2018 IEEE INTERNATIONAL WORKSHOP ON MACHINE LEARNING FOR SIGNAL PROCESSING, SEPT. 17–20, 2018, AALBORG, DENMARK

SINGLE-CHANNEL EEG CLASSIFICATION BY MULTI-CHANNEL TENSOR SUBSPACE LEARNING AND REGRESSION

Simon Van Eyndhoven[*†], *Martijn Boussé*[*], *Borbála Hunyadi*[*†],
Lieven De Lathauwer[*‡], *Sabine Van Huffel*[*†]

ABSTRACT

The classification of brain states using neural recordings such as electroencephalography (EEG) finds applications in both medical and non-medical contexts, such as detecting epileptic seizures or discriminating mental states in brain-computer interfaces, respectively. Although this endeavor is well-established, existing solutions are typically restricted to lab or hospital conditions because they operate on recordings from a set of EEG electrodes that covers the whole head. By contrast, a true breakthrough for these applications would be the deployment 'in the real world', by means of wearable devices that encompass just one (or a few) channels. Such a reduction of the available information inevitably makes the classification task more challenging. We tackle this issue by means of a multilinear subspace learning step (using data from *multiple* channels during training) and subsequently solving a regression problem with a low-rank structure to classify new trials (using data from only a *single* channel during testing). We demonstrate the feasibility of this approach on EEG data recorded during a mental arithmetic task.

Index Terms— Brain-computer interface (BCI), multilinear algebra, subspace learning, tensor, tensor regression, wearable electroencephalography (EEG)

1. INTRODUCTION

Non-invasive neuroimaging techniques have received considerable interest in the last couple of decades because they allow to conveniently investigate, classify, or detect a wide range of physiological phenomena taking place in the brain. Among these techniques, electroencephalography (EEG) has become the workhorse in many applications, thanks to its low cost, relatively easy setup of sensors on the scalp, and excellent temporal resolution. Well-known examples of such EEG-based applications are brain computer interfaces (BCI) to restore communication for disabled people, perform clinical research or develop consumer products for e.g. mental state monitoring [1]. Another important application is the automatic detection of seizures in epileptic patients [2].

What unites most of the research trends in these very diverse domains is the focus on lab- or clinical-grade EEG systems, restricting the use in settings outside the lab. For example, EEG measurement setups are often bulky and require substantial manual intervention to use, and more importantly employ a large set of electrodes that covers almost the whole head. Evidently, these systems cannot simply be used outside the lab or clinic. One important hurdle on the way to convenient EEG devices for daily use is a (drastic) reduction of the number of electrodes [3, 4]. This is not only a hardware-related issue: many established signal processing and machine learning approaches for EEG owe their success precisely to the fact that they exploit the spatial diversity that is present in the traditional, multi-channel measurement [5, 6]. Hence, existing processing techniques might have to be reengineered or replaced by new approaches for use in current and future EEG devices that employ only one or a few channels. Several attempts in this direction have already been made: some authors have proposed algorithms that find 'good' subsets of EEG channels to perform certain tasks [7, 8] or derive complex features from the time courses recorded at a few channels, that aim to compensate in some way for the loss of spatial diversity [9].

Here, we take a different approach, and train a model using all available channels but classify new trials using only single-channel data. This follows the realization that the restriction on the number of channels may present itself *only* at the time of deployment of an EEG system. In other words: it can be acceptable in many cases to have a calibration or training session where a larger set of channels is used to prepare or fine-tune an EEG device, after which it is taken into use ('testing phase') with only a low number of channels. In this paper, we present an EEG classification method based on existing work [10] that exploits multi-channel information during a training phase, and can operate on a single-channel EEG signal during testing. Firstly, since there are multiple modes of variation (channels × time points × trials × ...), the data are best represented as a tensor, which is a higher order general-

[*] KU Leuven, Department of Electrical Engineering (ESAT), STADIUS Center for Dynamical Systems, Signal Processing and Data Analytics, Kasteelpark Arenberg 10, 3001 Leuven, Belgium, [†] imec, Leuven, Belgium, [‡] KU Leuven Kulak, Group Science, Engineering and Technology, 8500 Kortrijk, Belgium. The research leading to these results has received funding from the European Research Council under the European Union's Seventh Framework Programme (FP7/2007-2013) / ERC Advanced Grant: BIOTENSORS (n° 339804). This paper reflects only the authors' views and the Union is not liable for any use that may be made of the contained information.

978-1-5386-5477-4/18/$31.00 ©2018 IEEE

ization of vectors and matrices (first and second order tensors, respectively) [11, 12]. Training then consists of a multilinear or tensor subspace learning step, in which appropriate bases are found for all modes of the data. During testing, it is then assumed that unseen trials approximately lie in the same multilinear subspaces as the training data. Hence, to perform classification, a data segment of a test trial is regressed simultaneously onto several subspaces/bases found during training; the resulting coefficients have a low-rank structure (because of the multilinear setting [13]) and inform a decision about class membership. This generic workflow has previously been introduced for face recognition and irregular heartbeat classification [10, 14, 15]. In this paper, we propose several extensions to this framework, to address difficulties that arise in the case of lower signal quality, such as in EEG data. Most importantly, we leverage the class labels of the available training trials to perform the subspace learning step in a supervised fashion, after tensorizing the data using a time delay embedding method. This approach is inspired by the common spectral patterns method presented in [16] and allows to identify bases in which the data differ substantially between classes, in contrast to existing approaches, which use a simple unsupervised subspace learning step. Additionally, we robustify the classification step by non-arbitrarily fixing a sign ambiguity and by employing support vector machines [17] and tree ensemble methods instead of a nearest neighbor approach.

To demonstrate the feasibility of the novel method, we apply it on publicly available EEG data that were recorded from subjects performing a mental arithmetic task [18].

2. CLASSIFYING SINGLE-CHANNEL DATA AFTER MULTI-CHANNEL TRAINING

2.1. Notation and preliminaries

We mostly follow the convention in [11, 12] for the algebraic notation. We denote scalars, vectors, matrices and tensors by lower case (e.g. a), lower case boldface (e.g. $\mathbf{a}$), upper case boldface (e.g. $\mathbf{A}$) and upper case calligraphic letters (e.g. $\mathcal{A}$), respectively. An Nth order tensor $\mathcal{A} \in \mathbb{R}^{I_1 \times I_2 \times \cdots \times I_N}$ is a multiway array that varies over N modes (e.g. sensors, time points, trials, ...) with dimensions I_1, I_2, ... , I_N, respectively. Slices of a tensor are obtained by fixing the index for one or several modes. When indices of all but one mode are fixed, and hence only variation over one mode is considered, we use the term *fiber*, which is a generalization for matrix rows and columns. A tensor $\mathcal{A}$ can be matricized by stacking all fibers of a certain mode-n in a matrix. This is also known as the mode-n unfolding of $\mathcal{A}$ and is written as $\mathbf{A}_{(n)}$. Equivalently, the vectorization of $\mathcal{A}$ is obtained by stacking all mode-1 fibers in a long vector and is written as vec($\mathcal{A}$). The mode-n multiplication of a tensor $\mathcal{A}$ and a matrix $\mathbf{V}$ is denoted by $\mathcal{A} \times_n \mathbf{V}$ and consists of the left-multiplication of all mode-n fibers of $\mathcal{A}$ with $\mathbf{V}$. The mode-n rank of a tensor is the rank of the mode-n unfolding of the tensor. The set of mode-n ranks (jointly called *multilinear rank*) is in general different from the tensor rank, which is the minimal number of rank-1 tensors to fully describe the tensor. Such a collection of rank-1 terms is then called a canonical polyadic decomposition. Every rank-1 tensor in its turn is the outer product of N vectors, denoted by $\mathbf{a}^{(1)} {\scriptstyle\otimes} \mathbf{a}^{(2)} {\scriptstyle\otimes} \ldots {\scriptstyle\otimes} \mathbf{a}^{(N)}$. The Kronecker product of two matrices $\mathbf{A} \in \mathbb{R}^{I_1 \times I_2}$ and $\mathbf{B} \in \mathbb{R}^{J_1 \times J_2}$ is indicated by $\otimes$ and defined as the $I_1 J_1 \times I_2 J_2$ matrix

$$\mathbf{A} \otimes \mathbf{B} = \begin{bmatrix} \mathbf{A}(1,1)\mathbf{B} & \cdots & \mathbf{A}(1,I_2)\mathbf{B} \\ \vdots & \ddots & \vdots \\ \mathbf{A}(I_1,1)\mathbf{B} & \cdots & \mathbf{A}(I_1,I_2)\mathbf{B} \end{bmatrix} \tag{1}$$

A useful relationship between the the outer and Kronecker product is $\text{vec}(\mathbf{a} {\scriptstyle\otimes} \mathbf{b}) = \mathbf{b} \otimes \mathbf{a}$.

2.2. Training phase: subspace learning

Training of the classifier consists of finding appropriate subspaces or bases for all modes, in which the training trials' data can be represented. Afterwards, data from a new trial can be regressed or projected onto these subspaces, and the resulting coefficients can be fed into a classifier that estimates the class membership. Hence, ideally, subspaces should be sought in which the coefficients of trials from the same class lie close to each other, but far from coefficients of other trials.

2.2.1. Tensorization of the data

While a subject participates in several trials of a BCI paradigm, EEG data are being collected at D electrodes or channels, with T time points for each of P trials. Hence, the data are naturally represented as a third order tensor $\mathcal{X} \in \mathbb{R}^{D \times T \times P}$, with modes channels × time points × trials, indexed by d, t and p, respectively. We may expand the data by stacking $L - 1$ delayed versions of the signal at every channel along a fourth mode, increasing the order of the tensor by one. This method is known as time delay embedding and was proposed for multi-channel EEG classification in [16]. Mathematically, the new tensor $\mathcal{X}^e \in \mathbb{R}^{D \times T \times L \times P}$ is obtained as $\mathcal{X}^e(d,t,l,p) = \mathcal{X}(d, t-(l-1)\tau, p),\ \forall\, d,t,l,p$, where L is known as the embedding dimension and τ is the introduced delay, and has modes channels × time points × lags × trials. Every mode has an associated subspace or base in which the mode-n fibers can be represented.

2.2.2. Unsupervised subspace learning

A popular way to find the multilinear subspaces of a higher order dataset $\mathcal{A} \in \mathbb{R}^{I_1 \times I_2 \times \cdots \times I_N}$ is the multilinear singular value decomposition (MLSVD), defined as

$$\mathcal{A} = \mathcal{S} \times_1 \mathbf{U}^{(1)} \times_2 \mathbf{U}^{(2)} \cdots \times_N \mathbf{U}^{(N)} \tag{2}$$

where every $\mathbf{U}^{(n)} \in \mathbb{R}^{I_n \times I_n}$ is a unitary matrix, and the core tensor $\mathcal{S} \in \mathbb{R}^{I_1 \times I_2 \times \cdots \times I_N}$ is ordered and all-orthogonal [11, 12]. It is a higher order generalization of the singular value decomposition (SVD), and as such the (truncated) matrices $\mathbf{U}^{(n)}$ contain bases that capture as much variance of $\mathcal{A}$ along mode n as possible. They can be found by applying an SVD on unfolded versions of the tensor in every mode. As this approach disregards class label info, there is no guarantee whatsoever that the data of trials from different classes are represented (sufficiently) differently in these bases.

2.2.3. Supervised subspace learning

Discriminative subspaces might be found by taking into account the label information from the training trials. This is the philosophy of the popular common spatial patterns (CSP) algorithm [5], in which a set of spatial filters $\mathbf{w}_i$ is applied to a multi-channel EEG dataset $\mathbf{X}$ such that the filtered output signals $\mathbf{y}_i = \mathbf{w}_i^T \mathbf{X}$ have a high power during trials of one condition, but a low power for trials of other conditions. The output power of several CSP filters can then be used as features for classification, assuming that changes in (frequency-specific) power in the EEG are related to the 'brain states' of interest. This approach was extended to the common spatio-spectral patterns (CSSP) algorithm, whereby time delay embedding is used to extend the set of channels, as described earlier, leading to filters that aggregate signals over multiple channels *and* lags [16]. The matrix of CSSP filters $\mathbf{W}$ can be found as the solution to the generalized eigenvalue problem

$$\mathbf{\Sigma}_2\, \mathbf{W} = (\mathbf{\Sigma}_1 + \mathbf{\Sigma}_2)\, \mathbf{W}\mathbf{\Lambda}\,, \tag{3}$$

in which $\mathbf{\Sigma}_1$ and $\mathbf{\Sigma}_2$ are the covariance matrices of the EEG data belonging to class 1 and 2, respectively, and $\mathbf{\Lambda}$ is a diagonal matrix containing the generalized eigenvalues λ_i, which lie between zero and one. Filters $\mathbf{w}_i$ associated with a λ_i that is close to one have a higher output power during trials of class 2 compared to trials of class 1, which are amplified by filters with a λ_i that is closer to zero.

As we aim to classify new trials based on single-channel EEG, we cannot aggregate multiple signals spatially, but only temporally, i.e. by taking a weighted sum of the single-channel EEG over multiple lags (after time delay embedding). Hence, the covariances in (3) are computed on the unfolded data $\mathbf{X}^e_{(3)} \in \mathbb{R}^{L\times DTP}$, leading to finite impulse response (FIR) filters that can be applied over the third mode (lags):

$$\mathcal{Y} = \mathcal{X}^e \times_3 \mathbf{W}^T \tag{4}$$

The multilinear bases of the filtered data $\mathcal{Y}$ are thus more discriminative than those of the original data $\mathcal{X}^e$, as the power of time series in $\mathcal{Y}$ correlates better to the class labels.

2.2.4. Feature extraction

The time-varying power of an EEG signal is computed by taking the sum of squares of every time course in $\mathcal{Y}$ in sliding, non-overlapping windows of length m. The logarithm of the power values is used as a feature for classification and stored in a feature tensor $\tilde{\mathcal{Y}}$. Ultimately, the MLSVD of this tensor is computed to obtain the training bases $\mathbf{U}_{\text{chan}} \in \mathbb{R}^{D\times D}$, $\mathbf{U}_{\text{time}} \in \mathbb{R}^{\frac{T}{m}\times\frac{T}{m}}$, $\mathbf{U}_{\text{filt}} \in \mathbb{R}^{L\times L}$ and $\mathbf{U}_{\text{trial}} \in \mathbb{R}^{P\times P}$ and a core tensor $\tilde{\mathcal{S}}_y \in \mathbb{R}^{D\times\frac{T}{m}\times L\times P}$ that explains the interaction between basis vectors of the different modes.

2.3. Testing phase: tensor regression and classification

2.3.1. Regression with a low-rank solution

Note that, before truncation of the bases, the time varying power over the spectral filter outputs $\tilde{\mathcal{Y}}^{c,p} \in \mathbb{R}^{1\times T\times L\times 1}$ at every channel d, in every trial p, admits a representation as

$$\tilde{\mathcal{Y}}^{d,p} = \tilde{\mathcal{S}}_y \times_1 \mathbf{u}^T_{\text{chan},d} \times_2 \mathbf{U}_{\text{time}} \times_3 \mathbf{U}_{\text{filt}} \times_4 \mathbf{u}^T_{\text{trial},p} \tag{5}$$

in which $\mathbf{u}^T_{\text{chan},d}$ and $\mathbf{u}^T_{\text{trial},p}$ represent the dth and pth row of $\mathbf{U}_{\text{chan}}$ and $\mathbf{U}_{\text{trial}}$, respectively. Unfolding over the second and third mode yields

$$\begin{aligned} \tilde{\mathbf{Y}}^{d,p}_{(2,3)} &= (\mathbf{U}_{\text{filt}} \otimes \mathbf{U}_{\text{time}})\, \tilde{\mathbf{S}}_{y(2,3)}\, (\mathbf{u}_{\text{trial},p} \otimes \mathbf{u}_{\text{chan},d}) \\ \tilde{\mathbf{Y}}^{d,p}_{(2,3)} &= \mathbf{B}\, (\mathbf{u}_{\text{trial},p} \otimes \mathbf{u}_{\text{chan},d}) \end{aligned} \tag{6}$$

This equation describes the data $\tilde{\mathbf{Y}}^{d,p}_{(2,3)}$ in the unfolded multilinear subspace $\mathbf{B}$ using coefficients $\mathbf{u}_{\text{chan},d}$ and $\mathbf{u}_{\text{trial},p}$. As was noted in section 2.1, the Kronecker product $\mathbf{u}_{\text{trial},p} \otimes \mathbf{u}_{\text{chan},d}$ can be seen as the vectorization of a rank-1 matrix $\mathbf{u}_{\text{chan},d} \otimes \mathbf{u}_{\text{trial},p}$. For this reason, expressions such as (6) have been referred to as Kronecker Product Equations (KPE) [14] or Linear Systems with Canonical Polyadic Decomposition-constrained solution [10]. To classify a new trial q using data from EEG channel f, we first construct the matrix $\tilde{\mathbf{Z}}^{f,q}_{(2,3)}$ by computing the features as explained in previous section (analogously to the training data). Subsequently, we solve the following set of equations for $\hat{\mathbf{u}}_{\text{chan}}$ and $\hat{\mathbf{u}}_{\text{trial}}$:

$$\tilde{\mathbf{Z}}^{f,q}_{(2,3)} = \mathbf{B}\, (\hat{\mathbf{u}}_{\text{trial}} \otimes \hat{\mathbf{u}}_{\text{chan}}) \tag{7}$$

This corresponds to performing a regression of the new data $\tilde{\mathbf{Z}}^{f,q}_{(2,3)}$ onto the multilinear basis $\mathbf{B}$, in which the solution has a low-rank structure by virtue of the Kronecker product between the coefficients [13]. The vector $\hat{\mathbf{u}}_{\text{trial}}$ then holds coefficients that express the new data segment in the subspace of the trial mode, and can be used for classification.

2.3.2. Classification

In [10, 14], new data instances are classified by comparing the estimated coefficients $\hat{\mathbf{u}}_{\text{trial}}$ with rows of the trained matrix $\mathbf{U}_{\text{trial}}$ and assigning the label of the closest matching training data instance. This nearest neighbor (NN) approach may be adjusted or replaced by more robust classification methods that take the coefficients $\hat{\mathbf{u}}_{\text{trial}}$ as input features. In this paper, we compare the performance of seven different methods: K-nearest neighbors (`knn`) with K equal to 1, 3, or 5, least-squares support vector machines with a linear kernel (`svm-lin`) or radial basis function kernel (`svm-rbf`), random forests with 100 trees (`randfor`) and gradient boosting using 10 shallow trees that are trained with AdaBoost (`treeboost`). The classifiers are trained using the labeled rows of $\mathbf{U}_{\text{trial}}$ as features.

2.3.3. Determining the sign of the coefficients

When solving the tensor regression problem, a sign ambiguity and scaling ambiguity remains, i.e., the values in $\hat{\mathbf{u}}_{\text{trial}}$ can be scaled by an arbitrary factor (that can be negative), if the values in $\hat{\mathbf{u}}_{\text{chan}}$ are counterscaled by the same factor. Since the scale and sign can have a large impact on the classification, it is crucial to fix them. Although algebraically the sign cannot be retrieved when solving the regression problem, it may be inferred from the training data set. Namely, we may look at the matrices $\mathbf{U}_{\text{trial}}$ and $\mathbf{U}_{\text{chan}}$ and try to find 'anchoring variables', i.e. columns of those matrices that maximize

$$\text{score}_{i,\text{trial}} = \left| \sum_{c=1}^{C} \frac{\mu_{c,i}}{\sigma_{c,i}} \right| \quad , \quad \text{score}_{j,\text{chan}} = \left| \frac{\mu_j}{\sigma_j} \right| \tag{8}$$

Here, $\mu_{c,i}$ and $\sigma_{c,i}$ are the mean and standard deviation over values from the ith column of $\mathbf{U}_{\text{trial}}$ that belong to class c. Analogously, μ_j and σ_j are the mean and standard deviation over values from the jth column of $\mathbf{U}_{\text{chan}}$. A large score for a latent variable i indicates that this variable has a consistent sign, i.e. that its values lie sufficiently far from zero and are either positive and negative, and are not informative for the class label. Hence, the latent variable that maximizes $\text{score}_{i,\text{trial}}$ provides an 'anchor' whose sign dictates the sign of the corresponding entry of $\hat{\mathbf{u}}_{\text{trial}}$, and thus of the whole vector. Analogously, the latent variable j with the largest score $\text{score}_{j,\text{chan}}$ can be used to correct the sign of the vector $\hat{\mathbf{u}}_{\text{chan}}$. As a third criterion, we may impose that $\hat{\mathbf{u}}_{\text{chan}}$, that was estimated for a certain channel f, should have a positive cosine similarity with the corresponding row of $\mathbf{U}_{\text{chan}}$: negative values of $\mathbf{U}_{\text{chan}}(f,:)\hat{\mathbf{u}}_{\text{chan}}$ suggest that the sign of $\hat{\mathbf{u}}_{\text{chan}}$ should be flipped. We use the majority vote of these three criteria to determine the sign of the coefficient pair. The scaling ambiguity is resolved by scaling $\hat{\mathbf{u}}_{\text{trial}}$ and all rows of $\mathbf{U}_{\text{trial}}$ to unit norm.

3. EXPERIMENTAL RESULTS ON A MENTAL TASK

3.1. Public EEG dataset

We apply the method on a recently published dataset with EEG recordings of subjects that participate in a mental arithmetic task (details can be found in [18]). During the experiment, subjects complete sixty trials of ten seconds each, in half of which they are instructed to repeatedly subtract two numbers, and in half of which they rest. In this paper, we used the 30-channel EEG recordings of 14 subjects from this dataset, which are sampled at 200 Hz.

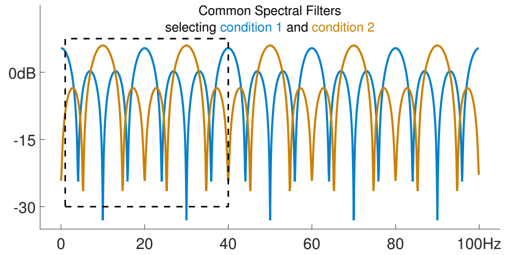

Fig. 1: During training, common spectral pattern (CSP) filters are tuned to frequency bands in which the spectral power is discriminative for the class. For a particular subject, the filter-which maximizes variance for condition 1 (the task) selects a band around 20 Hz. On the other hand, the filter that focuses on the other class of trials selects complementary bands.

3.2. Preprocessing and analysis

The EEG data of all subjects were band-pass filtered between 0.5-40 Hz, and eyeblink artifacts were estimated and removed using independent component analysis (ICA), as implemented in the SOBI algorithm from EEGLAB [19]. The traditional, full-channel common spatial patterns analysis was first ran on every subject's recording, to identify the channels in which the mental task manifested itself the strongest. This information is conveyed by the rows of the matrix $\mathbf{W}^{-1}$ [5]. We selected $C = 10$ channels per subject with the highest root-mean-square value (averaged over all rows of $\mathbf{W}^{-1}$). The parameters for the subsequent time delay embedding step were chosen as $\tau = 10$ samples and $L = 6$ after tuning on data from a subject whose results are left out of the current analysis.

We estimated L common spectral pattern filters by means of (3) and applied them to the data as in (4). Finally, the logarithm of the signal power was computed in windows of 1 second, leading to a feature tensor $\tilde{\mathcal{Y}} \in \mathbb{R}^{10\times10\times6\times60}$. The set of equations in (7) were solved using a state-of-the-art Gauss-Newton algorithm in Tensorlab [20]. In [10], it was derived that under mild conditions, the number of equations should be as least as high as the number of unknown coefficients to ensure that a unique solution exists, i.e. $TL \geqslant \dim(\mathbf{U}_{\text{trial}}) + \dim(\mathbf{U}_{\text{chan}})$. Hence, the basis matrices $\mathbf{U}_{\text{chan}}$ and $\mathbf{U}_{\text{trial}}$ were truncated to dimensions 3 and 5, respectively, upon inspection of the multilinear singular values of of $\tilde{\mathcal{Y}}$. Least squares SVMs with linear or RBF kernel were trained on the rows of the truncated matrix $\mathbf{U}_{\text{trial}}$ using LS-SVMlab [17].

3.3. Results and discussion

Trials were classified using the seven classifiers described in section 2.3 in a 15-fold cross-validation setting that was repeated ten times per subject. Note that the CSP filters use label info and are trained as part of the cross-validation procedure as well [5]. In fig. 1 we inspect (for one subject) the transfer functions of those filters belonging to the most extreme generalized eigenvalues. Because of the time delay embedding step, the columns of $\mathbf{W}$ each have L coefficients that

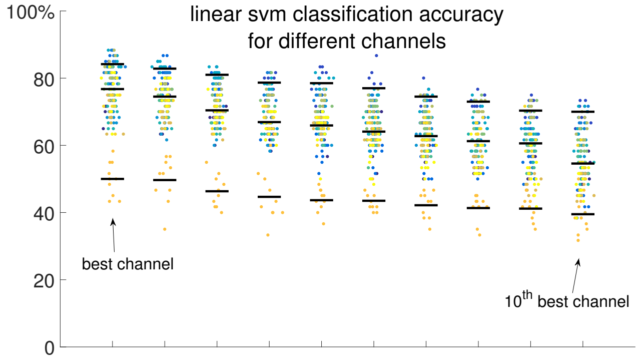

Fig. 2: Good single-channel classification accuracy can be attained by the tensor subspace learning and regression method. The performance depends on the selected channel, and differs between subjects (coded by color). This is indicated by black lines at the mean accuracy (over all repetitions) of subjects with the highest, median, and lowest performance.

map one-to-one to FIR filter coefficients that are spaced by τ samples. E.g. a column $\mathbf{w}_i$ gives rise to an impulse response $\mathbf{h}_i = \mathbf{w}_i \otimes [\,1\;0\;\ldots\;0\,]^T$ with L degrees of freedom, where the second vector contains $\tau - 1$ trailing zeros. From the blue trace, we deduce that the mental arithmetic task provokes neural activity which differs most saliently from the 'rest' condition in a band between approximately 17-23 Hz. On the other hand, the orange filter is tuned towards the 'rest' trials and selects complementary frequency bands. The classification accuracy varies between subjects and channels used during testing. This is shown in fig. 2 for the linear kernel SVM, in which the accuracies of different repetitions are shown, in distinct colors per subject, over channels. Note that the 10 selected channels can be different for every subject, and are grouped based on their ranking, although in our analysis there were some electrodes in common for the majority of subjects, located over the left temporal lobe, and the frontal lobe near the centerline. The mean accuracy over repetitions of the best-performing subject was 84% and 70% at the best and worst channel of the set, respectively. For the worst-performing subject, these metrics were 50% and 40%. Across all subjects, the single-channel classification method presented here had a performance which was only 6.6 ± 17% lower than the performance of the common spatial pattern classification using all thirty channels, conducted as in [18]; for some subjects the single-channel method performed even remarkably better. We conclude that for well-chosen subjects, our pipeline is robust and may be used for single-channel classification, with an acceptable loss in performance over a full-channel BCI.

A comparison between the seven evaluated classifiers is presented in fig. 3 for the best channel of every subject. We managed to boost the classification accuracy by a six percent, through the use of linear kernel SVMs instead of the simple nearest neighbor approach, which was used in [10, 14]. Although KNN classifiers, RBF-SVMs are able to model nonlinear decision boundaries, they are more sensitive to outliers – which are almost invariably present in these EEG datasets – than linear classifiers. The random forest yielded intermediate performance, whereas the tree boosting method was unreliable, potentially also due to outlier sensitivity of AdaBoost. Since the best sensitivity and specificity are attained at different channels, improved classification is possible by relying on two or more channels, as in [14].

We observed that testing accuracy was in general several percent lower than validation accuracy (not shown). This drop is due to the feature computation process: for the training (and validation) trials, features can be readily extracted from the rows of $\mathbf{U}_{\text{trial}}$, whereas for the test trials, an intermediate step (tensor regression to obtain the coefficients of the test trials in the precomputed subspace) is needed. Due to this extra step, the features of test trials are prone to estimation error, which explains a lower performance.

Here, we used the same L for all subjects, although this parameter could be optimized individually. We found that the performance is relatively stable when choosing closeby values of L, e.g. 5 or 7, but that the variability of the performance increases for higher L. This effect is due to two factors: for higher L, the regression problem in (7) has more observations and is hence 'more overdetermined', leading to a more robust estimation. However, the accompanying risk is that the spectral filters $\mathbf{W}$ start to overfit the training data, which undermines the performance for some subjects' data.

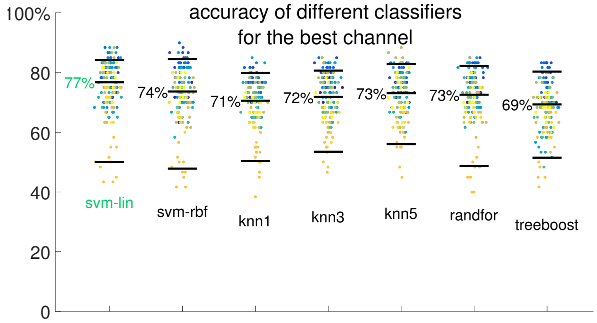

Fig. 3: Support vector machine (SVM) classifiers, especially the linear kernel SVM, are more robust than the K-nearest neighbor (KNN) classifiers. Tree boosting yielded the least reliable classification in the current analysis, followed by the simple `knn1` classifier, used in [10, 14]. Colors and black levels have the same interpretation as in fig. 2.

4. CONCLUSION

We developed a pipeline for classifying brain states based on data from a single EEG channel, after a calibration phase in which information of multiple channels is exploited. This key aspect renders it suitable for use in wearable devices, where the number of electrodes is limited. The pipeline relies on 1) a supervised subspace identification step, in which spectral filters are found that extract frequency bands that are discriminative for the classification task, 2) solving a tensor regression problem with a low-rank structure and 3) the application of a classifier on the estimated regression coefficients. Although we demonstrated its success for the classification of trials of a mental task in a BCI context, the approach is generic; it may be applicable for the detection of e.g. epileptic seizures as well, as these also induce spectral changes, or also for non-EEG data. The proposed framework can still be improved, by using a few channels instead of only one [14], and by rigorously tuning the parameters in the pipeline (e.g. L, τ, C, $\dim(\mathbf{U}_{\text{chan}})$, $\dim(\mathbf{U}_{\text{trial}})$) with cross-validation. Based on the 'equations versus unknowns' trade-off, a high L and T are warranted, although this incurs the risk of overfitting (as discussed in previous section), or increases the latency of the pipeline, respectively. Alternatively, $\dim(\mathbf{U}_{\text{chan}})$ and $\dim(\mathbf{U}_{\text{trial}})$ should be low, though this might compromise the regression's fit, and as $\dim(\mathbf{U}_{\text{trial}})$ equals the number of features, it limits the classifiers' learning capabilities. During training, dictionary learning methods could further aid to find suitable subspaces that are robust to outliers [21].

5. REFERENCES

[1] P. Brunner, L. Bianchi, et al., "Current trends in hardware and software for brain–computer interfaces (BCIs)," *Journal of neural engineering*, vol. 8, no. 2, pp. 025001, 2011.

[2] U. R. Acharya, S. V. Sree, et al., "Automated EEG analysis of epilepsy: a review," *Knowledge-Based Systems*, vol. 45, pp. 147–165, 2013.

[3] L. Liao, C. Lin, et al., "Biosensor technologies for augmented brain–computer interfaces in the next decades," *Proceedings of the IEEE*, vol. 100, no. Special Centennial Issue, pp. 1553–1566, 2012.

[4] D. Looney, P. Kidmose, et al., "The in-the-ear recording concept: User-centered and wearable brain monitoring," *IEEE pulse*, vol. 3, no. 6, pp. 32–42, 2012.

[5] B. Blankertz, R. Tomioka, et al., "Optimizing spatial filters for robust EEG single-trial analysis," *IEEE Signal processing magazine*, vol. 25, no. 1, pp. 41–56, 2008.

[6] F. Lotte, M. Congedo, et al., "A review of classification algorithms for EEG-based brain–computer interfaces," *Journal of neural engineering*, vol. 4, no. 2, pp. R1, 2007.

[7] T. N. Lal, M. Schroder, et al., "Support vector channel selection in BCI," *IEEE transactions on biomedical engineering*, vol. 51, no. 6, pp. 1003–1010, 2004.

[8] B. Lou, B. Hong, et al., "Bipolar electrode selection for a motor imagery based brain–computer interface," *Journal of Neural Engineering*, vol. 5, no. 3, pp. 342, 2008.

[9] S. Liang, C. Kuo, et al., "Automatic stage scoring of single-channel sleep EEG by using multiscale entropy and autoregressive models," *IEEE Transactions on Instrumentation and Measurement*, vol. 61, no. 6, pp. 1649–1657, 2012.

[10] M. Boussé, N. Vervliet, et al., "Linear systems with a canonical polyadic decomposition constrained solution: Algorithms and applications," Tech. Rep., 17-01, ESAT-STADIUS, KU Leuven, Leuven, Belgium, 2017.

[11] N. D. Sidiropoulos, L. De Lathauwer, et al., "Tensor decomposition for signal processing and machine learning," *IEEE Transactions on Signal Processing*, vol. 65, no. 13, pp. 3551–3582, 2017.

[12] A. Cichocki, D. Mandic, et al., "Tensor decompositions for signal processing applications: From two-way to multiway component analysis," *IEEE Signal Processing Magazine*, vol. 32, no. 2, pp. 145–163, 2015.

[13] H. Zhou, L. Li, and H. Zhu, "Tensor regression with applications in neuroimaging data analysis," *Journal of the American Statistical Association*, vol. 108, no. 502, pp. 540–552, 2013.

[14] M. Boussé, G. Goovaerts, et al., "Irregular heartbeat classification using Kronecker Product Equations," in *Proc. of the 39th Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC 2017, Jeju Island, South Korea)*, 2017, pp. 438–441.

[15] M. A. O. Vasilescu and D. Terzopoulos, "Multilinear image analysis for facial recognition," in *Object recognition supported by user interaction for service robots*, 2002, vol. 2, pp. 511–514.

[16] S. Lemm, B. Blankertz, et al., "Spatio-spectral filters for improving the classification of single trial EEG," *IEEE transactions on biomedical engineering*, vol. 52, no. 9, pp. 1541–1548, 2005.

[17] J. AK Suykens, T. Van Gestel, and J. De Brabanter, *Least squares support vector machines*, World Scientific, 2002.

[18] J. Shin, A. von Lühmann, et al., "Open access dataset for EEG + NIRS single-trial classification," *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, vol. 25, no. 10, pp. 1735–1745, 2017.

[19] A. Delorme and S. Makeig, "EEGLAB: an open source toolbox for analysis of single-trial EEG dynamics including independent component analysis," *Journal of neuroscience methods*, vol. 134, no. 1, pp. 9–21, 2004.

[20] N. Vervliet, O. Debals, et al., "Tensorlab 3.0," Available online, Mar. 2016. URL: www.tensorlab.net.

[21] Carlos A Loza and Jose C Principe, "A robust maximum correntropy criterion for dictionary learning," in *Proc. of the 26th IEEE International Workshop on Machine Learning for Signal Processing (MLSP 2016, Vietri sul Mare, Salerno, Italy)*, 2016, pp. 1–6.