Beyond Comon’s Identifiability Theorem for Independent Component Analysis

Riccardo Boscolo\(^1\), Hong Pan\(^2\), and Vwani P. Roychowdhury\(^1\)

\(^1\) Electrical Engineering Department, University of California, Los Angeles
Los Angeles, CA 90095
\{riccardo, vwani\}@ee.ucla.edu

\(^2\) Functional Neuroimaging Laboratory, Department of Psychiatry, Weill Medical College, Cornell University, New York, NY 10021
hop2001@med.cornell.edu

Abstract. In this paper, Comon’s conventional identifiability theorem for Independent Component Analysis (ICA) is extended to the case of mixtures where several gaussian sources are present. We show, in an original and constructive proof, that using the conventional mutual information minimization framework, the separation of all the non-gaussian sources is always achievable (up to scaling factors and permutations). In particular, we prove that a suitably designed optimization framework is capable of seamlessly handling both the case of one single gaussian source being present in the mixture (separation of all sources achievable), as well as the case of multiple gaussian signals being mixed together with non-gaussian signals (only the non-gaussian sources can be extracted).

1 Introduction

In his fundamental work [1], Comon showed that the separation of a set of stationary signals, instantaneously and linearly mixed, is always possible, as long as the mixing matrix has full rank, and at the most one of the original signals is gaussian distributed. This result is often cited in the literature as Comon’s identifiability theorem for ICA, and it represents a well-known and widely mentioned result in the blind signal separation field. Although the theorem holds strictly only when a functional of the probability density function of the reconstructed signals is used as contrast function, most ICA algorithms are based on contrast functions of this type, such as the mutual information between the reconstructed signals, or its equivalent counterparts, i.e. the InfoMax principle, or the maximum likelihood (ML) principle.

In recent years, Cruces et al. [2][3][4] investigated several criteria for the extraction of a subset of sources from a linear mixture, both in the instantaneous case, and in the case of convolutive mixtures. In particular, it was shown in [3],

© Springer-Verlag Berlin Heidelberg 2002
that a suitably designed entropy minimizing framework can be used to extract the non-gaussian sources, from mixtures containing an arbitrary number of gaussian distributed signals. The authors also introduced a moment-based iterative algorithm that minimizes an approximation of the contrast function derived from this principle.

In this paper, we derive a novel proof of Comon’s identifiability theorem, and extend the theorem to the case of multiple gaussian sources being mixed with non-gaussian sources. All the results are derived from investigating the properties of the optimization problem associated with minimizing the mutual information between the reconstructed signals. In particular, we prove that, regardless of the number of gaussian sources in the mixture, the resulting objective function always has extrema that yield the separation of the non-gaussian sources (up to scaling and permutations), and the gaussian components are irrelevant in determining such extrema.

## 2 Separation Principle and Objective Function Definition

We make the conventional assumption that \( N \) independent and stationary source signals \( (s_1, \ldots, s_N) \) are mixed by an unknown, full-rank mixing matrix \( A \), resulting in a set of mixtures given by \( x = As \). The reconstruction of the original sources is attempted from the mixture data through a linear projection of the type \( y = Bx \). Following the mutual information minimization principle, common to most ICA frameworks, we seek the matrix \( B \), solution of the optimization problem \(^3\):

\[
B_{\text{opt}} = \arg\min_B I(y_1, \ldots, y_N) \tag{1}
\]

Using basic information theory equalities, (1) becomes:

\[
\min_B \sum_{i=1}^{N} H(y_i) - \log |\det B| - H(x), \tag{2}
\]

where \( H(a) = -\int p_a(u) \log p_u(u) du \). If we assume that the mixture data has been sphered, i.e. \( \text{Cov}(xx^T) = I \), we can restrict the search space for the unmixing matrix \( B \) to the manifold of orthogonal matrices [5]. The problem can be simplified as:

\[
\min_B \sum_{i=1}^{N} H(y_i) \tag{3}
\]

\[\text{s.t.} \quad BB^T = I, \tag{4}\]

since \( \log |\det(B)| \equiv 1 \), and \( H(x) \) is a constant with respect to \( B \). The equality constraints (4) define a sub-group of the Stiefel manifold for the case of square

\[^3I(y_1, \ldots, y_N) \triangleq \int p_y(y) \log \prod_{i=1}^{N} \frac{p_{y_i}(y_i)}{p_y(y)} dy\]
Beyond Comon’s Identifiability Theorem for Independent Component Analysis

matrices. If we define \( F(B) \triangleq \sum_{i=1}^{N} H(y_i) \), then the gradient of the cost function defined on such manifold is given by [6]:

\[
\nabla_m F(B) \triangleq \nabla F(B) - B \nabla F(B)^T B.
\]

(5)

where \( \nabla F(B) \) is the conventional gradient of \( F(B) \) in the Euclidean space:

\[
\nabla F(B) \triangleq \left[ \frac{\partial F(B)}{\partial b_{ij}} \right] = \begin{bmatrix} \nabla H(y_1) \\ \vdots \\ \nabla H(y_N) \end{bmatrix}.
\]

(6)

The extrema of the optimization problem (3) are given by all the matrices that satisfy the condition:

\[
\nabla_m F(B) = 0 \quad \Rightarrow \quad \nabla F(B)B^T = B \nabla F(B)^T,
\]

(7)

since \( BB^T = I \).

3 Extending Comon’s Identifiability Theorem

In this section, an alternative proof of Comon’s well-known theorem on ICA identifiability [1] is derived, and it is extended to the case where more than one gaussian source is present in the mixture. Under the modeling assumption of Section 2, we consider mixtures of \( N \) independent sources \( s_1, \ldots, s_N \), with probability density function \( f_{s_1}, \ldots, f_{s_N} \), \( M \) of which are gaussian distributed.

We make the further assumption that the mixing matrix \( A \) is the \( N \times N \) identity matrix. This is not a restrictive assumption, since, if the mixture data is spherered, the solution spaces associated to any two full rank mixing matrices simply map to each other through an orthogonal transformation [7]. The generic reconstructed signal can be written as:

\[
y_i = b_{i1}s_1 + b_{i2}s_2 + \ldots + b_{iN}s_N \quad i = 1, \ldots, N,
\]

(8)

and its differential entropy is given by:

\[
H(y_i) = -\int_{-\infty}^{\infty} f_{y_i}(u) \log f_{y_i}(u) du,
\]

(9)

where, because of the independence between the sources:

\[
f_{y_i}(u) = \frac{1}{|b_{i1}|} f_{s_1} \left( \frac{u}{b_{i1}} \right) \ast \frac{1}{|b_{i2}|} f_{s_2} \left( \frac{u}{b_{i2}} \right) \ast \ldots \ast \frac{1}{|b_{iN}|} f_{s_N} \left( \frac{u}{b_{iN}} \right).
\]

(10)

The components of the gradient of \( H(y_i) \) with respect to \( b_i \) (ith row of \( B \)) can be computed as:
\[
\frac{\partial H(y_i)}{\partial b_{ij}} = -\int_{-\infty}^{\infty} (1 + \log f_{y_i}(u)) \frac{\partial f_{y_i}(u)}{\partial b_{ij}} \, du \quad (11)
\]

To make explicit the dependence of the entropy \( H(y_i) \) on \( b_i \), define \( h(b_i) \triangleq H(y_i) \). In order to satisfy the first-order conditions given by (7), we must have that:

\[
\begin{bmatrix}
\nabla h(b_1) \\
\vdots \\
\nabla h(b_N)
\end{bmatrix}
= 
\begin{bmatrix}
b_1 \\
\vdots \\
b_N
\end{bmatrix}
\begin{bmatrix}
\nabla h(b_1) \\
\vdots \\
\nabla h(b_N)
\end{bmatrix}^T.
\] (12)

The resulting set of equations is equivalent to the following set of \( N(N-1) \) equalities:

\[
\nabla h(b_k)b_k^T = \nabla h(b_l)b_l^T \quad k, l = 1, \ldots, N \quad (k \neq l). \quad (13)
\]

Using expression (11), we get:

\[
\int_{-\infty}^{\infty} \log f_{y_i}(u) \left[ b_{i1} \frac{\partial f_{y_i}(u)}{\partial b_{k1}} + \cdots + b_{iN} \frac{\partial f_{y_i}(u)}{\partial b_{kN}} \right] du =
\]

\[
= \int_{-\infty}^{\infty} \log f_{y_i}(u) \left[ b_{k1} \frac{\partial f_{y_i}(u)}{\partial b_{i1}} + \cdots + b_{kN} \frac{\partial f_{y_i}(u)}{\partial b_{iN}} \right] du.
\] (14)

The computation of \( \frac{\partial f_{y_i}(u)}{\partial b_{ij}} \) can be efficiently carried out in the frequency domain. Using the conventional definition of characteristic function of a random variable [8]:

\[
\Phi_X(\omega) \triangleq \mathcal{F}\{f_X(x)\} = \int_{-\infty}^{\infty} f_X(x)e^{-j\omega x} \, dx,
\] (15)

we have from (10), using the convolution theorem:

\[
\Phi_{y_i}(\omega) = \Phi_{s_1}(b_{i1}\omega)\Phi_{s_2}(b_{i2}\omega) \cdots \Phi_{s_N}(b_{iN}\omega) \quad i = 1, \ldots, N.
\] (16)

If we assume that the pdfs \( f_{s_i} \) are continuous functions, with continuous derivatives almost everywhere, we can exchange the order of the integral and the derivative, and compute \( \frac{\partial f_{y_i}(u)}{\partial b_{ij}} \) as follows:

\[
\frac{\partial f_{y_i}(u)}{\partial b_{ij}} = \mathcal{F}^{-1}\{\omega \Phi_{s_1}(b_{i1}\omega) \cdots \Phi_{s_i}(b_{ij}\omega) \cdots \Phi_{s_N}(b_{iN}\omega)\}\quad (17)
\]

where \( \mathcal{F}^{-1} \) denotes the inverse fourier transform operator. The conditions imposed by (14) are satisfied, in particular, when:

\[
b_{i1} \frac{\partial f_{y_1}(u)}{\partial b_{k1}} + \cdots + b_{iN} \frac{\partial f_{y_N}(u)}{\partial b_{kN}} = 0 \quad k, l = 1, \ldots, N \quad (k \neq l).
\] (18)
If we substitute (17) into (18), and, under the assumption that all the characteristic functions are non-zero for every \( \omega \), we divide by \( \Phi_{s_1}(b_{k1}\omega) \cdots \Phi_{s_N}(b_{kN}\omega) \) the resulting expression, we obtain:

\[
\frac{\omega b_{11}\Phi'_{s_1}(b_{k1}\omega)}{\Phi_{s_1}(b_{k1}\omega)} + \cdots + \frac{\omega b_{1N}\Phi'_{s_N}(b_{kN}\omega)}{\Phi_{s_N}(b_{1N}\omega)} = 0 \quad k, l = 1, \ldots, N \quad (k \neq l). \quad (19)
\]

Notice that if and only if \( f_{s_i} \) is a gaussian pdf it holds that:

\[
\Phi'_{s_i}(\alpha \omega) = -\alpha \omega \Phi_{s_i}(\alpha \omega).
\]

Therefore in the special case where \( M = N \), i.e. all the original sources have a gaussian distribution, (19) simplifies as:

\[
-(b_{k1}b_{11} + \ldots + b_{kN}b_{1N})\omega^2 = -b_k^T b_\omega \omega^2 = 0 \quad k, l = 1, \ldots, N \quad (k \neq l), \quad (21)
\]

which are always satisfied because of the orthogonality constraints. Therefore, if all sources are gaussian, the resulting objective is a constant with respect to the elements of an arbitrary orthogonal unmixing matrix, and the separation is not possible.

When \( M \) is strictly less than \( N \), in order to simplify the notation, we can assume that the first \( M \) sources, \( (s_1, \ldots, s_M) \), are gaussian distributed. The equations in (19) can be simplified as:

\[
-\omega^2(b_{11}b_{k1} + \cdots + b_{1M}b_{kM}) + \frac{\omega b_{1M+1}\Phi'_{s_{M+1}}(b_{kM+1}\omega)}{\Phi_{s_{M+1}}(b_{1M+1}\omega)} + \cdots + \frac{\omega b_{1N}\Phi'_{s_N}(b_{kN}\omega)}{\Phi_{s_N}(b_{1N}\omega)} = 0 \quad k, l = 1, \ldots, N \quad (k \neq l) \quad (22)
\]

The subset of orthogonal matrices that satisfy this set of equalities is given by:

\[
B = \begin{bmatrix} Q & 0 \\ 0 & P \end{bmatrix}, \quad (23)
\]

where \( Q \) is an arbitrary \( M \times M \) orthogonal matrix, and \( P \) is a generalized permutation matrix. Notice, in fact, that:

\[
\Phi'_{s_i}(b_{ij}\omega)|_{b_{ij}=0} = -f_E[s_i] = 0 \quad i = 1, \ldots, N, \quad (24)
\]

if the sources are zero-mean\(^4\). This result shows that minima of the optimization problem, that was derived from the separation principle (1), appear in correspondence of matrices \( B \) that result in separation of the non-gaussian sources. Therefore, we proved the following theorem:

\(^4\) In general this is not a restriction because the mean can always be removed during pre-processing of the mixtures.
Theorem 1 (Extended ICA Identifiability Theorem). Given \( N \) independent and stationary signals \( s_1, \ldots, s_N \), \( M < N \) of which are gaussian distributed, the \( N - M \) non-gaussian distributed signals can be reconstructed, up to scaling and permutations, from any linear mixture of the type \( x = As \), where \( A \) is a full-rank \( N \times N \) matrix, solving the following optimization problem:

\[
\min_B \sum_{i=1}^{N} H(y_i) \quad \text{s.t.} \quad BB^T = I.
\]

(25)

Notice that in the summation (25), the index is up to \( N \) since the number of non-gaussian sources is not assumed to be known a-priori, thus preserving the "blindness" of the approach to the underlying distribution of the mixed signals.

4 Conclusions

An extension to the conventional identifiability theorem for ICA is introduced and rigorously proved. We show that, even when an arbitrary number of gaussian sources is included in the set of independent signals, the conventional mutual information minimization framework is still capable of separating all the non-gaussian signals, without requiring an ad-hoc ICA implementation. In particular, the main result of this paper is shown by investigating the properties of the extremum of the optimization problem derived from the separation principle.

References