PERFORMANCE AND IMPLEMENTATION OF INVARIANT SOURCE SEPARATION ALGORITHMS

Jean-François Cardoso

CNRS / ENST / GdR TdSI.
46 rue Barrault, 75634 Paris, France.
cardoso@sig.enst.fr http://sig.enst.fr/~cardoso/stuff.html

ABSTRACT

This paper focuses on the *equivariant nature* of source separation : the unknown parameter of source separation is an invertible *matrix i.e.* it belongs to a multiplicative group. In this instance, inference theory calls for 'equivariant' estimation. This paper discusses some consequences of equivariance with respect to implementation and performance of source separation algorithms.

1. SOURCE SEPARATION

Source separation is receiving increasing attention in both signal processing and neural network literature since the seminal work of Jutten and Hérault [1]. The model of source separation is that of n statistically independent signals whose m (possibly noisy) linear combinations are observed; the problem consists in recovering the original signals from their mixture. The 'blind' qualification refers to the coefficients of the mixture: no *a priori* information is assumed to be available about them. This feature makes the blind approach extremely versatile because it does not rely on modeling the underlying physical phenomena.

This paper focuses on the *equivariant nature* of source separation : the unknown parameter of source separation is an invertible *matrix i.e.* it belongs to a multiplicative group. In this instance, inference theory calls for 'equivariant' estimation. This paper discusses some consequences of equivariance with respect to implementation and performance of source separation algorithms.

1.1. Model and assumptions

In this paper, we will consider the simplest source separation model where any additive noise can be neglected and the number of mixtures is equal to the number of sources. The signal model then is that of a n-dimensional time series $\mathbf{x}_t$ in the form :

$$\mathbf{x}_t = A\mathbf{s}_t \qquad t = 1, 2, \cdots$$

where $\mathbf{x}_t$ and $\mathbf{s}_t$ are $n \times 1$ vectors and A is a $n \times n$ matrix. The components of $\mathbf{s}_t$ are often termed 'source signals' and matrix A is the 'mixing matrix'.

Assumptions. For the purpose of source separation, the following assumptions are made

- A1 Matrix A is invertible.
- A2 Source signals are ergodic zero-mean processes.
- A3 Source signals are statistically independent.
- A4 Source signals have unit variance.

Assumption 3 is the key ingredient for source separation. It is a strong statistical hypothesis but a physically very plausible one when the source signals originate from physically separated systems. Assumption 4 only is a *normalization convention* because the amplitude of each source signal can be incorporated into A. Assumptions 2, 3 and 4 imply that

$$R_s \stackrel{\text{def}}{=} \mathrm{E}\,[\,\mathbf{s}_t\mathbf{s}_t^{\mathrm{T}}] = I \tag{1}$$

It is important to realize that without additional information, the outputs of a separating matrix cannot be ordered because the ordering of the source signals is itself immaterial (conventional): source signals can be at best recovered up to a permutation and a change of sign. To simplify exposition, it is assumed throughout that this indetermination can be fixed in one way or another, possibly using a priori information. The issue of indetermination is addressed at length in [2].

1.2. On-line/Off line source separation

Adaptive source separation consists in updating an $n \times n$ matrix B_t each time a new data sample is received:

$$B_{t+1} = B_t - \mu_t f(B_t, \mathbf{x}_t). \tag{2}$$

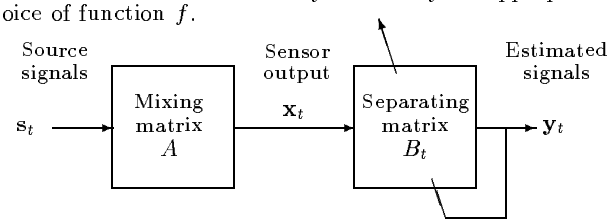
Function $f(\cdot,\cdot)$ defines the algorithm and $\{\mu_t\}$ is a sequence of positive numbers balancing convergence speed and accuracy. The output $\mathbf{y}_t$:

$$\mathbf{y}_t = B_t\mathbf{x}_t$$

should get as close as possible to the source signals, *i.e.* the *global system* C_t:

$$C_t \stackrel{\text{def}}{=} B_tA$$

should be driven to the identity matrix by an appropriate choice of function f.



Batch source separation consists in computing an estimate $\widehat{A}$ of A from of a batch X_T of T samples:

$$X_T \stackrel{\text{def}}{=} [\mathbf{x}_1, \ldots, \mathbf{x}_T]. \tag{3}$$

The estimate $\widehat{A}$ returned by any particular estimator is a function of the observations. This may be denoted as

$$\widehat{A} = \mathcal{A}(X_T)$$

where the mapping $\mathcal{A}$ corresponds to a given estimator. Source signals $\mathbf{s}_t$ are then estimated as $\hat{\mathbf{s}}_t = \mathbf{y}_t = (\widehat{A})^{-1}\mathbf{x}_t$. As for adaptive algorithms, we define a global system $\widehat{C}$ by

$$\widehat{C} = (\widehat{A})^{-1}A$$

which should be as close as possible to the identity matrix, since

$$\hat{\mathbf{s}}_t = (\widehat{A})^{-1}\mathbf{x}_t = (\widehat{A})^{-1}A\mathbf{s}_t = \widehat{C}\mathbf{s}_t.$$

1.3. Some approaches to source separation

In the seminal contributions by Jutten and Hérault, the separating matrix is determined in such a way that the output $\mathbf{y}$ has independent components. Typically, this is 'tested' by the decorrelation between each output and some non-linear functions of each other output. Thus, a stationary point of the algorithm is such that $\mathrm{E}y_i\phi_j(y_j) = 0$ for $i \neq j$. Similar stationarity conditions are found in many other adaptive algorithms : see for instance [3, 4, 5, 6]. Note that if the separating matrix is not constrained, there are n^2 unknown parameters to be determined and thus n^2 scalar stationarity conditions have to be satisfied. These conditions can be collected into:

$$\mathrm{E}H(\mathbf{y}) = 0 \tag{4}$$

where function $H(\cdot)$ is matrix-valued and each entry of the $n \times n$ matrix $H(\mathbf{y})$ depends on the coordinates of $\mathbf{y}$. An example of such a function is eq. (6).

Blind separation can also be based on the optimization of contrast functions. In the context of source separation, these were introduced by Comon [7] as functions of the distribution of $\mathbf{y}$ which are to be optimized under the whiteness constraint: $R_y = \mathrm{E}\mathbf{y}\mathbf{y}^{\mathrm{T}} = I$. Comon suggests minimizing the squared cross-cumulants of $\mathbf{y}$ (see also [8]). A similar (and asymptotically equivalent) contrast which can be efficiently optimized by a Jacobi-like algorithm, especially in the complex case, is described in [9].

In many cases, simpler orthogonal contrasts may be exhibited. For instance, if all the sources have a negative kurtosis, the minimization of $\sum_{i=1,n} \mathrm{E}|y_i|^4$ subject to $R_y = I$ is achieved only when B is a separating matrix. This is a strongly reminiscent of 4th-order objectives used in blind equalization [10]. When the 4th-order moments are estimated from a batch X_T of T samples, one can show that the resulting estimated signals $Y_T = [\mathbf{y}_1, \ldots, \mathbf{y}_T]$ satisfy :

$$\frac{1}{T}\sum_{t=1,T} H(\mathbf{y}_t) \;=\; 0 \tag{5}$$

where $H(\mathbf{y})$ is the $n \times n$ matrix with (i,j)-th entry given by

$$[H(\mathbf{y})]_{ij} = y_i^3 y_j - y_i y_j^3 + y_i y_j - \delta_{ij}. \tag{6}$$

This is a particularly interesting example, because it provides a connection between stationarity condition (4) of an adaptive algorithm and the 'estimating equation' (5) of a batch algorithm.

Other batch estimation techniques based on higher-order cumulants are used together with a prewhitening strategy in Tong and al. [2, 11]; fourth-order-only approaches are investigated in [12, 13]; purely second-order is possible if the sources have different spectra as investigated in [14, 15, 16, 2] and also in [17] in an adaptive implementation.

2. EQUIVARIANCE AND UNIFORM PERFORMANCE

The notion of *equivariance* (see for instance Lehman [18]) is of relevance when there is a 'natural' transformation group operating on the observations. In the case of source separation, the relevant transformation is left multiplication of the observed vector signals by invertible matrices. Let M denote an $n \times n$ invertible matrix and consider the operation of multiplying the batch of observations X_T by M:

$$MX_T = M\,[\mathbf{x}_1, \ldots, \mathbf{x}_T] \;=\; [M\mathbf{x}_1, \ldots, M\mathbf{x}_T].$$

Denote $S_T = [\mathbf{s}_1, \ldots, \mathbf{s}_T]$ the T realizations of the source signals that give rise to the observed batch X_T. We have $X_T = AS_T$ and we note that the effect of left multiplying X_T by M is equivalent to changing the mixing matrix A into MA since $MX_T = M(AS_T) = (MA)S_T$. Thus, a transformation on the data is equivalent to a transformation of the unknown parameter (namely: the mixing matrix A).

Hence, in the context of source separation, a 'natural' transformation group to be considered is left multiplication by the set all invertible $n \times n$ matrices. Section 2.1. explores the class of estimators which are consistent with these transformations.

2.1. Equivariant off-line separation

Equivariant estimators are consistent with a given transformation group whenever a transformation on the data induces a 'corresponding' transformation of the estimates. In the context of source separation, the precise definition of equivariant estimation is as follows (a more detailed exposition may be found in [19]). Recall that a particular estimator for source separation may be seen as an application $\mathcal{A}$ mapping any data set X_T to an estimate $\widehat{A}$ of A as denoted by eq. (3). An estimator $\mathcal{A}$ is said to be *equivariant* if it satisfies

$$\mathcal{A}(MX_T) = M\mathcal{A}(X_T) \tag{7}$$

for any invertible $n \times n$ matrix M.

The equivariance property has an immediate consequence on the performance of source separation. Consider applying an equivariant algorithm $\mathcal{A}$ to the mixture $X_T = AS_T$. One has

$$\widehat{A} = \mathcal{A}(X_T) = \mathcal{A}(AS_T) = A\mathcal{A}(S_T).$$

Hence we find that the global estimated system is

$$\widehat{C} \stackrel{\mathrm{def}}{=} (\widehat{A})^{-1}A = (\mathcal{A}(AS_T))^{-1}A = \mathcal{A}(S_T)^{-1}. \tag{8}$$

We arrive at the simple but crucial result that the global system $\widehat{C}$ estimated by an equivariant algorithm is $\mathcal{A}(S_T)^{-1}$, *i.e.* it does *not* depend on the mixing matrix but only on the particular realization S_T of the source signals. Thus, in terms of signal separation, the performance of an equivariant algorithm *does not depend at all on the mixing matrix*. We call this property *uniform performance* of batch equivariant estimators.

2.2. Equivariant on-line separation

There is no standard definition of equivariance in the case of adaptive algorithms. Hence, we adopt a definition which is convenient in the context of source separation by adapting the crucial property (8) of off-line equivariant estimators. According to this property, the global system estimated by an off-line equivariant estimator depends only on the particular realization of the sources.

Accordingly, an on-line separation algorithm is said to be equivariant when the updating rule of B_t is such that the global system $C_t = B_t A$ is updated in $C_{t+1} = B_{t+1}A$, where C_{t+1} *depends only on the previous value* C_t *and on the current realization of the source signal* $\mathbf{s}_t$.

The key point here is that, by mere definition, the behavior of an adaptive equivariant algorithm does not depend on the particular mixing matrix A. Thus, its performance (stability conditions, convergence speed, residual error and its tuning (choice of adaptation steps, design of the function H (see below)) can be studied and optimized independently of the mixture. Actually, it suffices to study the algorithm for $A = I$ *i.e.* when there is no actual mixture. This is one of the benefit of the uniform performance property.

3. IMPLEMENTATION

3.1. Adaptive implementations

It is not difficult to see that an adaptive algorithm in the form (2) is equivariant if and only if it can be recast as

$$B_{t+1} = B_t - \mu_t H(\mathbf{y}_t) B_t \tag{9}$$

where $H(\mathbf{y})$ is an $n \times n$ matrix-valued function of $\mathbf{y}$. Indeed, right multiplication of eq. (9) by matrix A yields the evolution equation of the global system $C_t = B_t A$:

$$C_{t+1} = C_t - \mu_t H(C_t \mathbf{s}_t) C_t \tag{10}$$

where we have used $\mathbf{y}_t = B_t \mathbf{x}_t = B_t A \mathbf{s}_t = C_t \mathbf{s}_t$. This last equation (10) shows that the global system evolves independently of the particular value of the mixing matrix.

Adaptation rule (9) is termed a 'serial update', because it reads equivalently $B_{t+1} = (I - \mu_t H(\mathbf{y}_t)) B_t$. This latter form evidences that B_t is updated by 'plugging' matrix $I - \mu_t H(\mathbf{y}_t)$ at the *output* of the current system B_t to get the updated system B_{t+1}. In this sense, serial updating is the general class of matrix updating which is consistent with equivariance since i) the transformation group relevant to source separation is left matrix multiplication and ii) system B_t is serially updated in (9) by *left multiplication* by matrix $I - \mu_t H(\mathbf{y}_t)$, not depending on A.

3.2. Batch (off-line) implementation

Estimating equations in the form (5) can be solved in an iterative fashion quite similar to the adaptive solution of section 3.1.. Consider the following algorithm

1. Initialize $\mathbf{y}_t = \mathbf{x}_t$ for $t = 1, \ldots, T$.
2. $\widehat{H} := T^{-1} \sum_{t=1}^{T} H(\mathbf{y}_t)$
3. $\mathbf{y}_t := (I - \mu \widehat{H})\, \mathbf{y}_t$ for $t = 1, \ldots, T$.
4. Stop if $\|\widehat{H}\|$ is 'small enough', otherwise go to step 2.

This is obviously the block counterpart of the 'serial update' algorithm (9). It stops when eq. (5) is verified. Note that it works by updating the signals themselves, without updating a separating matrix or without the need to explicitly construct an estimate of the mixing matrix A. However, it is easily seen that the implicit estimate of A obtained by this algorithm is equivariant, indeed.

This particular technique for solving (5) may be further improved by resorting to Newton-like algorithms as described in [20] where specific approximations to a true Newton algorithm are also discussed.

3.3. The estimating equation

Both the batch and the adaptive algorithms described above depend for their successful implementation of a specific vector-to-matrix mapping $\mathbf{y} \rightarrow H(\mathbf{y})$. The adaptive algorithm is a stochastic solver of equation $\mathrm{E}H(\mathbf{y}) = 0$ while the batch algorithm solves the same equation with the expectation replaced by a sample average in eq (5).

Both the batch algorithm and the adaptive algorithm are equivariant and thus enjoy the uniform performance property. It is desirable to have uniform performance, but it is even better to have uniformly *good* performance. This depends on choosing an appropriate function $H(\mathbf{y})$. A good choice implies that a separating matrix actually is an attractor of the algorithms and that the residual error is as small as possible. We discuss below two classes of H functions.

Symmetric forms. A function H appropriate for source separation is defined in eq. (6). It is a particular instance of the class of function of the form :

$$[H(\mathbf{y})]_{ij} = \phi_i(y_i) y_j - y_i \phi_j(y_j) + y_i y_j - \delta_{ij} \qquad 1 \le i, j \le n. \tag{11}$$

where for each i, function ϕ_i operates on y_i, the i-th output of the separating matrix.

Note that the stationary condition is $\mathrm{E}H(\mathbf{y}) = 0$ and that this matrix equation can be decomposed into its symmetric and skew-symmetric parts. The symmetric part of the equation implies that $\mathrm{E}y_i y_j = \delta_{ij}$ for all i and j. In other words, it implies that the output vector $\mathbf{y}$ have uncorrelated components with unit variance. The functions ϕ are to be taken non linear in order provide additional constraints. A typical example is the cubic functions appearing in eq. (6). This symmetric form actually stems from optimizing contrast function under a decorrelation constraint. See [19, 21] for more details.

By design, a H function like (11) satisfies $\mathrm{E}H(\mathbf{s}) = 0$ so that $B = A^{-1}$ is a stationary point of the adaptive algorithm. The *local asymptotic stability* of this stationary point depends on the choice of the ϕ_i functions in relation to the distribution of the sources. The asymptotic theory shows that the local stability is guaranteed provided that the

$$\text{Stability condition:} \quad \kappa_i + \kappa_j > 0 \quad \text{for} \quad 1 \le i, j \le n \tag{12}$$

holds. We have defined the following moments :

$$\kappa_i \stackrel{\text{def}}{=} \mathrm{E}\left[\phi_i'(s_i) - s_i \phi_i(s_i)\right] \tag{13}$$

and we note that for cubic functions : $\phi_i(y) = y^3$, the moment κ_i is (the opposite of) the 4th-order cumulant of the i-th source (recall that source signals are normalized to $\mathrm{E}s_i^2 = 1$).

Non symmetric forms. A different class of H functions are obtained from statistical considerations. Assume that the i-th source signal is i.i.d. with a differentiable probability density function p_i and define the 'score function' $\psi_i = -p_i'/p_i$. Then the maximum likelihood estimate of A is such that eq. (5) is satisfied with H defined as

$$[H(\mathbf{y})]_{ij} = \psi_i(y_i) y_j - \delta_{ij} \qquad 1 \le i, j \le n. \tag{14}$$

Because of the optimality properties of maximum likelihood estimation, a very good behavior is expected from such a function if the model holds (i.i.d. signals with the assumed distributions).

A generalization of this form of H function is as above, by using other non-linear functions. See the excellent paper [6] for a thorough investigation of this class of H functions.

4. ASYMPTOTIC PERFORMANCE

To get a simple performance characterization, we consider an asymptotic analysis in the case of i.i.d. source signals. Since the proportion of the q-th signal contained in the estimate of the p-th signal is given by the (p, q)-th entry of the global system, the asymptotic performance in terms of source separation will be characterized by the asymptotic variance of the off-diagonal terms of the global system. We thus get pair-wise rejection rates which correspond to inter-symbol interference (ISI) in the terminology of channel equalization.

Due to lack of space, we restrict ourselves to the case of identically distributed signals and identical non-linear functions: $\phi_1 = \ldots = \phi_n = \phi$. The asymptotic performance will depend on an extra non-linear moment :

$$\gamma \stackrel{\text{def}}{=} \mathrm{E}[\phi^2(s)] \; - \mathrm{E}^2[\phi(s)s] \tag{15}$$

where s is any of the s_i's. Note that γ is positive by the Cauchy-Schwartz inequality because of the normalization convention $\mathrm{E}s_i^2 = 1$.

4.1. Performance of batch algorithms

Performance of off-line algorithms in terms of interference rejection are characterized by the index

$$\mathrm{ISI}_{pq} \stackrel{\mathrm{def}}{=} \lim_{T\to\infty} T\ E|[\widehat{C}]_{pq}|^2 \quad \forall p \neq q. \tag{16}$$

The asymptotic analysis yields:

$$\mathrm{ISI} = \mathrm{ISI}_{pq} = \frac{1}{4} + \frac{\gamma}{2\kappa^2}. \tag{17}$$

The first term 1/4 may be shown to stem from the implicit decorrelation constraint. In fact, it may be shown [22] that any source separation technique requiring that the output is uncorrelated has its ISI lower bounded by this term. The second term depends on the distribution of the sources an may be minimized by an appropriate choice of of the non-linear function ϕ.

4.2. Performance of adaptive algorithms

Similarly, the asymptotic performance of an adaptive algorithm may be characterized by

$$\mathrm{ISI}_{pq} \stackrel{\mathrm{def}}{=} \lim_{\mu\to 0} \mu\ E_\mu|[C_t]_{pq}|^2 \quad \forall p \neq q. \tag{18}$$

where E_μ means that the expectation is taken with the adaptation step kept fixed at value μ and where it is understood that expectation is taken 'after convergence'. The rejection rates are [21, 23] :

$$\mathrm{ISI} = \mathrm{ISI}_{pq} = \frac{1}{4} + \frac{\gamma}{2\kappa}. \tag{19}$$

Again we have, as in the off-line case, the lower bound

$$\mathrm{ISI} \geq \frac{\mu}{4}. \tag{20}$$

We note that this bound is reached when $s = \pm 1$ with equal probability and ϕ is an odd function because then $\gamma = 0$.

There is a slight difference between the asymptotic ISI in batch and off-line algorithms. This should be considered by keeping in mind that adaptive algorithms have to optimized by balancing the ISI *and* the convergence speed.

5. CONCLUSION

We have tried to present the application of equivariance concepts to source separation in a unified manner for both adaptive and batch algorithms. We have not considered algorithm based on an estimating equation $\mathrm{E}H(\mathbf{y}) = 0$ because they are the simplest to analyze. In forthcoming work, we plan to include more general estimating equations that still yield equivariant estimates. Preliminary results are available in [22].

REFERENCES

[1] J. Hérault, C. Jutten, and B. Ans, "Détection de grandeurs primitives dans un message composite par une architecture de calcul neuromimétique en apprentissage non supervisé," in *Proc. GRETSI*, (Nice, France), pp. 1017–1020, 1985.

[2] L. Tong, R. Liu, V. Soon, and Y. Huang, "Indeterminacy and identifiability of blind identification," *IEEE Tr. on CS*, vol. 38, pp. 499–509, May 1991.

[3] J. Karhunen and J. Joutsensalo, "Representation and separation of signals using nonlinear PCA type learning," *Neural Networks*, vol. 7, no. 1, pp. 113–127, 1993.

[4] A. Chichocki and L. Moszczynski, "New learning alforithms for blind separation of sources," *Electronic Letters*, vol. 28, pp. 1986–1987, 1992.

[5] E. Moreau and O. Macchi, "New self-adaptive algorithms for source separation based on contrast functions," in *Proc. IEEE SP Workshop on Higher-Order Stat., Lake Tahoe, USA*, pp. 215–219, 1993.

[6] D.-T. Pham, P. Garrat, and C. Jutten, "Separation of a mixture of independent sources through a maximum likelihood approach," in *Proc. EUSIPCO*, pp. 771–774, 1992.

[7] P. Comon, "Independent component analysis," in *Proc. Int. Workshop on Higher-Order Stat., Chamrousse, France*, pp. 111–120, 1991.

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

[9] J.-F. Cardoso and A. Souloumiac, "Blind beamforming for non Gaussian signals," *IEE Proceedings-F*, vol. 140, pp. 362–370, Dec. 1993.

[10] O. Shalvi and E. Weinstein, "New criteria for blind deconvolution of nonminimum phase systems (channels)," *IEEE IT*, vol. 36, no. 2, pp. 312–321, 1990.

[11] L. Tong, Y. Inouye, and R. Liu, "Waveform preserving blind estimation of multiple independent sources," *IEEE Tr. on SP*, vol. 41, pp. 2461–2470, July 1993.

[12] J.-F. Cardoso, "Iterative techniques for blind source separation using only fourth order cumulants," in *Proc. EUSIPCO*, pp. 739–742, 1992.

[13] J.-F. Cardoso, "Fourth-order cumulant structure forcing. Application to blind array processing," in *Proc. 6th SSAP workshop on statistical signal and array processing*, pp. 136–139, Oct. 1992.

[14] L. Féty and J. P. Van Uffelen, "New methods for signal separation," in *Proc. of 4th Int. Conf. on HF radio systems*, (London), pp. 226–230, IEE, Apr. 1988.

[15] D. Pham and P. Garat, "Séparation aveugle de sources temporellement corrélées," in *Proc. GRETSI*, pp. 317–320, 1993.

[16] K. Abed Meraim, A. Belouchrani, J.-F. Cardoso, and Éric Moulines, "Asymptotic performance of second order blind source separation," in *Proc. ICASSP*, vol. 4, pp. 277–280, Apr. 1994.

[17] S. V. Gerven and D. V. Compernolle, "On the use of decorrelation in scalar signal separation," in *Proc. ICASSP*, (Adelaide, Australia.), 1994.

[18] E. L. Lehmann, *Testing statistical hypothesis*. Wiley pub. in statistics, John Wiley, 1959.

[19] J.-F. Cardoso, "The equivariant approach to source separation," in *Proc. NOLTA*, pp. 55–60, 1995.

[20] J.-F. Cardoso, A. Belouchrani, and B. Laheld, "A new composite criterion for adaptive and iterative blind source separation," in *Proc. ICASSP*, vol. 4, pp. 273–276, Apr. 1994.

[21] J.-F. Cardoso and B. Laheld, "Equivariant adaptive source separation." submitted to IEEE S.P., 1994.

[22] J.-F. Cardoso, "On the performance of source separation algorithms," in *Proc. EUSIPCO*, (Edinburgh), pp. 776–779, Sept. 1994.

[23] B. Laheld and J.-F. Cardoso, "Adaptive source separation without prewhitening," in *Proc. EUSIPCO*, (Edinburgh), pp. 183–186, Sept. 1994.