BLIND SIGNAL SEPARATION BASED ON THE DERIVATIVES OF THE OUTPUT CUMULANTS AND A CONJUGATE GRADIENT ALGORITHM

Rubén Martín-Clemente[†], *Carlos G. Puntonet*[‡], *José I. Acha*[†]

[†]Área de Teoría de la Señal y Comunicaciones, Universidad de Sevilla
E.S. Ingenieros, Avda. de los Descubrimientos s/n., 41092-Sevilla, SPAIN
ruben@cica.es, acha@viento.us.es
[‡]Departamento de Arquitectura y Tecnología de Computadores, Universidad de Granada
Facultad de Ciencias, E-18071 Granada, SPAIN
carlos@atc.ugr.es

ABSTRACT

In this paper it is proven that the estimation of the separating system can be based on the cancellation of some second partial derivatives of the output cross-cumulants. We propose a new contrast function that must be optimized on the Stiefel manifold. A conjugate gradient method is used in order to obtain a fast convergence speed.

1. INTRODUCTION

This paper deals with the separation of *instantaneous linear mixtures* of the sources, which is widely considered to be the core problem in many scenarios. It is also closely related to the Independent Component Analysis (ICA) approach, whose goal is to transform the data linearly in order to obtain transformed variables which are as statistically independent 'as possible'.

In the past ten years, many solutions to this problem have been proposed, starting from the seminal work of Jutten and Herault [10]. In recent times, independence criteria that are based on Information-Theoretic models have attracted a great deal of attention: for example, think of entropy maximization[3] and minimization of the mutual information between the outputs of the separation system[1]. We would point out that one must estimate the marginal distributions of the sources in order to use these criteria. By contrast, cumulant-based approaches are universal in the sense that they do not require any *a priori* knowledge of the probability density functions of the signals, although they may be computationally intensive. Depending on the scenario, one approach may be better than the other.

ICA [6], JADE [5] and FastICA [9] are representative cumulant-based algorithms. In order to smooth out the effect of additive noise on the estimated cumulants, both ICA and JADE solve a large amount of statistical equations [16], whereas FastICA uses a reduced set of cumulants (the kurtoses of the signals) in order to have a very low computational cost. Our work can be placed within the latter approach. We derive a small set of sufficient equations for BSS; nevertheless, in contrast to FastIca, we explore a non-deflation procedure to extract the sources.

The next Section of this paper is devoted to present the basic notation which will be used in the sequel. Section 3 proposes a new set of necessary and sufficient conditions for BSS. In Section 4, in view of the preceeding results, a new contrast function is presented. This contrast is minimized by using a conjugate gradient algorithm within the framework of the Stiefel manifold (Sections 5 and 6). Experimental results are given in Section 7. Finally, Section 8 is concerned with the conclusions.

2. PROBLEM STATEMENT AND NOTATION

Throughout the paper, the mixture model is represented by the equation:

$$\mathbf{x}(t) = \mathbf{A}\mathbf{s}(t) \tag{1}$$

where **s**(*t*) = $[s_1(t),...,s_N(t)]^T$ is a vector of *N* source signals, **A** = (a_{ij}) is an unknown $N \times N$ invertible *mixing matrix* and vector **x**(*t*) collects the observed signals, being the only data available. The aim of BSS is to determine a $N \times N$ *separating matrix* **B** = (b_{ij}) such that:

$$\mathbf{y}(t) = \mathbf{B}\mathbf{x}(t) = \mathbf{G}\mathbf{s}(t) \tag{2}$$

is an estimate of the source signals, *i.e.*, the *global matrix* **G** = (g_{ij}) has one and only one non-zero coefficient per row and column, where **G** = **A** · **B**. In this case, **G** is said to be a *'generalized permutation matrix'*. The only assumptions are that the sources are *statistically independent*, *stationary*, *unit-variance* and *zero-mean*. In addition, *at most* one source is *gaussian-distributed*.

3. DIFFERENTIATING THE OUTPUT CUMULANTS

Comon ([6], Theorem 11) showed that **G** is a *generalized permutation matrix* if the N components of $\mathbf{y}(t)$ are pairwise independent. Then, the cross-cumulants of any order between different outputs $y_i(t)$ and $y_j(t)$ vanish. Let us consider the output cross-cumulant $cum_{31}(y_i(t), y_j(t)) = cum(y_i(t), y_i(t), y_i(t), y_j(t))$. By using the multi-linearity property of the cumulants, it can be expanded as follows:

$$cum_{31}(y_i(t), y_j(t)) = \sum_{l=1}^{N} g_{il}^3 \, g_{jl} \, \kappa_l \qquad (3)$$

where κ_l is the kurtosis (fourth-order cumulant) of the *l*-th source. By setting (3) to zero for all $i \neq j$, we obtain a set of equations which are obviously satisfied when **G** is a *generalized permutation matrix*. Our first contribution is to show that these equations can be simplified by differentiation without losing any fundamental property, *i.e.*, consider the following second-order derivatives of (3):

$$\Delta_{ij} \triangleq \frac{1}{6} \frac{\partial^2 c_{31}}{\partial b_{ij}^2} = \sum_{l=1}^{N} g_{il} \, g_{jl} \, a_{jl}^2 \, \kappa_l, \quad (i \neq j) \qquad (4)$$

which can be written as

$$\Delta_{ij} = \sum_{l=1}^{N} \sum_{m=1}^{N} b_{il} b_{jm} cum(x_l, x_m, x_j, x_j), \qquad (5)$$

where we have used *cum*(x_l,x_m,x_j,x_j)= $\sum_p a_{lp} a_{mp} a_{jp}^2 \kappa_p$. The key point is that the *set of equations* Δ_{ij} = 0 for all $i \neq j$ *can provide us with a set of necessary and sufficient conditions to asssure the source separation*.

Proof. Trivial non-separating solutions, in which the separating matrix is not full-row rank, are avoided by assuming that the signals in **x**(*t*) are uncorrelated and have unit variance since, in this case, **A**, **B** and **G** must necesarily be orthogonal matrices. This assumption can be easily accomplished by means of the so-called *whitening* or *sphering* of the original mixtures[6].

Observe that (4) can be written compactly in a matrix form as:

$$\Delta_{ij} = \mathbf{g}_i^T \Gamma_j \mathbf{g}_j \qquad (6)$$

where $\mathbf{g}_k^T$ is the *k*-th row of **G** and Γ_j is the diagonal matrix whose (*l*,*l*)-entry is equal to $a_{jl}^2 \kappa_l$. If at most one source is gaussian (i.e., its kurtosis equals zero), we can also assume without loss of generality that each diagonal element of Γ_j is different (i.e. Γ_j has no repeated eigenvalues) since **A** (and thus Γ_j) can be properly adjusted by multipliying the observations **x**(*t*) by any orthogonal matrix.

Since both **A** and **B** are orthogonal matrices, it follows that $\mathbf{g}_i^T \mathbf{g}_j = \delta_{ij}$, where δ_{ij} stands for the Kronecker delta. Now, we prove our main result:

(*Necessity*). If **B** is a separating matrix, then each vector $\mathbf{g}_i$ is a different canonical vector. Therefore, it follows from (6) that Δ_{ij} = 0 for all *i*, *j* ($i \neq j$).

(*Sufficiency*). Let us assume that Δ_{ij} = 0 for all $i \neq j$. The vector $\Gamma_j \mathbf{g}_j$ can be expressed as a linear combination of the vectors $\mathbf{g}_i$ since the set { $\mathbf{g}_1$, $\mathbf{g}_2$, ... , $\mathbf{g}_n$ } forms an orthonormal basis of the space. Therefore:

$$\Gamma_j \mathbf{g}_j = \Delta_{jj} \mathbf{g}_j + \sum_{i \neq j} \Delta_{ij} \mathbf{g}_i \qquad (7)$$

where Δ_{ij} = $\mathbf{g}_i^T \Gamma_j \mathbf{g}_j$, as before, and Δ_{jj} = $\mathbf{g}_j^T \Gamma_j \mathbf{g}_j$. Since, Δ_{ij} = 0 for all $i \neq j$, we obtain that $\Gamma_j \mathbf{g}_j$ = $\Delta_{jj} \mathbf{g}_j$, which implies that each $\mathbf{g}_i$ is an eigenvector of a diagonal matrix, i.e., a canonical vector. As a consequence, **B** is a separating matrix. ♠

The proposed equations are simpler than the original ones: in contrast to (5), notice that (3) depends on as-high-as-fourth order powers of **B** through the relation **G** = **B A**. Interestingly, it can be easily shown that the equations $\Delta_{ij} = 0$ ($i \neq j$) are a subset of those equations which are solved by JADE[14]. In the two-source case, a direct solution can be easily obtained [12]. Finally, it should be emphasized that additional equations can be obtained by differentiating other cumulants, as will be shown in a future paper.

4. SEPARATION CRITERION

Instead of solving directly the equations, we propose to minimize the following cost function:

$$f(\mathbf{B}) = \sum_{i \neq j} \Delta_{ij}^2, \qquad (8)$$

where $\mathbf{x}(t)$ satisfies $E\{\mathbf{x}^T(t)\mathbf{x}(t)\} = \mathbf{I}$ (see above), being **I** the identity matrix, and **B** is constrained to the set of matrices such that

$$\mathbf{B}^T \mathbf{B} = \mathbf{I} \qquad (9)$$

i.e., **B** belongs to the group of orthogonal matrices. Generally speaking, the constraint surface (9) is known as the Stiefeld manifold[1]. It has been shown [7, 2] that the gradient of $f(\mathbf{B})$ at **B** on this manifold is given by:

$$\nabla f(\mathbf{B}) = \partial f(\mathbf{B}) - \mathbf{B} \, (\partial f(\mathbf{B}))^T \, \mathbf{B} \qquad (10)$$

where $\partial f(\mathbf{B})$ is the $N \times N$ matrix of partial derivatives of $f(\mathbf{B})$ with respect to the elements of **B**, i.e.,

$$(\partial f(\mathbf{B}))_{ij} = \frac{\partial f(\mathbf{B})}{\partial b_{ij}} \qquad (11)$$

[1] The Stiefel manifold consists of all the $n \times m$ matrices **Q** which verify that $\mathbf{Q}^T$ **Q** equals the identity matrix. If $n = m$ ($\equiv N$ in our case), we have the orthogonal group.

whose calculation is straightforward in view of (5).

5. INSIGHT INTO THE GRADIENT ON THE STIEFEL MANIFOLD

Some properties of the above-defined gradient are quite illustrative:

a. Consider small deviations of **B** in direction $-\nabla f(\mathbf{B})$, *i.e.*:

$$\mathbf{B} \to \tilde{\mathbf{B}} \equiv \mathbf{B} - \mu \, \nabla f(\mathbf{B}) \tag{12}$$

where $\mu > 0$. If matrix **B** is orthogonal, then it is straightforward to check that the translation (12) preserves the orthogonality constraint, in the sense that $\tilde{\mathbf{B}}\,\tilde{\mathbf{B}}^T = \mathbf{I} + o(\mu^2)$.

b. The first order Taylor expansion of $f(\mathbf{B})$ gives

$$f(\mathbf{B} + \Delta\mathbf{B}) = f(\mathbf{B}) + < \partial f(\mathbf{B}) | \Delta\mathbf{B} > + o(\Delta\mathbf{B}) \tag{13}$$

where $< \mathbf{M}|\mathbf{N} >$ = TRACE$[\mathbf{M}^T\mathbf{N}]$ stands for the standard inner product of matrices. If **B** is modified into $\tilde{\mathbf{B}}$, by virtue of (12) and (13) we obtain that

$$f(\tilde{\mathbf{B}}) = f(\mathbf{B}) - \mu < \partial f(\mathbf{B}) | \nabla f(\mathbf{B}) > + o(\mu) \tag{14}$$

Since

$$< \partial f(\mathbf{B}) | \nabla f(\mathbf{B}) > \equiv \frac{1}{2} < \nabla f(\mathbf{B}) | \nabla f(\mathbf{B}) > \tag{15}$$

is always *non-negative*, it follows that $f(\mathbf{B})$ is *decreased* by the translation (12). Identity (15) is readily obtained by expanding

$$< \partial f(\mathbf{B}) | \nabla f(\mathbf{B}) >$$

and using TR$[\mathbf{B}^T \partial f \, \partial f^T \, \mathbf{B}]$ = TR$[\partial f^T \, \mathbf{B}\,\mathbf{B}^T \partial f]$ = TR$[\partial f^T \, \partial f]$ and TR$[\mathbf{B}^T \partial f \mathbf{B}^T \partial f]$ = TR$[\partial f^T \mathbf{B} \partial f^T \mathbf{B}]$, where TR$[\cdot]$ denotes the TRACE operator.

c. Consequently, the gradient-step method for minimizing $f(\mathbf{B})$ is given by:

$$\mathbf{B}_{t+1} = \mathbf{B}_t - \mu \, \nabla f(\mathbf{B}_t) = \mathbf{B}_t - \mu \, \mathbf{H}(\mathbf{B}_t) \, \mathbf{B}_t \tag{16}$$

where $\mu > 0$ and $\mathbf{H}(\mathbf{B}_t) \equiv \partial f(\mathbf{B}_t)\mathbf{B}_t^T - \mathbf{B}_t \, \partial f(\mathbf{B}_t)^T$. We would point out that algorithm (16) is an EASI-type method [4] for the serial update of matrix **B**.

6. OPTIMIZATION METHOD

Different algorithms can be chosen to minimize $f(\mathbf{B})$. We have used a *conjugate gradient method* [8], which is an improved version of the abovementioned gradient-step algorithm. On the Stiefeld manifold, it is as follows (the algorithm has been adapted from[7]):

CONJUGATE GRADIENT METHOD ON THE STIEFEL MANIFOLD

```
1 Let $\mathbf{B}_0$ be the initial point and set $\mathbf{G}_0 = \nabla f(\mathbf{B}_0)$ and $\mathbf{F}_0 = -\mathbf{G}_0$.
2 For $t = 0, 1, \ldots$
    2.1 Let $\mathbf{B}_{t+1} = \mathbf{B}_t \exp(\mu \, \mathbf{B}_t^T \mathbf{F}_t)$, where parameter $\mu > 0$. When $\mu$ is small enough, observe that it can be replaced with $\mathbf{B}_{t+1} \simeq \mathbf{B}_t + \mu \mathbf{F}_t$.
    2.2 Compute $\mathbf{G}_{t+1} = \nabla f(\mathbf{B}_{t+1})$
    2.3 Compute the new search direction
            $\mathbf{F}_{t+1} = -\mathbf{G}_{t+1} + \gamma_t \mathbf{F}_t \exp(\mu \, \mathbf{B}_t^T \mathbf{F}_t)$,
        where $\gamma_t = \frac{<\mathbf{G}_{t+1}|\mathbf{G}_{t+1}>}{<\mathbf{G}_t|\mathbf{G}_t>}$ (which gives a Fletcher-Reeves conjugate gradient formulation [7]).
    2.4 Reset $\mathbf{F}_{t+1} = -\mathbf{G}_{t+1}$ (the steepest descent direction) if $t+1 \equiv 0 \; mod \; N\frac{N-1}{2}$.
```

It should be emphasized that both the above algorithm and the gradient-step method have a similar complexity, in terms of their practical implementations. Indeed, the dominant task is, by far, the computation of the statistics of the observations. In [13], we have successfully used a *simulated annealing* algorithm in order to obtain initial conditions for the conjugate gradient algorithm.

7. COMPUTER SIMULATIONS

Computer simulations have been carried out in order to corroborate the validity of the proposed procedure. Figure 1 shows the time-course of the mean signal to noise ratio of the estimated sources, averaged over ten independent experiments, for a mixture of six uniform sources. The cumulants of the observed signals were estimated over 5000 samples. Mixing matrices were randomly chosen. The parameter μ of the algorithm was set to one. The experiment reveals a fast convergence speed.

After a few (~ 50) iterations, the complete matrix which relates the sources and the outputs was typically as

$$\begin{bmatrix} -0.13 & -0.01 & -0.01 & 0.02 & 0.01 & \boxed{0.98} \\ 0.01 & 0.05 & -0.00 & -0.04 & \boxed{0.99} & -0.02 \\ \boxed{0.98} & -0.05 & 0.01 & 0.00 & 0.00 & 0.14 \\ 0.06 & \boxed{0.99} & 0.09 & 0.00 & -0.06 & 0.01 \\ 0.00 & -0.01 & 0.02 & \boxed{-0.99} & -0.03 & 0.03 \\ -0.00 & 0.07 & \boxed{-0.99} & -0.03 & -0.00 & -0.00 \end{bmatrix}$$

which shows the successful separation.

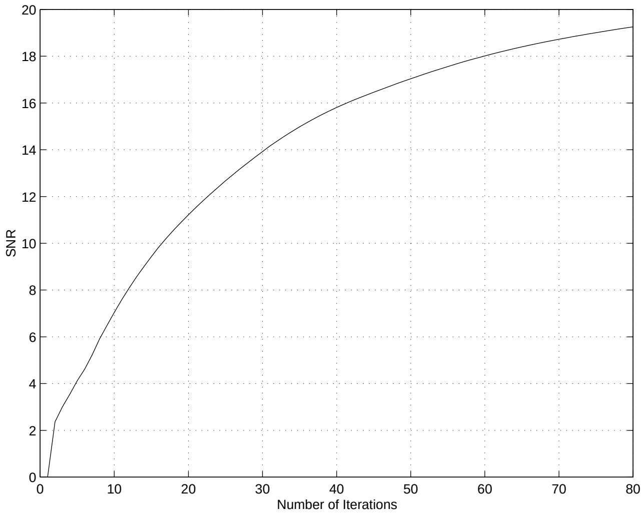

Fig. 1. Mean Signal to Noise Ratio. Mixture of six uniform sources.

8. SUMMARY AND CONCLUSIONS

We have presented quadratic equations for BSS which satisfy: *i.)* under a mild condition, they do not have spurious roots and *ii.)* Δ_{ij} is a quadratic function of the coefficients of *B* whereas the cumulant $cum_{31}(y_i(t), y_j(t))$ depends on fourth-order powers of these coefficients. To simplify the cumulant-based equations by means of differentiation seems to be a general procedure. We are currently studying its extension to the convolutive-mixture problem.

Imposing the orthogonality constraints in (9) is problematic in practice. Alternating projection methods have been widely used for this purpose[15]. Nevertheless, it is difficult to prove their convergence[11]. On the other hand, we use a conjugate gradient method on the Stiefel manifold which naturally preserves the constraint.

9. REFERENCES

[1] S.Amari, A. Cichocki and H.H. Yang ”A new Learning Algorithm for Blind Source Separation”, in *Advances in Neural Information Processing 8*, Cambridge, MA:MIT Press, 757-763, 1996.

[2] S.Amari ”Natural gradient learning for over- and under-complete bases in ICA”, in *Neural Computation*, vol. 11, 1875-1883, 1999.

[3] A.J.Bell and T. Sejnowski ”An Information-Maximization Approach to blind separation and blind deconvolution”, *Neural Computation*, vol. 7, pp.1129-1159, 1995.

[4] J.-F. Cardoso and B. Laheld ”Equivariant Adaptive Source Separation”, *IEEE Transactions on Signal Processing*, vol. 45, No. 2, pp.433-444, 1996.

[5] J.-F. Cardoso and A. Souloumiac ”Blind Beamforming for Non Gaussian Signals”, *IEE Proc.-F*, vol. 140, No. 6, pp.362-270, Dec. 1993.

[6] P. Comon, ”Independent Component Analysis - A New Concept?”, *Signal Processing*, vol. 36, No. 3, pp.287-314, 1994.

[7] A. Edelman, T. A. Arias and S. T. Smith ”The Geometry of Algorithms with Orthogonality Constraints”, *SIAM Journal on Matrix Analysis Applications*, vol. 20, No. 2, pp.303-353, 1998.

[8] P. E. Gill, W. Murray and M. H. Wright, ”Practical Optimization”, *Academic Press*, 1981.

[9] A. Hyvarinen and E.Oja ”A fast fixed-point algorithm for independent component analysis”, *Neural Computation*, vol. 9, pp.1483-1492, 1997.

[10] C. Jutten and J. Herault, ”Blind Separation of Sources, Part I: an adaptive algorithm based on neuromimetic architecture”, *Signal Processing*, vol. 24, pp.1-10, 1991.

[11] R. J. Marks II, ”Alternating projections onto convex sets”, in *Deconvolution of Images and Spectra*, P. A. Jansson, Ed., Academic, 1997.

[12] R. Martín-Clemente and J.I.Acha, ”Blind Separation of Sources using a New Polynomial Equation”, *Electronics Letters*, Vol.33, No.1, pp.176-177,1997.

[13] R. Martín-Clemente, C. G. Puntonet and J.I.Acha, ”A Conjugate Gradient Method and Simulated Annealing for Blind Separation of Sources”, in *6th International Work-Conference on Artificial and Natural Neural Networks, IWANN2001*, Granada, Spain, pp.810-816 (Part II), June 2001.

[14] R. Martín-Clemente, J.I.Acha and C. G. Puntonet, ”Blind Separation of Sources by Differentiating the Output Cumulants and using Newton’s Method”, in *International Conference on Artificial Neural Networks, ICANN2001*, Vienna, Austria, August 2001, to appear.

[15] E. Oja ”Nonlinear PCA Criterion and Maximum Likelihood in Independent Component Analysis”, in *First Intern. Workshop on Independent Component Analysis and Signal Separation*, Aussois, France, pp. 143-148, Jan. 1999.

[16] J.Zhu, X.-R.Cao and Z.Ding, ”An algebraic principle for blind separation of white non-Gaussian sources”, *Signal Processing*, vol. 76, No. 2, pp.105-116, 1999.