Blind Source Separation with the Weighted Mixed Statistics Algorithm

M. Klajman and A. G. Constantinides ¤
Communications and Signal Processing Group
Imperial College of Science, Technology and Medicine
London SW7 2BT

ABSTRACT

Depending on the character of the signals, most Blind Source Separation algorithms exploit either the second order or fourth order statistics of the signals. In this paper we present a novel weighted mixed statistics algorithm which performs signi¯cantly better than the single type statistics algorithms. As the algorithm is a generalisation of the single type statistics algorithm, it requires less prior information. Estimating functions are used in order to derive the weights. We provide simulations to show the enhanced performance of the weighted mixed statistics approach, even in mixtures were the signals contain no temporal information.

1 Introduction

Blind Source Separation (BSS) is concerned with recovering the original unknown sources from their observed mixture. The algorithm operates blindly in the sense that, except for statistical independence, no a-priori information about either the sources or the transmission medium is known. Most BSS algorithms operate in two steps. First the observed data is whitened. Then the transformation that relates the unknown sources to the decorrelated mixture is found as a pure rotation, since the sources and the decorrelated observations are white vectors. In algebraic BSS methods, this rotation is found by unitarily diagonalising a set of matrices. If the source signals have di®erent spectral contents or are non stationary, these matrices can be constructed from Second Order Statistics (SOS) only [1]. If the sources are white, one must resort to Higher Order Statistics (HOS) [2]. Note however, that HOS based separation is only possible if the sources are independent and non Gaussian. Thus, a certain amount of prior knowledge about the sources is necessary in order to choose the appropriate algorithm. In this paper we propose to generalise the BSS solution. The mixing matrix is found by jointly diagonalising a weighted combination of non-zero-lag whitened covariance matrices and the eigenmatrices of the fourth order cumulants. Hence, no prior information about the sources is necessary at all. Moreover, using a weighted combination of di®erent statistics, separation can be achieved in cases where algorithms relying on single order statistics will fail. The $JADE_{TD}$ algorithm [3], uses a combination of several non-zero-lag covariance matrices and the fourth order cross-cumulants in order to separate a mixture of three white Laplacian and three coloured Gaussian signals. However, the issue of weighting was not addressed. Simulations will show that suitable convex weights can improve performance signi¯cantly. Merging statistics ad hoc is adequate enough for separating the mixture described in [3]. The HOS will identify the Laplacian sources and virtually ignore the Gaussian ones, the SOS concentrate on the coloured Gaussian sources and do not take any notice of the Laplacians. Thus, each of the statistics identify di®erent signals and $JADE_{TD}$ is therefore able to separate all the sources completely. As a matter of fact, the sources could be separated more e±ciently by using two di®erent types of extraction techniques independently, such as FastICA, which will extract the Laplacians, and a type of power method, which will identify the Gaussians. In [4], the authors suggest a type of weighted combined statistics approach based on the Kolmogoro® complexity. Our algorithm makes direct use of the second order covariance matrices and the fourth order eigenmatrices.

2 Problem Formulation

The instantaneous noiseless BSS problem, with the assumption of an equal number of sources and sensors, can be described mathematically as follows:

$$x(t) = As(t) \tag{1}$$

In this context, the vector $s(t) = [s_1(t); :::; s_N(t)]^T$ contains the original sources, the vector $x(t) = [x_1(t); :::; x_N(t)]^T$ contains the array output sampled at time t and A is the $N £ N$ mixing matrix or transfer function between the sources and sensors. In this paper, only real sources will be considered. For the sake

¤This work was carried out as part of Technology Group TG10 of the MOD Corporate Research Program, supported by the Defence Evaluation Research Agency.

of simplicity, we will assume that the mixing matrix A is orthogonal. This is by no means a restriction as any arbitrary mixing matrix can be factored as $A = W^{-1}U$, where W and U are a whitening and orthogonal matrix respectively. The whitener can be found by using standard principal component analysis techniques. An implicit assumption in BSS is that the sources have unity power. Separation is achieved if a vector y can be found so that

$$y(t) = Bx(t) = PDs(t) \tag{2}$$

where B is the unmixing matrix, P is a permutation matrix and D is a diagonal matrix. Hence, we can estimate the sources up to their order and their power.

3 Finding the Mixing Matrix

3.1 Second Order Statistics

Let $R_x(\tau)$ and $R_s(\tau)$ be the covariance matrices of the mixture and the sources, then:

$$R_x(\tau) = AR_s(\tau)A^T \tag{3}$$

As the sources are independent, $R_s(\tau)$ will be diagonal at all lags τ. Hence, equation (3) shows that A can be found by performing a unitary eigendecomposition of $R_x(\tau)$. The eigenvectors obtained will be the columns of the mixing matrix and the eigenvalues a permutation of the autocorrelation of the sources at lag τ. Note that $R_x(0)$ can not be used as $R_x(0) = AR_s(0)A^T = AA^T = I$ and does not provide any additional information to estimate the mixing matrix. In theory, any covariance matrix at non-zero lag is sufficient to estimate the mixing matrix. In practice, however, it is useful to use a set of covariance matrices as this would enhance the statistical efficiency of the algorithm and prevent an unfortunate choice of lags [1].

3.2 Higher Order Statistics

The cumulant matrices of the observations are defined as:

$$\forall M \quad Q_x(M) : q_{ij} = \sum_{k,l} \mathrm{cum}\{x_i, x_j, x_k, x_l\} m_{kl} \tag{4}$$

where m_{kl} are the elements of an arbitrary matrix M. The cumulant matrix is simply a linear combination of two dimensional slices of a fourth order tensor. Due to the multilinearity and the additivity property of the cumulants, (4) can be written as:

$$\forall M \quad Q_x(M) = A\Lambda_M A^T \tag{5}$$

where $\Lambda_M = \mathrm{diag}\left(\kappa_1 a_1^T M a_1, \ldots, \kappa_N a_N^T M a_N\right)$, κ_p is the kurtosis of the pth source and a_p is the pth column of A. Equation (5) shows that the mixing matrix can be found as a unitary diagonaliser of the cumulant matrices for any arbitrary matrix M. In practice, it is advisable to use the columns of the mixing matrix A for M, i.e. $M = a_p a_q^T$ for all $1 \le p, q \le N$. Substituting this in (5) will yield the so called eigenmatrices of x, which for all p = q become

$$Q_x(a_p a_p^T) = \kappa_p a_p a_p^T \tag{6}$$

and zero for all other indices. A method of constructing the eigenmatrices without any knowledge of the mixing matrix A, can be found in [2]. Jointly diagonalising the different eigenmatrices will yield an estimation of the mixing matrix.

3.3 Mixed Statistics

From (6) it is clear that the N eigenmatrices are of the same dimensions as the covariance matrices, namely $N \times N$. Moreover, ignoring noise and estimation errors, the mixing matrix found by diagonalising the eigenmatrices is precisely the one found by diagonalising the covariance matrices, up to a permutation of the columns. Let K be the number of non zero lag covariance matrices to be used and L = K + N. We then form L matrices G_l:

$$G_l = \begin{cases} \rho \times Q_x(a_p a_p^T) & \text{for } 1 \le l \le N \\ (1-\rho) \times R_x(\tau_{l-N}) & \text{for } N < l \le L \end{cases} \tag{7}$$

The mixing matrix is found by jointly diagonalising the matrices G_l with $l = 1, \ldots, L$ using the approximate joint diagonalisation algorithm proposed in the appendix of [1] and [2]. Thus, the mixing matrix is estimated on basis of the HOS and SOS. The JADE_{TD} algorithm suggested in [3] is actually a special case of our algorithm, where $\rho = 0.5$. Note that a large value for ρ means that more emphasis is put in the HOS, whereas a small value for ρ means that the SOS are considered more reliable. The next step is to determine ρ.

4 Decision Rule

Devising a decision rule that will assign the different weights is a difficult task as it must rely on some kind of performance criterion. Finding a performance criterion that does not use any prior information about the mixing matrix or the sources, is on itself not a trivial task. Traditionally, the fourth order cross-cumulants or second order cross-cumulants are used. If the sources are successfully separated, these should be close to zero. But these cannot be used here, as they are closely related to the respective objective functions of the JADE and the SOBI algorithm. Clearly, the performance criterion based on the fourth order cumulant will be biased towards the JADE algorithm, while a criterion based on the covariance matrices will favour the SOBI algorithm. It is crucial to find a measure of separation that does not, or at least not entirely, rely on the second or fourth order statistics. An ideal choice would be Herault's and Jutten's approximate independence test suggested

in [5] as it relies on odd order moment only. However, the test is only applicable to sources that have even distributions. Many neural network type blind source separation algorithm make use of estimating functions. One of the most used estimating function is:

$$F(y;B) = I - f(y)y^T \qquad (8)$$

where $F(\cdot)$ is a $N \times N$ matrix function, the vector y contains the estimated source signals, I denotes the identity matrix and $f(\cdot)$ is a nonlinear function [6]. The estimating function becomes zero as B approaches the separating matrix. Expanding $f(\cdot)$ as a Taylor series and taking the expectancy, (8) can be written elementwise as:

$$E\{F(y;B)\}_{ij} = -f_1 E\{y_i y_j\} - \ldots - f_k E\{y_i^k y_j\} - \ldots \quad \text{for } i \neq j \qquad (9)$$

where $E\{\cdot\}$ is the expectation and $\{D\}_{ij}$ denotes the ijth element of the matrix D. When y_i and y_j are independent, this becomes:

$$E\{F(y;B)\}_{ij} = -f_1 E\{y_i\}E\{y_j\} - \ldots - f_k E\{y_i^k\}E\{y_j\} - \ldots \quad \text{for } i \neq j \qquad (10)$$

As the source signals are assumed to be zero mean, $E\{y_j\} = 0$ and at independence $E\{F(y;B)\}$ becomes zero. If either $E\{f(y)\}$ or $E\{g(y)\}$ is zero then:

$$F(y;B) = I - f(y)g(y)^T \qquad (11)$$

where $g(y)$ is a nonlinear function different from $f(y)$. Using (11) or (8), we can develop a criterion that tests the independence between two components:

$$h(y_i;y_j) = f(y_i)g(y_j) \qquad (12)$$

where y_i denotes the ith estimated source signal and $f(\cdot)$ a nonlinear function. Depending on the nonlinearities and the source distributions, $g(y)$ is either a nonlinear function, as in (11), or the identity function as in (8). Independence is restored when $h(y_i;y_j) = 0$.

5 Simulations

The MSfαg signal is a discrete i.i.d. signal that takes its values in the set $\{-1;0;\alpha\}$ with respective probabilities $\pi_1 = 1/(1+\alpha^2)$; $(\alpha-1)/\alpha$; $1/\alpha + \alpha^2$. We first demonstrate the advantage of weighting. Three different signals, MSf2:1g, MSf2:3g and MSf2:8g, of sample length 3000 are generated. Temporal information is introduced by passing the source signals through an AR(2) filter with complex conjugate poles at $\exp(\pm 0{:}16j)$, $\exp(\pm 0{:}17j)$ and $\exp(\pm 0{:}165j)$. As the spectra of the signals are very similar, SOS based separation is very unlikely to succeed. The estimated cumulants of the sources are 0:33, $-0{:}11$ and 0:39. From (6) it is evident that the eigenmatrices are a function of the source kurtosis. As the kurtosis are small, the eigenmatrices will not provide enough information to estimate the mixing matrix. The three sources are mixed by a 3×3 matrix with elements picked from a Gaussian distribution. Next we form 100 sets of weighted covariance matrices and eigenmatrices using (7) with ω ranging from zero to one. An estimation of the mixing matrix is then found by applying the joint diagonalisation algorithm to each set individually. The performance is measured by the PI(P) index, given by: $PI(P) = \sum_{i=1}^{N}\left(\sum_{j=1}^{N}\frac{|p_{ij}|}{\max_k(|p_{ik}|)} - 1\right) + \sum_{j=1}^{N}\left(\sum_{i=1}^{N}\frac{|p_{ij}|}{\max_k(|p_{kj}|)} - 1\right)$, where $\max_k(|p_{ik}|)$ denotes the maximum per row and $\max_k(|p_{kj}|)$ the maximum per column and $P = BA$, where B is the separating matrix. The PI(P) index measures the distance of the matrix P to a scaled permutation matrix. Figure 1 shows the performance versus the weight of different statistics.

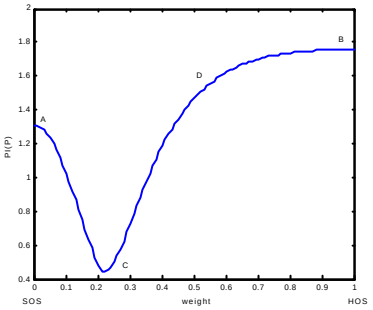

Figure 1: Quality of separation versus the weight of the statistics for three filtered MSfαg sources

From figure 1, we see that correct weighting is crucial for optimal separation. The JADE algorithm (B) achieves a PI(P) of 1:76, for SOBI (A), the PI(P) measures 1:30. Bracketed letters refer to the same letters in the figure. At the optimal weight (C), the PI(P) is 0:44, hence a big improvement in performance. Note that the JADE_{TD} (D) algorithm, which implicitly assumes a weighting of 0:5, performs not much better than the JADE (B) algorithm, and actually worse then SOBI (A). We next demonstrate the use of a decision function. Two signals of a 1000 sample length are created using MSf2:1g and MSf2:4g. The signals are then mixed by a 2×2 matrix with elements picked from a Gaussian distribution. Again a graph is constructed that shows the PI(P) versus the weights, as in figure 1.

Figure 2 shows that using only one type of statistics (point A and B) does not separate the sources sufficiently. However, using the combined statistics approach with $\omega = 0{:}14$ (C), unmixes the sources com-

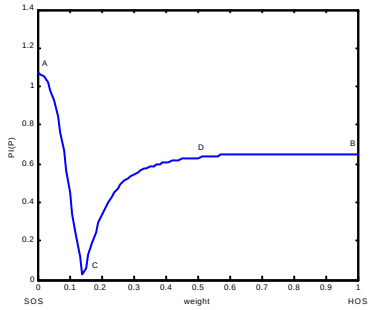

Figure 2: Quality of separation versus the weight of the statistics for two unfiltered MSfg sources

pletely. A PI(P) of 0:03 is obtained. The JADE$_{TD}$ approach (D) does not obtain in this case a better separation than the ordinary JADE algorithm (B). We next demonstrate the use of the independence test given in (12). The estimated source signals y_1 and y_2 were reconstructed from the different unmixing matrices. Then the function $h = f(y_1)\,g(y_2)$ was calculated for each weight. We have performed two experiments, one with $h_1 = \tan^{-1}(y_1) \pounds y_2$ and one with $h_2 = \tan^{-1}(y_1) \pounds \sin(y_2)$. We expect the former to perform better, as neither $E\{f(y_1)\}$ nor $E\{g(y_2)\}$ are zero.

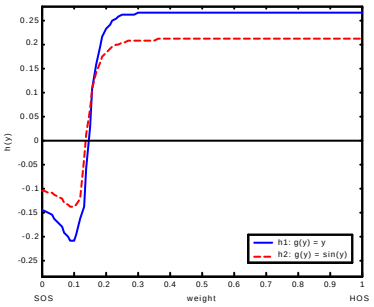

Figure 3: h_1 (solid) and h_2 (dashed) versus the weight of the statistics for two unfiltered MSfg sources.

Figure 3 shows that the points where $h = 0$ are very close to the points where separation is optimal. Recall that optimal separation is achieved when $\rho = 0{:}14$. Close inspection shows that h_1 and h_2 are zero at $\rho = 0{:}1464$ and $\rho = 0{:}1375$ respectively. At these points PI(P) = 0:045 and PI(P) = 0:06. Hence by using the weighted statistics algorithm with a decision function we have improved performance by at least one order in magnitude! And this without any additional prior information. An important note is of order here. One might ask how it is possible to use SOS as there is no temporal information in the sources. Due to noise and estimation errors, it is impossible to find the exact mixing matrix. Using only one type of statistics, or even a fixed combination of different statistics, will yield one possible estimation of the mixing matrix. A weighted combination however, will give an infinite amount of possible estimated mixing matrices. Based on an independent criterion, the decision rule then selects the optimal weighting and hence the most suitable matrix

6 Conclusion

The benefits of using a weighted combination of different types of statistics have been discussed. A decision rule that assigns the different weights was then derived. Simulations have shown the enhanced performance of the new algorithm. The example also confirmed that advantage may be gained from using combined statistics even if the sources do not contain any temporal information. Implementation aspects are beyond the scope of this paper. It suffices here to say that the independence criterion can be formulated as a constraint. The problem can then be solved using Lagrange programming neural networks. For more information about the use of a constraint in BSS, the reader is referred to [7].

References

[1] A. Belouchrani, K. Abed-Meraim, J. Cardoso, and E. Moulines, "A blind source separation technique using second order statistics," IEEE Transactions on Signal Processing, vol. 45, no. 2, pp. 434{444, 1997.

[2] J. F. Cardoso and A. Soulomiac, "Blind beamforming for non-gaussian signals," IEE PROCEEDINGS-F, vol. 140, pp. 362 { 370, December 1993.

[3] P. P. K. R. Muller and A. Ziehe, "JADE TD; combining higher order statistics and temporal information for blind source separation (with noise)," in ICA-99, pp. 87 { 92, 1999.

[4] A. Hyvarinen, "Complexity pursuit: Combining nongassianity and autocorrelations for signal separation," ICA - 2000, pp. 175 { 180, 2000.

[5] C. Jutten and J. Herault, "Blind separation of sources, part i: An adaptive algorithm based on neuromimietic architecture," Signal Processing, vol. 24, pp. 1 { 10, 1991.

[6] S. I. Amari and A. Cichocki, "Adaptive blind signal processing - neural network approaches," Proceedings of the IEEE, vol. 86, pp. 2026 { 2048, October 1998.

[7] M. Klajman and A. G. Constantinides, "A combined statistics cost function for blind and semi - blind source separation," in Proceedings of ICA 2001, 2001.