Agrégation de méthodes statistiques pour la sélection de variables corrélées et en grande dimension

Bérangère Bastien, Anne Gégout-Petit, Aurélie Muller-Gueudin

Agrégation de méthodes statistiques pour Sélection de variables corrélées & en grande dimension

Bérangère Bastien[1]
Anne Gégout-Petit[2], Aurélie Muller-Gueudin[2]

[1] Transgène, Strasbourg
[2] Université de Lorraine, Inria Nancy, équipe BIGS.

AgroParisTech, 21 mai 2019

Sommaire

Problématique

Données transcriptomiqes de patients atteints de cancer du poumon :

- $n = 37$ patients en stade avancé du cancer, ayant reçu une chimiothérapie.
- $m =$51 336 variables transcriptomiques, dépendantes entre elles.
- Une variable réponse : issue du traitement par chimiothérapie (binaire ou quantitative).

Objectifs

1. Sélection et tri des variables liées à la réponse
2. Visualisation des profils de patients

Sommaire

Sélection de variables en grande dimension

Quelques méthodes existantes :

- Procédures de tests multiples
- Régression pénalisée
- Forêts aléatoires

Dépendance des variables ⇒

- Tests multiples : FDR difficile à contrôler (biaisé, instable)
- Régression et forêts aléatoires : résultats perturbés

Dépendance des variables, effet sur les tests multiples :

Pour $i = 1, \ldots, m$, $\quad H_0^i : X_i$ n'influence pas la réponse Y.

- Si X_i indépendants : sous H_0, p-valeurs $\underset{H_0}{\sim} \mathcal{U}([0,1])$.
- Sinon :

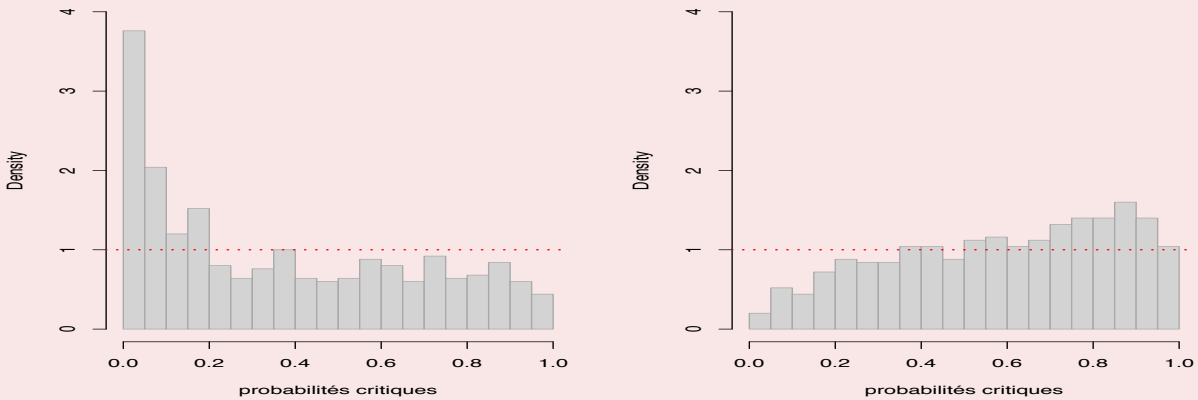


FIGURE 1 – Histogramme des p-values sous H_0 avec des X_i correlés, exemple avec 2 jeux de données (Friguet 2010).

Comment prendre en compte la dépendance

Factor Analysis for Multiple Testing (Friguet 2010)

- Décrire la dépendance entre variables via un petit nombre $q \ll m$ de facteurs communs $\mathbf{Z}$

- Pour $k = 1, \ldots, m$,

$$X_k = \underbrace{f_k(Y)}_{\text{regression}} + \underbrace{\mathbf{Z}\mathbf{b}_k' + \varepsilon_k}_{\text{FA pour les résidus}}$$

- Hypothèses :
  - $\mathbb{V}(\mathbf{Z}) = \mathbb{I}_q$
  - $\mathbb{V}(\epsilon) = \Psi$ (diagonale)
  - $\Sigma := \mathbb{V}(\mathbf{X}|Y) = BB' + \Psi$
  - $\mathbb{V}(\mathbf{X}|Y, \mathbf{Z}) = \Psi$
  - $\text{Cov}(\varepsilon_k, Z_j) = 0, \forall k, \forall j$

Décorrélation des variables par FAMT

$$\underbrace{\mathbf{X}}_{\text{données de départ}} \xrightarrow{\text{FAMT}} \underbrace{\mathbf{X}^*}_{\text{données ajustées}}$$

- $X_k = f_k(Y) + \mathbf{Z}\mathbf{b}'_k + \varepsilon_k$
- $\Longrightarrow X^*_k = X_k - \mathbf{Z}\mathbf{b}'_k = f_k(Y) + \varepsilon_k$
- $\mathbb{V}(\mathbf{X}^*|Y) = \Psi$ (diagonale)
- Estimation de $\mathbf{Z}, \mathbf{B}, \Psi$ et donc $\mathbf{X}^*$ par algo EM.

Limites de la procédure FAMT

- FAMT en échec pour décorréler m =50 000 variables
- gènes correlés par blocs (blocs $\Longleftrightarrow$ processus/pathways biologiques)
- FAMT a de meilleurs résultats si on l'applique par blocs, plutôt que sur $\mathbf{X}$ tout entier (Bastien 2018)

Prétraitement : structure de covariance des variables

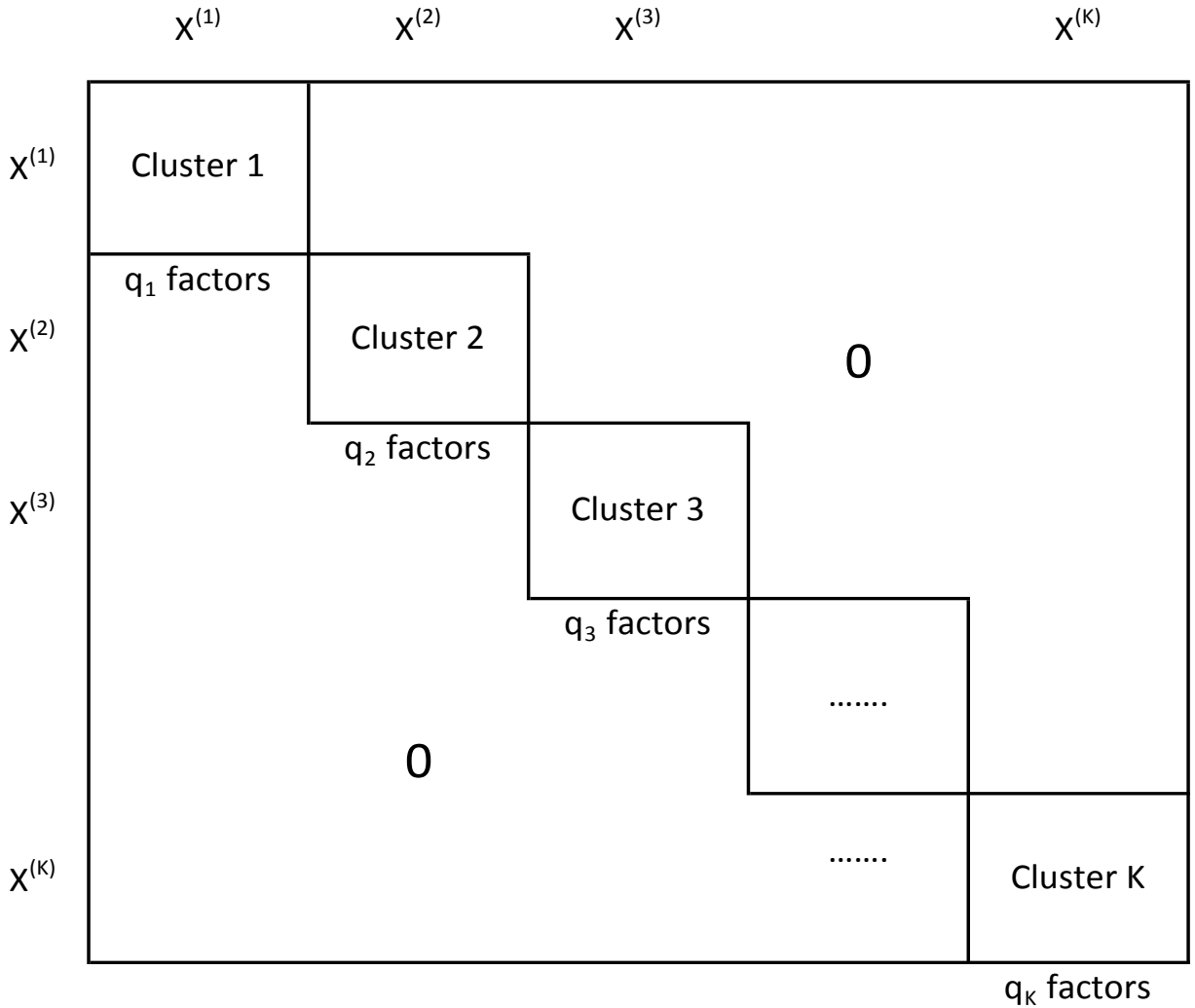
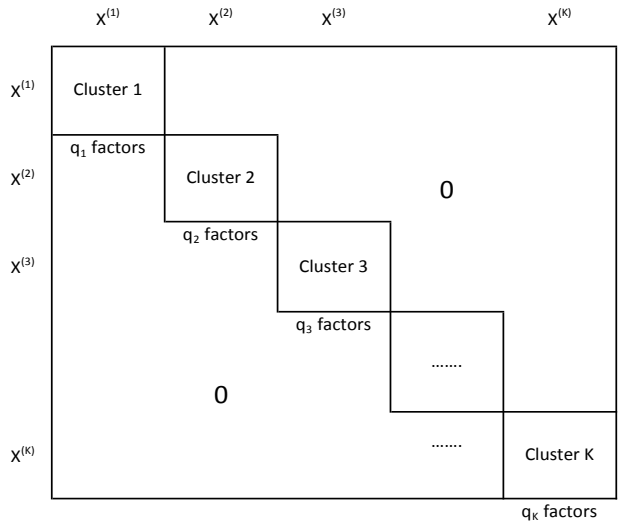


FIGURE 2 – Hypothèse sur la matrice de covariance des variables

Prétraitement : structure de covariance des variables



Hypothèses :

- $\mathbf{X} = (\underbrace{X_1^{(1)}, \ldots, X_{m_1}^{(1)}}_{\mathbf{X}^{(1)}}, \ldots, \underbrace{\ldots, X_i^{(k)}, \ldots}_{\mathbf{X}^{(i)}}, \ldots, \underbrace{\ldots, X_{m_K}^{(K)}}_{\mathbf{X}^{(K)}})$
- Indépendance des K clusters.
- Modèle de dépendance en facteurs dans chaque cluster :
  $X_i^{(k)} = f_i^{(k)}(Y) + \mathbf{Z}^{(k)}\mathbf{b}_i^{(k)\prime} + \varepsilon_i^{(k)}$

Détection des clusters

Algorithme `ClustOfvar` (Chavent et al, 2011)

Soit une partition $P_K = (C_1, \ldots, C_K)$ de $\mathbf{X}$, on définit

- la variable synthétique s_k du cluster C_k :

$$s_k := \underset{u \in \mathbb{R}^n}{\arg\max} \left\{ \sum_{X_j \in C_k} r^2(u, X_j) \right\}$$

- l'homogénéité H du cluster C_k :

$$H(C_k) := \sum_{X_j \in C_k} r^2(s_k, X_j) = \underbrace{\lambda_1^k}_{1^e \text{v.p. de ACP}}$$

- l'homogénéité $\mathcal{H}$ de la partition P_K :

$$\underbrace{\mathcal{H}(P_K)}_{\text{à maximiser}} := \sum_{k=1}^{K} H(C_k) = \sum_{k=1}^{K} \lambda_1^k$$

Synthèse de l'étape Prétraitement :

1. Détection de clusters, via `ClustOfVar`
2. Décorrélation dans les clusters, via `FAMT`

A l'issue de cette étape, on note $\mathbf{X}^*$ le jeu de données corrigé.

Example sur un jeu de données Peptides (800 variables)

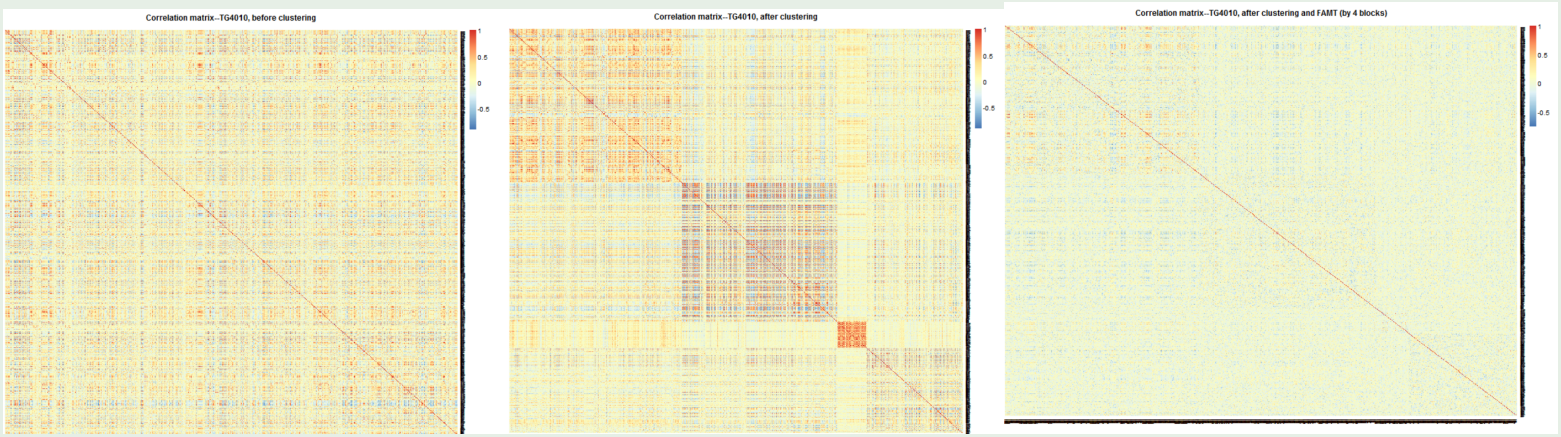


FIGURE 2 – Heatmaps des corrélations : avant clustering ; après clustering ; après FAMT par blocs.

Etape suivante : sélection de variables

Agrégation de méthodes statistiques

- On applique L méthodes statistiques de sélection

- Chaque variable X_j^* obtient un score $S_j \in \{0, \ldots, L\}$
  où $S_j =$ nombre de sélections parmi les L méthodes.

  $\hookrightarrow$ Si $S_j = L$, la variable X_j influence $+++$ Y

  $\hookrightarrow$ Si $S_j = 0$, la variable X_j n'influence pas Y

On peut ensuite trier les variables selon leurs scores.

Sélection de variables par tests multiples

Taux d'erreurs en tests multiples

réalité \ conclusion du test	non-rejet de H_0	rejet de H_0	Total
H_0	U_t	V_t	m_0
H_1	T_t	S_t	m_1
Total	$m - R_t$	R_t	m

FIGURE 3 – Nombres de bonnes décisions et erreurs d'une procédure de tests multiples pour un seuil de rejet t pour les pvaleurs.

- $\text{FWER}_t = \mathbb{P}\left(V_t \geqslant 1\right)$: Family Wise Error Rate
- $\text{FDR}_t = \mathbb{E}\left(\frac{V_t}{R_t}\right)$: False Discovery Rate

Les procédures présentées ci-après supposent que les tests sont indépendants.

Procédures de tests multiples

1. Bonferroni (1936) : seuillage des pvaleurs à $t = \alpha/m$, très conservatif ! Cela assure que FWER$\leqslant \alpha$.

2. Benjamini & Hochberg (1995) : tri des p-values $p_{(1)} \leqslant p_{(2)} \leqslant p_{(3)} \cdots \leqslant p_{(m)}$. Seuillage des pvaleurs à $t = p_{(i^*)}$ tel que $i^* = \underset{i=1,\ldots,m}{\text{argmax}} \{p_{(i)} \leqslant \alpha i/m\}$. Cela assure que FDR$\leqslant \alpha$. Conservatif.

3. Storey & Tibshirani (2003) : q-value$(p_{(i)})$ = proportion de faux positifs parmi les variables ayant des p-values $< p_{(i)}$. Seuillage des q-values à α. Cela assure que FDR$\leqslant \alpha$.

4. Aubert *et al* (2004) : localFDR(i) = proportion de faux positifs parmi les variables ayant des p-values proches de $p_{(i)}$.

5. Friguet *et al* (2010) : `FAMT` ici, avec 0 facteur. Calculs de pvaleurs via F-tests, puis procédure BH avec estimation de π_0.

↪ Chaque procédure fournit sa sélection de variables.

Sélection de variables par régression Lasso

Les variables sélectionnées sont $\{X_i : \widehat{\beta}_i \neq 0\}$ où

- Cas Y continue :

$$\widehat{\beta} = \underset{\beta \in \mathbb{R}^m}{\text{argmin}} \{ \|Y - X\beta\|^2 + \lambda \|\beta\|_1 \}$$

- Cas Y binaire (ou multinomiale) :

$$\widehat{\beta} = \underset{\beta \in \mathbb{R}^m}{\text{argmax}} \{ \underbrace{\ell(\beta; (Y, X))}_{\text{vraisemblance}} - \lambda \|\beta\|_1 \}$$

où ℓ est la vraisemblance du modèle logistique (ou multinomial)

$\hookrightarrow$ Package `glmnet`, Hastie *et al (2010)*

Sélection de variables par Forêts aléatoires

Régression non paramétrique, basée sur une agrégation d'arbres de décision.

Calcul de l'**importance** [1] de chaque variable.

Deux types de sélection :

- sélection de toutes les variables importantes, même redondantes (but : interprétation).
- sélection d'un petit nombre de variables importantes (but : modèle de prédiction parcimonieux).

$\hookrightarrow$ Package `VSURF`, Genuer *et al* (2015)

1. Importance de la variable X_j = augmentation de l'erreur prédite par la forêt si les observations de la variable X_j sont permutées

8 méthodes de sélection

1. Bonferroni
2. Benjamini et Hochberg
3. q-value
4. local FDR
5. `FAMT`
6. Forêts aléatoires : étape d'interprétation
7. Forêts aléatoires : étape de prédiction
8. Régression LASSO

$\Rightarrow$ Pour chaque variable X_j, on a son score $S_j \in \{0, \ldots, 8\}$: nb de sélections.

Sommaire

Simulations

- $n = 60$ observations
- Y binaire ($\frac{n}{2}$ sujets avec $Y = 1$ et $\frac{n}{2}$ sujets avec $Y = 0$)
- $\mathbf{X} = (\mathbf{X}^{(k)})_{k=1,\dots,4}$ = 4 clusters indépendants de 400 variables

- Variables corrélées dans chaque cluster, via des facteurs latents $\mathbf{Z}$:
  1. Scénario 1 : $\mathbf{X} = \mathbf{f}(Y) + \mathbf{Z}\mathbf{B}' + \epsilon$
  2. Scénario 2 :
     $\mathbf{X} = \delta + \mathbf{Z}\mathbf{B}' + \epsilon$, où $\delta \sim$ loi qui dépend de Y

Rappel : $\Sigma = \mathbf{B}\mathbf{B}' + \Psi$ et la force de la corrélation est donnée par $\frac{\text{trace}\left(\mathbf{B}\mathbf{B}'\right)}{\text{trace}(\Sigma)}$

Scénario 1 : $X_j = \delta_j \mathbf{1}_{Y=0} + \mathbf{Z}b'_j + \epsilon_j$

Dans le 1er cluster :
400 variables $=$ 40 variables influentes $+$ 360 variables de bruit

- $\delta_j = 1.5$ pour $j = 1, \ldots, 10$
- $\delta_j = 1$ pour $j = 11, \ldots, 20$
- $\delta_j = 0.75$ pour $j = 21, \ldots, 30$
- $\delta_j = 0.5$ pour $j = 31, \ldots, 40$
- $\delta_j = 0$ pour $j = 41, \ldots, 400$.

Même schéma dans les 3 autres clusters.

$\Longrightarrow$ Total : 160 variables influentes.

Scénario 2 : $X_j = \delta_j + \mathbf{Z}b'_j + \epsilon_j$

Dans le 1er cluster :
400 variables = 60 variables influentes + 340 variables de bruit

- $\delta_j \sim 0.7\mathcal{N}(3y, 1) + 0.3\mathcal{N}(0, 1)$, pour $j = 1, \ldots, 10$
- $\delta_j \sim 0.7\mathcal{N}(2y, 1) + 0.3\mathcal{N}(0, 1)$, pour $j = 11, \ldots, 20$
- $\delta_j \sim 0.7\mathcal{N}(y, 1) + 0.3\mathcal{N}(0, 1)$, pour $j = 21, \ldots, 30$
- $\delta_j \sim 0.3\mathcal{N}(3y, 1) + 0.7\mathcal{N}(0, 1)$, pour $j = 31, \ldots, 40$
- $\delta_j \sim 0.3\mathcal{N}(2y, 1) + 0.7\mathcal{N}(0, 1)$, pour $j = 41, \ldots, 50$
- $\delta_j \sim 0.3\mathcal{N}(y, 1) + 0.7\mathcal{N}(0, 1)$, pour $j = 51, \ldots, 60$

- $\delta_j = 0$ pour $j = 61, \ldots, 400.$

Même schéma dans les 3 autres clusters.

$\Longrightarrow$ Total : 240 variables influentes.

Objectif

Déterminer les variables différentiellement exprimées dans les 2 groupes $Y = 0$ et $Y = 1$.

$\hookrightarrow$ via $m = 1600$ tests de Wilcoxon

Intérêt du prétraitement

On compare 3 procédures de prétraitement

1. rien n'est fait sur $\mathbf{X}$
2. décorrélation de $\mathbf{X}$ via `FAMT` :

$$\mathbf{X} \xrightarrow{\text{FAMT}} \mathbf{X}^*$$

3. détection des 4 clusters via `ClustOfVar`, puis décorrélation dans chaque cluster via `FAMT` :

$$\mathbf{X} \xrightarrow{\text{FAMT}} \mathbf{X}^{**}$$

$\hookrightarrow$ Tests de Wilcoxon sur les 3 datasets issus de ces 3 procédures

$\hookrightarrow$ Sélection des variables X_j dont les p-valeurs sont $< 5\%$.

$\hookrightarrow$ 100 runs de $(\mathbf{X}, Y)$ pour chaque scénario.

Comparaison des 3 procédures (scénario 1)

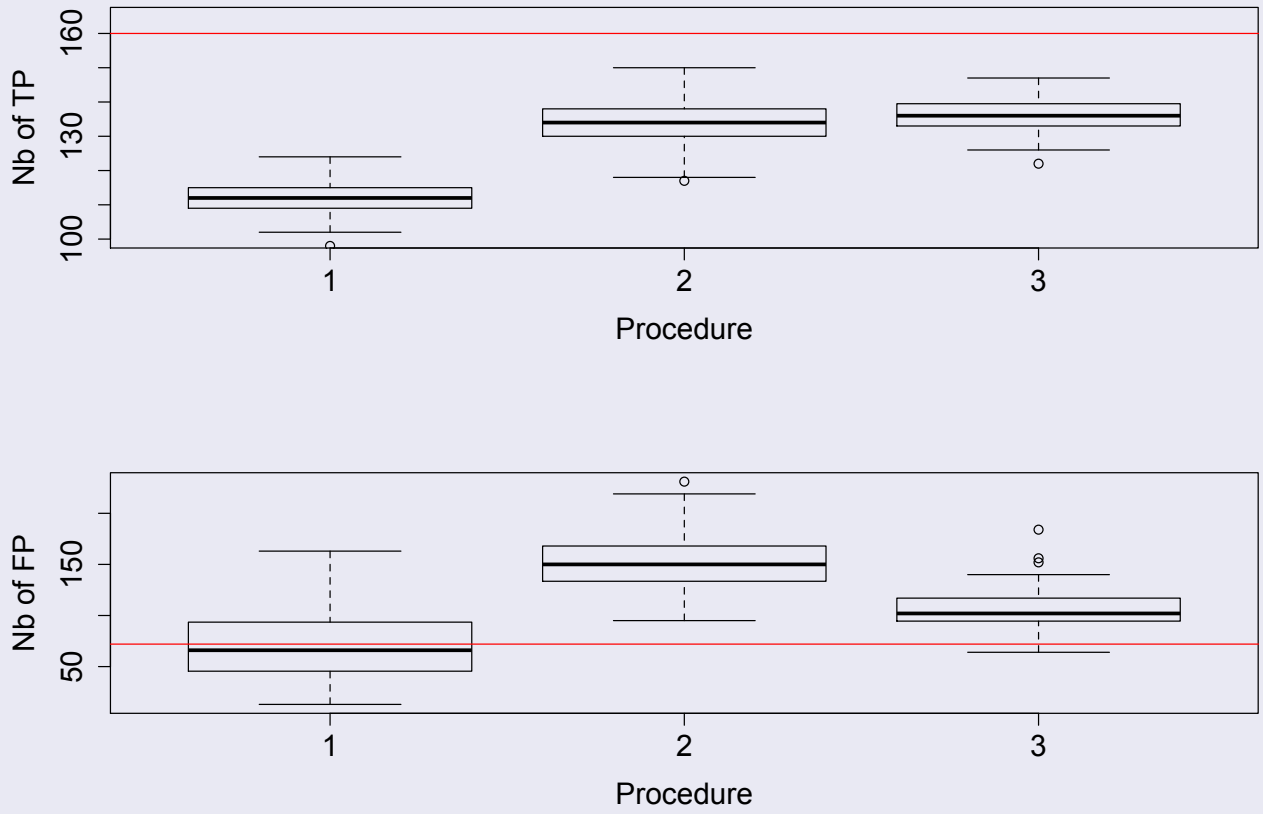


FIGURE 4 – Scénario 1. Boxplots calculés sur $N = 100$ runs de $(\mathbf{X}, Y)$. Lignes rouges : nombres attendus de TP (160) et FP $(5\% \times (1600 - 160))$

$\hookrightarrow$ Amélioration du taux de FP, par rapport à `FAMT`.

Comparaison des 3 procédures (scénario 2)

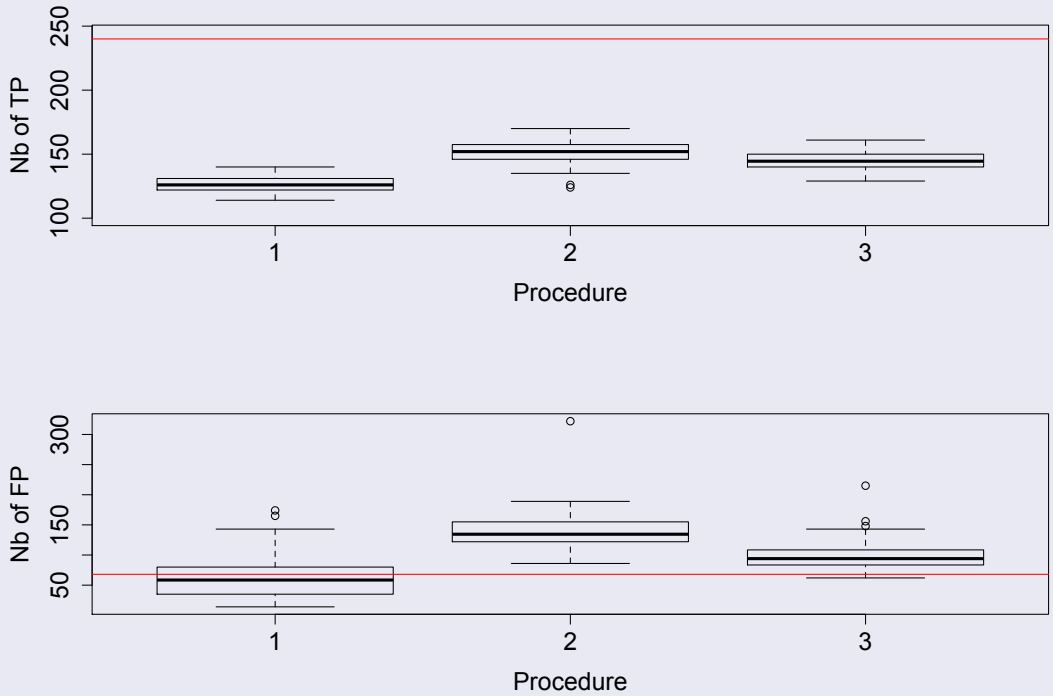
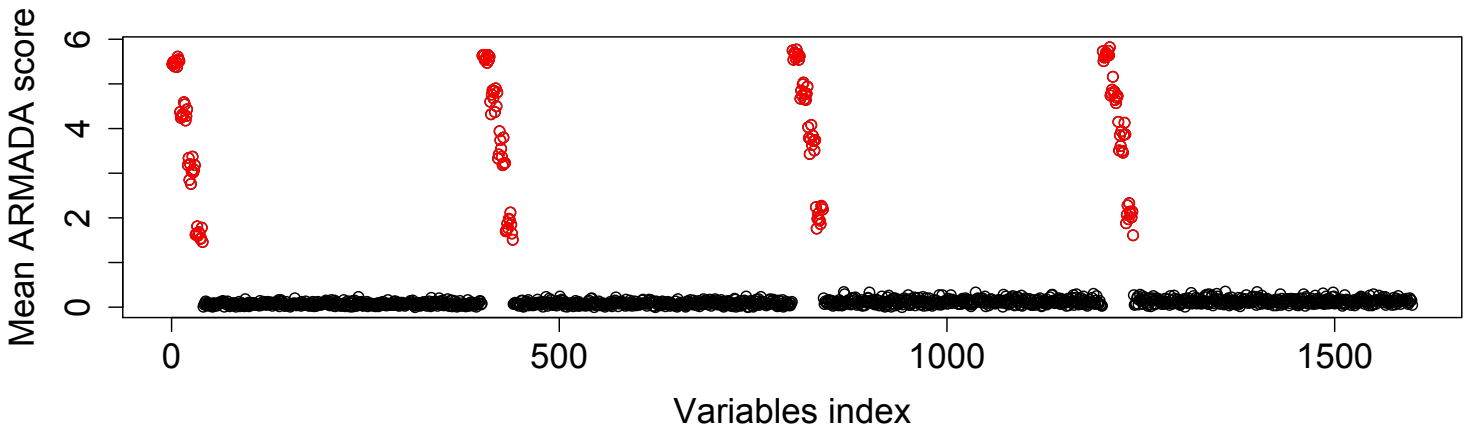


FIGURE 5 – Scénario 2. Boxplots calculés sur $N = 100$ runs de $(\mathbf{X}, Y)$. Lignes rouges : nombres attendus de TP (240) et FP $(5\% \times (1600 - 240))$

↪ Amélioration du taux de FP, par rapport à `FAMT`.

Scores obtenus après prétraitement



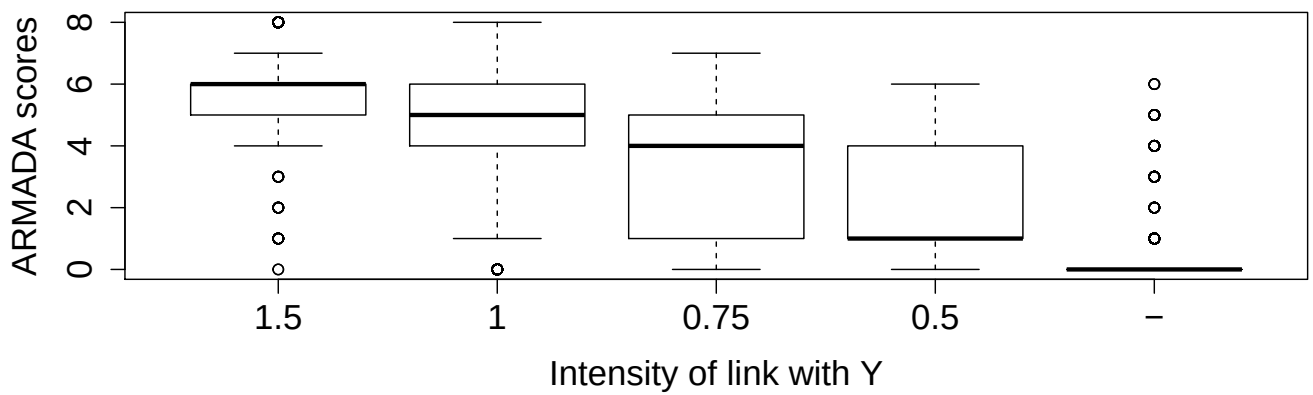


FIGURE 6 – Scénario 1 : moyennes et boxplots des scores calculés sur $N = 100$ runs de $(\mathbf{X}, Y)$.

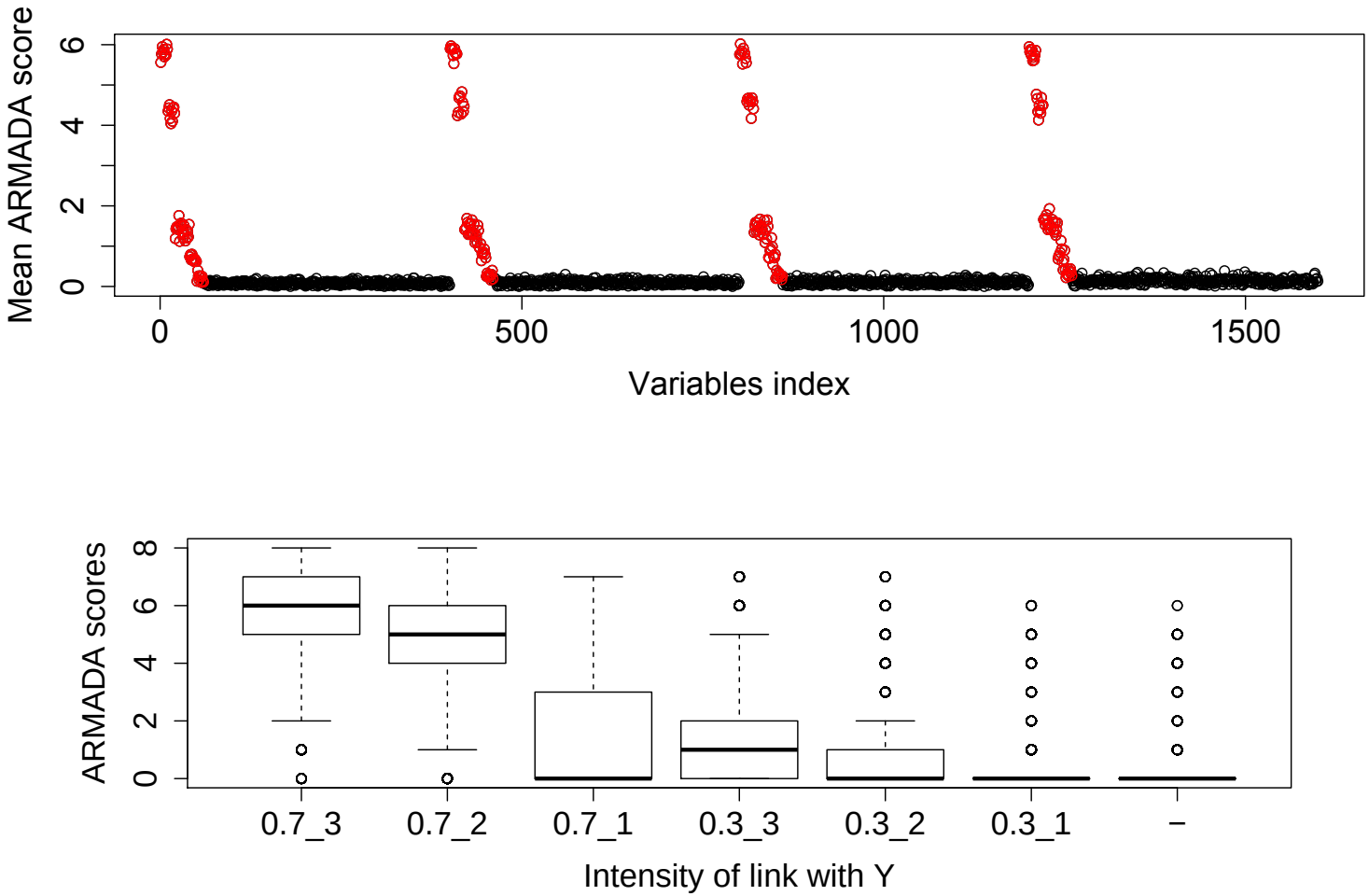
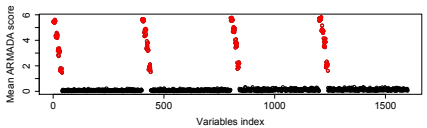


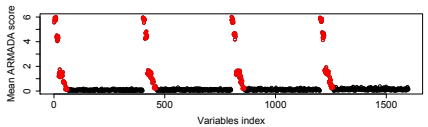
FIGURE 7 – Scénario 2 : moyennes et boxplots des scores calculés sur $N = 100$ runs de $(\mathbf{X}, Y)$.

Commentaires

- Tri des variables, selon l'intensité du lien avec Y.
- Scénario 1 : variables influentes clairement séparées du bruit.



- Scénario 2 : variables de faible influence : plus délicates à détecter.



- 95% des variables de bruit ont obtenu un score 0 (dans chaque cas).
- L'utilisateur choisit le seuil $s \in \{1, \ldots, 8\}$ de sélection des variables.

Comparaison avec d'autres méthodes de sélection

- `armada` : sélection de X_j si $S_j \geqslant 1$.
- Wilcoxon : sélection de X_j si pvaleur$(j) < \alpha$.
- `FAMT` : sélection de X_j si pvaleur ajustée $(j) < \alpha$.

	`armada`	Wilcoxon	`FAMT`
"1.5"	0.99 (0.04)	0.99 (0.07)	0.99 (0.02)
"1"	0.97 (0.15)	0.85 (0.35)	0.95 (0.20)
"0.75"	0.91 (0.27)	0.62 (0.48)	0.82 (0.38)
"0.5"	0.79 (0.40)	0.33 (0.47)	0.52 (0.49)
-	0.05 (0.23)	0.05 (0.22)	0.10 (0.30)

TABLE 1 – Taux moyens de sélection (avec écart-types), scénario 1, sur $N = 100$ runs de $(\mathbf{X}, Y)$.

$\hookrightarrow$ respect du taux de FP (5%) et meilleur taux de TP.

	armada	Wilcoxon	FAMT
(0.7-3)	0.99 (0.08)	0.99 (0.07)	0.99 (0.04)
(0.7-2)	0.92 (0.27)	0.92 (0.26)	0.96 (0.17)
(0.7-1)	0.44 (0.49)	0.43 (0.49)	0.58 (0.49)
(0.3-3)	0.54 (0.49)	0.41 (0.49)	0.61 (0.48)
(0.3-2)	0.32 (0.46)	0.28 (0.45)	0.41 (0.49)
(0.3-1)	0.12 (0.32)	0.12 (0.33)	0.19 (0.39)
-	0.05 (0.23)	0.05 (0.22)	0.09 (0.29)

TABLE 2 – Taux moyens de sélection (avec écart-types), scénario 2, sur $N = 100$ runs de $(\mathbf{X}, Y)$.

$\hookrightarrow$ respect du taux de FP (5%) et compétitif avec FAMT pour les TP.

Comparaison de méthodes : courbes ROC

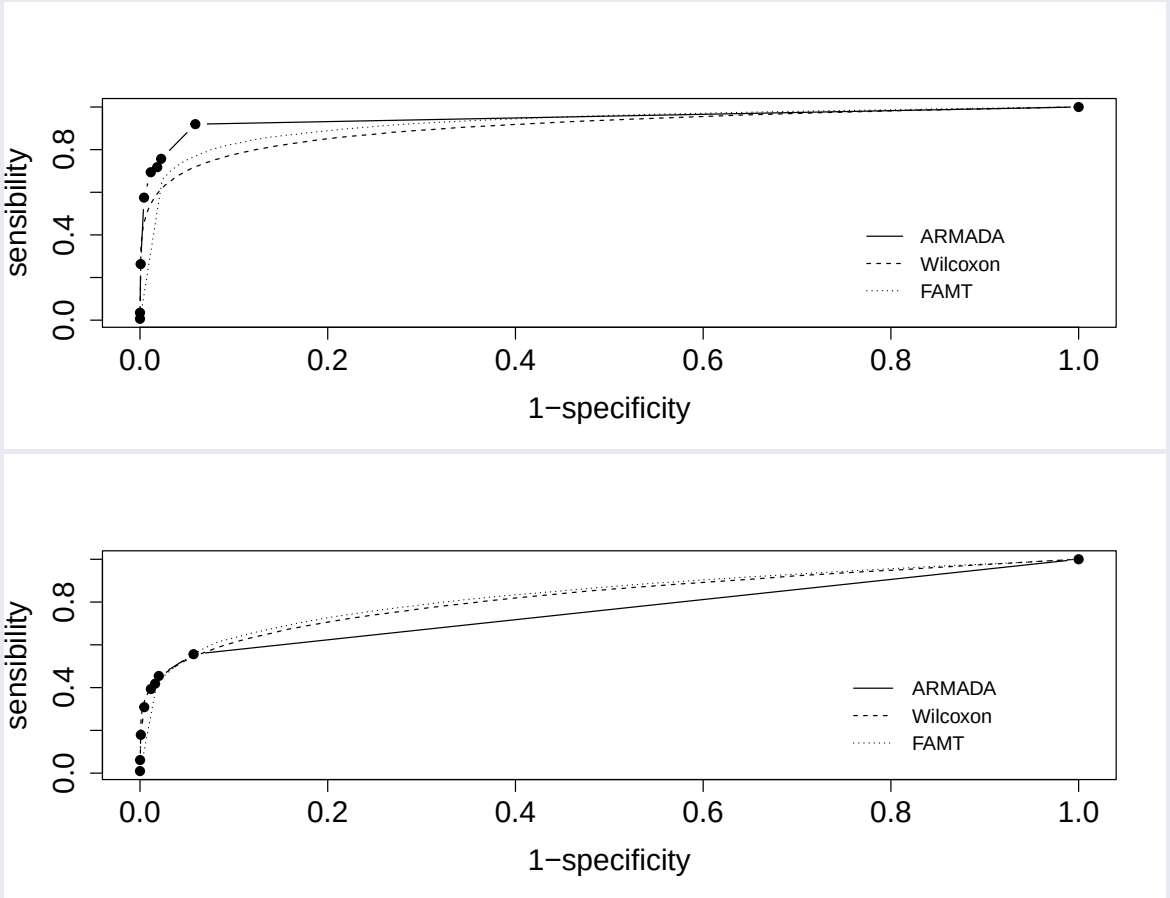


FIGURE 8 – Courbes ROC, simulation 1 (dessus), simulation 2 (dessous). Moyennes sur 100 courbes ROC obtenues sur 100 runs de $(\mathbf{X}, Y)$.

↪ Courbe ROC `armada` au dessus des autres.

Simulation en régression

Scénario

- $\mathbf{X}$ de la même taille que précédemment
- Corrélation croissante dans les 4 clusters (nulle dans le 1er cluster, forte dans le dernier cluster)

- 15 variables influentes dans les clusters 1, 2, 3 :

$$\begin{aligned} Y = & 50X_1^{(1)} + 40X_2^{(1)} + 30X_3^{(1)} + 20X_4^{(1)} + 10X_5^{(1)} \\ & + 50X_1^{(2)} + 40X_2^{(2)} + 30X_3^{(2)} + 20X_4^{(2)} + 10X_5^{(2)} \\ & + 50X_1^{(3)} + 40X_2^{(3)} + 30X_3^{(3)} + 20X_4^{(3)} + 10X_5^{(3)} + \epsilon \end{aligned}$$

avec $\epsilon \sim \mathcal{N}(0, 1)$, indépendant.

- Corrélation dans les clusters $\Longrightarrow Y$ indirectement lié à d'autres variables

Intérêt du prétraitement

On compare 3 procédures

1. rien n'est fait sur $\mathbf{X}$
2. décorrélation de $\mathbf{X}$ via `FAMT`
3. détection des 4 clusters, via `ClustOfVar`, puis décorrélation dans chaque cluster via `FAMT`

Puis tests de Pearson sur les 3 datasets issus de ces 3 procédures

$\hookrightarrow$ Sélection des variables dont les p-valeurs sont $< 5\%$.

Comparaison des 3 procédures

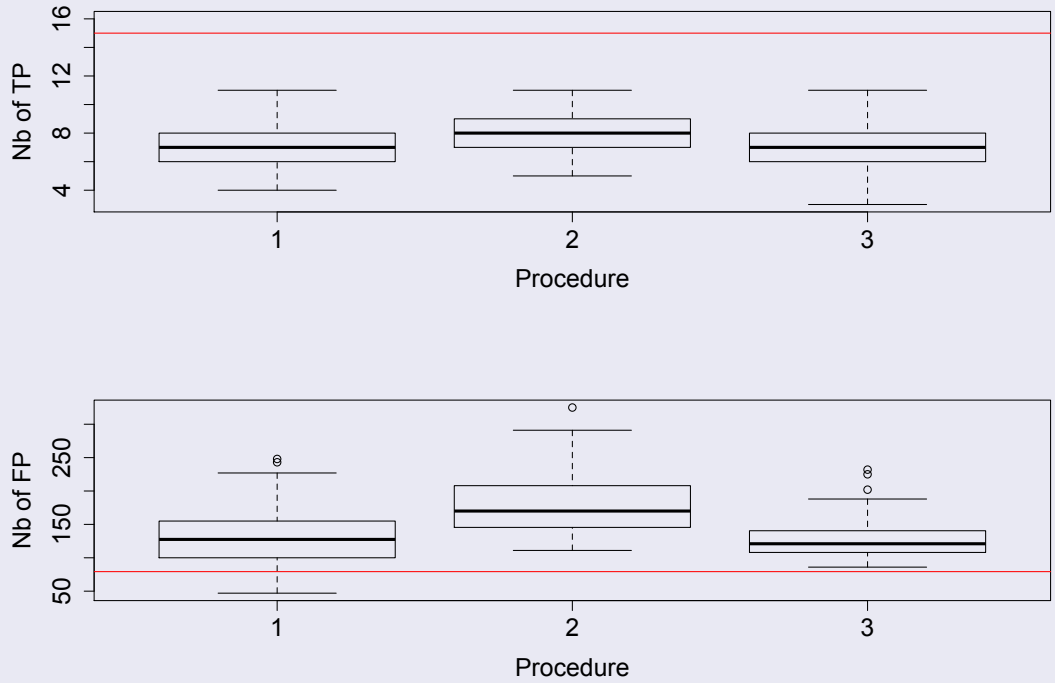
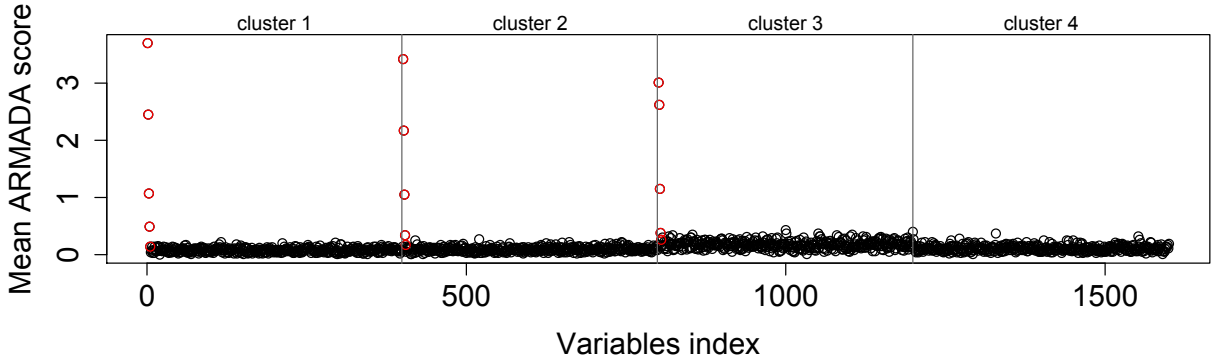


FIGURE 9 – Boxplots calculés sur $N = 100$ runs de $(\mathbf{X}, Y)$. Lignes rouges : nombres attendus de TP (15) et FP $(5\% \times (1600 - 15))$

$\hookrightarrow$ Les taux de FP sont les plus petits et les moins variables avec notre méthode.

Sélection des variables



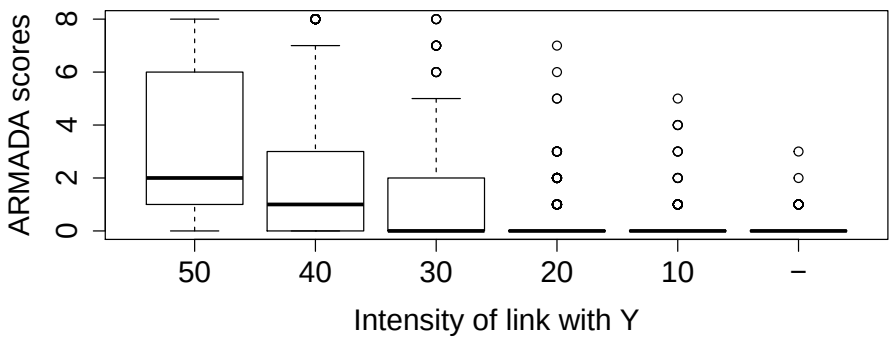


FIGURE 10 – Moyennes et boxplots des scores calculés sur $N = 100$ runs de $(\mathbf{X}, Y)$.

↪ En régression, la corrélation dans les clusters "floute" les scores.

Comparaison avec d'autres méthodes de sélection

Sélection de variables par 3 méthodes :

- `armada` : sélection de X_j si $S_j \geqslant 1$.
- Pearson : sélection de X_j si pvaleur$(j) < \alpha$.
- `FAMT` : sélection de X_j si pvaleur ajustée $(j) < \alpha$.

	ARMADA	Pearson	FAMT
50	0.83 (0.37)	0.82 (0.38)	0.87 (0.33)
40	0.71 (0.45)	0.64 (0.48)	0.75 (0.43)
30	0.45 (0.50)	0.46 (0.50)	0.55 (0.50)
20	0.22 (0.41)	0.25 (0.43)	0.31 (0.46)
10	0.11 (0.32)	0.15 (0.36)	0.17 (0.38)
-	0.07 (0.26)	0.08 (0.27)	0.11 (0.31)

TABLE 3 – Taux moyens de sélection (avec écart-types), sur $N = 100$ runs de $(\mathbf{X}, Y)$.

$\hookrightarrow$ Compétitif avec `FAMT` pour la détection des TP. Taux de FP proche de 5%.

Comparaison de méthodes : courbes ROC

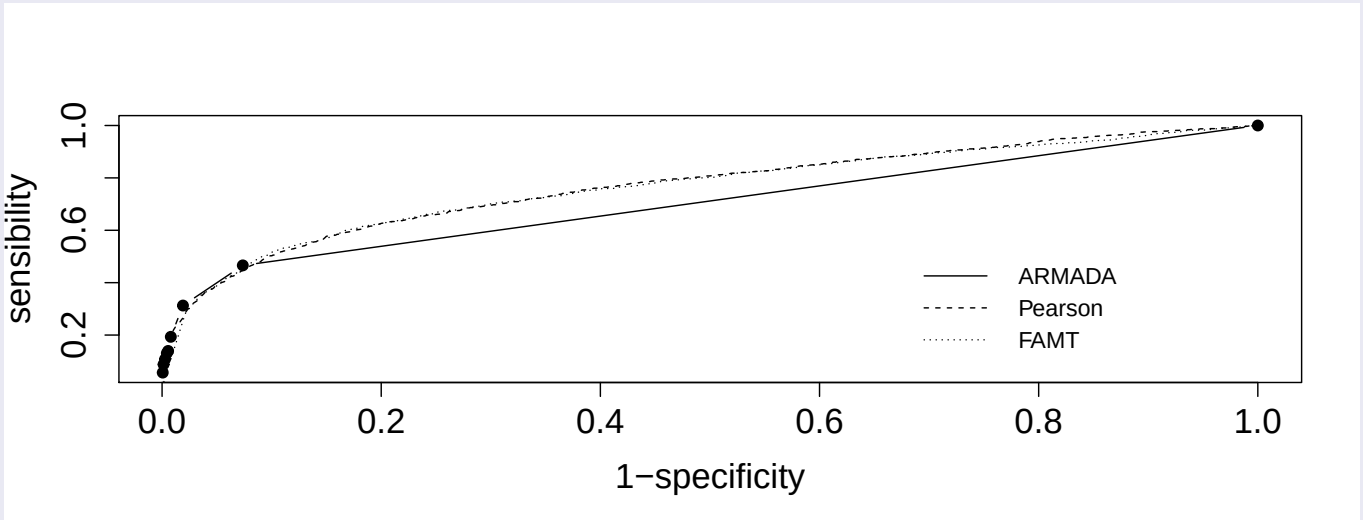


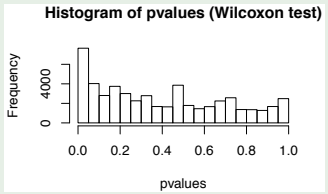
FIGURE 11 – Courbes ROC en régression. Moyennes sur 100 courbes ROC obtenues sur 100 runs de $(\mathbf{X}, Y)$.

↪ Courbe ROC `armada` au dessus des autres.

Sommaire

Data

Données transcriptomiques, patients traités par chimio

- $n = 13$ décédés à 12 mois $+ 24$ vivants à 12 mois $= 37$ patients
- $m = 51\ 336$ variables
- 

Filtre des variables avec Wilcoxon-pvalue $> 5\%$ $\Longrightarrow m = 6810$ variables

1. Etude en classification : $Y = 1/0$ (pour "décédé"/"vivant" à 12 mois après la chimio)
2. Etude en régression : $Y =$ temps de survie après la chimio (pas de censure)

Résultats

Classification $Y = 0/1$

Score	0	1	2	3	4	5	6	7	8
Nb variables	2827	553	460	596	1170	888	306	10	0

TABLE 4 – Distribution des scores des variables.

Régression $Y =$ durée de survie

Score	0	1	2	3	4	5	6	7	8
Nb variables	3988	89	456	509	984	692	86	5	1

TABLE 5 – Distribution des scores des variables.

	Classification score							
Regression score	0	1	2	3	4	5	6	7
0	2227	328	273	337	531	257	34	1
1	41	7	3	9	17	10	2	0
2	131	35	39	52	119	71	9	0
3	119	48	44	50	117	114	17	0
4	174	65	56	86	256	241	102	4
5	119	64	40	57	116	176	116	4
6	15	4	4	5	12	19	26	1
7	1	2	1	0	1	0	0	0
8	0	0	0	0	1	0	0	0

TABLE 6 – Répartition des scores des variables. 342 variables ont des scores $\geqslant 5$ en classification **et** en régression.

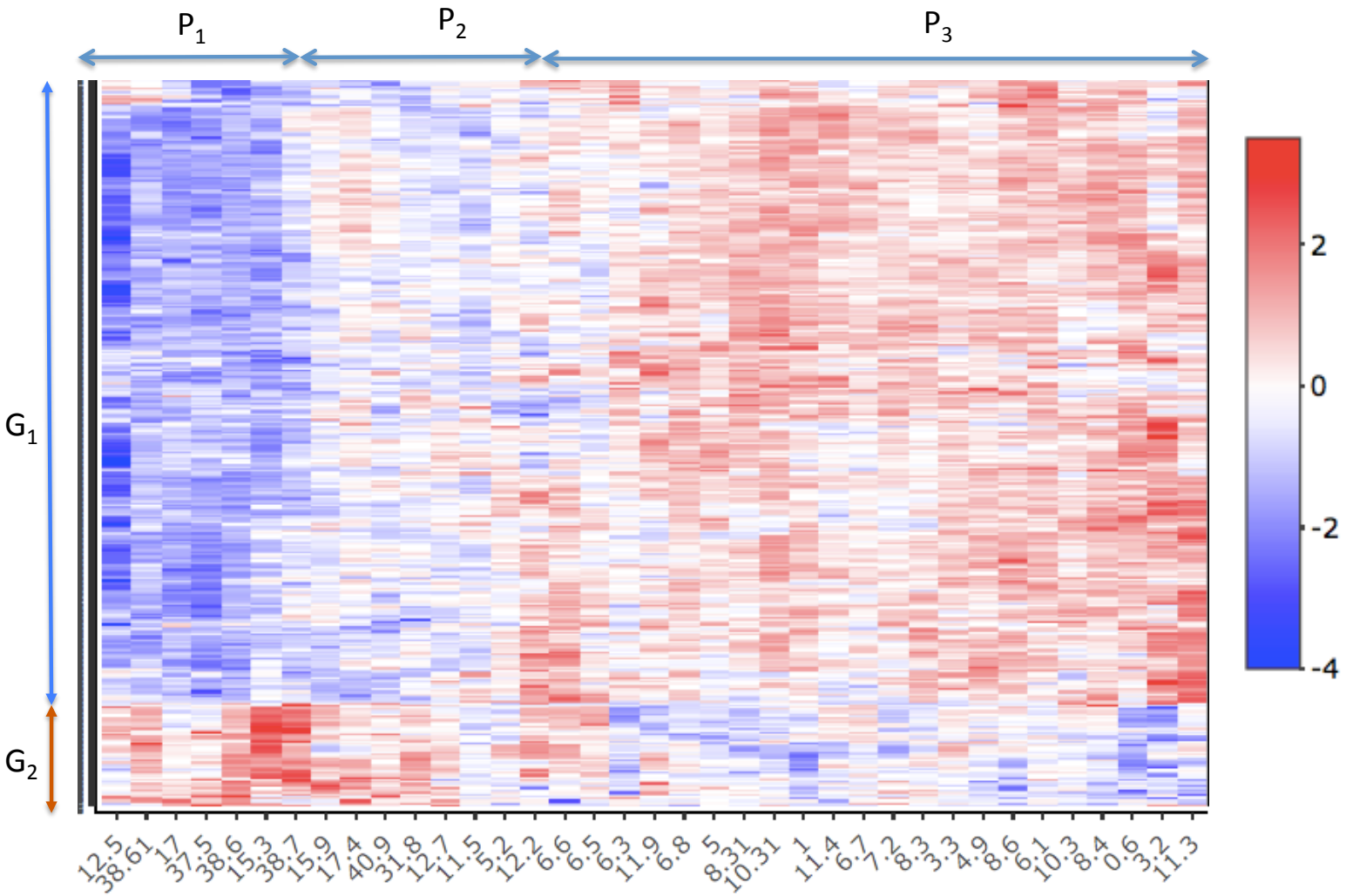


FIGURE 12 – Heatmap des 342 variables ayant des scores $\geqslant 5$ en classification **et** en régression. 1 colonne = 1 patient (marqué par son temps de survie), 1 ligne = 1 variable.

Robustesse des résultats

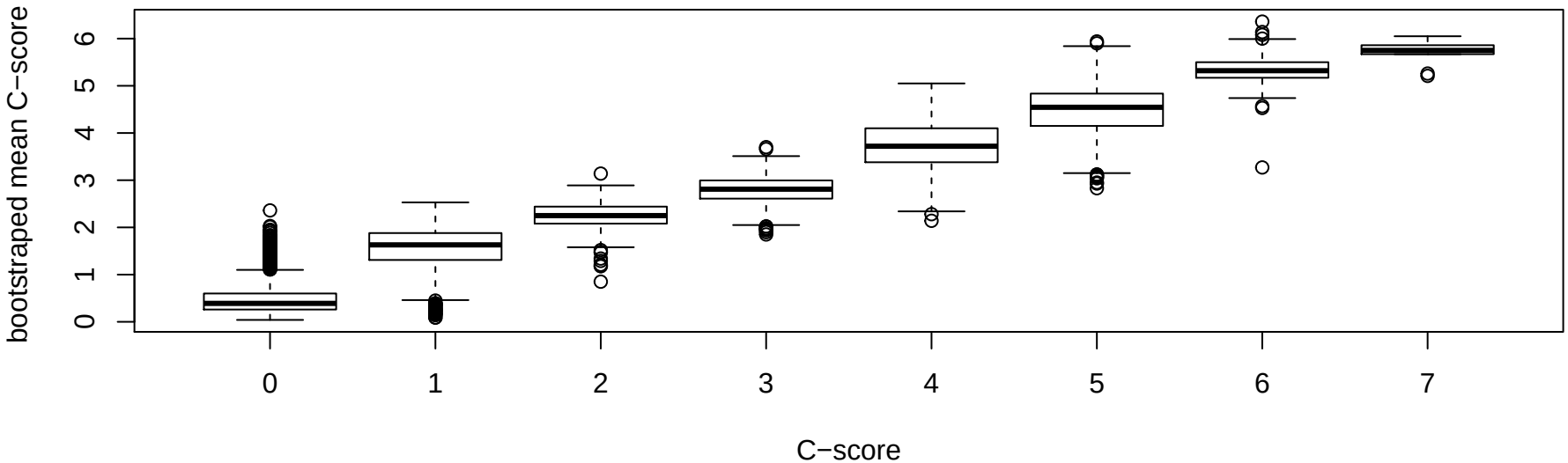


FIGURE 13 – Distribution des moyennes des C-scores obtenues sur $B = 100$ échantillons bootstrap, versus les scores originaux, pour toutes les $m = 6810$ variables.

Robustesse des résultats

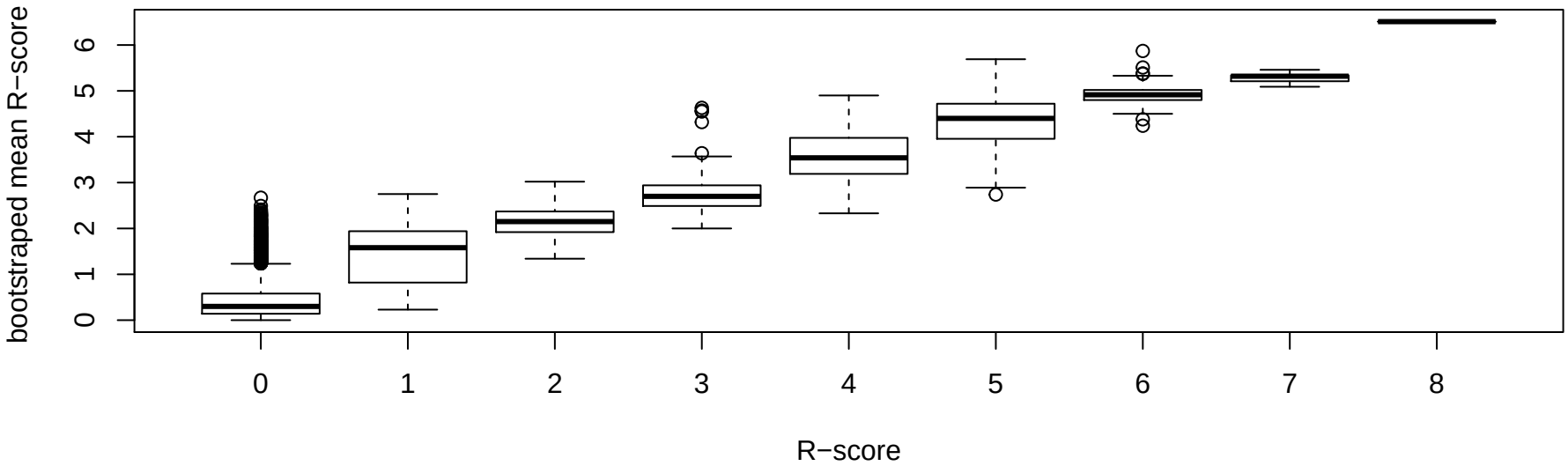


FIGURE 14 – Distribution des moyennes des R-scores obtenues sur $B = 100$ échantillons bootstrap, versus les scores originaux, pour toutes les $m = 6810$ variables.

Conclusion

- Sélection de variables (ici gènes) liées à une variable d'intérêt (ici issue d'un traitement)
- Variable d'intérêt binaire, multinomiale, ou continue

- Visualisation, sur tous les patients, des gènes sélectionnés
- Classification de profils génétiques de patients
- $\Longrightarrow$ développement de la médecine personnalisée....

- A été utilisé par des biologistes pour comprendre la fonction biologique des gènes liés à ER36 (article en cours)

Package `armada` disponible sur le CRAN.

Références ClustOfVar et FAMT


Chavent, Kuentz, Liquet, Saracco (2012),
ClustOfVar : an R package for the clustering of variables, J.of Statistical Software


Friguet, Causeur (2010)
Estimation of the proportion of true null hypotheses in high-dimensional data under dependence, CSDA

Références Tests multiples


Aubert, Bar-Hen, Daudin, Robin (2005)
Comparaisons multiples pour les microarrays, Journal de la SFDS


Benjamini, Hochberg (1995)
Controlling the false discovery rate : a practical and powerful approach to multiple testing, JRSS B


Bonferroni (1936)
Teoria statistica delle classi e calcolo delle probabilità.
Pubblicazioni del R Istituto Superiore si Scienze Economiche e Comerciali di Firenze.


Storey, Tibshirani (2003)
Statistical significance for genomewide studies, PNAS

Référence Forêts aléatoires et Régression Lasso


Genuer, Poggi, Tuleau-Malot (2010)
Variable selection using random forests, Pattern Recognition Letters


Friedman, Hastie, Tibshirani (2010)
Regularization paths for generalized linear models via coordinate descent, Journal of statistical software

Merci pour votre attention


Bastien, Chakir, Gégout-Petit, Muller-Gueudin, Shi (2018)
A statistical methodology to select covariates in high-dimensional data under dependence. Application to the classification of genetic profiles associated with outcome of a non-small-cell lung cancer treatment.