Journal of VLSI Signal Processing 31, 31–44, 2002

Geometrical Structures of FIR Manifold and Multichannel Blind Deconvolution

L.-Q. ZHANG, A. CICHOCKI* AND S. AMARI
Brain-Style Information Systems Research Group, RIKEN Brain Science Institute, Wako shi, Saitama 351-0198, Japan

Received April 26, 2000; Accepted August 31, 2001

Abstract. In this paper we study geometrical structures of the manifold of *Finite Impulse Response* (FIR) filters, and develop a natural gradient learning algorithm for blind deconvolution. First, A Lie group structure is introduced to the FIR manifold and the Riemannian metric is then derived by using the isometric property of the Lie group. The natural gradient on the FIR manifold is obtained by introducing a nonholonomic transformation. The Kullback-Leibler divergence is introduced as the measure of mutual independence of the output signals of the demixing model and a feasible cost function is derived for blind deconvolution. An efficient learning algorithm is presented based on the natural gradient approach and its stability analysis is also provided. Finally, we give computer simulations to demonstrate the performance and effectiveness of the proposed natural gradient algorithm.

Keywords: blind deconvolution, independent component analysis, Lie group, Riemannian metric, natural gradient, mutual information, learning algorithm, stability

1. Introduction

Independent component analysis (ICA) is to find a linear transformation of a vector of sensor signals (random variables) to new random variables that are maximally statistically independent. Since Jutten and Herault [1] introduced the concept of blind source separation (BSS), ICA has attracted considerable attention in the fields of signal processing and neural networks. There are diverse applications of ICA, covering from telecommunications [2], acoustics [3, 4], image enhancement [5], seismic data processing [6] to biomedical signal processing (EEG/MEG) [7, 8]).

Comon [9] first formulated the BSS problem in the framework of independent component analysis and presented cost functions by approximating mutual information of sensor signals. Cichocki and Unbehauen [10, 11] presented a robust learning algorithm, which is still one of the most popular ICA algorithms. Bell and Sejnowski elucidated blind source separation in the infomax principle. The natural gradient algorithm (equivalently the relative gradient algorithm) was developed by Amari et al. [12] and Cardoso and Laheld [13]. The fixed point algorithm (FastICA) was presented by Hyvarinen and Oja [14]. Stability analysis of the learning algorithms was also provided [13, 15]. Furthermore, some theoretical problems, such as convergence and efficiency of the learning algorithms were solved in the framework of the semiparametric model [16].

On the other hand, blind deconvolution/equalization was also developed independently in the field of communications [17–19]. The blind deconvolution/equalization is different from the traditional system identification and estimation. Generally, blind deconvolution uses only the sensor signals to estimate the original source signals, without knowing the source signals and the mixing system. Various algorithms, such as Bussgang algorithms [17–20], higher-order statistics approach [21, 22], information-theoretic approaches

*On leave from the Warsaw University of Technology, Poland.

[23–25] and the subspace method [26–28] have been developed for solving the blind deconvolution problem. Identifiability of blind deconvolution has also been discussed for single input multiple output (SIMO) systems [28, 29] and multiple input multiple output (MIMO) systems [27, 30, 31]. The blind deconvolution model was also extended to dynamical systems and the state space approach [32] was developed for blind deconvolution in the dynamical environment. The problem of blind separation of nonlinear mixtures was also discussed in [33].

However, most theoretical works, such as the geometrical structures of demixing model, stability and convergence, treat only blind source separation of instantaneous mixtures [13, 15, 16, 34, 35], and it is only recently that the natural gradient approach has been developed for multichannel blind deconvolution [23], where the demixing model is a doubly infinite impulse response (IIR) filter. Usually, it is necessary to use a doubly FIR filter as a demixing model in applications. The geometrical structures on the doubly FIR filter are different from the doubly IIR filter. The doubly FIR filters do not have self-closed operations, such as multiplications and inverse. Therefore, it is difficult to introduce the geometrical structures for the doubly FIR filters by the same procedure as for the doubly IIR filters.

In this paper, we surmount the difficulty by decomposing the doubly FIR filter into the product of two FIR filters. The geometrical structures are defined on the differential manifold of nonsingular FIR filters. First, a Lie group structure is introduced to the FIR manifold and Riemannian metric is then derived by using the isometric property of the Lie group. A novel approach is developed to derive the natural gradient on the FIR manifold. Introducing a nonholonomic transformation, we obtain an explicit expression of the natural gradient. The blind deconvolution problem is then formulated as an optimization problem and an efficient learning algorithm is developed using the natural gradient approach. Stability analysis of the natural gradient algorithm is provided for the first time in blind deconvolution case. Finally, we give computer simulations to demonstrate performance and effectiveness of the proposed learning algorithm.

2. Problem Formulation

As a convolutive mixing model, we consider a multichannel *linear time-invariant* (LTI) system of the form

$$\mathbf{x}(k) = \sum_{p=0}^{\infty} \mathbf{H}_p \mathbf{s}(k-p), \tag{1}$$

where $\mathbf{H}_p$ is an $n \times n$-dimensional matrix of mixing coefficients at time-lag p, called the impulse response at time p, $\mathbf{s}(k) = (s_1(k), \ldots, s_n(k))^T$ an n-dimensional vector of source signals, mutually independent and identically distributed, and $\mathbf{x}(k) = (x_1(k), \ldots, x_n(k))^T$ is an n-dimensional vector of sensor signals. For simplicity, we use the notation

$$\mathbf{H}(z) = \sum_{p=0}^{\infty} \mathbf{H}_p z^{-p}, \tag{2}$$

where z is the z-transform variable or the delay operator, defined by $z^{-1}\mathbf{x}(k) = \mathbf{x}(k-1)$. $\mathbf{H}(z)$ is called the mixing filter. Thus the mixing model (1) can be rewritten in the operator form

$$\mathbf{x}(k) = \mathbf{H}(z)\mathbf{s}(k). \tag{3}$$

The goal of multichannel blind deconvolution is to retrieve the source signals only using the sensor signals $\mathbf{x}(k)$ and some knowledge of source distributions and statistics. Generally, we carry out the blind deconvolution with another multichannel LTI and noncausal system of the form

$$\mathbf{y}(k) = \mathbf{W}(z)\mathbf{x}(k) = \sum_{p=-\infty}^{\infty} \mathbf{W}_p \mathbf{x}(k-p), \tag{4}$$

where $\mathbf{W}(z) = \sum_{p=-\infty}^{\infty} \mathbf{W}_p z^{-p}$, $\mathbf{y}(k) = [y_1(k), \ldots, y_n(k)]^T$ is an n-dimensional vector of the outputs and $\mathbf{W}_p$ is an $n \times n$-dimensional coefficient matrix at time lag p, which are the parameters determined during training. The system (4) is usually known as the demixing model, and $\mathbf{W}(z)$ is called the demixing filter. See Fig. 1 for illustration of the blind deconvolution problem.

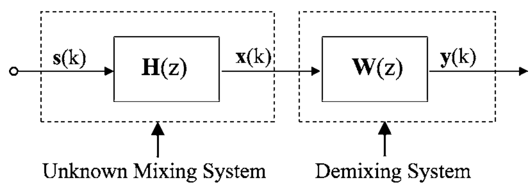

Figure 1. Illustration of blind deconvolution problem.

The global transfer function is defined by

$$\mathbf{G}(z) = \mathbf{W}(z)\mathbf{H}(z). \tag{5}$$

Generally speaking, blind deconvolution does not seek the inverse of the mixing filter. In blind deconvolution, we cannot observe the vector $\mathbf{s}(k)$ of original signals and the unknown mixing filter $\mathbf{H}(z)$ as well. This implies that there are three types of inherent ambiguities in the solution to blind deconvolution problem. We cannot identify the order in arranging the components $s_1(k), \ldots, s_n(k)$ into the vector $\mathbf{s}(k)$, the time origin and the magnitude of each component $s_i(k)$. Therefore, the task of blind deconvolution is to find a demixing filter $\mathbf{W}(z)$ such that

$$\mathbf{G}(z) = \mathbf{W}(z)\mathbf{H}(z) = \mathbf{P}\mathbf{\Lambda}\mathbf{D}(z), \tag{6}$$

where $\mathbf{P} \in \mathbf{R}^{n\times n}$ is a permutation matrix, $\mathbf{D}(z) = \text{diag}\{z^{-d_1}, \ldots, z^{-d_n}\}$, and $\mathbf{\Lambda} \in \mathbf{R}^{n\times n}$ is a nonsingular diagonal matrix.

The objective of blind deconvolution is to find a demixing filter $\mathbf{W}(z)$ such that its output signals $\mathbf{y}(k)$ are maximally spatially mutually independent and temporarily *i.i.d.*. In practice, we have to implement the blind deconvolution problem with a doubly *finite impulse response* (FIR) filter

$$\mathbf{W}(z) = \sum_{p=-N}^{N} \mathbf{W}_p z^{-p}, \tag{7}$$

where N is the length of the demixing filter. It is difficult to directly introduce geometrical structures, such as the Lie group and Riemannian metric to the doubly FIR manifold. Instead, we first introduce geometrical structures on the one-sided FIR filter manifold, and derive an efficient learning algorithm for the FIR filters. Using the filter decomposition approach [36], we can decompose a doubly FIR filter into the product of two one-sided FIR filters, $\mathbf{W}(z) = \mathbf{L}(z)\mathbf{R}(z^{-1})$, where $\mathbf{L}(z) = \sum_{p=0}^{N} \mathbf{L}_p z^{-p}$ is a causal FIR filter and $\mathbf{R}(z^{-1}) = \sum_{p=0}^{N} \mathbf{R}_p z^{p}$ an anti-causal FIR filter. Refer to [36] for more details. Therefore, the learning algorithm for one-sided FIR filters can used to train both filters $\mathbf{L}(z)$ and $\mathbf{R}(z^{-1})$, respectively.

3. Geometrical Structures of Nonsingular FIR Manifold

In this section we introduce some geometrical structures, such as the Lie group and Riemannian metric, to the manifold of FIR filters. Such structures are useful for the derivation of efficient learning algorithms [34].

The set of all FIR filters $\mathbf{W}(z)$ of length N, having the constraint $\mathbf{W}_0$ is nonsingular, is denoted by

$$\mathcal{M}(N) = \left\{ \mathbf{W}(z) \mid \mathbf{W}(z) = \sum_{p=0}^{N} \mathbf{W}_p z^{-p},\ \det(\mathbf{W}_0) \neq 0 \right\}. \tag{8}$$

$\mathcal{M}(N)$ is a manifold of dimension $n^2(N+1)$, which is referred to as *the nonsingular FIR manifold*. It is easy to prove that the manifold is a differential manifold. The tangent space at $\mathbf{W}(z) \in \mathcal{M}(N)$ is given by

$$\mathcal{T}_{\mathbf{W}}(\mathcal{M}(N)) = \left\{ \mathbf{P}(z) \mid \mathbf{P}(z) = \sum_{p=0}^{N} \mathbf{P}_p z^{-p} \right\}. \tag{9}$$

In general, multiplication of two filters in $\mathcal{M}(N)$ makes a new filter with length $2N$. This means that multiplication in ordinary sense is not self-closed in the manifold $\mathcal{M}(N)$. In order to explore possible geometrical structures of $\mathcal{M}(N)$, we define the algebraic operations of filters in the Lie group framework.

3.1. Lie Group

In the manifold $\mathcal{M}(N)$, Lie operations *multiplication* $*$ and *inverse* $\dagger$ are defined as follows: for any $\mathbf{B}(z), \mathbf{C}(z) \in \mathcal{M}(N)$,

$$\mathbf{B}(z) * \mathbf{C}(z) = \sum_{p=0}^{N} \sum_{q=0}^{p} \mathbf{B}_q \mathbf{C}_{(p-q)} z^{-p}, \tag{10}$$

$$\mathbf{B}^{\dagger}(z) = \sum_{p=0}^{N} \mathbf{B}_p^{\dagger} z^{-p}, \tag{11}$$

where $\mathbf{B}_p^{\dagger}$ are recurrently defined by $\mathbf{B}_0^{\dagger} = \mathbf{B}_0^{-1}$, $\mathbf{B}_p^{\dagger} = -\sum_{q=1}^{p} \mathbf{B}_{p-q}^{\dagger} \mathbf{B}_q \mathbf{B}_0^{-1}$, $p = 1, \ldots, N$. It is not difficult to verify that both $\mathbf{B}(z) * \mathbf{C}(z)$ and $\mathbf{B}^{\dagger}$ still remain in the manifold $\mathcal{M}(N)$ and the manifold $\mathcal{M}(N)$ forms a Lie Group with the above operations. The identity element is $\mathbf{E}(z) = \mathbf{I}$. Moreover, the Lie group has the following properties

$$\mathbf{A}(z) * (\mathbf{B}(z) * \mathbf{C}(z)) = (\mathbf{A}(z) * \mathbf{B}(z)) * \mathbf{C}(z), \tag{12}$$

$$\mathbf{B}(z) * \mathbf{B}^{\dagger}(z) = \mathbf{B}^{\dagger}(z) * \mathbf{B}(z) = \mathbf{I}. \tag{13}$$

In fact the Lie multiplication of two $\mathbf{B}(z), \mathbf{C}(z) \in \mathcal{M}(N)$ is the truncated form of the ordinary multiplication up to order N, that is

$$\mathbf{B}(z) * \mathbf{C}(z) = [\mathbf{B}(z)\mathbf{C}(z)]_N, \tag{14}$$

where $[\mathbf{B}(z)]_N$ is a truncating operator such that any terms with orders higher than N in the polynomial $\mathbf{B}(z)$ are omitted.

3.2. Riemannian Metrics

A Lie group has an important property that admits an invariant Riemannian metric. Let $\mathcal{T}_\mathbf{W}(\mathcal{M}(N))$ be the tangent space of $\mathcal{M}(N)$ at $\mathbf{W}(z)$, and $\mathbf{P}(z), \mathbf{Q}(z) \in \mathcal{T}_\mathbf{W}(\mathcal{M}(N))$ be the tangent vectors. We introduce the inner product with respect to $\mathbf{W}(z)$ as $\langle \mathbf{P}(z), \mathbf{Q}(z)\rangle_{\mathbf{W}(z)}$ in the following way. Since $\mathcal{M}(N)$ is a Lie group, any $\mathbf{B}(z) \in \mathcal{M}(N)$ defines an onto-mapping: $\mathbf{W}(z) \rightarrow \mathbf{W}(z) * \mathbf{B}(z)$. The multiplication transformation maps a tangent vector $\mathbf{P}(z)$ at $\mathbf{W}(z)$ to a tangent vector $\mathbf{P}(z) * \mathbf{B}(z)$ at $\mathbf{W}(z) * \mathbf{B}(z)$. Therefore we can define a Riemannian metric on $\mathcal{M}(N)$, such that the right multiplication transformation is isometric, that is, it preserves the Riemannian metric on $\mathcal{M}(N)$,

$$\begin{aligned}&\langle \mathbf{P}(z), \mathbf{Q}(z)\rangle_{\mathbf{W}(z)} \\ &\quad = \langle \mathbf{P}(z) * \mathbf{B}(z), \mathbf{Q}(z) * \mathbf{B}(z)\rangle_{\mathbf{W}(z)*\mathbf{B}(z)},\end{aligned} \tag{15}$$

for any $\mathbf{P}(z), \mathbf{Q}(z) \in \mathcal{T}_\mathbf{W}(\mathcal{M}(N))$. If we define the inner product at the identity $\mathbf{E}(z)$ by

$$\langle \mathbf{P}(z), \mathbf{Q}(z)\rangle_{\mathbf{E}(z)} = \sum_{p=0}^{N} \mathrm{tr}\big(\mathbf{P}_p \mathbf{Q}_p^T\big), \tag{16}$$

then $\langle \mathbf{P}(z), \mathbf{Q}(z)\rangle_{\mathbf{W}(z)}$ is automatically induced by

$$\langle \mathbf{P}(z), \mathbf{Q}(z)\rangle_{\mathbf{W}(z)} = \langle \mathbf{P}(z) * \mathbf{W}(z)^\dagger, \mathbf{Q}(z) * \mathbf{W}(z)^\dagger\rangle_{\mathbf{E}(z)}. \tag{17}$$

The Riemannian metric of the differential manifold $\mathcal{M}(N)$, denoted by $\mathcal{G}(\mathbf{W})$, can be explicitly calculated from Eq. (17) by introducing the Kronecker (tensor) product and the *vec* operator [37]. Due to its complexity, we will not pursuit further on the explicit expression of $\mathcal{G}(\mathbf{W})$. Actually, Eq. (17) provides us sufficient information to derive the natural gradient algorithm for blind deconvolution.

4. Natural Gradient

The FIR manifold $\mathcal{M}(N)$ is in the Riemannian space. The ordinary gradient is not optimal direction for minimizing a cost function defined on the Riemannian space. The steepest search direction is given by the natural gradient. It has been demonstrated that the natural gradient approach is an efficient technique for solving iterative estimation problems [34]. In this section, using the Lie group on $\mathcal{M}(N)$ we introduce a novel approach to derive the natural gradient without calculating the inverse of the Riemannian metric.

Assume that $l(\mathbf{W}(z))$ is a cost function defined on the manifold $\mathcal{M}(N)$. Since the parameters are in a matrix format, we define the ordinary gradient and natural gradient in the same matrix format. The ordinary gradient is denoted by

$$\nabla l(\mathbf{W}(z)) = \frac{\partial l(\mathbf{W}(z))}{\partial \mathbf{W}(z)} = \sum_{p=0}^{N} \frac{\partial l(\mathbf{W}(z))}{\partial \mathbf{W}_p} z^{-p}, \tag{18}$$

where

$$\frac{\partial l(\mathbf{W}(z))}{\partial \mathbf{W}_p} = \left(\frac{\partial l(\mathbf{W}(z))}{\partial \mathbf{W}_{p,ij}}\right)_{n\times n}, \quad p = 0, 1, \ldots, N. \tag{19}$$

The natural gradient $\tilde{\nabla} l(\mathbf{W}(z))$ is defined as the steepest ascent direction of the cost function $l(\mathbf{W}(z))$ as measured by the Riemannian metric on $\mathcal{M}(N)$.

In order to derive the natural gradient on the manifold $\mathcal{M}(N)$, we introduce the following notations. The operator *vec* transforms a matrix $\mathbf{A} = [\mathbf{a}_1, \mathbf{a}_2, \ldots, \mathbf{a}_n]$ to a vector $vec(\mathbf{A})$, defined by

$$vec(\mathbf{A}) = \big[\mathbf{a}_1^T, \mathbf{a}_2^T, \ldots, \mathbf{a}_n^T\big]^T. \tag{20}$$

We further define the *vec* operator for a filter $\mathbf{P}(z)$ in $\mathcal{T}_\mathbf{W}(\mathcal{M}(N))$ as

$$vec(\mathbf{P}(z)) = vec([\mathbf{P}_0, \mathbf{P}_1, \ldots, \mathbf{P}_N]). \tag{21}$$

Here, we present an alternative way to derive the natural gradient on the manifold $\mathcal{M}(N)$. According to the definition of the natural gradient [34], we have

$$vec(\tilde{\nabla} l(\mathbf{W}(z)) = \mathcal{G}(\mathbf{W})^{-1} vec(\nabla l(\mathbf{W}(z)). \tag{22}$$

For any FIR filter $\mathbf{P}(z) = \sum_{p=0}^{N} \mathbf{P}_p z^{-p}$ in the tangent space $\mathcal{T}_\mathbf{W}(\mathcal{M}(N))$, we take the inner product with $\mathbf{P}(z)$

on the both sides of the above equation

$$\begin{aligned}\langle \mathbf{P}(z), \nabla l(\mathbf{W}(z)\rangle_{\mathbf{E}(z)} \\ &= \langle vec(\mathbf{P}(z)), \mathcal{G}(\mathbf{W}) vec(\tilde{\nabla} l(\mathbf{W}(z))\rangle \\ &= \langle \mathbf{P}(z), \tilde{\nabla} l(\mathbf{W}(z))\rangle_{\mathbf{W}(z)}. \end{aligned} \quad (23)$$

Lemma 1. *The natural gradient* $\tilde{\nabla} l(\mathbf{W}(z))$ *of the cost function* $l(\mathbf{W}(z))$ *satisfies the following equation*

$$\langle \mathbf{P}(z), \tilde{\nabla} l(\mathbf{W}(z))\rangle_{\mathbf{W}(z)} = \langle \mathbf{P}(z), \nabla l(\mathbf{W}(z))\rangle_{\mathbf{E}(z)}, \quad (24)$$

for any $\mathbf{P}(z) \in \mathcal{T}_{\mathbf{W}}(\mathcal{M}(N))$.

The Eq. (24) has an geometrical interpretation: if we consider the filter $\mathbf{P}(z)$ as an element in $\mathcal{T}_{\mathbf{W}}(\mathcal{M}(N))$, then the inner product of $\mathbf{P}(z)$ and $\tilde{\nabla} l(\mathbf{W}(z))$ at $\mathbf{W}(z)$ is independent of $\mathbf{W}(z)$. This property is called isometry. Actually, this lemma provides us a new way to calculate the natural gradient. Using Eqs. (17) and (24), we can derive the natural gradient $\tilde{\nabla}(\mathbf{W}(z))$ in the following way,

$$\begin{aligned}&\langle vec(\mathbf{P}(z)), vec(\nabla l(\mathbf{W}(z))\rangle \\ &= \langle vec(\mathbf{P}(z)), vec(\tilde{\nabla} l(\mathbf{W}(z)) * \mathbf{W}^{-1}(z) \\ &\quad * \mathbf{W}^{-T}(z^{-1}))\rangle, \end{aligned} \quad (25)$$

for any $\mathbf{P}(z)$ in $\mathcal{T}_{\mathbf{W}}(\mathcal{M}(N))$. Comparing the two sides of the above equation, we obtain

$$\tilde{\nabla} l(\mathbf{W}(z)) = \nabla l(\mathbf{W}(z)) * \mathbf{W}^T(z^{-1}) * \mathbf{W}(z). \quad (26)$$

In fact for blind deconvolution, it will become much easier to calculate the natural gradient if we introduce a new differential variable,

$$d\mathbf{X}(z) = d\mathbf{W}(z) * \mathbf{W}^{\dagger}(z) = [d\mathbf{W}(z)\mathbf{W}^{-1}(z)]_N. \quad (27)$$

Consider the differential $dl(\mathbf{W}(z))$ with respect to $\mathbf{X}(z)$ and $\mathbf{W}(z)$, respectively,

$$\begin{aligned} dl(\mathbf{W}(z)) &= \left\langle \frac{\partial l(\mathbf{W}(z))}{\partial \mathbf{X}(z)}, d\mathbf{X}(z)\right\rangle \\ &= \left\langle \frac{\partial l(\mathbf{W}(z))}{\partial \mathbf{W}(z)}, d\mathbf{W}(z)\right\rangle \\ &= \left\langle \frac{\partial l(\mathbf{W}(z))}{\partial \mathbf{W}(z)} * \mathbf{W}^T(z^{-1}), d\mathbf{X}(z)\right\rangle. \end{aligned} \quad (28)$$

From the above equation, we deduce

$$\frac{\partial l(\mathbf{W}(z))}{\partial \mathbf{X}(z)} = \frac{\partial l(\mathbf{W}(z))}{\partial \mathbf{W}(z)} * \mathbf{W}^T(z^{-1}). \quad (29)$$

Substituting the above relation into (26), we obtain the following theorem

Theorem 1. *The natural gradient on the differential manifold* $\mathcal{M}(N)$ *is given by*

$$\tilde{\nabla} l(\mathbf{W}(z)) = \frac{\partial l(\mathbf{W}(z))}{\partial \mathbf{X}(z)} * \mathbf{W}(z). \quad (30)$$

It should be noted that $d\mathbf{X}(z) = [d\mathbf{W}(z)\mathbf{W}^{-1}(z)]_N$ is a nonholonomic basis, which has a definite geometrical meaning and proves to be useful in blind separation algorithms [15]. In fact, the differential $d\mathbf{X}(z)$ defines a channel error with respect to the variation of the output of the demixing model,

$$\begin{aligned} d\mathbf{y}(k) &= d\mathbf{W}(z)\mathbf{x}(k) \\ &= d\mathbf{W}(z) * \mathbf{W}^{\dagger}(z) * \mathbf{W}(z)\mathbf{x}(k) \\ &= d\mathbf{X}(z)\mathbf{y}(k). \end{aligned} \quad (31)$$

With this properties of the reparameterization, we can develop learning algorithms with the equivariance property [13].

5. Measure of Independence

The purpose of blind deconvolution is to find a FIR filter $\mathbf{W}(z)$ such that the output of the demixing model is maximally mutually independent and temporarily *i.i.d.*. The Kullback-Leibler Divergence has been used as a cost function for blind deconvolution [23] to measure the mutual independence of the output signals. Assume that $p_{\mathbf{y}}(\mathbf{y})$ is the joint probability density function of random variable $\mathbf{y}$, and $p_i(y_i)$ is the i-th marginal probability density function of y_i. The Kullback-Leibler Divergence between $p_{\mathbf{y}}(\mathbf{y})$ and $q_{\mathbf{y}}(\mathbf{y}) = \prod_{i=1}^n p_{y_i}(y_i)$ is given by

$$\mathcal{D}(p, q) = \int p_{\mathbf{y}}(y) \log\left(\frac{p_{\mathbf{y}}(\mathbf{y})}{\prod_{i=1}^n p_i(y_i)}\right) d\mathbf{y}, \quad (32)$$

or equivalently, we rewrite it into the mutual information form

$$l(\mathbf{W}) = -H(\mathbf{y}, \mathbf{W}) + \sum_{i=1}^n H(y_i, \mathbf{W}), \quad (33)$$

where $H(\mathbf{y}, \mathbf{W}) = -\int p(\mathbf{y}, \mathbf{W}) \log p(\mathbf{y}, \mathbf{W})\, d\mathbf{y}$, $H(y_i, \mathbf{W}) = -\int p_i(y_i) \log p_i(y_i)\, dy_i$. The divergence $l(\mathbf{W})$ is a nonnegative functional, which measures the mutual independence of the output signals $y_i(k)$. The output signals $\mathbf{y}$ are mutually independent if and only if $l(\mathbf{W}) = 0$. Therefore, the Kullback-Leibler Divergence $\mathcal{D}(p, \prod_{i=1}^n p_i(y_i))$ can be used as a cost function for blind deconvolution. However, there are several unknowns in the cost function: the joint probability density function $p_{\mathbf{y}}(\mathbf{y})$ and the marginal probability density functions $p_i(y_i)$. In the appendix, we show that the entropy $H(\mathbf{y})$ can be simplified as

$$H(\mathbf{y}) = -\log|\det(\mathbf{W}_0)| + const. \quad (34)$$

Therefore, the cost function derived from the mutual information becomes

$$l(\mathbf{y}, \mathbf{W}(z)) = -\log|\det(\mathbf{W}_0)| - \sum_{i=1}^n H(y_i, \mathbf{W}). \quad (35)$$

In order to implement the statistical on-line learning, we reformulate the cost function as

$$l(\mathbf{y}, \mathbf{W}(z)) = -\log|\det(\mathbf{W}_0)| - \sum_{i=1}^n \log q(y_i), \quad (36)$$

where $q(y_i)$ is an estimator of the true probability density function of the source signal. Actually, the choice of the distribution $q(y_i)$ is equivalent to the choice of its corresponding activation function. We will discuss the problem further in the following sections.

6. Learning Algorithm

In this section, we apply the stochastic natural gradient approach to derive a learning algorithm for online training FIR filters. First, we introduce the following lemma.

Lemma 2 ([2]). *If the matrix* $\mathbf{W}_0$ *is nonsingular*,

$$d\log|\det(\mathbf{W}_0)| = \mathrm{tr}\big(d\mathbf{W}_0\mathbf{W}_0^{-1}\big), \quad (37)$$

where tr *is the trace operation of matrices.*

For the gradient of $l(\mathbf{y}, \mathbf{W}(z))$ with respect to $\mathbf{W}(z)$, we calculate the total differential $dl(\mathbf{y}, \mathbf{W}(z))$

$$\begin{aligned} dl(\mathbf{y}, \mathbf{W}(z)) &= d\left(-\log|\det(\mathbf{W}_0)| - \sum_{i=1}^n \log q(y_i)\right) \\ &= -\mathrm{tr}\big(d\mathbf{W}_0\mathbf{W}_0^{-1}\big) + \varphi(\mathbf{y})^T d\mathbf{y}, \end{aligned} \quad (38)$$

where $\varphi(\mathbf{y})$ is a vector of nonlinear activation functions,

$$\varphi_i(y_i) = -\frac{d\log q_i(y_i)}{dy_i} = -\frac{q_i'(y_i)}{q_i(y_i)}. \quad (39)$$

By introducing the nonholonomic transform (27), we rewrite Eq. (38) as

$$dl(\mathbf{y}, \mathbf{W}(z)) = -\mathrm{tr}(d\mathbf{X}_0) + \varphi(\mathbf{y})^T d\mathbf{X}(z)\mathbf{y}. \quad (40)$$

From the above equation, we easily obtain the partial derivatives of $l(\mathbf{y}, \mathbf{W}(z))$ with respect to $\mathbf{X}(z)$,

$$\frac{\partial l(\mathbf{y}, \mathbf{W}(z))}{\partial \mathbf{X}_p} = -\delta_{0,p}\mathbf{I} + \varphi(\mathbf{y})\mathbf{y}^T(k-p), \quad p = 0, \ldots, N \quad (41)$$

Using the natural gradient descent learning rule, we present a novel learning algorithm as follows

$$\begin{aligned} \Delta\mathbf{W}_p &= -\eta\sum_{q=0}^p \frac{\partial l(\mathbf{y}, \mathbf{W}(z))}{\partial \mathbf{X}_q}\mathbf{W}_{p-q} \\ &= \eta\sum_{q=0}^p (\delta_{0,q}\mathbf{I} - \varphi(\mathbf{y})\mathbf{y}^T(k-q))\mathbf{W}_{p-q}, \end{aligned} \quad (42)$$

for $p = 0, 1, \ldots, N$, where η is the learning rate. In particular, the learning algorithm for $\mathbf{W}_0$ is described by

$$\Delta\mathbf{W}_0 = \eta(\mathbf{I} - \varphi(\mathbf{y})\mathbf{y}^T)\mathbf{W}_0. \quad (43)$$

It is worth noting that algorithm (42) is essentially different from the one in [23]. First, algorithm (42) has a triangle structure, i.e. $\Delta\mathbf{W}_p$ depends only on $\mathbf{W}_0, \ldots, \mathbf{W}_p$, given output signals $\mathbf{y}(k)$. Secondly, from the implementation point of view, we need to decompose the demixing filter into the product of two FIR filters if it is noncausal. The decomposition provides us a new way to train the demixing filter such that its coefficients decay as N becomes larger.

The algorithm (42) has two important properties, uniform performance (the equivariant property) [13] and non-singularity of $\mathbf{W}_0$. In multichannel blind deconvolution, an algorithm is equivariant if its dynamical behavior depends on the global transfer function $\mathbf{G}(z) = \mathbf{W}(z) * \mathbf{H}(z)$, but not on the specific mixing filter $\mathbf{H}(z)$. In fact the learning algorithm (42) has the equivariant property in the Lie group sense. Multiplying both sides of Eq. (42) by the mixing filter $\mathbf{H}(z)$ in the Lie group sense, we obtain

$$\Delta\mathbf{G}(z) = -\eta \frac{\partial l(\mathbf{y}, \mathbf{W}(z))}{\partial \mathbf{X}(z)} * \mathbf{G}(z). \tag{44}$$

where $\mathbf{G}(z) = \mathbf{W}(z) * \mathbf{H}(z)$. From Eq. (41) we know $\frac{\partial l(\mathbf{y},\mathbf{W}(z))}{\partial \mathbf{X}(z)}$ is formally independent of the mixing channel $\mathbf{H}(z)$. This means that the algorithm (42) is equivariant.

Another important property of the learning algorithm (43) is that it keeps the non-singularity of $\mathbf{W}_0$ provided the initial $\mathbf{W}_0$ is nonsingular [38, 39]. In fact if we denote the inner product of two matrices by $\langle \mathbf{A}, \mathbf{B}\rangle = \mathrm{tr}(\mathbf{A}^T\mathbf{B})$, we can easily calculate the derivative of the determinant $|\mathbf{W}_0|$ in the following way

$$\frac{d|\mathbf{W}_0|}{dt} = \left\langle \frac{\partial |\mathbf{W}_0|}{\partial \mathbf{W}_0}, \frac{d\mathbf{W}_0}{dt} \right\rangle = \left\langle |\mathbf{W}_0|\mathbf{W}_0^{-T}, \frac{d\mathbf{W}_0}{dt} \right\rangle \tag{45}$$

$$= \mathrm{tr}\left(|\mathbf{W}_0|\mathbf{W}_0^{-1}(\mathbf{I} - \varphi(\mathbf{y})\mathbf{y}^T)\mathbf{W}_0\right)$$
$$= \mathrm{tr}\left(\mathbf{I} - \varphi(\mathbf{y})\mathbf{y}^T)|\mathbf{W}_0|\right). \tag{46}$$

This equation results in

$$|\mathbf{W}_0(t)| = |\mathbf{W}_0(0)| \exp\left(\int_0^t \mathrm{tr}(\mathbf{I} - \varphi(\mathbf{y}(\tau))\mathbf{y}^T(\tau))d\tau \right). \tag{47}$$

Therefore the matrix $\mathbf{W}_0$ is nonsingular whenever the initial matrix $\mathbf{W}_0(0)$ is nonsingular.

This means that the learning algorithm (42) keeps the filter $\mathbf{W}(z)$ on the manifold $\mathcal{M}(N)$ if the initial filter is on the manifold. This property implies that the equilibrium point of the learning algorithm satisfy the following equations

$$E\{\varphi(\mathbf{y}(k))\mathbf{y}^T(k-p)\} = 0, \quad \text{for } p = 1, \ldots, N, \tag{48}$$

$$E\{\mathbf{I} - \varphi(\mathbf{y}(k))\mathbf{y}^T(k)\} = 0. \tag{49}$$

The nonlinear activation function $\varphi(\mathbf{y})$ originally is defined by (39). The choice of $\varphi(\mathbf{y})$ depends on both the statistics of the source signals and stability conditions of the learning algorithm.

7. Stability of Learning Algorithm

In this section, we analyze stability of the learning algorithm (42). Since the learning algorithm for updating $\mathbf{W}_p$, $p = 0, 1, \ldots, N$, is a linear combination of $\mathbf{X}_p$, $p = 0, 1, \ldots, N$, the stability of learning algorithm for $\mathbf{X}_k$, $k = 0, 1, \ldots, N$ implies the stability of the learning algorithm (42). Suppose that the separating signals $\mathbf{y} = (y_1, \ldots, y_n)^T$ are not only spatially mutually independent but also temporally independent and identically distributed. Now consider the learning algorithm for updating $\mathbf{X}_p$ in continuous time way,

$$\frac{d\mathbf{X}_p}{dt} = \eta(\delta_{0,p}\mathbf{I} - \varphi(\mathbf{y}(k)\mathbf{y}(k-p)^T), \quad p = 0, 1, \ldots, N. \tag{50}$$

To analyze the asymptotic properties of the learning algorithm, we take expectation on the above equation

$$\frac{d\mathbf{X}_p}{dt} = \eta(\delta_{0,p}\mathbf{I} - E[\varphi(\mathbf{y})\mathbf{y}^T(k-p)]), \quad p = 0, 1, \ldots, N. \tag{51}$$

If the variational matrix at equilibrium point is negative definite, then system (51) is stable in the vicinity of the equilibrium point. Taking a variation $\delta\mathbf{X}_p$ on $\mathbf{X}_p$, we have

$$\frac{d\delta\mathbf{X}_p}{dt} = -\eta E[\varphi'(\mathbf{y})\delta\mathbf{y}\mathbf{y}^T(k-p) + \varphi(\mathbf{y}(k))\delta\mathbf{y}^T(k-p)], \quad p = 0, 1, \ldots, N, \tag{52}$$

where $\delta\mathbf{y}(k-p) = \delta\mathbf{W}(z)\mathbf{u}(k-p) = \delta\mathbf{X}(z)\mathbf{y}(k-p) = \sum_{j=0}^{N} \delta\mathbf{X}_p\mathbf{y}(k-p-j)$. Using the mutual independence and i.i.d. properties of the output signals y_i, $i = 1, \ldots, n$ and the normalized condition (49), we deduce

$$\frac{d\delta\mathbf{X}_0}{dt} = -\eta\left(E[(\varphi'(\mathbf{y})\delta\mathbf{X}_0\mathbf{y})\mathbf{y}^T] + \delta\mathbf{X}_0^T\right), \tag{53}$$

$$\frac{d\delta\mathbf{X}_p}{dt} = -\eta(E[(\varphi'(\mathbf{y})\delta\mathbf{X}_p\mathbf{y}(k-p)\mathbf{y}(k-p)^T]), \quad p = 1, \ldots, N. \tag{54}$$

The above two equation systems can be rewritten into the following component form:

$$\frac{d\delta\mathbf{X}_{0,ij}}{dt} = -\eta\left(\kappa_i\sigma_j^2\delta\mathbf{X}_{0,ij} + \delta\mathbf{X}_{0,ji}\right), \tag{55}$$

$$\frac{d\delta\mathbf{X}_{0,ji}}{dt} = -\eta\left(\kappa_j\sigma_i^2\delta\mathbf{X}_{0,ji} + \delta\mathbf{X}_{0,ij}\right), \tag{56}$$

for $i \neq j$, and

$$\frac{d\delta\mathbf{X}_{0,ii}}{dt} = -\eta(m_i + 1)\delta\mathbf{X}_{0,ii}, \quad (57)$$

$$\frac{d\delta\mathbf{X}_{p,ij}}{dt} = -\eta\kappa_i\sigma_j^2\delta\mathbf{X}_{p,ij}, \quad (58)$$

for $p = 1, \ldots, N$, and $i, j = 1, \ldots, n$, where

$$m_i = E\left[\varphi'(y_i)y_i^2\right],\ \kappa_i = E[\varphi_i'(y_i)],\ \sigma_i^2 = E[|y_i|^2], \quad i = 1, \ldots, n. \quad (59)$$

For any $i \neq j$, (55) and (56) are a self-closed sub-systems, their stability conditions are given by

$$\kappa_i > 0, \quad \text{for } i = 1, \ldots, n, \quad (60)$$

$$\kappa_i\kappa_j\sigma_i^2\sigma_j^2 > 1, \quad \text{for } i, j = 1, \ldots, n. \quad (61)$$

Similarly, the stability conditions for (57) and (58) are as follows

$$m_i + 1 > 0, \quad \text{for } i = 1, \ldots, n, \quad (62)$$

$$\kappa_i > 0, \quad \text{for } i = 1, \ldots, n. \quad (63)$$

In summary, we have the following theorem

Theorem 2. *The stability conditions for* (53) *and* (54) *are*

$$m_i + 1 > 0, \quad (64)$$

$$\kappa_i > 0, \quad (65)$$

$$\kappa_i\kappa_j\sigma_i^2\sigma_j^2 > 1, \quad (66)$$

for all i, j $(i \neq j)$.

The stability conditions are identical to the ones derived by Amari et al. [15] for instantaneous blind source separation. Two families of activation functions have been discussed in [15]. In general, the nonlinear activation function $\varphi(y) = y^3 + \alpha y$, $0 < \alpha \ll 1$, is good for sub-Gaussian signals, and $\varphi(y) = \tanh(\gamma y)$, $0 < \gamma < 2$, is good for super-Gaussian signals, respectively. Refer to [15] for the detailed analysis for the instantaneous mixture.

8. Simulations

In order to implement the natural gradient algorithm, it is necessary to estimate first the length of the demixing filter. The model selection criteria, such as the Minimum Description Length (MDL) and Akaike Information-theoretic Criterion (AIC) can be used to select the model length N. Generally, the choice of the length N of the demixing filter usually depends on the mixing filter and error tolerance of recovered signals. Although we do not know the mixing filter in the blind deconvolution, we can estimate roughly the length N of the demixing filter by MDL criterion. The overestimate of N will not affect significantly the outcome of the natural gradient learning, but will increase the computing cost. Computer simulations show that the parameters in the overestimate range will automatically converge to zero.

To evaluate the performance of the proposed learning algorithms, we employ the multichannel inter-symbol interference, denoted by M_{ISI}, as a criterion,

$$M_{ISI} = \sum_{i=1}^{n} \frac{\sum_{j=1}^{n}\sum_{p=0}^{N}|\mathbf{G}_{pij}| - \max_{p,j}|\mathbf{G}_{pij}|}{\max_{p,j}|\mathbf{G}_{pij}|} + \sum_{j=1}^{n} \frac{\sum_{i=1}^{n}\sum_{p=0}^{N}|\mathbf{G}_{pij}| - \max_{p,i}|\mathbf{G}_{pij}|}{\max_{p,i}|\mathbf{G}_{pij}|}. \quad (67)$$

It is easy to show that $M_{ISI} = 0$ if and only if $\mathbf{G}(z)$ is of the form (6).

We give two examples to demonstrate the behavior and performance of the natural gradient algorithm (42). In both examples the mixing models are the three channel ARMA model

$$\mathbf{x}(k) + \sum_{i=1}^{10}\mathbf{A}_i\mathbf{x}(k-i) = \mathbf{B}_0\mathbf{s}(k) + \sum_{i=1}^{10}\mathbf{B}_i\mathbf{s}(k-i) + \mathbf{v}(k), \quad (68)$$

where $\mathbf{s} \in \mathbf{R}^3$ is the vector of source signals, $\mathbf{x} \in \mathbf{R}^3$ is the vector of sensor signals and $\mathbf{v} \in \mathbf{R}^3$ is the vector of noises. The matrices $\mathbf{A}_i$ and $\mathbf{B}_i$ are randomly chosen such that the mixing system is stable. The nonlinear activation function is chosen $\varphi(y) = y^3 + 0.01y$.

Example 1. In this simulation, the sources $\mathbf{s}$ are chosen to be i.i.d signals uniformly distributed in the range $(-1, 1)$, which are generated by computer, and $\mathbf{v}$ are the Gaussian noises with zero mean and a covariance matrix $0.1\,\mathbf{I}$. We perform 100 trials to demonstrate the natural gradient learning performance. In each trial,

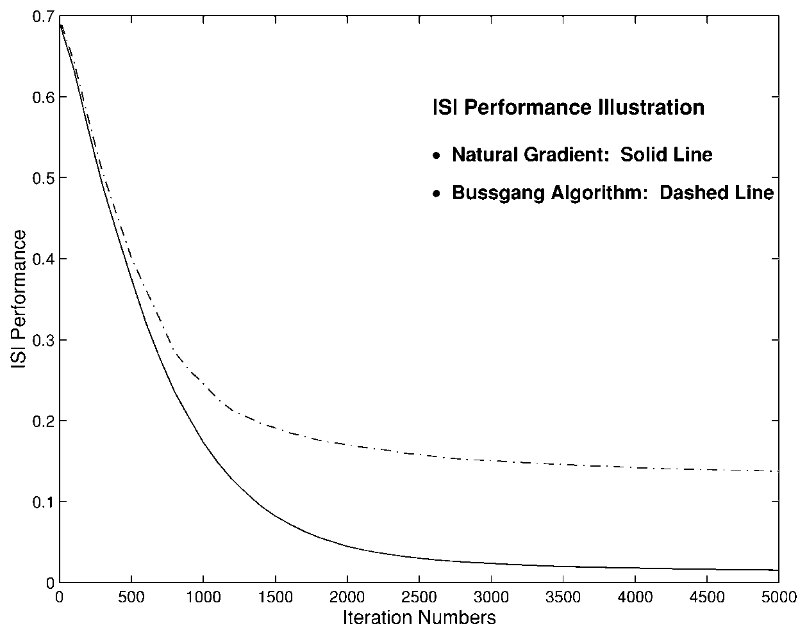

Figure 2. M_{ISI} performance of the natural gradient algorithm.

the mixing model is a 3 channel stable and minimum phase ARMA model, which is randomly chosen by computer. We employ both the natural gradient algorithm and the Bussgang algorithm [20] to train the demixing system. The simulation shows that the natural gradient learning algorithm can easily recover the source signals in the sense of (6).

Figure 2 illustrates 100 trial ensemble average M_{ISI} performance of the natural gradient learning algorithm and the Bussgang algorithm. It is observed that the natural gradient algorithm usually needs less than 2000 iterations to obtain satisfactory results, while the Bussgang algorithm needs more than 20000 iterations since there is a long plateau in the Bussgang learning.

Example 2. In this example, we give a computer simulation with a non-minimum phase mixing model, and compare the natural gradient algorithm with the IIR method in [23]. Assume that source signals are i.i.d quadrature amplitude modulated (QAM). The noise $\mathbf{v}(k)$ is Gaussian and zero mean with a covariance matrix $0.1\mathbf{I}$. The mixing model is a 3 channel non-minimum phase ARMA model, which is randomly generated by computer. Figure 3 plots the coefficients of transfer function $\mathbf{H}(z)$. Since the mixing system is a non-minimum phase system, we cannot find a causal filter to inverse the mixing system. Thus it is not appropriate to apply directly the natural gradient algorithm to the demixing model. In order to surmount the difficulty, we decompose the demixing filter in the following form [36]

$$\mathbf{W}(z) = \mathbf{L}(z)\mathbf{R}(z^{-1}), \tag{69}$$

where $\mathbf{L}(z) = \sum_{p=0}^{N} \mathbf{L}_p z^{-p}$ is a causal FIR filter and $\mathbf{R}(z^{-1}) = \sum_{p=0}^{N} \mathbf{R}_p z^{p}$ is an anti-causal FIR filter. We can see that both $\mathbf{L}(z)$ and $\mathbf{R}(z^{-1})$ are one-sided FIR filter. Refer to [36] for more details. Now, we apply the natural gradient algorithm both to $\mathbf{L}(z)$ and $\mathbf{R}(z^{-1})$.

Figure 4 illustrates the coefficients of the global transfer function $\mathbf{G}(z) = \mathbf{W}(z) * \mathbf{H}(z)$ after 3000 iterations. Figure 5 shows the output signal constellations of the demixing system by using the natural gradient algorithm (42) at three different time intervals: the first row plots the output signals from iteration $k = 1$ to 200, the second row plots those from $k = 1001$ to 1200 and the third row plots those from $k = 2001$ to 2200, respectively. Figure 6 shows the corresponding output signal constellations of the demixing system by using the IIR method in [23]. It is observed that the method proposed in this paper has better performance than the one in [23] both in convergence rate and stability. From a large number of computer simulations, we also see that the algorithm (42) has a wider convergent range than the one in [23]. It is worth noting that the output signals may converge to the characteristic QAM

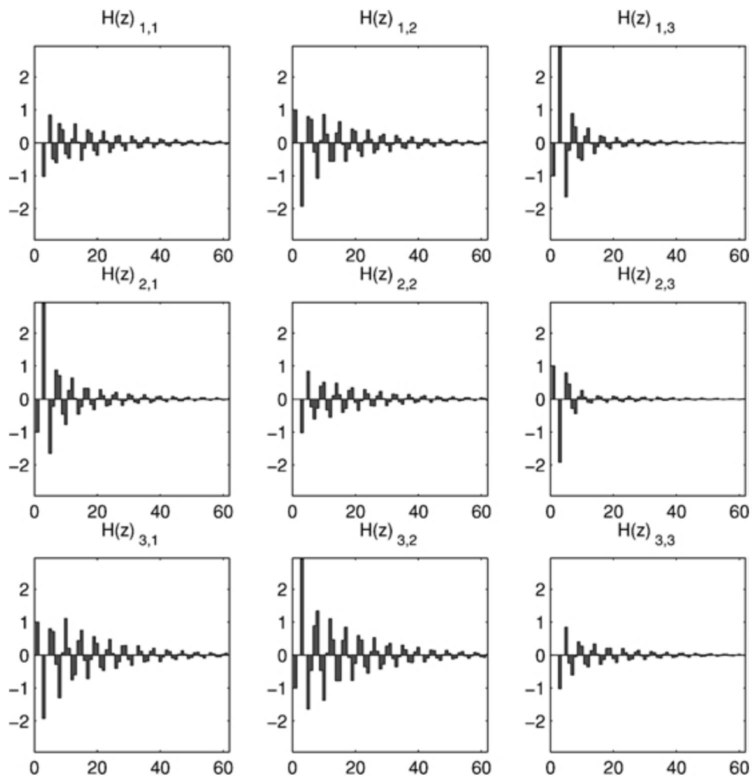

Figure 3. The transfer function $\mathbf{H}(z)$ of the mixing filter.

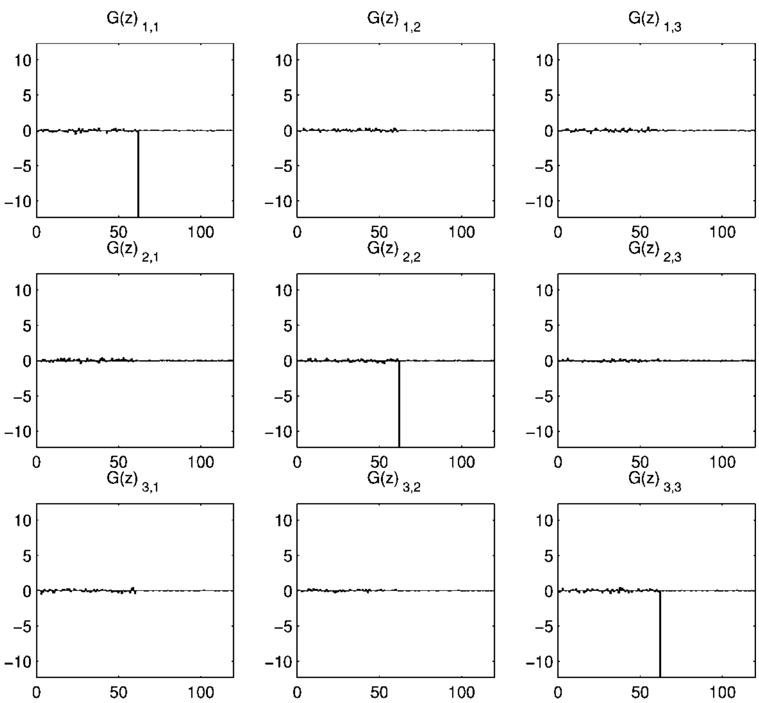

Figure 4. The global transfer function $\mathbf{G}(z)$ after convergence.

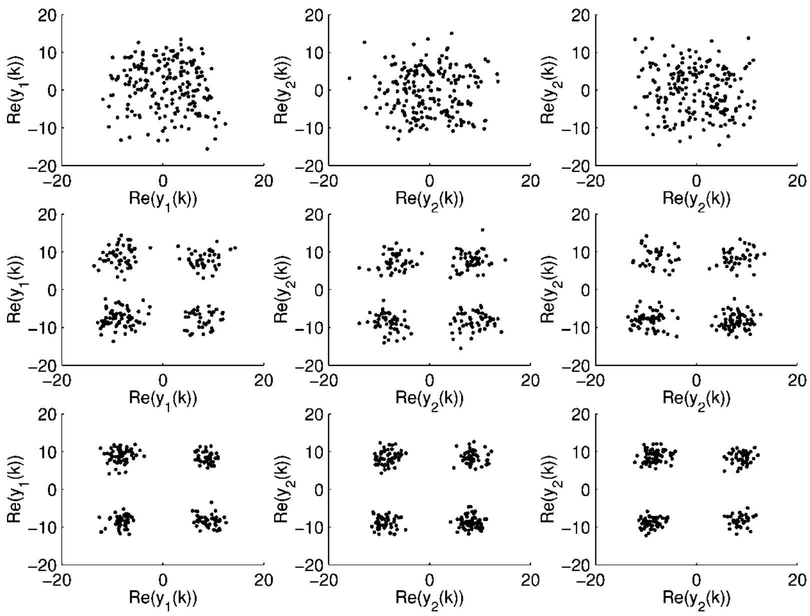

Figure 5. The output signal constellations of the demixing system by the algorithm (42) to treat the nonminimum phase filter.

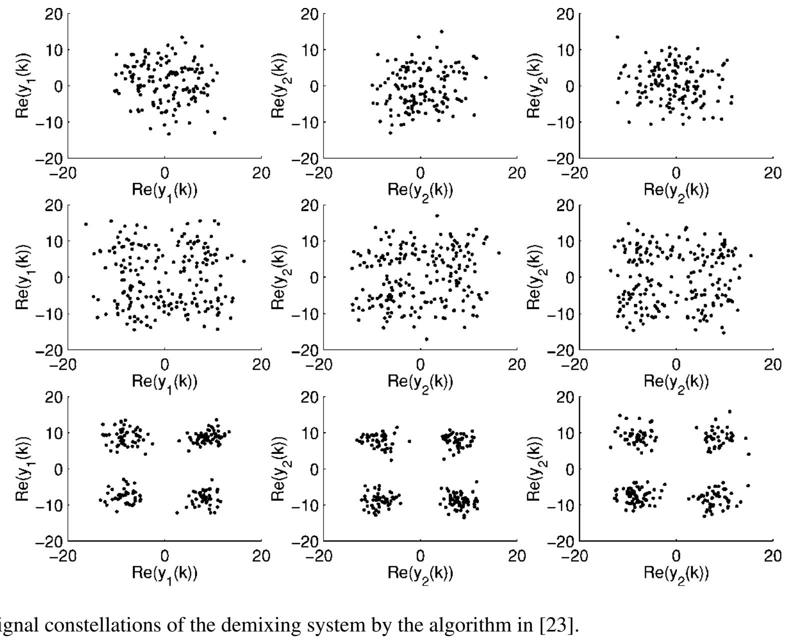

Figure 6. The output signal constellations of the demixing system by the algorithm in [23].

constellation, up to an amplitude and phase rotation factors ambiguities.

9. Conclusion

In this paper we have investigated geometrical structures of the FIR manifold and developed an efficient learning algorithm for blind deconvolution. The demixing filter manifold is not in Euclidean but in Riemannian space. In the Riemannian space, the steepest ascent direction is defined by the natural gradient, which is related to the Riemannian metric. In this paper, we deduce the Riemannian metric by using the isometric property of the Lie group and develop a new approach for deriving the natural gradient. Although the derivation is complicated, the expression of the natural gradient is concise and easy to implement for blind deconvolution. Computer simulations show that the natural gradient algorithm has much better performance of learning than the ordinary gradient algorithm. It should be noted here that the natural gradient approach can also be used to treat the nonminimum phase filter.

Appendix: Derivation of Cost Function

In order to derive the cost function for blind deconvolution, we have to calculate the entropy $H(\mathbf{y}, \mathbf{W})$ of $\mathbf{y}$. Consider $\mathbf{y}$ as a stochastic process $\mathbf{y}(k)$, $k = 1, 2, \ldots$. The entropy is defined by [40]

$$H(\mathbf{y}) = \lim_{L\to\infty} \frac{1}{L} H(\mathbf{y}(1), \ldots, \mathbf{y}(L)). \tag{70}$$

Lemma 3. *If* $\mathbf{s} = \{\mathbf{s}(k)\}$ *is a stochastic process, and* $\mathbf{s}(1), \mathbf{s}(2), \ldots$ *are i.i.d. random variables, then*

$$\begin{aligned} H(\mathbf{s}) &= \lim_{L\to\infty} \frac{H(\mathbf{s}(1), \mathbf{s}(2), \ldots, \mathbf{s}(L))}{L} \\ &= \lim_{L\to\infty} \frac{LH(\mathbf{s}(1))}{L} = H(\mathbf{s}(1)). \end{aligned} \tag{71}$$

Lemma 4. *If* $\mathbf{x}$ *is an n-dimensional vector of random variables and* $\mathbf{A} \in \mathbf{R}^{n\times n}$ *is a nonsingular matrix, then*

$$H(\mathbf{A}\mathbf{x}) = \log|\det(\mathbf{A})| + H(\mathbf{x}). \tag{72}$$

Now we consider the global transfer function in Lie group sense,

$$\mathbf{y}(k) = \mathbf{G}(z)\mathbf{s}(k) = [\mathbf{W}(z)\mathbf{H}(z)]_N \mathbf{s}(k), \tag{73}$$

We consider n observations $\{x_i(k)\}$ and n output signals $\{y_i(k)\}$ with length L.

$$\mathcal{S}(L) = \begin{bmatrix} \mathbf{s}(1) \\ \mathbf{s}(2) \\ \vdots \\ \mathbf{s}(L) \end{bmatrix}, \quad \mathcal{Y}(L) = \begin{bmatrix} \mathbf{y}(1) \\ \mathbf{y}(2) \\ \vdots \\ \mathbf{y}(L) \end{bmatrix}, \tag{74}$$

Since the asymptotic property of $\mathbf{y}$ do not depend on the initial conditions of $\mathbf{s}(k)$, we set $\mathbf{s}(k) = \mathbf{0}$, for $k = 0, -1, \ldots, -N+1$. Under this condition, the vector $\mathcal{Y}(L)$ is a linear transformation of $\mathcal{S}(L)$,

$$\mathcal{Y}(L) = \mathcal{W}\mathcal{S}(L), \tag{75}$$

where $\mathcal{W}$ is given by

$$\mathcal{W} = \begin{bmatrix} \mathbf{G}_0 & \mathbf{0} & \cdots & \mathbf{0} & \mathbf{0} \\ \mathbf{G}_1 & \mathbf{G}_0 & \cdots & \mathbf{0} & \mathbf{0} \\ \vdots & \mathbf{G}_1 & \ddots & \vdots & \vdots \\ \mathbf{0} & \mathbf{0} & \ddots & \mathbf{G}_0 & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \cdots & \mathbf{G}_1 & \mathbf{G}_0 \end{bmatrix}. \tag{76}$$

Here we presume that the delay length N of FIR filter $\mathbf{W}(z)$ is much smaller than L. Hence we have

$$H(\mathcal{Y}(L)) = (\log|\det(\mathcal{W})| + H(\mathcal{S}(L))). \tag{77}$$

According to the lemma 3 and 4, we calculate the entropy $H(\mathbf{y})$

$$\begin{aligned} H(\mathbf{y}) &= \lim_{L\to\infty} \frac{1}{L} H(\mathbf{y}(1), \ldots, \mathbf{y}(L)) \\ &= \lim_{L\to\infty} \frac{1}{L} H(\mathcal{Y}(L)) \\ &= \lim_{L\to\infty} \frac{1}{L} (\log|\det(\mathcal{W})| + H(\mathcal{S}(L))) \\ &= \log|\det(\mathbf{G}_0)| + H(\mathbf{s}(1)) \\ &= \log|\det(\mathbf{W}_0)| + (\log|\det(\mathbf{H}_0)| + H(\mathbf{s}(1))). \end{aligned} \tag{78}$$

The last two terms can be removed from the cost function for blind deconvolution, because they do not depend on the demixing filter $\mathbf{W}(z)$.

References

1. C. Jutten and J. Herault, "Blind Separation of Sources, Part I: An Adaptive Algorithm Based on Neuromimetic Architecture," *Signal Processing*, vol. 24, 1991, pp. 1–10.
2. S. Haykin (Ed.), *Blind Deconvolution*, Englewood Cliffs, NJ: Prentice-Hall, 1994.
3. T.W. Lee, A.J. Bell, and R. Lambert, "Blind Separation of Delayed and Convolved Sources," in *Advances in Neural Information Processing Systems*, vol. 9, Cambridge, MA: MIT Press, 1996, pp. 758–764.
4. K. Torkkola, "Blind Separation of Convolved Sources Based on Information Maximization," in *Proc. of the 1996 IEEE Workshop Neural Networks for Signal Processing (NNSP'96)*, vol. 6,

S. Usui, Y. Tohkura, S. Katagiri, and E. Wilson (Eds.), New York, NY: IEEE Press, 1996, pp. 423–432.

5. W. Kasprzak and A. Cichocki, "Hidden Image Separation from Incomplete Image Mixtures by Independent Component Analysis," in *Proc. of 13th Int. Conf. on Pattern Recognition (ICPR'96)*, IEEE Computer Society Press, 1996, pp. 394–398.
6. D. Donoho, *On Minimum Entropy Deconvolution*, Applied Times Series Analysis II, San Diego, CA: Academic Press, 1981.
7. S. Makeig, A.J. Bell, Tzyy-Ping Jung, and T.J. Sejnowski, "Independent Component Analysis of Electroencephalographic Data," in *Advances in Neural Information Processing Systems*, D.S. Touretzky, M.C. Mozer, and M.E. Hasselmo (Eds.), vol. 8, Cambridge, MA: The MIT Press, 1996, pp. 145–151.
8. R. Vigario, V. Jousmaki, M. Hamalainen, R. Hari, and E. Oja, "Independent Component Analysis for Identification of Artifacts in Magnetoencephalographic Recordings," in *Advances in Neural Information Processing System*, vol. 10, 1997, pp. 229–235.
9. P. Comon, "Independent Component Analysis: A New Concept?" *Signal Processing*, vol. 36, 1994, pp. 287–314.
10. A. Cichocki and R. Unbehauen, "Robust Neural Networks with On-line Learning for Blind Identification and Blind Separation of Sources," *IEEE Trans. Circuits and Systems I: Fundamentals Theory and Applications*, vol. 43, no. 11, 1996, pp. 894–906.
11. A. Cichocki, R. Unbehauen, and E. Rummert, "Robust Learning Algorithm for Blind Separation of Signals," *Electronics Letters*, vol. 30, no. 17, 1994, pp. 1386–1387.
12. S. Amari, A. Cichocki, and H.H. Yang, "A New Learning Algorithm for Blind Signal Separation," in *Advances in Neural Information Processing Systems (NIPS'95)*, vol. 8, G. Tesauro, D.S. Touretzky, and T.K. Leen (Eds.), Cambridge, MA: The MIT Press, 1996, pp. 757–763.
13. J.-F. Cardoso and B. Laheld, "Equivariant Adaptive Source Separation," *IEEE Trans. Signal Processing*, vol. SP-43, 1996, pp. 3017–3029.
14. A. Hyvärinen and E. Oja, "A Fast Fixed Point Algorithm for Independent Component Analysis," *Neural Computation*, vol. 9, 1997, pp. 1483–1492.
15. S. Amari, T. Chen, and A. Cichocki, "Stability Analysis of Adaptive Blind Source Separation," *Neural Networks*, vol. 10, 1997, pp. 1345–1351.
16. S. Amari and J.-F. Cardoso, "Blind Source Separation—Semiparametric Statistical Approach," *IEEE Trans. Signal Processing*, vol. 45, 1997, pp. 2692–2700.
17. R.W. Lucky, "Techniques for Adaptive Equalization of Digital Communication Systems," *Bell Sys. Tech. J.*, vol. 45, 1966, pp. 255–286.
18. Y. Sato, "Two Extensional Applications of the Zero-Forcing Equalization Method," *IEEE Trans. Commun.*, vol. COM-23, 1975, pp. 684–687.
19. J.R. Treichler and B.G. Agee, "A New Approach to Multipath Correction of Constant Modulus Signals," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-31, 1983, pp. 349–372.
20. S. Bellini, "Bussgang Techniques for Blind Deconvolution and Equalization," in S. Haykin (Ed.), *Blind Deconvolution*, Englewood Cliffs, NJ: Prentice Hall, 1994, pp. 8–59,
21. O. Shalvi and E. Weinstein, "New Criteria for Blind Deconvolution of Nonminimum Phase Systems (Channels)," *IEEE Trans. Info. Theory*, vol. 36, 1990, pp. 312–321.
22. J.A. Cadzow, "Blind Deconvolution vis Cumulant Extrema," *IEEE Signal Processing Mag.*, vol. 13, 1996, pp. 24–42.
23. S. Amari, S. Douglas, A. Cichocki, and H. Yang, "Novel On-line Algorithms for Blind Deconvolution using Natural Gradient Approach," in *Proc. 11th IFAC Symposium on System Identification, SYSID'97*, Kitakyushu, Japan, July 8–11, 1997, pp. 1057–1062.
24. A.J. Bell and T.J. Sejnowski, "An Information Maximization Approach to Blind Separation and Blind Deconvolution," *Neural Computation*, vol. 7, 1995, pp. 1129–1159.
25. Y. Hua, "Fast Maximum Likelihood for Blind Identification of Multiple FIR Channels," *IEEE Trans. Signal Processing*, vol. 44, 1996, pp. 661–672.
26. K. Abed-Meraim, J.F. Cardoso, A. Gorokhov, P. Loubaton, and E. Moulines, "On Subspace Methods for Blind Identification of SIMO-FIR Systems," *IEEE Trans. on Signal Processing*, vol. 45, 1997, pp. 42–56.
27. A. Gorokhov and P. Loubaton, "Blind Identification of MIMO-FIR System: A Generalized Linear Prediction Approach," *Signal Processing*, vol. 73, 1999, pp. 105–124.
28. E. Moulines, P. Duhamel, J.F. Cardoso, and S. Mayrargue, "Subspace Methods for the Blind Identification of Multichannel FIR Filters," *IEEE Trans. Signal Processing*, vol. 43, 1995, pp. 516–525.
29. L. Tong, G. Xu, and T. Kailath, "Blind Identification and Equalization Base on Second-Order Statistics: A Time Domain Approach," *IEEE Trans. Information Theory*, vol. 40, 1994, pp. 340–349.
30. J.K. Tugnait and B. Huang, "Multistep Linear Predictors-Based Blind Identification and Equalization of Multiple-Input Multiple-Output Channels," *IEEE Trans. on Signal Processing*, vol. 48, 2000, pp. 26–38.
31. Y. Hua and J.K. Tugnait, "Blind Identifiability of FIR-MIMO Systems with Colored Input using Second Order Statistics," *IEEE Signal Processing Letters*, vol. 7, 2000, pp. 348–350.
32. L. Zhang and A. Cichocki, "Blind Separation of Filtered Source using State-Space Approach," in *Advances in Neural Information Processing Systems*, vol. 11, M.S. Kearns, S.A. Solla, and D.A. Cohn (Eds.), Cambridge, MA: MIT Press, 1999, pp. 648–654.
33. H. Yang, S. Amari, and A. Cichocki, "Information-Theoretic Approach to Blind Separation of Sources in Non-Linear Mixture," *Signal Processing*, vol. 64, no. 3, 1998, pp. 291–300.
34. S. Amari, "Natural Gradient Works Efficiently in Learning," *Neural Computation*, vol. 10, 1998, pp. 251–276.
35. L. Zhang, A. Cichocki, and S. Amari, "Natural Gradient Algorithm for Blind Separaiton of Overdetermined Mixture with Additive Noise," *IEEE Signal Processing Letters*, vol. 6, no. 11, 1999, pp. 293–295.
36. L. Zhang, A. Cichocki, and S. Amari, "Multichannel Blind Deconvolution of Nonminimum Phase Systems using Information Backpropagation," in *Proceedings of the Fifth International Conference on Neural Information Processing (ICONIP'99)*, Perth, Australia, Nov. 16–20, 1999, pp. 210–216.
37. J. Schott, *Matrix Analysis for Statistics*, New York: John Wiley & Sons, 1991.
38. H. Yang, "Serial Undating Rule for Blind Separation Derived from the Method of Scoring," *IEEE Trans. Signal Processing*, vol. 47, no. 8, 1999, pp. 2279–2295.
39. H. Yang and S. Amari, "Adaptive On-line Learning Algorithms for Blind Separation: Maximum Entropy and Minimal Mutual Information," *Neural Comput.*, vol. 9, 1997, pp. 1457–1482.

40. T. Cover and J. Thomas, *Elements of Inofrmation Theory*, New York: John Wiley & Sons, 1991.

Liqing Zhang received the B.S. degree in Mathematics from Hangzhou University, China, in 1983 and Ph.D. degree in Computer Science from Zhongshan University, China, in 1988. After graduation, he joined South China University of Technology, as a lecturer. From 1990 to 1995, he served as Associate Professor and in 1995, he was promoted to Full Professor in College of Electronic and Information Engineering, South China University of Technology. He is currently working in the RIKEN Brain Science Institute, Japan as a Research Scientist.

His research interests include Learning Theory, Adaptive Systems, Biologically-Inspired Neural Networks and Computational Neuroscience. He is author or coauthor of more than 70 scientific papers in journals and conference proceedings.

Andrzej Cichocki received the M.Sc. (with honors), Ph.D., and Habilitate 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.

Since 1995 he has been working in the Brain Science Institute RIKEN (Japan), as a team leader for laboratory for Open Information Systems and currently as a head of laboratory for Advanced Brain Signal Processing. His current research interests include neural networks, biomedical signal and image processing, especially analysis and processing of multi-sensory EEG/MEG, PET, and array of microphones data.

Shun-ichi Amari graduated from the University of Tokyo in 1958, majoring in mathematical engineering. He is Professor-Emeritus at the University of Tokyo and Vice Director of RIKEN Brain Science Institute, Saitama, Japan. He has been engaged in research in wide areas of mathematical engineering and applied mathematics, such as topological network theory, differential geometry of continuum mechanics, pattern recognition, mathematical foundations of neural networks, and information geometry.

Dr. Amari served as President of the International Neural Network Society, Council member of Bernoulli Society for Mathematical Statistics and Probability Theory, and Vice President of the Institute of Electrical, Information and Communication Engineers. He was founding Coeditor-in Chief of Neural Networks. He has received the Japan Academy Award, IEEE Neural Networks Pioneer Award, IEEE Emanuel R. Piore Award, Caianiello Memorial Award, Neurocomputing best paper award, and IEEE Signal Processing Society best award, among many others.