Sélection de variables par la régression ridge adaptative

Jean-Michel Bécu
Christophe Ambroise, Yves Grandvalet, Cyril Dalmasso

Laboratoire Heuristique et Diagnostic des Systèmes Complexes, Heudiasyc (UTC)
Laboratoire de Mathématique et Modélisation dEvry, LaMME

NetBio 2014

Plan

- **Contexte** Prédiction de variables sur des données biologiques (GWAS, micro-array) avec $n \ll p$ et des structures de variables corrélées
- **Problème** Sélection de variables dans des problèmes de régression en grande dimension
- **Objectif** Test de significativité sur les prédicteurs de l'elastic-net et du lasso (méthodes parcimonieuses)

- **Contexte** Prédiction de variables sur des données biologiques (GWAS, micro-array) avec $n \ll p$ et des structures de variables corrélées
- **Problème** Sélection de variables dans des problèmes de régression en grande dimension
- **Objectif** Test de significativité sur les prédicteurs de l'elastic-net et du lasso (méthodes parcimonieuses)
  - Le coefficient estimé de ma variable j est-il significativement différent de 0 ?
  - Cette variable j apporte-t-elle significativement de l'information dans mon modèle ?

- Retrouver les variables associées au phénomène à expliquer
- Maximiser la **sensibilité** : retrouver les vrais positifs
- Contrôler le **taux de faux positifs (FDR)**

		DÉCISION	
		H_1	H_0
RÉALITÉ	H_1	VP	FN
	H_0	FP	VN

$$\mathrm{SEN} = \mathbb{E}\left[\frac{VP}{VP+FN}\mathbb{I}_{\{(VP+FN)>0\}}\right] \quad , \quad \mathrm{FDR} = \mathbb{E}\left[\frac{FP}{VP+FP}\mathbb{I}_{\{(VP+FP)>0\}}\right]$$

Ridge Adaptative - Pourquoi ?

- Proche de l'adaptive lasso sur le principe de transfert d'information pour une méthode en deux étapes
- Contrairement aux autres méthodes de régressions pénalisées il existe en théorie des test sur la significativité des coefficients (Halawa 99, Buhlmann 12).
- La régression ridge est capable de réestimer exactement les coefficients tels qu'ils sont produits par le lasso ou l'elastic-net

Introduction

Exemple - Données GWAS (Dalmasso 08)

- Association de la charge virale du HIV avec les SNPs (Single Nucleotide Polymorphism) du chromosome 6 chez l'humain
- SNP : variations du génome servant de marqueurs pour différencier des individus, de réponse à un médicament ou la sensibilité à une maladie
- Retrouver les SNPs associés à une forte/faible charge virale
- Étude portant sur 20 000 SNP et 605 individus séropositifs
- Corrélations entre SNP pour des raisons biologiques et spatiales

Introduction

Exemple - Données GWAS (Dalmasso 08)

- Association de la charge virale du HIV avec les SNPs (Single Nucleotide Polymorphism) du chromosome 6 chez l'humain
- SNP : variations du génome servant de marqueurs pour différencier des individus, de réponse à un médicament ou la sensibilité à une maladie
- Retrouver les SNPs associés à une forte/faible charge virale
- Étude portant sur 20 000 SNP et 605 individus séropositifs
- Corrélations entre SNP pour des raisons biologiques et spatiales

$\Rightarrow$ Ici stability selection et bolasso ne fonctionnent pas ! !

Plan

- Deux Étapes
  - Ajustement de modèle : "Screening"
    - Choix d'un sous ensemble de variables $\hat{\mathcal{S}}$ de taille p' par régression pénalisée et validation croisée : $\hat{\mathcal{S}} = \{j : \hat{\beta}_j \neq 0\}$ pour $j \in 1...p$

- Deux Étapes
  - Ajustement de modèle : "Screening"
    - Choix d'un sous ensemble de variables $\hat{\mathcal{S}}$ de taille p' par régression pénalisée et validation croisée : $\hat{\mathcal{S}} = \{j : \hat{\beta}_j \neq 0\}$ pour $j \in 1...p$
    - $p \gg n \quad \rightarrow p' < n(lasso)$
    - $\mathbb{P}(\mathcal{S}^\star \subset \hat{\mathcal{S}}) \rightarrow 1$ quand $n \rightarrow +\inf$

- Deux Étapes
  - Ajustement de modèle : "Screening"
    - Choix d'un sous ensemble de variables $\hat{\mathcal{S}}$ de taille p' par régression pénalisée et validation croisée : $\hat{\mathcal{S}} = \{j : \hat{\beta}_j \neq 0\}$ pour $j \in 1...p$
    - $p \gg n \quad \rightarrow p' < n(lasso)$
    - $\mathbb{P}(\mathcal{S}^\star \subset \hat{\mathcal{S}}) \rightarrow 1$ quand $n \rightarrow +\inf$
    - Les p' variables restantes ont également des poids spécifiques ($\hat{\beta}_j$ pour $j \in \hat{\mathcal{S}}$)

- Deux Étapes
  - Ajustement de modèle : "Screening"
    - Choix d'un sous ensemble de variables $\hat{\mathcal{S}}$ de taille p' par régression pénalisée et validation croisée : $\hat{\mathcal{S}} = \{j : \hat{\beta}_j \neq 0\}$ pour $j \in 1...p$
    - $p \gg n \quad \rightarrow p' < n(lasso)$
    - $\mathbb{P}(\mathcal{S}^\star \subset \hat{\mathcal{S}}) \rightarrow 1$ quand $n \rightarrow +\inf$
    - Les p' variables restantes ont également des poids spécifiques ($\hat{\beta}_j$ pour $j \in \hat{\mathcal{S}}$)
  - Sélection de variables : "Cleaning"
    - Ajustement d'un modèle par la regréssion **ridge** prenant en compte les poids
      des p' variables issues du "screening"
    - Test d'hypothèses sur les variables dans ce nouveau modèle
    - Sélection des variables significatives (cadre des tests multiples)

Régression pénalisée

Elastic-net

$$\hat{\boldsymbol{\beta}}(\lambda) = \underset{\boldsymbol{\beta} \in \mathbb{R}^p}{\arg\min}\, J(\boldsymbol{\beta}) + \left(\lambda_1 \|\boldsymbol{\beta}\|_1 + \lambda_2 \|\boldsymbol{\beta}\|_2^2\right) \quad ,$$

- Méthode de régression multivariée pénalisée

Régression pénalisée

Elastic-net

$$\hat{\boldsymbol{\beta}}(\lambda) = \underset{\boldsymbol{\beta} \in \mathbb{R}^p}{\arg\min}\, J(\boldsymbol{\beta}) + \left(\lambda_1 \|\boldsymbol{\beta}\|_1 + \lambda_2 \|\boldsymbol{\beta}\|_2^2\right) \quad ,$$

- Méthode de régression multivariée pénalisée
- Parcimonieuse
- Importance du choix des pénalités λ_1 et λ_2

Régression pénalisée

Elastic-net

$$\hat{\boldsymbol{\beta}}(\lambda) = \underset{\boldsymbol{\beta} \in \mathbb{R}^p}{\arg\min}\, J(\boldsymbol{\beta}) + \left(\lambda_1 \|\boldsymbol{\beta}\|_1 + \lambda_2 \|\boldsymbol{\beta}\|_2^2\right) \quad ,$$

- Méthode de régression multivariée pénalisée
- Parcimonieuse
- Importance du choix des pénalités λ_1 et λ_2
- Difficulté de quantifier une incertitude sur les $\hat{\boldsymbol{\beta}}_j$

Régression pénalisée

Elastic-net

$$\hat{\boldsymbol{\beta}}(\lambda) = \underset{\boldsymbol{\beta} \in \mathbb{R}^p}{\arg\min}\, J(\boldsymbol{\beta}) + \left(\lambda_1 \|\boldsymbol{\beta}\|_1 + \lambda_2 \|\boldsymbol{\beta}\|_2^2\right) \quad ,$$

- Méthode de régression multivariée pénalisée
- Parcimonieuse
- Importance du choix des pénalités λ_1 et λ_2
- Difficulté de quantifier une incertitude sur les $\hat{\boldsymbol{\beta}}_j$
- Si $\lambda_2 = 0$ alors cela correspond á la régression lasso

Régression pénalisée

Ridge

$$\hat{\beta}^{\lambda}_{ridge} = \underset{\beta}{\text{argmin}}(||y - X\beta||_2^2 + \lambda \sum_j \beta_j^2)$$

- Méthode de régression multivariée pénalisée
- Non parcimonieuse
- Importance du choix de la pénalité λ

Régression pénalisée

Ridge

$$\hat{\beta}^{\lambda}_{ridge} = \underset{\beta}{\text{argmin}}(||y - X\beta||_2^2 + \lambda \sum_j \beta_j^2)$$

- Méthode de régression multivariée pénalisée
- Non parcimonieuse
- Importance du choix de la pénalité λ
- Il existe des tests d'hypothèses trés approximatifs sur les $\hat{\beta}_j$

Régression pénalisée

Équivalence Ridge et Elastic-Net (Grandvalet 99)

$$\hat{\beta}^{\lambda}_{enet} = \underset{\boldsymbol{\beta}\in\mathbb{R}^p}{\arg\min}\, J(\boldsymbol{\beta}) + \left(\lambda_1 \left\|\boldsymbol{\beta}\right\|_1 + \lambda_2 \left\|\boldsymbol{\beta}\right\|_2^2\right)$$

$$\omega_j = \frac{\lambda_1}{|\hat{\beta}^{\lambda}_{enet(j)}|} + \lambda_2$$

$$\hat{\beta}^{\lambda}_{enet} = \underset{\beta}{\text{argmin}}\ J(\beta) + \sum_j \omega_j \beta_j^2$$

Screen and clean

Adapté de Wasserman 09

Recherche du support $\mathcal{S}^\star$: ensemble des coefficients non-nuls

$\mathcal{D} = \{(x_i, y_i)\}_{i=1}^n$ coupé en deux sous-ensembles indépendants $\mathcal{D}_1$ et $\mathcal{D}_2$ (de taille $\frac{n}{2}$)

$$\overbrace{\mathcal{D} = \mathcal{D}_1 \cup \mathcal{D}_2 \xrightarrow[\text{fit model}]{\text{step I}(\mathcal{D}_1)} \{\mathcal{S}_\lambda\}_{\lambda\in\Lambda} \xrightarrow[\text{select model}]{\text{step II}(\mathcal{D}_1)} \mathcal{S}_{\hat{\lambda}}}^{\text{screen}} \overbrace{\xrightarrow[\text{test support}]{\text{step III}(\mathcal{D}_2)} \hat{\mathcal{S}}}^{\text{clean}}$$

$$\overbrace{\mathcal{D} = \mathcal{D}_1 \cup \mathcal{D}_2 \xrightarrow[\text{fit model}]{\text{step I}(\mathcal{D}_1)} \{\mathcal{S}_\lambda\}_{\lambda\in\Lambda} \xrightarrow[\text{select model}]{\text{step II}(\mathcal{D}_1)} \mathcal{S}_{\hat{\lambda}}}^{\text{screen}} \overbrace{\xrightarrow[\text{test support}]{\text{step III}(\mathcal{D}_2)} \hat{\mathcal{S}}}^{\text{clean}}$$

Ajustement des modèles candidats par l'elastic net sur $\mathcal{D}_1$

$$\{\mathcal{S}_\lambda\}_{\lambda\in\Lambda} = \{\mathcal{S}_{\hat{\lambda}} : \lambda \in \Lambda\}$$

Choix de $\mathcal{S}_{\hat{\lambda}}$ par validation croisée sur $\mathcal{D}_1$.

Screen and clean

Cleaning - Regression ridge

$$\overbrace{\mathcal{D} = \mathcal{D}_1 \cup \mathcal{D}_2 \xrightarrow[\text{fit model}]{\text{step I}(\mathcal{D}_1)} \{\mathcal{S}_\lambda\}_{\lambda\in\Lambda} \xrightarrow[\text{select model}]{\text{step II}(\mathcal{D}_1)} \mathcal{S}_{\hat{\lambda}}}^{\text{screen}} \overbrace{\xrightarrow[\text{test support}]{\text{step III}(\mathcal{D}_2)} \hat{\mathcal{S}}}^{\text{clean}}$$

Régression ridge sur $\mathcal{D}_2$ avec les pénalités ω_j correspondantes pour $\mathcal{S}_{\hat{\lambda}}$.

$$\hat{\beta} = \underset{\beta}{\text{argmin}}\ J(\beta) + \sum_j \omega_j \beta_j^2$$

Test de Student : Test sur la significativité des coefficients, $H_0 : \hat{\beta}_j = 0$

$$T_j = \frac{\hat{\beta}_j - \mathbb{E}(\hat{\beta}_j)}{var(\hat{\beta}_j)} \sim \mathcal{T}_{n-p}$$

Test de Fisher : Test sur l'importance de la variable j dans le modèle de régression, $H_0 : \hat{y}^{\Omega} - \hat{y}_j^{\omega} = 0$

$$\begin{aligned}
\hat{y}^{\Omega} &= X(X^\top X + \Omega)^{-1} X^\top y \\
\hat{y}_j^{\omega} &= X_{\setminus j}(X_{\setminus j}^\top X_{\setminus j} + \Omega_{\setminus j})^{-1} X_{\setminus j}^\top y \\
F_j &= \frac{\left\| \hat{y}^{\Omega} - \hat{y}_j^{\omega} \right\|^2}{\left\| y - \hat{y}^{\Omega} \right\|^2} \frac{ddl2}{ddl1} \sim \mathcal{F}_{ddl1,ddl2}
\end{aligned}$$

Screen and clean

Comparaison des procédures de test ($\alpha = 5\%$)

Simulation design	IND		BLOCK		GROUP	
	FPR	SEN	FPR	SEN	FPR	SEN
standard *t*-test	8.0	94.0	12.4	93.1	5.8	95.7
standard *F*-test	9.9	93.1	11.8	89.6	14.8	73.0

Screen and clean

Comparaison des procédures de test ($\alpha = 5\%$)

Simulation design	IND		BLOCK		GROUP	
	FPR	SEN	FPR	SEN	FPR	SEN
standard *t*-test	8.0	94.0	12.4	93.1	5.8	95.7
standard *F*-test	9.9	93.1	11.8	89.6	14.8	73.0

- Tests exacts en régression lineaire non pénalisée (ols)

Screen and clean

Comparaison des procédures de test ($\alpha = 5\%$)

Simulation design	IND		BLOCK		GROUP	
	FPR	SEN	FPR	SEN	FPR	SEN
standard *t*-test	8.0	94.0	12.4	93.1	5.8	95.7
standard *F*-test	9.9	93.1	11.8	89.6	14.8	73.0

- Tests exacts en régression lineaire non pénalisée (ols)
- Tests inappropriés en régression pénalisée

Screen and clean

Test de permutation

Soient B permutations de x_j on obtient x_j^b, alors

$$X_{j(b)} = (x_1, x_{j-1}, x_j^b, x_{j+1}, x_p)$$

Screen and clean

Test de permutation

Soient B permutations de x_j on obtient x_j^b, alors

$$X_{j(b)} = (x_1, x_{j-1}, x_j^b, x_{j+1}, x_p)$$
$$\hat{y}^{\Omega}_{j(b)} = X_{j(b)}(X_{j(b)}^\top X_{j(b)} + \Omega)^{-1} X_{j(b)}^\top y$$

Screen and clean

Test de permutation

Soient B permutations de x_j on obtient x_j^b, alors

$$
\begin{aligned}
X_{j(b)} &= (x_1, x_{j-1}, x_j^b, x_{j+1}, x_p) \\
\hat{y}^{\Omega}_{j(b)} &= X_{j(b)} (X_{j(b)}^{\top} X_{j(b)} + \Omega)^{-1} X_{j(b)}^{\top} y \\
F_{j(b)} &= \frac{\| \hat{y}^{\Omega}_{j(b)} - \hat{y}^{\omega}_{j} \|^2}{\| y - \hat{y}^{\Omega}_{j(b)} \|^2}
\end{aligned}
$$

Screen and clean

Test de permutation

Soient B permutations de x_j on obtient x_j^b, alors

$$
\begin{aligned}
X_{j(b)} &= (x_1, x_{j-1}, x_j^b, x_{j+1}, x_p) \\
\hat{y}^{\Omega}_{j(b)} &= X_{j(b)} (X_{j(b)}^{\top} X_{j(b)} + \Omega)^{-1} X_{j(b)}^{\top} y \\
F_{j(b)} &= \frac{\| \hat{y}^{\Omega}_{j(b)} - \hat{y}^{\omega}_{j} \|^2}{\| y - \hat{y}^{\Omega}_{j(b)} \|^2} \\
Fn(F_j) &= \frac{1}{B} \sum_{b=1}^{B} I(F_{j(b)} \leq F_j) \\
\mathbb{P}_j &= 1 - Fn(F_j)
\end{aligned}
$$

Screen and clean

Test de permutation

Soient B permutations de x_j on obtient x_j^b, alors

$$
\begin{aligned}
X_{j(b)} &= (x_1, x_{j-1}, x_j^b, x_{j+1}, x_p) \\
\hat{y}^{\Omega}_{j(b)} &= X_{j(b)} (X_{j(b)}^{\top} X_{j(b)} + \Omega)^{-1} X_{j(b)}^{\top} y \\
F_{j(b)} &= \frac{\| \hat{y}^{\Omega}_{j(b)} - \hat{y}^{\omega}_{j} \|^2}{\| y - \hat{y}^{\Omega}_{j(b)} \|^2} \\
Fn(F_j) &= \frac{1}{B} \sum_{b=1}^{B} I(F_{j(b)} \leq F_j) \\
\mathbb{P}_j &= 1 - Fn(F_j)
\end{aligned}
$$

- $X_j(b)$ non corrélée à y
- $X_j(b)$ non corrélée aux autres variables explicatives

Screen and clean

Comparaison des procédures de test ($\alpha = 5\%$)

Simulation design	IND		BLOCK		GROUP	
	FPR	SEN	FPR	SEN	FPR	SEN
standard t-test	8.0	94.0	12.4	93.1	5.8	95.7
standard F-test	9.9	93.1	11.8	89.6	14.8	73.0
permutation F-test	5.1	92.4	3.9	86.7	3.9	62.3

Screen and clean

Comparaison des procédures de test ($\alpha = 5\%$)

Simulation design	IND		BLOCK		GROUP	
	FPR	SEN	FPR	SEN	FPR	SEN
standard *t*-test	8.0	94.0	12.4	93.1	5.8	95.7
standard *F*-test	9.9	93.1	11.8	89.6	14.8	73.0
permutation *F*-test	5.1	92.4	3.9	86.7	3.9	62.3

- Test approximatif par un test de permutation sur la statistique de Fisher
- Notre test dans le cadre de l'ensemble des simulations nous offre un contrôle de l'erreur de type-*I*

Plan

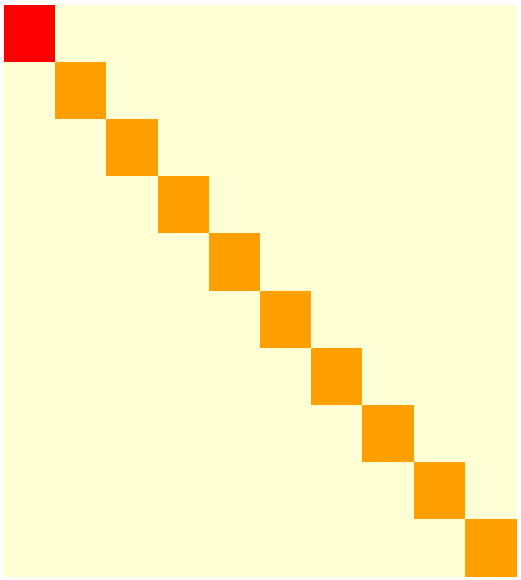
Jeux de données

Définition

- **IND** variables indépendantes
  - p variables indépendamment distribuées selon une loi normale
- **BLOCK**
  - Les p variables suivent une distribution normale $\mathcal{N}(0, \Sigma)$, où $\rho = 0.5$ pour les variables d'un même bloc. Les variables relevantes sont distribuées aléatoirement dans les B blocs.
- **GROUP**
  - Les p variables suivent une distribution normale $\mathcal{N}(0, \Sigma)$, où $\rho = 0.5$ pour les variables d'un même bloc. Un des B blocs est composé uniquement de variables pertinentes.

Jeux de données

GROUP



Matrice de corrélation

Jeux de données

Paramètres

- Pour toutes les simulations dans cette présentation
  - $n = 250$
  - $p = 500$
  - Nombre de variables explicatives $(p_{eff}) = 25$
  - $\beta_{relevant} \sim \mathcal{U}(10^{-1}, 1)$
  - Ratio signal sur bruit (SNR) $= 4$
  - $\rho = 0.5$
  - Taille des blocs $= 25$

Lasso vs. Lasso + Ridge

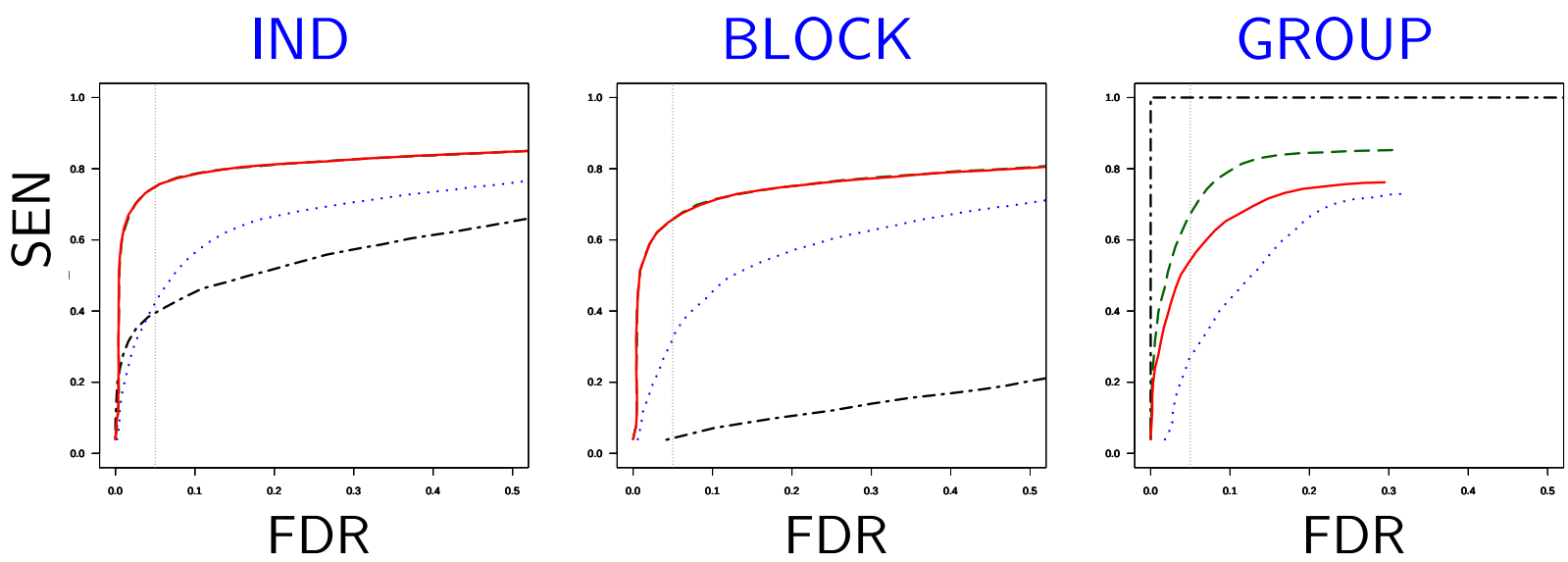
Effet cleaning

Simulation design		IND		BLOCK		GROUP	
		FDR	SEN	FDR	SEN	FDR	SEN
Lasso	*screening*	75.9	87.0	75.1	83.8	32.0	86.2
	cleaning	3.9	75.9	3.1	64.9	1.2	42.2
E.-Net	*screening*	76.9	87.0	76.1	84.1	37.0	90.4
	cleaning	3.9	76.1	3.3	64.9	1.2	66.8

Comparaison méthodes

. . . . Lasso + OLS (Wasserman 09) . _ . Univarié

- - E.-net + Ridge — Lasso + Ridge



Comparaison des méthodes (2)

FDR visé 5%

Simulation design	IND		BLOCK		GROUP	
	FDR	SEN	FDR	SEN	FDR	SEN
E.-Net + R	3.9	76.1	3.3	64.9	1.2	66.8
Lasso + R	3.9	75.9	3.1	64.9	1.2	42.2
Lasso +OLS	4.4	48.7	5.2	40.0	1.9	20.3
Univar	5.2	40.4	86.4	71.0	4.7	100.0

Données GWAS (Dalmasso 08)

- $p = 20000, n = 605$
- Forte corrélations

SNP	Genomic Region	E.-Net–AR	E.-Net–OLS	Univar
rs10484554	MHC	2.9	21.9	0.003
rs2523619	MHC	5.8	97.0	0.2
rs2395029	MHC	9.7	62.0	1.3
rs6923486	other	13.1	17.9	99.5

Plan

Conclusion

- Importance du transfert des poids ω
- Méthode performante pour tester la significativité des variables sélectionnées par l'elastic-net ou le lasso.
- La ridge pour l'étape de cleaning est plus efficace que l'OLS

  meilleure sensibilité $\leftarrow$ poids adaptatifs
- Mise en exergue des problématiques des tests d'hypothèse pour la régression ridge
- Le choix de la méthode de référence ("screen") est primordial

Perspectives

- Utiliser le boostrap pour estimer la distribution des coefficients de la ridge
  (Crivelli95 ,Charterjee10)
- "screen and clean" pour l'inférence (Yu 13)
- Réećhantillonage (Meinshausen 09, Beinrucker 11) → stabilité
- Extension au fused-lasso, group-lasso, ggm
- Classification (analyse discriminante linéaire)

MERCI

Test de Student : Test sur la significativité des coefficients, $H_0 : \hat{\beta}_j = 0$

$$T_j = \frac{\hat{\beta}_j - \mathbb{E}(\hat{\beta}_j)}{var(\hat{\beta}_j)}$$

où $var(\hat{\beta}_j)$ dépend de $\|y - \hat{y}\|^2$

- $\mathbb{E}(\hat{\beta}) = 0 \Rightarrow$ FAUX $\Rightarrow$ Estimateurs biaisés
- $\|y - \hat{y}\|^2 \sim \chi^2_{n-tr(H)} \Rightarrow$ FAUX $\Rightarrow H^2 \neq H$
- $\hat{\beta}_j \perp\!\!\!\perp \|y - \hat{y}\|^2 \Rightarrow$ FAUX $\Rightarrow H^2 \neq H$

Test de Fisher : Test sur l'importance de la variable j dans le modèle de régression, $H_0 : \hat{y}^{\Omega} - \hat{y}_j^{\omega} = 0$

$$\hat{y}^{\Omega} = X(X^{\top}X + \Omega)^{-1}X^{\top}y$$
$$\hat{y}_j^{\omega} = X_{\setminus j}(X_{\setminus j}^{\top}X_{\setminus j} + \Omega_{\setminus j})^{-1}X_{\setminus j}^{\top}y$$
$$F_j = \frac{\|\hat{y}^{\Omega} - \hat{y}_j^{\omega}\|^2}{\|y - \hat{y}^{\Omega}\|^2}\frac{ddl2}{ddl1} \sim \mathcal{F}_{ddl1,ddl2}$$

- $\|\hat{y}^{\Omega} - \hat{y}_j^{\omega}\|^2 \sim \chi^2_{ddl1} \Rightarrow$ FAUX
- $\|y - \hat{y}^{\Omega}\|^2 \sim \chi^2_{ddl2} \Rightarrow$ FAUX
- $\|\hat{y}^{\Omega} - \hat{y}_j^{\omega}\|^2 \perp\!\!\!\perp \|y - \hat{y}^{\Omega}\|^2 \Rightarrow$ FAUX