Joint Bayesian Decomposition of a Spectroscopic Signal Sequence

Vincent Mazet

***Abstract*—This letter addresses the problem of decomposing a sequence of spectroscopic signals: data are a series of (energy or electromagnetic) spectra and we aim to estimate the peak parameters (centers, amplitudes and widths). The key idea is to perform the decomposition of the whole sequence and to impose the parameters to evolve smoothly through the sequence. The problem is set within a Bayesian framework whose posterior distribution is sampled using a Markov chain Monte Carlo simulated annealing algorithm. Simulations conducted on synthetic data illustrate the performance of the method.**

***Index Terms*—Spectroscopic signal sequence; spectrum decomposition; Bayesian inference; Markov chain Monte Carlo (MCMC) method; Gibbs sampler; simulated annealing.**

I. Introduction

A SPECTROSCOPIC signal[1] represents the repartition of particles or electromagnetic waves with respect to their energy, wavelength, etc. It typically gathers several peaks whose positions and areas give informations about the chemical composition of the sample analysed. Some chemical experiments involve several spectra of the same sample acquired at different instants or temperatures [1], [2]. Thus, data are a sequence of different spectra, although two contiguous spectra (a spectrum and the one acquired at the next or previous instant or temperature) are similar because they gather the same peaks which slowly evolve between them (Fig. 1 shows a simulated sequence). In this work, we aim to decompose each spectrum of the sequence, that is to estimate the centers, amplitudes and widths of the peaks (it amounts to the positions and areas).

The decomposition can be done sequentially: the spectra are decomposed separately, without taking into account the similarity between contiguous spectra. We outlined in [3] that decomposing each spectrum independently from the others may lead to unsatisfactory results. Indeed, the decomposition of two contiguous (and, thus, different) spectra may lead to two very different decompositions which is physically impossible because the spectra evolve smoothly. Also, it would be useful to follow the peaks through the sequence so as to obtain an evolution of each peak: this implies to classify the peaks, each class corresponding to a unique peak in one spectrum. A sequential decomposition cannot achieve this point, and a post-processing should be performed.

Thus, we propose to perform the decomposition by taking into account the neighborhood of each spectrum, that is to use for each spectrum the decomposition of its two contiguous spectra. This can be done by decomposing the whole sequence. This approach has two advantages:

- a spectrum decomposition is aided by the contiguous decompositions since they are imposed to be similar (so it directs the optimization);
- furthermore, the estimation is more coherent than that obtained with a sequential decomposition because it is similar to its neighbors.

Several methods have been proposed for decomposing spectra [4]–[6], but, to our knowledge, nobody has addressed the problem of the joint decomposition of a sequence of spectra.

In this letter, the problem is set in a Bayesian framework since it has already been successfully proposed to ensure the decomposition of a unique spectrum [4]–[6] and it allows to take the neighborhood into account by imposing a regularization of the solution. The model and the priors are detailed in section II. The Maximum A Posteriori (MAP) estimator is obtained using a Markov chain Monte Carlo (MCMC) simulated annealing method (section III). Finally, the performance of the joint decomposition is compared with a sequential decomposition on synthetic data (section IV).

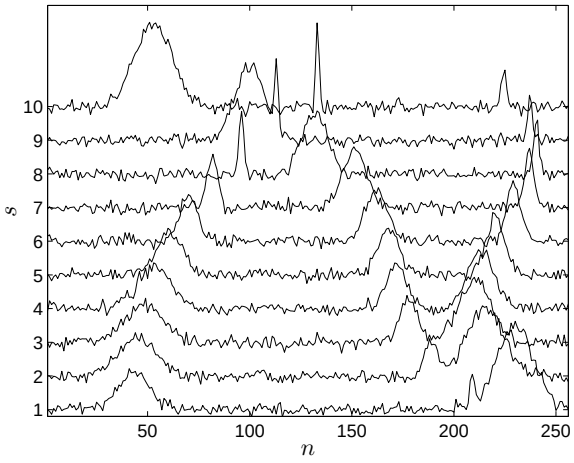

Fig. 1. A simulated sequence of $S = 10$ spectra with $N = 256$ samples and $K = 3$ peaks.

II. Problem Formulation & Bayesian Inference

The spectrum length is denoted N and the number of spectra in the sequence is denoted S. Each spectrum $\boldsymbol{y}_s \in \mathbb{R}^N$ ($s \in \{1, \ldots, S\}$) is modeled as a noisy sum of K peaks [5], [6]:

$$\forall s \in \{1, \ldots, S\}, \qquad \boldsymbol{y}_s = \sum_{k=1}^{K} \boldsymbol{\phi}(\boldsymbol{\theta}_{s,k}) + \boldsymbol{b}_s \qquad (1)$$

V. Mazet is with LSIIT (UMR 7005 University of Strasbourg & CNRS), Boulevard Sébastien Brant, BP 10413, 67412 Illkirch cedex, France. Phone: +33 3 68 85 44 91. Fax: +33 3 68 85 44 55. E-mail: vincent.mazet@unistra.fr.

[1]In the sequel, we will use instead the shorter term "spectrum".

where $\boldsymbol{\phi}(\boldsymbol{\theta}_{s,k}) \in \mathbb{R}^N$ is the parametric expression of the kth peak in spectrum s whose parameters are gathering into $\boldsymbol{\theta}_{s,k}$ and $\boldsymbol{b}_s \in \mathbb{R}^N$ is an additive noise modeling measurement errors and model uncertainties. The parametric expression of the peaks has to be chosen according to the experiment (it is generally a Gaussian). The peak parameters are its center $c_{s,k}$, amplitude $a_{s,k}$ and other shape parameters; in this letter, we deal with a unique shape parameter: the peak width $w_{s,k}$. Note also that the peak number in each spectrum (K) is supposed to be constant and known; this is a limitation of the proposed method since the peak number is in fact unknown and may vary through the sequence (an extension to this work would be to estimate K using a trans-dimensional MCMC approach [7]).

Besides, two contiguous spectra being similar, their peaks are also similar (i.e. $\boldsymbol{\theta}_{s,\bullet} \approx \boldsymbol{\theta}_{r,\bullet}$ if s and r are the index of two neighboring spectra). Thus one can introduce a label $z_{s,k} \in \{1,\dots,K\}$ for each peak such that the peaks with the same label evolve smoothly. Formally, it reads:

$$\boldsymbol{\theta}_{s,k} \approx \boldsymbol{\theta}_{r,l} \quad \text{if and only if} \quad z_{s,k} = z_{r,l} \quad \text{and} \quad |s-r| = 1.$$

For convenience, we will note in the sequel $\boldsymbol{\theta}^l$ the vectors $\{\boldsymbol{\theta}_{s,k}\}$ such that $z_{s,k} = l$ for all $s \in \{1,\dots,S\}$ and $k \in \{1,\dots,K\}$.

Thus, $\boldsymbol{\theta}_{s,k} = \{c_{s,k}, a_{s,k}, w_{s,k}, z_{s,k}\}$ gathers the center, amplitude, width and label of the kth peak in the sth spectrum: they are the unknown parameters to be estimated.

A. Priors

The noise is supposed to be white and Gaussian with variance $r_{\boldsymbol{b}}$:

$$\forall s,n, \quad b_{s,n}|r_{\boldsymbol{b}} \sim \mathcal{N}(0, r_{\boldsymbol{b}}) \tag{2}$$

where $n \in \{1,\dots,N\}$ denotes the samples and $\mathcal{N}(\mu,\sigma^2)$ is a Gaussian distribution with mean μ and variance σ^2.

The peak centers with the same label $l \in \{1,\dots,K\}$ are supposed to evolve slowly through the sequence. Thus, we choose a smoothness prior to regularize the solution. We also impose $c_{s,k} \in [1,N]$ so that the peaks lie in the observation interval. These priors can be expressed by [8]

$$p(\boldsymbol{c}^l|\boldsymbol{z}) \propto \exp\left(-\frac{1}{2r_{\boldsymbol{c}}}\|D\boldsymbol{c}^l\|^2\right) \mathbb{I}_{[1,N]}(\boldsymbol{c}^l) \tag{3}$$

where $r_{\boldsymbol{c}}$ is an hyperparameter controlling the prior strength, $\|\cdot\|$ is the L^2 norm, D is a difference operator (discrete derivative), and $\mathbb{I}_E(\boldsymbol{x})$ is the indicator function: $\mathbb{I}_E(\boldsymbol{x}) = 1$ if $\boldsymbol{x} \in E^M$ (M being the size of $\boldsymbol{x}$), 0 otherwise. A second-order derivative promotes the evolution to be linear:

$$\forall s \in \{2,\dots,S-1\}, \quad (D\boldsymbol{c}^l)_s = c^l_{s-1} - 2c^l_s + c^l_{s+1}. \tag{4}$$

On the other hand, when s is fixed (i.e., considering a unique spectrum), the peak centers $\boldsymbol{c}_{s,k}$ are supposed to be independent. Thus, the prior on $\boldsymbol{c}$ reads:

$$p(\boldsymbol{c}|\boldsymbol{z}) \propto \exp\left(-\frac{1}{2r_{\boldsymbol{c}}}\sum_{l=1}^{K}\|D\boldsymbol{c}^l\|^2\right) \mathbb{I}_{[1,N]}(\boldsymbol{c}). \tag{5}$$

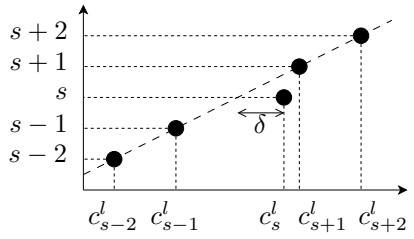

Fig. 2. Smoothness variances $r_{\boldsymbol{c}}$, $r_{\boldsymbol{a}}$ and $r_{\boldsymbol{w}}$ can be set according to allowed curvature of the peak parameters. This illustrates an example for $\boldsymbol{c}^l$: δ is the difference between the actual value c^l_s and the linear evolution defined by the neighboring parameters, i.e $(c^l_{s-1} + c^l_{s+1})/2$.

In the same way, the peak amplitudes $\boldsymbol{a} = \{a_{s,k}\}$ and the peak widths $\boldsymbol{w} = \{w_{s,k}\}$ are supposed to evolve slowly through the sequence and to be positive. Thus their priors are:

$$p(\boldsymbol{a}|\boldsymbol{z}) \propto \exp\left(-\frac{1}{2r_{\boldsymbol{a}}}\sum_{l=1}^{K}\|D\boldsymbol{a}^l\|^2\right) \mathbb{I}_{[0,+\infty[}(\boldsymbol{a}), \tag{6}$$

$$p(\boldsymbol{w}|\boldsymbol{z}) \propto \exp\left(-\frac{1}{2r_{\boldsymbol{w}}}\sum_{l=1}^{K}\|D\boldsymbol{w}^l\|^2\right) \mathbb{I}_{[0,+\infty[}(\boldsymbol{w}). \tag{7}$$

A priori, there is no information available on the labels. So they are distributed according to a uniform distribution:

$$\forall s, \qquad \boldsymbol{z}_s \sim \mathcal{U}_{\mathcal{S}_K} \tag{8}$$

where $\mathcal{S}_K$ is the permutation group of the set $\{1,\dots,K\}$.

A common choice for the noise variance ($r_{\boldsymbol{b}}$) prior is the Jeffreys distribution which is unfortunately non-proper. Therefore, we adopt an inverse gamma prior for $r_{\boldsymbol{b}}$:

$$r_{\boldsymbol{b}} \sim \mathcal{IG}(\alpha_b, \beta_b) \tag{9}$$

whose parameters are both set close to zero ($\alpha_b = \beta_b \ll 1$) which is in fact the limit case corresponding to the Jeffreys prior.

The three other hyperparameters ($r_{\boldsymbol{c}}$, $r_{\boldsymbol{a}}$ and $r_{\boldsymbol{w}}$) are fixed a priori: considering that they are related to the smoothness of the peak evolution (the smaller they are, the smoother is the solution), a heuristic method can set them. For instance, suppose that all peak centers would have a perfect linear evolution except c^l_s (Fig. 2). Then $\|D\boldsymbol{c}^l\|^2 = 6\delta^2$ where δ is the difference between the actual value c^l_s and the linear evolution defined by the neighboring parameters (that is $(c^l_{s-1} + c^l_{s+1})/2$). Then we can set $r_{\boldsymbol{c}} = 6\delta^2$ where δ is the mean variation to a linear evolution of the peak centers.

B. Posteriors

The posterior integrability cannot been checked mathematically because of its complexity, but we took care to use integrable priors so a sufficient condition to prove the posterior integrability is fulfilled. It is sampled using a Gibbs sampler (section III) which consists in simulating each variable according to its conditional posterior distribution [9]. We then obtain the following conditional posterior distributions.

The centers and widths are distributed $\forall s,k$ according to

$$c_{s,k}|\cdots \sim \exp\left(-\frac{\|\boldsymbol{e}_s\|^2}{2r_{\boldsymbol{b}}} - \frac{\|D\boldsymbol{c}^l\|^2}{2r_{\boldsymbol{c}}}\right) \mathbb{I}_{[1,N]}(c_{s,k}), \tag{10}$$

TABLE I
COEFFICIENTS $A_{s,k}$ AND $B_{s,k}$ RELATED TO $a_{s,k}$ ($l = z_{s,k}$).

s	$A_{s,k}$	$B_{s,k}$
1	1	$-4a_2^l + 2a_3^l$
2	5	$-4a_1^l - 8a_3^l + 2a_4^l$
$3, \ldots, S-2$	6	$2a_{s-2}^l - 8a_{s-1}^l - 8a_{s+1}^l + 2a_{s+2}^l$
$S-1$	5	$-4a_S^l - 8a_{S-2}^l + 2a_{S-3}^l$
S	1	$-4a_{S-1}^l + 2a_{S-2}^l$

$$w_{s,k}|\cdots \sim \exp\left(-\frac{\|\boldsymbol{e}_s\|^2}{2r_{\boldsymbol{b}}} - \frac{\|D\boldsymbol{w}^l\|^2}{2r_{\boldsymbol{w}}}\right)\mathbb{I}_{[0,+\infty[}(w_{s,k}) \quad (11)$$

with $l = z_{s,k}$ and $\boldsymbol{e}_s = \boldsymbol{y}_s - \sum_j \boldsymbol{\phi}(\boldsymbol{\theta}_{s,j})$.

The amplitude posterior is a Gaussian distribution with positive support:

$$a_{s,k}|\cdots \sim \exp\left(-\frac{(a_{s,k}-\mu_{s,k})^2}{2\rho_{s,k}}\right)\mathbb{I}_{[0,+\infty[}(a_{s,k}) \quad (12)$$

where

$$\mu_{s,k} = \frac{\rho_{s,k}}{r_{\boldsymbol{b}}}\big(\boldsymbol{e}_s + \boldsymbol{\psi}(\boldsymbol{\theta}_{s,k})\big)^T\boldsymbol{\psi}(\boldsymbol{\theta}_{s,k}) - \frac{\rho_{s,k}}{2r_{\boldsymbol{a}}}B_{s,k}, \quad (13)$$

$$\rho_{s,k} = \frac{r_{\boldsymbol{a}}r_{\boldsymbol{b}}}{r_{\boldsymbol{b}}A_{s,k} + r_{\boldsymbol{a}}\boldsymbol{\psi}(\boldsymbol{\theta}_{s,k})^T\boldsymbol{\psi}(\boldsymbol{\theta}_{s,k})}, \quad (14)$$

$A_{s,k}$ and $B_{s,k}$ are summarized in Table I and $\boldsymbol{\psi}(\boldsymbol{\theta}_{s,k}) = \boldsymbol{\phi}(\boldsymbol{\theta}_{s,k})/a_{s,k}$.

The labels are distributed according to

$$\boldsymbol{z}_s|\cdots \sim \exp\left(-\sum_l\left(\frac{\|D\boldsymbol{c}^l\|^2}{2r_{\boldsymbol{c}}} + \frac{\|D\boldsymbol{a}^l\|^2}{2r_{\boldsymbol{a}}} + \frac{\|D\boldsymbol{w}^l\|^2}{2r_{\boldsymbol{w}}}\right)\right)\mathbb{I}_{\mathcal{S}_K}(\boldsymbol{z}_s). \quad (15)$$

Finally, the noise variance posterior is:

$$r_{\boldsymbol{b}}|\cdots \sim \mathcal{IG}\left(\frac{NS}{2} + \alpha_b, \sum_s \frac{\|\boldsymbol{e}_s\|^2}{2} + \beta_b\right). \quad (16)$$

III. MCMC SIMULATED ANNEALING ALGORITHM

We aim at finding the MAP estimator but the posterior is too complex to derive closed-form expressions for this estimator because of its high-dimension and the numerous local minima. Thus, as in [4]–[6], a MCMC algorithm generates samples that are asymptotically distributed according to the posterior.

The easiest way to sample the posterior consists in using a (random sweep) Gibbs sampler[2] [9], [10]: the unknowns are sampled according to the conditional posteriors (section II-B). To be more precise, the centers (resp. widths) are sampled using a random walk Metropolis-Hastings algorithm [9] whose proposal distribution is a Gaussian centered on the current value with support $[1, N]$ (resp. $[0, +\infty[$). The labels are sampled using a Metropolis-Hastings algorithm whose proposal distribution is uniform on the permutation group (a direct simulation has to be avoided because it is time-consuming when K increases). Finally, the amplitudes and the noise variance can be sampled directly [9].

Yet, it is known that, in practice, the Gibbs sampler can be trapped in a local minimum [9]. Therefore, it is included within a simulated annealing scheme to find the MAP estimator [10], [11]. This technique involves simulating at iteration i the distribution p^{1/T_i} where the "temperature" T_i is set so that, under weak regularity assumptions on p, p^{1/T_i} resembles to a uniform distribution at a high temperature and concentrates itself on its global maximum when the temperature decreases. We use a geometric decrease:

$$T_i = \alpha T_{i-1} \quad (17)$$

where $\alpha = (T_f/T_1)^{1/I}$ and I is the iteration number. The MAP estimate is the sample of the Markov chain which maximizes the posterior $p(\boldsymbol{c}, \boldsymbol{a}, \boldsymbol{w}, \boldsymbol{z}, r_{\boldsymbol{b}}|\boldsymbol{y})$.

Finally, let's talk about the practical implementation of the method. It appears that storing the labels has two drawbacks: first, it needs memory to save the labels; and second, computing the conditional posteriors of $c_{s,k}$, $a_{s,k}$ and $w_{s,k}$ needs to find the peaks with the same label which is time-consuming. Instead, we prefer to permute the peak themselves so that the label is implicitly coded by the peak order[3].

IV. SIMULATIONS RESULTS

This section studies the accuracy of the joint decomposition method on the synthetic spectra shown Fig. 1. It is a sequence of $S = 10$ spectra with $N = 256$ samples and $K = 3$ Gaussian peaks:

$$\boldsymbol{\phi}(\boldsymbol{\theta}_{s,k})_n = a_{s,k}\exp\left(-\frac{(n-c_{s,k})^2}{2w_{s,k}^2}\right). \quad (18)$$

The mean signal-to-noise ratio is approximately 15 dB. The initial and final temperatures are set to $T_1 = 50$ and $T_I = 1$ respectively. The smoothness variances were set to $6\delta^2$ where the mean variation δ equals 1 % of the dynamic of each parameter. Supposing that the dynamic of $\boldsymbol{c}$, $\boldsymbol{a}$ and $\boldsymbol{w}$ are respectively equal to N, $\max(\boldsymbol{y})$ and $N/10$, it yields:

$$r_{\boldsymbol{c}} \approx 39, \qquad r_{\boldsymbol{a}} \approx 4\cdot 10^{-3}, \qquad r_{\boldsymbol{w}} \approx 0.4.$$

The method was implemented in Matlab 7.10.0 on a 2.5-GHz Pentium E5200 with Ubuntu 10.04; the code and data are available on-line at http://lsiit-miv.u-strasbg.fr/mazet/jointdec.

Fig. 3 shows the MAP estimation for $I = 5000$ iterations. The estimation is clearly in good agreement with the actual values of the peak parameters. Indeed, the mean errors on the centers, amplitudes and widths are 0.15, 0.04 and 0.13 respectively, while the maximal errors are 0.59, 0.21 and 0.61 respectively. Also, the labels are well estimated since the evolution of each peak is smooth, and the estimation agrees with the actual labels.

To quantify the proposed method benefits, we also compare the joint decomposition (JD) with a sequential decomposition (SD) in which each spectrum is processed separately from the other. The two decompositions handle the same model and parameters, but the labels are not estimated in the SD since the classification needs to deal the whole sequence. For the

[2]An alternative could be a Metropolis-Hastings algorithm if deriving the conditionals is not possible.

[3]This looks like the label switching problem [12]. However, the problem here is different because we just want to avoid the storage of the $S \times K$ matrix $\boldsymbol{z}$ and the search for the peaks with the same label. Moreover, we do not try to avoid permutations between the variables through the iterations since only one iteration is used to compute the estimation.

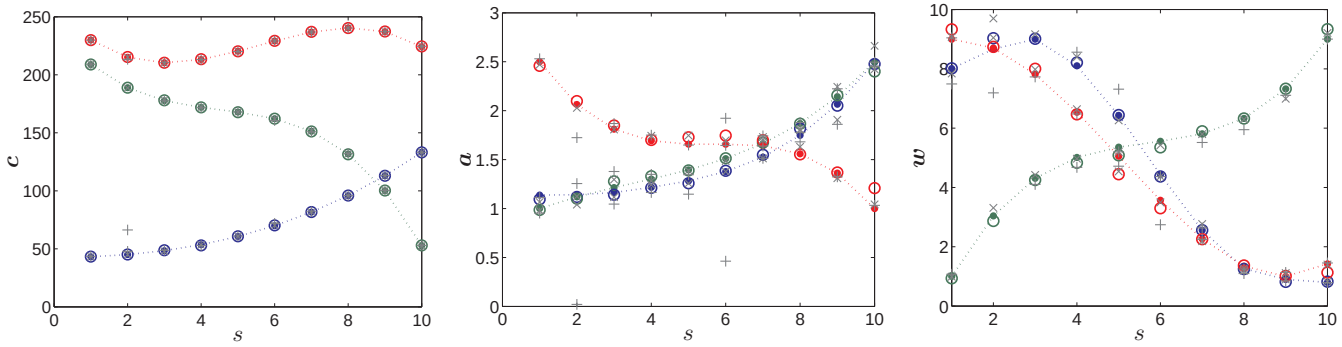

Fig. 3. Estimation of the peak parameters in the sequence of spectra shown Fig. 1. From left to right: centers, amplitudes and widths. Big dots and dotted lines represent the ground truth; circles, plus signs and crosses represent respectively the estimations of the joint decomposition (each color corresponds to a label), the sequential decomposition with and without simulated annealing. Some plus signs and crosses may be outside the axes.

TABLE II
PERFORMANCES OF THE JOINT DECOMPOSITION (JD) AND SEQUENTIAL DECOMPOSITION (SD) APPROACHES WITH OR WITHOUT SIMULATED ANNEALING (SA) ON 20 SEQUENCES (MSE: MEAN SQUARE ERROR).

	MSE	♯ iteration	Computation time
JD	$1.76 \cdot 10^{-2}$	5000	97.33 s
SD with SA	$7.20 \cdot 10^{-2}$	5700	98.07 s
SD, no SA	$11.11 \cdot 10^{-2}$	5700	96.18 s

same reason, the priors do not model a smooth variation of the peak parameters. Actually, they are $\forall s, k$:

$$c_{s,k} \sim \mathbb{I}_{[1,N]}(c_{s,k}), \tag{19}$$

$$a_{s,k} \sim \frac{1}{\sqrt{2\pi\sigma_{\boldsymbol{a}}^2}} \exp\left(-\frac{a_{s,k}^2}{2\sigma_{\boldsymbol{a}}^2}\right) \mathbb{I}_{[0,+\infty[}(a_{s,k}), \tag{20}$$

$$w_{s,k} \sim \mathbb{I}_{[0,+\infty[}(w_{s,k}). \tag{21}$$

The amplitude prior is a normal distribution with positive support to avoid divergence and $\sigma_{\boldsymbol{a}}^2 = 1$. A Metropolis-Hastings within Gibbs algorithm is also used for sampling. We choose to set the iteration number of the two methods so as to get the same computation time (about 97 s): this yields $I = 5700$ iterations for the SD. The estimations are plotted in Fig. 3.

In addition, the two methods were run on 20 sequences with $K = 4$; the performances are summarized in Table II. It appears that the mean square error (MSE) with the SD approach is worse than the one with the JD, either a simulated annealing is implemented or not. This can be explained by the fact that the SD may miss some peaks; on the contrary, the smoothness prior of the JD favor the search for peaks on certain zones in the solution space: in other terms, the estimation obtained for a spectrum is used to help the estimation of a contiguous spectrum. Moreover, the JD approach also performs the classification of the peaks which cannot be handled by the SD approach. Nevertheless, one can use a post-processing for classifying the estimated peaks (as in [3]), but the bad estimation of the peak parameters leads to incorrect classification. This confirms the superiority of the JD method to process a sequence of spectra.

V. CONCLUSION

We presented a new approach to estimate the peak parameters in a sequence of spectroscopic signals. The key idea is to achieve a joint decomposition by processing the whole sequence instead of each spectrum separately. This is done within a Bayesian framework in which a smoothness prior is imposed on the peak parameters and a MCMC simulated annealing algorithm is implemented to compute a MAP estimation. Simulations show that the proposed approach is relevant compared to a sequential decomposition scheme. Besides, a joint decomposition not only gives a good estimation but also performs the classification of the peaks. Future works will be mainly dedicated to the case where the peak number varies.

REFERENCES

[1] N. Dobigeon, S. Moussaoui, J.-Y. Tourneret, and C. Carteret, "Bayesian separation of spectral sources under non-negativity and full additivity constraints," *Signal Processing*, vol. 89, no. 12, pp. 2657–2669, 2009.
[2] A. Masson, L. Poisson, M.-A. Gaveau, B. Soep, J.-M. Mestdagh, V. Mazet, and F. Spiegelman, "Dynamics of highly excited barium atoms deposited on large argon clusters. I. General trends," *The Journal of Chemical Physics*, vol. 133, no. 5 (054307), 2010.
[3] V. Mazet, C. Collet, and B. Vollmer, "Decomposition and classification of spectral lines in astronomical radio data cubes," in *Proceedings of the 16th Scandinavian Conference on Image Analysis*, 2009, pp. 189–198.
[4] R. Fischer and V. Dose, "Analysis of mixtures in physical spectra." in *Bayesian methods*, 2001, pp. 145–154.
[5] S. Gulam Razul, W. Fitzgerald, and C. Andrieu, "Bayesian model selection and parameter estimation of nuclear emission spectra using RJMCMC," *Nuclear Instruments and Methods in Physics Research A*, vol. 497, pp. 492–510, 2003.
[6] N. Haan and S. Godsill, "Bayesian models for DNA sequencing," in *IEEE International Conference on Acoustics, Speech, and Signal Processing*, 2002, pp. IV–4020–IV–4023.
[7] P. Green, *Highly Structured Stochastic Systems*. Oxford University Press, 2003, ch. Trans-dimensional Markov chain Monte Carlo.
[8] D. Geman and C. Yang, "Nonlinear image recovery with half-quadratic regularization," *IEEE Transactions on Image Processing*, vol. 4, no. 7, pp. 932–946, 1995.
[9] C. Robert and G. Casella, *Monte Carlo statistical methods*, 2nd ed. Springer, 2004.
[10] S. Geman and D. Geman, "Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 6, pp. 721–741, 1984.
[11] S. Kirkpatrick, C. Gelatt Jr., and M. Vecchi, "Optimization by simulated annealing," *Science*, vol. 220, no. 4598, pp. 671–680, 1983.
[12] A. Jasra, C. Holmes, and D. Stephens, "Markov chain Monte Carlo methods and the label switching problem in Bayesian mixture modeling," *Statistical Science*, vol. 20, no. 1, pp. 50–67, 2005.