Journal of Machine Learning Research 6 (2005) 453–482 Submitted 6/04; Published 4/05

Estimating Functions for Blind Separation When Sources Have Variance Dependencies

Motoaki Kawanabe NABE@FIRST.FHG.DE
Fraunhofer FIRST.IDA
Kekuléstrasee 7
12489 Berlin, Germany

Klaus-Robert Müller KLAUS@FIRST.FHG.DE
Fraunhofer FIRST.IDA
Kekuléstrasse 7
12489 Berlin, Germany and
Department of Computer Science
University of Potsdam
August-Bebel-Strasse 89
14482 Potsdam, Germany

Editor: Aapo Hyvärinen

Abstract

A blind separation problem where the sources are not independent, but have variance dependencies is discussed. For this scenario Hyvärinen and Hurri (2004) proposed an algorithm which requires no assumption on distributions of sources and no parametric model of dependencies between components. In this paper, we extend the semiparametric approach of Amari and Cardoso (1997) to variance dependencies and study estimating functions for blind separation of such dependent sources. In particular, we show that many ICA algorithms are applicable to the variance-dependent model as well under mild conditions, although they should in principle not. Our results indicate that separation can be done based only on normalized sources which are adjusted to have stationary variances and is not affected by the dependent activity levels. We also study the asymptotic distribution of the quasi maximum likelihood method and the stability of the natural gradient learning in detail. Simulation results of artificial and realistic examples match well with our theoretical findings.

Keywords: blind source separation, variance dependencies, independent component analysis, semiparametric statistical models, estimating functions

1. Introduction

Blind methods of source separation have been successfully applied to many areas of science (e.g. Hyvärinen et al., 2001b; Olshausen and Field, 1996; Makeig et al., 1997; Vigario, 1997; Ziehe et al., 2000; Thi and Jutten, 1995; Cardoso, 1998a; Parra and Spence, 2000; Cardoso, 2003; Meinecke et al., 2005). The basic model assumes that the observed signals are linear superpositions of underlying hidden source signals. Let us denote the n source signals by the vector

$s(t) = (s_1(t), \ldots, s_n(t))^\top$, and the observed signals by $x(t) = (x_1(t), \ldots, x_m(t))^\top$. In this paper,[1] we will focus on real-valued signals. The mixing can be expressed as the equation

$$x(t) = As(t),$$

where $A = (a_{ij})$ denotes the mixing matrix. For simplicity, we consider the case where the number of source signals equals that of observed signals ($n = m$). Both the sources $s(t)$ and the mixing matrix A are unknown, and the goal is to estimate them based on the observation $x(t)$ alone.

In most blind source separation (BSS) methods, the source signals are assumed to be statistically independent. Blind source separation based on such a model is called independent component analysis (ICA). By using non-Gaussianity of the sources, the mixing matrix can be estimated and the source signals can be extracted under appropriate conditions. There are also further approaches of BSS, that are, for example, based on second-order statistics and algorithms exploiting nonstationarity. The second-order methods are applicable to the case where the source signals have (lagged) auto-correlation. Provided that components have nonstationary, smoothly changing variances, the model can be estimated as well by algorithms based on nonstationarity of signals.

Among many extensions of the basic ICA models, several researchers have studied the case where the source signals are not independent (for example, Cardoso, 1998b; Hyvärinen et al., 2001a; Bach and Jordan, 2002; Valpola et al., 2003, see also references in Hyvärinen and Hurri, 2004). The dependencies either need to be exactly known beforehand, or they are simultaneously estimated by the algorithms. Recently, a novel idea called double-blind approach was introduced by Hyvärinen and Hurri (2004). In contrast to previous work, their method requires no assumption on the distributions of the sources and no parametric model of the dependencies between the components. They simply assume that the sources are dependent only through their variances and that the sources have temporal correlation. In the Topographic ICA (Hyvärinen et al., 2001a), the dependencies of the sources are also caused only by their variances, but in contrast to the double blind case, they are determined by a prefixed neighborhood relation. It should be noted that for such dependent component models identifiability results have not been theoretically established so far, while identifiability of multidimensional ICA was proven by Theis (2004).

A statistical basis of ICA was established by Amari and Cardoso (1997). They pointed out that the ICA model is an example of semiparametric statistical models (Bickel et al., 1993; Amari and Kawanabe, 1997a,b) and studied estimating functions for it. In particular, they showed that the quasi maximum likelihood (QML) estimation and the natural gradient learning give a correct solution regardless of the true source densities which satisfy certain mild conditions. In this paper, we extend their approach to the BSS problem considered in Hyvärinen and Hurri (2004). Investigating estimating functions for the model, we show that many of ICA algorithms based on the independence assumption can achieve consistent solutions in a local sense, even if there exist variance dependencies, which is astonishing and seems somewhat counterintuitive. We remark that estimating functions are concerned with local consistency ('consistency' will denote its local version in the following) and in general have spurious solutions. For a few algorithms, even global consistency has been proven by different principles (for example, Hyvärinen and Hurri, 2004). Nevertheless, our result goes beyond existing ones, because it covers most types of BSS algorithms and can give asymptotic distributions. The main message of this paper is that most ICA algorithms can be proven to be consistent in our framework *although* the data is *not* independent. So they must effectively

1. This is an extended version of Kawanabe and Müller (2004) presented at ICA2004.

use some concept beyond independence. Thus our consistency results indicate that separation can be done based only on normalized sources which are adjusted to have stationary variances and is not affected by the dependent activity levels.

This paper is organized as follows. At first, we define the variance-dependent model in Section 2 and explain estimating functions, a useful tool for discussing semiparametric estimators in Section 3. In Section 4, we discuss the relation between estimating functions for the ICA model and those for the variance-dependent BSS model in general. It is shown that these algorithms work properly, *even if* there exist spatiotemporal variance dependencies. Among several ICA algorithms, the quasi maximum likelihood method and its online version, the natural gradient learning are discussed in detail. We study the asymptotic distributions of the quasi maximum likelihood method (Section 5.1) and the stability of the natural gradient learning (Section 5.2). We also give a brief summary about several other ICA algorithms from our viewpoint in Section 5.3. Detailed discussion can be found in Appendix A. The theoretical insights are underlined by several numerical simulations in Section 6. In particular, we carried out two experiments, where we extract two speech signals with high variance dependencies. It is sometimes believed that ICA algorithms work for mixture of acoustic signals or natural images because the data are sparse and often disjoint. Our results show that they can also separate even highly coherent signals, and our theoretical analysis can thus help to understand the reason.

2. Variance-Dependent BSS Model

Hyvärinen and Hurri (2004) formalized the probabilistic framework of variance-dependent blind separation. Let us assume that each source signal $s_i(t)$ is a product of non-negative activity level $v_i(t)$ and underlying i.i.d. signal $z_i(t)$, that is,

$$s_i(t) = v_i(t)z_i(t). \tag{1}$$

We remark that the sequences of the vectors $s = (s_1, \ldots, s_n)^\top$, $v = (v_1, \ldots, v_n)^\top$ and $z = (z_1, \ldots, z_n)^\top$ are considered as multivariate random processes in this paper. In practice, the activity levels $v_i(t)$ are often dependent among different signals and each observed signal is expressed as

$$x_i(t) = \sum_{j=1}^{n} a_{ij} v_j(t) z_j(t), \qquad i = 1, \ldots, n,$$

where $v_i(t)$ and $z_i(t)$ satisfy:

(i) $v_i(t)$ and $z_j(t')$ are independent for all i, j, t and t',

(ii) each $z_i(t)$ is i.i.d. in time for all i, the random vector $z = (z_1, \ldots, z_n)^\top$ is mutually independent,

(iii) $z_i(t)$ have zero mean and unit variance for all i.

No assumption on the distribution of z_i is made except (iii). Regarding the general activity levels v_i's, $v_i(t)$ and $v_j(t)$ are allowed to be statistically dependent, and furthermore, no particular assumption on these dependencies is made (double blind situation). We refer to this framework as the variance-dependent BSS model in this paper. Figure 1 shows an example of the sources s used in the model. As stated in the assumption (ii) above, the normalized signals z_1 and z_2 are mutually independent.

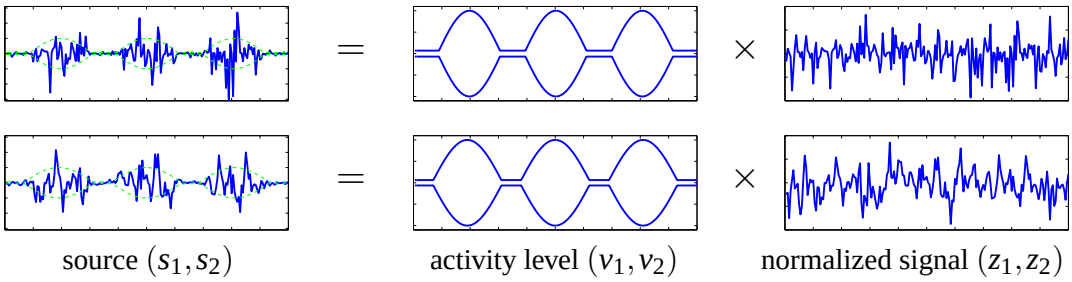

Figure 1: Sources (s_1, s_2) with variance dependencies in the variance-dependent BSS model. In the middle panels both v_i and $-v_i$ are plotted for clarity.

However, since the sequences z_1 and z_2 are multiplied by extremely dependent activity levels v_1 and v_2, respectively, the short-term variance of the source signals s_1 and s_2 are highly correlated.

Hyvärinen and Hurri (2004) proposed an algorithm which maximizes the objective function

$$J(W) = \sum_{i,j} [\widehat{\mathrm{cov}}([w_i^\top u(t)]^2, [w_j^\top u(t-\Delta t)]^2)]^2,$$

where $\widehat{\mathrm{cov}}$ denotes the sample covariance, $W = (w_1, \ldots, w_n)^\top$ is constrained to be orthogonal and where $u(t)$ is obtained by preprocessing the signal $x(t)$ by spatial whitening. It was proved that the objective function J is maximized when WA equals a signed permutation matrix, if the matrix $K = (K_{ij}) = (\mathrm{cov}\{s_i^2(t), s_j^2(t-\Delta t)\})$ is of full rank. This method shows good performance as long as there exist temporal variance dependencies and the data is not spoiled by outliers (see Meinecke et al., 2004).

It is important to remark that the nonstationary algorithm by Pham and Cardoso (2000) was also designed for the same source model (1), except that $v_i(t)$'s are assumed to be deterministic and slowly varying. However, it is straightforward to show validity of this algorithm, when $v_i(t)$'s are (slowly-varying) random sequences.

3. Semiparametric Statistical Models and Estimating Functions

Amari and Cardoso (1997) established a statistical basis of the ICA problem. They pointed out that the standard ICA model[2]

$$p(X|B, \rho_s) = |\det B|^T \prod_{t=1}^{T} \prod_{i=1}^{n} \rho_{s_i}\{b_i^\top x(t)\} \tag{2}$$

is an example of semiparametric statistical models (Bickel et al., 1993; Amari and Kawanabe, 1997a,b), where $B = (b_1, \ldots, b_n)^\top = A^{-1}$ is the demixing matrix to be estimated and $\rho_s(s) = \prod_{i=1}^{n} \rho_{s_i}(s_i)$ is the density of the sources s. Notations used in the following sections are also summarized in Table 1. As the function ρ_s in this model, semiparametric statistical models contain infinite dimensional or functional nuisance parameters which are difficult to estimate. Moreover, they even disturb inference on parameters of interest.

2. Since the sources are assumed to be i.i.d. in time, people consider the distribution of one sample x instead of the entire sequence X.

$x(t) = (x_1(t), \ldots, x_n(t))^\top$	observed data at t
$X = (x(1), \ldots, x(T))$	whole sequence of the observed data
$s(t) = (s_1(t), \ldots, x_n(t))^\top$	source signals at t
$v(t) = (v_1(t), \ldots, v_n(t))^\top$	general activity levels of the sources $s(t)$
$V = (v(1), \ldots, v(T))$	whole sequence of the activity levels
$z(t) = (z_1(t), \ldots, z_n(t))^\top$	normalized source signals by the activity levels $v(t)$
A	$n \times n$ mixing matrix
$B = (b_{ij}) = (b_1, \ldots, b_n)^\top$	demixing matrix which is equivalent to A^{-1}
$\rho_z(z) = \prod_{i=1}^{n} \rho_{z_i}(z_i)$	density of the normalized source signals z
$\rho_V(V)$	density of the entire sequence $V = (v(1), \ldots, v(T))$ of the activity levels
$y(t) = Bx(t)$	extracted sources by the demixing matrix B
$F(x, B)$ or $\bar{F}(X, B)$	estimating function which is an $n \times n$ matrix-valued function of the data and the parameter B
$\mathrm{vec}(F)$ $= (F_{11}, \ldots, F_{n1}, \ldots, F_{1n}, \ldots, F_{nn})^\top$	vectorization operator

Table 1: List of notations used in the variance-dependent BSS model

In the variance-dependent BSS model which we consider, the sources $s(t)$ are decomposed of two components, the normalized signals $z(t) = (z_1(t), \ldots, z_n(t))^\top$ and the general activity levels $v(t) = (v_1(t), \ldots, v_n(t))^\top$. Since the former has mutual independence like the ICA model, the density of the data X is factorized as

$$p(X|V; B, \rho_z) = |\det B|^T \prod_{t=1}^{T} \prod_{i=1}^{n} \frac{1}{v_i(t)} \rho_{z_i}\left\{ \frac{b_i^\top x(t)}{v_i(t)} \right\}, \tag{3}$$

when $V = (v(1), \ldots, v(T))$ is fixed. Therefore, the marginal distribution can be expressed as

$$p(X|B, \rho_z, \rho_V) = \int p(X|V; B, \rho_z) \rho_V(V) \mathrm{d}V, \tag{4}$$

where the density ρ_V of V becomes an extra nuisance function.

In order to construct valid estimators for such semiparametric models, estimating functions were introduced by Godambe (1976). Let us consider a general semiparametric model $p(x|\theta, \rho)$, where θ is an r-dimensional parameter of interest and ρ is a nuisance parameter. An r-dimensional vector valued function $f(x, \theta)$ is called an estimating function, when it satisfies the following conditions

for any θ and ρ (Godambe, 1991),

$$\begin{aligned}
&\mathrm{E}[\, f(x,\theta) \,|\theta,\rho] = \mathbf{0},\\
&|\det Q| \neq 0, \qquad \text{where } Q = \mathrm{E}\left[\frac{\partial}{\partial\theta} f(x,\theta) \,\middle|\, \theta,\rho \right],\\
&\mathrm{E}\left[\, \|f(x,\theta)\|^2 \,\middle|\, \theta,\rho \right] < \infty,
\end{aligned}$$

where $\mathrm{E}[\cdot|\theta,\rho]$ denotes the expectation over x with the density $p(x|\theta,\rho)$ and $\|\cdot\|$ is the Euclidean norm. Suppose i.i.d. samples $x(1),\ldots,x(T)$ are obtained from the model $p(x|\theta^*,\rho^*)$. If such a function exists, by solving the estimating equation

$$\sum_{t=1}^{T} f(x(t),\widehat{\theta}) = \mathbf{0}, \tag{5}$$

we can get an estimator $\widehat{\theta}$ with good asymptotic property. Such an estimator that is a solution of an estimating equation as (5) is called an M-estimator in statistics (Huber, 1981). It can be regarded as an extension of the maximum likelihood method for parametric models. The M-estimator $\widehat{\theta}$ is consistent regardless of the true nuisance parameter ρ^*, when the sample size T goes to infinity. Moreover, it is asymptotically Gaussian distributed, that is, $\widehat{\theta} \sim N(\theta^*, \mathrm{Av})$, where Av denotes the asymptotic variance computed by the following equation

$$\mathrm{Av} = \mathrm{Av}(\theta^*,\rho^*) = \frac{1}{T} Q^{-1}\, \mathrm{E}\left[\, f(x,\theta)\, f^{\top}(x,\theta) \,\middle|\, \theta^*,\rho^* \right]\, (Q^{-1})^{\top},$$

and $Q = Q(\theta^*,\rho^*) = \mathrm{E}\left[\frac{\partial}{\partial\theta} f(x,\theta) \,\middle|\, \theta^*,\rho^* \right]$. We remark that the asymptotic variance Av depends on the true parameters (θ^*,ρ^*), but not on the data $x(1),\ldots,x(T)$. As we will explain in Section 4.2, notions of estimating functions and M-estimators were extended to non i.i.d cases.

Although estimating functions are useful for semiparametric models, it is non-trivial to find such functions. Amari and Kawanabe (1997a,b) studied this problem from a geometrical point of view and gave a guideline for discussing estimating functions.

The asymptotic result guarantees theoretically that the estimator $\widehat{\theta}$ derived from the estimating function converges to the true parameter θ^* under mild conditions. However, we should remark that the asymptotic variance Av of the estimator depends on the true nuisance parameter ρ^*. For example, when the matrix Q is almost singular at ρ^*, it can happen that the asymptotic variance Av becomes very large. This may cause some practical problem, that is, the estimate from finite samples can be no longer close to the true parameter. We will revisit this issue in Section 6.

Furthermore, online algorithms with similar consistency property can also be constructed from estimating functions,

$$\theta_{t+1} = \theta_t - \eta_t\, f(x(t),\theta_t), \tag{6}$$

$$\theta_{t+1} = \theta_t - \eta_t\, R(\theta_t)\, f(x(t),\theta_t), \tag{7}$$

where $R(\theta)$ is an $n \times n$ nonsingular matrix and depends only on θ. We remark that the functions $f(x,\theta)$ and $R(\theta) f(x,\theta)$ give the same estimating equation, if $R(\theta)$ has the inverse matrix and does not depend on the data x. Such functions are called equivalent estimating functions. It is also easy to see that the online algorithms (6) and (7) have the same equilibria points. However, their dynamics are different. The stability of such online learning was investigated by Amari et al. (1997).

4. General Properties of Estimating Functions for Blind Separation

In this section we will at first review estimating functions for the ICA model (2) (see also Amari and Cardoso, 1997; Cardoso, 1997) and then discuss our contribution, that is, by defining estimating functions for the variance-dependent BSS model (3) and (4).

4.1 Estimating Functions for Ordinary Blind Source Separation

In case of the ICA model, the parameter of interest is the $n \times n$ matrix $B = A^{-1}$ and hence it is convenient to write the estimating functions in $n \times n$ matrix form $F(x,B)$. The conditions of estimating functions are reshaped accordingly as

$$\mathrm{E}[F(x,B)\ |B,\rho_s] = 0, \tag{8}$$

$$|\det Q| \neq 0, \qquad \text{where } Q = \mathrm{E}\left[\frac{\partial \mathrm{vec}\{F(x,B)\}}{\partial \mathrm{vec}(B)} \ \middle|\ B,\rho_s \right], \tag{9}$$

$$\mathrm{E}\left[\ \|F(x,B)\|_F^2\ \middle|\ B,\rho_s\right] < \infty, \tag{10}$$

where $\mathrm{vec}(F) = (F_{11},\ldots,F_{n1},\ldots,F_{1n},\ldots,F_{nn})^\top$ is the vectorization of matrices and $\|\cdot\|_F$ denotes Frobenius norm. It should be noted that both in usual ICA models and in the variance-dependent BSS model, scales and orders of the sources cannot be determined, that is, two matrices B and PDB indicate the same distribution, when P and D are a permutation and a diagonal matrix respectively (Comon, 1994).[3] Therefore, we can find any matrix in the equivalence class, so for notational convenience we will fix scales as the constraints (25) later.[4]

One of the standard ICA algorithms originates from maximum likelihood estimation, which is asymptotically the best method if the density ρ_s is known. Because in the semiparametric model ρ_s is unknown and difficult to estimate, the idea is to use instead the maximum likelihood estimation under a prefixed density $\widetilde{\rho}_s$. The method is called the quasi maximum likelihood estimation, since the fixed $\widetilde{\rho}_s$ does not coincide with the true one. The estimator $\widehat{B}$ is derived from the equation

$$\sum_{t=1}^{T} \left[I - \varphi\{y(t)\} y^\top(t) \right] = 0, \tag{11}$$

where $y(t) = \widehat{B}x(t)$ is the estimator of the sources, $\varphi(y) = (\varphi_1(y_1),\ldots,\varphi_n(y_n))^\top$ and

$$\varphi_i(y_i) = -\frac{\mathrm{d}}{\mathrm{d}y_i} \log \widetilde{\rho}_{s_i}(y_i).$$

For the nonlinear function $\varphi_i(y_i)$,

$$\varphi_i(y_i) = \tanh(c y_i), \qquad c > 0, \tag{12}$$

$$\varphi_i(y_i) = y_i^3, \tag{13}$$

are often employed. The function $F(x,B) = I - \varphi\{Bx\}(Bx)^\top$ in (11) is an example of estimating functions for the ICA model, provided that it satisfies (9) and (10). It is trivial to show that it fulfills (8). Another example is the function

$$F(x,B) = Bx\,(Bx)^\top - I + (Bx)\ g^\top(Bx) - g(Bx)\ (Bx)^\top$$

3. It is clear that the variance-dependent BSS model has at least such indeterminacy. On the other hand, the identifiability in this case has not been proved so far.
4. We ignore the permutation indeterminacy P, since it's locally not problematic.

for FastICA (see (37) in Appendix A.1), where $g(\cdot)$ is a vector valued non-linear function as $\varphi(\cdot)$. We remark that this procedure can also be derived from minimum mutual information (Yang and Amari, 1997) and infomax principle (Bell and Sejnowski, 1995).

In general the quasi maximum likelihood estimator is no longer consistent because of misspecified distribution. However, in the ICA model (2), Amari and Cardoso (1997) found that the quasi maximum likelihood method and its online version (the natural gradient learning) give an asymptotically consistent estimator, provided that $F(x,B) = I - \varphi\{Bx\}(Bx)^\top$ satisfies (9) and (10). In particular, we remark that the assumed distribution $\widetilde{\rho}_s$ is not equal to the true one. This research has motivated us to investigate also such semiparametric procedures for the variance-dependent BSS model (3) and (4). In particular, we will show in Section 5.1 that the quasi maximum likelihood method (11) still gives a consistent estimator even under this extended situation.

4.2 Estimating Functions for Variance-Dependent Blind Source Separation

In the variance-dependent BSS model, in contrast to the ICA model studied by Amari and Cardoso (1997), the data sequence $X = (x(1),\ldots,x(T))$ is not i.i.d. in time, but might have time dependencies. Therefore, we have to consider more general functions $\bar{F}(X,B)$ of the whole sequence X. General estimating functions $\bar{F}(X,B)$ must satisfy

$$\mathrm{E}[\bar{F}(X,B)\ |B,\rho_z,\rho_V] = 0, \tag{14}$$

$$|\det Q| \neq 0, \qquad \text{where } Q = \mathrm{E}\left[\ \frac{\partial \mathrm{vec}\{\bar{F}(X,B)\}}{\partial \mathrm{vec}(B)}\ \middle|\ B,\rho_z,\rho_V\right], \tag{15}$$

$$\mathrm{E}\left[\ \|\bar{F}(X,B)\|_F^2\ \middle|\ B,\rho_z,\rho_V\right] < \infty, \tag{16}$$

for all (B,ρ_z,ρ_V). An M-estimator $\widehat{B}$ can be derived from the estimating equation

$$\bar{F}(X,\widehat{B}) = 0. \tag{17}$$

Suppose that the data X is subject to $p(X|B^*,\rho_z^*,\rho_V^*)$ defined by (3) and (4).

Theorem 1 *If the function $\bar{F}(X,B)$ satisfies the conditions (14) – (16) and appropriate regularity conditions such as Condition 2.6 in Sørensen (1999), the M-estimator $\widehat{B}$ derived from the equation (17) is asymptotically Gaussian distributed* $\mathrm{vec}(\widehat{B}) \sim N(\mathrm{vec}(B^*),\mathrm{Av})$, *where*

$$\begin{aligned}
\mathrm{Av} &= \mathrm{Av}(B^*,\rho_z^*,\rho_V^*) = Q^{-1}\Sigma(Q^{-1})^\top, \\
\Sigma &= \Sigma(B^*,\rho_z^*,\rho_V^*) = E\left[\mathrm{vec}\{\bar{F}(X,B^*)\}\mathrm{vec}\{\bar{F}(X,B^*)\}^\top\middle|\ B^*,\rho_z^*,\rho_V^*\right] \\
Q &= Q(B^*,\rho_z^*,\rho_V^*) = \mathrm{E}\left[\ \frac{\partial \mathrm{vec}\{\bar{F}(X,B^*)\}}{\partial \mathrm{vec}(B)}\ \middle|\ B^*,\rho_z^*,\rho_V^*\right].
\end{aligned} \tag{18}$$

Proof See Sørensen (1999).

Now, we investigate the relation between estimating functions for the ICA model and those for the variance-dependent BSS model. Let $F(x,B)$ be an estimating function for the ICA model. In the ICA context it is often the case that such estimating functions satisfy

$$\mathrm{E}[F_{ij}(x,DB)\ |B,\rho_s] = 0, \qquad i \neq j, \tag{19}$$

for any diagonal matrix D, that is, its off-diagonal parts (19) hold for all matrices equivalent to B. The scale factor D is determined usually by the diagonal parts of condition (8)

$$\mathrm{E}[F_{ii}(x,B)\ |B,\rho_s] = 0,$$

in the ordinary ICA model. We will soon present the equation for fixing the scale factor D in the variance-dependent BSS model.

Let us consider the function

$$\bar{F}(X,B) = \sum_{t=1}^{T} F(x(t),B), \tag{20}$$

which is used in estimating equations for the ICA model.[5] We can show that this function becomes a candidate of estimating functions for the variance-dependent BSS model.

Proposition 2 *The function $\bar{F}(X,B)$ defined in (20) satisfies condition (14), provided that $F(x,B)$ is an estimating function for the ICA model and fulfills (19). Furthermore, if the additional assumption*

$$\mathrm{E}\left[\ \|F(x(t),B)\|_F^2\ \middle|\ B,\rho_z,\rho_V\right] < \infty, \qquad \forall t \tag{21}$$

holds, condition (16) is also satisfied.

Proof Taking expectations of the off-diagonal terms of (14), we get

$$\begin{aligned}
\mathrm{E}\left[\ \bar{F}_{ij}(X,DB)\ \middle|\ B,\rho_z,\rho_V\right] &= \mathrm{E}\left[\ \sum_{t=1}^{T} \mathrm{E}\left[\ F_{ij}(x(t),DB)\ \middle|\ V;B,\rho_z\right]\ \middle|\ \rho_V\right] \\
&= \mathrm{E}\left[\ \sum_{t=1}^{T} E\left[\ F_{ij}(x(t),DB)\ \middle|\ B,\rho_{s|v(t)}\right]\ \middle|\ \rho_V\right]
\end{aligned}$$

where $\rho_{s|v(t)}$ is the density function of $s(t)$ when its activity level is fixed at $v(t)$, that is,

$$\rho_{s|v(t)}(s) = \prod_{i=1}^{n} \frac{1}{v_i(t)}\ \rho_{z_i}\left\{\frac{s_i}{v_i(t)}\right\}.$$

We remark that the expectation $\mathrm{E}[\cdot|V;B,\rho_z]$ ($\mathrm{E}[\cdot|B,\rho_{s|v(t)}]$) is taken over $z(t)$ (resp. $s(t)$) under fixed activity levels V, while $\mathrm{E}[\cdot|\rho_V]$ denotes the expectation over the activity level V. Because (19) holds for any ρ_s, we can prove

$$\mathrm{E}\left[\ \bar{F}_{ij}(X,DB)\ \middle|\ B,\rho_z,\rho_V\right] = 0,$$

for all diagonal matrices D. If we select the scale factor D such that the diagonal terms

$$\mathrm{E}\left[\bar{F}_{ii}(X,B)\ |B,\rho_z,\rho_V\right] = 0$$

hold, $\bar{F}$ satisfies the unbiasedness condition (14). We furthermore note that this scaling is different from that in the ICA model presented before, and the expectation $\mathrm{E}\left[\ F_{ii}(x(t),B)\ |B,\rho_{s|v(t)}\right]$ at each time t can be non-zero in general.

5. We remark that some of ICA/BSS algorithms (for example, TDSEP/SOBI) are not based on estimating functions in this class. Because it is not easy to discuss them in such a general form, we deal with other classes separately in Appendix A.

The left hand side of Eq. (16) can be expressed as

$$\begin{aligned}
&\mathrm{E}\left[\, \|\bar{F}(X,B)\|_F^2 \,\middle|\, B,\rho_z,\rho_V\right] \\
&= \mathrm{E}\left[\sum_{t,t'} \mathrm{E}\left[\mathrm{tr}\{F(x(t),B)F^\top(x(t'),B)\} \,\middle|\, V;B,\rho_z\right] \,\middle|\, \rho_V\right] \\
&= \mathrm{E}\left[\sum_{t} \mathrm{E}\left[\, \|F(x(t),B)\|_F^2 \,\middle|\, B,\rho_{s|v(t)}\right] \,\middle|\, \rho_V\right] \\
&\quad + \mathrm{E}\left[\sum_{t\neq t'}\sum_{i=1}^{n} \mathrm{E}\left[\, F_{ii}(x(t),B) \,|\, B,\rho_{s|v(t)}\right] \mathrm{E}\left[\, F_{ii}(x(t'),B) \,\middle|\, B,\rho_{s|v(t')}\right] \,\middle|\, \rho_V\right], \qquad (22)
\end{aligned}$$

where we used the fact that $x(t)$ and $x(t')$ $(t \neq t')$ are independent for fixed V. From assumption (21), the first term of Eq. (22) is finite.

$$\begin{aligned}
&\sum_t \mathrm{E}\left[\, \mathrm{E}\left[\, \|F(x(t),B)\|_F^2 \,\middle|\, B,\rho_{s|v(t)}\right] \,\middle|\, \rho_V\right] \\
&\quad = \sum_t \mathrm{E}\left[\, \|F(x(t),B)\|_F^2 \,\middle|\, B,\rho_z,\rho_V\right] < \infty \qquad (23)
\end{aligned}$$

We remark that condition (10) does not necessarily imply assumption (21). Let us define $c(v(t)) := \mathrm{E}\left[\, \|F(x(t),B)\|_F^2 \,\middle|\, B,\rho_{s|v(t)}\right]$. Since

$$\left|\mathrm{E}\left[\, F_{ii}(x(t),B) \,|\, B,\rho_{s|v(t)}\right]\right| \leq \sqrt{\mathrm{E}\left[\, F_{ii}^2(x(t),B) \,\middle|\, B,\rho_{s|v(t)}\right]} \leq \sqrt{c(v(t))},$$

the second term of Eq. (22) (called r in the following) can be bounded as

$$\begin{aligned}
|r| &< n\sum_{t\neq t'} \mathrm{E}\left[\sqrt{c(v(t))}\sqrt{c(v(t'))} \,\middle|\, \rho_V\right] \\
&\leq n\left\{ \sum_t \sqrt{\mathrm{E}\left[\, c(v(t)) \,|\, \rho_V\right]} \right\}^2 \\
&\leq nT \sum_t \mathrm{E}\left[\, c(v(t)) \,|\, \rho_V\right].
\end{aligned}$$

Here we used Schwarz's inequality twice. Because of Eq. (23), this bound is also finite. □

The basic idea of this proof is that the situation becomes similar to the ordinary ICA model, if the activity levels V are fixed. Unfortunately, the other conditions are difficult to be proven in this general form. For example, the second condition can be transformed in the similar way as

$$\begin{aligned}
&\mathrm{E}\left[\frac{\partial \mathrm{vec}\{\bar{F}(X,B)\}}{\partial \mathrm{vec}(B)} \,\middle|\, B,\rho_z,\rho_V\right] \\
&\quad = \sum_{t=1}^{T} \mathrm{E}\left[\mathrm{E}\left[\frac{\partial \mathrm{vec}\{F(x(t),B)\}}{\partial \mathrm{vec}(B)} \,\middle|\, B,\rho_{s|v(t)}\right] \,\middle|\, \rho_V\right]. \qquad (24)
\end{aligned}$$

Even if each term $\mathrm{E}\left[\frac{\partial \mathrm{vec}\{F(x(t),B)\}}{\partial \mathrm{vec}(B)} \,\middle|\, B,\rho_{s|v(t)}\right]$ is non-singular, it may still be possible that the sum (24) becomes singular. However this is in practice an extremely rare case.

5. Consistency Results for Variance Dependent Blind Source Separation Using the Estimating Function Framework

We will use the estimating function framework to prove consistency results for (i) the quasi maximum likelihood type methods (for example, Pham and Garrat, 1997; Bell and Sejnowski, 1995), (ii) the natural gradient learning for ICA (for example, Amari, 1998) and (iii) various other ICA algorithms such as FastICA (Hyvärinen and Oja, 1997), TDSEP/SOBI (Ziehe and Müller, 1998; Belouchrani et al., 1997), 'Sepagaus' (Pham and Cardoso, 2000) and JADE (Cardoso and Souloumiac, 1993).

5.1 Asymptotic Distribution of the Quasi Maximum Likelihood Estimator

In this section, it is shown that the quasi maximum likelihood method (11) as for example Pham and Garrat (1997); Bell and Sejnowski (1995) still gives a consistent estimator even under the extended model (3) and (4). For convenience, we fix the scales of the recovered signals as

$$\mathrm{E}\left[\sum_{t=1}^{T}\varphi_i\{b_i^\top x(t)\}\,b_i^\top x(t)\,\middle|\,B,\rho_z,\rho_V\right]=T, \tag{25}$$

for $i=1,\ldots,n$. Then (14) is automatically fulfilled for the diagonal terms. We remark that by this constraints the length of b_i's may depend on the nuisance parameters (ρ_z,ρ_V), but this does not change the following discussion, because the scales can be fixed arbitrarily.

Since the function $F(x,B)=I-\varphi\{Bx\}(Bx)^\top$ obviously satisfies (19), we already know from Theorem 2 that the function

$$\bar{F}^{\mathrm{QML}}(X,B)=\sum_{t=1}^{T}\left[I-\varphi\{y(t)\}y^\top(t)\right]$$

satisfies the conditions (14) and (16) under the assumption

$$\mathrm{E}\left[\varphi_i^2\{y_i(t)\}y_j^2(t)\,\middle|\,B,\rho_z,\rho_V\right]<\infty,\qquad\forall i,\,j,\,t, \tag{26}$$

where $y(t)$ denotes the extracted sources $Bx(t)$. The additional assumption imposes mild restriction on the distribution of the activity levels V. For example, when the density ρ_V has extremely heavy tails, the left hand side of Eq. (16) becomes infinite, even if condition (10) is fulfilled. Thus, we need assumptions like (26) to exclude such unusual cases.

For better understanding, we directly analyze the off-diagonal terms of (14)

$$\begin{aligned}
&\mathrm{E}\left[\sum_{t=1}^{T}\varphi_i\{y_i(t)\}y_j(t)\,\middle|\,B,\rho_z,\rho_V\right]\\
&=\sum_{t=1}^{T}\mathrm{E}\left[\,\mathrm{E}\left[\varphi_i\{v_i(t)z_i(t)\}\;v_j(t)z_j(t)\,\middle|\,V;B,\rho_z\right]\,\middle|\,\rho_V\right]\\
&=\sum_{t=1}^{T}\mathrm{E}\left[\,\mathrm{E}\left[\varphi_i\{v_i(t)z_i(t)\}\,|V;B,\rho_z\right]\mathrm{E}\left[v_j(t)z_j(t)\,\middle|\,V;B,\rho_z\right]\,\middle|\,\rho_V\right]\;=\;0.
\end{aligned}$$

The second equality follows from the fact that z_i and z_j are independent for fixed V.[6]

6. This unbiasedness in fact holds under a wider condition $\mathrm{E}[s_i(t)|s_j(\cdot),\ j\neq i]=0$.

To prove condition (15) and compute the asymptotic variance (18), we calculate the $n^2 \times n^2$ matrix Q. If we use the non-holonomic basis $d\chi = dBB^{-1}$ (Amari et al., 2000), Q is expressed as

$$Q = \mathrm{E}\left[\left. \frac{\partial \mathrm{vec}(\bar{F}^{\mathrm{QML}})}{\partial \mathrm{vec}(\chi)} \right| B, \rho_z, \rho_V \right] \bar{B}^{-1},$$

where $\bar{B} = (\bar{B}_{ij;kl})$ and $\bar{B}_{ij;kl} = \delta_{ik} b_{lj}$. Fortunately, the matrix $\mathrm{E}\left[\frac{\partial \mathrm{vec}(\bar{F}^{\mathrm{QML}})}{\partial \mathrm{vec}(\chi)}\right]$ turns out to have a simple structure such that only the following $2n^2 - n$ components are non-zero,

$$\begin{aligned}
\mathrm{E}\left[\frac{\partial \bar{F}_{ii}^{\mathrm{QML}}}{\partial \chi_{ii}}\right] &= -\sum_{t=1}^{T} \mathrm{E}\left[m_i\{v_i(t)\} \right] - T, \\
\begin{pmatrix} \mathrm{E}\left[\frac{\partial \bar{F}_{ij}^{\mathrm{QML}}}{\partial \chi_{ij}}\right] & \mathrm{E}\left[\frac{\partial \bar{F}_{ij}^{\mathrm{QML}}}{\partial \chi_{ji}}\right] \\ \mathrm{E}\left[\frac{\partial \bar{F}_{ji}^{\mathrm{QML}}}{\partial \chi_{ij}}\right] & \mathrm{E}\left[\frac{\partial \bar{F}_{ji}^{\mathrm{QML}}}{\partial \chi_{ji}}\right] \end{pmatrix} &= -\begin{pmatrix} \sum_{t=1}^{T} \mathrm{E}[k_i\{v_i(t)\} v_j^2(t)] & T \\ T & \sum_{t=1}^{T} \mathrm{E}[k_j\{v_j(t)\} v_i^2(t)] \end{pmatrix},
\end{aligned}$$

in which we employed the following quantities

$$\begin{aligned}
k_i\{v_i(t)\} &= \mathrm{E}\left[\dot{\varphi}_i\{v_i(t) z_i(t)\} \,|V; B, \rho_z \right], \\
m_i\{v_i(t)\} &= v_i^2(t)\, \mathrm{E}\left[\dot{\varphi}_i\{v_i(t) z_i(t)\} z_i^2(t) \,\middle|V; B, \rho_z \right],
\end{aligned}$$

and $\dot{\varphi}_i$ is the derivative of φ_i. Hence, it is not difficult to check non-singularity of this matrix, and if this is the case, the condition (15) holds. We can also explicitly calculate the inverse matrix $Q^{-1} = \bar{B}\left(\mathrm{E}\left[\frac{\partial \mathrm{vec}(\bar{F}^{\mathrm{QML}})}{\partial \mathrm{vec}(\chi)}\right]\right)^{-1}$ that appears in the asymptotic variance (18), because we only have to invert the 2×2 matrices.

Finally, the variance of the estimating function can be computed as

$$\begin{aligned}
&\mathrm{E}\left[\bar{F}_{ij}^{\mathrm{QML}} \bar{F}_{kl}^{\mathrm{QML}} | B, \rho_z, \rho_V \right] \\
&= \begin{cases} \sum_{t,t'} \mathrm{cov}\left[\varphi_i\{y_i(t)\} y_i(t),\ \varphi_k\{y_k(t')\} y_k(t') \right], & i = j,\ k = l \\ \sum_{t} \mathrm{E}\left[\varphi_i\{y_i(t)\} \varphi_k\{y_k(t)\} y_j^2(t) \right], & j = l,\ i \neq j \text{ or } k \neq l \\ \sum_{t} \mathrm{E}\left[\varphi_i\{y_i(t)\} y_i(t) \varphi_j\{y_j(t)\} y_j(t) \right], & i = l,\ j = k,\ i \neq j \end{cases}
\end{aligned}$$

which is slightly more complicated than the standard ICA model. Summing up the discussion above, we get the following theorem.

Theorem 3 *Suppose that the conditions*

$$\sum_{t=1}^{T} \mathrm{E}\left[m_i\{v_i(t)\} \right] + T \neq 0, \qquad \forall i, \tag{27}$$

$$\det \begin{pmatrix} \sum_{t=1}^{T} \mathrm{E}[k_i\{v_i(t)\} v_j^2(t)] & T \\ T & \sum_{t=1}^{T} \mathrm{E}[k_j\{v_j(t)\} v_i^2(t)] \end{pmatrix} \neq 0, \qquad \forall i \neq j, \tag{28}$$

and assumption (26) hold. Then the function $\bar{F}^{QML}(X,B)$ *satisfies the conditions (14) – (16) and becomes an estimating function. In that case, the quasi maximum likelihood estimator* $\widehat{B}^{\mathrm{QML}}$ *derived from the equation* $\bar{F}^{QML}(X,\widehat{B}^{\mathrm{QML}})=0$ *is consistent regardless of the true nuisance functions* (ρ_z^*,ρ_V^*) *under appropriate regularity conditions.*

5.2 Stability of the Natural Gradient Learning

In neural networks and machine learning, online leaning is often preferred to batch learning because of computational efficiency, less memory and adaptability (see, for example, Müller et al., 1998; Murata et al., 2002). The natural gradient learning (Amari, 1998)

$$B(t+1)=B(t)+\eta(t)\left[I-\varphi\{y(t)\}y^{\top}(t)\right]B(t), \tag{29}$$

is an online algorithm based on the quasi maximum likelihood method, where $y(t)=B(t)x(t)$ is the current estimator of the sources and $\eta(t)$ is an appropriate learning constant.

Following the discussion in Amari et al. (1997), we will study the stability of the natural gradient learning for the variance-dependent BSS model. For the sake of simplicity, they analyzed a continuous version of the algorithm (29)

$$\dot{B}(t)=\mu(t)\left[I-\varphi\{y(t)\}y^{\top}(t)\right]B(t), \tag{30}$$

where $\dot{B}(t)$ denotes time derivative of the matrix $B(t)$, $\mu(t)=\eta(t)/\tau$ and τ means the sampling period. Suppose that the marginal distributions of the activity levels $v(t)$ are identical in time. For example, when the sequence V is generated from an AR process, this holds approximately after it reaches the equilibrium distribution. Although the random variables $v(t)$'s (activity levels) have an identical marginal distribution in time, their realization can fluctuate from time to time and weak nonstationary structures can be found in the observed signals. Unfortunately, it is difficult to eliminate this rather strong assumption. If we apply the online algorithm (29) to data with highly nonstationary variances like speech, the scale factor of the demixing matrix B changes substantially from time to time and never converges. This makes the current stability analysis impossible. It might be possible to discuss these cases by considering only the equivalence class, but it is out of the scope of the current paper.

In order to fix the scales of the sources, we impose constraints

$$\mathrm{E}\left[\varphi_i\{b_i^{\top}x(t)\}\ b_i^{\top}x(t)\right]=1, \qquad \forall i. \tag{31}$$

Note that the marginal distribution of $x(t)$ is identical in time t and the equilibrium points B_0 of the equation (30) satisfy

$$\mathrm{E}\left[I-\varphi\{y_0(t)\}y_0^{\top}(t)\right]=0, \tag{32}$$

where $y_0(t)=B_0x(t)$. With a similar calculation as in Section 5.1, we can show that the function

$$F^{\mathrm{NG}}(x,B)=I-\varphi(y)y^{\top}$$

satisfies the unbiasedness condition (8) of estimating functions. This means that the true demixing matrix B^* satisfies the equilibrium equation (32), that is, B^* becomes an equilibrium point of the flow (30). However, it does not guaranteed that $B(t)$ converges to B^* even locally.

Let us fix the stochastic process $V = \{v(t),\, t \geq 0\}$ of the activity levels at first and consider the conditional expected version of the learning equation

$$\dot{B}(t) = \mu(t)\, \mathrm{E}\left[\, I - \varphi\{y(t)\} y^{\top}(t) \,\middle|\, V \right]\, B(t).$$

By linearizing it at the equilibrium point B^*, we have the variational equation

$$\mathrm{vec}\{\delta\dot{B}(t)\} = \mu(t)\, \frac{\partial \mathrm{vec}\{\, \mathrm{E}\left[F^{\mathrm{NG}}(x(t), B^*) \,\middle|\, V \right] B^*\}}{\partial \mathrm{vec}(B)}\, \mathrm{vec}\{\delta B(t)\},$$

where $\delta B(t)$ is a small perturbation. Therefore, we have to check the eigenvalues of the operators $\frac{\partial \mathrm{vec}\{\, \mathrm{E}\left[F^{\mathrm{NG}}(x(t),B^*) \,\middle|\, V\right] B^*\}}{\partial \mathrm{vec}(B)}$ for each $t \geq 0$. If all eigenvalues have negative real parts, then the equilibrium B^* is asymptotically stable for the fixed activity levels V. Since the matrix can be expressed as

$$\frac{\partial \mathrm{vec}\{\, \mathrm{E}\left[\, F^{\mathrm{NG}}(x(t), B^*) \,\middle|\, V \right] B^*\}}{\partial \mathrm{vec}(B)} = \bar{B}^*\, \frac{\partial \mathrm{vec}\left(\mathrm{E}\left[F^{\mathrm{NG}} \middle| V\right]\right)}{\partial \mathrm{vec}(\chi)}\, (\bar{B}^*)^{-1}, \tag{33}$$

where $\bar{B}^* = (\bar{B}^*_{ij;kl}) = (\delta_{ik} b^*_{lj})$, and derivative w.r.t. χ corresponds to the non-holonomic basis $\mathrm{d}\chi = \mathrm{d}B B^{-1}$. Because the left hand side of (33) is a similar transformation of $\frac{\partial \mathrm{vec}(\mathrm{E}[F^{\mathrm{NG}}|V])}{\partial \mathrm{vec}(\chi)}$, their eigenvalues are the same. Fortunately, as is the case of the quasi maximum likelihood, the matrix $\frac{\partial \mathrm{vec}(\mathrm{E}[F^{\mathrm{NG}}|V])}{\partial \mathrm{vec}(\chi)}$ has a simple structure such that only the following $2n^2 - n$ components are non-zero,

$$\begin{aligned}
\frac{\partial \mathrm{E}[F^{\mathrm{NG}}_{ii}|V]}{\partial \chi_{ii}} &= -m_i\{v_i(t)\} - 1 \\
\begin{pmatrix} \dfrac{\partial \mathrm{E}[F^{\mathrm{NG}}_{ij}|V]}{\partial \chi_{ij}} & \dfrac{\partial \mathrm{E}[F^{\mathrm{NG}}_{ij}|V]}{\partial \chi_{ji}} \\ \dfrac{\partial \mathrm{E}[F^{\mathrm{NG}}_{ji}|V]}{\partial \chi_{ij}} & \dfrac{\partial \mathrm{E}[F^{\mathrm{NG}}_{ji}|V]}{\partial \chi_{ji}} \end{pmatrix} &= -\begin{pmatrix} k_i\{v_i(t)\}\, v_j^2(t) & 1 \\ 1 & k_j\{v_j(t)\}\, v_i^2(t) \end{pmatrix}
\end{aligned}$$

Therefore, the matrix $\frac{\partial \mathrm{vec}(\mathrm{E}[F^{\mathrm{NG}}|V])}{\partial \mathrm{vec}(\chi)}$ at time t has eigenvalues only with negative real parts, if and only if

$$m_i\{v_i(t)\} + 1 > 0 \tag{34}$$

$$k_i\{v_i(t)\} > 0 \tag{35}$$

$$v_i^2(t)\, v_j^2(t)\, k_i\{v_i(t)\}\, k_j\{v_j(t)\} > 1 \tag{36}$$

for all i, j $(i \neq j)$.

Theorem 4 *If the stochastic process $V = \{v(t),\, t \geq 0\}$ of the activity levels satisfies the conditions (34) – (36) with probability* 1 *as for the true parameter $(B^*, \rho_z^*, \rho_V^*)$, then the true demixing matrix B^* becomes an asymptotically stable equilibrium of the flow (30) with probability* 1.

Although asymptotic stability could be proved under weaker conditions, we summarize the discussion as Theorem 4 for simplicity. In order to understand the result better, we revisit the

examples presented in Amari et al. (1997). The conditions turn out to be much harder than those by Amari et al. (1997) because of the fluctuating activity levels.

Example 1. Let us consider the following odd activation function

$$\varphi_i(y_i) = |y_i|^p \mathrm{sign}(y_i)$$

for $p = 1, 2, \ldots$. The conditions (34) and (35) are automatically satisfied for any fixed $v_i(t) > 0$.

$$\begin{aligned} m_i\{v_i(t)\} &= p\, v_i^{p+1}(t)\, \mathrm{E}\left[\, |z_i(t)|^{p+1}\right] > 0 \\ k_i\{v_i(t)\} &= p\, v_i^{p-1}(t)\, \mathrm{E}\left[\, |z_i(t)|^{p-1}\right] > 0 \end{aligned}$$

The condition (36) becomes

$$p^2\, v_i^{p+1}(t)\, v_j^{p+1}(t)\, \mathrm{E}\left[\, |z_i(t)|^{p-1}\right]\, \mathrm{E}\left[\, |z_j(t)|^{p-1}\right] > 1.$$

By introducing Gray's norm

$$\gamma_{pi} = \frac{\mathrm{E}[\, |z_i|^{p+1}]}{\mathrm{E}[\, |z_i|^2]\, \mathrm{E}[\, |z_i|^{p-1}]}$$

and taking notice of the normalization constraints (31), that is, $\mathrm{E}[\, |z_i|^{p+1}] = \left(\mathrm{E}[\, v_i^{p+1}]\right)^{-1}$, finally we obtain

$$\gamma_{pi}\gamma_{pj} < p^2\, \frac{\min\limits_t v_i^{p+1}(t)}{\mathrm{E}[\, v_i^{p+1}]}\, \frac{\min\limits_t v_j^{p+1}(t)}{\mathrm{E}[\, v_j^{p+1}]}.$$

For the cubic function $\varphi_i(y_i) = y_i^3$, not as in the ICA model, the condition that all signals are sub-Gaussian

$$\gamma_{3i} = \frac{\mathrm{E}[\, |z_i|^4]}{\left(\mathrm{E}[\, |z_i|^2]\right)^2} < 3$$

is not enough, but the variation of activity levels v_i from (1) should be taken into account.

Example 2. Let us consider a symmetrical sigmoidal function

$$\varphi_i(y_i) = \tanh(\beta y_i).$$

The conditions (34) and (35) can be checked easily. Unfortunately, in this case we can only do a rather coarse analysis as follows. Let us assume $\beta \ll 1$ so that the approximation

$$\varphi_i(y_i) \approx \beta y_i - \frac{1}{3}(\beta y_i)^3 + \frac{2}{15}(\beta y_i)^5$$

holds with high probability. Then, we can express the condition (36) as

$$\beta^2 v_i^2(t)\, v_j^2(t)\, \mathrm{E}\left[1 - (\beta y_i)^2 + \frac{2}{3}(\beta y_i)^4 \middle| V\right]\, \mathrm{E}\left[1 - (\beta y_j)^2 + \frac{2}{3}(\beta y_j)^4 \middle| V\right] > 1.$$

Because $1 - t^2 + 2t^4/3 > t^4/3$, we get a stronger condition

$$\frac{\beta^{10}}{9}\, v_i^2(t)\, v_j^2(t)\, \mathrm{E}\left[\, y_i^4 \middle| V\right]\, \mathrm{E}\left[\, y_j^4 \middle| V\right] > 1.$$

From a rough approximation of (31), the relation $\beta \approx \left(\mathrm{E}[v_i^2]\right)^{-1}$ is derived. Therefore, if all approximations are accurate enough, we finally get a sufficient condition of (36) like

$$\gamma_{3i}\gamma_{3j} > \frac{9}{\beta^4}\left(\frac{\mathrm{E}[v_i^2]}{\min_t v_i^2}\right)^3\left(\frac{\mathrm{E}[v_j^2]}{\min_t v_j^2}\right)^3.$$

In contrast to the ordinal ICA model without variance dependence, the condition that all signals are super-Gaussian may not be enough, but each kurtosis γ_{3i} should be much larger than 3.[7]

5.3 Properties of Other BSS Algorithms

Although we concentrated on estimating functions of the form (20), we can deal with more general functions and investigate other ICA algorithms within the framework of estimating functions and asymptotic estimating functions (see also Cardoso, 1997). Such analysis helps to check whether these algorithms may give valid solutions regardless of the nuisance densities (ρ_z, ρ_V). We remark that our extension enables us to analyze algorithms based on temporal structure such as TDSEP/SOBI (Ziehe and Müller, 1998; Belouchrani et al., 1997). Since it is quite technical, the detailed discussion is put in Appendix A, where the unbiasedness condition (14) of estimating functions is examined for these algorithms under the variance-dependent BSS model. We briefly summarize the consequences in Table 2. Estimators by all algorithms listed below are derived from estimating equations which satisfy the unbiasedness condition at least asymptotically. When the other conditions are taken into account, TDSEP/SOBI never works for the variance-dependent BSS model, because sources have no lagged auto-correlations. ICA algorithms using non-Gaussianity such as FastICA and JADE are not working, if sources are Gaussian. The double blind algorithm (Hyvärinen and Hurri, 2004) cannot be applied to the case where the variance structures of sources are the same or there is no temporal variance-dependency. The nonstationary algorithm by Pham and Cardoso (2000) is not applicable to the case where time courses of the activity levels are proportional to each other. Of course, such a theoretical analysis tells us only about the possibility of failure. In practice, algorithms do not always return valid answers, because of local minima and numerical instability of their learning process. Nevertheless, this theoretical analysis can explain the results of our numerical experiments in the next section.

6. Numerical Experiments

We carried out experiments with several artificial and more realistic data sets for several BSS algorithms. The eight batch algorithms and the online versions of the quasi maximum likelihood methods listed in Table 3 were applied to those data sets. Note that our goal is not primarily an algorithm comparison but the experiments serve to demonstrate the correctness of our theoretical analysis.

For evaluating the results, we used the index defined by Amari et al. (1996)

$$\mathrm{AmariIndex}(B, A^*) = \sum_{i=1}^{n}\left\{\frac{\sum_{j=1}^{n}|C_{ij}|}{\max_k |C_{ik}|} - 1\right\} + \sum_{j=1}^{n}\left\{\frac{\sum_{i=1}^{n}|C_{ij}|}{\max_k |C_{kj}|} - 1\right\},$$

7. This different result corrects a calculation in Amari et al. (1997).

algorithm	unbiasedness	unavailable cases
FastICA Hyvärinen (1999)	yes	Sources are Gaussian.
double blind Hyvärinen and Hurri (2004)	asymptotically	Variance structures are same or there is no temporal variance-dependency.
JADE Cardoso and Souloumiac (1993)	asymptotically	Sources are Gaussian.
TDSEP/SOBI Ziehe and Müller (1998) Belouchrani et al. (1997)	yes	always (since we consider here only the case without auto-correlations)
nonstationary Pham and Cardoso (2000)	yes	Time course of the activity levels are proportional to each other.

Table 2: Availability of other ICA and BSS algorithms

QML(tanh)	quasi maximal likelihood method with the hyperbolic tangent nonlinearity
QML(pow3)	quasi maximal likelihood method with the cubic nonlinearity
Online(tanh)	online version of QML(tanh) with learning rate $\eta(t) = \frac{0.1}{(1+t/20)}$
Online(pow3)	online version of QML(pow3) with learning rate $\eta(t) = \frac{0.25}{(1+t/20)}$
'DoubleBlind'	the double blind algorithm by Hyvärinen and Hurri (2004)
JADE	JADE algorithm
FastICA(tanh)	FastICA with the hyperbolic tangent nonlinearity
FastICA(pow3)	FastICA with the cubic nonlinearity
TDSEP/SOBI	TDSEP/SOBI algorithm
'Sepagaus'	The 'sepagaus' algorithm for nonstationary signals by Pham and Cardoso (2000)

Table 3: ICA and BSS algorithms used in the experiments

where A^* is the true mixing matrix and $C = BA^*$. If $B = PD(A^*)^{-1}$ with a permutation matrix P and a diagonal matrix D, then $\mathrm{AmariIndex}(B, A^*) = 0$.

6.1 Artificial Data Sets

In all artificial data sets, five source signals of various types with length $T = 10000$ were generated and data after multiplying a random 5×5 mixing matrix were observed. We made 100 replications for each setting and compute the demixing matrix for each replication. The first data set was made according to the experiments in Hyvärinen and Hurri (2004). The activity levels $v(t)$ were generated from a multivariate AR(1) model, where outliers larger than three times standard deviations from the means were reduced to these bounds. The normalized signals z_i's were i.i.d. sub-Gaussian random variables which are signed fourth-order roots of zero-mean uniform variables. The medians of the 100 replications are summarized in the row 'ar_subG' of Table 4 with the measure of deviation (3rd-quantile $-$ 1st-quantile)/2. As was pointed out by Hyvärinen and Hurri (2004), only 'DoubleBlind' gave small AmariIndex. Because the marginal distribution of the source signal $s_i(t)$ looks like a Gaussian, all algorithms based on indices favouring non-Gaussianity failed. Even though the determinant in the left hand side of (28) is close to 0, all the assumptions are satisfied and the local consistency theorem is still valid. However, this does not directly mean that the estimated demixing matrix converges globally to the true one. In this case, many local optima can make the algorithms fail. This could also be understood from the fact that the contrast functions based on non-Gaussianity become almost flat and thus are very difficult to optimize. In the experiments, we observed that part of the true sources were often extracted correctly.

Although all the algorithms except for 'DoubleBlind' did not work for the first difficult example, the theoretical study in principle tells that many ICA and BSS algorithms are also applicable to the variance-dependent BSS problem. So in fact the failure of the algorithms except 'DoubleBlind' can be solely explained by the particular choice of the data set which is in contrast to prior findings in Hyvärinen and Hurri (2004). In the second example, uniform random variables were used as z_i's instead of sub-Gaussian ones. The marginal distribution of the source signal $s_i(t)$ looks Laplacian. Therefore, as was shown in the row 'ar_uni' of Table 4, the algorithms QML(tanh) and FastICA(tanh), which are suitable for super-Gaussian sources, always give correct answers. The algorithms 'DoubleBlind', JADE and FastICA(pow3) based on 4-th order moments also worked except several failures due to outliers. We got admissible results by the nonstationary BSS algorithm 'Sepagaus', if an appropriate smoothing window was chosen.

In the third and the fourth data, the activity levels $v_i(t)$ are sinusoidal functions with different frequencies.

$$v_i(t) = 1 + 0.9 \sin\left(\frac{(13+i)\pi t}{8000}\right), \qquad i = 1, \ldots, 5$$

For the normalized signals z_i, Laplacian and the sub-Gaussian i.i.d. random variables were used in the third and the fourth examples, respectively. In the super-Gaussian case (the row 'sin_supG' of Table 4), the six algorithms except QML(pow3) and TDSEP/SOBI worked properly. 'Sepagaus' showed best performance, and QML(tanh) and FastICA(tanh) based on the hyperbolic tangent nonlinearity gave better results than 'DoubleBlind', JADE and FastICA(pow3) with 4-th order moments. On the other hand, in the sub-Gaussian case (the row 'sin_subG' of Table 4), the six algorithms except QML(tanh) and TDSEP/SOBI returned admissible results. 'Sepagaus' also showed

	QML(tanh)	QML(pow3)	Online(tanh)	Online(pow3)	'DoubleBlind'
ar_subG	8.25 (1.85)	11.32 (2.84)	14.59 (2.29)	14.75 (2.32)	0.52 (0.10)
ar_uni	0.30 (0.04)	27.77 (0.32)	0.51 (0.08)	23.40 (1.88)	0.70 (0.16)
sin_supG	0.17 (0.02)	29.97 (0.26)	0.39 (0.05)	28.74 (0.96)	0.79 (0.13)
sin_subG	19.21 (0.24)	0.32 (0.05)	21.51 (2.08)	0.57 (0.30)	0.27 (0.03)
com_supG	0.39 (0.06)	28.37 (0.27)	0.64 (0.09)	25.67 (1.74)	6.45 (1.56)
com_subG	26.53 (0.55)	0.14 (0.02)	27.00 (2.41)	0.28 (0.05)	22.05 (1.96)
exp_supG	0.35 (0.05)	28.43 (0.45)	0.59 (0.07)	22.84 (2.06)	7.63 (1.88)
uni_subG	27.38 (0.17)	0.13 (0.02)	27.24 (1.27)	0.27 (0.04)	18.56 (1.66)
sss	0.03	3.82	0.06 (0.01)	2.79 (0.53)	0.02
v12	0.01	3.73	0.06	2.89 (0.04)	0.21

	JADE	FastICA(tanh)	FastICA(pow3)	TDSEP/SOBI	'Sepagaus'
ar_subG	10.79 (1.88)	9.25 (1.98)	12.52 (2.05)	15.07 (1.96)	1.19 (0.48)
ar_uni	0.66 (0.14)	0.38 (0.05)	0.73 (0.14)	14.92 (2.37)	0.85 (0.22)
sin_supG	0.43 (0.07)	0.23 (0.03)	0.41 (0.07)	15.31 (2.04)	0.08 (0.01)
sin_subG	0.31 (0.04)	0.68 (0.14)	0.33 (0.05)	15.70 (1.94)	0.08 (0.01)
com_supG	0.84 (0.16)	0.48 (0.07)	0.87 (0.14)	16.02 (2.05)	1.28 (0.19)
com_subG	26.49 (0.86)	27.04 (0.38)	26.65 (0.17)	16.23 (2.01)	27.08 (0.40)
exp_supG	1.24 (0.23)	0.44 (0.06)	1.20 (0.22)	16.47 (1.81)	1.28 (0.20)
uni_subG	0.17 (0.03)	0.18 (0.03)	0.18 (0.03)	16.20 (1.78)	27.08 (0.33)
sss	0.02	0.19 (0.04)	0.09 (0.01)	0.01	0.01
v12	0.19	0.17 (0.02)	0.08 (0.09)	0.14	0.01

Table 4: AmariIndex of the estimators. The values are the medians of 100 replications with the measure of deviation, $(\text{3rd-quantile} - \text{1st-quantile})/2$

best performance, and all four algorithms with 4-th order moments showed better performance than the FastICA(tanh).

The double blind algorithm ('DoubleBlind') by Hyvärinen and Hurri (2004) does not work when (i) all v_i's have same temporal structure, and (ii) there exist no temporal dependencies in v_i's. 'Sepagaus' does not have a guarantee to separate sources either, because smoothed sequences of the activity levels are nearly proportional to each other (see Table 2). The fifth and sixth data set are examples of the case (i), where $v_i(t)$ are the same sinusoidal functions.

$$v_i(t) = 1 + 0.9 \sin\left(\frac{\pi t}{500}\right), \qquad i = 1, \ldots, 5$$

As in the third and the fourth examples, Laplace and the sub-Gaussian i.i.d. random variables were used for the normalized signals z_i. As in the row 'com_supG' of Table 4, the five algorithms except QML(pow3), 'DoubleBlind' and TDSEP/SOBI worked properly. Among them, QML(tanh) and FastICA(tanh) had better performance. 'DoubleBlind' gave poor results, because the matrix $\widetilde{K}_{ij} = \widehat{\mathrm{cov}}\{s_i^2(\cdot), s_j^2(\cdot - 1)\}$ is almost singular. In the sub-Gaussian case, it looks quite difficult to distinguish the sources visually. Unfortunately, we could not demix them correctly except with QML(pow3) as shown in the row 'com_subG' of Table 4. In order to check why other algorithms

with the local consistency did not work, we carried out extra experiments with larger sample size T. When $T = 200000$, the AmariIndices of the estimated demixing matrices by JADE are below 0.11, 92 times out of 100 repetition. On the other hand, both FastICA methods returned valid results almost always (AmariIndices are below 0.22, 89 times for FastICA(tanh) and 100 times for FastICA(pow3)), if $T = 50000$ and the algorithms start from the true demixing matrix. Therefore, we think that the global convergence is not achieved in these cases, because of finite sample size effects and local optima.

The seventh and the eighth data sets are examples of the case (ii), where $v(t)$ is i.i.d. in time t. In the former example, we transform 5 independent exponential random variables linearly such that v_i and v_j have correlation 0.9, and z_i's were i.i.d Laplace random variables. On the other hand, in the latter example, $v(t)$ was generated from 5 uniform random variables by the same linear transformation and z_i's were the i.i.d sub-Gaussian random variables. As one can see in the row 'exp_supG' of Table 4, the results are similar to the data set 'com_supG'. On the other hand, in the sub-Gaussian case summarized in the row 'uni_subG' of Table 4, QML(pow3), JADE, FastICA(tanh) and FastICA(pow3) gave correct results, but 'Sepagaus' showed very poor performance. We remark that in both cases, 'DoubleBlind' did not work as was expected.

We would now like to digest the results from Table 4 and relate them to our theoretical findings. We have shown that all algorithms except for TDSEP/SOBI have the local consistency for most of the given data. However, this does not directly mean that they converge globally to the true solution. Although we hope that algorithms with a local consistency work properly, we sometimes see significant deviations from this expectation in practice as in Table 4. The algorithmic failures are caused by local optima as pointed out above for the data set 'ar_subG', or more importantly to numerical stability and convergence issues. For example, since learning algorithms like gradient descent are used for QML(tanh) and QML(pow3), desired solutions (equilibria) turn out to be instable for sub-Gaussian (QML(tanh)) and super-Gaussian signals (QML(pow3)). In our data sets, 'ar_uni', all data sets with 'supG' and acoustic signals are super-Gaussian, while all data sets with 'subG' except 'ar_subG' are sub-Gaussian. One can see the clear pattern in the columns QML(tanh) and QML(pow3). The online version Online(tanh) and Online(pow3) had slightly degraded performance with appropriate learning rate, if the batch version QML(tanh) and QML(pow3) worked, respectively. On the other hand, although FastICA uses similar criteria for non-Gaussianity, it employs a kind of Newton's method and so the desired solutions are automatically better stabilized. In the columns FastICA(tanh) and FastICA(pow3), except for the difficult case 'com_subG' and nearly Gaussian case 'ar_subG', both algorithms succeeded.

6.2 Variance-Dependent Speech Signals

Next we will deal with more realistic data sets. Speech and audio signals have often been used as sources $s(t)$ even for experiments of the instantaneous ICA model. In order to check whether variance-dependency matters to many ICA and BSS algorithms, we applied BSS algorithms to speech signals which have strong variance-dependency.

In the first experiments 'sss',[8] we took two speech signals with length $T = 120976$, where one speaker says digits from 1 to 10 in English, and the other speaker counts at the same time in Spanish. We used the separated signals of their second demo as the sources, because their separation quality is good enough. Figure 2 shows the sources and the estimators of their activity levels with

8. The signals were downloaded from `http://inc2.ucsd.edu/~tewon/`.

an appropriate smoother. We inserted one short pause at different positions of both sequences to make correlation of the activity levels of the modified signals much larger (0.65). In the second experiments 'v12',[9] we took two speech signals from Japanese text ($T = 48000$). Figure 3 shows the sources and the estimators of the activity levels. We extended and shorten each syllable of the second sequence and tuned its amplitude such that the two sources have high variance-dependency. Correlation of the activity levels of the arranged signals becomes 0.74.

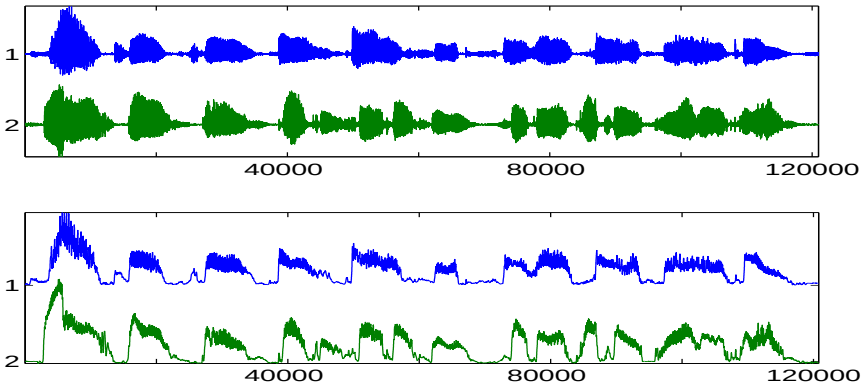

Figure 2: The sources of the data set 'sss' and the estimators of their activity levels. The upper panel contains the signals showing counting from 1 to 10 in English and Spanish. The lower panel shows their activity levels with an appropriate smoother.

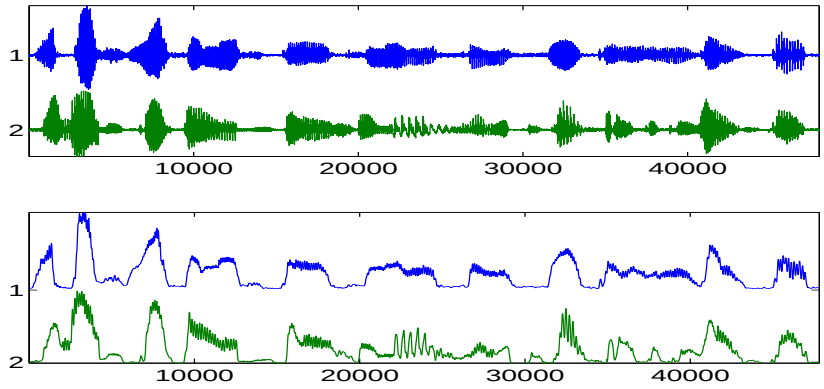

Figure 3: The sources of the data set 'v12' and the estimators of their activity levels. The upper panel are signals from Japanese sentences. The lower panel shows their activity levels with an appropriate smoother.

A 2×2 mixing matrix A was randomly generated 100 times and 100 different mixtures of the source signals were made. The results are summarized in the rows 'sss' and 'v12' of Table 4. In

9. The signals can be downloaded by `http://www.islab.brain.riken.go.jp/~mura/ica/v1.wav` and `v2.wav`.

both experiments, QML(tanh), JADE, TDSEP/SOBI and 'Sepagaus' always worked, while FastICA(tanh) and FastICA(pow3) gave admissible results except for several cases. Although TDSEP/SOBI is not applicable to the variance-dependent BSS model, it also returned correct results. This means that the speech signals are not perfectly matching the model Eq. (4), but the sources have furthermore a lagged autocorrelation. QML(tanh) always returned wrong answers, because speech is usually super-Gaussian.

7. Conclusions

In this paper, we discussed semiparametric estimation for blind source separation, when sources have variance dependencies. Hyvärinen and Hurri (2004) introduced the double blind setting where, in addition to source distributions, dependencies between components are not restricted by any parametric model. In the presence of these two nuisance parameters (densities of activity level and underlying signal), they proposed an algorithm based on lagged 4-th order cumulants. Although their algorithm works well in many cases, it fails if (i) all v_i's have similar temporal structure, or (ii) there exist no temporal dependencies in v_i's. Furthermore it also suffers from outliers.

Extending the semiparametric approach (Amari and Cardoso, 1997) under variance dependencies, we investigated estimating functions for the variance-dependent BSS model. In particular, we proved that the quasi maximum likelihood estimator is derived from an estimating function, and is hence consistent regardless of the true nuisance densities (which satisfy certain mild conditions). We also analyzed other ICA algorithms within the framework of (asymptotic) estimating functions and showed that many of them can separate sources with coherent variances. This is in contrast to previous understanding of the mechanisms underlying ICA algorithms. Theoretically we have shown that at least asymptotically all BSS algorithms except for TDSEP/SOBI have the local consistency, thus they should succeed on a given mixed data. However, local consistency does not necessarily guarantee global convergence to the true solution and we sometimes see significant deviations from this expectation in practice. The algorithmic failures are due to many local optima and more importantly due to numerical stability and convergence issues.

Although almost all ICA and BSS algorithms could not give correct answers in the numerical experiment of Hyvärinen and Hurri (2004), we showed here that this was mainly a matter of the specific choice of the data set. In fact, most ICA and BSS algorithms also work well in many other benchmark examples that use dependent data. In particular, we carried out two experiments with highly variance-dependent speech signals. Despite the dependence typically found in speech, most ICA and BSS algorithms yield excellent separation results and our theoretical analysis can help to understand the reason for this fact. We conjecture that it is not the coarse amplitude structure (e.g. from dependence) that matters for BSS but the statistical fine structure of the signals.

In this paper, we only tested existing ICA and BSS algorithms and pointed out that some of them are applicable to the variance-dependent BSS model. Future research will go one step further and construct more efficient or robust semiparametric algorithms. Note also that in practice, it is important to analyze how to select the best BSS method for a specific, say, variance-dependent data

set. We think that suitable methods might be developed along the lines of Meinecke et al. (2002) or Harmeling et al. (2004).

Acknowledgments

The authors acknowledge A. Ziehe, S. Harmeling, F. Meinecke and N. Murata for valuable discussions, and A. Hyvärinen and three anonymous reviewers for useful comments to improve this paper. We furthermore thank the PASCAL Network of Excellence (EU #506778) and DFG (SFB 618) for partial funding.

Appendix A. Comments on Other Selected BSS Algorithms

We will discuss in the following the local consistency of ICA/BSS algorithms except the quasi maximum likelihood method.

A.1 FastICA

FastICA is one of the standard algorithms for blind source separation. Let $u(t) = C^{-1/2}x(t)$ be the whitened data, where $C = \frac{1}{T}\sum_{t=1}^{T} x(t)x^\top(t)$ is the sample covariance. FastICA gives the demixing matrix $W = (w_1, \ldots, w_n)^\top$ which maximizes the total non-Gaussianity

$$\sum_{i=1}^{n} \frac{1}{T} \sum_{t=1}^{T} G\{w_i^\top u(t)\}$$

under the orthogonality condition $WW^\top = I$. We use, in the following the notation W for the demixing matrix after whitening in order to distinguish it from the total demixing matrix $B = WC^{-1/2}$ including whitening process. Here G is a nonlinear function which is introduced to approximate the negentropy (Hyvärinen et al., 2001b). By solving the constrained optimization problem, we see that the estimator of W must satisfy the estimating equation

$$\sum_{t=1}^{T} \Big[y(t)y^\top(t) - I + y(t)g^\top\{y(t)\} - g\{y(t)\}y^\top(t) \Big] = 0 \tag{37}$$

where $y(t) = Wu(t)$. If we write the total demixing matrix as $B = WC^{-1/2}$, $y(t)$ can be expressed as $Bx(t)$. The vector function $g(y)$ consists of the derivatives $g(y_i) = G'(y_i)$, that is, $g(y) = (g(y_1), \ldots, g(y_n))^\top$. The functions (12) and (13) are also used as the function g. We remark that the equation (37) is equivalent to

$$\sum_{t=1}^{T} \Big[y(t)g^\top\{y(t)\} - g\{y(t)\}y^\top(t) \Big] = 0, \tag{38}$$

$$\sum_{t=1}^{T} \Big[y(t)y^\top(t) - I \Big] = 0, \tag{39}$$

because the left hand side of (38) is antisymmetric, while that of (39) is symmetric. If we determine the scales of the sources such that

$$\mathrm{E}\left[\sum_{t=1}^{T}\{b_i^\top x(t)\}^2 \middle| B,\rho_z,\rho_V\right] = T, \qquad i=1,\ldots,n, \tag{40}$$

then it is easy to show that the expectations of the left hand side of (38) and (39) vanish regardless of the nuisance functions ρ_z and ρ_V, in the same way as for the quasi maximum likelihood method. This means that the left hand side of (37) satisfies the unbiasedness condition (14) of estimating functions. If the other regularity conditions hold, it becomes an estimating function and the estimator $\widehat{B}$ derived from it converges to the correct demixing matrix $B^* = (A^*)^{-1}$ with a permutation matrix P and a diagonal matrix D. Although the estimating function is similar to that of the quasi maximum likelihood, FastICA algorithm is based on the Newton's algorithm, and therefore, it has globally more stable dynamics than the natural gradient learning.

A.2 The Double Blind Algorithm by Hyvärinen and Hurri (2004)

Hyvärinen and Hurri (2004) proposed an algorithm for separating sources under the double blind situation. The estimator is obtained by maximizing

$$J(W) = \sum_{i,j}\left[\widehat{\mathrm{cov}}\{y_i^2(\cdot), y_j^2(\cdot-\Delta t)\}\right]^2,$$

under the orthogonality condition $WW^\top = I$, where

$$\widehat{\mathrm{cov}}\{y_i^2(\cdot), y_j^2(\cdot-\Delta t)\} = \frac{1}{T-\Delta t}\sum_{t=\Delta t+1}^{T} y_i^2(t)y_j^2(t-\Delta t) - 1.$$

Let us assume that

$$\begin{aligned}
&\widehat{\mathrm{cum}}\{s_i(\cdot), s_j(\cdot), s_k(\cdot-\Delta t), s_l(\cdot-\Delta t)\} \\
&:= \frac{1}{T-\Delta t}\sum_{t=\Delta t+1}^{T} s_i(t)s_j(t)s_k(t-\Delta t)s_l(t-\Delta t) - \frac{1}{T^2}\sum_{t=1}^{T} s_i(t)s_j(t)\sum_{t=1}^{T} s_k(t)s_l(t) \\
&\quad -\frac{1}{(T-\Delta t)^2}\sum_{t=\Delta t+1}^{T} s_i(t)s_k(t-\Delta t)\sum_{t=\Delta t+1}^{T} s_j(t)s_l(t-\Delta t) \\
&\quad -\frac{1}{(T-\Delta t)^2}\sum_{t=\Delta t+1}^{T} s_i(t)s_l(t-\Delta t)\sum_{t=\Delta t+1}^{T} s_j(t)s_k(t-\Delta t) \\
&= \begin{cases} K_{ik}+o_p(1), & i=j,\ k=l \\ o_p(1), & \text{otherwise} \end{cases}
\end{aligned}$$

that is, the empirical cumulants of the source signal $s(t) = (A^*)^{-1}x(t)$ converge to their expectation, where

$$K_{ij} = \frac{1}{T-\Delta t}\sum_{t=\Delta t+1}^{T}\mathrm{E}\left[s_i^2(t)s_j^2(t-\Delta t)\right] - \frac{1}{T^2}\sum_{t=1}^{T}\mathrm{E}\left[s_i^2(t)\right]\sum_{t=1}^{T}\mathrm{E}\left[s_j^2(t)\right].$$

By ignoring higher-order terms, we get

$$J = \sum_{i,j,k,l} (q_{ik}^2 K_{kl} q_{jl}^2)^2$$

where $Q = (q_{ij}) = BA^*$ and $B = WC^{-1/2}$ indicates the demixing matrix without whitening. Provided that the matrix $K = (K_{ij})$ is non-singular, the quantity J is maximized when Q is a signed permutation matrix, that is, by maximizing the criterion J we can estimate the true demixing matrix $B^* = (A^*)^{-1}$ up to signed permutation matrices. This also means that the algorithm does not work if there is no temporal covariance dependencies (for example, the data sets 'exp_supG' and 'uni_subG' in our experiment), or all sources have exactly same temporal covariance dependencies (for example, the data sets 'com_supG' and 'com_subG' in our experiment).

Although the authors have already given its validity as mentioned above, we will check its estimating equation. By solving the constrained optimization problem, we see that the estimator is obtained from the estimating equation

$$\widehat{F}(X, \widehat{B}) = 0, \tag{41}$$

where

$$\begin{aligned} \widehat{F}_{ij}(X,B) &= \sum_{t=1}^{T} \{y_i(t) y_j(t) - \delta_{ij}\} + \sum_{t=\Delta t+1}^{T} \left[\sum_l (\widehat{K}_{il} - \widehat{K}_{jl}) y_l^2(t-\Delta t) y_i(t) y_j(t) \right. \\ &\quad \left. + \sum_l (\widehat{K}_{li} - \widehat{K}_{lj}) y_l^2(t) y_i(t-\Delta t) y_j(t-\Delta t) \right]. \end{aligned} \tag{42}$$

and $\widehat{K}_{ij} = \widehat{\mathrm{cov}}\{y_i^2(\cdot), y_j(\cdot - \Delta t)\}$. By replacing $\widehat{K}_{ij}$ with K_{ij}, let us define the function

$$\begin{aligned} F_{ij}(X,B) &= \sum_{t=1}^{T} \{y_i(t) y_j(t) - \delta_{ij}\} + \sum_{t=\Delta t+1}^{T} \left[\sum_l (K_{il} - K_{jl}) y_l^2(t-\Delta t) y_i(t) y_j(t) \right. \\ &\quad \left. + \sum_l (K_{li} - K_{lj}) y_l^2(t) y_i(t-\Delta t) y_j(t-\Delta t) \right]. \end{aligned} \tag{43}$$

Suppose that $F(X,B)$ is an estimating function which fulfills $F(X,B) = O_p(T^{1/2})$, when B is the true parameter. If the function $\widehat{F}(X,B)$ satisfies

$$\widehat{F}(X, \widetilde{B}) = F(X, \widetilde{B}) + o_p(T^{1/2}) \tag{44}$$

for any $\widetilde{B}$ such that $\|\widetilde{B} - B\| = O(T^{-1/2})$, it can be shown that the residual does not matter to the asymptotic property of the estimator and the solution $\widehat{B}$ of (41) is asymptotically equivalent to that of the equation $F(X,B) = 0$ (see Cardoso, 1997). In fact, we can prove (44) under mild conditions, that is, the difference between the functions (42) and (43) can be neglected. Therefore, we will check whether $F(X,B)$ actually satisfies the conditions of estimating functions. If we take the constraints (40) to determine the scales of the sources, the unbiasedness condition (14) follows from uncorrelatedness of the sources and

$$\begin{aligned} &\sum_{t=\Delta t+1}^{T} \sum_l (K_{il} - K_{jl}) \, \mathrm{E}\left[y_l^2(t-\Delta t) y_i(t) y_j(t) \right] \\ &= \sum_{t=\Delta t+1}^{T} \sum_l (K_{il} - K_{jl}) \, \mathrm{E}\left[v_l^2(t-\Delta t) v_i(t) v_j(t) \right] \mathrm{E}\left[z_l^2(t-\Delta t) \right] \mathrm{E}\left[z_i(t) z_j(t) \right] = 0, \end{aligned}$$

$$\begin{aligned}
&\sum_{t=\Delta t+1}^{T} \sum_{l} (K_{li} - K_{lj}) \, \mathrm{E}\left[y_l^2(t) y_i(t-\Delta t) y_j(t-\Delta t) \right] \\
&= \sum_{t=\Delta t+1}^{T} \sum_{l} (K_{li} - K_{lj}) \, \mathrm{E}\left[v_l^2(t) v_i(t-\Delta t) v_j(t-\Delta t) \right] \, \mathrm{E}\left[z_l^2(t) \right] \, \mathrm{E}\left[z_i(t-\Delta t) z_j(t-\Delta t) \right] \\
&= 0.
\end{aligned}$$

We remark that the expectations are taken with respect to $p(X|B,\rho_z,\rho_V)$, and therefore $y(t) = Bx(t) = s(t)$ holds. If the other regularity condition holds, $\widehat{F}(X,B)$ turns out to be an asymptotic estimating function which is asymptotically equivalent to an estimating function and the estimator $\widehat{B}$ converges to the correct demixing matrix $B^* = (A^*)^{-1}$.

A.3 JADE

Although in a rigorous sense, the asymptotic properties of JADE should be analyzed as in the previous section (see also Cardoso, 1997), its consistency can be shown more easily (as suggested by one of the anonymous reviewers). Suppose that the contrast function of JADE

$$J_{\mathrm{JADE}}(W) = \sum_{ijkl \neq iikl} |\widehat{\mathrm{cum}}(y_i, y_j, y_k, y_l)|^2$$

uniformly converges to the ideal contrast function

$$J^*_{\mathrm{JADE}}(W) = \sum_{ijkl \neq iikl} |\mathrm{cum}(y_i, y_j, y_k, y_l)|^2$$

on the set of orthogonal matrices W such that $WW^\top = I$, where $\widehat{\mathrm{cum}}$ and cum denote the empirical and the expected cumulant tensor, respectively. Then, the minimum of the $J_{\mathrm{JADE}}(W)$ converges to that of $J^*_{\mathrm{JADE}}(W)$. If W is the true demixing matrix and y_i's are extracted signals with W, the components $K_{ijkl} := \mathrm{cum}(y_i, y_j, y_k, y_l)$ of the expected cumulant are zero except for $i = j = k = l$ or $i = j \neq k = l$ or $i = l \neq j = k$. Thus, one needs only to show that the estimating equation is associated to the minimization of $2\sum_{i \neq j} |\mathrm{cum}(y_i, y_j, y_i, y_j)|^2$ under the orthogonality constraints which is satisfied when y_i equals the true sources (up to a scaling and a permutation). The estimating equation is

$$\begin{aligned}
\sum_{t=1}^{T} \Bigg\{ & \mathrm{E}[y_i(t) y_j(t)] - \delta_{ij} \\
& + \sum_{k \neq i} K_{ikik} \mathrm{E}[y_k^2(t) y_i(t) y_j(t)] - \sum_{k \neq j} K_{jkjk} \mathrm{E}[y_k^2(t) y_i(t) y_j(t)] \Bigg\} = 0,
\end{aligned} \tag{45}$$

which can be seen to be satisfied, when y_i's equal to the true sources. We remark that the same formula as (45) can be obtained after the rigorous analysis. The function which is associated with the asymptotic estimating function (see (43)) becomes

$$F_{ij}(X,W) = \sum_{t=1}^{T} \left\{ y_i(t) y_j(t) - \delta_{ij} + \sum_{k \neq i} K_{ikik} y_k^2(t) y_i(t) y_j(t) - \sum_{k \neq j} K_{jkjk} y_k^2(t) y_i(t) y_j(t) \right\}.$$

A.4 TDSEP/SOBI

Let us define lagged covariance matrices of $x(t)$

$$\begin{aligned} R(\Delta t) &= \frac{1}{T}\sum_{t=\Delta t+1}^{T} \mathrm{E}\left[x(t)x^{\top}(t-\Delta t)\right] \\ &= A^{*}\left\{\frac{1}{T}\sum_{t=\Delta t+1}^{T} \mathrm{E}\left[s(t)s^{\top}(t-\Delta t)\right]\right\}(A^{*})^{\top}. \end{aligned}$$

When the sources s_i's are mutually independent and have temporal covariance structure, the demixing matrix $PD(A^{*})^{-1}$ can diagonalize all lagged covariance matrices $R(\Delta t)$, where P is a permutation matrix and D is a diagonal matrix. This property has been used in blind separation methods with second order statistics (Tong et al., 1991; Belouchrani et al., 1997; Ziehe and Müller, 1998).

In the variance-dependent BSS model, for any i, j, t and $\Delta t \geq 1$,

$$\mathrm{E}\left[s_i(t)s_j(t-\Delta t)\right] = \mathrm{E}\left[v_i(t)v_j(t-\Delta t)\right]\mathrm{E}\left[z_i(t)z_j(t-\Delta t)\right] \;=\; 0.$$

Therefore, $R(\Delta t) = 0$ for $\Delta t \geq 1$, that is, we cannot get any information about the mixing matrix A from lagged covariance matrices $R(\Delta t)$. This is why TDSEP does not work for this model.

References

S. Amari. *Differential Geometrical Methods in Statistics*. Springer Verlag, Berlin, 1985.

S. Amari. Natural gradient works efficiently in learning. *Neural Computation*, 10(2):251–276, 1998.

S. Amari and J.-F. Cardoso. Blind source separation—semiparametric statistical approach. *IEEE Trans. on Signal Processing*, 45(11):2692–2700, 1997.

S. Amari, T.-P. Chen, and A. Cichocki. Stability analysis of adaptive blind source separation. *Neural Networks*, 10(8):1345–1351, 1997.

S. Amari, T.-P. Chen, and A. Cichocki. Nonholonomic orthogonal learning algorithms for blind source separation. *Neural Computation*, 12:1463–1484, 2000.

S. Amari, A. Cichocki, and H. H. Yang. A new learning algorithm for blind source separation. In *Advances in Neural Information Processing Systems 8*, pages 757–763. MIT Press, 1996.

S. Amari and M. Kawanabe. Estimating functions in semiparametric statistical models. In I. V. Basawa et al., editor, *Selected Proceedings of the Symposium on Estimating Functions*, volume 32 of *IMS Lecture Notes–Monograph Series*, pages 65–81, 1997a.

S. Amari and M. Kawanabe. Information geometry of estimating functions in semiparametric statistical models. *Bernoulli*, 3:29–54, 1997b.

F. R. Bach and M. I. Jordan. Tree-dependent component analysis. In *Uncertainty in Artificial Intelligence: Proceedings of the Eighteenth Conference (UAI-2002)*, 2002.

A. J. Bell and T. J. Sejnowski. An information-maximization approach to blind separation and blind deconvolution. *Neural Computation*, 7:1129–1159, 1995.

A. Belouchrani, K. Abed Meraim, J.-F. Cardoso, and E. Moulines. A blind source separation technique based on second order statistics. *IEEE Trans. on Signal Processing*, 45(2):434–444, 1997.

P. J. Bickel, C. A. J. Klaassen, Y. Ritov, and J. A. Wellner. *Efficient and Adaptive Estimation for Semiparamtric Models*. John Hopkins Univ. Press, Baltimore, MD, 1993.

J.-F. Cardoso. Estimating equations for source separation. In *Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP'97)*, volume 5, pages 3449–3452, Munich, Germany, 1997.

J.-F. Cardoso. Blind signal separation: statistical principles. *Proc. of the IEEE*, 86(10):2009–2025, 1998a.

J.-F. Cardoso. Multidimensional independent component analysis. In *Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP'98)*, Seattle, WA, 1998b.

J.-F. Cardoso. Independent component analysis of the cosmic microwave background. In *Proc.Int. Symposium on Independent Component Analysis and Blind Signal Separation (ICA2003)*, pages 1111–1116, Nara, Japan, 2003.

J.-F. Cardoso and A. Souloumiac. Blind beamforming for non Gaussian signals. *IEE Proceedings-F*, 140:362 – 370, 1993.

P. Comon. Independent component analysis—a new concept? *Signal Processing*, 36:287–314, 1994.

V. P. Godambe. Conditional likelihood and unconditional optimum estimating equations. *Biometrika*, 63:277–284, 1976.

V. P. Godambe, editor. *Estimating Functions*. Oxford Univ. Press, New York, 1991.

S. Harmeling, F. Meinecke, and K.-R. Müller. Injecting noise for analysing the stability of ica components. *Signal Processing*, 84:255–266, 2004.

P. J. Huber. *Robust Statistics*. Wiley, New York, 1981.

A. Hyvärinen. Fast and robust fixed-point algorithms for independent component analysis. *IEEE Trans. on Neural Networks*, 10(3):626–634, 1999.

A. Hyvärinen, P. O. Hoyer, and M. Inki. Topographic independent component analysis. *Neural Computation*, 13(7):1527–1558, 2001a.

A. Hyvärinen and J. Hurri. Blind separation of sources that have spatiotemporal variance dependencies. *Signal Processing*, 84, 2004. 247–254.

A. Hyvärinen, J. Karhunen, and E. Oja. *Independent Component Analysis*. Wiley, 2001b.

A. Hyvärinen and E. Oja. A fast fixed-point algorithm for independent component analysis. *Neural Computation*, 9(7):1483–1492, 1997.

Ch. Jutten and J. Herault. Blind separation of sources, part I: An adaptive algorithm based on neuromimetic architecture. *Signal Processing*, 24:1–10, 1991.

M. Kawanabe and K.-R. Müller. Estimating functions for blind separation when source have variance-dependencies. In C. G. Puntonet and A. Prieto, editors, *Proc. Int. Workshop on Independent Component Analysis and Blind Signal Separation (ICA2004)*, pages 136 – 143, Granada, Spain, 2004.

M. Kawanabe and N. Murata. Independent component analysis in the presence of Gaussian noise based on estimating functions. In *Proc. Int. Workshop on Independent Component Analysis and Blind Signal Separation (ICA2000)*, pages 279–284, Helsinki, Finland, 2000.

S. Makeig, T-P. Jung, D. Ghahremani, A. J. Bell, and T. J. Sejnowski. Blind separation of event-related brain responses into independent components. *Proc. Natl. Acad. Sci. USA*, 94:10979–10984, 1997.

F. Meinecke, S. Harmeling, and K.-R. Müller. Robust ICA for super-Gaussian sources. In C. G. Puntonet and A. Prieto, editors, *Proc. Int. Workshop on Independent Component Analysis and Blind Signal Separation (ICA2004)*, pages 217 – 224, Granada, Spain, 2004.

F. Meinecke, A. Ziehe, M. Kawanabe, and K.-R. Müller. A resampling approach to estimate the stability of one- or multidimensional independent components. *IEEE Transactions on Biomedical Engineering*, 49(12):1514–1525, 2002.

F. Meinecke, A. Ziehe, J. Kurths, and K.-R. Müller. Measuring phase synchronization of superimposed signals. *Physical Review Letters*, 2005.

K.-R. Müller, N. Murata, A. Ziehe, and S.-I. Amari. *On-line learning in Switching and Drifting environments with application to blind source separation*, pages 93–110. On-line learning in neural networks. Cambridge University Press, 1998.

K.-R. Müller, R. Vigário, F. Meinecke, and A. Ziehe. Blind source separation techniques for decomposing evoked brain signals. *International Journal of Bifurcation and Chaos*, 14(2):773–791, 2004.

N. Murata, M. Kawanabe, A. Ziehe, K.-R. Müller, and S.-I. Amari. On-line learning in changing environments with applications in supervised and unsupervised learning. *Neural Networks*, 15 (4-6):743–760, 2002.

B. A. Olshausen and D. J. Field. Emergence of simple-cell receptive field by learning a sparse code for natural images. *Nature*, 381:607–609, 1996.

L. Parra and C. Spence. Convolutive blind source separation of non-stationary sources. *IEEE Trans. on Speech and Audio Processing*, 8:320–327, 2000.

D.-T. Pham. Blind separation of instantaneous mixture sources via an independent component analysis. *IEEE Trans. on Signal Processing*, 44(11):2768–2779, 1996.

D.-T. Pham and J.-F. Cardoso. Blind separation of instantaneous mixtures of non-stationary sources. In *Proc. Int. Workshop on Independent Component Analysis and Blind Signal Separation (ICA2000)*, pages 187–193, Helsinki, Finland, 2000.

D.-T. Pham and P. Garrat. Blind separation of mixture of independent sources through a quasi-maximum likelihood approach. *IEEE Trans. on Signal Processing*, 45(7):1712–1725, 1997.

D.-T. Pham, P. Garrat, and C. Jutten. Separation of a mixture of independent sources through a maximum likelihood approach. In *Proc. EUSIPCO*, pages 771–774, 1992.

M. Sørensen. On asymptotics of estimating functions. *Brazilian Journal of Probability and Statistics*, 13:111–136, 1999.

F. J. Theis. Uniqueness of complex and multidimensional independent component analysis. *Signal Processing*, 84(5):951–956, 2004.

H.-Lan Nguyen Thi and Ch. Jutten. Blind source separation for convolutive mixtures. *Signal Processing*, 45:209–229, 1995.

L. Tong, R. W. Liu, V. Soon, and Y. F. Huang. Indeterminacy and identifiability of blind identification. *IEEE Transactions on Circuits and Systems*, 38:499–509, 1991.

H. Valpola, M. Harva, and J. Karhunen. Hierachical models of variance sources. In *Proc.Int. Symposium on Independent Component Analysis and Blind Signal Separation (ICA2003)*, Nara, Japan, 2003.

R. N. Vigario. Extraction of ocular artefacts from EEG using independent component analysis. *Electroencephalography and clinical Neurophysiology*, 103:395–404, 1997.

H. H. Yang and S.-I. Amari. Adaptive on-line learning algorithms for blind separation: Maximum entropy and minimum mutual information. *Neural Computation*, 9(7):1457–1482, 1997.

A. Ziehe and K.-R. Müller. TDSEP – an efficient algorithm for blind separation using time structure. In L. Niklasson, M. Bodén, and T. Ziemke, editors, *Proceedings of the 8th International Conference on Artificial Neural Networks (ICANN'98)*, pages 675 – 680, 1998.

A. Ziehe, K.-R. Müller, G. Nolte, B.-M. Mackert, and G. Curio. Artifact reduction in magnetoneurography based on time-delayed second order correlations. *IEEE Transactions on biomedical Engineering*, 47:75–87, 2000.