

MISER SUR LA SPARSITÉ

Sahir Bhatnagar, candidat au doctorat, Biostatistique, Université de McGill

En collaboration avec Karim Oualkacha, Yi Yang et Celia Greenwood

le 18 décembre 2017

INTRODUCTION

MON CHEMINEMENT

- 2008: Baccalauréat en actuariat (Concordia)

MON CHEMINEMENT

- 2008: Baccalauréat en actuariat (Concordia)
- 2009: Aon Hewitt - Régimes de retraite

MON CHEMINEMENT

- 2008: Baccalauréat en actuariat (Concordia)
- 2009: Aon Hewitt - Régimes de retraite
- 2010: Associé de la Société des Actuaires (ASA)

MON CHEMINEMENT

- 2008: Baccalauréat en actuariat (Concordia)
- 2009: Aon Hewitt - Régimes de retraite
- 2010: Associé de la Société des Actuaires (ASA)
- 2012: Maîtrise en Biostatistique (Queen's)

MON CHEMINEMENT

- 2008: Baccalauréat en actuariat (Concordia)
- 2009: Aon Hewitt - Régimes de retraite
- 2010: Associé de la Société des Actuaires (ASA)
- 2012: Maîtrise en Biostatistique (Queen's)
- 2013: Doctorat en Biostatistique (McGill, diplôme prévu mai 2018)

MON CHEMINEMENT

- 2008: Baccalauréat en actuariat (Concordia)
- 2009: Aon Hewitt - Régimes de retraite
- 2010: Associé de la Société des Actuaires (ASA)
- 2012: Maîtrise en Biostatistique (Queen's)
- 2013: Doctorat en Biostatistique (McGill, diplôme prévu mai 2018)
- 2016: Wellcome Trust Sanger Institute (Cambridge)

MON CHEMINEMENT

- 2008: Baccalauréat en actuariat (Concordia)
- 2009: Aon Hewitt - Régimes de retraite
- 2010: Associé de la Société des Actuaires (ASA)
- 2012: Maîtrise en Biostatistique (Queen's)
- 2013: Doctorat en Biostatistique (McGill, diplôme prévu mai 2018)
- 2016: Wellcome Trust Sanger Institute (Cambridge)
- 2017: Chargé de cours théorie des probabilités et l'inférence statistique (McGill)

MON CHEMINEMENT

- 2008: Baccalauréat en actuariat (Concordia)
- 2009: Aon Hewitt - Régimes de retraite
- 2010: Associé de la Société des Actuaires (ASA)
- 2012: Maîtrise en Biostatistique (Queen's)
- 2013: Doctorat en Biostatistique (McGill, diplôme prévu mai 2018)
- 2016: Wellcome Trust Sanger Institute (Cambridge)
- 2017: Chargé de cours théorie des probabilités et l'inférence statistique (McGill)
- 2017: Consultation en statistiques avec des chercheurs au CHUM, l'Hôpital juif, CUSM

MON CHEMINEMENT

- 2008: Baccalauréat en actuariat (Concordia)
- 2009: Aon Hewitt - Régimes de retraite
- 2010: Associé de la Société des Actuaires (ASA)
- 2012: Maîtrise en Biostatistique (Queen's)
- 2013: Doctorat en Biostatistique (McGill, diplôme prévu mai 2018)
- 2016: Wellcome Trust Sanger Institute (Cambridge)
- 2017: Chargé de cours théorie des probabilités et l'inférence statistique (McGill)
- 2017: Consultation en statistiques avec des chercheurs au CHUM, l'Hôpital juif, CUSM
- sahirbhatnagar.com

APERÇU

1. Un exemple justificatif

1. Un exemple justificatif
2. Contexte sur les méthodes de pénalisation lasso et groupe lasso

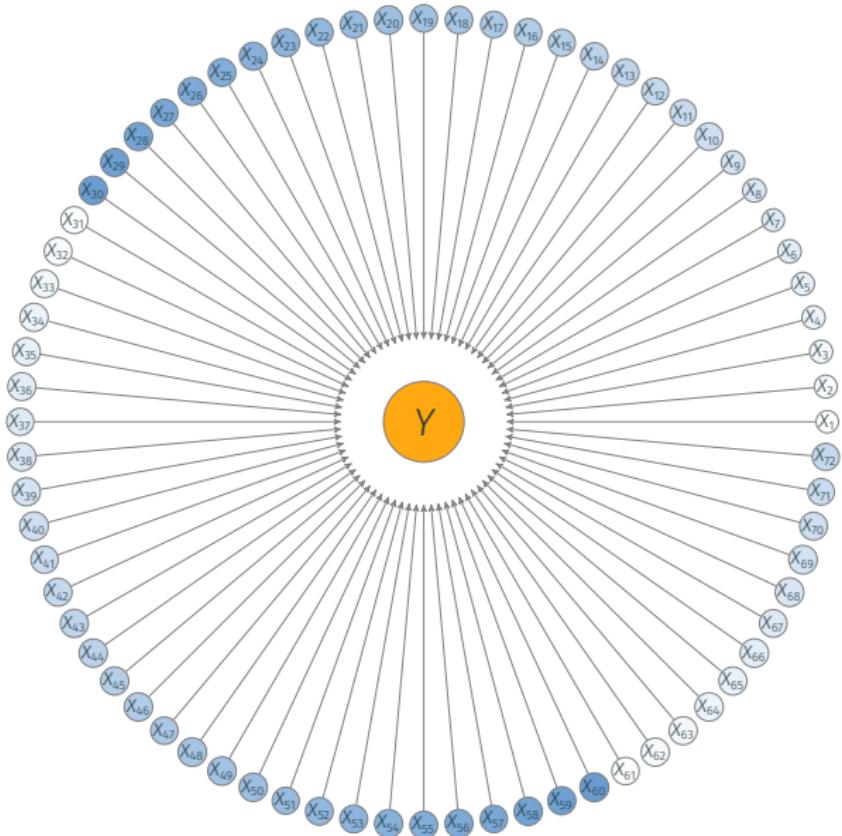
APERÇU

1. Un exemple justificatif
2. Contexte sur les méthodes de pénalisation lasso et groupe lasso
3. Portrait global de mes paquets R

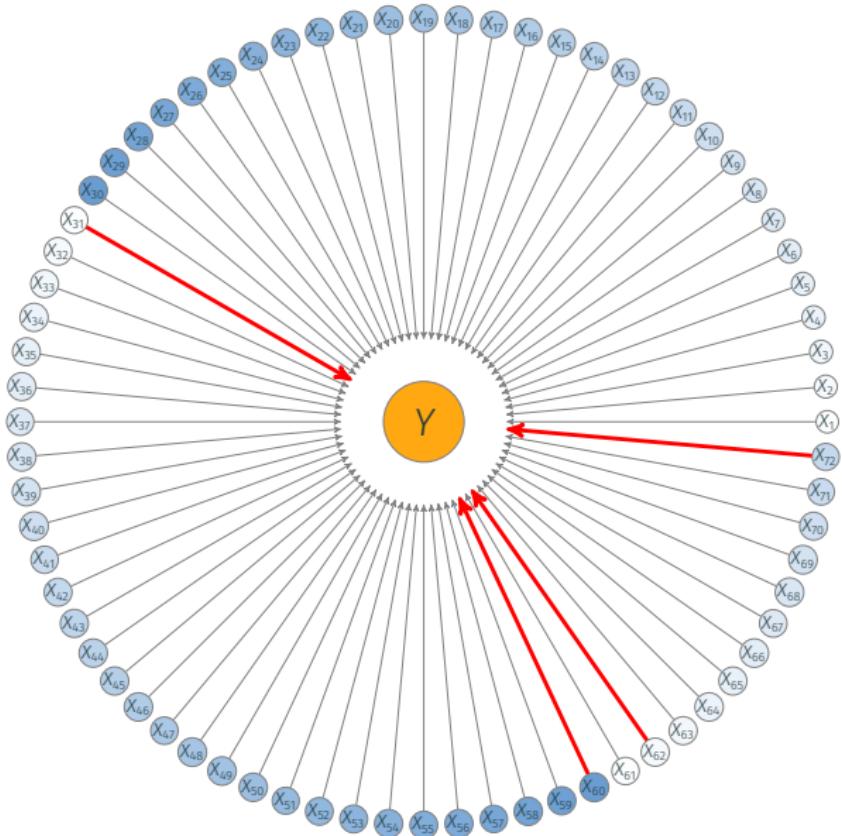
1. Un exemple justificatif
2. Contexte sur les méthodes de pénalisation lasso et groupe lasso
3. Portrait global de mes paquets R
4. Présentation de deux de nos méthodes de pénalisation: **sail** et **ggmix**

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.

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 ($N \ll p$).
- 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.

UN EXEMPLE JUSTIFICATIF

VARIABLES EXPLICATIVES DU SALAIRE DANS LA LNH²



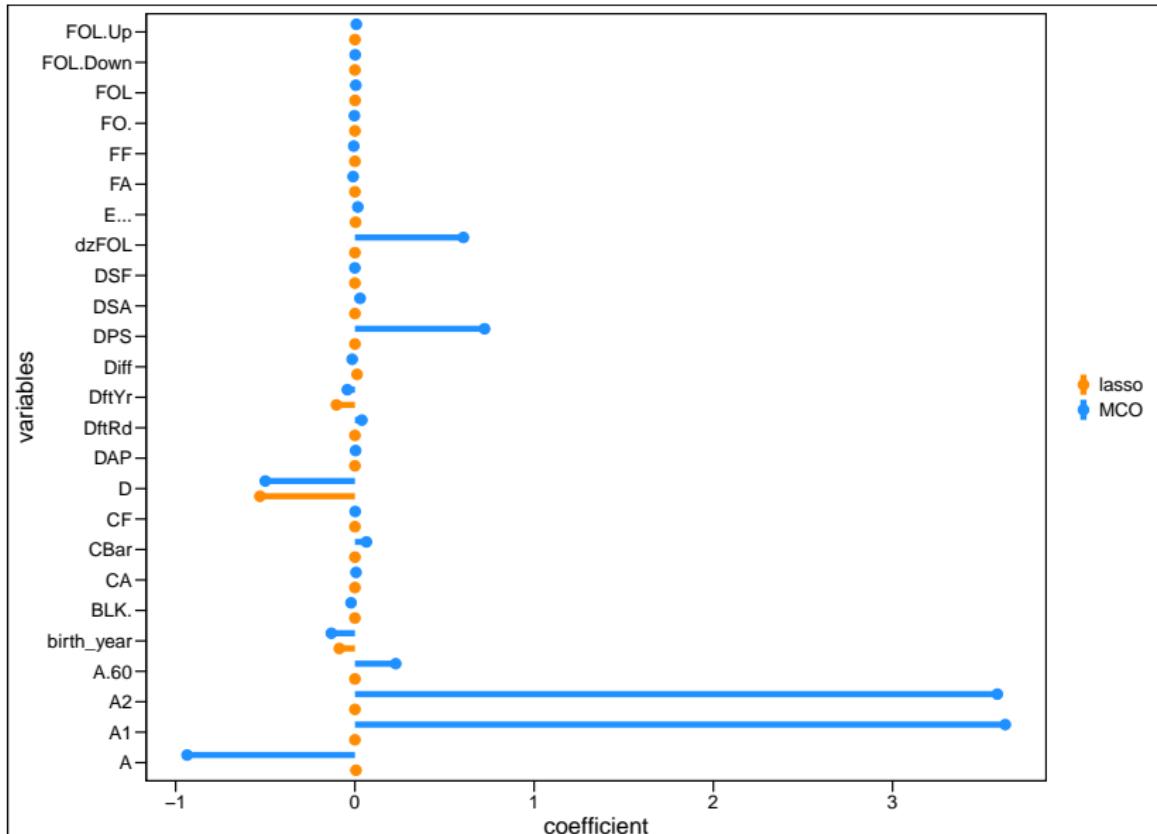
²<https://www.kaggle.com/camnugent/nhl-salary-data-prediction-cleaning-and-modelling>

APPRENTISSAGE SUPERVISÉ

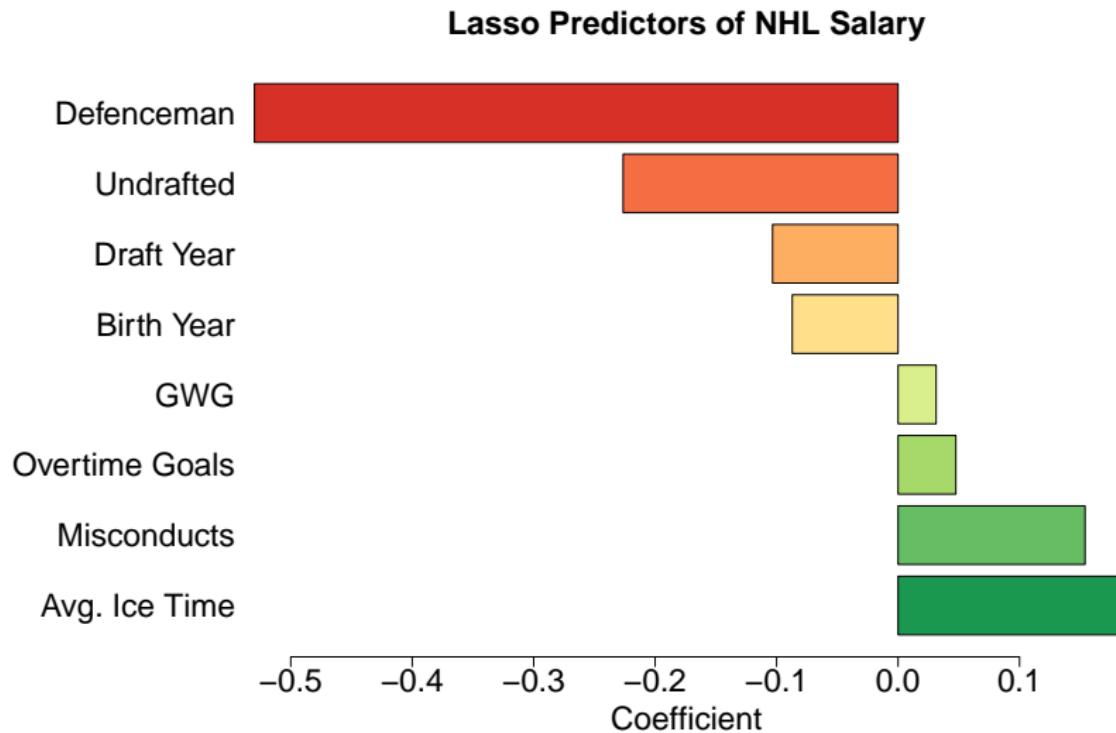
■ Apprendre la fonction f



COEFFICIENTS DES MOINDRES CARRÉS ORDINAIRES (MCO) ET LASSO



VARIABLES EXPLICATIVES SÉLECTIONNÉES PAR LE LASSO



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:

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

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

- 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

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

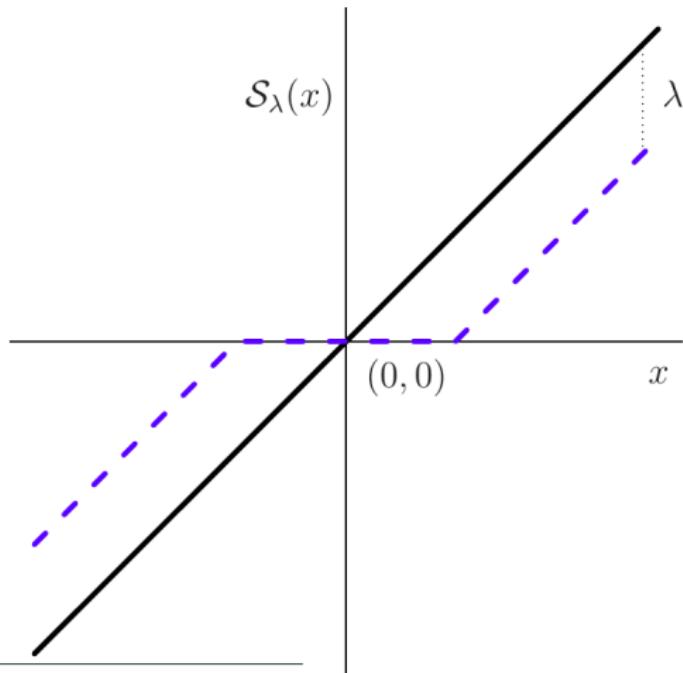
$$\hat{\beta}^{lasso} = \arg \min_{\beta} \frac{1}{2} \sum_{i=1}^n (y_i - x_i \beta)^2 + \lambda |\beta| \quad (1)$$

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

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

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



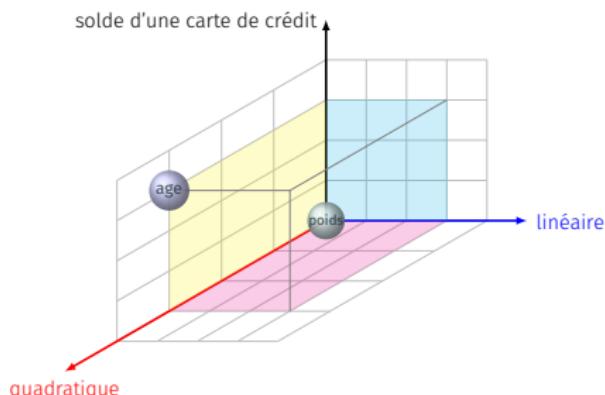
¹Hastie et al. Statistical learning with sparsity: the lasso and generalizations. CRC press, (2015).

CHOISIR LA COMPLEXITÉ DU MODÈLE

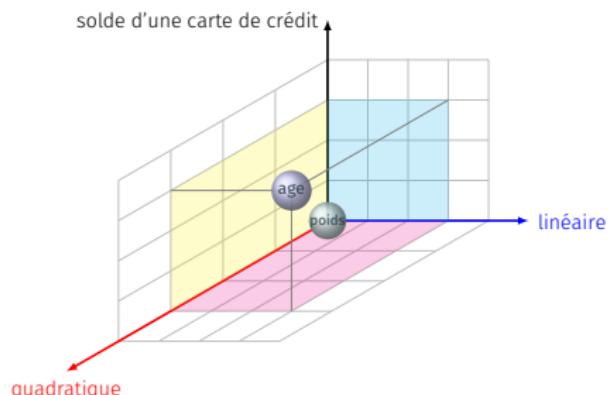
LE LASSO POUR DES GROUPES DE VARIABLES EXPLICATIVES

L'estimateur du **groupe lasso** est:

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



(a) Lasso



(b) Groupe lasso

NOS LOGICIELS

UN APERÇU DE NOS PAQUETS R

- **eclust** – Bhatnagar et al. (2017, Genetic Epidemiology)
<https://cran.r-project.org/package=eclust>
- **sail** – Bhatnagar, Yang and Greenwood (2017+, preprint)
<https://github.com/sahirbhatnagar/sail>
- **gmmix** – Bhatnagar, Oualkacha, Yang, Greenwood (2017+, preprint)
<https://github.com/sahirbhatnagar/gmmix>
- **casebase** – Bhatnagar¹, Turgeon¹, Yang, Hanley and Saarela (2017+, preprint)
<https://cran.r-project.org/package=casebase>

¹co-auteurs

UN APERÇU DE NOS PAQUETS R

	eclust	sail	gmmix	casebase
Modèle				
Moindres carrés	✓	✓	✓	
Classification binaire	✓			
Analyse de survie				✓
Penalité				
Ridge	✓		✓	✓
Lasso	✓	✓	✓	✓
Elastic Net	✓		✓	✓
Group Lasso		✓	✓	
Particularité				
Interactions	✓	✓		✓
Modélisation flexible	✓	✓		✓
Effets aléatoires				✓
Données	(x, y, e)	(x, y, e)	(x, y, Ψ)	(x, t, δ)

sail: L'APPRENTISSAGE DES
INTERACTIONS NON-LINÉAIRES AYANT
LA PROPRIÉTÉ D'HÉRÉDITÉ FORTE

MOTIVATION 1: LA PROPRIÉTÉ D'HÉRÉDITÉ FORTE

$$Y = \beta_0 \cdot \mathbf{1} + \sum_{j=1}^p \beta_j X_j + \beta_E X_E + \sum_{j=1}^p \alpha_j X_E X_j + \varepsilon$$

¹Chipman. Canadian Journal of Statistics (1996)

²McCullagh and Nelder. Generalized Linear Models (1983)

³Cox. International Statistical Review (1984)

MOTIVATION 1: LA PROPRIÉTÉ D'HÉRÉDITÉ FORTE

$$Y = \beta_0 \cdot 1 + \sum_{j=1}^p \beta_j X_j + \beta_E X_E + \sum_{j=1}^p \alpha_j X_E X_j + \varepsilon$$

La propriété d'héritéité forte¹

$$\hat{\alpha}_j \neq 0 \quad \Rightarrow \quad \hat{\beta}_j \neq 0 \quad \text{et} \quad \hat{\beta}_E \neq 0$$

la propriété d'héritéité faible¹

$$\hat{\alpha}_j \neq 0 \quad \Rightarrow \quad \hat{\beta}_j \neq 0 \quad \text{ou} \quad \hat{\beta}_E \neq 0$$

- La propriété d'héritéité forte est utile pour l'interprétation².
- Les grands effets principaux sont plus susceptibles d'entraîner des interactions modestes³.

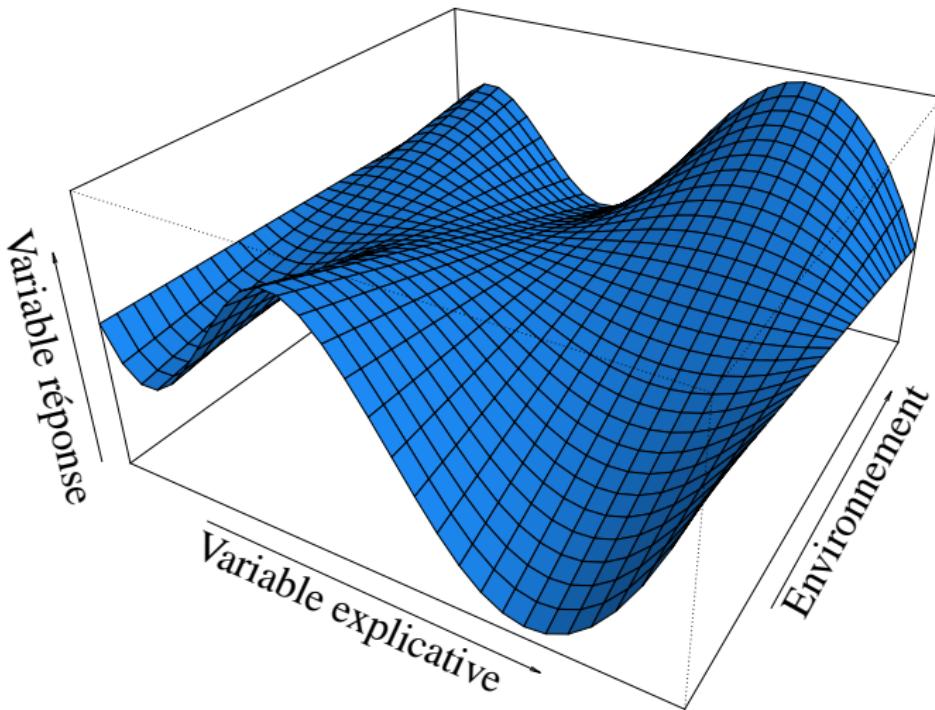
¹Chipman. Canadian Journal of Statistics (1996)

²McCullagh and Nelder. Generalized Linear Models (1983)

³Cox. International Statistical Review (1984)

MOTIVATION 2: INTERACTIONS NON-LINÉAIRES

Interaction non-linéaire



LE LASSO AVEC DES INTERACTIONS

- $Y \rightarrow$ variable réponse
- $X_E \rightarrow$ environnement
- $X_j \rightarrow$ variables fixes, $j = 1, \dots, p$

$$Y = \beta_0 \cdot \mathbf{1} + \sum_{j=1}^p \beta_j X_j + \beta_E X_E + \sum_{j=1}^p \alpha_j X_E X_j + \varepsilon$$

$$\operatorname{argmin}_{\beta_0, \beta, \alpha} \mathcal{L}(Y; \Theta) + \lambda(\|\beta\|_1 + \|\alpha\|_1)$$

MODÈLES AYANT LA PROPRIÉTÉ D'HÉRÉDITÉ FORTE: ÉTAT ACTUEL DE LA RECHERCHE

Particularité	Modèle	Logiciel
Linéaire	CAP (Zhao et al. 2009, <i>Ann. Stat</i>)	X
	SHIM (Choi et al. 2009, <i>JASA</i>)	X
	hiernet (Bien et al. 2013, <i>Ann. Stat</i>)	hierNet(x, y)
	GRESH (She and Jiang 2014, <i>JASA</i>)	X
	FAMILY (Haris et al. 2014, <i>JCGS</i>)	FAMILY(x, z, y)
	glinternet (Lim and Hastie 2015, <i>JCGS</i>)	glinternet(x, y)
	RAMP (Hao et al. 2016, <i>JASA</i>)	RAMP(x, y)
Non-linéaire	VANISH (Radchenko and James 2010, <i>JASA</i>)	X
	sail (Bhatnagar et al. 2017+)	sail(x, e, y)

NOTRE CONTRIBUTION POUR LES EFFETS NON-LINÉAIRES

On considère l'expansion avec des splines:

$$f_j(x_j) = \sum_{\ell=1}^{p_j} \psi_{j\ell}(x_j) \beta_{j\ell}$$

$$f(x_1) = \underbrace{\begin{bmatrix} \psi_{11}(x_{11}) & \psi_{12}(x_{12}) & \cdots & \psi_{11}(x_{15}) \\ \vdots & \vdots & \cdots & \vdots \\ \vdots & \vdots & \cdots & \vdots \\ \psi_{11}(x_{i1}) & \psi_{12}(x_{i2}) & \cdots & \psi_{11}(x_{i5}) \\ \vdots & \vdots & \cdots & \vdots \\ \vdots & \vdots & \cdots & \vdots \\ \psi_{11}(x_{N1}) & \psi_{12}(x_{N2}) & \cdots & \psi_{11}(x_{N5}) \end{bmatrix}}_{\Psi_1}_{N \times 5} \times \underbrace{\begin{bmatrix} \beta_{11} \\ \beta_{12} \\ \beta_{13} \\ \beta_{14} \\ \beta_{15} \end{bmatrix}}_{\theta_1}_{5 \times 1}$$

- $\theta_j = (\beta_{j1}, \dots, \beta_{jp_j}) \in \mathbb{R}^{p_j}$
- $\alpha_j = (\alpha_{j1}, \dots, \alpha_{jp_j}) \in \mathbb{R}^{p_j}$
- $\Psi_j \rightarrow n \times p_j$ matrices de $\psi_{j\ell}$
- En pratique, on utilise des **bsplines** avec 5 degrés de liberté.

Modèle

$$Y = \beta_0 \cdot 1 + \sum_{j=1}^p \Psi_j \theta_j + \beta_E X_E + \sum_{j=1}^p X_E \Psi_j \alpha_j + \varepsilon$$

sail: PROPRIÉTÉ D'HÉRÉDITÉ FORTE

Reparamétrisation¹

$$\alpha_j = \gamma_j \beta_E \theta_j$$

Modèle

$$Y = \beta_0 \cdot \mathbf{1} + \sum_{j=1}^p \Psi_j \theta_j + \beta_E X_E + \sum_{j=1}^p \gamma_j \beta_E X_E \Psi_j \theta_j + \varepsilon$$

Fonction de perte

$$\operatorname{argmin}_{\beta_E, \boldsymbol{\theta}, \boldsymbol{\gamma}} \mathcal{L}(Y; \Theta) + \lambda_\beta \left(w_E |\beta_E| + \sum_{j=1}^p w_j \|\theta_j\|_2 \right) + \lambda_\gamma \sum_{j=1}^p w_{jE} |\gamma_j|$$

¹Choi et al. JASA (2010)

ALGORITHMME

BLOCK RELAXATION (DE LEEUW, 1994)

Algorithm 1: Block Relaxation Algorithm

Définir le compteur d'itération $k \leftarrow 0$, valeurs initiales $\Theta^{(0)}$;

for pour chaque paire $(\lambda_\beta, \lambda_\gamma)$ **do**

repeat

$$\boldsymbol{\gamma}^{(k+1)} \leftarrow \operatorname{argmin}_{\boldsymbol{\gamma}} Q_{\lambda_\beta, \lambda_\gamma} (\boldsymbol{\gamma}, \beta_E^{(k)}, \boldsymbol{\theta}^{(k)})$$

$$\boldsymbol{\theta}^{(k+1)} \leftarrow \operatorname{argmin}_{\boldsymbol{\theta}} Q_{\lambda_\beta, \lambda_\gamma} (\boldsymbol{\theta}, \beta_E^{(k)}, \boldsymbol{\gamma}^{(k+1)})$$

$$\beta_E^{(k+1)} \leftarrow \operatorname{argmin}_{\beta_E} Q_{\lambda_\beta, \lambda_\gamma} (\boldsymbol{\theta}^{(k+1)}, \beta_E, \boldsymbol{\gamma}^{(k+1)})$$

$$k \leftarrow k + 1$$

until à la convergence;

end

Fonction de perte

$$\operatorname{argmin}_{\beta_E, \theta, \gamma} \mathcal{L}(Y; \Theta) + \lambda_\beta \left(w_E |\beta_E| + \sum_{j=1}^p w_j \|\theta_j\|_2 \right) + \lambda_\gamma \sum_{j=1}^p w_{jE} |\gamma_j|$$

¹<https://github.com/sahirbhatnagar/sail>

Fonction de perte

$$\operatorname{argmin}_{\beta_E, \theta, \gamma} \mathcal{L}(Y; \Theta) + \lambda_\beta \left(w_E |\beta_E| + \sum_{j=1}^p w_j \|\theta_j\|_2 \right) + \lambda_\gamma \sum_{j=1}^p w_{jE} |\gamma_j|$$

Un lasso modifié

$$\operatorname{argmin}_\gamma \mathcal{L}(Y; \Theta) + \lambda_\beta \left(w_E |\beta_E| + \sum_{j=1}^p w_j \|\theta_j\|_2 \right) + \lambda_\gamma \sum_{j=1}^p w_{jE} |\gamma_j|$$

¹<https://github.com/sahirbhatnagar/sail>

Fonction de perte

$$\operatorname{argmin}_{\beta_E, \theta, \gamma} \mathcal{L}(Y; \Theta) + \lambda_\beta \left(w_E |\beta_E| + \sum_{j=1}^p w_j \|\theta_j\|_2 \right) + \lambda_\gamma \sum_{j=1}^p w_{jE} |\gamma_j|$$

¹<https://github.com/sahirbhatnagar/sail>

Fonction de perte

$$\operatorname{argmin}_{\beta_E, \theta, \gamma} \mathcal{L}(Y; \Theta) + \lambda_\beta \left(w_E |\beta_E| + \sum_{j=1}^p w_j \|\theta_j\|_2 \right) + \lambda_\gamma \sum_{j=1}^p w_{jE} |\gamma_j|$$

Un groupe lasso modifié

$$\operatorname{argmin}_{\beta_E, \theta} \mathcal{L}(Y; \Theta) + \lambda_\beta \left(w_E |\beta_E| + \sum_{j=1}^p w_j \|\theta_j\|_2 \right) + \lambda_\gamma \sum_{j=1}^p w_{jE} |\gamma_j|$$

¹<https://github.com/sahirbhatnagar/sail>

SIMULATIONS

SCÉNARIOS DE SIMULATIONS

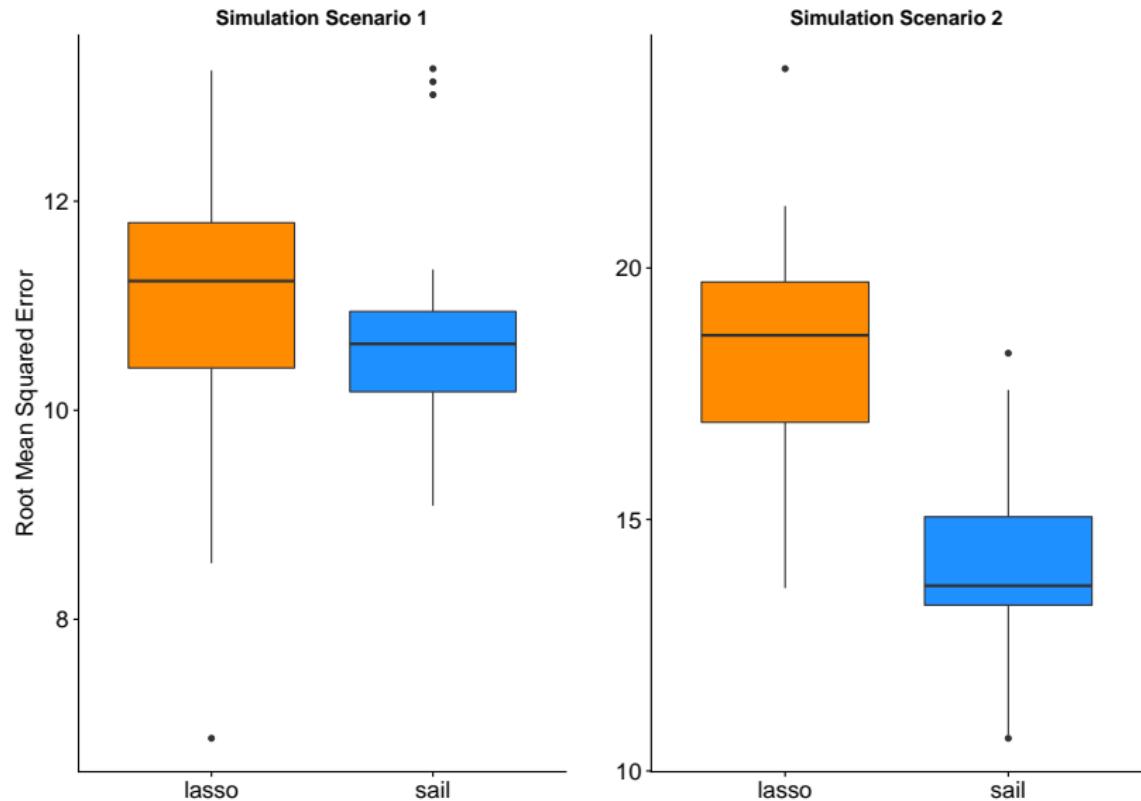
Scénario 1: «facile»

- $Y = \sum_{j=1}^5 f(X_j) + X_E + E \times (f(X_1) + f(X_2))$
- $f(\cdot)$ → B-splines avec 5 degrés de liberté
- $\theta_j \sim \mathcal{N}(0, 1)$
- $N = 200, p = 25$
- $25 \times 5 \times 2 + 1 = 251$ paramètres à estimer

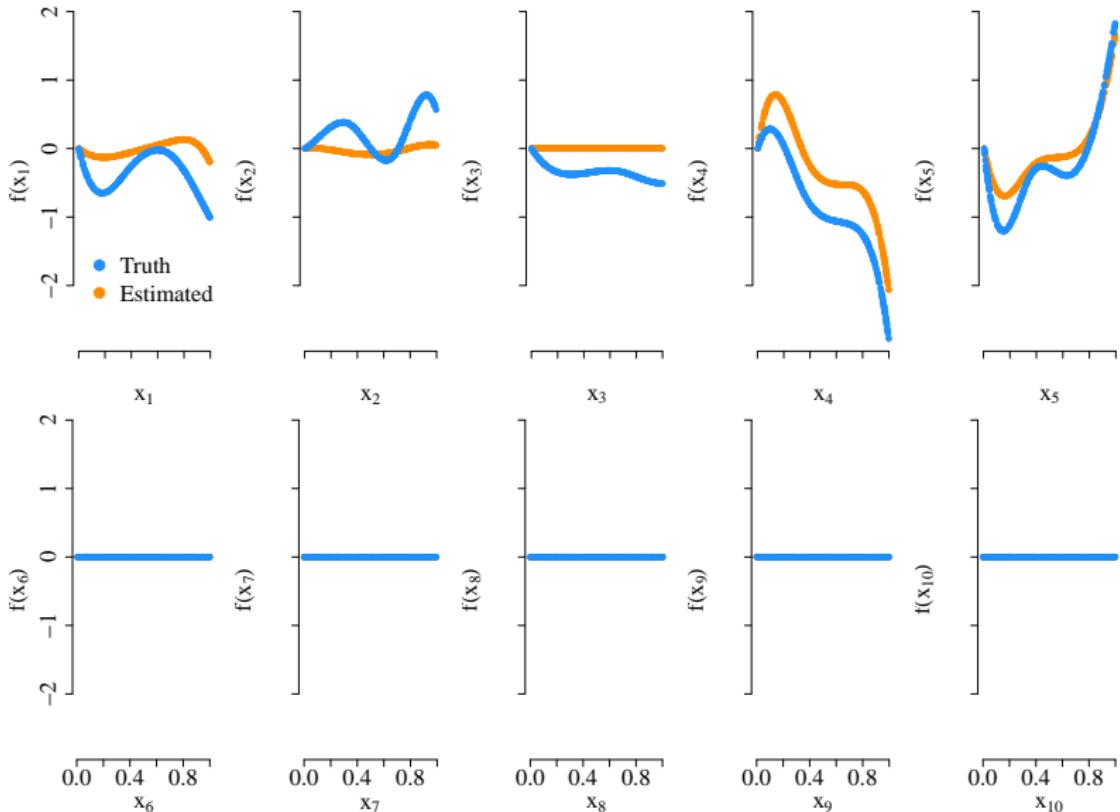
Scénario 2: «difficile»

- $Y = \sum_{j=1}^4 f(X_j) + X_E + E \times (f(X_3) + f(X_4))$
- $f(X_1)$ → linéaire
- $f(X_2)$ → quadratique
- $f(X_3)$ → sinusoïdal
- $f(X_4)$ → sinusoïdal compliqué
- $N = 200, p = 25$

ERREUR QUADRATIQUE MOYENNE (25 SIMULATIONS)

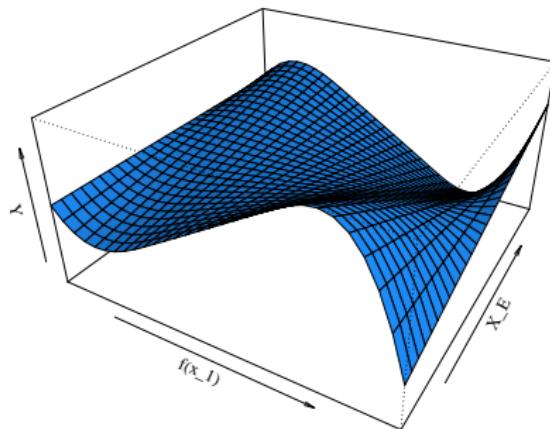


SCÉNARIO 1: LES EFFETS PRINCIPAUX

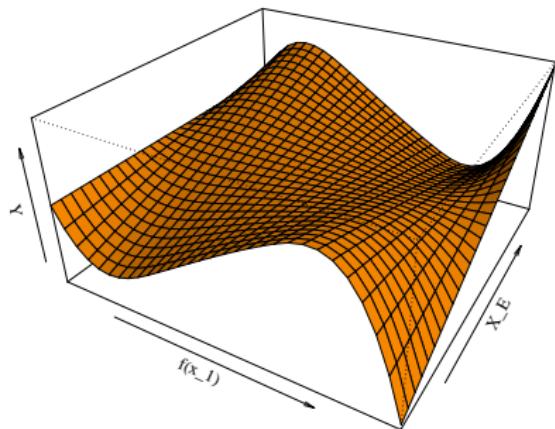


SCÉNARIO 1: LES INTERACTIONS

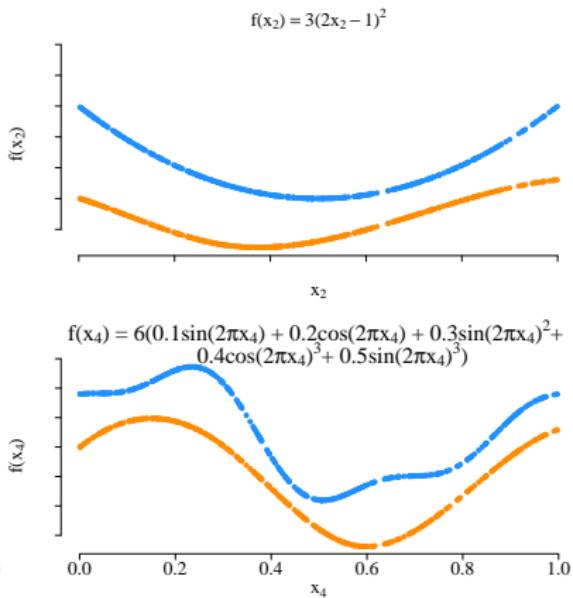
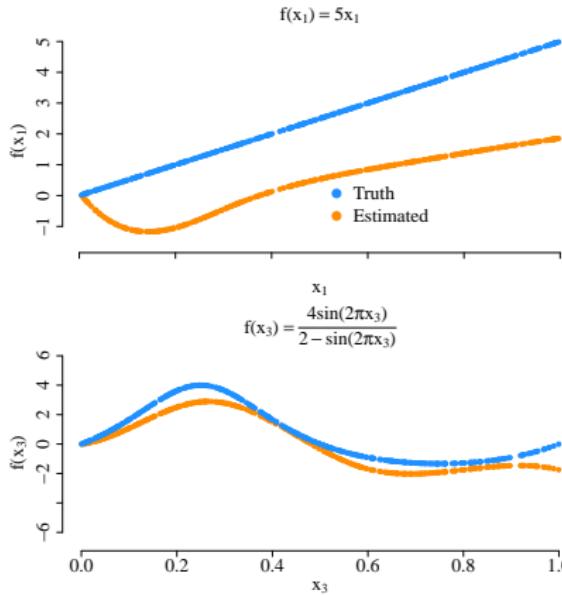
Truth



Estimated $X_E \circ f(X_I)$

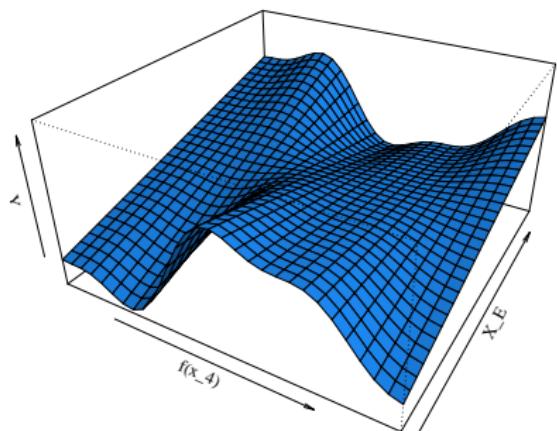


SCÉNARIO 2: LES EFFETS PRINCIPAUX

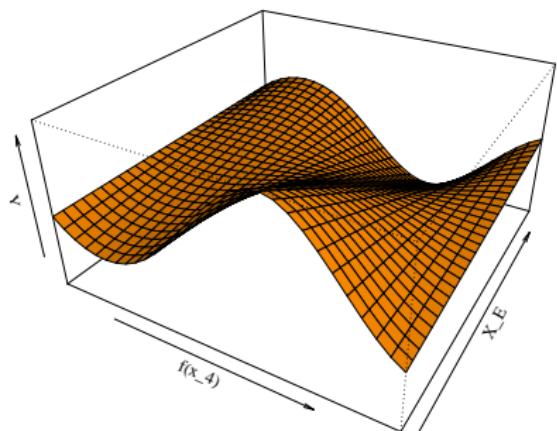


SCÉNARIO 2: LES INTERACTIONS

Truth

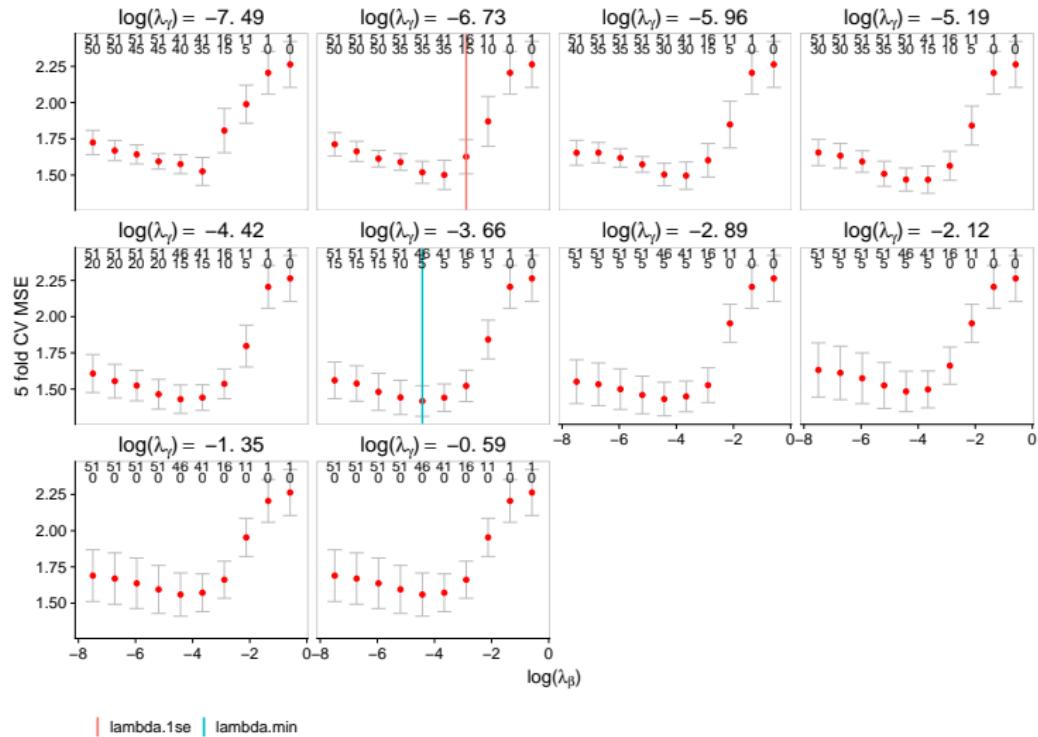


Estimated $X_E * f(X_A)$



sail PAQUET R: RÉSULTATS DE LA VALIDATION CROISÉE

```
sail:::plot(cvfit)
```



| lambda.1se | lambda.min

RÉSULTATS SUR UN VRAI JEUX DE DONNÉES

LA MALADIE D'ALZHEIMER

- En 2016, il y a environ **564 000 Canadiens** atteints de la maladie d'Alzheimer ou d'une maladie apparentée.
- À peu près 25 000 nouveaux cas chaque année.
- Selon les prévisions, il y en aura **937 000** d'ici 2031, soit une augmentation de 66 %.
- Les coûts totaux du système de soins de santé et ceux à la charge des aidants se chiffrent à un montant estimatif de **10,4 milliards de dollars par an**.
- D'ici à 2031, ce chiffre augmentera de 60 % pour atteindre **16,6 milliards de dollars**.

¹<http://www.alzheimer.ca/fr/Home/Get-involved/Advocacy/Latest-info-stats>

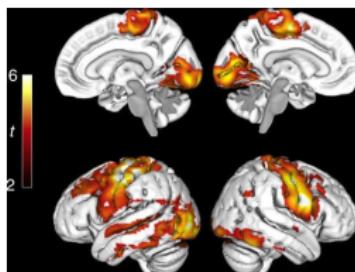
INTERACTION ENTRE LA BÊTA-AMYLOÏDE ET LE GÈNE APOE

- La présence de bêta-amyloïde sont les signes caractéristiques de la maladie d'Alzheimer.
- Il existe une corrélation importante entre le gène APOE et la maladie d'Alzheimer.

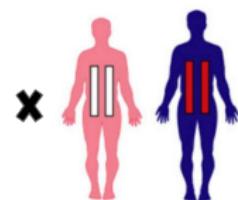
Mini-Mental State Examination



Amyloid Beta acide aminé



Gène APOE

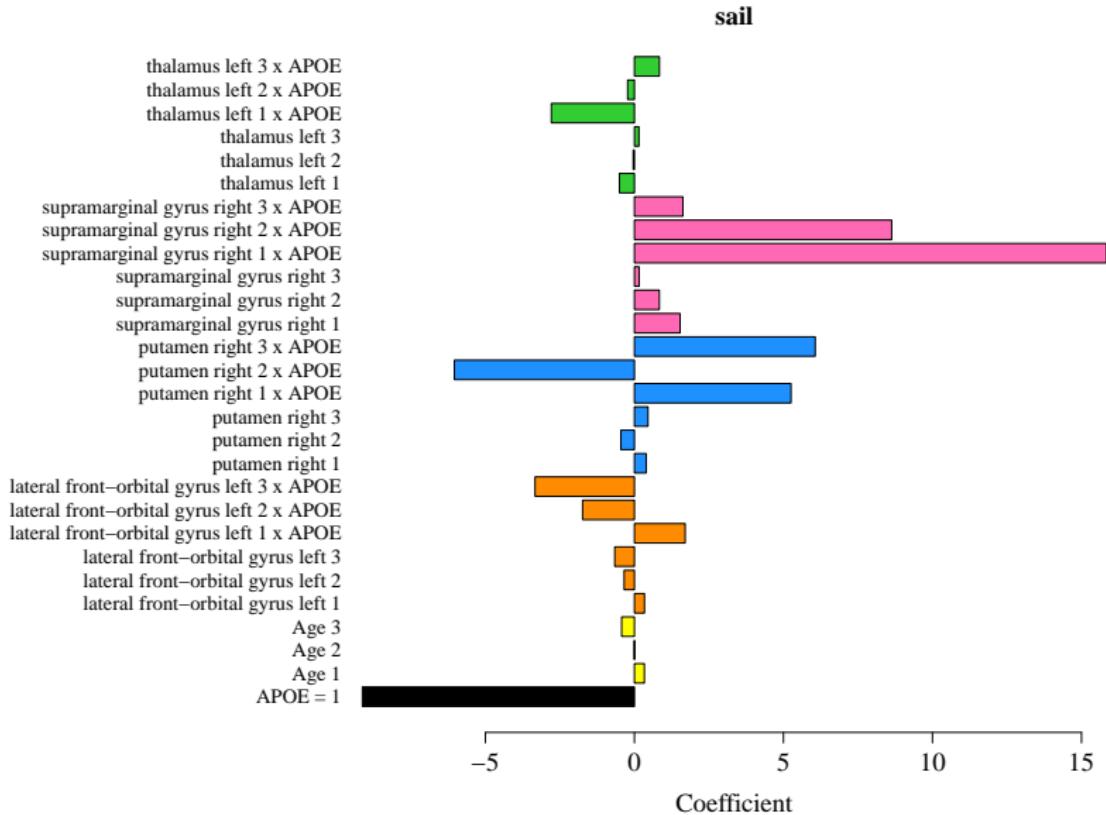


$$Y_{343 \times 1}$$

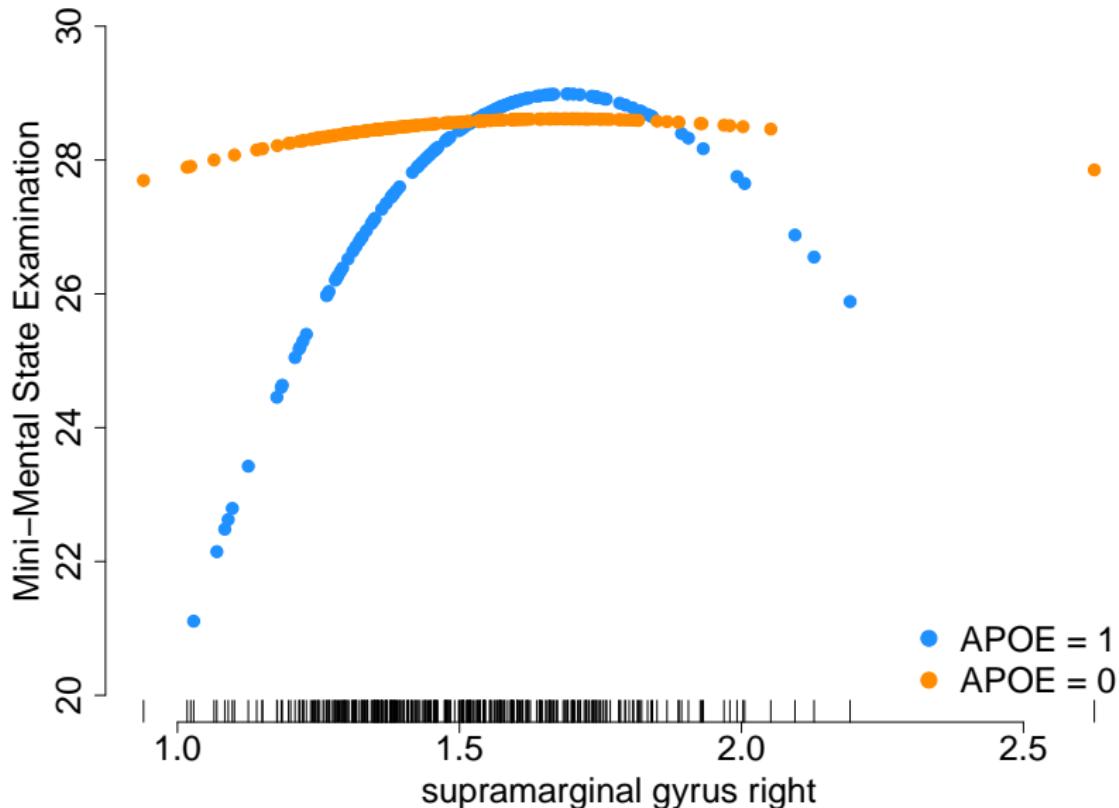
$$X_{343 \times 96}$$

$$E_{343 \times 1}$$

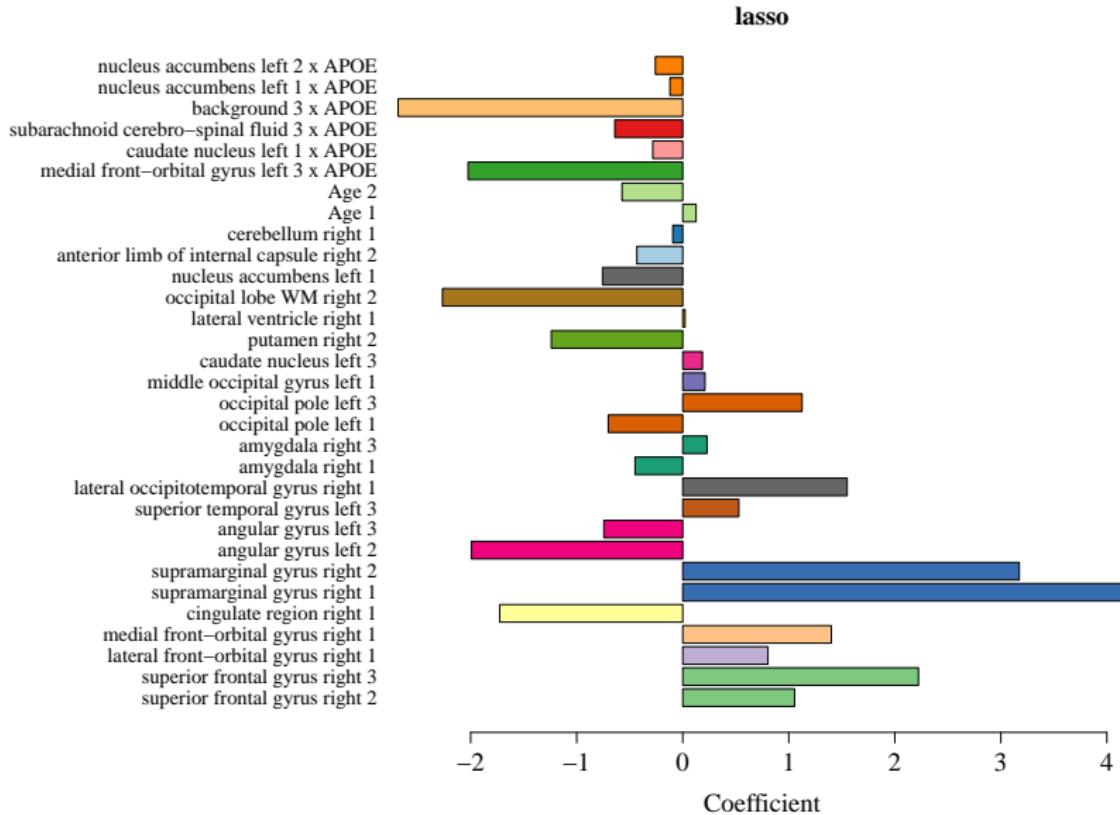
sail: LES COEFFICIENTS NON-ZÉROS



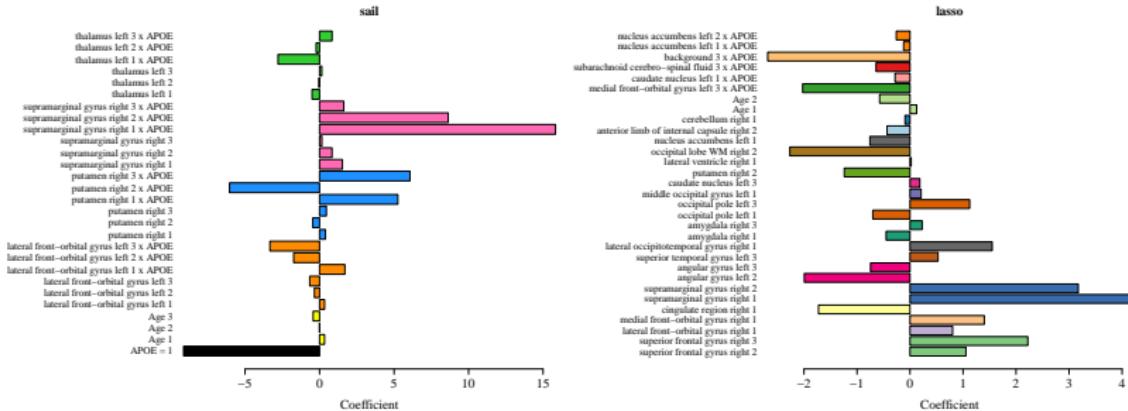
sail: INTERACTIONS AVEC LA RÉGION supramarginal gyrus



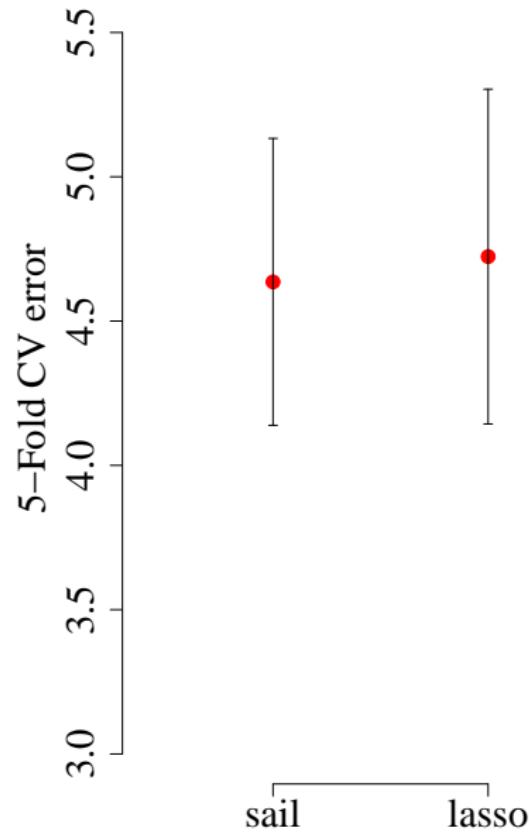
lasso: LES COEFFICIENTS NON-ZÉROS



COMPARAISON DES COEFFICIENTS: sail vs. lasso



ERREUR QUADRATIQUE MOYENNE DE LA VALIDATION CROISÉE



ggmix: LES MODÈLES MIXTES AVEC LE GROUPE LASSO

UN JEU DE DONNÉES JUSTIFICATIF: UK BIOBANK

- 500,000 individus, au-delà de 40 millions de variables explicatives.
- Grand nombre de variables réponses (ex. maladie, densité minérale osseuse).
- **Objectif:** Quelles variables explicatives sont associées à la variable réponse?



UN JEU DE DONNÉES JUSTIFICATIF: DEUX PROBLÈMES

	ID	Response	Gene1	Gene2	Gene3	Gene4	Gene5	Gene6
1	2610781	-1.255	1	2	0	0	0	1
2	4114347	-0.339	1	2	0	2	0	1
3	4399930	-0.6	1	2	1	1	0	1
4	2081319	0.809	1	2	0	1	0	2
5	1347380	0.279	2	2	0	0	0	0
6	3262449	-0.421	2	2	0	1	0	1
7	4870063	-0.454	2	2	0	0	0	2
8	1141212	1.383	2	2	1	1	1	0
9	2997954	-2.29	1	2	0	0	0	1
10	5805218	2.289	1	2	0	1	1	1

1: DES GROUPES DE VARIABLES PEUVENT AFFECTER LA VARIABLE RÉPONSE

ID	Response	Gene1	Gene2	Gene3	Gene4	Gene5	Gene6
1	2610781	-1.255	1	2	0	0	0
2	4114347	-0.339	1	2	0	2	0
3	4399930	-0.6	1	2	1	1	0
4	2081319	0.809	1	2	0	1	0
5	1347380	0.279	2	2	0	0	0
6	3262449	-0.421	2	2	0	1	0
7	4870063	-0.454	2	2	0	0	0
8	1141212	1.383	2	2	1	1	1
9	2997954	-2.29	1	2	0	0	0
10	5805218	2.289	1	2	0	1	1

2: LES INDIVIDUS SONT LIÉS

- Les observations sont corrélées, mais cette relation est inconnue.
- Cependant, elle peut être estimée à partir des données.

ID	Response	Gene1	Gene2	Gene3	Gene4	Gene5	Gene6
1 2610781	-1.255	1	2	0	0	0	1
2 4114347	-0.339	1	2	0	2	0	1
3 4399930	-0.6	1	2	1	1	0	1
4 2081319	0.809	1	2	0	1	0	2
5 1347380	0.279	2	2	0	0	0	0
6 3262449	-0.421	2	2	0	1	0	1
7 4870063	-0.454	2	2	0	0	0	2
8 1141212	1.383	2	2	1	1	1	0
9 2997954	-2.29	1	2	0	0	0	1
10 5805218	2.289	1	2	0	1	1	1

ggmix: ÉTUDE DE SIMULATION

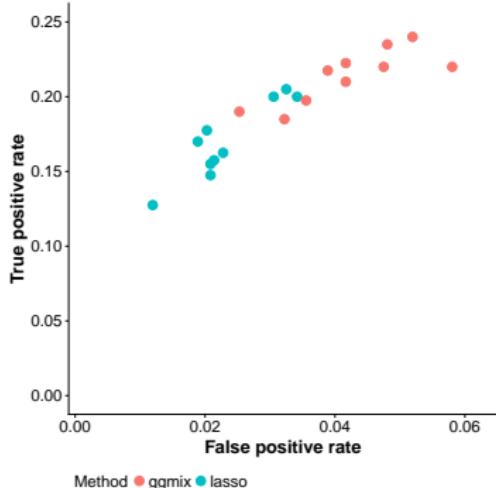


Figure 1: Vrai positif vs faux positif (10 simulations). $N = 1000$, $p = 5000$, $p_{active} = 500$

Lasso	ggmix
32.8 (0.87)	26.7 (1.06)

Table 1: Erreur quadratique moyenne (écart-type)

LES DONNÉES

- Variable réponse: $\mathbf{Y} = (y_1, \dots, y_n) \in \mathbb{R}^n$
- Variables explicatives: $\mathbf{X} = (\mathbf{x}_1, \dots, \mathbf{x}_n)^T \in \mathbb{R}^{n \times p}$, où $p \gg n$
- Matrices de similarité: $\Phi \in \mathbb{R}^{n \times n}$
- Coefficients: $\boldsymbol{\beta} = (\beta_1, \dots, \beta_p)^T \in \mathbb{R}^p$
- Effet aléatoire: $\mathbf{b} = (b_1, \dots, b_n) \in \mathbb{R}^n$
- Erreur: $\boldsymbol{\varepsilon} = (\varepsilon_1, \dots, \varepsilon_n) \in \mathbb{R}^n$

LES DONNÉES

- Variable réponse: $\mathbf{Y} = (y_1, \dots, y_n) \in \mathbb{R}^n$
- Variables explicatives: $\mathbf{X} = (\mathbf{x}_1, \dots, \mathbf{x}_n)^T \in \mathbb{R}^{n \times p}$, où $p \gg n$
- Matrices de similarité: $\Phi \in \mathbb{R}^{n \times n}$
- Coefficients: $\boldsymbol{\beta} = (\beta_1, \dots, \beta_p)^T \in \mathbb{R}^p$
- Effet aléatoire: $\mathbf{b} = (b_1, \dots, b_n) \in \mathbb{R}^n$
- Erreur: $\boldsymbol{\varepsilon} = (\varepsilon_1, \dots, \varepsilon_n) \in \mathbb{R}^n$

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{b} + \boldsymbol{\varepsilon} \quad (2)$$

$$\mathbf{b} \sim \mathcal{N}(0, \eta\sigma^2\Phi) \quad \boldsymbol{\varepsilon} \sim \mathcal{N}(0, (1 - \eta)\sigma^2\mathbf{I})$$

- σ^2 et $\eta \in [0, 1]$ partagent la variance entre \mathbf{b} et $\boldsymbol{\varepsilon}$
- $\mathbf{Y}|(\boldsymbol{\beta}, \eta, \sigma^2) \sim \mathcal{N}(\mathbf{X}\boldsymbol{\beta}, \eta\sigma^2\Phi + (1 - \eta)\sigma^2\mathbf{I})$

LA FONCTION DE VRAISEMBLANCE

- Le log de la vraisemblance négative est donné par:

$$-\ell(\Theta) \propto \frac{n}{2} \log(\sigma^2) + \frac{1}{2} \log(\det(V)) + \frac{1}{2\sigma^2} (Y - X\beta)^T V^{-1} (Y - X\beta)$$

où $V = \eta\Phi + (1 - \eta)\mathcal{I}$ ¹

¹Pirinen et al. Annals of Applied Statistics (2013)

LA FONCTION DE VRAISEMBLANCE

- Le log de la vraisemblance négative est donné par:

$$-\ell(\Theta) \propto \frac{n}{2} \log(\sigma^2) + \frac{1}{2} \log (\det(V)) + \frac{1}{2\sigma^2} (Y - X\beta)^T V^{-1} (Y - X\beta)$$

où $V = \eta\Phi + (1 - \eta)\mathcal{I}$ ¹

- Supposons que nous avons une matrice de rang inférieur $K \in \mathbb{R}^{n \times k}$ ($k < n$), pour calculer la matrice de similarité factorisée $\Phi = KK^T$

¹Pirinen et al. Annals of Applied Statistics (2013)

LA FONCTION DE VRAISEMBLANCE

- Le log de la vraisemblance négative est donné par:

$$-\ell(\Theta) \propto \frac{n}{2} \log(\sigma^2) + \frac{1}{2} \log(\det(V)) + \frac{1}{2\sigma^2} (Y - X\beta)^T V^{-1} (Y - X\beta)$$

où $V = \eta\Phi + (1 - \eta)\mathcal{I}$ ¹

- Supposons que nous avons une matrice de rang inférieur $K \in \mathbb{R}^{n \times k}$ ($k < n$), pour calculer la matrice de similarité factorisée $\Phi = KK^T$
- Soit $K = U\Lambda V^T$ la décomposition en valeurs singulières de K , alors

$$\Phi = U_1\Lambda\Lambda U_1^T = U_1\Sigma U_1^T$$

où $U_1 \in \mathbb{R}^{n \times k}$ est la matrice composée des vecteurs propres correspondant au k valeurs propres différentes de zéro.

¹Pirinen et al. Annals of Applied Statistics (2013)

ESTIMATEUR DE VRAISEMBLANCE PÉNALISÉ

- Le log de la vraisemblance négative peut alors être exprimé

$$-\ell(\Theta) \propto \frac{n}{2} \log(\sigma^2) + \frac{1}{2} \left(\sum_{i=1}^k \log(1 + \eta(\Sigma_i - 1)) + (n - k) \log(1 - \eta) \right) + \frac{1}{2} \left\{ (\mathbf{Y} - \mathbf{X}\boldsymbol{\beta})^T \left[\frac{1}{\sigma^2(1 - \eta)} \left(\mathbf{I}_n - \mathbf{U}_1 \left(\frac{1 - \eta}{\eta} \Sigma_1^{-1} + \mathbf{I}_k \right)^{-1} \mathbf{U}_1^T \right) \right] (\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}) \right\}$$

ESTIMATEUR DE VRAISEMBLANCE PÉNALISÉ

- Le log de la vraisemblance négative peut alors être exprimé

$$-\ell(\Theta) \propto \frac{n}{2} \log(\sigma^2) + \frac{1}{2} \left(\sum_{i=1}^k \log(1 + \eta(\Sigma_i - 1)) + (n - k) \log(1 - \eta) \right) + \frac{1}{2} \left\{ (\mathbf{Y} - \mathbf{X}\boldsymbol{\beta})^T \left[\frac{1}{\sigma^2(1 - \eta)} \left(\mathbf{I}_n - \mathbf{U}_1 \left(\frac{1 - \eta}{\eta} \Sigma_1^{-1} + \mathbf{I}_k \right)^{-1} \mathbf{U}_1^T \right) \right] (\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}) \right\}$$

- Définir la fonction de perte:

$$Q_\lambda(\Theta) = -\ell(\Theta) + \lambda \sum_j P_j(\beta_j)$$

- $P_j(\cdot)$ est la pénalité sur les $\beta_1, \dots, \beta_{p+1}$
- L'estimateur $\hat{\Theta}_\lambda$ est obtenu par

$$\hat{\Theta}_\lambda = \arg \min_{\Theta} Q_\lambda(\Theta)$$

GROUPE LASSO POUR LES MODÈLES MIXTES

- Soit les variables explicatives $X \in \mathbb{R}^{n \times p}$ appartenant à un des K groupes distinct prédéfini de taille p_k .
- $\beta_{(k)}$ est la parti de β qui correspond à la groupe k

GROUPE LASSO POUR LES MODÈLES MIXTES

- Soit les variables explicatives $\mathbf{X} \in \mathbb{R}^{n \times p}$ appartenant à un des K groupes distinct prédéfini de taille p_k .
- $\boldsymbol{\beta}_{(k)}$ est la parti de $\boldsymbol{\beta}$ qui correspond à la groupe k
- Nous considérons l'estimateur pénalisé par le groupe lasso

$$\min_{\boldsymbol{\beta}} L(\boldsymbol{\beta} | \mathsf{D}) + \lambda \sum_{k=1}^K w_k \|\boldsymbol{\beta}_{(k)}\|_2, \quad (3)$$

- où

$$L(\boldsymbol{\beta} | \mathsf{D}) = \frac{1}{2} [\mathbf{Y} - \widehat{\mathbf{Y}}]^T \mathbf{W} [\mathbf{Y} - \widehat{\mathbf{Y}}] \quad (4)$$

$\widehat{\mathbf{Y}} = \sum_{j=1}^p \beta_j \mathbf{x}_j$, D représente les données courants $\{\mathbf{Y}, \mathbf{X}\}$, et

$$\mathbf{W}_{n \times n} = \frac{1}{\sigma^2(1-\eta)} \left(\mathbf{I}_n - \mathbf{U}_1 \left(\frac{1-\eta}{\eta} \Sigma_1^{-1} + \mathbf{I}_k \right)^{-1} \mathbf{U}_1^T \right) \quad (5)$$

DESCENTE GROUPÉE: EXPLOITATION DE LA SPARSITÉ

On minimise la fonction de perte

$$\frac{1}{2} [\mathbf{Y} - \widehat{\mathbf{Y}}]^\top \mathbf{W} [\mathbf{Y} - \widehat{\mathbf{Y}}] + \lambda \sum_{k=1}^K w_k \|\boldsymbol{\beta}^{(k)}\|_2$$

Au cours de chaque sous-itération, optimisez $\boldsymbol{\beta}^{(k)}$. Révisez $\boldsymbol{\beta}^{(k')} = \widetilde{\boldsymbol{\beta}}^{(k')}$ pour $k' \neq k$ à leur valeurs courantes.

1. Initialisation: $\widetilde{\boldsymbol{\beta}}$
2. Descente cyclique en groupe: pour $k = 1, 2, \dots, K$, révisez $\boldsymbol{\beta}^{(k)}$ en minimisant la fonction de perte

$$\widetilde{\boldsymbol{\beta}}^{(k)}(\text{new}) \leftarrow \arg \min_{\boldsymbol{\beta}^{(k)}} L(\boldsymbol{\beta} \mid \mathbf{D}) + \lambda w_k \|\boldsymbol{\beta}^{(k)}\|_2$$

3. Réitérez (2) jusqu'à la convergence.

LA CONDITION <>

$$\arg \min_{\beta^{(k)}} \frac{1}{2} [\mathbf{Y} - \hat{\mathbf{Y}}]^T \mathbf{W} [\mathbf{Y} - \hat{\mathbf{Y}}] + \lambda \sum_{k=1}^K w_k \|\beta^{(k)}\|_2 \quad (6)$$

- Malheureusement, il n'y a pas de forme explicite pour (6)

¹Yang and Zou. Statistical Computing (2014)

LA CONDITION <>

$$\arg \min_{\beta^{(k)}} \frac{1}{2} [\mathbf{Y} - \hat{\mathbf{Y}}]^T \mathbf{W} [\mathbf{Y} - \hat{\mathbf{Y}}] + \lambda \sum_{k=1}^K w_k \|\beta^{(k)}\|_2 \quad (6)$$

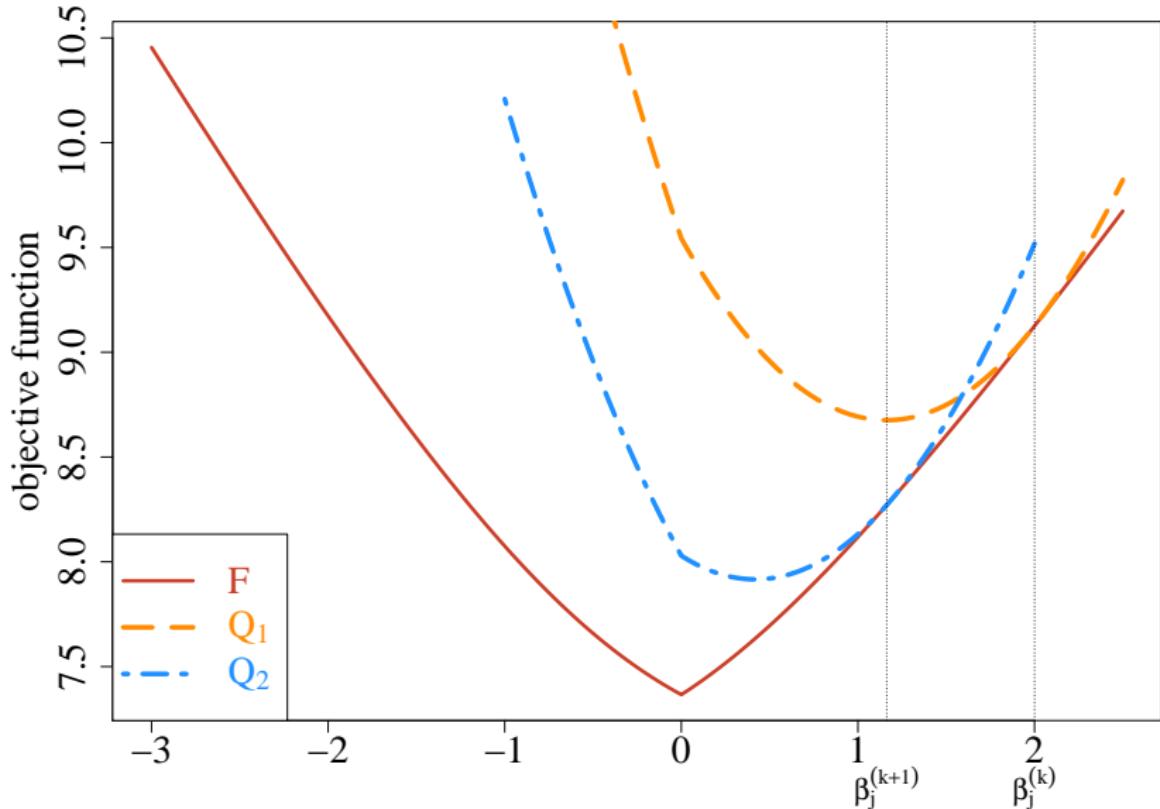
- Malheureusement, il n'y a pas de forme explicite pour (6)
- Cependant, la fonction de perte $L(\beta | D)$ satisfait à la condition «quadratic majorization» (QM)¹, puisqu'il existe
 - une matrice $p \times p$, $\mathbf{H} = \mathbf{X}^T \mathbf{W} \mathbf{X}$, et
 - $\nabla L(\beta | D) = -(\mathbf{Y} - \hat{\mathbf{Y}})^T \mathbf{W} \mathbf{X}$

qui dépend seulement des données D , tel que pour tous β, β^* ,

$$L(\beta | D) \leq L(\beta^* | D) + (\beta - \beta^*)^T \nabla L(\beta^* | D) + \frac{1}{2} (\beta - \beta^*)^T \mathbf{H} (\beta - \beta^*)$$

¹Yang and Zou. Statistical Computing (2014)

DESCENTE DE COORDONNÉES GÉNÉRALISÉE (GCD)



DESCENTE DE COORDONNÉES GÉNÉRALISÉE PAR GROUPE

- Révisez β par groupe

$$\beta - \tilde{\beta} = (\underbrace{0, \dots, 0}_{k-1}, \beta^{(k)} - \tilde{\beta}^{(k)}, \underbrace{0, \dots, 0}_{K-k})$$

DESCENTE DE COORDONNÉES GÉNÉRALISÉE PAR GROUPE

- Révisez β par groupe

$$\beta - \tilde{\beta} = (\underbrace{0, \dots, 0}_{k-1}, \beta^{(k)} - \tilde{\beta}^{(k)}, \underbrace{0, \dots, 0}_{K-k})$$

- Il suffit alors de calculer la fonction de majoration par groupe

$$L(\beta | D) \leq L(\tilde{\beta} | D) - (\beta^{(k)} - \tilde{\beta}^{(k)})^\top U^{(k)} + \frac{1}{2} \gamma_k (\beta^{(k)} - \tilde{\beta}^{(k)})^\top (\beta^{(k)} - \tilde{\beta}^{(k)})$$

$$U^{(k)} = \frac{\partial}{\partial \beta^{(k)}} L(\beta | D) = -(\gamma - \hat{\gamma})^\top W X_{(k)}$$

$$H^{(k)} = \frac{\partial^2}{\partial \beta^{(k)} \partial \beta_{(k)}^\top} L(\beta | D) = X_{(k)}^\top W X_{(k)}$$

- $\gamma_k = \text{eigen}_{\max}(H^{(k)})$

DESCENTE DE COORDONNÉES GÉNÉRALISÉE PAR GROUPE

- Révisez β par groupe

$$\beta - \tilde{\beta} = (\underbrace{0, \dots, 0}_{k-1}, \beta^{(k)} - \tilde{\beta}^{(k)}, \underbrace{0, \dots, 0}_{K-k})$$

- Il suffit alors de calculer la fonction de majoration par groupe

$$L(\beta | D) \leq L(\tilde{\beta} | D) - (\beta^{(k)} - \tilde{\beta}^{(k)})^\top U^{(k)} + \frac{1}{2} \gamma_k (\beta^{(k)} - \tilde{\beta}^{(k)})^\top (\beta^{(k)} - \tilde{\beta}^{(k)})$$

$$U^{(k)} = \frac{\partial}{\partial \beta^{(k)}} L(\beta | D) = -(\gamma - \hat{\gamma})^\top W X_{(k)}$$

$$H^{(k)} = \frac{\partial^2}{\partial \beta^{(k)} \partial \beta_{(k)}^\top} L(\beta | D) = X_{(k)}^\top W X_{(k)}$$

- $\gamma_k = \text{eigen}_{\max}(H^{(k)})$

- Révisez $\tilde{\beta}^{(k)}$ qui a une forme explicite:

$$\tilde{\beta}^{(k)}(\text{new}) = \frac{1}{\gamma_k} \left(U^{(k)} + \gamma_k \tilde{\beta}^{(k)} \right) \left(1 - \frac{\lambda w_k}{\| U^{(k)} + \gamma_k \tilde{\beta}^{(k)} \|_2} \right)_+$$

DISCUSSION

Forces

- Sélection des interactions non-linéaires en conservant la propriété de l'hérédité forte quand $p \gg N$.
- `sail` permet une modélisation flexible des variables explicatives.
- `gmmix` est la première implémentation du groupe lasso pour les modèles mixtes.

FORCES ET FAIBLESSES

Forces

- Sélection des interactions non-linéaires en conservant la propriété de l'hérédité forte quand $p \gg N$.
- **sail** permet une modélisation flexible des variables explicatives.
- **gmmix** est la première implémentation du groupe lasso pour les modèles mixtes.

Faiblesses

- **sail** peut actuellement gérer seulement $E \cdot f(X)$ ou $f(E) \cdot X$.
- Ne permet pas $f(X_1, E)$, ni $f(X_1, X_2)$.
- L'implémentation actuelle de **sail** est lente en raison de la validation croisée pour les 2 paramètres de réglage.
- L'empreinte mémoire est un problème pour **sail** et **gmmix**

- Est-ce que les deux paramètres de réglage sont vraiment nécessaires ?

$$\lambda \left\{ (1 - \alpha) \left[w_E |\beta_E| + \sum_{j=1}^p w_j \|\theta_j\|_2 \right] + \alpha \sum_{j=1}^p w_{jE} |\gamma_j| \right\}$$

- Faible propriété héréditaire $\rightarrow \alpha_j = \gamma_j(|\beta_j| + |\beta_E|)$.
- Implémenter l'algorithme ADMM. Calcul distribué (GPU).
- Autres pénalités (SCAD, MCP).
- Variable réponse binaire.

REMERCIEMENTS



RÉFÉRENCES

- Radchenko, P., & James, G. M. (2010). Variable selection using adaptive nonlinear interaction structures in high dimensions. *Journal of the American Statistical Association*, 105(492), 1541-1553.
- Choi, N. H., Li, W., & Zhu, J. (2010). Variable selection with the strong heredity constraint and its oracle property. *Journal of the American Statistical Association*, 105(489), 354-364.
- Chipman, H. (1996). Bayesian variable selection with related predictors. *Canadian Journal of Statistics*, 24(1), 17-36.
- Friedman, J., Hastie, T., & Tibshirani, R. (2010). Regularization paths for generalized linear models via coordinate descent. *Journal of statistical software*, 33(1)
- Yang, Y., & Zou, H. (2015). A fast unified algorithm for solving group-lasso penalize learning problems. *Statistics and Computing*, 25(6), 1129-1141
- De Leeuw, J. (1994). Block-relaxation algorithms in statistics. In *Information systems and data analysis* (pp. 308-324). Springer Berlin Heidelberg.

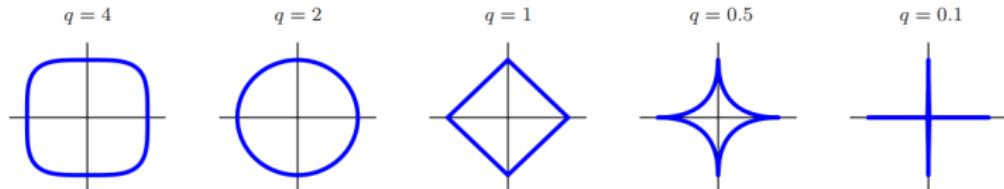
APPENDIX

WHY THE L1 NORM ?

- For a fixed real number $q \geq 0$ consider the criterion

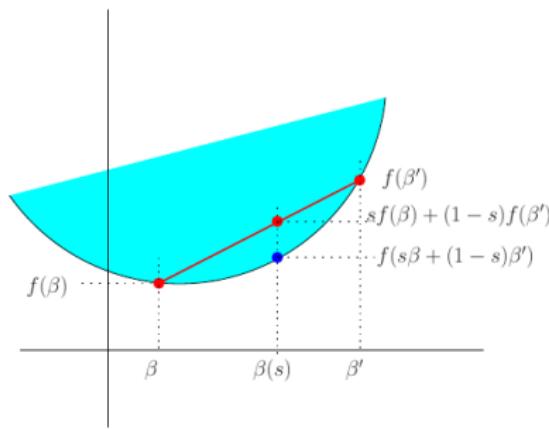
$$\tilde{\beta} = \arg \min_{\beta} \left\{ \sum_{i=1}^n \left(y_i - \beta_0 - \sum_{j=1}^p x_{ij} \beta_j \right)^2 + \lambda \sum_{j=1}^p |\beta_j|^q \right\}$$

- Why do we use the ℓ_1 norm? Why not use the $q = 2$ (Ridge) or any ℓ_q norm?

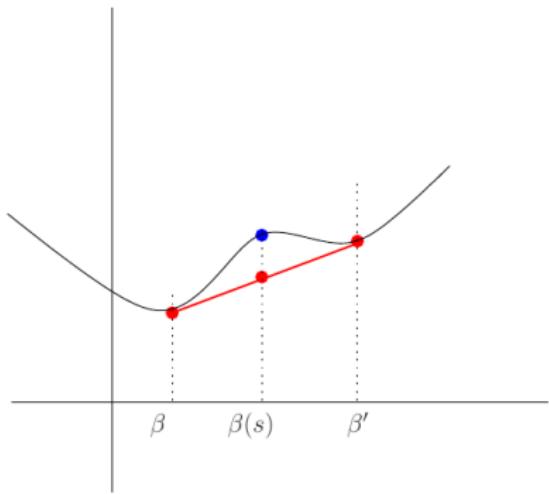


- $q = 1$ is the smallest value that yields a sparse solution **and** yields a **convex** problem → scalable to high-dimensional data
- For $q < 1$ the constrained region is **nonconvex**

CONVEX FUNCTION



(a)

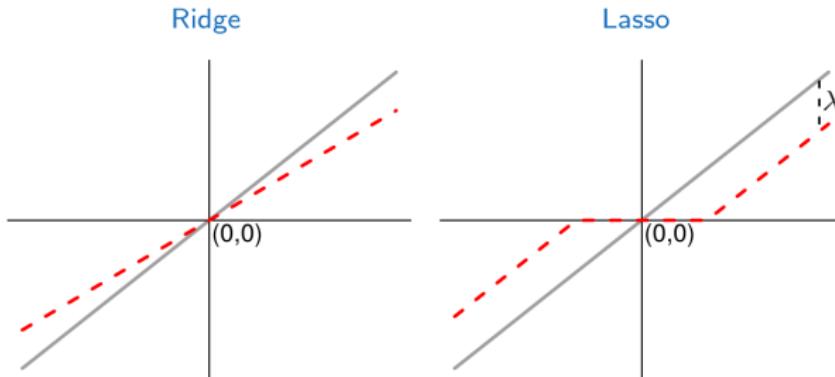


(b)

RIDGE VS. LASSO

- Estimators of β_j in terms of the least-squares estimate $\hat{\beta}_j^{LS}$ for an orthonormal model matrix X

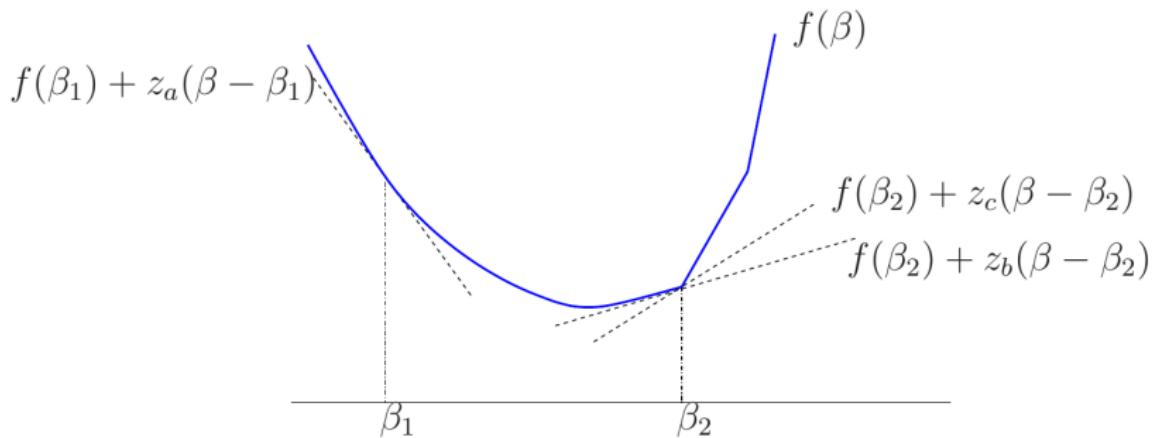
q	Estimator	Formula
1	Lasso	$\text{sign}(\hat{\beta}_j^{LS})(\hat{\beta}_j^{LS} - \lambda)_+$
2	Ridge	$\hat{\beta}_j^{LS}/(1 + \lambda)$



SUBGRADIENT

- A basic property of differentiable convex functions is that the first-order tangent approximation always provides a lower bound.
- The notion of subgradient is based on a natural generalization of this idea. In particular, given a convex function $f : \mathbb{R}^p \rightarrow \mathbb{R}$, a vector $z \in \mathbb{R}^p$ is said to be a *subgradient* of f at β if

$$f(\beta') \geq f(\beta) + \langle z, \beta' - \beta \rangle, \quad \text{for all } \beta' \in \mathbb{R}^p$$



ALGORITHM FOR ggmix

Algorithm 2: Block Relaxation Algorithm

Set the iteration counter $k \leftarrow 0$, initial values for the parameter vector $\Theta^{(0)}$ and convergence threshold ϵ ;

for $\lambda \in \{\lambda_{\max}, \dots, \lambda_{\min}\}$ **do**

repeat

$$\boldsymbol{\beta}^{(k+1)} \leftarrow \arg \min_{\boldsymbol{\beta}} Q_{\lambda} \left(\boldsymbol{\beta}, \eta^{(k)}, \sigma^2^{(k)} \right)$$

$$\eta^{(k+1)} \leftarrow \arg \min_{\eta} Q_{\lambda} \left(\boldsymbol{\beta}^{(k+1)}, \eta, \sigma^2^{(k)} \right)$$

$$\sigma^2^{(k+1)} \leftarrow \arg \min_{\sigma^2} Q_{\lambda} \left(\boldsymbol{\beta}^{(k+1)}, \eta^{(k+1)}, \sigma^2 \right)$$

$k \leftarrow k + 1$

until convergence criterion is satisfied: $\|\Theta^{(k+1)} - \Theta^{(k)}\|_2 < \epsilon$;

end

COORDINATE DESCENT

- Minimize the objective function

$$\min_{\beta} F(\beta) \equiv L(\beta_1, \dots, \beta_p) + p_\lambda(\beta)$$

- $L : \mathbb{R}^p \rightarrow \mathbb{R}$ is differentiable convex, $p_\lambda : \mathbb{R}^p \rightarrow \mathbb{R}$ bounded from below but not necessarily smooth. e.g. $p_\lambda(\beta) = \lambda \sum_{j=1}^p |\beta_j|$

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

1. Initialization: $\tilde{\beta}$
2. Cyclic coordinate descent: for $j = 1, 2, \dots, p$, update β_j by minimizing the objective function

$$\tilde{\beta}_j^{new} \leftarrow \arg \min_{\beta_j} F(\beta_j | \beta_k = \tilde{\beta}_k, k \neq j)$$

3. Repeat (2) till convergence.

QUADRATIC MAJORIZATION CONDITION

Empirical loss

$$L(\beta \mid D) = \frac{1}{n} \sum_{i=1}^n \Phi(y_i, \beta^\top x_i)$$

Φ satisfies the [QM condition](#), if and only if:

1. $L(\beta \mid D)$ is [differentiable](#) as a function of β .
2. Can find H (p by p), may only depend on the data D , such that for all β, β^*

$$L(\beta \mid D) \leq L(\beta^* \mid D) + (\beta - \beta^*)^\top \nabla L(\beta^* \mid D) + \frac{1}{2} (\beta - \beta^*)^\top H (\beta - \beta^*)$$

Loss	$-\nabla L(\beta \mid D)$	H
Least squares	$\frac{1}{n} \sum_{i=1}^n (y_i - x_i^\top \beta) x_i$	$X^\top X / n$
Logistic regression	$\frac{1}{n} \sum_{i=1}^n y_i x_i \frac{1}{1 + \exp(y_i x_i^\top \beta)}$	$\frac{1}{4} X^\top X / n$
Squared hinge loss	$\frac{1}{n} \sum_{i=1}^n 2y_i x_i (1 - y_i x_i^\top \beta)_+$	$4X^\top X / n$
Huberized hinge loss	$\frac{1}{n} \sum_{i=1}^n y_i x_i \text{hsvm}'(y_i x_i^\top \beta)$	$\frac{2}{\delta} X^\top X / n$

VERIFYING QUADRATIC MAJORIZATION CONDITION

Lemma 1 (Yang and Zou, 2014)

Assume $\Phi(y, f)$ is differentiable with respect to f and write

$$\Phi'_f = \frac{\partial \Phi(y, f)}{\partial f}, \quad \nabla L(\boldsymbol{\beta} | D) = \frac{1}{n} \sum_{i=1}^n \Phi'_f(y_i, \mathbf{x}_i^\top \boldsymbol{\beta}) \mathbf{x}_i.$$

1. If Φ'_f is Lipschitz continuous with constant C such that

$$|\Phi'_f(y, f_1) - \Phi'_f(y, f_2)| \leq C|f_1 - f_2| \quad \forall y, f_1, f_2,$$

then the QM condition holds for Φ and $H = \frac{2C}{n} \mathbf{X}^\top \mathbf{X}$.

2. If $\Phi''_f = \frac{\partial \Phi^2(y, f)}{\partial f^2}$ exists and

$$\Phi''_f \leq C_2 \quad \forall y, f,$$

then the QM condition holds for Φ and $H = \frac{C_2}{n} \mathbf{X}^\top \mathbf{X}$.

STRICT DESCENT PROPERTY OF GMD

Proposition (Yang and Zou, 2014)

- If $\tilde{\beta}^{(k)}$ (new) $\neq \tilde{\beta}^{(k)}$ then

$$L(\tilde{\beta}^{(k)} \text{ (new)} | D) + \lambda w_k \|\tilde{\beta}^{(k)} \text{ (new)}\|_2 < L(\tilde{\beta} | D) + \lambda w_k \|\tilde{\beta}^{(k)}\|_2$$

the objective function is strictly decreased after updating all k in a cycle.

- If $\tilde{\beta}^{(k)}$ (new) $= \tilde{\beta}^{(k)}$ for all k , then the solution must satisfy the KKT conditions:

$$-U^{(k)} + \lambda w_k \cdot \frac{\tilde{\beta}^{(k)}}{\|\tilde{\beta}^{(k)}\|_2} = 0 \quad \text{if } \tilde{\beta}^{(k)} \neq 0,$$

$$\left\| U^{(k)} \right\|_2 \leq \lambda w_k \quad \text{if } \tilde{\beta}^{(k)} = 0,$$

which means that the algorithm converges and finds the right answer.

SEPARABILITY AND COORDINATE DESCENT

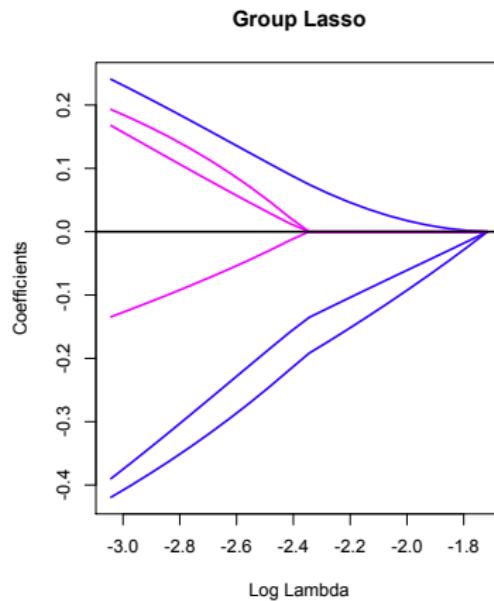
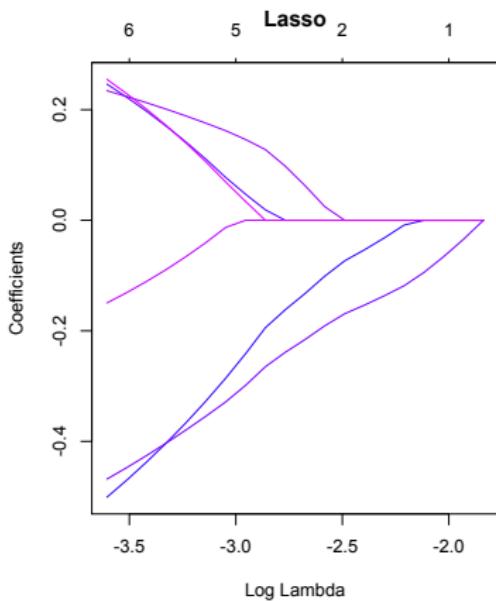
$$f(\beta_1, \dots, \beta_p) = g(\beta_1, \dots, \beta_p) + \sum_{j=1}^p h_j(\beta_j)$$

- $g : \mathbb{R}^p \rightarrow \mathbb{R}$ is differentiable and convex, and the univariate functions $h_j : \mathbb{R} \rightarrow \mathbb{R}$ are convex **but not necessarily differentiable**.
- Tseng (2001) shows that for any convex cost function f with separable structure, the coordinate descent algorithm is guaranteed to converge to the global minimizer
- The key property underlying this result is the separability of the nondifferentiable component $h(\beta) = \sum_{j=1}^p h_j(\beta_j)$, as a sum of functions of each individual parameter.

¹Tseng. Journal of optimization theory and applications (2001)

LASSO VS. GROUP LASSO

- Logistic regression with group lasso: $n = 50, p = 6$.
- Group lasso: specify $(\beta_1, \beta_2, \beta_3), (\beta_4, \beta_5, \beta_6)$. Variable selection at the group level.
- Solution path: view β as function of λ .



GROUP LASSO MOTIVATION

- Categorical predictors (factors): dummy variables
- Additive Model: $\sum_{k=1}^K f_k(x^{(k)}) \approx \sum_{k=1}^K \sum_{m=1}^M \beta_{km} h_m(x^{(k)})$
 - ex. birth weight predicted by the mother's age and weight, Age, Age², Age³ and Weight, Weight², Weight³

Group lasso partitions the variable coefficients into K groups

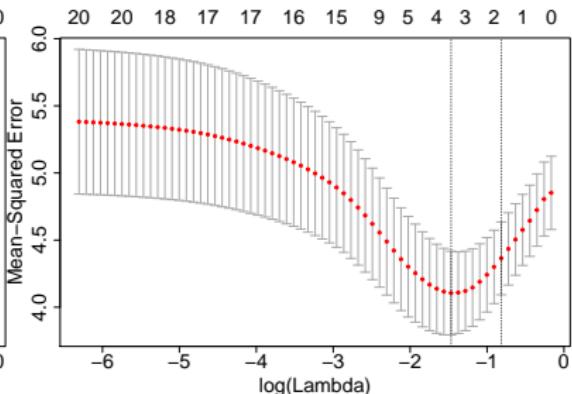
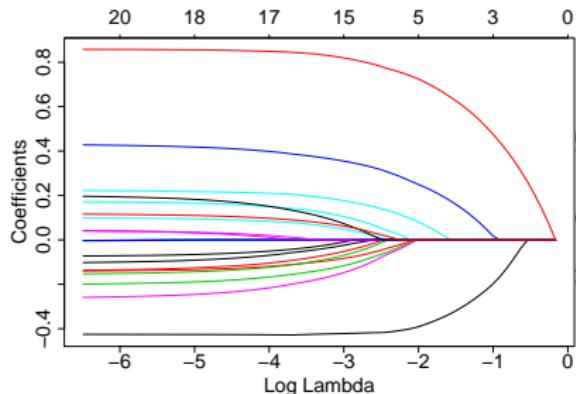
$$\boldsymbol{\beta} = ([\boldsymbol{\beta}^{(1)}]^\top, [\boldsymbol{\beta}^{(2)}]^\top, \dots, [\boldsymbol{\beta}^{(K)}]^\top)^\top$$

Extended from the lasso penalty, the group lasso estimator is:

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

CHOOSING MODEL COMPLEXITY

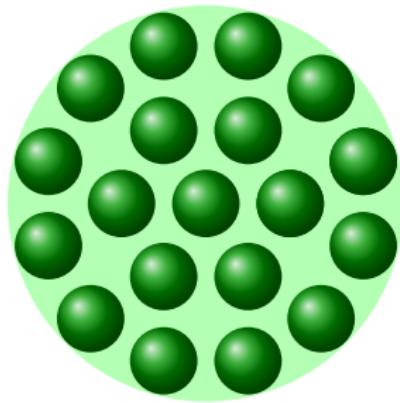
- The tuning parameter λ controls the **complexity** of the model
- **Generalization ability of the model:** we select the λ that gives the most accurate model for predicting independent test data from the same population



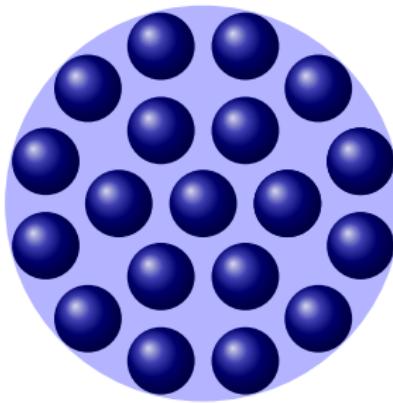
SIMULATION STUDY

- $\mathbf{X}^{(test)}$: 4k SNPs from UK Biobank genotyped data from 1k randomly sampled individuals
- $\mathbf{X}^{(causal)}$: random sample of 400 SNPs from $\mathbf{X}^{(test)}$
- $\mathbf{X}^{(kinship)}$: random sample of 4k SNPs used to construct the kinship matrix (Φ) including the causal SNPs
- $\beta_j \sim Unif(0.1, 0.3)$ for $j = 1, \dots, 400$
- $Y = \beta_0 + \sum_{j=1}^{400} \beta_j \mathbf{x}_j^{(causal)} + P + E$
- $P \sim \mathcal{N}(0, 0.6 \cdot \Phi_{n \times n})$, $E \sim \mathcal{N}(0, 0.4 \cdot \mathbf{I}_{n \times n})$

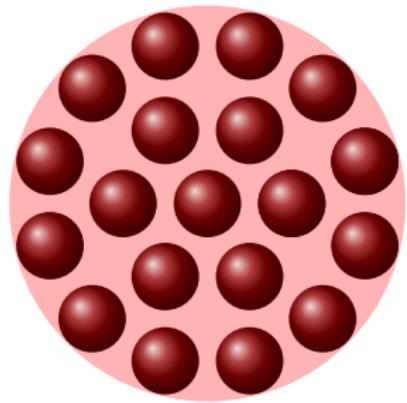
MOTIVATION



(a) Retired

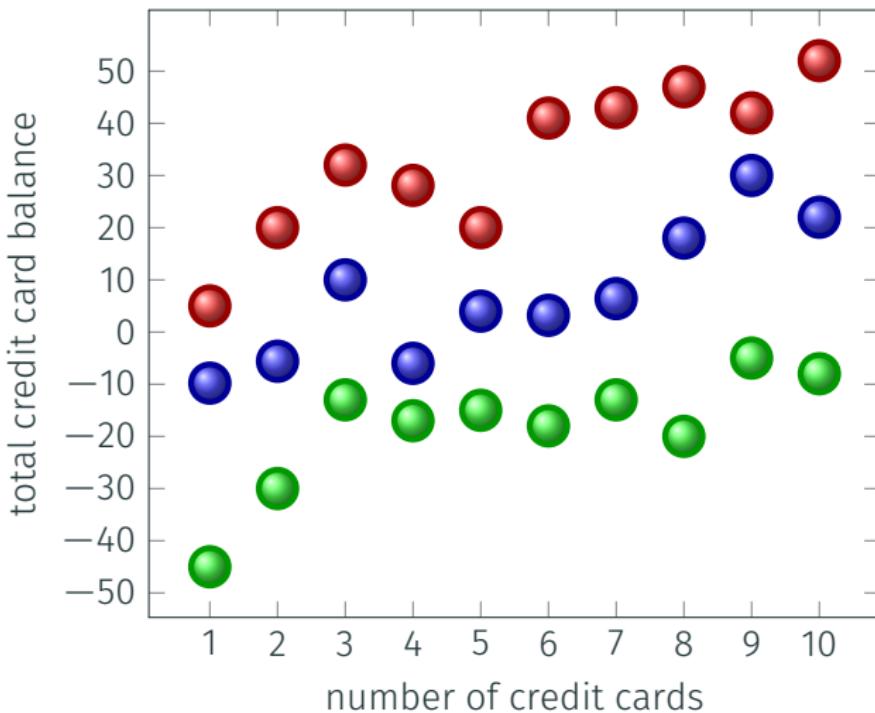


(b) Employed

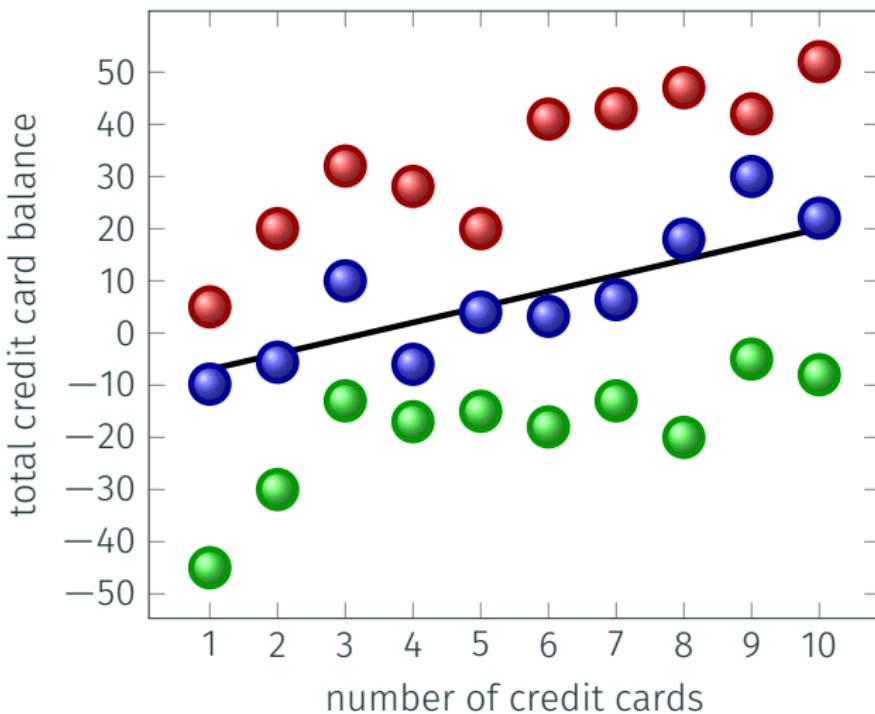


(c) Students

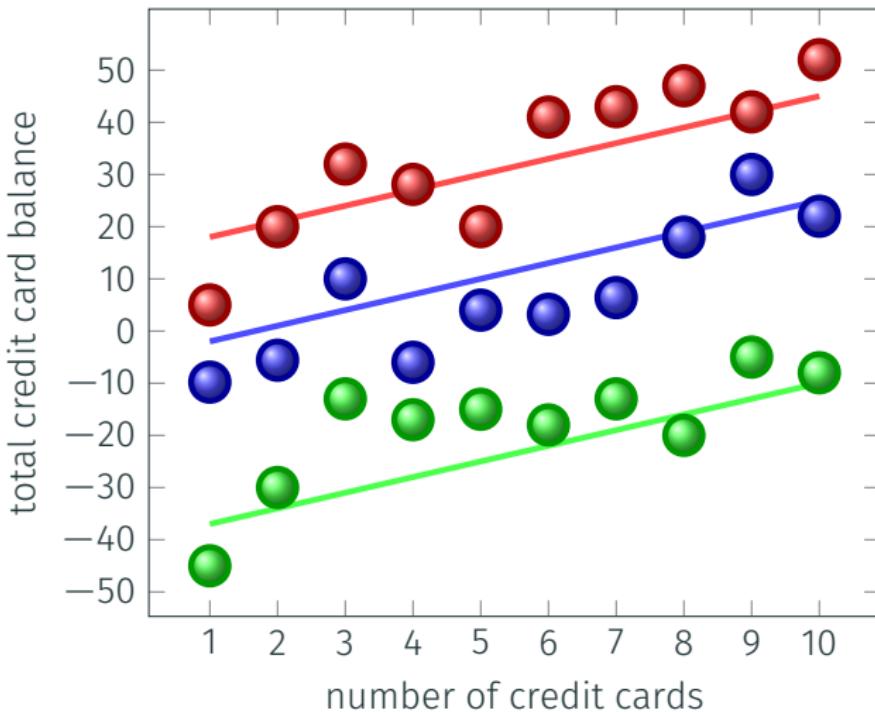
LINEAR MODEL



LINEAR MODEL

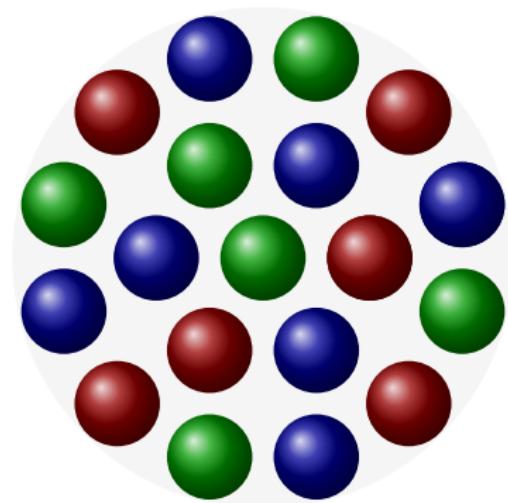


RANDOM INTERCEPT

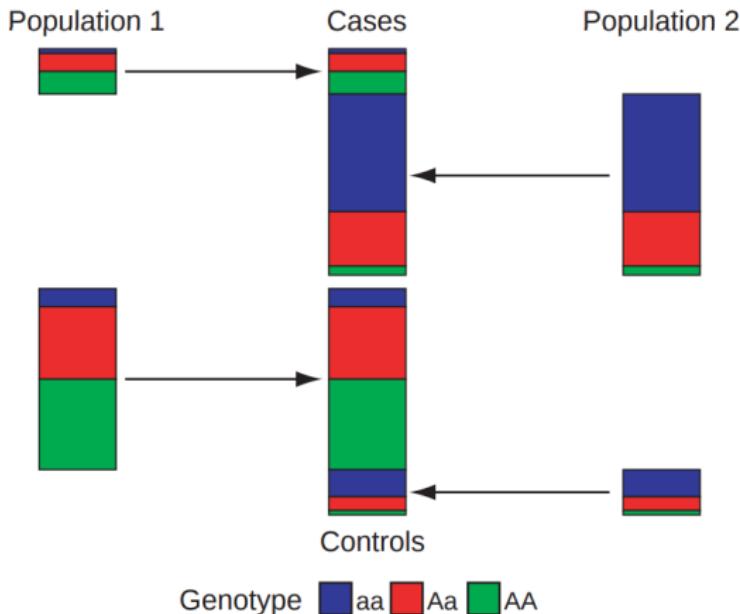


WHEN GROUPING INFORMATION IS UNKNOWN

- In our applications, the grouping information is unknown
- It must be estimated from the data



MOTIVATION



¹Marchini et al. Nature genetics (2004)

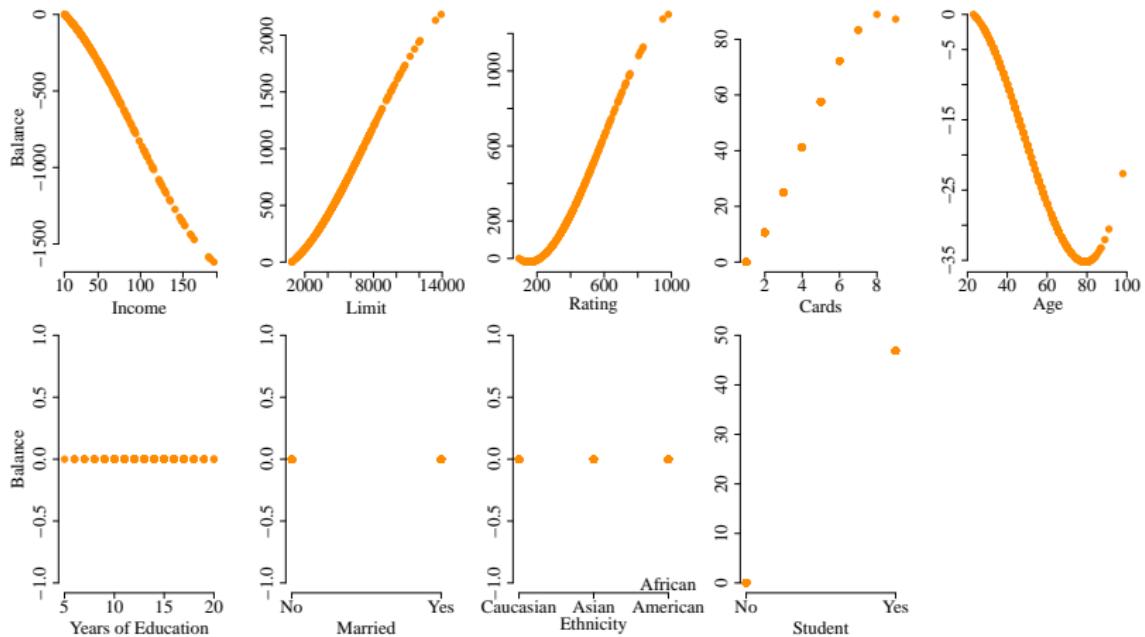
sail: RÉSULTATS SUR UN VRAI JEUX DE DONNÉES

- Credit dataset¹: $Y_{400 \times 1} \rightarrow$ Credit card balance, $X_{400 \times 9} \rightarrow$ predictors

¹ISLR package on CRAN

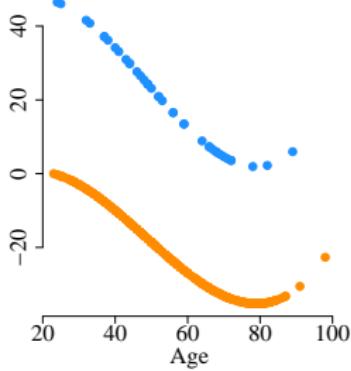
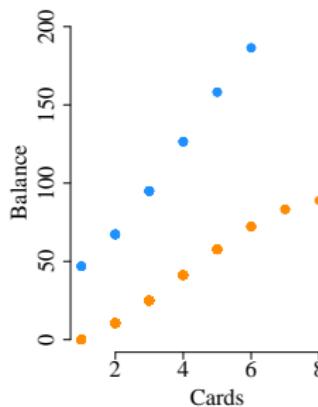
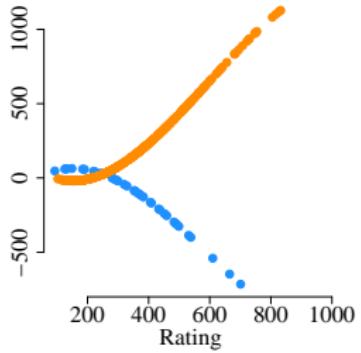
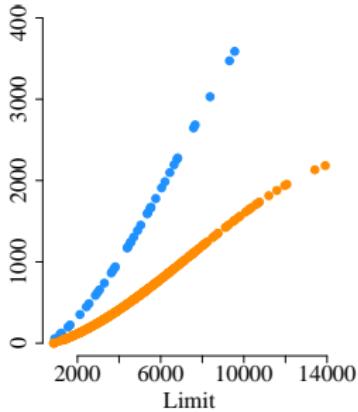
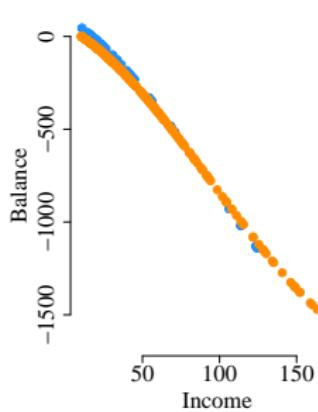
sail: RÉSULTATS SUR UN VRAI JEUX DE DONNÉES

■ Credit dataset¹: $Y_{400 \times 1} \rightarrow$ Credit card balance, $X_{400 \times 9} \rightarrow$ predictors



¹ISLR package on CRAN

sail: INTERACTIONS AVEC LA VARIABLE Student



Which color
represents
a Student?

sail: INTERACTIONS AVEC LA VARIABLE Student

