NON-NEGATIVE SOURCE SEPARATION USING THE MAXIMUM LIKELIHOOD APPROACH

Saïd Moussaoui, David Brie

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

Cédric Carteret

LCPME, UMR 7564 UHP-CNRS
3, rue de Vandoeuvre, 54600, Villers-lès-Nancy, France
carteret@lcpe.cnrs-nancy.fr

ABSTRACT

This papers addresses the problem of non-negative source separation using the maximum likelihood approach. It is shown that this approach can be effective by considering that the sources are distributed according to a density having a non-negative support from which an adequate non-linear separating function can be derived. In the particular of spectroscopic data which is our main concern, a good candidate is the Gamma distribution which allows to encode both non-negativity and sparsity of the source signals. Numerical experiments are used to assess the performances of the method.

1. INTRODUCTION

This work is motivated by the need to apply blind source separation methods to the processing of spectral data sets resulting from the analysis of multicomponent chemical substances by spectroscopy. The analysis aims at identifying the chemical composition of the substances and at evaluating the amount of each pure component. According to Beer-Lambert law, the measured data are a linear combination of the unknown pure component spectra. This is a source separation problem where the source signals correspond to the pure component spectra and the amounts of these components are deduced from the mixing coefficients. Both the source signals and the mixing coefficients are non-negative, so the problem is referred to as a non-negative source separation.

The observation model assumes that m observed signals, noted $\boldsymbol{x}_t = [x_1(t), \ldots, x_m(t)]^T$, are a non-negative linear combination of p non-observable non-negative source signals $\boldsymbol{s}_t = [s_1(t), \ldots, s_p(t)]^T$

$$\boldsymbol{x}_t = \boldsymbol{A}\,\boldsymbol{s}_t + \boldsymbol{e}_t, \tag{1}$$

where $\boldsymbol{A}$ is the $(m \times p)$ mixing matrix whose columns define an evolution profile of the amount of each source in the mixtures and $\boldsymbol{e}_t$ is a vector of measurement errors and model uncertainties. Having n samples, $(t = 1, \ldots, n)$, and using matrix notations, the whole mixing model is expressed as

$$\boldsymbol{X} = \boldsymbol{A}\,\boldsymbol{S} + \boldsymbol{E}, \tag{2}$$

where $\boldsymbol{X}$ is the $(m \times n)$ data matrix, $\boldsymbol{S}$ is a $(p \times n)$ matrix containing the source signals in its rows. By assuming a known number of components, the separation problem consists in estimating the source signals and the mixing coefficients that reproduce the data according to the mixing model (2) and fulfil the non-negativity constraints.

To address this problem several approaches are possible. The first one is based on non-negative least squares estimation using either alternate optimization techniques [1, 2] or non-negative matrix factorization methods [3]. However, this approach suffers from the problem of non-uniqueness of the solution, since considering only the non-negativity constraint does not ensure the uniqueness of the decomposition, unless under some restrictive conditions [4, 5]. Penalized least squares estimation methods allow to select one particular solution, among the admissible ones, by considering additional assumptions [6, 7, 8]. Assuming the mutual independence of the non-negative sources leads to non-negative independent component analysis [9] and the proposed algorithms are suitable for *well grounded* and orthogonal sources [10]. As pointed in [11], using an ICA method enforcing source orthogonality such as JADE [12] may lead to negative estimates when the available samples of the source signals present a spatial correlation.

The maximum likelihood source separation approach does not impose source orthogonality but requires the choice of appropriate non-linear functions to get the most independent sources whose statistical distributions are close to the density functions from which the non-linear functions are derived. In this paper we consider the case of Gamma distributions which are good models for spectral data allowing to encode both non-negativity as well as sparsity of the sources [8, 13].

2. MAXIMUM LIKELIHOOD NON-NEGATIVE SOURCE SEPARATION

The case of noise-free observation model and a square mixing ($m = p$) is considered. If the number of observations is greater than the number of sources ($m > p$), a dimension reduction is necessary in order to use the maximum likelihood approach. This reduction can be performed using a $(p \times m)$ random transformation matrix before applying the maximum likelihood approach.

2.1. Problem Statement

Under the assumptions of mutually independent and i.i.d (independent and identically distributed) source signals, the maximum likelihood estimate of the separating matrix, $\boldsymbol{B} = \boldsymbol{A}^{-1}$ is defined as

$$\hat{\boldsymbol{B}} = \arg\max_{\boldsymbol{B}} p(\boldsymbol{X}|\boldsymbol{B}), \tag{3}$$

or equivalently as [14, 15]

$$\hat{\boldsymbol{B}} = \arg\min_{\boldsymbol{B}} \Big\{ -\log|\det(\boldsymbol{B})| - \frac{1}{n}\sum_{t=1}^{n}\sum_{j=1}^{p} \log p_{s_j}\big([\boldsymbol{B}\,\boldsymbol{x}_t]_j\big) \Big\}, \tag{4}$$

The optimization of this criterion can be achieved by the relative/natural gradient algorithm [16, 17] which leads to the iterative update rule

$$\hat{\boldsymbol{B}}^{(r+1)} = \hat{\boldsymbol{B}}^{(r)} - \mu \left[\frac{1}{n}\sum_{t=1}^{n} \boldsymbol{\psi}\left(\hat{\boldsymbol{s}}_t^{(r)}\right)\left(\hat{\boldsymbol{s}}_t^{(r)}\right)^T - \boldsymbol{I} \right] \hat{\boldsymbol{B}}^{(r)}, \tag{5}$$

where $\hat{\boldsymbol{s}}_t^{(r)} = \hat{\boldsymbol{B}}^{(r)}\boldsymbol{x}_t$ and $\boldsymbol{\psi}^T(\boldsymbol{z}) = [\psi_1(z_1), ..., \psi_n(z_n)]$ and the non-linear separating functions $\psi_j(z)$ are defined by $\psi_j(z) = -\dfrac{\partial}{\partial z}\log p_{z_j}(z)$. Ideally, the non-linear functions should be obtained from the probability density functions of the source signals [15, 16]. However, unless in some applications [18], these distributions are not known *a priori*. So, they are approximated using either high order statistics [14] or a parametric density function model [15, 19]. The choice of a density model allows to specify some information on the structure of the source signal distributions. Example of parametric models are the generalized Gaussian distribution, the Student distribution and the mixture of Gaussians. Taking generalized Gaussian and Student distributions implicitly assumes unimodal symmetrical source signals and the mixture of Gaussians allows to consider multimodal and asymmetrical distributions. However, none of these models allows to explicitly account for the non-negativity.

2.2. Synthesis of an Appropriate Non-linear Function

To separate non-negative sources, the non-linear function may be obtained through an assignment of a parametric non-negative support density function to the statistical distribution of the sources. However, the choice of a particular model needs additional informations or assumptions on the source signals. In the case of spectral data, one can assume sparse non-negative sources which can be represented by a Gamma distribution model. The Gamma density is defined by

$$p(s|\alpha,\beta) = \frac{\beta^\alpha}{\Gamma(\alpha)}\, s^{\alpha-1}\, \exp\{-\beta s\}\, \mathbb{I}_{s\geq 0}, \tag{6}$$

where $\alpha > 0$ and $\beta > 0$ are its shape and rate parameters and $\mathbb{I}_{s\geq 0}$ is the indicator function. The choice of this distribution is motivated by two advantages ; firstly it explicitly accounts for non-negativity since $p(s < 0) = 0$ and its parameters allow to adjust its asymmetrical shape to fit the distribution of the source signals. The non-linear function deduced from this model is given by

$$\psi(s;\alpha,\beta) = \beta - (\alpha-1)\, s^{-1}, \quad \text{for } s > 0. \tag{7}$$

Since during the optimization negative intermediate estimates may occur, this function is rectified as

$$\psi(s;\alpha,\beta) = \beta - (\alpha-1)/\max(s,\epsilon), \quad \forall\, s \in \mathbb{R}. \tag{8}$$

where ϵ is set to a small value (10^{-3}). An example of Gamma distribution shape and the corresponding non-linear function are shown in figure 1.

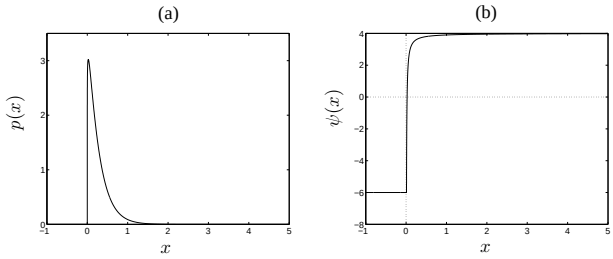

Fig. 1. (a) Gamma distribution for $\alpha = 1.1$ and $\beta = 4$ and (b) related non-linear function for $\epsilon = 0.01$.

2.3. Algorithm Stability

According the results of [20, 21] a sufficient condition for the algorithm stability is

$$\mathrm{E}[\psi'(\tilde{s}_j)\tilde{s}_j^2] - \mathrm{E}[\psi(\tilde{s}_j)\tilde{s}_j] > 0 \quad \forall\, j = 1, \ldots, p \tag{9}$$

where $\mathrm{E}[f(\tilde{s}_j)]$ are the non-linear moments of the centered sources

$$\tilde{s}_j = s_j - \mathrm{E}[s_j] = s_j - \alpha_j/\beta_j.$$

The calculation of all these moments yields

$$\mathrm{E}[\psi'(\tilde{s}_j)\tilde{s}_j^2] = \mathrm{E}\left[\frac{(\alpha_j - 1)}{\tilde{s}_j^2} \cdot \tilde{s}_j^2\right] = (\alpha_j - 1); \qquad (10)$$

and

$$\mathrm{E}[\psi(\tilde{s}_j)\tilde{s}_j] = \mathrm{E}\left[\beta_j \tilde{s}_j - (\alpha_j - 1)\right] = (1 - \alpha_j). \qquad (11)$$

Consequently, a sufficient condition for the algorithm stability is

$$\alpha_j > 1, \quad j = 1, \ldots, p. \qquad (12)$$

From this stability condition, it turns out that, due to the optimization method used, only Gamma distribution having shape parameter $\alpha > 1$ can be considered. We believe this is the main shortcoming of such an optimization scheme. We come back to that point in section 3.3.

2.4. Estimation of the Gamma distribution parameters

In order to handle the problem of choosing manually the values of $\boldsymbol{\theta} = [\alpha_1, \ldots, \alpha_p, \beta_1, \ldots, \beta_p]^T$, these parameters are estimated jointly with the separating matrix by maximizing the joint likelihood

$$\left(\hat{\boldsymbol{B}}, \hat{\boldsymbol{\theta}}\right) = \arg\max_{\boldsymbol{B}, \boldsymbol{\theta}} p(\boldsymbol{X}|\boldsymbol{B}, \boldsymbol{\theta}), \qquad (13)$$

The minimization of the resulting criterion is obtained by a gradient based algorithm which updates at each iteration the separating matrix estimate using the last estimate of the parameters and uses this new estimate to update the prior model parameters using a gradient descent algorithm. The separating matrix being updated using equation 5, the Gamma prior hyperparameters are updated as follows: let $\{s(t)\}_{t=1}^n$ a sequence of non-negative random variables assumed to be Gamma distributed. We want to estimate the parameters of the Gamma density (α, β) from the realization $\boldsymbol{s} = \{s(t)\}_{t=1}^n$. The likelihood is expressed as

$$p(\boldsymbol{s}|\alpha, \beta) = \left(\frac{\beta^\alpha}{\Gamma(\alpha)}\right)^n \prod_{t=1}^n \left(s_t^{\alpha-1}\right) \exp\left[-\sum_{t=1}^n \beta s(t)\right], \qquad (14)$$

and the related criterion

$$\mathcal{J}(\boldsymbol{s}|\alpha, \beta) = -\frac{1}{n} \log p(\boldsymbol{s}|\alpha, \beta), \qquad (15)$$

is given by

$$\mathcal{J}(\boldsymbol{s}|\alpha, \beta) = -\alpha \log \beta + \log \Gamma(\alpha) - (\alpha - 1)\frac{1}{n}\sum_{t=1}^n \log s(t) + \frac{1}{n}\sum_{t=1}^n \beta s(t). \qquad (16)$$

At each r-th iteration of the algorithm, the parameters are updated using the gradient algorithm according to

$$\begin{cases} \alpha^{(r+1)} = \alpha^{(r)} - \rho_\alpha \nabla_\alpha \mathcal{J}(\boldsymbol{s}|\alpha^{(r)}, \beta^{(r)}), \\ \beta^{(r+1)} = \beta^{(r)} - \rho_\beta \nabla_\beta \mathcal{J}(\boldsymbol{s}|\alpha^{(r+1)}, \beta^{(r)}), \end{cases} \qquad (17)$$

where $(\rho_\alpha, \rho_\beta)$ are positive learning parameters and the gradients $(\nabla_\alpha, \nabla_\beta)$ are given as

$$\begin{cases} \nabla_\alpha \mathcal{J}(\boldsymbol{s}|\alpha, \beta) = -\log \beta + \Psi(\alpha) - \frac{1}{n}\sum_{t=1}^n \log s(t) \\ \nabla_\beta \mathcal{J}(\boldsymbol{s}|\alpha, \beta) = -\frac{\alpha}{\beta} + \frac{1}{n}\sum_{t=1}^n s(t), \end{cases}$$

where $\Psi(\alpha) = \frac{d}{d\alpha}\Gamma(\alpha)$ is the digamma function [22]. According to the algorithm stability conditions 12 and the definition of the Gamma density, the values of α_j and β_j are forced to satisfy $\alpha_j \geq (1 + \epsilon)$ and $\beta_j \geq \epsilon$ where ϵ is set to a small value (10^{-3}).

3. NUMERICAL EXAMPLE

The source signals and mixing coefficients are simulated in such a way to get mixture signals similar to spectometric data sets. Each source signal is obtained by a superposition of $K = 15$ of Gaussian and Lorentzian shapes with different randomly chosen parameters (location, amplitude and width). The mixing coefficients are also simulated to get evolution profiles similar to what we get in kinetic reactions. Figure 2 shows an example of $p = 3$ normalized (unit variance) source signals of $n = 1000$ samples and the mixing coefficient evolution profiles for $m = 3$ observations.

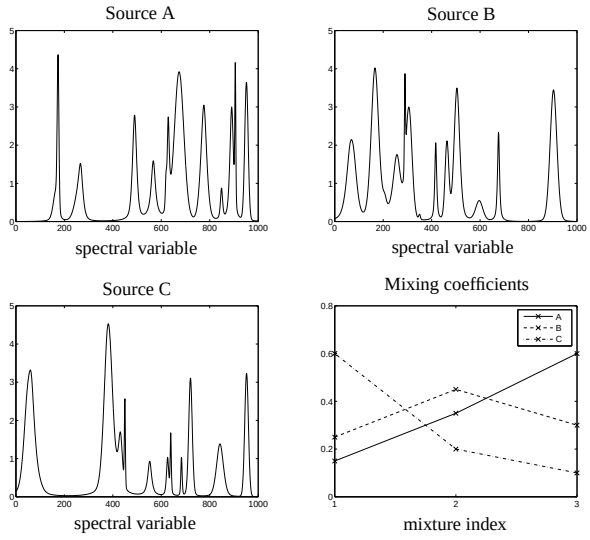

Fig. 2. Simulated source signals and mixing coefficients.

3.1. Performance Measures

To assess the separation performances of the tested algorithms, we use the normalized mean square error, noted $\mathcal{E}_{s_j}$, and defined by

$$\mathcal{E}_{s_j} = \left(\sum_{t=1}^{n} \left(s_j(t) - \hat{s}_j(t) \right)^2 \right) \Big/ \left(\sum_{t=1}^{n} s_j(t)^2 \right), \quad (18)$$

where $\hat{s}_j$ is the estimate of the j-th source s_j. This index measures the quality of the source reconstruction. To asses the estimation quality of the mixing matrix, we use the performance index noted $\mathcal{PI}$ and defined by

$$\mathcal{PI} = \frac{1}{2p\,(p-1)} \sum_{i=1}^{p} \left\{ \left(\sum_{k=1}^{p} \frac{|g_{ik}|^2}{\max\limits_{\ell} |g_{i\ell}|^2} - 1 \right) + \left(\sum_{k=1}^{p} \frac{|g_{ki}|^2}{\max\limits_{\ell} |g_{\ell i}|^2} - 1 \right) \right\}, \quad (19)$$

where g_{ij} is the (i,j)-th element of the matrix $\boldsymbol{G} = \hat{\boldsymbol{B}}\boldsymbol{A}$.

3.2. Noise-free Case

The empirical covariance matrix of these sources

$$\hat{\boldsymbol{R}}_s = \begin{bmatrix} 1.00 & 0.13 & -0.18 \\ 0.13 & 1.00 & -0.22 \\ -0.18 & -0.22 & 1.00 \end{bmatrix}, \quad (20)$$

is not diagonal showing that the available samples of the sources present a significant spatial correlation, therefore applying a separation algorithm enforcing orthogonality leads to an incorrect solution, as illustrated in figure 3 where negative estimates obtained with JADE algorithm [12] can be seen. Due to the non orthogonality of the available samples of the sources the NNICA algorithm may also lead to negative estimates, as illustrated in figure 4.

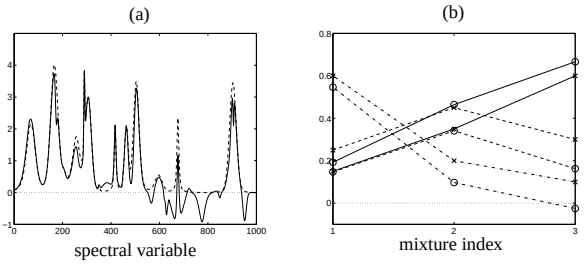

Fig. 3. JADE results: (a) Estimated second source signal (the true source is shown in dotted line) and (b) estimated (circles) and true (cross) mixing coefficients.

On the other hand, applying methods, such as NMF, which are based only on non-negativity do not provide a unique solution since the necessary conditions for the uniqueness (see [5]) are not satisfied (it can be checked out in the mixing coefficient profiles since there in no zero coefficient). Figure 5 shows some admissible solutions obtained by several different initializations of the NMF algorithm. The point is that we cannot know in advance the one which will be recovered.

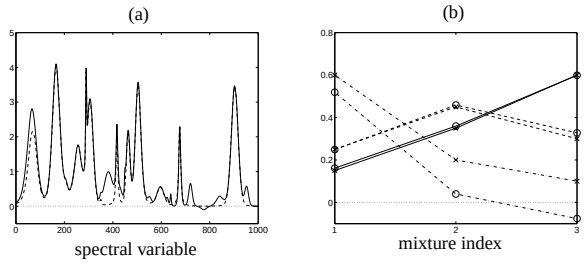

Fig. 4. NNICA results: (a) Estimated second source signal (the true source is shown in dotted line) and (b) estimated (circles) and true (cross) mixing coefficients.

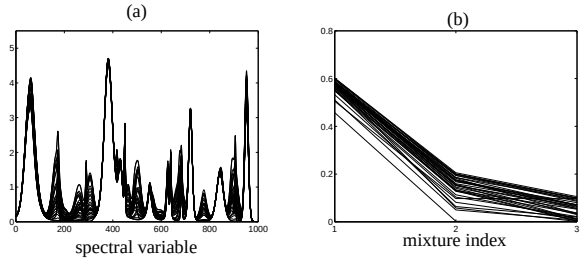

Fig. 5. NMF results: (a) Estimates of the third source signal and (b) estimated mixing coefficients of the third source.

Figure 6 shows the estimation results using the maximum likelihood approach with a Gamma distribution model (the method is termed as MLPSS for *maximum likelihood positive source separation*). It appears that the non-negativity of the mixing coefficients is ensured even if this constraint has not been taken into account by the separation algorithm. This result can be explained by two points. Firstly, the distribution of the source signals is well fitted by the gamma distribution resulting in a correct estimation of the sources. Secondly, the mixing model is noise free, so a good estimation of the source signals ensures the non-negativity of the mixing coefficients.

The performances of the applied methods are summarized in table 1 and the comparison shows that, in the case considered, the maximum likelihood separation approach with gamma distribution model leads to better performances.

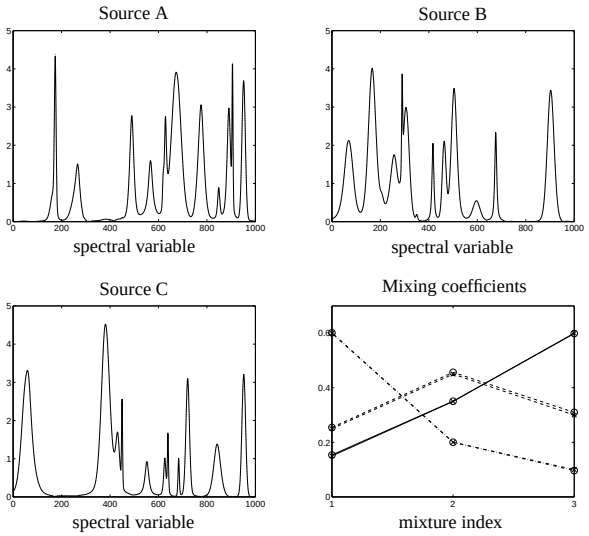

Fig. 6. MLPSS results: Estimated source signals and mixing coefficients using the maximum likelihood approach (the true source signals are shown in dotted lines). The algorithm was run with $\mu = 0.001, \rho_\alpha = \rho_\beta = 0.01$ and iterated for $r_{max} = 10000$ iterations.

Table 1. Comparison of the performances. The performance indexes are expressed in decibels (dB)

	JADE	NMF	NNICA	MLPSS
$\mathcal{E}_A$	-10.58	-16.20	-15.11	-35.88
$\mathcal{E}_B$	-10.96	-15.88	-14.19	-42.25
$\mathcal{E}_C$	-17.40	-14.14	-39.41	-40.30
$\mathcal{PI}$	-12.06	-15.67	-15.90	-38.32

3.3. Case of Noisy Observations

In order to discuss the performances of the method with respect to the noise level, a random Gaussian noise is added to the mixtures. The noise level is assessed using the signal to noise ratio (SNR) defined for each i-th mixture as the ratio between the variances of the noise-free mixture signal and the i-th noise sequence. Figure 7 shows the evolution the performance indexes reached by the NNICA, MLPSS and BPSS (method based on assigning a Gamma distribution prior on both the source signals and mixing coefficients and estimation using Monte Carlo Markov chain sampling [23]) algorithms with respect to the noise level. For high SNR, the use of a Gamma distribution prior either using maximum likelihood or Bayesian estimation approaches perform better than NNICA. In spite of the fact that BPSS and MLPSS use the same prior model on the source signals, they yield different results for high SNR which can be explained by the restriction of the Gamma distribution shape parameter to be greater than unity in order to ensure the gradient algorithm stability. When the noise level increases, the performance of the MLPSS decreases since the maximum likelihood approach does not take into account the additive noise in the observation model. To overcome these limitations, an alternative method would be to use expectation-maximization algorithms, allowing to jointly estimate the mixing coefficients and the noise covariance matrix and to consider sources with Gamma shape parameter $0 < \alpha < 1$.

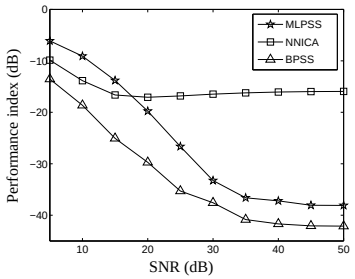

Fig. 7. Influence of the noise level on the performance index reached by NNICA, MLPSS and BPSS methods.

4. CONCLUSION

The aim of this paper was to show that the maximum likelihood approach can be applied successfully to the separation of non-negative sources, provided that adequate non-linear functions are used. The non-negativity of the source signals is accounted through the synthesis of a non-linear function derived from non-negative support probability density functions as prior models of their distributions. In the case of spectral data, the source signals are non-negative and sparse. These informations are accounted by using a Gamma distribution model which due to the stability condition should be constrained to have a shape parameter $\alpha > 1$. The resulting algorithm is illustrated through a simulation example in the noise-free case and compared with other available approaches. The results of the separation using the proposed approach are good when the actual source distributions can be well approximated by a Gamma density, as it the case for spectral data. If not, other distributions should be considered. The analysis of the performances in the case of noisy observations pointed out the need to use alternative methods allowing to jointly estimate the mixing coefficients and the source signals. In this context, expectation-maximization algorithms can be used to account for the noisy observation model and a Bayesian source separation approach can be used to incorporate the mixing coefficient non-negativity.

Acknowledgments

This work is supported by the ”Region Lorraine” and the CNRS.

5. REFERENCES

[1] R. Tauler, B. Kowalski, and S. Fleming, "Multivariate curve resolution applied to spectral data from multiple runs of an industrial process," *Analytical Chemistry*, vol. 65, pp. 2040–2047, 1993.

[2] R. Bro and S. De Jong, "A fast non-negativity constrained least squares algorithm," *Journal of Chemometrics*, vol. 11, pp. 393–401, 1997.

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

[4] D. Donoho and V. Stodden, "When does non-negative matrix factorization give a correct decomposition into parts?" in *proceedings of NIPS*, 2003.

[5] S. Moussaoui, D. Brie, and J. Idier, "Non-negative source separation: Range of admissible solutions and conditions for the uniqueness of the solution," in *proceedings of ICASSP*, 2005, pp. 289–292.

[6] K. Sasaki, S. Kawata, and S. Minami, "Estimation of component spectral curves from unknown mixture data," *Applied Optics*, vol. 23, no. 12, pp. 1955–1959, 1984.

[7] P. Paatero, "Least squares formulation of robust non-negative factor analysis," *Chemomerics and Intelligent Laboratory Systems*, vol. 37, pp. 23–35, 1997.

[8] S. Moussaoui, A. Mohammad-Djafari, D. Brie, and O. Caspary, "A Bayesian method for positive source separation," in *proceedings of ICASSP*, 2004, pp. 485–488.

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

[10] ——, "Conditions for non-negative independent component analysis," *Signal Processing Letters*, vol. 9, no. 6, pp. 177–180, 2002.

[11] D. Nuzillard and J. Nuzillard, "BSS applied to non-orthogonal signals," in *proceedings of ICA*, 1999.

[12] J.-F. Cardoso and A. Souloumiac, "Blind beamforming for non Gaussian signals," *IEE Proceedings-F*, vol. 140, no. 6, pp. 362–370, 1993.

[13] S. A. Abdallah and M. D. Plumbley, "Polyphonic transcription by non-negative sparse coding of power spectra," in *Proceedings ISMIR*, 2004, pp. 318–325.

[14] M. Gaeta and J.-L. Lacoume, "Source separation without prior knowledge: The maximum likelihood solution," in *Proceedings of EUSIPCO*, 1990.

[15] D.-T. Pham, P. Garrat, and C. Jutten, "Separation of a mixture of independent sources through a maximum likelihood approach," in *proceedings of EUSIPCO*, 1992.

[16] J.-F. Cardoso and B. H. Laheld, "Equivariant adaptive source separation," *IEEE Transactions on Signal Processing*, vol. 44, no. 12, pp. 3017–3030, 1996.

[17] S.-I. Amari, A. Cichocki, and H. Yang, "A new learning algorithm for blind source separation," in *proceedings of NIPS*, 1996.

[18] A. Belouchrani and J.-F. Cardoso, "Maximum likelihood source separation for discrete sources," in *proceedings of EUSIPCO*, 1994, pp. 768–771.

[19] B. Pearlmutter and L. Parra, "A context-sensitive generalization of ICA," in *proceedings of NIPS*, 1996.

[20] S.-I. Amari, T.-P. Chen, and A. Cichocki, "Stability analysis of adaptive blind source separation," *Neural Networks*, vol. 10, no. 8, pp. 1345–1351, 1997.

[21] J.-F. Cardoso, "On the stability of source separation algorithms," *Journal of VLSI Signal Processing*, vol. 26, pp. 7–14, 2000.

[22] M. Abrahamovitz and I. Stegun, *Handbook of Mathematical Functions*. New York: Dover Publications, 1972.

[23] S. Moussaoui, D. Brie, A. Mohammad-Djafari, and C. Carteret, "Separation of non-negative mixture of non-negative sources using a Bayesian approach and MCMC sampling," 2004, to appear in IEEE Transactions on Signal Processing.