Approche bayésienne pour l'analyse de mélanges en spectroscopie

Saïd Moussaoui*, Cédric Carteret, Ali Mohammad-Djafari***, Olivier Caspary*, David Brie*, Bernard Humbert****

** Centre de Recherche en Automatique de Nancy (CRAN), CNRS UMR 7039, Faculté des Sciences et Techniques, B.P. 239, 54506 Vandœuvre-lès-Nancy, France. said.moussaoui@cran.uhp-nancy.fr*
*** Laboratoire de Chimie Physique et Microbiologie pour l'Environnement (LCPME), CNRS UMR 7564, 405, rue de Vandœuvre, 54600 Villers-lès-Nancy, France. cerdric.carteret@lcpe-cnrs.fr*
**** Laboratoire des Signaux et Systèmes (LSS), CNRS-SUPELEC-UPS, Plateau de Moulon, 91192 Gif-sur-Yvette Cedex, France. djafari@lss-supelec.fr*

RÉSUMÉ. Cet article s'intéresse au problème d'analyse de mélanges linéaires en spectroscopie vibrationnelle. L'objectif du traitement est la reconstruction des spectres des composantes pures et l'estimation de leurs concentrations. La méthode proposée consiste en l'utilisation d'une approche bayésienne pour la séparation en modélisant la distribution statistique des spectres et des concentrations par des densités de probabilité gamma afin de prendre en compte la non-négativité. Cette approche est illustrée sur un exemple synthétique.

MOTS-CLÉS : Analyse de mélanges, séparation de sources, estimation bayésienne, non-négativité.

1. Introduction

Le modèle de mélange en spectroscopie [MAL 02] suppose l'existence de *n composantes pures* dont on observe *m mélanges linéaires*. L'analyse du mélange consiste en la reconstruction des spectres des composantes pures et l'estimation de leurs concentrations dans chaque mélange observé. En traitement du signal, ce problème est connu sous le nom de *séparation de sources* [MOH 99] où les *sources*, dans le cas de la spectroscopie, correspondraient aux composantes pures. Le modèle peut donc s'écrire sous la forme :

$$D = C\,S + E$$

où les données mesurées sont représentées dans la matrice D, de dimension $m \times N$, où m est le nombre de spectres mesurés et N le nombre de variables d'observation disponibles (nombre d'onde, fréquence, etc.) pour chaque mesure, les spectres des composantes pures sont contenues dans la matrice S, de dimension $m \times N$ où n est le nombre de composantes pures, C représente la matrice de mélange, de dimension $m \times n$, et dont chaque coefficient C_{ij} correspond à la concentration de la source S_{jk} dans la mesure D_{ik}. E est une matrice de dimension $m \times N$, dont les composantes E_{ik} correspondent aux erreurs de mesures et incertitudes du modèle. D'une façon plus précise le problème de séparation de sources ou encore d'analyse de mélanges peut être formulé de la façon suivante : à partir des données D et en utilisant toutes les informations disponibles sur les signaux sources et le mélange, retrouver les spectres S des composantes pures et leurs concentrations C.

Les méthodes d'analyse de mélanges existantes en chimiométrie se basent d'abord sur une première étape de décomposition par différentes méthodes algébriques (SIMPLISMA [WIN 91], OPA [CUE 96], PCA[BU 00]) ensuite une procédure de moindres carrés alternés (ALS [TAU 95]) est appliquée afin d'imposer la contrainte de non-négativité sur les spectres et des concentrations. Au lieu d'utiliser ces deux étapes, nous proposons dans cet article une méthode issue de l'approche inférence bayésienne dont l'avantage principal réside dans la possibilité d'incorporer des connaissances *a priori* sur les variables à estimer (sources et concentrations). Dans cette première approche, nous utilisons comme *a priori* la non-négativité des spectres et des concentrations.

3. Méthode proposée

Afin de rappeler le principe de l'estimation bayésienne, considérons la résolution d'un problème du type $y = f(x) + b$ où on cherche à retrouver x à partir de la mesure de y et connaissant la fonction $f(x)$. b étant un bruit additif représentant les erreurs de mesure et incertitudes du modèle. Le principe de base de l'estimation bayésienne consiste d'abord en la formulation de l'expression de la densité de probabilité *a posteriori* de la variable à estimer, en utilisant le théorème de Bayes, à partir du modèle de mesure (qui donne la *vraisemblance*) et de la modélisation statistique de la variable recherchée par une densité de probabilité *a priori* qui permet de coder l'information disponible sur x. On écrit donc :

$$\underset{\text{densité a posteriori}}{p(x/y)} = \underset{\text{vraisemblance}}{p(y/x)} \times \underset{\text{densité a priori}}{p(x)} / \underset{\text{constante}}{p(y)}$$

La solution par la méthode du maximum *a posteriori* (MAP) s'obtient par la valeur de x qui maximise la densité de probabilité *a posteriori*, c'est-à-dire, la valeur la plus probable au vue des donnés mesurées et compte tenu des connaissances disponibles *a priori*. En pratique, au lieu de maximiser cette loi de probabilité, on cherche à minimiser un critère $J(x) = -\log p(x/y)$. On note que si l'on n'utilise pas de densité *a priori* explicite sur la variable recherchée, on revient à l'estimateur du maximum de vraisemblance qui dans certains cas donne la même solution que la méthode des moindres carrés. Dans la suite de cette section, nous explicitons la démarche suivie pour l'analyse de mélanges en utilisant cette approche bayésienne.

3.1. Modélisation statistique des spectres

La solution retenue pour représenter la non-négativité consiste en l'utilisation de distributions Gamma pour modéliser la distribution statistique de chaque spectre pur et sa concentration dans le mélange. Nous rappelons l'expression et les propriétés d'une loi Gamma

$$p(x/\alpha,\beta) = \frac{\beta^{\alpha}}{\Gamma(\alpha)} x^{\alpha-1} \exp(-\beta x) \qquad pour \quad (\alpha,\beta) > 0 \quad et \quad x \geq 0$$

où $\Gamma(\alpha)$ représente la fonction Gamma.

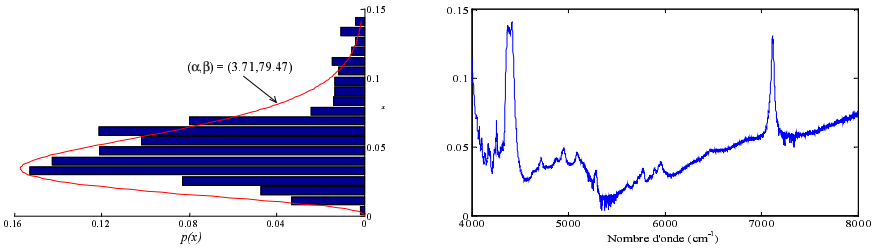

Figure1. *Exemple de modélisation de la distribution statistique d'un spectre par une densité Gamma*

La moyenne d'une variable distribuée suivant une loi Gamma de paramètre (α,β) vaut (α/β) et la variance vaut (α/β^2), Donc à partir de la moyenne et de la variance d'une variable distribuée selon une loi Gamma, on peut obtenir une estimée de ses paramètres (α,β) par la *méthode des moments* [MAR 03].

3.2. Modélisation du bruit

Le bruit est modélisé par une distribution Gaussienne de moyenne nulle et de variance σ_e^2.

$$p(e_{jk} / \sigma_e^2) = \frac{1}{\sqrt{2\pi}\sigma_e} \exp\left(-\frac{e_{jk}^2}{2\sigma_e^2}\right) \quad pour\ i = 1,\ldots,m \quad et \quad k = 1,\ldots,N.$$

3.3. Critère résultant

L'estimateur du MAP conduit à la minimisation d'un critère $J(S,C)$ donné par :

$$J(S,C) = -\log p(S,C/D) = \Phi_L(S,C) + \Phi_{P1}(S) + \Phi_{P2}(C).$$

Le critère $\Phi_L(S,C) = \frac{1}{2\sigma_b^2}\sum_{k=1}^{N}\sum_{i=1}^{m}\left[(D_{ik} - [CS]_{ik})^2\right]$ correspond au critère des moindres carrés classique et les critères $\Phi_{P1}(S) = \sum_{j=1}^{n}\left[(1-\alpha_j)\sum_{k=1}^{N}\log S_{jk} + \beta_j\sum_{k=1}^{N}s_{jk}\right]$ et $\Phi_{P2}(C) = \sum_{j=1}^{n}\left[(1-\lambda_j)\sum_{i=1}^{m}\log C_{ij} + \gamma_j\sum_{i=1}^{m}C_{i,j}\right]$ sont des termes de régularisation qui permettent de pénaliser les valeurs négatives des spectres et concentrations. Les coefficients $\{\alpha_j, \beta_j, \lambda_j, \gamma_j$ pour $j=1,...,n\}$ sont les paramètres des lois Gamma associées aux sources et concentrations dont l'estimation est effectuée par la méthode des moments. Il faut noter que la forme du critère présente une similitude avec celui présenté par Paatero [PAA 97] dans le cadre de la méthode PMF (*Positive Matrix Factorization*). Néanmoins, nous pouvons constater des différences, notamment le fait que les termes de régularisation $\mathbf{\Phi}_{P1}$ et $\mathbf{\Phi}_{P2}$ résultent d'une modélisation statistique des sources et concentrations. De plus, les poids associés à ces termes peuvent être différents pour les différentes sources et concentrations et s'obtiennent d'une façon non-supervisée (automatique) .

La minimisation de ce critère est effectuée par une optimisation conjointe par rapport aux sources et ensuite par rapport aux concentrations, en utilisant l'algorithme du gradient :

$$\begin{cases} S^{(r)} = S^{(r-1)} - \mu_S{}^{(r)}\,\nabla_S J\left(S^{(r-1)}, C^{(r-1)}\right); \\ C^{(r)} = C^{(r-1)} - \mu_C{}^{(r)}\,\nabla_C J\left(S^{(r)}, C^{(r-1)}\right); \end{cases} \quad pour \;\; r = 1,\ldots,r_{\max},$$

où μ_S et μ_C sont des pas d'adaptation positifs qui contrôlent la convergence de l'algorithme du gradient. Ces pas sont choisis à chaque itération de telle sorte à minimiser le critère global. $r_{\max}$ est le nombre maximal d'itérations nécessaires pour la convergence. A chaque itération r de l'algorithme d'optimisation conjointe, les paramètres des distributions Gamma sont estimés par la méthode des moments et la variance du bruit par la méthode du maximum de vraisemblance. Par la suite nous faisons référence à cet algorithme par la dénomination BPSS pour *Bayesian Positive Source Separation.*

4. Application à un mélange synthétique

Afin de tester la méthode proposée, nous avons synthétisé un mélange à partir de deux composantes pures dont les spectres purs sont connus (figure 2). Les concentrations sont choisies de façon à avoir un mélange de dix mesures (m = 10). Le nombre de point d'observation N = 1000. L'analyse est effectuée avec les méthodes BPSS, SIMPLISMA et ALS, afin de comparer leurs performances. La figure (3) montre que la méthode présentée permet d'estimer correctement les composantes pures et leurs concentrations et ne présente pas de composantes négatives, ce qui n'est pas le cas avec la méthode SIMPLISMA, qui necessite ainsi un post-traitement par ALS pour éliminer les composantes négatives. Par ailleurs, une comparaison des temps de traitement sur un même calculateur montre qu'à résultats équivalents, BPSS nécessite un temps de calcul 10 fois moindre que SIMPLISMA-ALS.

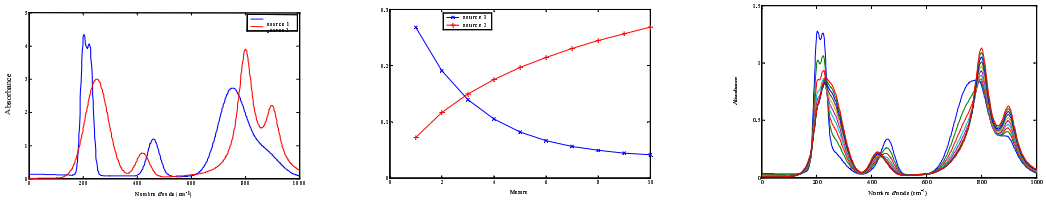

Figure 2. *Synthèse du mélange*

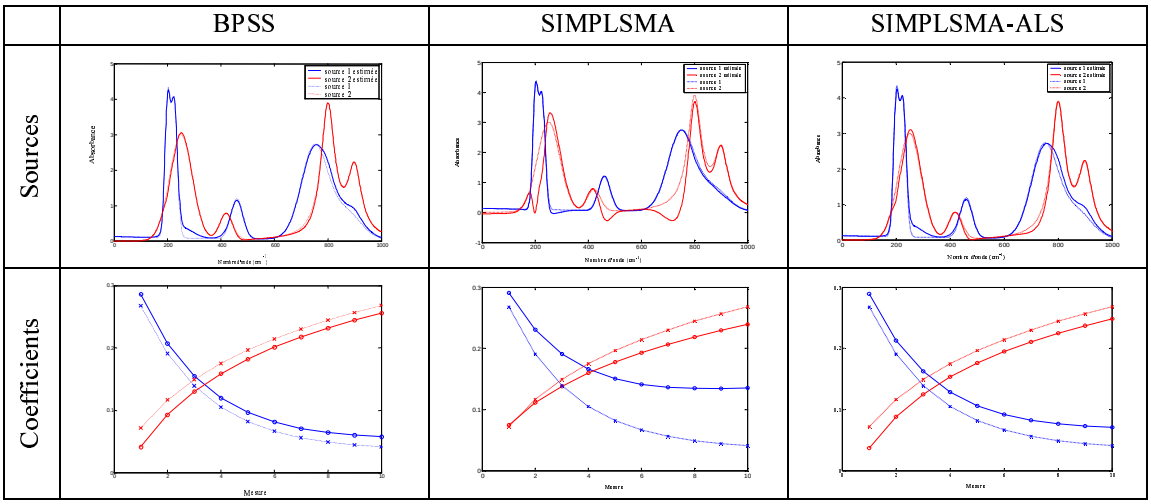

Figure 3. *Comparaison entre les estimées (continu) et les valeurs vraies (discontinu)*

5. Conclusion

Il est clair qu'il est prématuré de généraliser des conclusions à partir d'un seul mélange, car il reste à étudier la convergence et l'efficacité de la méthode dans un contexte plus difficile. Cependant, ce résultat illustre que l'approche proposée est une direction intéressante pour l'analyse de mélanges en spectroscopie. En effet l'approche inférence bayésienne offre la possibilité d'utiliser, en plus des données mesurées, les différentes connaissances théoriques comme informations *a priori* permettant d'adapter la méthode de traitement au contexte d'analyse du mélange et aux propriétés des signaux d'intérêt (unimodalité des spectres, profil d'évolution des concentrations, type de perturbations, etc.).

6. Bibliographie

[BU 00] BU D., BROWN C. W., "Self-modeling mixture analysis by interactive principal component analysis ", *Applied Spectroscopy,* vol. 54, 2000, p. 1241-1221.

[CUE 96] CUESTA SANCHEZ F., TOFT J., VAN DEN BOGAERT B., MASSART D. L., " Orthogonal projection method applied to peak purity assessment ", *Annals of Chemometrics.* Vol. 68, 1996, p. 79-85.

[MAL 02] MALINOWSKI E. R., *Factor Analysis in chemistry*, Willey Interscience. New York, 3rd ed. 2002.

[MAR 02] MARTINEZ W. L., MARTINEZ A. R., *Computational statistics handbook with Matlab,* Chapman & Hall. 2002.

[MOH 99] MOHAMMAD-DJAFARI A., " A Bayesian approach for source separation ", *19th International Workshop on Maximum Entropy and Bayesian Methods (MaxEnt 99),* Boise, Idaho, USA, Aug.4-8, 1999.

[PAA 97] PAATERO P., " Least square formulation of robust non-negative factor analysis ", C*hemometrics and Intelligent. Laboratory Systems,* vol. 37, 1997, p. 23-35.

[TAU 95] TAULER R., SMILDE A., KOWALSKI B., " Selectivity, local rank, three-way data analysis and ambiguity in multivariate curve resolution ", *Journal of Chemometrics,*vol. 9, 1995, p. 31-58.

[WIN 91] WINDIG W., GUILMENT J., "Interactive self-modeling mixture analysis ", *Annals of Chemometrics,* vol.63, 1991, p. 1425-1432.