Minimal Distortion Principle for Blind Source Separation

Kiyotoshi MATSUOKA[*1] and Satoshi NAKASHIMA[*2]

[*1]Department of Brain Science and Engineering and [*2] Department of Control Engineering

Kyusyu Institute of Technology

Sensui 1-1, Tobata, Kitakyushu, Japan

Abstract: Blind source separation (BSS) is a method for recovering a set of source signals from the observation of their mixtures without any prior knowledge about the mixing process. In BSS the definition of a source signal has an inherent indeterminacy; any linear transform of a source signal can also be considered a source signal. Due to this indeterminacy, there are an infinite number of valid separators that can extract the source signals. This paper proposes a principle for choosing an optimal separator among them in a certain sense. The optimal choice is made such that the observed signals are the least subjected to distortion by the separator. The proposed normalization has some favorable features, particularly for BSS of convolutive mixture.

I. Introduction

Blind source separation (BSS) is a method for recovering a set of source signals from the observation of their mixtures without any prior knowledge about the mixing process. It has been receiving a great deal of attention from various fields as a new signal processing technique. In view of the level of complexity, the mixing process can be classified into two types: instantaneous mixture and convolutive mixture. In this paper we deal with convolutive mixture in general.

Inherently BSS has two kinds of indeterminacy. One is the indeterminacy in the numbering of the sources and the other is that in the scaling or normalization. The latter indeterminacy is more essential and will be addressed in this paper. The indeterminacy has usually been considered unsubstantial, but it cannot be overlooked in view of actual implementations and applications of BSS. This paper addresses a basic question of what kind of normalization of the separator is optimal. An answer will be given based on the Minimal Distortion Principle. Below we describe what the principle is, why it is important, and how it can be implemented.

II. Mathematical Nomenclature

Here we summarize several notations appearing in the following sections. Matrix $\mathbf{X}=\left[x_{ij}\right]$ and transfer function matrix $\mathbf{X}(z)=\sum_{k=-\infty}^{\infty}\mathbf{X}_k z^{-k}$ below are always square matrices, and the coefficients $\mathbf{X}_k$ of $\mathbf{X}(z)$ are all real-valued.

- Frequency transfer function $\mathbf{X}(e^{2\pi jf})$ associated with $\mathbf{X}(z)$ is denoted by $\tilde{\mathbf{X}}(f)$. If $\tilde{\mathbf{X}}(f)$ is nonsingular for every *f*, then $\mathbf{X}(z)$ is said to be nonsingular.
- The conjugate of the transpose of matrix **X** is denoted as $\mathbf{X}^{\dagger}$. The same notation is also used for transfer function matrix $\mathbf{X}(z)$ as $\mathbf{X}^{\dagger}(z)\triangleq\mathbf{X}^{T}(z^{-1})$.
- $\|\mathbf{x}\|$ represents the Euclidean norm of vector **x**. $\operatorname{tr}\mathbf{X}$ stands for the trace of matrix **X**.
- The Frobenius norm of matrix **X** is defined as $\|\mathbf{X}\|\triangleq\left(\operatorname{tr}\mathbf{X}\mathbf{X}^{\dagger}\right)^{1/2}=\left(\sum_{i,j}|x_{ij}|^{2}\right)^{1/2}$. Also the norm of transfer function matrix $\mathbf{X}(z)$ is defined as $\|\mathbf{X}(z)\|\triangleq\left(\sum_{k=-\infty}^{\infty}\|\mathbf{X}_k\|^{2}\right)^{1/2}$ or equivalently $\|\mathbf{X}(z)\|\triangleq\left(\int_{-1/2}^{1/2}\|\tilde{\mathbf{X}}(f)\|^{2}df\right)^{1/2}$.
- $\operatorname{diag}\{d_1,\dots,d_N\}$ or $\operatorname{diag}\{d_i\}$ represents the diagonal matrix which has diagonal entries $d_1,\dots,d_N$. $\operatorname{diag}\mathbf{X}$ ($\operatorname{off-diag}\mathbf{X}$) sets every nondiagonal (diagonal)

entry of matrix **X** to be zero.

- $\delta(\tau) \triangleq 1\ (\tau = 0),\ 0\ (\tau \neq 0)$.

III. The Mixing Process and the Demixing Process

Let us consider a situation where statistically independent random signals $s_i(t)$ ($i = 1,\cdots, N$) are generated by N sources and their mixtures are observed by N sensors. It is assumed that every source signal $s_i(t)$ is a stationary random process with zero mean, and the sensors' outputs $x_i(t)$ ($i = 1,\cdots, N$) are given by a linear mixing process:

$$\mathbf{x}(t) = \sum_{\tau=0}^{\infty} \mathbf{A}_\tau \mathbf{s}(t-\tau) = \mathbf{A}(z)\mathbf{s}(t), \qquad (1)$$

where $\mathbf{s}(t) \triangleq [s_1(t),\ \cdots, s_N(t)]^T$, $\mathbf{x}(t) \triangleq [x_1(t),\ \cdots, x_N(t)]^T$, and $\mathbf{A}(z) \triangleq \sum_{\tau=0}^{\infty} \mathbf{A}_\tau z^{-\tau}$. It is known that, in order to realize BSS, at most one source signal is allowed to be Gaussian.

For the mixing process we assume two conditions: $\sum_{\tau=0}^{\infty} \|\mathbf{A}_\tau\| < \infty$ and nonsingularity of $\mathbf{A}(z)$. The first condition states that the mixing process is stable, and the second one claims that $\mathbf{A}(z)$ must be invertible (though the inverse $\mathbf{A}^{-1}(z)$ may not be a causal system).

To recover the source signals from the sensor signals, we consider a demixing process (which will be referred to as the separator) of the form

$$\mathbf{y}(t) = \sum_{\tau=-\infty}^{\infty} \mathbf{W}_\tau \mathbf{x}(t-\tau) = \mathbf{W}(z)\mathbf{x}(t), \qquad (2)$$

where $\mathbf{y}(t) \triangleq [y_1(t),\ \cdots, y_N(t)]^T$ and $\mathbf{W}(z) \triangleq \sum_{\tau=-\infty}^{\infty} \mathbf{W}_\tau z^{-\tau}$.

If the mixing process $\mathbf{A}(z)$ is known beforehand, the source signals can be recovered by setting as $\mathbf{W}(z) = \mathbf{A}^{-1}(z)$, of course. Essential difficulty in BSS is that $\mathbf{A}(z)$ or $\mathbf{A}^{-1}(z)$ must be estimated from the observed data $\mathbf{x}(t)$ only. Besides, the impulse response $\{\mathbf{W}_\tau\}$ might need to take a noncausal form in general, i.e., $\mathbf{W}_\tau \neq \mathbf{O}$ $(\tau < 0)$.

In BSS the definition of the source signals has an indeterminacy. Namely, if $s_1(t)$, …, $s_N(t)$ are source signals, their arbitrarily linear-filtered signals $e_1(z)s_1(t)$, …, $e_N(z)s_N(t)$ can also be considered source signals because they are also mutually independent. The mixing process is then $\mathbf{A}(z)\mathrm{diag}\{e_1^{-1}(z),\ldots,e_N^{-1}(z)\}$. There is no way to distinguish between $\{s_i(t)\}$ and $\{e_i(z)s_i(t)\}$ because the only information we are given a priori is the fact that the sources are mutually independent and the mixing process is a linear one.

IV. Minimal Distortion Principle

We call a separator of the following form a valid separator:

$$\mathbf{W}(z) = \mathbf{D}(z)\mathbf{A}^{-1}(z), \qquad (3)$$

where $\mathbf{D}(z)$ is an arbitrary nonsingular diagonal matrix; $\mathbf{D}(z) = \mathrm{diag}\{d_i(z)\}$. If the separator is valid, each of the source signals appears at an output terminal of the separator, though it is subjected to a linear transformation $d_i(z)$. [More generally we should define a valid separator as $\mathbf{W}(z) = \mathbf{P}\mathbf{D}(z)\mathbf{A}^{-1}(z)$, where $\mathbf{P}$ is a permutation matrix, but we consider only the case of $\mathbf{P} = \mathbf{I}$ to make the description below simple.]

In BSS, all the valid separators are usually considered essentially equivalent. However the following separator has a special meaning:

$$\mathbf{W}^*(z) \triangleq \mathrm{diag}\,\mathbf{A}(z)\cdot\mathbf{A}^{-1}(z) \text{(i.e.,} \mathbf{D}(z) = \mathrm{diag}\,\mathbf{A}(z)) \quad (4)$$

We call this separator the optimal (valid) separator. It should be noted that this definition of the optimal separator has no indeterminacy; it is uniquely determined independently of in the indeterminacy in the definition of the source signals because the following holds for any diagonal matrix $\mathbf{E}(z)$:

$$\mathrm{diag}\,\mathbf{A}(z)\mathbf{E}(z)\cdot\left(\mathbf{A}(z)\mathbf{E}(z)\right)^{-1} = \mathrm{diag}\,\mathbf{A}(z)\cdot\mathbf{A}(z)^{-1}. \quad (5)$$

The optimal separator $\mathbf{W}^*(z)$ can be characterized by either of the following two propositions.

Proposition 1: The optimal separator $\mathbf{W}^*(z)$ is the valid separator that minimizes $\|\mathbf{W}(z)\mathbf{A}(z) - \mathbf{A}(z)\|^2$.

(Proof) It is easy to show

$$\begin{aligned}\|\mathbf{W}(z)\mathbf{A}(z) - \mathbf{A}(z)\|^2 &= \|\mathbf{D}(z) - \mathbf{A}(z)\|^2 \\ &= \int_{-1/2}^{1/2} \|\tilde{\mathbf{D}}(f) - \tilde{\mathbf{A}}(f)\|^2\, df\end{aligned}. \qquad (6)$$

So, we have only to consider the minimization of

2

$\left\|\tilde{\mathbf{D}}(f)-\tilde{\mathbf{A}}(f)\right\|^2$ with respect to $\tilde{\mathbf{D}}(f)$ for each f. We find that $\tilde{\mathbf{D}}(f)=\operatorname{diag}\tilde{\mathbf{A}}(f)$ or $\mathbf{D}(z)=\operatorname{diag}\mathbf{A}(z)$ minimizes (6). Substituting this into (3), we obtain (4).□

Proposition 2: The optimal separator $\mathbf{W}^*(z)$ is the valid separator that minimizes $E\left[\|\mathbf{y}(t)-\mathbf{x}(t)\|^2\right]$.

(Proof) It is easy to derive

$$E\left[\|\mathbf{y}(t)-\mathbf{x}(t)\|^2\right] \tag{7}$$

$$=\int_{-1/2}^{1/2}\operatorname{tr}\left(\tilde{\mathbf{D}}(f)-\tilde{\mathbf{A}}(f)\right)\mathbf{\Phi}_{\mathbf{s}}(f)\left(\tilde{\mathbf{D}}(f)-\tilde{\mathbf{A}}(f)\right)^{\dagger}df$$

where $\mathbf{\Phi}_{\mathbf{s}}(f)$ is the power spectrum of $\mathbf{s}(t)$, i.e, the Fourier transform of the auto-correlation matrix $E[\mathbf{s}(t)\mathbf{s}^T(t+\tau)]$. We have only to consider the minimization of $tr\left(\tilde{\mathbf{D}}(f)-\tilde{\mathbf{A}}(f)\right)\mathbf{\Phi}_{\mathbf{s}}(f)\left(\tilde{\mathbf{D}}(f)-\tilde{\mathbf{A}}(f)\right)^{\dagger}$ with respect to $\tilde{\mathbf{D}}(f)$ for each f. From the fact that $\mathbf{\Phi}_{\mathbf{s}}(f)$ is diagonal, we find that $\tilde{\mathbf{D}}(f)=\operatorname{diag}\tilde{\mathbf{A}}(f)$ or $\mathbf{D}(z)=\operatorname{diag}\mathbf{A}(z)$ gives the minimum of (7).□

These two propositions state the minimal distortion principle in two manners. Namely, the optimal separator is determined such that the overall transfer function $\mathbf{W}(z)\mathbf{A}(z)$ be as close to $\mathbf{A}(z)$ as possible, or equivalently the separator's output $\mathbf{y}(t)$ be as close to $\mathbf{x}(t)$ as possible. The optimal separator can also be characterized as a direct constraint on matrix $\mathbf{W}(z)$.

Proposition 3: The optimal separator $\mathbf{W}^*(z)$ is the valid separator that satisfies

$$\operatorname{diag}\mathbf{W}^{-1}(z)=\mathbf{I}. \tag{8}$$

(Proof) This equation implies that $\operatorname{diag}\mathbf{A}(z)\mathbf{D}^{-1}(z)=\mathbf{I}$. This leads to $\mathbf{D}(z)=\operatorname{diag}\mathbf{A}(z)$.□

The optimal separator has some properties that are favorable in actual implementation of BSS.

(i) The separator's output then becomes $\mathbf{y}(t)$ $=\operatorname{diag}\mathbf{A}(z)\cdot\mathbf{A}^{-1}(z)\mathbf{A}(z)\mathbf{s}(t)=\operatorname{diag}\mathbf{A}(z)\cdot\mathbf{s}(t)$. This implies that output $y_i(t)$ is $a_{ii}(z)s_i(t)$, which is the i-th source that would be observed at the i-th sensor when there were no other source signals. This property will be convenient for interpretation of the signals separated and later processing.

(ii) The optimal separator does not depend on the properties of the sources; it depends on the mixing process $\mathbf{A}(z)$ only. So, even for such nonstationary signals as voices, the optimal separator is invariant with time as long as the mixing process is fixed.

(iii) In actual implementation, the separator needs to be realized with an FIR filter. It is desirable that the filter's degree is as low as possible. Based on the minimal distortion principle, the separator is chosen such that the separator's output becomes as close to the sensor's output as possible. So, it can be expected that the separator will be realized with a relatively low degree.

Including the pioneering work by Herault and Jutten some studies on BSS have considered a separator of feedback structure;

$$\mathbf{y}(t)=\mathbf{x}(t)-\bar{\mathbf{W}}(z)\mathbf{y}(t), \tag{10}$$

where $\bar{\mathbf{W}}(z)$ is a matrix whose diagonal elements are all zeros. This is equivalent to putting $\mathbf{W}(z)$ $=\left(\mathbf{I}+\bar{\mathbf{W}}(z)\right)^{-1}$ in a feedforward-type separator, leading to $diag\,\mathbf{W}^{-1}(z)=\mathbf{I}$. So, the present normalization itself is not a new idea. What we want to stress is that the constraint (8) can be derived from the minimal distortion principle (Propositions 1 and 2). It is hard to design a feedback-type separator while its stability is secured, particularly for non-minimum phase mixing processes. Using Proposition 2, we can incorporate the constraint (8) easily in a multi-dimensional FIR filter, which is guaranteed to be stable.

V. An Implementation of the Minimal Distortion Principle

Here, we want to show how the proposed principle is implemented. We start with an approach proposed by Amari et al. [1] Define

$$I(\mathbf{W}(z))\triangleq-\sum_{i=1}^{N}E[\log q_i(y_i(t))]-h[\mathbf{y}(t)], \tag{10}$$

where $h[\mathbf{y}(t)]$ is the entropy *rate* of $\mathbf{y}(t)$ and $q_i(u)$ is a pdf assumed for source signal $s_i(t)$. If the source signals are iid (or liner processes in general) and $q_i(u)$ approximates well the real pdf of $s_i(t)$, then minimizing $I(\mathbf{W}(z))$ provides a valid solution. In actual computation, however, the separator must be embodied by a FIR filter as $\mathbf{W}(z) \triangleq \sum_{\tau=0}^{L} \mathbf{W}_\tau z^{-\tau}$. The minimization is then performed by the following iterative calculation (natural gradient learning):

$$\Delta \mathbf{W}_\tau = \alpha\left\{\mathbf{W}_\tau - \varphi(\mathbf{y}(t-L))\sum_{r=0}^{L} \mathbf{y}^T(t-L-\tau+r)\mathbf{W}_r\right\}, \quad (11)$$

where $\varphi(\mathbf{y}(t-L)) \triangleq [\varphi_1(y_1(t-L)),...,\varphi_N(y_N(t-L))]^T$ and φ_i is defined as $\varphi_i(u) \triangleq -d\log q_i(u)/du$. α is a small positive constant.

This algorithm however has some problems:

(i) The separator's outputs are made iid. Namely, the recovered signals will become white and hence they might be far different from the source signals observed at the sensors.

(ii) For the same reason an unnecessarily high degree FIR filter is required in general.

(iii) When the source signals are nonstationary, $\mathbf{W}(z)$ will fluctuate with time.

(iv) This algorithm induces an instability when the number of the sources is over-estimated.

To overcome these problems, Choi et al. [2] introduces a nonholonomic constraint to the algorithm. Let $d\mathbf{W}(z)$ be a tangent vector at $\mathbf{W}(z)$ and define

$$d\mathbf{V}(z) = \sum_k d\mathbf{V}_k z^{-k} \triangleq d\mathbf{W}(z)\mathbf{W}^{-1}(z). \quad (12)$$

Choi et al. [2] propose the nonholonomic constraint as $diag\, d\mathbf{V}_0 = \mathbf{0}$. We here extend the constraint to $diag\, d\mathbf{V}(z) = \mathbf{0}$. This modifies (11) as

$$\Delta \mathbf{W}_\tau = -\alpha\sum_{r=0}^{L}\left\{\text{off-diag}\,\varphi(\mathbf{y}(t-L))\mathbf{y}^T(t-L-\tau+r)\right\}\mathbf{W}_r \quad (13)$$

According to this algorithm, each output $y_i(t)$ of the separator becomes indeterminate with respect to linear transformation.

To some extent this algorithm alleviates the problems in algorithm (11). However, the indeterminacy introduced itself induces a numerical instability; the final value of $\mathbf{W}(z)$ depends on its initial value and moreover it fluctuates as the iterative modification proceeds.

The minimal distortion principle (Proposition 2) gives a solution. We superimpose the (natural) gradient of $E\left[\|\mathbf{y}(t)-\mathbf{x}(t-L/2)\|^2\right]$ to (13) as

$$\Delta \mathbf{W}_\tau = -\alpha\sum_{r=0}^{L}\{\text{off-diag}\,\varphi(\mathbf{y}(t-L))\mathbf{y}^T(t-L-\tau+r) + \beta\left(\mathbf{y}(t-L)-\mathbf{x}(t-3L/2)\right)\mathbf{y}^T(t-L-\tau+r)\}\mathbf{W}_r \quad (14)$$

Parameter β must be a sufficiently small positive constant. This algorithm gives the desired separator, independently of the initial condition of $\mathbf{W}(z)$.

VI. An Example

Here, we show a computer simulation. The mixing process is a two-input, two-output system given by

$$\mathbf{A}(z) = \begin{bmatrix} 1 & 0.5z^{-1} \\ 0.5z^{-1} & 1 \end{bmatrix}.$$

In the first simulation the source signals are $s_i(t) = g(z)u_i(t)$ $(i=1,2)$, where $u_i(t)$ is a binary-valued iid signal with $\Pr\{u_i(t) = \pm 1\} = 1/2$ and

$$g(z) = 0.864 + 0.094z^{-1} - 0.852z^{-2} + 0.873z^{-3}. \quad (15)$$

For φ_i, we use $\varphi_i(u) = u^3$, implying $q_i(u) \propto e^{-u^4/4}$, which is a sub-Gaussian distribution. The degree of the separator and the initial values of $\{\mathbf{W}_\tau\}$ are set as $L = 40$ and $\mathbf{W}_\tau = \delta(\tau - L/2)\mathbf{I}$, respectively. Parameters α and β are chosen as 1.0×10^{-5} and 0.1, respectively. Fig. 1(a) is the impulse responses of $\mathbf{W}(z)$ and $\mathbf{W}(z)\mathbf{A}(z)$ obtained by the algorithm (11); Fig. 1(b) is the result obtained by the proposed algorithm (14). Fig. 2 is the result obtained when $g(z)$ is changed as

$$g(z) = 1.749 + 0.133z^{-1} + 0.325z^{-2} - 0.794z^{-3}. \quad (16)$$

Compared to the result of (11), the result obtained by our algorithm (14) has better features:

(i) $\mathbf{W}(z)$ does not depend on $g(z)$, i.e., the property of the source signals.

(ii) $\mathbf{W}(z)\mathbf{A}(z) \approx z^{-L/2}\mathbf{I} = z^{-L/2}\mathrm{diag}\ \mathbf{A}(z)$, implying that the source signals observed at the sensors appear at the separator's output (with a dead time $L/2$).

(iii) The impulse response $\{\mathbf{W}_\tau\}$ is localized around $\tau = L/2 = 20$. This suggests that the degree L of the separator may be reduced to less than 10 in this case.

Next we consider other source signals, which are binary-valued signals generated by a Markov chain given by the following conditional probabilities:

$$\begin{aligned} \Pr\{s_i(t) = \pm 1 \mid s_i(t-1) = \pm 1\} &= (1+c)/2, \\ \Pr\{s_i(t) = \mp 1 \mid s_i(t-1) = \pm 1\} &= (1-c)/2, \end{aligned} \quad (17)$$

where parameter c takes a value between 0 and 1. When c is equal to 0, source $s_i(t)$ is a linear (iid) process. As the parameter c increases, nonlinearity of $s_i(t)$ is enhanced. Ohata and Matsuoka [3] show that the algorithm (11) does not give a valid solution stably when c exceeds around 0.75.

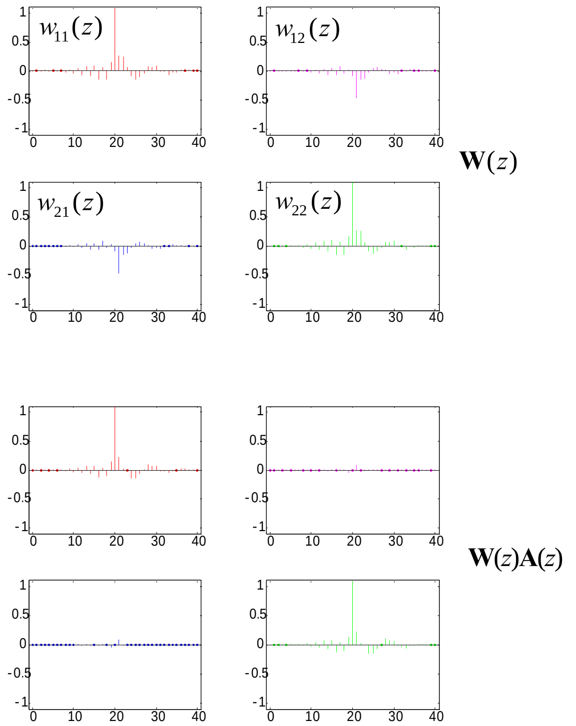
The two algorithms are applied to the sources with c = 0.7 and 0.8. The overall impulse responses are shown in Fig. 3. While algorithm (11) does not give a valid solution for c = 0.8, algorithm (14) provides the optimal separator stably. Theoretical study on stability is a future work.

References

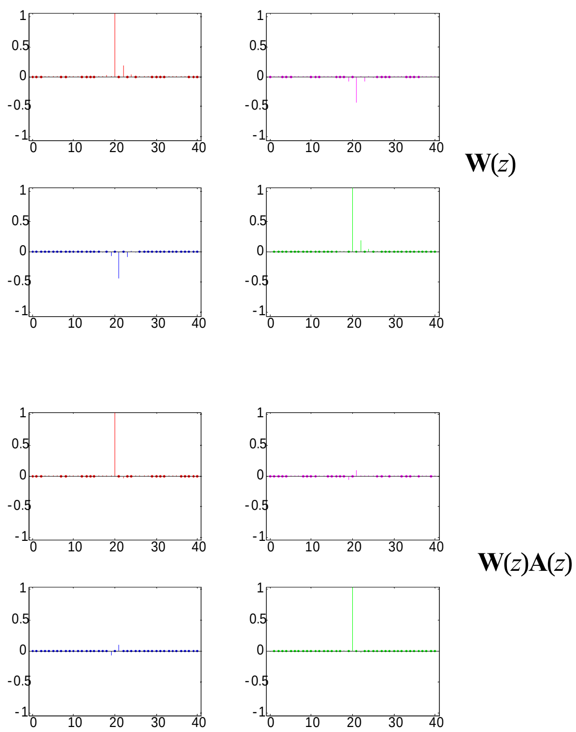
[1] S. Amari, S. C. Douglas, A. Cichocki and H. H.Yang, "Multichannel blind deconvolution and equalization using the natural gradient", Proc. IEEE International Workshop on Wireless Communication, pp. 101-104, 1997.

[2] S. Choi, S. Amari, A. Cichocki, and R. Liu, "Natural gradient learning with a nonholonomic constraint for blind deconvolution of multiple channels", Proc. International Workshop on Independent Component Analysis and Blind Signal Separation (ICA'99), pp. 371-376, 1999.

[3] M. Ohata and K. Matsuoka, "Stability analyses of a couple of blind separation algorithms when the sources are nonlinear processes", to appear.

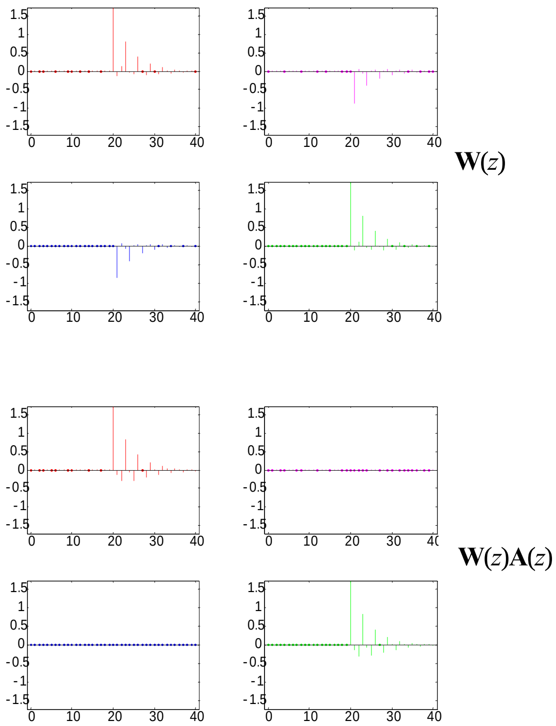


(a) The conventional method

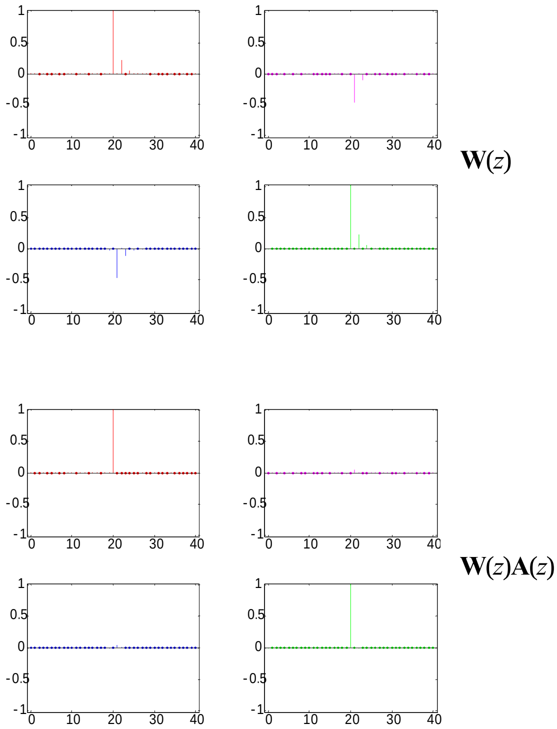


(b) The proposed method

Fig.1 The impulse response of **W**(z) and **W**(z)**A**(z) obtained by the conventional method (a) and the proposed method (b). The sources are given by (15).

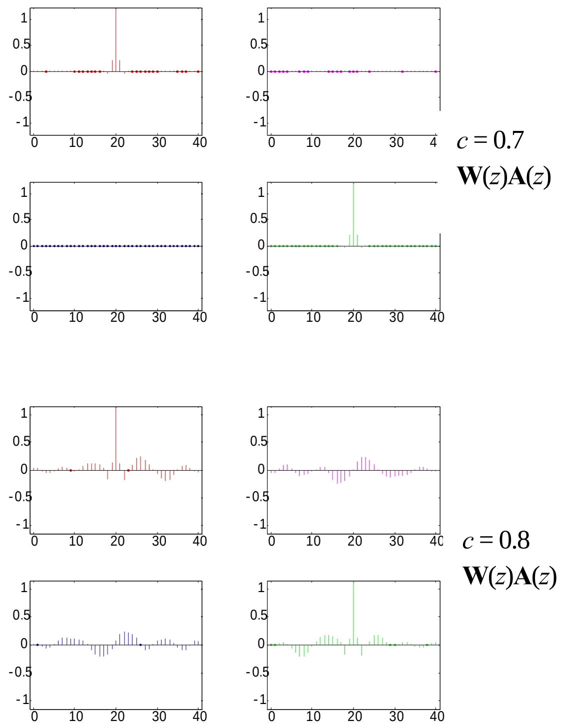


(a) The conventional method

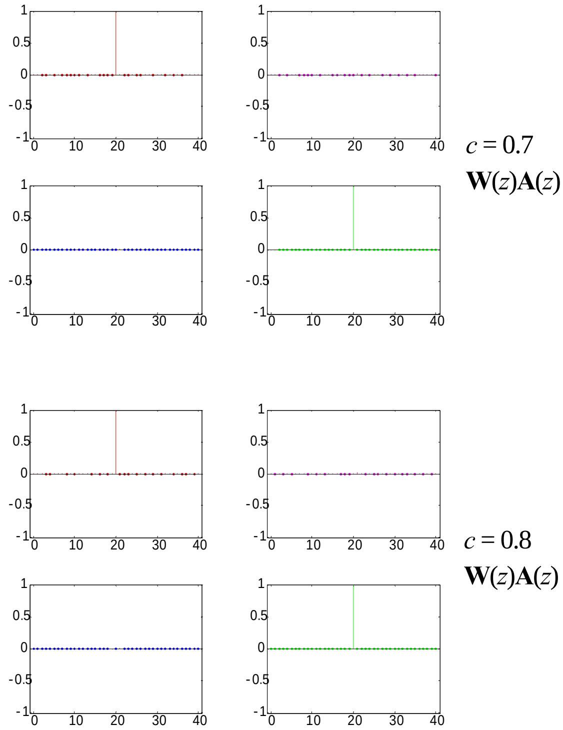


(b) The proposed method

Fig.2 The impulse response of $\mathbf{W}(z)$ and $\mathbf{W}(z)\mathbf{A}(z)$ obtained by the conventional method (a) and the proposed method (b). The sources are given by (16).



(a) The conventional method



(b) The proposed method

Fig.3 The impulse response of $\mathbf{W}(z)\mathbf{A}(z)$ obtained by the conventional method (a) and the proposed method (b). The sources are given by (17) with c = 0.7 and c = 0.8.