ESANN'2004 proceedings - European Symposium on Artificial Neural Networks
Bruges (Belgium), 28-30 April 2004, d-side publi., ISBN 2-930307-04-8, pp. 211-216

Towards a Local Separation Performances Estimator Using Common ICA Contrast Functions ?

Frédéric Vrins, Cédric Archambeau and Michel Verleysen*
Université catholique de Louvain (UCL) - Machine Learning Group
Place du Levant 3, 1348 Louvain-la-Neuve, Belgium
{vrins, archambeau, verleysen}@dice.ucl.ac.be

Abstract. In most ICA algorithms, the separation performances are estimated through the evaluation of a *contrast function* Φ, used in the update rule of elements of the unmixing matrix. In particular situations, optimizing Φ does not lead to optimize the extraction of each source, one by one. However, in some applications, one can be interested in quantifying the extraction performances of a specific signal. In this paper, we emphasize that none of the usual Φ's could be directly applied, without further precautions, to evaluate the performances of the local separation.

1 Introduction

Blind Source Separation (BSS) consists in recovering m independent sources ($\mathbf{s} = [s_1, \ \ldots, s_m]^T$) only from $n \geq m$ mixtures of them (denoted by $\mathbf{x} = [x_1, \ldots, x_n]^T$). If $n > m$, these observations must be projected to an m-dimensional subspace. In the case where the mixture is linear, instantaneous and noiseless, we have: $\mathbf{x} = \mathbf{A}\mathbf{s}$. This problem can be solved by Independent Component Analysis (ICA), which allows to recover estimates of the sources: $\mathbf{y} = \hat{\mathbf{s}} = \mathbf{W}\mathbf{x} \simeq \mathbf{P}\mathbf{D}\mathbf{s}$, where $\mathbf{y}$ are the outputs of the algorithm, $\mathbf{P}$ and $\mathbf{D}$ are a permutation and a diagonal matrix, respectively. In ICA algorithms, sources are recovered when a *contrast function* Φ (CF) is maximized. In order to achieve this, update rules for $\mathbf{W}$ were derived. Indirectly, the evaluation of Φ at the final step of the algorithm can be seen as a measure of the *global* separation performances (GSP). Nevertheless, in some situations, maximizing the GSP does not imply that each source is separated in the best way, and one may be interested in extracting a particular signal as well as possible (for denoising applications, fetal electrocardiogram extraction, etc.). This is the reason why a *local* separation performance (LSP) estimator could be useful. How can we identify when a specific source $\hat{s}_i$ is as close as possible from the associated source s_i without using the original sources $\mathbf{s}$, which are -supposed to be- unknown ?

*M.V. is Research Associate of the Belgian National Funds for the Scientific Research.

ESANN'2004 proceedings - European Symposium on Artificial Neural Networks
Bruges (Belgium), 28-30 April 2004, d-side publi., ISBN 2-930307-04-8, pp. 211-216

In this paper, we recall the link between ICA, independence and nongaussianity; The most used CF's are recalled. Next, we discuss on the feasibility of using such classical global CF's as an estimator of LSP. Some simulations results are given in section 4, before a conclusion.

2 Independence and Nongaussianity

2.1 Independence as a GSP estimator

To derive CF's, some authors have used the main hypothesis on the s_i : they are mutually independent. This implies that the product of the marginal density functions $f_{s_i}(s_i)$ (pdf's) equals the joint density function $f_{\mathbf{s}}(\mathbf{s})$ (jpdf) of the sources. Below, we detail the CF's based on the cancellation of the higher-order cumulants and on mutual information.

2.1.1 Higher-orders cross-cumulants (JADE)

It is well-known that signals y_i and y_j are uncorrelated if their covariance is zero, implying that $E\{y_i y_j\} = E\{y_i\}E\{y_j\}$. Independence requires that $E\{g_i(y_i)g_j(y_j)\} = E\{g_i(y_i)\}E\{g_j(y_j)\}$ for any g_i non-linear function (cross-cumulants of all orders must be null). *Whitening* cancels the cross-cumulants up to two (uncorrelatedness), and JADE [1] diagonalizes the tensor of 4^{th} order cumulants $C_{ijkl}(y)$, which is in practice a good approximation of independence. The CF used in JADE is defined as: $\Phi_{Jade} \doteq -\sum_{ijkl \neq ijkk} C^2_{ijkl}(\mathbf{y})$.

2.1.2 Mutual Information (Infomax - Infomut)

The Infomax algorithm, derived by Bell and Sejnowski relies on the Linsker's principle [2]. Sources are recovered by maximizing the mutual information $I(\mathbf{y}, \mathbf{x})$ propagated in a network (parameterized by ω) of inputs $\mathbf{x}$ and outputs $\mathbf{y}$. Actually, the authors prove that this could be reached by the maximization of the joint entropy $h(\mathbf{y})$ between outputs: $\frac{\partial I(\mathbf{x},\mathbf{y})}{\partial \omega} = \frac{\partial h(\mathbf{y})}{\partial \omega}$. Applying directly this principle leads us inevitably to infinity by finding an arbitrarily large unmixing matrix $\mathbf{W}$ [3]. Thus we prefer maximizing $h\big(\mathbf{g}(\mathbf{y})\big)$ instead of $h(\mathbf{y})$, where $[\mathbf{g}(\mathbf{y})]_i = g_i(y_i)$. The g_i's are non-linear functions mapping $\mathbb{R} \mapsto [0, 1]$ and increasing monotonously (*e.g.* a sigmoid). It can be shown that the joint entropy is equal to the difference between the sum of the marginal ones and the mutual information [4]: $h\big(\mathbf{g}(\mathbf{y})\big) = \sum_{i=1}^{m} h\big(g_i(y_i)\big) - I\big(\mathbf{g}(\mathbf{y})\big)$.

In theory, it is clear that maximizing the joint entropy of $g(\mathbf{y})$ is not *equivalent* to minimizing the mutual information between the outputs $I\big(\mathbf{g}(\mathbf{y})\big)$. However, in practice (even there exists no proof of that), maximizing $\Phi_{Infomax} \doteq h\big(\mathbf{g}(\mathbf{y})\big)$ seems maximizing $\Phi_{Infomut} \doteq -I(\mathbf{g}(\mathbf{y}))$ [5] for supergaussian sources[1].

[1] An extended version of the algorithm (*Extended Infomax*) allows a good separation for subgaussian signals as well. Furthermore, The CF of Infomax is equivalent to the *Maximum Likelihood Approach* one [3].

ESANN'2004 proceedings - European Symposium on Artificial Neural Networks
Bruges (Belgium), 28-30 April 2004, d-side publi., ISBN 2-930307-04-8, pp. 211-216

2.2 Nongaussianity as a GSP estimator

The *maximum nongaussianity* approach is based on two major results from information theory [6]. First, the Central Limit Theorem (CLT) says that if u is a sum of n random independent variables ($n \to \infty$), each one having an arbitrary probability distribution, then $f_u(u)$ tends to a Gaussian function. In other words, it means that the pdf of a mixture is *closer to a Gaussian* than the pdf of each independent variable involved in the mixture. Secondly, the Maximum Differential Entropy of a Gaussian variable expresses the fact that a Gaussian variable x_G is the one (among all unbounded variables x) which has the highest differential entropy for a given variance ($\sigma_x^2 = \sigma_{x_G}^2$): $h(x) \leq h(x_G) = \frac{1}{2}\log(2\pi e)\sigma_x^2$, with equality if and only if x is Gaussian. Combining those two principles, we can say that in order to find one source, we have to find an output which minimizes entropy. In the following subsections, we give CF's to *measure* nongaussianity.

2.2.1 Kurtosis (MaxKurt)

The kurtosis $\kappa(y)$ ($\in [-2, \infty[$) of a variable y tells us if $f_y(y)$ is spikier (*supergaussian*: $\kappa(y) > 0$) or flatter (*subgaussian*: $\kappa(y) < 0$) than the Gaussian pdf ($\kappa(y) = 0$). The kurtosis is classically defined as: $\kappa(y) = E\{y^4\} - 3E\{y^2\}^2$. For almost all non-Gaussian pdf's, the kurtosis is strictly different from zero. For example, $|\kappa(y)|$ or $\kappa^2(y)$ could thus be used as a measure of nongausianity [6] of $f_y(y)$: $\Phi_{Kurt_1} \doteq \sum_{i=1}^m |\kappa(y_i)|$ or $\Phi_{Kurt_2} \doteq \sum_{i=1}^m \kappa(y_i)^2$.

2.2.2 (Neg)entropy (FastICA)

The Kullback-Leibler divergence (KLd, [4]) could be used to measure nongaussianity in order to derive a CF: $\Phi_{KL} \doteq \sum_{i=1}^m \int f_{y_i}(\zeta) \log \frac{f_{y_i}(\zeta)}{f_{x_G}(\zeta)} d\zeta$. Another criterion, equivalent to this latter CF, was derived: the *negentropy*, noted J. The negentropy of a variable y_k is defined as the difference between $h(x_G)$ of variance $\sigma_{x_G}^2 = \sigma_{y_k}^2$ and the differential entropy of y_k: $J(y_k) = h(x_G) - h(y_k) = \frac{1}{2}\log(2\pi e)\sigma_{x_G}^2 - h(y_k)$. Actually, the exact computation of differential entropy and KLd requires a high computational time, and the analytic expression of the pdf. Good approximations of J were derived which considerably decrease the computational time. FastICA [6] uses the following: $\hat{J}(y_k) \propto (E\{g(y_k)\} - E\{g(x_G)\})^2$, where g is a non-linear function, chosen according to the input signals $\mathbf{x}$. The aim of FastICA is the same as MaxKurt: drive far away the pdf of the m outputs (estimated sources) y_i from Gaussian ones. The CF Φ of FastICA is the previous approximation of negentropy applied to the output signals: $\Phi_{FastICA} \doteq \sum_{i=1}^m \hat{J}(y_i)$.

The *Minimum Marginal Entropy* algorithm relies on these principles and on the entropy power inequality [7]: $h(x+y) \geq \max\big(h(x), h(y)\big)$, where x and y are two independent variables.

3 Towards a LSP estimator ?

As explained in the introduction, it could be interesting to measure the extraction quality of each $\hat{s}_k$, one by one. Indeed, even if the GSP estimator indicate a failure

ESANN'2004 proceedings - European Symposium on Artificial Neural Networks
Bruges (Belgium), 28-30 April 2004, d-side publi., ISBN 2-930307-04-8, pp. 211-216

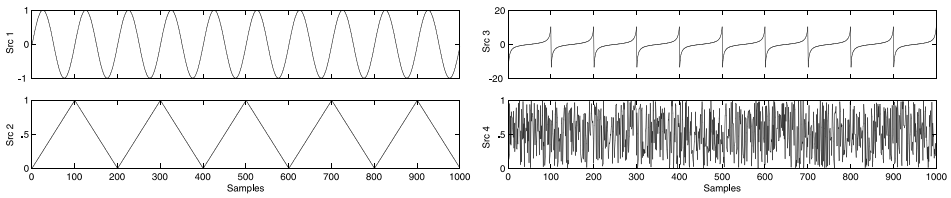

Figure 1: Four original slightly dependent sources.

of the global convergence, a specific source could be quite well recovered. A natural question is the following: *If the previous contrast functions are used to measure the GSP, could they also be used to estimate the quality of the separation of each source independently ?*

The CF's based on independence are estimated *between* outputs, and are thus not appropriate to estimate the LSP. On the other hand, algorithms based on nongaussianity are intrinsically different: this criterion is measured *on each signal*. The evaluation of each $\hat{J}(\hat{s}_k)$, $h(\hat{s}_k)$ or $\kappa(\hat{s}_k)$ seems thus to be a good - and easy - way to estimate the LSP of $\hat{s}_k$.

Nevertheless, the independence assumption on the s_i's is very critical. The violation of this hypothesis implies that a mixture of the s_i's is not necessarily more Gaussian than each one taken separately: the nongaussianity of $\hat{s}_k$ can no more be linked to the correlation of $\hat{s}_k$ with the associated source!

4 Simulations and results

In spite of a low dependence between sources, it happens in practice that the global separation is satisfying, but in this case the evaluation of the local one -using the same CF- could be irrelevant (see previous section). Here, we illustrate these problems when global convergence is nearly reached. In the following simulations, the output signals were ordered with respect to their correlation with the original sources $\rho(s_i, \hat{s}_i)$ $(1 \leq i \leq m = 4)$, in such a way that the evolution of the separation of each $\hat{s}_i$ has sense.

Two cases are analyzed: four random (independent) signals and four slightly dependent sources (non-zero correlations, see Figure 1) are mixed, the mixtures being polluted by uniform noise (resp. $SNR = 32$dB and $SNR = 60$dB, where SNR is defined as $SNR = 10 \log\left(\frac{\sigma^2_{signal}}{\sigma^2_{noise}}\right)$). Note that the absolute value of the correlation coefficients for the independent sources is bounded by 0.02 and for the slightly dependent sources, this maximum is 0.08.

We give results of simulations using Φ_{Jade}. We have plotted the evolution of each $\rho(s_i, \hat{s}_i)$, $h(\hat{s}_i)$ and $\kappa(\hat{s}_i)$ *versus* n the number of inputs x_i randomly chosen (among a set of 15 available noisy mixtures) projected by PCA on a 4-dimensional subspace, as explained in the introduction. As the mixtures are noisy, the information contained in $n > 4$ observations is useful to extract the original sources. For the random sources (Figure 2), $h(\hat{s}_i)$ and $\kappa(\hat{s}_i)$ give the same information as $\rho(s_i, \hat{s}_i)$. However, the LSP estimation of the low-dependent sources through the

ESANN'2004 proceedings - European Symposium on Artificial Neural Networks
Bruges (Belgium), 28-30 April 2004, d-side publi., ISBN 2-930307-04-8, pp. 211-216

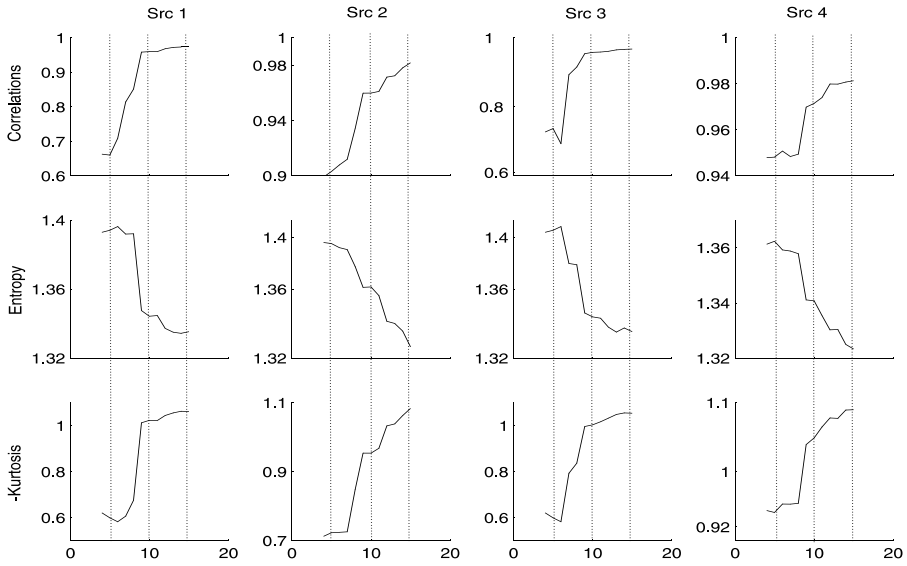

Figure 2: Four uniform independent sources mixed with additive noise on the mixtures: evolution of correlations (between orignal and associated estimated source), $h(\hat{s}_i)$ *vs* n and $-\kappa(\hat{s}_i)$ *vs* n the number of observed signals considered.

entropy or the kurtosis is not reliable, as shown the comparison with the correlation curves (see Figure 3, columns 2 and 3).

5 Conclusion

When the assumptions on the model are respected, the global separation leads to the optimization of the extraction of each source. However, in real-world applications, the model of ICA is rarely fully respected. For this reason, it could be interesting to identify when a specific source is extracted in the best way, especially in the particular case where the hypothesis of statistical independence between the sources is violated (*e.g.* how blindly chose $n = 10$ -the optimal value of n for the extraction of the third source- in Figure 3 ?). We have shown that in such cases, the usual CF's in ICA are not able to estimate perfectly the quality of the extraction of each source. Note that in many applications, the measure of the local separation performance should be done on a specific signal (*i.e.* the *desired* source). This identification requires an *a priori* knowledge on it (pdf, temporal structure, ...). Other problems occur when convergence failed: if we observe between step t (where it is assumed that $\hat{s}_k^t \simeq s_k$) and step $t+1$ that $h(\hat{s}_k^t) \geq h(\hat{s}_k^{t+1})$, does it means that s_k is better recovered at step $t+1$? Not necessarily: $\hat{s}_k^{t+1}$ could become a mixture of two low-entropic sources implying that $h(\hat{s}_k^{t+1}) \geq h(y^*)$, with $y^* = \alpha s_p + \beta s_q$, $k \neq p \neq q$ and even for $\sigma^2_{\hat{s}_k} = \sigma^2_{y^*}$!

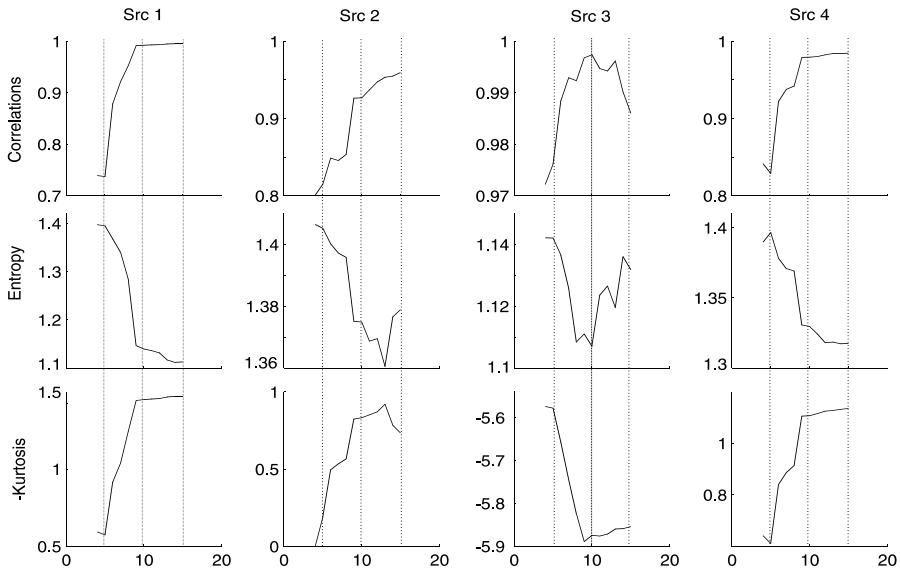

Figure 3: Four original slightly dependent sources (shown on Fig. 1) mixed with additive noise on the mixtures: evolution of correlations (between orignal and associated estimated source), $h(\hat{s}_i)$ *vs* n and $-\kappa(\hat{s}_i)$ *vs* n the number of observed signals considered.

Despite these problems, it must be stressed that the contrast functions used in the previous section seem to be very convenient (under constraint that the sources are mutually independent), even if the mixtures are noisy. Indeed, each correlation curve $\rho(y_i)$ (the ideal criterion of source extraction, but not blind), is directly linked to the evolution of these contrast functions $\Phi(y_i)$.

References

[1] J.-F. Cardoso. Source separation using higher order moments. *Proceedings of the IEEE, Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP'89)*, pages 2109–2112, 1989, Glasgow, England.

[2] A. J. Bell and T. J. Sejnowski. An information-maximisation approach to blind separation and blind deconvolution. *Neural Computation*, 7(6):1129–1159, 1995.

[3] J.-F. Cardoso. Infomax and maximum likelihood for blind source separation. *IEEE Signal Processing Letters*, 4(4):112–114, April 1997.

[4] T. M. Cover and J. A. Thomas. *Elements of information theory.* Wiley and sons, 1991.

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

[6] A. Hyvärinen and al. *Independent component analysis.* John Willey ans Sons, Inc., New York, 2001.

[7] S. Cruces, A. Cichocki, and S. Amari. The minimum entropy and cumulants based contrast functions for blind source extraction. In J. Mira and A. Prieto, editors, *IWANN'01, LNCS 2085*, pages 786–793. Springer-Verlag 2001, 2001.