Independent Component Analysis in the presence of Gaussian Noise

Motoaki Kawanabe and Noboru Murata

METR 2000-02 March 2000

Independent Component Analysis in the presence of Gaussian Noise

Motoaki Kawanabe* Noboru Murata†

March 28, 2000.

Abstract

The problem of estimating the statistical model of independent component analysis in the presence of Gaussian noise is considered. Because of the additive noise, a combination of factor analysis and a noise-free ICA algorithm doesn't give a consistent estimator of the mixing matrix. In this paper, following the semiparametric statistical approach to the noise-free ICA model by Amari and Cardoso(1997), we propose a method of estimating the mixing matrix consistently even if the additive noise exists. The proposed algorithm consists of two stages: First find the factor subspace by means of factor analysis, and then determine the directions of independent components based on an estimating function in this semiparametric model.

*Department of Mathematical Engineering and Information Physics, Faculty of Engineering, University of Tokyo, Hongo 7-3-1, Bunkyo-ku, Tokyo 113-8656, Japan.

†Lab for Information Synthesis, Brain-Style Information System Group, RIKEN Brain Science Institute, Hirosawa 2-1, Wako, Saitama 351-0198, Japan.

1 Introduction

Independent component analysis (ICA) uses a statistical model where observed data are expressed as a linear combination of statistically independent random variables. Since Jutten & Herault(1991) published the first algorithm for the blind source separation, a lot of new ideas and algorithms have been proposed by researchers on signal processing and neural networks. These algorithms were rationalized theoretically by Amari and Cardoso(1997) in the framework of semiparametric statistical models (Bickel et al.,1993).

Many papers on ICA treat the following simplest case. Let $\boldsymbol{s} = (s_1, \ldots, s_n)^{\mathrm{T}}$ be a vector of n source signals whose components are mutually stochastically independent. Let

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

be an observed mixed signal vector, where we assume A is an unknown $n \times n$ invertible matrix, and the probability distribution $\kappa(\boldsymbol{s})$ of $\boldsymbol{s}$ is unknown except that the n source signals are mutually independent. When a sequence of observed signals $\{\boldsymbol{x}(1), \ldots, \boldsymbol{x}(T)\}$ is given, ordinal ICA algorithms are designed for estimating the mixing matrix A within this model and recovering the original signals by

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

where W corresponds to the inverse of A.

However in realistic situations such as in MEG data analysis, it is not rare that certain measurement noises are added after mixing source signals. Ordinal ICA algorithms perform worse as the noise level increases and it is very difficult to derive meaningful outcomes. Therefore investigation of the ICA model with additive noise

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

becomes one of the most important topics now. There exist several papers which handled ICA models with measurement noise. Both the maximum joint likelihood method (Hyvärinen,1999) and the maximum marginal likelihood method (Attias,1999) work on the condition that the distributions of the source signals are known. The bias removal learning algorithm proposed by Chichocki et al.(1998) assumes that the amplitude of noise is small. The JADE algorithm (Cardoso et al.,1993) and the Fast ICA algorithm with Gaussian moments (Hyvärinen,1999) are semiparametric methods which give an estimate of the mixing matrix without knowledge of the

unobserved source distributions. Although the original algorithms assume that the noise variance is known or sphere, they can be used with quasi-whitening by factor analysis even for the noisy ICA model where the noise variance is unknown.

In this paper, we explain a semiparametric approach for the noisy ICA model where desired estimators are described in terms of estimating functions. Then we propose a noisy ICA algorithm which consists of two stages: First find the factor subspace by means of factor analysis, and then determine the directions of independent components based on an estimating function in this semiparametric model.

This paper is organized as follows. In section 2, some notions and definitions such as semiparametric models, estimating functions, and multivariate Hermite polynomials are prepared. Then theorems about estimating functions in the noisy ICA model are summarized in section 3. Because it is not the aim of this paper to discuss the estimating functions generally and thoroughly, further detail will be presented in a forthcoming paper. Moreover those who are interested in the proposed algorithm can skip this section. After factor analysis are explained briefly in section 4, Section 5 is devoted for explaining a noisy ICA algorithm based on an estimating function and the reason why it gives a consistent estimator in this semiparametric situation. In section 6, we investigate performance of the presented algorithm via numerical experiments. Relationships to other noisy ICA algorithms are discussed in the final section.

2 Mathematical Preliminaries

2.1 Semiparametric Models and Estimating Functions

Let us consider a sequence of n-dimensional random vectors generated from an ICA model with additive noise

$$\boldsymbol{x}(t) = A\boldsymbol{s}(t) + \boldsymbol{\xi}(t), \qquad t = 1, \ldots, T. \tag{2.1}$$

where A is an unknown $n \times m$ matrix, and the vector $\boldsymbol{\xi}$ is measurement noise. And $\boldsymbol{s}(t) = (\, s_1(t), \ldots, s_m(t)\,)^{\mathrm{T}}$ is a sequence of m unobserved source signals which are mutually independent. Although there are a lot of papers considering time-dependent sources in the noise-free ICA model, we assume for simplicity that the source signal vectors $\boldsymbol{s}(t)$ are independent and identically distributed in time, and sometimes the index of time is omitted in the

following. The joint probability density function $\kappa(\boldsymbol{s})$ of $\boldsymbol{s}$ is then factorized as

$$\kappa(\boldsymbol{s}) = \prod_{i=1}^{n} \kappa_i(s_i) \tag{2.2}$$

where $\kappa_i(s_i)$ is the density function of the i-th signal s_i. We consider the semiparametric situation that the function forms of $\kappa_1, \ldots, \kappa_m$ are unknown except for

$$\mathrm{E}_{\kappa_i}[s_i] = 0, \qquad i = 1, \ldots, m. \tag{2.3}$$

With respect to the additive noise, we assume that the random vector $\boldsymbol{\xi}$ is independent from the sources $\boldsymbol{s}$ and subjects to a multivariate normal distribution $N(\mathbf{0}, \Sigma)$ where $\Sigma = \mathrm{diag}(\sigma_i^2)$.

The density function of observed data $\boldsymbol{x}$ can be expressed as

$$p(\boldsymbol{x}; A, \Sigma, \kappa) = \int p(\boldsymbol{x}|\boldsymbol{s}; A, \Sigma)\, \kappa(\boldsymbol{s}) \mathrm{d}\boldsymbol{s}, \tag{2.4}$$

$$p(\boldsymbol{x}|\boldsymbol{s}; A, \Sigma) = \frac{1}{(2\pi)^{n/2}|\Sigma|^{1/2}} \exp\left[-\frac{1}{2}(\boldsymbol{x} - A\boldsymbol{s})^{\mathrm{T}}\Sigma^{-1}(\boldsymbol{x} - A\boldsymbol{s})\right]. \tag{2.5}$$

The equation (2.4) is a mixture of the normal distributions (2.5) by the unknown function κ. This is a semiparametric model, where the mixing matrix A and the noise variance Σ are parameters of interest and the density κ is a nuisance parameter in a function space. We remark that this parameterization is redundant as follows. Let $D = \mathrm{diag}(d_1, \ldots, d_m)$ be any diagonal matrix and put $\widetilde{A} = AD^{-1}$, $\widetilde{\boldsymbol{s}} = D\boldsymbol{s}$, $\widetilde{\kappa}_i(\widetilde{s}_i) = \kappa_i(\widetilde{s}_i/d_i)/d_i$, then the densities correspond to these two parameters are the same. Therefore, it is necessary to impose certain appropriate restrictions on the mixing matrix A or the scale of the source signals $\boldsymbol{s}$. For instance, we can add restrictions

$$\mathrm{E}_{\kappa_i}\left[s_i^2\right] = 1, \qquad i = 1, \ldots, m. \tag{2.6}$$

Estimating functions introduced by Godambe(1976) provide a general framework for discussing semiparametric estimators. Let us consider general semiparametric models in the form of $\{\, p(x; \boldsymbol{\theta}, \kappa) \,\}$, where $\boldsymbol{\theta}$ is the r-dimensional parameter to be estimated and κ is a nuisance parameter which belongs to an infinite dimensional or a function space. A r-dimensional vector function $\boldsymbol{f}(x, \boldsymbol{\theta})$ that does not depend on κ is called an estimating function when the following conditions are satisfied for all $\boldsymbol{\theta}$ and all κ.

$$\mathrm{E}_{\theta,\kappa}[\boldsymbol{f}(x, \boldsymbol{\theta})] = \mathbf{0} \tag{2.7}$$

$$\det|K| \neq 0, \qquad \text{where} \quad K = \mathrm{E}_{\theta,\kappa}\left[\frac{\partial}{\partial \boldsymbol{\theta}} \boldsymbol{f}(x,\boldsymbol{\theta})\right] \tag{2.8}$$

$$\mathrm{E}_{\theta,\kappa}\left[\boldsymbol{f}(x,\boldsymbol{\theta})\,\boldsymbol{f}^{\mathrm{T}}(x,\boldsymbol{\theta})\right] < \infty \tag{2.9}$$

If such an estimating function $\boldsymbol{f}(x,\boldsymbol{\theta})$ exists, we can obtain an M-estimator from given i.i.d. data $x_1,\ldots,x_n$ by solving the estimating equation

$$\sum_{i=1}^{n} \boldsymbol{f}(x_i,\widehat{\boldsymbol{\theta}}) = \mathbf{0}. \tag{2.10}$$

It can be shown under some additional regularity conditions that the M-estimator is consistent whatever κ is. Its covariance matrix is given asymptotically by

$$V\left[\widehat{\boldsymbol{\theta}}\right] = \frac{1}{n} K^{-1} \mathrm{E}_{\theta,\kappa}\left[\boldsymbol{f}\,\boldsymbol{f}^{\mathrm{T}}\right]\left(K^{-1}\right)^{\mathrm{T}}. \tag{2.11}$$

2.2 Decomposition into Regressor and Residual

The noisy ICA model (2.1) at a time has some analogy with the general regression model, where A is a matrix of explanatory variables, $\boldsymbol{s}$ is regression coefficients and Σ is assumed to be known. We will prepare and use some notions and terminologies of the regression analysis in order to investigate the estimating functions for the noisy ICA model.

Given the source signals the conditional density (2.5) can be decomposed as

$$p(\boldsymbol{x}|\boldsymbol{s};A,\Sigma) = \frac{1}{(2\pi)^{n/2}|\Sigma|^{1/2}} \exp\left[-\frac{1}{2}(\boldsymbol{y}-\boldsymbol{s})^{\mathrm{T}} V^{-1}(\boldsymbol{y}-\boldsymbol{s}) - \frac{1}{2}\boldsymbol{z}^{\mathrm{T}}\Sigma^{-1}\boldsymbol{z}\right] \tag{2.12}$$

where we define

$$\boldsymbol{y}(\boldsymbol{x};A,\Sigma) \equiv (A^{\mathrm{T}}\Sigma^{-1}A)^{-1}A^{\mathrm{T}}\Sigma^{-1}\boldsymbol{x} \in \boldsymbol{R}^m, \tag{2.13}$$

$$\boldsymbol{z}(\boldsymbol{x};A,\Sigma) \equiv \left\{I - A(A^{\mathrm{T}}\Sigma^{-1}A)^{-1}A^{\mathrm{T}}\Sigma^{-1}\right\}\boldsymbol{x} \in \boldsymbol{R}^n, \tag{2.14}$$

and $V = (v_{ij}) \equiv (A^{\mathrm{T}}\Sigma^{-1}A)^{-1}$. By the decomposition theorem $\boldsymbol{y}$ can be regarded as a sufficient statistics for $\boldsymbol{s}$ under fixed A and Σ, while $\boldsymbol{z}$ is an ancillary statistics (remind that estimating functions are functions of the parameters A and Σ). Furthermore $\boldsymbol{y}$ and $\boldsymbol{z}$ are independent from each other. This can be shown by the orthogonal property explained after the

equation (2.18). Because $W \equiv (A^{\mathrm{T}}\Sigma^{-1}A)^{-1}A^{\mathrm{T}}\Sigma^{-1}$ satisfies $WA = I_m$ (a generalized inverse matrix A), $\boldsymbol{y}$ can be expressed as

$$\boldsymbol{y} \;=\; \boldsymbol{s} + \boldsymbol{\zeta}, \qquad \boldsymbol{\zeta} \equiv W\boldsymbol{\xi} \sim N(\boldsymbol{0}, V). \tag{2.15}$$

Therefore, $\boldsymbol{y}$ is subject to $N(\boldsymbol{s}, V)$ for given $\boldsymbol{s}$. On the other hand, $\boldsymbol{z}$ does not depend on $\boldsymbol{s}$ and distributes with an $(n-m)$-dimensional degenerated normal distribution ($A^{\mathrm{T}}\Sigma^{-1}\boldsymbol{z} = 0$).

$$\boldsymbol{z} \;\sim\; N(\boldsymbol{0}, \Gamma), \qquad \Gamma \;=\; (\gamma_{ij}) \;=\; \Sigma - AVA^{\mathrm{T}} \tag{2.16}$$

The marginal distribution (2.4) which is integration of (2.5) with the density κ of $\boldsymbol{s}$ can be decomposed as

$$\begin{aligned} p(\boldsymbol{x}; A, \Sigma, \kappa) \;&=\; \frac{1}{(2\pi)^{n/2}|\Sigma|^{1/2}} \exp\left[-\frac{1}{2}\boldsymbol{z}^{\mathrm{T}}\Sigma^{-1}\boldsymbol{z}\right] \\ &\quad \times \int \exp\left[-\frac{1}{2}(\boldsymbol{y}-\boldsymbol{s})^{\mathrm{T}}V^{-1}(\boldsymbol{y}-\boldsymbol{s})\right] \kappa(\boldsymbol{s})\mathrm{d}\boldsymbol{s}. \end{aligned} \tag{2.17}$$

In the same way $\boldsymbol{y}$ can be regarded as sufficient statistics of the nuisance function κ, $\boldsymbol{z}$ as ancillary statistics, and they are independent. We remark that these properties hold without mutually independence of the source signals s_i.

The data $\boldsymbol{x}$ can be decomposed into the projection to the subspace spanned by the column vectors of A and its orthogonal complement

$$\boldsymbol{x} \;=\; A\boldsymbol{y} + \boldsymbol{z} \tag{2.18}$$

where orthogonality is defined by the metric Σ^{-1} (inverse of noise variance) and expressed as $(A\boldsymbol{y})^{\mathrm{T}}\Sigma^{-1}\boldsymbol{z} = 0$. Using the terminologies of regression analysis, $\boldsymbol{y}$ corresponds to the weighted least square or minimum variance unbiased estimator of $\boldsymbol{s}$ and $\boldsymbol{z}$ corresponds to the residual.

2.3 Multivariate Hermite Polynomials

Multivariate Hermite polynomials will be used to specify the estimating functions in the noisy ICA model. They are extensions of well known Hermite polynomials to multivariate normal distributions. We express the density function of the m-variate normal distribution with mean $\boldsymbol{0}$ and covariance matrix $V = (v_{ij})$ as

$$\phi(\boldsymbol{y}; V) = \frac{1}{(2\pi)^{m/2}|V|^{1/2}} \exp\left(-\frac{1}{2}\boldsymbol{y}^{\mathrm{T}}V^{-1}\boldsymbol{y}\right). \tag{2.19}$$

Let us put two differential operators, $D_i \equiv \partial/\partial y_i$ and $\widetilde{D}_i \equiv \partial/\partial \widetilde{y}_i$, where $\widetilde{\boldsymbol{y}} = V^{-1}\boldsymbol{y}$. Then two types of multivariate Hermite polynomials are defined as follows.

Definition 1 Covariant Hermite Polynomials H and Contravariant Hermite Polynomials $\widetilde{H}$

$$H_{r_1 \dots r_m}(\boldsymbol{y}; V) = (-D_1)^{r_1} \cdots (-D_m)^{r_m} \phi(\boldsymbol{y}; V)/\phi(\boldsymbol{y}; V) \quad (2.20)$$

$$\widetilde{H}_{r_1 \dots r_m}(\boldsymbol{y}; V) = (-\widetilde{D}_1)^{r_1} \cdots (-\widetilde{D}_m)^{r_m} \phi(\boldsymbol{y}; V)/\phi(\boldsymbol{y}; V) \quad (2.21)$$

Here the subscripts $r_1, \dots, r_m$ of these polynomials mean the power exponents.

In this paper, we express the power exponents as $\boldsymbol{r} = (r_1, \dots, r_m)$ in the vector form and adopt the simplified notations

$$\begin{aligned} \boldsymbol{r}! &\equiv r_1! \cdots r_m!, \\ \boldsymbol{t}^{\boldsymbol{r}} &\equiv t_1^{r_1} \cdots t_m^{r_m}. \end{aligned}$$

The generating functions of multivariate Hermite polynomials are

$$\sum_{\boldsymbol{r}} \frac{\boldsymbol{t}^{\boldsymbol{r}}}{\boldsymbol{r}!} H_{\boldsymbol{r}}(\boldsymbol{y}; V) = \frac{\phi(\boldsymbol{y} - \boldsymbol{t})}{\phi(\boldsymbol{y})} = \exp\left\{ \boldsymbol{t}^{\mathrm{T}} V^{-1} \boldsymbol{y} - \frac{1}{2} \boldsymbol{t}^{\mathrm{T}} V^{-1} \boldsymbol{t} \right\}, \quad (2.22)$$

$$\sum_{\boldsymbol{r}} \frac{\boldsymbol{t}^{\boldsymbol{r}}}{\boldsymbol{r}!} \widetilde{H}_{\boldsymbol{r}}(\boldsymbol{y}; V) = \exp\left\{ \boldsymbol{y}^{\mathrm{T}} \boldsymbol{t} - \frac{1}{2} \boldsymbol{t}^{\mathrm{T}} V \boldsymbol{t} \right\}. \quad (2.23)$$

These are used for proofs of the next lemma and other properties.

Lemma 1 $\{H_{\boldsymbol{r}}\}$ and $\{\widetilde{H}_{\widetilde{\boldsymbol{r}}}\}$ form mutually orthogonal polynomials.

$$\int H_{\boldsymbol{r}}(\boldsymbol{y}; V) \widetilde{H}_{\widetilde{\boldsymbol{r}}}(\boldsymbol{y}; V) \phi(\boldsymbol{y}; V) \mathrm{d}\boldsymbol{y} = \begin{cases} 0 & \text{if } \boldsymbol{r} \neq \widetilde{\boldsymbol{r}} \\ \boldsymbol{r}! & \text{if } \boldsymbol{r} = \widetilde{\boldsymbol{r}} \end{cases}. \quad (2.24)$$

3 Estimating Functions in the Noisy ICA Model

In this section we will explain estimating functions in the noisy ICA model briefly. As the most important property is unbiasedness (2.7) for any nuisance parameter κ, the most part in this section are used in order to characterize scalar functions which satisfy the same unbiasedness as (2.7). These unbiased functions are candidates for components of estimating functions.

Indeed, if we can collect $n \times m + n$ unbiased functions (the same number as the parameter to be estimated) which satisfy the other conditions (2.8) and (2.9) in all, then the set of the functions becomes an estimating function.

Let us express model assumptions concretely. On the density function κ of the source $\boldsymbol{s}$, the following constraints must be imposed.

$$\text{condition I:} \qquad \int \kappa_i(s_i)\mathrm{d}s_i \;=\; 1, \qquad i = 1, \ldots, m \tag{3.1}$$

$$\int s_i \kappa_i(s_i)\mathrm{d}s_i \;=\; 0, \qquad i = 1, \ldots, m \tag{3.2}$$

The former is normalization of density, while the latter comes from the condition (2.3). If we restrict the mixing matrix A to cancel the redundancy of the noisy ICA model, we should discuss under the condition I. On the other hand, in case we restrict scales of the source signals, constraints on variances (2.6) are employed in addition to the necessary constraints,

$$\text{condition II:} \qquad \text{condition I} + \int s_i^2 \kappa_i(s_i)\mathrm{d}s_i = 1, \qquad i = 1, \ldots, m. \tag{3.3}$$

Let us define $F^{\perp}_{A,\Sigma}$ as the set of functions whose conditional expectation for given $\boldsymbol{y} = \boldsymbol{y}(\boldsymbol{x}; A, \Sigma)$ are zero.

$$F^{\perp}_{A,\Sigma} \equiv \{\boldsymbol{f}(\boldsymbol{x});\ \mathrm{E}_{A,\Sigma}[\boldsymbol{f}(\boldsymbol{x})|\boldsymbol{y}] = 0\} \tag{3.4}$$

We remark that the conditional distribution of $\boldsymbol{x}$ for given $\boldsymbol{y}$ does not depend on the nuisance parameter κ because of (2.17). It can be shown that the unbiased scalar functions consist of elements of $F^{\perp}_{A,\Sigma}$, the unbiased functions of $\boldsymbol{y} = \boldsymbol{y}(\boldsymbol{x}; A, \Sigma)$ and their linear combinations. Roughly speaking, the former contribute to estimating the signal subspace, while the latter will do for pursuit of the independent component directions.

Theorem 1 Under the condition I, the set of the scalar unbiased functions are expressed as

$$F^{\perp}_{A,\Sigma} \oplus \{f(\boldsymbol{y}; A, \Sigma)\,;\ \text{satisfy (3.6)}\} \tag{3.5}$$

$$\mathrm{E}_{A,\Sigma}\left[f(\boldsymbol{y}; A, \Sigma)|\boldsymbol{s}\right] \;=\; \sum_{i=1}^{m} s_i\, \nu_i(\boldsymbol{s}_{-i}) \tag{3.6}$$

where $\oplus$ means the direct sum and ν_i is an arbitrary function of $\boldsymbol{s}_{-i} \equiv (s_1, \ldots, s_{i-1}, s_{i+1}, \ldots, s_m)$.

Proof Any scalar function $f(\boldsymbol{x}; A, \Sigma)$ can be decomposed as

$$f(\boldsymbol{x}; A, \Sigma) = \mathrm{E}_{A,\Sigma}[f(\boldsymbol{x}; A, \Sigma)|\boldsymbol{y}] + \{\, f(\boldsymbol{x}; A, \Sigma) - \mathrm{E}_{A,\Sigma}[f(\boldsymbol{x}; A, \Sigma)|\boldsymbol{y}] \,\} \quad (3.7)$$

where the first term of the right hand side of (3.7) is a function of $\boldsymbol{y} = \boldsymbol{y}(\boldsymbol{x}; A, \Sigma)$ and the second term is an element of $F^{\perp}_{A,\Sigma}$, that is, its conditional expectation vanishes. The uniqueness of such decomposition is obvious. For any element $\widetilde{f}(\boldsymbol{x}; A, \Sigma)$ of $F^{\perp}_{A,\Sigma}$,

$$\mathrm{E}_{A,\Sigma,\tilde{\kappa}} \left[\, \widetilde{f}(\boldsymbol{x}; A, \Sigma) \,\right] = \mathrm{E}_{A,\Sigma,\tilde{\kappa}} \left[\, \mathrm{E}_{A,\Sigma} \left[\widetilde{f}(\boldsymbol{x}; A, \Sigma) \middle| \boldsymbol{y} \right] \,\right] \;=\; 0$$

holds for all $\tilde{\kappa}$. This indicates that $F^{\perp}_{A,\Sigma}$ is included in the set of the scalar unbiased functions.

With respect to functions of $\boldsymbol{y}$, a function $f(\boldsymbol{y}; A, \Sigma)$ whose conditional expectation for given $\boldsymbol{s}$ can be expressed as (3.6) is unbiased for all κ, because the components s_i are mutually independent and have zero mean. Proof of the converse is omitted here. □

Remark Under the condition II, (3.6) is replaced by

$$\mathrm{E}_{A,\Sigma}\,[f(\boldsymbol{y}; A, \Sigma)|\boldsymbol{s}] \;=\; \sum_{i=1}^{m} s_i\, \nu_i(\boldsymbol{s}_{-i}) + \sum_{i=1}^{m} (s_i^2 - 1)\, \lambda_i(\boldsymbol{s}_{-i}) \qquad (3.8)$$

It is difficult to determine general form of functions of $\boldsymbol{y}$ whose conditional expectation for given $\boldsymbol{s}$ can be expressed as (3.6). However we can describe polynomials concretely which have this property. For simplicity, we assume that any moment of the source signals s_i exist in this section.

Theorem 2 Under the condition I, the set of the unbiased polynomials are expressed as

$$\begin{aligned} I^{Y}_{A,\Sigma} \;=\; & \mathrm{span}\{\widetilde{H}_{\boldsymbol{r}}(\boldsymbol{y}; V);\ \boldsymbol{r} = (r_1, \ldots, r_m), \\ & \qquad \text{at least one of the indices } \boldsymbol{r} \text{ is equal to } 1\} \end{aligned} \qquad (3.9)$$

Proof Polynomials of $\boldsymbol{y}$ are spanned by the covariant Hermite polynomials $\{H_{\boldsymbol{r}}(\boldsymbol{y}; V)\}$ or the contravariant Hermite polynomials $\{\widetilde{H}_{\boldsymbol{r}}(\boldsymbol{y}; V)\}$. From the generating function of the covariant Hermite polynomials, the following equation holds,

$$\frac{\phi(\boldsymbol{y} - \boldsymbol{s}; V)}{\phi(\boldsymbol{y}; V)} \;=\; \sum_{\boldsymbol{q}} \frac{\boldsymbol{s}^{\boldsymbol{q}}}{\boldsymbol{q}!}\, H_{\boldsymbol{q}}(\boldsymbol{y}; V).$$

Using this equation and lemma 1, the conditional expectation of a contravariant Hermite polynomial $\widetilde{H}_{\boldsymbol{r}}(\boldsymbol{y}; V)$ becomes

$$\begin{aligned} \mathrm{E}_{A,\Sigma}[\widetilde{H}_{\boldsymbol{r}}(\boldsymbol{y}; V)|\boldsymbol{s}] &= \int \widetilde{H}_{\boldsymbol{r}}(\boldsymbol{y}; V)\phi(\boldsymbol{y} - \boldsymbol{s}; V)\mathrm{d}\boldsymbol{y} \\ &= \int \widetilde{H}_{\boldsymbol{r}}(\boldsymbol{y}; V)\phi(\boldsymbol{y}; V) \sum_{\boldsymbol{q}} \frac{\boldsymbol{s}^{\boldsymbol{q}}}{\boldsymbol{q}!} H_{\boldsymbol{q}}(\boldsymbol{y}; V)\mathrm{d}\boldsymbol{y} \\ &= \boldsymbol{s}^{\boldsymbol{r}} = s_1^{r_1} \cdots s_m^{r_m}. \end{aligned}$$

Therefore if an Hermite polynomial is unbiased, one of its power indices $r_1, \ldots, r_m$ must be 1. It means that the set of unbiased polynomials of $\boldsymbol{y}$ is described as $I^Y_{A,\Sigma}$. □

Examples of unbiased polynomials are

$$y_j y_k - v_{jk}, \qquad j < k, \tag{3.10}$$
$$y_j^3 y_k - 3v_{jj} y_j y_k - 3v_{jk} y_j^2 + 3v_{jj} v_{jk}, \qquad j \neq k, \tag{3.11}$$
$$y_j^2 y_k y_l - v_{jj} y_k y_l - v_{kl} y_j^2 - 2v_{jk} y_j y_l - 2v_{jl} y_j y_k + v_{jj} v_{kl} + 2v_{jk} v_{jl}. \tag{3.12}$$

We call them $(1,1)$-type, $(3,1)$-type and $(2,1,1)$-type respectively.

Remark Under the condition II, for instance the following polynomials satisfy the unbiasedness.

$$\widetilde{H}_{B_j^2 \boldsymbol{0}}(\boldsymbol{y}; V) - 1 = y_j^2 - v_{jj} - 1, \quad \boldsymbol{0} = (0, \ldots, 0) \tag{3.13}$$
$$\widetilde{H}_{B_j^2 \boldsymbol{r}}(\boldsymbol{y}; V) - \widetilde{H}_{\boldsymbol{r}}(\boldsymbol{y}; V), \quad \boldsymbol{r} = (r_1, \ldots, r_{j-1}, 0, r_{j+1}, \ldots, r_m) \tag{3.14}$$

where B_j means an operator which add 1 to the j-th suffix r_j $(r_j \to r_j + 1)$.

We can construct unbiased functions other than polynomials. Some functions which are product of a polynomial and a Gaussian density become unbiased. These functions appeared in the Fast ICA with Gaussian moments (Hyvärinen,1999).

Theorem 3 For any $d > 0$ and any integer r,

$$\{y_j H_r(y_k; d) + v_{jk} H_{r+1}(y_k; d)\}\, \phi(y_k; d), \qquad j \neq k \tag{3.15}$$

is unbiased for all κ.

Finally we briefly deal with the assumption (2.8) of estimating functions. If (2.8) is not satisfied it may happen with non zero probability that we can not determine an estimator even locally and the M-estimator is not guaranteed to have the good asymptotic properties as described in section 2. A necessary condition of (2.8) is that any components $f_i(x, \boldsymbol{\theta})$ $(i = 1, \ldots, r)$ of an estimating function $\boldsymbol{f}(x, \boldsymbol{\theta})$ should satisfy

$$\exists j \quad \text{s.t.} \quad \mathrm{E}_{\theta,\kappa}\left[\, f_i(x, \boldsymbol{\theta})\, u_j(x; \boldsymbol{\theta}, \kappa)\,\right] \neq 0 \tag{3.16}$$

for all $i = 1, \ldots, r$, where $u_j \equiv \partial \log p(x; \boldsymbol{\theta}, \kappa)/\partial \theta_j$ is the the score function of θ_j (Amari and Kawanabe,1997). In information geometry (Amari,1985), $\mathrm{E}[\, f_i\, u_j\,]$ in the equation the equation (3.16) is employed as the inner product of these functions. Therefore, to put the necessary condition differently, any components of an estimating function are not orthogonal to the score functions of the parameter to be estimated. Roughly speaking, only score functions carry the sufficient information to estimate the parameters, therefore if f_i is orthogonal to all the scores, it means f_i does not include any information. In the noisy ICA model, the score functions of the parameter A and Σ are expressed as

$$\begin{aligned} U_A &= (\, u_{a_{ij}}\,) \\ &= \frac{\partial}{\partial A} \log p \;=\; \Sigma^{-1} \boldsymbol{x}\, \mathrm{E}[\boldsymbol{s}^{\mathrm{T}} | \boldsymbol{y}] - \Sigma^{-1} A\, \mathrm{E}[\boldsymbol{s}\boldsymbol{s}^{\mathrm{T}} | \boldsymbol{y}], \qquad (3.17) \\ u_{\sigma_i^2} &= \frac{\partial}{\partial \sigma_i^2} \log p \\ &= -\frac{1}{2\sigma_i^2} + \frac{x_i^2}{2\sigma_i^4} - \frac{x_i}{\sigma_i^4} \sum_j a_{ij} \mathrm{E}[s_j | \boldsymbol{y}] + \frac{1}{2\sigma_i^4} \sum_{j,k} a_{ij} a_{ik} \mathrm{E}[s_j s_k | \boldsymbol{y}]. \qquad (3.18) \end{aligned}$$

We characterize here which kind of polynomials in the set $I^Y_{A,\kappa}$ are not orthogonal to all these score functions and available for components of estimating functions.

We can obtain the inner product of contravariant Hermite polynomials and the score functions.

Theorem 4 The inner product of a contravariant Hermite polynomials $\widetilde{H}_{\boldsymbol{r}}$ and the score functions $u_{a_{ij}}$, $u_{\sigma_i^2}$ can be expressed as

$$\mathrm{E}\left[\widetilde{H}_{\boldsymbol{r}}\, u_{a_{ij}}\right] \;=\; \sum_h r_h w_{hi}\, \boldsymbol{\mu}^{(B_h^{-1} B_j \boldsymbol{r})}, \tag{3.19}$$

$$\mathrm{E}\left[\widetilde{H}_{\boldsymbol{r}} u_{\sigma_i^2}\right] = \frac{1}{2}\left\{\sum_h r_h(r_h-1)w_{hi}^2\,\boldsymbol{\mu}^{(B_h^{-2}\boldsymbol{r})} + \sum_{h\neq\tilde{h}} r_h r_{\tilde{h}} w_{hi} w_{\tilde{h}i}\,\boldsymbol{\mu}^{(B_h^{-1}B_{\tilde{h}}^{-1}\boldsymbol{r})}\right\}, \tag{3.20}$$

where $W = VA^{\mathrm{T}}\Sigma^{-1} = (A^{\mathrm{T}}\Sigma^{-1}A)^{-1}A^{\mathrm{T}}\Sigma^{-1}$ is a generalized inverse matrix of A, μ's denote moments of the source signals.

$$\boldsymbol{\mu}^{(\boldsymbol{r})} = \mu_1^{(r_1)}\cdots\mu_m^{(r_m)}, \qquad \mu_i^{(r_i)} = \mathrm{E}\left[s_i^{r_i}\right]$$

B_j means an operator which replaces r_j with r_j+1 as defined before, and B_j^{-1} is its inverse operator which replace r_j with r_j-1 (define $\mu_i^{(r_i)}=0$ for $r_i<0$).

From this theorem and $\mu_i^{(1)}=0$, next corollary follows.

Corollary 5

1. If $\boldsymbol{r}=(1,0,\ldots,0)$ (or its permutations), $\widetilde{H}_{\boldsymbol{r}}$ is orthogonal to all score functions.
2. If $\boldsymbol{r}=(1,r_2,\ldots,r_m)$, $r_j=0$ or 2, (or its permutations), $\widetilde{H}_{\boldsymbol{r}}$ is orthogonal to $u_{\sigma_i^2}$.
3. If the number of 1 included in $\boldsymbol{r}$ is greater than three, $\widetilde{H}_{\boldsymbol{r}}$ is orthogonal to all score functions.

Furthermore, we assume that $s_1,\ldots,s_m$ have symmetric distributions around the origin. Then for any odd integer j, $\mu_i^{(j)}=0$ holds.

Corollary 6 Suppose $s_1,\ldots,s_m$ are symmetrically distributed around the origin. Score functions and $\widetilde{H}_{\boldsymbol{r}}$ are orthogonal except for

$$\boldsymbol{r}=(1,r_2,r_3,\ldots,r_m), \qquad r_2:\text{odd},\ \ r_3,\ldots,r_m:\text{even}. \tag{3.21}$$

or its permutation. If $\boldsymbol{r}$ is (3.21), the inner products of score functions and $\widetilde{H}_{\boldsymbol{r}}$ are expressed as

$$\mathrm{E}\left[\widetilde{H}_{\boldsymbol{r}} u_{a_{i1}}\right] = r_2 w_{2i}\,\mu_1^{(2)}\mu_2^{(r_2-1)}\mu_3^{(r_3)}\cdots\mu_m^{(r_m)}, \tag{3.22}$$

$$\mathrm{E}\left[\widetilde{H}_{\boldsymbol{r}} u_{a_{i2}}\right] = w_{1i}\,\mu_2^{(r_2+1)}\mu_3^{(r_3)}\cdots\mu_m^{(r_m)}, \tag{3.23}$$

$$\mathrm{E}\left[\widetilde{H}_{\boldsymbol{r}} u_{a_{ij}}\right] = 0, \qquad (j \geq 3), \tag{3.24}$$

$$\mathrm{E}\left[\widetilde{H}_{\boldsymbol{r}} u_{\sigma_i^2}\right] = r_2 w_{1i} w_{2i}\,\mu_2^{(r_2-1)}\mu_3^{(r_3)}\cdots\mu_m^{(r_m)}. \tag{3.25}$$

The examples (3.10) $\sim$ (3.12) are the simplest polynomials that are not orthogonal to all the score functions in the situation of Corollary 6. Examples of the polynomials that cannot be used as the estimating function are $(2,1)$-type Hermite polynomials. The inner products of score functions and $(2,1)$-type Hermite polynomials

$$y_j^2 y_k - v_{jj} y_k - 2v_{jk} y_j, \qquad j < k, \tag{3.26}$$

are zero except for

$$\mathrm{E}\left[\widetilde{H}_{\boldsymbol{r}} u_{a_{ij}}\right] = w_{ki}\,\mu_j^{(3)} \tag{3.27}$$

because $\mu_j^{(r_j-1)} = \mu_j^{(1)} = 0$. When $s_1, \ldots, s_m$ are symmetrically distributed around the origin, they are orthogonal to all score functions because $\mu_j^{(3)} = 0$. Therefore, these polynomials do not contain sufficient information in order to determine the parameter.

4 A Noisy ICA Algorithm Based on Estimating Functions

Now we propose an algorithm which is a combination of factor analysis and an estimating function method for the noisy ICA model. The latter part can be regarded as a modification of Jutten & Herault's procedure with the concept of the estimating function.

1. Find the factor subspace by using factor analysis such as the unweighted least squares method (ULS) or the maximum likelihood method (ML). Let $\left(A^{(0)}, \Sigma^{(0)}\right)$ be the solution derived by factor analysis.

2. Calculate initial estimates of source signals and their conditional covariances.

$$\boldsymbol{y}^{(0)}(t) = \left\{(A^{(0)})^{\mathrm{T}}(\Sigma^{(0)})^{-1}A^{(0)}\right\}^{-1}(A^{(0)})^{\mathrm{T}}(\Sigma^{(0)})^{-1}\boldsymbol{x}(t) \tag{4.1}$$

$$V^{(0)} = \left\{(A^{(0)})^{\mathrm{T}}(\Sigma^{(0)})^{-1}A^{(0)}\right\}^{-1} \tag{4.2}$$

3. Let Q be an $m \times m$ transformation matrix to the direction of the independent components, that is, the mixing matrix is expressed as $A = A^{(0)}Q^{-1}$ and

$$\boldsymbol{y}(t) = \left\{A^{\mathrm{T}}(\Sigma^{(0)})^{-1}A\right\}^{-1} A^{\mathrm{T}}(\Sigma^{(0)})^{-1}\boldsymbol{x}(t) = Q\boldsymbol{y}^{(0)}(t) \tag{4.3}$$

$$V = \left\{A^{\mathrm{T}}(\Sigma^{(0)})^{-1}A\right\}^{-1} = QV^{(0)}Q^{\mathrm{T}} \tag{4.4}$$

The matrix Q can be determined by the following estimating equations $(i \neq j)$

$$\sum_{t=0}^{T}\left\{y_i^3(t)y_j(t) - 3v_{ij}y_i^2(t) - 3v_{ii}y_i(t)y_j(t) + 3v_{ii}v_{ij}\right\} = 0 \tag{4.5}$$

$$\Leftrightarrow \sum_{b,c,d,e} q_{ib}q_{ic}q_{id}q_{ie}\sum_{t=0}^{T}\left\{y_b^{(0)}(t)y_c^{(0)}(t)y_d^{(0)}(t)y_e^{(0)}(t) - 3v_{be}^{(0)}y_c^{(0)}(t)y_d^{(0)}(t) - 3v_{bc}^{(0)}y_d^{(0)}(t)y_e^{(0)}(t) + 3v_{bc}^{(0)}v_{de}^{(0)}\right\} = 0 \tag{4.6}$$

with appropriate additional constraints such as

$$\sum_{t=1}^{T}\left\{y_i^2(t) - v_{ii} - 1\right\} = 0, \qquad i = 1, \ldots, m, \tag{4.7}$$

$$\sum_{j=1}^{m} q_{ij}^2 = 1, \qquad i = 1, \ldots, m. \tag{4.8}$$

We will explain why this algorithm gives a consistent estimator regardless of the density κ of the source signals $\boldsymbol{s}$ in terms of estimating functions.

4.1 Factor Analysis and Prewhitening

In factor analysis, the model is also expressed as

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

where A is called the factor loading matrix and $\boldsymbol{s}$ the factors. However, distributional assumption on the factors $\boldsymbol{s}$ are different. In case of factor analysis, the factors $\boldsymbol{s}$ are often supposed to have a normal distribution $N(\boldsymbol{0}, I_m)$ and information of only second order moments are employed. So we can only determine the factor subspace which is spanned by the column

vectors of the factor loading matrix A, and a base of this subspace is selected by another criterion such as othomax and oblimin in order to obtain explainable factors. On the other hand, we assume that at least $m-1$ signals is subject to non-normal distributions in the noisy ICA model. Under this assumption, the mixing matrix A is identified up to permutation and scaling of its columns. In order to avoid the redundancy of the model, we constrain the signals to have unit variances

$$\mathrm{E}\left[\boldsymbol{s}\boldsymbol{s}^{\mathrm{T}}\right] = I_m. \tag{4.9}$$

We remark that this is also imposed in orthogonal factor models.

Although the distributional assumptions are different from the noisy ICA model, factor analysis can be used to determine the factor subspace and the noise variance Σ. They can be estimated correctly, because the covariance matrix of $\boldsymbol{x}$ is also expressed as

$$\Psi = \mathrm{E}\left[\boldsymbol{x}\boldsymbol{x}^{\mathrm{T}}\right] = AA^{\mathrm{T}} + \Sigma, \tag{4.10}$$

in the noisy ICA model. We note that the covariance matrix Ψ is unaffected by multiplying any $m \times m$ orthogonal matrix P on the right of A. An additional constraint may be imposed to determine A uniquely. Let us express the sample covariance matrix as

$$S = \frac{1}{T}\sum_{t=1}^{T} \boldsymbol{x}(t)\boldsymbol{x}(t)^{\mathrm{T}}. \tag{4.11}$$

Several algorithms for estimating the factor loadings and the noise variances have been proposed so far. We employ two major algorithms, the unweighted least squares algorithm (ULS) and the maximum likelihood algorithm (ML).

In the ULS method, estimators are defined as the minimizer of the quadratic loss criterion

$$F_u(\Psi) = \mathrm{tr}(S-\Psi)^2 \tag{4.12}$$

Differentiating the criterion (4.12) by the parameters, we obtain the following estimating equations

$$(S-\Psi)A = 0 \tag{4.13}$$

$$\sigma_i^2 = s_{ii} - \sum_j a_{ij}^2 \tag{4.14}$$

Further, in order to determine the rotation uniquely for the present, we assume that

$$\mathrm{offdiag}(A^{\mathrm{T}}A) = 0 \tag{4.15}$$

and the diagonal elements are in decreasing order. Then the estimators can be derived by the Newton method where an eigen value decomposition is available for calculating increments because of (4.15).

The ML method gives estimators which minimize the following loss criterion

$$F_M(\Psi) = \mathrm{tr}(S\Psi^{-1}) - \log|S\Psi^{-1}| - n, \tag{4.16}$$

which is derived from the negative log likelihood under the assumption that $\boldsymbol{s}$ and $\boldsymbol{\xi}$ distribute normally. Differentiating the likelihood criterion (4.16) by the parameters, we obtain the following estimating equations

$$(S\Psi^{-1} - I_n)A = 0 \tag{4.17}$$

$$\sigma_i^2 = s_{ii} - \sum_j a_{ij}^2 \tag{4.18}$$

Further, in order to determine the rotation uniquely for the present, we assume that

$$\mathrm{offdiag}(A^{\mathrm{T}}\Sigma^{-1}A) = 0 \tag{4.19}$$

and the diagonal elements are in decreasing order. Then the estimators can be derived by the Newton method where an eigen value decomposition is available for calculating increments because of (4.19).

The $n \times n$ matrix valued function $\boldsymbol{x}\boldsymbol{x}^{\mathrm{T}} - \Psi = \boldsymbol{x}\boldsymbol{x}^{\mathrm{T}} - AA^T - \Sigma$ which appears in the estimating equations (4.13), (4.14), (4.17) and (4.18) is unbiased, because $\mathrm{E}[\boldsymbol{x}\boldsymbol{x}^{\mathrm{T}}] = AA^T + \Sigma$ holds for any distribution in the noisy ICA model. Therefore, if we can obtain an estimator of the factor subspace and the noise variance from these estimating equations, they will be consistent regardless of the density κ of the source signal $\boldsymbol{s}$. In reference to the previous section, let us decompose this function as

$$\begin{aligned} \boldsymbol{x}\boldsymbol{x}^{\mathrm{T}} - \Psi &= (A\boldsymbol{y} + \boldsymbol{z})(A\boldsymbol{y} + \boldsymbol{z})^{\mathrm{T}} - AA^{\mathrm{T}} - \Sigma \\ &= A(\boldsymbol{y}\boldsymbol{y}^{\mathrm{T}} - V - I_m)A^{\mathrm{T}} + A\boldsymbol{y}\boldsymbol{z}^{\mathrm{T}} + \boldsymbol{z}\boldsymbol{y}^{\mathrm{T}}A^{\mathrm{T}} + \boldsymbol{z}\boldsymbol{z}^{\mathrm{T}} - \Gamma \end{aligned} \tag{4.20}$$

where $\Gamma = \Sigma - AVA^{\mathrm{T}}$ is the degenerated covariance matrix of $\boldsymbol{z}$. Then, the components of the second, third and fourth term belong to $(F_{A,\Sigma,\kappa})^{\perp}$. The offdiagonal components of the first term $(\boldsymbol{y}\boldsymbol{y}^{\mathrm{T}} - V - I_m)$ are $(1,1)$-type contravariant Hermite polynomials, while the diagonal components become unbiased under the additional condition (3.3).

Finally, we discuss the results of factor analysis when the sample size T is large. We express the true parameters with superscript $*$ which indicate the generating model of the observed samples. Let us define a transformation $A^\dagger = A^* Q^*$ where for each procedure Q^* is an orthogonal matrix defined as follows. If the ULS method is used, Q^* consists of the eigen vectors of $(A^*)^{\mathrm{T}} A^*$, i.e.

$$(A^*)^{\mathrm{T}} A^* = Q^* \Delta (Q^*)^{\mathrm{T}}. \qquad (\Delta \text{ is a diagonal matrix}) \tag{4.21}$$

When the ML method is used, we define Q^* as the eigen vectors of $(A^*)^{\mathrm{T}} (\Sigma^*)^{-1} A^*$, i.e.

$$(A^*)^{\mathrm{T}} (\Sigma^*)^{-1} A^* = Q^* \Delta (Q^*)^{\mathrm{T}}. \qquad (\Delta \text{ is a diagonal matrix}) \tag{4.22}$$

The sample covariance matrix S converges to the covariance matrix $\Psi^* = A^*(A^*)^{\mathrm{T}} + \Sigma^* = A^\dagger (A^\dagger)^{\mathrm{T}} + \Sigma^*$ of the true model as the sample size T goes to infinity. Therefore it can be shown that the parameter $(A^\dagger, \Sigma^*)$ asymptotically minimizes the criterion F_u (or F_M) and $(A^\dagger)^{\mathrm{T}} A^\dagger = \Delta$ (or $(A^\dagger)^{\mathrm{T}} (\Sigma^*)^{-1} A^\dagger = \Delta$) becomes a diagonal matrix. Suppose that $\Psi^* = A^*(A^*)^{\mathrm{T}} + \Sigma^*$ is identifiable, that is, parameter (A, Σ) that satisfies $AA^{\mathrm{T}} + \Sigma = \Psi^*$ can be determined uniquely except for a rotation matrix.

Theorem 7 The estimator $(A^{(0)}, \Sigma^{(0)})$ derived by the ULS method or the ML method converges to $(A^\dagger, \Sigma^*)$ as T goes to infinity.

This theorem leads to the fact that $\boldsymbol{y}^{(0)}$ constructed by factor analysis can be regarded as quasi-whitened data of $\boldsymbol{x}$. When the sample size T is very large, $\boldsymbol{y}^{(0)}$ can be approximately expressed as

$$\boldsymbol{y}^{(0)} \doteq (Q^*)^{\mathrm{T}} \boldsymbol{s} + \boldsymbol{\zeta}^{(0)}$$

where Q^* is the orthogonal matrix and $\boldsymbol{\zeta}^{(0)}$ is a linear transformation of the additive noise $\boldsymbol{\xi}$. The variance of the signal part $(Q^*)^{\mathrm{T}} \boldsymbol{s}$ included in $\boldsymbol{y}^{(0)}$ becomes an identity matrix I_m. The Quasi-whitening is also used in noisy ICA algorithms proposed so far, though the noise variance Σ is assumed to be known or sphere. As described here, we can carry out the quasi-whitening by factor analysis even if Σ is unknown.

4.2 Pursuit of independent component directions

After estimating the factor subspace and the noise variance, we must determine the directions of the independent components. Here we show that we

can estimate the correct transformation matrix to the independent components by solving the estimating equations. Assuming that T is very large, we consider for simplicity that $(A^{(0)}, \Sigma^{(0)}) = (A^{\dagger}, \Sigma^{*})$ holds. Since

$$A^{*} \;=\; A^{\dagger}(Q^{*})^{\mathrm{T}} \;=\; A^{(0)}(Q^{*})^{\mathrm{T}}, \tag{4.23}$$

the correct transformation is expressed as $Q = Q^{*}$ or $Q = PDQ^{*}$ where D is any diagonal matrix and P is any permutation matrix (remember $A = A^{(0)}Q^{-1}$). Due to the additional conditions (4.7) or (4.8) we can derive Q^{*} except for indefiniteness of sign and order of independent components. We ignore this indefiniteness here. It is also possible to construct estimation procedures on the restricted set of orthogonal matrices.

Let us express the estimates of the source signals and their conditional covariances at the true parameters (A^{*}, Σ^{*}) as

$$\boldsymbol{y}^{*}(t) \;=\; \left\{(A^{*})^{\mathrm{T}}(\Sigma^{*})^{-1}A^{*}\right\}^{-1}(A^{*})^{\mathrm{T}}(\Sigma^{*})^{-1}\boldsymbol{x}(t), \tag{4.24}$$

$$V^{*} \;=\; \left\{(A^{*})^{\mathrm{T}}(\Sigma^{*})^{-1}A^{*}\right\}^{-1}. \tag{4.25}$$

Even if we know the true parameters (A^{*}, Σ^{*}), it is shown that the weighted least squares estimates $\boldsymbol{y}^{*}(t)$ or the best linear predictors

$$\widetilde{\boldsymbol{y}}^{*}(t) \;=\; \left\{I_m + (A^{*})^{\mathrm{T}}(\Sigma^{*})^{-1}A^{*}\right\}^{-1}(A^{*})^{\mathrm{T}}(\Sigma^{*})^{-1}\boldsymbol{x}(t) \tag{4.26}$$

$$\;=\; \left\{I_m + (A^{*})^{\mathrm{T}}(\Sigma^{*})^{-1}A^{*}\right\}^{-1}(A^{*})^{\mathrm{T}}(\Sigma^{*})^{-1}A^{*}\boldsymbol{y}^{*}(t) \tag{4.27}$$

are not mutually independent. In fact, the covariance matrices are

$$\mathrm{Var}\,[\,\boldsymbol{y}^{*}(t)\,] \;=\; I_m + V^{*} \tag{4.28}$$

$$\mathrm{Var}\,[\,\widetilde{\boldsymbol{y}}^{*}(t)\,] \;=\; (I_m + V^{*})^{-1} \tag{4.29}$$

which are not diagonal matrices in general. Therefore applying a noise-free ICA algorithm to the quasi-whitened data $\boldsymbol{y}^{(0)}$ leads to an inconsistent estimator of the mixing matrix because it forces dependent random variables to be mutually independent.

From the equation (4.23), the relationships

$$\boldsymbol{y}^{(0)}(t) \;=\; (Q^{*})^{\mathrm{T}}\boldsymbol{y}^{*}(t) \tag{4.30}$$

$$\boldsymbol{y}^{*}(t) \;=\; Q^{*}\boldsymbol{y}^{(0)}(t) \tag{4.31}$$

hold between the initial value and the estimates at the true parameters of the source signals. We remark that these expressions have some analogies to the

noise-free ICA model: $\boldsymbol{y}^{(0)}(t)$ corresponds to $\boldsymbol{x}(t)$, $\boldsymbol{y}^*(t)$ does to $\boldsymbol{s}(t)$, $(Q^*)^{\mathrm{T}}$ does to A, and Q does to W. These analogies are helpful to understand the latter part of our algorithms as modified Jutten & Herault's algorithms.

Then we show that the solution of the estimating equation is guaranteed to converge to Q^* because of property of estimating functions.

Theorem 8 The transformations $Q = PDQ^*$ for any diagonal matrices D and any permutation matrix P satisfy unbiasedness of the non-diagonal terms of the estimating function. From the additional constraints (4.7) or (4.8), the solution Q^* is selected in these transformations.

Proof If $Q = Q^*$, $\boldsymbol{y} = \boldsymbol{y}^*$ and $V = V^*$. From the distributional assumptions, we can show that the random vector $\boldsymbol{y} = \boldsymbol{y}^*$ is subject to the normal distribution $N(\boldsymbol{s}, V^*)$ for given $\boldsymbol{s}$. Therefore we get the conditional moments as follows.

$$\begin{aligned}
\mathrm{E}\left[\boldsymbol{y}^* \middle| \boldsymbol{s}\right] &= \boldsymbol{s} \\
\mathrm{E}\left[\boldsymbol{y}^* (\boldsymbol{y}^*)^{\mathrm{T}} \middle| \boldsymbol{s}\right] &= \boldsymbol{s}\boldsymbol{s}^{\mathrm{T}} + V^* \\
\mathrm{E}\left[(y_i^*)^3 y_j^* \middle| \boldsymbol{s}\right] &= s_i^3 s_j + 3v_{ij}^* s_i^2 + 3v_{ii}^* s_i s_j + 3v_{ii}^* v_{ij}^*
\end{aligned}$$

Then the expectation of the estimating function becomes

$$\mathrm{E}\left[(y_i^*)^3 y_j^* - 3v_{ij}^* (y_i^*)^2 - 3v_{ii}^* y_i^* y_j^* + 3v_{ii}^* v_{ij}^*\right] = \mathrm{E}\left[s_i^3 s_j\right] = 0, \quad i \neq j$$

because of mutual independence and the zero-mean assumption of $\boldsymbol{s}$.

When $Q = DQ^*$ for $D = \mathrm{diag}(d_{ii})$, $y_i = d_{ii} y_i^*$, and $v_{ij} = d_{ii} v_{ij}^* d_{jj}$. Then because

$$\begin{aligned}
& y_i^3 y_j - 3v_{ij} y_i^2 - 3v_{ii} y_i y_j + 3v_{ii} v_{ij} \\
&= d_{ii}^3 d_{jj} \left\{(y_i^*)^3 y_j^* - 3v_{ij}^* (y_i^*)^2 - 3v_{ii}^* y_i^* y_j^* + 3v_{ii}^* v_{ij}^*\right\},
\end{aligned}$$

unbiasedness holds again.

$$\mathrm{E}\left[y_i^3 y_j - 3v_{ij} y_i^2 - 3v_{ii} y_i y_j + 3v_{ii} v_{ij}\right] = d_{ii}^3 d_{jj} \mathrm{E}\left[s_i^3 s_j\right] = 0, \quad i \neq j$$

Unbiasedness can be derived in the same manner even in case of $Q = PD^*Q^*$.

When we take $Q = DQ^*$, it can be shown that the elements d_{ii} of D must be equal to ± 1 from the additional constraints (4.7) or (4.8).

$$\begin{aligned} \mathrm{E}\left[y_i^2 - v_{ii} - 1\right] &= d_{ii}^2 \mathrm{E}\left[s_i^2\right] - 1 = 0 \\ \left(\text{or} \quad \sum_{j=1}^{m} q_{ij}^2 = d_{ii}^2 \sum_{j=1}^{m} (q_{ij}^*)^2 = d_{ii}^2 = 1\right) \\ \Longleftrightarrow \quad d_{ii}^* = \pm 1, \quad i = 1, \ldots, m. \end{aligned}$$

Therefore with the additional constraints, the expectation of the estimating function becomes zero only at $Q = Q^*$ except for changing signs and orders of its rows. If we assume that the 8th moments of the source signals exist, square integrability (2.9) of this estimating function is satisfied. By calculating the matrix K in (2.8), nonsingularity of K can be proved. □

5 Numerical Experiments

In order to evaluate performance of the proposed algorithm and compare that of other algorithms, we carried out the following numerical experiments. For source signals we synthesize five different acoustic sounds (synthesized music instruments, male voices) $\boldsymbol{s} = (s_1, \ldots, s_5)^{\mathrm{T}}$ whose size are 48000 (see Figure 1). We normalized the source signals so that they have unit variances. At each trial a 15×5 mixing matrix A was randomly generated so that each component was subject to the standard normal distribution independently and source signals $\boldsymbol{s}$ were mixed with this matrix A at first. Then, to the mixed signals we added a Gaussian noise $\boldsymbol{\xi}$ where the standard deviation of each components was determined by a uniform random number on $(0, 6)$. Generating each data set in this way, we made 500 sets of such samples. We estimate the mixing matrices and the noise covariances, applying the following ICA algorithms to the quasi-whitened data $\boldsymbol{y}^{(0)}$ by factor analysis.

FastICA The FastICA algorithm with the kurtosis contrast function.

JADE The JADE algorithm without pre-whitening by PCA (quasi-whitening was done by factor analysis already).

Akuzawa Akuzawa's quasi-Newton algorithm which was used to obtain initial estimators of our algorithm (Akuzawa,2000).

EF The proposed algorithm constructed from the $(3,1)$-type estimating function.

We computed the matrices $R = (r_{ij})$ of crosstalk ratios in order to compare the performance of the estimators. These matrices are constructed by normalizing each row of the following matrix $\widetilde{R}$ so that the maximum absolute value of the components is one and replacing the components with maximum absolute value by zero.

$$\widetilde{R} = (\widehat{A}^{\mathrm{T}}\widehat{\Sigma}^{-1}\widehat{A})^{-1}\widehat{A}^{\mathrm{T}}\widehat{\Sigma}^{-1}A^{*}, \tag{5.1}$$

were $(\widehat{A}, \widehat{\Sigma})$ are the final estimator, and A^{*} denotes the true value. We show several criterions calculated from the matrices $R = (r_{ij})$ of crosstalk ratios from Table 1 to Table 4.

Table 1. Frobenius norm of R

method	mean	s.d.	min	max
FastICA	0.6779	0.4993	0.0555	2.4384
JADE	0.0492	0.0612	0.0146	1.3640
Akuzawa	0.0660	0.0640	0.0369	1.4149
EF	0.0526	0.0546	0.0301	1.2246

Table 2. maximum absolute value of R

method	mean	s.d.	min	max
FastICA	0.3512	0.2713	0.0230	0.9923
JADE	0.0247	0.0346	0.0083	0.7653
Akuzawa	0.0402	0.0351	0.0164	0.7610
EF	0.0297	0.0322	0.0141	0.7191

Table 3. mean of $\sum_j |r_{ij}|$ (i=1,. . . ,5)

method	mean	s.d.	min	max
FastICA	0.4406	0.3238	0.0395	1.7234
JADE	0.0349	0.0376	0.0103	0.8359
Akuzawa	0.0433	0.0368	0.0218	0.8065
EF	0.0356	0.0309	0.0181	0.6869

Table 4. maximum of $\sum_j |r_{ij}|$ (i=1,. . . ,5)

method	mean	s.d.	min	max
FastICA	0.7752	0.6208	0.0542	3.0131
JADE	0.0517	0.0733	0.0154	1.6308
Akuzawa	0.0689	0.0724	0.0350	1.5837
EF	0.0531	0.0454	0.0285	0.9848

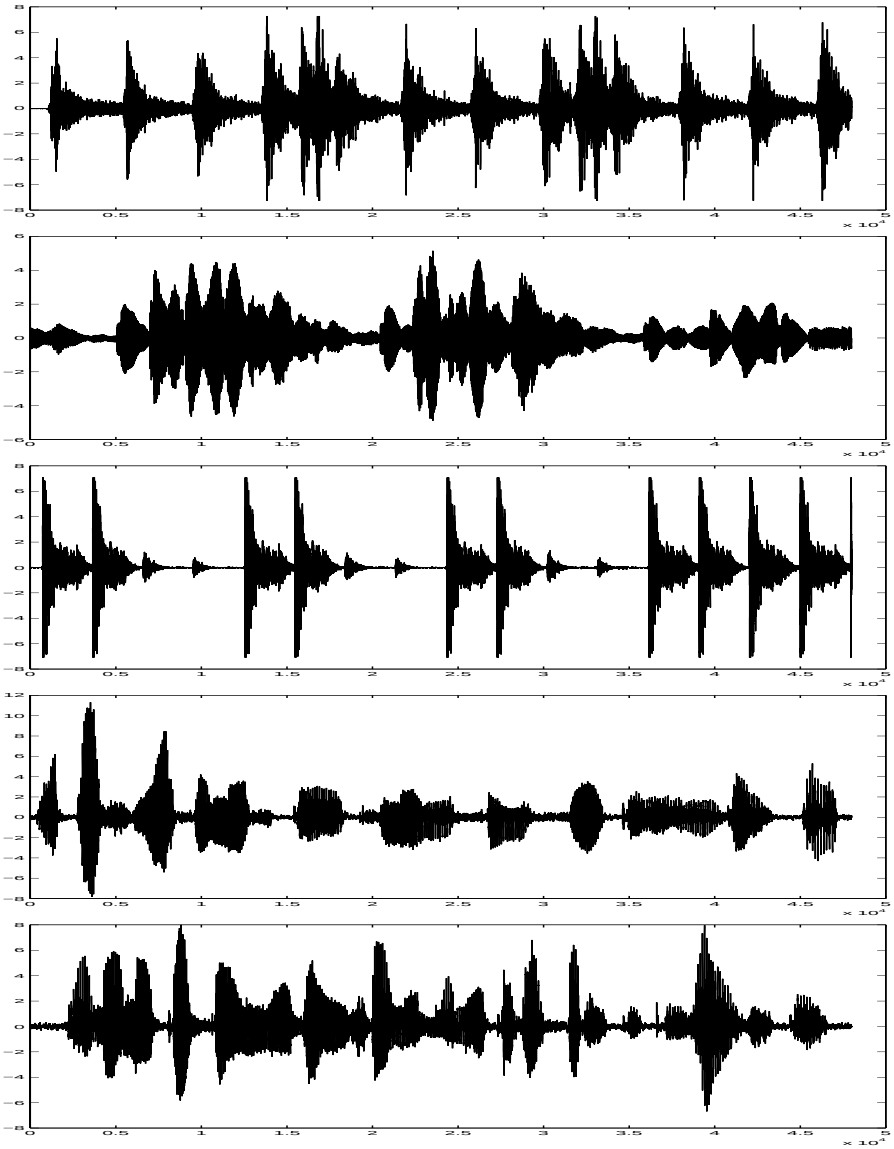

Figure 1: source signals

All of these criterions indicate similar results. We write down conclusions obtained from this numerical experiment.

1. The FastICA combined with factor analysis gives a biased estimator of the mixing matrix in the presence of Gaussian noise. This is because the FastICA used here does not take the additive noise into account.

2. As lead by the theoretical consideration in the present paper, our algorithm, Akuzawa's algorithm, and the JADE give desirable estimators. Our algorithm have almost same performance as the JADE.

3. Estimators of Akuzawa's algorithm are little bit worse than other two methods because the source signals used here are not mutually independent indeed.

4. The means of CPU time spended for calculating estimators are 7.1045, 13.8989, 5.8864 and 2.7563(sec) respectively. Therefore, including the time for calculating initial estimators, our algorithm is much faster than the JADE algorithm. The difference may be much clear in case of lager dimensional data.

6 Discussion

In this paper, we discussed a general form of estimating functions in the noisy ICA model, following the semiparametric statistical approach by Amari and Cardoso(1997). Then a noisy ICA algorithm which gives a consistent estimator of the mixing matrix and the noise variance was proposed. This algorithm consists of two steps: prewhitening by factor analysis and pursuit of the independent component directions with the estimating equations. The $(3,1)$-type Hermite polynomials used in the second step can be regarded as a modification of the estimating function which appears in Jutten-Herault algorithm.

The results of the numerical experiments support the theoretical consideration. They indicate that the proposed algorithm give estimates whose bias are very small, while a noise-free ICA algorithm with quasi-whitening by factor analysis lead to inconsistent estimates. The JADE algorithm also give consistent estimates, because it uses cross-kurtosis tensors which are not influenced by Gaussian noises. We note that our algorithm can be justified in terms of cross-kurtosis too. When the ML method of factor analysis is

used, the final estimates of A and Σ satisfy the estimating equations (4.17), (4.18) and

$$\frac{1}{T}\sum_{t=1}^{T}\left[\{y_j(t)\}^3 y_k(t) - 3v_{jj}y_j(t)y_k(t) - 3v_{jk}\{y_j(t)\}^2 + 3v_{jj}v_{jk}\right] = 0, \quad (6.1)$$

for any $j \neq k$. From (4.17) we derive

$$v_{jj} = \frac{1}{T}\sum_{t=1}^{T}\{y_j(t)\}^2 - 1 \quad (6.2)$$

$$v_{jk} = \frac{1}{T}\sum_{t=1}^{T} y_j(t)y_k(t), \qquad j \neq k. \quad (6.3)$$

Substituting these equations, it can be shown that the equations (6.1) are equivalent to the estimates of the following cross-kurtosis.

$$\begin{aligned}&\widehat{\mathrm{cum}}(y_j, y_j, y_j, y_k)\\ &\equiv \frac{1}{T}\sum_{t=1}^{T}\{y_j(t)\}^3 y_k(t) - 3\left[\frac{1}{T}\sum_{t=1}^{T}\{y_j(t)\}^2\right]\left[\frac{1}{T}\sum_{t=1}^{T} y_j(t)y_k(t)\right] = 0\end{aligned} \quad (6.4)$$

This shows close relation to the JADE algorithm, that is, our algorithm also search for the directions of the independent components by using the fact that some of cross-kurtosis tensors should vanish. Extending this discussion, we conjecture that estimators of the JADE algorithm have lager variances than those of our algorithm, because the JADE algorithm contains polynomials without information of the parameters.

For convenience of explanation, we assumed that the additive noise $\boldsymbol{\xi}$ is subject to the normal distribution $N(\mathbf{0}, \Sigma)$. It is necessary to extend this distributional assumption and other model assumptions. At least, the algorithm proposed here still has consistency in the semiparametric sense under the weaker assumption that $\boldsymbol{\xi}$ has the same 4th order moments structure as the normal distribution. Furthermore, when the number of sensors are much more than that of sources, our algorithm is expected to have good performance even if the additive noise is not Gaussian. The reason is that the noise part included in the quasi-whitened data $\boldsymbol{y}^{(0)}$ is approximately normally distributed from the central limit theorem.

In this paper we only checked performance of our algorithm and compared to that of a few existing algorithms via numerical examples. Theoretical analysis of their performance should be studied in the future. Although

the presented framework of estimating functions is useful for investigation of the noisy ICA model, it is also important to develop other kind of algorithms such as the bias removal method and analyze their performance.

References

[1] Akuzawa,T. (2000). extended quasi-Newton method for the ICA. submitted.

[2] Amari,S. (1985). *Differential-Geometrical Method in Statistics*. New York: Springer-Verlag, **28**.

[3] Amari,S., and Cardoso,J.-F. (1997). Blind source separation — semiparametric statistical approach. *IEEE Trans. on Signal Processing*, **45**, 2692–2700.

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

[5] Amari,S., and Kawanabe,M. (1997). Information geometry of estimating functions in semiparametric statistical models. in *Estimating Functions*, V.P.Godambe, Ed., IMS Monograph Series.

[6] Attias,H. (1999). Independent Factor Analysis. *Neural Computation*, **11**, 803-851.

[7] Bickel,P.J., Klaassen,C.A.J., Ritov,V., and Wellner,J.A. (1993). *Efficient and Adaptive Estimation in Semiparametric Models*. Baltimore: Johns Hopkins University Press.

[8] Cardoso,J.-F., and Souloumiac,A. (1993). Blind beamforming for non Gaussian signals. *IEE-Proceedings-F*, **140**, 362-370.

[9] Cichocki,A., Douglas,S.C., and Amari,S. (1998). Robust techniques for independent component analysis with noisy data. *Neurocomputing*, **22**, 113-129.

[10] Godambe,V.P. Ed. (1991). *Estimating Functions*. New York: Oxford Univ. Press.

[11] Hyvärinen,A. (1998). Independent Component Analysis in the Presence of Gaussian Noise by Maximizing Joint Likelihood. *Neurocomputing*, **22**, 49-67.

[12] Hyvärinen,A. (1999). Gaussian Moments for Noisy Independent Component Analysis. *IEEE Signal Processing Letters*, **6**, 145–147.

[13] Hyvärinen,A. (1999). Survey on independent component analysis. *Neural Computing Surveys*, **2**, 94–128.

[14] Jutten,C., and Herault,J. (1991). Separation of sources, Part I. *Signal Processing*, **24**, 1–10.