ELSEVIER

Signal Processing 81 (2001) 2535–2553

SIGNAL PROCESSING

www.elsevier.com/locate/sigpro

Semiparametric model and superefficiency in blind deconvolution

L.-Q. Zhang *, S. Amari, A. Cichocki

Brain-style Information Systems Research Group, RIKEN Brain Science Institute, Hirosawa 2-1, Wako shi, Saitama 351-0198, Japan

Received 7 June 1999; received in revised form 24 July 2000; accepted 27 July 2001

Abstract

In this paper, we study convergence and efficiency of the batch estimator and natural gradient algorithm for blind deconvolution. First, the blind deconvolution problem is formulated in the framework of a semiparametric model, and a family of estimating functions is derived for blind deconvolution. To improve the learning efficiency of the online algorithm, explicit standardized estimating functions are given and within this framework the superefficiency of batch learning and online natural gradient learning is proven. © 2001 Elsevier Science B.V. All rights reserved.

Keywords: Blind deconvolution; Independent component analysis; Semiparametric model; Estimating function; Superefficiency; Natural gradient

1. Introduction

Blind separation/deconvolution is of increasing importance in areas such as telecommunications, speech, image enhancement and biomedical signal processing [8,13,20–22,24,25,30–32,38,40,47] Refer to papers [7,19,26,41] for more details. Various algorithms, such as Bussgang algorithms [14,34,37,43], higher-order statistics approach [39,17], information-theoretic approaches [9,13,28] and the subspace method [1,25,35] have been developed for solving the blind deconvolution problem. Identifiability of blind deconvolution has also been discussed for single input multiple output (SIMO) systems [35,42] and multiple input multiple output (MIMO) systems [25,44,29]. In general, the second order statistical methods rely on the separability of noise and signal subspaces, which requires some prior knowledge on the length of the unknown channels to be identified. The performance of the algorithms is still not satisfactory in presence of noise when the length of the unknown channels is not well estimated. On the other hand, the high order statistical methods can be effective in presence of noise under appropriate initialization, but may suffer from low convergence and local convergence. The efficiency of statistical learning algorithms has not been covered in the previous works on blind deconvolution. It is the purpose of this paper to develop fast and efficient algorithms based on the high order statistics and to analyze the convergence and efficiency of the learning algorithms.

[*] Corresponding author. Tel.: +81-48-467-9665; fax: +81-48-467-9694.

E-mail address: zha@bsp.brain.riken.go.jp (L.-Q. Zhang).

0165-1684/01/$ - see front matter © 2001 Elsevier Science B.V. All rights reserved.

PII: S0165-1684(01)00146-3

Nomenclature

$\mathbf{s}(k)$	source signal vector
$\mathbf{x}(k)$	sensor signal vector
$\mathbf{y}(k)$	recovered signal vector
$\mathbf{H}(z)$	convolutive mixing filter
$\mathbf{W}(z)$	demixing filter
$\mathscr{M}(N)$	FIR filters manifold
$*$	Lie multiplication in $\mathscr{M}(N)$
$\dagger$	Lie inverse in $\mathscr{M}(N)$
$\tilde{\nabla} l(\mathbf{W}(z))$	natural gradient of a cost function
$\mathbf{X}(z)$	nonholonomic parameterization variable
θ	parameter of interest in semiparametric model
ξ	nuisance parameter in semiparametric model
$r(\boldsymbol{s})$	probability density function of source signals $\mathbf{s}$
$\mathscr{T}^{\mathscr{N}}_{\mathbf{W}(z),r}$	nuisance tangent space at $(\mathbf{W}(z), r)$
$p(\mathbf{x}; \mathbf{W}, r)$	probability density function of sensor signal $\mathbf{x}$
$\varphi_i(s)$	activation function
$\mathbf{F}(\mathbf{x}, \mathbf{W}(z))$	estimating function for blind deconvolution
$\mathscr{K}(z)$	derivative operator
$\mathbf{F}^*(\mathbf{x}, \mathbf{W}(z))$	standardized estimating function

A semiparametric statistical model concerns a family of probability distributions specified by a finite dimensional parameter of interest and an infinite-dimensional nuisance parameter [15]. Amari and Kumon [12] suggest approaching semiparametric statistical models via estimating functions and understanding their geometries and efficiencies in terms of information geometry [2,36]. Amari and Cardoso [5] have also applied information geometry to blind source separation and derived an admissible class of estimating functions including efficient estimators. They show that the manifold of mixtures is m-curvature free, so that algorithms of blind separation can be designed without attention to source probability functions. The theory of semiparametric model is also applied to derive efficiency and superefficiency of demixing learning algorithms [4]. See also [6,20] for stability of demixing algorithms.

Most theories treat only blind source separation of instantaneous mixtures and it is only recently that the natural gradient approach has been proposed for multichannel blind deconvolution [9,45]. Amari et al. [9] discuss the geometric structures of the IIR filter manifold, to develop an efficient learning algorithm for blind deconvolution. However, in most practical implementations, it is necessary to employ a filter of finite length as a demixing model. Zhang et al. [45] directly investigate the geometric structures of the FIR filter manifold and derive the natural gradient algorithm for training FIR filters. Stability analysis for natural gradient learning is also provided.

The present paper will examine further convergence and efficiency of the batch estimator and natural gradient learning for blind deconvolution via the semiparametric statistical model and estimating functions [15]. First, we introduce the geometrical properties of the manifold of the FIR filters based on the Lie group structure and formulate the blind deconvolution problem within the framework of the semiparametric model deriving a family of estimating functions for blind deconvolution. We then analyze the efficiency of the batch estimator based on estimating function — obtaining its convergence rate. Finally, we prove that both batch learning and natural gradient learning are superefficient under given nonsingular conditions.

Further information on information geometry is given in Ref. [2,36] and that for semiparametric statistical model in [5,15,11].

2. Problem formulation

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

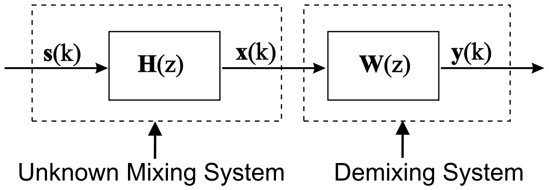

Fig. 1. Illustration of blind deconvolution problem.

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, $\mathbf{s}(k)=(s_1(k),\ldots,s_n(k))^{\mathrm{T}}$ is an n-dimensional vector of source signals, zero-mean and *independent and identically distributed* (i.i.d.), and $\mathbf{x}(k)=(x_1(k),\ldots,x_n(k))^{\mathrm{T}}$ 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. $\mathbf{H}(z)$ is usually called the mixing filter, unknown in blind deconvolution.

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

$$\mathbf{y}(k)=\sum_{p=-\infty}^{\infty}\mathbf{W}_p\mathbf{x}(k-p), \tag{3}$$

where $\mathbf{y}(k)=(y_1(k),\ldots,y_n(k))^{\mathrm{T}}$ 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 (double-side) z-transform of $\mathbf{W}_p$ is denoted by $\mathbf{W}(z)$, which is called the demixing filter. See Fig. 1 for illustration of the blind deconvolution problem.

The objective of blind deconvolution is to make the output signals $\mathbf{y}(k)$ of the demixing model maximally spatially mutually independent and temporarily i.i.d.. In this paper, we employ the semiparametric model to derive a family of estimating functions and develop efficient learning algorithms for training the demixing filter $\mathbf{W}(z)$. Finally, we analyze the convergence and efficiency of the learning algorithms.

In practice, we have to implement the blind deconvolution problem with a *finite impulse response* (FIR) filter

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

where N is the length of the demixing filter. In general, the multiplication of two filters of form (4) will enlarge the filter length. Below, we will investigate some geometrical structures of the FIR manifold.

2.1. *Recoverability*

It is possible to ask if there exists an FIR filter $\mathbf{W}(z)$ such that the output of the demixing model recovers the source signals and in what sense the source signals are recovered. For simplicity, assume that the mixing filter is an FIR filter

$$\mathbf{H}(z)=\sum_{p=0}^{L}\mathbf{H}_p z^{-p},\ \det(\mathbf{H}_0)\neq 0. \tag{5}$$

Assume that $\mathbf{H}(z)$ has no zeros on the unit circle. If we consider the FIR filter $\mathbf{H}(z)$ as a matrix of polynomials of z, the determinant of $\mathbf{H}(z)$ is given by

$$\det(\mathbf{H}(z)) \\ =\det(\mathbf{H}_0)\prod_{p=1}^{L_1}(1-a_p z^{-1})\prod_{p=1}^{L_2}(1-b_p z^{-1}), \tag{6}$$

where L_1 and L_2 are certain natural numbers, $0<\|a_p\|<1$, for $p=1,\ldots,L_1$ and $\|b_p\|>1$ for $p=1,\ldots,L_2$. Usually, a_p, b_p are referred to the zeros of the FIR filter $\mathbf{H}(z)$. If all the zeros are located in the interior of the unit circle, the filter $\mathbf{H}(z)$ is minimum-phase. Otherwise, the filter $\mathbf{H}(z)$ is nonminimum-phase.

Now the inverse $\mathbf{H}^{-1}(z)=\sum_{p=-\infty}^{\infty}\bar{\mathbf{H}}_p z^{-p}$, can be calculated by

$$\bar{\mathbf{H}}_p=\frac{1}{2\pi\mathrm{i}}\oint\mathbf{H}^{\#}(z)\det(\mathbf{H}(z))^{-1}z^{p-1}\,\mathrm{d}z \quad \text{for } p=-\infty,\ldots,+\infty, \tag{7}$$

where $\mathbf{H}^{\#}(z)$ is the adjoint matrix of $\mathbf{H}(z)$. It is not difficult to verify that the coefficient matrices $\bar{\mathbf{H}}_p$ satisfy the following decay condition:

$$\|\bar{\mathbf{H}}_p\|\leqslant \text{const.}\left[\max_{j,k}(\|a_j\|,\|b_k\|^{-1})\right]^{|p|}, \tag{8}$$

where $\|\bar{\mathbf{H}}_p\|=\{tr(\bar{\mathbf{H}}_p^{\mathrm{T}}\bar{\mathbf{H}}_p)\}^{1/2}$ is a matrix norm. Generally, the inverse filter of $\mathbf{H}(z)$ is a noncausal filter of infinity length. In practice, we usually employ a noncausal filter of finite length as a demixing model. The approximation will introduce a model error in blind deconvolution. If we make the length of the demixing filter sufficiently large, the model error will became negligible due to decay (8). By using the time delay transform or filter decomposition approach [46], we thus need to view the causal FIR filter only as a demixing model. In the next section, we define precisely the sense of the recovered signals in the Lie group framework.

2.2. Indeterminacy

In the following discussion, we presume that both the mixing filter $\mathbf{H}(z)$ and demixing filter $\mathbf{W}(z)$ are causal FIR filters of length N. The global transfer function is defined by

$$\mathbf{G}(z)=[\mathbf{W}(z)\mathbf{H}(z)]_N, \tag{9}$$

where $[\,\cdot\,]_N$ is a truncating operator such that any terms with orders higher than N in the polynomial are omitted. Generally speaking, blind deconvolution does not seek an exact inverse filter of the mixing filter. In blind deconvolution, we cannot observe the vector $\mathbf{s}(k)$ of original signals and the unknown mixing filter $\mathbf{H}(z)$. This implies three inherent ambiguities in the solution to the 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 of each component $s_i(k)$ and the magnitude of each component $s_i(k)$. Therefore, the blind deconvolution task is to find a demixing filter $\mathbf{W}(z)$ such that

$$\mathbf{G}(z)=[\mathbf{W}(z)\mathbf{H}(z)]_N=\mathbf{P}\boldsymbol{\Lambda}\mathbf{D}(z), \tag{10}$$

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

3. Geometrical structures on FIR manifold

Geometrical structures, such as the Riemannian metric on the parameter space, can help us develop efficient learning algorithms for training parameters. The commonly used gradient descent learning is not optimal in minimizing a cost function defined on Riemannian space. The steepest search direction is given by the natural gradient. It has been demonstrated that the natural gradient search scheme is an efficient approach for solving iterative parameter estimation problems [3]. In order to develop an efficient learning algorithm for blind deconvolution, we first explore some geometrical properties of the manifold of FIR filters.

3.1. The FIR manifold

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

$$\mathscr{M}(N)=\left\{\mathbf{W}(z)\,|\,\mathbf{W}(z)=\sum_{p=0}^{N}\mathbf{W}_p z^{-p},\det(\mathbf{W}_0)\neq 0\right\}. \tag{11}$$

$\mathscr{M}(N)$ is a manifold of dimension $n^2(N+1)$. The tangent space of $\mathscr{M}(N)$ at $\mathbf{W}(z)$, denoted by $\mathscr{T}\mathscr{M}_{\mathbf{W}(z)}$, is given by $\mathscr{T}\mathscr{M}_{\mathbf{W}(z)}=\{\mathbf{X}(z)\,|\,\mathbf{X}(z)=\sum_{p=0}^{N}\mathbf{X}_p z^{-p}\}$, where $\mathbf{X}_p$, $p=0,1,\ldots,N$ are $n\times n$ matrices. In general, multiplication of two filters in $\mathscr{M}(N)$ will enlarge the filter length. This makes it difficult to introduce the Riemannian structure to the manifold of multichannel FIR filters. In order

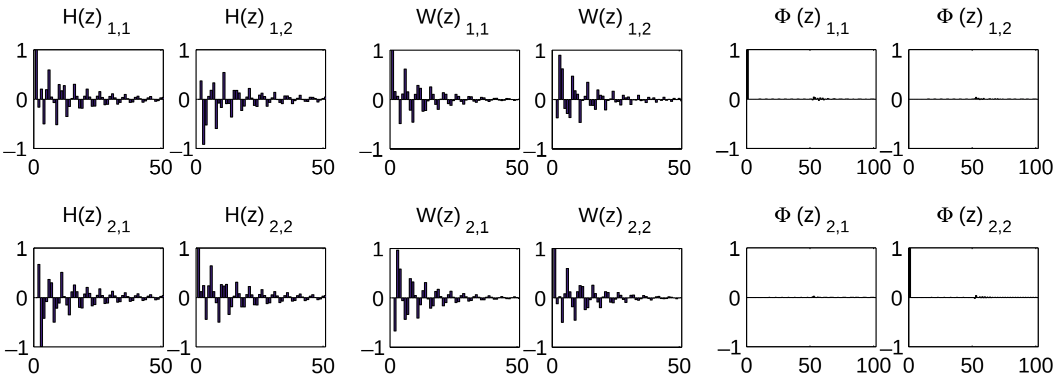

Fig. 2. Illustration of the Lie group inverse of an FIR filter, where $\mathbf{H}(z)$ is an FIR filter of length 50, $\mathbf{W}(z)$ is the Lie group inverse of $\mathbf{H}(z)$, and $\mathbf{\Phi}(z)=\mathbf{W}(z)\mathbf{H}(z)$ is the composite transfer function.

to explore possible geometrical structures of $\mathscr{M}(N)$ which will lead to effective learning algorithms for $\mathbf{W}(z)$, we define the algebraic operations of filters in the Lie group framework.

3.2. Lie group

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

$$\mathbf{B}(z)*\mathbf{C}(z)=\sum_{p=0}^{N}\sum_{q=0}^{p}\mathbf{B}_q\mathbf{C}_{(p-q)}z^{-p}, \tag{12}$$

$$\mathbf{B}^{\dagger}(z)=\sum_{p=0}^{N}\mathbf{B}_p^{\dagger}z^{-p}, \tag{13}$$

where $\mathbf{B}_p^{\dagger}$ are recurrently defined by $\mathbf{B}_0^{\dagger}=\mathbf{B}_0^{-1}$, $\mathbf{B}_p^{\dagger}=-\sum_{q=1}^{p}\mathbf{B}_{p-q}^{\dagger}\mathbf{B}_q\mathbf{B}_0^{-1}$, $p=1,\dots,N$. With these operations, both $\mathbf{B}(z)*\mathbf{C}(z)$ and $\mathbf{B}^{\dagger}(z)$ still remain in the manifold $\mathscr{M}(N)$. It is easy to verify that the manifold $\mathscr{M}(N)$ with the above operations forms a Lie Group [16,23]. The identity element is $\mathbf{E}(z)=\mathbf{I}$, where $\mathbf{I}$ is the identity matrix. In fact the Lie multiplication of two $\mathbf{B}(z),\mathbf{C}(z)\in\mathscr{M}(N)$ 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.

The geometrical interpretation of the Lie group inverse is illustrated in Fig. 2, where $\mathbf{H}(z)$ is a two channel filter of length $N=50$, $\mathbf{W}(z)=\mathbf{H}^{\dagger}(z)$ is the Lie group inverse filter of length 50 and the composite transfer function $\mathbf{\Phi}(z)=\sum_{p=0}^{2N}\mathbf{\Phi}_p z^{-p}=\mathbf{W}(z)\mathbf{H}(z)$ is a filter of length $2N$. In this figure, subfigure $\mathbf{H}(z)_{11}$ plots the subchannel transfer function $H_{11}(z)=\sum_{p=0}^{N}h_{p,11}z^{-p}$, where the horizontal axis indicates the time delays $p=0,\dots,N$, and vertical axis indicates the magnitude $h_{p,11}$. From this illustration, we see that the composite transfer function $\mathbf{\Phi}(z)$ is not the exact identity matrix, there still exist small fluctuations in coefficients $\mathbf{\Phi}_p$, for $p>N$. The fluctuations will be negligible if we make the length N of $\mathbf{W}(z)$ sufficiently large. However, considering the multiplication in the Lie group sense, we have $\mathbf{G}(z)=\mathbf{W}(z)*\mathbf{H}(z)=\mathbf{I}$. In the following discussion, we consider the global transfer function in the Lie group sense $\mathbf{G}(z)=\mathbf{W}(z)*\mathbf{H}(z)$.

3.3. Natural gradient

The Lie Group has an important property that admits an invariant Riemannian metric [23]. Using the Lie group structure, we derive the natural gradient of a cost function $l(\mathbf{W}(z))$ defined on the manifold $\mathscr{M}(N)$

$$\tilde{\nabla} l(\mathbf{W}(z)) = \frac{\partial l(\mathbf{W}(z))}{\partial \mathbf{X}(z)} * \mathbf{W}(z), \tag{15}$$

where $\mathbf{X}(z)$ is a nonholonomic variable [6], defined by the following equation:

$$\mathrm{d}\mathbf{X}(z) = \mathrm{d}\mathbf{W}(z) * \mathbf{W}^{\dagger}(z) = [\mathrm{d}\mathbf{W}(z)\mathbf{W}^{-1}(z)]_N. \tag{16}$$

There are two ways to calculate the $\partial l(\mathbf{W}(z))/\partial \mathbf{X}(z)$. One is to evaluate it by the following relation:

$$\frac{\partial l(\mathbf{W}(z))}{\partial \mathbf{X}(z)} = \frac{\partial l(\mathbf{W}(z))}{\partial \mathbf{W}(z)} * \mathbf{W}^{\mathrm{T}}(z^{-1}). \tag{17}$$

See Section 5 for the detailed derivation. The other way is to directly calculate it by using the following property:

$$\mathrm{d}\mathbf{y}(k) = \mathrm{d}\mathbf{W}(z)\mathbf{x}(k) = \mathrm{d}\mathbf{X}(z)\mathbf{y}(k). \tag{18}$$

From the above equation, we see that the differential $\mathrm{d}\mathbf{X}(z)$ defines a channel variation with respect to variation of output of the demixing model. This property is critical for the derivation of learning algorithms with equivariance. See [4,20] for instantaneous mixtures.

4. Semiparametric models for blind deconvolution

In order to study convergence and efficiency of the batch estimator and natural gradient learning for blind deconvolution, we first introduce a basic theory of semiparametric models, and formulate the blind deconvolution problem within its framework.

4.1. Semiparametric model

Consider a general statistical model $\{p(\mathbf{x};\boldsymbol{\theta},\boldsymbol{\xi})\}$, where $\mathbf{x}$ is a random variable whose probability density function is specified by two parameters, $\boldsymbol{\theta}$ and $\boldsymbol{\xi}$, $\boldsymbol{\theta}$ being the parameter of interest, and $\boldsymbol{\xi}$ being the nuisance parameter. When the nuisance parameter is of infinite dimension or of functional degrees of freedom, the statistical model is called a semiparametric model [15].

The gradient vectors of the log likelihood

$$\boldsymbol{u}(\boldsymbol{x},\boldsymbol{\theta},\boldsymbol{\xi}) = \frac{\partial}{\partial \boldsymbol{\theta}} \log p(\mathbf{x};\boldsymbol{\theta},\boldsymbol{\xi}), \tag{19}$$

are called the score functions of the parameter of interest or shortly $\boldsymbol{\theta}$-score. In order to discuss the geometrical properties of the statistical manifold, we introduce the following function space:

$$\mathscr{H}_{\theta,\xi} = \{w(\boldsymbol{x}) | E_{\theta,\xi}[w(\boldsymbol{x})] = 0,\ E_{\theta,\xi}[w(\boldsymbol{x})^2] < \infty\}, \tag{20}$$

where $E_{\theta,\xi}$ denotes the expectation with respect to $p(\mathbf{x};\boldsymbol{\theta},\boldsymbol{\xi})$. The set $\mathscr{H}_{\theta,\xi}$ is a linear space admitting a Hilbert space structure with the inner product

$$\langle w_1(x), w_2(x)\rangle = E_{\theta,\xi}[w_1(x)w_2(x)]. \tag{21}$$

The components $u_i(\mathbf{x},\boldsymbol{\theta},\boldsymbol{\xi})$ of the $\boldsymbol{\theta}$-score are in $\mathscr{H}_{\theta,\xi}$, provided the Fisher information exists.

In the semiparametric model, it is difficult to estimate both the parameters of interest and nuisance parameters simultaneously, since $\boldsymbol{\xi}$ is of infinite degrees of freedom. The semiparametric approach suggests use of an estimating function to estimate the parameters of interest, regardless of the nuisance parameters. In general, the estimating function is a vector function, independent of nuisance parameters $\boldsymbol{\xi}$, satisfying certain conditions [15,5]. Generally speaking, it is not easy to find an estimating function. Amari and Kawanabe [10] studied the information geometry of estimating functions and provided a novel approach that we follow in this paper to find a family of estimating functions for blind deconvolution.

4.2. Semiparametric formulation for blind deconvolution

We now formulate blind deconvolution within the framework of semiparametric models. The joint probability density function $p(\mathbf{x};\mathbf{W},r)$ of sensor signal $\boldsymbol{x}$ is determined by the probability density function $r(\boldsymbol{s})$ and the demixing filter $\mathbf{W}(z)$. From the statistical point of view, the problem is to estimate $\mathbf{W}(z)$ or $\mathbf{H}^{\dagger}(z)$ from the observed data and the estimate includes two unknowns: one is the demixing filter $\mathbf{W}(z)$ which is the parameter of interest, and the other is the probability density function $r(\boldsymbol{s})$ of sources, which is the nuisance parameter in the present case. For the blind deconvolution problem, we usually assume that source signals are zero-mean,

$$E[s_i] = 0 \quad \text{for } i = 1, \ldots, n. \tag{22}$$

In addition, we generally impose constraints on the recovered signals to remove the indeterminacy,

$$E[k_i(s_i)] = 0 \quad \text{for } i = 1, \ldots, n. \tag{23}$$

A typical example of the constraint is $k_i(s_i) = s_i^4 - 1$.

Since the source signals are spatially mutually independent and temporally iid, the pdf $r(\boldsymbol{s})$ can be factored into the product form

$$r(\boldsymbol{s}) = \prod_{i=1}^{n} r(s_i). \tag{24}$$

The nuisance parameter $r(\boldsymbol{s})$, the probability density function of the source signals, is in a function space. In the semiparametric approach, it is not necessary to estimate the nuisance parameter. The problem reduces to finding a suitable estimating function. Remarkable progress has been made recently in the theory of semiparametric models [10,15] and it has been shown that the efficient score itself is an estimating function for blind separation. In this paper, we utilize the theory to derive a family of estimating functions.

5. Efficient score

In this section, we give an explicit form of the score function of interest parameter, by using a local nonholonomic reparameterization. We then derive an efficient score by projecting the score function into the subspace orthogonal to the nuisance tangent space.

5.1. Score function matrix and its representation

Assume that the mixing filter $\mathbf{H}(z)$ is in $\mathscr{M}(N)$. The blind deconvolution problem is to find a demixing FIR filter $\mathbf{W}(z)$ such that the output $\mathbf{y}(k)$ of the demixing model is maximally spatially mutually independent and temporarily i.i.d.. To this end, we first define score functions of log-likelihood with respect to $\mathbf{W}(z)$. Since the mixing model is a matrix FIR filter, we write an estimating function in the same matrix filter format

$$\mathbf{F}(\mathbf{x}; \mathbf{W}(z)) = \sum_{p=0}^{N} \mathbf{F}_p(\mathbf{x}; \mathbf{W}) z^{-p}, \tag{25}$$

where $\mathbf{F}_p(\mathbf{x}; \mathbf{W})$ are matrix functions of $\mathbf{x}$ and $\mathbf{W} = [\mathbf{W}_0, \mathbf{W}_1, \ldots, \mathbf{W}_N]$.

Now consider the $\mathbf{W}$-score function, which is a filter in $\mathscr{T}\mathscr{M}(N)$, defined by

$$\frac{\partial \log p(\mathbf{y}; \mathbf{W}, r)}{\partial \mathbf{W}(z)} = \sum_{p=0}^{N} \frac{\partial \log p(\mathbf{y}; \mathbf{W}, r)}{\partial \mathbf{W}_p} z^{-p}, \tag{26}$$

where $p(\mathbf{y}; \mathbf{W}, r)$ is the probability density function of $\mathbf{y}$, and $\partial \log p(\mathbf{y}; \mathbf{W}, r)/\partial \mathbf{W}_p$ denotes the gradient in matrix form, whose (i,j)-element is defined by $\partial \log p(\mathbf{y}; \mathbf{W}, r)/\partial \mathbf{W}_{pij}$.

Using Cardoso's relative gradient technique [20], we reparameterize the filter in a small neighborhood of the true mixing filter $\mathbf{H}(z)$ by using a new variable matrix filter as

$$\mathbf{H}(z) * (\mathbf{I} - \mathbf{X}(z)), \tag{27}$$

where $\mathbf{I}$ is the identity element of the manifold $\mathscr{M}(N)$. The variation $\mathbf{X}(z)$ represents a local coordinate system at the neighborhood $\mathscr{N}_{\mathbf{H}}$ of $\mathbf{H}(z)$ in the manifold $\mathscr{M}(N)$. The variation $\mathrm{d}\mathbf{H}(z)$ of $\mathbf{H}(z)$ is represented as $\mathrm{d}\mathbf{H}(z) = -\mathbf{H}(z) * \mathrm{d}\mathbf{X}(z)$ in terms of $\mathrm{d}\mathbf{X}(z)$. Letting $\mathbf{W}(z) = \mathbf{H}^{\dagger}(z)$, we obtain

$$\mathrm{d}\mathbf{X}(z) = \mathrm{d}\mathbf{W}(z) * \mathbf{W}^{\dagger}(z), \tag{28}$$

a nonholonomic differential variable [45] since (28) is not integrable. Denote the inner product of any two filters $\mathbf{X}(z)$ and $\mathbf{Y}(z)$ in tangent space $\mathscr{T}\mathscr{M}_{\mathbf{W}(z)}$ by $\langle \mathbf{X}(z), \mathbf{Y}(z) \rangle = \sum_{p=0}^{N} tr(\mathbf{X}_p^{\mathrm{T}} \mathbf{Y}_p)$. Consider the differential $\mathrm{d} \log p(y; \mathbf{W}, r)$ with respect to the new variables,

$$\mathrm{d} \log p(\mathbf{y}; \mathbf{W}, r) = \left\langle \frac{\partial \log p(\mathbf{y}; \mathbf{W}, r)}{\partial \mathbf{X}(z)}, \mathrm{d}\mathbf{X}(z) \right\rangle. \tag{29}$$

On the other hand, using relation (28), we have

$$\begin{aligned}&\mathrm{d}\log p(\mathbf{y};\mathbf{W},r)\\ &\quad=\left\langle\frac{\partial\log p(\mathbf{y};\mathbf{W},r)}{\partial\mathbf{W}(z)},\mathrm{d}\mathbf{W}(z)\right\rangle\\ &\quad=\left\langle\frac{\partial\log p(\mathbf{y};\mathbf{W},r)}{\partial\mathbf{W}(z)}*\mathbf{W}^{\mathrm{T}}(z^{-1}),\mathrm{d}\mathbf{X}(z)\right\rangle. \quad (30)\end{aligned}$$

Comparing the two equations (29) and (30), and using the invariant property of the differential expression, we deduce

$$\frac{\partial\log p(\mathbf{y};\mathbf{W},r)}{\partial\mathbf{X}(z)}=\frac{\partial\log p(\mathbf{y};\mathbf{W},r)}{\partial\mathbf{W}(z)}*\mathbf{W}^{\mathrm{T}}(z^{-1}). \quad (31)$$

Using the relation (18), we evaluate the score function at $\mathbf{X}(z)=\mathbf{0}$

$$\left.\frac{\partial\log p(\mathbf{y};\mathbf{W},r)}{\partial\mathbf{X}_{p,ij}}\right|_{\mathbf{X}(z)=\mathbf{0}}=\varphi_i(s_i(k))s_j(k-p), \quad (32)$$

where $\varphi_i(s_i)=-\,\mathrm{d}\log(r_i(s_i))/\mathrm{d}s_i$, $i=1,\ldots,n$. This can also be re-written in the compact form

$$\begin{aligned}\mathbf{U}(\mathbf{x};\mathbf{W}(z),r)&=\sum_{p=0}^{N}\mathbf{U}_p z^{-p}\\ &=\sum_{p=0}^{N}\boldsymbol{\varphi}(\mathbf{s})\mathbf{s}^{\mathrm{T}}(k-p)z^{-p}, \quad (33)\end{aligned}$$

where $\boldsymbol{\varphi}(\mathbf{s})=(\varphi_1(s_1),\ldots,\varphi_n(s_n))^{\mathrm{T}}$, and $\mathbf{s}$ is the source signal vector. It should be noted that the score function $\mathbf{U}(\mathbf{x};\mathbf{W}(z),r)$ generally depends on the sensor signals $\mathbf{x}(k)$ and the demixing filter $\mathbf{W}(z)$. However, by introducing the nonholonomic reparameterization, we derive a score function that only depends on output of the demixing model or the global transfer function $\mathbf{G}(z)$. This property is called the equivariance in blind separation of instantaneous mixtures [20]. The relative or the natural gradient of a cost function on the Riemannian manifold can be automatically derived from this nonholonomic representation [8,20,45].

5.2. Efficient scores

In general, the space spanned by the components of the score function (33) is not orthogonal to the nuisance tangent space, denoted by $\mathcal{T}^{\mathcal{N}}_{\mathbf{W}(z),r}$. A detailed discussion of the nuisance tangent space in the blind separation problem can be found in [5]. Since the nuisance tangent space $\mathcal{T}^{\mathcal{N}}_{\mathbf{W}(z),r}$ in blind deconvolution is the same as that in blind separation, we directly use the result in [5].

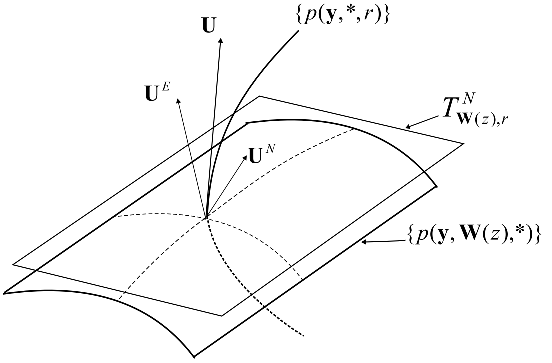

Fig. 3. Orthogonal decomposition of score functions.

Lemma 1 (Amari and Cardoso [5]). *The nuisance tangent space $\mathcal{T}^{\mathcal{N}}_{\mathbf{W}(z),r}$ is the linear space spanned by the nuisance score functions,*

$$\mathcal{T}^{\mathcal{N}}_{\mathbf{W}(z),r}=\left\{\sum_{i=1}^{n}c_i\alpha_i(s_i)\right\}, \quad (34)$$

where c_i are coefficients and α_i are arbitrary functions satisfying

$$\begin{aligned}&E_{r_i}[\{\alpha_i(s_i)\}^2]<\infty,\quad E_{r_i}[s_i\alpha_i(s_i)]=0,\\ &E_{r_i}[k_i(s_i)\alpha_i(s_i)]=0.\end{aligned} \quad (35)$$

The efficient scores, denoted by $\mathbf{U}^{\mathrm{E}}(\mathbf{x};\mathbf{W}(z),r)$, can be obtained by projecting the score function to the subspace orthogonal to the nuisance tangent space $\mathcal{T}^{\mathcal{N}}_{\mathbf{W}(z),r}$. See Fig. 3 for the illustration of orthogonal decomposition of score functions. Now we denote $u_{p,ij}=\varphi_i(s_i(k))s_j(k-p)$.

Lemma 2. *The off-diagonal elements $u_{0,ij}$, $i\neq j$, and the delay elements $u_{p,ij}$, $p\geqslant 1$, of the score functions are orthogonal to the nuisance tangent space $\mathcal{T}^{\mathcal{N}}_{\mathbf{W}(z),r}$.*

Proof. The inner product of $u_{0,ij}$ and any element $\sum c_l\alpha_l(y_l) \in \mathcal{T}^{\mathcal{N}}_{\mathbf{W}(z),r}$,

$$\left\langle u_{0,ij}, \sum c_l\alpha_l(s_l)\right\rangle = \sum c_l\langle \varphi(s_i)s_j, \alpha_l(s_l)\rangle \qquad (36)$$

vanishes because of mutual independence and the zero-mean of s_i and (35). Similarly, using the iid property of $\mathbf{s}$, we can prove that the inner product of $u_{p,ij}$ and any element $\sum c_l\alpha_l(s_l) \in \mathcal{T}^{N}_{\mathbf{H}(z),r}$ also vanishes for $p \geqslant 1$. □

Lemma 3. *The projection of $u_{0,ii}$ to the subspace orthogonal to the nuisance tangent space $\mathcal{T}^{\mathcal{N}}_{\mathbf{W}(z),r}$ is of the form*

$$w(s_i) = c_1 s_i + c_2 k_i(s_i), \qquad (37)$$

where c_i are any constants.

Proof. Using Lemma 1 and the arbitrariness of the $\alpha(s_i)$, $i = 1,\dots,n$, we see that the efficient scores in the diagonal elements $u_{0,ii}$ are given by (37). □

In summary we have the following theorem.

Theorem 1. *The efficient score, $\mathbf{U}^{\mathrm{E}}(\mathbf{x};\mathbf{W}(z),r)$ is expressed by*

$$\mathbf{U}^{\mathrm{E}}(\mathbf{x};\mathbf{W}(z),r) = \sum_{p=0}^{N} \mathbf{U}^{\mathrm{E}}_p z^{-p}, \qquad (38)$$

where

$$\mathbf{U}^{\mathrm{E}}_p = \boldsymbol{\varphi}(\mathbf{y})\mathbf{y}^{\mathrm{T}}(k-p) \quad \textit{for } p \geqslant 1; \qquad (39)$$

$$\mathbf{U}^{\mathrm{E}}_0 = \begin{cases} \boldsymbol{\varphi}(\mathbf{y})\mathbf{y}^{\mathrm{T}} & \textit{for off-diagonal elements}, \\ c_1 y_i + c_2 k_i(y_i) & \textit{for diagonal elements}. \end{cases} \qquad (40)$$

6. Estimating function and standardized estimating function

In this section, we derive a family of estimating functions and standardized estimating functions for blind deconvolution.

For instantaneous mixture, it has been proven [11] that the semiparametric model for blind separation is information m-curvature free. This is also true in multichannel blind deconvolution. As a result, the efficient score function is an estimating function. The derivative operator $\mathcal{K}(z) = E[\partial\mathbf{F}(\mathbf{x},\mathbf{W}(z))/\partial\mathbf{X}(z)]$ is a tensor filter, represented by

$$\mathcal{K}(z) = \sum_{p=0}^{N} \mathcal{K}_p z^{-p}. \qquad (41)$$

See Appendix A.2 for detailed derivation. We take the following notations

$$n_i = E[s_i^2\varphi_i'(s_i)], \quad \kappa_i = E[\varphi_i'(s_i)], \quad \sigma_i^2 = E[s_i^2], \qquad (42)$$

$$\gamma_{ij} = \kappa_i\kappa_j\sigma_i^2\sigma_j^2 - 1, \quad l_i = E[\varphi(s_i)]. \qquad (43)$$

Lemma 4. *The coefficients of operator $\mathcal{K}(z) = \sum_{p=0}^{N}\mathcal{K}_p z^{-p}$ can be expressed by*

$$\mathcal{K}_{p,ij,lm} = E[\varphi'(s_i(k))s_j^2(k-p)]\delta_{il}\delta_{jm} + \delta_{im}\delta_{jl}\delta_{0p}. \qquad (44)$$

Furthermore, if the following conditions are satisfied

$$\kappa_i \neq 0, \quad \kappa_i\kappa_j\sigma_i^2\sigma_j^2 - 1 \neq 0, \quad n_i + 1 \neq 0, \qquad (45)$$

then the derivative operator $\mathcal{K}(z)$ is invertible.

The proof is given in Appendix A.2. Therefore, we derive a family of estimating functions for blind deconvolution

$$\mathbf{F}(\mathbf{x}(k),\mathbf{W}(z)) = \sum_{p=0}^{N} \boldsymbol{\varphi}(\mathbf{y}(k))\mathbf{y}(k-p)^{\mathrm{T}}z^{-p} - \mathbf{I}, \qquad (46)$$

where $\mathbf{y}(k) = \sum_{p=0}^{N}\mathbf{W}_p\mathbf{x}(k-p)$, and $\boldsymbol{\varphi}$ is a vector of given activation functions, provided that the derivative operator $\mathcal{K}(z) = E[\partial\mathbf{F}(\mathbf{x},\mathbf{W}(z))/\partial\mathbf{X}(z)]$ is invertible. The estimating function is the efficient score function, when $c_1 = 0$, $c_2 = 1$ and $k_i(y_i) = \varphi_i(y_i)y_i - 1$.

The semiparametric approach suggests use of the following estimating equation [5,18] for parameters of interest,

$$\sum_{k=1}^{t} \mathbf{F}(\mathbf{x}(k), \mathbf{W}(z)) = \mathbf{0}. \tag{47}$$

The estimator obtained from (47) is called an M-estimator. An M-estimator is consistent, that is, the estimator $\mathbf{W}_t(z)$ converges to the true value as t tends to infinity without reference to $r(\mathbf{s})$. The estimating function is not unique, since that for any nonsingular linear operator $\mathscr{R}(z)$ mapping from $\mathscr{M}(N)$ to $\mathscr{M}(N)$, $\mathscr{R}(z)\mathbf{F}(\mathbf{x}, \mathbf{W}(z))$ is also an estimating function. It has already been established that the two estimating functions are equivalent in the sense that the derived batch estimators give exactly the same solution. This defines an equivalent class of estimating functions that are essentially the same in batch estimation. However, when we consider online learning, the learning dynamics is not equivalent and this necessitates introduction of an estimating function that will make the learning algorithm more stable and efficient. To this end, we introduce the concept of standardized estimating function. The standardized estimating function [4] is defined as follows: if the derivative operator $\mathscr{K}(z) = E[\partial \mathbf{F}(\mathbf{x}, \mathbf{W}(z))/\partial \mathbf{X}(z)]$ is an identity operator, the estimating function is called the standardized estimating function.

Lemma 5. *Given any estimating function* $\mathbf{F}(\mathbf{x}, \mathbf{W}(z))$, *if the operator* $\mathscr{K}(z)$ *is invertible*, *then*

$$\mathscr{K}^{-1}(z)\mathbf{F}(\mathbf{x}, \mathbf{W}(z)) \tag{48}$$

is a standardized estimating function.

The proof is not difficult. Using Lemma 5 we can derive a family of standardized estimating functions for the blind deconvolution problem.

Theorem 2. *Given an estimating function of form* (46), *the standardized estimating function is expressed by*

$$\mathbf{F}^*(\mathbf{x}, \mathbf{W}(z)) = \sum_{p=0}^{N} \mathbf{F}^*_p(\mathbf{x}, \mathbf{W}(z)) z^{-p}, \tag{49}$$

where

$$F^*_{0,ii} = \frac{1}{n_i + 1}\{\varphi_i(y_i)y_i - 1\} \quad \text{for } i = 1, \ldots, n, \tag{50}$$

$$F^*_{0,ij} = \frac{1}{\gamma_{ij}}\{\kappa_j \sigma_i^2 \varphi_i(y_i)y_j - \varphi_j(y_j)y_i\} \quad \text{for } i \neq j, \tag{51}$$

$$F^*_{p,ij} = \varphi_i(y_i)y_j(k-p)/(\kappa_i \sigma_j^2) \quad \text{for } p \geqslant 1. \tag{52}$$

Proof. In order to compute the inverse of the operator $\mathscr{K}(z)$, we consider the following equation:

$$\mathscr{K}(z)\mathbf{F}^*(\mathbf{x}, \mathbf{W}(z)) = \mathbf{F}(\mathbf{x}, \mathbf{W}(z)). \tag{53}$$

Using expression (44), we can rewrite (53) into the following component form

$$(n_i + 1)F^*_{0,ii} = F_{0,ii} \quad \text{for } i = 1, \ldots, n, \tag{54}$$

$$\kappa_i \sigma_j^2 F^*_{0,ij} + F^*_{0,ji} = F_{0,ij} \quad \text{for } i, j = 1, \ldots, n,\ i \neq j, \tag{55}$$

$$\kappa_i \sigma_j^2 F^*_{p,ij} = F_{p,ij} \quad \text{for } p \geqslant 1,\ i, j = 1, \ldots, n. \tag{56}$$

Solving the above equations, we obtain the results. □

There are some advantages to use the standardized estimating function in on-line learning. The natural gradient learning is given by

$$\Delta \mathbf{W}(z) = -\eta_k \mathbf{F}^*(\mathbf{x}, \mathbf{W}(z)) * \mathbf{W}(z). \tag{57}$$

It can be proved that the true solution $\mathbf{W}(z) = \mathbf{H}^\dagger(z)$ is always the stable equilibrium of the natural gradient learning above, provided conditions (45) are satisfied. The property is called universal convergence. See [6,4] for further information. The statistics in (42) and (43) require on-line estimate so as to implement learning algorithm (57). In particular, if the source signals are binary, taking values $1, -1$, we can calculate the statistics for the standardized estimating function. if we choose the cubic function $\varphi_i(y_i) = y_i^3$ as activation function, the statistics are evaluated by

$$n_i = 3, \quad \kappa_i = 3, \quad \sigma_i^2 = 1, \quad \gamma_{ij} = 8. \tag{58}$$

Therefore, the standardized estimating function can be given explicitly.

7. Performances of learning

In this section, we investigate the convergence rate of the learning process by using an estimating function. The basic idea is to use the two well-known theorems: the law of large numbers and the central limit theorem.

The Kronecker (tensor) product [27] is used extensively in the following discussion. The Kronecker product of two matrices $\mathbf{A}$ and $\mathbf{B}$ is denoted by $\mathbf{A}\otimes\mathbf{B}$. For any two matrix filters $\mathbf{X}(z)$ and $\mathbf{Y}(z)$ in $\mathscr{M}(N)$, their Kronecker product is defined as follows:

$$\mathbf{X}(z)\otimes\mathbf{Y}(z)=\sum_{p=0}^{N}(\mathbf{X}_p\otimes\mathbf{Y}_p)z^{-p}. \tag{59}$$

We now examine the statistical error analysis of the batch estimator. Suppose $\mathbf{W}(z)$ is the true solution to the estimating equation and $\mathbf{W}_t(z)$ is the solution to the empirically averaged equation

$$\sum_{k=1}^{t}\mathbf{F}(\mathbf{x}(k),\mathbf{W}(z))=0. \tag{60}$$

The estimator error is given by $\Delta\mathbf{W}_t(z)=\mathbf{W}_t(z)-\mathbf{W}(z)$. In order to simplify the analysis, we define the relative error in the nonholonomic form

$$\Delta\mathbf{X}_t(z)=\Delta\mathbf{W}_t(z)*\mathbf{W}^{\dagger}(z), \tag{61}$$

where $\mathbf{W}^{\dagger}(z)$ is the Lie group inverse of the filter $\mathbf{W}(z)$. The relative error has an explicit meaning, defining a global channel error in the following way:

$$\Delta\mathbf{y}=\Delta\mathbf{W}_t(z)\mathbf{x}=\Delta\mathbf{X}_t(z)\mathbf{s}. \tag{62}$$

Expanding (60) with respect to $\mathbf{X}(z)$, we have

$$\sum_{k=1}^{t}\mathbf{F}(\mathbf{x}(k),\mathbf{W}(z))+\sum_{k=1}^{t}\frac{\partial\mathbf{F}(\mathbf{x}(k),\mathbf{W}(z))}{\partial\mathbf{X}(z)}\Delta\mathbf{X}_t(z)=\mathbf{0}. \tag{63}$$

We rewrite the above equation in another form

$$\frac{1}{t}\sum_{k=1}^{t}\frac{\partial\mathbf{F}(\mathbf{x}(k),\mathbf{W}(z))}{\partial\mathbf{X}(z)}\Delta\mathbf{X}_t(z)=-\frac{1}{\sqrt{t}}\frac{1}{\sqrt{t}}\sum_{k=1}^{t}\mathbf{F}(\mathbf{x}(k),\mathbf{W}(z)). \tag{64}$$

According to the law of large numbers, we have the following estimation:

$$\frac{1}{t}\sum_{k=1}^{t}\frac{\partial\mathbf{F}(\mathbf{x}(k),\mathbf{W}(z))}{\partial\mathbf{X}(z)}=\mathscr{K}(z)+O\left(\frac{1}{\sqrt{t}}\right), \tag{65}$$

where $\mathscr{K}(z)=E[\partial\mathbf{F}(\mathbf{x},\mathbf{W}(z))/\partial\mathbf{X}(z)]$, is a filter to filter operator, given by Lemma 4. In the following discussion, we presume that conditions (45) are satisfied. On the other hand, since $\mathbf{F}(\mathbf{x},\mathbf{W}(z))$ is an estimating function, its expectation vanishes. The central limit theorem guarantees that

$$\frac{1}{\sqrt{t}}\sum_{k=1}^{t}\mathbf{F}(\mathbf{x}(k),\mathbf{W}(z)) \tag{66}$$

converges in distribution to the normal random variable matrix, denoted by $\mathbf{V}(z,t)$, with mean $\mathbf{0}$ and covariance matrix

$$\mathscr{G}(z)=\sum_{p=0}^{N}\mathscr{G}_p z^{-p}=\sum_{p=0}^{N}E[\mathbf{F}_p(\mathbf{x},\mathbf{W}(z))\otimes\mathbf{F}_p^{\mathrm{T}}(\mathbf{x},\mathbf{W}(z))]z^{-p}, \tag{67}$$

where $\otimes$ is the Kronecker product of two matrices.

Lemma 6. *The covariance of the error measured in term of $\Delta\mathbf{X}_t(z)$ of the estimator $\mathbf{W}_t(z)$ is asymptotically given by*

$$E[\Delta\mathbf{X}_t(z)\otimes\Delta\mathbf{X}_t^{\mathrm{T}}(z)]=\frac{1}{t}\mathscr{K}^{-1}(z)\mathscr{G}(z)\mathscr{K}^{-T}(z)+O\left(\frac{1}{t^2}\right). \tag{68}$$

Proof. Substituting (65) into (64) and using (67), we have

$$\Delta \mathbf{X}_t(z) = -\frac{1}{\sqrt{t}}\mathscr{K}^{-1}\{\mathbf{V}(z,t)\} + O\left(\frac{1}{t}\right). \qquad (69)$$

Taking the covariance of $\Delta\mathbf{X}_t(z)$, we have estimation (68). □

From Lemma 6 we can estimate the covariance of the recovered signals.

Lemma 7. *For $i \neq j$, the cross covariance is given by*

$$V_{ij}^t = E[y_i y_j] = \sum_{l=1}^{n}\sum_{p=0}^{N} E[\Delta X_{p,il}^t \Delta X_{p,jl}^t]\sigma_l^2, \qquad (70)$$

where $\sigma_l^2 = E[s_l^2]$.

Proof. Using the i.i.d. property of $\mathbf{s}(t)$ and relation (62), we have, for $i \neq j$

$$\begin{aligned} E[y_i y_j] &= E[(s_i + \Delta y_i)(s_j + \Delta y_j)] \\ &= \sum_{l=1}^{n} E[\Delta X_{il}(z) s_l(t) \Delta X_{jl}(z) s_l(t)] \\ &= \sum_{l=1}^{n}\sum_{p=0}^{N} E[\Delta X_{p,il}^t \Delta X_{p,jl}^t]\sigma_l^2. \quad \square \qquad (71) \end{aligned}$$

From Lemmas 6 and 7 we know that generally, the covariance $V_{ij}(t) = E[y_i y_j]$ $(i \neq j)$ vanishes at rate $1/t$, as t tends to infinity.

8. Superefficiency of batch estimator

Amari [4] proves that in the instantaneous case, the covariance $V_{ij}(t) = E[y_i y_j]$ $(i \neq j)$ vanishes at rate $1/t^2$ under certain simple conditions. This property is called superefficiency. In this section, we prove that superefficiency remains valid in blind deconvolution.

Suppose that $\mathbf{F}^*(\mathbf{x}, \mathbf{W}(z))$ is a standardized estimating function:

$$E[\Delta\mathbf{X}_t(z) \otimes \Delta\mathbf{X}_t^{\mathrm{T}}(z)] = \frac{1}{t}\mathscr{G}^*(z) + O\left(\frac{1}{t^2}\right), \qquad (72)$$

where $\mathscr{G}^*(z) = \mathscr{K}^{-1}(z)\mathscr{G}(z)\mathscr{K}^{-T}(z) = E[\mathbf{F}^*(\mathbf{x}, \mathbf{W}(z)) \otimes \mathbf{F}^{*\mathrm{T}}(\mathbf{x}, \mathbf{W}(z))]$.

Lemma 8. *The coefficients of $\mathscr{G}^*(z)$ are expressed by*

$$G_{0,il,jl}^* = c_{il}c_{jl}\sigma_i^2\sigma_j^2\sigma_l^2 k_l^2 l_i l_j \quad \text{for } i \neq j,\ j \neq l,\ l \neq i, \qquad (73)$$

$$G_{0,ii,ji}^* = \frac{1}{n_i + 1} c_{ji}\kappa_i\sigma_j^2 l_j E[s_i^2\varphi_i(s_i)] \quad \text{for } i \neq j, \qquad (74)$$

$$G_{p,il,jl}^* = \frac{l_i l_j}{\kappa_i \kappa_j} \quad \text{for } p \geqslant 1,\ i,j = 1,\dots,n. \qquad (75)$$

Proof. Using the expression of $\mathbf{F}^*(\mathbf{x}, \mathbf{W}(z))$ in Theorem 2, we derive the result by direct calculation. □

Theorem 3. *A batch estimator is superefficient when the following condition is satisfied*

$$l_i = E[\varphi_i(s_i)] = 0 \quad \text{for } i = 1,\dots,n. \qquad (76)$$

Proof. Using Lemma 8 and (76), we have

$$G_{p,il,jl}^* = 0 \quad \text{for } i \neq j,\ p = 0,\dots,N,\ l = 1,\dots,n. \qquad (77)$$

Write the estimate (72) in component form,

$$E[\Delta X_{p,il}^t \Delta X_{p,jl}^t] = \frac{1}{t}G_{p,il,jl}^* + O\left(\frac{1}{t^2}\right) = O\left(\frac{1}{t^2}\right), \qquad (78)$$

for $i \neq j$. The combination of (70) and (78) leads to the following estimation

$$V_{ij}^t = \sum_{l=1}^{n}\sum_{p=0}^{N} E[\Delta X_{p,il}^t \Delta X_{p,jl}^t]\sigma_l^2 = O\left(\frac{1}{t^2}\right). \qquad (79)$$

This proves our result. □

9. Superefficiency in on-line learning

Now we turn to superefficiency in on-line learning. The natural gradient learning algorithm is described as follows [45]:

$$\mathbf{W}_{t+1}(z) = \mathbf{W}_t(z) - \eta_t \mathbf{F}(\mathbf{x}(t), \mathbf{W}_t(z)) * \mathbf{W}_t(z), \quad (80)$$

where $\mathbf{F}(\mathbf{x}, \mathbf{W}(z))$ is an estimating function in the form (46). For simplicity, we reparameterize the demixing model in the nonholonomic form $\mathrm{d}\mathbf{X}(z) = \mathrm{d}\mathbf{W}(z) * \mathbf{W}^{\dagger}(z)$ and the learning algorithm for $\mathbf{X}(z)$ is described by

$$\mathbf{X}_{t+1}(z) = \mathbf{X}_t(z) - \eta_t \mathbf{F}(\mathbf{x}(t), \mathbf{W}_t(z)). \quad (81)$$

The local stability conditions for the learning algorithm (80) is given by the following lemma.

Lemma 9 (Zhang et al. [45]). *If for* $i, j = 1, \ldots, n$,

$$n_i + 1 > 0, \quad \kappa_i > 0, \quad \kappa_i \kappa_j \sigma_i^2 \sigma_j^2 > 1, \quad (82)$$

then the natural gradient learning algorithm (80) *is locally stable.*

From statistical learning theory, provided the learning algorithm is stable, the expectation of $\mathbf{W}_t(z)$ converges to the optimal solution exponentially when the learning rate η_t is fixed to a small constant η. However, even when t is large, $\mathbf{W}_t(z)$ still fluctuates around optimal value for a fixed learning rate η. For the instantaneous mixture, Amari [4] analyses the covariance matrices of $\Delta\mathbf{X}_t$ and proves the superefficiency in on-line learning. In this section we extend the result to blind deconvolution.

Theorem 4. *When η is sufficiently small and $\mathbf{W}_t(z)$ converges to the true solution, the covariance matrix of the relative error $\Delta\mathbf{X}_t(z) = (\mathbf{W}_t(z) - \mathbf{W}(z)) * \mathbf{W}^{\dagger}(z)$ converges to*

$$E[\Delta\mathbf{X}(z) \otimes \Delta\mathbf{X}^{\mathrm{T}}(z)] = \eta \mathscr{Y}(z) + O(\eta^2), \quad (83)$$

where $\mathscr{Y}(z)$ is a 4-dimensional tensor filter, defined by the solution of

$$\mathscr{K}(z)\mathscr{Y}(z) + \mathscr{Y}(z)\mathscr{K}^{\mathrm{T}}(z) = \mathscr{G}(z). \quad (84)$$

Proof. See Appendix A.3. □

For on-line learning, superefficiency is defined in a similar way: a learning algorithm is superefficient if the cross covariance $V_{ij}^t = E[y_i(t) y_j(t)]$, $i \neq j$, is of the order η^2 for sufficiently large t.

Theorem 5. *Assume that stability conditions* (82) *are satisfied. Superefficiency holds for the natural gradient learning algorithm* (80) *when the following conditions are satisfied*:

$$l_i = E[\varphi_i(s_i)] = 0 \quad \textit{for } i = 1, \ldots, n. \quad (85)$$

Proof. From Lemma 7 and Theorem 4, we have

$$V_{ij}^t = \sum_{l=1}^{n} \sum_{p=0}^{N} E[\Delta X_{p,il}^t \Delta X_{p,jl}^t] \sigma_l^2$$
$$= \eta \sum_{l=1}^{n} \sum_{p=0}^{N} \mathscr{Y}_{p,il,jl} \sigma_l^2 + O(\eta^2). \quad (86)$$

It is proven in Appendix A.3 that for any $i \neq j$, $\mathscr{Y}_{p,il,jl} = 0$. This yields

$$V_{ij}^t = O(\eta^2) \quad \text{for } i \neq j. \quad \square \quad (87)$$

From the arguments above we can see that superefficiency of both batch estimator and natural gradient algorithm require (85) and fortunately the commonly used activation functions, such as the cubic function and the hyperbolic tangent function, satisfy these conditions.

10. Computer simulations

To evaluate performance of the proposed learning algorithms, we employ the multichannel intersymbol interference [33], denoted by M_{ISI}, as a criteria,

$$M_{\mathrm{ISI}} = \sum_{i=1}^{n} \frac{|\sum_{j=1}^{n} \sum_{p=0}^{N} |\mathbf{G}_{pij}|^2 - \max_{p,j} |\mathbf{G}_{pij}|^2}{\max_{p,j} |\mathbf{G}_{pij}|^2}$$
$$+ \sum_{j=1}^{n} \frac{|\sum_{i=1}^{n} \sum_{p=0}^{N} |\mathbf{G}_{pij}|^2 - \max_{p,i} |\mathbf{G}_{pij}^2|}{\max_{p,i} |\mathbf{G}_{pij}|^2}. \quad (88)$$

It is easy to show that $M_{\mathrm{ISI}} = 0$ if and only if $\mathbf{G}(z)$ is of the form (10). In order to remove the effect

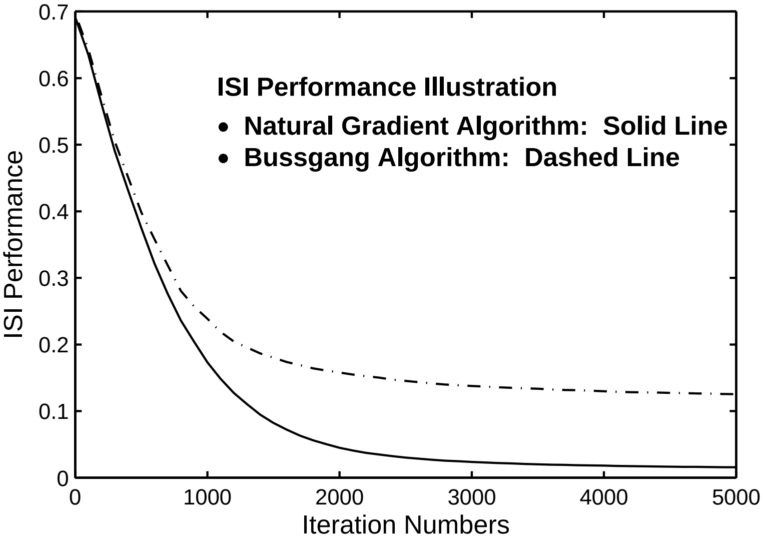

Fig. 4. M_{ISI} performance of the natural gradient algorithm.

of a single numerical trial on evaluating the performance of algorithms, we use the ensemble average approach, that is, in each trial we obtain a time sequence of M_{ISI}, and take average of ISI performance to evaluate the performance of the algorithms.

A large number of computer simulations have been performed to evaluate validity and performance of the proposed natural gradient algorithm and we give two examples to demonstrate the behavior and performance of algorithm (80). In both examples the mixing model is a multichannel ARMA model as follows:

$$\mathbf{x}(k) + \sum_{i=1}^{N} \mathbf{A}_i \mathbf{x}(k-i) = \sum_{i=0}^{N} \mathbf{B}_i \mathbf{s}(k-i) + \mathbf{v}(k), \qquad (89)$$

where $\mathbf{x}, \mathbf{s}, \mathbf{v} \in \mathbf{R}^3$. The matrices $\mathbf{A}_i$ and $\mathbf{B}_i$ are randomly chosen such that the mixing system is stable. The nonlinear activation function is chosen to be $\boldsymbol{\varphi}(y) = y^3$.

Example 1. In this simulation, we randomly generate ARMA model of form (89) by computer, and employ natural gradient algorithm (80) and the Bussgang algorithm [14] to train the demixing filter, respectively. The source signals **s** are randomly generated i.i.d. signals uniformly distributed in the range $(-1, 1)$, and $\boldsymbol{v}$ is chosen as Gaussian noise with zero mean and covariance matrix $0.1\boldsymbol{I}$.

Fig. 4 illustrates 100 trial ensemble average M_{ISI} performance of the natural gradient 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 20 000 since there is a long plateau in the Bussgang equalizer.

Example 2. Assume that source signals are i.i.d. quadrature amplitude modulated (QAM). The transfer function of the randomly chosen mixing system is plotted in Fig. 5, which is assumed to be unknown during learning. We use the natural gradient algorithm with the standardized estimating function to training the demixing filter. The learning rate is set to 0.001.

Fig. 6 illustrates the output signal constellations of the natural gradient learning in three time intervals $1 \leqslant k \leqslant 300$, $1001 \leqslant k \leqslant 1300$ and $2001 \leqslant k \leqslant 2300$, respectively. It is worth noting that the output signals converge to the characteristic QAM constellation, up to an amplitude and phase rotation factors ambiguities.

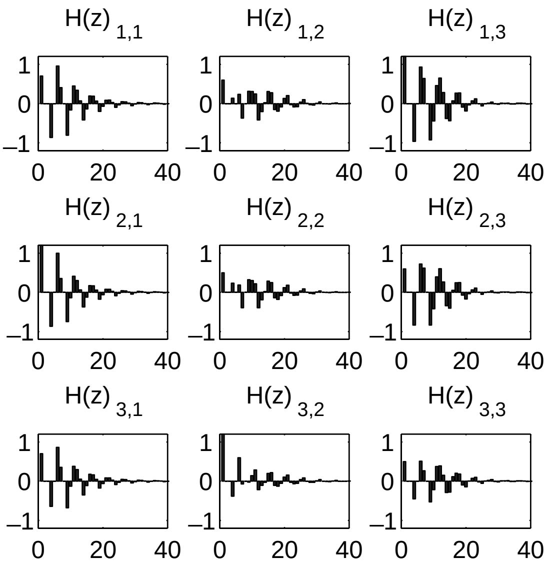

Fig. 5. The coefficients of $\mathbf{H}(z)$ of the mixing systems.

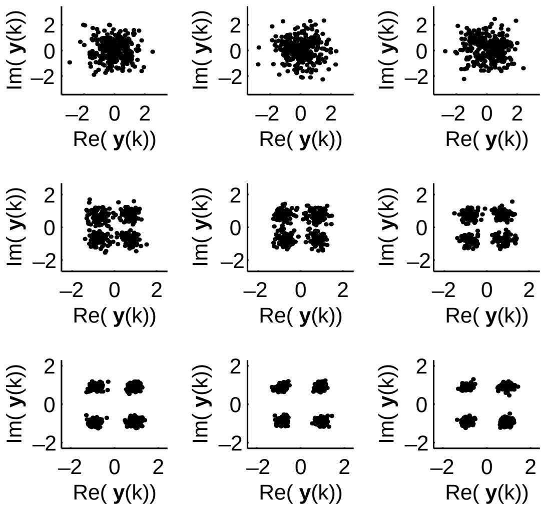

Fig. 6. Output constellations.

11. Conclusion

In this paper, we present the semiparametric approach to blind deconvolution, and study the convergence and efficiency of the batch estimator and natural gradient learning. First, multichannel blind deconvolution is formulated in the framework of the semiparametric model and a family of estimating functions and standardized estimating functions are derived by using efficient score functions. The advantage of using the semiparametric approach is that we do not need to estimate the nuisance parameters — the probability density functions of source signals in blind deconvolution. It is inferred from the theory of estimating functions that the batch estimator of the estimating equation converges to the true solution as the number of observed data tends to infinity. If stability conditions are satisfied, the natural gradient learning also converges to the true solution whatever the probability density function of the source signals. The superefficiency of both the batch estimator and natural gradient learning is proven when conditions (85) are satisfied. Finally, computer simulations are given to demonstrate the validity and effectiveness of the natural gradient approach.

Appendix A

11.1. Definition

In this appendix, we introduce some basic definitions and concepts in this paper. Assume that $\mathbf{X}(z)=\sum_{p=0}^{N}\mathbf{X}_p z^{-p}$ is a filter in $\mathscr{M}(N)$, and $l(\mathbf{X}(z))$ is a cost function defined on $\mathscr{M}(N)$. The derivative of $l(\mathbf{X}(z))$ with respect to a matrix $\mathbf{X}_p=(X_{p,ij})_{n\times n}$ is defined by

$$\frac{\partial l(\mathbf{X}(z))}{\partial \mathbf{X}_p}=\left(\frac{\partial l(\mathbf{X}(z))}{\partial X_{p,ij}}\right)_{n\times n}. \tag{A.1}$$

The derivative of $l(\mathbf{X}(z))$ with respect to a filter $\mathbf{X}(z)$ is defined by

$$\frac{\partial l(\mathbf{X}(z))}{\partial \mathbf{X}(z)}=\sum_{p=0}^{N}\frac{\partial l(\mathbf{X}(z))}{\partial \mathbf{X}_p}z^{-p}. \tag{A.2}$$

The estimating function for blind deconvolution is denoted by

$$\mathbf{F}(\mathbf{y},\mathbf{X}(z))=\sum_{p=0}^{N}\mathbf{F}_p(\mathbf{y},\mathbf{X}(z))z^{-p}, \tag{A.3}$$

where $\mathbf{F}_p \in \mathbf{R}^{n\times n}$, $p=0,\dots,N$ are matrix functions on $\mathscr{M}(N)$. Given p,q, the derivative $\partial \mathbf{F}_p/\partial \mathbf{X}_q$ is a 4-dimensional tensor, defined by $\partial \mathbf{F}_p/\partial \mathbf{X}_q = (\partial F_{p,ij}/\partial X_{q,lk})_{n\times n\times n\times n}$. For any matrix $\mathbf{Y}\in \mathbf{R}^{n\times n}$, the operation $(\partial \mathbf{F}_p/\partial \mathbf{X}_q)\mathbf{Y}$ is defined by $(\partial \mathbf{F}_p/\partial \mathbf{X}_q)\mathbf{Y} = \sum_{l,k}(\partial \mathbf{F}_p/\partial X_{q,lk})Y_{lk}$ Therefore, the derivative $\partial \mathbf{F}(\mathbf{y},\mathbf{X}(z))/\partial \mathbf{X}(z)$ is an operator mapping $\mathscr{M}(N)$ to $\mathscr{M}(N)$, defined by

$$\frac{\partial \mathbf{F}(\mathbf{y},\mathbf{X}(z))}{\partial \mathbf{X}(z)}\mathbf{Y}(z) = \sum_{p=0}^{N}\sum_{q=0}^{N}\frac{\partial \mathbf{F}_p}{\partial \mathbf{X}_q}\mathbf{Y}_q z^{-p} \tag{A.4}$$

for any filter $\mathbf{Y}(z)\in\mathscr{M}(N)$.

11.2. Representation of operator $\mathscr{K}(z)$

We derive the explicit form of operator $\mathscr{K}(z)$ and its inverse $\mathscr{K}^{-1}(z)$ and give a definition of the transpose $\mathscr{K}^{\mathrm{T}}(z)$ of $\mathscr{K}(z)$ here. Assume that the recovered signal $\mathbf{y}(k)$ is spatially mutually independent and temporally i.i.d..

Lemma 10. *For any $p \neq q$,*

$$E\left[\frac{\partial \mathbf{F}_p}{\partial \mathbf{X}_q}\right] = \mathbf{0}. \tag{A.5}$$

Proof. By definition, $F_{p,ij} = \varphi(y_i)y_j(k-p) - \delta_{0,p}$. Using the i.i.d properties of $\mathbf{y}(k)$ and relation (18), we have, for $p\neq q$,

$$E\left[\frac{\partial F_{p,ij}}{\partial X_{q,lm}}\right] = E\left[\varphi'(y_i)\frac{\partial y_i(k)}{\partial X_{q,lm}}y_j(k-p) + \varphi(y_i)\frac{\partial y_j(k-p)}{\partial X_{q,lm}}\right] = 0. \tag{A.6}$$

Proposition 1. *The derivative operator $\mathscr{K}(z)$ can be represented as*

$$\mathscr{K}(z) = \sum_{p=0}^{N}\mathscr{K}_p z^{-p} = \sum_{p=0}^{N} E\left[\frac{\partial \mathbf{F}_p}{\partial \mathbf{X}_p}\right] z^{-p}, \tag{A.7}$$

which maps $\mathbf{Y}(z)\in\mathscr{M}(N)$ to $\mathscr{K}(z)\mathbf{Y}(z) = \sum_{p=0}^{N}\mathscr{K}_p\mathbf{Y}_p z^{-p}$. Furthermore, the coefficients of $\mathscr{K}(z)$ are given by

$$\mathscr{K}_{p,ij,lm} = E[\varphi'(y_i(k))y_j^2(k-p)]\delta_{il}\delta_{jm} + \delta_{im}\delta_{jl}\delta_{0p}. \tag{A.8}$$

Proof. From definition (A.4) and using (A.5), we have

$$\mathscr{K}(z)\mathbf{Y}(z) = \sum_{p=0}^{N}\sum_{q=0}^{N} E\left[\frac{\partial \mathbf{F}_p}{\partial \mathbf{X}_q}\right]\mathbf{Y}_q z^{-p} = \sum_{p=0}^{N} E\left[\frac{\partial \mathbf{F}_p}{\partial \mathbf{X}_p}\right]\mathbf{Y}_p z^{-p}. \tag{A.9}$$

Using the i.i.d properties of $\mathbf{y}(k)$ and (18), we have

$$E\left[\frac{\partial F_{p,ij}}{\partial X_{p,lm}}\right] = E\left[\varphi'(y_i)\frac{\partial y_i(k)}{\partial X_{p,lm}}y_j(k-p) + \varphi(y_i)\frac{\partial y_j(k-p)}{\partial X_{p,lm}}\right] = E[\varphi'(y_i(k))y_j^2(k-p)]\delta_{il}\delta_{jm} + \delta_{im}\delta_{jl}\delta_{0p}. \tag{A.10}$$

The result follows.

In order to calculate the inverse of $\mathscr{K}(z)$, consider the following equation:

$$\mathscr{K}(z)\mathbf{X}(z) = \mathbf{Y}(z), \tag{A.11}$$

where $\mathbf{X}(z)$ and $\mathbf{Y}(z)\in\mathscr{M}(N)$. Substitute (A.8) into (A.11), and write it in component form

$$(n_i+1)X_{0,ii} = Y_{0,ii} \quad \text{for } i=1,\dots,n, \tag{A.12}$$

$$\kappa_i\sigma_j^2 X_{0,ij} + X_{0,ji} = Y_{0,ij} \quad \text{for } i,j=1,\dots,n,\ i\neq j, \tag{A.13}$$

$$\kappa_i\sigma_j^2 X_{p,ij} = Y_{p,ij} \quad \text{for } p\geqslant 1,\ i,j=1,\dots,n. \tag{A.14}$$

We can directly solve $X_{0,ii}$ and $X_{p,ij}$ from (A.12) and (A.14). For $X_{0,ij}$, $i\neq j$, we can write (A.12) in the following 2×2 self-closed subsystem

$$\begin{bmatrix}\kappa_i\sigma_j^2 & 1\\ 1 & \kappa_j\sigma_i^2\end{bmatrix}\begin{bmatrix}X_{0,ij}\\ X_{0,ji}\end{bmatrix} = \begin{bmatrix}Y_{0,ij}\\ Y_{0,ji}\end{bmatrix}. \tag{A.15}$$

If $\gamma_{ij} = \kappa_i\kappa_j\sigma_i^2\sigma_j^2 - 1 \neq 0$, we can uniquely solve the above equations. Therefore, we have the following result. □

Proposition 2. *If $n_i+1 \neq 0, \kappa_i \neq 0, \gamma_{ij} = \kappa_i\kappa_j\sigma_i^2\sigma_j^2 - 1 \neq 0$, then the operator $\mathscr{K}(z)$ is invertible and the inverse $\mathscr{K}^{-1}(z) = \sum_{p=0}^{N} \mathscr{R}_p z^{-p}$ is expressed by*

$$\mathscr{R}_{0,ii,lm} = \frac{1}{n_i+1}\delta_{il}\delta_{im}, \tag{A.16}$$

$$\mathscr{R}_{p,ij,lm} = \frac{1}{\kappa_i\sigma_j^2}\delta_{il}\delta_{jm}, \tag{A.17}$$

$$\mathscr{R}_{0,ij,lm} = \frac{1}{\gamma_{ij}}(\kappa_j\sigma_i^2\delta_{il}\delta_{jm} - \delta_{im}\delta_{jl}). \tag{A.18}$$

Now we give a definition of the transpose operation of tensor filters. The transpose of a tensor filter $\mathscr{K}(z)$ is given by

$$\mathscr{K}^{\mathrm{T}}(z) = \sum_{p=0}^{N} \mathscr{K}_p^{\mathrm{T}} z^{-p}, \tag{A.19}$$

where $\mathscr{K}_p^{\mathrm{T}} = (\mathscr{K}_{p,lm,ij})$, given $\mathscr{K}_p = (\mathscr{K}_{p,ij,lm})$.

11.3. Derivation and properties of operator $\mathscr{Y}(z)$

We derive the estimation (83) and discuss some properties of $\mathscr{Y}(z)$ here. Using the nonholonomic reparameterization, we have the following on-line learning rule:

$$\mathbf{X}_{t+1}(z) = \mathbf{X}_t(z) - \eta\mathbf{F}(\mathbf{x}(t), \mathbf{W}_t(z)). \tag{A.20}$$

By the Taylor expansion, we have

$$\Delta\mathbf{X}_{t+1}(z) = \Delta\mathbf{X}_t(z) - \eta\left(\mathbf{F}(\mathbf{x}(t), \mathbf{W}(z)) + \frac{\partial\mathbf{F}(\mathbf{x}(t), \mathbf{W}(z))}{\partial\mathbf{X}(z)}\Delta\mathbf{X}_t(z)\right). \tag{A.21}$$

The error covariance at time t is denoted by

$$\mathscr{V}^t(z) = E[\Delta\mathbf{X}_t(z) \otimes \Delta\mathbf{X}_t^{\mathrm{T}}(z)]. \tag{A.22}$$

Substituting expansion (A.21) into (A.22), we have

$$\begin{aligned}\mathscr{V}^{t+1}(z) = \mathscr{V}^t(z) &- \eta(\mathscr{K}(z)E[\Delta\mathbf{X}_t(z) \otimes \Delta\mathbf{X}_t^{\mathrm{T}}(z)] \\ &+ E[\Delta\mathbf{X}_t(z) \otimes \Delta\mathbf{X}_t^{\mathrm{T}}(z)]\mathscr{K}^{\mathrm{T}}(z)) \\ &+ \eta^2 E[\mathbf{F}(\mathbf{x}(t), \mathbf{W}(z)) \otimes \mathbf{F}^{\mathrm{T}}(\mathbf{x}(t), \mathbf{W}(z))] \\ &+ O(\eta^3).\end{aligned} \tag{A.23}$$

Therefore, when $\mathbf{W}_t(z)$ converges to $\mathbf{W}(z)$, for sufficiently large t, we have

$$\mathscr{K}(z)\mathscr{V}^t(z) + \mathscr{V}^t(z)\mathscr{K}^{\mathrm{T}}(z) = \eta\mathscr{G}(z) + O(\eta^2). \tag{A.24}$$

Assume that the filter operator $\mathscr{P}(z)$, defined by

$$\mathscr{P}(z)\mathscr{Y}(z) = \mathscr{K}(z)\mathscr{Y}(z) + \mathscr{Y}(z)\mathscr{K}^{\mathrm{T}}(z), \tag{A.25}$$

is invertible. Combining (A.24) and (84) we obtain the estimation (83).

Proposition 3. *Assume that the filter operator $\mathscr{P}(z)$ is invertible. If the following conditions are satisfied*

$$l_i = E[\varphi_i(y_i)] = 0 \quad \textit{for } i = 1, \ldots, n, \tag{A.26}$$

then for any $i \neq j$,

$$\mathscr{Y}_{p,il,jl} = 0. \tag{A.27}$$

Proof. Under condition (A.26), we have, for $i \neq j$,

$$\mathscr{G}_{p,il,jl} = 0 \quad \text{for } p = 0, \ldots, N. \tag{A.28}$$

From (84), we have

$$\mathscr{K}_p\mathscr{Y}_p + \mathscr{Y}_p\mathscr{K}_p^{\mathrm{T}} = \mathscr{G}_p \quad \text{for } p = 0, \ldots, N. \tag{A.29}$$

Rewriting the above equation system into component form, we see that the system can be separated into 2 × 2 or 4 × 4 self-closed subsystems. If $p = 0,\ i = l,\ j \neq l$, we solve $\mathscr{Y}_{p,ii,ji}$ by the following subsystem:

$$\begin{bmatrix} n_i + \kappa_{ij} + 1 & 1 \\ 1 & n_i + \kappa_{ji} + 1 \end{bmatrix} \begin{bmatrix} \mathscr{Y}_{0,ii,ij} \\ \mathscr{Y}_{0,ii,ji} \end{bmatrix} = \begin{bmatrix} \mathscr{G}_{0,ii,ij} \\ \mathscr{G}_{0,ii,ji} \end{bmatrix}. \tag{A.30}$$

If the following nonsingular conditions are satisfied

$$(n_i + \kappa_{ij} + 1)(n_i + \kappa_{ji} + 1) - 1 \neq 0 \quad \text{for } i \neq j, \tag{A.31}$$

we deduce that

$$\mathscr{Y}_{0,ii,ji} = 0, \tag{A.32}$$

for $i \neq j$. Similarly, we can verify that (A.27) holds for any other cases if $\mathscr{P}(z)$ is invertible. □

References

[1] K. Abed-Meraim, J.F. Cardoso, A. Gorokhov, P. Loubaton, E. Moulines, On subspace methods for blind identification of SIMO-FIR systems, IEEE Trans. Signal Process. 45 (1997) 42–56.

[2] S. Amari, in: Differential–geometrical methods in statistics, Lecture Notes in Statistics, Vol. 28, Springer, Berlin, 1985.

[3] S. Amari, Natural gradient works efficiently in learning, Neural Comput. 10 (1998) 251–276.

[4] S. Amari, Superefficiency in blind source separation, IEEE Trans. Signal Process. 47 (4) (April 1999) 936–944.

[5] S. Amari, J.-F. Cardoso, Blind source separation — semiparametric statistical approach, IEEE Trans. Signal Process. 45 (November 1997) 2692–2700.

[6] S. Amari, T. Chen, A. Cichocki, Stability analysis of adaptive blind source separation, Neural Networks 10 (1997) 1345–1351.

[7] S. Amari, A. Cichocki, Adaptive blind signal processing — neural network approaches, Proc. IEEE 86 (10) (1998) 2026–2048.

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

[9] S. Amari, S. Douglas, A. Cichocki, H. Yang, Novel on-line algorithms for blind deconvolution using natural gradient approach, in: Proceedings of the 11th IFAC Symposium on System Identification, SYSID'97, Kitakyushu, Japan, July 8–11 1997, pp. 1057–1062.

[10] S. Amari, M. Kawanabe, Information geometry of estimating functions in semiparametric statistical models, Bernoulli 3 (1) (1997) 29–54.

[11] S. Amari, M. Kawanabe, Estimating functions in semiparametric statistical models, in: I.V. Basawa, V. Godambe, R. Taylor (Eds.), Estimating Functions, Monograph Series, Vol. 32, IMS, 1998, pp. 65–81.

[12] S. Amari, M. Kumon, Estimation in the presence of infinitely many nuisance parameters in semiparametric statistical models, Ann. Statist. 16 (1988) 1044–1068.

[13] A. Bell, T. Sejnowski, An information maximization approach to blind separation and blind deconvolution, Neural Comput. 7 (1995) 1129–1159.

[14] S. Bellini, Bussgang techniques for blind deconvolution and equalization, in: S. Haykin (Ed.), Blind Deconvolution, Prentice-Hall, New Jersey, 1994, pp. 8–59.

[15] P. Bickel, C. Klaassen, Y. Ritov, J. Wellner, Efficient and Adaptive Estimation for Semiparametric Models, The Johns Hopkins University Press, Baltimore and London, 1993.

[16] W.M. Boothby, An Introduction to Differential Manifolds and Riemannian Geometry, Academic Press, Inc., New York, 1986.

[17] J.A. Cadzow, Blind deconvolution vis cumulant extrema, IEEE Signal Process. Mag. 13 (1996) 24–42.

[18] J.-F. Cardoso, Estimating equations for source separation, in: Proceedings of the ICASSP'97, Vol. 5, Munich, 1997, pp. 3449–3452.

[19] J.-F. Cardoso, Blind signal separation: statistical principles, Proc. IEEE 86 (10) (1998) 2009–2025.

[20] J.-F. Cardoso, B. Laheld, Equivariant adaptive source separation, IEEE Trans. Signal Process. SP-43 (December 1996) 3017–3029.

[21] A. Cichocki, R. Unbehauen, E. Rummert, Robust learning algorithm for blind separation of signals, Electron. Lett. 30 (17) (1994) 1386–1387.

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

[23] L. Conlon, Differential Manifolds, Birkhauser, Boston, 1993.

[24] N. Delfosse, P. Loubaton, Adaptive blind separation of independent sources: a deflation approach, Signal Processing 45 (1995) 59–83.

[25] A. Gorokhov, P. Loubaton, Blind identification of MIMO-FIR system: a generalized linear prediction approach, Signal Processing 73 (1999) 105–124.

[26] S. Haykin, Unsupervised Adaptive Filtering, Vol. II: blind deconvolution, Wiley, New York, 2000.

[27] R.A. Horn, C.R. Johnson, Topics in Matrix Analysis, Cambridge University Press, Cambridge, 1991.

[28] Y. Hua, Fast maximum likelihood for blind identification of multiple FIR channels, IEEE Trans. Signal Process. 44 (1996) 661–672.

[29] Y. Hua, J. Tugnait, Blind identifiability of FIR-MIMO systems with colored input using second order statistics, IEEE Signal Process. Lett. 7 (2000) 348–350.

[30] A. Hyvarinen, E. Oja, A fast fixed-point algorithm for independent component analysis, Neural Comput. 9 (7) (1997) 1483–1492.

[31] C. Jutten, J. Herault, Blind separation of sources, Part I: an adaptive algorithm based on neuromimetic architecture, Signal Processing 24 (1991) 1–10.

[32] J. Karhunen, P. Pajunen, E. Oja, The nonlinear PCA criterion in blind source separation: relations with other approaches, Neurocomputing 22 (1998) 5–20.

[33] R. Lambert, Multichannel blind deconvolution: FIR matrix algebra and separation of multipath mixtures, Ph.D. Thesis, University of Southern California, 1995.

[34] R.W. Lucky, Techniques for adaptive equalization of digital communication systems, Bell Sys. Tech. J. 45 (1966) 255–286.

[35] E. Moulines, P. Duhamel, J.F. Cardoso, S. Mayrargue, Subspace methods for the blind identification of multichannel FIR filters, IEEE Trans. Signal Process. 43 (1995) 516–525.

[36] M. Murray, J. Rice, Differential Geometry and Statistics, Chapman & Hall, New York, 1993.

[37] Y. Sato, Two extensional applications of the zero-forcing equalization method, IEEE Trans. Commun. COM-23 (1975) 684–687.

[38] E. Serpedin, A. Chevreuil, G. Giannakis, P. Loubaton, Blind channel and carrier frequency offset estimation using periodic modulation precoders, IEEE Trans. Signal Process. 48 (2000) 2389–2405.

[39] O. Shalvi, E. Weinstein, New criteria for blind deconvolution of nonminimum phase systems (channels), IEEE Trans. Inform. Theory 36 (1990) 312–321.

[40] L. Tong, R. Liu, V. Soon, Y. Huang, Indeterminacy and identifiability of blind identification, IEEE Trans. Circuits, Systems 38 (5) (May 1991) 499–509.

[41] L. Tong, S. Perreau, Multichannel blind identification: from subspace to maximum likelihood methods, Proc. IEEE 86 (8) (1998) 1951–1968.

[42] L. Tong, G. Xu, T. Kailath, Blind identification and equalization base on second-order statistics: a time domain approach, IEEE Trans. Inform. Theory 40 (1994) 340–349.

[43] J.R. Treichler, B.G. Agee, A new approach to multipath correction of constant modulus signals, IEEE Trans. Acoust. Speech, Signal Process. ASSP-31 (1983) 349–372.

[44] J. Tugnait, B. Huang, Multistep linear predictors-based blind identification and equalization of multiple-input multiple-output channels, IEEE Trans. Signal Process. 48 (2000) 26–38.

[45] L. Zhang, A. Cichocki, S. Amari, Geometrical structures of FIR manifold and their application to multichannel blind deconvolution, in: Proceeding of the International IEEE Workshop on Neural Networks for Signal Processing (NNSP'99), Madison, Wisconsin, August 23–25 1999, pp. 303–312.

[46] L. Zhang, A. Cichocki, 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, November 16–20 1999, pp. 210–216.

[47] L. Zhang, A. Cichocki, S. Amari, Natural gradient algorithm for blind separation of overdetermined mixture with additive noise, IEEE Signal Process. Lett. 6 (11) (1999) 293–295.