Deep L1-PCA of Time-Variant Data with Application to Brain Connectivity Measurements

Giovanna Orrù[1], Tiziana Cattai[1,2,3], Stefania Colonnese[1], Gaetano Scarano[1],
Fabrizio De Vico Fallani[2,3], Panos Markopoulos[4], Dimitris Pados[5]
[1]*Dept. of Information Engineering, Electronics and Telecommunication, Sapienza University of Rome, Italy*,
[2]*Inria Paris,Paris, France*
[3]*Institut du Cerveau et de la Moelle Épinière, ICM, Inserm, U 1127, CNRS, UMR 7225, Sorbonne Université, F-75013, Paris, France*
[4]*Dept. of Electrical and Microelectronic Engineering, Rochester Institute of Technology, Rochester,NY 14623 USA*
[5]*Dept. of Computer and Electrical Engineering & Computer Science, Florida Atlantic University, Boca Raton, FL 33431, USA*
Corresponding author: stefania.colonnese@uniroma1.it

***Abstract*—L1-Principal Component Analysis (L1-PCA) is a powerful computational tool to identify relevant components in data affected by noise, outliers, partial disruption and so on. Relevant efforts have been made to adapt its powerful summarization capacity to time variant data, e.g. in tracking the evolution of L1-PCA components. Here, we analyze a layered version of L1-PCA, to which we refer to as Deep L1-PCA. Deep L1-PCA is obtained by recursive application of two stages: estimation of L1-PCA basis and extraction of the first rank projector. The Deep L1-PCA is applied to repeated EEG connectivity measures and it proves relevant for identifying outliers, changes, and stable components. Moreover, at each layer, an in-depth analysis of the mean square error between the data applied at the input layer and the output projector is provided. The Deep L1-PCA allows to cope with outliers of different temporal extent as well as to extract the relevant common component at a reduced computational cost.**

***Index Terms*—L1-norm, PCA, outliers, first rank component extraction, tensor-based representation of biomedical data**

I. Introduction

L1-Principal Component Analysis (L1-PCA) is a powerful computational tool to identify relevant components in data affected by noise, outliers, and partial disruption [1]. Relevant efforts have been made to adapt its powerful summarization capacity to time variant data, e.g. in tracking the evolution of L1-PCA components [2]. Tracking usually applies a moving window and it presents computational advantages since only the innovation components are estimated. Nonetheless, this approach requires a preliminary selection of the window width, which is set according to the time duration of the transitory phenomena to be discarded. If the whole dataset is available, a different hierarchical approach can be applied, referred to as Deep L1-PCA in the following. A deep computational architecture is exploited in [3] to extract Euclidean norm PCA features for face recognition purposes. Therein, the impact of the layering parameters (depth, extent) are not discussed. In this paper we present a Deep L1-PCA obtained by recursive application of data partitioning and L1-PCA analysis with first rank ($K = 1$) component extraction : $\mathbf{R} = \underset{RR^T=I_K}{\operatorname{argmax}} \|X^T U_{:,1:K}\|_1, \quad K = 1.$

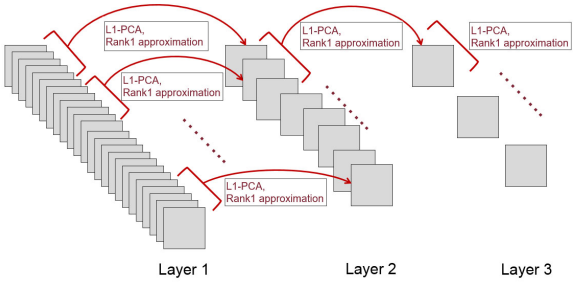

Fig. 1. Deep L1-PCA computation structure.

An in-depth analysis of the structure of the approximation error within each layer and among different layers shows that Deep L1-PCA copes with outliers of different extent while extracting the relevant common component at a reduced computational cost. The approach is applied to repeated brain connectivity measurements and it proves relevant for identifying stable components, outliers, and minimum required number of measurements while avoiding any normalization pre-processing step.

II. Deep L1-PCA Computing Architecture

Before turning to mathematics, herein we illustrate the core of the Deep L1-PCA computational architecture. In a nutshell, at the first layer, the original data series is partitioned, and for each subgroup the first L1-Principal Component (L1-PC) is computed. The first rank L1-PCs of all subgroups are then collected to build a new group, which is applied at the input of the second layer. The second layer applies the process of partitioning and analysis to the new group, and produces the input at the next layer. The operation is repeated as many times as necessary to end up, eventually, with only one global principal component. Let us remark that from the second layer on, the data series elements are the unit-norm L1-PCs that

978-9-0827-9703-9/19/$31.00 ©2019 IEEE

we've identified in the previous layer. Therefore, the new input vectors are identified apart from a scalar factor, and differ in nature and range from the original data. This procedure is exemplified in Fig. 1, where the input data series is given by a set of bidimensional sequences.

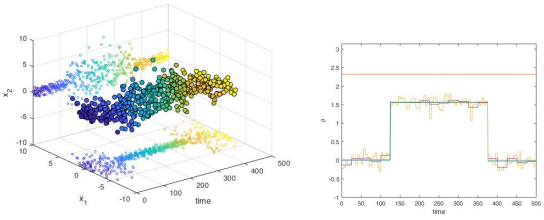

Fig. 2. Time-variant 2D normal data and underlying PDF orientation.

More in general, the input data are a set of N samples of D-dimensional vectors, and are collected in the data matrix $\mathbf{X} = [\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n \ldots, \mathbf{x}_N] \in \mathbb{R}^{D \times N}$. At the k-th layer, the input data is partitioned and N_k size-w_k batches $\mathbf{R}_k(n)$, $n = 1, \ldots, N_k$, are identified. For $k = 1$, $\mathbf{R}_1(n)$ represents the n-th batch of the original data. For $k > 1$, $\mathbf{R}_k(n)$ contains w_k of the N_{k-1} PCs found in layer $k-1$. Specifically, the n-th batch $\mathbf{R}_k(n)$ collects the data with index in the set $S_k(n) = \{(n-1) \cdot w_k + 1, \ldots, n \cdot w_k\}$. It holds that $|S_k(n)| = w_k \ \ \forall n$ and $\sum_{n=1}^{N_k} |S_k(n)| = N_k \cdot w_k = N_{k-1}$, $N_0 = N$ and $w_0 = 1$. Then, the n-th batch, denoted as $\mathbf{R}_k(n)$, is processed and the L1-PC $\mathbf{r}_k(n)$, $n = 1, \ldots, N_k$, is computed, as follows in (1). Thus, a reduction of the input data set occurs at each layer, replacing every n-th w_k-size batch with its own principal component. For N_k the number of batches at layer k, it holds $N_k = N \cdot \prod_{j=1}^{k} \frac{1}{w_j}$.[1]

The L1-PC $\mathrm{r}_k(n)$ is computed resorting to the fast estimator in [2], as

$$\mathbf{r}_k(n) = \mathbf{R}_k(n)\mathbf{b}_k(n) \cdot \Delta_k(n) \tag{1}$$

where the optimal binary vector $\mathbf{b}_k(n)$ is computed as $\underset{b \in \{\pm 1\}^{w_k}}{\operatorname{argmax}} \|\mathbf{R}_k(n)\mathbf{b}\|_2$ and $\Delta_k(n) = \dfrac{1}{\|\mathbf{R}_k(n)\mathbf{b}_k(n)\|_2}$. Thereby $\mathbf{r}_k(n)$ is a normalized version of $\mathbf{R}_k(n)$ $\mathbf{b}_k(n)$. This leads to N_k separate principal components, each one of them being the only first L1-PC obtained from the correspondent batch. This basic operation is repeated several times, until we reach the final layer K.

It is possible to express the Deep L1-PCA at every layer with a compact formula, using the Khatri-Rao product.[2]

To elaborate, we consider the following three matrices:
$\mathbf{B}_k = [\mathbf{b}_{k-1}(1),\ \mathbf{b}_{k-1}(2),\ \ldots,\ \mathbf{b}_{k-1}(n),\ \ldots,\ \mathbf{b}_{k-1}(N_{k-1})] \in \{\pm 1\}^{w_{k-1} \times N_{k-1}}$ grouping the above mentioned optimal $\mathbf{b}_{k-1}(n) \in \mathbb{R}^{w_k - 1}$
$\mathbf{R}_k = [\mathbf{r}_{k-1}(1),\ \mathbf{r}_{k-1}(2),\ \ldots,\ \mathbf{r}_{k-1}(n),\ \ldots,\ \mathbf{r}_{k-1}(N_{k-1})] \in \mathbb{R}^{D \times N_{k-1}}$ collecting all the N_{k-1} PCs found in layer $k-1$ for $k > 1$ and set to $\mathbf{R}_1 = \mathbf{X}$ for $k = 1$
$\mathbf{\Delta}_k = \mathrm{diag}([\Delta_{k-1}(1), \Delta_{k-1}(2), \ldots, \Delta_{k-1}(N_{k-1})])$
Applying the Khatri-Rao product to $\mathbf{B}_k \in \mathbb{R}^{w_{k-1} \times N_{k-1}}$, we find

$$\mathbf{I}_{N_{k-1}} \odot \mathbf{B}_k = \begin{bmatrix} \mathbf{b}_{k-1}(1) & \mathbf{0}_{w_{k-1}} & \ldots & \mathbf{0}_{w_{k-1}} \\ \mathbf{0}_{w_{k-1}} & \mathbf{b}_{k-1}(2) & \ldots & \mathbf{0}_{w_{k-1}} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{0}_{w_{k-1}} & \mathbf{0}_{w_{k-1}} & \ldots & \mathbf{b}_{k-1}(N_{k-1}) \end{bmatrix}$$

It is possible to compactly express the whole set of layer 1 projectors as the results of matrix multiplication between the original data set $\mathbf{X}$ and a "Khatri-Rao size-modulated" maximizer collection $\mathbf{B}_2$:

$$\begin{aligned} \mathbf{R}_2 &= [r_1(1), \ldots, r_1(N_1)] \\ &= [\mathbf{R}_1(1), \ldots, \mathbf{R}_1(N_1)] \cdot (\mathbf{I}_{N_1 \times N_1} \odot \mathbf{B}_2) \cdot \mathbf{\Delta}_2 \\ &= \mathbf{X} \cdot (\mathbf{I}_{N_1 \times N_1} \odot \mathbf{B}_2) \cdot \mathbf{\Delta}_2 \\ \mathbf{R}_2{}_{D \times N_1} &= \mathbf{X}_{D \times N} \Big(\underbrace{\mathbf{I}_{N_1 \times N_1} \odot \mathbf{B}_2{}_{w_1 \times N_1}}_{N_1 \cdot w_1 \times N_1 = N \times N_1} \Big) \cdot \mathbf{\Delta}_2{}_{N_1 \times N_1} \end{aligned} \tag{2}$$

and, for $k = 2, 3, \ldots, K$:

$$\mathbf{R}_k = \mathbf{R}_{k-1} \cdot (\mathbf{I}_{N_{k-1}} \odot \mathbf{B}_k) \cdot \mathbf{\Delta}_k \qquad \in \mathbb{R}^{D \times w_{k-1}} \tag{3}$$

Then, $\mathbf{R}_k$ can be expressed as

$$\begin{aligned} \mathbf{R}_k &= \mathbf{X} \cdot (\mathbf{I}_{N_1 \times N_1} \odot \mathbf{B}_2) \cdot \mathbf{\Delta}_2 \cdot \ldots \cdot (\mathbf{I}_{N_{k-1} \times N_{k-1}} \odot \mathbf{B}_k) \cdot \mathbf{\Delta}_k \\ &= \mathbf{X} \cdot \prod_{j=2}^{k} [(\mathbf{I}_{N_{j-1} \times N_{j-1}} \odot \mathbf{B}_j) \cdot \mathbf{\Delta}_j] \end{aligned} \tag{4}$$

At the last considered layer, the "global" principal component will be extracted from the last subset:

$$\mathbf{r}_K = \mathbf{R}_K \cdot (\mathbf{I}_{N_K \times N_K} \odot \mathbf{B}_{K+1}) \cdot \mathbf{\Delta}_{K+1} \tag{5}$$

where $N_K = 1$, $\mathbf{B}_{K+1} = \mathbf{b}_K(1) \in \{\pm 1\}^{N_{K-1} \times 1}$ and $\mathbf{\Delta}_{K+1} = \Delta_K(1) \in \mathbb{R}^+$. It is clear that the Deep L1-PCA i) differs from the results of overall L1-PCA and ii) conveys a further information, that is the relative distance of L1-PCA solutions found at the intermediate layers.
The Deep L1-PCA is exemplified on the random data set in Fig. 2, which plots 500 samples $\mathbf{X} = [\mathbf{x}_1, \cdots, \mathbf{x}_{500}]$ of a zero-mean bivariate normal distribution with marginal

[1] From now on we assume a simple scenario in which it holds that repeated partitions of the data lead to a single global component, i.e. $N = \prod_{k=1}^{K} w_k$. In some circumstances, on the other hand, the size N of the original dataset could not be equally divided by the chosen w_k and this could lead to up to $w_k - 1$ elements to be dealt with separately in some way.

[2] The Khatri-Rao product of matrices $\mathbf{F} \in \mathbb{R}^{C \times E}$ and $\mathbf{G} \in \mathbb{R}^{D \times E}$ is defines as $\mathbf{F} \odot \mathbf{G} = (F_1 \otimes G_1, \ldots, F_E \otimes G_E)$, with $\otimes$ being the Kronecker product [5], [6].

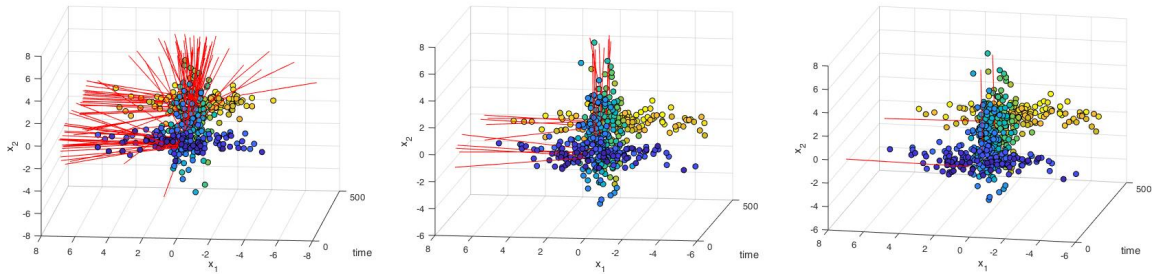

Fig. 3. Time-variant data and first order L1-PCA approximation at different layers.

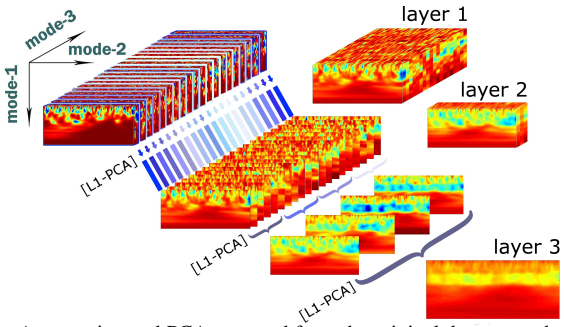

Fig. 4. Aggregation and PCA executed from the original data set to the single common "global" principal component (layer 3).

variances $1/4, 4$, and with principal axis rotation varying according to a pulse $\theta(t)$ represented in Fig. 2, (green line). The original data set $\mathbf{X}$ are parsed to produce the first layer L1-PCA basis vectors (red arrows in Fig. 3 left); the results of the "parallel" L1-PCs are applied at the input of the next layer and processed to produce the second and eventually the third layer (red arrows in Fig. 3 center and right). Finally, the global deep L1-PC $\mathbf{r}^K$, $K = 3$, is obtained. In further results on this dataset, not reported here for compactness' sake, we have compared the first basis vectors obtained at each layer by the deep L1-PCA and by the classical L1-PCA computed on the corresponding original measurements. The unitary energy PCs computed in the two cases differ by a mean square difference in within 10^{-1} - 10^{-4}. Still, let us observe that the mean square difference is expected to depend on the temporal variability of the original data distribution.

III. Application to Brain Connectivity Measurements

Deep L1-PCA can be used to extract robust information across different trials and to identify outliar trials. The signals of interest were provided by the Berlin BCI group and they were collected during brain computer interface (BCI) task. Seven healthy subjects contributed with their motor imagery performances, where visualizing the movements stimulates the brain area responsible for the motor activation and modulates the brain activity, recorded as an EEG signal filtered between 0.05 and 200 Hz and digitized at 1000 Hz with 16 bit (0.1 μV) accuracy. Continuous EEG were recorded during motor tasks imagery session without feedback [4]. Attention was drawn on F4 and CP6 electrodes which are in frontal and central-parietal brain areas which are supposed to be involved in the task. Suppose $\mathbf{y}_{F4}$ and $\mathbf{y}_{CP6}$ are the vectors containing the corresponding EEG signals. Their coherence and phase lag can be analyzed as a function both of time and frequency through by computing their spectral wavelet coherence map:

$$Coh_{\mathbf{y}_{F4},\mathbf{y}_{CP6}} = \frac{|S(C^*_{\mathbf{y}_{F4}}(a,b)\ C_{\mathbf{y}_{CP6}}(a,b))|^2}{S(|C_{\mathbf{y}_{F4}}(a,b)|^2)\ S(|C_{\mathbf{y}_{CP6}}(a,b))|^2)} \quad (6)$$

where $\mathrm{C}_{\mathbf{y}_{F4}}$(a,b) and $\mathrm{C}_{\mathbf{y}_{CP6}}$(a,b) represent the continuous wavelet transforms of time series $\mathbf{y}_{F4}$ and $\mathbf{y}_{CP6}$ at varying scales a and positions b, while S is a smoothing factor in time and scale. Specifically, the coherence between the two electrodes depends on frequencies and time measured with respect to a per-trial stimulus delivery instant [9]. Examples labelled as $1, 2, \ldots, 10$ appear in of Fig. 4.

The measurements in motor imagery conditions have been carried out on a total of 100 repetitions (trial), and are collected in a third-order tensor $\mathcal{A} \in \mathbb{R}^{L \times M \times N}$, comprising modes: frequency span $0 - 32$ [Hz] ($L = 85$), time duration of each trial $M = 400$ [ms], and number of trials considered $N = 100$. Here is shown the application of Deep L1-PCA on the vectorized versions of the coherency maps, which had the goal of extracting a common component and identify the minimum number of trials on which it is consistently estimated, as well as discriminate outliers and changes. The first layer aggregation is shown in Fig.4: the original data tensor $\mathcal{A}$ lays at the starting layer and $N_1 = 20$ batches are obtained at the first layer by slicing it every $w_1 = 5$ wavelet coherence maps, each of which is then represented as a column vector of the input data matrix $\mathbf{X}$.

Every subgroup $\mathbf{X}(n)$ spans the original data set from the "$(n-1) \cdot w_1 + 1$"-th to the "$n \cdot w_1$"-th (1-5, 6-10, ..., 96-100) element, and it undergoes L1-PCA to yield the corresponding first L1-PC. The resulting $N_1 = 20$ L1-PCs form the input at the second layer. The partition and L1-PCA estimation process is carried out until the final layer ($K = 3$).

It is worth noting that for non-negative entries (i.e., $[\mathbf{x}]_i \geq 0$) $\mathbf{1}_M = \underset{\mathbf{b}\in\{\pm 1\}^M}{\operatorname{argmax}} \|\mathbf{Xb}\|_2^2$. This implies that the L1-PC is a scaled version of the mean, in the presented work it can be considered in the perspective of being a layered computation of the mean, for appropriate intermediate scalings of the PCs.

We analyze here the Mean Square value of the Error (MSE) between the L1-PCs with respect to the originating batches at different layers. Firstly, the layer 1 L1-PCs set and the N_0 elements of the input dataset are considered. In Fig. 5 we show the comparison between L1-PCs and the original data elements as well as the MSE scatterplot summarizing the results achieved for different L1-PCA width $w_{k=1} = 5, 10, 20$. Secondly, each element in layer 2 was confronted with its originating batch in layer 1 and the MSE values obtained were plotted into Fig. 6. Similar results, not reported her for compactness' sake, have been obtained at layer 3. In Figs. 5,6, we recognize that the largest mean and spread of the MSE is observed between layer 1 L1-PCs and original data, consistently across all widths w_k, whereas the MSE show a decreasing trend on consecutive layers, since the local L1-PCs approximate the global L1-PC for any considered w.

Next, we plot the MSE between each and every L1-PC in a layer and the corresponding set of wavelet coherence maps from which it derives. In Fig. 7 the plots are organized in different columns corresponding to the L1-PC layers, and rows corresponding to the selected width w. The plots are equipped with a series of vertical lines so as to outline the correspondence between the considered L1-PC and the original data batches; no lines appear when the final single L1-PC is compared with each element of the starting dataset.

From this analysis, we recognize that for any width w, the global principal component obtained at the end of the grouping and PCA process turned out to be substantially the same, i.e. the overall obtained final component is consistent across any initially chosen aggregating interval width (see Fig. 8). Thereby, the estimate achieved at the first layer is a good approximation of the one emerging at the final layer. In application, this can be exploited to select the width (number of trials) of interest, indicating that a stable estimate is obtained considering a reduced number of trials.

Finally, focusing on the 20 L1-PC elements at layer 1, we analyze the Minkowski distance $D_{ij} = (\prod_{p=1}^{P} |r_i(p) - r_j(p)|)^{1/k}, k = 1, P = L \cdot M$ between layer 1 PCs. In Fig. 9 we display the distance matrix as a heatmap; zero values in the principal diagonal are displayed in black, dark red hues represent closer distances whereas largest distances appear in white. The Minkowski distance matrix highlights the L1-PC component which mostly "differs" from the others. The representation allows to identify $w_1 = 5$ coherence maps corresponding to mismatching behaviours (see Fig.10). The same analysis, repeated at higher layers, allows to identify longer and longer groups of trials differing from the common behaviour. Thereby, the Deep L1-PCA computational architecture provides a tool to highlight ouliers of different temporal extent.

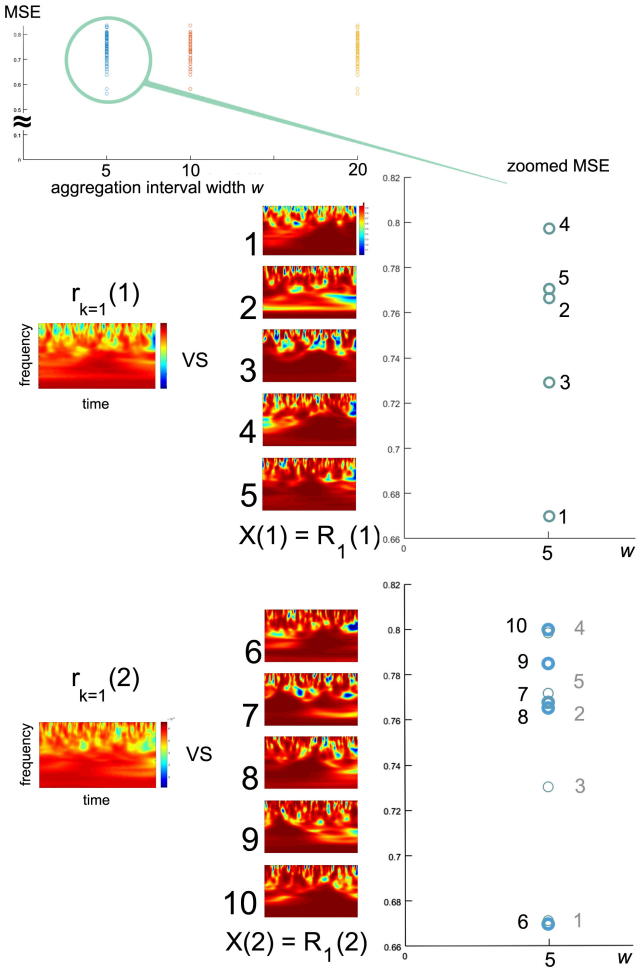

Fig. 5. Top: MSE between layer 1 L1-PCs and corresponding batches from the original data. Botton: zoomed first steps of how MSE (layer k vs. layer $(k-1)$) scatterplot points are generated, for $k = 1$.

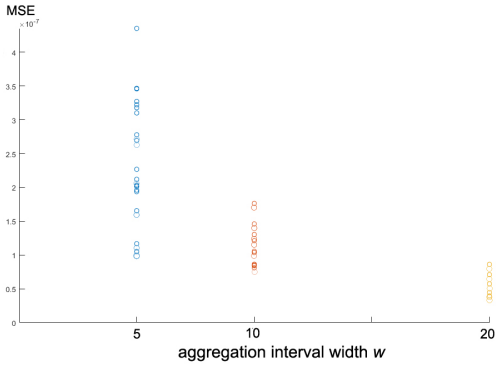

Fig. 6. MSE between every layer 2 L1-PC and its corresponding layer 1 subgroup/batch vs. various aggregation interval widths $w = 5, 10, 20$.

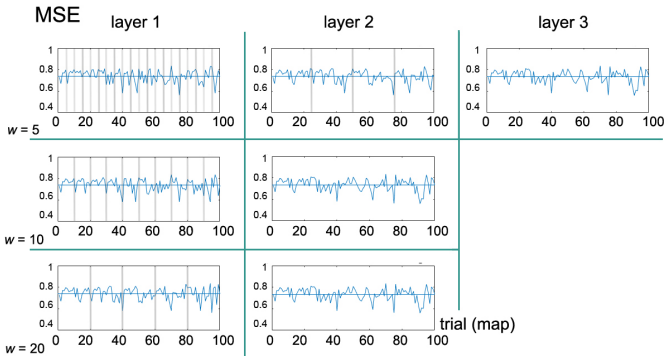

Fig. 7. MSE between L1-PCA at every layer and their corresponding maps in the original dataset.

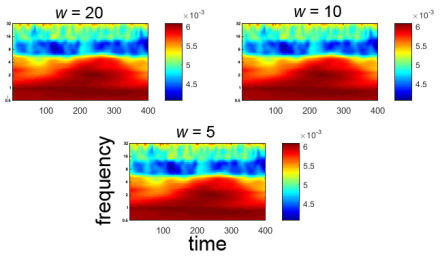

Fig. 8. Global components (layer 3) for each of the grouping parameters w.

To sum up, applying the Deep L1-PCA to the repeated measurements (trials) has allowed us to i) identify the common component(s) at a reduced computational cost, ii) robustly identify outlier trials, and iii) assess the minimum number of trials on which the global component is consistently estimated; besides, this has been accomplished by iv) getting rid of unnecessary normalization tasks, which are usually performed on different acquisitions: the inherently normalization of the L1-PCA basis avoids computation of local data features such as variance or mean, which usually act as nuisance parameters and affect the performances of further processing/detection stages.

IV. Conclusion

In this work we have proposed a layered version of L1-PCA, to which we refer to as Deep L1-PCA. The first layer acts on the original data, which are partitioned in groups; on each

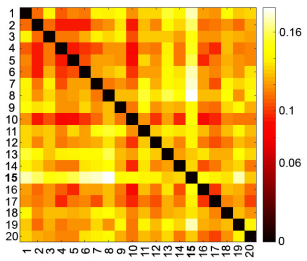

Fig. 9. Layer 1 Minkowski distances matrix heatmap.

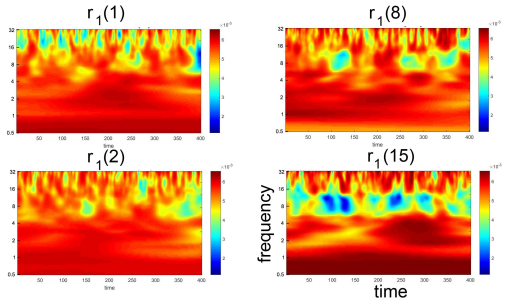

Fig. 10. Layer 1 L1-PCA typical PCs ($\mathbf{r}_{k=1}(n)$) - $n = 1, 2, 8$ vs. the differing element $n = 15$ (bottom right).

group, the L1-PCA basis is estimated and the first component is extracted. The first components are collected and form the basis of the next layer. Then, by recursive application of a more and more compact representation is obtained. At each stage, an inherent normalization is carried out, and Deep L1-PCa is robust to local data features such as variance or mean, which usually act as nuisance parameters and affect the performances of further processing/classification stages.The Deep L1-PCA is applied to repeated brain connectivity measurements and it proves relevant for identifying outliers changes and stable components, as well as to assess the minimum number of measurements on which the global component is consistently estimated. Further work is devoted to analyze the theoretical differences between the global component emerging from Deep L1-PCA and the global component obtained by straightforward L1-PCA.

References

[1] P.P. Markopoulos, D. Mayur, S.Andreas **Adaptive L1-norm principal-component analysis with online outlier rejection**, IEEE Journal of Selected Topics in Signal Processing [Vol. 12, No. 6, 2018]

[2] P.P. Markopoulos, S. Kundu, S. Chamadia and D.A. Pados **Efficient L1-Norm Principal-Component Analysis via Bit Flipping**, IEEE Transactions on Signal Processing [Vol. 65, Iss. 16, August 2017]

[3] V.E. Liong, J. Lu and G. Wang, **Face recognition using Deep PCA**, 2013 9th International Conference on Information, Communications & Signal Processing, Tainan, 2013, pp. 1-5.

[4] B. Blankertz, G. Dornhege, M. Krauledat, K. Müller, and G. Curio, **The non-invasive Berlin Brain-Computer Interface: Fast acquisition of effective performance in untrained subjects**, NeuroImage, 37(2):539-550, 2007

[5] T. Kolda, B. Bader, **Tensor Decompositions and Applications**, SIAM review [Vol. 51, No. 3, 2009]

[6] A. Cichocki, D. Mandic, A-H. Phan, C. Caiafa, G. Zhou, Q. Zhao, and L. De Lathauwer **Tensor Decompositions for Signal Processing Applications From Two-way to Multiway Component Analysis**, IEEE Signal Processing Magazine [Vol. 32, Iss. 2, March 2015]

[7] C. Torrence, G.P. Compo **A Practical Guide to Wavelet Analysis**, Bulletin of the American Meteorological Society, [Vol. 79, No. 1, January 1998]

[8] I. Daubechies, **The Wavelet Transform, Time-Frequency Localization and Signal Analysis**, IEEE Transaction on Information Theory, [Vol. 36, No. 5, September 1990]

[9] A. Klein, T. Sauer, A. Jedynak, W. Skrandies **Conventional and Wavelet Coherence Applied to Sensory-Evoked Electrical Brain Activity**, IEEE Transactions on Biomedical Engineering, [Vol: 53, No. 2, February 2006]

[10] N. Vervliet, O. Debals, L. Sorber, M. Van Barel, L. De Lathauwer **https://www.tensorlab.net**