Blind Signal Extraction of Arbitrarily Distributed, but Temporally Correlated Signals - A Neural Network Approach

Ruck THAWONMAS†, Nonmember and Andrzej CICHOCKI‡†, Member

SUMMARY In this paper, we discuss a neural network approach for blind signal extraction of temporally correlated sources. Assuming autoregressive models of source signals, we propose a very simple neural network model and an efficient on-line adaptive algorithm that extract, from linear mixtures, a temporally correlated source with an arbitrary distribution, including a colored Gaussian source and a source with extremely low value (or even zero) of kurtosis. We then combine these extraction processing units, with deflation processing units to extract such sources sequentially in a cascade fashion. Theory and simulations show that the proposed neural network successfully extracts all arbitrarily distributed, but temporally correlated source signals from linear mixtures.

key words: blind source separation and extraction, neural networks, on-line adaptive algorithms

1. Introduction

Blind source separation has opened a new paradigm of signal processing with high potential for being applied to various areas of science and engineering [1-3,5-7,9,12,14-21,23]. There are in general two approaches for recovering the original source signals from their linear mixtures, namely, the simultaneous separation approach [1,3,5,9,12,15,18-21] and the extraction approach [6,7,10,11,14,17,23]. In the separation approach, separating algorithms perform separation of all the source signals simultaneously while in the extraction approach, the source signals are extracted one-by-one by eliminating the already extracted sources from their mixtures with deflation techniques. Except for a few neural algorithms in [1],[15],[18],[19], and [20] as well as non-neural methods surveyed in [21] that consider the temporal structures in source signals, most of the existing algorithms can not separate mixtures of sources with extremely low kurtosis or colored Gaussian sources. Furthermore, to separate mixtures of superand sub-Gaussian signals, it is necessary to use adaptive or switching nonlinearities which are controlled via computationally intensive measures, such as estimation of the sign of kurtosis of recovered signals. Moreover, the sign of kurtosis can change over time for non-stationary signals.

In this paper, adopting the extraction approach, we introduce a new technique for recovering one-by-one arbitrarily distributed and statistically independent signals, under a condition that signals are not i.i.d. (independent and identically distributed). We first give an extraction criterion that copes with the temporal structure in a source. A very simple neural network extraction model and an efficient on-line adaptive algorithm are then derived. A similar approach has been developed by S. Amari in [1] for simultaneous blind separation but his algorithm is different to ours. In addition, we prove analytically that the proposed criterion has no spurious extrema. This type of global convergence analysis has not been given for separating algorithms coping with temporally correlated sources in [1],[15],[18],[19], [20], and [21]. For extraction of multiple source signals, we propose a neural network that combines in a cascade fashion these extraction processing units with other processing units for deflation proposed in [7] and [23]. Although the employed neural network is linear, the developed associated on-line learning algorithms are nonlinear and biologically plausible. In fact the proposed learning algorithms are generalized Hebbian learning rule [8] and [13]. In general, linear neural networks [4] are often used because 1) they are well-suited for some specific applications† and 2) they provide on-line learning capabilities which can be easily implemented using current VLSI technologies. Simulations verifying the ability of the proposed technique to extract sequentially temporally correlated sources with arbitrary distributions are provided.

2. Blind Signal Extraction

In this section, we first give a formulation of the problem. We then present a processing unit (see Fig. 1) and an associated learning algorithm that extract a temporally correlated source signal with an arbitrary distribution.

†The author is with Dept. of Information Systems Engineering, Kochi University of Technology, Tosayamada-cho, Kochi 782-8502, Japan.
‡†The author is with the Laboratory for Open Information Systems, Brain Science Institute, RIKEN, 2-1 Hiro-sawa, Wako, Saitama 351-0198, Japan. He is on leave from the Warsaw University of Technology, Poland.

††Besides blind source separation, other promising applications of linear neural networks are, for example, principal, minor, and independent component analysis and total least-square problem (see, e.g., [8] and [13]).
2.1 Problem Formulation

The task of the blind extraction problem is that of recovering one by one unknown source signals from sensor or measured signals described by \( \mathbf{x}(t) = \mathbf{A}s(t) \), where \( \mathbf{x}(t) = [x_1, x_2, \ldots, x_n]^T \) is an available \( n \times 1 \) sensor vector, \( s(t) = [s_1, s_2, \ldots, s_m]^T \) is an \( m \times 1 \) unknown source vector of mutually independent sources, and \( \mathbf{A} = [\mathbf{a}_1, \mathbf{a}_2, \ldots, \mathbf{a}_m] \) is an \( n \times m \) unknown full-rank mixing matrix with column vectors \( \mathbf{a}_j = [a_{1j}, a_{2j}, \ldots, a_{nj}]^T \) for \( j = 1, 2, \ldots, m \). In other words, given observation vector \( \mathbf{y}(t) \), it is necessary to find new series \( \mathbf{y}(t) = [y_1, y_2, \ldots, y_m]^T \) which are estimates of the source signals \( s(t) \). We would like to point out that, a successfully recovered source signal may have a reversed sign with any nonzero magnitude scaling factor, compared to the corresponding original source. In addition, the order of extracted signals need not be the same as the order of source signals. In other words, there are two intrinsic ambiguities (interdeterminacies) associated with blind separation or extraction of sources, i.e., arbitrary scaling and arbitrary permutation [2].

Each source signal \( s_i \) is assumed to have zero mean and be stochastically independent from each other. We assume further that each source signal is described by a stable auto-regressive (AR) model:

\[
s_i(t) = \xi_i(t) - \sum_{k=1}^{L} b_{ik} s_i(t-k), \tag{1}
\]

where \( \xi_i(t) \) for \( i = 1, 2, \ldots, m \) are zero-mean innovation i.i.d. sequences and \( b_{ik} \) ensure stability of the system. Note that \( \xi_i(t) \) and \( b_{ik} \) are unknown.

To cope with ill-conditioning problems (when a mixing matrix is ill-conditioned and/or source signals have different energy), we apply preprocessing (pre-whitening) to the sensor signals \( \mathbf{x} \) in the form of

\[
\mathbf{x}_1 = \mathbf{Qx},
\]

where \( \mathbf{Q} \in \mathbb{R}^{m \times n} \) is a decorrelation matrix ensuring that the covariance matrix \( \mathbf{R}_{\mathbf{x}_1}\mathbf{x}_1 = E(\mathbf{x}_1 \mathbf{x}_1^T) = \mathbf{I}_m \) is identity matrix.

2.2 One-unit Blind Extraction

To model the temporal structures of source signals, we consider a linear neural network cascaded with an adaptive FIR filter (cf. Fig. 1), where the input-output relations of the network and the filter are given, respectively, as follows:

\[
y_1(t) = \mathbf{w}_1(t)^T \mathbf{x}_1(t) = \sum_{j=1}^{n} w_{1j} x_{1j}(t)
\]

and

\[
\tilde{y}_1(t) = y_1(t) + \sum_{k=1}^{L} \tilde{w}_{1k} y_1(t-k) = \tilde{\mathbf{w}}_1(z^{-1})y_1(t),
\]

where \( \tilde{\mathbf{w}}_1(z^{-1}) = 1 + \sum_{k=1}^{L} \tilde{w}_{1k} z^{-k} \) and \( z^{-k} \) is the time delay operator such that \( z^{-k}f(t) = f(t-k) \) for any function \( f(t) \).

According to the AR model of the source signals in (1), the filter output can be represented as \( \tilde{y}_1(t) = y_1(t) - \tilde{y}_1(t) \), with \( \tilde{y}_1(t) = -\sum_{k=1}^{L} \tilde{w}_{1k} y_1(t-k) \) defined as an error or estimator of an innovation source \( \xi_i(t) \). The mean square error \( E[\tilde{y}_1^2(t)] \) achieves a minimum \( c_1^2 E[\xi_i^2(t)] \), where \( c_1 \) is a positive scaling constant, if and only if \( y_1 = \pm c_1 s_i \) for any \( i \in \{1, \ldots, m\} \) or \( y_1 = 0 \) holds. To prevent the latter trivial case, we need a constant to bound \( E[y_1^2(t)] \) to, say, 1. We can formulate this constrained criterion as

\[
\min_1 J_1(\mathbf{w}_1, \tilde{\mathbf{w}}_1) = \frac{1}{2} E[\tilde{y}_1^2] + \beta_1 (1 - E[\tilde{y}_1^2])^2,
\]

where \( \beta_1 > 0 \) is a constant factor. The standard stochastic gradient descent method leads to the learning algorithms for vector \( \mathbf{w}_1 \) and scalars \( \tilde{w}_{1k} \), respectively,

\[
\frac{d\mathbf{w}_1}{dt} = -\mu_1 \frac{\partial J_1(\mathbf{w}_1, \tilde{\mathbf{w}}_1)}{\partial \mathbf{w}_1},
\]

\[
= -\mu_1 (t) \tilde{y}_1(t) \tilde{\mathbf{x}}_1(t) + \mu_1(t) \beta_1 (1 - m_2(y_1(t))) y_1(t) \mathbf{x}_1(t), \tag{3}
\]

and

\[
\frac{d\tilde{w}_{1k}}{dt} = -\tilde{\mu}_1(t) \frac{\partial J_1(\mathbf{w}_1, \tilde{\mathbf{w}}_1)}{\partial \tilde{w}_{1k}},
\]

\[
= -\tilde{\mu}_1(t) \tilde{y}_1(t) y_1(t-k), \tag{4}
\]

where \( \tilde{\mathbf{x}}_1(t) = \mathbf{x}_1(t) + \sum_{k=1}^{L} \tilde{w}_{1k} \mathbf{x}_1(t-k) = \tilde{\mathbf{w}}_1(z^{-1})\mathbf{x}_1(t) \), and \( \mu_1 \) as well as \( \tilde{\mu}_1 \) are learning rates. The second-order moment \( m_2(y_1(t)) = E[y_1(t)^2] \) can be estimated on-line using the following averaging formula

\[
\frac{dm_2(y_1(t))}{dt} = \mu_1(t) [y_1^2(t) - m_2(y_1(t))], \tag{5}
\]
Theorem 1: The developed learning algorithms in (3) and (4) always converge to a desired solution giving $y_i = \pm c_1 s_i$, where $i \in \{1, \ldots, m\}$ and $c_1$ is a positive scaling constant.

Theorem 1 says that the cost function in (2) has no spurious minima. As a result, it is guaranteed that a desired solution can always be reached independent of initial conditions or, in other words, a source signal having temporal structures can always be extracted from the mixtures.

3. Neural Network for Multi-unit Blind Extraction

For extraction of multiple source signals, we discuss a neural network architecture (see Fig. 2) that connects, in a cascade fashion, extraction processing units and other processing units of different type for deflation described in [7] and [23]. In this cascade architecture, a jth deflation processing unit deflates (eliminates) the newly extracted source signal $y_j$, yielded by the jth extraction processing unit, from the mixtures $x_j = [x_{j1}, x_{j2}, \ldots, x_{jn}]^T$ and feeds the resulting outputs $x_{j+1}$, as new mixtures, to the next $(j+1)$th extraction processing unit which then extracts another source signal, if any. It has been analytically shown in [23] that the resulting outputs $x_{j+1}$ of the jth extraction processing unit do not include the already extracted signals $\{y_1, \ldots, y_j\}$ by the following linear transformation

$$x_{j+1}(t) \doteq x_j(t) - \tilde{a}_j(t)y_j(t),$$

which minimizes the loss function

$$\tilde{J}_j(\tilde{a}_j) = \frac{1}{2} E[x_{j+1}^2],$$

where $\tilde{a}_j = [\tilde{a}_{j1}, \tilde{a}_{j2}, \ldots, \tilde{a}_{jn}]^T$ and $x_j = [x_{j1}, x_{j2}, \ldots, x_{jn}]^T$. In (6), $y_j = w_j^T x_j$ is the output of the jth extraction processing unit whose weights $w_j = [w_{j1}, w_{j2}, \ldots, w_{jn}]^T$ are updated according to the learning rule in (9). Using Theorem 1 in Sect. 2.2 together with Theorem 2 given in [23], we can say that at each extraction processing unit, a temporally correlated source signal that is different to those already obtained previously can always be extracted from the mixtures.

For $\tilde{a}_j$, we obtain the following updating rule by applying the stochastic gradient descent method to (7)

$$\frac{d\tilde{a}_j}{dt} = -\mu_j(t) \frac{\partial \tilde{J}_j(\tilde{a}_j)}{\partial \tilde{a}_j},$$

where $\mu_j(t) > 0$ is a learning rate.

Applying the standard stochastic gradient descent method to a generalized criterion of (2) for the jth extraction processing unit, i.e., $J_j(w_j, \tilde{w}_j) = \frac{1}{2} E[y_j^2] + \beta_j (1 - E[y_j^2])^2$, we obtain the following learning algorithms for vectors $w_j$ ($j = 1, 2, \ldots, m$) and scalars $\tilde{w}_{jk}$, respectively,

$$\frac{dw_j}{dt} = -\mu_j(t)\tilde{y}_j(t)x_j(t),$$

and

$$\frac{d\tilde{w}_{jk}}{dt} = -\tilde{\mu}_j(t)\tilde{y}_j(t)(y_j(t) - k),$$

where

$$w_j = [w_{j1}, w_{j2}, \ldots, w_{jn}]^T,$$

$$\tilde{y}_j(t) = y_j(t) + \sum_{k=1}^L \tilde{w}_{jk}y_j(t - k) = \tilde{w}_j(z^{-1})y_j(t),$$

$$\tilde{x}_j(t) = x_j(t) + \sum_{k=1}^L \tilde{w}_{jk}x_j(t - k) = \tilde{w}_j(z^{-1})x_j(t),$$

$$\frac{dm_2(y_j(t))}{dt} = \tilde{\mu}_j(t)[y_j^2(t) - m_2(y_j(t))],$$

$\beta_j > 0$ is a constant factor, and $\mu_j(t)$, $\tilde{\mu}_j(t)$ as well as $\tilde{\mu}_j(t)$ are learning rates.

4. Experimental Results

We now verify the performance of the proposed algorithms via simulations. Due to limit of space, we give experimental results from four examples. In each example, source signals are mixed with a mixing matrix $A$ whose elements are randomly selected in the range...
[1, 1]. The length \( L \) of the FIR filter, for each extraction processing unit, is set to 50\(^{\dagger} \). All synaptic weights are initialized with random values in the range \([-0.1, 0.1]\). The initial learning rates for \( \mu_j(t) \), \( \bar{\mu}_j(t) \), \( \mu_j(t) \), and \( \bar{\mu}_j(t) \) are set to 0.01. During the learning process, these learning rates are exponentially decreased to zero. According to assumption B3 in Appendix A, optimal values of \( \beta_j \) require prior knowledge of the source signals. However, since the role of the constraint term in (2) is to prevent the trivial case where \( y_j \rightarrow 0 \), \( \beta_j(t) \) can be set in practice to \( \beta_j(t) = \frac{1}{m_2(y_j(0))} \) with \( m_2(y_j(0)) = 1 \). Extraction and deflation are performed in a quasi parallel manner with the delay of 5000 time steps for each extracted signal \( y_j \).

To show qualitatively the performance of the proposed algorithms, we use the performance index which is defined at the ith extraction processing unit by

\[
P_I = \frac{1}{m} \left( \sum_{j=1}^{m} \frac{\hat{e}_{ij}^2}{e_{ij}^2}, -1 \right),
\]

where

\[
\hat{e}_i = w_i^T \hat{A}_i = [\hat{e}_{i1}, \hat{e}_{i2}, \ldots, \hat{e}_{im}],
\]

\[
\hat{e}_{ij} = \max(\hat{e}_{ij}) \quad \text{for } j = 1, \ldots, m,
\]

\[
\hat{A}_i = (I - \hat{a}_{i-1} w_{i-1}^T) \hat{A}_{i-1},
\]

and

\[
\hat{A}_i = \begin{cases} \text{QA}, & \text{when mixed signals are whitened} \\ \text{A}, & \text{otherwise.} \end{cases}
\]

The better quality of the extracted source signal at the ith extraction processing unit, compared to the original source signal, the smaller is the value of \( P_I \).

4.1 Mixtures of Colored Gaussian Signals

Three colored Gaussian signals are used here. Each colored Gaussian signal is generated by passing Gaussian sequences with variance 1 through an FIR filter of length 10 whose elements were randomly chosen between -1 and 1\(^{\dagger} \). The normalized kurtosis of the resulting signals \( s_1 \), \( s_2 \), and \( s_3 \) are close to zero, i.e., 0.02, -0.02, and -0.06, respectively. These signals are mixed with the randomly chosen mixing matrix

\[
A = \begin{pmatrix}
0.82 & -0.90 & -0.62 \\
-0.54 & -0.84 & 0.69 \\
-0.52 & 0.28 & -0.65
\end{pmatrix}.
\]

\(^{\dagger}\) According to our experience, \( L \) should be sufficiently large in order to capture the temporal structures in sources, and determination of \( L \) is problem dependent. Use of too small \( L \) might sometimes prevents us from successful extraction of the source signals.

\(^{\dagger}\) Colored Gaussian signals used in subsequent examples are also generated in the same fashion.

Fig. 3 shows, from top to bottom, the original source signals \( s = [s_1, s_2, s_3]^T \), the whitened mixed signals \( x_1 = [x_{11}, x_{12}, x_{13}]^T \), and the extracted signals \( y = [y_1, y_2, y_3]^T \). The performance indexes are \( P_{I1} = 0.00002, P_{I2} = 0.00006, \) and \( P_{I3} = 0.00011 \). Visual comparison of the original source signals and the extracted signals (with \( y_1 = -s_1, y_2 = s_3, \) and \( y_3 = -s_2 \)), together with the performance indexes, confirms the validity of the proposed algorithms.

For this set of source signals, a number of other algorithms for blind extraction; such as the fixed point algorithm [14], the KneqNet [11], and our another online algorithm employing the optimization criterion based on normalized kurtosis [7] and [23]; have failed to recover the original sources from the mixtures. This is due to the fact that those algorithms require source signals to be different from Gaussian or, in other words, to have non-zero kurtosis.

4.2 Mixtures of Natural Speech Signals and a Colored Gaussian Signal

Two natural speech signals\(^{\ddagger}\), i.e., an English word /hello/ (\( s_1 \) with normalized kurtosis = 3.44) and a Japanese word /moshimoshi/ (\( s_3 \) with normalized kurtosis = 6.13), and a colored Gaussian signal (\( s_2 \) with normalized kurtosis = -0.003) are mixed by the same mixing matrix in the above example. The original source signals and the whitened mixed signals are shown in the top and middle rows of Fig. 4, respectively. The extracted signals are shown in the bottom row of Fig. 4.

\(^{\ddagger}\) It is known that speech signals have temporal structures.
which reveals that the recovered signals (with $y_1 = s_2$, $y_2 = -s_3$, and $y_3 = -s_1$) are very close to the original sources. The performance indexes are $PI_1 = 0.0037$, $PI_2 = 0.0045$, and $PI_3 = 0.0042$.

4.3 Mixtures of Non-i.i.d. Signals and i.i.d. Sequences

For this example, three non-i.i.d. signals, i.e., a sub-Gaussian signal $s_2$, a colored Gaussian signal $s_3$, and a super-Gaussian signal $s_4$, are mixed with two i.i.d. sequences, one being a uniform random noise $s_1$, and the other being a Gaussian random noise $s_5$. The normalized kurtosis of $s_1, s_2, s_3, s_4,$ and $s_5$ are -1.22, -2.00, -0.04, 0.41, and 0.07, respectively. These signals are mixed with the randomly chosen mixing matrix

$$
A = \begin{pmatrix}
-0.13 & -0.35 & -0.09 & -0.25 & 0.17 \\
0.69 & 0.24 & 0.60 & -0.03 & 0.05 \\
-0.63 & 0.77 & -0.73 & 0.94 & -0.67 \\
0.02 & 0.52 & -0.87 & -0.32 & -0.03 \\
-0.10 & 0.77 & -0.25 & -0.49 & -0.01
\end{pmatrix}.
$$

Our aim here is to show an interesting property of the proposed algorithms that, given mixtures of non-i.i.d. signals (or signals with temporal structures) and i.i.d. signals, only non-i.i.d. signals will be extracted from the mixtures. We conjecture that this property holds in general. A theoretical proof of this property is left as an open problem.

Fig. 5 shows, from top to bottom, the original source signals $s = [s_1, s_2, s_3, s_4, s_5]^T$, the whitened mixed signals $x = [x_{11}, x_{12}, x_{13}, x_{14}, x_{15}]^T$, and the extracted signals at the first three extraction processing units, $y = [y_1, y_2, y_3]^T$. The performance indexes are $PI_1 = 0.0030$, $PI_2 = 0.0000$, and $PI_3 = 0.0077$. Both performance indexes as well as visual comparison of the original source signals and the extracted signals (with $y_1 = -s_2, y_2 = -s_4,$ and $y_3 = -s_3$) confirm the validity...
of the aforementioned conjecture.

4.4 Mixtures of Image Signals

In this section, we further test our conjecture with image signals for a more difficult case where the number of sensors is greater than the number of sources. Three 512 × 512 images are used, where the image \(s_2\) is a natural image. This image has temporal correlations when scanned in one dimension. The other two images are interferences artificially generated from Gaussian i.i.d. noises \(s_1\) and binary i.i.d. sequences \(s_3\). The normalized kurtosis of images \(s_1\), \(s_2\), and \(s_3\) are 0.02, 0.31 and -2.00, respectively. These image signals are mixed with the randomly chosen non-square mixing matrix

\[
A = \begin{pmatrix}
-0.97 & -0.16 & 0.68 \\
0.49 & 0.69 & -0.96 \\
-0.11 & 0.05 & 0.36 \\
0.87 & -0.59 & -0.24 \\
-0.07 & 0.34 & 0.66
\end{pmatrix}.
\]

Fig. 6 shows, from top to bottom, the original images \(s = [s_1, s_2, s_3]^T\), the mixed images \(x = [x_{11}, x_{12}, x_{13}, x_{14}, x_{15}]^T\), and the extracted signal at the first processing unit \(y_1\). The performance index at the first processing unit is 0.00004. As can be seen, the natural image has been successfully extracted at this unit.

5. Conclusions

We have proposed a neural network for blind signal extraction of temporally correlated, but arbitrarily distributed source signals. The proposed neural network can also extract source signals whose distribution are close to Gaussian, subject to a condition that such signals have temporal structures. For both sub- and super-Gaussian signals, the same learning rule can be used. This was verified by theoretical considerations and simulations given in this paper. In addition, simulation results indicate that the neural network can extract non-i.i.d. signals out of mixtures of those signals and other i.i.d. sequences. These three features make the proposed network promising for practical use, e.g., in the area of non-invasive medical diagnosis and biomedical signal analysis, such as EEG, MEG, and ECG.

Acknowledgment

The authors wish to thank Prof. A.J. Lawrence, the corresponding Associate Editor, and anonymous referees for their valuable comments.

References

works: a survey,” *IEEE Trans. on Neural Networks*, vol. 6, no. 4, pp. 837-858, 1995.
separation method for multiesfer communications,” *IEEE
signal extraction in order specified by stochastics properties,”
[8] A. Cichocki and R. Unbehauen, Neural Networks for Opti-
with on-line learning for blind identification and blind sepa-
ration of sources,” *IEEE Trans. Circuits and Systems-I*,
vol. 43, pp. 894-906, Nov. 1996.
independent sources: A deflation approach,” *Signal Process-
deconvolution,” *Proc. of the 1998 IEEE Workshop on Neu-
ral Networks for Signal Processing*, IEEE Press, New York,
pp. 5-12, 1998.
network model for exploratory data analysis and independent
component analysis,” *IEEE Trans. Neural Networks*, vol. 9,
no. 6, pp. 1495-1501, 1998.
[13] S. Haykin, Neural Networks - A Comprehensive Founda-
independent component analysis,” *Neural Computation*,
on temporal structure of signals,” *Proc. of 1998 Intern-
[16] Y. Inoue, “Blind deconvolution of multichannel linear time-
invariant systems on nonminimum phase,” *Statistical Meth-
ods in Control and Signal Processing*, Eds. T. Katayama and
traction of one component of a linear mixture with a sin-
gle neuron,” *IEEE Trans. Neural Networks*, vol. 9, no. 1,
for blind separation of nonstationary signals,” *Neural Net-
independent signals using time delayed correlations,” *Phys.
blind source separation: a context-sensitive generalization of
linear, instantaneous combinations,” *Digital Signal Process-
ing*, vol. 6, pp. 5-16, 1996.
principal-component analysis,” *Europhys. Lett.*, vol. 10,
no. 7, pp. 693-698.

Appendix A: Global Convergence Analysis
-Proof of Theorem 1

Preliminaries:
Let us denote a row vector \( \mathbf{v}_1 = [v_{11}, v_{12}, \ldots, v_{1m}] \) = \( \mathbf{w}_1^T \mathbf{A} \). The output \( y_1(t) \) and error \( \tilde{y}_1(t) \) can now be expressed as \( y_1(t) = \sum_{j=1}^{m} v_{1j} s_j(t) \) and
\( \tilde{y}_1(t) = \tilde{\mathbf{w}}_1 (z^{-1}) \sum_{j=1}^{m} v_{1j} s_j(t) = \sum_{j=1}^{m} v_{1j} \xi_j(t) \), re-
spectively, where \( \xi_j(t) = \tilde{\mathbf{w}}_1 (z^{-1}) s_j(t) = \xi_j(t) + \sum_{k=1}^{L} (\tilde{w}_{1k} - b_{jk}) s_j(t-k) \). The 1st extraction process-
ing unit successfully extracts a source signal if the condi-
tion \( \mathbf{v}_1 = \pm \mathbf{c}_j \mathbf{e}_j \), holds, where \( \mathbf{e}_j \) is an 1 \( \times m \) row vector whose elements are zero except the jth element
being one, and \( c_j > 0 \) is a positive scaling constant. If this condition holds, the output signal \( y_1 \) converges to
\( \pm \ve_j s_j \).

Next, before we show our analytical results, we re-
call here some useful facts that we will use throughout
the proof of Theorem 1. Since zero-mean sources \( s_i \) and \( s_j \) are independent from each other, and so are \( \xi_i \) and \( \xi_j \), we have the following property P1:

\[
E[s_i(t)s_j(t+k)] = E[\xi_i(t)\xi_j(t+k)] = 0,
\]

\[
E[e_i(t)e_j(t+k)] = E[e_i(t)e_j(t+k)] = 0,
\]

for any integer \( k \) and \( i \neq j \). In addition, according to the AR model, we can say that \( \xi_i(t) \) and \( s_i(t-k) \), for \( k = 1, 2, \ldots \), are independent from each other since \( s_i(t-k) \) depends on past values \( \xi_i(t-k), s_i(t-k-1), \ldots \), \( \xi_i(t-k) \) is not among them. As a result, we also have the following prop-
erty P2:

\[
E[\xi_i(t)s_i(t-k)] = E[\xi_i(t)s_i(t-k)] = 0,
\]

for \( k = 1, 2, \ldots \). Exploiting the above properties, we can express the mean square error \( E[\tilde{y}_1^2(t)] \) as

\[
E[\tilde{y}_1^2(t)] = \sum_{j=1}^{m} \epsilon_j^2 E[\tilde{\xi}_j^2(t)], \tag{A.1}
\]

where

\[
E[\tilde{\xi}_j^2(t)] = E[(\tilde{\mathbf{w}}_1 (z^{-1}) s_j(t))^2] = E[\xi_j^2(t)] + E[(\sum_{k=1}^{L} (\tilde{w}_{1k} - b_{jk}) s_j(t-k))^2]. \tag{A.2}
\]

The Proof:
For the proof given below, we employ the following as-
sumptions.
B1: \( E[s_j(t)^2] = 1 \) \( \forall j \). This assumption is always feasible because the source signals have zero-mean and differences in the power can be absorbed by the mixing matrix \( A \).

B2: \( E[\xi_i^2] = E[\xi_j^2] \), where \( i \neq j \), if there exists \( \bar{w}_1(z^{-1}) \) satisfying \( E[[\bar{w}_1(z^{-1})s_j]^2] = E[\xi_i^2] \) and \( E[[\bar{w}_1(z^{-1})s_j]^2] = E[\xi_j^2] \). This assumption implies that the innovations of source signals with the same temporal structures must have different variances. Note that signals such as i.i.d. signals violate this assumption\(^1\).

B3: \( \beta_1 \gg E[\epsilon_j^2] \) \( \forall j \). How to choose proper values for penalty parameter \( \beta_1 \) in practice has been discussed in Sect. 4.

The cost function \( J_1(\mathbf{w}_1, \bar{\mathbf{w}}_1) \) given in (2) can be represented in terms of \( \mathbf{v}_1 \) and \( \bar{\mathbf{w}}_1 \) as follows:

\[
J_1(\mathbf{v}_1, \bar{\mathbf{w}}_1) = \frac{1}{2} E[[\bar{w}(z^{-1}) \sum_{j=1}^{m} v_j s_j(t)]^2] + \frac{\beta_1}{4} (1 - E[\sum_{j=1}^{m} v_j s_j(t)])^2,
\]

or, after simplification of the second term on the right-hand side according to property P1 and assumption B1,

\[
J_1(\mathbf{v}_1, \bar{\mathbf{w}}_1) = \frac{1}{2} E[[\bar{w}(z^{-1}) \sum_{j=1}^{m} v_j s_j(t)]^2] + \frac{\beta_1}{4} (1 - \sum_{j=1}^{m} v_j^2).
\]

(A-3)

Theorem 1 can, thereby, be reformulated as follows:

\textit{All local minima with respect to \( \mathbf{v}_1 \) of \( J_1(\mathbf{v}_1, \bar{\mathbf{w}}_1) \) belong to the set of all \( \mathbf{v}_1 \) satisfying the condition \( \mathbf{v}_1 = \pm c_1 \mathbf{e}_1 \), where \( c_1 \) is a positive scaling constant.}

Here, to analyze the stability of equilibrium points, we differentiate (A-3) with respect to \( \mathbf{v}_1 \) and with respect to \( \bar{\mathbf{w}}_1 \). After exploitation of property P1, we have at equilibria, respectively,

\[
\frac{\partial J_1(\mathbf{v}_1, \bar{\mathbf{w}}_1)}{\partial v_{1i}} = E[[\bar{w}(z^{-1})s_i(t)]^2] v_{1i} - \beta_1 (1 - \sum_{j=1}^{m} v_j^2) v_{1i} = 0,
\]

(A-4) and

\[
\frac{\partial J_1(\mathbf{v}_1, \bar{\mathbf{w}}_1)}{\partial \bar{w}_{1k}} = E[[\bar{w}(z^{-1}) \sum_{j=1}^{m} v_j s_j(t)]
\]

\[\{ \sum_{j=1}^{m} v_j s_j(t - k) \} = 0. \quad (A-5)\]

There are three types of critical points that we have to examine their stability, namely,

- **Type I:** the points \( \mathbf{v}_1 \) with two or more non-zero elements.

- **Type II:** the trivial point \( \mathbf{v}_1 = [0, \ldots, 0] \).

- **Type III:** the points \( \pm c_1 \mathbf{e}_1 \), where \( c_1 \) is a positive constant.

Note that only type-III critical points are our desired solutions. To perform a global convergence analysis, we need to show that type-I and -II critical points are not stable while proving that those of type-III are stable. We first prove in the following that type-I critical points are not stable. Let assume here without loss of generality that at this point \( \mathbf{v}_1(t) = [c_{11}, c_{12}, 0, \ldots, 0] \), where \( c_{11} \neq 0 \) and \( c_{12} = 0 \). From (A-1) and (A-3), the cost function at time \( t \) is

\[
J_1(t) = \frac{1}{2} \{ E[\epsilon_1^2(t)] c_{11}^2 + E[\epsilon_2^2(t)] c_{12}^2 \} + \frac{\beta_1}{4} (1 - (c_{11}^2 + c_{12}^2))^2.
\]

(A-6)

Below we will show that by small perturbation of \( \mathbf{v}_1 \) and \( \bar{\mathbf{w}}_1 \), \( J_1(t + 1) - J_1(t) < 0 \). From (A-4), since \( v_{11} \neq 0 \) and \( v_{12} = 0 \), we have

\[
E[\epsilon_1^2(t)] = E[\epsilon_2^2(t)] = 1 - (c_{11}^2 + c_{12}^2).
\]

(A-7)

By using property P1, (A-5) becomes

\[
\frac{\partial J_1(\mathbf{v}_1, \bar{\mathbf{w}}_1)}{\partial \bar{w}_{1k}} = c_{11}^2 E[\epsilon_{11}(t)] s_{1}(t - k) + c_{12}^2 E[\epsilon_{12}(t)] s_{2}(t - k) = 0.
\]

(A-8)

Equation (A-8) indicates that there are two cases that we have to consider, namely,

- **Case 1:** There exist at least one \( k \) such that both \( E[\epsilon_{11}(t)] s_{1}(t - k) \) and \( E[\epsilon_{12}(t)] s_{2}(t - k) \) are not zero.

- **Case 2:** \( E[\epsilon_{11}(t)] s_{1}(t - k) = E[\epsilon_{12}(t)] s_{2}(t - k) = 0 \) \( \forall k \).

For case 1, assuming without loss of generality that \( k = 1 \), we perturb \( \mathbf{v}_1(t) \) and \( \bar{\mathbf{w}}_1(t) \) with \( \Delta \mathbf{v}_1 \) and \( \Delta \bar{\mathbf{w}}_1 \), respectively, such that

\[
\mathbf{v}_1(t + 1) = [\sqrt{c_{11}^2 + \delta}, \sqrt{c_{12}^2 - \delta^2}, 0, \ldots, 0],
\]

and

\[
\bar{\mathbf{w}}_1(t + 1) = \begin{cases} \bar{w}_{1k}(t) - \delta E[\epsilon_{11}(t)] s_{1}(t - 1), & \text{for } k = 1 \\ \bar{w}_{1k}(t), & \text{for } k = 2, \ldots, L, \end{cases}
\]

\[\{ \sum_{j=1}^{m} v_j s_j(t - k) \}

\[\{ \sum_{j=1}^{m} v_j s_j(t - k) \} = 0. \quad (A-5)\]

\[\{ \sum_{j=1}^{m} v_j s_j(t - k) \} = 0. \quad (A-5)\]
where $\delta$ and $\tilde{\delta}$ are small positive numbers. Since $\frac{dE[e_{11}^2]}{dt} = \sum_{k=1}^{L} \frac{dE[e_{12}^2]}{\partial u_{1k}} \frac{\partial \hat{z}_{1k}}{dt}$, we can readily calculate $E[e_{11}(t+1)]$ and $E[e_{12}(t+1)]$, respectively, as follows:

$$E[e_{11}(t+1)] = E[e_{11}(t)] - 2\tilde{\delta}E^2[e_{11}(t)]s_1(t-1)$$

(A.9)

and

$$E[e_{12}(t+1)] = E[e_{12}(t)] - 2\delta E^2[e_{11}(t)]s_1(t-1)$$

(A.10)

From (A.8), $E[e_{12}(t)]s_2(t-1) = -\frac{c^2}{c_{12}}E[e_{11}(t)]s_1(t-1)$, (A.10), therefore, becomes

$$E[e_{12}(t+1)] = E[e_{12}(t)] + 2\frac{c^2}{c_{12}}E^2[e_{11}(t)]s_1(t-1)$$

(A.11)

The cost function at time $t+1$ is

$$J_1(t+1) = \frac{1}{2} E[e_{11}^2(t+1)](c_{11}^2 + \delta^2) + E[e_{12}^2(t+1)](c_{12}^2 - \delta^2) + \frac{\beta_j}{4}(1 - (c_{11}^2 + c_{12}^2)^2)$$

Using (A.9) and (A.11), we thus have

$$J_1(t+1) - J_1(t) = \frac{1}{2} \delta^2(E[e_{12}^2(t)] - E[e_{12}^2(t)]) - \frac{\delta}{\tilde{\delta}}(E[e_{11}(t)]s_1(t-1) - (1 + \frac{c_{12}}{c_{12}}))$$

From (A.7), $E[e_{12}^2(t)] = E[e_{12}^2(t)]$, thus $J_1(t+1) - J_1(t) < 0$. This means that $v_i(t) = |c_{11}, c_{12}, 0, \ldots, 0|$ for case 1 cannot be a local minimum.

For case 2, i.e., $E[e_{11}(t)]s_1(t-k) = E[e_{12}(t)]s_2(t-k) = 0 \ \forall k$, recalling property $P_2$, we have

$$E\left[\left\{\sum_{j=1}^{L} (\tilde{w}_{1k} - b_{1k})s_1(t) - j\right\}s_1(t-k)\right] = 0$$

(A.12)

Multiplying (A.12) by $(\tilde{w}_{1k} - b_{1k})$ and summing up the resulting terms for all $k$, we eventually get

$$E\left[\left\{\sum_{k=1}^{L} (\tilde{w}_{1k} - b_{1k})s_1(t-k)\right\}^2\right] = 0$$

From (A.2), this means that $E[e_{11}^2(t)] = E[e_{12}^2(t)]$. Similarly, it can be shown that $E[e_{12}^2(t)] = E[e_{12}^2(t)]$. Thus, from (A.7), $E[e_{12}^2(t)] = E[e_{12}^2(t)]$, which is against assumption $B_2$. This case can hence be omitted from consideration.

In the following, we examine the stability of type-II and -III critical points. Let $u_{1i} = [v_{11}, \ldots, v_{1m}, \tilde{w}_{11}, \ldots, \tilde{w}_{1L}]$. Using (A.4) and (A.5), we derive the Hessian matrix $H$

$$H = \frac{\partial^2 J_1(u_{1i})}{\partial u_{1i}, \partial u_{1j}}$$

at these critical points as follows. For the diagonal elements of $H$,

$$h_{ii} = \begin{cases} 
2v_i, & \text{for } i = 1, \ldots, m \text{ and } j = 1, \ldots, m \\
0, & \text{for } i = 1, \ldots, m \text{ and } j > m \\
\beta(1 - \sum_{j=1}^{m} v_{1j}^2 - 2\tilde{v}_{1i}^2), & \text{for } i \leq m \\
E\left[\sum_{j=1}^{m} v_{1j} s_j(t - (i-m))^2\right], & \text{otherwise}.
\end{cases}$$

(A.13)

For non-diagonal elements of $H$,

$$h_{ij} = \begin{cases} 
2v_i v_j, & \text{for } i \leq m \text{ and } j \leq m \\
0, & \text{for } i > m \text{ and } j < m \\
\beta(1 - \sum_{j=1}^{m} v_{1j}^2 - 2\tilde{v}_{1i}^2), & \text{for } i \leq m \text{ and } j > m \\
E\left[\sum_{j=1}^{m} v_{1j} s_j(t - (i-m))^2\right] + \sum_{j=1}^{m} v_{1j} s_j(t - (i-m)), & \text{for } i > m \text{ and } j \leq m.
\end{cases}$$

(A.14)

Now we are ready to perform the stability analysis. For type-II, i.e., the trivial point $v_i = 0$, the diagonal elements $h_{ii}$ for $i = 1, \ldots, m$ are negative due to assumption $B_3$. The Hessian $H$ is not positive semi-definite for this type. The point $v_i = 0$ is therefore unstable.

Finally, we examine type-III critical points, i.e., the points $\pm c_1 e_r$, where $c_1$ is a positive scaling factor. Let assume without loss of generality that $i^* = 1$ satisfying the following condition$^{11}$

$$E\left[\sum_{j=1}^{m} v_{1j}^2(t) s_j(t)^2\right] - E\left[\sum_{j=1}^{m} \tilde{w}_{1j}^2(t) s_j(t)^2\right] \geq 0$$

for $i = 1, \ldots, m$. Note also that, from (A.4), $E\left[\sum_{j=1}^{m} v_{1j}^2(t) s_j(t)^2\right] = \beta(1 - c_1^2)$. Therefore, the Hessian $H$ for this type is as follows:

$^{11}$A similar representation for a system of coupled differential equations, such as the one defined by eqs. (A.4) and (A.5), can be found in [22].

$^{11}$Since it can be shown that $E\left[\sum_{j=1}^{m} \tilde{w}_{1j}^2(t) s_j(t)^2\right] = E[c_1^2(t)]$, as done in the proof for case 2 of type-I critical points, such $i^*$ always exists. For example, $i^* = \arg\min_{i=1}^{m} E[c_i^2]$.
simplify the non-diagonal elements of the Hessian as

\[
\frac{\partial^2 E}{\partial x_i \partial x_j} = \left\{ \begin{array}{ll} 
2\beta c_i^2, & \text{for } i = 1 \\
E\left[ (\tilde{w}_1(z^{-1})s_1(t))^2 \right], & \text{for } 1 < i \leq m \\
-c_i E\left[ (\tilde{w}_1(z^{-1})s_1(t))^2 \right], & \text{for } m < i.
\end{array} \right.
\tag{A.15}
\]

and

\[
\frac{\partial^2 E}{\partial x_i \partial x_j} = \left\{ \begin{array}{ll} 
0, & \text{for } i \leq m \text{ and } j \leq m \\
2c_i E\left[ (\tilde{w}_1(z^{-1})s_1(t))s_1(t-(j-m)) \right], & \text{for } i = 1 \text{ and } j > m \\
0, & \text{for } 1 < i \leq m \text{ and } j > m \\
c_i^2 E\left[ (s_1(t-(i-m))s_1(t-(j-m)) \right], & \text{for } i > m \text{ and } j > m \\
c_i E\left[ \tilde{w}_1(z^{-1})s_1(t) \right]s_1(t-(i-m)) + s_1(t-(i-m))\tilde{w}_1(z^{-1})s_1(t), & \text{for } i > m \text{ and } j \leq m.
\end{array} \right.
\tag{A.16}
\]

From (A.5), \( c_i^2 E[(\tilde{w}_1(z^{-1})s_1(t))s_1(t-k)] = 0 \) for \( k = 1, \ldots, L \). Using also property P1, we can further simplify the non-diagonal elements of the Hessian as follows:

\[
\frac{\partial^2 E}{\partial x_i \partial x_j} = \left\{ \begin{array}{ll} 
0, & \text{for } i \leq m \\
0, & \text{for } i > m \text{ and } j \leq m \\
c_i^2 E[s_1(t-(i-m))s_1(t-(j-m))], & \text{for } i > m \text{ and } j > m.
\end{array} \right.
\tag{A.17}
\]

The Hessian of type-III critical points is positive definite because \( \mathbf{x}^T \mathbf{H} \mathbf{x} \) given by

\[
\mathbf{x}^T \mathbf{H} \mathbf{x} = 2\beta c_i^2 x_i^2 + \sum_{i=2}^{m} \left[ E[ (\tilde{w}_1(z^{-1})s_1(t))^2 ] - E[ (\tilde{w}_1(z^{-1})s_1(t))^2 ] \right] x_i^2 + c_i^2 E\left[ \sum_{i=m+1}^{m+L} x_i s_1(t-(i-m))^2 \right]
\]

is greater than zero for all nonzero \( \mathbf{x} = [x_1, \ldots, x_{m+L}]^T \). Hence, the desired extracting solutions are stable. This ends the proof. \( \square \)

Ruck Thawonmas received the B.Eng. degree in Electrical Engineering from Chulalongkorn University, Thailand, in Mar. 1987, the M.Eng. degree in Information Science from Ibaraki University, Japan, in Mar. 1990, and the D.Eng. degree in Information Engineering from Tohoku University, Japan, in Jan. 1994. Dr. Thawonmas is now Associate Professor of the Department of Information Systems Engineering, Tohoku University, Japan, where he has joined since Apr. 1, 1999. Prior to joining Tohoku, he was Visiting (HIVIPS) Researcher at Hitachi Research Laboratory, Hitachi, Ltd., from Jan. 1994 to Mar. 1996; Special Postdoctoral Researcher at the Brain Information Processing Group, Frontier Research Program, Institute of Physical and Chemical Research (RIKEN), from Apr. 1996 to Sep. 1997; Assistant Professor at the Department of Computer Hardware, the University of Aizu, from Oct. 1997 to Jul. 1998; BSI Researcher at the Brain-Style Information Systems Research Group, the Brain Science Institute, RIKEN from Aug. 1998 to Mar. 1999. His research interests lie in soft data analysis, in other words, data analysis using soft computing techniques such as neural networks and fuzzy logic. During his graduate study in Japan, he was a recipient of the Japanese Government (Monbusho) Scholarship from Apr. 1987 to Mar. 1993. He was Research Fellow at Tohoku University being granted by the Tohoku Kaihatsu Memorial Foundation from Apr. 1993 to Jan. 1994.

Andrzej Cichocki received the M.Sc. (with honors), Ph.D., and Habilitation Doctorate (Dr.Sc.) degrees, all in electrical engineering, from Warsaw University of Technology (Poland) in 1972, 1975, and 1982, respectively. Since 1972, he has been with the Institute of Theory of Electrical Engineering and Electrical Measurements at the Warsaw University of Technology, where he became a full Professor in 1991. He is the co-author of two books: MOS Switched-Capacitor and Continuous-Time Integrated Circuits and Systems (Springer-Verlag, 1989) and Neural Networks for Optimization and Signal Processing (Teubner-Wiley, 1993/94) and more than 150 research papers. He spent at University Erlangen-Nuernberg (Germany) a few years as Alexander Humboldt Research Fellow and Guest Professor, at Lehrstuhl fuer Allgemeine und Theorische Elektrotechnik directed by Professor Rolf Unbehauen. In 1995-1996 he has worked as a head of the laboratory for Artificial Brain Systems, at Frontier Research Program, RIKEN, Japan. He is currently working as a head of the laboratory for Open Information Systems, at Brain Science Institute, RIKEN in the Brain-Style Information Processing Group directed by professor Shun-ichi Amari. His current research interests include adaptive semi-blind signal processing, especially intelligent blind/sparse processing of biomedial signals, dynamic independent component analysis, neural networks and nonlinear dynamic systems theory.