

## Non-orthogonal Support-Width ICA

John A. Lee<sup>1</sup>, Frédéric Vrins<sup>1</sup> and Michel Verleysen<sup>1</sup> \*

1 – Université catholique de Louvain – Machine Learning Group  
Place du Levant 3, B-1348 Louvain-la-Neuve, Belgium  
<http://www.ucl.ac.be/mlg/>

**Abstract.** Independent Component Analysis (ICA) is a powerful tool with applications in many areas of blind signal processing; however, its key assumption, i.e. the statistical independence of the source signals, can be somewhat restricting in some particular cases. For example, when considering several images, it is tempting to look on them as independent sources (the picture subjects are different), although they may actually be highly correlated (subjects are similar). Pictures of several landscapes (or faces) fall in this category. How to separate mixtures of such pictures? This paper proposes an ICA algorithm that can tackle this apparently paradoxical problem. Experiments with mixtures of real images demonstrate the soundness of the approach.

### 1 Introduction

For two decades, Independent Component Analysis (ICA) [1, 2] has brought an elegant solution to many Blind Source Separation problems (BSS). Examples of applications are the cocktail party problem, ECG/EEG/MEG processing [3, 4], signal denoising, etc. The key assumption in ICA is the statistical independence of the source signals. Starting from there, many mixtures models (instantaneous, convolutive, linear, post-non-linear) can be developed and several algorithm can be implemented (gradient ascent, fixed-point, joint diagonalization, etc.).

Sometimes however, statistical independence may be a restricting assumption. Indeed, assume that the source vector  $\mathbf{s}$  consists of composite random variables that can be written as

$$s_i = \begin{cases} c_i & \text{with probability } \alpha \\ p_i & \text{with probability } 1 - \alpha \end{cases}, \quad (1)$$

where  $0 \leq \alpha \leq 1$ . In the last equation, the random variables  $c_i$  are strongly correlated whereas the  $p_i$  are fully independent. Intuitively, all sources can be seen as independent variations of a common pattern. This situation may be encountered when dealing with images. Several images may share a common underlying pattern, like e.g. various pictures of landscapes or identity pictures. It is clear that in each of these sets, images are correlated although we would like to consider them to be independent. Within that framework, separating mixtures of such images becomes a difficult task, at least with usual ICA algorithms, in which mixtures are systematically decorrelated, either by prewhitening or by

---

\*JAL and MV are respectively a Scientific Research Worker and a Research Director with the Belgian FNRS (Fonds National de la Recherche Scientifique).

the ICA algorithm itself. Solving the above problem requires leaving aside the common patterns  $c_i$  and focusing on the independent ones  $p_i$ . For this purpose, the approach followed in this paper assumes that  $c_i$  and  $p_i$  influence the source pdfs in different locations of their supports. More precisely, it is assumed that i)  $c_i$  act exclusively on the inner part of the  $i$ th source support, ii) tails and bounds of the  $i$ th source support are determined solely by  $p_i$  and iii) pdf tails are sharply cut. This situation is not uncommon in digital image processing, because image are often under- and/or over-exposed.

The remainder of this paper is organized as follows. After this introduction, Section 2 reviews the classical mixture model of ICA. Section 3 describes an ICA contrast function that exploits information contained in the pdf tails. Next, Section 4 presents the NOSWICA algorithm, which can tackle the above-mentioned problem. Section 5 shows some experimental results of NOSWICA. Finally, conclusions are drawn in Section 6.

## 2 Mixture model

Within the ICA framework, it is usually assumed that sources are mixed in an instantaneous and linear way. This model can be written as follows:

$$\mathbf{x} = \mathbf{A}\mathbf{s} , \quad (2)$$

where  $\mathbf{x}$  is the vector of observed signals,  $\mathbf{s}$  the vector of independent sources and  $\mathbf{A}$  the mixing matrix. Assuming that sources are statistically independent allows one to identify  $\mathbf{A}$  by solving blindly the reverse model:

$$\mathbf{y} = \mathbf{B}\mathbf{x} = \mathbf{B}\mathbf{A}\mathbf{s} . \quad (3)$$

In this equation, the unmixing matrix  $\mathbf{B}$  is such that  $\mathbf{C} = \mathbf{B}\mathbf{A} = \mathbf{P}\mathbf{D}$ , where  $\mathbf{P}$  and  $\mathbf{D}$  are respectively a permutation matrix and a diagonal one.

Typically, most ICA algorithms proceed in two stages: first, mixtures are whitened, i.e. are decorrelated and standardized; second, full independence is reached by maximizing some contrast function. Prewhitening allows decomposing the unmixing matrix into the product  $\mathbf{B} = \mathbf{W}^T\mathbf{V}$ , where  $\mathbf{V}$  is the whitening matrix and  $\mathbf{W}$  is the ICA matrix, constrained to be orthogonal. Prewhitening also allows rewriting the unmixing model (3) as

$$\mathbf{y} = \mathbf{W}^T\mathbf{V}\mathbf{x} = \mathbf{W}^T\mathbf{z} , \quad (4)$$

where both  $\mathbf{y}$  and  $\mathbf{z}$  have zero mean and unit variance. The orthogonality constraint on the ICA matrix  $\mathbf{W}$  makes the ICA problem much easier to solve. Indeed, there are fewer parameter to identify and in particular orthogonality prevents the algorithm from extracting twice the same source.

## 3 The support width measure as ICA contrast

A contrast is a function  $\mathcal{C}$  of one or all ICA outputs that respectively measures the 'level of non-Gaussianity' or the 'level of independence'. Maximizing the contrast

allows solving the ICA problem, either at once or by extracting one source at a time. Usual contrasts for all outputs are for example minus the mutual information [5]. Contrasts for one output is typically the absolute value of the normalized kurtosis, the negentropy or any other measure of ‘non-Gaussianity’.

Previous work [6] has shown that for bounded sources, the support width measure (SWM) of an ICA output can be used as contrast. The support  $\Omega(Y)$  of a random variable  $y$  is the subset of the domain where its pdf is non-zero. If the support is non-convex, then the convex hull  $\bar{\Omega}(Y)$  of the support may be used instead. The support width is the length of the shortest interval that contains the (convex hull of the) support. Formally, this can be written using the Lebesgue measure of sets  $\mu[\cdot]$ :

$$\mathcal{C}(y_i) = -\mu[\bar{\Omega}(y_i)] = -\mu[\bar{\Omega}(\mathbf{w}_i^T \mathbf{z})] , \quad (5)$$

where  $\mathbf{w}_i$  is the  $i$ th column of the matrix  $\mathbf{W}$ . Maximizing  $\mathcal{C}$  w.r.t.  $\mathbf{w}_i$  under the constraint  $\mathbf{w}_i^T \mathbf{w}_i = 1$  allows extracting one source.

Practically, pdfs are unknown and only samples of the observed variables are provided. In this case, the support width can be estimated using order statistics. Assuming that observations of  $y_i$  are sorted in the list  $[y_i^{(1)}, y_i^{(2)}, \dots, y_i^{(N-1)}, y_i^{(N)}]$ , then a good estimator of the support width is

$$\hat{\mathcal{C}}(y_i) = \frac{1}{m} \sum_{j=1}^m y_i^{(m)} - \frac{1}{m} \sum_{j=1}^m y_i^{(N+1-m)} , \quad (6)$$

where  $1 \leq m \ll \lfloor N/2 \rfloor$ . Taking  $m = 1$  amounts to measuring the interval between the minimum and maximum observed values. Taking  $m > 1$  leads to a more robust estimator (see [7] for more details).

## 4 The NOSWICA algorithm

The SWM contrast is not everywhere differentiable. As a consequence, most usual ICA algorithm based on gradient ascent or fixed-point approaches would fail to maximize it. A deflation procedure that can handle non-differentiable contrasts is detailed in [8]. It relies on a naive but yet efficient trial-and-error optimization scheme. The association of that procedure with the SWM contrast is called SWICA.

As many other ICA algorithms, SWICA works on prewhitened mixtures. Unfortunately, in the source model described in the introduction, prewhitening does not help because sources are correlated. Maintaining the orthogonality constraint on  $\mathbf{W}$  would then amount to imposing that  $\mathbf{y}$  is white, what obviously contradicts the source model. At least, whitening remains useful to make variance constant in all directions, but it may not be stated that  $\mathbf{W}$  is orthogonal anymore: only the constraint  $\mathbf{w}_i^T \mathbf{w}_i = 1$  for all  $i$  remains acceptable. As a consequence, orthogonality has to be replaced with some kind of penalty in order to avoid extracting repeatedly the same source. In the case of the SWM contrast, geometrical arguments help to identify the ideal penalty.



the dashed parallelogram is always longer than for the solid one:

$$B = \frac{\mu_2}{|\sin \theta|} \leq \frac{\hat{\mu}_2}{|\sin \hat{\theta}|} = \hat{B} = B + h \left| \cot \theta - \cot \hat{\theta} \right|. \quad (7)$$

These inequalities allow concluding that  $A \leq \hat{A}$  and  $A = \hat{A}$  iff  $\hat{\theta} = \theta + k\pi$ . Differentiating  $\hat{A}$  w.r.t.  $\hat{\theta}$  also shows that no other (local) minimum exists. Therefore, as the constant  $h$  can be omitted, minimizing  $\hat{B}$  w.r.t.  $\hat{\mathbf{w}}_2$  (instead of  $\hat{A}$ ) allows determining  $\mathbf{w}_2$ , the direction of the second source.

The above reasoning can be extended to more than two dimensions. For example, assume that two sources have been extracted and that  $\mathbf{w}_1$  and  $\mathbf{w}_2$  are known. Assume also that  $\mathbf{U} = [\mathbf{u}_1, \mathbf{u}_2]$  is an orthonormal basis of the subspace spanned by  $\mathbf{w}_1$  and  $\mathbf{w}_2$ . Then, minimizing  $\hat{B} = \hat{\mu}_3 / |\sin \hat{\theta}|$  w.r.t.  $\hat{\mathbf{w}}_3$ , where  $\hat{\mu}_3 = \mu[\hat{\Omega}(\hat{\mathbf{w}}_3^T \mathbf{z})]$  and  $|\sin \hat{\theta}| = \|\hat{\mathbf{w}}_3 - \mathbf{U}\mathbf{U}^T \hat{\mathbf{w}}_3\|$  leads to the right solution, without any risk of converging on  $\mathbf{w}_1$  or  $\mathbf{w}_2$ . In summary, the deflation algorithm of NOSWICA consists of the following steps:

- Whiten the mixtures and initialize the ICA matrix:  $\mathbf{z} \leftarrow \mathbf{V}\mathbf{x}$ ,  $\mathbf{W} \leftarrow \mathbf{I}$ .  
 (Whitening is used here for obtaining mixtures with similar variances; visually, this allows avoiding cases where the parallelogram would be too flat, making the estimator of the support width not robust enough.)
- Minimize the support width in order to recover a first source:  
 $\mathbf{w}_1 \leftarrow \arg \min_{\hat{\mathbf{w}}_1} \mu[\hat{\Omega}(\hat{\mathbf{w}}_1^T \mathbf{z})]$ ,  $\mathbf{U} \leftarrow [\mathbf{w}_1]$ .
- To extract other sources, minimize the penalized support and update  $\mathbf{U}$ :  
 $\mathbf{w}_i \leftarrow \arg \min_{\hat{\mathbf{w}}_i} \mu[\hat{\Omega}(\hat{\mathbf{w}}_i^T \mathbf{z})] / \|\hat{\mathbf{w}}_i - \mathbf{U}\mathbf{U}^T \hat{\mathbf{w}}_i\|$   
 $\mathbf{U} \leftarrow [\mathbf{U}, \mathbf{w}_i - \mathbf{U}\mathbf{U}^T \mathbf{w}_i / \|\mathbf{w}_i - \mathbf{U}\mathbf{U}^T \mathbf{w}_i\|]$  (append a column)

From a practical point of view, the support estimator in (6) seems to be the most robust and the optimization procedure described in [8] gives good results.

## 5 Experiments

*Toy example.* Two sources have been artificially generated according to the model in (1), with  $\alpha = 0.5$ . The data set is then an equiprobable mixture of samples of both the  $c_i$  and  $p_i$ . Obtaining the two  $p_i$  is easy: each  $p_i$  is drawn from a uniform distribution in the interval  $(-1, +1)$ . On the other hand, the two  $c_i$  are built as follows. First, two independent zero-mean unit-variance Gaussian distributions are sampled. Next, the obtained vectors are mixed by premultiplying them with the matrix  $[0.2, 0.4; 0.4, 0.2]$ , in order to correlate their components. Finally, all values greater than one in absolute value are replaced with  $\pm 1$ . The covariance matrix of the resulting sources is  $\Sigma_{\mathbf{ss}} = [0.25, 0.07; 0.07, 0.25]$ . The first plot in Fig. 2 shows 500 points obtained according to the above building scheme. The second plot displays mixtures of these sources. The third plot shows the whitened mixtures. The three last plots illustrate the results of FastICA [9] (v2.5, deflation, pow3), SWICA [6, 8] ( $m = 5$  in (6)) and NOSWICA





