INDEPENDENT COMPONENT ANALYSIS BASED ON MARGINAL ENTROPY APPROXIMATIONS

Juan J. Murillo-Fuentes, Rafael Boloix-Tortosa, Susana Hornillo-Mellado
Area. de teoria de la Señal y Comunicaciones, Universidad de Sevilla, Spain[1]
murillo@us.es
Vicente Zarzoso[2]
Department of Electrical Engineering and Electronics, The University of Liverpool, UK
vicente@liv.ac.uk

ABSTRACT

The problem of blind source separation (BSS) can be solved through the statistical tool of independent component analysis (ICA). The present contribution reviews recent solutions to ICA contrasts based on the minimization of marginal entropy (ME). In the two-signal case, a novel estimator, so-called sinusoidal ICA (SICA), is obtained by approximating Comon's 4th-order cumulant based contrast function. Interestingly, SICA as well as analogous methods scattered across the literature are particular instances of a class of closed-form solutions gathered under the name of general weighted estimator (GWE). In the n-dimensional case, $n > 2$, these elementary estimators are applied over the input components in pairs, as in the Jacobi optimization (JO) technique for matrix diagonalization. The reduction of the computational burden of JO for ICA is addressed. Adaptive (on-line) versions are briefly considered as well. A simple simulation experiment illustrates the good performance of the approximate ME approach.

KEYWORDS: array signal processing, blind source separation, higher order statistics, independent component analysis, unsupervised learning.

1. INTRODUCTION

Blind separation of sources (BSS) consists of extracting a non-observable set of signals, the so-called sources, from another set of observable signals generated as mixtures of the sources. One of the approaches to achieve separation is to project the observations into a basis where the components are statistically independent. This is the aim of independent analysis (ICA) [1]. In its simplest (noiseless) form, the problem accepts the following matrix formulation. The entries of the *mixture* vector $\boldsymbol{x}(t) = [x_1(t), \ldots, x_m(t)]^T$ observed at the sensor output at time instant t, where $(\cdot)^T$ represents the transpose operator, are instantaneous linear combinations of a set of zero-mean unknown *source signals* $\boldsymbol{s}(t) = [s_1(t), \ldots, s_n(t)]^T$, $m \geq n$:

$$\boldsymbol{x}(t) = \boldsymbol{A}\boldsymbol{s}(t). \tag{1}$$

The $(m \times n)$ matrix $\boldsymbol{A}$, called *transfer* or *mixing matrix*, is also unknown. In general, if $\boldsymbol{s}(t)$ is a stationary ergodic random sequence, the mixing matrix is full column rank, the sources are statistically independent and at most one of them is Gaussian, it is possible to compute a separation matrix which extracts the source components up to, perhaps,

[1] Thanks Spanish government MCYT TIC2003-03781 for funding.

[2] Supported through a Post-doctoral Research Fellowship awarded by the Royal Academy of Engineering of the UK.

irrelevant permutation and scaling. Since the source amplitudes are immaterial, one may assume, without loss of generality, that the source variance is unity: $\mathrm{E}[\boldsymbol{s}\boldsymbol{s}^H] = \boldsymbol{I}$, where $\boldsymbol{I}$ denotes the identity matrix and symbol $(\cdot)^H$ represents the conjugate-transpose operator.

The separation is typically carried out in two steps. First, whitening or standardization (often carried out via principal component analysis, PCA) consists of projecting the observed vector on the signal subspace, second-order decorrelation and power normalization, yielding signals $\{z_i(t)\}_{i=1}^n$ such that $\mathrm{E}[\boldsymbol{z}\boldsymbol{z}^H] = \boldsymbol{I}$. As a result, the source and whitened vectors are necessarily related through a unitary transformation:

$$\boldsymbol{z}(t) = \boldsymbol{V}\boldsymbol{s}(t). \tag{2}$$

Thus, the problem reduces to the computation of matrix $\boldsymbol{V}$. In the second step, ICA aims to compute matrix $\boldsymbol{V}$ that makes the entries of the output vector $\boldsymbol{y}(t) = \boldsymbol{V}^H\boldsymbol{z}(t)$ as independent as possible. This process is able to retrieve the sources up to scalings and permutations, i.e., accomplishes the source separation [1]. Although many approaches have been proposed to date [2], this paper focuses on those minimizing contrast (or cost) functions based on the output differential entropy. We first deal with the two-dimensional case to later develop solutions for $n > 2$ dimensions based on Jacobi optimization (JO). The convenience of initializing the Jacobi iterations by computing the relevant statistics beforehand will be discussed. Finally, adaptive versions will also be outlined.

2. THE 2-DIMENSIONAL ME CONTRASTS

The minimization of statistical dependence of the projected observation results in, among other criteria, the minimization of the mutual information (MI) between outputs or, equivalently, of their marginal entropies (ME). For zero-mean unit-variance decorrelated mixtures, these approaches reduce to the computation of the unitary transformation that either cancels the cross-cumulants or maximizes the marginal cumulants [3, 4]. Under orthonormal transformations both approaches are tantamount. We will focus on ME minimization in the sequel [1, 4].

A contrast function $\phi(\boldsymbol{y})$ is a functional in the output components whose minimization yields the ICA solution [1]. If $\boldsymbol{s}$ has independent components then $\phi(\boldsymbol{s}) \leq \phi(\boldsymbol{A}\boldsymbol{s})$, $\forall \boldsymbol{A}$ non-singular, with equality iff $\boldsymbol{A}$ can be decomposed as the product of a permutation and an invertible diagonal matrix. Besides, $\phi(\boldsymbol{y})$ is invariant to permutation and scaling of the components in $\boldsymbol{y}$.

If the maximum likelihood (ML) contrast [5] is minimized under the whitening constraint for all possible distributions, we obtain the ME contrast [6, 7]

$$\phi^{\mathrm{ME}}(\boldsymbol{y}) = \sum_i \mathrm{H}[y_i] \tag{3}$$

where $\mathrm{H}[\cdot]$ denotes the differential entropy. Using the Edgeworth expansion [8] to approximate ϕ^{ME} by a function of 4th-order cumulants, we obtain [1, 7]

$$\phi_{24}^{\mathrm{ME}}(\boldsymbol{y}) = -\sum_i (C_{iiii}^{y})^2 \tag{4}$$

where $C_{iiii}^{y} = \mathrm{E}[y_i^4] - 3$ represents the 4th-order marginal cumulant (kurtosis) of y_i. This contrast is discriminant over the set of random vectors $\boldsymbol{y}$ with at most one non-kurtic component [1]. Functional ϕ_{24}^{ME} is an *orthogonal* contrast [4], for it assumes (2nd-order) decorrelated outputs and is thus to be optimized under the unitary constraint.

In the real-valued two-dimensional case, matrix $\boldsymbol{V} = \boldsymbol{V}(\theta)$ performs a planar rotation of angle θ. The source pair $\boldsymbol{s}(t) = [s_p(t),\, s_q(t)]^T$ can be expressed in polar coordinates as

$(r(t), \alpha(t))$. Accordingly, the whitened sensor-output pair $\boldsymbol{z}(t) = [z_p(t),\, z_q(t)]^T$ accepts the polar representation $(r(t), \beta(t))$, where, by virtue of eqn. (2), $\beta(t) = \alpha(t) + \theta$. Therefore, the BSS/ICA problem is reduced to the estimation of angular parameter θ.

In [1], the value of θ which minimizes the ME criterion (4) is obtained among the roots of a 4th-degree polynomial. However, certain simplifications are possible, leading to approximate closed-form solutions which avoid polynomial rooting. For instance, the ML solution of [9] is valid for sources with the same long-tailed distribution, the extended ML (EML) estimator of [10] can be used when the source kurtosis sum is not null, the alternative EML (AEML) estimator of [11] when the sources kurtoses are different, whereas the approximate ML (AML) approach of [12] combines the latter two expressions in a more robust estimator that may be used for any source distribution. Other methods of this type can be found in [13, 14, 15].

Interestingly, all these approximate closed-form estimators can be gathered under a general expression, the so called *general weighted estimator (GWE)* [16, 15], given by:

$$\hat{\theta}_{\mathrm{GWE}}(\omega_\gamma, \omega_\xi) = \frac{1}{4}\angle\big(\omega_\xi \omega_\gamma \xi_\gamma + (1-\omega_\xi)\xi_\eta\big), \qquad 0 < \omega_\xi < 1, \quad \omega_\gamma = \{\pm 1, \gamma\} \tag{5}$$

where

$$\xi_\gamma = \mathrm{E}[r^4(t)e^{j4\beta(t)}] \qquad \xi_\eta = \mathrm{E}^2[r^4(t)e^{j2\beta(t)}] \qquad \gamma = \mathrm{E}[r^4(t)] - 8 \tag{6}$$

are the *centroids* associated to the whitened outputs. In the above equations, $\angle(\cdot)$ supplies the principal value of its argument, and $j = \sqrt{-1}$. The sample estimates of (6) are used in practice.

As realized in [16, 15], approximations to the ME contrast minimization solution are found as particular cases of the GWE for different values of ω_ξ and ω_γ. In effect: $\hat{\theta}_{\mathrm{EML}} = \hat{\theta}_{\mathrm{GWE}}(\gamma,\, 1)$ [10]; $\hat{\theta}_{\mathrm{AEML}} = \hat{\theta}_{\mathrm{GWE}}(\cdot,\, 0)$ [11]; $\hat{\theta}_{\mathrm{AML}} = \hat{\theta}_{\mathrm{GWE}}(\gamma,\, 1/3)$ [12]; $\hat{\theta}_{\mathrm{MaSSFOC}} = \hat{\theta}_{\mathrm{GWE}}(\gamma,\, 1/2)$ [13]; or the estimators in [17, 7, 18, 9] as $\hat{\theta}_{\mathrm{GWE}}(\pm 1, 1)$. Also, it is shown in [14, 15] that a sinusoidal approximation to $\phi_{24}^{\mathrm{ME}}(\theta)$ in (4) yields the angle estimator

$$\hat{\theta}_{\mathrm{SICA}} = \hat{\theta}_{\mathrm{GWE}}(\gamma,\, 3/7) \tag{7}$$

which is consequently referred to as sinusoidal ICA (SICA).

The optimum values of $(\omega_\gamma,\, \omega_\xi)$ can be computed from certain source statistics [16], if these are available. Such weights are optimum in that they characterize the minimum (asymptotic) variance estimator of the GWE class. Although we have focused on real-valued mixtures, an expression for the weighted estimator also exists for complex mixtures [16]. Other closed-form formulae, comprising cumulant orders other than four, are studied in [19, 20].

3. N-DIMENSIONAL CASE

The GWE estimator in eqn. (5) was designed as an analytic solution to the two-dimensional ICA problem. Based on the Jacobi optimization (JO) technique for matrix diagonalization [21], Comon extended the solution of contrast (4) to the n-dimensional scenario, for an arbitrary source number $n > 2$. JO operates pairwise, minimizing the two-dimensional contrast for every whitened-signal pair in turn over several sweeps until convergence. For each pair of outputs, the unitary matrix is properly updated with the angle that minimizes the contrast. The algorithm can be summarized as follows [1, 22]:

Algorithm 1 n-*dimensional GWE using non-initialized Jacobi optimization: JO-GWE.*

1. *Whitening. Compute the whitened outputs* $\boldsymbol{z}$. *Initialize the separator output as* $\boldsymbol{y} = \boldsymbol{z}$ *and set sweep number* $c = 1$.

2. *Sweep c. For all $g = n(n-1)/2$ signal pairs ($1 \leq p < q \leq n$), do*

   (a) *Set $[z_p, z_q]^T = [y_p, y_q]^T$ and compute the Givens angle $\theta_{pq} = \theta_{\mathrm{GWE}}$ as in eqn. (5).*

   (b) *If $|\theta_{pq}| > \theta_{\min}$, update the solution with angle θ_{pq}.*

3. *End? If the number of sweeps c satisfies $c = 1 + \sqrt{n}$ or no angle $|\theta_{pq}| > \theta_{\min}$, stop. Otherwise sweep again over Step 2 with $c = c + 1$.*

In [7] the algorithms only stops when the set of g Givens rotations have all been updated by a value under a threshold $\theta_{\min}$. No limit is set on the number of sweeps c. The value $\theta_{\min}$ is selected so that rotations by a smaller angle are not considered as statistically significant. Typically, $\theta_{\min} = 10^{-2}/\sqrt{N}$ where N is the number of samples. In [1], the algorithm stops after going through Step 2 more than $1 + \sqrt{n}$ times. We use this threshold in our approach and we have set $\theta_{\min} = \pi/360$ (i.e., 0.5^o) to avoid useless computations.

In Steps 2.a–2.b we may estimate θ_{pq} and update the solution in two fashions [1, 22]:

- In the standard JO approach [1], at every Step 2.a the statistics (6) are computed from the current values of the output samples. Angle θ_{pq} is estimated and used to update (rotate) the outputs in Step 2.b until convergence.

- In the initialized JO (IJO) approach [22], the pertinent statistics are computed from the whitened outputs before entering the Jacobi-like iterations. The statistics are then 'rotated' according to the current estimate of $\boldsymbol{V}$ in order to calculate the centroids (6) at every Step 2.a of the algorithm. Matrix $\boldsymbol{V}$ is then updated with the estimated θ_{pq}. The output signals are extracted from the final value of $\boldsymbol{V}$ after convergence.

Reference [22] discusses which of these two methods should be used. As both approaches provide the same solution accuracy, the focus is on the computational burden. In the JO, the statistics are recalculated from the available samples at every step. Thus, the computational complexity of this method is mainly determined by the sample size. In the IJO approach, by contrast, the larger the number of sources, the larger the dimension of matrix $\boldsymbol{V}$ and the larger the number of operations spent in the rotations of the statistics. Hence, IJO's computations may soon be dominated by the number of sources. As a result, the approach to be used depends on a suitable compromise between the sample size and the number of sources. This decision leads to the Optimized JO (OJO) method [22].

Similarly, adaptive versions may also be designed according to two criteria. Both approaches are based on the adaptive learning of the statistics as follows:

- In the adaptive EML (ad-EML) method [23], the standard JO approach [1] is taken by adaptively updating the statistics (6). This involves learning these three values for every signal pair pq and every sweep c. Since the statistics at sweep c depend on their learning at previous sweeps, which, in turn, depends on the trajectories of the separator output, the method may not converge properly for a large number of sources.

- In the adaptive IJO (AIJO) approach [24], we just adaptively learn the statistics with every new sample. In parallel, every N new samples we may compute the solution $\boldsymbol{V}$ by means of the JO. Since the learning of the moments and the computation of the solution are decoupled, convergence is guaranteed if we succes in the adaptive learning of the statistics, provided that the JO also converges.

4. EXPERIMENTAL RESULTS

The comparative performance of the GWE is illustrated in this section with a simple simulation experiment. Estimator $\hat{\theta}_{\mathrm{SICA}}$ in (7) is used as GWE [15], and is compared to JADE [3, 25], the 4th-order based ME method by Comon [1], and the Fast-ICA [26, 27] algorithm. The same whitening procedure was implemented in all methods. The Fast-ICA [27] was executed with the parameters by default, including stabilization. We mix $n = 6$ zero-mean unit-variance signals with different distributions: uniform, Laplacian ($\mu = 0.1$), Rayleigh (B=1), exponential ($\mu = 1$), Gaussian, and lognormal ($\sigma = 0.1$). The mixing-matrix entries are uniformly distributed random numbers in the range $[-1, 1]$ and we assume the mixture to be noiseless.

We used OJO to decide between JO and IJO; under the conditions of this simulation, the latter method had the lowest computational cost. Fig. 1 shows the performance variation as a function of the sample size N. The interference-to-signal ratio (ISR) [28], averaged over 1000 independent experiments, is used as an objective separation index. As expected, SICA-OJO and Comon's ME present a similar performance. Though close to OJO's, the performance of JADE and Fast-ICA are slightly worse than that of the other two methods.

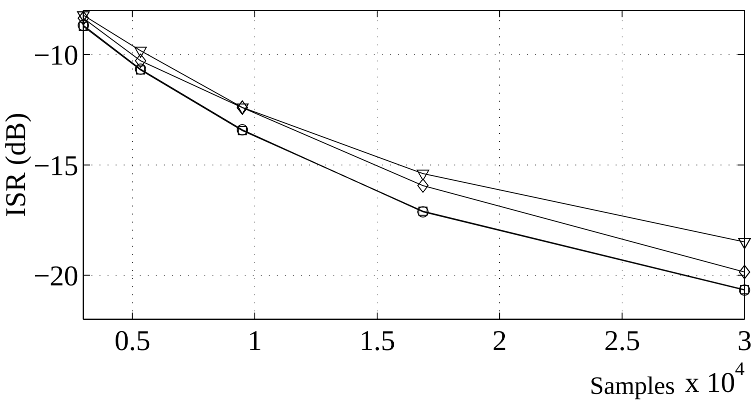

Figure 1: Mean ISR for SICA-OJO (□), Comon's ME (○), JADE (◇) and Fast-ICA (▽).

5. CONCLUSIONS

This paper has reviewed the approximate solutions to ICA contrasts relying on ME minimization. These solutions consist of formulae which provide, in the two-dimensional case, a direct estimate of the separation parameter by 'plugging in' certain statistics (centroids) of the whitened sensor output, without polynomial rooting or costly iterative optimization. The GWE provides a general expression for many of the 4th-order cumulant based approximations to the ME contrast encountered in the literature. In the n-dimensional case, $n > 2$, a Jacobi-like iterative procedure may be employed, either in the standard form originally proposed by Comon (JO) or suitably initialized (IJO), depending on their relative computational complexity. The performance of these approximate solutions are up to the mark of more elaborate ICA techniques at a fraction of the computational cost.

6. ACKNOWLEDGEMENTS

V. Zarzoso carried out part of this work while on leave at Laboratoire I3S, Sophia Antipolis, France. He gratefully acknowledges Prof. Pierre Comon for his kind hospitality.

7. REFERENCES

[1] P. Comon, "Independent component analysis, a new concept?" *Signal Processing*, vol. 36, no. 3, pp. 287–314, Apr 1994, http://www.i3s.unice.fr/~comon/matlab.html.

[2] A. Mansour and M. Kawamoto, "Ica papers classified according to their applications and performances," *IEICE Trans. Fundamentals*, vol. E86-A, no. 3, pp. 620–633, Mar 2003.

[3] J. F. Cardoso and A. Souloumiac, "Blind beamforming for non gaussian signals," *Proceedings IEE F*, vol. 140, no. 6, pp. 362–370, Dec 1993.

[4] J. F. Cardoso, "Blind signal separation: Statistical principles," *Proc. of the IEEE*, vol. 86, no. 10, pp. 2009–2025, Oct 1998.

[5] M. Gaeta and J. L. Lacoume, "Source separation without a priori knowledge: the maximum likelihood solution," in *Proc. EUSIPCO*, vol. V, 1990, pp. 621–624.

[6] D. Pham, "Blind separation of instantaneous mixture of sources via an independent component analysis," *IEEE Trans. on Signal Processing*, vol. 44, no. 11, pp. 2668–2779, 1996.

[7] J. F. Cardoso, "High-order contrasts for independent component analysis," *Neural Computation*, vol. 11, no. 1, pp. 157–192, Jan 1999.

[8] A. Stuart and J. K. Ord, *Kendall's Advanced Theory of Statistics*, 6th ed. London: Edward Arnold, 1994, vol. I.

[9] F. Harroy and J.-L. Lacoume, "Maximum likelihood estimators and cramer-rao bounds in source separation," *Signal Processing*, vol. 55, no. 2, pp. 167–177, 1996.

[10] V. Zarzoso and A. K. Nandi, "Blind separation of independent sources for virtually any source probability density function," *IEEE Trans. on Signal Processing*, vol. 47, no. 9, pp. 2419–2432, 1999.

[11] V. Zarzoso, A. K. Nandi, F. Herrmann, and J. Millet-Roig, "Combined estimation scheme for blind source separation with arbitrary source PDFs," *IEE Electronic Letters*, vol. 37, no. 2, pp. 132–133, 2001.

[12] M. Ghogho, A. Swami, and T. Durrani, "Approximate maximum likelihood blind source separation with arbitrary source pdfs," in *Proc. SSAP'00*, Pocono Manor Inn, Pennsylvania, Aug 2000.

[13] F. Herrmann and A. Nandi, "Blind separation of linear instantaneous mixture using close forms estimators," *Signal Processing*, vol. 81, pp. 1537–1556, 2001.

[14] J. Murillo-Fuentes and F. González-Serrano, "Independent component analysis with sinusoidal fouth-order contrast," in *Proc. ICASSP'01*, vol. V, Salt Lake City, USA, May 2001, pp. 2785–2788.

[15] J. J. Murillo-Fuentes and F. J. Gonzlez-Serrano, "A sinusoidal contrast function for the blind separation of statistically independent sources," *IEEE Trans. on Signal Processing. Accepted for publication*, 2004.

[16] V. Zarzoso, F. Herrmann, and K. Nandi, "Weighted closed-form estimators for blind source separation," in *Proc. SSP01*, vol. 4, Singapore, Aug 2001, pp. 2141–2146.

[17] P. Comon and E. Moreau, "Improved contrast dedicated to blind separation in communications," in *IEEE International Conference on Acoustics, Speach and Signal Processing*, vol. V, Munich, Germany, 1997, pp. 3453–3456.

[18] E. Moreau and O. Macchi, "Higher order contrast for self-adaptive source separation," *International Journal of Adaptive Control and Signal Processing*, vol. 10, no. 1, pp. 19–46, Jan 1996.

[19] V. Zarzoso and A. K. Nandi, "Closed-form estimators for blind separation of sources — part I: real mixtures," *Wireless Personal Communications*, vol. 21, no. 1, pp. 5–28, Apr. 2002.

[20] ——, "Closed-form estimators for blind separation of sources — part II: complex mixtures," *Wireless Personal Communications*, vol. 21, no. 1, pp. 29–48, Apr. 2002.

[21] G. H. Golub and C. F. Van Loan, *Matrix Computations*, 3rd ed. Baltimore, MD: The John Hopkins University Press, 1996.

[22] J. J. Murillo-Fuentes, R. Boloix, and F. J. González-Serrano, "Initialized jacobi optimization in independent component analysis," in *Proc. ICA2003*, Nara, Japan, Apr 2003, pp. 1053–1058.

[23] V. Zarzoso and A. K. Nandi, "Adaptive blind source separation for virtually any source probabiliy density function," *IEEE Trans. on Signal Processing*, vol. 48, no. 2, pp. 477–488, 2000.

[24] J. J. Murillo-Fuentes, R. Boloix, and F. J. González-Serrano, "Adaptive initialized jacobi optimization in independent component analysis," in *Proc. ICA2003*, Nara, Japan, Apr 2003, pp. 1065–1070.

[25] J. F. Cardoso, "Web page: Jade for real-valued data," WWW, ftp://sig.enst.fr/pub/jfc/Algo/Jade/jadeR.m.

[26] A. Hyvärinen, "Fast and robust fixed-point algorithms for independent component analysis," *IEEE Trans. on Neural Networks*, vol. 10, no. 3, pp. 626–634, 1999.

[27] J. Hurri, H. Gävert, J. Särelä, and A. Hyvärinen, "FastICA version 2.1, web page," 1998, http://www.cis.hut.fi/projects/ica/fastica /index.shtml.

[28] V. Zarzoso and A. K. Nandi, "Blind source separation," in *Blind Estimation Using Higher-Order Statistics*, A. K. Nandi, Ed. Boston, MA: Kluwer Academic Publishers, 1999, ch. 4, pp. 167–252.