Higher-order PLS for Classification of ERPs with Application to BCIs

Qibin Zhao[*†], Liqing Zhang[†], Jianting Cao [‡] and Andrzej Cichocki [*]
[*] Brain Science Institute, RIKEN, Japan
[†] Department of Computer Science and Engineering, Shanghai Jiao Tong University, China
[†] Saitama Institute of Technology, Saitama, Japan

***Abstract*—The EEG signals recorded during Brain Computer Interfaces (BCIs) are naturally represented by multi-way arrays in spatial, temporal, and frequency domains. In order to effectively extract the underlying components from brain activities which correspond to the specific mental state, we propose the higher-order PLS approach to find the latent variables related to the target labels and then make classification based on latent variables. To this end, the low-dimensional latent space can be optimized by using the higher-order SVD on a cross-product tensor, and the latent variables are considered as shared components between observed data and target output. The EEG signals recorded under the P300-type affective BCI paradigm were used to demonstrate the effectiveness of our new approach.**

I. Introduction

Brain computer interfaces (BCIs) are communication systems that enable subjects to transmit their intention to computers through decoding of brain activity [1], which can be used to assist patients with severe neuromuscular disabilities. The P300 event related potential (ERP), evoked in scalp-recorded electroencephalography (EEG) by external stimuli, based spelling machine is one of the most successful BCI paradigm first introduced by Farwell and Donchin [2]. In addition, a number of variations of P300-based BCI have been explored such as an apparent motion and color onset paradigm [3], the checkerboard paradigm [4] and the auditory oddball ERP [5]–[7]. We have also investigated the three oddball BCI paradigms utilizing randomly flashed images of objects, faces and emotional faces [8]. The subjects were requested to perform three different mental tasks, i.e., visual attention, face recognition (identification), emotion discrimination, corresponding to three types of images. The main objective was to find the ERP waveforms elicited by oddball faces or emotional faces stimuli and whether it is feasible to apply face-related ERPs for BCI paradigm. In contrast to the classical P300-based BCI, we investigate the multiple ERP components (e.g., VPP, N170, N250 and LPP) modulated by several different stimuli and mental tasks related to face identity and emotion recognition.

The analysis of single-trial ERP suffers from the superposition of task-relevant signals by task-unrelated brain activities, resulting in a low signal-to-noise (SNR) of the observed single-trial responses [9]. The strengths and inherent pitfalls of machine learning algorithms for decoding brain states have been systematically reviewed in [10]. Several classification techniques have demonstrated notable performance for the P300-based BCI, including stepwise linear discriminant analysis (SWLDA) support vector machines (SVM). Linear Discriminant Analysis (LDA) using a linear hyperplane to separate data from two classes under the assumption of normal distribution is widely used for BCI designs as it has been shown to be one of the most efficient classifier, especially for the P300-based BCI [11]. Although the speed and accuracy of P300 BCIs have been significantly improved by various signal processing methods [12], [13], the single-trial classification of P300 ERP remains a challenging problem due to the trial-to-trial variability. The method of partial least squares (PLS) [14] has been been found to be a useful dimension reduction technique as well as principal component analysis (PCA). The PLS can be considered as penalized canonical correlation analysis (CCA), with basically a PCA in the X space and a PCA in the Y space providing the penalties. The widely used data analysis tool for classification based on the PLS model are PLSDA [15] or OPLSDA [16], [17], in which the dependent variable is chosen to represent the class membership. This technique was successfully applied in analyzing the microarray data [18] and metabolomics data [19]. The ERP data for the BCI are characterized by both spatial and temporal variables, which are often high dimension and naturally represented as a higher-order tensor. The corresponding tensor decomposition methods [20], [21] are more suitable for modeling such data. In this paper, we introduce a higher-order partial least squares discriminant analysis (HOPLSDA) method, based on multilinear subspace regression, to extract discriminant features from multi-way arrays (tensor) representation of multichannel ERPs and to make classification for single-trial ERP recorded under our new BCI paradigm.

The rest of paper is organized as follows: Section II describes the experimental paradigm used for affective BCI and the preprocessing of the brain data. A new multilinear discriminant analysis method is described in Section III. The classification performance and results were presented in Section IV followed by the conclusion of this study.

II. Data requisition and preprocessing

A. Experimental setup

The standard stimuli, employed in the classical P300 speller, are intensifications ("flashing") of the characters. The subject attends to the desired character by silently counting the number of intensification. When the row or column containing the character lights up it elicits a P300 wave, which can be detected from the EEG. To investigate whether face-evoked potential can also be applied in an oddball BCI paradigm, stimulation ("flashing") was made by presenting a picture of face over the symbols instead of simple intensification of the letters, the subject's task was to focus only on the target stimulus from a stimuli sequence. For each target, two sequences of stimuli were presented, where each sequence contained a random series of of nine stimuli (one for each symbol).

We collected data under three experimental conditions and three types of pictures, i.e. objects, faces, emotional faces, served as stimuli, respectively. In condition 1, the subjects were asked to focus on the target item and count the number of flashing, instead of highlighting the target arrow, the images from objects group were shown randomly at each of 9 positions; In condition 2, the images from faces group were utilized for flashed targets and the subjects were asked to perform the face recognition tasks when the desired target is flashed; In condition 3, the images from emotional faces group were presented as flashed targets and the subjects were asked to perform emotion discrimination tasks whenever the desired target is flashed.

B. Preprocessing

We first filtered the signal between 0.1 and 20 Hz and then extracted the epochs corresponding to each stimulus. To avoid phase shifts the filter was applied both forward and backward in time. To ensure a reliable artifact rejection, these epochs were baseline corrected, after which all trials containing amplitudes exceeding $+/-75\mu V$ were removed. The baseline correction is performed by subtracting the average amplitude during pre-stimulus interval from the whole epoch.

If we denote each EEG epoch by $\mathbf{X} \in \mathbb{R}^{J\times K}$ with J channels and K time samples, the dimensionality of the input data amounts to $J \times K (e.g., 16 \times 256)$, while the number of training samples (epochs) is typically rather small, up to a few hundred samples. Moreover, EEG is contaminated by various sources of the artifacts or interfering noise, while task relevant discriminative information is often concentrated in a low dimensional subspace. Consequently, to avoid the overfitting of classifier, the dimensionality of the data needs to be significantly reduced, and informative features have to be extracted.

III. Higher order partial least squares discriminant analysis (HOPLSDA)

The features of EEG epoches are typically derived from spatial, spectral, and temporal domains of raw EEG signals, while the most classical classifiers are based on 2D matrices with one dimension of samples and another dimension of feature vectors. Thus, the features from multi-domains have to be concatenated, as a result the dimensionality of features becomes extremely high and spatial structure information has been destroyed. Therefore, we use multi-way arrays (tensors) to represent EEG data and apply multilinear discriminant analysis in tensor space for classification.

The multilinear regression model, termed higher-order partial least squares (HOPLS), operates by modeling independent data $\underline{\mathbf{X}}$ with a special MSVD (i.e., rank-$(1, L_2, \ldots, L_N)$ decomposition) while dependent data $\mathbf{Y}$ is modeled with rank-one decomposition. This allows us to find the optimal subspace approximation of $\underline{\mathbf{X}}$, in which the independent and dependent variables share a common set of latent vectors on one specific mode (i.e., samples mode). For the three-way independent variables $\underline{\mathbf{X}} \in \mathbb{R}^{I\times J\times K}$ and a dependent variable $\mathbf{Y} \in \mathbb{R}^{I\times M}$, with the same sample size I, we have

$$\underline{\mathbf{X}} = \sum_{r=1}^{R} \underline{\mathbf{G}}_r \times_1 \mathbf{t}_r \times_2 \mathbf{P}_r \times_3 \mathbf{Q}_r + \underline{\mathbf{E}}_R, \quad (1)$$

$$\mathbf{Y} = \sum_{r=1}^{R} d_{rr} \mathbf{t}_r \mathbf{c}_r^T + \mathbf{F}_R, \quad (2)$$

where R is the number of latent vectors, $\mathbf{t}_r \in \mathbb{R}^I$ is the rth latent vector, $\mathbf{P}_r \in \mathbb{R}^{J\times L_2} (J \geqslant L_2), \mathbf{Q}_r \in \mathbb{R}^{K\times L_3} (K \geqslant L_3)$ are loading matrices corresponding to the latent vector $\mathbf{t}_r$ in mode-2 and mode-3 respectively. Tensors $\underline{\mathbf{G}}_r \in \mathbb{R}^{1\times L_2\times L_3}$ are core tensors describing the interaction level between each latent vector and a set of loading vectors on each mode.

If we define $\underline{\mathbf{Z}} = \underline{\mathbf{X}} \times_1 \mathbf{Y}$, the parameters $\mathbf{P}, \mathbf{Q}, \mathbf{c}$ can be learned by maximizing the objective function

$$\max_{\mathbf{c},\mathbf{P},\mathbf{Q}} \|\underline{\mathbf{Z}} \times_1 \mathbf{c}^T \times_2 \mathbf{P}^T \times_3 \mathbf{Q}^T\|^2,$$
$$\text{s. t.} \quad \mathbf{P}^T\mathbf{P} = \mathbf{I}_{L_2}, \mathbf{Q}^T\mathbf{Q} = \mathbf{I}_{L_3} \text{ and } \|\mathbf{c}\| = 1, \quad (3)$$

indicating that instead of decomposing $\underline{\mathbf{X}}$ directly, we may opt to find a rank-$(1, L_2, L_3)$ tensor decomposition of $\underline{\mathbf{Z}}$ by keeping only one component in the first mode, and the L_2 and L_3 components in the other two modes.

Since MSVD is performed by looking for orthogonal coordinate transformations of $R^{I_1}, R^{I_2}, \ldots, R^{I_N}$ that induce a particular representation of the higher order tensor [22], the optimization problem in (3) can be solved by using MSVD on the Nth-order tensor $\underline{\mathbf{Z}}$. After estimating all the factors $\mathbf{P}, \mathbf{Q}$ and $\mathbf{c}$, the core tensor $\underline{\mathbf{G}} \in \mathbb{R}^{1\times L_2\times L_3}$ with mode-1 size of 1 can be computed as

$$\underline{\mathbf{G}} = \underline{\mathbf{Z}} \times_1 \mathbf{c}^T \times_2 \mathbf{P}^T \times_3 \mathbf{Q}^T. \quad (4)$$

The latent vector $\mathbf{t}$ can be estimated in the least squares sense, yielding the best approximation of $\underline{\mathbf{X}}$ with given factors $\mathbf{P}, \mathbf{Q}$ and core tensor $\underline{\mathbf{G}}$. Then, upon defining $\mathbf{G}_{(1)}^{+} \in \mathbb{R}^{L_2 L_3\times 1}$ as the Moore-Penrose pseudo-inverse of $\mathbf{G}_{(1)}$, the normalized latent vector $\mathbf{t}$ can be calculated as

$$\mathbf{t} = (\underline{\mathbf{X}} \times_2 \mathbf{P}^T \times_3 \mathbf{Q}^T)_{(1)} \mathbf{H}_{(1)}^{+}, \ \mathbf{t} = \mathbf{t}/\|\mathbf{t}\|. \quad (5)$$

The above described procedure finds the first set of parameters (i.e., r=1) of $\mathbf{P}_r, \mathbf{Q}_r$ and $\mathbf{c}_r$ which extract the first common latent vectors of $\mathbf{t}_r$. The same procedure can be carried out repeatedly after deflation of both $\underline{\mathbf{X}}$ and $\mathbf{Y}$ until the appropriate number of components is obtained.

Let $\underline{\mathbf{X}} \in \mathbb{R}^{I \times J \times K}$ denotes EEG data with I epochs, J channels and K time samples, $\mathbf{Y} \in \mathbb{R}^{I \times M}$ denote dummy variables describing group membership of corresponding EEG epochs M classes. The coding of dummy matrix is represented as

$$y_{im} = \begin{cases} 1, & \mathbf{X}_i \in \text{class}_m; \\ 0, & \text{otherwise.} \end{cases} \tag{6}$$

Since our goal is to perform dimension reduction and extract the discriminant features from the higher-order tensor $\underline{\mathbf{X}}$ according to the class information encoded in $\mathbf{Y}$, HOPLS can be applied to model such data and extract the common latent variables $\mathbf{T}$ which explains variance of $\underline{\mathbf{X}}$ and $\mathbf{Y}$ as much as possible. In other words, $\mathbf{T}$ is considered as the underlying features in a new tensor subspace, spanned by a set of loading matrices $\mathbf{P}, \mathbf{Q}$, which have the maximum covariance with the corresponding class labels. Therefore, the latent variables are the most discriminant features for classification and the number of latent variables is much smaller than the number of original features in tensor $\mathbf{X}$. Since HOPLS is specially suited to deal with a much larger number of variables than observations and with multicollinearity, which are two of main problems encountered when analysing ERP data, it is reasonable to expect that HOPLS should perform well for discriminant analysis.

IV. Results

Typically the ERP classification has been performed separately in the temporal and spatial domain, for instance, classification on temporal features is to determine which channels contribute most to the discrimination task, and classification on spatial features demonstrates which time intervals are most important. This investigation provides a way to interpret the spatio-temporal patterns of EEG exploited by the classifier. Actually, the spatio-temporal features of ERP are concatenated for classification resulting in the problem of much larger variables than samples and difficulties of interpretation of discriminative patterns in spatial and temporal domains. Therefore, in this study, HOPLSDA was applied to extract the most discriminant features based on tensor subspace regression model followed by classification on relatively small number of latent features.

As the P300-based BCIs are actually based on the binary classification of epochs, i.e. target vs. non-target, each epoch corresponding to the flash of one specific symbol. The epoch corresponding to the largest classification confidence is considered as online output of BCI. Therefore, for offline analysis, we investigated binary classification based on epochs and evaluated the performance using 5×5-fold cross validation. Since the number of training samples for target and non-target are unbalanced, the true positive rate (TPR or sensitivity), false positive rate (FPR) and true negative rate (TNR or specificity) provide better evaluation than classification accuracy.

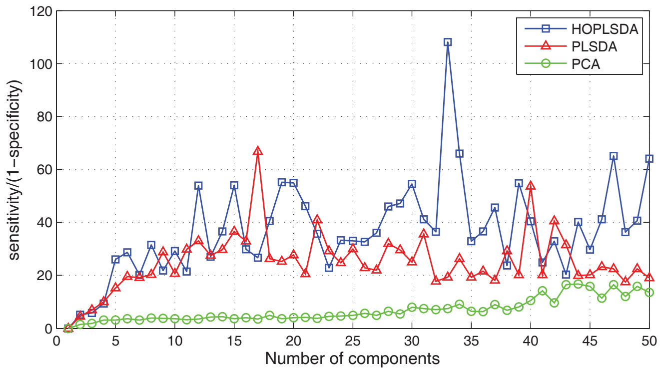

Fig. 1. 5×5-fold cross-validation performance of HOPLSDA, PLSDA and PCA with varying number of the latent components

To investigated the discriminant ability of our proposed HOPLSDA, PLSDA and PCA were also employed for feature extraction or dimension reduction followed by LDA/SVM classifier. There are several tuning parameters in HOPLSDA, which need to learn from the training data, such as number of latent components and number of loadings in spatial and temporal mode. Fig. 1 depicts the offline performance of HOPLSDA, PLSDA, and PCA with varying number of latent components, illustrating the superiority of HOPLSDA with respect to discriminant ability as compared to PLSDA and PCA. The optimal number of latent components are 33 for HOPLSDA, 17 for PLSDA and 43 for PCA. The optimal loading number for HOPLSDA is $L_2 = 4$ in the spatial mode and $L_3 = 2$ in the temporal mode. The data distribution in the latent variable space, i.e. feature space, are shown in fig. 2. Observe that HOPLSDA is able to find the new tensor subspace providing the most discriminant features not only for training data but also for test data. PLSDA also performed well on the training data but lack of generalization ability on the test data, which might related to the overfitting problem due to the small samples-to-feature ratios. It is also interesting that the direction for minimum within-class covariance is almost same with the direction for maximum inter-class covariance, illustrating clearly the power of HOPLS for discriminant analysis.

Another promising property of HOPLSDA is the meaningful interpretation of discriminant patterns with regard to spatial and temporal domains. According to Eq. (1) in HOPLSDA, $\mathbf{P}_r$ and $\mathbf{Q}_r$ represents the discriminant patterns corresponding to the rth latent component in spatial and temporal domains, respectively, providing the plausible neurophysiology interpretation. Fig. 3 shows the spatial and temporal discriminant patterns with regard to the first latent component. Since Tucker model is employed in HOPLSDA, the optimal number of spatial loadings is ($L_2 = 4$) and the number of temporal loadings ($L_3 = 2$). It is clear that ERP components (e.g. VPP, N170, N250, LPP) were separated in temporal domains. For instance, the temporal loading of object stimuli illustrates

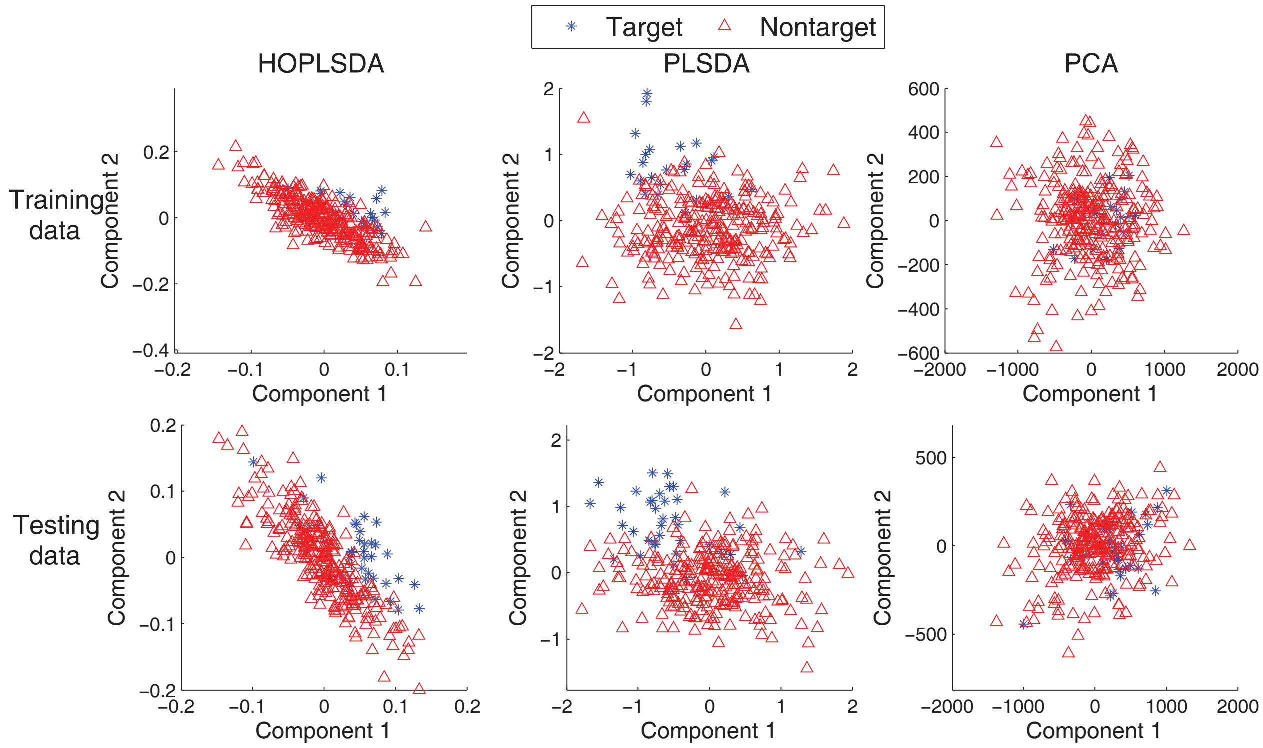

Fig. 2. The feature distribution in latent variables space for HOPLSDA, PLSDA, and PCA. Observe that HOPLSDA provided the most discriminant features either on training data or on test data, followed by PLSDA which performed well on training data but lack of generalization ability. Both HOPLS and PLSDA outperformed PCA with respect to discriminant ability.

the most discriminant ability of LPP at 400-800ms and P300 at 200-400ms; the temporal patterns of face stimuli mostly lies in N170 and N250. For affective face stimuli, it is quite clear that LPP at 400-800ms is the most important feature, N170 and P300 also make contribute for classification. The similar spatial patterns under these three experimental conditions were observed, including centro-parietal and occipitio-temporal region. These results implies that multiple ERP components are modulated to some extend by the specific mental task. ERP components related to visual processing and emotion processing have been used together, proving the best classification performance, in the proposed affective paradigm.

V. Conclusions

In summary, a new tensor-based HOPLS method was applied for discriminative analysis of ERPs evoked from the affective BCI paradigm. The proposed approach can not only extract low-dimensional features by collaborative spatio-temporal multilinear transformation, but also provides us meaningful interpretation for the spatial and temporal patterns. The classification performance illustrated the superiority of proposed feature extraction framework and further demonstrated the effectiveness of our affective BCI.

Acknowledgment

The work was supported by NSFC international cooperation program (grant number 61111140019 and 61202155), JSPS under the Japan - China Scientific Cooperation Program and JSPS Grants-in-Aid for Scientific Research (grant number 24700154).

References

[1] J. Wolpaw, N. Birbaumer, D. McFarland, G. Pfurtscheller, and T. Vaughan, "Brain-computer interfaces for communication and control," *Clinical Neurophysiology*, vol. 113, no. 6, pp. 767–791, 2002.

[2] L. Farwell and E. Donchin, "Talking off the top of your head: toward a mental prosthesis utilizing event-related brain potentials," *Electroencephalography and Clinical Neurophysiology*, vol. 70, no. 6, pp. 510–523, 1988.

[3] S. Martens, N. Hill, J. Farquhar, and B. Schölkopf, "Overlap and refractory effects in a brain-computer interface speller based on the visual P300 event-related potential," *Journal of Neural Engineering*, vol. 6, p. 026003, 2009.

[4] G. Townsend, B. LaPallo, C. Boulay, D. Krusienski, G. Frye, C. Hauser, N. Schwartz, T. Vaughan, J. Wolpaw, and E. Sellers, "A novel P300-based brain-computer interface stimulus presentation paradigm: Moving beyond rows and columns," *Clinical Neurophysiology*, vol. 121, no. 7, pp. 1109–1120, 2010.

[5] A. Furdea, S. Halder, D. Krusienski, D. Bross, F. Nijboer, N. Birbaumer, and A. Kübler, "An auditory oddball (P300) spelling system for brain-computer interfaces," *Psychophysiology*, vol. 46, no. 3, pp. 617–625, 2009.

[6] F. Nijboer, A. Furdea, I. Gunst, J. Mellinger, D. McFarland, N. Birbaumer, and A. Kübler, "An auditory brain-computer interface (bci)," *Journal of neuroscience methods*, vol. 167, no. 1, pp. 43–50, 2008.

[7] M. Schreuder, B. Blankertz, and M. Tangermann, "A new auditory multi-class brain-computer interface paradigm: spatial hearing as an informative cue," *PLoS One*, vol. 5, no. 4, p. e9813, 2010.

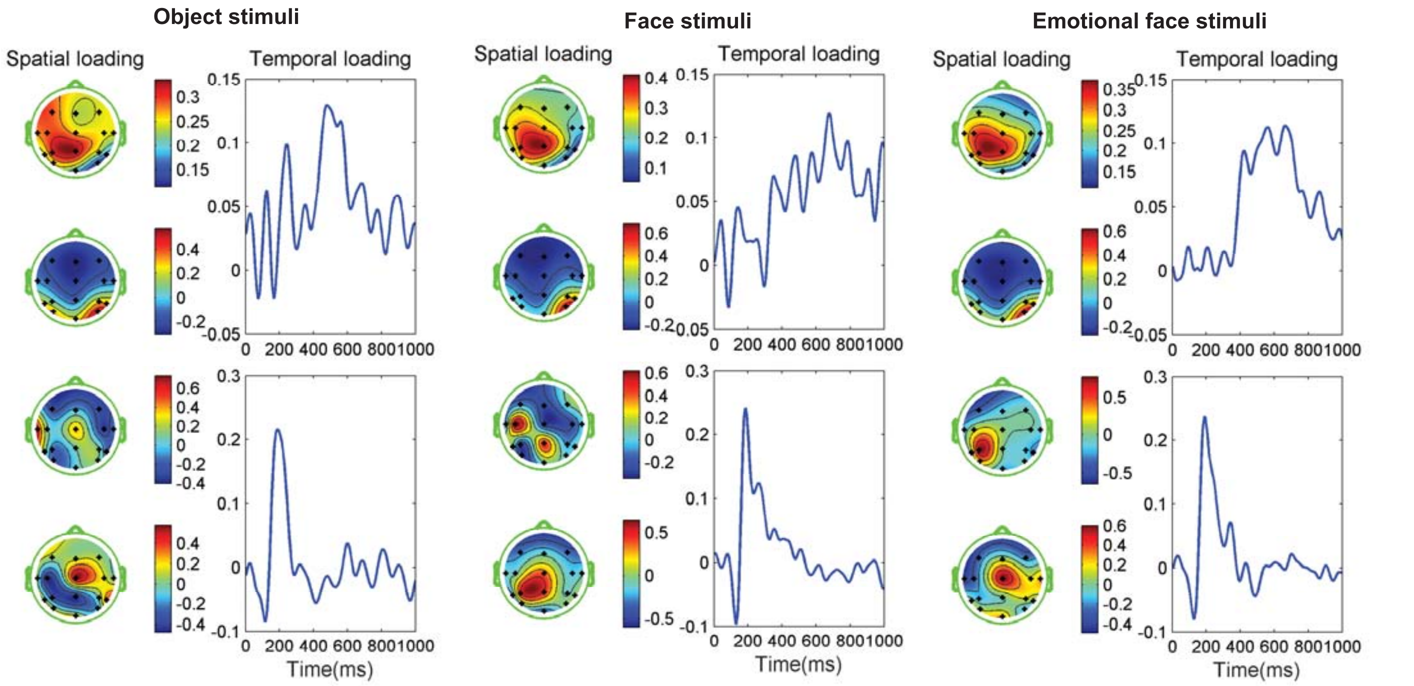

Fig. 3. Spatial and temporal patterns for the first latent component of HOPLSDA using EEG recorded from three experimental conditions.

[8] Q. Zhao, A. Onishi, Y. Zhang, J. Cao, L. Zhang, and A. Cichocki, "A novel oddball paradigm for affective bcis using emotional faces as stimuli," in *Neural Information Processing*. Springer, 2011, pp. 279–286.

[9] B. Blankertz, S. Lemm, M. Treder, S. Haufe, and K. M "uller, "Single-trial analysis and classification of erp components–a tutorial," *Neuroimage*, vol. 56, no. 2, pp. 814–825, 2011.

[10] S. Lemm, B. Blankertz, T. Dickhaus, and K. M "uller, "Introduction to machine learning for brain imaging." *NeuroImage*, vol. 56, no. 2006, pp. 387–399, 2010.

[11] D. Krusienski, E. Sellers, F. Cabestaing, S. Bayoudh, D. McFarland, T. Vaughan, and J. Wolpaw, "A comparison of classification techniques for the p300 speller," *Journal of neural engineering*, vol. 3, p. 299, 2006.

[12] A. Lenhardt, M. Kaper, and H. Ritter, "An adaptive P300-based online brain–computer interface," *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, vol. 16, no. 2, pp. 121–130, 2008.

[13] P. Xu, P. Yang, X. Lei, and D. Yao, "An enhanced probabilistic LDA for multi-class brain computer interface," *PloS One*, vol. 6, no. 1, p. e14634, 2011.

[14] H. Wold, "Soft modelling by latent variables: The non-linear iterative partial least squares (NIPALS) approach," *Perspectives in Probability and Statistics*, pp. 117–142, 1975.

[15] M. Barker and W. Rayens, "Partial least squares for discrimination," *Journal of chemometrics*, vol. 17, no. 3, pp. 166–173, 2003.

[16] M. Bylesjö, M. Rantalainen, O. Cloarec, J. Nicholson, E. Holmes, and J. Trygg, "Opls discriminant analysis: combining the strengths of pls-da and simca classification," *Journal of Chemometrics*, vol. 20, no. 8-10, pp. 341–351, 2006.

[17] J. Trygg and S. Wold, "Orthogonal projections to latent structures (o-pls)," *Journal of Chemometrics*, vol. 16, no. 3, pp. 119–128, 2002.

[18] M. Pérez-Enciso and M. Tenenhaus, "Prediction of clinical outcome with microarray data: a partial least squares discriminant analysis (pls-da) approach," *Human genetics*, vol. 112, no. 5, pp. 581–592, 2003.

[19] J. Westerhuis, E. van Velzen, H. Hoefsloot, and A. Smilde, "Multivariate paired data analysis: multilevel plsda versus oplsda," *Metabolomics*, vol. 6, no. 1, pp. 119–128, 2010.

[20] A. Cichocki, R. Zdunek, A. H. Phan, and S. I. Amari, *Nonnegative Matrix and Tensor Factorizations*. John Wiley & Sons, 2009.

[21] T. Kolda and B. Bader, "Tensor decompositions and applications," *SIAM Review*, vol. 51, no. 3, pp. 455–500, 2009.

[22] L. De Lathauwer, B. De Moor, and J. Vandewalle, "A multilinear singular value decomposition," *SIAM Journal on Matrix Analysis and Applications*, vol. 21, no. 4, pp. 1253–1278, 2000.