ISSN 0918-2802

Technical Report

A regularized discriminative framework for EEG analysis with application to brain-computer interfacing

Ryota Tomioka & Klaus-Robert Müller

TR09-0002 February

Department of Computer Science
Tokyo Institute of Technology
Ôokayama 2-12-1 Meguro Tokyo 152-8552, Japan
http://www.cs.titech.ac.jp/

©The author(s) of this report reserves all the rights.

Abstract

We propose a regularized discriminative framework for signal analysis of electroencephalography (EEG) in the context of brain-computer interfacing (BCI). The proposed approach 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.

Keywords: brain-computer interfacing, discriminative learning, regularization, group-lasso, spatio-temporal factorization, dual spectral norm, trace norm, convex optimization, P300 speller, discriminative modeling of brain imaging signals

1 Introduction

Brain-computer interfacing (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 restore 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 imagination 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

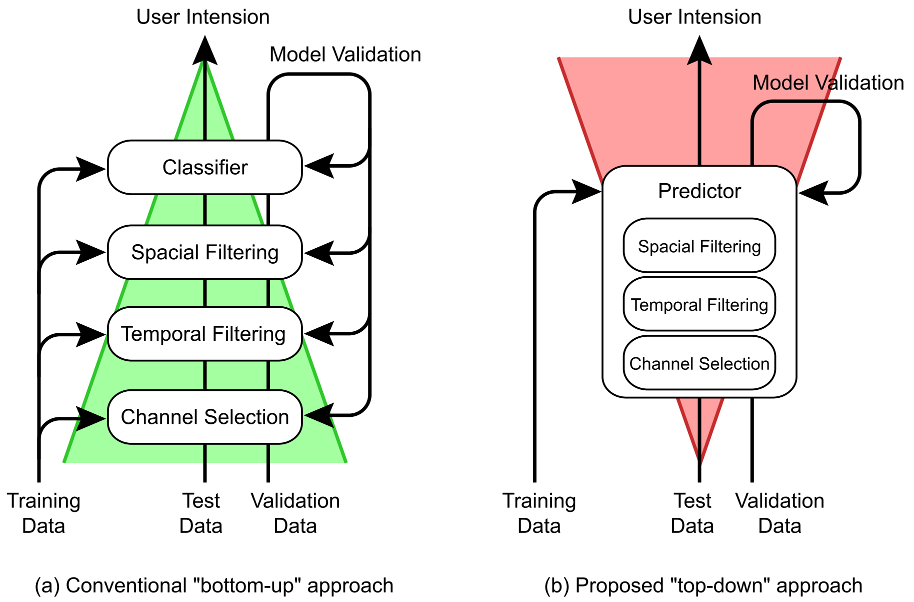


Figure 1: Comparison of the proposed "top-down" approach to the conventional "bottom-up" approach. Conventionally EEG signal has been analyzed in a bottom-up manner; different feature extraction techniques have been applied sequentially that capture different aspects of the data and often based on different criteria that are loosely connected to the prediction of user intention; high predictive accuracy can be obtained by fine tuning the feature extractors, e.g., using cross validation. The proposed "top-down" approach focuses on predicting user intention; it is based on two criteria, namely, the empirical prediction performance and the regularizer. Suitably chosen regularizers automatically induce sparse decomposition of the signal, which corresponds to conventional feature extraction and provides insights into the behavior of the predictor.

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)). We term this approach a bottom-up approach: given a large collection of single-trial EEG data (bottom) better representations of the data have been constructed in a step-by-step manner to finally obtain the classification output at the top (see Fig. 1).

Less obvious questions arise if we view this classifier from the top to bottom. How can the classifier achieve optimality with respect to the training data? How do we control the complexity of the classifier? How can we gain explanatory insights into the behavior of the estimated classifier? The last two questions are trivial when we build the classifier in a bottom-up manner. However in the bottom-up approach it is difficult to guarantee the overall optimality.

Our paper contributes to answering the above questions by presenting a *discriminative* approach based on the top-to-bottom view (see Fig. 1 (b)). In other words we focus on how well we *predict* the intention of a user in a probabilistic sense; thus we can assess the optimality of the classifier with respect to the training data while preserving transparency. The learning algorithm is naturally derived from the predictor model. For the complexity control and interpretation, we employ the concept of regularization (Tikhonov and Arsenin, 1977). It is known that through an appropriate regularization we can obtain a sequence of classifiers along the whole spectrum of complexity measured by the regularizer. Moreover, we employ two specific regularizers that induce different types of sparsity, which as we see below have nice physiological interpretation. The group-lasso type regularizers (Yuan and Lin, 2006; Haufe et al., 2008) produce sparsity in a group-wise manner; they can be used to select informative electrodes or temporal basis functions. The dual spectral regularizer (Fazel et al., 2001; Tomioka and Aihara, 2007) produce a low-rank weight matrix. The resulting low-rank weight matrix can be interpreted for example as a linear combination of small number of pairs of spatio-temporal filters. The issue of complexity control, feature extraction, and the interpretability of the resulting predictor is now tackled in a unified and systematic manner under the roof of a regularized empirical risk minimization problem. Earlier studies either had to solve a separate optimization problem (e.g., common spatial pattern (Ramoser et al., 2000; Blankertz et al., 2008)) or had to fix the number of features a priori (Tomioka et al., 2007; Farquhar et al., 2006; Dyrholm et al., 2007; Christoforou et al., 2008). Finally the proposed framework is applied to two BCI problems and shows improved classification performance while additionally providing explanation into how the classifier works.

This paper is organized as follows. In Sec. 2.1 our discriminative learning approach is presented. In Sec. 2.2, the framework is applied to the P300 speller BCI problem. In Sec. 2.3, the framework is applied to the problem of predicting self-paced finger tapping. The results for the two BCI problems are given in Secs. 3 and 4, 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 the highest performance in comparison 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 dual spectral 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 Sec. 5. Concluding remarks are given in Sec. 6

2 Materials and methods

2.1 Signal analysis framework

In this section we present our discriminative learning framework for brain-computer interfacing. 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 Secs. 2.1.1, 2.1.2, and 2.1.3, respectively.

2.1.1 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 contrast to the commonly employed bottom-up approach we look at he 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:

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

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 it; 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 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)):

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

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

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 constraint minimization problem:

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

The second term $\Omega(\theta)$ is called the regularizer and it measures the complexity of the parameter configuration θ. C is a hyperparameter 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 $\Theta_{C_1} \subset \Theta_{C_2} \subset \Theta_{C_3} \subset \cdots$ 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. 2) 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:

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

In fact, under the predictor q_θ 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:

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

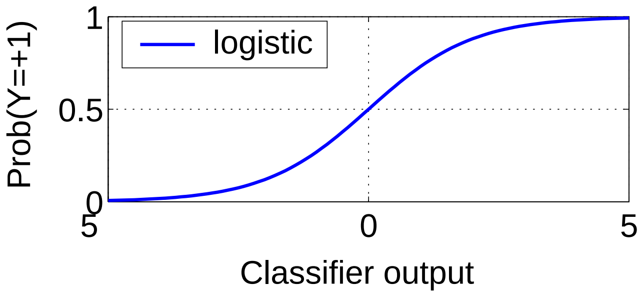


Figure 2: Logistic function (see Eq. (4)).

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 Sec. 3 and another model that uses the logistic link function in Sec. 4. 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 Sec. 5.

2.1.2 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 the 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:

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

where $\Xi^{(1)}$ is the first order term ($\boldsymbol{X}$ in the above first order model) and $\Xi^{(2,k)} = \mathtt{cov}(\boldsymbol{X}^{(k)})$ is the covariance matrix[2] of a short segment of band-pass filtered EEG $\boldsymbol{X}^{(k)}$ for $k = 1, \ldots, K$. We call $\boldsymbol{X}$ the augmented input matrix and the corresponding $\boldsymbol{W}$ the augmented weight matrix. The normalization factor η_* is defined as the square root of the total variance[3] of each block element, i.e., $\eta_* = \sqrt{\sum_{j,k} \mathsf{var}\left(\Xi^*(j,k)\right)}$ where $* \in \{(1), (2,1), \ldots, (2,K)\}$. 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 dual spectral regularization (see Sec. 2.1.3) 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 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).

2.1.3 Regularization

In this section we preset three types of regularizers $\Omega(\theta)$ 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 Euclidian norm of the weight matrix viewed as a vector.

The next two regularizers induce different types of *sparsity* in the weight matrix. The second regularizer is defined as the “linear sum of group-wise norms”, where the group is defined a priori. This type of regularization is known as group lasso (Yuan and Lin, 2006) and can be considered as a generalization of ℓ_1-regularization (known as lasso (Tibshirani, 1996)), which is used to obtain sparse solutions; that is it shrinks most of the coefficients to zero. In the case of group lasso, the sparsity is induced in a group-wise manner; i.e., when a group is switched off (one of the group norms equals zero) all the variables in that group are simultaneously switched off. We consider two variations of this regularizer; 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

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

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

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 Euclidian norm. In Eq. (9) each row is grouped together. Similarly in Eq. (10) each column is grouped together. Thus the two regularizers induce sparsity in the electrode-wise (row-wise), and the time-point-wise (column-wise) manners, respectively.

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

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

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 dual spectral 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 dual spectral 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:

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

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 dual spectral 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

[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).

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 Sec. 5).

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 dual spectral 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).

2.2 P300 speller BCI

In this section we apply the general framework presented in Sec. 2.1 to a brain-controlled spelling system known as P300 speller. The design of the spelling system is reviewed in Sec. 2.2.1. The probabilistic predictor model tailored for the P300 speller system is proposed in Sec. 2.2.2. The details about preprocessing can be found in Sec. 2.2.3.

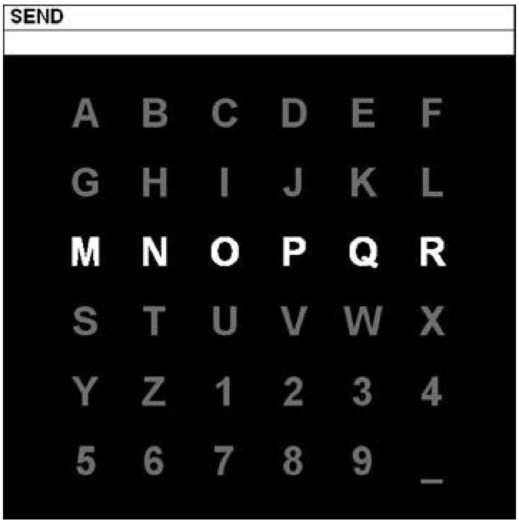
2.2.1 P300 speller system

Here we briefly describe the P300 speller system designed by Farwell & Donchin (Farwell and Donchin, 1988). The subjects are presented a 6×6 table of 36 letters on the screen (see Fig. 3); 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 in a random order. It is known that the subject's brain shows a characteristic reaction called P300 when the row or column is intensified that includes the letter on which the subject is placing his focus. Thus we can predict the letter that the subject is trying to spell by detecting the P300 response. Each intensification lasts 100ms with an interval of 75ms afterwards; the intensifications of all 6 rows and 6 columns (in a random order) are repeated 15 times; hence one letter takes $175\text{ms} \times 12 \times 15 = 31.5\text{sec}$. Note that the period of intensification (175ms) is shorter than the expected reaction of the brain (300ms). Thus the intervals we analyze are usually overlaps of several intensifications.

2.2.2 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

[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.



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	_

Figure 3: Table of letters shown on the display in the P300 speller system (Farwell and Donchin, 1988). The third row is intensified.

as follows:

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

where col(y) and row(y) are the indices of the column and the row of the letter y on the display (see Fig. 3). 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:

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

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

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

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:

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

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

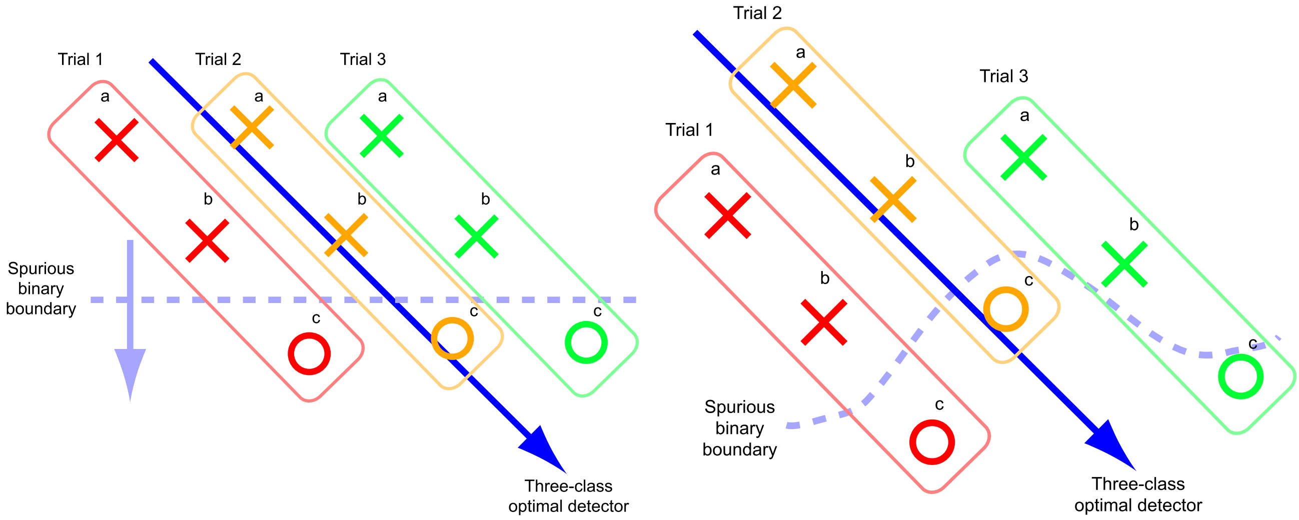
As we have seen in the previous section, the above model is used *simultaneously* for decoding the letter and learning the parameter $\boldsymbol{W}$; according to Eq. (2)) the loss function is defined as follows:

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

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. 4(a)). In other words, only the detector output 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. 4(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).

2.2.3 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 channel EEG amplifier.



(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.

Figure 4: 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).

We low-pass filter the signal at 20Hz, down sample the signal to 60Hz, and cut out an interval of 600ms from the onset of each intensification as an *epoch* $\boldsymbol{X}_{(l)} \in \mathbb{R}^{T\times C}$ where $T = 37$ and $C = 64$ ($l = 1,\ldots,12$). A *trial* $\boldsymbol{X} \in (\mathbb{R}^{T\times C})^{12}$ consists of 12 epoches 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 $\boldsymbol{P}^s$ and $\boldsymbol{P}^t$ to the low-pass filtered signal $\boldsymbol{X}^{\mathrm{LP}}_{(l)}$ as $\boldsymbol{X}_{(l)} = \boldsymbol{P}^t\boldsymbol{X}^{\mathrm{LP}}_{(l)}\boldsymbol{P}^s$. The spatial and temporal preprocessing matrices $\boldsymbol{P}^s$ and $\boldsymbol{P}^t$ are defined as follows. For the channel selection regularizer and the temporal basis selection regularizer, we choose $\boldsymbol{P}^s = \mathrm{diag}(\sigma^s_1,\ldots,\sigma^s_C)^{-1}$ and $\boldsymbol{P}^t = \mathrm{diag}(\sigma^t_1,\ldots,\sigma^t_T)^{-1}$ where σ^s_c and σ^t_t is the square-root 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 dual spectral norm regularizers, we chose $\boldsymbol{P}^s = \boldsymbol{\Sigma}^{s-1/4}$ and $\boldsymbol{P}^t = \boldsymbol{\Sigma}^{t-1/4}$, where $\boldsymbol{\Sigma}^s$ and $\boldsymbol{\Sigma}^t$ are the pooled covariance matrices in the spatial and temporal domain defined as follows:

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

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

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 $\boldsymbol{\Sigma}^s$ and $\boldsymbol{\Sigma}^t$. 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.

2.3 Self-paced finger tapping problem

In this section, the general framework presented in Sec. 2.1 is applied to the problem of single-trial prediction of self-paced finger tapping. The problem and the dataset is outlined in Sec. 2.3.1. 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 source 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 Sec. 2.3.2.

2.3.1 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.

2.3.2 Preprocessing and predictor model for the self-paced finger tapping problem

EEG is recorded from 28 electrodes at sampling frequency 1000Hz and down-sampled to 100Hz. The raw signal matrix $\boldsymbol{X}^{\mathrm{raw}} \in \mathbb{R}^{T \times C}$ is a short segment of multi-channel EEG recording starting 630ms and ending 130ms 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 20Hz and preprocessed as $\boldsymbol{X} = \boldsymbol{P}^t \boldsymbol{X}^{\mathrm{LP}} \boldsymbol{P}^s$ with $\boldsymbol{P}^t = \boldsymbol{\Sigma}^{t^{-1/4}}$ and $\boldsymbol{P}^s = \boldsymbol{\Sigma}^{s^{-1/4}}$ as in the P300 speller problem (see Sec. 2.2.3).

The second function consists of both first-order term and a wide band (7-30Hz) 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 7-30Hz and preprocessed with a spatial whitening matrix $\boldsymbol{\Sigma}^{s^{-1/2}}$, i.e., $\Xi^{(2,1)} = \boldsymbol{\Sigma}^{s^{-1/2}} \mathrm{cov}\left(\boldsymbol{X}^{\mathrm{BP}}\right) \boldsymbol{\Sigma}^{s^{-1/2}}$.

Finally the last function consists of the first-order term and two second-order terms that capture the alpha band (7-15Hz) and the beta band (15-30Hz) 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 7-15Hz and 15-30Hz, 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 transients. We test the Frobenius norm regularizer as the base line as well as the proposed dual spectral norm regularizer Our aim is to simultaneously learn and select few informative spatio-temporal filters in a systematic manner.

3 Results: P300 speller BCI

The result of the proposed framework applied to P300 speller BCI (Sec. 2.2) is given in this section. Additionally discussion including the interpretation provided by the proposed two sparse regularizers is given in 3.2.

3.1 Performance

Figures 5 (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) dual spectral 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. Note that we calculate the cross-validation accuracy for each number of repetitions M. 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 dual spectral 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.

3.2 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, the dual spectral norm

[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.

[7] The number of active components is defined as follows: given the weight matrix $\boldsymbol{W}$ let $s_1, \ldots, s_r$ 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 dual spectral regularizer.) #active components $= |s_j : s_j > 0.01 \max_j(s_j)|$.

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).

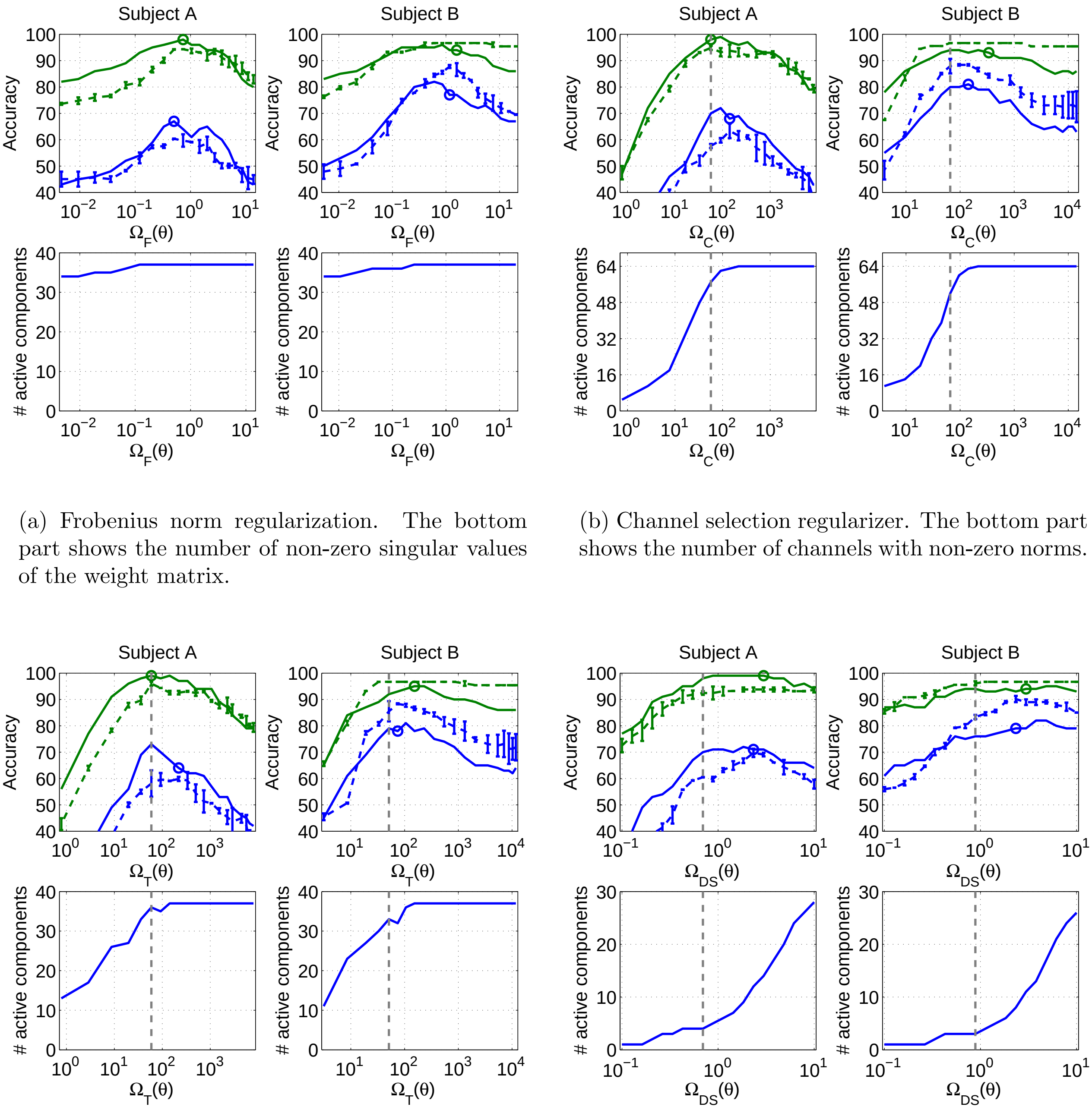
Subject	Frobenius	CSR	TSR	DS	R&G
A ($M = 5$)	67	68	64	*71*	**72**
A ($M = 15$)	*98*	*98*	**99**	**99**	97
B ($M = 5$)	77	**81**	78	*79*	75
B ($M = 15$)	93	93	*95*	94	**96**
mean ($M = 5$)	72	*74.5*	71	**75**	73.5
mean ($M = 15$)	95.5	95.5	**97**	*96.5*	*96.5*

regularizer seems to slightly outperform the winner when two subjects are averaged. Interestingly the improvement is more pronounced for small number of repetitions (e.g., subject B, $M = 5$). This observation can be confirmed in Fig. 6 where the performance of the proposed model with the dual spectral regularizer is compared against that of Rakotomamonjy and Guigue (2008) at various number of repetitions M. The proposed model is consistently better for $M < 10$. Note that the performance of Rakotomamonjy and Guigue (2008) is interpolated in between $M = 1, 2, 3, 4, 5, 10, 13, 15$.

The advantage of the proposed model is not only the classification accuracy. 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. 7 and 8 for subjects A and B, respectively. The first two plots (Fig. 7 (a)(b) and Fig. 8 (a)(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. 7(c) and 8(c)) show the leading singular vectors of the weight matrices obtained with the dual spectral regularization. We first perform singular-value decomposition of the low-rank weight matrix as $\boldsymbol{W} = \boldsymbol{U}\mathrm{diag}(\sigma_1, \ldots, \sigma_{\mathrm{r}})\boldsymbol{V}^\top$ where $\boldsymbol{U}$ is a $T \times r$ matrix and $\boldsymbol{V}$ is a $C \times r$ matrix. Then we define a *spatial filter* $\boldsymbol{w}_j$ and a *spatial pattern* $\boldsymbol{a}_j$ as follows:

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

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 $\boldsymbol{a}_j$ is orthogonal to every $\boldsymbol{w}_{j'}$ for $j' \neq j$. Similarly a temporal filter and a temporal pattern is defined from $\boldsymbol{U}$ and $\boldsymbol{P}^t$. Finally, the spatial filter, spatial pattern, and the temporal pattern are plotted from left to right



(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.

Figure 5: 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. 7 and 8.

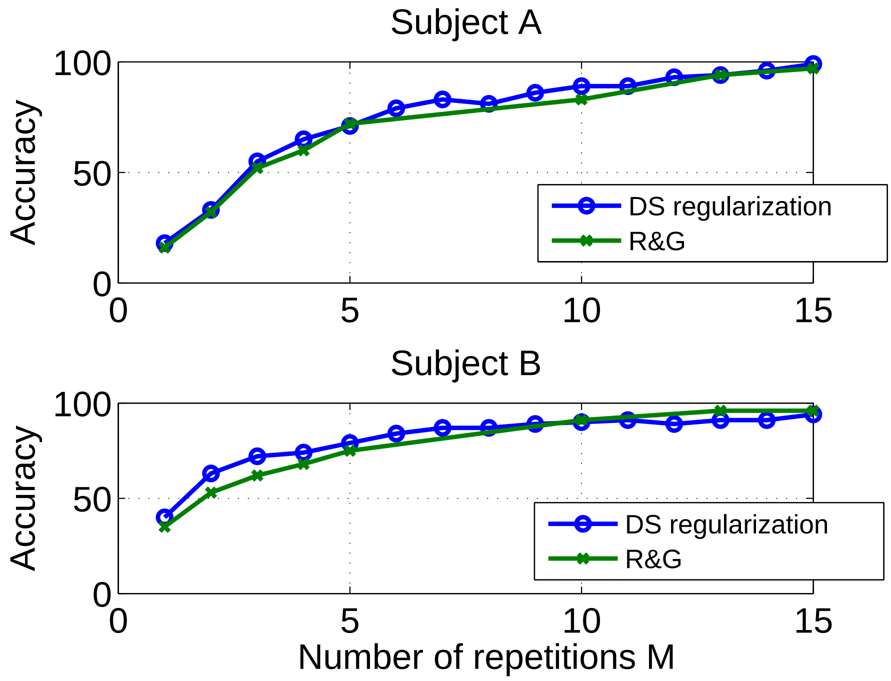


Figure 6: The accuracy of the proposed dual spectral regularization compared to Rakotomamonjy and Guigue (2008) at variety of number of repetitions.

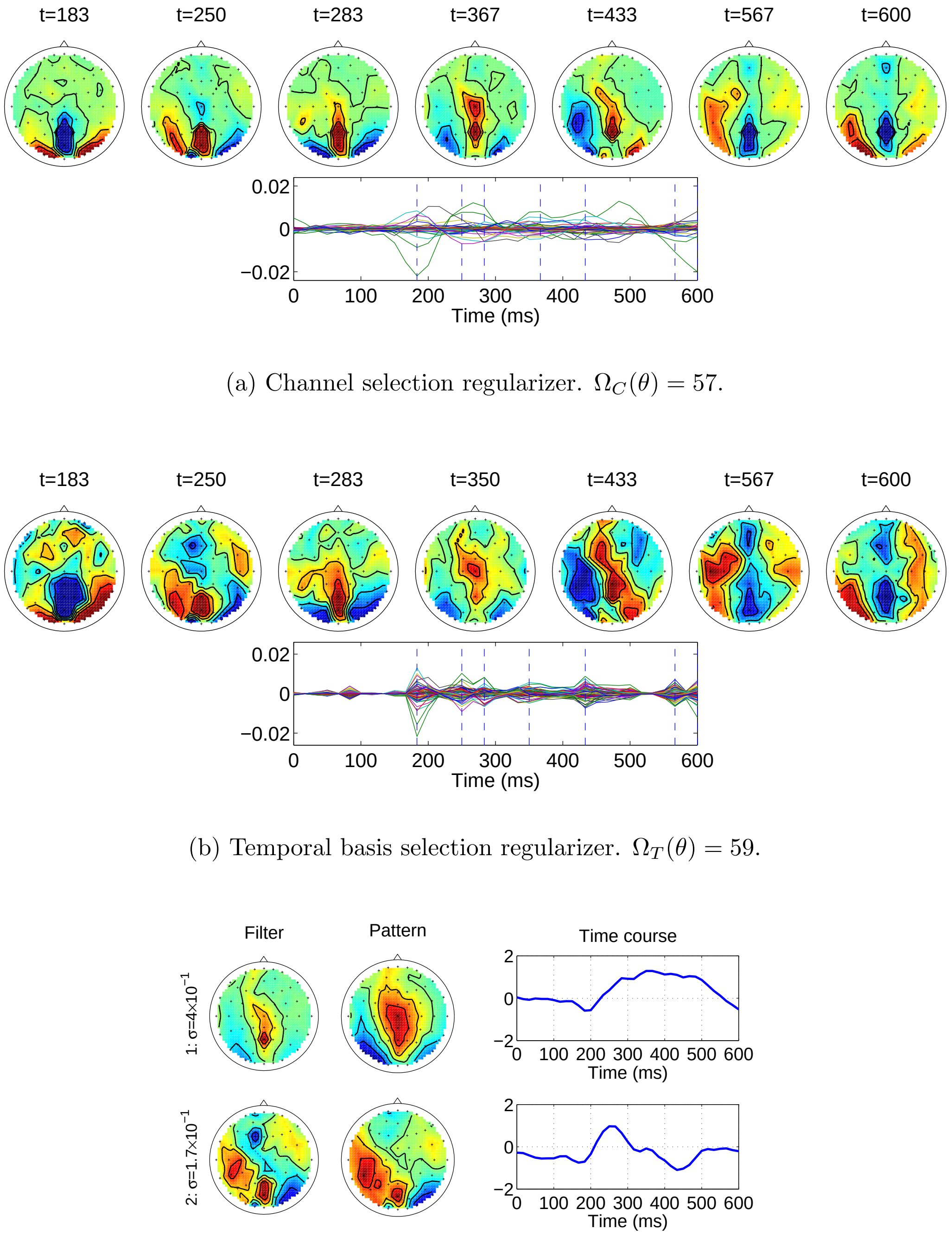
for each left/right singular vector pairs of the leading singular values from top to bottom. The spatial filters/patterns are plotted in the same way as above. The temporal patterns, which are T dimensional vectors, are plotted along time. The singular value is also shown vertically at the left end of each row.

The channel selection regularizer (see Figs. 7(a) and 8(a)) 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.

Comparing the temporal profile of the weight matrices (see e.g. the plot at the bottom of Figs. 7(a) and 7(b)), we can see that the temporal-basis selection regularizer is good at temporally localizing the discriminative information. For subject A (Fig. 7(b)), there is a negative peak at 183ms, and a broad positive component from 350ms to 500ms, which correspond to the early occipital component and the late parietal-central component mentioned above. For subject B (Fig. 8(b)), there is only a small negative peak at 150ms however there is a strong parietal peak at 217ms and a sustained discriminability from 300ms to 450ms which has more central focus.

The dual spectral 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. 7(c) and 8(c)) confirm our earlier observation that there are two major discriminative components: the early occipital component (the second row in Fig. 7(c) and the first two rows in Fig. 8(c)) and the late central component (the first row in Fig. 7(c) and the third row in Fig. 8(c)). 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 217ms. 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 217ms can be seen clearer in Fig. 9 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.

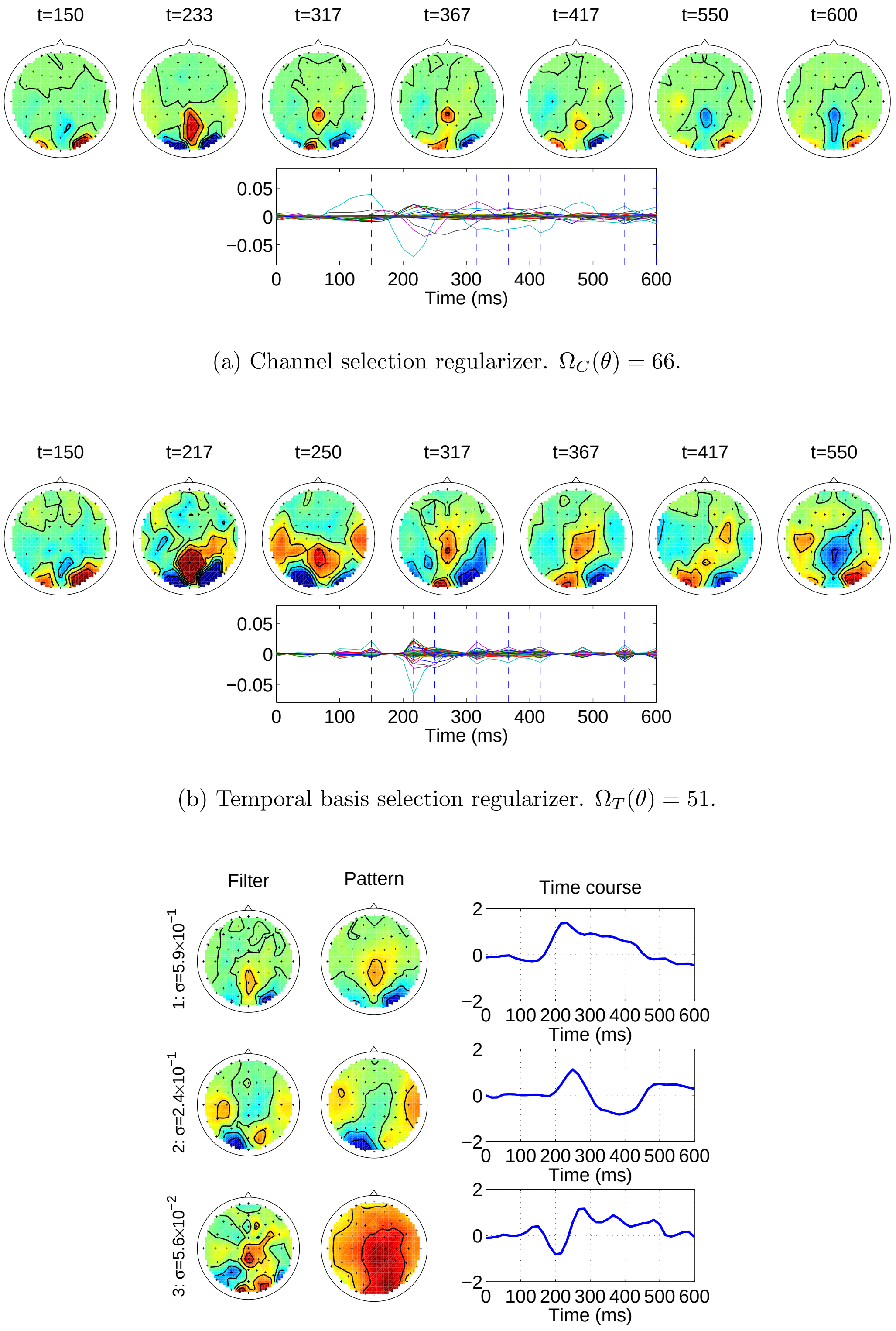


(a) Channel selection regularizer. $\Omega_C(\theta) = 57$.

(b) Temporal basis selection regularizer. $\Omega_T(\theta) = 59$.

(c) Dual spectral regularizer. $\Omega_*(\theta) = 0.70$.

Figure 7: Spatial/temporal profile of subject A.



(a) Channel selection regularizer. $\Omega_C(\theta) = 66$.

(b) Temporal basis selection regularizer. $\Omega_T(\theta) = 51$.

(c) Dual spectral regularizer. $\Omega_*(\theta) = 0.89$.

Figure 8: Spatial/temporal profile of subject B.

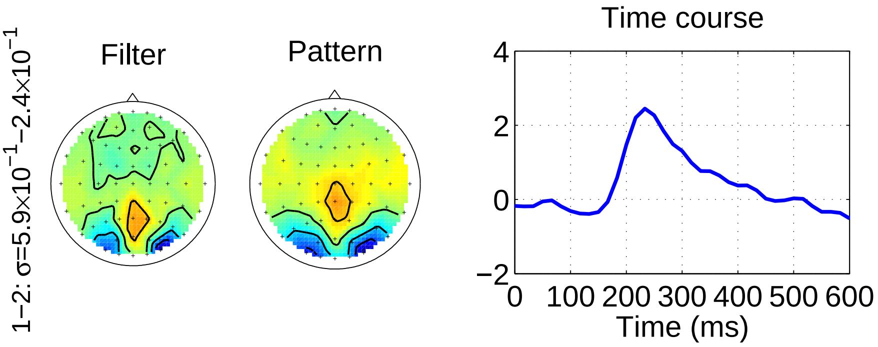


Figure 9: Spatial/temporal profile of subject B with the dual spectral regularizer. The first two components in Fig. 8(c) are merged proportional to their singular values.

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

	first order model	wide-band (7-30Hz) second-order model	double-band (7-15Hz; 15-30Hz) second-order model	Zhang *et al.* (1st-order +wide-band (10-33Hz) +19 selected channels)	SOBDA (1st-order +wide-band (10-33Hz))
#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)		

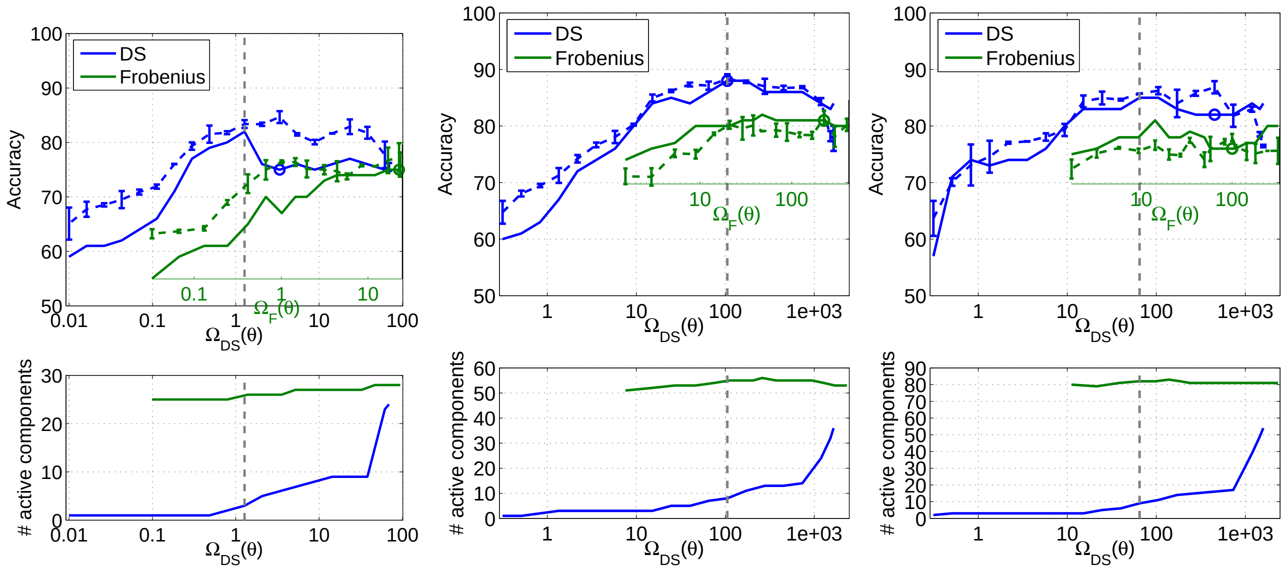
4 Results: self-paced finger tapping dataset

The result of the proposed framework applied to the self-paced finger tapping dataset (Sec. 2.3) is given in Sec. 4.1 and a discussion including the visualization of spatial/temporal filter pairs obtained from the dual spectral regularization is given in Sec. 4.2.

4.1 Performance

Figures 10(a), 10(b), and 10(c) show the classification accuracy of the proposed three detector models with the Frobenius and dual spectral norm regularization. The 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 regularization. The accuracies obtained at the selected regularization constants are marked with circles. The accuracy is plotted at the complexity measured by the dual spectral 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 dual spectral norm of the Frobenius regularized model grows monotonically with the regularization constant.

The performance obtained with the two regularizers is summarized in Tab. 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 dual spectral regularization which also achieved the highest with respect to the cross validation accuracy.



(a) First order model.

(b) Wide band Second order modelwith a first-order term and a wide-band (7-30Hz) second-order term.

(c) Double band second order model with a first-order term, alpha (7-15Hz) and beta (15-30Hz) band second-order terms.

Figure 10: 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. 11–13) are marked with vertical gray dashed lines. See Sec. 2.3.2 for the definition of the three proposed models.

4.2 Discussion

In Fig. 10(a) we can see that the performance of the dual spectral 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. 10(b) and 10(c).

Incorporating the wide-band (7-30Hz) second order term significantly improves the performance (see Fig. 10(b)) 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 (7-15Hz) and beta-band (15-30Hz) separately (see Fig. 10(c)). One possible explanation is over-fitting. 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 Tab. 2. The number of parameters are calculated as follows: for the first order model it is 28(channels) $\times$ 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 Zhang *et al.*(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[8] 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 Sec. 5).

The spatial/temporal profiles of the three proposed models are visualized in Figs. 11-13. See Sec. 3.2 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

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

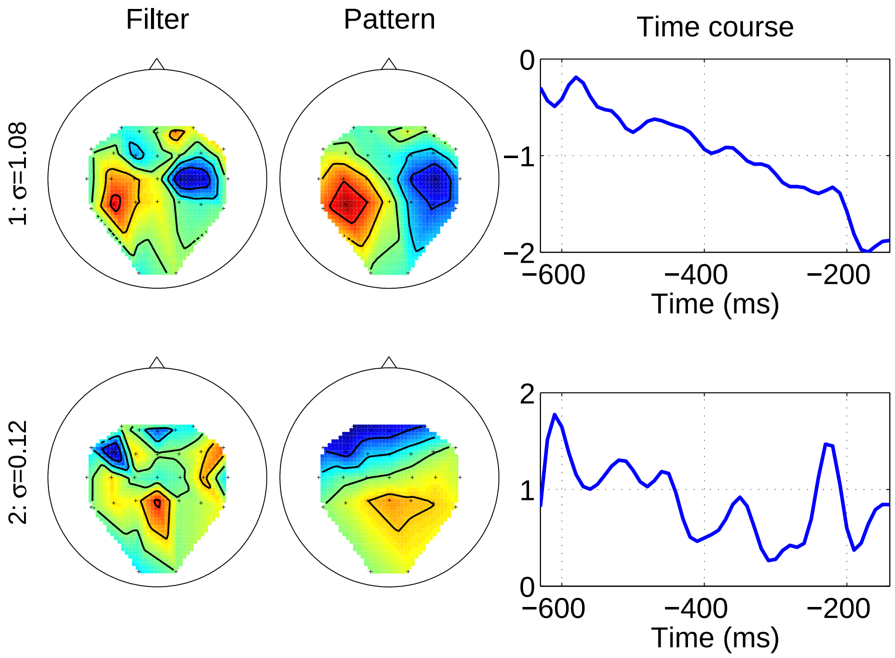


Figure 11: Spatial/temporal profile of the proposed first order model at $\Omega_{DS}(\theta) = 1.27$ (see Fig. 10(a)). The spatial filter, spatial pattern and temporal pattern that corresponds to the first two singular values of the weight matrix are shown.

this component to the classifier is one order smaller than the first component.

In Fig. 12, 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[9] 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. 13, we can find spatial filter/pattern pairs that look similar to the ERD/ERS components in Fig. 12 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. 14 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] Since the block weight matrix associate to the second-order component (see Eq. (7)) is symmetric, we show the eigenvalues instead of singularvalues for the second-order components.

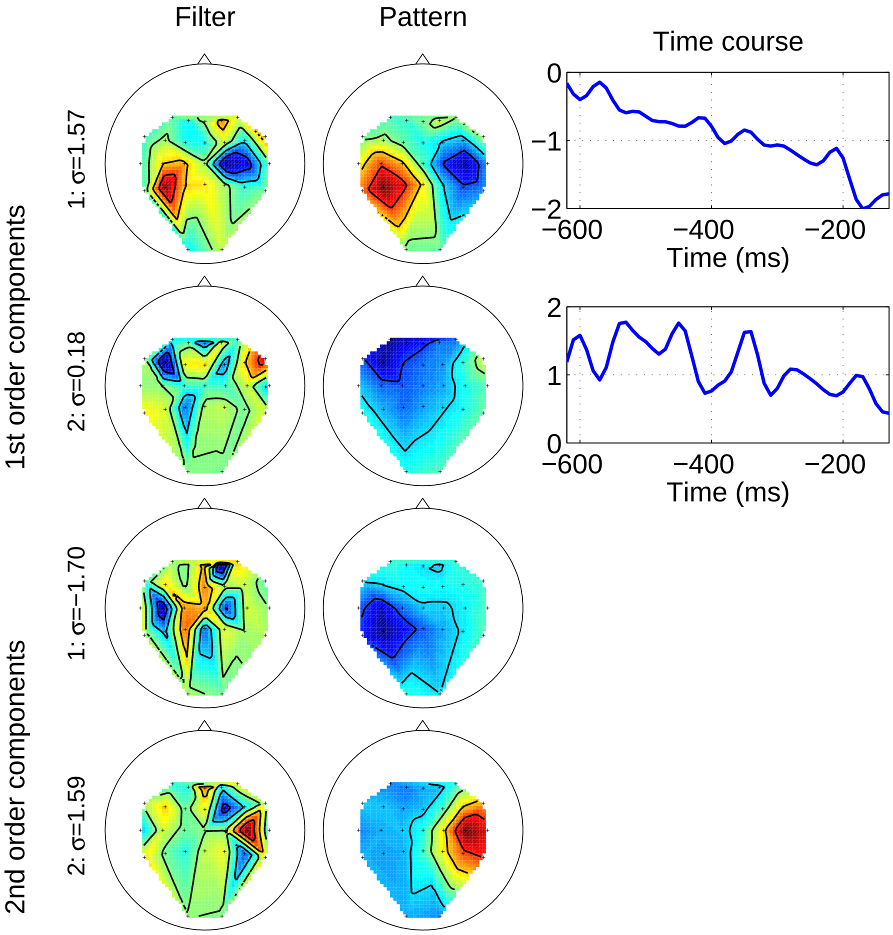


Figure 12: Spatial/temporal profile of the proposed wide-band second order model at $\Omega_{DS}(\theta) = 106$ (see Fig. 10(b)). The first two rows show the spatial filter, spatial pattern, and temporal pattern of the first order components. The last two rows show the spatial filter and pattern of the second-order components (7-30Hz). Note that there is no temporal structure for the second order components.

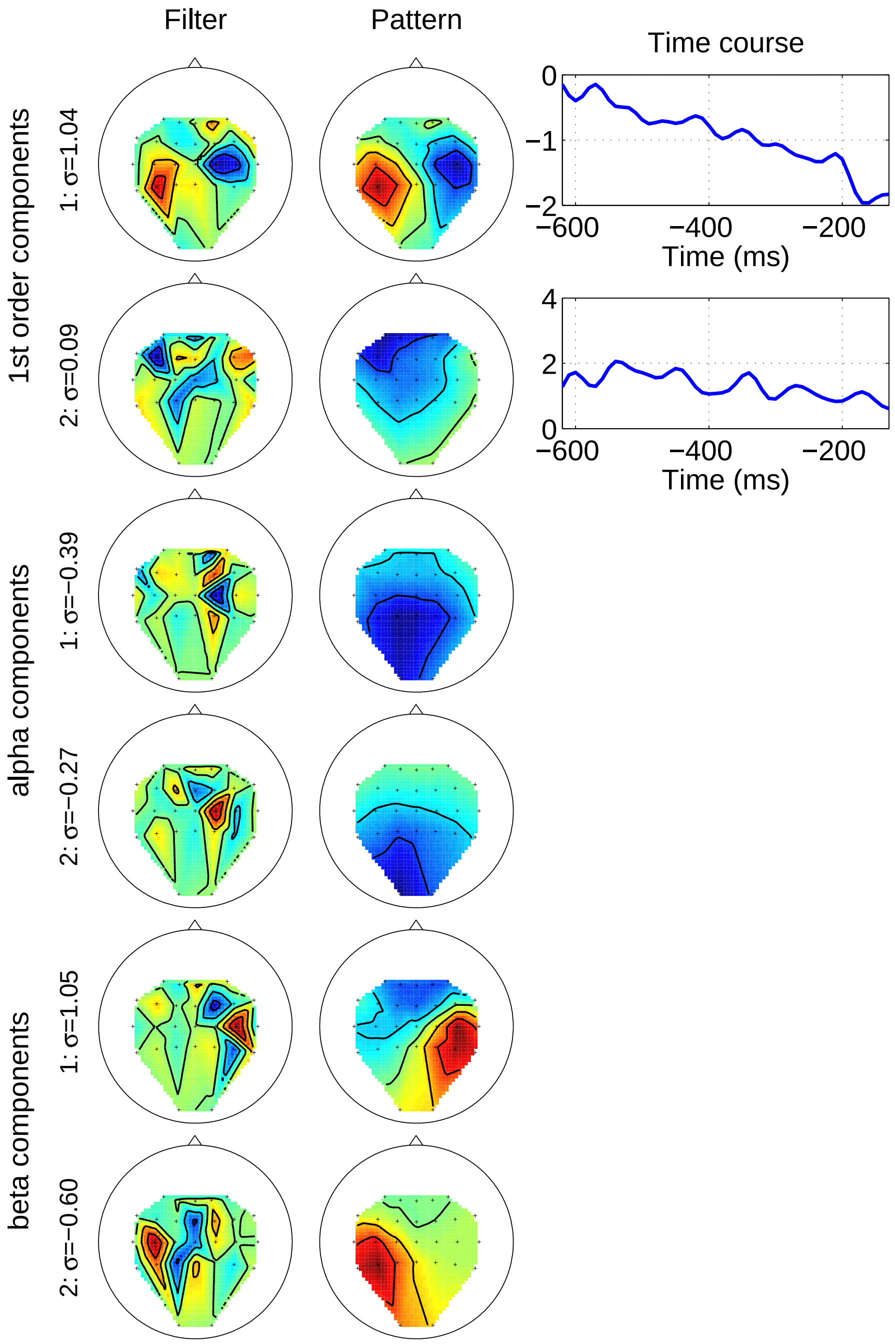


Figure 13: Spatial/temporal profile of the proposed double-band second order model at $\Omega_{DS}(\theta) = 65.0$ (see Fig. 10(c)). The first two rows show the spatial filter, spatial pattern, and temporal pattern of the first order components. The last four rows show the spatial filter and pattern of the alpha-band (7-15Hz) second order components (rows 3–4) and the beta-band (15-30Hz) second order components (rows 5–6). Note that there is no temporal structure for the second order terms.

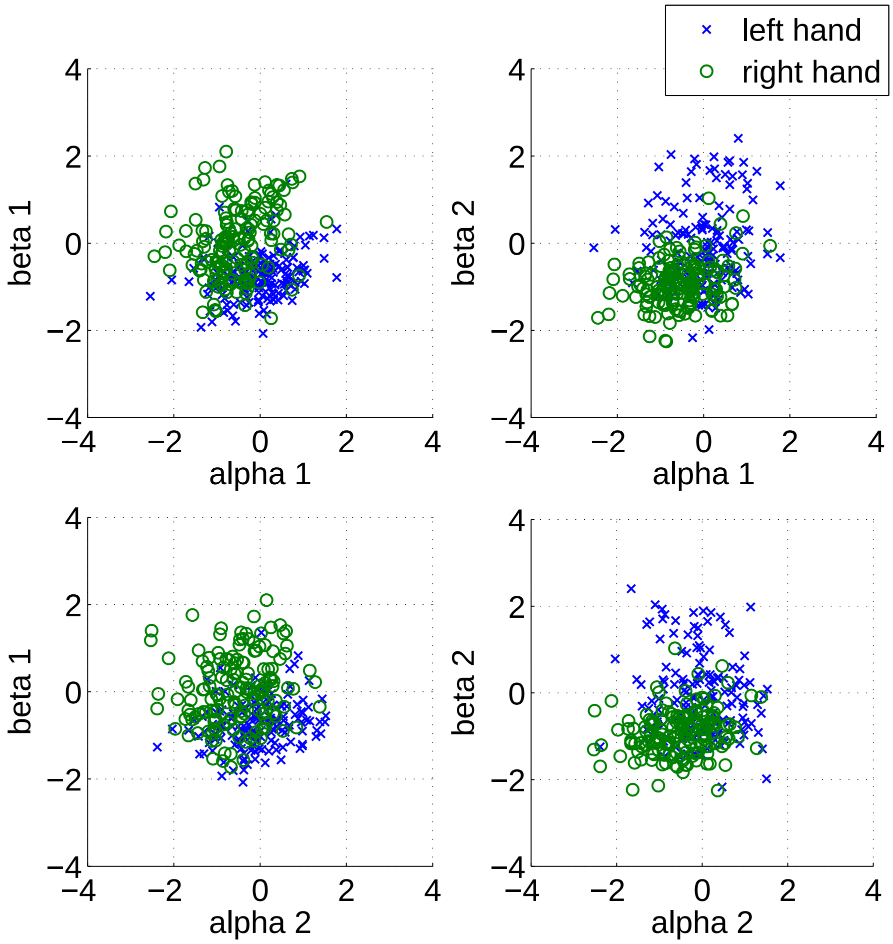


Figure 14: Comparison of the four features obtained from alpha and beta bands. The log powers of the spatially filtered training signals are plotted for the last four spatial filters shown in Fig. 13. Two filters are obtained from the alpha-band and are shown along the horizontal axes. Two filters are 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.

5 Discussion on earlier discriminative approaches

In this section we review earlier studies on discriminative modeling. In all the studies the logistic loss function (Eq. (5)) is used except (Farquhar et al., 2006) in which they used the hinge loss function. In addition, in all the studies, the (squared) Frobenius norm regularizer is used except (Tomioka and Aihara, 2007) in which they used the dual spectral norm regularizer. Thus the major difference arises in the parameterization of the detector function $f_\theta(\boldsymbol{X})$ where θ is the list of parameters for each model.

5.1 First 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):

$$f_\theta(\boldsymbol{X}) = \sum_{j=1}^{J} \boldsymbol{u}_j^\top \boldsymbol{X} \boldsymbol{v}_j + \beta_0 = \mathrm{Tr}\left(\boldsymbol{U}^\top \boldsymbol{X} \boldsymbol{V}\right) + \beta_0, \tag{17}$$

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 the 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. Thus BDA can be considered as a reduced rank regression model (Velu and Reinsel, 1998).

As the regularizer, the authors used the Frobenius norm on the coefficients $\{\boldsymbol{u}_j\}_{j=1}^J$ 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 $\boldsymbol{P}^s = \boldsymbol{\Sigma}^{s-1/4}$ and $\boldsymbol{P}^t = \boldsymbol{\Sigma}^{t-1/4}$ (see Sec. 2.2.3) 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. 7–9). We argue 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 $\boldsymbol{\Sigma}^s$ and $\boldsymbol{\Sigma}^t$.

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 dual spectral norm of $\boldsymbol{W}$ i.e.,

$$\|\boldsymbol{W}\|_* = \frac{1}{2} \min_{\boldsymbol{W}=\boldsymbol{U}\boldsymbol{V}^\top} (\|\boldsymbol{U}\|_F^2 + \|\boldsymbol{V}\|_F^2) \tag{18}$$

where $\|\cdot\|_*$ and $\|\cdot\|_F$ are the dual spectral 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 dual spectral regularization (see also Weimer et al. (2008)). 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)).

5.2 Second order feature based BCI

One of the most successful approach in motor-imagination 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., 2006; Blankertz et al., 2008):

$$f_\theta(\boldsymbol{X}) = \sum_{j=1}^{J} \beta_j \log(\boldsymbol{w}_j^\top \boldsymbol{X}^\top \boldsymbol{B}_j \boldsymbol{B}_j^\top \boldsymbol{X} \boldsymbol{w}_j) + \beta_0, \tag{19}$$

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 the 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 $\boldsymbol{B}_j$ is chosen a priori or based on some heuristics (Blankertz et al., 2008). Second, the spatial filter is obtained from solving the generalized eigenvalue problem. Third, the classifier weights $\{\beta_j\}_{j=1}^J$ are obtained from Fisher's linear discriminant analysis.

Several studies have been done based on this detector model and related models. Farquhar et al. (2006) proposed to learn all the above coefficients[10] *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 $\{\beta_j\}_{j=1}^J$. However this approach leads to a *non-convex* optimization problem which may suffer from multiple local minima and poor convergence property.

It is shown that the above model can be reformulated into a convex optimization problem (Tomioka et al., 2007; Tomioka and Aihara, 2007). Two simplifications are necessary: first we fix the temporal filter coefficient $\boldsymbol{B}_j$ as a constant $\boldsymbol{B}$; second the logarithm is omitted. Now we use the following identity:

$$\sum_{j=1}^{J} \beta_j (\boldsymbol{w}_j^\top \boldsymbol{X}^\top \boldsymbol{B}\boldsymbol{B}^\top \boldsymbol{X} \boldsymbol{w}_j) = \mathrm{Tr}\left(\boldsymbol{W}^\top \boldsymbol{X}^\top \boldsymbol{B}\boldsymbol{B}^\top \boldsymbol{X}\right),$$

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

where $\boldsymbol{W} = \sum_{j=1}^{J} \beta_j \boldsymbol{w}_j \boldsymbol{w}_j^\top \in \mathbb{R}^{C \times C}$. Finally we obtain:

$$f_\theta(\boldsymbol{X}) = \left\langle \boldsymbol{W}, \boldsymbol{X}^\top \boldsymbol{B}\boldsymbol{B}^\top \boldsymbol{X} \right\rangle + \beta_0, \tag{20}$$

where $\boldsymbol{W}$ is the weight matrix and β_0 is the bias term. Now the detector function f is a linear function of the coefficient matrix $\boldsymbol{W}$ in Eq. (20); thus the optimization problem (Eq. (3)) becomes *convex*, which can be reliably optimized (e.g., Boyd and Vandenberghe (2004)). Moreover, the dual spectral norm regularization is employed in (Tomioka and Aihara, 2007) in order to obtain a low-rank weight matrix $\boldsymbol{W}$; in fact, it was demonstrated that good classification performance is obtained with only a few spatial filters $\boldsymbol{w}_j$. Furthermore, the low-rank model obtained through the dual spectral regularization showed higher performance than fixed rank=2 model proposed in (Tomioka et al., 2007). This is confirmed by the observation that typically the dual spectral regularization chose rank=4 or 5. Note that again by Eq. (18) the rank=2 model in (Tomioka et al., 2007) can be considered as the fixed-rank approximation of the proposed second order model (with no first order term) with the dual spectral regularization.

5.3 First and second order features

Christoforou et al. (2008) extended the first order BDA (Eq. (17)) and proposed the following second-order BDA (SOBDA) model:

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

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. (17) and $\boldsymbol{w}_k \in \mathbb{R}^C$ and $\boldsymbol{B} \in \mathbb{R}^{T \times T}$ are the second order spatial and temporal filter coefficients as in Eq. (19). The difference from the popular CSP-inspired model in Eq. (19) is that the logarithm on the second order term is omitted and the temporal filter matrix $\boldsymbol{B}$ is common to all spatial components $k = 1, \ldots, K$. Moreover β_k was set to some fixed value from $\{+1, -1\}$ in order to avoid redundancy in the parameterization. Similarly to Farquhar et al. (2006), they learned 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 see that using the dual spectral norm regularization, this problem can also be reformulated as a convex optimization problem when the second order temporal filter matrix $\boldsymbol{B}$ is kept constant. Let two weight matrices $\boldsymbol{W}^{(1)}$ and $\boldsymbol{W}^{(2)}$ be $\boldsymbol{W}^{(1)} = \boldsymbol{U}\boldsymbol{V}^\top$ and $\boldsymbol{W}^{(2)} = \sum_{k=1}^{K} \beta_j \boldsymbol{w}_k \boldsymbol{w}_k^\top$. Now we can rewrite Eq. (21) as follows:

$$f_\theta(\boldsymbol{X}) = \left\langle \boldsymbol{W}^{(1)}, \boldsymbol{X} \right\rangle + \left\langle \boldsymbol{W}^{(2)}, \boldsymbol{X}^\top \boldsymbol{B}\boldsymbol{B}^\top \boldsymbol{X} \right\rangle + \beta_0. \tag{22}$$

It is easy to see that Eq. (22) can also be written as Eq. (6) by using the block diagonal concatenation (Eq. (7)). Thus using Eq. (18), SOBDA can be considered as a fixed-rank approximation of our proposed dual spectral regularization on the augmented weight matrix.

6 Conclusion

In this article we have proposed a novel unified framework for signal-analysis in EEG/MEG 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/MEG 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 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 our proposed training procedure (Sec. 2.1.1) 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 activity is spatially localized, volume conduction spreads it 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 regularizer 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 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. 7 and 8. 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 dual spectral 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 dual spectral 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 dual spectral regularization and the sum-of-squared-Euclidian-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.

Acknowledgment: 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 01IBE01A (BCI III) and 01GQ0415 (BCCNB-A4), 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 (Tools for brain-computer interaction). This publication only reflects the authors' views. We thank Benjamin Blankertz, Stefan Haufe, Kazuyuki Aihara, Motoaki Kawanabe, and Masashi Sugiyama for valuable discussions. Part of this work was done while the authors stayed at Fraunhofer FIRST.

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. In: ICML '07: Proceedings of the 24th international conference on Machine learning. ACM Press, New York, NY, USA, pp. 17–24.

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.

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., Dornhege, G., Krauledat, M., Curio, G., Müller, K.-R., 2007. The non-invasive Berlin Brain-Computer Interface: Fast acquisition of effective performance in untrained subjects. NeuroImage 37 (2), 539–550.

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., 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, see also the webpage: `http://ida.first.fhg.de/projects/bci/competition_ii/`.

Blankertz, B., Müller, K.-R., Krusienski, D., Schalk, G., Wolpaw, J. R., Schlögl, A., Pfurtscheller, G., del R. Millán, J., 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://ida.first.fhg.de/projects/bci/competition_iii/`.

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.

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

Dyrholm, M., Christoforou, C., Parra, L. C., 2007. Bilinear Discriminant Component Analysis. J. Mach. Learn. Res. 8, 1097–1111.

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.

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 Working Papers in Phonetics 16, 1–84.

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., Bießmann, F., Schölkopf, B., 2009. Effects of stimulus type and of error-correcting 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.

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.

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), 1439–50.

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

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., Christoforou, C., Gerson, A., Dyrholm, M., Luo, A., Wagner, M., Philiastides, M., Sajda, P., 2008. Spatiotemporal Linear Decoding of Brain State. Signal Processing Magazine, IEEE 25 (1), 107–115.

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

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., 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.

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.

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.

Rennie, J. D. M., Srebro, N., 2005. Fast Maximum Margin Matrix Factorization for Collaborative Prediction. In: 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. Clinical Neurophysiology 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. Journal of the Royal Statistical Society: 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. In: ICML '07: Proceedings of the 24th international conference on Machine learning. ACM Press, pp. 895–902.

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.

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

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

Velu, R., Reinsel, G. C., 1998. Multivariate Reduced-Rank Regression: Theory and Applications. Springer.

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.

Weimer, M., Karatzoglou, A., Le, Q., Smola, A., 2008. Cofi rank - maximum margin matrix factorization for collaborative ranking. In: Platt, J., Koller, D., Singer, Y., Roweis, S. (Eds.), Advances in Neural Information Processing Systems 20. MIT Press, Cambridge, MA, pp. 1593–1600.

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.

Yuan, M., Ekici, A., Lu, Z., Monteiro, R., 2007. Dimension reduction and coefficient estimation in multivariate linear regression. Journal of the Royal Statistical Society: Series B 69 (3), 329–346.

Yuan, M., Lin, Y., 2006. Model selection and estimation in regression with grouped variables. Journal of the Royal Statistical Society: Series B 68 (1), 49–67.