PRIOR INFORMATION ABOUT MIXING MATRIX IN BSS-ICA FORMULATION

Jorge Igual, Luis Vergara

Departamento Comunicaciones
Universidad Politécnica Valencia
Camino de Vera, S/N, Valencia 46023, Spain
e-mail: jigual@dcom.upv.es; lvergara@dcom.upv.es

ABSTRACT

In BSS-ICA few assumptions are considered. With respect to the mixing matrix usually nothing is supposed. We present in this paper an extension of the BSS-ICA problem: when we have a prior statistical information about the elements of the mixing matrix. We will model this information with a matrix of probability density functions (pdf) and we will analyze how traditional and new solutions to BSS-ICA problem can be adjusted to the new problem.

1. INTRODUCTION

The problem of Blind Source Separation (BSS) consists on recovering a set of independent source signals from linear mixtures of them [1]. The matrix that relates the source signals with the mixed observed signals is called the mixing matrix and nothing is supposed about it, excepting it is non-singular (for a square mixing matrix). However, in some applications, we have a prior information about some of its elements, not about its structure as it is usual in array signal processing. Obviously, if this information is disposable, it would be interesting to include it in the statement of the problem. In fact, although the definition of ICA does not assume any condition about mixing matrix or sources, it is usual that with the aim of overcome the indeterminations inherent to it, supposing that sources are unit-variance or columns of mixing matrix are unit-norm (we will suppose in this paper the first hypothesis, normalized sources).

The new problem is called AKICA (A priori Knowledge ICA) and must be considered as an extension of ICA, where no new assumptions are made about sources or mixing matrix, but a new statistical information about the elements of the mixing matrix is included, so the concept of blindness is partially lost and a bayessian approach to the problem is possible.

The way this information is introduced is explained in Section 2. In Section 3 we analyze the AKICA problem and how prior information about **A** is transformed during the AKICA solution. As an example, we focus on the infomax version of AKICA, and simulate in Section 4 the algorithm in the one-dimensional case. Finally some conclusions and extensions are presented in Section 5.

2. AKICA PROBLEM STATEMENT

For the *2x2* real noiseless instantaneous mixture, the observations **y** *2x1* and unknown zero-mean unit-variance statistically independent sources **x** *2x1* are linearly related by the mixing matrix **A**:

$$\mathbf{y}=\mathbf{A}\mathbf{x} \tag{1}$$

The ICA of vector **y** tries to recover the sources up to a permutation and a sign using the statistical independence of the unit-variance sources:

$$\mathbf{s}=\mathbf{F}\mathbf{y} \tag{2}$$

where **s** is the vector of recovered sources *2x1* and **F** the *2x2* recovering matrix. If **F** is the solution, then **FA**=**PD**, where **P** is a *2x2* permutation matrix and **D** a sign diagonal matrix.

The aim of AKICA is the same: to recover independent sources or to obtain mixing matrix. For this goal, we add a new information to the statement of the problem. This prior information is referred to the elements of the mixing matrix and is mathematically formulated with a matrix of pdf, $p(\mathbf{A})$, where element (i,j), $i,j=1,2$, is $p(a_{ij})$, the pdf of (i,j) element of the mixing matrix.

First of all, we must note that ICA problem can be considered as a particular case of AKICA, when $p(\mathbf{A})$ is a matrix of uniform random variables, so the prior information is not meaningful. On the other side, when an element of **A** is known, it can be considered as an unknown with a delta-like pdf centered in the correct value.

Secondly, the indetermination inherent to ICA disappears in AKICA. The introduction of a prior pdf brings about that one of the possible solutions is more probable than the others, so the sources are recovered in the correct order and sign, i.e. **FA**=**I**.

However, a new problem may appear: if our prior information is incorrect, estimated matrix **F** may be very far from the solution, so **FA**≠**I**, and the recovered signals are statistically dependent. In this case the solution will be a commitment between the prior information and the degree of statistical independence of the recovered signals.

3. AKICA SOLUTION

ICA solutions can be classified in adaptive and batch type algorithms. Most of them obtain the solution in two steps. Firstly, statistics of second order are used to decorrelate and normalize the observations. Secondly, higher order statistics approximate the statistical independence, normally up to fourth order. This is because the mixing matrix can be factored $\mathbf{A}=\mathbf{L}^{-1}\mathbf{Q}$, where $\mathbf{L}$ is the whitening matrix and $\mathbf{Q}$ a Givens rotation matrix,

$$\mathbf{Q}=\begin{bmatrix}\cos\alpha & \operatorname{sen}\alpha\\ -\operatorname{sen}\alpha & \cos\alpha\end{bmatrix}=\frac{1}{\sqrt{1+\theta^2}}\begin{bmatrix}1 & \theta\\ -\theta & 1\end{bmatrix}=\mathbf{LA} \qquad (3)$$

so $\mathbf{u}=\mathbf{L}\mathbf{y}$, $E\{\mathbf{u}\mathbf{u}^T\}=\mathbf{I}$, and the recovered sources $\mathbf{s}=\mathbf{Q}^T\mathbf{u}$.

We will employ a two-step algorithm to solve the AKICA problem. In second step, we have to estimate only the rotation angle that decorrelated sources must be rotated in order to obtain independent sources. If first step is carried out, our prior knowledge is transformed in a prior information about rotation angle.

$$p(q_{ij})=\int_{-\infty}^{\infty} p(a_{1j},\frac{q_{ij}-l_{i1}a_{1j}}{l_{i2}})da_{1j}/|l_{i2}| \qquad (4)$$

where q_{ij}, l_{ij}, a_{ij} are the elements (i,j) of $\mathbf{Q}$, $\mathbf{L}$ and $\mathbf{A}$, respectively. If the elements of a column of $\mathbf{A}$ are statistically independent, (4) can be expressed as the convolution of r.v. $b_{1j}=l_{i1}a_{1j}, b_{2j}=l_{i2}a_{2j}$

$$p_{q_{ij}}(q_{ij})=\int_{-\infty}^{\infty} p_{a_{1j}}(a_{1j})\cdot p_{a_{2j}}(\frac{q_{ij}-l_{i1}a_{1j}}{l_{i2}})da_{1j}/|l_{i2}| \qquad (5)$$

As we can see in (5) the prior information is transformed in a prior knowledge about matrix $\mathbf{Q}$ that depends on the whitening matrix. This first step is usually called PCA and can be achieved by different ways. We review now some of these methods [2] and how they affect to (5).

- Whitening based on eigenvalue decomposition. Let $\mathbf{R}$ be the covariance matrix of the observed mixed signals $\mathbf{R}=E\{\mathbf{y}\mathbf{y}^T\}$ (zero-mean signals); $\mathbf{R}$ can be factored as $\mathbf{E}^T\mathbf{R}\mathbf{E}=\Lambda^2$, where the columns of $\mathbf{E}$ are the eigenvectors of $\mathbf{R}$ and the elements of the diagonal matrix Λ^2 are the eigenvalues. In this case, the whitening matrix is $\mathbf{L}=\Lambda^{-1}\mathbf{E}^T$.
- Whitening based on singular value decomposition. Let $\mathbf{y}=\mathbf{U}\Delta\mathbf{V}^T$ be the SVD of matrix $\mathbf{y}$ *2xT*, being T the number of observations; then $\mathbf{L}=\sqrt{T}\Delta^{-1}\mathbf{U}^T$.
- Whitening based on triangular lower-upper decomposition. Let $\mathbf{R}=\mathbf{G}\mathbf{D}^2\mathbf{G}^T$ be the lower-upper decomposition of $\mathbf{R}$, where $\mathbf{G}$ is a unit lower triangular matrix and $\mathbf{D}$ is diagonal; then $\mathbf{L}=\mathbf{D}^{-1}\mathbf{G}^{-1}$.
- Whitening based on triangular QR factorization. A possible QR factorization is defined as $\mathbf{y}^T=\mathbf{Q}_1\mathbf{R}_1$ where $\mathbf{Q}_1$ is a matrix *Tx2* whose columns are orthonormal, $\mathbf{R}_1$ a upper triangular matrix *2x2* ("economy size" decomposition). In this case $\mathbf{L}=\sqrt{T}\mathbf{R}_1^{-T}$.

SVD and QR work directly on the data matrix, so numerically they are better than the others, which need to estimate the covariance matrix $\mathbf{R}$. Eigenvalue and SVD decomposition, and triangular upper-lower and QR factorizations are easily related [2].

If we employ a triangular decomposition method in the whitening step, (5) is simplified:

$$\mathbf{Q}=\mathbf{L}\mathbf{A}=\begin{bmatrix}l_{11} & 0\\ l_{21} & l_{22}\end{bmatrix}\mathbf{A} \qquad (6)$$

so $l_{11}a_{11}=q_{11}$ and (5) reduces to:

$$p_{q_{11}}(q_{11})=\frac{p_{a_{11}}(q_{11}/l_{11})}{|l_{11}|} \qquad (7)$$

From (7) we deduce that the pdf of q_{11} is a scaled version of pdf of a_{11}, so, depending on the modulus of l_{11}, the whitening step will have transformed the prior information about a_{11} in a sharper or flatter information about Givens matrix $\mathbf{Q}$.

The disadvantage of triangular decomposition is clear when the prior information of $\mathbf{A}$ is in the element that corresponds to the null in the matrix $\mathbf{L}$. In this case (7) is not useful and we have to apply (5).

Finally, it is usual in the *2x2* case to parameterize the matrix $\mathbf{Q}$ as in (3). This fact implies that our prior information is about the rotation angle, and can be expressed as $p_\alpha(\alpha)=p_{q_{11}}(\cos\alpha)\cdot|\operatorname{sen}\alpha|$.

The second step of AKICA solution obtains the angle α. In [3] some classical BSS solutions are revisited and reformulated to adapt them to the new problem. The idea is that we can consider all of them as algorithms that try to minimize/maximize an objective function, so this function can be modified to include the prior information.

We present here the analysis of one of them, the infomax solution, an information theoretic approach.

3.1 Infomax solution

The infomax [4] solution maximizes the joint entropy $H[g(\mathbf{s})]$ of the recovered signals transformed by a sigmoidal function g. The learning algorithm is:

$$\Delta \mathbf{F} \alpha \ \frac{\partial H(\mathbf{r})}{\partial \mathbf{F}} = E\left[\frac{\partial \ln|J|}{\partial \mathbf{F}}\right] \tag{8}$$

where $\mathbf{r} = [g(s_1), g(s_2)]^T$ and $|J|$ the absolute value of the jacobian of the transformation $\mathbf{y} \rightarrow \mathbf{r}$.

$$J = \det\begin{bmatrix} \frac{\partial r_1}{\partial y_1} & \frac{\partial r_1}{\partial y_2} \\ \frac{\partial r_2}{\partial y_1} & \frac{\partial r_2}{\partial y_2} \end{bmatrix} \tag{9}$$

The stochastic version is:

$$\Delta \mathbf{F} \alpha \ \mathbf{F}^{-T} + \hat{\mathbf{r}} \mathbf{y}^T \tag{10}$$

with $\hat{\mathbf{r}} = [\hat{r}_1, \hat{r}_2]^T$, $\hat{r}_i = \frac{\partial}{\partial r_i}\frac{\partial r_i}{\partial s_i}$

For example:

$$r_i = \frac{1}{1+e^{-s_i}}, \ \hat{r}_i = 1 - 2r_i \tag{11.1}$$

or

$$r_i = tanh(s_i), \ \hat{r}_i = -2r_i \tag{11.2}$$

If we use the natural or relative gradient [5], the learning rule is:

$$\Delta \mathbf{F} \alpha \frac{\partial H(\mathbf{r})}{\partial \mathbf{F}} \mathbf{F}^T \mathbf{F} = \left(\mathbf{I} + \hat{\mathbf{r}} \mathbf{s}^T\right) \mathbf{F} \tag{12}$$

which avoids to invert the matrix and provides uniform performance.

As **FA**=**I**, our prior knowledge is transformed in a prior information about recovering matrix:

$$p(\mathbf{F}) = p(\mathbf{A}^{-1}) = \begin{bmatrix} p(a_{22} / \det \mathbf{A}) & p(-a_{12} / \det \mathbf{A}) \\ p(-a_{21} / \det \mathbf{A}) & p(a_{11} / \det \mathbf{A}) \end{bmatrix} \tag{13}$$

The new function to be maximized is $\ln|J| \cdot p(\mathbf{F})$. Using the stochastic gradient ascent technique and (11.1), we obtain the new learning rule:

$$\Delta F_{ij} \ \alpha \left(\frac{cof(F_{ij})}{\det \mathbf{F}} + y_j(1-2r_i)\right) \cdot p(F_{ij}) + \\ + \ln\left|\det(\mathbf{F}) \prod_k r_k(1-r_k)\right| \cdot p'(F_{ij}) \tag{14}$$

where $cof(F_{ij})$ is $(-1)^{i+j}$ times the determinant of matrix obtained eliminating row *i* and column *j* of **F**.

The problem is that we know something about some elements of **A**, but not about its determinant. However, we know that $\det \mathbf{F} = 1/\det \mathbf{A}$, so we can obtain an estimate of it from the separating matrix **F** calculated imposing only statistical independence (10) (or (12) if we use natural gradient).

Nevertheless, as we explain in previous section, most of BSS algorithms improve if we first decorrelate the observed signals. If this step is carried out, we have only to estimate the rotation matrix **Q** and our prior knowledge is modeled by $p(\mathbf{Q}^T)$, which can be obtained from (4). The final adaptation rule is (substituting $\mathbf{F} = \mathbf{Q}^T$ and $\mathbf{F}^{-T} = \mathbf{Q}^T$ in (10))

$$\Delta \mathbf{Q}^T \alpha \left(\mathbf{Q}^T + \hat{\mathbf{r}} \mathbf{u}^T\right) \cdot p(\mathbf{Q}^T) + \ln|J| \cdot p'(\mathbf{Q}^T) \tag{15}$$

4. SIMULATION

In order to simplify, we will apply our theoretical formulation to the one-dimensional case. Matrix **F** is reduced to a scalar *f* and mixed and recovered signals are *s* and *y*, respectively. We want to obtain the value of *f* that maximizes the entropy of a sigmoidally transformed version of $s = f \cdot y$.

In this case, infomax algorithm (10) is (we include the time index):

$$f[n+1] = f[n] + \lambda[n]\left(\frac{1}{f[n]} + y[n]\hat{r}[n]\right) \tag{16}$$

where λ is the adaptation step, that can be adaptable. The new learning rule is:

$$f[n+1] = f[n] + \lambda[n]\left\{\left(\frac{1}{f[n]} + y[n]\hat{r}_i[n]\right) \cdot p(f[n]) + \right. \\ \left. + \log|f[n] \cdot r_i[n] \cdot (1 - r_i[n])| \cdot \frac{dp(f[n])}{df}\right\} \tag{17}$$

The solution is the *f* that transforms the r.v. *y* in a uniform r.v. r_i (11.1 or 11.2) (maximum entropy).

In our simulation, we suppose that *y* is a Gaussian r.v. and *T*=2000 observations are available. Our prior information is modeled by a Gaussian $p(f)$ with variable mean and unit-variance.

In figure 1, we show the convergence of *f* for different values of the mean of *f*. In solid line, the traditional infomax or Bell solution; in dotted line the solution for the new algorithm for a *f* with mean 1.6 and in dashed line the new algorithm for a *f* with mean 0.75. As we see, when our prior knowledge is meaningful (mean 1.6), the solution is achieved faster than the infomax algorithm with no prior information. However, when the mean of *f* is far away from the value that maximizes the entropy of *r* (mean 0.75), the value of the estimated *f* moves between the maximum entropy solution and maximum prior information *f*=1.3.

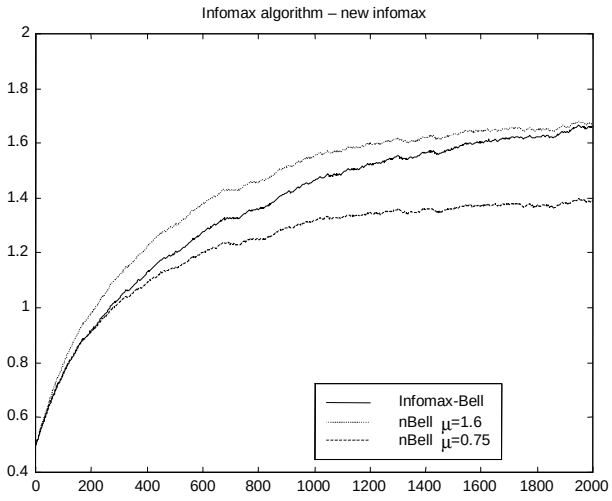

Figure 1. Estimated f vs number of observations, for different $p(f)$. Solid line, infomax or Bell solution. Dotted and dashed line, new infomax estimator (called "nBell"), for a Gaussian $p(f)$ with different mean and unit-variance.

In figure 2 we show the histogram of y, the histogram of the signal r for the infomax or correct mean new algorithm estimator (same histogram in the convergence) and the histogram for the estimator obtained with the new algorithm and mean 0.75 As we can see, the histogram for $f = 1.6$ corresponds to a uniform r.v., as we were expecting.

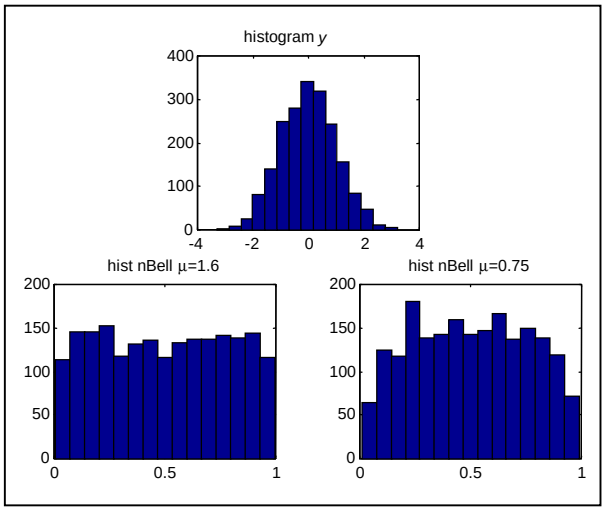

Figure 2. Up, histogram of y. Down-left histogram of r for infomax or new infomax solution with mean 1.6. Down-right histogram of r for new infomax solution with mean 0.75.

5. CONCLUSIONS

We have shown in this paper a theoretical extension of BSS-ICA problem, including in the formulation a prior information about the mixing matrix. We have explained how this information can be mathematically expressed by a matrix of pdfs.

Considering a two-step analysis of the new problem (called AKICA), we have studied how the whitening step modifies our prior knowledge, and how it is transferred to the second step (rotation imposing statistical independence).

We have focused on the infomax solution as an example of traditional solution and how its adaptation rule can be easily modified. Other solutions based on classical BSS methods are explained in [3] and a new statistical approach for the second step can be found in [6].

6. REFERENCES

[1] P. Comon, "Independent component analysis, a new concep?", *Signal Processing*, Vol. 36, No. 3, April 1994, pp 287-314.

[2] C. Therrien, *Discrete random signals and statistical signal processing*, Englewood Cliffs, NJ. Prentice-Hall, 1992.

[3] J. Igual, L. Vergara, Modified BSS algorithms including prior statistical information about mixing matrix, *submitted to SSAP 2000.*

[4] A.J. Bell, T. J. Sejnowsky., "An information maximization approach to blind separation and blind deconvolution", *Neural computation*, 7, pp 1129-1159.

[5] S. Amari, "*Natural gradient works efficiently in learning*", Neural computation, 1996.

[6] J. Igual, L. Vergara, "A MAP solution to BSS", *IEEE Signal Processing Workshop on HOS*, June 1999, Israel, pp 112-115.