Contents lists available at ScienceDirect

NeuroImage

journal homepage: www.elsevier.com/locate/ynimg

A regularized discriminative framework for EEG analysis with application to brain–computer interface

Ryota Tomioka [a,b,c,*], Klaus-Robert Müller [c,d]

[a] Department of Mathematical Informatics, University of Tokyo, 7-3-1, Hongo, Bunkyo-ku, Tokyo, 113-8656, Japan
[b] Department of Computer Science, Tokyo Institute of Technology, 2-12-1-W8-74, O-okayama, Meguro-ku, Tokyo, 152-8552, Japan
[c] Machine Learning Group, Department of Computer Science, TU Berlin, Franklinstr. 28/29, 10587 Berlin, Germany
[d] Bernstein Centers for Neurotechnology and Computational Neuroscience Berlin, Germany

ARTICLE INFO

Article history:
Received 13 February 2009
Revised 7 July 2009
Accepted 17 July 2009
Available online 29 July 2009

Keywords:
Brain–computer interface
Discriminative learning
Regularization
Group-lasso
Spatio-temporal factorization
Dual spectral norm
Trace norm
Convex optimization
P300 speller
Discriminative modeling of brain imaging signals

ABSTRACT

We propose a framework for signal analysis of electroencephalography (EEG) that unifies tasks such as feature extraction, feature selection, feature combination, and classification, which are often independently tackled conventionally, under a regularized empirical risk minimization problem. The features are automatically learned, selected and combined through a convex optimization problem. Moreover we propose regularizers that induce novel types of sparsity providing a new technique for visualizing EEG of subjects during tasks from a discriminative point of view. The proposed framework is applied to two typical BCI problems, namely the P300 speller system and the prediction of self-paced finger tapping. In both datasets the proposed approach shows competitive performance against conventional methods, while at the same time the results are easier accessible to neurophysiological interpretation. Note that our novel approach is not only applicable to Brain imaging beyond EEG but also to general discriminative modeling of experimental paradigms beyond BCI.

Introduction

Brain–computer interface (BCI) is a rapidly growing field of research combining neurophysiological insights, statistical signal analysis, and machine learning (Wolpaw et al., 2002; Dornhege et al., 2007; Curran and Stokes, 2003; Kübler et al., 2001; Birbaumer et al., 1999; Penny et al., 2000; Parra et al., 2002; Pfurtscheller et al., 2006; Blankertz et al., 2006a, 2007). The goal of BCI research is to build a communication channel from the brain to computers bypassing peripheral nerves and muscle activity (Wolpaw et al., 2002). This can help people who have damage in their peripheral pathway to recover their communication abilities (e.g. Birbaumer et al. (1999); Kübler et al. (2001); Nicolelis (2003); Hochberg et al. (2006)).

Among different techniques for the noninvasive measurement of the human brain, the electroencephalography (EEG) is commercially affordable and has excellent temporal resolution which enables real-time interaction through BCI. Thus our primary focus in this paper is on EEG-based BCI but the techniques presented can also be applied to other brain imaging techniques such as magnetoencephalography (MEG) or fMRI. Note furthermore that discriminative techniques are a valuable tool for a computational analysis of neuroscience experiments beyond BCI (e.g. Haynes and Rees (2006); Parra et al. (2005)).

Based on a short segment of EEG called a trial, the signal analysis in BCI aims to predict the brain state of a user out of prescribed options (e.g. foot vs. left-hand motor imagery vs. rest). In machine learning terms, this is a multi-class classification problem. The challenge in EEG-based BCI is the low spatial resolution caused by volume conduction, the high artifact and outlier content of the signal and the mass of data that makes the application of conventional statistical analysis difficult. Therefore many studies have focused on how to extract a small number of task informative features from the data (see e.g. Dornhege et al. (2007); Blankertz et al. (2008)) that can be fed into some relatively simple classifiers; commonly used are linear spatial filtering methods (e.g., common spatial pattern (Ramoser et al., 2000; Blankertz et al., 2008) or independent component analysis (Hyvärinen et al., 2001)) coupled with heuristic frequency band selection (Blankertz et al., 2008) or band weighting (Tomioka et al., 2006b; Wu et al., 2008). One of the shortcomings of the feature extraction approaches is the strong and hard-to-control inductive bias

* Corresponding author. Department of Mathematical Informatics, University of Tokyo, 7-3-1, Hongo, Bunkyo-ku, Tokyo, 113-8656, Japan. Fax: +81 3 5841 6897.
E-mail addresses: tomioka@mist.i.u-tokyo.ac.jp (R. Tomioka), krm@cs.tu-berlin.de (K.-R. Müller).

1053-8119/$ – see front matter © 2009 Elsevier Inc. All rights reserved.
doi:10.1016/j.neuroimage.2009.07.045

that limits their application to rather specific experimental paradigms that they are developed for. Another approach is the *discriminative approach* that tries to optimize the classifier coefficients from the training data under a unified criterion(Dyrholm and Parra, 2006; Dyrholm et al., 2007; Christoforou et al., 2008; Farquhar et al., 2006; Tomioka et al., 2007; Tomioka and Aihara, 2007). The theoretical advantage of the discriminative approach is that the coefficients (e.g., spatial filter and temporal filter) are jointly optimized under a single criterion. Moreover, inductive bias can be controlled in a principled manner through *regularization* (Tikhonov and Arsenin, 1977; Vapnik, 1998). However many previous studies had to solve non-convex optimization problems (Dyrholm et al., 2007; Tomioka et al., 2007), which can be challenging because of multiple local minima and difficulty in terminating the learning algorithms.

In this paper, we contribute to the discriminative approach in the following three issues. First, we combine probabilistic data-fit criteria with sparse regularizers. The proposed regularizers naturally induce sparse or factorized models through a *convex* optimization problem; moreover the number of components is automatically determined. Second, we propose a probabilistic decoding model for P300 evoked response based BCI; in addition we show that the decoding model can be instantly converted into a loss function that is used for the training of the classifier; thus no intermediate goal such as binary classification needs to be imposed. Finally, we show how first-order and second-order information in the signal (see Christoforou et al. (2008)) can be combined and selected in a systematic manner through the dual spectral (DS) regularization (Fazel et al., 2001; Srebro et al., 2005; Tomioka and Aihara, 2007). The issue of complexity control, feature extraction, and the interpretability of the resulting model is now tackled in a unified and systematic manner under the roof of a convex regularized empirical risk minimization problem.

This paper is organized as follows. In Signal analysis framework section, our discriminative learning framework is presented. In P300 speller BCI section, the proposed framework is applied to the P300 speller BCI problem. In Self-paced finger tapping problem section, the framework is applied to the problem of predicting self-paced finger tapping. The results for the two BCI problems are given in Results: P300 speller BCI section and Results: self-paced finger tapping dataset section, respectively. On the P300 problem, the proposed approach shows comparable performance to the winner of the BCI competition (Blankertz et al., 2006b; Rakotomamonjy and Guigue, 2008) using only a loss criterion derived from a novel predictor model and regularization. Different aspects of the discriminative information captured by different regularizers are discussed. On the self-paced problem, the proposed approach shows competitive performance to the winner of the competition (Blankertz et al., 2004; Wang et al., 2004) and recently proposed second-order bilinear discriminant analysis model (Christoforou et al., 2008). Our proposed DS regularization provides a principled way of learning, selecting, and combining different sources of information. Short discussions are given at the end of each section. Earlier studies on discriminative approaches to BCI are discussed in Discussion on earlier discriminative approaches section. Concluding remarks are given in Conclusion section.

Materials and methods

Signal analysis framework

In this section we present our discriminative learning framework for brain–computer interface. The framework consists of three components. The first is a probabilistic predictor model that is used for both *decoding* the intention of a user[1] and *learning* the predictor model from a collection of trials. The second component is the design of a detector function. The last component is how to appropriately control the complexity of the detector function. These three issues are presented in Discriminative learning, Detector function, and Regularization sections, respectively.

Discriminative learning

In any BCI system, the goal of signal analysis is to construct a function that predicts the intention of a user from his/her brain signal. In our discriminative approach we are interested in the whole function from the brain signal to the probability distribution over possible user intention, which we call a predictor. When we deal with this type of probabilistic predictor we are facing two tasks. First, how to decode the intention of a user given the brain signal and the predictor. Second, how to *learn* the predictor from a collection of labeled examples. The answers to these questions are derived naturally from probability theory and statistics.

Let $\boldsymbol{X} \in \mathcal{X}$ be the input brain signal and let $q(Y|\boldsymbol{X})$ be the *predictor*, which assigns probabilities to the user's command $Y \in \mathcal{Y}$ given the brain signal $\boldsymbol{X}$. The task of decoding is to find the most likely command $\hat{y}$ given the input $\boldsymbol{X}$ and the predictor q as follows:

$$\hat{y} = \underset{y \in \mathcal{Y}}{\operatorname{argmax}}\, q\,(Y = y\,|\boldsymbol{X}). \tag{1}$$

The task of learning is to find a predictor from a suitably chosen collection of candidates, which we call a *model*, and we assume that a model is parameterized by a parameter vector $\theta \in \Theta$. We denote the predictor specified by θ as q_θ; thus the model is a set $\{q_\theta : \theta \in \Theta\}$. In order to say how a predictor q_θ compares to another predictor $q_{\theta'}$, it is necessary to define a loss function. We can consider the probability that the predictor assigns to each possible user intention y as the payoff the predictor can obtain if the actual intention coincides with the prediction; the predictor can choose its strategy between equally distributing the probability mass over all the possible outcomes and concentrating it on a single output that is based on the brain signal $\boldsymbol{X}$. This payoff is commonly measured in the logarithmic scale. The loss function is thus defined as the negative logarithmic payoff (or the Shannon information content in information theory (MacKay, 2003)) as follows:

$$\ell((\boldsymbol{X}, y), \theta) = -\log q_\theta(Y = y|\boldsymbol{X}), \tag{2}$$

where $\boldsymbol{X}$ is the brain signal and y is the true intention of the user. Thus the loss is smaller if the predictor predicts the actual intention of the user with high confidence.

Suppose we are given a collection of input signal $\boldsymbol{X}_i$ and true intention y_i, which we denote $\{\boldsymbol{X}_i, y_i\}_{i=1}^n$. It is reasonable to choose the parameter θ that minimizes the empirical average of losses (see MacKay (2003, Chap. 39)):

$$L_n(\theta) = \frac{1}{n}\sum_{i=1}^{n} \ell((\boldsymbol{X}_i, y_i), \theta).$$

However, often the complexity of the class of predictors q_θ is very large and the minimization of $L_n(\theta)$ leads to overfitting due to small sample size. Therefore, we learn the parameter θ by solving the following constrained minimization problem:

$$\underset{\theta \in \Theta}{\text{minimize}} \;\; L_n(\theta) \;\text{subject to}\; \Omega(\theta) \leq C. \tag{3}$$

The second term $\Omega(\theta)$ is called the regularizer and it measures the complexity of the parameter configuration θ. C is a hyper-parameter that controls the complexity of the model and is selected by cross-validation. A complexity function induces a nested sequence of subsets $\Theta_C := \{\theta \in \Theta : \Omega(\theta) \leq C\}$, which is parameterized by the bound C on the complexity; i.e., $C_1 < C_2 < C_3 < \cdots$ implies $\Theta_{C_1} \subset \Theta_{C_2} \subset \Theta_{C_3} \subset \cdots$

[1] Here also other neuroscience paradigms than BCI can readily be used.

and vice versa. Therefore we can consider a sequence of predictors that we obtain through the learning framework (Eq. (3)) at monotonically increasing level of complexity (see Vapnik (1998)).

If we suppose that the training examples $\{\boldsymbol{X}_i, y_i\}_{i=1}^{n}$ are sampled independently and identically from some probability distribution $p(\boldsymbol{X}, Y)$, the above function $L_n(\theta)$ can be considered as the empirical version of the following function $L(\theta)$:

$$L(\theta) = D(p(Y|\boldsymbol{X})||q_\theta(Y|\boldsymbol{X})) + H(p(Y|\boldsymbol{X})),$$

where $D(p||q)$ is the Kullback–Leibler divergence between two probability distributions p and q (see e.g., MacKay (2003); Bishop (2007)); the second term is the conditional entropy of Y given $\boldsymbol{X}$ and is a constant that does not depend on the model parameter θ.

Logistic model. For example, the logistic regression model is a popular model in a binary decision setting. The logistic model assumes the user command Y to be either one of the two possibilities; e.g., $Y = -1$ and $Y = +1$ for left and right-hand movement, respectively. The logistic predictor q_θ is defined through a latent function f_θ; we define a real valued function f_θ which outputs a positive number if $Y = +1$ is more likely than $Y = -1$ and vice versa. Then a logistic function $u(z) = 1/(1 + \exp(-z))$ (see Fig. 1) is applied to the output $f_\theta(\boldsymbol{X})$ to convert it into the probability of $Y = +1$ given $\boldsymbol{X}$; similarly applying the logistic function to $-f(\boldsymbol{X})$ gives the probability of $Y = -1$ given $\boldsymbol{X}$. Thus we have the following expression for the predictor:

$$q_\theta(Y = y|\boldsymbol{X}) = \frac{1}{1 + \exp(-yf_\theta(\boldsymbol{X}))} \quad (y \in \{-1, +1\}). \tag{4}$$

In fact, under the predictor $q_{\theta'}$ defined above, the log likelihood ratio of $Y = +1$ to $Y = -1$ given $\boldsymbol{X}$ is precisely the latent function value $f_\theta(\boldsymbol{X})$ as follows:

$$\log\frac{q_\theta(Y = +1|\boldsymbol{X})}{q_\theta(Y = -1|\boldsymbol{X})} = f_\theta(\boldsymbol{X}).$$

The loss function for the logistic model is called the logistic loss and can be written as follows:

$$\ell_L((\boldsymbol{X}, y), \theta) = \log\left(1 + e^{-yf_\theta(\boldsymbol{X})}\right), \tag{5}$$

which is obtained by taking the negative logarithm of Eq. (4). As shown above, it is often a useful strategy to construct a model as a combination of a class of functions that converts the input signal into a scalar value and a link function that converts this value into the probability of the command Y. In fact, we study models with a multi-class extension of logistic link function in P300 speller BCI section and another model that uses the logistic link function in Self-paced finger tapping problem section. The function f_θ is called a *detector* in this article because in the BCI context it captures some characteristic spatio-temporal activity in the brain; a class of functions parameterized by $\theta \in \Theta$ is called a detector model. Furthermore, we review different recent approaches in modeling detector functions f_θ in Discussion on earlier discriminative approaches section.

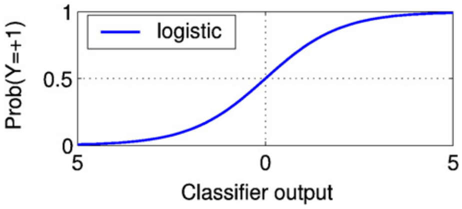


Fig. 1. Logistic function (see Eq. (4)).

Detector function

We use the following linear detector function throughout this article:

$$f_\theta(\boldsymbol{X}) = \langle \boldsymbol{W}, \boldsymbol{X} \rangle + b, \tag{6}$$

where $\theta := (\boldsymbol{W}, b)$, $\boldsymbol{W}$ is a matrix of some appropriate size and $b \in \mathbb{R}$ is a bias term. $\langle \boldsymbol{W}, \boldsymbol{X} \rangle = \sum_{ij} \boldsymbol{W}(i, j)\ \boldsymbol{X}(i, j)$ is the inner product between two matrices $\boldsymbol{X}$ and $\boldsymbol{W}$ ($\boldsymbol{W}(i, j)$ denotes the (i, j) element of a matrix $\boldsymbol{W}$).

In the simplest case, $\boldsymbol{X}$ is a short segment of appropriately filtered EEG signal with C channels and T sampled time-points, i.e., $\boldsymbol{X}$ and $\boldsymbol{W}$ are both $T \times C$ matrices. The detector is called the first-order detector in this case. This model can be used to detect slow change in the cortical potential (Blankertz et al., 2006a) and evoked response such as P300 (Farwell and Donchin, 1988) and the error potential (Schalk et al., 2000).

When we are also interested in the second-order information such as variance and covariance, we can set $\boldsymbol{X}$ as a block diagonal concatenation of these terms as follows:

$$\boldsymbol{X} = \begin{pmatrix} \frac{1}{\eta_1}\Xi^{(1)} & & & \\ & \frac{1}{\eta_{2,1}}\Xi^{(2,1)} & & \\ & & \ddots & \\ & & & \frac{1}{\eta_{2,K}}\Xi^{(2,K)} \end{pmatrix}, \tag{7}$$

where $\Xi^{(1)}$ is the first-order term ($\boldsymbol{X}$ in the above first-order model) and $\Xi^{(2,k)} = cov(\boldsymbol{X}^{(k)})$ is the covariance matrix[2] of a short segment of band-pass filtered EEG $\boldsymbol{X}^{(k)}$ for $k = 1, \dots, K$. Here we consider K second-order terms that are filtered by different (possibly overlapped) band-pass filters. We call $\boldsymbol{X}$ the augmented input matrix and the corresponding $\boldsymbol{W}$ the augmented weight matrix. The normalization factor η_* is introduced in order to prevent biasing the selection of terms with large power or large size; it is defined as the square root of the total variance[3] of each block element, i.e., $\eta_* = \sqrt{\sum_{j,k} \mathrm{var}(\Xi^*(j,k))}$ where $* \in \{(1), (2, 1), \dots, (2, K)\}$. This choice is motivated by the common practice in the ℓ_1-regularization (or lasso (Tibshirani, 1996)) to standardize each feature to unit variance. In fact, when all the block diagonal matrices are 1×1, the DS regularization (see Regularization section) reduces to lasso and the above η_* reduces to the standard deviation of each feature.

It can be shown that when we learn the augmented weight matrix $\boldsymbol{W}$ under suitable regularization (see Eq. (3)), the weight matrix turns out to have the same block diagonal structure as the input $\boldsymbol{X}$. This model is called the second-order detector. This model can be used to detect oscillatory features such as event-related desynchronization which is useful in detecting real or imagined movement (Pfurtscheller and da Silva, 1999; Pfurtscheller et al., 2000; Blankertz et al., 2006a, 2008). In these tasks it is known that both the slow change in the cortical potential and the event related desynchronization are useful features to predict the movement (Dornhege et al., 2004; Wang et al., 2004; Christoforou et al., 2008). Our contribution is to combine these features in the block diagonal form in Eq. (7).

[2] cov denotes the sample covariance matrix of the row vectors of a matrix (MATLAB cov function).

[3] var denotes element-wise sample variance with respect to a collection of matrices $\Xi_i^*(i = 1, \dots, n)$.

Regularization

In this section we preset three types of regularizers (θ) in our learning framework (Eq. (3)).

The first regularizer is the standard Frobenius norm of the weight matrix as follows:

$$\Omega_F(\theta) = \|\boldsymbol{W}\|_F = \sqrt{\langle \boldsymbol{W}, \boldsymbol{W} \rangle}, \tag{8}$$

In other words, it is the Euclidean norm of the weight matrix viewed as a vector.

The next three regularizers induce different types of *sparsity* in the weight matrix. The first two of them are defined as the "linear sum of group-wise norms", where the group is defined a priori. We assume a simple first-order detector in which the columns correspond to electrodes and rows correspond to sampled time-points; the two regularizers are called channel selection regularizer and temporal-basis selection regularizer and are defined as follows:

$$\Omega_C(\theta) = \sum_{c=1}^{C} \|\boldsymbol{W}(:,c)\|_2, \tag{9}$$

$$\Omega_T(\theta) = \sum_{t=1}^{T} \|\boldsymbol{W}(t,:)\|_2, \tag{10}$$

where $\boldsymbol{W}(:,c)$ denotes the c-th column vector of the weight matrix $\boldsymbol{W}$, $\boldsymbol{W}(t,:)$ denotes the t-th row vector of $\boldsymbol{W}$ and $\|\cdot\|_2$ is the vector Euclidean norm. In Eq. (9) each row is grouped together. Similarly in Eq. (10) each column is grouped together. Thus analogous to ℓ_1-regularization (known as lasso (Tibshirani, 1996)) the two regularizers induce sparsity in the electrode-wise (row-wise), and the time-point-wise (column-wise) manners, respectively. This type of regularization is known as group-lasso (Yuan and Lin, 2006) or M-FOCUSS (Cotter et al., 2005) and recently also applied to the reconstruction of focal vector fields (Haufe et al., 2008).

The last regularizer is defined as the linear sum of singular-values of the weight matrix $\boldsymbol{W}$, which is called the dual spectral (DS) norm (Fazel et al., 2001)[4].

$$\Omega_{\mathrm{DS}}(\theta) = \|\boldsymbol{W}\|_* := \sum_{j=1}^{r} \sigma_j(\boldsymbol{W}), \tag{11}$$

where $\sigma_j(\boldsymbol{W})$ is the j-th singular value of the weight matrix $\boldsymbol{W}$ and r is the rank of $\boldsymbol{W}$. The DS regularization can be considered as another generalization of the ℓ_1-regularization; it induces sparsity in the singular-value spectrum of the weight matrix $\boldsymbol{W}$. That is, it induces low-rank matrix $\boldsymbol{W}$. Similarly to group-lasso, when a singular-component is switched off, all the degrees of freedom associated to that component (i.e., left and right singular vectors) are simultaneously switched off. However in contrast to group-lasso regularizer, there is no notion of any group a priori. The DS regularization automatically tunes the feature detectors as well as the rank of $\boldsymbol{W}$. It is also interesting to contrast the dual spectral regularizer to the Frobenius norm regularizer (Eq. (8)). The Frobenius norm can be rewritten as the ℓ_2-norm on the singular-value spectrum as follows:

$$\begin{aligned}\Omega_F(\theta) &= \sqrt{\mathrm{Tr}(\boldsymbol{W}^\top \boldsymbol{W})} \\ &= \sqrt{\textstyle\sum_{j=1}^{r} \sigma_j^2(\boldsymbol{W})},\end{aligned} \tag{12}$$

where we used the fact that the trace of a positive semidefinite matrix is equal to the sum of its eigenvalues which equals the sum of squared singular values of $\boldsymbol{W}$. Comparing Eq. (12) and Eq. (11), we can understand the Frobenius norm and the DS norm as the ℓ_2 and ℓ_1-norm on the singular-value spectrum of a matrix, respectively. In machine learning literature, the low-rank enforcing property of the dual spectral norm is well known and has been used in applications such as collaborative filtering (Srebro, 2004; Srebro et al., 2005; Rennie and Srebro, 2005; Abernethy et al., 2006), multi-class classification (Amit et al., 2007), multi-output prediction (Argyriou et al., 2007, 2008; Yuan et al., 2007). It has been also successfully applied to the classification of motor-imagery based BCI (Tomioka and Aihara (2007), see also Discussion on earlier discriminative approaches section).

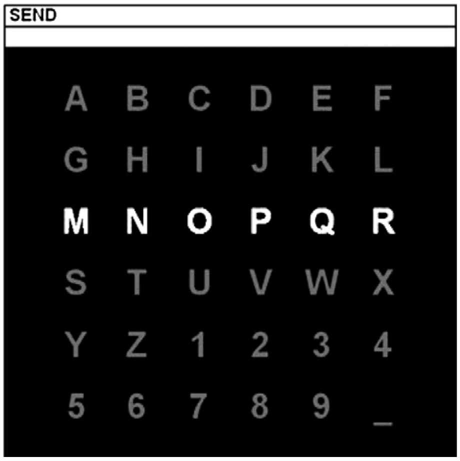
All the above regularizers give rise to some conic constraints in Eq. (3). The Frobenius and group-lasso-type regularizers (Eqs. (8)–(10)) induce the second-order cone constraint and the DS regularizer (Eq. (11)) induces the positive semidefinite cone constraint. In fact, mathematically these cones are understood as generalizations of the positive-orthant cone induced by the ℓ_1- (lasso) regularizer (Faraut and Koranyi, 1995). Some algorithmic details of the minimization in Eq. (3) are presented in Appendix A.

P300 speller BCI

In this section we apply the general framework presented in Signal analysis framework section to a brain-controlled spelling system known as P300 speller. The design of the spelling system is reviewed in P300 speller system section. The probabilistic predictor model tailored for the P300 speller system is proposed in Predictor model for P300 speller section. The details about preprocessing can be found in Signal acquisition and preprocessing section.

P300 speller system

Here we briefly describe the P300 speller system designed by Farwell and Donchin (1988). The subjects are presented a 6×6 table of 36 letters on the screen (see Fig. 2); they are instructed to focus on the letter they wish to spell for some specified period for each letter; during that period the rows and columns of the table are intensified (more specifically highlighted) in a random order. It is known that the subject's brain shows a characteristic reaction with a latency of about 300 ms called P300 when the row or column that is intensified includes the letter on which the subject is placing his/her focus. Thus by detecting the P300 response, we can predict the letter that the subject is trying to spell. Each intensification lasts 100 ms with an interval of 75 ms afterwards; the intensifications of all 6 rows and 6 columns (in a random order) are repeated 15 times; hence one letter takes 175 ms × 12 × 15 = 31.5 s. Note that the period of intensification (175 ms) is shorter than the expected reaction of the brain (300 ms). Thus the intervals we analyze are usually overlaps of several intensifications.



SEND

A B C D E F

G H I J K L

M N O P Q R

S T U V W X

Y Z 1 2 3 4

5 6 7 8 9 _

Fig. 2. Table of letters shown on the display in the P300 speller system (Farwell and Donchin, 1988). The third row is intensified.

[4] It is also known as the trace norm (Srebro et al., 2005), the Ky-Fan r-norm (Yuan et al., 2007), and the nuclear norm (Boyd and Vandenberghe, 2004).

Predictor model for P300 speller

Let the alphabet $\mathcal{A}$ be the set of all letters in the table, a trial $\boldsymbol{X}$ be a list of epochs[5] $\boldsymbol{X}=(\boldsymbol{X}_{(1)},\ldots,\boldsymbol{X}_{(12)}),\boldsymbol{X}_{(l)}\in\mathbb{R}^{T\times C}$ be the short segment of multi-channel EEG recorded after each intensification (1–6 corresponds to columns and 7–12 corresponds to rows), where C is the number of channels and T is the number of sampled time-points, and y be the true letter that the subject intends to spell during the intensifications. Inspired by Farwell and Donchin (1988) we model the predictive probability over 36 candidate letters proportional to the exponential of the sum of detector function outputs at the two corresponding row and column intensifications as follows:

$$q_\theta(y|\boldsymbol{X}) = \frac{\exp\left(f_\theta\left(\boldsymbol{X}_{(\mathrm{col}(y))}\right)+f_\theta\left(\boldsymbol{X}_{(\mathrm{row}(y)+6)}\right)\right)}{\sum_{y'\in\mathcal{A}}\exp\left(f_\theta\left(\boldsymbol{X}_{(\mathrm{col}(y'))}\right)+f_\theta\left(\boldsymbol{X}_{(\mathrm{row}(y')+6)}\right)\right)}, \tag{13}$$

where col(y) and row(y) are the indices of the column and the row of the letter y on the display (see Fig. 2). It is easy to see that the above Eq. (13) can be decomposed into a direct product of two six-class multinomial distribution as follows:

$$q_\theta(y|\boldsymbol{X}) = \frac{e^{f_\theta\left(\boldsymbol{X}_{(\mathrm{col}(y))}\right)}}{\sum_{l=1}^{6}e^{f_\theta\left(\boldsymbol{X}_{(l)}\right)}}\cdot\frac{e^{f_\theta\left(\boldsymbol{X}_{(\mathrm{row}(y)+6)}\right)}}{\sum_{l=7}^{12}e^{f_\theta\left(\boldsymbol{X}_{(l)}\right)}}. \tag{14}$$

Here f_θ $(\boldsymbol{X}_{(l)})$ is a first-order detector that outputs a scalar value for each intensification as follows:

$$f_\theta\left(\boldsymbol{X}_{(l)}\right) = \langle\boldsymbol{W},\boldsymbol{X}_{(l)}\rangle, \quad (l=1,\ldots,12), \tag{15}$$

where the weight matrix $\boldsymbol{W}$ has T rows and C columns. The bias term is omitted because the probability distribution in Eq. (14) is invariant to a constant shift of Eq. (15). Note that the parameter $\boldsymbol{W}$ is shared among all inputs $\boldsymbol{X}_{(l)}$ ($l=1,\ldots,12$). Another difference between the proposed predictor model (Eq. (14)) and the general multi-class likelihood (Bishop, 2007) is that the l-th output value only depends on the l-th input matrix $\boldsymbol{X}_{(l)}$. Furthermore, let a subtrial be the collection of six epochs within a trial with either row ($l=1,\ldots,6$) or column ($l=7,\ldots,12$) intensifications; thus a trial consists of two subtrials. Note that the contribution of the subtrials to the predictor (Eq. (14)) is independent of each other. Thus mathematically Eq. (14) is equivalent to P300 speller for six letters with two times as many trials. Note that our proposed predictor model (Eq. (13)) can also accommodate novel coding schemes for P300 speller proposed in Hill et al. (2009).

For the decoding, according to Eq. (1), we maximize the posterior probability $q(y|\boldsymbol{X})$ given $\boldsymbol{X}$ with respect to y as follows:

$$\begin{aligned}\hat{y} &= \underset{y\in\mathcal{A}}{\operatorname{argmax}}\log q_\theta(y|\boldsymbol{X})\\ &= \underset{y\in\mathcal{A}}{\operatorname{argmax}}\left(f_\theta\left(\boldsymbol{X}_{(\mathrm{col}(y))}\right)+f_\theta\left(\boldsymbol{X}_{(\mathrm{row}(y)+6)}\right)\right),\end{aligned} \tag{16}$$

which is simply to choose the column and row with maximum response.

As we have seen in the previous section, the above model is used *simultaneously* for decoding the letter and learning the parameter $\boldsymbol{W}$; according to Eq. (2)) the loss function is defined as follows:

$$\begin{aligned}\ell((\boldsymbol{X},y),\theta) = &-f_\theta\left(\boldsymbol{X}_{(\mathrm{col}(y))}\right)+\log\left(\sum_{l=1}^{6}e^{f_\theta\left(\boldsymbol{X}_{(l)}\right)}\right)\\ &-f_\theta\left(\boldsymbol{X}_{(\mathrm{row}(y)+6)}\right)+\log\left(\sum_{l=7}^{12}e^{f_\theta\left(\boldsymbol{X}_{(l)}\right)}\right).\end{aligned}$$

The above model contrasts sharply to the conventional approach that first trains a binary classifier that detects P300 response and then combines them to predict the letter (see e.g., Rakotomamonjy and Guigue (2008)) in the following way. The proposed multinomial model is normalized in a subtrial-wise manner whereas the conventional binary approach is normalized in an epoch-wise manner. More specifically, we have a budget of probability one for each subtrial that we can distribute over the epochs within the *subtrial* whereas the conventional binary approach has the same budget for each *epoch* which is distributed between the possibility that it contains P300 response or not. This epoch-wise normalization imposes stronger constraint on the detector function than our subtrial-wise normalization. In fact, the conventional binary approach tries to separate all the positive epochs (which contains P300 response) from all the negative epochs (which contains no P300 response) whereas the proposed subtrial-wise multinomial approach tries to align the positive epoch in front of the negative epochs in the same subtrial (see Fig. 3(A)). In other words, only the detector output

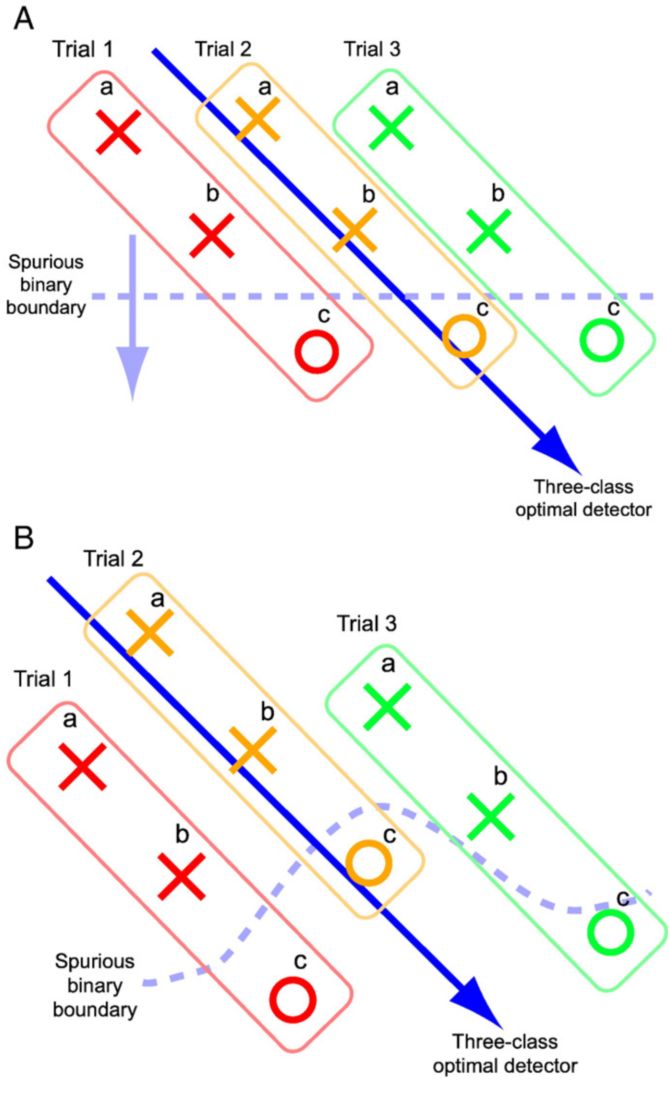


Fig. 3. Schematic comparison of our trial-wise multinomial detection approach and the conventional epoch-wise binary detection approach. Suppose the alphabet $\mathcal{A}$ consists of three letters, "a", "b", and "c" and we have three trials containing three epochs each (i.e., response after the intensification of "a", "b", and "c"). The true letter is "c" for all the three trials. Thus "a" and "b" are negative epochs (marked with crosses) and "c" are positive epochs (marked with circles). (A) Conventional binary model learns strict boundary whereas the proposed multinomial model only learns alignment. (B) The binary decision boundary can be nonlinear even when the optimal detector function is linear.

[5] In this section we reserve the term *trial* for a collection of short segments of EEG (called *epoch*) recorded after different intensifications for each character.

value in a positive epoch relative to the negative epochs *in the same subtrial* matters for the proposed model. Furthermore, even when the optimal detector function is linear, the binary decision boundary can be nonlinear as in Fig. 3(B). Moreover there is no class bias problem which arise in the conventional binary detection approach because the whole (sub)-trial is fed jointly to the predictor. Furthermore we can directly measure the letter predictor accuracy for model selection without introducing auxiliary performance measure as in Rakotomamonjy and Guigue (2008).

Signal acquisition and preprocessing

We use the P300 dataset (dataset II) provided by Jonathan R. Wolpaw, Gerwin Schalk, and Dean Krusienski in the BCI competition III (Blankertz et al., 2006b). The dataset includes two subjects namely A and B. The signal is recorded with a 64 channel EEG amplifier. We low-pass filter the signal at 20 Hz, down sample the signal to 60 Hz, and cut out an interval of 600 ms from the onset of each intensification as an *epoch* $\boldsymbol{X}_{(l)} \in \mathbb{R}^{T \times C}$ where $T=37$ and $C=64$ ($l=1, \dots, 12$). A *trial* $\boldsymbol{X} \in (\mathbb{R}^{T \times C})^{12}$ consists of 12 epochs and is assigned a single letter $y \in \mathcal{A}$. For each letter, trials (each consisting of 12 epochs) are repeated 15 times. These repetitions are simply considered as separate training examples; of course the first trial and the last trial for one letter might have different statistical character but the detector would regard this difference as inner-class variability and would become invariant as possible to the difference. Since the training set consists of 85 letters, we have $15 \cdot 85 = 1275$ training examples consisting of 12 epochs.

Before applying the learning algorithm (Eq. (3)), we apply preprocessing matrices $\boldsymbol{P}^{\mathrm{s}}$ and $\boldsymbol{P}^{\mathrm{t}}$ to the low-pass filtered signal $\boldsymbol{X}(_{l})^{\mathrm{LP}}$ as $\boldsymbol{X}_{(l)} = \boldsymbol{P}^{t}\mathbf{X}_{(l)}^{\mathrm{LP}}\,\boldsymbol{P}^{\mathrm{s}}$. The spatial and temporal preprocessing matrices $\boldsymbol{P}^{\mathrm{s}}$ and $\boldsymbol{P}^{t}$ are defined as follows. For the channel selection regularizer and the temporal-basis selection regularizer, we choose $\boldsymbol{P}^{\mathrm{s}} = \mathrm{diag}\,(\sigma_1^{\mathrm{s}}, \dots, \sigma_C^{\mathrm{s}})^{-1}$ and $\boldsymbol{P}^{\mathrm{t}} = \mathrm{diag}(\sigma_1^{t}, \dots, \sigma_T^{t})^{-1}$ where σ_c^{s} and σ_t^{t} are the square roots of the average variance of the c-th channel and the t-th time-point, respectively. This choice approximately normalizes each channel and time-point to unit variance. However it does not mix different channels or different time-points because we aim to select a few informative ones from them. For the Frobenius norm and DS norm regularizers, we chose $\boldsymbol{P}^{\mathrm{s}} = \sum^{\mathrm{s}-1/4}$ and $\boldsymbol{P}^{t} = \sum^{t-1/4}$, where $\sum^{s}$ and $\sum^{t}$ are the pooled covariance matrices in the spatial and temporal domain defined as follows:

$$\Sigma^{s} = \frac{1}{12n}\sum_{i=1}^{n}\sum_{l=1}^{12}\mathrm{cov}\left(\boldsymbol{X}_{i(l)}^{\mathrm{LP}}\right),$$

$$\Sigma^{t} = \frac{1}{12n}\sum_{i=1}^{n}\sum_{l=1}^{12}\mathrm{cov}\left(\boldsymbol{X}_{i(l)}^{\mathrm{LP}^{\top}}\right).$$

The exponent $-1/4$ is empirically found to produce a signal matrix $\boldsymbol{X}_{(l)}$ that has approximately unit variance for every element. This is because the variance of the raw signal is counted both in $\sum^{s}$ and $\sum^{t}$. In contrast to the spatial/temporal selection regularizer, there is no need to restrict the preprocessing matrices to a diagonal form because the goal is to choose a few informative pairs of spatial and temporal filters.

The test data consists of 100 letters; also 12 different intensifications are repeated 15 times (in a random order) in the test set. We report the results of (a) averaging all the 15 repetitions ($M=15$) and (b) averaging only the first 5 repetitions ($M=5$) in the prediction of each letter.

Self-paced finger tapping problem

In this section, the general framework presented in Signal analysis framework section is applied to the problem of single-trial prediction of self-paced finger tapping. The problem and the dataset is outlined in Problem setting section. In contrast to the P300 speller system, because the problem is binary classification, the choice of link function is rather simple. The challenge is how to incorporate different sources of information, namely the slow change in the cortical potential and the event-related modulation of rhythmic activity, in a principled manner. To this end, three detector functions are presented in Preprocessing and predictor model for the self-paced problem section.

Problem setting

In the self-paced finger tapping dataset (dataset IV, BCI competition 2003 (Blankertz et al., 2004)), the goal is to predict the type of upcoming voluntary finger movement before it actually occurs (Blankertz et al., 2002). EEG of a subject was recorded while the subject was typing certain keys on the keyboard at his/her own choice at the average speed of 1 key stroke per second. The subject used either the index finger or the little finger of the left hand or the right hand. Here the task is to predict whether the upcoming key press is by the left or right hand according to the task at the competition.

Preprocessing and predictor model for the self-paced problem

EEG is recorded from 28 electrodes at sampling frequency 1000Hz and down-sampled to 100 Hz. The raw signal matrix $\boldsymbol{X}^{\mathrm{raw}} \in \mathbb{R}^{T \times C}$ is a short segment of multi-channel EEG recording starting 630 ms and ending 130 ms before each key press, where $C=28$ and $T=50$. The training set contains in total 316 trials which consists of 159 left-hand and 157 right-hand trials.

Since the problem is binary we use the logistic predictor model (Eq. (4)); thus the decoding is carried out by simply taking the sign of the detector function as follows:

$$\hat{y} = \begin{cases} +1 & \text{if } f_{\theta}(\boldsymbol{X}) \geq 0, \\ -1 & \text{if } f_{\theta}(\boldsymbol{X}) < 0. \end{cases}$$

For the learning of the detector function the logistic loss function (Eq. (5)) is used in Eq. (3).

For the detector function we propose three models. The first function is a simple first-order model that only captures the slow change in the potential. Thus the weight matrix $\boldsymbol{W}$ in Eq. (6) has T rows and C columns. The input matrix $\boldsymbol{X}^{\mathrm{raw}}$ is low-pass filtered at 20 Hz and preprocessed as $\boldsymbol{X} = \boldsymbol{P}^{t}\mathbf{X}^{\mathrm{LP}}\boldsymbol{P}^{\mathrm{s}}$ with $\boldsymbol{P}^{\mathrm{t}} = \sum^{t-1/4}$ and $\boldsymbol{P}^{\mathrm{s}} = \sum^{s-1/4}$ as in the P300 speller problem (see Signal acquisition and preprocessing section in P300 speller BCI section).

The second function consists of both first-order term and a wide-band (7–30 Hz) second-order covariance term which are concatenated along the diagonal of the input matrix (see Eq. (7)). It is called the *wide-band second-order model*. The first-order term $\Xi^{(1)}$ is preprocessed in the same way as the above first-order model; the second-order term is band-pass filtered at 7–30 Hz and preprocessed with a spatial whitening matrix $\sum^{s-1/2}$, i.e., $\Xi^{(2,1)} = \sum^{s-1/2}\mathrm{cov}(\boldsymbol{X}^{\mathrm{BP}})\sum^{s-1/2}$.

Finally the last function consists of the first-order term and two second-order terms that capture the alpha-band (7–15 Hz) and the beta-band (15–30 Hz) which again form the augmented input matrix by block diagonal concatenation (see Eq. (7)). It is called the *double-band second-order model*. Similarly, the first-order term $\Xi^{(1)}$ is preprocessed in the same way as the above models; the two second-order terms $\Xi^{(2,1)}$ and $\Xi^{(2,2)}$ are band-pass filtered at 7–15 Hz and 15–30 Hz, respectively, and spatially whitened individually.

All the temporal filtering mentioned above is done using the MATLAB function filtfilt[6] because it minimizes the effect of start-up

[6] filtfilt performs zero-phase digital filtering by applying the filter in both the forward and reverse directions. This is necessary in this dataset because the signal is provided as a collection of 500ms long segments.

transients. We test the Frobenius norm regularizer as the base line as well as the proposed DS norm regularizer. Our aim is to simultaneously learn and select few informative spatio-temporal filters in a systematic manner.

Results: P300 speller BCI

The result of the proposed framework applied to P300 speller BCI (P300 speller BCI section) is given in this section. Additionally discussion including the interpretation provided by the proposed three sparse regularizers is given in Discussion section.

Performance

Figs. 4 (A–D) show the performance of the proposed decoding model with (A) Frobenius norm, (B) channel selection regularizer, (C) temporal-basis selection regularizer, and (D) DS norm regularizer, respectively. The classification accuracy (solid line) obtained at the regularization constant chosen by cross-validation is marked with a circle. The cross-validation accuracy is also shown as dashed lines with error bars; we show the mean and standard deviation of two runs of 10-fold cross-validation. Note that we compute the character recognition accuracy for each number of repetitions M on the

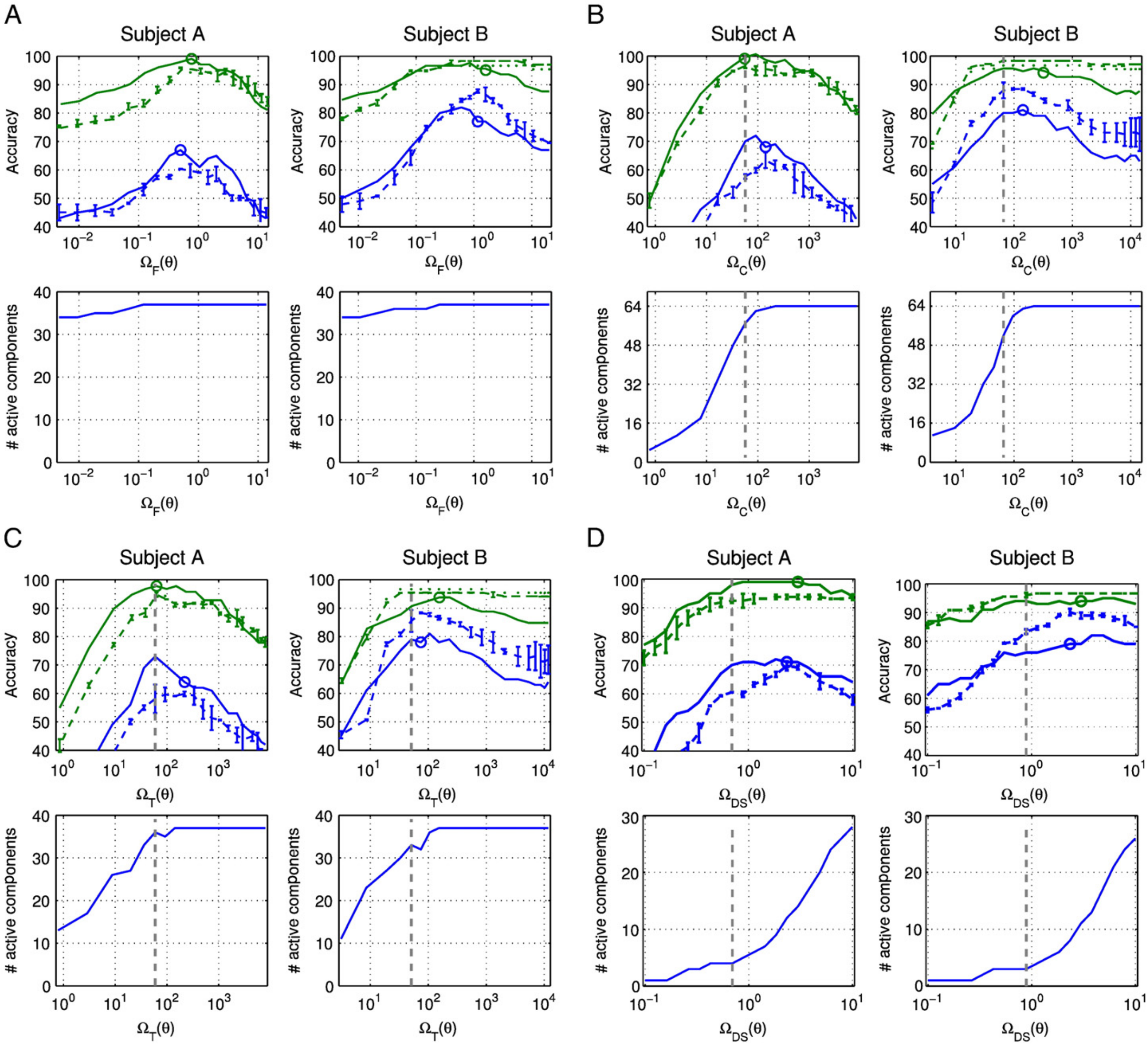


Fig. 4. Classification accuracy and the number of active components obtained with different regularizers. Top part of each figure: the blue and green lines correspond to 5 repetitions ($M = 5$) and 15 repetitions ($M = 15$), respectively. The dashed lines with error bars show the cross-validation performance. The solid lines show the test performance. Bottom part of each figure: number of active components. The vertical dashed lines show the regularization constant chosen for the visualization in Figs. 6 and 7. (A) Frobenius norm regularization. The bottom part shows the number of non-zero singular values of the weight matrix. (B) Channel selection regularizer. The bottom part shows the number of channels with non-zero norms. (C) Temporal-basis selection regularizer. The bottom part shows the number of temporal bases with non-zero norms. (D) Dual spectral norm regularizer. The bottom part shows the number of non-zero singular values of the weight matrix.

Table 1
Classification accuracy in % obtained with four regularizers namely channel selection regularizer (CSR, Eq. (9)), temporal-basis selection regularizer (TSR, Eq. (10)), and the dual spectral norm regularizer (DS, Eqs. (11)), compared against the winner of the competition (R&G).

Subject	Frobenius	CSR	TSR	DS	R&G
A ($M=5$)	67	*68*	64	**71**	*68*
A ($M=15$)	*98*	*98*	**99**	**99**	*98*
B ($M=5$)	77	**81**	78	*79*	76
B ($M=15$)	93	93	**95**	*94*	**95**
mean ($M=5$)	*72*	*74.5*	71	**75**	72
mean ($M=15$)	95.5	95.5	**97**	*96.5*	*96.5*

validation set. Thus we can choose the best model depending on the target information transfer rate.

In addition, the number of active components[7] is shown at the bottom of each plot. The plot is almost flat for the Frobenius norm regularizer, which employs no selection mechanism. The number of components falls sharply for the channel selection regularizer and the temporal-basis selection regularizer but it seems that the selection occurs at the cost of performance reduction. In contrast, the number of components (rank) can be greatly reduced with little cost until some point for the DS regularization.

Table 1 summarizes the test accuracy obtained at the selected regularization constant. The result from Rakotomamonjy and Guigue (2008) who won the competition is also shown. Bold and italic numbers are the best and the second best accuracy for each subject and number of repetitions. Note that the results of the winner are slightly different from the ones that are available from the official BCI competition website; this is because we recomputed the results using the scripts provided by the winner[8].

Discussion

The performance of the proposed regularizers are comparable to that of the winner of the competition except for the Frobenius norm regularizer. In addition, in Fig. 5, the performance of the proposed model with the DS regularizer is compared against that of the winner (R&G) at various number of repetitions M. Although it is difficult to draw any conclusion from such a small test, for both subjects the proposed method is competitive to Rakotomamonjy and Guigue (2008) for most M especially at small number of repetitions. This can be partly explained by the fact that we choose the regularization constant C using the character recognition accuracy on the validation set for each M whereas an auxiliary measure for model selection based on the binary classification model is used in Rakotomamonjy and Guigue (2008), which does not take the number of repetitions into account.

We should note that although the normalization by $\sum^{t\,-1/4}$ and $\sum^{s\,-1/4}$ from both sides seems sensible from dimensionality consideration and its separability, in principle this should be also considered as a hyper-parameter that needs to be selected based on the training data. In fact, we obtained a lower performance by normalizing each element of the input matrices to unit variance, which is actually not preferable because it cannot be separated into the spatial part and the temporal part as the above normalization (results not shown).

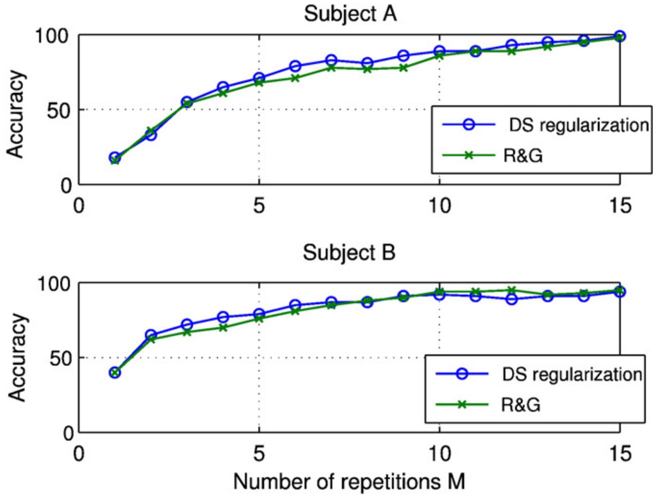


Fig. 5. The accuracy of the proposed dual spectral regularization compared to Rakotomamonjy and Guigue (2008) at variety of number of repetitions.

Different types of sparsity induced by the regularizers are useful in understanding how classifiers work and also understanding inter-subject variability. The weight matrices obtained with the three sparsity inducing regularizer are visualized in Figs. 6 and 7 for subjects A and B, respectively. The first two plots (Figs. 6A, B) and Figs. 7A, B) show the weight matrix including the preprocessing matrices $\boldsymbol{P}^t$ and $\boldsymbol{P}^s$ defined as $\boldsymbol{W}^{\mathrm{raw}}=\boldsymbol{P}^t\boldsymbol{W}\,\boldsymbol{P}^s$ which has again T rows and C columns. The upper plot shows the temporal slice of $\boldsymbol{W}^{\mathrm{raw}}$ at the time-point shown above. The temporal slice $\boldsymbol{W}^{\mathrm{raw}}(t,:)$ is color coded as blue–green–red from negative to positive and since it is a C dimensional vector, it is mapped on a scalp viewed from above (nose pointing upwards). The lower plot shows the spatial slice $\boldsymbol{W}^{\mathrm{raw}}(:,c)$ for every electrode along time. The last plots (Figs. 6C and 7C) show the leading singular vectors of the weight matrices obtained with the DS regularization. We first perform singular-value decomposition of the low-rank weight matrix as $\boldsymbol{W}=\boldsymbol{U}\mathrm{diag}(\sigma_1,\dots,\sigma_{\mathrm{r}})\boldsymbol{V}^\top$ where $\boldsymbol{U}$ is a $T\times r$ matrix and $\boldsymbol{V}$ is a $C\times r$ matrix. Then we define a *spatial filter* $\boldsymbol{w}_j^s$ and a spatial pattern $\boldsymbol{a}_j^s$ as follows:

$$\boldsymbol{w}_j^s=\boldsymbol{P}^s\boldsymbol{V}(:,j),\quad \boldsymbol{a}_j^s=\left(\boldsymbol{P}^s\right)^{-1}\boldsymbol{V}(:,j)\quad (j=1,\dots,r).$$

A spatial filter is a coefficient vector applied to the raw (low-pass filtered) signal as part of the classifier. On the other hand, the spatial pattern of a given spatial filter is the EEG activity that is optimally captured by the corresponding spatial filter. That is $\boldsymbol{a}_j^s$ is orthogonal to every $\boldsymbol{w}_{j'}^s$ for $j'\neq j$. Similarly a temporal filter $\boldsymbol{w}_j^t$ and a temporal pattern $\boldsymbol{a}_j^t$ is defined from $\boldsymbol{U}$ and $\boldsymbol{P}^t$. Now we have a decomposition of the raw coefficient matrix $\boldsymbol{W}^{\mathrm{raw}}$ that includes the preprocessing and classifier coefficient as $\boldsymbol{W}^{\mathrm{raw}}=\sum_{j=1}^r\sigma_j\mathbf{w}_j^t\mathbf{w}_j^s$. Note that using the spatial/temporal filters we can decompose the first-order model (Eq. (15)) as follows:

$$f_\theta\left(\boldsymbol{X}_{(l)}\right)=\langle\boldsymbol{W},P^t\boldsymbol{X}_{(l)}^{\mathrm{LP}}P^s\rangle=\sum_{j=1}^r\sigma_j\boldsymbol{w}_j^{t^\top}\boldsymbol{X}_{(l)}^{\mathrm{LP}}\boldsymbol{w}_j^s.$$

Moreover, one can assume the following generative model for the band-pass filtered signal $\boldsymbol{X}_{(l)}^{\mathrm{LP}}$:

$$\boldsymbol{X}_{(l)}^{\mathrm{LP}}=\sum_{j=1}^r\eta_j^{(l)}\boldsymbol{a}_j^t\boldsymbol{a}_j^{s\top}+\boldsymbol{N}^{(l)}.$$

where $\boldsymbol{a}_j^s$ and $\boldsymbol{a}_j^t$ functions as a fixed spatial/temporal-basis function, $\eta_j^{(l)}$ is a task-related component, and $\boldsymbol{N}^{(l)}$ is the non-task-related component (i.e., $\boldsymbol{w}_j^{t^\top}\boldsymbol{N}^{(l)}\boldsymbol{w}_j^s=0(\forall j)$). Thus the spatial/temporal

[7] The number of active components is defined as follows: given the weight matrix $\boldsymbol{W}$ let $s_1,\dots,s_r$ be the component norms (column-wise norms for the channel selection regularizer, row-wise norms for the temporal-basis selection regularizer and the singular values for the DS regularizer.) #active components $=|s_j:s_j>0.01\max_j(s_j)|$.

[8] Scripts are available from http://asi.insa-rouen.fr/enseignants/~arakotom/code/bciindex.html.

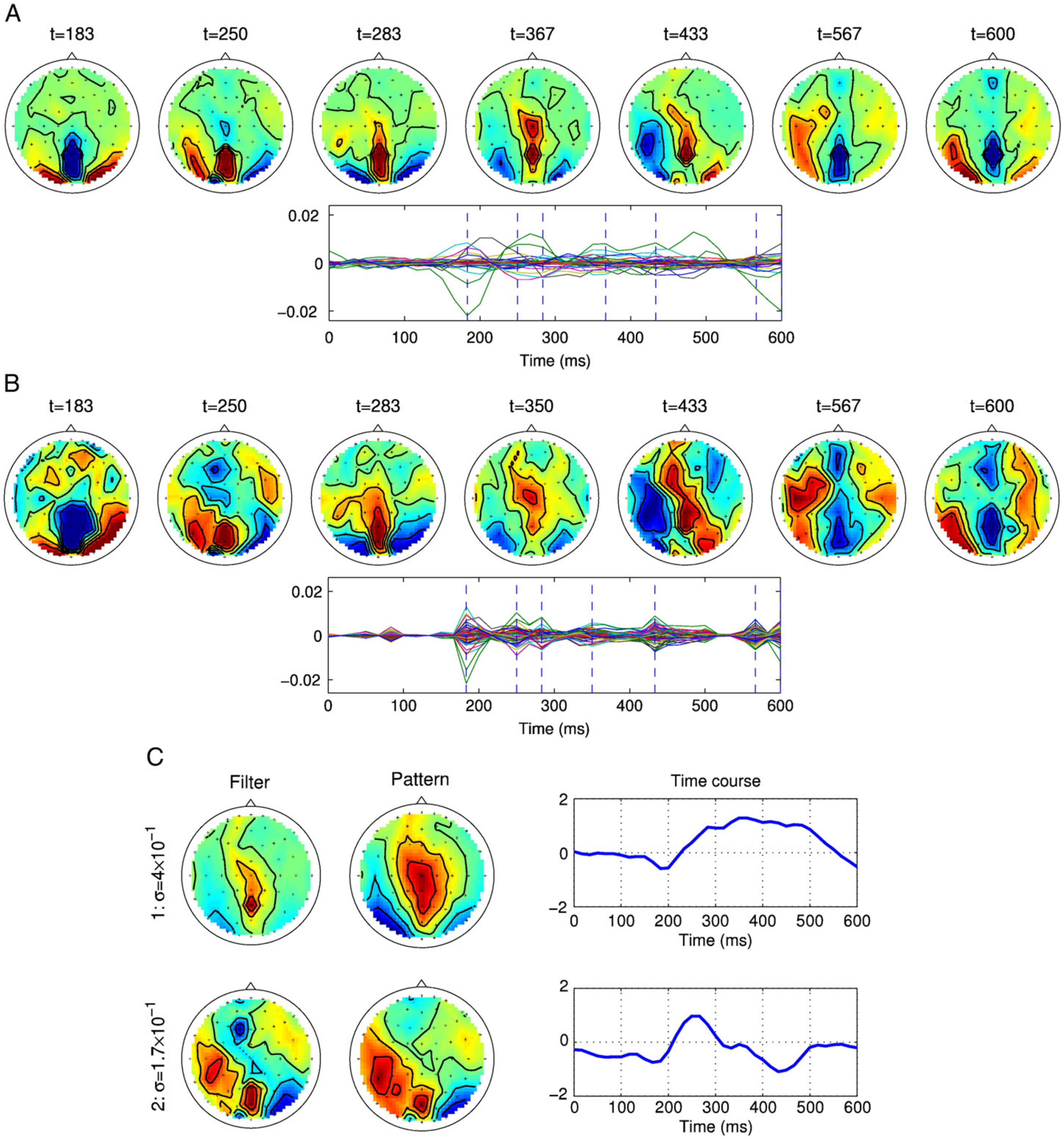


Fig. 6. Spatial/temporal profile of subject A. (A) Channel selection regularizer. $\Omega_C(\theta) = 57$. (B) Temporal-basis selection regularizer. $\Omega_T(\theta) = 59$. (C) Dual spectral regularizer. $\Omega_*(\theta) = 0.70$.

patterns $\boldsymbol{a}_j$ and filters $\boldsymbol{w}_j$ provide forward and backward view on the generation of task-related EEG activities, respectively (Parra et al., 2005; Blankertz et al., 2008). Finally, the spatial filter, spatial pattern, and the temporal pattern are plotted from left to right for each left/right singular vector pairs of the leading singular values from top to bottom. The spatial filters/patterns are plotted in the same way as above. The temporal patterns, which are T dimensional vectors, are plotted along time. The singular value is also shown vertically at the left end of each row.

The channel selection regularizer (see Figs. 6A and 7A) is good at spatially localizing the discriminative information. For both subjects A and B we can see occipital focus in the early phase and more parietal-central focus in the later phase.

On the other hand, the temporal-basis selection regularizer localizes the discriminative information in the temporal domain. For subject A (Fig. 6B bottom), there is a prominent negative peak at 183 ms and a broad positive component from 350 ms to 500 ms, which roughly agree with the early occipital component and the late

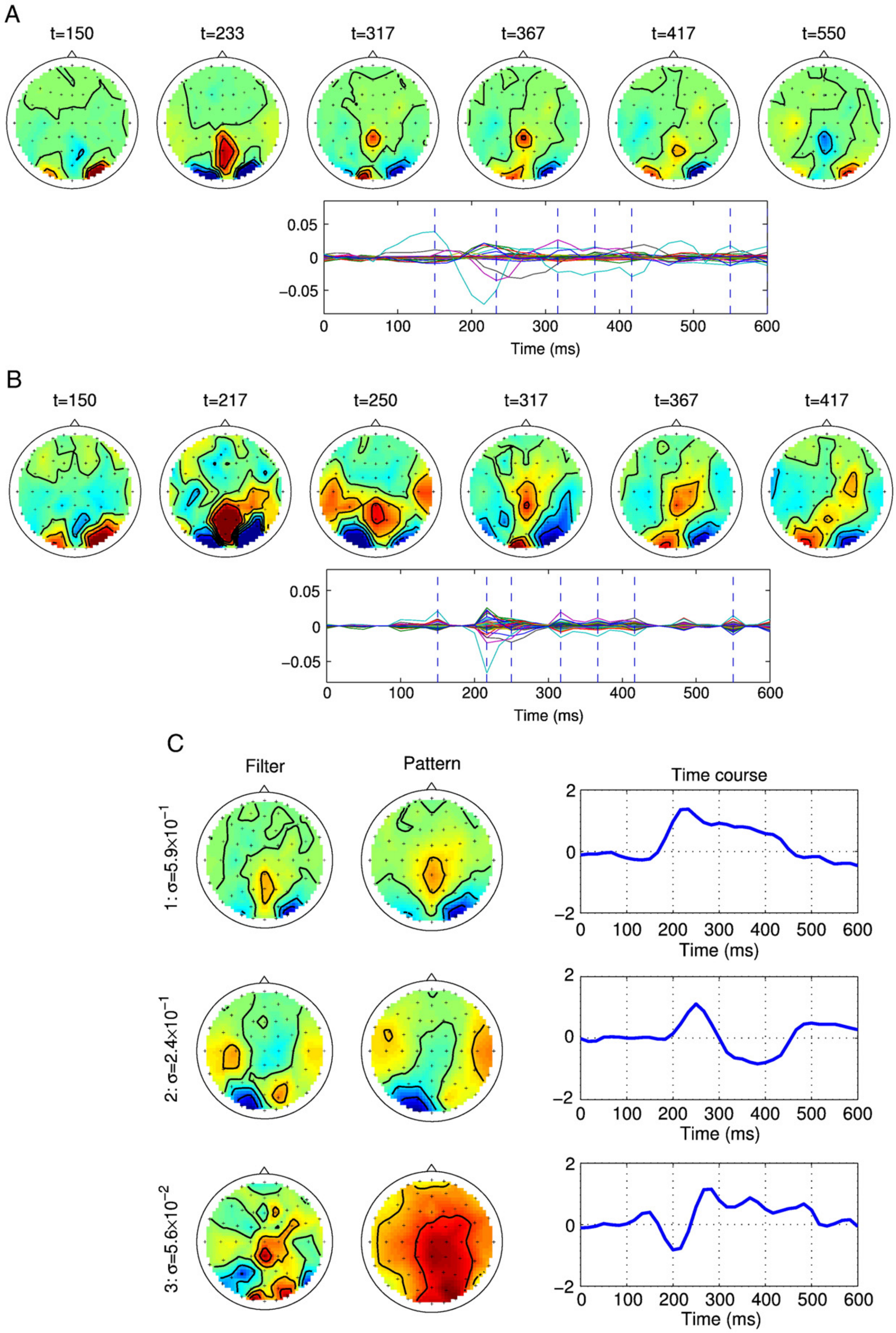


Fig. 7. Spatial/temporal profile of subject B. (A) Channel selection regularizer. $\Omega_C(\theta)=66$. (B) Temporal-basis selection regularizer. $\Omega_T(\theta)=51$. (C) Dual spectral regularizer. $\Omega_*(\theta)=0.89$.

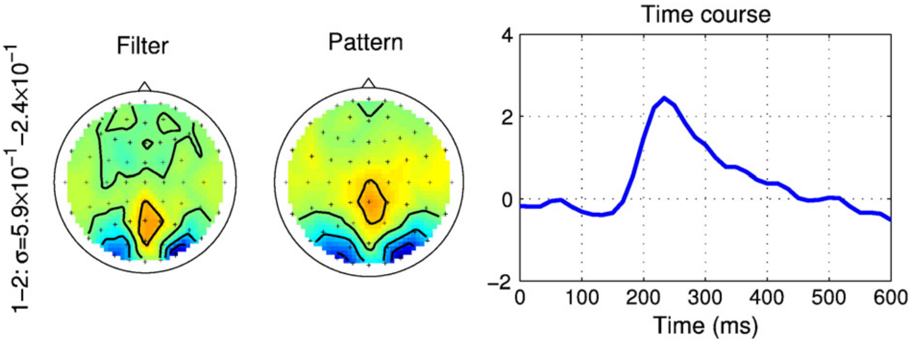


Fig. 8. Spatial/temporal profile of subject B with the dual spectral regularizer. The first two components in Fig. 7C are merged proportional to their singular values.

parietal–central component mentioned above. For subject B (Fig. 7B bottom), the strong negative peak sits around 217 ms and, similarly to subject A, a sustained discriminability can be observed from 300 ms to 450 ms. Note however that for both subjects the spatial localization cannot be seen as clearly as in the channel selection regularizer.

The DS regularizer provides a small number of pairs of spatial and temporal filters that show both spatial and temporal localization of the discriminative information in a compact manner. The two plots (Figs. 6C and 7C) confirm our earlier observation that there are two major discriminative components: the early occipital component (the second row in Fig. 6C and the first two rows in Fig. 7C) and the late central component (the first row in Fig. 6C and the third row in Fig. 7C). From the magnitude of the singular values, it seems that the classifier relies more on the late sustained component for subject A whereas for subject B it relies more on the early component around 217 ms. Interestingly the early component was split into the first two components for subject B. The spatial focus in the occipital area and the temporal focus around 217 ms can be seen clearer in Fig. 8 where we plotted the first two components mixed proportional to their singular values.

Note that our findings are consistent with the study by Krusienski et al. (2008) in which they reported that the combination of central and posterior electrode provided the best performance in average over seven subjects.

Finally we note that the application of the proposed model in an online BCI is efficient because of the linearity of the detector function Eq. (15); low-pass filtering can be applied to the one dimensional signal obtained by applying the classifier in an online manner.

Results: self-paced finger tapping dataset

The result of the proposed framework applied to the self-paced finger tapping dataset (Self-paced finger tapping problem section) is given in Performance section and a discussion including the visualization of spatial/temporal filter pairs obtained from the DS regularization is given in Discussion section.

Performance

Figs. 9A–C show the classification accuracy of the proposed three detector models with the Frobenius and DS norm regularization. The

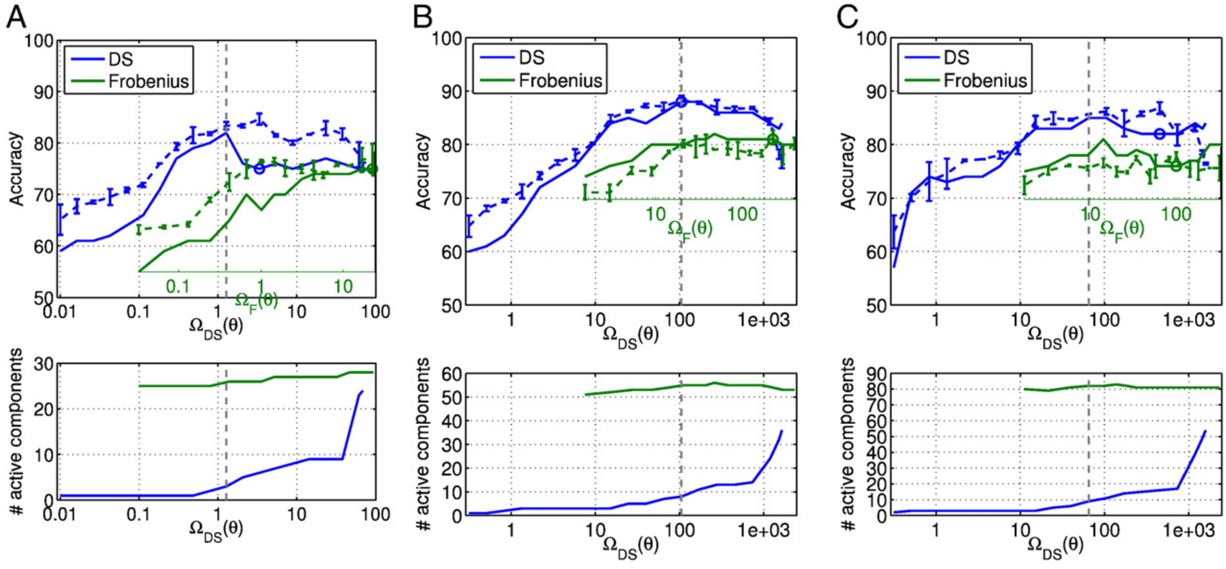


Fig. 9. Classification accuracies of the three proposed models with two different regularizers. Top plots: the accuracies obtained from the dual spectral regularizer (blue curves) and the Frobenius norm regularizer (green curves) are shown against the complexity of the resulting classifiers measured by the dual spectral norm. The solid curves show the test accuracy (in %). The dashed curves with error bars show the cross-validation accuracy. Bottom plots: the ranks of the weight matrices obtained from the two regularizers are shown against the dual spectral norm of the obtained classifiers. The complexity of the classifiers that are used in the visualization (see Figs. 10–12) are marked with vertical gray dashed lines. See Preprocessing and predictor model for the self-paced problem section for the definition of the three proposed models. (A) First-order model. (B) Wide-band second order model with a first-order term and a wide-band (7–30 Hz) second-order term. (C) Double-band second-order model with a first-order term, alpha (7–15 Hz) and beta (15–30 Hz) band second-order terms.

Table 2
Comparison of the complexity (in terms of the number of parameters) and the performance of three proposed models and two earlier studies.

	First-order model	Wide-band (7–30 Hz) second-order model	Double-band (7–15 Hz; 15–30 Hz) second-order model	Wang et al. (2004) (1st-order +wide-band (10–33 Hz) +19 selected channels)	SOBDA (1st-order +wide-band (10–33 Hz))
#parameters	1401 (433)	1807 (341)	2213 (559)	282	135
DS	75 (85)	88 (88)	82 (87)	84	87
Frobenius	75 (77)	81 (81)	76 (78)		

In the first row the number of parameters is shown (see main text for the derivation); the number of active parameters is also shown in parenthesis for the proposed models. The classification accuracy is shown in %. For the proposed models the accuracy obtained with two regularization strategies are shown. The cross-validation accuracy used for the selection of the regularization constant is shown inside parentheses.

2×10 fold cross-validation accuracy used for the selection of the regularization constant is also shown as a dashed curve with error bars for each detector model and regularizer. The accuracies obtained at the selected regularization constants are marked with circles. The accuracy is plotted at the complexity measured by the DS norm for the classifiers obtained with the two regularizers. This is done in order to compare the performance of the two classifiers at the same complexity. The original complexity measure of the Frobenius norm regularized classifiers is also shown as second axis in each figure. Note that this is only possible when the DS norm of the Frobenius regularized model grows monotonically with the regularization constant.

The performance obtained with the two regularizers is summarized in Table 2. The performance of the winner of the competition (Wang et al., 2004) and a recently proposed bilinear discriminant analysis (Christoforou et al., 2008) is also shown. The best accuracy 88% is obtained with the wide-band second-order model with the DS regularization which also achieved the highest with respect to the cross-validation accuracy.

Discussion

In Fig. 9A we can see that the performance of the DS norm regularizer is higher than the Frobenius norm regularizer over the whole range of complexity. The performance of the two regularizers converges to the same value when the highest complexity is allowed. Indeed the training loss $L_n(\theta)$ is less than 10^{-10} at the highest complexity. Thus the difference in the regularizer plays almost no role. Similar trends can also be seen in Figs. 9B and 9C.

Incorporating the wide-band (7–30 Hz) second-order term significantly improves the performance (see Fig. 9B) as reported earlier in (Dornhege et al., 2004; Wang et al., 2004; Christoforou et al., 2008). However the performance is reduced if we allow further flexibility by dealing with the alpha-band (7–15 Hz) and beta-band (15–30 Hz) separately (see Fig. 9C). One possible explanation is overfitting. In addition, the cross-validation failed to predict the drop in the accuracy above $\Omega_{DS}(\theta)$>100. Strong correlation between the alpha and beta-band may also account for the poor performance; i.e., dealing with the two bands separately may not provide more information in comparison to the increased dimensionality.

In addition, the dimensionalities of the proposed detector models are compared to those of the two earlier studies in Table 2. The number of parameters is calculated as follows: for the first-order model it is 28(channels)×50(time-points) 1(bias)=1401; for the second-order model adding 406 (the degree of freedom of 28×28 symmetric matrix) it is 1807; for the double-band second-order model it is 2213 with an additional 406. For Wang et al. (2004), since they used a rank=2 first-order term with 4 time-points $((28+4)\cdot 2=64)$, a rank=6 wide-band second-order term with 4 time-points $((28+4)\cdot 6=192)$, hand-chosen 19 electrodes with a fixed temporal filter (19), and 3 classifier weights and 4 bias terms, it is 282. For SOBDA (Christoforou et al., 2008), since they used a rank=1 first-order term with 50 time-points and a rank=2 second-order term with no temporal information, and a single bias term, it is $28+50+28\cdot 2+1=135$. Although, the raw dimensionality of the proposed models are higher than those of the two earlier studies, the numbers of active parameters[9] at the selected regularization constant (shown inside the parentheses) are of the same order as the earlier studies. Importantly for the proposed models, the rank is *automatically tuned* through the regularization. Similar models which in contrast had to *fix the rank* a priori have been employed in earlier studies (see Wang et al. (2004); Christoforou et al. (2008) and Discussion on earlier discriminative approaches section).

The spatial/temporal profiles of the three proposed models are visualized in Figs. 10–12. See Discussion section in Results: P300 speller BCI section for the definition of spatial/temporal filters and patterns. The top two components obtained from the first-order term seems to be preserved from the simple first-order model to the most complex double-band second-order model. The first component clearly focuses on the lateralized readiness potential. This can be seen from the bipolar structure of the spatial pattern (two peaks with opposite signs on left and right motor cortices) as well as the temporal profile that drops monotonically towards the key press. The meaning of the second component is not obvious. From the downward trend along time, we conjecture that it also detects some part of the readiness potential that is not captured by the first component though the contribution of this component to the classifier is one order smaller than the first component.

In Fig. 11, we can find typical spatial patterns for event-related (de)-synchronization (ERD/ERS (Pfurtscheller and da Silva, 1999). The first second-order component (third row) captures ERD in the right-hand area (which can be seen from the negative sign of the eigenvalue[10] shown next to the filter) and the second second-order component (forth row) captures ERD in the left-hand area.

Interestingly this discriminability is mainly due to the beta-band. In Fig. 12, we can find spatial filter/pattern pairs that look similar to the ERD/ERS components in Fig. 11 in the bottom two rows (components obtained from the beta-band) though the order is reversed. Then what are the two alpha components (rows 3–4) doing? From the spatial filters they might seem to be focusing on the right motor cortex which delivers the ERD in the left-hand trials. However the negative signs of the eigenvalues and the spatial patterns suggest that these components detect ERD in the right-hand trials. We confirm this in Fig. 13 where we plot the log powers of the spatially filtered beta-band features against those of alpha-band features. Indeed both alpha-band features show lower magnitude in the right-hand trials than in the left-hand trials.

[9] The number of active parameters is calculated from the rank of the weight matrix, i.e., rank = r matrix of size R × C has $(R + C)r - r^2$ active parameters.

[10] Since the block weight matrix associate to the second-order component (see Eq. (7)) is symmetric, we show the eigenvalues instead of singular values for the second-order components.

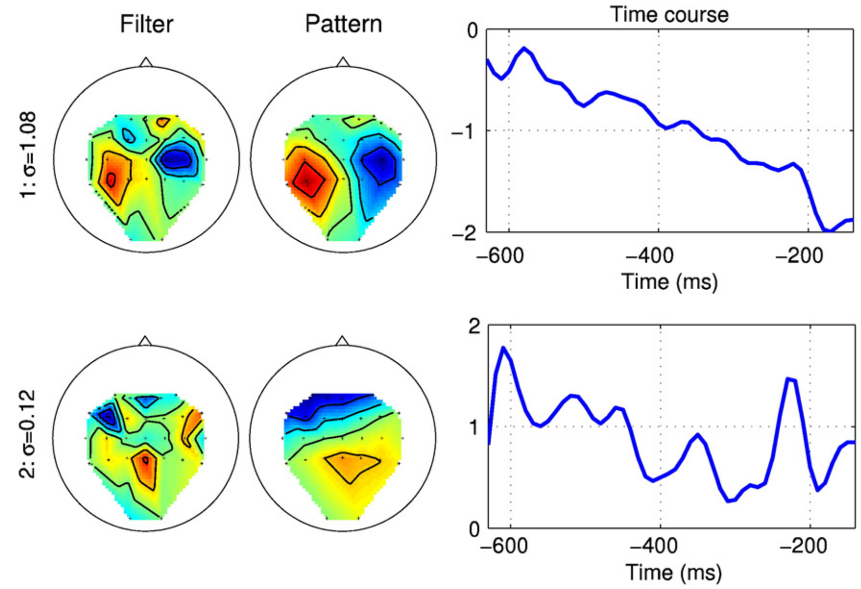


Fig. 10. Spatial/temporal profile of the proposed first-order model at $\Omega_{DS}(\theta) = 1.27$ (see Fig. 9A). The spatial filter, spatial pattern and temporal pattern that correspond to the first two singular values of the weight matrix are shown.

The second-order models can also be applied efficiently online in the case of DS regularization because the coefficient matrix is typically low rank. We can decompose the second-order weight matrix $\boldsymbol{W}^{(2)}$ as $\boldsymbol{W}^{(2)} = \sum_{j=1}^{r} \lambda_j \boldsymbol{w}_j^{(2)} \boldsymbol{w}_j^{(2)\top}$, where $\boldsymbol{w}_j^{(2)} = \Sigma^{s-1/2} \boldsymbol{V}(:, j)$ is a spatial filter as in Discussion section in Results: P300 speller BCI section, and compute the dot product as $\langle \boldsymbol{W}^{(2)}, \mathrm{cov}(\boldsymbol{X}^{\mathrm{BP}}) \rangle = \sum_{j=1}^{r}$

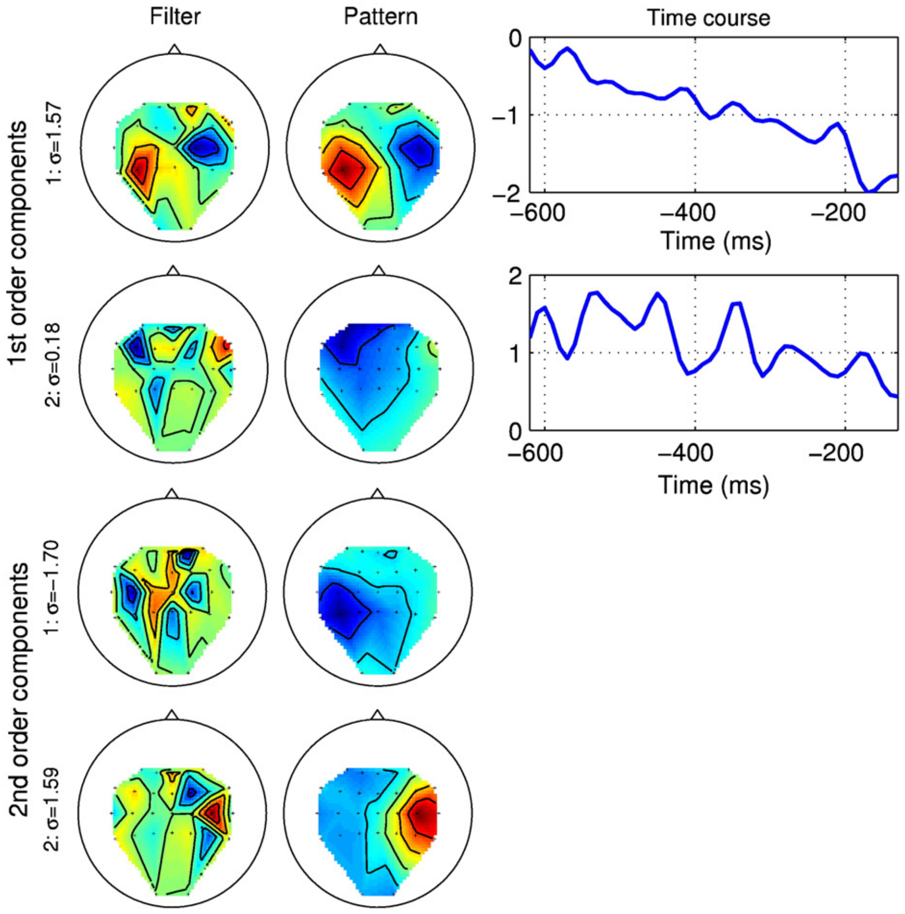


Fig. 11. Spatial/temporal profile of the proposed wide-band second-order model at $\Omega_{DS}(\theta) = 106$ (see Fig. 9B). The first two rows show the spatial filter, spatial pattern, and temporal pattern of the first-order components. The last two rows show the spatial filter and pattern of the second-order components (7–30 Hz). Note that there is no temporal structure for the second-order components.

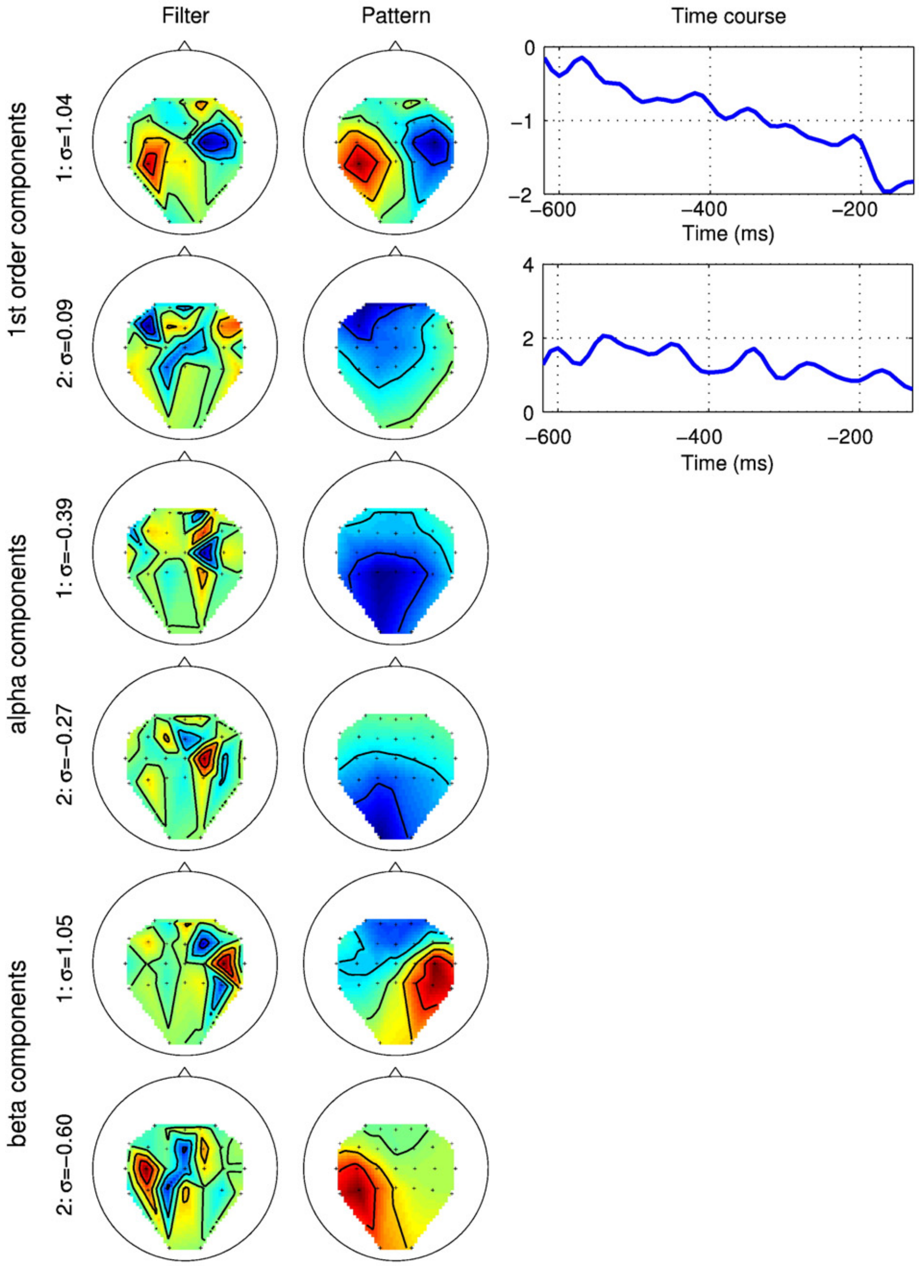


Fig. 12. Spatial/temporal profile of the proposed double-band second-order model at $\Omega_{DS}(\theta) = 65.0$ (see Fig. 9C). The first two rows show the spatial filter, spatial pattern, and temporal pattern of the first-order components. The last four rows show the spatial filter and pattern of the alpha-band (7–15 Hz) second-order components (rows 3–4) and the beta-band (15–30 Hz) second-order components (rows 5–6). Note that there is no temporal structure for the second-order terms.

$\lambda_j \mathrm{var}(h^{\mathrm{BP}}(\boldsymbol{w}_j^\top \boldsymbol{X}^{\mathrm{raw}}))$, where h^{BP} denotes the band-pass filtering and var denotes the short-time variance estimate.

Discussion on earlier discriminative approaches

In this section we review earlier studies on discriminative modeling. Major difference arises in the parameterization of the detector function $f_\theta(\boldsymbol{X})$.

Second order feature based BCI

One of the most successful approach in motor-imagery based BCI is common spatial pattern (CSP) (see Fukunaga (1990); Koles (1991); Ramoser et al. (2000) and also Dornhege et al. (2004); Lemm et al. (2005); Dornhege et al. (2006, 2007); Blankertz et al. (2008) for various extensions). A commonly used form of CSP based detector model can be written as follows (Tomioka et al., 2006a; Blankertz et al., 2008):

$$f_\theta(\boldsymbol{X}) = \sum_{j=1}^{J} \beta_j \log\left(\boldsymbol{w}_j^\top \boldsymbol{X}^\top \boldsymbol{B}_j \boldsymbol{B}_j^\top \boldsymbol{X} \boldsymbol{w}_j\right) + \beta_0, \qquad (17)$$

where $\boldsymbol{X} \in \mathbb{R}^{T \times C}$ is a short segment of multi-channel EEG measurement with C channels and T sampled time-points; $\boldsymbol{B}_j \in \mathbb{R}^{T \times T}$ are temporal filters, $\boldsymbol{w}_j \in \mathbb{R}^C$ are spatial filters, $\{\beta_j\}_{j=1}^J$ are weighting coefficients of the J features, and β_0 is a bias term. CSP is a dimensionality reduction method based on a generalized eigenvalue problem (Fukunaga, 1990; Koles, 1991).

In the conventional CSP based approach, thus the classifier is trained in three steps. First, the temporal filter coefficients ***Bj*** is chosen a priori or based on some heuristics (Blankertz et al., 2008). Second, the spatial filter is obtained from solving the generalized eigenvalue

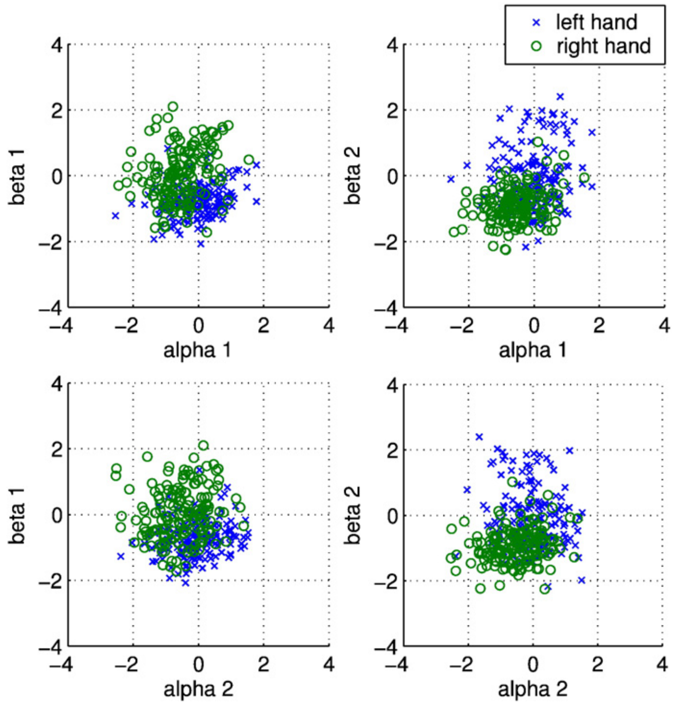


Fig. 13. Comparison of the four features obtained from alpha and beta-bands. The log powers of the spatially filtered training signals are plotted for the last four spatial filters shown in Fig. 12. Two filters are obtained from the alpha-band and are shown along the horizontal axes. Two filters are obtained from the beta-band and are shown along the vertical axes. Training examples that correspond to left and right hand trials are shown as blue crosses and green circles, respectively.

problem. Third, the classifier weights $\{\beta_j\}_{j=1}^{J}$ are obtained from Fisher's linear discriminant analysis.

Several studies used this detector model and related models. Farquhar et al. (2006) proposed to learn all the above coefficients[11] *jointly* with the hinge loss and the Frobenius norm regularization for coefficients $\{\boldsymbol{w}_j\}_{j=1}^{J}$, $\{\boldsymbol{B}_j\}_{j=1}^{J}$, and $\{\beta_j\}_{j=1}^{J}$. In Tomioka et al. (2007), the logarithm in Eq. (17) is omitted and two spatial filters $\boldsymbol{w}_j$ $(j=1, 2)$ is optimized under the logistic loss (with the temporal filter coefficients $\boldsymbol{B}_j$ being kept constant). The resulting model is similar to SOBDA proposed in Christoforou et al. (2008) (see next subsection). However these approaches lead to a *non-convex* optimization problem which may suffer from multiple local minima and poor convergence property.

In Tomioka and Aihara (2007), the DS regularization is introduced and the following model is assumed:

$$f_\theta(\boldsymbol{X}) = \langle \boldsymbol{W}, \boldsymbol{X}^\top \boldsymbol{X} \rangle + \beta_0, \tag{18}$$

where Eq. (18) is obtained from Eq. (17) by omitting the logarithm, and the temporal filter coefficient Bj (assumed to be constant), and denoting $\boldsymbol{W} = \sum_{j=1}^{J} \beta_j \boldsymbol{w}_j \boldsymbol{w}_j^\top$. It was demonstrated that using the DS regularization, good classification performance is obtained with only a few spatial filters $\boldsymbol{w}_j$. Interestingly, the DS regularization typically chose rank = 4 or 5 which roughly corresponds to a common practice in the CSP based approach (Ramoser et al., 2000).

First/second-order feature based BCI

Dyrholm et al. proposed the following *bilinear* detector model which they call bilinear discriminant component analysis (BDA) model (Dyrholm and Parra, 2006; Dyrholm et al., 2007):

$$f_\theta(\boldsymbol{X}) = \sum_{j=1}^{J} \boldsymbol{u}_j^\top \boldsymbol{X} \upsilon_j + \beta_0 = \mathrm{Tr}\left(\boldsymbol{U}^\top \boldsymbol{X} \boldsymbol{V}\right) + \beta_0, \tag{19}$$

where $\boldsymbol{X} \in \mathbb{R}^{C \times T}$ is a short segment of multi-channel EEG measurement with C channels and T sampled time-points; $\theta = (\{\boldsymbol{u}_j\}_{j=1}^{J}, \{\boldsymbol{v}_j\}_{j=1}^{J}, \beta_0)$ where $\boldsymbol{U} \in \mathbb{R}^{C \times J}$ and $\boldsymbol{V} \in \mathbb{R}^{T \times J}$ are temporal and spatial filter coefficients and β_0 is a bias term; $\boldsymbol{u}_j$ and $\boldsymbol{v}_j$ are the j-th row of $\boldsymbol{U}$ and $\boldsymbol{V}$, respectively. The number of spatial-temporal filter pairs J is usually chosen much smaller than C and T.

As the regularizer, the authors used the Frobenius norm on the coefficients $\{\boldsymbol{u}_j\}_{\mathrm{j}=1}^{J}$ and $\{\boldsymbol{v}_j\}_{j=1}^{J}$ as follows:

$$\Omega_{\mathrm{BDA}}(\theta) = \frac{1}{2}\left(\|\boldsymbol{U}\|_F^2 + \|\boldsymbol{V}\|_F^2\right),$$

where we omit the smoothing kernels used in Dyrholm et al. (2007) because they can be applied to the signal as $\boldsymbol{X} = \boldsymbol{K}^{t1/2}\ \boldsymbol{X}^{\mathrm{raw}}\ \boldsymbol{K}^{s1/2}$ where $\boldsymbol{K}^t$ and $\boldsymbol{K}^s$ are the smoothing kernels for $\boldsymbol{U}$ and $\boldsymbol{V}$, respectively. Note that in contrast to Dyrholm et al., our preprocessing matrices $\boldsymbol{P}^s = \sum^{s-1/4}$ and $\boldsymbol{P}^t = \sum^{t-1/4}$ (see Signal acquisition and preprocessing section) can be interpreted as *inverse smoothing* of the spatial/temporal filters if we assume that the *input signal is smooth*. In fact the spatial filters that we obtain typically have Laplacian type shapes (see e.g. Figs. 6–8). However note that it only (approximately) normalizes the correlation spectrum and it does not emphasize any frequency component as the true Laplacian operator. We believe that this inverse smoothing of the coefficients is useful in optimally detecting a smooth signal such as P300 evoked response, provided that its correlation structure is well captured in the covariance matrices $\sum^s$ and $\sum^t$. Note that the above mentioned inverse smoothing is analogous to the $\sum^{-1}$ term in the well known linear discriminant analysis (see e.g., Hastie et al. (2001)); linear discriminant coefficient vector is given as $\boldsymbol{w} = \sum^{-1}(\boldsymbol{\mu}_+ - \boldsymbol{\mu}_-)$ where the positive (negative) samples follow normal distributions with mean vector $\boldsymbol{\mu}_+$ ($\boldsymbol{\mu}_-$) and a covariance matrix $\sum$; if we also consider $\boldsymbol{\mu}_+$ and $\boldsymbol{\mu}_-$ as random variables that have the covariance $\sum$, then $\boldsymbol{w}$ has the covariance $\sum^{-1}$. Note also that the preprocessing matrices are calculated from the whole collection of epochs without any class information; in fact empirically the estimates are quite stable.

A remarkable fact about the above regularizer is that when J is sufficiently large the sum of squared Frobenius norms for $\boldsymbol{u}_j$ and $\boldsymbol{v}_j$ is equivalent to the DS norm of $\boldsymbol{W}$ i.e.,

$$\|\boldsymbol{W}\|_* = \frac{1}{2} \min_{\boldsymbol{W} = \boldsymbol{U}\boldsymbol{V}^\top} \left(\|\boldsymbol{U}\|_F^2 + \|\boldsymbol{V}\|_F^2\right) \tag{20}$$

where $\|\cdot\|_*$ and $\|\cdot\|_F$ are the DS norm and the Frobenius norm, respectively (see Srebro et al. (2005)). Thus BDA can be considered as a fixed-rank approximation of the proposed first-order model with the DS regularization (see also Recht et al. (2007)). Note however that typically BDA is used with extremely small J (Dyrholm et al., 2007; Christoforou et al., 2008) in which case the solutions will not coincide.

BDA was applied to the self-paced finger tapping dataset from BCI competition 2003 and a rapid serial visual presentation experiment (see Dyrholm et al. (2007); Parra et al. (2008)).

Christoforou et al. (2008) extended the first-order BDA (Eq. (19)) and proposed the following second-order BDA (SOBDA) model:

$$f_\theta(\boldsymbol{X}) = \mathrm{Tr}\left(\boldsymbol{U}^\top \boldsymbol{X} \boldsymbol{V}\right) + \sum_{k=1}^{K} \beta_k \left(\boldsymbol{w}_k \boldsymbol{X}^\top \boldsymbol{B}\boldsymbol{B}^\top \boldsymbol{X} \boldsymbol{w}_k\right) + \beta_0, \tag{21}$$

[11] In addition they proposed to jointly learn the temporal windowing function which is omitted here for simplicity.

where $\boldsymbol{U} \in \mathbb{R}^{C \times J}$ and $\boldsymbol{V} \in \mathbb{R}^{T \times J}$ are the first-order temporal and spatial filter coefficients as in Eq. (19) and $\boldsymbol{w} \in \mathbb{R}^{C}$ and $\boldsymbol{B} \in \mathbb{R}^{T \times T}$ are the second-order spatial and temporal filter coefficients. They directly optimized all the coefficients $\boldsymbol{U}$, $\boldsymbol{V}$, $\boldsymbol{w}_k$ and $\boldsymbol{B}$ with the logistic loss function and squared Frobenius norm penalty on the coefficients. We can see that when the temporal filter matrix $\boldsymbol{B}$ is kept constant, Eq. (21) can be written in the form of Eq. (6) by using the block diagonal concatenation (Eq. (7)).

Conclusion

In this article we have proposed a novel unified framework for signal analysis in EEG-based BCI. The proposed framework focuses on probabilistic predictors from which the decoding and learning algorithms are naturally deduced. The proposed framework includes conventional binary single-trial EEG classification as a special case but it is oriented to the final goal in BCI i.e., to predict the intention of a user in contrast to the training of a binary classifier as an intermediate step. This is very much in the spirit of Vapnik (1998): solving the problem directly instead of an indirect multi-step procedure. Moreover, the issues of feature learning, feature selection, and feature combination are addressed through regularization. This allows us to perform feature learning *jointly* with the training of the predictor model in a convex optimization framework. Note that although the proposed training procedure (Discriminative learning section) might seem exotic to some EEG practitioners, the resulting detector function is linear and the decoding procedures (see e.g., Eq. (16)) have the intuitive forms as in the previous studies (Farwell and Donchin, 1988; Krusienski et al., 2008).

In the P300 speller problem we have demonstrated how the learning algorithm derived from a natural predictor model can be different from the conventionally used binary classification approach. In fact, we have shown that the epoch-wise normalization imposed by the conventional approach may make it difficult to find a simple detector function. Furthermore different regularizers have revealed different aspects of the localization of the discriminative information. The spatial localization was investigated through the channel selection regularizer. Although the number of electrodes did not significantly reduce without compromising the performance, the plots have shown strong focus on occipital to central area. The low performance of the strongly regularized models may be attributed to volume conduction effects. Even if the source activities are spatially localized, volume conduction spreads them over a wide area, making it difficult to recover the activity from a small number of electrodes. The temporal localization was similarly investigated through the temporal-basis selection regularizer. Interestingly the temporal profiles have shown stronger inter-subject variability than the spatial profile. The dual spectral regularization has revealed both spatial and temporal profiles in a compact manner. All three regularizers performed comparable to the winner of the BCI competition while the dual spectral regularizer being competitive. However from the point of view of understanding the classifier, the three regularizers provided complimentary views that made it possible to find a consistent neurophysiological interpretation for each subject. The use of, say, the channel selection regularizer alone would not have allowed us to gain such insights. It is also important to mention that the complimentary views were particularly useful in deciding the complexity in plot Figs. 6 and 7. Strongly regularized predictors tend to be over-simplified and the plots do not account for the success at the more complex predictors selected by the cross-validation. On the other hand, the predictor at the complexity selected by the cross-validation did not always provide the best intuition.

In the self-paced finger tapping problem we have addressed the issue of how to learn, select, and combine features from different sources. We have employed the DS regularization on the augmented weight matrix. The input feature matrices were concatenated along the diagonal to form an augmented input feature matrix. The low-rank factorized predictor obtained from the DS regularization always outperformed the naive Frobenius norm regularization. Moreover, the proposed model has shown the highest performance in comparison to the winner of the BCI competition as well as the recently proposed second-order bilinear discriminant model.

Recent discriminative approaches are also discussed and the connection between our DS regularization and the sum-of-squared-Euclidean-norms regularization with fixed number of components used in (Dyrholm et al., 2007; Tomioka et al., 2007; Christoforou et al., 2008) is also pointed out. However often these models are used with only an extremely small number of components in which case the above equivalence does not hold.

The key idea in our approach is to focus on directly predicting the intention of a user. This enabled us to approach decoding and learning in a unified and systematic manner and to avoid intermediate steps. Note that this idea applies not only to other BCI paradigms including invasive BCIs but also to general discriminative neurophysiological paradigms even beyond EEG.

Furthermore we have shown that our discriminative approach can be considered as a novel visualization technique of the brain activity of a subject during tasks since it focuses on the basic components that are useful in predicting the intention of the subject. It reveals the most relevant piece of discriminant information in the subject's brain activity. Other types of decomposition problems such as multi-way tensor factorization (Harshman, 1970; Mørup et al., 2008) may also be tackled in a similar manner from the discriminative point of view considered in this work.

Acknowledgments

This work was supported in part by grants of Japan Society for the Promotion of Science (JSPS) through the Global COE program (Computationism as a Foundation for the Sciences), Bundesministerium für Bildung und Forschung (BMBF), FKZ 01IB001A (brainwork), 01GQ0850 (BFNT) and 01GQ0415 (BCCNB), Deutsche Forschungsgemeinschaft (DFG), MU 987/3-1 (Vital-BCI), the FP7-ICT Programme of the European Community, under the PASCAL2 Network of Excellence, ICT-216886 and project ICT-2008-224631 (TOBI). This publication only reflects the authors' views. We thank Benjamin Blankertz, Stefan Haufe, Kazuyuki Aihara, Motoaki Kawanabe, Masashi Sugiyama, David Wipf, Srikantan Nagarajan, and Hagai Attias for valuable discussions. Part of this work was done while the authors stayed at Fraunhofer FIRST.

Appendix A. Details of the algorithm

We used the projected gradient method described in Kim et al. (2006); Tomioka and Sugiyama (2008) for the optimization of Eq. (3) with the DS regularizer (Eq. (11)). The efficiency of the projected gradient method varies depending on the regularization constant C; it is faster for strong regularization (small C) and slower for weak regularization (large C.) For the P300 problem in Results: P300 speller BCI section, it takes about 5–6 min to obtain the solution for a single regularization constant around the best value $C \simeq 5$ on a workstation with two 3.3 GHz dual core Xeon processors and 8GB of RAM.

The channel selection regularizer and the temporal-basis selection regularizer (Eqs. (9) and (10)) are rewritten into the following linear penalty formulation:

$$\underset{\theta}{\text{minimize}}\, L_n(\theta) + \lambda\Omega(\theta), \tag{22}$$

and the steepest descent subgradient method described in Andrew and Gao (2007) (see also Yu et al. (2008)) was used for the optimization with 20 log-linearly spaced candidates of λ from the interval [0.01, 100]. The above formulation is equivalent to Eq. (3) in the sense that for any regularization constant C if the solution of Eq. (3) is unique, then there is a $\lambda > 0$ in Eq. (22) that yields the same solution with $\Omega(\theta) = C$; conversely for any $\lambda > 0$, Eq. (3) with $C = \Omega(\theta^*)$ gives the same solution, where θ^* is the solution of Eq. (22) for the given λ.

For the Frobenius norm regularizer (Eq. (8)), we rewrite Eq. (3) into the following squared penalty formulation:

$$\underset{\theta}{\text{minimize}} \quad L_n(\theta) + \lambda \|\boldsymbol{W}\|_F^2,$$

and the limited memory BFGS method (Nocedal and Wright, 1999) was used. The above squared norm formulation is again equivalent to Eq. (3).

We have recently been developing a new optimization algorithm for all the sparse regularizers in the linear penalty formulation (Eq. (22)) (Tomioka and Sugiyama, in press). The code is also available from http://www.ibis.t.u-tokyo.ac.jp/ryotat/dal/. However the new algorithm is still an early release still to be improved; note that this novel optimizer was not used for the computation of the results in this paper.

References

Abernethy, J., Bach, F., Evgeniou, T., Vert, J.-P., September 2006. Low-rank matrix factorization with attributes. Tech. Rep. N-24/06/MM, Ecole des mines de Paris, France.

Amit, Y., Fink, M., Srebro, N., Ullman, S., 2007. Uncovering shared structures in multiclass classification. ICML '07: Proceedings of the 24th International Conference on Machine learning. ACM Press, New York, NY, USA, pp. 17–24.

Andrew, G., Gao, J., 2007. Scalable training of L1-regularized log-linear models. Proc. of the 24th International Conference on Machine learning. ACM, New York, NY, USA, pp. 33–40.

Argyriou, A., Evgeniou, T., Pontil, M., 2007. Multi-task feature learning. In: Schölkopf, B., Platt, J., Hoffman, T. (Eds.), Advances in Neural Information Processing Systems 19. MIT Press, Cambridge, MA, pp. 41–48.

Argyriou, A., Micchelli, C.A., Pontil, M., Ying, Y., 2008. A spectral regularization framework for multi-task structure learning. In: Platt, J., Koller, D., Singer, Y., Roweis, S. (Eds.), Advances in Neural Information Processing Systems 20. MIT Press, Cambridge, MA, pp. 25–32.

Birbaumer, N., Ghanayim, N., Hinterberger, T., Iversen, I., Kotchoubey, B., Kübler, A., Perelmouter, J., Taub, E., Flor, H., 1999. A spelling device for the paralysed. Nature 398, 297–298.

Bishop, C.M., 2007. Pattern Recognition and Machine Learning. Springer.

Blankertz, B., Curio, G., Müller, K.-R., 2002. Classifying single trial EEG: Towards brain computer interfacing. In: Diettrich, T.G., Becker, S., Ghahramani, Z. (Eds.), Advances in Neural Inf. Proc. Systems (NIPS 01), Vol. 14, pp. 157–164.

Blankertz, B., Müller, K.-R., Curio, G., Vaughan, T.M., Schalk, G., Wolpaw, J.R., Schlögl, A., Neuper, C., Pfurtscheller, G., Hinterberger, T., Schröder, M., Birbaumer, N., 2004. The BCI competition 2003: Progress and perspectives in detection and discrimination of EEG single trials. IEEE Trans. Biomed. Eng. 51 (6), 1044–1051.

Blankertz, B., Dornhege, G., Krauledat, M., Müller, K.R., Kunzmann, V., Losch, F., Curio, G., 2006a. The Berlin brain–computer interface: EEG-based communication without subject training. IEEE Trans. Neural Sys. Rehab. Eng. 14 (2), 147–152.

Blankertz, B., Müller, K.-R., Krusienski, D., Schalk, G., Wolpaw, J.R., Schlögl, A., Pfurtscheller, G., Millán, J. del R., Schröder, M., Birbaumer, N., 2006b. The BCI competition III: Validating alternative approaches to actual BCI problems. IEEE Trans. Neural Sys. Rehab. Eng. 14 (2), 153–159 see also the webpage: http://www.bbci.de/competition/iii/.

Blankertz, B., Dornhege, G., Krauledat, M., Müller, K.-R., Curio, G., 2007. The non-invasive Berlin brain-computer interface: fast acquisition of effective performance in untrained subjects. NeuroImage 37 (2), 539–550.

Blankertz, B., Tomioka, R., Lemm, S., Kawanabe, M., Müller, K.-R., 2008. Optimizing spatial filters for robust EEG single-trial analysis. IEEE Signal Proc. Magazine 25 (1), 41–56.

Boyd, S., Vandenberghe, L., 2004. Convex Optimization. Cambridge University Press.

Christoforou, C., Sajda, P., Parra, L.C., 2008. Second order bilinear discriminant analysis for single trial EEG analysis. In: Platt, J., Koller, D., Singer, Y., Roweis, S. (Eds.), Advances in Neural Information Processing Systems 20. MIT Press, Cambridge, MA, pp. 313–320.

Cotter, S.F., Rao, B.D., Engan, K., Kreutz-Delgado, K., 2005. Sparse solutions to linear inverse problems with multiple measurement vectors. IEEE Trans. Signal Process 53 (7), 2477–2488.

Curran, E.A., Stokes, M.J., 2003. Learning to control brain activity: a review of the production and control of EEG components for driving brain–computer interface (BCI) systems. Brain Cogn. 51, 326–336.

Dornhege, G., Blankertz, B., Curio, G., Müller, K.-R., Jun. 2004. Boosting bit rates in non-invasive EEG single-trial classifications by feature combination and multi-class paradigms. IEEE Trans. Biomed. Eng. 51 (6), 993–1002.

Dornhege, G., Blankertz, B., Krauledat, M., Losch, F., Curio, G., Müller, K.-R., 2006. Combined optimization of spatial and temporal filters for improving brain–computer interfacing. IEEE Trans. Biomed. Eng. 53 (11), 2274–2281.

Dornhege, G., Millán, J. del R., Hinterberger, T., McFarland, D., Müller, K.-R. (Eds.), 2007. Towards Brain-Computer Interfacing. MIT Press.

Dyrholm, M., Parra, L.C., 2006. Smooth bilinear classification of EEG. In: Proceedings of the IEEE 2006 International Conference of the Engineering in Medicine and Biology Society.

Dyrholm, M., Christoforou, C., Parra, L.C., 2007. Bilinear discriminant component analysis. J. Mach. Learn. Res. 8, 1097–1111.

Faraut, J., Koranyi, A., 1995. Analysis on Symmetric Cones. Oxford University Press.

Farquhar, J., Hill, J., Schölkopf, B., 2006. Learning optimal EEG features across time, frequency and space. In NIPS 2006 workshop *Current Trends in Brain-Computer Interfacing*.

Farwell, L., Donchin, E., 1988. Talking off the top of your head: toward a mental prosthesis utilizing event-related brain potentials. Electroencephalogr. Clin. Neurophysiol. 70 (6), 510–523.

Fazel, M., Hindi, H., Boyd, S.P., 2001. A Rank minimization heuristic with application to minimum order system approximation. In: Proceedings of the American Control Conference.

Fukunaga, K., 1990. Introduction to Statistical Pattern Recognition, 2nd Edition. Academic Press, Boston.

Harshman, R., 1970. Foundations of the PARAFAC procedure: models and conditions for an "explanatory" multi-modal factor analysis. UCLA Work. Pap. Phon. 16, 1–84.

Hastie, T., Tibshirani, R., Friedman, J., 2001. The Elements of Statistical Learning. Springer, Verlag.

Haufe, S., Nikulin, V.V., Ziehe, A., Müller, K.-R., Nolte, G., 2008. Combining sparsity and rotational invariance in EEG/MEG source reconstruction. NeuroImage 42 (2), 726–738.

Haynes, J.D., Rees, G., 2006. Decoding mental states from brain activity in humans. Nat. Rev. Neurosci. 7 (7), 523–534.

Hill, J., Farquhar, J., Martens, S., Bieβmann, F., Schölkopf, B., 2009. Effects of stimulus type and of errorcorrecting code design on bci speller performance. In: Koller, D., Schuurmans, D., Bengio, Y., Bottou, L. (Eds.), Advances in Neural Information Processing Systems 21. MIT Press, Cambridge, MA, pp. 665–672.

Hochberg, L.R., Serruya, M.D., Friehs, G.M., Mukand, J.A., Saleh, M., Caplan, A.H., Branner, A., Chen, D., Penn, R.D., Donoghue, J.P., 2006. Neuronal ensemble control of prosthetic devices by a human with tetraplegia. Nature 442, 164–171.

Hyvärinen, A., Karhunen, J., Oja, E., 2001. Independent Component Analysis. Wiley Interscience.

Kim, Y., Kim, J., Kim, Y., 2006. Blockwise sparse regression. Stat. Sinica 16, 375–390.

Koles, Z.J., 1991. The quantitative extraction and topographic mapping of the abnormal components in the clinical EEG. Electroencephalogr. Clin. Neurophysiol. 79, 440–447.

Krusienski, D., Sellers, E., McFarland, D., Vaughan, T., Wolpaw, J., 2008. Toward enhanced P300 speller performance. J. Neurosci. Meth. 167 (1), 15–21.

Kübler, A., Kotchoubey, B., Kaiser, J., Wolpaw, J., Birbaumer, N., 2001. Brain-computer communication: unlocking the locked in. Psychol. Bull. 127 (3), 358–375.

Lemm, S., Blankertz, B., Curio, G., Müller, K.-R., 2005. Spatio-spectral filters for improved classification of single trial EEG. IEEE Trans. Biomed. Eng. 52 (9), 1541–1548.

MacKay, D.J., 2003. Information Theory, Inference and Learning Algorithms. Cambridge University Press.

Mørup, M., Hansen, L.K., Arnfred, S.M., Lim, L., Madsen, K.H., 2008. Shift invariant multilinear decomposition of neuroimaging data. NeuroImage 42 (4), 1439–1450.

Nicolelis, M.A.L., 2003. Brain-machine interfaces to restore motor function and probe neural circuits. Nat. Rev. Neurosci. 4 (5), 417–422.

Nocedal, J., Wright, S., 1999. Numerical Optimization. Springer.

Parra, L., Alvino, C., Tang, A.C., Pearlmutter, B.A., Yeung, N., Osman, A., Sajda, P., 2002. Linear spatial integration for single trial detection in encephalography. NeuroImage 7 (1), 223–230.

Parra, L.C., Spence, C.D., Gerson, A.D., Sajda, P., 2005. Recipes for the linear analysis of EEG. NeuroImage 28 (2), 326–341.

Parra, L., Christoforou, C., Gerson, A., Dyrholm, M., Luo, A., Wagner, M., Philiastides, M., Sajda, P., 2008. Spatiotemporal linear decoding of brain state. IEEE Signal Process Mag. 25 (1), 107–115.

Penny, W.D., Roberts, S.J., Curran, E.A., Stokes, M.J., June 2000. EEG-based communication: a pattern recognition approach. IEEE Trans. Rehab. Eng. 8 (2), 214–215.

Pfurtscheller, G., da Silva, F.H.L., Nov 1999. Event-related EEG/MEG synchronization and desynchronization: basic principles. Clin. Neurophysiol. 110 (11), 1842–1857.

Pfurtscheller, G., Neuper, C., Guger, C., Harkam, W., Ramoser, R., Schlögl, A., Obermaier, B., Pregenzer, M., June 2000. Current trends in Graz brain-computer interface (BCI). IEEE Trans. Rehab. Eng. 8 (2), 216–219.

Pfurtscheller, G., Müller-Putz, G.R., Schlögl, A., Graimann, B., Scherer, R., Leeb, R., Brunner, C., Keinrath, C., Lee, F., Townsend, G., Vidaurre, C., Neuper, C., June 2006. 15 years of BCI research at Graz University of Technology: current projects. IEEE Trans. Neural Sys. Rehab. Eng. 14 (2), 205–210.

Rakotomamonjy, A., Guigue, V., 2008. BCI Competition III : Dataset II — Ensemble of SVMs for BCI P300 speller. IEEE Trans. Biomed. Eng. 55 (3), 1147–1154.

Ramoser, H., Müller-Gerking, J., Pfurtscheller, G., 2000. Optimal spatial filtering of single trial EEG during imagined hand movement. IEEE Trans. Rehab. Eng. 8 (4), 441–446.

Recht, B., Fazel, M., Parrilo, P.A., June 2007. Guaranteed Minimum-Rank Solutions of Linear Matrix Equations via Nuclear Norm Minimization. Tech. Rep. arXiv:0706.4138v1 [math.OC].

Rennie, J.D.M., Srebro, N., 2005. Fast maximum margin matrix factorization for collaborative prediction. ICML '05: Proceedings of the 22nd International Conference on Machine learning. ACM Press, New York, NY, USA, pp. 713–719.

Schalk, G., Wolpaw, J.R., McFarland, D.J., Pfurtscheller, G., 2000. EEG-based communication: presence of an error potential. Clin. Neurophysiol. 111 (12), 2138–2144.

Srebro, N., 2004. Learning with Matrix Factorizations. Ph.D. thesis, Massachusetts Institute of Technology.

Srebro, N., Rennie, J.D.M., Jaakkola, T.S., 2005. Maximum-margin matrix factorization. In: Saul, L.K., Weiss, Y., Bottou, L. (Eds.), Advances in Neural Information Processing Systems 17. MIT Press, Cambridge, MA, pp. 1329–1336.

Tibshirani, R., 1996. Regression shrinkage and selection via the lasso. J. R. Stat. Soc. Series B 58 (1), 267–288.

Tikhonov, A.N., Arsenin, V.Y., 1977. Solutions of Ill-Posed Problems. V. H. Winston & Sons.

Tomioka, R., Aihara, K., 2007. Classifying matrices with a spectral regularization. ICML '07: Proceedings of the 24th International Conference on Machine learning. ACM Press, pp. 895–902.

Tomioka, R., Sugiyama, M., 2008. Sparse learning with duality gap guarantee. In NIPS 2008 workshop *Optimization for Machine Learning*.

Tomioka, R., Sugiyama, M., in press. Dual Augmented Lagrangian Method for Efficient Sparse Reconstruction. IEEE Signal Processing Letters.

Tomioka, R., Dornhege, G., Nolte, G., Aihara, K., Müller, K.-R., 2006a. Optimizing Spectral Filters for Single Trial EEG Classification. Lecture Notes in Computer Science, Vol. 4174. Springer Berlin, Heidelberg, pp. 414–423.

Tomioka, R., Dornhege, G., Nolte, G., Blankertz, B., Aihara, K., Müller, K.-R., July 2006b. Spectrally weighted common spatial pattern algorithm for single trial EEG classification. Tech. Rep. METR-2006-40, Department of Mathematical Informatics, University of Tokyo.

Tomioka, R., Aihara, K., Müller, K.-R., 2007. Logistic regression for single trial eeg classification. In: Schölkopf, B., Platt, J., Hoffman, T. (Eds.), Advances in Neural Information Processing Systems 19. MIT Press, Cambridge, MA, pp. 1377–1384.

Vapnik, V.N., 1998. Statistical Learning Theory. Wiley-Interscience.

Wang, Y., Zhang, Z., Li, Y., Gao, X., Gao, S., Yang, F., 2004. BCI competition 2003—data set IV: An algorithm based on CSSD and FDA for classifying single-trial EEG. IEEE Trans. Biomed. Eng. 51 (6), 1081–1086.

Wolpaw, J.R., Birbaumer, N., McFarland, D.J., Pfurtscheller, G., Vaughan, T.M., 2002. Brain–computer interfaces for communication and control. Clin. Neurophysiol. 113, 767–791.

Wu, W., Gao, X., Hong, B., Gao, S., 2008. Classifying single-trial EEG during motor imagery by iterative spatio-spectral patterns learning (ISSPL). IEEE Trans. Biomed. Eng. 55 (6), 1733–1743.

Yu, J., Vishwanathan, S.V.N., Günter, S., Schraudolph, N.N., 2008. A Quasi-Newton Approach to Nonsmooth Convex Optimization. ArXiv:0804.3835v3 [stat.ML].

Yuan, M., Lin, Y., 2006. Model selection and estimation in regression with grouped variables. J. R. Stat. Soc. Series B 68 (1), 49–67.

Yuan, M., Ekici, A., Lu, Z., Monteiro, R., 2007. Dimension reduction and coefficient estimation in multivariate linear regression. J. R. Stat. Soc. Series B 69 (3), 329–346.