TWO SPATIO-TEMPORAL DECORRELATION LEARNING ALGORITHMS AND THEIR APPLICATION TO MULTICHANNEL BLIND DECONVOLUTION

Seungjin CHOI[†], *Andrzej CICHOCKI*[‡], *Shun-ichi AMARI*[‡]

[†] School of Electrical and Electronics Engineering, Chungbuk National University, KOREA

[‡] Brain-style Information Systems Group, Brain Science Institute, RIKEN, JAPAN

ABSTRACT

In this paper we present and compare two different spatio-temporal decorrelation learning algorithms for updating the weights of a linear feedforward network with FIR synapses (MIMO FIR filter). Both standard gradient and the natural gradient are employed to derive the spatio-temporal decorrelation algorithms. These two algorithms are applied to multichannel blind deconvolution task and their performance is compared. The rigorous derivation of algorithms and computer simulation results are presented.

1. INTRODUCTION

Multichannel blind deconvolution (MBD) is a fundamental problem encountered in a variety of applications such as wireless communications, image processing, array processing, and some biomedical applications. Let us define an m dimensional vector of observations, $\mathbf{x}(k)$ and an n dimensional vector of sources, $\mathbf{s}(k)$ as

$$\begin{aligned} \mathbf{x}(k) &= [x_1(k) \cdots x_m(k)]^T, \\ \mathbf{s}(k) &= [s_1(k) \cdots s_n(k)]^T. \end{aligned} \tag{1}$$

With this definition, the observation vector $\mathbf{x}(k)$ is assumed to be generated from an unknown source vector $\mathbf{s}(k)$ through the unknown MIMO FIR filter $\mathbf{H}(z)$ i.e.,

$$\mathbf{x}(k) = \mathbf{H}(z)\mathbf{s}(k) + \mathbf{v}(k), \tag{2}$$

where $\mathbf{v}(k)$ is an m dimensional additive white Gaussian noise vector that is assumed to be statistically independent of the source vector $\mathbf{s}(k)$. The FIR polynomial matrix $\mathbf{H}(z)$ is described as

$$\mathbf{H}(z) = \sum_{p=0}^{M} \mathbf{H}_p z^{-p}, \tag{3}$$

where z^{-p} is the delay operator such that $z^{-p}\mathbf{s}(k) = \mathbf{s}(k-p)$ and M is the order of the given MIMO FIR channel. We assume that source signals $\{s_i(k)\}$ are spatially independent and temporally i.i.d.

The task of multichannel blind deconvolution is to recover the source vector $\mathbf{s}(k)$ from the observation vector $\mathbf{x}(k)$, up to possibly scaled, reordered, and delayed estimates, i.e., $\hat{\mathbf{s}}(k) = \mathbf{P}\mathbf{\Lambda}\mathbf{D}(z)\mathbf{s}(k)$, where $\mathbf{P} \in \mathbb{R}^{n\times n}$ is a permutation matrix, $\mathbf{\Lambda} \in \mathbb{R}^{n\times n}$ is a nonsingular diagonal scaling matrix, and $\mathbf{D}(z)$ is a diagonal matrix given by

$$\mathbf{D}(z) = \mathrm{diag}\{z^{-d_1}, \cdots, z^{-d_n}\}. \tag{4}$$

In other words, the objective of multichannel blind deconvolution is to design a multichannel equalizer so that the global system $\mathbf{G}(z)$ (which combines the effect of channel and equalizer) has a decomposition of the following form:

$$\mathbf{G}(z) = \mathbf{P}\mathbf{\Lambda}\mathbf{D}(z). \tag{5}$$

For an finite order MIMO FIR channel, not every channel matrix $\mathbf{H}(z)$ has a decomposition (5). A channel matrix $\mathbf{H}(z)$ is said to be signal-separable [16] if there exist an equalizer (the inverse of the channel) so that $\mathbf{G}(z)$ has a decomposition (5). Sufficient conditions for signal-separability have been investigated by Massey and Sain [17] and Tugnait [20]. It usually requires strictly more sensors than sources, i.e., $m > n$. Throughout this paper, we will consider the case where the channel $H(z)$ satisfies the signal-separability conditions (see [20] for detailed signal-separability conditions). In addition, we neglect the effect of additive noise vector $\mathbf{v}(k)$, although the computer simulation was conducted with considering the noise vector.

2. WHY SPATIO-TEMPORAL DECORRELATION FOR MBD

It was shown [12, 13] that if the channel $\mathbf{H}(z)$ is signal-separable, then in the absence of additive noise $\mathbf{v}(k)$, spatio-temporal decorrelation can deconvolve the MIMO channel up to the instantaneous mixtures of source signals which can be further separated by independent component analysis [15, 11, 10, 5, 3, 6]. Let us define by $\mathbf{W}(z)$ an multivariate FIR filter for spatio-temporal decorrelation and by $\mathbf{U}$ a demixing matrix. Then $\mathbf{U}\mathbf{W}(z)\mathbf{H}(z)$ has a decomposition (5). Linear prediction method was employed in [12, 13] where some prior knowledge is required to find an innovation vector. Later, the spatio-temporal anti-Hebbian rule [7, 9] for a linear feedback network was developed for spatio-temporal decorrelation.

We would like to mention the advantages of this approach over other existing MBD methods.

- Spatio-temporal decorrelation is able to deconvolve the channel up to instantaneous mixtures of sources, so the number of sources can be easily detected via principal component analysis (PCA), whereas the number of sources is assumed to be known in [14, 18, 4, 19].
- This approach with the proposed algorithms is computationally efficient over the successive estimation [20], the sequential extraction [8].
- Spatio-temporal decorrelation is based on linear learning, so the convolutive mixtures of arbitrary-distributed sources can be separated. We found out this approach is efficient for the mixtures of super-Gaussian (sparse) sources, whereas most existing algorithms are focused on sub-Gaussian sources.

In this paper, we consider an linear feedforward network with FIR synapses (see Figure 1) and derive two efficient spatio-temporal decorrelation algorithms using both standard gradient and the natural gradient [1].

3. SPATIO-TEMPORAL DECORRELATION ALGORITHMS

We derive two algorithms which are able to minimize statistical dependence, although we are interested in decorrelation. We treat spatio-temporal decorrelation algorithms as a special case of the derived algorithms. As will be shown here, spatio-temporal decorrelation algorithms can be obtained using Gaussian density model.

3.1. Standard Gradient

We consider a linear feedforward network with finite order FIR synapses (multivariate FIR filter, see Figure 1) whose m dimensional output vector, $\mathbf{y}(k)$ is described as

$$\mathbf{y}(k) = \sum_{p=0}^{L} \mathbf{W}_p(k)\mathbf{x}(k-p), \tag{6}$$

where $\{\mathbf{W}_p(k)\}$ are synaptic weight matrices. We define $\mathbf{W}(z,k)$ as

$$\mathbf{W}(z,k) = \sum_{p=0}^{L} \mathbf{W}_p(k) z^{-p}. \tag{7}$$

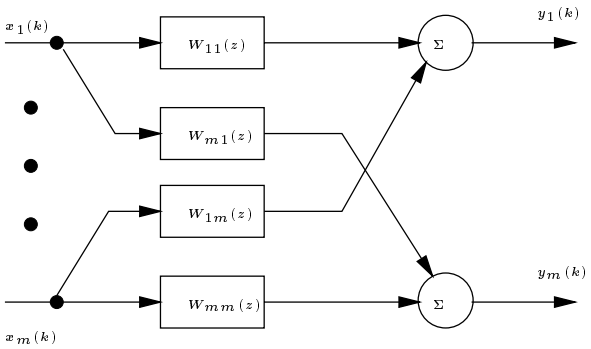


Figure 1: A linear feedforward network with FIR synapses.

We consider m observations $\{x_i(k)\}$ and m output signals $\{y_i(k)\}$ over a N-point time block. Let us define the following vectors:

$$\mathcal{X} = [\mathbf{x}^T(0) \cdots \mathbf{x}^T(N-1)]^T,$$
$$\mathcal{Y} = [\mathbf{y}^T(0) \cdots \mathbf{y}^T(N-1)]^T.$$

The coefficient matrices $\{\mathbf{W}_p(k)\}$ should be updated in such a way that the filter output signals are as spatio-temporally independent as possible, i.e., the joint probability density of $\mathcal{Y}$ is factored into the product of marginal densities:

$$\begin{aligned} p(\mathcal{Y}) &= \prod_{i=1}^{m}\prod_{k=0}^{N-1} q_i(y_i(k)) \\ &= \prod_{i=1}^{m} \{q_i(y_i)\}^N. \end{aligned} \tag{8}$$

In the second equality is the result of the assumption on identical distribution.

As an optimization function, we choose the Kullback-Leibler divergence which is an asymmetric measure of distance between two different probability distributions. Then, the risk $R(\mathbf{W}(z,k))$ (the optimization function) is given by

$$\begin{aligned} R(\mathbf{W}(z,k)) &= E\{L(\mathbf{W}(z,k))\} \\ &= \frac{1}{N}\int p(\mathcal{Y}) \log \frac{p(\mathcal{Y})}{\prod_{i=1}^{m}\{q_i(y_i)\}^N} d\mathcal{Y}. \end{aligned} \tag{9}$$

To derive the relation between $p(\mathcal{X})$ and $p(\mathcal{Y})$, we write (6) in a matrix form,

$$\mathcal{Y} = \mathcal{W}\mathcal{X}, \tag{10}$$

where $\mathcal{W}$ is given by

$$\mathcal{W} = \begin{bmatrix} \mathbf{W}_0 & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{W}_1 & \mathbf{W}_0 & \cdots & \mathbf{0} \\ \vdots & & & \vdots \\ \mathbf{W}_{N-1} & \mathbf{W}_{N-2} & \cdots & \mathbf{W}_0 \end{bmatrix} \tag{11}$$

The length of delay, L in the FIR filter is much smaller than N, i.e., $\mathbf{W}_{L+1} = \cdots = \mathbf{W}_{N-1} = \mathbf{0}$. The input-output equation (10) written in a matrix form, leads to the following relation between $p(\mathcal{X})$ and $p(\mathcal{Y})$:

$$p(\mathcal{Y}) = \frac{p(\mathcal{X})}{|\det \mathbf{W}_0^N|}, \tag{12}$$

where det denotes the determinant of a matrix. Invoking the relation (12), our loss function $L(\mathbf{W}(z,k))$ is given by

$$L(\mathbf{W}(z,k)) = -\log|\det \mathbf{W}_0| - \sum_{i=1}^{m} \log q_i(y_i). \tag{13}$$

Note that $p(\mathcal{X})$ was not included in (13) because it does not depend on the parameter matrix $\{\mathbf{W}_p(k)\}$.

Using the stochastic gradient descent, we can derive the following algorithm:

$$\begin{aligned} \Delta\mathbf{W}_p(k) &= -\eta_k \frac{dL(\mathbf{W}(z,k))}{d\mathbf{W}_p} \\ &= \eta_k\{\mathbf{W}_p^{-T}(k)\delta_p - \varphi(\mathbf{y}(k))\mathbf{x}^T(k-p)\}, \end{aligned} \tag{14}$$

where $\eta_k > 0$ is a learning rate and δ_p is the Kronecker delta equal to 1 if $p = 0$, otherwise it is zero. The $\varphi(\mathbf{y}(k))$ is a elementwise function defined as

$$\varphi(\mathbf{y}(k)) = [\varphi_1(y_1(k)), \ldots, \varphi_m(y_m(k))]^T, \tag{15}$$

where

$$\varphi_i(y_i) = -\frac{\partial \log q_i(y_i)}{\partial y_i}. \tag{16}$$

In order to avoid the computation of the inverse of the matrix $\mathbf{W}_0(k)$, we postmultiply (14) by $\mathbf{W}_0^T(k)\mathbf{W}_0(k)$. In addition, we add a constraint so that the magnitude of $\{y_i(k)\}$ is not controlled by the algorithm. Specially this constraint is efficient for overdetermined case. The resulting learning algorithm for updating $\mathbf{W}_0(k)$ has the form:

$$\Delta\mathbf{W}_0(k) = \eta_k\{\mathbf{\Gamma}(k) - \mathbf{y}(k)\mathbf{x}^T(k)\mathbf{W}_0^T(k)\}\mathbf{W}_0(k), \tag{17}$$

where $\mathbf{\Gamma}(k)$ is a diagonal matrix whose ith diagonal element is equal to the ith diagonal element of the matrix $\mathbf{y}(k)\mathbf{x}^T(k)\mathbf{W}_0^T(k)$. And for $p \neq 0$, the learning algorithm is

$$\Delta \mathbf{W}_p(k) = -\eta_k \mathbf{y}(k)\mathbf{x}^T(k-p). \quad (18)$$

As a special case of the algorithm (17), (18), linear learning ($\varphi(y_i) = y_i$) can obtained using the Gaussian density model, i.e.,

$$q_i(y_i) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}y_i^2}. \quad (19)$$

3.2. Natural Gradient

The derivation of the spatio-temporal decorrelation algorithm using the natural gradient was motivated by Amari *et al*'s previous work [4] in which the only complete case ($m = n$) was considered. In this derivation, we use the technique described in [4] and also incorporate a nonholonomic constraint [2] into the algorithm.

To determine an learning algorithm which minimizes the loss function (13), we calculate an infinitesimal increment,

$$dL(\mathbf{W}(z,k)) = L(\mathbf{W}(z,k) + d\mathbf{W}(z,k)) - L(\mathbf{W}(z,k)), \quad (20)$$

corresponding to an increment $d\mathbf{W}(z,k)$. Simple algebraic and differential calculus yields

$$dL(\mathbf{W}(z,k)) = \varphi^T(\mathbf{y}(k))d\mathbf{V}(z,k)\mathbf{y}(k) - \mathrm{tr}\{d\mathbf{V}_0(k)\}, \quad (21)$$

where $d\mathbf{V}(z,k)$ is defined as

$$d\mathbf{V}(z,k) = d\mathbf{W}(z,k)\mathbf{W}^{-1}(z,k), \quad (22)$$

and tr$\{\cdot\}$ denotes the trace operation. This gives the following learning algorithm in terms of $d\mathbf{V}(z,k)$ to minimize (13),

$$\begin{aligned} \Delta \mathbf{V}_p(k) &= -\eta_k \frac{dL\mathbf{W}(z,k)}{d\mathbf{V}_p(k)} \\ &= \eta_k \{\mathbf{I}\delta_p - \varphi(\mathbf{y}(k))\mathbf{y}^T(k-p)\}. \end{aligned} \quad (23)$$

The stationary points of (23) satisfy

$$E\{\varphi_i(y_i(k))y_i(k)\} = 1. \quad (24)$$

In other words, the learning algorithm (23) forces $\{y_i(k)\}$ to have constant magnitude. This might be a problem for $m > n$ if we do not know the number of source signals. To avoid this drawback, we follow the proposal on a nonholonomic constraint that was applied to blind source separation [2]. We propose to replace the identity matrix by a $m \times m$ diagonal matrix $\mathbf{\Lambda}_p(k)$ whose ith diagonal element is given by $\varphi_i(y_i(k))y_i(k-p)$. Then, the modified algorithm is

$$\Delta \mathbf{V}_p(k) = \eta_k \{\mathbf{\Lambda}_p(k)\delta_p - \varphi(\mathbf{y}(k))\mathbf{y}^T(k-p)\}. \quad (25)$$

Therefore, the learning algorithm in terms of $d\mathbf{W}(z,k)$ to minimize (13) has the form

$$\begin{aligned} \Delta \mathbf{W}_p(k) &= \sum_{r=0}^{L} \Delta \mathbf{V}_{p-r}(k)\mathbf{W}_r(k) \\ &= \eta_k \{\mathbf{\Lambda}_0(k)\mathbf{W}_p(k) \\ &- \varphi(\mathbf{y}(k-L)) \sum_{r=0}^{L} \mathbf{W}_{L-r}^T(k)\mathbf{y}(k-p-r)\}. \end{aligned} \quad (26)$$

Note that as in [4], the second term in the right-hand side of (26) is computed using the values delayed by L to avoid the noncausality.

4. COMPUTER SIMULATIONS

We present one exemplary computer simulation result here. Two source signals consist of random variables that are uniformly distributed over the binary set $\{+1, -1\}$. Three convolutive mixtures were generated through the following multivariate FIR channel:

$$\mathbf{x}(k) = \mathbf{H}_0\mathbf{s}(k) + \mathbf{H}_5\mathbf{s}(k-5) + \mathbf{H}_{10}\mathbf{s}(k-10) + \mathbf{v}(k), \quad (27)$$

where

$$\mathbf{H}_0 = \begin{bmatrix} -.9239 & .9924 \\ -.8228 & -.9845 \\ .7987 & -.3449 \end{bmatrix}, \mathbf{H}_5 = \begin{bmatrix} -.4465 & .5576 \\ .4386 & -.5137 \\ .5345 & -.5119 \end{bmatrix},$$

and

$$\mathbf{H}_{10} = \begin{bmatrix} -.0768 & -.1865 \\ -.2917 & -.0029 \\ -.1011 & -.0511 \end{bmatrix}.$$

The white Gaussian noise was added by the level of SNR=20dB. Two spatio-temporal decorrelation algorithms (17)-(18) and (26) were tested. For both algorithms, the linear learning, i.e., $\varphi_i(y_i) = y_i$ was used. The length of delay, L was set $L = 20$. The constant learning rate $\eta_k = .0005$ was used for both algorithms.

The output $\mathbf{y}(k)$ was fed into a linear feedforward network described by

$$\mathbf{z}(k) = \mathbf{U}(k)\mathbf{y}(k), \quad (28)$$

for further separation. Any ICA algorithm can be applied to update $\mathbf{U}(k)$. In this simulation, we have used

$$\mathbf{U}(k+1) = \mathbf{U}(k) + \eta_k \{\mathbf{I} - f^T(\mathbf{z}(k))\mathbf{z}(k)\}\mathbf{U}(k), \quad (29)$$

where $f(\mathbf{z}(k))$ is a elementwise cubic nonlinear function, i.e.,

$$f_i(z_i(k)) = |z_i(k)|^2 z_i(k). \quad (30)$$

To detect the number of sources, we have checked the spread of eigenvalues of the covariance matrix, $E\{\mathbf{y}(k)\mathbf{y}^T(k)\} = \mathbf{Q}\mathbf{\Sigma}\mathbf{Q}^T$ (see Figure 2). Three eigenvalue were 1.58, 1.28, and .03, so we can decide that there are two sources.

For performance measure, we have computed mean square error (MSE) after the arbitrary delay induced by the algorithm is eliminated. MSE with respect to each recovered signal was computed using a 50 point rectangular window (see Figures 3 and 4). The averaged value of MSE over the duration [10000,15000] was summarized in Table 1. It can be observed that the performance of the natural gradient learning in this task is slightly better than that of standard gradient.

5. CONCLUSIONS

We have presented two spatio-temporal decorrelation learning algorithms for updating the linear feedforward network with FIR synapses. The algorithms have been derived from an information-theoretic viewpoint using both standard gradient and the natural gradient. We incorporate a nonholonomic constraint into the natural gradient algorithm so that the resulting algorithm tolerates the overdetermined case. The demonstration of the algorithms was shown by applying them to multichannel blind deconvolution task.

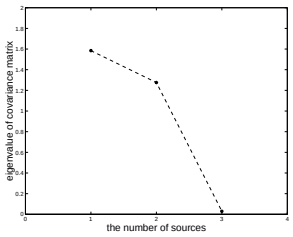
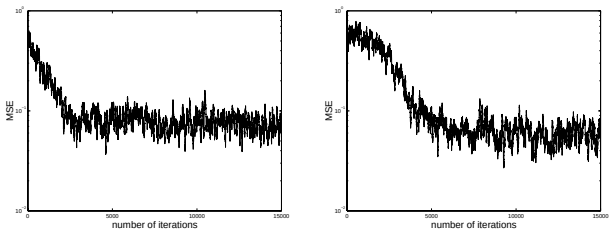


Figure 2: The spread of eigenvalues of covariance matrix R_{yy}.



(a) (b)

Figure 3: The mean squared error: (a) the MSE of $z_1(k)$ w.r.t $s_1(k)$; (b) the MSE of $z_2(k)$ w.r.t $s_2(k)$.

6. REFERENCES

[1] S. Amari. Natural gradient works efficiently in learning. *Neural Computation*, 10:251–276, 1998.

[2] S. Amari, T. P. Chen, and A. Cichocki. Nonholonomic orthogonal learning algorithms for blind source separation. *Neural Computation*, 1998, to appear.

[3] S. Amari, A. Cichocki, and H. H. Yang. A new learning algorithm for blind signal separation. In D. S. Touretzky, M. C. Mozer, and M. E. Hasselmo, editors, *Advances in Neural Information Processing Systems*, volume 8, pages 757–763. MIT press, 1996.

[4] S. Amari, S. C. Douglas, A. Cichocki, and H. H. Yang. Multichannel blind deconvolution and equalization using the natural gradient. In *The First Signal Processing Workshop on Signal Processing Advances in Wireless Communications*, pages 101–104, Paris, France, 1997.

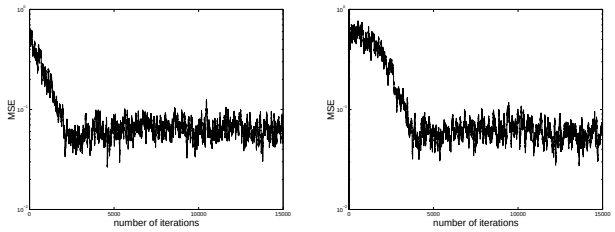
[5] A. Bell and T. Sejnowski. An information maximisation approach to blind separation and blind deconvolution. *Neural Computation*, 7:1129–1159, 1995.

[6] J. F. Cardoso and B. H. Laheld. Equivariant adaptive source separation. *IEEE Trans. Signal Processing*, 44:3017–3030, 1996.

[7] S. Choi and A. Cichocki. Blind signal deconvolution by spatio-temporal decorrelation and demixing. In J. Principe, L. Gile, N. Morgan, and E. Wilson, editors, *Neural Networks for Signal Processing 7*, pages 426–435. IEEE, 1997.

[8] S. Choi and A. Cichocki. On-line sequential multichannel blind deconvolution: A deflation approach. In *IEEE DSP Workshop*, Utah, 1998.

[9] S. Choi, A. Cichocki, and S. Amari. Blind equalization of SIMO channels via spatio-temporal anti-Hebbian learning rule. In *IEEE Workshop on Neural Networks for Signal Processing*, pages 93–102, Cambridge, UK, 1998.



(a) (b)

Figure 4: The mean squared error: (a) the MSE of $z_1(k)$ w.r.t $s_1(k)$; (b) the MSE of $z_2(k)$ w.r.t $s_2(k)$.

	Standard Gradient	Natural Gradient
MSE of z_1 w.r.t s_1	-11.14dB	-11.85dB
MSE of z_2 w.r.t. s_2	-12.12dB	-12.31dB

Table 1: The MSE performance.

[10] A. Cichocki, R. Unbehauen, and E. Rummert. Robust learning algorithm for blind separation of signals. *Electronics Letters*, 43(17):1386–1387, 1994.

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

[12] N. Delfosse and P. Loubaton. Adaptive blind separation of convolutive mixtures. In *Proc. ICASSP*, pages 2940–2943, 1996.

[13] S. Icart and R. Gautier. Blind separation of convolutive mixtures using second and fourth order moments. In *Proc. ICASSP*, pages 3018–3021, 1996.

[14] Y. Inouye. Blind deconvolution of multichannel linear time-invariant systems of nonminimum phase. In T. Katayama and S. Sugimoto, editors, *Statistical Methods in Control and Signal Processing*. Marcel Dekker, 1997.

[15] C. Jutten and J. Herault. Blind separation of sources, part i: An adaptive algorithm based on neuromimetic architecture. *Signal Processing*, 24:1–10, 1991.

[16] R. Liu and H. Luo. Direct blind separation of independent non-gaussian signals with dynamic channels. In *International Workshop on Cellular Neural Networks and their Applications*, 1998.

[17] J. Massey and M. Sain. Inverses of linear sequential circuits. *IEEE Trans. Computers*, 17:330–337, 1968.

[18] C. B. Papadias and A. J. Paulraj. A constant modulus algorithm for multiuser signal separation in the presence of delay spread using antenna arrays. *IEEE Signal Processing Letters*, 4(6):178–181, 1997.

[19] A. Touzni, I. Fijalkow, M. G. Larimore, and J. R. Treichler. A globally convergent approach for blind MIMO adaptive deconvolution. In *Proc. ICASSP*, pages 2385–2388, 1998.

[20] J. Tugnait. Identification and deconvolution of multichannel linear non-gaussian processes using higher order statistics and inverse filter criteria. *IEEE Trans. Signal Processing*, 45:658–672, 1997.