Contrôle de la Family-Wise Error Rate (FWER). Application en Métabolomique

P. Tardivel, R. Servien, C. Canlet, L. Debrauwer, M. Tremblay-Franco, D. Concordet

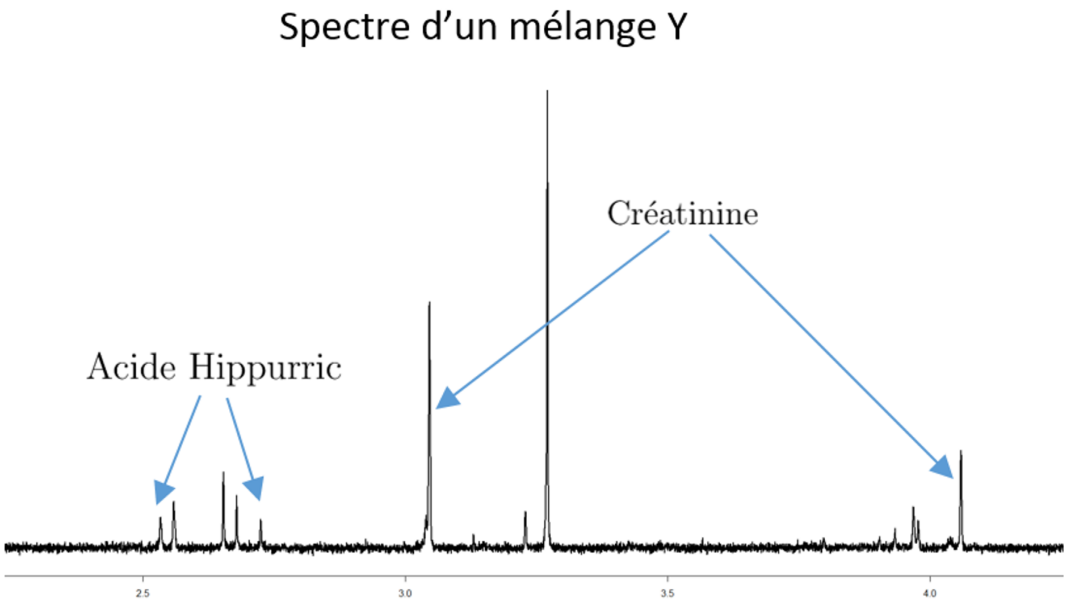
UMR 1331 Toxalim,
INRA - ENVT,
Toulouse

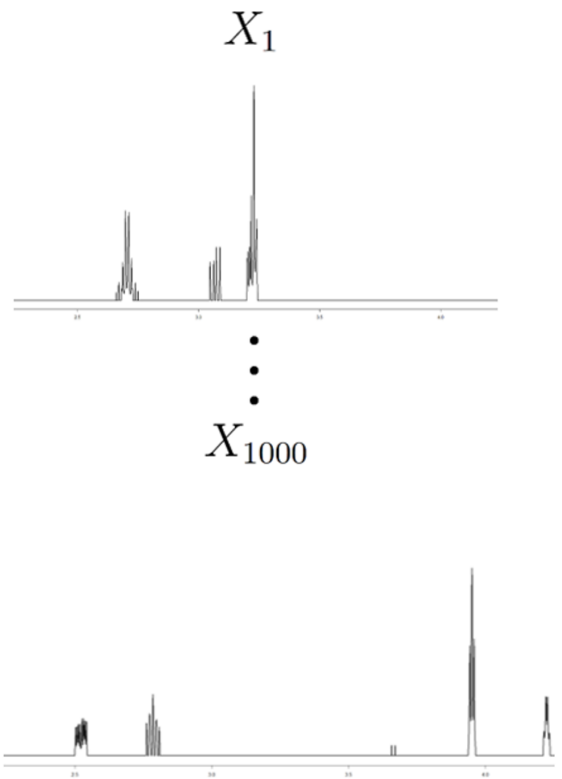
28 Avril 2017

La métabolomique

La RMN est une machine permettant l'identification de métabolites dans un mélange complexe (urine, cellules, plasma...)

Identification





Modélisation du spectre d'un mélange

Le spectre du mélange est une combinaison linéaire bruitée de spectres de métabolites.

$$\begin{aligned} Y &= \beta_1^* X_1 + \cdots + \beta_p^* X_p + \varepsilon, \text{ avec } \forall i \in \{1, \ldots, p\}, \|X_i\|_1 = 1 \\ Y &= X\beta^* + \varepsilon, \text{ avec } X = (X_1 | \ldots | X_p) \text{ et } \beta^* = (\beta_1^*, \ldots, \beta_p^*) \end{aligned}$$

Un métabolite absent du mélange a un paramètre nul.
On souhaite déterminer l'ensemble $\mathcal{A}$ des métabolites qui interviennent dans la composition du mélange

$$\mathcal{A} := \{i \in [\![1, p]\!] \mid \beta_i^* \neq 0\}$$

Tester la présence d'un métabolite

Objectif : Tester pour chaque métabolite l'hypothèse nulle $\beta_i^* = 0$ et contrôle de la FWER à un niveau α prescrit.

La FWER est la probabilité que la procédure de test multiple rejette une hypothèse nulle à tort.

La "puissance" est la proportion de rejets à raison parmi les éléments de l'active set.

Test multiple via un estimateur de type lasso

Lorsque X est une matrice $n \times p$ avec $n < p$ les méthodes traditionnelles ne sont pas utilisables. Développement de procédure de test multiple avec un estimateur de type lasso.

- Les noeuds du lasso developpé par Lockhart et al. [2014] permettent de construire une statistique de test [G'Sell et al., 2015] qui contrôle le FDR
- Le Slope Lasso développé par Bogdan et al. [2015] contrôle le FDR fonctionne lorsque X quasiment orthogonale

Test multiple lorsque X est de plein rang

Le knockoff lasso développé par Janson and Su [2016], Barber and Candès [2015] fonctionne pour une matrice X de plein rang. Cette estimateur permet de construire une procédure qui contrôle la FWER.

L'estimateur lasso a pour expression

$$\hat{\beta}(\lambda) = \underset{\beta \in \mathbb{R}^p}{\operatorname{argmin}} \left\{ \|Y - X\beta\|^2 + \lambda \|\beta\|_1 \right\}$$

On rejette l'hypothèse nulle $\beta_i^* = 0$ dès que $\hat{\beta}_i(\lambda) \neq 0$. La FWER de cette procédure est

$$\mathbb{P}(\exists i \notin \mathcal{A}, \hat{\beta}_i(\lambda) \neq 0)$$

La "puissance" de cette procédure est

$$\frac{1}{card(\mathcal{A})} \sum_{i \in \mathcal{A}} \mathbb{1}_{\hat{\beta}_i(\lambda) \neq 0}$$

X a des colonnes orthogonales : choix du λ

Lorsque X a des colonnes orthogonales ($X'X = \text{diag}(d_1, \ldots, d_p)$), l'estimateur lasso a une expression explicite [Hastie et al., 2009]

$$\hat{\beta}(\lambda) := \text{sign}(\hat{\beta}_i^{\text{ols}}) \left(|\hat{\beta}_i^{\text{ols}}| - \frac{\lambda}{d_i} \right)_+$$

Choix de λ_0 pour avoir un contrôle de la FWER à un niveau α

$$\lambda_0 := 1 - \alpha \text{ quantile de } \max\{d_1|Z_1^{\text{ols}}|, \ldots, d_p|Z_p^{\text{ols}}|\}$$

avec $Z^{\text{ols}} \sim \mathcal{N}(0, \sigma^2(X'X)^{-1})$

X quelconque

On applique une transformation linéaire $U \in G$ qui orthogonalise X

$$UY = UX\beta + U\varepsilon \text{ avec } (UX)'UX = \text{diag}(d_1, \ldots, d_p)$$

$\hat{\beta}^U(\lambda)$ est l'estimateur lasso standard de ce modèle modifié

- $\hat{\beta}^{\text{ols}}(U) = ((UX)'UX)^{-1}(UX)'Y$ estimateur des moindres carrés de ce modèle
- $\lambda_0(U)$ paramètre de régularisation de ce modèle

L'estimateur lasso a une expression explicite

$$\hat{\beta}_i^U(\lambda_0(U)) = \text{sign}(|\hat{\beta}_{\text{i}}^{\text{ols}}(\text{U})|)(\hat{\beta}_{\text{i}}^{\text{ols}}(\text{U}) - \lambda_0(\text{U})/\text{d}_{\text{i}})_+.$$

On note $\phi(U) := \phi(\lambda_0(U)/d_1, \ldots, \lambda_0(U)/d_p)$, avec ϕ une norme
Pour avoir une procédure puissante, on veut que $\phi(U)$ soit petit.

Transformation linéaire optimale

Théorème (Tardivel et al.)

ϕ une norme sur $\mathbb{R}^p$, Il existe $U^* \in G$ tel que

$$\forall U \in G, \phi(U^*) \leq \phi(U).$$

De plus, $\mathrm{var}(\hat{\beta}^{ols}(\mathrm{U}^*)) = \sigma^2(\mathrm{X}^\mathrm{T}\mathrm{X})^{-1}$

La construction de cette transformation linéaire est donnée dans Tardivel et al. [2016]

Récapitulatif pour l'emploi de notre procédure

1. On se donne une norme ϕ (en pratique la norme l^1)
2. On détermine la transformation linéaire optimale U^*
3. On se ramène au cas ou la matrice de planification est orthogonale
4. On obtient l'estimateur lasso
   $\hat{\beta}_i^{U^*}(\lambda) = \mathrm{sign}(\hat{\beta}_i^{ols}(U^*))(|\hat{\beta}_i^{ols}(U^*)| - \lambda/d_i^*)_+$
5. On calcule $\lambda_0(U^*)$ par simulation

Notre procédure de test multiple rejette $\beta_i^* = 0$ dès que

$$\hat{\beta}_i^{U^*}(\lambda_0(U^*)) \neq 0$$

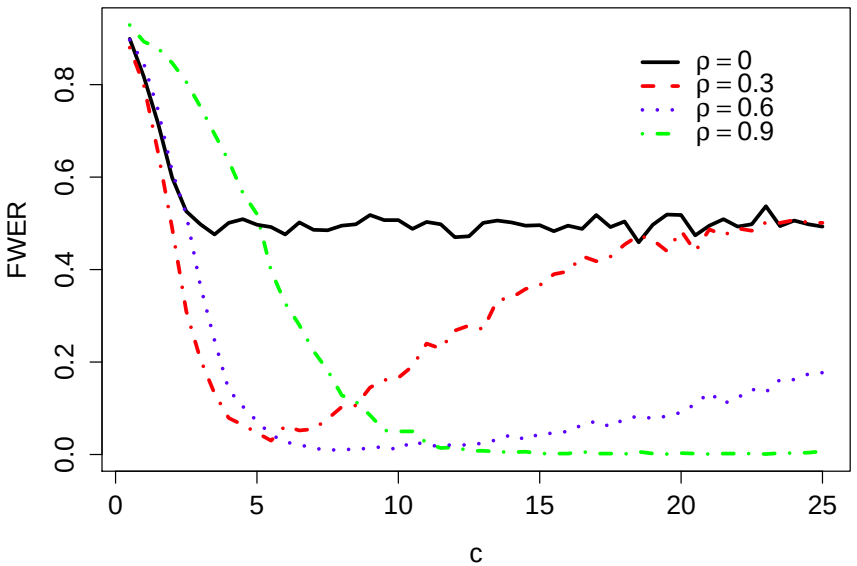
Comparaison avec le knockoff lasso (FWER=0.5)

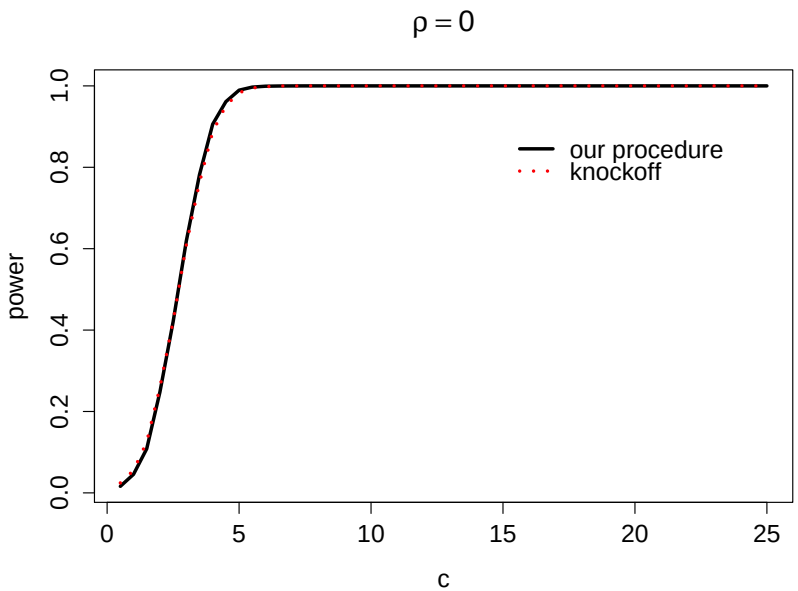
- $\beta_1^* = \cdots = \beta_{10}^* = c, \beta_{11}^* = \cdots = \beta_{100}^* = 0$
- $\sigma^2(X^TX)^{-1} = \Sigma$ avec $\forall i, \Sigma_{i,i} = 1,\ i \neq j, \Sigma_{i,j} = \rho$.

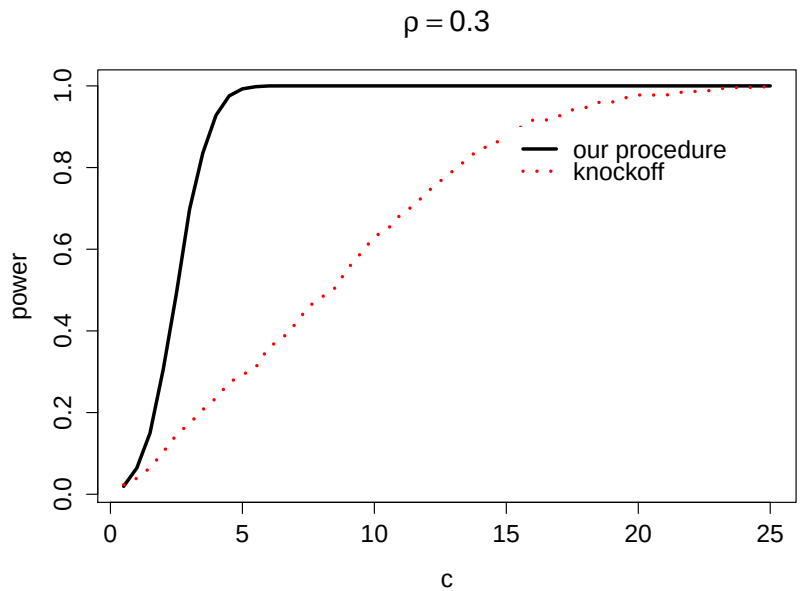
Le contrôle de la FWER de notre méthode ne dépend pas de c :

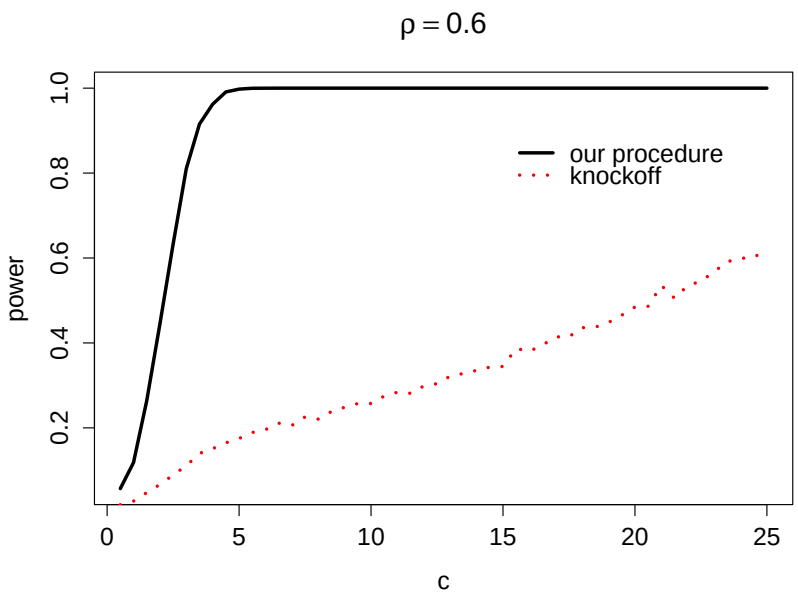
ρ	0	0.3	0.6	0.9
FWER	0.462	0.477	0.482	0.477

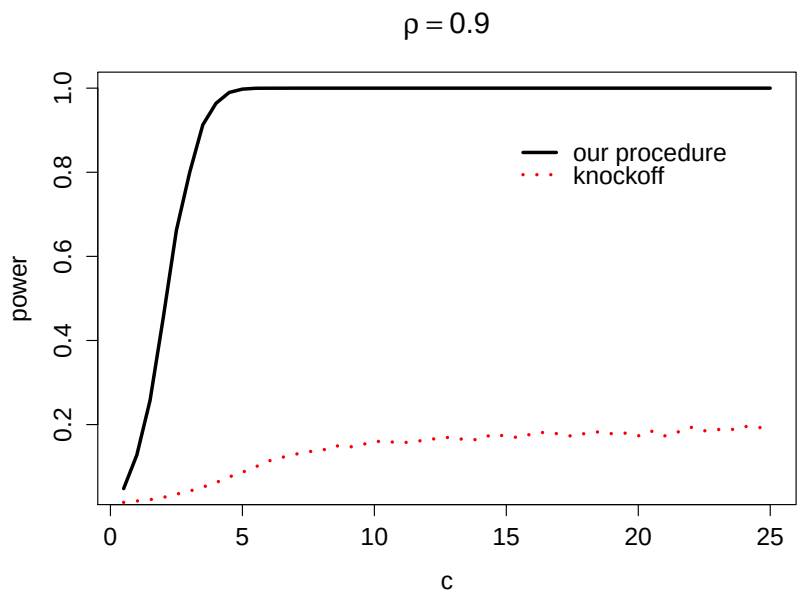
Le contrôle de la FWER de la méthode knockoff dépend de c :




ρ = 0
power
c
our procedure
knockoff


ρ = 0.3
power
c
our procedure
knockoff

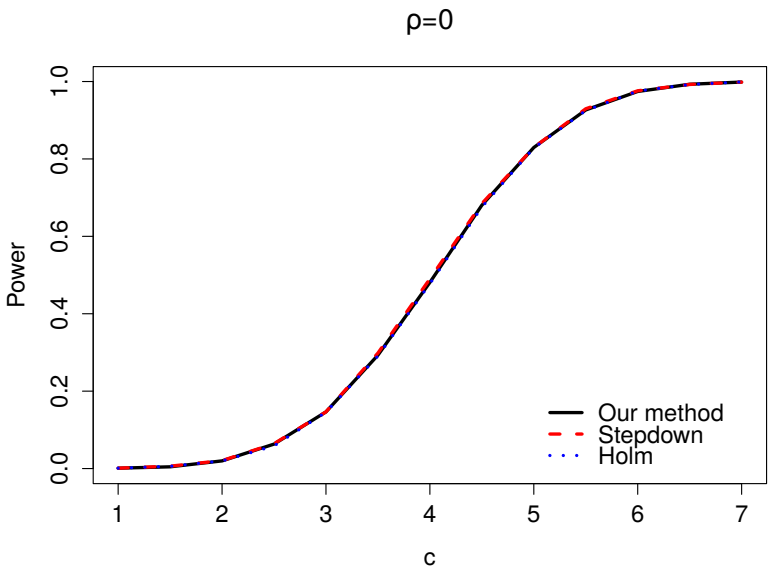

ρ = 0.6
power
c
our procedure
knockoff

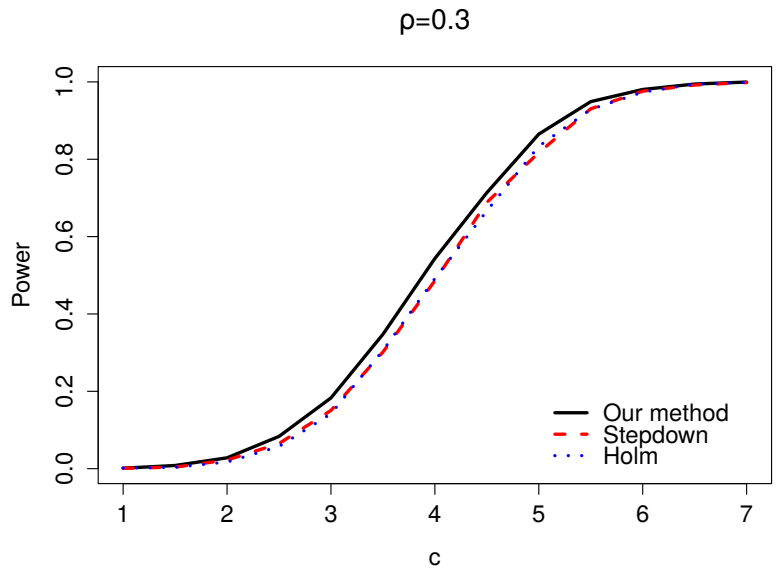

ρ = 0.9
power
c
our procedure
knockoff

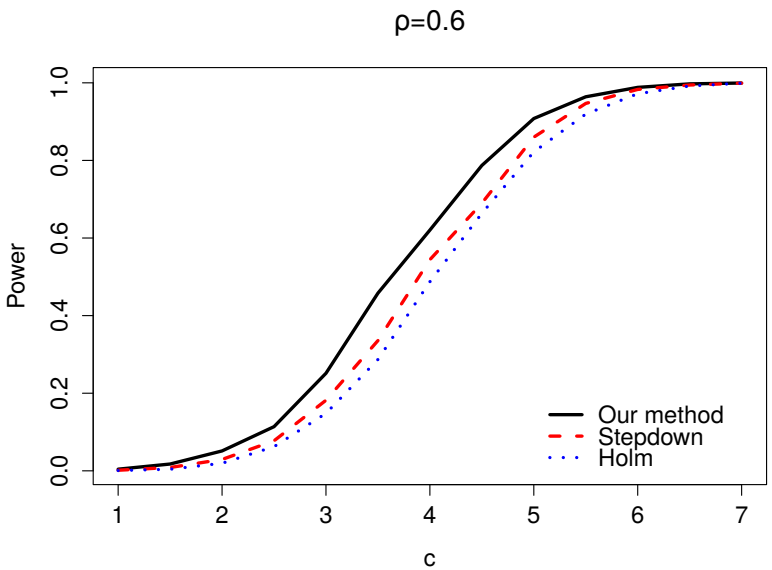
Comparaison avec les procédures standard (FWER=0.05)

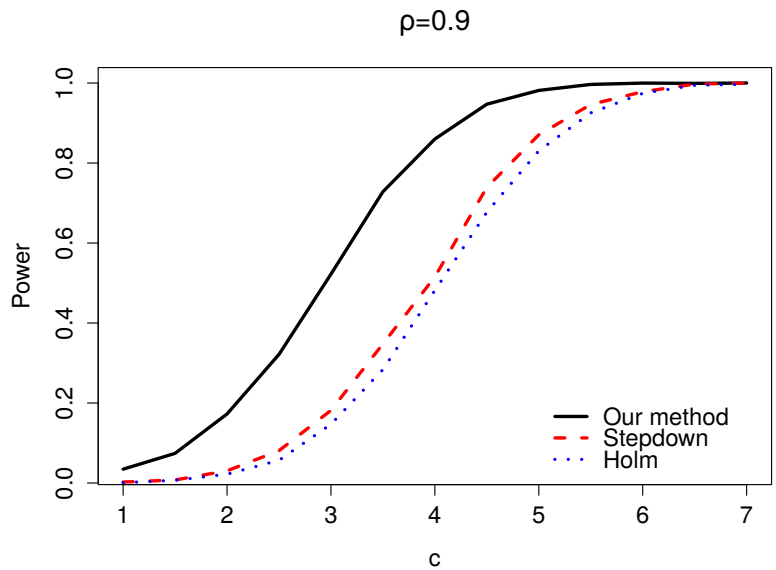
- $\beta_1^* = \cdots = \beta_{20}^* = c, \beta_{21}^* = \cdots = \beta_{1000}^* = 0$
- $\sigma^2(X^TX)^{-1} = diag(\Sigma, Id_{500})$ avec $\Sigma \in M_{500}(\mathbb{R})$ telle que $\forall i, \Sigma_{i,i} = 1$ et $i \neq j, \Sigma_{i,j} = \rho$

	Contrôle de la FWER			
	$\rho = 0$	$\rho = 0.3$	$\rho = 0.6$	$\rho = 0.9$
Holm	0.0496	0.0430	0.034	0.0286
Generic stepdown	0.0491	0.0498	0.0491	0.0505
Our procedure	0.0483	0.0487	0.0502	0.0540


ρ=0
Power
c
Our method
Stepdown
Holm

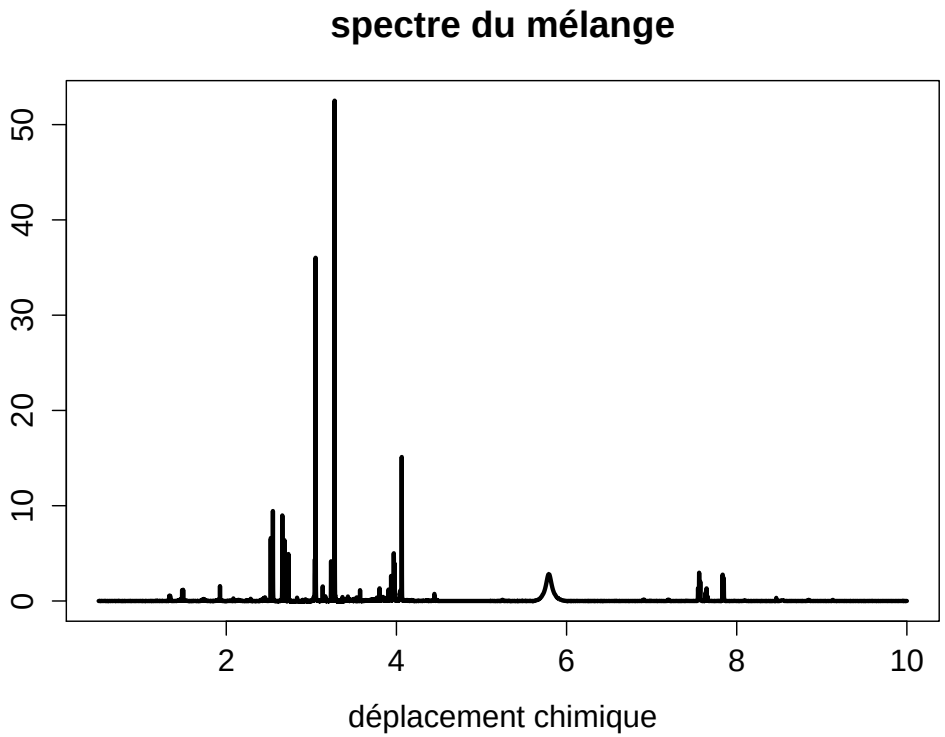

ρ=0.3
Power
c
Our method
Stepdown
Holm


ρ=0.6
Power
c
Our method
Stepdown
Holm


ρ=0.9
Power
c
Our method
Stepdown
Holm

Application : identification de métabolites dans un mélange

Analyse en aveugle d'un mélange de composition connue à l'aide d'une bibliothèque contenant 175 spectres [Tardivel et al., 2017]



Comparaison avec Bayesil, Chenomx, BATMAN et Metabohunter

Tableau des quantifications

Les proportions sont relatives à celle de la Créatinine

Composition	Proportion	NOUS	Bayesil	Chenomx	BATMAN
Creatinine	1	1	1	1	1
Citric acid	0.434	0.693	12.38	0	0.089
Hippuric acid	0.338	0.344	0	0.312	0.072
⋮	⋮	⋮	⋮	⋮	⋮
Ascorbic acid	0.156	0	0	0	0.568

Tableau donnant les bonnes et mauvaises identifications

	I et P	I et NP	NI et P	NI et NP
NOUS	17/21	10	4/21	145
MetaboHunter	4/21	51	17/21	795
Batman	21/21	125	0/21	1
Bayesil	12/21	17	7/21	53
Chenomx	15/21	48	6/21	269

I=Identifié, NI=Non Identifié,
P=Présent dans le mélange, NP=Non Présent dans le mélange

Temps de calcul de notre méthode : environ 2 minutes

Conclusion et perspective

- Développement d'une procédure de test multiple qui contrôle la FWER.
- Procédure puissante par rapport aux méthodes existantes.
- Développement d'une méthode d'identification de métabolites performante.

On souhaiterait étendre nos résultats pour une matrice de planification X de grande dimension.

Rina Foygel Barber and Emmanuel J Candès. Controlling the false discovery rate via knockoffs. The Annals of Statistics, 43(5) : 2055–2085, 2015.

Małgorzata Bogdan, Ewout van den Berg, Chiara Sabatti, Weijie Su, and Emmanuel J Candès. Slope - adaptive variable selection via convex optimization. The Annals of Applied Statistics, 9(3) : 1103–1140, 2015.

Max Grazier G'Sell, Stefan Wager, Alexandra Chouldechova, and Robert Tibshirani. Sequential selection procedures and false discovery rate control. Journal of the Royal Statistical Sociey : Series B (Statistical Methodology), 78(2) :423–444, 2015.

Trevoir Hastie, Rob Tibshirani, and Jerome Friedman. The Elements of Statistical Learning : Data Mining, Inference, and Prediction, Second Edition. Springer Series in Statistics. Springer, 2009. ISBN 9780387848587.

Lucas Janson and Weijie Su. Familywise error rate control via knockoffs. Electronic Journal of Statistics, 10(1) :960–975, 2016.

Richard Lockhart, Jonathan Taylor, Ryan J Tibshirani, and Robert Tibshirani. A significance test for the lasso. The Annals of Statistics, 42(2) :413–468, 2014.

Patrick JC Tardivel, Rémi Servien, and Didier Concordet. Non-asymptotic active set properties of lasso-type estimators in small-dimension. 2016.

Patrick J.C. Tardivel, Cécile Canlet, Gaëlle Lefort, Marie Tremblay-Franco, Laurent Debrauwer, Didier Concordet, and Rémi Servien. ASICS : an automatic method for identification and quantification of metabolites in NMR 1D ^{1}H spectra. Submitted, 2017.