Nonlinear Blind Source Separation Applied to a Simple Bijective Model

Shahram Hosseini[1], Yannick Deville[1], Sonia El Amine[1], and Hicham Saylani[2]

[1] Institut de Recherche en Astrophysique et Planétologie, Université de Toulouse, UPS-CNRS-OMP, 14 Av. Edouard Belin, 31400 Toulouse, France
[2] Laboratoire d'Electronique, de Traitement du Signal et de Modélisation Physique, Faculté des Sciences, Université Ibnou Zohr, BP. 8061, 80000 Agadir, Maroc
{shosseini,ydeville}@irap.omp.eu, sonia_elamine@yahoo.fr, h.saylani@uiz.ac.ma

Abstract. This paper deals with nonlinear Blind Source Separation (BSS) applied to a simple bijective "toy" model. Our objective is to better understand the difficulties encountered in nonlinear BSS, especially when estimating the parameters of mixing or separating structures. The results of this study and the proposed solutions may then be used by the BSS researchers dealing with actual nonlinear physical models. The simulation results confirm the usefulness of our proposed solutions.

1 Introduction

Blind Source Separation (BSS) aims at restoring source signals from their mixtures when the mixing parameters are unknown. While linear BSS has been widely studied, little work is available about nonlinear BSS. It is well known that the independence hypothesis is not sufficient for separating general nonlinear mixtures because of the very large indeterminacies which make the problem ill-posed [1]. A natural idea for reducing these indeterminacies is to constrain the structure of mixing and separating models to belong to a certain set of transformations [2]. Thus, the problem should be studied separately for each considered mixing structure. Even in this simplified case, nonlinear BSS is much more difficult than linear BSS because of the following problems:

1. most nonlinear models are not bijective so that even in the non-blind case when the mixing parameters are known, it is not possible to retrieve the sources in a unique manner without supplementary assumptions,
2. even when the mixing model is known, it is not always possible to find an analytical expression for its inverse,
3. the study of the identifiability and separability of nonlinear mixtures is a hard task and should be done model-by-model to determine which families of source distributions are not separable for each nonlinear model,
4. the blind estimation of the parameters in mixing (or separating) structure is another issue which is generally more difficult than in linear BSS. In particular, the matrix-based estimation algorithms can no longer be used.

A. Elmoataz et al. (Eds.): ICISP 2012, LNCS 7340, pp. 157–165, 2012.

The goal of our paper is the last issue, *i.e.* parameter estimation. The papers addressing this problem may be classified in the following categories[1]:

- the papers considering the models which may be reduced to a linear model using some transformations (*e.g.* [1], [5]). The estimating methods proposed in these papers are especially developed for the particular considered model and cannot be generalized to other models,
- the papers studying non-bijective mixing models (*e.g.* [6], [7]). Since in this case there are several difficult problems to handle simultaneously, these papers do not focus especially on parameter estimation,
- the papers addressing this issue in general, without considering practical examples (*e.g.* [8]).

In this paper, we address the problem in the case of bijective models with known inverse and study in particular a simple "toy" model. Thus, we can focus our efforts on the parameter estimation. Although this model does not fit any known physical system, we believe this study will be useful for the BSS researchers dealing with other actual nonlinear physical models.

2 Problem Statement

Consider the mixing equation $\mathbf{x} = \boldsymbol{\mathcal{F}}(\mathbf{s}, \boldsymbol{\theta}^*)$ where $\mathbf{s} = [s_1, \cdots, s_K]^T$ is the vector of K independent unknown sources, $\mathbf{x} = [x_1, \cdots, x_K]^T$ is the vector of K observations and $\boldsymbol{\mathcal{F}}$ is a bijective parametric function, defined by the unknown parameter vector $\boldsymbol{\theta}^*$. Denote $\boldsymbol{\mathcal{G}}$ the inverse of $\boldsymbol{\mathcal{F}}$ so that $\mathbf{s} = \boldsymbol{\mathcal{G}}(\mathbf{x}, \boldsymbol{\theta}^*)$. BSS may possibly be achieved by constructing the separating model

$$\mathbf{y} = \boldsymbol{\mathcal{G}}(\mathbf{x}, \boldsymbol{\theta}) \tag{1}$$

and looking for a parameter vector $\boldsymbol{\theta}$ which makes the components of $\mathbf{y} = [y_1, \cdots, y_K]^T$ independent. It is clear that $\boldsymbol{\theta}^*$ is one of the solutions which provides the original sources. The other possible solutions depend on the indeterminacies involved in the problem. To make the components of $\mathbf{y}$ independent, we can minimize the mutual information criterion defined as $I = E[\log f_{\mathbf{y}}(\mathbf{y})] - \sum_{k=1}^{K} E[\log f_{y_k}(y_k)]$ where $f_{\mathbf{y}}$ and f_{y_k} are respectively the joint and the marginal probability density functions (pdf) of the variables y_k. Since the model is supposed bijective, (1) yields $f_{\mathbf{y}}(\mathbf{y}) = f_{\mathbf{x}}(\mathbf{x})/|J|$ where J is the Jacobian of the separating model. Then, we obtain

$$I = E[\log f_{\mathbf{x}}(\mathbf{x})] - E[\log(|J|)] - \sum_{k=1}^{K} E[\log f_{y_k}(y_k)] \tag{2}$$

To minimize this criterion using an optimization algorithm we need to compute its gradient and possibly its Hessian with respect to the parameter vector $\boldsymbol{\theta}$. As shown in [1], the gradient reads

[1] In this classification, we do not consider the non model-based papers like [3] and [4].

$$\frac{dI}{d\boldsymbol{\theta}} = -E\left[\frac{1}{J}\frac{dJ}{d\boldsymbol{\theta}}\right] + \sum_{k=1}^{K} E\left[\psi_{y_k}(y_k)\frac{dy_k}{d\boldsymbol{\theta}}\right] \tag{3}$$

where $\psi_{y_k}(y_k) = -d\log f_{y_k}(y_k)/dy_k$ is the score function of y_k. Then, the element (i,j) of the Hessian matrix $\mathbf{H}$ can be obtained as follows:

$$H_{ij} = \frac{d}{d\theta_j}\frac{dI}{d\theta_i} = -E\left[\frac{d}{d\theta_j}(\frac{1}{J}\frac{dJ}{d\theta_i})\right] + \sum_{k=1}^{K} E\left[\frac{d\psi_{y_k}(y_k)}{dy_k}\frac{dy_k}{d\theta_j}\frac{dy_k}{d\theta_i} + \psi_{y_k}(y_k)\frac{d}{d\theta_j}\frac{dy_k}{d\theta_i}\right] \tag{4}$$

The mutual information may be minimized using e.g. the gradient descent algorithm $\boldsymbol{\theta}_{new} = \boldsymbol{\theta}_{old} - \mu\frac{dI}{d\boldsymbol{\theta}}$ or the Newton algorithm $\boldsymbol{\theta}_{new} = \boldsymbol{\theta}_{old} - \mathbf{H}^{-1}\frac{dI}{d\boldsymbol{\theta}}$. The online (stochastic) version of the gradient descent algorithm may be obtained by removing the expected values in (3).

The score functions required in the equations must be estimated from the outputs y_1 and y_2 and be updated at each iteration of the optimization algorithm. They may be estimated for example using the approach proposed in [9] which consists in writing $\psi_{y_k}(y_k) = \sum_{m=1}^{M} c_{km}\phi_m(y_k)$ where $\phi_m(y_k)$ are some basis functions and in computing the coefficients c_{km} by solving the following equation

$$\mathbf{G}_k[c_{k1}, \cdots, c_{kM}]^T = \mathbf{g}_k \tag{5}$$

where $\mathbf{G}_k = E[\boldsymbol{\phi}(y_k)\boldsymbol{\phi}(y_k)^T]$, $\mathbf{g}_k = E[\boldsymbol{\phi}'(y_k)]$ with $\boldsymbol{\phi}(y_k) = [\phi_1(y_k), \cdots, \phi_M(y_k)]^T$ and $\boldsymbol{\phi}'(y_k)$ its derivative with respect to y_k. This derivative may also be used for estimating the score function derivatives required in (4).

An online estimate of the score functions may be obtained [10] at each time t by updating the matrices $\mathbf{G}_k$ and the vectors $\mathbf{g}_k$ using

$$\mathbf{G}_k(t) = \rho\mathbf{G}_k(t-1) + (1-\rho)\boldsymbol{\phi}(y_k)\boldsymbol{\phi}(y_k)^T\ ,\ \mathbf{g}_k(t) = \rho\mathbf{g}_k(t-1) + (1-\rho)\boldsymbol{\phi}'(y_k) \tag{6}$$

where ρ is a "forgetting factor" (for example equal to $(t-1)/t$), then solving $\mathbf{G}_k(t)[c_{k1}(t), \cdots, c_{kM}(t)]^T = \mathbf{g}_k(t)$ to find the coefficients $c_{km}(t)$.

In the following sections, we study a simple toy problem to show the different practical aspects of nonlinear BSS.

3 A Simple Bijective Model

We consider the following inverse structure

$$s_1 = a^* x_1^3 + b^* x_2 \qquad , \qquad s_2 = -b^* x_1 + a^* x_2. \tag{7}$$

This model, which is defined by the parameter vector $\boldsymbol{\theta}^* = [a^*, b^*]^T$ is bijective if $b^* \neq 0$ (and if $b^* = 0$ but $a^* x_1 \neq 0$) : in this case its Jacobian $3a^{*^2}x_1^2 + b^{*^2}$ is always positive. The above equations yield $a^* x_1^3 + (b^{*^2}/a^*)x_1 + (b^*/a^*)s_2 - s_1 = 0$

which can be solved using Cardan's formula with respect to x_1 to obtain one of the two mixing equations

$$x_1 = \left(\frac{-q}{2} + \sqrt{\Delta}\right)^{1/3} + \left(\frac{-q}{2} - \sqrt{\Delta}\right)^{1/3} \tag{8}$$

where $\Delta = \frac{q^2}{4} + \frac{p^3}{27}$, $q = \left((b^*/a^*)s_2 - s_1\right)/a^*$ and $p = (b^*/a^*)^2$. The other mixture may then be obtained using

$$x_2 = (s_2 + b^* x_1)/a^*. \tag{9}$$

BSS may be achieved by constructing the separating structure

$$y_1 = ax_1^3 + bx_2 \qquad , \qquad y_2 = -bx_1 + ax_2 \tag{10}$$

and minimizing the mutual information of y_1 and y_2 with respect to $\boldsymbol{\theta} = [a, b]^T$. This model yields $J = 3a^2x_1^2 + b^2$, $dJ/d\boldsymbol{\theta} = [6ax_1^2, 2b]^T$, $dy_1/d\boldsymbol{\theta} = [x_1^3, x_2]^T$, $dy_2/d\boldsymbol{\theta} = [x_2, -x_1]^T$. Using (3) and (4), we obtain the following expressions for the gradient and the Hessian

$$\begin{aligned}
\frac{dI}{d\boldsymbol{\theta}} &= E\left[\frac{-1}{3a^2x_1^2 + b^2}[6ax_1^2, 2b]^T + \psi_{y_1}(y_1)[x_1^3, x_2]^T + \psi_{y_2}(y_2)[x_2, -x_1]^T\right] \\
\mathbf{H} &= E\left[\frac{1}{J^2}\begin{pmatrix} -6x_1^2J + (6ax_1^2)^2 & 12abx_1^2 \\ 12abx_1^2 & -2J + (2b)^2 \end{pmatrix}\right] + E\left[\frac{d\psi_{y_1}(y_1)}{dy_1}\begin{pmatrix} x_1^6 & x_1^3x_2 \\ x_1^3x_2 & x_2^2 \end{pmatrix}\right] \\
&\quad + E\left[\frac{d\psi_{y_2}(y_2)}{dy_2}\begin{pmatrix} x_2^2 & -x_1x_2 \\ -x_1x_2 & x_1^2 \end{pmatrix}\right].
\end{aligned} \tag{11}$$

From (7) and (10), it is evident that if $\boldsymbol{\theta} = k[a^*, b^*]^T$, then $y_1 = ks_1$ and $y_2 = ks_2$ so that y_1 and y_2 are independent and minimize the criterion. One of the solutions to fix this indeterminacy consists in defining $s_i' = s_i/a^*$ and $c^* = b^*/a^*$ which yields the inverse model

$$s_1' = x_1^3 + c^* x_2 \qquad , \qquad s_2' = -c^* x_1 + x_2 \tag{12}$$

and the corresponding separating structure

$$y_1 = x_1^3 + cx_2 \qquad , \qquad y_2 = -cx_1 + x_2 \tag{13}$$

In this case, there is only one parameter to estimate so that the gradient and the Hessian are scalars: $\frac{dI}{dc} = E\left[\frac{-2c}{3x_1^2+c^2} + \psi_{y_1}(y_1)x_2 - \psi_{y_2}(y_2)x_1\right]$, $\mathbf{H} = \frac{dI^2}{d^2c} = E\left[\frac{-6x_1^2+2c^2}{(3x_1^2+c^2)^2} + \psi'_{y_1}(y_1)x_2^2 + \psi'_{y_2}(y_2)x_1^2\right]$.

4 Local Minima and Separability of the Model

In a first experiment, we mixed two 10000-sample independent, zero-mean and unit-variance, uniformly distributed sources s_1 and s_2 using the mixing model

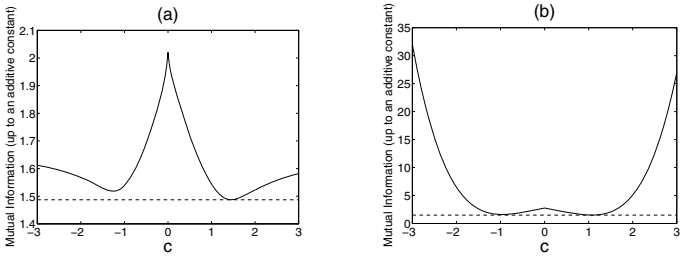

Fig. 1. Estimation of mutual information (up to an additive constant) (a) with pdf shape re-estimated for each value of c. (b) with pdf shape estimated for $c = 1$.

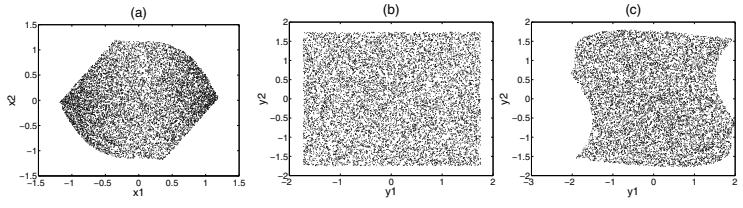

Fig. 2. scatter plots of (a) mixtures, (b) output components corresponding to the good local minimum, (c) output components corresponding to the bad local minimum

(8-9) and the parameters $a^* = 1$ and $b^* = 1.5$ which is equivalent to the inverse model (12) with $c^* = 1.5$ and $s'_i = s_i$. Then, we constructed the separating model (13), varied the parameter c between -3 and 3, and for each value of c estimated the mutual information of the outputs (up to the additive constant $E[\log f_{\mathbf{x}}(x)]$). The result is shown in Fig. 1.a. As can be seen, this function has two local minima but only one of them (which is also the global minimum) corresponds to the actual value of the parameter and provides the independent components corresponding to the actual sources[2]. When initialized with negative values of c, the optimization algorithms like gradient descent or Newton converge towards this "bad" local minimum. However, this value may be rejected a posteriori using an independence test. Fig. 2 shows the scatter plots of the mixtures and of the output components corresponding to these minima.

Note also that in practice, at each iteration of an optimization algorithm, one first estimates (using the current value of the parameter c) the coefficients c_{km} in (5) which determine the shape of score functions and related pdf, then freezes them and performs a minimization step for the mutual information related to these pdf with respect to c. Since the estimated pdf change during successive iterations, the shape of the function to be minimized changes too. For example, Fig. 1.b shows the mutual information as a function of c in the above example

[2] Note that replacing x_1^3 by x_1 in (12) and (13) yields a linear model for which the criterion has two good local minima at c^* and $-1/c^*$ leading respectively to $[y_1, y_2]^T = [s_1, s_2]^T$ and $[y_1, y_2]^T = [-s_2, s_1]^T/c^*$.

(with c^*= 1.5) corresponding to the coefficients c_{km} estimated using the value $c = 1$. As can be seen this function is not the same as in Fig. 1.a. The practical optimization is then more difficult than what is suggested by Fig. 1.a. This example also shows the sensitivity of the method to the estimation of score functions: if the estimated score functions are not updated in the following iterations, the optimization algorithm converges towards the minimum of Fig. 1.b, i.e. c= 1.07.

The separability of our one-parameter model may be formulated as follows: are there a family of source distributions and a value of the parameter c in the separating model (13) for which y_1 and y_2 are independent but contain mixtures of s_1 and s_2? To answer this question, one has to solve an independence conservation functional equation [8]. Here, we only try to respond partially to this question by the following experiment: we consider the generalized Gaussian distribution family defined by the parameter α. For the values of α between 0.5 and 20 we generated the mixtures x_1 and x_2 for $a^* = 1$ and $b^* = 1.5$ (so that $c^* = 1.5$), then the outputs y_1 and y_2 using (13) for the values of c between -20 and 20. For each value of α, we estimated the mutual information I as a function of c (like in Fig. 1.a) and verified if there was a value of c different from c^* for which $I \simeq 0$. Since we did not find such values, we can say "experimentally" that the model is separable for generalized Gaussian distributions.

5 Simulation Results

The first two lines of Table 1 compare the batch versions of the gradient descent (with a constant learning rate $\mu = 0.02$) and Newton methods applied to the mixtures generated as in Section 4. The algorithms were run 100 times corresponding to 100 different source signals and 100 different initial random values of the parameter c (uniformly distributed over [0.1, 2.1]). The score functions were estimated using the method described in Section 2 with $\phi_m(y_k) = y_k^{m-1}$ for $m = 1, \cdots, 5$. In each simulation, the algorithm was stopped if $|c_{new} - c_{old}| < 10^{-6}$ and the performance was measured using the Signal to Interference Ratio (SIR) criterion defined by $SIR = \frac{1}{2}\sum_{i=1}^{2} 10\log_{10}\frac{E[s_i^2]}{E[(s_i-\hat{s_i})^2]}$ where $\hat{s_i}$ is the estimate of s_i computed using the final estimate of the parameter c. We also tested the initial model with two parameters (Eq. 7-10) using $a^* = 4$, $b^* = 6$ (so that $b^*/a^* = 1.5$) and the parameters a and b initialized with positive random values. In this case, we estimate the two parameters simultaneously. The sources may be then estimated only up to a scaling factor. The last two lines of

Table 1. Comparing batch versions of gradient and Newton algorithms

	Mean(SIR)	Std(SIR)	Iterations per simulation	time per simulation
Gradient (1 param)	52.6 dB	9.6 dB	1208	17.87 sec
Newton (1 param)	52.6 dB	9.6 dB	111	1.7 sec
Gradient (2 param)	60.5 dB	7.8 dB	75	1.7 sec
Newton (2 param)	60.5 dB	7.8 dB	4	0.2 sec

Table 1 show the results. The SIR was computed after normalizing the estimated sources so that they had the same variances and signs as the actual sources. Note that in the Newton method, the Hessian $\mathbf{H}$ may be badly conditioned and even negative-definite. To avoid this problem, after the eigenvalue decomposition of $\mathbf{H} = \mathbf{Q}\boldsymbol{\Lambda}\mathbf{Q}^T$, the eigenvalues smaller than a positive value δ (chosen equal to 10^{-4} in our experiments) are replaced by δ [11].

As can be seen, this separating model leads to better results. This is probably because the parameters space defined by two parameters has a better shape so that the optimization algorithms converge better towards its minimum. In all the experiments, the Newton algorithm is much less time consuming than the gradient algorithm while providing the same performance.

Then, we tested the stochastic gradient algorithm. When using this algorithm, we have two principal problems: carefully choosing the learning rate (because the algorithm is extremely sensitive to this choice), and carefully estimating the score functions. The choice of learning rate μ is widely discussed in the neural networks literature [12]. We found that the following adaptation rule for updating the learning rate gives good results

$$\mu_i(t) = \mu_i(t-1).max\left(0.5,\ \ 1 + q\nabla_i(t)\frac{\overline{\nabla_i}(t-1)}{\overline{\nabla_i^2}(t)}\right) \tag{14}$$

where $\nabla_i = \frac{dI}{d\theta_i}$, $\overline{\nabla_i}(t-1) = \rho\overline{\nabla_i}(t-2)+(1-\rho)\nabla_i(t-1)$ and $\overline{\nabla_i^2}(t) = \rho\overline{\nabla_i^2}(t-1)+(1-\rho)\nabla_i^2(t)$ with ρ a forgetting factor. The main idea is to increase the learning rate when the new gradient points in the same direction as the average past gradient $\overline{\nabla_i}(t-1)$ (normalized by the average of the squared gradient to make it better conditioned), and to decrease it otherwise. The multiplier is limited below by 0.5 to guard against very small (or even negative) factors. In our experiments, we used $\rho = \frac{t-1}{t}$, $q = 1.5$ and $\mu(0) = 0.02$.

We also need a new estimation of the score functions at each time t. Our experiments show that the approach proposed at the end of Section 2 and based on Eq. (6) does not give good results because at the first stages of the algorithm the estimation of c and consequently the estimation of the score functions are bad. Using (6), this bad estimation of the score functions does not change significantly afterwards. Hence, we propose another approach which consists in updating the score functions at each time t from all the past data, using $\mathbf{G}_k(t) = \frac{1}{t}\sum_{n=1}^{t}\boldsymbol{\phi}(y_k(n))\boldsymbol{\phi}(y_k(n))^T$ and $\mathbf{g}_k(t) = \frac{1}{t}\sum_{n=1}^{t}\boldsymbol{\phi}'(y_k(n))$.

We repeated the experiment with the one-parameter model using the stochastic gradient algorithm and the signals containing 10000 samples. The mean and the standard deviation of the SIR using 10 simulations were 42.0 dB and 15.2 dB with a runtime of about 90 seconds for each simulation. Figure 3 shows the evolution of the parameter c and the learning rate μ in one of the simulations. The same experiment using the two-parameters model led to an average SIR of 46.9 dB with a standard deviation of 6.3 dB and about 140 sec per simulation.

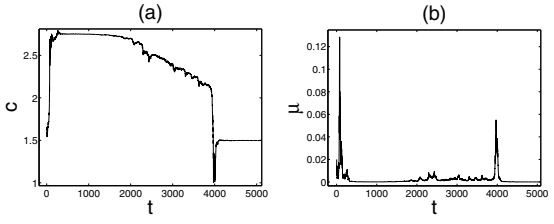

Fig. 3. Evolution of (a) the parameter c and (b) the learning rate μ in the stochastic algorithm versus sample index t

6 Conclusion

In this paper, we studied the BSS problem for one of the simplest bijective nonlinear models. Even for this simple model, the problem is much more difficult than linear BSS because of the existence of spurious local minima, the high sensitivity of the optimization algorithms to the estimation of score functions, the importance of parameter tuning in these algorithms, etc. We proposed solutions to cope with these problems which may be helpful for the future works using more realistic models. More experiments using constrained optimization algorithms are required for treating the case of non-bijective models.

References

1. Taleb, A., Jutten, C.: Source separation in post-nonlinear mixtures. IEEE Trans. on Signal Processing 47(10), 2807–2820 (1999)
2. Jutten, C., Babaie-Zadeh, M., Hosseini, S.: Three easy ways for separating nonlinear mixtures? Signal Processing 84(2), 217–229 (2004)
3. Almeida, L.B.: MISEP - linear and nonlinear ICA based on mutual information. Journal of Machine Learning Research 4, 1297–1318 (2003)
4. Zhang, K., Chan, L.: Kernel-Based Nonlinear Independent Component Analysis. In: Davies, M.E., James, C.J., Abdallah, S.A., Plumbley, M.D. (eds.) ICA 2007. LNCS, vol. 4666, pp. 301–308. Springer, Heidelberg (2007)
5. Eriksson, J., Koivunen, V.: Blind identifiability of class of nonlinear instantaneous ICA models. In: Proc. of EUSIPCO 2002, Toulouse, vol. 2, pp. 7–10 (September 2002)
6. Hosseini, S., Deville, Y.: Blind Maximum Likelihood Separation of a Linear-Quadratic Mixture. In: Puntonet, C.G., Prieto, A.G. (eds.) ICA 2004. LNCS, vol. 3195, pp. 694–701. Springer, Heidelberg (2004), ERRATUM: `http://arxiv.org/abs/1001.0863`
7. Duarte, L.T., Jutten, C.: Blind Source Separation of a Class of Nonlinear Mixtures. In: Davies, M.E., James, C.J., Abdallah, S.A., Plumbley, M.D. (eds.) ICA 2007. LNCS, vol. 4666, pp. 41–48. Springer, Heidelberg (2007)
8. Taleb, A.: A generic framework for blind source separation in structured nonlinear models. IEEE Trans. Sig. Proc. 50(8), 1819–1830 (2002)

9. Pham, D.T., Garat, P.: Blind separation of mixtures of independent sources through a quasi maximum likelihood approach. IEEE Trans. Sig. Proc. 45(7), 1712–1725 (1997)
10. Pham, D.T.: Séparation aveugle de mélange instantanée de sources à l'aide de fonctions séparatrices ajustées. In: Proc. GRETSI 1997, pp. 969–972 (September 1997)
11. Nocedal, J., Wright, S.: Numerical optimization. Springer (2006)
12. Bishop, C.M.: Neural networks for pattern recognition, Oxford (1995)