Hierarchical Independent Component Analysis: a multi-resolution non-orthogonal data-driven basis

Piercesare Secchi[a], Simone Vantini[a], Paolo Zanini[a]

[a]*MOX – Department of Mathematics, Politecnico di Milano*

Abstract

A new method named Hierarchical Independent Component Analysis is presented, particularly suited for dealing with two problems regarding the analysis of high-dimensional and complex data: dimensional reduction and multi-resolution analysis. It takes into account the Blind Source Separation framework, where the purpose is the research of a basis for a dimensional reduced space to represent data, whose basis elements represent physical features of the phenomenon under study. In this case orthogonal basis could be not suitable, since the orthogonality introduce an artificial constraint not related to the phenomenological properties of the analyzed problem. For this reason this new approach is introduced. It is obtained through the integration between Treelets and Independent Component Analysis, and it is able to provide a multi-scale non-orthogonal data-driven basis. Furthermore a strategy to perform dimensional reduction with a non orthogonal basis is presented and the theoretical properties of Hierarchical Independent Component Analysis are analyzed. Finally HICA algorithm is tested both on synthetic data and on a real dataset regarding electroencephalographic traces.

Keywords: Blind Source Separation, Dimensional reduction, Electroencephalography, Independent Component Analysis, Multi-resolution analysis, Treelets

☆Paolo Zanini (corresponding author) – Postal address: MOX – Department of Mathematics, Politecnico di Milano, piazza Leonardo da Vinci 32, 20133, Milano, Italy; email address: paolo.zanini@polimi.it; tel: +39 02 2399 4604.

1. Introduction

The statistical analysis of high-dimensional and complex data often requires the solution of two related issues: a data-driven dimensional reduction and a meaningful multiscale approximation. We look for a basis generating a space of small dimension where to represent data. We long for basis elements which are representative of the significant features of the phenomenon under study; some of these may involve a great number of the primitive variables describing the data set while others may be restricted only to a few. Hence a multi-resolution analysis is desirable. In this paper we propose a new method for the construction of a multi-scale non-orthogonal data-driven basis.

We frame the subject as a Blind Source Separation problem (BSS) [3]. Let $\mathbf{X} \in \mathbb{R}^p$ be a random vector and assume the existence of a vector $\mathbf{S} \in \mathbb{R}^K$ representing $K \leq p$ latent random sources and such that

$$\mathbf{X} = C\mathbf{S}, \tag{1}$$

where C is an unknown $p \times K$ matrix of real numbers whose columns constitute a basis of a K-dimensional subspace of $\mathbb{R}^p$. If the rows of the $n \times p$ matrix $\mathbb{X}$ collect n observed realizations $\mathbf{x}_1, ..., \mathbf{x}_n \in \mathbb{R}^p$ of the random vector $\mathbf{X}$ while the rows of the $n \times K$ matrix $\mathbb{S}$ represent the corresponding unobserved realizations of the latent random vector $\mathbf{S}$, model (1) implies that

$$\mathbb{X} = \mathbb{S}C^T. \tag{2}$$

A BSS problem consists in estimating C and $\mathbb{S}$, given $\mathbb{X}$. In this paper we analyze the EEG traces of patients affected by alcoholism and we consider a single patient. The brain signals from p electrodes are recorded at n instants of time in the matrix $\mathbb{X}$. Aim of the analysis is to decompose these signals in a linear combination between K reference brain maps (i.e., the columns of C) and its related temporal activation profiles (i.e., columns of $\mathbb{S}$), focusing in particular on the spatial maps. Hence this problem fits in the BSS framework.

Many approaches are commonly used to solve a BSS problem. The most common is Principal Component Analysis (PCA). PCA is a powerful method to find optimal subspaces where to represent data, but it presents some drawbacks. First, PCA yields an orthonormal basis; in many circumstances orthogonality is a desirable property but in some it introduces an artificial constraint not related to the phenomenological characteristics of the analyzed problem. Indeed basis elements provided by PCA might not represent

physical features of the phenomenon under study. Moreover PCA is a global method not suitable for multi-resolution analysis since each basis element most often results in a linear combination of all the primitive variables. Independent Component Analysis (ICA) [7] solves a BSS problem and provides a non-orthogonal basis for data representation. It is widely used in a huge kind of differente application as, for instance, biomedical signals analysis, hyperspectral imaging, astronomy etc... [2, 4, 9, 13]. The ICA model assumes independence between the random sources which are components of the vector $\mathbf{S}$ and produces a non orthogonal basis - an estimate of the columns of the matrix C in (2) - such that the data scores on the basis elements - estimates of the columns of $\mathbb{S}$ - are as much independent as possible. Like PCA, ICA is a global method not suitable for multi-scale analysis. If we consider the EEG problem, multi-resolution of the spatial brain maps is an interest property. Indeed, some brain activities could involve the whole brain, while others only a localized part of it. In this sense, wavelets are commonly used [11, 10] to generate a localized and multi-scale basis for data representation. Their main limitation is that the wavelet basis is not data-driven, since basis elements are fixed, regardless of the data. The Treelets algorithm is an efficient and recent approach that avoids this problem [8]. The Treelets algorithm generates a multi-scale orthonormal data-driven basis yielding a hierarchical tree that, at each level, represents data through an orthonormal basis. Thus the problem of interpretability of basis elements due to the exogenously imposed constraint of orthogonality still holds. We here propose a new approach able to provide a multi-scale non orthogonal data-driven basis through the integration between ICA and Treelets: we call it Hierarchical Independent Component Analysis (HICA).

The paper is organized as follows. In section 2, we briefly describe Independent Component Analysis and the Treelets algorithm in order to introduce HICA in the second part of the section. In section 3, we consider a procedure for data dimensional reduction with a non-orthogonal basis that will be used in HICA. Then, in section 4, we present some theoretical properties of the HICA method. In section 5, we show some simulations which validate the algorithm proposed. Finally, in section 6, we present a case study of EEG traces. All the simulations and the analyses of real data are carried out using `R` statistical software [15]. Furthermore we developed a `R` package implementing HICA algorithm, named `fastHICA` and available on the CRAN website [16].

2. Hierarchical Independent Component Analysis

In the first part of this section we describe the main ideas concerning ICA and Treelets, since HICA is obtained by integrating these two approaches. Then we introduce the HICA algorithm.

Independent Component Analysis is a method commonly used to solve Blind Source Separation problems. Consider model (1) and assume $K = p$. Given the data matrix $\mathbb{X}$, ICA looks for estimates of the basis matrix C and of the source matrix $\mathbb{S}$ in model (2), such that the columns of $\mathbb{S}$ could be taken as samples of the independent components of $\mathbf{S}$.

The ICA model presents two ambiguities. The first is label switching. The second is due to the fact that the independent components $S_1, ..., S_K$ of the vector $\mathbf{S}$ (i.e., the sources) are identifiable only up to multiplicative constants. Hence, for identifiability, the variances of the independent components are usually constrained to be 1; without loss of generality, we also assume that both the vector $\mathbf{X}$ and the vector $\mathbf{S}$ have zero mean. Moreover it is common to preprocess data by whitening $\mathbf{X}$ through a transformation matrix D. The covariance matrix of the transformed vector $\mathbf{Z} = D\mathbf{X}$ is required to be the identity, i.e., $E[\mathbf{Z}\mathbf{Z}'] = I$; for instance, $\mathbf{Z}$ is found by standardizing the principal components of $\mathbf{X}$. Therefore model (1) becomes $\mathbf{Z} = (DC)\mathbf{S}$. Since $E[\mathbf{S}\mathbf{S}'] = I$, one then derives

$$I = E[\mathbf{Z}\mathbf{Z}'] = E[DC\mathbf{S}\mathbf{S}'C'D'] = DCE[\mathbf{S}\mathbf{S}']C'D' = (DC)(DC)'.$$

Hence $C^* = DC$ is orthogonal. Once the optimal rotation C^* has been found, C is obtained as $D^{-1}C^*$.

Existence of a basis for data representation through independent components is not guaranteed (differently from a representation through uncorrelated components which always exists, and it is found by PCA). In practical problems, the estimate of the matrix C^* is obtained through the minimization of the empirical dependence between the columns of $\mathbb{S}$. In [6], it is shown that C^* can be found by maximizing the non-gaussianity of the sources $S_1, ..., S_K$. This simplifies the ICA optimization problem and suggests some suitable numerical algorithm for its solution. In this paper all analyses will be carried out with the fastICA algorithm, which maximizes a non-gaussianity measure (e.g., the absolute value of the kurtosis) through a fast fixed-point procedure. Details about the fastICA algorithm are presented in [6].

By comparing the ICA solution with that provided by PCA, we note that while PCA yields a basis whose elements are conveniently arranged

for dimensional reduction, this is not so for ICA which is useless for this purpose. A common approach to circumvent this difficulty, and to allow for the number K of independent components to be much smaller than the number p of primitive variables, is to first project data into the K-dimensional space generated by the first K principal directions. Then, ICA is carried out in this reduced K-dimensional space.

The Treelets algorithm generates a multi-resolution orthonormal basis for data representation, like wavelets, but the basis is data-driven. The Treelets algorithm yields a hierarchical tree that at each level, between the group of active variables given by the previous level, replaces the two more correlated variables through a pair-wise Principal Component decomposition. The procedure consists of an iterative algorithm with $p - 1$ steps. At each step three operations are performed:

1. compute the correlations between couples of variables, between the active variables provided by the previous step, and search for the two variables with the highest correlation;

2. compute a Principal Component Analysis in the space of the two selected variables;

3. store the second principal direction - that will not be processed in the following step and will become one of the treelet components - while the first principal direction replaces the two original variables in the active variables set.

At each level of aggregation $l = 0, ..., p - 1$, the algorithm provides a multi-resolution data-driven orthogonal basis $B^{(l)}$, able to catch internal structural features of the data.

The two methods presented above are useful to reduce the complexity of high-dimensional problems and to detect relevant features of the data. However some problems still hold. ICA, as PCA, is a global method that produces a non-sparse basis. Hence it is not suitable for a multi-resolution analysis. Treelets provide a multi-resolution but orthonormal basis, whose elements can be unrelated to the phenomenological characteristics of the problem under study. Hierarchical Independent Component Analysis, instead, aims at the construction of a multi-resolution non orthogonal data-driven basis through the integration between ICA and Treelets with the idea

of inheriting the returns of both ICA and Treelets over PCA. Basically it consists in replacing in the Treelet algorithm the pair-wise Principal Component Analysis step with a pair-wise Independent Component Analysis step. With respect to this manuscript wording, we should indeed refer to Treelet analysis as Hierarchical Principal Component Analysis (HPCA). Anyhow we preferred to keep the authors' original wording (i.e., Treelets).

A more detailed description of the HICA algorithm is now in order. First we need to define a suitable similarity measure between two random variables. According to the ICA procedure, we search for a measure that is greater when the dependence between two variables is larger. In particular we consider the distance correlation, a measure of dependence introduced in [19], and based on the distance covariance. Let X_1 and X_2 be two random variables and let $\phi_{X_1}(t)$ and $\phi_{X_2}(s)$ be their characteristic functions, while $\phi_{(X_1,X_2)}(t,s)$ is the characteristic function of the random vector $(X_1,\, X_2)'$. Then, the distance covariance between X_1 and X_2 is the non-negative number $\mathcal{V}(X_1, X_2)$ defined as

$$\mathcal{V}(X_1, X_2) = \left(\frac{1}{c^2} \int_{\mathbb{R}^2} \frac{|\phi_{(X_1,X_2)}(t,s) - \phi_{X_1}(t)\phi_{X_2}(s)|^2}{t^2 s^2} dtds \right)^{\frac{1}{2}},$$

where $c = \frac{\pi}{\Gamma(1)}$ and $\Gamma(\cdot)$ is the complete gamma function. If we indicate with $\mathcal{V}(X_1) = \mathcal{V}(X_1, X_1)$, the distance correlation between two random variables X_1 and X_2 is defined as

$$\mathcal{R}(X_1, X_2) = \frac{\mathcal{V}(X_1, X_2)}{\sqrt{\mathcal{V}(X_1)\mathcal{V}(X_2)}}.$$

Note $0 \leq \mathcal{R}(X_1, X_2) \leq 1$ and $\mathcal{R}(X_1, X_2)$ can be considered to be a measure of dependence between X_1 and X_2 in the sense that $\mathcal{R}(X_1, X_2)$ is equal to 0 if and only if X_1 and X_2 are independent random variables. Moreover distance variance and distance covariance have some properties that will be used in the following. In particular:

1. if X_1 and X_2 are independent random variables, then $\mathcal{V}(X_1 + X_2) \leq \mathcal{V}(X_1) + \mathcal{V}(X_2)$;
2. if $(X_{11}, X_{21})'$ and $(X_{12}, X_{22})'$ are independent random vectors, then $\mathcal{V}(X_{11} + X_{12}, X_{21} + X_{22}) \leq \mathcal{V}(X_{11}, X_{21}) + \mathcal{V}(X_{12}, X_{22})$.

We now describe the HICA algorithm. At level $l = 0$ of the hierarchical tree each component $X_1, ..., X_p$ of the random vector $\mathbf{X}$ is represented by itself, the basis matrix $B^{(0)}$ is indeed the canonical basis of dimension p and

the coordinates vector $\mathbf{Y}^{(0)} = (Y_1^{(0)}, ..., Y_p^{(0)})'$ corresponds to the primitive variables (i.e., $Y_i^{(0)} = X_i$). Define $\mathfrak{A}$ to be a set of indices of the active variables, initializing $\mathfrak{A}^{(0)} = \{1, ..., p\}$, and compute the sample similarity matrix $\widehat{R}^{(0)}$, where $\widehat{R}_{ij}^{(0)} = \mathcal{R}(Y_i^{(0)}, Y_j^{(0)})$. Then, for $l = 1, ..., p-1$, repeat the following three steps:

1. find the two most similar variables. In particular set:

$$(\alpha, \beta) = \arg \max_{i<j \in \mathfrak{A}^{(l-1)}} \widehat{R}_{ij}^{(l-1)};$$

2. compute an Independent Component Analysis of the variables $Y_\alpha^{(l-1)}$ and $Y_\beta^{(l-1)}$:

$$\begin{aligned} Y_\alpha^{(l-1)} &= c_{11}^{(l)} S_1 + c_{12}^{(l)} S_2, \\ Y_\beta^{(l-1)} &= c_{21}^{(l)} S_1 + c_{22}^{(l)} S_2. \end{aligned} \tag{3}$$

   The idea is to replace $Y_\alpha^{(l-1)}$ with S_1 and $Y_\beta^{(l-1)}$ with S_2. Hence define the matrix

$$\widetilde{C}^{(l)} = \begin{bmatrix} 1 & \cdots & 0 & \cdots & 0 & \cdots & 0 \\ \vdots & \ddots & \vdots & & \vdots & & \vdots \\ 0 & \cdots & \widetilde{c}_{11}^{(l)} & \cdots & \widetilde{c}_{12}^{(l)} & \cdots & 0 \\ \vdots & & \vdots & \ddots & \vdots & & \vdots \\ 0 & \cdots & \widetilde{c}_{21}^{(l)} & \cdots & \widetilde{c}_{22}^{(l)} & \cdots & 0 \\ \vdots & & \vdots & & \vdots & \ddots & \vdots \\ 0 & \cdots & 0 & \cdots & 0 & \cdots & 1 \end{bmatrix},$$

   where $\widetilde{c}_{11}^{(l)}$ and $\widetilde{c}_{22}^{(l)}$ are, respectively, in position (α, α) and (β, β). The elements $\widetilde{c}_{ij}^{(l)}$ correspond to the $c_{ij}^{(l)}$ in (3), normalized such that $\widetilde{C}^{(l)}$ has columns with unitary norm. $\widetilde{C}^{(l)}$ represents the non orthogonal transformation identified by ICA. The new basis matrix and coordinates vector become $B^{(l)} = B^{(l-1)}\widetilde{C}^{(l)}$ and $\mathbf{Y}^{(l)} = (\widetilde{C}^{(l)})^{-1}\mathbf{Y}^{(l-1)}$, respectively. The similarity matrix $\widehat{R}^{(l)}$ is then updated accordingly;

3. order the two new variables according to their variances. If the variance of $Y_\alpha^{(l)}$ is greater than the variance of $Y_\beta^{(l)}$, store the variable $Y_\beta^{(l)}$ and,

at the next step, consider only $Y_\alpha^{(l)}$ as a possible candidate for a new aggregation. This corresponds to remove the index β from the set $\mathfrak{A}$ of the active variables, defining $\mathfrak{A}^{(l)} = \mathfrak{A}^{(l-1)} \setminus \{\beta\}$. Otherwise store $Y_\alpha^{(l)}$ and set $\mathfrak{A}^{(l)} = \mathfrak{A}^{(l-1)} \setminus \{\alpha\}$.

The algorithm provides, at each level of aggregation l, a non orthogonal basis matrix $B^{(l)} = B^{(0)}\widetilde{C}^{(1)} \cdots \widetilde{C}^{(l)}$ - an estimate of the basis matrix C - and a coordinates vector $\mathbf{Y}^{(l)} = \widetilde{C}^{(l)^{-1}} \cdots \widetilde{C}^{(1)^{-1}}\mathbf{Y}^{(0)}$, which is an estimate of the scores matrix $\mathbb{S}$.

3. Selection of the level of the tree and dimensional reduction with a non-orthogonal basis

The HICA algorithm generates p different matrices $B^{(0)}, ..., B^{(p-1)}$ as estimates of the basis matrix C. Obviously one cannot take into account all these different estimates, but it is reasonable to choose only one (or some) of them for the analysis. The more natural choice is to consider the estimate related to the maximum height of the tree, $l = p - 1$, but alternatively one can choose any of the basis given at the different levels l. At a generic level l, $B^{(l)}$ is composed by the l elements stored in the previous steps and the $p - l$ elements corresponding to variables of the active set $\mathfrak{A}^{(l)}$ that would be ready for aggregation in the following steps. Let A^l be a partition of $\{1, ..., p\}$ in $p - l$ sets named A_i^l, with $i = 1, ..., p - l$. By construction each basis element of $B^{(l)}$ is defined on a different set A_i^l of the partition (i.e., the positions of the non-zero values of each basis element correspond to the indexes of one of the set A_i^l). Since at each level a new variable is generated as a linear combination of two variables of the active set, the number of sets that form the partition is reduced by aggregating two of them. Hence at a specific level l the basis elements stored in the previous steps of the algorithm are defined on subsets of the A_i^l. Therefore we can divide basis elements of $B^{(l)}$ into $p-l$ different groups, according to the $p-l$ different sets of the partition. For this reason we can relate the different basis $B^{(l)}$ to different degrees of sparsity, where the different degrees refer to the different cardinalities of partitions. In particular, the lower is the level l considered, the greater is the degree of sparsity of the basis taken into account (i.e., greater is the cardinality of the partition).

Once a specific basis $B^{(l)}$ is chosen, another important aspect to consider is dimensional reduction. In particular we need to select the dimension K

(with $K \leq p$) of a suitable subspace to represent data, choosing only K basis elements.

To jointly face these two problems (i.e., the choice of the degree of sparsity and the K "best" basis elements) we consider the energy, an index related to the fraction of variance explained by a basis. We now first describe the energy index, focusing on the non trivial case of its evaluation for a non-orthogonal basis. Then we propose a strategy to choose a suitable dimension K to represent data in a reduced space and, given K, we show how to select a specific basis $B^{(l)}$ and its K basis elements.

Consider a basis $C = [\mathbf{c}_1; ...; \mathbf{c}_p]$, not necessarily orthogonal. Let $\mathcal{I}_K = \{i_1, i_2, ..., i_K\}$ be one of the $\binom{p}{K}$ subsets of the index set $\{1, ..., p\}$ with cardinality K, and let $C_{\mathcal{I}_K} = [\mathbf{c}_{i_1}; ...; \mathbf{c}_{i_K}]$. Let $\mathbf{X}^{C_{\mathcal{I}_K}} = C_{\mathcal{I}_K}(C_{\mathcal{I}_K}^T C_{\mathcal{I}_K})^{-1} C_{\mathcal{I}_K}^T \mathbf{X}$ be the orthogonal projection of $\mathbf{X}$ on the space spanned by $C_{\mathcal{I}_K}$, where $\mathbf{X} \in \mathbb{R}^p$ is a random vector with zero mean. Then we define

$$\gamma(C_{\mathcal{I}_K}) = \frac{E[\|\mathbf{X}^{C_{\mathcal{I}_K}}\|^2]}{E[\|\mathbf{X}\|^2]} = \frac{tr(\Sigma C_{\mathcal{I}_K}(C_{\mathcal{I}_K}^T C_{\mathcal{I}_K})^{-1} C_{\mathcal{I}_K}^T)}{tr(\Sigma)},$$

being $\Sigma = E[\mathbf{X}\mathbf{X}^T] = Cov(\mathbf{X})$, and we call $\gamma(C_{\mathcal{I}_K})$ the energy associated to the basis $C_{\mathcal{I}_K}$. At this point we define $\Gamma_K(C)$ as the maximum energy among all the $\binom{p}{K}$ energies associated to the K-dimensional subspaces spanned by all possible subsets of cardinality K of the basis matrix C:

$$\Gamma_K(C) = \max_{\mathcal{I}_K \subseteq \{1,...,p\}} \gamma(C_{\mathcal{I}_K}). \tag{4}$$

If C is non orthogonal the evaluation of $\Gamma_K(C)$ may become cumbersome. The non orthogonality, in fact, implies that the elements of the best $K-1$-dimensional space are not necessary a subset of the elements of the best K-dimensional space. Hence we need to search for the optimal basis between al the possible $\binom{p}{K}$ combinations. This can be done if $\binom{p}{K}$ is small, but when it increases the computation may become unfeasible. Therefore a selection strategy is needed. This is true not only for HICA, but whenever dealing with non orthogonal basis. For instance we present a forward selection strategy, that is one the most used in the literature. It can be easily computed and, in practical problems, produces reasonable approximations of the K-dimensional subspace with maximal energy $\Gamma_K(C)$. However any other selection strategy can be used. Considering a given K, for forward selection we start by calculating the energy $\gamma([\mathbf{c}_k])$ for each basis element and

we set the maximum energy element as the first element of the basis. Let it be $\mathbf{c}_{(1)}$. Then we look for the second basis element, named $\mathbf{c}_{(2)}$, such that

$$\mathbf{c}_{(2)} = \arg \max_{\mathbf{c}_j \neq \mathbf{c}_{(1)}} \gamma([\mathbf{c}_{(1)}; \mathbf{c}_j]).$$

Once $\mathbf{c}_{(1)}, ..., \mathbf{c}_{(k)}$ have been identified, $\mathbf{c}_{(k+1)}$ is found accordingly:

$$\mathbf{c}_{(k+1)} = \arg \max_{\mathbf{c}_j \neq \mathbf{c}_{(1)}, ..., \mathbf{c}_{(k)}} \gamma([\mathbf{c}_{(1)}; ...; \mathbf{c}_{(k)}; \mathbf{c}_j]),$$

and the procedure continues until $\mathbf{c}_{(K)}$ is found.

Remark 1. If C is an orthonormal matrix, the exact solution of the optimization problem (4) can be found efficiently since we do not need to evaluate all the $\binom{p}{K}$ energies $\gamma(C_{\mathcal{I}_K})$. Indeed, let $W = [\mathbf{w}_1; ...; \mathbf{w}_p]$ be an orthonormal basis, $\Gamma_K(W)$ is found by computing, for $j = 1, ..., p$,

$$\gamma([\mathbf{w}_j]) = \frac{E[(\mathbf{w}_j^T \mathbf{X})^2]}{E[\|\mathbf{X}\|^2]} = \frac{\mathbf{w}_j^T \Sigma \mathbf{w}_j}{tr(\Sigma)} = \frac{\sum_{i=1}^p \lambda_i (\mathbf{w}_j^T \mathbf{e}_i)^2}{\sum_{i=1}^p \lambda_i},$$

where λ_i and $\mathbf{e}_i$ are the eigenvalues and the eigenvectors of Σ. After sorting the basis elements according to their energy, such that $\gamma([\mathbf{w}_{(1)}]) \geq \gamma([\mathbf{w}_{(2)}]) \geq \cdots \geq \gamma([\mathbf{w}_{(p)}])$, $\Gamma_K(W)$ is obtained by summing the first K energy terms. In particular:

$$\Gamma_K(W) = \frac{tr(\Sigma W_K W_K^T)}{tr(\Sigma)} = \sum_{k=1}^{K} \gamma([\mathbf{w}_{(k)}]),$$

where $W_K = [\mathbf{w}_{(1)}; ...; \mathbf{w}_{(K)}]$. This is the same procedure adopted in [8], for finding the elements of the K-dimensional basis and also coincides with the criterium used in PCA to order the principal directions. Indeed, if $E = [\mathbf{e}_1; ...; \mathbf{e}_p]$ is the matrix whose columns are the eigenvectors of Σ, $\gamma([\mathbf{e}_j]) = \frac{\sum_{i=1}^p \lambda_i (\mathbf{e}_j^T \mathbf{e}_i)^2}{\sum_{i=1}^p \lambda_i} = \frac{\lambda_j}{\sum_{i=1}^p \lambda_i}$ and $\Gamma_K(E) = \sum_{i=1}^K \gamma([\mathbf{e}_{(k)}]) = \frac{\sum_{k=1}^K \lambda_{(k)}}{\sum_{i=1}^p \lambda_i}$.

We now focus on the energy index as a tool to perform dimensional reduction and to find the best basis between the p estimates provided by HICA. The choice of K is not related to this specific algorithm. We can decide on the best value for K considering only the maximum height tree basis (i.e., considering only $\Gamma_K(B^{(p-1)})$), or also through other dimension reduction method. Once K has been determined, we compute $\Gamma_K(B^{(l)})$, for $l = 0, ..., p-1$, and

we choose the best basis B_{best} according the same criterium adopted in [8], for Treelets:

$$B_{best} = \arg \max_{B_l : 0 \leq l \leq p-1} \Gamma_K(B_l).$$

This argmax is not necessarily unique. Indeed, at a specific level, say $l = p - k$, we have k elements (corresponding to the variables in the active set $\mathfrak{A}^{(p-k)}$) that in the following steps are merged together. It is straightforward to show that, if the best k-dimensional space was generated by these k elements, the quantity $\Gamma_k(B^{(p-k)})$ would not increase in the next levels, since, even if two of these elements are merged together, the space spanned by the new elements is the same. In general at level $p - k$ the best k-dimensional space need not be generated by the k active variables. However from the level when all variables of the active set $\mathfrak{A}^{(l)}$ constitute the best k-dimensional basis, the quantity $\Gamma_k(B^{(l)})$ does not increase. Hence we could have more than one basis with the same energy.

The choice suggested in [8], is to take into account the basis with the smallest l. Such proceeding could however discard solutions which are able to better catch the underlying structure of the problem. Indeed, all the basis with the same energy are, in principle, equally valid, and all basis with the same highest energy Γ_K should be considered. However, since they have different degrees of sparsity (the lower is the level, the higher is the degree of sparsity), some of them can be more preferable, but this is a case-specific choice. In the examples of section 5 and in the real case study of section 6, we will deepen the analysis of this issue.

4. Theoretical results

In this section we analyze the consistency of HICA when data are generated by K independent groups of sources with disjoint supports plus some noise. The consistency of treelets for K uncorrelated groups has been proved in [8]. Here we want to show that the hierarchical nature of HICA is able to highlight possible grouping structures of the p original variables, where the structure is defined in terms of dependence. Specifically we consider a situation where the p primitive variables are divided into K groups, with dependent variables within groups and weakly dependent variables between groups. We want to show that HICA is well suited for representing and catching the underlying structure of this kind of data, providing at level $p - K$ loading vectors whose supports are defined on the different groups. We show

this result in Lemma 2, and Theorem 3, after the discussion of a preliminary property in Lemma 1. The proofs are provided in Appendix A.

We start by dealing with an issue directly connected to the fact that the fastICA algorithm is grounded on non-gaussianity measures. In some special situations the directions maximizing kurtosis, a well-known non-gaussianity measure, can be found analytically, as it is proved in the following Lemma.

Lemma 1. *Let T be a random variable such that $kurt(T) \neq 0$ and let E be a gaussian random variable. Set $\boldsymbol{Z} = (T, E)'$ and assume that T and E are independent. Let $\boldsymbol{w} = (w_1, w_2)'$ be a vector of unitary norm. The absolute value of the kurtosis of the random variable $\boldsymbol{w}'\boldsymbol{Z}$ is maximized by $\boldsymbol{w}_{max} = (1, 0)'$.*

We now deal with p non-gaussian random variables identical but for an additive gaussian noise, in order to show that, in this particular case, HICA provides a constant loading vector at the final level $l = p - 1$, thus gathering the common component.

Lemma 2. *Let T be a random variable with 0 mean, $kurt(T) \neq 0$ and such that $\mathcal{V}(T) = 1$. Let $\boldsymbol{X} = (X_1, ..., X_p)' \in \mathbb{R}^p$ be a random vector such that, for $\sigma^2, \sigma_e^2 > 0$,*

$$X_i = \sigma^2 T + \sigma_e^2 E_i, \qquad i = 1, ..., p,$$

with E_i a random gaussian noise such that, for $i, j = 1, ...p$, $i \neq j$, $\mathcal{V}(E_i) = 1$, $\mathcal{V}(E_i, E_j) = 0$ and $\mathcal{V}(E_i, T) = 0$. At each level $1 \leq l \leq p - 1$ the HICA decomposition reads:

$$B^{(l)} = [\boldsymbol{c}_1^{(l)}; ...; \boldsymbol{c}_{p-l}^{(l)}; \widetilde{\boldsymbol{c}}_1^{(l)}; ...; \widetilde{\boldsymbol{c}}_l^{(l)}],$$
$$\boldsymbol{Y}^{(l)} = (Y_1^{(l)}, ..., Y_{p-l}^{(l)}, \widetilde{Y}_1^{(l)}, ..., \widetilde{Y}_l^{(l)})',$$

where $\boldsymbol{c}_i^{(l)} = \frac{1}{\sqrt{|A_i^l|}} I_{A_i^l}$ and $Y_i^{(l)} = \frac{|A_i^l|}{\sqrt{|A_i^l|}} \sigma^2 T + \frac{\sigma_e^2}{\sqrt{|A_i^l|}} E_i^{(l)}$, with $\mathcal{V}(E_i^{(l)}) \leq |A_i^l|$ and $\mathcal{V}(Y_i^{(l)}) \leq \sqrt{|A_i^l|}(\sigma^2 + \sigma_e^2)$ $\forall i = 1, ..., p - l$ (the sets A_i^l have been defined in Section 3, and $I_{A_i^l}$ is a vector with ones for the elements of the set A_i^l and 0 otherwise). In particular, at the level $l = p - 1$, $\boldsymbol{c}_1^{(p-1)} = (\frac{1}{\sqrt{p}}, ..., \frac{1}{\sqrt{p}})'$ and $Y_1^{(p-1)} = \frac{p}{\sqrt{p}} \sigma^2 T + \frac{\sigma_e^2}{\sqrt{p}} E_1^{(p-1)}$, with $\mathcal{V}(E_1^{(p-1)}) \leq p$ and $\mathcal{V}(Y_1^{(p-1)}) \leq \sqrt{p}(\sigma^2 + \sigma_e^2)$.

Lemma 2, is instrumental for proving Theorem 3, the main theoretical result of the paper.

Theorem 3. *Let $T_1, ..., T_K$ be random variables with 0 mean, non zero kurtosis and such that $\mathcal{V}(T_k) = 1$, $k = 1, ..., K$. Let $\boldsymbol{X} = (X_{11}, ..., X_{1p_1}, ..., X_{K1}, ..., X_{Kp_K})' \in \mathbb{R}^p$ be a random vector such that, for $\sigma_1^2, ..., \sigma_K^2, \sigma_e^2 > 0$,*

$$X_{ji} = \sigma_j^2 T_j + \sigma_e^2 E_{ji},$$

with E_{ji} random gaussian noise such that $\mathcal{V}(E_{ji}) = 1$, $\mathcal{V}(E_{ji}, E_{hl}) = 0$ and $\mathcal{V}(E_{ji}, T_h) = 0$ for $j, h = 1, ..., K$, $i = 1, ..., p_j$ and $l = 1, ..., p_h$. Furthermore set
$\mathcal{V}(\sigma_j^2 T_j, \sigma_h^2 T_h) = \sigma_{jh}$ and assume that

$$\max_{1 \leq j,h \leq K} \left(\frac{\sigma_{jh}}{\sigma_j \sigma_h} \right) < \frac{c(\sigma_e)}{1 + \delta^2}, \tag{5}$$

with $\delta = \frac{\sigma_e}{\min_{1 \leq j \leq K} \sigma_j}$ and $c(\sigma_e)$ a constant such that $0 < c(\sigma_e) \leq 1$ and $c(\sigma_e) \overset{\sigma_e \to 0}{\longrightarrow} 1$. Then, at level $l = p - K$, the HICA decomposition reads:

$$\begin{aligned} B^{(p-K)} &= [\boldsymbol{c}_1^{(p-K)}; ...; \boldsymbol{c}_K^{(p-K)}; \widetilde{\boldsymbol{c}}_1^{(p-K)}; ...; \widetilde{\boldsymbol{c}}_{p-K}^{(p-K)}], \\ \boldsymbol{Y}^{(p-K)} &= (Y_1^{(p-K)}, ..., Y_K^{(p-K)}, \widetilde{Y}_1^{(p-K)}, ..., \widetilde{Y}_{p-K}^{(p-K)})', \end{aligned}$$

where $\boldsymbol{c}_i^{(p-K)} = \frac{1}{\sqrt{|F_i|}} I_{F_i}$ and $Y_i^{(p-K)} = \frac{|F_i|}{\sqrt{|F_i|}} \sigma_i^2 T_i + \frac{\sigma_e^2}{\sqrt{|F_i|}} E_i^{(p-K)}$, with $\mathcal{V}(E_i^{(p-K)}) \leq |F_i|$, $\mathcal{V}(Y_i^{(p-K)}) \leq \sqrt{|F_i|}(\sigma_i^2 + \sigma_e^2)$ and $F_i = \{i1, ..., ip_i\}$, for $i = 1, ..., K$.

This results states that if variables are dependent according to an approximate block structure where variables in the same block are exchangeable and strongly dependent while variables in different blocks are weakly dependent, then HICA is able to uncover this feature providing loading vectors constants on each block and null elsewhere.

Remark 2. As pointed out by one of the reviewers, these theoretical results do not deal with more complex situations where mixtures came from different sources, or when the Gaussian component is one of the latent source. Indeed, we chose to investigate this simpler setting in order to prove the consistency of HICA algorithm following the line depicted in [8]. Nevertheless, we explored the performance of HICA in more complex situations through numerical simulations presented in Section 5.

5. Comparison among PCA, ICA, Treelets, and HICA on synthetic data

In this section we will present some simulated examples to compare PCA, ICA, Treelets, and HICA performances in different scenarios. For all scenarios we consider the following latent variable model:

$$\mathbf{X} = \sum_{k=1}^{3} \mathbf{c}_k S_k + \sigma \mathbf{E} \ , \tag{6}$$

where $\mathbf{X}$ is the observed p-variate random vector, $\mathbf{c}_k$ represent the columns of the basis matrix C (i.e., the unknown basis elements), S_k are unobserved non-gaussian random variables, and $\mathbf{E}$ is a p-variate gaussian vector (with $\mathbf{0}$ mean and identity covariance matrix) acting as a noise term. Our purpose is to use PCA, ICA, Treelets, and HICA to obtain an estimate for the basis matrix C from a sample of size n drawn from model (6).

In detail, we investigate four different scenarios exploring different structures of dependence and orthogonality of the components (i.e., dependent/indepedent sources S_k and orthogonal/non-orthogonal basis elements $\mathbf{c}_k$):

Scenario A: Orthogonal and independent latent components.

Scenario B: Orthogonal and dependent latent components.

Scenario C: Non-orthogonal and independent latent components.

Scenario D: Non-orthogonal and dependent latent components.

Below, we focus on scenarios B, C, and D, respectively. Scenario A is not discussed since, as expected, all four methods are effective in estimating the model in this trivial case. In the next scenarios we always consider light tailed distributions and a value of p relatively small, in order to obtain clearer plots. Simulations with heavy tailed distributions, a larger number of variables, and more complex settings have been performed, and they are provided in the supplementary material available online.

Scenario B: Orthogonal and dependent latent components. We first consider an example similar to the one presented in [8], where $p = 10$ random variables are obtained by linear combinations of three dependent - and thus

correlated - random sources such that the basis elements $\mathbf{c}_1, \mathbf{c}_2$, and $\mathbf{c}_3$ are non-overlapping - and thus orthogonal -. In particular we set:

$$S_1 \sim U([0, b_1]) \quad \perp\!\!\!\perp \quad S_2 \sim U([0, b_2]), \quad S_3 = a_1 S_1 + a_2 S_2,$$

with $b_1 = 20$, $b_2 = 15$, $a_1 = 2$, $a_2 = 1$, and $\sigma = 1$. The basis elements $\mathbf{c}_k$ are defined on disjoint subsets, specifically:

$$\mathbf{c}_1 = (1\ 1\ 1\ 1\ 0\ 0\ 0\ 0\ 0\ 0)',$$
$$\mathbf{c}_2 = (0\ 0\ 0\ 0\ 1\ 1\ 1\ 1\ 0\ 0)',$$
$$\mathbf{c}_3 = (0\ 0\ 0\ 0\ 0\ 0\ 0\ 0\ 1\ 1)'.$$

Finally, we sample $n = 1000$ independent realizations from the model.

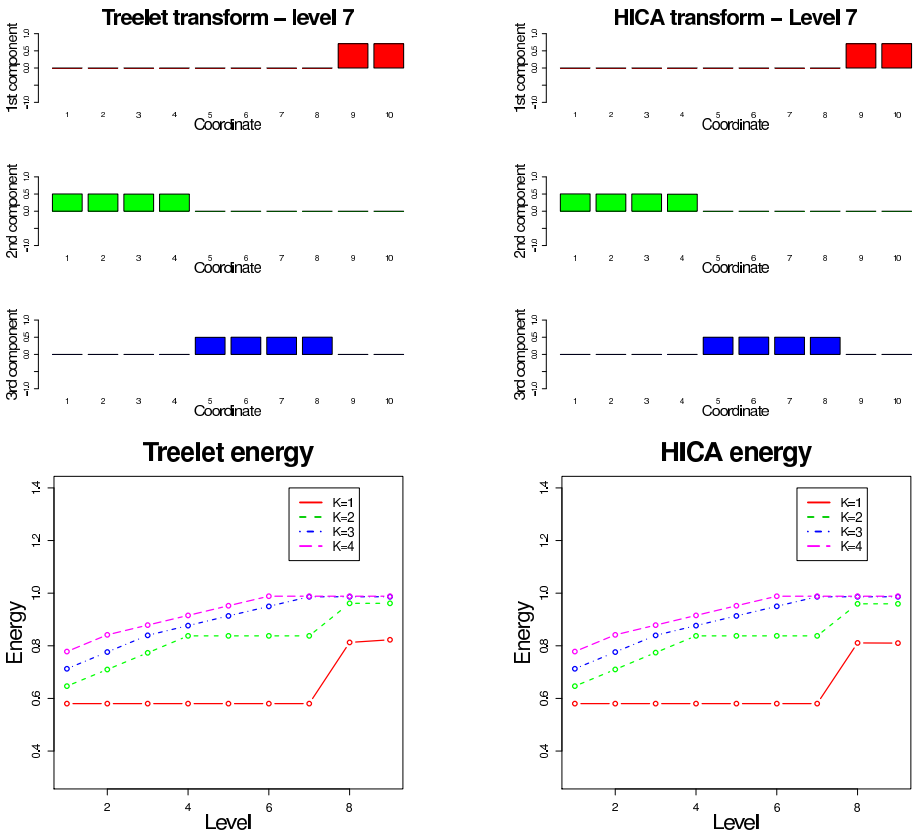

Figure 1: Scenario B: Orthogonal and dependent latent components. Top panels report the basis elements provided by Treelets (left) and HICA (right) when $K = 3$ and $l = 7$. The bottom panels report the energy as a function of l and K for Treelets decompositions (left) and HICA decompositions (right).

This is an example in which neither PCA nor ICA is expected to target the correct model being the three sources neither uncorrelated nor independent. On the contrary, both Treelets and HICA can detect the correct model if the chosen level of aggregation is $l = 7$ (i.e., 3 disjoint supports) and the chosen number of latent sources is $K = 3$. As shown in the bottom panels of Figure 1, this choice of l and K is the one suggested by the criterion presented in [8], and is among the ones suggested by the criterion suggested in section 3. This latter criterion supports indeed $K = 3$ and $l = 7, 8, 9$ as candidate values.

Scenario C: Non-orthogonal and independent latent components. The previous example presents a situation in which hierarchical methods (i.e., Treelets and ICA) can outperform non-hierarchical methods (i.e., PCA and ICA). We now consider a complementary scenario in which ICA-inspired methods (i.e., ICA and HICA) can outperform PCA-inspired methods (i.e., PCA and Treelets). In this scenario $p = 6$, the basis elements $\mathbf{c}_1$ and $\mathbf{c}_2$ are overlapping and non-orthogonal and sources S_1, S_2, and S_3 are independent. In particular:

$$S_1 \sim U([0, b_1]) \quad \perp\!\!\!\perp \quad S_2 \sim U([0, b_2]) \quad \perp\!\!\!\perp \quad S_3 \sim U([0, b_3],$$

with $b_1 = b_2 = b_3 = 20$ and $\sigma = 1$. The basis elements $\mathbf{c}_k$ are defined as follows:

$$\mathbf{c}_1 = (1\ 1\ 0\ 0\ 0\ 0)',$$
$$\mathbf{c}_2 = (1\ 1\ 1\ 1\ 0\ 0)',$$
$$\mathbf{c}_3 = (0\ 0\ 0\ 0\ 1\ 1)'.$$

Finally, we sample $n = 1000$ independent realizations from the model. Of course in this scenario, PCA and Treelets cannot target the right solution being the basis elements non-orthogonal. ICA instead targets the right solution being the sources independent. Figure 2 shows that also HICA can detect the right solution if $K = 3$ and $l = 4$ (i.e., 2 disjoint supports).

Note that the criterion proposed in [8], would have suggested $K = 3$ and $l = 3$ (i.e., 3 disjoint supports) which would have taken to a misidentification of the model for HICA as well (see top panels of Figure 2). This example confirms what suggested in our criterion: once K is chosen, all values of l providing the maximal energy are candidate values and not just the minimum one. Good representations are indeed obtained using HICA with $K = 3$ and $l = 4, 5$.

Scenario D: Non-orthogonal and dependent latent components. We finally present a situation in which HICA outperforms PCA, ICA, and Treelets. This last scenario is simply obtained by setting latent components both non-orthogonal and dependent. In this case indeed, PCA cannot target the correct model being the sources non-orthogonal and dependent, ICA cannot target the correct model being the sources dependent, Treelets cannot target the correct model being the sources non-orthogonal. HICA remains the only method having the chance to target the correct model.

We here set $p = 6$, the basis elements $\mathbf{c}_1$ and $\mathbf{c}_2$ are overlapping (and thus non-orthogonal) and the three sources S_1, S_2, and S_3 dependent. In particular:

$$S_1 \sim U([0, b_1]) \quad \perp\!\!\!\perp \quad S_2 \sim U([0, b_2]) \quad S_3 = S_1 + S_2 + U,$$

and $U \sim U([0, b_3]$, with $b_1 = b_2 = 20$, $b_3 = 1$, and $\sigma = 1$, while basis elements $\mathbf{c}_k$ are the same defined as in Scenario C.

As shown in the bottom panels of Figure 3, we can draw the same conclusions of Scenario C with respect to the choice of K and l: $K = 3$ and $l = 3, 4, 5$ are good candidate values. Once again (top panels of Figure 3) $l = 3$, the value suggested by the criterion proposed in [8], is not the best choice. Although in this case, neither HICA is able to exactly catch the right configuration, HICA with $K = 3$ and $l = 4$ of course provides the closest representation: second and third components are very well detected with some bias in the estimation of the first component.

These simulated examples suggest that when dealing non-Gaussian latent components (even non-orthogonal and/or dependent) HICA always performs better than or equally to PCA, ICA, and Treelets. Moreover, as expected by theory, they discourage the use of PCA and ICA when components are dependent and the use of PCA and Treelets when components are non-orthogonal. A summary of the "win situations" for the four methods that can be drawn from the simulations is reported in Table 1.

Simulations also show that the criterion proposed in [8], for the choice of K and l might take to a misdetection of the model. For a given value of K the more proper approach seems indeed to consider as candidate values for the level of aggregation l all values providing the maximal energy and not necessarily the minimum one.

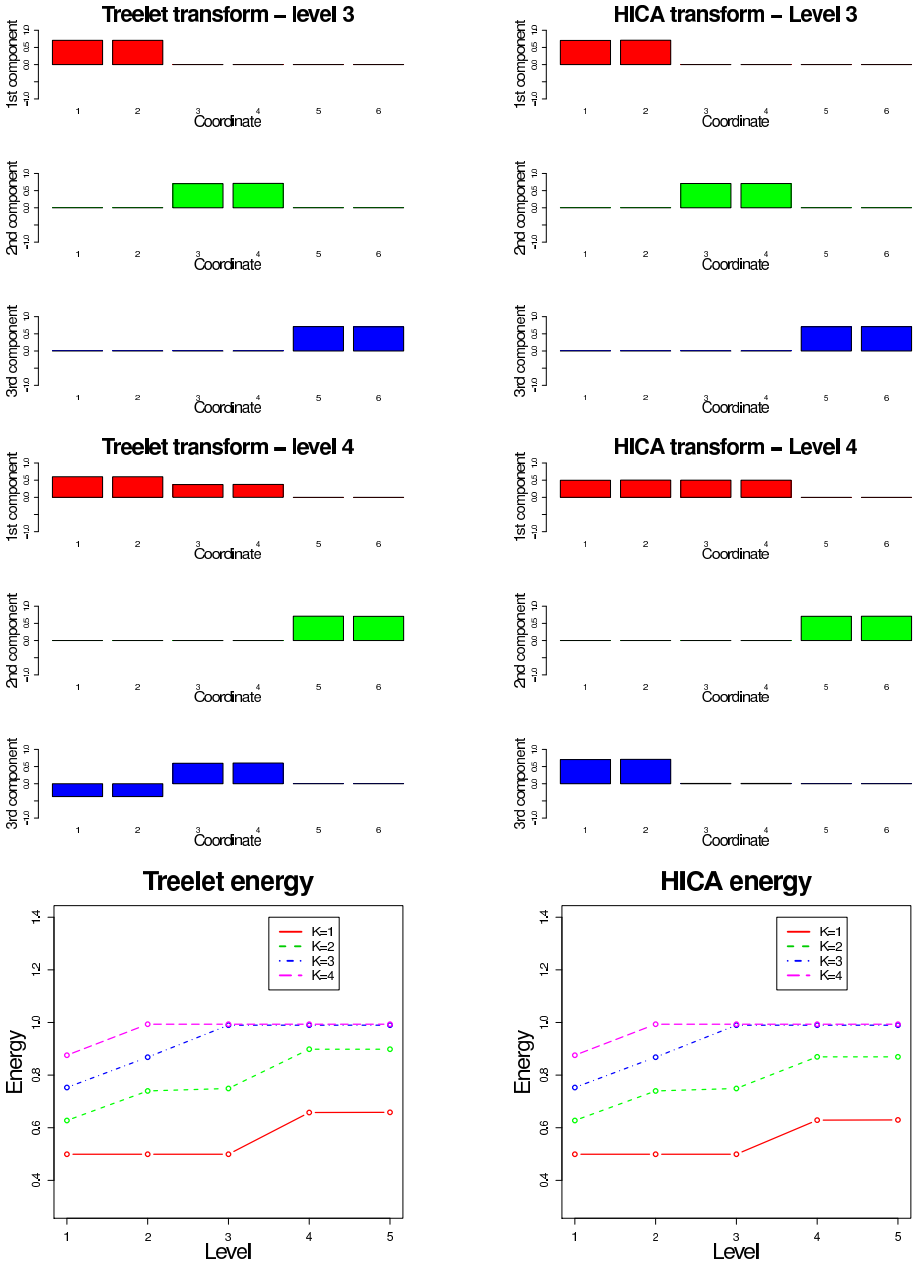

Figure 2: Scenario C: Non-orthogonal and independent latent components. Top panels report the basis elements provided by Treelets (left) and HICA (right) when $K = 3$ and $l = 3$. Middle panels report the basis elements provided by Treelets (left) and HICA (right) when $K = 3$ and $l = 4$. The bottom panels report the energy as a function of l and K for Treelets decompositions (left) and HICA decompositions (right).

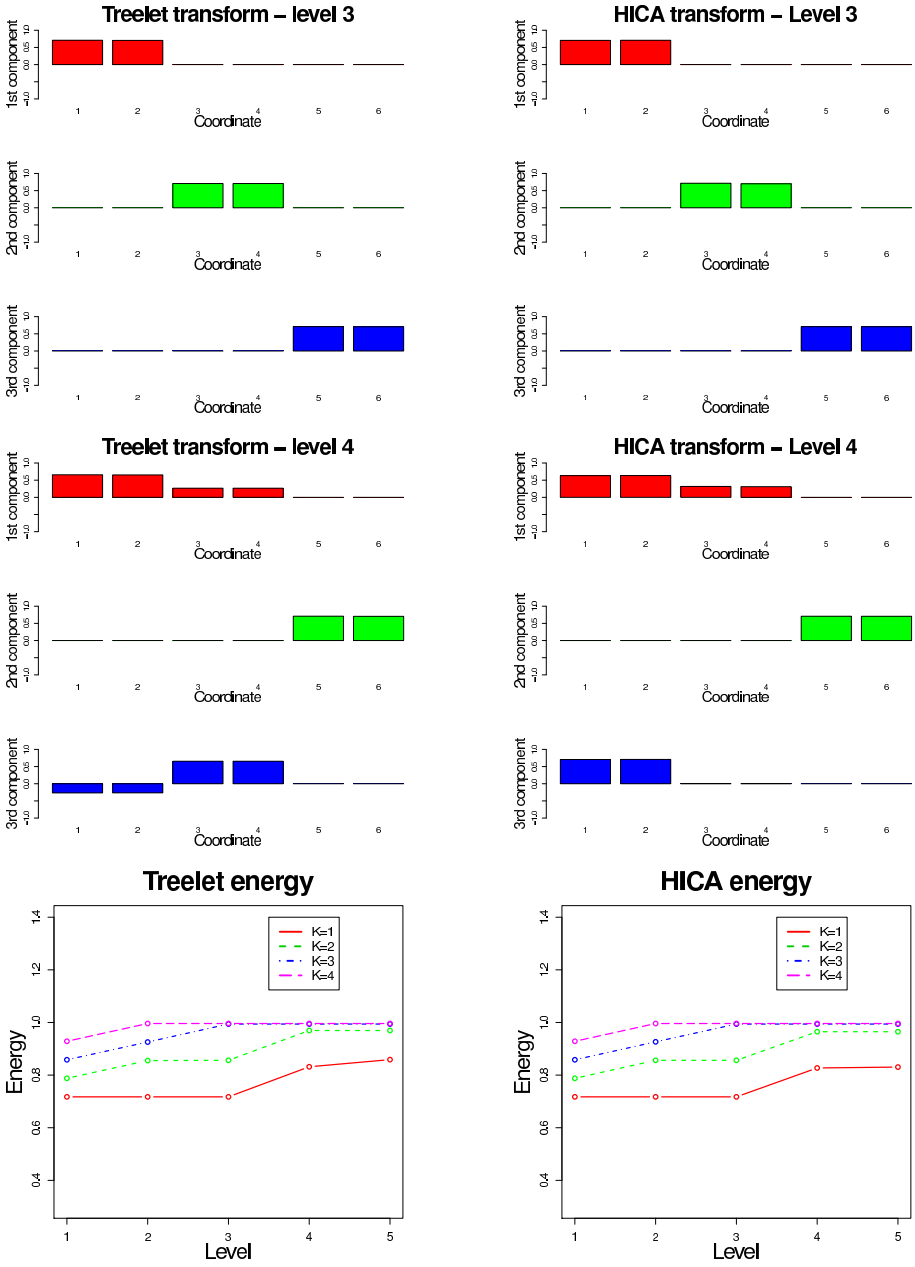

Figure 3: Scenario D: Non-orthogonal and dependent latent components. Top panels report the basis elements provided by Treelets (left) and HICA (right) when $K = 3$ and $l = 3$. Middle panels report the basis elements provided by Treelets (left) and HICA (right) when $K = 3$ and $l = 4$. The bottom panels report the energy as a function of l and K for Treelets decompositions (left) and HICA decompositions (right).

Table 1: Summary of the "win situations" for PCA, ICA, Treelets, and HICA with respect to orthogonality/non-orthogonality (O/NO) and dependence/independence (D/I) of the latent components.

	PCA	ICA	Treelets	HICA
A: O - I	win	win	win	win
B: O - D			win	win
C: NO - I		win		win
D: NO - D				win

6. Case study: Analysis of EEG signals

We now apply HICA to a BSS real data problem by analyzing EEG traces of patients affected by alcoholism. The multi-resolution and non-orthogonal properties characterizing the HICA solution, allow to obtain interpretable and meaningful results that provide noticeable improvements in terms of phenomenological interpretation.

EEG datasets are widely studied through statistical methods [12, 14]. The data analyzed in this paper are courtesy of the online UCI Machine Learning Repository [1]. For each patient in the study, measurements from 61 electrodes out of 64 placed on the scalp are available. The electrodes are located at standard sites [17, 20]. For each electrode, the recorded signal measures the electrode electric potential with respect to some reference electrode and describes the electrical activity of the brain in the neighborhood of the electrode across time. We observe these signal at $n = 256$ equally spaced instants along a time span of 1 second. This sample represents the n realizations of a random vector $\mathbf{X}$ in $\mathbb{R}^p$ with $p = 61$, that is the number of electrodes considered in the study. The analysis consist in the decomposition of the original variables through model (2). We implement the HICA algorithm to solve this BSS problem and we compare the results obtained by HICA with those provided by Treelets, ICA and PCA. As an example, in Figure 4 we show some relevant basis elements identified by these methods for one patient. The subject was exposed to two stimuli. Specifically, the patient was shown two pictures chosen from the 1980 Snodgrass and Vanderwart set presented in [18]. The two stimuli were presented in a matched condition (i.e., the subject has been asked to look at the same picture twice).

We consider $K = 5$ components, since it is sufficient to explain a great portion of variability and to show interesting and interpretable results. For

PCA we show the first 5 principal components, for ICA the results obtained with the fastICA algorithm selecting 5 sources, while for Treelets and HICA we need to select a level. For both methods the energy for $K = 5$ reaches the maximum at $l = p-6 = 55$. Hence, as discussed in Section 3, any level from $l = p - 6 = 55$ to $l = p - 1 = 60$ is a good candidate, and the choice should be problem-specific. In this case-study we aim to obtain a decomposition with basis elements defined on few electrodes, which can identify little brain regions devoted to specific features. For this reason we choose to select the level $l = 55$ and we show the 5 components found by the energy criterium. Since $l = p - 6$, we expect to find basis elements whose supports are defined on no more than six different sets of variables.

Multi-resolution methods yield localized basis elements. This is a very interesting property, since it highlights components defined on localized brain regions and allows to identify more precisely the areas involved in the task. PCA and ICA, instead, yield more general and unspecific components, possibly difficult to read. Even when they seem to catch localized information, basis elements are not so clearly defined since they involve the entire set of variables. This is apparent in the fourth row of Figure 4, where HICA and Treelets select a single electrode (i.e., a single variable). This electrode clearly represents some noise either related to facial muscles activity or due to an unexpected saturation of the electrode. The related components identified by ICA and PCA, even though highlighting the same electrode, present more complex loadings diffused on other electrodes. The first row of Figure 4 reveals very similar components for HICA and and Treelets. Both analysis identify the associative activity in the frontal brain area, that is the area which processes the information related to similarities and differences between the two pictures. This crucial component is not caught by PCA and ICA. The main difference between HICA and Treelets regards instead the second and the fifth row in Figure 4. While Treelets yield an unfocused result, with components involving all the occipital cerebral hemisphere (i.e., one component averaging over the entire occipital part and the other contrasting the right and the left activity in the occipital part), HICA splits this information in two separate parts. The HICA component shown in the second row is related to the primary visual cortex, the first area reached by visual information, which analyzes it in terms of shape and pattern recognition. Then the information flow goes to the internal area of the occipital hemisphere, which associates to the stimulus specific features like color, direction or origin. This area is identified only by HICA, specifically by the

component in the fifth row of Figure 4.

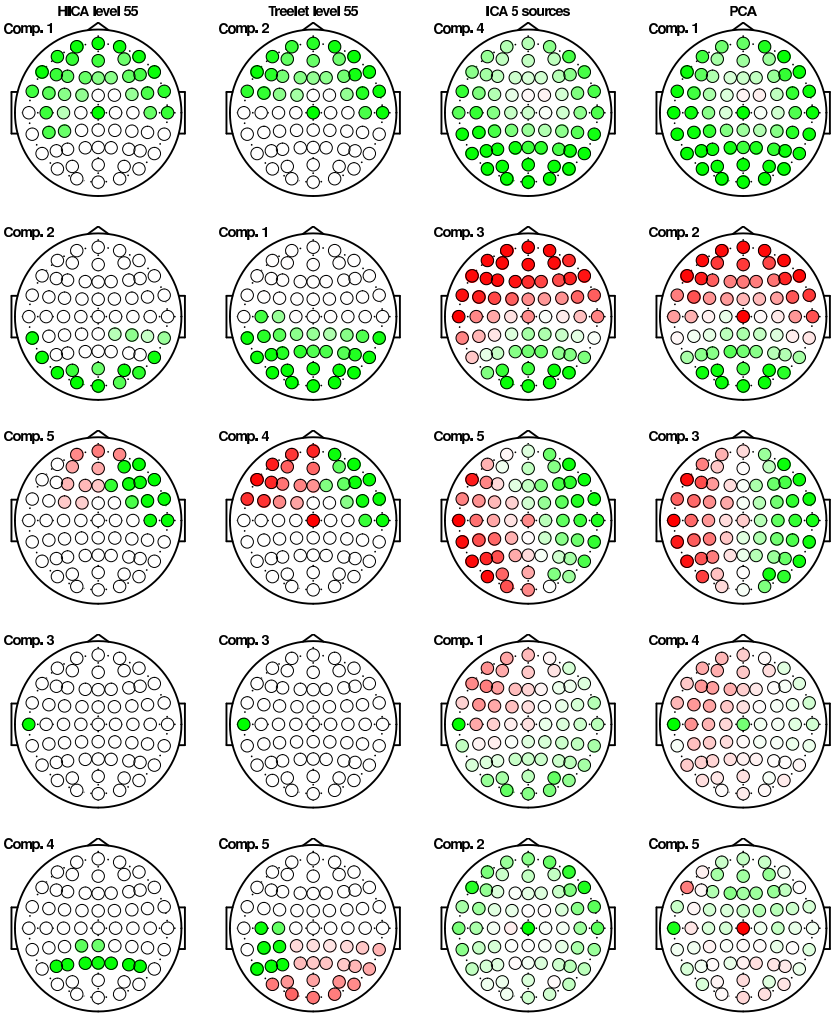

Figure 4: First five loadings found out by HICA (first column on the left), Treelets (second column), ICA (third column) and PCA (fourth column).

As suggested by a referee, we can also evaluate the mutual information reduction (MIR) index, in order to verify the efficiency of decomposition of EEG signals [5]. MIR is evaluated efficiently only in those cases when a complete decomposition is performed (i.e., $K = p$). We, instead, perform also dimension reduction, and thus we could not directly compute the MIR associated to the decomposition presented above. To make a comparison via

MIR we therefore performed a complete decompositions, and these could in principle be different from those described in the paper, where they are obtained with non-orthogonal basis, like those provided by fastICA and HICA. Hence we first check that the main features illustrated were still present in the complete decompositions. Then we evaluate the MIR of the complete decompositions, obtaining results in support of HICA. Indeed HICA, with a MIR of 9 Kb/s, outperforms PCA and Treelets, both providing a MIR around 0.88 Kb/s. Of course HICA provides a MIR lower than fastICA (28.75 Kb/s), since fastICA is precisely designed to find a data representation minimizing the mutual information of the sources. This shows that, with respect to its "multi-resolution" competitor (i.e., Treelets), HICA provides a better decomposition in terms of independence of the sources, in addition to a more interpretable basis elements.

7. Conclusion

We presented a new method for the construction of a multi-resolution non-orthogonal data-driven basis, appropriate to deal with high-dimensional and complex data. Non-orthogonality allows for basis elements with a physical interpretation, while multi-resolution provides basis elements able to catch very localized data features. The new HICA algorithm is obtained by merging the Treelet and the ICA algorithms. We illustrated the details of the HICA algorithm and propose a forward selection strategy to perform data-driven dimensional reduction with a non-orthogonal basis. Both the HICA algorithm and the dimensional reduction procedure have been implemented in the R package `fastHICA` [16].

Furthermore, we proved the consistency of the HICA algorithm. Indeed we proved that when the primitive variables are dependent according to a block structure such that between-block dependencies are weaker than within-block dependencies, HICA identifies the underlying block structure. The analysis of synthetic data suggests that when dealing non-Gaussian latent components (even non-orthogonal and/or dependent) HICA always performs better than or equally to PCA, ICA, and Treelets supporting the claim that HICA inherits the returns of both ICA and Treelets over PCA. Moreover, as expected by theory, simulations discourage the use of PCA and ICA when the latent components are dependent and the use of PCA and Treelets when the latent components are non-orthogonal. Simulations also show that the criterion proposed in [8], for the choice of K and l might take to a mis-

detection of the model. For a given value of K the more proper approach seems indeed to consider as candidate values for the level of aggregation l all values providing the maximal energy and not necessarily the minimum one. Finally, the analysis of EEG traces shows the possible returns of using in real applications methods providing multi-resolution and non-orthogonal representations of the phenomenon under investigation.

Appendix A: Proofs

PROOF OF LEMMA 1. For simplicity we consider T and E to be zero mean and unit variance random variables. The kurtosis of a zero mean and unit variance random variable Y is $kurt(Y) = E[Y^4] - 3$. If Y is gaussian, $kurt(Y) = 0$. Moreover if Y_1 and Y_2 are independent random variables and α e β real parameters, $kurt(\alpha Y_1 + \beta Y_2) = \alpha^4 kurt(Y_1) + \beta^4 kurt(Y_2)$. Hence:

$$\begin{aligned} &|kurt(\mathbf{w}'\mathbf{Z})| = |kurt(w_1 T + w_2 E)| = \\ &= |w_1^4 kurt(T) + w_2^4 kurt(E)| = |w_1^4 kurt(T)|. \end{aligned} \tag{7}$$

Since $kurt(T) \neq 0$, (7) is maximized by $w_1 = \pm 1$ (and $w_2 = 0$ because $\mathbf{w}$ is a vector of unitary norm). □

PROOF OF LEMMA 2. Suppose that the aggregation between variables follows the scheme:
$\{\cdots\{\{X_1, X_2\}, X_3\}\cdots, X_p\}$. Hence, at level $l = 1$ we aggregate:

$$\begin{aligned} X_1 &= \sigma^2 T + \sigma_e^2 E_1, \\ X_2 &= \sigma^2 T + \sigma_e^2 E_2. \end{aligned}$$

The whitening procedure of ICA, transforms the vector $\mathbf{X} = (X_1\, X_2)'$ in a new vector $\mathbf{Z} = (Z_1\, Z_2)'$ such that

$$\begin{aligned} Z_1 &= \tfrac{X_1 + X_2}{a} = \tfrac{2\sigma^2 T + \sigma_e^2 (E_1 + E_2)}{a}, \\ Z_2 &= \tfrac{X_1 - X_2}{b} = \tfrac{\sigma_e^2 (E_1 - E_2)}{b}, \end{aligned}$$

where a and b are, respectively, the standard deviations of $X_1 + X_2$ and $X_1 - X_2$. We observe that Z_1 is a non gaussian variable, while Z_2 is gaussian. Because of Lemma 1, the rotation found by fastICA in the whitened space coincides with the identity matrix. According to the selection criterium and

taking into account the normalization of the matrix $\widetilde{C}^{(1)}$ in step 2 of the HICA algorithm, we obtain $\mathbf{c}_1^{(1)} = (\frac{1}{\sqrt{2}}\frac{1}{\sqrt{2}}0\cdots 0)'$ and $Y_1^{(1)} = \frac{1}{\sqrt{2}}X_1 + \frac{1}{\sqrt{2}}X_2 = \frac{2}{\sqrt{2}}\sigma^2 T + \frac{\sigma_e^2}{\sqrt{2}}E^{(1)}$, where $E^{(1)} = E_1 + E_2$ and $\mathcal{V}(E^{(1)}) \leq \mathcal{V}(E_1) + \mathcal{V}(E_2) \leq 2$. Furthermore $\mathcal{V}(Y_1^{(1)}) \leq \mathcal{V}(\frac{2}{\sqrt{2}}\sigma^2 T) + \mathcal{V}(\frac{\sigma_e^2}{\sqrt{2}}E_1^{(1)}) \leq \sqrt{2}(\sigma^2 + \sigma_e^2)$. At level $l = 2$ we aggregate:

$$Y_1^{(1)} = \frac{2}{\sqrt{2}}\sigma^2 T + \frac{\sigma_e^2}{\sqrt{2}}E_1^{(l)},$$
$$X_3 = \sigma^2 T + \sigma_e^2 E_3.$$

The whitening procedure provides a vector $\mathbf{Z} = (Z_1\, Z_2)'$ such that

$$Z_1 = \frac{\sqrt{2}Y_1^{(1)} + X_3}{a'} = \frac{3\sigma^2 T + \sigma_e^2(E_1^{(1)} + E_3)}{a'},$$
$$Z_2 = \frac{Y_1^{(1)} - \sqrt{2}X_3}{b'} = \frac{\sqrt{2}\sigma_e^2(E_1^{(1)}/2 - E_3)}{b'},$$

where a' and b' are, respectively, the standard deviations of $\sqrt{2}Y_1^{(1)} + X_3$ and $Y_1^{(1)} - \sqrt{2}X_3$. Once again, Lemma 1, implies that the rotation provided by fastICA is the identity and according to the selection criterium and to the normalization of $\widetilde{C}^{(2)}$ we have

$$\mathbf{c}_1^{(2)} = \begin{bmatrix} \frac{1}{\sqrt{2}} & 0 \\ \frac{1}{\sqrt{2}} & 0 \\ 0 & 1 \\ 0 & 0 \\ \vdots & \vdots \\ 0 & 0 \end{bmatrix} \begin{bmatrix} \sqrt{\frac{2}{3}} \\ \frac{1}{\sqrt{3}} \end{bmatrix} = \begin{bmatrix} \frac{1}{\sqrt{3}} \\ \frac{1}{\sqrt{3}} \\ \frac{1}{\sqrt{3}} \\ 0 \\ \vdots \\ 0 \end{bmatrix},$$

and $Y_1^{(2)} = \sqrt{\frac{2}{3}}Y_1^{(1)} + \frac{1}{\sqrt{3}}X_3 = \frac{3}{\sqrt{3}}\sigma^2 T + \frac{\sigma_e^2}{\sqrt{3}}E_1^{(2)}$, where $E_1^{(2)} = E_1^{(1)} + E_3$ and $\mathcal{V}(E_1^{(2)}) \leq \mathcal{V}(E_1^{(1)}) + \mathcal{V}(E_3) \leq 3$. Moreover $\mathcal{V}(Y_1^{(2)}) \leq \mathcal{V}(\frac{3}{\sqrt{3}}\sigma^2 T) + \mathcal{V}(\frac{\sigma_e^2}{\sqrt{3}}E_1^{(2)}) \leq \sqrt{3}(\sigma^2 + \sigma_e^2)$. Iterating, we obtain the lemma when the aggregation scheme is $\{\cdots\{\{X_1, X_2\}, X_3\}\cdots, X_p\}$.

For a general aggregation scheme, at the level $l + 1 = 2, ..., p - 1$ we aggregate:

$$Y_i^{(l)} = \frac{m}{\sqrt{m}}\sigma^2 T + \frac{\sigma_e^2}{\sqrt{m}}E_i^{(l)},$$
$$Y_j^{(l)} = \frac{n}{\sqrt{n}}\sigma^2 T + \frac{\sigma_e^2}{\sqrt{n}}E_j^{(l)},$$

with $A_i^l \cap A_j^l = \emptyset$ and $m + n = l + 2$. The whitening procedure provides a vector $\mathbf{Z} = (Z_1\, Z_2)'$ such that

$$Z_1 = \frac{\sqrt{m}Y_i^{(l)} + \sqrt{n}Y_j^{(l)}}{a''} = \frac{(m+n)\sigma^2 T + \sigma_e^2(E_i^{(l)} + E_j^{(l)})}{a''},$$
$$Z_2 = \frac{\sqrt{n}Y_i^{(l)} - \sqrt{m}Y_j^{(l)}}{b''} = \frac{\sqrt{mn}(E_i^{(l)}/m - E_j^{(l)}/n)}{b''},$$

where a'' and b'' are, respectively, the standard deviations of $\sqrt{m}Y_i^{(l)} + \sqrt{n}Y_j^{(l)}$ and $\sqrt{n}Y_i^{(l)} - \sqrt{m}Y_j^{(l)}$. Then

$$\mathbf{c}_i^{(l+1)} = \begin{bmatrix} 0 & 0 \\ \vdots & \vdots \\ 0 & 0 \\ \frac{1}{\sqrt{m}} & 0 \\ \vdots & \vdots \\ \frac{1}{\sqrt{m}} & 0 \\ 0 & 0 \\ \vdots & \vdots \\ 0 & 0 \\ 0 & \frac{1}{\sqrt{n}} \\ \vdots & \vdots \\ 0 & \frac{1}{\sqrt{n}} \\ 0 & 0 \\ \vdots & \vdots \\ 0 & 0 \end{bmatrix} \begin{bmatrix} \sqrt{\frac{m}{m+n}} \\ \sqrt{\frac{n}{m+n}} \end{bmatrix} = \begin{bmatrix} 0 \\ \vdots \\ 0 \\ \frac{1}{\sqrt{m+n}} \\ \vdots \\ \frac{1}{\sqrt{m+n}} \\ 0 \\ \vdots \\ 0 \\ \frac{1}{\sqrt{m+n}} \\ \vdots \\ \frac{1}{\sqrt{m+n}} \\ 0 \\ \vdots \\ 0 \end{bmatrix},$$

and $Y_i^{(l+1)} = \frac{\sqrt{m}Y_i^{(l)} + \sqrt{n}Y_j^{(l)}}{\sqrt{m+n}} = \frac{m+n}{\sqrt{m+n}}\sigma^2 T + \frac{\sigma_e^2}{\sqrt{m+n}}E_i^{(l+1)}$, where $E_i^{(l+1)} = E_i^{(l)} + E_j^{(l)}$ and $\mathcal{V}(E_i^{(l+1)}) \leq \mathcal{V}(E_i^{(l)}) + \mathcal{V}(E_j^{(l)}) \leq m + n$. Moreover $\mathcal{V}(Y_i^{(l+1)}) \leq \mathcal{V}(\frac{m+n}{\sqrt{m+n}}\sigma^2 T) + \mathcal{V}(\frac{\sigma_e^2}{\sqrt{m+n}}E_i^{(l+1)}) \leq \sqrt{m+n}(\sigma^2 + \sigma_e^2)$. The result now follows by induction. □

Proof of Theorem 3. Assume that, at a generic level $l < p - K$ of the tree, random variables from different blocks have not been merged. Hence, from Lemma 2, any two variables in the active set $\mathfrak{A}$ have the form:

$$Y_u^{(l)} = \frac{m}{\sqrt{m}}\sigma_u^2 T_u + \frac{\sigma_e^2}{\sqrt{m}}E_u^{(l)},$$
$$Y_v^{(l)} = \frac{n}{\sqrt{n}}\sigma_v^2 T_v + \frac{\sigma_e^2}{\sqrt{n}}E_v^{(l)},$$

with $\mathbf{c}_u^{(l)} = \left(0 \cdots 0 \, \frac{1}{\sqrt{m}} \cdots \frac{1}{\sqrt{m}} \, 0 \cdots 0\right)'$,
$\mathbf{c}_v^{(l)} = \left(0 \cdots 0 \, \frac{1}{\sqrt{n}} \cdots \frac{1}{\sqrt{n}} \, 0 \cdots 0\right)'$ and $\mathbf{c}_u^{(l)}$, $\mathbf{c}_v^{(l)}$ have non-zero elements relative to two disjoint subsets of indexes A_u^l, A_v^l with $|A_u^l| = m$, $|A_v^l| = n$. Let $\delta_k = \frac{\sigma_e}{\sigma_k}$. We now consider two different cases. In the first case $A_u^l \subseteq F_i$ e $A_v^l \subseteq F_j$ $(i \neq j)$. Hence:

$$
\begin{aligned}
Y_u^{(l)} &= \tfrac{m}{\sqrt{m}}\sigma_i^2 T_i + \tfrac{\sigma_e^2}{\sqrt{m}} E_i^{(l)}, \\
Y_v^{(l)} &= \tfrac{n}{\sqrt{n}}\sigma_j^2 T_j + \tfrac{\sigma_e^2}{\sqrt{n}} E_j^{(l)}.
\end{aligned}
$$

Let $\sqrt{m}\sigma_i^2 + \widetilde{\sigma}_m^2 = \mathcal{V}(Y_u^{(l)})$ $(\widetilde{\sigma}_m^2 \leq \sqrt{m}\sigma_e^2)$ and $\sqrt{n}\sigma_i^2 + \widetilde{\sigma}_n^2 = \mathcal{V}(Y_v^{(l)})$ $(\widetilde{\sigma}_n^2 \leq \sqrt{n}\sigma_e^2)$. In this case, the distance covariance and distance correlation between $Y_u^{(l)}$ and $Y_v^{(l)}$ are, respectively:

$$
\begin{aligned}
\mathcal{V}(Y_u^{(l)}, Y_v^{(l)}) \leq \quad & \mathcal{V}(\tfrac{m}{\sqrt{m}}\sigma_i^2 T_i, \tfrac{n}{\sqrt{n}}\sigma_j^2 T_j) + \\
& + \mathcal{V}(\tfrac{\sigma_e^2}{\sqrt{m}} E_i^{(l)}, \tfrac{\sigma_e^2}{\sqrt{n}} E_j^{(l)}) \leq \sqrt[4]{mn}\sigma_{ij}, \\
\mathcal{R}(Y_u^{(l)}, Y_v^{(l)}) = \quad & \tfrac{\mathcal{V}(Y_u^{(l)}, Y_v^{(l)})}{\sqrt{\mathcal{V}(Y_u^{(l)})\mathcal{V}(Y_v^{(l)})}} \leq \\
& \leq \tfrac{\sqrt[4]{mn}\sigma_{ij}}{\sqrt{\sqrt{m}\sigma_i^2 + \widetilde{\sigma}_m^2}\sqrt{\sqrt{n}\sigma_j^2 + \widetilde{\sigma}_n^2}} = \\
= \quad & \tfrac{\sqrt[4]{mn}\sigma_{ij}}{\sqrt[4]{mn}\sigma_i\sigma_j\sqrt{1 + \frac{\widetilde{\sigma}_m^2}{\sqrt{m}\sigma_i^2}}\sqrt{1 + \frac{\widetilde{\sigma}_n^2}{\sqrt{n}\sigma_j^2}}} \leq \tfrac{\sigma_{ij}}{\sigma_i\sigma_j}.
\end{aligned}
$$

In the second case A_u^l, A_v^l are subsets of the same F_k. Hence

$$
\begin{aligned}
Y_u^{(l)} &= \tfrac{m}{\sqrt{m}}\sigma_k^2 T_k + \tfrac{\sigma_e^2}{\sqrt{m}} E_{k1}^{(l)}, \\
Y_v^{(l)} &= \tfrac{n}{\sqrt{n}}\sigma_k^2 T_k + \tfrac{\sigma_e^2}{\sqrt{n}} E_{k2}^{(l)}.
\end{aligned}
$$

Let $\sqrt[4]{mn}\sigma_k^2 c(\sigma_e) = \mathcal{V}(Y_u^{(l)}, Y_v^{(l)}) \leq \sqrt[4]{mn}\sigma_k^2$, with $c(\sigma_e)$ a constant such that $0 < c(\sigma_e) \leq 1$ and $c(\sigma_e) \overset{\sigma_e \to 0}{\longrightarrow} 1$. Furthermore $\mathcal{V}(Y_u^{(l)}) \leq \sqrt{m}(\sigma_k^2 + \sigma_e^2)$ and $\mathcal{V}(Y_v^{(l)}) \leq \sqrt{n}(\sigma_k^2 + \sigma_e^2)$. Therefore the distance correlation between $Y_u^{(l)}$ and

$Y_v^{(l)}$ are, respectively:

$$\begin{aligned}\mathcal{R}(Y_u^{(l)}, Y_v^{(l)}) &= \frac{\mathcal{V}(Y_u^{(l)}, Y_v^{(l)})}{\sqrt{\mathcal{V}(Y_u^{(l)})\mathcal{V}(Y_v^{(l)})}} \geq \\ &\geq \frac{\sqrt[4]{mn}\sigma_k^2 c(\sigma_e)}{\sqrt{\sqrt{m}(\sigma_k^2+\sigma_e^2)}\sqrt{\sqrt{n}(\sigma_k^2+\sigma_e^2)}} \geq \\ &\geq \frac{\sqrt[4]{mn}\sigma_k^2 c(\sigma_e)}{\sqrt[4]{mn}\sigma_k^2\sqrt{\left(1+\delta_k^2\right)^2}} = \frac{c(\sigma_e)}{1+\delta_k^2}.\end{aligned}$$

Since, from (5), the maximum distance correlation between variables belonging to different blocks is lower than the minimum distance correlation between variables belonging to the same block, aggregation involves variables relative to the same block and this proves the theorem.

Furthermore, if the noise variance is not too large, the K dimensional space that explains the most part of the variability is that spanned by the K basis elements related to the K blocks. Then the energy criterium identifies those elements. □

Acknowledgements

The authors want to thank dr. Fabio Rotondi and Davide Rossi Sebastiano, MD at *C. Besta Neurological Institute*, IRCCS Foundation - Milano for their invaluable advice in the phenomenological interpretation of the basis elements of the EEG case study.

References

[1] Bache, K., Lichman, M.: UCI Machine Learning Repository. Irvine, CA: University of California, School of Information and Computer Science. http://archive.ics.uci.edu/ml (2013)

[2] Chattopadhyay, A. K., Mondal, S., Chattopadhyay, T.: Independent Component Analysis for the objective classification of globular clusters of the galaxy NGC 5128. Comput. Stat. Data An. 57, 17–32 (2013)

[3] Comon, P., Jutten, C.: Handbook of Blind Source Separation: Independent Component Analysis and Applications, Academic Press (2010)

[4] De Luca, M., Beckmann, C. F., De Stefano, N., Matthews, P. M., Smith, S. M.: fMRI resting state networks define distinct modes of long-distance interactions in the human brain. Neuroimage 29, 1359-1367 (2006)

[5] Delorme, A., Palmer, J., Onton, J., Oostenveld, R., Makeig, S.: Independent EEG Sources are Dipolar. PLoS ONE 7, 1–14 (2012)

[6] Hyvarinen, A., Oja, E.: Independent Component Analysis: Algorithms and Applications. Neural Networks 13, 411–430 (2000)

[7] Hyvarinen, A., Karhunen, J., Oja, E.: Independent Component Analysis, Wiley, New York (2001)

[8] Lee, A. B., Nadler, B., Wasserman L.: Treelets - an adaptive multi-scale basis for sparse unordered data. Ann. Appl. Stat. 2, 435–471 (2008)

[9] Li, C., Yin, J., Zhao, J.: Using Improved ICA Method for Hyperspectral Data Classification. Arab. J. Sci. 39, 181–189 (2014)

[10] Ogden, R. T.: Essential Wavelets for Statistical Applications and Data Analysis, Birkhauser, Boston (1997)

[11] Mallat, S. G.: A theory for multiresolution signal decomposition: the wavelet representation. IEEE T. Pattern Anal. 11, 674–693 (1989)

[12] Maris, E., Oostenveld, R.: Nonparametric statistical testing of EEG- and MEG-data. J. Neurosci. Meth. 164, 177–190 (2007)

[13] McKeown, M. J., Sejnowski, T. J.: Independent Component Analysis of fMRI Data: Examining the Assumptions. Hum. Brain Mapp. 6, 368-372 (1998)

[14] Pardalos, P. M., Yatsenko, V., Sackellares, J. C., Shiau, D. S., Chaovalitwongse, W., Iasemidis, L., D.: Analysis of EEG data using optimization, statistics, and dynamical system techniques. Comput. Stat. Data An. 44, 391–408 (2003)

[15] R Development Core Team. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL http://www.R-project.org/ (2014)

[16] Secchi, P., Vantini, S., Zanini, P.: fastHICA: Hierarchical Independent Component Analysis: a multi-scale sparse non-orthogonal data-driven basis. R package version 1.0. http://CRAN.R-project.org/package=fastHICA (2014)

[17] Sharbrough, F., Chatrian, G. E., Lesser, R.P., Lders, H., Nuwer, M., Picton, T.W.: American Electroencephalographic Society Guidelines for Standard Electrode Position Nomenclature. J. Clin. Neurophysiol. 8, 200–202 (1991)

[18] Snodgrass, J. G., Vanderwart, M.: A standardized set of 260 pictures: norms for the naming agreement, familiarity, and visual complexity. J. Exp. Psychol-Hum. L. 6, 174–215 (1980)

[19] Székely, G. J., Rizzo, M. L., Bakirov, N. K.: Measuring and testing dependence by correlation of distances. Ann. Stat. 35, 2769–2794 (2007)

[20] Zhang, X.L., Begleiter, H., Porjesz, B., Wang, W., Litke, A.: Event related potentials during object recognition tasks. Brain Res. Bull. 38, 531–538 (1995)