Blind Source Separation—Semiparametric Statistical Approach

Shun-ichi Amari, *Fellow, IEEE,* and Jean-François Cardoso, *Member, IEEE*

***Abstract*— The semiparametric statistical model is used to formulate the problem of blind source separation. The method of estimating functions is applied to this problem. It is shown that an estimator of the mixing matrix or its learning version can be described in terms of an estimating function. The statistical efficiencies of these algorithms are studied. The main results are as follows.**

1) **The space consisting of all the estimating functions is derived.**
2) **The space is decomposed into the orthogonal sum of the admissible part and redundant ancillary part. For any estimating function, one can find a better or equally good estimator in the admissible part.**
3) **The Fisher efficient (that is, asymptotically best) estimating functions are derived.**
4) **The stability of learning algorithms is studied.**

***Index Terms*— Blind source separation, estimating function, independent component analysis, learning algorithm, semiparametric statistical model.**

I. Introduction

SINCE the proposal of Jutten and Hérault [14], blind source separation is one of the most active areas connecting signal processing and neural networks. Although a lot of new ideas and heuristic algorithms have so far been proposed [6], [10], [12], [16], [17], [21], yet there still remain fundamental problems such as statistical efficiency and convergence of learning algorithms, which must be be studied further.

The present paper elucidates these problems from the statistical point of view. Blind source separation is formulated in the framework of semiparametric statistical models (Bickel *et al.* [11]). We apply the estimating function theory of semiparametric statistics (Amari and Kawanabe [8], [9]) and analyze the efficiency of an estimator by obtaining the error covariance matrix. We then obtain the estimating function that gives asymptotically the best estimator, that is, the Fisher efficient estimator. We also discuss the convergence property of learning algorithms.

Let us formulate the problem. Let $\boldsymbol{s} = (s_1, s_2, \cdots, s_m)^T$ be a vector of m source signals whose components are stochastically independent. Let

$$\boldsymbol{x} = A\boldsymbol{s}$$

be an observed mixed signal vector, where we assume A is an unknown $m \times m$ invertible matrix, and the probability distribution $r(\boldsymbol{s})$ of $\boldsymbol{s}$ is unknown except that the m source signals are independent. Let $\{\boldsymbol{x}(1), \boldsymbol{x}(2), \cdots, \boldsymbol{x}(n)\}$ be a sequence of observed signals. Our task is to obtain a good estimate W of A^{-1} and recover the original signals by

$$\boldsymbol{y}(t) = W\boldsymbol{x}(t).$$

As will be shown later, there are indeterminacies in this problem, and A^{-1} itself cannot be identified. What we can do is to estimate a rescaled and permuted version of the m source signals.

Most of the algorithms proposed so far (Amari *et al.* [6]; Bell and Sejnowski [10]; Common [14]; Jutten and Hérault [16]) heuristically find a matrix $F(\boldsymbol{x}, W)$, which is often of the form $F(\boldsymbol{x}, W) = G(\boldsymbol{y})$ for some matrix G with $\boldsymbol{y} = W\boldsymbol{x}$, that satisfies

$$E[F(\boldsymbol{x}, W)] = 0. \tag{1.1}$$

Here, the expectation E is taken with respect to $\boldsymbol{x} = A\boldsymbol{s}$, where $W = A^{-1}$, and (1.1) is required to hold whatever source probability distribution $r(\boldsymbol{s})$ is used, as long as s_i's are independent. Such a function is called an estimating function.

Given n observations $\boldsymbol{x}(1), \cdots, \boldsymbol{x}(n)$, an estimator $\hat{W}$ is obtained by solving the estimating equation

$$\sum_{i=1}^{n} F\{\boldsymbol{x}(i), W\} = 0 \tag{1.2}$$

which is obtained by replacing the expectation (1.1) by the arithmetic mean over the observed data. The related learning rule is given by

$$W_{t+1} = W_t - \eta_t F\{\boldsymbol{x}(t), W_t\}. \tag{1.3}$$

It is not a trivial task to obtain an estimating function satisfying (1.1) because this should hold for any independent $r(\boldsymbol{s})$. The present paper solves the following fundamental problems based on the results of Amari and Kawanabe [8], [9]. The main results are as follows.

1) We derive the set of all the estimating functions explicitly. The set forms a linear space.
2) The asymptotic covariance matrix of an estimator $\hat{W}$ obtained from an estimating function $F(\boldsymbol{x}, W)$ is explicitly calculated.

Manuscript received June 5, 1997. The associate editor coordinating the review of this paper and approving it for publication was Prof. Jenq-Neng Hwang.

S. Amari is with the RIKEN Frontier Research Program, Saitama, Japan (e-mail: amari@brain.riken.go.jp).

J.-F. Cardoso is with the Signal Department, École National Superieure des Télécommunications, CNRS, Paris, France.

Publisher Item Identifier S 1053-587X(97)08057-4.

3) The Fisher efficient estimator is derived from the optimal F.
4) We decompose the space of estimating functions into the orthogonal sum of an admissible part and ancillary part. The admissible part is spanned by the functions of the form

$$F(\boldsymbol{x}, W) = \varphi(\boldsymbol{y})\boldsymbol{y}^T \tag{1.4}$$

except for the diagonal elements, where

$$\varphi(\boldsymbol{y}) = [\varphi_1(y_1), \cdots, \varphi_m(y_m)]^T$$

with arbitrary function φ_i, and $\boldsymbol{y}^T$ denotes the transposition of a column vector $\boldsymbol{y}$.
5) For any estimating function F, one can always find a better or at least equally good estimating function in the admissible part. This implies that a general form of estimating function such as

$$F(\boldsymbol{x}, W) = \varphi(\boldsymbol{y})\psi(\boldsymbol{y})^T$$

or more complicated forms, include redundant ineffective parts and that a better F is found in the class of form (1.4).
6) When F is an estimating function that gives a consistent estimator $\hat{W}$, the local stability of the related learning equation (1.3) is not guaranteed. We discuss this problem and give converging estimating functions.

The background of the present paper is information geometry [1], [3], that is, a differential geometrical theory of the manifold of probability distributions. We do not state the mathematical details in the present paper. For further discussions on information geometry, see Amari [1], Amari [3], Amari and Kawanabe [9], Murray and Rice [19], etc.

II. Blind Source Separation as Semiparametric Statistical Inference

Let us consider m signal sources that generate signal sequences $s_i(t)$, $i = 1, \cdots, m$ at discrete times $t = 1, 2, \cdots$. It is assumed that m signals s_i are independent. It is also assumed that $s_i(t)$ are iid time series, that is, $s_i(t)$ and $s_i(t')$ are independent and identically distributed (iid) when $t \neq t'$, but this assumption can easily be relaxed. Our analysis can be extended to strongly mixing ergodic sources. Let us denote the signals by a column vector $\boldsymbol{s}(t)$. The joint probability density function $r(\boldsymbol{s})$ of $\boldsymbol{s}(t)$ is then written as

$$r(\boldsymbol{s}) = \prod_{i=1}^{m} r_i(s_i) \tag{2.1}$$

which does not depend on t, where $r_i(s)$ is the probability density function of the ith signal s_i. We assume that the source signals are not Gaussian, or at most, one is Gaussian.

We assume that these source signals are not observed directly. Instead, their instantaneous linear mixtures

$$x_j(t) = \sum_{i=1}^{m} A_i^j s_i(t), \qquad j = 1, \cdots, m$$

are observed. The mixing matrix $A = (A_i^j)$ is unknown, but it is nonsingular and time independent. The observed signal vector $\boldsymbol{x} = (x_i)$ is then written as

$$\boldsymbol{x}(t) = A\boldsymbol{s}(t) \tag{2.2}$$

and the original $\boldsymbol{s}(t)$ is written as

$$\boldsymbol{s}(t) = A^{-1}\boldsymbol{x}(t). \tag{2.3}$$

Let $D_n = \{\boldsymbol{x}(1), \cdots, \boldsymbol{x}(n)\}$ be a data set consisting of n observed signals. The blind separation problem is to recover original signals $S_n = \{\boldsymbol{s}(1), \cdots, \boldsymbol{s}(n)\}$ from D_n without any knowledge of A and the probability distributions $r_i(s)$. If we can estimate A or A^{-1} from the data D_n, S_n is recovered by (2.3). Note that A itself is not identifiable from the data. When signal s_i is multiplied by a scalar c_i, this is equivalent to rescaling the corresponding column of A by c_i. Therefore, the scale of each signal remains undetermined. The usual convention is to assume that the sources s_i satisfy the normalization conditions

$$E[(s_i)^2] = 1 \tag{2.4}$$

where E denotes the expectation with respect to $r_i(s)$, and matrix A is unconstrained. We write the condition in a more general form

$$E[k(s_i)] = 0 \tag{2.5}$$

by using a function k, which in the case of (2.4) is

$$k(s) = s^2 - 1.$$

It is possible to use a general normalization condition other than (2.4). We further assume that s_i has zero mean

$$E[s_i] = 0. \tag{2.6}$$

It should be noted that m sources $\{s_1, \cdots, s_n\}$ cannot be recovered in their exact order. We do not know which is the first signal and which is the second. Hence, we can identify the source signals except for their scales and permutations.

The joint probability density function $p_X(\boldsymbol{x})$ of $\boldsymbol{x}$ is determined by A and $r(\boldsymbol{s})$. It is given by

$$p_X(\boldsymbol{x}; A, r) = |A|^{-1} r(A^{-1}\boldsymbol{x}) \tag{2.7}$$

where $|A|$ is the absolute value of the determinant of matrix A. From the statistical point of view, the problem is to estimate A or A^{-1} from the observed iid data D_n. This includes two unknowns: One is the mixing matrix A, which we want to estimate, and the other is a function r, which we do not care about because it is not necessary for recovering $\boldsymbol{s}(t)$. Such a function r is called the nuisance parameter, whereas A is called the parameter of interest. Here, r is a function that can be decomposed into the product (2.1) of m unknown functions r_i. When the nuisance parameter is of infinite dimensions or functional degrees of freedom, the statistical model (2.7) is called a semiparametric statistical model (Bickel *et al.* [11]).

The estimating function method uses a function $F(\boldsymbol{x}, W)$, where $W = A^{-1}$, to obtain an estimator $\hat{W}$ by solving

$$\sum_t F\{\boldsymbol{x}(t), W\} = 0.$$

The matrix function F should be designed without knowing r because we do not have precise knowledge about r. Since the nuisance parameter r is of infinite dimensions, it is difficult to estimate it from a finite number of observations. There has been a remarkable progress recently in the theory of semiparametric statistics (Amari and Kawanabe [9]; Bickel *et al.* [11]). The next section is devoted to a brief introduction to the results of the geometrical theory of estimating functions in semiparametric models (Amari and Kawanabe [8], [9]).

III. Estimating Functions in Semiparametric Statistical Models

A. Efficient Score and Cramér–Rao Theorem

Consider a general statistical model $\{p(x; \theta, \xi)\}$, where the probability density function of a random variable x is specified by two vector parameters θ and ξ, θ being the parameter of interest, and ξ being the nuisance parameter. In this subsection, we assume that both θ and ξ are of finite dimensions so that the classical statistical theory is applicable to this case. Given n iid observations $D_n = \{x_1, \cdots, x_n\}$, the maximum likelihood estimator (MLE) is known to be an efficient estimator, that is, asymptotically the best estimator. Let $\hat{\theta}$ and $\hat{\xi}$ be the MLE that maximizes the likelihood function

$$\prod_{i=1}^{n} p(x_i; \theta, \xi).$$

The gradient vectors of log likelihood

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

are called the score functions of the parameter of interest or θ score and the nuisance score function or ξ score, respectively. Here, $\partial/\partial\theta$ is the gradient whose components are $\partial/\partial\theta_i$, θ_i being the components of θ, and so on. The MLE is given by solving the likelihood equations

$$\sum_{i=1}^{n} \boldsymbol{u}(x_i; \theta, \xi) = 0, \quad \sum_{i=1}^{n} \boldsymbol{v}(x_i; \theta, \xi) = 0. \tag{3.1}$$

The joint covariance matrix (or the expected squared error) of estimator $\hat{\theta}$ and $\hat{\xi}$ is defined by

$$V = \begin{bmatrix} E[(\hat{\theta} - \theta)(\hat{\theta} - \theta)^T] & E[(\hat{\theta} - \theta)(\hat{\xi} - \xi)^T] \\ E[(\hat{\xi} - \xi)(\hat{\theta} - \theta)^T] & E[(\hat{\xi} - \xi)(\hat{\xi} - \xi)^T] \end{bmatrix}$$

where θ and ξ are the true parameter, and E denotes the expectation. Then, for any unbiased estimators $\hat{\theta}$ and $\hat{\xi}$, the Cramér–Rao theorem gives the lower bound of estimation errors by the inequality

$$V \geq \frac{1}{n} G^{-1} \tag{3.2}$$

where G^{-1} is the inverse of the Fisher information matrix G defined by

$$G = \begin{bmatrix} G_u & G_{uv} \\ G_{vu} & G_v \end{bmatrix} = \begin{bmatrix} E[\boldsymbol{u}\boldsymbol{u}^T] & E[\boldsymbol{u}\boldsymbol{v}^T] \\ E[\boldsymbol{v}\boldsymbol{u}^T] & E[\boldsymbol{v}\boldsymbol{v}^T] \end{bmatrix}. \tag{3.3}$$

Here, the matrix inequality is in the sense of positive definiteness. For the MLE $\hat{\theta}$ and $\hat{\xi}$, the equality holds asymptotically so that the MLE is an efficient estimator, that is, asymptotically the best estimator.

The covariance matrix of the parameter of interest is denoted by

$$V_\theta = E[(\hat{\theta} - \theta)(\hat{\theta} - \theta)^T].$$

By taking the θ part of the matrix G^{-1}, for the MLE $\hat{\theta}$, this is given asymptotically by

$$V_\theta = \frac{1}{n} (G_u - G_{uv} G_v^{-1} G_{vu})^{-1}. \tag{3.4}$$

This V_θ is asymptotically best when the unknown nuisance parameter ξ exists. When the nuisance parameter ξ is known, the covariance matrix of the MLE $\hat{\theta}$ of θ is given asymptotically by

$$V_\theta^* = \frac{1}{n} G_u^{-1}. \tag{3.5}$$

Comparing (3.4) with (3.5), we see that the term $G_{uv} G_v^{-1} G_{vu}$ represents the loss of information caused by the existence of unknown nuisance parameter ξ.

B. Efficient Score

A geometrical interpretation of the Cramér–Rao theorem is shown by again using the finite dimensional θ and ξ. This gives a good introduction to the infinite-dimensional semiparametric case. We consider the set of functions $w(x)$ of x

$$H_{\theta,\xi} = \{w(x) \mid E_{\theta,\xi}[w(x)] = 0 \quad E_{\theta,\xi}[\{w(x)\}^2] < \infty\} \tag{3.6}$$

where $E_{\theta,\xi}$ denotes the expectation with respect to $p(x; \theta, \xi)$. This is a linear space defined at each (θ, ξ). The linear space $H_{\theta,\xi}$ is a Hilbert space, where the inner product of $w_1(x), w_2(x) \in H_{\theta,\xi}$ is defined by

$$\langle w_1(x), w_2(x) \rangle = E_{\theta,\xi}[w_1(x) w_2(x)] \tag{3.7}$$

where we show by $E_{\theta,\xi}$ explicitly that the expectation is taken with respect to $p(x; \theta, \xi)$.

The components $u_i(x; \theta, \xi)$ and $v_i(x; \theta, \xi)$ of the θ score $\boldsymbol{u}$ and ξ score $\boldsymbol{v}$ belong to $H_{\theta,\xi}$, provided the Fisher information exists because, for example

$$\begin{aligned} E[u_i(x; \theta, \xi)] &= \int p(x; \theta, \xi) \frac{\partial}{\partial \theta_i} \log p(x; \theta, \xi)\, dx \\ &= \int \frac{\partial}{\partial \theta_i} p(x; \theta, \xi)\, dx \\ &= \frac{\partial}{\partial \theta_i} \int p(x; \theta, \xi)\, dx = 0. \end{aligned}$$

Geometrically speaking, $w(x) \in H_{\theta,\xi}$ represents a small deviation of the probability density function from $p(x; \theta, \xi)$ to

$$p(x; \theta, \xi)\{1 + \varepsilon w(x)\}$$

where ε is a small constant. Since this is again a probability distribution, it must integrate to 1, showing that $E_{\theta,\xi}[w(x)] = 0$ is required. The set $H_{\theta,\xi}$ of all such deviations $w(x)$ can

be regarded as the "tangent space" at $p(x; \theta, \xi)$ of the set of all the probability distributions. Among them, $u_i(x; \theta, \xi)$ and $v_i(x; \theta, \xi)$ represent the deviations caused by small changes in θ_i and in ξ_i, respectively. Any other $w(x)$ includes a deviation that cannot be realized by changing θ and ξ.

Since u_i denotes a change in the direction of θ_i, we can interpret it as the tangent direction along the coordinate curve θ_i. The subspace $T^N_{\theta,\xi}$ spanned by the ξ-score functions $v_i(x; \theta, \xi)$ is called the nuisance tangent space at (θ, ξ) because it represents the directions caused by changes in ξ. Let us decompose the θ-score function $u_i(x; \theta, \xi)$ as a sum of the component included in $T^N_{\theta,\xi}$ and that orthogonal to it. The component in the direction of $T^N_{\theta,\xi}$ is obtained by projecting u_i to it and is the $v(x)$ that minimizes

$$\|u_i(x; \theta, \xi) - v(x)\|^2 = \langle u_i - v, u_i - v\rangle, \qquad v \in T^N_{\theta,\xi}.$$

This is written as

$$v(x) = \sum_{j,k} \langle u_i, v_j\rangle G^{-1}_{jk} v_k(x; \theta, \xi)$$

where G^{-1}_{jk} is the inverse of the ξ part of the Fisher information matrix

$$G_v = \langle v_j, v_k\rangle.$$

The component orthogonal to $T^N_{\theta,\xi}$ is then given by

$$u^E_i(x; \theta, \xi) = u_i(x; \theta, \xi) - v(x).$$

The vector function $\boldsymbol{u}^E(x; \theta, \xi)$ composed of (u^E_i) is called the efficient score. This is the vector whose components are orthogonal to $T^N_{\theta,\xi}$ and is explicitly written as

$$\boldsymbol{u}^E(x; \theta, \xi) = \boldsymbol{u}(x; \theta, \xi) - G_{uv} G_v^{-1} \boldsymbol{v}(x; \theta, \xi).$$

The matrix

$$G^E = E[\boldsymbol{u}^E (\boldsymbol{u}^E)^T]$$

is called the efficient Fisher information matrix. It is given by

$$G^E = G_u - G_{uv} G_v^{-1} G_{vu} \tag{3.8}$$

so that the Cramér–Rao bound (3.4) is rewritten as

$$V_\theta \geq \frac{1}{n} G_E^{-1}. \tag{3.9}$$

The equality asymptotically holds for the covariance matrix of the MLE $\hat{\theta}$. When the value of the nuisance parameter is known, the Fisher information matrix is

$$G_u = E_{\theta,\xi}[\boldsymbol{u}\boldsymbol{u}^T].$$

It is larger than G_E, as shown by (3.8), in the sense of the positive definiteness

$$G_u \geq G^E$$

and it quantifies the cost of having to estimate nuisance parameters. The Fisher information G_u represents the magnitude of the θ score $\boldsymbol{u}$. In the case when the unknown nuisance parameter ξ exists, $\boldsymbol{u}$ is decomposed as the sum of $\boldsymbol{u}_E$ whose components are orthogonal to $T^N_{\theta,\xi}$ and $(\boldsymbol{u} - \boldsymbol{u}_E)$ whose components belong to $T^N_{\theta,\xi}$. The inequality (3.9) shows that the effective part of the θ-score function is only that orthogonal to the ξ score, and the part parallel to the ξ score becomes ineffective when ξ is unknown.

C. Estimating Functions

In the case of a semiparametric model, ξ is infinite dimensional so that the joint estimation of θ and ξ is difficult. Here, we state the method of estimating functions to obtain a consistent estimator $\hat{\theta}$ of the parameter of interest. In this section, the nuisance parameter ξ may be of infinite dimensions or a function. A vector function $\boldsymbol{z}(x, \theta)$ that does not depend on ξ is called an estimating function when the following conditions are satisfied for all θ and ξ.

1) $E_{\theta,\xi}[\boldsymbol{z}(x, \theta)] = 0$ (3.10)

2) $\det |K| \neq 0$, where $K = E_{\theta,\xi}\left[\dfrac{\partial}{\partial\theta}\boldsymbol{z}(x, \theta)\right]$ (3.11)

3) $E_{\theta,\xi}[\boldsymbol{z}(x, \theta)\boldsymbol{z}^T(x, \theta)] < \infty$. (3.12)

Given an estimating function $\boldsymbol{z}(x, \theta)$, by replacing the expectation in (3.10) by the empirical average over data D_n, we obtain the estimating equation

$$\sum_{i=1}^{n} \boldsymbol{z}(x_i, \theta) = 0. \tag{3.13}$$

The estimator obtained from (3.13) is called an M estimator. An M estimator is consistent, where $\hat{\theta}$ converges to the true value as n tends to infinity, whatever ξ is. Its covariance matrix is given asymptotically by

$$V_\theta = \frac{1}{n} K^{-1} E_{\theta,\xi}[\boldsymbol{z}\boldsymbol{z}^T](K^{-1})^T \tag{3.14}$$

where the dependence of V_θ on ξ is not explicitly denoted. Matrix V_θ gives the asymptotic evaluation of the estimator obtained from an estimating function $\boldsymbol{z}(x, \theta)$.

Let $R(\boldsymbol{\theta})$ be a nonsingular matrix depending on $\boldsymbol{\theta}$. Then, it is easy to see that when $\boldsymbol{z}(\boldsymbol{x}, \boldsymbol{\theta})$ is an estimating function, $R(\boldsymbol{\theta})\boldsymbol{z}(\boldsymbol{x}, \boldsymbol{\theta})$ is also an estimating function. The two estimating functions $\boldsymbol{z}$ and $R\boldsymbol{z}$ give the same estimating equations and the same estimator. In this sense, $\boldsymbol{z}$ and $R\boldsymbol{z}$ are regarded as equivalent. However, the two learning equations

$$\boldsymbol{\theta}_{t+1} = \boldsymbol{\theta}_t - \eta_t \boldsymbol{z}(\boldsymbol{x}_t, \boldsymbol{\theta}_t)$$
$$\boldsymbol{\theta}_{t+1} = \boldsymbol{\theta}_t - \eta_t R(\boldsymbol{\theta}_t)\boldsymbol{z}(\boldsymbol{x}_t, \boldsymbol{\theta}_t)$$

have different dynamic characteristics, even though they have the same equilibria points. We will study this in a later section.

The method of estimating functions can easily be applied to the semiparametric model where the nuisance parameter ξ is of functional degrees of freedom and is difficult to estimate. However, it is not easy to find an estimating function in general.

Amari and Kawanabe [8], [9] proposed the information-geometrical theory of estimating functions and solved the following problems.

1) Obtain a condition that guarantees the existence of estimating functions.

2) Obtain the linear function space of all the estimating functions.
3) Obtain a condition that guarantees that the optimal M estimator is Fisher efficient.

The results of Amari and Kawanabe [8], [9] are applied to the blind separation problem. It was shown that the manifold of the probability distributions (2.7) of blind separation is information-curvature free (Amari and Kawanabe [9]), implying the following results.

1) The set of M estimators includes the Fisher efficient estimator. Therefore, the best estimator is obtained by the method of estimation functions.
2) The efficient score vector function $\boldsymbol{u}^E(x;\, \theta,\, \xi_0)$, where ξ_0 is an arbitrarily fixed function, is an estimating function satisfying, in particular

$$E_{\theta,\xi}[\boldsymbol{u}^E(x;\, \theta,\, \xi_0)] = 0$$

for any ξ_0 and ξ (at least locally). This is the optimal estimating function when the true nuisance parameter is ξ_0. It should be remarked that $\boldsymbol{u}^E$ is not an estimating function in the general case.

We need to calculate the efficient score $\boldsymbol{u}^E$ in our semiparametric model $p_X(\boldsymbol{x};\, A,\, r)$ of blind separation.

IV. Efficient Scores

A. Score Function Matrix

The present paper applies the results by Amari and Kawanabe [9] to the blind separation problem. The parameter of interest θ is the mixing matrix A, and the nuisance parameter ξ is a function r. Hence, an estimating function in the present problem is a matrix $F(\boldsymbol{x};\, A)$ that satisfies

$$E_{A,\,r}[F(\boldsymbol{x};\, A)] = 0 \tag{4.1}$$

for all A and r, together with the additional conditions given by (3.11) and (3.12). Here, $E_{A,\,r}$ denotes the expectation with respect to $p_X(\boldsymbol{x};\, A,\, r)$.

Here, we give typical examples of estimating functions. Let $\boldsymbol{y} = W\boldsymbol{x}$, and let $\varphi(\boldsymbol{y})$ be a vector defined by

$$\varphi(\boldsymbol{y}) = [\varphi_1(y_1),\, \cdots,\, \varphi_m(y_m)]^T$$

where φ_i are adequate nonlinear functions. Then, the following matrix

$$F(\boldsymbol{x},\, W) = \Lambda - \boldsymbol{\varphi}(\boldsymbol{y})\boldsymbol{y}^T$$

where Λ is a diagonal matrix, is an estimating function under adequate regularity conditions because

$$E[\varphi(y_i)y_j] = \lambda_i\delta_{ij}$$

holds when the expectation is taken at the true W, that is, when y_i and y_j become independent. These types of estimating functions are widely used and derived from various heuristic ideas by Jutten and Hérault [16], the max-entropy principle [10], [20], minimizing cross cumulants [13], independent component analysis [6], [14], and so on. The natural gradient [2], [6] or the relative gradient [12] of the likelihood precisely is in the form $F(\boldsymbol{x},\, W) = [\Lambda - \varphi(\boldsymbol{y})\boldsymbol{y}^T]W$ and is an estimating function as well.

Let $H_{A,\,r}$ be the Hilbert space of functions that satisfy

$$H_{A,\,r} = \{f(\boldsymbol{x}) \mid E_{A,\,r}[f(\boldsymbol{x})] = 0,\, E_{A,\,r}[f^2] < \infty\}. \tag{4.2}$$

We now define the A-score functions. The A-score function is a matrix valued function defined by

$$U(\boldsymbol{x};\, A,\, r) = \frac{\partial}{\partial A}\log p(\boldsymbol{x};\, A,\, r) \tag{4.3}$$

where $\partial/\partial A$ denotes the gradient operator whose elements are given by $(\partial/\partial A_i^j)\log p_X(\boldsymbol{x};\, A,\, r)$. We will give its explicit form in Section IV-C.

B. Nuisance Score

Next, we define the nuisance scores and the nuisance tangent space $T_{A,\,r}^N$ spanned by them. Since the nuisance parameter is a function of the form

$$r(\boldsymbol{s}) = \prod r_i(s_i)$$

we consider a small deviation of each distribution function $r_i(s)$ in the direction of function $\alpha_i(s)$. This deviation is given by a one-parameter family

$$r_i(s;\, \alpha_i,\, \delta) = r_i(s)\{1 + \delta\alpha_i(s)\} \tag{4.4}$$

where δ ($0 \le \delta < \varepsilon$) is a parameter denoting the magnitude of deviation. Here, $\alpha_i(s)$ is a function satisfying

$$E_{A,\,r}[\{\alpha_i(s)\}^2] < \infty. \tag{4.5}$$

We call the one-parameter family $r_i(s;\, \alpha_i,\, \delta)$ the path of deviation of r_i in the direction of α_i.

Since $r_i(s;\, \alpha_i,\, \delta)$ is a probability density function of s satisfying (2.5) and (2.6), we easily find by differentiation of these constraints that α_i must also satisfy

$$E_{A,\,r}[s\alpha_i(s)] = 0 \tag{4.6}$$

$$E_{A,\,r}[k(s)\alpha_i(s)] = 0. \tag{4.7}$$

The nuisance score function in the direction of deviation $\boldsymbol{\alpha} = [\alpha_1(s_1),\, \cdots,\, \alpha_m(s_m)]$ is defined by

$$v(\boldsymbol{x};\, \boldsymbol{\alpha}) = \frac{d}{d\delta}\log p\{\boldsymbol{x};\, A,\, r(\cdot;\, \boldsymbol{\alpha},\, \delta)\}|_{\delta=0} \tag{4.8}$$

where $r(\cdot;\, \boldsymbol{\alpha},\, \delta)$ is a function given by

$$r(\boldsymbol{s};\, \boldsymbol{\alpha},\, \delta) = \prod_{i=1}^{m} r_i(s_i;\, \alpha_i,\, \delta). \tag{4.9}$$

The nuisance score (4.8) is explicitly calculated as

$$v(\boldsymbol{x};\, \boldsymbol{\alpha}) = \sum_i \frac{d}{d\delta}\log r_i(s_i;\, \alpha_i,\, \delta)|_{\delta=0} = \sum_i \alpha_i(s_i).$$

Therefore, the tangent nuisance space $T_{A,\,r}^N$ is the linear space spanned by the nuisance score functions

$$T_{A,\,r}^N = \text{span}\,\{v(\boldsymbol{x};\, \boldsymbol{\alpha})\} = \left\{\sum_{i=1}^{m} c_i\alpha_i(s_i)\right\} \tag{4.10}$$

where c_i are coefficients, and α_i are arbitrary functions satisfying (4.5)–(4.7).

C. Reparameterization to Calculate Scores

We have used the mixing matrix A as the parameter of interest. However, it is convenient to reparameterize it by using another parameter matrix E as follows. Let us fix A_0, and put

$$A(E) = A_0(I - E) \tag{4.11}$$

where I is the identity matrix. Then, E plays the role of a local coordinate system at a neighborhood N_{A_0} of A_0 in the space of all the nonsingular mixing matrices. The origin $E = 0$ corresponds to A_0. A small change dE of E corresponds to a small change $dA = -A_0 dE$ of A. It is easy to show that the gradient in terms of A at A_0 is expressed as

$$\frac{\partial}{\partial A} \log p(\boldsymbol{x}; A_0, r) = -(A_0^{-1})^T \left\{ \frac{\partial \log p(\boldsymbol{x}; A, r)}{\partial E} \bigg|_{E=0} \right\}. \tag{4.12}$$

By using this parameter E, whatever the true parameter A_0 is, the neighborhood N_{A_0} is mapped to a neighborhood N_I of the unit matrix so that, in terms of the parameter E, inference about A can be analyzed in the form not depending on a specific A. This is the equivariant property (Cardoso and Laheld [12]). The relative or the natural Riemannian gradient (Amari [2], Amari *et al.* [6], Cardoso and Laheld [12]) amounts to differentiating with respect to the local parameter E.

The score functions $U = (u_{ij})$ in terms of E are given by

$$u_{ij} = \frac{\partial}{\partial E_{ij}} \log p_X\{\boldsymbol{x}; A(E), r\}|_{E=0}. \tag{4.13}$$

Putting

$$\varphi_i(s) = \frac{d}{ds} \log r_i(s) \tag{4.14}$$

we have

$$\frac{\partial}{\partial E_{ij}} \log p_X\{\boldsymbol{x}; A(E), r\}|_{E=0} = \varphi_i(s_i)s_j + \delta_{ij}$$

where δ_{ij} is the Kronecker delta. This can be written in the matrix-vector form as

$$\frac{\partial}{\partial E} \log p_X(\boldsymbol{x}; A, r) = \varphi(\boldsymbol{s})\boldsymbol{s}^T + I \tag{4.15}$$

where $\varphi(\boldsymbol{s})$ represents the column vector whose components are $\varphi_i(s_i)$, $i = 1, \cdots, m$, and $\boldsymbol{s} = A^{-1}\boldsymbol{x}$.

D. Efficient Scores

The efficient scores $U^E(\boldsymbol{s}; A, r)$ or its elements $u_{ij}^E(\boldsymbol{s}; A, r)$ are obtained by projecting the scores u_{ij} to the subspace orthogonal to the nuisance tangent space $T_{A,r}^N$. To this end, we show the following lemma.

Lemma 1: The off-diagonal elements $u_{ij}(\boldsymbol{s}; A, r)$, $i \neq j$ of the scores are orthogonal to the nuisance tangent space $T_{A,r}^N$.

Proof: The inner product of u_{ij} and an element $\sum c_k\alpha_k \in T_{A,r}^N$ are calculated as

$$\left\langle u_{ij}(\boldsymbol{s}; A, r), \sum c_k\alpha_k(s_k) \right\rangle = \sum c_k \langle \varphi_i(s_i)s_j, \alpha_k(s_k)\rangle.$$

This vanishes because s_i's are mutually independent, and

$$\langle \varphi_i(s_i)s_j\alpha_k(s_k)\rangle = E[\varphi_i(s_i)s_j\alpha_k(s_k)] = 0$$

whenever $i \neq j$.

The lemma shows that the off-diagonal elements of the efficient scores are given by $u_{ij}^E = u_{ij}$. Next, we calculate the projection of the diagonal elements u_{ii}.

Lemma 2: The projection of u_{ii} to the space orthogonal to $T_{A,r}^N$ is of the form

$$w(s_i) = c_1 k(s_i) + c_2 s_i \tag{4.16}$$

where c_1 and c_2 are constants.

Proof: Because of the conditions (4.6) and (4.7), any nuisance score $\alpha_i(s_i)$ is orthogonal to the linear space spanned by s_i and $k(s_i)$ but is otherwise free. Therefore, $u_{ii} = \varphi_i(s_i)s_i + 1$ can be decomposed in two orthogonal terms

$$u_{ii} = w(s_i) + \{u_{ii} - w(s_i)\} \tag{4.17}$$

where $w(s_i)$ is a linear combination of s_i and $k(s_i)$, and $u_{ii} - w$ is orthogonal to $w(s_i)$ and can be regarded as some nuisance score $\alpha_i(s_i)$. Hence, the diagonal components of the efficient score reduce to the form (4.16). The coefficients c_1 and c_2 are easily determined.

Theorem 1: The efficient score $U^E(\boldsymbol{x}; A, r)$ is given by

$$U^E(\boldsymbol{x}; A, r) = \begin{cases} \varphi(\boldsymbol{s})\boldsymbol{s}^T, & \text{for off-diagonal elements} \\ C_1 k(\boldsymbol{s}) + C_2 \boldsymbol{s}, & \text{for diagonal elements} \end{cases} \tag{4.18}$$

where C_1 and C_2 are diagonal matrices, and $k(\boldsymbol{s}) = [k(s_1), \cdots, k(s_m)]^T$.

V. Estimating Functions and Their Efficiencies

A. The Space of Estimating Functions

As is shown in Theorem 1, the diagonal elements u_{ii}^E of the efficient score have a different form from the off-diagonal elements u_{ij}^E. This fact is also true of the components of an estimating function. We can treat them separately: We discuss the off-diagonal elements; the diagonal elements are trivial because they are of the form $c_1 k(s_i) + c_2 s_i$ at any r.

Let us define the linear space spanned by the off-diagonal components of the efficient score $U^E(\boldsymbol{x}; A, r)$

$$T_{A,r}^E = \text{span } \{\varphi_i(s_i; r)s_j \, (i \neq j)\}. \tag{5.1}$$

Since function φ depends on r, it is explicitly denoted as $\varphi(s_i; r)$. Let us define the space T_A^E, which is spanned by all the $T_{A,r}^E$ for all r

$$T_A^E = \text{span } \{\varphi_i(s_i; r)s_j \, (i \neq j) \text{ for any } r\}. \tag{5.2}$$

Let us define the ancillary space $T_{A,r}^A$ at (A, r), which consists of all the functions satisfying

$$E_{A,r}[a(\boldsymbol{x})] = 0 \tag{5.3}$$

whose components are orthogonal to $T^N_{A,r}$ and $T^E_{A,r}$. We define T^A_A, which consists of those satisfying (5.3) and orthogonal to $T^E_{A,r}$ and $T^N_{A,r}$ at all r.

It has been shown in Amari and Kawanabe [9] that the semiparametric model (2.7) is information-m-curvature free. This implies that

$$E_{A,r}[U^E(\boldsymbol{x}; A, r')] = 0 \tag{5.4}$$

for any r and r'. This property does not hold for a general semiparametric model. It holds for the source separation model because for the off-diagonal elements, we have

$$E_{A,r}[\varphi_i(s_i; r')s_j] = 0$$

(since s_i and s_j are independent) and for the diagonal elements, u^E_{ii} is equal to $c_1 k(s_i) + c_2 s_i$ so that

$$E_{A,r}[u^E_{ii}(s_i)] = 0$$

always holds because of (2.5) and (2.6). This guarantees the following theorem.

Theorem 2: A matrix function $F(\boldsymbol{x}, A)$ or its equivalence $R(A)F(\boldsymbol{x}, A)$ is an estimating function when it is a sum of a linear combination of the elements of $U^E(\boldsymbol{x}; A, r_0)$'s for some r_0's and of a matrix whose elements belong to T^A_A.

We can decompose an estimating function $F(\boldsymbol{x}; A)$ as a sum

$$F(\boldsymbol{x}; A) = C(A, r)U^E(\boldsymbol{x}; A, r) + B(\boldsymbol{x}; A, r) \tag{5.5}$$

at each (A, r), where $C(A, r)$ is an arbitrary nonsingular linear operator such that (3.11) holds, and the components of B belong to $T^A_{A,r}$. This decomposition depends on r. It is useful for evaluating the Fisher efficiency of the related M estimator.

B. Efficiencies of Estimators

The asymptotic covariance matrix of an estimator derived from an estimating function matrix is evaluated by (3.14). Since the parameter of interest θ is matrix A, its covariance matrix is represented by a fourth-order tensor. Here, we state general results by using the parameter θ and ξ instead of A and r, thus avoiding complicated notations of tensors.

Let us decompose an estimating function $\boldsymbol{z}(x, \theta)$ as a sum

$$\boldsymbol{z}(x, \theta) = \boldsymbol{z}_E(x; \theta, \xi) + \boldsymbol{z}_A(x; \theta, \xi) \tag{5.6}$$

where $\boldsymbol{z}_E$ and $\boldsymbol{z}_A$ belong to the space of the efficient scores and the ancillary space, respectively, at a point (θ, ξ). Since components of $\boldsymbol{z}_E$ are orthogonal to those of $\boldsymbol{z}_A$ in the function space, we have

$$E_{\theta,\xi}[\boldsymbol{z}\boldsymbol{z}^T] = E_{\theta,\xi}[\boldsymbol{z}_E\boldsymbol{z}_E^T] + E_{\theta,\xi}[\boldsymbol{z}_A\boldsymbol{z}_A^T]. \tag{5.7}$$

Since $\boldsymbol{z}_E$ is the efficient score at (θ, ξ), we have from (3.14)

$$G^E = K^T(E_{\theta,\xi}[\boldsymbol{z}_E\boldsymbol{z}_E^T])^{-1}K$$

so that the covariance matrix V_θ of this estimator is decomposed as

$$V_\theta = (G^E)^{-1} + K^{-1}E_{\theta,\xi}[\boldsymbol{z}_A\boldsymbol{z}_A^T](K^{-1})^T. \tag{5.8}$$

This leads us to another important theorem.

Theorem 3: When the true nuisance parameter is r_0, the estimator using the estimating function $U^E(\boldsymbol{x}; A, r_0)$ is asymptotically Fisher efficient.

The theorem is not trivial, even though we do not know r_0. It guarantees that the Fisher efficient estimator is obtained if we can estimate r_0. Moreover, even if we misspecify r_0 and use $U^E(\boldsymbol{x}; A, \hat{r})$, where $\hat{r}$ is different from the true r_0, this still gives a consistent estimator of A. When $\hat{r}$ is close to r_0, its efficiency is good. This suggests that we use adaptive methods of obtaining good estimators. One is to use a parametric family of estimating functions such as

$$U(\boldsymbol{x}; A) = \sum_{k=1}^{p} c_k U^E\{\boldsymbol{x}; A, r_k\}$$

where r_k $(k = 1, \cdots, p)$ are adequately chosen probability distributions, and determine the coefficients c_k from the data (see Lindsay [18] and Pham *et al.* [22]). The other is to use a parametric family of distribution $r(\boldsymbol{s}, \boldsymbol{\eta})$, where $\boldsymbol{\eta}$ is a finite-dimensional parameter. It is then easy to estimate $\hat{\boldsymbol{\eta}}$ from the data. Since the true r_0 does not necessarily belong to $\{r(\boldsymbol{s}, \boldsymbol{\eta})\}$, $\hat{r} = r(\boldsymbol{s}, \hat{\boldsymbol{\eta}})$ does not converge to the true r_0. However, the estimating function $U^E(\boldsymbol{x}; A, \hat{r})$ gives a good performance. This was used in Amari *et al.* [6], where the third and fourth cumulants are used as $\boldsymbol{\eta}$.

We state another important consequence of the theorem. Let us consider a more general function of the form like $\varphi_i(s_i)\psi(s_j)$ or $\sum c_k\varphi_i(s_i)\psi(s_j)\pi(s_k)$, etc. They do not belong to $T^E_{A,r}$ or linear combinations of its elements. It looks as though these functions are more general than those of the form $\varphi_i(s_i)s_j$, and it has been suggested that such functions might increase efficiency of the estimator. This is not true. The simplest form of $\varphi_i(s_i)s_j$ and its concomitant $\varphi_j(s_j)s_i$ (except for diagonal elements) is the best one. The point is how to choose $\varphi(s_i)$.

Theorem 4: The estimator obtained from the general form $U(\boldsymbol{x}; A)$ other than (5.2) always has a larger asymptotic variance than its projection to T^E_A of (5.2).

Note, however, that the theorem characterizes only the local asymptotic behavior of estimates. Other estimating functions may be preferred for global properties (like absence of spurious roots).

VI. Learning Algorithm

Learning algorithms are used in most proposals for blind source separation. Let W_t be an estimator of A^{-1} at time t, and let us put

$$\boldsymbol{y}_t = W_t\boldsymbol{x}_t. \tag{6.1}$$

The learning algorithm is written as

$$W_{t+1} = W_t - \eta_t F(\boldsymbol{y}_t, W_t) \tag{6.2}$$

where η_t is a learning rate, and F is a matrix specifying the learning rule. This can be rewritten in the equivariant form (Cardoso and Laheld [12]) in the natural gradient form [2] as

$$W_{t+1} = W_t - \eta_t F(\boldsymbol{y}_t, W_t)W_t.$$

Let $F(\boldsymbol{y}; W)$ be an arbitrary estimating function where $W = A^{-1}$. Since

$$E_{A,r}[F(\boldsymbol{y}; A^{-1})] = 0$$

the true $W = A^{-1}$ is an equilibrium of the averaged version of (6.2). However, there is no guarantee that it is stable. Indeed, let us put

$$\tilde{F}(\boldsymbol{y}; W) = R(W)F(\boldsymbol{y}; W) \tag{6.3}$$

where $R(W)$ is a linear operator that maps a matrix to a matrix. Then, the two estimating equations

$$\sum F(\boldsymbol{y}_t; W) = 0 \quad \text{and} \quad \sum \tilde{F}(\boldsymbol{y}_t; W) = 0$$

are essentially the same so that they give the same estimator. Therefore, $\tilde{F}$ is an estimating function giving the same asymptotic efficiency. However, their stability may be different in the case of the learning equations of (6.2).

When F is a gradient of a potential function and the true solution is a local minimum, the local stability is guaranteed. Moreover, it is known that the relative or natural gradient has a good convergence property as well as the equivariant property. However, an estimating function matrix F is not, in general, of the gradient form. Even when it is of the gradient form, it is not, in general, guaranteed that the true solution is the only local minimum. Therefore, the entropy maximization method [10], the cumulant minimization method, etc., do not work for arbitrary source distributions, although they give estimating functions.

How can we derive a converging learning rule from a given F? Let us define the invariant estimating function by

$$F^*(\boldsymbol{y}; W) = K^{-1}(W)F(\boldsymbol{y}; W) \tag{6.4}$$

where K is an operator defined by

$$K(W) = \frac{\partial}{\partial W} E[F(\boldsymbol{y}; W)]. \tag{6.5}$$

It is easy to show that two equivalent estimating functions F and $R(W)F$ have the same invariant F^*. The importance of F^* is shown by the Theorem 5.

Theorem 5: The demixing matrix W is a stable equilibrium of the equation

$$W_{t+1} = W_t - \eta_t F^*(\boldsymbol{y}_t, W_t). \tag{6.6}$$

Proof: By putting $\Delta W_t = W_t - W$, we linearize $F^*(\boldsymbol{y}; W_t)$ at the true W

$$F^*(\boldsymbol{y}, W_t) = F^*(\boldsymbol{y}, W) + \frac{\partial F^*}{\partial W} \Delta W_t.$$

The expectation of $F^*(\boldsymbol{y}, W)$ vanishes, and the expectation of $\partial F^*/\partial W$ is the identity operator because of

$$E\left[\frac{\partial F^*}{\partial W}\right] = \frac{\partial K^{-1}(W)}{\partial W} E[F(\boldsymbol{y}; W)] + K^{-1}E\left[\frac{\partial F}{\partial W}\right]$$
$$= K^{-1}K = \text{identity}.$$

Hence, (6.6) is written as

$$\Delta W_{t+1} = (1 - \eta_t)\Delta W_t + \eta_t D$$

where D is the fluctuating term whose expectation is 0. Hence, for an adequate η_t, the stochastic approximation theory guarantees the convergence of ΔW_t to 0 or W_t to W under a certain regularity conditions on D.

It is not difficult in many cases to calculate explicit forms of $K(W)$ and its inverse at $W = A^{-1}$ (see Cardoso and Laheld [12]). Explicit forms of K and K^{-1} are also given in [5]. Moreover, Amari [4] used them to prove that the "superefficiency" of estimation and learning holds under certain conditions.

VII. Conclusions

The present paper uses the information geometry of estimating functions in a semiparametric statistical model to elucidate the statistical and geometrical structures of blind source separation. We have given general solutions to this problem and studied the asymptotic efficiencies of the algorithms. The optimal estimator is included in this class. Most existing methods use estimating functions. We give a general form of admissible estimating functions, which clearly proves that some general types of estimating functions are redundant. The optimal estimating function, however, depends on the unknown probability distribution $r(\boldsymbol{s})$ of sources. The method of estimating functions works well even if we misspecify the true $r(\boldsymbol{s})$. This leads to the adaptive method of obtaining a good estimating function. We have also studied the local convergence of the learning algorithms to the true solutions. The present theoretical framework can be applied to multiterminal blind deconvolution [7], [15].

Acknowledgment

The authors thank Dr. A. Back at RIKEN for his editorial assistance.

References

[1] S. Amari, *Differential-Geometrical Method in Statistics*. New York: Springer-Verlag, 1985, vol 28.
[2] ——, "Natural gradient works efficiently in learning," *Neural Comput.*, to be published.
[3] ——, "Information geometry," in *Geometry and Nature*, H. Necka and J.-P. Bourguignon, Eds., *Contemporary Math.*, vol. 203, pp. 81–95, 1997.
[4] ——, "Superefficiency in blind source separation," submitted for publication.
[5] S. Amari, T.-P. Chen, and A. Cichocki, "Stability analysis of adaptive blind source separation," *Neural Networks*, to be published.
[6] S. Amari, A. Cichocki, and 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. Cambridge, MA: MIT Press, 1996, pp. 757–763, vol. 8.
[7] S. Amari, S. C. Douglas, A. Cichocki, and H. H. Yang, "Novel on-line adaptive learning algorithms for blind deconvolution using the natural gradient approach," in *Proc. IFAC Symp. Syst. Identification*, Kitakyushu-shi, 1997, pp. 1055–1062.
[8] S. Amari and M. Kawanabe, "Estimating functions in semiparametric statistical models," in *Estimating Functions*, V. P. Godambe, Ed., 1997, IMS Monograph Series.
[9] ——, "Information geometry of estimating functions in semi-parametric statistical models," *Bernoulli*, vol. 3, pp. 29–54, 1997.
[10] A. J. Bell and T. J. Sejnowski, "An information-maximization approach to blind separation and blind deconvolution," *Neural Comput.*, vol. 7, pp. 1129–1159, 1995.
[11] P. J. Bickel, C. A. J. Klaassen, Y. Ritov, and J. A. Wellner, *Efficient and Adaptive Estimation for Semiparametric Models.* Baltimore, MD: Johns Hopkins Univ. Press, 1993.

[12] J.-F. Cardoso and B. Laheld, "Equivariant adaptive source separation," *IEEE Trans. Signal Processing*, vol. 44, pp. 3017–3030, Dec. 1996.
[13] J.-F. Cardoso and A. Souloumiac, "Blind beamforming for non Gaussian signals," *Proc. IEEE*, vol. 140, pp. 362–370, Dec. 1993.
[14] P. Common, "Independent component analysis, a new concept?," *Signal Process.*, vol. 36, pp. 287–314, 1994.
[15] A. Gorokhov and J.-F. Cardoso, "Equivariant blind deconvolution of MIMO-FIR channels," in *Proc. Signal Process. Adv. Wireless Commun. Workshop*, Paris, France, 1997, pp. 89–92.
[16] C. Jutten and J. Hérault, "Independent component analysis versus PCA," in *Proc. EUSIPCO*, 1988, pp. 643–646.
[17] J. Karhunen and J. Joutsensalo, "Representation and separation of signals using nonlinear PCA type learning," *Neural Networks*, vol. 7, no. 1, pp. 113–127, 1993.
[18] B. G. Lindsay, "Using emprical partially Bayes inference for increased efficiency," *Ann. Statist.*, vol. 13, pp. 914–931, 1985.
[19] M. K. Murray and J. W. Rice, *Differential Geometry and Statistics.* New York: Chapman & Hall, 1993.
[20] J.-P. Nadal and N. Parga, "Nonlinear neurons in the low-noise limit: A factorial code maximizes information transfer," *Network*, vol. 7, pp. 565–581, 1994.
[21] D. T. Pham, "Blind separation of instantaneous mixtures of sources via an independent component analysis," *IEEE Trans. Signal Processing*, vol. 44, pp. 2768–2779, 1996.
[22] D. T. Pham, P. Garrat, and C. Jutten, "Separation of a mixture of independent sources through a maximum likelihood approach," in *Proc. EUSIPCO*, 1992, pp. 771–774.

Shun-ichi Amari (F'94) was born in Tokyo, Japan, on January 3, 1936. He graduated from the University of Tokyo in 1958, majoring in mathematical engineering, and received the Dr.Eng. degree from the University of Tokyo in 1963.

He was an Associate Professor at Kyushu University, an Associate and then Full Professor at the Department of Mathematical Engineering and Information Physics at the University of Tokyo, and is now Professor-emeritus at the University of Tokyo. He is the Director of the Brain Information Processing Group, Riken Frontier Research Program. He has been engaged in research in wide areas of mathematical engineering or applied mathematics, such as topological network theory, differential geometry of continuum mechanics, pattern recognition, mathematical foundations of neural networks, and information geometry.

Dr. Amari is Past President of the International Neural Network Society, was Vice President of IEICE, and served as a founding Coeditor-in-Chief of *Neural Networks*. He received the Japan Academy Award, the IEEE Neural Networks Pioneer Award, and IEEE Emanuel R. Piore Award, among many others.

Jean-François Cardoso (M'91) was born in Tunis, Tunisia, on March 1, 1958. He received the Agrégation de Physique degree from the École Normale Supérieure de Saint-Cloud, France, in 1981 and the Doctorat de Physique degree from the University of Paris, France, in 1984.

He currently is with the CNRS and works in the "Signal" Department, Télécom Paris (former ENST). His research interests are in statistical signal processing with emphasis on (blind) array processing and connections to information theory.