Robustesse des approches chimiométriques pour la reconstruction de profils moléculaires

C. PAULUS [(1)] , G. STRUBEL [(1)] , L. GERFAULT [(1)] , P. GRANGEAT [(1)]

[(1)] LETI, MINATEC, Département des microTechnologies pour la Biologie et la Santé, CEA-GRENOBLE,

17 Rue des Martyrs, 38054 GRENOBLE Cedex 9, France, caroline.paulus@cea.fr

Résumé – Ce papier traite des approches chimiométriques pour la reconstruction de profils moléculaires. La quantification de protéines du sang est réalisée à partir du traitement par analyse factorielle de spectrogrammes issus d'une chaîne d'analyse contenant une colonne de nano-chromatographie et un spectromètre de masse. Nous nous intéressons plus particulièrement à la comparaison de la robustesse des méthodes de régression de type Unfold-PLS, N-PLS et PARAFAC vis-à-vis du problème de décalage temporel des pics contenus dans les spectrogrammes. Les méthodes multidimensionnelles type N-PLS et PARAFAC fournissent de meilleurs résultats ce qui permet d'envisager une quantification des protéines avec une plus grande tolérance sur le recalage des pics en temps de rétention.

Abstract – This paper deals with molecular profile reconstruction for blood proteins quantification. We focus on factor analysis treatment of spectrograms coming from the analysis chain combining liquid chromatography and mass spectrometry. Quantification of protein is performed using chemometrics approaches and more precisely Unfold-PLS, N-PLS and PARAFAC algorithms. We compare the robustness of these calibration methods with respect to the problem of retention peaks time shift. Multidimensional methods (N-PLS and PARAFAC) provide better results which enable to consider the quantification problem with a greater tolerance on retention time shift.

1. Contexte et problématique

La problématique de ce papier s'insère dans le projet LOCCANDIA [1] qui vise à permettre une détection précoce du cancer du pancréas à partir d'une chaîne d'analyse protéomique allant de l'échantillon de sang au diagnostic et combinant la biologie, les nanotechnologies et le traitement de l'information. Nous nous intéressons à la partie traitement des données, appelée reconstruction de profils moléculaires (FIG. 1). Cette étape doit permettre d'estimer la concentration de certaines protéines cibles (caractéristiques de la maladie) présentes dans l'échantillon sanguin à partir de spectrogrammes 2D obtenus en sortie de la chaîne d'analyse. Une dimension de cette image est le temps de rétention, associée à la séparation sur une nano colonne de chromatographie liquide (nano-LC), l'autre dimension correspond au rapport masse sur charge des protéines obtenu par un spectromètre de masse.

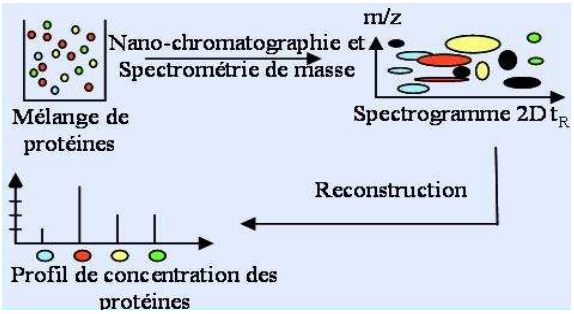

FIG. 1 : Reconstruction d'un profil moléculaire

La méthodologie proposée se base sur une approche de type problème inverse associé à un modèle fonctionnel direct reliant les inconnues (les concentrations des protéines cibles) aux mesures (les spectrogrammes). Nous supposons que notre système est linéaire et reproductible. Nous nous intéressons aux approches chimiométriques. La chimiométrie [2] fait référence à des techniques mathématiques liées aux traitements de données issues d'instrumentation dédiée à la chimie. Le but est d'extraire l'information pertinente de données physico-chimiques mesurées en construisant et en exploitant un modèle multivariable. Celui-ci relie les variables *Xvar*, dont l'estimation est délicate (les concentrations) aux variables *Yvar* facilement mesurables (les spectrogrammes). Deux étapes sont nécessaires. La première, appelée **étalonnage**, correspond à la construction du modèle mathématique à partir de mélanges d'étalonnage (*Xvar* et *Yvar* sont connues). La deuxième étape, appelée **prédiction**, permet pour des mélanges de composition inconnue d'estimer les concentrations *Xvar* à partir des données mesurées *Yvar* et du modèle construit à l'étape d'étalonnage. Les variables sont liées par la relation tensorielle suivante :

$$\underline{Y} = \underline{H} \times_1 X \qquad (1)$$

où $\underline{\mathbf{Y}}$ est le tenseur des mesures de taille (M, J, K), $\mathbf{X}$ la matrice des concentrations de taille (M, N), $\underline{\mathbf{H}}$ le tenseur d'interaction contenant les spectrogrammes théoriques des protéines cibles en concentration unitaire de taille (N, J, K) et $\times_1$ le produit n-mode[1] [3] selon la première dimension du tenseur. J et K correspondent aux dimensions d'un spectrogramme, M au nombre de mélanges et N au nombre de protéines cibles. $\underline{\mathbf{H}}$ est estimé à l'étape d'étalonnage. Ce modèle peut aussi s'écrire sous forme 2D (les spectrogrammes sont dépliés en longs vecteurs) :

[1] Le produit n-mode permet de réaliser le produit d'un tenseur par une matrice. Le résultat est un tenseur possédant autant de modes que le tenseur de départ mais dont la taille d'une de ces dimensions a été modifiée.

$$Y = HX \quad (2)$$

où **Y**, **H** et **X** sont respectivement des matrices de taille (JK, M), (JK, N) et (N, M).

Un problème important lié aux données chromatographiques est l'**incertitude sur la position en temps de rétention des pics du spectrogramme** [4]. D'une analyse à une autre et pour une même protéine, la position des pics le long de la dimension chromatographique est susceptible de varier. Bien que des méthodes d'alignements permettent de recaler ces pics [4], des imprécisions sur la position peuvent tout de même subsister. Cette incertitude introduit un biais dans l'estimation des concentrations. Nous proposons de comparer la robustesse de différentes méthodes chimiométriques vis-à-vis de l'incertitude sur la position des pics. Les méthodes testées sont la méthode **« Unfold – PLS » (Partial Least Square)** et deux méthodes multidimensionnelles : **N-PLS et PARAFAC (Parallel Factor Analysis)**.

2. Méthodes de prédiction par analyse factorielle

L'analyse factorielle réalise une diminution du nombre des variables *Yvar* en un nombre réduit de combinaisons linéaires des *Yvar*, appelées facteurs, dans le but de concentrer l'information utile dans un espace de plus petite dimension et permettre ainsi une régularisation de l'inversion en réduisant le nombre d'inconnus.

2.1 Unfold-Partial Least Square (Unfold-PLS)

La méthode PLS a été développée et popularisée en science analytique par Wold [5]. « Unfold-PLS » est identique à la régression PLS mais est appliquée à des données tensorielles réarrangées sous forme vectorielle (cf. Eq. (2)). Cette technique, proche de la méthode PCR (« Principal Component Regression »), prend en compte à la fois l'information sur les concentrations *Xvar* et les spectrogrammes *Yvar* lors de l'étape d'étalonnage. La méthode PLS réalise donc une extraction des vecteurs propres à partir des matrices de concentration **X** et des matrices de mesures **Y** des mélanges d'étalonnage.

Il existe plusieurs types de régression PLS. La plus simple, appelé PLS1, s'applique au cas où **x** est un simple vecteur (une seule protéine cible par exemple). Dans le cas où l'on recherche plusieurs protéines cibles, X correspond à une matrice et on utilise alors l'algorithme PLS2. Nous allons décrire successivement les étapes d'étalonnage et de prédiction de l'algorithme PLS1.

Pour chacun des K facteurs à inclure dans le modèle, les étapes de l'algorithme d'**étalonnage** sont les suivantes:

- Initialisation : $Y_0 = Y$ et $x_0 = x$
- Une composante de la matrice de changement de base "loading vector" w_k est calculée pour déterminer la composante de $\mathbf{Y_k}$ la plus corrélée aux variations de concentrations $\mathbf{x_k}$:

$$\max_{w_k}\left[\operatorname{cov}(t_k, x_{k-1}) \;\middle|\; t_k = Y_{k-1}w_k \quad et \quad \|w_k\| = 1\right] \quad (3)$$

Soit après résolution de la fonctionnelle :

$$w_k = \frac{{Y_{k-1}}^T x_{k-1}}{\left\|{Y_{k-1}}^T x_{k-1}\right\|} \quad (4)$$

- Le projeté "score" est ensuite estimé en projetant $\mathbf{Y_{k-1}}$ sur $\mathbf{w_k}$:

$$t_k = Y_{k-1}w_k \quad (5)$$

- Le vecteur "loading" des spectrogrammes $\mathbf{v_k}$ est déterminé par régression de $\mathbf{Y_{k-1}}$ sur $\mathbf{w_k}$ et celui des concentrations $\mathbf{q_k}$ en régressant $\mathbf{x_{k-1}}$ sur $\mathbf{w_k}$:

$$v_k = {Y_{k-1}}^T t_k ({t_k}^T t_k)^{-1} \quad (6)$$

$$q_k = {x_{k-1}}^T t_k ({t_k}^T t_k)^{-1} \quad (7)$$

- Les résidus $\mathbf{Y_k}$ et $\mathbf{x_k}$ sont formées par déflation en enlevant les effets de ce facteur aux données (on retranche l'information déjà modélisée):

$$Y_k = Y_{k-1} - t_k v_k^T \quad (8)$$

$$x_k = x_{k-1} - t_k q_k^T \quad (9)$$

- Reprendre à la deuxième étape jusqu'à la détermination des K facteurs souhaités.

Une fois le modèle estimé, il est utilisé pour l'étape de **prédiction**. Lors de cette étape, le modèle et les spectrogrammes **Y** sont utilisés afin de prédire les concentrations des protéines cibles dans les mélanges de prédiction. Ainsi, on extrait successivement de **Y** les informations correspondant à chaque facteur avec lesquelles on construit la concentration **x** (initialement nulle) :

- Calcul du "score" : $t_k = Y_k w_k$
- Calcul de la contribution de ce « score » sur la concentration :

$$x_{k+1} = x_k + t_k q_k^T \quad (10)$$

- Calcul du résidu:

$$Y_{k+1} = Y_k - t_k v_k^T \quad (11)$$

- Itération jusqu'à k=K

L'utilisation simultanée de l'information sur **X** et **Y** à l'étape d'étalonnage permet d'obtenir de meilleurs résultats de prédiction qu'avec la méthode PCR qui n'utilise que l'information sur **Y**. Considérons des mélanges complexes de protéines pour l'étalonnage dont la matrice de concentration ne contient que l'information de concentration des protéines cibles. Si l'on réalise l'analyse en composantes principales de ce mélange alors les composantes obtenues peuvent ne contenir que très peu d'information relative aux protéines cibles si celles-ci sont en minorité dans le mélange. Étant donné que la PLS recherche l'espace des facteurs le plus conforme aux variables **X** et **Y**, sa prédiction est meilleure.

2.2 Méthodes multidimensionnelles

Les méthodes multidimensionnelles [6] permettent de garder la cohérence tridimensionnelle des données car elles traitent les données tensorielles sous leur forme initiale (cf. eq. (1)). Dans notre cas, $\underline{\underline{\mathbf{Y}}}$ est un tenseur d'ordre 3 (1ère dimension : mélange, 2ème dimension : rapport masse sur charge, 3ème dimension : temps de rétention).

2.2.1 N-way Partial Least Square (N-PLS)

La méthode N-PLS [7] est une extension de la méthode PLS aux tableaux d'ordre supérieur. Elle correspond à une décomposition du tenseur $\underline{\underline{\mathbf{Y}}}$ d'étalonnage en un ensemble de tenseurs de rang 1. Les modèles obtenus sont généralement beaucoup plus simples et robustes et permettent de réduire le risque de surmodélisation par rapport aux approches de type unfold-PLS.

La régression N-PLS correspond à la décomposition du tenseur $\underline{\underline{\mathbf{Y}}}$ de taille (M, J, K) en triades (1 vecteur « score » **t** et 2 vecteurs « weight » $\mathbf{w^J}$,$\mathbf{w^K}$, un selon chaque mode). Le modèle de décomposition de $\underline{Y}$ est le suivant :

$$y_{ijk} = t_i w_j^{\ J} w_k^{\ K}$$

Le problème revient à trouver un couple de vecteurs (w^J,w^K) engendrant un vecteur « score » de covariance avec **x** maximale. Le critère est donc le suivant :

$$\max_{w_J w_K}\left[\operatorname{cov}(t,x)\middle| t_i = \sum_{j=1}^{J}\sum_{k=1}^{K} y_{ijk} w_j^{\ J} w_k^{\ K} et \left\|w^J\right\| = \left\|w^K\right\| = 1\right] \quad (12)$$

Soit après résolution de la fonctionnelle :

$$\left(w^J, s, w^K\right) = svd(Z,1) \text{ avec } z_{jk} = \sum_{i=1}^{M} x_i y_{ijk} \quad (13)$$

où svd(**Z**,1) correspond à la première composante de la décomposition en valeurs singulières (Singular Value Decomposition) de **Z**. L'algorithme est ensuite très similaire à celui de la PLS : calcul de $\mathbf{w^J}$ et $\mathbf{w^K}$, calcul du vecteur **t**, régression puis calcul des résidus par déflation.

2.2.2 Parallel Factor Analysis (PARAFAC)

La décomposition PARAFAC [8] correspond à une généralisation de l'analyse en composantes principales (PCA) aux tenseurs d'ordre N. Elle fut initiée par Harshman en 1970 [9]. Cependant, certaines caractéristiques diffèrent de la décomposition à l'ordre deux. Tout d'abord, la décomposition PARAFAC est unique. De plus, à la différence de la PCA, les facteurs ne sont pas estimés successivement. Enfin, la décomposition utilise à la fois les mélanges d'étalonnage et de prédiction et ne tient pas compte de l'information contenue dans **X** (à la différence des méthodes PLS et N-PLS). Par conséquent, si nous devons estimer les concentrations d'un nouveau mélange, il est alors nécessaire de décomposer entièrement le nouveau tenseur de données. L'avantage de ce type de méthode est qu'elle permet d'obtenir des modèles simples, robustes et facilement interprétables. Un modèle de décomposition PARAFAC d'un tenseur $\underline{\underline{\mathbf{Y}}}$ d'ordre 3 est donné à partir de trois matrices de décomposition : **A**, **B** et **C**. Si F est le nombre de facteurs, le modèle trilinéaire minimise la somme des carrés des résidus e_{ijk} du modèle suivant :

$$y_{ijk} = \sum_{f=1}^{F} a_{ij} b_{jf} c_{kf} + e_{ijk} \quad (14)$$

Un avantage de PARAFAC par rapport aux autres méthodes est l'unicité de sa décomposition. Cette unicité induit que le modèle estimé ne peut subir une rotation sans dégrader l'ajustement entre les données et le modèle.

Afin de pouvoir exprimer le modèle PARAFAC de façon simple, il est nécessaire d'utiliser le produit de *Khatri-Rao*. Ce produit est définit pour deux matrices ayant le même nombre de colonnes F par :

$$C\left|\otimes\right|B = \left[c_1 \otimes b_1, c_2 \otimes b_2, \cdots, c_F \otimes b_F\right] \quad (15)$$

où $\mathbf{c_i}$ et $\mathbf{b_i}$ sont les i èmes colonnes respectivement des matrices **C** et **B**.

En utilisant ce produit, on peut écrire :

$$Y^{(M\times JK)} = A(C\left|\otimes\right|B)^T = \sum_{f=1}^{F} a_f (c_f^T \otimes b_f^T) \quad (16)$$

où $Y^{(M\times JK)}$ correspond aux données 3 dimensions concaténées sous forme d'une matrice de taille (M, JK). **A** est estimée en résolvant la fonctionnelle suivante :

$$\min_{A,B,C} \left\| Y^{(M\times JK)} - A(C\left|\otimes\right|B)^T \right\|_F^{\ 2} \quad (17)$$

Si l'on note $Z = (C\left|\otimes\right|B)$, une estimation de **A** est donnée par:

$$A = Y^{(M\times JK)} Z (Z^T Z)^{-1} \quad (18)$$

De la même façon, **B** et **C** peuvent être déterminées de façon unique. La décomposition PARAFAC peut alors être réalisée en appliquant l'algorithme ALS (Alternating Least Square) aux trois matrices dépliées issues du tenseur des données $\underline{\underline{\mathbf{Y}}}$:

1. Choix du nombre de composants F
2. Initialisation de B et C
3. On pose $Z = (C\left|\otimes\right|B)$ et on cherche **A** tel que: $A = Y^{(M\times JK)} Z (Z^T Z)^{-1}$
4. On pose $Z = (A\left|\otimes\right|C)$ et on cherche **B** tel que : $B = Y^{(J\times MK)} Z (Z^T Z)^{-1}$
5. On pose $Z = (B\left|\otimes\right|A)$ et on cherche **C** tel que: $C = Y^{(K\times MJ)} Z (Z^T Z)^{-1}$
6. Retour à l'étape 3 jusqu'à un meilleur ajustement.

Le temps nécessaire à la convergence de l'algorithme est fortement dépendant des valeurs initiales pour **B** et **C**.

3. Application : test de la robustesse vis-à-vis du décalage des pics chromatographiques

Nous avons testé la robustesse des méthodes Unfold-PLS, N-PLS et PARAFAC vis-à-vis des erreurs de recalage des pics en temps de rétention. Les méthodes classiques

basées sur la pseudo inverse sont très sensibles à cette incertitude.

Les données proposées ont été obtenues à partir de simulations. Les mélanges contiennent deux protéines à différentes concentrations de temps de rétention respectifs 15 et 30 minutes. Les spectrogrammes enregistrés sur 298 points en temps de rétention et 1001 points en masse contiennent deux pics dont la position peut être légèrement différente d'un mélange à l'autre. Cinq mélanges sont disponibles. Pour trois de ces mélanges, les quantités de protéines sont connues et servent à réaliser l'étape d'étalonnage. La protéine 1 et la protéine 2 sont respectivement en quantité (1,2,3) et (4,8,2) fmol dans les 3 mélanges d'étalonnage. Deux autres mélanges « test » sont utilisés pour l'étape de prédiction. Le but est donc d'estimer la quantité de protéines dans les deux mélanges « test » tout en faisant varier la position en temps de rétention des pics associés à ces protéines dans les cinq mélanges. Les variations de position suivent une loi normale de moyenne 15 et 30 min et d'écart type variable entre 0 et 0,5 min. Les résultats présentés correspondent à l'erreur absolue moyenne (400 réalisations) sur l'estimation de la quantité des 2 protéines dans les deux mélanges « test » en fonction d'un facteur proportionnel à l'amplitude des variations de la position des pics par rapport à leur position initiale.

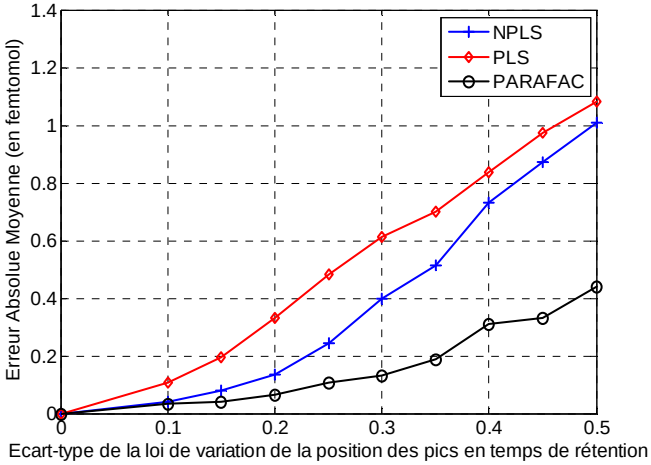

FIG. 2 : Erreur sur l'estimation de la concentration de la protéine 1 dans un des mélanges « test »

La FIG. 2 correspond à l'erreur absolue moyenne pour l'estimation de la quantité de l'une des deux protéines dans un des mélanges « test » (quantité théorique : 5 fmol). Les résultats obtenus montrent la robustesse de PARAFAC par rapport aux méthodes « unfold-PLS » et N-PLS pour de faibles variations de la position des pics. Cette différence peut s'expliquer par le fait que PARAFAC prend en compte simultanément les mélanges d'étalonnage et de prédiction à la différence des autres méthodes pour lesquelles le modèle mathématique est construit seulement à partir des mélanges d'étalonnage. De plus, on constate la supériorité des méthodes N-dimensionnelles (N-PLS et PARAFAC) proposant des modèles simples et robustes et permettant ainsi de réduire le risque de surmodélisation par rapport aux méthodes classiques à 2 dimensions (PLS).

Cependant, sur ce jeu de données simulées, les protéines quantifiées sont les espèces majoritaires du mélange. Dans ce cas, l'intérêt de la méthode N-PLS vis-à-vis de PARAFAC n'est pas démontré. En effet, pour notre application où nous quantifions des protéines en faible concentration, il est nécessaire de prendre en compte à la fois les données mesurées et les concentrations d'entrée à l'étape d'étalonnage. Des tests récents sur des données réelles ont montré la supériorité de N-PLS [10].

Conclusion

Dans ce papier, nous avons comparé la robustesse de différentes méthodes d'analyse factorielle vis-à-vis de l'incertitude sur la position des pics dans le signal chromatographique. Les méthodes testées sont la méthode Unfold–PLS et deux méthodes multidimensionnelles : N-PLS et PARAFAC. Les méthodes N-PLS et PARAFAC s'avèrent être beaucoup plus robustes face à ces variations et permettent ainsi d'avoir une plus grande tolérance sur le recalage des pics.

Remerciements

Ce travail est financée en partie par la commission européenne sous le numéro de contrat FP6/2005/IST/5/034202 associé au projet LOCCANDIA.

Références

1. *LOCCANDIA European Project - http://www.loccandia.eu - FP6/2005/IST/5/034202*
2. Wold, S., *Chemometrics; what do we mean with it, and what do we want from it?* Chemometrics and Intelligent Laboratory Systems, 1995. **30**: p. 109-115.
3. De Lathauwer, L., *Signal Processing based on multilinear algebra*. 1997, K. U. Leuven: Belgium.
4. Hilario, M., et al., *Processing and classification of protein mass spectra.* Mass Spectrom Rev, 2006. **25**(3): p. 409--449.
5. Wold, S., K. Esbensen, and P. Geladi, *Principal component analysis.* Chemometrics and Intelligent Laboratory Systems, 1987. **2**(1-3): p. 37-52.
6. Bro, R., *Multi-way analysis in the food industry: models, algorithms, and applications*. 1998, University of Amsterdam.
7. Bro, R., *Multiway calibration. Multilinear PLS.* Journal of Chemometrics, 1996. **10**(1): p. 47-61.
8. Bro, R., *PARAFAC. Tutorial and applications.* Chemometrics and Intelligent Laboratory Systems, 1997. **38**(2): p. 149-171.
9. Harshman, R.A., *Foundations of the Parafac procedure: models and conditions for an explanatory multimodal factor analysis.* UCLA Working papers in phonetics, 1970. **16**: p. 1-84.
10. Paulus, C., et al. *Chromatographic alignment combined with chemometrics profile reconstruction approaches applied to LC-MS data* in *IEEE Engineering in Medicine and Biology Conference*. 2007. Lyon, France.