

Single-trial analysis and classification of ERP components – A tutorial

Benjamin Blankertz^{a,b,*}, Steven Lemm^b, Matthias Treder^a, Stefan Haufe^a, Klaus-Robert Müller^a

^a Berlin Institute of Technology, Machine Learning Laboratory, Berlin, Germany

^b Fraunhofer FIRST, Intelligent Data Analysis Group, Berlin, Germany

ARTICLE INFO

Article history:

Received 15 December 2009

Revised 14 June 2010

Accepted 18 June 2010

Available online 28 June 2010

Keywords:

EEG
ERP
BCI
Decoding
Machine learning
Shrinkage
LDA
Spatial filter
Spatial pattern

ABSTRACT

Analyzing brain states that correspond to event related potentials (ERPs) on a single trial basis is a hard problem due to the high trial-to-trial variability and the unfavorable ratio between signal (ERP) and noise (artifacts and neural background activity). In this tutorial, we provide a comprehensive framework for decoding ERPs, elaborating on linear concepts, namely spatio-temporal patterns and filters as well as linear ERP classification. However, the bottleneck of these techniques is that they require an accurate covariance matrix estimation in high dimensional sensor spaces which is a highly intricate problem. As a remedy, we propose to use shrinkage estimators and show that appropriate regularization of linear discriminant analysis (LDA) by shrinkage yields excellent results for single-trial ERP classification that are far superior to classical LDA classification. Furthermore, we give practical hints on the interpretation of what classifiers learned from the data and demonstrate in particular that the trade-off between goodness-of-fit and model complexity in regularized LDA relates to a morphing between a difference *pattern* of ERPs and a spatial *filter* which cancels non task-related brain activity.

© 2010 Elsevier Inc. All rights reserved.

Introduction

Designated as one of the final frontiers of science, understanding brain function is a challenge that keeps attracting scientists from a multitude of disciplines. Early research efforts culminated in the emergence of computational neuroscience, the principal theoretical method for investigating the mechanisms of the nervous system. In particular, the interest in *modelling* single-trial behavior of the human brain has rapidly grown in the past decades. Nowadays the scope of modern neuroscience has been widened to *decoding* single-trial encephalogram data with respect to the identification of mental states or human intentions. This branch of research is strongly influenced by the development of an effective communication interface connecting the human brain and a computer (Dornhege et al., 2007; Kübler and Kotchoubey, 2007; Kübler and Müller, 2007; Wolpaw, 2007; Birbaumer, 2006; Pfurtscheller et al., 2005; Curran and Stokes, 2003; Wolpaw et al., 2002; Kübler et al., 2001), which finally also attracted the machine learning community to the field (Blankertz et al., 2002; Vidaurre and Blankertz, 2010; Hong et al., 2009; Müller et al., 2008; Blankertz et al., 2008b, 2007a, 2006; Parra et al., 2008, 2003; Wang et al., 2004; Tomioka and Müller, 2010). In this context the ability to perform single-trial classification of EEG data received much attention. But there is also interest from the basic research in single-

trial analysis of event-related potentials (ERPs), where mainly the question of trial-to-trial variability is addressed (e.g., Ratcliff et al., 2009).

Generally, the analysis of single-trial responses suffers from the superposition of task-relevant signals by task-unrelated brain activities, resulting in a low signal-to-noise ratio (SNR) of the observed single-trial responses. Here, in the context of single-trial classification of ERPs, we refer to the ERPs as the *signals* and to all non-phase-locked neural activity as well as to non-neural artifacts as interfering *noise*. Accordingly, the major goal of data processing prior to the classification of single-trial ERPs is to enhance their SNR significantly, in other words, isolating the phase-locked ERP *signal* from the interfering *noise*. To distinguish signals of interest from the interfering noise, different feature extraction methods have been applied, including temporal and spatial filters. Here, the most prevalent techniques are bandpass, notch or Laplace filters as well as principle component analysis (PCA) and more sophisticated techniques such as wavelet denoising (Quiroga and Garcia, 2003) and blind source separation (BSS) techniques (Cardoso and Souloumiac, 1993; Comon, 1994; Belouchrani et al., 1997; Makeig et al., 1997; Ziehe et al., 2000; Lemm et al., 2006). To these extracted features, different classification techniques have been applied, that can be either assigned to linear or non-linear methods. Among the non-linear methods the support vector machine is the most powerful method applied to ERP classification (Müller et al., 2001; Meinicke et al., 2003; Rakotomamonjy and Guigue, 2008). However, there is an ongoing debate whether the classification of single-trial EEG requires a non-linear

* Corresponding author. Berlin Institute of Technology, Machine Learning Laboratory, Berlin, Germany.

E-mail address: benjamin.blankertz@tu-berlin.de (B. Blankertz).

model or if a linear model is sufficient given an appropriate feature extraction (Müller et al., 2003). However, regardless of the particular techniques employed for feature extraction or classification, there is substantial variability in the classification accuracy both between subjects (Guger et al., 2003, 2009; Blankertz et al., 2007a; Krauledat et al., 2008; Dickhaus et al., 2009; Allison et al., 2009) and within subjects during the course of an experiment (Shenoy et al., 2006). It was shown in online studies that adaptation techniques can help to cope with the corresponding changes of the data distributions (Vidaurre et al., 2006, 2007; Vidaurre and Blankertz, 2010). Furthermore, there are other techniques that have been found promising in the same respect in offline studies, namely explicitly modeling the distribution change (cf. Sugiyama et al., 2007), restricting the feature space to the stationary part only (cf. von Büna et al., 2009), or enforcing invariance properties in the feature extract step (e.g. Blankertz et al., 2008a).

The rest of the paper is structured as follows. First, we introduce an EEG sample data set that is used throughout this paper for illustration and validation purpose and thereupon we define spatial, temporal, and spatio-temporal features. In *Spatial filters and spatial patterns* section, we introduce the concept of spatial patterns and filters within the framework of the linear EEG model and give a first argument on why an effective spatial filter will typically look much different from a pattern. Then, we discuss LDA and the plausibility of the assumptions underlying the optimality criterion in the context of EEG. Furthermore, we provide an illustrative simulation as another argument for the fundamental difference between spatial patterns and filters. After that, we introduce the important concept of regularization of the empirical covariance matrix by shrinkage and a method to determine the optimal shrinkage parameter. In *Classification of ERP components* section, the introduced concepts of machine learning are applied to one example data set to illustrate the interpretation of the classification method. An extensive validation of the proposed method on 13 data sets is provided in *Empirical evaluation* section including a comparison of the performance with state-of-the-art methods. Finally, we summarize the findings in a conclusion.

Example data set

We introduce an example EEG data set that we use throughout the paper to exemplify feature extraction and classification methods. The data set stems from a calibration recording for an attention-based typewriter. It provides a good show-case, because it comprises a sequence of ERP components that reflect different brain processes, related to visual processing of the physical stimulus properties as well as higher cognitive components associated with more abstract processing of the visual event. This data set will serve as a touchstone for the machine learning techniques presented in this paper. Furthermore, we will argue that these techniques can also be used as a tool to study the underlying visuo-cognitive processes.

Attention-based typewriting was introduced by Farwell and Donchin (1988). The so called Matrix Speller consists of a 6×6 symbol matrix wherein symbols are arranged within rows and columns. Throughout the course of a trial, the rows and columns are intensified ('flashed') one after the other in a random order. Since the neural processing of a stimulus can be modulated by attention, the ERP elicited by target intensifications is different from the ERP elicited by nontarget intensifications. Several improvements on the original approach have been suggested, see Martens et al. (2009), Hong et al. (2009), Salvaris and Sepulveda (2009), Krusienski et al. (2008), Li et al. (2006).

The present data is based on a novel variant of Hex-o-Spell (Treder and Blankertz, 2010), a mental typewriter which was originally controlled by means of motor imagery (Williamson et al., 2009; Blankertz et al., 2007b; Müller and Blankertz, 2006) while the new variant is based on ERPs. The Hex-o-Spell realizes selection of a target symbol in two recurrent steps. At the first level, the group containing

the target symbol is selected, followed by the selection of the target symbol itself at the second level. Groups of symbols (level 1) and single symbols (level 2) are arranged in six circles that are placed at the corners of an (invisible) hexagon, see Fig. 1. For the selection at each level, the circles are intensified in random order (10 blocks of intensifications of each of the 6 discs).

Intensification was realized by upsizing the disc including the symbol (s) by 62.5% for 100 ms with a stimulus onset asynchrony of 166 ms, see Fig. 1. Participants were instructed to fixate the target symbol (checked online with an eye tracker) while silently counting the number of intensifications of the target symbol. For a detailed comparison between Matrix Speller and the ERP-based Hex-o-Spell, in particular concerning the issue of eye gaze, see Treder and Blankertz (2010).

The data set used in this paper was recorded in an offline calibration mode, that means brain signals are not classified online and the application behaves as if all decisions are classified correctly. A text field at the top of the screen displays the words that are to be 'written' with the current symbol shown in red and larger font size, see Fig. 1. After the first level stimulation is finished, the second level starts automatically with the correct subgroup. Such calibration is recorded in order to train the classifier and optimize parameters.

From the classification point of view, the task can be reduced to the discrimination of brain responses to target stimuli (intensifications of discs that the user draws attention to) and nontarget stimuli. If this binary discrimination problem can be solved, the system can obviously determine the intended letter. Single-trial classification of these brain responses in most users lacks accuracy. Ten repetitions of stimulus sequences are performed in the calibration, which allows to investigate the optimal number of repetitions for the use in subsequent online experiments.

ERPs for the Hex-o-Spell data set in one particular participant¹ are shown in Fig. 2. Spatial distributions are shown for several subcomponents, determined from those time intervals that are shaded in the ERP curve on top. As the figure suggests, some components are related to the processing of the physical properties of the stimulus in primary visual areas in occipital cortex (the according component labels are tagged with the prefix 'v' for visual). Other components reflect more abstract processes of attentional gating and event detection and they are characterized by a more central topographic distribution.

Since all of these components can be modulated by attention and eye fixation, all intervals are potentially informative regarding the classification task. However, the fact that trial-to-trial variability of ERP amplitudes is extremely high in both target and nontarget classes (not shown in the graph) renders the classification problem a hard one. It requires the application of sophisticated machine learning methods, as introduced next.

Features of ERP classification

ERP components are characterized by their temporal evolution and the corresponding spatial potential distributions. As raw material, we have the spatio-temporal data matrix² $\mathbf{X}^{(k)} \in \mathbb{R}^{M \times T}$ for each trial k with M being the number of channels and T the number of sampled time points. For the formalism in classification, features are considered to be column vectors, which are obtained by concatenation of all columns of $\mathbf{X}^{(k)}$ into $\mathbf{x} \in \mathbb{R}^{M \cdot T}$. Time is typically sampled in equidistant intervals, but it might be beneficial to select specific intervals of

¹ For illustration purpose here and in *Linear classification* and *Classification of ERP components* sections, we use single-subject data from Hex-o-Spell data set, while the validation in *Empirical evaluation* section is performed on data from all 13 participants and for both typewriter types, Hex-o-Spell and the Matrix Speller.

² Throughout this manuscript, italic characters refer to scalars such as t and M ; a bold lower case character as in \mathbf{a} denotes a vector; matrices are indicated by uppercase bold characters, such as \mathbf{A} ; while the transpose of a vector and a matrix is consistently denoted by \mathbf{a}^T and \mathbf{A}^T , respectively.

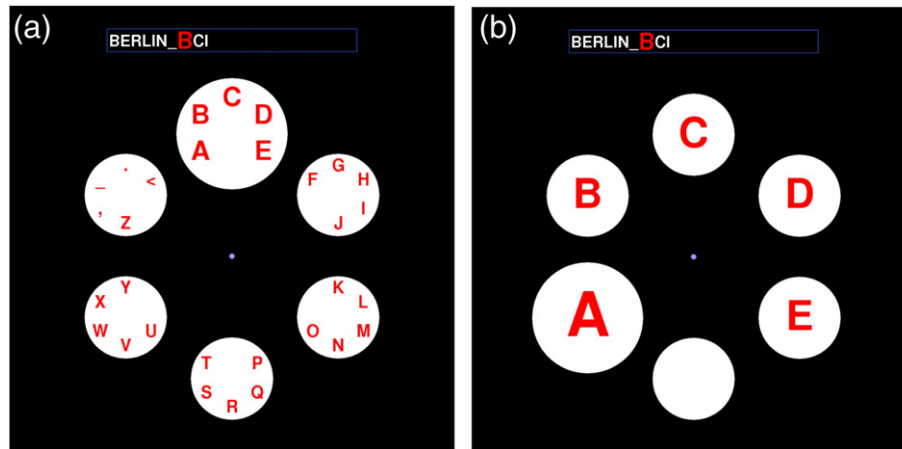


Fig. 1. Experimental design: Symbols can be selected in the mental typewriter Hex-o-Spell in a two level procedure. (a) At the first level, a disc containing a group of 5 symbols is selected. (b) For the second level, the symbols of the selected group are distributed to all discs (animated transition). The empty disc can be used as 'undo' to return to the group level without selection. Selection of the backspace symbol '<' allows to erase the last written symbol.

interest corresponding to subcomponents or an ERP complex, see [Classification in the spatio-temporal domain](#) section. Therefore, we give the following formal definition of spatio-temporal features. We denote the scalp potential at channel c and time point t within trial k by $x_c^{(k)}(t)$ (superscript k will be omitted). Given a subset of channels $C = \{c_1, \dots, c_M\}$ we define $\mathbf{x}_C(t) = [x_{c_1}(t), \dots, x_{c_M}(t)]^T$ as the vector of potential values for the subset of channels at time point t .

By concatenation of those vectors for all time points t_1, \dots, t_T of one trial one obtains spatio-temporal features

$$\mathbf{X}(C) = [\mathbf{x}_C(t_1), \dots, \mathbf{x}_C(t_T)]. \quad (1)$$

It can be helpful for classification to reduce the dimensionality of those features by averaging across time in certain intervals. These can

be equally spaced (subsampling), or specifically chosen, preferably such that each interval contains one component of the ERP in which the spatial distribution is approximately constant, see [Classification in the spatio-temporal domain](#) section. Therefore, we use the following more general definition. Let $\mathcal{T}_1, \dots, \mathcal{T}_I$ sets of time points (each \mathcal{T}_i typically being an interval). Given selected channels C and time intervals $\mathcal{T} = \langle \mathcal{T}_i \rangle_{i=1, \dots, I}$, we define the *spatio-temporal* feature as

$$\mathbf{X}(C, \mathcal{T}) = [\text{mean}(\mathbf{x}_C(t))_{t \in \mathcal{T}_1}, \dots, \text{mean}(\mathbf{x}_C(t))_{t \in \mathcal{T}_I}]. \quad (2)$$

For the case that there is only one time interval (i.e., $I = 1$), we call the features (purely) *spatial* (but it relates to the temporal domain). In this case, the dimensionality of the feature coincides with the number of channels. The feature of a single trial is the average scalp potential

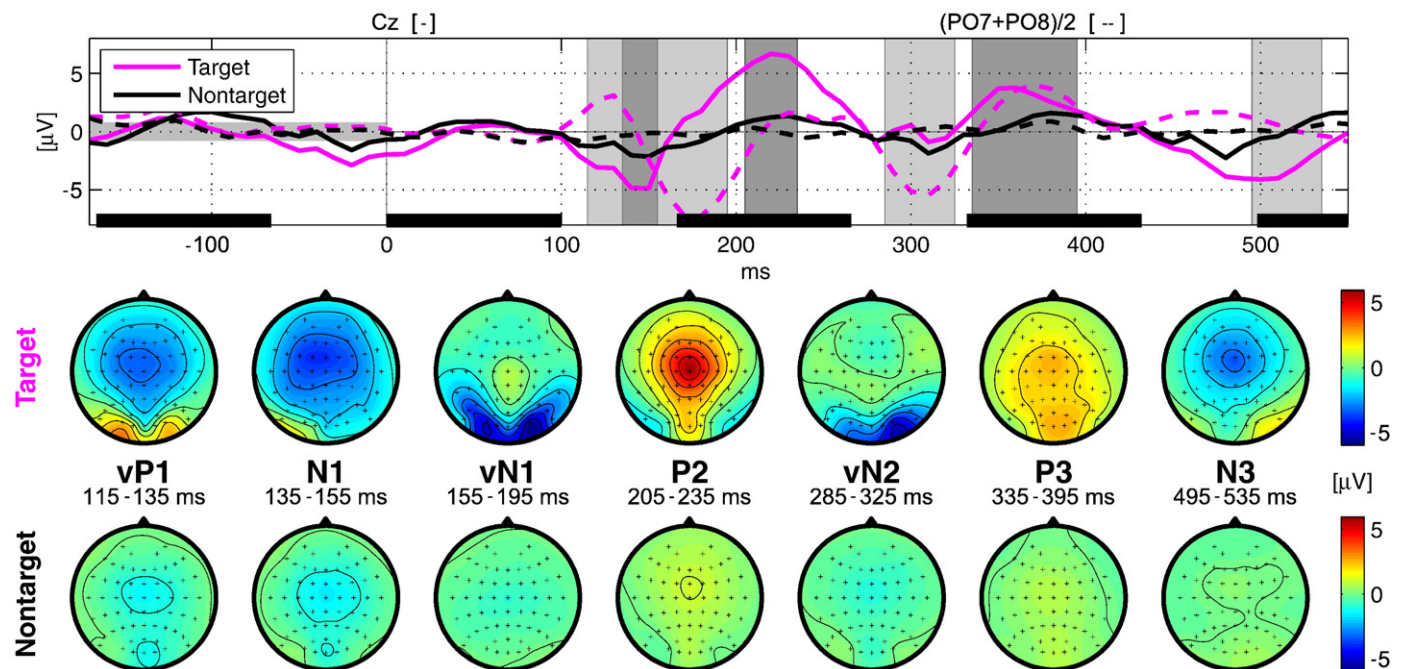


Fig. 2. ERPs in the attention based typewriter. Every 166 ms one disc of the typewriter was intensified for 100 ms (indicated by solid bars at the bottom of the time course plot). According to whether or not it was the to-be-attended disc, trials were labelled as targets or nontargets, respectively. Brain responses to target stimuli are characterized by a sequence of different ERP components. The graph on top shows the time course of the ERPs at channel Cz and the average of channels PO7 and PO8. The baseline was taken -166 to 0 ms prestimulus indicated by the horizontal gray bar around the x-axis. The trial-to-trial standard deviation is between 6.7 and $9 \mu\text{V}$ at channel Cz and between 4.7 and $6.5 \mu\text{V}$ at channel $(\text{PO7} + \text{PO8})/2$. The two rows below display scalp topographies for the two experimental conditions at the time intervals that are shaded in the ERP plot. The maps show a top view on the head with nose up where the averaged scalp potentials obtained for each electrode are spatially interpolated and coded in color.

within the given time interval (at all given channels). Otherwise, when \mathcal{C} is a singleton $\{c_1\}$ (i.e., $M = 1$), we say \mathbf{X} is a *temporal feature* (at channel c_1). In this case, the feature of a single trial is the ‘time course’ of scalp potentials at the given channel, sampled at time intervals $\mathcal{T}_1, \dots, \mathcal{T}_T$.

To discuss the distribution of ERP features, we consider a simplified model. We model each single trial $\mathbf{x}^{(k)}(t)$ as the superposition of a (phase-locked) event-related source $\mathbf{s}(t)$ (= ERP) which is constant in every single trial and ‘noise’ (non phase-locked activity) $\mathbf{n}^{(k)}(t)$, which is assumed to be identically independent distributed (iid) according to a Gaussian distribution $\mathcal{N}(0, \Sigma_n)$ (for a fixed t).

$$\mathbf{x}^{(k)}(t) = \mathbf{s}(t) + \mathbf{n}^{(k)}(t) \quad \text{for all trials } k = 1, \dots, K \quad (3)$$

Under this model assumption, the spatial feature, say, at time point t_0 is Gaussian distributed according to $\mathcal{N}(\mathbf{s}(t_0), \Sigma_n)$. This means that ERP features are Gaussian distributed with the mean being the ERP pattern and the covariance being the covariance of ongoing background EEG, see Fig. 3.

The assumption of the ERP being the same within each trial is known to be invalid. E.g., amplitude and latency of the P3 component depend on many factors that vary from trial to trial, such as time since last target and attention. In this case, the covariance of the data distribution comprises also the trial-to-trial variation of the ERP. Nevertheless, these variations in the ERP are typically small compared to the background noise, such that Eq. (3) still holds well enough to provide a useful model.

Spatial filters and spatial patterns

The basic macroscopic model of EEG generation (Nunez and Srinivasan, 2005) assumes the tissue to be a resistive medium and hence only considers effects of volume conduction, while neglecting the marginal capacitive effects (Stinstra and Peters, 1998). Subject to these prerequisites, a single current source $\mathbf{s}(t)$ contributes linearly to the scalp potential $\mathbf{x}(t)$, i.e.,

$$\mathbf{x}(t) = \mathbf{A}\mathbf{s}(t), \quad (4)$$

where the propagation vector $\mathbf{a} \in \mathbb{R}^M$ represents the individual coupling strengths of the source \mathbf{s} to the M surface electrodes. In general, the

propagation vector \mathbf{a} depends on three factors; the conductivity of the intermediary layers (brain tissue, skull, skin); the spatial location and orientation of the current source within the brain; and the impedances and locations of the scalp electrodes. In order to model the contribution of multiple source signals $\mathbf{s}(t) = (s_1(t), s_2(t), \dots)^T$ to the surface potential, the propagation vectors of the individual sources are aggregated into a matrix $\mathbf{A} = [\mathbf{a}_1, \mathbf{a}_2, \dots]$ and the overall surface potential results in

$$\mathbf{x}(t) = \mathbf{A}\mathbf{s}(t) + \mathbf{n}(t). \quad (5)$$

This model also incorporates an additive term $\mathbf{n}(t)$, which comprises any contribution not described by the matrix \mathbf{A} . Although some of the originating sources might be of neocortical origin, $\mathbf{n}(t)$ is conventionally conceived as *noise*, emphasizing that those activities are not subject of the investigation. The propagation matrix \mathbf{A} is often called the *forward model*, as it relates the source activities to the signals acquired at the different sensors. In this regard, the propagation vector \mathbf{a} of a source \mathbf{s} is often referred to as the *spatial pattern* of \mathbf{s} , and can be visualized by means of a scalp map.

The reverse process of relating the sensor activities to originating sources, is called *backward modeling* and aims at computing a linear estimate of the source activity from the observed sensor activities:

$$\hat{\mathbf{s}}(t) = \mathbf{W}^T \mathbf{x}(t). \quad (6)$$

A source $\hat{\mathbf{s}}$ is therefore obtained as a linear combination of the spatially distributed information from the multiple sensors, i.e., $\hat{\mathbf{s}}(t) = \mathbf{w}^T \mathbf{x}(t)$. In general, the estimation of the *backward model* corresponds to a reconstruction of the sources by means of inverting the *forward model*. In case of a non-invertible matrix \mathbf{A} a solution can be obtained only approximatively. For example, the recovering of the sources given by the forward model \mathbf{A} by means of a least mean squares estimator, that is

$$\mathbf{W} = \arg \min_{\mathbf{W}} \sum_t \|\mathbf{V}^T \mathbf{A}\mathbf{s}(t) - \mathbf{s}(t)\|^2, \quad (7)$$

leads to a solution of the pseudo-inverse of \mathbf{A} , i.e.,

$$\hat{\mathbf{W}}^T = (\mathbf{A}^T \mathbf{A})^{-1} \mathbf{A}^T. \quad (8)$$

However, generally the forward model is not known and the goal of backward modelling is to find a suitable linear transformation \mathbf{W}^T of the data which leads to an improved signal to noise ratio (SNR) of signals of interest. Accordingly, the rows \mathbf{w}^T of the matrix \mathbf{W}^T are commonly referred to as *spatial filters*. Although the spatial filters \mathbf{w}^T can be visualized by means of scalp maps, their interpretation is not as straightforward as for the spatial patterns, cf. also the discussion in [Understanding linear classifiers](#) section. To see this, consider the following simplified noise free example of two sources \mathbf{s}_1 and \mathbf{s}_2 given with their corresponding propagation vectors \mathbf{a}_1 and \mathbf{a}_2 , respectively. The task is to recover the source \mathbf{s}_1 from the observed mixture $\mathbf{x} = \mathbf{a}_1 \mathbf{s}_1 + \mathbf{a}_2 \mathbf{s}_2$. In general, a linear filter \mathbf{w}^T yields $\mathbf{w}^T \mathbf{x} = \mathbf{w}^T \mathbf{a}_1 \mathbf{s}_1 + \mathbf{w}^T \mathbf{a}_2 \mathbf{s}_2$. If we suppose the two propagation vectors to be orthogonal, i.e., $\mathbf{a}_1^T \mathbf{a}_2 = 0$, it follows immediately that the best linear filter is given by $\mathbf{w}^T = \mathbf{a}_1^T$ and \mathbf{s}_1 can be perfectly recovered (up to scaling) by $\hat{\mathbf{s}}_1 = \|\mathbf{a}_1\|^2 \mathbf{s}_1$. Hence in case of orthogonal sources the best *filter* corresponds to the propagation direction of the source, i.e., a *pattern*. However, changing the setting slightly and assuming non-orthogonal propagation vectors, i.e., $\mathbf{a}_1^T \mathbf{a}_2 \neq 0$, the signal along the direction \mathbf{a}_1 also consists of a portion of \mathbf{s}_2 . In order to obtain the optimal filter to recover \mathbf{s}_1 , the filter has to be orthogonal to the interfering source \mathbf{s}_2 , i.e., $\mathbf{w}^T \mathbf{a}_2 = 0$, while having a positive inner product $\mathbf{w}^T \mathbf{a}_1 > 0$. Thus the optimal filter has to take the propagation vectors of both sources into account and is given by the first row of the pseudo-inverse of $[\mathbf{a}_1 \mathbf{a}_2]$. Consequently, the spatial filters depend on the scalp distributions not just of the reconstructed source, but also on the

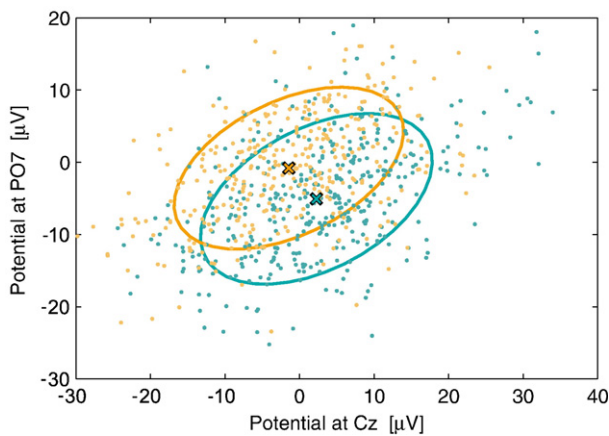


Fig. 3. Distribution of ERP Features. Scalp potentials of each single trial have been averaged across time within the interval of the visual N1 component (155–195 ms). The dots show the distribution of those values for channels Cz and PO7 separately for the two conditions ‘target’ (cyan) and ‘nontarget’ (orange). Means of the two distributions are marked by crosses. Classwise covariance matrices have been used to indicate a fitted Gaussian distribution by equidensity contours at 1.5 times standard deviation. Note, that the covariance of both conditions is very similar.

distribution of the other sources. Moreover, the signal that is recovered by a spatial filter \mathbf{w}^T also captures the portion of the noise that is collinear with the source estimate: $\hat{\mathbf{s}}(t) = \mathbf{s}(t) + \mathbf{w}^T \mathbf{n}(t)$. Hence a spatial filter which optimizes the SNR of a signal of interest must be approximately orthogonal against interfering sources and noise signals.

For a comprehensive review on spatial filters and linear analysis of EEG we refer to Parra et al. (2005) and for a tutorial on optimal spatial filters for features based on oscillatory brain activity (cf. Blankertz et al., 2008c).

Linear classification

In this paper, we demonstrate how a basic classification algorithm, Linear Discriminant Analysis (LDA), can become a powerful tool for the classification of ERP components when endowed with a technique called shrinkage for the use with high dimensional features. This technique is simple to implement, computationally cheap, easy to apply, and yet—to our experience—gives impressive results that are at least on the same level with state-of-the-art classification methods that are more complex, see Empirical evaluation section. Of course, the applicability of LDA with shrinkage is by no means restricted to classification of ERP components but is a general technique for linear classification.

Linear discriminant analysis

For known Gaussian distributions with the same covariance matrix for all classes, it can be shown that Linear Discriminant Analysis (LDA) is the optimal classifier in the sense that it minimizes the risk of misclassification for new samples drawn from the same distributions (Duda et al., 2001). Note that LDA is equivalent to Fisher Discriminant Analysis and Least Squares Regression (Duda et al., 2001). The optimality criterion relies on three assumptions:

- (1) *Features of each class are Gaussian distributed.* Due to our experience, features of ERP data satisfy this condition quite well, see also Features of ERP classification section, and the method is quite robust to deviations from the normality assumption. For other type of features one can often find a preprocessing to approximately achieve Gaussian distributions,

see Dornhege et al. (2004), Lemm et al. (2011). But, e.g., inherently multi modal data would require a particular preprocessing or classification.

- (2) *Gaussian distributions of all classes have the same covariance matrix.* This assumption implies the linear separability of the data. It is approximately satisfied for many ERP data sets as the ongoing activity is typically independent of the different conditions under investigation, see also Fig. 3 and the discussion regarding distributions of ERP features in Features of ERP classification section. But this is not necessarily so. When comparing ERPs related to different visual cues, the visual alpha rhythm may be modulated by each type of cue in a different way. Fortunately, LDA is quite robust in cases of different covariance matrices. On the other hand, modelling a separate covariance matrix for each class leads to a nonlinear separation (quadratic discriminant analyses, for QDA, see Duda et al., 2001; Friedman, 1989; Vidaurre et al., 2006) and for kernel based learning see (Müller et al., 2001) that is much more sensible to errors in the estimation of the covariance matrices and therefore often yields inferior results to LDA unless a large number of training samples is available.

In the current data set, we verified the assumption by inspecting scalp topographies of the principal components corresponding to the largest eigenvalues of the covariance matrices, see Fig. 4.

- (3) *True class distributions are known.* This assumption is obviously never fulfilled in any real application. Means and covariance matrices of the distributions have to be estimated from the data. Due to the inevitable errors in those estimates the optimality statement might not hold at all, even when the Assumptions (1) and (2) are met quite well. This is typically the case, when the number of training samples is low compared to the dimensionality of the features. For this case, we will see later that shrinkage estimators are an excellent option to apply, see Shrinkage for improved classification section.

It is our firm belief, that it is vital to comprehend all aspects of the classification, and thus to be able to follow all steps of the process in the analysis for a particular data set, rather than to treat the method as a black box. To this end, we start by discussing linear classification interpreting what was 'learned' by the classifier. In this respect, it is

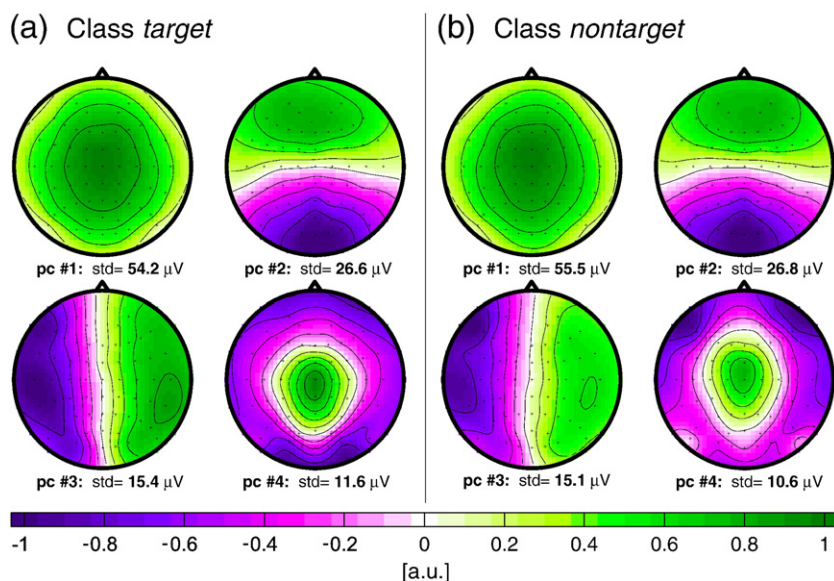


Fig. 4. Verification of the assumption of equal covariance matrices. An eigenvalue decomposition was performed for the covariance matrices of the two conditions *target* and *nontarget* (for spatial features at 220 ms post stimulus, but results for other time points/intervals look similar). The eigenvectors (or principal components) corresponding to the four largest eigenvalues are visualized as scalp topographies (as the scaling is arbitrary, the maximum absolute value is chosen to be 1 and the scale of the colormap ranges from -1 to 1 [arbitrary unit]). The similarity of the corresponding maps for the two classes can be seen as an indication that the covariance matrices of the two classes can be assumed to be equal.

important to remember the notions of filters and patterns, see [Spatial filters and spatial patterns](#) section. For illustration purpose, we utilize so-called spatial features, that were introduced in [Features of ERP classification](#) section.

Understanding linear classifiers

Binary linear classifiers can be characterized by a projection vector \mathbf{w} and a bias term b referring to the separating hyperplane $\mathbf{w}^\top \mathbf{x} + b = 0$. The classification function assigns a given input $\mathbf{x} \in \mathbb{R}^N$ the class label according to $\text{sign}(\mathbf{w}^\top \mathbf{x} + b)$. The projection vector of LDA is defined as

$$\mathbf{w} = \hat{\Sigma}_c^{-1}(\hat{\mu}_2 - \hat{\mu}_1) \quad (9)$$

where μ_i is the estimated mean of the class i and $\hat{\Sigma}_c = \frac{1}{2}(\hat{\Sigma}_1 + \hat{\Sigma}_2)$ is the estimated common covariance matrix, i.e., the average of the class-wise empirical covariance matrices $\hat{\Sigma}_i$, see Eq. (11) below.

In the following, we consider applying LDA to spatial features (see [Features of ERP classification](#) section) and provide an insightful interpretation in terms of spatial filters and patterns and a good visualization. But the principles apply to other types of features, too.

A linear classifier that was trained on *spatial features* can also be regarded as a *spatial filter* (cf. [Spatial filters and spatial patterns](#) section). Thus, the linear classifier may be interpreted as a “backward model” to recover the signal of discriminative sources, cf. also [Parra et al. \(2003\)](#). Let $\mathbf{w} \in \mathbb{R}^M$ be the weight vector and $\mathbf{Y} \in \mathbb{R}^{M \times \text{time points}}$ be continuous EEG signals. Then

$$\mathbf{Y}_f := \mathbf{w}^\top \mathbf{Y} \in \mathbb{R}^{1 \times \text{time points}} \quad (10)$$

is the result of spatial filtering: each channel of \mathbf{Y} is weighted with the corresponding component of \mathbf{w} and summed up. The weight vector \mathbf{w} of the classifier can be displayed as scalp map, but its interpretation requires some background knowledge, see also the discussion in [Spatial filters and spatial patterns](#) section. Taking the interval of the P2 component (205–235 ms) of our example data set, the difference pattern of the two conditions (target stimulus minus nontarget stimulus) displays a typical pattern with broad central focus, see right topography of [Fig. 3](#). In contrast, the classifier resp. spatial filter has an intricate and more complex structure that has also weights of substantial magnitude, e.g., in the occipital region which does not contribute to the P2 component itself, see left topography of [Fig. 5](#). Note, that the difference pattern in [Fig. 5](#) (right) corresponds to the LDA classifier under the assumption that the class covariance matrices are spherical, i.e. $\hat{\Sigma}^{-1} \sim \mathbf{I}$, with \mathbf{I} being the identity matrix.

Clearly this non-smoothness of the filter topography seen in [Fig. 5](#) (left) might evoke disbelief in researchers that are used to look at

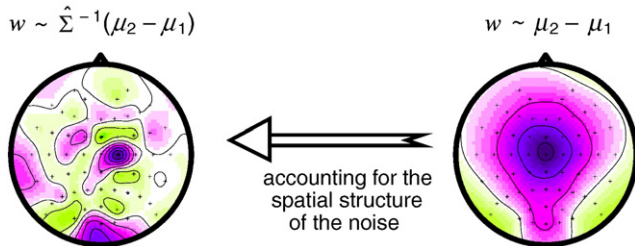


Fig. 5. Weight vector of a linear classifier displayed as scalp topography. When a linear classifier is trained on spatial features (see [Features of ERP classification](#) section), the obtained weight vector can coherently be displayed as scalp topography. In this example, a classifier was trained on the spatial feature extracted from the time interval 205–235 ms. The topography on the left displays the spatial filter calculated with LDA. The right maps shows the “filter” that was obtained under the assumption that the common covariance matrix is spherical. Note, that in this particular case the filter is the pattern of the difference of the two ERPs.

patterns only. Therefore, we will give an illustrative simulation to point out, that indeed the optimal filter *typically* requires a more intricate structure and may have substantial weight on electrodes that do not provide good discriminability when looking at [Fig. 5](#) (right). Note, that these channels can contribute to a reduction of noise in the informative channels. To illustrate this argument, we use artificially generated data. Two dimensional data was drawn from two Gaussian distributions simulated potentials recorded at two scalp electrodes FCz and CPz. Both channels thus contribute to the discrimination of the two classes, see [Fig. 6](#) (a). Classification of this 2D data yields 15% error. Now we add one source of disturbance, namely occipital alpha rhythm which is not related to the classification task. Simulating volume conduction, we add simulated alpha activity with a factor of 0.4 to CPz and with factor 0.2 to FCz, see [Fig. 6](#) (b). Since this introduces noise to the classification task, the error rises to 37% for classification on channels FCz and CPz on these disturbed signals, see [Fig. 6](#) (c). But when we add the third channel Oz to the feature vector, the original classification accuracy can be re-established on the 3D data (to be accurate, it was 16% in our simulation). Although channel Oz itself is not informative for the discrimination of the two classes, it is required for good classification. This demonstrates why a classifier on the three dimensional data will have nonzero weight on channel Oz, which is itself not discriminative.

Shrinkage for improved classification

The standard estimator for a covariance matrix is the empirical covariance (see Eq. (11) below). This estimator is unbiased and has good properties under usual conditions. But for high-dimensional data with only a few data points given, the estimation may become imprecise, because the number of unknown parameters that have to be estimated is quadratic in the number of dimensions.

This leads to a systematic error: Large eigenvalues of the original covariance matrix are estimated too large, and small eigenvalues are estimated too small; see [Fig. 7](#). This error in the estimation degrades classification performance (and renders LDA far from being optimal). Shrinkage is a common remedy for compensating the systematic bias (Stein, 1956) of the estimated covariance matrices (e.g., [Friedman, 1989](#)):

Let $\mathbf{x}_1, \dots, \mathbf{x}_n \in \mathbb{R}^d$ be n feature vectors and let

$$\hat{\mu} = \frac{1}{n} \sum_{i=1}^n \mathbf{x}_i \quad \text{and} \quad \hat{\Sigma} = \frac{1}{n-1} \sum_{i=1}^n (\mathbf{x}_i - \hat{\mu})(\mathbf{x}_i - \hat{\mu})^\top \quad (11)$$

be the unbiased estimator of the mean and the covariance matrix (empirical covariance matrix). In order to counterbalance the estimation error, $\hat{\Sigma}$ is replaced by

$$\tilde{\Sigma}(\gamma) := (1-\gamma)\hat{\Sigma} + \gamma\nu\mathbf{I} \quad (12)$$

for a tuning parameter $\gamma \in [0, 1]$ and ν defined as average eigenvalue $\text{trace}(\hat{\Sigma})/d$ of $\hat{\Sigma}$ with d being the dimensionality of the feature space. Then the following holds. Since $\hat{\Sigma}$ is positive semi-definite we have an eigenvalue decomposition $\hat{\Sigma} = \mathbf{V}\mathbf{D}\mathbf{V}^\top$ with orthonormal \mathbf{V} and diagonal \mathbf{D} . Due to the orthogonality of \mathbf{V} we get

$$\tilde{\Sigma} = (1-\gamma)\mathbf{V}\mathbf{D}\mathbf{V}^\top + \gamma\nu\mathbf{I} = (1-\gamma)\mathbf{V}\mathbf{D}\mathbf{V}^\top + \gamma\nu\mathbf{V}\mathbf{I}\mathbf{V}^\top = \mathbf{V}((1-\gamma)\mathbf{D} + \gamma\nu\mathbf{I})\mathbf{V}^\top$$

as eigenvalue decomposition of $\tilde{\Sigma}$. That means

- $\tilde{\Sigma}$ and $\hat{\Sigma}$ have the same eigenvectors (columns of \mathbf{V})
- extreme eigenvalues (large or small) are modified (shrunk or elongated) towards the average ν .
- $\gamma=0$ yields unregularized LDA, $\gamma=1$ assumes spherical covariance matrices.

Using LDA with such modified covariance matrix is termed covariance-regularized LDA, regularized LDA or LDA with shrinkage.

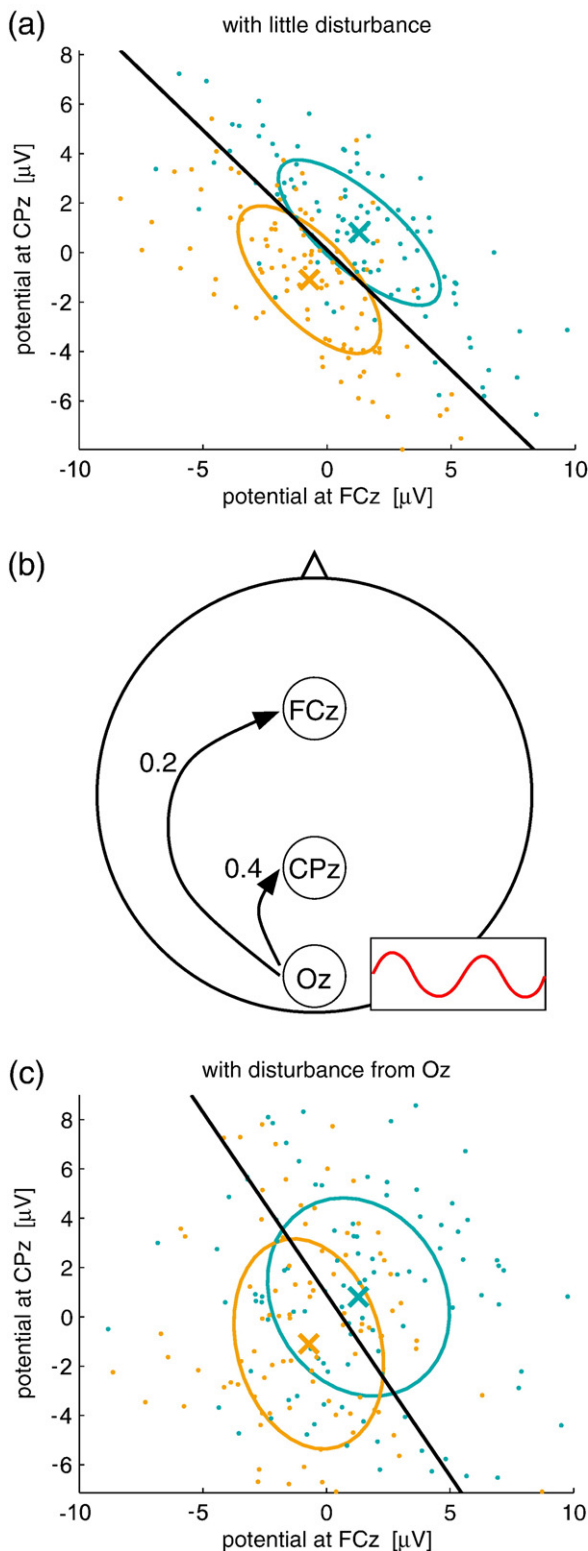


Fig. 6. Understanding spatial filters. (a) Two dimensional Gaussian distributions were generated to simulate scalp potentials at two electrodes with relative high signal-to-noise ratio. (b) A disturbing signal is generated: simulated visual alpha at channel Oz, which is propagated to channels CPz and FCz. (c) This results in a substantially decreased separability in the original two dimensional space. However when classifying 3D data that includes data from sensor Oz, it becomes possible to subtract the noise for the informative channels CPz and FCz and classification becomes possible with the same accuracy as for 'undisturbed data' in (a).

For the trade-off with respect to the shrinkage parameter, recently an analytic method to calculate the optimal shrinkage parameter for certain directions of shrinkage was found (Ledoit and Wolf, 2004; see also Vidaurre et al., 2009 for the first application in BCI). This approach aims at minimizing the Frobenius norm between the shrunk covariance matrix and the *unknown* true covariance matrix Σ . In effect, large sample-to-sample variance of entries in the empirical covariance are penalized, i.e., lead to stronger shrinkage. When we denote by $(\mathbf{x}_k)_i$ resp. $(\hat{\mu})_i$ the i -th element of the vector \mathbf{x}_k resp. $\hat{\mu}$ and denote by s_{ij} the element in the i -th row and j -th column of $\hat{\Sigma}$ and define

$$z_{ij}(k) = ((\mathbf{x}_k)_i - (\hat{\mu})_i)((\mathbf{x}_k)_j - (\hat{\mu})_j),$$

then the optimal parameter for shrinkage towards identity (as defined by Eq. (12)) can be calculated as Schäfer and Strimmer, 2005b)

$$\gamma^* = \frac{n}{(n-1)^2} \frac{\sum_{i,j=1}^d \text{var}_k(z_{ij}(k))}{\sum_{i \neq j} s_{ij}^2 + \sum_i (s_{ii} - v)^2}. \quad (13)$$

The usage of Eq. (13) for the selection of the shrinkage parameter does not necessarily lead to better classification results, compared to a selection based on, e.g., cross-validation, but is easier to implement and computationally much cheaper. Using regularized LDA for retraining the classifier after each trial during online operation as in Vidaurre and Blankertz (2010) would hardly be possible without the analytical solution.

With respect to classification of ERP features, we know from Fig. 5 that the shrinkage parameter γ controls the weight vector of the corresponding classifier between the two extreme poles of spatial pattern (difference between ERPs) for $\gamma = 1$ and a spatial filter that incorporates the spatial structure of the noise for $\gamma = 0$. The latter case is optimal, if the covariance of the noise is known. The empirical estimates that can be calculated from training data are error prone, in particular for high-dimensional feature spaces. In this case, the automatic selection of the shrinkage parameter provides a good trade-off of using the spatial structure of the noise without overfitting to the particular statistics found in the training set.

The method of regularized LDA with shrinkage parameter selected by the analytical solution of Schäfer and Strimmer (2005b) is called shrinkage LDA.

Classification of ERP components

We start by exploring ERP classification separately in the temporal and in the spatial domain. The purpose of classification on temporal features is to determine which channels contribute most to the discrimination task. And classification on spatial features demonstrates which time intervals are most important. Taken together, this investigation provides a good idea of which components of the EEG is exploited by the classifier, and gives a better understanding of the data and the classification process. For the actual classification task, the full information of spatio-temporal features should be exploited, as demonstrated in Classification in the spatio-temporal domain section.

Classification in the temporal domain

ERPs exhibit a particular time course at electrodes over its generators. The temporal features calculated from single channels have been introduced in Features of ERP classification section.

This single channel data does (in most cases) not contain sufficient information for a competitive classification. Nevertheless, we can gain useful information in the following way. For each single channel the classification accuracy that can be obtained from temporal features with ordinary LDA is determined (e.g. by cross validation, cf. Lemm et al., 2011). The resulting accuracy values can be visualized as scalp

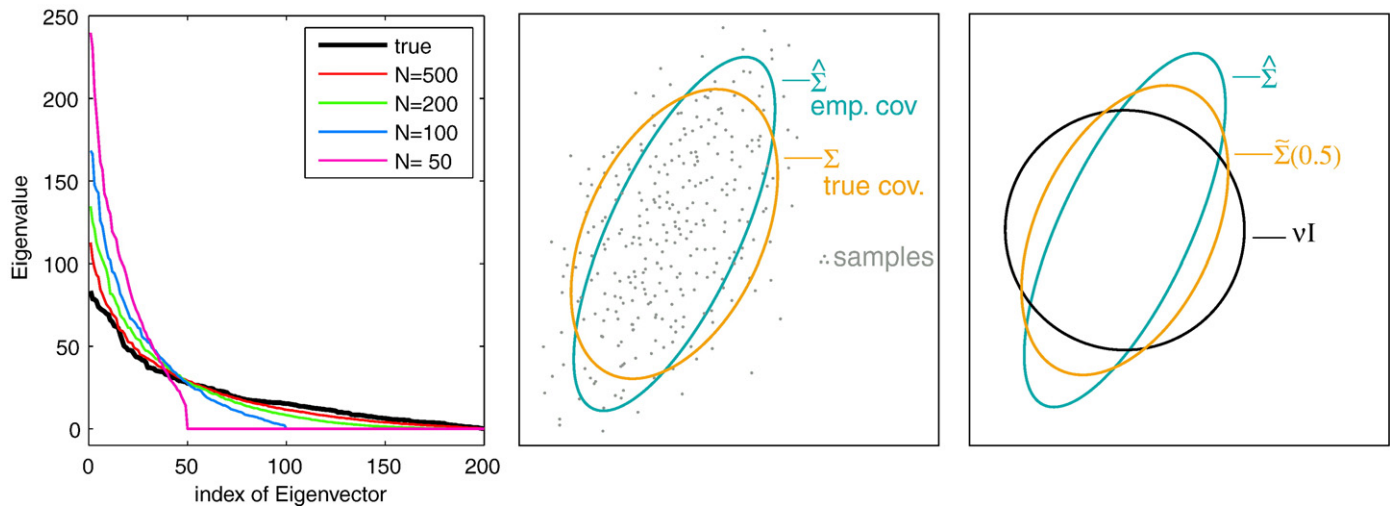


Fig. 7. Left: Eigenvalue spectrum of a given covariance matrix (bold line) and eigenvalue spectra of covariance matrices estimated from a finite number of samples drawn ($N=50, 100, 200, 500$) from a corresponding Gaussian distribution. Middle: Data points drawn from a Gaussian distribution (gray dots; $d=200$ dimensions, two dimensions selected for visualization) with true covariance matrix indicated by an orange colored ellipsoid, and estimated covariance matrix in cyan. Right: An approximation of the true covariance matrix can be obtained as a linear interpolation between the empirical covariance matrix and a sphere of appropriate size.

topography and provide thus a survey of the spatial distribution of discriminative information, see Fig. 8 (a).

In this case, we can see that relevant information originates from central locations (P3), but even much better discriminability is provided by occipital regions (vN2 component). The fact that classification in the matrix speller is mainly based on visual evoked potentials rather than the P3 was only recently reported (Treder and Blankertz, 2010; Bianchi et al., 2010). The separability map found here is completely in line with neurophysiological plausibility, but in other paradigms it might indicate the need for further preprocessing of the signals.

Classification in the spatial domain

ERPs exhibit a particular spatial distribution during the peak times of their subcomponents. Spatial features calculated from single time points (or potential values obtained by averaging within a given time interval) have been introduced in Features of ERP classification section. Depending on the experimental setting, classification of such spatial feature may already yield powerful classification, given an appropriate selection of the time interval. But there is often a complex of several ERP components that contribute to the classification. In that case, spatio-temporal features can enhance the result, see Classification in the spatio-temporal domain section.

For spatial features, a classifier with 'maximum' shrinkage ($\gamma=1$) uses the pattern of the difference of the two classes as projection

vector, see also Fig. 5. This corresponds to a rather smooth topography, and might therefore seem neurophysiologically more plausible. The more intricate spatial filters we get with little shrinkage account for the spatial structure of the noise and hold therefore the potential of more accurate classification, see Fig. 9 and the discussion in Understanding linear classifiers section. In this sense, the shrinkage parameter reflects our belief in the estimation of the noise. If the noise (covariance matrix) can be estimated very well, it should be taken into account for classification without restriction ($\gamma=0$, i.e., ordinary LDA). But if the spatial structure of the noise cannot be reliably estimated from the training data one should disbelieve it and shrink towards the no-noise assumption ($\gamma=1$) in order to avoid overfitting. The procedure for the selection of the shrinkage parameter γ provides a trade-off, that was found to work well for classification of ERP components.

Fig. 11 (left part) shows the classification results of spatial features extracted at different time intervals for the example data set using ordinary LDA (the right panel of that figure is explained in the next section).

Classification in the spatio-temporal domain

A good way to get an overview of where the discriminative information lies in the spatio-temporal plane is to visualize a matrix of separability measures to the spatio-temporal features of the

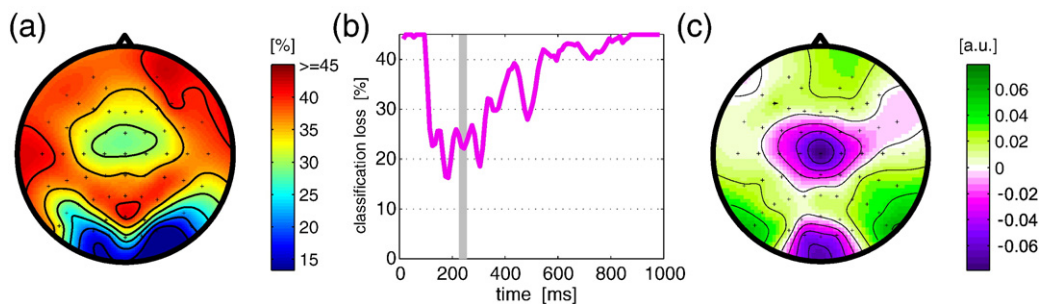


Fig. 8. Temporal and spatial classification. (a) The classification error of temporal features extracted from the time interval 115–535 ms was determined for each single channel. The obtained values are visualized here as scalp topography by spatially interpolating the values assigned to each channel position and displaying the result as color coded map. (b) The classification error of spatial features was determined for each time interval of 30 ms duration, shifted from 0 to 1000 ms. (c) Classifier as spatial filter: A linear classifier was trained on spatial features extracted from the time interval 220–250 ms (shaded in subplot (b)) of the running example data set. The resulting weight vector can be visualized as a topography and can be regarded as a spatial filter.

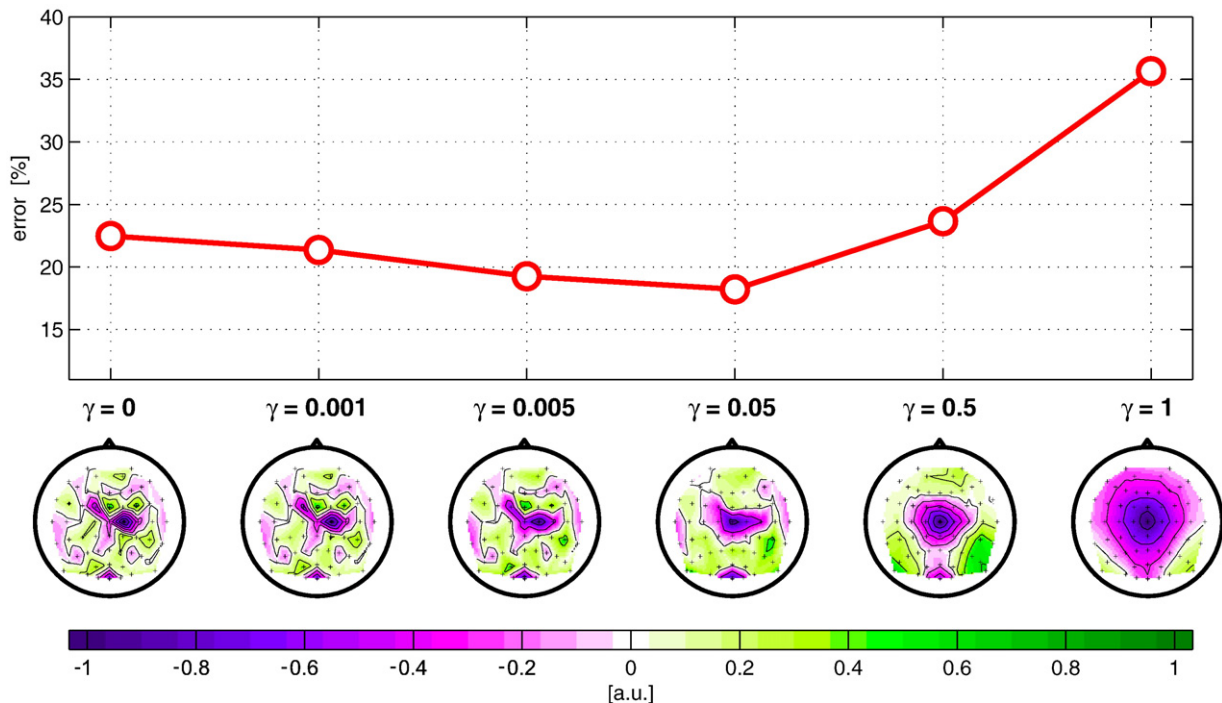


Fig. 9. Results for different values of shrinkage parameter γ . Classification accuracy of spatial features (time interval 205–235 ms) was evaluated for different values of shrinkage parameter γ . The normal vector of the classifier's separating hyperplane w can be regarded as a spatial filter (see [Spatial filters and spatial patterns](#) and [Understanding linear classifiers](#) sections) and is visualized here for each value of γ as scalp map (arbitrary unit). For $\gamma = 1$, the normal vector w is proportional to $\mu_2 - \mu_1$, while it is proportional to $\hat{\Sigma}^{-1}(\mu_2 - \mu_1)$ in the case without shrinkage ($\gamma = 0$). Note, that for illustration reasons, shrinkage was applied here to spatial features, i.e., data which is not so high-dimensional. Therefore, the gain obtained by shrinkage is not as huge as it often is for higher dimensional data, see below the result for spatio-temporal features.

experimental conditions. More specifically, this matrix is obtained by calculating a separability index separately for each pair of channel and time point. In this paper, we will use signed- r^2 -values: The pointwise biserial correlation coefficient (r -value) is defined as

$$r(x) := \frac{\sqrt{N_1 \cdot N_2}}{N_1 + N_2} \frac{\text{mean}\{x_i | y_i = 1\} - \text{mean}\{x_i | y_i = 2\}}{\text{std}\{x_i\}}, \quad (14)$$

and the signed- r^2 -values are $\text{sgn} - r^2(x) := \text{sign}(x) \cdot r(x)^2$. Alternatively, other measures of separability of distributions can be used, such as Fisher score (Duda et al., 2001; Müller et al., 2004), Student's t -statistic (Student, 1908; Müller et al., 2004) area under the ROC curve (Green and Swets, 1966) or rank-biserial correlation coefficient (Cureton, 1956). The latter two measures have the advantage that they do not rely on the assumption that the class distributions are Gaussian. On the other hand, since LDA assumes Gaussian distributions, it seems appropriate to make the same assumption in feature selection.

The analysis result for our example data set is shown in Fig. 10. Based on the displayed matrix of separability indices (here, r^2 values), it is possible to determine a set of time intervals ($\{T_1, \dots, T_n\}$ in our notions of [Features of ERP classification](#) section) that are good candidates for classification. Since within each interval the average across time is calculated, see Eq. (2), the intervals should be chosen such that within each interval the spatial pattern is fairly constant. Due to the high dimensionality of spatio-temporal features it is important to use the shrinkage technique presented in [Shrinkage for improved classification](#) section for classification.

Fig. 11 (right part) shows the classification results for spatio-temporal features. While ordinary LDA suffered from overfitting and obtained a result of 25% error, which is worse than most single interval results, shrinkage greatly reduced the error to only 4%. While the results of ordinary LDA was disappointing in this particular example data set, we would like to point out, that LDA is in many cases a powerful and robust classification method. In the example used

here, the number of training samples was low (750) compared to the dimensionality of the features ($7 \times 55 = 385$), since this requires the estimation of $385 \times 384/2 = 73,920$ parameters in the covariance matrix. Additionally, to the bias in the estimation of the covariance matrix, there is also a numerical problem in the inversion of a badly conditioned matrix. There are ways to cope with these numerical issues, e.g., involving singular value decomposition, but since that is also kind of regularization, we did not use such a method for comparison here.

Finally, we would like to point out that the weight vector of a linear classifier trained on spatio-temporal features can be visualized similar as for spatial features, see Fig. 8 (c), but as a *sequence* for scalp maps that represent the resulting spatial filters for each of the chosen time intervals T_i .

Empirical evaluation

Finally, we demonstrate the effect of shrinkage on ERP detection performance and present classification results, validated on data of all 13 participants for both types of speller paradigms, Hex-o-Spell and the Matrix Speller, see [Example data set](#) section.

In this context, we restrict the analysis to the binary classification problem *target vs. non-target* and provide validation results for a varying number of training samples, which nicely demonstrates the effect of degrading performance in cases of small samples-to-feature ratios and its remedy by shrinkage.

We compare shrinkage LDA with ordinary LDA and also with stepwise LDA (SWLDA, Draper and Smith, 1966), because the latter is the state-of-the-art algorithm for the Matrix Speller resp.ERP classification in some BCI groups (e.g., Krusienski et al., 2008, 2006; Farwell and Donchin, 1988). In Krusienski et al. (2006) different classification algorithms have been compared with the conclusion that SWLDA has the greatest potential for the usage in the Matrix Speller. The SWLDA classifier was extended in Johnson and Krusienski

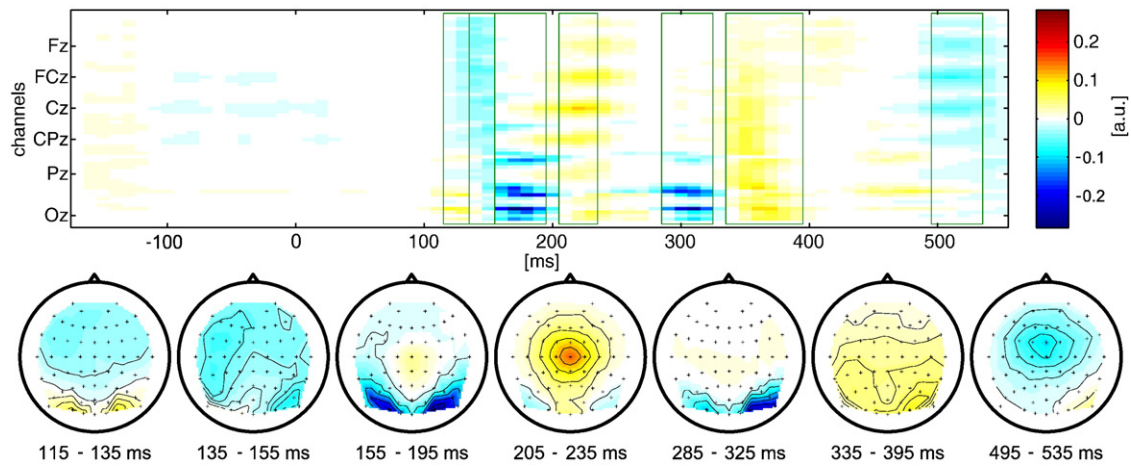


Fig. 10. Visualization of signed r^2 -matrix. For the spatio-temporal features, signed r^2 -values of targets minus nontarget single trial ERPs were calculated and displayed as a color coded matrix. A heuristic was used to determine the indicated time intervals, which accumulate high r^2 values and each time interval has a fairly constant spatial pattern of r^2 -values. The time-averaged r^2 -values for each of those intervals are also visualized as scalp topography. Note, that the matrix visualization also shows propagation of components. The visual N1 component with peak around 175 ms at electrodes PO7 and PO8 seems to originate from frontal areas around 115 ms (light blue band going from top (= frontal areas) diagonally down to the bottom (= occipital areas)). The central P2 component peaking around 220 ms is initially more focussed to the central area and propagates from there to frontal and occipital sites.

(2009) into an ensemble framework with some improvement in performance, evaluated for the case of sufficient training data.

This SWLDA classifier is basically another regularized version of LDA, performing dimensionality reduction by restricting the number of used features (variables). Model estimation for SWLDA is done in a greedy manner by iteratively inserting and removing features from the model based on statistical tests until the maximal number of active variables is reached. The method has three free parameters to choose: two p -values guiding insertion and removal of variables, and the maximal number of variables. We here use $p_{ins} = 0.1$ and $p_{rem} = 0.15$, and a maximum of 60 predictors, as recommended in Krusienski et al. (2006).

Data preprocessing and feature selection

The continuous EEG recordings of each of the 13 participants were epoched and aligned to the onsets of intensification. A baseline

correction was done based on the average EEG activity within 170 ms directly preceding the stimulus. Epochs were split into a training and a test set as described below. The training data were used to heuristically determine a set of seven subject-specific discriminative time intervals (see [Classification in the temporal domain](#) section), which were constrained to lie between 50 ms and 650 ms poststimulus (covering all components of potential interest, like P1, N1, P2, N3, P3, N3, see Fig. 2). For each trial, spatio-temporal features (cf. [Features of ERP classification](#) section) have been calculated in the following way. The mean activity in the selected intervals for 55 scalp electrodes gave rise to 55×7 matrices $X^{(1)}, \dots, X^{(n)}$, that were stacked into feature vectors $x^{(1)}, \dots, x^{(n)}$ of dimension $p = 385$.

Performance evaluation setting

To analyze the performance of the classifiers in difficult settings of small samples-to-feature ratios, we used a varying number of training

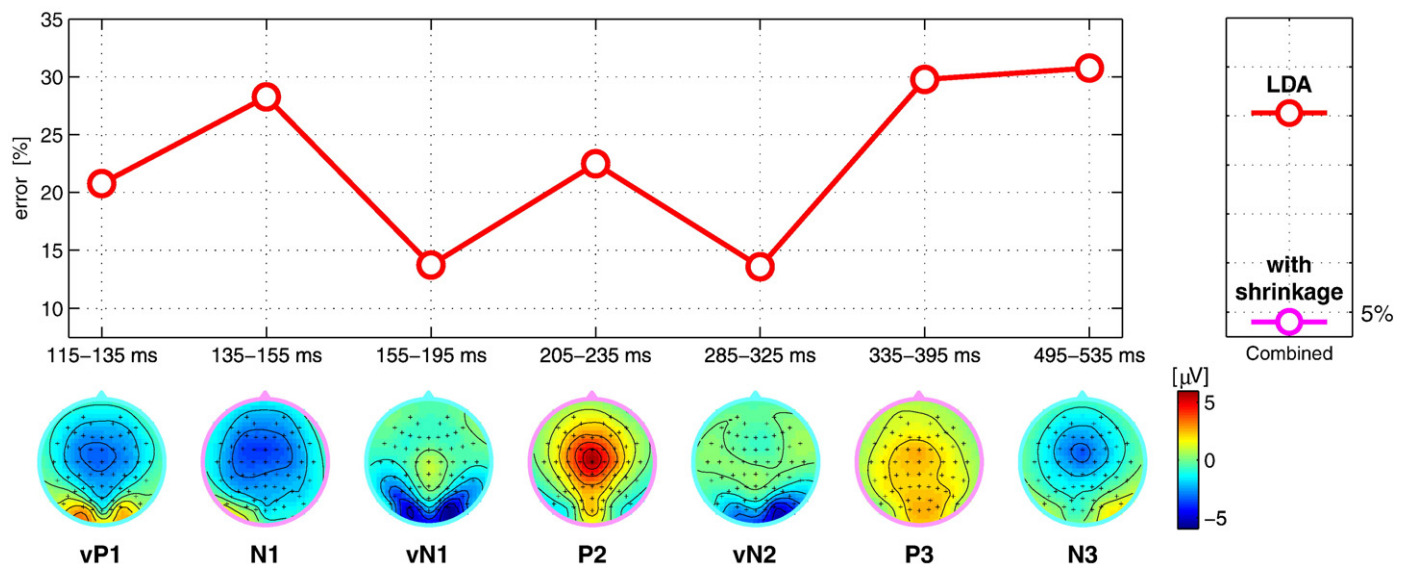


Fig. 11. Classification error: (left) using spatial features extracted from different time intervals; (right) using spatio-temporal features obtained from all eight time intervals. While LDA (red circle) obtains a result that is worse than most single interval results due to overfitting, shrinkage (magenta colored circle) greatly reduces the error to about 4%. The intervals correspond to the shaded areas in Fig. 2. Bottom row: Scalp maps show the spatial distribution of the ERP components in the corresponding time intervals.

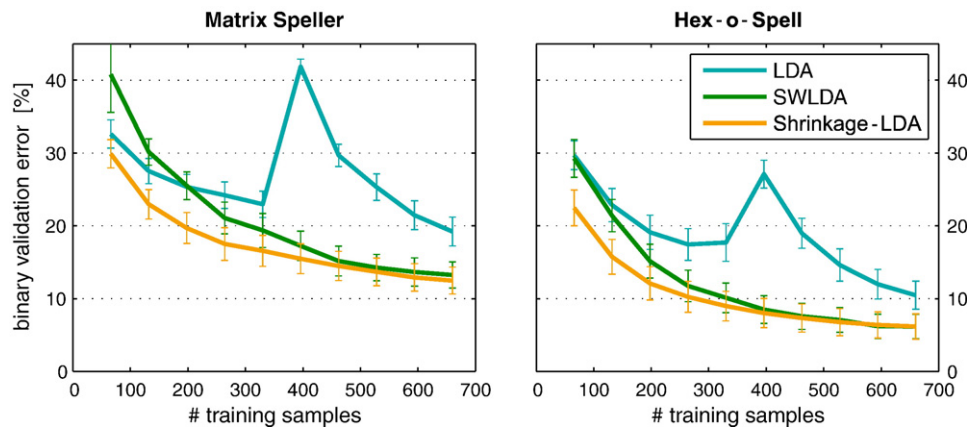


Fig. 12. Classification results on spatio-temporal features for a variable number of training samples. Three different variants of linear discriminant classifiers have been trained on spatio-temporal features of $p = 385$ dimensions. The number of training samples was varied from $n = 50$ to 650. The strange peaking behaviour of the LDA performance near $n = p$ is discussed in the main text. *Left:* Results for the Matrix Speller. *Right:* Results for Hex-o-Spell.

samples (n) between 50 and 650. Splitting the data into training and test set was done in a chronological fashion, taking the first part as training data and the second part as test data. Note, that the maximum number of samples corresponds to 6 symbol selection steps with 10 intensifications. The parameters of the linear classifiers are estimated from the training set, and applied on the test samples. The classifiers are optimized to provide good separation between target epochs (participant focuses on intensified row/column/disc) and non-target epochs (participant focuses on different item). The validation error is calculated on all left-out trials. This procedure was performed for all 13 participants separately and the error rates were averaged.

Results

Fig. 12 depicts the results in the described validation setting. In cases with $p \gg n$, Shrinkage-LDA clearly outperforms the other methods. For $p < n$ the performance of SWLDA converges towards Shrinkage-LDA, while ordinary LDA needs considerably more training samples for stable operation. The peaking behaviour of the LDA performance near the ratio $n/p = 1$ looks strange, but is well known in the machine learning literature, see Raudys and Duin (1998), Schäfer and Strimmer (2005a). It is due to a number of eigenvalues being very small but nonzero, which leads unfavorable numerical effects in the inversion of the empirical covariance matrix using the pseudo-inverse, see Krämer (2009).

Conclusion

When analyzing BCI data, we typically examine the spatial patterns and filters that allow to classify a certain brain state. In this tutorial, we identified an intuitive relation between patterns and filters in the context of regularized LDA. Furthermore, we gave two arguments for the different nature of filters in contrast to patterns, which should provide a better understanding and interpretation of spatial filters.

Mathematically, a key ingredient of the proposed algorithm was an accurate covariance matrix estimate, which is particularly hard in high dimensions. The use of a higher number of channels is potentially advantageous for ERP classification, but this improvement can only be turned to practice if the employed classification algorithm uses a proper regularization technique that allows to handle high dimensional features albeit a small number of training samples. While this insight is well-known in statistics, the practical use of shrinkage for ERP decoding is novel and yields substantial improvements in predictive accuracy.

Future work will explore shrinkage estimation also for combinations of imaging features (see e.g. Dornhege et al., 2004) and for

correlative measurements between different image modalities such as local field potentials (LFP) and fMRI (Bießmann et al., 2009). Furthermore we will consider nonlinear variants, where shrinkage could be performed in a kernel feature space (Schölkopf et al., 1998, 1999; Müller et al., 2001).

Acknowledgments

We are very grateful to Nicole Krämer (Weierstrass Institute for Applied Analysis and Stochastics, Berlin) for pointing us to the analytic solution of the optimal shrinkage parameter for regularized linear discriminant analysis.

Furthermore, we are indebted to two reviewers and our colleges in the Berlin BCI group who gave valuable comments on earlier versions of the manuscript.

The studies were partly supported by the *Bundesministerium für Bildung und Forschung* (BMBF), Fkz 01IB001A/B, 01GQ0850, by the German Science Foundation (DFG, contract MU 987/3-1), by the European Union under the PASCAL2 Network of Excellence, ICT-216886. This publication only reflects the authors' views. Funding agencies are not liable for any use that may be made of the information contained herein.

References

- Allison, B., Lüth, T., Valbuena, D., Teymourian, A., Volosyak, I., Gräser, A., 2009. BCI demographics: How many (and what kinds of) people can use an SSVEP BCI? *IEEE Trans. Neural Syst. Rehabil. Eng.* 18 (2), 107–116.
- Belouchrani, A., Abed-Meraim, K., Cardoso, J.-F., Moulines, E., 1997. A blind source separation technique using second-order statistics. *IEEE Trans. Signal Process.* 45 (2), 434–444 Feb.
- Bianchi, L., Sami, S., Hillebrand, A., Fawcett, I.P., Quitadamo, L.R., Seri, S., 2010. Which physiological components are more suitable for visual ERP based brain-computer interface? A preliminary MEG/EEG study. *Brain Topogr.* 23, 180–185 Jun.
- Bießmann, F., Meinecke, F.C., Gretton, A., Rauch, A., Rainer, G., Logothetis, N., Müller, K.-R., 2009. Temporal kernel canonical correlation analysis and its application in multimodal neuronal data analysis. *Mach. Learn.* 79 (1–2), 5–27 URL <http://www.springerlink.com/content/e1425487365v2227>.
- Birbaumer, N., 2006. Brain-computer-interface research: coming of age. *Clin. Neurophysiol.* 117, 479–483 Mar.
- 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., Lemm, S., Krauledat, M., Curio, G., Müller, K.-R., 2006. The Berlin Brain-Computer Interface: machine learning based detection of user specific brain states. *J. Univ. Comput. Sci.* 12 (6), 581–607.
- Blankertz, B., Dornhege, G., Krauledat, M., Müller, K.-R., Curio, G., 2007a. The non-invasive Berlin Brain-Computer Interface: fast acquisition of effective performance in untrained subjects. *Neuroimage* 37 (2), 539–550 URL <http://dx.doi.org/10.1016/j.neuroimage.2007.01.051>.
- Blankertz, B., Krauledat, M., Dornhege, G., Williamson, J., Murray-Smith, R., Müller, K.-R., 2007b. A note on brain actuated spelling with the Berlin Brain-Computer

- Interface. In: Stephanidis, C. (Ed.), *Universal Access in HCI, Part II, HCI 2007*. Vol. 4555 of LNCS. Springer, Berlin Heidelberg, pp. 759–768.
- Blankertz, B., Kawanabe, M., Tomioka, R., Hohlfele, F., Nikulin, V., Müller, K.-R., 2008a. Invariant common spatial patterns: Alleviating nonstationarities in brain–computer interfacing. In: Platt, J., Koller, D., Singer, Y., Roweis, S. (Eds.), *Advances in Neural Information Processing Systems 20*. MIT Press, Cambridge, MA, pp. 113–120.
- Blankertz, B., Losch, F., Krauledat, M., Dornhege, G., Curio, G., Müller, K.-R., 2008b. The Berlin Brain–Computer Interface: accurate performance from first-session in BCI-naïve subjects. *IEEE Trans. Biomed. Eng.* 55 (10), 2452–2462 URL <http://dx.doi.org/10.1109/TBME.2008.923152>.
- Blankertz, B., Tomioka, R., Lemm, S., Kawanabe, M., Müller, K.-R., 2008c. Optimizing spatial filters for robust EEG single-trial analysis. *IEEE Signal Process. Mag.* 25 (1), 41–56 Jan. URL <http://dx.doi.org/10.1109/MSP.2008.4408441>.
- Cardoso, J.-F., Souloumiac, A., 1993. Blind beamforming for non gaussian signals. *IEE Proc.-F* 140, 362–370.
- Comon, P., 1994. Independent component analysis, a new concept? *Signal Process.* 36 (3), 287–314.
- Cureton, E.E., 1956. Rank-biserial correlation. *Psychometrika* 21 (3), 287–290.
- 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.
- Dickhaus, T., Sannelli, C., Müller, K.-R., Curio, G., Blankertz, B., 2009. Predicting BCI performance to study BCI illiteracy. *BMC Neurosci.* 10 (Suppl 1), P84.
- Dornhege, G., Blankertz, B., Curio, G., Müller, K.-R., 2004. Boosting bit rates in non-invasive EEG single-trial classifications by feature combination and multi-class paradigms. *IEEE Trans. Biomed. Eng.* 51 (6), 993–1002 Jun. URL <http://dx.doi.org/10.1109/TBME.2004.827088>.
- Dornhege, G., Millán, J. del R., Hinterberger, T., McFarland, D., Müller, K.-R. (Eds.), 2007. *Toward Brain–Computer Interfacing*. MIT Press, Cambridge, MA.
- Draper, N., Smith, H., 1966. *Applied regression analysis*. Wiley series in probability and mathematical statistics. Wiley, New York.
- Duda, R.O., Hart, P.E., Stork, D.G., 2001. *Pattern Classification*, 2nd Edition. Wiley & Sons.
- 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, 510–523.
- Friedman, J.H., 1989. Regularized discriminant analysis. *J. Am. Stat. Assoc.* 84 (405), 165–175.
- Green, M.D., Swets, J.A., 1966. *Signal detection theory and psychophysics*. Krieger, Huntington, NY.
- Guger, C., Edlinger, G., Harkam, W., Niedermayer, I., Pfurtscheller, G., 2003. How many people are able to operate an EEG-based Brain–Computer Interface (BCI)? *IEEE Trans. Neural Syst. Rehabil. Eng.* 11 (2), 145–147.
- Guger, C., Daban, S., Sellers, E., Holzner, C., Krausz, G., Caraballona, R., Gramatica, F., Edlinger, G., 2009. How many people are able to control a P300-based brain–computer interface (BCI)? *Neurosci. Lett.* 462, 94–98 Oct.
- Hong, B., Guo, F., Liu, T., Gao, X., Gao, S., 2009. N200-speller using motion-onset visual response. *Clin. Neurophysiol.* 120, 1658–1666 Sep.
- Johnson, G.D., Krusienski, D.J., 2009. Ensemble SWLDA classifiers for the P300 speller. *Proceedings of the 13th International Conference on Human–Computer Interaction. Part II*. Springer-Verlag, Berlin, Heidelberg, pp. 551–557.
- Krämer, N., 2009. On the peaking phenomenon of the lasso in model selection. Preprint available: <http://arxiv.org/abs/0904.4416>.
- Krauledat, M., Tangermann, M., Blankertz, B., Müller, K.-R., 2008. Towards zero training for brain–computer interfacing. *PLoS ONE* 3 (8), e2967 Aug.
- Krusienski, D.J., Sellers, E.W., Cabestaing, F., Bayoudh, S., McFarland, D.J., Vaughan, T.M., Wolpaw, J.R., 2006. A comparison of classification techniques for the P300 speller. *J. Neural Eng.* 3 (4), 299–305 Dec.
- Krusienski, D., Sellers, E., McFarland, D., Vaughan, T., Wolpaw, J., 2008. Toward enhanced P300 speller performance. *J. Neurosci. Meth.* 167, 15–21 Jan.
- Kübler, A., Kotchoubey, B., 2007. Brain–computer interfaces in the continuum of consciousness. *Curr. Opin. Neurol.* 20, 643–649 Dec.
- Kübler, A., Müller, K.-R., 2007. An introduction to brain computer interfacing. In: Dornhege, G., Millán, J. del R., Hinterberger, T., McFarland, D., Müller, K.-R. (Eds.), *Toward Brain–Computer Interfacing*. MIT press, Cambridge, MA, pp. 1–25.
- Kübler, A., Kotchoubey, B., Kaiser, J., Wolpaw, J., Birbaumer, N., 2001. Brain–computer communication: unlocking the locked in. *Psychol. Bull.* 127 (3), 358–375.
- Ledoit, O., Wolf, M., 2004. A well-conditioned estimator for large-dimensional covariance matrices. *J. Multivar. Anal.* 88, 365–411.
- Lemm, S., Curio, G., Hlushchuk, Y., Müller, K.-R., 2006. Enhancing the signal to noise ratio of ICA-based extracted ERPs. *IEEE Trans. Biomed. Eng.* 53 (4), 601–607 April.
- Lemm, S., Blankertz, B., Dickhaus, T., Müller, K.-R., 2011. Introduction to machine learning for brain imaging. *NeuroImage* 56, 387–399.
- Li, H., Li, Y., Guan, C., 2006. An effective BCI speller based on semi-supervised learning. *Conf. Proc. IEEE Eng. Med. Biol. Soc. 1*, 1161–1164.
- Makeig, S., Jung, T.-P., Bell, A.J., Ghahramani, D., Sejnowski, T.J., 1997. Blind separation of auditory event-related brain responses into independent components. *PNAS* 94, 10979–10984.
- Martens, S.M., Hill, N.J., Farquhar, J., Schölkopf, B., 2009. Overlap and refractory effects in a brain–computer interface speller based on the visual P300 event-related potential. *J. Neural Eng.* 6, 026003 Apr.
- Meinicke, P., Kaper, M., Hoppe, F., Heumann, M., Ritter, H., 2003. Improving transfer rates in brain computer interfacing: a case study. In: Becker, S.T.S., Obermayer, K. (Eds.), *Advances in Neural Information Processing Systems 15*, pp. 1107–1114.
- Müller, K.-R., Blankertz, B., 2006. Toward noninvasive brain–computer interfaces. *IEEE Signal Process. Mag.* 23 (5), 125–128 September.
- Müller, K.-R., Mika, S., Rätsch, G., Tsuda, K., Schölkopf, B., 2001. An introduction to kernel-based learning algorithms. *IEEE Neural Netw.* 12 (2), 181–201 May.
- Müller, K.-R., Anderson, C.W., Birch, G.E., 2003. Linear and non-linear methods for brain–computer interfaces. *IEEE Trans. Neural Syst. Rehabil. Eng.* 11 (2), 165–169.
- Müller, K.-R., Krauledat, M., Dornhege, G., Curio, G., Blankertz, B., 2004. Machine learning techniques for brain–computer interfaces. *Biomed. Tech.* 49 (1), 11–22.
- Müller, K.-R., Tangermann, M., Dornhege, G., Krauledat, M., Curio, G., Blankertz, B., 2008. Machine learning for real-time single-trial EEG-analysis: From brain–computer interfacing to mental state monitoring. *J. Neurosci. Meth.* 167 (1), 82–90 URL <http://dx.doi.org/10.1016/j.jneumeth.2007.09.022>.
- Nunez, P.L., Srinivasan, R., 2005. *Electric Fields of the Brain: The Neurophysics of EEG*, 2nd Edition. Oxford University Press, USA, December.
- Parra, L., Alvino, C., Tang, A., Pearlmutter, B., Yeung, N., Osman, A., Sajda, P., 2003. Single-Trial Detection in EEG and MEG: Keeping it Linear. *Neurocomputing* 52–54, 177–183 Jun.
- Parra, L.C., Spence, C.D., Gerson, A.D., Sajda, P., 2005. Recipes for the linear analysis of EEG. *Neuroimage* 28 (2), 326–341.
- Parra, L., Christoforou, C., Gerson, A., Dyrholm, M., Luo, A., Wagner, M., Philastides, M., Sajda, P., 2008. Spatiotemporal linear decoding of brain state. *IEEE Signal Process. Mag.* 25 (1), 107–115.
- Pfurtscheller, G., Neuper, C., Birbaumer, N., 2005. Human Brain–Computer Interface. In: Riehle, A., Vaadia, E. (Eds.), *Motor Cortex in Voluntary Movements*. CRC Press, New York, pp. 367–401. Ch. 14.
- Quiroga, R., Garcia, H., 2003. Single-trial event-related potentials with wavelet denoising. *Clin. Neurophysiol.* 114, 376–390.
- Rakotomamonjy, A., Guigue, V., 2008. BCI competition III: dataset II- ensemble of SVMs for BCI P300 speller. *IEEE Trans. Biomed. Eng.* 55, 1147–1154 Mar.
- Ratcliff, R., Philastides, M.G., Sajda, P., 2009. Quality of evidence for perceptual decision making is indexed by trial-to-trial variability of the EEG. *Proc. Natl. Acad. Sci. U. S. A.* 106, 6539–6544 Apr.
- Raudys, S., Duin, R., 1998. Expected classification error of the fisher linear classifier with pseudo-inverse covariance matrix. *Pattern Recognit. Lett.* 19, 385–392.
- Salvaris, M., Sepulveda, F., 2009. Visual modifications on the P300 speller BCI paradigm. *J. Neural Eng.* 6, 046011 Aug.
- Schäfer, J., Strimmer, K., 2005. An empirical bayes approach to inferring large-scale a gene association networks. *Bioinformatics* 21 (6), 754–764.
- Schäfer, J., Strimmer, K., 2005a. A shrinkage approach to large-scale covariance matrix estimation and implications for functional genomics. *Stat. Appl. Genet. Mol. Biol.* 4 Article32.
- Schölkopf, B., Smola, A., Müller, K.-R., 1998. Nonlinear component analysis as a kernel eigenvalue problem. *Neural Comput.* 10 (5), 1299–1319.
- Schölkopf, B., Mika, S., Burges, C., Knirsch, P., Müller, K.-R., Rätsch, G., Smola, A., 1999. Input space vs. feature space in kernel-based methods. *IEEE Trans. Neural Netw.* 10 (5), 1000–1017 September.
- Shenoy, P., Krauledat, M., Blankertz, B., Rao, R.P.N., Müller, K.-R., 2006. Towards adaptive classification for BCI. *J. Neural Eng.* 3 (1), R13–R23 URL <http://dx.doi.org/10.1088/1741-2560/3/1/R02>.
- Stein, C., 1956. Inadmissibility of the usual estimator for the mean of a multivariate normal distribution. *Proc. 3rd Berkeley Sympos. Math. Statist. Probability* 1, pp. 197–206.
- Stinstra, J., Peters, M., 1998. The volume conductor may act as a temporal filter on the ECG and EEG. *Med. Biol. Eng. Comput.* 36, 711–716.
- Student, 1908. The probable error of a mean. *Biometrika* 6, 1–25.
- Sugiyama, M., Krauledat, M., Müller, K.-R., 2007. Covariate shift adaptation by importance weighted cross validation. *J. Mach. Learn. Res.* 8, 1027–1061.
- Tomioka, R., Müller, K.R., 2010. A regularized discriminative framework for EEG based communication. *Neuroimage* 49, 415–432.
- Treder, M.S., Blankertz, B., 2010. (C)overt attention and speller design in visual attention based brain–computer interfaces. *Behav. Brain Funct.* 6, 28 May.
- Vidaurre, C., Blankertz, B., 2010. Towards a cure for BCI illiteracy. *Brain Topogr.* 23, 194–198 open Access. URL <http://dx.doi.org/10.1007/s10548-009-0121-6>.
- Vidaurre, C., Schlögl, A., Cabeza, R., Scherer, R., Pfurtscheller, G., 2006. A fully on-line adaptive BCI. *IEEE Trans. Biomed. Eng.* 53 (6), 1214–1219.
- Vidaurre, C., Schlögl, A., Cabeza, R., Scherer, R., Pfurtscheller, G., 2007. Study of on-line adaptive discriminant analysis for EEG-based brain computer interfaces. *IEEE Trans. Biomed. Eng.* 54 (3), 550–556.
- Vidaurre, C., Krämer, N., Blankertz, B., Schlögl, A., 2009. Time domain parameters as a feature for eeg-based brain computer interfaces. *Neural Netw.* 22, 1313–1319.
- von Büna, P., Meinecke, F.C., Király, F., Müller, K.-R., 2009. Finding stationary subspaces in multivariate time series. *Phys. Rev. Lett.* 103, 214101.
- 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, 1081–1086 Jun.
- Williamson, J., Murray-Smith, R., Blankertz, B., Krauledat, M., Müller, K.-R., 2009. Designing for uncertain, asymmetric control: interaction design for brain–computer interfaces. *Int. J. Hum. Comput. Stud.* 67 (10), 827–841.
- Wolpaw, J., 2007. Brain–computer interfaces as new brain output pathways. *J. Physiol.* 579, 613–619 Mar.
- Wolpaw, J.R., Birbaumer, N., McFarland, D.J., Pfurtscheller, G., Vaughan, T.M., 2002. Brain–computer interfaces for communication and control. *Clin. Neurophysiol.* 113 (6), 767–791.
- Ziehe, A., Müller, K.-R., Nolte, G., Mackert, B.-M., Curio, G., 2000. Artifact reduction in magnetoneurography based on time-delayed second-order correlations. *IEEE Trans. Biomed. Eng.* 47 (1), 75–87 January.