ESANN'2006 proceedings - European Symposium on Artificial Neural Networks
Bruges (Belgium), 26-28 April 2006, d-side publi., ISBN 2-930307-06-4.

Non-orthogonal Support-Width ICA

John A. Lee[1], Frédéric Vrins[1] and Michel Verleysen[1] *

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

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

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 ith source support, ii) tails and bounds of the ith 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:

$$\mathbf{x} = \mathbf{A}\mathbf{s} \ , \tag{2}$$

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:

$$\mathbf{y} = \mathbf{B}\mathbf{x} = \mathbf{B}\mathbf{A}\mathbf{s} \ . \tag{3}$$

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

$$\mathbf{y} = \mathbf{W}^T\mathbf{V}\mathbf{x} = \mathbf{W}^T\mathbf{z} \ , \tag{4}$$

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 $\mu[\cdot]$:

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

where $\mathbf{w}_i$ is the ith 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 $[y_i^{(1)}, y_i^{(2)}, \ldots, y_i^{(N-1)}, y_i^{(N)}]$, then a good estimator of the support width is

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

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.

Considering two sources that are bounded and independent, the convex hull of their joint support is a rectangle. Similarly, two linear mixtures of these sources live in a parallelogram. If the sources are fully independent, the orthogonality constraint on $\mathbf{W}$ is meaningful because whitening these mixtures transforms the parallelogram back to a (rotated and scaled) rectangle. However, if sources are composite as in (1), this property does not hold anymore and whitening merely leads to another (rotated and scaled) parallelogram[1]. For the sake of simplicity, it can then be assumed without loss of generality that $\mathbf{z} = \mathbf{x}$ and $\mathbf{B} = \mathbf{W}^T = [1, 0; \cos\theta, \sin\theta]$ where θ is the mixing angle. In this case, the joint support corresponds to the solid parallelogram in Fig. 1. Under these assumptions, recovering the direction of the first source is trivial ($\mathbf{w}_1 = [1, 0]^T$). This can be achieved by minimizing the support without any constraint or penalty. On the other hand, finding the direction of the second source proves more difficult: minimizing the estimated support without penalty could lead to the first source once again. An alternative approach would be to minimize the area of the joint support instead of the width of the marginal support. The true area of the joint support (the solid parallelogram in Fig. 1) can be written as a function of the two marginal supports: $A = \mu_1\mu_2/|\sin\theta|$, where μ_i is a short-hand notation for $\mu[\bar{\Omega}(s_i)]$. The last formula expresses the area as the intersection of two infinite-length strips, with crossing angle θ and widths equal to μ_1 and μ_2 respectively. Knowing only the estimate of the marginal support in direction $\hat{\mathbf{w}}_2$, A can be approximated by $\hat{A} = \mu_1\hat{\mu}_2/|\sin\hat{\theta}|$, where $\hat{\mu}_2 = \mu[\bar{\Omega}(\hat{\mathbf{w}}_2^T\mathbf{z})]$. This estimate corresponds to the area of the dashed parallelogram in Fig. 1: the second strip is not perfectly known but must be wide enough to cover the true support. By construction, the height h is shared by the two parallelograms and equals the support of the first source ($h = \mu_1$). On the other hand, the base of

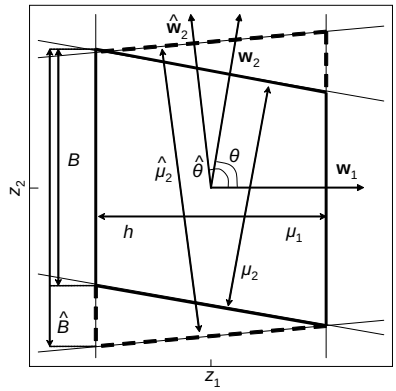

Fig. 1: Schematic view of the joint support of two mixtures of two bounded sources. The solid parallelogram is the true convex hull of the joint support; the dashed one represents the hypothetic support when one source direction is perfectly known and the second one is only approximated. Recovering the second source may be achieved by finding the angle that minimizes the area of the dashed parallelogram.

[1]As some source components in (1) may be correlated, the symmetry axes of their component-wise joint supports are not necessarily aligned. See an example in Fig. 2: whitening leads to a 'tradeoff', where both the independent uniform components and the correlated Gaussian ones influence the final result.

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

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

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 $\hat{\mu}_3 = \mu[\bar{\Omega}(\hat{\mathbf{w}}_3^T\mathbf{z})]$ 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:
  $\mathbf{w}_1 \leftarrow \arg\min_{\hat{\mathbf{w}}_1} \mu[\bar{\Omega}(\hat{\mathbf{w}}_1^T\mathbf{z})]$, $\mathbf{U} \leftarrow [\mathbf{w}_1]$.
- To extract other sources, minimize the penalized support and update $\mathbf{U}$:
  $\mathbf{w}_i \leftarrow \arg\min_{\hat{\mathbf{w}}_i} \mu[\bar{\Omega}(\hat{\mathbf{w}}_i^T\mathbf{z})]/\|\hat{\mathbf{w}}_i - \mathbf{U}\mathbf{U}^T\mathbf{w}_i\|$
  $\mathbf{U} \leftarrow [\mathbf{U}, \mathbf{w}_i - \mathbf{U}\mathbf{U}^T\mathbf{w}_i/\|\mathbf{w}_i - \mathbf{U}\mathbf{U}^T\mathbf{w}_i\|]$ (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 ± 1. The covariance matrix of the resulting sources is $\mathbf{\Sigma_{ss}} = [0.25, 0.07; 0.07, 0.25]$. 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

($m = 5$). FastICA does not recover any source, SWICA extracts the first one but misses the second one because of the orthogonality constraint. Only NOSWICA yields the expected result.

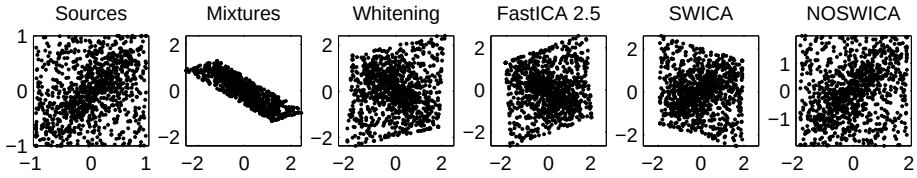

Fig. 2: Toy example with computer-generated composite distributions. The six plots show respectively the two sources, two random mixtures, the whitened mixtures and results of FastICA, SWICA and NOSWICA.

Mixtures of real images. For this experiment, color pictures of the three authors are downloaded from `http://www.dice.ucl.ac.be`. Pictures are cropped in order to share the same size (192 by 144 pixels) and converted to grayscale using `rgb2gray` in Matlab, as shown in Fig.3. Next, images are transposed and rows of pixels are concatenated to obtain three row vectors that contain the observations of each source. Finally, those vectors are standardized so that sources have zero mean and unit variance. The leftmost plots in Fig. 3 show the histograms of the sources. Computing the covariance matrix of these sources leads to

$$\mathbf{\Sigma_{ss}} = \begin{bmatrix} 1.00 & -0.27 & -0.25 \\ -0.27 & 1.00 & 0.51 \\ -0.25 & 0.51 & 1.00 \end{bmatrix} . \tag{8}$$

Pictures corresponding to three random linear combinations of the sources are shown in the third column of Fig. 3. The three rightmost columns show the results of FastICA (v2.5, deflation, pow3), SWICA ($m = 500$) and NOSWICA ($m = 500$). As in the case of the toy example, NOSWICA clearly outperforms the two other algorithms. The quality of the results may be assessed using a performance index (PI) that reflects the residual interferences between the recovered sources. In order to compute the latter, it is assumed that both the sources and ICA outputs are standardized. For a given ICA output vector $\mathbf{y}$, the equation system $\mathbf{y} = \mathbf{Cs}$ is solved and for each source the PI is computed as

$$\mathrm{PI}(y_i) = \frac{\sum_j |c_{ij}|}{\max_j |c_{ij}|} - 1 \ , \tag{9}$$

where c_{ij} is the entry of $\mathbf{C}$ at the crossing of the i-th row and j-th column. In Fig. 4, the three algorithms are run 1000 times with different mixtures and histograms of the PIs are displayed for each source or for all of them (average PI). As it can also be seen in Table 1, the average result of NOSWICA is excellent. On the other hand, its robustness needs some improvement: the standard deviation

of the PIs is not negligible. Because the orthogonality constraint is relaxed in NOSWICA, the space of solutions is larger than for the two other algorithms. This could account for the result variability.

6 Conclusions

Most ICA algorithms rely on the statistical independence of the unknown sources to be recovered. A slightly different model is proposed in this paper, where sources have composite distributions. More precisely, it is assumed that only some parts of these composite distributions are independent whereas others may be heavily correlated.

The adoption of this model gives rise to several consequences. Firstly, separating the sources requires defining ad hoc contrast functions, that take into account the independent parts of the sources and neglect their correlated parts. As shown in this paper, the support width measure is a possible answer to this issue, in the case of bounded sources. Secondly, the prewhitening step involved in many ICA algorithms becomes partly useless here. Indeed, decorrelating the mixtures would contradict the source model.

A non-orthogonal ICA algorithm for the optimization of the support width measure, called NOSWICA, has been developed to identify the parameters of the proposed model. Experiments show the soundness of the approach in the case of mixtures of real images.

References

[1] P. Comon. Independent component analysis, a new concept? *Signal Processing*, 36(3):287–314, 1994.

[2] A. Hyvärinen, J. Karhunen, and E. Oja. *Independent component analysis*. John Willey ans Sons, Inc., New York, 2001.

[3] L. De Lathauwer, B. De Moor, and J. Vandewalle. Fetal electrocardiogram extraction by blind source subspace separation. *IEEE Trans. Biomed. Eng.*, 47(5):567–572, 2000.

[4] R. Vigario, J. Sarela, V. Jousmaki, M. Hamalainen, and E. Oja. Independent component approach to the analysis of eeg and meg recordings. *IEEE trans. on biomedical engineering*, 47(5):589–593, 2000.

[5] H.H. Yang and S. Amari. Adaptive on-line learning algorithms for blind separation - maximum entropy and minimum mutual information. *Neural computation*, 9:1457–1482, 1997.

[6] F. Vrins and M. Verleysen. SWM: A class of convex contrasts for source separation. In *proceedings of ICASSP'05*, pages V.161–V.164, Philadelphia (USA), 2005.

[7] F. Vrins and M. Verleysen. Quasi-ranges estimators for minimum support independent component analysis, 2006. Submitted to ICA 06, Charleston (USA).

[8] J.A. Lee, F. Vrins, and M. Verleysen. A simple ica algorithm for non-differentiable contrasts. In *Proc. European Signal Processing Conference (EUSIPCO 2005)*, Antalya (Turkey), September 2005.

[9] A. Hyvärinen. Fast and robust fixed-point algorithms for independent component analysis. *IEEE Trans. on Neural Networks*, 10(3):626–634, 1999.

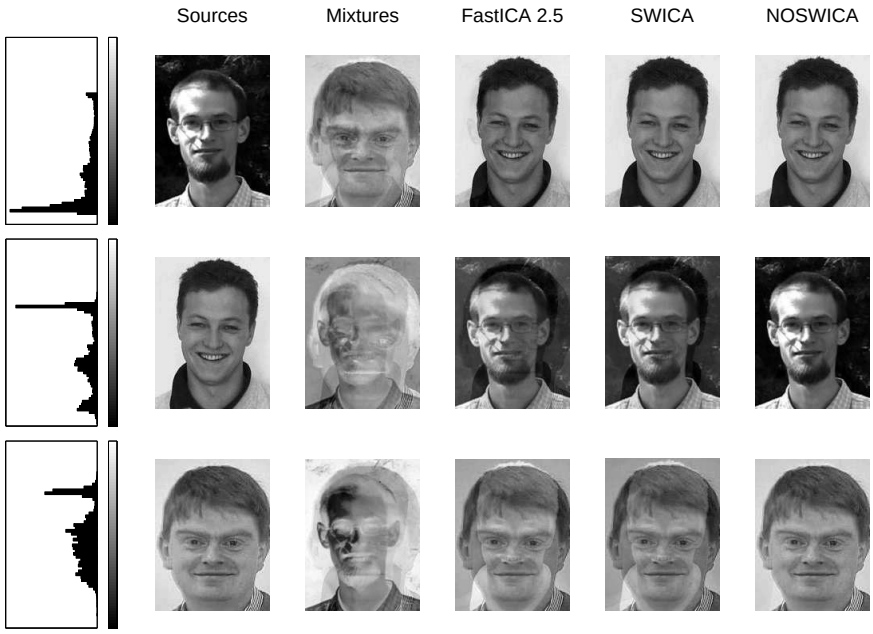

Fig. 3: Mixtures of real images (pictures of the three authors' faces). The two leftmost columns shows the sources and their histogram. The third column consists of three random mixtures of the sources. The three rightmost columns display results of FastICA, SWICA and NOSWICA. All images are displayed using a grayscale that fits to their own intensity range.

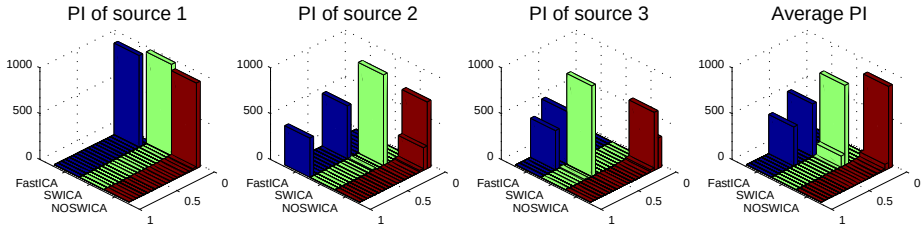

Fig. 4: Histograms of PI for the example in Fig. 3. Lowest values are the best.

PI	FastICA 2.5	SWICA	NOSWICA
Source 1	0.1472 (*0.0204*)	*0.0423* (0.0306)	0.0555 (0.0888)
Source 2	0.6528 (0.2464)	0.3126 (*0.0125*)	*0.0599* (0.0583)
Source 3	0.6684 (0.0810)	0.5928 (*0.0172*)	*0.0926* (0.0542)
Average	0.4895 (0.1159)	0.3159 (*0.0172*)	*0.0693* (0.0463)

Table 1: Mean and standard deviation (in parentheses) of the performance indexes for the example in Fig. 3, over 1000 trials with different random mixtures. Best values are italicized.