Robust Common Spatial filters with a Maxmin Approach

Motoaki Kawanabe[1†*], **Wojciech Samek**[2†], **Klaus-Robert Müller**[2,3*], **and Carmen Vidaurre**[2]

[1]ATR Brain Information Communication Research Laboratory Group, 2-2-2 Hikaridai Seika-cho, Soraku-gun, Kyoto 619-0288, Japan.

[2]Department of Machine Learning, Berlin Institute of Technology (TU Berlin), Marchstr. 23, 10587 Berlin, Germany.

[3]Department of Brain and Cognitive Engineering, Korea University, Anam-dong, Seongbuk-gu, Seoul 136-713, Korea.

† M. Kawanabe and W. Samek contributed equally to this work.

Keywords: Common Spatial Patterns, Brain-Computer Interfacing, Robustness.

Abstract

Electroencephalographic signals are known to be non-stationary and easily affected by artifacts, therefore their analysis requires methods that can deal with noise. In this work we present a way to robustify the popular Common Spatial Patterns (CSP) algorithm under a maxmin approach. In contrast to standard CSP that maximizes the variance ratio between two conditions based on a single estimate of the class covariance matrices, we propose to robustly compute spatial filters by maximizing the minimum variance ratio within a prefixed set of covariance matrices called the tolerance set. We show that this kind of maxmin optimization makes CSP robust to outliers and reduces its tendency to overfit. We also present a data driven approach to construct a tolerance set that captures the actual variability of the covariance matrices over time and show its ability to reduce non-stationarity of the extracted features and to significantly improve the classification accuracy. We test the spatial filters derived with this approach and compare them to standard CSP and to a state-of-the-art method on a real world brain-computer interface (BCI) data set in which we expect substantial fluctuations caused by environmental differences. Finally we investigate the advantages and limitations of the maxmin approach with simulations.

1 Introduction

Brain-computer interfaces (BCI) (Wolpaw et al., 2002; del R Millán, 2003; Dornhege et al., 2007) are systems that translate the users intent, coded by a small set of mental tasks, into control actions such as computer applications or prostheses. In order to translate the brain activity into commands, it is necessary to extract meaningful features from the acquired signals. One of the most popular tools to extract information from the brain signals is the calculation of Common Spatial Patterns (CSP) (e.g. (Fukunaga, 1990; Blankertz et al., 2008c; Tomioka and Müller, 2009; Lotte and Guan, 2011; Reuderink, 2011)). This data driven approach optimizes spatial filters for each subject individually.

CSP analysis is embedded in machine learning methods (Müller et al., 2001; Montavon et al., 2013; Parra et al., 2005; Dyrholm et al., 2007). Over the last years, machine learning has led to significant advances in the analysis and modeling of neural signals (Lemm et al., 2011; Blankertz et al., 2007; Wolpaw et al.,

2002). In EEG-BCI experimentation, it has reduced the time needed for user's neurofeedback training from several days to only a few sessions (Blankertz et al., 2008b, 2010; Vidaurre et al., 2007). Typically, collecting examples of EEG signals during which the user is cued to perform repeatedly a small number of e.g. motor imagery tasks (Müller et al., 2008) is sufficient to adapt the system to the subject and start the feedback. During feedback the users are enabled to transfer information through their brain activity and control applications. However, there are several aspects in which BCI research can profit from improvement, see the 'Challenges' section of (Nijholt et al., 2008).

One of them is to adapt the system to the changing signals (Buttfield et al., 2006; Li and Guan, 2006; Shenoy et al., 2006; Vidaurre et al., 2006; Blumberg et al., 2007; Sugiyama et al., 2007; Vidaurre et al., 2007; Wang et al., 2007; Lu et al., 2009; Vidaurre et al., 2011b). Another approach is to robustify the system against non task-related fluctuations and/or non-stationarity of the measured EEG signals. One of the first methods that aims to robustify the feature extraction process by applying regularization to CSP was proposed in (Blankertz et al., 2008a). The method acquires information about the noise in the data by using recordings from additional sessions. A recently proposed algorithm called stationary CSP (sCSP) (Samek et al., 2012b) applies a data-driven strategy to penalize non-stationary directions in the spatial filter computation process. A similar method that uses Kullback-Leibler divergence to measure the changes in the data has been proposed in (Arvaneh et al., 2013). Other approaches (Lu et al., 2010; Lotte and Guan, 2010; Kang et al., 2009; Samek et al., 2013) use inter-subject information to regularize the solution towards stationarity and robustness. Two-step approaches (von Bünau et al., 2010; Samek et al., 2012a) have also been suggested for computing stationary features. They first estimate and remove the non-stationary contributions and apply CSP to the remaining part of the data in a second step.

In this manuscript we propose an approach to robustify CSP without using additional recordings or data from other users. Our method is inspired by (Kim et al., 2006), where a maxmin approach to Fisher discriminant analysis (FDA) was applied for robust classification. In particular, the maxmin FDA is guaranteed to have higher discriminative power for *any fluctuations within a prefixed tolerance set*. Following this philosophy of robustifying a method we develop a maxmin version of CSP and propose a novel approach to compute a tolerance set that captures the variability in the data. This tolerance set robustifies our CSP variant to non-stationarities in the data. In contrast to the FDA case (Kim et al., 2006), we show that for a certain class of tolerance sets the solution for the CSP maxmin problem can be derived analytically and computed as a generalized eigenvalue problem. In contrast to our prior conference contributions (Kawanabe and Vidaurre, 2009; Kawanabe et al., 2009), in this paper we present the maxmin CSP approach in detail (including proofs) and extensively evaluate it using a large number of BCI datasets. Furthermore, we compare maxmin CSP to a state-of-the-art method, stationary CSP, and evaluate the advantages and limitations of both techniques with simulations. Finally, we also thoroughly discuss and interpret the non-stationarities found in our data.

The paper is organized as follows: At first, we explain neurophysiological background and spatial filtering techniques for EEG classification tasks. In Section 3, we present the maxmin approach for robust Common Spatial Patterns. There, the following two maxmin CSP procedures are proposed: an algorithm with universal (data-independent) tolerance sets and an alternative with tolerance sets depending on the actual non-stationarity. In Section 4 we investigate the advantages and limitations of our method using simulations and discuss its relation to a state-of-the-art algorithm called stationary CSP. We evaluate and compare maxmin CSP to two baseline methods, namely CSP and stationary CSP, on a large BCI data set with 80 subjects in Section 5. Furthermore, we illustrate the robustness property of our method using data from a particular subject and discuss the non-stationarities in the data set. Finally, Section 6 concludes this work.

2 Spatial Filters for EEG Classification Tasks

2.1 Neurophysiologial Background

Many EEG-BCIs, are based on motor imagery. Commonly, participants using these systems are asked to perform the imagination of hands, feet or mouth movements. Motor imagery alters the rhythmic activity that can be measured in the EEG over the sensorimotor cortex. Many EEG rhythms are called idle rhythms because they are generated by large populations of cortical neurons that fire in rhythmical

synchrony when they are not engaged in a specific task. Oscillations with a fundamental frequency between 9 and 13 Hz can be observed over motor and sensorimotor areas in most subjects (the μ-rhythm). These sensorimotor rhythms (SMRs) are attenuated in the corresponding cortical area when a motor task (e.g. movement or motor imagery) takes place. As this effect is due to loss of synchrony in the neural populations, it is termed event-related desynchronization (ERD). The increase of oscillatory EEG (i.e., the reestablishment of neuronal synchrony) is called event-related synchronization (ERS), see (Pfurtscheller and da Silva, 1999).

For distinguishing motor imagery tasks of different body parts it is necessary to recognize the different spatial localizations of SMR modulations. The locations over the sensorimotor cortex are related to corresponding parts of the body. For example, left and right hand are localized in the contralateral hemisphere, i.e., right and left motor cortex, respectively. Thus, spatial filters are an essential step for a meaningful feature extraction and posterior classification of motor intentions. One of the most popular and successful algorithms for calculating spatial filters is Common Spatial Patterns (CSP). Given two distributions in a high-dimensional space (corresponding in our case to two different mental tasks), the CSP algorithm finds directions (i.e., spatial filters) that maximize variance for one class and simultaneously minimize variance for the other class (Blankertz et al., 2008c). Since computation of band-power is equivalent to the assessment of variance of band-pass filtered signals, this criterion reflects well the underlying physiology of ERD/ERS effects.

2.2 Common Spatial Pattern

Mathematically CSP analysis works as follows. Let $\overline{\Sigma}_+$ and $\overline{\Sigma}_-$ be the (average) covariance matrices of the band-pass filtered EEG signals of two different motor imagery tasks. These two matrices are simultaneously diagonalized such that the eigenvalues of $\overline{\Sigma}_+$ and $\overline{\Sigma}_-$ sum to 1. This can be done by calculating the generalized eigenvectors W:

$$\overline{\Sigma}_+ W = (\overline{\Sigma}_+ + \overline{\Sigma}_-) W D. \tag{1}$$

Here, the diagonal matrix D contains the (generalized) eigenvalues of $\overline{\Sigma}_+$ (defined such that they are between 0 and 1) and the column vectors of W are the filters w's for computing the CSP features. The best discrimination is provided by those filters with high eigenvalues (large variance for condition 1 and small variance for condition 2) and by filters with low eigenvalues (vice versa). Therefore, the common practice in a classification setting is to use several eigenvectors from both ends of the eigenvalue spectrum for feature computation. Alternatively, the solution for the eigenvector with the largest eigenvalue can also be obtained by maximizing the Rayleigh quotient:

$$\underset{w \in \mathbb{R}^C}{\text{maximize}} \qquad \frac{w^\top \overline{\Sigma}_+ w}{w^\top (\overline{\Sigma}_+ + \overline{\Sigma}_-) w}. \tag{2}$$

This correspondence is often useful for algorithmic considerations.

2.3 Stationary Common Spatial Pattern

The class covariance matrices $\overline{\Sigma}_+$ and $\overline{\Sigma}_-$ used in CSP can vary substantially because of non task-related fluctuations and/or non-stationarity of the EEG signals (Krauledat, 2008; Shenoy et al., 2006; von Bünau et al., 2009; Samek et al., 2012b; Grosse-Wentrup et al., 2011), as well as because of artefacts in the data. In BCI applications, it is thus important to robustify the features against such task unrelated fluctuations and artefactual trials. The stationary Common Spatial Patterns (sCSP) (Samek et al., 2012b) method regularizes the CSP solution towards stationarity, i.e. its goal is to compute filters that not only provide a large variance ratio between two conditions but at the same time it aims to keep the variance estimation along the projected direction as stable as possible across trials. In other words it prefers filters that constantly provide a high variance ratio over filters that focus on single events with very high variance ratio, e.g. electrode artefacts. For that the method computes penalty matrices Δ_+ and Δ_- as

$$\Delta_\pm = \frac{1}{K} \sum_{k=1}^{K} \mathcal{F}\left(\Sigma_\pm^{(k)} - \overline{\Sigma}_\pm\right), \tag{3}$$

where $\mathcal{F}$ is an operator to make symmetric matrices positive definite by flipping the sign of negative eigenvalues, N represents the number of trials and $\Sigma_{\pm}^{(k)}$ and $\overline{\Sigma}_{\pm}$ denote the covariance matrix of trial k and the average class covariance matrix, respectively. By adding this quantity to the denominator of the CSP objective function we penalize spatial filters that extract non-stationary features. The Rayleigh quotient summarizes to

$$\underset{w\in\mathbb{R}^C}{\text{maximize}} \qquad \frac{w^\top \overline{\Sigma}_+ w}{w^\top(\overline{\Sigma}_+ + \overline{\Sigma}_- + \lambda(\Delta_+ + \Delta_-))w}, \tag{4}$$

so that the sCSP filters can be computed by solving a generalized eigenvalue problem.

3 The Maxmin Approach to Spatial Filters

The maxmin approach (Kim et al., 2006) that was successfully applied to robustify FDA has inspired the current work for constructing robust CSP filters. The key idea is that, instead of employing only two single covariance matrices, we consider convex sets $\mathcal{S}_+$ and $\mathcal{S}_-$ for the class covariances Σ_+ and Σ_-, respectively. These sets, we call them tolerance sets, specify the tolerance regions of fluctuations around the class covariances. We define these sets as balls in the space of $C \times C$ symmetric positive definite matrices centered at $\overline{\Sigma}_+$ and $\overline{\Sigma}_-$

$$\begin{aligned} \mathcal{S}_+ &= \left\{\Sigma_+ \middle| \Sigma_+ \succeq 0,\ \|\Sigma_+ - \overline{\Sigma}_+\| \leq \delta_+ \right\}, \\ \mathcal{S}_- &= \left\{\Sigma_- \middle| \Sigma_- \succeq 0,\ \|\Sigma_- - \overline{\Sigma}_-\| \leq \delta_- \right\}, \end{aligned} \tag{5}$$

where $\|\cdot\|$ denotes an appropriate norm of the matrix space and δ_+ and δ_- stand for the radii of the balls. One possibility of such norms is

$$\|X\|_P^2 := \mathrm{Tr}\left(P^{-1}XP^{-1}X\right) \tag{6}$$

for any symmetric matrix X, where P is a $C \times C$ symmetric positive definite matrix specifying the shape of the balls. This type of norms is derived from the Riemannian metric on the manifold of symmetric positive definite matrices and takes into account its intrinsic geometric structures (Ohara et al., 1996). Note that when $P = I$, we arrive at the standard 'Frobenius' norm. In the following, we call $\mathcal{S}_{\pm}$ with Frobenius norm as 'universal tolerance sets' (see Section 3.1). Another example of such norms is defined by a matrix PCA of the locally-averaged covariances. It leads to data-driven tolerance sets taking into accounts their non-stationary fluctuations (see Section 3.2). Once the tolerance sets $\mathcal{S}_{\pm}$ are fixed, based on the maxmin framework, robust CSP filters can be constructed by maximizing the worst case (minimum) Rayleigh quotient within all possible covariance matrices in the tolerance regions yielding the following optimization problems

$$\max_{w\neq 0} \quad \min_{\Sigma_+\in\mathcal{S}_+,\,\Sigma_-\in\mathcal{S}_-} \frac{w^\top \Sigma_+ w}{w^\top(\Sigma_+ + \Sigma_-)w}, \tag{7}$$

$$\max_{w\neq 0} \quad \min_{\Sigma_+\in\mathcal{S}_+,\,\Sigma_-\in\mathcal{S}_-} \frac{w^\top \Sigma_- w}{w^\top(\Sigma_+ + \Sigma_-)w}. \tag{8}$$

Note that if the radii of the tolerance sets shrink to zero, then this approach converges to standard CSP.

The idea behind maxmin CSP is illustrated in Fig. 1. The panels (a), (b) and (c) show covariance matrices Σ_+ and Σ_- of three different sessions. One can see that the data from these sessions show large variability, especially class 2 (solid line). The panel (d) shows schematically the space of symmetric positive definite matrices, where the covariance matrices in (a), (b) and (c) are represented as points (three $\times$ and three $\circ$). Classical CSP simply averages covariances $\overline{\Sigma}_+$ and $\overline{\Sigma}_-$. In contrast, in the maxmin framework, we consider ellipsoids (namely the tolerance sets $\mathcal{S}_+$ and $\mathcal{S}_-$) around the averages which can capture non-stationary fluctuations of the covariances Σ_+ and Σ_- to some extent. From both ellipsoids, a pair of the worst case covariances is obtained for each optimization problem (7) or (8). The maxmin CSP spatial filters are computed by maximizing the variance ratio (applying CSP) for this pair of covariance matrices.

At the end of this section we discuss why considering the worst case covariance matrices from the tolerance set leads to robust estimation. Before that we introduce two different maxmin CSP algorithms, one that uses an universal tolerance set and one that uses a tolerance set that captures the non-stationarities in the data.

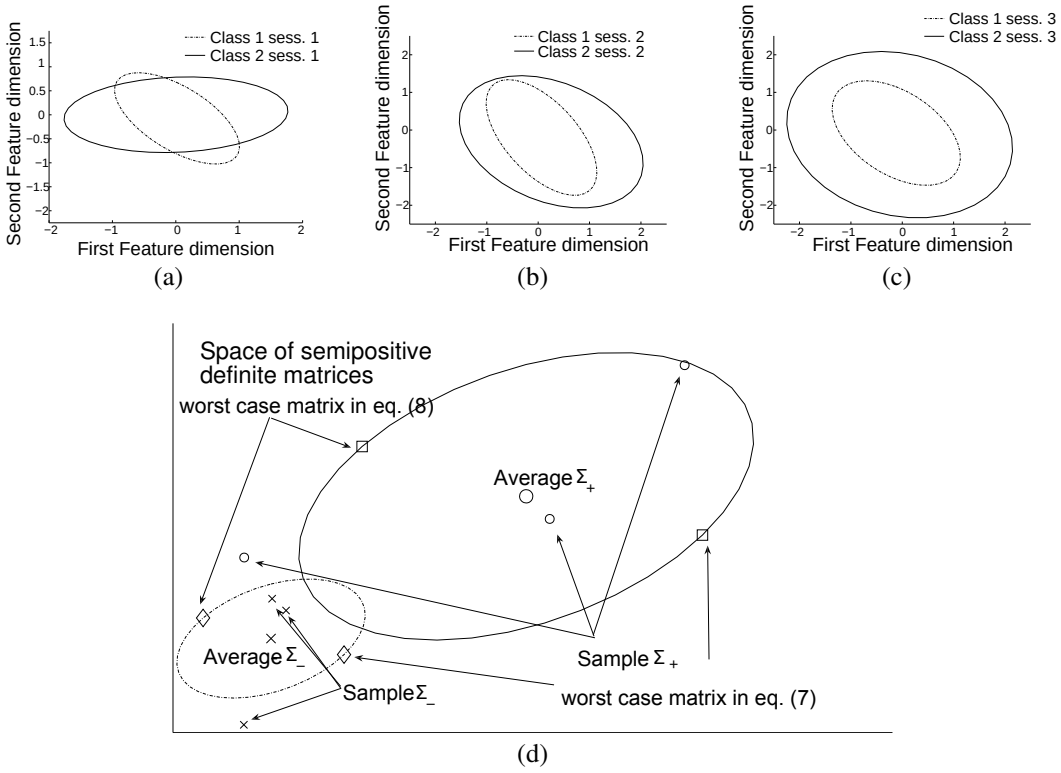

Figure 1: Figs. (a), (b) and (c) represent Σ_+ and Σ_- at different time points. Mean of the features is 0, as it is bandpass filtered data. Fig. (d) represents the previous matrices as points in the space of symmetric positive definite matrices. The ellipsoids in Fig. (d) are the tolerance sets S_+ and S_- centered at the average matrices $\overline{\Sigma}_+$ and $\overline{\Sigma}_-$, respectively. From both ellipsoids, a pair of the worst case covariances is obtained for each optimization problem (7) or (8).

3.1 maxmin CSP with Universal Tolerance Sets

A tolerance set is called universal if it is a sphere, i.e. it has the same extent in all directions form the center. Such sets can be constructed by using the Frobenius norm, i.e. that define in Eq. (6) with $P = I$. In the following we show that the worst-case covariances in the universal tolerance set can be determined explicitly, even in the general case, when the norms specifying the tolerance sets $\mathcal{S}_\pm$ are defined as Eq. (6) with $P = P_\pm$, respectively. The proof of Lemma 1 can be found in Appendix.

Lemma 1 *For the sets $\mathcal{S}_+$ and $\mathcal{S}_-$ defined in Eq.* (5) *with the norm Eq.* (6) *at $P = P_+$ and $P = P_-$, respectively, the worst case Rayleigh quotient becomes*

$$\frac{w^\top \left(\overline{\Sigma}_+ - \delta_+ P_+\right) w}{w^\top \left(\overline{\Sigma}_+ + \overline{\Sigma}_- - \delta_+ P_+ + \delta_- P_-\right) w}, \qquad \forall w, \tag{9}$$

if $\overline{\Sigma}_+ - \delta_+ P_+ \succeq 0$.

Thus the maxmin CSP approach with univeral tolerance sets can also be computed as generalized eigenvalue problem. Our prior work used this approach (Kawanabe and Vidaurre, 2009). It may be important to mention that the generalized eigenvalue problem (9) coincides with the regularized CSP (Lotte and Guan, 2011) when the radius δ_+ of the tolerance set $\mathcal{S}_+$ is zero. Therefore, application of maxmin principle to spatial filter add a new viewpoint to the regularized CSP, which can help to design the regularization matrix.

3.2 maxmin CSP with Data-driven Tolerance Sets

Although the maxmin CSP with the identity matrix improved the original CSP in our experiments (Kawanabe and Vidaurre, 2009), the corresponding tolerance sets do not well capture the actual variability of the covariance matrices over time. We conjecture that the performance increase is analogous to the fact that Bayesian regularization helps even with non-informative priors. If we have extra (prior) information about possible fluctuations as is the case with the real world BCI data in (Blankertz et al., 2008a), the covariance of the distortions may be used for the matrices P_+ and P_- and the problem can be solved as a generalized eigenvalue problem by applying Eq. (9). If we do not have such extra knowledge we may still analyze the non-stationarity in the data and to take such information into account for calculating robust filters. In (Kawanabe et al., 2009), we briefly introduced a variant of the maxmin CSP with tolerance sets that capture exactly this variability. In this section, we will explain it more in detail. The algorithm for this maxmin variant consists of three steps: 1) construction of data-driven tolerance sets using matrix PCA, 2) computation of the worst case covariances and 3) derivation of the maxmin filters. Ideally, we should alternate the step 2) and 3) until convergence (see Algorithm 1).

In the first step we construct the tolerance sets by using matrix PCA. Krauledat (Krauledat, 2008) analyzed session-to-session variability of BCI data based on this technique. Let $\{\Sigma_\pm^{(k)}\}_{k=1}^K$ be sets of locally-averaged covariance matrices in time, where in our experiment we will use trial-wise covariances (without averaging) and local averages using a tenth of the total data. We would like to fit the tolerance regions (the ellipsoids in Fig. 1.(d)) so that they match the variability of the covariances $\{\Sigma_\pm^{(k)}\}_{k=1}^K$. To do so, we need to find directions of large fluctuation, which can be done by PCA. At first, $C \times C$ matrices are transformed to C^2-dimensional vectors by stacking all the column vectors into one big column vector (the vectorization operation denoted often with vec). Then, the covariance of the extended vectors for each class

$$\frac{1}{n-1}\sum_{k=1}^{K} \mathrm{vec}\left(\Sigma_\pm^{(k)} - \overline{\Sigma}_\pm\right) \mathrm{vec}\left(\Sigma_\pm^{(k)} - \overline{\Sigma}_\pm\right)^\top \tag{10}$$

is calculated and its eigen decomposition is obtained

$$\left[\frac{1}{n-1}\sum_{k=1}^{K} \mathrm{vec}\left(\Sigma_\pm^{(k)} - \overline{\Sigma}_\pm\right) \mathrm{vec}\left(\Sigma_\pm^{(k)} - \overline{\Sigma}_\pm\right)^\top\right] \boldsymbol{v} \;=\; \lambda \boldsymbol{v}, \tag{11}$$

$$\lambda \;=\; \frac{1}{n-1}\sum_{k=1}^{K} \left\{\mathrm{vec}\left(\Sigma_\pm^{(k)} - \overline{\Sigma}_\pm\right)^\top \boldsymbol{v}\right\}^2,$$

where λ and $\boldsymbol{v}$ with $\|\boldsymbol{v}\| = 1$ denote an eigenvalue and eigenvector of the extended covariance in Eq. (10). Finally, the C^2-dimensional eigenvectors are transformed back to $C \times C$ matrices by finding V such that $\boldsymbol{v} = \mathrm{vec}(V)$ (the inverse operation of the vectorization vec). Suppose that $\lambda_\pm^{(i)}$ and $V_\pm^{(i)}$ are eigenvalues and matrices by class-wise PCA. Let us express general covariance matrices as

$$\Sigma_\pm \;=\; \overline{\Sigma}_\pm + \Delta_\pm \;=\; \overline{\Sigma}_\pm + \sum_i \alpha_\pm^{(i)} V_\pm^{(i)}$$

and define a PCA-based norm for each class by

$$\|\Delta_\pm\|_{\mathrm{PCA}}^2 := \sum_i \frac{\left(\alpha_\pm^{(i)}\right)^2}{\lambda_\pm^{(i)}}. \tag{12}$$

This norm allows larger variations in the directions with large eigenvalues. We also remark that this cannot be transformed into the form (6) and the maximin principle leads to a spatial filtering method quite different from the regularized CSP. Thus, the tolerance sets with the PCA norm

$$\mathcal{S}_\pm \;:=\; \left\{\Sigma_\pm = \overline{\Sigma}_\pm + \Delta_\pm \;\middle|\; \Delta_\pm = \sum_i \alpha_\pm^{(i)} V_\pm^{(i)},\;\; \|\Delta_\pm\|_{\mathrm{PCA}} \le \delta_\pm\right\}, \tag{13}$$

become ellipsoid which fit nicely to the distributions of $\{\Sigma_\pm^{(k)}\}_{k=1}^K$.

The second step of our algorithm consists of identifying the worst case covariances in the tolerance sets for fixed w. As described in the proof of Lemma 1 (see Appendix), the worst case covariances in (7) can be obtained by optimizing separately

$$\min_{\Sigma_+ \in \mathcal{S}_+} w^\top \Sigma_+ w = \min_{\boldsymbol{\alpha}_+} \sum_i \alpha_+^{(i)} w^\top V_+^{(i)} w, \tag{14}$$

$$\max_{\Sigma_- \in \mathcal{S}_-} w^\top \Sigma_- w = \max_{\boldsymbol{\alpha}_-} \sum_i \alpha_-^{(i)} w^\top V_-^{(i)} w, \tag{15}$$

under the norm and the positive definiteness constraints

$$\|\Sigma_\pm - \overline{\Sigma}_\pm\|_{\mathrm{PCA}}^2 = \sum_i \frac{\left(\alpha_\pm^{(i)}\right)^2}{\lambda_\pm^{(i)}} \leq \delta_\pm^2, \tag{16}$$

$$\Sigma_\pm = \overline{\Sigma}_\pm + \sum_i \alpha_\pm^{(i)} V_\pm^{(i)} \succeq 0. \tag{17}$$

If we ignore the condition (17), the solutions of the constrained optimization problems can be obtained analytically as

$$\alpha_+^{(i)} = \frac{-\delta_+ \lambda_+^{(i)} w^\top V_+^{(i)} w}{\sqrt{\sum_i \lambda_+^{(i)} \left(w^\top V_+^{(i)} w\right)^2}}, \tag{18}$$

$$\alpha_-^{(i)} = \frac{\delta_- \lambda_-^{(i)} w^\top V_-^{(i)} w}{\sqrt{\sum_i \lambda_-^{(i)} \left(w^\top V_-^{(i)} w\right)^2}}. \tag{19}$$

When $\overline{\Sigma}_\pm + \sum_i \alpha_\pm^{(i)} V_\pm^{(i)}$ violate (17), we truncate the negative eigenvalues of the worst case covariances to zero (Laub and Müller, 2004). Note that the worst case covariances in (8) can be obtained in an analogous way.

Finally, once the worst case covariances are obtained, we can update the maxmin filters by maximizing the variance ratio (applying CSP) of these matrices. In contrast to the previous case (Eq. (9)), we need to iterate the second and third steps, since the worst case covariances depend on the current filter w. However, in our experiments a single update from the original CSP works sufficiently well.

Algorithm 1 maxmin CSP with data-driven tolerance sets for the optimization problem (7)

1: **Input**: $\overline{\Sigma}_\pm,\ \{\Sigma_\pm^{(k)}\}_{k=1}^K,\ \delta_\pm$
2: Compute the eigen decomposition $\{\lambda_\pm^{(i)}, V_\pm^{(i)}\}_i$ from the set of covariances $\{\Sigma_\pm^{(k)}\}_{k=1}^K$ for each class by the matrix PCA (11).
3: Compute the ordinal CSP w_0 by Eq. (2) for the initial filter.
4: **repeat**
5: For the current filter w, find the worst-case covariance matrices in the tolerance sets $\mathcal{S}_\pm$ (13), i.e. calculate the coefficients $\alpha_\pm^{(i)}$ in $\Sigma_\pm = \overline{\Sigma} + \sum_i \alpha_\pm^{(i)} V_\pm^{(i)}$ by Eqs. (18) and (19).
6: Make the worst-case covariances $\Sigma_\pm = \overline{\Sigma} + \sum_i \alpha_\pm^{(i)} V_\pm^{(i)}$ non-negative definite by truncating their eigenvalues at 0, if necessary.
7: Compute CSP with the worst case covariances $\Sigma_\pm$ obtained above and update w.
8: **until** convergence
9: **Output**: w

3.3 Why is maxmin CSP robust?

This subsection discusses the robustness property of the maxmin approach. Although it seems counterintuitive to compute spatial filters from the worst covariance matrices in the tolerance region, this approach

increases robustness. There are two reasons for this, namely an implicit shrinkage of the sample covariance matrix and the separation of intrinsic variability and outlier effects.

From Eq. (9) one can see that maximizing the worst covariance matrices from a tolerance region can be interpreted from a shrinkage perspective (see (Ledoit and Wolf, 2004; Lotte and Guan, 2011)). Although it is not equivalent to standard shrinkage methods, the underlying idea is very similar. It is well known (Ledoit and Wolf, 2004) that the largest eigenvalues of the empirical covariance matrix are overestimated, whereas the smallest ones are underestimated. This poses a severe problem for CSP as it maximizes the variance between two conditions, i.e. over- and underestimation effects may largely influence the solution. Our maxmin CSP approach reduces the (relative) impact of the largest eigenvalues in the numerator of the Rayleigh quotient by subtracting $\delta_+ P_+$ from $\overline{\Sigma}_+$. Furthermore it reduces the (relative) impact of the smallest eigenvalues in the denominator by computing $\overline{\Sigma}_- + \delta_- P_-$. Note that this effect stabilizes the Rayleigh quotient and is mainly responsible for the robustness of our CSP variant.

The second robustness effect can be attributed to the way how the tolerance regions are computed. Note that the size of the tolerance region determines the trade-off between robustness and optimality of the solution. If the size shrinks to zero we arrive at the standard CSP solution and lose the robustness property, whereas if it is too large then the CSP estimation is becomes very poor. By optimizing the solution within a tolerance region of certain shape and extend we implicitly separate the intrinsic variability in the data and outlier effects. In other words the tolerance sets capture some of the variability present in the data and ignore the trials that are too far away from the average covariance matrix. This separation between natural noise that should be included in the computation and outliers that should be excluded from it further increases robustness and is neither performed by standard CSP nor by sCSP. Note that the way the tolerance region is computed also affects the robustness of the algorithm. In this paper we propose to construct it by estimating the non-stationary directions, however, other ways of constructing it could be considered as well.

There is also an intuitive reason why maximizing the worst solution is advisable. The maxmin CSP provides a cautious estimate of the band power ratio between two conditions in the sense that the spatial filters perform well for all trials that fall into the tolerance region. Thus it reduces the tendency to overfit, i.e. relying on few very strong (artefactual) trials that may dominate the sample covariance matrix. Note that if we would maximize the maximum Rayleigh quotient, then we would most likely arrive at solutions that concentrate on few (artefactual) events with a very high power ratio but do not generalize well. In contrast, maximizing the minimum prefers robust solutions in the tolerance set.

4 Simulations

This section investigates the robustness of maxmin CSP in two scenarios, namely one artefact and one non-stationarity scenario. In the light of the results we discusses the advantages and limitations of our method with respect to sCSP.

4.1 Robustness of maxmin CSP

We generate a toy data set containing training and test recordings. In more detail, a 10-dimensional signal $x(t)$ is generated as noisy mixture of two discriminative $s^d(t)$ and eight non-discriminative $s^n(t)$ sources

$$x(t) \quad = \quad A \begin{bmatrix} s^d(t) \\ s^n(t) \end{bmatrix} \; + \; \epsilon, \tag{20}$$

where A is a random rotation matrix and ϵ an i.i.d. noise term sampled from a zero mean Gaussian distribution $\mathcal{N}(0, 2)$ with variance 2. The first discriminative source is sampled from $\mathcal{N}(0, 1.8)$ in condition 1 and $\mathcal{N}(0, 0.2)$ in condition 2. The second discriminative source is sampled from $\mathcal{N}(0, 0.6)$ in condition 1 and $\mathcal{N}(0, 1.4)$ in condition 2. The non-discriminative sources are sampled from $\mathcal{N}(0, 1)$ irrespective of condition. We sample 200 data points per trial and 50 trials per condition for the training data set and 50 trials per condition for the test data set. Note that all sources generate data with an average variance of 1. In order to study the robustness of our algorithm to artefacts we add very strong noise to the data with a small probability of 0.01. We add the artefacts by sampling ϵ from $\mathcal{N}(0, 30)$.

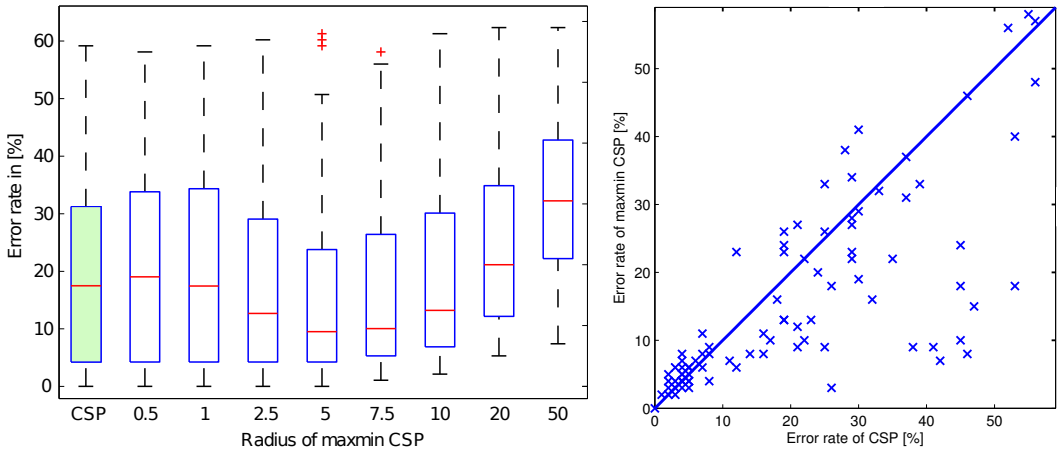

Figure 2: Application of maxmin CSP to toy data containing artefacts. Left Panel: Boxplot showing error rates of CSP (green box) and maxmin CSP with different radii. Right Panel: Scatter plot of CSP and maxmin CSP (radius 5) error rates. The maxmin CSP approach outperforms CSP for all points lying below the solid line.

The left panel of Fig. 2 shows the classification error rates for CSP (green box) and maxmin CSP with universal tolerance sets with varying radius (x-axis). The results are based on 100 repetitions. Since we have two discriminative sources we filter the data with two spatial filters and compute log-variance features and train a Linear Discriminant Analysis (LDA) classifier. One can see from the figure that for small radii the maxmin CSP and CSP performance coincides, whereas maxmin CSP is more robust for a certain radius, but the error rates increase when considering a too large tolerance set. The performance difference between CSP and maxmin CSP with radius 5 is highly significant when applying the one-sided Wilcoxon signed rank test, the p-value is 0.0009. The right panel of Fig. 2 shows a scatter plot for CSP and the best maxmin CSP (radius 5). Each cross in the plot represents one run in our experiment, the x-axis denotes the CSP error rates and the y-axis shows the maxmin CSP error rates. Thus our method outperform CSP for all points lying below the solid line.

4.2 Simulations with real EEG

Here we evaluate the proposed maxmin CSP with constructed data from real EEG. We show that maxmin CSP is not only robust to artefacts, but also to non-stationary changes in the data.

Maxmin CSP was trained on motor imagery data, using tolerance sets based on the PCA approach and without using any a priori information on the type of disturbance present in the test data. The test data was constructed using motor imagery data to which activity related to increased occipital alpha, more specifically a recording where the user had his eyes closed, was added with a back-projection of 5 ICA components and using 3 different factors (0.5, 1 and 2). The right panel of Fig. 3 displays the classifier output obtained classifying the constructed test data. The performance of the original CSP is severely deteriorated with increased alpha mixed. In contrast, the proposed maxmin CSP method shows a more stable performance, even when the filters were computed without using information on the type of noise added on the test set. As expected, when no stationarities or noise are found ($\alpha = 0$) the maxmin CSP solution performs slightly worse than CSP, with 11.8% of error versus 10%. However, when disturbances appear, the performance of maxmin CSP is clearly better than that of CSP.

On the left of Fig. 3 the pattern from original CSP (top) and maxmin CSP (down) are shown. The pattern of the original CSP has positive weights at the right occipital side which might be susceptible to modulations on the alpha band, whereas the corresponding maxmin CSP has much lower weights in the same area, and therefore is more robust against this noise. In conclusion, the maxmin CSP approach offers a more robust solution than CSP in case of non-stationarities but also works properly even if the

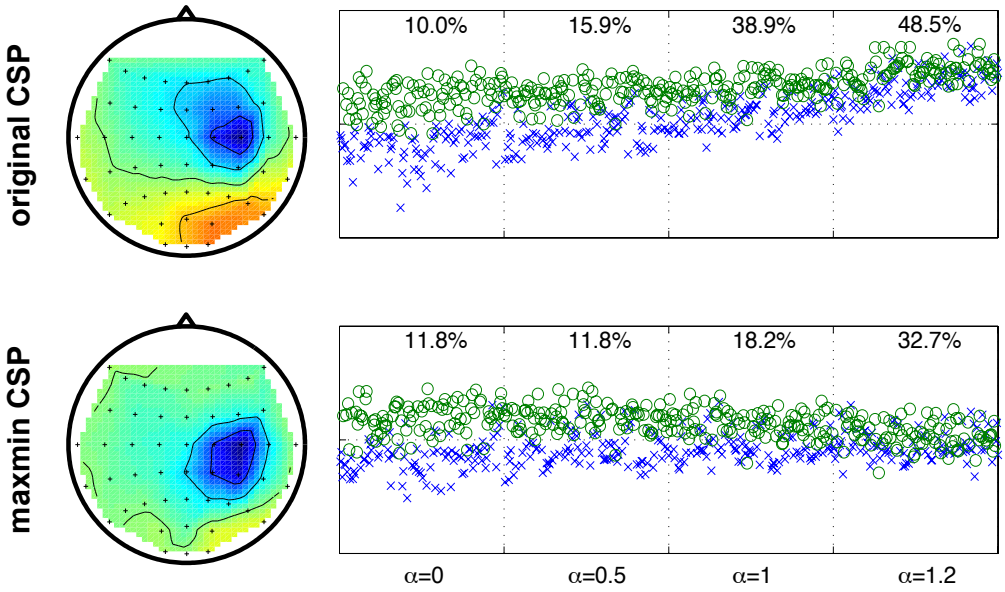

Figure 3: Comparison of CSP (top) and maxmin CSP (bottom) on EEG data with artificially increased occipital alpha. The plots on the right show the classifier output on the simulation data where different degrees of alpha have been added in time (factors 0, 0.5, 1, 2). The plots on the left show the pattern coefficients topographically mapped on the scalp from original CSP (top) and maxmin CSP (bottom). The non-stationarity is represented by an increase in the alpha activity in the visual cortex (occipital location) that was added using an eyes open/eyes closed recording.

data is not very variable and without including information about the type of noise that will be present in the future.

4.3 Relations to sCSP

Stationary CSP has also been shown (Samek et al., 2012b) to robustify the solution against artefacts in the data and to increase stationarity of the features, thus from a conceptual point of view both methods are very similar.

One principle difference between both methods is that maxmin CSP performs optimization on part of the data only, i.e. it disregards everything as outliers that is too far away from the center of the tolerance regions. This has the advantage that it better captures the intrinsic variability of the data. Furthermore the maxmin CSP solution can be regarded as the CSP solution after applying a particular kind of shrinkage to the covariance matrices. Thus it regularizes the covariance matrices and not the spatial filters, therefore the solution has a clear interpretation.

On the other hand, stationary CSP computes the non-stationarity term from all trials, i.e. it does not differentiate between intrinsic variability and outliers. It also applies regularization to the spatial filters not to the covariance matrix and uses a heuristic, namely flipping the sign of negative eigenvalues. Note that this heuristic may fail[1], i.e. it does not capture the relevant non-stationarity in certain cases.

[1] This example comes from private communication with the authors of (Arvaneh et al., 2013).

For instance, assume we have the following matrices

$$\overline{\Sigma}_+ = \begin{bmatrix} 0.9 & 0.15 \\ 0.15 & 0.1 \end{bmatrix}, \overline{\Sigma}_- = \begin{bmatrix} 0.1 & 0 \\ 0 & 0.9 \end{bmatrix}$$

$$\Sigma_+^{(1)} = \begin{bmatrix} 0.9 & 0.05 \\ 0.05 & 0.1 \end{bmatrix}, \Sigma_+^{(2)} = \begin{bmatrix} 0.9 & 0.25 \\ 0.25 & 0.1 \end{bmatrix}.$$

If we aim to maximize the ratio between the variance of class $+$ and $-$ and simultaneously want to minimize non-stationarity then we should use the filter $w = [1\ \ 0]^\top$. Considering the class differences in the off-diagonal elements of $\overline{\Sigma}_+$ and $\overline{\Sigma}_-$ leads to a higher Rayleigh quotient (therefore it is preferred by CSP), but introduces non-stationarity between trials. In our example the sCSP penalty matrix is

$$\Delta = 0.5 \cdot \mathcal{F}\left(\Sigma_+^{(1)} - \overline{\Sigma}_+\right) + 0.5 \cdot \mathcal{F}\left(\Sigma_+^{(2)} - \overline{\Sigma}_+\right) = \begin{bmatrix} 0.1 & 0 \\ 0 & 0.1 \end{bmatrix},$$

thus it does not penalize the off-diagonal elements. Therefore neither CSP nor sCSP will extract the filter $w = [1\ \ 0]^\top$.

Our maxmin CSP method captures the variability of the covariance matrices of class $+$ and (with appropriate radii) is able to extract the stationary filter $w = [1\ \ 0]^\top$. Note that this filter is the maxmin solution in the tolerance set created by PCA because P_- is the zero matrix (no changes between trials) and $P_+ = \begin{bmatrix} 0 & \alpha \\ \alpha & 0 \end{bmatrix}$ reduces to a line. Since removing the off-diagonal elements from $\overline{\Sigma}_+$ decreases the Rayleigh quotient $w = [1\ \ 0]^\top$ will be the maxmin solution. In other words maximizing the minimum translates to minimizing the impact of the off-diagonal elements in this example. Thus preferring the maxmin solution in this case means to prefer the spatial filter that does not profit from the (non-stationary) class differences in the off-diagonal elements.

5 Experimental Results

In this section we provide an offline evaluation of the proposed algorithm on a high number of BCI users where transfer between calibration and feedback occurs. We compare the maxmin approach with data-driven tolerance sets to two baseline methods, namely CSP and stationary CSP. Furthermore we analyse the robustness property of our method and discuss the non-stationarities in the data.

5.1 Dataset and Experimental Setup

The data were recorded by the University of Tübingen and the Berlin Institute of technology in a common project. The datasets used for this study consist of a calibration and a feedback session recorded in a single day for each of 80 BCI users (Blankertz et al., 2010). The volunteers performed motor imagery first in a calibration session and then in a feedback operation in which they had to control a 1D cursor application. Coarsely, three categories of users were observed: users for whom (I) a classifier could be successfully trained and who performed feedback with good accuracy; (II) a classifier could be successfully trained, but feedback did not work well (because there are changes between the calibration and the feedback step that can affect the EEG signals, making the feedback fail); (III) no classifier with acceptable accuracy could be trained after the calibration. Whereas users of category II have difficulties with the transition from off-line to on-line operation, participants of category III do not show the expected modulation of sensorimotor rhythms (SMRs): either no idle SMR was observed over motor areas, or it was not attenuated during motor imagery.

The EEG signal was recorded from 119 Ag/AgCl electrodes, band-pass filtered between 0.05 and 200 Hz and downsampled to 100 Hz. After discarding the electrodes situated on the laterals, and close to the forehead and neck a set of 85 electrodes densely covering the motor cortex was used. Then, an estimate of the most discriminative frequency band and time segment for each subject was individually selected as done in (Blankertz et al., 2008c). After filtering the data spatially with four CSP, sCSP or maxmin CSP filters (two per class) we compute the log-variance features and train a Linear Discriminant Analysis

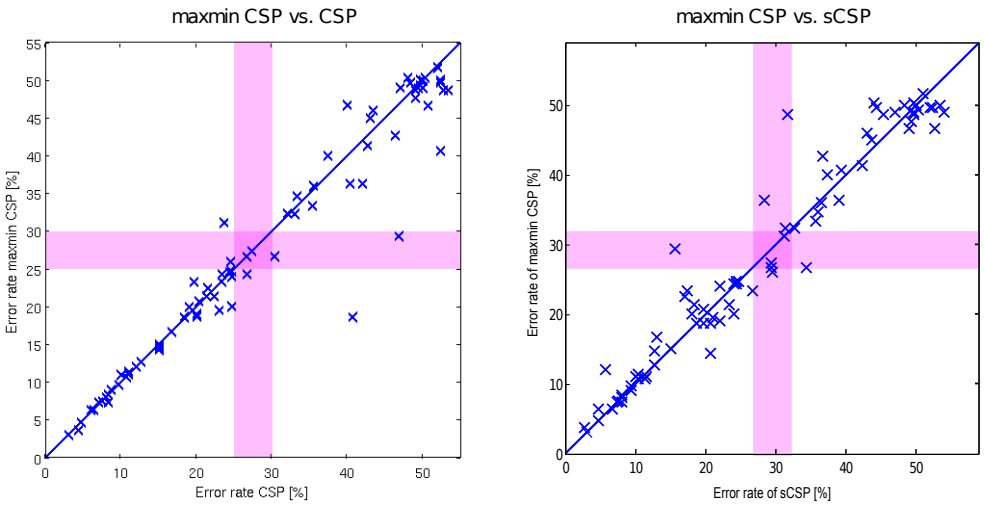

Figure 4: Left Panel: Scatter plot of error rates from CSP features (x-axis) versus maxmin CSP features (y-axis). The shaded pink areas respresent the error rage where BCI users start to feel control of the interface (for both conditions). All values below the diagonal indicate that maxmin CSP outperforms CSP. Right Panel: Scatter plot of error rates from sCSP features (x-axis) versus maxmin CSP features (y-axis).

(LDA) classifier. The hyperparameters of maxmin CSP, δ_+ and $\delta-$, were determined from $[0, \ldots, 1]$ by 10-fold cross-validation on the training set. A local set of matrices is computed either from single trials or from groups of 15 trials each. The selection of either one is as well performed in the training data (within the cross-validation procedure) as an additional hyperparameter. The regularization parameter λ of sCSP were determined from the set $[0, 2^{-8}, \ldots, 2^{-1}, 2^{0}]$ (as in (Samek et al., 2012b)) by 10-fold cross-validation on the training set. Also here we computed the penalty matrix from a local set of matrices from single trials or from groups of 15 trials and selected the best option within the cross-validation procedure.

5.2 Performance Comparison

The left panel of Fig. 4 depicts the scatter plot of errors obtained with maxmin CSP (y-axis) versus standard CSP (x-axis). The pink shaded band identifies a range of the error rate which divides users with and without BCI control. One can clearly see that the maxmin approach helps users without BCI control (error rate $>$30%) during the feedback application. This improvement is significant according to a Wilcoxon test (one sided, p-value=0.027). Note that users who gain most when using maxmin CSP features are those with low accuracy values. The right panel of this figure shows the same results for maxmin CSP vs. sCSP. Since both approaches penalize non-stationarity we can not improve over the sCSP results on average, the p-value of the one sided Wilcoxon test is 0.3343. So what is the advantage of using maxmin CSP ?

In Section 4.3 we showed that the sCSP heuristic may fail in certain cases, whereas maxmin CSP has a proper interpretation in terms of shrinkage and relies on a solid and well understood mathematical ground (maxmin principle). Furthermore the maxmin CSP solution is guaranteed to perform well on the whole tolerance region, e.g. for most training trials, but this may be not true for the sCSP filters. So if one needs to guarantee a certain minimum level of performance, then maxmin CSP is the method of choice. Finally we expect maxmin CSP to perform well in small-sample settings due to its regularization property, i.e. the over- and underestimation of the largest and smallest eigenvalue is much larger in small-sample settings.

5.3 Reduction of Non-Stationarity

In the following we investigate why maxmin CSP improves classification performance over the CSP baseline. Figure 5 reveals that the maxmin CSP patterns computed from the training set (middle row) are

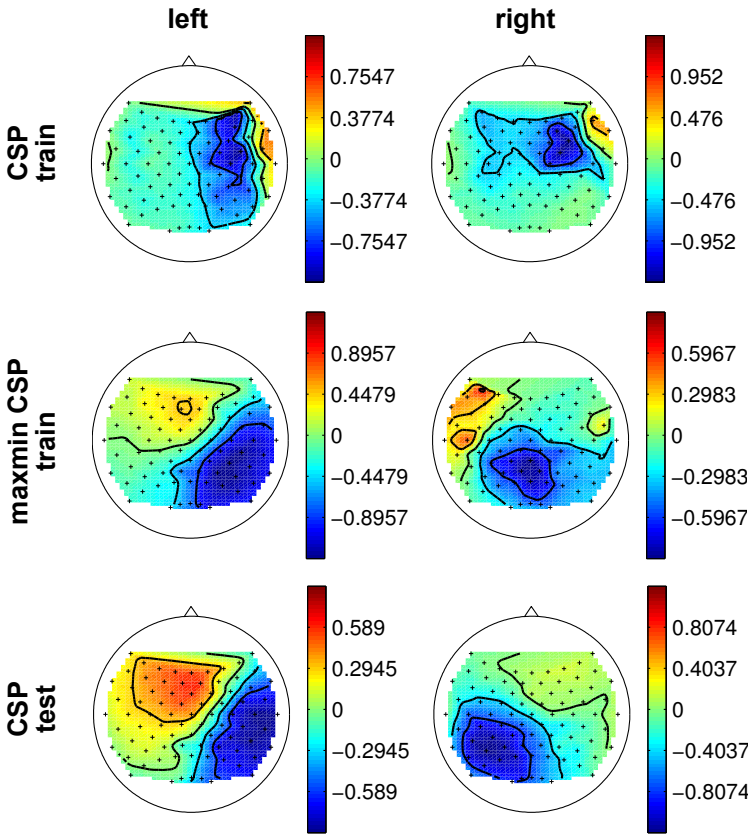

Figure 5: CSP patterns for one user that benefited from using maxmin CSP. Top row: CSP patterns of the training set. Middle row: maxmin CSP patterns of the training set. Bottom row: CSP patterns from the test set.

much more similar to those of the feedback data (bottom row) than the ones obtained with the standard CSP method (top row). A visual inspection of the CSP patterns shows that they do not well represent BCI-related activity, but are rather influenced by artefacts in the frontal and temporal electrodes. The maxmin patterns on the other hand are not estimated from the average class covariance matrices (that may be affected by outliers), but from a tolerance region that captures the intrinsic variability of the data but discards artefacts. Since the maxmin filters work well for all covariance matrices within the tolerance regions, they do not overfit to single artefactual events. Therefore they are more robust to outliers in the frontal and temporal electrodes, thus better capture the underlying BCI-related activity that is also present in test data. By computing the distance (with the dot product) from the filters computed using the training data (maxmin CSP and CSP) to the filters computed on the testing data (CSP), one can also quantify the increased robustness and invariance to changes of the maxmin CSP filters.

Thus we have showed that filters obtained with maxmin CSP generalize better, i.e. that they are more similar to those of the test data than the CSP traditional features. This increased robustness against noise and non-stationarity is especially helpful for BCI users with low BCI control. To investigate this further, we compute the Kullback-Leibler divergence[2] between the training and test feature distribution for maxmin CSP and CSP features. These values are presented as averages in Table 1 and as a function of time in the upper row of Figure 6.

[2]The Kullback-Leibler Divergence between two Gaussians $\mathcal{N}_1(\mu_1, \Sigma_1)$ and $\mathcal{N}_2(\mu_2, \Sigma_2)$ is defined as $D_{\mathrm{KL}}(\mathcal{N}_0 \| \mathcal{N}_1) = \frac{1}{2}\left(\mathrm{tr}\left(\Sigma_1^{-1}\Sigma_0\right) + (\mu_1 - \mu_0)^\top \Sigma_1^{-1}(\mu_1 - \mu_0) - \ln\left(\frac{\det \Sigma_0}{\det \Sigma_1}\right) - k\right)$.

Table 1: Kullback-Leibler divergence between training and testing features for CSP and maxmin CSP. The smaller the value, the most similar are the features between the training and the testing sets.

	CSP	maxmin CSP
Class 1	2.20	1.34
Class 2	2.62	2.40

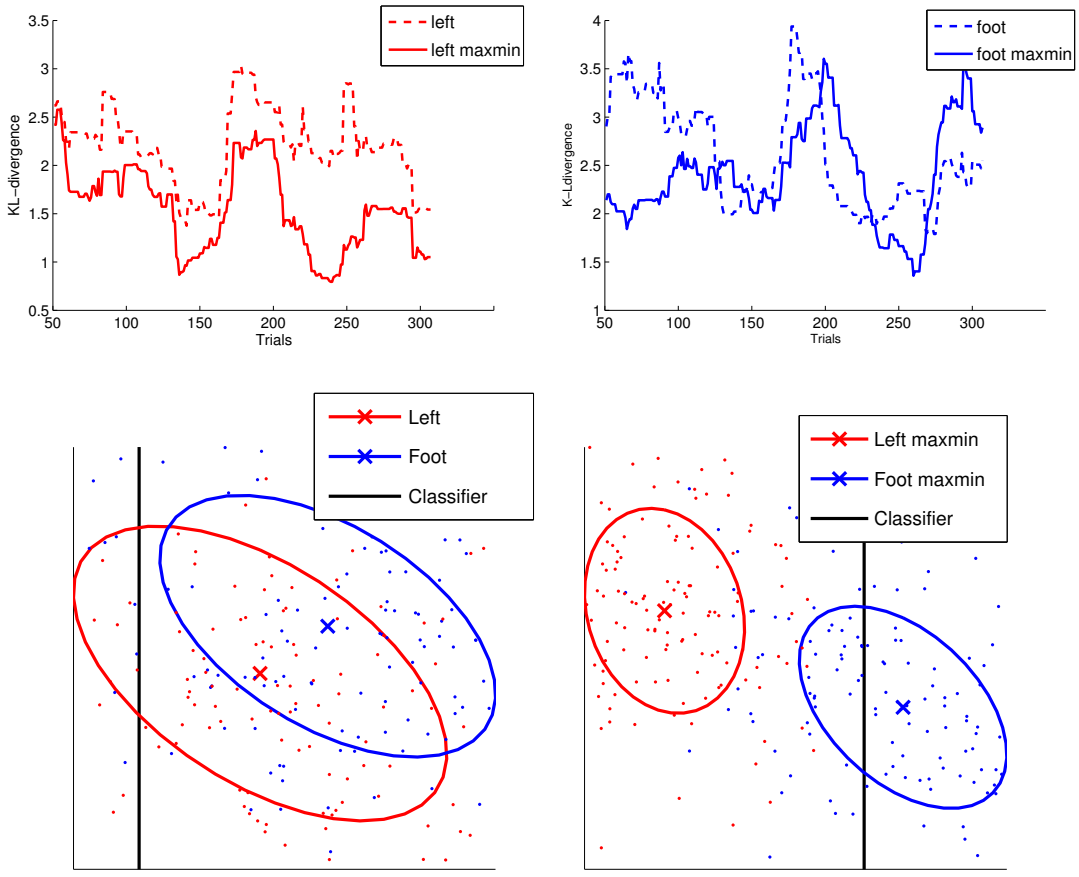

Figure 6: Upper row: Kullback-Leibler divergence between training features and those within sliding windows of testing sessions for CSP and maxmin CSP and the classes 'left hand' (left panel) and 'foot' (right panel). Bottom row: Features projected on the plane spanned by the normal vector of the corresponding classifier's hyperplane and the largest PCA-component of the feedback data. The black line is the classifier trained with the calibration data. The features are computed by CSP (left panel) and maxmin CSP (right panel).

From Table 1 we observe that the training features computed with maxmin CSP are more similar to those of the test set than those for classical CSP. Thus, maxmin CSP has robustified the features against the non-stationarity of the data (see also discussion of stationarizing nonstationary data (von Bünau et al., 2009; Sugiyama and Kawanabe, 2011)). However, the Kullback-Leibler divergence does not provide information about the separability of the test features. This is available in the bottom row of Figure 6 which is calculated by projecting the data on the plane spanned by the normal vector of the corresponding classifier's hyperplane and the largest PCA-component of the feedback data. It shows that the maxmin CSP test features (right panel) are more separable than those of the usual CSP (left panel). Note that the shift in the classifier bias can be easily adjusts by applying unsupervised bias adaptation methods

(Vidaurre et al., 2011a). However, it can not be adjusted when using CSP as a rotation in the feature distributions occur.

6 Conclusions

BCI data is contaminated by a variety of noise sources, artifacts, non-stationarities and outliers that make it indispensable to strive for more robust learning methods. In this paper we proposed a novel algorithm for robust spatial filtering drawing inspiration from (Kim et al., 2006).

In particular, we analyze the worst case performance among possible class covariance matrices and optimize the respective CSP-like filters based on such a robust criterion. We assume ellipsoids in matrix space for the sets of covariances, then the algorithm can be elegantly reduced to a generalized eigenvalue problem similar to the original CSP, but with modified covariance matrices. The simulations presented in this paper show that the maxmin CSP framework is indeed more robust as it allows to better transfer BCI classifier knowledge from the calibration to the feedback sessions, allowing users with low BCI control to achieve a better performance.

It becomes an important research direction to improve the robustness and usability of BCI systems. For instance, there exists a certain amount of users who cannot generate discriminative task-related brain signals in every trial. Recently, Sannelli et al. (Sannelli et al., 2009) develop a method to prune such undesirable trials by comparing true labels and their predictions. The authors computed CSP only with reliable trials which can improve BCI classification performance and stabilize the features at the same time. Our method strives for the same goal by an overall robustness applying a mathematical construction, the maxmin principle; there is no need to estimate the reliability of each of the trials separately.

In future work we will investigate several other ways for estimating the tolerance sets. Although the current maxmin CSP algorithms assume ellipsoids as the tolerance sets, in principle, it is possible to consider non-convex sets with cluster structures which may be closer to actual variabilities of training data. We will investigate whether such extension can improve BCI performances. Furthermore, instead of capturing within-session non-stationarity one could also focus on the shift between training and test distributions by applying matrix PCA (e.g. Krauledat, 2008) to the difference between the covariance matrix estimated on training data and the one estimated on test data (assuming some small amount of test data is available). Such strategy would potentially provide spatial filtersl that are robust with respect to session-to-session variations. In order to obtain subject-independent solutions one could also apply maxmin CSP with tolerance sets capturing the variability between subjects.

Acknowledgement

This work was supported in part by the German Research Foundation (GRK 1589/1), in part by the Federal Ministry of Education and Research (BMBF) under the project Adaptive BCI (FKZ 01GQ1115), in part by the World Class University Program through the National Research Foundation of Korea funded by the Ministry of Education, Science, and Technology, under Grant R31-10008 and in part by the Japan Science and Technology Agency (German-Japanese cooperation program on computational neuroscience), the Ministry of Education, Culture, Sports, Science and Technology (Grant-in-Aid for Scientific Research B, 24300093) and by the Ministry of Internal Affairs and Communications. This publication only reflects the authors views. Funding agencies are not liable for any use that may be made of the information contained herein.

Appendix

Proof of Lemma 1

Proof It is easy to see that

$$\min_{\Sigma_+\in\mathcal{S}_+,\Sigma_-\in\mathcal{S}_-} \frac{w^\top\Sigma_+ w}{w^\top\Sigma w} = \frac{\min_{\Sigma_+\in\mathcal{S}_+} w^\top\Sigma_+ w}{\min_{\Sigma_+\in\mathcal{S}_+} w^\top\Sigma_+ w + \max_{\Sigma_-\in\mathcal{S}_-} w^\top\Sigma_- w}.$$

for all fixed w. Thus, we will show that

$$\begin{aligned}\max_{\Sigma_-\in\mathcal{S}_-} w^\top\Sigma_- w &= w^\top\left(\overline{\Sigma}_- + \delta_- P_-\right) w,\\ \min_{\Sigma_+\in\mathcal{S}_+} w^\top\Sigma_+ w &= w^\top\left(\overline{\Sigma}_+ - \delta_+ P_+\right) w.\end{aligned}$$

Let $\Delta_+ := \Sigma_+ - \overline{\Sigma}_+$ and $\Delta_- := \Sigma_- - \overline{\Sigma}_-$. Then, the optimization problem can be rewritten as

$$\begin{aligned}&\max \mathrm{Tr}\left(\Delta_- w w^\top\right) \quad \text{s.t. } \mathrm{Tr}\left(P_-^{-1}\Delta_- P_-^{-1}\Delta_-\right) \le \delta_-^2,\\ \Leftrightarrow\ &\max \mathrm{Tr}\left\{\widetilde{\Delta}_- \left(P_-^{1/2} w w^\top P_-^{1/2}\right)\right\}\\ &\quad \text{s.t. } \mathrm{Tr}\left(\widetilde{\Delta}_-^2\right) \le \delta_-^2 \qquad (21)\\ &\min \mathrm{Tr}\left(\Delta_+ w w^\top\right) \quad \text{s.t. } \mathrm{Tr}\left(P_+^{-1}\Delta_+ P_+^{-1}\Delta_+\right) \le \delta_+^2\\ \Leftrightarrow\ &\min \mathrm{Tr}\left\{\widetilde{\Delta}_+ \left(P_+^{1/2} w w^\top P_+^{1/2}\right)\right\}\\ &\quad \text{s.t. } \mathrm{Tr}\left(\widetilde{\Delta}_+^2\right) \le \delta_+^2, \qquad (22)\end{aligned}$$

where $\widetilde{\Delta}_- := P_-^{-1/2}\Delta_- P_-^{-1/2}$ and $\widetilde{\Delta}_+ := P_+^{-1/2}\Delta_+ P_+^{-1/2}$. Symmetric matrices form a metric space with $\langle A, B\rangle = \mathrm{Tr}(AB)$ as the inner product between two symmetric matrices A and B, and $\|A\| = \left\{\mathrm{Tr}(A^2)\right\}^{1/2}$ as the norm. Schwarz's inequality in this case boils down to $\left\{\mathrm{Tr}(AB)\right\}^2 \le \mathrm{Tr}(A^2)\mathrm{Tr}(B^2)$, where the equality holds if and only if $B \propto A$. Therefore, the maximum of Eq. (21) and the minimum of Eq. (22) are attained at $\Delta_-^* = \frac{\delta_-}{w^\top P_- w} P_- w w^\top P_-$ and $\Delta_+^* = -\frac{\delta_+}{w^\top P_+ w} P_+ w w^\top P_+$. The corresponding quadratic forms become

$$\begin{aligned}w^\top\Delta_-^* w &= \delta_-\, w^\top P_- w,\\ w^\top\Delta_+^* w &= -\,\delta_+\, w^\top P_+ w,\end{aligned}$$

which leads to Eq. (9). □

Maxmin theorem for the single filter case

In this section, we will show the minimax theorem, i.e. the order of the maximization over w and the minimization over Σ_+ and Σ_- can be exchanged and both maxmin and minmax problems give the same solution. Let $\Sigma := \Sigma_+ + \Sigma_-$, for convenience. We will write the objective function as

$$R\left(w, \Sigma_+, \Sigma\right) := \frac{w^\top\Sigma_+ w}{w^\top\Sigma w} \qquad (23)$$

and the joint tolerance region as

$$\mathcal{V} := \left\{(\Sigma_+, \Sigma)\,|\Sigma_+ \in \mathcal{S}_+,\ \Sigma \in \mathcal{S}_+ + \mathcal{S}_-\right\}. \qquad (24)$$

The first optimization problem of Eq. (7) can be expressed as

$$\begin{aligned}&\text{maximize} \quad \min_{(\Sigma_+,\Sigma)\in\mathcal{V}} R(w, \Sigma_+, \Sigma)\\ &\text{subject to} \quad w \neq 0\end{aligned} \qquad (25)$$

In the FDA case, the objective function

$$f\left(w, \mu_+, \mu_-, \Sigma_+, \Sigma_-\right) := \frac{\{w^\top(\mu_+ - \mu_-)\}^2}{w^\top(\Sigma_+ + \Sigma_-)w}$$

is convex with respect to the parameters $(\mu_+, \mu_-, \Sigma_+, \Sigma_-)$ for any fixed w. In our CSP case, the tolerance set $\mathcal{V}$ is convex. However, the objective function $R(w, \Sigma_+, \Sigma)$ is neither convex nor concave with respect to Σ_+ and Σ for any fixed $w \neq 0$.

Theorem 1 *Let* $w_{(\Sigma_+,\Sigma)} := \underset{w^\top \Sigma w = 1}{\operatorname{argmax}} R(w, \Sigma_+, \Sigma)$ *and*

$$(\Sigma_+^*, \Sigma^*) := \underset{(\Sigma_+,\Sigma)\in\mathcal{V}}{\operatorname{argmin}} R(w_{(\Sigma_+,\Sigma)}, \Sigma_+, \Sigma). \tag{26}$$

Then, at (Σ_+^*, Σ^*) *and* $w^* := w_{(\Sigma_+^*,\Sigma^*)}$ *the minimax property*

$$\begin{aligned} R\left(w^*, \Sigma_+^*, \Sigma^*\right) &= \max_{w\neq 0} \min_{(\Sigma_+,\Sigma)\in\mathcal{V}} R(w, \Sigma_+, \Sigma) \\ &= \min_{(\Sigma_+,\Sigma)\in\mathcal{V}} \max_{w\neq 0} R(w, \Sigma_+, \Sigma), \end{aligned} \tag{27}$$

and the saddle point property

$$\begin{aligned} R\left(w, \Sigma_+^*, \Sigma^*\right) \leq R(w^*, \Sigma_+^*, \Sigma^*) \leq R(w^*, \Sigma_+, \Sigma), \\ \forall w \in \mathbb{R}^C \backslash \{0\}, \quad \forall (\Sigma_+, \Sigma) \in \mathcal{V}. \end{aligned} \tag{28}$$

hold.

Proof The minimax property (27) can be obtained from the saddle point property (28). The first inequality of Eq. (28) is trivial. Since (Σ_+^*, Σ^*) minimize the generalized eigenvalue $R(w_{(\Sigma_+,\Sigma)}, \Sigma_+, \Sigma)$

$$(w^*)^\top \left\{ (\Sigma_+ - \Sigma_+^*) - \frac{(w^*)^\top \Sigma_+^* w^*}{(w^*)^\top \Sigma^* w^*} (\Sigma - \Sigma^*) \right\} w^* \geq 0 \tag{29}$$

in a neighborhood of (Σ_+^*, Σ^*). Let us take any matrices $(\Sigma_+, \Sigma) \in \mathcal{V}$. Then,
$(\,(1-t)\Sigma_+^* + t\Sigma_+,\ (1-t)\Sigma^* + t\Sigma)$ is included in the neighborhood of (Σ_+^*, Σ^*) for $0 < t \ll 1$. Therefore, from Eq. (29), we have

$$\begin{aligned} &(w^*)^\top \left\{ t(\Sigma_+ - \Sigma_+^*) - \frac{(w^*)^\top \Sigma_+^* w^*}{(w^*)^\top \Sigma^* w^*} t(\Sigma - \Sigma^*) \right\} w^* \geq 0 \\ \Leftrightarrow \quad &\frac{(w^*)^\top \Sigma_+ w^*}{(w^*)^\top \Sigma w^*} \geq \frac{(w^*)^\top \Sigma_+^* w^*}{(w^*)^\top \Sigma^* w^*}. \end{aligned}$$

□

References

Arvaneh, M., Guan, C., Ang, K. K., and Quek, C. (2013). Optimizing spatial filters by minimizing within-class dissimilarities in electroencephalogram-based brain-computer interface. *IEEE Transactions on Neural Networks and Learning Systems*, 24(4):610–619.

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

Blankertz, B., Kawanabe, M., Tomioka, R., Hohlefeld, F. U., Nikulin, V., and Müller, K.-R. (2008a). Invariant common spatial patterns: Alleviating nonstationarities in brain-computer interfacing. In *Advances in Neural Information Processing Systems 20*, pages 113–120. MIT Press.

Blankertz, B., Losch, F., Krauledat, M., Dornhege, G., Curio, G., and Müller, K.-R. (2008b). The Berlin Brain-Computer Interface: Accurate performance from first-session in BCI-naive subjects. *IEEE Transactions on Biomedical Engineering*, 55(10):2452–2462.

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

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

Blumberg, J., Rickert, J., Waldert, S., Schulze-Bonhage, A., Aertsen, A., and Mehring, C. (2007). Adaptive classification for brain computer interfaces. In *Proceedings of the International IEEE Conference of the IEEE Engineering in Medicine and Biology Society*, pages 2536–2539.

Buttfield, A., Ferrez, P., and del R. Millán, J. (2006). Towards a robust bci: error potentials and online learning. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 14(2):164–168.

del R Millán, J. (2003). Brain-computer interfaces. *Handbook of Brain Theory and Neural Networks*.

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

Dyrholm, M., Christoforou, C., and Parra, L. C. (2007). Bilinear Discriminant Component Analysis. *Journal of Machine Learning Research*, 8:1097–1111.

Fukunaga, K. (1990). *Introduction to statistical pattern recognition*. Academic Press, San Diego, 2nd edition.

Grosse-Wentrup, M., Schölkopf, B., and Hill, J. (2011). Causal influence of gamma oscillations on the sensorimotor rhythm. *NeuroImage*, 56(2):837–842.

Kang, H., Nam, Y., and Choi, S. (2009). Composite common spatial pattern for subject-to-subject transfer. *IEEE Signal Processing Letters*, 16(8):683–686.

Kawanabe, M. and Vidaurre, C. (2009). Improving BCI performance by modified common spatial patterns with robustly averaged covariance matrices. In *Proceedings of IWANN 09, Part I, LNCS*, pages 279–282.

Kawanabe, M., Vidaurre, C., Scholler, S., Blankertz, B., and Müller, K.-R. (2009). Robust common spatial filters with a maxmin approach. In *Proceedings of the International IEEE Conference of the IEEE Engineering in Medicine and Biology Society*, pages 2470–2473.

Kim, S.-J., Magnani, A., and Boyd, S. P. (2006). Robust fisher discriminant analysis. In *Advances in Neural Information Processing Systems 18*, pages 659–666. MIT Press.

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

Laub, J. and Müller, K.-R. (2004). Feature discovery in non-metric pairwise data. *Journal of Machine Learning*, 5(Jul):801–818.

Ledoit, O. and Wolf, M. (2004). A well-conditioned estimator for large-dimensional covariance matrices. *Journal of Multivariate Analysis*, 88(2):365–411.

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

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

Lotte, F. and Guan, C. (2010). Learning from other subjects helps reducing Brain-Computer interface calibration time. In *Proceedings of the International IEEE Conference on Acoustics, Speech, and Signal Processing*, pages 614–617.

Lotte, F. and Guan, C. (2011). Regularizing common spatial patterns to improve bci designs: Unified theory and new algorithms. *IEEE Transactions on Biomedical Engineering*, 58(2):355–362.

Lu, H., Eng, H.-L., Guan, C., Plataniotis, K., and Venetsanopoulos, A. (2010). Regularized common spatial pattern with aggregation for eeg classification in small-sample setting. *IEEE Transactions on Biomedical Engineering*, 57(12):2936–2946.

Lu, S., Guan, C., and Zhang, H. (2009). Unsupervised brain computer interface based on intersubject information and online adaptation. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 17(2):135–145.

Montavon, G., Braun, M., Krüger, T., and Müller, K.-R. (2013). Analyzing local structure in kernel-based learning: Explanation, complexity, and reliability assessment. *IEEE Signal Processing Magazine*, 30(4):62–74.

Müller, K.-R., Mika, S., Rätsch, G., Tsuda, K., and Schölkopf, B. (2001). An introduction to kernel-based learning algorithms. *IEEE Transactions on Neural Networks*, 12(2):181–201.

Müller, K.-R., Tangermann, M., Dornhege, G., Krauledat, M., Curio, G., and Blankertz, B. (2008). Machine learning for real-time single-trial EEG-analysis: From brain-computer interfacing to mental state monitoring. *Journal of Neuroscience Methods*, 167(1):82–90.

Nijholt, A., Tan, D., Pfurtscheller, G., Brunner, C., del R Millán, J., Allison, B., Grainmann, B., Popescu, F., Blankertz, B., and Müller, K.-R. (2008). Brain-computer interfacing for intelligent systems. *IEEE Intelligent Systems*, 23(3):72–79.

Ohara, A., Suda, N., and Amari, S. (1996). Dualistic differential geometry of positive definite matrices and its applications to related problems. *Linear Algebra and its Applications*, 247(0):31–53.

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

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

Reuderink, B. (2011). *Robust brain-computer interfaces*. PhD thesis, University of Twente.

Samek, W., Meinecke, F. C., and Müller, K.-R. (2013). Transferring subspaces between subjects in brain-computer interfacing. *IEEE Transactions on Biomedical Engineering*, 60(8):2289–2298.

Samek, W., Müller, K.-R., Kawanabe, M., and Vidaurre, C. (2012a). Brain-computer interfacing in discriminative and stationary subspaces. In *Proceedings of the International IEEE Conference of the IEEE Engineering in Medicine and Biology Society*, pages 2873–2876.

Samek, W., Vidaurre, C., Müller, K.-R., and Kawanabe, M. (2012b). Stationary common spatial patterns for brain-computer interfacing. *Journal of Neural Engineering*, 9(2):026013.

Sannelli, C., Braun, M., and Müller, K.-R. (2009). Improving bci performance by task-related trial pruning. *Neural Networks*, 22(9):1295–1304. Brain-Machine Interface.

Shenoy, P., Krauledat, M., Blankertz, B., Rao, R. P., and Müller, K.-R. (2006). Towards adaptive classification for BCI. *Journal of Neural Engineering*, 3(1):R13–R23.

Sugiyama, M. and Kawanabe, M. (2011). *Machine learning in non-stationary environments: Introduction to covariate shift adaptation*. MIT Press, Cambridge, MA.

Sugiyama, M., Krauledat, M., and Müller, K.-R. (2007). Covariate shift adaptation by importance weighted cross validation. *Journal of Machine Learning Research*, 8:985–1005.

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

Vidaurre, C., Kawanabe, M., von Bünau, P., Blankertz, B., and Müller, K.-R. (2011a). Toward unsupervised adaptation of lda for brain-computer interfaces. *IEEE Transactions on Biomedical Engineering*, 58(3):587–597.

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

Vidaurre, C., Schlögl, A., Cabeza, R., Scherer, R., and Pfurtscheller, G. (2006). A fully on-line adaptive bci. *IEEE Transactions on Biomedical Engineering*, 53(6):1214–1219.

Vidaurre, C., Schlögl, A., Cabeza, R., Scherer, R., and Pfurtscheller, G. (2007). Study of on-line adaptive discriminant analysis for eeg-based brain computer interfaces. *IEEE Transactions on Biomedical Engineering*, 54(3):550–556.

von Bünau, P., Meinecke, F., Scholler, S., and Müller, K.-R. (2010). Finding stationary brain sources in eeg data. In *Proceedings of the International IEEE Conference of the IEEE Engineering in Medicine and Biology Society*, pages 2810–2813.

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

Wang, Y., Hong, B., Gao, X., and Gao, S. (2007). Implementation of a brain-computer interface based on three states of motor imagery. In *Proceedings of the International IEEE Conference of the IEEE Engineering in Medicine and Biology Society*, pages 5059–5062.

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