CHROMATOGRAPHIC SIGNAL PROCESSING FOR PAH IN METHANOL SOLUTION

François BERTHOLON, Olivier HARANT*, Louise FOAN*, Séverine VIGNOUD*, Christian JUTTEN†, and Pierre GRANGEAT**
*Université Grenoble Alpes, F-38000 Grenoble, France,
CEA, Leti, MINATEC Campus, 17 Rue des martyrs F-38054 Grenoble Cedex 9, France
†Université Grenoble Alpes, F-38000 Grenoble, France,
Grenoble Image Parole Signal Automatique lab, GIPSA lab UMR 5216 CNRS, 11 Rue des Mathématiques, 38400 Saint-Martin-d'Hères, France

ABSTRACT

In this paper we describe two methods to estimate the concentration of polycyclic aromatic hydrocarbons (PAHs) in a methanol solution, from a gas chromatography analysis. We present an innovative stochastic forward model based on a molecular random walk. To infer on PAHs concentration profiles, we use two inversion methods. The first one is a Bayesian estimator using a MCMC algorithm and Gibbs sampling. The second one is a sparse representation method with non-negativity constraint on the mixture vector based on the decomposition of the signal on a dictionary of chromatographic impulse response functions as defined by the forward model. Some results provided by those two methods are finally shown with a comparison of the computational and the quantification performances.

Index Terms— Gas chromatography, Bayesian estimation, Monte Carlo Markov Chain (MCMC), Sparse Representation, Dictionary, FOCUSS Algorithm.

1. INTRODUCTION

Analysis mixture is nowadays essential in pollutant detection and quantification for instance to monitor the air we breathe or the water we drink. One of the usual processes for this analysis is to separate gas. In this paper the separation of components of a gas mixture is done using a chromatographic system. Our interest is focused on Polycyclic Aromatic Hydrocarbons (PAHs) in methanol solvent. In this communication, we present a method to infer jointly on the concentration of each PAH and on other unknown parameters. In the next section, the Gas Chromatography (GC) - Flame Ionization Detector (FID) system is described. Then, we introduce two inversion schemes to retrieve the concentration of each component of the mixture from the chromatographic signal. We use a stochastic forward model based on the molecular random walk principle as described by Giddings and Eyring [1]. We introduce two methods to invert this model. The first one is a Bayesian parameters estimation scheme based on a microscopic model. The second one is a sparse representation method based on a dictionary of macroscopic parametric chromatographic responses built from Gidding and Eyring's model. Finally, we compare those two methods, in terms of computational time and quantification performances on experimental data.

2. THE GC-FID SYSTEM DESCRIPTION

The gas chromatography system used is composed of an injector, a 30 meters length 5MS chromatography column and a sensor to acquire a signal. In our study, the sensor is a flame ionization detector. As illustrated in Figure 1, the injector pushes a gas mixture thanks to a carrier gas into the column. It also helps to volatilize the solvent. The internal surface of the column is coated with a layer called *stationary phase* where molecules of organic compounds are adsorbed for a random time. The molecules are carried within the column by a carrier gas called *mobile phase*. The carrier gas, usually helium, has no interaction with the stationary phase. During the run through the column, a molecule will undergo a serie of adsorption and desorption steps. The number of these steps is depending on the nature of the molecule. The more it is adsorbed on the stationary phase, the more this molecule spends time in the column. This affinity rate with the stationary phase defines the total time t_R spent by a molecule in the column before elution, called retention time. This contributes to the separation power of the column. The adjusted retention time is defined as the total adsorbed time of a molecule. For sake of simplicity, we consider the retention time t as the adjusted retention time. It is also used for gas chromatography studies.

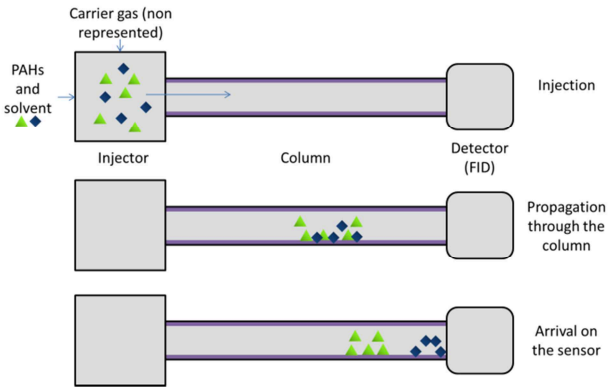

Figure 1: Principle of the chromatographic system

978-0-9928626-3-3/15/$31.00 ©2015 IEEE

Once the molecules exit the column, they arrive onto the FID where they are burned. The signal is acquired with two electrodes around the flame of pyrolysis, which are sensible to the ions generated by the combustion. The sensibility of this sensor is linked to the Carbon number of the burnt chemical entity.

3. SIGNAL MODEL AND PREPROCESSING

Knowing the chemical nature of the gas, the recorded signal is considered directly proportional to the number of molecules. Let's quantify the number of molecules arriving on the sensor. According to the Giddings and Eyring's model [1, 2], the probability that a molecule exits at time t is given by:

$$P^{GE}(t,\boldsymbol{\theta}) = 2\frac{\mu}{\sigma^2}\sqrt{\frac{\mu}{t}}\, I_1\left(4\frac{\mu}{\sigma^2}\sqrt{\mu t}\right) e^{-2\frac{\mu}{\sigma^2}(t+\mu)}. \qquad (1)$$

Where:

- $\boldsymbol{\theta} = (\mu, \sigma)$ is the set of parameters of distribution (1)
- μ corresponds to the mean
- σ corresponds to the standard deviation
- $I_1(\cdot)$ denotes the first order first kind Bessel function.

The response of the sensor is considered as being perfect, that is to say the impulse response of the FID is modeled by a Dirac distribution.

Finally, the probability $P^{molecule}(t)$ that an unknown molecule from the sample, except those from the solvent which is in excess, exits the column at time t is given by:

$$P^{molecule}(t) = \sum_{j=1}^{M} c_j\, P^{GE}\left(t,\boldsymbol{\theta_j}\right), \qquad (2)$$

where j is the index of chemical entity, c_j its proportion, M the total number of gases in the mixture and $\boldsymbol{\theta_j}$ the parameters vector of retention time distribution for chemical entity j. We then propose to model the electric signal g, expressed in Volt unit, as:

$$g(t) = \alpha \sum_{j=1}^{M} c_j\, P^{GE}\left(t,\boldsymbol{\theta_j}\right) + b(t), \qquad (3)$$

where α is the coefficient of proportionality expressed in Volts by molecules and $b(\cdot)$ is the noise function.

Figure 2 shows a signal example.

In this figure, we observe a baseline including a first saturated peak which exponentially decays. It is the elution peak solvent, the methanol. Our interest is focused on the other peaks corresponding to chemical entities present in the injected gas mixture.

We suppress this baseline and reduce the noise with a preprocessing filter. We use two successive averaging filters with two sliding rectangular windows at different scales. The size of the windows depends on the signal. This is not an automatic processing. The preprocessed signal is noted $g_{pp}(t)$ and is represented in Figure 3.

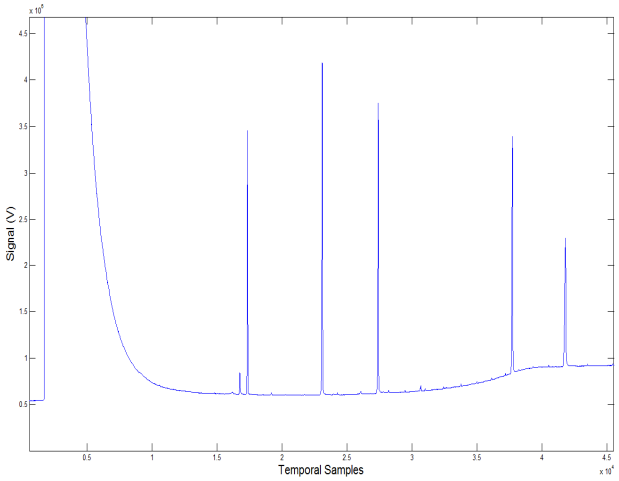

Figure 2: Example of GC-FID signal

Moreover, the signal of interest starts after the saturated solvent first peak and ends just after the last peak. In this example we keep 32 500 signal samples.

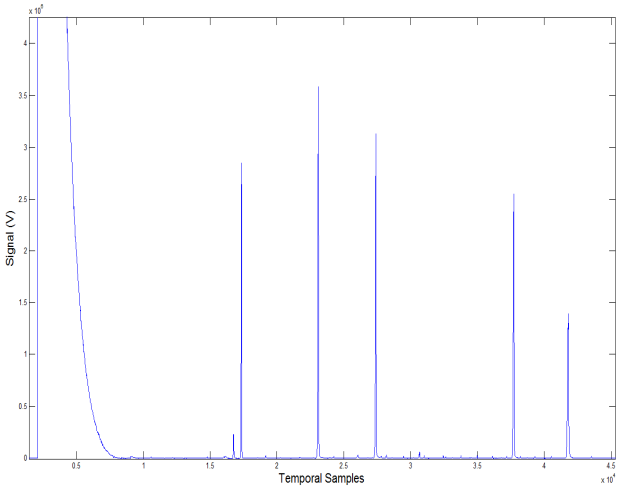

Figure 3: GC-FID signal after preprocessing

4. INVERSION

We propose to investigate two inversion methods: the first one is Bayesian inference based on a microscopic model, while the other one is a sparse representation of the signal on a redundant dictionary based on a macroscopic model.

4.1. Bayesian inversion

From this preprocessed signal we construct a population of N retention times of some molecules extracted from the gas mixture analyzed. We define $\boldsymbol{y} = \{t_i\,, i = 1..N\}$ as the list of retention times supposed independent. We set the size N of the population to be a compromise between statistical approximation quality and computational time. When N tends toward infinity the normalized histogram of all this population corresponds to the normalized preprocessed signal.

Practically to construct $\boldsymbol{y}$, we sample random retention times under the normalized preprocessed signal distribution $\frac{g_{pp}(t)}{\int g_{pp}(t)\,dt}$ which is modeled by the distribution (2).

Once $\boldsymbol{y}$ is sampled, we need to estimate the unknown parameters of the signal: $\boldsymbol{\theta_j}$ and c_j.

We use a posterior expectation estimator to estimate simultaneously $\boldsymbol{\mu}, \boldsymbol{\sigma}$ and $\boldsymbol{c} = (c_j)$.

$$\widehat{\boldsymbol{\Theta}} = \int_{\Omega} \boldsymbol{\Theta}\, p(\boldsymbol{\Theta}|\boldsymbol{y}) d\boldsymbol{\Theta},$$

where Ω is the domain of the possible values for the vector, $\boldsymbol{y}$ the observations and $\boldsymbol{\Theta} = (\boldsymbol{\mu}, \boldsymbol{\sigma}, \boldsymbol{c})$.

In order to apply Bayes' rule, $p(\boldsymbol{\Theta}|\boldsymbol{y}) \propto p(\boldsymbol{\Theta})\, p(\boldsymbol{y}|\boldsymbol{\Theta})$, we define the likelihood and settle the priors as in [3, 4].

The likelihood on the retention time list $\boldsymbol{y}$ is given by:

$$p(\boldsymbol{y}|\boldsymbol{\Theta}) = \prod_{i=1}^{N} \sum_{j=1}^{M} c_j\, P^{GE}(t_i, \boldsymbol{\theta_j}). \qquad (4)$$

We define the following prior distributions:

$$p(\boldsymbol{\mu}) = \mathcal{N}\left(\boldsymbol{m_\mu}, \frac{1}{2}\boldsymbol{I}\right), \qquad (5)$$

$$p(\boldsymbol{c}) = Di\left(\frac{1}{M}, \dots, \frac{1}{M}\right), \qquad (6)$$

$$p(\boldsymbol{\sigma}) = \mathcal{N}\left(\boldsymbol{m_\sigma}, \frac{1}{2}\boldsymbol{I}\right), \qquad (7)$$

where $\mathcal{N}$ denote the normal distribution and Di the Dirichlet distribution. The choice of Dirichlet distribution ensures the property $\sum_j^M c_j = 1$ and $c_j \geq 0$.

The hyper-parameters $\boldsymbol{m_\mu}$, $\boldsymbol{m_\sigma}$ **and** $\boldsymbol{M}$ of these priors are fixed constants.

To define $\boldsymbol{m_\mu}$ we use the `findpeaks` matlab function applied on the preprocessed signal. This function detects peaks in a signal by thresholding the differences between successive samples of the signal.

Besides to define $\boldsymbol{m_\sigma}$ we implement a method of moment around each peak. This is possible because the peaks are widely separated in this case thanks to the column.

The dimension M of those hyper-parameters is determined by the number of peaks. This is a needed prior knowledge. Those hyper-parameters values provide first guess of the peak parameters. The Bayesian approach delivers fine tuning taking into account those approximate values and the exact measurement according to the molecular model.

The analytical expression of the posterior distribution is unknown. So in order to compute the posterior mean estimator we use a Markov Chain Monte Carlo (MCMC) algorithm. We note $\left(\boldsymbol{\Theta}^{(k)}\right)$ the Markov Chain. So we compute the estimator with the formula:

$$\widehat{\boldsymbol{\Theta}} = \frac{1}{K} \sum_{\mathrm{k}=K_0+1}^{K_0+K} \boldsymbol{\Theta}^{(k)},$$

where K_0 corresponds to the number of warming iterations and $K_0 + K$ corresponds to the total number of iterations.

To sample the Markov Chain under the posterior distribution, we use a Gibbs sampling which is described in Algorithm 1.

- **Initialization** : $\boldsymbol{\mu}^{(1)}, \boldsymbol{\sigma}^{(1)}, \boldsymbol{c}^{(1)}$
- **For** $k = 2$ to $K + K_0$
  - **Sample** $p\left(\boldsymbol{\mu}^{(k)}|Y, \boldsymbol{\sigma}^{(k-1)}, \boldsymbol{c}^{(k-1)}\right)$
  - **Sample** $p\left(\boldsymbol{c}^{(k)}|Y, \boldsymbol{\mu}^{(k)}, \boldsymbol{\sigma}^{(k-1)}\right)$
  - **Sample** $p\left(\boldsymbol{\sigma}^{(k)}|Y, \boldsymbol{\mu}^{(k)}, \boldsymbol{c}^{(k)}\right)$
- **End For**.

Algorithm 1: Gibbs sampling for the Bayesian algorithm

The parameters need to be sampled under their posterior distributions. As every posterior distribution is unknown we implement a Metropolis Hastings (MH) algorithm step. It consists in sampling a possible value $\boldsymbol{\Theta}^{(k+1)}$ knowing $\boldsymbol{\Theta}^{(k)}$under an arbitrary density. Then an acceptance ratio a_r is computed from those 2 values. The possible value is kept as the chosen value for $\boldsymbol{\Theta}^{(k+1)}$ with a probability a_r. To sample a possible value, the arbitrary probability density used corresponds for each parameter to its prior distribution. In that case the acceptance ratios always correspond to a likelihood ratio which is then quickly computed.

4.2. Sparse representation on a dictionary

The preprocessed signal can be seen as a linear combination of sources. Sparse representation allows determining the best combination. Each source produces an elementary signal also called *atom*. Those atoms are gathered in a dictionary. According to our model we propose to construct a dictionary $\boldsymbol{\mathcal{D}}$ of signals by choosing a structured grid for the 2 dimensions of the parameters space with regular sampling step on each dimension:

$$\boldsymbol{\mathcal{D}} = \left\{ P^{GE}(t_i, \boldsymbol{\theta_j}) \right\}_{i,j}. \qquad (8)$$

The signal model (3) hence becomes:

$$\boldsymbol{y^{ss}} = \boldsymbol{\mathcal{D}}\, \boldsymbol{s} + \boldsymbol{b}, \qquad (9)$$

where $\boldsymbol{y^{ss}}$ denotes the sampled normalized signal, s the mixture vector to estimate and b the noise.

Moreover, for physical reasons we introduce a non-negativity constraint upon $\boldsymbol{s}$.

In term of minimization, we rewrite the inverse problem to solve as:

$$\hat{\boldsymbol{s}} = \min_{\boldsymbol{s}} \{ \|\boldsymbol{y^{ss}} - \boldsymbol{\mathcal{D}}\, \boldsymbol{s}\|_2^2, \quad s.t : \boldsymbol{s} > 0 \text{ and } \min \|\boldsymbol{s}\|_0 \}, \qquad (10)$$

where: $\|\cdot\|_2$ denotes the L_2 norm and $\|\cdot\|_0$ the L_0 norm.

To solve this equation, we use a FOCal Underdetermined System Solution (FOCUSS) algorithm [5]. This algorithm

can easily take into account the non-negativity constraint compared to others algorithms like SL0 [6, 7].
The principle of the algorithm is to find iteratively a low resolution estimation of the signal by minimizing the quadratic quantity (10) and then prune this solution in a sparse signal representation. The pruning is achieved with a minimization of an L_p norm and by an affine scaled tranformation. We need to choose the power p value close to zero.

Finally the concentration vector $\boldsymbol{c}$ is estimated by normalizing the concentration of non-zeros values of the vector $\boldsymbol{s}$.

5. RESULTS

5.1. Bayesian algorithm

For computational reasons the algorithm was applied with $N = 603$ molecules whose retention times are drawn from the signal of interest shown in Figure 2. Figure 5 shows the convergence of the log-likelihood applied with each $\boldsymbol{\Theta}^{(k)}$. The stability of the likelihood highlights the convergence of the Markov Chain. We fixed the number of warming iterations at 20000.

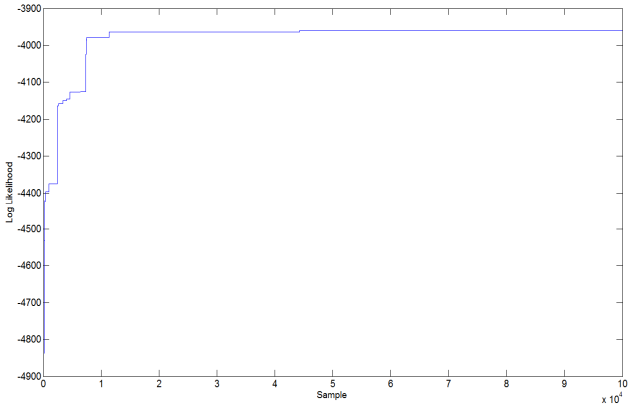

Figure 5: Log-likelihood with iteration

5.2. Sparse representation algorithm

Our dictionary is composed of 63 300 atoms of 25 250 temporal samples, i.e. $\boldsymbol{\mathcal{D}}$ is 63 300 x 25 250 size. Figure 6 shows a rather good fit of the model to the preprocessed signal. For readability only two peaks are showed. The power on the constraint is settled at $p = 0{,}01$.

The FOCUSS algorithm needs estimation of the noise variance. We estimate it on a portion of the preprocessed signal between 10 and 20 min where there is no peak in the signal.

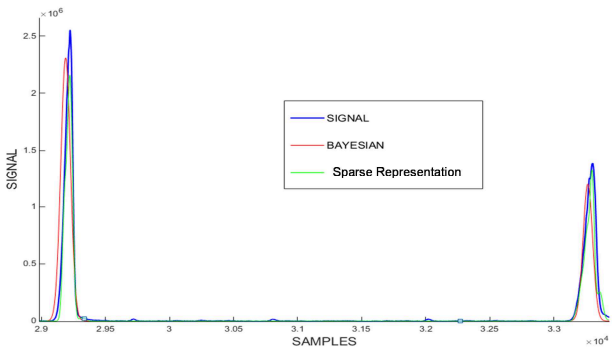

Figure 6: Display of signal and fit for each method.

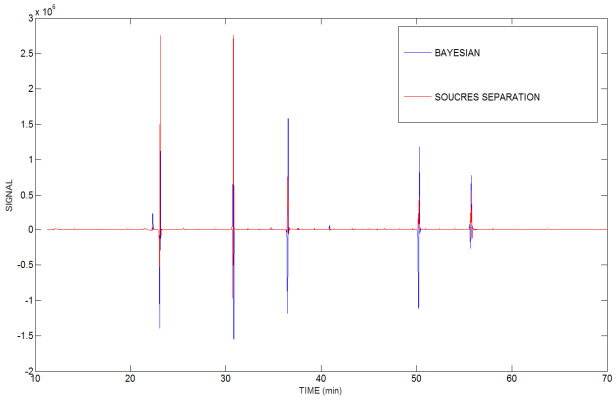

Figure 7: Display of subtractions between signal and fits.

5.3. Algorithm comparison

We fix the parameters of both algorithms at the values selected in sections 5.1 and 5.2.
Let us examine the quantification performances. As we can see in Figure 7, the first little parasite peak present in the signal is not detected by Bayesian algorithm. This is explained by the fact that we need the number of peaks as a prior knowledge in this Bayesian estimation. This is an issue of this kind of Bayesian parametric inference compared with the sparse representation method. Also this parasite peak introduces a bias on the neighboring estimated peak. The molecules associated to this peak are indeed clustered with the molecules of the neighboring peak.

The algorithms were applied on GC signals for analysis of a mixture of pollutants in methanol. We did these experiments with different dilution rates (1, 10 and 100 μg/L). The references values of the proportions are the areas under the peaks. By the way we suppose that the ratios of PAHs aren't modified by the different parts of the system. The estimation errors depend also on the PAHs dilution. Indeed the more the concentration of pollutant in the solvent is important the more the signal to noise ratio (SNR) is high. Figure 8 sums up the concentration values obtained by those two algorithms and the real values. We observe that sparse representation provides, in mean, better estimation than Bayesian algorithm.

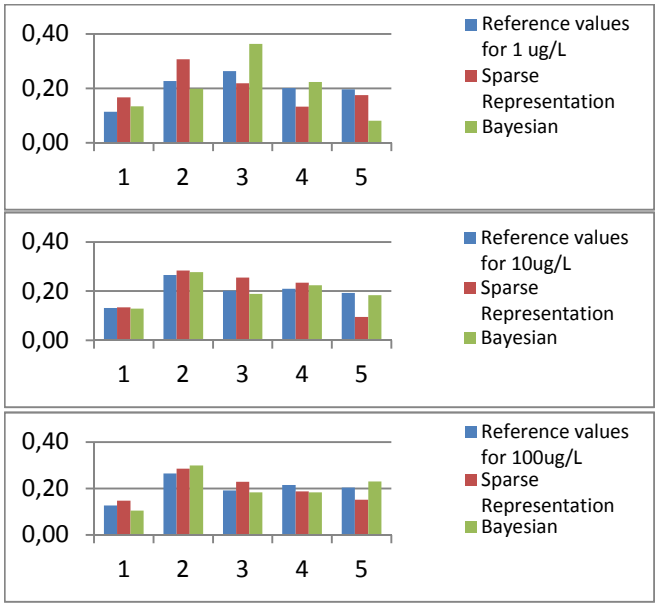

Figure 8: Ratios values, reference in blue and estimate by Bayesian in green and sparse representation in red.

The relative root mean square error used is the square root of the quadratic error between the true and estimated signal divided by the power of the signal. It is given in the following table for each dilution and for each method.

	Bayesian	Sparse representation
Dilution 1 μg/L	1.64	0.68
Dilution 10 μg/L	1.12	0.71
Dilution 100 μg/L	0.58	0.47

Table 1: Relative root mean square error for each dilution

We observe the better performances of the sparse representation. It's caused by the fact that our Bayesian algorithm is not designed for an unknown number of peaks which is not the case with dictionary. However in some applications the number of peaks is known.
Also the errors are sensitive to the position peak estimation (μ).

Let us now compare the computing performances of the two algorithms which have been run on the same computer. For the sparse representation algorithm we don't take into account the computational time used for constructing the dictionary. The computational time for our Bayesian algorithm is $\mathcal{O}(N^2\,(K+K_0))$ elementary operations, whereas the sparse representation needs $\mathcal{O}(N^3)$ elementary operations. The following table shows the computational time for the lowest SNR, and the required memory resources we observed on our computer (Intel Xeon 2 GHz CPU) using the Matlab software. The convergence is fastest for Bayesian method with higher SNR [8].

	Bayesian	Sparse representation
Computational time	2596 sec	9855 sec
Memory resources	500 Mo	100 Go

Table 2: Algorithms comparison on computational time

We note that the sparsity of the dictionary is not respected in our case. For computational reasons the number of atoms is not much larger than the number of temporal samples. FOCUSS algorithm needs indeed some memory resources, which limit the size of the dictionary. The dictionary used requires 21 Go of RAM.

6. CONCLUSION

To sum up, we have presented a stochastic molecular model of GC-FID signal. We have proposed two inversion methods to compute the concentration profiles. We have compared their quantification and computational performances. The main results are that this parametric Bayesian estimation is faster and requires less memory resources than sparse representation. The quantification by sparse representation with possibility of estimating the number of peaks M gives in this case a better estimation of relative concentrations than the Bayesian method. This is due to an error on the a priori peak number caused by the existence of a contaminant peak.
In perspective, we propose to improve the sparse algorithm by introducing an iterative processing with an adaptation of the resolution of the dictionary at each iteration, starting from a dictionary with a low temporal resolution up to an adapted and high resolution dictionary. Thus it will become possible to add sparsity in the dictionary only in the temporal regions of interest. The Bayesian algorithm will also be improved by implementing a non-parametric Bayesian algorithm. This will allow analyzing the chromatographic signal with an unknown number of peaks [8].

REFERENCES

[1] J. C. Giddings et H. Eyring, «A Molecular Dynamic Theory of Chromatography,» *The Journal of physical Chemistry* 1955, vol 59, pp. 416-421.

[2] A. Felinger, «Molecular dynamic theories in chromatography,» *Journal of Chromatography A,* vol. 1184, pp. 20-41, 2008.

[3] G. Strubel, J. F. Giovannelli, C. Paulus, L. Gerfault et P. Grangeat, «Bayesian estimation for molecular profile reconstruction in proteomics based on liquid chromatography and mass spectrometry,» chez *Engineering in Medicine and Biology Society, 2007. EMBS 2007. 29th Annual International Conference of the IEEE*, 2007.

[4] P. Szacherski, J.-F. Giovannelli, L. Gerfault, P. Mahe, J.-P. Charrier, A. Giremus, B. Lacroix et P. Grangeat, «Classification of Proteomic MS Data as Bayesian Solution of an Inverse Problem,» *Access, IEEE,* vol. 2, pp. 1248-1262, 2014.

[5] I. F. Gorodnitsky et B. D. Rao, «Sparse signal reconstruction from limited data using FOCUSS: A re-weighted minimum norm algorithm,» *IEEE Trans. Signal Processing,* pp. 600-616, 1997.

[6] M. B.-Z. G. Hosein Mohimani et C. Jutten, «Fast Sparse Representation based on Smoothed l0 Norm,» *HAL Archives ouvertes,* 2007.

[7] J. Murray et K. Kreutz-Delgado, «An improved FOCUSS-based learning algorithm for solving sparse linear inverse problems,» chez *Signals, Systems and Computers, 2001. Conference Record of the Thirty-Fifth Asilomar Conference on*, 2001.

[8] F.Bertholon et al., «From molecular model to sparse representation of chromatographic signals with an unknown number of peaks» *EMBC 2015*.