Méthodes multivariées combinant ondelettes et analyse en composantes principales pour le débruitage de données issues de spectrométrie de masse

Elise Mostacci, Caroline Truntzer, Hervé Cardot, Patrick Ducoroy

METHODES MULTIVARIEES COMBINANT ONDELETTES ET ANALYSE EN COMPOSANTES PRINCIPALES POUR LE DEBRUITAGE DE DONNEES ISSUES DE SPECTROMETRIE DE MASSE

Elise Mostacci [1,3,4], Caroline Truntzer [1,2], Hervé Cardot [3,4], Patrick Ducoroy [1,2]

[1] Clinical and Innovation Proteomic Platform, Dijon, F-21000, France
[2] Centre Hospitalier Universitaire, Dijon, F-21000, France
[3] Institut de Mathématiques de Bourgogne, UMR CNRS 5584, Dijon, F-21000, France
[4] Université de Bourgogne, Dijon, F-21000, France

Mots clé: protéomique clinique / débruitage / spectrométrie de masse / ACP / ondelettes

Résumé:
L'identification de nouveaux biomarqueurs diagnostiques ou pronostiques est un des objectifs majeurs en recherche clinique. L'utilisation des technologies à haut débit comme la spectrométrie de masse est prometteuse pour l'identification de tels marqueurs. A partir d'un prélèvement de sang ou de tumeur par exemple, cette technologie permet de traduire sous forme de spectres le profil protéique des individus. Le signal biologique observé dans les spectres est masqué par différentes sources de variabilités techniques, qu'une phase préalable de prétraitement doit permettre de retirer. La méthode classique permettant de retirer le bruit aléatoire de mesure de ce signal combine la méthodologie des ondelettes et un seuillage, ceci spectre à spectre. L'utilisation des ondelettes permet la séparation du bruit aléatoire (coefficients de détail) et du signal biologique (coefficients d'approximation). Le seuillage des coefficients de détails permet ensuite d'annuler un certain nombre d'entre eux.
Nous proposons dans ce travail d'améliorer le débruitage classique des données en tenant compte de la structure commune des signaux. Nous avons pour cela adapté aux données spectrométriques deux méthodes de débruitage qui combinent les ondelettes, le seuillage et les analyses en composantes principales classique ou creuse. Les méthodes proposées ont été évaluées et comparées à la méthode de seuillage univariée classique, ceci pour des données réelles et simulées. Il a été montré que l'ajout d'une étape de réduction de la dimension sur les approximations par une analyse en composantes principales en plus d'un seuillage classique sur les détails améliore le débruitage.

Abstract:
The identification of new diagnostic or prognostic biomarkers is one of the main aims of clinical research. The use of high-throughput technology like mass spectrometry is a promising tool for the detection of such biomarkers. Using samples of blood or tumor for example, mass spectrometry measurements reflect through spectra the proteomic profiles of the individuals under study. The observed biological signal is contaminated by different sources of technical variability that can be removed by a prior pre-processing step. Wavelet methodology associated with thresholding is usually used for this purpose. The use of wavelets allows the separation of the random noise (detail coefficients) and the biological signal of interest (approximation coefficients). The thresholding of detail coefficients then allows the shrinking of small coefficients to zero.
In this work we adapted to spectrometric data two multivariate denoising methods that combine wavelets, thresholding rule and principal component analysis (PCA) or sparse PCA. The proposed methods were evaluated and compared with the classical thresholding denoising method using both real and simulated datasets. It was shown that taking into account common structures of the signals by adding a dimension reduction step on approximation coefficients through PCA provided more effective denoising when combined with soft thresholding on detail coefficients.

Introduction

La protéomique clinique a pour objectif l'identification de nouveaux biomarqueurs diagnostiques ou pronostiques. La spectrométrie de masse est une technologie à haut débit de choix pour la protéomique clinique. Les données issues de spectrométrie de masse se présentent sous la forme de spectres de plusieurs dizaines de milliers de points de mesure qui reflètent le profil protéique d'un patient. Lors de leur acquisition, ces spectres sont affectés par différentes sources de variabilité qui masquent le signal biologique et qui doivent donc être retirées lors d'une phase de prétraitement.

La technique classiquement utilisée pour l'élimination du bruit est basée sur la décomposition en ondelettes de chacun des spectres. Par ailleurs, l'Analyse en Composantes Principales (ACP) est une technique classiquement utilisée pour éliminer les composantes non informatives contenues dans des données de grande dimension.

Nous proposons dans ce travail d'améliorer le débruitage classique des données issues de spectrométrie de masse grâce à une approche multivariée qui combine les ondelettes, le seuillage et les analyses en composantes principales classique ou creuse.

Dans la suite du document, nous commencerons par présenter la méthode de débruitage classique combinant ondelettes et seuillage doux. Nous poursuivrons par la description des méthodes multivariées que nous proposons puis par leur comparaison. Enfin, nous discuterons des résultats.

Matériel et méthode

1) Décomposition en ondelettes et seuillage doux

La technique de débruitage classique combinant ondelettes et seuillage doux (soft thresholding) a pour objectif de séparer le bruit aléatoire du signal biologique. Cette technique peut être décomposée en 3 étapes. La première consiste à calculer la décomposition du signal jusqu'au niveau J en utilisant la transformée en ondelettes [1, 2]. Deux types de coefficients sont alors obtenus: les approximations et les détails ; les approximations décrivent la forme globale du signal tandis que les détails décrivent les variations plus fines. Les détails de faible intensité correspondent à la partie bruitée du signal. La seconde étape consiste alors à seuiller les détails par un seuillage doux qui pénalise les coefficients.

Pour un seuil δ fixé, le seuillage doux des détails de niveau j est réalisé selon la formule suivante

$$\tilde{d}_{j,k} = Sign(d_{j,k})(|d_{j,k}| - \delta)_+ \quad (1)$$

Le seuil généralement utilisé est le seuil universel global proposé par Donoho et Johnstone [3]

$$\delta = \hat{\sigma}\sqrt{2\log(p)} \quad (2)$$

où p est la longueur du signal et $\hat{\sigma}$ une estimation du bruit basée sur l'écart absolu median (mad) des détails de niveau 1

$$\hat{\sigma} = \sqrt{2}mad(d_1)/0.6745 \quad (3)$$

Pour améliorer le débruitage, nous avons choisi de modifier ce seuil pour qu'il prenne en compte l'ordre de grandeur des détails.

$$\delta = \hat{\sigma}\sqrt{2\log(p) * \max|d_1|} \quad (4)$$

Cette modification empirique a permis d'améliorer les résultats du débruitage dans notre étude.

Dans la suite, le terme SOFT fait référence à la technique de seuillage doux qui utilise le seuil (4).

Enfin, la dernière étape consiste à reconstruire le signal par la transformée en ondelettes inverse à partir des détails seuillés et des approximations.

Nous avons choisi d'utiliser la transformée en ondelettes MODWT (Maximal Overlap Discret Transform) ainsi que la transformée inverse correspondante (IMODWT) pour réaliser la décomposition et la reconstruction des signaux.

2) MW-PCA

L'ACP est une méthode classiquement utilisée pour extraire l'information de données de grande dimension. Elle permet de réduire le nombre de variables par projection des données dans un espace de dimension inférieure. Les nouvelles variables, combinaisons linéaires des variables d'origine, sont appelées composantes principales.

Dans le cadre du débruitage de signaux multicanaux, Aminghafari *et al.* [4] proposent la méthode MW-PCA qui associe l'ACP au débruitage par ondelettes. Les auteurs proposent de seuiller les détails après leur projection dans la base de l'ACP, puis d'utiliser l'ACP sur les approximations pour extraire les caractéristiques les plus importantes des spectres.

Les étapes de la méthode MW-PCA sont les suivantes :

I. Réaliser la transformée MODWT de chaque spectre jusqu'au niveau *J*. On obtient 2*J* matrices de taille 2^jpxn : $D_1,\ldots, D_J$ et $A_1,\ldots, A_J$ contenant respectivement les détails et les approximations pour chaque niveau de décomposition.

II. Définir $\sum$ estimation de la matrice de variance covariance du bruit basée sur les détails de niveau 1. Calculer la matrice V telle que $\sum=U\Lambda V^t$ où $D=diag(\lambda_i)_{1\le i\le n}$. Projeter chaque niveau de détail dans la base de l'ACP : D_jV, j=1,…,*J*. Appliquer un seuillage doux en utilisant le seuil

$$\delta_i = \sqrt{2\lambda_i \log(p) * \max|dv_i|} \qquad (6)$$

où dv_i est la $i^{ème}$ colonne de D_jV. Comme dans la méthode SOFT la formule du seuil proposée par les auteurs a été modifiée pour tenir compte de l'ordre de grandeur des détails.

III. Réaliser une ACP sur les approximations de niveau J: A_J^t.

IV. Réaliser le changement de base des détails avec V^t.

V. Reconstruire les signaux via la transformée inverse IMODWT.

Pour tester la pertinence de l'ACP sur les approximations dans le cas de données issues de spectrométrie de masse, nous proposons deux versions de cette méthode : MW-PCA1 (suppression de l'étape III) et MW-PCA2 (étapes I à V).

3) MW-SPCA

Dans l'ACP toutes les variables contribuent à chacune des composantes principales, et tous les facteurs sont non nuls. Certains auteurs ont alors proposé l'ACP creuse (Sparse Principal Component Analysis) dont l'objectif est non seulement de réduire la dimension mais aussi de réduire le nombre de variables exprimées dans les composantes. Parmi les algorithmes disponibles pour l'ACP creuse, Zou *et al.* [5] proposent d'écrire l'ACP comme un problème d'optimisation de type régression qui dans le cas des données de grande dimension se ramène à un seuillage de type doux. L'algorithme est donc rapide.

Soit X, de taille nxp, une matrice centrée-réduite dont la décomposition en valeurs singulières est $X=U\Lambda V^t$. L'algorithme SPCA de Zou *et al.* est le suivant :

1. Initialiser $\alpha=V[,1:q]$, les facteurs des q premières composantes principales classiques.
2. Pour α fixé et pour j=1,2,…,q, appliquer le seuillage doux à chaque β_j

$$\beta_j = Sign(\alpha_j^t X^t X)\left(\left|\alpha_j^t X^t X\right| - \delta\right)_+ \qquad (7)$$

3. Pour chaque β, calculer la décomposition en valeurs singulières de $X^tX\beta=UDV^t$, puis actualiser $\alpha= UV^t$.
4. Répéter les étapes 2 et 3 jusqu'à la convergence de β.

5. Normaliser $\tilde{V}_j = \frac{\beta_j}{|\beta_j|}$, j=1,…q.

Le choix du seuil δ prend en compte le nombre de facteurs annulés sur les composantes ainsi que le pourcentage de variance expliquée par les composantes principales ainsi modifiées [5]. Le nombre optimal de composantes est celui qui maximise le pourcentage de variance expliquée. L'algorithme est disponible dans la librairie elasticnet R [6] disponible sur le site CRAN (http://cran.r-project.org).

Nous avons choisi de combiner un débruitage classique sur les détails et une ACP creuse sur les approximations. L'algorithme MW-SPCA est décrit ci-dessous :

I. Réaliser la transformée MODWT de chaque spectre jusqu'au niveau *J*.
II. Réaliser un seuillage doux des détails jusqu'au niveau *J*.
III. Réaliser l'algorithme SPCA pour les approximations de niveau *J*. Reconstruire les approximations à partir des composantes creuses.
IV. Reconstruire par IMODWT les signaux débruités à partir des détails et des approximations modifiés.

4) Critères d'évaluation des méthodes

Toutes ces méthodes ont été comparées selon deux critères : le ratio signal sur bruit (S/N) et la reproductibilité des pics (R).
Nous avons choisi la définition du S/N proposée par Aminghafari *et al.* [4], basée sur la taille des spectres et une estimation du bruit. Pour un signal f

$$S/N(f) = 20\log_{10}\left(\max|f| / \hat{\sigma}(f)\right) \quad (8)$$

où $\hat{\sigma}(f) = mad(d_1)$ et d_1 les détails de niveau 1 du signal f. Plus le S/N est élevé plus la proportion du signal par rapport au bruit est forte. L'amélioration du S/N est évaluée en comparant le S/N avant (S/N_{ori}) et après débruitage (S/N_{fin})

$$S/N = \frac{S/N_{fin}}{S/N_{orig}} * 100 \quad (9)$$

La détermination de la reproductibilité des pics détectés est une autre façon d'évaluer l'efficacité du processus de débruitage. La reproductibilité des pics est mesurée en calculant la quantité suivante :

$$R = \frac{Nombre\ de\ pics\ communs}{Nombre\ total\ de\ pics} * 100 \quad (10)$$

Le nombre total de pics correspond au nombre de pics détectés sur tous les spectres alors que le nombre de pics communs correspond aux pics détectés dans au moins 50% des spectres. Plus R augmente plus la reproductibilité augmente, ce qui traduit un débruitage performant.
La librairie R MassSpecWavelet (www.bioconductor.org) développée par Du *et al.* [7, 8] a été utilisée pour la détection des pics.
La visualisation des spectres après débruitage a été utilisée pour s'assurer de la qualité des spectres reconstruits. Nous avons vérifié que le débruitage n'était pas trop fort et que la forme globale des spectres était préservée.

5) Données

Les méthodes ont été évaluées sur des données réelles et simulées. Les 3 jeux de données réelles proviennent d'échantillons de plasma ou de cellules et sont répartis en 2 groupes diagnostics comme décrit dans le tableau 1. Le jeu de données simulé a été obtenu en suivant le procédé proposé par Morris *et al.* [9]. Il est composé de 186 spectres également répartis en 2 groupes diagnostiques.
Les spectres ont été alignés et normalisés avant le débruitage et la détection des pics.

	Origine des échantillons	Groupe diagnostique	Nombre de Spectres par groupes	Nombre de réplicats
Data1	Plasma	rechute/non rechute	24/24	4
Data2	Cellule	marqueurs de stress	72/104	8
Data3	Plasma	2 types de cellules	72/152	8

Tableau 1. Caractéristiques des jeux de données réelles.

Résultats

Les méthodes ont été comparées pour le niveau de décomposition J=3, que nous avons jugé optimal grâce à une étude préliminaire. Les résultats du S/N pour le jeu de données réelles Data2-G1 et pour le jeu de données simulées G1 sont présentés figure 1. Des résultats semblables ont été obtenus pour les autres jeux de données. La figure 2 représente les valeurs de la reproductibilité des pics obtenues pour les différents jeux de données réelles et simulées (G1).

La comparaison de MW-PCA1 à SOFT a permis d'évaluer la contribution d'un débruitage multivarié des détails par rapport à un débruitage univariée. Les figures 1 et 2 montrent que le caractère multivarié de la méthode MW-SPCA1 n'a pas permis d'améliorer les résultats. A noter que les performances de l'ACP creuse pour le traitement des détails ont aussi été évaluées mais elle s'est révélée moins performante que les méthodes SOFT et MW-PCA1.

La comparaison de MW-PCA2 et MW-SPCA aux autres méthodes a permis d'évaluer l'intérêt de la réduction de dimension des approximations en plus du seuillage des détails. L'utilisation de ces deux méthodes a permis d'améliorer le S/N et la reproductibilité des pics quelque soit le jeu de données. Ainsi, la réduction de la dimension des approximations a été bénéfique.

Enfin la comparaison de MW-PCA2 avec MW-SPCA a permis d'évaluer le besoin de choisir seulement quelques coefficients lors de la réduction de la dimension des approximations. Cette question est plus délicate puisque les résultats de MW-PCA2 et MW-SPCA sont proches. Les deux méthodes ont leurs avantages et leurs inconvénients: l'ACP classique est rapide et exige seulement le choix d'un paramètre, le nombre de composantes à conserver. L'ACP creuse est plus coûteuse en temps de calcul et exige l'évaluation du nombre de composantes à conserver ainsi que de la pénalité δ qui peut être long et difficile. Cette méthode, cependant, fournit une réduction de dimension plus efficace et cela peut être utile si le but est de construire ensuite un modèle pronostique sur un nombre réduit de variables.

Pour résumer, d'une part, les résultats sur les détails montrent qu'il n'existe pas de structure de corrélation dans le bruit aléatoire de mesure; il est uniformément réparti dans tous les spectres. D'autre part, il y a bénéfice à prendre en compte la structure de dépendance qui existe entre les spectres lorsque l'on considère leur forme globale.

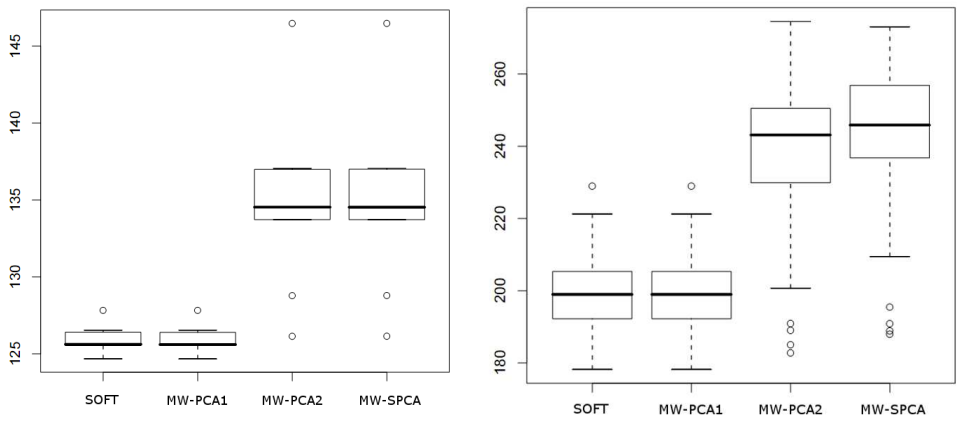

Figure1. Boîtes à moustache représentant le S/N pour les méthodes SOFT, MW-PCA(1-2) et MW-SPCA. A gauche : Data2-G1. A droite : données simulées.

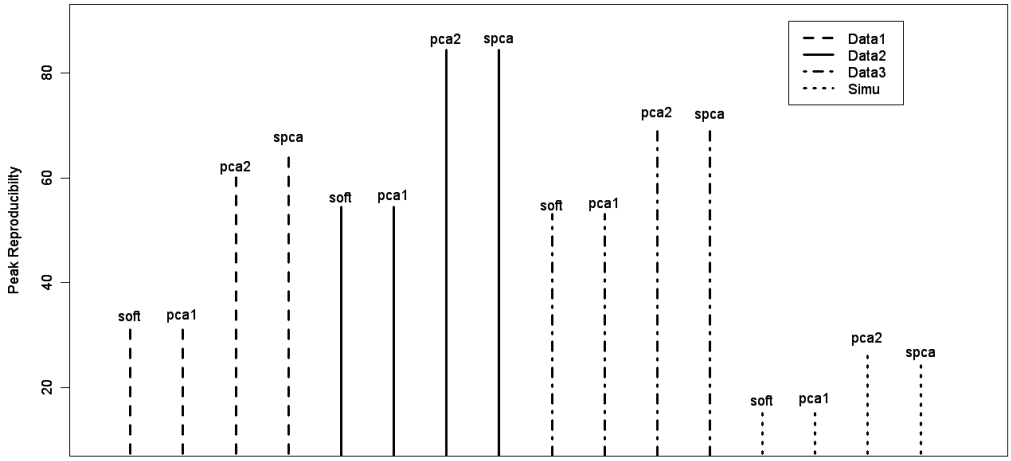

Figure2. Représentation des valeurs de la reproductibilité des pics obtenues pour les données réelles et simulées avec les méthodes SOFT, MW-PCA(1-2) et MW-SPCA.

Conclusion

Dans ce travail, nous nous sommes intéressés à des méthodes de débruitage multivariées qui traitent les deux types de coefficients d'ondelettes. Par l'intermédiaire de 3 méthodes nous avons évalué l'intérêt de 1) utiliser un débruitage multivarié plutôt qu'un débruitage univarié sur les détails et 2) utiliser une réduction de la dimension sur les approximations par ACP classique ou ACP creuse. L'évaluation des méthodes sur des jeux de données réelles et simulées nous a permis de montrer que la méthode consistant à combiner un seuillage doux sur les détails et une réduction de la dimension avec une ACP classique ou creuse permettait d'optimiser le débruitage.

Bibliographie

[1] Daubechies, I. (1992) *Ten Lectures On Wavelets*, CBMS-NSF Regional Conference Series in Applied Mathematics.

[2] Mallat, S. (1999) *A Wavelet Tour Of Signal Processing*. Ed. Academic Press.

[3] Donoho, D. et Johnstone, I. (1994) Ideal spatial adaptation by wavelet shrinkage, *Biometrika*, 81, 425-455.

[4] Aminghafari, M., Cheze, N. et Poggi, J-M. (2006) Multivariate denoising using wavelets and principal component analysis, *Computational Statistics & Data Analysis*, 50, 2381-2398.

[5] Zou, H., Hastie, T. et Tibshirani, R. (2006) Sparse Principal Component Analysis, *Journal of Computational and Graphical Statistics*, 15, 262-286.

[6] Zou, H. et Hastie, T. (2005) Regularization and variable selection via the elastic net. *Royal Statistical Society*, 67, 301-320.

[7] Zou, H. et Hastie, T. (2007) *elasticnet: Elastic Net regularization and variable selection*, R package version 1.0.4.

[8] Cruz-Marcelo, A., Guerra, R., Vannucci, M., Li, Y., Lau, C. et Man, T-K (2008) Comparison of algorithms for pre-processing of SELDI-TOF mass spectrometry data, *BMC bioinformatics*, 24, 2129-2136.

[9] Morris, J., Coombes, K., Koomen, J., Baggerly, K. et Kobayashi, R. (2005) Feature extraction and quantification for mass spectrometry in biomedical applications using the mean spectrum, *Bioinformatics*, 21, 1764-1775.