

Tests d'hypothèse en régressions pénalisées pour la sélection de variables en bioinformatique

Jean-Michel Bécu

Christophe Ambroise, Yves Grandvalet, Cyril Dalmasso

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

JOURNÉES MAS 2014



1 Introduction

2 Méthode

- Régression pénalisée
- Screen and Clean
- Test d'hypothèses

3 Simulation

- Jeu de données
- Effet Cleaning
- Comparaison méthodes

4 Conclusion

Introduction

Problème considéré - Données biologiques

- **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)

Introduction

Sélection de variables : critères

- 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

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

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

1 Introduction

2 Méthode

- Régression pénalisée
- Screen and Clean
- Test d'hypothèses

3 Simulation

- Jeu de données
- Effet Cleaning
- Comparaison méthodes

4 Conclusion

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

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

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

- Deux Étapes

- Ajustement de modèle : "Screening"

- Choix d'un sous ensemble de variables \hat{S} de taille p' par régression pénalisée et validation croisée : $\hat{S} = \{j : \hat{\beta}_j \neq 0\}$ pour $j \in 1 \dots p$
 - $p \gg n \rightarrow p' < n$ (*lasso*)
 - $\mathbb{P}(S^* \subset \hat{S}) \rightarrow 1$ quand $n \rightarrow +\infty$
 - Les p' variables restantes ont également des poids spécifiques ($\hat{\beta}_j$ pour $j \in \hat{S}$)

- Sélection de variables : "Cleaning"

- Nouvel ajustement d'un modèle 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)

$$\hat{\beta}(\lambda) = \arg \min_{\beta \in \mathbb{R}^p} J(\beta) + (\lambda_1 \|\beta\|_1 + \lambda_2 \|\beta\|_2^2) ,$$

- Méthode de régression multivariée pénalisée

$$\hat{\beta}(\lambda) = \arg \min_{\beta \in \mathbb{R}^p} J(\beta) + (\lambda_1 \|\beta\|_1 + \lambda_2 \|\beta\|_2^2) ,$$

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

$$\hat{\beta}(\lambda) = \arg \min_{\beta \in \mathbb{R}^p} J(\beta) + (\lambda_1 \|\beta\|_1 + \lambda_2 \|\beta\|_2^2) ,$$

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

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

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

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

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

Régression pénalisée

Équivalence Ridge et Elastic-Net (Grandvalet 99)

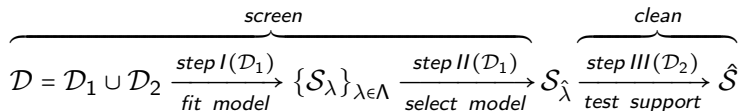
$$\hat{\beta}_{enet}^{\lambda} = \arg \min_{\beta \in \mathbb{R}^p} J(\beta) + (\lambda_1 \|\beta\|_1 + \lambda_2 \|\beta\|_2^2)$$

$$\omega_j = \frac{\lambda_1}{|\hat{\beta}_j(\lambda)|} + \lambda_2$$

$$\hat{\beta}_{enet}^{\lambda} = \arg \min_{\beta} J(\beta) + \sum_j \omega_j \beta_j^2$$

Recherche du support \mathcal{S}^* : 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 $\frac{n}{2}$)



Screen and clean

Screening

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

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

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

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

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

Screen and clean

Cleaning - Test de significativité (t -test)

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

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

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

Screen and clean

Cleaning - Test de significativité (t -test)

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

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

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

- $\mathbb{E}(\hat{\beta}) = 0$

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

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

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

- $\mathbb{E}(\hat{\beta}) = 0$
- $\|y - \hat{y}\|^2 \sim \chi_{n-\text{tr}(H)}^2$

Screen and clean

Cleaning - Test de significativité (t -test)

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

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

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

- $\mathbb{E}(\hat{\beta}) = 0$
- $\|y - \hat{y}\|^2 \sim \chi_{n-\text{tr}(H)}^2$
- $\hat{\beta}_j \perp \|y - \hat{y}\|^2$

Screen and clean

Cleaning - Test de significativité (F -test)

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

$$\hat{y}^\Omega = X(X^\top X + \Omega)^{-1} X^\top y$$

$$\hat{y}_j^\omega = X_{\setminus j} (X_{\setminus j}^\top X_{\setminus j} + \Omega_{\setminus j})^{-1} X_{\setminus j}^\top y$$

$$F_j = \frac{\|\hat{y}^\Omega - \hat{y}_j^\omega\|^2}{\|y - \hat{y}^\Omega\|^2} \frac{ddl2}{ddl1} \sim \mathcal{F}_{ddl1, ddl2}$$

Screen and clean

Cleaning - Test de significativité (F -test)

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

$$\hat{y}^\Omega = X(X^\top X + \Omega)^{-1} X^\top y$$

$$\hat{y}_j^\omega = X_{\setminus j} (X_{\setminus j}^\top X_{\setminus j} + \Omega_{\setminus j})^{-1} X_{\setminus j}^\top y$$

$$F_j = \frac{\|\hat{y}^\Omega - \hat{y}_j^\omega\|^2}{\|y - \hat{y}^\Omega\|^2} \frac{ddl2}{ddl1} \sim \mathcal{F}_{ddl1, ddl2}$$

- $\|\hat{y}^\Omega - \hat{y}_j^\omega\|^2 \sim \chi_{ddl1}^2$
- $\|y - \hat{y}^\Omega\|^2 \sim \chi_{ddl2}^2$

Screen and clean

Cleaning - Test de significativité (F -test)

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

$$\hat{y}^\Omega = X(X^\top X + \Omega)^{-1} X^\top y$$

$$\hat{y}_j^\omega = X_{\setminus j} (X_{\setminus j}^\top X_{\setminus j} + \Omega_{\setminus j})^{-1} X_{\setminus j}^\top y$$

$$F_j = \frac{\|\hat{y}^\Omega - \hat{y}_j^\omega\|^2}{\|y - \hat{y}^\Omega\|^2} \frac{ddl2}{ddl1} \sim \mathcal{F}_{ddl1, ddl2}$$

- $\|\hat{y}^\Omega - \hat{y}_j^\omega\|^2 \sim \chi_{ddl1}^2$
- $\|y - \hat{y}^\Omega\|^2 \sim \chi_{ddl2}^2$
- $\|\hat{y}^\Omega - \hat{y}_j^\omega\|^2 \perp \|y - \hat{y}^\Omega\|^2$

Screen and clean

Comparaison des procédures de test

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

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

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 inapproprié en régression pénalisée

Screen and clean

Cleaning - Test de significativité (t -test)

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

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

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

- $\mathbb{E}(\hat{\beta}) = 0$
- $\|y - \hat{y}\|^2 \sim \chi_{n-\text{tr}(H)}^2$
- $\hat{\beta}_j \perp \|y - \hat{y}\|^2$

Screen and clean

Cleaning - Test de significativité (t -test)

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

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

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

- $\mathbb{E}(\hat{\beta}) = 0 \Rightarrow$ **FAUX** \Rightarrow Estimateurs biaisés
- $\|y - \hat{y}\|^2 \sim \chi_{n-\text{tr}(H)}^2$
- $\hat{\beta}_j \perp \|y - \hat{y}\|^2$

Screen and clean

Cleaning - Test de significativité (t -test)

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

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

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

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

Screen and clean

Cleaning - Test de significativité (t -test)

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

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

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

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

Screen and clean

Cleaning - Test de significativité (F -test)

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

$$\hat{y}^\Omega = X(X^\top X + \Omega)^{-1} X^\top y$$

$$\hat{y}_j^\omega = X_{\setminus j} (X_{\setminus j}^\top X_{\setminus j} + \Omega_{\setminus j})^{-1} X_{\setminus j}^\top y$$

$$F_j = \frac{\|\hat{y}^\Omega - \hat{y}_j^\omega\|^2}{\|y - \hat{y}^\Omega\|^2} \frac{ddl2}{ddl1} \sim \mathcal{F}_{ddl1, ddl2}$$

- $\|\hat{y}^\Omega - \hat{y}_j^\omega\|^2 \sim \chi_{ddl1}^2$
- $\|y - \hat{y}^\Omega\|^2 \sim \chi_{ddl2}^2$
- $\|\hat{y}^\Omega - \hat{y}_j^\omega\|^2 \perp \|y - \hat{y}^\Omega\|^2$

Screen and clean

Cleaning - Test de significativité (F -test)

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

$$\hat{y}^\Omega = X(X^\top X + \Omega)^{-1} X^\top y$$

$$\hat{y}_j^\omega = X_{\setminus j} (X_{\setminus j}^\top X_{\setminus j} + \Omega_{\setminus j})^{-1} X_{\setminus j}^\top y$$

$$F_j = \frac{\|\hat{y}^\Omega - \hat{y}_j^\omega\|^2}{\|y - \hat{y}^\Omega\|^2} \frac{ddl2}{ddl1} \sim \mathcal{F}_{ddl1, ddl2}$$

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

Screen and clean

Cleaning - Test de significativité (F -test)

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

$$\hat{y}^\Omega = X(X^\top X + \Omega)^{-1} X^\top y$$

$$\hat{y}_j^\omega = X_{\setminus j} (X_{\setminus j}^\top X_{\setminus j} + \Omega_{\setminus j})^{-1} X_{\setminus j}^\top y$$

$$F_j = \frac{\|\hat{y}^\Omega - \hat{y}_j^\omega\|^2}{\|y - \hat{y}^\Omega\|^2} \frac{ddl2}{ddl1} \sim \mathcal{F}_{ddl1, ddl2}$$

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

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

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

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

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

Screen and clean

Test de permutation

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

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

$$\hat{y}_{j(b)}^\Omega = X_{j(b)} (X_{j(b)}^\top X_{j(b)} + \Omega)^{-1} X_{j(b)}^\top y$$

$$F_{j(b)} = \frac{\|\hat{y}_{j(b)}^\Omega - \hat{y}_j^\omega\|^2}{\|y - \hat{y}_{j(b)}^\Omega\|^2}$$

$$Fn(F_j) = \frac{1}{B} \sum_{b=1}^B I(F_{j(b)} \leq F_j)$$

$$\mathbb{P}_j = 1 - Fn(F_j)$$

Screen and clean

Test de permutation (2)

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

Test de permutation (2)

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

1 Introduction

2 Méthode

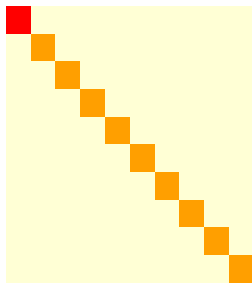
- Régression pénalisée
- Screen and Clean
- Test d'hypothèses

3 Simulation

- Jeu de données
- Effet Cleaning
- Comparaison méthodes

4 Conclusion

- **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 pertinentes 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.



- 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	<i>screening</i>	75.9	87.0	75.1	83.8	32.0	86.2
	<i>cleaning</i>	3.9	75.9	3.1	64.9	1.2	42.2
E.-Net	<i>screening</i>	76.9	87.0	76.1	84.1	37.0	90.4
	<i>cleaning</i>	3.9	76.1	3.3	64.9	1.2	66.8

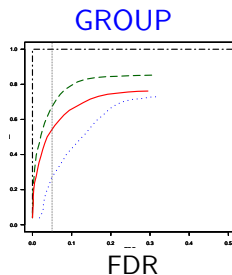
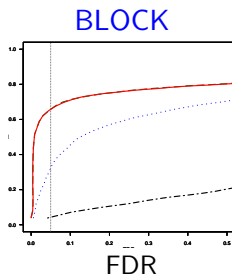
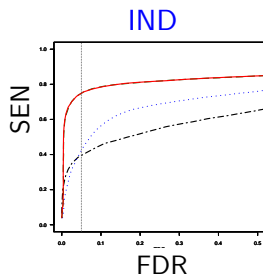
Comparaison méthodes

.... Lasso + OLS (Wasserman 09)

- - - E.-net + Ridge

. _ . Univarié

— 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

1 Introduction

2 Méthode

- Régression pénalisée
- Screen and Clean
- Test d'hypothèses

3 Simulation

- Jeu de données
- Effet Cleaning
- Comparaison méthodes

4 Conclusion

- 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 ← poids adaptatifs
- Mise en exergue des problématiques des tests d'hypothèse pour la régression ridge.

- Stratégies de permutation plus efficaces que celle proposée ?
- Pas d'échantillonnage entre les étapes de "screen and clean" (Yu 13)
- Rééchantillonnage (Meinshausen 09, Beinrucker 11) → stabilité
- Extension au group-lasso
- Classification (analyse discriminante linéaire)

MERCI