A BAYESIAN METHOD FOR POSITIVE SOURCE SEPARATION

Saïd Moussaoui, David Brie, Olivier Caspary

CRAN CNRS UMR 7039, UHP
B.P. 239, 54506 Vandœuvre-lès-Nancy, France
{*firstname.lastname*}*@cran.uhp-nancy.fr*

Ali Mohammad-Djafari

LSS CNRS-SUPELEC-UPS
91192, Gif-sur-Yvette cedex, France
djafari@lss-supelec.fr

ABSTRACT

This paper considers the problem of source separation in the particular case where both the sources and the mixing coefficients are positive. The proposed method addresses the problem in a Bayesian framework. We assume a Gamma distribution for the spectra and the mixing coefficients. This prior distribution enforces the non-negativity. This leads to an original method for positive source separation. A simulation example is presented to illustrate the effectiveness of the method.

1. INTRODUCTION

In analytical chemistry, spectral data resulting from sample analysis often present mixtures, i.e the measures are a linear combination of pure spectra. Pure spectra are needed to identify the sample constituents (qualitative analysis) and mixing coefficients are used to assess their concentrations (quantitative analysis). The mixture analysis can be formalized as a source separation problem on which many attention has been paid during the last two decades. See for example the surveys of [1, 2].

The linear instantaneous mixture model assumes that the m observed signals are a linear combination of n unknown sources, at each t (t can represent either time, frequency, wavenumber, etc.):

$$\boldsymbol{x}_t = \mathbf{A}\,\boldsymbol{s}_t + \boldsymbol{n}_t, \qquad (1)$$

where $\boldsymbol{s}_t$ denotes the $n \times 1$ source vector, $\boldsymbol{x}_t$ the $m \times 1$ vector containing the measured data, $\boldsymbol{n}_t$ a $m \times 1$ vector of an additive noise, $\mathbf{A}$ is a $m \times n$ unknown mixing matrix. The source separation aims at estimating the source signals $\boldsymbol{s} = \{\boldsymbol{s}_t\}_{t=1}^{N}$ and the mixing matrix $\mathbf{A}$, from the measured data $\boldsymbol{x} = \{\boldsymbol{x}_t\}_{t=1}^{N}$. This is an ill posed inverse problem since there are an infinity of solutions. To achieve separation, additional prior information and assumptions about the mixing process and sources are necessary. A common assumption firstly introduced is the statistical independence of the sources leading to *Independent Component Analysis* (ICA) algorithms [3]. In the case of spectroscopic mixtures, a very strong *a priori* knowledge is the non-negativity of both sources and mixing coefficients. To incorporate this information one can use an ICA method and optimize a contrast function under the source non-negativity constraint [4]. However, since ICA methods produce an unmixing matrix, which is the (pseudo) inverse of the mixing matrix, the non-negativity of the mixing coefficients cannot be ensured explicitly. This is the main shortcoming of this approach. Other methods consist in optimizing the least squares error under the non-negativity constraint, leading to algorithms differing on the manner how non-negativity constraint is introduced. In particular, the NMF algorithm (*Non-negative Matrix Factorization*) of Lee and Seung [5] achieves the decomposition by constructing a gradient descent algorithm over the objective function and updates iteratively spectra and concentration estimates under the non–negativity constraint. The procedure of Tauler et *al.* [6] performs an *Alternating Least Squares* (ALS) estimation where the non–negativity is hardly imposed between successive iterations. However, we believe that Bayesian estimation methods are more suitable in such an application because of the possibility to take into account explicitly the non-negativity information. The main idea of the Bayesian approach for source separation [2] is to use not only the likelihood $f(\boldsymbol{x}|\boldsymbol{s}, \mathbf{A})$ but also any prior knowledge one may have on the sources $\boldsymbol{s}$ and the matrix $\mathbf{A}$ through the assignment of prior distributions $p(\boldsymbol{s})$ and $p(\mathbf{A})$.

This paper is organized as follows: section 2 presents the proposed method for positive signal separation using Gamma priors for sources and mixing coefficients. A simulation example is presented in section 3.

2. POSITIVE SOURCE SEPARATION

2.1. Posterior Density

The noise is assumed to be zero mean, Gaussian, i.i.d (independent and identically distributed) and independent of the source signals. The sources s_j are supposed statistically i.i.d and distributed as Gamma distributions of parameters $\{\alpha_j, \beta_j\}_{j=1}^{n}$. These parameters are considered constant for

0-7803-8484-9/04/$20.00 ©2004 IEEE

each source but may differ from one source to another. The Gamma density is used to take into account non-negativity and its parameters allow a better fit to the spectra distribution. To incorporate the mixing coefficient non-negativity, each column j of the mixing matrix is also assumed distributed as a Gamma density of parameters $\{\lambda_j, \gamma_j\}_{j=1}^n$. These parameters are considered equal for each column j that corresponds to the variation of the source j concentrations. The Gamma density is expressed by:

$$\mathcal{G}(z;\alpha,\beta) = \frac{\beta^\alpha}{\Gamma(\alpha)}\, z^{\alpha-1}\, e^{-\beta z}\, \mathbb{I}_{[0,+\infty]}(z). \tag{2}$$

where $\Gamma(\alpha)$ is the Gamma function. This distribution allows to encode non-negativity since $p(z<0)=0$.

Using Bayes theorem and considering the vector $\boldsymbol{\theta}$ of hyperparameters containing the noise variance σ^2 and the gamma density parameters $\{\alpha_j, \beta_j, \gamma_j, \lambda_j\}_{j=1}^n$, the posterior law is expressed as:

$$\begin{aligned}\pi\left(\boldsymbol{s}, \mathbf{A}|\boldsymbol{x}, \boldsymbol{\theta}\right) &\propto \prod_{t=1}^{N} \mathcal{N}\left(\boldsymbol{x}_t - \mathbf{A}\,\boldsymbol{s}_t, \sigma^2 \mathbf{I}_m\right)\\ &\times \prod_{t=1}^{N}\prod_{j=1}^{n} \mathcal{G}(s_j(t);\alpha_j,\beta_j) \times \prod_{i=1}^{m}\prod_{j=1}^{n} \mathcal{G}(a_{ij};\lambda_j,\gamma_j).\end{aligned} \tag{3}$$

2.2. Joint MAP Estimation

The problem now is the posterior law maximization or equivalently the minimization of the resulting objective function $\Phi(\boldsymbol{s}, \mathbf{A}|\boldsymbol{\theta}) = -\log \pi\left(\boldsymbol{s}, \mathbf{A}|\mathrm{x}, \boldsymbol{\theta}\right)$, which takes the form:

$$\Phi(\boldsymbol{s}, \mathbf{A}|\boldsymbol{\theta}) = \Phi_L(\boldsymbol{s}, \mathbf{A}|\boldsymbol{\theta}) + \Phi_{P1}(\boldsymbol{s}|\boldsymbol{\theta}) + \Phi_{P2}(\mathbf{A}|\boldsymbol{\theta}), \tag{4}$$

where the terms Φ_L, Φ_{P1}, and Φ_{P2} are given by:

$$\Phi_L = \frac{1}{2\sigma^2}\sum_{t=1}^{N}\sum_{i=1}^{m}\left[x_i(t) - [\mathbf{A}\boldsymbol{s}]_i(t)\right]^2, \tag{5}$$

$$\Phi_{P1} = \sum_{t=1}^{N}\sum_{j=1}^{n}\left[(1-\alpha_j)\,\log s_j(t) + \beta_j s_j(t)\right], \tag{6}$$

$$\Phi_{P2} = \sum_{i=1}^{m}\sum_{j=1}^{n}\left[(1-\lambda_j)\,\log a_{ij} + \gamma_j\, a_{ij}\right]. \tag{7}$$

The first term Φ_L can be seen as a data fitting measure, while the two last terms are regularization terms that penalize the negative values of $\mathbf{A}$ and $\boldsymbol{s}$. Note that this criterion is similar to the one minimized in the PMF method (*Positive matrix factorization*) [7]. But our approach can be seen as a generalization of the PMF method since the regularization parameters differ from one source to another.

The separation is achieved by solving the following optimization problem:

$$\left(\hat{\boldsymbol{s}}, \hat{\mathbf{A}}\right) = \arg\min_{\boldsymbol{s},\mathbf{A}} \Phi\left(\boldsymbol{s}, \mathbf{A}|\boldsymbol{\theta}\right). \tag{8}$$

Our strategy to perform this optimization is to use an alternating iterative descent procedure, updating, at each iteration r, the source estimate $\hat{\boldsymbol{s}}^{(r+1)}$ using the latest estimate of $\mathbf{A}$, then the mixing matrix estimate $\hat{\mathbf{A}}^{(r+1)}$ using the latest estimate of $\boldsymbol{s}$. The minimization at each step is carried out using a relative gradient based algorithm [1]:

$$\begin{cases} \hat{\boldsymbol{s}}^{(r+1)} &= \hat{\boldsymbol{s}}^{(r)} - \mu_s^{(r+1)}\,\nabla_s\Phi\left(\boldsymbol{s}^{(r)}, \hat{\mathbf{A}}^{(r)}\right) \odot \hat{\boldsymbol{s}}^{(r)},\\ \hat{\mathbf{A}}^{(r+1)} &= \hat{\mathbf{A}}^{(r)} - \mu_a^{(r+1)}\,\nabla_A\Phi\left(\hat{\boldsymbol{s}}^{(r+1)}, \mathbf{A}^{(r)}\right) \odot \hat{\mathbf{A}}^{(r)}, \end{cases}$$

where $\odot$ represents the point–wise multiplication, $\mu_s^{(r+1)}$ and $\mu_a^{(r+1)}$ are positive learning parameters that control the update rate. A golden section search method is used at each iteration to find the optimal value of these learning parameters. $\nabla_s\Phi$ and $\nabla_A\Phi$ are the Gradient of the criterion with respect to $\boldsymbol{s}$ and $\mathbf{A}$ expressed as:

$$\begin{aligned}\nabla_s\Phi\big(\boldsymbol{s}, \mathbf{A}\big) &= -\frac{1}{\sigma^2}\mathbf{A}^T\left(\boldsymbol{x} - \mathbf{A}\boldsymbol{s}\right) + \mathbf{B} + \mathbf{F} \oslash \boldsymbol{s},\\ \nabla_A\Phi\big(\boldsymbol{s}, \mathbf{A}\big) &= -\frac{1}{\sigma^2}\left[\boldsymbol{x} - \mathbf{A}\boldsymbol{s}\right]\,\boldsymbol{s}^T + \mathbf{L} + \mathbf{G} \oslash \mathbf{A}.\end{aligned}$$

The symbol $\oslash$ stands for point–wise division and the matrices $\mathbf{B}, \mathbf{F}, \mathbf{G}, \mathbf{L}$ are obtained by:

$$\begin{aligned}\mathbf{B} &= [\beta_1;\ldots;\beta_n]^T \otimes \mathbf{1}_{1\times N},\\ \mathbf{F} &= [1-\alpha_1;\ldots;1-\alpha_n]^T \otimes \mathbf{1}_{1\times N},\\ \mathbf{G} &= [\gamma_1;\ldots;\gamma_n]^T \otimes \mathbf{1}_{1\times m},\\ \mathbf{L} &= [1-\lambda_1;\ldots;1-\lambda_n]^T \otimes \mathbf{1}_{1\times m},\end{aligned}$$

where $\otimes$ represents the kronecker product and $\mathbf{1}_{p\times q}$ a $p \times q$ ones matrix.

2.3. Hyperparameter Assessment

In practice, the hyperparameters are not available. Therefore, for an unsupervised learning, one has to estimate them from the data. In this paper, the noise variance and the Gamma distribution parameters are estimated as follows:

$a)$ *Noise variance*

The estimated sources, mixing matrix and the measured data being given, the noise variance can be estimated by maximizing the posterior distribution $\pi(\sigma|\boldsymbol{x}, \mathbf{A}, \boldsymbol{s})$ which has the following expression:

$$\begin{aligned}\pi\left(\sigma^{-2}|\boldsymbol{x}, \mathbf{A}, \boldsymbol{s}\right) \propto \left(\frac{1}{\sigma^2}\right)^{\frac{mN}{2}} \exp\left\{-\frac{1}{2\sigma^2}\|\boldsymbol{x} - \mathbf{A}\boldsymbol{s}\|^2\right\}\\ \times\, p\left(\sigma^{-2}\right).\end{aligned} \tag{9}$$

The prior for the noise variance σ^2 is an inverse Gamma, which corresponds to assigning a Gamma distribution for σ^{-2}:

$$\sigma^{-2} \sim \mathcal{G}(\alpha_\sigma^o, \beta_\sigma^o), \tag{10}$$

leading to an *a posteriori* given by:

$$(\sigma^{-2}|\boldsymbol{x}, \mathbf{A}, \boldsymbol{s}) \sim \mathcal{G}(\alpha_\sigma^{post}, \beta_\sigma^{post}), \tag{11}$$

$$\alpha_\sigma^{post} = \alpha_\sigma^o + \frac{mN}{2}, \tag{12}$$

$$\beta_\sigma^{post} = \beta_\sigma^o + \frac{1}{2}\|\boldsymbol{x} - \mathbf{A}\boldsymbol{s}\|^2, \tag{13}$$

then the maximum is reached for :

$$\left(\hat{\sigma}^{-2}\right)^{(r+1)} = \frac{\alpha_\sigma^o + \frac{mN}{2} - 1}{\beta_\sigma^o + \frac{1}{2}\left\|\boldsymbol{x} - \mathbf{A}^{(r+1)}\boldsymbol{s}^{(r+1)}\right\|^2}. \tag{14}$$

The parameters α_σ^o, β_σ^o are chosen according to an *a priori* noise level and variance. Note that this approach transforms the original problem of choosing σ^2 in that of choosing $(\alpha_\sigma^o, \beta_\sigma^o)$. But the point is that this last choice is by no way as crucial as the choice of σ^2 is.

b) Source hyperparameters $\{\alpha_j, \beta_j\}_{j=1}^n$

The estimated sources being given, their associated Gamma distribution parameters $\{\alpha_j, \beta_j\}_{j=1}^n$ are estimated as follows:

The posterior distribution $\pi(\beta_j|s_j)$ is given by:

$$\pi(\beta_j|s_j, \alpha_j) \propto \beta_j^{N\alpha_j} \exp\left\{-\beta_j \sum_{t=1}^N s_j(t)\right\} \times p(\beta_j). \tag{15}$$

Therefore, one can note that the conjugate prior for the parameter β_j is a Gamma density:

$$\beta_j \sim \mathcal{G}(\alpha_{\beta_j}^o, \beta_{\beta_j}^o), \tag{16}$$

leading to an *a posteriori* Gamma distribution:

$$(\beta_j|s_j(t), \alpha_j) \sim \mathcal{G}(\alpha_{\beta_j}^{post}, \beta_{\beta_j}^{post}), \tag{17}$$

with parameters:

$$\alpha_{\beta_j}^{post} = \alpha_{\beta_j}^o + N\alpha_j + 1, \tag{18}$$

$$\beta_{\beta_j}^{post} = \beta_{\beta_j}^o + \sum_{t=1}^N s_j(t). \tag{19}$$

The maximum is then reached for:

$$\hat{\beta}_j^{(r+1)} = \frac{\alpha_{\beta_j}^o + N\hat{\alpha}_j^{(r)}}{\beta_{\beta_j}^o + \sum_{t=1}^N s_j^{(r+1)}(t)} \tag{20}$$

For the hyperparameter $\{\alpha_j\}_{j=1}^n$ assessment, we consider $\mu_j = \alpha_j/\beta_j$. The law $\pi\left(\alpha_j|s_j, \mu_j\right)$ takes the form:

$$\pi\left(\alpha_j|s_j, \mu_j\right) = \prod_{t=1}^N \frac{\alpha_j^{\alpha_j}}{\mu_j^{\alpha_j}\Gamma(\alpha_j)}\, s_j^{\alpha_j - 1}(t) \times \exp\left\{-\frac{\alpha_j}{\mu_j} s_j(t)\right\} p(\alpha_j). \tag{21}$$

By assigning a Gamma prior for α_j of parameters $\alpha_{\alpha_j}^o$ and $\beta_{\alpha_j}^o$, this posterior density takes the form:

$$\pi\left(\alpha_j|s_j, \mu_j\right) = \frac{\alpha_j^{N\alpha_j}}{\mu_j^{N\alpha_j}\Gamma^N(\alpha_j)} \prod_{t=1}^N s_j(t)^{\alpha_j - 1} \times \exp\left\{-\frac{\alpha_j}{\mu_j}\sum_{t=1}^N s_j(t)\right\} \alpha_j^{\alpha_{\alpha_j}^o - 1} \exp\left\{-\beta_{\alpha_j}^o \alpha_j\right\}, \tag{22}$$

The maximization of this density and using a second order approximation of the first derivative of $\log\Gamma(\alpha_j)$:

$$\frac{d\log\Gamma(\alpha_j)}{d\alpha_j} = \log\alpha_j - \frac{1}{2\alpha_j} - \frac{1}{12\alpha_j^2} + ..., \tag{23}$$

yields the MAP estimate of $\{\alpha_j\}_{j=1}^n$:

$$\hat{\alpha}_j^{(r+1)} = \frac{\frac{N}{2} + \alpha_{\alpha_j}^o - 1}{\frac{\alpha_{\alpha_j}^o}{\beta_{\alpha_j}^o} + \sum_{t=1}^N \left[\frac{s_j^{(r+1)}(t)}{\mu_j^{(r)}} - \log\frac{s_j^{(r+1)}(t)}{\mu_j^{(r)}} - 1\right]}, \tag{24}$$

c) Mixing coefficient hyperparameters $\{\alpha_j, \lambda_j\}_{j=1}^n$

Since the mixing coefficients are also assigned by gamma densities as prior laws, their hyperparameters are estimated by generalizing the results obtained for the sources:

$$\hat{\gamma}_j^{(r+1)} = \frac{\alpha_{\gamma_j}^o + m\,\hat{\lambda}_j^{(r)}}{\beta_{\gamma_j}^o + \sum_{i=1}^m a_{ij}^{(r+1)}}, \tag{25}$$

$$\hat{\lambda}_j^{(r+1)} = \frac{\frac{m}{2} + \alpha_{\lambda_j}^o - 1}{\frac{\alpha_{\lambda_j}^o}{\beta_{\lambda_j}^o} + \sum_{i=1}^m \left[\frac{a_{ij}^{(r+1)}}{\nu_j^{(r)}} - \log\frac{a_{ij}^{(r+1)}}{\nu_j^{(r)}} - 1\right]}, \tag{26}$$

where $\nu_j^{(r)} = \lambda_j^{(r)}/\gamma_j^{(r)}$.

3. EXPERIMENT

To illustrate the method applicability, we consider an simulation example which consists in analyzing a mixture of three sources. The mixture is obtained by constructing three synthetic spectra and considering nineteen measures with mixing coefficients chosen in such a way to have a realistic evolution. Gaussian noise is added to have a signal to noise ratio equal to 50 dB. Figure 1 shows the resulting mixture. To discuss the result accuracy, we use the global system matrix $\mathbf{G} = \hat{\mathbf{A}}^{-1}\mathbf{A}$ which should tend to a permutation (due to the order indetermination) of the identity matrix when the sources are separated correctly. The empirical source covariance matrix is:

$$\hat{\mathbf{R}}_{\boldsymbol{s}} = \begin{bmatrix} 1.000 & 0.516 & 0.386 \\ 0.516 & 1.000 & -0.105 \\ 0.386 & -0.105 & 1.000 \end{bmatrix}. \tag{27}$$

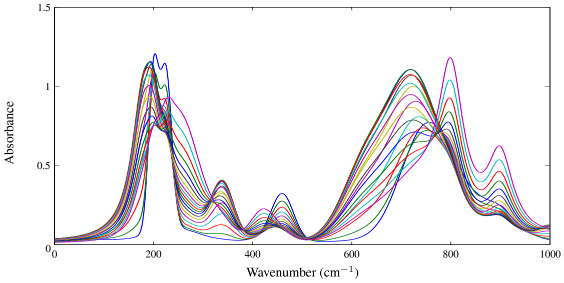


Fig. 1: Mixture data

The off-diagonal terms of this covariance matrix are non null. We deduce that the available samples of the sources are spatially correlated, so the independence assumption is not sufficient for the spectra reconstruction. This explains the failure in applying directly an ICA algorithm, which gives sources and mixing coefficients having negative values. This is confirmed by the global system matrix resulting from the analysis by JADE algorithm [1]:

$$\mathbf{G} = \begin{bmatrix} -0.499 & 0.836 & 1.030 \\ 1.263 & -0.412 & -0.280 \\ -0.127 & 0.856 & -0.480 \end{bmatrix}. \qquad (28)$$

The results obtained by applying the proposed method to the mixture analysis are presented in figure 2. We can see that source spectra and mixing coefficients are estimated without apparition of negative values. Concerning the separation performances, the resulting global system matrix is:

$$\mathbf{G} = \begin{bmatrix} 1.028 & -0.027 & -0.011 \\ 0.014 & 0.996 & 0.137 \\ -0.018 & 0.089 & 1.020 \end{bmatrix}, \qquad (29)$$

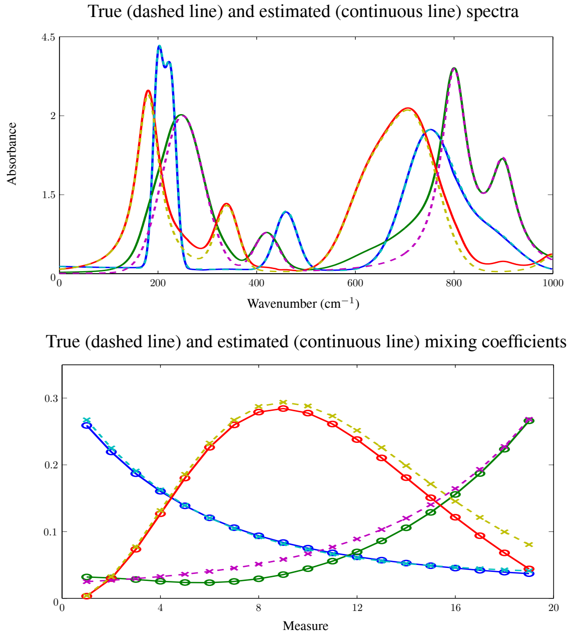
which is very close to the identity matrix.

4. CONCLUSION

In this paper, the Bayesian theory for source separation has been applied to the particular case of positive sources and mixing coefficients. The non-negativity has been considered explicitly by assigning Gamma density as priors for the sources and for the mixing coefficients. We showed the superior performances of the proposed method compared to the classical JADE algorithm. Future works concern comparing this method performances with that of available algorithms such as NMF, PMF and ALS.

5. ACKNOWLEDGEMENTS

This work is supported by the ”Région Lorraine” and the CNRS. The authors are very indebted to Dr. Cédric Carteret and Pr. Bernard Humbert, from the Laboratory of Chemistry Physics and Microbiology for the Environment (LCPME, UHP), for insightful discussions.


True (dashed line) and estimated (continuous line) spectra

True (dashed line) and estimated (continuous line) mixing coefficients

Fig. 2: Mixture analysis results

6. REFERENCES

[1] J. F. Cardoso, “Blind signal separation: statistical principles,” *Proceedings of the IEEE*, vol. 9, no. 10, pp. 2009–2025, 1998.

[2] A. Mohammad-Djafari, “A Bayesian approach to source separation,” in 19th *International workshop on maximum entropy and bayesian methods (MaxEnt 99)*, Boise, Idaho, USA, 1999.

[3] A. Hyvärinen, J. Karhunen, and E. Oja, *Independent Component Analysis*, John Wiley, New York, 2001.

[4] M. D. Plumbley, “Algorithms for non–negative independent component analysis,” *IEEE Transactions on Neural Networks*, vol. 14, no. 3, pp. 534–543, 2003.

[5] D. D. Lee and H. S. Seung, “Learning the parts of objects by non–negative matrix factorization,” *Nature*, vol. 401, pp. 788–791, 1999.

[6] R. Tauler, A. Izquierdo-Ridorsa, and E. Casassas, “Simultaneous analysis of several spectroscopic titrations with self-modelling curve resolution,” *Chemometrics and Intelligent Laboratory Systems*, vol. 18, no. 3, pp. 293–300, 1993.

[7] P. Paatero and U. Tapper, “Positive matrix factorization: A non–negative factor model with optimal utilization of error estimates of data values,” *Environmetrics*, vol. 5, pp. 111–126, 1994.