LETTER Communicated by Aapo Hyvarinen

A New Concept for Separability Problems in Blind Source Separation

Fabian J. Theis
fabian@theis.name
Institute of Biophysics, University of Regensburg, 93040 Regensburg, Germany

The goal of blind source separation (BSS) lies in recovering the original independent sources of a mixed random vector without knowing the mixing structure. A key ingredient for performing BSS successfully is to know the indeterminacies of the problem—that is, to know how the separating model relates to the original mixing model (separability). For linear BSS, Comon (1994) showed using the Darmois-Skitovitch theorem that the linear mixing matrix can be found except for permutation and scaling. In this work, a much simpler, direct proof for linear separability is given. The idea is based on the fact that a random vector is independent if and only if the Hessian of its logarithmic density (resp. characteristic function) is diagonal everywhere. This property is then exploited to propose a new algorithm for performing BSS. Furthermore, first ideas of how to generalize separability results based on Hessian diagonalization to more complicated nonlinear models are studied in the setting of postnonlinear BSS.

1 Introduction

In independent component analysis (ICA), one tries to find statistically independent data within a given random vector. An application of ICA lies in blind source separation (BSS), where it is furthermore assumed that the given vector has been mixed using a fixed set of independent sources. The advantage of applying ICA algorithms to BSS problems in contrast to correlation-based algorithms is that ICA tries to make the output signals as independent as possible by also including higher-order statistics.

Since the introduction of independent component analysis by Hérault and Jutten (1986), various algorithms have been proposed to solve the BSS problem (Comon, 1994; Bell & Sejnowski, 1995; Hyvärinen & Oja, 1997; Theis, Jung, Puntonet, & Lang, 2002). Good textbook-level introductions to ICA are given in Hyvärinen, Karhunen, and Oja (2001) and Cichocki and Amari (2002).

Separability of linear BSS states that under weak conditions to the sources, the mixing matrix is determined uniquely by the mixtures except for permutation and scaling, as showed by Comon (1994) using the Darmois-Skitovitch theorem. We propose a direct proof based on the concept of

separated functions, that is, functions that can be split into a product of one-dimensional functions (see definition 1). If the function is positive, this is equivalent to the fact that its logarithm has a diagonal Hessian everywhere (see lemma 1 and theorem 1). A similar lemma has been shown by Lin (1998) for what he calls block diagonal Hessians. However, he omits discussion of the separatedness of densities with zeros, which plays a minor role for the separation algorithm he is interested in but is important for deriving separability. Using separatedness of the density, respectively, the characteristic function (Fourier transformation), of the random vector, we can then show separability directly (in two slightly different settings, for which we provide a common framework). Based on this result, we propose an algorithm for linear BSS by diagonalizing the logarithmic density of the Hessian. We recently found that this algorithm has already been proposed (Lin, 1998), but without considering the necessary assumptions for successful algorithm application. Here we give precise conditions for when to apply this algorithm (see theorem 3) and show that points satisfying these conditions can indeed be found if the sources contain at most one gaussian component (see lemma 5). Lin uses a discrete approximation of the derivative operator to approximate the Hessian. We suggest using kernel-based density estimation, which can be directly differentiated. A similar algorithm based on Hessian diagonalization has been proposed by Yeredor (2000) using the characteristic function of a random vector. However, the characteristic function is complex valued, and additional care has to be taken when applying a complex logarithm. Basically, this is well defined locally only at nonzeros. In algorithmic terms, the characteristic function can be easily approximated by samples (which is equivalent to our kernel-based density approximation using gaussians before Fourier transformation). Yeredor suggests joint diagonalization of the Hessian of the logarithmic characteristic function (which is problematic because of the nonuniqueness of the complex logarithm) evaluated at several points in order to avoid the locality of the algorithm. Instead of joint diagonalization, we use a combined energy function based on the previously defined separator, which also takes into account global information but does not have the drawback of being singular at zeros of the density, respectively, characteristic function. Thus, the algorithmic part of this article can be seen as a general framework for the algorithms proposed by Lin (1998) and Yeredor (2000).

Section 2 introduces separated functions, giving local characterizations of the densities of independent random vectors. Section 3 then introduces the linear BSS model and states the well-known separability result. After giving an easy and short proof in two dimensions with positive densities, we provide a characterization of gaussians in terms of a differential equation and provide the general proof. The BSS algorithm based on finding separated densities is proposed and studied in section 4. We finish with a generalization of the separability for the postnonlinear mixture case in section 5.

2 Separated and Linearly Separated Functions

Definition 1. *A function* $f : \mathbb{R}^n \to \mathbb{C}$ *is said to be* separated, *respectively,* linearly separated, *if there exist one-dimensional functions* $g_1, \ldots, g_n : \mathbb{R} \to \mathbb{C}$ *such that* $f(\mathbf{x}) = g_1(x_1) \cdots g_n(x_n)$ *respectively* $f(\mathbf{x}) = g_1(x_1) + \cdots + g_n(x_n)$ *for all* $\mathbf{x} \in \mathbb{R}^n$.

Note that the functions g_i are uniquely determined by f up to a scalar factor, respectively, an additive constant. If f is linearly separated, then $\exp f$ is separated. Obviously the density function of an independent random vector is separated. For brevity, we often use the tensor product and write $f \equiv g_1 \otimes \cdots \otimes g_n$ for separated f, where for any functions h, k defined on a set U, $h \equiv k$ if $h(\mathbf{x}) = k(\mathbf{x})$ for all $\mathbf{x} \in U$.

Separatedness can also be defined on any open parallelepiped $(a_1, b_1) \times \cdots \times (a_n, b_n) \subset \mathbb{R}^n$ in the obvious way. We say that f is locally separated at $\mathbf{x} \in \mathbb{R}^n$ if there exists an open parallelepiped U such that $\mathbf{x} \in U$ and $f|U$ is separated. If f is separated, then f is obviously everywhere locally separated. The converse, however, does not necessarily hold, as shown in Figure 1.

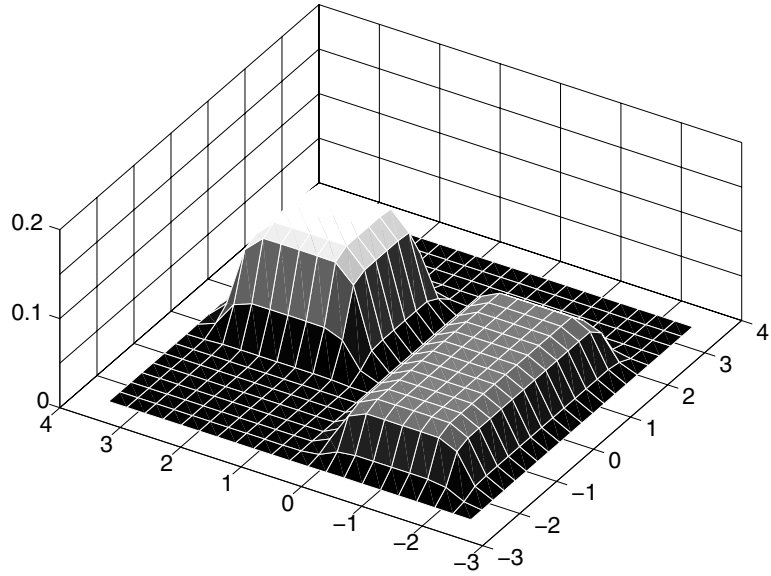

Figure 1: Density of a random vector $\mathbf{S}$ with a locally but not globally separated density. Here, $p_{\mathbf{S}} := c\chi_{[-2,2]\times[-2,0]\cup[0,2]\times[1,3]}$ where χ_U denotes the function that is 1 on U and 0 everywhere else. Obviously, $p_{\mathbf{S}}$ is not separated globally, but is separated if restricted to squares of length < 1. Plotted is a smoothed version of $p_{\mathbf{S}}$.

The function f is said to be positive if f is real and $f(\mathbf{x}) > 0$ for all $\mathbf{x} \in \mathbb{R}^n$, and nonnegative if f is real and $f(\mathbf{x}) \geq 0$ for all $\mathbf{x} \in \mathbb{R}^n$. A positive function f is separated if and only if $\ln f$ is linearly separated.

Let $\mathcal{C}^m(U, V)$ be the ring of all m-times continuously differentiable functions from $U \subset \mathbb{R}^n$ to $V \subset \mathbb{C}$, U open. For a $\mathcal{C}^m$-function f, we write $\partial_{i_1} \cdots \partial_{i_m} f := \partial^m f / \partial x_{i_1} \cdots \partial x_{i_m}$ for the m-fold partial derivatives. If $f \in \mathcal{C}^2(\mathbb{R}^n, \mathbb{C})$, denote with the symmetric $(n \times n)$-matrix $\mathbf{H}_f(\mathbf{x}) := \left(\partial_i \partial_j f(\mathbf{x})\right)_{i,j=1}^n$ the Hessian of f at $\mathbf{x} \in \mathbb{R}^n$.

Linearly separated functions can be classified using their Hessian (if it exists):

Lemma 1. *A function $f \in \mathcal{C}^2(\mathbb{R}^n, \mathbb{C})$ is linearly separated if and only if $\mathbf{H}_f(\mathbf{x})$ is diagonal for all $\mathbf{x} \in \mathbb{R}^n$.*

A similar lemma for block diagonal Hessians has been shown by Lin (1998).

Proof. If f is linearly separated, its Hessian is obviously diagonal everywhere by definition.

Assume the converse. We prove that f is separated by induction over the dimension n. For $n = 1$, the claim is trivial. Now assume that we have shown the lemma for $n - 1$. By induction assumption, $f(x_1, \ldots, x_{n-1}, 0)$ is linearly separated, so

$$f(x_1, \ldots, x_{n-1}, 0) = g_1(x_1) + \cdots + g_{n-1}(x_{n-1})$$

for all $x_i \in \mathbb{R}$ and some functions g_i on $\mathbb{R}$. Note that $g_i \in \mathcal{C}^2(\mathbb{R}, \mathbb{C})$.

Define a function $h : \mathbb{R} \to \mathbb{C}$ by $h(y) := \partial_n f(x_1, \ldots, x_{n-1}, y)$, $y \in \mathbb{R}$, for fixed $x_1, \ldots, x_{n-1} \in \mathbb{R}$. Note that h is independent of the choice of the x_i, because $\partial_n \partial_i f \equiv \partial_i \partial_n f$ is zero everywhere, so $x_i \mapsto \partial_n f(x_1, \ldots, x_{n-1}, y)$ is constant for fixed $x_j, y \in \mathbb{R}$, $j \neq i$. By definition, $h \in \mathcal{C}^1(\mathbb{R}, \mathbb{C})$, so h is integrable on compact intervals. Define $k : \mathbb{R} \to \mathbb{C}$ by $k(y) := \int_0^y h$. Then

$$f(x_1, \ldots, x_n) = g_1(x_1) + \cdots + g_{n-1}(x_{n-1}) + k(x_n) + c,$$

where $c \in \mathbb{C}$ is a constant, because both functions have the same derivative and $\mathbb{R}^n$ is connected. If we set $g_n := k + c$, the claim follows.

This lemma also holds for functions defined on any open parallelepiped $(a_1, b_1) \times \cdots \times (a_n, b_n) \subset \mathbb{R}^n$. Hence, an arbitrary real-valued $\mathcal{C}^2$-function f is locally separated at $\mathbf{x}$ with $f(\mathbf{x}) \neq 0$ if and only if the Hessian of $\ln |f|$ is locally diagonal.

For a positive function f, the Hessian of its logarithm is diagonal everywhere if it is separated, and it is easy to see that for positive f, the converse

also holds globally (see theorem 1(ii)). In this case, we have for $i \neq j$,

$$0 \equiv \partial_i \partial_j \ln f \equiv \frac{f \partial_i \partial_j f - (\partial_i f)(\partial_j f)}{f^2},$$

so f is separated if and only if

$$f \partial_i \partial_j f \equiv (\partial_i f)(\partial_j f)$$

for $i \neq j$ or even $i < j$. This motivates the following definition:

Definition 2. *For $i \neq j$, the operator*

$$R_{ij} : \mathcal{C}^2(\mathbb{R}^n, \mathbb{C}) \to \mathcal{C}^0(\mathbb{R}^n, \mathbb{C})$$
$$f \mapsto R_{ij}[f] := f \partial_i \partial_j f - (\partial_i f)(\partial_j f)$$

*is called the ij-*separator.

Theorem 1. *Let $f \in \mathcal{C}^2(\mathbb{R}^n, \mathbb{C})$.*

i. If f is separated, then $R_{ij}[f] \equiv 0$ for $i \neq j$ or, equivalently,

$$f \partial_i \partial_j f \equiv (\partial_i f)(\partial_j f) \tag{2.1}$$

holds for $i \neq j$.

ii. If f is positive and $R_{ij}[f] \equiv 0$ holds for all $i \neq j$, then f is separated.

If f is assumed to be only nonnegative, then f is locally separated but not necessarily globally separated (if the support of f has more than one component). See Figure 1 for an example of a nonseparated density with $R_{12}[f] \equiv 0$.

Proof of Theorem 1.i. If f is separated, then $f(\mathbf{x}) = g_1(x_1) \cdots g_n(x_n)$ or short $f \equiv g_1 \otimes \cdots \otimes g_n$, so

$$\partial_i f \equiv g_1 \otimes \cdots \otimes g_{i-1} \otimes g_i' \otimes g_{i+1} \otimes \cdots \otimes g_n$$

and

$$\partial_i \partial_j f \equiv g_1 \otimes \cdots \otimes g_{i-1} \otimes g_i' \otimes g_{i+1} \otimes \cdots \otimes g_{j-1} \otimes g_j' \otimes g_{j+1} \otimes \cdots \otimes g_n$$

for $i < j$. Hence equation 2.1 holds.

ii. Now assume the converse and let f be positive. Then according to the remarks after lemma 1, $\mathbf{H}_{\ln f}(\mathbf{x})$ is everywhere diagonal, so lemma 1 shows that $\ln f$ is linearly separated; hence, f is separated.

Some trivial properties of the separator R_{ij} are listed in the next lemma:

Lemma 2. *Let $f, g \in \mathcal{C}^2(\mathbb{R}^n, \mathbb{C})$, $i \neq j$ and $\alpha \in \mathbb{C}$. Then*

$$R_{ij}[\alpha f] = \alpha^2 R_{ij}[f]$$

and

$$R_{ij}[f+g] = R_{ij}[f] + R_{ij}[g] + f\partial_i\partial_j g + g\partial_i\partial_j f - (\partial_i f)(\partial_j g) - (\partial_i g)(\partial_j f).$$

3 Separability of Linear BSS

Consider the noiseless linear instantaneous BSS model with as many sources as sensors:

$$\mathbf{X} = \mathbf{AS}, \tag{3.1}$$

with an independent n-dimensional random vector $\mathbf{S}$ and $\mathbf{A} \in \mathrm{Gl}(n)$. Here, $\mathrm{Gl}(n)$ denotes the general linear group of $\mathbb{R}^n$, that is, the group of all invertible $(n \times n)$-matrices.

The task of linear BSS is to find $\mathbf{A}$ and $\mathbf{S}$ given only $\mathbf{X}$. An obvious indeterminacy of this problem is that $\mathbf{A}$ can be found only up to scaling and permutation because for scaling $\mathbf{L}$ and permutation matrix $\mathbf{P}$,

$$\mathbf{X} = \mathbf{ALPP}^{-1}\mathbf{L}^{-1}\mathbf{S},$$

and $\mathbf{P}^{-1}\mathbf{L}^{-1}\mathbf{S}$ is also independent. Here, an invertible matrix $\mathbf{L} \in \mathrm{Gl}(n)$ is said to be a scaling matrix if it is diagonal. We say two matrices $\mathbf{B}, \mathbf{C}$ are equivalent, $\mathbf{B} \sim \mathbf{C}$, if $\mathbf{C}$ can be written as $\mathbf{C} = \mathbf{BPL}$ with a scaling matrix $\mathbf{L} \in \mathrm{Gl}(n)$ and an invertible matrix with unit vectors in each row (permutation matrix) $\mathbf{P} \in \mathrm{Gl}(n)$. Note that $\mathbf{PL} = \mathbf{L}'\mathbf{P}$ for some scaling matrix $\mathbf{L}' \in \mathrm{Gl}(n)$, so the order of the permutation and the scaling matrix does not play a role for equivalence. Furthermore, if $\mathbf{B} \in \mathrm{Gl}(n)$ with $\mathbf{B} \sim \mathbf{I}$, then also $\mathbf{B}^{-1} \sim \mathbf{I}$, and, more generally if $\mathbf{BC} \sim \mathbf{A}$, then $\mathbf{C} \sim \mathbf{B}^{-1}\mathbf{A}$. According to the above, solutions of linear BSS are equivalent. We will show that under mild assumptions to $\mathbf{S}$, there are no further indeterminacies of linear BSS.

$\mathbf{S}$ is said to have a gaussian component if one of the S_i is a one-dimensional gaussian, that is, $p_{S_i}(x) = d\exp(-ax^2 + bx + c)$ with $a, b, c, d \in \mathbb{R}$, $a > 0$, and $\mathbf{S}$ has a deterministic component if one S_i is deterministic, that is, constant.

Theorem 2 (Separability of linear BSS). *Let $\mathbf{A} \in \mathrm{Gl}(n)$ and $\mathbf{S}$ be an independent random vector. Assume one of the following:*

i. $\mathbf{S}$ has at most one gaussian or deterministic component, and the covariance of $\mathbf{S}$ exists.

ii. **S** *has no gaussian component, and its density* $p_{\mathbf{S}}$ *exists and is twice continuously differentiable.*

Then if **AS** *is again independent,* **A** *is equivalent to the identity.*

So **A** is the product of a scaling and a permutation matrix. The important part of this theorem is assumption i, which has been used to show separability by Comon (1994) and extended by Eriksson and Koivunen (2003) based on the Darmois-Skitovitch theorem (Darmois, 1953; Skitovitch, 1953). Using this theorem, the second part can be easily shown without $\mathcal{C}^2$-densities.

Theorem 2 indeed proves separability of the linear BSS model, because if $\mathbf{X} = \mathbf{AS}$ and **W** is a demixing matrix such that **WX** is independent, then $\mathbf{WA} \sim \mathbf{I}$, so $\mathbf{W}^{-1} \sim \mathbf{A}$ as desired.

We will give a much easier proof without having to use the Darmois-Skitovitch theorem in the following sections.

3.1 Two-Dimensional Positive Density Case. For illustrative purposes we will first prove separability for a two-dimensional random vector **S** with positive density $p_{\mathbf{S}} \in \mathcal{C}^2(\mathbb{R}^2, \mathbb{R})$. Let $\mathbf{A} \in \mathrm{Gl}(2)$. It is enough to show that if **S** and **AS** are independent, then either $\mathbf{A} \sim \mathbf{I}$ or **S** is gaussian.

S is assumed to be independent, so its density factorizes:

$$p_{\mathbf{S}}(\mathbf{s}) = g_1(s_1)g_2(s_2),$$

for $\mathbf{s} \in \mathbb{R}^2$. First, note that the density of **AS** is given by

$$p_{\mathbf{AS}}(\mathbf{x}) = |\det \mathbf{A}|^{-1} p_{\mathbf{S}}(\mathbf{A}^{-1}\mathbf{x}) = c g_1(b_{11}x_1 + b_{12}x_2) g_2(b_{21}x_1 + b_{22}x_2)$$

for $\mathbf{x} \in \mathbb{R}^2$, $c \neq 0$ fixed. Here, $\mathbf{B} = (b_{ij}) = \mathbf{A}^{-1}$. **AS** is also assumed to be independent, so $p_{\mathbf{AS}}(\mathbf{x})$ is separated.

$p_{\mathbf{S}}$ was assumed to be positive; then so is $p_{\mathbf{AS}}$. Hence, $\ln p_{\mathbf{AS}}(\mathbf{x})$ is linearly separated, so

$$\partial_1 \partial_2 \ln p_{\mathbf{AS}}(\mathbf{x}) = b_{11}b_{12}h_1''(b_{11}x_1 + b_{12}x_2) + b_{21}b_{22}h_2''(b_{21}x_1 + b_{22}x_2) = 0$$

for all $\mathbf{x} \in \mathbb{R}^2$, where $h_i := \ln g_i \in \mathcal{C}^2(\mathbb{R}^2, \mathbb{R})$. By setting $\mathbf{y} := \mathbf{Bx}$, we therefore have

$$b_{11}b_{12}h_1''(y_1) + b_{21}b_{22}h_2''(y_2) = 0 \tag{3.2}$$

for all $\mathbf{y} \in \mathbb{R}^2$, because **B** is invertible.

Now, if **A** (and therefore also **B**) is equivalent to the identity, then equation 3.2 holds. If not, then **A**, and hence also **B**, have at least three nonzero entries. By equation 3.2 the fourth entry has to be nonzero, because the

h_i'' are not zero (otherwise $g_i(y_i) = \exp(ay_i + b)$, which is not integrable). Furthermore,

$$b_{11}b_{12}h_1''(y_1) = -b_{21}b_{22}h_2''(y_2)$$

for all $\mathbf{y} \in \mathbb{R}^2$, so the h_i'' are constant, say, $h_i'' \equiv c_i$, and $c_i \neq 0$, as noted above. Therefore, the h_i are polynomials of degree 2, and the $g_i = \exp h_i$ are gaussians ($c_i < 0$ because of the integrability of the g_i).

3.2 Characterization of Gaussians. In this section, we show that among all densities, respectively, characteristic functions, the gaussians satisfy a special differential equation.

Lemma 3. *Let $f \in \mathcal{C}^2(\mathbb{R}, \mathbb{C})$ and $a \in \mathbb{C}$ with*

$$af^2 - ff'' + f'^2 \equiv 0. \tag{3.3}$$

Then either $f \equiv 0$ or $f(x) = \exp\left(\frac{a}{2}x^2 + bx + c\right)$, $x \in \mathbb{R}$, with constants $b, c \in \mathbb{C}$.

Proof. Assume $f \not\equiv 0$. Let $x_0 \in \mathbb{R}$ with $f(x_0) \neq 0$. Then there exists a nonempty interval $U := (r, s)$ containing x_0 such that a complex logarithm log is defined on $f(U)$. Set $g := \log f|U$. Substituting $\exp g$ for f in equation 3.3 yields

$$a\exp(2g) - \exp(g)(g'' + g'^2)\exp(g) + g'^2\exp(2g) \equiv 0,$$

and therefore $g'' \equiv a$. Hence, g is a polynomial of degree ≤ 2 with leading coefficient $\frac{a}{2}$.

Furthermore,

$$\lim_{x \to r+} f(x) \neq 0$$

$$\lim_{x \to s-} f(x) \neq 0,$$

so f has no zeros at all because of continuity. The argument above with $U = \mathbb{R}$ shows the claim.

If, furthermore, f is real nonnegative and integrable with integral 1 (e.g., if f is the density of a random variable), then f has to be the exponential of a real-valued polynomial of degree precisely 2; otherwise, it would not be integrable. So we have the following corollary:

Corollary 1. *Let X be a random variable with twice continuously differentiable density p_X satisfying equation 3.3. Then X is gaussian.*

If we do not want to assume that the random variable has a density, we can use its characteristic function (Bauer, 1996) instead to show an equivalent result:

Corollary 2. *Let X be a random variable with twice continuously differentiable characteristic function $\widehat{X}(x) := E_X(\exp ixX)$ satisfying equation 3.3. Then X is gaussian or deterministic.*

Proof. Using $\widehat{X}(0) = 1$, lemma 3 shows that $\widehat{X}(x) = \exp(\frac{a}{2}x^2 + bx)$. Moreover, from $\widehat{X}(-x) = \overline{\widehat{X}(x)}$, we get $a \in \mathbb{R}$ and $b = ib'$ with real b'. And $|\widehat{X}| \leq 1$ shows that $a \leq 0$. So if $a = 0$, then X is deterministic (at b'), and if $a \neq 0$, then X has a gaussian distribution with mean b' and variance $-a^{-1}$.

3.3 Proof of Theorem 2. We will now prove linear separability; for this, we will use separatedness to show that some source components have to be gaussian (using the results from above) if the mixing matrix is not trivial. The main argument is given in the following lemma:

Lemma 4. *Let $g_i \in \mathcal{C}^2(\mathbb{R}, \mathbb{C})$ and $\mathbf{B} \in \mathrm{Gl}(n)$ such that $f(\mathbf{x}) := g_1 \otimes \cdots \otimes g_n(\mathbf{Bx})$ is separated. Then for all indices l and $i \neq j$ with $b_{li}b_{lj} \neq 0$, g_l satisfies the differential equation 3.3 with some constant a.*

Proof. f is separated, so by theorem 1i.

$$R_{ij}[f] \equiv f\partial_i\partial_j f - (\partial_i f)(\partial_j f) \equiv 0 \tag{3.4}$$

holds for $i < j$. The ingredients of this equation can be calculated for $i < j$ as follows:

$$\begin{aligned}
\partial_i f(\mathbf{x}) &= \sum_k b_{ki} g_1 \otimes \cdots \otimes g_k' \otimes \cdots \otimes g_n(\mathbf{Bx}) \\
(\partial_i f)(\partial_j f)(\mathbf{x}) &= \sum_{k,l} b_{ki}b_{lj}(g_1 \otimes \cdots \otimes g_k' \otimes \cdots \otimes g_n) \\
&\quad \times (g_1 \otimes \cdots \otimes g_l' \otimes \cdots \otimes g_n)(\mathbf{Bx}) \\
\partial_i\partial_j f(\mathbf{x}) &= \sum_k b_{ki}(b_{kj} g_1 \otimes \cdots \otimes g_k'' \otimes \cdots \otimes g_n \\
&\quad + \sum_{l \neq k} b_{lj} g_1 \otimes \cdots \otimes g_k' \otimes \cdots \otimes g_l' \otimes \cdots \otimes g_n)(\mathbf{Bx}).
\end{aligned}$$

Putting this in equation 3.4 yields

$$\begin{aligned}
0 &= (f\partial_i\partial_j f - (\partial_i f)(\partial_j f))(\mathbf{x}) \\
&= \sum_k b_{ki}b_{kj}((g_1 \otimes \cdots \otimes g_n)(g_1 \otimes \cdots \otimes g_k'' \otimes \cdots \otimes g_n) \\
&\quad - (g_1 \otimes \cdots \otimes g_k' \otimes \cdots \otimes g_n)^2)(\mathbf{Bx}) \\
&= \sum_k b_{ki}b_{kj} g_1^2 \otimes \cdots \otimes g_{k-1}^2 \otimes \left(g_k g_k'' - g_k'^2\right) \otimes g_{k+1}^2 \otimes \cdots \otimes g_n^2(\mathbf{Bx})
\end{aligned}$$

for $\mathbf{x} \in \mathbb{R}^n$. $\mathbf{B}$ is invertible, so the whole function is zero:

$$\sum_k b_{ki}b_{kj} g_1^2 \otimes \cdots \otimes g_{k-1}^2 \otimes \left(g_k g_k'' - g_k'^2\right) \otimes g_{k+1}^2 \otimes \cdots \otimes g_n^2 \equiv 0. \tag{3.5}$$

Choose $\mathbf{x} \in \mathbb{R}^n$ with $g_k(x_k) \neq 0$ for $k = 1, \ldots, n$. Evaluating equation 3.5 at $(x_1, \ldots, x_{l-1}, y, x_{l+1}, \ldots, x_n)$ for variable $y \in \mathbb{R}$ and dividing the resulting one-dimensional equation by the constant $g_1^2(x_1) \cdots g_{l-1}^2(x_{l-1}) g_{l+1}^2(x_{l+1}) \cdots g_n^2(x_n)$ shows

$$b_{li}b_{lj}\left(g_l g_l'' - g_l'^2\right)(y) = -\left(\sum_{k \neq l} b_{ki}b_{kj} \frac{g_k g_k'' - g_k'^2}{g_k^2}(x_k)\right) g_l^2(y) \tag{3.6}$$

for $y \in \mathbb{R}$. So for indices l and $i \neq j$ with $b_{li}b_{lj} \neq 0$, it follows from equation 3.6 that there exists $a \in \mathbb{C}$ such that g_k satisfies the differential equation $a g_l^2 - g_l g_l'' + g_l'^2 \equiv 0$, that is, equation 3.3.

Proof of Theorem 2. *i.* $\mathbf{S}$ is assumed to have at most one gaussian or deterministic component and existing covariance. Set $\mathbf{X} := \mathbf{AS}$.

We first show using whitening that $\mathbf{A}$ can be assumed to be orthogonal. For this, we can assume $\mathbf{S}$ and $\mathbf{X}$ to have no deterministic component at all (because arbitrary choice of the matrix coefficients of the deterministic components does not change the covariance). Hence, by assumption, $\mathrm{Cov}(\mathbf{X})$ is diagonal and positive definite, so let $\mathbf{D}_1$ be diagonal invertible with $\mathrm{Cov}(\mathbf{X}) = \mathbf{D}_1^2$. Similarly, let $\mathbf{D}_2$ be diagonal invertible with $\mathrm{Cov}(\mathbf{S}) = \mathbf{D}_2^2$. Set $\mathbf{Y} := \mathbf{D}_1^{-1}\mathbf{X}$ and $\mathbf{T} := \mathbf{D}_2^{-1}\mathbf{S}$, that is, normalize $\mathbf{X}$ and $\mathbf{S}$ to covariance $\mathbf{I}$. Then

$$\mathbf{Y} = \mathbf{D}_1^{-1}\mathbf{X} = \mathbf{D}_1^{-1}\mathbf{AS} = \mathbf{D}_1^{-1}\mathbf{A}\mathbf{D}_2\mathbf{T}$$

and $\mathbf{T}$, $\mathbf{D}_1^{-1}\mathbf{A}\mathbf{D}_2$ and $\mathbf{Y}$ satisfy the assumption, and $\mathbf{D}_1^{-1}\mathbf{A}\mathbf{D}_2$ is orthogonal because

$$\begin{aligned}\mathbf{I} &= \mathrm{Cov}(\mathbf{Y}) \\ &= E(\mathbf{Y}\mathbf{Y}^\top) \\ &= E(\mathbf{D}_1^{-1}\mathbf{A}\mathbf{D}_2\mathbf{T}\mathbf{T}^\top\mathbf{D}_2\mathbf{A}^\top\mathbf{D}_1^{-1}) \\ &= (\mathbf{D}_1^{-1}\mathbf{A}\mathbf{D}_2)(\mathbf{D}_1^{-1}\mathbf{A}\mathbf{D}_2)^\top.\end{aligned}$$

So without loss of generality, let $\mathbf{A}$ be orthogonal.

Now let $\widehat{\mathbf{S}}(\mathbf{s}) := E_{\mathbf{S}}(\exp i\mathbf{s}^\top\mathbf{S})$ be the characteristic function of $\mathbf{S}$. By assumption, the covariance (and hence the mean) of $\mathbf{S}$ exists, so $\widehat{\mathbf{S}} \in \mathcal{C}^2(\mathbb{R}^n, \mathbb{C})$ (Bauer, 1996). Furthermore, since $\mathbf{S}$ is assumed to be independent, its characteristic function is separated: $\widehat{\mathbf{S}} \equiv g_1 \otimes \cdots \otimes g_n$, where $g_i \equiv \widehat{S}_i$. The characteristic function of $\mathbf{A}\mathbf{S}$ can easily be calculated as

$$\widehat{\mathbf{A}\mathbf{S}}(\mathbf{x}) = E_{\mathbf{S}}(\exp i\mathbf{x}^\top\mathbf{A}\mathbf{S}) = \widehat{\mathbf{S}}(\mathbf{A}^\top\mathbf{x}) = g_1 \otimes \cdots \otimes g_n(\mathbf{A}^\top\mathbf{x})$$

for $\mathbf{x} \in \mathbb{R}^n$. Let $\mathbf{B} := (b_{ij}) = \mathbf{A}^\top$. Since $\mathbf{A}\mathbf{S}$ is also assumed to be independent, $f(\mathbf{x}) := \widehat{\mathbf{A}\mathbf{S}}(\mathbf{x}) = g_1 \otimes \cdots \otimes g_n(\mathbf{B}\mathbf{x})$ is separated.

Now assume that $\mathbf{A} \not\sim \mathbf{I}$. Using orthogonality of $\mathbf{B} = \mathbf{A}^\top$, there exist indices $k \neq l$ and $i \neq j$ with $b_{ki}b_{kj} \neq 0$ and $b_{li}b_{lj} \neq 0$. Then according to lemma 4, g_k and g_l satisfy the differential equation 3.3. Together with corollary 2, this shows that both S_k and S_l are gaussian, which is a contradiction to the assumption.

ii. Let $\mathbf{S}$ be an n-dimensional independent random vector with density $p_{\mathbf{S}} \in \mathcal{C}^2(\mathbb{R}^n, \mathbb{R})$ and no gaussian component, and let $\mathbf{A} \in \mathrm{Gl}(n)$. $\mathbf{S}$ is assumed to be independent, so its density factorizes $p_{\mathbf{S}} \equiv g_1 \otimes \cdots \otimes g_n$. The density of $\mathbf{A}\mathbf{S}$ is given by

$$p_{\mathbf{A}\mathbf{S}}(\mathbf{x}) = |\det\mathbf{A}|^{-1}p_{\mathbf{S}}(\mathbf{A}^{-1}\mathbf{x}) = |\det\mathbf{A}|^{-1}g_1 \otimes \cdots \otimes g_n(\mathbf{A}\mathbf{x})$$

for $\mathbf{x} \in \mathbb{R}^n$. Let $\mathbf{B} := (b_{ij}) = \mathbf{A}^{-1}$. $\mathbf{A}\mathbf{S}$ is also assumed to be independent, so

$$f(\mathbf{x}) := |\det\mathbf{A}|p_{\mathbf{A}\mathbf{S}}(\mathbf{x}) = g_1 \otimes \cdots \otimes g_n(\mathbf{B}\mathbf{x})$$

is separated.

Assume $\mathbf{A} \not\sim \mathbf{I}$. Then also $\mathbf{B} = \mathbf{A}^{-1} \not\sim \mathbf{I}$, so there exist indices l and $i \neq j$ with $b_{li}b_{lj} \neq 0$. Hence, it follows from lemma 4 that g_l satisfies the differential equation 3.3. But g_l is a density, so according to corollary 1 the lth component of $\mathbf{S}$ is gaussian, which is a contradiction.

4 BSS by Hessian Diagonalization

In this section, we use the theory already set out to propose an algorithm for linear BSS, which can be easily extended to nonlinear settings as well. For this, we restrict ourselves to using $\mathcal{C}^2$-densities. A similar idea has already been proposed in Lin (1998), but without dealing with possibly degenerated eigenspaces in the Hessian. Equivalently, we could also use characteristic functions instead of densities, which leads to a related algorithm (Yeredor, 2000).

If we assume that $\mathrm{Cov}(\mathbf{S})$ exists, we can use whitening as seen in the proof of theorem 2i (in this context, also called principal component analysis) to reduce the general BSS model, equation 3.2, to

$$\mathbf{X} = \mathbf{AS} \tag{4.1}$$

with an independent n-dimensional random vector $\mathbf{S}$ with existing covariance $\mathbf{I}$ and an orthogonal matrix $\mathbf{A}$. Then $\mathrm{Cov}(\mathbf{X}) = \mathbf{I}$. We assume that $\mathbf{S}$ admits a $\mathcal{C}^2-$density $p_{\mathbf{S}}$. The density of $\mathbf{X}$ is then given by

$$p_{\mathbf{X}}(\mathbf{x}) = p_{\mathbf{S}}(\mathbf{A}^\top \mathbf{x})$$

for $\mathbf{x} \in \mathbb{R}^n$, because of the orthogonality of $\mathbf{A}$. Hence,

$$p_{\mathbf{S}} \equiv p_{\mathbf{X}} \circ \mathbf{A}.$$

Note that the Hessian of the composition of a function $f \in \mathcal{C}^2(\mathbb{R}^n, \mathbb{R})$ with an $n \times n$-matrix $\mathbf{A}$ can be calculated using the Hessian of f as follows:

$$\mathbf{H}_{f\circ\mathbf{A}}(\mathbf{x}) = \mathbf{A}\mathbf{H}_f(\mathbf{Ax})\mathbf{A}^\top.$$

Let $\mathbf{s} \in \mathbb{R}^n$ with $p_{\mathbf{S}}(\mathbf{s}) > 0$. Then locally at $\mathbf{s}$, we have

$$\mathbf{H}_{\ln p_{\mathbf{S}}}(\mathbf{s}) = \mathbf{H}_{\ln p_{\mathbf{X}}\circ\mathbf{A}}(\mathbf{s}) = \mathbf{A}\mathbf{H}_{\ln p_{\mathbf{X}}}(\mathbf{As})\mathbf{A}^\top. \tag{4.2}$$

$p_{\mathbf{S}}$ is assumed to be separated, so $\mathbf{H}_{\ln p_{\mathbf{S}}}(\mathbf{s})$ is diagonal, as seen in section 2.

Lemma 5. *Let $\mathbf{X} := \mathbf{AS}$ with an orthogonal matrix $\mathbf{A}$ and $\mathbf{S}$, an independent random vector with $\mathcal{C}^2$-density, and at most one gaussian component. Then there exists an open set $U \subset \mathbb{R}^n$ such that for all $\mathbf{x} \in U$, $p_{\mathbf{X}}(\mathbf{x}) \neq 0$ and $\mathbf{H}_{\ln p_{\mathbf{X}}}(\mathbf{x})$ has n different eigenvalues.*

Proof. Assume not. Then there exists no $\mathbf{x} \in \mathbb{R}^n$ at all with $p_{\mathbf{X}}(\mathbf{x}) \neq 0$ and $\mathbf{H}_{\ln p_{\mathbf{X}}}(\mathbf{x})$ having n different eigenvalues because otherwise, due to continuity, these conditions would also hold in an open neighborhood of $\mathbf{x}$.

Using equation 4.2 the logarithmic Hessian of $p_{\mathbf{S}}$ has at every $\mathbf{s} \in \mathbb{R}^n$ with $p_{\mathbf{S}}(\mathbf{s}) > 0$ at least two of the same eigenvalues, say, $\lambda(\mathbf{s}) \in \mathbb{R}$. Hence, since $\mathbf{S}$ is independent, $\mathbf{H}_{\ln p_{\mathbf{S}}}(\mathbf{s})$ is diagonal, so locally,

$$\left(\ln p_{S_i}\right)''(\mathbf{s}) = \left(\ln p_{S_j}\right)''(\mathbf{s}) = \lambda(\mathbf{s})$$

for two indices $i \neq j$. Here, we have used continuity of $\mathbf{s} \mapsto \mathbf{H}_{\ln p_{\mathbf{S}}}(\mathbf{s})$ showing that the two eigenvalues locally lie in the same two dimensions i and j. This proves that $\lambda(\mathbf{s})$ is locally constant in directions i and j. So locally at points $\mathbf{s}$ with $p_{\mathbf{S}}(\mathbf{s}) > 0$, S_i and S_j are of the type $\exp P$, with P being a polynomial of degree ≤ 2. The same argument as in the proof of lemma 3 then shows that p_{S_i} and p_{S_j} have no zeros at all. Using the connectedness of $\mathbb{R}$ proves that S_i and S_j are globally of the type $\exp P$, hence gaussian (because of $\int_{\mathbb{R}} p_{S_k} = 1$), which is a contradiction.

Hence, we can assume that we have found $\mathbf{x}^{(0)} \in \mathbb{R}^n$ with $\mathbf{H}_{\ln p_{\mathbf{X}}}(\mathbf{x}^{(0)})$ having n different eigenvalues (which is equivalent to saying that every eigenvalue is of multiplicity one), because due to lemma 5, this is an open condition, which can be found algorithmically. In fact, most densities in practice turn out to have logarithmic Hessians with n different eigenvalues almost everywhere. In theory however, U in lemma 5 cannot be assumed to be, for example, dense or $\mathbb{R}^n \setminus U$ to have measure zero, because if we choose p_{S_1} to be a normalized gaussian and p_{S_2} to be a normalized gaussian with a very localized small perturbation at zero only, then U cannot be larger than $(-\varepsilon, \varepsilon) \times \mathbb{R}$.

By diagonalization of $\mathbf{H}_{\ln p_{\mathbf{X}}}(\mathbf{x}^{(0)})$ using eigenvalue decomposition (principal axis transformation), we can find the (orthogonal) mixing matrix $\mathbf{A}$. Note that the eigenvalue decomposition is unique except for permutation and sign scaling because every eigenspace (in which $\mathbf{A}$ is only unique up to orthogonal transformation) has dimension one. Arbitrary scaling indeterminacy does not occur because we have forced $\mathbf{S}$ and $\mathbf{X}$ to have unit variances. Using uniqueness of eigenvalue decomposition and theorem 2, we have shown the following theorem:

Theorem 3 (BSS by Hessian calculation). *Let* $\mathbf{X} = \mathbf{AS}$ *with an independent random vector* $\mathbf{S}$ *and an orthogonal matrix* $\mathbf{A}$. *Let* $\mathbf{x} \in \mathbb{R}^n$ *such that locally at* $\mathbf{x}$, $\mathbf{X}$ *admits a* C^2*-density* $p_{\mathbf{X}}$ *with* $p_{\mathbf{X}}(\mathbf{x}) \neq 0$. *Assume that* $\mathbf{H}_{\ln p_{\mathbf{X}}}(\mathbf{x})$ *has n different eigenvalues (see lemma 5). If*

$$\mathbf{E}\mathbf{H}_{\ln p_{\mathbf{X}}}(\mathbf{x})\mathbf{E}^{\top} = \mathbf{D}$$

is an eigenvalue decomposition of the Hessian of the logarithm of $p_{\mathbf{X}}$ *at* $\mathbf{x}$, *that is,* $\mathbf{E}$ *orthogonal,* $\mathbf{D}$ *diagonal, then* $\mathbf{E} \sim \mathbf{A}$, *so* $\mathbf{E}^{\top}\mathbf{X}$ *is independent.*

Furthermore, it follows from this theorem that linear BSS is a local problem as proven already in Theis, Puntonet, and Lang (2003) using the restriction of a random vector.

4.1 Example for Hessian Diagonalization BSS. In order to illustrate the algorithm of local Hessian diagonalization, we give a two-dimensional example. Let $\mathbf{S}$ be a random vector with densities

$$p_{S_1}(s_1) = \frac{1}{2}\chi_{[-1,1]}(s_1)$$
$$p_{S_2}(s_2) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}s_2^2\right)$$

where $\chi_{[-1,1]}$ is one on $[-1, 1]$ and zero everywhere else. The orthogonal mixing matrix $\mathbf{A}$ is chosen to be

$$\mathbf{A} = \frac{1}{\sqrt{2}}\begin{pmatrix} 1 & 1 \\ -1 & 1 \end{pmatrix}.$$

The mixture density $p_{\mathbf{X}}$ of $\mathbf{X} := \mathbf{AS}$ then is ($\det \mathbf{A} = 1$),

$$p_{\mathbf{X}}(\mathbf{x}) = \frac{1}{2\sqrt{2\pi}}\chi_{[-1,1]}\left(\frac{1}{\sqrt{2}}(x_1 - x_2)\right) \exp\left(-\frac{1}{4}(x_1 + x_2)^2\right),$$

for $\mathbf{x} \in \mathbb{R}^2$. $p_{\mathbf{X}}$ is positive and $\mathcal{C}^2$ in a neighborhood around $\mathbf{0}$. Then

$$\partial_1 \ln p_{\mathbf{X}}(\mathbf{x}) = \partial_2 \ln p_{\mathbf{X}}(\mathbf{x}) = -\frac{1}{2}(x_1 + x_2)$$
$$\partial_1^2 \ln p_{\mathbf{X}}(\mathbf{x}) = \partial_2^2 \ln p_{\mathbf{X}}(\mathbf{x}) = \partial_1\partial_2 \ln p_{\mathbf{X}}(\mathbf{x}) = -\frac{1}{2}$$

for $\mathbf{x}$ with $|\mathbf{x}| < \frac{1}{2}$, and the Hessian of the logarithmic densities is

$$\mathbf{H}_{\ln p_{\mathbf{X}}}(\mathbf{x}) = -\frac{1}{2}\begin{pmatrix} 1 & 1 \\ 1 & 1 \end{pmatrix}$$

independent on $\mathbf{x}$ in a neighborhood around $\mathbf{0}$. Diagonalization of $\mathbf{H}_{\ln p_{\mathbf{X}}}(\mathbf{0})$ yields

$$\begin{pmatrix} -1 & 0 \\ 0 & 0 \end{pmatrix},$$

and this equals $\mathbf{A}\mathbf{H}_{\ln p_{\mathbf{X}}}(\mathbf{0})\mathbf{A}^\top$, as stated in theorem 3.

4.2 Global Hessian Diagonalization Using Kernel-Based Density Approximation. In practice, it is usually not possible to approximate the density locally with sufficiently high accuracy, so a better approximation using the typically global information of $\mathbf{X}$ has to be found. We suggest using kernel-based density estimation to get an energy function with minima at the BSS solutions together with a global Hessian diagonalization in the following.

The idea is to construct a measure for separatedness of the densities (hence independence) based on theorem 1. A possible measure could be the norm of the summed-up separators $\sum_{i<j} R_{ij}[f]$. In order for this to be calculable, we choose only a set of points $\mathbf{p}^{(i)}$ where we evaluate the difference and minimize $\sum_k \sum_{i<j} R_{ij}[f](\mathbf{p}^{(k)})^2$ at those points. Although in the linear noiseless case, calculation of the Hessian at only one point would be enough, using an energy function of this type ensures using global information of the densities while averaging over possible local errors.

First, we need to approximate the density function. For this, let $\mathbf{X} \in \mathbb{R}^n$ be an n-dimensional random vector with ν independent and identically distributed samples $\mathbf{x}^{(1)}, \ldots, \mathbf{x}^{(\nu)} \in \mathbb{R}^n$. Let

$$\begin{aligned} \varphi : \mathbb{R}^n &\to \mathbb{R} \\ \mathbf{x} &\mapsto \frac{1}{\sigma^n \sqrt{(2\pi)^n}} \exp\left(-\frac{1}{2\sigma^2}\|\mathbf{x}\|^2\right) \end{aligned}$$

be the n-dimensional-centered independent gaussian with fixed variance $\sigma^2 > 0$. For ease of notation, denote $\kappa := \frac{1}{2\sigma^2}$.

Define the approximated density $\hat{p}_{\mathbf{X}}$ of $\mathbf{X}$ by

$$\hat{p}_{\mathbf{X}}(\mathbf{x}) := \frac{1}{\nu} \sum_{i=1}^{\nu} \varphi(\mathbf{x} - \mathbf{x}^{(i)}). \tag{4.3}$$

If $l \to \infty$, $\hat{p}_{\mathbf{X}}$ converges to $p_{\mathbf{X}}$ in the space of all integrable functions if σ is chosen appropriately. This can be shown using the central limit theorem. Figure 2 depicts the approximation of a Laplacian using equation 4.3.

The partial derivatives of φ can be calculated as

$$\begin{aligned} \partial_i \varphi(\mathbf{x}) &= -2\kappa x_i \varphi(\mathbf{x}) \\ \partial_i \partial_j \varphi(\mathbf{x}) &= 4\kappa^2 x_i x_j \varphi(\mathbf{x}) \end{aligned} \tag{4.4}$$

for $i \neq j$. φ is separated, so $R[\varphi] \equiv 0$. Note that $\hat{p}_{\mathbf{X}} \in \mathcal{C}^\infty(\mathbb{R}^n, \mathbb{R})$ is positive. So according to theorem 1 $\hat{p}_{\mathbf{X}}$ is separated if and only if $R_{ij}[\hat{p}_{\mathbf{X}}] \equiv 0$ for $i < j$. And since $\hat{p}_{\mathbf{X}}$ is an approximation of $p_{\mathbf{X}}$, separatedness of $\hat{p}_{\mathbf{X}}$ also induces approximate independence of $\mathbf{X}$.

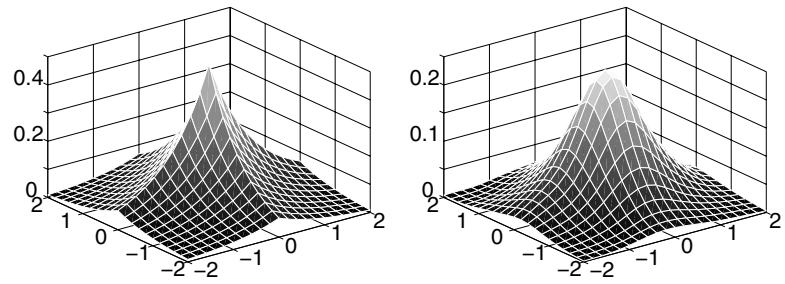

Figure 2: Independent Laplacian density $p_{\mathbf{S}}(\mathbf{s}) = \frac{1}{2}\exp(-|x_1| - |x_2|)$: theoretic (left) and approximated (right) densities. For the approximation, 1000 samples and gaussian kernel approximation (see equation 4.3) with standard deviation 0.37 were used.

$R_{ij}[\hat{p}_{\mathbf{X}}]$ can be calculated using lemma 2—here $R_{ij}[\varphi(\mathbf{x} - \mathbf{x}^{(k)})] \equiv 0$—and equation 4.4:

$$\begin{aligned}
R_{ij}[\hat{p}_{\mathbf{X}}](\mathbf{x}) &= \frac{1}{\nu^2} R_{ij}\left[\sum_{k=1}^{\nu} \varphi(\mathbf{x} - \mathbf{x}^{(k)})\right] \\
&= \frac{1}{\nu^2} \sum_{k \neq l} \varphi\left(\mathbf{x} - \mathbf{x}^{(k)}\right) \partial_i \partial_j \varphi\left(\mathbf{x} - \mathbf{x}^{(l)}\right) \\
&\quad - \left(\partial_i \varphi\right)\left(\mathbf{x} - \mathbf{x}^{(k)}\right) \left(\partial_j \varphi\right)\left(\mathbf{x} - \mathbf{x}^{(l)}\right) \\
&= \frac{4\kappa^2}{\nu^2} \sum_{k \neq l} \varphi\left(\mathbf{x} - \mathbf{x}^{(k)}\right) \varphi\left(\mathbf{x} - \mathbf{x}^{(l)}\right) \left(x_i^{(k)} - x_i^{(l)}\right) \left(x_j - x_j^{(l)}\right) \\
&= \frac{4\kappa^2}{\nu^2} \sum_{k < l} \varphi\left(\mathbf{x} - \mathbf{x}^{(k)}\right) \varphi\left(\mathbf{x} - \mathbf{x}^{(l)}\right) \left(x_i^{(k)} - x_i^{(l)}\right) \left(x_j^{(k)} - x_j^{(l)}\right).
\end{aligned}$$

This function is zero for $i < j$ if and only if $\hat{p}_{\mathbf{X}}$ is separated. For linear BSS, it would be enough to check this at one point in general position (see theorem 3), but for robustness, we want to require $R_{ij}[\hat{p}_{\mathbf{X}}]$ to be zero (or as close to zero as possible) at all sample points. So the desired independence estimator can be calculated as

$$E(\mathbf{x}^1, \ldots, \mathbf{x}^{(n)}) := E := \sum_{m=1}^{\nu} \sum_{i<j} (R_{ij}[\hat{p}_{\mathbf{X}}](\mathbf{x}^{(m)}))^2;$$

hence,

$$E = (\sigma^2\nu)^{-4} \sum_m \sum_{i<j} \left(\sum_{k<l} \varphi(\mathbf{x} - \mathbf{x}^{(k)})\varphi(\mathbf{x} - \mathbf{x}^{(l)})(x_i^{(k)} - x_i^{(l)})(x_j^{(k)} - x_j^{(l)}) \right)^2 .$$

Minimizing the function

$$\begin{aligned} \varepsilon : \mathrm{Gl}(n) &\to \mathbb{R} \\ \mathbf{W} &\mapsto E(\mathbf{W}\mathbf{x}^1, \ldots, \mathbf{W}\mathbf{x}^{(n)}) \end{aligned}$$

then yields the desired demixing matrix with $\mathbf{W}^{-1} \sim \mathbf{A}$ according to theorem 2. ε can be minimized using the usual techniques—for example, global search, gradient descent, or fixed-point search. Figure 3 shows the energy function of an example mixture of two Laplacians. E is minimal at the points where $\mathbf{WX}$ is independent.

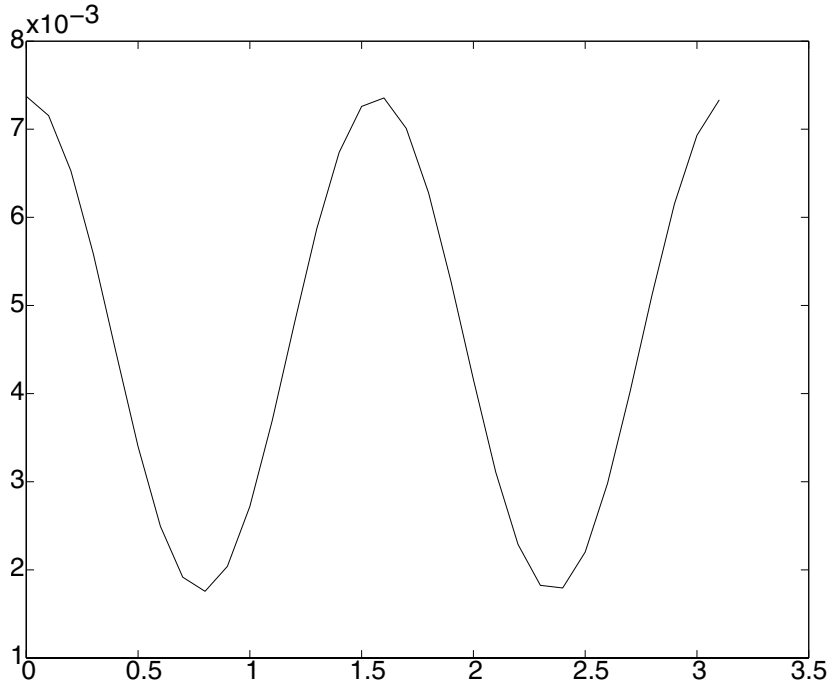

Figure 3: Energy function $\mathbf{W} \mapsto E(\mathbf{WX})$ of a mixture $\mathbf{X}$ of two Laplacians using as mixing matrix $\mathbf{A}$ a rotation by 45 degrees. One hundred samples were used, and E is plotted in steps of 0.1. The minima of E clearly lie at $\frac{1}{4}\pi$ and $\frac{3}{4}\pi$, as desired.

Note that E represents a new approximate measure of independence. Therefore, the linear BSS algorithm can now be readily generalized to nonlinear situations by finding an appropriate parameterization of the possibly nonlinear separating model.

The proposed algorithm basically performs a global diagonalization of the logarithmic Hessian after prewhitening. Interestingly, this is similar to traditional BSS algorithms based on joint diagonalization such as JADE (Cardoso & Souloumiac, 1993) using cumulant matrices or AMUSE (Tong, Liu, Soan, & Huang, 1991) and SOBI (Belouchrani, Meraim, Cardoso, & Moulines, 1997) employing time decorrelation. Instead of using a global energy function as proposed above, we could therefore also jointly diagonalize a given set of Hessians (respectively, separator matrices, as above; see also Yeredor, 2000). Another relation to previously proposed ICA algorithms lies in the kernel approximation technique. Gaussian or generalized gaussian kernels have already been used in the field of independent component analysis to model the source densities (Lee & Lewicki, 2000; Habl, Bauer, Putonet, Rodriguez-Alvarez, & Lang, 2001), thus giving an estimate of the score function used in Bell-Sejnowski-type semiparametric algorithms (Bell, & Sejnowski, 1995) or enabling direct separation using a maximum likelihood parameter estimation. Our algorithm also uses density approximation, but employs this for the mixture density, which can be problematic in higher dimensions. A different approach not involving density approximation is a direct sample-based Hessian estimation similar to Lin (1998).

5 Separability of Postnonlinear BSS

In this section, we show how to use the idea of Hessian diagonalization in order to give separability proofs in nonlinear situations, more precisely, in the setting of postnonlinear BSS. After stating the postnonlinear BSS model and the general (to the knowledge of the author, not yet proven) separability theorem, we will prove postnonlinear separability in the case of random vectors with distributions that are somewhere locally constant and nonzero (e.g., uniform distributions). A possible proof of postnonlinear separability has been suggested by Taleb and Jutten (1999); however, the proof applies only to densities with at least one zero and furthermore contains an error rendering the proof applicable only to restricted situations.

Definition 3. *A function* $\mathbf{f} : \mathbb{R}^n \to \mathbb{R}^n$ *is called* diagonal *if each component* $f_i(\mathbf{x})$ *of* $\mathbf{f}(\mathbf{x})$ *depends on only the variable* x_i.

In this case, we often omit the other variables and write $\mathbf{f}(x_1, \ldots, x_n) = (f_1(x_1), \ldots, f_n(x_n))$; so $\mathbf{f} \equiv f_1 \times \cdots \times f_n$ where $\times$ denotes the Cartesian product.

Consider now the postnonlinear BSS model,

$$\mathbf{X} = \mathbf{f}(\mathbf{AS}), \tag{5.1}$$

where again $\mathbf{S}$ is an independent random vector, $\mathbf{A} \in \mathrm{Gl}(n)$, and $\mathbf{f}$ is a diagonal nonlinearity. We assume the components of $\mathbf{f}$ to be injective analytical functions with invertible Jacobian at every point (locally diffeomorphic).

Definition 4. *An invertible matrix $\mathbf{A} \in \mathrm{Gl}(n)$ is said to be* mixing *if $\mathbf{A}$ has at least two nonzero entries in each row.*

Note that if $\mathbf{A}$ is mixing, then $\mathbf{A}'$, $\mathbf{A}^{-1}$, and $\mathbf{ALP}$ for scaling matrix $\mathbf{L}$ and permutation matrix $\mathbf{P}$ are also mixing.

Postnonlinear BSS is a generalization of linear BSS, so the indeterminacies of postnonlinear ICA contain at least the indeterminacies of linear BSS: $\mathbf{A}$ can be reconstructed only up to scaling and permutation. In the linear case, affine linear transformation is ignored. Here, of course, additional indeterminacies come into play because of translation: f_i can be recovered only up to a constant. Also, if $\mathbf{L} \in \mathrm{Gl}(n)$ is a scaling matrix, then

$$\mathbf{f}(\mathbf{AS}) = (\mathbf{f} \circ \mathbf{L})((\mathbf{L}^{-1}\mathbf{A})\mathbf{S}),$$

so $\mathbf{f}$ and $\mathbf{A}$ can interchange scaling factors in each component. Another obvious indeterminacy could occur if $\mathbf{A}$ is not general enough. If, for example, $\mathbf{A} = \mathbf{I}$, then $\mathbf{f}(\mathbf{S})$ is already independent, because independence is invariant under diagonal nonlinear transformation; so $\mathbf{f}$ cannot be found in this case. If we assume, however, that $\mathbf{A}$ is mixing, then we will show that except for scaling interchange between $\mathbf{f}$ and $\mathbf{A}$, no more indeterminacies than in the affine linear case exist.

Theorem 4 (separability of postnonlinear BSS). *Let $\mathbf{A}, \mathbf{W} \in \mathrm{Gl}(n)$ be mixing, $\mathbf{h} : \mathbb{R}^n \to \mathbb{R}^n$ be a diagonal bijective function with analytical locally diffeomorphic components, and $\mathbf{S}$ be an independent random vector with at most one gaussian component and existing covariance. If $\mathbf{W}(\mathbf{h}(\mathbf{AS}))$ is independent, then there exists a scaling matrix $\mathbf{L} \in \mathrm{Gl}(n)$ and $\mathbf{p} \in \mathbb{R}^n$ with $\mathbf{LA} \sim \mathbf{W}^{-1}$ and $\mathbf{h} \equiv \mathbf{L}+\mathbf{p}$.*

If analyticity of the components of $\mathbf{h}$ is not assumed, then $\mathbf{h} \equiv \mathbf{L} + \mathbf{p}$ can only hold on $\{\mathbf{As} | p_{\mathbf{S}}(\mathbf{s}) \neq 0\}$.

If $\mathbf{f} \circ \mathbf{A}$ is the mixing model, $\mathbf{W} \circ \mathbf{g}$ is the separating model. Putting the two together, we get the above mixing-separating model. Since $\mathbf{A}$ has to be assumed to be mixing, we can assume $\mathbf{W}$ to be mixing as well because the inverse of a matrix that is mixing is again mixing. Furthermore, the mixing-separating model is assumed to be bijective—hence, $\mathbf{A}$ and $\mathbf{W}$ invertible and $\mathbf{h}$ bijective—because otherwise trivial solutions as, for example, $\mathbf{h} \equiv c$ for a constant $c \in \mathbb{R}$, would also be solutions.

We will show the theorem in the case of $\mathbf{S}$ and $\mathbf{X}$ with components having somewhere locally constant nonzero $\mathcal{C}^2$-densities. An alternative geometric idea of how to prove theorem 4 for bounded sources in two dimensions is mentioned in Babaie-Zadeh, Jutten, and Nayebi (2002) and extended in Theis and Gruber (forthcoming). Note that in our case, as well as in the above restrictive cases, the assumption that $\mathbf{S}$ has at most one gaussian component holds trivially.

Proof of Theorem 4 *(with Locally-Constant Nonzero $\mathcal{C}^2$-Densities).* Let $\mathbf{h} = h_1 \times \cdots \times h_n$ with bijective $\mathcal{C}^\infty$-functions $h_i : \mathbb{R} \to \mathbb{R}$. We have to show only that the h_i' are constant. Then $\mathbf{h}$ is affine linear, say, $\mathbf{h} \equiv \mathbf{L} + \mathbf{p}$, with diagonal matrix $\mathbf{L} \in \mathrm{Gl}(n)$ and a vector $\mathbf{p} \in \mathbb{R}^n$. Hence, $\mathbf{WLAS} + \mathbf{Wp}$, and then $\mathbf{WLAS}$ is independent, so using linear separability, theorem 2i, $\mathbf{WLA} \sim \mathbf{I}$, therefore, $\mathbf{LA} \sim \mathbf{W}^{-1}$.

Let $\mathbf{X} := \mathbf{W}(\mathbf{h}(\mathbf{AS}))$. The density of this transformed random vector is easily calculated from $\mathbf{S}$:

$$p_{\mathbf{X}}(\mathbf{Wh}(\mathbf{As})) = |\det \mathbf{W}|^{-1} |h_1'((\mathbf{As})_1)|^{-1} \cdots |h_n'((\mathbf{As})_n)|^{-1} |\det \mathbf{A}|^{-1} p_{\mathbf{S}}(\mathbf{s})$$

for $\mathbf{s} \in \mathbb{R}^n$. $\mathbf{h}$ has by assumption an invertible Jacobian at every point, so the h_i' are either positive or negative; without loss of generality, $h_i' > 0$. Furthermore, $p_{\mathbf{X}}$ is independent, so we can write

$$p_{\mathbf{X}} \equiv g_1 \otimes \cdots \otimes g_n.$$

For fixed $\mathbf{s}^0 \in \mathbb{R}^n$ with $p_{\mathbf{S}}(\mathbf{s}^0) > 0$, there exists an open neighborhood $U \subset \mathbb{R}^n$ of $\mathbf{s}^0$ with $p_{\mathbf{S}}|U > 0$ and $p_{\mathbf{S}}|U \in \mathcal{C}^2(U, \mathbb{R})$. If we define $f(\mathbf{s}) := \ln\left(|\det \mathbf{W}|^{-1} |\det \mathbf{A}|^{-1} p_{\mathbf{S}}(\mathbf{s})\right)$ for $\mathbf{s} \in U$, then

$$\begin{aligned} f(\mathbf{s}) &= \ln(h_1'((\mathbf{As})_1) \cdots h_n'((\mathbf{As})_n) g_1((\mathbf{Wh}(\mathbf{As}))_1) \cdots g_n((\mathbf{Wh}(\mathbf{As}))_n)) \\ &= \sum_{k=1}^{n} \ln h_k'((\mathbf{As})_k) + \zeta_k((\mathbf{Wh}(\mathbf{As}))_k) \end{aligned}$$

for $\mathbf{x} \in \mathbb{R}^n$ where $\zeta_k := \ln g_k$ locally at s_k^0. $p_{\mathbf{S}}$ is separated, so

$$\partial_i \partial_j f \equiv 0 \tag{5.2}$$

for $i < j$. Denote $\mathbf{A} =: (a_{ij})$ and $\mathbf{W} =: (w_{ij})$. The first derivative and then the nondiagonal entries in the Hessian of f can be calculated as follows ($i < j$):

$$\partial_i f(\mathbf{s}) = \sum_{k=1}^{n} a_{ki} \frac{h_k''}{h_k'}((\mathbf{As})_k) + \zeta_k'((\mathbf{Wh}(\mathbf{As}))_k) \left(\sum_{l=1}^{n} w_{kl} a_{li} h_l'((\mathbf{As})_l) \right)$$

$$\begin{aligned}\partial_i\partial_j f(\mathbf{s}) = &\sum_{k=1}^{n} a_{ki}a_{kj}\frac{h_k'h_k''' - h_k''^2}{h_k'^2}((\mathbf{As})_k)\\ &+ \zeta_k''((\mathbf{Wh}(\mathbf{As}))_k)\left(\sum_{l=1}^{n} w_{kl}a_{li}h_l'((\mathbf{As})_l)\right)\left(\sum_{l=1}^{n} w_{kl}a_{lj}h_l'((\mathbf{As})_l)\right)\\ &+ \zeta_k'((\mathbf{Wh}(\mathbf{As}))_k)\left(\sum_{l=1}^{n} w_{kl}a_{li}a_{lj}h_l''((\mathbf{As})_l)\right).\end{aligned}$$

Substituting $\mathbf{y} := \mathbf{As}$ and using equation 5.2, we finally get the following differential equation for the h_k:

$$\begin{aligned}0 = &\sum_{k=1}^{n} a_{ki}a_{kj}\frac{h_k'h_k''' - h_k''^2}{h_k'^2}(y_k)\\ &+ \zeta_k''((\mathbf{Wh}(\mathbf{y}))_k)\left(\sum_{l=1}^{n} w_{kl}a_{li}h_l'(y_l)\right)\left(\sum_{l=1}^{n} w_{kl}a_{lj}h_l'(y_l)\right)\\ &+ \zeta_k'((\mathbf{Wh}(\mathbf{y}))_k)\left(\sum_{l=1}^{n} w_{kl}a_{li}a_{lj}h_l''(y_l)\right)\end{aligned} \tag{5.3}$$

for $\mathbf{y} \in V := \mathbf{A}(U)$.

We will restrict ourselves to the simple case mentioned above in order to solve this equation. We assume that the h_k are analytic and that there exists $\mathbf{x}^0 \in \mathbb{R}^n$ where the demixed densities g_k are locally constant and nonzero. Consider the above calculation around $\mathbf{s}^0 = \mathbf{A}^{-1}(\mathbf{h}^{-1}(\mathbf{W}^{-1}\mathbf{x}^0))$.

Choose the open set V such that the g_k are locally constant nonzero on $\mathbf{h}(\mathbf{W}(V))$. Then so are the $\zeta_k' = \ln g_k$, and therefore

$$0 = \sum_{k=1}^{n} a_{ki}a_{kj}\frac{h_k'h_k''' - h_k''^2}{h_k'^2}(y_k)$$

for $\mathbf{y} \in V$. Hence, there exist open intervals $I_k \subset \mathbb{R}$ and constants $b_k \in \mathbb{R}$ with

$$a_{ki}a_{kj}\left(h_k'h_k''' - h_k''^2\right) \equiv d_k h_k'^2$$

on I_k (here, $d_k = \sum_{l\neq k} a_{li}a_{lj}\frac{h_l'h_l''' - h_l''^2}{h_l'^2}(y_l)$ for some (and then any) $\mathbf{y} \in V$).

By assumption, $\mathbf{W}$ is mixing. Hence, for fixed k, there exist $i \neq j$ with $a_{ki}a_{kj} \neq 0$. If we set $c_k := \frac{b_k}{a_{ki}a_{kj}}$, then

$$c_k h_k'^2 - h_k'h_k''' + h_k''^2 \equiv 0 \tag{5.4}$$

on I_k. h_k was chosen to be analytic, and equation 5.4 holds on the open set I_k, so it holds on all $\mathbb{R}$. Applying lemma 3 then shows that either $h_k' \equiv 0$ or

$$h_k'(x) = \pm \exp\left(\frac{c_k}{2}x^2 + d_k x + e_k\right), x \in \mathbb{R} \tag{5.5}$$

with constants $d_k, e_k \in \mathbb{R}$. By assumption, h_k is bijective, so $h_k' \not\equiv 0$.

Applying the same arguments as above to the inverse system

$$\mathbf{S} = \mathbf{A}^{-1}(\mathbf{h}^{-1}(\mathbf{W}^{-1}\mathbf{X}))$$

and using the fact that also $p_{\mathbf{S}}$ is somewhere locally constant nonzero shows that equation 5.5 also holds for $(h_k^{-1})'$ with other constants. But if both the derivatives of h_k and h_k^{-1} are of this exponential type, then $c_k = d_k = 0$, and therefore h_k is affine linear for all $k = 1, \ldots, n$, which completes the proof of postnonlinear separability in this special case.

Note that in the above proof, local positiveness of the densities was assumed in order to use the equivalence of local separability with the diagonality of the logarithm of the Hessian. Hence, these results can be generalized using theorem 1 in a similar fashion as we did in the linear case with theorem 2. Hence, we have proven postnonlinear separability also for uniformly distributed sources.

6 Conclusion

We have shown how to derive the separability of linear BSS using diagonalization of the Hessian of the logarithmic density, respectively, characteristic function. This induces separated, that is, independent, sources. The idea of Hessian diagonalization is put into a new algorithm for performing linear independent component analysis, which is shown to be a local problem. In practice, however, due to the fact that the densities cannot be approximated locally very well, we also propose a diagonalization algorithm that takes the global structure into account. In order to show the use of this framework of separated functions, we finish with a proof of postnonlinear separability in a special case.

In future work, more general separability results of postnonlinear BSS could be constructed by finding more general solutions of the differential equation 5.3. Algorithmic improvements could be made by using other density approximation methods like mixture of gaussian models or by approximating the Hessian itself using the cumulative density and discrete approximations of the differential. Finally, the diagonalization algorithm can easily be extended to nonlinear situations by finding appropriate model parameterizations; instead of minimizing the mutual information, we minimize the absolute value of the off-diagonal terms of the logarithmic Hessian.

The algorithm has been specified using only an energy function; gradient and fixed-point algorithms can be derived in the usual manner.

Separability in nonlinear situations has turned out to be a hard problem—illposed in the most general case (Hyvärinen & Pajunen, 1999)—and not many nontrivial results exist for restricted models (Hyvärinen & Pajunen, 1999; Babaie-Zadeh et al., 2002), all only two-dimensional. We believe that this is due to the fact that the rather nontrivial proof of the Darmois-Skitovitch theorem is not at all easily generalized to more general settings (Kagan, 1986). By introducing separated functions, we are able to give a much easier proof for linear separability and also provide new results in nonlinear settings. We hope that these ideas will be used to show separability in other situations as well.

Acknowledgments

I thank the anonymous reviewers for their valuable suggestions, which improved the original manuscript. I also thank Peter Gruber, Wolfgang Hackenbroch, and Michaela Theis for suggestions and remarks on various aspects of the separability proof. The work described here was supported by the DFG in the grant "Nonlinearity and Nonequilibrium in Condensed Matter" and the BMBF in the ModKog project.

References

Babaie-Zadeh, M., Jutten, C., & Nayebi, K. (2002). A geometric approach for separating post non-linear mixtures. In *Proc. of EUSIPCO '02* (Volume 2, pp. 11–14). Toulouse, France.

Bauer, H. (1996). *Probability theory*. Berlin: Walter de Gruyter.

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

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

Cardoso, J.-F., & Souloumiac, A. (1993). Blind beamforming for non gaussian signals. *IEEE Proceedings–F*, *140*(6), 362–370.

Cichocki, A., & Amari, S. (2002). *Adaptive blind signal and image processing*. New York: Wiley.

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

Darmois, G. (1953). Analyse générale des liaisons stochastiques. *Rev. Inst. Internationale Statist.*, *21*, 2–8.

Eriksson, J., & Koivunen, V. (2003). Identifiability and separability of linear ica models revisited. In *Proc. of ICA 2003*, (pp. 23–27). Nara, Japan.

Habl, M., Bauer, C., Puntonet, C., Rodriguez-Alvarez, M., & Lang, E. (2001). Analyzing biomedical signals with probabilistic ICA and kernel-based source

density estimation. In M. Sebaaly, (Ed.) *Information science innovations (Proc.ISI'2001)* (pp. 219–225). Alberta, Canada: ICSC Academic Press.

Hérault, J., & Jutten, C. (1986). Space or time adaptive signal processing by neural network models. In J. Denker (Ed.), *Neural networks for computing: Proceedings of the AIP Conference* (pp. 206–211). New York: American Institute of Physics.

Hyvärinen, A., Karhunen, J., & Oja, E. (2001). *Independent component analysis*. New York: Wiley.

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

Hyvärinen, A., & Pajunen, P. (1999). Nonlinear independent component analysis: Existence and uniqueness results. *Neural Networks*, *12*(3), 429–439.

Kagan, A. (1986). New classes of dependent random variables and a generalization of the Darmois-Skitovitch theorem to several forms. *Theory Probab. Appl.*, *33*(2), 286–295.

Lee, T., & Lewicki, M. (2000). The generalized gaussian mixture model using ICA. In *Proc. of ICA 2000* (pp. 239–244). Helsinki, Finland.

Lin, J. (1998). Factorizing multivariate function classes. In M. Kearns, M. Jordan, & S. Solla (Eds.), *Advances in neural information processing systems*, *10*, (pp. 563–569). Cambridge, MA: MIT Press.

Skitovitch, V. (1953). On a property of the normal distribution. *DAN SSSR*, *89*, 217–219.

Taleb, A., & Jutten, C. (1999). Source separation in post non linear mixtures. *IEEE Trans. on Signal Processing*, *47*, 2807–2820.

Theis, F., & Gruber, P. (forthcoming). Separability of analytic postnonlinear blind source separation with bounded sources. In *Proc. of ESANN 2004*. Evere, Belgium: d-side.

Theis, F., Jung, A., Puntonet, C., & Lang, E. (2002). Linear geometric ICA: Fundamentals and algorithms. *Neural Computation*, *15*, 1–21.

Theis, F., Puntonet, C., & Lang, E. (2003). Nonlinear geometric ICA. In *Proc. of ICA 2003* (pp. 275–280). Nara, Japan.

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

Yeredor, A. (2000). Blind source separation via the second characteristic function. *Signal Processing*, *80*(5), 897–902.

Received June 27, 2003; accepted March 8, 2004.