Inférence bayésienne non paramétrique pour l'analyse chromatographique de polluants dans l'eau

Olivier HARANT, Louise FOAN, François BERTHOLON, Séverine VIGNOUD, Pierre GRANGEAT

Université Grenoble Alpes, F-38000 Grenoble, France,
CEA, Leti, MINATEC Campus, 17 rue des martyrs, F-38054 Grenoble Cedex 9 France
prénom.nom@cea.fr

Résumé – Les modèles bayésiens non-paramétriques et en particulier les modèles de mélange de processus de Dirichlet offrent une grande souplesse pour l'inférence sur des mélanges de distributions dont le nombre de composantes est inconnu. Nous présentons ici une méthode d'inférence sur la concentration d'un nombre inconnu de polluants dans l'eau à partir de signaux chromatographiques abordés comme une loi de distribution sur les temps de rétention d'un ensemble de molécules. Des résultats sur des mélanges d'Hydrocarbures Aromatiques Polycycliques (HAP) à différentes concentrations sont présentés.

Abstract – Non-parametric Bayesian Models and more specificaly Dirichlet Process Mixture Models are quite powerfull for infering on mixtures of distributions where the number of components is unknown. This paper deals with a method for infering on the concentration of an unknown number of contaminants in water from chromatographic signals described as a distribution of retention time on a set of molecules. Some promising results are finaly shown on various Polycyclic Aromatic Hydracarbons (PAH) mixtures.

1 Introduction

De nombreux dispositifs analytiques fournissent des signaux de sortie de type histogramme de *temps d'arrivée* de particules, de molécules ou d'événements. Ils représentent généralement une succession de pics dont la forme caractérise la distribution non normalisée des temps d'arrivée et dont l'aire quantifie les particules, molécules ou événement associés qu'on cherche à identifier et quantifier. Les signaux de chromatographie, processus de chimie analytique de séparation des analytes d'un mélange liquide ou gazeux, font partie de cette famille de signaux. Le temps de parcours de la colonne chromatographique par les molécules, appelé *temps de rétention*, dépend de la nature des molécules, ce qui permet de les séparer.

Ce papier aborde la problématique de quantification d'analytes d'un mélange d'un point de vue microscopique afin de mieux caractériser les interactions moléculaires. L'approche proposée vise à décomposer le signal observé en une somme de signaux élémentaires caractérisant chacun un analyte. Des méthodes de séparation de source ont été appliquées en chimie analytique mais elles nécessitent d'avoir à disposition plusieurs signaux et de connaître le nombre de composantes [1]. Ceci n'est pas envisageable dans le contexte de notre étude d'échantillons d'eau prélevés dans le milieu naturel. La reconstruction de profils protéomiques a également été mis en œuvre dans un cadre bayésien paramétrique [2] avec un nombre de pics connu. Nous proposons ici d'introduire les méthodes bayésiennes non paramétriques basées sur les modèles de mélanges de processus de Dirichlet (DPMM) [3] pour estimer les proportions de toutes les composantes d'un mélange sans en connaître le nombre et à partir d'un unique signal chromatographique abordé d'un point de vue microscopique.

La section suivante introduit le modèle moléculaire d'une colonne de chromatographie qui justifie le modèle Normal utilisé dans l'inférence sur DPMM décrite dans la troisième section. La quatrième section présente des résultats d'estimation non paramétrique de concentrations de polluants à différentes dilutions.

2 Modèle microscopique de chromatographie

2.1 Marche aléatoire

Une molécule parcourant la colonne de chromatographie subit une succession d'états désorbés et adsorbés. Les temps passés dans chacun de ces états sont des variables aléatoires qui, dans des conditions expérimentales données, dépendent de la nature de la molécule. Le temps de rétention τ d'une famille de molécules est une variable aléatoire définie comme la somme de tous les temps des états adsorbés et désorbés avant sa sortie de colonne. Il est distribué selon une famille de lois dont les paramètres dépendent de la nature de la molécule. Un chromatogramme représente l'histogramme des temps de rétention des molécules du mélange. Différents modèles microscopiques ont été proposés pour modéliser la loi de τ [4]. Mais dans notre cas où les temps de rétention sont supérieurs à la minute, ces modèles convergent vers la loi normale qui sera utilisée ici. Nous

proposons de reformuler l'analyse du signal comme une problématique de classification d'un ensemble de N molécules dont les temps de rétentions $\boldsymbol{\tau}$, tirés selon le signal chromatographique $y(\tau)$ normalisé, sont représentatives du mélange étudié.

2.2 Modèle de mélange

Soit $\boldsymbol{\tau} = \{\tau_1, \cdots, \tau_N\}$ une série d'observations des temps de rétentions de N molécules indépendantes échantillonnées par la méthode de la transformée inverse à partir du signal chromatographique $y(\tau)$. N est choisi arbitrairement. Le dit signal normalisé s'écrit comme la loi de mélange inconnue suivante :

$$\frac{y(\tau)}{\int y(t)\mathrm{d}t} = p(\tau|\mathbf{c}, \boldsymbol{\Theta}) = \sum_{k=1}^{K} c_k p_\tau(\tau|\boldsymbol{\theta}_k^*), \tag{1}$$

où $\boldsymbol{\Theta} = [\boldsymbol{\theta}_1^*, \cdots, \boldsymbol{\theta}_K^*]^T$ est le vecteur des paramètres des K composantes du mélange, K étant inconnu. $\boldsymbol{\theta}_k^* = (\mu_k^*, {\sigma^2}_k^*)$ dans le cas du modèle gaussien. $\mathbf{c} = [c_1, \cdots, c_K]^T$ est le vecteur des proportions qui satisfait $\sum_{k=1}^{K} c_k = 1,\ c_k > 0$.

3 Inférence

Nous cherchons à estimer les proportions $\mathbf{c}$ des analytes présents dans le mélange. Cette estimation se fait indirectement par l'inférence du vecteur d'index de classe $\hat{\mathbf{z}} = [\hat{z}_1, \cdots, \hat{z}_N]^T$ qui décrit à quelle classe appartient chaque molécule. L'écriture du modèle de mélange (1) comme un mélange de processus de Dirichlet permet de formaliser cette inférence par un échantillonnage de Gibbs du vecteur $\mathbf{z}$.

3.1 Les modèles de mélanges de Processus de Dirichlet

Dans le cas gaussien, la loi Normale-Inverse-Wishart est bien adaptée comme loi *a priori* G_0 du vecteur de paramètres $\boldsymbol{\theta}_k^*$. Chaque τ_i est tiré selon la loi d'une composante k du mélange : $\tau_i \sim p_\tau(\tau_i|\boldsymbol{\theta}_i)$ où $\boldsymbol{\theta}_i \in \{\boldsymbol{\theta}_1^*, \cdots, \boldsymbol{\theta}_K^*\}$. $\boldsymbol{\theta}_i$ prend la valeur $\boldsymbol{\theta}_k^*$ avec la probabilité c_k :

$$\boldsymbol{\theta}_i \sim G \triangleq \sum_{k=1}^{K} c_k \delta(\boldsymbol{\theta}_k^*),$$

où $\delta(\,\cdot\,)$ est la distribution de Dirac. La distribution de G, discrète, est une distribution sur les distributions appelée *Processus de Dirichlet* (DP) à partir de laquelle les DPMM sont définis par [5] :

$$\begin{aligned} \tau_i|\boldsymbol{\theta}_i &\sim p_\tau(\tau_i|\boldsymbol{\theta}_i) \\ \boldsymbol{\theta}_i|G &\sim G \\ G &\sim DP(\alpha, G_0) \end{aligned} \tag{2}$$

où α est un hyper paramètre qui influence l'espérance du nombre de classes K qui pour un nombre N d'observations grand est $E[K|\alpha, N] \approx \alpha \log(1 + N/\alpha)$. Les temps de rétention observés sont donc modélisables par des DPMM dont l'inférence sur leurs paramètres est détaillée dans la section suivante.

3.2 Collapsed Gibbs Sampling

Nous décrivons ici une boucle de Gibbs particulière appelée *Collapsed Gibbs Sampling* (CGS) qui permet d'échantillonner directement $\mathbf{z}$ [5]. La loi *a posteriori* de z_i s'écrit :

$$p(z_i = k|\mathbf{z}_{-i}, \boldsymbol{\tau}, \alpha, G_0) \propto p(z_i = k|\mathbf{z}_{-i}, \alpha) p(\tau_i|\boldsymbol{\tau}_{k,-i}, G_0), \tag{3}$$

où $\tau_{k,-i}$ est l'ensemble des observations de classe k sans l'observation i et $\mathbf{z}_{-i}$ est le vecteur des indices de classe des observations autre que l'observation i.

Le **premier terme** correspond à la probabilité d'appartenance à une des $K+1$ classes :

$$p(z_i = k|\mathbf{z}_{-i}, \alpha) = \begin{cases} \frac{N_k}{\alpha + N - 1} & \text{si } k \in [1, K], \\ \frac{\alpha}{\alpha + N - 1} & \text{si } k = K + 1. \end{cases}$$

où N_k est le nombre d'observations de la classe k.

Le **second terme** $p(\tau_i|\boldsymbol{\tau}_{k,-i}, G_0)$ décrit la probabilité *a posteriori* d'une observation. En choisissant des lois *a priori* $p(\mathbf{c}|\alpha)$ et $p(\boldsymbol{\Theta}|G_0)$ conjuguées, on montre que $\mathbf{c}$ et $\boldsymbol{\Theta}$ peuvent être marginalisés et ce terme calculé notamment dans le cas d'une vraisemblance gaussienne des observations [6, p. 135]. Un tirage multinomial paramétré par les probabilités (3) ainsi calculées permet finalement d'attribuer l'index de classe z_i.

L'hyper paramètre α influençant directement le nombre de classes K, nous avons intégré son inférence dans la boucle de Gibbs [7]. Le CGS est initialisé par un Processus du Restaurant Chinois [6, p. 884]. Après $B-1$ itérations de chauffe, la proportion de chaque classe $c_k^{(g)}$ par rapport aux N molécules du mélange est estimée à chaque itération g à partir des échantillons $\mathbf{z}^{(g)}$ d'index de classe comme $c_k^{(g)} = \frac{1}{N} \sum_{i=1}^{N} \mathbb{1}_{z_i^{(g)}=k}$ où $\mathbb{1}$ est la fonction indicatrice.

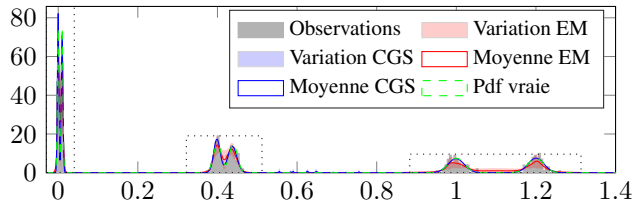
La proportion $\hat{c}_k$ est estimée comme la moyenne *a posteriori* des proportions $\hat{c}_k = \frac{1}{G-B+1} \sum_{g=B}^{G} c_k^{(g)}$. Cet estimateur est sensible à la présence de *permutations d'index* mais elles sont négligeables dans nos exemples. Des méthodes plus robustes d'estimation *a posteriori* de classification à partir des échantillons de Gibbs sont détaillées dans [8].

4 Résultats

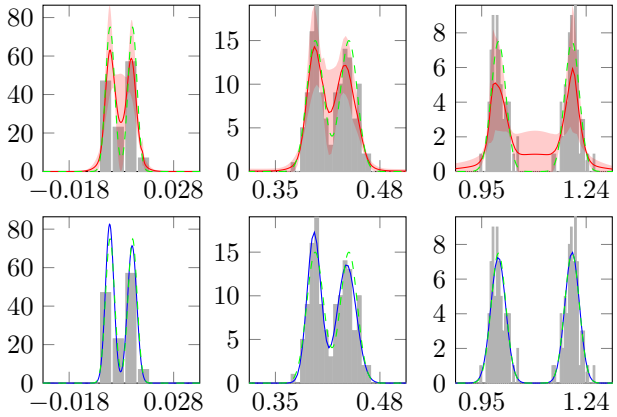
4.1 Sur chromatogrammes simulés

$N = 400$ échantillons de temps de rétention ont été tirés selon un mélange de six lois normales de moyennes et variances représentatives d'un chromatogramme réel i.e. la variance augmente avec le temps de rétention moyen et les premiers pics peuvent être très proches. L'histogramme de ce mélange de gaussiennes est illustré sur la Figure 1. La moyenne de 100 estimations de la loi de mélange (1) et sa variance y sont également représentées dans les cas d'une estimation par espérance a posteriori des échantillons de CGS et d'autre part par estimation de type EM (Expectation Maximisation) initialisée avec $K = 6$. L'algorithme EM est instable et ne distingue parfois pas certains pics comme en témoigne la plage de variation importante

des distributions de la Figure 1b du haut. Au delà des itérations de chauffe, les échantillons de Gibbs ont une variance très faible et un biais limité comme en témoigne la quasi absence de variation sur la Figure 1b du bas. On peut remarquer sur la Figure 1a des classes de très faible proportion parfois détectées par l'algorithme non paramétrique autour de la moyenne 0.6.



(a) Histogramme des données, moyennes et plages de variation à $\pm\sigma$ des densités de probabilités du mélange estimées.



(b) Zoom sur les pics d'intérêt. (haut) EM, (bas) CGS.

FIGURE 1 – Estimations des paramètres d'un mélange de 6 gaussiennes issues de 100 exécutions d'EM et de CGS.

4.2 Sur chromatogrammes réels

Les teneurs en HAP dans l'eau peuvent être déterminées grâce à l'extraction des analytes dans un solvant approprié à l'analyse chromatographique. Nous avons analysé par chromatographie en phase gazeuse couplée à un détecteur par ionisation de flamme (GC FID) des solutions de 5 HAP dans le méthanol contenant les composés suivants : acenaphtène (ACE), anthracène (ANT), fluoranthène (FTN), benzo(a)pyrène (B(a)P) et indéno(1,2,3-cd)pyrène (IND). Afin de souligner la robustesse de la méthode dans deux cas de rapport signal à bruit extrêmes, nous présentons ici les résultats de l'inférence sur deux de ces solutions à 1 et $100\mu g/mL$ (Figure 2a). Les analyses ont été réalisées avec une colonne 5MS ($30m$, $0.25mm$, $0.25\mu m$) soumise à une pression de $12psi$ ($1mL/min$) et un gradient de température de $5°C/min$ de $50°C$ à $300°C$. L'injection de $0.5\mu L$ de solvant est exécutée à $250°C$ en mode splitless pulsé ($40psi$ sur $12s$). Le FID est chauffé à $300°C$. Dans les résultats suivants, les proportions prises comme références sont les aires mesurées par un expert sous chacun des pics du signal sans ligne de base.

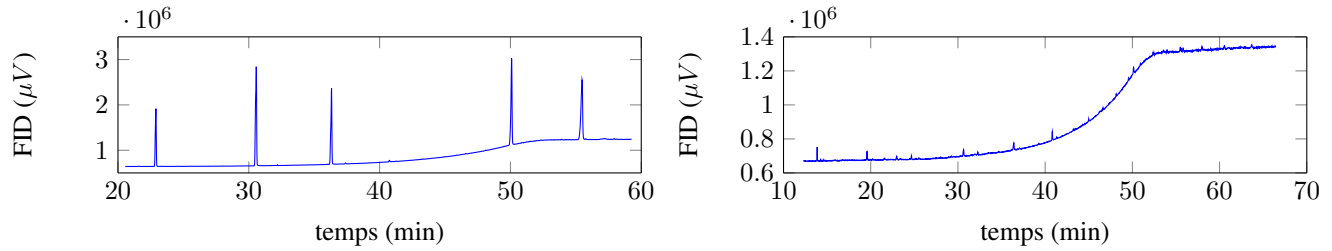
Les boites à moustache de la Figure 2c représentent les distributions des proportions de chaque classe $c_k^{(g)}$ des itérations de Gibbs relativement au nombre total de molécules de HAP détectées. Les valeurs médianes des concentrations estimées sont remarquablement proches des valeurs de référence. Les classes d'intérêt, i.e. des HAP, sont identifiées comme celles dont les moments des temps de rétention $\boldsymbol{\tau}_k$ correspondants sont les plus proches de ceux de leur loi *a priori* définis par le modèle de marche aléatoire [4].

Sur le signal chromatographique du mélange des 5 HAP à $100\mu g/mL$, le rapport signal sur bruit est élevé et la distinction des pics est évidente (Figure 2a de gauche). L'échantillonnage CGS converge très rapidement vers la loi stationnaire *a posteriori* de $\mathbf{z}$ comme en témoigne l'évolution de la log-vraisemblance des observations sur la Figure 2b. Dans ce cas les pics correspondants aux 5 HAP ont correctement été segmentés. Notons que le signal est complètement décrit par l'expression (1) qui intègre donc la ligne de base. Cela se retrouve sur la Figure 2c de gauche où l'interférent I_2 dont la proportion est élevée et les interférents I_1 et I_4 de variances élevées (resp. σ^2 =), font référence à la ligne de base. L'interférent I_3 correspond lui à un pic inconnu bien marqué.

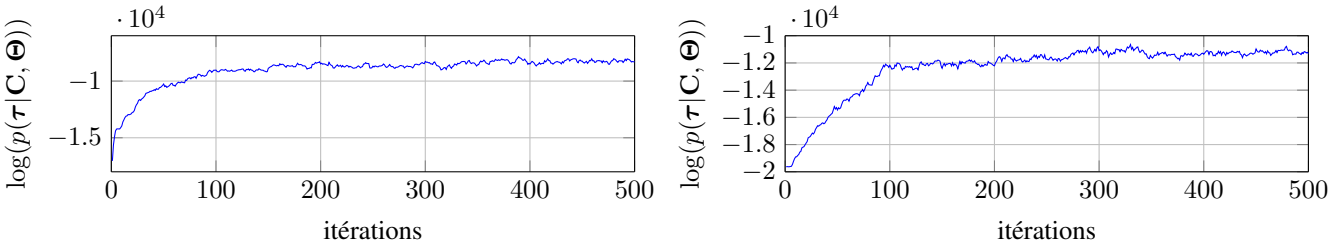
Avec une concentration des HAP de l'ordre de $1\mu g/mL$, les pics d'intérêt ressortent peu d'une ligne de base prépondérante et des pics d'interférents divers (Figure 2a de droite) : nous sommes proches de la limite de détection. Pour permettre une inférence avec un nombre d'échantillons N limité, un filtrage grossier de type moyenne glissante a été fait pour soustraire la ligne de base *basse fréquence* au signal. Ce traitement a pour conséquence d'atténuer voire de supprimer des termes de la distribution du mélange (1) correspondant à la ligne de base mais n'influence pas les termes correspondants aux pics. Dans ce cas de figure, l'inférence converge vers une trentaine de composantes dont certaines de variance élevée. Ceci traduit une distribution vague associée à la partie uniforme du bruit de fond (la ligne de base résiduelle). Les autres composantes identifient en revanche la plupart des pics présents sur le signal. La Figure 2c illustre les distributions des proportions des composantes de variances inférieures à 1. Celles-ci rassemblent les HAP d'intérêt mais aussi quelques interférents I. Le nombre N limité n'a ici pas permis la détection de certains pics comme l'ACE qui est assimilé au bruit de fond et n'est pas quantifié. Ceci justifie la plus faible valeur moyenne de la vraisemblance des observations après convergence par rapport au cas du mélange à $100\mu g/mL$.

5 Conclusion

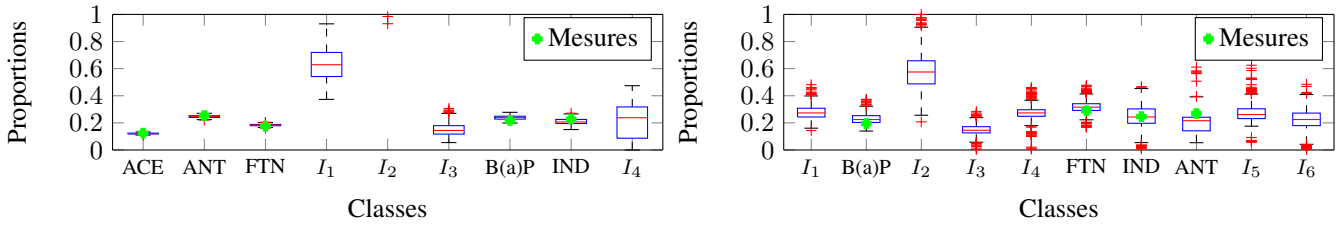
Nous avons proposé de reformuler l'analyse de signaux chromatographiques comme une problématique de décomposition d'une loi de distribution sur les temps de rétention de molécules en une somme de lois de distributions associées à chaque pic chromatographique. Ce point de vue associé à une approche



(a) Signaux chromatographiques $y(\tau)$.



(b) Évolution de la log-vraisemblance des observations $\boldsymbol{\tau}$.



(c) Boîtes à moustaches des proportions estimées par rapport à la quantité totale des classes de HAP détectés et mesures de référence.

FIGURE 2 – Inférence sur 5000 molécules d'une solution de HAP à $100\mu g/mL$ (gauche) et $1\mu g/mL$ (droite).

bayésienne non-paramétrique a permis d'estimer la concentration de mélanges de polluants dans un solvant à partir d'un seul signal chromatographique et sans hypothèse sur le nombre d'analytes. Cette approche trouve tout son intérêt dans un contexte d'analyse en environnement ouvert sur des mélanges non contrôlés tels que l'eau dans un milieu naturel. La justesse des estimations par rapport aux valeurs de référence est encourageante en particulier pour des mélanges à faibles concentrations, proches de la limite de détection. Cette méthode permet en outre une estimation jointe de l'ensemble des composantes du signal, y compris la ligne de base. Ceci rend superflu l'étape de prétraitement pour soustraire la ligne de base ou la limite à un prétraitement grossier.

Les résultats présentés ici s'appuient sur un modèle gaussien simple de la forme des pics chromatographiques. Cette première approximation offre des résultats satisfaisants et réduit grandement la complexité mathématique. Toutefois il est intéressant d'élargir cette approche aux modèles de temps de rétention plus complexes qui prennent en compte la diffusion au sein de la colonne [4] et son conditionnement en température.

La souplesse d'une telle approche offre un champ applicatif couvrant tous les signaux de type *histogrammes de temps d'arrivée* tels que ceux issus de la spectrométrie de masse. Enfin cette approche univariée est généralisable aux signaux multidimensionnels tels que les signaux de chromatographie en phase gazeuse couplée à la spectrométrie de masse.

Références

[1] L. Duarte, S. Moussaoui, and C. Jutten, "Source separation in chemical analysis : Recent achievements and perspectives," *Signal Processing Magazine, IEEE*, vol. 31, pp. 135–146, May 2014.

[2] P. Szacherski, J.-F. Giovannelli, L. Gerfault, P. Mahe, J.-P. Charrier, A. Giremus, B. Lacroix, and P. Grangeat, "Classification of proteomic ms data as bayesian solution of an inverse problem," *Access, IEEE*, vol. 2, pp. 1248–1262, 2014.

[3] T. S. Ferguson, "A bayesian analysis of some nonparametric problems," *The annals of statistics*, pp. 209–230, 1973.

[4] A. Felinger, "Molecular dynamic theories in chromatography," *Journal of Chromatography A*, vol. 1184, no. 1-2, pp. 20 – 41, 2008. 50 Years Journal of Chromatography.

[5] R. M. Neal, "Markov chain sampling methods for dirichlet process mixture models," *Journal of Computational and Graphical Statistics*, vol. 9, no. 2, pp. 249–265, 2000.

[6] K. P. Murphy, *Machine learning : a probabilistic perspective*. MIT press, 2012.

[7] M. D. Escobar and M. West, "Bayesian density estimation and inference using mixtures," *Journal of the american statistical association*, vol. 90, no. 430, pp. 577–588, 1995.

[8] A. Fritsch and K. Ickstadt, "Improved criteria for clustering based on the posterior similarity matrix," *Bayesian Anal.*, vol. 4, pp. 367–391, 06 2009.