Uniqueness of linear factorizations into independent subspaces

Harold W. Gutch[a,c,*], Fabian J. Theis[b,c]

[a] *Max Planck Institute for Dynamics and Self-Organization, Bunsenstr. 10, 37073 Göttingen, Germany*
[b] *CMB, Institute of Bioinformatics and Systems Biology, German Research Center for Environmental Health, 85764 Neuherberg, Germany*
[c] *Technical University of Munich, Arcisstrasse 21, 80333 München, Germany*

Abstract

Given a random vector $\mathbf{X}$, we address the question of linear separability of $\mathbf{X}$, that is, the task of finding a linear operator $\mathbf{W}$ such that we have $(\mathbf{S}_1, \ldots, \mathbf{S}_M) = (\mathbf{WX})$ with statistically independent random vectors $\mathbf{S}_i$. As this requirement alone is already fulfilled trivially by $\mathbf{X}$ being independent of the empty rest, we require that the components be not further decomposable. We show that if $\mathbf{X}$ has finite covariance, such a representation is unique up to trivial indeterminacies. Related algorithms, however with fixed dimensionality of the subspaces, have already been successfully employed in biomedical applications, such as separation of fMRI recorded data. Based on the presented uniqueness result, it is now clear that also subspace dimensions can be determined in a unique and therefore meaningful fashion, which shows the advantages of Independent Subspace Analysis in contrast to methods like Principal Component Analysis.

Keywords: statistical independence, independent component analysis, independent subspace analysis, separability, inverse models
2000 MSC: 62E10, 62H25

1. Introduction

Assume a random vector $\mathbf{S}$ consisting of statistically independent components S_i, none of which is Gaussian (normally distributed). If the components of a linear mixing $\mathbf{AS}$ then again are statistically independent, one can show that $\mathbf{A}$ is at most the product of a permutation and scaling within the components, which originally was shown using the Darmois-Skitovitch theorem [8, 17, 23]. Under the additional assumption of finite covariance of $\mathbf{S}$, one may additionally allow at most one of the S_i to be Gaussian [7]. The assumption of finite

*Corresponding author
Email addresses: harold.gutch@ds.mpg.de (Harold W. Gutch), fabian.theis@helmholtz-muenchen.de (Fabian J. Theis)

covariance actually is not required for this to hold [10], but if one assumes it, a simpler proof is possible, based on the idea that the characteristic function of $\mathbf{S}$ factorizes, so its logarithm has a diagonal Hessian almost everywhere [25].

This property of random variables has driven the development of algorithms performing so-called Independent Component Analysis (ICA) under some approximations of statistical independence [2, 7, 16, 29]. Exact cost functions, so-called contrasts such as mutual information, are difficult to estimate in practice, where the random vector in question is only known up to some finite precision, so many approximations have been extensively studied. Such algorithms have been successfully used in various fields, e.g. signal processing, biomedical imaging and analysis of financial data, where it was argued that the data sets to be analyzed can be approximated well enough by modeling them as random variables mixed in a linear fashion, see [6, 14] and references therein.

Apart from the question of validity when transferring the mathematical theory to real life data sets, another problem is apparent: What if a given random vector $\mathbf{X}$ has no such representation? This motivates the question if the original claim has a straight-forward extension to higher dimensions: If we write $(\mathbf{S}_1, \ldots, \mathbf{S}_M) := \mathbf{WX}$ with independent random vectors $\mathbf{S}_i$ of which at most one is Gaussian, again, are these unique up to permutation and invertible linear transformations (the multidimensional translation of scaling) within the $\mathbf{S}_i$? The task of finding a basis in which a random vector $\mathbf{X}$ has this property is usually denoted Independent Subspace Analysis (ISA), as one typically reads the independent random vectors gained here as data subsets or data subspaces [5]. Obviously this task requires some minimality constraint, as, given such an independent representation, we might arbitrarily group together some of the components and thus of course get two representations differing in more than just the permutation and linear transformations. Our minimality constraint is the inability to decompose any of the components even further, a property we call *irreducibility* of the components.

Our main result is the following uniqueness theorem:

Theorem 1.1. *The decomposition of a random vector $\mathbf{X}$ with existing covariance into independent, irreducible components is unique up to order and invertible transformations within the components and an invertible transformation in the possibly higher-dimensional Gaussian component.*

This manuscript is structured as follows: In the second section we define the notation used and the framework we work in and state a few simple lemmata used. The main part of this manuscript, the third section, consists of the proof of Theorem 1.1. In the fourth section we shed some light on the practical usefulness of this result, compare it with literature and address some open questions.

Parts of this work were presented at the ICA 2007 conference stating Theorem 1.1, however without proof [11].

2. Definition of Independent Subspace Analysis

We will define the analyzed model and review a few properties of characteristic functions. We restrict this analysis to the real case, that is, real valued random vectors and real linear mixings thereof, although extensions to the complex case are possible.

2.1. Notation

In order to be able to quickly differentiate between scalars and vectors, scalars are depicted in regular font, e.g. $x \in \mathbb{R}$ while vectors and matrices are depicted in bold font, e.g. $\mathbf{x} \in \mathbb{R}^n$. Random values and vectors are always depicted in uppercase letters, e.g. S and $\mathbf{S}$, and we will only need the two letters $\mathbf{S}$, and $\mathbf{X}$ (and regular typeface versions thereof) for these; all other uppercase letters used represent real valued matrices. In order to keep the notation as simple as possible, vectors will often be written in rows, e.g. $\mathbf{x} = \mathbf{A}(\mathbf{v}_1, \mathbf{v}_2)$ instead of $\mathbf{x} = \mathbf{A}(\mathbf{v}_1^\top, \mathbf{v}_2^\top)^\top$. The symbol ∂_i denotes the i-th partial derivative operator, so for a function depending on $\mathbf{x} = (x_1, \ldots, x_n)$, we have $\partial_i = \frac{\partial}{\partial x_i}$. We write $\mathbf{d}f$ for the differential of an $f \in C^1$, and we depict the Hessian of an $f \in C^2$ with $\mathbf{H}_f$, that is $\mathbf{e}_i^\top \mathbf{H}_f \mathbf{e}_j = \partial_i \partial_j f$. We write $\mathbf{d}f|_{\mathbf{x}}$ instead of $(\mathbf{d}f)(\mathbf{x})$, the differential of f evaluated at $\mathbf{x}$, and similarly $\mathbf{H}_f|_{\mathbf{x}}$ instead of $\mathbf{H}_f(\mathbf{x})$, the Hessian of f evaluated at $\mathbf{x}$.

2.2. Irreducibility

Let us now introduce the key notion of irreducibility and point out the special role of Gaussian random vectors.

Definition 2.1. *An n-dimensional random vector $\mathbf{X}$ is said to be* reducible *if it can be written as $\mathbf{X} = \mathbf{A}(\mathbf{S}_1, \mathbf{S}_2)$ with some invertible $n \times n$-matrix $\mathbf{A}$, a k-dimensional random vector $\mathbf{S}_1$ and an $(n-k)$-dimensional random vector $\mathbf{S}_2$, where $\mathbf{S}_1$ is independent of $\mathbf{S}_2$. A random vector that is not reducible is called* irreducible*.*

Remark 2.1. *For example, any n-dimensional Gaussian random vector is reducible if $n > 1$: Gaussians are fully defined by their first and second order moments, so here independence is equivalent to decorrelation, and for every random vector $\mathbf{X}$ with finite covariance there is some invertible matrix $\mathbf{A}$ such that $\mathbf{AX}$ is decorrelated (see Lemma 3.1). Therefore an n-dimensional Gaussian can always be fully reduced to one-dimensional components.*

Obviously both properties, irreducibility and reducibility are preserved under any invertible linear transformation.

A decomposition $(\mathbf{X}_1, \ldots, \mathbf{X}_L) = \mathbf{X}$ of a random vector $\mathbf{X}$ is said to be *independent*, if the random vectors $\mathbf{X}_j$ $(j = 1, \ldots, L)$ are mutually statistically independent. It is said to be *irreducible*, if the vectors $\mathbf{X}_j$ additionally are irreducible.

Remark 2.2. *It is straight-forward to see that for any random vector* $\mathbf{X}$ *there is some invertible matrix* $\mathbf{A}$ *such that* $\mathbf{AX} = (\mathbf{X}_1, \ldots, \mathbf{X}_L)$ *is an irreducible decomposition: either* $\mathbf{X}$ *already is irreducible or there is some invertible* $\mathbf{A}$ *such that* $\mathbf{AX} = (\mathbf{X}_1, \mathbf{X}_2)$ *with independent* $\mathbf{X}_1$, $\mathbf{X}_2$. *If these two are irreducible, we are finished, otherwise we proceed to decompose whichever of the two still is reducible. After a finite number* $L < \dim(\mathbf{X})$ *of steps, we are left with irreducible components.*

Having established the existence of such a decomposition, we define a normalized version of it:

Definition 2.2. *Assume an* n*-dimensional random vector* $\mathbf{X}$ *and an invertible* $n \times n$ *matrix* $\mathbf{A}$ *such that* $\mathbf{S} = \mathbf{AX}$ *can be subdivided into* $\mathbf{S} = (\mathbf{S}_1, \ldots, \mathbf{S}_M)$ *where*

(i) $\dim(\mathbf{S}_i) \leq \dim(\mathbf{S}_j)$ *for any* $1 \leq i < j \leq M$,

(ii) the random vectors $\mathbf{S}_k$ *are mutually independent,*

(iii) at most one of the $\mathbf{S}_k$ *is Gaussian,*

(iv) all non-Gaussian $\mathbf{S}_k$ *are irreducible,*

then, with $\mathbf{m} := (\dim(\mathbf{S}_1), \ldots, \dim(\mathbf{S}_m))$, *an ordered partition of* $\dim(\mathbf{S})$, *the pair* $(\mathbf{A}, \mathbf{m})$ *is called an Irreducible Subspace Analysis (ISA) of* $\mathbf{X}$ *and the random vectors* $\mathbf{S}_i$ *are called the irreducible components of* $(\mathbf{A}, \mathbf{m})$.

Note that we have gathered all independent one-dimensional Gaussians into a single, higher-dimensional Gaussian component, an idea that was introduced in [3, 4]. Here, observe that any two-dimensional rotation maps two independent Gaussians again onto two independent Gaussians, a fact that also holds for higher dimensions. Therefore it is always possible for two ISAs of a random vector $\mathbf{X}$ to differ in the matrix component by a rotation in the higher-dimensional Gaussian, so the irreducible components of the Gaussian are also unique only up to this indeterminacy.

We note that according merely to the definition, a given $\mathbf{X}$ may have several ISAs, differing in either the basis $\mathbf{A}$, the sizes $\mathbf{m}$ or both.

2.3. The characteristic function and its properties

In the following we will work extensively with the characteristic function of a random vector, so we shortly review its definition and some elementary properties.

Definition 2.3. *Let* $\mathbf{X}$ *be an* n*-dimensional random vector. Then the* characteristic function *of* $\mathbf{X}$ *is defined as* $\widehat{\mathbf{X}}(\mathbf{x}) := E\{\exp(i\mathbf{X}^\top \mathbf{x})\}$ *where* $\mathbf{x} \in \mathbb{R}^n$.

The characteristic function has similar properties to the density when it comes to statistic independence: Assume $\mathbf{X} = (\mathbf{X}_1, \mathbf{X}_2)$ where $\mathbf{X}_1$ and $\mathbf{X}_2$ are

independent. Then their joint characteristic function is equal to the product of the marginal characteristic functions [12]:

$$\widehat{\mathbf{X}}(\mathbf{x}) = \widehat{\mathbf{X}_1}(\mathbf{x}_1)\widehat{\mathbf{X}_2}(\mathbf{x}_2) \ .$$

A local logarithm (note that $\widehat{\mathbf{X}}$ is complex valued) of $\widehat{\mathbf{X}}$ – this logarithm is also known as the second characteristic function [31] – then splits into the sum of the local logarithms of the marginal characteristic functions. Characteristic functions always exist, whereas not all random variables admit a density. If $\mathbf{A}$ is an invertible matrix and $\mathbf{S}$ is a random vector, then

$$\widehat{\mathbf{AS}}(\mathbf{x}) = E\big(\exp(i\mathbf{S}^\top\mathbf{A}^\top\mathbf{x})\big) = \widehat{\mathbf{S}}(\mathbf{A}^\top\mathbf{x}) \ .$$

The characteristic function of a random vector has a simple connection to its moments (given that they exist). If $\mathbf{X}$ is a random vector then its $(n_1, \ldots, n_k)$-th moment can be calculated as follows (see e.g. [12]):

$$E\{X_1^{n_1} \cdots X_k^{n_k}\} = i^{-m} \frac{\partial^m}{\partial x_1^{n_1} \cdots \partial x_k^{n_k}} \widehat{\mathbf{X}}(\mathbf{x})|_0 \tag{1}$$

with $m = n_1 + \cdots + n_k$.

Among all random variables, the ones with the simplest representation are Gaussians. For example, a one dimensional random variable X with known density $p_X > 0$ is Gaussian if and only if $\ln p_X$ is a polynomial of degree at most 2, i.e. if $(\ln p_X)'' = 0$. This property can be used to extract Gaussian components from larger random vectors [30].

2.4. Complex differentiation and linear transformations

In the main proof, we will iteratively extract components. For this we make use of the following three lemmata, the proofs of which are omitted as they are straight-forward.

Lemma 2.1 (differential of a product)**.** *Assume twice differentiable functions $f_k : \mathbb{R}^n \to \mathbb{C}$ $(k = 1, \ldots, M)$ and (not necessarily different) differential operators ∂_i and ∂_j. Then*

$$f\partial_i\partial_j f - (\partial_i f)(\partial_j f) = \sum_{k=1}^{M} \big(\prod_{l\neq k} f_l^2\big) \Big[f_k\partial_i\partial_j f_k - (\partial_i f_k)(\partial_j f_k)\Big]$$

where $f : \mathbb{R}^n \to \mathbb{C}$ is defined by $f(\mathbf{x}) := \prod_{k=1}^{M} f_k(\mathbf{x})$.

Lemma 2.2 (directional derivative and Hessian)**.** *Let $g(\mathbf{x}) := f(\mathbf{A}(\mathbf{x}))$ where $\mathbf{A}$ is an $(m \times n)$-matrix over $\mathbb{R}$ and f is a twice differentiable function on $\mathbb{R}^m$. Then for every $\mathbf{x} \in \mathbb{R}^n$*

(i) $(\partial_i g)(\mathbf{x}) = \mathbf{d}f|_{\mathbf{Ax}}\mathbf{a}_i$

(ii) $(\partial_i \partial_j g)(\mathbf{x}) = (\mathbf{a}_i)^\top \mathbf{H}_f|_{\mathbf{A}\mathbf{x}} \mathbf{a}_j$

where $\mathbf{a}_i$ *denotes the* i*-th column of* $\mathbf{A}$*.*

Lemma 2.3 (logarithmic Hessian)**.** *Assume* $U \subset \mathbb{R}^n$ *and* $f : U \to \mathbb{C}$ *to be a twice continuously differentiable function with some* $\mathbf{x} \in U$ *such that* $f(\mathbf{x}) \neq 0$*. Then, in some neighborhood of* $\mathbf{x}$*,*

$$\mathbf{H}_f = f\mathbf{H}_g + f(\mathbf{d}g)^\top(\mathbf{d}g)$$

where $g := \log f$ *is a local complex logarithm.*

3. Uniqueness of ISA

We will now show Theorem 1.1 in a number of steps. Without loss of generality we assume $\mathbf{X}$ to be centered (zero-mean). As the existence of an ISA of $\mathbf{X}$ holds, we may additionally assume to already have one:

Assumption A1. *Let* $\mathbf{X}$ *be an* n*-dimensional random vector with mean* 0 *and finite covariance and* $(\mathbf{A}, \mathbf{m})$ *be an ISA of* $\mathbf{X}$ *with irreducible components* $\mathbf{S}_1, \ldots, \mathbf{S}_M$*.*

The claim of Theorem 1.1 is equivalent to the claim that for any other ISA $(\mathbf{A}', \mathbf{m}')$ of $\mathbf{X}$, actually $\mathbf{m}' = \mathbf{m}$ and $(\mathbf{A}'\mathbf{A}^{-1})$ can be written as the product of an invertible block-diagonal matrix with blocks of size $m_1, \ldots, m_M$ and a block-permutation matrix, swapping at most blocks of the same size corresponding to non-Gaussians. In terms of the irreducible components of $(\mathbf{A}, \mathbf{m})$ and $(\mathbf{A}', \mathbf{m}')$, this is equivalent to them being mapped to each other by such a product.

3.1. The Gaussian Subspace

Recently the idea of Non-Gaussian Component Analysis (NGCA), or Non-Gaussian Subspace Analysis (NGSA), was proposed, where one separates a higher dimensional distribution into two independent parts, one of them being a high dimensional Gaussian (the *Gaussian subspace*) and the rest (the *Non-Gaussian subspace*) [3, 4]. If the Gaussian subspace is maximal (i.e. there is no way to split off an independent Gaussian from the Non-Gaussian subspace), this decomposition is unique up to transformations within each of the two subspaces.

In order to simplify notation, we from now on will use this convention:

Definition 3.1. *Two random vectors* $\mathbf{X}$ *and* $\mathbf{S}$ *are called* equivalent *(in symbols:* $\mathbf{X} \sim \mathbf{S}$*) if there is some invertible* $\mathbf{A}$ *such that* $\mathbf{X} = \mathbf{A}\mathbf{S}$*.*

The following theorem establishes uniqueness of the decomposition.

Theorem 3.1 (Uniqueness of NGSA)**.** *Assume* $\mathbf{X}$*, a random vector with existing covariance and two arbitrary decompositions* $\mathbf{X} = \mathbf{A}(\mathbf{X}_N, \mathbf{X}_G) = \mathbf{B}(\mathbf{S}_N, \mathbf{S}_G)$ *such that*

(i) $\mathbf{X}_G$ *and* $\mathbf{S}_G$ *are higher-dimensional Gaussians,*

(ii) $\mathbf{X}_N$ *and* $\mathbf{X}_G$ *are independent and so are* $\mathbf{S}_N$ *and* $\mathbf{S}_G$*,*

(iii) the decompositions are maximally reduced, in the sense that there is no projection $\mathbf{M}$ *such that the first component of* $\mathbf{M}\mathbf{X}_N$ *is a Gaussian independent of the rest, and similarly for* $\mathbf{S}_N$*.*

Then $\mathbf{X}_N \sim \mathbf{S}_N$ *and* $\mathbf{X}_G \sim \mathbf{S}_G$*.*

For a proof of this theorem, we refer to [30]. We note that as a special case this also includes deterministic components, which can be seen as Gaussians with variance 0.

This theorem shows that both the maximally Gaussian subspace and the rest are essentially unique. It therefore suffices to show uniqueness of ISA for the non-Gaussian component of $\mathbf{X}$, that is, the part of $\mathbf{X}$ that contains no independent Gaussian, and we may restrict our analysis to random vectors of this kind. It will furthermore be of use in the following to assume $\mathbf{X} = (\mathbf{X}_1, \mathbf{X}_2)$ with independent $\mathbf{X}_1, \mathbf{X}_2$.

Assumption A2. *Let* $\mathbf{X} = (\mathbf{X}_1, \mathbf{X}_2)$ *such that* $\mathbf{X}_1$ *and* $\mathbf{X}_2$ *are independent and such that for no* $\mathbf{B} \in \mathrm{Gl}(m)$ *the projection* $(\mathbf{B}\mathbf{X})_1$ *is normal and independent of the rest of* $(\mathbf{B}\mathbf{X})$*.*

If both (A1) and (A2) hold, then obviously none of the components $\mathbf{S}_i$ are normally distributed.

3.2. Whitening

We will now show that we may assume $\mathbf{X}$ and $\mathbf{S}$ to be decorrelated, which implies $\mathbf{A}$ being orthonormal. The proofs in this section hold in a more general setting than covered by our current assumptions (A1) and (A2).

Lemma 3.1. *Assume an* n*-dimensional random vector* $\mathbf{X}$*. Then there is an invertible* $(n \times n)$ *matrix* $\mathbf{T}$ *such that* $\mathbf{T}\mathbf{X}$ *is decorrelated.*

Proof. As $\mathrm{Cov}(\mathbf{X})$ is symmetric, there is some $\mathbf{T}$ such that $\mathbf{T}\,\mathrm{Cov}(\mathbf{X})\mathbf{T}^\top$ is diagonal. Then

$$\begin{aligned}\mathrm{Cov}(\mathbf{T}\mathbf{X}) &= E\{\mathbf{T}\mathbf{X}(\mathbf{T}\mathbf{X})^\top\} - E\{\mathbf{T}\mathbf{X}\}E\{\mathbf{T}\mathbf{X}\}^\top = \\ &= \mathbf{T}E\{\mathbf{X}\mathbf{X}^\top\}\mathbf{T}^\top - \mathbf{T}E\{\mathbf{X}\}E\{\mathbf{X}\}^\top\mathbf{T}^\top = \mathbf{T}\,\mathrm{Cov}(\mathbf{X})\mathbf{T}^\top\end{aligned}$$

is diagonal, so $\mathbf{T}\mathbf{X}$ is decorrelated. □

We may furthermore rescale the non-deterministic components of $\mathbf{T}\mathbf{X}$ and find a transformation such that every component of $\mathbf{T}\mathbf{X}$ has variance 1 or 0, that is, $\mathrm{Cov}(\mathbf{T}\mathbf{X})$ contains ones somewhere on the diagonal and is zero everywhere else (due to independence the covariance is zero outside of the block-diagonal).

Let us now bring this to use:

Lemma 3.2. *Assume* $\mathbf{S} = (\mathbf{S}_1, \ldots, \mathbf{S}_M)$ *with existing covariance and independent random vectors* $\mathbf{S}_k$*, such that for no invertible* $\mathbf{B}$*, the projection* $(\mathbf{BS})_1$ *is deterministic. Assume furthermore an invertible* $\mathbf{A}$ *and* $\mathbf{X} := \mathbf{AS}$ *where* $\mathbf{X} = (\mathbf{X}_1, \ldots, \mathbf{X}_L)$ *with independent* $\mathbf{X}_j$*. Then there are* $\mathbf{S}'_k \sim \mathbf{S}_k$ *(*$k = 1, \ldots, M$*),* $\mathbf{X}'_j \sim \mathbf{X}_j$ *(*$j = 1, \ldots, L$*), and an orthonormal* $\mathbf{A}'$ *such that every* $\mathbf{S}'_k$ *and* $\mathbf{X}'_j$ *is decorrelated and* $(\mathbf{X}'_1, \ldots, \mathbf{X}'_L) = \mathbf{A}'(\mathbf{S}'_1, \ldots, \mathbf{S}'_M)$*.*

Proof. As $\mathbf{S}$ has existing covariance, so does $\mathbf{X} = \mathbf{AS}$. Using (3.1), we may first decorrelate every component of $\mathbf{X}$ and $\mathbf{S}$, modifying $\mathbf{A}$ accordingly. We here scale our decorrelation matrices such that after decorrelation we have unit covariance everywhere (this can be done due to the assumption of non-deterministic projections). As this operation was performed fully within the single components, for the decorrelated $\mathbf{S}'_k$ and $\mathbf{X}'_j$, we have $\mathbf{S}'_k \sim \mathbf{S}_k$ and $\mathbf{X}'_j \sim \mathbf{X}_j$. Then, setting $\mathbf{S}' := (\mathbf{S}'_1, \ldots, \mathbf{S}'_M)$, $\mathbf{X}' := (\mathbf{X}'_1, \ldots, \mathbf{X}'_L)$, and letting $\mathbf{A}'$ be the modified $\mathbf{A}$, we have

$$\mathbf{I} = \mathrm{Cov}(\mathbf{X}') = \mathrm{Cov}(\mathbf{A}'\mathbf{S}') = \mathbf{A}' \,\mathrm{Cov}(\mathbf{S}')\mathbf{A}'^{\top} = \mathbf{A}'\mathbf{I}\mathbf{A}'^{\top}$$

so $\mathbf{A}'$ is orthonormal. □

Given our assumptions (A1) and (A2), decorrelating the independent components $\mathbf{X}_j$ and $\mathbf{S}_k$ this way does not lose any generality: the independence of the components still holds, and performing decorrelation in this manner we may therefore now assume the following:

Assumption A3. *Assume* $\mathrm{Cov}(\mathbf{X}) = \mathbf{I} = \mathrm{Cov}(\mathbf{S})$*, and hence* $\mathbf{A}$ *to be orthonormal.*

3.3. Uniqueness of the non-Gaussian decomposition

Theorem 3.2. *Assume (A1 - A3). Then there is a permutation* π *of* $\{1, \ldots, M\}$ *and some index* $1 \leq k < M$ *such that* $\mathbf{X}_1 \sim (\mathbf{S}_{\pi(1)}, \ldots, \mathbf{S}_{\pi(k)})$ *and* $\mathbf{X}_2 \sim (\mathbf{S}_{\pi(k+1)}, \ldots, \mathbf{S}_{\pi(M)})$*.*

Uniqueness of ISA can easily be established using this theorem, and most of the rest of this section will be devoted to its proof. Before proving it, we will show how it implies uniqueness of ISA.

Assumption A4. *Assume* $\mathbf{X} = (\mathbf{X}_1, \ldots, \mathbf{X}_L)$ *with irreducible, independent* $\mathbf{X}_j$*.*

Again, similarly to above, assuming (A1) and (A4), decorrelation of all $\mathbf{X}_j$ and $\mathbf{S}_k$ may be performed without any loss of generality, so (A3) does not conflict with these in any way, and given any combination of (A1), (A2) and (A4) we may always additionally assume (A3) (decorrelation of the independent components and orthonormality of the mixing matrix).

Theorem 3.3. *Assume (A1 - A4). Then* $L = M$ *and for every* $1 \leq k \leq M$ *there is some* $1 \leq j \leq L$ *such that* $\mathbf{S}_k \sim \mathbf{X}_j$*.*

Proof. We choose the component of minimal size among $\mathbf{S}_k$ and $\mathbf{X}_j$. Without loss of generality, we may assume this to be $\mathbf{X}_1$. Let us gather the other components of $\mathbf{X}$ to a single random vector:

$$\mathbf{X}' := (\mathbf{X}_2, \ldots, \mathbf{X}_L)$$

Then $(\mathbf{X}_1, \mathbf{X}')$ is an independent (but not irreducible if $L > 2$) decomposition and according to Theorem 3.2, $\mathbf{X}_1$ is equivalent to a combination of some $\mathbf{S}_k$. As $\mathbf{X}_1$ is the smallest component in the $\mathbf{S}_k$ and $\mathbf{X}_j$, it has to be equivalent to a single $\mathbf{S}_k$ for some k, and $\mathbf{X}'$ is equivalent to the concatenation of the other components of $\mathbf{S}$:

$$\mathbf{X}' \sim (\mathbf{S}_1, \ldots, \mathbf{S}_{k-1}, \mathbf{S}_{k+1}, \ldots, \mathbf{S}_M)$$

We remove $\mathbf{S}_k$ and $\mathbf{X}_1$ and proceed iteratively to get $M = L$, and for every k some j such that $\mathbf{S}_k \sim \mathbf{X}_j$. □

Let us now prove Theorem 3.2. We split up $\mathbf{A}$ into submatrices $\mathbf{A}_{jk}$ of size $\dim(\mathbf{X}_j) \times \dim(\mathbf{S}_k)$, so

$$\mathbf{X}_j = \sum_{k=1}^{M} \mathbf{A}_{jk} \mathbf{S}_k \tag{2}$$

and, equivalently,

$$\mathbf{S}_k = \sum_{j=1}^{2} \mathbf{A}_{jk}^\top \mathbf{X}_j \tag{3}$$

since $\mathbf{A}$ is orthonormal. The claim of Theorem 3.2 is equivalent to the claim that in every pair of matrices $\{\mathbf{A}_{1k}, \mathbf{A}_{2k}\}$ one of the two is zero. It suffices to show this for $k = 1$ as the proofs for the other cases are fully analogous. As $\mathbf{A}$ has full rank, $\operatorname{rank}(\mathbf{A}_{11}) + \operatorname{rank}(\mathbf{A}_{21}) \geq \dim(\mathbf{S}_1)$. If we have equality in this relation, we can easily show that both $\mathbf{A}_{11}$ and $\mathbf{A}_{2k}$ being non-zero implies that $\mathbf{S}_1$ is reducible:

Lemma 3.3. *Assume a random vector $\mathbf{S}_1$ and two non-zero matrices $\mathbf{A}_1, \mathbf{A}_2$ such that $\operatorname{rank}(\mathbf{A}_1) + \operatorname{rank}(\mathbf{A}_2) = \dim(\mathbf{S}_1) = \operatorname{rank}(\mathbf{A}_1^\top \ \mathbf{A}_2^\top)$. If we can write*

$$\mathbf{S}_1 = (\mathbf{A}_1^\top \ \mathbf{A}_2^\top) \begin{pmatrix} \mathbf{X}_1 \\ \mathbf{X}_2 \end{pmatrix}$$

with independent random vectors $\mathbf{X}_1$ and $\mathbf{X}_2$, then $\mathbf{S}_1$ is reducible.

Proof. Let $D := \dim(\mathbf{S}_1)$ and $d := \dim\big(\ker(\mathbf{A}_1)\big)$. Using the rank-nullity theorem twice, we have

$$\begin{aligned} \dim\big(\ker(\mathbf{A}_2)\big) &= \dim(\mathbf{S}_1) - \operatorname{rank}(\mathbf{A}_2) = \operatorname{rank}(\mathbf{A}_1) = \\ &= \dim(\mathbf{S}_1) - \dim(\ker(\mathbf{A}_1)) = D - d \end{aligned}$$

and we can then find a linearly independent set $\{\mathbf{v}_1, \ldots, \mathbf{v}_{D-d}\}$ such that $\mathbf{A}_2 \mathbf{v}_j = 0$ for any $1 \leq j \leq D - d$. We also can find a linearly independent

set $\{\mathbf{v}_{D-d+1}, \ldots, \mathbf{v}_D\}$ such that $\mathbf{A}_1\mathbf{v}_i = 0$ for any $D-d+1 \leq i \leq D$. These two sets are guaranteed to be linearly independent, as $\operatorname{rank}(\mathbf{A}_1^\top \ \mathbf{A}_2^\top) = \dim(\mathbf{S}_1)$ and as $\mathbf{A}_1$, $\mathbf{A}_2$ were assumed to be non-zero, neither set is empty. Using these vectors, we define

$$\mathbf{B} := \begin{pmatrix} \mathbf{v}_1^\top \\ \vdots \\ \mathbf{v}_D^\top \end{pmatrix} .$$

Then

$$\begin{aligned} \mathbf{B}\mathbf{S}_1 = (\mathbf{B}\mathbf{A}_1^\top \ \mathbf{B}\mathbf{A}_2^\top) \begin{pmatrix} \mathbf{X}_1 \\ \mathbf{X}_2 \end{pmatrix} = \\ = \begin{pmatrix} \mathbf{B}_1 & 0 \\ 0 & \mathbf{B}_2 \end{pmatrix} \begin{pmatrix} \mathbf{X}_1 \\ \mathbf{X}_2 \end{pmatrix} = \begin{pmatrix} \mathbf{B}_1\mathbf{X}_1 \\ \mathbf{B}_2\mathbf{X}_2 \end{pmatrix} \end{aligned}$$

with some full rank matrices $\mathbf{B}_1$ and $\mathbf{B}_2$. It follows that $\mathbf{S}_1$ is reducible, as $\mathbf{X}_1$ and $\mathbf{X}_2$ are independent and $\mathbf{B}$ is invertible. □

The other case contradicts the assumption that $\mathbf{S}$ contains no independent, normally distributed component:

Lemma 3.4. *Assume (A1 - A3). If* $\operatorname{rank}(\mathbf{A}_{11}) + \operatorname{rank}(\mathbf{A}_{21}) > \dim(\mathbf{S}_1)$*, then* $\mathbf{S}_1$ *contains an independent Gaussian.*

Note that this does not necessarily imply reducibility of $\mathbf{S}_1$, as it might simply be just a one-dimensional Gaussian. In order to prove this claim, we need the following technical lemma, the proof of which we have moved to the appendix.

Lemma 3.5. *Assume* $L, M \in \mathbb{N}$*, functions*

$$f_k : \mathbb{R}^{m_k} \to \mathbb{C} \ \ (k = 1, \ldots, L)$$

and

$$g_j : \mathbb{R}^{n_j} \to \mathbb{C} \ \ (j = 1, \ldots, M)$$

and matrices $\mathbf{A}_{jk} : \mathbb{R}^{m_k} \to \mathbb{R}^{n_j}$ *where* $\sum_{k=1}^{L} m_k = \sum_{j=1}^{M} n_j$ *such that*

(i) the functions f_k, g_j *are twice continuously differentiable*

(ii) the matrix $\mathbf{A} := (\mathbf{A}_{ij})_{i,j}$ *is invertible*

(iii) for any $\mathbf{x} = (\mathbf{x}_1, \ldots, \mathbf{x}_L) \in (\mathbb{R}^{m_1} \times \cdots \times \mathbb{R}^{m_L})$

$$\prod_{k=1}^{L} f_k(\mathbf{x}_k) = \prod_{j=1}^{M} g_j(\mathbf{A}_j\mathbf{x}) \tag{4}$$

where $\mathbf{A}_j = (\mathbf{A}_{j1} \ldots \mathbf{A}_{jL})$.

Then, for any two indices $1 \leq i < j \leq L$ *and for any* M*-tuple of points* $(\mathbf{y}_1, \ldots, \mathbf{y}_M) \in \mathbb{R}^{n_1} \times \cdots \times \mathbb{R}^{n_M}$ *where* $g_k(\mathbf{y}_k) \neq 0$ *(*$k = 1, \ldots, M$*), locally*

$$\sum_{k=1}^{M} \mathbf{A}_{ki}^{\top} (\mathbf{H}_{\log(g_k)}|_{\mathbf{y}_k}) \mathbf{A}_{kj} = 0$$

where log *is a (local) complex logarithm and* $\mathbf{H}_f$ *denotes the Hessian of* f*.*

Using this, we now can prove Lemma 3.4:

Proof. Let us define $D := \dim(\mathbf{S}_1)$, and $d_i := \dim(\ker(\mathbf{A}_{i1}))$ $(i = 1, 2)$. Then $d_1 + d_2 = (D - \operatorname{rank}(\mathbf{A}_{11})) + (D - \operatorname{rank}(\mathbf{A}_{21})) < D$. We then choose vectors $\mathbf{v}_2, \ldots, \mathbf{v}_{d_1+1}$ that form a basis of $\ker(\mathbf{A}_{11})$ and similarly $\mathbf{v}_{D-d_2+1}, \ldots, \mathbf{v}_D$ (note here that $D - d_2 + 1 > d_1 + 1$) that form a basis of $\ker(\mathbf{A}_{21})$. As the matrix $\mathbf{A}$ is invertible, $\ker(\mathbf{A}_{11})$ and $\ker(\mathbf{A}_{21})$ are disjoint, so the set of vectors $\{\mathbf{v}_2, \ldots, \mathbf{v}_{d_1+1}, \mathbf{v}_{D-d_2+1}, \ldots, \mathbf{v}_D\}$ is linearly independent. Now choose vectors $\mathbf{v}_k (k = d_1 + 2, \ldots, D - d_2)$ – this set might be empty – such that the vectors $\mathbf{v}_2, \ldots, \mathbf{v}_D$ are linearly independent, and finally choose a $\mathbf{v}_1$ orthogonal to $\operatorname{span}(\{\mathbf{v}_2, \ldots, \mathbf{v}_D\})$. We define $\mathbf{T}_0 := (\mathbf{v}_1, \ldots, \mathbf{v}_D)$, and then

$$\mathbf{T} := \begin{pmatrix} \mathbf{T}_0 & 0 \\ 0 & \mathbf{I} \end{pmatrix},$$

where $\mathbf{I}$ is the $(\dim(\mathbf{S}) - D)$-dimensional identity. The first column of $\mathbf{T}$ is orthogonal to the other columns, so the first row of $\mathbf{T}^{-1}$ is orthogonal to the other rows of $\mathbf{T}^{-1}$. Now

$$\mathbf{X} = \mathbf{AS} = \mathbf{AT}(\mathbf{T}^{-1}\mathbf{S})$$

where $\mathbf{T}^{-1}$ is an operation purely within $\mathbf{S}_1$. Therefore we may replace $\mathbf{A}$ with $\mathbf{AT}$ and $\mathbf{S}$ with $\mathbf{T}^{-1}\mathbf{S}$. Note that due to our choice of $\mathbf{T}$ we do not have full decorrelation of $\mathbf{S}_1$ anymore; only $(\mathbf{S}_1)_1$ is decorrelated from the other components of $\mathbf{S}_1$ due to the first row of the transformation $\mathbf{T}^{-1}$ being orthogonal to its other rows. Now, the columns of $\mathbf{A}_{11}$ with indices $2, \ldots, d_1 + 1$ contain only zeros, and so do the columns of $\mathbf{A}_{21}$ with indices $D - d_2 + 1, \ldots, D$. The other columns of $\mathbf{A}_{11}$ have full rank, and so do the the other columns of $\mathbf{A}_{21}$.

Let us now turn to the characteristic functions of $\mathbf{X}$ and $\mathbf{S}$. Due to their independent decompositions, we have

$$\prod_{k=1}^{2} \widehat{\mathbf{X}_k}(\mathbf{x}_k) = \widehat{\mathbf{X}}(\mathbf{x}) = \widehat{\mathbf{AS}}(\mathbf{x}) = \widehat{\mathbf{S}}(\mathbf{A}^{\top}\mathbf{x}) = \prod_{j=1}^{M} \widehat{\mathbf{S}_j}(\mathbf{A}_j^{\top}\mathbf{x}) \ .$$

For now we fix an $\mathbf{x} = (\mathbf{x}_1, \ldots, \mathbf{x}_M)$ where $\widehat{\mathbf{S}_k}(\mathbf{A}_k^{\top}\mathbf{x}_k) \neq 0$ (this exists as $\widehat{\mathbf{S}}(0) = 1$). As $\mathbf{S}$ and $\mathbf{X}$ have existing covariance, their characteristic functions are twice continuously differentiable, so all assumptions of Lemma 3.5 are fulfilled, therefore locally

$$0 = \sum_{k=1}^{M} \mathbf{A}_{1k} \mathbf{H}_{\mathbf{h}_k}|_{\mathbf{s}_k} \mathbf{A}_{2k}^{\top}$$

where $\mathbf{s}_k := \mathbf{A}_k\mathbf{x}$ and $\mathbf{h}_k := \log(\widehat{\mathbf{S}_k})$ is a local logarithm. In this sum, every summand depends on a different variable $\mathbf{s}_k$, therefore every summand is constant: For every k

$$\mathbf{A}_{1k}\mathbf{H}_{\mathbf{h}_k}|_{\mathbf{s}_k}\mathbf{A}_{2k}^\top = \mathbf{C}_k \tag{5}$$

with some constant matrices $\mathbf{C}_k$, and in particular the first summand is constant:

$$\mathbf{A}_{11}\mathbf{H}_{\mathbf{h}_1}|_{\mathbf{s}_1}\mathbf{A}_{21}^\top = \mathbf{C}\ . \tag{6}$$

Now, there are invertible matrices $\mathbf{B}_1$ and $\mathbf{B}_2$ such that the first $D-d_2$ rows of $\mathbf{B}_2\mathbf{A}_{21}$ consist of the first $D-d_2$ unit vectors, and such that the k-th row of $\mathbf{B}_1\mathbf{A}_{11}$ consists of the k-th unit vector, where $k = 1, d_1+2, \ldots, D$:

$$\begin{aligned}
&\mathbf{e}_j^\top\mathbf{B}_2\mathbf{A}_{21} = \mathbf{e}_j^\top \ \ (j = 1, \ldots, D-d_2);\\
&\mathbf{e}_j^\top\mathbf{B}_2\mathbf{A}_{21} = 0 \ \ (j = D-d_2+1, \ldots, D)
\end{aligned}$$

and

$$\begin{aligned}
&\mathbf{e}_k^\top\mathbf{B}_1\mathbf{A}_{11} = \mathbf{e}_k^\top \ \ (k = 1, d_1+2, \ldots, D);\\
&\mathbf{e}_k^\top\mathbf{B}_1\mathbf{A}_{11} = 0 \ \ (k = 2, \ldots, d_1+1)\ .
\end{aligned}$$

Multiplying equation (6) with $\mathbf{B}_1$ from the left and $\mathbf{B}_2^\top$ from the right gives us

$$\mathbf{B}_1\mathbf{A}_{11}\mathbf{H}_{h_1}|_{\mathbf{s}_1}(\mathbf{B}_2\mathbf{A}_{21})^\top = \mathbf{C} \tag{7}$$

for some matrix $\mathbf{C}$. Multiplication of equation (7) first with $\mathbf{e}_1^\top$ from the left and $\mathbf{e}_j$ $(j = 1, \ldots, D-d_2)$ from the right gives us:

$$\partial_1\partial_j h_1(\mathbf{s}_1) = \mathbf{e}_1^\top\mathbf{H}_{h_1}|_{\mathbf{s}_1}\mathbf{e}_j = c_j$$

with some constants c_j and $j \in \{1, \ldots, D-d_2\}$. Similarly, multiplication of equation (7) with $\mathbf{e}_1$ from the right and $\mathbf{e}_k^\top$ $(k = d_1+2, \ldots, D)$ from the left gives us:

$$\partial_1\partial_k h_1(\mathbf{s}_1) = \partial_k\partial_1 h_1(\mathbf{s}_1) = \mathbf{e}_k^\top\mathbf{H}_{h_1}|_{\mathbf{s}_1}\mathbf{e}_1 = c_k$$

with some constants c_k and $k \in \{d_1+2, \ldots, D\}$. We have assumed $d_1 < D-d_2$, so also $d_1+2 \leq D-d_2+1$, so all in all:

$$\partial_1\partial_k h_1(\mathbf{s}_1) = c_k$$

for all $1 \leq k \leq D$. Integrating the k-th such equation by s_k tells us that

$$\partial_1 h_1(\mathbf{s}_1) = c_k s_k + g(s_1, \ldots, s_{k-1}, s_{k+1}, \ldots, s_D)$$

for some g independent of s_k, and so together $\partial_1 h_1(\mathbf{s}_1) = \sum_{k=1}^D c_k s_k + C$ with some constant C. Integration then shows

$$h_1(\mathbf{s}_1) = \sum_{k=1}^D c_k s_k s_1 + c_0 s_1 + g(s_2, \ldots, s_D)$$

(note that we have redefined c_1 here) with some continuous function g that does not depend on s_1, and we then get

$$\widehat{\mathbf{S}_1}(\mathbf{s}_1) = \exp(\sum_{k=1}^{D} c_k s_1 s_k + c_0 s_1 + g(s_2, \ldots, s_D)) \ .$$

Such a representation exists in a neighborhood of any $\mathbf{s}$ where $\widehat{\mathbf{S}}(\mathbf{s}) \neq 0$. This is an open condition, so the set of all such $\mathbf{s}$ is open again. But due to continuity the representation above also holds in the closure of this set, hence in a clopen set. As $\widehat{\mathbf{S}}(0) = 1 \neq 0$, this set cannot be empty, therefore this has to hold everywhere.

Let us now use what we know on the first two moments (expectation and covariance) to infer some information on the constants c_k. As $\mathbf{S}_1$ is centered and we know the first row (and column) of its covariance matrix, let us calculate the according expressions in terms of $\widehat{\mathbf{S}_1}$ using equation (1) from section 2.3:

$$\frac{\partial}{\partial s_1}\widehat{\mathbf{S}_1} = \widehat{\mathbf{S}_1}(\mathbf{s}_1)(2c_1 s_1 + \sum_{k=2}^{D} c_k s_k + c_0)$$

so plugging in $\mathbf{s} = 0$, we see $c_0 = 0$. Next, for $j \neq 1$, we have

$$\frac{\partial}{\partial s_j}\widehat{\mathbf{S}_1} = \widehat{\mathbf{S}_1}(\mathbf{s}_1)(c_j s_1 + \frac{\partial}{\partial s_j} g(s_2, \ldots, s_D))$$

so, again plugging in $\mathbf{s} = 0$, we get $\frac{\partial}{\partial s_j} g(s_2, \ldots, s_D)|_{\mathbf{s}=0} = 0$ for all $j \neq 1$. Now we use the fact that $E\{\mathbf{S}_{11}\mathbf{S}_{1j}\} = 0$ for all $j \neq 1$:

$$\frac{\partial^2}{\partial s_1 \partial s_j}\widehat{\mathbf{S}_1} = \widehat{\mathbf{S}_1}(\mathbf{s}_1)(c_j s_1 + \frac{\partial}{\partial s_j} g(s_2, \ldots, s_D))(2c_1 s_1 + \sum_{k=2}^{D} c_k s_k) + c_j \widehat{\mathbf{S}_1}(\mathbf{s}_1)$$

so, plugging in $\mathbf{s} = 0$, we get:

$$0 = c_j$$

so $c_j = 0$ for all $j \neq 1$. Altogether we have:

$$\widehat{\mathbf{S}_1}(\mathbf{s}_1) = \exp(c_1 s_1^2 + g(s_2, \ldots, s_D)) = \exp(c_1 s_1^2) \exp(g(s_2, \ldots, s_D)) \ .$$

As $\widehat{\mathbf{S}_1}$ factorizes this way, the first component of $\mathbf{S}_1$ is a Gaussian independent of the rest. For the sake of completeness, similarly to above, $\mathbf{S}_1$ being white implies c_2 to be $-1/2$, but for our proof it suffices to notice that the first component of $\mathbf{S}_1$ is a Gaussian independent of the rest of $\mathbf{S}_1$, which is what we wanted to show. □

Let us summarize the proof of Theorem 3.2:

Proof. The claim is equivalent to the fact that in equation (2), for every k, one of the two $\mathbf{A}_{k1}$ and $\mathbf{A}_{k2}$ is zero, or, equivalently, $\operatorname{rank}(\mathbf{A}_{k1}) = 0$ or $\operatorname{rank}(\mathbf{A}_{k2}) = 0$. We have $\operatorname{rank} \mathbf{A}_{k1}, \operatorname{rank} \mathbf{A}_{k2} \leq \dim(\mathbf{S}_k)$, and as $\mathbf{A}$ is invertible, $\operatorname{rank}(\mathbf{A}_{k1}) + \operatorname{rank}(\mathbf{A}_{k2}) \geq \dim(\mathbf{S}_k)$.

Assume first $\operatorname{rank}(\mathbf{A}_{k1}) + \operatorname{rank}(\mathbf{A}_{k2}) = \dim(\mathbf{S}_k)$. Equation (3) tells us that $\mathbf{S}_k = \mathbf{A}_{k1}^\top \mathbf{X}_1 + \mathbf{A}_{k1}^\top \mathbf{X}_2$ and now we are facing exactly the assumptions of Lemma 3.3 so $\mathbf{S}_k$ is reducible, contradicting the assumption of irreducibility of all $\mathbf{S}_k$.

If $\operatorname{rank}(\mathbf{A}_{k1}) + \operatorname{rank}(\mathbf{A}_{k2}) > \dim(\mathbf{S}_k)$, all assumptions of Lemma 3.4 are fulfilled, so $\mathbf{S}_k$ contains (or is) an independent Gaussian, contradicting the assumption of non-Gaussianity of $\mathbf{X}$. □

Finally, collecting Theorem 3.3 and Theorem 3.1, we conclude with the uniqueness of ISA in general:

Theorem 3.4. *The linear decomposition of a random vector $\mathbf{X}$ with existing covariance into independent, irreducible subspaces is unique up to order of the components, invertible transformations purely within the total Gaussian subspace and invertible transformations within the non-Gaussian subspaces.*

In other words, this means that the set of vector subspaces $\mathbf{X}$ is projected to is unique, but for any vector subspace, we may freely choose a basis.

4. Discussion and Conclusion

The fact that ICA is only applicable to random vectors with a factorization into 1-dimensional components has given rise to the development of generalizations where certain dependencies in the random vectors are allowed. This question was originally raised in [5], where it was argued that ICA algorithms make the data set as independent as possible, after which one simply has to gather the linear components together that turn out still to be dependent after this step, and thus get what then was called Multidimensional Independent Component Analysis (MICA). At this point it was not clear if the nice separability property of ICA holds in this setting, as no formal proof for this claim was available. One of the main contributions of a formal proof for higher dimensional generalizations of ICA was the proof of separability of k-ISA [27], where all subspaces were assumed to have the same size. The proof of this work was based on a multivariate extension of the Darmois-Skitovitch theorem, where the random vectors in question have fixed size k, and the additional assumption is made that the mixing matrix is k-admissible, that is, the aligned $k \times k$ submatrices are either invertible or zero, however this theorem cannot be used in the fully general context where the subspaces are allowed to have arbitrary dimensionality.

Dropping all assumptions on subspace sizes, it is easy to see that any given random vector can be represented by a linear mixture of statistically independent random vectors. The additional assumption of irreducibility of these random vectors restricts them to essentially only one set of such subspaces, so these can be seen as *the* linear factors, or subspaces that together form the original

random vector. In any linear mixing model of random vectors it is therefore now possible to assume that the random vector is given as the unique representation of irreducible subspaces, thus easing further analysis.

While this may be an interesting uniqueness result from the pure point of statistics, one may also see this as a structural result: it is now clear that the density of any random vector $\mathbf{X}$ with existing covariance can be represented uniquely as a product

$$p(\mathbf{x}) = p_1(\mathbf{A}_1\mathbf{x}) \dots p_n(\mathbf{A}_n\mathbf{x}) p_{\text{Gauss}}(\mathbf{A}_G\mathbf{x})$$

up to the ambiguities discussed (permutation and linear factors). This result has a wide range of applications, e.g. in the fields of signal processing, biomedical imaging and analysis of financial data. Independent Component Analysis (ICA) has proven to be a valuable tool here, see e.g. [6, 14, 32]. If one assumes that the data set to be analyzed is generated by linear mixings of some independent, unknown underlying sources, which can be modeled by random variables, the fact that ICA is separable tells us that any demixing of the observations into independent components recovers the real sources (up to permutation and scaling). From a theoretical data analysis point of view, the assumption of independence of the sources is very restrictive: Even if one knows that the observations are linear mixings of the sources, what happens if the sources are not fully independent? Does the whole idea of ICA completely break down then if one has non-trivial dependencies? As we have seen, this is not the case, and we actually have a straight-forward extension of ICA to ISA, replacing the independent components with independent irreducible subspaces of arbitrary dimension. It is worthwhile to point out that for random vectors $\mathbf{S}$ fulfilling the ICA assumption of complete mutual independence, one can show that a representation of $\mathbf{S}$ that is merely pairwise independent already is equivalent to $\mathbf{S}$ itself (that is, there is a one-to-one correspondence between the pairwise independent random vectors and the components of $\mathbf{S}$). When showing uniqueness of what is nowadays more commonly referred to as ISA, we assumed mutual independence of the irreducible subspaces, but the open question whether pairwise independence is sufficient is interesting. But even before the question of uniqueness of (mutual) ISA was answered satisfactorily, it already gave rise to the development of algorithms [15, 18, 21], performing the ISA task. Similar extensions had already been performed for Principal Component Analysis (PCA) – which only employs statistics up to the second order, i.e. correlations – where techniques are known that extract not only a single principal component, but rather a whole principal component subspace [19]. As PCA uses only first and second order moments, it is impossible to uniquely define self-consistent independent subspaces in this context. Therefore methods to extract such subspaces here have to make use of additional information e.g. by ordering the principal components by power (variance) and then extracting the subspace of the strongest k components. On the other hand, ISA makes use of all higher order statistics, so in this context the term "subspace" actually has a unique meaningful denotation and the sizes of the subspaces arise purely from the definition of independence and irreducibility. The conjecture that ISA can be solved by applying standard ICA algorithms

and grouping the recovered random variables that then still show dependencies has been shown for some distributions [24]. Based on this conjecture ISA algorithms performing joint block diagonalization [26, 28] or making use of the fact that algorithmic outputs will have a large variability within the subspaces [32] have been developed, but so far there is no algorithmic approach where full convergence in the general setting has been proven.

When considering practical implementations, it is important to point out another fact. In our theoretical work, we have assumed perfect knowledge of the distribution of the given signal, whereas for practical considerations one can only estimate the distribution of the signal up to some precision determined among other things by the number of available samples. In practice one can therefore not expect to estimate full independence of subspaces, even if from a theoretical point of view these actually are independent, e.g. if they represent different independent biophysical processes in the human body. Even in the ICA case, this is an important question and has found attention only recently [9, 13, 20] although in this setting the original question of course allows a slightly easier handling, as one simply assumes to know the number of subspaces and their dimensionality (in ICA one deals only with 1-dimensional subspaces). For ISA this question has to deserve a lot more attention. Theory answers the question of uniqueness without consideration of the sizes and the number of subspaces, so it would be good to have algorithms that do the same. Furthermore, similar to PCA, for data analysis, the estimation of the dimension of relevant subspaces is non-trivial and we can expect to generalize some of the many existing approaches developed in this easier setting, such as Minimum Description Length (MDL) [22] or Akaike's Information Criterion (AIC) [1] in PCA.

Acknowledgements

The authors acknowledge financial support by the German Ministry for Education and Research (BMBF) via the Bernstein Center for Computational Neuroscience (BCCN) Göttingen under Grant No. 01GQ0430. They thank Florian Blöchl and Claudia Czado for careful proofreading and valuable comments.

Appendix A.

Proof of Lemma 3.5. Let us first fix two indices i and j in different subspaces. More formally, we fix two different subspace indices $1 \leq i' < j' \leq M$ and then take any i and j such that $\sum_{k=1}^{i'-1} m_k < i \leq \sum_{k=1}^{i'} m_k$ and $\sum_{k=1}^{j'-1} m_k < j \leq \sum_{k=1}^{j'} m_k$. Then, if F is the function defined on the left hand side of equation (4),

$$F(\mathbf{x})(\partial_i \partial_j F)(\mathbf{x}) - (\partial_i F)(\mathbf{x})(\partial_j F)(\mathbf{x}) = 0$$

as x_i and x_j appear in different factors of F. So the result of the same operations applied to the right hand side also is 0. Using Lemma 2.1 this tells us that

$$0 = \sum_{k=1}^{M} \big(\prod_{l\neq k} g_k(\mathbf{A}_k\mathbf{x}_k)\big)^2 \Big[g_k(\mathbf{A}_k\mathbf{x})\big(\partial_i\partial_j(g_k\circ\mathbf{A}_k)\big)(\mathbf{x}) - \\ - \big(\partial_i(g_k\circ\mathbf{A}_k)\big)(\mathbf{x})\big(\partial_j(g_k\circ\mathbf{A}_k)\big)(\mathbf{x}) \Big] .$$

So the right hand side of equation (4) is also equal to zero for all i, j fulfilling the inequalities above. Performing this for all such i, j, that is, for all i in the i'-th subspace and for all j in the j'-th subspace, and collecting the expressions into a single matrix and substituting $\mathbf{y}_i := \mathbf{A}_i\mathbf{x}$, we get after using Lemma 2.2:

$$0 = \sum_{k=1}^{M} \big(\prod_{l\neq k} g_k(\mathbf{y}_k)\big)^2 \mathbf{A}_{ki'}^\top \Big[g_k(\mathbf{y}_k)\mathbf{H}_{g_k}|_{\mathbf{y}_k} - \\ - (\mathbf{d}g_k|_{\mathbf{y}_k})^\top (\mathbf{d}g_k|_{\mathbf{y}_k}) \Big] \mathbf{A}_{kj'}$$

for all $\mathbf{y} = (\mathbf{y}_1, \ldots, \mathbf{y}_M)$ that can be written as $\mathbf{y} = \mathbf{A}\mathbf{x}$. Due to the invertibility of $\mathbf{A}$ this is the case everywhere. The functions g_k are twice continuously differentiable, hence continuous. Let us fix $\mathbf{y}_1, \ldots, \mathbf{y}_N$ where for all $k \in \{1, \ldots, M\}$ we have $g_k(\mathbf{y}_k) \neq 0$, which, due to continuity, is then also the case in some environment of the $\mathbf{y}_k$, where we may choose a complex logarithm. Then, with Lemma 2.3, locally

$$0 = \sum_{k=1}^{M} \big(\prod_{l=1}^{M} g_k(\mathbf{y}_k)\big)^2 \mathbf{A}_{ki'}^\top \mathbf{H}_{\log(g_k)}|_{\mathbf{y}_k} \mathbf{A}_{kj'} = \\ = \sum_{k=1}^{M} \mathbf{A}_{ki'}^\top \mathbf{H}_{\log(g_k)}|_{\mathbf{y}_k} \mathbf{A}_{kj'}$$

where the last equality holds as $g_k(\mathbf{y}_k) \neq 0$ for all k. □

[1] H. Akaike. A new look at the statistical model identification. *Automatic Control, IEEE Transactions on*, 19(6):716–723, 1974.

[2] A. J. Bell and T. J. Sejnowski. An information-maximization approach to blind separation and blind deconvolution. *Neural computation*, 7(6):1129–1159, 1995.

[3] G. Blanchard, M. Kawanabe, M. Sugiyama, V. Spokoiny, and K.-R. Müller. In search of non-gaussian components of a high-dimensional distribution. *The Journal of Machine Learning Research*, 7:247–282, Dec 2006.

[4] G. Blanchard, M. Sugiyama, M. Kawanabe, V. Spokoiny, and K.-R. Müller. Non-gaussian component analysis: a semi-parametric framework for linear dimension reduction. *Proc. NIPS*, pages 131–138, Jan 2006.

[5] J.-F. Cardoso. Multidimensional independent component analysis. *Proc. ICASSP*, 4:1941–1944, 1998.

[6] A. Cichocki and S.-I. Amari. *Adaptive blind signal and image processing*. John Wiley & Sons, 2002.

[7] P. Comon. Independent component analysis, a new concept? *Signal Processing*, 36(3):287–314, Jan 1994.

[8] G. Darmois. Analyse générale des liaisons stochastiques: etude particulière de l'analyse factorielle linéaire. *Revue de l'Institut International de Statistique / Review of the International Statistical Institute*, 21(1/2):2–8, Jan 1953.

[9] S. C. Douglas, Z. Yuan, and E. Oja. Average convergence behavior of the fastica algorithm for blind source separation. *Proc. ICA*, pages 790–798, 2006.

[10] J. Eriksson and V. Koivunen. Identifiability and separability of linear ICA models revisited. *Proc. ICA*, pages 23–27, 2003.

[11] H. W. Gutch and F. J. Theis. Independent subspace analysis is unique, given irreducibility. *Proc. ICA*, pages 49–56, 2007.

[12] W. Härdle and L. Simar. *Applied Multivariate Statistical Analysis*. Springer Verlag, 2007.

[13] J. M. Herrmann and F. J. Theis. Statistical analysis of sample-size effects in ICA. *Proc. IDEAL*, pages 416–425, Jan 2007.

[14] A. Hyvärinen, J. Karhunen, and E. Oja. *Independent Component Analysis*. John Wiley & Sons, 2001.

[15] A. Hyvärinen and U. Köster. FastISA: A fast fixed-point algorithm for independent subspace analysis. *Proc. ESANN*, pages 371–376, 2006.

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

[17] A. M. Kagan, Y. V. Linnik, and C. R. Rao. *Characterization Problems in Mathematical Statistics*. Wiley, New York, 1973.

[18] Y. Nishimori, S. Akaho, and M. D. Plumbley. Riemannian optimization method on the flag manifold for independent subspace analysis. *Proc. ICA*, pages 295–302, Jan 2006.

[19] E. Oja. Neural networks, principal components, and subspaces. *International Journal of Neural Systems*, 1(1):61–68, 1989.

[20] E. Ollila, H.-J. Kim, and V. Koivunen. Compact cramér-rao bound expression for independent component analysis. *Signal Processing, IEEE Transactions on*, 56(4):1421–1428, 2008.

[21] B. Póczos and A. Lőrincz. Independent subspace analysis using k-nearest neighborhood distances. *Proc. ICANN*, pages 163–168, Jul 2005.

[22] J. Rissanen. Modeling by shortest data description. *Automatica*, 14(5):465–471, Jan 1978.

[23] V. P. Skitovich. Linear forms of independent random variables and the normal distribution law. *Izv. Akad. Nauk SSSR Ser. Mat.*, 18(2):185–200, Jul 1954.

[24] Z. Szabó, B. Póczos, and A. Lőrincz. Separation theorem for k-independent subspace analysis with sufficient conditions. *arXiv*, math.ST, Jan 2006.

[25] F. J. Theis. A new concept for separability problems in blind source separation. *Neural computation*, 16:1827–1850, Jan 2004.

[26] F. J. Theis. Blind signal separation into groups of dependent signals using joint block diagonalization. *Proc. ISCAS*, pages 5878–5881, Jan 2005.

[27] F. J. Theis. Multidimensional independent component analysis using characteristic functions. *Proc. EUSIPCO*, Jan 2005.

[28] F. J. Theis. Towards a general independent subspace analysis. *Proc. NIPS*, pages 1361–1368, Jan 2006.

[29] F. J. Theis, A. Jung, C. G. Puntonet, and E. W. Lang. Linear geometric ICA: Fundamentals and algorithms. *Neural computation*, 15(2):419–439, 2003.

[30] F. J. Theis and M. Kawanabe. Uniqueness of non-gaussian subspace analysis. *Proc. ICA*, pages 917–925, Jan 2006.

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

[32] J. Ylipaavalniemi and R. Vigário. Analyzing consistency of independent components: an fMRI illustration. *Neuroimage*, 39(1):169–180, 2008.