Paper reference:
Y. Deville, S. Hosseini, A. Deville, ”Effect of indirect dependencies on maximum likelihood and information theoretic blind source separation for nonlinear mixtures”, Signal Processing, vol. 91, issue 4, pp. 793-800, April 2011.

Available online at:
http://dx.doi.org/10.1016/j.sigpro.2010.08.012 .

Effect of indirect dependencies on maximum likelihood and information theoretic blind source separation for nonlinear mixtures

Yannick Deville[a,*], Shahram Hosseini[a], Alain Deville[b]

[a]*Laboratoire d'Astrophysique de Toulouse-Tarbes, Université de Toulouse, CNRS, 14 Av. Edouard Belin, 31400 Toulouse, France.*
[b]*IM2NP, Université de Provence, Centre de Saint-Jérôme, 13397 Marseille Cedex 20, France.*

Abstract

Two major approaches for Blind Source Separation (BSS) are respectively based on the maximum likelihood (ML) principle and mutual information (MI) minimization. They have been mainly studied for simple linear mixtures. We here show that they additionally involve indirect functional dependencies for general nonlinear mixtures. Moreover, the notations commonly employed by the BSS community in calculations performed for these methods may become misleading when using them for nonlinear mixtures, due to the above-mentioned dependencies. In this paper, we first explain this phenomenon for arbitrary nonlinear mixing models. We then accordingly correct two previously published methods for specific nonlinear mixtures, where indirect dependencies were mistakenly ignored. This paper therefore opens the way to the application of the ML and MI BSS methods to many specific mixing models, by providing general tools to address such mixtures and explicitly showing how to apply these tools to practical cases.

Keywords: independent component analysis, maximum likelihood, information theory, mutual information, indirect functional dependencies, nonlinear mixture

1. Introduction

Blind source separation (BSS) consists in restoring a vector $s(t)$ of N unknown source signals from a vector $x(t)$ of P observed signals (here with $P = N$), which are derived from $s(t)$ through an unknown mixing function g, i.e. [5], [12]

$$x(t) = g(s(t)). \tag{1}$$

The main class of methods which has been proposed to this end is based on the assumed statistical independence of the source signals and is called Independent Component Analysis (ICA). Various principles have been reported for performing ICA[1]. This especially includes the maximum likelihood (ML) approach. ML-based ICA is attractive, because it takes advantage of the general good statistical properties of ML estimation, beyond the specific scope of ICA [5], [12]. The ML approach for ICA was mainly developed for linear instantaneous mixtures. It was especially introduced by Gaeta and Lacoume [9], and Pham and Garat [14]. Other linear instantaneous ICA methods were developed by using information theoretic criteria, mainly by minimizing the mutual information (MI) of the estimated sources [4]. Although they were initially based on different signal processing tools, these ML and MI approaches turned out to yield similar methods.

More recently, several authors extended the ML and MI approaches[2] to some specific classes of nonlinear mixtures (see especially [6],[8],[10],[16] ; see also [17] for a more general framework, or [1] for a method which is not dedicated

*Corresponding author. Phone: +33 5 61 33 28 24, fax: +33 5 61 33 28 40.
Email addresses: `ydeville@ast.obs-mip.fr` (Yannick Deville), `shosseini@ast.obs-mip.fr` (Shahram Hosseini), `alain.deville@univ-provence.fr` (Alain Deville)

[1] For linear instantanous mixtures, various ICA-related methods were proposed, in addition to the ML and MI approaches which are considered hereafter. Several of these other ICA-related approaches are based on second-order or higher-order moments or cumulants: see e.g. COM2/ICA [4], JADE [3], SOBI [2] and the kurtosis-based version of FastICA [11].

[2] Apart from ML and MI, various other BSS approaches were also proposed for nonlinear mixtures: see e.g. the overviews in [5], [13].

to a specific class of mixtures). This extension requires some care, because the considered cost functions contain indirect dependencies, and the notations used for ML and MI by the BSS community may become ambiguous, or even lead to false interpretation, for *nonlinear* mixtures. This paper aims at clarifying this point for general nonlinear mixing models, and at correcting associated errors which appeared in part of the studies that were previously reported for some specific nonlinear mixing models.

The remainder of this paper is therefore organized as follows. In Section 2, we detail the effect of indirect dependencies when using the ML approach for general (possibly nonlinear) mixing models. Then, in Section 3, we show that this phenomenon disappears for linear mixtures, and we explain how it must be taken into account to correct a previously reported method for linear-quadratic mixtures. The MI approach yields similar phenomena and is therefore then addressed somewhat more briefly: first, in Section 4, we show how to handle indirect dependencies in the MI approach for general mixtures ; then, in Section 5, we correct a previously reported method for specific mixtures, and we comment about linear mixtures as a spin-off. Conclusions are drawn from this investigation in Section 6.

2. Maximum likelihood approach for general mixtures

2.1. Separation criterion

We here consider a general mixing function g, i.e. we only assume that it is bijective and memoryless: the observed vector at time t, i.e. $x(t)$, only depends on the source vector at the same time, i.e. $s(t)$. Moreover, each source signal is assumed to be independent and identically distributed (i.i.d.), as usual in the ML approach. In this framework, the original mixing model (1) may be reformulated by only considering a single time t and by using the corresponding random source vector S and random observed vector X, which reads

$$X = g(S). \tag{2}$$

The joint probability density functions (pdf) of these vectors are respectively denoted as f_S and f_X. The pdf f_S is fixed (and possibly unknown). Since we assume that the mixing function g is bijective, we have

$$f_X(x) = \frac{f_S(s)}{|J_g(s)|} \tag{3}$$

where $s = g^{-1}(x)$ and $J_g(s)$ is the Jacobian of g, i.e. the determinant of the Jacobian matrix $\frac{\partial g}{\partial s}$, whose element (i, j) is equal to $\frac{\partial g_j}{\partial s_i}$ [12]. Taking the logarithm of (3), and assuming the sources to be mutually statistically independent, we obtain

$$\ln f_X(x) = \sum_{i=1}^{N} \ln f_{S_i}(s_i) - \ln |J_g(s)| \tag{4}$$

where s_i are the components of s and $f_{S_i}(s_i)$ are the marginal source pdf.

The mixing function g is assumed to belong to a given class of functions and to have a set of parameters, whose values are unknown. Similarly, the separating system used to restore the sources from the observations corresponds to a function h belonging to a given class, and its parameter values must be selected so as to achieve $h = g^{-1}$ (examples are provided below in Section 3). Eq. (1) then yields

$$s(t) = h(x(t)). \tag{5}$$

The ML approach may be used to estimate either the parameters of g or those of h. The set of parameters of the considered function, i.e. of g or h, is denoted as $\theta = [\theta_1, \ldots, \theta_K]^T$ hereafter, where ^{T} stands for transpose. When θ consists of the parameters of g, Eq. (1) focuses on the signals (i.e. sources and observations), and it hides the fact that the observations also depend on θ. This additional dependency can be made explicit, by rewriting (1) as

$$x(t) = g(s(t), \theta). \tag{6}$$

Similarly, when θ consists of the parameters of h, Eq. (5) may be rewritten as

$$s(t) = h(x(t), \theta). \tag{7}$$

Eq. (7) thus explicitly shows that the separating system outputs (which are equal to the source values in the considered configuration, where $h = g^{-1}$) are functions of the observations and of the vector θ of parameters of the separating system. Therefore, for a given observed vector $x(t)$, Eq. (7) shows that the source vector $s(t)$ may be considered as a function of θ. This topic is addressed in more detail further in this paper (see especially paragraph after Eq. (19)), because we show below that expressions (6) and (7) of the mixing and separating models are better suited than (1) and (5) to the ML approach considered in this paper.

Given M samples of the observed vector X, the ML estimator of θ is obtained as the value $\hat{\theta}_{ML}$ of θ which maximizes the joint pdf of all these observations, called the likelihood, which is equal to

$$L = f_X(x_1(1), \ldots, x_N(1), \cdots, x_1(M), \ldots, x_N(M)) \tag{8}$$

where $x_j(m)$ is the value of the m-th sample of the j-th observation. Since we assumed that the source signals are i.i.d. and the mixing model is memoryless, each observed signal is also i.i.d, so that

$$\begin{aligned} L &= \prod_{m=1}^{M} f_X(x_1(m), \ldots, x_N(m)) && (9) \\ \text{and} \qquad \ln L &= \sum_{m=1}^{M} \ln f_X(x_1(m), \ldots, x_N(m)). && (10) \end{aligned}$$

Maximizing L is equivalent to maximizing the (normalized) log-likelihood $\mathcal{L} = \frac{1}{M} \ln L$. Thanks to (10), $\mathcal{L}$ may be denoted as

$$\mathcal{L} = E_t[\ln f_X(x_1(t), \ldots, x_N(t))], \tag{11}$$

using the temporal averaging operator over the set of available data, which is denoted as $E_t[.]$. Eq. (4) then yields

$$\mathcal{L} = \sum_{i=1}^{N} E_t[\ln f_{S_i}(s_i(t))] - E_t[\ln |J_g(s(t))|]. \tag{12}$$

2.2. Gradient of log-likelihood

Determining the value of θ which maximizes $\mathcal{L}$ involves each of the derivatives of $\mathcal{L}$ with respect to one component θ_k of θ, while all its other components are constant. Each such derivative is denoted $\frac{d\mathcal{L}}{d\theta_k}$ hereafter. The column vector composed of these derivatives is called the gradient of $\mathcal{L}$ with respect to θ and is denoted $\frac{d\mathcal{L}}{d\theta}$ hereafter. These notations are used for the sake of clarity: although the BSS community most often denotes this gradient as $\frac{\partial\mathcal{L}}{\partial\theta}$ and each of its components as $\frac{\partial\mathcal{L}}{\partial\theta_k}$, we will not use the latter notations, because we will show that they may be misleading for *nonlinear* mixtures.

The above gradient is first used to express a necessary condition for a value $\hat{\theta}_{ML}$ of θ to maximize $\mathcal{L}$. This condition reads

$$\left.\frac{d\mathcal{L}}{d\theta}\right|_{\theta=\hat{\theta}_{ML}} = 0 \tag{13}$$

or, in scalar form

$$\left.\frac{d\mathcal{L}}{d\theta_k}\right|_{\theta=\hat{\theta}_{ML}} = 0 \qquad \forall\, k = 1, \ldots, K. \tag{14}$$

This gradient is also used in the so-called gradient ascent algorithm, which is a simple procedure for numerically optimizing θ so as to (locally) maximize $\mathcal{L}$, by means of the iterative adaptation rule

$$\theta(n+1) = \theta(n) + \mu \left.\frac{d\mathcal{L}}{d\theta}\right|_{\theta=\theta(n)} \tag{15}$$

where μ is a positive adaptation gain.

Whereas the above description of the first steps of the ML approach is a rather conventional prerequisite of our analysis, the next step consists in deriving the explicit expression of the above-defined gradient and deserves some care for nonlinear mixtures, as will now be shown. Eq. (12) is used to determine each derivative $\frac{d\mathcal{L}}{d\theta_k}$, by taking into account that $\frac{d \ln |J_g|}{d\theta_k} = \frac{1}{J_g}\frac{dJ_g}{d\theta_k}$, and by introducing the score functions of the sources[3], defined as

$$\psi_{s_i}(u) = -\frac{d \ln f_{S_i}(u)}{du} \qquad \forall i = 1, \ldots, N. \tag{16}$$

Eq. (12) thus yields

$$\frac{d\mathcal{L}}{d\theta_k} = -\sum_{i=1}^{N} E_t[\psi_{s_i}(s_i)\frac{ds_i}{d\theta_k}] - E_t[\frac{1}{J_g}\frac{dJ_g}{d\theta_k}] \tag{17}$$

or, in column vector form

$$\frac{d\mathcal{L}}{d\theta} \quad = \quad -E_t[\frac{ds}{d\theta}\psi_S(s)] - E_t[\frac{1}{J_g}\frac{dJ_g}{d\theta}] \tag{18}$$

$$\text{where} \quad \psi_S(s) \quad = \quad \left[\psi_{s_1}(s_1), \ldots, \psi_{s_N}(s_N)\right]^T \tag{19}$$

and $\frac{ds}{d\theta}$ is the matrix whose element (i, j) is equal to $\frac{ds_j}{d\theta_i}$.

The calculation of the term $\frac{dJ_g}{d\theta_k}$ of (17) then deserves care and led to an error in a previously published paper, as detailed in Section 3.2. One should realize that, when applying the ML approach to any BSS configuration, the log-likelihood $\mathcal{L}$ is considered for the fixed set of observed vectors. The only independent variable in this approach is the column vector θ of mixing or separating parameters to be estimated. The source vectors are dependent variables, here linked to the observations and to θ by (6) or (7). The overall variations of the log-likelihood $\mathcal{L}$ with respect to θ result from two types of terms contained in the expression of $\mathcal{L}$, i.e. (i) the terms involving θ itself and (ii) the terms involving the source signals $s_1, \ldots, s_N$, which are here considered as functions of θ and may therefore be denoted as $s_1(\theta), \ldots, s_N(\theta)$ for the sake of clarity. Similarly, the log-likelihood, which appears in the left-hand side of (12), may be denoted as $\mathcal{L}(\theta, s_1(\theta), \ldots, s_N(\theta))$ for the sake of clarity. In order to determine the location of the maximum of this log-likelihood, one should then consider the *total* derivatives of $\mathcal{L}(\theta, s_1(\theta), \ldots, s_N(\theta))$ with respect to each parameter θ_k. We therefore denoted these derivatives as $\frac{d\mathcal{L}}{d\theta_k}$ in (17). On the contrary, the notation with *partial* derivatives, i.e. $\frac{\partial\mathcal{L}}{\partial\theta_k}$, often used for these quantities in the BSS community may be misleading, as confirmed below.

The above comment is of importance for the term $\dfrac{dJ_g}{d\theta_k}$ in (17) because, for general nonlinear mixing models, the Jacobian J_g contains the above-defined two types of dependencies with respect to θ_k, i.e. (i) *direct dependencies* due to the terms of J_g which explicitly contain θ_k and (ii) *indirect dependencies* due to the terms of J_g which depend on the source signals, which themselves depend on θ_k in the ML approach. We here have to consider the *total* derivative $\dfrac{dJ_g}{d\theta_k}$, which takes into account both types of dependencies, and which therefore reads[4]

$$\frac{dJ_g}{d\theta_k} = \frac{\partial J_g}{\partial\theta_k} + \sum_{i=1}^{N}\frac{\partial J_g}{\partial s_i}\frac{ds_i}{d\theta_k}, \tag{20}$$

[3] In practice, the pdf of the sources are most often unknown and their score functions are estimated, as explained e.g. in [12].

[4] Each derivative $\frac{dJ_g}{d\theta_k}$ is ”total” only with respect to the considered component θ_k of θ, i.e. it takes into account all variations of J_g with respect to that component θ_k, while all other components of θ are kept constant. For the sake of clarity, we could therefore denote that derivative $\left(\frac{dJ_g}{d\theta_k}\right)_{\theta\backslash\{\theta_k\}}$, to show that all components of θ except θ_k are constant. However, this would decrease readability. Therefore, in all this paper we omit the notation $(.)_{\theta\backslash\{\theta_k\}}$, but it should be kept in mind that each considered derivative with respect to θ_k is calculated with all other components of θ constant. Then, in this framework, what we have to distinguish are: (i) the total derivative due to the variations of θ_k and of all s_i, and (ii) the partial derivative (i.e. with all components of θ except θ_k fixed, and all s_i fixed). We then have to use two different notations for these two types of derivatives, such as $\frac{dJ_g}{d\theta_k}$ and $\frac{\partial J_g}{\partial\theta_k}$ in (20). These two types of notations are commonly used in the literature for functions which depend (i) on a single independent variable, i.e. time, and (ii) on other variables which themselves depend on time, such as coordinate variables: see e.g. http://en.wikipedia.org/wiki/Total_derivative or [15]. We here extend this concept to a configuration which involves several independent variables, i.e. all components θ_k (and, again, other variables which themselves depend on the independent variables, i.e. all s_i). We here keep the same types of notations as in the standard case involving a single independent variable. The MI approach described in Section 4 yields the same type of comment.

or, in column vector form

$$\frac{dJ_g}{d\theta} = \frac{\partial J_g}{\partial \theta} + \frac{ds}{d\theta}\frac{\partial J_g}{\partial s}. \tag{21}$$

The term $\dfrac{\partial J_g}{\partial \theta_k}$ of (20) is the *partial* derivative of J_g with respect to θ_k, calculated by considering that the source signals s_i are constant. This clearly shows that the total derivative, which appears in the left-hand side of (20) and which is the quantity that we aim at determining here, should be denoted $\frac{dJ_g}{d\theta_k}$, as in the current paper: instead, if it was denoted $\dfrac{\partial J_g}{\partial \theta_k}$ as is usually done in the BSS community, this would yield two problems. First, it would not be possible to write (20) as above, because the same notation would be used for its left-hand side and for only the first term of its overall right-hand side. Then, and more dangerously, starting from the inadequate notation $\dfrac{\partial J_g}{\partial \theta_k}$ for the overall quantity to be determined (i.e. left-hand side of (20)), one would be led to mistakenly interpret it as the *partial* derivative of J_g with respect to θ_k, and to calculate it by considering that the source signals are constant. One would thus forget all other terms, i.e. all $\frac{\partial J_g}{\partial s_i}\frac{ds_i}{d\theta_k}$ in the right-hand side of (20). This error was made in a previously published paper. In the next section, we therefore show how this error should be corrected and, more generally speaking, how the ML method may be applied to various mixing models[5], including linear ones.

3. Applications of maximum likelihood approach

To apply the ML approach defined in Section 2 to a given mixing model, we now just have to determine the expressions of the Jacobian J_g of this mixture and of its total derivatives (20) with respect to the mixing or separating parameters θ_k. This then makes it possible to derive the expressions of the gradient components (17) associated with this mixing model, as all factors $\frac{ds_i}{d\theta_k}$ required in (17) and (20) may be obtained as explained in the Appendix. This gradient may then e.g. be used in the gradient ascent algorithm (15). For the sake of clarity, we first briefly show how this approach is related to already known results for linear mixtures.

3.1. Linear mixtures

The simplest BSS configuration corresponds to linear instantaneous mixtures. The mixing model (2) then reads

$$X = AS, \tag{22}$$

where A is a square, supposedly invertible, unknown, mixing matrix. The ML method may be used to estimate the inverse of this mixing matrix , i.e. $B = A^{-1}$. The variables called θ_k in the above discussion then consist of the elements of B. This matrix B is used as the separating system, in order to restore the sources by computing them according to

$$y(t) = Bx(t). \tag{23}$$

The Jacobian of any mixing model was defined above (after (3)). For the mixing model (22), it reads

$$J_g(s) = \det A^T = \frac{1}{\det B} \qquad \forall\ s. \tag{24}$$

Our main comment is that, for this specific case of *linear* mixtures, $J_g(s)$ *does not depend* on the source signals s_i. Therefore, the total derivative $\frac{dJ_g}{d\theta_k}$ in (20) is here only composed of the partial derivative $\frac{\partial J_g}{\partial \theta_k}$. This may be expressed in a compact way, by gathering all these scalar derivatives in matrices [12], which yields

$$\frac{dJ_g}{dB} = \frac{\partial J_g}{\partial B}. \tag{25}$$

[5] The application of the ML approach to a specific nonlinear mixing model reported in [6] does not contain any explicitly false expression. However, it is ambiguous because is does not detail all notations and expressions for derivatives. This ambiguity can be easily solved by using the approach described in the current paper.

Therefore, for the specific case of linear mixtures, the distinction between total and partial derivatives is not an issue. This is probably the reason why this distinction has not been considered in detail up to now, since most BSS investigations were restricted to linear mixtures. We will now show that things become different for nonlinear mixtures, by considering a practical example.

3.2. Linear-quadratic mixtures

The configuration considered in [10] involves two observations which are linear-quadratic mixtures of two sources, i.e.

$$x_1 = s_1 - l_1 s_2 - q_1 s_1 s_2 \tag{26}$$

$$x_2 = s_2 - l_2 s_1 - q_2 s_1 s_2 \tag{27}$$

where l_1, l_2, q_1, q_2 are unknown mixing parameters. This model is a specific version of the additive-target mixtures (ATM) which were defined in [7], together with associated separating structures. Here, we need not describe the separating structure used in [10], because the ML approach of [10] that we want to address is used to estimate the parameters of the *mixing* model. Therefore, we here have $\theta = [l_1, l_2, q_1, q_2]^T$. The Jacobian J_g of the mixing model is derived from (26)-(27), which yield

$$J_g = 1 - l_1 l_2 - (q_2 + l_2 q_1) s_1 - (q_1 + l_1 q_2) s_2. \tag{28}$$

We then aim at computing the gradient of J_g defined in (21). Its first term is derived from (28), which yields

$$\frac{\partial J_g}{\partial \theta} = -\Big[l_2 + q_2 s_2, l_1 + q_1 s_1, l_2 s_1 + s_2, s_1 + l_1 s_2 \Big]^T. \tag{29}$$

Similarly, (28) results in

$$\frac{\partial J_g}{\partial s} = -[q_2 + l_2 q_1, q_1 + l_1 q_2]^T. \tag{30}$$

The factor $\frac{ds}{d\theta}$ to be used in (21) is then obtained from Eq. (A.5) derived in the appendix, where (26)-(27) yield

$$\frac{\partial g}{\partial \theta} = \begin{bmatrix} -s_2 & 0 & -s_1 s_2 & 0 \\ 0 & -s_1 & 0 & -s_1 s_2 \end{bmatrix}^T \tag{31}$$

$$\text{and} \quad \frac{\partial g}{\partial s} = \begin{bmatrix} 1 - q_1 s_2 & -l_2 - q_2 s_2 \\ -l_1 - q_1 s_1 & 1 - q_2 s_1 \end{bmatrix}. \tag{32}$$

Eq. (A.5) thus results in

$$\frac{ds}{d\theta} = \frac{1}{J_g} \begin{bmatrix} (1 - q_2 s_1) s_2 & (l_1 + q_1 s_1) s_1 & (1 - q_2 s_1) s_1 s_2 & (l_1 + q_1 s_1) s_1 s_2 \\ (l_2 + q_2 s_2) s_2 & (1 - q_1 s_2) s_1 & (l_2 + q_2 s_2) s_1 s_2 & (1 - q_1 s_2) s_1 s_2 \end{bmatrix}^T. \tag{33}$$

Using (29), (33) and (30), Eq. (21) eventually becomes

$$\frac{dJ_g}{d\theta} = \begin{bmatrix} -(l_2 + q_2 s_2) - (q_2 + l_2 q_1)(1 - q_2 s_1) s_2 / J_g - (q_1 + l_1 q_2)(l_2 + q_2 s_2) s_2 / J_g \\ -(l_1 + q_1 s_1) - (q_2 + l_2 q_1)(l_1 + q_1 s_1) s_1 / J_g - (q_1 + l_1 q_2)(1 - q_1 s_2) s_1 / J_g \\ -(l_2 s_1 + s_2) - (q_2 + l_2 q_1)(1 - q_2 s_1) s_1 s_2 / J_g - (q_1 + l_1 q_2)(l_2 + q_2 s_2) s_1 s_2 / J_g \\ -(l_1 s_2 + s_1) - (q_2 + l_2 q_1)(l_1 + q_1 s_1) s_1 s_2 / J_g - (q_1 + l_1 q_2)(1 - q_1 s_2) s_1 s_2 / J_g \end{bmatrix}. \tag{34}$$

This is the correct expression of $\frac{dJ_g}{d\theta}$, then used to derive the expression of $\frac{d\mathcal{L}}{d\theta}$ (see (18)). On the contrary, the set (29) of partial derivatives was mistakenly used in [10], as if it were the factor $\frac{dJ_g}{d\theta}$ used in the expression of $\frac{d\mathcal{L}}{d\theta}$ (see (17) in [10]).

For the sake of completeness, we eventually provide the explicit expression of $\frac{d\mathcal{L}}{d\theta}$, defined in (18), which results from (34) and (33) (and which therefore replaces (17) of [10]), i.e.

$$\begin{aligned}
\frac{d\mathcal{L}}{d\theta} \;=\; & -E_t\Big[\Big(\psi_{s_1}(s_1)(1-q_2s_1)s_2+\psi_{s_2}(s_2)(l_2+q_2s_2)s_2 \\
& -(l_2+q_2s_2)-(q_2+l_2q_1)(1-q_2s_1)s_2/J_g-(q_1+l_1q_2)(l_2+q_2s_2)s_2/J_g\Big)/J_g, \\
& \Big(\psi_{s_1}(s_1)(l_1+q_1s_1)s_1+\psi_{s_2}(s_2)(1-q_1s_2)s_1 \\
& -(l_1+q_1s_1)-(q_2+l_2q_1)(l_1+q_1s_1)s_1/J_g-(q_1+l_1q_2)(1-q_1s_2)s_1/J_g\Big)/J_g, \\
& \Big(\psi_{s_1}(s_1)(1-q_2s_1)s_1s_2+\psi_{s_2}(s_2)(l_2+q_2s_2)s_1s_2 \\
& -(l_2s_1+s_2)-(q_2+l_2q_1)(1-q_2s_1)s_1s_2/J_g-(q_1+l_1q_2)(l_2+q_2s_2)s_1s_2/J_g\Big)/J_g, \\
& \Big(\psi_{s_1}(s_1)(l_1+q_1s_1)s_1s_2+\psi_{s_2}(s_2)(1-q_1s_2)s_1s_2 \\
& -(l_1s_2+s_1)-(q_2+l_2q_1)(l_1+q_1s_1)s_1s_2/J_g-(q_1+l_1q_2)(1-q_1s_2)s_1s_2/J_g)\Big)/J_g\Big]^T.
\end{aligned} \tag{35}$$

4. Mutual information minimization for general mixtures

The MI-based BSS approach leads to the same type of phenomenon as above for nonlinear mixtures. We therefore describe it more briefly hereafter, again for an arbitrary bijective memoryless mixing function g. The sources are assumed to be mutually statistically independent and stationary, so that we omit the considered time index t in all signals hereafter. The separating system corresponds to a function h, i.e. its output vector y reads

$$y = h(x). \tag{36}$$

h is assumed to belong to a given class of functions and to have a vector $\theta = [\theta_1, \ldots, \theta_K]^T$ of parameters, that we aim at estimating so as to achieve $h = g^{-1}$. The criterion used to this end here consists in minimizing the mutual information, denoted $I(Y)$, of the vector Y of random variables Y_i associated with the output signal samples y_i of the separating system at time t. Denoting $H(.)$ marginal and joint differential entropies, we have

$$I(Y) \;=\; \left(\sum_{i=1}^{N} H(Y_i)\right) - H(Y). \tag{37}$$

Moreover, $H(Y) = -E\{\ln f_Y(Y)\}$, where $E\{.\}$ stands for expectation. Eq. (36) and (37) therefore yield

$$I(Y) \;=\; \left(\sum_{i=1}^{N} H(Y_i)\right) - H(X) - E\{\ln |J_h|\} \tag{38}$$

where J_h is the Jacobian[6] of the separating function h, defined in the same way as J_g above.

In order to determine the value of θ which minimizes $I(Y)$, we consider the gradient of $I(Y)$ with respect to θ. Its components read as follows, using [16] for differential entropy derivatives, and taking into account that, in this investigation, the observations are fixed and $H(X)$ is therefore a constant[7]

[6] For the sake of readability, we use the same notation, i.e. J_h, for (i) the sample value of this Jacobian associated with signal sample values y_i (see e.g. (49)) and (ii) the random variable defined by this quantity when considered as a function of the random variables Y_i (see e.g. (38)). To know whether we are considering the sample value of J_h or the associated random variable in an equation, one just has to check whether that equation involves the sample values y_i or the associated random variables Y_i.

[7] One may therefore equivalently minimize $C(Y) = I(Y) + H(X)$ instead of $I(Y)$, e.g. as in [8]. $\frac{dC(Y)}{d\theta_k}$ is then also expressed as in (39).

$$\frac{dI(Y)}{d\theta_k} = \left(\sum_{i=1}^{N} E\{\psi_{Y_i}(Y_i)\frac{dY_i}{d\theta_k}\}\right) - E\{\frac{1}{J_h}\frac{dJ_h}{d\theta_k}\} \qquad (39)$$

$$\text{where} \quad \psi_{Y_i}(u) = -\frac{d\ln f_{Y_i}(u)}{du} \qquad \forall\ i = 1, \ldots, N \qquad (40)$$

are the score functions of the outputs of the separating system, denoting $f_{Y_i}(.)$ the pdf of these signals. The term $\frac{dJ_h}{d\theta_k}$ in (39) again deserves some care, because J_h in general contains (i) direct dependencies with respect to θ and (ii) dependencies with respect to the separating system outputs y_i, which yield indirect dependencies with respect to θ. We here have to consider the *total* derivative $\frac{dJ_h}{d\theta_k}$, which takes into account both types of dependencies, and which therefore reads

$$\frac{dJ_h}{d\theta_k} = \frac{\partial J_h}{\partial \theta_k} + \sum_{i=1}^{N} \frac{\partial J_h}{\partial y_i}\frac{dy_i}{d\theta_k}. \qquad (41)$$

In this expression, $\dfrac{\partial J_h}{\partial \theta_k}$ is the *partial* derivative of J_h with respect to θ_k, calculated by considering that the signals y_i are constant (in addition to the fact that all components of θ except θ_k are also constant).

In [8], the variations of J_h with respect to all y_i were forgotten, i.e. $\dfrac{\partial J_h}{\partial \theta_k}$ was used instead of $\frac{dJ_h}{d\theta_k}$ in (39). We show how to correct that error in the next section. That section also illustrates a general phenomenon: for many nonlinear mixing models, the analytical expressions of the inverse (i.e. separating) model h and therefore of its Jacobian J_h cannot be derived. However, those for h^{-1} can (they are nothing but those for g, but expressed vs. the signals and separating coefficients involved in h, as illustrated below). The Jacobian J_h is then calculated as $J_h = (J_{h^{-1}})^{-1}$. That expression could also be used to directly simplify (38) and (39).

The above presentation also shows that the ML and MI approaches are closely related for the considered general mixing model: replacing sample temporal averaging by expectation in (12) (based on ergodicity and considering $M \rightarrow +\infty$), and the unknown source signals by their estimates available as outputs of the separating system, the log-likelihood $\mathcal{L}$ is replaced by

$$\mathcal{L}_2 = \sum_{i=1}^{N} E\{\ln f_{Y_i}(Y_i)\} - E\{\ln|J_{h^{-1}}|\} = -\sum_{i=1}^{N} H(Y_i) + E\{\ln|J_h|\} = -I(Y) - H(X). \qquad (42)$$

Therefore, maximizing $\mathcal{L}_2$ is equivalent to minimizing $I(Y)$, since $H(X)$ is a constant.

5. Applications of mutual information minimization approach

5.1. Mixtures with power terms

The investigation in [8] concerns a specific nonlinear BSS problem which involves two observed signals, derived from two source signals through the nonlinear function defined as

$$\begin{aligned} x_1 &= s_1 + a_{12}(s_2)^k & (43) \\ x_2 &= s_2 + a_{21}(s_1)^{\frac{1}{k}} & (44) \end{aligned}$$

where a_{12} and a_{21} are two unknown mixing coefficients and k is a known integer.

The separating structure used in [8] to process such mixtures was derived from the structure for linear-quadratic mixtures proposed e.g. in [7],[10]. It belongs to the general class of structures proposed in [7] for the ATM class of mixing models, which includes the specific model (43)-(44). The separating structure of [8] has internal adaptive coefficients w_{12} and w_{21}, which here compose θ. For each time t, this structure determines the output vector $y = [y_1, y_2]^T$ from its current internal coefficients and from the current observation vector x. To this end, it iteratively updates its output according to

$$\begin{aligned} y_1(n+1) &= x_1 - w_{12}(y_2(n))^k & (45) \\ y_2(n+1) &= x_2 - w_{21}(y_1(n))^{\frac{1}{k}}. & (46) \end{aligned}$$

The convergence of this recurrence therefore corresponds to a state such that $y_i(n+1) = y_i(n) = y_i$ for $i \in \{1, 2\}$. Eq. (45)-(46) then yield

$$x_1 = y_1 + w_{12} y_2^k \tag{47}$$

$$x_2 = y_2 + w_{21} y_1^{\frac{1}{k}} \tag{48}$$

which is the expression of the above-mentioned function h^{-1}. This yields

$$J_h = \frac{1}{J_{h^{-1}}} = \frac{1}{1 - w_{12} w_{21} y_1^{\frac{1}{k}-1} y_2^{k-1}}. \tag{49}$$

All partial derivatives of J_h involved in (41) are then easily derived from (49) and read

$$\frac{\partial J_h}{\partial w_{12}} = \frac{w_{21} y_1^{\frac{1}{k}-1} y_2^{k-1}}{[1 - w_{12} w_{21} y_1^{\frac{1}{k}-1} y_2^{k-1}]^2} \tag{50}$$

$$\frac{\partial J_h}{\partial w_{21}} = \frac{w_{12} y_1^{\frac{1}{k}-1} y_2^{k-1}}{[1 - w_{12} w_{21} y_1^{\frac{1}{k}-1} y_2^{k-1}]^2} \tag{51}$$

$$\frac{\partial J_h}{\partial y_1} = \frac{w_{12} w_{21} \left(\frac{1}{k} - 1\right) y_1^{\frac{1}{k}-2} y_2^{k-1}}{[1 - w_{12} w_{21} y_1^{\frac{1}{k}-1} y_2^{k-1}]^2} \tag{52}$$

$$\frac{\partial J_h}{\partial y_2} = \frac{w_{12} w_{21} y_1^{\frac{1}{k}-1} (k-1) y_2^{k-2}}{[1 - w_{12} w_{21} y_1^{\frac{1}{k}-1} y_2^{k-1}]^2}. \tag{53}$$

The last terms required in the complete expressions in (39) and (41) are all four derivatives $\frac{dy_i}{dw_{k\ell}}$. Two of them are obtained by computing the total derivatives of (47)-(48) with respect to w_{12} (for fixed observations) and solving the resulting two linear equations in $\frac{dy_i}{dw_{12}}$ (this is the same as in [8], but with total derivative *notations*[8]). This yields

$$\frac{dy_1}{dw_{12}} = \frac{-y_2^k}{1 - w_{12} w_{21} y_1^{\frac{1}{k}-1} y_2^{k-1}} \tag{54}$$

$$\frac{dy_2}{dw_{12}} = \frac{w_{21} \frac{1}{k} y_1^{\frac{1}{k}-1} y_2^k}{1 - w_{12} w_{21} y_1^{\frac{1}{k}-1} y_2^{k-1}}. \tag{55}$$

Similarly, computing the total derivatives of (47)-(48) with respect to w_{21} eventually yields

$$\frac{dy_1}{dw_{21}} = \frac{w_{12} k y_1^{\frac{1}{k}} y_2^{k-1}}{1 - w_{12} w_{21} y_1^{\frac{1}{k}-1} y_2^{k-1}} \tag{56}$$

$$\frac{dy_2}{dw_{21}} = \frac{-y_1^{\frac{1}{k}}}{1 - w_{12} w_{21} y_1^{\frac{1}{k}-1} y_2^{k-1}}. \tag{57}$$

Gathering all above results directly yields the correct expressions of the total derivatives $\frac{dJ_h}{dw_{k\ell}}$ in (41), then used to derive the expressions of overall gradient components $\frac{dI(Y)}{dw_{k\ell}}$ in (39), which are not detailed here for the sake of brevity. On the contrary, the set of partial derivatives (50)-(51) was mistakenly used in [8], as if it were the factor $\frac{dJ_h}{dw_{k\ell}}$ of the expression of $\frac{dI(Y)}{dw_{k\ell}}$ (see (8) in [8]).

[8] We here reconsider the approach of [8] in order to show which of its steps should be corrected. Instead, for determining all derivatives $\frac{dy_i}{dw_{k\ell}}$, an alternative approach consists in reformulating the appendix of the current paper, especially (A.5), for the BSS method studied here.

5.2. Linear mixtures

When restricting oneself to 2 mixtures of 2 sources, the specific case of linear mixtures may be obtained as a spin-off of the above investigation: Eq. (43)-(44) show that, when $k = 1$, the mixing model becomes linear. Besides, as shown by (52)-(53), we then have

$$\frac{\partial J_h}{\partial y_1} = 0 \tag{58}$$

$$\frac{\partial J_h}{\partial y_2} = 0. \tag{59}$$

In (41), the total derivative $\frac{dJ_h}{dw_{k\ell}}$ is then equal to the partial derivative $\frac{\partial J_h}{\partial w_{k\ell}}$. This clearly shows that, for the MI-based method too, the problems due to the distinction between these two derivatives concern *nonlinear* mixtures.

6. Conclusion

In the literature, the BSS methods based on ML and MI have been mainly studied for linear mixtures up to now. In this paper, we showed that these methods are more complex for nonlinear mixtures, because they usually lead one to calculate the analytical expressions of the derivatives of the cost functions on which these methods are based (likelihood or information), and these functions involve indirect dependencies. Moreover, the notations commonly employed by the BSS community in such calculations may become misleading when using them for nonlinear mixtures, due to the above-mentioned dependencies. In this paper, we first described the effect of indirect dependencies when using the ML and MI approaches for general (possibly nonlinear) mixing models. We also showed that this effect disappears in the specific case of linear mixtures, which is the reason why it has not been addressed in detail up to now. We eventually focused on two specific nonlinear mixing models, for which two BSS methods were previously proposed. We showed that these methods contain an error because they did not take indirect dependencies into account. We showed how to fix this error and we thus derived the correct expressions of the gradient of the considered cost functions. This paper therefore opens the way to the application of the ML and MI BSS methods to many mixing models, by providing general tools to address such mixtures and explicitly showing how to apply these tools to practical cases.

Appendix A. Derivation of $\frac{ds}{d\theta}$

Let us first analyze the variations of all components of s when a single parameter θ_k is varied, for a fixed observed vector x. Denoting g_i the components of g, Eq. (6) here reads

$$g_i(s, \theta) = x_i = constant \qquad \forall\ i = 1, \ldots, N. \tag{A.1}$$

Each $g_i(s, \theta)$ may be considered as a function which depends on θ_k both directly, i.e. due to θ, and indirectly, i.e. through s. Considering the total derivative of this (constant) function with respect to θ_k therefore yields

$$\frac{\partial g_i}{\partial \theta_k} + \sum_{j=1}^{N} \frac{\partial g_i}{\partial s_j}\frac{ds_j}{d\theta_k} = \frac{dg_i}{d\theta_k} = 0 \qquad \forall\ i = 1, \ldots, N \tag{A.2}$$

from which we then derive $\frac{\partial g_i}{\partial \theta_k}$ with respect to the other terms of (A.2). Gathering all these expressions of $\frac{\partial g_i}{\partial \theta_k}$ for $i = 1, \ldots, N$ in the row vector $\frac{\partial g}{\partial \theta_k}$, Eq. (A.2) yields

$$\frac{\partial g}{\partial \theta_k} = -\frac{ds}{d\theta_k}\frac{\partial g}{\partial s} \tag{A.3}$$

where $\frac{\partial g}{\partial s}$ is the Jacobian matrix of g that we defined after (3), and $\frac{ds}{d\theta_k}$ is the row vector composed of all $\frac{ds_j}{d\theta_k}$, for $j = 1, \ldots, N$. Then gathering, as adjacent matrix rows, the row vectors $\frac{\partial g}{\partial \theta_k}$ corresponding to all parameters θ_k, Eq. (A.3) yields in matrix form

$$\frac{\partial g}{\partial \theta} = -\frac{ds}{d\theta}\frac{\partial g}{\partial s} \tag{A.4}$$

where $\frac{ds}{d\theta}$ is the matrix whose element (i, j) is equal to $\frac{ds_j}{d\theta_i}$. Eq. (A.4) eventually yields

$$\frac{ds}{d\theta} = -\frac{\partial g}{\partial \theta}\left(\frac{\partial g}{\partial s}\right)^{-1}. \tag{A.5}$$

This makes it possible to derive $\frac{ds}{d\theta}$ by only resorting to the partial derivatives of the *mixing* model g, whose analytical expression is assumed to be known, i.e. without using the *inverse* (i.e. separating) model $h = g^{-1}$, whose analytical expression cannot be derived from that of g for many nonlinear models.

References

[1] L.B. Almeida, MISEP - Linear and nonlinear ICA based on mutual information, Journal of Machine Learning Research. 4 (2003) 1297-1318.

[2] A. Belouchrani, K. Abed-Meraim, J.-F. Cardoso, E. Moulines, A blind source separation technique using second-order statistics, IEEE Transactions on Signal Processing 45 (1997) 434-444.

[3] J.-F. Cardoso, A. Souloumiac, Blind beamforming for non Gaussian signals, IEE Proceedings F, 140 (1993) 362-370.

[4] P. Comon, Independent Component Analysis, a new concept ? Signal Processing, 36 (1994) 287-314.

[5] P. Comon and C. Jutten Eds., Handbook of blind source separation. Independent component analysis and applications, Academic Press, Oxford, 2010.

[6] Y. Deville, A. Deville, Maximum likelihood blind separation of two quantum states (qubits) with cylindrical-symmetry Heisenberg spin coupling. Proceedings of ICASSP 2008, pp. 3497-3500, Las Vegas, Nevada, USA, March 30 - April 4, 2008.

[7] Y. Deville, S. Hosseini, Recurrent networks for separating extractable-target nonlinear mixtures. Part I: non-blind configurations. Signal Processing, 89 (2009) 378-393.

[8] L. T. Duarte, C. Jutten, A mutual information minimization approach for a class of nonlinear recurrent separating systems. IEEE International Workshop on Machine Learning for Signal Processing, pp. 122-127, Thessaloniki, Greece, 27-29 Aug. 2007.

[9] M. Gaeta, J.-L. Lacoume, Source separation without a priori knowledge: the maximum likelihood solution. European Signal Processing Conference (EUSIPCO), pp. 621-624, 1990.

[10] S. Hosseini, Y. Deville, Blind maximum likelihood separation of a linear-quadratic mixture. Proceedings of ICA 2004, pp. 694-701, Springer-Verlag, vol. LNCS 3195, Granada, Spain, Sept. 22-24, 2004.

[11] A. Hyvärinen, E. Oja, A fast fixed-point algorithm for independent component analysis, Neural Computation 9 (1997) 1483-1492.

[12] A. Hyvärinen, J. Karhunen, E. Oja, Independent Component Analysis, Wiley, New York, 2001.

[13] C. Jutten, J. Karhunen, Advances in Blind Source Separation (BSS) and Independent Component Analysis (ICA) for Nonlinear Mixtures. International Journal of Neural Systems, 14 (2004) 267-292.

[14] D. T. Pham, P. Garat, Blind separation of mixture of independent sources through a quasi-maximum likelihood approach. IEEE Transactions on Signal Processing, 45 (1997) 1712-1725.

[15] N. Piskunov, Differential and integral calculus, Ed. P. Noordhoff, Groningen, 1965 / N. Piskounov, Calcul différentiel et intégral, vol. 1, chap. VIII, & 10, Ed. Mir, Moscow, 1980, French transl., 11th edition, Ellipses, 1993.

[16] A. Taleb, C. Jutten, Source separation in post-nonlinear mixtures. IEEE Transactions on Signal Processing, 47 (1999) 2807-2820.

[17] A. Taleb, A generic framework for blind source separation in structured nonlinear models. IEEE Transactions on Signal Processing, 50 (2002) 1819-1830.