Journal of Machine Learning Research 11 (2010) 665-685 Submitted 6/08; Revised 10/09; Published 2/10

Second-Order Bilinear Discriminant Analysis

Christoforos Christoforou CCHRISTOFOROU@RKILEADERS.COM
R.K.I Leaders Limited
Agias Triados 26A
7100, Aradippou, Cyprus

Robert Haralick HARALICK@GC.CUNY.EDU
Department of Computer Science
Graduate Center, City University of New York
New York, NY 10011, USA

Paul Sajda PSAJDA@COLUMBIA.EDU
Department of Biomedical Engineering
Columbia University
New York, NY 10027, USA

Lucas C. Parra PARRA@ENGR.CCNY.CUNY.EDU
Department of Biomedical Engineering
City College, City University of New York
New York, NY 10031, USA

Editor: Mikio Braun

Abstract

Traditional analysis methods for single-trial classification of electro-encephalography (EEG) focus on two types of paradigms: phase-locked methods, in which the amplitude of the signal is used as the feature for classification, that is, event related potentials; and second-order methods, in which the feature of interest is the power of the signal, that is, event related (de)synchronization. The process of deciding which paradigm to use is *ad hoc* and is driven by assumptions regarding the underlying neural generators. Here we propose a method that provides an unified framework for the analysis of EEG, combining first and second-order spatial and temporal features based on a bilinear model. Evaluation of the proposed method on simulated data shows that the technique outperforms state-of-the art techniques for single-trial classification for a broad range of signal-to-noise ratios. Evaluations on human EEG—including one benchmark data set from the Brain Computer Interface (BCI) competition—show statistically significant gains in classification accuracy, with a reduction in overall classification error from 26%-28% to 19%.

Keywords: regularization, classification, bilinear decomposition, neural signals, brain computer interface

1. Introduction

The work presented in this paper is motivated by the analysis of functional brain imaging signals recorded via electroencephalography (EEG). EEG is measured across time and typically at multiple scalp locations, providing a spatio-temporal data set of the underlying neural activity. In addition, these measurements are often taken over multiple repetitions or trials, where trials may differ in the type of stimulus presented, the task given to the subject, or the subject's response. Analysis of these

signals is often expressed as a single-trial classification problem. The goal for the classifier is to determine from the EEG data which stimulus was presented or how the subject responded. Many of these classification techniques were originally developed in the context of Brain Computer Interfaces (BCI) but are now more widely used to interpret activity associated with neural processing.

In the case of BCI algorithms (Wolpaw et al., 2002; Birbaumer et al., 1999; Blankertz et al., 2002, 2003) the aim is to decode brain activity on a single-trial basis in order to provide a direct control pathway between a user's intentions and a computer. Such an interface could provide "locked in patients" a more direct and natural control over a neuroprosthesis or other computer applications (Birbaumer et al., 1999). Furthermore, by providing an additional communication channel for healthy individuals, BCI systems can be used to increase productivity and efficiency in high-throughput tasks (Gerson et al., 2006; Parra et al., 2008).

Single-trial discriminant analysis has also been used as a research tool to study the neural correlates of behavior. By extracting activity that differs maximally between two experimental conditions, the typically low signal-to-noise ratio of EEG can be overcome. The resulting discriminant components can be used to identify the spatial origin and time course of stimulus/response specific activity, while the improved SNR can be leveraged to correlate variability of neural activity across trials to behavioral variability and behavioral performance (Philiastides et al., 2006; Gerson et al., 2006; Philiastides and Sajda, 2006) In essence, discriminant analysis adds to the existing set of multi-variate statistical tools commonly used in neuroscience research (ANOVA, Hoteling T^2, Wilks' Λ test, etc.).

1.1 Traditional EEG Analysis

In EEG the signal-to-noise ratio (SNR) of individual channels is low, often at, or below -20dB. To overcome this limitation, all analysis methods perform some form of averaging, either across repeated trials, across time, or across electrodes. Traditional EEG analysis averages signals across many repeated trials for each individual electrode. Typical in this case is to average the measured potentials following stimulus presentation, thereby canceling uncorrelated noise that is not reproducible from one trial to the next. This averaged activity, called an event related potential (ERP), captures activity that is time-locked to the stimulus presentation but cancels induced oscillatory activity that is not locked in phase to the timing of the stimulus. Alternatively, many studies compute the oscillatory activity in specific frequency bands by filtering and squaring the signal prior to averaging. Induced changes in oscillatory activity are termed event related synchronization or desynchronization (ERS/ERD) Pfurtscheller and da Silva (1999).

Surprisingly, discriminant analysis methods developed thus far by the machine learning community have followed this dichotomy: First order methods in which the amplitude of the EEG signal is considered to be the feature of interest in classification—corresponding to ERP—and second-order methods in which the power of the feature is considered to be of importance for classification—corresponding to ERS/ERD. First order methods include temporal filtering and thresholding (Birbaumer et al., 1999), Fisher linear discriminants (Parra et al., 2005; Blankertz et al., 2002), hierarchical linear classifiers (Gerson et al., 2006) and bilinear discriminant analysis (Dyrholm et al., 2007; Tomioka and Aihara, 2007). Second-order methods include logistic regression with a quadratic term (Tomioka et al., 2007) and the well known common spatial patterns method (CSP) (Ramoser et al., 2000) and its variants: common spatio-spectral patterns (CSSP) (Lemm et al., 2005), and common sparse spectral spatial patterns (CSSSP) (Dornhege et al., 2006).

In the past, the process for choosing features for classification has been *ad hoc* and driven primarily by prior knowledge and/or assumptions regarding the underlying neurophysiology and task. From a machine-learning point of view, it seems limiting to commit *a priori* to only one type of feature. Instead, it would be desirable for the analysis method to extract the relevant neurophysiological activity *de novo* with minimal prior expectations.

In this paper we present a new framework that combines both first and second-order features in the analysis of EEG. Through a bilinear formulation, the method can simultaneously identify spatial linear components as well as temporal modulation of activity. These spatio-temporal components are identified such that their first and second-order statistics are maximally different between two conditions. Further, through the bilinear formulation, the method exploits the spatio-temporal nature of the EEG signals and provides a reduced parametrization of the high dimensional data space. We show that a broad set of state-of-the-art EEG analysis methods can be characterized as special cases under this bilinear framework. Simulated EEG data is then used to evaluate performance of the different methods under varying signal strengths. We conclude the paper with a performance comparison on human EEG. In all instances the performance of the present method is comparable or superior to the existing state-of-the-art.

2. Second-Order Bilinear Discriminant Analysis

To introduce the new method we start by formally defining the classification problem in EEG. We then present the bilinear model, discuss interpretation in the context of EEG, and establish a link to current analysis methods. The section concludes with the optimization criterion and regularization approaches. As the title of this section suggests, we termed our method Second-Order Bilinear Discriminant Analysis (SOBDA).

2.1 Problem Setting

Suppose that we are given examples of brain activity as a set of trials $\{\mathbf{X}_n, y_n\}_{n=1}^N, \mathbf{X}_n \in \mathbb{R}^{D\times T}, y_n \in \{-1,1\}$, where for each example n the matrix $\mathbf{X}_n$ corresponds to the EEG signal with D channels and T time samples and y_n indicates the class to which this example corresponds. The class label may indicated one of two conditions, that is, imagined right or left hand movement, stimulus or non-stimulus control conditions, etc. Given these examples the task is then to predict the class label y for a new trial with data $\mathbf{X}$.

2.2 Second-order Bilinear Model

To solve this problem we propose the following discriminant function

$$f(\mathbf{X};\theta) = C\,\mathrm{Trace}\left(\mathbf{U}^\top\mathbf{X}\mathbf{V}\right) + (1-C)\,\mathrm{Trace}\left(\Lambda\mathbf{A}^\top\mathbf{X}\mathbf{B}\mathbf{B}^\top\mathbf{X}^\top\mathbf{A}\right) + w_o\,, \tag{1}$$

where the parameters are $\theta = \{\mathbf{U}\in\mathbb{R}^{D\times R}, \mathbf{V}\in\mathbb{R}^{T\times R}, \mathbf{A}\in\mathbb{R}^{D\times K}\,\mathbf{B}\in\mathbb{R}^{T\times T'}, w_o\in\mathbb{R}, \Lambda\in diag(K)|\ \lambda_{ii}\in\{-1,+1\}, C\in[0,1]$. Some of these parameters may be specified using prior knowledge as will be discussed later. The scalars R, K and T' are chosen by the user and denote the rank of matrix $\mathbf{U}$,$\mathbf{V}$ $\mathbf{A}$ and $\mathbf{B}$. Typically we use $T' = T$. The goal will be to use the N examples to optimize these parameters such that the discriminant function takes on positive values for examples with $y_n = +1$ and negative values for $y_n = -1$. To accomplish this we will use a standard probabilistic

formalism—logistic regression—which will permit us to incorporate regularization criteria as prior probabilities on the parameter as will be explained in Sections 2.6 and 2.8.

2.3 Interpretation and Rationale of the Model

The discriminant criterion defined in (1) is the sum of a linear and a quadratic term, each combining bilinear components of the EEG signal. The first term can be interpreted as a spatio-temporal projection of the signal that captures the first-order statistics of the signal. Specifically, the columns $\mathbf{u}_r$ of $\mathbf{U}$ represent R linear projections in space (rows of $\mathbf{X}$). Similarly, each of the R columns of $\mathbf{v}_k$ in matrix $\mathbf{V}$ represent linear projections in time (columns of $\mathbf{X}$). By re-writing the term as:

$$\text{Trace}(\mathbf{U}^\top \mathbf{X}\mathbf{V}) = \text{Trace}(\mathbf{V}\mathbf{U}^\top \mathbf{X}) = \text{Trace}(\mathbf{W}^\top \mathbf{X}),$$

where we defined, $\mathbf{W} = \mathbf{U}\mathbf{V}^\top$, it is easy to see that the bilinear projection is a linear combination of elements of $\mathbf{X}$ with a rank-R constraint on $\mathbf{W}$. This expression is linear in $\mathbf{X}$ and thus captures directly the amplitude of the signal. In particular, the polarity of the signal (positive evoked response versus negative evoked response) will contribute to discrimination if it is consistent across trials. This term, therefore, captures phase-locked event related potential in the EEG signal. This bilinear projection reduces the number of model parameters of $\mathbf{W}$ from $D \times T$ dimensions to $R \times (D+T)$ which is a significant dimensionality reduction that alleviates the problem of over-fitting in parameters estimation given the small training set size. This projection assumes that the generators of class-dependent variances in the data have a low-rank contribution to each data matrix $\mathbf{X}$. This holds true in EEG data, where an electrical current source which is spatially static in the brain will give a rank-one contribution to the spatio-temporal $\mathbf{X}$ (Dyrholm and Parra, 2006).

The second term of Equation (1) is the power of spatially and temporally weighted signals and thus captures the second-order statistics of the signal. As before, each column of matrix $\mathbf{A}$ and $\mathbf{B}$ represent components that project the data in space and time respectively. Depending on the structure one enforces in matrix $\mathbf{B}$, different interpretations of the model can be achieved. In the general case where no structure on $\mathbf{B}$ is assumed, the model captures a linear combination of the elements of a rank-T' second-order matrix of the signal $\mathbf{X}\mathbf{B}(\mathbf{X}\mathbf{B})^\top$. In the case where Toeplitz structure is enforced on $\mathbf{B}$ (see Section 2.7), then $\mathbf{B}$ defines a temporal filter on the signal and the model captures powers of the filtered signal. Further, by allowing $\mathbf{B}$ to be learned from the data, we may be able to identify new frequency bands that have so far not been identified in novel experimental paradigms. The spatial weights $\mathbf{A}$ together with the Trace operation ensure that the power is measured, not in individual electrodes, but in some component space that may reflect activity distributed across several electrodes. The diagonal matrix Λ partitions the K spatial components (i.e., K columns of $\mathbf{A}$) into those that contribute power positively and those that contribute power negatively to the total sum. Since each column of $\mathbf{A}$ measures the power from different sources, then by multiplying the expression with Λ we capture the difference in power between different spatial components. As motivation consider the task of distinguishing between imagined left versus right hand movements. It is known that imagining a movement of the left hand reduces oscillatory activity over the motor cortex of the right hemisphere, while an imagined right-hand movement reduces oscillations over the left motor cortex. Each of these cortical areas will be captured by a different spatial distribution in the EEG. If we limit the columns of $\mathbf{A}$ to two, then these columns may capture the power of oscillatory activity over the right and left motor cortex respectively. One would like one of these two terms to contribute positively providing evidence of the observation belonging to the first class,

while the second should contribute negatively, supporting the observations coming from the second class. This can be achieved with the proper choice of Λ. Finally, the parameter C defines a convex combination of the first-order term and the second-order term. $C = 1$ indicates that the discriminant activity is dominated by the first-order features; $C = 0$ indicates that the activity is dominated by second-order features, and any value in between denotes the importance of one component versus the other.

2.4 Incorporating Prior Knowledge into the Model

Realizing that the parameters of the SOBDA model have a physical meaning (i.e., $\mathbf{u}_r$ and $\mathbf{a}_r$ map the sensor signal to a current-source space, $\mathbf{v}_r$ are temporal weight on a source signal and $\mathbf{b}_r$ can be arranged to represent a temporal filter) it becomes intuitive for the experimenter to incorporate prior knowledge of an experimental setup in the model. If the signal of interest is known to be in a specific frequency band, one can fix matrix $\mathbf{B}$ to capture only the desired frequency band. For example, $\mathbf{B}$ can be fixed to a Toeplitz matrix with coefficients corresponding to an 8Hz-12Hz band-pass filter, then the second-order term is able to extract power in the alpha-band which is known to be modulated during motor related tasks. It is often the case that experimenters have a hypothesis about the temporal profile of the signal of interest, for example the P300 signal or the N170 are known EEG responses with a positive peak at 300ms or negative peak at 170ms and are associated with surprise or processing of faces respectively. In such a scenario the experimenter can fix the temporal profile parameter $\mathbf{V}$ to emphasize time samples around the expected location of the peak activity and optimize over the rest of the parameters. The model also provides the ability to integrate information from fMRI studies. fMRI has high spatial resolution and can provide locations within the brain that may be known to participate in the processing during a particular experimental paradigm. This location information can be incorporated into the present model by fixing the spatial parameters $\mathbf{u}_r$ and $\mathbf{a}$ to reflect a localized source (often approximated as a current dipole). The remaining temporal parameters of the model can then be optimized.

2.5 SOBDA as a Generalized EEG Analysis Framework

The present model provides a generic framework that encompasses a number of popular EEG analysis techniques. The following list identifies some of the algorithms and how they relate to the model used in the SOBDA framework:

- Set $C = 1$, $R = 1$ and choose temporal component $\mathbf{v}$ to select a time window of interest (i.e., set $v_j = 1$ if j is inside the window of interest,$v_j = 0$ otherwise). Learn the spatial filters $\mathbf{u}$. This exactly corresponds to averaging over time and classifying in the sensor space as in Parra et al. (2002, 2005)

- Set $C = 1$ and select some $R > 1$ and choose the component vectors $\mathbf{v}_r$ to select multiple time windows of interest as in 1. Learn for each temporal window the corresponding spatial vector $\mathbf{u}_r$ from examples separately and then combine these components by learning a linear combination of the elements. This corresponds to the multiple window hierarchical classifier as in Gerson et al. (2006) and Parra et al. (2008)

- Set $C = 1$, $R = D$ while constraining $\mathbf{U}$ to be a diagonal matrix and select, separately for each channel, the time window $\mathbf{v}_r$ which is most discriminative. Then train the diagonal terms of

U resulting in a latency dependent spatial filter (Luo and Sajda, 2006a). Alternatively, in the first step, use feature selection to find the right set of time windows $\mathbf{v}_r$ simultaneously for all channels (Luo and Sajda, 2006b).

- Set $C = 1$,$R = 1$ and learn the spatial and temporal components $\mathbf{u}, \mathbf{v}$ simultaneously. This reduces to the rank-one bilinear discriminant as in Dyrholm and Parra (2006)

- Select $C = 1$ and some $R > 1$ and learn all columns of the spatial and temporal projection matrix **U** and **V** simultaneously. This results in the *Bilinear Discriminant Component Analysis (BDCA)* (Dyrholm et al., 2007).

- Set $C = 0$, $K = 2$ and fix **B** to a Toeplitz structure encoding a specific frequency band and set the diagonal of Λ to be $[1 - 1]$. Then learn the spatial component **A**. This then reduces to the logistic regression with a quadratic term (Tomioka et al., 2007) which is related to the Common Spatial Patters (CSP) algorithm of Ramoser et al. (2000).

- Define $\hat{\mathbf{X}}$ to be the concatenation of **X** with itself delayed in time by τ samples, where τ is specified by the user, fix **B** to a Toeplitz structure, $C = 0$, and $\mathbf{A} \in \mathbb{R}^{2D \times 2}$, learn the matrix **A**. This configuration can be related to the Common Spatio-Spectral Pattern algorithm of Lemm et al. (2005).

2.6 Logistic Regression

To optimize the model parameters $\mathbf{U}, \mathbf{V}, \mathbf{A}$ and **B** we use a Logistic Regression (LR) formalism. The probabilistic formalism is particularly convenient when imposing additional statistical properties on the coefficients such as smoothness or sparseness. In addition, in our experience, linear LR performs well in strongly overlapping high-dimensional data-sets and is insensitive to outliers, the later being of particular concern when including quadratic features.

Under the Logistic Regression model the probability that a trial belongs to class y after seeing data **X** is given by the class posterior probability

$$P(y|\mathbf{X};\theta) = \frac{1}{1+e^{-yf(\mathbf{X};\theta)}}\,.$$

With this definition, the discriminant criterion given by the log-odds ratio of the posterior class probability

$$\log\frac{P(y=+1|\mathbf{X})}{P(y=-1|\mathbf{X})} = f(\mathbf{X};\theta)\,,$$

is simply the discriminant function which we chose to define in (1) as a sum of linear and quadratic terms. The Likelihood of observing the N examples under this model is then given by

$$L(\theta) = -\sum_{n=1}^{N} \log(1+e^{-y_n f(\mathbf{X}_n;\theta)})\,. \tag{2}$$

Training consists of maximizing this likelihood using a gradient assent algorithm. Analytic gradi-

ents of the log likelihood (2) with respect to the various parameters are given by:

$$\begin{aligned}
\frac{\partial L(\theta)}{\partial \mathbf{u}_r} &= C\sum_{n=1}^{N} y_n \pi_n \mathbf{X}_n \mathbf{v}_r\,, \\
\frac{\partial L(\theta)}{\partial \mathbf{v}_r} &= C\sum_{n=1}^{N} y_n \pi_n \mathbf{u}_r \mathbf{X}_n\,, \\
\frac{\partial L(\theta)}{\partial \mathbf{a}_r} &= 2(1-C)\lambda_r \sum_{n=1}^{N} y_n \pi_n \mathbf{X}_n \mathbf{B}\mathbf{B}^\top \mathbf{X}_n^\top \mathbf{a}_r\,, \qquad (3) \\
\frac{\partial L(\theta)}{\partial \mathbf{b}_t} &= 2(1-C)\sum_{n=1}^{N} y_n \pi_n \mathbf{X}^\top \mathbf{A}\Lambda\mathbf{A}^\top \mathbf{X}\mathbf{b}_t\,, \qquad (4)
\end{aligned}$$

where we define

$$\pi_n = 1 - P(y_n|\mathbf{X}_n) = \frac{e^{-y_n f(\mathbf{X}_n;\theta)}}{1+e^{-y_n f(\mathbf{X}_n;\theta)}}\,,$$

and $\mathbf{u}_i, \mathbf{v}_i, \mathbf{a}_i$ and $\mathbf{b}_i$ correspond to the i_{th} columns of $\mathbf{U}, \mathbf{V}, \mathbf{A}$ and $\mathbf{B}$ respectively.

2.7 Enforcing Structure on B

If matrix $\mathbf{B}$ is constrained to have a circular Toeplitz structure then it can be represented as $\mathbf{B} = \mathbf{F}^{-1}\mathbf{D}\mathbf{F}$, where $\mathbf{F}$ denotes the orthonormal Fourier matrix with $\mathbf{F}^H = \mathbf{F}^{-1}$, and $\mathbf{D}$ is a diagonal complex-valued matrix of Fourier coefficients. In such a case we can re-write Equations (3) and (4) as

$$\begin{aligned}
\frac{\partial L(\theta)}{\partial \mathbf{a}_r} &= 2(1-C)\sum_{n=1}^{N} y_n \pi_n \mathbf{X}_n \mathbf{F}^{\mathrm{H}} \mathbf{D}\mathbf{D}^{\mathrm{H}} \mathbf{F}\mathbf{X}_n^\top \mathbf{a}_r\,. \\
\frac{\partial L(\theta)}{\partial d_i} &= 2(1-C)\sum_{n=1}^{N} y_n \pi_n \left(\mathbf{F}\mathbf{X}_n^\top \mathbf{A}\Lambda\mathbf{A}^\top \mathbf{X}_n \mathbf{F}^{\mathrm{H}}\right)_{ii} d_i\,.
\end{aligned}$$

and the parameters are now optimized with respect to Fourier coefficients $d_i = (\mathbf{D})_{i,i}$. An iterative gradient descent optimization procedure can be used to solve the minimization above.

This way of modeling $\mathbf{B}$ opens up a new perspective on the capabilities of the model. These last two equations are equally applicable for any choice of orthonormal basis $\mathbf{F}$. For example, the columns of $\mathbf{F}$ can represent a set of wavelet basis vectors. We note that a wavelet basis can be thought of as time-frequency representation of the signal; hence, proper selection of a wavelet basis allows for the method to not only capture the stationary power of the signal, but also the local changes in power within the T samples of matrix $\mathbf{X}$.

2.8 Regularization

Due to the high dimensional space in which the model lies and the limited samples available during training (typically in the order of 100), a maximum likelihood estimate of the parameters will over-train the data and have poor generalization performance. To ensure good generalization performance additional regularization criteria are required. The probabilistic formulation of Logistic Regression can incorporate regularization terms as prior probabilities resulting in maximum a posteriori (MAP) estimates.

We choose Gaussian process priors (Rasmussen and Williams, 2005) on the various parameters of the model and ensure smoothness by choosing the proper covariance matrices. Spatial and temporal smoothness is typically a valid assumption in EEG (Penny et al., 2005). Specifically, the spatial components of the model (i.e., columns of **U**, and **A**) follow a normal distribution with $\mathbf{u}_i \sim \mathcal{N}(\mathbf{0}, \mathbf{K}_u)$, $\mathbf{a}_i \sim \mathcal{N}(\mathbf{0}, \mathbf{K}_a)$ where the covariance matrices $\mathbf{K}_u$ and $\mathbf{K}_a$ define the degree and form of the smoothness of **u** and **a**. This is done through choice of covariance function: Let r be a spatial or temporal measure in context of **X**. For instance r is a measure of spatial distance between data acquisition sensors, or a measure of time difference between two samples in the data. Then a covariance function $k(r)$ expresses the degree of correlation between any two points with that given distance. For example, a class of covariance functions that has been suggested for modeling smoothness in physical processes, the Matérn class, is given by:

$$\mathrm{k}_{\text{Matérn}}(r) = \frac{2^{1-\nu}}{\Gamma(\nu)} \left(\frac{\sqrt{2\nu} r}{l} \right)^{\nu} \mathrm{B} \left(\frac{\sqrt{2\nu} r}{1} \right) ,$$

where l is a length-scale parameter, and ν is a shape parameter. Parameter l can be roughly though of as the distance within which points are significantly correlated (Rasmussen and Williams, 2005). The parameter ν defines the degree of ripple. The covariance matrix **K** is then built by evaluating the covariance function

$$(\mathbf{K})_{ij} = \sigma^2 \, \mathrm{k}_{\text{Matérn}}(r_{ij})$$

where $r_{i,j}$ denotes the physical distance of sensor-i from sensor-j, and σ^2 defines the overall scale parameter. Similarly, the Gaussian prior can be used on the columns of the temporal matrix **V** (i.e., $mv \sim \mathcal{N}(\mathbf{0}, \mathbf{K}_v)$). The Matérn function was preferred because it allows for a low parametrization of the covariance matrix (two parameters define the entire covariance), but also because of the physical and intuitive interpretation of its parameters. Specifically the parameter l is associated with the physical concept of distance between measurements (either in space or time). This understanding of the parameters is useful since it allows for an educated search strategy in setting the proper values for these parameters.

Regularizing logistic regression amounts to minimizing the negative log-likelihood plus the negative log-priors, which can be written as:

$$\arg \min_{\mathbf{U},\mathbf{V},\mathbf{A},\mathbf{B},w_o} -L(\theta) + \frac{1}{2} \left(\sum_{r=1}^{R} \mathbf{u}_r^\top \mathbf{K}_u^{-1} \mathbf{u}_r + \mathbf{v}_r^\top \mathbf{K}_v^{-1} \mathbf{v}_r + \sum_{k=1}^{K} \mathbf{a}_k^\top \mathbf{K}_a^{-1} \mathbf{a}_k + \sum_{t=1}^{T'} \mathbf{b}_t^\top \mathbf{K}_t^{-1} \mathbf{b}_t \right) , \qquad (5)$$

where we ignored constants that have no effect in the optimization. The covariances of these priors are given by $\mathbf{K}_u, \mathbf{K}_a \in \mathbb{R}^{D \times D}$ and $\mathbf{K}_v, \mathbf{K}_b \in \mathbb{R}^{T \times T}$ and control the smoothness of the parameter space. In the case of the spectral regularization we use the identity matrix for the covariance, $\mathbf{K}_b = \sigma^2 \mathbf{I}$, since the smoothness assumption does not necessarily hold in the spectral domain.

Following Rasmussen and Williams (2005) the shape parameter was chosen to be $\nu = 100$ for the spatial components and $\nu = 2.5$ for the temporal components. Reasonable choices for the length-scale parameter l may be 25ms, 50ms or 100ms and in space 1cm, 2cm, and 3cm. Cross-validation was used to select among these choices. The overall scale parameters σ were chose to be the same for space and time components, but allowed to take on separate values for the first and second order component. We used a line-search procedure in combination with cross-validation to select appropriate values for σ.

2.9 Optimization

Optimization (5) is achieved using a coordinate decent type algorithm (Nielsen, 2005) with parameters $\mathbf{U}, \mathbf{V}$ and $\mathbf{A}, \mathbf{B}$ optimized separately. We obtain analytic expressions for both the gradient and the Hessian of the function, however, in the optimization only the gradient information is used.[1] We first optimize the parameters $\mathbf{U}$ and $\mathbf{V}$, then optimize parameters $\mathbf{A}$ and $\mathbf{B}$ and finally perform a line search to determine the value of C.

Given that the optimization function is non-convex, the gradient decent method only finds local minima. In fact, the performance of SOBDA is particularly sensitive to the starting conditions of the spectral parameter $\mathbf{d}$ (parameter $\mathbf{d}$ enters the model when enforcing a Toeplitz structure on $\mathbf{B}$, see section 2.7.), while it is quite robust to the choice of initial conditions for the remaining parameters $\mathbf{U}$, $\mathbf{V}$ and $\mathbf{A}$. A common technique in global optimization is to use parameter seeding and multiple runs of the optimization procedure. For most parameters it was sufficient to try a few random initial starting points. However, for the spectral parameter we found it important to initialize to a frequency band that was expected to carry useful information, for example, 8Hz-30Hz. Note that the present learning task falls into the class of bi-convex optimization problems for which efficient algorithms have been developed (Floudas, 1997).

3. Results

We evaluated our algorithm on 3300 simulated data sets as well as 6 real EEG recordings, including a data set used in the Brain Computer Interface Competitions II (Blankertz et al., 2004). The simulation aims to quantify the algorithm's performance on a broad spectrum of conditions and various noise levels, as well as to compare the extracted spatial, temporal and frequency components with ground truth. The evaluation on real data set compares the cross-validation performance of the proposed method with three popular methods used in EEG analysis and BCI. Results show that our method outperformed these methods, decreasing the overall classification error rates from 26%-28% to 19%. For the data set of the BCI competition we also report performance results on the independent test set and compare to the previous results.

The three methods we will compare with are Bilinear Discriminant Component Analysis (BDCA) (Dyrholm et al., 2007), Common Spatial Patterns (CSP) (Ramoser et al., 2000), and Matrix Logistic Regression (MLR) (Tomioka et al., 2007). For the evaluation on the 6 real EEG data sets, we further compare our method to the trace norm regularized Matrix Logistic Regression (sMLR) (Tomioka and Aihara, 2007). These may be considered current state-of-the art methods in EEG single-trial analysis. In our evaluation we use a rank one approximation for the BDCA as in Dyrholm et al. (2007). We implemented CSP following the description of Ramoser et al. (2000). We used two spatial patterns (SP) and employ a logistic regression classifier on the resulting SP. In the case of MLR we use the rank-2 approximation as described in the corresponding paper (Tomioka et al., 2007). For sMLR we used the implementation provide in Tomioka and Aihara (2007). Since CSP,MLR and sMLR require the data to be band-pass filtered to the frequency of interest, data sets where filtered in the range of 8Hz-30Hz for these two methods. For our algorithm we use rank-1 for the first-order parameters $\mathbf{U}$ and $\mathbf{V}$ with $R = 1$. For the spatial parameter $\mathbf{A}$ we set $K = 2$ allowing for two spatial patterns, while we enforce a Toeplitz structure on $\mathbf{B}$. We initialize the parameters

1. We discard the Hessian information because of its computational cost and the non-convexity of the optimization function. The Hessian of a non-convex function would need to be approximated by a positive definite matrix in each iteration.

$\mathbf{U}, \mathbf{V}$ and $\mathbf{A}$ by a random assignment. While we initialize the matrix $\mathbf{B}$ to encode a band-pass filter in the range of $8Hz - 30Hz$ as in the case of CSP, MLR, sMLR. As discussed in Section 2.7, enforcing a Toeplitz structure on $\mathbf{B}$ implies a representation of $\mathbf{B}$ in the form $\mathbf{B} = \mathbf{F}^{-1}\mathbf{D}\mathbf{F}$, where $\mathbf{F}$ denotes the orthonormal Fourier matrix with $\mathbf{F}^{\mathrm{H}} = \mathbf{F}^{-1}$, and $\mathbf{D}$ is a diagonal complex-valued matrix of Fourier coefficients. In our implementation, we optimize the coefficients of the matrix $\mathbf{D}$ instead of $\mathbf{B}$ directly.

3.1 Simulated EEG Data

Simulated data for a two-class problem was generated using standard EEG simulation software (GmbH, 2006). This software can generate electrode measurements under the assumption of dipolar current sources in the brain. We used 3 dipoles at three different locations, with one dipole used to generate evoked response activity, one dipole to generate induced oscillatory activity, and one dipole to generate unrelated noise/interference. The first dipole's component simulates a P300 evoked response potential (ERP) signal. We used a half-sinusoid lasting 125ms with the peak positioned at 300ms after trial-onset and a trial-to-trial Gaussian temporal jitter with standard deviation of 10ms. The second dipole's component simulates ERS/ERD in the frequency band of 8Hz to 30Hz. A variable signal in this frequency band was generated by bandpass filtering an uncorrelated Gaussian process. The third dipole was used to generate noise in the source space representing brain activity that is not related to the evoked/induced activity. Electric potentials at $D = 31$ electrode locations were generated corresponding to 500ms of EEG signal sampled at 100Hz ($T = 50$ samples). In addition to this rank-one noise we added noise to each sensor representing other sources of noise (muscle activity, skin potentials, inductive noise, amplifier noise, etc.). All noise sources were white. Trials belonging to the first class ($y_n = +1$) contained the ERP and ERD/ERS source signals scaled appropriately to achieve a specified SNR for each data set. The second class was generated by only including the noise with no ERP or ERD/ERS activity. A data set is specified by indicating the SNR for the ERP component and the SNR for the ERD/ERS component. A total of 500 trials for each class were generated for each classification problem. The SNR of the ERP component is in the range of -33dB to -13dB, and in the range of -22dB to -10dB for the oscillatory component. This is a very broad range in terms of SNR. We note that -20dB translates to the signal being 10 times smaller than the noise. ERP signals are known to be as low as $-20dB$ so this evaluation captures some extreme cases of SNR. We generated 35 data sets for each combination of SNR resulting to a total of 3300 data sets.

3.2 Performance Results on Simulated Data

The simulation results are summarized in Figure 1. The top two rows show the performance of each of the methods as a function of the SNR. The contours of the classification performance for each method as a function of the SNR of the first-order and the second-order components are shown. It is clear that BDCA performance is only affected by the noise in the linear term while CSP and MLR performance only changes as a function of the second-order component's SNR. SOBDA however, uses both first and second-order terms, hence performs well in data sets where at least one of the components has reasonable SNR. This finding confirms that SOBDA performs well in a broader range of SNRs than the other three competitive methods. The third row in 1 shows the difference in classification performance between SOBDA vs (BDCA,CSP,MLR).

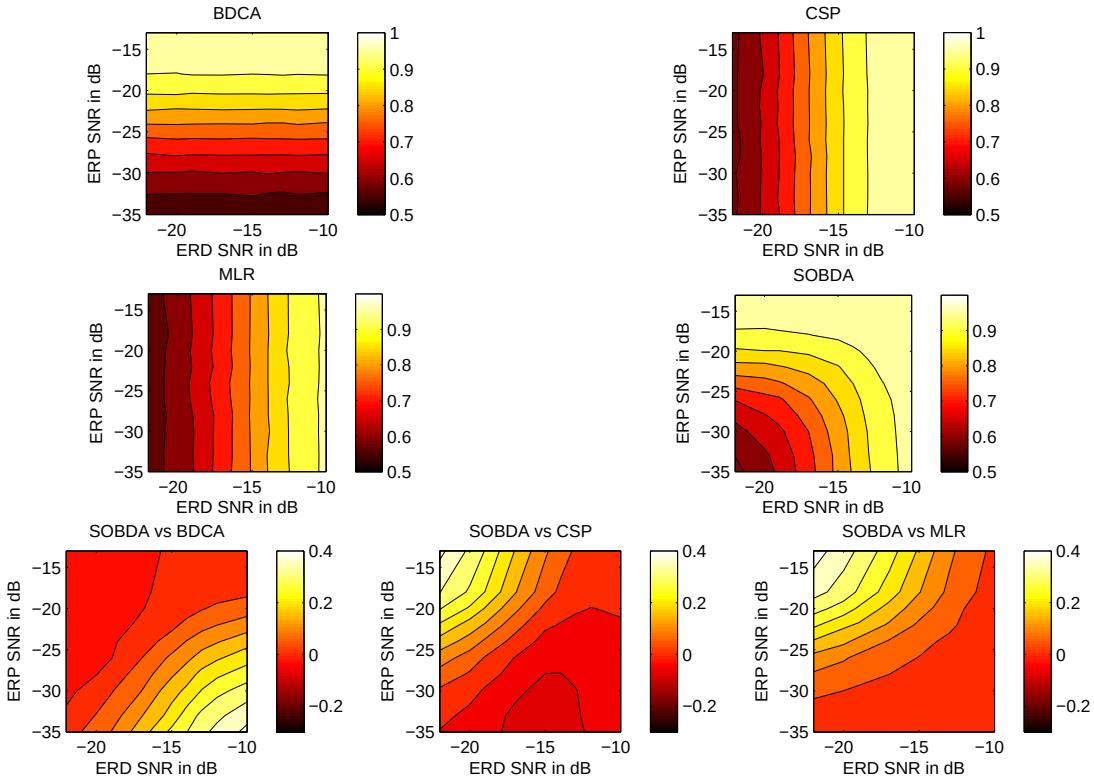

Figure 1: Performance results on simulated data. *Second and third row:* Probability of correct classification (Pc) as a function of the component's SNR. SOBDA equi-performance contours span larger area in the SNR space than any of the other three algorithms. *Third row:* Difference in Pc performance between SOBDA and each of the three methods as a function of components SNR.

As a decomposition method, SOBDA extracts spatial, temporal and frequency components. The advantage of simulated data is that we can now compare the extracted information to ground truth. The component recovered for one of the data sets at $-22dB$ and $-15dB$ is shown in figure 2. The first row shows the extracted temporal component $\mathbf{U}$ and the frequency component $\mathbf{d}$.[2] We can see that the method extracted a temporal component with a peak at 300ms which is exactly the signal used in the simulation data design. Similarly, the frequency band extracted shows a higher amplitude in the range of 8Hz-30Hz which is the band used to generate the oscillatory component. The spatial components extracted and the corresponding dipole used in the model generation are shown in rows two and three in the figure. It is clear that the topography of the extracted components is similar for the first and second-order components. The last column of the figure captures the second-order oscillatory component and the dipole of the rank one noise. Visual inspection allows one to give neurological interpretations to the extracted components. Further, the results can be used as input to

2. $\mathbf{d}$ the vector of diagonal elements of matrix $\mathbf{D}$, such that $\mathbf{B} = \mathbf{F}^{\mathrm{H}}\mathbf{D}\mathbf{F}$

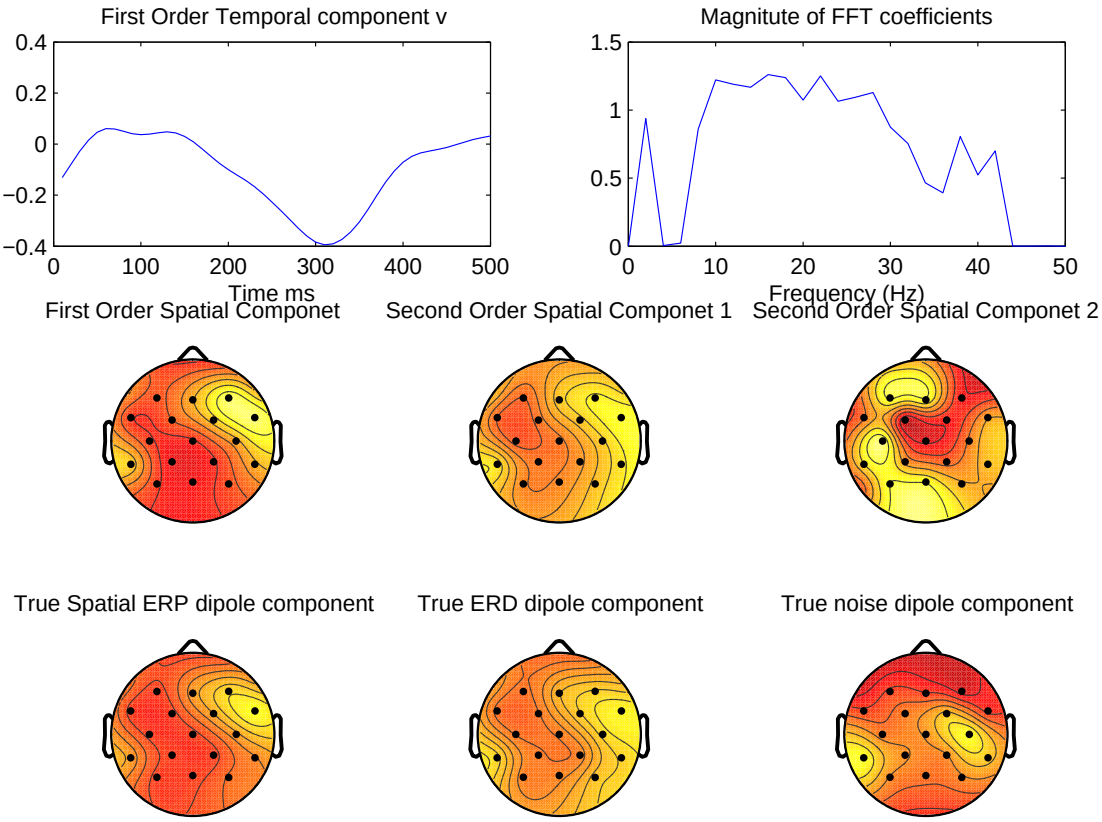

Figure 2: Extracted components on simulated data set with first-order SNR at $-22dB$ and second-order SNR at $-15dB$. *Top row:* Extracted temporal weight of linear term (left) and frequency weights of quadratic term (right). *Center row:* Extracted spatial weights. *Bottom row:* Distribution of electric potentials corresponding to the three dipoles used during stimulus generation.

a source localization algorithm, or as a guide to reduce the number of electrodes in a brain computer interface.

3.2.1 GENERIC INITIALIZATION OF FREQUENCY COMPONENT

In the evaluation presented above, we initialized the matrix **B** to encode a band-pass in the range of 8Hz - 30Hz as it was the case for CSP and MLR. In this section we demonstrate the ability of the proposed SOBDA in cases where no initialization information is available. Specifically, we evaluated the SOBDA algorithm on a simulated data set using the process described above, but this time we initialize matrix B to a high-pass filter with cut of frequency at 1 Hz. High pass filtering is a standard preprocessing steps in EEG that removes the DC power. Figure 3 shows the temporal and frequency component obtained from SOBDA. As it is evident from the figure, the resulting frequency component has higher weights for frequencies in the band 8Hz-30Hz, which is the band used to generate the power component in the simulated data. Thus the proposed method is able optimize the frequency band even in cases where we use a generic initialization of the matrix **B**.

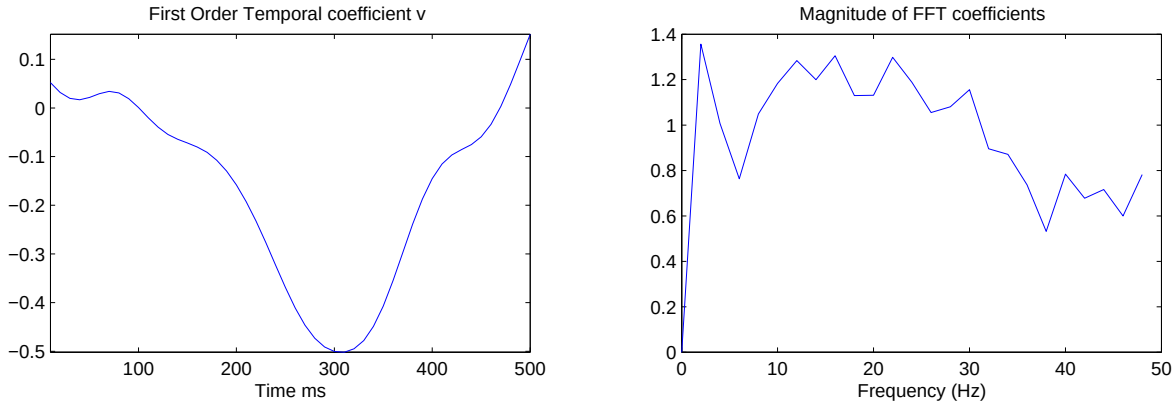

Figure 3: Discriminant coefficients on simulated data set with first-order SNR at $-22dB$ and second-order SNR at $-15dB$. The Fourier coefficients were initialized to a high-pass filter with cut off frequency at 1 Hz *Left figure:* Extracted temporal weight of linear term.*Right figure:* Magnitude of the Fourier coefficients in $\mathbf{D}$, such that $\mathbf{B} = \mathbf{F}^{\mathrm{H}}\mathbf{D}\mathbf{F}$.

3.3 Human Subject EEG

To evaluate the performance of the proposed method on real data we first applied the algorithm to an EEG data set that was made available through The BCI Competition 2003 (Blankertz et al., 2004, Data Set IV). EEG was recorded on 28 channels for a single subject performing "self-paced key typing", that is, pressing with the index and little fingers corresponding keys in a self-chosen order and timing. Key-presses occurred at an average speed of 1 key per second. Trial matrices were extracted by epoching the data starting 630ms before each key-press. A total of 416 epochs were recorded, each of length 500ms. For the competition, the first 316 epochs were used for classifier training, while the remaining 100 epochs were used as a test set. Data was recorded at 1000Hz with a pass-band between 0.05 and 200Hz, then down sampled to 100Hz sampling rate.

For this experiment, the matrix $\mathbf{B}$ was fixed to a Toeplitz structure that encodes a 10Hz-33Hz bandpass filter and only the parameters $\mathbf{U}, \mathbf{V}, \mathbf{A}$ and w_0 were trained. The number of columns of $\mathbf{U}$ and $\mathbf{V}$ were set to $R = 1$ and the number of columns for $\mathbf{A}$ was set to $K = 2$. The selection of these parameters is motivated by the task at hand. Specifically, we are looking for one ERP components associated with the *readiness* potential that is, the slow increase in amplitude before an actual hand movement. In the case of the second-order term involving the parameter $\mathbf{A}$ we set $K = 2$ because we are interested in finding the modulation of oscillatory activity associated with the different movements of the movements of the hands. Hands and fingers are represented in somato-sensory cortex covering different areas and will hence modulate activity in distinct spatial profiles. In order to detect the power difference of these two components we set, $\Lambda = [1, 0; 0, -1]$, in agreement with the original approach of Wolpaw et al. (2002).

The temporal filter was selected based on prior knowledge of the relevant frequency band. This demonstrates the flexibility of our approach to either incorporate prior knowledge when available or extract it from data otherwise. Regularization parameters where chosen via a five fold-cross validation procedure as described in Section 2.8.

Benchmark performance was measured on the test set which had not been used during either training or cross-validation. The number of misclassified trials in the test set was 13 which places

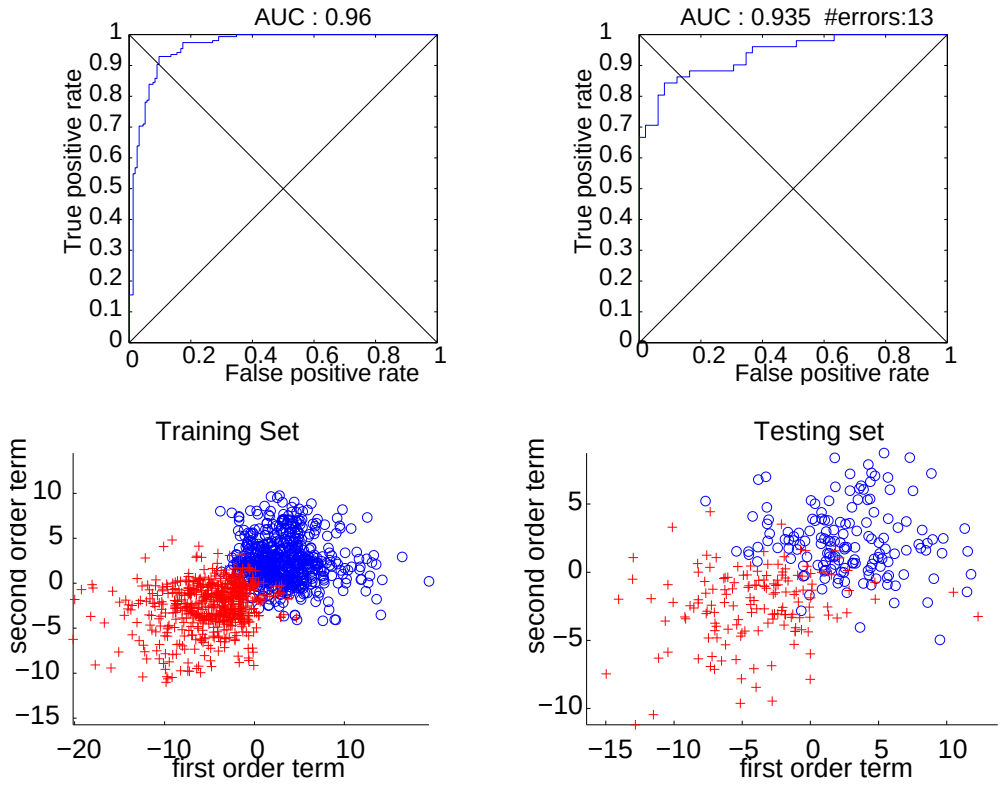

Figure 4: Results on human EEG for BCI. *Top row:* Cross-validation performance shown as ROC curve with area under the curve of 0.96 for the benchmark data set (left) and 0.93 for the independent test set (right). There were a total of 13 errors on unseen data, which is less than any of the results previously reported. *Bottom row:* Scatter plot of the first-order term vs second-order term of the model, on the training and testing set for the benchmark data set ('+' left key, and 'o' right key). It is clear that the two types of features contain independent information that can help improve the classification performance.

our method in a new first place ranking, based on the results of the competition (Blankertz et al., 2004). The receiver-operator characteristic curve (ROC) for cross-validation and for the independent test set are shown in Figure 4. The Figure also shows the contribution of the linear and quadratic terms for every trial for the two types of key-presses.

To further validate our method we performed our own EEG recordings asking subjects now to respond with the left and right index fingers. We obtain five more data sets with the same number of electrodes. For each data set and each algorithm we performed 20 repetitions of a five-fold cross-validation procedure. Each repetition uses a different partitioning of the data. For the cross-validation evaluation of these data sets, we initialized (but did not fix) matrix $\mathbf{B}$ to a Toeplitz structure that encodes a 10Hz-33Hz bandpass filter and trained over all parameters $\mathbf{U}, \mathbf{V}, \mathbf{A}, \mathbf{B}$ and w_0.[3] The number of columns of $\mathbf{U}$ and $\mathbf{V}$ were set to 1, where two columns were used for $\mathbf{A}$. This corresponds to the parameter configuration of $R = 1, K = 2$ and $T' = T$.

3. We remind the reader that in the actual implementation we optimize the Fourier coefficients $\mathbf{D}$ instead of matrix $\mathbf{B}$

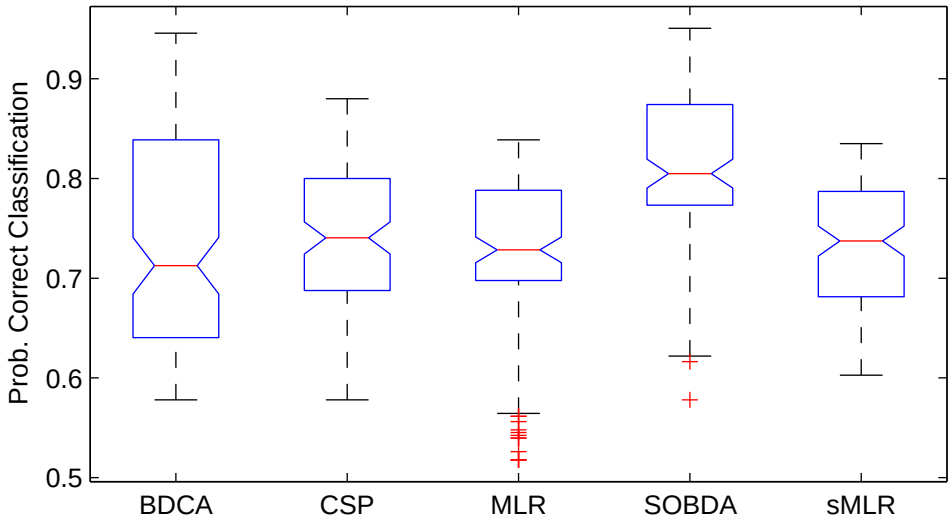

Figure 5: Estimate of the spread of the probability of correct identification from multiple cross-validation repetitions. Lines show lower quartile, median, and upper quartile values for each of the methods on all data sets. + symbols represent outliers.

Figure 5 shows performance distribution across these bootstrap repetitions using a standard boxplot. The mean performance and standard deviation of each data set and algorithm are summarized in table 1. The reduction in the overall classification error is from 26%-28% to 19%. In the mean, SOBDA outperforms competitive methods in five out of the six data sets, while achieving a comparable performance on data set 2. The performance optained with SOBDA is comparable to performance gains that may be obtained by combining existing first and second order methods (e.g., CSP and BDCA—data not shown).

Figure 6 shows the extracted components for 3 of the 6 data sets. We note that in all three cases the extracted components follow the general shape of the pre-motor or readiness potential (a.k.a. Bereitschafts potential) which known to precede a voluntary muscle movement. In addition, for two of the data sets, the frequency weightings suggest that alpha band activity also provides discriminant information for this task. This finding is consistent with the changes in the μ rhythm—that is, alpha-band activity localized over the motor cortex and associated with motor planning and execution. This demonstrates the ability of our method to learn first and second-order features that are consistent with, and can be linked to existing knowledge of the underlying neuronal signal generators.

Experiment	BDCA	CSP	MLR	SOBDA	sMLR
1	0.84 ± 0.011	0.8 ± 0.017	0.82 ± 0.011	$\mathbf{0.88 \pm 0.013}$	0.78 ± 0.0089
2	0.69 ± 0.037	$\mathbf{0.84 \pm 0.017}$	0.77 ± 0.028	0.83 ± 0.021	0.82 ± 0.012
3	$\mathbf{0.63 \pm 0.018}$	0.62 ± 0.016	0.55 ± 0.02	$\mathbf{0.63 \pm 0.017}$	0.62 ± 0.015
4	0.72 ± 0.021	0.78 ± 0.015	0.77 ± 0.015	$\mathbf{0.79 \pm 0.018}$	0.76 ± 0.021
5	0.64 ± 0.018	0.7 ± 0.022	0.7 ± 0.011	$\mathbf{0.78 \pm 0.013}$	0.73 ± 0.0097
6	0.93 ± 0.01	0.7 ± 0.016	0.72 ± 0.01	$\mathbf{0.94 \pm 0.0089}$	0.68 ± 0.0056
Mean	0.7412	0.7388	0.7213	0.8068	0.7316

Table 1: Probability of correct identification for the six EEG data sets obtained by each of the four methods. The last row indicates the percentage of decrease in the classification error achieved by SOBDA compared to each one of the methods. $\pm$ range indicates one standard deviation for results of multiple cross-validation repetitions.

4. Rank-Selection

In our results, we selected the rank for the parameters $\mathbf{U}$ and $\mathbf{V}$ to be one (i.e., $R = 1$) and the rank for the parameter $\mathbf{A}$ to be two (i.e., $K = 2$). The selection of these parameters was motivated in Section 3.3. Specifically, in the current experimental paradigm, we are looking for one ERP components associated with the *readiness* potential, that is, the slow increase in amplitude before an actual hand movement. The search for a single component suggests setting $R = 1$, one spatio-temporal component. In the case of the second-order term involving the parameter $\mathbf{A}$ we set the $K = 2$ because we are interested in finding two components corresponding to the two different spatial profiles of the two classes. To validate our selection for these parameters, we preformed repeated cross-validation evaluation of our algorithm for different configurations of the parameters R and K. The parameter R was tested for values $\{1,2,3,4\}$ while parameter K was tested for $\{2,4\}$. The results of this evaluation are summarized in Figure 7. The Figure 7.a shows the mean cross-validation performance of the SOBDA algorithm across all real-EEG data sets for all configurations of the parameters R and K. It is evident from this figure that configuration $R = 1$, $K = 2$ corresponds to the best selection for these parameters on average for these data sets. The Figure 7.b shows the cross-validation performance of the SOBDA algorithm for each data set separately and for all configuration of the parameters R and K. The cross-validation procedure can be used to determine or validate the configuration of parameters R and K in cases were no prior knowledge is available about the signal of interest.

5. Conclusion

In this paper we presented a new method called Second-Order Bilinear Discriminant Analysis (SOBDA) for analyzing EEG signals on a single-trial basis. The method combines linear and quadratic features thus encompassing and extending a number of existing EEG analysis methods. We evaluated the SOBDA algorithm in both simulated and real human EEG data sets. We show a reduction in the classification error on human EEG when comparing our method to the state-of-the-art. The results on simulated data characterize the operational range of these algorithms in terms of SNR and shows that the proposed algorithm operates well where other methods fail. The parametrization

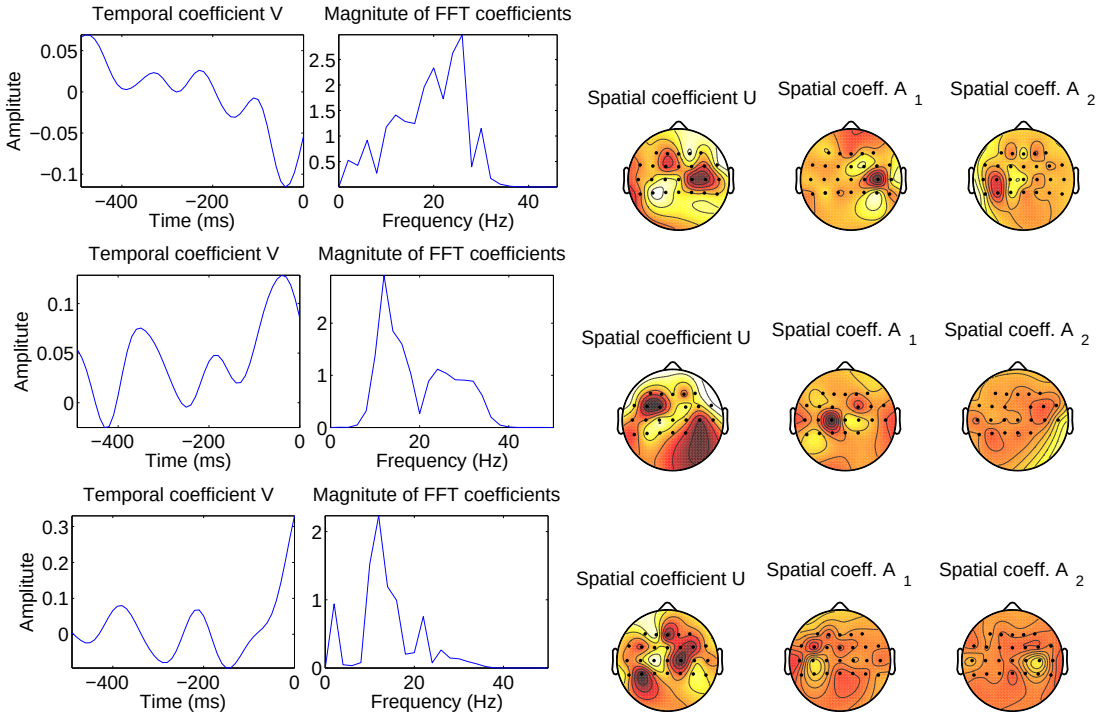

Figure 6: Extracted components in EEG for data sets 6, 4, and 3. *Left:* Temporal weights of linear component (first column) and and frequency weights of quadratic component (second column). *Right:* Spatial weights of linear component (third column) and two spatial weights for second-order spatial components (fourth and fifth column).

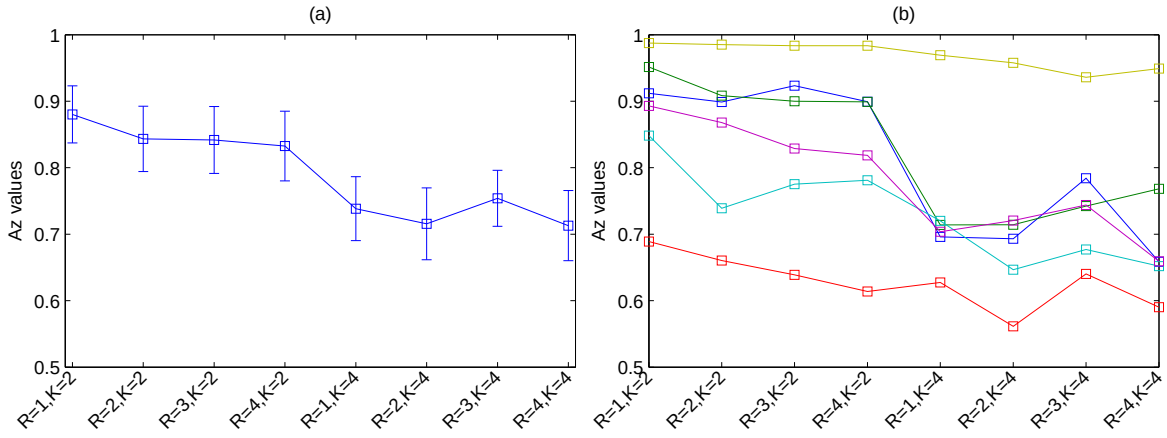

Figure 7: Cross-validation performance of SOBDA on the six real-EEG data sets used in the evaluation, at various configuration of the parameters R and K. (a) The mean cross-validation performance across data sets at various configuration of the parameters R and K.(b) Cross-validation performance for each of the data sets at various configuration of the parameters R and K.

of the discriminant criterion is intuitive, allowing one to incorporate prior knowledge as well as to derive spatial, temporal, and spectral information about the underlying neurological activity.

6. Derivations

In this section we derive the analytic gradient formulas of the negative log-likelihood function defined in (2). In general the gradient with respect to any of the variables can be expressed as:

$$\begin{aligned}\frac{\partial L(\theta)}{\partial \theta} &= -\sum_{n=1}^{N} \frac{\partial \log(1+e^{-y_n f(\mathbf{X}_n;\theta)})}{\partial \theta} \\ &= -\sum_{n=1}^{N} \frac{1}{1+e^{-y_n f(\mathbf{X};\theta)}} \frac{\partial\{1+e^{-y_n f(\mathbf{X}_n;\theta)}\}}{\partial \theta} \\ &= \sum_{n=1}^{N} y_n \frac{e^{-y_n f(\mathbf{X};\theta)}}{1+e^{-y_n f(\mathbf{X};\theta)}} \frac{\partial f(\mathbf{X}_n;\theta)}{\partial \theta},\end{aligned}$$

Now one has to take the specific derivatives with respect to each of the variables in θ is:

The gradient with respect to $\mathbf{u}_r$, the rth column of $\mathbf{U}$.

$$\begin{aligned}\frac{\partial\{f(\mathbf{X}_n;\theta)+w0\}}{\partial \mathbf{u}_r} &= C\frac{\partial\{\mathrm{Trace}\,\mathbf{U}^\top \mathbf{X}_n \mathbf{V}\}}{\partial \mathbf{u}_r} \\ &= C\frac{\partial\{\sum_{r'=1}^{R} \mathbf{u}_{r'}^\top \mathbf{X}_n \mathbf{v}_{r'}\}}{\partial \mathbf{u}_r} \\ &= C\mathbf{X}_n \mathbf{v}_r .\end{aligned}$$

The gradient with respect to $\mathbf{v}_r$, the *rth* column of $\mathbf{V}$ is:

$$\begin{aligned}\frac{\partial\{f(\mathbf{X}_n;\theta)+w0\}}{\partial \mathbf{v}_r} &= C\frac{\partial\{\mathrm{Trace}\,\mathbf{U}^\top \mathbf{X}_n \mathbf{V}\}}{\partial \mathbf{v}_r} \\ &= C\frac{\partial\{\sum_{r'=1}^{R} \mathbf{u}_{r'}^\top \mathbf{X}_n \mathbf{v}_{r'}\}}{\partial \mathbf{v}_r} \\ &= C\mathbf{u}_r^\top \mathbf{X}_n .\end{aligned}$$

The gradient with respect to $\mathbf{a}_r$, the *rth* column of $\mathbf{A}$ is:

$$\begin{aligned}\frac{\partial\{f(\mathbf{X}_n;\theta)+w0\}}{\partial \mathbf{a}_r} &= (1-C)\frac{\partial\{\mathrm{Trace}\,\mathbf{A}^\top(\mathbf{X}_n\mathbf{B})(\mathbf{X}_n\mathbf{B})^\top\mathbf{A}\}}{\partial \mathbf{a}_r} \\ &= (1-C)\frac{\partial\{\sum_{r'=1}^{K} \lambda_{r'}\mathbf{a}_{r'}^\top(\mathbf{X}_n\mathbf{B})(\mathbf{X}_n\mathbf{B})^\top\mathbf{a}_{r'}\}}{\partial \mathbf{a}_r} \\ &= 2(1-C)\lambda_r(\mathbf{X}_n\mathbf{B})(\mathbf{X}_n\mathbf{B})^\top\mathbf{a}_r ,\end{aligned}$$

The gradient with respect to $\mathbf{b}_r$, the *rth* column of $\mathbf{B}$ is:

$$\begin{aligned}
\frac{\partial\{f(\mathbf{X}_n;\theta)+w0\}}{\partial \mathbf{b}_r} &= (1-C)\frac{\partial\{\text{Trace}\,\Lambda\mathbf{A}^\top(\mathbf{X}_n\mathbf{B})(\mathbf{X}_n\mathbf{B})^\top\mathbf{A}\}}{\partial \mathbf{b}_r} \\
&= (1-C)\frac{\partial\{\text{Trace}\,\mathbf{B}^\top\mathbf{X}_n^\top\mathbf{A}\Lambda\mathbf{A}^\top\mathbf{X}_n\mathbf{B}\}}{\partial \mathbf{b}_r} \\
&= (1-C)\frac{\partial\{\sum_{r'=1}^{K}\mathbf{b}_{r'}^\top\mathbf{X}_n^\top\mathbf{A}^\top\Lambda\mathbf{A}^\top\mathbf{X}_n\mathbf{b}_{r'}\}}{\partial \mathbf{b}_r} \\
&= 2(1-C)(\mathbf{X}_n^\top\mathbf{A}\Lambda\mathbf{A}^\top\mathbf{X}_n)\mathbf{b}_r\,.
\end{aligned}$$

Acknowledgments

This work was funded by DARPA (contract #:NBCHCD80029).

References

N. Birbaumer, N. Ghanayim, T. Hinterberger, I. Iversen, B. Kotchoubey, A. Kubler, J. Perelmouter, E. Taub, and H. Flor. A spelling device for the paralysed. *Nature*, 398(6725):297–8, Mar FebruaryMay 1999.

B. Blankertz, G. Curio, and K. Müller. Classifying single trial EEG: Towards brain computer interfacing. In T. G. Diettrich, S. Becker, and Z. Ghahramani, editors, *Advances in Neural Information Processing Systems 14*. MIT Press, 2002., 2002.

B. Blankertz, G. Dornhege, C. Schfer, R. Krepki, J. Kohlmorgen, K. Müller, V. Kunzmann, F. Losch, and G. Curio. Boosting bit rates and error detection for the classification of fast-paced motor commands based on single-trial EEG analysis. *IEEE Trans. Neural Sys. Rehab. Eng.*, 11(2): 127–131, 2003.

B. Blankertz, K.-R. Müller, G. Curio, T.M. Vaughan, G. Schalk, J.R. Wolpaw, A. Schlogl, C. Neuper, G. Pfurtscheller, T. Hinterberger, M. Schroder, and N. Birbaumer. The bci competition 2003: progress and perspectives in detection and discrimination of EEG single trials. *Biomedical Engineering, IEEE Transactions on*, 51(6):1044–1051, 2004.

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

M. Dyrholm and L.C. Parra. Smooth bilinear classification of EEG. In *In Proc. 28th Annu. Int Conf. IEEE Engineering in Medicine and Biology Society*, 2006.

M. Dyrholm, C. Christoforou, and L.C. Parra. Bilinear discriminant component analysis. *J. Mach. Learn. Res.*, 8:1097–1111, 2007. ISSN 1533-7928.

C.A. Floudas. Deterministic global optimization in design, control, and computational chemistry. *In Proceedings: Large Scale Optimization with Applications. Part III: Optimal Design and Control, (L.T. Biegler, A. Conn, L. Coleman, and F. Santosa, Editors)*, pages 129–184, 1997.

A.D. Gerson, L.C. Parra, and P. Sajda. Cortically-coupled computer vision for rapid image search. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 14:174–179, June 2006.

MEGIS Software GmbH. BESA. http://www.besa.de/products/besa/, 2006.

S. Lemm, B. Blankertz, G. Curio, and K. Müller. Spatio-spectral filters for improving the classification of single trial EEG. *IEEE Trans Biomed Eng., 52(9):1541–8, 2005.*, 2005.

A. Luo and P. Sajda. Learning discrimination trajectories in EEG sensor space: Application to inferring task difficulty. *J. Neural Eng.*, 3:L1–L6, 2006a.

A. Luo and P. Sajda. Using single-trial EEG to estimate the timing of target onset during rapid serial visual presentation. In *Proc. Engineering in Medicine and Biology Society(EMBC2006)*, 2006b.

H. B. Nielsen. IMMOPTIBOX. General optimization software available at http://www.imm.dtu.dk/ hbn/immoptibox/, 2005.

L. Parra, C. Alvino, A. Tang, B. Pearlmutter, N. Young, A. Osman, and P. Sajda. Linear spatial integration for single-trial detection in encephalogra phy. *Neuroimage*, 17:223–230, 2002.

L.C. Parra, C.D. Spence, A.D Gerson, and P. Sajda. Recipes for the linear analysis of EEG. *Neuroimage*, 28(2):326–341, November 2005. ISSN 1053-8119.

L.C. Parra, C. Christoforou, A.D. Gerson, M. Dyrholm, A. Luo, M. Wagner, M.G. Philiastides, and P. Sajda. Spatiotemporal linear decoding of brain state: Application to performance augmentation in high-throughput tasks. *IEEE, Signal Processing Magazine*, January 2008.

W. D. Penny, N. J. Trujillo-Barreto, and K. J. Friston. Bayesian fMRI time series analysis with spatial priors. *NeuroImage*, 24:350362, 2005.

G. Pfurtscheller and F. H. Lopes da Silva. Event-related EEG/MEG synchronization and desynchronization: basic principles. *Clin Neurophysiol*, 110(11):1842–1857, November 1999. ISSN 1388-2457.

M.G. Philiastides and P. Sajda. Temporal characterization of the neural correlates of perceptual decision making in the human brain. *Cerebral Cortex*, 16(4), April 2006.

M.G. Philiastides, R. Ratcliff, and P. Sajda. Neural representation of task difficulty and decision making during perceptual categorization: A timing diagram. *Journal of Neuroscience*, 26(35): 8965–8975, August 2006.

H. Ramoser, J. Müller-Gerking, and G. Pfurtscheller. Optimal spatial filtering of single trial EEG during imagined hand movement. *IEEE Trans. Rehab. Eng.*, 8:441–446, December 2000. URL citeseer.ist.psu.edu/ramoser98optimal.html.

C.E. Rasmussen and C.K.I. Williams. *Gaussian Processes for Machine Learning (Adaptive Computation and Machine Learning)*. The MIT Press, 2005. ISBN 026218253X.

R. Tomioka and K. Aihara. Classifying matrices with a spectral regularization. In *ICML '07: Proceedings of the 24th international conference on Machine learning*, pages 895–902, New York, NY, USA, 2007. ACM. ISBN 978-1-59593-793-3. doi: http://www.ibis.t.u-tokyo.ac.jp/ryotat/lrds.

R. Tomioka, K. Aihara, and K. Müller. Logistic regression for single trial EEG classification. In B. Schölkopf, J. Platt, and T. Hoffman, editors, *Advances in Neural Information Processing Systems 19*, pages 1377–1384. MIT Press, Cambridge, MA, 2007.

J. R. Wolpaw, N. Birbaumer, D. J. McFarland, G. Pfurtscheller, and T. M. Vaughan. Brain-computer interfaces for communication and control. *Clin Neurophysiol*, 113(6):767–791, June 2002. ISSN 1388-2457.