

## Auto-Regressive Independent Process Analysis without Combinatorial Efforts\*

Zoltán Szabó, Barnabás Póczos\*\*, András Lőrincz\*\*\*

Department of Information Systems, Eötvös Loránd University,  
Pázmány Péter sétány 1/C, Budapest H-1117, Hungary  
WWW home page: <http://nipg.inf.elte.hu>  
e-mail: [szzoli@cs.elte.hu](mailto:szzoli@cs.elte.hu), [poczos@cs.ualberta.ca](mailto:poczos@cs.ualberta.ca), [andras.lorincz@elte.hu](mailto:andras.lorincz@elte.hu)

**Abstract** We treat the problem of searching for hidden multi-dimensional independent auto-regressive processes (Auto-Regressive Independent Process Analysis, AR-IPA). Independent Subspace Analysis (ISA) can be used to solve the AR-IPA task. The so-called separation theorem simplifies the ISA task considerably: the theorem enables one to reduce the task to one-dimensional Blind Source Separation (BSS) task followed by the grouping of the coordinates. However, the grouping of the coordinates still involves 2 types of combinatorial problems: (i) the number of the independent subspaces and their dimensions, and then (ii) the permutation of the estimated coordinates are to be determined. Here, we generalize the separation theorem. We also show a non-combinatorial procedure, which – under certain conditions – can treat these 2 combinatorial problems. Numerical simulations have been conducted. We investigate problems that fulfill sufficient conditions of the theory and also others that do not. The success of the numerical simulations indicates that further generalizations of the separation theorem may be feasible.

**Key words** Independent Component Analysis, Independent Process Analysis, auto-regressive processes

---

\* © Springer-Verlag London Limited 2009. The original publication is available at <http://dx.doi.org/10.1007/s10044-009-0174-x>.

\*\* Present address: Department of Computing Science, University of Alberta, Athabasca Hall, Edmonton, Canada, T6G 2E8

\*\*\* Corresponding author. Voice: (36-1) 209-0555 / 8473, Fax: (36-1) 381 2140.

## **Originality and Contribution**

Combinatorial explosion is a major obstacle for signal processing. Here, we show that for Auto-Regressive Independent Process Analysis (AR-IPA), a relative of Independent Component Analysis, this problem can be alleviated. Namely, our proposed method does not require previous knowledge about the dimensions of the hidden independent components. Independent Subspace Analysis (ISA) can be used to solve the AR-IPA task. We provide a rigorous theorem with sufficient conditions for the ISA task to be solved. Numerical studies outside of the domain of the theorem indicate the robustness of the method.

## 1 Introduction

Search for independent components is in the focus of research interest. There are important applications in this field, such as blind source separation, blind source deconvolution, feature extraction, and denoising. Thus, a variety of particular methods have been developed over the years. For a recent review on these approaches and for further applications, see [1, 2] and the references therein.

Traditionally, Independent Component Analysis (ICA) is one-dimensional in the sense that all sources are assumed to be independent real valued stochastic variables. The typical example of ICA is the so-called *cocktail-party problem*, where there are  $n$  sound sources and  $n$  microphones and the task is to separate the original sources from the observed mixed signals. However, applications where not all, but only certain groups of the sources are independent may have high relevance in practice. In this case, independent sources can be multi-dimensional. Consider the following generalizations of the cocktail-party problem. There could be independent groups of people talking about independent topics at a conference or more than one group of musicians may be playing at a party. This is the Independent Subspace Analysis (ISA) extension of ICA, also called Multi-dimensional Independent Component Analysis (MICA) [3].

Efforts have been made to develop ISA algorithms [3, 4, 5, 6, 7, 8, 9]. Theoretical problems concern mostly the estimation of the entropy or the mutual information. Entropy estimation by Edgeworth expansion [4] has been extended to more than 2 dimensions and has been used for clustering and mutual information testing [10]. Other recent approaches search for independent subspaces via kernel methods [6], joint block diagonalization [9],  $k$ -nearest neighbors [7], and geodesic spanning trees [8].

An important application of ISA is, e.g., the processing of EEG-fMRI data [4]. Clearly, most practical problems, alike to the analysis of EEG-fMRI signals, exhibit considerable temporal correlations. In such cases, one may take advantage of Auto-Regressive Independent Process Analysis (AR-IPA) [11], a generalization of ISA to auto-regressive (AR) processes, similar to the AR generalization of the original ICA problem [12]. A separation theorem [13] allows one, under certain conditions, to reduce the AR-IPA task to ICA and then to the search for the optimal permutation of the ICA elements.<sup>1</sup> Nonetheless, this permutation search is still a combinatorial problem and it is computationally intractable for large dimensional hidden sources. Further, AR-IPA methods [11, 13, 12] assume that the number of the processes and their respective dimensions are known in advance.

Our algorithm makes use of the ISA separation theorem [13], that we generalize here. The algorithm builds upon the temporal dependencies within the subspaces. It does not require previous knowledge about the number and the dimensions of the sub-processes and can considerably ease or may fully avoid the estimation of the multi-dimensional Shannon entropy, the tool of some of the previous methods [11, 13].

---

<sup>1</sup> The possibility of such a decomposition principle was suspected by [3], who based his conjecture on numerical experiments.

The paper is constructed as follows: Section 2 formulates the AR-IPA problem and its reduction to the ISA task. The ISA separation theorem and our novel AR-IPA identification algorithm are presented in Section 3. Numerical simulations are provided in Section 4. Short discussion can be found in Section 5 and conclusions are drawn in Section 6. We also provide an Appendix. It contains the details of the ISA separation theorem, including previous and novel results for the sake of self-containment.

## 2 The AR-IPA Model

We formulate the AR-IPA task (Section 2.1) and then reduce it to the ISA problem (Section 2.2). The uniqueness of the identification is also treated.

### 2.1 The AR-IPA Equations

Assume that we have  $M$  hidden and independent AR processes and that only the mixture of these  $M$  components is available for observation:

$$\mathbf{s}^m(t+1) = \mathbf{F}^m \mathbf{s}^m(t) + \mathbf{e}^m(t), \quad m = 1, \dots, M \quad (1)$$

$$\mathbf{s}(t) = [\mathbf{s}^1(t); \dots; \mathbf{s}^M(t)], \quad (2)$$

$$\mathbf{z}(t) = \mathbf{A} \mathbf{s}(t), \quad (3)$$

where  $\mathbf{s}(t)$  is the vector concatenated form of components  $\mathbf{s}^m(t)$ ,  $\mathbf{s}^m(t)$ ,  $\mathbf{e}^m(t) \in \mathbb{R}^{d^m}$ ,  $\mathbf{e}^m(t)$  is i.i.d. in  $t$ , and there is at most a single Gaussian among sources  $\mathbf{e}^m$ . Also  $I(\mathbf{e}^1, \dots, \mathbf{e}^M) = 0$ , where  $I$  stands for the mutual information of the arguments.  $\mathbf{F}^m (\neq \mathbf{0}) \in \mathbb{R}^{d^m \times d^m}$  is called the *predictive matrix* of the  $m^{\text{th}}$  process. The total dimension of the components is  $D := \sum_{m=1}^M d^m$ . Thus,  $\mathbf{s}(t), \mathbf{z}(t) \in \mathbb{R}^D$ . Matrix  $\mathbf{A} \in \mathbb{R}^{D \times D}$  is the so called *mixing matrix* which, according to our assumptions, is invertible. From now on, for the sake of notational simplicity, all  $d^m$ s are assumed to be equal ( $d^m = d \forall m$ ), but *all results* hold for the general case.

The goal of the AR-IPA problem is to estimate the original source  $\mathbf{s}(t)$  and the unknown mixing matrix  $\mathbf{A}$  (or its inverse  $\mathbf{W}$ , which is called the *separation matrix*) by using observations  $\mathbf{z}(t)$  only. If  $\forall \mathbf{F}^m = \mathbf{0}$  then the task reduces to the ISA task. The ICA task is recovered if both  $\forall \mathbf{F}^m = \mathbf{0}$  and  $d = 1$ .

### 2.2 Reduction of AR-IPA to ISA and Ambiguities

The identification of the AR-IPA model, alike to the identification of ICA and ISA models, is ambiguous. First, we shall reduce the AR-IPA task to the ISA task [12, 14, 11, 13] by means of *innovations*. The innovation of stochastic process  $\mathbf{u}(t)$  is the error of the optimal quadratic estimation of the process using its past, i.e.,

$$\tilde{\mathbf{u}}(t) := \mathbf{u}(t) - E[\mathbf{u}(t) | \mathbf{u}(t-1), \mathbf{u}(t-2), \dots], \quad (4)$$

where  $E[\cdot]$  is the expectation value operator. It is easy to see that for an AR process, the innovation is identical to the noise that drives the process. Therefore, constructing a block-diagonal matrix  $\mathbf{F}$  from matrices  $\mathbf{F}^m$ , the AR-IPA model assumes the following form

$$\mathbf{s}(t+1) = \mathbf{F}\mathbf{s}(t) + \mathbf{e}(t), \quad (5)$$

$$\mathbf{z}(t) = \mathbf{A}\mathbf{s}(t) = \mathbf{F}_z\mathbf{z}(t-1) + \mathbf{A}\mathbf{e}(t-1), \quad (6)$$

$$\tilde{\mathbf{z}}(t) = \mathbf{A}\mathbf{e}(t-1) = \mathbf{A}\tilde{\mathbf{s}}(t), \quad (7)$$

where  $\mathbf{F}_z = \mathbf{A}\mathbf{F}\mathbf{A}^{-1}$ . Equation (6) shows that the observation process  $\mathbf{z}(t)$  is also an AR process. Thus, applying ISA to innovation  $\tilde{\mathbf{z}}(t)$  of the observation, mixing matrix  $\mathbf{A}$  and thus  $\mathbf{e}(t)$  as well as  $\mathbf{s}(t)$  can be determined.

Concerning the ISA task, we can lessen the ambiguity of the problem (see, e.g., [15]) by assuming that both the noise source and the innovation of the observation are *white*, that is,  $E[\mathbf{e}] = \mathbf{0}$ ,  $E[\mathbf{e}\mathbf{e}^T] = \mathbf{I}_D$  and  $E[\tilde{\mathbf{z}}] = \mathbf{0}$ ,  $E[\tilde{\mathbf{z}}\tilde{\mathbf{z}}^T] = \mathbf{I}_D$ , where  $\mathbf{I}_D$  is the  $D$ -dimensional identity matrix, superscript  $T$  denotes transposition. Then, ambiguities are restricted to permutations of the subspaces and to arbitrary orthogonal transformations within the subspaces. Now, we have that mixing matrix  $\mathbf{A}$  and thus matrix  $\mathbf{W} = \mathbf{A}^{-1}$  are orthogonal, because:

$$\mathbf{I}_D = E[\tilde{\mathbf{z}}\tilde{\mathbf{z}}^T] = \mathbf{A}E[\mathbf{e}\mathbf{e}^T]\mathbf{A}^T = \mathbf{A}\mathbf{I}_D\mathbf{A}^T = \mathbf{A}\mathbf{A}^T. \quad (8)$$

### 3 ISA Separation Theorem and Consequences

First, we define the ISA cost function (Section 3.1), then we state the ISA separation theorem (Section 3.2). For the proof of the theorem, see Appendix (A.) This theorem forms the basis of our AR-IPA identification algorithm, which is introduced in Section 3.3. A novel extension of the separation theorem is also provided in Section 3.4.

#### 3.1 The ISA Cost Function

The ISA task can be viewed as the minimization of mutual information between the estimated components:

$$\min_{\mathbf{W} \in \mathcal{O}^D} I(\mathbf{y}^1, \dots, \mathbf{y}^M), \quad (9)$$

where  $\mathbf{y} = \mathbf{W}\mathbf{z}$ ,  $\mathbf{y} = [\mathbf{y}^1; \dots; \mathbf{y}^M]$  and  $\mathcal{O}^D$  denotes the space of the  $D \times D$  orthogonal matrices. Cost function  $I$  can be transcribed to another cost function that sums up Shannon's multi-dimensional differential entropy terms (see, e.g., [8] for details):

$$\min_{\mathbf{W} \in \mathcal{O}^D} J(\mathbf{W}) := \sum_{m=1}^M H(\mathbf{y}^m). \quad (10)$$

### 3.2 The ISA Separation Theorem

ICA and ISA problems can be formulated as the optimization of the above cost function  $J$ . Further, the following ISA separation theorem holds [13]:

**Theorem 1 (Separation Theorem for ISA)** *Let us suppose, that all the  $\mathbf{u} = [u_1; \dots; u_d] = \mathbf{s}^m$  components of source  $\mathbf{s}$  in the ISA task satisfy*

$$H\left(\sum_{i=1}^d w_i u_i\right) \geq \sum_{i=1}^d w_i^2 H(u_i), \forall \mathbf{w} : \sum_{i=1}^d w_i^2 = 1, \quad (11)$$

*and that the ICA cost function  $J(\mathbf{W}) = \sum_{i=1}^D H(y_i)$  has a minimum ( $\mathbf{W} \in \mathcal{O}^D$ ). Then it suffices to search for the minimum of the ISA task ( $\mathbf{W}_{\text{ISA}}$ ) as a permutation of a solution of the ICA task ( $\mathbf{W}_{\text{ICA}}$ ). In other words, it is enough to look for the ISA separation matrix in the following form*

$$\mathbf{W}_{\text{ISA}} = \mathbf{P}\mathbf{W}_{\text{ICA}}, \quad (12)$$

*where  $\mathbf{P} (\in \mathbb{R}^{D \times D})$  is a permutation matrix to be determined. (Proof is provided in Appendix A.)*

*Note 1* We do not suppose in the separation theorem that matrix  $\mathbf{W}_{\text{ICA}}$  is unique – apart from the ambiguities of the ICA solution. According to our experiences this assumption is not necessary (see Section 4).

Thus, if noises  $\mathbf{e}^m$  of processes  $\mathbf{s}^m$  satisfy condition (11), then the AR-IPA model can be estimated as follows:

1. observe  $\mathbf{z}(t)$  and estimate the AR model,
2. whiten the innovation of the AR process and perform ICA on it,
3. solve the combinatorial problem: search for the permutation of the ICA components that minimizes the cost function  $J$ .

This is a viable approach. The cross-entropy (CE) method [16] seems very efficient in the last step [13]. However, the CE method (i) still assumes that the number of the sub-processes and their respective dimensions are known, (ii) it requires costly multi-dimensional entropy estimations, and (iii) the optimization concerns  $\frac{D!}{d^1! \dots d^M!}$  permutations even if the dimensions of the subspaces are given. As we show below, under certain conditions, the separation theorem offers a solution to the AR-IPA task; previous knowledge of the dimensions of the subspaces is not necessary, global search in permutation space as well as the computer time consuming estimation of multi-dimensional entropies can be circumvented.

### 3.3 Non-combinatorial AR-IPA Algorithm Using the Separation Theorem

We make the following observations:

- Eq. (6) prescribes how to transform the predictive matrix of an AR process [ $\mathbf{z}(t) = \mathbf{A}\mathbf{s}(t)$ ] to another basis. Thus, the predictive matrix of the hidden process  $\mathbf{s}(t) = \mathbf{W}\mathbf{z}(t)$  ( $\mathbf{W} = \mathbf{A}^{-1}$ ) is

$$\mathbf{F} = \mathbf{W}\mathbf{F}_z\mathbf{W}^{-1} = \mathbf{W}\mathbf{F}_z\mathbf{W}^T, \quad (13)$$

where the orthogonality of matrix  $\mathbf{W}$  was taken into account. Thus, making use of the separation theorem, matrix

$$\mathbf{W}_{\text{ICA}}\mathbf{F}_z\mathbf{W}_{\text{ICA}}^T \quad (14)$$

– apart from (possible) permutations – is equal to the block-diagonal predictive matrix  $\mathbf{F}$  of the source  $\mathbf{s}$ .

- It then follows that connected groups of the coordinates of the hidden source can be recovered by collecting the elements belonging to the same block in  $\mathbf{F}$ . In practice, the estimation of matrix  $\mathbf{F}$  (i.e., matrix  $\hat{\mathbf{F}}$ ), is only nearly block-diagonal (apart from permutation). Thus, we say that two coordinates  $i$  and  $j$  are  $\hat{\mathbf{F}}$ -‘connected’ if  $\max(|\hat{F}_{ij}|, |\hat{F}_{ji}|) > \epsilon$ , where,  $\hat{F}_{ij}$  denotes the  $(i, j)^{th}$  coordinate of matrix  $\hat{\mathbf{F}}$ , and, in the ideal case,  $\epsilon = 0$ . Then we can group the  $\hat{\mathbf{F}}$ -‘connected’ coordinates into separate subspaces using the following algorithm:
  - Choose an arbitrary coordinate  $i$  ( $1 \leq i \leq D$ ) and group all  $j \neq i$  coordinates to it, which are  $\hat{\mathbf{F}}$ -‘connected’ with it.
  - Choose an arbitrary and not yet grouped coordinate. Find its connected coordinates recursively. Group them together.
  - Continue until all components are grouped.

This *gathering procedure* is fast. In the worst case, it is quadratic in  $D$ .

Summing up, the pseudo-code of our non-combinatorial AR-IPA algorithm is provided below. Estimations of  $\mathbf{F}_z$ ,  $\mathbf{W}_{\text{ICA}}$ ,  $\mathbf{s}(t)$ ,  $\mathbf{F}$  are denoted by  $\hat{\mathbf{F}}_z$ ,  $\hat{\mathbf{W}}_{\text{ICA}}$ ,  $\hat{\mathbf{s}}(t)$ ,  $\hat{\mathbf{F}}$ , respectively and the estimation of the innovation of process  $\mathbf{z}(t)$  is denoted by  $\tilde{\mathbf{z}}(t)$ , where the hat sign is neglected for the sake of notational simplicity:

- AR fit to  $\mathbf{z}(t)$ : innovation  $\tilde{\mathbf{z}}(t)$  is approximated by making use of the estimation  $\hat{\mathbf{F}}_z$ .
- ICA on whitened  $\tilde{\mathbf{z}}(t)$ : hidden source  $\hat{\mathbf{s}}$  is estimated by means of the estimated separation matrix  $\hat{\mathbf{W}}_{\text{ICA}}$ :

$$\hat{\mathbf{s}}(t) := \hat{\mathbf{W}}_{\text{ICA}}\tilde{\mathbf{z}}(t). \quad (15)$$

- Estimation of the predictive matrix of source  $\mathbf{s}(t)$  [see Eq. (14)]:

$$\hat{\mathbf{F}} = \hat{\mathbf{W}}_{\text{ICA}}\hat{\mathbf{F}}_z\hat{\mathbf{W}}_{\text{ICA}}^T. \quad (16)$$

- Grouping of  $\hat{\mathbf{F}}$ -‘connected’ processes  $\hat{s}_i$  ( $\hat{s}_i$  is the  $i^{th}$  coordinate of  $\hat{\mathbf{s}}$ ).

For our method, computational time is negligible as compared to the CE based AR-IPA technique of [13]. Namely, in the present technique one needs at most  $O(D^2)$  simple ‘ $\max(|\hat{F}_{ij}|, |\hat{F}_{ji}|) > \epsilon$ ’-like comparisons to execute the gathering procedure. By contrast, the computational time of the CE-based technique using

k-nearest neighbor method for entropy estimation [13] scales with hidden dimension  $D$  and sample number  $T$  as follows: Our CE-based optimization assumes that the  $d^m$  dimensions of the hidden subspaces are known. Without this knowledge the CE method becomes inefficient and one has to try all possible combinations for the dimensions of the subspaces. The possible set of subspace dimensions is constrained by the full dimension  $D$  of the problem:  $D = d^1 + \dots + d^M$  ( $d^m > 0, M \leq D$ ). The number of these possibilities is given by the so-called partition function  $f(D)$ , i.e., the number of sets of positive integers that sum up to  $D$ . The value of  $f(D)$  grows quickly with the argument. Asymptotic behavior is known [17, 18]:  $f(D) \sim \exp(\pi\sqrt{2D/3})/(4D\sqrt{3})$  as  $D \rightarrow \infty$ . For given subspace dimensions, CE optimizes iteratively, say in  $max_{it}$  steps. In each iterative step new permutation samples are drawn. The number of these samples is constant,  $c$  times the number of the parameters of the optimization problem [16]. These parameters in the CE permutation-optimization correspond to the transition probabilities of a Markov-chain. Thus, for dimension  $D$ , there are  $D(D-1)$  of them. For each permutation drawn, one has to compute the CE score function from the multi-dimensional Shannon-entropy. The entropy can be estimated by computing the  $T^2$  Euclidean distances between the  $T$  training samples [19]. N-body techniques [20] offer efficient estimations and they require only  $O(T \log(T))$  computations. Putting together, CE based gathering requires *previous knowledge about the dimensions of the subspaces* and for a given set of subspace dimensions the number of computations is in the order of  $O(D^2 T \log(T))$ . By contrast, for the method we suggest here, the gathering requires  $O(D^2)$  computations without any knowledge about the dimension of the hidden subspaces.

### 3.4 Sufficient Conditions for the Separation Theorem

Sufficient conditions of the separation theorem with respect to the  $e^m$  noise processes of Eq. (11) are considered in this subsection. We review known results and provide an extension of the theorem. (For details of the proofs, see Section B of the Appendix.)

**3.4.1 The w-EPI Condition** First, consider the so called Entropy Power Inequality (EPI)

$$e^{2H(\sum_{i=1}^L u_i)} \geq \sum_{i=1}^L e^{2H(u_i)}, \quad (17)$$

where  $u_1, \dots, u_L \in \mathbb{R}$  denote continuous stochastic variables. This inequality holds for example, for independent continuous variables [21], but it may hold for other variables, too.

Let  $S^L$  denote the unit sphere of the  $L$ -dimensional space, that is

$$S^L := \{\mathbf{w} \in \mathbb{R}^L : \|\mathbf{w}\| = 1\}, \quad (18)$$

where  $\|\cdot\|$  denotes the Euclidean norm, i.e., for  $\mathbf{w} \in \mathbb{R}^L$   $\|\mathbf{w}\|^2 := \sum_{i=1}^L w_i^2$ .

**Definition 1** If continuous stochastic variables  $u_1, \dots, u_L \in \mathbb{R}$  satisfy the following entropy inequalities

$$e^{2H(\sum_{i=1}^L w_i u_i)} \geq \sum_{i=1}^L e^{2H(w_i u_i)}, \forall \mathbf{w} \in S^L, \quad (19)$$

then we say that they satisfy the w-EPI condition.

**Proposition 1** The w-EPI property of variables  $e^m$  implies inequality (11) (see Lemma 1).

**3.4.2 Takano's Dependency Criterion** For constant  $\mathbf{w} \in S^d$ , the w-EPI relation reduces to the EPI property [Eq. (17)]. In [22], sufficient condition is provided to satisfy the EPI condition. The condition, which is restricted to the 2-dimensional case, is based on the weak dependencies of the variables. The constraint of  $d = 2$  may be generalized to higher dimensions, but we are not aware of such generalizations.

### 3.4.3 Spherically Symmetric Sources

**Definition 2 (Spherically symmetric variable)** A stochastic variable  $\mathbf{u} \in \mathbb{R}^d$  is called spherically symmetric (or shortly spherical), if its density function is not modified by any rotation. Formally, if

$$\mathbf{u} \stackrel{\text{distr}}{=} \mathbf{O}\mathbf{u}, \quad \forall \mathbf{O} \in \mathcal{O}^d, \quad (20)$$

where  $\stackrel{\text{distr}}{=}$  denotes equality in distribution.

A spherical stochastic variable has a density function and this density function takes constant values on concentric spheres around the origin.

**Proposition 2** For spherically symmetric variables with finite covariance Eq. (11) holds. Further, the stronger w-EPI property [Eq. (19)] also holds and with equality between the two sides ( $\forall \mathbf{w} \in S^d$ ). (See Proof B.2 in the Appendix.)

*Note 2* Spherical variables as well as their non-degenerate affine transforms, the so called elliptical variables (which are equivalent to spherical ones from the point of view of ISA) are thoroughly treated in [23, 24].

**3.4.4 Sources Invariant to 90° Rotation** We have seen that stochastic variables with density functions invariant to orthogonal transformations (spherical variables) satisfy the conditions of the separation theorem. For mixtures of 2-dimensional components ( $d = 2$ ), invariance to 90° rotation suffices. This condition is our novel extension to previous conditions of the ISA separation theorem.

The result: if in the AR-IPA tasks variables  $\mathbf{u} = (u_1, u_2)(= e^m) \in \mathbb{R}^2$  are invariant to 90° rotation, i.e., their density function  $f$  satisfies invariances

$$f(u_1, u_2) = f(-u_2, u_1) = f(-u_1, -u_2) = f(u_2, -u_1) \quad (\forall \mathbf{u} \in \mathbb{R}^2), \quad (21)$$

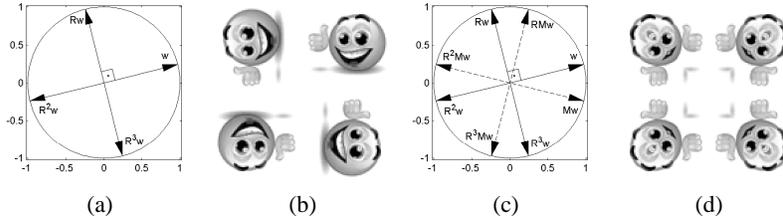


Figure 1: Illustration: Density functions (for variables  $\mathbf{e}^m$ ) invariant to  $90^\circ$  rotation or permutation and sign changes. (a) and (c): density function  $f$  takes identical values at the arrowheads. Matrix  $\mathbf{R}$  and matrix  $\mathbf{M}$  are  $90^\circ$  ccw rotation and reflection to axis  $x$ , respectively. (b) and (d): examples for (a) and (c), respectively.

then the AR-IPA task defined by these variables can be identified by means of the ISA separation theorem. The formulation of this theorem requires care and it can be found in the Appendix (Theorem 2). The proof of the theorem is also provided there.

*Note 3* An important special case of condition (21) is the invariance to permutation and sign changes. Then, for density function  $f$  the following holds

$$f(\pm u_1, \pm u_2) = f(\pm u_2, \pm u_1) \quad (\forall \mathbf{u} \in \mathbb{R}^2). \quad (22)$$

In other words, there exists function  $g : \mathbb{R}^2 \rightarrow \mathbb{R}$ , which is symmetric in its variables and

$$f(\mathbf{u}) = g(|u_1|, |u_2|) \quad (23)$$

holds. In particular, density functions of spherical or  $L^p$ -norm spherical variables [25], such as

$$f(\mathbf{u}) = g\left(\sum_i |u_i|^p\right) \quad (p > 0) \quad (24)$$

also belong to the realm of the theorem.

Illustrations for the theorem are shown in Fig. 1.

*3.4.5 Summary of Sufficient Conditions* Our results for the ISA separation theorem are summarized in Table 1.

## 4 Illustrations

The AR-IPA identification algorithm of Section 3.3 is illustrated below. Test cases are introduced in Section 4.1. The quality of the solutions will be measured by the normalized Amari-error, that we call the Amari-index (Section 4.2). Numerical results are provided in Section 4.3.

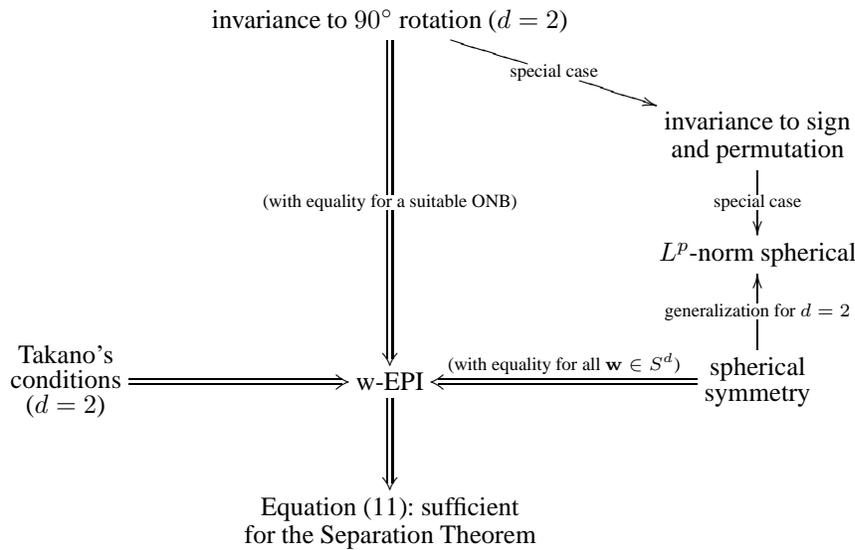


Table 1: Relations amongst sufficient conditions for the ISA separation theorem.

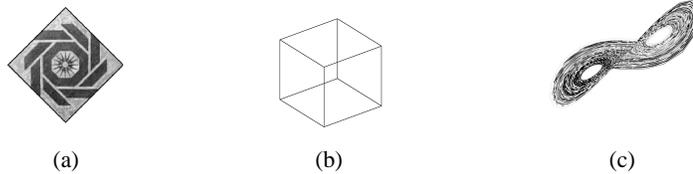


Figure 2: Databases: (a) and (b) depicts examples for the density function of  $\mathbf{e}^m$ , (c) shows an example for  $\mathbf{s}^m$  [see AR-IPA equations (1)–(3)]. (a): *mosaic* database is made of 2D components. Density function of  $\mathbf{e}^m$  is invariant to  $90^\circ$  rotation. (b) and (c) concern tests outside of the conditions of the separation theorem. (b): *3D-geom* database: density functions of  $\mathbf{e}^m$  are identically distributed on 3D geometrical structures. (c): *Lorenz* database: components  $\mathbf{s}^m$  are not AR processes, but 3D chaotic systems; trajectories of Lorenz attractors.

#### 4.1 Databases

Four databases were defined to study the AR-IPA identification algorithm. The first 3 of them are illustrated in Fig. 2.

**4.1.1 Mosaic** This test has 2-dimensional source components generated from images of mosaics. AR components were constructed through driving noise

sources  $\mathbf{e}^m$ . Stochastic noise sources  $\mathbf{e}^m$  were generated by sampling 2-dimensional coordinates proportional to the corresponding pixel intensities. That is, 2-dimensional images of mosaics were considered as density functions. These density functions were invariant to  $90^\circ$  rotation and thus they met condition (11) of the separation theorem.  $M = 4$  was chosen. This database is called *mosaic*.

**4.1.2 3-dimensional Geometric Forms** This example, the *3D-geom* database, is outside of the domain of the sufficient conditions of the separation theorem. Variables  $\mathbf{e}^m$  were uniform distribution variables on 3-dimensional geometric forms ( $d = 3$ ). Three AR processes were generated from the noises ( $M = 3$ ) and were mixed before observation.

**4.1.3 Lorenz Attractor** In this example non-AR sources were used. The sources ( $\mathbf{s}^m$ ) correspond to 3-dimensional ( $d = 3$ ) deterministic chaotic time series, the so called Lorenz attractor [26] with different initial points  $(x_0, y_0, z_0)$  and with different speeds. The Lorenz attractor is described by the following ordinary differential equations:

$$\epsilon \dot{x}(t) = a(y(t) - x(t)), \quad (25)$$

$$\epsilon \dot{y}(t) = x(t)(b - z(t)) - y(t), \quad (26)$$

$$\epsilon \dot{z}(t) = x(t)y(t) - cz(t), \quad (27)$$

where  $\epsilon$  denotes the inverse of the speed of the processes. The database was created by using the standard  $a = 10, b = 27, c = \frac{8}{3}$  parametrization, the differential equations were computed by Euler's method, and  $M = 3$  components were used. This database is called *Lorenz*.

**4.1.4 Led Zeppelin** Here, hidden sources were real world data, stereo Led Zeppelin songs.<sup>2</sup> 8 kHz sampled portions of four songs (Bring It On Home, Heartbreaker, Communication Breakdown, How Many More Times) made the hidden  $\mathbf{s}^m$ s. Thus, the dimension of the components  $d$  was 2 and the number of the components  $M$  was 4.

## 4.2 Normalized Amari-error, the Amari-index

The precision of our algorithm was measured by the normalized Amari-error, that we call the Amari-index. The Amari-error is a 'classical' gauge for the quality of the ICA methods [27]. The normalized form [28] is advantageous, because different ICA methods can be judged on equal footing. The Amari-error has been adapted to the ISA task [9,29]. The performance of the method was gauged by the Amari-index [13]. The index measures, for a given matrix, how close it is to a block-diagonal structure.

The optimal estimation of the AR-IPA model provides matrix  $\mathbf{B} := \mathbf{W}\mathbf{A}$ , a block-permutation matrix made of  $d \times d$  sized blocks. Let us decompose matrix

---

<sup>2</sup> <http://rock.mididb.com/ledzeppelin/>

$\mathbf{B} \in \mathbb{R}^{D \times D}$  into  $d \times d$  blocks:  $\mathbf{B} = [\mathbf{B}^{ij}]_{i,j=1,\dots,M}$ . Let  $b^{ij}$  denote the sum of the absolute values of the elements of matrix  $\mathbf{B}^{ij} \in \mathbb{R}^{d \times d}$ . Then the Amari-index is given as [13]:

$$r(\mathbf{B}) := \frac{1}{2M(M-1)} \left[ \sum_{i=1}^M \left( \frac{\sum_{j=1}^M b^{ij}}{\max_j b^{ij}} - 1 \right) + \sum_{j=1}^M \left( \frac{\sum_{i=1}^M b^{ij}}{\max_i b^{ij}} - 1 \right) \right]. \quad (28)$$

where the normalization extends the one-dimensional definition [28] to higher dimensions. Now, for matrix  $\mathbf{B}$  we have that  $0 \leq r(\mathbf{B}) \leq 1$ , and  $r(\mathbf{B}) = 0$  if, and only if  $\mathbf{B}$  is a block-permutation matrix with  $d \times d$  sized blocks. Thus,  $r = 0$  corresponds to perfect estimation (0% error),  $r = 1$  is the worst estimation (100% error).

### 4.3 Simulations

Results on databases *mosaic*, *3D-geom*, *Lorenz* and *Led Zeppelin* are provided here. Our gauge to measure the quality of the results is the Amari-index (Section 4.2) that we computed by averaging over 50 randomly chosen computations. These experimental studies concerned the following problems:

1. The quality of the gathering procedure depends on threshold parameter  $\varepsilon$ . We studied the estimation error, the Amari-index, as a function of sample number. The  $\varepsilon$  values were preset to reasonably good values.
2. We studied the optimal domain for the  $\varepsilon$  values. We looked for the dynamic range, i.e., the ratio of the highest and lowest ‘good  $\varepsilon$  values’: We divided interval  $[0, F_{max}]$  ( $F_{max} := \max_{i,j} |\hat{F}_{ij}|$ ) into 200 equal parts. For different sample numbers in all databases at each division point we used the gathering procedure to group the ICA elements. For each of the 50 random trials we have computed the Amari-indices separately. For the smallest Amari-index, we determined the corresponding interval of  $\varepsilon$ ’s, these are the ‘good  $\varepsilon$  values’. Then we took the ratio of the largest and smallest  $\varepsilon$  values in this set and averaged the ratios over the 50 runs. The average is called the dynamic range.

In our simulations, sample number  $T$  of observations  $\mathbf{z}(t)$  was varied between 1,000 and 200,000. Mixing matrix  $\mathbf{A}$  was random and orthogonal matrix. In the case of the *mosaic* and *3D-geom* tests, quantities  $\mathbf{F}$ , and  $\mathbf{e}$  were drawn randomly, and only stable AR processes were allowed. In the *Lorenz* and *Led Zeppelin* tests, hidden processes were estimated for different  $1 \leq p \leq 10$  depth AR process assumptions. For AR order  $p > 1$ , we used quantities  $|\hat{F}_{ij}| := \sum_{k=1}^p |\hat{F}_{ij}^k|$ , where  $\hat{F}_{ij}^k$  is the  $ij^{th}$  element of matrix  $\hat{\mathbf{F}}^k$  ( $k = 1, \dots, p$ ) and estimated  $\hat{\mathbf{F}}$ -‘connectedness’ and  $F_{max}$  by means of these quantities. For all datasets, we used the method described in [30,31] for the identification of the AR process. We used fastICA [32] on the estimated innovations.

Our results are summarized in Fig. 3 and Fig. 4. According to Fig. 3, there are good  $\varepsilon$  parameters for the  $\hat{\mathbf{F}}$ -‘connectedness’ already for 10,000 – 20,000

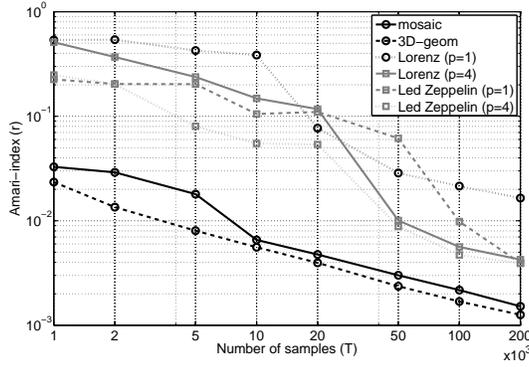


Figure 3: Amari-index as a function of sample number for the *3D-geom*, *mosaic*, *Lorenz* and *Led Zeppelin* databases on log-log scale.  $p$ : the order of the AR estimation for the *Lorenz* and the *Led Zeppelin* datasets.

samples: our method can find the hidden components with high precision. For  $T = 10,000$  samples Fig. 5 and Fig. 6 illustrates this for the *mosaic* and the *3D-geom* databases, respectively. Figure 7 illustrates that for sample number  $T = 10,000$  and for  $p = 1$  assumption on the AR process, we can get reasonable estimations for the *Lorenz* database. Figure 3 shows that usually for this database more samples are necessary. Figure 3 also shows that by increasing the sample number the Amari-index decreases; it is 0.15% for the *mosaic*, 0.13% for the *3D-geom* databases, whereas for the *Lorenz* and *Led Zeppelin* tests, it is 1.65% and 0.39% for  $p = 1$ , 0.42% and 0.38% for  $p = 4$ , respectively on the average for 200,000 samples. For the *mosaic* and the *3D-geom* tests, the Amari-index decreases according to power law  $r(T) \propto T^{-c}$  ( $c > 0$ ). In these tests hidden sources  $s$  are true AR processes as opposed to the other tests, where the AR assumption was only an approximation. The power law decline is manifested by straight line on log-log scale. The slopes of these straight lines are very close to one another. In the *Lorenz* and *Led Zeppelin* tests, AR estimations with  $p > 4$  did not improve the results. Figure 4 demonstrates that for larger sample numbers threshold parameter  $\varepsilon$  that determines the  $\hat{F}$ -‘connected’ property can be chosen from a broader domain; the dynamic range grows. For the *mosaic* and the *3D-geom* databases this is 80 and 86, for the *Lorenz* test it is 14 ( $p = 1$ ), 28 ( $p = 4$ ), for the *Led Zeppelin* test it is 6.7 ( $p = 1$ ), and 11 ( $p = 4$ ) for 200,000 samples. The dynamic range did not improve for the *Lorenz* and the *Led Zeppelin* tests by assuming AR processes with order larger than 4.

## 5 Discussion

We have introduced a novel AR-IPA identification algorithm. Previous AR-IPA algorithms [11, 13, 12] assumed that the number of the hidden processes (the components) as well as the dimensions of the components are known. In these meth-

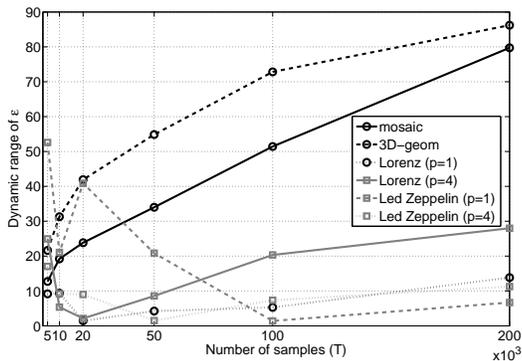


Figure 4: Dynamic range as a function of sample number for the *3D-geom*, *mosaic*, *Lorenz* and *Led Zeppelin* databases.  $p$ : the order of the AR estimation for the *Lorenz* and the *Led Zeppelin* datasets.

ods computation time consuming, multi-dimensional entropy estimations were applied [11, 13]. The constraints and the computation load can be eased by our novel method.

In our approach, first the AR-IPA task is reduced to an ISA problem by using the concept of innovations. Then we make use of the ISA separation theorem and reduce the ISA problem to an ICA task *and* to the search of optimal permutation. Although the permutation problem can be solved, e.g., by the efficient CE method [16], it requires previously known dimensions of the subspaces and further, it requires the estimation of multi-dimensional entropy terms or the estimation of mutual information.

These difficulties can be reduced by our method:

1. We recover up to permutation the coordinates of hidden processes ( $s^m$ ) using the separation theorem.
2. We group the one-dimensional processes, which are ‘connected’.

‘Connectedness’ has been defined through the estimated predictive matrix  $\hat{\mathbf{F}}$ . (We assumed that  $\mathbf{F}^m \neq 0$ .) Matrix  $\mathbf{F}$ , the block-diagonal matrix formed by matrices  $\mathbf{F}^m$  can be estimated by means of the estimated separation matrix  $\hat{\mathbf{W}}_{ICA}$  and the estimated predictive matrix  $\hat{\mathbf{F}}_z$  of the observed AR process [see Eq. (16)].

The true matrix  $\mathbf{F}$  is block-diagonal [ $\mathbf{F} = \text{blockdiag}(\mathbf{F}^1, \dots, \mathbf{F}^M)$ ] according to the AR-IPA model [Eq. (1)–(3)] and we can estimate it by matrix  $\hat{\mathbf{F}}$  up to permutation provided that the separation theorem holds. This property is exploited through our definition of  $\hat{\mathbf{F}}$ -‘connectedness’ of the  $i^{\text{th}}$  and the  $j^{\text{th}}$  coordinates using quantity  $\max(|F_{ij}|, |F_{ji}|)$ . The concept of  $\hat{\mathbf{F}}$ -‘connectedness’ enabled a fast *gathering procedure*.

Our AR-IPA algorithm needs extension if one of the matrices  $(\mathbf{W}_{ICA}^m) \mathbf{F}^m (\mathbf{W}_{ICA}^m)^T$  is itself a block-diagonal matrix made of more than 1 block, or if it becomes block-

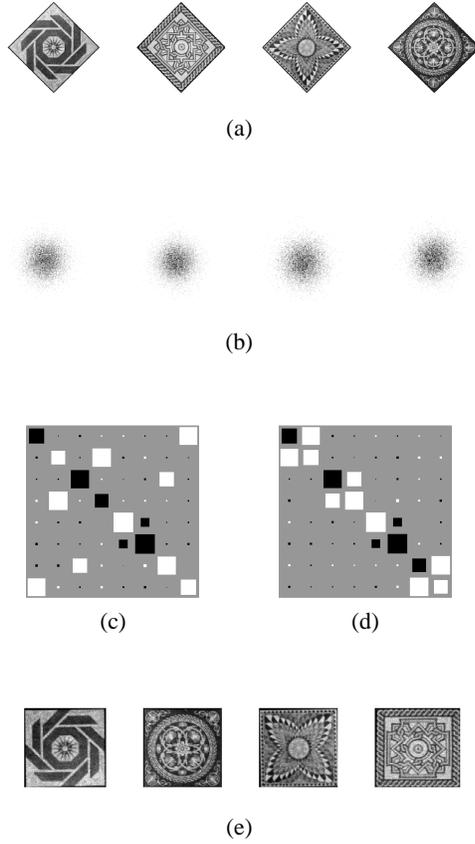


Figure 5: AR-IPA estimation on the *mosaic* database. (a): densities of  $e^m$ . (b): 10,000 samples of the observed mixed signals  $[\mathbf{z}(t)]$ . This forms the input of our algorithm. (c): Hinton-diagram of matrix  $\hat{\mathbf{F}}$  belonging to the ICA coordinates of the estimated AR process  $\hat{\mathbf{s}}(t)$ . Colors: white-positive, black-negative value. (d): same as (c) after reordering according to groups connected by matrix  $\hat{\mathbf{F}}$ . (e): estimated noise sources  $\hat{e}^m$  illustrated on 1 million sample points as uncovered by the reordering procedure. Note that noise components ( $e^m$ ) are recovered, but only up to permutation and orthogonal transformation.

diagonal during the estimation process.<sup>3</sup> In this case our gathering procedure based on the measure of the connectedness of matrix  $\hat{\mathbf{F}}$  is insufficient. Other methods, such as the estimation of mutual information  $I(\hat{s}_i, \hat{s}_j)$  may be used instead to group the components. Consider, however,  $d$ -dimensional stochastic variables that are dependent, but for which any of the  $d - 1$  dimensional subset of the coordinates is independent. (For a construction of ‘*all-(d-1)-independent variable*’, see

<sup>3</sup>  $\mathbf{W}_{ICA}^m$  denotes the component of separation matrix  $\mathbf{W}_{ICA}$  that corresponds to the  $m^{th}$  sub-process.

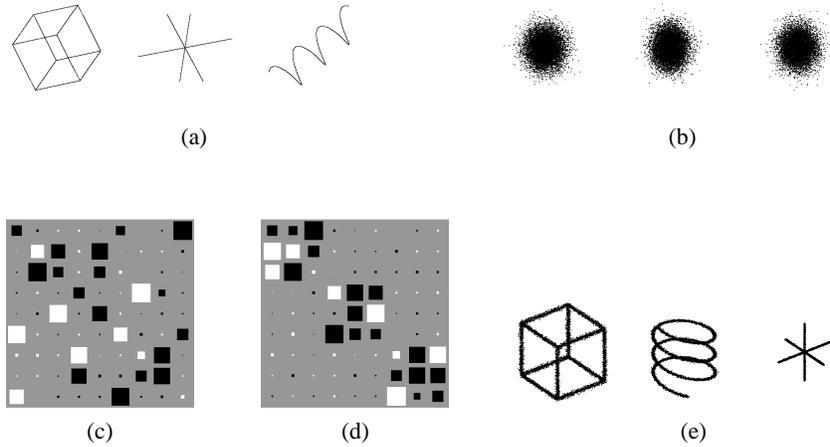


Figure 6: AR-IPA estimation on the *3D-geom* database. (a): densities of  $e^m$ . (b): 10,000 samples of the observed mixed signals  $[z(t)]$ . (c): Hinton-diagram of matrix  $\hat{F}$  belonging to the ICA coordinates of the estimated AR process  $\hat{s}(t)$ . (d): same as (c) after reordering according to groups connected by matrix  $\hat{F}$ . (e): estimated noise sources ( $\hat{e}^m$ ).

[8, 13].) For such sources, previous methods are unsatisfactory and estimation of the  $d$ -dimensional Shannon-entropy, or that of the mutual information may be necessary [6, 8, 7].

Our numerical simulations have some indications that deserve further investigations.

1. The crude but fast  $\hat{F}$ -‘connectedness’ criterion may be able to group the appropriate components. We found that the block-diagonal feature of the estimated predictive matrix increased upon increasing the number of samples. This experience was pronounced for the Lorenz dynamical system.
2. ICA components can be grouped for sources which are outside of our sufficient conditions for the separation theorem. It seems that further generalization of the separation theorem should be possible.
3. AR estimation can be used as a trick to attempt to separate more complex, e.g., mixed chaotic processes or songs. It then may enable higher order estimations and time series prediction in lower dimensional spaces.

## 6 Conclusions

We have shown that oftentimes the separation of *multi-dimensional and mixed* processes is feasible without knowing the number and the dimensions of the hid-

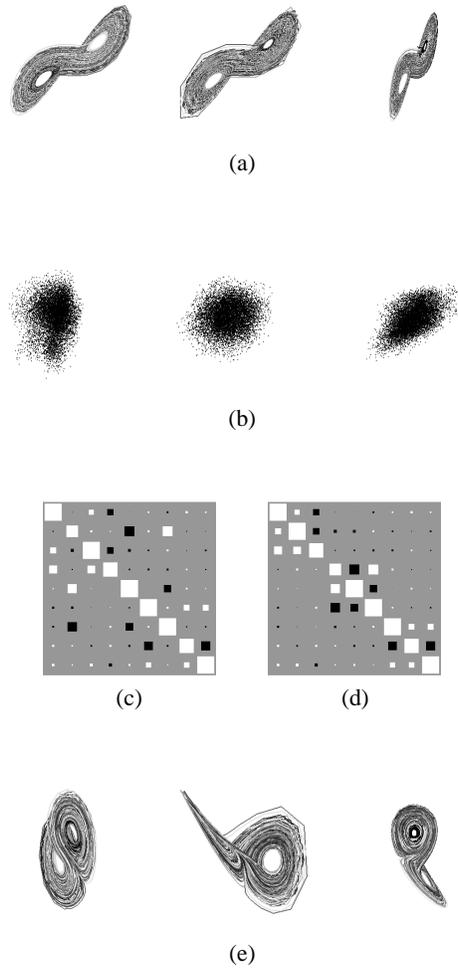


Figure 7: AR-IPA estimation on the *Lorenz* database with  $p = 1$  order AR assumption. (a): non-AR hidden processes  $\mathbf{s}^m$ , (b): observed mixed 10,000 sample points  $[\mathbf{z}(t)]$ . (c): Hinton-diagram of matrix  $\hat{\mathbf{F}}$  belonging to the ICA coordinates of the estimated AR process  $\hat{\mathbf{s}}(t)$ . (d): same as (c) after reordering according to groups connected by matrix  $\hat{\mathbf{F}}$ . (e): Estimation of the sources ( $\hat{\mathbf{s}}^m$ ).

den processes and without the computation time consuming estimations of multi-dimensional entropy or mutual information terms. We have shown a new sufficient condition for the ‘Separation Theorem’ that forms the basis of our algorithm. Numerical simulations point to the possibility of further extensions of the conditions of the theorem. Our AR estimation can be used to separate more complex, e.g., chaotic processes or mixed songs that may considerably lessen the estimation and prediction problems of time series.

## Appendix

The theorems that we present here concern the ISA task that we gain after reducing the AR-IPA task and thus, here  $\mathbf{s}^m = \mathbf{e}^m$  ( $m = 1, \dots, M$ ). Since the ISA task also concerns source, but these sources exhibit the i.i.d. property, thus we shall use notation  $\mathbf{s}^m$ . In the present work, the differential entropy  $H$  is defined by the logarithm of base  $e$ .

### A The ISA Separation Theorem (Proof)

The main idea of our ISA separation theorem is that the ISA task may be accomplished in two steps under certain conditions. In the first step ICA is executed. The second step is search for the optimal permutation of the ICA components.

If EPI [see Eq. (17)] is satisfied (on  $S^L$ ) then a further inequality holds:

**Lemma 1** *Suppose that continuous stochastic variables  $u_1, \dots, u_L \in \mathbb{R}$  satisfy the w-EPI condition [see Eq. (19)]. Then, they also satisfy*

$$H\left(\sum_{i=1}^L w_i u_i\right) \geq \sum_{i=1}^L w_i^2 H(u_i), \forall \mathbf{w} \in S^L. \quad (29)$$

*Note 4* w-EPI holds, for example, for independent variables  $u_i$ , because independence is not affected by multiplication with a constant.

*Proof* Assume that  $\mathbf{w} \in S^L$ . Applying  $\ln$  on condition (19), and using the monotonicity of the  $\ln$  function, we can see that the first inequality is valid in the following inequality chain

$$\begin{aligned} 2H\left(\sum_{i=1}^L w_i u_i\right) &\geq \ln\left(\sum_{i=1}^L e^{2H(w_i u_i)}\right) = \ln\left(\sum_{i=1}^L e^{2H(u_i)} \cdot w_i^2\right) \\ &\geq \sum_{i=1}^L w_i^2 \cdot \ln\left(e^{2H(u_i)}\right) = \sum_{i=1}^L w_i^2 \cdot 2H(u_i). \end{aligned} \quad (30)$$

Then,

1. we used the relation [21]:

$$H(w_i u_i) = H(u_i) + \ln(|w_i|) \quad (31)$$

for the entropy of the transformed variable. Hence

$$e^{2H(w_i u_i)} = e^{2H(u_i) + 2\ln(|w_i|)} = e^{2H(u_i)} \cdot e^{2\ln(|w_i|)} = e^{2H(u_i)} \cdot w_i^2. \quad (32)$$

2. In the second inequality, we utilized the concavity of  $\ln$ .

Now we shall use Lemma 1 to proceed. The separation theorem will be a corollary of the following claim:

**Proposition 3** Let  $\mathbf{y} = [\mathbf{y}^1; \dots; \mathbf{y}^M] = \mathbf{y}(\mathbf{W}) = \mathbf{W}\mathbf{s}$ , where  $\mathbf{W} \in \mathcal{O}^D$ ,  $\mathbf{y}^m$  is the estimation of the  $m^{\text{th}}$  component of the ISA task. Let  $y_i^m$  be the  $i^{\text{th}}$  coordinate of the  $m^{\text{th}}$  component. Similarly, let  $s_i^m$  stand for the  $i^{\text{th}}$  coordinate of the  $m^{\text{th}}$  source. Let us assume that the  $\mathbf{s}^m$  sources satisfy condition (11). Then

$$\sum_{m=1}^M \sum_{i=1}^d H(y_i^m) \geq \sum_{m=1}^M \sum_{i=1}^d H(s_i^m). \quad (33)$$

*Proof* Let us denote the  $(i, j)^{\text{th}}$  element of matrix  $\mathbf{W}$  by  $W_{i,j}$ . Coordinates of  $\mathbf{y}$  and  $\mathbf{s}$  will be denoted by  $y_i$  and  $s_i$ , respectively. Let  $\mathcal{G}^1, \dots, \mathcal{G}^M$  denote indices belonging to the  $1^{\text{st}}, \dots, M^{\text{th}}$  subspaces, respectively, that is,  $\mathcal{G}^1 := \{1, \dots, d\}, \dots, \mathcal{G}^M := \{D-d+1, \dots, D\}$ . Now, writing the elements of the  $i^{\text{th}}$  row of matrix multiplication  $\mathbf{y} = \mathbf{W}\mathbf{s}$ , we have

$$y_i = \sum_{j \in \mathcal{G}^1} W_{i,j} s_j + \dots + \sum_{j \in \mathcal{G}^M} W_{i,j} s_j \quad (34)$$

and thus,

$$H(y_i) = H\left(\sum_{j \in \mathcal{G}^1} W_{i,j} s_j + \dots + \sum_{j \in \mathcal{G}^M} W_{i,j} s_j\right) \quad (35)$$

$$= H\left\{\sum_{m=1}^M \left[\left(\sum_{l \in \mathcal{G}^m} W_{i,l}^2\right)^{\frac{1}{2}} \frac{\sum_{j \in \mathcal{G}^m} W_{i,j} s_j}{\left(\sum_{l \in \mathcal{G}^m} W_{i,l}^2\right)^{\frac{1}{2}}}\right]\right\} \quad (36)$$

$$\geq \sum_{m=1}^M \left[\left(\sum_{l \in \mathcal{G}^m} W_{i,l}^2\right) H\left(\frac{\sum_{j \in \mathcal{G}^m} W_{i,j} s_j}{\left(\sum_{l \in \mathcal{G}^m} W_{i,l}^2\right)^{\frac{1}{2}}}\right)\right] \quad (37)$$

$$= \sum_{m=1}^M \left[\left(\sum_{l \in \mathcal{G}^m} W_{i,l}^2\right) H\left(\sum_{j \in \mathcal{G}^m} \frac{W_{i,j}}{\left(\sum_{l \in \mathcal{G}^m} W_{i,l}^2\right)^{\frac{1}{2}}} s_j\right)\right] \quad (38)$$

$$\geq \sum_{m=1}^M \left[\left(\sum_{l \in \mathcal{G}^m} W_{i,l}^2\right) \sum_{j \in \mathcal{G}^m} \left(\frac{W_{i,j}}{\left(\sum_{l \in \mathcal{G}^m} W_{i,l}^2\right)^{\frac{1}{2}}}\right)^2 H(s_j)\right] \quad (39)$$

$$= \sum_{j \in \mathcal{G}^1} W_{i,j}^2 H(s_j) + \dots + \sum_{j \in \mathcal{G}^M} W_{i,j}^2 H(s_j) \quad (40)$$

The above steps can be justified as follows:

1. (35): Eq. (34) was inserted into the argument of  $H$ .
2. (36): New terms were added for Lemma 1.
3. (37): Sources  $\mathbf{s}^m$  are independent of each other and this independence is preserved upon mixing *within* the subspaces, and we could also use Lemma 1, because  $\mathbf{W}$  is an orthogonal matrix.

4. (38): Nominators were transferred into the  $\sum_j$  terms.
5. (39): Variables  $\mathbf{s}^m$  satisfy condition (11) according to our assumptions.
6. (40): We simplified the expression after squaring.

Using this inequality, summing it for  $i$ , exchanging the order of the sums, and making use of the orthogonality of matrix  $\mathbf{W}$ , we have

$$\sum_{i=1}^D H(y_i) \geq \sum_{i=1}^D \sum_{m=1}^M \left[ \sum_{j \in \mathcal{G}^m} W_{i,j}^2 H(s_j) \right] \quad (41)$$

$$= \sum_{m=1}^M \left[ \sum_{j \in \mathcal{G}^m} \left( \sum_{i=1}^D W_{i,j}^2 \right) H(s_j) \right] \quad (42)$$

$$= \sum_{j=1}^D H(s_j). \quad (43)$$

*Note 5* The proof holds if the dimensions of the subspaces are not equal. The same is true for the ISA separation theorem.

Having this proposition, now we prove our main theorem (Theorem 1).

*Proof* ICA minimizes the l.h.s. of Eq. (33), that is, it minimizes  $\sum_{m=1}^M \sum_{i=1}^d H(y_i^m)$ . The set of minima is invariant for permutations and sign changes and according to Proposition 3,  $\{s_i^m\}$  – that is the coordinates of components  $\mathbf{s}^m$  of the ISA task – belong to the set of minima.

## B Sufficient Conditions of the Separation Theorem

In the separation theorem, we assumed that relation (11) is fulfilled for the  $\mathbf{s}^m$  sources. Below, we present sufficient conditions – together with proofs – when this inequality is fulfilled.

### B.1 w-EPI

According to Lemma 1, if the w-EPI property [i.e., (19)] holds for sources  $\mathbf{s}^m$ , then inequality (11) holds, too.

### B.2 Spherically Symmetric Sources

We shall make use of the following well-known property of spherically symmetric variables [23,24]:

**Lemma 2** *Let  $\mathbf{v}$  denote a  $d$ -dimensional variable, which is spherically symmetric. Then the projection of  $\mathbf{v}$  onto lines through the origin have identical univariate distribution.*

**Lemma 3** *The expectation value and the variance of a  $d$ -dimensional  $\mathbf{v}$  spherically symmetric variable are*

$$E[\mathbf{v}] = \mathbf{0}, \quad (44)$$

$$\text{Var}[\mathbf{v}] = c(\text{onstant}) \cdot \mathbf{I}_d. \quad (45)$$

*Proof* Here, we show that the w-EPI property is fulfilled with equality for spherical sources. According to (44)–(45), spherically symmetric sources  $\mathbf{s}^m$  have zero expectation values and up to a constant multiplier they also have identity covariance matrices:

$$E[\mathbf{s}^m] = \mathbf{0}, \quad (46)$$

$$\text{Var}[\mathbf{s}^m] = c^m \cdot \mathbf{I}_d. \quad (47)$$

Note that our constraint on the ISA task, namely that covariance matrices of the  $\mathbf{s}^m$  sources should be equal to  $\mathbf{I}_d$ , is fulfilled up to constant multipliers.

Let  $P_{\mathbf{w}}$  denote the projection to straight line with direction  $\mathbf{w} \in S^d$ , which crosses the origin, i.e.,

$$P_{\mathbf{w}} : \mathbb{R}^d \ni \mathbf{u} \mapsto \sum_{i=1}^d w_i u_i \in \mathbb{R}. \quad (48)$$

In particular, if  $\mathbf{w}$  is chosen as the canonical basis vector  $\mathbf{e}_i$  (all components are 0, except the  $i^{\text{th}}$  component, which is equal to 1), then

$$P_{\mathbf{e}_i}(\mathbf{u}) = u_i. \quad (49)$$

In this interpretation, w-EPI ((19)) is concerned with the entropies of the projections of the different sources onto straight lines crossing the origin. The l.h.s. projects to  $\mathbf{w}$ , whereas the r.h.s. projects to the canonical basis vectors. Let  $\mathbf{u}$  denote an arbitrary source, i.e.,  $\mathbf{u} := \mathbf{s}^m$ . According to Lemma 2, distribution of the spherical  $\mathbf{u}$  is the same for all such projections and thus their entropies are identical. That is,

$$\sum_{i=1}^d w_i u_i \stackrel{\text{distr}}{=} u_1 \stackrel{\text{distr}}{=} \dots \stackrel{\text{distr}}{=} u_d, \quad \forall \mathbf{w} \in S^d, \quad (50)$$

$$H\left(\sum_{i=1}^d w_i u_i\right) = H(u_1) = \dots = H(u_d), \quad \forall \mathbf{w} \in S^d. \quad (51)$$

Thus:

- l.h.s. of w-EPI is equal to  $e^{2H(u_1)}$ .
- r.h.s. of w-EPI can be written as follows:

$$\sum_{i=1}^d e^{2H(w_i u_i)} = \sum_{i=1}^d e^{2H(u_i)} \cdot w_i^2 = e^{2H(u_1)} \sum_{i=1}^d w_i^2 = e^{2H(u_1)} \cdot 1 = e^{2H(u_1)} \quad (52)$$

At the first step, we used identity (32) for each of the terms. At the second step, (51) was utilized. Then term  $e^{2H(u_1)}$  was pulled out and we took into account that  $\mathbf{w} \in S^d$ .

*Note 6* We note that sources of spherically symmetric distribution have already been used in the context of ISA in [33]. In that work, a generative model was assumed. According to the assumption, the distribution of the norms of sample projections to the subspaces were independent. This way, the task was restricted to spherically symmetric source distributions, which is a special case of the general ISA task.

### B.3 Sources Invariant to 90° Rotation

By definition, spherical variables are invariant to orthogonal transformations [see Eq. (20)]. For mixtures of 2-dimensional components ( $d = 2$ ), much milder condition, invariance to 90° rotation, suffices. First, we observe that:

*Note 7* In the ISA separation theorem, it is enough if some orthogonal transformation of the  $\mathbf{s}^m$  sources,  $\mathbf{C}^m \mathbf{s}^m$  ( $\mathbf{C}^m \in \mathcal{O}^d$ ) satisfy the condition (11). In this case, the  $\mathbf{C}^m \mathbf{s}^m$  variables are extracted by the permutation search after the ICA transformation. Because the ISA identification has ambiguities up to orthogonal transformation in the respective subspaces, this is suitable. In other words, for the ISA identification the existence of an Orthonormal Basis (ONB) for each  $\mathbf{u} := \mathbf{s}^m \in \mathbb{R}^d$  components is sufficient, on which the

$$h : \mathbb{R}^d \ni \mathbf{w} \mapsto H[\langle \mathbf{w}, \mathbf{u} \rangle] \quad (53)$$

function takes its minimum. (Here, the  $\langle \mathbf{w}, \mathbf{u} \rangle := \sum_{i=1}^d w_i u_i$  stochastic variable is the projection of  $\mathbf{u}$  to the direction  $\mathbf{w}$ .) In this case, the entropy inequality (11) is met with equality on the elements of the ONB.

Now we present our theorem concerning to the  $d = 2$  case.

**Theorem 2** *Let us suppose, that the density function  $f$  of stochastic variable  $\mathbf{u} = (u_1, u_2) (= \mathbf{s}^m) \in \mathbb{R}^2$  exhibits the invariance*

$$f(u_1, u_2) = f(-u_2, u_1) = f(-u_1, -u_2) = f(u_2, -u_1) \quad (\forall \mathbf{u} \in \mathbb{R}^2), \quad (54)$$

*that is, it is invariant to 90° rotation. If function  $h(\mathbf{w}) = H[\langle \mathbf{w}, \mathbf{u} \rangle]$  has minimum on the set  $\{\mathbf{w} \geq \mathbf{0}\} \cap S^2$ , it also has minimum on an ONB. (Relation  $\mathbf{w} \geq \mathbf{0}$  concerns each coordinates.) Consequently, the ISA task can be identified by the use of the separation theorem.*

*Proof* Let

$$\mathbf{R} := \begin{bmatrix} 0 & -1 \\ 1 & 0 \end{bmatrix} \quad (55)$$

denote the matrix of 90° ccw rotation. Let  $\mathbf{w} \in S^2$ .  $\langle \mathbf{w}, \mathbf{u} \rangle \in \mathbb{R}$  is the projection of variable  $\mathbf{u}$  onto  $\mathbf{w}$ . The value of the density function of the stochastic variable  $\langle \mathbf{w}, \mathbf{u} \rangle$  in  $t \in \mathbb{R}$  (we move  $t$  in direction  $\mathbf{w}$ ) can be calculated by integration starting from the point  $\mathbf{w}t$ , in direction perpendicular to  $\mathbf{w}$

$$f_{y=y(\mathbf{w})=\langle \mathbf{w}, \mathbf{u} \rangle}(t) = \int_{\mathbf{w}^\perp} f(\mathbf{w}t + \mathbf{z}) d\mathbf{z}. \quad (56)$$

Using the supposed invariance of  $f$  and the relation (56) we have

$$f_y(\mathbf{w}) = f_y(\mathbf{R}\mathbf{w}) = f_y(\mathbf{R}^2\mathbf{w}) = f_y(\mathbf{R}^3\mathbf{w}), \quad (57)$$

where ‘=’ denotes the equality of functions. Consequently, it is enough to optimize  $h$  on the set  $\{\mathbf{w} \geq \mathbf{0}\}$ . Let  $\mathbf{w}_{min}$  be the minimum of function  $h$  on the set  $S^2 \cap \{\mathbf{w} \geq \mathbf{0}\}$ . According to Eq. (57),  $h$  takes constant and minimal values in the

$$\{\mathbf{w}_{min}, \mathbf{R}\mathbf{w}_{min}, \mathbf{R}^2\mathbf{w}_{min}, \mathbf{R}^3\mathbf{w}_{min}\}$$

points.  $\{\mathbf{v}_{min}, \mathbf{R}\mathbf{v}_{min}\}$  is a suitable ONB in Note 7.

*Acknowledgements* This research has been supported by the EC NEST ‘Perceptual Consciousness: Explication and Testing’ grant under contract 043261. Opinions and errors in this manuscript are the author’s responsibility, they do not necessarily reflect those of the EC or other project members.

## References

1. Hyvärinen, A., Karhunen, J., Oja, E.: Independent Component Analysis. John Wiley & Sons (2001)
2. Cichocki, A., Amari, S.: Adaptive blind signal and image processing. John Wiley & Sons (2002)
3. Cardoso, J.: Multidimensional independent component analysis. In: International Conference on Acoustics, Speech, and Signal Processing (ICASSP ’98). Volume 4. (1998) 1941–1944
4. Akaho, S., Kiuchi, Y., Umeyama, S.: MICA: Multimodal independent component analysis. In: International Joint Conference on Neural Networks (IJCNN ’99). Volume 2. (1999) 927–932
5. Vollgraf, R., Obermayer, K.: Multi-dimensional ICA to separate correlated sources. In: Neural Information Processing Systems (NIPS 2001). Volume 14., Cambridge, MA, MIT Press (2001) 993–1000
6. Bach, F.R., Jordan, M.I.: Beyond independent components: Trees and clusters. Journal of Machine Learning Research **4** (2003) 1205–1233
7. Póczos, B., Lőrincz, A.: Independent subspace analysis using k-nearest neighborhood distances. Artificial Neural Networks: Formal Models and their Applications (ICANN 2005) **3697** (2005) 163–168
8. Póczos, B., Lőrincz, A.: Independent subspace analysis using geodesic spanning trees. In: International Conference on Machine Learning (ICML 2005). Volume 119., New York, NY, USA, ACM Press (2005) 673–680
9. Theis, F.J.: Blind signal separation into groups of dependent signals using joint block diagonalization. In: International Society for Computer Aided Surgery (ISCAS 2005). Volume 6. (2005) 5878–5881
10. Van Hulle, M.M.: Edgeworth approximation of multivariate differential entropy. Neural Computation **17** (2005) 1903–1910
11. Póczos, B., Takács, B., Lőrincz, A.: Independent subspace analysis on innovations. In: European Conference on Machine Learning (ECML 2005). Volume 3720 of LNAI., Springer-Verlag (2005) 698–706

12. Hyvärinen, A.: Independent component analysis for time-dependent stochastic processes. In: International Conference on Artificial Neural Networks (ICANN '98), Berlin, Springer-Verlag (1998) 541–546
13. Szabó, Z., Póczos, B., Lőrincz, A.: Cross-entropy optimization for independent process analysis. In: Independent Component Analysis and Blind Signal Separation (ICA 2006). Volume 3889 of LNCS., Springer (2006) 909–916
14. Cheung, Y., Xu, L.: Dual multivariate auto-regressive modeling in state space for temporal signal separation. *IEEE Transaction on Systems, Man, Cybernetics—Part B* **33** (2003) 386–398
15. Theis, F.J.: Uniqueness of complex and multidimensional independent component analysis. *Signal Processing* **84** (2004) 951–956
16. Rubinstein, R.Y., Kroese, D.P.: *The Cross-Entropy Method*. Springer (2004)
17. Hardy, G.H., Ramanujan, S.I.: Asymptotic formulae in combinatory analysis. *Proceedings of the London Mathematical Society* **17** (1918) 75–115
18. Uspensky, J.V.: Asymptotic formulae for numerical functions which occur in the theory of partitions. *Bulletin of the Russian Academy of Sciences* **14** (1920) 199–218
19. Costa, J.A., Hero, A.O.: Manifold learning using k-nearest neighbor graphs. In: International Conference on Acoustic Speech and Signal Processing (ICASSP 2004). Volume 4. (2004) 988–991
20. Gray, A.G., Moore, A.W.: 'N-Body' problems in statistical learning. In: Proceedings of NIPS. (2000) 521–527
21. Cover, T.M., Thomas, J.A.: *Elements of information theory*. John Wiley and Sons, New York, USA (1991)
22. Takano, S.: The inequalities of Fisher information and entropy power for dependent variables. In: Proceedings of the 7th Japan-Russia Symposium on Probability Theory and Mathematical Statistics. (1995)
23. Fang, K.T., Kotz, S., Ng, K.W.: *Symmetric multivariate and related distributions*. Chapman and Hall (1990)
24. Frahm, G.: *Generalized elliptical distributions: Theory and applications*. PhD thesis, University of Köln (2004)
25. Gupta, A.K., Song, D.:  $L^p$ -norm spherical distribution. *Journal of Statistical Planning and Inference* **60** (1997)
26. Lorenz, E.N.: Deterministic nonperiodic flow. *Journal of Atmospheric Sciences* **20** (1963) 130–141
27. Amari, S., Cichocki, A., Yang, H.H.: A new learning algorithm for blind signal separation. *Advances in Neural Information Processing Systems* **8** (1996) 757–763
28. Bach, F.R., Jordan, M.I.: Kernel independent component analysis. *Journal of Machine Learning Research* **3** (2002) 1–48
29. Theis, F.J.: Multidimensional independent component analysis using characteristic functions. In: European Signal Processing Conference (EUSIPCO 2005). (2005)
30. Neumaier, A., Schneider, T.: Estimation of parameters and eigenmodes of multivariate autoregressive models. *ACM Transactions on Mathematical Software* **27** (2001) 27–57
31. Schneider, T., Neumaier, A.: Algorithm 808: ARfit - a matlab package for the estimation of parameters and eigenmodes of multivariate autoregressive models. *ACM Transactions on Mathematical Software* **27** (2001) 58–65
32. Hyvärinen, A., Oja, E.: A fast fixed-point algorithm for independent component analysis. *Neural Computation* **9** (1997) 1483–1492
33. Hyvärinen, A., Hoyer, P.O.: Emergence of phase and shift invariant features by decomposition of natural images into independent feature subspaces. *Neural Computation* **12** (2000) 1705–1720