Classification non supervisée et sélection de variables dans les modèles mixtes fonctionnels. Applications à la biologie moléculaire.

Madison Giacofci
LJK (Grenoble)

Sophie Lambert-Lacroix (TIMC - Grenoble)
Franck Picard (LBBE - Lyon)

Soutenance de thèse
22 octobre 2013

Plan de la présentation

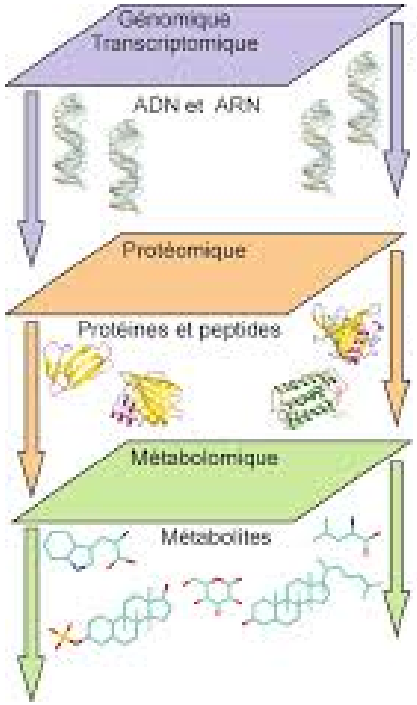
Plan de la présentation

Contexte de la biologie moléculaire

Le statisticien dispose à l'heure actuelle de quantités de données toujours plus grandes

$\Rightarrow$ Fléau de la grande dimension

Ceci est particulièrement vrai dans le domaine de la biologie moléculaire avec le développement des technologies "omiques"



Les questions statistiques restent standards :

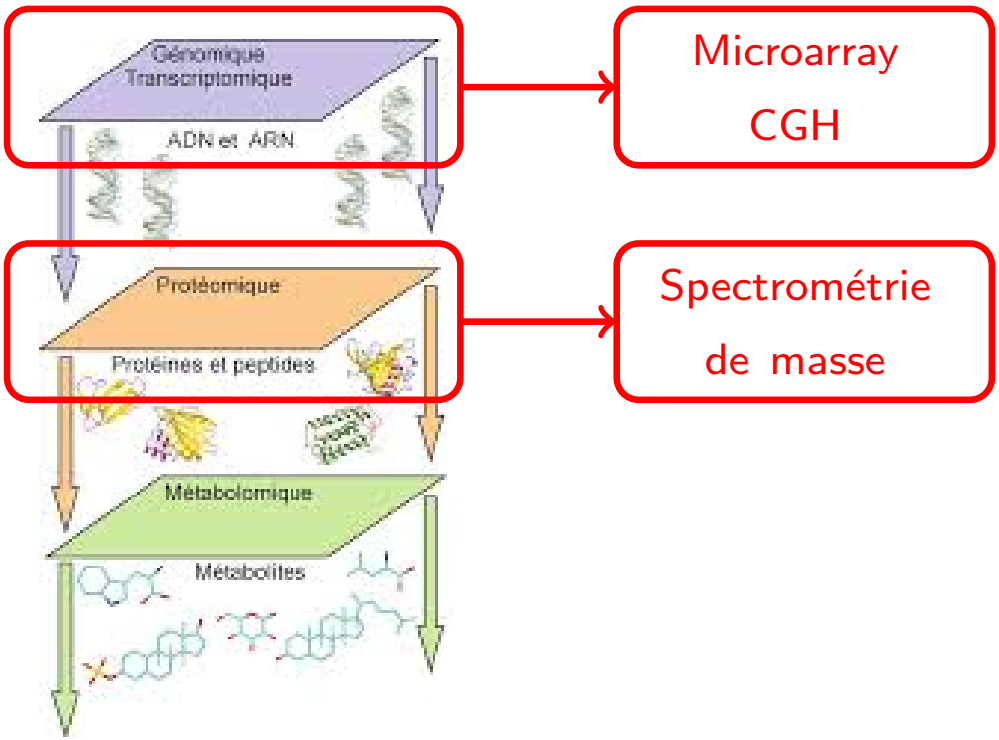
- classification (supervisée ou non)
- modélisation
- estimation
- ...

Contexte de la biologie moléculaire

Le statisticien dispose à l'heure actuelle de quantités de données toujours plus grandes

$\Rightarrow$ Fléau de la grande dimension

Ceci est particulièrement vrai dans le domaine de la biologie moléculaire avec le développement des technologies "omiques"



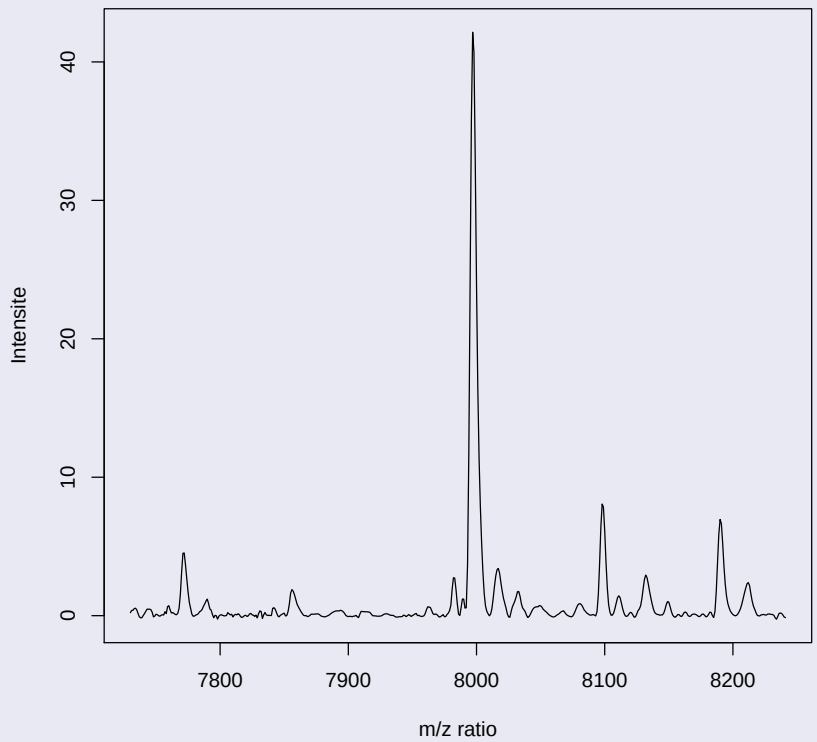
Les questions statistiques restent standards :

- classification (supervisée ou non)
- modélisation
- estimation
- ...

Données de spectrométrie de masse

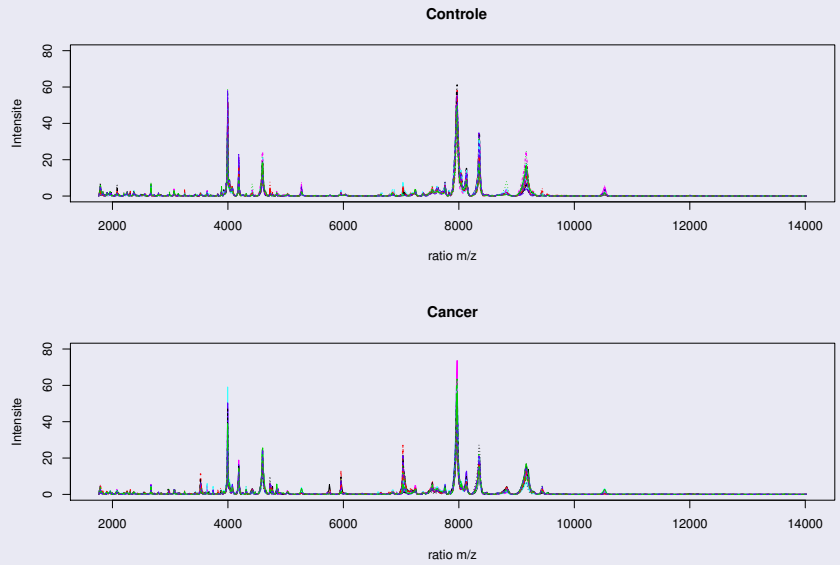
Données liées à l'étude du protéome permettant d'obtenir une image plus complète des processus biologiques impliqués dans le développement d'une pathologie

- Chaque abscisse caractérise un unique peptide par son ratio m/z
- La hauteur d'un pic est associée à la quantité d'un peptide au sein de l'échantillon
- Les spectres sont caractérisés par la présence de pics
- Les signaux produits contiennent entre 10^3 et 10^6 points de mesures



Données de cancer de l'ovaire (Petricoin et al., 2002)

- Composé de 253 spectres sous 2 conditions
  - Individus sains (91)
  - Atteints d'un cancer (162)
- Chaque spectre est composé de la mesure de l'intensité de 15154 peptides



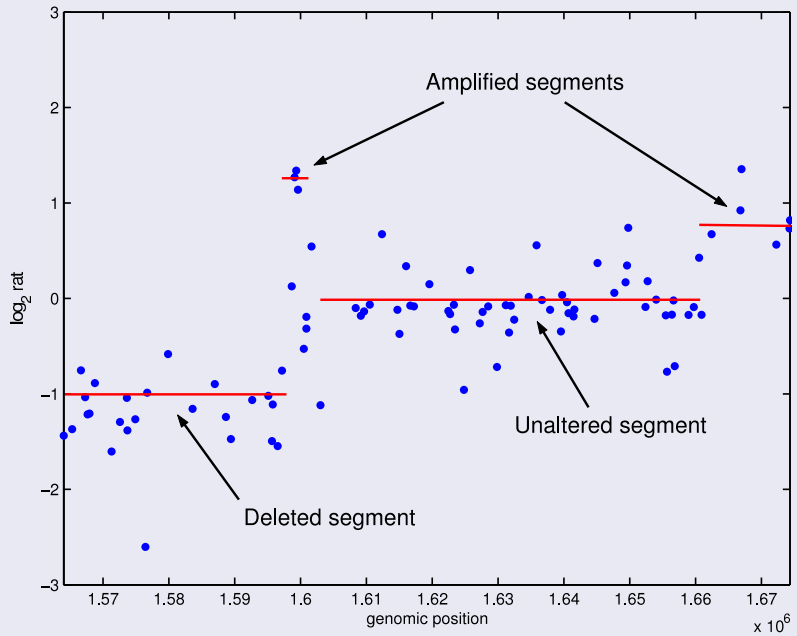
Enjeux statistiques

- Découverte de sous-groupes à partir des données moléculaires
  ▷ Classification non supervisée
- Profil protéomique moyen
  ▷ Estimation/Reconstruction
- Prise en compte d'une variabilité inter-individuelle
  ▷ Modélisation

Données de microarray CGH

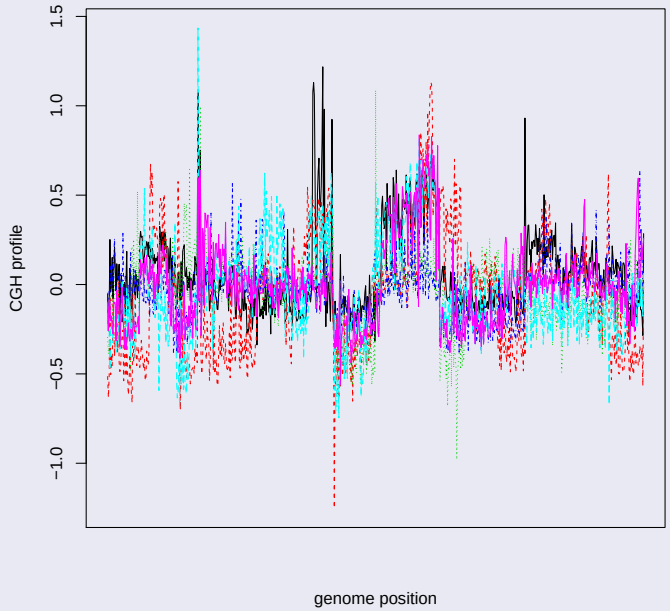
Données liées à l'étude du génome permettant de détecter la présence d'éventuelles aberrations génomiques

- Chaque point représente le ratio du nombre de copies d'une portion d'ADN entre une cellule de contrôle et une cellule à tester
- L'analyse d'un génome individuel entier produit des jeux de données contenant entre 10^3 et 10^5 points
- Les spectres obtenus sont de forme constante par morceaux



Données de cancer du sein (Fridlyand et al., 2006)

- Composé de 66 profils CGH d'individus atteints d'un cancer du sein
- Chaque spectre est composé de 2044 ratios de nombre de copies des gènes
- Données cliniques comme le grade, le stade, la récurrence, données de survie...



Enjeux statistiques

- Recherche de variantes inconnues de la pathologie
  ▷ Classification non supervisée
- Reconstruction des profils génomiques ▷ Segmentation/Estimation
- Quantification des variabilités ▷ Modélisation mixte

Modélisation de données omiques

- Chaque individu est représenté par un spectre de données mesurées à haut-débit, de manière régulière

▷ Cadre de modélisation fonctionnelle

- Chaque spectre individuel est caractérisé par la présence de discontinuités qui sont informatives

▷ Utilisation des bases d'ondelettes

- La nature des données observées "implique" la présence d'une grande variabilité due à l'individu

▷ Contexte des modèles mixtes

Plan de la présentation

Modèle fonctionnel

Approche fonctionnelle (Ramsay et Silverman, 1997)

On parle de données fonctionnelles lorsque :

- les observations sont mesurées sur une grille fine et régulière
- l'unité d'observation idéale est la courbe

Modèle de régression fonctionnelle

Les données sont vues comme N courbes observées de manière bruitée sur une grille de discrétisation $(t_1, \ldots, t_M)$ telle que pour $i = 1, \ldots, N$ et pour $m = 1, \ldots, M$:

$$Y_i(t_m) = \mu(t_m) + E_i(t_m), \qquad E(t_m) \sim \mathcal{N}(0, \sigma_E^2)$$

$\Rightarrow$ Objectif : Retrouver la fonction μ à partir de l'observation de signaux bruités.

Modèle mixte fonctionnel

Modélisation de la présence d'une variabilité spécifique due à l'individu : introduction d'effets aléatoires fonctionnels dans le modèle fonctionnel

$$Y_i(t_m) = \mu(t_m) + U_i(t_m) + E_i(t_m)$$

où $U_i \sim \mathcal{N}(0, K(s,t))$ est un processus Gaussien centré, indépendant de E_i.

Modèle de classification non supervisée de courbes

Dans un contexte de classification non supervisée, les N individus sont supposés provenir de L groupes différents.

$$Y_i(t_m)|\{\zeta_{i\ell} = 1\} = \mu_\ell(t_m) + U_i(t_m) + E_i(t_m)$$

où $\zeta_{i\ell} = 1$ si l'individu i est dans la classe ℓ.

Approche non paramétrique

Contexte non-paramétrique

On ne spécifie pas de forme particulière pour les fonctions du modèle

▷ Le problème se place en dimension infinie

Approche usuelle

Projeter le modèle sur une base fonctionnelle bien choisie

Exemples : Bases de splines, polynomiales, de Fourier, d'ondelettes

Avantages des ondelettes

- ▷ Modélisation de courbes possédant des irrégularités
- ▷ Représentation parcimonieuse des signaux réguliers
- ▷ Propriétés décorrélantes
- ▷ Algorithme de décomposition efficace
- ▷ Contrôle fin de la régularité

Ondelettes : quelques brefs rappels

Construction des bases d'ondelettes

Construction de bases orthonormées de $L^2(\mathbb{R})$ obtenues en dilatant et translatant une fonction d'échelle ϕ et une ondelette mère ψ :

$$\left\{\phi_{j_0 k}(t), k = 0, \ldots, 2^{j_0} - 1; \;\; \psi_{jk}(t), j \geq j_0, k = 0, \ldots, 2^j - 1\right\}$$

avec $\psi_{jk}(t) = 2^{\frac{j}{2}} \psi(2^j t - k)$ et $\phi_{jk}(t) = 2^{\frac{j}{2}} \phi(2^j t - k)$

Projection dans la base d'ondelettes

Toute fonction $f \in L^2(\mathbb{R})$ peut être décomposée dans la base d'ondelettes :

$$f(t) = \sum_{k=0}^{2^{j_0}-1} c^*_{j_0 k} \phi_{j_0 k}(t) + \sum_{j \geq j_0} \sum_{k=0}^{2^j - 1} d^*_{jk} \psi_{jk}(t)$$

où $c^*_{j_0 k} = \langle f, \phi_{j_0 k} \rangle$ et $d^*_{jk} = \langle f, \phi_{jk} \rangle$ sont, respectivement, les coefficients d'échelle et d'ondelettes théoriques.

Transformée en ondelettes discrètes

En présence de signaux discrets $\mathbf{Y} = (Y(t_1), \ldots, Y(t_M))$, un outil populaire est donné par la transformée en ondelettes discrètes (DWT) (Mallat, 2008) :

$$\underset{[M\times M]}{\mathbf{W}} \underset{[M\times 1]}{\mathbf{Y}} = \begin{bmatrix} \mathbf{c} \\ \mathbf{d} \end{bmatrix}$$

avec $\mathbf{W}$ matrice de filtres dépendant de la base d'ondelettes

$(\mathbf{c}, \mathbf{d})$ sont les coefficients d'échelle et d'ondelettes empiriques de la décomposition :

$$\begin{aligned} \mathbf{c} &\simeq \sqrt{M} \times \mathbf{c}^* \\ \mathbf{d} &\simeq \sqrt{M} \times \mathbf{d}^* \end{aligned}$$

Représentation du modèle dans le domaine des ondelettes

Sachant que $\{\zeta_{i\ell} = 1\}$:

$$W Y_i(t_m) = W \mu_\ell(t_m) + W U_i(t_m) + W E_i(t_m)$$

$$\Longleftrightarrow \quad \begin{bmatrix} \mathbf{c}_i \\ \mathbf{d}_i \end{bmatrix} = \begin{bmatrix} \boldsymbol{\alpha}_\ell \\ \boldsymbol{\beta}_\ell \end{bmatrix} + \begin{bmatrix} \boldsymbol{\nu}_i \\ \boldsymbol{\theta}_i \end{bmatrix} + \boldsymbol{\varepsilon}_i$$

où $\begin{cases} \boldsymbol{\varepsilon}_i \sim \mathcal{N}(0, \sigma_\varepsilon^2 \mathbf{I}) \\ \begin{bmatrix} \boldsymbol{\nu}_i \\ \boldsymbol{\theta}_i \end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \mathbf{G} = \begin{bmatrix} \mathbf{G}_{\boldsymbol{\nu}} & 0 \\ 0 & \mathbf{G}_{\boldsymbol{\theta}} \end{bmatrix} \right) \end{cases}$ et $\varepsilon^i \perp (\boldsymbol{\nu}_i^T, \boldsymbol{\theta}_i^T)^T$

Dans le domaine des ondelettes, le modèle mixte fonctionnel se ramène à un modèle linéaire mixte diagonal

Modélisation de la matrice **G**

Motivations

On souhaite une modélisation de la variabilité des effets aléatoires permettant d'obtenir :

- une structure "simple" tout en assurant une flexibilité suffisante
- des effets fixes et aléatoires partageant la même régularité (Antoniadis & Sapatinas, 2007)

Idée naturelle

Proposer un modèle pour la fonction $K(s,t)$ et en déduire des conditions sur la matrice **G**, mais cela entraîne en général des difficultés pour :

- le contrôle du nombre de paramètres du modèle
- le contrôle de la régularité des trajectoires

Modélisation de la matrice **G**

Motivations

On souhaite une modélisation de la variabilité des effets aléatoires permettant d'obtenir :

- une structure "simple" tout en assurant une flexibilité suffisante
- des effets fixes et aléatoires partageant la même régularité (Antoniadis & Sapatinas, 2007)

Idée naturelle

Proposer un modèle pour la fonction $K(s,t)$ et en déduire des conditions sur la matrice **G**, mais cela entraîne en général des difficultés pour :

- le contrôle du nombre de paramètres du modèle
- le contrôle de la régularité des trajectoires

On préfère donner une modélisation de la matrice **G** dans le domaine des ondelettes suivant l'idée d'Antoniadis & Sapatinas (2007)

Hypothèses considérées

Structure de la matrice

G est supposée diagonale

▷ justifiée par la propriété décorrélante des ondelettes (Frazier et al, 1992)

Décroissance des paramètres de variances

Les termes diagonaux de **G** ont une décroissance exponentielle vis-à-vis de l'échelle j, tels que :

$$[\mathbf{G}_{\boldsymbol{\theta}}]_{jk} = 2^{-j\eta}\gamma^2$$

▷ Permet d'assurer que les effets fixes et aléatoires seront dans le même espace fonctionnel

Théorème - Abramovich et al (1998)

Supposons que $\mu(t) \in B^s_{\rho_1,\rho_2}$ et $\mathbb{V}(\theta^i_{jk}) = \gamma^2 2^{-j\eta}$ alors

$$U_i(t) \in B^s_{\rho_1,\rho_2} \Leftrightarrow \begin{cases} \eta = 2s+1 \text{ pour } 1 \leq \rho_1 \leq \infty \text{ et } \rho_2 = \infty \\ \eta > 2s+1 \text{ sinon} \end{cases}$$

Hypothèses considérées

Structure de la matrice

G est supposée diagonale

▷ justifiée par la propriété décorrélante des ondelettes (Frazier et al, 1992)

Décroissance des paramètres de variances

Les termes diagonaux de **G** ont une décroissance exponentielle vis-à-vis de l'échelle j, tels que :

$$[\mathbf{G}_{\boldsymbol{\theta}}]_{jk} = 2^{-j\eta}\gamma^2_{\ell,jk}$$

▷ Permet d'assurer que les effets fixes et aléatoires seront dans le même espace fonctionnel

Théorème - Abramovich et al (1998)

Supposons que $\mu(t) \in B^s_{\rho_1,\rho_2}$ et $\mathbb{V}(\theta^i_{jk}) = \gamma^2_{\ell,jk} 2^{-j\eta}$ alors

$$U_i(t) \in B^s_{\rho_1,\rho_2} \Leftrightarrow \begin{cases} \eta = 2s+1 \text{ pour } 1 \leq p \leq \infty \text{ et } \rho_2 = \infty \\ \eta > 2s+1 \text{ sinon} \end{cases}$$

Plan de la présentation

Plan de la présentation

Objectif et procédure globale

Objectif principal

Retrouver les labels individuels $\left(\zeta_{i\ell}\right)_{i=1,\ldots,N}^{\ell=1,\ldots,L}$

Procédure globale en deux étapes

- Étape de réduction de dimension
  - Nécessaire due à la grande dimension des données considérées
  - Basée sur les propriétés de parcimonie des ondelettes
- Étape de classification non supervisée
  - Estimation des paramètres du modèle par maximum de vraisemblance

Cette partie fait l'objet d'une publication et d'un package `curvclust` disponible sur le CRAN

Giacofci,M., Lambert-Lacroix,S., Marot,G., Picard,F. (2013). Wavelet-based clustering for mixed-effects functional models in high dimension. *Biometrics 69*(1), 31-40.

Détails de la procédure

Étape de réduction de dimension

Inspiré d'une stratégie proposée par Antoniadis et al (2008) consistant en :

- Un seuillage individuel dur comme défini par Donoho et Johnstone (1994)
- L'union des coefficients sélectionnés

Étape de classification non supervisée

Estimation des paramètres du modèle par maximum de vraisemblance :

- Réalisée au moyen de l'algorithme EM (Dempster et al, 1977)
- Avec deux types de variables non observées
  - Labels individuels $\boldsymbol{\zeta}$
  - Effets aléatoires $(\boldsymbol{\nu}, \boldsymbol{\theta})^T$
- Les labels individuels sont finalement déterminés par une régle de *Maximum A Posteriori* (MAP)

Choix du nombre de groupes

Basé sur l'utilisation d'un critère BIC

Plan de la présentation

Création de jeux de données synthétiques

Objectifs

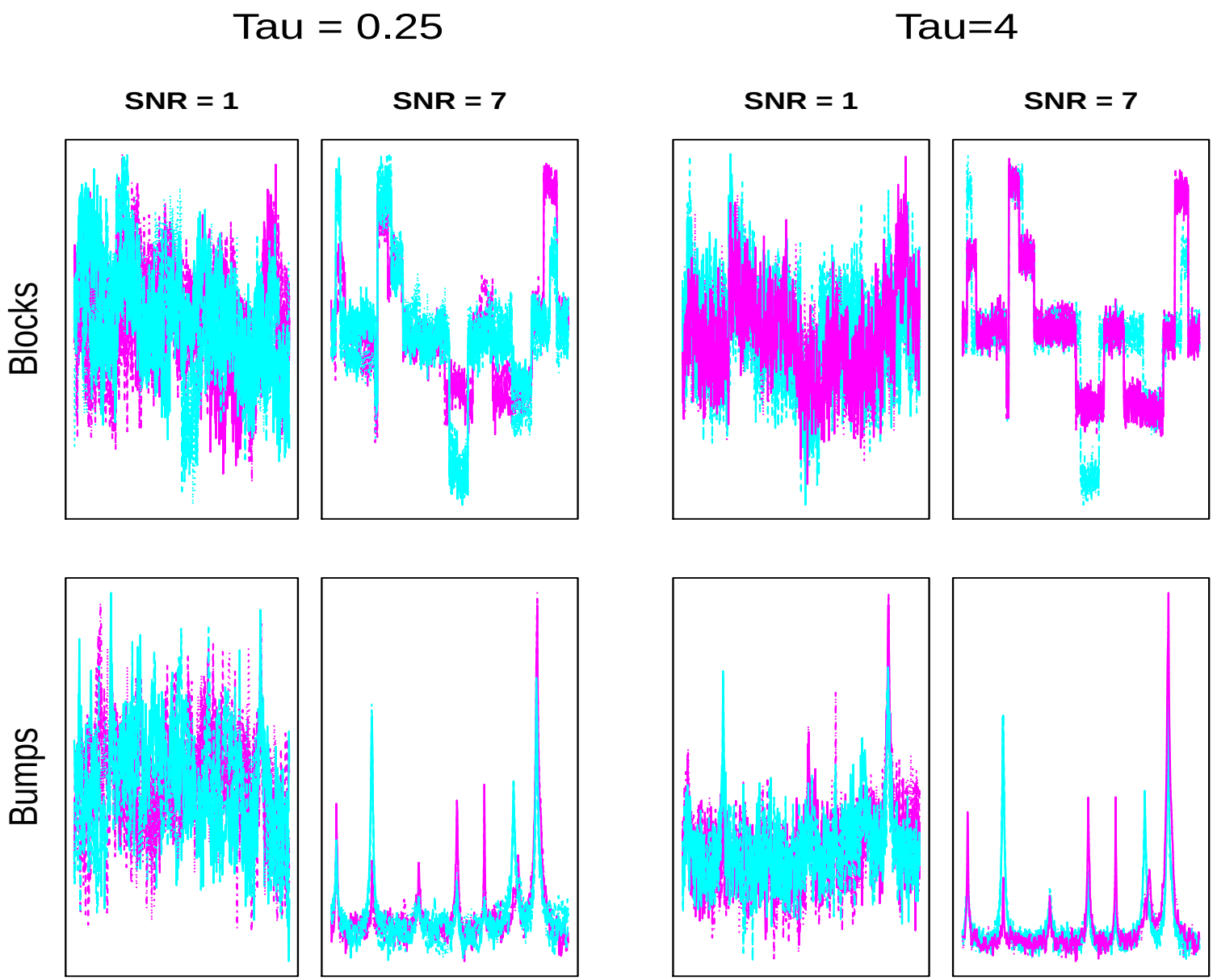
- Création systématique de jeux de données à partir de paramètres canoniques
- Meilleure exploration de l'univers des simulations
- Comparaison équitable des différentes procédures

Principe

- Simulation des jeux de données dans le domaine des ondelettes
- Définition d'un SNR et d'un τ_U fonctionnel pour le contrôle des niveaux de variabilité :

$$\begin{cases} \mathrm{SNR}^2 = \dfrac{1}{M\sigma_E^2} \displaystyle\sum_{\ell=1}^{L} \pi_\ell \left(\sum_{k=0}^{2^{j_0}-1} \alpha_{j_0 k \ell}^2 + \sum_{j \geq j_0} \sum_{k=0}^{2^j-1} \beta_{jk\ell}^2 \right) \\ \tau_U = \sigma_E^2 / \left(\gamma_\nu^2 + \dfrac{\gamma_\theta^2}{1 - 2^{(1-\eta)}} \right) \end{cases}$$

Exemple de jeux de données simulés (2 groupes)



Cadre de simulation

Paramètres

- $N = 50$, $M = 512$, $L = 2$, SNR $\in \{1, 3, 5, 7\}$ et $\tau_U \in \{1/4, 1, 4\}$

Procédures comparées

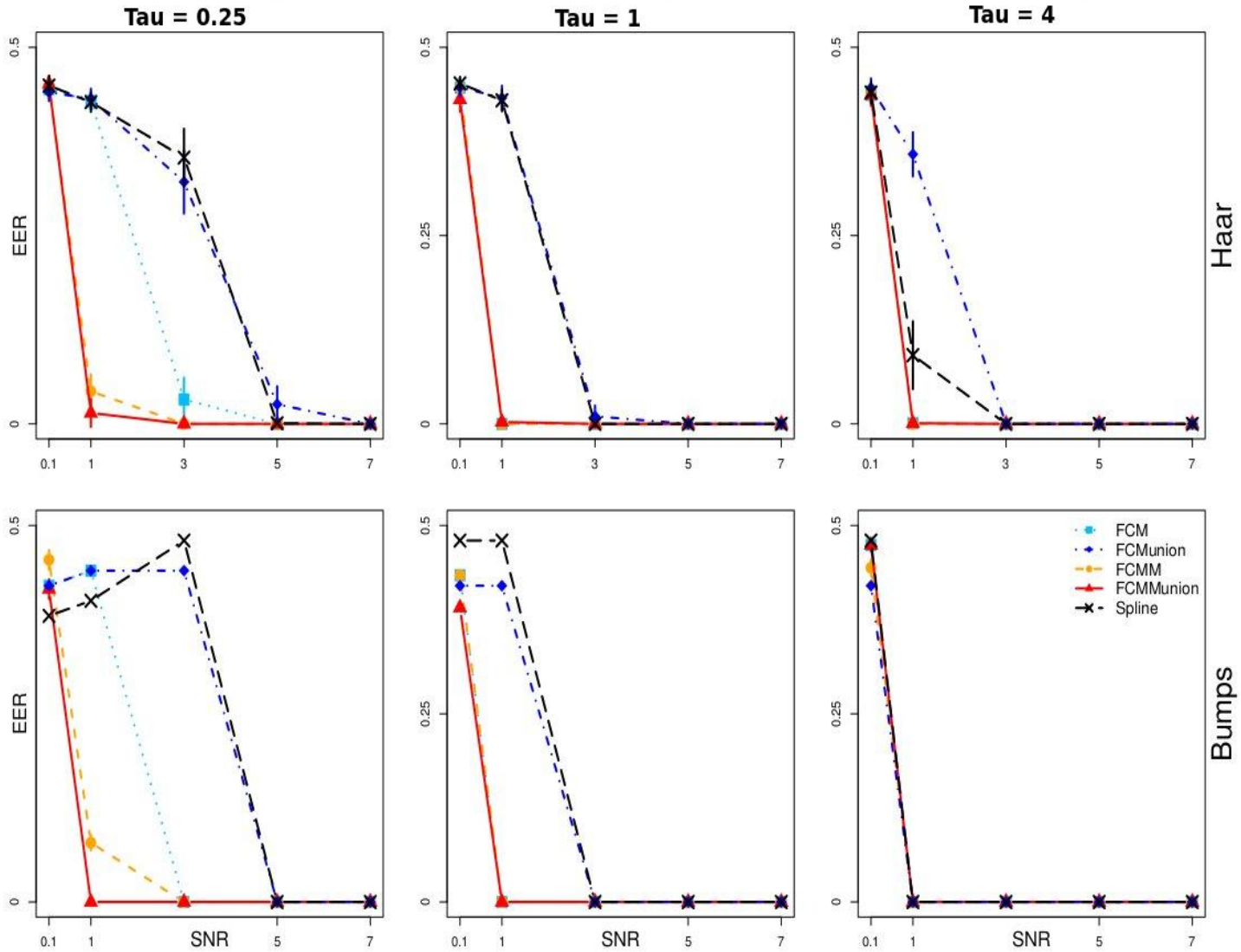
Notre procédure [FCMM] est comparée à 3 autres procédures :

- FCMM sans effets aléatoires [FCM]
  - Bénéfice de la prise en compte des effets aléatoires ?
- FCMM sans réduction de dimension [FCMMnone]
  - Quel est l'effet de la réduction de dimension ?
- Procédure basée sur les splines (James & Sugar, 2003) **[Spline]**
  - Quels sont les performances offertes par les splines ?

Critère de comparaison

$$EER = \frac{1}{n} \sum_{i=1}^{n} \sum_{\ell}^{L} \mathbb{I}_{\{\widehat{\zeta}_{i\ell} \neq \zeta_{i\ell}\}}$$

Résultats


Tau = 0.25
Tau = 1
Tau = 4
EER
SNR
Haar
Bumps
FCM
FCMunion
FCMM
FCMMunion
Spline

Résultats - données de spectrométrie de masse

- Nécessité d'une étape de pré-traitement (normalisation des données, alignement des pics), couteuse d'un point de vue numérique (Antoniadis et. al, 2007)
- Restriction à $M = 8192$ points de discrétisation

Méthode	FCM	FCMM	FCMM.gr	FCMM.jk
EER - alignement global	38%	24%	24%	23%
EER - alignement par groupe	20%	21%	22%	0.4%

Interprétations

- Amélioration des résultats par la prise en compte des effets aléatoires
- Effet important de l'étape d'alignement des données
- Résultat encourageant : un seul individu mal classé pour des variances dépendant de la position
  ▷ Suggère une configuration parcimonieuse de la variabilité individuelle ?

Résultats - données de microarray CGH

Contexte

- Approches existantes : classification sur la base de résultats de segmentation (van Wieringen & van de Wiel (2008))
- Variabilité inter-individuelle peu étudiée sur ce type de données
- Sur ce jeu de données particulier : existence de plusieurs classifications (reliées aux données de survie)

Principales conclusions

- Notre procédure trouve plus de groupes que l'étude originale
- Un groupe commun avec l'étude originale (associé à la meilleure survie)
- L'estimation du SNR et τ_U a posteriori montre des variabilités très importantes ($\approx 10^{-4}$)

  ▷ Nécessité de disposer d'un nombre beaucoup plus important d'individus dans une optique de prédiction

Plan de la présentation

Problématiques

Cadre d'estimation : $L = 1$

Une fois les groupes formés, on s'intéresse à la problématique de l'estimation au sein d'un groupe homogène

Estimation dans les modèles mixtes

- Estimation des effets fixes : objectif premier du statisticien car traduit le comportement moyen au sein du groupe
- Estimation des effets aléatoires : pour permettre une meilleure compréhension des sources de variabilité et une meilleure estimation des effets fixes

Contexte fonctionnel

La modélisation adoptée sur les données nous conduit à attendre :

- Une représentation parcimonieuse de l'effet fixe
- Une représentation parcimonieuse des effets aléatoires fonctionnels
  $\Rightarrow$ La sélection des effets aléatoires est réalisée par l'intermédiaire de la sélection des variances associées γ^2_{jk}

Plan de la présentation

Estimation de l'effet fixe

Cadre fonctionnel

L'objectif est de proposer un estimateur $\widehat{\mu}(t)$ de l'effet fixe fonctionnel $\mu(t)$

- C'est un des intérêts premiers du praticien car μ représente le profil moyen associé à un groupe homogène d'individus

Réinterprétation du modèle

$$\begin{bmatrix} \mathbf{c}_i \\ \mathbf{d}_i \end{bmatrix} = \begin{bmatrix} \boldsymbol{\alpha} \\ \boldsymbol{\beta} \end{bmatrix} + \widetilde{\boldsymbol{\varepsilon}}_i \qquad \widetilde{\varepsilon}_{i,jk} \sim \mathcal{N}(0, \gamma_{jk}^2 2^{-j\eta} + \sigma_\epsilon^2)$$

$\widetilde{\boldsymbol{\varepsilon}}_i$: effet individuel + bruit de mesure

- Problème de régression non-paramétrique dans un contexte hétéroscédastique en présence de répétitions individuelles

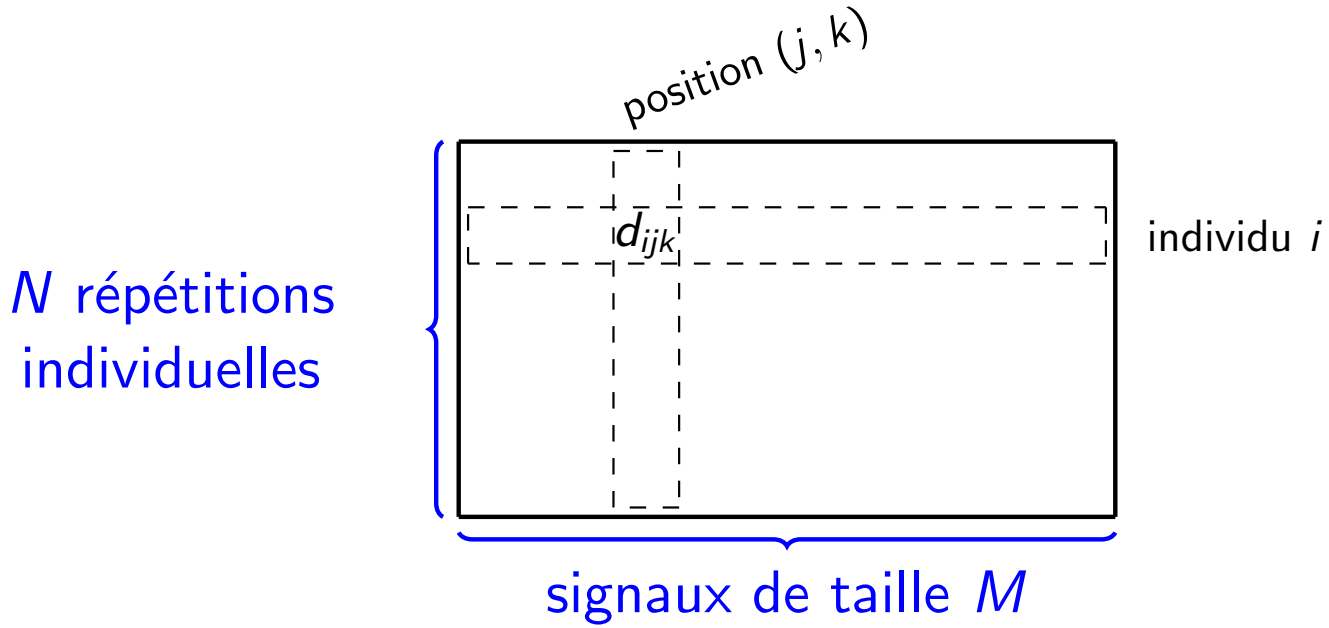
Méthodes usuelles dédiées : techniques de seuillage

- Seuillages doux et durs (Donoho & Johnstone, 1994)
- Seuillage SCAD (Antoniadis & Fan, 2001)

Seuillage hétéroscédastique

Idée générale

- Étendre le seuil universel à un cadre hétéroscédastique $\lambda = \widehat{\sigma}_{jk}\sqrt{2\log M}$ (seuil dépendant de la position)
- Les paramètres σ_{jk} sont estimés de manière empirique grâce à la présence de répétitions individuelles
- Propriétés de reconstruction de l'estimateur fonctionnel $\widehat{\mu}$?



Propriétés de reconstruction de l'estimateur $\widehat{\mu}$?

Théorème

- $\mu \in B^s_{pq}$, $s' = s - \frac{1}{p} + \frac{1}{2} > 0$, $p \geq 1$, $q \geq 1$ et $s > 1/p$
- $\widehat{\mu}$ estimateur résultant du seuillage hétéroscédastique
- $\left(\sigma^2_{jk}\right)_{(j,k)}$ bornées dont on dispose d'estimateurs $\sqrt{N}$-consistants

$$\text{Alors : } \mathbb{E}\left(\|\widehat{\mu} - \mu\|^2_{L^2}\right) = \underbrace{\mathcal{O}\left[\frac{M}{N \log M}\right]}_{\mathbf{T_1}} + \underbrace{\mathcal{O}\left[\left(\frac{\log M}{MN}\right)^{\frac{2s}{2s+1}}\right]}_{\mathbf{T_2}} + \underbrace{\mathcal{O}\left[\left(\frac{\log M}{M}\right)^{2s'}\right]}_{\mathbf{T_3}}$$

Convergence

- La vitesse de convergence vers le vrai paramètre est contrôlée en imposant des conditions sur le ratio M/N,
  - Pour $(M/\log M)^{\frac{4s+1}{2s+1}} = \mathcal{O}(N)$: Convergence en $\left[\frac{\log M}{M}\right]^{\frac{2s}{2s+1}}$

$\Rightarrow$ On retrouve la vitesse near-minimax usuelle

Limites de l'approche marginale

Vitesse de convergence

- Idéalement, le taux de convergence de l'estimateur $\widehat{\mu}$ devrait être comparé à la vitesse minimax obtenue dans notre cadre

  $\Rightarrow$ Intuitivement, cette vitesse devrait se comporter comme

$$\mathcal{O}\left\{\max\left[\left(\frac{\log M}{MN}\right)^{\frac{2s}{2s+1}}, \left(\frac{\log M}{M}\right)^{2s'}\right]\right\}$$

Sélection des variances des effets aléatoires

- Difficultés à proposer une sélection satisfaisante des paramètres de variances dans cette approche basée sur une procédure non itérative

  $\Rightarrow$ Peut conduire à une mauvaise spécification des effets aléatoires

Plan de la présentation

Approche par vraisemblance pénalisée

Idée générale

- Tirer parti de l'équivalence entre seuillage et régression pénalisée dans le cadre orthogonal

Sélection des effets fixes [Antoniadis et Fan (01) // Fan et Peng (04)]

- Critère de vraisemblance pénalisée pour l'estimation/sélection des coefficients d'effets fixes

$$\ell(\boldsymbol{\beta}) = -\log \mathcal{L}(\mathbf{d}, \boldsymbol{\theta}; \mathbf{G}, \boldsymbol{\beta}, \sigma_\epsilon^2) + \text{pen}_{\text{SCAD}}(\boldsymbol{\beta}, \lambda_1)$$

avec λ_1 paramètre de régularisation.

Approche par vraisemblance pénalisée

Idée générale

- Tirer parti de l'équivalence entre seuillage et régression pénalisée dans le cadre orthogonal

Sélection des effets fixes

$$\ell(\beta) = -\log \mathcal{L}(\mathbf{d}, \theta; \mathbf{G}, \beta, \sigma_\epsilon^2) + \text{pen}_{\text{SCAD}}(\beta, \lambda_1)$$

Double sélection des effets fixes et aléatoires

- Extension du critère pour l'estimation/sélection des coefficients d'effets fixes et des variances des effets aléatoires

$$\ell(\beta, \gamma) = -\log \mathcal{L}(\mathbf{d}, \theta; \mathbf{G}, \beta, \sigma_\epsilon^2) + \text{pen}_{\text{SCAD}}(\beta, \lambda_1) + \text{pen}_{\text{SCAD}}(\gamma, \lambda_2)$$

avec λ_1 et λ_2 paramètres de régularisation.

Propriétés de sélection

Notations

Soit $\Upsilon = (\beta^T, \gamma^T)^T = (\Upsilon_1^T, \Upsilon_2^T)^T$ avec Υ_1 ensemble des paramètres non nuls et Υ_2 ensemble des paramètres nuls

Propriété d'oracle

- L'optimisation du critère de vraisemblance pénalisée conduit à un estimateur $\widehat{\Upsilon}$ possédant la propriété d'oracle, à savoir :
  - $\widehat{\Upsilon_2} = 0$ presque sûrement (le bon modèle est atteint presque sûrement)
  - $\widehat{\Upsilon_1}$ est asymptotiquement normal
- Cadre de double asymptotique où $N \to \infty$, $M \to \infty$ et $\frac{M^5}{N} \to 0$

Limites

- Ratio $M^5/N \to 0$ peu réaliste dans un cadre fonctionnel
  - ▷ Peu de réponses dans la littérature actuelle (Kim et al., 2008 ; Fan et Li, 2012)
- Propriétés de reconstruction de l'effet fixe fonctionnel dans ce cadre ?

Algorithme d'optimisation

Optimisation du critère de vraisemblance pénalisée

- Nécessite une reparamétrisation des effets aléatoires θ_i (Chen et Dunson, 2003) pour des contraintes techniques
- Basée sur une variante ECM de l'algorithme EM : Maximisations conditionnelles au cours de l'étape M

Mise à jour de l'effet fixe

- ▷ Se ramène à un seuillage SCAD des données, corrigées des prédictions des effets aléatoires

Mise à jour des variances des effets aléatoires

- ▷ Se ramène à un seuillage SCAD des données centrées et "normalisées" par les prédictions des effets aléatoires

Plan de la présentation

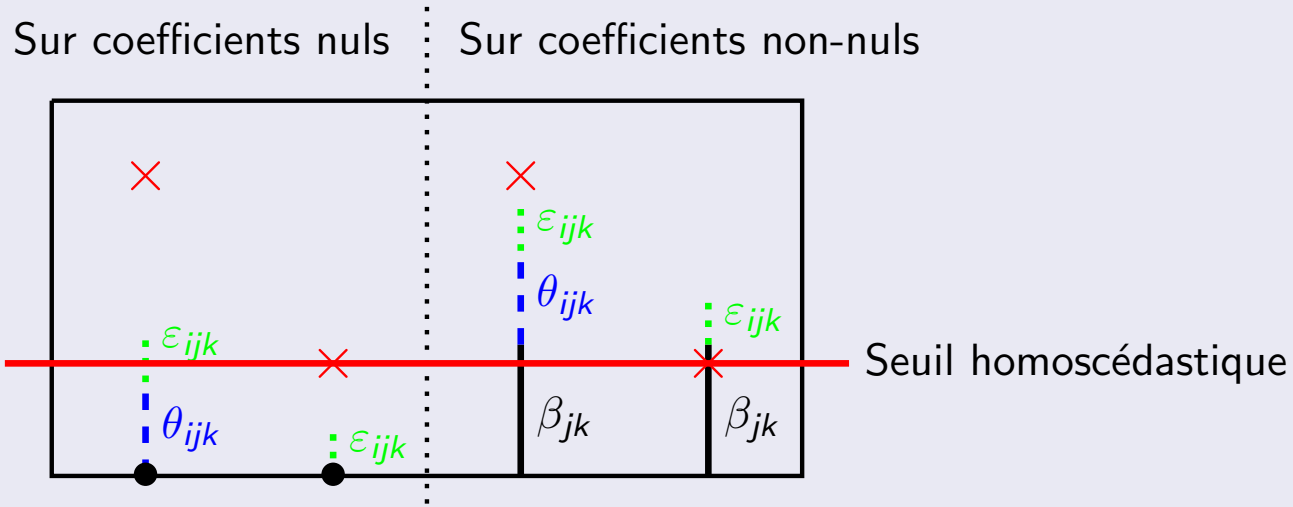
Configurations étudiées

Création de jeux de données synthétiques

- Selon le même principe que pour la partie classification
- Possédant une représentation parcimonieuse :
  - de l'effet fixe fonctionnel
  - des effets aléatoires fonctionnels par le biais des variances associées

Configurations étudiées

- A - Effets aléatoires situés sur les coefficients nuls de l'effet fixe
- B - Effets aléatoires situés sur les coefficients non-nuls de l'effet fixe



Approches comparées

Paramètres

- $N = 100$, $M = 512$, SNR $\in \{1, 3, 5, 7\}$ et $\tau_U \in \{0.1, 1/4, 1, 4\}$

Procédures

- Seuillage homoscédastique SCAD usuel [Homoscédastique]
- Seuillage hétéroscédastique avec estimation empirique des variances [Type moment]
- Estimation par vraisemblance pénalisée, avec une étape de relaxation (Meinhausen, 2007) [Minimax.relaxe]

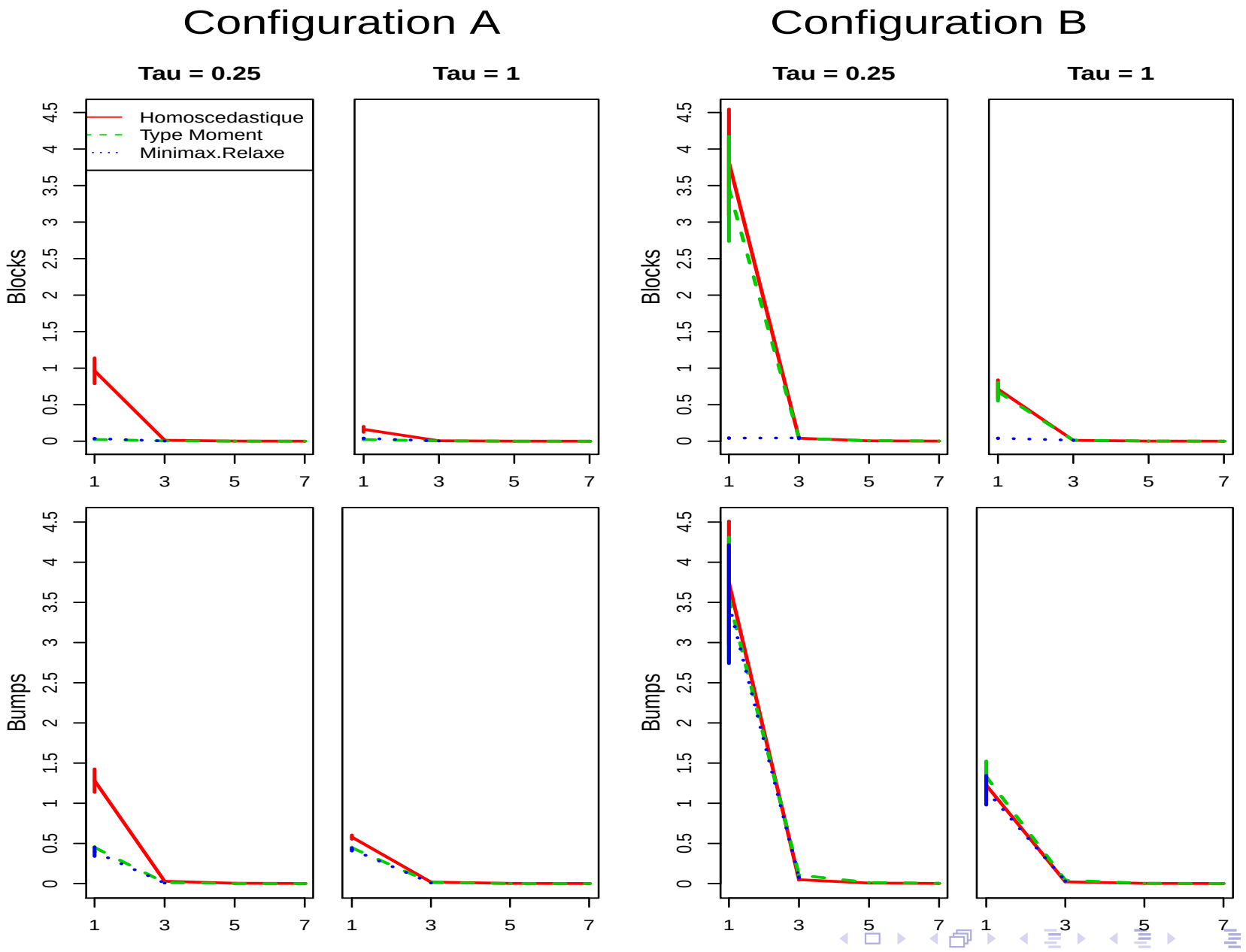
Critères de comparaison

- Reconstruction : Écart Quadratique Moyen

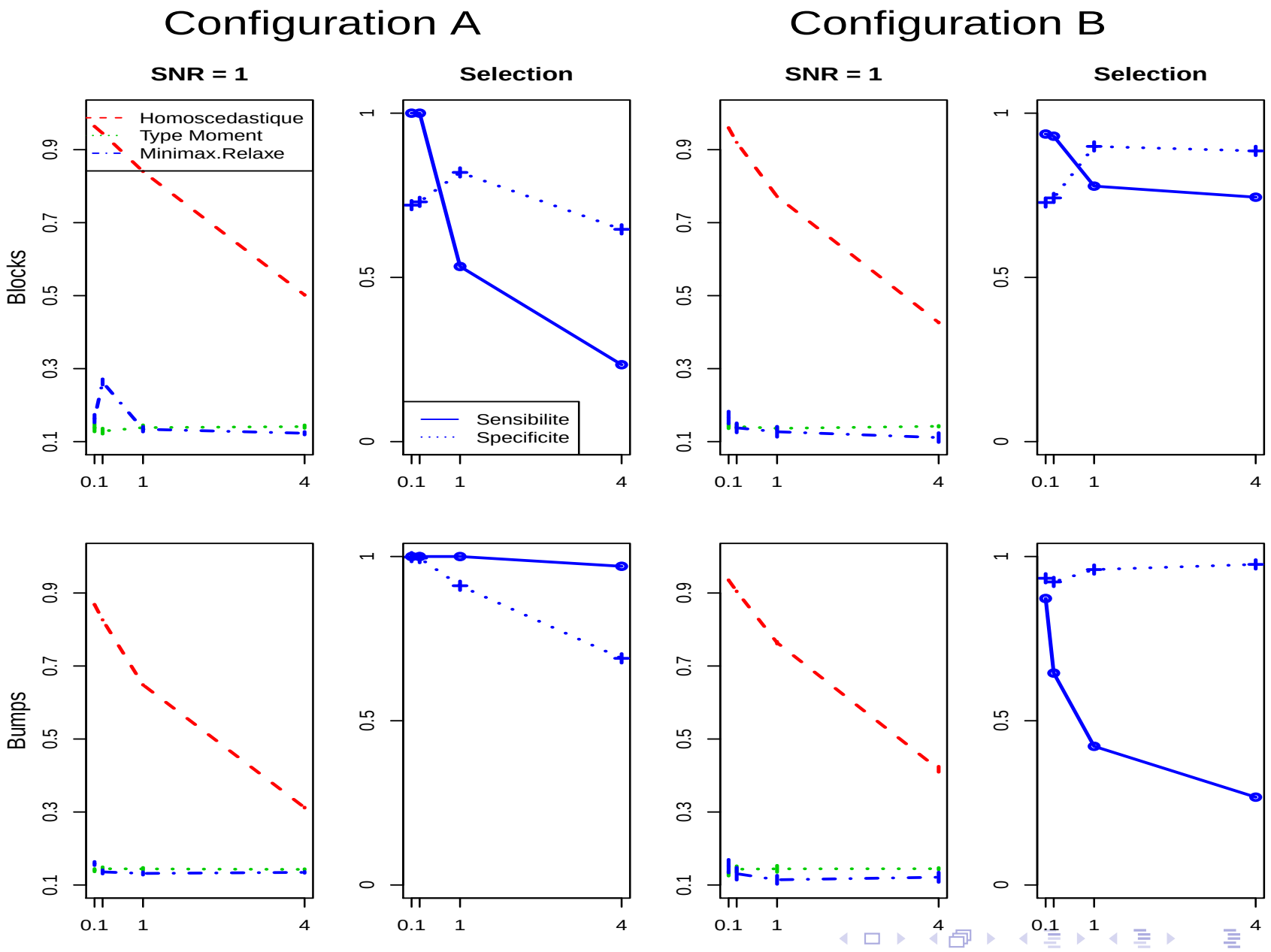
$$\text{EQM}_{\boldsymbol{\mu}} = \sqrt{\frac{\|\widehat{\boldsymbol{\mu}} - \boldsymbol{\mu}_0\|_2^2}{\|\boldsymbol{\mu}_0\|_2^2}} \quad \text{et} \quad \text{EQM}_{\boldsymbol{\sigma}} = \sqrt{\frac{\|\widehat{\boldsymbol{\sigma}} - \boldsymbol{\sigma}_0\|_2^2}{\|\boldsymbol{\sigma}_0\|_2^2}},$$

- Sélection : critères de sensibilité et spécificité

Résultats - EQM associé à l'effet fixe fonctionnel



Résultats - Sélection des variances des effets aléatoires



Résultats - Temps d'exécution pour $\tau_U = 0.1$

			Configuration A	Configuration B
Blocks	0.1	Sélection	168.9 (9.101)	192.7 (14.64)
		Type Moment	0.159 (0.019)	0.135 (0.039)
	1	Sélection	231.9 (14.89)	349.1 (243.2)
		Type Moment	0.138 (0.027)	0.199 (0.041)
Bumps	0.1	Sélection	277.4 (25.02)	475.9 (170.1)
		Type Moment	0.157 (0.031)	0.171 (0.032)
	1	Sélection	561.2 (86.54)	476.1 (73.09)
		Type Moment	0.141 (0.021)	0.179 (0.032)

Les temps d'exécution des différentes procédures sont donnés en secondes

Plan de la présentation

Conclusions et perspectives

Conclusions

Nous nous sommes concentrés sur l'étude des modèles mixtes fonctionnels du point de vue :

- de la classification non supervisée
- de l'estimation et de la sélection de variables

Perspectives

- Partie classification non supervisée
  - Seuillage dans les modèles de mélanges Gaussiens
  - Alignement et classification simultanée dans le cas de la spectrométrie de masse
- Partie estimation/sélection
  - Calcul de la vitesse minimax en présence de répétitions individuelles
  - Extension des résultats asymptotiques au cas $M > N$
  - Propriétés de reconstruction de l'estimateur de l'effet fixe fonctionnel
  - Applications à des données réelles

Merci de votre attention !