Apprentissage et Sélection de Variables: (I) le cas linéaire

Nicolas Verzelen, Alexis Joly

INRA,Inria

M2 MIASH

Sommaire

1. Introduction : Apprentissage Statistique
   - Formalime de l'apprentissage supervisé
   - Fléau de la dimension
   - Modèles Linéaires
2. Sélection de variables par critères AIC/BIC et validation Croisée
3. Lasso
   - Préambule : Ridge
   - Définition du Lasso
   - Lasso et Grande dimension
   - Aspects Algorithmiques
4. Extensions du Lasso
   - Lasso et Modèles linéaires Généralisés
   - Lasso et Estimation Structurée

Bibliographie

Deux livres :

- Gareth James, Daniela Witten, Trevor Hastie and Robert Tibshirani
  An Introduction to Statistical Learning (2013) Springer
  Le livre est disponible en ligne gratuitement : `http://www.statlearning.com`
- Trevor Hastie, Robert Tibshirani, and Martin Wainwright
  Statistical Learning with Sparsity : The Lasso and Generalizations (2015) CRC Press
  Le livre est disponible en ligne gratuitement :
  `https://web.stanford.edu/~hastie/StatLearnSparsity/`

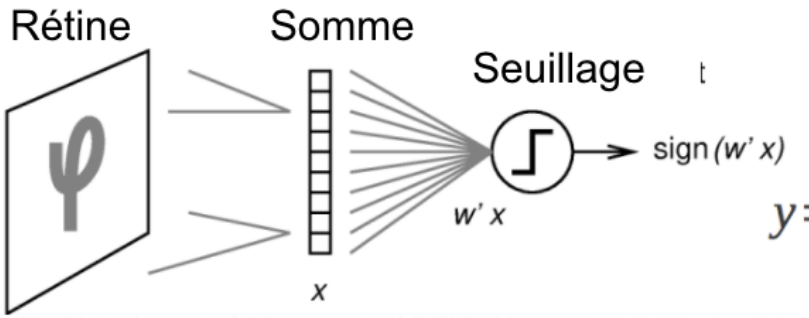
Problèmes d'apprentissage statistique

1. Identifier les facteurs de risque du cancer de la prostate
2. Prédire si une personne est sujette aux crises cardiaques, à partir de mesures cliniques, son régime et des données démographiques
3. Personnaliser un système de détection de spam email
4. Lecture de code postal écrit à la main
5. Classification d'échantillons de tissus dans différents types de cancer, en fonction de données d'expression de gènes
6. Établir une relation entre salaires et variables démographiques
7. Classifier les pixels d'une image satellite

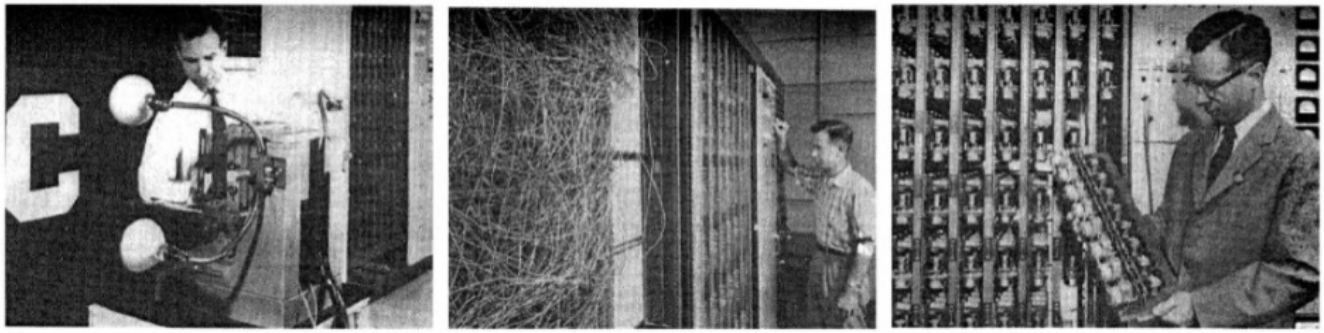
1957, le perceptron (”le première machine apprenante”)

Un “neurone” simulé avec des poids synaptiques adaptatifs

- Permet de calculer une somme pondérée des entrées
- Donne en sortie une valeur égale à +1 si la somme dépasse un certain seuil, -1 sinon



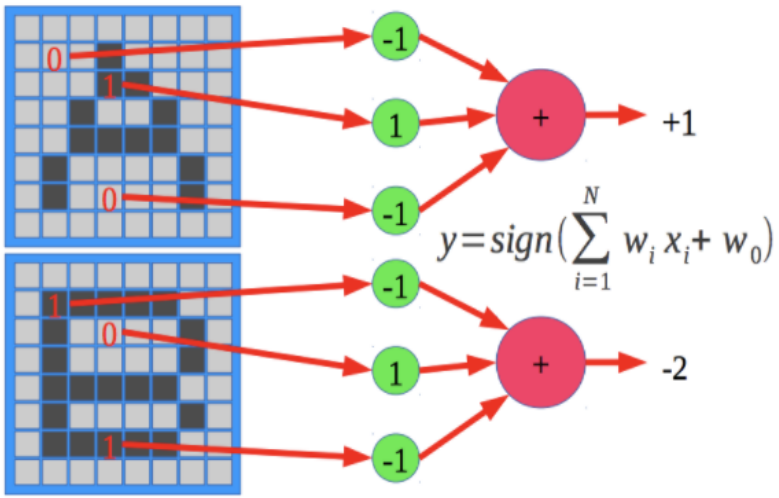
$$y = sign\left(\sum_{i=1}^{N} W_i X_i + b\right)$$



1957, le perceptron (”le première machine apprenante”)

Exemple d’apprentissage supervisé: reconnaissance des lettres “A” et “B”

- Objectif: trouver la valeur des poids qui donne +1 pour “A” et -1 pour B
- Ensemble d’apprentissage = (A ,+1) (A ,+1) (A ,+1) (B ,-1) (B ,-1)



Problème d'apprentissage supervisé

Point de départ

Y : réponse, variable dépendante, cible
X : données d'entrée, variables explicatives, co-variables,...

Régression : Y est quantitative, continue
Classification : Y est qualitative, discrète

Données d'apprentissage
$\mathcal{d}_1^n := \{(x_1, y_1), \ldots, (x_n, y_n)\}$
avec $x_i \in \mathscr{X}$ quelconque (souvent $\mathbb{R}^p$),
$y_i \in \mathscr{Y}$ pour $i = 1, \ldots, n$.

Objectifs

À partir de la base de données, on voudrait

- prédire le plus précisément possible la sortie y pour une nouvelle entrée x.
- comprendre quelle(s) co-variables influe sur la réponse, et comment
- (évaluer la qualité des inférences et prédictions)

Sommaire

1. Introduction : Apprentissage Statistique
   - Formalime de l'apprentissage supervisé
   - Fléau de la dimension
   - Modèles Linéaires
2. Sélection de variables par critères AIC/BIC et validation Croisée
3. Lasso
   - Préambule : Ridge
   - Définition du Lasso
   - Lasso et Grande dimension
   - Aspects Algorithmiques
4. Extensions du Lasso
   - Lasso et Modèles linéaires Généralisés
   - Lasso et Estimation Structurée

Modèle statistique non paramétrique

On suppose que d_1^n est l'observation d'un n-échantillon
$D_1^n = \{(X_1, Y_1), \ldots, (X_n, Y_n)\}$ d'une loi conjointe P sur $\mathscr{X} \times \mathscr{Y}$ inconnue
On suppose que x est une observation de la variable X, (X, Y) étant un couple aléatoire de loi conjointe P indépendante de D_1^n.

Definition

Une **règle de prédiction/ régression ou discrimination** est une fonction (mesurable) $f : \mathscr{X} \mapsto \mathscr{Y}$ qui associe la sortie $f(x)$ à l'entrée $x \in \mathscr{X}$.

Qualité de prédiction

Soit $l : \mathscr{Y} \times \mathscr{Y} \to \mathbb{R}_+$ une fonction de perte (i.e. $l(y, y) = 0$ et $l(y, y') > 0$ pour $y \neq y'$, par exemple :

- $l(y, y') = |y - y'|^q$ en régression réelle
  (perte absolue si $q = 1$, perte quadratique si $q = 2$)
- $l(y, y') = 1_{y \neq y'}$ en discrimination binaire.

Le **risque** - ou l'**erreur de généralisation** - d'une règle de prédiction f est défini par :

$$R_P(f) = \mathbb{E}_{(X,Y)\sim P}[l(Y, f(X))]$$

Si $\mathscr{F}$ désigne l'ensemble des règles de prédiction possibles, quelles sont les règles de prédiction optimales au sens de la minimisation du risque sur $\mathscr{F}$, c'est à dire les règles f^* telles que $R_P(f^*) = \inf_{f \in \mathscr{F}} R_P(f)$?

Régression réelle et discrimination

On appelle **fonction de régression** la fonction $\eta^* : \mathscr{X} \to \mathscr{Y}$ définie par $\eta^*(x) = \mathbb{E}[Y|X = x]$

Cas de la régression réelle

$$\mathscr{Y} = \mathbb{R}, \quad l(y, y') = (y - y')^2$$

Théorème

La fonction de régression η^* *vérifie* $R_P(\eta^*) = \inf_{f \in \mathscr{F}} R_P(f)$

Cas de la discrimination binaire

$$\mathscr{Y} = \{-1, 1\}, \quad l(y, y') = 1_{y \neq y'} = |y - y'|/2 = (y - y')^2/4$$

On appelle **règle de Bayes** toute fonction ϕ^* de $\mathscr{F}$ telle que pour tout $x \in \mathscr{X}$, $\mathbb{P}(Y = \phi^*(x)|X = x) = \max_{y \in \mathscr{Y}} \mathbb{P}(Y = y|X = x)$.

Théorème

Si ϕ^* est une règle de Bayes, alors $R_P(\phi^*) = \inf_{f \in \mathscr{F}} R_P(f)$.

Théorème

La règle de discrimination plug in définie par

$$\phi_{\eta^*}(x) = \textit{signe}(\eta^*(x)) = 1_{\eta^*(x)\geqslant 0} - 1_{\eta^*(x)<0}$$

est une règle de Bayes

Dans tous les cas, ces règles de prédiction optimales dépendent de P !

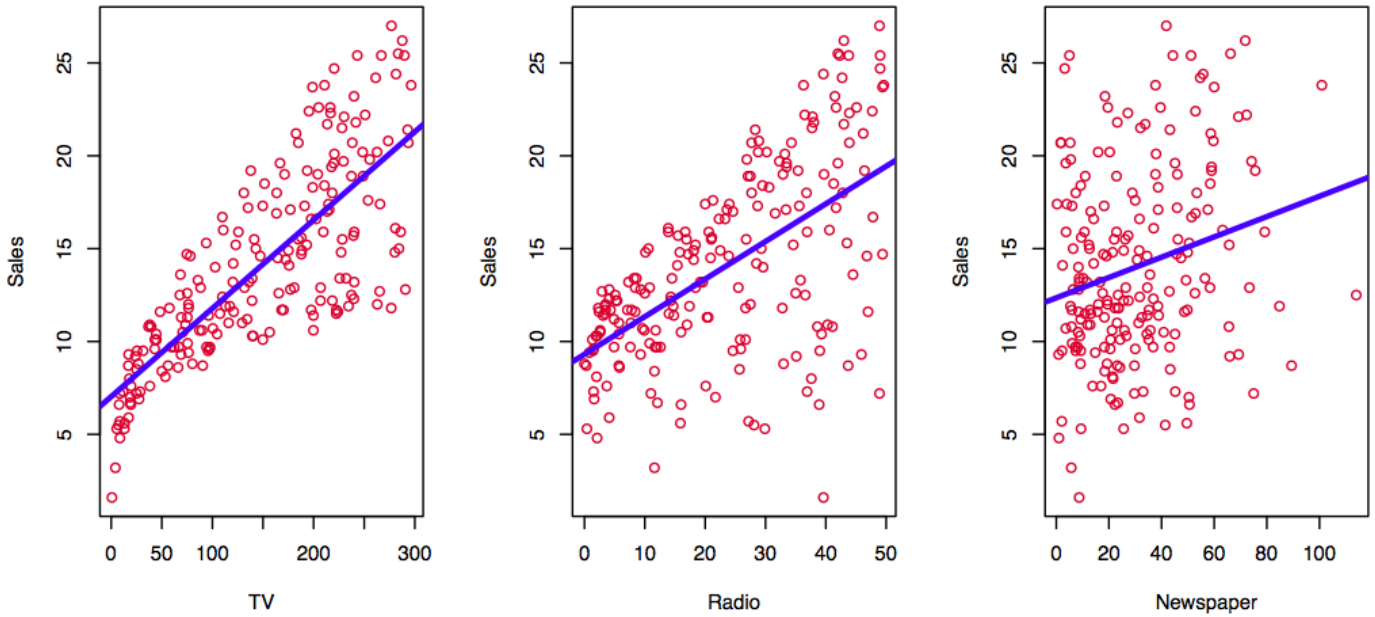
$\rightsquigarrow$ Nécessité de construire des règles - ou algorithmes - de prédiction qui ne dépendent pas de P mais des données d_1^n.

Le **risque moyen** d'un algorithme de prédiction $\widehat{f}$ construit sur d_1^n est défini par

$$\mathscr{R}(\widehat{f}) = \mathbb{E}_{D_1^n \sim P^{\otimes n}}\left[\mathbb{E}_{(X,Y)\sim P}[l(Y, \widehat{f}(X))]\right]$$

Remarque : On veut utiliser des procédures $\widehat{f}$ telle que $\mathscr{R}(\widehat{f})$ est le plus petit possible.
Le risque moyen d'un algorithme dépend de P inconnu ! Comment faire en pratique ?

Soyons un peu plus concret



Sur le graphique :
ventes (`Sales`) en fonction de

- pub TV (`TV`)
- pub radio (`Radio`)
- pub presse (`Newspaper`)

Les trois lignes bleues sont trois régression linéaire simple

Objectif : prédire `Sales` en utilisant les trois variable

i.e : construire une fonction `f` tel que

$$\texttt{Sales} \approx \texttt{f}(\texttt{TV}, \texttt{Radio}, \texttt{Newspaper})$$

Cadre

Notations

- `Sales` est la réponse. On la note Y. $\mathscr{Y} = \mathbb{R}$.
- `TV`, `Radio`, `Newspaper` sont les covariables. On les note $X^{(1)}, X^{(2)}, X^{(3)}$.
- Le vecteur des covariables

$$X = \begin{pmatrix} X^{(1)} \\ X^{(2)} \\ X^{(3)} \end{pmatrix}$$

On a $\mathscr{X} = \mathbb{R}^3$.

La fonction de régression $\eta^*(x) = \mathbb{E}(Y|X = x)$ minimise le risque quadratique ($l(y, y') = (y - y')^2$) .
Si on définit $\varepsilon = Y - \eta^*(X)$, alors on a

$$Y = \eta^*(X) + \varepsilon \ ,$$

avec $\mathbb{E}(\varepsilon|X) = 0$.

ε capture l'information de Y non prédictible par X.

La décomposition ci-dessus ne **repose sur aucune hypothèse** sur le lien entre Y et X (si ce n'est l'intégrabilité de Y).

Pourquoi estimer une règle $\widehat{f}$ proche de η^* ?

- Prédiction : faire des **prédictions** précises de `Sales` (Y) pour une nouvelle valeurs de (`TV,Radio,Newspaper`) à de nouveaux points $X = x$

- Explication ($\neq$ Causalité !) : On peut comprendre quelles composantes de $X = (X^{(1)}, \ldots, X^{(p)})$ sont importantes pour expliquer Y (exemple : `Experience` et `Diplome` ont une grande influence sur le `Salaire`, mais le `Statut marital` en a peu, ou pas)

Comment estimer la règle optimale à partir de d_1^n ?

Quelques exemples en **régression** :

- Moyenne mobile :

$$\widehat{f}_h(x) = \frac{\sum_{i=1}^n Y_i 1_{X_i \in \mathcal{N}_h(x)}}{\sum_{i=1}^n 1_{X_i \in \mathcal{N}_h(x)}} ,$$

  avec la convention $0/0 = 0$ et où $\mathcal{N}_h(x)$ est un *voisinage* de x de rayon h.

- Modèle linéaires : estimer η^* par une fonction de la forme

$$\beta_0 + \beta_1 X^{(1)} + \cdots + \beta_p X^{(p)}$$

- Modèles additif généralisés (GAM) (voir livre ISLR)) : estimer par une fonction de la forme

$$g_1(X^{(1)}) + \cdots + g_p(X^{(p)}) ,$$

  où $g_1, \ldots, g_p$ sont des fonctions non-linéaires (ex : polynômes, splines,...)

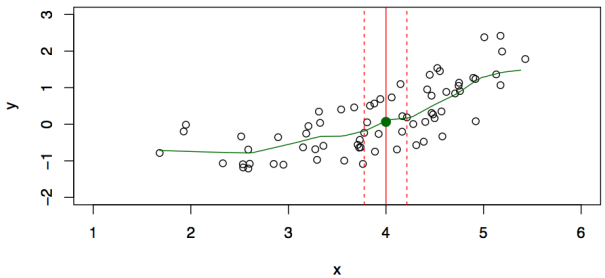
Comment choisir une méthode ?

Décomposition Biais-variance

Considérons le risque **moyen d'une procédure** $\widehat{f}$.

$$\mathscr{R}(\widehat{f}) = \underbrace{\mathbb{E}_{X\sim P_X}\left(\eta^*(X) - \mathbb{E}_{D_1^n\sim P^{\otimes n}}(\widehat{f}(X))\right)^2}_{\text{Biais}} + \underbrace{\mathbb{E}_{X\sim P_X}\left[\text{Var}_{D_1^n\sim P^{\otimes n}}(\widehat{f}(X))\right]}_{\text{Variance}}$$
$$+ \underbrace{R_P(\eta^*)}_{\text{Irréductible}}$$

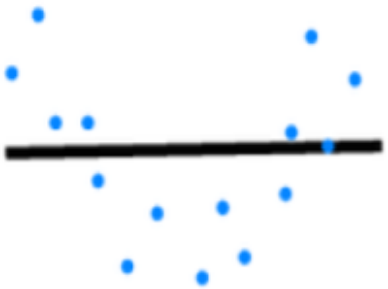
Exemple de la moyenne Mobile :



$\rightsquigarrow$ une bonne procédure doit exhiber un bon compromis biais-variance.

Problème : Le biais dépend de la distribution P (inconnue) des données !
$\rightsquigarrow$ la meilleure procédure (en termes de risques) est inconnue.

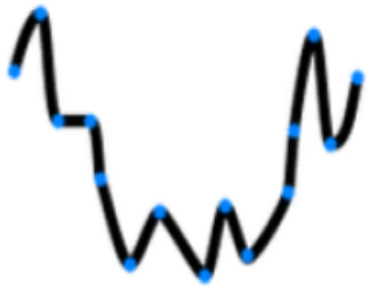
Compromis Biais-variance (1)



Underfitting

= sous-ajustement
= sous-apprentissage
= biais élevé

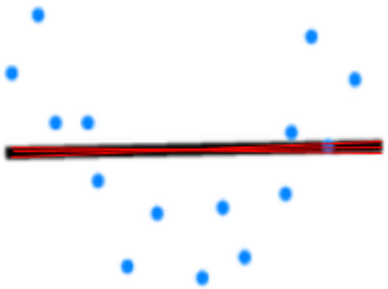
Desired



Over-fitting

= sur-ajustement
= sur-apprentissage
= biais faible

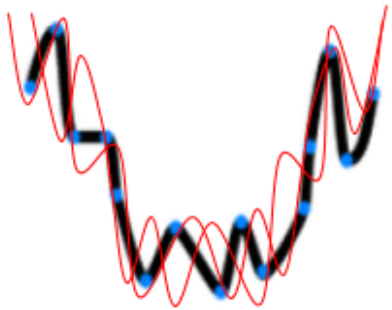
Compromis Biais-variance (2)



Underfitting

= sous-ajustement
= sous-apprentissage
= biais élevé
= variance faible

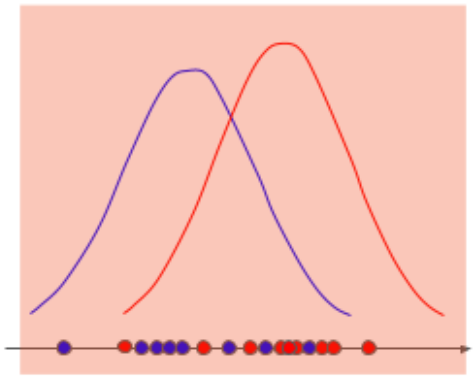
Desired



Over-fitting

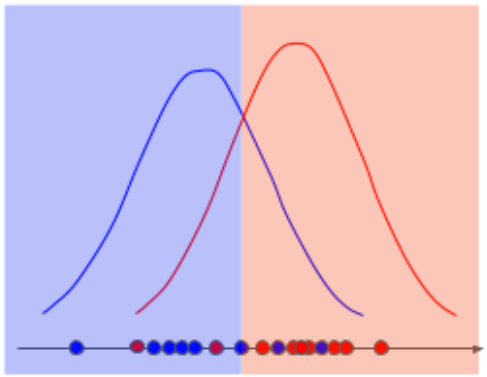
= sur-ajustement
= sur-apprentissage
= biais faible
= variance élevée

Compromis Biais-variance (3)

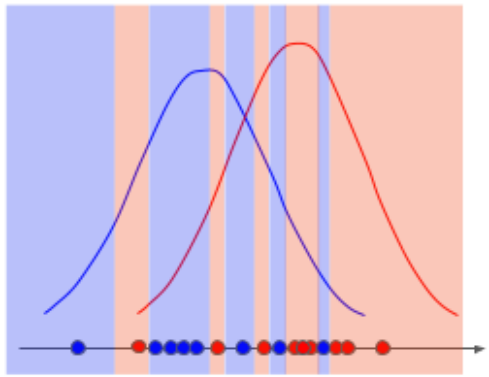


Underfitting

= sous-ajustement
= sous-apprentissage
= biais élevé
= sur-généralisation



Desired



Over-fitting

= sur-ajustement
= sur-apprentissage
= biais faible
= sur-mémorisation

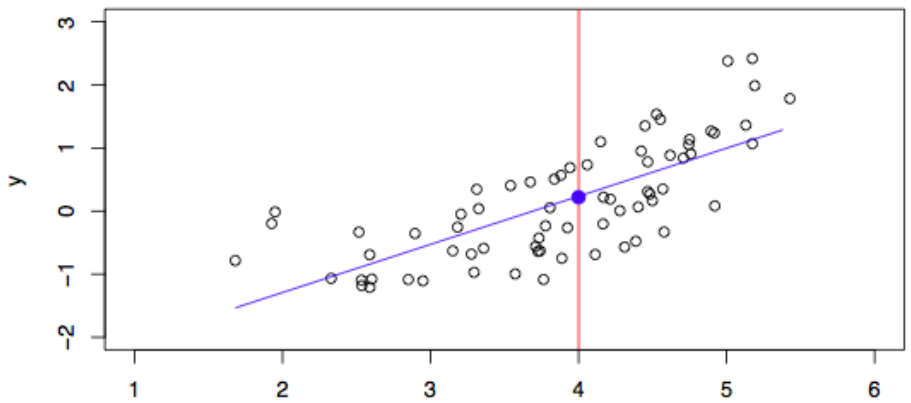
Modèles Linéaires

Les modèles *linéaires*. Construire une règle de régression dans la famille $\mathscr{F}_L \subset \mathscr{F}$.

$$\mathscr{F}_L := \{f(X) = \beta_0 + \beta_1 X^{(1)} + \cdots + \beta_p X^{(p)}; \beta \in \mathbb{R}^{p+1}\}$$

- La dimension de $\mathscr{F}_L$ vaut $p + 1$.
- Lorsque la perte est quadratique, la règle $\widehat{f}_L$ est généralement ajustée par le critère des moindres carrés sur d_1^n.

Exemple : $\widehat{f}_L(x) = \widehat{\beta}_0 + \widehat{\beta}_1 x$

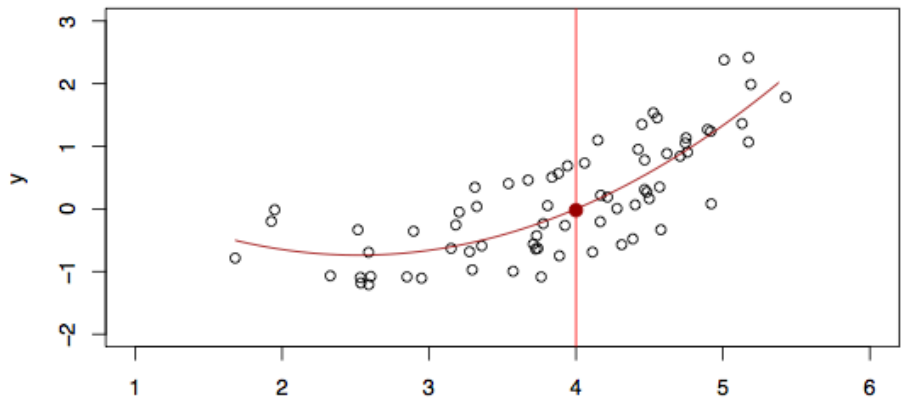


En général, $\eta^* \notin \mathscr{F}_L$: les modèle linéaires ne sont *presque jamais corrects*. Néanmoins, ils peuvent fournir bonnes approximations interprétables de la vraie fonction $\eta^*(X)$

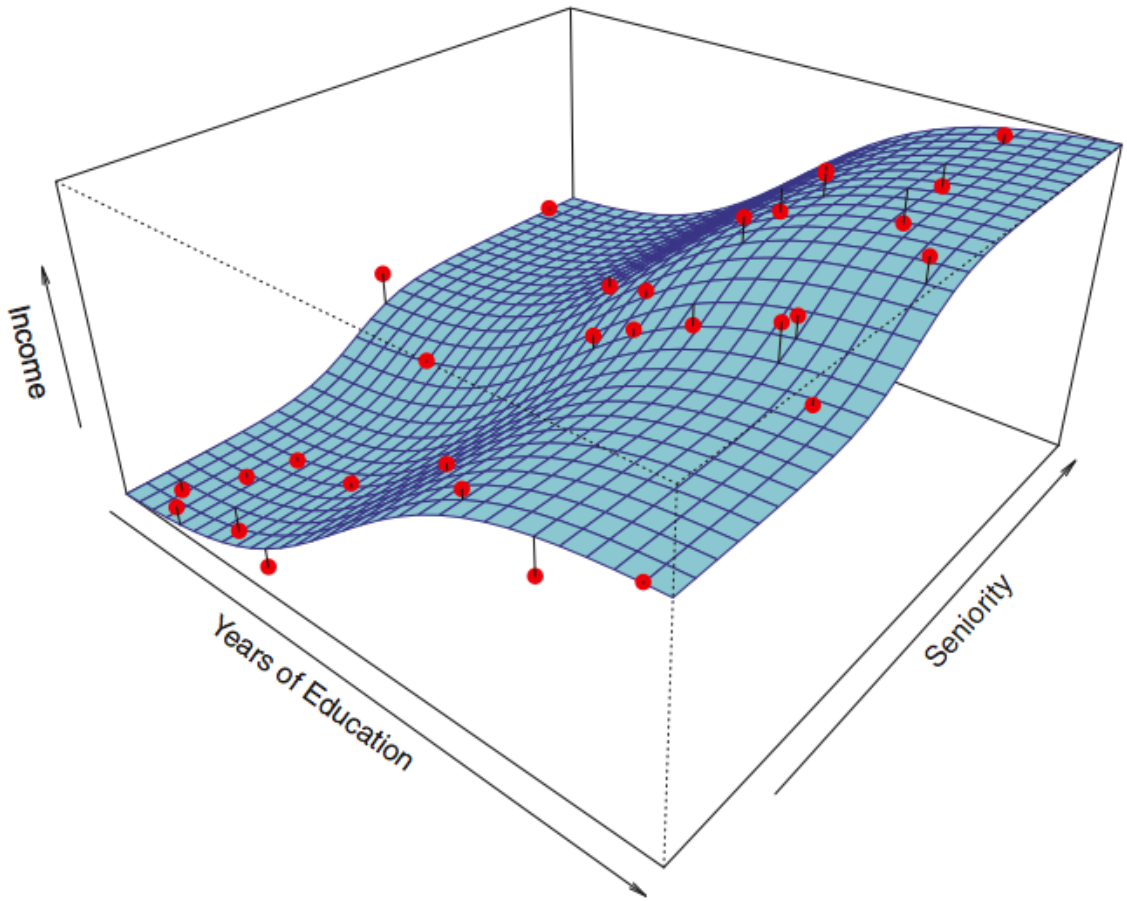
Modèle paramétrique (suite)

Modèle quadratique :

$$\mathscr{F}_Q := \{f(X) = \beta_0 + \beta_1 X + \beta_2 X^2\}$$



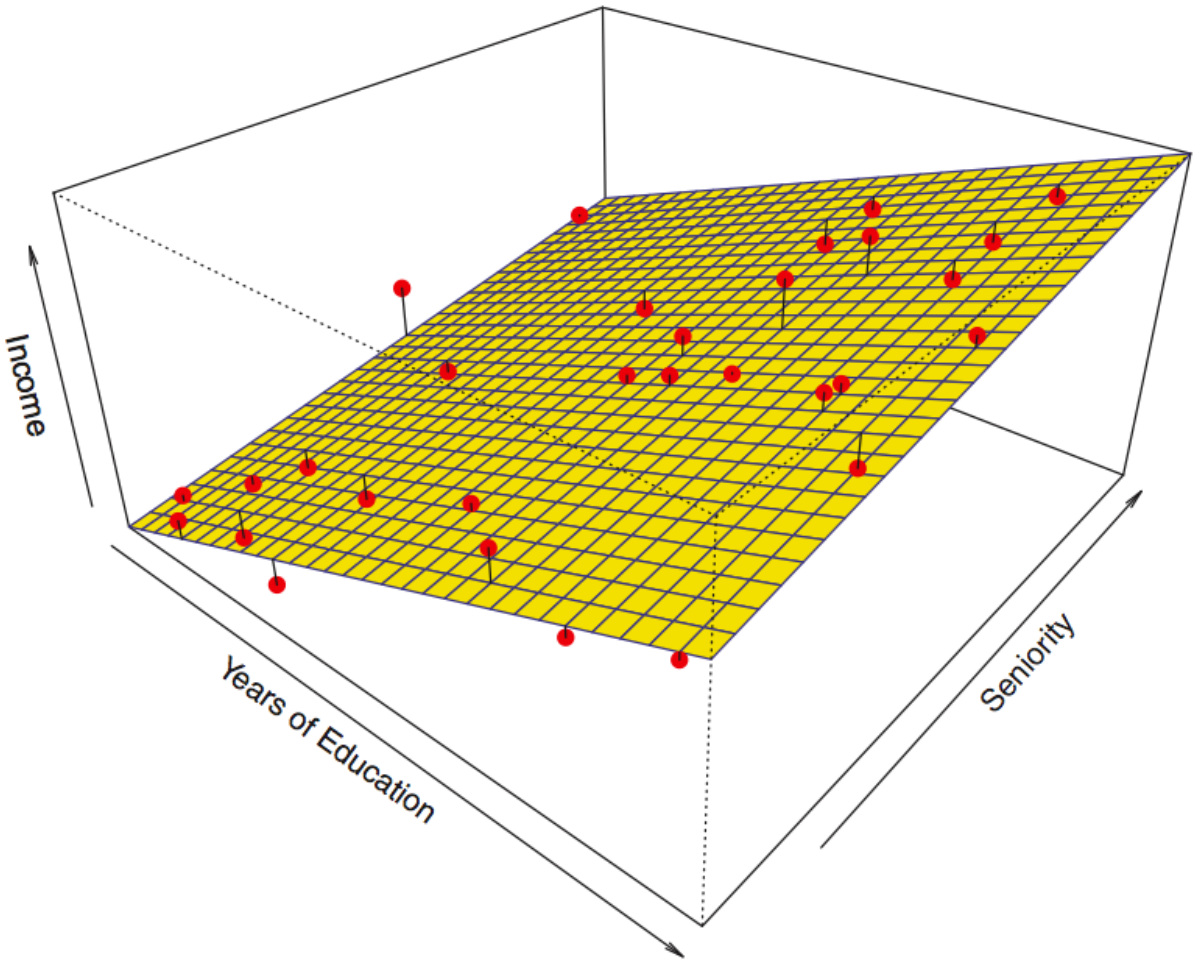
Un exemple simulé



$$\texttt{income} = \eta^*(\texttt{education}, \texttt{seniority}) + \varepsilon$$

La véritable fonction η^* est la surface bleue

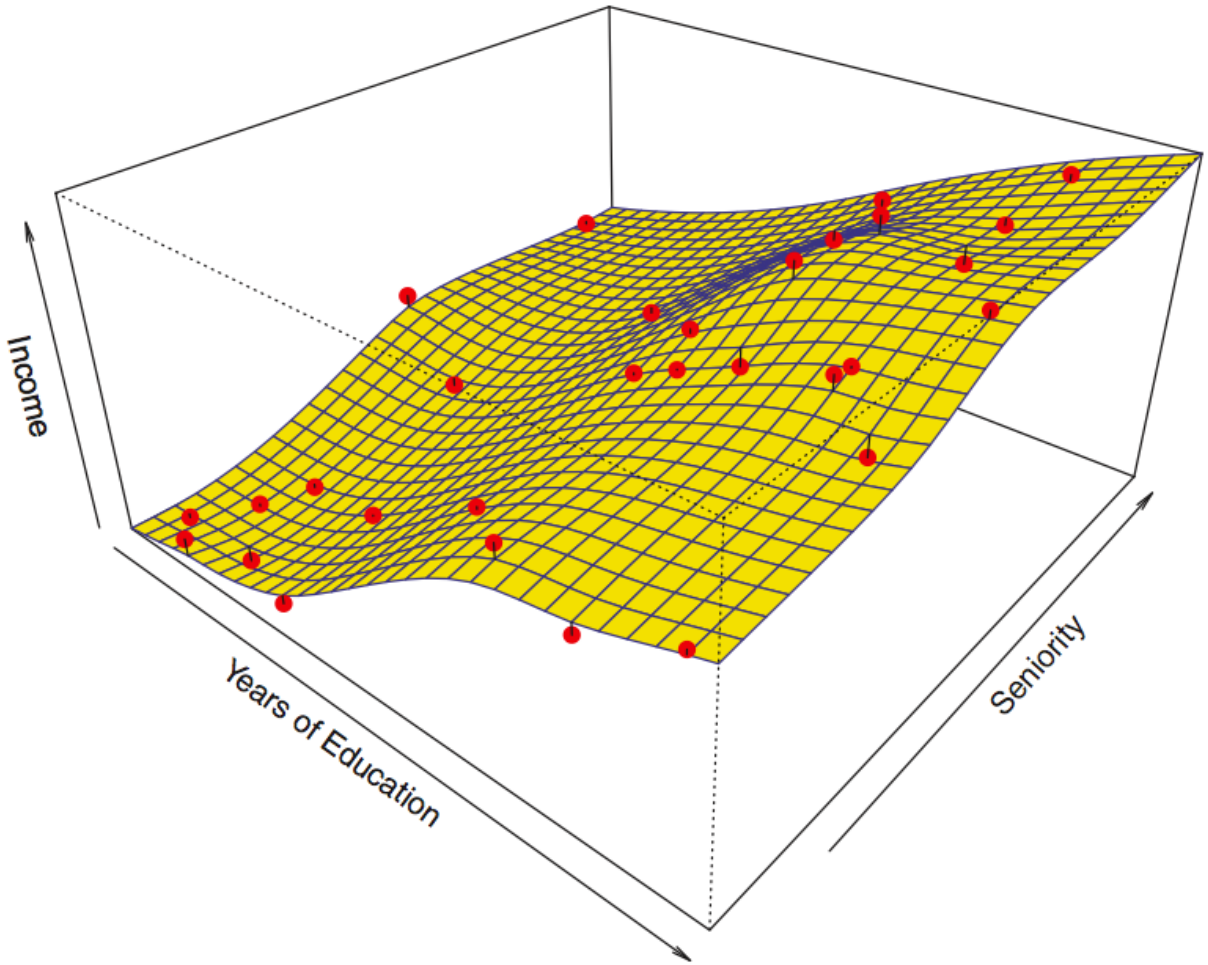
Un exemple simulé : modèle linéaire



$$\widehat{f_L}(\texttt{education}, \texttt{seniority}) = \widehat{\beta_0} + \widehat{\beta_1} \times \texttt{education} + \widehat{\beta_2} \times \texttt{seniority}$$

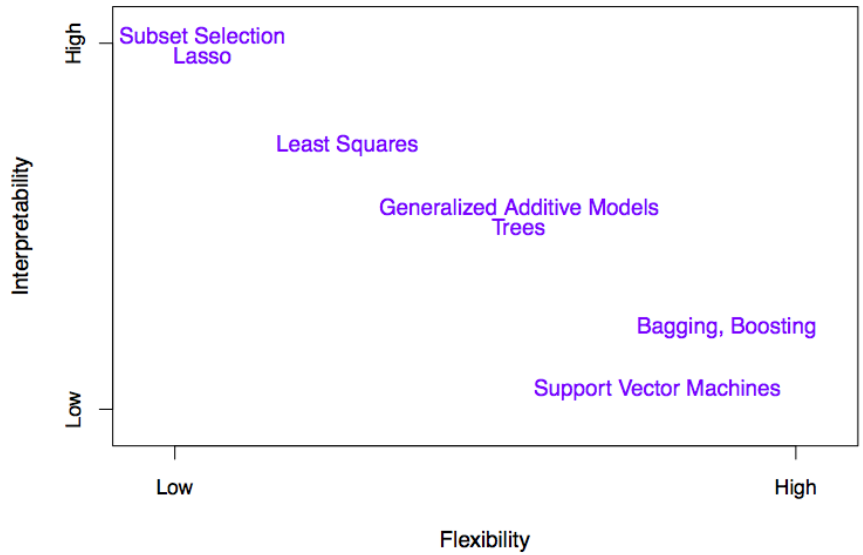
ajustée aux données

Un exemple simulé : un modèle de splines



Des méthodes, des compromis

- ajustement *vs*
  sur-ajustement (variance trop grande)
  vs sous-ajustement (biais trop grand)
- Flexibilité *vs*
  interprétabilité des résultats
  (ex : moyenne mobile vs. modèle linéaire)



Toutes une gamme de méthodes

En résumé

- Il n'existe pas de méthode universellement meilleure que les autres.
- Sélectionner une approche nécessite de savoir les comparer, i.e. d'estimer le risque de plusieurs règles
  ($\rightsquigarrow$ ex : Validation croisée)
- Le choix de la méthode dépend aussi des objectifs du statisticiens (interprétation).

Sommaire

1. Introduction : Apprentissage Statistique
   - Formalime de l'apprentissage supervisé
   - Fléau de la dimension
   - Modèles Linéaires
2. Sélection de variables par critères AIC/BIC et validation Croisée
3. Lasso
   - Préambule : Ridge
   - Définition du Lasso
   - Lasso et Grande dimension
   - Aspects Algorithmiques
4. Extensions du Lasso
   - Lasso et Modèles linéaires Généralisés
   - Lasso et Estimation Structurée

Fléau de la dimension (1) : méthodes basées sur des voisinages.

Ex : Moyenne mobile :

(ou méthode à noyaux)

$$\widehat{f}_h(x) = \frac{\sum_{i=1}^n Y_i 1_{X_i \in \mathcal{N}_h(x)}}{\sum_{i=1}^n 1_{X_i \in \mathcal{N}_h(x)}}$$

- Ces méthodes, basées sur des moyennes autour des voisins sont plutôt bonnes si
  - petite dimension $p \leqslant 4$
  - grand échantillon $n \gg p$

Ces méthodes peuvent être *mauvaises* quand p est grand.

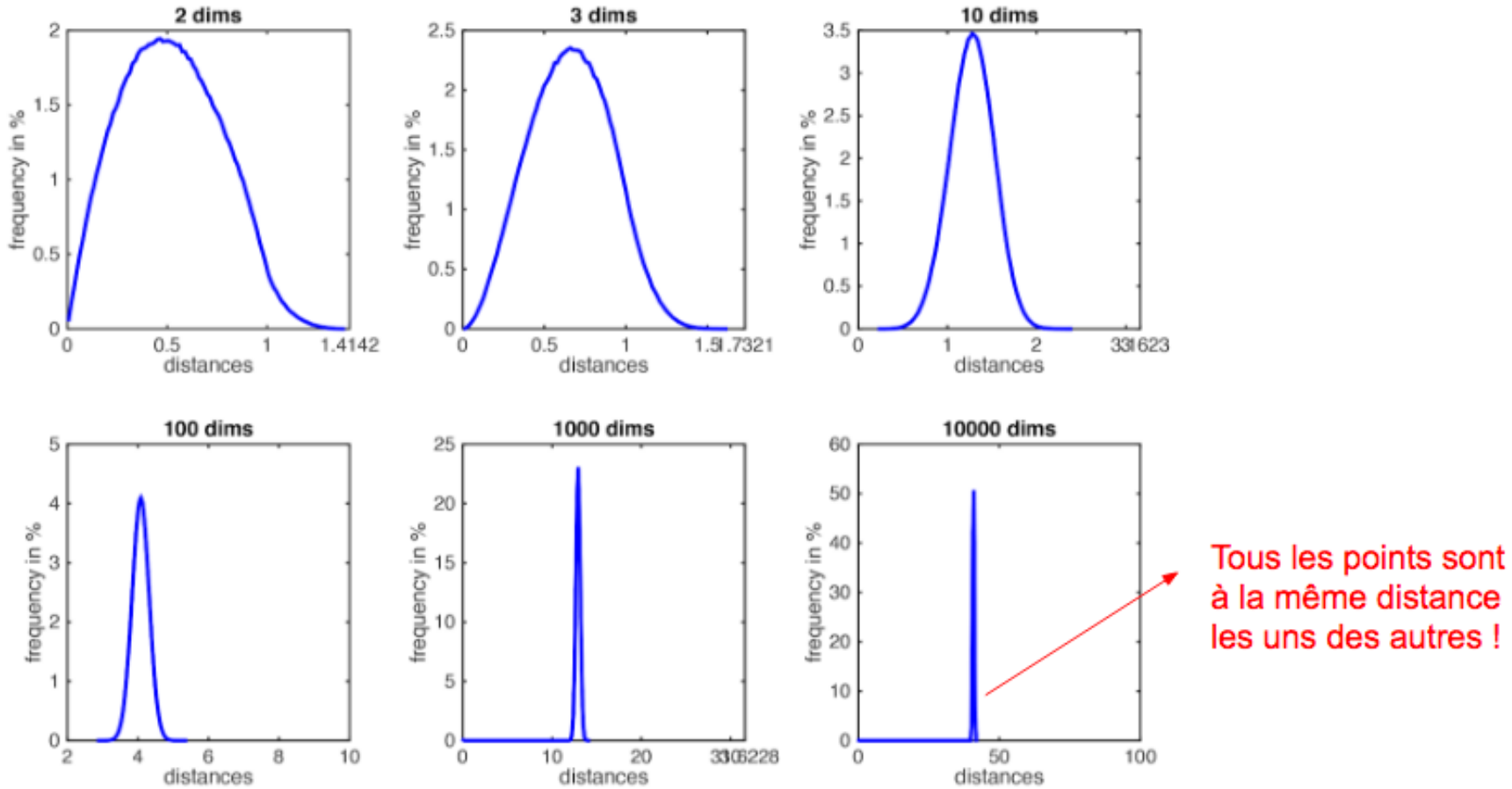
Raison. le *fléau de la dimension*.
Les voisins les plus proches peuvent être éloignés en grande dimension

- Il faut une quantité raisonnable de valeurs de y_i à moyenner pour que $\widehat{f}_h(x)$ ait une faible variance
- En grande dimension, pour obtenir cette quantité d'observation, il faut s'éloigner beaucoup de x.

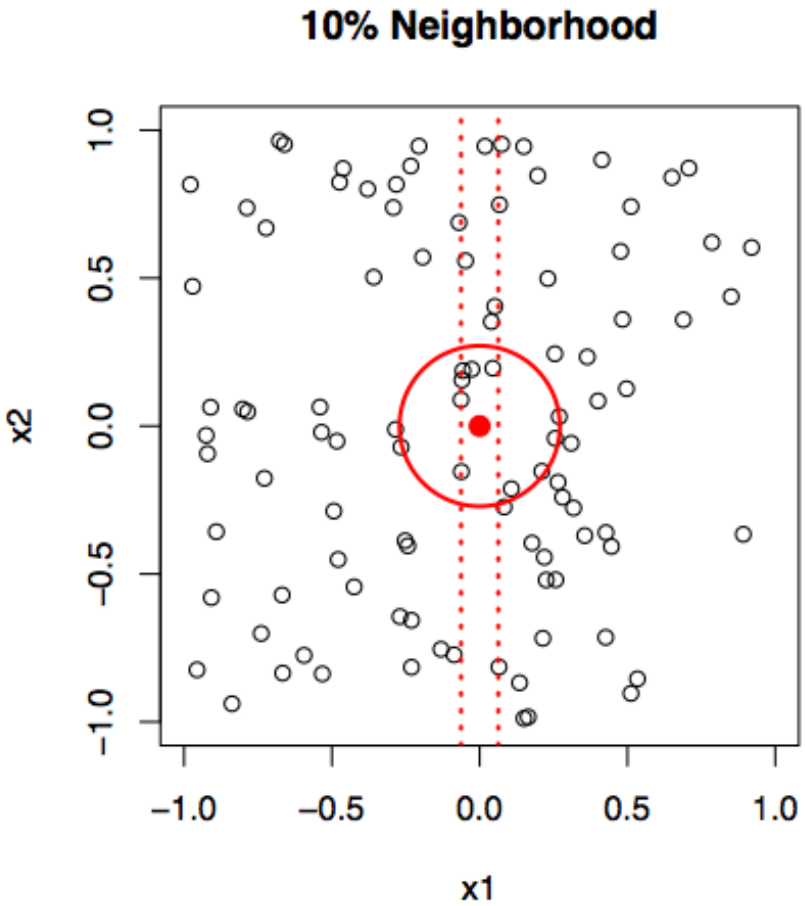
On perd l'idée de moyenne **locale** autour de $X = x$.

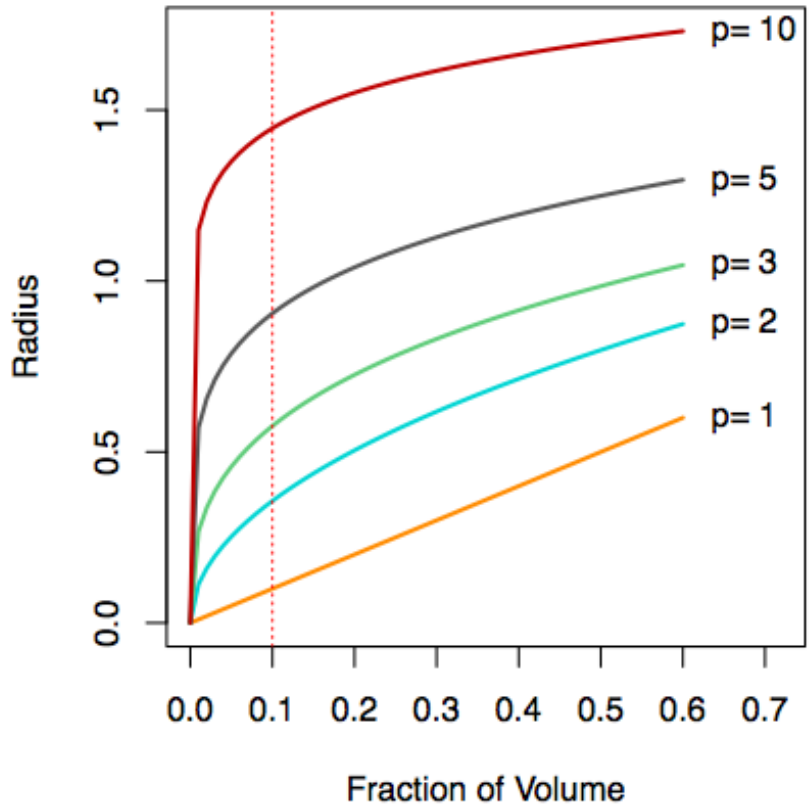
Le fléau de la dimension (2)

Distribution de la distance de points tirés aléatoirement dans un espace multidimensionnel



Le fléau de la dimension (3)

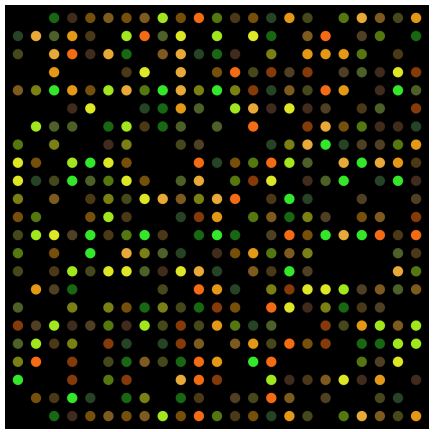




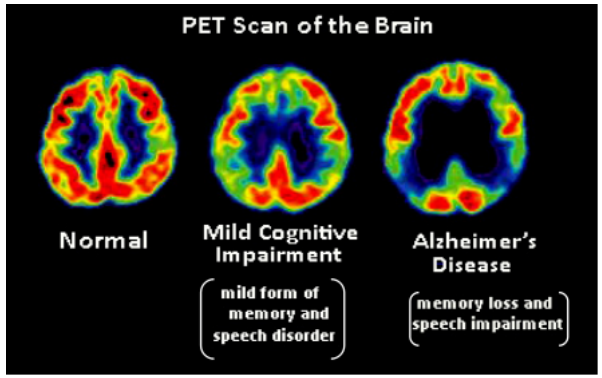
Fléau de la dimension (4) : problèmes réels en grande dimension

On cherche à prédire Y par p covariables $(X^{(1)}, \ldots, X^{(p)})$.

Quelques **problèmes** modernes :



Expériences Microarray
$n \approx 100$, $p \approx 1000$
Obj. : Relier un caractère à l'expression des gènes



IRM Fonctionnel de cerveau
$n \approx 100$, $p > 10000$
Obj. : Relier un caractère à l'activité de certaines zones du cerveaux (voxels)

$p + 1$ paramètres $> n$ Observations :
$\rightsquigarrow$ non unicité de l'estimateur des moindres carrés, Grande variance !

Sommaire

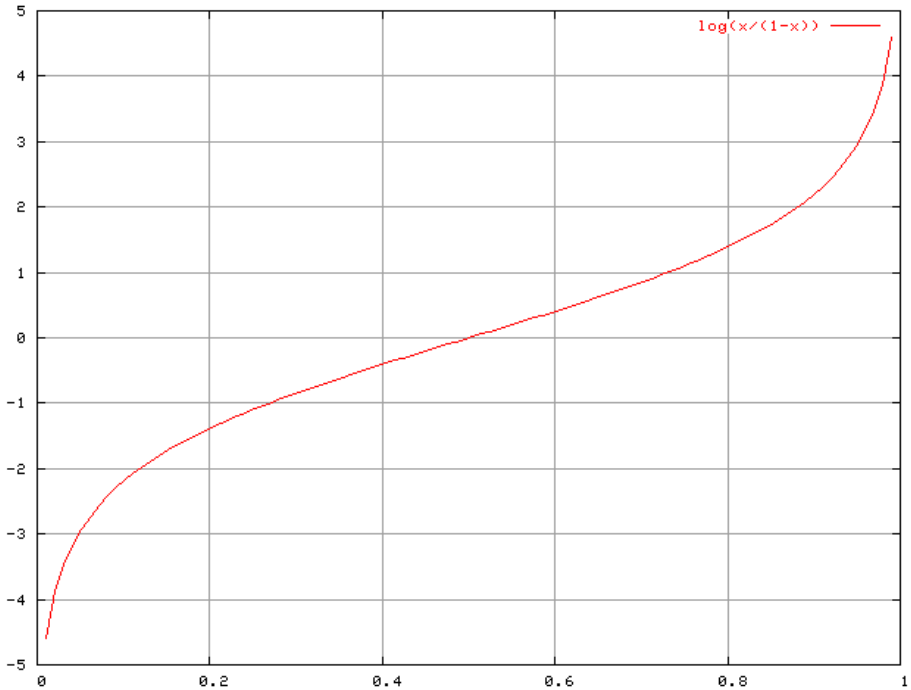
1. Introduction : Apprentissage Statistique
   - Formalime de l'apprentissage supervisé
   - Fléau de la dimension
   - Modèles Linéaires
2. Sélection de variables par critères AIC/BIC et validation Croisée
3. Lasso
   - Préambule : Ridge
   - Définition du Lasso
   - Lasso et Grande dimension
   - Aspects Algorithmiques
4. Extensions du Lasso
   - Lasso et Modèles linéaires Généralisés
   - Lasso et Estimation Structurée

Modèles linéaires et logistiques

- Rappelons que le modèle linéaire cherche à expliquer Y grâce à un modèle de la forme

$$\beta_0 + \beta_1 X^{(1)} + \cdots + \beta_p X^{(p)} \tag{1}$$

- La plupart des approches décrites ici s'étendent simplement au modèle de régression logistique pour lequel on modélise $\text{logit}\left(\mathbb{P}\left[Y = 1 | (X^{(1)}, \ldots, X^{(p)})\right]\right)$ sous la forme (1).



Notations vectorielles

Dans la suite, on notera l'échantillon D_1^n du modèle de régression linéaire sous la forme :

$$\mathbf{Y} = \mathbf{X}\beta^* + \boldsymbol{\varepsilon} \ ,$$

où $\mathbf{Y} = \begin{pmatrix} Y_1 \\ \dots \\ Y_n \end{pmatrix}$, $\mathbf{X}_{i,j} = X_i^{(j)}$, $\beta^* = \begin{pmatrix} \beta_1^* \\ \dots \\ \beta_p^* \end{pmatrix}$ et $\boldsymbol{\varepsilon} = \begin{pmatrix} \varepsilon_1 \\ \dots \\ \varepsilon_n \end{pmatrix}$

ATTENTION : Pour simplifier la présentation des méthodes, on supposera parfois que $\beta_0^* = 0$.

On peut facilement ajouter ce paramètre en ajoutant la colonne constante $\begin{pmatrix} 1 \\ \dots \\ 1 \end{pmatrix}$ à la matrice de design $\mathbf{X}$. Dans les packages R décrits dans ce cours, le coefficient d'ordonnée à l'origine β_0^* est toujours estimé.

Défendons les modèles linéaires

- Malgré leur simplicité, le modèle linéaire a des avantages en termes d'**interprétabilité** et souvent il fournit de bonnes **performances prédictives**

Critère des moindres carrés

Le modèle linéaire est généralement ajusté par le critère des moindres carré. Si on note $l(y, y') = (y - y')^2$ la perte quadratique,

$$\widehat{\beta} \in \arg\min_{\beta} \frac{1}{n} \sum_{i=1}^{n} l\left(Y_i, \sum_{j=1}^{p} X_i^{(j)} \beta_j\right)$$

Expression alternative en notation matricielles

$$\widehat{\beta} \in \arg\min_{\beta \in \mathbb{R}^p} \|\mathbf{Y} - \mathbf{X}\beta\|_2^2$$

Si $\mathbf{X}^T\mathbf{X}$ est inversible, alors $\widehat{\beta} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T Y$.

Moindres carrés et minimisation du risque empirique

Plus généralement, le critères moindres carrés fait partie de la famille des méthodes d'ajustement par **minimisation du risque empirique**. Soit $F \subset \mathscr{F}$ une collection de règle de prédictions

$$\widehat{f} \in \arg\min_{f \in F} \widehat{R}_n(f) \ ,$$

où $\widehat{R}_n(f) = \frac{1}{n} \sum_{i=1}^{n} l(Y_i, f(X_i))$.

Sommaire

1. Introduction : Apprentissage Statistique
   - Formalime de l'apprentissage supervisé
   - Fléau de la dimension
   - Modèles Linéaires
2. Sélection de variables par critères AIC/BIC et validation Croisée
3. Lasso
   - Préambule : Ridge
   - Définition du Lasso
   - Lasso et Grande dimension
   - Aspects Algorithmiques
4. Extensions du Lasso
   - Lasso et Modèles linéaires Généralisés
   - Lasso et Estimation Structurée

Pourquoi considérer des alternatives aux moindres carrés ?

Dans ce cours, nous étudierons des alternatives aux critères des moindres carrés (et plus généralement de minimisation du risque empirique) :

- **Pour améliorer le risque (diminuer la variance)** : en particulier lorsque $\mathrm{p} > \mathrm{n}$.
- **Pour l'interprétation des modèles** : en supprimant les covariables inutiles, c'est-à-dire en annulant les coefficients correspondants, on obtient un modèle qui s'interprète plus facilement.

Que ce soit pour augmenter la précision ou pour améliorer l'interprétabilité, nous allons chercher à sélectionner des variables.

Deux classes de méthodes

- **Selection d'un sous-ensemble**. Nous identifions un sous-ensemble des p prédicteurs pour lesquels nous pensons qu'ils sont en lien avec la réponse. Nous ajustons ensuite un modèle par moindres carrés sur le sous-ensemble réduit.
- **Régularisation**. Nous ajustons un modèle sur l'ensemble complet des p prédicteurs, mais les coefficients estimés sont tirés vers 0 par rapport à l'estimateur des moindres carrés. Cette méthode de **régularisation** réduit la variance, et peut aussi aider à sélectionner les variables.

Sélection de modèles

Objectif : Sélectionner un sous-ensemble de variables explicatives qui explique au mieux Y. Soit $m \subset \{1, \dots, p\}$ un sous-ensemble d'indices. On note

$$\widehat{\beta}_m \in \arg \min_{\beta, \text{supp}(\beta) \subset m} \|Y - X\beta\|_2^2 ,$$

l'estimateur des moindres carrés sur des paramètres β dont toutes les coordonnées en dehors de m sont fixés à zéro.

Il correspond à l'estimateur des moindres carrés dans le modèle linéaire dont seules les variables $X^{(j)}$, $j \in m$ ont été gardée.

Le problème de la sélection de modèle est le suivant. Etant donnée une collection $\mathcal{M} = \{m_1, \dots, m_r\}$ de modèles, on veut sélectionner le modèle $m^* \in \mathcal{M}$ tel que

$$\mathbb{E}_{(Y,X)}\left[\left(Y - \sum_{j=1}^{p} X^{(j)} (\widehat{\beta}_{m^*})_j\right)^2\right] = R(\widehat{f}_{m^*}) \text{ est le plus petit possible ,}$$

où $\widehat{f}_m(X) = \sum_{j=1}^{p} X^{(j)} (\widehat{\beta}_m)_j$.

Deux exemples de problèmes de sélection de modèles

1. **Sélection ordonnée**. Supposons qu'il existe un ordre naturel sur les covariables $X^{(1)}$, ... $X^{(p)}$.
   Exemple : régression polynomiale $X^{(1)} = X$, $X^{(2)} = X^2$, ..., $X^{(k)} = X^k$.
   L'objectif est de sélectionner le "meilleur" degré du polynôme pour prédire Y.
   $\mathscr{M} := \{\{1\}, \{1,2\}, \{1,2,3\}, \ldots, \}$
2. **Sélection compléte**. On veut choisir les "meilleurs" covariables $X^{(1)}, \ldots X^{(p)}$.
   $\mathscr{M} = \mathscr{P}(\{1, \ldots, p\})$, l'ensemble des parties de $\{1, \ldots, p\}$.

Comment sélectionner un bon modèle

- Sélectionner le modèle qui minimise l'erreur d'entrainement

$$\widehat{\mathrm{R}}_n(\widehat{\mathrm{f}}_m) = \frac{1}{n}\|\mathbf{Y} - \mathbf{X}\widehat{\beta}_m\|_2^2$$

  est une mauvaisee idée. Pourquoi ?

- L'objectif étant de choisir un modèle dont le risque $\mathrm{R}(\widehat{\mathrm{f}}_m)$ est le plus petit possible, il es naturel de vouloir estimer ce risque pour chaque $\widehat{\mathrm{f}}_m$, $m \in \mathscr{M}$. Deux approches s'offrent à nous :
  1. Estimer le risque en *ajustant* l'erreur d'entrainement pour tenir compte du biais dû au sur-apprentissage
  2. Estimer *directement* l'erreur de test, par une approche de validation ou une approche de validation croisée

Pénalisation : AIC (C_p) et BIC

- Ces techniques corrigent l'erreur d'entrainement par la taille du modèle, et peuvent être utilisées pour sélectionner des modèles de dimension différentes.

$$\widehat{m} \in \arg\min_{m \in \mathscr{M}} \frac{1}{n} \|\mathbf{Y} - \mathbf{X}\widehat{\beta}_m\|_2^2 + \text{pen}(m)$$

où pen : $\mathscr{M} \to \mathbb{R}^+$ est une pénalité qui va pénaliser les plus grands modèles.
Dans la suite, on va voir deux (ou trois) fonctions de pénalités différentes.

$$\begin{aligned} \text{pen}_{AIC}(m) = \text{pen}_{C_p}(m) &= 2\widehat{\sigma}_m^2 \frac{|m|}{n} \\ \text{pen}_{BIC}(m) &= \log(n)\widehat{\sigma}_m^2 \frac{|m|}{n}, \end{aligned}$$

où $\widehat{\sigma}_m^2 = \|\mathbf{Y} - \mathbf{X}\widehat{\beta}_m\|_2^2/n$.

Justification de ces pénalités : heuristique de Mallows

Sortons un peu du cadre d'apprentissage statistique et considérons et le modèle de régression linéaire :

$$\mathbf{Y} = \mathbf{X}\beta^* + \varepsilon$$

où $\varepsilon \sim \mathcal{N}(0, \sigma^2 \mathrm{I}_n)$ et $\mathbf{X}$ est supposé **déterministe**.
Considèrons le critère suivant

$$\mathrm{Crit}_{C_p}(m) := \frac{1}{n}\|\mathbf{Y} - \mathbf{X}\widehat{\beta}_m\|_2^2 + 2\frac{\sigma^2}{n}|m|$$

Notons β_m la meilleure approximation de β^* dans m :
$\beta_m \in \arg\min_{\beta,\ \mathrm{supp}(\beta)\subset m} \|\mathbf{X}\beta - \mathbf{X}\beta^*\|_2^2$.

Heuristique de Mallows (suite)

Proposition

Supposons que $X_m^T X_m$ *est inversible. Alors,*

$$\begin{aligned} \mathbb{E}\left[\|Y - X\widehat{\beta}_m\|_2^2\right] &= \|X\beta_m - X\beta^*\|_2^2 + \sigma^2(n - |m|) \\ \mathbb{E}\left[\|X\beta^* - X\widehat{\beta}_m\|_2^2\right] &= \|X\beta_m - X\beta^*\|_2^2 + \sigma^2|m| \end{aligned}$$

Donc $\mathbb{E}[\mathrm{Crit}_{C_p}(m)] = \mathbb{E}\left[\|X\beta^* - X\widehat{\beta}_m\|_2^2\right] + n\sigma^2$

La C_p de Mallows est un estimateur sans-biais du risque (à design fixe) de $\widehat{\beta}_m$!

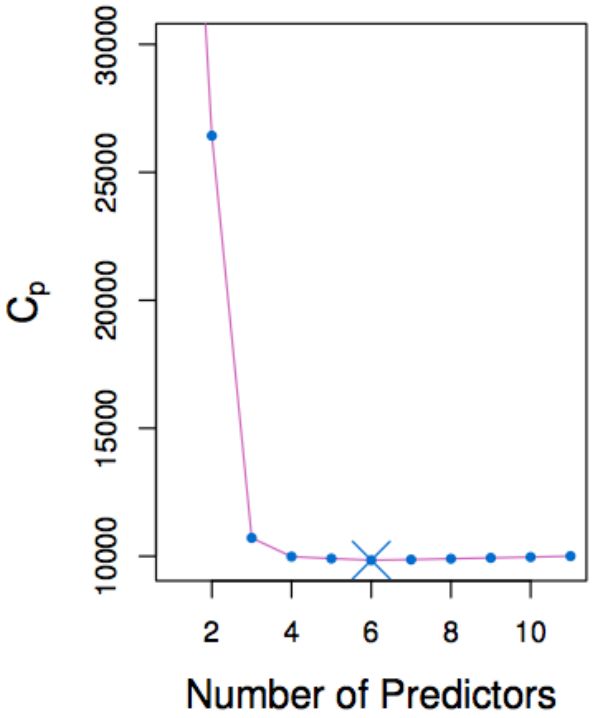
Remarque sur les pénalités AIC et BIC

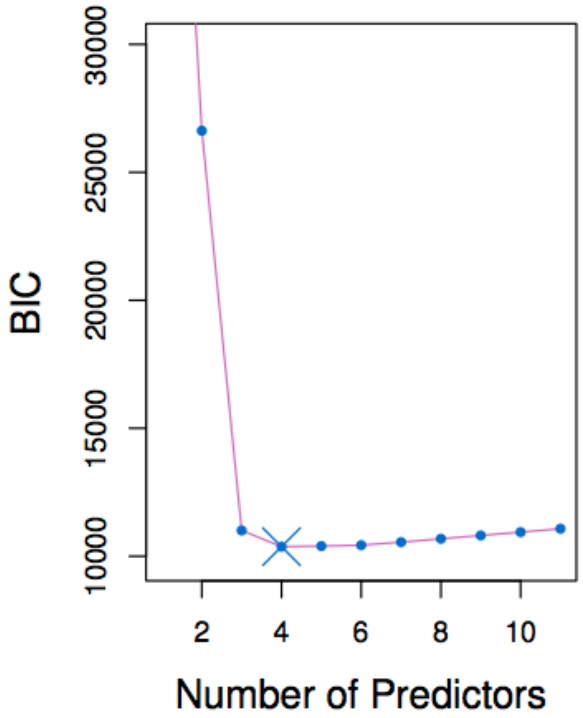
$$\begin{aligned} \mathrm{pen}_{\mathrm{AIC}}(m) &= 2\hat{\sigma}^2_m \frac{|m|}{n} \\ \mathrm{pen}_{\mathrm{BIC}}(m) &= \log(n)\hat{\sigma}^2_m \frac{|m|}{n} \ , \end{aligned}$$

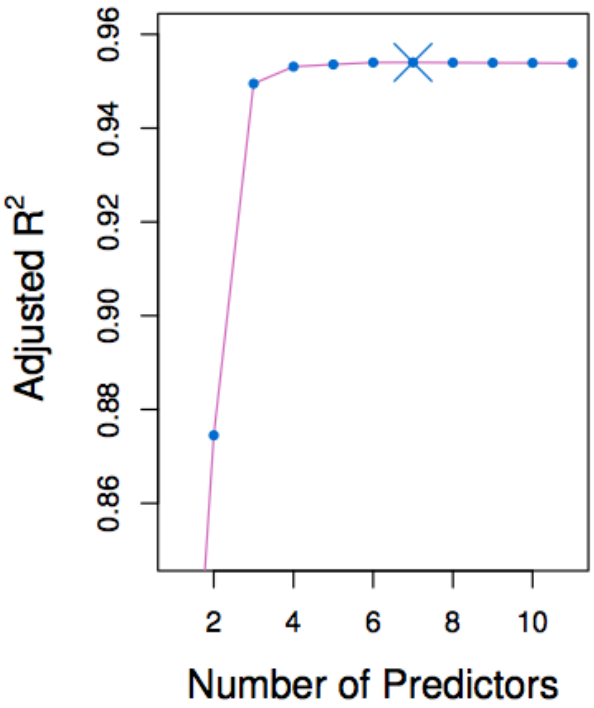
- Comme les C_p, BIC a tendance à être petit lorsque le risque est petit, et on choisit donc généralement le modèle qui a la plus petite valeur de BIC.
- Notons que BIC remplace le $2\,\hat{\sigma}^2$ utilisé par C_p par un terme $\log(n)\,\hat{\sigma}^2$ où n est le nombre d'observations.
- Puisque $\log(n) > 2$ dès que $n > 7$, le critère BIC pénalise plus les modèles de grandes dimensions. Les modèles choisis avec ce critère seront donc de dimension plus petite.

Exemple : jeu de données de crédit

La figure suivante montre C_p et BIC pour le meilleur modèle de chaque dimension pour le jeu de données de crédit







Comparaison de ces critères

- AIC sont des critères qui réalisent un compris biais-variance. Ils sont donc indiqués pour choisir un modèle que l'on souhaite utiliser pour prédire.
- BIC pénalise plus les modèles de grandes dimensions. C'est le seul critère à être consistent (i.e., à sélectionner le vrai modèle $\text{supp}(\beta^*)$ avec probabilité tendant vers 1 lorsque $n \to \infty$)
- BIC étant plus sélectif, on doit le préférer si l'on souhaite un modèle explicatif.
- Lorsque la taille de la base d'apprentissage est grande, préférer BIC (AIC fournit des modèles de trop grandes dimensions)
- **ATTENTION** : Lorsque p est grand (au moins de l'ordre de n), les pénalités BIC et AIC peuvent s'avérer trop petites et il faut recourir à d'autres pénalités.

Extensions à des modèles ajustés par maximum de vraisemblance (ex : régression logistique)

- Le critère **AIC** (Akaike Information Criterion) est défini pour une large de modèles ajustés par maximum de vraisemblance par :

$$\mathrm{Crit}_{\mathrm{AIC}}(m) = -2 \log \mathrm{L}(\widehat{\beta}_m) + 2\, \mathrm{d}_m$$

où L est la vraisemblance maximale pour le modèle de dimension d_m considéré.

- Le critère **BIC** (Bayesian Information Criterion) est défini par

$$\mathrm{Crit}_{\mathrm{BIC}}(m) = -2 \log \mathrm{L}(\widehat{\beta}_m) + \log(n)\, \mathrm{d}_m$$

Validation et validation croisée

Une alternative à la pénalisation est d'estimer le risque de chaque estimateur $\widehat{\beta}_m$ par validation croisée.
Le modèle $\widehat{m}$ est choisie comme minimiseur du critère suivant

$$\mathrm{Crit}_{\mathrm{CV}}(m) = \widehat{R}^{\mathrm{CV}}(\widehat{f}_m)\,,$$

où $\widehat{R}^{\mathrm{CV}}(\widehat{f}_m)$ est un estimateur par validation croisée du risque $R(\widehat{f}_m)$.

AVANTAGE : La sélection par validation croisée permet de sélectionner des estimateurs sans aucune hypothèse sur la distribution des données ou les procédures d'estimation.

INCONVENIENT : Plus coûteux en temps de calcul. Légérement moins efficace que la pénalisation lorsque la vraie distribution des données est celle d'un modèle linéaire gaussien.

Retour sur la sélection compléte de variables

1. Pour chaque valeur de k entre 1 et p :
   - Choisir le meilleur parmi ces $\binom{p}{k}$ modèles et le noter $\widehat{m}_k$.
2. Choisir le meilleur modèle parmi $\widehat{m}_1, \dots \widehat{m}_p$ en utilisant la validation croisée, ou AIC, ou BIC.

Cet algorithme est équivalent à la méthode présentée précédemment.

Sélection pas à pas

- Pour des raisons de calcul, la sélection du meilleur sous-ensemble ne peut pas être appliquée quand p est grand. *Pourquoi ?*
- Les méthodes *pas à pas*, qui n'explorent qu'une sous-partie de l'ensemble de tous les modèles possibles sont plus attirantes pour sélectionner le meilleur sous-ensemble.

Sélection pas à pas progressive

- La sélection progressive commence par le modèle nul et ajoute progressivement des prédicteurs au modèle, un par un, jusqu'à ce que l'on utilise tous les prédicteurs.
- En particulier, à chaque étape, la variable qui conduit à la meilleure amélioration du modèle est ajoutée.

En détail

Sélection pas à pas progressive

1. Noter $\widehat{\mathrm{m}}_0$ le **modèle nul**, qui ne contient aucun prédicteurs. Ce modèle prédit simplement la réponse Y avec $\mathbb{E}(Y)$, ou plutôt la moyenne empirique $\bar{Y}$.
2. Pour chaque valeur de k entre 0 et $\mathrm{p}-1$:
   1. Considérer tous les $(\mathrm{p}-\mathrm{k})$ modèles qui consistent à ajouter un prédicteur à $\widehat{\mathrm{m}}_\mathrm{k}$.
   2. Choisir le meilleur parmi ces $(\mathrm{p}-\mathrm{k})$ modèles et noter le $\widehat{\mathrm{m}}_{\mathrm{k}+1}$. Ici, le *meilleur* modèle est celui qui minimise le critére des moindres carrés.
3. Choisir le meilleur modèle parmi $\widehat{\mathrm{m}}_0, \widehat{\mathrm{m}}_1, \ldots \widehat{\mathrm{m}}_\mathrm{p}$ en utilisant la validation croisée, ou les AIC ou BIC.

Sélection pas à pas progressive (suite)

- L'avantage en terme de temps de calcul par rapport à la méthode exhaustive du meilleur sous-ensemble est claire.
- Rien ne garantit de trouver le meilleur modèle possible parmi les 2^p modèles.

Exemple : jeu de données crédit

Nb covar	Meilleur sous-ensemble	Sélection progressive pas à pas
1	`rating`	`rating`
2	`rating`, `income`	`rating`, `income`
3	`rating`, `income`, `student`	`rating`, `income`, `student`
4	`cards`, `income`, `student`, `limit`	`rating`, `income`, `student`, `limit`

Les trois premiers modèles sont identiques, mais le dernier est différent de ce qu'on trouve par sélection compléte.

Sélection pas à pas rétrograde

- Comme la sélection pas à pas progressive, la *sélection pas à pas rétrograde* propose une méthode efficace alternative au meilleur sous-ensemble.
- Cependant, contrairement à la méthode progressive, elle commence par le modèle complet, ajusté par moindres carrés, contenant les p prédicteurs, et les supprime un à un.

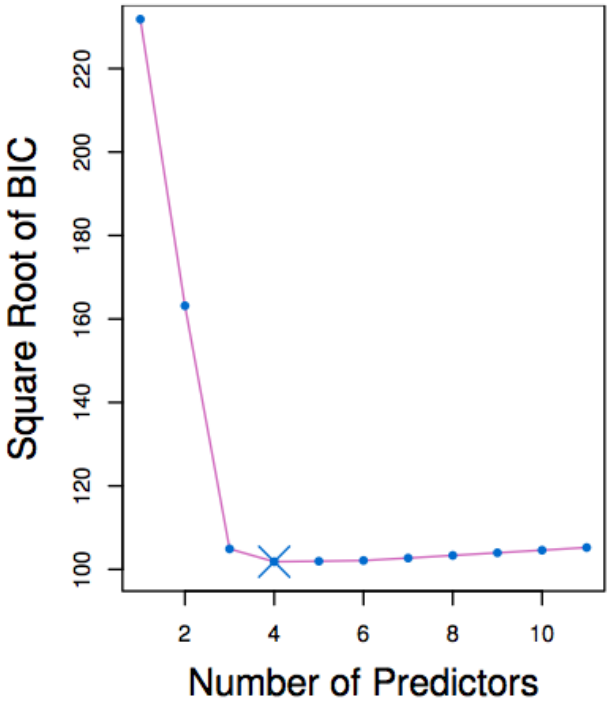
Sélection pas à pas rétrograde :détail

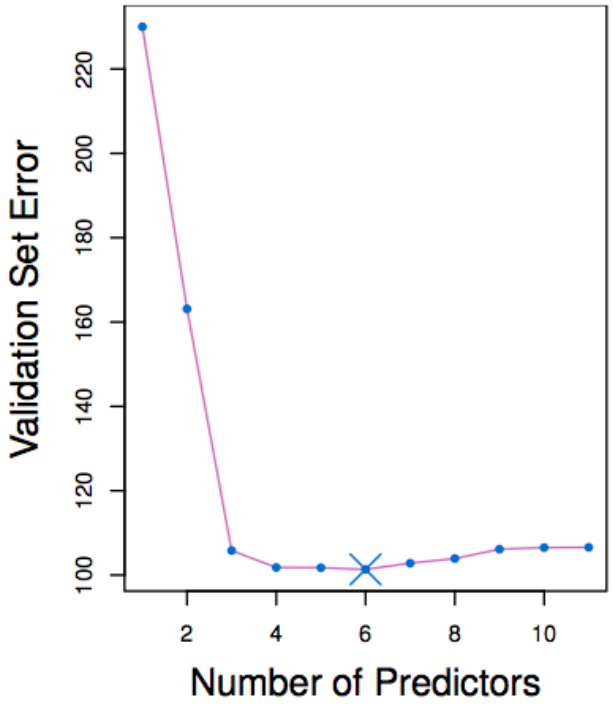
1. Noter $\widehat{\mathrm{m}}_{\mathrm{p}}$ le **modèle complet**, qui contient tous les p prédicteurs
2. Pour chaque valeur de k allant de p à 1 :
   1. Considérer tous les k modèles qui consistent à supprimer un prédicteur à $\widehat{\mathrm{m}}_{\mathrm{k}}$.
   2. Choisir le meilleur parmi ces k modèles et noter le $\widehat{\mathrm{m}}_{\mathrm{k}-1}$. Le *meilleur* modèle est celui qui minimise le critére des moindres carrés.
3. Choisir le meilleur modèle parmi $\widehat{\mathrm{m}}_0, \widehat{\mathrm{m}}_1, \ldots \widehat{\mathrm{m}}_{\mathrm{p}}$ en utilisant la validation croisée, AIC, ou BIC.

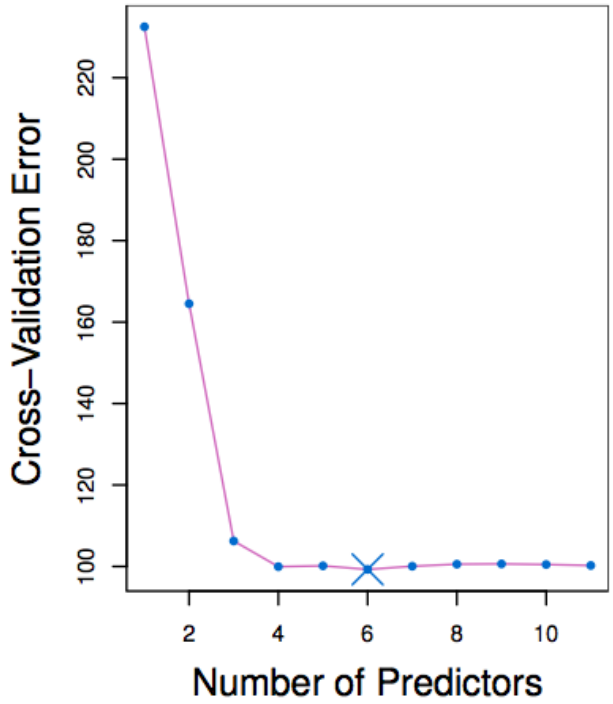
Sélection pas à pas rétrograde (suite)

- Comme la méthode progressive, la méthode rétrograde ne visite que $1 + p(p+1)/2$ modèles, et peut donc être appliquée dans des contextes où p est trop grand pour la méthode exhaustive.
- Comme la méthode progressive, la méthode rétrograde ne garantit pas de trouver le *meilleur* modèle.
- La méthode rétrograde suppose que *la taille de l'échantillon n est plus grande que le nombre de prédicteurs p* (pour pouvoir ajuster le modèle complet). En revanche, la méthode progressive peut s'arrêter à n covariables si $p > n$ et peut donc être utilisée dans un contexte plus large.

Exemple : jeu de données crédit







Commentaires

- L'erreur par validation a été estimée en mettant de côté un quart du jeu de données (tiré au hasard) pour valider. Les trois quarts restants servant à entrainer les modèles
- L'erreur de validation croisée a été calculée par une méthode à $\mathrm{k} = 10$ blocs. Dans ce cas, ces deux méthodes renvoient un modèle à 6 variables (de dimension 7, pourquoi ?).
- Cependant, les trois approches suggèrent que les modèles à 4, 5 ou 6 variables sont à peu près équivalents en terme d'erreur de test.

Sommaire

1 Introduction : Apprentissage Statistique
- Formalime de l'apprentissage supervisé
- Fléau de la dimension
- Modèles Linéaires

2 Sélection de variables par critères AIC/BIC et validation Croisée

3 Lasso
- Préambule : Ridge
- Définition du Lasso
- Lasso et Grande dimension
- Aspects Algorithmiques

4 Extensions du Lasso
- Lasso et Modèles linéaires Généralisés
- Lasso et Estimation Structurée

Méthode de Régularisation

Régression ridge et *Lasso*

- Les méthodes précédentes de choix de sous-ensembles utilisent les moindres carrés pour ajuster chacun des modèles en compétition.
- Alternativement, on peut ajuster un modèle contenant toutes les p covariables en utilisant une technique qui *contraint* ou *régularise* les estimations des coefficients, ou de façon équivalente, pousse les coefficients vers 0.

Préambule

Les variables explicatives observées $x^{(j)} = (x_1^{(j)}, \ldots, x_n^{(j)})$ sont centrées et standardisées (ie $\|x^{(j)}\|_2^2/n = 1$) et on suppose que $\beta_0^* = 0$ et $\overline{Y} = 0$ ce qui revient à estimer β_0^* par $\overline{Y}$ et à remplacer Y_i par $Y_i - \overline{Y}$.

Remarque : Une fois les paramètres ajustés pour les variables centrées et standardisées, on peut facilement revenir au modèles initial en transformer linéairement les paramètres.

Pénalisation l_q

Un estimateur par minimisation du risque empirique régularisé (pour la perte quadratique) est dans le cadre de la régression linéaire défini par

$$\widehat{\beta}_\lambda \in \arg\min_{\beta \in \mathbb{R}^p} \|\mathbf{Y} - \mathbf{X}\beta\|_2^2 + \lambda \|\beta\|_q^q$$

λ étant un paramétre positif, appelé paramètre de régularisation.

- $q = 2 \rightsquigarrow$ régression ridge
- $q = 1 \rightsquigarrow$ régression lasso

Régression ridge

L'estimateur est défini par

$$\widehat{\beta}_{\lambda}^{\text{ridge}} \in \arg\min_{\beta \in \mathbb{R}^p} \|\mathbf{Y} - \mathbf{X}\beta\|_2^2 + \lambda \|\beta\|_2^2$$

Proposition

- *Minimiser* $\|\mathbf{Y} - \mathbf{X}\beta\|_2^2 + \lambda\|\beta\|_2^2$ *en* $\beta \in \mathbb{R}^p$ *est équivalent à minimiser* $\|\mathbf{Y} - \mathbf{X}\beta\|_2^2$ *sous une contrainte de la forme* $\|\beta\|_2^2 \leqslant r(\lambda)$.
- *La matrice* $(\mathbf{X}^T\mathbf{X} + \lambda\mathbf{I})$ *est toujours définie positive, donc inversible et* $\widehat{\beta}_{\lambda}^{\text{ridge}} = (\mathbf{X}^T\mathbf{X} + \lambda\mathbf{I})^{-1}\mathbf{X}^T\mathbf{Y}$.

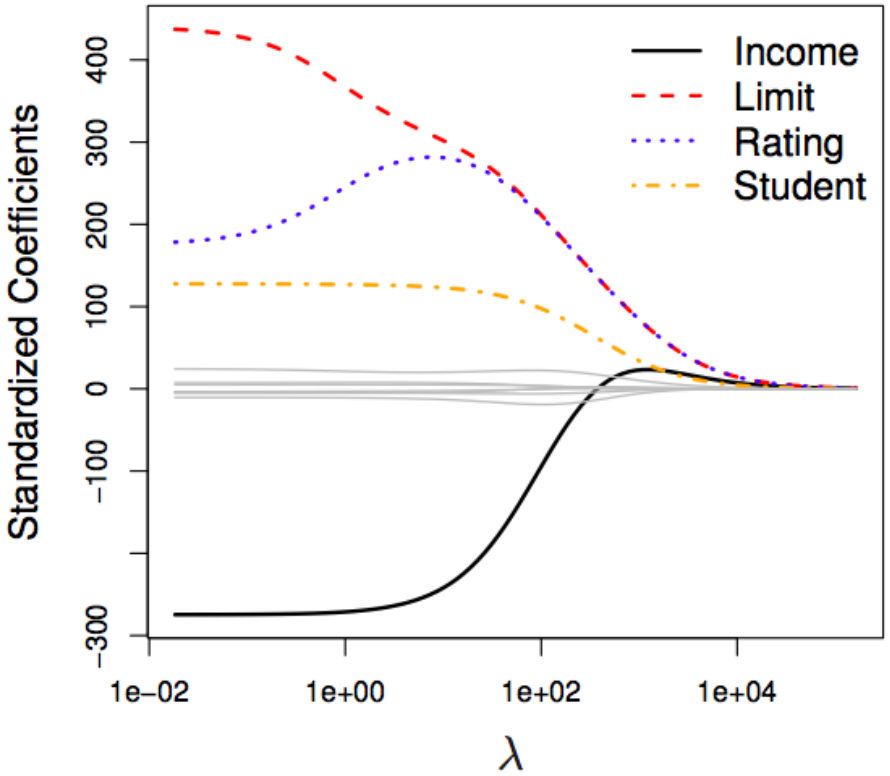
Remarque : L'estimateur $\widehat{\beta}_{\lambda}^{\text{ridge}}$ est biaisé mais sa variance est plus faible que celle de l'estimateur des moindres carrés.

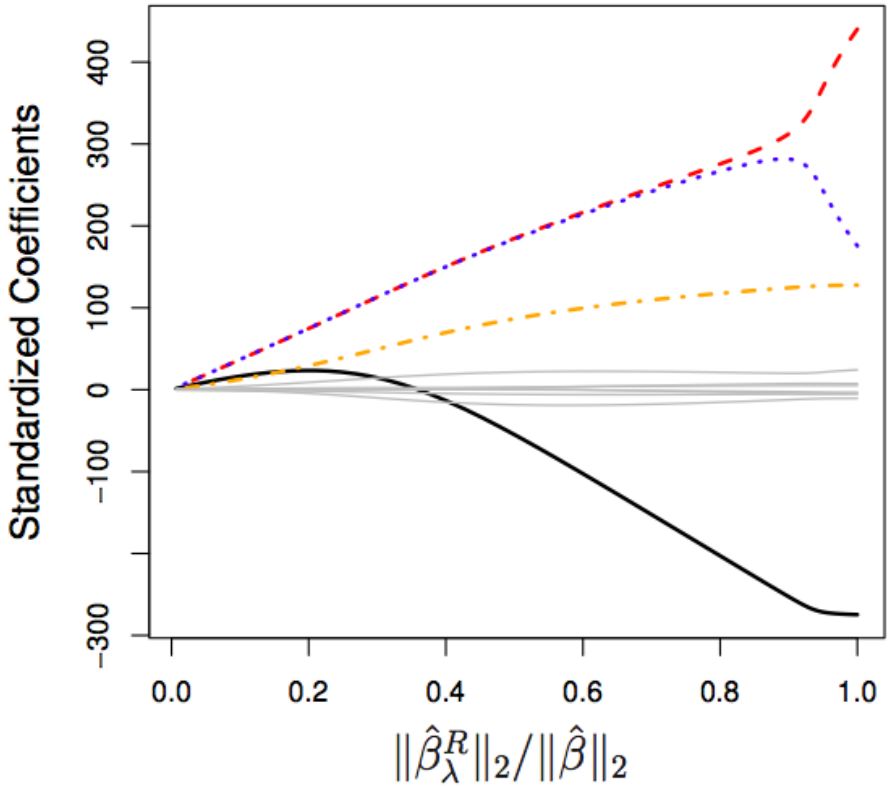
Rôle et ajustement du paramétre de régularisation

- Lorque $\lambda = 0$, $\widehat{\beta}_\lambda^{\text{ridge}}$ est l'estimateur des moindres carrés.
- Lorque $\lambda \to \infty$, $\widehat{\beta}_\lambda^{\text{ridge}}$ tend vers 0
- Lorque λ augmente, le biais de $\widehat{\beta}_\lambda^{\text{ridge}}$ a tendance à augmenter et la variance à diminuer $\Rightarrow$ Recherche d'un compromis

$\rightsquigarrow$ Choix usuel de λ par validation croisée V fold sur une grille finie de valeur de $\lambda > 0$.

Exemple jeu de données crédit



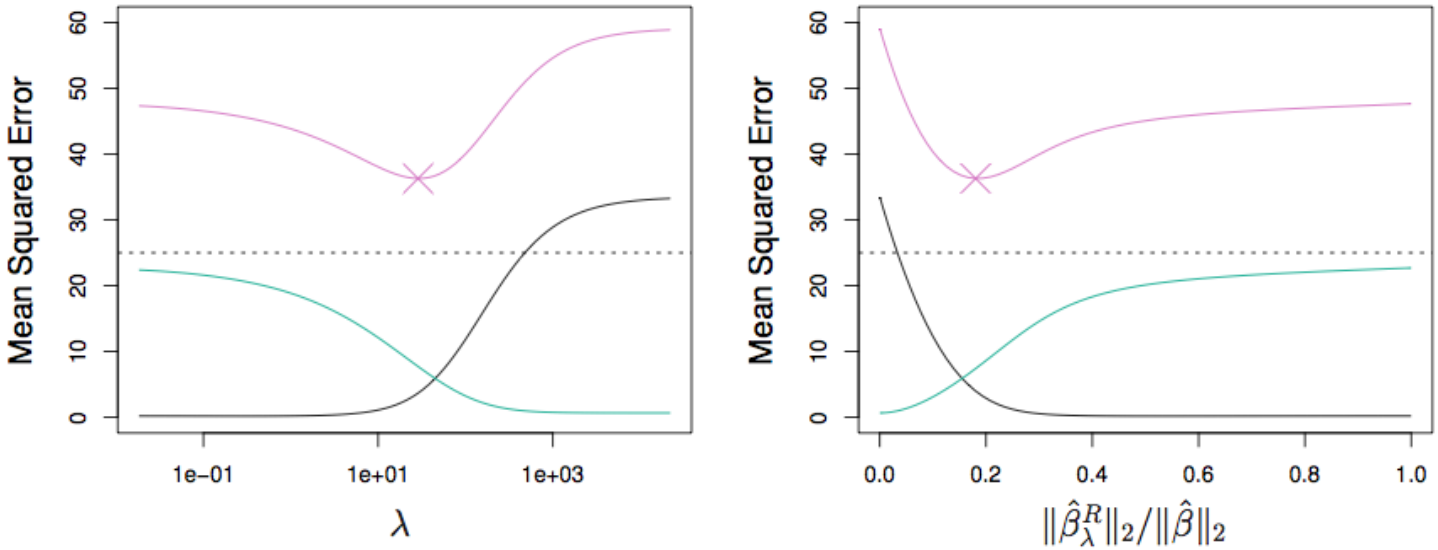


Commentaires

- À gauche, chaque courbe correspond à l'estimation des coefficients par régression ridge pour l'une des 10 variables, représentée en fonction de λ.
- À droite, l'axe des abscisses est maintenant le rapport entre la norme quadratique des coefficients estimés par régression ridge et les coefficients estimés par moindres carrés.

Pour la régression ridge ?

Compromis biais-variance



Données simulées : $\mathrm{n} = 50$, $\mathrm{p} = 45$, tous de coefficients non nuls. Biais au carré (en noir), variance (en vert) et erreur de test quadratique (en violet) pour la régression ridge.
Droite horizontale : erreur minimale.

Sommaire

1 Introduction : Apprentissage Statistique
- Formalime de l'apprentissage supervisé
- Fléau de la dimension
- Modèles Linéaires

2 Sélection de variables par critères AIC/BIC et validation Croisée

3 Lasso
- Préambule : Ridge
- Définition du Lasso
- Lasso et Grande dimension
- Aspects Algorithmiques

4 Extensions du Lasso
- Lasso et Modèles linéaires Généralisés
- Lasso et Estimation Structurée

La Régression Lasso

- La régression ridge a un inconvénient évident : contrairement à la sélection de variable, la régression ridge inclut tous les prédicteurs dans le modèle final.

L'estimateur LASSO (Least Absolute Selection and Shrinkage Operator) est défini pour $\lambda > 0$ par

$$\widehat{\beta}_{\lambda}^{\text{lasso}} \in \arg\min_{\beta \in \mathbb{R}^p} \|\mathbf{Y} - \mathbf{X}\beta\|_2^2 + \lambda \|\beta\|_1$$

La fonction $\mathscr{L} : \beta \mapsto \|\mathbf{Y} - \mathbf{X}\beta\|_2^2 + \lambda\|\beta\|_1$ est convexe, non différentiable. La solution du problème peut ne pas être unique.

Proposition

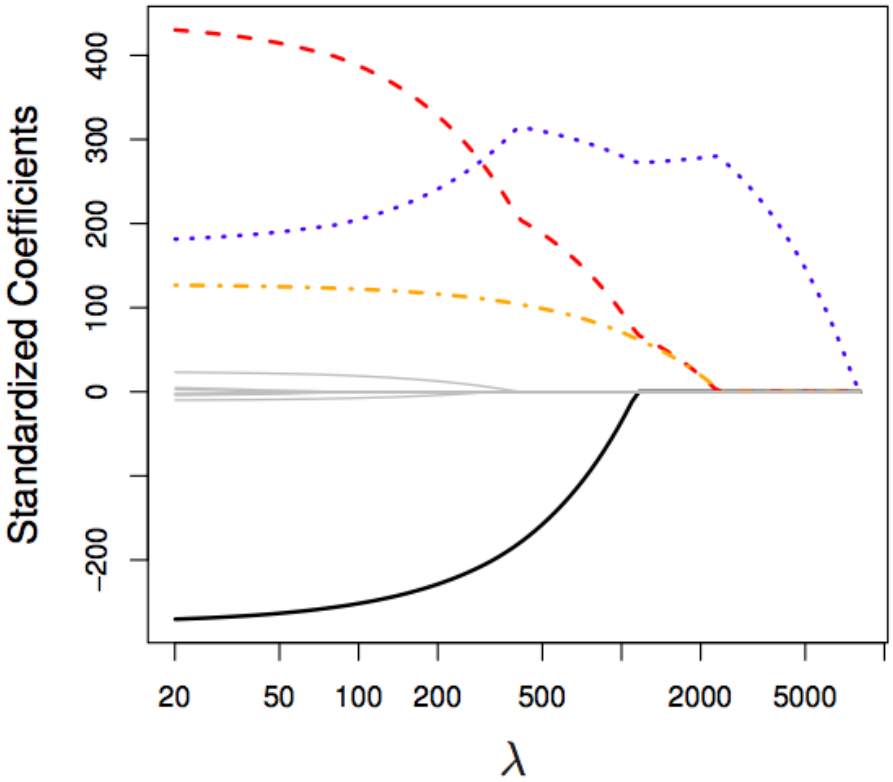
Minimiser $\|\mathbf{Y} - \mathbf{X}\beta\|_2^2 + \lambda\|\beta\|_1$ *en* $\beta \in \mathbb{R}^p$ *est équivalent à minimiser* $\|\mathbf{Y} - \mathbf{X}\beta\|_2^2$ *sous une contrainte de la forme* $\|\beta\|_1 \leqslant R_\lambda$ *pour une certaine quantité* R_λ.

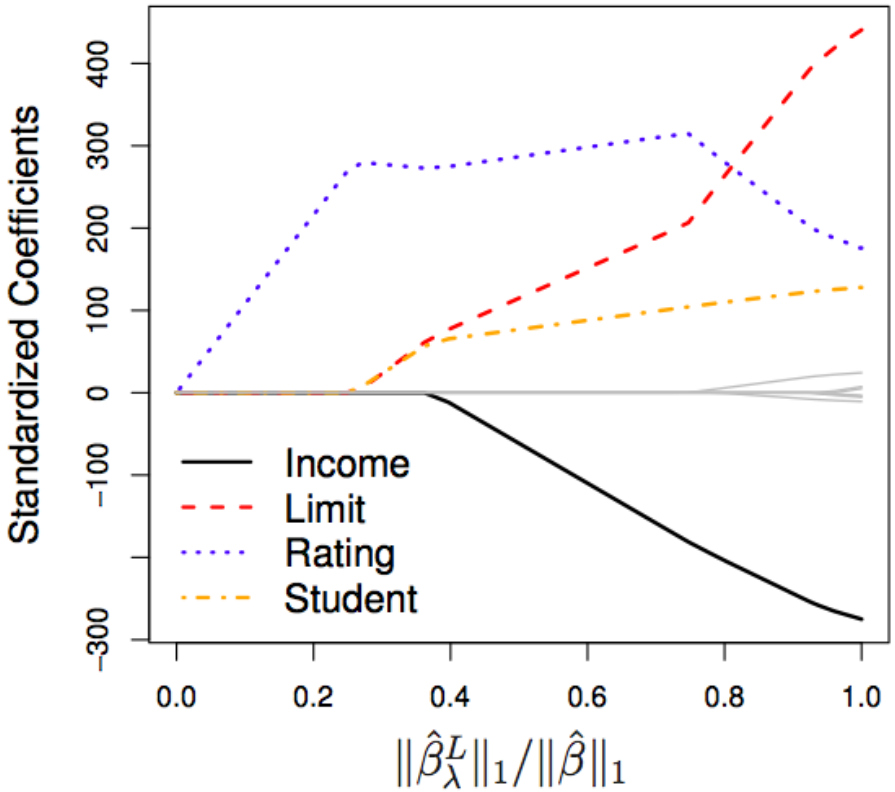
Preuve : Lagrangien

Le Lasso (suite)

- Comme pour la régression ridge, le Lasso tire les estimations des coefficients vers 0.
- Cependant, dans le cas du Lasso, la pénalité ℓ^1 a pour effet de forcer certains coefficients à s'annuler lorsque λ est suffisamment grand.
- Donc, le Lasso permet de faire de la *sélection de variable*.
- On parle de modèle creux (sparse), c'est-à-dire de modèles qui n'impliquent qu'un sous ensemble des variables.
- Comme pour la régression ridge, choisir une bonne valeur de λ est critique. Procéder par validation croisée.

Exemple : jeu de données crédit





Qu'est qui fait marcher le Lasso ?

Avec les multiplicateurs de Lagrange, on peut voir

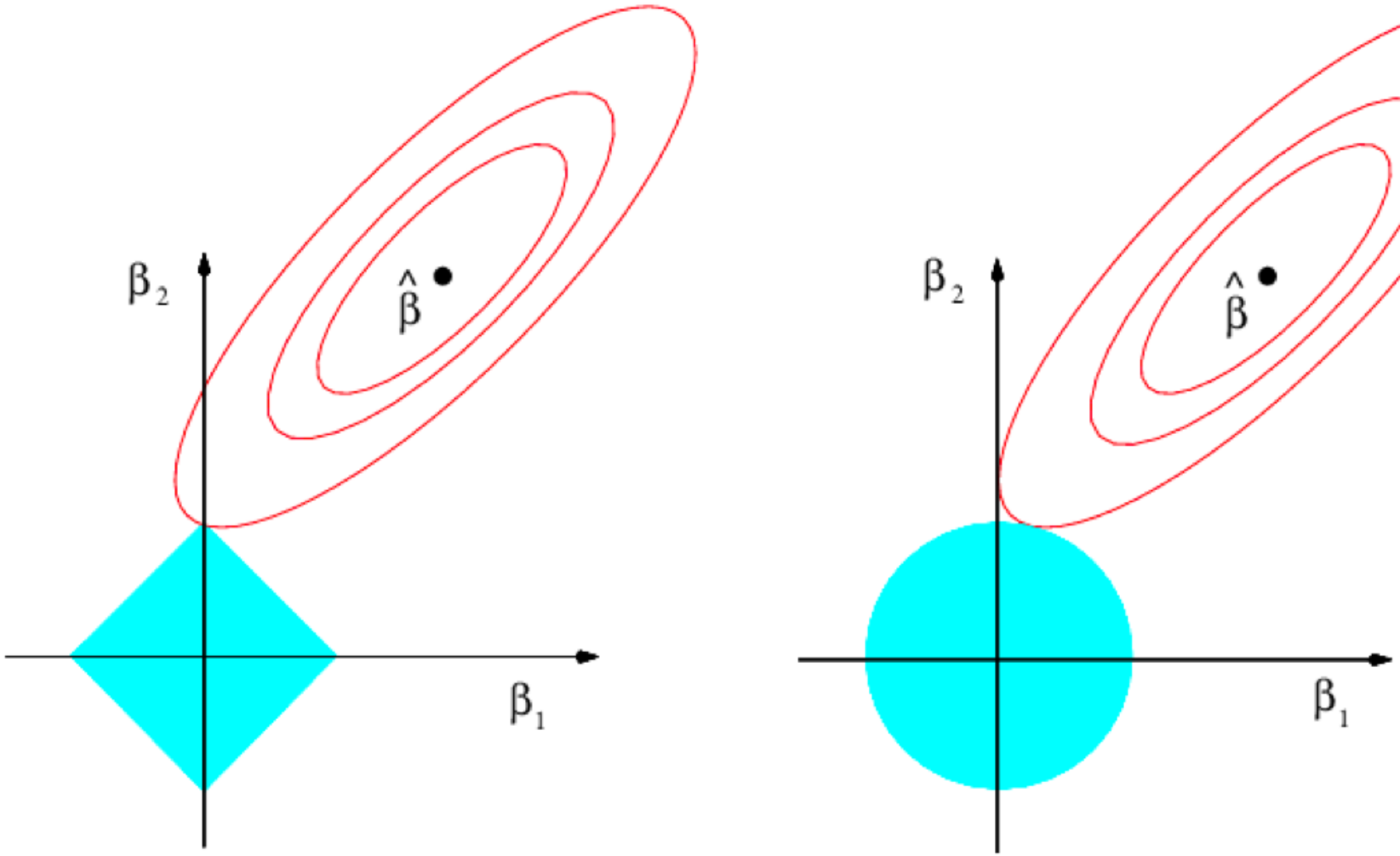
- La régression ridge comme

$$\text{minimise } \sum_{i=1}^{n}\left(Y_i - \sum_{j=1}^{p}\beta_j X_i^{(j)}\right)^2 \text{ sous la contrainte } \sum_{j=1}^{p}\beta_j^2 \leqslant s$$

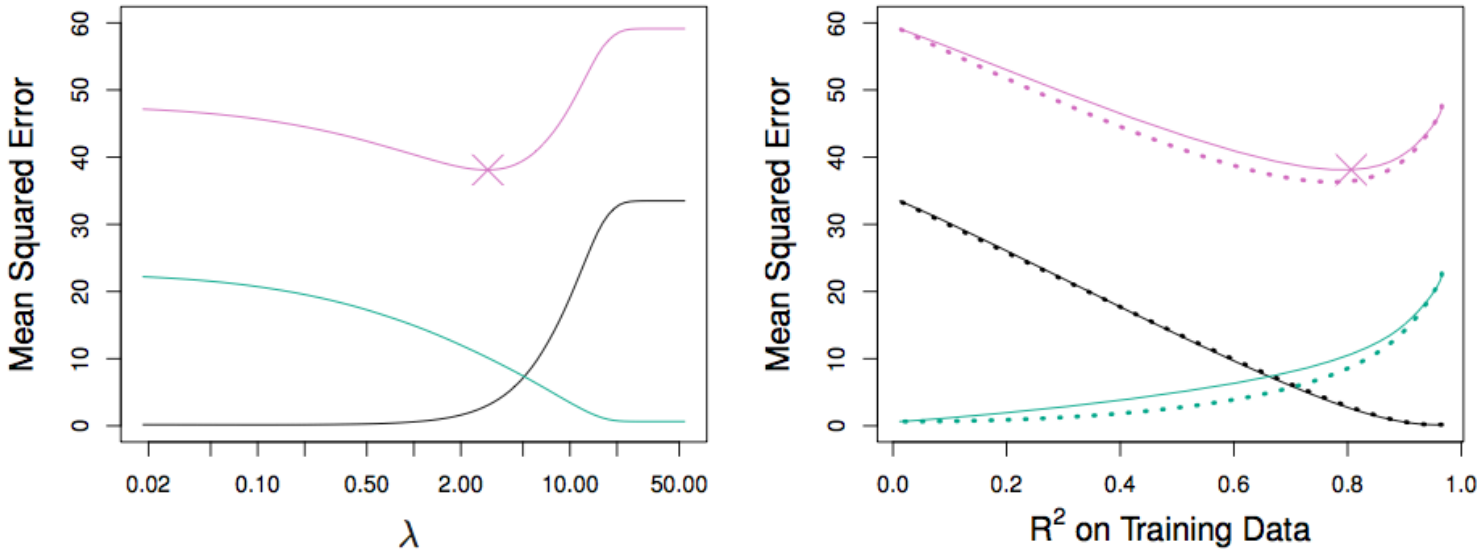
- Le Lasso comme

$$\text{minimise } \sum_{i=1}^{n}\left(Y_i - \sum_{j=1}^{p}\beta_j X_i^{(j)}\right)^2 \text{ sous la contrainte } \sum_{j=1}^{p}\left|\beta_j\right| \leqslant s$$

Le Lasso en image



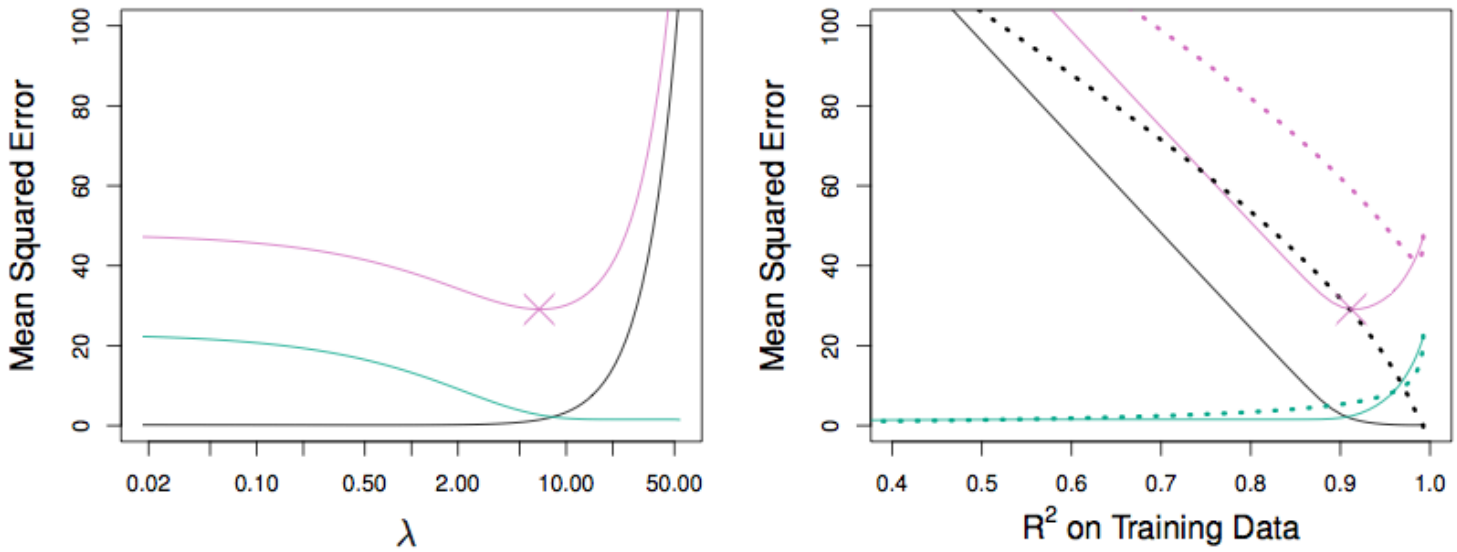
Comparaison du Lasso et de la régression ridge



À gauche, biais au carré (noir), variance (en vert) et erreur quadratique de test (violet) pour le Lasso sur données simulées.

À droite, comparaison du biais au carré, de la variance et de l'erreur de test quadratique pour le Lasso (traits plains) et la régression ridge (pointillés)

Comparaison du Lasso et de la régression ridge (suite)



À gauche, biais au carré (noir), variance (en vert) et erreur quadratique de test (violet) pour le Lasso sur données simulées (où seulement deux prédicteurs sont influents).

À droite, comparaison du biais au carré, de la variance et de l'erreur de test quadratique pour le Lasso (traits plains) et la régression ridge (pointillés)

Conclusions

- Ces deux exemples montrent qu'il n'y a pas de meilleur choix universel entre la régression ridge et le Lasso.
- En général, on s'attend à ce que le Lasso se comporte mieux lorsque la réponse est une fonction d'un nombre relativement faible de prédicteurs.
- Cependant, le nombre de prédicteurs reliés à la réponse n'est jamais connu *a priori* dans des cas concrets.
- Une technique comme la validation croisée permet de déterminer quelle est la meilleure approche.

Sommaire

1. Introduction : Apprentissage Statistique
   - Formalime de l'apprentissage supervisé
   - Fléau de la dimension
   - Modèles Linéaires
2. Sélection de variables par critères AIC/BIC et validation Croisée
3. Lasso
   - Préambule : Ridge
   - Définition du Lasso
   - Lasso et Grande dimension
   - Aspects Algorithmiques
4. Extensions du Lasso
   - Lasso et Modèles linéaires Généralisés
   - Lasso et Estimation Structurée

Bornes de risque

Notation $x^{(k)}$: k-ième colonne de $\mathbf{X}$. $\|\beta\|_0$ = Nbre de composantes non nulles de β.

Supposons que le modèle linéaire est vrai et que les erreurs sont gaussiennes :

$$\mathbf{Y} = \mathbf{X}\beta^* + \varepsilon \text{ avec } \varepsilon \sim \mathcal{N}(0, \sigma^2 \mathbf{I}_n)$$

Proposition (Koltchinski et al.(2010))

Pour $\lambda = 6\sigma\sqrt{\frac{2}{n}\log(p)}$ *avec probabilité au moins égale* $1 - 1/p$,

$$\|\mathbf{X}(\widehat{\beta}_\lambda^{Lasso} - \beta^*)\|_2^2 \leqslant \inf_{\beta \neq 0} \left\{ \|\mathbf{X}(\beta - \beta^*)\|_2^2 + \frac{36\sigma^2 \log(p)}{\kappa^2(\beta)} \sum_{j=1}^{p} \|\beta\|_0 \right\}$$

$\kappa(\beta)$ *est une constante de compatibilité : mesure le manque d'orthogonalité des colonnes de* $\mathbf{X}$.

Cas particulier : Si $\|\beta^*\|_0$ (nombre de composantes non nuls) vaut k alors (sous des conditions sur le design)

$$\|\mathbf{X}(\widehat{\beta}_\lambda^{\text{Lasso}} - \beta^*)\|_2^2 \leqslant c' \sigma^2 k \log(p)$$

Remarques :

- Si nous connaissions *à l'avance* les composantes non nulles de β^*, l'erreur de l'estimateurs moindres carrés restreint à ce modèle serait de $\sigma^2 k$.
- Le Lasso fait presque aussi bien que cet estimateur (oracle) en ne perdant qu'un facteur $\log(p)$.

Grande dimension et très grande dimension

- (sous des conditions sur le design), le resultat précédent est valable pour $p \gg n$.
  $\rightsquigarrow$ Estimer précisément β^* est possible en grande dimension ($p \gg n$) sous des hypothèses de parcimonie.
- En revanche si p est vraiment trop grand ou si β^* n'est pas parcimonieux, c'est si dire si

$$2\|\beta^*\|_0\left[1 + \log\left(\frac{p}{\|\beta^*\|_0}\right)\right] \geqslant n$$

alors, **aucune procédure statistique ne peut estimer précisément** β^*
(à moins d'avoir d'autres informations à priori sur β^*)

Sommaire

1 Introduction : Apprentissage Statistique
- Formalime de l'apprentissage supervisé
- Fléau de la dimension
- Modèles Linéaires

2 Sélection de variables par critères AIC/BIC et validation Croisée

3 Lasso
- Préambule : Ridge
- Définition du Lasso
- Lasso et Grande dimension
- Aspects Algorithmiques

4 Extensions du Lasso
- Lasso et Modèles linéaires Généralisés
- Lasso et Estimation Structurée

Coordinate descent

Proposition (Condition d'optimalité du premier ordre)

$\widehat{\beta}_\lambda^{lasso}$ *vérifie* $X^T X \widehat{\beta}_\lambda^{lasso} = X^T Y - \lambda \widehat{Z}/2$ *avec* $\widehat{Z}_j \in [-1,1]$ *et* $\widehat{Z}_j = signe([\widehat{\beta}_\lambda^{lasso}]_j)$ *si* $[\widehat{\beta}_\lambda^{lasso}]_j \neq 0$.

Pas de solution explicite.

Une approche pour calculer l'estimateur lasso : la descente par coordonnées.

Proposition

La fonction $\beta_j \mapsto \|Y - X\beta\|_2^2 + \lambda\|\beta\|_1$ *est minimum en* $\beta_j = R_j(1 - \lambda/(2|R_j|))_+/n$ *avec* $R_j = (x^{(j)})^T(Y - \sum_{k\neq j} \beta_k x^{(k)})$.

Exercice : le prouver.

Cyclic Coordinate descent pour le lasso

Algorithme de Coordinate descent

input : $X, Y, \lambda > 0$

begin

Initialiser $\beta = \beta_{in}$;

while β *n'a pas convergé* **do**

for $j = 1, \ldots, p$ **do**

Calculer $R_j = (x^{(j)})^T (Y - \sum_{k \neq j} \beta_k x^{(k)})$;

Calculer $\beta_j = R_j (1 - \lambda / (2|R_j|))_+ / n$

end

end

output: β

end

Sommaire

1. Introduction : Apprentissage Statistique
   - Formalime de l'apprentissage supervisé
   - Fléau de la dimension
   - Modèles Linéaires
2. Sélection de variables par critères AIC/BIC et validation Croisée
3. Lasso
   - Préambule : Ridge
   - Définition du Lasso
   - Lasso et Grande dimension
   - Aspects Algorithmiques
4. Extensions du Lasso
   - Lasso et Modèles linéaires Généralisés
   - Lasso et Estimation Structurée

Sommaire

1. Introduction : Apprentissage Statistique
   - Formalime de l'apprentissage supervisé
   - Fléau de la dimension
   - Modèles Linéaires
2. Sélection de variables par critères AIC/BIC et validation Croisée
3. Lasso
   - Préambule : Ridge
   - Définition du Lasso
   - Lasso et Grande dimension
   - Aspects Algorithmiques
4. Extensions du Lasso
   - Lasso et Modèles linéaires Généralisés
   - Lasso et Estimation Structurée

Jusqu'ici on utilisait l'information a priori que β^* était potentiellement parcimonieux (la plupart de ces coefficients sont nuls).

En modifiant la pénalité $\|\beta\|_1$, on peut prendre en compte d'autres informations a priori.

Group Lasso

Cadre : Les variables $X^{(j)}, j = 1, \ldots, p$ sont partitionnées en q groupes $G_1, \ldots, G_q$ de variables cohérentes.

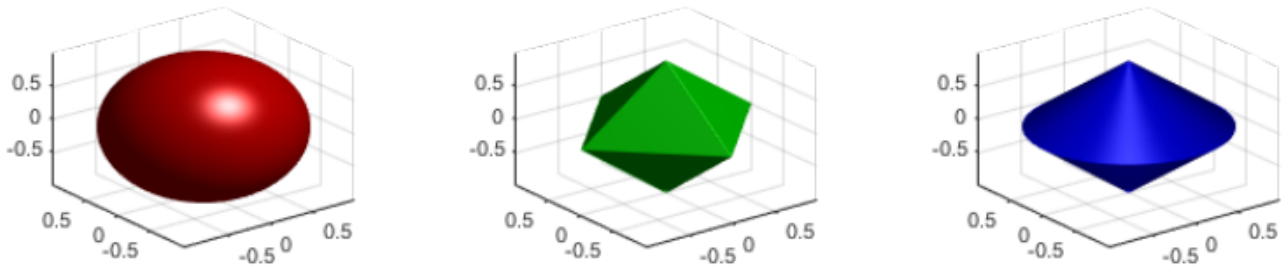
Exemple pour des IRMf : Groupes G_1, ... G_q correspondent à des petite régions du cerveaux.

Hypothèse structurelle : Les coefficients β_i^* dans un même groupe G_j sont soit tous nul, soit tous nos nuls.

Estimateur Group Lasso :

$$\widehat{\beta} \in \arg\min_{\beta} \|Y - X\beta\|_2^2 + \lambda \sum_{j=1}^{q} \|\beta_{G_j}\|_2 ,$$

où β_{G_j} est la restriction de β aux variables dans G_j.



(a) ℓ_2 norm (b) ℓ_1 norm (c) $\ell_{2,1}$ norm

Figure 3. Isosurface for three different regularization terms, with $\mu_\lambda = 1$. (a) Standard squared ℓ_2 norm. (b) ℓ_1 norm enforcing sparsity. (c) $\ell_{2,1}$ norm applied to the groups $\{1, 2\}$ and $\{3\}$ (without considering the scaling factors).

Fused Lasso

Cadre : Les variables $X^{(j)}, j = 1, \ldots, p$ sont naturellement ordonnées.

Exemple une série temporelle ($p = n$) :

$$Y_t = \beta_t^* + \varepsilon_t, \quad t = 1, \ldots, n$$

Hypothèse structurelle : Les coefficients β_t^* sont constants par morceaux
$\rightsquigarrow$ i.e. $\beta_t^* - \beta_{t-1}^*$ est souvent nul.

Estimateur Fused Lasso :

$$\widehat{\beta} \in \arg\min_{\beta} \|Y - X\beta\|_2^2 + \lambda \sum_{j=1}^{p} \|\beta_i - \beta_{i+1}\|_1 ,$$

Bilan

- Les méthodes de sélection de modèles sont essentielles pour l'analyse de données, et l'apprentissage statistique, en particulier avec de gros jeu de données contenant de nombreux prédicteurs.
- En grande dimension, la prise en compte de structure (parcimonie, parcimonie par groupe,...) dans les paramètres joue un rôle central.