Classification of structured EEG tensors using nuclear norm regularization: improving P300 classification

Borbála Hunyadi*†, Marco Signoretto*, Stefan Debener‡, Sabine Van Huffel*† and Maarten De Vos‡
*Department of Electrical Engineering (ESAT-SISTA), KU Leuven, Leuven, Belgium
†iMinds Future Health Department, Leuven, Belgium
‡Neuropsychology Lab, Department of Psychology, University of Oldenburg, Oldenburg, Germany

***Abstract*—Choosing an appropriate approach for single-trial EEG classification is a key factor in brain computer interfaces (BCIs). Here we consider an auditory oddball paradigm, recorded in normal indoor and walking outdoor conditions. The signal of interest, namely the P300 component of the event related potential (ERP), unlike noise, is a structured signal in the multidimensional space spanned by channels, time and frequency or possibly other types of features. Therefore, we apply spectral regularization using nuclear norm on a tensorial representation of the EEG data. Due to the a-priori structural information conveyed by the nuclear norm penalty, we expect an improved performance compared to traditional approaches, especially under noisy conditions and in case of small sample sizes.**

***Keywords*-spectral regularization; tensorial representation; nuclear norm; mobile BCI**

I. Introduction

Successful single-trial classification of EEG data lies within discriminative feature extraction combined with a suitable classifier. Both aspects have been tackled extensively in the literature.

Regarding classification approaches, linear discriminant analysis (LDA) is probably the most common approach in BCI research, and was shown to be optimal if proper preprocessing is applied. To limit the influence of outliers and account for the estimation error of the covariance matrix, regularized or shrinkage LDA can be used. Various different algorithms were successfully applied as well, such as Bayesian techniques, support vector machines, neural networks, etc. For an overview, see [1], [2]. Nevertheless, the features used as input to any classifier need to be carefully chosen to obtain real discriminative power.

A common approach for feature extraction in ERP-based BCI is the decimation and concatenation of the signals from each channel to a long feature vector. However, this leads to high dimensional input data, moreover, it ignores the spatial properties of multichannel data. Feature extraction approaches exploiting spatial information from the multichannel EEG include common spatial patterns (CSP) [3] and beamforming techniques [4]. Recently, higher order discriminant analysis was proposed to extract features capturing the multilinear structure in tensor expansion of EEG data [5].

Our classification approach finds an optimal discriminative subspace exploiting the inherent multidimensional structure of the EEG data through spectral regularization with nuclear norm without explicit feature extraction. The signal of interest, in our case the P300 ERP component in an auditory paradigm, unlike noise, is a structured signal in the multidimensional space spanned by channels, time, frequency or possibly other types of features. Therefore, this algorithm is expected to be robust under noisy conditions. Moreover, it was shown that exploiting tensor structure in the learning algorithm is especially useful for small samples sizes [6], which is a typical challenge in BCI reseach. Regularization with nuclear norm for matrix classification was proposed in [7], [8] for BCI data, and was successfully applied for epileptic seizure detection [9]. Nuclear norms were extended to the case of higher order arrays as a heuristic for multilinear rank constrained problems [10].

Here we consider an auditory oddball paradigm, recorded in normal indoor and walking outdoor conditions. Single trials are represented in a tensor, where Hankel matrices constructed from the decimated signals of each channel are stacked along the third mode. The performance of the proposed nuclear norm regularization technique applied on tensorial data is compared with two alternative approaches: (1) classification using nuclear norm regularization on matrices, where decimated signals from each channel constitute each row of a matrix, and (2) shrinkage LDA applied on concatenating the decimated signals from each channel.

II. Methods

A. Data acquisition

Sixteen healthy volunteers free of past or present neurological or psychiatric conditions participated. 616 standard and 84 deviant tones (600 and 1200 Hz, respectively) were presented in randomized order at a fixed interstimulus interval (1,000 ms) in two experimental conditions. Participants were asked to silently count the rare tones. In the indoor recording condition, they sat in a quiet office room; in the outdoor recording condition, they walked slowly on campus. No instructions were given for head and eye movements. The route consisted of different surfaces and the environment included substantial ambient noise.

The order of indoor and outdoor conditions was balanced across subjects. Stimulus presentation was controlled with OpenViBE. EEG was recorded with a 14-channel wireless EEG (128 Hz sampling rate; 0.1645 Hz band-pass; Emotiv: www.emotiv.com) connected to a state-of-the-art infracerebral electrode cap (www.easycap.de), with electrodes positioned at sites according to the standard 10-20 system. The resulting system is of small size and weight, could be tightly attached to the cap. The signal quality was improved due to the sintered Ag/AgCl electrodes. Data were analyzed offline using EEGLAB and MATLAB. Extended infomax independent component analysis (ICA) was used to semiautomatically attenuate eye blinks [11]. EEG data were 20 Hz low-pass filtered, and trial epochs were extracted (-200 to 700 ms) and baseline corrected (-200 to 0 ms) and re-referenced to the average of Tp9 and TP10. Trials with large or non-stereotypical artifacts were rejected. Standard trials preceding deviants were selected to equalize the number of standard and deviant trials, resulting in a minimum of 70 standard and deviant tones per subject.

B. Feature extraction

EEG data from all 12 electrodes was subdivided in 30 non-overlapping time windows comprising 0 to 700 ms. The mean signal amplitudes of each time window served as the initial features. Different organization of these initial features leads to the following three different data representations: (1) Features extracted from each electrode concatenated in a long vector of size $1 \times 12 \cdot 30$. (2) Features extracted from each electrode constitute each row of a matrix of size 12×30. (3) Features extracted from each electrode ($x_{ch}(n), n = 1, ..., N$) are organized in a Hankel matrix of dimensions $L \times M$, with $N = L + M - 1$, as follows:

$$H_{ch} = \begin{bmatrix} x_{ch}(1) & x_{ch}(2) & \dots & x_{ch}(M) \\ x_{ch}(2) & x_{ch}(3) & \dots & x_{ch}(M+1) \\ \vdots & \ddots & \ddots & \vdots \\ x_{ch}(L) & x_{ch}(3) & \dots & x_{ch}(N) \end{bmatrix}$$

The resulting matrices are stacked in a tensor of size $12 \times 16 \times 15$. The use of a Hankel matrix was motivated by its crucial role in subspace identification methods for linear time invariant systems, such as spectral estimation of signals as a sum of exponentially damped sinusoids [12].

C. Classification

In order to mimic an online scenario, the first half of the trials ($N_{tr} = 70$) were used for training, and the second half ($N_{te} = 70$) was used for testing. In order to test the influence of the training set size on the performance of the classifiers, different training set sizes were tested, ranging from 10 to 70. Note the relatively low number of available training samples in each case and the high dimensionality of our feature space. Modelling with such a dataset carries the risk of overfitting. Therefore, the following classification approaches were considered.

1) Linear discriminant analysis with shrinkage: In linear discriminant analysis (LDA) a crucial role is played by the estimation of the class covariance estimator. The empirical covariance matrix, however, is systematically biased towards larger eigenvalues in case of small sample sizes. Shrinkage LDA overcomes this error by replacing the empirical covariance matrix ($\hat{\Sigma}$) with:

$$\tilde{\Sigma}(\gamma) = (1-\gamma)\hat{\Sigma} + \gamma\nu\mathbf{I} \tag{1}$$

where ν is the average of the eigenvalues of $\hat{\Sigma}$, $\mathbf{I}$ is the identity matrix, and $\gamma \in [0, 1]$ is a tuning parameter. For the latter an analytical solution has been found recently. For further details about shrinkage LDA, see [1] and references therein. Shrinkage LDA was applied on the feature vector representation described above in section II-B.

2) Spectral regularization with nuclear norm: Regularization via nuclear norm conveys a-priori structural information and therefore facilitates good classification for small sample sizes. Therefore, it particularly fits the P300 classification problem, where the signal of interest has a multilinear structure and only a limited number of training samples are available. For classification of an N-order feature tensor we consider the following model:

$$\begin{aligned} \hat{y}(A, b) &= \langle A, X \rangle + b, \\ \langle A, X \rangle &= \sum_{I_1, I_2, \dots I_N} A_{i_1, i_2, \dots i_N} X_{i_1, i_2, \dots i_N} \end{aligned} \tag{2}$$

where $X \in R^{I_1 \times I_2 \times \dots I_N}$ is the input pattern, $A \in R^{I_1 \times I_2 \times \dots I_N}$ is the classifier tensor of the same size, and b is a bias term or threshold. Decisions are made according to $\text{sign}(\hat{y}) \in \{-1, 1\}$, where 1 and -1 correspond to target and non-target stimuli, respectively. The classifier, namely the pair (A, b), is found by solving a non-smooth convex optimization problem using nuclear norm penalty:

$$\min_{A,b} F(A, b) = f(A, b) + \mu \sum_{n=1}^{N} ||A_{<n>}||_* \tag{3}$$

$$f(A, b) = \sum_{n=1}^{N} (\hat{y_k}(A, b) - y_k)^2, \tag{4}$$

where $f(A, b)$ is the quadratic error function accounting for the misclassification. Model selection, namely the selection of the tuning parameter $\mu \geq 0$ was done according to the five-fold cross-validation of the misclassification error. Finally, $||A_{<n>}||_*$ is the nuclear norm of the n-mode matrix unfolding of the tensor A with singular values σ_i :

$$||A_{<n>}||_* = \sum_i \sigma_i. \tag{5}$$

The nuclear norm regularization approach was applied on the matrix and tensor based feature representation described in section II-B. We refer to these appoaches as mNNL and tNNL, respectively. For a CVX implementation of the algorithm solving (3), see [9].

III. Results

Figure 1 shows the performance of each classification approach in the inside and outside conditions using $N_{tr} = 70$ training samples for each subject. Each classifier performed better than chance level, although higher mean accuracies were achieved in the inside condition (LDA: 74%, mNNL: 77% and tNNL: 81%) than in the walking outside condition (LDA: 66%, mNNL: 70% and tNNL: 72%). Note that the P300 ERP component was significantly larger in the indoor condition [13]. In each condition both mNNL and tNNL outperformed LDA significantly ($p < .05$ and $p < .01$, respectively), moreover, in the inside condition tNNL outperformed mNNL significantly ($p < .05$).

Figure 2 shows the performance of each classification approach in the inside and outside conditions using different training set sizes. In the inside condition, using only 10 training samples, each classifier achieves a moderate performance, and improves significantly if more training samples are included ($p < .01$). However, in the outside condition, LDA fails to generalize from the additional training samples and maintains approximately constant performance with increasing training set size. In contrast, tNNL and mNNL are able to capture additional information even from such noisy samples and reach significantly higher performance for larger training sets ($p < .01$).

Finally, the feature weights obtained for a single subject in each classification approach is depicted in Figure 3. In case of LDA, the long feature vector is matricized in order to get a channel x time representation. For the tNNL approach the classifier tensor is unfolded along the second mode, resulting in a matrix representation where the features corresponding to the Hankel matrices obtained from each channel are concatenated. Note that the classifier matrix and tensor obtained by the mNNL and tNNL approaches clearly resemble P300 properties (maximal feature weights at channel Pz around 300ms after stimulus), as opposed to the unstructured feature weights obtained by LDA.

IV. Discussion and conclusions

In this work we consider the problem of classifying high dimensional, low sample size data, namely an auditory P300 paradigm recorded under normal and under noisy conditions. We showed that spectral regularization with nuclear norm outperforms shrinkage LDA in both conditions. Moreover, we demonstrate that under noisy conditions, while shrinkage LDA overfits and fails to benefit from additional training samples, spectral regularization can generalize from such a diverse manifold. Initially support vector machines with linear and with RBF kernels were also considered, however, as their performance was comparable to shrinkage LDA the results are not shown here. Due to its low computational cost, LDA remains a popular classification method. Computational complexity of the nuclear norm regularization approach scales with the order of the tensor. However, for small training set sizes and for second and third order tensors it remains within reasonable limits [14]. The strength of spectral regularization with nuclear norm penalty lies within conveying structural information from the higher order input arrays. In our study this means conveying spatiotemporal structure in the matrix case and spatial-temporal-spectral structure in the tensor case. Such a-priori structural information facilitates good generalization even in case of small sample sizes and in the presence of noise.

Acknowledgement

Support by research council kul: phd/postdoc grants, goa manet, pfv/10/002 (optec), fwo: g.0427.10n (integrated eeg-fmri), g.0108.11 (compressed sensing) iwt: tbm080658-mri (eeg-fmri), iminds 2013, iuap P719/ (dysco), cluster of excellence hearing4all (dfg).

References

[1] S. Lemm, B. Blankertz, T. Dickhaus, and K.-R. Müller, "Introduction to machine learning for brain imaging," *NeuroImage*, vol. 56, no. 2, pp. 387–99, 2011, multivariate Decoding and Brain Reading.

[2] F. Lotte, M. Congedo, A. Lécuyer, F. Lamarche, and B. Arnaldi, "A review of classification algorithms for eeg-based brain-computer interfaces," *J. Neural Eng.*, vol. 4, pp. 1–13, 2007.

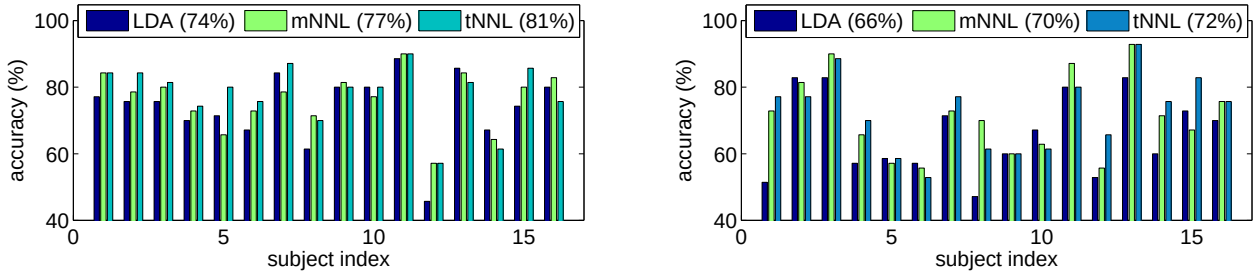
[3] H. Ramoser, J. Muller-Gerking, and G. Pfurtscheller, "Optimal spatial filtering of single trial eeg during imagined hand movement," *IEEE T. Rehabil. Eng.* , vol. 8, pp. 441–46, 2000.

[4] G. Pires, U. Nunes, and M. Castelo-Branco, "Statistical spatial filtering for a p300-based bci: Tests in able-bodied, and patients with cerebral palsy and amyotrophic lateral sclerosis," *J. Neurosci Meth.*, vol. 195, no. 2, pp. 270–81, 2011.

[5] A. Phan and A. Cichocki, "Tensor decompositions for feature extraction and classification of high dimensional datasets," *Nonlinear theory and its applications, IEICE*, vol. 1, pp. 37–68, 2010.

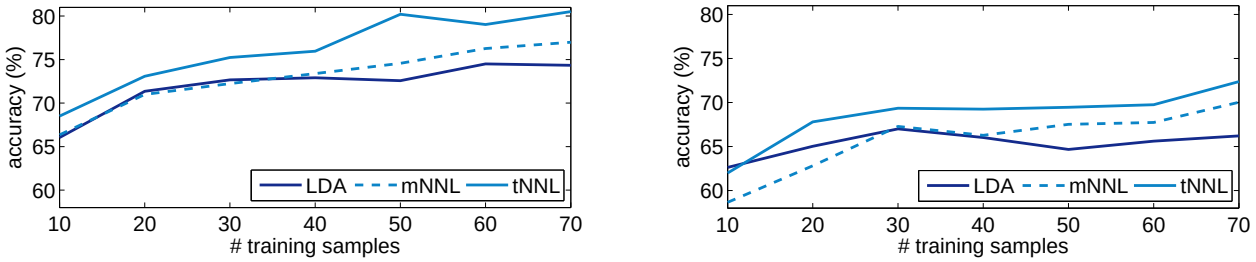
[6] X. He, D. Cai, and P. Niyogi, "Tensor subspace analysis," in *Advances in Neural Information Processing Systems (NIPS)*, Y. Weiss, B. Schölkopf, and J. Platt, Eds. MIT Press, 2006, pp. 499–506.

[7] R. Tomioka and K. Aihara, "Classifying matrices with a spectral regularization," in *Proc. of the 24th International Conference on Machine Learning (ICML'07)*, 2007, pp. 895–902.



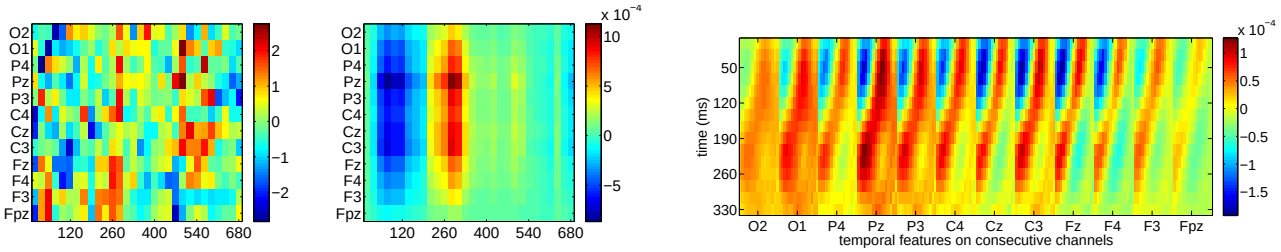
(a) Inside (b) Outside

Figure 1: Comparison of accuracies obtained with the different classification approaches for a training set size of $N_{tr} = 70$. Mean accuracies are shown in brackets.



(a) Inside (b) Outside

Figure 2: Comparison of accuracies of the different classification approaches with increasing training set size.



(a) LDA (b) mNNL (c) tNNL

Figure 3: Feature weights obtained for a single subject. In case of LDA, the long feature vector is matricized in order to get a channel x time representation. For the tNNL approach the classifier tensor is unfolded along the second mode, resulting in a matrix representation where the features corresponding to the Hankel matrices obtained from each channel are concatenated.

[8] R. Tomioka and K.R. M uller, "A regularized discriminative framework for EEG analysis with application to brain-computer interface" *NeuroImage*, vol. 49, pp. 415–32, 2010.

[9] B. Hunyadi, M. Signoretto, W. Van Paesschen, J. Suykens, S. Van Huffel, and M. De Vos, "Incorporating structural information from the multichannel eeg improves patient-specific seizure detection," *Clin. Neurophysiol.*, vol. 123, no. 12, pp. 2352–61, 2012.

[10] M. Signoretto, Q. Tran Dinh, L. De Lathauwer, and J. Suykens, "Learning with tensors : a framework based on convex optimization and spectral regularization," *Mach. Learn.*, accepted, 2013.

[11] M. De Vos, L. De Lathauwer, and S. Van Huffel, "Spatially constrained ica algorithm with an application in eeg processing," *Signal Processing*, vol. 91, no. 8, pp. 1963–72, 2011.

[12] S. Van Huffel, H. Chen, C. Decanniere, and P. Vanhecke, "Algorithm for time-domain nmr data fitting based on total least squares," *J. Magn. Reson., Series A*, vol. 110, no. 2, pp. 228–37, 1994.

[13] S. Debener, F. Minow, R. Emkes, K. Gandras, and M. De Vos, "How about taking a low-cost, small, and wireless eeg for a walk?" *Psychophysiology*, vol. 49, no. 11, pp. 1449–53, 2012.

[14] B. Hunyadi, M. De Vos, M. Signoretto, J. Suykens, W. Van Paesschen, and S. Van Huffel, "Automatic seizure detection incorporating structural information," in *Artificial Neural Networks and Machine Learning (ICANN)*, 2011, pp. 233–240.