

PAPER

# Approximate Maximum Likelihood Source Separation Using the Natural Gradient\*

Seungjin CHOI<sup>†</sup>, *Nonmember*, Andrzej CICHOCKI<sup>††</sup>, *Member*, Liqing ZHANG<sup>††</sup>, *Nonmember*,  
and Shunichi AMARI<sup>††</sup>, *Member*

**SUMMARY** This paper addresses a maximum likelihood method for source separation in the case of overdetermined mixtures corrupted by additive white Gaussian noise. We consider an approximate likelihood which is based on the Laplace approximation and develop a natural gradient adaptation algorithm to find a local maximum of the corresponding approximate likelihood. We present a detailed mathematical derivation of the algorithm using the Lie group invariance. Useful behavior of the algorithm is verified by numerical experiments.

**key words:** *Independent component analysis, maximum likelihood estimation, natural gradient, source separation, overdetermined mixtures.*

## 1. Introduction

Source separation is a statistical method which aims at recovering unknown sources from their linear instantaneous mixtures without any prior knowledge of the mixing process. It has drawn lots of attractions in signal processing and neural networks since it is a fundamental problem encountered in many practical applications such as speech/image processing, array processing, biomedical signal processing where multiple sensors are involved.

In the context of source separation, it is assumed that an  $m$ -dimensional observation vector  $\mathbf{x}(t) = [x_1(t) \cdots x_m(t)]^T$  is generated by

$$\mathbf{x}(t) = \mathbf{A}\mathbf{s}(t) + \mathbf{v}(t), \quad (1)$$

where  $\mathbf{A} \in \mathbb{R}^{m \times n}$  ( $m \geq n$ ) is called the *mixing matrix*,  $\mathbf{s}(t)$  is the  $n$ -dimensional vector whose elements are called *sources*, and  $\mathbf{v}(t)$  is the additive white Gaussian noise vector that is assumed to be statistically independent of  $\mathbf{s}(t)$ . The task of source separation is to

recover sources from sensor signals  $\mathbf{x}(t)$  without resorting to any prior knowledge except for the assumption of statistical independence of sources. Sources can be recovered blindly by either estimating the mixing matrix  $\mathbf{A}$  or its pseudo-inverse  $\mathbf{A}^\dagger$  (which is usually referred to as the demixing matrix  $\mathbf{W} = \mathbf{A}^\dagger$ ).

Two indeterminacies cannot be resolved in source separation without some prior knowledge. They include scaling and permutation ambiguities. Thus if the estimate of the mixing matrix,  $\hat{\mathbf{A}}$  satisfies  $\hat{\mathbf{A}}^\dagger \mathbf{A} = \mathbf{P}\mathbf{\Lambda}$  where  $\mathbf{P}$  is some permutation matrix,  $\mathbf{\Lambda}$  is some non-singular diagonal matrix, then  $(\hat{\mathbf{A}}, \hat{\mathbf{s}})$  and  $(\mathbf{A}, \mathbf{s})$  are said to be related by a waveform-preserving relation [1]. In zero noise limit,  $\hat{\mathbf{s}} = \hat{\mathbf{A}}^\dagger \mathbf{x}$  gives the exact recovery of source vector except for scaling and permutation ambiguities. In the presence of additive white Gaussian noise,  $\hat{\mathbf{s}} = \hat{\mathbf{A}}^\dagger \mathbf{x}$  is related to the best linear unbiased estimator in the presence of isotropic white Gaussian noise.

A variety of methods for source separation have been developed (see [2] and references therein). Most methods of source separation considered complete ( $m = n$ ) noise-free ( $\mathbf{v}(t) = \mathbf{0}$ ) mixtures. In such a case, maximum likelihood methods were well investigated [3], [4] and natural gradient adaptation algorithms were developed [5].

In this paper we consider the case of overdetermined ( $m > n$ ) noisy mixtures. In the framework of maximum likelihood, we employ the Laplace approximation to make the evaluation of the likelihood function to be mathematically tractable. Then we derive a natural gradient adaptation algorithm which estimates the mixing matrix  $\mathbf{A}$ . The resulting algorithm is referred to as Approximate Maximum Likelihood Source Separation (AMLSS).

Throughout this paper, the following assumptions are made:

- AS1 The mixing matrix  $\mathbf{A}$  has full column rank.
- AS2 Sources are mutually independent non-Gaussian stochastic processes with zero mean.
- AS3 Noise is isotropic Gaussian with zero mean and variance  $\sigma^2$  and is statistically independent of source.

The rest of this paper is organized as follows. In

Manuscript received August 10, 2000.

Manuscript revised February 7, 2001.

<sup>†</sup>The author is with the Department of Computer Science and Engineering, POSTECH, Korea

<sup>††</sup>The authors are with Brain-style Information Systems Research Group, BSI, RIKEN, Japan

\*The portion of this work was presented at IEEE SPAWC'01. This work was supported by Korea Ministry of Science and Technology under an International Cooperative Research Project and Brain Science and Engineering Research Program and by Ministry of Education of Korea for its financial support toward the Electrical and Computer Engineering Division at POSTECH through its BK21 program. The portion of this work was carried out when the first author visited RIKEN.

next section we briefly review the maximum likelihood estimation method for source separation in the case of complete noise-free mixtures. In Section 3, we consider the overdetermined noisy mixtures and propose an appropriate objective function in the framework of maximum likelihood estimation. We also derive an adaptive source separation algorithm using the natural gradient. In Section 4, we present numerical experimental results and compare the proposed method to some existing source separation algorithms. Finally conclusions are drawn in Section 5.

## 2. Complete Noise-free Mixtures

For the case of complete noise-free mixtures, one assumes the data model

$$\mathbf{x}(t) = \mathbf{A}\mathbf{s}(t), \quad (2)$$

where the number of sensors is equal to the number of sources, i.e.,  $m = n$ . A brief review of maximum likelihood source separation in such a case, is given below.

Let us consider a set of  $N$  independent observations,  $\mathcal{X} = \{\mathbf{x}(1), \mathbf{x}(2), \dots, \mathbf{x}(N)\}$ . Source signals  $\{s_i(t)\}$  are assumed to be statistically independent and their probability density functions are denoted by  $\{r_i(\cdot)\}$ . Then the likelihood function is given by

$$p(\mathcal{X}|\mathbf{A}, r) = \prod_{t=1}^N p(\mathbf{x}(t)|\mathbf{A}, r). \quad (3)$$

A single factor in the log-likelihood is given by

$$\begin{aligned} \log p(\mathbf{x}|\mathbf{A}, r) &= -\log |\det \mathbf{A}| + \log r(\mathbf{A}^{-1}\mathbf{x}) \\ &= -\log |\det \mathbf{A}| + \sum_{i=1}^n r_i(\hat{s}_i), \end{aligned} \quad (4)$$

where  $\hat{s}_i = [\mathbf{A}^{-1}\mathbf{x}]_i$ .

Or in terms of the demixing matrix  $\mathbf{W} = \mathbf{A}^{-1}$ , the log-likelihood can be written as

$$\log p(\mathbf{x}|\mathbf{W}, r) = \log |\det \mathbf{W}| + \sum_{i=1}^n r_i(\hat{s}_i). \quad (5)$$

The maximum likelihood estimator for the mixing matrix  $\mathbf{A}$  is

$$\hat{\mathbf{A}} = \max_{\mathbf{A}} \log p(\mathbf{x}|\mathbf{A}, r). \quad (6)$$

Or the maximum likelihood estimator for the demixing matrix  $\mathbf{W}$  is the one that maximize the log-likelihood function (5).

The natural gradient method was shown to be efficient in on-line learning and to find the steepest direction when the parameter space is the Riemannian manifold [6]. The maximum likelihood estimate for the demixing matrix  $\mathbf{W}$  can be found iteratively using the natural gradient algorithm that has the form

$$\begin{aligned} \mathbf{W}(t+1) \\ &= \mathbf{W}(t) + \eta_t \left\{ \mathbf{I} - \varphi(\hat{\mathbf{s}}(t))\hat{\mathbf{s}}^T(t) \right\} \mathbf{W}(t), \end{aligned} \quad (7)$$

where  $\eta_t > 0$  is a learning rate. The element-wise nonlinear function  $\varphi(\cdot)$  is the negative score function whose  $i$ th element is defined by

$$\varphi_i(\hat{s}_i) = -\frac{d \log r_i(\hat{s}_i)}{d \hat{s}_i}. \quad (8)$$

The detailed derivation of the algorithm (7) can be found in [5].

Alternatively the natural gradient algorithm to find the mixing matrix  $\mathbf{A}$  has the form

$$\mathbf{A}(t+1) = \mathbf{A}(t) - \eta_t \mathbf{A}(t) \left\{ \mathbf{I} - \varphi(\hat{\mathbf{s}}(t))\hat{\mathbf{s}}^T(t) \right\}. \quad (9)$$

Both algorithms (7) and (9) performs equally well in the case of complete noise-free mixtures.

### Remarks

- The algorithm (7) is one of widely-used source separation algorithms. The nonlinear information maximization [7] and the mutual information minimization [8] fall on this algorithm. It is referred to as the conventional source separation algorithm in this paper.
- As in most adaptive source separation algorithms, the shapes of nonlinear functions  $\{\varphi_i(\cdot)\}$  depend on the probability distributions of sources that are unknown in advance. Usually the hypothesized density model replaces the true distribution. One interesting aspect of maximum likelihood estimation method for source separation is that a reasonable mismatch between the hypothesized density and the true density does not degrade performance in the task of source separation [4].
- Various methods for the selection of the nonlinear functions  $\{\varphi_i(\cdot)\}$  have been developed [9]–[12].
- Another popular algorithm is the EASI [13] that has the form

$$\begin{aligned} \mathbf{W}(t+1) &= \mathbf{W}(t) + \eta_t \left\{ \mathbf{I} - \hat{\mathbf{s}}(t)\hat{\mathbf{s}}^T(t) \right. \\ &\quad \left. - \varphi(\hat{\mathbf{s}}(t))\hat{\mathbf{s}}^T(t) + \hat{\mathbf{s}}(t)\varphi^T(\hat{\mathbf{s}}(t)) \right\} \mathbf{W}(t). \end{aligned} \quad (10)$$

This algorithm was originally derived using the relative gradient method [13]. Later it was shown that this algorithm could also be derived by the natural gradient in Stiefel manifold [14].

- Zhang *et al.* [15] showed that the algorithm (7) was also valid for overdetermined mixtures.

## 3. Overdetermined Noisy Mixtures

This section describes the main contribution of this paper. We consider an objective function which is the log-likelihood based on the Laplace approximation and

derive an associated natural gradient adaptation algorithm. The same objective function as ours was also considered in [16] for learning overcomplete representation.

We consider the linear data model given in Eq. (1) with  $m > n$ . We assume that the noise vector  $\mathbf{v}(t)$  is isotropic Gaussian with mean 0 and variance  $\sigma^2$ , i.e., the probability density function of  $\mathbf{v}(t)$  is

$$p(\mathbf{v}) = \frac{1}{(2\pi\sigma^2)^{\frac{m}{2}}} \exp\left\{-\frac{1}{2\sigma^2}\mathbf{v}^T\mathbf{v}\right\}. \quad (11)$$

### 3.1 Objective Function

In the task of source separation, both mixing matrix  $\mathbf{A}$  and source vector  $\mathbf{s}(t)$  are unknown in contrast to the conventional parameter estimation problem. As in the case of noise-free mixtures, we treat the sources as nuisance parameters. A single factor of the likelihood function by marginalizing over the nuisance parameter space is given by

$$p(\mathbf{x}|\mathbf{A}) = \int_{\mathbf{s}} p(\mathbf{x}|\mathbf{A}, \mathbf{s})r(\mathbf{s})d\mathbf{s}. \quad (12)$$

In the case of complete noise-free mixtures, the conditional density  $p(\mathbf{x}|\mathbf{A}, \mathbf{s})$  is simply

$$p(\mathbf{x}|\mathbf{A}, \mathbf{s}) = \prod_{i=1}^m \delta\left(x_i - \sum_{j=1}^n a_{ij}s_j\right), \quad (13)$$

where  $a_{ij}$  denotes the  $(i, j)$ -element of  $\mathbf{A}$ . Thus the log-likelihood is simplified as Eq. (5). However in the presence of white Gaussian noise, the conditional density  $p(\mathbf{x}|\mathbf{A}, \mathbf{s})$  is Gaussian. In general the integral in (12) for non-Gaussian prior  $r(\mathbf{s})$  is intractable.

Let us define the energy function  $\mathcal{E}(\mathbf{s})$  as

$$\mathcal{E}(\mathbf{s}) = -\log p(\mathbf{x}|\mathbf{A}, \mathbf{s}) - \log r(\mathbf{s}). \quad (14)$$

With this definition, we can rewrite (12) in the form

$$p(\mathbf{x}|\mathbf{A}) = \int_{\mathbf{s}} \exp\{-\mathcal{E}(\mathbf{s})\} d\mathbf{s}. \quad (15)$$

We assume that  $\mathcal{E}(\mathbf{s})$  has a local quadratic form around a most probable value of  $\mathbf{s}$ , say  $\hat{\mathbf{s}}$ . Then we can use the Laplace approximation [17].

Taylor series expansion of  $\mathcal{E}(\mathbf{s})$  at  $\hat{\mathbf{s}}$  up to second order gives

$$\begin{aligned} \mathcal{E}(\mathbf{s}) &= \mathcal{E}(\hat{\mathbf{s}}) + (\mathbf{s} - \hat{\mathbf{s}})^T \nabla \mathcal{E}(\hat{\mathbf{s}}) \\ &\quad + \frac{1}{2}(\mathbf{s} - \hat{\mathbf{s}})^T \nabla^2 \mathcal{E}(\hat{\mathbf{s}})(\mathbf{s} - \hat{\mathbf{s}}), \end{aligned} \quad (16)$$

where

$$\begin{aligned} \nabla \mathcal{E}(\hat{\mathbf{s}}) &= \left. \frac{\partial \mathcal{E}(\mathbf{s})}{\partial \mathbf{s}} \right|_{\mathbf{s}=\hat{\mathbf{s}}}, \\ \nabla^2 \mathcal{E}(\hat{\mathbf{s}}) &= \left. \frac{\partial}{\partial \mathbf{s}} \left[ (\nabla \mathcal{E}(\mathbf{s}))^T \right] \right|_{\mathbf{s}=\hat{\mathbf{s}}}. \end{aligned} \quad (17)$$

Since the local quadratic approximation is made at the minimum  $\hat{\mathbf{s}}$  of the energy function (which is a most probable value of  $\mathbf{s}$ ), the first-order derivative  $\nabla \mathcal{E}(\hat{\mathbf{s}})$  in (16) is equal to zero.

Define  $\Delta \mathbf{s} = \mathbf{s} - \hat{\mathbf{s}}$ , then we have

$$\begin{aligned} p(\mathbf{x}|\mathbf{A}) &= \exp\{-\mathcal{E}(\hat{\mathbf{s}})\} \int_{\mathbf{s}} \exp\left\{-\frac{1}{2}\Delta \mathbf{s}^T \nabla^2 \mathcal{E}(\hat{\mathbf{s}})\Delta \mathbf{s}\right\} d\mathbf{s} \\ &= \exp\{-\mathcal{E}(\hat{\mathbf{s}})\} (2\pi)^{\frac{n}{2}} [\det(\nabla^2 \mathcal{E}(\hat{\mathbf{s}}))]^{-\frac{1}{2}}. \end{aligned} \quad (18)$$

Note that

$$p(\mathbf{x}|\mathbf{A}, \hat{\mathbf{s}}) = \frac{1}{(2\pi\sigma^2)^{\frac{m}{2}}} \exp\left\{-\frac{1}{2\sigma^2}\|\mathbf{x} - \mathbf{A}\hat{\mathbf{s}}\|^2\right\} \quad (19)$$

Then, the log-likelihood  $L$  is

$$\begin{aligned} L &= \log p(\mathbf{x}|\mathbf{A}) \\ &= -\mathcal{E}(\hat{\mathbf{s}}) + \frac{n}{2} \log 2\pi - \frac{1}{2} \log \det(\nabla^2 \mathcal{E}(\hat{\mathbf{s}})) \\ &= \log p(\mathbf{x}|\mathbf{A}, \hat{\mathbf{s}}) + \log r(\hat{\mathbf{s}}) + \frac{n}{2} \log 2\pi \\ &\quad - \frac{1}{2} \log \det(\nabla^2 \mathcal{E}(\hat{\mathbf{s}})) \\ &= -\frac{m}{2} \log 2\pi\sigma^2 - \frac{1}{2\sigma^2} \|\mathbf{x} - \mathbf{A}\hat{\mathbf{s}}\|^2 + \frac{n}{2} \log 2\pi \\ &\quad + \log r(\hat{\mathbf{s}}) - \frac{1}{2} \log \det(\nabla^2 \mathcal{E}(\hat{\mathbf{s}})). \end{aligned} \quad (20)$$

One can easily see that

$$\nabla^2 \mathcal{E}(\hat{\mathbf{s}}) = \frac{\mathbf{A}^T \mathbf{A}}{\sigma^2} - \nabla^2 \log r(\hat{\mathbf{s}}). \quad (21)$$

### 3.2 Natural Gradient Adaptation Algorithm

The log-likelihood based on the Laplace approximation is described in (20). Here we derive an adaptation algorithm for estimating  $\mathbf{A}$  that maximizes the log-likelihood (20).

The natural gradient was shown to be efficient in on-line learning and to find the steepest direction when the parameter space is Riemannian manifold [6]. The natural gradient learning was successfully applied to the task of blind source separation [8], [12], [18]–[20] and multichannel blind deconvolution [21]–[23]. However, all these methods were restricted to the case of complete noise-free mixtures.

Here we derive a natural gradient algorithm for estimating the mixing matrix  $\mathbf{A}$ . Some of the result here was motivated by Zhang *et. al* [15] where it was shown that even in the case of overdetermined mixtures, the natural gradient algorithm has the same form as given in (7) that was derived in the case of complete mixtures.

We define the manifold of the mixing matrices as  $Gl(m, n) = \{\mathbf{A} \in \mathbb{R}^{m \times n} | \text{rank}(\mathbf{A}) = n\}$ . For  $\mathbf{A} \in$

$Gl(m, n)$ , there exists an orthogonal matrix  $\mathbf{Q} \in \mathbb{R}^{m \times m}$  such that

$$\mathbf{A} = \mathbf{Q}^T \begin{bmatrix} \mathbf{A}_1 \\ \mathbf{A}_2 \end{bmatrix}, \quad (22)$$

where  $\mathbf{A}_1 \in \mathbb{R}^{n \times n}$  and  $\mathbf{A}_2 \in \mathbb{R}^{(m-n) \times n}$ .

The Lie group structure of the parameter space is a key ingredient in the calculation of the natural gradient [6]. In similar manner as in [15], we introduce two operations on the manifold  $Gl(m, n)$ ,

$$\mathbf{X} \circ \mathbf{Y} = \mathbf{Q}^T \begin{bmatrix} \mathbf{X}_1 \mathbf{Y}_1 \\ \mathbf{X}_2 \mathbf{Y}_1 + \mathbf{Y}_2 \end{bmatrix}, \quad (23)$$

and

$$\mathbf{X}^\# = \mathbf{Q}^T \begin{bmatrix} \mathbf{X}_1^{-1} \\ -\mathbf{X}_2 \mathbf{X}_1 \end{bmatrix}, \quad (24)$$

where

$$\mathbf{X} = \mathbf{Q}^T \begin{bmatrix} \mathbf{X}_1 \\ \mathbf{X}_2 \end{bmatrix}, \quad \mathbf{Y} = \mathbf{Q}^T \begin{bmatrix} \mathbf{Y}_1 \\ \mathbf{Y}_2 \end{bmatrix}. \quad (25)$$

The operator  $\circ$  denotes the multiplication and  $\#$  represents the inverse on the manifold  $Gl(m, n)$ .

The identity on  $Gl(m, n)$  is defined as

$$\mathbf{E} = \mathbf{Q}^T \begin{bmatrix} \mathbf{I}_n \\ \mathbf{0} \end{bmatrix}, \quad (26)$$

where  $\mathbf{I}_n$  denotes the  $n \times n$  identity matrix. It is easy to see that the manifold with these operations forms a Lie group.

One important property in the Lie group is the invariance of Riemannian metric. Let us define the tangent space of  $Gl(m, n)$  by  $T_{\mathbf{A}}$  and the tangent vectors by  $\mathbf{X}, \mathbf{Y} \in T_{\mathbf{A}}$ . We define the inner product between two tangent vectors  $\mathbf{X}$  and  $\mathbf{Y}$  at  $\mathbf{A}$  by  $\langle \mathbf{X}, \mathbf{Y} \rangle_{\mathbf{A}}$ . The Lie group invariance ensures

$$\langle \mathbf{X}, \mathbf{Y} \rangle_{\mathbf{A}} = \langle \mathbf{Z} \circ \mathbf{X}, \mathbf{Z} \circ \mathbf{Y} \rangle_{\mathbf{Z} \circ \mathbf{A}}, \quad (27)$$

for any left multiplication transformation  $\mathbf{Z} \in Gl(m, n)$  that is an onto mapping.

Now we calculate the natural gradient of the log-likelihood (20). The conventional gradient  $\nabla \log p(\mathbf{x}|\mathbf{A})$  (see Appendix for detailed derivation) is

$$\begin{aligned} \nabla \log p(\mathbf{x}|\mathbf{A}) &= -\mathbf{A} \left( \mathbf{A}^T \mathbf{A} \right)^{-1} \left\{ \mathbf{I} - \varphi(\hat{\mathbf{s}}) \hat{\mathbf{s}}^T \right\} \\ &\quad + \frac{1}{\sigma^2} \mathbf{P}_A^\perp \mathbf{x} \hat{\mathbf{s}}^T. \end{aligned} \quad (28)$$

We denote the natural gradient of the log-likelihood (20) by  $\tilde{\nabla} L$ .

The natural gradient  $\tilde{\nabla} L$  is defined by [6]

$$\left\langle \mathbf{X}, \tilde{\nabla} L \right\rangle_{\mathbf{A}} = \left\langle \mathbf{A}^\# \circ \mathbf{X}, \mathbf{A}^\# \circ \tilde{\nabla} L \right\rangle_{\mathbf{E}}. \quad (29)$$

Comparing both sides of (29), we have

$$\tilde{\nabla} L = \left\{ \mathbf{A} \mathbf{A}^T + \mathbf{N}_I \right\} \nabla L, \quad (30)$$

where

$$\mathbf{N}_I = \mathbf{Q}^T \begin{bmatrix} \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{I}_{m-n} \end{bmatrix} \mathbf{Q}. \quad (31)$$

Therefore the updating rule for  $\hat{\mathbf{A}}$  is given by

$$\begin{aligned} \Delta \mathbf{A} &= -\eta_t \left[ \mathbf{A} \mathbf{A}^T + \mathbf{N}_I \right] \\ &\quad \left[ \mathbf{A} \left( \mathbf{A}^T \mathbf{A} \right)^{-1} \left\{ \mathbf{I} - \varphi(\hat{\mathbf{s}}) \hat{\mathbf{s}}^T \right\} - \frac{1}{\sigma^2} \mathbf{P}_A^\perp \mathbf{x} \hat{\mathbf{s}}^T \right] \\ &= -\eta_t \mathbf{A} \left\{ \mathbf{I} - \varphi(\hat{\mathbf{s}}) \hat{\mathbf{s}}^T \right\} + \mathbf{C}, \end{aligned} \quad (32)$$

where

$$\begin{aligned} \mathbf{C} &= -\eta_t \mathbf{N}_I \left[ \mathbf{A} \left( \mathbf{A}^T \mathbf{A} \right)^{-1} \left\{ \mathbf{I} - \varphi(\hat{\mathbf{s}}) \hat{\mathbf{s}}^T \right\} \right. \\ &\quad \left. - \frac{1}{\sigma^2} \mathbf{P}_A^\perp \mathbf{x} \hat{\mathbf{s}}^T \right]. \end{aligned} \quad (33)$$

In [15], they explained which projection matrix  $\mathbf{N}_I$  was the best in the sense of minimizing the effect of additive noise. Although they considered a different cost function (based on mutual information minimization) and derived the algorithm for updating the demixing matrix  $\mathbf{W}$ , we can make a similar argument here. It was shown in [15] that the optimal projection  $\mathbf{N}_I$  is chosen in such a way that the matrix  $\mathbf{C}$  is vanished. In such a case, the updating rule is simplified as

$$\Delta \mathbf{A} = -\eta_t \mathbf{A} \left\{ \mathbf{I} - \varphi(\hat{\mathbf{s}}) \hat{\mathbf{s}}^T \right\}. \quad (34)$$

#### Algorithm Outline: AMLSS

- Given the current estimate of the mixing matrix,  $\hat{\mathbf{A}}(t)$ , we infer the source vector by

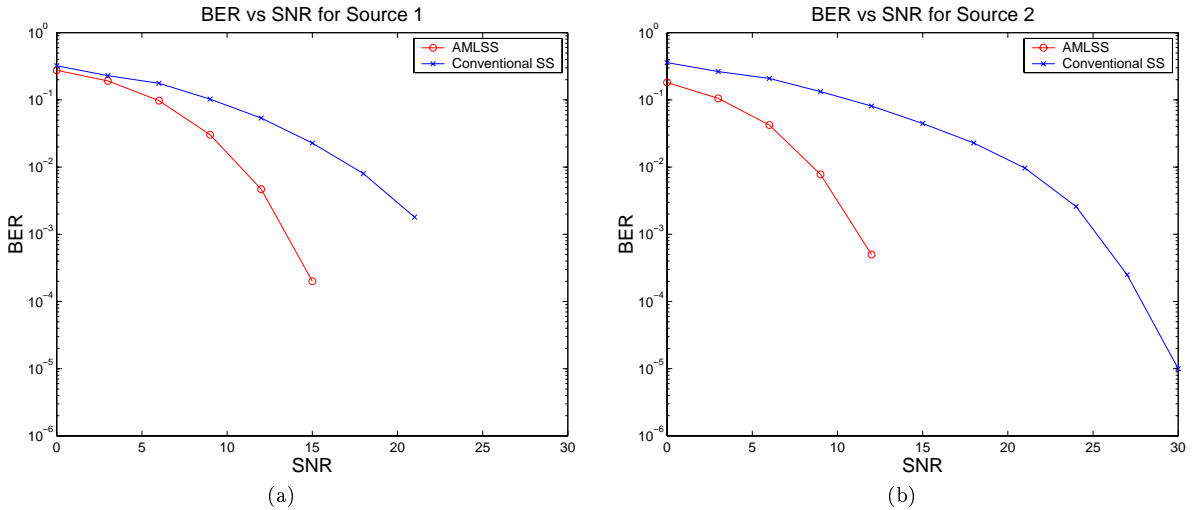
$$\hat{\mathbf{s}}(t) = \left( \hat{\mathbf{A}}^T(t) \hat{\mathbf{A}}(t) \right)^{-1} \hat{\mathbf{A}}^T(t) \mathbf{x}(t). \quad (35)$$

- Using  $\hat{\mathbf{s}}(t)$  and  $\hat{\mathbf{A}}(t)$ , we find the new estimate of the mixing matrix,  $\hat{\mathbf{A}}(t+1)$  by the algorithm (34) that can be written as

$$\begin{aligned} \hat{\mathbf{A}}(t+1) &= \hat{\mathbf{A}}(t) - \eta_t \hat{\mathbf{A}}(t) \left\{ \mathbf{I} - \varphi(\hat{\mathbf{s}}(t)) \hat{\mathbf{s}}^T(t) \right\}. \end{aligned} \quad (36)$$

- These two steps are repeated until  $\hat{\mathbf{A}}$  converges.

**Remark:** The AMLSS algorithm is very similar to the algorithm proposed by Lewicki and Sejnowski [16]. We consider the case of overdetermined (undercomplete) mixture, whereas underdetermined (overcomplete) mixture was considered in [16]. That is why we can use a simple least squares projection for inference. Moreover we present a rigorous derivation of the algorithm using the Lie group invariance which is not found in [16].



**Fig. 1** Bit error rate (BER) with respect to SNR in Experiment 1: (a) for source 1; (b) source 2.

## 4. Numerical Experiments

We demonstrate the useful behavior of our method, AMLSS, that is summarized in (35) and (36). We provide two simulation results, both of which consider the case of overdetermined noisy mixtures. The AMLSS is compared with the conventional source separation algorithm in (7).

### 4.1 Experiment 1

In this experiment, we use two binary sources and the mixing matrix  $\mathbf{A} \in \mathbb{R}^{3 \times 2}$  whose elements are drawn from standardized Gaussian distribution. We consider two source separation algorithms: (1) AMLSS; (2) conventional source separation in (7). In order to evaluate performance, we performed 10 different runs at each SNR (that varies from 0 dB to 30 dB) and calculated the average of bit error rate (BER).

Since the recovered source vector  $\hat{\mathbf{s}}$  contains the scaling and permutation ambiguities, the BER is calculated after removing these indeterminacies. For both algorithms (AMLSS and the conventional method), randomly-chosen initial value was assigned to  $\mathbf{A}(0)$  or  $\mathbf{W}(0)$ . The learning rate  $\eta_t = .001$  was used. We used the nonlinear function  $\varphi_i(\hat{s}_i) = |\hat{s}_i|^2 \hat{s}_i$  that is known to be effective when sources are sub-Gaussian [13].

The experimental result in terms of BER is shown in Fig. 1. The AMLSS outperforms the conventional method in both high and low SNR environments. In the environment where SNR is greater than 15dB, the AMLSS method produces 0 BER, whereas the conventional method resulted in a considerable amount of BER. This numerical experiment shows the high performance of our method when mixtures are noisy.

### 4.2 Experiment 2

The second numerical experiment was carried out for blind co-channel signal separation using an antenna array. We assume a uniform linear 4-element ( $m = 4$ ) antenna array with each element being half wavelength spaced. We consider  $n = 2$  digitally modulated QPSK sources with angles of arrival,  $0^\circ$  and  $20^\circ$ .

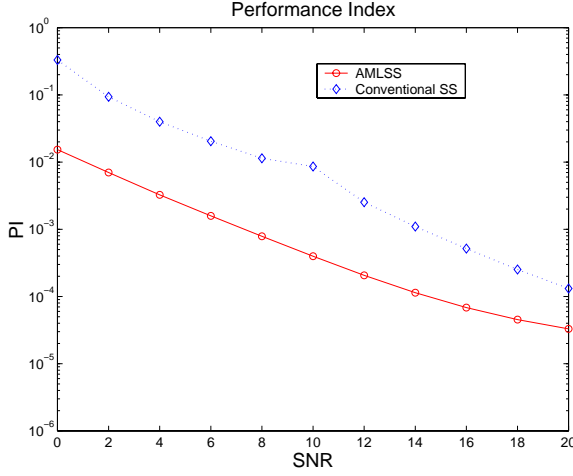
For performance evaluation, we use the performance index (PI) that is defined by

$$\text{PI} = \frac{1}{n(n-1)} \sum_{i=1}^n \left\{ \left( \sum_{k=1}^n \frac{|g_{ik}|}{\max_j |g_{ij}|} - 1 \right) + \left( \sum_{k=1}^n \frac{|g_{ki}|}{\max_j |g_{ji}|} - 1 \right) \right\}, \quad (37)$$

where  $g_{ij}$  is the  $(i, j)$ -element of the global system matrix  $\mathbf{G} = \hat{\mathbf{A}}\mathbf{A} = \mathbf{W}\mathbf{A}$ .

We evaluated the performance of AMLSS and the conventional method (7) in terms of PI with SNR varying from 0 dB to 20 dB (see Fig. 2). Since sources are complex, the transpose operator is replaced by the Hermitian in both algorithms. Randomly-chosen initial value was assigned to  $\mathbf{A}(0)$  or  $\mathbf{W}(0)$ . The learning rate  $\eta_t = .01$  was used. Once again, the cubic nonlinear function  $\varphi_i(\hat{s}_i) = |\hat{s}_i|^2 \hat{s}_i$  was used for both algorithms. At each SNR, we performed 100 different runs and evaluated the average of the performance index.

Fig. 2 shows that the AMLSS outperforms the conventional source separation method in whole range of SNR. This results from the objective function 20 which takes the additive white Gaussian noise into account and our new natural gradient adaptation algorithm.



**Fig. 2** Performance of AMLSS and the conventional source separation method (7) in experiment 2.

## 5. Conclusions

In this paper, we have presented a maximum likelihood method for source separation for the case of overdetermined noisy mixtures. In the framework of maximum likelihood estimation, we have employed the Laplace approximation to arrive at a mathematical tractable objective function. We have developed a natural gradient algorithm to find a local maximum of the approximate log-likelihood. Our algorithm, AMLSS, was presented and its performance was compared with the conventional method. Numerical experiments confirmed the useful behavior and high performance of the AMLSS in the case of overdetermined noisy mixtures.

## Appendix

Here we present the detailed calculation of the gradient of the log-likelihood function (20) with respect to  $\mathbf{A}$ . We assume that the most probable value of  $\mathbf{s}$  is inferred by  $\hat{\mathbf{s}} = (\mathbf{A}^T \mathbf{A})^{-1} \mathbf{A}^T \mathbf{x} = \mathbf{A}^\dagger \mathbf{x}$ .

Derivation of  $\frac{\partial}{\partial \mathbf{A}} [\|\mathbf{x} - \mathbf{A}\hat{\mathbf{s}}\|^2]$

We calculate the infinitesimal increment of  $\|\mathbf{x} - \mathbf{A}\hat{\mathbf{s}}\|^2$ ,

$$\begin{aligned} d\{\|\mathbf{x} - \mathbf{A}\hat{\mathbf{s}}\|^2\} &= d\left\{\mathbf{x}^T \mathbf{x} - 2\hat{\mathbf{s}}^T \mathbf{A}^T \mathbf{x} + \hat{\mathbf{s}}^T \mathbf{A}^T \mathbf{A} \hat{\mathbf{s}}\right\} \\ &= \hat{\mathbf{s}}^T d\mathbf{A}^T \mathbf{A} \hat{\mathbf{s}} + \hat{\mathbf{s}}^T \mathbf{A}^T d\mathbf{A} \hat{\mathbf{s}} - 2\hat{\mathbf{s}}^T d\mathbf{A}^T \mathbf{x}. \end{aligned} \quad (38)$$

Thus we have

$$\begin{aligned} \frac{\partial}{\partial \mathbf{A}} [\|\mathbf{x} - \mathbf{A}\hat{\mathbf{s}}\|^2] &= -2(\mathbf{x} - \mathbf{A}\hat{\mathbf{s}}) \hat{\mathbf{s}}^T \\ &= -2\mathbf{P}_A^\perp \mathbf{x} \hat{\mathbf{s}}^T, \end{aligned} \quad (39)$$

where  $\mathbf{P}_A^\perp = \mathbf{I} - \mathbf{A}\mathbf{A}^\dagger$  is the orthogonal projection matrix

Derivation of  $\frac{\partial}{\partial \mathbf{A}} [\log r(\hat{\mathbf{s}})]$

We define

$$\varphi_i(\hat{s}_i) = -\frac{d \log r_i(\hat{s}_i)}{d \hat{s}_i}, \quad (40)$$

and

$$\varphi(\hat{\mathbf{s}}) = [\varphi_1(\hat{s}_1), \dots, \varphi_n(\hat{s}_n)]^T. \quad (41)$$

With this definition, the infinitesimal increment of  $\log r(\hat{\mathbf{s}})$  is

$$\begin{aligned} d\{\log r(\hat{\mathbf{s}})\} &= d\left\{\sum_{i=1}^n \log r_i(\hat{s}_i)\right\} \\ &= -\varphi^T(\hat{\mathbf{s}}) d\hat{\mathbf{s}}. \end{aligned} \quad (42)$$

Note that

$$\begin{aligned} d\hat{\mathbf{s}} &= d\left[\left(\mathbf{A}^T \mathbf{A}\right)^{-1}\right] \mathbf{A}^T \mathbf{x} + \left(\mathbf{A}^T \mathbf{A}\right) d\mathbf{A}^T \mathbf{x} \\ &= -\left(\mathbf{A}^T \mathbf{A}\right)^{-1} \mathbf{A}^T d\mathbf{A} \hat{\mathbf{s}}. \end{aligned} \quad (43)$$

Hence we have

$$d\{\log r(\hat{\mathbf{s}})\} = -\varphi^T(\hat{\mathbf{s}}) \left(\mathbf{A}^T \mathbf{A}\right)^{-1} \mathbf{A}^T d\mathbf{A} \hat{\mathbf{s}}, \quad (44)$$

which gives

$$\frac{\partial}{\partial \mathbf{A}} [\log r(\hat{\mathbf{s}})] = \mathbf{A} \left(\mathbf{A}^T \mathbf{A}\right)^{-1} \varphi(\hat{\mathbf{s}}) \hat{\mathbf{s}}^T. \quad (45)$$

Derivation of  $\frac{\partial}{\partial \mathbf{A}} [\log \det \nabla^2 \mathcal{E}(\hat{\mathbf{s}})]$

We assume that the variance of noise,  $\sigma^2$  is small, i.e.,

$$\nabla^2 \mathcal{E}(\hat{\mathbf{s}}) \approx \frac{\mathbf{A}^T \mathbf{A}}{\sigma^2}. \quad (46)$$

Then,

$$\begin{aligned} d\{\log \det \nabla^2 \mathcal{E}(\hat{\mathbf{s}})\} &= \text{tr} \left\{ \left[\nabla^2 \mathcal{E}(\hat{\mathbf{s}})\right]^{-1} d\nabla^2 \mathcal{E}(\hat{\mathbf{s}}) \right\}. \end{aligned} \quad (47)$$

Note that

$$\begin{aligned} d\nabla^2 \mathcal{E}(\hat{\mathbf{s}}) &= \text{tr} \left\{ \left(\mathbf{A}^T \mathbf{A}\right)^{-1} d\mathbf{A}^T \mathbf{A} \right. \\ &\quad \left. + \left(\mathbf{A}^T \mathbf{A}\right)^{-1} \mathbf{A}^T d\mathbf{A} \right\}. \end{aligned} \quad (48)$$

Hence we have

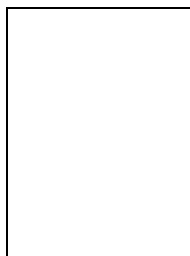
$$\frac{\partial}{\partial \mathbf{A}} [\log \det \nabla^2 \mathcal{E}(\hat{\mathbf{s}})] = 2\mathbf{A} \left(\mathbf{A}^T \mathbf{A}\right)^{-1}. \quad (49)$$

Combining the results in (39), (45), and (49), the gradient of the cost function (20) is given by

$$\begin{aligned} \nabla \log p(\mathbf{x}|\mathbf{A}) &= \frac{\partial}{\partial \mathbf{A}} [\log p(\mathbf{x}|\mathbf{A})] \\ &= -\mathbf{A} \left(\mathbf{A}^T \mathbf{A}\right)^{-1} \left\{ \mathbf{I} - \varphi(\hat{\mathbf{s}}) \hat{\mathbf{s}}^T \right\}. \end{aligned} \quad (50)$$

## References

- [1] L. Tong, Y. Inouye, and R. Liu, "Waveform-preserving blind estimation of multiple independent sources," *IEEE Trans. Signal Processing*, vol. 41, pp. 2461–2470, July 1993.
- [2] S. Haykin, *Unsupervised Adaptive Filtering: Blind Source Separation*. Prentice-Hall, 2000.
- [3] D. T. Pham, "Blind separation of instantaneous mixtures of sources via an independent component analysis," *IEEE Trans. Signal Processing*, vol. 44, no. 11, pp. 2768–2779, 1996.
- [4] J. F. Cardoso, "Infomax and maximum likelihood for source separation," *IEEE Signal Processing Letters*, vol. 4, pp. 112–114, Apr. 1997.
- [5] S. Amari and A. Cichocki, "Adaptive blind signal processing - neural network approaches," *Proc. of the IEEE, Special Issue on Blind Identification and Estimation*, vol. 86, no. 10, pp. 2026–2048, Oct. 1998.
- [6] S. Amari, "Natural gradient works efficiently in learning," *Neural Computation*, vol. 10, no. 2, pp. 251–276, 1998.
- [7] A. Bell and T. Sejnowski, "An information maximisation approach to blind separation and blind deconvolution," *Neural Computation*, vol. 7, pp. 1129–1159, 1995.
- [8] S. Amari, A. Cichocki, and H. H. Yang, "A new learning algorithm for blind signal separation," in *Advances in Neural Information Processing Systems* (D. S. Touretzky, M. C. Mozer, and M. E. Hasselmo, eds.), vol. 8, pp. 757–763, MIT press, 1996.
- [9] A. Cichocki, I. Sabala, S. Choi, B. Orsier, and R. Szupiluk, "Self-adaptive independent component analysis for sub-Gaussian and super-Gaussian mixtures with unknown number of source signals," in *Proc. Int. Symp. Nonlinear Theory and Applications*, pp. 731–734, 1997.
- [10] M. Girolami, "An alternative perspective on adaptive independent component analysis algorithms," *Neural Computation*, vol. 10, no. 8, pp. 2103–2114, Nov. 1998.
- [11] T. W. Lee, M. Girolami, and T. Sejnowski, "Independent component analysis using an extended infomax algorithm for mixed sub-Gaussian and super-Gaussian sources," *Neural Computation*, vol. 11, no. 2, pp. 609–633, 1999.
- [12] S. Choi, A. Cichocki, and S. Amari, "Flexible independent component analysis," *Journal of VLSI Signal Processing*, vol. 26, pp. 25–38, Aug. 2000.
- [13] J. F. Cardoso and B. H. Laheld, "Equivariant adaptive source separation," *IEEE Trans. Signal Processing*, vol. 44, no. 12, pp. 3017–3030, Dec. 1996.
- [14] S. Amari, "Natural gradient for over- and under-complete bases in ICA," *Neural Computation*, vol. 11, no. 8, pp. 1875–1883, 1999.
- [15] L. Q. Zhang, A. Cichocki, and S. Amari, "Natural gradient algorithm for blind separation of overdetermined mixtures with additive noise," *IEEE Signal Processing Letters*, vol. 6, pp. 293–295, Nov. 1999.
- [16] M. S. Lewicki and T. Sejnowski, "Learning overcomplete representation," *Neural Computation*, vol. 12, no. 2, pp. 337–365, 2000.
- [17] J. J. K. O. Ruanaidh and W. J. Fitzgerald, *Numerical Bayesian Methods Applied to Signal Processing*. Springer, 1996.
- [18] S. Amari, T. P. Chen, and A. Cichocki, "Stability analysis of learning algorithms for blind source separation," *Neural Networks*, vol. 10, no. 8, pp. 1345–1351, 1997.
- [19] S. Choi, A. Cichocki, and S. Amari, "Local stability analysis of flexible independent component analysis algorithm," in *Proc. ICASSP*, (Istanbul, Turkey), pp. 3426–3429, 2000.
- [20] S. Amari, T. P. Chen, and A. Cichocki, "Nonholonomic orthogonal learning algorithms for blind source separation," *Neural Computation*, vol. 12, no. 6, pp. 1463–1484, 2000.
- [21] S. Amari, S. C. Douglas, A. Cichocki, and H. H. Yang, "Multichannel blind deconvolution and equalization using the natural gradient," in *Proc. SPAWC*, (Paris, France), pp. 101–104, 1997.
- [22] S. Choi, S. Amari, A. Cichocki, and R. Liu, "Natural gradient learning with a nonholonomic constraint for blind deconvolution of multiple channels," in *Proc. ICA'99*, (Aussais, France), pp. 371–376, 1999.
- [23] S. Choi, A. Cichocki, and S. Amari, "Two spatio-temporal decorrelation algorithms and their application to blind deconvolution of multiple channels," in *Proc. ICASSP*, (Phoenix, Arizona), pp. 1085–1088, 1999.

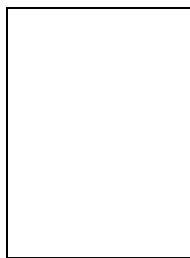


**Seungjin Choi** was born in Seoul, Korea, in 1964. He received the B.S. and M.S. degrees in electrical engineering from Seoul National University, Korea, in 1987 and 1989, respectively and the Ph.D degree in electrical engineering from the University of Notre Dame, Indiana, in 1996.

He was a Visiting Assistant Professor in the Department of Electrical Engineering at University of Notre Dame, Indiana during

the Fall semester of 1996. He was with the Laboratory for Artificial Brain Systems, RIKEN, Japan in 1997 and was an Assistant Professor in the School of Electrical and Electronics Engineering, Chungbuk National University from 1997 to 2000. He is currently Assistant Professor of Computer Science and Engineering at Pohang University of Science and Technology, and is affiliated with the Intelligent Multimedia Lab as well as the Brain Research Center. He has also been an adjunct senior researcher at Laboratory for Advanced Brain Signal Processing, BSI, Japan since 1998. His primary research interests include probabilistic/statistical learning, statistical (blind) signal processing, independent component analysis, data analysis, pattern recognition, and multiuser communications.

Dr. Choi is a Technical Committee member of IEEE Neural Networks for Signal Processing (NNSP)



**Andrzej Cichocki** was born in Poland on August 1947. He received the M.Sc.(with honors), Ph.D., and Habilitate Doctorate (Dr.Sc.) degrees, all in electrical engineering and computer science, from Warsaw University of Technology (Poland) in 1972, 1975, and 1982, respectively.

Since 1972, he has been with the Institute of Theory of Electrical Engineering and Electrical Measurements at the War-

saw University of Technology, where he became a full Professor in 1995.

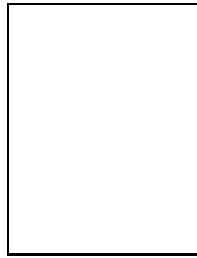
He is the co-author of two international books: *MOS Switched-Capacitor and Continuous-Time Integrated Circuits and Systems* (Springer-Verlag, 1989) and *Neural Networks for Optimization*

and Signal Processing (J Wiley and Teubner Verlag,1993/94) and author or co-author of more than hundred fifty (150) scientific papers.

He spent at University Erlangen-Nuernberg (GERMANY) a few years as Alexander Humboldt Research Fellow and Guest Professor. In 1995-96 he has been working as a Team Leader of the Laboratory for Artificial Brain Systems, at the Frontier Research Program RIKEN (JAPAN), in the Brain Information Processing Group directed by professor Shun-ichi Amari. Currently he is head of the laboratory for Open Information Systems in the Brain Science Institute, Riken, Wako-schi, JAPAN.

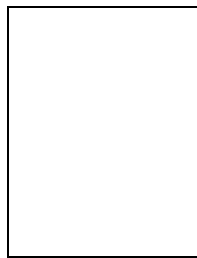
He is reviewer of several international Journals, e.g. IEEE Trans. on Neural Networks, Signal Processing, Circuits and Systems, Biological Cybernetics, Electronics Letters, Neurocomputing, Neural Computation. He is also member of several international Scientific Committees and the associated Editor of IEEE Transaction on Neural Networks (since January 1998). His current research interests include signal and image processing (especially blind signal/image processing), neural networks and their electronic implementations, learning theory and algorithms, independent and principal component analysis, optimization problems, circuits and systems theory and their applications, artificial intelligence.

He was founding Coeditor-in-Chief of Neural Networks. He has been awarded the Japan Academy Award, the IEEE Neural Networks Pioneer Award, and the IEEE Emanuel R. Piore Award.



**Liqing Zhang** received the B.S. degree in Mathematics from Hangzhou University and the Ph.D. degree in Computer Science from Zhongshan University, China, in 1983 and 1988, respectively. In 1988, he joined the Department of Automation at South China University of Technology, as a lecturer. From 1990 to 1995, he was an Associate Professor, in 1995 he was promoted to a full Professor in the College of Electronic and Informa-

tion Engineering at South China University of Technology. He is currently working in the Laboratory for Advanced Brain Signal Processing, RIKEN Brain Science Institute as a researcher. His research interests include blind signal processing and neural networks.



**Shun-ichi Amari** was born in Tokyo, Japan, on January 3, 1936. He graduated from the University of Tokyo in 1958, having majored in mathematical engineering, and he received the Dr.Eng. degree from the University of Tokyo in 1963.

He was an Associate Professor at Kyushu University, an Associate and then Full Professor at the Department of Mathematical Engineering and Information Physics, University of Tokyo, and is now

Professor-Emeritus at the University of Tokyo. He is the Director of the Brain-Style Information Systems Group, RIKEN Brain Science Institute, Saitama Japan. He has been engaged in research in wide areas of mathematical engineering and applied mathematics, such as topological network theory, differential geometry of continuum mechanics, pattern recognition, mathematical foundations of neural networks, and information geometry. Dr. Amari served as President of the International Neural Network Society, Council member of Bernoulli Society for Mathematical Statistics and Probability Theory, and Vice President of the Institute of Electrical, Information and Communication En-