Multichannel Blind Deconvolution of Nonminimum-Phase Systems Using Filter Decomposition

Liqing Zhang, Andrzej Cichocki, *Member, IEEE*, and Shun-ichi Amari, *Fellow, IEEE*

Abstract—In this paper, we present a new filter decomposition method for multichannel blind deconvolution of nonminimum-phase systems. With this approach, we decompose a doubly finite impulse response filter into a cascade form of two filters: a causal finite impulse response (FIR) filter and an anticausal FIR filter. After introducing a Lie group to the manifold of FIR filters, we discuss geometric properties of the FIR filter manifold. Using the nonholonomic transform, we derive the natural gradient on the FIR manifold. By simplifying the mutual information rate, we present a very simple cost function for blind deconvolution of nonminimum-phase systems. Subsequently, the natural gradient algorithms are developed both for the causal FIR filter and for the anticausal FIR filter. Simulations are presented to illustrate the validity and favorable learning performance of the proposed algorithms.

Index Terms—Blind deconvolution, independent component analysis, natural gradient, nonmimimum-phase systems.

I. INTRODUCTION

RECENTLY, blind deconvolution has attracted considerable attention in signal processing and neural network societies. The objective of blind deconvolution is to recover the original source signals and/or to estimate the channel filters, given the noisy measurements (the sensor signals), without the knowledge of the system transfer function. This kind of problem arises in various applications, such as digital telecommunications, speech enhancement, and biomedical signal processing. A number of methods have been developed to deal with the blind deconvolution problem. These methods include the Bussgang algorithms [1]–[4], the higher order statistics approach [5]–[11], and the second order statistical approach [12]–[16]. The high-order statistical methods explicitly exploit the high-order spectra to estimate the channel

transfer function and source signals. The second-order statistical approach has also been brought into consideration since Tong *et al.* [12] first presented the blind identifiability of single-input multiple-output (SIMO) linear channels from only the second-order statistics. Identifiability of blind deconvolution has been further discussed for SIMO systems [13] and multiple-input multiple-output (MIMO) systems [14]–[16]. For further information, see the two recent books in [17] and [18].

Generally speaking, second-order statistical methods rely on the separability of noise and signal subspaces, which requires some prior knowledge of the length of the unknown channels to be identified. When noise is present, high-order statistical methods can be effective under appropriate initialization but may suffer from slow convergence and local convergence.

The natural gradient, which was developed by Amari *et al.* [19], and the relative gradient developed by Cardoso *et al.* [20], improve learning efficiency in blind separation and blind deconvolution [21]. For doubly infinite impulse response (IIR) filters, the natural gradient algorithm was developed by Amari *et al.* [22]. However, in practice, it is necessary to use doubly finite impulse response (FIR) filters as demixing models. In contrast to doubly IIR filters, the doubly FIR filters do not have self-closed multiplication and inverse operations in the manifold of FIR filters with a fixed length. In general, the product of two FIR filters with a given length makes a new filter with a greater length, as does the inverse operation.

The main objective of this paper is to develop an efficient learning algorithm for training the doubly FIR filters for blind deconvolution. First, a doubly noncausal FIR filter is decomposed into two one-sided FIR filters, an anti-causal FIR filter, and a causal FIR filter. With this filter decomposition, we derive a very simple cost function for multichannel blind deconvolution of nonmimimum-phase systems. Some geometrical structures, such as the Lie group, on the differential manifold of one-sided FIR filters are discussed, and the natural gradient on the differential manifold is derived by introducing a nonholonomic transform. Then, the natural gradient learning algorithm is presented to train both the causal FIR filter and anticausal FIR filter. Results from simulations are presented to illustrate the validity and learning performance of the proposed algorithms.

The filter decomposition has two main purposes. One is to keep the demixing filter stable during training, and the other is to use the natural gradient algorithm for training one-sided FIR filters efficiently. We should stress that the filter decomposition approach is not only applicable to blind deconvolution but also to other identification problems of noncausal FIR filters.

II. PROBLEM FORMULATION

For a convolutive mixing model, we consider multichannel, linear time-invariant (LTI), and noncausal systems of the form

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

where $\mathbf{H}_p$ is an $n \times n$-dimensional matrix of mixing coefficients at time-lag p, which is called the impulse response at time p, $\mathbf{s}(k) = [s_1(k), \cdots, s_n(k)]^T$ is an n-dimensional vector of source signals with mutually independent components, and $\mathbf{x}(k) = [x_1(k), \cdots, x_n(k)]^T$ is the vector of the sensor signals. The objective of multichannel blind deconvolution is to retrieve the source signals using only the sensor signals $\mathbf{x}(k)$ and certain knowledge of the source signal distributions and statistics. To fulfil this task, we employ another multichannel LTI system as a demixing model

$$\mathbf{y}(k) = \sum_{p=-\infty}^{\infty} \mathbf{W}_p \mathbf{x}(k-p) \tag{2}$$

where $\mathbf{y}(k) = [y_1(k), \cdots, y_n(k)]^T$ is an n-dimensional vector of the outputs, and $\mathbf{W}_p$ is an $n \times n$-dimensional coefficient matrix at time lag p. We use the following notations for the mixing filter and demixing filter:

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

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

where z is the z-transform variable, as well as the back-shift operator in the sense $z^{-1}\mathbf{x}(k) = \mathbf{x}(k-1)$. Thus, the mixing and demixing model can be simply rewritten as

$$\mathbf{x}(k) = [\mathbf{H}(z)]\mathbf{s}(k), \quad \mathbf{y}(k) = [\mathbf{W}(z)]\mathbf{x}(k). \tag{5}$$

In order to ensure that the mixing filter is invertible, we impose the following constraints on the mixing filter.

1) The filter $\mathbf{H}(z)$ is stable, that is, the impulse response satisfies the absolutely summable condition

$$\sum_{p=-\infty}^{\infty} \|\mathbf{H}_p\|_F < \infty \tag{6}$$

where $\| \cdot \|_F$ denotes the Frobenius matrix norm.

2) The filter $\mathbf{H}(z)$ is of full rank on the unit circle $\gamma = \{z | |z| = 1\}$; this implies that it has no Smith zeros on the unit circle.

The set of filters $\mathbf{H}(z)$, satisfying the above two conditions, is referred to the invertible class as $\mathcal{A}$. The class $\mathcal{A}$ is closed with respect to the convolution and inverse operations. In the following discussion, we assume, for simplicity, that both the mixing and demixing filters are in the class $\mathcal{A}$. It is worth noting that the conditions imposed on the mixing filter are quite general. Here, we do not assume any other conditions, such as irreducibility. Such irreducibility of the mixing filter is necessary in subspace-based methods [12]–[16].

The global transfer function is defined by $\mathbf{G}(z) = \mathbf{W}(z)\mathbf{H}(z)$. The blind deconvolution task is to find a demixing filter $\mathbf{W}(z)$ such that

$$\mathbf{G}(z) = \mathbf{W}(z)\mathbf{H}(z) = \mathbf{P}\mathbf{\Lambda}\mathbf{D}(z) \tag{7}$$

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

In practice, it is necessary to use a doubly finite multichannel filter as a demixing model

$$\mathbf{W}(z) = \sum_{p=-N}^{N} \mathbf{W}_p z^{-p} \tag{8}$$

where N is a given positive integer. We denote the set of all doubly FIR filters of length N as $\mathcal{M}(N, N)$

$$\mathcal{M}(N, N) = \left\{ \mathbf{W}(z) | \mathbf{W}(z) = \sum_{p=-N}^{N} \mathbf{W}_p z^{-p} \right\}. \tag{9}$$

In general, the product of two filters in $\mathcal{M}(N, N)$ makes a filter with length $2N$, which no longer belongs in $\mathcal{M}(N, N)$. This makes it difficult to introduce the natural gradient on the manifold of doubly FIR filters.

Two key issues need to be resolved in training the noncausal demixing filter. First, how is the demixing filter $\mathbf{W}(z)$ kept stable during training? In other words, the trained demixing filter $\mathbf{W}_N(z)$ needs to be kept stable when its filter length N tends to infinity. Second, what learning algorithms should be used to efficiently train the demixing filter?

In attempting to seek a stable demixing filter, we suggest the decomposition of the doubly FIR filter into two one-sided FIR filters: one a causal FIR filter and the other an anticausal filter. For the one-sided causal FIR filter, the natural gradient algorithm has been developed [23]. However, such a approach cannot be directly used to train noncausal filters. In this paper, we will apply the filter decomposition and nonholonomic transform to solve the training problem of noncausal filters.

III. FILTER DECOMPOSITION

In order to explore the geometric structures on the manifold $\mathcal{M}(N, N)$ and develop an efficient learning algorithm for $\mathbf{W}(z)$, we present a novel filter decomposition approach and define operations for one-sided FIR filters in the Lie group framework.

For single channel blind equalization, a filter decomposition approach was proposed by Labat *et al.* [24]. They presented a novel unsupervised (blind) adaptive decision feedback equalizer via a cascade of four devices, whose main components are a purely recursive filter and a transversal filter. Nandi *et al.* [25] proposed alternative recursive filter structures for equalization of severely distorted channels by using decomposing the equalizer into a cascade of a recursive prewhitening filter adapted with second-order statistics and a phase equalizer adapted with higher order statistics. The main purpose of the cascade is to split the difficult task of unsupervised equalization into several, but easier, subtasks.

In this paper, we apply the filter decomposition approach to multichannel blind deconvolution problems. The purpose of the filter decomposition is to find an efficient way to train the demixing filter. To begin, let us decompose a noncausal filter $\mathbf{W}(z)$ in $\mathcal{M}(N, N)$ into a cascade form of two FIR filters

$$\mathbf{W}(z) = \mathbf{L}(z)\mathbf{R}(z^{-1}) \tag{10}$$

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 anticausal FIR filter with a constraint $\mathbf{R}_0 = \mathbf{I}$; $\mathbf{I}$ is the identity matrix. The coefficients of three filters satisfy the following relations:

$$\mathbf{W}_k = \sum_{p-q=k,\, 0 \leq p,\, q \leq N} \mathbf{L}_p \mathbf{R}_q, \quad \text{for } k = -N, \cdots, N. \tag{11}$$

The blind deconvolution process presented in this paper is illustrated in Fig. 1. With this decomposition, it becomes much easier to discuss the invertibility of doubly finite multichannel filters in the Lie group sense, which will be introduced in the next section.

Directly analyzing the structure of the inverse filter $\mathbf{H}^{-1}(z)$, we will find that it is plausible to use a cascade form of a one-sided filter and an anticausal filter as a demixing model. For simplicity, assume that the mixing filter is an FIR filter

$$\mathbf{H}(z) = \sum_{p=0}^{L} \mathbf{H}_p z^{-p}, \quad \det(\mathbf{H}_0) \neq 0. \tag{12}$$

Assume that $\mathbf{H}(z)$ has no null values on the unit circle. Considering the FIR filter $\mathbf{H}(z)$ as a matrix of polynomials of z, we can express the determinant of $\mathbf{H}(z)$ as

$$det(\mathbf{H}(z)) = c\, z^{-L_0} \prod_{p=1}^{L_1}(1 - a_p z^{-1}) \prod_{p=1}^{L_2}(1 - b_p z^{-1}) \tag{13}$$

where c is a nonzero constant, L_0, L_1, and L_2 are certain natural numbers, $0 < |a_p| < 1$ for $p = 1, \cdots, L_1$, and $|b_p| > 1$ for $p = 1, \cdots, L_2$. Usually, a_p and b_p are referred to as the zeros of filter $\mathbf{H}(z)$. If all the zeros are located in the interior of the unit circle on the complex plane, the filter $\mathbf{H}(z)$ is minimum phase. Otherwise, the filter $\mathbf{H}(z)$ is nonminimum phase. Now, the inverse of $\mathbf{H}(z)$ can be calculated by

$$\mathbf{H}^{-1}(z) = \mathbf{H}^{\#}(z) det(\mathbf{H}(z))^{-1} \tag{14}$$

where $\mathbf{H}^{\#}(z)$ is the adjoint matrix of $\mathbf{H}(z)$. On the other hand, the components of $det(\mathbf{H}(z))^{-1}$ can be expanded in the following way:

$$(1 - a_p z^{-1})^{-1} = \sum_{q=0}^{\infty} a_p^q z^{-q} \tag{15}$$

$$(1 - b_p z^{-1})^{-1} = -b_p^{-1} z \sum_{q=0}^{\infty} b_p^{-q} z^q. \tag{16}$$

This means that if the zero a_p is located in the interior of the unit circle, we can expand the factor $[1 - a_p z^{-1}]^{-1}$ as a causal
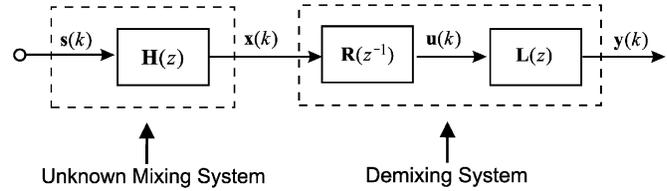
Fig. 1.   Illustration of filter decomposition for blind deconvolution.

filter; otherwise, we need to expand the factor $[1 - b_q z^{-1}]^{-1}$ as an anticausal filter. Hence, the inverse of $\mathbf{H}(z)$ is expressed by

$$\mathbf{H}^{-1}(z) = c^{-1} z^{L_0 + L_2} \prod_{p=1}^{L_2} (-b_p)^{-1} \mathbf{L}(z)\mathbf{R}(z^{-1}) \tag{17}$$

where

$$\mathbf{L}(z) = \sum_{r=0}^{\infty} \mathbf{L}_r z^{-r} = \mathbf{H}^{\#}(z) \prod_{p=1}^{L_1}(1 - a_p z^{-1})^{-1} \tag{18}$$

$$\mathbf{R}(z^{-1}) = \sum_{r=0}^{\infty} \mathbf{R}_r z^{r} = \prod_{p=1}^{L_2} \sum_{q=0}^{\infty} b_p^{-q} z^{q} \mathbf{I}. \tag{19}$$

It is easily seen from (18) and (19) that $\|\mathbf{L}_r\|_F$ and $\|\mathbf{R}_r\|_F$ decay exponentially to zero as r tends to infinity. The asymptotic decay rate is dominated by the poles closest to the unit circle. Generally, the inverse filter of $\mathbf{H}(z)$ is a noncausal filter of infinite length. Due to the asymptotic decay of these two one-sided filters, we can use two one-sided FIR filters to approximate filters $\mathbf{L}(z)$ and $\mathbf{R}(z^{-1})$, respectively. The approximation will cause a model error in blind deconvolution. If we make the length of the demixing filter sufficiently large, however, the model error will become negligible due to the decay properties of two filters $\mathbf{L}(z)$ and $\mathbf{R}(z^{-1})$. The length of the demixing filter depends on the zeros of the mixing filter that are closest to the unit circle. If the zeros of the mixing filter are close to the unit circle, then we need to choose a longer filter length according to the decay properties of the demixing filter.

From the above analysis, we see that the inverse of the mixing filter can be expressed in the form of filter decomposition (10). It should be pointed that if the FIR filter $\mathbf{H}(z)$ is a minimum-phase filter, then the inverse $\mathbf{H}^{-1}(z)$ will be a causal filter. In general, the filter $\mathbf{L}(z)$ is used to invert the minimum-phase portion, and $\mathbf{R}(z^{-1})$ is used to invert the maximum-phase portion of the mixing filter. Therefore, we can further assume that the causal FIR filter $\mathbf{L}(z)$ is minimum phase, i.e., its inverse is also a causal filter. Similarly, we assume that the inverse of the anticausal filter $\mathbf{R}(z^{-1})$ is also an anticausal filter.

In the following sections, we will develop the natural gradient algorithm to adjust the parameters of both the causal filter $\mathbf{L}(z)$ and anticausal filter $\mathbf{R}(z^{-1})$.

IV. GEOMETRIC STRUCTURES ON THE FIR MANIFOLD

In this section, we discuss the geometrical structures on the manifold of FIR filters $\mathbf{L}(z)$. First, we introduce a Lie group to

the manifold of FIR filters in order to define self-closed multiplication and inverse operations. Using the nonholonomic transform, we derive the natural gradient of a cost function defined on the manifold.

In the following discussion, we use the notation $\mathcal{M}(N)$ to denote the subset of $\mathcal{M}(N, 0)$, having the constraint that $\mathbf{L}_0$ is nonsingular. In order to discuss geometrical structures of nonsingular FIR manifold $\mathcal{M}(N)$, we first introduce the tangent space on manifold $\mathcal{M}(N)$. Given an FIR filter $\mathbf{L}(z) \in \mathcal{M}(N)$, a tangent vector, which is denoted by $\mathbf{X}(z)$, is an infinitesimal displacement at $\mathbf{L}(z)$ on manifold $\mathcal{M}(N)$. The set of tangent vectors at a point $\mathbf{L}(z)$ forms a vector space called the tangent space of $\mathcal{M}(N)$ at $\mathbf{L}(z)$. The tangent space of $\mathcal{M}(N)$ at $\mathbf{L}(z)$ is given by

$$\mathcal{T}_{\mathbf{L}(z)} = \left\{ \mathbf{X}(z) | \mathbf{X}(z) = \sum_{p=0}^{N} \mathbf{X}_p z^{-p} \right\} \quad (20)$$

where $\mathbf{X}_p \in R^{n \times n}$, $p = 0, 1, \cdots, N$ are arbitrary $n \times n$ matrices. The tangent space $\mathcal{T}_{\mathbf{L}(z)}$ at a point $\mathbf{L}(z)$ on manifold $\mathcal{M}(N)$ is intuitively the vector space obtained by locally linearizing $\mathcal{M}(N)$ around $\mathbf{L}(z)$.

The dimension of the tangent space $\mathcal{T}_{\mathbf{L}(z)}$ is the same as that of the manifold $\mathcal{M}(N)$. The tangent space $\mathcal{T}_{\mathbf{L}(z)}$ is an $(N+1)n^2$-dimensional Euclidean space. Because it is different from the filters in the manifold $\mathcal{M}(N)$, the first matrix $\mathbf{X}_0$ of the filter $\mathbf{X}(z) \in \mathcal{T}_{\mathbf{L}(z)}$ is not necessarily nonsingular.

We say that a manifold is Riemannian if it is equipped with the Riemannian metric [26]. There are at least two critical problems needed to be solved in the optimization problem on the Riemannian manifold. One is how to keep the updated filter on the manifold, and the second is to find the optimal search direction during iterations. In the Riemannian manifold, the derivative of the cost function is not the steepest ascent direction if the Riemannian metric is not the identity matrix. The natural gradient has been introduced to define the steepest ascent direction [21]. In the following sections, we will investigate the geometric structures of the nonsingular FIR manifold and find the natural gradient direction for the optimization problem of blind deconvolution.

A. Lie Group

A Lie group [26] is a group that is also a differential manifold such that for any $\mathbf{A}, \mathbf{B} \in \mathcal{M}$, the multiplication $(\mathbf{A}, \mathbf{B}) \rightarrow \mathbf{A} * \mathbf{B}$ and inverse $\mathbf{A} \rightarrow \mathbf{A}^\dagger$ are smooth maps. The Lie group approach has been successfully applied to matrix groups to derive efficient algorithms for optimization problems [21], [27]. Using the uniform properties of the Lie group, the natural gradient was derived on the matrix manifolds, such as nonsingular matrix [21], orthogonal matrix [28], [29], and nonsquare matrix [30]. In this paper, we further apply the Lie group approach to the manifold of FIR filters. The main purpose of introducing a Lie group to manifold $\mathcal{M}(N)$ is to define self-closed operations on the manifold $\mathcal{M}(N)$.

First, we introduce the Lie multiplication of two filters $\mathbf{A}(z)$, $\mathbf{B}(z) \in \mathcal{M}(N)$ in the following way:

$$\mathbf{A}(z) * \mathbf{B}(z) = [\mathbf{A}(z)\mathbf{B}(z)]_N \quad (21)$$

where $[\mathbf{B}(z)]_N$ is the truncating operator such that any terms with orders higher than N in the polynomial $\mathbf{B}(z)$ are omitted. Explicitly, we define the Lie multiplication as

$$\mathbf{A}(z) * \mathbf{B}(z) = \sum_{p=0}^{N} \sum_{q=0}^{p} \mathbf{A}_q \mathbf{B}_{(p-q)} z^{-p}. \quad (22)$$

Once the Lie multiplication is defined, the Lie inverse of a filter $\mathbf{B}(z)$, which is denoted by $\mathbf{B}^\dagger(z)$, is defined by solving the following equation:

$$\mathbf{B}(z) * \mathbf{B}^\dagger(z) = \mathbf{I} \quad (23)$$

where $\mathbf{I}$ is the identity matrix. With a simple calculation, we obtain the explicit expression of the Lie inverse as follows:

$$\mathbf{B}^\dagger(z) = \sum_{p=0}^{N} \mathbf{B}_p^\dagger z^{-p} \quad (24)$$

where $\mathbf{B}_p^\dagger$ $(p = 0, 1, \cdots, N)$ are recurrently defined by

$$\mathbf{B}_0^\dagger = \mathbf{B}_0^{-1}, \quad \mathbf{B}_1^\dagger = -\mathbf{B}_0^\dagger \mathbf{B}_1 \mathbf{B}_0^\dagger \quad (25)$$

$$\mathbf{B}_p^\dagger = -\sum_{q=1}^{p} \mathbf{B}_{p-q}^\dagger \mathbf{B}_q \mathbf{B}_0^\dagger, \quad p = 1, \cdots, N. \quad (26)$$

Now, it is easily verified that both $\mathbf{A}(z) * \mathbf{B}(z)$ and $\mathbf{B}^\dagger$ remain in the manifold $\mathcal{M}(N)$. The multiplication of $\mathbf{A}(z)$ and $\mathbf{B}(z)$ maps $(\mathbf{A}(z), \mathbf{B}(z))$ to $\mathbf{A}(z) * \mathbf{B}(z)$, whose elements are multivariable polynomials. From the definition of the Lie inverse, the elements of Lie inverse $\mathbf{B}^\dagger(z)$ are multivariable rational polynomials. We see that both the Lie multiplication and Lie inverse are smooth maps. Therefore, the manifold $\mathcal{M}(N)$ with the above operations forms a Lie Group. The identity element is $\mathbf{E}(z) = \mathbf{I}$. Moreover, the Lie group has the following properties:

1) $\mathbf{A}(z) * (\mathbf{B}(z) * \mathbf{C}(z)) = (\mathbf{A}(z) * \mathbf{B}(z)) * \mathbf{C}(z)$ (27)
2) $\mathbf{B}(z) * \mathbf{B}^\dagger(z) = \mathbf{B}^\dagger(z) * \mathbf{B}(z) = \mathbf{I}.$ (28)

The result that a left and right inverse of an FIR filter coincide is inherited trivially from the standard result of linear algebra. Here, we give a simple example to illustrate the Lie group operations. Consider $n = 1$, $N = 100$, and $\mathbf{A}(z) = 1 - az^{-1}$, $|a| < 1$. The Lie inverse is $\mathbf{A}^\dagger(z) = \sum_{p=0}^{N} a^p z^{-p}$. It is easy to verify

$$\mathbf{A}(z) * \mathbf{A}^\dagger(z) = [1 - a^{N+1} z^{-N-1}]_N = 1. \quad (29)$$

The geometric interpretation of the inverse filter in the Lie group sense is given as follows. Suppose $\mathbf{L}(z)$ is a minimum-phase FIR filter of length N, and $\mathbf{L}^\dagger(z)$ is its Lie inverse. We see that the product of the two filters $\mathbf{L}(z)\mathbf{L}^\dagger(z)$ in the ordinary sense is a filter with length $2N$, which is not the identity matrix. However, the truncated filter $[\mathbf{L}(z)\mathbf{L}^{-1}(z)]_N = \mathbf{L}(z) * \mathbf{L}^\dagger(z)$ is equal to the identity matrix. If the coefficients of both $\mathbf{L}(z)$ and its $\mathbf{L}^\dagger(z)$ decay exponentially, and their filter lengths are chosen sufficiently large, the truncation error will become negligible.

After defining Lie group, we can easily introduce to the FIR manifold a Riemannian metric. Furthermore, the natural gradient on the FIR manifold can be defined by the Riemannian

metric. For more detailed information about the Riemannian metric of an FIR manifold, see Zhang *et al.* [23]. In this paper, we mostly avoid sophisticated mathematical concepts and derivations and attempt to derive the natural gradient on the FIR filter via a nonholonomic transform. We will further give a geometric interpretation about the nonholonomic transform and natural gradient in the next section.

In the same way, we can introduce a Lie group to the manifold of the anticausal filters, which is denoted as $\mathcal{R}(N) = \{\mathbf{R}(z^{-1})|\mathbf{R}(z^{-1}) = \sum_{p=0}^{N} \mathbf{R}_p z^p,\ \mathbf{R}_0 = \mathbf{I}\}$. For any $\mathbf{P}(z^{-1})$, $\mathbf{Q}(z^{-1}) \in \mathcal{R}(N)$, we define the multiplication of the two filters as

$$\mathbf{P}(z^{-1}) * \mathbf{Q}(z^{-1}) = \sum_{p=0}^{N}\sum_{q=0}^{p}\mathbf{P}_q \mathbf{Q}_{(p-q)} z^p \qquad (30)$$

and the inverse of filter $\mathbf{P}(z^{-1})$ is given by solving the following equation:

$$\mathbf{P}(z^{-1}) * \mathbf{P}^{\dagger}(z^{-1}) = \mathbf{I}. \qquad (31)$$

It should be emphasized here that the Lie inverse $\mathbf{L}^{\dagger}(z)$ of filter $\mathbf{L}(z) \in \mathcal{M}(N)$ still lies in the manifold $\mathcal{M}(N)$, whereas the Lie inverse $\mathbf{R}^{\dagger}(z^{-1})$ is in the manifold $\mathcal{R}(N)$.

B. Natural Gradient

In this section, we introduce the natural gradient on manifold $\mathcal{M}(N)$ via a geometric approach. For a cost function $l(\mathbf{L}(z))$ defined on the Riemannian manifold $\mathcal{M}(N)$, the natural gradient $\tilde{\nabla}l(\mathbf{L}(z))$ is the steepest ascent direction of the cost function $l(\mathbf{L}(z))$. We use the following notations for the derivative of cost function $l(\mathbf{L}(z))$:

$$\frac{\partial l(\mathbf{L}(z))}{\partial \mathbf{L}(z)} = \sum_{p=0}^{N}\frac{\partial (\mathbf{L}(z))}{\partial \mathbf{L}_p}z^{-p} \qquad (32)$$

where $(\partial l(\mathbf{L}(z))/\partial \mathbf{L}_p) = (\partial l(\mathbf{L}(z))/\partial \mathbf{L}_{p,ij})_{n \times n}$, $p = 0, 1, \cdots, N$.

When the parameter space is Euclidean, the natural gradient becomes the ordinary derivative of the cost function. The main idea of this approach is to first define a search direction locally in a certain coordinate system such as the tangent space; then, the search direction is projected back onto the parameter space. Consider that $\mathbf{L}(z)$ is a local coordinate system and that $d\mathbf{L}(z)$ is a variation in the vicinity of $\mathbf{L}(z)$. In order to remove the effect of individual filter $\mathbf{L}(z)$ on the search direction, we define a map between the differential variables $d\mathbf{L}(z)$ and $d\mathbf{X}(z)$ for the blind deconvolution problem

$$d\mathbf{X}(z) = d\mathbf{L}(z) * \mathbf{L}^{\dagger}(z) = \left[d\mathbf{L}(z)\mathbf{L}^{-1}(z)\right]_N \qquad (33)$$

where d is the differential operator. The modified search direction has a uniform property that the tangent vector $d\mathbf{L}(z)$ in $\mathcal{T}_{\mathbf{L}(z)}$ is mapped onto the tangent vector $d\mathbf{X}(z)$ in $\mathcal{T}_{\mathbf{E}(z)}$ at the unit filter $\mathbf{E}(z)$. Thus, the search direction is almost independent of specific filter $\mathbf{L}(z)$ since the effect of $\mathbf{L}(z)$ is removed by $\mathbf{L}^{\dagger}(z)$ in the new search direction.

It should be noted that $d\mathbf{X}(z)$ is a nonholonomic form, which has a definite geometrical meaning and proves to be useful in blind separation algorithms [31]. A differential form $d\mathbf{Z}(z)$ is holonomic if there exists a differential map $\mathcal{G}$ from $\mathbf{L}(z)$ to $\mathbf{Z}(z)$ such that $\mathbf{Z}(z) = \mathcal{G}(\mathbf{L}(z))$. If it is not holonomic, we say that the differential form is nonholonomic.

The differential form $d\mathbf{X}(z)$ is well defined, which represents the modified search direction in the blind deconvolution setting. Now, we first define a search direction for $d\mathbf{X}(z)$. For a Euclidean space, the steepest descent direction is given by the derivative of the cost function, i.e.,

$$\nabla_{\mathbf{X}(z)}l(\mathbf{L}(z)) = \frac{\partial l(\mathbf{L}(z))}{\partial \mathbf{X}(z)}. \qquad (34)$$

Therefore, the gradient descent learning rule for $\mathbf{X}(z)$ is defined by

$$\Delta\mathbf{X}(z) = -\eta\nabla_{\mathbf{X}(z)}l(\mathbf{L}(z)) = -\eta\frac{\partial l(\mathbf{L}(z))}{\partial \mathbf{X}(z)}. \qquad (35)$$

Since the search is performed in the tangent space, we need to project the search direction back onto the original manifold via the nonholonomic transform (33). Recalling the relationship between $d\mathbf{X}(z)$ and $d\mathbf{L}(z)$, we obtain the updating rule for $\mathbf{L}(z)$ as follows:

$$\begin{aligned}\Delta\mathbf{L}(z) &= \Delta\mathbf{X}(z) * \mathbf{L}(z) \\ &= -\eta\frac{\partial l(\mathbf{L}(z))}{\partial \mathbf{X}(z)} * \mathbf{L}(z).\end{aligned} \qquad (36)$$

This search direction is consistent with the natural gradient defined by the Riemannian metric [23]

$$\tilde{\nabla}l(\mathbf{L}(z)) = \frac{\partial l(\mathbf{L}(z))}{\partial \mathbf{X}(z)} * \mathbf{L}(z). \qquad (37)$$

Using this derivation, we have a clear geometric interpretation of the natural gradient on the FIR filter manifold (see Fig. 2). It should be mentioned that the nonholonomic transform is related to the relative gradient and is important in order for learning algorithms to have the equi-variance property [20].

In the following discussion, we will employ the nonholonomic transform technique to derive efficient learning algorithms to train both the causal and anticausal FIR filters.

V. COST FUNCTION FOR NONCAUSAL FIR FILTER

The objective of blind deconvolution is to find a demixing filter $\mathbf{W}(z)$ such that the output of the demixing model is spatially mutually independent and temporally identically independently distributed (i.i.d.). In order to derive a cost function for blind deconvolution, we consider the output signals $y_i = \{y_i(k), k = 1, 2, \cdots\}$, $i = 1, \cdots, n$ as stochastic processes. For any i and k, $y_i(k)$ is considered to be a random variable. In this paper, we employ the Kullback–Leibler divergence to measure the mutual independence of the output signal. Assume that $q(\mathbf{y}(1), \cdots, \mathbf{y}(T))$ is the joint probability density function (pdf) of $\mathbf{y}(1), \cdots, \mathbf{y}(T)$, and $p_i(y_i(k))$ is the marginal
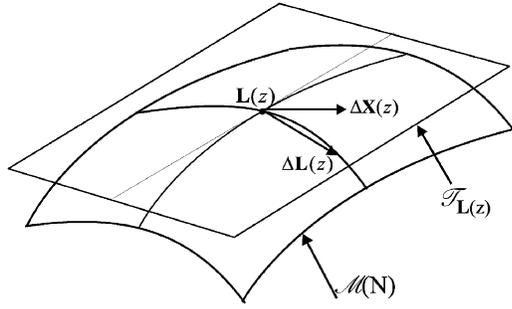
Fig. 2. Geometric interpretation of the natural gradient on the FIR filter manifold.

pdf of $y_i(k)$ for $i = 1, \cdots, n$ and $k = 1, \cdots, T$. The Kullback–Leibler divergence between q and $\prod_{i=1}^{n} \prod_{k=1}^{T} p(y_i(k))$ is defined by

$$
\begin{aligned}
KL = \int q\left(\mathbf{y}(1), \cdots, \mathbf{y}(T)\right) \\
\times \log \left(\frac{q\left(\mathbf{y}(1), \cdots, \mathbf{y}(T)\right)}{\prod_{i=1}^{n} \prod_{k=1}^{T} p\left(y_i(k)\right)} \right) \prod_{i=1}^{n} \prod_{k=1}^{T} dy_i(k)
\end{aligned} \tag{38}
$$

which measures the mutual independency and i.i.d. property between the stochastic processes $y_1, \cdots, y_n$. Since the pdfs q and p are unknown in the blind deconvolution setting, we need to simplify the cost function in order to implement the natural gradient approach. Amari *et al.* [32] introduce the following cost function for online statistical learning:

$$
\begin{aligned}
l\left(\mathbf{y}, \mathbf{W}(z)\right) = -\frac{1}{2\pi j} \oint_{\gamma} \log \left|\det\left(\mathbf{W}(z)\right)\right| z^{-1} dz \\
- \sum_{i=1}^{n} \log p_i\left(y_i(k)\right)
\end{aligned} \tag{39}
$$

where j is the imaginary unit of complex numbers, and the path integral is over the unit circle γ of the complex plane. Pham [33] simplifies the mutual information rate (38) via upbound estimation and derives the same cost function as (39). Taleb *et al.* [34] also present the same cost function for blind inversion of Wiener systems. Now, the problem becomes how to solve the first term of the cost function (39). Amari *et al.* [32] introduce a nonholonomic transform and derive a natural gradient algorithm for IIR filters. However, such a nonholonomic transform is not available for doubly FIR filters because the associative law fails in the manifold of doubly FIR filters.

In this paper, we use the filter decomposition approach and derive a very simple cost function for blind deconvolution. The cost function (39) can be also rewritten as

$$
l\left(\mathbf{y}, \mathbf{W}(z)\right) = -\frac{1}{2\pi} \int_{0}^{2\pi} \log \left|\det\left(\mathbf{W}(e^{j\theta})\right)\right| d\theta - \sum_{i=1}^{n} \log p_i(y_i). \tag{40}
$$

Assume that the demixing filter has the decomposition $\mathbf{W}_N(z) = \mathbf{L}(z)\mathbf{R}(z^{-1})$, where both $\mathbf{L}(z)$ and $\mathbf{R}(z^{-1})$ are one-sided FIR filters. With this filter decomposition, we have

$$
\log \left|\det\left(\mathbf{W}(e^{-j\theta})\right)\right| = \log \left|\det\left(\mathbf{L}(e^{-j\theta})\right)\right| + \log \left|\det\left(\mathbf{R}(e^{j\theta})\right)\right|. \tag{41}
$$

Because $\mathbf{L}(z)$ is assumed minimum phase and causal, we can explicitly calculate the following integral:

$$
\frac{1}{2\pi} \int_{0}^{2\pi} \log \left|\det\left(\mathbf{L}(e^{-j\theta})\right)\right| d\theta = \log \left|\det(\mathbf{L}_0)\right|. \tag{42}
$$

Similarly, because $\mathbf{R}(z^{-1})$ is assumed maximum phase and anticausal, we have

$$
\frac{1}{2\pi} \int_{0}^{2\pi} \log \left|\det\left(\mathbf{R}(e^{j\theta})\right)\right| d\theta = \log \left|\det(\mathbf{R}_0)\right| = 0 \tag{43}
$$

where $\mathbf{R}_0 = \mathbf{I}$. Combining (41) and (43) leads to the following lemma.

Lemma 1: If $\mathbf{W}(z) = \mathbf{L}(z)\mathbf{R}(z^{-1})$, then

$$
\frac{1}{2\pi} \int_{0}^{2\pi} \log \left|\det\left(\mathbf{W}(e^{-j\theta})\right)\right| d\theta = \log \left|\det(\mathbf{L}_0)\right|. \tag{44}
$$

In summary, we obtain the following cost function for blind deconvolution:

$$
l\left(\mathbf{y}, \mathbf{W}(z)\right) = -\log \left|\det(\mathbf{L}_0)\right| - \sum_{i=1}^{n} \log p_i(y_i). \tag{45}
$$

The first term in the cost function is introduced to prevent the matrix $\mathbf{L}_0$ from being singular.

VI. LEARNING ALGORITHMS

In this section, we develop learning algorithms both for the causal filter and anticausal filter via the nonholonomic transform. For the causal FIR filter, we employ the natural gradient algorithm to train $\mathbf{L}(z)$. Now, the key issue is how to train the anticausal filter $\mathbf{R}(z^{-1})$, which is critical to blind deconvolution of nonmimimum-phase systems.

Assume that the demixing filter $\mathbf{W}(z)$ has the decomposition (10). For simplicity, we introduce the intermediate variable $\mathbf{u}$, which is defined as

$$
\mathbf{u}(k) = \left[\mathbf{R}(z^{-1})\right] \mathbf{x}(k) \tag{46}
$$
$$
\mathbf{y}(k) = \left[\mathbf{L}(z)\right] \mathbf{u}(k). \tag{47}
$$

For an $n \times n$ matrix $\mathbf{A} = (a_{ij})_{n \times n}$, we use $\mathrm{tr}(\mathbf{A})$ to represent the trace of matrix $\mathbf{A}$, which is defined by $\mathrm{tr}(\mathbf{A}) = \sum_{i=1}^{n} a_{ii}$. To calculate the natural gradient of the cost function, we consider the differential of the cost function

$$
dl\left(\mathbf{y}, \mathbf{W}(z)\right) = -d\log \left|\det(\mathbf{L}_0)\right| - \sum_{i=1}^{n} dp_i(y_i, \mathbf{W}). \tag{48}
$$

Using the relation $d \log \left|\det(\mathbf{L}_0)\right| = \mathrm{tr}(d\mathbf{L}_0 \mathbf{L}_0^{-1})$ [21], we have

$$
dl\left(\mathbf{y}, \mathbf{W}(z)\right) = -\mathrm{tr}\left(d\mathbf{L}_0 \mathbf{L}_0^{-1}\right) + \boldsymbol{\varphi}(\mathbf{y})^T d\mathbf{y} \tag{49}
$$

where $\boldsymbol{\varphi}(\mathbf{y}) = (\varphi_1(y_1), \cdots, \varphi_n(y_n))^T$ is the vector of nonlinear activation functions, which is defined by

$$\varphi_i(y_i) = -\frac{d}{dy}\log\left(p_i(y_i)\right) = -\frac{p_i'(y_i)}{p_i(y_i)}, \quad \text{for } i = 1, \cdots, n. \tag{50}$$

Now, we introduce the nonholonomic transforms both for the causal filter $\mathbf{L}(z)$ and the anticausal filter $\mathbf{R}(z^{-1})$ as follows:

$$d\mathbf{X}(z) = d\mathbf{L}(z) * \mathbf{L}^{\dagger}(z) \tag{51}$$
$$d\mathbf{Y}(z^{-1}) = d\mathbf{R}(z^{-1}) * \mathbf{R}^{\dagger}(z^{-1}). \tag{52}$$

In particular

$$d\mathbf{X}_0 = d\mathbf{L}_0\mathbf{L}_0^{-1} \tag{53}$$
$$d\mathbf{R}_0 = \mathbf{0}, \quad d\mathbf{Y}_0 = \mathbf{0}. \tag{54}$$

The differential transform has a definite geometric interpretation and is very important for us to develop an learning algorithm with the equivariant property [20], [21]. Actually, considering the relation (47) and letting the anticausal filter $\mathbf{R}(z^{-1})$ be fixed, we have

$$d\mathbf{y}(k) = [d\mathbf{L}(z)]\,\mathbf{u}(k) = \left[d\mathbf{L}(z)\mathbf{L}^{\dagger}(z)\mathbf{L}(z)\right]\mathbf{u}(k) \tag{55}$$
$$= [d\mathbf{X}(z)]\,\mathbf{L}(z)\mathbf{u}(k) = [d\mathbf{X}(z)]\,\mathbf{y}(k). \tag{56}$$

This means that if we adopt $\mathbf{X}(z)$ as a local coordinate system, the variation of the channel output $d\mathbf{y}(k)$ depends only on the variation $[d\mathbf{X}(z)]$, given $\mathbf{y}(k)$. In other words, the search direction $d\mathbf{X}(z)$ does not depend explicitly on the mixing filter $\mathbf{H}(z)$ but depends on the variation of the channel output. This property allows us to derive a learning algorithm with the equivariant property [20]. Furthermore, the nonholonomic transforms not only make the derivation of the algorithms simple but also lead to much more efficient and reliable learning algorithms. Using the nonholonomic transforms, we can easily calculate

$$d\mathbf{y} = d\left[\mathbf{W}(z)\right]\mathbf{x}(k)$$
$$= \left[d\mathbf{L}(z)\mathbf{R}(z^{-1})\right]\mathbf{x}(k) + \left[\mathbf{L}(z)d\mathbf{R}(z^{-1})\right]\mathbf{x}(k)$$
$$= \left[d\mathbf{L}(z) * \mathbf{L}^{\dagger}(z) * \mathbf{L}(z)\right]\mathbf{u}(k)$$
$$\quad + \left[\mathbf{L}(z)d\mathbf{R}(z^{-1}) * \mathbf{R}^{\dagger}(z^{-1}) * \mathbf{R}(z^{-1})\right]\mathbf{x}(k)$$
$$= [d\mathbf{X}(z)]\,\mathbf{y}(k) + \left[\mathbf{L}(z)d\mathbf{Y}(z^{-1})\right]\mathbf{u}(k). \tag{57}$$

Substituting (53) and (57) into (49), we obtain

$$dl\left(\mathbf{y}, \mathbf{W}(z)\right) = -tr(d\mathbf{X}_0) + \boldsymbol{\varphi}(\mathbf{y})^T\left[d\mathbf{X}(z)\right]\mathbf{y}(k)$$
$$\quad + \boldsymbol{\varphi}(\mathbf{y})^T\left[\mathbf{L}(z)d\mathbf{Y}(z^{-1})\right]\mathbf{u}(k)$$
$$= -tr(d\mathbf{X}_0) + \boldsymbol{\varphi}(\mathbf{y})^T\left[d\mathbf{X}(z)\right]\mathbf{y}(k)$$
$$\quad + \boldsymbol{\varphi}(\mathbf{y})^T\sum_{p,q=0}^{N}\mathbf{L}_p d\mathbf{Y}_q\mathbf{u}(k-p+q). \tag{58}$$

Now, we easily obtain the derivatives of the cost function with respect to $\mathbf{X}(z)$ and $\mathbf{Y}(z^{-1})$

$$\frac{\partial l\left(\mathbf{y}, \mathbf{W}(z)\right)}{\partial \mathbf{X}_p} = -\delta_{0,p}\mathbf{I} + \boldsymbol{\varphi}\left(\mathbf{y}(k)\right)\mathbf{y}^T(k-p) \tag{59}$$

$$\frac{\partial l\left(\mathbf{y}, \mathbf{W}(z)\right)}{\partial \mathbf{Y}_q} = \sum_{p=0}^{N}\mathbf{L}_p^T\boldsymbol{\varphi}\left(\mathbf{y}(k)\right)\mathbf{u}^T(k-p+q) \tag{60}$$

for $p = 0, 1, \cdots, N; q = 1, \cdots, N$. As discussed in the previous section, the nonholonomic variable $\mathbf{X}(z)$ can be considered to be a coordinate system in the tangent space $\mathcal{T}_{\mathbf{L}(z)}$. Since the tangent space is Euclidean, the natural gradient is defined by the derivative of the cost function. Thus, the gradient descent algorithms for $\mathbf{X}(z)$ and $\mathbf{Y}(z^{-1})$ are given by

$$\Delta\mathbf{X}_p = -\eta\frac{\partial l\left(\mathbf{y}, \mathbf{W}(z)\right)}{\partial \mathbf{X}_p}$$
$$= \eta\left(\delta_{0,p}\mathbf{I} - \boldsymbol{\varphi}(\mathbf{y})\mathbf{y}^T(k-p)\right) \tag{61}$$
$$\Delta\mathbf{Y}_q = -\eta\frac{\partial l\left(\mathbf{y}, \mathbf{W}(z)\right)}{\partial \mathbf{Y}_q}$$
$$= -\eta\sum_{p=0}^{N}\mathbf{L}_p^T\boldsymbol{\varphi}(\mathbf{y})\mathbf{u}^T(k-p+q) \tag{62}$$

for $p = 0, 1, \cdots, N; q = 1, \cdots, N$. Using the differential relations (51) and (52), we present learning algorithms that update the filters $\mathbf{L}(z)$ and $\mathbf{R}(z^{-1})$ as follows:

$$\Delta\mathbf{L}(z) = \Delta\mathbf{X}(z) * \mathbf{L}(z) \tag{63}$$
$$\Delta\mathbf{R}(z^{-1}) = \Delta\mathbf{Y}(z^{-1}) * \mathbf{R}(z^{-1}). \tag{64}$$

The learning algorithm (63) can be rewritten in matrix form as

$$\Delta\mathbf{L}_p = -\eta\sum_{q=0}^{p}\frac{\partial l\left(\mathbf{y}, \mathbf{W}(z)\right)}{\partial \mathbf{X}_q}\mathbf{L}_{p-q}$$
$$= \eta\sum_{q=0}^{p}\left(\delta_{0,q}\mathbf{I} - \boldsymbol{\varphi}\left(\mathbf{y}(k)\right)\mathbf{y}^T(k-q)\right)\mathbf{L}_{p-q} \tag{65}$$

for $p = 0, 1, \cdots, N$. In particular, for $p = 0, 1$, we have

$$\Delta\mathbf{L}_0 = \eta\left(\mathbf{I} - \boldsymbol{\varphi}\left(\mathbf{y}(k)\right)\mathbf{y}^T(k)\right)\mathbf{L}_0 \tag{66}$$
$$\Delta\mathbf{L}_1 = \eta\left[\left(\mathbf{I} - \boldsymbol{\varphi}\left(\mathbf{y(k)}\right)\mathbf{y}^T(k)\right)\mathbf{L}_1\right.$$
$$\left. -\boldsymbol{\varphi}\left(\mathbf{y}(k)\right)\mathbf{y}^T(k-1)\right)\mathbf{L}_0\right]. \tag{67}$$

Similarly, we can also give the explicit expression of the natural gradient algorithm for the anticausal FIR filter $\mathbf{R}(z^{-1})$

$$\Delta\mathbf{R}_p = -\eta\sum_{q=0}^{p}\frac{\partial l\left(\mathbf{y}, \mathbf{W}(z)\right)}{\partial \mathbf{Y}_q}\mathbf{R}_{p-q}$$
$$= -\eta\sum_{q=1}^{p}\sum_{r=0}^{N}\mathbf{L}_r^T\boldsymbol{\varphi}(\mathbf{y})\mathbf{u}^T(k-r+q)\mathbf{R}_{p-q} \tag{68}$$

for $p = 1, \cdots, N$. In particular, for $p = 1$, we have

$$\Delta\mathbf{R}_1 = -\eta(\Delta\mathbf{Y}_1\mathbf{R}_0 + \Delta\mathbf{Y}_0\mathbf{R}_1)$$
$$= -\eta\left(\sum_{p=0}^{N}\mathbf{L}_p^T\boldsymbol{\varphi}\left(\mathbf{y}(k)\right)\mathbf{u}^T(k-p+1)\right)\mathbf{R}_0. \tag{69}$$

Theoretically, the function activation functions $\varphi_i(y_i)$, $(i = 1, \cdots, n)$ depend on the pdfs of sources, which are unknown in a blind deconvolution setting. It is not necessary to precisely estimate the pdfs of sources to train the demixing filter. One important factor in determining the activation functions is that the stability conditions of the learning algorithm must be satisfied

[23], [31]. For QAM signals, the cubic function is a good activation function. For more detailed information about the activation function selection, see Amari *et al.* [31].

It should be noted that the natural gradient algorithm for the causal FIR filter is consistent with the one for the instantaneous mixture. If the causal FIR filter becomes a matrix, the natural gradient algorithm (63) will reduce to the algorithm for the instantaneous mixture [19].

For the noncausal demixing filter, it is important for an algorithm to adjust the demixing filter such that its coefficients decay on both sides, i.e., $\mathbf{W}_p$ tends to zero if $p \to +\infty$ or $p \to -\infty$. The filter decomposition approach is designed to meet such requirements.

One interesting question is how the interaction between the two filters is captured correctly by using the derived algorithms. Theoretically, both algorithms are the gradient descent method, which are derived from minimizing the same cost function, whose minimum is the solution. Geometrically, if the cascade form of the two filter is considered as a two-layer network (see Fig. 3), the learning algorithm for $\mathbf{R}(z^{-1})$ is considered to be an error back propagation through channel $\mathbf{L}(z)$. First, we give an geometric interpretation of the ordinary gradient

$$\frac{\partial l\left(\mathbf{y}, \mathbf{W}(z)\right)}{\partial \mathbf{R}_q} = \sum_{p=0}^{N} \left\langle \frac{\partial l\left(\mathbf{y}, \mathbf{W}(z)\right)}{\partial \mathbf{u}(k-p)}, \frac{\partial \mathbf{u}(k-p)}{\partial \mathbf{R}_q} \right\rangle$$
$$= \sum_{p=0}^{N} \mathbf{L}_p^T \boldsymbol{\varphi}\left(\mathbf{y}(k)\right) \mathbf{x}(k-p+q))^T. \quad (70)$$

The derivation shows that the derivative of $l(\mathbf{y}, \mathbf{W}(z))$ with respect to $\mathbf{R}_q$ can be calculated through nodes $\mathbf{u}(k-p)$, $p = 0, 1, \cdots, N$. This is similar to the error back propagation in training multilayer neural networks. The natural gradient algorithm [23] is the steepest direction on the manifold of FIR filters, which is the modified direction from the ordinary gradient. Thus, the interaction between the two filters can be captured correctly by using the derived algorithms.

VII. NUMERICAL IMPLEMENTATION

In this section, we consider the efficient implementation of learning algorithms for blind deconvolution of nonmimimum-phase systems. We introduce the following notations:

$$\mathcal{X}(k) = \left[\mathbf{x}^T(k), \mathbf{x}^T(k+1), \ldots, \mathbf{x}^T(k+N)\right]^T \quad (71)$$
$$\mathcal{U}(k) = \left[\mathbf{u}^T(k), \mathbf{u}^T(k-1), \ldots, \mathbf{u}^T(k-N)\right]^T \quad (72)$$

and

$$\mathcal{R} = [\mathbf{R}_0, \ldots, \mathbf{R}_N], \quad \mathcal{L} = [\mathbf{L}_0, \ldots, \mathbf{L}_N]. \quad (73)$$

Then, we can rewrite (46) and (47) into the following matrix form:

$$\mathbf{u}(k) = \mathcal{R}\mathcal{X}(k) \quad (74)$$
$$\mathbf{y}(k) = \mathcal{L}\mathcal{U}(k). \quad (75)$$
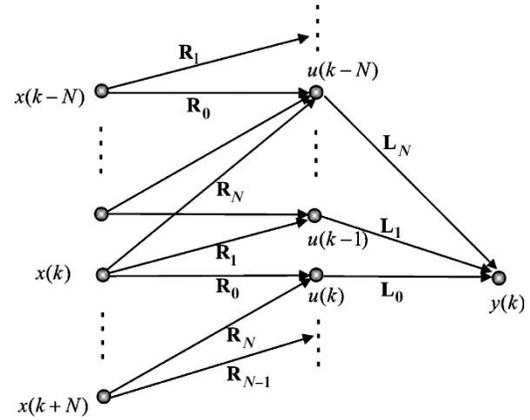
Fig. 3. Information back propagation through the two layer neural network.

A. Filter Length

The filter length of the demixing model is a critical parameter for blind deconvolution. Generally speaking, the longer filter length will give better performance in the absence of noise due to the decay properties of the demixing filter. However, a longer filter length will increase computing cost. To derive the best cost/benefit ratio, we need to introduce a criterion in order to choose an adequate filter length. Some statistical criteria for model selection, such as the minimum description length (MDL) [35] and the Akaike information-theoretic criterion (AIC) [36], can be used to determine the appropriate length of the demixing filter. In this paper, we employ a slightly modified minimum description length (MDL) as the criterion for filter length selection. We define the criterion as

$$\text{MDL}(N) = E\left[l\left(\mathbf{y}, \mathbf{W}(z)\right)\right] + \frac{n}{2}\log N. \quad (76)$$

The first term of the right side is slightly different from original MDL criterion in order to fit the model selection principle for blind deconvolution. The filter length N is chosen to minimize $\text{MDL}(N)$ over all natural integers.

B. Natural Gradient Learning for $\mathbf{L}(z)$

The natural gradient algorithm (63) can be rewritten equivalently in the following matrix form:

$$\Delta \mathcal{L}(k+1) = \eta(k)\mathcal{A}(k)\mathcal{L}(k) \quad (77)$$

where the matrix $\mathcal{A}(k)$ is defined as

$$\mathcal{A}(k) = \begin{bmatrix} \mathbf{U}_0 & \mathbf{0} & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{U}_1 & \mathbf{U}_0 & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{U}_2 & \mathbf{U}_1 & \mathbf{U}_0 & \cdots & \mathbf{0} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \mathbf{U}_N & \mathbf{U}_{N-1} & \mathbf{U}_{N-2} & \cdots & \mathbf{U}_0 \end{bmatrix} \quad (78)$$

and the matrices $\mathbf{U}_p$ are defined by

$$\mathbf{U}_p = \delta_{0,p}\mathbf{I} - \boldsymbol{\varphi}\left(\mathbf{y}(k)\right)\mathbf{y}^T(k-p), \quad \text{for } p = 0, \cdots, N. \quad (79)$$

From (77), it is observed that the natural gradient algorithm (63) is different from the one proposed in [22]. The difference comes from the fact that the matrix $\mathcal{A}(k)$ in this paper is a low-triangle

matrix, whereas in [22], it is a full one. This is because the effect of the noncausal part is included in the latter.

C. Natural Gradient Learning for Anticausal Filter $\mathbf{R}(z^{-1})$

The natural gradient algorithm for $\mathbf{R}(z^{-1})$ can be implemented in the same way

$$\Delta \mathcal{R}(k+1) = \eta(k)\mathcal{B}(k)\mathcal{R}(k) \tag{80}$$

where the matrix $\mathcal{B}(k)$ is defined as

$$\mathcal{B}(k) = \begin{bmatrix} \mathbf{0} & \mathbf{0} & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{V}_1 & \mathbf{0} & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{V}_2 & \mathbf{V}_1 & \mathbf{0} & \cdots & \mathbf{0} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \mathbf{V}_N & \mathbf{V}_{N-1} & \mathbf{V}_{N-2} & \cdots & \mathbf{0} \end{bmatrix} \tag{81}$$

and the matrices $\mathbf{V}_p$ are defined by

$$\mathbf{V}_p = -\sum_{r=0}^{N} \mathbf{L}_r^T \boldsymbol{\varphi}(\mathbf{y}) \mathbf{u}^T(k-r+p), \quad \text{for } p = 1, \cdots, N. \tag{82}$$

In the learning algorithm (80), we see that $\mathbf{R}_0$ is fixed to $\mathbf{I}$. Computer simulations show that the natural gradient algorithm has much better convergence properties and performance than the ordinary gradient algorithm.

VIII. SIMULATIONS

A number of computer simulations have been performed to investigate the validity and performance of the proposed natural gradient learning algorithms. We will illustrate simulation results in the following sections.

A. Performance Criterion

To evaluate the performance of the proposed learning algorithms, we employ the multichannel intersymbol interference [37], [38] denoted by M_{ISI} as a criterion

$$M_{\text{ISI}} = \sum_{i=1}^{n} \frac{\left| \sum_{j=1}^{n} \sum_{p=-N}^{N} |g_{p,ij}| - \max_{p,j} |g_{p,ij}| \right|}{\max_{p,j} |g_{p,ij}|}$$
$$+ \sum_{j=1}^{n} \frac{\left| \sum_{i=1}^{n} \sum_{p=-N}^{N} |g_{p,ij}| - \max_{p,i} |g_{p,ij}| \right|}{\max_{p,i} |g_{p,ij}|} \tag{83}$$

where $\mathbf{G}(z) = \left[\sum_{p=-N}^{N} g_{p,ij} z^{-p} \right]_{n \times n} = \mathbf{W}(z) * \mathbf{H}(z)$ is the global transfer function. It is easy to show that $M_{\text{ISI}} = 0$ if and only if $\mathbf{G}(z)$ is of the form (7). In order to remove the effect of a single numerical trial on evaluating the performance of algorithms, we use the ensemble average approach, that is, for each trial, we obtain a time sequence of M_{ISI}, and then, we take the average of M_{ISI} over differential trials to evaluate the performance of the algorithms.

B. Causal FIR Filters

In order to evaluate the performance of the learning algorithm (63), we randomly choose minimum-phase filters as the mixing
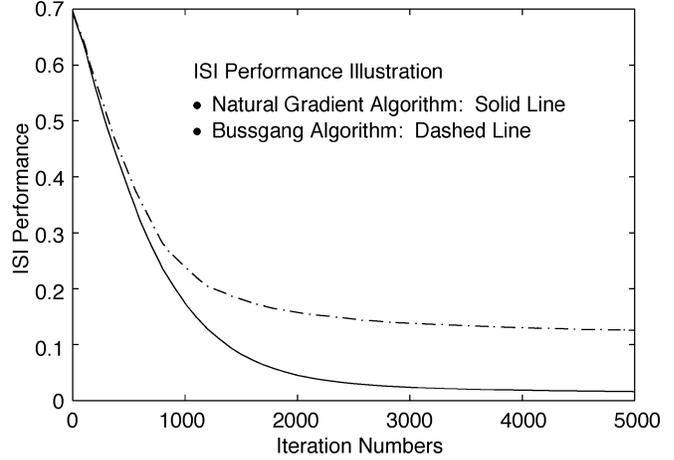
Fig. 4. M_{ISI} performance of the natural gradient algorithm.

model and use causal FIR filters to recover source signals. We present three examples to demonstrate the behavior and performance of the learning algorithm (63). In these examples, the mixing model is a multichannel ARMA model as follows:

$$\mathbf{u}(k) + \sum_{i=1}^{N} \mathbf{A}_i \mathbf{u}(k-i) = \sum_{i=0}^{N} \mathbf{B}_i \mathbf{s}(k-i) + \mathbf{v}(k) \tag{84}$$

where $\mathbf{u}$, $\mathbf{s}$, and $\mathbf{v} \in \mathbf{R}^3$. The matrices $\mathbf{A}_i$ and $\mathbf{B}_i$ are randomly chosen such that the mixing system is stable and minimum phase. The source signals $\mathbf{s}(k) = (s_1(k), s_2(k), s_3(k))^T$ are randomly generated i.i.d signals uniformly distributed in the range $(-1, 1)$, and $\mathbf{v}$ is the Gaussian noise with zero mean and a covariance matrix $0.1\mathbf{I}$. The nonlinear activation function is chosen to $\boldsymbol{\varphi}(y) = y^3$. The learning rate is another important factor to consider when implementing the natural gradient algorithm. The strategy used in our studies is to update the learning rate by $\eta(k+1) = \max\{0.9\eta(k), 10^{-4}\}$ for every 200 iterations. The initial value is set to $\eta(0) = 10^{-2}$.

Example 1: This example is considered to demonstrate the learning dynamics of the natural gradient algorithm. A large number of simulations show that the natural gradient algorithm can easily recover the source signals in the sense of $\mathbf{W}(z)\mathbf{H}(z) = \mathbf{P}\mathbf{\Lambda}\mathbf{D}(z)$ if the filter length is appropriately chosen. Fig. 4 illustrates a 100-trial ensemble average M_{ISI} performance of the natural gradient learning algorithm and the ordinary gradient algorithm. It can be seen that the natural gradient algorithm usually requires less than 2000 iterations to obtain satisfactory results, whereas the ordinary gradient algorithm requires more than 10 000 iterations to obtain satisfactory results since there is a long plateau in the ordinary gradient learning.

Example 2: This example is presented to show the noise tolerance of the natural gradient algorithm in the presence of noise levels ranging from -5 to 30 dB. There were 100 trials conducted with additive noise varying from -5 to 30 dB. Fig. 5 plots the average ISI index M_{ISI} for 100 trials versus the signal-to-noise ratio (SNR). We can see from this simulation that the average ISI index keeps below 0.1 if the SNR ≥ 7.

Example 3: Here, we consider an example to show how to choose the best filter length for the demixing model. We choose
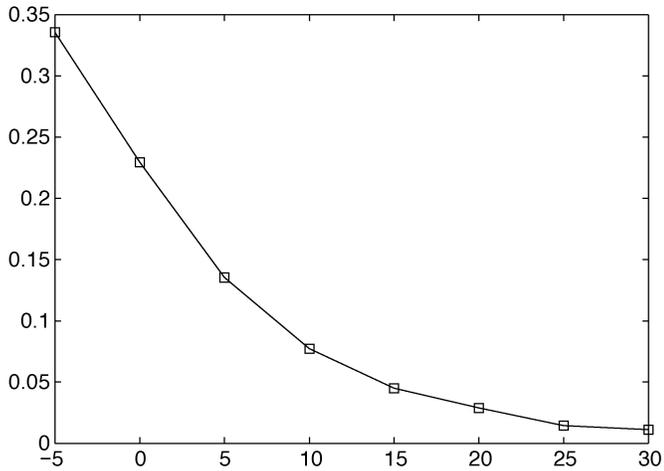
Fig. 5. M_{ISI} index for different noise levels from -5 to 30 dB. The horizontal axis indicates the noise level, and the vertical axis indicates the intersymbol inference index after convergence.
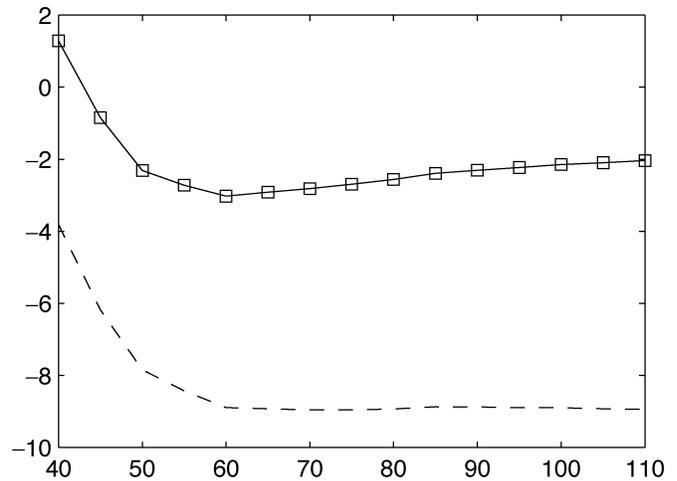
Fig. 7. MDL index and the cost function $E[l(\mathbf{y}, \mathbf{W}(z))]$ for different filter lengths of the demixing model. The continuous line shows the MDL index with respect to filter length N, and the dash line shows the cost function $E[l(\mathbf{y}, \mathbf{W}(z))]$ with respect to filter length N.
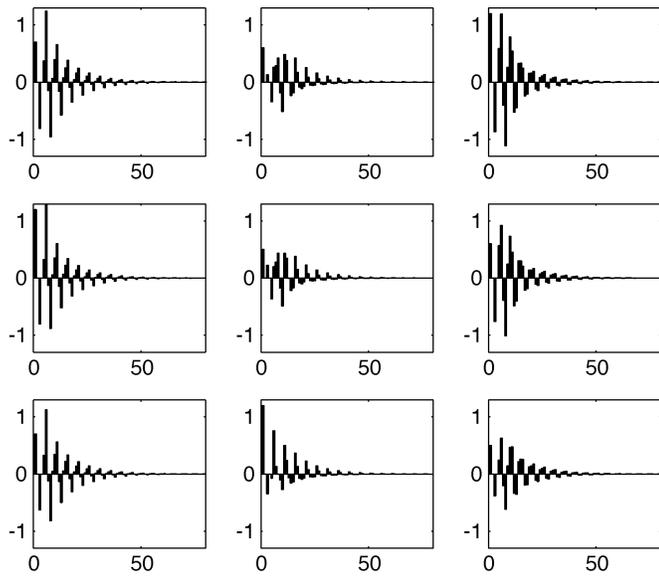
Fig. 6. Coefficients of the mixing filter $\mathbf{H}(z)$.
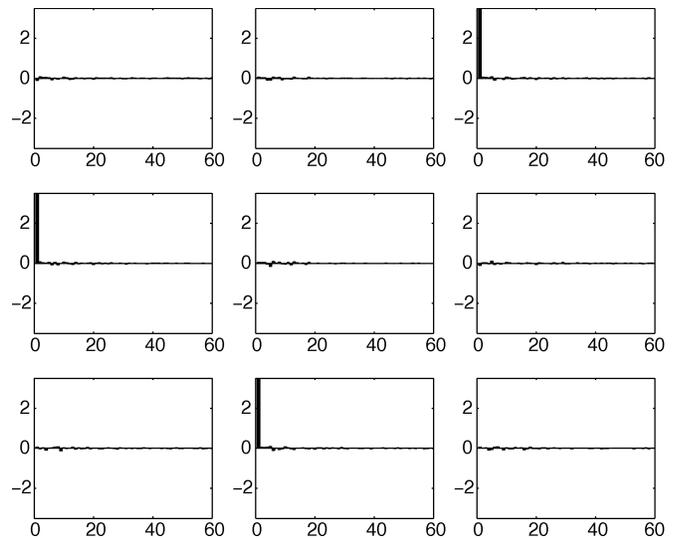
Fig. 8. Coefficients of $\mathbf{G}(z)$ of the causal system after 2000 iterations.

the ARMA model as a mixing system that is stable and minimum phase. Fig. 6 illustrates the coefficients of the transfer function $\mathbf{H}(z)$, where the (i, j)th subfigure plots the coefficients of the transfer function $H_{ij}(z) = \sum_{p=0}^{\infty} h_{p,ij} z^{-p}$ up to order 80.

We use the MDL criterion to choose the filter length. Fig. 7 shows the MDL index (upper curve) and the cost function $E[l(\mathbf{y}, \mathbf{W}(z))]$ (lower curve) for different filter lengths. We observed that the cost function does not decrease significantly if we further increase the filter length when $N \geq 60$. According to the MDL criterion, $N = 60$ is the best length for the demixing model.

Fig. 8 illustrates the coefficients of the global transfer function $\mathbf{G}(z) = \mathbf{W}(z)\mathbf{H}(z)$ after 2000 iterations, where the (i, j)th subfigure plots the coefficients of the transfer function $G_{ij}(z) = \sum_{p=0}^{\infty} g_{p,ij} z^{-p}$ up to order 80.

C. Nonmimimum-Phase Mixture Cases

In this simulation, the sensor signals are generated by the multichannel ARMA model (84), of which the matrices are randomly chosen such that the mixing system is stable and nonmimimum phase. The zero and pole distribution of the mixing system are plotted in Fig. 9. It is easy to verify that the system is stable and nonmimimum phase. In order to estimate the source signals, learning algorithms (77) and (80) have been used to train the demixing model. The filter length is selected using the MDL criterion. Figs. 10 and 11 illustrate the coefficients of the global transfer function $\mathbf{G}(z) = \mathbf{W}(z)\mathbf{H}(z)$ at the initial state and after 6000 iterations, respectively, where the (i, j)th subfigure plots the coefficients of the transfer function $G_{ij}(z) = \sum_{p=-N}^{N} g_{p,ij} z^{-p}$ ($\mathrm{N} = 60$). Figs. 12 and 13 show the coefficients of the causal filter $\mathbf{L}(z)$ and the anticausal filter $\mathbf{R}(z^{-1})$.
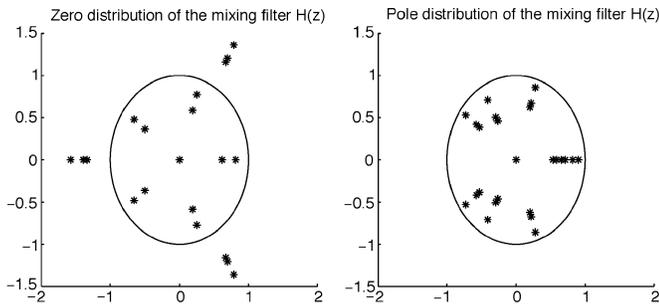
Fig. 9.   Zero and pole distributions of the mixing ARMA model.

Fig. 12.   Coefficients of the causal filter $\mathbf{L}(z)$.

Fig. 10.   Coefficients of $\mathbf{G}(z)$ of the noncausal system at the initial state.

Fig. 13.   Coefficients of the anticausal filter $\mathbf{R}(z^{-1})$.

decomposition approach. Computer simulations show that if the mixing system is nonmimimum phase, the Bussgang algorithm is very slow. There are significant differences in learning performance between the conventional transversal equalizers and the filter decomposition approach. The main reason for the difference is that the conventional equalizer does not utilize the invertibility property of the demixing filter. In the manifold of the invertible noncausal filter, the steepest ascent direction is not defined by the ordinary gradient. Therefore, the algorithm derived from the ordinary gradient does not give the best convergence rate.

The filter decomposition approach utilizes the invertibility property of the noncausal filter and introduces the natural gradient of causal filters as a search direction. Thus, the filter decomposition algorithm performs better than the conventional equalizer.

Fig. 11.   Coefficients of $\mathbf{G}(z)$ of the noncausal system after convergence.

It is easy to see that the coefficients of both filters decay as the delay number p increases.

In order to compare the learning performance of different algorithms, we also use the Bussgang algorithm to train an FIR filter with the same initialization and filter length as the filter

IX. CONCLUSION

A new filter decomposition approach is presented in this paper for multichannel blind deconvolution of nonmim-

imum-phase systems. Geometrical structures, such as the Lie group and the natural gradient, are introduced on the manifold of FIR filters. Natural gradient-based algorithms are developed to update the demixing filters. It should be mentioned that the filter decomposition approach is also applicable to other problems of estimating noncausal filters.

A large number of computer simulations have been performed to demonstrate the efficiency and performance of the proposed algorithms. It is observed from these simulations that the natural gradient algorithm usually requires only 2000 iterations to obtain satisfactory results, whereas the ordinary gradient algorithm requires more than 10 000 iterations to obtain satisfactory results since there is a long plateau in the ordinary gradient learning. Therefore, the learning plateau in blind deconvolution may be overcome by using the natural gradient approach.

REFERENCES

[1] Y. Sato, "Two extensional applications of the zero-forcing equalization method," *IEEE Trans. Commun.*, vol. COM-23, pp. 684–687, 1975.

[2] A. Benveniste, M. Goursat, and G. Ruget, "Robust identification of a nonmimimum-phase system: blind adjustment of a linear equalizer in data communications," *IEEE Trans. Automat. Contr.*, vol. AC-25, pp. 385–399, 1980.

[3] D. Godard, "Self-recovering equalization and carrier tracking in two-dimensional data communication systems," *IEEE Trans. Commun.*, vol. COM-28, pp. 1867–1875, 1980.

[4] 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, pp. 349–372, 1983.

[5] O. Shalvi and E. Weinstein, "New criteria for blind deconvolution of nonmimimum-phase systems (channels)," *IEEE Trans. Inform. Theory*, vol. 36, pp. 312–321, Mar. 1990.

[6] J. A. Cadzow, "Blind deconvolution vis cumulant extrema," *IEEE Signal Processing Mag.*, vol. 13, pp. 24–42, 1996.

[7] G. Giannakis and J. Mendel, "Identification of nonmimimum-phase systems using higher-order statistics," *IEEE Trans. Acoust. Speech, Signal Processing*, vol. 37, pp. 360–377, Mar. 1989.

[8] J. Mendel, "Turoiral on higher-order statistics (spectra) in signal processing and system theory: theoretical results and some applications," *Proc. IEEE*, vol. 79, pp. 278–305, Mar. 1991.

[9] S. Amari, S. Douglas, A. Cichocki, and H. Yang, "Novel on-line algorithms for blind deconvolution using natural gradient approach," in *Proc. 11th IFAC Symp. Syst. Ident.*, Kitakyushu, Japan, July 8–11, 1997, pp. 1057–1062.

[10] A. Bell and T. Sejnowski, "An information maximization approach to blind separation and blind deconvolution," *Neural Comput.*, vol. 7, pp. 1129–1159, 1995.

[11] Y. Hua, "Fast maximum likelihood for blind identification of multiple FIR channels," *IEEE Trans. Signal Processing*, vol. 44, pp. 661–672, Mar. 1996.

[12] L. Tong, G. Xu, and T. Kailath, "Blind identification and equalization base on second-order statistics: a time domain approach," *IEEE Trans. Inform. Theory*, vol. 40, pp. 340–349, Mar. 1994.

[13] 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, pp. 516–525, Feb. 1995.

[14] A. Gorokhov and P. Loubaton, "Blind identification of MIMO-FIR system: a generalized linear prediction approach," *Signal Process.*, vol. 73, pp. 105–124, 1999.

[15] J. Tugnait and B. Huang, "Multistep linear predictors-based blind identification and equalization of multiple-input multiple-output channels," *IEEE Trans. Signal Processing*, vol. 48, pp. 26–38, Jan. 2000.

[16] Y. Hua and J. Tugnait, "Blind identifiability of FIR-MIMO systems with colored input using second order statistics," *IEEE Signal Processing Lett.*, vol. 7, pp. 348–350, Dec. 2000.

[17] Z. Ding and Y. Li, *Blind Equalization and Identification*. New York: Marcel Dekker, 2000.

[18] S. Haykin, *Unsupervised Adaptive Filtering, Vol. II: Blind Deconvolution*. New York: Wiley, 2000.

[19] S. Amari, A. Cichocki, and H. Yang, "A new learning algorithm for blind signal separation," in *Advances in Neural Information Processing Systems 8 (NIPS* 95)*, G. Tesauro, D. Touretzky, and T. Leen, Eds. Cambridge, MA: MIT Press, 1996, pp. 757–763.

[20] J.-F. Cardoso and B. Laheld, "Equivariant adaptive source separation," *IEEE Trans. Signal Processing*, vol. 43, pp. 3017–3029, Dec 1996.

[21] S. Amari, "Natural gradient works efficiently in learning," *Neural Comput.*, vol. 10, pp. 251–276, 1998.

[22] S. Amari, S. Douglas, and A. Cichocki, "Multichannel blind deconvolution and source separation using the natural gradient," IEEE Trans. Signal Processing, to be published.

[23] L. Zhang, S. Amari, and A. Cichocki, "Semiparametric model and superefficiency in blind deconvolution," *Signal Process.*, pp. 2535–2553, 2001.

[24] J. Labat, O. Macchi, and C. Laot, "Adaptive decision feedback equalization: can you skip the training period?," *IEEE Trans. Commun.*, vol. 46, pp. 921–930, July 1998.

[25] A. Nandi and S. Anfinsen, "Blind equalization with recursive filter structures," *Signal Process.*, vol. 80, pp. 2151–2167, 2000.

[26] W. M. Boothby, *An Introduction to Differential Manifolds and Riemannian Geometry*. New York: Academic, 1986.

[27] S. Smith, "Geometric optimization methods for adaptive filtering," Ph.D. dissertation, Harvard Univ., Cambridge, MA, 1993.

[28] A. Edelman, T. Arias, and S. Smith, "The geometry of algorithms with orthogonal consrtaints," *SIAM J. Matrix Anal. Applicat.*, 1998.

[29] S. Amari, "Natural gradient for over- and under-complete bases in ICA," *Neural Comput.*, vol. 11, pp. 1883–1975, 1999.

[30] L. Zhang, A. Cichocki, and S. Amari, "Natural gradient algorithm for blind separaiton of overdetermined mixture with additive noise," *IEEE Signal Processing Lett.*, vol. 6, pp. 293–295, Nov. 1999.

[31] S. Amari, T. Chen, and A. Cichocki, "Stability analysis of adaptive blind source separation," *Neural Networks*, vol. 10, pp. 1345–1351, 1997.

[32] S. Amari and A. Cichocki, "Adaptive blind signal processing—neural network approaches," *Proc. IEEE*, vol. 86, pp. 2026–2048, Oct. 1998.

[33] D. Pham, "Mutual information approach to blind separation of stationary sources," in *Proc. First Intl. Workshop Independent Component Anal. Signal Separation*, Aussois, France, Jan. 1999, pp. 215–220.

[34] A. Taleb, J. Sole, and C. Jutten, "Quasinonparametric blind inversion of wiener systems," *IEEE Trans. Signal Processing*, vol. 49, pp. 917–924, May 2001.

[35] J. Rissanen, "Modeling by shortest data description," *Automatica*, vol. 14, pp. 465–471, 1978.

[36] H. Akaike, "A new look at the statistical model identification," *IEEE Trans. Automat. Contr.*, vol. 19, pp. 716–723, 1974.

[37] J. Tugnait, "Channel estimation and equlization using high-order statistics," in *Signal Processing Advances in Wireless and Mobile Communications*. Upper Saddle River: Prentice-Hall, 2000, vol. 1, pp. 1–40.

[38] Y. Inouye and S. Ohno, "Adaptive algorithms for implementing the single-stage criterion for multichannel blind deconvolution," in *Proc. Fifth Int. Conf. Neural Inform. Process.*, Kitakyushu, Japan, Oct. 21–23, 1998, pp. 733–736.

Liqing Zhang received the B.S. degree in mathematics from Hangzhou University, Hangzhou, China, in 1983 and the Ph.D. degree in computer sciences from Zhongshan University, Guangzhou, China, in 1988.

He became an Associate Professor in 1990 and then a Full Professor in 1995 at the Department of Automation, South China University of Technology, Ghangzhou. He joined Laboratory for Advanced Brain Signal Processing, RIKEN Brain Science Institute, Saitama, Japan, in 1997 as a Research Scientist. Since 2002, he has been with the Department of Computer Sciences, Shanghai Jiaotong University, Shanghai, China. His research interests include neuroinformatics, visual computing, adaptive systems, and statistical learning. He has published more than 80 papers.

Andrzej Cichocki (M'96) received the M.Sc. (with honors), Ph.D., and Habilitate Doctorate (Dr.Sc.) degrees, all in electrical engineering, from Warsaw University of Technology, Warsaw, Poland, in 1972, 1975, and 1982, respectively.

Since 1972, he has been with the Institute of Theory of Electrical Engineering, Measurements, and Information Systems, Warsaw University of Technology, where he became a full Professor in 1991. He is the co-author of three books: *Adaptive Blind Signal and Image Processing-Learning Algorithms and Applications* (New York: Wiley, 2002), *MOS Switched-Capacitor and Continuous-Time Integrated Circuits and Systems* (New York: Springer-Verlag, 1989), and *Neural Networks for Optimization and Signal Processing* (New York: Teubner-Wiley, 1993, 1994). He has also written more than 150 research papers. He spent a few years as Alexander Humboldt Research Fellow and Guest Professor at University Erlangen-Nuernberg, Nuernberg, Germany. Since 1995, he has been with the Brain Science Institute RIKEN, Saitama, Japan, as a team leader of the Laboratory for Open Information Systems and currently as a head of the Laboratory for Advanced Brain Signal Processing. His current research interests include optimization, bioinformatics, neurocomputing, and signal and image processing, especially analysis and processing of multi-sensory biomedical data.

Dr. Cichocki is currently an associate editor of IEEE TRANSACTION ON NEURAL NETWORKS and recently was the member of a core group who established a new IEEE Circuits and Systems Technical Committee for Blind Signal Processing. He is also a member of the Steering Committee of ICA workshops.

Shun-ichi Amari (F'94) graduated from the University of Tokyo, Tokyo, Japan, in 1958, majoring in mathematical engineering, and received the Dr. Eng. degree from the University of Tokyo in 1963.

He was an Associate Professor at Kyushu University, Kyushu, Japan, an Associate and then Full Professor with the Department of Mathematical Engineering and Information Physics, University of Tokyo. He is now Professor-Emeritus at the University of Tokyo. He is the Director of the 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. He was founding Coeditor-in-Chief of *Neural Networks*.

Dr. Amari served as President of the International Neural Network Society, as Council Member of Bernoulli Society for Mathematical Statistics and Probability Theory, and is President-Elect of the Institute of Electrical, Information, and Communication Engineers. He has been awarded the Japan Academy Award, the IEEE Neural Networks Pioneer Award, the IEEE Emanuel R. Piore Award, the Neurocomputing best paper award, and the IEEE Signal Processing Society best paper award, among many others.