Stationary Common Spatial Patterns for Brain-Computer Interfacing

Wojciech Samek[†‡], Carmen Vidaurre[†], Klaus-Robert Müller[†§¶], and Motoaki Kawanabe[♯]

[†]Machine Learning Group, Department of Computer Science, Berlin Institute of Technology, Franklinstr. 28/29, D-10587 Berlin, Germany
[‡]Intelligent Data Analysis Group, Fraunhofer Institute FIRST, Kekuléstr. 7, D-12489 Berlin, Germany
[§]Bernstein Fokus Neurotechnology, Berlin, Germany
[¶]Institute for Pure and Applied Mathematics, University of California Los Angeles, Los Angeles, CA 90095, USA
[♯]Advanced Telecommunications Research Institute International. 2-2-2 Hikaridai, Keihanna Science City, Kyoto 619-0288, Japan

E-mail:
wojciech.samek@campus.tu-berlin.de, carmen.vidaurre@tu-berlin.de, klaus-robert.mueller@tu-berlin.de, kawanabe@atr.jp

Abstract. Classifying motion intentions in Brain-Computer Interfacing (BCI) is a demanding task as the recorded EEG signal is not only noisy and has limited spatial resolution but it is also intrinsically non-stationary. The non-stationarities in the signal may come from many different sources, for instance electrode artifacts, muscular activity or changes of task involvement, and often deteriorate classification performance. This is mainly because features extracted by standard methods like Common Spatial Patterns (CSP) are not invariant to variations of the signal properties, thus should also change over time. Although many extensions of CSP were proposed to, for example, reduce the sensitivity to noise or incorporate information from other subjects, none of them tackles the non-stationarity problem directly. In this paper we propose a method which regularizes CSP towards stationary subspaces (sCSP) and show that this increases classification accuracy, especially for subjects who are hardly able to control a BCI. We compare our method with the state-of-the-art approaches on different data sets, show competitive results and analyse the reasons for the improvement.

1. Introduction

Brain-Computer Interface (BCI) systems (see e.g. [1, 2, 3, 4, 5, 6]) aim to decode the intent of a subject measured from brain signals, e.g. EEG is translated into control commands for a computer application or a neuroprosthesis. A popular paradigm for BCI communication is motor imagery (MI). In this paradigm the user encodes a command by imagining the execution of a movement with a particular limb, this alters the rhythmic activity in locations over the sensorimotor cortex which correspond to this limb and the BCI system detects these differences and decodes the intended command. A major problem in EEG-based BCI systems is the limited quality and resolution of the signal due to volume conduction effects, a low signal-to-noise ratio and the non-stationary nature of EEG [1].

Variations of the signal properties over time i.e. non-stationarities may arise from many sources and have different time scales, for instance changes in impedance occur when an electrode gets loose or the skin prepping gel dries out, muscular activity or eye movements lead to artifacts in the signal and we often observe a decreasing task involvement and changes in the user's background activity due to tiredness or lack of attention [1]. Further changes in the recorded EEG signal can be caused by differences between sessions e.g. no feedback in the calibration session vs. feedback in later sessions or small differences in electrode positions between sessions [7]. The result of all these variations is a feature distribution that changes over time. This violates the assumption of most statistical learning algorithms that data come from a non-changing underlying distribution and may therefore give rise to deteriorated classification performance [8].

A particularly popular and powerful signal processing technique used for feature extraction in EEG-based BCIs is Common Spatial Patterns (CSP) (see e.g. [9, 10]). The CSP algorithm computes spatial filters that aim at achieving optimal discrimination when using band power features, thus it increases the signal-to-noise ratio and reduces adverse effects of volume conduction [9]. However, CSP is also known to be sensitive to noise and prone to overfitting [11]. Several extensions of CSP have been proposed to overcome this problem. Some recent examples include the use of regularization and utilization of information from other subjects [12] or incorporation of additional measurements [13]. These extensions have shown their advantage to improve classification performance, however, none of them tackles the non-stationarity problem directly.

In this paper we propose a method called stationary Common Spatial Patterns (sCSP) which regularizes the CSP solution towards stationary subspaces i.e. we extend CSP to be invariant to non-stationarities in the data. In other words, our goal is to reduce variations of the extracted features as we assume that they come from processes which are not task-related like eye movements or electrode artifacts. We provide results on three different data sets and compare our approach to the state-of-the-art CSP methods and show that sCSP is competitive. In addition, the performance improvement is put into neuroscientific context for a selected subject.

This paper is organized as follows. In Section 2 we introduce the Common Spatial Patterns method, its extensions and our stationary CSP algorithm. Section 3 describes the experimental setup and presents and analyses the results. In Section 4 we conclude with a short summary and future research ideas.

1.1. Related Work

Recently, several methods have been proposed to robustify BCIs against noise and non-stationary changes in the data. The approaches can be divided into two main groups, namely methods extracting robust or invariant features and approaches adapting to changes in the data.

Many methods from the first group were proposed to robustify CSP against noise and artifacts. Lotte et al. [12] give an overview over different CSP variants. One promising approach is called Tikhonov Regularized CSP (TRCSP) where a multiple of the identity matrix is added to the CSP denominator in order to regularize the solution i.e. to restrict the norm of the filters. This avoids overfitting and is especially useful when only few labeled trials are available. The best method in [12] is called weighted Tikhonov Regularized CSP (WTRCSP) and applies a weighted regularization to the CSP filters and the weights are computed from other subjects. Grosse-Wentrup et al. [14] propose to compute CSP on regions of interest in order to incorporate a priori neurophysiological knowledge (see also [15]). Lotte and Guan [16] penalize spatially non smooth filters and Blankertz et al. [13] propose a method called invariant CSP which allows to learn invariances by incorporating extra measurements e.g. from an eye movement session. Several other variants of CSP exist which aim at robustifying the original CSP algorithm [7, 17, 18, 19]. A different approach extracts invariant features by projecting the data to a stationary subspace before applying CSP [20, 21]. Finally Tomioka and Müller [22] propose to learn, select and combine robust features and perform classification in an unified framework.

The other class of methods aiming at robustifying BCIs against noise and non-stationary changes are based on adaptation. In contrast to the approaches presented above, adaptive methods can handle changes which occur in subsequent sessions i.e. after CSP has been computed and the classifier has been trained. Some of the proposed approaches focus on bias adaptation [23, 24], others update classifiers or the CSP filters [25, 7]. A recent method which uses techniques for co-adaptive learning of user and machine was proposed by Vidaurre et al. [26]. Li et al. [27] and Sugiyama et al. [28] apply covariate shift adaptation to account for shifts of the features. The method of Hasan et al. [29] uses additional measures to improve the performance of the classifier when rotations in the feature space occur. Other work like Li and Guan [30] propose adaptive classifiers which are based on expectation-maximization procedure. Buttfield et al. [31] use supervised online learning for adaptation.

Our novel stationary Common Spatial Patterns method belongs to the first group as its main goal is to extract features that are robust and stationary. We first proposed

the stationary CSP algorithm in [32] and [33]. In this paper we describe the method in more detail, extend it and apply it to more data sets. Furthermore we perform extensive comparison to other regularization approaches like weighted and unweighted Tikhonov Regularized CSP [12], invariant CSP [13] and SSA+CSP [21].

2. Common Spatial Patterns Algorithms

Common Spatial Patterns (CSP) have been widely used in BCI systems [9, 10] as they are well suited to discriminate different motor imagery patterns. A CSP spatial filter $\mathbf{w}$ maximizes the variance of band-pass filtered EEG signals in one condition while minimizing it in the other condition (or equivalently minimizing the common variance). Since the variance of a band-pass filtered signal is equal to band power, CSP enhances the differences in band power between two conditions. The CSP problem can be solved for condition 1 by maximizing the Rayleigh quotient

$$R(\mathbf{w}) = \frac{\mathbf{w}^\top \mathbf{\Sigma}_1 \mathbf{w}}{\mathbf{w}^\top \{\mathbf{\Sigma}_1 + \mathbf{\Sigma}_2\} \mathbf{w}}, \tag{1}$$

where $\mathbf{\Sigma}_1$ and $\mathbf{\Sigma}_2$ are the average covariance matrices from class 1 and 2, respectively. Note that the maximization of the Rayleigh quotient can be reformulated as a constrained optimization problem

$$\max_{\mathbf{w}} \mathbf{w}^\top \mathbf{\Sigma}_1 \mathbf{w} \tag{2}$$

$$\text{subject to } \mathbf{w}^\top \{\mathbf{\Sigma}_1 + \mathbf{\Sigma}_2\} \mathbf{w} - \mathrm{C} = 0 \tag{3}$$

where C is an arbitrary constant (norm of $\mathbf{w}$ can be chosen so that equation 3 holds), and solved using Lagrange multipliers. The solution $\mathbf{w}_1^*$ satisfies $\mathbf{\Sigma}_1 \mathbf{w}_1^* = \lambda(\mathbf{\Sigma}_1 + \mathbf{\Sigma}_2)\mathbf{w}_1^*$, thus it has the form of a generalized eigenvalue problem where the generalized eigenvector with largest eigenvalue corresponds to the spatial filter $\mathbf{w}_1^*$ that maximizes the variance of class 1 while minimizing the common variance. One can show that the optimal filter $\mathbf{w}_2^*$ for condition 2 can also be obtained from equation 1, it is simply the generalized eigenvector with smallest eigenvalue [34].

2.1. Regularization of CSP

Regularization of the CSP objective function can be performed by adding a penalty term $P(\mathbf{w}) = \mathbf{w}^\top \mathbf{K} \mathbf{w}$ to the denominator of the Rayleigh quotient of equation 1 (see the work of Blankertz et al. [13] and Lotte et al. [12]). Note that in this case the best filter for class 2 does not equal the eigenvector with the smallest eigenvalue, i.e. the Rayleigh quotients needs to be maximized separately for each class

$$\mathbf{w}_1^* = \arg\max_{\mathbf{w}} \frac{\mathbf{w}^\top \mathbf{\Sigma}_1 \mathbf{w}}{\mathbf{w}^\top \{\mathbf{\Sigma}_1 + \mathbf{\Sigma}_2\} \mathbf{w} + \alpha P(\mathbf{w})}, \tag{4}$$

$$\mathbf{w}_2^* = \arg\max_{\mathbf{w}} \frac{\mathbf{w}^\top \mathbf{\Sigma}_2 \mathbf{w}}{\mathbf{w}^\top \{\mathbf{\Sigma}_1 + \mathbf{\Sigma}_2\} \mathbf{w} + \alpha P(\mathbf{w})}, \tag{5}$$

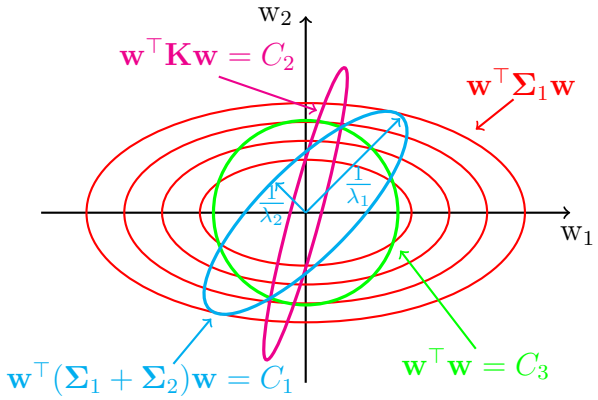

Figure 1. The computation of the CSP objective function can be reformulated as maximization of the objective $\mathbf{w}^\top\mathbf{\Sigma}_1\mathbf{w}$ subjected to certain constraints. The goal in this case is to find the largest red ellipse while satisfying the constraints represented by the other ellipses. The cyan curve stands for the standard CSP denominator i.e. it represents the common variance term. The magenta ellipse stands for the general penalty term $P(\mathbf{w}) = \mathbf{w}^\top\mathbf{K}\mathbf{w}$ used e.g. by our sCSP method. The green circle represents the Tikhonov regularization which mitigates the influence of artifacts and reduces the tendency to overfitting as filters with large norm are penalized. This kind of regularization stabilizes the solution when the matrix $\mathbf{\Sigma}_1+\mathbf{\Sigma}_2$ is not estimated properly or does not have full rank i.e. the eigenvalue λ_1 or λ_2 is zero. The constants C_1, C_2 and C_3 depend on the scale of $\mathbf{w}$ and are related with each other as $C = C_1 + C_2 + C_3$ can be arbitrarily chosen, but not all combination of C_1, C_2 and C_3 can be obtained by choosing the norm of $\mathbf{w}$.

The penalty term translates to an additional constraint in equation 3, but the maximization can still be performed by solving a generalized eigenvalue problem. Figure 1 visualizes the maximization process. Mika et al. [35] showed that this general form of regularization can be used to compute invariant features. In the following subsections five CSP variants with different penalty terms $P(\mathbf{w})$ are described.

2.2. Invariant CSP

Invariant CSP (iCSP) [13] allows to add general invariances to the CSP features by incorporating additional measurements like Electrooculography (EOG) or Electromyography (EMG) or using extra sessions for the computation of the penalty matrix $\mathbf{K}$. In order to robustify CSP features against eye movement artifacts, we compute the artifact covariance matrix $\mathbf{K}$ from EEG data recorded in an extra artifact session consisting of different eye movements, namely "eyes_open", "look_left", "look_right", "look_up" and "look_down". The extra recordings are filtered in the same frequency band as the motor imagery data, they are cut into epochs of 1.5 sec length and the average covariance matrix $\mathbf{K}$ is computed and normalized. Adding $P(\mathbf{w}) = \mathbf{w}^\top\mathbf{K}\mathbf{w}$ to the denominator of the CSP objective function results in features which are invariant against changes generated by eye movements.

2.3. Tikhonov Regularized CSP

Tikhonov Regularized CSP (TRCSP) penalizes solutions with large weights (see Lotte and Guan [12]). In this approach $\mathbf{K}$ is set to the identity matrix $\mathbf{I}$, thus $P(\mathbf{w})$ reduces to $||\mathbf{w}||^2$. The result of such regularization is mitigation of the influence of artifacts and a reduced tendency to overfitting as filters with large norm are avoided. It must be noted here that TRCSP penalizes all channels equally i.e. no distinction is made whether a channel captures a lot of BCI-related activity or is completely irrelevant. However, if a channel is likely to contain a lot of useful information, one should not prevent CSP from giving it high weights.

2.4. Weighted Tikhonov Regularized CSP

Weighted Tikhonov Regularized CSP (WTRCSP) [12] makes exactly the above distinction between channels i.e. it penalizes channels which are less important for BCI stronger than channels which capture a lot of BCI relevant information. This leads to a penalty matrix $\mathbf{K} = \mathbf{uI}$, where $\mathbf{u}$ is a vector capturing the level of penalty assigned to each channel and $\mathbf{I}$ is the identity matrix. The vector $\mathbf{u}$ can be obtained in a manual fashion using neurophysiological knowledge i.e. looking up the literature for brain regions (and thus channels) which are expected to be useful for a specific task. However, since it is not easy to select an appropriate penalty value by hand, one can compute $\mathbf{u}$ by using data from other subjects as done in [12]. The penalty value of a channel is simply set to the inverse of the average absolute value of the normalized weight of this channel in the CSP filters obtained from other subjects. Formally, this is

$$\mathbf{u} = \left(\frac{1}{2 \times N_f \times |\Omega|} \sum_{i \in \Omega} \sum_{f=1}^{2 \times N_f} \left| \frac{\mathbf{w}_f^i}{||\mathbf{w}_f^i||} \right| \right)^{-1}, \tag{6}$$

where $\mathbf{w}_f^i$ is the f-th spatial filter obtained using CSP (among the eigenvectors corresponding to the N_f largest and lowest eigenvalues) for the i-th additional subject available. Thus WTRCSP assigns higher penalties to channels with low average channel weights.

Note that none of the regularization techniques discussed so far has tackled the non-stationarity problem i.e. none of the approaches ensures that the obtained features are stationary.

2.5. SSA+CSP

Stationarity is a necessary assumption of many machine learning algorithms for optimal classification [8]. Therefore we introduce an additional preprocessing step which extracts the stationary part of the EEG signal before computing the CSP features. The underlying assumption is that the observed signal $\mathbf{x}(t)$ is a linear superposition of

stationary $\mathbf{s}^{\mathfrak{s}}(t)$ and non-stationary $\mathbf{s}^{\mathfrak{n}}(t)$ sources

$$\mathbf{x}(t) = A\,\mathbf{s}(t) = \begin{bmatrix} A^{\mathfrak{s}} & A^{\mathfrak{n}} \end{bmatrix} \begin{bmatrix} \mathbf{s}^{\mathfrak{s}}(t) \\ \mathbf{s}^{\mathfrak{n}}(t) \end{bmatrix}, \tag{7}$$

and that the BCI-related information is contained in the stationary subspace. Recently, von Bünau et al. [20] proposed a method called Stationary Subspace Analysis (SSA) to separate the $\mathfrak{s}$-sources from the $\mathfrak{n}$-sources. Before applying SSA we band-pass filter the EEG data as described in the next section and subsequently compute the CSP features on the stationary subspace (as done in [21]). Note that we combine trials of opposite classes into epochs that serve as input to SSA in order to assure that differences between both classes are not treated as non-stationarities and ignored. The dimensionality d of the stationary subspace is selected via cross-validation.

2.6. Stationary CSP

The goal of stationary CSP (sCSP) is to extract features that not only discriminate between two conditions, but are also stationary. In contrast to the two-step SSA+CSP approach presented above, sCSP optimizes discriminativity and stationarity in a single objective function. The main idea is to extract filters that maximize or minimize variances for two different conditions, but at the same time keep the variance estimation along the projected direction as stable as possible across trials. For that we introduce a measure of stationarity which is the sum of absolute differences between the projected average variance and the projected variance in k-th trial. Formally, the following quantity is minimized for each class c

$$D_c(\mathbf{w}) = \sum_k \left| \mathbf{w}^\top \mathbf{\Sigma}_c^{(k)} \mathbf{w} - \mathbf{w}^\top \mathbf{\Sigma}_c \mathbf{w} \right|, \tag{8}$$

where $\mathbf{\Sigma}_c^{(k)}$ is the covariance matrix of the k-th trial of class c and $\mathbf{\Sigma}_c$ is the average covariance matrix of class c. Figure 2 visualizes the quantity to minimize.

Since this penalty can not be introduced directly into the Rayleigh quotient‡, a related quantity $\overline{\mathbf{\Delta}}_c$ is used in this paper. For that we compute the difference between the covariance matrix of each trial and the global average $\mathbf{\Sigma}_1$ or $\mathbf{\Sigma}_2$ and ensure that the difference matrix is positive definite. Thus for each class we compute

$$\mathbf{\Delta}_1^{(k)} := \mathcal{F}\left(\mathbf{\Sigma}_1^{(k)} - \mathbf{\Sigma}_1 \right), \tag{9}$$

$$\mathbf{\Delta}_2^{(k)} := \mathcal{F}\left(\mathbf{\Sigma}_2^{(k)} - \mathbf{\Sigma}_2 \right), \tag{10}$$

where $\mathcal{F}$ is an operator to make symmetric matrices be positive definite. More precisely, if a symmetric matrix $\mathbf{M}$ has eigendecomposition $\mathbf{M} = \mathbf{V}\,\mathrm{diag}(d_i)\,\mathbf{V}^\top$, the operator

‡ The problem is that one can not take the $\mathbf{w}$ from the $|\cdot|$-function i.e. a Rayleigh quotient of the form $\frac{\mathbf{w}^T A \mathbf{w}}{\mathbf{w}^T B \mathbf{w}}$ is not obtained.

returns $\mathcal{F}(\mathbf{M}) = \mathbf{V}\,\mathrm{diag}(|d_i|)\,\mathbf{V}^\top$, i.e. the signs of all the negative eigenvalues are flipped. The intuition behind this operation is to ensure that the penalty term is always positive (similar to absolute value function in D_c), even in the case that power of a feature in the k-th trial is smaller than its global average.

Consequently instead using D_c we measure variations in the projected direction $\mathbf{w}$ as $\mathbf{w}^\top \sum_{k=1}^{K} \mathcal{F}(\mathbf{\Sigma}_c^{(k)} - \mathbf{\Sigma}_c)\mathbf{w}$. Although the quantities are not equal, they both measure absolute deviations, in the latter case before and in the case of D_c after projecting. In fact, it can be shown that our new measure is an upper bound for D_c§. In summary, we use $P(\mathbf{w}) = \mathbf{w}^\top(\overline{\mathbf{\Delta}}_1 + \overline{\mathbf{\Delta}}_2)\mathbf{w}$ as penalty term in sCSP with $\overline{\mathbf{\Delta}}_c := \frac{1}{K}\sum_{k=1}^{K} \mathbf{\Delta}_c^{(k)}$ being the average (positive definite) difference matrix of class c.

Two extensions will be applied to sCSP in this paper. First, we compute the covariance matrices not only on a trial-wise basis, but on local chunks of BCI data sequences (see [32]). More precisely, $\mathbf{\Sigma}_c^{(k)}$ does no longer denote the covariance matrix in k-th trial, but the covariance matrix estimated from k-th chunk. A chunk of size ν is a set of ν consecutive trials from the same class. By using chunks one can take into account non-stationarities on different time scales e.g. estimating the covariance matrix from individual trials allows to capture changes like muscular artifact which occur on a trial-by-trial basis whereas if the chunk size increases the focus shifts towards slower changes like variations of task involvement or electrode impedance. In this paper the best chunk size is selected via cross-validation. Note that the way we apply sCSP chunking here differs from [32]. Firstly, the maximal chunk size is larger here which allows to capture non-stationarities on a larger time-scale. Secondly, there is a difference in parameter selection when more than one parameter leads to the same lowest error rate (see Section 3.2) and finally we normalize the class covariance matrices and the penalty matrix in order to define a more meaningful range of regularization parameters.

It must be noted that sCSP aims at extracting stationary features, but it is not able to handle rank deficient matrices and does not reduce the tendency to overfitting as TRCSP does. Therefore, our second extension is to combine both approaches into a method called stationary Tikhonov Regularized CSP (sTRCSP). For that we maximize the following objective function

$$R(\mathbf{w}) = \frac{\mathbf{w}^\top \mathbf{\Sigma}_c \mathbf{w}}{\mathbf{w}^\top\{\mathbf{\Sigma}_1 + \mathbf{\Sigma}_2\}\mathbf{w} + \alpha P_{\text{sCSP}}(\mathbf{w}) + \beta P_{\text{TRCSP}}(\mathbf{w})}, \tag{11}$$

where $P_{\text{sCSP}}(\mathbf{w})$ is the penalty term of sCSP, $P_{\text{TRCSP}}(\mathbf{w})$ is the penalty term of TRCSP and α and β are determined by cross-validation.

§ Let $\mathbf{V}\mathbf{D}\mathbf{V}^\top$ be the eigendecomposition of the difference matrix $\mathbf{\Sigma}_c^{(k)} - \mathbf{\Sigma}_c$. In order to prove $\left|\mathbf{w}^\top(\mathbf{\Sigma}_c^{(k)} - \mathbf{\Sigma}_c)\mathbf{w}\right| \leq \mathbf{w}^\top\mathcal{F}(\mathbf{\Sigma}_c^{(k)} - \mathbf{\Sigma}_c)\mathbf{w}$ we introduce $\mathbf{u} = \mathbf{V}^\top\mathbf{w}$. With that $\left|\mathbf{u}^\top\mathbf{D}\mathbf{u}\right| \leq \mathbf{u}^\top|\mathbf{D}|\mathbf{u}$ or $\left|u_1^2 d_1 + u_2^2 d_2 + \ldots + u_n^2 d_n\right| \leq u_1^2|d_1| + u_2^2|d_2| + \ldots + u_n^2|d_n|$ which follows from Jensen's inequality.

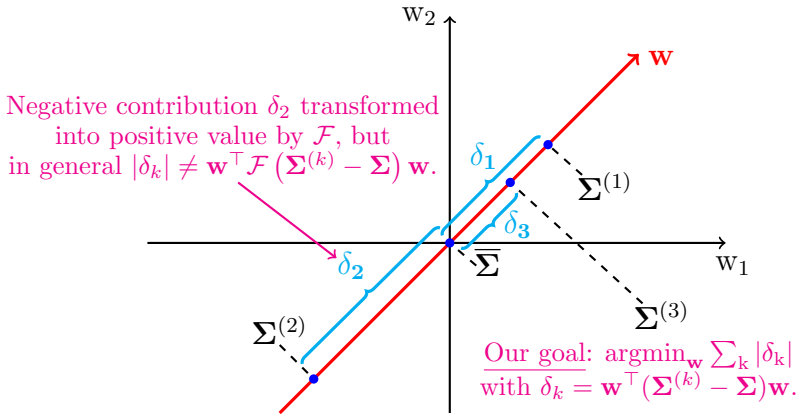

Figure 2. Visualization of the idea behind stationary CSP. The goal of sCSP is to find a projecting direction $\mathbf{w}$ with stable variance (features) for each class i.e. to minimize the quantity $D_c(\mathbf{w}) = \sum_k |\mathbf{w}^\top(\mathbf{\Sigma}_c^{(k)} - \mathbf{\Sigma}_c)\mathbf{w}|$. Since this can not be minimized directly inside the Rayleigh quotient, we introduce an operator $\mathcal{F}$ which recomputes the difference matrix $\mathbf{\Sigma}_c^{(k)} - \mathbf{\Sigma}_c$ by converting negative variations into positive ones i.e. it makes the difference matrix positive definite by flipping the sign of all negative eigenvalues. The intuition behind this transformation is to take the "absolute value" before projecting, which is an upper bound of the absolute value after projecting. This figure shows the projection direction $\mathbf{w}$ (red line) and the projection of the covariance matrices on $\mathbf{w}$ i.e. the variance which is explained by $\mathbf{w}$. Furthermore we see the variations δ_i between the average and trial-wise covariance matrices after projection.

3. Evaluation

3.1. Data Sets

The experiments in this paper are based on three different data sets containing EEG signals recorded while subjects perform motor imagery (MI).

3.1.1. Data set IVa, BCI Competition III This data set [36] from BCI Competition III [37] contains EEG signals from five healthy subjects performing right hand and foot MI without feedback. Two types of visual cues, a letters appearing behind a fixation cross and a randomly moving object, shown for 3.5 s were used to indicate the target class. The presentation of target cues were intermitted by periods of random length, 1.75 to 2.25 s, in which the subject could relax. The EEG signal was recorded from 118 Ag/AgCl electrodes, band-pass filtered between 0.05 and 200 Hz and downsampled to 100 Hz, so that 280 trials are available for each subject, among which 168, 224, 84, 56 and 28 compose the training set for subject A1, A2, A3, A4 and A5 respectively, the remaining trials composing their test set.

3.1.2. Data set IIa, BCI Competition IV This data set [38] from BCI Competition IV [39] consists of EEG recordings from 22 Ag/AgCl electrodes and nine healthy subjects performing left hand, right hand, foot and tongue MI without feedback. In this study

we only use the left and right hand motor imagery data. Two sessions on different days were recorded for each subject. Each session is comprised of 6 runs separated by short breaks. One run consists of 48 trials (12 for each of the four possible classes), yielding a total of 288 trials per session. The subjects were sitting in a comfortable armchair in front of a computer screen. At the beginning of a trial ($t = 0$ s), a fixation cross appeared on the black screen. In addition, a short acoustic warning tone was presented. After two seconds ($t = 2$ s), a cue in the form of an arrow pointing either to the left, right, down or up appeared and stayed on the screen for 1.25 s. This prompted the subjects to perform the desired motor imagery task. No feedback was provided. The subjects were asked to carry out the motor imagery task until the fixation cross disappeared from the screen at $t = 6$ s. A short break followed where the screen was black again. The signals were sampled with 250 Hz and bandpass-filtered between 0.5 Hz and 100 Hz. Both the training and the testing set contain 72 trials per class.

3.1.3. Data set from Vital BCI Project This data set [40] comes from a joint study with University Tübingen and contains EEG recordings from 80 healthy volunteers (41 female, age 29.9±11.5 years; 4 left-handed) performing motor imagery tasks with the left and right hand or with the feet. The subjects were sitting in a comfortable chair with arms lying relaxed on armrests. Brain activity was recorded from the scalp with multi-channel EEG amplifiers using 119 Ag/AgCl electrodes in an extended 10-20 system sampled at 1000 Hz (downsampled to 100 Hz) with a band-pass from 0.05 to 200 Hz. First, the subjects performed a calibration recording in which every 8s one of three different visual cues (arrows pointing left, right, down) indicated to the subject which type of motor imagery to perform: left/right hand or foot. Three runs with 25 trials of each motor condition were recorded. Then, two of the classes were selected and the subjects performed feedback with three runs of 100 trials each, although for some subjects only one or two runs were recorded. Each trial of feedback started with a period of 2s with a black fixation cross in the center of a gray screen. Then an arrow appeared behind the cross to indicate the target direction of that trial and 1s later the cross turned purple and started moving according to the classifier output. After 4s of cursor movement the cross froze at the final position and turned black again. Two seconds later the cross was reset to the center position and the next trial began. Both sessions were recorded on the same day. All subjects in this study are BCI novices.

3.2. Experimental Setup

For the BCI competition data sets the same preprocessing is applied as in Lotte and Guan [12] i.e. the time segment located from 0.5s to 2.5s after the cue instructing the subject to perform MI is extracted and the signal is band-pass filtered in 8-30 Hz using a 5-th order Butterworth filter. For the Vital BCI data we do not use fixed preprocessing, but select the best binary task-combination and estimate the most discriminative frequency band and time segment (typically 750-3500 ms relative to the

Table 1. Comparison of classification error rates for data set IVa and IIa from BCI Competition III and IV, respectively. The best results for each subject are displayed in bold characters. The overall performance of stationary CSP is better than of the other approaches, especially for subjects lacking BCI efficiency.

	BCI Competition III					BCI Competition IV									Overall		
Subject	A1	A2	A3	A4	A5	B1	B2	B3	B4	B5	B6	B7	B8	B9	Mean	Median	Std
CSP	33.9	**3.6**	41.8	11.2	**19.0**	7.6	48.6	**2.8**	34.0	42.4	32.6	**19.4**	6.3	**6.3**	22.1	19.2	16.3
TRCSP	25.9	**3.6**	41.8	20.1	**19.0**	7.6	48.6	**2.8**	34.7	42.4	30.6	**19.4**	6.3	8.3	22.2	19.8	15.5
WTRCSP	20.5	5.4	41.8	15.2	**19.0**	7.6	**40.3**	**2.8**	33.3	41.0	**31.3**	**19.4**	6.3	9.0	20.9	19.2	14.2
SSA+CSP	33.9	5.4	39.8	9.4	**19.0**	7.6	43.1	**2.8**	29.9	40.3	32.6	30.6	6.3	**6.3**	21.9	24.5	15.2
sCSP	23.2	**3.6**	**33.7**	**8.5**	19.8	**6.9**	45.8	**2.8**	**27.8**	41.0	33.3	20.1	**4.2**	6.9	**19.8**	20.0	14.8
sTRCSP	**17.9**	5.4	37.2	10.2	20.6	7.6	49.3	**2.8**	29.2	**38.2**	**30.3**	**19.4**	6.9	8.3	20.2	**18.7**	14.5

presentation of the visual cue) for each subject using calibration data (as done in [9]).

For the experiments we manually select 68 electrodes‖ densely covering the motor cortex. We do not apply any manual or automatic rejection of trials or electrodes and use three filters per class for feature extraction as recommended in [9]. As classifier we apply Linear Discriminant Analysis (LDA) and use error rate to measure performance. In order to set a meaningful range for the regularization parameters, we normalize the covariance matrices $\mathbf{\Sigma}_c$ and $\mathbf{K}$ by dividing them by their traces. The α and β parameters are selected from the set of 10 candidates $\{0, 2^{-8}, ..., 2^{-1}, 2^0\}$ by 5-fold cross-validation on the calibration data. In addition, for sCSP we select the best chunk size ν from $\{1, 5, 10\}$ on the calibration data, for sTRCSP we use a fixed chunk size of $\nu = 5$ in order to save computation time. Note that the parameter selection is performed according to error rate, but if more than one parameter value leads to the same lowest error, we select the parameter with highest Fisher Score between the classification output and the true label¶. All results in this paper are offline results.

3.3. Performance Comparison

At first we compare the performances of our new methods with three state-of-the art approaches and the CSP baseline. Table 1 shows the error rates of CSP, TRCSP, WTRCSP, SSA+CSP, sCSP and sTRCSP for the data set IVa and IIa. Note that the invariant CSP method could not be tested as additional measurements e.g. from an eye movement session were not available.

From the results we see that on average both stationary CSP methods perform

‖ The electrodes F*, FFC*, FC*, CFC*, C*, CCP*, CP*, PCP*, P*, PPO* and PO* are used. Electrode higher than 6 according to the International 10-20 system are discarded.

¶ The Fisher Score between two random variables X and Y is defined as $\mathrm{FS(X,Y)} = \frac{(\mathrm{E[X]-E[Y]})^2}{\mathrm{Var[X]+Var[X]}}$ with $E[\cdot]$ and $Var[\cdot]$ being the mean and variance operator respectively.

better than the other approaches. The difference in performance between sCSP and CSP, SSA+CSP or TRCSP is significant up to 95% (using one-sided Wilcoxon signed-rank test) and sTRCSP is significantly better than CSP and TRCSP when excluding the participants with very small error rates ($< 10\%$). There is no significant difference between WTRCSP and stationary CSP, however, our methods do not use information from other subjects. One observation from the results is that participants who perform well with CSP benefit less from applying regularization than subjects who can hardly control a BCI. The largest improvement can be achieved for subjects A1, A3, B4 and B5. Users lacking BCI efficiency often have a low signal-to-noise ratio and an artifactual and non-stationary signal, thus CSP may fail to capture neurophysiologically meaningful information and to produce stable and discriminative features. Regularization alleviates this problem as it weakens the influence of artifacts, avoids overfitting and/or reduces variations of the features. It must be noted that we were not able to reproduce the results of Lotte and Guan [12] for all subjects. The baseline CSP performance of subjects A3, A4 and A5 is significantly better in our experiments, even when using recordings from all electrodes and applying the same normalization as Lotte and Guan. Since improving a better baseline is more challenging, we did not further investigate the deviations.

As a second experiment we tested all methods (including iCSP) on a larger data set, namely the Vital BCI data set containing 80 BCI novices performing motor imagery. The results are visualized in Figure 3 and Figure 4. The first figure compares error rates of different methods using scatter plots. We see that sTRCSP is the winning method and greatly improves classification results, it is even superior to WTRCSP which uses information from other subjects. The mean (median) error rates of the methods are as follows: CSP = 29.5% (31%), TRCSP = 26.9% (23.8%), WTRCSP = 26.7% (22.8%), iCSP = 29.1% (29%), SSA+CSP = 29.4% (29.8%), sCSP = 27.4% (25.6%) and sTRCSP = 26.2% (22.7%). From Figure 4 which shows a boxplot of classification performances one can see that the median performance (red line) is very similar for sTRCSP, TRCSP and WTRCSP, but sTRCSP has a higher 25% quantile. In other words sTRCSP (and sCSP) significantly improves classification performance of subjects lacking BCI efficiency. Astonishingly, invariant CSP and SSA+CSP lead to a performance decrease (lower 75% quantile) for participants who perform well with CSP. We compare the effects of regularization and analyse the reasons for improvement and deterioration in the next subsection.

As before the Wilcoxon signed-rank test is used to evaluate significance. Table 2 shows the p-values (one-sided) of different comparisons using either all subjects or dividing them into three groups based on their error rates: 0% - 15%, 15% - 30% and above 30%. Note that the null hypothesis of the signed-rank test states that the median of the distribution of error rate differences is zero. Therefore one needs to consider Figure 3 in order to arrive at the hypothesis "Method A is better than Method B". As before we see from Table 2 that most significant improvements are obtained for subjects in the above 30% error rate group. Furthermore it seems advantageous to combine Tikhonov Regularization and stationarity. In fact, sTRCSP outperforms both

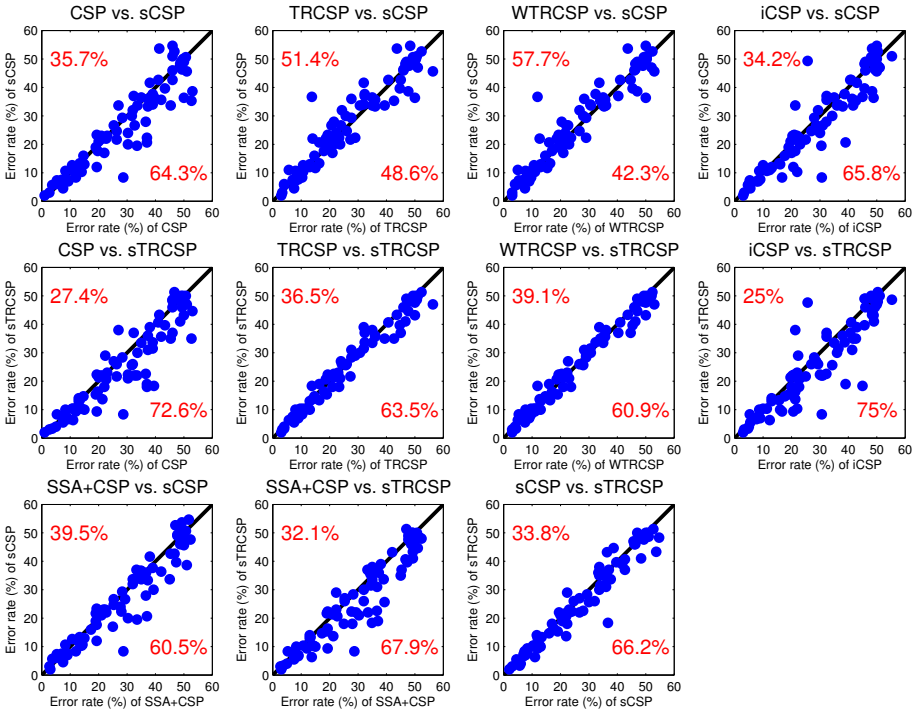

Figure 3. Nine scatter plots comparing the error rates of different methods on the Vital BCI data set. Each subject is represented by a blue dot. If the percentage of points below the diagonal is lower, the method reported in the y-axis performs better. sTRCSP outperforms all other approaches. sTRCSP outperforms all other approaches.

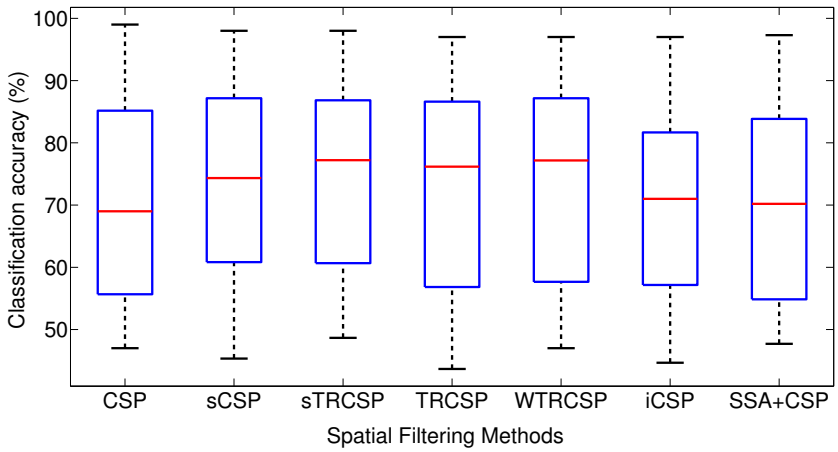

Figure 4. Comparison of the classification accuracies using box plots for the Vital BCI data set. The red lines denote the median classification performance, the lower and upper sides of the blue box represent the 25% and 75% quantiles and the black line stands for the outliers. The median values of sCSP, sTRCSP, TRCSP and WTRCSP are very similar, but the stationary CSP methods perform better on subjects with low classification accuracy i.e. they have a higher 25% quantile.

sCSP and TRCSP as it finds the right trade-off between discriminative ability of the filters, stationarity of the features and robustness of the estimation. It must be noted

Table 2. Overview of Wilcoxon signed-rank test p-values (one-sided) for different comparisons. Bold values indicate significance with 5% level. Grouping is performed based on the error rates of CSP and Figure 3 is considered to build the hypothesis *"Method A is better than Method B"*.

Comparison	0-15	15-30	>30	all
sCSP better than CSP	.3081	.2065	**.0021**	**.0015**
sCSP better than TRCSP	.5000	.4952	.1197	.1473
WTRCSP better than sCSP	.3969	.2312	.0887	.0685
sCSP better than iCSP	**.0461**	.1222	**.0087**	**.0019**
sTRCSP better than CSP	.0679	.0894	**.0000**	**.0000**
sTRCSP better than TRCSP	.0851	.2305	.0844	**.0303**
sTRCSP better than WTRCSP	.1986	.4156	**.0050**	.0596
sTRCSP better than iCSP	**.0139**	.0784	**.0000**	**.0000**
sCSP better than SSA+CSP	.0584	.2673	**.0045**	**.0015**
sTRCSP better than SSA+CSP	**.0182**	.1745	**.0000**	**.0000**
sTRCSP better than sCSP	.2065	.1600	**.0022**	**.0009**

that although its mean absolute improvement over CSP is 3.3%, the decrease in error rate for some subjects is more significant, e.g. for ten subjects it is larger than 10% and the maximum improvement is 21.1%. The standard deviation over all subjects is 6.0%. The most popular parameters for sTRCSP are $\alpha = 2^{-5}$ and $\beta = 2^{-8}$. Fixing the parameters to these values for all subjects still gives a significant improvement over CSP with mean (median) error rate of 26.9% (25.2 %). However, there is no significant difference to TRCSP and WTRCSP (with cross-validation) in this case.

3.4. Effects of Regularization

In this subsection we investigate the regularization effects of TRCSP, iCSP, SSA+CSP and sCSP. We will not include the weighted TRCSP method and sTRCSP in the analysis as they do not represent an own regularization concept, but are modifications/combinations of sCSP and TRCSP regularization. The analysis is conducted with subject 30 performing left vs. foot motor imagery as this user shows one of the largest improvements in terms of classification accuracy and the limited CSP performance is not due to a few bad outlier trials, but the reasons are more general and the effects of regularization can be relatively easily visualized. Similar analysis was conducted with several subjects and is available at: `http://www.user.tu-berlin.de/wojwoj/research/jne2012.html`. An overview over the error rates and the parameters selected by cross-validation is given in Table 3.

Before comparing the different regularization approaches, one should understand the sources and effects of the non-stationarities present in subject 30's data. The scalp

Table 3. Overview over subject 30's error rates and the parameters (regularization parameter α, dimensionality of stationary subspace d and chunk size ν) selected by cross-validation for different regularization approaches.

	CSP	TRCSP	iCSP	SSA+CSP	sCSP
Error rates	37%	23.3%	39.7%	28.0%	22.3%
Selected Parameters	-	$\alpha = 2^{-2}$	$\alpha = 2^{-4}$	$d = 53$	$\alpha = 2^{-2}$, $\nu = 10$

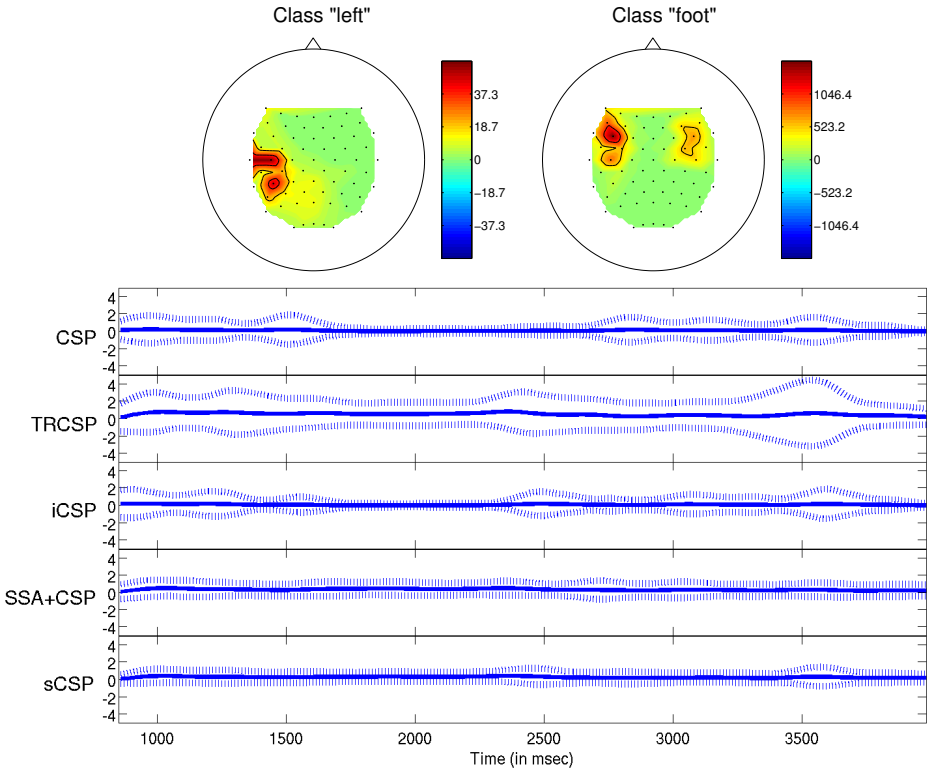

Figure 5. Illustration of non-stationarities for subject 30. Upper Figure: Non-stationarity map for both motor imagery classes "left" and "foot". For each electrode the variance of the variances of the trials is computed and plotted. Especially in the frontal, temporal and occipital areas one can see large changes. Lower Figure: The bandpass filtered EEG signal is projected using the best performing (normalized) CSP, TRCSP, iCSP, SSA+CSP and sCSP filter and the mean and standard deviation is computed at each time point over all 75 trials of class "foot". The solid line represents the mean and the dashed lines stand for the mean ± standard deviation interval. The selected time interval is 850 to 3970 msec. We see that both sCSP and SSA+CSP produce a stable signal with small standard deviation, whereas in the case of CSP, TRCSP and iCSP the variations between trials are larger and non-constant.

plots in the upper part of Figure 5 visualize the variations for each electrode and each class measured as the variance of the variances of the trials. This measure gives an idea of where the most significant variations of the EEG can be observed. The largest changes can be found in frontal, temporal and occipital locations. The lower part of

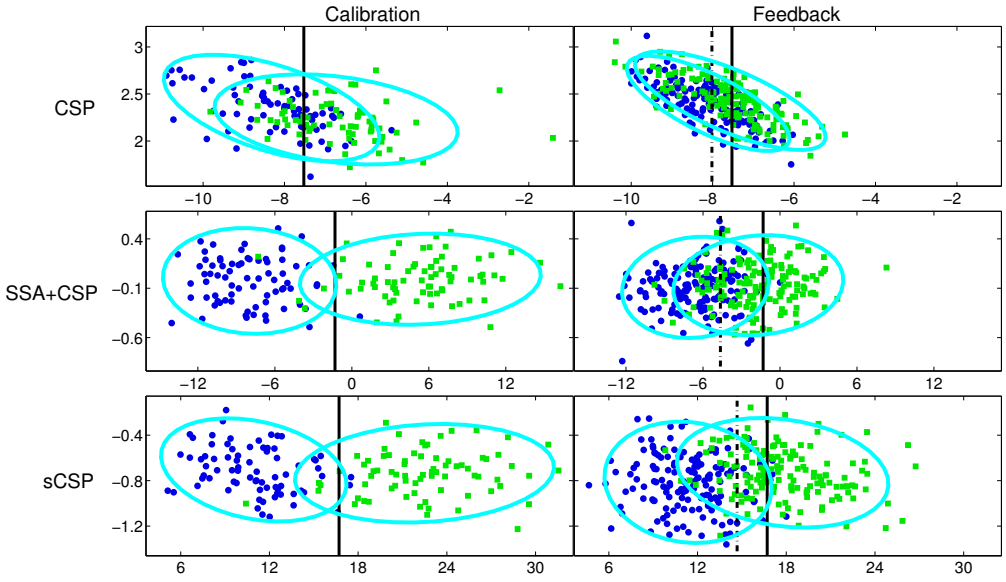

Figure 6. Calibration and test features for CSP, SSA+CSP and sCSP for both motor imagery classes "left" (blue) and "foot" (green) for subject 30. The features are projected to the normal vector of the classifier hyperplane (x-axis) and the largest PCA-component (y-axis) of the calibration data. The vertical black line represents the classifier bias which is computed in calibration phase, the dashed line stands for the optimal feedback bias. We see that the CSP features do not only undergo a shift between calibration and feedback phase, but their distribution changes. This rotation can not be resolved by adapting the bias. The SSA+CSP and sCSP features are more robust, but the SSA+CSP features have a larger overlap between both classes in feedback phase, thus are less discriminative than in the case of sCSP.

the figure shows the variations of the projected signal for the class "foot". A clear difference can be observed between sCSP and SSA+CSP on the one hand and CSP, TRCSP and iCSP on the other hand. In the first case the variations are small and do not change over time, thus the extracted features are stationary. In contrast, for the other methods the variations are larger and non-constant which results in non-stationary features. Although stationarity of the signal seems to be important, it is not sufficient to explain the classification results of subject 30. For instance, applying TRCSP and sCSP leads to very similar classification performance although the TRCSP signal varies significantly, on the other hand the SSA+CSP signal is very stationary, but the results are worse than in the sCSP case.

In the following we study the impact of the non-stationarities on the features. Figure 6 shows subject 30's calibration and test features for CSP, SSA+CSP and sCSP. The features, blue points correspond to left hand and green points to foot motor imagery, are projected to the normal vector of the classifier hyperplane (x-axis) and the largest PCA-component (y-axis) of the calibration data. The black line represents the bias used by the classifier (computed in calibration phase), the dashed black line stands for the optimal feedback bias. The first observation that can be made from Figure 6 is that in the calibration phase the SSA+CSP and sCSP features are more separable than the

CSP features. This indicates that in the case of SSA+CSP and sCSP the filters better capture information that helps to discriminate between left hand and foot MI, thus they are potentially neurophysiologically more meaningful. We will see later that this is indeed the case for subject 30. Another interesting point in Figure 6 is the difference in the feature distributions between calibration and feedback. In the case of CSP the features not only undergo a translation, but the shape of the distribution changes. In contrast, in the case of SSA+CSP and sCSP there is a shift, but no significant rotation occurs. A shift in the features can be easily resolved by unsupervised adaptation of the bias (see [24, 33]), whereas it is much harder to adapt to rotations. Although there is no significant difference with respect to discriminativity between the SSA+CSP and sCSP features in calibration phase, the overlap between the "left" and "foot" class in feedback phase is smaller for sCSP i.e. the sCSP features are more separable. This may be due to the fact that SSA+CSP (in contrast to sCSP) contains an unsupervised step which may discard discriminative information. In summary, one can say that sCSP mitigates the influence of non-stationarities in subject 30's EEG, thus produces more separable and stable features than CSP (and SSA+CSP).

After studying the effects of non-stationarities on the signal and the features, we are now going to compare the different regularization approaches in more detail. Figure 7 shows the spatial filters, i.e. a vector of weights that are assigned to each electrodes, and spatial patterns , i.e. a vector containing the source activities to the signals acquired at the different sensors (cf. Blankertz et al. [41]), with best classification performance for CSP, TRCSP, iCSP, SSA+CSP and sCSP. It can be seen that CSP and iCSP fail to extract meaningful BCI-related filters, therefore have the largest error rates. The reason for the limited performance of CSP is the noise and the artifacts contained in subject 30's signal (see Figure 5) since after removing 7 noisy electrodes and 44 artifactual trials, the overall error rate of CSP decreases from 37% to 19.3%. We will see later that although artifact removal helps for subject 30, it is in general inferior to sCSP and sTRCSP. The fact that CSP is prone to overfitting and can be negatively affected by artifacts is well known, but regularization approaches mitigate it. However, although iCSP uses regularization, it is not able to extract a meaningful filter as the penalty matrix is computed in an extra eye movement session, but eye movements do not seem to be the main source of non-stationarities in subject 30. Furthermore, during the eye-movement measure, an idle mu rhythm might appear in the sensorimotor cortex, thus penalizing these regions may remove discriminant information. In contrast to CSP and iCSP, the other three methods are able to extract filters that are related to left hand MI. The TRCSP filter is very smooth due to regularization of the norm, but it is still affected by noise in electrode FC3. The SSA+CSP filter is less affected by artifacts in the frontal electrodes than TRCSP as it removes the non-stationarities by applying SSA, however, its performance is still worse than that of sCSP which effectively manages to reduce the impact of artifacts in electrode FC3 and to extract the "cleanest" left hand MI filter. Among all methods only sCSP optimizes for both discriminativity and stationarity simultaneously.

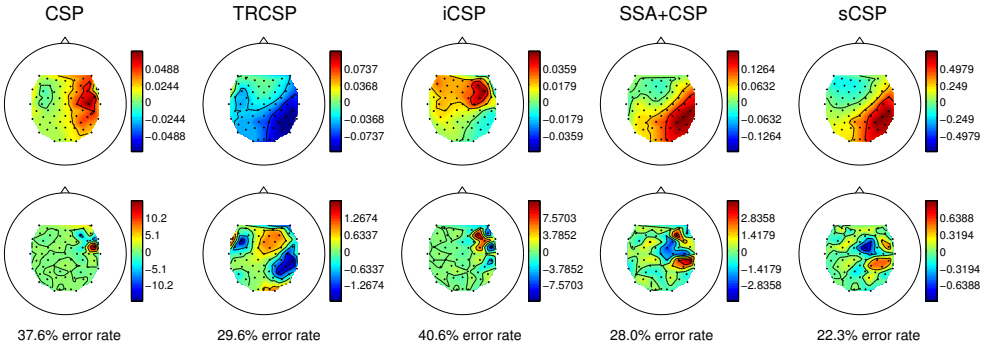

Figure 7. Illustration of the best performing patterns (upper row) and filters (lower row) for subject 30 performing left vs. foot motor imagery. The error rates of the corresponding dimensions can be seen at the bottom. Although in all cases the activation patterns can be interpreted as left hand motor imagery activation, i.e. there is a dipole-like activation over the right motor cortex, they differ significantly. In the case of CSP and iCSP the filters fail to capture the correct pattern as they are adversely affected by non-stationarities (see Figure 8). On the other hand TRCSP and SSA+CSP provide more meaningful filters, thus perform better than CSP and iCSP, however, they are still affected by artifacts in electrode FC3. Our sCSP method penalizes non-stationarities and extracts a clean left hand motor imagery filter.

With the filters in Figure 7 we can explain the large variations in the projected signal of CSP, iCSP and TRCSP (see Figure 5) and the difference in the feature distribution of CSP and sCSP (see Figure 6). Only sCSP and SSA+CSP extract filters that capture BCI-related activity and mitigate the influence of artifactual electrodes. These filters produce stable features as the BCI-related activity is stationary. In contrast, the other methods extract information from locations that are affected by artifacts, consequently the features are more non-stationary. In the case of TRCSP the filter is still discriminative, thus there is no significant performance deterioration.

One can gain more insight into the regularization effects by studying Figure 8 which visualizes locations that are being penalized by sCSP, iCSP and SSA+CSP. In the first two rows we apply PCA to the penalty matrix of sCSP and iCSP in order to obtain the non-stationarity patterns. In the case of sCSP the largest regularization is applied to the left frontal and temporal electrodes and the right frontal and occipital areas. These location are highly non-stationary (see Figure 5) and deteriorate performance of CSP, thus penalizing these electrodes helps to extract neurophysiologically more meaningful filters. In the second row one can see the effects of regularization with iCSP. Since the penalized locations do not coincide with the non-stationary regions in the calibration data, iCSP is not able to improve classification. On the contrary, it deteriorates performance by removing potentially discriminant information from central electrodes (probably because of idle mu rhythm over sensorimotor cortex). The bottom row of Figure 8 shows four projections to the non-stationary subspace computed by SSA. Also here we see that areas are penalized that coincide with the non-stationarities in the data, therefore SSA+CSP extracts a more meaningful filter than CSP. Although SSA+CSP

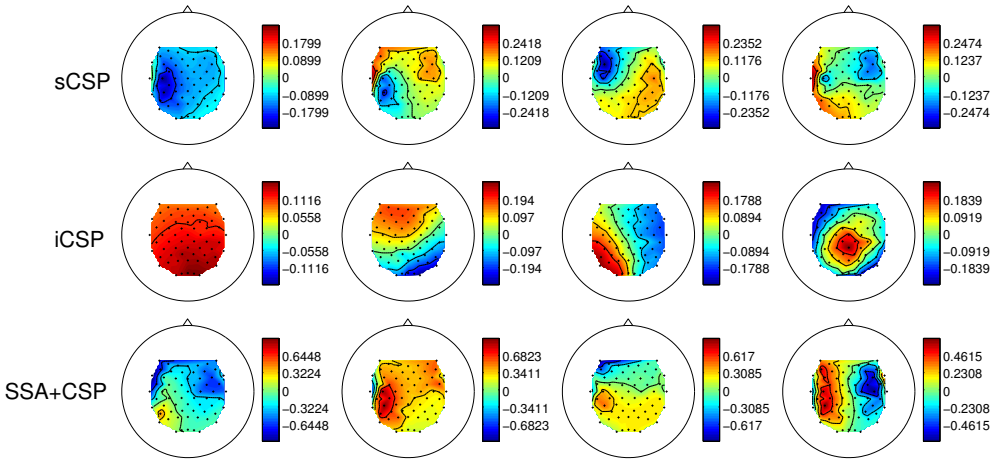

Figure 8. Illustration of the regularization effects of sCSP, iCSP and SSA+CSP for subject 30. The first row shows the eigenvectors of the penalty matrix $(\overline{\mathbf{\Delta}}_1 + \overline{\mathbf{\Delta}}_2)$ i.e. the penalty weights that are put on the electrodes by sCSP. There is a correspondence between regions with high penalty and locations with large variations (see Figure 5). The eigenvectors of the iCSP penalty matrix are plotted in the second row. Since the penalty matrix was computed from an extra session, it does not capture non-stationaries appearing during the experiment. In contrast, iCSP even removes potentially important information as it penalizes central electrodes which are discriminative for the "foot" class. This deteriorates classification performance. The bottom row visualizes the non-stationary activity patterns that are removed by SSA. As in the case of sCSP, SSA penalizes regions with large variations, thus it is also able to improve classification performance.

improves classification accuracy for subject 30, it performs worse than sCSP on average. This is mainly because SSA is an unsupervised method and may remove BCI-related information, whereas sCSP optimizes stationarity and discriminativity simultaneously.

3.5. Regularization vs. Artifact Rejection

The last subsection showed that regularization methods may improve subject 30's classification performance by mitigating the effects of artifacts. In the following we study the relationship between regularization and artifact rejection.

The goal of artifact rejection is to clean up the data by removing artifactual trials and noisy electrodes before computing the features. In this sense artifact rejection is more flexible than regularization as it can remove individual trials and electrodes. On the other hand rejection is a 0/1 decision and may result in information loss whereas regularization methods like sCSP penalize combinations of electrodes and may remove more complex non-stationarities than simple electrode artifacts. This is in spirit to weighed averaging found in ERP analysis [42, 43]. It should be noted that sCSP is more flexible than WTRCSP and TRCSP as it captures the non-stationarities present in the data (i.e. it is data-driven) and allows to penalize complex patterns whereas (W)TRCSP only applies an (weighted) uniform penalty. There is a significant correlation of .6421

(.7015) between the performance increase of sCSP and the gain of (W)TRCSP i.e. there exist common regularization effects. On the other hand sCSP and TRCSP complement each other as a combination of both gives significantly better results than each of the method on its own. In fact, sTRCSP combines the data-driven regularization of sCSP with the robustifying uniform penalty of TRCSP.

In the following we compare sCSP with a combination of artifact rejection and CSP using all 80 subjects. Artifactual trials and electrodes are identified in a heuristic manner by analysing the variance of the EEG signal (see [9]). In addition, an expert examines the data and if necessary manually removes trials or electrodes. Although, applying CSP on the cleaned data reduces the mean (median) error rate from 29.5% (31%) to 28.6% (28.3%), both sCSP and sTRCSP (without artifact rejection) perform significantly better with p-values .0254 and .0001. A weak correlation of .3514 between the performance gain of CSP with artifact rejection and sTRCSP exist indicating some similarity between both artifact rejection and regularization. However, note that removing trials and electrodes not only improves classification results, but it can also deteriorate performance e.g. for subjects 18, 24 and 49. In summary one can say that sCSP is more than an artifact rejection method as it identifies non-stationarities on different time scales (by using chunks) and mitigates the influence of (combinations of) noisy electrodes without removing them. Since it is data-driven, it adapts to non-stationarities present in the data (in contrast to TRCSP).

3.6. Using Stationary CSP in Practice

In this subsection we comment on several issues relevant for the application of sCSP and its extensions in practice.

The first question which is relevant for the practical application is whether one should use a fixed chunk size or select it via cross-validation ? Since the optimal chunk size depends on the time-scale of the non-stationarities present in the data, we expect to obtain better results when choosing the chunk size for each user individually. Using a chunk size which is smaller than the non-stationarities present in the data results in unreliable and noisy estimation of the covariance matrix and may lead to features which are invariant against task-related fluctuations, thus may be suboptimal. On the other hand when using too large chunks one may average out the important non-stationarities, thus the extracted features will not be robust against them. Indeed in our Vital BCI experiments all chunk sizes were selected quite uniformly, chunk size 1 was selected 36 times, 5 was selected in 20 cases and 10 was selected for 24 subjects. We could also observe that selecting the chunk size by cross-validation on average outperforms a fixed setting. In the case of chunk size 1 and 5 the average (relative) gain of using cross-validation for chunk size selection is 2%, in the case of 10 it is slightly larger, namely 3%. Especially subjects with limited BCI performance showed a preference for smaller chunk sizes. In fact, these users very often suffer from non-stationaries that can only be captured on small time-scales like muscle or electrode artifacts. Among the fixed chunk

sizes the trial-wise case works best. So if it is required to reduce computation time, one should use a chunk size of 1. In general, however, we recommend to individually select this parameter.

A different question is whether one should use the extended version of sCSP, namely sTRCSP, or the original one without Tikhonov Regularization ? On the Vital BCI data sTRCSP clearly outperformed sCSP, in the case of BCI Competition it had a better median performance, but a little worse mean error rate. It must be noted that there is a difference between the BCI Competition data sets and the Vital BCI data in terms of the noise level and artifacts contained. Furthermore, in the case of Vital BCI the users are BCI novices and the frequency band and time interval are user optimized, thus the risk of overfitting might be higher. On such data sets Tikhonov Regularization reduces the tendency to overfit and robustifies the filters, therefore on the Vital BCI data set TRCSP improves classification performance compared to CSP. In contrast, on the BCI Competition data set there is less noise and therefore less need to robustify the filters, thus using TRCSP does not increase performance. For that reason sTRCSP does not outperform sCSP, but is on par. In general we recommend to use sTRCSP as it combines both the advantages of Tikhonov Regularization and stationarity.

Regarding the trade-off between computation complexity and performance, in the case of sCSP or sTRCSP the additional computation does not exceed the limit of practicability. Spatial filters are usually computed in the break between calibration and feedback phase, or maybe during a break after some feedback runs. Therefore, the selection of one or two additional parameters which may last a couple of minutes more is acceptable. If one used the method in an online setting were the filters are computed and adapted regularly then the selection of sCSP or sTRCSP with fixed parameters would be a better choice.

Finally there is a question about the benefits of sTRCSP over WTRCSP. On the BCI Competition data sets sTRCSP is significantly better according to sign rank test when excluding subject B2, whereas in the case of Vital BCI sTRCSP performs significantly better than WTRCSP for the group of subjects with error rate $> 30\%$ and is almost significantly better in total ($p = .0596$). The average relative improvement of sTRCSP over WTRCSP is 2%. However, not only the better performance is an argument for sTRCSP, but also the fact that it does not use information from other subjects. This is important as recording data from similar experiments is always time consuming and costly. Additionally, in contrast to WTRCSP, one can also apply sTRCSP as a tool for analysis of non-stationarities.

4. Conclusion

In this paper we presented an approach which regularizes the CSP solution towards stationary subspaces i.e. extracts features that are invariant to variations of the signal properties. We compared this method with the state-of-the-art approaches and observed a significant performance gain, especially for subjects lacking BCI efficiency.

Furthermore we combined the stationary CSP method with Tikhonov Regularization in order to robustify the features and avoid overfitting. We showed that one reason for the performance improvement of sCSP and sTRCSP is the penalization of non-stationary electrodes which can corrupt the CSP filters. When computed on larger chunk sizes stationary CSP can potentially reduce shifts which occur at a longer time scale e.g. changes in task involvement or changes in electrode impedance. We analysed the relations between regularization methods and artifact rejection and showed that sCSP provides significantly better results than semi-automatic artifact rejection.

Unlike other methods, such as invariant CSP, our method is completely data-driven and does not need additional recordings or models of the expected changes occurring in the EEG. In future research we will study other data-driven regularization criteria and investigate ways to combine the information contained in different penalty matrices and across different imaging modalities (see Bießmann et al. [44]). Furthermore we plan to analyse the variability of the sCSP patterns over time in order to gain more insights into the nature of the changes.

Acknowledgment

This work was supported by the German Research Foundation (GRK 1589/1), the European Union under the project TOBI (FP7-ICT-224631) and the Federal Ministry of Economics and Technology of Germany under the project THESEUS (01MQ07018). MK is supported by Japanese Ministry of Internal Affairs and Communications (the Network BMI project). 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

[1] G. Dornhege, J. del R. Millán, T. Hinterberger, D. McFarland, and K.-R. Müller, editors. *Toward Brain-Computer Interfacing*. MIT Press, Cambridge, MA, 2007.

[2] 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, 2002.

[3] A. Kübler, B. Kotchoubey, J. Kaiser, J. Wolpaw, and N. Birbaumer. Brain-computer communication: unlocking the locked in. *Psychological bulletin*, 127(3):358–375, 2001.

[4] N. Birbaumer, N. Ghanayim, T. Hinterberger, I. Iversen, B. Kotchoubey, A. Kübler, J. Perelmouter, E. Taub, and H. Flor. A spelling device for the paralysed. *Nature*, 398(6725):297–298, 1999.

[5] S. Lemm, B. Blankertz, T. Dickhaus, and K.-R. Müller. Introduction to machine learning for brain imaging. *NeuroImage*, 56(2):387–399, 2011.

[6] J. del R. Millán, R. Rupp, G. Müller-Putz, R. Murray-Smith, C. Giugliemma, M. Tangermann, C. Vidaurre, F. Cincotti, A. Kübler, R. Leeb, C. Neuper, K.-R. Müller, and D. Mattia. Combining brain-computer interfaces and assistive technologies: State-of-the-art and challenges. *Frontiers in Neuroscience*, 4:161, doi:10.3389/fnins.2010.00161, 2010.

[7] M. Krauledat. *Analysis of Nonstationarities in EEG signals for improving Brain-Computer Interface performance*. PhD thesis, Technische Universität Berlin, 2008.

[8] V. N. Vapnik. *Statistical Learning Theory*. Wiley-Interscience, September 1998.

[9] B. Blankertz, R. Tomioka, S. Lemm, M. Kawanabe, and K.-R. Müller. Optimizing Spatial filters for Robust EEG Single-Trial Analysis. *IEEE Signal Proc. Magazine*, 25(1):41–56, 2008.

[10] H. Ramoser, J. Müller-Gerking, and G. Pfurtscheller. Optimal spatial filtering of single trial eeg during imagined hand movement. *IEEE Trans. Rehab. Eng.*, 8(4):441–446, 1998.

[11] B. Reuderink and M. Poel. Robustness of the common spatial patterns algorithm in the bci-pipeline. Technical Report TR-CTIT-08-52, Centre for Telematics and Information Technology, University of Twente, Enschede, July 2008.

[12] F. Lotte and C. Guan. Regularizing common spatial patterns to improve bci designs: Unified theory and new algorithms. *IEEE Trans. Biomed. Eng.*, 58(2):355 –362, February 2011.

[13] B. Blankertz, M. Kawanabe R. Tomioka, F. U. Hohlefeld, V. Nikulin, and K.-R. Müller. Invariant common spatial patterns: Alleviating nonstationarities in brain-computer interfacing. In *Ad. in NIPS 20*, pages 113–120, 2008.

[14] M. Grosse-Wentrup, C. Liefhold, K. Gramann, and M. Buss. Beamforming in noninvasive brain computer interfaces. *IEEE Trans. Biomed. Eng.*, 56(4):1209 –1219, April 2009.

[15] C. Sannelli, C. Vidaurre, K.-R. Müller, and B. Blankertz. Csp patches: an ensemble of optimized spatial filters. an evaluation study. *Journal of Neural Engineering*, 8(2):025012, 2011.

[16] F. Lotte and C. Guan. Spatially regularized common spatial patterns for eeg classification. In *Proc. of 20th Int. Conf. on Pattern Recognition*, ICPR '10, pages 3712–3715, 2010.

[17] S. Lemm, B. Blankertz, G. Curio, and K.-R. Müller. Spatio-spectral filters for improving the classification of single trial eeg. *IEEE Trans. Biomed. Eng*, 52:1541–1548, 2005.

[18] M. Kawanabe, C. Vidaurre, B. Blankertz, and K.-R. Müller. A maxmin approach to optimize spatial filters for eeg single-trial classification. In J. Cabestany, F. Sandoval, A. Prieto, and J. Corchado, editors, *Bio-Inspired Systems: Computational and Ambient Intelligence*, volume 5517 of *LNCS*, pages 674–682. Springer, 2009.

[19] G. Dornhege, B. Blankertz, M. Krauledat, F. Losch, G. Curio, and K.-R. Müller. Combined Optimization of Spatial and Temporal Filters for Improving Brain-Computer Interfacing. *IEEE Trans. on Biomed. Eng.*, 53(11):2274–2281, November 2006.

[20] P. von Bünau, F. C. Meinecke, F. C. Király, and K.-R. Müller. Finding Stationary Subspaces in Multivariate Time Series. *Physical Review Letters*, 103(21):214101+, November 2009.

[21] P. von Bünau, F.C. Meinecke, S. Scholler, and K.-R. Müller. Finding stationary brain sources in eeg data. In *IEEE Int. Conf. of Engineering in Medicine and Biology Society (EMBC)*, pages 2810 –2813, September 2010.

[22] R. Tomioka and K.-R. Müller. A regularized discriminative framework for EEG analysis with application to brain-computer interface. *NeuroImage*, 49(1):415–432, July 2009.

[23] P. Shenoy, M. Krauledat, B. Blankertz, R. P. Rao, and K.-R. Müller. Towards adaptive classification for BCI. *Journal of neural engineering*, 3(1):R13–R23, March 2006.

[24] C. Vidaurre, M. Kawanabe, P. von Bünau, B. Blankertz, and K.-R. Müller. Toward unsupervised adaptation of lda for brain-computer interfaces. *IEEE Trans. Biomed. Eng.*, 58(3):587 –597, March 2011.

[25] R. Tomioka, J.N. Hill, B. Blankertz, and K. Aihara. Adapting spatial filter methods for nonstationary bcis. In *Proc. of Workshop on Information-Based Induction Sciences (IBIS)*, pages 65 – 70, 2006.

[26] C. Vidaurre, C. Sannelli, K.-R. Müller, and B. Blankertz. Machine-learning-based coadaptive calibration for brain-computer interfaces. *Neural Comp.*, 23(3):791–816, 2011.

[27] Y. Li, H. Kambara, Y. Koike, and M. Sugiyama. Application of covariate shift adaptation techniques in brain-computer interfaces. *IEEE Trans. Biomed. Eng.*, 57(6):1318–24, 2010.

[28] M. Sugiyama, M. Krauledat, and K.-R. Müller. Covariate shift adaptation by importance weighted cross validation. *J. Mach. Learn. Res.*, 8:985–1005, December 2007.

[29] B. Awwad Shiekh Hasan and J.Q. Gan. Unsupervised adaptive gmm for bci. In *Prof. of Int. IEEE EMBS Conf. on Neural Engineering (NER '09)*, pages 295 –298, May 2009.

[30] Y. Li and C. Guan. An extended em algorithm for joint feature extraction and classification in brain-computer interfaces. *Neural Comp.*, 18(11):2730–2761, 2006.

[31] A. Buttfield, P.W. Ferrez, and J. del R. Millán. Towards a robust bci: error potentials and online learning. *IEEE Trans. Neural. Syst. Rehabil. Eng.*, 14(2):164 –168, june 2006.

[32] W. Wojcikiewicz, C. Vidaurre, and M. Kawanabe. Stationary common spatial patterns: Towards robust classification of non-stationary eeg signals. In *36th IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP '11)*, pages 577–580, 2011.

[33] W. Wojcikiewicz, C. Vidaurre, and M. Kawanabe. Improving classification performance of bcis by using stationary common spatial patterns and unsupervised bias adaptation. In E. Corchado, M. Kurzynski, and M. Wozniak, editors, *Hybrid Artificial Intelligent Systems*, volume 6679 of *LNCS*, pages 34–41. Springer, 2011.

[34] H. Ramoser, J. Müller-Gerking, and G. Pfurtscheller. Optimal spatial filtering of single trial eeg during imagined hand movement. *IEEE Trans. Rehab. Eng*, 8(4):441–446, 1998.

[35] S. Mika, G. Rätsch, J. Weston, B. Schölkopf, A. Smola, and K.-R. Müller. Invariant feature extraction and classification in kernel spaces. In *Ad. in NIPS 12*, pages 526–532, 2000.

[36] G. Dornhege, B. Blankertz, G. Curio, and K.-R. Müller. Boosting bit rates in noninvasive eeg single-trial classifications by feature combination and multiclass paradigms. *IEEE Trans. Biomed. Eng.*, 51(6):993 –1002, june 2004.

[37] B. Blankertz, K.-R. Müller, D.J. Krusienski, G. Schalk, J.R. Wolpaw, A. Schlogl, G. Pfurtscheller, J. del R. Millán, M. Schroder, and N. Birbaumer. The bci competition iii: validating alternative approaches to actual bci problems. *IEEE Trans. on Neural Syst. and Rehabil. Eng.*, 14(2):153 –159, june 2006.

[38] M. Naeem, C. Brunner, R. Leeb, B. Graimann, and G. Pfurtscheller. Seperability of four-class motor imagery data using independent components analysis. *Journal of Neural Engineering*, 3(3):208, 2006.

[39] M. Tangermann, K.-R. Müller, A. Aertsen, N. Birbaumer, C. Braun, C. Brunner, R. Leeb, C. Mehring, K. J. Miller, G. Müller-Putz, G. Nolte, G. Pfurtscheller, H. Preissl, G. Schalk, A. Schlögl, C. Vidaurre, S. Waldert, and B. Blankertz. Review of the bci competition iv. *Frontiers in Neuroscience*, 2012.

[40] B. Blankertz, C. Sannelli, S. Halder, E. M. Hammer, A. Kübler, K.-R. Müller, G. Curio, and T. Dickhaus. Neurophysiological predictor of smr-based bci performance. *NeuroImage*, 51(4):1303–1309, 2010.

[41] B. Blankertz, S. Lemm, M. S. Treder, S. Haufe, and K.-R. Müller. Single-trial analysis and classification of ERP components – a tutorial. *NeuroImage*, 56(2):814–825, 2011.

[42] S. J. Luck. *An Introduction to the Event-Related Potential Technique (Cognitive Neuroscience)*. The MIT Press, August 2005.

[43] A. Bezerianos, N. Laskaris, S. Fotopoulos, and P. Papathanasopoulos. Data dependent weighted averages for recording of evoked potential signals. *Electroencephalography and Clinical Neurophysiology/Evoked Potentials Section*, 96(5):468 – 471, 1995.

[44] F. Bießmann, S. M. Plis, F. C. Meinecke, T. Eichele, and K.-R. Müller. Analysis of multimodal neuroimaging data. *IEEE Rev. Biomed. Eng.*, 4:26 – 58, 2011.