_{2i}$

$$\hookrightarrow y_i \approx x_{1i} + 3x_{1i} + \varepsilon_i$$

$$\approx 4x_{1i} + \varepsilon_i$$

$$\approx 4x_{2i} + \varepsilon_i$$

$$\approx 2x_{1i} - 2x_{2i} + \varepsilon_i$$

$$\hat{y} = \hat{\alpha}_0 + \hat{\alpha}_1 x_1 + \hat{\alpha}_2 x_2 + \dots$$

$$\hat{y} \approx -72x_{1i} + 68x_{2i} + \epsilon_i$$

Régression avec variables corélées et/ou en grande dimension

Rq: grande dimension (n petit devant p)
 \Rightarrow faible corrélation

Pb : corrélation (vraie ou fausse) entre les variables $\Rightarrow V(\hat{\beta})$ très grande,
 $\hat{\beta}$ non significatif.

Solutions

① sélection de variables

algorithme de type "step wise selection"

(marche bien si p est pas très grande, qq dizaines)

② construire de nouvelles méta-variables

qui résument l'information des variables

- régression sur les composantes principales de l'ACP

(pas top d'un point de vue interprétation
(car on ne travaille plus sur les variables de départ)

- régression PLS
(idem ...)

③ régression généralisée (Ridge, LASSO, ...)

- λ : . . . \rightarrow \hat{R} : \hat{t}_n

Régression linéaire classique, on trouve $\hat{\beta}_j$

$$\text{OLS} = \sum_{i=1}^n \left(y_i - \left(\alpha + \sum_{j=1}^p \hat{\beta}_j x_{ij} \right) \right)^2 \text{ sont minimum}$$

Pour éviter que les $\hat{\beta}_j$ prennent de trop grande valeurs (due à $V(\hat{\beta}_j)$ grande ...), on va les imposer de prendre des valeurs trop grande, en cherchant $\hat{\beta}$

RIDGE Régression

$$\sum_{i=1}^n \left(y_i - \left(\alpha + \sum_j \hat{\beta}_j x_{ij} \right) \right)^2 + \lambda \sum \hat{\beta}_j^2 \text{ sont minimum}$$

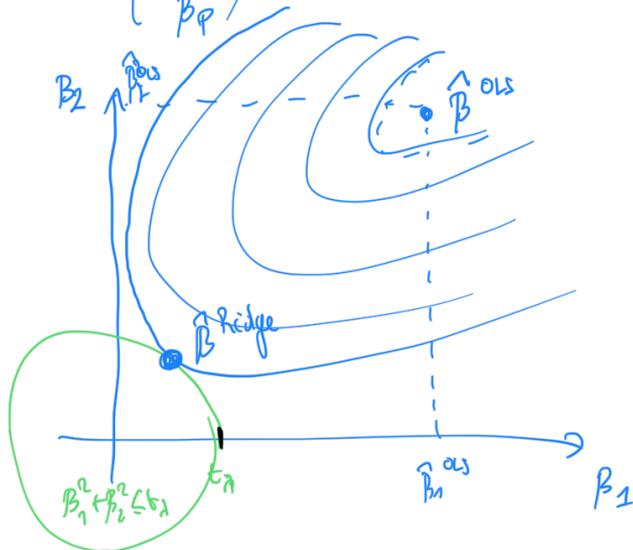
ou de façon équivalente

$$\sum_{i=1}^n \left[\underbrace{\left(y_i - \left(\alpha + \sum_j \hat{\beta}_j x_{ij} \right) \right)^2}_{\text{la contrainte}} \right] \text{ sont minimum sous } \sum \hat{\beta}_j^2 \leq t_\lambda$$

On peut montrer que la solution à ce pb est

$$\hat{\beta} = \begin{pmatrix} \hat{\beta}_1 \\ \vdots \\ \hat{\beta}_p \end{pmatrix} = \frac{(X'X + \lambda I)^{-1} X'Y}{\cancel{\text{determinant}}} \quad *$$

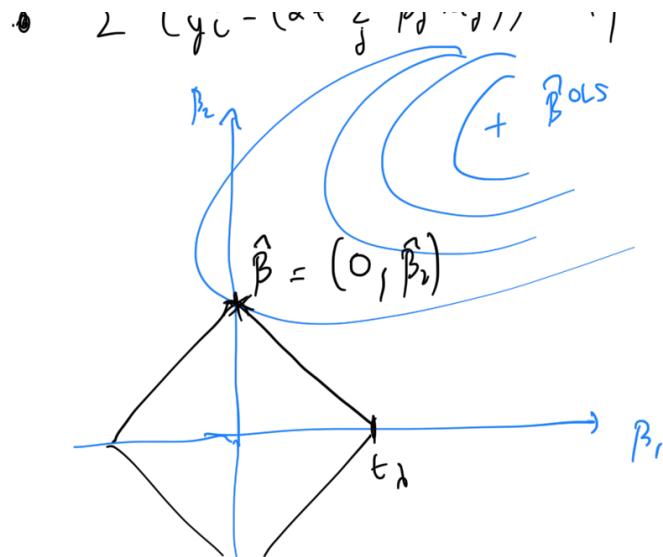
$p=2$



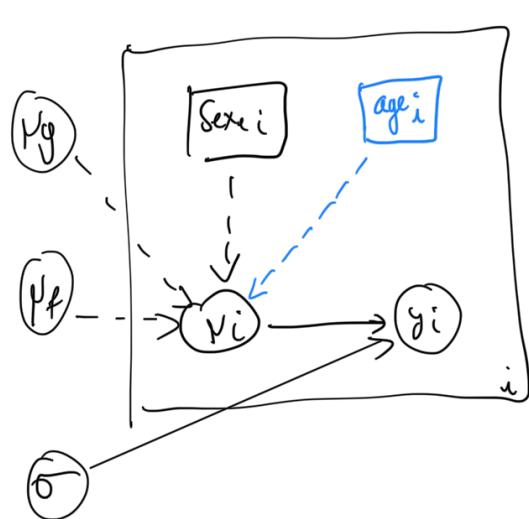
OLS: moindre carré classique

Régression LASSO

$$\sum (y_i - (\alpha + \sum_j \beta_j x_{ij}))^2 \text{ ta } \sum |\beta_j| \leq t_\lambda$$



permet de réduire la variance de $\hat{\beta}$ (et donc d'être bon en prediction, comme ridge) et en plus de faire de la sélection de variable (bon en interprétation,



$$y_i \sim N(\mu, \sigma^2)$$

$$\mu_i = \begin{cases} \mu_f & \text{si sexe=f} \\ \mu_g & \text{si sexe=g} \end{cases} + \beta \cdot (age_i - \bar{age})$$

$$\beta \sim U(0, 500)$$