

Miser sur la sparsité

Sahir Rai Bhatnagar

Department of Epidemiology, Biostatistics, and Occupational Health
Department of Diagnostic Radiology

<https://sahirbhatnagar.com/>

SÉSÀME
15 mars 2019



Aperçu

- Les modèles classiques
- Miser sur la sparsité
- Un exemple justificatif
- Contexte de la méthode lasso
- Extensions

Les modèles classiques

Les modèles classiques

- Une belle et puissante boîte à outils pour analyser des jeux de données où la taille de l'échantillon est beaucoup plus grande que le nombre de variable explicative ($n > p$).
- ex: Modèle linéaire généralisée (GLM), modèle additif généralisé (GAM), analyse discriminante linéaire (LDA), support vector machines (SVM).

$$\mathbf{X}_{n \times p} = \begin{bmatrix} X_{11} & X_{12} & \cdots & X_{1p} \\ X_{21} & X_{12} & \cdots & X_{1p} \\ X_{31} & X_{12} & \cdots & X_{1p} \\ \vdots & \vdots & \vdots & \vdots \\ X_{n1} & X_{12} & \cdots & X_{np} \end{bmatrix}$$

	Sepal.Length	Sepal.Width	Petal.Length	Petal.Width	Species
1	5.1	3.5	1.4	0.2	setosa
2	4.9	3.0	1.4	0.2	setosa
3	4.7	3.2	1.3	0.2	setosa
4	4.6	3.1	1.5	0.2	setosa
5	5.0	3.6	1.4	0.2	setosa
6	5.4	3.9	1.7	0.4	setosa
7	4.6	3.4	1.4	0.3	setosa
8	5.0	3.4	1.5	0.2	setosa
9	4.4	2.9	1.4	0.2	setosa
10	4.9	3.1	1.5	0.1	setosa
11	5.4	3.7	1.5	0.2	setosa
12	4.8	3.4	1.6	0.2	setosa
13	4.8	3.0	1.4	0.1	setosa
14	4.3	3.0	1.1	0.1	setosa
15	5.8	4.0	1.2	0.2	setosa

Modèle linéaire

- Données: $(\mathbf{x}_1, y_1), \dots, (\mathbf{x}_n, y_n)$ iid de

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \varepsilon$$

$$E(\varepsilon | \mathbf{X}) = 0 \text{ et } \dim(\mathbf{X}) = p$$

- L'estimateur des moindres carrés ordinaires:

$$\hat{\boldsymbol{\beta}} = \arg \min_{\boldsymbol{\beta}} \|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|^2$$

$$\hat{\boldsymbol{\beta}}_{MCO} = (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{y}$$

Modèle linéaire

- Données: $(\mathbf{x}_1, y_1), \dots, (\mathbf{x}_n, y_n)$ iid de

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \varepsilon$$

$$E(\varepsilon | \mathbf{X}) = 0 \text{ et } \dim(\mathbf{X}) = p$$

- L'estimateur des moindres carrés ordinaires:

$$\hat{\boldsymbol{\beta}} = \arg \min_{\boldsymbol{\beta}} \|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|^2$$

$$\hat{\boldsymbol{\beta}}_{MCO} = (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{y}$$

- Comment trouver les variables (\mathbf{x}_j) pertinentes?

Meilleure sélection de sous-ensembles (Beal et al. 1967, Biometrika)

- $\widehat{\boldsymbol{\beta}} = \arg \min_{\boldsymbol{\beta}} \|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|^2 + \lambda \|\boldsymbol{\beta}\|_0$
- $\|\boldsymbol{\beta}\|_0 = \sum_{j=1}^p \mathbf{1}\{\beta_j \neq 0\}$
- $\lambda \geq 0$ est un paramètre de réglage qui contrôle la taille du modèle
- Le calcul de tous les modèles de sous-ensembles possibles est un problème d'optimisation combinatoire (NP difficile)
- Beaucoup d'instabilité dans le processus de sélection (Breiman, 1996)

Meilleure sélection de sous-ensembles (Beal et al. 1967, Biometrika)

Predictor set	model
None of $x_1 x_2 x_3 x_4$	$E(Y) = \beta_0$
x_1	$E(Y) = \beta_0 + \beta_1 x_1$
x_2	$E(Y) = \beta_0 + \beta_2 x_2$
x_3	$E(Y) = \beta_0 + \beta_3 x_3$
x_4	$E(Y) = \beta_0 + \beta_4 x_4$
$x_1 x_2$	$E(Y) = \beta_0 + \beta_1 x_1 + \beta_2 x_2$
$x_1 x_3$	$E(Y) = \beta_0 + \beta_1 x_1 + \beta_3 x_3$
$x_1 x_4$	$E(Y) = \beta_0 + \beta_1 x_1 + \beta_4 x_4$
$x_2 x_3$	$E(Y) = \beta_0 + \beta_2 x_2 + \beta_3 x_3$
$x_2 x_4$	$E(Y) = \beta_0 + \beta_2 x_2 + \beta_4 x_4$
$x_3 x_4$	$E(Y) = \beta_0 + \beta_3 x_3 + \beta_4 x_4$
$x_1 x_2 x_3$	$E(Y) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3$
$x_1 x_2 x_4$	$E(Y) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_4 x_4$
$x_1 x_3 x_4$	$E(Y) = \beta_0 + \beta_1 x_1 + \beta_3 x_3 + \beta_4 x_4$
$x_2 x_3 x_4$	$E(Y) = \beta_0 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4$

Régression ridge

(Hoerl & Kennard 1970, Technometrics)

- $\hat{\boldsymbol{\beta}} = \arg \min_{\boldsymbol{\beta}} \|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|^2 + \lambda \|\boldsymbol{\beta}\|_2^2$
- $\|\boldsymbol{\beta}\|_2^2 = \sum_{j=1}^p \beta_j^2$
- $\hat{\boldsymbol{\beta}}_{Ridge} = (\mathbf{X}^\top \mathbf{X} + \lambda \mathbf{I})^{-1} \mathbf{X}^\top \mathbf{y} \rightarrow \text{solution exacte}$

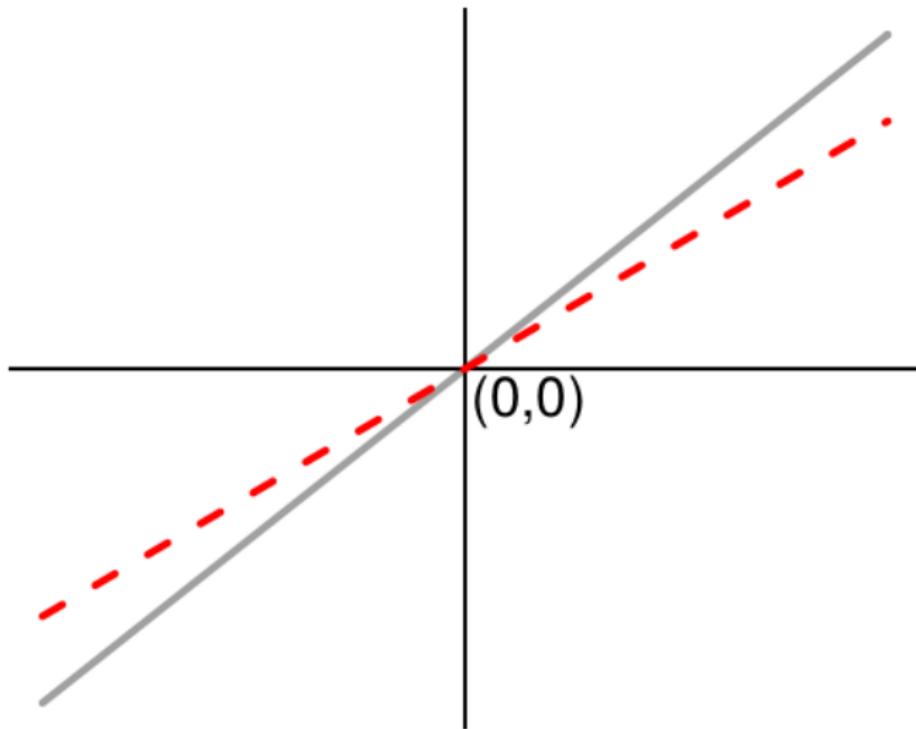
Régression ridge

(Hoerl & Kennard 1970, Technometrics)

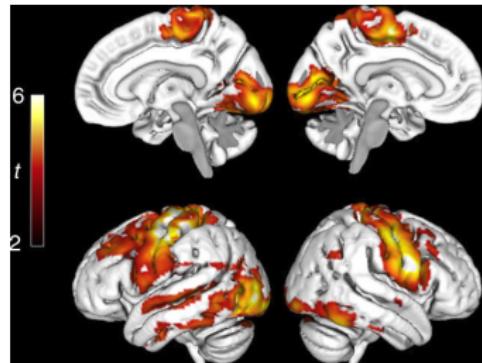
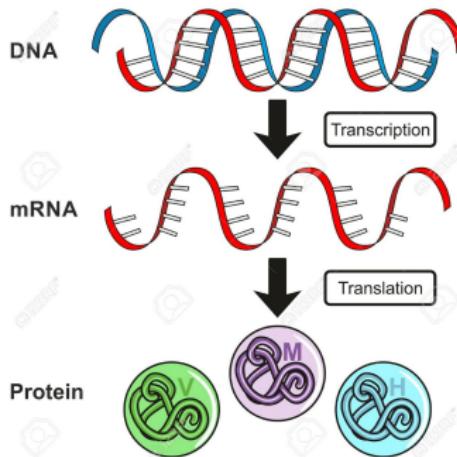
- $\hat{\beta} = \arg \min_{\beta} \|\mathbf{y} - \mathbf{X}\beta\|^2 + \lambda \|\beta\|_2^2$
- $\|\beta\|_2^2 = \sum_{j=1}^p \beta_j^2$
- $\hat{\beta}_{Ridge} = (\mathbf{X}^\top \mathbf{X} + \lambda \mathbf{I})^{-1} \mathbf{X}^\top \mathbf{y} \rightarrow \text{solution exacte}$
- $\hat{\beta}_{MCO} = (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{y}$
- Soit $\mathbf{X}^\top \mathbf{X} = \mathbf{I}_{p \times p}$

$$\hat{\beta}_{j(Ridge)} = \frac{\hat{\beta}_{j(MCO)}}{1 + \lambda}$$

Ridge



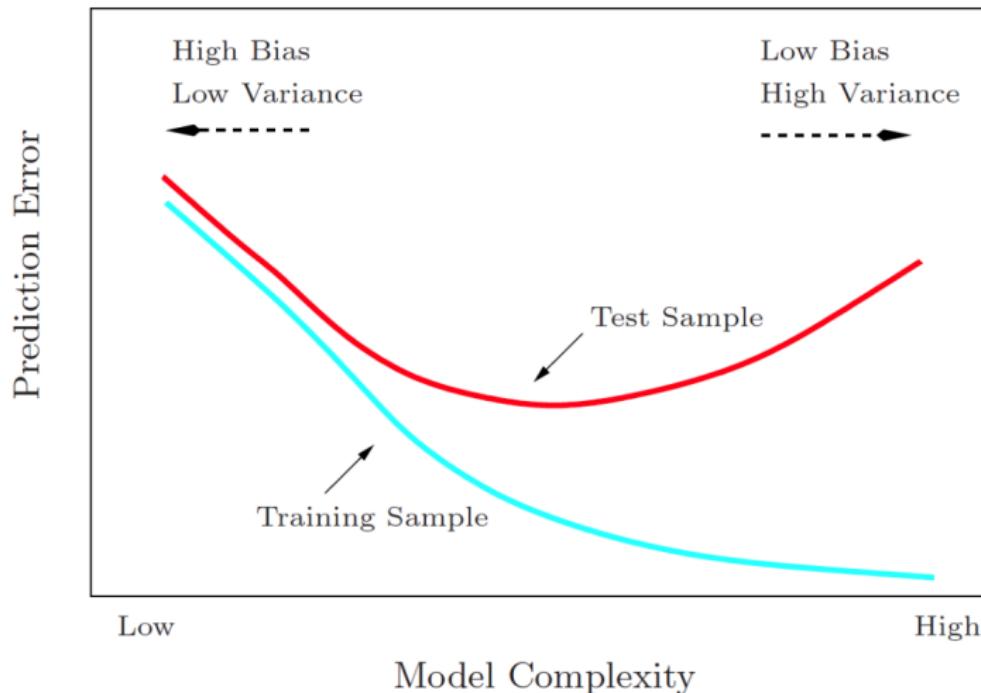
Les données de grande dimension ($n \ll p$)



$$\mathbf{X}_{n \times p} = \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1p} \\ \vdots & \vdots \\ x_{n1} & x_{12} & \cdots & x_{np} \end{bmatrix}$$

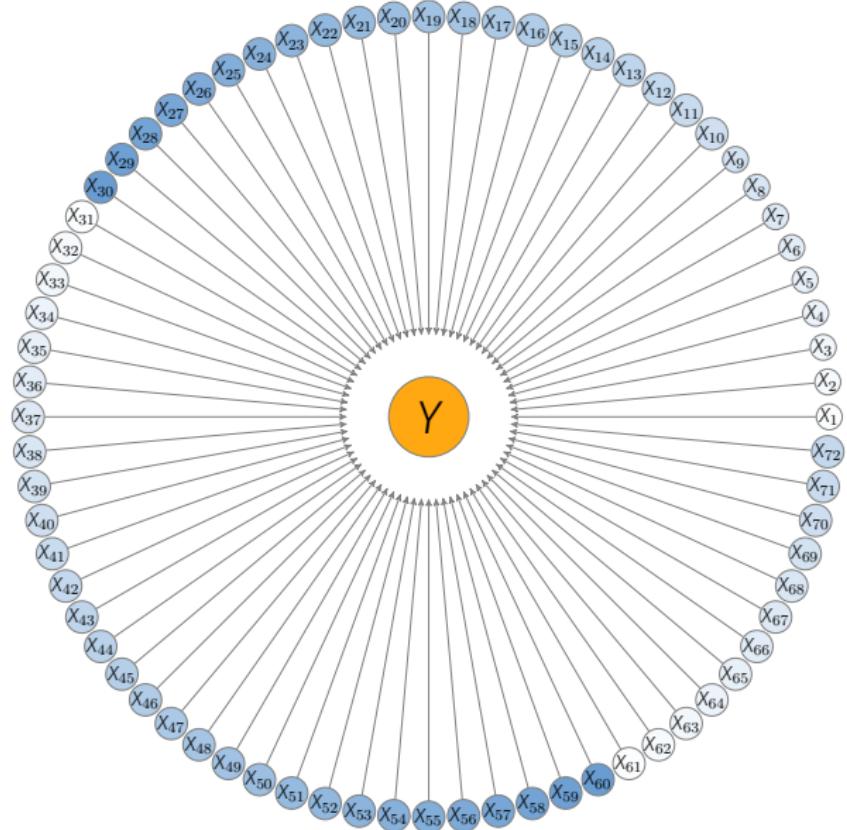
Compromis biais-variance

- La principale limitation de la régression ridge est le fait que tous ses coefficients sont non nuls → modèle complexe

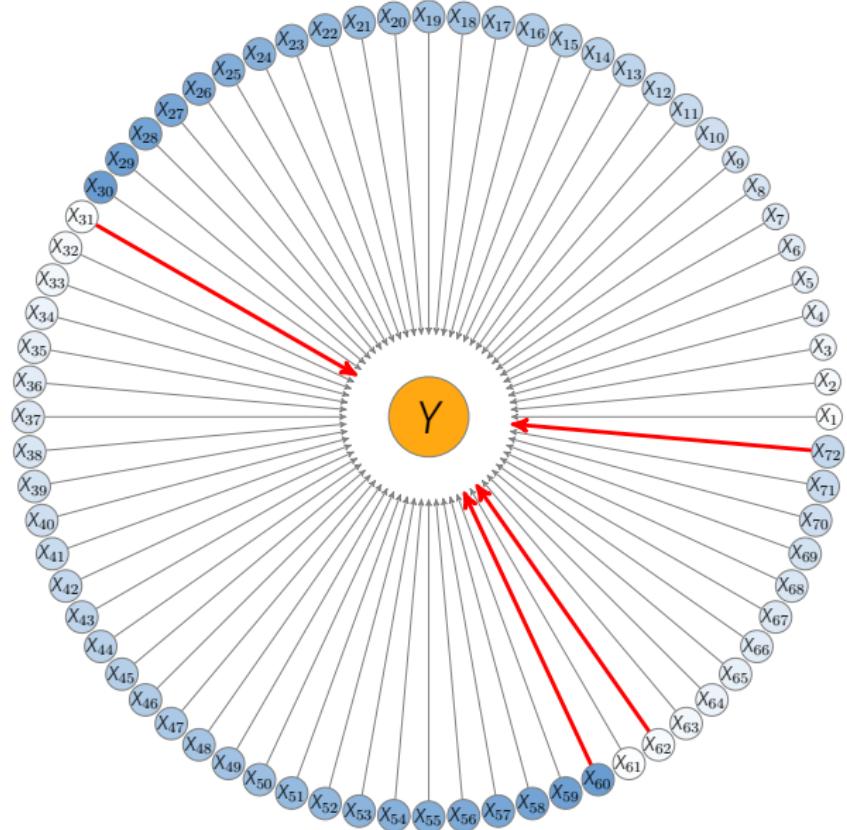


Miser sur la sparsité

Miser sur la sparsité



Miser sur la sparsité



Miser sur la sparsité

Utilisez une procédure qui fonctionne bien pour les problèmes sparse, car aucune procédure ne fonctionne bien pour les problèmes denses.¹

¹The elements of statistical learning. Springer series in statistics, 2001.

Miser sur la sparsité

Utilisez une procédure qui fonctionne bien pour les problèmes sparse, car aucune procédure ne fonctionne bien pour les problèmes denses.¹

- Un modèle statistique sparse est un modèle pour lequel seulement un petit nombre de variables explicatives jouent un rôle important.
- Hypothèse de parcimonie: peu de variables sont pertinentes pour les données de grande dimension ($N << p$).
- β est “creux”
- Les modèles sparse peuvent être plus rapides à calculer, plus faciles à comprendre et produire des prédictions plus stables.

¹The elements of statistical learning. Springer series in statistics, 2001.

Réfléchissons-nous

Comment organiseriez-vous une réunion de 20 personnes?

Comment organiseriez-vous une réunion de 20 personnes?

March 2017												
11 participants	Thu 9	Fri 10	Sat 11	Sun 12	Mon 13	Tue 14	Wed 15	Thu 16	Fri 17	Sat 18	Sun 19	
JayZ	✓	✓	✓			✓			✓	✓	✓	
Evan										✓	✓	✓
Omar	✓	✓		✓		✓			✓	✓	✓	
Caitlin	✓	✓	✓						✓	✓	✓	
Austin	✓	✓	✓									
Ethan			✓	✓					✓		✓	
Max	✓	✓	✓			✓			✓	✓	✓	
Tycho	✓	✓	✓	✓		✓			✓	✓	✓	
Janavi Chaddha		✓	✓	✓		✓	✓		✓	✓	✓	
Charlotte										✓	✓	
Darshanye	✓	✓			✓			✓	✓			
Your name	□	□	□	□	□	□	□	□	□	□	□	□
5:00 PM – 9:00 PM												
Thu 9	Fri 10	Sat 11	Sun 12	Mon 13	Tue 14	Wed 15	Thu 16	Fri 17	Sat 18	Sun 19		
7	8	7	4	0	6	1	0	7	8	9	2	
March 2017												

Les médecins misent aussi sur la sparsité

Les médecins misent aussi sur la sparsité



Un exemple justificatif

Variables explicatives du salaire dans la LNH²



²<https://www.kaggle.com/camnugent/nhl-salary-data-prediction-cleaning-and-modelling>

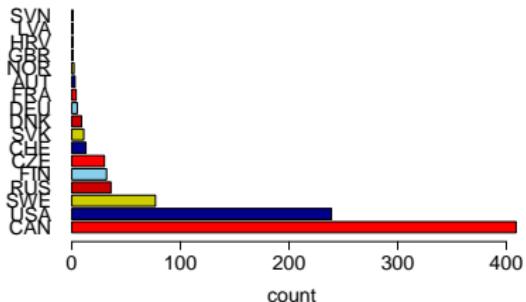
Apprentissage supervisé

■ Apprendre la fonction f

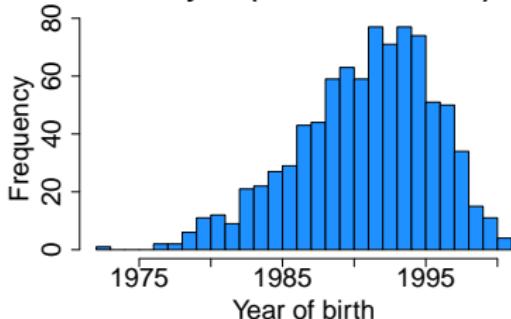


Variables explicatives du salaire des joueurs de la LNH

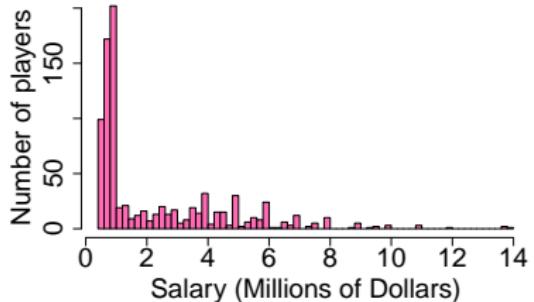
Country



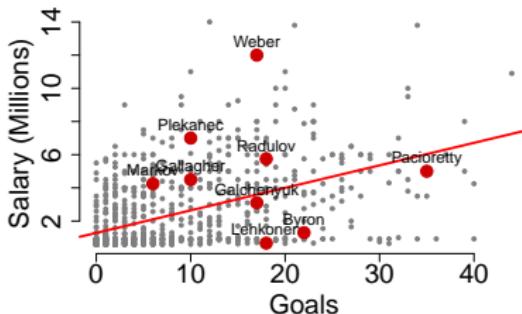
Birth year (2016/2017 season)



NHL Salary Distribution: 2016/2017

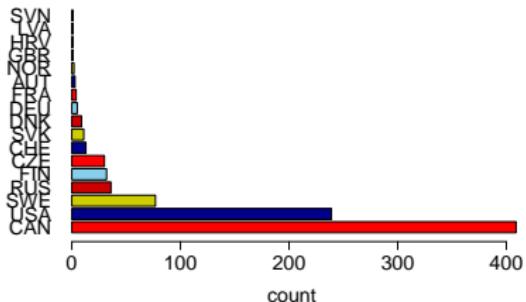


Linear Regression Fit

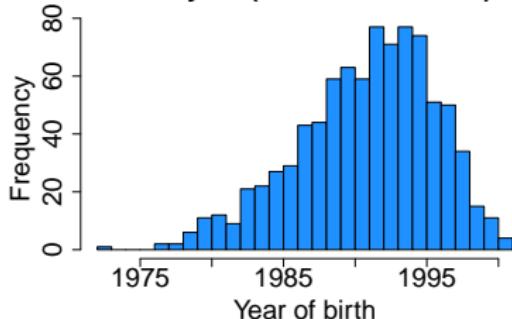


Variables explicatives du salaire des joueurs de la LNH

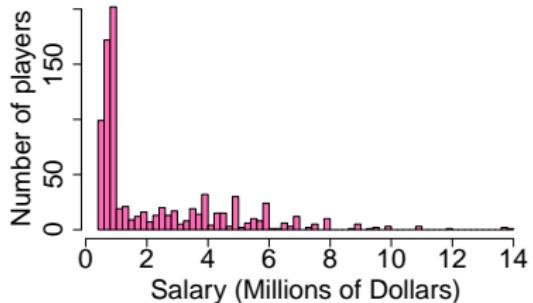
Country



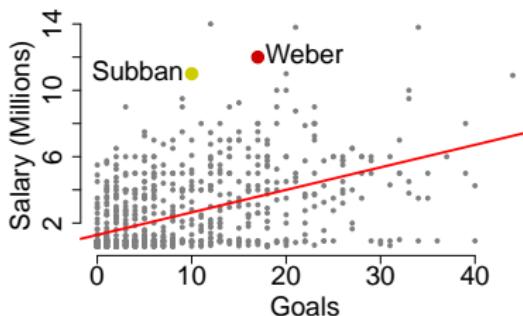
Birth year (2016/2017 season)



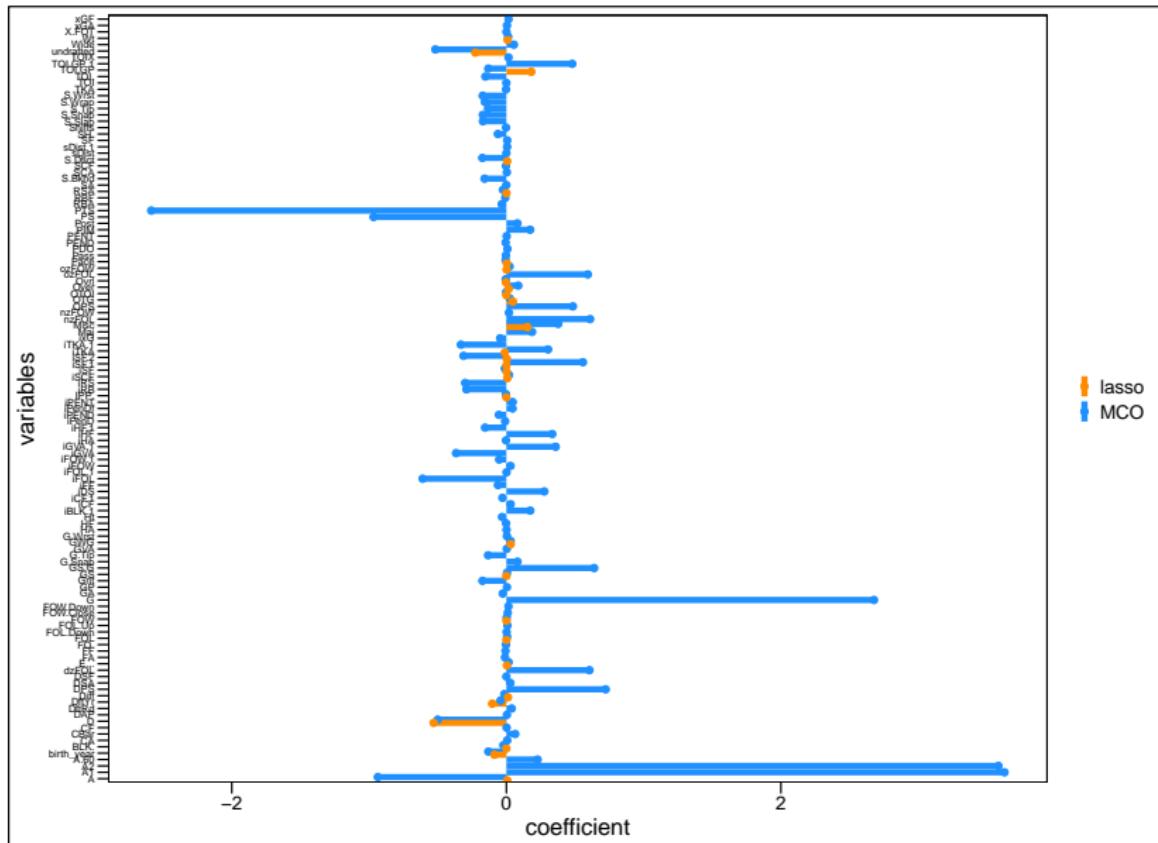
NHL Salary Distribution: 2016/2017



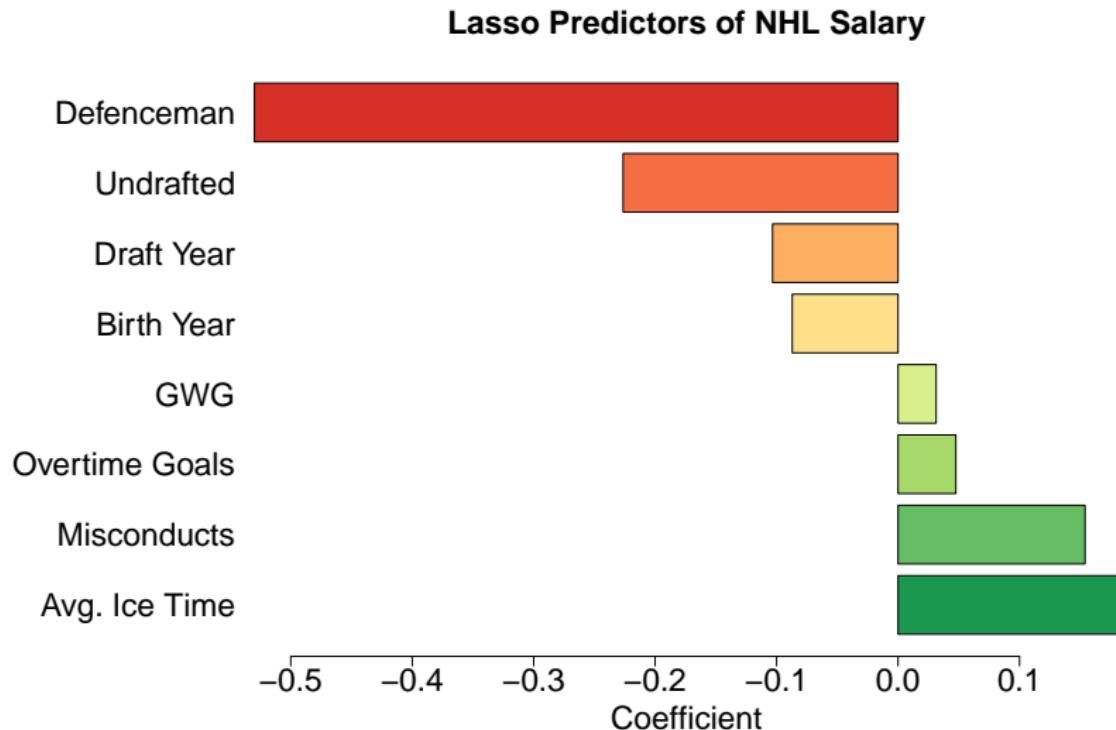
Linear Regression Fit



Coefficients des moindres carrés ordinaires (MCO) et lasso



Variables explicatives sélectionnées par le lasso



Contexte de la méthode lasso
(Tibshirani. *JRSSB*, 1996)

Contexte de la méthode lasso

- Variables explicatives: x_{ij} , $j = 1, \dots, p$, variable réponse: y_i , $i = 1, \dots, n$
- Supposons que les x_{ij} sont standardisés $\rightarrow \sum_i x_{ij}/n = 0$ et $\sum_i x_{ij}^2 = 1$.

¹Tibshirani. JRSSB (1996)

Contexte de la méthode lasso

- Variables explicatives: $x_{ij}, j = 1, \dots, p$, variable réponse: $y_i, i = 1, \dots, n$
- Supposons que les x_{ij} sont standardisés $\rightarrow \sum_i x_{ij}/n = 0$ et $\sum_i x_{ij}^2 = 1$. La fonction de perte du lasso¹ est:

$$\hat{\beta}^{lasso} = \arg \min_{\beta} \frac{1}{2} \sum_{i=1}^n \left(y_i - \sum_{j=1}^p x_{ij} \beta_j \right)^2$$

sujet à $\sum_{j=1}^p |\beta_j| \leq s, \quad s > 0$

¹Tibshirani. JRSSB (1996)

Contexte de la méthode lasso

- Variables explicatives: $x_{ij}, j = 1, \dots, p$, variable réponse: $y_i, i = 1, \dots, n$
- Supposons que les x_{ij} sont standardisés $\rightarrow \sum_i x_{ij}/n = 0$ et $\sum_i x_{ij}^2 = 1$. La fonction de perte du lasso¹ est:

$$\hat{\beta}^{lasso} = \arg \min_{\beta} \frac{1}{2} \sum_{i=1}^n \left(y_i - \sum_{j=1}^p x_{ij} \beta_j \right)^2$$

$$sujet \ à \ \sum_{j=1}^p |\beta_j| \leq s, \quad s > 0$$

- La version de Lagrange du problème, pour $\lambda > 0$

$$\hat{\beta}^{lasso} = \arg \min_{\beta} \frac{1}{2} \sum_{i=1}^n \left(y_i - \sum_{j=1}^p x_{ij} \beta_j \right)^2 + \lambda \sum_{j=1}^p |\beta_j|$$

¹Tibshirani. JRSSB (1996)

La solution du lasso en fonction de l'estimateur MCO

- Considérez une variable explicative et une variable réponse: $\{(x_i, y_i)\}_{i=1}^n$.

La solution du lasso en fonction de l'estimateur MCO

- Considérez une variable explicative et une variable réponse: $\{(x_i, y_i)\}_{i=1}^n$. La fonction de perte du lasso est:

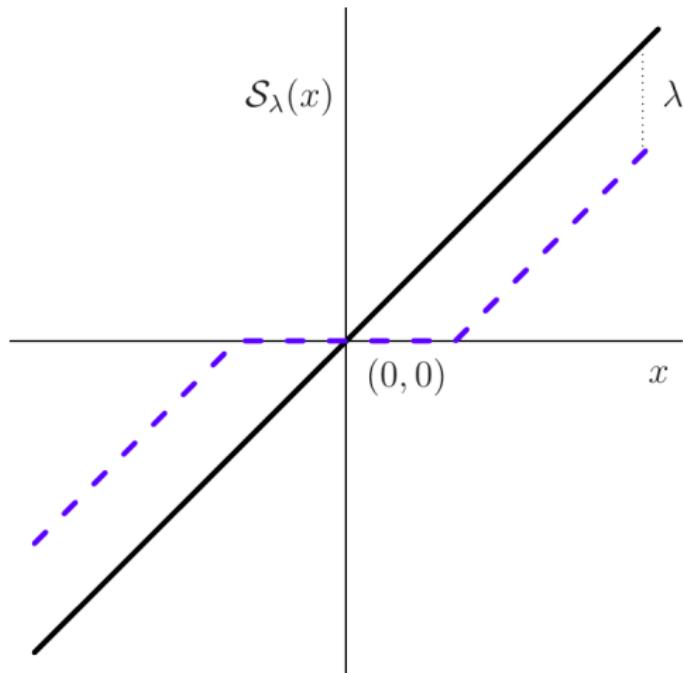
$$\hat{\beta}^{lasso} = \arg \min_{\beta} \frac{1}{2} \sum_{i=1}^n (y_i - x_i \beta)^2 + \lambda |\beta| \quad (1)$$

- Si la variable explicative est **standardisée**, la solution du lasso (1) est une fonction de l'estimateur MCO $\hat{\beta}^{MCO}$

$$\begin{aligned}\hat{\beta}^{lasso} &= S_{\lambda}(\hat{\beta}^{MCO}) = \text{sign}(\hat{\beta}^{MCO}) (|\hat{\beta}^{MCO}| - \lambda)_+ \\ &= \begin{cases} \hat{\beta}^{MCO} - \lambda, & \hat{\beta}^{MCO} > \lambda \\ 0 & |\hat{\beta}^{MCO}| \leq \lambda \\ \hat{\beta}^{MCO} + \lambda & \hat{\beta}^{MCO} \leq -\lambda \end{cases}\end{aligned}$$

La solution du lasso en fonction de l'estimateur MCO

- Lorsque la variable explicative est **standardisée**, le lasso va réduire la solution MCO vers zéro par le facteur λ



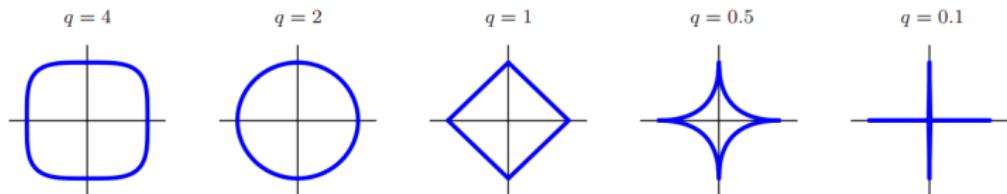
¹Hastie et al. Statistical learning with sparsity: the lasso and generalizations. CRC press, (2015).

Pourquoi la norme ℓ_1 ?

- Pour $q \geq 0$, évaluons le critère

$$\tilde{\boldsymbol{\beta}} = \arg \min_{\boldsymbol{\beta}} \left\{ \sum_{i=1}^n \left(y_i - \beta_0 - \sum_{j=1}^p x_{ij} \beta_j \right)^2 + \lambda \sum_{j=1}^p |\beta_j|^q \right\}$$

- Pourquoi utilisons-nous la norme ℓ_1 et non la $q = 2$ (Ridge) ou toute autre norme ℓ_q ?



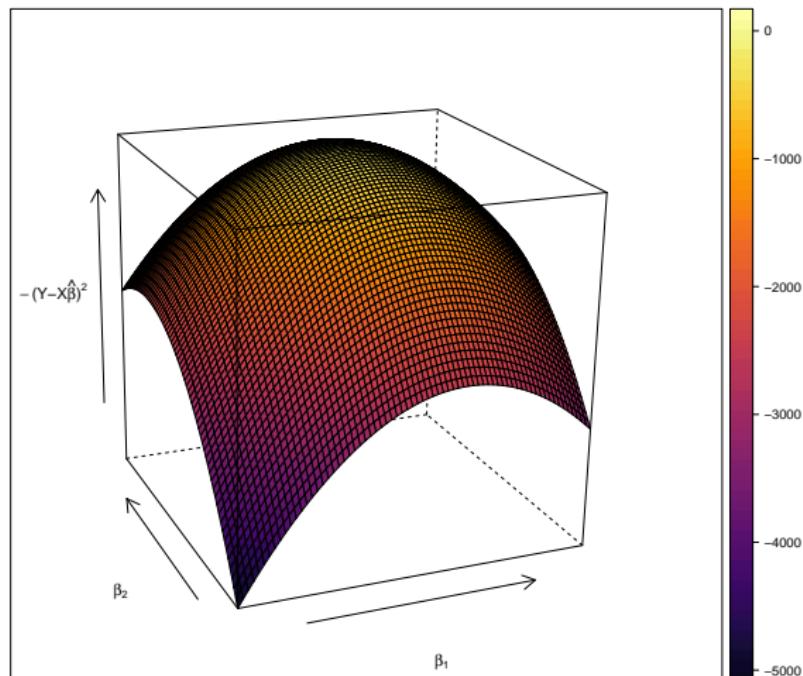
- $q = 1$ est la plus petite valeur qui donne une solution sparse **et** donne un problème **convex** → données de grande dimensions
- Pour $q < 1$ la région contrainte est **non-convexe**

Choisir la complexité du modèle

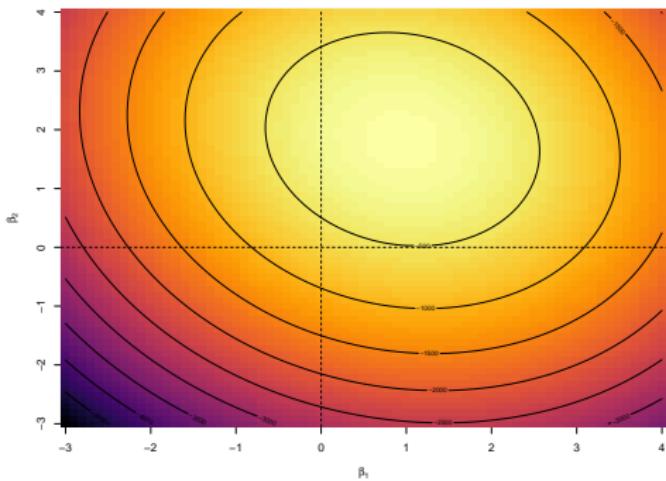
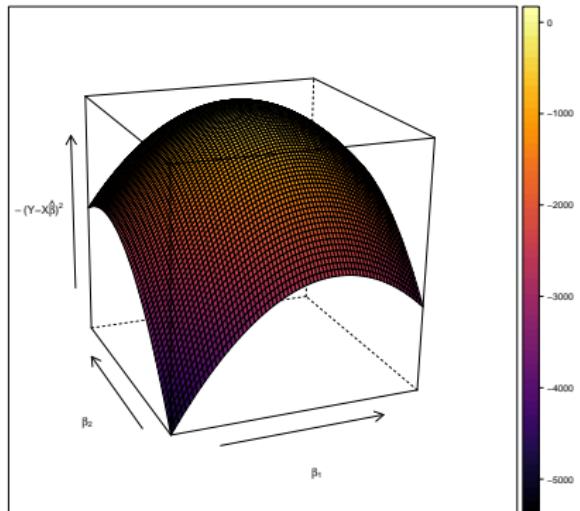
La surface de la somme des carrés résiduelles

- Considérons le modèle suivant avec deux variables explicatives (y est centrée)

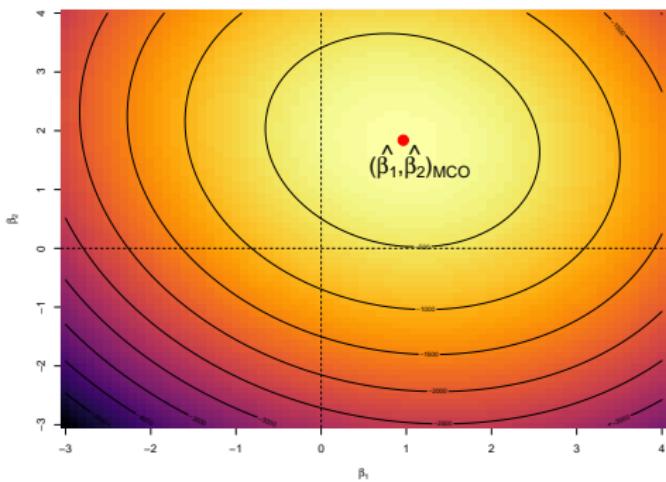
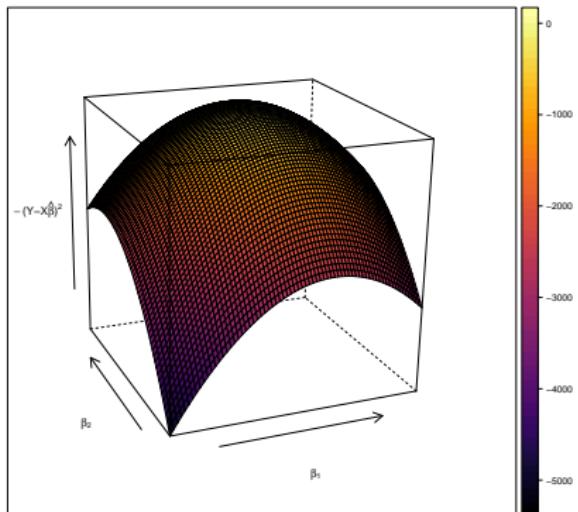
$$\mathbf{y} = \beta_1 \mathbf{x}_1 + \beta_2 \mathbf{x}_2 + \boldsymbol{\varepsilon}$$



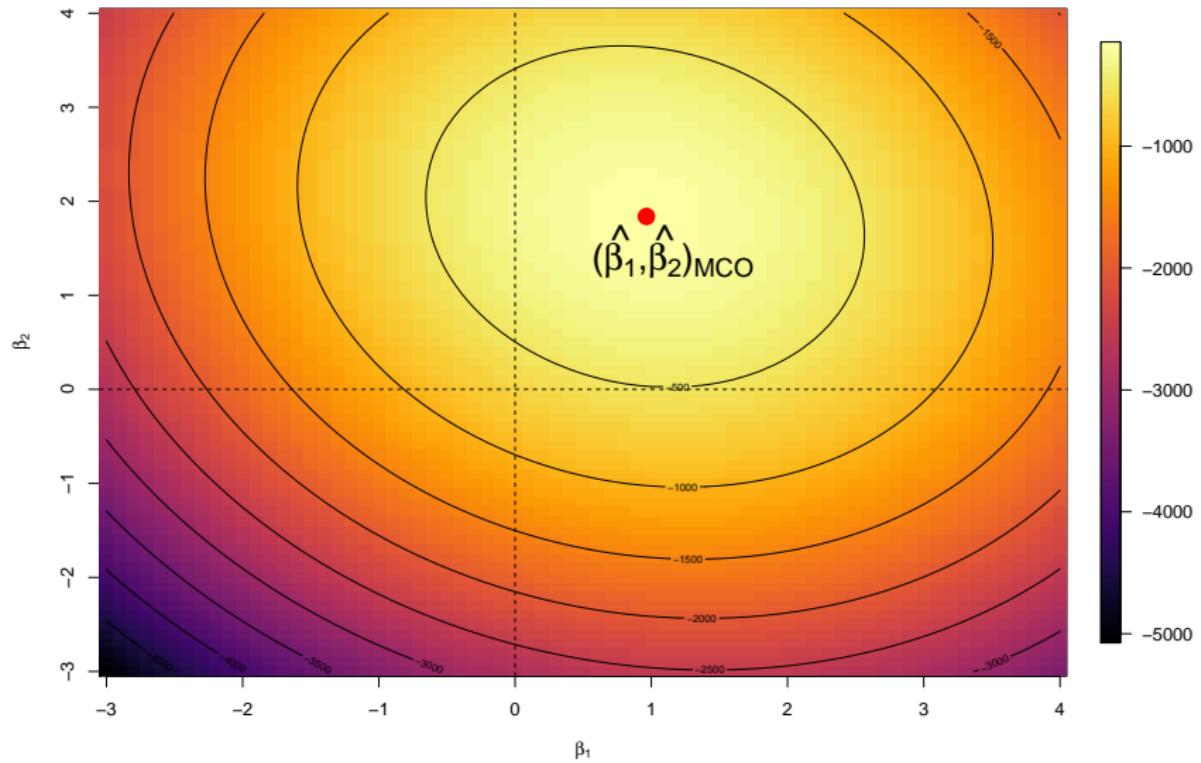
Courbe de niveau de la somme des carrés résiduelles



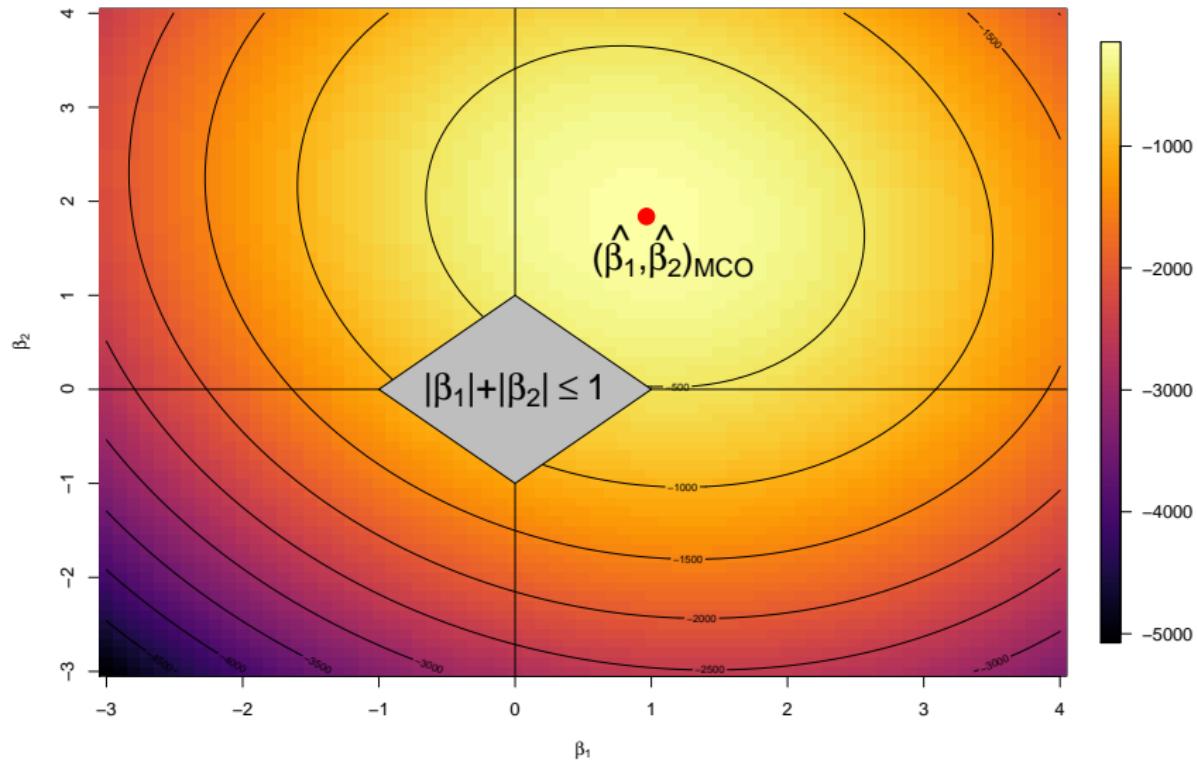
Courbe de niveau de la somme des carrés résiduelles



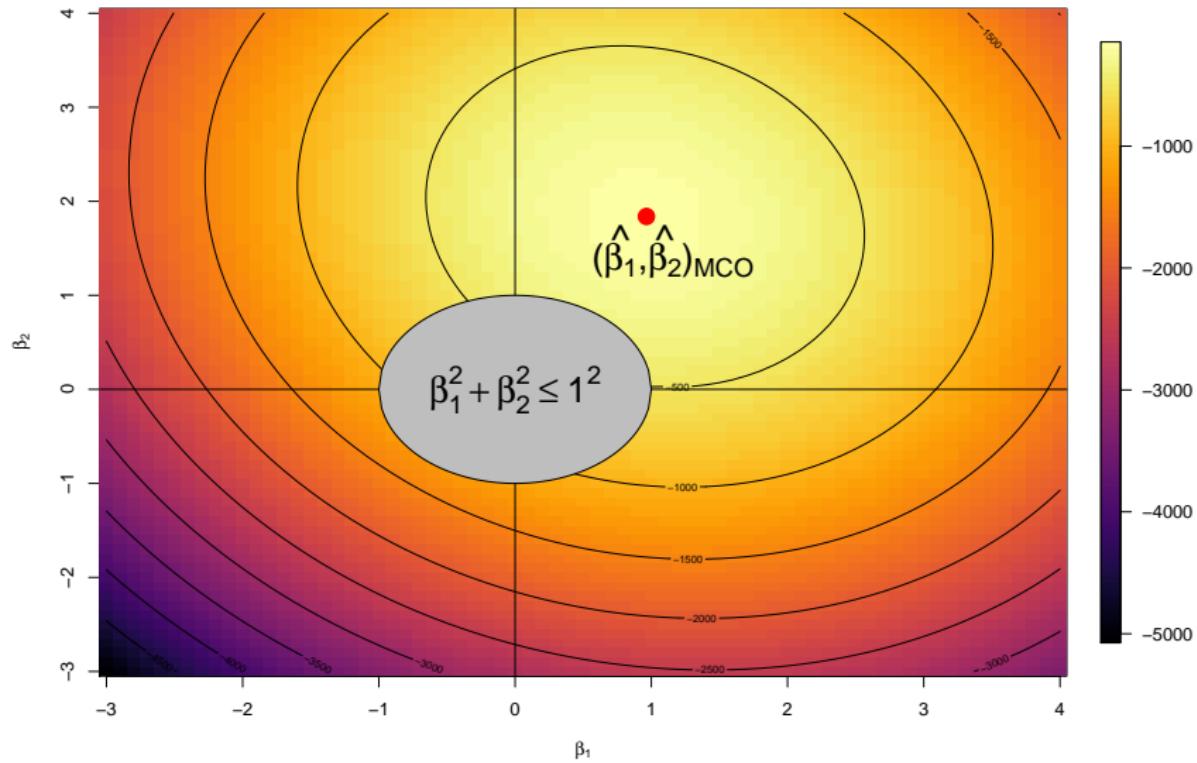
Courbe de niveau de la somme des carrés résiduelles



La région de contrainte du lasso



La région de contrainte du ridge



Lasso vs. ridge

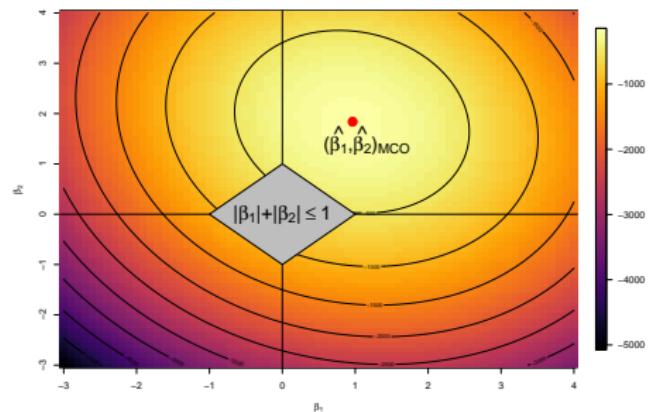


Fig. 1: lasso

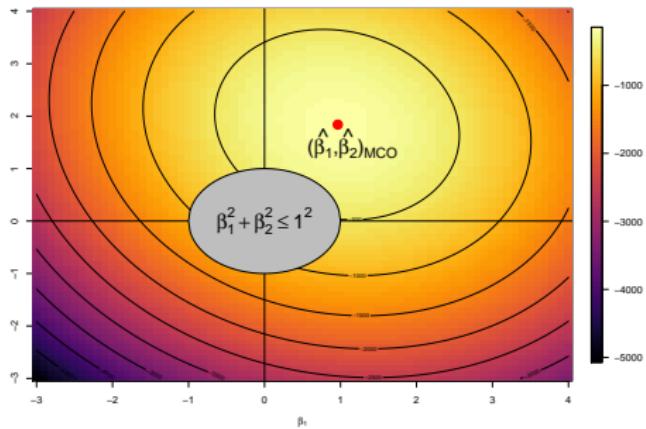
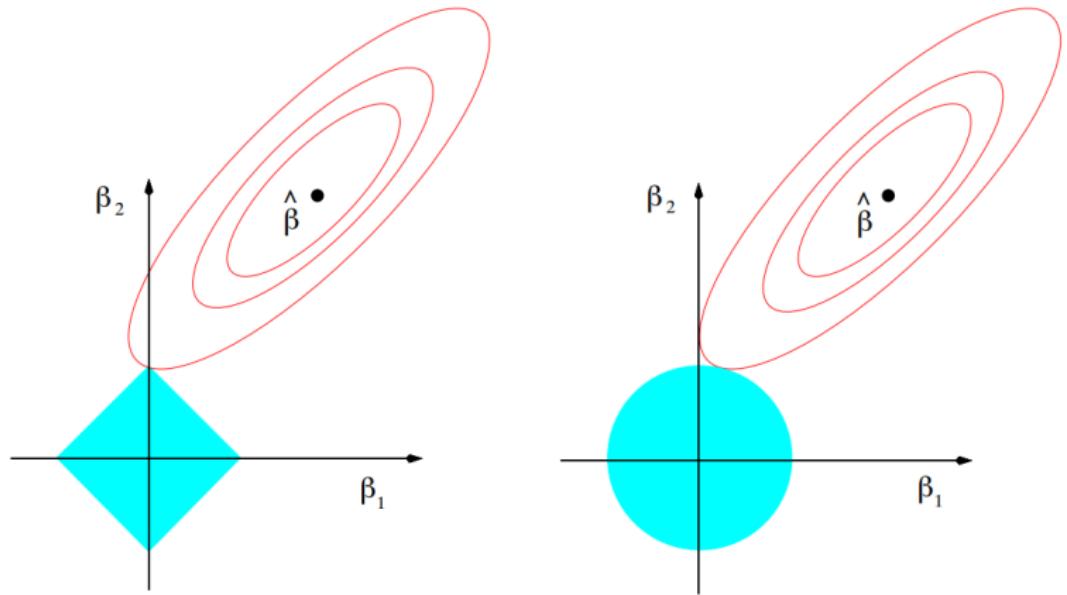


Fig. 2: ridge

Version «classique»



Conditions d'optimalité

L'équation de score et l'équation de score pénalisée

- Dans la théorie statistique classique, la dérivée de la fonction de log-vraisemblance ($\mathcal{L}(\theta)$) s'appelle la fonction de score, et les estimateurs de vraisemblance maximum sont trouvés en réglant cette dérivée égale à zéro, donnant ainsi les équations de score:

$$0 = \frac{\partial}{\partial \theta} \mathcal{L}(\theta)$$

L'équation de score et l'équation de score pénalisée

- Dans la théorie statistique classique, la dérivée de la fonction de log-vraisemblance ($\mathcal{L}(\theta)$) s'appelle la fonction de score, et les estimateurs de vraisemblance maximum sont trouvés en réglant cette dérivée égale à zéro, donnant ainsi les équations de score:

$$0 = \frac{\partial}{\partial \theta} \mathcal{L}(\theta)$$

- Étendre cette idée aux fonction de log-vraisemblance pénalisées implique de prendre les dérivés de fonctions de perte de la forme:

$$\mathbf{Q}(\theta) = \underbrace{\mathcal{L}(\theta)}_{\text{vraisemblance}} + \underbrace{P(\theta)}_{\text{pénalité}}$$

donnant les équations de score pénalisée

L'équation de score pénalisée

- Pour la régression ridge, la vraisemblance pénalisée est partout différentiable, et l'extension aux équations de score pénalisées est simple

$$\hat{\boldsymbol{\beta}}^{ridge} = \arg \min_{\boldsymbol{\beta}} \frac{1}{2} \|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|_2^2 + \lambda \|\boldsymbol{\beta}\|_2^2$$

L'équation de score pénalisée

- Pour la régression ridge, la vraisemblance pénalisée est partout différentiable, et l'extension aux équations de score pénalisées est simple

$$\hat{\boldsymbol{\beta}}^{ridge} = \arg \min_{\boldsymbol{\beta}} \frac{1}{2} \|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|_2^2 + \lambda \|\boldsymbol{\beta}\|_2^2$$

- Pour le lasso la vraisemblance pénalisée n'est pas différentiable, spécifiquement, non différentiables à zéro → sous-différentiels sont nécessaire pour les caractériser

$$\hat{\boldsymbol{\beta}}^{lasso} = \arg \min_{\boldsymbol{\beta}} \mathbf{Q}(\theta) = \arg \min_{\boldsymbol{\beta}} \frac{1}{2} \|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|_2^2 + \lambda \|\boldsymbol{\beta}\|_1$$

L'équation de score pénalisée

- Pour la régression ridge, la vraisemblance pénalisée est partout différentiable, et l'extension aux équations de score pénalisées est simple

$$\hat{\boldsymbol{\beta}}^{ridge} = \arg \min_{\boldsymbol{\beta}} \frac{1}{2} \|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|_2^2 + \lambda \|\boldsymbol{\beta}\|_2^2$$

- Pour le lasso la vraisemblance pénalisée n'est pas différentiable, spécifiquement, non différentiables à zéro → sous-différentiels sont nécessaire pour les caractériser

$$\hat{\boldsymbol{\beta}}^{lasso} = \arg \min_{\boldsymbol{\beta}} \mathbf{Q}(\theta) = \arg \min_{\boldsymbol{\beta}} \frac{1}{2} \|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|_2^2 + \lambda \|\boldsymbol{\beta}\|_1$$

- Soit $\partial \mathbf{Q}(\theta)$ le sous-différentiel de \mathbf{Q} , les équations de score pénalisées sont:

$$0 \in \partial \mathbf{Q}(\theta)$$

Conditions d'optimalité de Karush-Kuhn-Tucker (KKT)

- Dans la littérature sur l'optimisation, les équations résultantes sont connu sous le nom de Karush-Kuhn-Tucker (KKT)

Conditions d'optimalité de Karush-Kuhn-Tucker (KKT)

- Dans la littérature sur l'optimisation, les équations résultantes sont connu sous le nom de Karush-Kuhn-Tucker (KKT)
- Pour les problèmes d'optimisation convexes tels que le lasso, les conditions KKT sont à la fois nécessaires et suffisantes pour caractériser la solution

Conditions d'optimalité de Karush-Kuhn-Tucker (KKT)

- Dans la littérature sur l'optimisation, les équations résultantes sont connu sous le nom de Karush-Kuhn-Tucker (KKT)
- Pour les problèmes d'optimisation convexes tels que le lasso, les conditions KKT sont à la fois nécessaires et suffisantes pour caractériser la solution
- L'idée est simple: pour résoudre $\widehat{\beta}^{lasso}$, nous remplaçons simplement le dérivé avec le sous-gradient et le vraisemblance avec le vraisemblance pénalisée

Conditions KKT pour le lasso

- Résultat: $\hat{\beta}^{lasso}$ minimize la fonction de perte du lasso si et seulement si il répond aux conditions KKT:

$$\frac{1}{n} \mathbf{x}_j^\top (\mathbf{y} - \mathbf{X}\hat{\beta}) = \lambda \text{sign}(\hat{\beta}_j) \quad \hat{\beta}_j \neq 0$$

$$\frac{1}{n} |\mathbf{x}_j^\top (\mathbf{y} - \mathbf{X}\hat{\beta})| \leq \lambda \quad \hat{\beta}_j = 0$$

Conditions KKT pour le lasso

- **Résultat:** $\hat{\beta}^{lasso}$ minimize la fonction de perte du lasso si et seulement si il répond aux conditions KKT:

$$\frac{1}{n} \mathbf{x}_j^\top (\mathbf{y} - \mathbf{X}\hat{\beta}) = \lambda \text{sign}(\hat{\beta}_j) \quad \hat{\beta}_j \neq 0$$

$$\frac{1}{n} |\mathbf{x}_j^\top (\mathbf{y} - \mathbf{X}\hat{\beta})| \leq \lambda \quad \hat{\beta}_j = 0$$

- La corrélation entre un prédicteur et les résidus, $\mathbf{x}_j^\top (\mathbf{y} - \mathbf{X}\hat{\beta})/n$, doit dépasser un certain minimum seuil λ avant qu'il soit inclus dans le modèle
- Lorsque cette corrélation est inférieure à λ , $\hat{\beta}_j = 0$

Algorithme

Algorithme pour le lasso

- Les conditions KKT nous permettent seulement de vérifier une solution
- Ils n'aident pas nécessairement à trouver la solution dans un premier temps

L'algorithme coordinate descent¹

- L'algorithme coordinate descent est simplement d'optimiser une fonction de perte par rapport à un seul paramètre à la fois, de façon itérative, à travers tous les paramètres jusqu'à ce que la convergence atteint

¹Fu (1998), Friedman et al. (2007), Wu and Lange (2008)

L'algorithme coordinate descent¹

- L'algorithme coordinate descent est simplement d'optimiser une fonction de perte par rapport à un seul paramètre à la fois, de façon itérative, à travers tous les paramètres jusqu'à ce que la convergence atteint
- Ceci est particulièrement adaptée aux problèmes, tels que le lasso, qui a une forme explicite dans un seul dimension mais manque une dans les dimensions supérieures

¹Fu (1998), Friedman et al. (2007), Wu and Lange (2008)

L'algorithme coordinate descent

- Envisageons de minimiser \mathbf{Q} par rapport à β_j , tout en traitant temporairement les autres coefficients de régression $\boldsymbol{\beta}_{-j}$ comme fixe:

$$\mathbf{Q}(\beta_j | \boldsymbol{\beta}_{-j}) = \frac{1}{2n} \sum_{i=1}^n \left(y_i - \sum_{k \neq j} x_{ij} \beta_k - x_{ij} \beta_j \right)^2 + \lambda |\beta_j| + \lambda \sum_{k \neq j} |\beta_k|$$

L'algorithme coordinate descent

- Envisageons de minimiser \mathbf{Q} par rapport à β_j , tout en traitant temporairement les autres coefficients de régression $\boldsymbol{\beta}_{-j}$ comme fixe:

$$\mathbf{Q}(\beta_j | \boldsymbol{\beta}_{-j}) = \frac{1}{2n} \sum_{i=1}^n \left(y_i - \sum_{k \neq j} x_{ij} \beta_k - x_{ij} \beta_j \right)^2 + \lambda |\beta_j| + \lambda \sum_{k \neq j} |\beta_k|$$

$$\tilde{\beta}_j = \arg \min_{\beta_j} \mathbf{Q}(\beta_j | \boldsymbol{\beta}_{-j}) = S_\lambda(\tilde{z}_j) = \begin{cases} \tilde{z}_j - \lambda, & \tilde{z}_j > \lambda \\ 0 & |\tilde{z}_j| \leq \lambda \\ \tilde{z}_j + \lambda & \tilde{z}_j < -\lambda \end{cases}$$

L'algorithme coordinate descent

- Envisageons de minimiser \mathbf{Q} par rapport à β_j , tout en traitant temporairement les autres coefficients de régression $\boldsymbol{\beta}_{-j}$ comme fixe:

$$\mathbf{Q}(\beta_j | \boldsymbol{\beta}_{-j}) = \frac{1}{2n} \sum_{i=1}^n \left(y_i - \sum_{k \neq j} x_{ij} \beta_k - x_{ij} \beta_j \right)^2 + \lambda |\beta_j| + \lambda \sum_{k \neq j} |\beta_k|$$

$$\tilde{\beta}_j = \arg \min_{\beta_j} \mathbf{Q}(\beta_j | \boldsymbol{\beta}_{-j}) = S_\lambda(\tilde{z}_j) = \begin{cases} \tilde{z}_j - \lambda, & \tilde{z}_j > \lambda \\ 0 & |\tilde{z}_j| \leq \lambda \\ \tilde{z}_j + \lambda & \tilde{z}_j < -\lambda \end{cases}$$

- $\tilde{r}_{ij} = y_i - \sum_{k \neq j} x_{ik} \tilde{\beta}_k$ $\tilde{z}_j = n^{-1} \sum_{i=1}^n x_{ij} \tilde{r}_{ij}$
- $\{\tilde{r}_{ij}\}_{i=1}^n$ sont les résidus partiels par rapport au covariable j , et \tilde{z}_j est l'estimateur MCO basée sur $\{\tilde{r}_{ij}, x_{ij}\}_{i=1}^n$

La convergence

- L'analyse numérique des problèmes d'optimisation de la forme

$$\mathbf{Q}(\theta) = \mathcal{L}(\theta) + P(\theta)$$

a montré que le coordinate descent converge vers une solution de la vraisemblance pénalisée à condition que:

La convergence

- L'analyse numérique des problèmes d'optimisation de la forme

$$Q(\theta) = \mathcal{L}(\theta) + P(\theta)$$

a montré que le coordinate descent converge vers une solution de la vraisemblance pénalisée à condition que:

- ▶ la fonction $\mathcal{L}(\beta)$ soit différentiable et
- ▶ la fonction de pénalité $P_\lambda(\beta)$ est séparable →

$$P_\lambda(\beta) = \sum_j P_\lambda(\beta_j)$$

La convergence

- L'analyse numérique des problèmes d'optimisation de la forme

$$Q(\theta) = \mathcal{L}(\theta) + P(\theta)$$

a montré que le coordinate descent converge vers une solution de la vraisemblance pénalisée à condition que:

- ▶ la fonction $\mathcal{L}(\beta)$ soit différentiable et
- ▶ la fonction de pénalité $P_\lambda(\beta)$ est séparable \rightarrow
$$P_\lambda(\beta) = \sum_j P_\lambda(\beta_j)$$

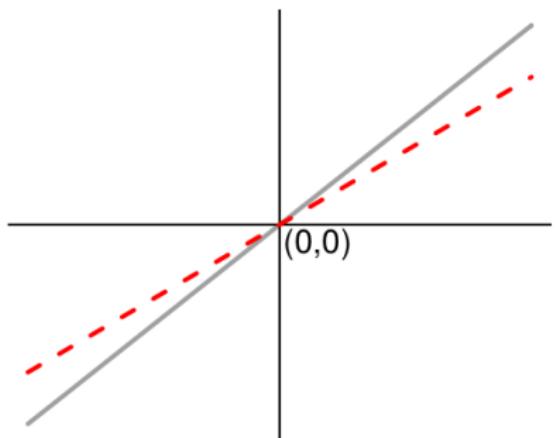
- Le lasso satisfait à ces deux critères

Autres pénalité pour la sélection des variables

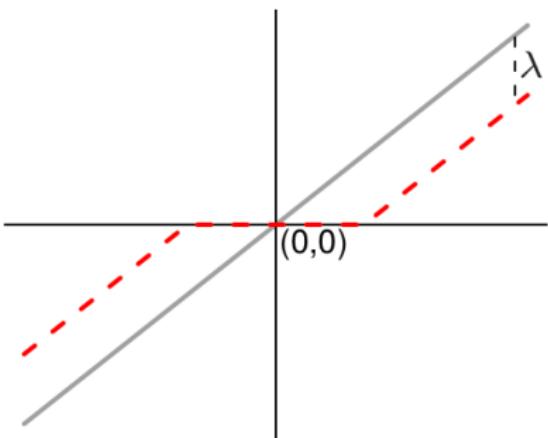
Le biais du lasso

q	Estimator	Formula
1	Lasso	$\text{sign}(\hat{\beta}_j^{\text{LS}})(\hat{\beta}_j^{\text{LS}} - \lambda)_+$
2	Ridge	$\hat{\beta}_j^{\text{LS}} / (1 + \lambda)$

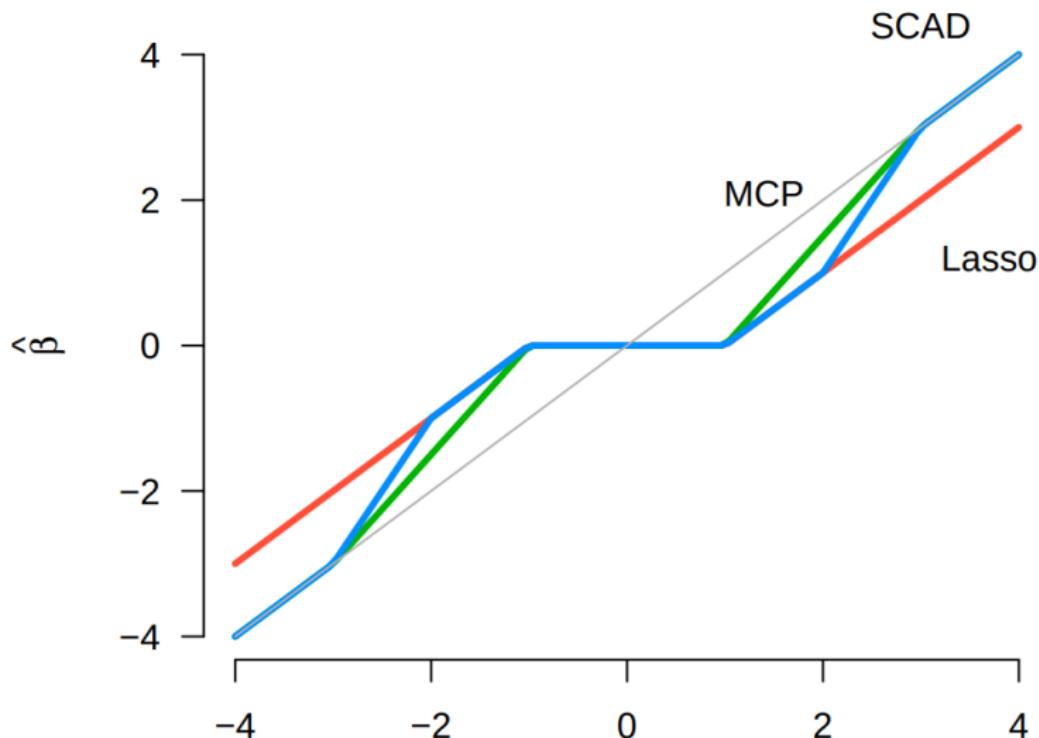
Ridge



Lasso



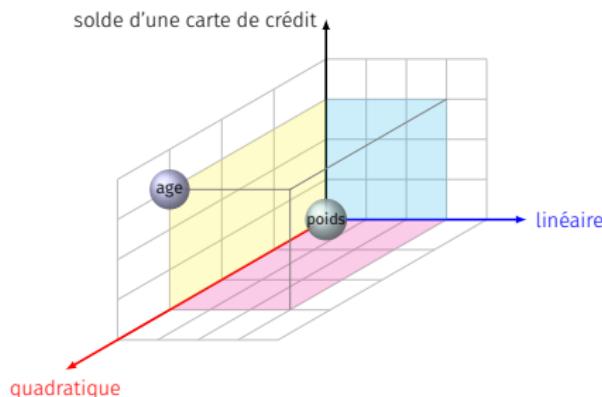
SCAD (Fan et Li, JASA, 2001), MCP (Zhang, Ann. Stat., 2010)



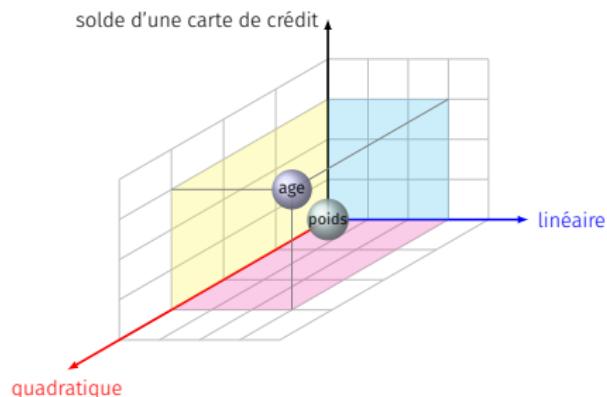
Le lasso pour des groupes de variables explicatives

Utile pour des groupes de variables (facteur avec > 2 catégories, Age, Age²). L'estimateur du **groupe lasso** est:

$$\min_{(\beta_0, \boldsymbol{\beta})} \frac{1}{2} \|\mathbf{y} - \beta_0 - \mathbf{X}\boldsymbol{\beta}\|_2^2 + \lambda \sum_{k=1}^K \sqrt{p_k} \|\boldsymbol{\beta}^{(k)}\|_2 \quad p_k - \text{taille de group}$$



(a) Lasso



(b) Groupe lasso

Discussion

- Les méthodes de sélection de variables constituent un domaine de recherche actif en raison de leurs applications répandues dans les problèmes de données de grande dimension
- Une limitation de ces méthodes est le manque d'outils d'inférence

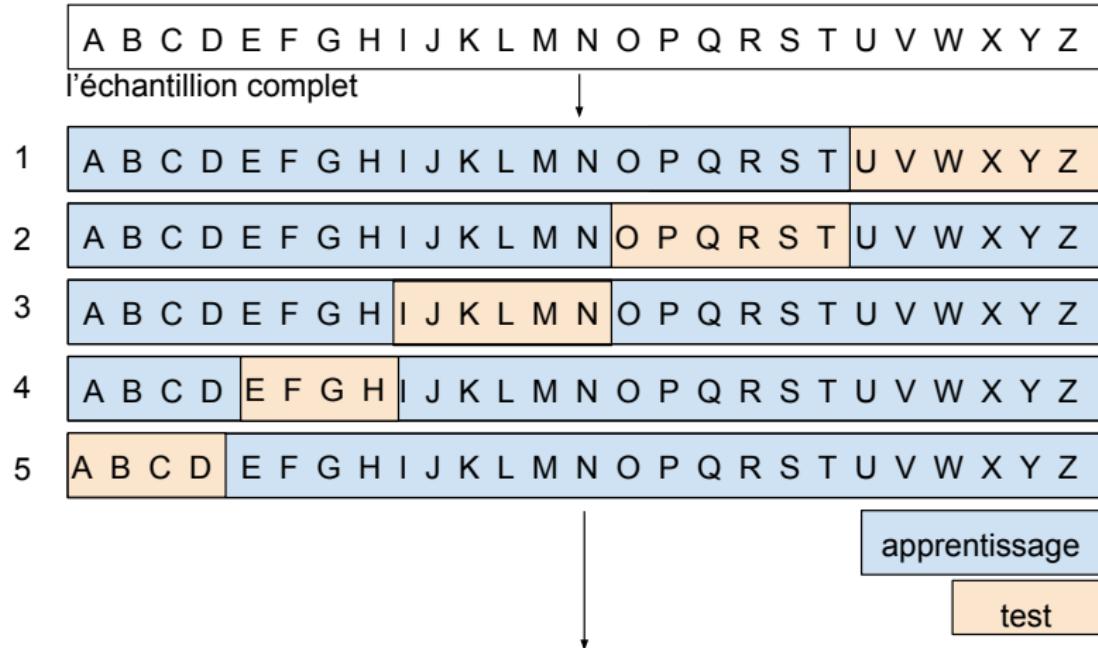
Références

- Fan, J. and Li, R., 2001. Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American statistical Association*, 96(456), pp.1348-1360.
- Tibshirani, R., 1996. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, 58(1), pp.267-288.
- Friedman, J., Hastie, T., Höfling, H. and Tibshirani, R., 2007. Pathwise coordinate optimization. *The annals of applied statistics*, 1(2), pp.302-332.
- Buhlmann, P. & van de Geer, S. (2011), *Statistics for High-Dimensional Data*, Springer.
- Breheny, P. [BIOS 7240 class notes \(accessed March 15, 2019\)](#).
- Tibshirani, R. [A Closer Look at Sparse Regression \(accessed March 15, 2019\)](#).
- Gaillard, P. and Rudi, A. [Introduction to Machine Learning \(accessed March 15, 2019\)](#).
- Hastie, T., Tibshirani, R. & Friedman, J. (2009), *The Elements of Statistical Learning; Data Mining, Inference and Prediction*, Springer. Second edition.
- Hastie, T., Tibshirani, R. & Wainwright, M. (2015), *Statistical Learning with Sparsity: the Lasso and Generalizations*, Chapman & Hall.

slides available at

<https://sahirbhatnagar.com/talks/>

Contexte sur la validation croisée



$$CV(\alpha) = \frac{1}{5} \sum_{v=1}^5 MSE_v^{(test)}$$

SCAD

$$p'(|\beta|; \lambda) = \lambda \text{sign}(\beta_j) \left\{ I_{(|\beta_j| \leq \lambda)} + \frac{(a\lambda - |\beta_j|)_+}{(a-1)\lambda} I_{(|\beta_j| > \lambda)} \right\}, \quad a > 2$$

The penalty is expressed in terms of its derivative. The SCAD is a combination of the HARD, LASSO, and Clipped penalties.
This leads to the solution

$$\hat{\beta}_{j,SCAD} = \begin{cases} \text{sign}(\hat{\beta}_{j,OLS})(|\hat{\beta}_{j,OLS}| - \lambda)_+ & |\hat{\beta}_{j,OLS}| \leq 2\lambda \\ \frac{(a-1)\hat{\beta}_{j,OLS} - \text{sign}(\hat{\beta}_{j,OLS})a\lambda}{a-2} & 2\lambda < |\hat{\beta}_{j,OLS}| \leq a\lambda \\ \hat{\beta}_{j,OLS} & |\hat{\beta}_{j,OLS}| > a\lambda \end{cases}$$

$$p(|\beta|_j; \lambda, \gamma) = \begin{cases} \lambda|\beta_j| - \frac{|\beta_j|^2}{2\gamma} & |\beta_j| \leq \gamma\lambda \\ \frac{\gamma\lambda^2}{2} & |\beta_j| > \gamma\lambda \end{cases}$$