ESANN'2001 proceedings - European Symposium on Artificial Neural Networks
Bruges (Belgium), 25-27 April 2001, D-Facto public., ISBN 2-930307-01-3, pp. 39-44

More on stationnary points in Independent Component Analysis

Vincent Vigneron[1] Ludovic Aubry [2]

[1]MATISSE-SAMOS UMR 8595
90, rue de Tolbiac
75634 Paris cedex 13

[2]CEMIF
40 rue du Pelvoux
91020 Evry Courcouronnes

Abstract. In this paper, we will focus on the problem of blind source separation for *independent and identically distributed* variables (*iid*). The problem may be stated as follows: we observe a linear (unknown) mixture of k *iid* variables (the sources), and we want to recover either the sources or the linear mapping. We give online stability conditions of the algorithm using the eigenvalues of the hessian matrix of the pseudo-likelihood matching our set of observations.

1 Introduction

The process of discovering the linear mapping is called Blind Source Separation (BSS), because we don't want to make any asumption on the distribution of the sources, except that they are mutually independent. A review of several methods recovering the mapping can be found in [2]. Tee-Wong Lee shows that many of these methods are¡equivalent to finding the maximum of entropy of the observations assuming a given distribution of the sources [6]. Cardoso & Amari show that the mixture can be recovered even with wrong assumptions on the distributions, provided they satisfy a few stability conditions.

We consider the case where we observe a d-dimensional random vector X. The vector X is obtained from a linear (non-random) mapping of a source variable S. Thus we have $X(t) = AS(t)$ for all of our observations. Let $S(t) = [s_1(t), \dots, s_d(t)]$ be the sources such that for each instant t, the source $s_i(t)$ has the probability density fuction (pdf) p_i (independent of t). With the (*only*) assumption that S has independent components, its joint pdf is $p = \prod_{i=1}^{d} p_i$.
Let observe n realisations $x(t)$ of the variables X such that $x(t) = A \cdot s(t)$, for $t = 1, \dots, n$ and an unknown $d \times d$ matrix A[1]. Further, we will propose a model distribution for S, noted $g = \prod_{i=1}^{d} g_i$. Several papers from Comon [5], Cardoso

[1]If we assume there is as much sources as captors.

ESANN'2001 proceedings - European Symposium on Artificial Neural Networks
Bruges (Belgium), 25-27 April 2001, D-Facto public., ISBN 2-930307-01-3, pp. 39-44

[3], Amari [1] show that it is possible to recover a satisfying estimation of the matrix A^{-1} even when g is different from p, under certain conditions (such as g_i being sub-gaussian if p_i is).
In order to obtain an estimator of A, we have to minimize a given criterion, most of them are based on entropy and mutual information. Typically we woud minimize the criterion $H(BX)$, or $\mathbb{E}[\log g(BX)]$ with respect to B. The best choice here would be to take for g the distribution of the sources[2]. Such a criterion is minimal when each coordinate of the vector BX are independent and the distribution of $G(BX)$ is uniform (G is the cumulative density function of g).
Works from Cardoso [3] and Cardoso & Amari [4] give conditions on g and p so that the minimisation algorithm is stable. The conditions show that a broad range of function g can be used.

2 Maximum likelihood

Let us define a likelihood function, matching our set of observations: let $\{\Omega, \mathcal{A}, (G_B)_{B\in GL_d(\mathbb{R})}\}$ be a statistical model of the sources s. In our case, we don't know the true distribution P ($P(ds) = p(s)ds$) of the sources s, and we don't assume that P belongs to our parametric model $(G_B)_{B\in GL_d(\mathbb{R})}$. That is we are not doing maximum likelihood estimation. Nevertheless, we will define a pseudo log-likelihood with:

$$U_n(B) = -\frac{1}{n}\sum_{i=1}^{n} \log(|\det B| g(BX_i)),$$

where the variables $(X_i)_{i=1,\dots,n}$ are the observations, and $g(y) = \prod_{k=1}^{d} g_k(x_k)$, is the density of $G_B(dx) = |\det B|g(Bx)dx$. This sequence of functions of the observations is called a *contrast processus* if it verifies some simple properties. The main condition is that it converge in probability toward a contrast function whose minimum is our solution. In fact, the requirement on $U_n(B)$ is a bit broader, because as we will see later, we only need that its gradient cancels at the solution point in order to define a sequence of estimators of A^{-1}.

Lemma 1 *if $\mathbb{E}[|\log(|\det B|g(BX))|] < \infty$, we have the convergence in probability P of :*

$$\begin{aligned}\lim_{n\to\infty} U_n(B) &= -\mathbb{E}[\log(|\det B|g(BX))] \\ &= -\int \log(|\det B|g(BAS))p(s)ds \\ &\geq C(B, A^{-1}) \\ &= \mathcal{K}(g_{BA}\|p) + \mathcal{H}(p) + \log|\det B|. \qquad (1)\end{aligned}$$

[2]Recently, *quasi-optimal methods* with online estimation of pdf's and of score functions have been published.

ESANN'2001 proceedings - European Symposium on Artificial Neural Networks
Bruges (Belgium), 25-27 April 2001, D-Facto public., ISBN 2-930307-01-3, pp. 39-44

$C(B, A^{-1})$ *is called a* contrast function *if* $B \to C(B, A^{-1})$ *has a strict minimum at the point* $B = A^{-1}$. *The processus* $U_n(B)$ *is called a* contrast processus, *and* $\hat{B}_n = \inf_B U_n(B)$ *is called the* contrast estimate.
From inequality (1), it is clear that $C(B, A^{-1}) \geq \mathcal{H}(p) + \log|\det B|$ *with equality only if the distributions* g_{BA} *and* p *are the same. Yet, we have* $g \neq p$ *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 is no general conditions ensuring that

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

but we can enumerate several necessary conditions.

We show that the function $C(B, A^{-1})$ has several minima of the form ΛA^{-1}.

3 Stationnary points

We call stationnary points, the matrices of $GL_d(\mathbb{R})$, such that $dU_n(B) = 0$. Those points are good candidates for maxima and minima of our contrast function. Furthermore, we show that there exists such points that are always solutions to our problems.
Let's compute the total differential of our contrast process with respect to the inverse mixing matrix B:

$$U_n(B) = -\frac{1}{n} \sum_{i=1}^{n} \log(|\det B g(BX_i))$$

then

$$dU_n(B) = -\mathrm{Trace}(dB \cdot B^{-1}) - \frac{1}{n} \sum_{i=1}^{n} \phi^T(BX_i) d(BX_i)$$

with $\phi(x_1, \ldots, x_d) \leq -\left[\frac{g'(x_1)}{g(x_1)}, \ldots, \frac{g'(x_d)}{g(x_d)}\right]^T$.

***Remark* 1** *If we define the mapping* dW *as* $dW = dB \cdot B^{-1}$, *and* $Y_i = BX_i$, *we have*

$$\begin{aligned} dU_n(B) = & -\mathit{Trace}(dW) - \tfrac{1}{n}\textstyle\sum_{i=1}^{n} \phi^T(Y_i) dB B^{-1} X_i \\ = & -\mathit{Trace}(dW) - \tfrac{1}{n}\textstyle\sum_{i=1}^{n} \phi^T(Y_i) dW Y_i. \end{aligned}$$

This mapping does not correspond to a change of variable, although it represents a local change of coordinate. As the only points of interest for us 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 stationnary points.*

From the differential of dU_n we have :

$$\frac{\partial U_n(B)}{\partial B} = -B^{-T} - \frac{1}{n} \sum_{i=1}^{n} \phi(BX_i) X_i^T.$$

Let us define $\hat{B}_n$ as a solution of $\hat{B}_n^{-T} + \frac{1}{n}\sum_{i=1}^n \phi(\hat{B}_n X_i)X_i^T = 0$ or $I_d + \frac{1}{n}\sum_{i=1}^n \phi(\hat{B}_n X_i)(\hat{B}_n X_i)^T = 0$. In order C to be a contrast function (according to Comon's definition [5]), it must verify :

$$\begin{array}{rcll} \nabla_B C(B, A^{-1}) = 0 & \Leftrightarrow & \mathbb{E}[-\nabla l_n(B)]B^T & = 0 \\ & \Leftrightarrow & I_d + \mathbb{E}[\phi(BX)(BX)^T] & = 0 \end{array}$$

for matrices of the form ΛA^{-1} where Λ is the product of a diagonal (scaling) matrix and a permutation $(\delta_{i,\sigma(j)})$. This is equivariant property as introduced by Comon [5]: 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 the set of $\lambda_{i,j}$ be solutions of the integral equations $1+\mathbb{E}[\phi_i(\lambda_{i,j}s_j)\lambda_{i,j}s_j] = 0$. For any permutation σ of $\{1,\ldots,d\}$, we define Λ_σ the matrix whose components are $\lambda_{i,\sigma(i)\delta_{\sigma(i),j}}$. Let $B_\sigma = \Lambda_\sigma A^{-1}$, then:

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

that is for each element (i,j) we have $\delta_{i,j} + \mathbb{E}[\phi_j(\Lambda_\sigma S)(\Lambda S)_j^T] = 0$. The left member of the equation:

$$D_{i,j} = \left\{ \begin{array}{ll} 0 & \text{if } i \neq j \\ 1 + \mathbb{E}[\phi_i(\lambda_{i,\sigma(i)}s_{\sigma(i)})\lambda_{j,\sigma(j)}s_{\sigma(j)}] = 0 & \text{if } i = j \end{array} \right.$$

We yet have to prove the existence of such solutions (or some conditions on the distributions p and g) and the uniqueness.

3.1 Stability conditions

We just showed B_σ is a good candidate for a local minimum, we need to prove that the hessian matrix $\mathbb{E}[-\nabla^2 l_n(B_\sigma)]$ is positive definite. This may not always be the case, however. Amari *et al.* [1] proposed a modification of the algorithm so that the hessian becomes positive definite. Let us first examine the one-dimensional case for which

$$\frac{\partial U_n(B)}{\partial B} = -B^{-T} - \frac{1}{n}\sum_{i=1}^n \phi(BX_i)X_i.$$

and

$$\frac{\partial^2 U_n(B)}{\partial B^2} = \frac{1}{B^2} - \frac{1}{n}\sum_{i=1}^n B\phi'(BX_i)X_i^2.$$

For such stability at point B_σ such that $\frac{1}{B} + \mathbb{E}[\phi(BX)X] = 0$, we need that $\frac{1}{B^2} - [B\phi'(BX_i)X_i^2] \geq 0$.

ESANN'2001 proceedings - European Symposium on Artificial Neural Networks
Bruges (Belgium), 25-27 April 2001, D-Facto public., ISBN 2-930307-01-3, pp. 39-44

3.2 Hessian matrix form

Noting that $\frac{\partial b_{k\ell}^{-T}}{\partial b_{ij}} = -b_{jk}^{-1} b_{\ell i}^{-1}$, we can compute the Hessian as follows:

$$\begin{aligned} H_{ijk\ell}(B) &= \frac{\partial^2 U_n(B)}{\partial b_{ij} \partial b_{k\ell}} = -\frac{\partial}{\partial b_{ij}} \left[b_{k\ell}^{-1} + \frac{1}{n} \sum_{r=1}^{n} \phi_k(BX^r) X_\ell^r \right] \\ &= b_{\ell i}^{-1} b_{jk}^{-1} - \frac{1}{n} \sum_{r=1}^{n} \phi'(BX^r) X_\ell^r X_j^r \delta_{ik}. \end{aligned}$$

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

$$\mathbb{E}[H_{ijk\ell}(B)] = (A\Lambda^{-1})_{\ell i} (A\Lambda^{-1})_{jk} - \mathbb{E}[\phi_k'(\Lambda S) X_\ell X_j] \delta_{ik}$$

Assuming Λ is a diagonal matrix (Λ is solution of $I_d + \mathbb{E}[\phi(\Lambda S)(\Lambda S)^T] = 0$):

$$\mathbb{E}[H_{ijk\ell}(B)] = a_{\ell i} a_{jk} \frac{1}{\lambda_i \lambda_k} - \sum_p a_{\ell p} a_{jp} \mathbb{E}[\phi_k'(\lambda_k s_k) s_p^2] \delta_{ik}$$

Because $p \neq q$ implies $\mathbb{E}[\phi_k'(\lambda_k s_k) s_p s_q] = 0$. We note that if $Q(B) = \sum_{ijk\ell} b_{ij}$ $b_{k\ell} H_{ijk\ell}$ is a positive definite quadratic form, so $W \to Q(WA^{-1})$, which can also be written $Q(WA^{-1}) = \sum_{ijk\ell} \sum_{pq} w_{ip} w_{kq} a_{pj}^{-1} a_{q\ell}^{-1} H_{ijk\ell} = \sum_{ipkq} w_{ip} w_{kq} U_{ipkq}$, with $U_{ipkq} = \sum_{j\ell} a_{pj}^{-1} a_{q\ell}^{-1} H_{ijk\ell}$.
So it is equivalent to prove that

$$\begin{aligned} \sum_{k,\ell} a_{uj}^{-1} a_{v\ell}^{-1} \mathbb{E}[H_{ijk\ell}(\Lambda A^{-1})] &= \sum_{j,\ell} a_{uj}^{-1} a_{v\ell}^{-1} a_{\ell i} a_{jk} \frac{1}{\lambda_i \lambda_k} - \\ &\quad - \sum_{j,\ell} a_{uj}^{-1} a_{v\ell}^{-1} a_{\ell p} a_{jp} \mathbb{E}[\phi_k'(\lambda_k s_k) s_p^2] \delta_{ik}, \\ U_{ijk\ell} &= \delta_{jk} \delta_{i\ell} \frac{1}{\lambda_i \lambda_j} - \mathbb{E}[\phi_k'(\lambda_k s_k) s_j^2] \delta_{j\ell} \delta_{ik} \end{aligned}$$

is positive definite. If we rewrite the matrices M_{ij} as the vector $\theta = [M_{12}, M_{21}, \ldots, M_{ij}, M_{ji}, \ldots, M_{dd}]^T$, hence the transformed hessian $U_{ijk\ell}$ has a matrix form as $U_{ijk\ell} = \delta_{jk} \delta_{i\ell} \frac{1}{\lambda_i \lambda_k} - \mathbb{E}[\phi_i'(\lambda_i s_i) s_j^2] \delta_{j\ell} \delta_{ik}$

$$\mathcal{U} = \begin{bmatrix} \ddots & & & & & \\ & \mathcal{U}_{ijij} & \mathcal{U}_{jiij} & & 0 & \\ & \mathcal{U}_{ijji} & \mathcal{U}_{jiji} & & & \\ & & & \ddots & & 0 \\ & 0 & & & \mathcal{U}_{iiii} & \\ & & & 0 & & \ddots \end{bmatrix}$$

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

$$\begin{aligned} U_{ij} &= \begin{pmatrix} \kappa_{ij} & \alpha_{ij} \\ \alpha_{ij} & \kappa_{ij} \end{pmatrix} \\ U_i &= \mathcal{U}_{iiii} = \alpha_{ij} + \kappa_{ij}. \end{aligned}$$

if $\kappa_{ij} = -\mathbb{E}[\phi_i'(\lambda_i s_i) s_j^2]$, $\alpha_{ij} = \frac{1}{\lambda_i \lambda_j}$. The simplified solution where $\lambda_i = 1$ is

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

ESANN'2001 proceedings - European Symposium on Artificial Neural Networks
Bruges (Belgium), 25-27 April 2001, D-Facto public., ISBN 2-930307-01-3, pp. 39-44

From the two equations (3.2) and (3.2), we can now give the stability conditions, that is $U_i < 0$ and the eigenvalues of $U_{ij} < 0$. The eigenvalues of U_{ij} are the solutions of

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

i.e.

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

We need that $\mathrm{Re}(x_1) < 0$ and $\mathrm{Re}(x_2) < 0$. In our case, x_1 and x_2 are real numbers because U_{ij} is symetric. Thus, since $x_1 > x_2$,

$$\begin{array}{rcl} x_1 < 0 & \Leftrightarrow & -(\kappa_{ij} + \kappa_{ji}) > \sqrt{(\kappa_{ij} - \kappa_{ji})^2 + 4\alpha_{ij}^2} > 0 \\ & \Leftrightarrow & (\kappa_{ij} + \kappa_{ji})^2 > (\kappa_{ij} - \kappa_{ji})^2 + 4\alpha_{ij}^2 \\ & \Leftrightarrow & -\kappa_{ij}\kappa_{ji} > \alpha_{ij}^2 \\ & \text{and} & \\ & & \mathbb{E}\left[\phi_i'(\lambda_i s_i)(\lambda_j s_j)^2\right]\mathbb{E}\left[\phi_j'(\lambda_j s_j)(\lambda_i s_i)^2\right] > 1 \end{array}$$

4 Conclusion

This last formula allows us to check the stability conditions online, by estimating the values $\mathbb{E}[\phi_i'(\lambda_i s_i)]$ and $\mathbb{E}[(\lambda_j s_j)^2]$ with respectively $\frac{1}{n}\sum_{k=1}^{n} \phi_i'(\hat{B}_n X_k)$ and $\frac{1}{n}\sum_{k=1}^{n} \phi_i'(\hat{B}_n X_k)^2$.

References

[1] Amari, S., Chen, T.P., and Cichocki, A. (1997). *Stability analysis of learning algorithms for Blind Source Separation*, Neural Networks, Vol. 10, N^0 8, pp 1345–1351.

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

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

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

[5] Comon, P. Independent Component Analysis, a new concept ? *Signal Processing*, **36** , 287–314.

[6] Lee, T.W. (1998). *Independent Component Analysis and applications.* Kluwer academic publisher.