Analyse de données de spectrométrie de masse

A. Antoniadis, J. Bigot, S. Lambert-Lacroix, F. Letué

Laboratoire Jean Kuntzmann, Grenoble

octobre 2009

Plan de l'exposé

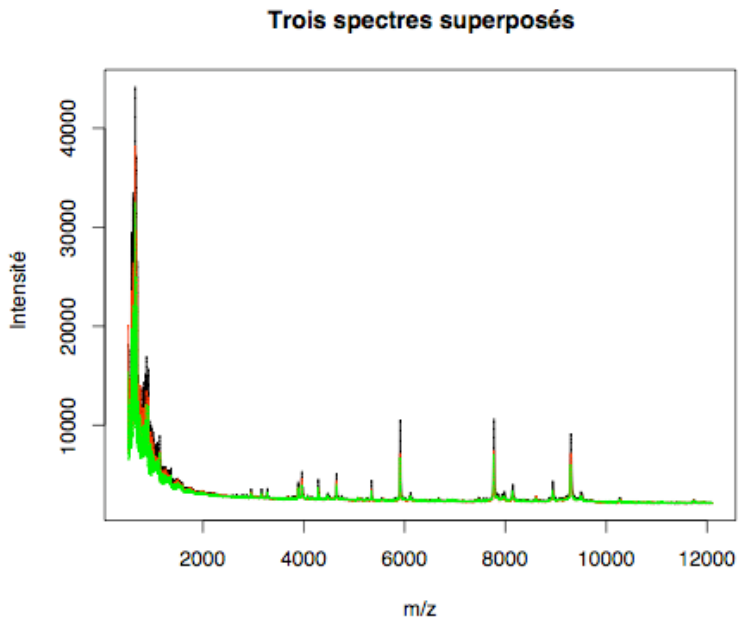
- Acquisition de données SELDI-TOF (ou MALDI-TOF)
- Pré-traitements et premiers problèmes statistiques
  - Débruitage
  - Suppression du bruit de fond
  - Normalisation, alignement et quantification des spectres
- Inférence statistique
  - Analyse de la variance fonctionnelle
  - Modèles à effets aléatoires
  - Extraction des biomarqueurs
  - Apprentissage et techniques de classification.

Acquisition des données

- Des protéines purifiées ou partiellement purifiées extraites d'un échantillon biologique (sérum par exemple) sont mélangées à un acide qui permet au mélange de se cristalliser lorsqu'il sèche.
- Le mélange est alors appliqué sur une lame d'acier inoxydable pré-traitée qui retient à sa surface de manière préférentielle des classes particulières de protéines selon leurs propriétés physio-chimiques.
- L'échantillon est placé dans un tube à vide et le cristal est soumis à un rayonnement laser, entraînant une ionisation des protéines et un détachement (en phase gazeuse).
- Les molécules de protéines ionisées en phase gazeuse sont alors soumises à un bref champ électrique qui produit une accélération des ions dans le tube et un détecteur au bout du tube enregistre le temps de vol.

Exemple de spectres

Spectre typique : enregistrement séquentiel du nombre d'ions qui arrivent sur un détecteur avec les valeurs de leurs m/z. Les pics correspondent aux diverses protéines.



Avantages

- la préparation des échantillons biologiques est rapide ;
- la production des ions révèle des molécules de protéines avec très peu de fragmentation ;
- le champ de détection des masses sur charge s'étend sur des intervalles pouvant aller de 2000 à 100000 daltons, avec une précision de 1/10000.

Malgré ces avantages, il persiste néanmoins plusieurs problèmes dans la quantification des protéines dans les échantillons biologiques, inhérents au processus d'ionisation même avec pour conséquence une grande variabilité dans les intensités enregistrées, même pour des données répétées.

Les premiers problèmes statistiques

- **Fléau de la dimension :** un spectre typique contient plus de 10000 mesures d'intensité et on ne dispose relativement que de peu de spectres (individus) ;
- **Hétéroscedastisticité :** les mesures d'un spectre présentent une dispersion (ou échelle) variable en fonction de l'intensité enregistrée, rendant des comparaisons difficiles ;
- **Alignement :** lorsque l'expérience est répétée ou porte sur des échantillons biologiques de même nature, il arrive que les pics enregistrés ne soient pas alignés. Tout processus de moyennisation est alors rendu impossible sans alignement préalable.
- **Détection de pics :** Identifier les pics importants pour des études de différentiation.

Modélisation statistique d'un spectre

L'idée est de considérer que chaque spectre est constitué de la superposition de trois composantes : le signal des pics, un bruit de fond lisse et un bruit aléatoire additif de mesure.

$$Y(m/z) = \overbrace{B(m/z)}^{\text{bdf}} + \underbrace{N}_{\text{facteur de normalisation}} \overbrace{S(m/z)}^{\text{signal des pics}} + \underbrace{\epsilon(m/z)}_{\text{bruit}},$$

$$\epsilon_i(m/z) \sim N(0, \sigma^2(m/z))$$

But : Isoler le signal d'intérêt S

Débruitage

- On commence par débruiter (supprimer le bruit) par des méthodes non paramétriques. Pour cela on utilise un débruitage par ondelettes puisque les ondelettes sont des fonctions de base permettant de représenter de façon parcimonieuse des fonctions composées de pics. On utilise la transformée en ondelettes invariante par translation (TIWT) de sorte que le résultat ne dépende pas de l'endroit où on commence à traiter le signal.
- Pourquoi cela marche-il ? Le signal est caractérisé par un petit nombre de coefficients alors que le bruit est réparti sur tous les coefficients .
- Le seuillage enlève le bruit sans trop affecter le signal. Les ondelettes marchent beaucoup mieux que les méthodes à noyau ou les splines, qui ont tendance à atténuer les intensités des pics lors du débruitage.

Correction du bruit de fond

Bruit de fond : signal lisse, attribuable à la surcharge du détecteur. Il est estimé de manière plus stable après débruitage. Plusieurs procédures existent dans la littérature

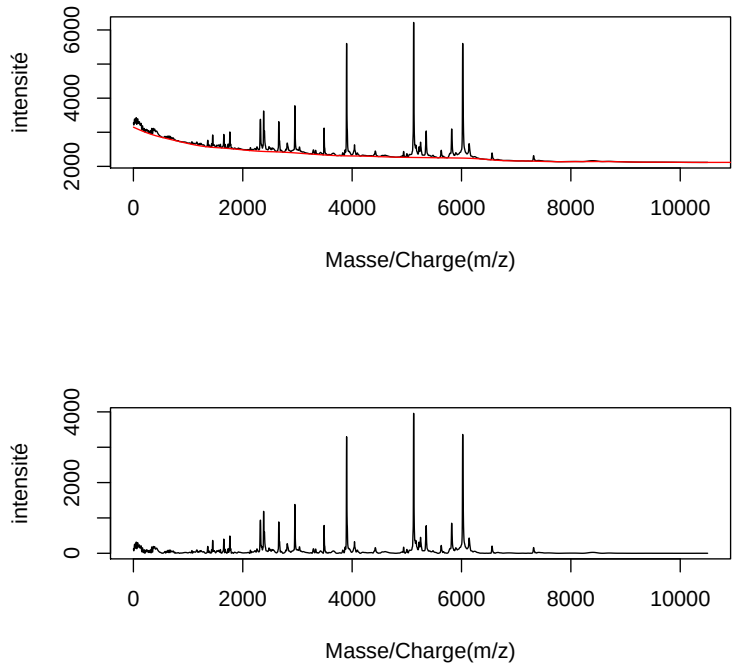
- **Filtrage digital**. Habituellement elles produisent des artefacts et des distorsions des signaux sous-jacents.
- **Rejet automatique des pics**. Ces algorithmes ajustent certaines fonctions (polynômes, splines) sur des régions du spectre ne contenant que du bruit de fond et pas de pics. Ils sont médiocres pour des spectres dont le bruit de fond varie fortement.
- **Décomposition en ondelettes**. Le spectre est décomposé dans une base d'ondelettes orthogonales et le bruit de fond est estimé par la projection sur l'espace d'approximation de résolution la plus grossière.

Estimation du bruit de fond

Idée pour estimer le bruit de fond : Le bruit de fond est concentré sur les points bas du spectre. On l'estime donc en utilisant une régression quantile pénalisée (version pondérée asymétrique de la valeur absolue des résidus). Le bruit de fond (supposé lisse) est décomposé dans une base de fonctions B-splines sur des noeuds équidistants. Afin de représenter le bruit de façon parcimonieuse nous préférons utiliser une version pénalisée de type Lasso.

Estimation du bruit de fond

Un exemple d'estimation du bruit de fond. En haut le spectre débruité ; en bas le spectre corrigé du bruit de fond.

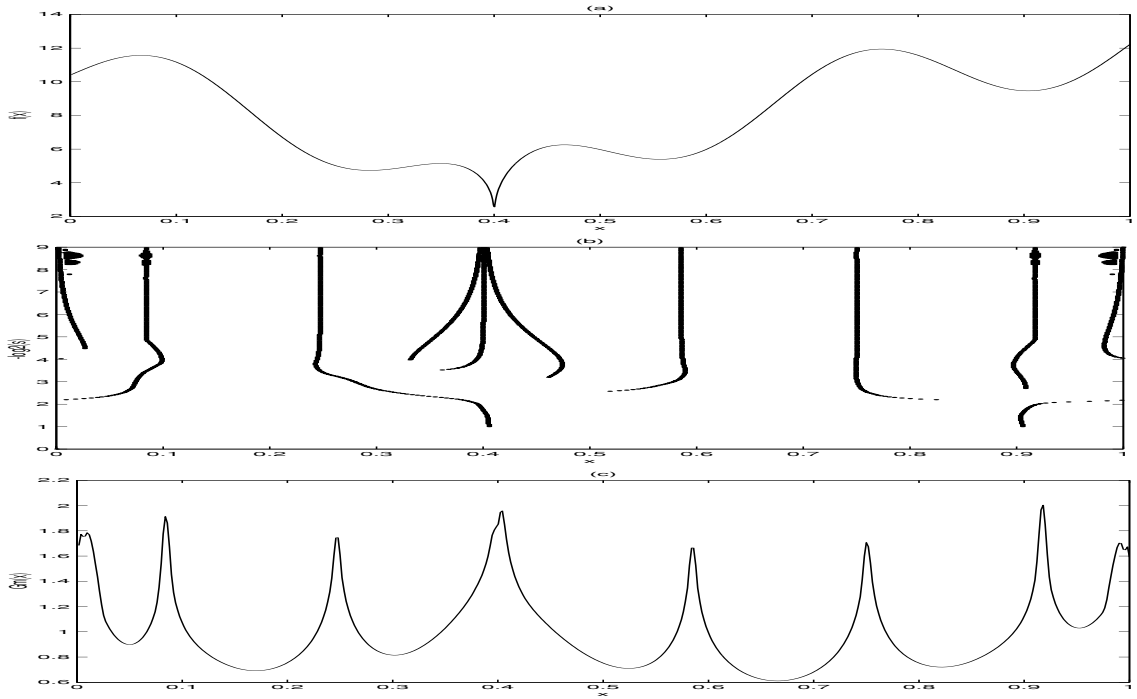


Alignement de deux spectres par ondelettes

- Définition des landmarks (ensemble de points caractérisant la forme du signal : maxima, minima ...) ;
- Extraction des landmarks d'un spectre à partir de son observation ;
- Déterminer les landmarks qui sont communs (et qui doivent se correspondre) ;
- Calcul des transformations qui alignent ces landmarks.
- Déformation des signaux à l'aide de ces transformations.

Exemple

Les landmarks d'un signal 1D peuvent être caractérisés par la propagation des maxima et des zero-crossings de sa transformée continue en ondelettes.

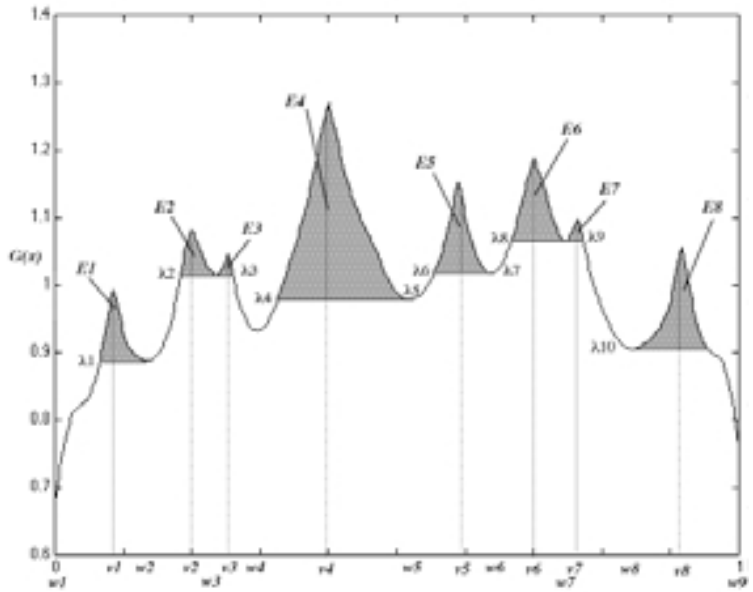


Intensité structurelle

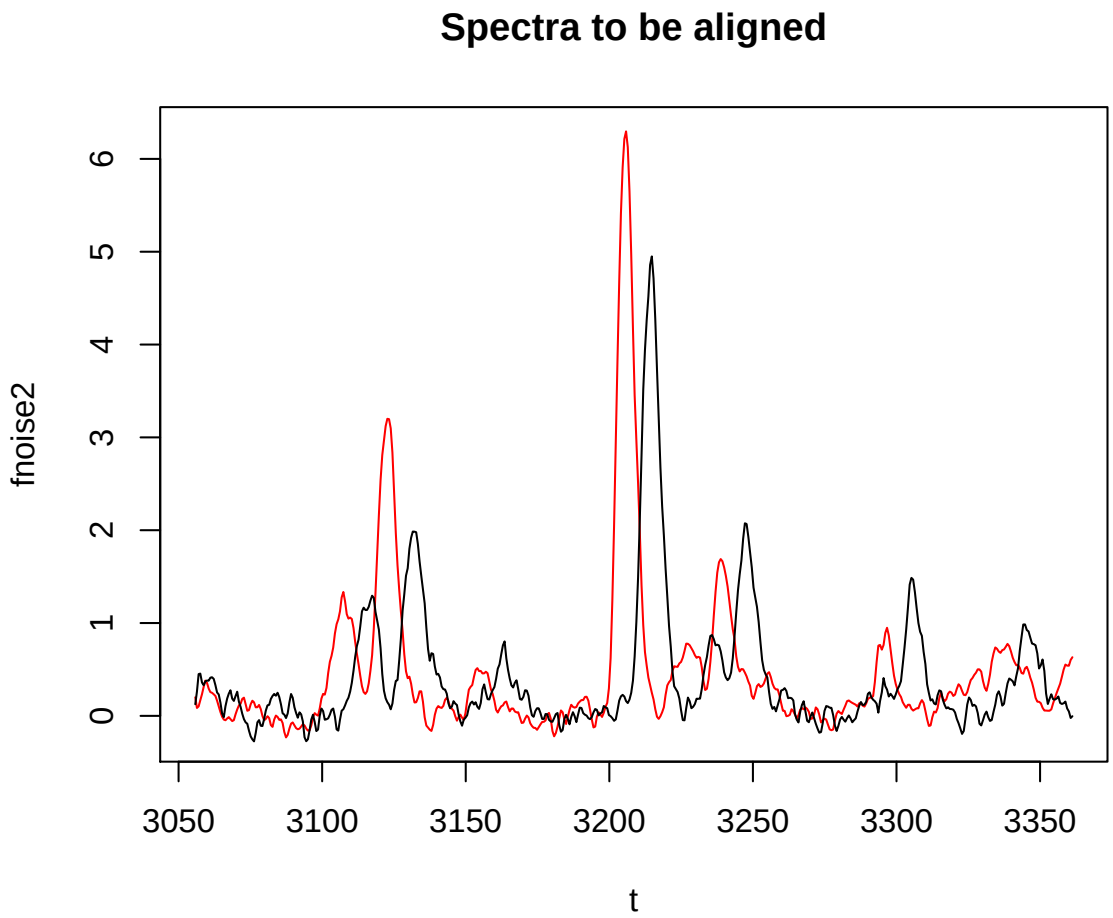
- Pour identifier les landmarks nous avons utilisé l'approche non paramétrique proposée dans sa thèse par Jéremie Bigot (2003).
- La notion d'intensité structurelle permet d'identifier les limites des lignes lorsqu'elle se propagent aux fines échelles. Elle correspond en quelque sorte à la "densité" des zéros et des extrema (en module) d'une représentation en ondelette le long de plusieurs échelles.
- Les modes des intensités structurelles sont localisés au voisinage des landmarks du signal correspondant et l'intensité structurelle est également un outil efficace pour éliminer les erreurs d'estimation dues aux fluctuations du bruit.

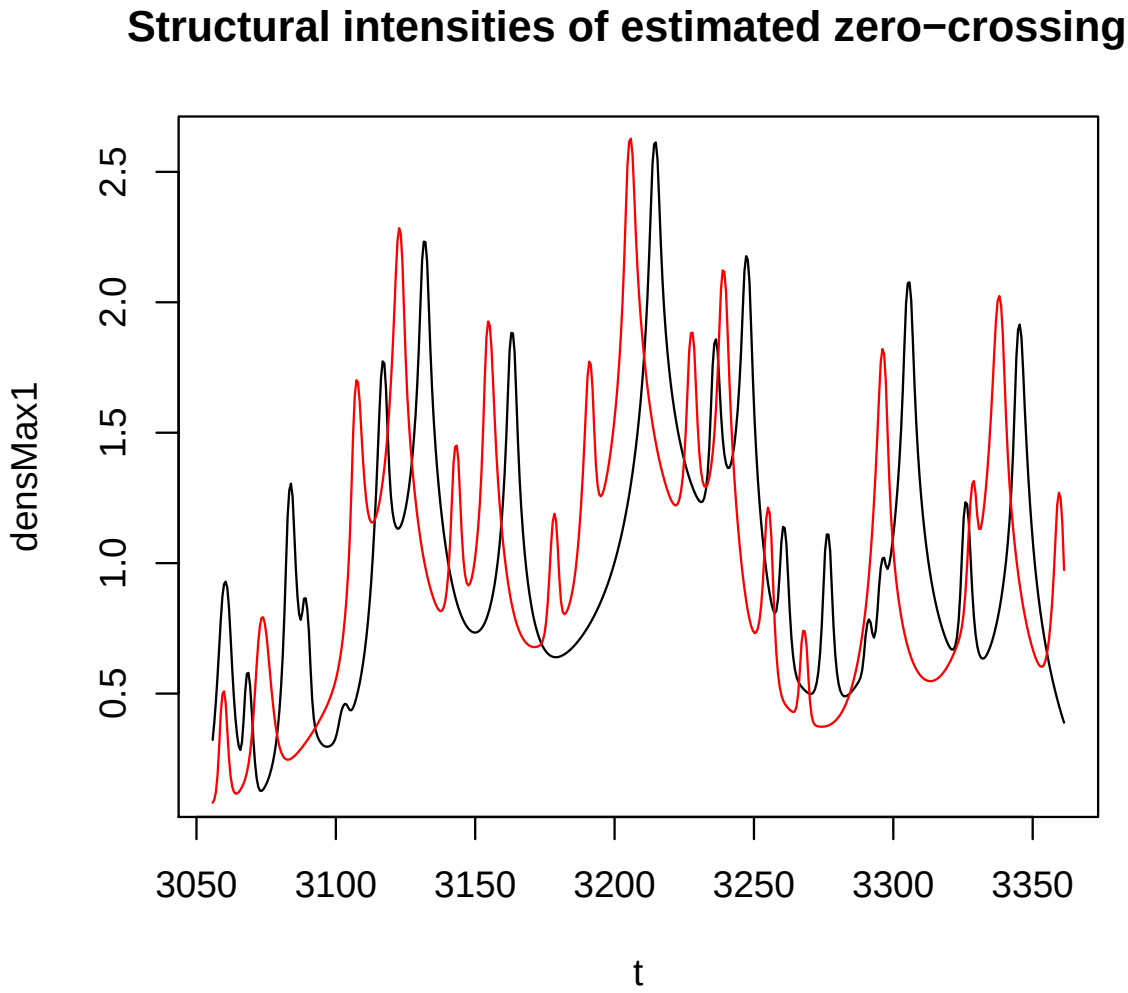
Excessive mass

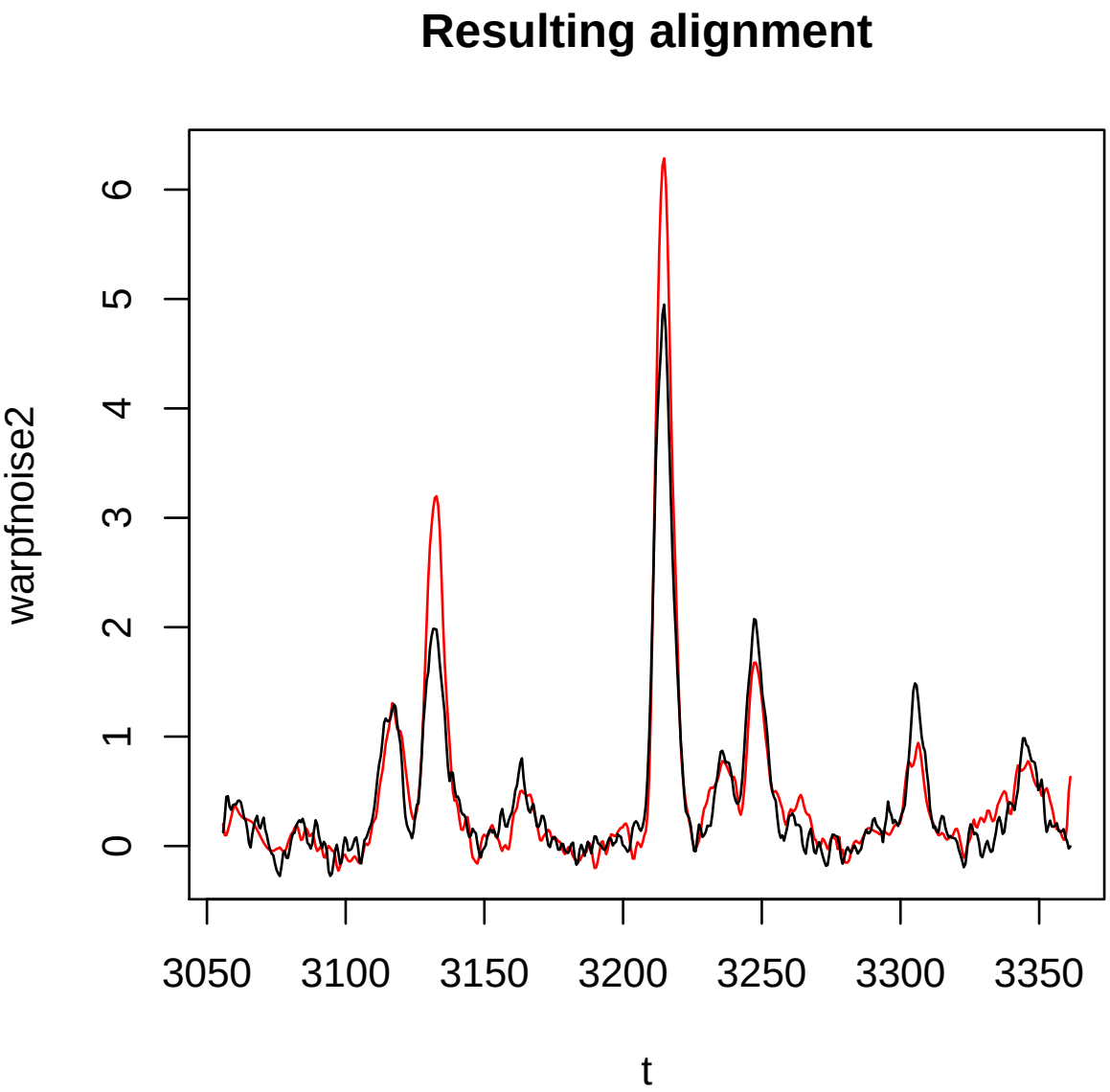
Pour détecter les extrema on utilise une approche “excessive mass” :



L’alignement final est alors obtenu en estimant de manière non paramétrique la fonction de déformation qui aligne deux ensembles de N landmarks ordonnés appartenant à deux spectres (transformation non-rigide développée par Chui et Rangarajan (2000)).


Spectra to be aligned
fnoise2
0
1
2
3
4
5
6
3050
3100
3150
3200
3250
3300
3350
t


Structural intensities of estimated zero−crossings
densMax1
0.5
1.0
1.5
2.0
2.5
3050
3100
3150
3200
3250
3300
3350
t


Resulting alignment
warpfnoise2
0
1
2
3
4
5
6
3050
3100
3150
3200
3250
3300
3350
t

Alignements de plusieurs spectres

- On propose de prendre pour spectre référant un spectre moyen (obtenu comme moyenne des spectres débruités) ;
- **Approche de Morris et al., 2005 :** elle utilise un spectre référant obtenu en débruitant après avoir fait la moyenne. Cette méthode suppose que les spectres sont déjà bien calibrés (loi des grands nombres). On montre les bénéfices obtenus en débruitant avant de faire la moyenne lorsque ce n'est pas le cas.
- Notre approche a été illustrée sur données réelles et comparée avec l'alignement par splines de (Jeffries, 2005). Le problème avec ce type d'algorithme est le choix des pics initiaux pour l'alignement.

Analyse de la variance fonctionnelle

- **Objectifs :** Une fois le pré-traitement effectué, on veut comparer les spectres d'individus "normaux" et d'individus malades.
- Comme il s'agit de données fonctionnelles il est naturel d'utiliser des analyses de variance adaptées (FANOVA, Abramovich et al., 2004). Cette méthode permet de réduire la dimension via la représentation creuse en ondelettes des spectres
- **Intérêt :** Évite d'estimer au préalable les pics. Même en admettant que le problème de la sélection des pics soit résolu, le nombre de pics sélectionnés peut être important et le problème des comparaisons multiples avec données dépendantes est alors posé.

Exemple de données traitées

- Spectrométre de type SELDI-TOF (Ciphergen) à partir d'échantillons de sérum de personnes saines du sexe féminin et de sérum de femmes atteintes d'un cancer de l'ovaire.
- Données : Petricoin *et al.* (2002) (08- 07-02).
- Chaque spectre représente l'expression de 15154 peptides définis par leur ratios m/z. Les valeurs de m/z (en Daltons par coulomb) sont identiques pour tous les spectres.
- 253 individus sous 2 conditions :
  - Normale (91)
  - Cancer (162)

Question

Identifier des différences caractéristiques dans les protéines exprimées entre patientes atteintes d'un cancer et les patientes saines à l'aide du serum.

Modèle à effet aléatoire non paramétrique

Modèle de base : $Y_i(t)$ $(i = 1, 2, \ldots, n = n_1 + n_2)$ est le spectre du $i$ème sujet en le point t et peut être modélisé par

$$Y_i(t) = \mathbf{X}_i \boldsymbol{\beta}(t) + \alpha^{(i)}(t) + \epsilon_i(t),$$

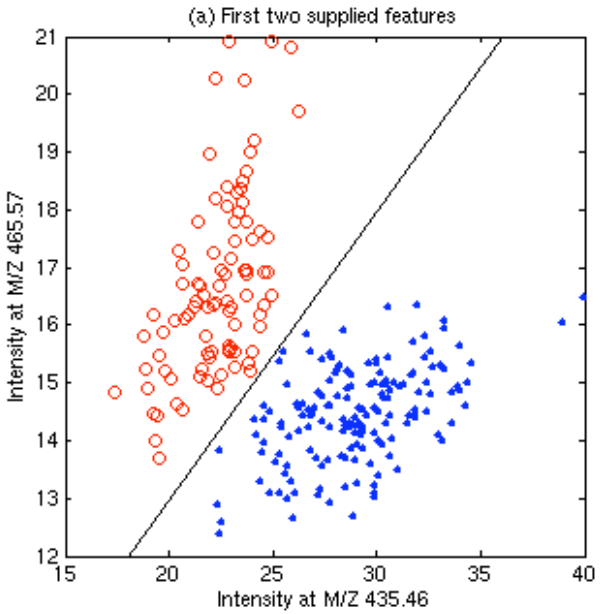
avec $n = 253$ et

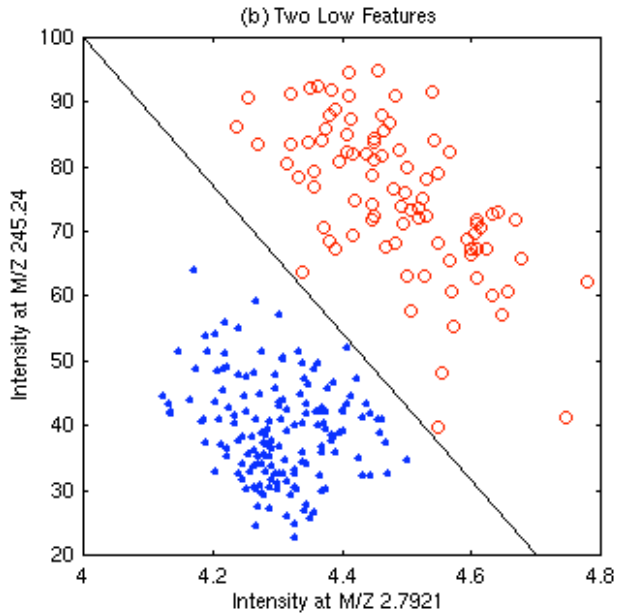
- $\boldsymbol{\beta}(t) = (\beta_1(t), \beta_2(t))^{\mathrm{T}}$ le vecteur des fonctions spectres moyens (*effet fixe*)
- $\alpha^{(i)}(t)$ des fonctions *aléatoires* stochastiquement indépendantes modélisées comme réalisations de processus gaussiens centrés de noyaux de covariances paramétriques
- $\mathbf{X}_i = (X_i[1], X_i[2])$ les vecteurs du plan d'expérience , et
- $\epsilon_i(t)$ des bruits blancs indépendants des $\alpha^{(i)}(t)$.

Modélisation par ondelettes

- La modélisation des effets fixes et aléatoires par ondelettes permet une représentation plus ou moins irréguliére tant pour les effets fixes que pour les effets aléatoires et permet de s'assurer que ces derniers prennent leurs valeurs dans le même espace fonctionnel.
- On se ramène à un modèle linéaire à effets mixtes à une composante de la variance pour lequel les effets fixes sont paramétrisés par les coefficients d'ondelette des $\beta_k(t)$ $(k = 1, 2)$ et les effets aléatoires par les coefficients des $\alpha^{(i)}(t)$ $(i = 1, 2, \ldots, n)$.
- On peut tester la présence d'effet aléatoire en utilisant les résultats développés par Antoniadis et Sapatinas (2005) pour réaliser ce test. Il en est de même pour tester le contraste d'égalité des effets fixes (égalité des coefficients des fonctions β_1 et β_2) .

Les spectres étant différenciés selon la condition et cela en prenant compte l'hétérogenéité des individus, on peut se servir des statistiques de test pour identifier les protéines les plus discriminantes.





Analyse discriminante linéaire en ne retenant que deux pics (données de Petricoin et al.)