Approximate Canonical Correlation Analysis for common/specific subspace decompositions

Radu Ranta, Steven Le Cam, Baptiste Chaudet, Louise Tyvaert, Louis Maillard, Sophie Colnat-Coulbois, Valérie Louis-Dorr

Approximate Canonical Correlation Analysis for common/specific subspace decompositions

Radu Ranta[a,*], Steven Le Cam[a], Baptiste Chaudet[a], Louise Tyvaert[a,b], Louis Maillard[a,b], Sophie Colnat-Coulbois[a,b], Valérie Louis-Dorr[a]

[a] *CRAN, Université de Lorraine, CNRS, Nancy, F-54000, France*
[b] *Neurology Service, Université de Lorraine, CHRU Nancy, F-54000, France*

Abstract

The objective of this paper is to present a new technique for jointly decomposing two sets of signals. The proposed method is a modified version of Canonical Correlation Analysis (CCA), which automatically identifies from the two (*a priori* noisy) data-sets, having the same number of samples but potentially different number of variables (measurements), an approximate bisector common subspace and its complementary specific subspaces. Within these subspaces, common and specific parts of the signals can be reconstructed and analysed separately. The method we propose here can also be seen as an extension of other joint decomposition methods based on "stacking" the analysed data sets, but, unlike these methods, we propose a "stacked basis" approach and we show its relationship with the CCA. The proposed method is validated with convincing results on simulated data and applied successfully on (stereo-)electroencephalographic signals, either for artefact cancelling or for identifying common and specific activities for two different physiological conditions (sleep - wake).

Keywords: subspace correlation, joint decomposition, EEG

1. Introduction

In multiple applications (including biomedical signal processing, which is our main motivation) it is of high interest to separate the recorded signals in sources or sets of sources (components) with different properties. Among the most common situations, one might need to separate between artefacts and signals of interest, between regular (background) signals vs. condition specific ones (task/disease related), etc. While classical (more or less blind) source separation procedures tackle this problem by decomposing a given data-set (one multi-channel recording) in so-called sources, it has been shown that using two or more data-sets simultaneously improves source separation. The data sets can be "trials" potentially recording different conditions but also recordings using different modalities, see *e.g.*, [1, 2]. The confrontation of these several data-sets is expected to provide more information, or *diversity*, about the underlying variables (activities) to be identified. Each trial/modality is expected to bring added value (complementary information) on these hidden variables and to reduce the degrees of freedom of the decomposition by providing their own additional constraints (insight on the data).

While for multiple data-sets the analysis can be done by methods extending the Canonical Correlation Analysis (CCA) [3] such as MCCA (multiple CCA) [4] or its recent extensions such as Independent Vector Analysis (IVA) [5, 6] or Joint Blind Source Separation (JBSS) [1], the most common situations consisting in two data sets can be tackled by CCA. These cases include comparisons of data sets recorded in two different conditions ("control" and "task") or one recorded data-set and a given "dictionary", either simulated or extracted itself from the data [7, 8, 9, 10][1].

*Corresponding author
Email address: radu.ranta@univ-lorraine.fr (Radu Ranta)

[1] It is noteworthy that MUSIC based algorithms for EEG source localization can also be seen as special cases of CCA, where one of the data-sets is the precomputed lead-field matrix (head model) [11, 12]. Besides, they can be further adapted for model-based separation and localization of "task" and "control" sources [13].

In this paper we propose a new methodology, inspired from the algebraical approaches to CCA [4], for identifying sources or more precisely basis (canonical) vectors separating two linearly related data-sets in common and specific subspaces. Related ideas were independently proposed by different authors [14, 15, 16]. We will discuss the differences between the method proposed in this paper and the cited references further in the paper. Non-linear relations between data-sets, such as explored by Kernel CCA [17] or neural networks implementations such as Deep CCA [18] are outside the scope of the paper, although we suggest some future research directions at the end of this paper.

We start by examining a strict constraint: we look for a common subspace, defined by a basis shared by the considered data sets. Such as in the MUSIC or CCA framework, the method is based on evaluating the correlations between the subspaces spanned by two matrices. The strictly common subspace can then be defined as the perfectly aligned vectors of the two basis (one for each subspace), *i.e.*, those sharing a correlation of 1. The specific parts of each data set are then identified as the orthogonal complement of each set with respect to the common subset. In real applications though, the constraint of unitary correlation is too strong, as the data are likely to be distorted due to frequency filtering or noise/artifact impacting the signals. In this case, the correlations between the basis vectors of the two data subspaces will not perfectly be 1 and no common basis can be found (angles between the components will not be perfectly 0). We propose then a method for identifying an approximate common subspace, built as the truncated bisector subspace of the two data sets. The size of the common subspace is determined through an eigenvalues ratio based test [19].

The usefulness and the performance of the approach is evaluated in details on simulated signals and two examples of application are shown on real electrophysiological signals.

2. Common/Specific subspace decomposition (CSSD)

We start our development by recalling classical multi-channel models, using the conventions employed in source separation / ICA literature. A data-set consists in an $M \times N$ matrix, where each row represents a signal recorded by one channel/electrode (M being thus the number of signals) and each column a time instant (N being the number of time samples). All along the paper, we assume that the number of samples N is *a priori* much higher than the number of channels M ($N \gg M$), such as sample covariance matrices are reliable estimates of the true covariances (and so are the estimated canonical correlation coefficients, see further in the paper). We consider in the following two sets of signals S_1 and S_2, possibly with different number of channels M ($M_1 \neq M_2$) but with the same number of time samples N, issued from two different measurements (two recording sessions of the same phenomenon, two sets of sensors with different partial views on the process, etc.). Zero mean signals are assumed all over the paper.

The signals from the set S_1 can then be modeled as a full row-rank[2] $M_1 \times N$ matrix $\mathbf{X}_1$, with $M_1 \ll N$. As the M_1 vectors representing the individual signals are linearly independent, they span an M_1-dimensional subspace of $\mathbb{R}^N$. A basis for this subspace can be obtained for example by whitening: after computing the product $\mathbf{\Phi}_1 = \mathbf{X}_1\mathbf{X}_1^T$, the basis vectors (*i.e.*, the whitened signals) $\mathbf{Z}_1$ are obtained as follows:

$$\mathbf{Z}_1 = \mathbf{D}_1^{-1/2}\mathbf{U}_1^T\mathbf{X}_1, \tag{1}$$

with $\mathbf{D}_1$ and $\mathbf{U}_1$ respectively the diagonal matrix of eigenvalues and the eigenvector matrix of the matrix $\mathbf{\Phi}_1 = \mathbf{U}_1\mathbf{D}_1\mathbf{U}_1^T$.

Strictly the same modelling is valid for the M_2 signals $\mathbf{X}_2$ and the corresponding basis $\mathbf{Z}_2$.

Three distinct situations are possible. In the first case, the two subspaces are disjoint, *i.e.*, they are orthogonal:

$$\mathbf{Z}_1\mathbf{Z}_2^T = \mathbf{0}_{M_1 \times M_2} \tag{2}$$

From a signal point of view, S_1 signals are orthogonal to the S_2 signals, which might be interpreted as there is no common information between the two of them.

[2] We assume all along this paper that both sets of measurements are linearly independent, thus full row-rank. If this hypothesis is not fulfilled, the dimension of the measurement space can be reduced, for one or for both sets, by some PCA technique (see *e.g.*, [20]).

Second situation, the two subspaces are overlapped. Assuming that $M_1 \geq M_2$, this means that the entire set S_2 signals can be obtained as a linear combination of S_1 signals, *i.e.*, there exists a linear transform $\mathbf{T}$ such as $\mathbf{X}_2 = \mathbf{T}\mathbf{X}_1$. Equivalently, the basis $\mathbf{Z}_2$ can be rotated onto a subset of M_2 vectors of the $\mathbf{Z}_1$ basis.

$$\mathbf{Z}_2 = \begin{bmatrix}\mathbf{I}_{M_2} & \mathbf{0}_{M_2\times(M_1-M_2)}\end{bmatrix}\mathbf{R}\mathbf{Z}_1 \tag{3}$$

with $\mathbf{R}$ an orthogonal (rotation) matrix and $\begin{bmatrix}\mathbf{I}_{M_2} & \mathbf{0}_{M_2\times(M_1-M_2)}\end{bmatrix}$ a selection matrix. Physiologically, this case might be interpreted as the situation when all the sources responsible for the signals in S_1 are also present or seen by the S_2 channels.

The most common situation appears when the two subspaces $\mathbf{X}_1$ (basis $\mathbf{Z}_1$) and $\mathbf{X}_2$ (basis $\mathbf{Z}_2$) are neither orthogonal nor overlapped. In this case, their relationships are measured by some angles quantifying the degree of correlation between them. The cosines of these angles are strictly similar to the canonical correlations given by the Canonical Correlation Analysis (CCA) [3]. Some of the angles might be null or close to zero, in which case the two subspaces might be decomposed in a common part (subspace) and specific subspaces having different possible relationships among them. Our aim explore these cases and propose a particular decomposition of $\mathbf{X}_1$ and $\mathbf{X}_2$ such as:

$$\mathbf{X}_1 = \mathbf{X}_{1c} + \mathbf{X}_{1s} \text{ and } \mathbf{X}_2 = \mathbf{X}_{2c} + \mathbf{X}_{2s} \tag{4}$$

where the subscripts $1c$ and $2c$ indicate respectively parts of the S_1 and S_2 signals laying in the common subspace (which might be interpreted as due to common sources), while $1s$ and $2s$ indicate respectively the part of the S_1 and S_2 signals linearly independent among them and orthogonal to the common sub-subspace. Using source separation notations, one can write the full model by stacking the two data sets:

$$\begin{bmatrix}\mathbf{X}_1\\ \mathbf{X}_2\end{bmatrix} = \begin{bmatrix}\mathbf{A}_{1c}\\ \mathbf{A}_{2c}\end{bmatrix}\mathbf{Y}_c + \begin{bmatrix}\mathbf{X}_{1s}\\ \mathbf{X}_{2s}\end{bmatrix} \tag{5}$$

In source separation language, $\mathbf{Y}_c$ are called "sources" and they are usually at least orthogonal, if not independent as in ICA. These sources do not necessarily have a physical of physiological interpretation, especially for biomedical data such as EEG, where brain sources have a clear neuroscientific interpretation and they are usually not independent. We are not interested here in source estimation, but in revealing common parts of two different data sets, *i.e.*, the intersection of their respective subspaces. Therefore, $\mathbf{Y}_c$ will designate further in the paper common basis vectors or basis of the common subspace. The dimension of this common subspace, *i.e.*, the number of rows in $\mathbf{Y}_c$, will be noted on M_c. Consequently, the dimensions of the mixing matrices $\mathbf{A}_{ic}$ ($i = 1,\ 2$) will be $M_i \times M_c$, while $\mathbf{X}_{ic}$ and $\mathbf{X}_{ic}$ will have the dimensions of the original data sets.

Note that this model is different from the ones presented in [16], where the number of rows in the two data sets is the same and where the joint ICA step assumes a common mixing matrix $\mathbf{A}_c$. On the other hand, the model (5) is similar to the ones proposed in [14] and [15], our decomposition method being different, as it will be described in the next subsections.

2.1. Ideal decompositions algorithm

The developments presented in this section are a brief recall of one of the classical algebraic approaches for finding the canonical variates (scores) and the projection matrices in CCA (see *e.g.*, [21]). A first step consists in finding the parts of the S_1 and S_2 signals laying in the common subspace, the complements being simply obtained by subtraction from (4). Intuitively, the common subspace can be found rotating the two measurement basis $\mathbf{Z}_1$ and $\mathbf{Z}_2$ (1) onto one another: the common vectors will constitute a basis $\mathbf{Y}_c$ of the common subspace.

An elegant solution is provided by Golub and Loan [22, p. 603-605]. The aim is to find two rotation matrices $\mathbf{V}_Z$ ($M_1 \times M_1$) and $\mathbf{W}_Z$ ($M_2 \times M_2$) that will align the subspaces spanned by $\mathbf{Z}_1$ and $\mathbf{Z}_2$ (*i.e.*, will maximize the number of common basis vectors). This can be done by singular value decomposition of the $M_1 \times M_2$ matrix $\mathbf{Z}_1\mathbf{Z}_2^T$ (assume that $M_1 \geq M_2$, otherwise permute):

$$\mathbf{Z}_1\mathbf{Z}_2^T = \mathbf{V}_Z\mathbf{\Sigma}\mathbf{W}_Z^T \tag{6}$$

Let

$$\begin{aligned}\mathbf{Y}_1 &= \mathbf{V}_Z^T\mathbf{Z}_1\\ \mathbf{Y}_2 &= \mathbf{W}_Z^T\mathbf{Z}_2\end{aligned} \tag{7}$$

be the rotated versions of the $\mathbf{Z}_1$ and $\mathbf{Z}_2$ basis. It is easy to see that

$$\mathbf{Y}_1\mathbf{Y}_2^T = \mathbf{\Sigma}, \tag{8}$$

thus that a vector k in $\mathbf{Y}_1$ is orthogonal to all vectors $j \neq k$ from $\mathbf{Y}_2$ and that the angle between vectors having the same index k in both $\mathbf{Y}_1$ and $\mathbf{Y}_2$ basis is given by the k-th singular value σ_k. In particular, if $\sigma_k = 1$, the two basis vectors are aligned (their cosine/cross-correlation is 1). In other words (see Theorem 12.4.2 from [22]), the common subspace is defined by the vectors (the rows) $\mathbf{Y}_c$ from $\mathbf{Y}_1$ (or, equivalently, from $\mathbf{Y}_2$) corresponding to unitary singular values in $\mathbf{\Sigma}$. Let M_c be the number of unitary singular values (which are also the greatest, as $\mathbf{Z}$ vectors have all unit norm). Then

$$\mathbf{Y}_1 = \begin{bmatrix} \mathbf{Y}_c \\ \mathbf{Y}_{1s} \end{bmatrix} \tag{9}$$

where $\mathbf{Y}_c$ corresponds to the first M_c rows of $\mathbf{Y}_1$ and $\mathbf{Y}_{1s}$ constitute a basis for $\mathbf{X}_{1s}$, the specific part of $\mathbf{X}_1$ signals.

The part of the S_1 signals common with the S_2 is obtained combining (1) and (9) (or projecting $\mathbf{X}_1$ on $\mathbf{Y}_c$)

$$\mathbf{X}_{1c} = \mathbf{U}_1\mathbf{D}_1^{1/2}\mathbf{V}_{Z,1..M_c}\mathbf{Y}_c = \mathbf{A}_{1c}\mathbf{Y}_c \tag{10}$$

while the complementary part can be obtained by simple subtraction:

$$\mathbf{X}_{1s} = \mathbf{X}_1 - \mathbf{X}_{1c} \tag{11}$$

or, as long as (9) holds, by projection:

$$\mathbf{X}_{1s} = \mathbf{U}_1\mathbf{D}_1^{1/2}\mathbf{V}_{Z,M_c+1..M_1}\mathbf{Y}_{1s} = \mathbf{A}_{1s}\mathbf{Y}_{1s} \tag{12}$$

In equation (10), $\mathbf{V}_{Z,1..M_c}$ corresponds to the first M_c rows of the orthogonal matrix $\mathbf{V}_Z$ from (6).

Strictly similar equations can be written for S_2 set of signals, *i.e.*, for obtaining $\mathbf{X}_{2c}$ and $\mathbf{X}_{2s}$.

It is useful to recall that, in the general case, $\mathbf{X}_{1s} \not\perp \mathbf{X}_{2s}$, as $\mathbf{X}_{1s} \not\perp \mathbf{X}_2$ and $\mathbf{X}_{2s} \not\perp \mathbf{X}_1$. Indeed, we do not have simultaneously $\mathbf{Y}_c \perp \mathbf{Y}_{1s}$, $\mathbf{Y}_c \perp \mathbf{Y}_{2s}$ and $\mathbf{Y}_{1s} \perp \mathbf{Y}_{2s}$. The non-orthogonality between the subspaces spanned by $\mathbf{Y}_{1s}$ and $\mathbf{Y}_{2s}$ is revealed by the non-unitary non-null singular values in (8). These values represent cosines of the principal angles between the two subspaces [22] and reveal the existing correlations between the "uncommon" signals belonging to the two sets.

To conclude on the proposed common-specific subspace decomposition method proposed here, it is noteworthy that, even if the mixing model is similar to those proposed in [14, 15], the solution is different, as we perform a SVD after stacking the basis of the data subspaces and not the data themselves. This leads to a common subspace that is not biased by the correlation structure of the original datasets, which remain present even after centering and scaling.

2.2. Common subspace for the overdetermined case

The previous approach is based on identifying the unitary singular values in $\mathbf{\Sigma}$ (8), revealing activities due to common sources projected on both sets of signals S_1 and S_2. This is an ideal particular case, only possible when the dimension of the common subspace plus the dimension of the specific subspaces equals the dimension of the signal spaces (N_1 and N_2). Indeed, assume that

$$\mathbf{X} = \mathbf{X}_c + \mathbf{X}_s = \mathbf{A}_c\mathbf{Y}_c + \mathbf{A}_s\mathbf{Y}_s$$

with $\mathbf{Y}_c$ and $\mathbf{Y}_s$ the disjoint basis of common and specific subspaces, of dimensions M_c and M_s respectively (we drop in this paragraph the indices of the two subspaces 1 and 2). It is easy to see that if $M_c + M_s > M$, than the basis $\mathbf{Z}$ of the measured signals $\mathbf{X}$ looses some information because of the loss in dimension.

In the general case, it is perfectly possible that $M_c + M_s > M$ (a trivial case is obtained when $\mathbf{Y}_s$ are independent random noises, with $\mathbf{A}_s$ equal or not to the identity matrix). In this case, none of diagonal elements of $\mathbf{\Sigma}$ (*i.e.*, the cosines of the principal angles) is equal to 1. On the other hand, if some of the correlations are close enough to 1, one might suspect that common sources exist, but they are somehow distorted when recorded by the electrodes, either because of filtering effects or by perturbing specific sources (possibly more than electrodes, *i.e.*, projected from a

higher dimension subspace) with different amplitudes and thus varying signal to noise ratio. A more general solution can then be proposed if one accepts the existence of a common subspace of dimension M_c if M_c values of the singular values in $\mathbf{\Sigma}$ (eq. 8) are great enough.

A first problem that appears at this point is how to find a basis for this common subspace. This basis is given by Proposition 1 below, which gives a new interpretation of the canonical variates given by Kettenring [4]. The proof is based on an intermediate result, detailed in the Appendix.

Proposition 1. Let $\mathbf{Y}$ be the the $(M_1+M_2)\times N$ matrix obtained by "stacking" $\mathbf{Y}_1$ and $\mathbf{Y}_2$ from (7):

$$\mathbf{Y} = \begin{bmatrix}\mathbf{Y}_1\\ \mathbf{Y}_2\end{bmatrix} = \mathbf{V}\mathbf{S}\mathbf{W}^T \tag{13}$$

with $\mathbf{V}$, $\mathbf{S}$, $\mathbf{W}$ obtained by its SVD decomposition. By construction, $\mathbf{W}$ is a basis for the space spanned together by $\mathbf{Y}_1$ and $\mathbf{Y}_2$. The basis $\mathbf{Y}_c$ of the common subspace between $\mathbf{Y}_1$ and $\mathbf{Y}_2$ (9) is identical to the first M_c vectors (columns) in $\mathbf{W}$, corresponding to singular values of $\sqrt{2}$. An approximation of the common subspace is given by the the columns in $\mathbf{W}$ corresponding to singular values in $\mathbf{S}$ smaller than but close to $\sqrt{2}$. These columns define a bisector subspace between $\mathbf{Y}_1$ and $\mathbf{Y}_2$.

Proof First, it's easy to see that

$$\mathbf{Y}\mathbf{Y}^T = \mathbf{M} = \mathbf{V}\mathbf{\Lambda}\mathbf{V}^T, \tag{14}$$

where $\mathbf{M}$ is given by (24), $\mathbf{V}$ by (32) and $\mathbf{\Lambda}$ is the diagonal matrix having $\boldsymbol{\lambda}$ (27) on the diagonal.

As $\mathbf{V}$ is orthogonal and observing that the singular values in $\mathbf{S}$ are $\sqrt{\lambda_i}$ (27), one can write:

$$\mathbf{V}^T\mathbf{Y} = \mathbf{S}\mathbf{W}^T$$

Take the i-th row ($i \le r$):

$$\mathbf{v}_i^T \begin{bmatrix}\mathbf{Y}_1\\ \mathbf{Y}_2\end{bmatrix} = \sqrt{\lambda_i}\mathbf{w}_i^T, \tag{15}$$

where $\mathbf{w}_i$ is the i-th right singular vector (column of $\mathbf{W}$). Consequently, as $\mathbf{Y}_1$ has exactly M_1 rows,

$$\mathbf{w}_i^T = \frac{1}{\sqrt{2\lambda_i}}(\mathbf{y}_{1,i}+\mathbf{y}_{2,i}), \tag{16}$$

with $\mathbf{y}_{1,i}$ and $\mathbf{y}_{2,i}$ the i-th rows of $\mathbf{Y}_1$ and $\mathbf{Y}_2$ respectively.

It is now easy to notice that $\mathbf{w}_i^T$ is the bisector between $\mathbf{y}_{1,i}$ and $\mathbf{y}_{2,i}$: right multiplication by either $\mathbf{y}_{1,i}^T$ or $\mathbf{y}_{2,i}^T$ gives $\frac{1}{\sqrt{2\lambda_i}}(1+\sigma_i) = \sqrt{\frac{\lambda_i}{2}}$, which is the cosine of the angle α_i between $\mathbf{y}_{1,i}$ (or $\mathbf{y}_{2,i}$) and $\mathbf{w}_i$. Recalling that $\sigma_i = \lambda_i - 1$ is the cosine of the angle β_i between $\mathbf{y}_{1,i}$ and $\mathbf{y}_{2,i}$, simple trigonometric identity shows that $\alpha_i = \beta_i/2$.

Finally, a singular value of $\sqrt{2}$ implies $\lambda_i = 2$ ($\sigma_i = 1$), thus identical $\mathbf{y}_{1,i}$ and $\mathbf{y}_{2,i}$, which will also be identical to $\mathbf{w}_i$ (16).

The previous reasoning yields an approximate common subspace defined by the basis $\hat{\mathbf{Y}}_c = \mathbf{W}_{1:M_c}^T$, which is simply a basis of the bisector space between $\mathbf{Y}_1$ and $\mathbf{Y}_2$, in other words a normalized version[3] of $(\mathbf{Y}_1+\mathbf{Y}_2)$ (a possible way for determining M_c is proposed in the next subsection). Another way of interpreting $\mathbf{Y}_c$ is as a set of features of an "average representative individual" or a centroid of a class, if one uses CCA for discriminative purposes such as proposed by [23, 24].It is important to note that it does not allow us to determine specific subspaces $\hat{\mathbf{Y}}_{1s}$ and $\hat{\mathbf{Y}}_{2s}$ orthogonal to $\hat{\mathbf{Y}}_c$. Consequently, if common parts $\hat{\mathbf{X}}_{1c}$ and $\hat{\mathbf{X}}_{2c}$ can be obtained by projecting $\mathbf{X}_1$, respectively $\mathbf{X}_2$, onto $\hat{\mathbf{Y}}_c$ by

$$\hat{\mathbf{X}}_{1c} = \hat{\mathbf{Y}}_c\hat{\mathbf{Y}}_c^T\mathbf{X}_1 \;\text{ and }\; \hat{\mathbf{X}}_{2c} = \hat{\mathbf{Y}}_c\hat{\mathbf{Y}}_c^T\mathbf{X}_2, \tag{17}$$

[3]The proof is immediate, in two steps: (1) as all vectors in $\mathbf{Y}_1$ and $\mathbf{Y}_2$ have unit length, any resultant vector $\tilde{\mathbf{w}}_i = \mathbf{y}_{1,i}+\mathbf{y}_{2,i}$ will be the bisector between $\mathbf{y}_{1,i}$ and $\mathbf{y}_{2,i}$, with a length equal to $\sqrt{2(1+\rho_i)}$, where ρ_i is the cosine of the i^{th} principal angle; (2) as $\mathbf{Y}_1\mathbf{Y}_2^T$ is diagonal by construction (8), $(\mathbf{Y}_1+\mathbf{Y}_2)(\mathbf{Y}_1+\mathbf{Y}_2)^T$ will also be diagonal, meaning that the individual vectors $\mathbf{y}_{1,i}+\mathbf{y}_{2,i}$ are orthogonal between them for different i.

the specific parts $\hat{\mathbf{X}}_{1s}$ and $\hat{\mathbf{X}}_{2s}$ can only be obtained by subtraction (11) and they are no longer orthogonal to the common parts (one cannot simultaneously obtain specific parts orthogonal to the common part and perfectly reconstruct the original data, as $\hat{\mathbf{Y}}_c$ is no longer an exact sub-basis of the original data sets).

Finally, it is important to note that the inherent sign ambiguity in the SVD can be tackled as proposed in [25], in order to avoid hazardous reconstruction of the common/specific subspaces by (10) and (11).

2.3. Common subspace dimension

An auxiliary question is the value of the threshold ρ used for determining the common subspace dimension, *i.e.*, the number of σ_i singular values in $\mathbf{\Sigma}$ "close" to 1 (6). This is significantly different from looking for non-null correlations: in our case, all components (sources) of the two datasets can be in principle correlated among them, for example $\mathbf{Y}_1$ or $\mathbf{Y}_2$ in equation 7. In other words, we are looking for a set of basis vectors corresponding to the common subspace and not to the whole signal subspace, so we cannot directly use model order selection criteria such as Akaike's AIC or Minimum Description Length (MDL) [20, 26, 27]. More recent and elaborated model selection methods, robust to small sample sizes also (*e.g.*, [28] or [29]) are not adequate neither, for the same reasons. This is a meaningful situation in many applications, as for example in EEG analysis: different brain areas produce correlated activities, but some sources propagate to all electrodes. Studying correlations between the former after separating them from the latter might reveal important functional information on the brain.

A first solution for detecting common sources is a simple thresholding of the singular values in $\mathbf{\Sigma}$ (6). As these values stand for correlations between the two subspaces S_1 and S_2, one might simply impose some meaningful user-chosen threshold smaller than 1, in order to be robust to noise. Indeed, the threshold, whose aim is to select big-enough correlations, should take into account the maximum possible correlation between two noisy data sets. Even if one assumes two identical sets of vectors $\mathbf{Y}_1$ and $\mathbf{Y}_2$, each one perturbed by white independent noise with a given signal-to-noise ratio (*SNR*), the maximal correlations are limited by this *SNR*. One can show that the correlation coefficient of two identical signals perturbed by noise equals:

$$CC_{max} = \frac{SNR^2}{1+SNR^2} \tag{18}$$

(here, the *SNR* is not considered in dB, but as the ratio between signal and noise standard deviations). For example, for $SNR = \sqrt{2}$ (*i.e.*, 3dB), the maximum correlation will be $2/3$. In other words, if one suspects strong noise, the threshold should be lowered.

This thresholding can be written directly for the eigenvalues λ_i from (27). As these values stand for variances of the principal vectors, we can fix a threshold ρ standing for a minimum correlation acceptable between two basis vectors in order to be accepted as similar (for example, a value of $\lambda_1 > \rho = 1.9$ is equivalent to a correlation between the first basis vectors $\sigma_1 > 0.9$). This value can be modulated by the signal to noise ratio if some prior knowledge is available.

On the other hand, although we do not explore here explicitly the small sample size (*M* comparable to *N*)[4] it was shown for example in [28] that the sample correlation values could be over-estimated from the data, meaning that we could have an overestimated dimension of the common space if they pass above the chosen threshold. But, if the methods proposed in [28] or [29] are effective for estimating the correct number of non-null correlation coefficients, they do not explicitly provide unbiased estimates of these coefficients.

We thus proceed by a more or less heuristic argument, at the basis of a second method for determining the common space dimension. Recall that we are only interested in correlations theoretically equal to 1. In practice, we consider that all singular values that are not significantly different from the first one indicate common basis vectors (sources), if of course the first value is itself significantly above noise. The second method aims thus to partially avoid the arbitrary thresholding proposed by the first one by detecting a first gap or jump among the eigenvalues λ_i. We propose then to use a factor analysis based criterion introduced by [19], which yields the dimension of the factors subspace by simply taking the maximum of the ratio of two successive eigenvalues:

$$\hat{N}_c = \text{argmax}_k \frac{\lambda_k}{\lambda_{k+1}} \tag{19}$$

[4] In most of the pratical applications we are interested in, the sampling frequency and the length of the recordings is high enough with respect to the number of channels (it is in fact more current to have redundant channels and thus low column rank than low row rank).

equivalent to the maximum of the Fisher statistic testing for equality between two successive eigenvalues (variances of the principal components $\mathbf{W}$).

It is noteworthy that we have also evaluated two other common subspace dimension detection techniques (equality of two successive eigenvalues, or equality of the different eigenvalues with the first one, both by classical Fisher hypothesis tests with different non-detection levels α). We do not report here the results, as they are (in average) slightly below the eigevalues ratio method on our simulated signals (see next section).

3. Results

3.1. Simulation

In order to illustrate our approach, we have simulated two sets of signals of different sizes $M_1 = 7$ and $M_2 = 5$. The signals were constructed as follows: a set of $M_c = 3$ signals of length N were considered as common sources and they were projected using random matrices $M_1 \times M_c$ (respectively $M_2 \times M_c$) to form the common parts $\mathbf{X}_{1c}$ and $\mathbf{X}_{2c}$ of the two sets $\mathbf{X}_1$ and $\mathbf{X}_2$. Specific signals were also generated separately between the two subspaces, ensuring that the maximum subspace correlation between them was 0.2. In the simplest configuration, both signal sets were well determined, with dimensions of the specific subspaces of $M_{1s} = M_1 - M_c$ respectively $M_{2s} = M_2 - M_c$, thus an Xb_1 subspace of dimension M_1, an Xb_2 subspace of dimension M_2 and an overall dimension of both subspaces of $M_1 + M_2 - M_c$. The dimensions of each specific subspace was next increased by adding specific signals, from 1 specific signal per subspace to $M_1 + M_2$ specific signals per subspace, yielding thus maximal dimensions for the specific subspaces of $M_{1s} = 2M_1 + M_2 - M_c$ and $M_{2s} = 2M_2 + M_1 - M_c$, *i.e.*, an overall dimension of $3(M_1 + M_2) - M_c$. In the well determined case, signals $\mathbf{X}_1$ are obtained by mixing M_1 sources, signals $\mathbf{X}_2$ by mixing M_2 sources, $M_c = 3$ of them being common[5]. In this case, one should obtain M_c canonical correlation coefficients equal to 1 and thus potentially perfectly recoverable common subspace. In the worst case, we deal with highly overdetermined signals $\mathbf{X}_1$ and $\mathbf{X}_2$, obtained by projecting $2M_1 + M_2$ independent signals onto the M_1 dimensional subspace of $\mathbf{X}_1$ (respectively $2M_2 + M_1$ signals onto the M_2 dimensional subspace of $\mathbf{X}_2$), M_c of them being common.

Moreover, we have varied the common to specific power ratio (which could be interpreted as a signal to noise or interference ratio) from 20 to 0 dB ($SNR \in \{\infty,\ 20,\ 10,\ 3,\ 0\}$dB) by varying the gains K_1 and K_2 in equations (20) and (21).

$$\mathbf{X}_1 = \mathbf{X}_{1c} + K_1\mathbf{X}_{1s} = \mathbf{A}_{1c}\mathbf{Y}_c + K_1\mathbf{A}_{1s}\mathbf{Y}_{1s} \quad (20)$$

$$\mathbf{X}_2 = \mathbf{X}_{2c} + K_2\mathbf{X}_{2s} = \mathbf{A}_{2c}\mathbf{Y}_c + K_2\mathbf{A}_{2s}\mathbf{Y}_{2s} \quad (21)$$

where the dimensions of $\mathbf{Y}_{1s}$ and $\mathbf{Y}_{2s}$ varied as explained above.

One hundred simulations were done randomly varying the mixing matrices $\mathbf{A}$ (uniform between -1 and 1), but also the common and specific sources, which were randomly chosen from a simulated sources database created as follows: we have taken a set of intracerebral signals (SEEG) recorded at the Nancy University Hospital (CHU Nancy) (195 signals at 2048 Hz sampling rate), that we have whitened (eq. 1) in order to obtain unitary norm orthogonal sources constituting the mentioned database. The length N of all signals was 8196 (4 seconds), much higher that M_1 or M_2. Finally, in order to test the influence of unbalanced datasets, $\mathbf{X}_2$ amplitudes were multiplied by a factor of 2.

The CSSD procedure was then applied with the aim of estimating common ($\hat{\mathbf{X}}_{1c}$ and $\hat{\mathbf{X}}_{2c}$) and specific ($\hat{\mathbf{X}}_{1s}$ and $\hat{\mathbf{X}}_{2s}$) parts of the two signal sets $\mathbf{X}_1$ and $\mathbf{X}_2$, using (17). Three types of estimates were computed: the first one by informing the algorithm with the correct subspace dimension ($M_c = 3$), the second one using the proposed ER test (19) and the third one using a fixed threshold $\rho = 1 + 0.9$ for determining the common subspace dimension[6]. The three types of estimates were indexed respectively with o, e, and ρ. We present below averaged results over the two signal sets $\mathbf{X}_1$ and $\mathbf{X}_2$, indexed as explained before *e.g.*, $\hat{\mathbf{X}}_{1c,o}$, $\hat{\mathbf{X}}_{1c,e}$ or $\hat{\mathbf{X}}_{1c,\rho}$ for the common parts of the signals from the first set.

[5] Note that this is the situation modeled by JIVE [14], where the ranks of the specific subspaces are smaller than the number of measurements.

[6] A fourth intermediate estimation was done using a threshold depending on the SNR (18)

$$\rho = 1 + 0.9CC_{max} \quad (22)$$

but the results are not presented here for conciseness.

We have also compared the 3 CSSD estimates (o, e, and ρ) with the results obtained by JIVE [14] and DISCO-SCA [15]. It is important to note though that, among the two cited algorithms, DISCO-SCA does not specifically aim to reconstruct the parts of the original data sets lying in the common and specific subspaces, but the "basis/sources" $\mathbf{Y}$ of these subspaces. The algorithm proceeds in two steps: a first one consisting in an SVD of the stacked data, followed by a selection of the common basis vectors (dimension of the common space M_c) and a supplementary rotation. On the other hand, our main motivation is of denoising type, *i.e.*, we aim to estimate common an specific parts of the original signals by reconstructing them once the respective subspace basis are found. In other words, the supplementary rotation in DISCO-SCA is irrelevant for our application.

Unlike DISCO-SCA, JIVE [14] has the same objectives as our CSSD method. Still, the model is different: first, the rank of the specific subspace of every dataset in JIVE is assumed to be $M - M_C$, implying that the original data mixture is full-rank and it does not stem from a projection on a lower dimensional subspace (which is allowed in our method). On the other hand, it includes independent white noise, while in our model the "noise" results from the projection on the measurement space of the specific sources (and it can thus be correlated). Finally, from an algorithmic point of view, the first step of JIVE is similar to DISCO-SCA, that is an SVD of the stacked data (optionally normalized) and common space rank selection. This initial step is followed by an iterative procedure that stems from the successive rank estimation steps, allowing to take into account the noise. In the absence of the rank re-estimation (if the algorithm converges after the first iteration), the reconstructed common/specific parts are thus the same as those obtained by DISCO-SCA (on raw or normalized data).

Consequently, we only compare here the results of the different versions of CSSD with the common first step of [14, 15] consisting in an SVD of the stacked data, normalized or not, followed by a selection of the M_c common sources by rank thresholding. For a fair comparison, we have used the same three estimates for M_c as for CSSD. The estimated reconstructed signals are indexed in the following with j and jn for the reconstructions obtained respectively by simply stacking the (possibly normalized) $\mathbf{X}$ sets before SVD. For example, the reconstruction of the common signals of the first data set using the informed subspace dimension and normalized JIVE will be noted $\hat{\mathbf{X}}_{1c,o,jn}$.

The quality of the different estimations was evaluated by the mean correlation coefficient (*i.e.*, averaged over the M_1, respectively M_2 signals), indexed by the signal set (1 or 2), subspace type (common c or specific s) and thresholding method (ρ, e, and o), as for example in:

$$CC_{1c,\rho} \quad = \quad \frac{1}{M_1} \sum_{i=1}^{M_1} \frac{\mathbf{x}_{1c,i}^T \hat{\mathbf{x}}_{1c,\rho,i}}{\|\mathbf{x}_{1c,i}\| \|\hat{\mathbf{x}}_{1c,\rho,i}\|} \tag{23}$$

where the index i in *e.g.*, $\mathbf{x}_{1c,i}$ denotes the i-th signal (row) in $\mathbf{X}_{1c}$. Note that (23) can also be computed using the raw measured signals instead of the denoised estimates, (*e.g.*, by replacing $\hat{\mathbf{X}}_{1c,\rho}$ with $\mathbf{X}_1$), and one might consider that all flavours of CSSD that improve these raw signals values are useful.

One hundred simulations were performed using the previously described setup. In order to synthesise the results, we averaged the correlation coefficients (23) over the two data sets: for example, $CC_{c,o}$ represents the average correlation over the two data sets for the common signals reconstructed using CSSD and the informed common subspace dimension, while $CC_{c,o,j}$ is the JIVE/DISCO-SCA version, on non-normalized data (results on the separate datasets are similar). The CSSD results are synthetically presented in figure 1 for the three different common space dimension choice methods, for different signal to noise ratios and total dimension of the original "source" space (see also discussion below).

We first focus on CSSD results only, for different conditions and space dimension detection procedures. As it can be seen, for high SNRs (first column in all subfigures), that is for low power specific parts, the estimated common part is very precisely recovered, regardless of the dimension of the specific subspace by all methods (including not doing anything and estimating the common part by the measurements themselves, $CC_{c,m}$). The specific part on the other hand, is almost invisible in the measurements (first column of $CC_{s,m}$), but quite properly estimated by the different versions of CSSD, including for high dimensional problems (last row, 0.82 minimal correlation for $CC_{s,o}$ and $CC_{s,e}$, 0.79 for $CC_{s,\rho}$). Likewise, if the mixtures are well conditioned (*i.e.*, the specific subspace dimension $M_s = M - M_c$), all common and specific parts are very well estimated (first rows of all subfigures).

If the dimension of the common subspace is correctly informed ($CC_{c,o}$ and $CC_{s,o}$), all signals (common and specific) are well recovered, with better performances for low dimensional specific subspaces and high SNR. But even the worst performances are 0.86 for $CC_{c,o}$ (high dimension and low SNR - lower right corner of the second subfigure

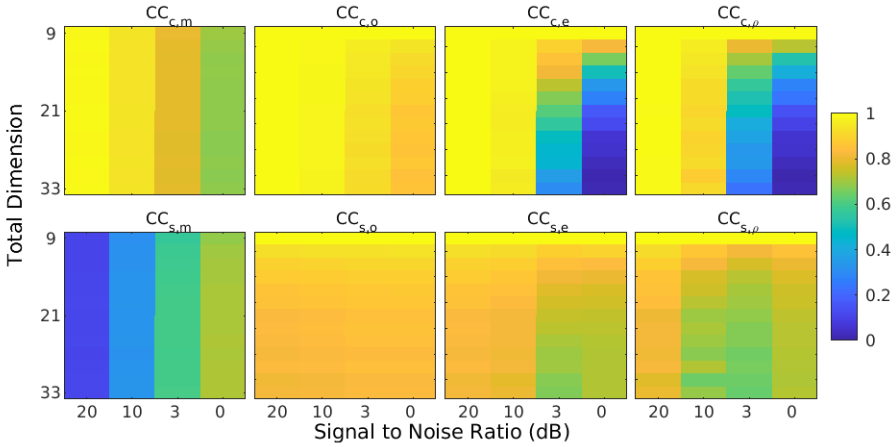


Figure 1: Correlation coefficients between simulated and estimated common parts $\mathbf{X}_c$ and $\hat{\mathbf{X}}_c$ (first row) and specific parts $\mathbf{X}_s$ and $\hat{\mathbf{X}}_s$ (second row), averaged over $\mathbf{X}_1$ and $\mathbf{X}_2$. From left to right, the estimates were the measurements themselves, the informed dimension estimator (index o), the ER based dimension detection (index e) and the fixed threshold dimension detection (index ρ). For each subfigure, the original common + specific subspaces dimension increases from up to down (well determined in the upper part, *i.e.*, $\mathbf{X}_1$ of dimension M_1 and $\mathbf{X}_2$ of dimension M_2, thus a total dimension of, severely underdetermined in the lower part ($2M_1+M2$)), while the signal to noise ratio (common to specific) decreases from left to right.

on row 1) and 0.82 $CC_{s,o}$ (high dimension and high SNR - lower left corner of the second subfigure on row 2).

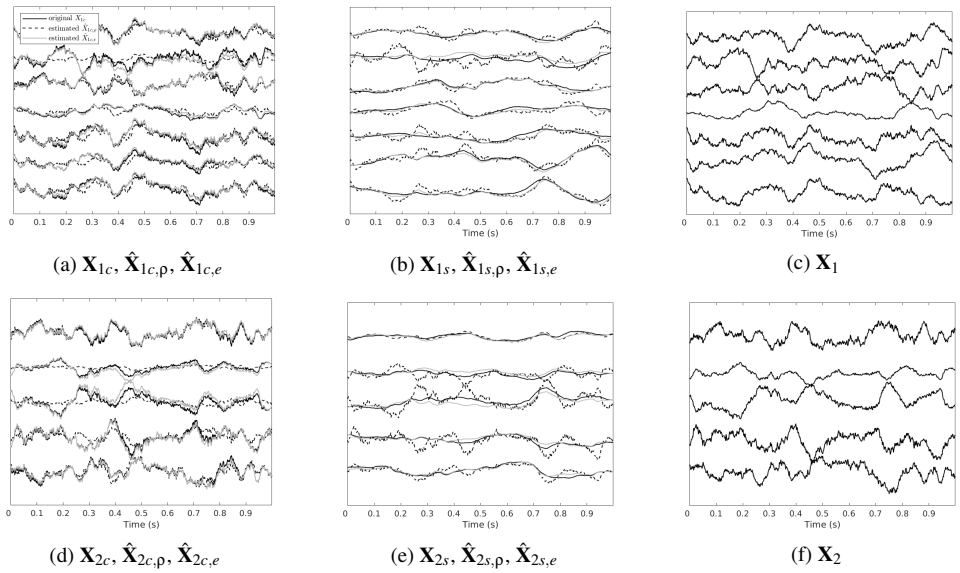
Among the two uninformed procedures, the eigenvalue ratio automatic detection procedure performs better than a fixed threshold, but both fail for high-dimensional mixtures and low SNR. It is useful to mention though that the maximum ER procedure was not applied alone: more specifically, we only considered that a common subspace exists ($M_c \neq 0$) if the highest eigenvalue was above the fixed threshold $\rho = 1.9$. In most of the cases, when $\lambda_1 > \rho$, the dimension of the common subspace was correctly estimated. Of course, in the opposite situation, the estimated $M_c = 0$ (lower right corner of subfigures $CC_{c,e}$, $CC_{s,e}$, $CC_{c,\rho}$ and $CC_{s,\rho}$. Still, for 10 dB SNR, the common and specific spaces were rather correctly estimated by the ER based procedure even for high dimensional mixtures (minimum correlation coefficients $CC_{1c,e} = 0.96$ and $CC_{1s,e} = 0.80$), with lower performances for the fixed threshold method ($CC_{1c,\rho} = 0.91$ and $CC_{1s,\rho} = 0.61$). For 3 dB, correlations above 0.8 between the estimates and original subspaces were obtained for moderately overdetermined mixtures (constructed for example, for $\mathbf{X}_1$, by mixing $M_1 + M_c$ sources), while for 0 dB, equivalent performances can only be obtained if the the total original dimensions remains low (for one 0 dB supplementary source for each data set, the correlation coefficient $CC_{c_e} = 0.82$, rightmost element of the second row of the third subfigure on row 1). These results are valid for the ER automatic subspace dimension detection and they are lower for the fixed threshold[7].

An example of the different estimated signals is given Figure 2. For this example, the eigenvalues $\boldsymbol{\lambda}$ (27) are plotted Figure 3.

If the figure 1 presents detailed CSSD results over different combinations of SNR and original space dimension, we also summarize averaged results in tables 1 and 2, in order to compare the CSSD based reconstructions to those obtained using concurrent algorithms such as JIVE or DISCO-SCA (as explained above, only the first SVD based step of these methods is relevant for this comparison). The two tables present two complementary points of view.

The first one neglects the effect of the SNR by averaging over all specific subspaces powers, but illustrates the effect of the dimension of these original specific subspaces (how many specific sources are indeed interfering with the common ones). As it can be seen, if the number of interfering specific sources is high, all algorithms are equivalent, regardless of the method of estimation of the common subspace dimension (bottom rows of the table). Actually, only if the dimension of the common subspace is correctly informed (columns indexed wih o), CSSD and JIVE/DISCO-SCA improve the estimation of the common signals with respect to the raw data, with a consistently better CSSD

[7] If the threshold is adapted to the *a priori* known SNR (eq. 22), the results come very close to the informed dimension version of the method, $CC_{c,o}$ and $CC_{s,o}$. Including an estimate of the noise power in the method is thus a promising possibility, but it is outside the scope of this paper.



(a) $\mathbf{X}_{1c}$, $\hat{\mathbf{X}}_{1c,\rho}$, $\hat{\mathbf{X}}_{1c,e}$ (b) $\mathbf{X}_{1s}$, $\hat{\mathbf{X}}_{1s,\rho}$, $\hat{\mathbf{X}}_{1s,e}$ (c) $\mathbf{X}_1$

(d) $\mathbf{X}_{2c}$, $\hat{\mathbf{X}}_{2c,\rho}$, $\hat{\mathbf{X}}_{2c,e}$ (e) $\mathbf{X}_{2s}$, $\hat{\mathbf{X}}_{2s,\rho}$, $\hat{\mathbf{X}}_{2s,e}$ (f) $\mathbf{X}_2$

Figure 2: Example of simulated signals $\mathbf{X}_1$ and $\mathbf{X}_2$ (dimensions $M_1 = 7$ and $M_2 = 5$). The total dimension of the original source space was set to 17 (common source space of dimension $M_c = 3$, specific source spaces for $\mathbf{X}_1$ and $\mathbf{X}_2$ of dimensions $M_{1s} = M_{2s} = 7$). The SNR (common/specific) was set to 3dB (for more details, see simulation description in text). Estimated common and specific parts are plotted in subfigures (a), (b) for $\mathbf{X}_1$ and (d), (e) for $\mathbf{X}_2$. For this example, the automatic eigenvalues ratio procedure (19) yielded the correct common subspace dimension $\hat{M}_c = 3$, while the fixed threshold procedure selected only the first eigenvalue (see fig 3). The mean correlation coefficients were $CC_{1c,e} = 0.96$, $CC_{1c,\rho} = 0.80$, $CC_{1s,e} = 0.92$, $CC_{1s,\rho} = 0.78$, $CC_{2c,e} = 0.95$, $CC_{2c,\rho} = 0.70$, $CC_{2s,e} = 0.92$, $CC_{2s,\rho} = 0.80$. Subfigures (c) and (f) illustrate the two signal sets, sums of (a)+(b) and (d)+(e) respectively.

(and better JIVE estimates on normalized data than on raw data). On the other hand, if the M_c estimate fails (which is the case when averaging over all SNRs and for highly dimensional specific subspaces, *i.e.*, in a big right lower corner of the table), all algorithms fail, with correlation coefficients below those in the first column. CSSD keeps though a small advantage when the number of specific sources remain relatively low, regardless of the SNR (upper right corner). Interestingly, applying the SVD-only initial step of JIVE/DISCO-SCA yields better performances when the number of specific perturbing sources increases (column $\mathrm{CC_{c,o,jn}}$): the higher the number of specific signals in the mixture, the closer the specific part comes to white noise and the easier the common sources are isolated on the first singular vectors.

The second table neglects the effect of the total subspace dimensionality by averaging, but illustrates the effect of the SNR. As it can be seen, if the SNR is high, all algorithms perform correctly, except for the fixed threshold version (last three columns), when for noises starting from a 10 dB SNR the *CC* falls below the *CC* between the simulated common signals $\mathbf{X}_c$ and the raw mixed ones $\mathbf{X}$ (*i.e.*, below the value in the first column). But again, if M_c is correctly informed,, CSSD is consistently equivalent or it outperforms JIVE/DISCO-SCA for all SNRs, including 3 and 0 dB.

Finally, a third table 3 compares the CSSD and JIVE/DISCO-SCA using the well-conditioned noisy model proposed in [14]. More precisely, the rank of the common plus the rank of the specific subspaces equals the rank of the simulated datasets (this situation corresponds to the first row of table 1). The norms of the common and specific components for each signal (each row of $\mathbf{X}_1$ and $\mathbf{X}_2$ were equal (0 dB in the previous simulation) and white noise was added directly to the measurements $\mathbf{X}_i$, with different signal to noise ratios (Inf, 20, 10, 3 and 0 dB). Only the informed rank decomposition was performed (without using any model dimension estimation). As it can be seen, CSSD largely dominates stacked data SVD methods [14, 15]. A possible explanation stems from another feature of the considered model: in our simulation, specific subspaces are (weakely) correlated among them, while in [14] (Supplementary Material), the row-space intersection of specific subspaces is assumed null, which in the two datasets situation is implies orthogonal specific subspaces.

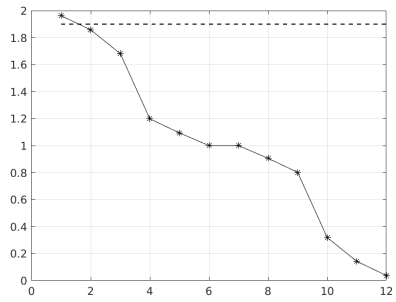


Figure 3: Example of eigenvalues $\boldsymbol{\lambda}$ (27) corresponding to the principal components $\mathbf{W}$ from Proposition 1. For the simulation details, see text and legend of Figure 2. The dashed line represents the fixed threshold $\rho = 1.9C$, which selects a monodimensional subspace $\mathbf{Y}_c$ ($\hat{M}_c = 1$), while the ER test based procedure (section 2.3) detects the correct common space dimension $\hat{M}_c = 3$. The resulted reconstructed signals are given Figure 2

Dim	$CC_{c,m}$	$CC_{c,o}$	$CC_{c,o,j}$	$CC_{c,o,jn}$	$CC_{c,e}$	$CC_{c,e,j}$	$CC_{c,e,jn}$	$CC_{c,\rho}$	$CC_{c,\rho,j}$	$CC_{c,\rho,jn}$
9	0.86	**1.00**	0.88	0.89	**1.00**	0.88	0.89	**1.00**	0.88	0.89
11	0.85	**0.98**	0.89	0.90	**0.93**	0.85	0.88	**0.88**	0.82	0.86
13	0.85	**0.97**	0.89	0.91	**0.88**	0.83	0.85	**0.80**	0.79	0.79
15	0.85	**0.96**	0.90	0.92	**0.83**	0.79	0.82	0.75	0.73	0.75
17	0.85	**0.95**	0.90	0.92	**0.76**	0.74	0.75	0.69	0.68	0.69
19	0.85	**0.95**	0.90	0.93	0.74	0.72	0.74	0.68	0.67	0.68
21	0.85	**0.94**	0.90	0.93	0.66	0.65	0.66	0.62	0.61	0.62
23	0.85	**0.94**	0.90	0.93	0.66	0.65	0.66	0.61	0.61	0.61
25	0.85	**0.94**	0.91	0.93	0.62	0.61	0.62	0.58	0.58	0.58
27	0.85	**0.94**	0.91	0.93	0.63	0.62	0.63	0.59	0.58	0.59
29	0.85	**0.94**	0.91	0.93	0.61	0.60	0.61	0.57	0.56	0.57
31	0.85	**0.94**	0.91	0.93	0.60	0.59	0.60	0.56	0.56	0.57
33	0.85	0.94	0.91	0.94	0.57	0.57	0.57	0.54	0.54	0.54

Table 1: Correlation coefficients *CC* between the simulated signals $\mathbf{X}_c$ (obtained from simulated common sources) and their estimates $\hat{\mathbf{X}}_c$, averaged over all SNRs and the two data sets $\mathbf{X}_1$ and $\mathbf{X}_2$. The first column indicates the total dimension of the source space (see also main text and the results from figure 1). The second column displays the *CC* between the simulated signals and the raw data. Next, by groupes of three columns, we show the results obtained for different estimation methods of the common subspace dimension (informed: index *o*; eigenvalues ratio: index *e*; fixed threshold: index ρ). The three columns inside each group correspond to the evaluated algorithm (first column: CSSD, no index; second column: JIVE/DISCO-SCA on raw data; third column: JIVE/DISCO-SCA on normalized data

SNR	$CC_{c,m}$	$CC_{c,o}$	$CC_{c,o,j}$	$CC_{c,o,jn}$	$CC_{c,e}$	$CC_{c,e,j}$	$CC_{c,e,jn}$	$CC_{c,\rho}$	$CC_{c,\rho,j}$	$CC_{c,\rho,jn}$
20	0.99	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
10	0.94	0.99	0.98	0.99	0.97	0.97	0.97	0.92	0.92	0.93
3	0.79	**0.94**	0.87	0.91	0.63	0.59	0.61	0.52	0.50	0.52
0	0.68	**0.89**	0.75	0.80	0.32	0.25	0.27	0.29	0.23	0.25

Table 2: Correlation coefficients *CC* between the simulated signals $\mathbf{X}_c$ (obtained from simulated common sources) and their estimates $\hat{\mathbf{X}}_c$, averaged over all source space dimensions and the two data sets $\mathbf{X}_1$ and $\mathbf{X}_2$.

SNR	$CC_{c,m}$	$CC_{c,o}$	$CC_{c,o,j}$	$CC_{c,o,jn}$
Inf	0.69	**1.00**	0.70	0.70
20	0.69	**0.98**	0.70	0.70
10	0.66	**0.92**	0.69	0.69
3	0.56	**0.81**	0.63	0.65
0	0.48	**0.73**	0.56	0.60

Table 3: Correlation coefficients *CC* between the simulated signals $\mathbf{X}_c$ and their estimates $\hat{\mathbf{X}}_c$ for well conditioned noisy setups (see text), averaged over the two data sets $\mathbf{X}_1$ and $\mathbf{X}_2$.

3.2. Application on real signals

Different applications can be imagined for the presented method in analyzing electro-physiological signals. We focus here on brain electrophysiological data (EEG and SEEG, see below). Electroencephalography (EEG) is one of the most widely used modalities to explore the electrical brain activity, using sensors placed over the scalp. Its invasive counterpart, stereo-electroencephalography (SEEG), is used in clinical routine for exploring the deep brain structures of drug-resistant epileptic patients with the help of intra-cerebral electrodes, consisting in 10 to 15 sensors equally spaced (2.5 mm) on a 0.8 mm diameter shaft, surgically implanted through the skull in the brain. These investigations help in localizing and delimiting the epileptic zone to be removed by subsequent surgery. In parallel to this primary clinical objective, EEG/SEEG recordings can also be used to explore the brain responses in controlled conditions, related to a specific tasks (cognitive or motor) or to a provoked stimulus. The stimuli can be visual, tactile, etc ..., but also electrical. For example, during clinical investigations, epileptic events as well as cortico-cortical evoked potentials (CCEP) can be also induced by Trans- or Intra-Cranial/ Cerebral Stimulation (ICS) [30].

All the data used here to illustrate our method was recorded at the University Hospital (CHU) Nancy, France, during routine epilepsy evaluation using a Micromed® (Italy) system. The patients gave their informed consent.

Two application examples of CSSD decomposition are presented, on two types of data:

- surface EEG data, with eye blink and muscle artefacts. The goal of this application is to cancel the extracerebral artefacts contaminating the signals (for other techniques, see *e.g.*, [31, 32, 33, 34]). The two datasets are sets of distant electrodes, placed over the two hemispheres. The hypothesis is that extracerebral artefacts contaminate both sets, while intracerebral sources, although possibly correlated, are not common, because they originate from different hemispheres;

- SEEG data recorded during electrical intra-cerebral stimulation (ICS) in two different conditions of the patient (awake and sleeping). In some cases (at specific locations and for specific amplitudes), the ICS generates a deterministic activation of the neurons in its cortical target and in connected areas [30]. It has been shown that these responses differ between non-rapid eye movement (NREM) sleep and wakefulness [35]. According to [36], a possible explanation is the connectivity breakdown during sleep, compared to wakefulness, when a complex chain of causal interactions is triggered and propagates through a distributed network of cortical areas. The goal of this application is thus to determine if the responsive brain areas differ between the two conditions.

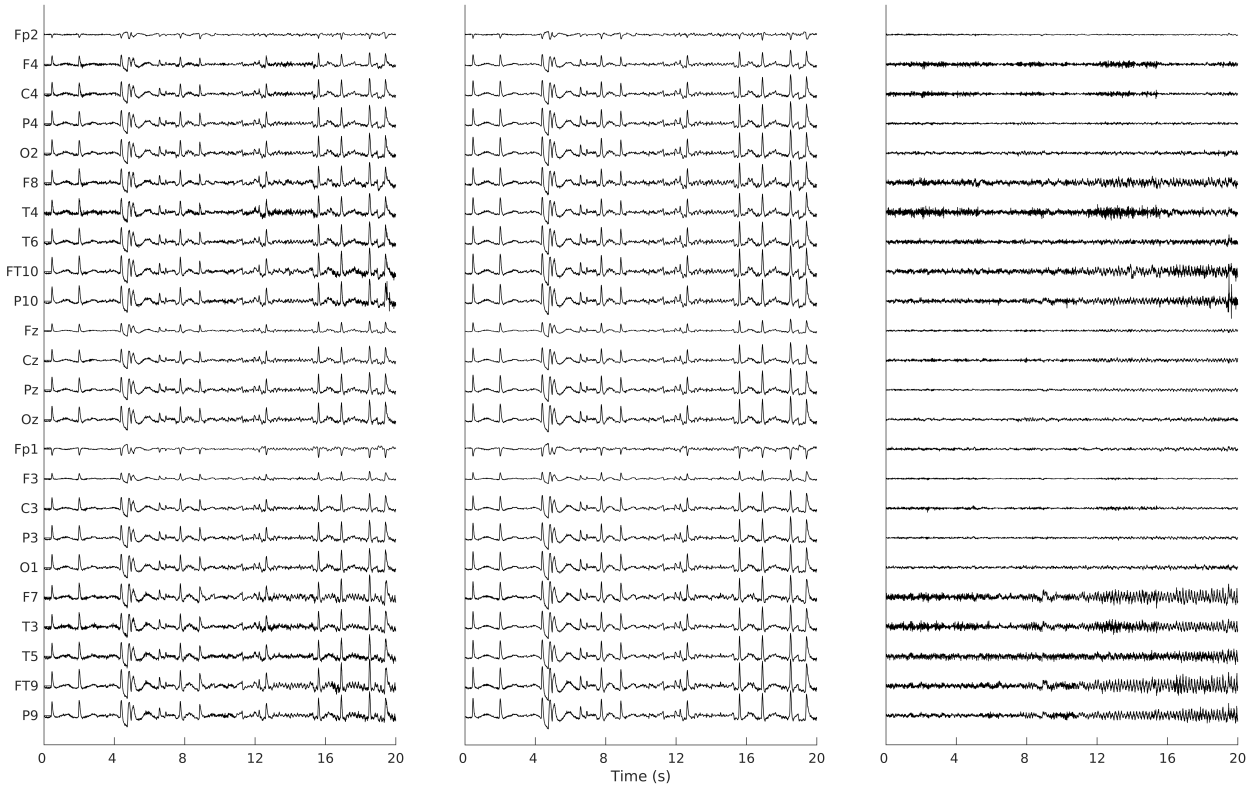
3.2.1. Artefacted EEG data

We present here an example of a 20 seconds surface EEG consisting of 24 electrodes recorded at a sampling frequency of 256 Hz. The raw data is presented figure 4, left panel. We have considered the first 10 electrodes (even numbers, right hemisphere) as $\mathbf{X}_1$ and the last 10 as $\mathbf{X}_2$ (odd numbers, left hemisphere). We applied CSSD with an automatic detection of the common space dimension (19) in order to obtain the common space of the two data sets, deemed as artefacts and displayed in figure 4, central panel (the detected dimension was $\hat{M}_c = 2$). Next, a second CSSD was applied considering the raw EEG as $\mathbf{X}_1$ and the artefacts as $\mathbf{X}_2$ (in this case, obviously $\mathbf{X}_2$ is included in $\mathbf{X}_1$). We obtained thus a specific subspace $\mathbf{X}_{1s}$ containing cleaned signals, shown in figure 4, right panel. The procedure, fully automatic and fast, allows to uncover the epileptic seizure, involving mainly the left hemisphere and starting around second 10. Extensive tests of the quality of the cleaned data are outside the scope of this paper, but it is clear that the CSSD based procedure can at least be used as a first attempt of data cleaning in order to ease the medical interpretation.

3.2.2. SEEG during ICS

The signals were collected from a patient implanted with 13 SEEG electrodes, 11 in the right hemisphere and 2 in the left hemisphere. In all, 115 sensors were placed inside different brain structures. All signals were recorded at a sampling frequency of 1024 Hz. The employed ICS consisted in a biphasic current pattern of very short duration (1.05 ms), having an amplitude of 1 mA and injected periodically with a 1 Hz frequency during 30 s[8].

[8] For detailed models of this type of current stimulation, both from a time and frequency point of view, see [37, 38, 39].

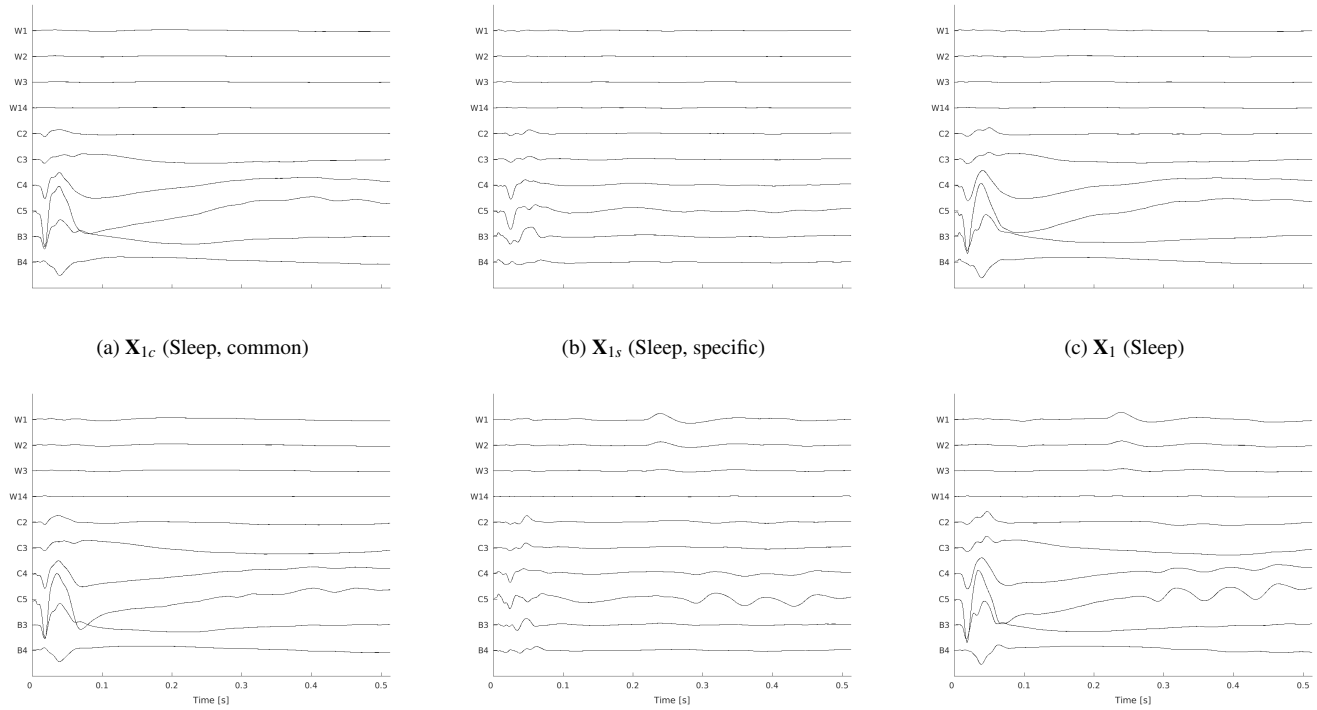


(a) $\mathbf{X}_1$: raw EEG (b) $\mathbf{X}_2$: artefact signals (reconstructed) (c) $\mathbf{X}_{1s}$: artefact free EEG

Figure 4: Artefact cancelling using CSSD (for the details of the procedure see main text)

We focused on two data sets, consisting in similar location and amplitude ICS for NREM sleep and wakefulness states. The ICS was applied in the posterior hippocampus, between B1-2 sensors. In order to avoid the ICS artifact effects on the analysis of the CCEP, previous studies [30] focused on a time interval starting 15 ms after the current injection. In our case, we canceled the artifact separately for each data-set by as described in [10]. Next, we applied the proposed approximated CSSD approach between the two data-sets (sleep and wake signals) in order to identify common and specific activities.

As the cortical responses (CCEP) were in principle deterministic, we extracted windows of one second length, centered on the stimulation instant (30 windows, as each session consists in 30 ICS pulses delivered with a period of one second). As in BCI applications, these "single trials" were next averaged to obtain averaged versions of the CCEP. These averaged versions (on sensors of interest) are plotted in Figure 5. As it can be seen, the stimulation triggered rapid activities both in sleep and awake states, from 10 to 50ms after the stimulation, spreading over a large number of electrodes (especially powerful on the sensor C5 - anterior hippocampus). During awake state, a later evoked activity was observed over the profound sensors of the sensor W1 (posterior cingulate, 240 ms latency), as well as oscillating activities (high alpha / low beta rythms) on the sensor C5 (300 ms latency). Inversely during the NREM sleep state, these activities were not observed.



(a) $\mathbf{X}_{1c}$ (Sleep, common) (b) $\mathbf{X}_{1s}$ (Sleep, specific) (c) $\mathbf{X}_1$ (Sleep)

(d) $\mathbf{X}_{2c}$ (Awake, common) (e) $\mathbf{X}_{2s}$ (Awake, specific) (f) $\mathbf{X}_2$ (Awake)

Figure 5: Common (dimension $M_c = 9$) and specific subspaces of NREM sleep (up) and wake (down) signals. Estimated common and specific parts are plotted in subfigures (a), (b) for sleep and (d), (e) for awake states. Subfigures (c) and (f) plot the raw signals (averaged). See discussion in the main text

4. Conclusion

This paper proposes a modified CCA approach called CSSD, whose aim is to automatically identify common and specific parts of two sets of signals. More precisely, the common part is defined by a basis of a common subspace, bisector between the subspaces of the two original data-sets, whose dimension can be automatically estimated using an eigenvalues ratio (ER) test. The main novelty of the proposed method is a "stacked basis" approach, different from the "stacked data" approach proposed by other methods aiming to estimate common components of several sets of signals such as JIVE [14] and DISCO-SCA [15].

We illustrate, on simulated signals, the performances of the method and compare them with these methods. The results show that, for two data sets, CSSD is at least equal or better than the concurrent methods for estimating the common and specific parts of the two sets of signals and it is able to deal with difficult situations, when specific subspaces are correlated. We present next two examples of application, the first one being a fully automatic fast extracerebral artefact rejection for EEG signals, while the second one analyses SEEG signals recorded in two conditions (sleep vs. wake) during an intra-cerebral electrical stimulation protocol, the aim being to discover sleep vs. wake common and specific activated brain regions.

Possible other applications will be explored elsewhere, namely for example common and specific parts between simultaneously recorded signals using two modalities (*e.g.*, SEEG / EEG / MEG) or two signals subsets. Besides answering quantitatively to difficult questions such as which are the activities visible to two different modalities, assessing common activities could be a necessary pre-processing step for multi-modality localization problems by inverse problem solving.

Finally, we most note that the application that initially motivated our work (electrophysiological brain signal analysis) uses linear models. This is motivated by physical considerations (quasi-static Maxwell equations as they are

applied to model electrical fields of the brain). Nevertheless, one can imagine non-linear relations between different modalities (for example fMRI and (S)EEG). Another interesting application could be comparing two datasets when the positions of the measuring electrodes are not exactly controlled or some uncontrolled physiological modifications have appeared (for example between recordings taken during time-distant sessions). A promising research direction could be to adapt nonlinear techniques such as Kernel-CCA or Deep CCA [17, 18] for estimating an approximate "nonlinear bisector space".

References

[1] Li, Y.O., Adali, T., Wang, W., Calhoun, V.D.: 'Joint blind source separation by multiset canonical correlation analysis', *IEEE Transactions on Signal Processing*, 2009, **57**, (10), pp. 3918–3929

[2] Lahat, D., Adali, T., Jutten, C.: 'Multimodal data fusion: An overview of methods, challenges, and prospects', *Proceedings of the IEEE*, 2015, **103**, (9), pp. 1449–1477

[3] Hotelling, H.: 'Relations between two sets of variates', *Biometrika*, 1936, **28**, (3/4), pp. 321–377

[4] Kettenring, J.R.: 'Canonical analysis of several sets of variables', *Biometrika*, 1971, **58**, (3), pp. 433–451

[5] Kim, T., Eltoft, T., Lee, T.W. In: Rosca, J., Erdogmus, D., Príncipe, J.C., Haykin, S., editors. 'Independent vector analysis: An extension of ica to multivariate components'. (Berlin, Heidelberg: Springer Berlin Heidelberg, 2006. pp. 165–172

[6] Adali, T., Anderson, M., Fu, G.S.: 'Diversity in independent component and vector analyses: Identifiability, algorithms, and applications in medical imaging', *IEEE Signal Processing Magazine*, 2014, **31**, (3), pp. 18–33

[7] Lin, Z., Zhang, C., Wu, W., Gao, X.: 'Frequency recognition based on canonical correlation analysis for SSVEP-based BCIs', *IEEE Transactions on Biomedical Engineering*, 2007, **54**, (6), pp. 1172–1176

[8] Nakanishi, M., Wang, Y., Wang, Y.T., Jung, T.P.: 'A comparison study of canonical correlation analysis based methods for detecting steady-state visual evoked potentials', *PLOS ONE*, 2015, **10**, (10), pp. 1–18

[9] Spüler, M., Walter, A., Rosenstiel, W., Bogdan, M.: 'Spatial filtering based on canonical correlation analysis for classification of evoked or event-related potentials in EEG data', *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 2014, **22**, (6), pp. 1097–1103

[10] Bhattacharyya, A., Ranta, R., Le Cam, S., Louis.Dorr, V., Tyvaert, L., Colnat.Coulbois, S., et al.: 'A multi-channel approach for cortical stimulation artefact suppression in depth eeg signals using time-frequency and spatial filtering', *IEEE Transactions on Biomedical Engineering*, 2018, pp. 1–1. in print

[11] Mosher, J.C., Lewis, P.S., Leahy, R.M.: 'Multiple dipole modeling and localization from spatio-temporal MEG data', *IEEE Transactions on Biomedical Engineering*, 1992, **39**, (6), pp. 541–557

[12] Mosher, J.C., Leahy, R.M.: 'Recursive MUSIC: A framework for EEG and MEG source localization', *IEEE Transactions on Biomedical Engineering*, 1998, **45**, (11), pp. 1342–1354

[13] Ermer, J.J., Mosher, J.C., Huang, M., Leahy, R.M.: 'Paired MEG data set source localization using recursively applied and projected (RAP) MUSIC', *IEEE Transactions on Biomedical Engineering*, 2000, **47**, (9), pp. 1248–1260

[14] Lock, E.F., Hoadley, K.A., Marron, J.S., Nobel, A.B.: 'Joint and individual variation explained (JIVE) for integrated analysis of multiple data types', *The annals of applied statistics*, 2013, **7**, (1), pp. 523–542

[15] Schouteden, M., Van.Deun, K., Wilderjans, T.F., Van.Mechelen, I.: 'Performing DISCO-SCA to search for distinctive and common information in linked data', *Behavior Research Methods*, 2014, **46**, pp. 576–587

[16] Adali, T., Akhonda, M.A.B.S., Calhoun, V.D.: 'Ica and iva for data fusion: An overview and a new approach based on disjoint subspaces', *IEEE Sensors Letters*, 2019, **3**, (1), pp. 1–4

[17] Hardoon, D.R., Szedmak, S., Shawe.Taylor, J.: 'Canonical Correlation Analysis: An Overview with Application to Learning Methods', *Neural Computation*, 2004, **16**, (12), pp. 2639–2664. Available from: https://doi.org/10.1162/0899766042321814

[18] Andrew, G., Arora, R., Bilmes, J., Livescu, K. 'Deep canonical correlation analysis'. In: Dasgupta, S., McAllester, D., editors. Proceedings of the 30th International Conference on Machine Learning. vol. 28 of *Proceedings of Machine Learning Research*. (Atlanta, Georgia, USA: PMLR, 2013. pp. 1247–1255. Available from: http://proceedings.mlr.press/v28/andrew13.html

[19] Ahn, S.C., Horenstein, A.R.: 'Eigenvalue ratio test for the number of factors', *Econometrica*, 2013, **81**, (3), pp. 1203–1227

[20] Cichocki, A., Amari, S.: 'Adaptive Blind Signal and Image Processing Learning Algorithms and Applications'. (John Wiley & Sons, New York, USA, 2002)

[21] Uurtio, V., Monteiro, J.a.M., Kandola, J., Shawe.Taylor, J., Fernandez.Reyes, D., Rousu, J.: 'A tutorial on canonical correlation methods', *ACM Comput Surv*, 2017, **50**, (6). Available from: https://doi.org/10.1145/3136624

[22] Golub, G.H., Loan, C.F.V.: 'Matrix Computation'. (John Hopkins University Press, 1996)

[23] Kim, T.K., Kittler, J., Cipolla, R.: 'Discriminative learning and recognition of image set classes using canonical correlations', *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2007, **29**, (6), pp. 1005–1018

[24] Sakar, C.O., Kursun, O.: 'Discriminative feature extraction by a neural implementation of canonical correlation analysis', *IEEE Transactions on Neural Networks and Learning Systems*, 2017, **28**, (1), pp. 164–176

[25] Bro, R., Acar, E., Kolda, T.G.: 'Resolving the sign ambiguity in the singular value decomposition', *Journal of Chemometrics*, 2008, **22**, (2), pp. 135–140

[26] Wax, M., Kailath, T.: 'Detection of signals by information theoretic criteria', *IEEE Transactions on Acoustics, Speech and Signal Processing*, 1985, **33**, (2), pp. 387–392

[27] Williams, D.B.: 'Counting the degrees of freedom when using AIC and MDL to detect signals', *IEEE Transactions on Signal Processing*, 1994, **42**, (11), pp. 3282–3284

[28] Song, Y., Schreier, P.J., Ramírez, D., Hasija, T.: 'Canonical correlation analysis of high-dimensional data with very small sample

support', *Signal Processing*, 2016, **128**, pp. 449 – 458. Available from: http://www.sciencedirect.com/science/article/pii/S0165168416300834
[29] Lameiro, C., Hasija, T., Marrinan, T., Schreier, P.J. 'Estimating the number of correlated components based on random projections'. In: ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). (, 2019. pp. 5152–5156
[30] Krieg, J., Koessler, L., Jonas, J., Colnat.Coulbois, S., Vignal, J.P., Bénar, C.G., et al.: 'Discrimination of a medial functional module within the temporal lobe using an effective connectivity model: A CCEP study', *NeuroImage*, 2017, **161**, (Supplement C), pp. 219 – 231
[31] LeVan, P., Urrestarazu, E., Gotman, J.: 'A system for automatic artifact removal in ictal scalp EEG based on independent component analysis and Bayesian classification', *Clinical Neurophysiology*, 2006, **117**, (4), pp. 912–927
[32] Krishnaveni, V., Jayaraman, S., Gunasekaran, A., Ramadoss, K.: 'Automatic Removal of Ocular Artifacts using JADE Algorithm and Neural Network', *International Journal of Intelligent Systems and Technologies*, 2006, **1**, (4), pp. 322–333
[33] Shao, S., Shen, K., Ong, C., Wilder.Smith, E., Li, X.: 'Automatic EEG Artifact Removal: A Weighted Support-Vector-Machine Approach With Error Correction', *IEEE Transactions on Biomedical Engineering*, 2009, **56**, (2), pp. 336 – 344
[34] Romo.Vázquez, R., Vélez.Pérez, H., Ranta, R., Dorr, V.L., Maquin, D., Maillard, L.: 'Blind source separation, wavelet denoising and discriminant analysis for EEG artefacts and noise cancelling', *Biomedical Signal Processing and Control*, 2012, **7**, (4), pp. 389 – 400
[35] Pigorini, A., Sarasso, S., Proserpio, P., Szymanski, C., Arnulfo, G., Casarotto, S., et al.: 'Bistability breaks-off deterministic responses to intracortical stimulation during non-REM sleep', *Neuroimage*, 2015, **112**, pp. 105–113
[36] Massimini, M., Ferrarelli, F., Huber, R., Esser, S.K., Singh, H., Tononi, G.: 'Breakdown of cortical effective connectivity during sleep', *Science*, 2005, **309**, (5744), pp. 2228–2232
[37] Hofmanis, J., Caspary, O., Louis.Dorr, V., Ranta, R., Maillard, L.: 'Denoising depth EEG signals during DBS using filtering and subspace decomposition', *IEEE Transactions on Biomedical Engineering*, 2013, **60**, (10), pp. 2686–2695
[38] Trebaul, L., Rudrauf, D., Job, A.S., Mălîia, M.D., Popa, I., Barborica, A., et al.: 'Stimulation artifact correction method for estimation of early cortico-cortical evoked potentials', *Journal of Neuroscience Methods*, 2016, **264**, pp. 94 – 102
[39] Ranta, R., Le Cam, S., Tyvaert, L., Louis.Dorr, V.: 'Assesing human brain impedance using simultaneous surface and intracerebral recordings', *Neuroscience*, 2017, **343**, pp. 411–422

Appendix

We give here a concise version of a result from [4] and a modified proof.

Lemma (Kettenring). Consider the symmetric matrix

$$\mathbf{M} = \begin{bmatrix} \mathbf{I}_{M_1} & \mathbf{\Sigma} \\ \mathbf{\Sigma}^T & \mathbf{I}_{M_2} \end{bmatrix} \tag{24}$$

with $\mathbf{\Sigma}$ a $M_1 \times M_2$, $(M_1 \geq M_2)$ matrix having r non-null sorted diagonal elements $1 \geq \sigma_i > 0$ for all i, $1 \leq i \leq r \leq M_2$ $(\sigma_i \geq \sigma_{i+1})$. Its first r eigenvalues λ_i $(i = 1..r)$ are given by $1+\sigma_i$. The corresponding eigenvectors $\mathbf{v}_i$ have only two non-null elements, $v_i(i) = v_i(M_1+i) = \frac{\sqrt{2}}{2}$.

Proof First, use the fact that adding an identity matrix to a square matrix adds 1 to its eigenvalues. We need then to compute the eigenvalues μ of the simpler matrix

$$\mathbf{Q} = \begin{bmatrix} \mathbf{0}_{N_1} & \mathbf{\Sigma} \\ \mathbf{\Sigma}^T & \mathbf{0}_{N_2} \end{bmatrix} \tag{25}$$

which are given by roots of the characteristic polynomial

$$\begin{aligned} P(\mu) &= \det \begin{bmatrix} \mu\mathbf{I}_{M_1} & -\mathbf{\Sigma} \\ -\mathbf{\Sigma}^T & \mu\mathbf{I}_{M_2} \end{bmatrix} = \mu^{M_1} \det\left(\mu\mathbf{I}_{M_2} - \frac{1}{\mu}\mathbf{\Sigma}^T\mathbf{\Sigma}\right) \\ &= \mu^{(M_1+M_2-2r)} \prod_{i=1}^{r} (\mu^2 - \sigma_i^2) \end{aligned} \tag{26}$$

In the last derivation, we used the block determinant property:

$$\det \begin{bmatrix} \mathbf{A} & \mathbf{B} \\ \mathbf{C} & \mathbf{D} \end{bmatrix} = \det(\mathbf{A}) \det(\mathbf{D} - \mathbf{C}\mathbf{A}^{-1}\mathbf{B})$$

Adding 1 to the μ_i roots will give the desired eigenvalues λ_i, grouped below in a vector $\boldsymbol{\lambda}$:

$$\boldsymbol{\lambda} = [\underbrace{1+\sigma_1 \ldots 1+\sigma_r}_{r} \; \underbrace{1 \ldots 1}_{M_1+M_2-2r} \; \underbrace{1-\sigma_r \ldots 1-\sigma_1}_{r}] \tag{27}$$

Computing a solution for the eigenvectors is straightforward. Consider a $\mathbf{v}_i$ among the first r eigenvectors ($1 \leq i \leq r$, corresponding to $\lambda_i = 1 + \sigma_i$). Then,

$$\begin{aligned} (\lambda_i \mathbf{I}_{M_1+M_2} - \mathbf{M})\mathbf{v}_i &= \mathbf{0} \Leftrightarrow \\ (\sigma_i \mathbf{I}_{M_1+M_2} - \mathbf{Q})\mathbf{v}_i &= \mathbf{0} \end{aligned} \tag{28}$$

Taking row i will lead to $\sigma_i \mathbf{v}_i(i) - \sigma_i \mathbf{v}_i(M_1 + i) = 0$, thus

$$\mathbf{v}_i(i) = \mathbf{v}_i(M_1 + i) \tag{29}$$

Applying the same approach for the last r eigenvectors ($M_1 + M_2 - r + 1 \leq i \leq M_1 + M_2$) will lead to

$$\mathbf{v}_i(i) = -\mathbf{v}_i(M_1 + i) \tag{30}$$

Finally, for $\lambda_i = 1$, we'll have

$$\mathbf{Q}\mathbf{v}_i = \mathbf{0}, \tag{31}$$

which implies that the first and the last r elements of $\mathbf{v}_i$ are 0 and the middle elements, between $r+1$ and $M_1 + M_2 - r$ are undetermined.

A direct solution to the eigenvectors problem respecting all the previously listed constraints (29), (30) and (31) is given by the following: for the first and the last r vectors, fix all the elements except $\mathbf{v}_i(i)$ and $\mathbf{v}_i(M_1 + i)$ to 0. As their norm is by definition 1, the solution is given by $\mathbf{v}_i(i) = \frac{\sqrt{2}}{2}$. For the remaining eigenvectors $\mathbf{v}_i$ (i between $r+1$ and $M_1 + M_2 - r$), it is sufficient to choose one non null element equal to 1, ensuring that these non null elements are not on the same row. The complete eigenvector matrix $\mathbf{V}$ is given by:

$$\mathbf{V} = \begin{bmatrix} \frac{\sqrt{2}}{2}\mathbf{I}_r & \mathbf{0}_{r,M_1-r} & \mathbf{0}_{r,M_2-r} & \frac{\sqrt{2}}{2}\mathbf{I}_r \\ \mathbf{0}_{M_1-r,r} & \mathbf{I}_{M_1-r} & \mathbf{0}_{M_1-r,M_2-r} & \mathbf{0}_{M_1-r,r} \\ \frac{\sqrt{2}}{2}\mathbf{I}_r & \mathbf{0}_{r,M_1-r} & \mathbf{0}_{r,M_2-r} & -\frac{\sqrt{2}}{2}\mathbf{I}_r \\ \mathbf{0}_{M_2-r,r} & \mathbf{0}_{M_2-r,M_1-r} & \mathbf{I}_{M_2-r} & \mathbf{0}_{M_2-r,r} \end{bmatrix} \tag{32}$$

which concludes the proof.