General conditions of stability in blind source separation models and score function selection

Vincent Vigneron, Ludovic Aubry, Christian Jutten

General conditions of stability in Blind Source Separation models and score function selection

Vincent Vigneron[a,b,c]*, Ludovic Aubry[b] and Christian Jutten[a].

[a]LIS, avenue Félix Viallet, 38031 Grenoble cedex, France.
[b]MATISSE-SAMOS, UMR-CNRS, Panthéon-Sorbonne, Paris, France.
[c]LSC, 40 rue du Pelvoux, 91020 Courcouronnes, France.

Received 9 September 2001, accepted 6 February 2003.

Abstract

This contribution contains a theoretical analysis on asymptotic stability requirements in *blind source separation* (BSS) algorithms. BSS extracts independent component signals from their mixtures without knowing either the mixing coefficients or the probability distributions of the source signals. It is known that some algorithms work surprisingly well. "blind" means that no *a prior* information is assumed to be available both on the mixture and on the sources. This feature make BSS approach versatile because it is not relying on the modeling of some physical phenomena. Nevertheless, few papers mention either convergence or stability of the estimators in the case where one make wrong assumptions on the distribution of the sources. This paper presents and discusses stability conditions for BSS algorithms to avoid spurious stationary points in the case of instantaneous mixtures of independent and identically distributed sources.

*Corresponding author. Fax: +33-1-44-77-19-22. *Email addresses*: `vvigne@lis.inpg.fr` (Vincent Vigneron), `Christian.Jutten@lis.inpg.fr` (Christian Jutten), `ludal@club-internet.fr` (Ludovic Aubry). This work was supported by the French Research Agency (CNRS) and the BLISS project, EU IST-1999-14190.

1 The source separation approach

The problem addressed here is the recovery of m unknown sources, assumed mutually independent, from the observation of a set of linear mixtures.

1.1 Notations and assumptions

Uppercase letters denote random variables, lowercase letters realisations. Consider here the simplest noiseless case where n primary source signals $S(t) = [S_1(t), \ldots, S_n(t)]^T$ are observed only through $m(\geq n)$ instantaneous mixtures of these signals $X(t) = [X_1(t), \ldots, X_n(t)]^T$, given by (for each time instant t)

$$X(t) = AS(t) \tag{1}$$

where $A = (a_{ij})$ is the unknown nonsingular $m \times n$ matrix which does not depend on time t. The symbol T is the transpose operator. This problem is closely related to *independent component analysis* (ICA) introduced by Comon [8].

In the following, it is assumed that the signals are stationary, hence we will omit the time index. The i-th component of S is denoted $S_i(t)$ or S_i (and similarily for the other vectors) and has the probability density function (pdf) $p_{S_i}(S_i)$. With the only assumption that S has independent components, its joint pdf is $p(S) = \prod_{i=1}^n p_{S_i}(S_i)$. Let observe T realizations $x(t)$ of $X(t)$ such that $x(t) = As(t)$. We shall only consider the *noiseless case*. The following assumptions hold throughout:

1. components of $S(t)$ are mutually independent variables with zero mean iid random variables, with non Gaussian marginal distributions and such that
   $\forall i, \mathbb{E}[S_i] = 0$ ($\mathbb{E}[.]$ denote the expectation operator).

2. matrix A is a square full rank matrix ($n = m$).

3. $\hat{S} = BX$ is an estimator of the source signals, which is achieved as soon as BA is a $n \times n$ matrix with exactly one non-zero entry in each row and each column.

Let $g(S) = \prod_{i=1}^n g_{S_i}(S_i)$ (denoted g for simplicity) be a proposed model distribution for S.

1.2 Source separation criterion

Numerous techniques have been designed to identify the mixing matrix A from only observations (see for a review Hyvärinen *et al.* [9] and Lee [10]). Several papers from Cardoso [4], Comon [8], Cardoso and Amari [6] show that it is possible to recover a satisfying estimation of the matrix B even when g is different from p, under certain conditions (such as g_{S_i} being sub-Gaussian if p_{S_i} is). If we know $B = A^{-1}$, the original signals are easily recovered by $BX(t)$.
Therefore, the problem is to estimate B using $X(1), \ldots, X(T)$ from the statistical point of view. The mathematical framework of ICA is formulated in Comon [8]. The basic idea is to minimize the dependency among the output components. The dependency is usually measured by the Kullback-Leibler divergence between the joint and marginal distributions of the outputs or by the related *mutual information* of $BX(t)$, noted $I(BX)$. Typically, we would minimize the criterion $I(BX)$ w.r.t. B. Such a criterion is minimal when components of BX are independent, where $G(\cdot)$ denote the cumulative density function (cdf) associated to g. However, to compute such a criterion in the case of linear mixtures, we need an approximation of the distribution of the sources (the best choice here would be to take for g the exact distribution of the sources). Although a crude approximation of score function does not affect a lot algorithm convergence in linear mixtuures, it no longer holds for nonlinear ones [14].

2 Likelihood of the ICA model

It is not difficult to derive the likelihood in the noise-free ICA model. Let the domain $\{\Omega, \mathcal{A}, (G_B)_{B \in GL_d(\mathbb{R})}\}$ be a statistical model of the sources S: Ω a set of events, $\mathcal{A}$ an algebre on Ω, and $(G_B)_{B \in GL_d(\mathbb{R})}$ an application $\Omega \to \mathbb{R}^n$, differentiable and parametrized by B. In our case, we don't know the true distribution p of the sources, and we don't assume that p belongs to the parametric model $(G_B)_{B \in GL_d(\mathbb{R})}$. That is we are *not* doing maximum likelihood (ML) estimation. Nevertheless, let define a *pseudo log-likelihood* :

$$U_T(B) = -\frac{1}{T} \sum_{t=1}^{T} \log\left(|\det B| g(BX_t)\right), \tag{2}$$

where the vectors $X_t, t = 1, \ldots, T$ are the observations (with the abusive notation $X_t = X(t)$), and g is the density of $G_B(dx) = |\det B| g(Bx)dx$. Statistically, $U_T(B)$ is a *contrast process* if it converges in probability toward a *contrast function* whose minimum is the solution [8]. Recall that, if the minimum is not unique, the discretized form of the criterion remain a *local contrast*, under mild conditions (see Pham [12]). The convergence of the minima towards local minimum of the contrast function has yet not been fully studied and is a major motivation for this work (see [13] for a very first stability analysis of the adaptive Herault-Jutten network). Major contributions in stability analysis are due also to Von Hoff[16, 17], Deville [18], Macchi and Moreau [19], Fort [20] for the special case where the number of source is two, or more generaly by Cardoso and Laheld [21] in the case of n sources where skew symmetric components are major parts of the learning algorithm. In fact, the requirement on $U_T(B)$ is a bit broader since one only needs that its gradient vanishes at the solution in order to define an estimator of A^{-1}.

Lemma 1 *If* $\mathbb{E}[|\log(|\det B| g(BX)))|] < \infty$, $U_T(B)$ *converges in probability (C.P.):*

$$\begin{aligned}
\lim_{t\to\infty} U_T(B) &\overset{C.P.}{\to} -\mathbb{E}[\log(|\det B| g(BX))] \\
&= -\int \log(|\det B| g(BX)) p(s) ds \\
&\geq C(B, A^{-1}) \\
&= K(g\|p) + H(p) + \log|\det A|. \qquad (3)
\end{aligned}$$

where $K(g\|p)$ *denotes the Kullback-Leibler distance between the approximate* g_B *and the true distribution* p *and* $H(p)$ *the entropy of* p. $\triangle$

The process $U_T(B)$ is called a *contrast function*, and $\hat{B}_T = \inf_B U_T(B)$ the *contrast estimate*. $C(B, A^{-1})$ is called a *contrast function* if the application $C(B, A^{-1})$ has a *strict minimum* at the point $B = A^{-1}$.
From inequality (3), it is clear that $C(B, A^{-1}) \geq H(X) + \log|\det A|$ with equality iff the distributions g (the pdf associated to G) and p are the same. Yet, $g \neq p$ and we need to prove that $B = \Lambda A^{-1}$, where Λ is the product of a scale matrix and a permutation matrix, is a minimum.
There are no general (necessary and sufficient) conditions ensuring that

$$\mathbb{E}[|\log(|\det B| g(BX)))|] < \infty, \qquad (4)$$

but we can enumerate several necessary conditions. In the next section, we recall former results given separatly by [1, 2, 12] and demonstrate theoretically why the function $C(B, A^{-1})$ has several minima of the form ΛA^{-1}.

2.1 Candidates for constrast function maxima

Stationary points are the matrices of $GL_d(\mathbb{R})$, such that $dU_T(B) = 0$. Those points are good candidates for maxima and minima of our contrast function [15]. Furthermore, such points are always solutions to our problems.The proof is given in the following. The total differential of our contrast process with respect to the inverse mixing matrix B can be written:

$$\begin{aligned} dU_T(B) &= -d(\log|\det B|) - \frac{1}{T}\sum_{t=1}^{T}\frac{1}{g(BX_t)}d(g(BX_t)) \\ &= -\mathrm{Trace}(dBB^{-1}) - \frac{1}{T}\sum_{t=1}^{T}\phi(BX_t)d(BX_t) \end{aligned}$$

with $\phi(u,\ldots,v) = -\left[\frac{g'(u)}{g(u)},\ldots,\frac{g'(v)}{g(v)}\right]^T$. Denoting $dW = dBB^{-1}$ and $Y_t = BX_t$, we have

$$dU_T(B) = -\ \mathrm{Trace}(dW) - \frac{1}{T}\sum_{t=1}^{T}\phi^T(Y_t)dBB^{-1}Y_t. \tag{5}$$

The mapping dW does not correspond to a change of variable, although it represents a local change of coordinate. As the only points of interest are those of the form ΛA^{-1}, we will see that with the change of parameters $W = BA^{-1}$, the Hessian matrix has a block diagonal form at each stationary points. From the differential of dU_T we have :

$$\frac{\partial dU_T(B)}{\partial B} = -B^{-T} - \frac{1}{T}\sum_{t=1}^{T}\phi(BX_t)X_t^T. \tag{6}$$

Let $\hat{B}_T$ be a solution of $\hat{B}_T^{-1^T} + \frac{1}{T}\sum_{t=1}^{T}\phi(BX_t)X_t^T = 0$ or $I_d + \frac{1}{T}\sum_{t=1}^{T}\phi(B_TX_t)(B_TX_t)^T = 0$. According to Comon's definition [8], C is a contrast function if

$$I_d + \mathbb{E}[\phi(BX)(BX)^T] = 0, \tag{7}$$

for matrices of the form ΛA^{-1} where Λ is the product of a diagonal (scaling) matrix and a permutation matrix $(\delta_{i,\sigma(j)})$. We only need (and can) recover the mixing matrix up to a permutation, thus we only require unicity of the minima up to a permutation and scaling of the matrix A.
Let $\lambda_{i,j}$ be the set of solutions of the integral equations $1+\mathbb{E}[\phi_i(\lambda_{i,j})\lambda_{j,i}s_j] = 0, \forall i,j \in \{1,\ldots,n\}$. For any permutation σ of $\{1,\ldots,n\}$, we define Λ_σ the matrix whose components are $\lambda_{i,\sigma(i)}\delta_{\sigma(i),j}$. Then let $B_\sigma A = \Lambda_\sigma^{-1}$. Yet,

$$I_d + \mathbb{E}[\phi(B_\sigma X)(B_\sigma X)^T] = I_d + \mathbb{E}[\phi(\Lambda_\sigma S)(\Lambda S)^T] \tag{8}$$

That is for each element (i,j) we have $\delta_{i,j} + \mathbb{E}[\phi_i(\Lambda_\sigma S)(\Lambda S)_j^T] = 0$. Further computations lead to:

$$\begin{aligned} D_{i,j} &= \delta_{i,j} + \mathbb{E}\left[\phi_i\left(\sum_k \lambda_{i,\sigma(i)}\delta_{\sigma(i),k}S_k\right)(\Lambda_\sigma S)_j^T\right] \\ &= \delta_{i,j} + \mathbb{E}[\phi_i(\lambda_{i,\sigma(i)}S_{\sigma(i)})\lambda_{j,\sigma(j)}S_{\sigma(j)}] \\ &= \begin{cases} 0 & \text{if } i \neq j \\ 1+\mathbb{E}[\phi_i(\lambda_{i,\sigma(i)}S_{\sigma(i)})\lambda_{j,\sigma(j)}S_{\sigma(j)}] = 0 & \text{if } i = j \end{cases} \end{aligned} \tag{9}$$

2.2 Existence of solutions

In section 2.1, we demonstrated that B_σ is a good candidate for a local minimum. Let us prove the *existence* of such solutions (or some conditions on the distributions p and g) and the *unicity*. In this goal, we examine if the Hessian matrix $\mathbb{E}[\frac{\partial^2 U_T(B)}{\partial B^2}]$ is positive definite. This may not always be the case. In [3], Amari proposed a modification of the algorithm so that the Hessian becomes always positive definite. If such a stability point B_σ exists (satisfying $B^{-1} + \mathbb{E}[\phi(BX)X] = 0$), this is achieved if $\frac{\partial^2 U_T(B)}{\partial B^2} \geq 0$, *i.e.*

$$B^{-2} - \frac{1}{T}\sum_{t=1}^{T}\phi'(BX_t)X_t^T X_t \geq 0. \tag{10}$$

2.2.1 Hessian matrix form

Noting that $\frac{\partial b^{-1}_{k\ell}}{\partial b_{ij}} = -b^{-1}_{jk} b^{-1}_{\ell i}$, the entries of the Hessian matrix $\nabla^2_B U$ can be written

$$\frac{\partial^2 U_T(B)}{\partial b_{ij}\partial b_{k\ell}} = -\frac{\partial}{\partial b_{ij}}\left[b^{-1}_{k\ell} + \frac{1}{T}\sum_{t=1}^{T}\phi_k(BX_t)(X_t)_\ell\right] \tag{11}$$

$$(\nabla^2_B U)_{ijk\ell} = b^{-1}_{\ell i} b^{-1}_{jk} - \frac{1}{T}\sum_{t=1}^{T}\phi'(BX_t)(X_t)_\ell (X_t)_j \delta_{ik}. \tag{12}$$

B_σ is a strict minimum of $C(B, A^{-1})$ iff $\mathbb{E}[\nabla^2_B]$ is positive definite, *i.e.*

$$\mathbb{E}[(\nabla^2_B U)_{ijk\ell}] = (\Lambda A^{-1})^{-1}_{\ell i}(\Lambda A^{-1})^{-1}_{jk} - \mathbb{E}[\phi'(\Lambda S)X_\ell X_j]\delta_{ik} \tag{13}$$

$$= (A\Lambda^{-1})_{\ell i}(A\Lambda^{-1})_{jk} - \mathbb{E}[\phi'(\Lambda S)X_\ell X_j]\delta_{ik} \geq 0 \tag{14}$$

Assume Λ is a diagonal matrix (recall that Λ is the solution of $I_d + \mathbb{E}[\phi(\Lambda S)(\Lambda S)^T] = 0$), hence

$$\mathbb{E}[(\nabla^2_B U)_{ijk\ell}] = a_{\ell i}a_{jk}\frac{1}{\lambda_i\lambda_k} - \mathbb{E}[\phi'_k(\lambda_k S_k)X_\ell X_j]\delta_{ik} \tag{15}$$

$$= a_{\ell i}a_{jk}\frac{1}{\lambda_i\lambda_k} - \sum_{p,q} a_{\ell p}a_{jq}\mathbb{E}[\phi'_k(\lambda_k S_k)S_p S_q]\delta_{ik} \tag{16}$$

$$= a_{\ell i}a_{jk}\frac{1}{\lambda_i\lambda_k} - \sum_{p} a_{\ell p}a_{jq}\mathbb{E}[\phi'_k(\lambda_k S_k)S_p^2]\delta_{ik} \tag{17}$$

because $p \neq q \Rightarrow \mathbb{E}[\phi'_k(\lambda_k S_k)S_p S_q] = 0$. Note that if $Q(B) = \sum_{ijk\ell} b_{ij}b_{k\ell}(\nabla^2_B U)_{ijk\ell}$ is a positive definite quadratic form, then $W \to Q(WA^{-1})$ and

$$Q(WA^{-1}) = \sum_{ijk\ell}\sum_{pq} w_{ip}w_{kp}a^{-1}_{pj}a^{-1}_{q\ell}(\nabla^2_B U)_{ijk\ell} \tag{18}$$

$$= \sum_{ipkq} w_{ip}w_{kp}U_{ipkq}, \tag{19}$$

with $U_{ipkq} = \sum_{j\ell} a^{-1}_{pj}a^{-1}_{q\ell}(\nabla^2_B U)_{ijk\ell}$.

So it is equivalent to prove that

$$\sum_{k,\ell} a_{uj}^{-1} a_{v\ell}^{-1} \mathbb{E}[(\nabla_B^2 U)_{ijk\ell}(\Lambda A^{-1})] = \sum_{j,\ell} a_{uj}^{-1} a_{v\ell}^{-1} a_{\ell i} a_{jk} \frac{1}{\lambda_i \lambda_k} -$$
$$- \sum_{j,\ell} a_{uj}^{-1} a_{v\ell}^{-1} a_{\ell p} a_{jp} \mathbb{E}[\phi_k'(\lambda_k S_k) S_p^2] \delta_{ik}$$
$$U_{iukv} = \delta_{uk}\delta_{vi} \frac{1}{\lambda_i \lambda_k} - \sum_p \delta_{up}\delta_{vp} \mathbb{E}[\phi_k'(\lambda_k S_k) S_p^2] \delta_{ik} \tag{20}$$
$$U_{ijk\ell} = \delta_{jk}\delta_{i\ell} \frac{1}{\lambda_i \lambda_k} - \mathbb{E}[\phi_k'(\lambda_k S_k) S_j^2] \delta_{j\ell}\delta_{ik} \tag{21}$$

is positive definite. Suppose the transformed Hessian $U_{ijk\ell}$

$$U_{ijk\ell} = \delta_{jk}\delta_{i\ell} \frac{1}{\lambda_i \lambda_k} - \mathbb{E}[\phi_i'(\lambda_i S_i) S_j^2] \delta_{j\ell}\delta_{ik} \tag{22}$$

has a matrix form as

$$U = \begin{bmatrix} \ddots & & & & & \\ & U_{ijij} & U_{jiij} & & 0 & \\ & U_{ijji} & U_{jiji} & & & \\ & & & \ddots & & 0 \\ & 0 & & & U_{iiii} & \\ & & & 0 & & \ddots \end{bmatrix} \tag{23}$$

which leads to define for all $i < j$:

$$\kappa_{ij} = -\mathbb{E}[\phi_i'(\lambda_i S_i) S_j^2] \tag{24}$$
$$\alpha_{ij} = \frac{1}{\lambda_i \lambda_j} \tag{25}$$
$$U_{ij} = U_{ijij} = \begin{pmatrix} \kappa_{ij} & \alpha_{ij} \\ \alpha_{ij} & \kappa_{ji} \end{pmatrix} \tag{26}$$
$$U_i = U_{iiii} = \alpha_{ij} + \kappa_{ij}, \tag{27}$$

The simplified solution where $\lambda_i = 1$ is

$$U_{ij} = \begin{pmatrix} -\mathbb{E}[\phi_i'(S_i)]\mathbb{E}[S_j^2] & 1 \\ 1 & -\mathbb{E}[\phi_j'(S_j)]\mathbb{E}[S_i^2] \end{pmatrix}$$
$$U_i = 1 - \mathbb{E}[\phi_i'(\lambda_i S_i) S_i^2]$$

It is straightforward to derive from equations (26) and (27) the stability conditions, (i) $U_i < 0$ (see Amari *et al.* [3]), (ii) the real part of the eigenvalues of U_{ij} are negatives (U_{ij} being symetric, its eigenvalues are real). The eigenvalues of U_{ij} are the solutions of

$$(\kappa_{ij} - x)(\kappa_{ji} - x) - \alpha_{ij}^2 = 0 \tag{28}$$

i.e.

$$x_{1,2} = \frac{1}{2}\left(\kappa_{ij} + \kappa_{ji} \pm \sqrt{(\kappa_{ij} - \kappa_{ji})^2 + 4\alpha_{ij}^2}\right). \tag{29}$$

The conditions $x_1 < 0$ and $x_2 < 0$ implies from Eq. 29 that $x_1 + x_2 = \kappa_{ij} + \kappa_{ji}) < 0$ and

$$\kappa_{ij} + \kappa_{ji} < \pm\sqrt{(\kappa_{ij} - \kappa_{ji})^2 + 4\alpha_{ij}^2} \tag{30}$$

$$(\kappa_{ij} + \kappa_{ji})^2 > (\kappa_{ij} - \kappa_{ji})^2 + 4\alpha_{ij}^2 \tag{31}$$

or $-\kappa_{ij}\kappa_{ji} > \alpha_{ij}^2$, which is equivalent to write

$$-\mathbb{E}\left[\,\phi_i'(\lambda_i S_i)(\lambda_j S_j)^2\,\right]E\left[\,\phi_j'(\lambda_j S_j)(\lambda_i S_i)^2\right] > 1. \tag{32}$$

Formula (32) allows to check the stability conditions online (see [3]), by estimating the values $\mathbb{E}[\phi_i'(\lambda_i S_i)]$ and $\mathbb{E}[(\lambda_j S_j)^2]$ with respectively $\frac{1}{T}\sum_{t=1}^{T}\phi_i'(\hat{B}_T X_t)$ and $\frac{1}{T}\sum_{t=1}^{T}(\hat{B}_T X_t)^2$. Similar but less general proof can be found in [7].

3 More on stationary points

In this section, we exhibit cases where we can obtain false solutions. In the next section, we will show that the algorithm converges towards *any* point of stability, we have already shown that some interesting stable points are good solution of our problem. However, we need to ensure that we don't converge toward any other stationary point that does not correspond to interesting solutions.

3.1 Unicity

Let g be a probability density function with enough regularity such that $\log g$ has a power series expansion which converges over all its domain of definition, say $\log g(X) = \sum_{k=0}^{\infty} \lambda_k X^k$. Then the score function ϕ is

$$\phi(X) = \frac{g'(X)}{g(X)} = \frac{d}{dX} \log g(X), \tag{33}$$

which can be rewriten $\phi(X) = \sum_{k=1}^{\infty} k\lambda_k X^{k-1} = \sum_{k=1}^{\infty} \mu_k X^{k-1}$ if $\mu_k = k\lambda_k$. Thus solving the equation $1 + \mathbb{E}[\phi(\lambda X)\lambda X] = 0$ reduces to solve

$$1 + \sum_{k=0}^{\infty} \mu_k \lambda^k \mathbb{E}[X^k] = 0. \tag{34}$$

As g is a pdf over $\mathbb{R}$, then necessarily $\eta(\{X | g(X) > 1\}) < 1$ (η being a Lebesgue measure). As a particular case, we will focus on densities of the exponential form $g(X) = C \exp(P(X))$, where C is a normalization constant, and P is a polynomial. In order for g to be a probability density function, it must verify $\int_{-\infty}^{+\infty} \exp(P(X)) dX = \frac{1}{C}$. A sufficient and necessary condition is that the largest nonzero power of P be even and associated with a negative coefficient. Hence, we add a condition on P such that $\int_{-\infty}^{+\infty} CX \exp(P(X)) dX = 0$. With the usual conventions, $Y = BX$, $X = AS$ ($Y(t) = [Y_1(t), \ldots, Y_n(t)]^T$ is a random variable). Let's denote $M = BA$. Since the sources S are supposed to be independent, we try to obtain all possible solutions of M such that

$$1 + \mathbb{E}[\phi(MS)(MS)^T] = 0. \tag{35}$$

Now let us see what we obtain if P is a simple polynomial, for example $P(X) = -\frac{1}{2k} X^{2k}$. Then $\phi(X) = -X^{2k-1}$, with k an integer. Solving Eq. (35) is similar to solve for each couple (i,j) the equation $\delta_{ij} - \mathbb{E}[Y_i^{2k} Y_j] = 0$. Using the multilinearity property of the moments and mutually independence, we have

$$\delta_{ij} = \sum_{a_1,\ldots,a_k,b} M_{i,a_1} \ldots M_{i,a_k} M_{jb} \mathbb{E}[S_{a_1} \ldots S_{a_k} S_b] \tag{36}$$

$$\delta_{ij} = \sum_{a_1,\ldots,a_k,b} M_{i,a_1} \ldots M_{i,a_k} M_{jb} \delta_{a_1,\ldots,a_k,b} \tag{37}$$

$$\delta_{ij} = \sum_{a} M_{i,a}^k M_{j,a} \tag{38}$$

Thus, for $\phi(X) = \sum_{p=0}^{k} \lambda_p X^p$, we have

$$\delta_{ij} = \sum_a \left(\sum_{p=0}^{k} \lambda_p M_{i,a}^p \right) M_{j,a}. \tag{39}$$

Example 1 *Consider the case the case of 2 signals with* $\phi(x) = -x^3$ *and consider the product matrix* $M = BA = \begin{bmatrix} a & b \\ c & d \end{bmatrix}$. *Equation (35) can be written in this case*

$$\begin{bmatrix} a^3 & b^3 \\ c^3 & d^3 \end{bmatrix} \begin{bmatrix} a & c \\ b & d \end{bmatrix} = \begin{bmatrix} a^4 + b^4 & a^3c + b^3d \\ c^3a + d^3b & c^4 + d^4 \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \tag{40}$$

By simple algebraic operations, one can prove that the following table contains **some** *exact solutions*

a	b	c	d
1	0	0	1
1	0	0	-1
1	0	0	ρ
ρ	0	0	1
ρ	0	0	-1
ρ	0	0	ρ
-1	0	0	-1
-1	0	0	ρ
0	1	1	0
0	1	-1	0
0	1	ρ	0
0	ρ	1	0
0	ρ	-1	0
0	ρ	ρ	0
0	-1	-1	0
0	-1	ρ	0
$-\rho$	ρ	ρ	ρ
ρ	ρ	ρ	$-\rho$
ρ	ρ	ρ	ρ

where ρ *is the imaginary number (such that* $\rho^2 = -1$*). One can easily verify that the combinations* $a = c = d = -b = \frac{1}{\sqrt[4]{2}}$, $a = -c = d = b = \frac{1}{\sqrt[4]{2}}$ *and*

$a = c = -d = b = \frac{1}{\sqrt[4]{2}}$ are also solutions of the previous equation. The last three solutions are not simple permutations each one from each other and are still mixing the sources.
But consider the use of the following function $g(x) = C\exp(-\frac{x^4}{8})$ the score function is $\phi(x) = -\frac{1}{2}x^3$ and the last equation becomes

$$\frac{1}{2}\begin{bmatrix} a^4 + b^4 & a^3c + b^3d \\ c^3a + d^3b & c^4 + d^4 \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}. \tag{41}$$

In that case, the matrix $M = \begin{bmatrix} 1 & 1 \\ -1 & 1 \end{bmatrix}$ is a trivial solution which gives false results [11].
Not so simple computations leads to a solution of the more general equation

$$\begin{bmatrix} a^{k+1} + b^{k+1} & a^kc + b^kd \\ c^ka + d^kb & c^{k+1} + d^{k+1} \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \tag{42}$$

that is:

$$a = b = \sqrt[k+1]{\frac{1}{2}} \tag{43}$$

$$c = -d = \sqrt[k+1]{\frac{1}{2}} \tag{44}$$

Whether increasing the number of sources will likely lead to more or less of such mixing solutions is not completely clear. However, as the dimension increases, we are looking for roots of polynomials of more variables. Lastly, we have to look at the Gaussian case, where $\phi(x) = -x$. The roots of the equation are solutions of $MM^T = \mathcal{I}_d$, that is we can obtain any rotation of the sources as a solution. It is clear that for a particular function ϕ, we may have a infinite number of spurious solutions, which is a well-known result [8].

3.2 Conditions for unicity

Previous section 3.1 demonstrates that, with some kinds of g functions, stationary points exist which do not correspond to the solutions. In this section we propose a theoretical criterion to choose the correct pdf $g = p$. If we know the distribution of sources, we can try to find the solutions of (see Eq. 35)

$$I = -\mathbb{E}[\frac{p'(X)}{p(X)}X^T] \tag{45}$$

which can be formulated also as

$$\delta_{ij} = -\int \frac{p'_{S_i}(MS)}{p_{S_i}(MS)}(MS)_j p(S) dS \tag{46}$$

$$= -\int \frac{p'_{S_i}(u)}{p_{S_i}(u)} u_j |\det M^{-1}| p(M^{-1}u) du, \tag{47}$$

3.2.1 Sensitivity on the function g

Suppose that pdf p is such that no possible wrong solutions are possible, then we can try to evaluate the sensibility of the solution obtained using g when the difference $\|p - g\|$ vanishes. Let define for one function f and for some definitive positive matrix Γ the function

$$T_{p,f}(\Gamma) = \int_u f(u) |\det \Gamma| p(\Gamma u) du. \tag{48}$$

Remark that the form of $T_{p,f}$ is similar to the one given in Eq. 47. And let Γ_f be a solution of the equation $T_{p,f}(\Gamma) = 0$. Note that T is linear in f.

Let define $d_{\Gamma,f,h}$ a distance measure between two distribution, such that

$$d_{\Gamma,f,h} = \|T_{p,f}(\Gamma) - T_{p,g}(\Gamma)\| \tag{49}$$

$$= \left\| \int_u (f(u) - h(u)) |\det \Gamma| p(\Gamma u) du \right\| \tag{50}$$

$$\leq \int_u \|f(u) - h(u)\| \, |\det \Gamma| p(\Gamma u) du. \tag{51}$$

Using infinite norm $\|\cdot\|_\infty$:

$$d_{\Gamma,f,h} \leq \|f(u) - h(u)\|_\infty \times 1, \tag{52}$$

because $\int_u p(x)dx = 1$. Moreover, the p-norm (with $p > 1$) $\|\cdot\|_p$ applied to Eq. (50) using Hölder inequality gives

$$d_{\Gamma,f,h} \leq \|f(u) - h(u)\|_p \left(\int_u (|\det \Gamma| p(\Gamma u))^q du \right)^{\frac{1}{q}}. \tag{53}$$

Because $\int_u (|\det \Gamma| p(\Gamma u))^q du = |\det \Gamma|^{q-1} \int_v p^q(v) dv$, the right term of Eq. (53) measure the distance between f and h using $\|\cdot\|_p$-norm, Then for all $p \geq 1$ and $\frac{1}{p} + \frac{1}{q} = 1$ (p could be ∞),

$$d_{\Gamma,f,h} \leq \|f(u) - h(u)\|_p |\det \Gamma|^{q-1} \|p\|_q, \tag{54}$$

for instance,

$$d_{\Gamma,f,h} \leq \|f(u) - h(u)\|_1 |\det \Gamma| \|p\|_\infty. \tag{55}$$

***Remark* 1** *Of course, in our case, we would have* $f(Y) = 1 + \frac{p'(Y)}{p(Y)} Y^T$ *and* $h(Y) = 1 + \phi(Y) Y^T$. □

Further, if Γ_f is unique solution of $T_{p,f}(\Gamma) = 0$, and if the mapping $\Gamma \to T_{p,f}(\Gamma)$ defines a continuous application from $\mathbb{R}^{n\times n}$ to $\mathbb{R}$, then Γ should converge to Γ_f as $\|f - h\| \to 0$.

3.2.2 Exploiting the Jacobian sensitivity

An other assumption is that $d_{\Gamma,f,h}$ is invertible and derivable at the first order (at least, we have computed an explicit form in the section 3.1). In that case, we can find a neighborhood $\mathcal{V}$ of Γ_f inside which $d_{\Gamma,f,h}$ is invertible. The Taylor expansion of $d_{\Gamma,f,h}^{-1}$ around 0 at the first order is:

$$d_{\Gamma,f,h}^{-1}(0) = d_{\Gamma,f,h}^{-1}(0) + {d_{\Gamma,f,h}^{-1}}'(0)\alpha + \mathcal{O}(\alpha), \tag{56}$$

where $\mathcal{O}(\cdot)$ is the integral rest of the formula and $\alpha = d_{\Gamma,f,h}$. $d_{\Gamma,f,h} = 0$ if $\Gamma = \Gamma_f$. Then

$$\Gamma = \Gamma_f + [d'_{\Gamma,f,h}(0)]^{-1}\alpha + \mathcal{O}(\alpha). \tag{57}$$

Back to BSS problems, we can replace:

$$\Gamma = \hat{B}A \tag{58}$$

$$\Gamma_f = I \tag{59}$$

$$\alpha = d_{\Gamma,f,h} \tag{60}$$

$$d'_{\Gamma,f,h} = \mathcal{H}. \tag{61}$$

Recall that $\mathcal{H}$ has the same matrix representation as the Hessian in the previous section, but it is seen as a linear mapping instead of a quadratic form. From the equation (57), we draw:

$$\hat{B}A = I + \mathcal{H}^{-1}\alpha + \mathcal{O}(\alpha) \tag{62}$$

We can straigthforward derive if $\lim_{\alpha \to 0} \mathcal{O}(\alpha) = 0$

$$\left\| \hat{B}A - I \right\| \leq \|\mathcal{H}^{-1}\| \|\alpha\|, \tag{63}$$

we can *avoid* wrong stationary points if we choose a *close enough* function g. Replacing α in Eq. (63) by its value resp. in Eq. (60) and (55), we find:

$$\left\| \hat{B}A - I \right\| \leq \|\mathcal{H}^{-1}\| \left(\|f - h\|_p \left| \det(\hat{B}A - I) \right|^{q-1} \|p\|_q \right)$$

Choosing $p = \infty$ and $q = 1$, it comes (after grouping into the same member):

$$\left\| \hat{B}A - I \right\| \leq \|\mathcal{H}^{-1}\| \; \|f - h\|_\infty. \tag{64}$$

The final theorem (64) which put in evidence the factor $\mathcal{H}^{-1}$ is useful and advantageous for many different reasons among which the following: (i) it provides a short summary of the way the iterative scheme progresses, (ii) it allows score functions selection, (iii) it proposes stability conditions of the criterion used,etc...

4 Conclusion

The last equation provides a boundary on the approximation of the separating matrix in case of BSS problems. It shows theoretically that we can avoid wrong stationary points *(1)* for a close enough score function even if the model distribution chosen is wrong whether $\|\mathcal{H}\|$ is big or not,*(2)* for a crude score function estimated if $\|\mathcal{H}\|$ is small. This point should be the subject of a further study as well as the efficiency of the algorithm.

References

[1] S.I. Amari and A. Cichoki and H.H. Yang, “A new learning algorithm for blind signal separation”, Neural Information Processing Systems, Eds D.S. Toureysky et al., pp. 757-763, 1995.

[2] S.I. Amari , “Superefficiency in blind source separation”,*IEEE Signal Processing*, Vol. 47(4), pp. 936-944, 1999.

[3] S.I. Amari, T.P. Chen and A. Cichoki, “Stability analysis of adaptive blind source separation”, Neural Networks, Vol. 10(8), pages 1345-1351, 1997.

[4] J.-F. Cardoso, “Statistical principles of source separation”, In *Proc. of SYSID’97, 11th IFAC Symposium on system identification*, Fukuoka (Japan), pages 1837–1844, 1997.

[5] J.-F. Cardoso, “On the stability of some source separation algorithms”, In *Proc. of the 1998 IEEE SP workshop on Neural Networks for Signal Processing (NNSP’98)*, pp. 13–22, 1998.

[6] J.-F. Cardoso and S.I. Amari, ”Maximum likelihood source separation: equivariance and adaptativity”. In *Proc. of SYSID’97, 11th IFAC Symposium on system identification*, Fukuoka (Japan), pages 1063–1068, 1997.

[7] J.-F. Cardoso, “Blind signal separation: statistical principles”, *Proceedings of the IEEE*, vol. 90, n. 8, pp. 2009-2026, Oct. 98, Special Issue on Blind Identification and Estimation, R.-W. Liu and L. Tong editors.

[8] P. Comon, “Independent component analysis. A new concept ?”, *Signal Processing*, V. 36, pp. 287-314, 1994.

[9] A. Hyvärinen, J. Karhunen and E. Oja, “Independent Component Analysis”, Wiley, 2001.

[10] T.W. Lee, *Independent Component Analysis and applications*, Kluwer, 1998.

[11] A. Mansour and C. Jutten, “What should we say about the kurtosis?”, *IEEE Signal Processing letters*, Vol. 6(12), pp. 321-322, December 1999.

[12] D.T. Pham, “Blind separation of instantaneous mixtures of sources based on order statistics”, *IEEE Signal Processing*, Vol. 48(2), pp. 1712-1725, 2000.

[13] E. Sorouchyari, “Blind separation of sources, Part III: stability analysis”, *Signal Processing*, 24, pp. 21–29, 1991.

[14] A. Taleb and C. Jutten,"Source separation in post-nonlinear mixtures", IEEE transactions on Ssignal Processing, 47(10):2807–2820, October 1999.

[15] V. Vigneron and L. Aubry, "More on stationary points in independent component analysis", *ESANN'2001 Proceedings*, Bruges, 25-27, April 2001.

[16] T.P. Von Hoff , A. Lindgren and A. N. Kaelin, "Transpose properties in the stability and performance of the clasic adaptive algorithms for blind source separation and deconvolution", *Signal Processing*, 80(9), 1807-1822, 2000.

[17] T.P. Von Hoff , " Transpose properties in the stability and performance of the classic adaptive algorithms for blind source separation and deconvolution", PhD Thesis, ETH Zurich, 2000.

[18] Y. Deville, " A unified stability analysis of the Herault-Jutten source separation neural network", *Signal Processing*, 51(3), pp. 229-233, 1996.

[19] O. Macchi and E. Moreau, "Self-adaptive source separation, Part I: Covergence analysis of a direct Linear Network controled by the Herault-Jutten algorithm", *IEEE Trans. on Signal Processing*, 45(4), 1997.

[20] J.C. Fort, "Stabilit de l'algorithme de sparation de sources de Jutten et Hrault", *Traitement du signal* (in French), 8(1), 1992.

[21] J.F. Cardoso and B. Laheld, "Equivariant adaptive source separation", *IEEE Trans. on Signal Processing*, 44(12), 1996.