INVESTIGATIONS ON CONTRAST FUNCTIONS FOR BLIND SOURCE SEPARATION BASED ON NON-GAUSSIANITY AND SPARSITY MEASURES

M. Sahmoudi and K. Abed-Meraim

TSI, Telecom-Paris, 37 rue Dareau, 75014 Paris, France

ABSTRACT
In this paper, we provide a systematic method to construct contrast functions through the use of sub- or super-additive functionals. The used sub- or super-additive functionals are applied to the distributions of the extracted sources to quantify the degree of non-Gaussianity or sparsity. In this work, we assume a completely blind scenario where one knows only the observations and the existence of at most one Gaussian independent component in the mixture. However, there is no a priori information about the mixing matrix or about the source density. Some practical examples of useful contrast functions are introduced and discussed in order to illustrate the usefulness of the proposed approach.

1. INTRODUCTION
In this work, we consider the standard linear instantaneous mixture model given by \( x = As \) where \( A \) is an unknown \( n \times m \) mixing matrix, \( x \) denotes the observation vector and \( s \) represents the source vector. The blind source separation (BSS) problem consists of finding a \( m \times n \) separating matrix \( B \) such that the components of \( y = Bx \) are independent. Note that in this work we consider BSS under orthogonality constraint assuming implicitly that a whitening step has already been performed. Thus, in the sequel, we suppose that \( B \) is an orthogonal matrix. The concept of contrast function for source separation has been first presented in [4]. A contrast function for source separation is a real valued function of the distribution of a random vector which is maximized when the source separation is achieved. To characterize mathematically a contrast function we use the following definition.

Definition 1 A functional \( F \) is a contrast function if and only if it satisfies the two requirements:

\[ F(Cs) \geq F(s) \] or (possibly) \( F(Cs) \leq F(s) \) for any independent random vector \( s \) and any invertible matrix \( C \).

\[ F(Cs) = F(s) \] holds if and only if \( C = PD \), where \( P \) and \( D \) are a permutation and a diagonal matrix, respectively.

Intuitively, a measure of dependence between the components of \( Bx \) would be a contrast, but there may be others. It should be noted that the construction of a contrast function is only a first step toward a separation procedure. Indeed, a contrast function needs to be associated with an efficient algorithm that estimates the underlying statistics and achieves the optimization problem. It should be however pointed out that the ease of the estimation and of the minimizing algorithm, both in term of implementation and computational cost, should be taken into account, besides performance considerations, in assessing the final separation method. For that, we will propose an efficient algorithm for minimizing the resulting empirical contrast. As existing examples of contrast functions, it has been shown in [4] and that the sum of the 4-th order cross-cumulants of the components is a contrast. Other contrast functions can be found in [12, 1], [8]. Note that the ideas of this paper are inspired from the projection pursuit methodology described in [7]. In this paper, Huber used sub- and super-additive functionals to define a statistical test of normality. Similarly, we use these classes of functionals to define some index of non-Gaussianity or sparsity. Thus minimizing the proposed criteria may be viewed as maximizing the non-Gaussianity or the sparsity of data. Finally we should note that during the revision period of this paper, we have found out that the idea of using sub- and super-additivity in BSS has been used by Pham in a previous work [13].

Remark 1 Some heuristic arguments that sub- and super-additivity functionals go together with non-Gaussianity and sparsity measures, are as follows:

- The cumulants which are widely used as measure of non-Gaussianity or sparsity (using kurtosis) are additive (sub-and super-additive) functionals.
- The exponential (power) Shannon entropy defined by

\[ H(x) = \exp \left\{ -2 \int \log f(x) dx \right\} \] (1)

where \( f \) is the density function of \( x \), which is commonly used for non-gaussianity measure is super-additive [3].

Definition 2 A functional \( F \) of the distribution of a random variable \( X \), denoted by \( F(X) \), is said to be scale equivariant if \( F(aX) = |a|F(X) \) for any real number \( a \). Without loss of generality, we can assume that \( F \geq 0 \).

2. ORTHOGONALITY CONSTRAINT
Principal Component Analysis (PCA) or whitening consists of transforming the observation vector into a white vector which entries are decorrelated. However, it is well known that the PCA job is not sufficient for separating the sources. To observe this, consider a square BSS model (i.e. \( n = m \)). For estimating the \( n \times n \) matrix \( A \), taking into account \( n \) scale ambiguities, we must determine \( n(n - 1) \) unknown coefficients. The second-order decorrelation constraints give \( n(n - 1)/2 \) equations, which is not sufficient for determining \( A \). This explains why Gaussian sources cannot be separated as they are characterized only by first and second order statistics. It fact, the second-order independence (whitening),

\footnote{These arguments follow the same procedure as in [15]. Furthermore, inspired from the proof of the fact that the minimum dispersion is a contrast function, this approach is a direct generalization of the previous one.}
solves the BSS problem up to an orthogonal transformation. Indeed, if $W$ is a whitening matrix satisfying $WE(xx^T)W^T = I$, then $UW$ is also a whitening matrix for any unitary matrix $U$. Thus one can say that whitening solves half of the problem of BSS, the other half being the estimation of the appropriate unitary matrix $U$. Because whitening is a very simple and standard procedure, much simpler than any BSS algorithms, it is a good idea to reduce the complexity of the problem this way. The remaining half of the parameters has to be estimated by some other method. This fact shows that for finding the other required equations, other information may be used such as higher-order statistics (HOS) [12] and/or fractional-order statistics (FLOS) [15]. Even in cases where whitening is not explicitly required, it is recommended, since it reduces the number of free parameters and considerably increases the performance, especially with high-dimensional data.

3. SUPER-ADDITIVITY BASED CONTRAST FUNCTIONS

Definition 3 A functional $\mathcal{G}$ of the distribution of a random variable $X$, denoted by $\mathcal{G}(X)$, is said to be $\sigma$-super-additive if

$$\mathcal{G}^\sigma(X + Y) \geq \mathcal{G}^\sigma(X) + \mathcal{G}^\sigma(Y)$$

(2)

for any two independent random variables $X$ and $Y$.

Theorem 1 Let $\mathcal{G}$ be a positive valued, $\sigma$-super-additive and scale equivariant functional and assume that the mixing matrix $A$ is orthogonal and that $\sigma$ is a real number such that $0 < \sigma < 2$. We assume also that the sparsity index $\mathcal{G}$ is always nonnegative at the considered sources, i.e. $\mathcal{G}(s_i) > 0 \forall i$. Then, the following objective function

$$\mathcal{C}(B) = - \sum_{i=1}^m \mathcal{G}^\sigma(y_i) \text{ where } y = Bx$$

(3)

is a contrast function for blind separation of linear mixtures under the orthogonality constraint of the mixing matrix $B$.

Proof

- Let write $C := BA$. Then, letting $C_{ij}$ be the general element of $C$ and $s_j$ be the components of $s$, one has $y_i = \sum_{j=1}^m C_{ij}s_j$. Using the $\sigma$-super-additivity and the scale equivariant properties of the functional $\mathcal{G}$, we have

$$-\mathcal{G}^\sigma(y_i) \leq -\sum_{j=1}^m \mathcal{G}^\sigma(C_{ij}s_j) \leq -\sum_{j=1}^m |C_{ij}|^\sigma \mathcal{G}^\sigma(s_j)$$

Summing this quantity for all output components and using the fact that $|C_{ij}|^\sigma \geq |C_{ij}|^2$ since $\sigma < 2$ and $|C_{ij}|^2 \leq \sum_{i=1}^m |C_{ij}|^2 = 1$ due to the orthogonality constraint, one gets

$$-\sum_{i=1}^m \mathcal{G}^\sigma(y_i) \leq -\sum_{i=1}^m \sum_{j=1}^m |C_{ij}|^\sigma \mathcal{G}^\sigma(s_j) \leq -\sum_{j=1}^m (\sum_{i=1}^m |C_{ij}|^2) \mathcal{G}^\sigma(s_j) = -\sum_{j=1}^m \mathcal{G}^\sigma(s_j)$$

So

$$\mathcal{C}(y) = \mathcal{C}(Cs) \leq \mathcal{C}(s)$$

(4)

Thus, requirement R1 is fulfilled.

- Finally, the equality $\mathcal{C}(Cs) = \mathcal{C}(s)$ requires that $|C_{ij}|^\sigma = |C_{ij}|^2$. It implies that $C_{ij}$ has exactly one nonzero component $C_{ij} = \pm 1$. Since $C$ is orthogonal, it means that $C = DP$, where $D$ denotes a diagonal matrix with entries $\pm 1$, and $P$ the permutation matrix associated to the permutation $i(1), \ldots, i(m)$. Clearly the equality in (4) is attained if $C$ is a permutation matrix. On the other hand if $C$ is not a separating matrix, then $\exists (i,j)$ such that $|C_{ij}| < 1$ and hence $|C_{ij}|^\sigma > |C_{ij}|^2$ meaning that the inequality in (4) is strict. This proves that $\mathcal{C}(B) = -\sum_{i=1}^m \mathcal{G}^\sigma(y_i)$ is a contrast function.

4. SUB-ADDITIVITY BASED CONTRAST FUNCTIONS

Definition 4 A functional $\mathcal{F}$ of the distribution of a random variable $X$, denoted by $\mathcal{F}(X)$, is said to be $\sigma$-sub-additive if

$$\mathcal{F}^\sigma(X + Y) \leq \mathcal{F}^\sigma(X) + \mathcal{F}^\sigma(Y)$$

(5)

for any two independent random variables $X$ and $Y$.

Theorem 2 Let $\mathcal{F}$ be a positive-valued $\sigma$-sub-additive and scale equivariant functional and assume that the mixing matrix $A$ is orthogonal. We assume also that the Gaussianity index $\mathcal{G}$ is always nonnegative at the considered sources, i.e. $\mathcal{F}(s_i) > 0 \forall i$. Let $\sigma$ be a real number such that $\sigma > 2$. Then, the following objective function

$$\mathcal{C}(B) = \sum_{i=1}^m \mathcal{F}^\sigma(y_i) \text{ where } y = Bx$$

(6)

is a contrast for blind separation of linear instantaneous mixtures under the orthogonality constraint of the mixing matrix $B$.

Proof

- To prove the first contrast function requirement, remember that with the same notations as in the proof of theorem 1, one has $y_i = \sum_{j=1}^m C_{ij}s_j$. Hence, using the equivariant and the sub-additivity properties of $\mathcal{F}$ we have

$$\sum_{i=1}^m \mathcal{F}^\sigma(y_i) = \sum_{i=1}^m \mathcal{F}^\sigma \left( \sum_{j=1}^m C_{ij}s_j \right) \leq \sum_{i=1}^m \sum_{j=1}^m \mathcal{F}^\sigma(C_{ij}s_j) \leq \sum_{i=1}^m \sum_{j=1}^m |C_{ij}|^\sigma \mathcal{F}^\sigma(s_j) \leq \sum_{i=1}^m \sum_{j=1}^m |C_{ij}|^2 \mathcal{F}^\sigma(s_j) \leq \sum_{j=1}^m \mathcal{F}^\sigma(s_j).$$
Similarly to theorem 1, the equality is attained if \( C \) is a generalized permutation matrix. This proves that \( C(B) = \sum_{i=1}^{m} S_{\alpha}^{\gamma} (y_i) \) is a contrast function.

\[ \text{Remark 2} \] It is possible to relax the condition on the source signals in the theorem (1) et theorem (2) in such a way to assume \( G(s_i) > 0 \) (resp. \( F(s_i) > 0 \)) for all \( i \) with strict inequality except for at most one source signal.

Thus to separate orthogonal linear mixtures, one needs only to find appropriate sub- or super-additive scale equivariant functional. Examples of such functionals are given next.

**5. EXAMPLES OF SUB OR SUPER ADDITIVITY BASED CONTRAST FUNCTIONS**

It is important to note that the statistical independence and sparsity are different criteria. However, for alpha-stable distributed random signals both sparsity and non-Gaussianity are consistent in the sense that the independence and sparsity criteria coincide for super-Gaussian signals with positive kurtosis. In that case, the connection between the two concepts comes from the fact that mixing independent signals decreases the sparsity. In this paper, we observe that the super-additivity is a measure used essentially for sparsity while sub-additivity is used for measuring non-Gaussianity. This fact is illustrated by the following examples.

**5.1. Contrast functions for sparsity measure**

The relationship between Independent Component Analysis (ICA) and sparsity is studied by many authors [10]. As argued in [6], the choice of the sparseness measure is not a minor detail but may have far-reaching implications on the structure of a solution. Below we present two examples of such a measure.

- **Alpha-stable scale contrast function**

  Here, we consider a linear mixture of \( \alpha \)-stable signals with the same characteristic exponent \( \alpha \), dispersion \( \gamma \) and scale functional \( S = \gamma^\alpha \). It is well known that \( S \) is scale equivariant \( S(aY) = a^{\alpha}S(Y) = |a|S(Y) \). Note that the scale functional is also \( \sigma \)-super-additive for any \( \sigma \geq \alpha \). To prove this, let us consider two independent r.v. \( X \) and \( Y \), then we have

  \[
  S(X + Y) = (\gamma_{X+Y})^\alpha = (\gamma_X + \gamma_Y)^\alpha \geq (\gamma_X)^\alpha + (\gamma_Y)^\alpha.
  \]

The last inequality valid for \( \sigma \geq \alpha \) follows from the fact that for non negative numbers \( u, v, r \), with \( r \geq 1 \) one has \( (u + v)^r \geq u^r + v^r \), because \( (u+v)^r - u^r = \int_0^{u+v} rt^{r-1}dt = \int_0^u r(t+u)^{r-1}dt + \int_0^v r(t+u)^{r-1}dt = v^r \).

From theorem 1 and for \( \sigma < 2 \) the sum of scale of all output BSS model

\[
C(B) = -\sum_{i=1}^{m} S_{\alpha, \gamma}^{\sigma} (y_i)
\]

defines a contrast function that can separate linear alpha-stable mixtures. In the algorithm derivation step one can estimate \( S_{\alpha, \gamma}^{\sigma} \).

using one of existing method to estimate the dispersion \( \gamma_{\alpha} \) [2]. In particular, for \( \sigma = \alpha \), we retrieve the minimum dispersion criterion (MDC) which we introduced in [15].

- **\( \ell_p \)-norm based contrast functions for \( p \leq 1 \)**

  Let \( F(Y) \) be the so called \( \ell_p \) norm of the random variable \( Y \) which has been used as a sparsity measure and exploited for the sources separation [9, 10]. Using the super-additivity concept, we provide here a rigorous proof justifying its use in BSS. Note that for \( p < 1 \), the functional \( F(Y) = \|Y\|_p \) is super-additive for positive r.v. and scale equivariant. Note also that sparse sources can be defined as sources which have disjoint time-support and hence \( \|X + Y\|_p = \|X\|_p + \|Y\|_p \) for any r.v. \( X \) and \( Y \) satisfying the above sparseness property. Differently, we can prove mathematically that for \( p \leq 1 \), we have \( \|X + Y\|_p \geq \|X\|_p + \|Y\|_p \) for positive r.v. \( X \) and \( Y \). Thus, \( F(Y) = \|Y\|_p \) is "nearby" 1-super additive. Following theorem 2 results, the \( \ell_p \)-norm criterion

\[
C_p(B) = \sum_{i=1}^{m} \|y_i\|_p
\]

is a contrast function for sparse signals separation.

**Remark 3**

- It is worth to emphasis that the existence of fractional lower-order moments imply that the \( \ell_p \)-norm contrast function can separate mixtures of heavy-tailed \( \alpha \)-stable signals with different characteristic exponent \( \alpha_k \) by choosing \( p < \min\{1, \min_k(\alpha_k)\} \).

- The important question then is "can a sparsity criterion, like criterion (8), lead to source separation?". If the source vector is sufficiently sparse, then it can be successfully recovered. Otherwise, blind separation can be implemented in the time-frequency domain [11]. Indeed, if the assumption is not satisfied for analyzed data, a preprocessing (e.g., wavelet transformation [6]) to the known data matrix \( x \) is necessary in order that the assumption holds in the transform domain.\(^4\)

**5.2. Contrast functions for non-Gaussianity measure**

A classical measure of non-Gaussianity, is based on the cumulants which present a lot of suitable properties like as equivariance and additivity.

- **Cumulant - based contrast functions**

  Many authors have used higher order cumulants for the separation of non-Gaussian sources. We revisit this class of contrast functions using the sub-additivity property. Let assume that the \( k \)-th order cumulant (with \( k > 2 \)) of the sources are non-zero, i.e. \( \text{Cum}_k(s_i) > 0 \). Let \( F(Y) \) denote \( \text{Cum}_k(Y) = |\text{Cum}_k(Y)|^{1/k} \) denote the \( k \)-th order cumulant of the random variable \( Y \). This functional is scale equivariant and \( k \)-sub-additive. Indeed, using the additivity property of the cumulants, we have for independent \( X \) and \( Y \) r.v.

\[
F_k(X + Y) = |\text{Cum}_k(X + Y)| = |\text{Cum}_k(X) + \text{Cum}_k(Y)| \leq |\text{Cum}_k(X)| + |\text{Cum}_k(Y)|
\]

Hence, from theorem 2, \( \sum_{i=1}^{m} |\text{Cum}_k(Y_i)| \) is a contrast function.

\(^3\)For \( p < 1 \), \( \ell_p \)-norm is not a valid norm, (it does not satisfy the triangle inequality) but we choose to keep the same terminology for convenience.

\(^4\)The applied preprocessing depends on the nature of analyzed data.
6. MAXIMIZATION USING JACOBI-GRADIENT ALGORITHM

As known and used in [15], every orthogonal matrix can be parameterized in terms of Givens rotation angles, each of which defines a rotation in a single plane of the high-dimensional vector space. Then, these individual rotations can be cascaded to span the whole set of rotation matrices. Every rotation matrix has a unique set of Givens rotation angles that characterize it. In n-dimensions, a Givens rotation matrix in the plane formed by the i-th and j-th axes is denoted by \( \Omega_{ij} \). A rotation matrix is then formed from these sparse matrices according to

\[
B = \prod_{p=1}^{m-1} \prod_{q=p+1}^{m} \Omega_{pq}
\]  

(9)

The multiplication order can be always taken from the left or from the right. It is not crucial to the generality of this formula as long as we maintain the same order when taking the derivative of the matrix with respect to a rotation angle.

6.1. Optimization & algorithm

Our aim is to solve the previously mentioned constrained optimization problem, which becomes unconstrained if Givens angles are used: Let \( \theta_{ij} \), \( k = 1, \ldots, m - 1 \) and \( \Theta \) be the Givens rotation angles that form our parameter vector \( \Theta \). To derive a nice and fast algorithm, we propose here to combine the Jacobi-like decomposition of Givens rotations and the Gradient algorithm using a numerical computation for searching \( \Theta \). The so called Jacobi-Gradient algorithm can be summarized as follows:

Jacobi-Gradient Algorithm

\begin{itemize}
  \item [Step 1.] Initialize Givens angles randomly.
  \item [Step 2.] Estimate robustly from data, especially if the sources are in noisy environment, the considered sources statistics used in the contrast function.
  \item [Step 3.] Calculate the gradient of the cost function with respect to the Givens angles. The gradient is \( \Theta = \frac{\partial C(B)}{\partial \Theta} \).
  \item [Step 4.] Update the Givens angles using gradient ascent
    \[
    \Theta(k + 1) = \Theta(k) - \eta \frac{\partial C(B)}{\partial \Theta}
    \]
  \item [Step 5.] Go back to step 3 and continue until convergence.
\end{itemize}

6.2. Complexity

A key concern in many adaptive algorithms is the computational complexity. It is clear that if the multiplications in (9) are performed from left, the first output is only affected from the Givens angles with indices \( \theta_{ij} \), \( q = 2, \ldots, m \), the second is affected by all the angles \( \theta_{ij} \), \( i = 1, \ldots, m \). Thus, if we wish to extract the m source component, we only need to adapt the angles \( \theta_{ij} \), \( i = 1, \ldots, m \), \( j = i + 1, \ldots, m \) which makes a total of \( m^2 - m(m + 1)/2 \) parameters, that is less than \( Km(m + 1)/2 \) (\( K \) being the number of iterations) parameters required in many Jacobi-like algorithms. But then, we will have to evaluate either the sin and cos of all these parameters once. In addition, the necessary matrix vector multiplications in the algorithm will be performed at each iteration, which amount \( O(Km(m + 1)/2) \) flops per iteration.

7. CONCLUDING REMARKS

In this paper, a general and unified mathematical framework to construct contrast functions, using super and sub additivity, have been proposed. More precisely, super-additivity is used to measure sparsity while sub-additivity is exploited to measure the non-Gaussianity which are two known criteria for BSS.

Some practical and illustrative examples of the above concepts are provided. In particular, robust contrast functions were derived for blind separation of heavy-tailed sources. By using the proposed approach, we avoid working explicitly with the high dimensionality of the joint densities.

In addition, coupling Jacobi and Gradient optimization techniques, a nice implementation was proposed to maximize (or minimize) the previous contrast functions under orthogonality constraint.

8. REFERENCES