

The minimum entropy and cumulants based contrast functions for blind source extraction

Sergio CRUCES¹, Andrzej CICHOCKI², and Shun-ichi AMARI²

¹ Signal Processing Group, Camino Descubrimientos, 41092-Seville, SPAIN,
sergio@cica.es,

WWW home page: <http://viento.us.es/~sergio>

² Brain Science Institute, RIKEN, 2-1 Hirosawa, Wako-shi, Saitama, 351-0198
JAPAN {cia,amari}@brain.riken.go.jp,

WWW home page: <http://www.bsp.brain.riken.go.jp/>

Abstract. In this paper we address the problem of blind source extraction of a subset of “interesting” independent sources from a linear convolutive or instantaneous mixture. The interesting sources are those which are independent and, in a certain sense, are sparse and far away from Gaussianity. We show that in the low-noise limit and when none of the desired sources is Gaussian, the minimum entropy and cumulants based approaches can solve the problem. These criteria, with roots in Blind Deconvolution and in Projection Pursuit, will be proposed here for the simultaneous blind extraction of a group of independent sources. Then, we suggest simple algorithms which, working on the Stiefel manifold perform maximization of the proposed contrast functions.

1 Introduction

In the recent years the criteria for blind source separation (BSS) of independent and non-Gaussian sources have been an active field of research [15, 8, 4, 16, 6, 3]. The blind separation problem considers the case where a certain number of sources is linearly combined to give the observations and only from these observations we try to recover all the possible sources.

More recently, blind source extraction, the problem of recovering or extraction of only a subset of the most “interesting” independent sources from the linear mixture has gained increasing attention due to its practical applications in communications [22] and in biomedical engineering [20].

The first approaches on blind source extraction can be traced back to the work of Donoho and many others [13, 21] in the single input single output blind deconvolution problem where one is interested in blindly recovering a non-minimum phase filtered non-Gaussian signal which is independent and identically distributed (i.i.d). Independently, but nearly at the same time, another field coined under the name of Projection Pursuit [17] was interested in the spatial counterpart of this problem. The motivation here was to find “interesting” low-dimensional projections of high-dimensional data sets. Both fields arrived to the common conclusion that the pursuit of the most non-Gaussian projection allows

to extract one of the independent sources. More recently, but with the same aim, other related criteria and algorithms has been developed in the context of blind source separation and independent component analysis [12, 7, 18, 22, 19]. However, in most cases the developed methods allow only the extraction of the sources one by one or all of them simultaneously.

The main objective of this paper is to extend the concepts and approaches proposed by Amari *et al.* [1, 2] and Cruces *et al.* [11] to the case of the simultaneous extraction of several “interesting sparse” sources without the need of employing a deflation procedure. The organization of the paper is as follows. Section 2 discuss mixture signal model and some basic but useful results. Section 3 presents the minimum entropy contrast for blind source extraction while section 4 presents a contrast function based on higher order cumulants. Section 5 extends some of the previous results to the problem of blind deconvolution and section 6 presents family of the algorithms that can perform the optimization of the proposed contrast functions. Finally, section 7 summarizes the conclusions of the paper.

2 Signal model and basic results

Let us consider a vector of N unknown statistically independent source signals $\mathbf{s} = [s_1, \dots, s_N]^T$ with zero mean and normalized covariance $Cov(\mathbf{s}) = \mathbf{R}_{ss} = E[\mathbf{s}\mathbf{s}^T] = \mathbf{I}_N$. These signals are linearly mixed by unknown nonsingular mixing matrix \mathbf{A} as

$$\mathbf{x} = \mathbf{A}\mathbf{s}, \quad (1)$$

where \mathbf{x} is the available vector of observations (sensor signals).

Without loss of generality we assume that the unknown mixing matrix \mathbf{A} is orthogonal. Note, that the orthogonality of the mixing matrix ($\mathbf{A}\mathbf{A}^T = \mathbf{I}_N$) can be always enforced by simply performing pre-whitening of the original observations.

In order to extract $E < N$ sources, the observations will be further processed by an $E \times N$ semi-orthogonal separating matrix \mathbf{U} satisfying relationship $\mathbf{U}\mathbf{U}^T = \mathbf{I}_E$ which yields to the outputs vector (or estimated sources)

$$\mathbf{y} = \mathbf{U}\mathbf{x} = \mathbf{G}\mathbf{s} \quad (2)$$

where $\mathbf{G} = \mathbf{U}\mathbf{A}$ is also the semi-orthogonal $E \times N$ global transfer matrix of mixing-separating system.

It is well known [8, 5] that for Gaussian stationary sources, independence is not a sufficient condition to separate them. This result is a direct consequence of the Darmois-Skitovich theorem which is presented below.

Theorem 1 (Darmois-Skitovich). *Let $\mathbf{s} = [s_1[n], s_2[n], \dots, s_N[n]]^T$ be an N -dimensional random vector ($N \geq 2$) whose components are mutually independent and consider the two outputs obtained from a linear combination of these sources*

$$\begin{aligned} y_1[n] &= G_{1,1}s_1[n] + G_{1,2}s_2[n] + \dots + G_{1,N}s_N[n] \\ y_2[n] &= G_{2,1}s_1[n] + G_{2,2}s_2[n] + \dots + G_{2,N}s_N[n] \end{aligned} \quad (3)$$

Assuming that $y_1[n]$ and $y_2[n]$ are independent, if for any index i $G_{1,i} \neq 0$ and $G_{2,i} \neq 0$ hold, then $s_i[n]$ will have a Gaussian distribution.

Thus, the blind source separation of $E = N$ sources can be identifiable, up to the arbitrary scaling and ordering indeterminacies, if and only if there is at most one Gaussian source in the mixture. This result is easily extended to the case of blind extraction for $E < N$ in the following corollary.

Corollary 1. *For a linear mixture of N stationary and independent sources, the extraction from the observations of any non-Gaussian subset with $E < N$ can be performed, up to the arbitrary ordering and scaling, if and only if, there are at most $N - E_{max}$ (where $E \leq E_{max}$) Gaussian sources in the mixture.*

Another useful result is the entropy power inequality [10] which provides a lower bound on the differential entropy of a sum of two random variables in terms of their individual differential entropies.

Theorem 2 (Entropy power inequality). *If a and b are independent continuous random variables then*

$$2^{2h(a+b)} \geq 2^{2h(a)} + 2^{2h(b)} \quad (4)$$

where $h(a) = -\int p_a(a) \log p_a(a) da$ is the differential entropy of a continuous random variable a with probability density function $p_a(a)$.

3 The minimum entropy contrast for blind extraction of groups of sources

The central limit theorem tells us that the linear mixture of N independent signals will become asymptotically Gaussian (as N grows towards ∞). This has suggested that the pursuit of non-Gaussianity can be a separating process and, since the Gaussian distribution for a fixed variance maximizes the entropy, one can intuitively think that by minimizing the entropy of the outputs while keeping the variance constant leads to separation of sources.

Projection Pursuit approaches [17] have shown that, indeed, such idea is correct and several algorithms for the blind extraction of a single source and for the simultaneous blind source separation of the whole set of sources from the mixture have been developed [14]. The following theorem will extend these results to consider of the simultaneous extraction of a specific subset of $E \leq E_{max} \leq N$ sources from the linear mixture.

Theorem 3. *Let us assume that the sources can be ordered by increasing value of the differential entropy (or uncertainty) as*

$$h(s_1) \leq \dots \leq h(s_E) < h(s_{E+1}) \leq \dots \leq h(s_N) . \quad (5)$$

If s_E is not Gaussian distributed, then the following objective function

$$\Psi_{ME}(\mathbf{y}) = -\sum_{i=1}^E h(y_i) \quad \text{subject to} \quad \text{Cov}(\mathbf{y}) = \mathbf{I}_E \quad (6)$$

is a contrast function whose global maxima correspond to sources with the smallest value of entropy (i.e., the least uncertain sources of the mixture), i.e., $\mathbf{y} = [s_1, \dots, s_E]^T$ up to an arbitrary reordering or permutation.

Proof. The proof of this theorem is based on the entropy power inequality. First note that $\text{Cov}(\mathbf{y}) = \mathbf{G}\mathbf{G}^T$. Taking into account that $y_i = \sum_{j=1}^N G_{ij}s_j$, and applying theorem 2 we can see that

$$2^{2h(y_i)} \geq \sum_{j=1}^N 2^{2h(G_{ij}s_j)} \quad (7)$$

$$= \sum_{j=1}^N G_{ij}^2 2^{2h(s_j)} \quad (8)$$

$$= [\mathbf{V}]_{ii} \quad (9)$$

where $\mathbf{V} = \mathbf{G}\mathbf{A}_1\mathbf{G}^T$ and \mathbf{A}_1 is the diagonal matrix with elements $[\mathbf{A}_1]_{ii} = 2^{2h(s_i)}$.

Now, expressing marginal entropies of the outputs in terms of the diagonal elements of matrix \mathbf{V} , and taking into account that $\sum_{j=1}^N G_{ij}^2 = 1$, we can rewrite the sum of marginal entropies as

$$\sum_{i=1}^E h(y_i) = \sum_{i=1}^E \frac{1}{2} \log(V_{ii}) \quad (10)$$

$$\geq \text{trace}\{\mathbf{G}\mathbf{A}\mathbf{G}^T\} \quad (11)$$

where \mathbf{A} is a diagonal matrix with elements $[\mathbf{A}]_{ii} = h(s_i)$, and the inequality between (10) and (11) follows from the concavity of the logarithm. Note that, from the semi-orthogonality of \mathbf{G} , the equality is attained if and only if \mathbf{V} is diagonal. Then, as a consequence of the Poincaré's separation theorem of matrix algebra we obtain that

$$\min_{\text{Cov}(\mathbf{y})=\mathbf{I}_E} \sum_{i=1}^E h(y_i) = \min_{\mathbf{G}\mathbf{G}^T=\mathbf{I}_E} \text{trace}\{\mathbf{G}\mathbf{A}\mathbf{G}^T\} = \sum_{i=1}^E h(s_i) \quad (12)$$

Thus the global maxima of the contrast function $\Psi_{ME}(\mathbf{y})$ are achieved for such matrices \mathbf{G} of which rows are orthogonal vectors that span the same subspace of the eigenvectors associated with the E lowest eigenvalues of \mathbf{A} , what enforces that $G_{ij} = 0 \forall j \geq E$.

The necessary and sufficient condition for the equality between (10) and (11) (\mathbf{V} being diagonal) will trivially hold for any subset of decorrelated Gaussian sources. However, from the given hypotheses and due to ordering the sources according increasing entropy, the first E sources are non-Gaussian and have the lowest possibly entropy. Thus, the condition for \mathbf{V} to be a diagonal matrix implies that each row of \mathbf{G} should contain only one nonzero element $+1$ or -1 . Then, \mathbf{G} is any generalized permutation matrix which extracts the E sources with the lowest individual differential entropy, i.e., \mathbf{G} can be reduced by row permutations to the form $[\mathbf{I}_E, \mathbf{0}]$. \square

The following lemma brings us another interpretation of the minimum entropy contrast which reveals, in a more explicit fashion, the pursuit of non-Gaussianity.

Lemma 1. *Let be $\mathbf{g} = [g_1, \dots, g_E]^T$ a vector of normalized Gaussian random variables. The maximization of*

$$\Psi_{ME}(\mathbf{y}) = - \sum_{i=1}^E h(y_i) \quad \text{subject to} \quad \text{Cov}(\mathbf{y}) = \mathbf{I}_E \quad (13)$$

is equivalent to the maximization of the following quasi-distance of the outputs marginal distributions from the Gaussianity

$$\sum_{i=1}^E KL(p_{y_i} || p_{g_i}) \quad \text{subject to} \quad \text{Cov}(\mathbf{g}) = \text{Cov}(\mathbf{y}) \quad (14)$$

where $KL(p_{y_i} || p_{g_i}) = \int p_{y_i} \log \frac{p_{y_i}}{p_{g_i}} dy_i$ denotes the Kullback-Leibler divergence (relative entropy) between the involved (marginal) probability density functions.

The proof of the lemma easily follows from the decomposition of the Kullback-Leibler divergence in terms of the individual differential entropies of the outputs $KL(p_{y_i} || p_{g_i}) = h(p_{g_i}) - h(p_{y_i})$ when the constraint $\text{Cov}(\mathbf{g}) = \text{Cov}(\mathbf{y})$ applies.

4 Contrast function based on higher order cumulants

In the preceding section we have observed how the negative of the differential entropy of the outputs can give us an index of non-Gaussianity. However, it is well known that this is not the only function we can use for this task. In particular, the absolute value of the higher order autocumulants $C_s^r = \text{Cum}(\underbrace{s, \dots, s}_{\times r})$ with $r > 2$

can also measure the departure from Gaussianity. This result is summarized in the following theorem.

Theorem 4. *When the sources signals are normalized such that $\text{cov}(\mathbf{s}) = \mathbf{I}_N$ and for a decreasing order arrangement of them with regard to the absolute values of their $(1 + \beta)$ -order autocumulants*

$$|C_{s_1}^{1+\beta}| \geq \dots \geq |C_{s_E}^{1+\beta}| > |C_{s_{E+1}}^{1+\beta}| \geq \dots \geq |C_{s_N}^{1+\beta}|, \quad (15)$$

if $|C_{s_E}^{1+\beta}| \neq 0$, the following function

$$\Psi_{Cum}(\mathbf{y}) = \frac{1}{1 + \beta} \sum_{i=1}^E |C_{y_i}^{1+\beta}| \quad \text{subject to} \quad \text{cov}(\mathbf{y}) = \mathbf{I}_E \quad (16)$$

is a contrast whose global maxima lead to the extraction of the first E sources with the largest $(1 + \beta)$ -order autocumulants.

Proof. We will only sketch the proof which is parallel to that of the minimum entropy contrast. After some straightforward simplifications and subject to the semi-orthogonality of \mathbf{G} we obtain that

$$\sum_{i=1}^E |C_{y_i}^{1+\beta}| \leq \sum_{j=1}^N |C_{s_j}^{1+\beta}| \sum_{i=1}^E G_{ij}^2 \quad (17)$$

Then, we can apply the Poincaré's separation theorem to observe that the maximum of (17) is given by

$$\max_{\text{cov}(\mathbf{y})=\mathbf{I}_E} \Psi_{Cum}(\mathbf{y}) = \frac{1}{1+\beta} \sum_{j=1}^E |C_{s_j}^{1+\beta}| \quad (18)$$

Here, the bound is only attained for that matrices \mathbf{G} that can be reduced by row permutations to the form $[\mathbf{I}_E, \mathbf{0}]$, i.e., \mathbf{G} is the extraction matrix of the first E sources. \square

5 Multichannel blind deconvolution

The previous results for blind source extraction can be also extended to solve the Multichannel Blind Deconvolution problem as is shown in the following theorem.

Theorem 5. *Consider N source random processes which are mutually independent and temporally i.i.d. (independent and identically distributed) and with cross-covariance $\text{Cov}_{\mathbf{s},\mathbf{s}}[n] = \delta[n] \mathbf{I}_E$. Let the function $\psi(\cdot) = \{\psi_{ME}(\cdot), \psi_{Cum}(\cdot)\}$ where $\psi_{ME}(\cdot) = \frac{1}{2} \log(2\pi e) - h(\cdot)$ and $\psi_{Cum}(\cdot) = |C_{(\cdot)}^{1+\beta}|$ with $\beta > 1$. Assuming that the sources can be ordered by decreasing value of the function $\psi(\cdot)$ of their random variables as*

$$\psi(s_1[n]) \geq \dots \geq \psi(s_E[n]) > \psi(s_{E+1}[n]) \geq \dots \geq \psi(s_N[n]) , \quad (19)$$

and if $\psi(s_E[n]) > 0$, the function

$$\Psi(\mathbf{y}[n]) = \sum_{i=1}^E \psi(y_i[n]) \quad \text{subject to} \quad \text{Cov}_{\mathbf{y},\mathbf{y}}[n] = \delta[n] \mathbf{I}_E \quad (20)$$

is a contrast whose global maxima are found at the extraction of the, in a certain sense, less Gaussian sources of the mixture, i.e., $\mathbf{y}[n] = [s_1[n], \dots, s_E[n]]^T$ up to a permutation and arbitrary individual delays.

It is interesting to note that theorem 5 includes several special cases already known in the literature. For $E = N = 1$ equation (20) includes the minimum entropy contrast for blind deconvolution (whose optimality has been analyzed by Donoho in [13]) and also includes the cumulants based contrast function proposed by Shalvi and Weinstein in [21]. For $E = N > 1$ equation (20) is the contrast for blind deconvolution proposed by Comon in [9]. Furthermore, Tugnait [22] and Inouye *et al.* [19] have analyzed similar cumulants based criteria to (20) in the case of $E = 1$ and $N > 1$.

6 Blind source extraction/deconvolution algorithms

When the observations are decorrelated ($Cov_{\mathbf{x},\mathbf{x}}[n] = \delta[n] \mathbf{I}_E$) and of zero mean, one of the possible methods to maximize the proposed contrast $\Psi(\mathbf{y}[n])$ is to employ the natural Riemannian gradient ascent in the Stiefel manifold [1] which preserves the outputs decorrelation constraint. The desired gradient, which is given by

$$\tilde{\nabla}_{\mathbf{U}[n]}\Psi = \nabla_{\mathbf{U}[n]}\Psi(\mathbf{y}[n]) - \mathbf{U}[n] * (\nabla_{\mathbf{U}[n]}\Psi(\mathbf{y}[n]))^T * \mathbf{U}[n], \quad (21)$$

where $*$ denotes the convolution operator, leads to Amari's gradient algorithm¹

$$\mathbf{U}^{(k+1)}[n] = \mathbf{U}^{(k)}[n] + \mu \left(\mathbf{R}_{\varphi,x}^{(k)}[n] - \mathbf{R}_{y,\varphi}^{(k)}[n] * \mathbf{U}^{(k)}[n] \right) \quad (22)$$

where $\mathbf{R}_{\varphi,x}^{(k)}[n] = E[\varphi(\mathbf{y}^{(k)}[l]) (\mathbf{x}[l-n])^T]$ and $\varphi(\mathbf{y}[n]) = [\frac{d\Psi}{dy_1[n]}, \dots, \frac{d\Psi}{dy_E[n]}]^T$. The exact expressions of $\varphi(\mathbf{y}[n])$ depend on the used criteria. When $\psi(\cdot) = \psi_{ME}(\cdot)$ approximations to these derivatives can be found in [8] and in [23] where the marginal p.d.f. of the outputs are truncated at low orders of the Edgeworth or Gram-Charlier expansions, respectively. By using the contrast function expressed by cumulants $\psi(\cdot) = \psi_{Cum}(\cdot)$, the learning algorithm takes a specific form

$$\mathbf{U}^{(k+1)}[n] = \mathbf{U}^{(k)}[n] + \mu \left(\mathbf{S}_y \mathbf{C}_{y,x}^{\beta,1}[n] - \mathbf{C}_{y,y}^{1,\beta}[n] \mathbf{S}_y * \mathbf{U}^{(k)}[n] \right) \quad (23)$$

where \mathbf{S}_y is the diagonal matrix with entries $[\mathbf{S}_y]_{ii} = \text{sign}(\mathbf{C}_{y_i}^{1+\beta}[0])$ and $\mathbf{C}_{y,x}^{\beta,1}[n]$ is the $(1+\beta)$ -order cross-cumulant matrix whose elements are given by $[\mathbf{C}_{y,x}^{\beta,1}[n]]_{ij} = Cum(y_i[l], \dots, y_i[l], x_j[l-n])$, similarly $[\mathbf{C}_{y,y}^{1,\beta}[n]]_{ij} = Cum(y_i[l], y_j[l-n], \dots, y_j[l-n])$. Note that the stochastic versions of these algorithms can be easily obtained.

7 Conclusions

The minimum entropy and cumulants based approaches for blind source extraction/deconvolution of temporal i.i.d. sources has been extended to allow the simultaneous recovery of groups with the least Gaussian sources (in a certain specific sense) from a linear mixture. The connections of this approach with other criteria have been shown and Amari's gradient algorithm has been suggested for the optimization of the proposed contrast functions.

References

1. S. Amari, "Natural gradient learning for over- and under-complete bases in ICA," *Neural Computation*, vol. 11, pp. 1875–1883, 1999.

¹ Algorithm (22) was originally proposed in [1, 2] to solve the blind separation problem for memoryless mixtures.

2. S. Amari, A. Cichocki, Y.Y. Yang, *Blind Signal Separation and Extraction*, chapter 3 in "Unsupervised Adaptive Filtering" Volume I, edited by S. Haykin, Wiley, 2000.
3. S. Amari, A. Cichocki, "Adaptive blind signal processing – neural network approaches," *Proceedings of the IEEE*, vol. 86, no. 10, pp. 2026–2048, 1998.
4. A.J. Bell, T.J. Sejnowski, "Blind separation and blind deconvolution: An information-theoretic approach," in *ICASSP*, 1995.
5. X-R. Cao, R-W. Liu, "General Approach to Blind Source Separation," in *IEEE Transactions on Signal Processing*, vol. 44(3), pp. 562-571, March 1996.
6. J. F. Cardoso, "Blind signal separation: Statistical principles," *Proceedings of the IEEE*, vol. 86, no. 10, pp. 2009–2025, 1998.
7. A. Cichocki, R. Thwonmas, S. Amari, "Sequential blind signal extraction in order specified by stochastic properties," *Electronics Letters*, vol. 33, no. 1, pp. 64–65, 1997.
8. P. Comon, "Independent component analysis, a new concept?," *Signal Processing*, vol. 3, no. 36, pp. 287–314, 1994.
9. P. Comon, "Contrasts for Multichannel Blind Deconvolution," *IEEE Signal Processing Letters*, vol. 3, no. 7, pp. 209–211, 1996.
10. T. M. Cover, J. A. Thomas, *Elements of Information Theory*, Wiley series in telecommunications. John Wiley, 1991.
11. S. Cruces, A. Cichocki, L. Castedo, "Blind source extraction in gaussian noise," in *proc. of the 2nd International Workshop on Independent Component Analysis and Blind Signal Separation (ICA '2000), Helsinki, Finland, June 2000*, pp. 63–68.
12. N. Delfosse, P. Loubaton, "Adaptive blind separation of independent sources: A deflation approach," *Signal Processing*, vol. 45, pp. 59–83, 1995.
13. D. Donoho, *On Minimum Entropy Deconvolution*, Applied Time Series Analysis II, D. F. Findley Editor, Academic Press, New York, 1981.
14. M. Girolami, C. Fyfe, *Negentropy and kurtosis as projection pursuit indices provide generalized ICA algorithms*, pp. 752–763, Boston, MA: MIT Press, 1996.
15. C. Jutten, J. Herault, "Blind separation of sources, part i: An adaptive algorithm based on neuromimetic architecture," *Signal Processing*, vol. 24, pp. 1–10, 1991.
16. J. Karhunen, E. Oja, L. Wang, R. Vigario, J. Koutsensalo, "A class of neural networks for independent component analysis," *IEEE Transactions on Neural Networks*, vol. 8, no. 3, pp. 486–503, May 97.
17. P.J. Huber, "Projection pursuit," *Annals of Statistics*, vol. 13, pp. 435–525, 1985.
18. A. Hyvarinen, E. Oja, "A fast fixed-point algorithm for independent component analysis," *Neural Computation*, vol. 9, pp. 1483–1492, 1997.
19. Y. Inouye, T. Sato, "Iterative algorithms based on multistage criteria for blind deconvolution," *IEEE Transactions on Signal Processing*, vol. 47, no. 6, pp. 1759–1764, June 1999.
20. T-P. Jung S. Makeig, A. Bell, T.J. Sejnowski, *Independent Component Analysis of Electroencephalographic Data*, vol. 8, pp. 145–151, M. Mozer *et al.*, Cambridge, MA: MIT Press, 1996.
21. O. Shalvi, E. Weinstein, "New criteria for blind deconvolution of nonminimum phase systems (channels)," *IEEE Transactions on Information Theory*, vol. 36, no. 2, pp. 312–321, 1990.
22. J. K. Tugnait, "Identification and deconvolution of multichannel linear non-Gaussian processes using higher order statistics and inverse filter criteria," *IEEE Transactions on Signal Processing*, vol. 45, no. 3, pp. 658–672, 1997.
23. H. H. Yang, S. Amari, "Adaptive on-line learning algorithms for blind source separation – maximum entropy and minimum mutual information," *Neural Computation*, 1997.